IISE Transactions

Taylor & Francis
Taylor & Francis Group


IISE Transactions

ISSN: 2472-5854 (Print) 2472-5862 (Online) Journal homepage: https://www.tandfonline.com/loi/uiie21

# Hierarchical sparse functional principal component analysis for multistage multivariate profile data

Kai Wang & Fugee Tsung

**To cite this article:** Kai Wang & Fugee Tsung (2021) Hierarchical sparse functional principal component analysis for multistage multivariate profile data, IISE Transactions, 53:1, 58-73, DOI: 10.1080/24725854.2020.1738599

**To link to this article:** https://doi.org/10.1080/24725854.2020.1738599

View supplementary material

Published online: 22 Apr 2020.

Submit your article to this journal

Article views: 294

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=uiie21

IISE TRANSACTIONS
2021, VOL. 53, NO. 1, 58–73
https://doi.org/10.1080/24725854.2020.1738599

Taylor & Francis
Taylor & Francis Group

Check for updates

# Hierarchical sparse functional principal component analysis for multistage multivariate profile data

Kai Wang[a] and Fugee Tsung[b]

[a]School of Management, Xi'an Jiaotong University, Xi'an, China; [b]Department of Industrial Engineering and Decision Analytics, Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong, Kowloon

**ABSTRACT**

Modern manufacturing systems typically involve multiple production stages, the real-time status of which can be tracked continuously using sensor networks that generate a large number of profiles associated with all process variables at all stages. The analysis of the collective behavior of the multistage multivariate profile data is essential for understanding the variance patterns of the entire manufacturing process. For this purpose, two major challenges regarding the high data dimensionality and low model interpretability have to be well addressed. This article proposes integrating Multivariate Functional Principal Component Analysis (MFPCA) with a three-level structured sparsity idea to develop a novel Hierarchical Sparse MFPCA (HSMFPCA), in which the stage-wise, profile-wise and element-wise sparsity are jointly investigated to clearly identify the informative stages and variables in each eigenvector. In this way, the derived principal components would be more interpretable. The proposed HSMFPCA employs the regression-type reformulation of the PCA and the reparameterization of the entries of eigenvectors, and enjoys an efficient optimization algorithm in high-dimensional settings. The extensive simulations and a real example study verify the superiority of the proposed HSMFPCA with respect to the estimation accuracy and interpretation clarity of the derived eigenvectors.

**ARTICLE HISTORY**
Received 3 January 2019
Accepted 16 May 2019

**KEYWORDS**
Hierarchical sparsity; process variation; sensor data; sparse PCA; variance decomposition

## 1. Introduction

Modern manufacturing systems typically involve numerous operating stages before delivering final products. Real examples are common in many industrial sectors including semiconductor, consumer electronics, automobile, aerospace, and so on (Linn *et al.*, 2002; Shu and Tsung, 2003; Xiang and Tsung, 2008; Shang *et al.*, 2014). In such complicated multistage processes, a product sequentially goes through a series of stages, and at each stage, it might be subject to a variety of process variables related to the mechanical, physical or chemical treatment. The collective behavior of these process variables from all stages fully characterizes the total variation of the underlying multistage production system, the analysis of which is crucial for quality evaluation and improvement.

Due to recent progress in information technology, online sensing is being increasingly deployed in current industrial practices, where the real-time process status is continuously tracked by time-ordered data known as *profiles* or *signals* (Paynabar *et al.*, 2013). By appropriately configuring sensor networks in the entire manufacturing process (Liu and Shi, 2013), we can obtain a large number of profiles associated with all process variables at all stages. These multiple profiles are called *multistage multivariate profile data* in this article. A concrete example is given in Figure 1, which considers a Physical Vapor Deposition (PVD) process that produces electronic panels by coating glass material with thin functional films. This process consists of three stages (pre-coating, coating and post-coating), and each stage includes several distinct-form profiles of different process variables (e.g., voltage, current and pressure). The lower half of Figure 1 depicts one individual sample or realization of the multistage multivariate profiles pertinent to one product (only two representative profiles are shown at each stage).

