Functional SVD for Big Data

Pan Chao

April 23, 2014

Outline

1. One-Way Functional SVD
   a) Interpretation
   b) Robustness
   c) CV/GCV
2. Two-Way Problem
3. Big data solution (split-and-recombine)
   a) Data split
   b) Recombine
4. Simulation
5. Summary and furture work
6. Reference

One-Way Problem

To estimate the functional principle component of the data, we can use SVD method. In order to incorporate the functional nature of the data, regularization is imposed on the estimates.

The first regularized principle component for functional data can be estimated by minimizing

$$\rho(Y - \mathbf{u}\mathbf{v}^T) + \lambda\|\mathbf{u}\|^2 \int v''^2,$$

where:

Y: data matrix, $Y_{ij} = Y_i(t_j)$.

$\mathbf{u}$: first left vector.

$\mathbf{v}$: first right vector, $\mathbf{v}_j = v(t_j)$.

Remarks:

1) Smoothness is controlled by the tuning parameter λ.
2) Adding $\|\mathbf{u}\|^2$ to make the problem scale-invariant [Huang et al. 2008].
3) $\rho(\cdot)$ is the loss function which measure the fidelity of the rank-1 approximation. If $\rho(\cdot) = \|\cdot\|^2$, then least square.
4) Due to the theory of smoothing spline [Green & Silverman], the integral has a matrix expression

$$\int \mathbf{v}''^2 = \mathbf{v}^T \Omega_v \mathbf{v}$$

where

$$\Omega_v = Q R^{-1} Q^T$$

Q and R are two banded matrices depending on the discretization of the data.

Smoothing Spline View

If Y has only one row, denoted by $\mathbf{y}$, then requiring $u = 1$ results in a standard smoothing spline problem:

$$\rho(\mathbf{x} - \mathbf{v}) + \lambda \mathbf{v}^T \Omega_v \mathbf{v}.$$

PCA View

If $\rho(\cdot) = \|\cdot\|_F^2$, the minimizer of $\mathbf{v}$ of the one-way problem without penalty is

$$\hat{\mathbf{v}} = \arg\max_{\mathbf{v}} \frac{\mathbf{v}^T Y^T Y \mathbf{v}}{\|\mathbf{v}\|^2},$$

which maximizing the variance of the projected data on $\mathbf{v}$. So $\hat{\mathbf{v}}$ is an estimated PC direction.

With one-way penalty imposed, [Huang et al. 2008]

$$\begin{aligned}\hat{\mathbf{v}} &= \arg\max_{\mathbf{v}} \frac{\mathbf{v}^T Y^T Y \mathbf{v}}{\|\mathbf{v}\|^2 + \lambda \mathbf{v}^T \Omega \mathbf{v}} \\ &= \arg\max_{\mathbf{v}} \frac{\mathbf{v}^T Y^T Y \mathbf{v}}{\mathbf{v}^T (I + \lambda\Omega) \mathbf{v}}\end{aligned}$$

Regression View [Zhang et al. 2013]

Let $\mathcal{Y}$ to be a column stack of the data matrix Y, and

$$\mathcal{U} = \begin{bmatrix} \mathbf{u} & 0 & \cdots & 0 \\ 0 & \mathbf{u} & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & 0 & \cdots & \mathbf{u} \end{bmatrix}$$

If the loss function $\rho(\cdot) = \|\cdot\|^2$ and $\mathbf{u}$ is given, the estimate of $\mathbf{v}$ for the one-way problem without penalty is given by a linear regression:

$$\hat{\mathbf{v}} = \left(\mathcal{U}^T\mathcal{U}\right)^{-1}\mathcal{U}^T\mathcal{Y}$$

