DOI: 10.1111/biom.128

Exponential Family Functional Data Analysis via a Low-Rank Model

Gen Li[1,*] **Jianhua Z. Huang,**[2] **and Haipeng Shen**[3]

[1]Department of Biostatistics, Mailman School of Public Health, Columbia University, New York, U.S.A.
[2]Department of Statistics, Texas A&M University, Texas, U.S.A.
[3]Faculty of Business and Economics, University of Hong Kong, Hong Kong, China
email: gl2521@cumc.columbia.edu

Summary. In many applications, non-Gaussian data such as binary or count are observed over a continuous domain and there exists a smooth underlying structure for describing such data. We develop a new functional data method to deal with this kind of data when the data are regularly spaced on the continuous domain. Our method, referred to as Exponential Family Functional Principal Component Analysis (EFPCA), assumes the data are generated from an exponential family distribution, and the matrix of the canonical parameters has a low-rank structure. The proposed method flexibly accommodates not only the standard one-way functional data, but also two-way (or bivariate) functional data. In addition, we introduce a new cross validation method for estimating the latent rank of a generalized data matrix. We demonstrate the efficacy of the proposed methods using a comprehensive simulation study. The proposed method is also applied to a real application of the UK mortality study, where data are binomially distributed and two-way functional across age groups and calendar years. The results offer novel insights into the underlying mortality pattern.

Key words: Functional principal component analysis; Generalized linear model; Mortality study; Singular value decomposition; Two-way functional data

1. Introduction

Functional data are frequently encountered in biomedical sciences, such as the study of temporal electrical activities in electroencephalogram and kinematic trajectories of planar reaching motions. Information is available over a continuum of time or location, while data are typically sampled at discrete time points. Functional Principal Component Analysis (FPCA) is commonly used to reduce the dimension and estimate the intrinsic structure of functional data (see Ramsay and Silverman, 2005; Ferraty and Vieu, 2006, and references therein). However, most existing FPCA methods assume observed functional data are continuous and follow a Gaussian distribution. In practice, observations at discrete time points may be binary or count-valued, despite smooth underlying structure. In this article, our goal is to generalize FPCA to non-Gaussian data.

In our motivating example, the UK mortality study, we aim to understand both the age effect and the historical trend of the mortality rate in UK. The available data are the population size and death count for each age group in each calendar year, sampled on a two-dimensional, dense, and regularly spaced grid. Since similar age groups or adjacent years should have similar underlying mortality rates, it is reasonable to view the mortality rate as a *smooth* bivariate function of age and calendar year (if both are treated as continuous variables). Such a viewpoint allows us to borrow strength across ages and years to estimate the underlying mortality structure. One may use FPCA to describe the age effect on and historical trend of the mortality rate. However, there are two major issues we need to address: on the one hand, the observed data are binomially distributed and not continuous; on the other hand, the underlying mortality rate function is smooth in both time domains –that is, we have two-way functional data (Huang et al., 2009).

Several FPCA methods have been developed to deal with non-Gaussian functional data. For example, Hall et al. (2008) proposed a Latent Gaussian Process (LGP) model to analyze irregularly spaced non-Gaussian observations, where the relations between longitudinal measurements are characterized by an underlying Gaussian process. Later, the method was generalized to rare events (Serban et al., 2013) and multilevel data (Goldsmith et al., 2015). More recently, Gertheiss et al. (2017) developed a Bayesian generalized FPCA (gfPCA) approach for modeling sparsely observed functional response curves. The method was only implemented for binary data. Although these methods can deal with non-Gaussian functional data, none of them accommodates two-way functional data. In addition, the computational cost is usually very high when the methods are applied to regularly spaced and dense data as in our motivating example.

To model two-way functional data, Huang et al. (2009) proposed a Functional Singular Value Decomposition (FSVD) method, which achieves structured estimation of the latent patterns in two domains via regularized matrix decomposition. A robust generalization which accommodates excessive outliers was proposed in Zhang et al. (2013). In particular, both studies investigated mortality rate patterns across ages and years. However, since both methods are only applicable to

Gaussian data, raw death counts and population sizes need to be processed and/or transformed before fed in the methods. The processing step neglects important distributional information such as heterogeneous variances of different values, and may lead to misleading conclusions. A more principled method is needed to deal with two-way non-Gaussian functional data.

In this article, we develop a new method, called Exponential Family FPCA (EFPCA), for analyzing regularly spaced, dense, and generalized functional data with one-way or two-way smoothness. We assume observed data are in the matrix form. We model each entry of the observed data matrix with a single-parameter exponential family distribution, which fully accounts for the nature of the data, and consider a low-rank and regularized decomposition of the underlying natural parameter matrix. We exploit an efficient penalized likelihood approach for model fitting. In addition, we devise a new rank selection method for estimating the latent rank of a generalized data matrix, which may have broad application beyond EFPCA. Estimated model parameters characterize the underlying smooth structure, and can be used for further statistical analyses such as prediction and inference. We apply the proposed method to the UK mortality data, and obtain interesting findings of the age effect on and historical trend of the mortality rate. To sum up, the main contributions of the article are as follows.

- We develop a new FPCA method for analyzing one/two-way generalized functional data, measured on a regularly spaced and dense grid.
- We devise a broadly applicable rank selection method for a non-Gaussian data matrix.
- The application to the UK mortality study offers novel insights and may stimulate further social studies.

The rest of the article is organized as follows. In Section 2, we introduce the EFPCA model and describe a penalized likelihood framework. In Section 3, we elaborate the model fitting algorithm. In Section 4, we introduce a cross validation method for estimating the latent rank of a generalized data matrix, and discuss the estimation of other tuning parameters. The simulation results and the UK mortality study are presented in Section 5 and 6. We conclude the article with discussions in Section 7. Technical details can be found in the online Supplementary Material.

2. Model

2.1. *Model Setup*

We focus on generalized functional data measured on a regular and dense grid. Let $\boldsymbol{X}$ be an $n \times p$ observed data matrix, with underlying smooth structure. In the classic one-way functional data analysis setting, n is the number of samples and p is the number of regular sampling points. In a two-way setting, both the row variables and column variables are discrete samples of some smooth latent process. Correspondingly, n and p both represent the numbers of regular sampling points in different domains, respectively. Let X_{ij} be the (i, j)th value of $\boldsymbol{X}$. We assume X_{ij} follows a single-parameter exponential family distribution (e.g., Poisson, Bernoulli, Gaussian, Binomial). In particular, if each entry of $\boldsymbol{X}$ follows a Binomial distribution, let $\boldsymbol{M}$ be a matched $n \times p$ matrix with the total numbers of trials. Each variable has the following probability density function:

$$f(x_{ij} \mid \theta_{ij}) = h(x_{ij}) \exp\left\{x_{ij}\theta_{ij} - b(\theta_{ij})\right\},$$

where θ_{ij} is the underlying natural parameter, $b(\cdot)$ is a distribution-specific convex function, and $h(\cdot)$ is a normalization function making the integration of the density function equal to 1. The first order derivative $b'(\theta_{ij})$ is equal to the expectation $\mathbb{E}(X_{ij})$ (denoted by μ_{ij}), and $g(\mu_{ij}) = b'^{-1}(\mu_{ij})$ is the canonical link function. Some commonly used exponential family distributions are listed in Table 1 with corresponding functions and parameters. We collect the underlying natural parameters in an $n \times p$ matrix $\boldsymbol{\Theta}$, which is the building block of the EFPCA model.

