The Annals of Applied Statistics
2013, Vol. 7, No. 3, 1540–1561
DOI: 10.1214/13-AOAS649

ROBUST REGULARIZED SINGULAR VALUE DECOMPOSITION WITH APPLICATION TO MORTALITY DATA

BY LINGSONG ZHANG, HAIPENG SHEN[1] AND JIANHUA Z. HUANG[2]

Purdue University, University of North Carolina and Texas A&M University

We develop a robust regularized singular value decomposition (RobRSVD) method for analyzing two-way functional data. The research is motivated by the application of modeling human mortality as a smooth two-way function of age group and year. The RobRSVD is formulated as a penalized loss minimization problem where a robust loss function is used to measure the reconstruction error of a low-rank matrix approximation of the data, and an appropriately defined two-way roughness penalty function is used to ensure smoothness along each of the two functional domains. By viewing the minimization problem as two conditional regularized robust regressions, we develop a fast iterative reweighted least squares algorithm to implement the method. Our implementation naturally incorporates missing values. Furthermore, our formulation allows rigorous derivation of leave-one-row/column-out cross-validation and generalized cross-validation criteria, which enable computationally efficient data-driven penalty parameter selection. The advantages of the new robust method over nonrobust ones are shown via extensive simulation studies and the mortality rate application.

1. Introduction. This paper develops a *robust* regularized singular value decomposition (SVD) method for *two-way functional* data. One-way functional data analysis (FDA) focuses on a population of curves or functions and has gained much attention in the last decade or so, as well documented in Ramsay and Silverman (2002, 2005) and Ferraty and Vieu (2006). Different from one-way functional data, two-way functional data are functions in two ways: both index domains I and J of the data matrix $\mathbf{X} = (x_{i,j})_{i\in I, j\in J}$ are structured with notions of smoothness, that is, both rows and columns of the data matrix can be viewed as discretizations of some underlying smooth functions [Huang, Shen and Buja (2009)]. For example, in our motivating Spanish mortality application (Section 4), the data matrix records mortality rates for different age groups between ages 0 and 110 (columns) in Spain from year 1908 to 2007 (rows). It is reasonable to consider the mortality rate as a

Received October 2012; revised February 2013.

[1]Supported in part by NIH/NIDA (1 RC1 DA029425-01) and NSF (CMMI-0800575, DMS-11-06912).

[2]Supported in part by NCI (CA57030), NSF (DMS-09-07170, DMS-10-07618, DMS-12-08952) and Award No. KUS-C1-016-04, made by King Abdullah University of Science and Technology (KAUST).

Key words and phrases. Cross-validation, functional data analysis, GCV, principal component analysis, robustness, smoothing spline.

smooth function of both age and time period. Similar two-way functional structure also exists in many other applications. For example, the network traffic pattern in Zhang et al. (2007) is a smooth function of time-of-the-day and calendar date; the call center customer patience in Huang, Shen and Buja (2009) is a smooth function of customer waiting time and time-of-the-day; and the magnetoencephalography signal in Tian and Li (2011) is a smooth function of signal recording time and brain spatial location.

Recently, Huang, Shen and Buja (2009) proposed a regularized singular value decomposition (RSVD) method for dimension reduction and feature extraction of two-way functional data. It is based on minimization of a regularized sum of squared reconstruction errors of a low-rank matrix approximation. Since the squared-error loss function is used to measure the size of reconstruction errors, the results of applying the RSVD are sensitive to outliers. Outliers in two-way functional data can appear in various forms, such as outlying cells, columns, rows or blocks (Section 3). For example, the Spanish mortality data contain two outlying time periods and, as we will demonstrate in Section 4, they significantly affect the estimation of the underlying smooth mortality trend across year when applying the RSVD. One major contribution of the current paper is to develop a robust regularized SVD method that can mitigate outlying effects in two-way functional data analysis, which, to the best of our knowledge, is the first of its kind.

To give some background on our proposed method for two-way functional data, we first review several relevant robust functional principal component analysis (PCA) methods that have been developed for analyzing one-way functional data. Locantore et al. (1999) proposed a robust PCA approach, which projects the data onto a sphere or an ellipse around a robust estimate of the center of the data, and then performs the usual PCA on the covariance matrix of the projected data. Gervini (2008) extended the approach of Locantore et al. (1999) to functional data, introduced the concepts of functional median and functional spherical principal components (PC), and established the corresponding robustness properties of the approach. Hyndman and Shahid Ullah (2007) and Hyndman and Shang (2009) used a projection pursuit (PP) approach for robust functional PCA; Bali et al. (2011) recently studied the asymptotic robustness properties of this PP approach in terms of influence function and breakdown point. On the other hand, Bai et al. (2008) proposed a supervised SVD technique, which can be combined with independent component analysis to improve the robustness of analyzing functional MRI brain images. Gervini (2009) considered irregularly and sparsely sampled functional data, used basis expansions to model the functional trajectories, and modeled the functional PC scores and the reconstruction errors using heavy-tailed distributions such as t or Cauchy to achieve robustness. All this work has focused on one-way functional data.

We now introduce some notation to facilitate the discussion of our proposed robust regularized SVD method for two-way functional data. Sometimes it is reasonable to use the term—functional SVD—instead of regularized SVD to emphasize the focus on functional data. We view the element x_{ij} of the $m \times n$ data

matrix $\mathbf{X}$ as evaluation of an underlying *smooth* function $X(\cdot,\cdot)$ on a rectangular grid of sampling points (y_i, z_j), where y_i $(i = 1,\ldots,m)$ are from a domain $\mathcal{Y}$ and z_j $(j = 1,\ldots,n)$ are from a domain $\mathcal{Z}$. According to Huang, Shen and Buja (2009), the RSVD for two-way functional data can be considered as fitting the following smooth rank-r approximation model for the two-way functional data:

$$X(y,z) = U_1(y)V_1(z) + U_2(y)V_2(z) + \cdots + U_r(y)V_r(z) + \varepsilon(y,z), \tag{1}$$

where $U_k(y)$ and $V_k(z)$ are smooth functions on their respective domains, and $\varepsilon(y,z)$ is a mean zero random noise. Model (1) can be thought of as a truncated version of the singular value decomposition of bivariate functions [Gervini (2010)], and the orthonormal constraints $\int V_k(z)V_l(z)\,dz = \delta_{kl}$, where δ is the Kronecker delta, are usually imposed for identifiability. The low-rank approximation formulation indicates that the proposed SVD method is useful for dimensionality reduction and feature selection. The smoothness requirement on $U_k(y)$ and $V_k(z)$ takes into account the underlying continuity of the functional data. It is important to note that the SVD formulation offers a symmetric treatment of the two domains. The existing robust functional PCA methods cannot be directly extended to two-way functional data, because PCA treats the rows and the columns asymmetrically. We are therefore led to the SVD which offers symmetric treatment.

To give a simple description of our approach, we focus on extracting the first pair of components in (1), $U_1(y)$ and $V_1(z)$, whose discretized realizations are, respectively, denoted as $\mathbf{u}_1 \equiv (U_1(y_1),\ldots,U_1(y_m))^T$ and $\mathbf{v}_1 \equiv (V_1(z_1),\ldots,V_1(z_n))^T$. Subsequent pairs are extracted sequentially after removing the effects of the preceding pairs. This sequential approach allows the different pairs of components to have differing smoothness. The extracted components should possess two desirable features—smoothness and robustness against outliers. We propose to solve the following problem:

$$(\mathbf{u}_1, \mathbf{v}_1) \equiv \underset{\mathbf{u},\mathbf{v}}{\operatorname{argmin}}\big\{\rho\big(\mathbf{X} - \mathbf{u}\mathbf{v}^T\big) + \mathcal{P}_\lambda(\mathbf{u},\mathbf{v})\big\}, \tag{2}$$

where $\mathbf{u}$ and $\mathbf{v}$ are m-dimensional and n-dimensional vectors, respectively, $\rho(\cdot)$ is a robust loss function, $\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u},\mathbf{v})$ is a two-way roughness penalty to ensure smoothness for the $\mathbf{u}$ and $\mathbf{v}$, and $\boldsymbol{\lambda}$ is a vector of penalty parameters.

This formulation is very general, allowing the flexibility in the choice of the loss function and the penalty function. Although various robust loss functions in the robust statistics literature [Huber and Ronchetti (2009)] can be used in our framework, we focus on a typical Huber's function for its easy implementation and fast computation. If the nonrobust squared-error loss is used, then the penalized criterion function in (2) reduces to the minimizing criterion for the RSVD of Huang, Shen and Buja (2009). By using a robust loss function, our framework essentially robustifies the RSVD method and, therefore, we refer to our approach as robust regularized SVD, or *RobRSVD* for short. On the other hand, without the penalty term, the criterion in (2) offers another way for robust SVD [Ammann

(1993), Liu et al. (2003)]; hence, RobRSVD can also be interpreted as smoothing of a robust SVD.

In this paper, we adopt the two-way roughness penalty function introduced in Huang, Shen and Buja (2009), which has several desirable properties for two-way regularization. Other choices of penalty functions are possible such as the ones that shrink the functional components to certain subspaces, for example, spaces of periodic functions. Our framework also offers one-way robust functional data analysis as a special case if one only imposes roughness penalty on one of the functional domains such as the one that corresponds to the row or the column of the data matrix. One important feature of our method is that it works directly with the raw observed data; there is no need to pre-smooth the raw data, nor to obtain a robust estimate of the high-dimensional covariance matrix, which can be computationally challenging for one-way functional data and even more technically difficult for two-way functional data.