This article aims to study the variation of the entire manufacturing process based on the multistage multivariate profile data. By proposing a novel variance decomposition methodology, we expect to potentially attribute the total variance to a few dominant variance patterns, each of which might only include a few informative stages and process variables, to facilitate a clear interpretation. In the literature on statistical process control, profile data analysis has been extensively studied for both Phase I knowledge discovery and Phase II process monitoring (Noorossana *et al.*, 2011). Starting from simple linear regressions for linear profiles (Kang and Albin, 2000), profile analytic tools have evolved to incorporate wavelet transformation (Chicken *et al.*, 2009), spline approximation (Chang and Yadama, 2010), nonparametric regressions (Zou *et al.*, 2008; Qiu *et al.*, 2010) and Functional Principal Component Analysis (FPCA) (Ramsay,

Kai Wang kwangai01@163.com
Supplemental data for this article is available online at www.tandfonline.com/uiie
Copyright © 2020 "IISE"

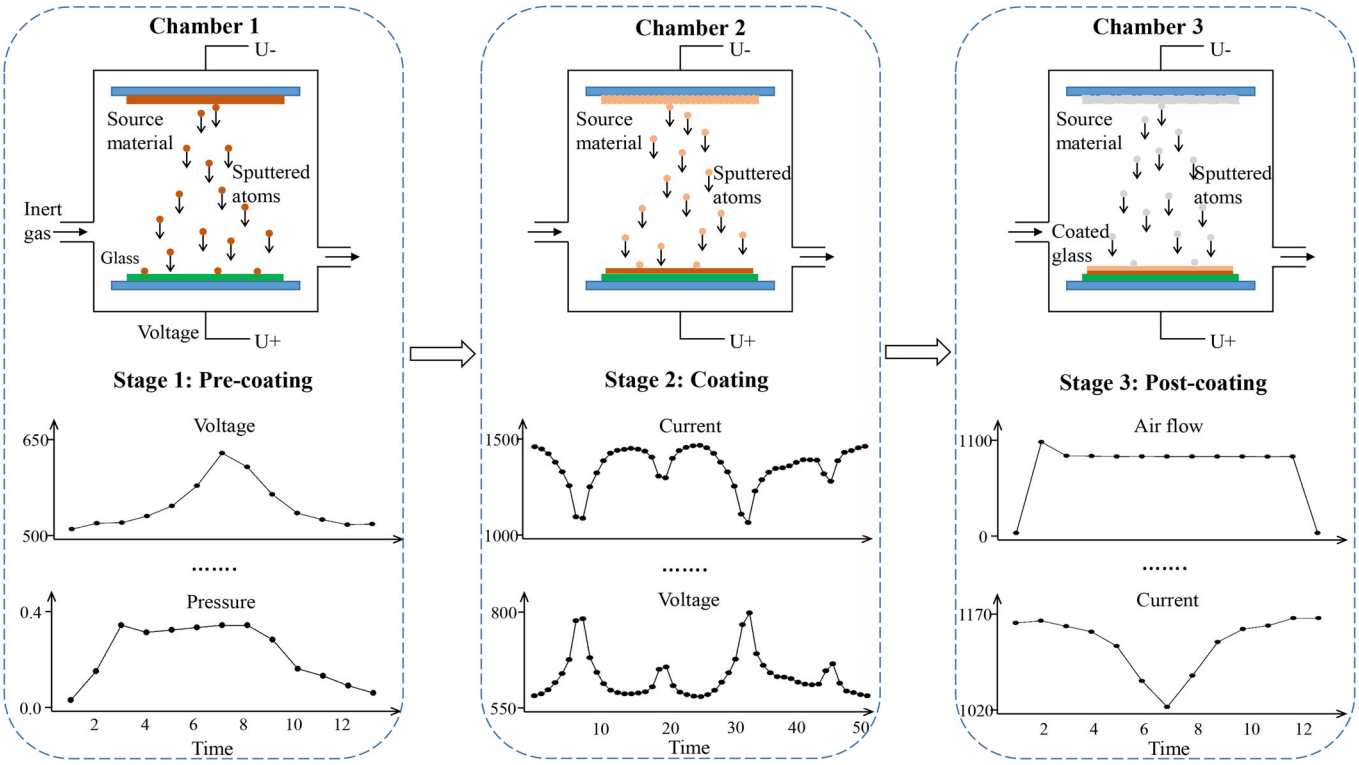


**Figure 1.** Multivariate profile data in the three-stage PVD process.

2005; Ding *et al.*, 2006; Colosimo and Pacella, 2007; Yu *et al.*, 2012) for general nonlinear profiles.