The fundamental assumption of the EFPCA model is that the natural parameter matrix $\boldsymbol{\Theta}$ is a collection of discrete observations of a one/two-way smooth bivariate function on a regular grid. More specifically, for one-way functional data, we consider a bivariate function $\Theta(i, s)$, where $i \in \{1, \cdots, n\}$ is a discrete index and s is a continuous index in a bounded domain S. The function is smooth with respect to s. For two-way functional data, we consider a bivariate function

Table 1
Functions and parameters for some commonly used exponential family distributions

	Mean μ	Natural param θ	$b(\theta)$	$g(\mu)$
Nature (with known variance)	μ	μ	$\frac{\theta^2}{2}$	μ
Poisson	λ	$\log\lambda$	$\exp(\theta)$	$\log(\mu)$
Bernoulli	p	$\log\frac{p}{1-p}$	$\log\{1+\exp(\theta)\}$	$\log\frac{\mu}{1-\mu}$
Binomial (with m trials)	mp	$\log\frac{p}{1-p}$	$m\log\{1+\exp(\theta)\}$	$\log\frac{\mu}{m-\mu}$

$\Theta(t,s)$, where t and s are continuous indices in bounded domains T and S, respectively. The function is smooth with respect to both t and s. The natural parameter matrix $\boldsymbol{\Theta}$ is the discretization of $\Theta(i,s)$ or $\Theta(t,s)$.

By the Karhunen–Loève theorem, we can approximate either bivariate function with a finite sum of the product of some univariate functions. Correspondingly, in the discrete setting, we assume $\boldsymbol{\Theta}$ has a low-rank (rank $r < \min(n,p)$) singular value decomposition (SVD)

$$\boldsymbol{\Theta} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T, \tag{1}$$

where $\boldsymbol{U}$ and $\boldsymbol{V}$ are $n \times r$ and $p \times r$ matrices with r left and right singular vectors respectively, and $\boldsymbol{D}$ is an $r \times r$ diagonal matrix with positive non-increasing singular values on the diagonal. Without loss of generality, hereafter we always assume $\boldsymbol{U}$ absorbs the singular values in $\boldsymbol{D}$.

In particular, for one-way functional data, each column of $\boldsymbol{V}$ is the discretization of some smooth function (Huang et al., 2008); for two-way functional data, each column of $\boldsymbol{V}$ and $\boldsymbol{U}$ is the discretization of some smooth function (Huang et al., 2009). As a result, we can write the matrix form of the proposed model as

$$g\{\mathbb{E}(\boldsymbol{X})\} = \boldsymbol{U}\boldsymbol{V}^T, \tag{2}$$

where $\mathbb{E}(\cdot)$ and $g(\cdot)$ are the entry-wise expectation operator and canonical link function, and the smoothness of underlying structure is captured by structured $\boldsymbol{V}$ (and $\boldsymbol{U}$).

2.2. *Penalized Likelihood*

In order to estimate the parameters $\{\boldsymbol{U}, \boldsymbol{V}\}$ in Model (2), we exploit a penalized likelihood approach. We introduce regularization to the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ to incorporate one/two-way smoothness. We particularly assume that discretized data points in $\boldsymbol{X}$ are independent, given the underlying natural parameter matrix $\boldsymbol{\Theta}$. The dependency between variables in both domains is fully characterized by the structure of $\boldsymbol{\Theta}$. Such assumption has been widely used in the literature (Collins et al., 2001; Hall et al., 2008; Serban et al., 2013; Goldsmith et al., 2015). As a result, the general form of the penalized likelihood function of the observed data matrix $\boldsymbol{X}$ can be expressed as

$$\mathcal{C}(\boldsymbol{U}, \boldsymbol{V}) = -\mathcal{L}(\boldsymbol{X} \mid \boldsymbol{\Theta}) + \mathcal{P}_{\boldsymbol{\lambda}}(\boldsymbol{U}, \boldsymbol{V}), \tag{3}$$

where $\mathcal{L}(\boldsymbol{X}|\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{j=1}^{p} \log f(x_{ij}|\theta_{ij})$ is the joint log likelihood function, and $\mathcal{P}_{\boldsymbol{\lambda}}(\boldsymbol{U}, \boldsymbol{V})$ is a roughness penalty on the columns of $\boldsymbol{U}$ and $\boldsymbol{V}$ with smoothing parameters in $\boldsymbol{\lambda}$. We particularly focus on the following penalty function:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\boldsymbol{U}, \boldsymbol{V}) = \frac{1}{2}\sum_{k=1}^{r} \lambda_k \boldsymbol{v}_k^T \boldsymbol{\Omega}_{\boldsymbol{v}} \boldsymbol{v}_k + \frac{1}{2}\sum_{k=1}^{r} \alpha_k \boldsymbol{u}_k^T \boldsymbol{\Omega}_{\boldsymbol{u}} \boldsymbol{u}_k, \tag{4}$$

where $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$ are the kth columns of $\boldsymbol{V}$ and $\boldsymbol{U}$; $\boldsymbol{\Omega}_{\boldsymbol{v}}$ and $\boldsymbol{\Omega}_{\boldsymbol{u}}$ are prespecified domain-specific penalty matrices; λ_k and α_k are layer-specific smoothing parameters for $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$, respectively ($k = 1, \cdots, r$). The quadratic penalties in the form $\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v}$ introduce regularization to the vector $\boldsymbol{v}$. Depending on the selection of the penalty matrix $\boldsymbol{\Omega}$, the penalty may be as simple as a sum of squared second differences $\boldsymbol{v}^T\boldsymbol{\Omega}\boldsymbol{v} = \sum(2v_j - v_{j-1} - v_{j+1})^2$, which encourages smoothness in $\boldsymbol{v}$. Throughout the article, we use the same setting of the penalty matrix as in Huang et al. (2008), which closely connects to smoothing spline (Green and Silverman, 1993). The smoothing parameters λ_k and α_k are different for different components and different layers, allowing for different degrees of smoothness. In particular, for two-way functional data, both λ_k and α_k are positive; for one-way functional data, we set $\alpha_k = 0$ to avoid regularization on the columns of $\boldsymbol{U}$.

The proposed model and penalized likelihood approach subsume many existing methods as special cases. On the one hand, if observed data follow a Gaussian distribution, (3) connects to the FSVD methods in Huang et al. (2008) (for one-way functional data) and Huang et al. (2009) (for two-way functional data). On the other hand, without the roughness penalty, (3) reduces to the exponential family principal component analysis (EPCA) method studied in Collins et al. (2001).