We develop an efficient iterative reweighted least squares (IRLS) algorithm to solve the minimization problem (2). Our algorithm iteratively updates $\mathbf{u}$ and $\mathbf{v}$ conditioning on the other, where each updating step can be viewed as a (regularized) robust regression. This view of (2) as conditional robust regressions suggests that many robust regression procedures can be used, such as the M-estimator [Huber and Ronchetti (2009)], the L_1 estimator [Croux et al. (2003)], the least median of squares (LMS) and the least trimmed squares (LTS) estimators [Rousseeuw (1984)], and the IRLS estimator [Heiberger and Becker (1992)]. We choose the IRLS estimator in this paper for the following two reasons. First, it enables us to interpret the conditional regularized robust regressions as regularized weighted least squares. Based on this interpretation, we can rigorously derive explicit shortcut formula for leave-one-row/column-out cross-validation and related generalized cross-validation (GCV) scores; hence, data-driven selection of the penalty parameters can be carried out very efficiently. Note that the selection of the penalty parameters for the row and column is naturally decoupled due to the conditional regression perspective. Second, the IRLS estimator is used due to its fast computation and comparable performance when compared against several other robust regression procedures, as shown by Shen, Zhu and Lee (2007). The alternating estimation procedure also suggests a natural way to incorporate missing values.

The remainder of the paper is organized as follows. Section 2 describes technical details of the RobRSVD method, including formulation, the IRLS algorithm, penalty parameter selection, treatment of missing values and interpolation of results in function space. Results of simulation studies are presented in Section 3 to compare the performance of RobRSVD with standard SVD and the regularized SVD (RSVD) of Huang, Shen and Buja (2009). Section 4 analyzes the motivating Spanish mortality application and demonstrates the practical advantages of RobRSVD over the other two methods.

2. The methodology. We describe the RobRSVD method in this section. Section 2.1 gives its formulation, Section 2.2 derives the IRLS algorithm, and Sections 2.3–2.6 discuss several implementation details.

2.1. *Formulation*. It is well known that the SVD can be viewed as finding a sequence of rank-one matrix approximations of a data matrix [Gabriel and Zamir (1979)]. We adapt this idea to define the RobRSVD as a method for obtaining a sequence of *robust regularized* rank-one matrix approximations. Our discussion focuses on obtaining the first pair of components. Subsequent pairs of components can be obtained by applying the method sequentially on the residuals from lower-rank approximations.

The first pair of singular vectors of a data matrix $\mathbf{X} = (x_{ij})_{m\times n}$ can be obtained by solving a least squares problem as

$$(\widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \underset{(\mathbf{u},\mathbf{v})}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|_F^2,$$

where $\mathbf{u}$ and $\mathbf{v}$ are $m \times 1$ and $n \times 1$ vectors, respectively, and $\|\cdot\|_F$ is the Frobenius norm of a matrix. For two-way functional data, the RSVD of Huang, Shen and Buja (2009) defines the regularized singular vectors as

$$(\widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \underset{(\mathbf{u},\mathbf{v})}{\operatorname{argmin}} \{\|\mathbf{X} - \mathbf{u}\mathbf{v}^T\|_F^2 + \mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u}, \mathbf{v})\}, \tag{3}$$

where $\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u}, \mathbf{v})$ is a regularization penalty and $\boldsymbol{\lambda}$ is a vector of regularization parameters. Huang, Shen and Buja (2009) suggested to use the following specific form of the penalty function:

$$\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u}, \mathbf{v}) = \lambda_{\mathbf{u}}\mathbf{u}^T\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{u} \cdot \|\mathbf{v}\|^2 + \lambda_{\mathbf{v}}\mathbf{v}^T\boldsymbol{\Omega}_{\mathbf{v}}\mathbf{v} \cdot \|\mathbf{u}\|^2 + \lambda_{\mathbf{u}}\mathbf{u}^T\boldsymbol{\Omega}_{\mathbf{u}}\mathbf{u} \cdot \lambda_{\mathbf{v}}\mathbf{v}^T\boldsymbol{\Omega}_{\mathbf{v}}\mathbf{v}, \tag{4}$$

where $\boldsymbol{\Omega}_{\mathbf{u}}$ and $\boldsymbol{\Omega}_{\mathbf{v}}$ are symmetric and nonnegative definite penalty matrices that apply, respectively, to the left and right singular vectors, and $\|\cdot\|$ is the Euclidean norm. The usual roughness penalties used in nonparametric smoothing literature can be adopted to define the penalty matrices [e.g., Green and Silverman (1994)]. This penalty function enjoys several desirable properties: (i) Invariance under scale transformations $\mathbf{u} \mapsto c\mathbf{u}$ and $\mathbf{v} \mapsto \mathbf{v}/c$ for some positive constant c; (ii) Equivariance under rescaling of $\mathbf{X}$ and the fit $\mathbf{u}\mathbf{v}^T$; (iii) For $\boldsymbol{\Omega}_u = 0$, the penalty specializes to the one-way penalty of Silverman (1996) for functional PCA. See Huang, Shen and Buja (2009) for more discussions.

To achieve robustness, we replace the squared-error loss in (3) with a robust loss function. Let $\rho(z)$ be a nonnegative, symmetric function that is increasing in $|z|$. With a slight abuse of notation, we also use $\rho(\cdot)$ to denote the summation over elementwise applications when the scalar function $\rho(\cdot)$ is applied to a matrix. A general loss function for rank-one approximation of the matrix $\mathbf{X}$ can be written as

$$\rho\left(\frac{\mathbf{X} - \mathbf{u}\mathbf{v}^T}{\sigma}\right) = \sum_{i=1}^{m}\sum_{j=1}^{n} \rho\left(\frac{x_{ij} - u_i v_j}{\sigma}\right),$$

where σ is a scale parameter measuring the variability in the approximation errors. For RobRSVD, we define the first pair of singular vectors as

$$(\widehat{\mathbf{u}}, \widehat{\mathbf{v}}) = \underset{(\mathbf{u},\mathbf{v})}{\operatorname{argmin}} R(\mathbf{u}, \mathbf{v}),$$

where

$$R(\mathbf{u}, \mathbf{v}) = \rho\left(\frac{\mathbf{X} - \mathbf{u}\mathbf{v}^T}{\sigma}\right) + \mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u}, \mathbf{v}) \tag{5}$$

and $\mathcal{P}_{\boldsymbol{\lambda}}(\mathbf{u}, \mathbf{v})$ is the penalty function defined in (4). The determination of the scale parameter σ will be discussed later in Section 2.4.

Our implementation uses the following Huber's function in defining $R(\mathbf{u}, \mathbf{v})$:

$$\rho_\theta(x) = \begin{cases} x^2, & \text{if } |x| \le \theta, \\ 2\theta|x| - \theta^2, & \text{if } |x| > \theta, \end{cases}$$

where θ is a parameter that controls the robustness level and a smaller value of θ usually leads to more robust estimation. Our implementation uses $\theta = 1.345$, the value commonly used in robust regression that produces 95% efficiency for normal errors [Huber and Ronchetti (2009)]. Our numerical studies suggested that the RobRSVD is not very sensitive to the choice of θ. Instead of the Huber function, other robust loss functions can be used as well, for example, the L_1 loss which gives similar estimates. We choose the Huber function due to its easier implementation and faster computation.

2.2. *Iterative reweighted penalized least squares algorithm.* Although $\rho(\cdot)$ is a convex function, $R(\mathbf{u}, \mathbf{v})$ is not convex with respect to the pair $(\mathbf{u}, \mathbf{v})$ and, thus, simultaneous optimization of $R(\mathbf{u}, \mathbf{v})$ over $\mathbf{u}$ and $\mathbf{v}$ is complicated. Note that, conditional on either $\mathbf{u}$ or $\mathbf{v}$, $R(\mathbf{u}, \mathbf{v})$ becomes a convex function of the other variable. This naturally suggests an iterative reweighted (penalized) least squares (IRLS) algorithm that alternately updates $\mathbf{u}$ and $\mathbf{v}$, assuming that the penalty parameters $\lambda_{\mathbf{u}}$ and $\lambda_{\mathbf{v}}$ are fixed values. This section gives the details of the algorithm, while the choice of penalty parameters will be discussed later in Section 2.3.

For notational simplicity we assume $\sigma = 1$, since otherwise σ can be absorbed into $\rho(\cdot)$. Let u_i denote the ith element in $\mathbf{u}$, and v_j denote the jth element of $\mathbf{v}$. Let $\mathbf{x}_j$ denote the jth column, and $\mathbf{x}^{(i)}$ denote the ith row of $\mathbf{X}$. Let $\operatorname{Svec}(\mathbf{X}) = (x_{11}, x_{21}, \ldots, x_{m1}, x_{12}, \ldots, x_{mn})^T$ be the column vector that is obtained by stacking the columns of $\mathbf{X}$. Furthermore, let $\psi(x) = \rho'(x)$, $W(x) = \psi(x)/x$, and $\mathbf{W} = (w_{ij})$, where $w_{ij} = W(x_{ij} - u_i v_j)$.

