Received: 21 March 2021 | Revised: 7 November 2021 | Accepted: 15 November 2021

DOI: 10.1111/biom.13611

BIOMETRIC PRACTICE

Biometrics A JOURNAL OF THE INTERNATIONAL BIOMETRIC SOCIETY WILEY

Supervised two-dimensional functional principal component analysis with time-to-event outcomes and mammogram imaging data

Shu Jiang[1] | **Jiguo Cao**[2] | **Bernard Rosner**[3] | **Graham A. Colditz**[1]

[1] Division of Public Health Sciences, Washington University School of Medicine in St. Louis, Missouri

[2] Department of Statistics and Actuarial Science, Simon Fraser University, Canada

[3] Channing Division of Network Medicine, Harvard Medical School, Massachusetts

Correspondence
Shu Jiang, Division of Public Health Sciences, Washington University School of Medicine in St. Louis, MO.
Email: jiang.shu@wustl.edu

Funding information
National Cancer Institute, Grant/Award Number: R37 CA256810; Breast Cancer Research Foundation, Grant/Award Number: BCRF 20-028; Canadian Network for Research and Innovation in Machining Technology, Natural Sciences and Engineering Research Council of Canada, Grant/Award Number: RGPIN-2018-06008

Abstract

Screening mammography aims to identify breast cancer early and secondarily measures breast density to classify women at higher or lower than average risk for future breast cancer in the general population. Despite the strong association of individual mammography features to breast cancer risk, the statistical literature on mammogram imaging data is limited. While functional principal component analysis (FPCA) has been studied in the literature for extracting image-based features, it is conducted independently of the time-to-event response variable. With the consideration of building a prognostic model for precision prevention, we present a set of flexible methods, supervised FPCA (sFPCA) and functional partial least squares (FPLS), to extract image-based features associated with the failure time while accommodating the added complication from right censoring. Throughout the article, we hope to demonstrate that one method is favored over the other under different clinical setups. The proposed methods are applied to the motivating data set from the Joanne Knight Breast Health cohort at Siteman Cancer Center. Our approaches not only obtain the best prediction performance compared to the benchmark model, but also reveal different risk patterns within the mammograms.

KEYWORDS
functional partial least squares, functional principal component analysis, image analysis, risk prediction, survival analysis

1 | INTRODUCTION

Breast cancer is the leading cancer diagnosis among women worldwide, resulting in over 2.1 million new cases diagnosed per year (Bray *et al.*, 2018). While chemoprevention of breast cancer is effective and recommended for high-risk women over 35 years of age (Visvanathan *et al.*, 2019), uptake has been low at around 8% of eligible women in routine clinic settings (Smith *et al.*, 2016), though evidence indicates that benefits will exceed harms for women who have a 5-year risk of 3% or higher (Freedman *et al.*, 2011). Improving risk stratification methods for routine clinical use offers potential to facilitate provider discussion of risk reduction options and the balance of risks and benefits.

Systematic review shows that in models predicting women's risk of breast cancer that were published from 2007 to 2019, the addition of mammographic breast density (BD) significantly increased discriminatory accuracy in 7 of 11 studies (Vilmun *et al.*, 2020). While patterns of breast parenchymal complexity, formed by the x-ray attenuation of fatty, fibroglandular, and stromal tissues, are known to be associated with breast cancer risk (Wolfe, 1976), mammogram BD only aims to measure the relative amount of fibroglandular tissue. Thus, there exist other unexplored features within mammograms, which limits our ability to fully capture the between patients heterogeneity in the breast tissue (Gastounioti *et al.*, 2016). New methods and better use of the extensive existing imaging data can make screening programs more efficient and better adapted to the classification of risk and appropriate screening and risk reduction strategies, ultimately making programs more cost efficient. However, the challenge lies in modeling these high-dimensional mammogram imaging data, since the total number of pixels will greatly exceed the number of subjects in the cohort, making the model nonidentifiable. An efficient dimension reduction method is thus needed.

Functional principal component analysis (FPCA) has been a popular dimension reduction technique for identifying major modes of variation among functional imaging data. Approaches for functional or smooth principal component analysis for tensor imaging data have been proposed in the literature; see Huang *et al.* (2009), Zipunnikov *et al.* (2011), Allen (2013), Lin *et al.* (2015), for example. However, the conventional FPCA does not consider the relationship between the functional predictor and the response variable. Partial least squares (PLS), on the other hand, is a widely adopted alternative in iteratively constructing new explanatory components using linear combinations of original variables to maximize the covariance between these components and the response variable. The functional version of the PLS (FPLS) has also been successfully employed in regression modeling with high-dimensional predictors (Reiss and Ogden, 2007). Due to the presence of right censoring in the survival outcome, there have been several different formulations of the Cox regression model to enable the PLS estimation; see Park *et al.* (2002), Li and Gui (2004), Nygård *et al.* (2008), Bastien *et al.* (2015), for example. To the best of our knowledge, no extension has been proposed under the FPLS framework to accommodate survival outcomes.

As noted above, right censoring is an added complication that is accompanied with time-to-event data. This article introduces a set of flexible methods to accommodate the censored outcomes that combine supervised image feature extraction and risk prediction. Throughout the article, we aim to demonstrate that one method is favored over the other under different clinical setups.

Specifically, the main contributions of this article are as follows. First, we propose a novel supervised FPCA (sFPCA) framework for extracting image features ordered by magnitude of association with the failure time. This newly proposed method is accompanied with an eigenvalue decomposition algorithm that provides a closed-form solution and is computationally efficient. Second, we extend the FPLS framework to accommodate right censored outcomes by utilizing the inverse probability censoring weights (IPCW). Third, we provide intuitions on situations where one method may be preferred over the other, as well as empirical investigations on their finite sample performance via intensive simulation studies. Last, we leverage new insights using the motivating mammogram imaging data from the Joanne Knight Breast Health cohort at Siteman Cancer Center. As the statistical literature addressing methods applied to mammogram imaging data is limited, we provide a detailed description on data preprocessing given the unique shape of the breast boundary. Comparisons across the benchmark model, unsupervised FPCA, FPLS, and sFPCA are demonstrated in this case study.

The rest of this article is organized as follows. In Section 2.1, we first introduce notation, model setup, and the implementation procedures under the sFPCA framework. Extension of the FPLS method to accommodate the censored outcomes is discussed in detail in Section 2.2. Connections with survival analysis utilizing the set of supervised image-based features extracted are made in Section 2.3. We investigate the finite sample performance of the proposed methods via simulation studies under Section 3 and present comprehensive results comparing the conventional FPCA, FPLS, and sFPCA methods. In Section 4, we apply our proposed methods to the motivating data set from the Joanne Knight Breast Health cohort at Siteman Cancer Center. We conclude the article with discussions in Section 5.