3. Algorithm

We devise a block-wise coordinate descent algorithm to optimize the penalized likelihood function in (3). Each step can be converted to a generalized linear model (GLM) with a regularization term, which can be efficiently fitted by an iteratively reweighted least square (IRLS) algorithm. Throughout the section, we assume the rank of the underlying natural parameter matrix is known, and the smoothing parameters are given. The estimation of the tuning parameters is discussed in the next section.

Notice that the roughness penalty (4) is separable with respect to different columns of $\boldsymbol{U}$ and $\boldsymbol{V}$. We develop an alternating algorithm to update one vector at a time while keeping everything else fixed. The algorithm alternates between the estimation of $\boldsymbol{v}_k$ and $\boldsymbol{u}_k$, and cycles through different layers until convergence. More specifically, let $\boldsymbol{U}_{-k}$ and $\boldsymbol{V}_{-k}$ denote the submatrices of $\boldsymbol{U}$ and $\boldsymbol{V}$ without the kth columns ($k = 1, \cdots, r$). With fixed $\boldsymbol{U}$ and $\boldsymbol{V}_{-k}$, if we ignore the regularization, we have the following GLM:

$$g[\mathbb{E}\{\mathrm{vec}(\boldsymbol{X})\}] = \mathrm{vec}(\boldsymbol{U}_{-\mathrm{k}}\boldsymbol{V}_{-\mathrm{k}}^{\mathrm{T}}) + (\boldsymbol{I}_{\mathrm{p}} \otimes \boldsymbol{u}_{\mathrm{k}})\boldsymbol{v}_{\mathrm{k}},$$

where $\mathrm{vec}(\boldsymbol{X})$ stacks the columns of $\boldsymbol{X}$, $\boldsymbol{I}_p$ is a $p \times p$ identity matrix (we shall drop the subscript when it does not cause any confusion), and $\otimes$ represents the Kronecker product. Namely, $\mathrm{vec}(\boldsymbol{X})$ is an $np \times 1$ response vector, $\mathrm{vec}(\boldsymbol{U}_{-\mathrm{k}}\boldsymbol{V}_{-\mathrm{k}}^{\mathrm{T}})$ is an offset term, $\boldsymbol{I} \otimes \boldsymbol{u}_k$ is an $np \times p$ design matrix, and $\boldsymbol{v}_k$ is a $p \times 1$ coefficient vector to be estimated. Therefore, the estimation of $\boldsymbol{v}_k$ can be formulated as the following optimization problem:

$$\min_{\boldsymbol{v}_k} -\mathrm{vec}(\boldsymbol{X})^{\mathrm{T}}\boldsymbol{\eta}_{\mathrm{k}} + 1^{\mathrm{T}}\mathrm{b}(\boldsymbol{\eta}_{\mathrm{k}}) + \frac{\lambda_{\mathrm{k}}}{2}\boldsymbol{v}_{\mathrm{k}}^{\mathrm{T}}\boldsymbol{\Omega}_{\boldsymbol{v}}\boldsymbol{v}_{\mathrm{k}}, \tag{5}$$

where $\boldsymbol{\eta}_k = \mathrm{vec}\left(\boldsymbol{U}_{-\mathrm{k}}\boldsymbol{V}_{-\mathrm{k}}^{\mathrm{T}}\right) + (\boldsymbol{I}_{\mathrm{p}} \otimes \boldsymbol{u}_{\mathrm{k}})\,\boldsymbol{v}_{\mathrm{k}}$ is the linear predictor. With a slight abuse of notation, let $b(\boldsymbol{\eta}_k)$ denote

entry-wise mappings of $\boldsymbol{\eta}_k$. The objective function is the penalized log likelihood corresponding to the GLM.

In general, the optimization problem (5) does not have an explicit solution. form, We use an quadratic approximation approach and exploit a regularized IRLS algorithm to solve (5). More specifically, $\boldsymbol{v}_k$ can be estimated by iteratively solving

$$\boldsymbol{v}_k^{(t+1)} = \arg\min_{\boldsymbol{v}_k} \| \boldsymbol{W}_k^{(t)\,1/2} \boldsymbol{y}_k^{(t)} - \boldsymbol{W}_k^{(t)\,1/2} (\boldsymbol{I} \otimes \boldsymbol{u}_k) \boldsymbol{v}_k \|_{\mathbb{F}}^2 + \lambda_k \boldsymbol{v}_k^T \boldsymbol{\Omega}_{\boldsymbol{v}} \boldsymbol{v}_k, \quad (6)$$

where $\boldsymbol{y}_k^{(t)} = (\boldsymbol{I} \otimes \boldsymbol{u}_k)\boldsymbol{v}_k^{(t)} + \{\mathrm{vec}(\boldsymbol{X}) - \mathrm{b}'(\boldsymbol{\eta}_{\mathrm{k}}^{(\mathrm{t})})\} \cdot \mathrm{b}''(\boldsymbol{\eta}_{\mathrm{k}}^{(\mathrm{t})})^{-1}$ is a working response vector with '$\cdot$' being the entrywise product; $\boldsymbol{W}_k^{(t)} = diag\{b''(\boldsymbol{\eta}_k^{(t)})\}$ is a diagonal weight matrix with diagonal values in $b''(\boldsymbol{\eta}_k^{(t)})$; $\boldsymbol{\eta}_k^{(t)} = \mathrm{vec}\left(\boldsymbol{U}_{-\mathrm{k}} \boldsymbol{V}_{-\mathrm{k}}^{\mathrm{T}}\right) + (\boldsymbol{I} \otimes \boldsymbol{u}_{\mathrm{k}})\, \boldsymbol{v}_{\mathrm{k}}^{(\mathrm{t})}$ is the working linear predictor; and $\| \cdot \|_{\mathbb{F}}$ represents the Frobenius norm. The updating formula (6) has a closed-form solution.

Similarly, when $\boldsymbol{V}$ and $\boldsymbol{U}_{-k}$ are fixed, the estimation of $\boldsymbol{u}_k$ can also be formulated as a penalized GLM problem, and solved by iteratively updating the following formula:

$$\begin{aligned} \boldsymbol{u}_k^{(t+1)} = \arg\min_{\boldsymbol{u}_k} & \| \widetilde{\boldsymbol{W}}_k^{(t)\,1/2} \widetilde{\boldsymbol{y}}_k^{(t)} - \widetilde{\boldsymbol{W}}_k^{(t)\,1/2} \\ & \times (\boldsymbol{I} \otimes \boldsymbol{v}_k) \boldsymbol{u}_k \|_{\mathbb{F}}^2 + \alpha_k \boldsymbol{u}_k^T \boldsymbol{\Omega}_{\boldsymbol{u}} \boldsymbol{u}_k, \end{aligned} \quad (7)$$

where the working response vector is $\widetilde{\boldsymbol{y}}_k^{(t)} = (\boldsymbol{I} \otimes \boldsymbol{v}_k)\boldsymbol{u}_k^{(t)} + \{\mathrm{vec}(\boldsymbol{X}^{\mathrm{T}}) - \mathrm{b}'(\widetilde{\boldsymbol{\eta}}_{\mathrm{k}}^{(\mathrm{t})})\} \cdot \mathrm{b}''(\widetilde{\boldsymbol{\eta}}_{\mathrm{k}}^{(\mathrm{t})})^{-1}$; the weight matrix is $\widetilde{\boldsymbol{W}}_k^{(t)} = diag\{b''(\widetilde{\boldsymbol{\eta}}_k^{(t)})\}$; and the working linear predictor is $\widetilde{\boldsymbol{\eta}}_k^{(t)} = \mathrm{vec}\left(\boldsymbol{V}_{-\mathrm{k}} \boldsymbol{U}_{-\mathrm{k}}^{\mathrm{T}}\right) + (\boldsymbol{I} \otimes \boldsymbol{v}_{\mathrm{k}})\, \boldsymbol{u}_{\mathrm{k}}^{(\mathrm{t})}$. For one-way functional data, we can simply set $\alpha_k = 0$, and formula (7) degenerates to the standard IRLS updating formula.