Now we consider optimization of $R(\mathbf{u}, \mathbf{v})$ over $\mathbf{v}$ given $\mathbf{u}$. Taking the derivative of $R(\mathbf{u}, \mathbf{v})$ in (5) with respect to v_j, we have

$$\frac{\partial R}{\partial v_j} = \sum_{i=1}^{m} w_{ij}(x_{ij} - u_i v_j)(-u_i) + \frac{\partial \mathcal{P}_{\lambda}(\mathbf{u}, \mathbf{v})}{\partial v_j}, \tag{6}$$

where

$$\frac{\partial \mathcal{P}_\lambda(\mathbf{u},\mathbf{v})}{\partial \mathbf{v}} = 2\{\mathbf{u}^T(I+\lambda_{\mathbf{u}}\Omega_{\mathbf{u}})\mathbf{u}(I+\lambda_{\mathbf{v}}\Omega_{\mathbf{v}}) - \mathbf{u}^T\mathbf{u}\}\mathbf{v}.$$

The root of $\partial R/\partial v_j = 0$ then gives us the optimizer with respect to v_j.

Let

$$\mathcal{Y} = \mathrm{Svec}(\mathbf{X}) = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \vdots \\ \mathbf{x}_n \end{pmatrix}, \qquad \mathcal{U} = \begin{pmatrix} \mathbf{u} & 0 & \cdots & 0 \\ 0 & \mathbf{u} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{u} \end{pmatrix},$$

$\mathcal{W} = \mathrm{diag}\{\mathrm{Svec}(\mathbf{W})\}$, and $\Omega_{\mathbf{v}|\mathbf{u}} = \mathbf{u}^T(I+\lambda_{\mathbf{u}}\Omega_{\mathbf{u}})\mathbf{u}(I+\lambda_{\mathbf{v}}\Omega_{\mathbf{v}}) - (\mathbf{u}^T\mathbf{u})I$. The equations $\partial R/\partial v_j = 0$ lead to

$$\mathcal{U}^T\mathcal{W}\mathcal{U}\mathbf{v} + 2\Omega_{\mathbf{v}|\mathbf{u}}\mathbf{v} = \mathcal{U}^T\mathcal{W}\mathcal{Y}.$$

Solving for $\mathbf{v}$, we obtain

$$\widehat{\mathbf{v}} = (\mathcal{U}^T\mathcal{W}\mathcal{U} + 2\Omega_{\mathbf{v}|\mathbf{u}})^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}, \tag{7}$$

which is the updating formula for $\mathbf{v}$ given $\mathbf{u}$. It is easy to see that this $\widehat{\mathbf{v}}$ minimizes the following penalized weighted sum of squares:

$$\widetilde{R}(\mathbf{u},\mathbf{v}) = (\mathcal{Y}-\mathcal{U}\mathbf{v})^T\mathcal{W}(\mathcal{Y}-\mathcal{U}\mathbf{v}) + \mathbf{v}^T\Omega_{\mathbf{v}|\mathbf{u}}\mathbf{v}. \tag{8}$$

The equation for the fitted value of $\mathcal{Y}$ is

$$\widehat{\mathcal{Y}} = \mathcal{U}\widehat{\mathbf{v}} = \mathcal{U}(\mathcal{U}^T\mathcal{W}\mathcal{U} + 2\Omega_{\mathbf{v}|\mathbf{u}})^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}.$$

Equivalently, we denote $\widehat{\mathcal{Y}} = \mathcal{H}\mathcal{Y}$ with the hat matrix $\mathcal{H}$ defined as

$$\mathcal{H} = \mathcal{U}(\mathcal{U}^T\mathcal{W}\mathcal{U} + 2\Omega_{\mathbf{v}|\mathbf{u}})^{-1}\mathcal{U}^T\mathcal{W}.$$

Similarly, let

$$\mathcal{Y}^* = \mathrm{Svec}(\mathbf{X}^T) = \begin{pmatrix} \mathbf{x}^{(1)} \\ \mathbf{x}^{(2)} \\ \vdots \\ \mathbf{x}^{(m)} \end{pmatrix}, \qquad \mathcal{V} = \begin{pmatrix} \mathbf{v} & 0 & \cdots & 0 \\ 0 & \mathbf{v} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{v} \end{pmatrix},$$

$\mathcal{W}^* = \mathrm{diag}\{\mathrm{Svec}(\mathbf{W}^T)\}$, and $\Omega_{\mathbf{u}|\mathbf{v}} = \mathbf{v}^T(I+\lambda_{\mathbf{v}}\Omega_{\mathbf{v}})\mathbf{v}(I+\lambda_{\mathbf{u}}\Omega_{\mathbf{u}}) - (\mathbf{v}^T\mathbf{v})I$. Setting $\partial R/\partial u_i = 0$, we have

$$\mathcal{V}^T\mathcal{W}^*\mathcal{V}\mathbf{u} + 2\Omega_{\mathbf{u}|\mathbf{v}}\mathbf{u} = \mathcal{V}^T\mathcal{W}^*\mathcal{Y}^*.$$

Solving for $\mathbf{u}$ gives the following updating formula for $\mathbf{u}$ given $\mathbf{v}$:

$$\widehat{\mathbf{u}} = (\mathcal{V}^T\mathcal{W}^*\mathcal{V} + 2\Omega_{\mathbf{u}|\mathbf{v}})^{-1}\mathcal{V}^T\mathcal{W}^*\mathcal{Y}^*. \tag{9}$$

This $\widehat{\mathbf{u}}$ also solves a penalized weighted least squares problem and the corresponding hat matrix is $\mathcal{H}^* = \mathcal{V}(\mathcal{V}^T\mathcal{W}^*\mathcal{V} + 2\Omega_{\mathbf{u}|\mathbf{v}})^{-1}\mathcal{V}^T\mathcal{W}^*$.

The IRLS algorithm takes the results from the SVD as the initial values, and alternately applies (7) and (9) until convergence. The convergence of the algorithm is guaranteed because each iteration step reduces the objective function, which has a lower bound. For identifiability, at the end of each iteration step, we normalize both $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ to have unit L_2 norm. Upon convergence, the normalizing constant obtained in the last iteration step will be the estimate for the corresponding singular value.

Note that the weighting matrix $\mathbf{W}$ needs to be updated at each iteration. The matrix computation in (7) and (9) can be efficiently implemented using the block diagonal structure of the matrices. Let $\mathbf{w}_{(j)}$ be the jth column of $\mathbf{W}$, and $\mathbf{w}^{(i)}$ be the ith row of $\mathbf{W}$. It can be shown that

$$
\begin{aligned}
\mathcal{U}^T\mathcal{W}\mathcal{U} &= \operatorname{diag}\Big\{\sum_j \mathbf{u}^T \operatorname{diag}(\mathbf{w}_j)\mathbf{u}\Big\},\\
\mathcal{V}^T\mathcal{W}^*\mathcal{V} &= \operatorname{diag}\Big\{\sum_i \mathbf{v}^T \operatorname{diag}(\mathbf{w}^{(i)})\mathbf{v}\Big\},\\
\mathcal{U}^T\mathcal{W}\mathcal{Y} &= \operatorname{diag}\Big\{\sum_j \mathbf{u}^T \operatorname{diag}(\mathbf{w}_j)\mathbf{x}_j\Big\},\\
\mathcal{V}^T\mathcal{W}^*\mathcal{Y}^* &= \operatorname{diag}\Big\{\sum_i \mathbf{v}^T \operatorname{diag}(\mathbf{w}^{(i)})\mathbf{x}^{(i)}\Big\}.
\end{aligned}
\tag{10}
$$

These identities help significantly simplify the matrix computation. Moreover, sparse matrix algorithms can be applied for efficient computation since both $\Omega_{\mathbf{u}|\mathbf{v}}$ and $\Omega_{\mathbf{v}|\mathbf{u}}$ are banded matrices.

2.3. *Penalty parameter selection.* Following Huang, Shen and Buja (2009), we nest penalty parameter selection inside the alternating algorithm that optimizes $\mathbf{u}$ for fixed $\mathbf{v}$, and $\mathbf{v}$ for fixed $\mathbf{u}$. Let $\widehat{\mathbf{v}}^* = (\mathcal{U}^T\mathcal{W}\mathcal{U})^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}$ denote the unregularized update of $\mathbf{v}$, that is, the update of $\mathbf{v}$ corresponding to $\lambda_{\mathbf{v}} = 0$. The GCV criterion for selecting $\lambda_{\mathbf{v}}$ conditional on $\lambda_{\mathbf{u}}$ is

$$
\operatorname{GCV}(\lambda_{\mathbf{v}}|\lambda_{\mathbf{u}}) = \frac{\|\widehat{\mathbf{v}} - \widehat{\mathbf{v}}^*\|^2/n}{(1 - \operatorname{tr}(\mathcal{H})/n)^2}.
$$

Let $\widehat{\mathbf{u}}^* = (\mathcal{V}^T\mathcal{W}^*\mathcal{V})^{-1}\mathcal{V}^T\mathcal{W}^*\mathcal{Y}^*$ denote the unregularized update of $\mathbf{u}$. The GCV criteria for selecting $\lambda_{\mathbf{u}}$ conditional on $\lambda_{\mathbf{v}}$ is

$$
\operatorname{GCV}(\lambda_{\mathbf{u}}|\lambda_{\mathbf{v}}) = \frac{\|\widehat{\mathbf{u}} - \widehat{\mathbf{u}}^*\|^2/m}{(1 - \operatorname{tr}(\mathcal{H}^*)/m)^2}.
$$

These GCV formulas can be derived as a modification of appropriately defined leave-one-row/column-out cross-validation criteria. Details of the derivation are given in Section 1 of the online supplemental article [Zhang, Shen and Huang (2013)]. We minimize the GCV criterion to select the optimal penalty parameters, which is done by using grid search in our implementation. Penalty parameter selection using the GCV formulas has much less computational complexity than directly using cross-validation. In our numerical experiments it usually took seconds for one entire iteration of the algorithm including penalty parameter selection.