2 | SUPERVISED FPCA AND FPLS WITH TIME-TO-EVENT OUTCOMES

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. We let $\mathbb{S}$ be a two-dimensional bounded domain, and $\boldsymbol{s} = (s_1, s_2)$ be a point in $\mathbb{S}$. Then we can define $\{Z_i(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathbb{S}\}$ to be the imaging data for individual i, $i = 1, \ldots, n$.

Under the functional framework, we assume that the observed $Z_i(\boldsymbol{s})$ are realizations of a stochastic process $\{Z(\boldsymbol{s}), \forall \boldsymbol{s} \in \mathbb{S}\}$ in a square integrable rectangle in $\mathbb{R}^2$: $L^2(\mathbb{S}) = \{f : \mathbb{S} \rightarrow \mathbb{R} | \mid \int_{\mathbf{s} \in \mathbb{S}} f(\mathbf{s})^2 d\mathbf{s} \mid < \infty\}$. This is a well-known Hilbert space with inner product defined as $\langle f, g \rangle = \int_{\mathbf{s} \in \mathbb{S}} f(\boldsymbol{s}) g(\boldsymbol{s}) d\boldsymbol{s}$. The mean and the covariance function are assumed to exist. If $\boldsymbol{Z}$ is a square integrable process in space with continuous covariance function, the process has a Karhunen-Loève expansion,

$$Z(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sum_{k=1}^{\infty} \xi_k \phi_k(\boldsymbol{s}) \ , \tag{1}$$

where $\phi_k(\boldsymbol{s})$ is the kth basis function, and $\xi_k = \langle \boldsymbol{Z}, \boldsymbol{\phi}_k \rangle$ are the corresponding scores that are assumed to be uncorrelated random variables with zero mean and finite variance σ_k^2, $k = 1, 2, \ldots$. Without loss of generality, we assume that $\boldsymbol{Z}$ is mean-centered, that is, $\mu(\boldsymbol{s}) = 0$, from here on. In practice, we usually work with the truncated Karhunen-Loève expansion to a relative small dimension K instead of the infinite dimensional object. In the typical case of an unsupervised FPCA method, for example, the ordering of the basis functions is ranked by the values of σ_k such that $\sigma_1^2 \geq \sigma_2^2 \geq \ldots \geq 0$. Therefore, the kth basis function represents the kth major source of variation in the functional data $Z(\boldsymbol{s})$ and may not be associated with the failure time.

To address this concern, we will first propose a supervised framework for the functional principal component analysis (sFPCA) in Section 2.1, and then extend the FPLS method to accommodate censored outcomes in Section 2.2.

2.1 | Supervised FPCA with time-to-event outcomes

Under the supervised framework, we propose to allocate a specific set of orthonormal basis functions $\boldsymbol{\phi} = (\boldsymbol{\phi}_1, \boldsymbol{\phi}_2, \ldots)$, that is, $||\boldsymbol{\phi}|| = 1$, $\langle \boldsymbol{\phi}_k, \boldsymbol{\phi}_{k'} \rangle = 0$ for $k < k'$, such that we can maximize

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

for $0 < \theta \leq 1$. The set of orthonormal basis functions $\boldsymbol{\phi}$ is called the supervised functional principal components (sFPCs). When $\theta = 1$, the objective function $Q(\boldsymbol{\phi})$ is equivalent to the conventional FPCA method. A trade-off between the variance and covariance function is accommodated by setting $\theta \neq 1$. The optimal value of θ can be estimated by conducting cross-validation on a user-specified tuning grid chosen to maximize the prediction accuracy. A suitable truncation to the first K terms can be found by plotting the number of basis functions versus $Q(\boldsymbol{\phi})$, where the point that exhibits an "elbow" can be chosen as a reasonable number.

Due to the presence of right censoring, we propose to reweight the covariance between the observed $\widetilde{T}$ and functional scores using the IPCW. Specifically, for the ith observation, the IPCW can be expressed as, $w_i = \triangle_i / \widehat{G}(\log(\widetilde{T}_i))$, where $\widehat{G}(t) = P(\log(C) > t)$, $i = 1, \ldots, n$ (Welchowski *et al.*, 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,

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

where $\overline{T} = \frac{1}{n} \sum_{i=1}^{n} w_i \log(\widetilde{T}_i)$. This is known to be a consistent estimator (Van der Laan and Robins, 2003).

Next, we introduce an eigenvalue decomposition algorithm to optimize (2) in solving for $\boldsymbol{\phi}$. Here, we assume that the values of θ and K are prespecified. We define the bivariate basis function $\boldsymbol{B}$ as the tensor product between the two univariate basis functions (such as B-splines), $\boldsymbol{B}^{(1)}(s_1)$ of length K_1 and $\boldsymbol{B}^{(2)}(s_2)$ of length K_2, as $\boldsymbol{B}(\boldsymbol{s}) = \boldsymbol{B}^{(1)}(s_1) \otimes \boldsymbol{B}^{(2)}(s_2)$, where $\boldsymbol{s} = (s_1, s_2)$, $\boldsymbol{s} \in \mathbb{S}$, with $\boldsymbol{B}(\boldsymbol{s})$ being dimension $K_+ = K_1 \times K_2$. The imaging observations can then be rewritten as $Z_i(\mathbf{s}) = \boldsymbol{a}_i^T \boldsymbol{B}(\mathbf{s})$, where $\boldsymbol{B}(\mathbf{s})$ denote the column vector $(\boldsymbol{B}_1(\mathbf{s}), \ldots, \boldsymbol{B}_{K_+}(\mathbf{s}))^T$ and $\boldsymbol{a}_i$ denote the coefficients of length $K_+ \times 1$.