After estimating all the columns of $\boldsymbol{V}$ and $\boldsymbol{U}$ in one iteration, we normalize the estimated $\boldsymbol{U}$ and $\boldsymbol{V}$ by applying the SVD to $\boldsymbol{U}\boldsymbol{V}^T$. Consequently, the respective parameters are uniquely defined up to a sign change (if the singular values are distinct). The model fitting algorithm is summarized in Algorithm 1. More details can be found in Section A of the Supplementary Material. We remark that with fixed tuning parameters, the algorithm always reduces the objective function value in (3) in each step, and hence is guaranteed to converge if the likelihood is bounded. However, the penalized likelihood function is biconvex with respect to $\boldsymbol{U}$ and $\boldsymbol{V}$, so the stationary point may not be the global optimal solution. We discuss about the choice of initial values in Section D of the Supplementary Material.

Algorithm 1 The Model Fitting Algorithm for EFPCA

```
Initialize U and V;
while Estimation has not reached convergence do
    for k = 1, ..., r do
        Estimate v_k by iteratively updating (6);
        Estimate u_k by iteratively updating (7);
    end for
    Set U = (u_1, ..., u_r) and V = (v_1, ..., v_r);
    Normalize U and V through SVD.
end while
```

4. Tuning Parameter Selection

Up to now, the rank of the natural parameter matrix $\boldsymbol{\Theta}$, and the smoothing parameters for the roughness penalties are assumed known. In practice, they usually need to be estimated from data. In total, there are $2r + 1$ tuning parameters for two-way functional data (or $r + 1$ tuning parameters for one-way functional data). In this section, we introduce a new cross validation (CV) method for estimating the latent rank r for non-Gaussian data, and discuss a data-driven approach for estimating the smoothing parameters.

4.1. *Rank Estimation*

There has been a rich body of literature on estimating the number of principal components for Gaussian data (see Bai and Ng, 2002; Kritchman and Nadler, 2008; Owen and Perry, 2009; Josse and Husson, 2012, for example). However, the extension to non-Gaussian data is lacking. We develop an N-fold CV approach to estimate the rank of $\boldsymbol{\Theta}$ for any exponential family distribution. The idea is to randomly divide the entries of a data matrix into N non-overlapping parts. Each time, we leave out one block, and estimate natural parameter matrices with different ranks using the remaining data. Subsequently, we compute the residuals of the left-out block, and summarize them as CV scores for different ranks. We repeat the procedure for N different blocks, and aggregate the N CV scores into a single score for each rank. The ranks that give smaller CV scores are preferred. Technical details of the proposed rank selection procedure can be found in Section C of the Supplementary Material.

4.2. *Smoothing Parameter Selection*

In the penalty function (4), we use different smoothing parameters for different vectors in different layers. This allows different components to have different degrees of smoothness. It has been shown in the Gaussian case that component-specific smoothing parameters lead to higher estimation accuracy and more interpretable results (Huang et al., 2008, 2009). However, it is computationally prohibitive to search over a $2r$-dimensional space and select the optimal parameters simultaneously. As a remedy, we devise a data-driven selection procedure and embed it into the model fitting algorithm.

More specifically, we update one smoothing parameter at a time. Since λ_k only comes into play in the estimation of $\boldsymbol{v}_k$, we select the best λ_k when we solve (6) for $\boldsymbol{v}_k$. Similarly, we select the best α_k when we solve (7) for and $\boldsymbol{u}_k$. The objective functions in (6) and (7) can be written as penalized least squares $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{v}\|_{\mathbb{F}}^2 + \lambda \boldsymbol{v}^T \boldsymbol{\Omega} \boldsymbol{v}$, whose leave-one-out cross validation (LOOCV) score has an explicit formula as

$$\mathrm{LOOCV}(\lambda) = \frac{\|\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}\|_{\mathbb{F}}^2}{n\{1 - tr(\boldsymbol{H})\}^2},$$

where $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{\Omega})^{-1}\boldsymbol{X}^T$ is a hat matrix, $tr(\cdot)$ is the trace of a square matrix, and n is the number of samples in $\boldsymbol{y}$. Thus, there is no need to refit the model to obtain CV scores. Consequently, the best smoothing parameter can be selected very efficiently using LOOCV. Moreover, we also consider a variant of the LOOCV procedure to achieve smoother estimation. More details can be found in Section B of the Supplementary Material.

With the above procedure, it is computationally efficient to update the optimal values for λ_k and α_k in (6) and (7) in each iteration. Although the optimal values may change over inner iterations (of updating $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$) and outer iterations (of updating $\boldsymbol{U}$ and $\boldsymbol{V}$), numerical studies suggest they usually remain constant over inner iterations and converge after a few outer iterations. In practice, one could also fix the smoothing parameters after a few outer iterations to guarantee and expedite the convergence of the overall fitting algorithm.

5. Simulation

In this section, we first present results of rank estimation in different settings. Then, we treat ranks as fixed and demonstrate the efficacy of the EFPCA method. In particular, we consider one-way and two-way non-Gaussian functional data. For one-way functional data, we compare the proposed method with the LGP method (Hall et al., 2008), the gfPCA method (Gertheiss et al., 2017), and the EPCA method (Collins et al., 2001). For two-way functional data where LGP and gfPCA are not applicable, we compare EFPCA with EPCA and an ad hoc approach where we first transform non-Gaussian data to be continuous, and then apply FSVD (Huang et al., 2009).

5.1. *Rank Estimation Results*

We consider four scenarios, where data matrices of size 100×50 are generated from Gaussian, Poisson, Bernoulli, and binomial (the number of trials is fixed to be 1000) distributions, respectively. In particular, in the first three scenarios, the underlying natural parameter matrices have rank 5, with singular values being $(45, 40, 35, 30, 25)$ for Gaussian, $(25, 25, 25, 20, 20)$ for Poisson, and $(120, 100, 90, 80, 60)$ for Bernoulli. We remark that the information contained in Bernoulli data is scarce, so we intentionally increase the signal level to make it detectable. In the last scenario, we consider a High-Rank Binomial setting where only the first few ranks are dominant, to mimic real world situations. The true rank is 8 and the singular values are $(50, 45, 40, 10, 8, 6, 4, 2)$.