2.4. *Estimation of* σ. We have fixed the scale parameter σ in our development so far. In practice, σ can be estimated from the data using residuals from a preliminary rank-one approximation of $\mathbf{X}$. Specifically, consider the residual matrix $R = (r_{ij}) = \mathbf{X} - \widehat{\mathbf{u}}\widehat{\mathbf{v}}^T$, where $\widehat{\mathbf{u}}\widehat{\mathbf{v}}^T$ is a rank-one matrix. The normalized Median Absolute Deviation (MAD), defined as

$$\widehat{\sigma} = \frac{1}{0.675}\operatorname{Med}_{ij}\bigl(|r_{ij}|, r_{ij} \neq 0\bigr), \tag{11}$$

provides an estimate of σ [Maronna, Martin and Yohai (2006)]. In (11), the $\widehat{\mathbf{u}}$ and $\widehat{\mathbf{v}}$ can be obtained using the SVD or by minimizing a robust loss function in rank-one approximation. We found that using the SVD works very well and there is no need to resort to a computationally more complicated robust loss function. The RobRSVD procedure can also be applied iteratively, where residuals from previous application are used to estimate the scale parameter, but our experience suggests that such iteration is usually not necessary. Hence, standard SVD is used to estimate the scale parameter for our numerical studies.

2.5. *Missing values*. In some situations, the data set may contain missing values, such as the mortality data set analyzed in this paper or sparse functional data as discussed in Yao, Müller and Wang (2005). The IRLS algorithm can still be applied with some slight modification on the updating equations (7) and (9). One approach is to redefine $\mathcal{Y}$, $\mathcal{U}$, $\mathcal{W}$, $\mathcal{Y}^*$, $\mathcal{V}$ and $\mathcal{W}^*$ by removing the rows/columns of these matrices that contain the missing entries. However, this approach is computationally inefficient, since the calculation of $\mathcal{U}^T\mathcal{W}\mathcal{U}$ and $\mathcal{V}^T\mathcal{W}^*\mathcal{V}$ cannot be simplified as in (10).

Below we develop a more efficient algorithm to deal with missing entries. We propose to iteratively impute the missing values and then apply the IRLS algorithm. Each missing entry X_{ij} is replaced by $\widehat{u}_i\widehat{v}_j$, where $\widehat{u}_i$ and $\widehat{v}_j$ are obtained from the previous iteration. The initial round of imputation can use either the row-wise mean of the nonmissing entries in the same row or the column-wise mean of the nonmissing entries in the same column. Our experience suggests that both initialization methods lead to the same results at convergence. Our proposed imputation approach can be thought as an application of the MM algorithm [Hunter and Lange (2004)], which has nice convergence properties; see Section 2 of the online

supplemental article for details [Zhang, Shen and Huang (2013)]. Similar iterative imputation approaches have been used in the literature; see, for example, Beckers and Rixen (2003), Martinez et al. (2009) and Lee, Huang and Hu (2010).

2.6. *Function space view.* So far our formulation of RobRSVD is in finite dimensions, although the use of regularization penalties implicitly assumes that there are underlying smooth functions. We now use the Reproducing Kernel Hilbert Space (RKHS) theory to extend our formulation to function spaces. We refer to a standard reference such as Wahba (1990) for the necessary background.

We assume $\mathbf{X} = (X(y_i, z_j))_{i=1,\ldots,n;\, j=1,\ldots,m}$ contains the evaluations of a realization of a random field $X(y, z)$ at (y_i, z_j), where y_i and z_j are distinct sampling points in the respective domains $\mathcal{Y}$ and $\mathcal{Z}$. Seeking a rank-one or product approximation $X(y, z) \simeq U(y)V(z)$ in function spaces, we assume that $U(y)$ and $V(z)$ are members of RKHSs $\mathcal{H}_u$ and $\mathcal{H}_v$ defined, respectively, on the domains $\mathcal{Y}$ and $\mathcal{Z}$. The RKHSs carry reproducing kernels $K_u(y_1, y_2)$ and $K_v(z_1, z_2)$, inner products $\langle U_1, U_2\rangle_u$ and $\langle V_1, V_2\rangle_v$, as well as norms $\|U\|_u$ and $\|V\|_v$, respectively. For arbitrary $\mathbf{u} = (u_1, \ldots, u_n)^T \in \mathbb{R}^n$ there is a unique $U \in \mathcal{H}_u$ interpolating $\mathbf{u}$, that is, satisfying $u_i = U(y_i)$ $(i = 1, \ldots, n)$ and having minimum norm $\|U\|_u$ among all interpolants. Moreover, this function is of the form $U(y) = \sum_{i=1,\ldots,n} c_i K_u(y_i, y)$, and $\|U\|_u^2 = \mathbf{u}^T \mathbf{\Omega}_u \mathbf{u}$, where $\mathbf{\Omega}_u = \mathbf{K}_u^{-1}$ and $\mathbf{K}_u = (K_u(y_{i'}, y_{i''}))_{i',i''=1,\ldots,n}$. The same argument yields $V(z) = \sum_{j=1,\ldots,m} d_j K_v(z_j, z)$ for given $\mathbf{v} \in \mathbb{R}^m$, and $\mathbf{\Omega}_v = \mathbf{K}_v^{-1}$. The function space version of the criterion $R(\mathbf{u}, \mathbf{v})$ (5) is (with some abuse of notation)

$$
\begin{aligned}
R(U, V) = \rho\left(\frac{\mathbf{X} - \mathbf{u}\mathbf{v}^T}{\sigma}\right) &+ \lambda_u \|U\|_u^2 \|\mathbf{v}\|^2 \\
&+ \lambda_v \|\mathbf{u}\|^2 \|V\|_v^2 + \lambda_u \|U\|_u^2 \cdot \lambda_v \|V\|_v^2,
\end{aligned} \tag{12}
$$

where $\mathbf{u} = (U(y_1), \ldots, U(y_n))^T$ and $\mathbf{v} = (V(z_1), \ldots, V(z_m))^T$.

The representer theorem argument [Kimeldorf and Wahba (1971)] shows that minimization of $R(U, V)$ in the RKHSs can be reduced to minimization of $R(\mathbf{u}, \mathbf{v})$ in the finite-dimensional space. Specifically, if $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ are minimizers of $R(\mathbf{u}, \mathbf{v})$, and $\tilde{U}$ and $\tilde{V}$ are their unique interpolants in RKHSs $\mathcal{H}_u$ and $\mathcal{H}_v$, then $\tilde{U}$ and $\tilde{V}$ are the minimizers of $R(U, V)$. This result suggests that our methodological discussions in finite-dimensional space are without loss of generality. An important application of this result, however, is that it allows us to extend the output vectors $\tilde{\mathbf{u}}$ and $\tilde{\mathbf{v}}$ to their function space counterparts $\tilde{U}$ and $\tilde{V}$ through the RKHS interpolation.

In the nonparametric smoothing literature, an integrated squared second derivative penalty is commonly used. Applying this penalty to our setting means using $\|U\|_u^2 = \int \{U''(y)\}^2 \, dy$ and $\|V\|_v^2 = \int \{U''(z)\}^2 \, dz$ in (12). The corresponding RKHSs $\mathcal{H}_u$ and $\mathcal{H}_v$ are Sobolev spaces of functions with reproducing kernels defined in Chapter 1 of Wahba (1990). On the other hand, for this special kind of

penalty, we do not need the machinery of RKHS for connecting finite-dimensional and functional spaces. We can resort to the standard results of natural cubic splines [see Chapter 1 of Green and Silverman (1994)]. There are closed-form expressions of the penalty matrices in terms of the evaluation points and interpolation formulas available; see Section 5 of Huang, Shen and Buja (2008).

3. Simulation studies. Three simulation studies were conducted to compare the performance of RobRSVD against the standard SVD and the RSVD of Huang, Shen and Buja (2009). The underlying true signal matrix was generated to be either rank one, or rank one with missing values, or rank two. A detailed analysis of the rank-one signal matrix is reported in Section 3.1. To save space, we only summarize the findings for the other two settings in Sections 3.2 and 3.3, and present details of the studies in Section 3 of the online supplemental document [Zhang, Shen and Huang (2013)].

3.1. *Rank-one signal matrix.* We consider the following rank-one two-way functional model:

$$X(y,z) = s_0 u_0(y) v_0(z) + \varepsilon(y,z), \tag{13}$$

where $s_0 = 773$ is a scalar, and the two functions are $u_0(y) = (\log 10/9)10^y$ and $v_0(z) = (1 + 1/\pi)^{-1} \sin(2\pi z)$ with $y \in [0,1]$, $z \in [0,1]$. Note that (13) is slightly different from the general two-way functional model (1) in that the two functions are now normalized: $\int_0^1 u_0^2(y)\,dy = 1$ and $\int_0^1 v_0^2(z)\,dz = 1$, which makes it necessary to have the scalar s_0. To simulate the functional data matrix, we consider 100 equal-spaced grids in either direction. The true two-way signal surface without any noise is plotted in panel (a) of Figure 1.

As a benchmark scenario, we consider the situation where the data have no outliers. In addition, we study four different scenarios that outliers can occur in two-way functional data: (1) random outlying cells, (2) outlying rows, (3) outlying blocks, and (4) diagonal outlying cells. Under each setting, the outliers are introduced as discussed below. Besides the outliers, independent Gaussian noises $\varepsilon(y,z)$ with mean 0 and variance σ^2 are added to the simulated data. We consider different variances: $\sigma^2 = 0.2, 0.5, 0.8, 1$. For each simulation setting, 100 simulation replications are performed. The surface plot of one random replication (with $\sigma^2 = 1$) is plotted in Figure 1 for each of the four outlying scenarios, respectively.