We assume that the set of sFPCs can be found by linear combinations of $\boldsymbol{B}(\mathbf{s})$ where it can be rewritten as $\boldsymbol{\phi}_k(\mathbf{s}) = \boldsymbol{b}_k^T \boldsymbol{B}(\mathbf{s})$, $\boldsymbol{b}_k = (b_{k,1}, \ldots, b_{k,K_+})^T$; $\boldsymbol{b} = (\boldsymbol{b}_1, \ldots, \boldsymbol{b}_{K_+})^T$. We can thus rewrite $\langle \boldsymbol{Z}_i, \boldsymbol{\phi} \rangle = \boldsymbol{b}^T \boldsymbol{M} \boldsymbol{a}_i$, where $\boldsymbol{M}$ is of dimension $K_+ \times K_+$, with the $\boldsymbol{M}(k, k') = \langle \boldsymbol{B}_k, \boldsymbol{B}_{k'} \rangle$. The variance can then be estimated empirically as $\text{var}(\langle \boldsymbol{Z}, \boldsymbol{\phi} \rangle) = \frac{1}{n} \boldsymbol{b}^T \boldsymbol{M} \boldsymbol{a}^T \boldsymbol{a} \boldsymbol{M} \boldsymbol{b}$. In the same spirit, the covariance can be estimated as $\text{cov}(\boldsymbol{Y}, \langle \boldsymbol{Z}, \boldsymbol{\phi} \rangle) = \frac{1}{n} \boldsymbol{b}^T \boldsymbol{M} \boldsymbol{a}^T (\boldsymbol{Y} \circ \boldsymbol{w})$, where $\boldsymbol{Y} = ((\log(\widetilde{T_1}) - \overline{T}), \ldots, (\log(\widetilde{T_n}) - \overline{T}))^T$, with $\boldsymbol{w} = (w_1, \ldots, w_n)^T$ being the IPCW weights, and $(\boldsymbol{v} \circ \boldsymbol{x})$ the element-wise multiplication between $\boldsymbol{v}$ and $\boldsymbol{x}$.

Finally, we can reconstruct the objective function in (2) as

$$Q(\boldsymbol{\phi}) = \frac{\boldsymbol{b}^T \boldsymbol{U} \boldsymbol{b}}{\boldsymbol{b}^T \boldsymbol{M} \boldsymbol{b}} \ , \tag{3}$$

where $\boldsymbol{U} = \frac{\theta}{n} \boldsymbol{M} \boldsymbol{a}^T \boldsymbol{a} \boldsymbol{M} + \frac{1-\theta}{n^2} \boldsymbol{M} \boldsymbol{a}^T (\boldsymbol{Y} \circ \boldsymbol{w})(\boldsymbol{Y} \circ \boldsymbol{w})^T \boldsymbol{a} \boldsymbol{M}^T$. Note that maximizing (2) is equivalent to maximizing, $\boldsymbol{\delta}^T (\boldsymbol{M}^{-1/2})^T \boldsymbol{U} \boldsymbol{M}^{-1/2} \boldsymbol{\delta}$ subject to $\boldsymbol{\delta}^T \boldsymbol{\delta} = \boldsymbol{I}$, where $\boldsymbol{\delta} = \boldsymbol{M}^{1/2} \boldsymbol{b}$. We can then estimate $\boldsymbol{\delta}_1, \ldots, \boldsymbol{\delta}_K$, $K \leq K_+$, by finding the leading K eigenvectors of the matrix $(\boldsymbol{M}^{-1/2})^T \boldsymbol{U} \boldsymbol{M}^{-1/2}$. As a result, we are able to estimate $\widehat{\boldsymbol{b}}_k = \boldsymbol{M}^{-1/2} \boldsymbol{\delta}_k$, and consequently the sFPCs as $\widehat{\phi}_k(\mathbf{s}) = \widehat{\boldsymbol{b}}_k^T \boldsymbol{B}(\mathbf{s})$ for $\mathbf{s} \in \mathbb{S}$, $k = 1, \ldots, K$.

2.2 | FPLS with time-to-event outcomes

In this subsection, we discuss an alternative to the proposed sFPCA method where we aim to extend the FPLS method to accommodate time-to-event outcomes that are typically associated with right censoring. Consistent with the proposed sFPCA method, we consider a log transformation of the observed time $\widetilde{T}_i$ such that $\log(\widetilde{T}_i)$ is on a real line. Under the survival framework, $\widetilde{T}_i$ is not observed for all individuals due to the presence of censoring. Ignoring the censored outcomes would give rise to biased estimates. We therefore define a proper outcome, $Y_i^\star = \triangle_i \log(\widetilde{T}_i)/\widehat{G}(\widetilde{T}_i)$ for all individual i, where the observed failure times are reweighted with IPCW similar to the previous section, $i = 1, \dots, n$ (Koul *et al.*, 1981).

Then, under the FPLS framework, we can estimate the first latent component $\xi_1 = \boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}_1$, such that,

$$\begin{aligned}\boldsymbol{r}_1 &= \text{argmax}_{\boldsymbol{r}} \frac{\text{cov}^2(\boldsymbol{\xi}, \boldsymbol{Y}^\star)}{\boldsymbol{r}^T\boldsymbol{r}} \\ &= \text{argmax}_{\boldsymbol{r}} \frac{\boldsymbol{r}^T\boldsymbol{Z}\boldsymbol{B}^T\boldsymbol{Y}^\star\boldsymbol{Y}^{\star T}\boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}}{\boldsymbol{r}^T\boldsymbol{r}}. \qquad (4)\end{aligned}$$

Consistent with the sFPCA framework, we also assume that $\boldsymbol{B}$ is a tensor product of B-splines, but this can be extended to other types of basis functions. The optimization for the kth latent component ξ_k follows that of (4), but is subjective to $\boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}_k \perp \boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}_{k'}$ for $k' < k$. From the setup of (4), we can imagine that as the percentage of censoring increases, solely maximizing the covariance function may lead to suboptimal results given the weak signal. This is discussed further and empirically investigated in simulation studies in the subsequent section.

Under the functional framework, we may also want to consider a roughness penalty function such that,

$$\boldsymbol{r}_1 = \text{argmax}_{\boldsymbol{r}} \frac{\boldsymbol{r}^T\boldsymbol{Z}\boldsymbol{B}^T\boldsymbol{Y}^\star\boldsymbol{Y}^{\star T}\boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}}{\boldsymbol{r}^T(\boldsymbol{1} + \lambda\boldsymbol{P})\boldsymbol{r}}\ , \qquad (5)$$

where $\lambda\boldsymbol{P}$ is positive semidefinite, λ is the tuning parameter, and $\boldsymbol{P}$ denotes the penalty function that penalizes the second-order derivative (Krämer *et al.*, 2008). Similar to (4), the optimization for the kth latent component ξ_k is subjective to $\boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}_k \perp \boldsymbol{Z}\boldsymbol{B}\boldsymbol{r}_{k'}$ for $k' < k$. The optimal value of λ can be found based on the cross-validation method (Ramsay and Silverman, 2005; Reiss and Ogden, 2007). For instance, we can choose λ for a given K such that $\boldsymbol{r}(\lambda(K))$ maximizes the prespecified prediction accuracy measure within the cross-validation study. Estimation of (4) and (5) can be naturally carried out using the well-developed NIPALS or SIMPLS algorithm (Wold, 1975). We note that for univariate responses, both the NIPALS and SIMPLS estimates are equivalent (Martens and Naes, 1992).