We conduct a 10-fold CV study for each simulation setting. The results are shown in Figure 1. In the first three settings, the average CV score across different folds reaches minimum at the correct rank ($r = 5$). In the High-Rank Binomial setting, the CV score keeps decreasing but has a clear "elbow" point at $r = 3$, indicating the first three ranks are dominant. In general, the proposed method can effectively estimate the rank of the natural parameter matrix underlying a generalized data matrix. Further investigation on how the number of folds affects the rank selection results is in Section C of the Supplementary Material.

5.2. *Simulation Settings*

We simulate data from Model (2), with $n = 100$, $p = 50$, and $r = 2$. For one-way functional data, Model (2) is equivalent to a discretized LGP model on a dense grid. The left singular vectors are filled with standard Gaussian random numbers; the right singular vectors are obtained by discretizing a quadratic function and a sine function on the same regularly spaced grid. For two-way functional data, we substitute left singular vectors with mutually orthogonal sinusoids evaluated on a regular and dense grid. All singular vectors are standardized (each left singular vector is also centered).

For one-way functional data, we consider two settings: Bernoulli and Poisson. The non-zero singular values are set to be $(40, 30)$ for each setting. For two-way functional data, we particularly focus on the binomial distribution under two settings: one has an underlying rank 2 with singular values $(40, 30)$, and the other has an underlying rank 12 where the first two singular values are $(40, 30)$ and the remaining singular values are all 1. It simulates an approximately low-rank scenario. The number of total trials for each entry is contained in a matched matrix $\boldsymbol{M}$. We particularly divide $\boldsymbol{M}$ into 4 equal-sized blocks as $\boldsymbol{M} = \begin{pmatrix} \boldsymbol{M}_1 & \boldsymbol{M}_2 \\ \boldsymbol{M}_3 & \boldsymbol{M}_4 \end{pmatrix}$, where $\boldsymbol{M}_1$ and $\boldsymbol{M}_4$ are filled with random integers from 10 to 20, and $\boldsymbol{M}_2$ and $\boldsymbol{M}_3$ are filled with random integers from 2 to 4. Subsequently, the data matrix $\boldsymbol{X}$ is generated from binomial distributions with the preset success rates and numbers of trials. In all settings, we always set the rank to be $r = 2$ when estimating model parameters.

We evaluate the performance of different methods using the following criteria:

$$\mathrm{MSE}_{\boldsymbol{V}} = \|\boldsymbol{V} - \widehat{\boldsymbol{V}}\|_{\mathbb{F}}, \quad \mathrm{Angle}_{\boldsymbol{V}} = \frac{180}{\pi}\arccos\{\sigma_0(\widehat{\boldsymbol{V}}^T\boldsymbol{V})\},$$
$$\mathrm{MSE}_{\boldsymbol{\Theta}} = \|\boldsymbol{\Theta} - \widehat{\boldsymbol{\Theta}}\|_{\mathbb{F}}, \tag{8}$$

where $\sigma_0(\boldsymbol{A})$ denotes the smallest singular value of $\boldsymbol{A}$. The second criterion measures the maximal principal angle between the subspaces spanned by the columns in $\boldsymbol{V}$ and the columns in $\widehat{\boldsymbol{V}}$. In two-way settings, we also evaluate the mean squared error (MSE) and maximal principal angle for left singular vectors (denoted by $\mathrm{MSE}_{\boldsymbol{U}}$ and $\mathrm{Angle}_{\boldsymbol{U}}$). In addition, we also compare the computing times of different methods.

5.3. *Simulation Results*

For each simulation setting, we conduct 100 simulation runs and summarize the results. Summary statistics of different criteria for the one-way simulation settings are shown in Table 2. In the Bernoulli setting, the proposed EFPCA method performs similarly to the LGP and gfPCA methods in terms of the estimation accuracy. All three methods significantly outperform the EPCA approach, which does not account for the underlying smooth structure in data. However, from the computational perspective, the proposed method is much more appealing than LGP and gfPCA. Our method can be fitted within a couple of seconds on a standard desktop computer (Intel i5 CPU (3.3GHz) with 8Gb RAM). The computational advantage of EFPCA persists in higher dimensions as well. In the Poisson setting where gfPCA is not applicable, the proposed method significantly outperforms the EPCA and LGP methods in all criteria (surprisingly, LGP is even worse than the non-regularized EPCA method). Again, the EFPCA method is computationally efficient, and more than 200 times faster than the LGP method.

The results of the two-way binomial setting are summarized in Table 3. In particular, the ad hoc procedure is as follows: first divide each value of $\boldsymbol{X}$ by the corresponding value in $\boldsymbol{M}$