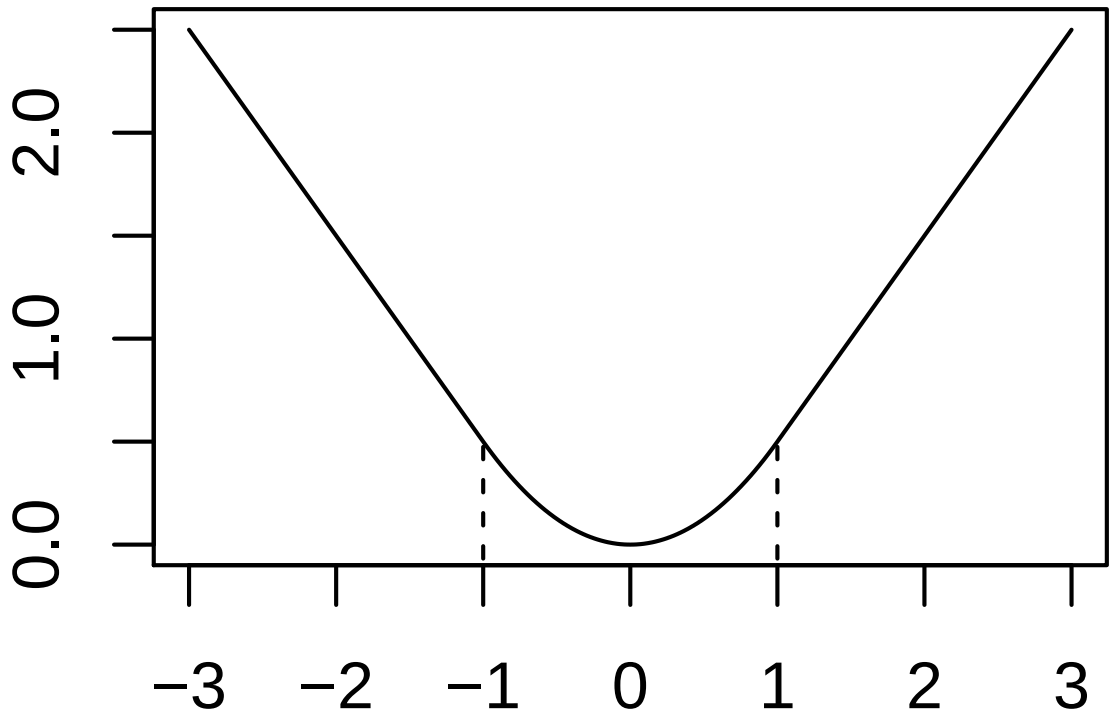
If the one-way penalty is imposed, then we have Ridge-regression type problem. The estimate of $\mathbf{v}$ is

$$\hat{\mathbf{v}} = \left(\mathcal{U}^T\mathcal{U} + \lambda\|\mathbf{u}\|^2\Omega\right)^{-1}\mathcal{U}^T\mathcal{Y}$$

Robust Version [Zhang et al. 2013]

Sometimes there may be outliners in the observed data, then we can use a more robust loss function $\rho(\cdot)$ to replace the usual quadratic loss. For example, Huber loss

$$\rho(x) = \begin{cases} \frac{x^2}{2}, & \text{if } |x| \leq \theta \\ \theta\left(|x| - \frac{\theta}{2}\right), & \text{o.w.} \end{cases}$$



Then the regression is analogous to a weighted least square problem

$$\hat{\mathbf{v}} = \left(\mathcal{U}^T \mathcal{W} \mathcal{U} + \lambda \|\mathbf{u}\|^2 \Omega\right)^{-1} \mathcal{U}^T \mathcal{W} \mathcal{Y},$$

where $\mathcal{W}$ is constructed from the weight matrix

$$W_{ij} = \frac{\rho'(y_{ij} - u_i v_j)}{y_{ij} - u_i v_j}.$$

Picking λ

The smoothing parameter λ can be chosen by leave-one-column Cross-Validation. [Huang et al . 2008] [Zhang et al. 2013]

No robust: Let $A_\lambda = (I + \lambda\Omega)^{-1}$, then the CV and GCV scores are:

$$CV(\lambda) = \frac{1}{m}\sum_{j=1}^{m}\left(\frac{\{(I - A_\lambda)\, Y^T\mathbf{u}\}_j}{1 - \{A_\lambda\}_{jj}}\right)^2$$

$$GCV(\lambda) = \frac{\frac{1}{m}\sum_{j=1}^{m}\|(I - A_\lambda)\, Y^T\mathbf{u}\|^2}{\left(1 - \mathbf{Tr}\left(A_\lambda\right)/m\right)^2}$$

Remark: If CV is defined to be leave-out-one-column, no computational simplification.