We now describe how the outliers are introduced for each simulation setting. Let $\mathbf{X}_0 = s_0 \mathbf{u}_0 \mathbf{v}_0^T$ denote the signal matrix (i.e., without any noise), where $\mathbf{u}_0$ (or $\mathbf{v}_0$) denotes the vector that contains the observed values of the function $u_0(y)$ [or $v_0(z)$] at the 100 equally-spaced grid points within $[0,1]$:

1. *Outlying cells*: Under this setting, we randomly select 100 cells in the data and replace their entries with outlying values. In particular, the values in the selected cells are randomly simulated from the uniform distribution with support $[C_1, 2C_1]$ with $C_1 = \max(\mathbf{X}_0)$.

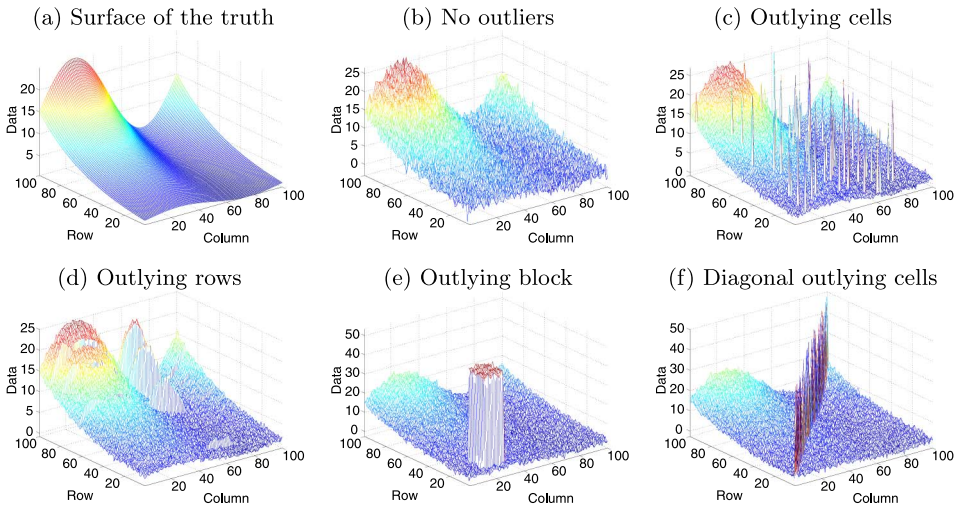

FIG. 1. *Rank-one simulation*: *The surface plots*. (a) *No noise and no outliers*, (b) *no outliers with noise*, (c) *random outlying cells*, (d) *outlying rows*, (e) *outlying blocks*, *and* (f) *diagonal outlying cells*.

2. *Outlying rows*: We randomly select five rows, and replace them by five new rows defined below. For each of the five randomly selected rows, we obtain the outlying curve by multiplying the corresponding $s_0 u_0(y)$ with a different function $v_1(z) = C(1 + \sin(4\pi z))$ with C being the normalizing constant. Note that the curve shapes of the outlying rows are different from the shape of the other rows.

3. *Outlying block*: We randomly select a continuous square block of cells at a randomly selected location, with the block size fixed as 10×10. Within the block, we shift the cells upward by adding a random amount, which is uniformly distributed on $[2C_1, 3C_1]$.

4. *Diagonal outliers*: We replace the diagonal entries of the matrix with values uniformly distributed between $[C_1, 2C_1]$. This setting mimics the cohort effects observed in the Spanish mortality data (Section 4).

The three methods, SVD, RSVD and RobRSVD, were applied to the 100 simulated data sets under each setting, and the best rank-one approximations were obtained to get the estimates for $\mathbf{u}_0$ and $\mathbf{v}_0$. The penalty parameters of the RSVD and RobRSVD were selected using the GCV method.

To compare various methods, we calculated the L_2 distance between the estimates and the truth for each simulated data. Figures 2 and 3 present the boxplots of the 100 distances for the three methods for $\mathbf{u}_0$ and $\mathbf{v}_0$, respectively, for each of the four noise levels and each of the outlier scenarios.

In summary, both figures clearly show that:

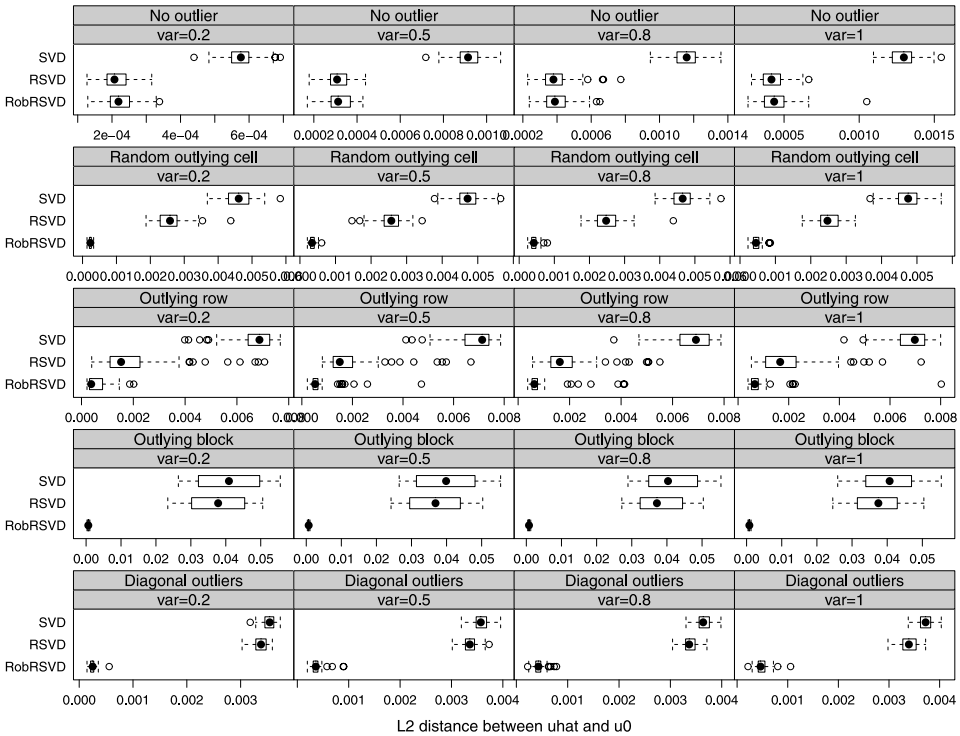

FIG. 2. *Rank-one simulation*: *Boxplots of the* L_2 *distance between* $\widehat{\mathbf{u}}_0$ *and* $\mathbf{u}_0$.

1. For the benchmark no-outlier cases, RobRSVD and RSVD perform comparably, and both are better than SVD due to smoothing regularization; this suggests that our RobRSVD does not lose much when the data contain no outliers.

2. For all the outlying settings, RobRSVD improves significantly over RSVD and SVD, which supports the robustness of RobRSVD against various kinds of outliers in two-way functional data. RobRSVD has the smallest median L_2 distance and variability across all the settings and the different noise levels.

We also calculated the estimated singular values $\widehat{s}_0$ and compared them with the true singular value $s_0 = 773$. For each method and each noise level, Figure 4 presents the boxplot of the 100 absolute differences between $\widehat{s}_0$ and s_0. The comparison shows that RobRSVD performs similarly with SVD and RSVD for cases with no outliers, while much better when there are outliers.

To get some ideas about individual estimation performance, Figure 5 compares the estimates obtained from the particular data sets shown in Figure 1, by plotting the differences between the estimated curves and the true curve [either $u_0(\cdot)$ or $v_0(\cdot)$]. As one can see, the RobRSVD method is again the clear winner. We also observe that the smoothing step in RSVD can mitigate the outlying effects to some extent in certain cases, but still cannot fully remove those effects. The additional

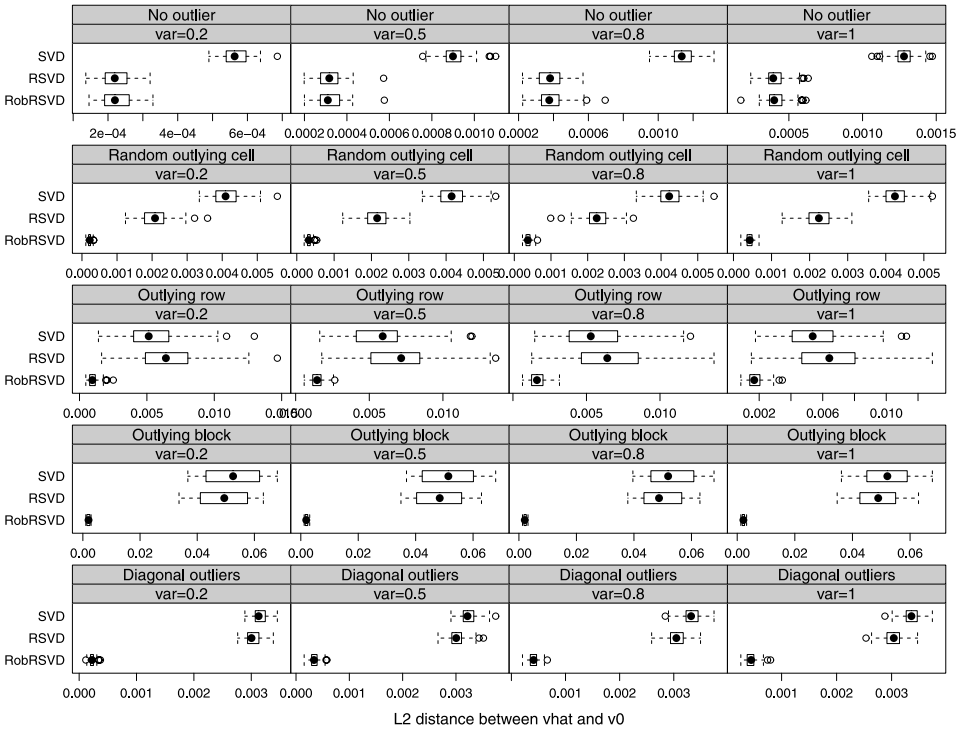

FIG. 3. *Rank-one simulation*: *Boxplots of the* L_2 *distance between* $\widehat{\mathbf{v}}_0$ *and* $\mathbf{v}_0$.

incorporation of robust loss function in RobRSVD further improves the robustness of RSVD.