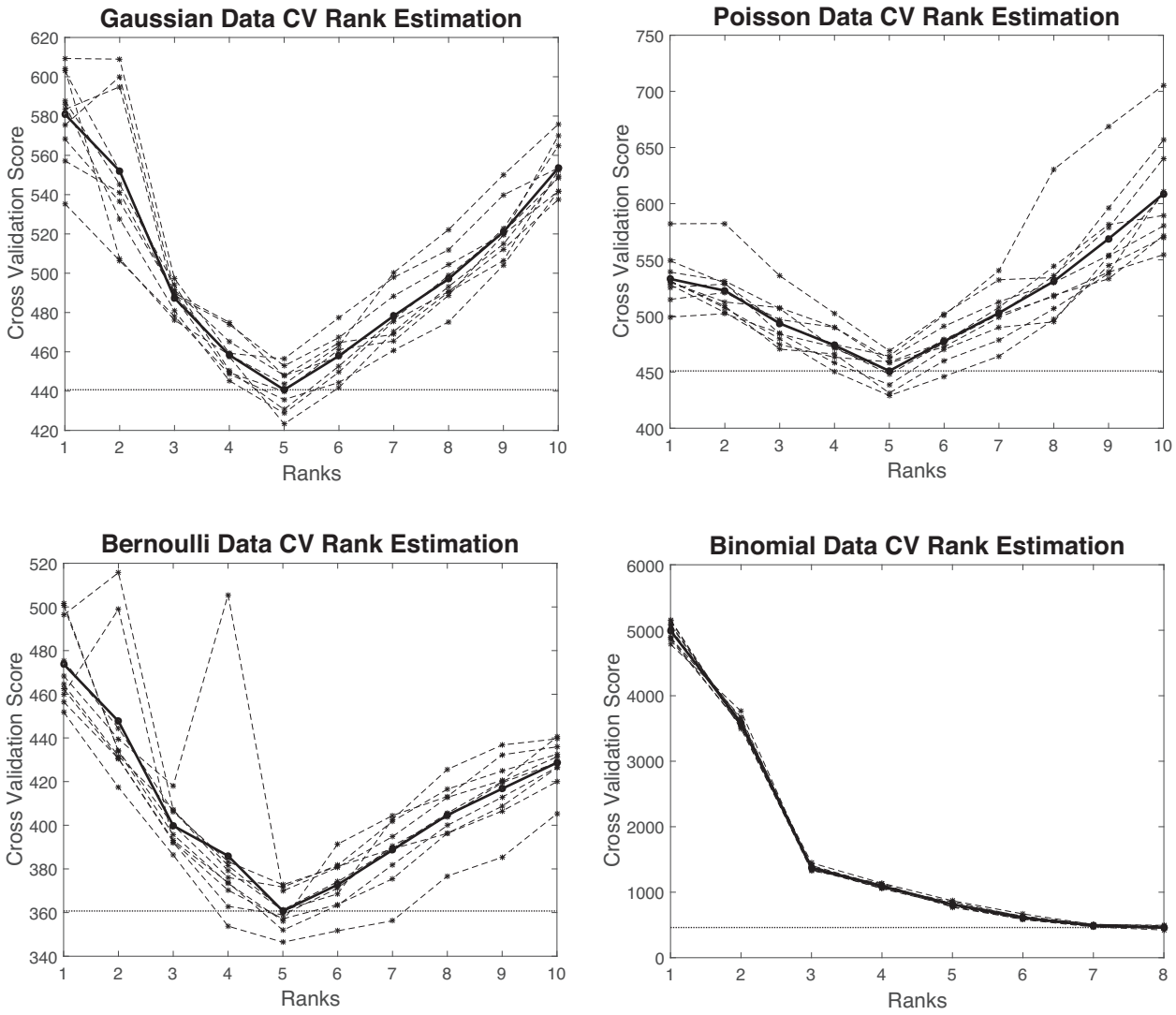

Figure 1. Simulation Study of Cross-Validation Rank Estimation: Each panel corresponds to the 10-fold cross validation result of a simulation setting (upper left: Gaussian; upper right: Poisson; bottom left: Bernoulli; bottom right: High-Rank Binomial). In each plot, the dashed lines with stars correspond to CV scores in different folds; the solid line with circles corresponds to average CV scores. The minimal average CV score is marked by a horizontal dotted line.

Table 2

One-way functional data simulation results for the Bernoulli setting and the Poisson setting. Median and median absolute deviation (in parenthesis) for the values of different criteria are calculated from 100 simulation runs. The best results are in boldface.

		EPCA	LGP	gfPCA	EFPCA (proposed)
Bernoulli setting	$\text{MSE}_{\boldsymbol{V}}$	0.83 (0.06)	**0.31** (0.06)	0.43 (0.09)	0.33 (0.09)
	$\text{Angle}_{\boldsymbol{V}}$	36.65 (3.08)	12.25 (2.27)	16.03 (2.57)	**12.15** (3.48)
	$\text{MSE}_{\boldsymbol{\Theta}}$	50.54 (2.21)	33.82 (0.87)	**29.13** (0.95)	34.30 (2.15)
	Time (sec)	**1.04** (0.19)	391.03 (9.25)	11.72 (0.39)	3.81 (0.59)
Poisson setting	$\text{MSE}_{\boldsymbol{V}}$	0.27 (0.01)	0.63 (0.05)	N/A	**0.13** (0.02)
	$\text{Angle}_{\boldsymbol{V}}$	11.74 (0.79)	13.15 (1.27)	N/A	**4.76** (0.82)
	$\text{MSE}_{\boldsymbol{\Theta}}$	16.69 (0.53)	31.73 (0.61)	N/A	**14.23** (0.59)
	Time (sec)	**1.66** (0.17)	316.46 (2.37)	N/A	2.78 (0.23)

Table 3

Two-way functional data simulation results for the binomial settings with low rank ($r = 2$) and approximately low rank ($r = 12$). Median and median absolute deviation (in parenthesis) for the values of different criteria are calculated from 100 simulation runs. The best results are in boldface.

		EPCA	Ad Hoc	EFPCA (proposed)
Binomial setting (low rank)	$\text{MSE}_{\boldsymbol{V}}$	0.23 (0.02)	0.15 (0.04)	**0.10** (0.02)
	$\text{Angle}_{\boldsymbol{V}}$	10.13 (0.71)	5.72 (1.18)	**3.61** (0.75)
	$\text{MSE}_{\boldsymbol{U}}$	0.32 (0.01)	0.28 (0.03)	**0.12** (0.02)
	$\text{Angle}_{\boldsymbol{U}}$	14.25 (0.68)	13.44 (1.57)	**5.06** (0.57)
	$\text{MSE}_{\boldsymbol{\Theta}}$	13.56 (0.45)	27.46 (1.48)	**4.41** (0.50)
	Time (sec)	1.03 (0.02)	**0.96** (0.12)	2.79 (0.17)
Binomial setting (approx. low rank)	$\text{MSE}_{\boldsymbol{V}}$	0.23 (0.02)	0.16 (0.03)	**0.10** (0.02)
	$\text{Angle}_{\boldsymbol{V}}$	9.83 (0.73)	6.19 (1.38)	**3.77** (0.67)
	$\text{MSE}_{\boldsymbol{U}}$	0.32 (0.01)	0.28 (0.03)	**0.11** (0.02)
	$\text{Angle}_{\boldsymbol{U}}$	14.17 (0.62)	13.19 (1.88)	**5.01** (0.53)
	$\text{MSE}_{\boldsymbol{\Theta}}$	13.85 (0.41)	27.71 (1.60)	**5.44** (0.32)
	Time (sec)	1.03 (0.01)	**0.96** (0.11)	3.67 (0.14)

to get an estimate of the success rate; then transform the success rate with the logit function; finally apply FSVD to the transformed success rate matrix. To avoid singularity, we threshold the estimated rates by 0.01 and 0.99 before transformation. In both settings, the EFPCA method is uniformly the best in terms of parameter estimation accuracy. Notice that the ad hoc procedure outperforms the EPCA method in the singular vector estimation, but has worse results in the natural parameter matrix estimation (i.e., higher $\text{MSE}_{\boldsymbol{\Theta}}$). This is because the ad hoc procedure cannot take into account the distributional information, and its singular value estimation is greatly affected by the preset thresholds of the entrywise estimated rates. Overall, the EFPCA method achieves the best estimation accuracy at a low computational cost.

6. UK Mortality Rate Study

In this section, we apply the EFPCA method to the UK mortality data. We focus on the male population in the main manuscript, and study the female population in Section E of the Supplementary Material. The observed data are the pair of population size (i.e., $\boldsymbol{M}$) and death count (i.e., $\boldsymbol{X}$) for each age group (from 0 to 95, $p = 96$) in each year (from 1922 to 2013, $n = 92$). The visualization of the raw data can also be found in Section E of the Supplementary Material. Given the population size, the number of deaths follows a binomial distribution. We use the proposed method to investigate mortality rate patterns across age groups and calendar years.

6.1. *Model Fitting*

We first estimate the rank of the underlying mortality rate structure. We use an eightfold CV approach as described in Section 4.1. The CV score plot is shown in Section E of the Supplementary Material. The CV score keeps declining as a function of the rank, with a clear "elbow" point at $r = 3$. This implies the true rank may be large, but the first three ranks capture most of the variability. The large rank may be due to the cohort effect, where diagonal entries of the observed data matrix are from the same birth cohort and thus interrelated. Regardless, we set the rank to be 3.