Robust: Let $A_\lambda = \mathcal{U}\left(\mathcal{U}^T\mathcal{W}\mathcal{U} + \lambda\|\mathbf{u}\|^2\Omega\right)^{-1}\mathcal{U}^T\mathcal{W}$, then

$$GCV(\lambda) = \frac{\frac{1}{m}\|\hat{\mathbf{v}} - \hat{\mathbf{v}}^*\|^2}{(1 - \mathbf{Tr}\left(A_\lambda\right)/m)^2},$$

where

$$\hat{\mathbf{v}} = \left(\mathcal{U}^T\mathcal{W}\mathcal{U} + \lambda\|\mathbf{u}\|^2\Omega\right)^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}$$
$$\hat{\mathbf{v}}^* = \left(\mathcal{U}^T\mathcal{W}\mathcal{U}\right)^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}$$

Two-Way Problem

If both directions of data are considered as functions, then a two-way version is to minimize

$$\rho(Y - \mathbf{u}\mathbf{v}^T) + \lambda_v \|\mathbf{u}\|^2 \int v''^2 + \lambda_u \|\mathbf{v}\|^2 \int u''^2 + \lambda_v \lambda_u \|u\| \|v\| \int v''^2 \int u''^2,$$

where we added the penalty for the second direction and an interaction is introduced.

$\rho(\cdot)$ can be chosen to be a robust loss.

Estimation [Zhang et al. 2013]

The estimates of $\mathbf{v}$ and $\mathbf{u}$ can be updated iteratively (IRLS) as:

$$\hat{\mathbf{v}} = \left(\mathcal{U}^T \mathcal{W} \mathcal{U} + 2\Omega_{\mathbf{v}|\mathbf{u}}\right)^{-1} \mathcal{U}^T \mathcal{W} \mathcal{Y}$$
$$\Omega_{\mathbf{v}|\mathbf{u}} = \mathbf{u}^T \left(I + \lambda_{\mathbf{u}} \Omega_{\mathbf{u}}\right) \mathbf{u} \left(I + \lambda_{\mathbf{v}} \Omega_{\mathbf{v}}\right) - \mathbf{u}^T \mathbf{u} I$$

$$\hat{\mathbf{u}} = \left(\mathcal{V}^T \mathcal{W}^* \mathcal{V} + 2\Omega_{\mathbf{u}|\mathbf{v}}\right)^{-1} \mathcal{V}^T \mathcal{W}^* \mathcal{Y}^*$$
$$\Omega_{\mathbf{u}|\mathbf{v}} = \mathbf{v}^T \left(I + \lambda_{\mathbf{v}} \Omega_{\mathbf{v}}\right) \mathbf{v} \left(I + \lambda_{\mathbf{u}} \Omega_{\mathbf{u}}\right) - \mathbf{v}^T \mathbf{v} I$$

The hat matrices are:

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

$$GCV(\lambda_{\mathbf{v}}|\lambda_{\mathbf{u}}) = \frac{1}{n}\left(\frac{\|\hat{\mathbf{v}} - \hat{\mathbf{v}}^*\|}{1 - \mathbf{Tr}\left(\mathcal{H}\right)/n}\right)^2$$

$$\hat{\mathbf{v}}^* = \left(\mathcal{U}^T\mathcal{W}\mathcal{U}\right)^{-1}\mathcal{U}^T\mathcal{W}\mathcal{Y}$$

$$GCV(\lambda_{\mathbf{u}}|\lambda_{\mathbf{v}}) = \frac{1}{m}\left(\frac{\|\hat{\mathbf{u}} - \hat{\mathbf{u}}^*\|}{1 - \mathbf{Tr}\left(\mathcal{H}^*\right)/m}\right)^2$$

$$\hat{\mathbf{u}}^* = \left(\mathcal{V}^T\mathcal{W}^*\mathcal{V}\right)^{-1}\mathcal{V}^T\mathcal{W}^*\mathcal{Y}^*$$

Remark: Leave-out-one-column/row CV criteria.