3.2. *Rank-one signal matrix with missing values*. Our motivating Spanish mortality data contain both outliers and missing values, which motivates us to investigate the performance of RobRSVD when there are missing values. For each simulated data set considered in Section 3.1, we randomly selected and deleted 100 cells from it to form a new data set with missing values. We used the imputation method described in Section 2.5 to estimate $\mathbf{u}$ and $\mathbf{v}$ for SVD, RSVD and RobRSVD.

The simulation results are reported in the online supplement. The comparison presented in Figures 1 and 2 there clearly shows that the RobRSVD remains to be the winner across all the settings considered.

3.3. *Rank-two signal matrix*. We also studied the situation where the true signal matrix is rank two, using a setting similar to what has been studied by Huang, Shen and Buja (2009). Similar to Section 3.1, we considered five simulation scenarios: no outliers, outlying cells, outlying rows, outlying block, and diagonal out-

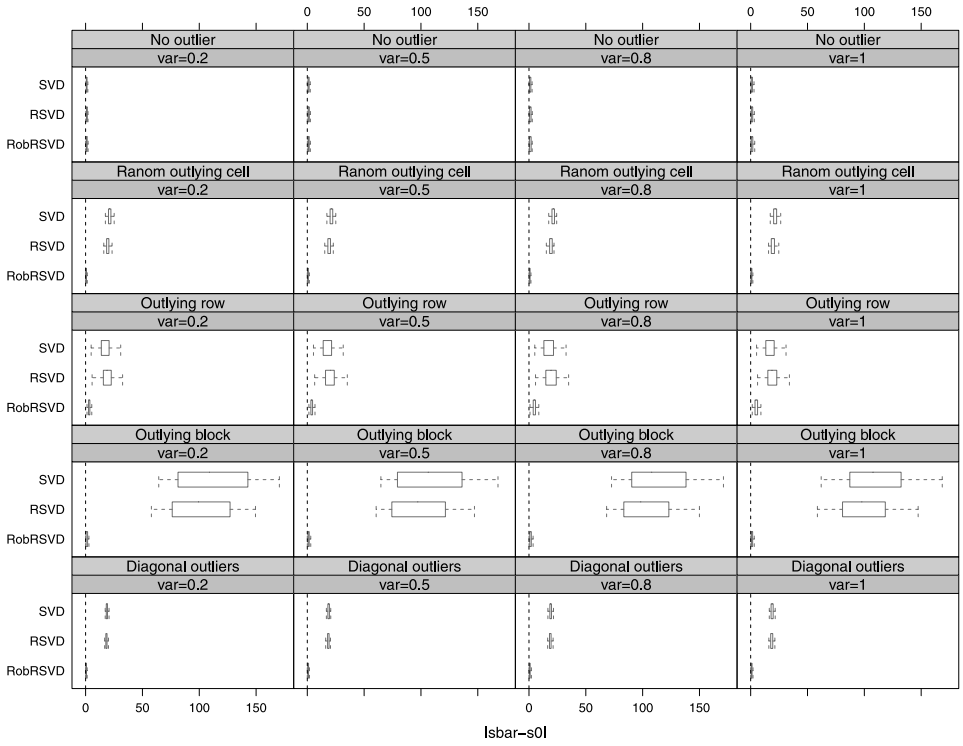

FIG. 4. *Rank-one simulation*: *Boxplots of* $|\widehat{s}_0 - s_0|$.

liers. Detailed descriptions can be found in the online supplement, with comparative results presented in Figures 3–5 there.

We used two measures to gauge the performance of estimating the rank-2 signal matrix. The first measure is $\|\widehat{\mathbf{X}}_0 - \mathbf{X}_0\|_F$, the Frobenius norm of the difference between the estimated best rank-two matrix $\widehat{\mathbf{X}}_0$ and the true signal matrix $\mathbf{X}_0$. The second measure the largest *principal angle* [Golub and Van Loan (1996)] between the true subspace and the subspace spanned by the corresponding singular vector estimates. Specifically, let $\mathbf{U} = \text{span}(U_1^*, U_2^*)$ denote the linear subspace spanned by $U_1^*(y)$ and $U_2^*(y)$ evaluated at the grid points and $\widehat{\mathbf{U}}$ be the corresponding estimate of this subspace. The principal angle between $\mathbf{U}$ and $\widehat{\mathbf{U}}$ can be computed as $\cos^{-1}(\rho) \times 180/\pi$, where ρ is the minimum eigenvalue of the matrix $Q_{\widehat{\mathbf{U}}}^T Q_{\mathbf{U}}$ where $Q_{\widehat{\mathbf{U}}}$ and $Q_{\mathbf{U}}$ are orthogonal basis matrices obtained by the QR decomposition of the matrices $\widehat{\mathbf{U}}$ and $\mathbf{U}$, respectively. RobRSVD performed the best in all cases with outliers under both distance measures, while RSVD and RobRSVD usually performed similarly and were better than SVD in cases without outliers.

4. The Spanish mortality data. In this section we analyze the Spanish mortality data using various methods to illustrate the benefits of our proposed RobRSVD method. The Spanish mortality data are available in the Human Mortality

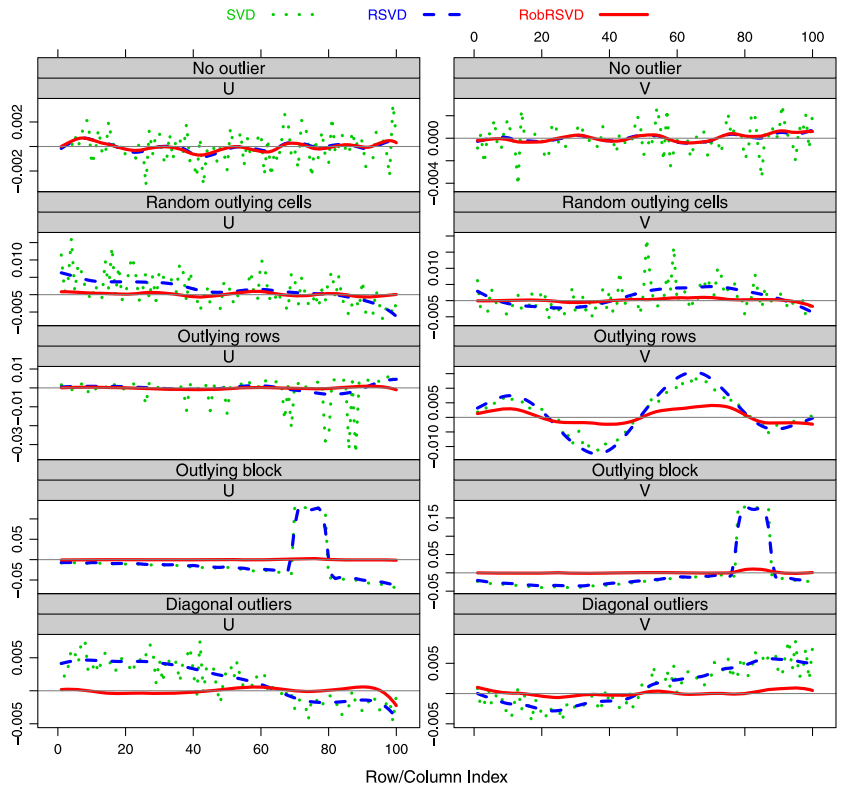

FIG. 5. *Rank-one simulation*: *Comparison of individual estimates obtained from the data plotted in Figure* 1. *The differences between the estimates and the truth are plotted. The RobRSVD method shows the most robust performance.*

Database [HMD (2011)]. This mortality data set was collected such that each row represents a year between 1908 and 2007, each column represents an age group from 0 to 110, and each cell records the mortality rate for a particular age group during that year. The data are naturally two-way functional, since each column vector is a time series of mortality rate of a given age group, and each row vector is a mortality curve of different age groups at a specific year.

Zhang et al. (2007) developed several visualization tools for exploring two-way functional data, which were used to analyze a subset of the Spanish mortality data. As a result, they identified a couple of interesting outlying time periods (i.e., rows in the data matrix):

- the 1918 Spanish flu pandemic, and
- the 1936–1939 Spanish Civil War,

both of which experienced the death of millions of Spanish people (in an unusual age distribution). In both cases, the mortality rate increased well above what the

normal yearly trend would have predicted, and the authors noted that the outlying years affected the estimation of the first few leading SVD components, which is consistent with our findings reported below.