We fit a rank-3 EFPCA model to the binomial data ($\boldsymbol{X}$, $\boldsymbol{M}$) with two-way smoothness. The algorithm converges very quickly after 10 iterations and takes less than 30 s in total. The three singular values of the underlying natural parameter matrix are 492.57, 16.07, and 10.74. The estimated loadings in $\boldsymbol{U}$ and $\boldsymbol{V}$, as well as pointwise 95% confidence intervals, are shown in Figure 2. The confidence intervals are obtained from a parametric bootstrap approach. More specifically, after fitting the model with the original data, we use the fitted model as a generative model to simulate binomial random numbers (with the fixed $\boldsymbol{M}$), and re-estimate the model parameters. The procedure is repeated 100 times, and we obtain 100 sets of parameter estimates. Subsequently, we calculate the mean and standard deviation of the estimates for each loading entry, and obtain a 95% pointwise confidence interval based on the Gaussian distribution. We remark that the confidence intervals in Figure 2 are too narrow to be visible, mainly due to the excessive numbers of trials in $\boldsymbol{M}$.

As a comparison, we follow the procedure in Huang et al. (2009), and directly apply FSVD (with the rank set to be 2) to the transformed data (i.e., ratios of death counts $\boldsymbol{X}$ and population sizes $\boldsymbol{M}$). The estimated loadings are shown in Figure 3. Note that the FSVD loadings and the EFPCA loadings are not on the same scale.

6.2. *Results*

The first age-specific loadings (i.e., $\boldsymbol{v}_1$) in both methods recover the well known mortality rate patterns: the mortality rate is relatively high for newborns; but it quickly drops to a lower level; after adolescence the mortality rate increases with age. Focusing on the natural parameters, the EFPCA method reveals more detailed patterns: the mortality rate reaches a minimum around 12; it increases to a local peak around

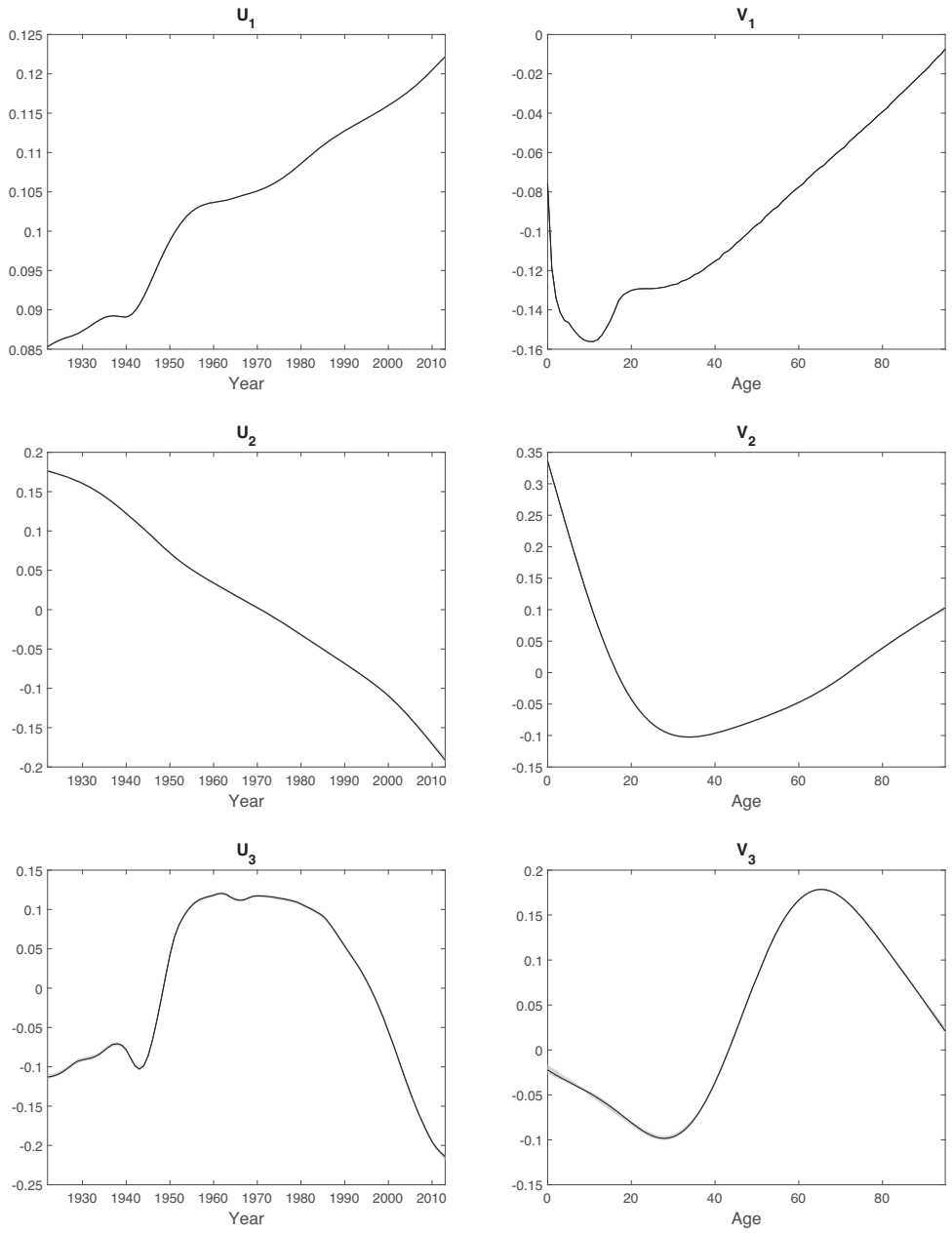

Figure 2. UK Male Mortality Study: EFPCA loadings (solid lines) and pointwise 95% confidence intervals (gray bands, very narrow) from a rank-3 model.

20 possibly due to various activity accidents in adolescence; after 30, the log odds of mortality increases almost linearly with age. The first time-varying loadings (i.e., $\boldsymbol{u}_1$) reflect the dominant time effect. Overall, the mortality rate declines gradually over decades (in EFPCA, because $\boldsymbol{v}_1$ is negative, larger values in $\boldsymbol{u}_1$ mean lower mortality rates). This probably attributes to the fast advancement of medical conditions and technologies. Furthermore, in Figure 2 we observe an obvious dip around year 1940, which coincides with the World War II. Since UK was actively involved in the war, the male mortality rates naturally went up beyond normal. This feature is missed by FSVD.

The subsequent loadings capture more subtle structures. The second pair of FSVD loadings mainly capture the variation for the very young and very old population over time. However, this is probably due to the much smaller population size of those age groups (which leads to larger variation) rather than richer mortality patterns. The FSVD does not account for the population size of different age groups, and thus cannot distinguish the two sources of variation. As a result, it may overlook important mortality patterns for the larger population.