Big Data and Parallelism for One-Way Problem

The number of rows is large while the number of columns is moderate. Equally-spaced common grids are assumed. The smoothing parameter is forced to be the same for all subsets.

1. Split data
2. Estimation for one block
3. Recombine
4. Simulation result

Split Data

$$Y = \begin{array}{|c|c|c|c|} \hline y_{11} & y_{12} & \cdots & y_{1m} \\ y_{21} & y_{22} & \cdots & y_{2m} \\ \vdots & & & \\ y_{n_1 1} & y_{n_1 2} & \cdots & y_{n_1 m} \\ \hline y_{(n_1+1)1} & y_{(n_1+1)2} & \cdots & y_{(n_1+1)m} \\ \vdots & & & \\ y_{(n_1+n_2)1} & y_{(n_1+n_2)2} & \cdots & y_{(n_1+n_2)m} \\ \hline \vdots & & & \\ \hline y_{(n-n_K)1} & y_{(n-n_K)2} & \cdots & y_{(n-n_K)m} \\ \vdots & & & \\ y_{n1} & y_{n2} & \cdots & y_{nm} \\ \hline \end{array}_{n \times m}$$

Remarks:

1. Each subset is a block of the data matrix.
2. The size of the subests are the same except the last one. (The total number of rows may not be divisible by an integer)

When penalty is imposed in one direction (say $\mathbf{v}$), a SVD direction can be estimated for each subset and they can be combined to recover the result when the whole matrix is used.

For k^{th} subset,

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

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

Recombining

$$\hat{\mathbf{v}}^{(c)} = \left[\left(\sum_{k=1}^{K} \mathcal{U}_k^T \mathcal{W}_k \mathcal{U}_k \right) + 2 \sum_{k=1}^{K} \Omega_{\mathbf{v}_k|\mathbf{u}_k} \right]^{-1} \sum_{k=1}^{K} \left[\left(\mathcal{U}_k^T \mathcal{W}_k \mathcal{U}_k + 2\Omega_{\mathbf{v}_k|\mathbf{u}_k} \right) \hat{\mathbf{v}}_k \right]$$

$$\hat{\mathbf{u}}^{(c)} = (\hat{\mathbf{u}}_1, \hat{\mathbf{u}}_2, \cdots, \hat{\mathbf{u}}_K)$$