2.3 | Survival analysis incorporating imaging data

In this subsection we introduce the linkage between the proposed sFPCA and FPLS with survival analysis. Under the sFPCA framework, we are able to obtain the functional scores of the imaging data using the inner product, $\widehat{\xi}_{ik} = \langle \boldsymbol{Z}_i, \widehat{\boldsymbol{\phi}}_k \rangle$, $k = 1, \dots, K$. With the FPLS method, we are also able to estimate the latent components, $\widetilde{\xi}_{ik} = \boldsymbol{Z}_i\boldsymbol{B}\widehat{\boldsymbol{r}}_k$, $k = 1, \dots, K$. Note that for notation simplicity, we use K to denote the number of components selected under both the sFPCA and FPLS methods, but they may vary. Further, assume that we also have a set of demographic predictors $\boldsymbol{X}_i$ of length $P \times 1$ in addition to the set of estimated $\widehat{\boldsymbol{\xi}}_i = (\widehat{\xi}_{i1}, \dots, \widehat{\xi}_{iK})^T$ or $\widetilde{\boldsymbol{\xi}}_i = (\widetilde{\xi}_{i1}, \dots, \widetilde{\xi}_{iK})^T$. Using $\widehat{\boldsymbol{\xi}}_i$ as an example, we can write the survival distribution at time t as,

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

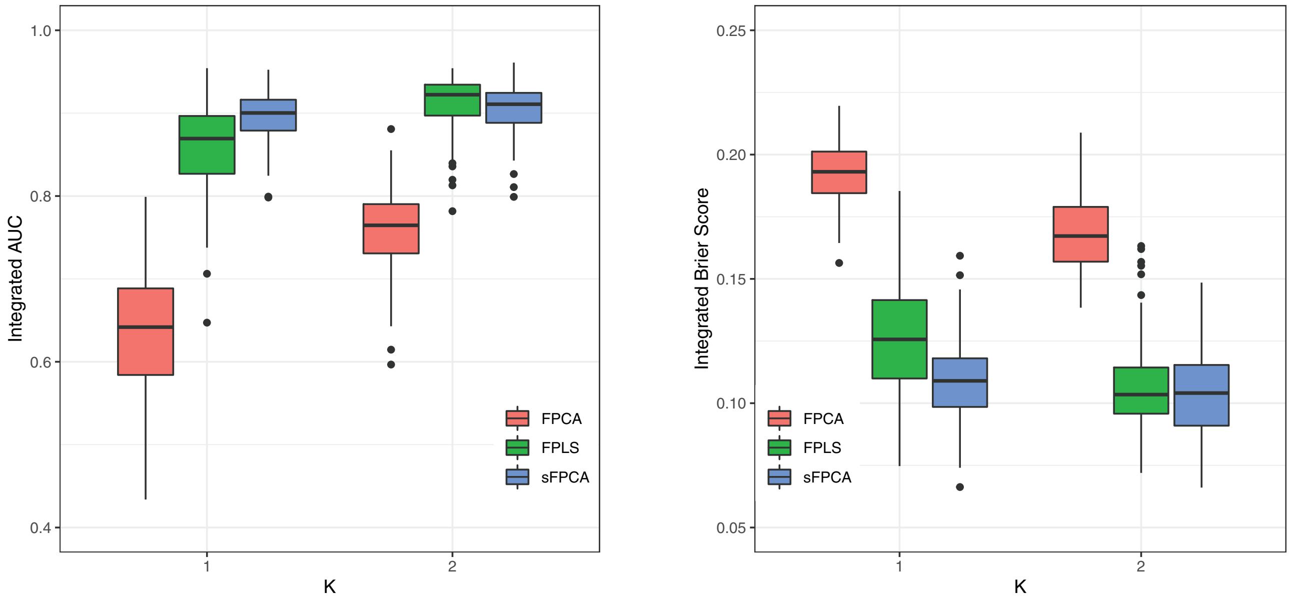
under the proportional hazard assumption, where $S_0(t) = \exp\{-\int_0^t h_0(u)du\}$ with $h_0(t)$ being the baseline hazard function, and $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_P)^T$, $\boldsymbol{\beta} = (\beta_1, \dots, \beta_K)^T$ denoting the regression coefficients.

3 | SIMULATION STUDIES

In this section, we describe our simulation setups and examine the finite sample performance of the proposed methods. We first simulate the 2D eigenfunctions from univariate orthonormal basis functions. Specifically, we simulate $K = 3$ FPCs, where the $\boldsymbol{B}_k$ are formed by tensor products of orthogonal Fourier basis functions on $[-\pi, \pi] \times [-\pi, \pi]$ (Happ and Greven, 2018). The three basis functions are illustrated in Figure S1 within the Supplemental Material.

We then simulate the individual-specific scores $\boldsymbol{a}_i = (a_{i1}, a_{i2}, a_{i3})^T$ with mean 0 and variance of diag(10, 8, 4). The observations are discretized using a grid of 32×32 equidistant points, resulting in a total number of 1024 pixels. Given the orthonormal basis functions and scores, the individual-specific images are generated from the model, $Z_i(\boldsymbol{s}) = \mu(\boldsymbol{s}) + \sum_{k=1}^{3} a_{ik}B_k(\boldsymbol{s})$, where we set $\mu(\boldsymbol{s}) = 0$ without loss of generality, for $\forall \boldsymbol{s} \in \mathbb{S}$.

We further consider a proportional hazards model in this simulation study such that the hazard function is expressed as, $h_i(t) = h_0(t)\exp\{\int_{\boldsymbol{s}\in\mathbb{S}} c(\boldsymbol{s})Z_i(\boldsymbol{s})d\boldsymbol{s}\}$, where the image coefficient is defined as a linear combination of all three basis functions $c(\boldsymbol{s}) = 0.25B_1(\boldsymbol{s}) + 0.5B_2(\boldsymbol{s}) + B_3(\boldsymbol{s})$. The maximum time to the end of the study has been set to 15 years. The baseline hazard function is assumed