In contrast, the subsequent EFPCA loadings are more interpretable and trustworthy. The second pair of loadings

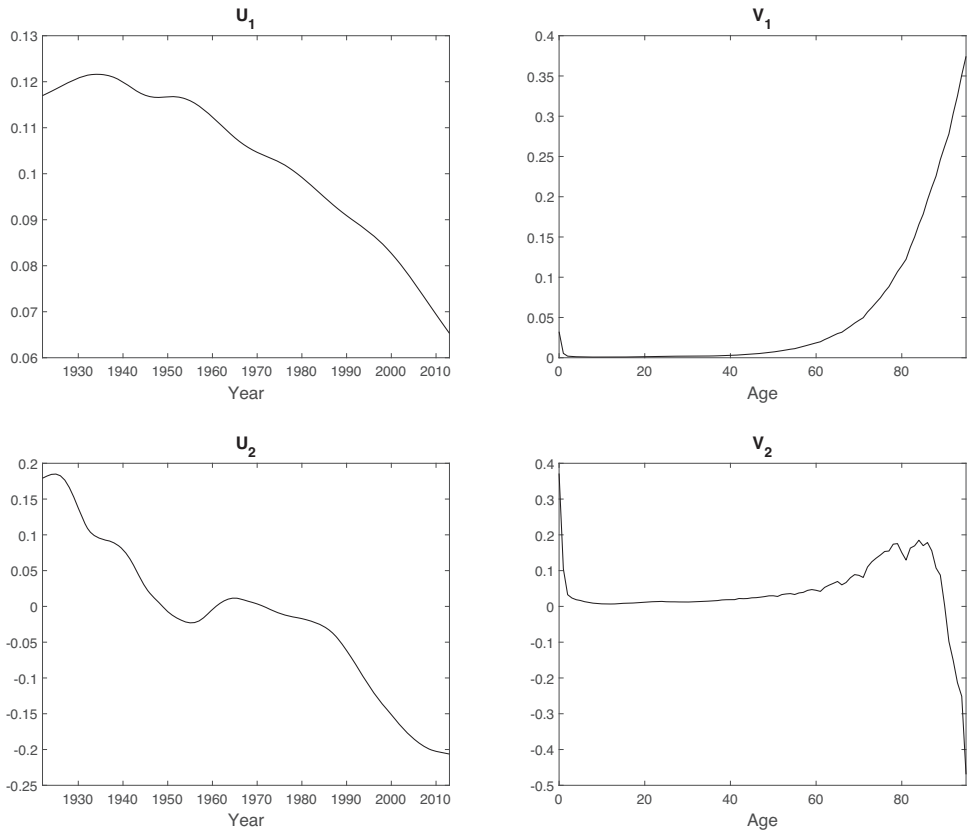

Figure 3. UK Male Mortality Study: FSVD loadings with rank 2 (directly applied to the transformed data).

identifies that the mortality rate for young people (younger than 15 or so) and old people (older than 60 or so) decreases over years, while the mortality rate for middle aged men has the opposite time-varying trend. A plausible explanation is that for very young and old people, the leading cause of death is the health related issues. Their mortality rate largely depends on the level of medical care, which gets steadily improved over time. However, for the middle aged working class, the leading causes of death include unintentional injury, homicide and suicide, and various diseases. With the fast development of the industry and deterioration of the environment, these factors become more and more problematic in the modern society. Consequently, after taking out the dominant pattern, the secondary mortality rate pattern shows clear contrast between the two groups. The third time-varying loading $\boldsymbol{u}_3$ has a clear dip between 1939 and 1945, corresponding to the World War II period. The third age-specific loading $\boldsymbol{v}_3$ indicates the mainly affected population are the young people between 20 and 40. This pair of loadings clearly captures the adverse effect of the war on male mortality.

All in all, the proposed method not only identifies the well known trend of mortality rate across age groups and years, but also reveal interesting patterns related to social environment changes and major historical events. It outperforms the FSVD in pattern recognition and interpretation. It has potential application in mortality forecasting and various social studies.

7. Discussion

In this article, we develop an EFPCA method for modeling generalized functional data with one/two-way smoothness. We devise a computationally efficient algorithm to fit the model. We also develop a new CV approach for estimating the rank of the underlying natural parameter matrix of non-Gaussian data. It suits well for the proposed method, and may have more general applications. Both simulation and real data studies demonstrate the efficacy and efficiency of the proposed method.

There are a few directions for future research. One direction is to extend the current framework to sparse generalized longitudinal observations. The proposed method is most suitable for dense data measured on a regular grid. For sparse and irregular measurements, one may use a common grid where each observation has multiple missing values, and exploit an expectation-maximization algorithm to deal with missing data. Another topic of interest is to explore the theoretical properties of the proposed framework. Since we model generalized functional data from a matrix factorization perspective, the standard large-sample asymptotics does not easily apply. More theoretical insights call for further investigation.

8. Supplementary Materials

Web Appendices and Figures referenced in Sections 3, 4, 5, 6 are available with this article at the *Biometrics* website on Wiley Online Library. The mortality data are

publicly available at http://www.mortality.org. The Matlab code for implementing the proposed method is available at https://github.com/reagan0323/EFPCA.

Acknowledgements

JH was partially supported by the National Science Foundation grant DMS-1208952. HS was partially supported by the National Science Foundation grants DMS-1106912 and DMS-1407655, the Ministry of Science and Technology Major Project of China 2017YFC1310900 and 2017YFC1310903, University of Hong Kong Stanley Ho Alumni Challenge Fund, HKU University Research Committee Seed Funding Award 104004215, and the Xerox UAC Foundation.

References

Bai, J. and Ng, S. (2002). Determining the number of factors in approximate factor models. *Econometrica* **70**, 191–221.

Collins, M., Dasgupta, S., and Schapire, R. E. (2001). A generalization of principal components analysis to the exponential family. *In Advances in Neural Information Processing Systems*, 617–624.

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

Gertheiss, J., Goldsmith, J., and Staicu, A.-M. (2017). A note on modeling sparse exponential-family functional response curves. *Computational Statistics & Data Analysis* **105**, 46–52.

Goldsmith, J., Zipunnikov, V., and Schrack, J. (2015). Generalized multilevel function-on-scalar regression and principal component analysis. *Biometrics* **71**, 344–353.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. Boca Raton: CRC Press.

Hall, P., Müller, H.-G., and Yao, F. (2008). Modelling sparse generalized longitudinal observations with latent gaussian processes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **70**, 703–723.

Huang, J. Z., Shen, H., and Buja, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electronic Journal of Statistics* **2**, 678–695.

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.

Josse, J. and Husson, F. (2012). Selecting the number of components in principal component analysis using cross-validation approximations. *Computational Statistics & Data Analysis* **56**, 1869–1879.

Kritchman, S. and Nadler, B. (2008). Determining the number of components in a factor model from limited noisy data. *Chemometrics and Intelligent Laboratory Systems* **94**, 19–32.

Owen, A. B. and Perry, P. O. (2009). Bi-cross-validation of the svd and the nonnegative matrix factorization. *The Annals of Applied Statistics* **3**, 564–594.

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

Serban, N., Staicu, A.-M., and Carroll, R. J. (2013). Multilevel cross-dependent binary longitudinal data. *Biometrics* **69**, 903–913.

Zhang, L., Shen, H., and Huang, J. Z. (2013). Robust regularized singular value decomposition with application to mortality data. *The Annals of Applied Statistics* **7**, 1540–1561.

Received February 2017. *Revised February* 2018.
Accepted March 2018.