Algorithm

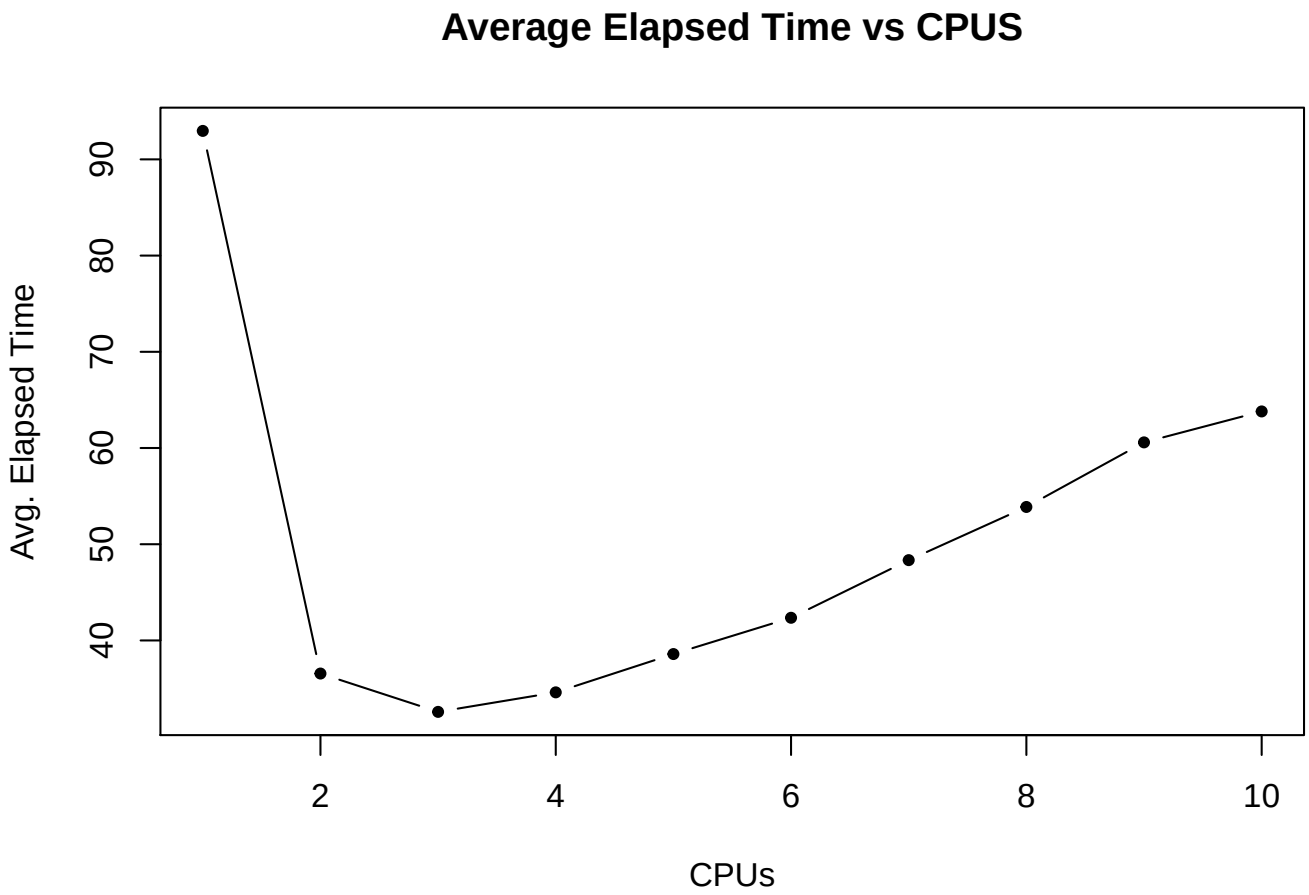
initialize $\mathbf{u} = \hat{\mathbf{u}}^{(0)}$ using SVD;
split Y and $\hat{\mathbf{u}}^{(0)}$;

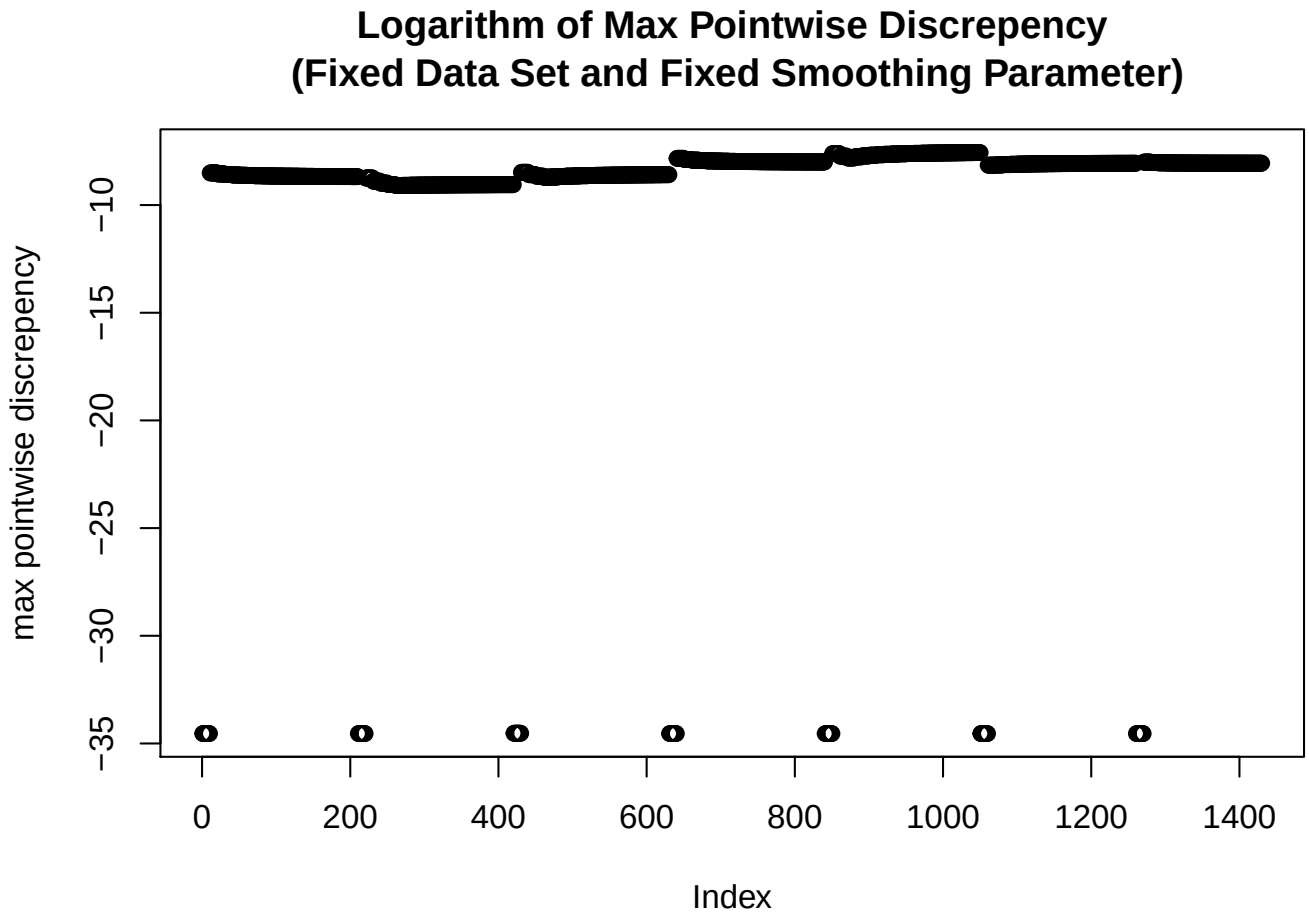
iter $= 0$;
tol $= 99999$;