(a) AUC

(b) BS

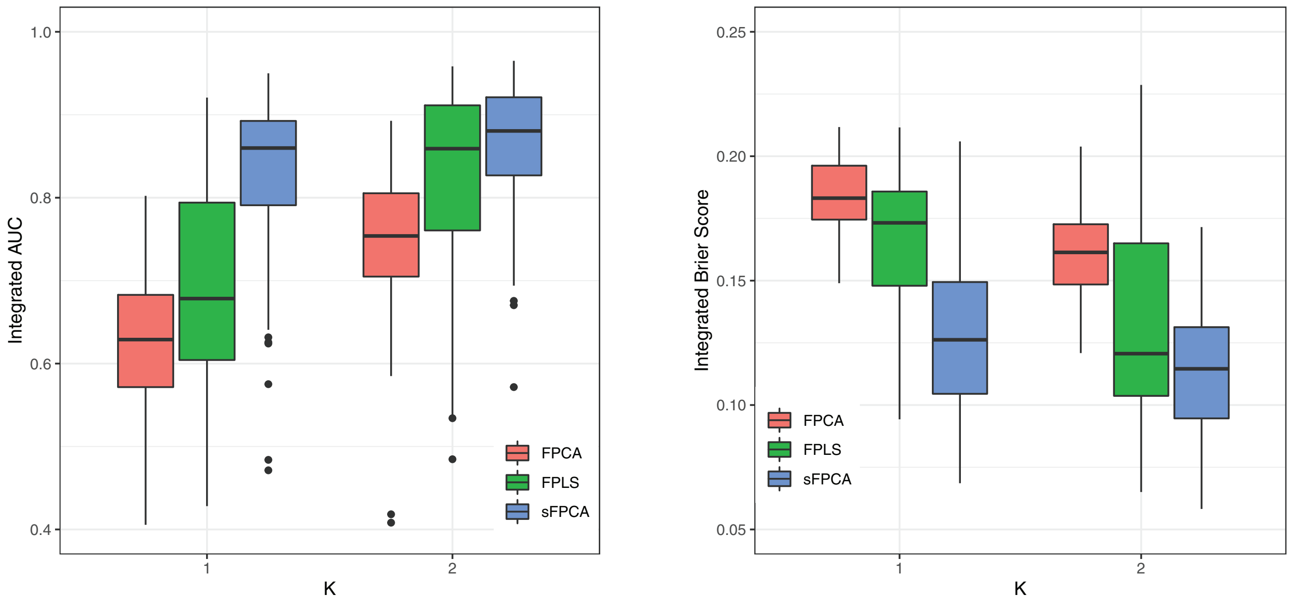
FIGURE 1 The boxplots for the estimated integrated area under the receiver operating characteristic (ROC) curve (AUC) and integrated Brier scores (BS) with FPCA and supervised FPCA under simulation scenario 2 with K number of FPCs/sFPCs; moderate censoring (30%). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

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 $\rho = 0.158$. The failure 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% (moderate) and 60% (heavy).

To avoid overfitting we have simulated 400 individuals per data set, of which 300 are used for training the model and the remaining 100 as validation set. Within the training model, the tuning parameter θ for the sFPCA method was chosen by conducting a fivefold cross-validation on a grid from $1e^{-3}$ to 1 with an equal increment of 0.05. Similarly, a fivefold cross-validation has been conducted to select the smoothness parameter λ from the grid of $(0, 1, 10, 10^2, 10^3)$ for the FPLS method. We have repeated the simulation study 100 times.

Figure 1 illustrates the simulation results under moderate censoring. This figure appears in color in the electronic version of this article, and any mention of color refers to that version. We have assessed both the model discrimination and calibration represented with the integrated area under the receiver operating characteristic (ROC) curve (AUC) (Uno *et al.*, 2007) and integrated Brier scores (Graf *et al.*, 1999; Gerds *et al.*, 2008). For a fair comparison, we have constrained all methods to utilize $K = 1$ followed by two basis functions. Given that the image coefficient surface $c(\mathbf{s})$ is a linear combination of three basis functions with increasing magnitude, we would expect that the conventional FPCA will not pick up the largest effect rising from the third basis function that is associated with the hazard. From Figure 1, we see that the conventional FPCA method (red) indeed retained the worst model discrimination and calibration performance along with bigger standard errors. The FPLS (green) and sFPCA (blue), on the other hand, outperforms the conventional FPCA by picking up the effect of the imaging coefficient that is associated with the failure time. We see that the sFPCA method outperforms FPLS when $K = 1$. But FPLS catches up with sFPCA when K increased to 2.

We further show the simulation results under heavy censoring (60%) in Figure 2. Overall, there is a decrease in prediction accuracy and an increase in the standard error across all three methods as expected. We note that the FPLS method shows the biggest standard error under heavy censoring due to the weak signal, that is, the covariance between the functional imaging predictor and the survival time reweighted by IPCW. On the other hand, we see that the proposed sFPCA retains the superior prediction performance under all settings with comparatively smaller standard error. When $K = 2$, the mean value for θ under moderate censoring is 0.002 (SD = 0.012) while the mean is 0.053 (SD = 0.151) under heavy censoring. This suggests that when the signal is weakened by heavy censoring,



(a) AUC

(b) BS

FIGURE 2 The 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 2 with K number of FPCs/sFPCs; heavy censoring (60%). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

the proposed trade-off structure between the variance and covariance function is more crucial. From a Monte Carlo study with 10,000 simulated individuals, we have empirically estimated the covariance between functional imaging predictor and failure time reweighted by IPCW at 0%, 30%, and 60% censoring rate. The signal, in this case the covariance with the outcome, has weakened by 55% when the censoring rate is 30%, and by 85% at 60% censoring rate as compared to no censoring. Note that we cannot directly infer that the covariance term within the trade-off structure (2) has dominated the optimization if θ is close to 0. This may be due to the difference in magnitude in the variance in relation to the covariance function and should be empirically investigated in the real data application for a reasonable interpretation.

In addition to the simulation results discussed above, we have also investigated the prediction performance under model misspecification. The prediction accuracy decreased across all methods as expected. The pattern of the results under different methods are similar to what we see here in this section. Interestingly, we see more variability in the FPLS method due to model misspecification. We have also compared the performance of the proposed method with the random survival forest. Further, we have done a plasmode simulation study mimicking the mammography data setting. Similar conclusions can be drawn under the plasmode simulation study with the setting discussed in this section where the proposed method retains superior prediction performance. Details on the setup and simulation results for model misspecification, comparison with the random survival forest, and plasmode simulation study are available in the online Supplemental Material.