One can view the mortality rate data as some normal mortality trend, a function of age group and year, contaminated with additive noises, including measurement errors and potential outliers. Hence, a good estimation method should be able to recover the underlying normal mortality varying pattern across age and year, with minimal effects of the noises including the outliers.

Before the formal analysis, we make two comments regarding the data. Following Zhang et al. (2007), the data were first transformed through $\log_2(X + 1/2)$ where X denotes the original mortality rate. There are missing values for the elder people in the data, and we employ the procedure discussed in Section 2.5 to automatically accommodate the missing values.

Figure 6 provides several functional views of the log-transformed data. Several interesting observations can be made from the plots. The mesh surface plot in panel (a) highlights the high mortality rates among the seniors that are older than 100. To better depict the mortality trend among people less than 100 years old, the zoomed surface plot in panel (b) shows the mortality rate pattern up to age 100: for a given year, the mortality rate generally decreases from infants to teenagers and adults younger than 60, and begins to increase when the age is over 60, which is the standard mortality pattern across age; for a given age group, the mortality rate decreases across the years, which reflects the improvement of life quality and health care; in addition, the decrease-across-year among younger people is more significant than for elder people. For the (zoomed) image plots on panels (c)–(d), we observe the cohort effects discussed by Zhang et al. (2007) showing up as the diagonal strips and, more importantly, the two outlying time periods appearing as horizontal strips: the 1918 flu pandemic affects all age groups, while the 1936–1939 civil war affects only those older than 20. The curve plots in panel (e) show the mortality rate as a function of age where each curve corresponds to a particular year, and in panel (f) show the mortality rate as a function of year where each curve is for a particular age.

To better understand the dominating modes of variation within the data, we use SVD, RSVD and RobRSVD to find (smooth) low-rank approximations for the data and compare their results. Let s_i be the ith singular value for the standard SVD. The ratio of s_i^2 over the Fronbenius norm of the data matrix represents the percentage of energy explained by the ith component. The percentage can be plotted in a scree plot as a useful visual aid for deciding the number of significant components. For the mortality data, the scree plot based on the SVD shows a clear knee at rank two, with the first two standard SVD components explaining 93.3% and 5.0% of the total energy, respectively, while the third component accounts for less than 1.0% of the total energy. Thus, we only look at the first two dominating pairs of functional components when we compare different methods.

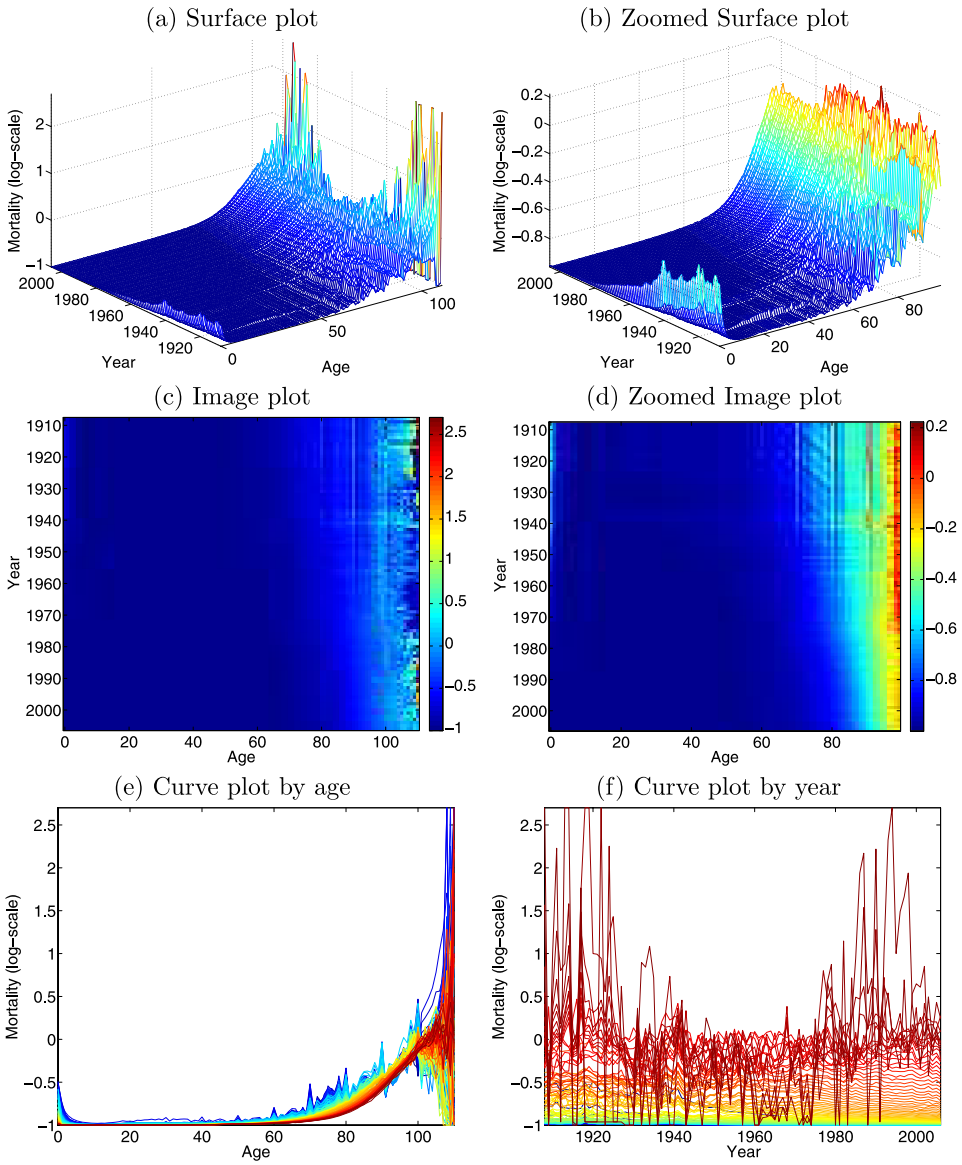

FIG. 6. *Various visualizations of the mortality data* (*log-scale*): (a) *the mesh surface plot*, (b) *the zoomed surface plot* (*up to age* 100), (c) *the image plot*, (d) *the zoomed image plot* (*up to age* 100), (e) *the curve plot versus age*, (f) *the curve plot versus year.*

Figure 7 compares the first left (regularized) singular vectors (RSVs) ($\mathbf{u}_1$) and the first right RSVs ($\mathbf{v}_1$), as well as the best rank-one two-way approximation from the three methods. Note that the first pair of RSVs explains the major mode of variation in the data. The green dotted-dash curves show the results of the regu-

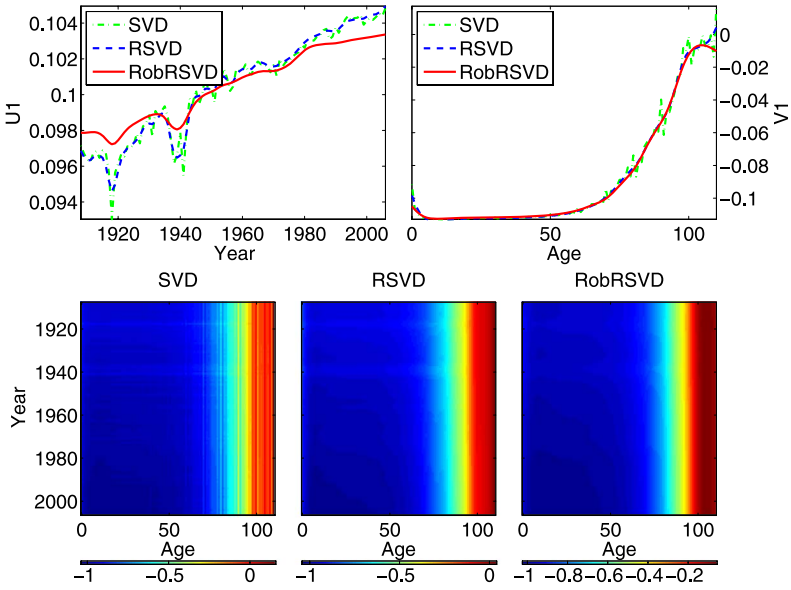

FIG. 7. *Comparison of the first pairs of* (*regularized*) *singular vectors. The left* (*regularized*) *singular vectors* ($\mathbf{u}_1$) *from SVD and RSVD are obviously affected by the two outlying time periods* (1918, 1936–1939).

lar SVD method, the blue dash ones correspond to the RSVD method, and the red solid curves are for our RobRSVD method. The RobRSVD left component shows a general smooth increasing trend from 1908 to 2007, while the corresponding right component resembles the standard smooth age-mortality curve. On the other hand, the left functional components from SVD and RSVD are rather wiggly and seriously affected by the two outlying time periods in 1918 and 1936–1939. The robustness of RobRSVD can also be seen from the image plots of the best rank-one approximation, the bottom row of Figure 7. For both SVD and RSVD approximations, the outlying years show up as horizontal strips to reflect the increased mortality rates across a wide range of age groups. Furthermore, the RobRSVD image plot shows a much smoother trend across age.

The second pair of (regularized) singular vectors is compared in Figure 8. In general, we observe that the RobRSVD component is smoother and more interpretable than the SVD and RSVD components, which tend to be wiggly and show effects from the outlying years. Note that the numerical scales of the colorbars for SVD/RSVD are much larger than those of RobRSVD, which are caused by the outliers appearing in the SVD/RSVD components. The second pair of the RobRSVD component highlights the contrast between people of age 50–100 and people older than 100 during two different time periods: before 1970, the older group has a lower mortality rate than the younger group, while after 1970, the comparison is reversed. This contrast can be clearly seen in the bottom right panel.