However, the methods above are all confined to a single or univariate profile. For multivariate profiles, a major challenge arises regarding the data dimensionality, since a profile is theoretically infinite-dimensional as a function, or at least high-dimensional in practice when being evaluated discretely at a fine covariate grid as a long vector. The curse of dimensionality is significantly exacerbated when many profiles are considered jointly, making the within-profile and between-profile correlation analysis extremely hard. Therefore, most existing works adopt the Multivariate FPCA (MFPCA) proposed in Ramsay (2005) and several of its variants for dimension reduction. Grasso *et al.* (2014) and Fang *et al.* (2017) used the Vectorized PCA (VPCA) where all profiles are first vectorized and concatenated into one single vector, and then the original PCA is applied to derive a few leading Principal Components (PCs). An alternative is the Multilinear PCA (MPCA) or Uncorrelated MPCA (UMPCA) (Paynabar *et al.*, 2013; Grasso *et al.*, 2014) which operates directly on tensor representations of multivariate profiles rather than on their vectorized versions. Assuming all profiles share similar patterns, Paynabar *et al.* (2016), Wang *et al.* (2018) and Zhang *et al.* (2018a) proposed using a multi-channel FPCA that decomposes each profile with the same set of orthonormal basis functions. Though achieving much lower dimensions, a severe limitation of these MFPCA methods is that they all output dense eigenvectors or loadings with all entries being nonzero, i.e., the PCs are linear combinations of all of the original process variables. This hinders the PCs being interpreted with clear practical meanings, generating no useful industrial knowledge about the variance patterns for practitioners.

To address high-dimensional settings and enhance model interpretability simultaneously, Sparse PCA (SPCA) has been proposed as an intuitively appealing solution where only significant entries are kept in an eigenvector. Instead of manually thresholding small-value entries of eigenvectors which may yield misleading results (Cadima and Jolliffe, 1995), Zou *et al.* (2006), Shen and Huang (2008), and Witten *et al.* (2009) performed PCA through minimizing reconstruction errors and imposed sparsity on eigenvectors by the $L_1$ or LASSO penalty. The SPCA has also been developed in terms of the $L_0$ penalty (d'Aspremont *et al.*, 2008) and a thresholding orthogonal iteration procedure (Ma, 2013). Allen (2013) and Chen and Lei (2015) exploited sparsity in FPCA for a univariate profile, but for multivariate profiles, there are few related works. One exception is a recent work in Zhang *et al.* (2018b) which combined the SPCA in Zou *et al.* (2006) and the multi-channel FPCA in Paynabar *et al.* (2016), and represented each profile with a selected set of orthonormal basis functions. In fact, they imposed sparsity on the PC scores rather than on the eigenvectors. This formulation works well for their profile monitoring purpose, but is not applicable to our variance decomposition problem where we expect sparsity in the eigenvectors such that the significant stages and process variables in each eigenvector can be clearly identified to improve the interpretation of the extracted variance patterns.

In this article, the sparsity idea and the MFPCA introduced above are combined together to explore the variation of the multistage multivariate profile data. The eigenvectors

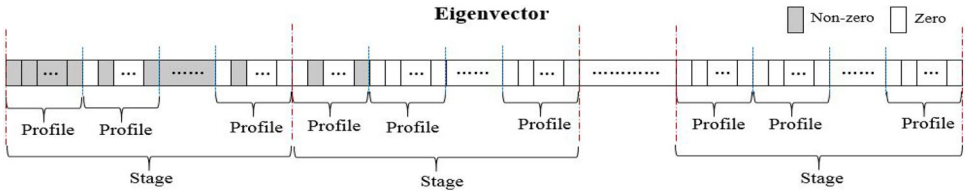


Figure 2. Example of a three-level hierarchical sparsity pattern in an eigenvector.