4 | APPLICATION TO THE JOANNE KNIGHT BREAST HEALTH COHORT AT SITEMAN CANCER CENTER

The data that motivated this study come from the Joanne Knight Breast Health Cohort at Siteman Cancer Center, Washington University School of Medicine in St. Louis (Colditz *et al.*, 2022). This cohort was established to link breast cancer risk factors, mammographic BD, and blood markers in a diverse population of women undergoing routine mammographic screening to estimate breast cancer incidence. The service uses the Hologic machines and provides the same screening for women regardless of insurance status and ability to pay. The same screening protocol and follow-up protocol are used across all women screened. Women were recruited in St. Louis, MO, from November 2008 to April 2012, and have been followed through October 2020. We have included 785 women who had a full field digital craniocaudal (CC) view mammogram available at the baseline. This subsample of women came from the case and control subcohort. Cases were diagnosed after the entry mammogram and blood draw

TABLE 1 Estimated median 5-year integrated receiver operating characteristic (ROC) curve (AUC) and integrated Brier Score (BS) with the empirical standard errors across the 10-fold cross-validation. Results are displayed under the benchmark model (only includes baseline age, BMI, MP, BD, age×MP, and BMI×MP), unsupervised FPCA, FPLS, and sFPCA method

	Partial likelihood				**Weighted partial likelihood**			
	AUC	**SE**	**BS**	**SE**	**AUC**	**SE**	**BS**	**SE**
Benchmark	0.628	0.107	0.067	0.021	0.613	0.089	0.067	0.021
FPCA	0.656	0.101	0.066	0.023	0.650	0.082	0.067	0.022
FPLS	0.668	0.103	0.066	0.023	0.662	0.098	0.067	0.022
sFPCA	0.685	0.095	0.066	0.022	0.679	0.090	0.067	0.021

and matched on age and year of entry to women who remained free from breast cancer. Of the 785 women, 246 have been diagnosed with breast cancer prior to the end of follow-up. Those who had a diagnosis of breast cancer within half-year of entry have been excluded as we are focused on risk prediction instead of diagnosis.

We have done some preprocessing prior to analyzing these imaging data. There exist several well-developed automated methods for preprocessing the mammograms; see Ou *et al.* (2011), Brandt *et al.* (2011), and Lee and Nishikawa (2019) for example. To minimize the noise caused by distinct breast size and position within the mammograms, we followed the approach proposed in Lee and Nishikawa (2019). Specifically, the breast area is first segmented using a tight rectangular box followed by soft tissue removal for parts not in the breast. Soft tissue in CC-view images includes parts of nonbreast tissue either above or below the true breast area, and these were removed prior to analysis. Then, each mammogram is resized to 500 × 800 pixels using the bicubic interpolation. After these preprocessing procedures, the breasts depicted in the mammograms were similar in size and orientation as illustrated within Section S2 of the Supplemental Material. As such, the pixel intensities were averaged between the mammogram images for the left and the right breast, which is a common practice for estimating mammogram density in the literature.

We consider the Cox proportional hazards model in this analysis,

$$h_i(t) = h_0(t)\exp\big(\alpha_1 \text{age}_i + \alpha_2 \text{BD}_i + \alpha_3 \text{MP}_i + \boldsymbol{\alpha}_4^T \text{BMI}_i + \boldsymbol{\alpha}_5^T \text{MP}_i * \text{BMI}_i + \alpha_6 \text{MP}_i * \text{age}_i + \boldsymbol{\beta}^T \widehat{\boldsymbol{\xi}}_i\big),$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_K)^T$ are the coefficients for the first K scores estimated from the proposed sFPCA $\widehat{\boldsymbol{\xi}}_i$ or FPLS $\widetilde{\boldsymbol{\xi}}_i$. The demographic variables that we consider in the proportional hazards model include baseline age, menopausal status ($MP_i = 1$ for postmenopausal woman), BD by treating the BIRADS mammogram density as a continuous variable, and three levels of BMI corresponding to (1) normal (BMI < 25), (2) overweight (BMI $\in [25, 30)$), and (3) obese (BMI ≥ 30). Thus, we have $\boldsymbol{\alpha}_4$ and $\boldsymbol{\alpha}_5$ as the vector of two coefficients, using normal BMI as the reference. It is crucial that model checking is carried out prior to adopting the proportional hazards model to prevent utilizing a poor model (Fleming and Harrington, 2011). The proportional hazards assumption is formally checked by inspecting the Schoenfeld residual plot for each of the baseline covariates and was deemed reasonable in this case study.

In this analysis, all mammograms have been mean-centered prior to feature extraction. We have adopted a 12×12 tensor product cubic B-spline basis functions $\boldsymbol{B}$ in this application. A 10-fold internal cross-validation was adopted to avoid overfitting. Under the sFPCA approach, the tuning parameter θ was selected from an equal-distance grid ranging from $1e^{-3}$ to 1 with an increment of 0.01 in a nested fivefold cross-validation. Similarly, a nested fivefold cross-validation has been conducted to select the smoothness parameter λ under the FPLS method from the grid of $(0, 1, 10, 1e^2, 1e^3, 1e^6)$. As discussed in Section 2, the number of components (K) can be chosen by looking at an elbow plot. An alternative approach is to adopt a two-dimensional grid search for K. For a given K, we can select an optimal $\theta(K)$ and $\lambda(K)$ such that it maximizes the prediction accuracy measure.

We now illustrate the prediction performance under different models in Table 1. Note that we have reported the prediction performances under both the standard Cox model with partial likelihood as well as the Cox model with a weighted partial likelihood due to the nested case control subcohort in this analysis (Samuelsen, 1997; Støer and Samuelsen, 2013). We start with the benchmark model where we only include baseline age, BMI, MP, BD, age×MP, and BMI×MP, without using any information from the images; this gives us the lowest estimated median 5-year AUC relative to competing methods (see Table 1). We then assessed the gain in model discrimination and calibration by adding in features extracted from the conventional FPCA, FPLS, and sFPCA. We can see that all three models achieved higher AUC when compared to the benchmark model. These results suggest that the extracted features from mammogram images can provide complementary information to BD in risk prediction.