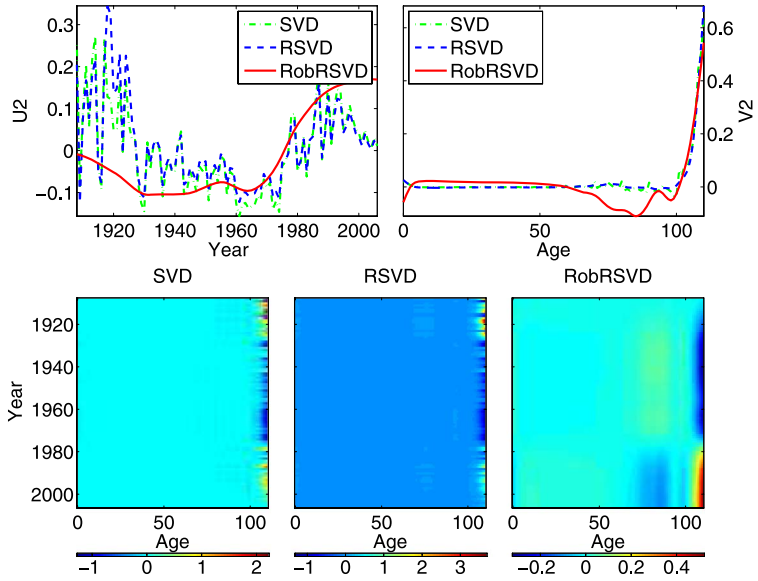

FIG. 8. *Comparison of the second pairs of (regularized) singular vectors. Note the contrast in the RobRSVD image plot around* 1970.

Figure 6 of the online supplement shows the 3-dimensional surface plots of the best rank-two approximations by the three methods, also indicating that the RobRSVD is least influenced by outlying observations.

Acknowledgments. We thank the Editor, the Associate Editor and the referees for invaluable comments and suggestions, which greatly improved the quality of this paper.

SUPPLEMENTARY MATERIAL

Supplemental notes for "Robust regularized singular value decomposition with application to mortality data" (DOI: 10.1214/13-AOAS649SUPP; .pdf). The supplemental notes include deviation of the GCV formula in this paper, an MM algorithm to handle missing value, two additional simulation examples in details, and one additional plot for the analysis of the mortality data.

REFERENCES

AMMANN, L. P. (1993). Robust singular value decompositions: A new approach to projection pursuit. *J. Amer. Statist. Assoc.* **88** 505–514. MR1224375

BAI, P., SHEN, H., HUANG, X. and TRUONG, Y. (2008). A supervised singular value decomposition for independent component analysis of fMRI. *Statist. Sinica* **18** 1233–1252. MR2468266

BALI, J. L., BOENTE, G., TYLER, D. E. and WANG, J.-L. (2011). Robust functional principal components: A projection-pursuit approach. *Ann. Statist.* **39** 2852–2882. MR3012394

BECKERS, J. and RIXEN, M. (2003). EOF calculations and data filling from incomplete oceanographic datasets. *J. Atmos. Oceanic Technol.* **20** 1839–1856.

CROUX, C., FILZMOSER, P., PISON, G. and ROUSSEEUW, P. J. (2003). Fitting multiplicative models by robust alternating regressions. *Stat. Comput.* **13** 23–36. MR1973864

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis*: *Theory and Practice*. Springer, New York. MR2229687

GABRIEL, K. R. and ZAMIR, S. (1979). Lower rank approximation of matrices by least squares with any choice of weights. *Technometrics* **21** 489–498.

GERVINI, D. (2008). Robust functional estimation using the median and spherical principal components. *Biometrika* **95** 587–600. MR2443177

GERVINI, D. (2009). Detecting and handling outlying trajectories in irregularly sampled functional datasets. *Ann. Appl. Stat.* **3** 1758–1775. MR2752157

GERVINI, D. (2010). The functional singular value decomposition for bivariate stochastic processes. *Comput. Statist. Data Anal.* **54** 163–172. MR2558467

GOLUB, G. H. and VAN LOAN, C. F. (1996). *Matrix Computations*, 3rd ed. Johns Hopkins Univ. Press, Baltimore, MD. MR1417720

GREEN, P. J. and SILVERMAN, B. W. (1994). *Nonparametric Regression and Generalized Linear Models*: *A Roughness Penalty Approach*. *Monographs on Statistics and Applied Probability* **58**. Chapman & Hall, London. MR1270012

HEIBERGER, R. M. and BECKER, R. A. (1992). Design of an S function for robust regression using iteratively reweighted least squares. *J. Comput. Graph. Statist.* **1** 181–196.

HMD (2011). Human mortality database. Available at www.mortality.org.

HUANG, J. Z., SHEN, H. and BUJA, A. (2008). Functional principal components analysis via penalized rank one approximation. *Electron. J. Stat.* **2** 678–695. MR2426107

HUANG, J. Z., SHEN, H. and BUJA, A. (2009). The analysis of two-way functional data using two-way regularized singular value decompositions. *J. Amer. Statist. Assoc.* **104** 1609–1620. MR2750581

HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust Statistics*, 2nd ed. Wiley, Hoboken, NJ. MR2488795

HUNTER, D. R. and LANGE, K. (2004). A tutorial on MM algorithms. *Amer. Statist.* **58** 30–37. MR2055509

HYNDMAN, R. J. and SHAHID ULLAH, M. (2007). Robust forecasting of mortality and fertility rates: A functional data approach. *Comput. Statist. Data Anal.* **51** 4942–4956. MR2364551

HYNDMAN, R. J. and SHANG, H. L. (2009). Forecasting functional time series. *J. Korean Statist. Soc.* **38** 199–211. MR2750314

KIMELDORF, G. and WAHBA, G. (1971). Some results on Tchebycheffian spline functions. *J. Math. Anal. Appl.* **33** 82–95. MR0290013

LEE, S., HUANG, J. Z. and HU, J. (2010). Sparse logistic principal components analysis for binary data. *Ann. Appl. Stat.* **4** 1579–1601. MR2758342

LIU, L., HAWKINS, D. M., GHOSH, S. and YOUNG, S. S. (2003). Robust singular value decomposition analysis of microarray data. *Proc. Natl. Acad. Sci. USA* **100** 13167–13172 (electronic). MR2016727

LOCANTORE, N., MARRON, J. S., SIMPSON, D. G., TRIPOLI, N., ZHANG, J. T. and COHEN, K. L. (1999). Robust principal component analysis for functional data. *TEST* **8** 1–73. MR1707596

MARONNA, R. A., MARTIN, R. D. and YOHAI, V. J. (2006). *Robust Statistics*: *Theory and Methods*. Wiley, Chichester. MR2238141

MARTINEZ, J. G., HUANG, J. Z., BURGHARDT, R. C., BARHOUMI, R. and CARROLL, R. J. (2009). Use of multiple singular value decompositions to analyze complex intracellular calcium ion signals. *Ann. Appl. Stat.* **3** 1467–1492. MR2752142

RAMSAY, J. O. and SILVERMAN, B. W. (2002). *Applied Functional Data Analysis*: *Methods and Case Studies*. Springer, New York. MR1910407

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

ROUSSEEUW, P. J. (1984). Least median of squares regression. *J. Amer. Statist. Assoc.* **79** 871–880. MR0770281

SHEN, H., ZHU, Z. and LEE, T. (2007). Robust estimation of the self-similarity parameter in network traffic using wavelet transform. *Signal Processing* **87** 2111–2124.

SILVERMAN, B. W. (1996). Smoothed functional principal components analysis by choice of norm. *Ann. Statist.* **24** 1–24. MR1389877

TIAN, T. S. and LI, Z. (2011). A spatio-temporal solution for the EEG/MEG inverse problem using group penalization methods. *Stat. Interface* **4** 521–533. MR2868834

WAHBA, G. (1990). *Spline Models for Observational Data. CBMS-NSF Regional Conference Series in Applied Mathematics* **59**. SIAM, Philadelphia, PA. MR1045442

YAO, F., MÜLLER, H.-G. and WANG, J.-L. (2005). Functional data analysis for sparse longitudinal data. *J. Amer. Statist. Assoc.* **100** 577–590. MR2160561

ZHANG, L., SHEN, H. and HUANG, J. (2013). Supplement to "Robust regularized singular value decomposition with application to mortality data." DOI:10.1214/13-AOAS649SUPP.

ZHANG, L., MARRON, J. S., SHEN, H. and ZHU, Z. (2007). Singular value decomposition and its visualization. *J. Comput. Graph. Statist.* **16** 833–854. MR2412485

L. ZHANG
DEPARTMENT OF STATISTICS
PURDUE UNIVERSITY
150 N. UNIVERSITY ST.
WEST LAFAYETTE, INDIANA 47906
USA
E-MAIL: lingsong@purdue.edu

H. SHEN
DEPARTMENT OF STATISTICS AND OPERATIONS RESEARCH
UNIVERSITY OF NORTH CAROLINA
CHAPEL HILL, NORTH CAROLINA 27599
USA
E-MAIL: haipeng@email.unc.edu

J. Z. HUANG
DEPARTMENT OF STATISTICS
TEXAS A&M UNIVERSITY
3143 TAMU
COLLEGE STATION, TEXAS 77843-3143
USA
E-MAIL: jianhua@stat.tamu.edu