while *localdiff>tol and iter<maxiter* **do**
    initialize parallelism;
        estimate $\hat{\mathbf{v}}_k^{(i)}$ using $\hat{\mathbf{u}}_k^{(i)}$;
    stop parallelism;
    recombine $\hat{\mathbf{v}}_k^{(i)}$'s to get $\hat{\mathbf{v}}^{(i+1)}$;
    initialize parallelism;
        estimate $\hat{\mathbf{u}}_k^{(i+1)}$;
    stop parallelism;
    recombine $\hat{\mathbf{u}}_k^{(i+1)}$'s to get $\hat{\mathbf{u}}^{(i+1)}$;
    $\hat{Y} = \hat{\mathbf{u}}^{(i+1)}\{\hat{\mathbf{v}}^{(i+1)}\}^T$;
    localdiff $=$ distance$(Y - \hat{Y})$;
end

Simulation

7 replicates, 1000×200 data matrix, number of cores $1 - 10$, smoothing parameters `seq(0, 2, by=0.1)`.




Logarithm of Max Pointwise Discrepency
(Fixed Data Set and Fixed Smoothing Parameter)
max pointwise discrepency
−10
−15
−20
−25
−30
−35
0
200
400
600
800
1000
1200
1400
Index

Summary

1. One-way functional SVD and its interpretation.
2. Estimation and smoothing parameter selection for a one-way problem.
3. Big data problem, i.e. too many curves, and parallelism.
4. Simulation results, efficiency and accuracy.

Future Work

1. Adding GCV for the one-way problem.
2. Parallelizing two-way problems.

References

Xueying Chen and Min-ge Xie.

A split-and-conquer approach for analysis of extraordinarily large data.

pages 1–35, 2012.

Jianhua Z. Huang, Haipeng Shen, and Andreas Buja.

Functional principal components analysis via penalized rank one approximation.

Electronic Journal of Statistics, 2(March):678–695, 2008.

Jianhua Z. Huang, Haipeng Shen, and Andreas Buja.

The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions.

Journal of the American Statistical Association, 104(488):1609–1620, December 2009.

L Zhang, H Shen, and JZ Huang.

Robust regularized singular value decomposition with application to mortality data.

The Annals of Applied Statistics, 2013.