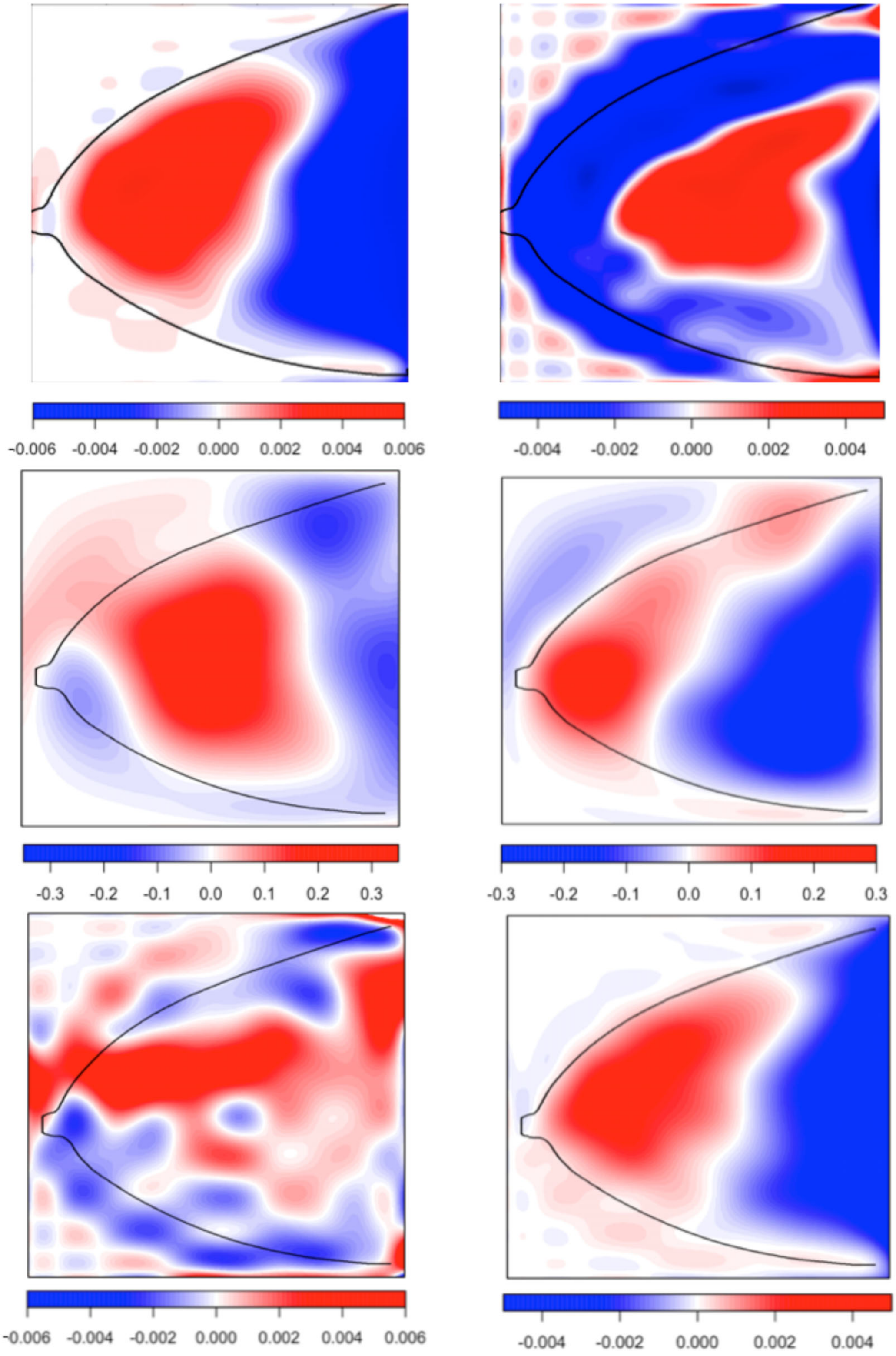


FIGURE 3 The first two components estimated with the conventional FPCA (top row), FPLS (middle row), and supervised FPCA (bottom row). This figure appears in color in the electronic version of this article, and any mention of color refers to that version

Nevertheless, it is apparent that the sFPCA retains a superior prediction performance under all submodels presented in Table 1. It is not surprising that the FPLS did not outperform the sFPCA due to the high censoring rate (66.7%); this behavior is consistent with the observations in our simulation studies.

Finally, we give some insights in Figure 3 on the first two estimated basis functions under FPCA (top row), FPLS (middle row), and sFPCA (bottom row). We have superimposed the border of the reference breast onto these plots to avoid overinterpreting the boundary effects. This figure appears in color in the electronic version of this article, and any mention of color refers to that version. In Figure 3, because the red regions represent positive estimates and the blue regions represent negative estimates, the corresponding scores or latent components can be interpreted as the weighted difference of mean-centered mammograms between the red and blue regions. Taking the sFPCA method as an example, because the kth score is represented by $\xi_k = \langle \boldsymbol{Z}, \boldsymbol{\phi}_k \rangle = \int_{\boldsymbol{s}\in\mathbb{S}} Z(\boldsymbol{s})\phi_k(\boldsymbol{s})d\boldsymbol{s}$, the term ξ_k can be interpreted as a weighted average of the mean-centered image $\boldsymbol{Z}$ with $\phi_k(\boldsymbol{s})$ being the weight, $k = 1, 2, \ldots, K$. Thus, the hazard function is affected by the difference of the mean-centered mammogram at the red regions in comparison with the blue regions. Notice that the first estimated basis function from the conventional FPCA method is close to the second estimated basis functions from the FPLS and sFPCA method, telling us that both of the FPLS and sFPCA also consider this pattern of risk to be associated with the hazard function.

5 | CONCLUSION AND DISCUSSION

Patterns of breast parenchymal complexity, formed by the x-ray attenuation of fatty, fibroglandular, and stromal tissues, are known to be associated with breast cancer (Wolfe, 1976). Mammogram density is one of the most well-known risk factors for breast cancer that aims to measure the relative amount of fibroglandular tissue in the breast, which limits our ability to fully capture the between-patient heterogeneity in the breast. Motivated by data from the Joanne Knight Breast Health cohort at Siteman Cancer Center, we propose a set of flexible methods, that is, sFPCA and FPLS, for image feature extraction while accommodating censored outcomes. We investigate the empirical performances of the proposed methods through comprehensive simulation studies and show that both of the proposed methods retain superior prediction performance in comparison to the conventional unsupervised FPCA. We have demonstrated that when the percentage of censoring is high and when the model is misspecified, sFPCA may be more robust than the FPLS method.

Application to the Joanne Knight Breast Health cohort has also revealed important insights. We have shown the sFPCA method has achieved superior prediction performance over the benchmark model, FPCA, and the FPLS method under a 10-fold internal cross-validation study. This finding suggests that we are able to refine the individual-specific risks using the additional image-based features. These newly defined features within mammograms could better discriminate women who are at higher and lower risk, which will facilitate tailored screening, thereby reducing unnecessary imaging and health care cost (Pashayan *et al.*, 2018).

While we have proposed using the proportional hazards model in this article, our modeling framework can be directly extended to other types of survival setups. For instance, the features extracted from the proposed methods can be directly used under the random survival forest (Ishwaran *et al.*, 2008; Jiang *et al.*, 2021) as shown in one of our simulation studies. Other radiomic feature-based methods such as deep neural networks can also be adopted (Wu *et al.*, 2019; McKinney *et al.*, 2020). As noted in our simulation studies, one major disadvantage for these "black-box" methods is that we will not be able to visualize the coefficient surface for the imaging data because the inner working of these methods are not transparent. Study involving an independent validation data set will remain as part of our future work.

ACKNOWLEDGMENTS

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

DATA AVAILABILITY STATEMENT

The data that support the findings in this article are available on request from Dr. Graham A. Colditz (colditzg@wustl.edu). The data are not publicly available due to privacy or ethical restrictions.

ORCID

Shu Jiang https://orcid.org/0000-0003-1464-4838
Jiguo Cao https://orcid.org/0000-0001-7417-6330

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)*, pp. 220–223. IEEE.

Bastien, P., Bertrand, F., Meyer, N. and Maumy-Bertrand, M. (2015) Deviance residuals-based sparse PLS and sparse kernel PLS regression for censored data. *Bioinformatics*, 31, 397–404.

Brandt, S.S., Karemore, G., Karssemeijer, N. and Nielsen, M. (2011) An anatomically oriented breast coordinate system for mammogram analysis. *IEEE Transactions on Medical Imaging*, 30, 1841–1851.

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.

Colditz, G.A., Bennett, D.L., Tappenden, J., Beers, C., Ackermann, N., Wu, N., et al. (2022) Joanne knight breast health cohort at siteman cancer center. *Cancer Causes & Control*, pp. 1–7.

Fleming, T.R. and Harrington, D.P. (2011) *Counting Processes and Survival Analysis*, volume 169. New York: John Wiley & Sons.

Freedman, A.N., Yu, B., Gail, M.H., Costantino, J.P., Graubard, B.I., Vogel, V.G., et al. (2011) Benefit/risk assessment for breast cancer chemoprevention with raloxifene or tamoxifen for women age 50 years or older. *Journal of Clinical Oncology*, 29, 2327.

Gastounioti, A., Conant, E.F. and Kontos, D. (2016) Beyond breast density: a review on the advancing role of parenchymal texture analysis in breast cancer risk assessment. *Breast Cancer Research*, 18, 1–12.

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.

Ishwaran, H., Kogalur, U.B., Blackstone, E.H. and Lauer, M.S. (2008) Random survival forests. *The Annals of Applied Statistics*, 2, 841–860.

Jiang, S., Xie, Y. and Colditz, G.A. (2021) Functional ensemble survival tree: dynamic prediction of Alzheimer's disease progression accommodating multiple time-varying covariates. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 70, 66–79.

Koul, H., Susarla, V. and Van Ryzin, J. (1981) Regression analysis with randomly right-censored data. *The Annals of Statistics*, 9, 1276–1288.

Krämer, N., Boulesteix, A.-L. and Tutz, G. (2008) Penalized partial least squares with applications to B-spline transformations and functional data. *Chemometrics and Intelligent Laboratory Systems*, 94, 60–69.

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, H. and Gui, J. (2004) Partial Cox regression analysis for high-dimensional microarray gene expression data. *Bioinformatics*, 20, i208–i215.

Lin, N., Jiang, J., Guo, S. and Xiong, M. (2015) Functional principal component analysis and randomized sparse clustering algorithm for medical image analysis. *PLoS ONE*, 10, e0132945.

Martens, H. and Naes, T. (1992) *Multivariate Calibration*. New York: John Wiley & Sons.

McKinney, S.M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., et al. (2020) International evaluation of an AI system for breast cancer screening. *Nature*, 577, 89–94.

Nygård, S., Borgan, Ø., Lingjærde, O.C. and Størvold, H.L. (2008) Partial least squares Cox regression for genome-wide data. *Lifetime Data Analysis*, 14, 179–195.

Ou, Y., Sotiras, A., Paragios, N. and Davatzikos, C. (2011) DRAMMS: deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15, 622–639.

Park, P.J., Tian, L. and Kohane, I.S. (2002) Linking gene expression data with patient survival times using partial least squares. *Bioinformatics*, 18, S120–S127.

Pashayan, N., Morris, S., Gilbert, F.J. and Pharoah, P.D. (2018) Cost-effectiveness and benefit-to-harm ratio of risk-stratified screening for breast cancer: a life-table model. *The Journal of the American Medical Association Oncology*, 4, 1504–1510.

Ramsay, J.O. and Silverman, B.W. (2005) *Functional Data Analysis*, 2nd edition. New York: Springer.

Reiss, P.T. and Ogden, R.T. (2007) Functional principal component regression and functional partial least squares. *Journal of the American Statistical Association*, 102, 984–996.

Samuelsen, S.O. (1997) A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika*, 84, 379–394.

Smith, S.G., Sestak, I., Forster, A., Partridge, A., Side, L., Wolf, M., et al. (2016) Factors affecting uptake and adherence to breast cancer chemoprevention: a systematic review and meta-analysis. *Annals of Oncology*, 27, 575–590.

Støer, N.C. and Samuelsen, S.O. (2013) Inverse probability weighting in nested case-control studies with additional matching—a simulation study. *Statistics in Medicine*, 32, 5328–5339.

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.

Vilmun, B.M., Vejborg, I., Lynge, E., Lillholm, M., Nielsen, M., Nielsen, M.B., et al. (2020) Impact of adding breast density to breast cancer risk models: a systematic review. *European Journal of Radiology*, 127, 109019.

Visvanathan, K., Fabian, C.J., Bantug, E., Brewster, A.M., Davidson, N.E., DeCensi, A., et al. (2019) Use of endocrine therapy for breast cancer risk reduction: ASCO clinical practice guideline update. *Journal of Clinical Oncology*, 37, 3152–3165.

Welchowski, T., Zuber, V. and Schmid, M. (2019) Correlation-adjusted regression survival scores for high-dimensional variable selection. *Statistics in Medicine*, 38, 2413–2427.

Wold, H. (1975) Soft modelling by latent variables: the non-linear iterative partial least squares (NIPALS) approach. *Journal of Applied Probability*, 12, 117–142.

Wolfe, J.N. (1976) Breast patterns as an index of risk for developing breast cancer. *American Journal of Roentgenology*, 126, 1130–1137.

Wu, N., Phang, J., Park, J., Shen, Y., Huang, Z., Zorin, M., et al. (2019) Deep neural networks improve radiologists' performance in breast cancer screening. *IEEE Transactions on Medical Imaging*, 39, 1184–1194.

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.

SUPPORTING INFORMATION

Web Appendices and Figures referenced in Sections 3 and 4 are available with this paper at the Biometrics website on Wiley Online Library. Software in the form of R code, together with a sample input data set and complete documentation is available on github repository (https://github.com/jj113/2D-sFPCA-FPLS).

How to cite this article: Jiang, S., Cao, J., Rosner, B., Colditz, G.A. Supervised two-dimensional functional principal component analysis with time-to-event outcomes and mammogram imaging data. *Biometrics*. 2022;1–11. https://doi.org/10.1111/biom.13611