corresponding to a few leading PCs that explain most of the data variance are encouraged to be sparse with clear interpretation. More creatively, for each PC that is a linear combination of all process variables from all stages, we simultaneously investigate: (i) the stage-wise sparsity, which allows the retention of only a few informative stages; (ii) the profile-wise sparsity, which then allows the selection of only several informative profiles at the retained stages; and (iii) the element-wise sparsity, which finally enables identifying significant elements or local regions in the selected profiles (see Figure 2 for instance). For the multistage multivariate profile data in our context, this three-level hierarchical sparsity will greatly facilitate interpreting the eigenvectors in an insightful manner. Note that our sparsity configuration is related to the group sparse settings commonly seen in variable selection problems (Yuan and Lin, 2006; Zhou and Zhu, 2010; Simon *et al.*, 2013; Khan *et al.*, 2015; Paynabar *et al.*, 2015), where variables are grouped and a two-level sparsity is typically sudied, allowing for a few active groups and a few active variables in each selected group. By considering an additional level of sparsity in the hierarchy, our methodology is expected to produce more parsimonious and accurate results for the multistage multivariate profile data (see Section 3 for evidence).

To sum up, our contributions are as follows:

1. A pioneering Hierarchical Sparse MFPCA (HSMFPCA) for the multistage multivariate profile data is proposed by regression-type reformulating and reparameterizing the original MFPCA.
2. A highly-efficient numerical optimization algorithm equipped with closed-form updating equations at each iteration is developed to overcome the challenge of high dimensionality.
3. Several useful practical guidelines are given with respect to profile data standardization, PC number selection and penalty parameter tuning.
4. Extensive simulations and a real industrial example study are performed to verify the superiority of our proposed HSMFPCA in respect of estimation accuracy and model interpretability.

In practice, by applying our methodology, practitioners are able to gain a better understanding of the variation patterns of the entire manufacturing process, and the quality improvement efforts can thus be better allocated to the key stages and process variables.

The remainder of this article is organized as follows. The proposed HSMFPCA and the designed algorithm are explained in detail in Section 2. Then extensive simulations and a real example of the PVD process are presented in Section 3 and 4, respectively. Section 5 concludes this work. Some technical details are given in the Appendices, and supplementary material is provided online.

## 2. Methodology

### 2.1. Notations and MFPCA

Let $x_{i,sj}(t)$ $(i = 1, ..., N, s = 1, ..., S, j = 1, ..., M_s)$ denote the $j$th profile at the $s$th stage in the $i$th sample, and there are $S$ stages and $M = \sum_{s=1}^{S} M_s$ profiles from all stages. The time index $t$ is just used to indicate $x_{i,sj}(t)$ being a profile, and for ease of exposition, $t \in [0, 1]$ is assumed for all profiles at different stages, which is later shown to have little influence on our derivation.

In addition to being defined as seeking a linear subspace where the variance of projected data is maximized, PCA can also be formulated as sequentially minimizing the reconstruction errors (Bishop, 2006; Shen and Huang, 2008). Let all $M$ profiles in the $i$th sample be denoted by a function vector $\mathbf{x}_i(t) = (x_{i,11}(t), ..., x_{i,1M_1}(t), ..., x_{i,S1}(t), ..., x_{i,SM_S}(t))^T$. Suppose that the eigenfunction vector corresponding to the first PC is $\mathbf{v}(t) = (v_{11}(t), ..., v_{1M_1}(t), ..., v_{S1}(t), ..., v_{SM_S}(t))^T$, which can be obtained as the solution to the following problem:

$$\begin{aligned} &\min_{\mathbf{v}(t)} \sum_{i=1}^{N} \int_0^1 ||\mathbf{x}_i(t) - \mathbf{v}(t)c_i||^2 dt \\ &\text{s.t. } \langle \mathbf{v}(t), \mathbf{v}(t) \rangle = 1, \end{aligned} \tag{1}$$

where $c_i = \langle \mathbf{x}_i(t), \mathbf{v}(t) \rangle = \int_0^1 \mathbf{x}_i(t)^T \mathbf{v}(t) dt = \sum_{s=1}^{S} \sum_{j=1}^{M_s} \int_0^1 x_{i,sj}(t) v_{sj}(t) dt$ is the PC score of the $i$th sample in the framework of the MFPCA (Ramsay, 2005; Fang *et al.*, 2017). Hence, $\mathbf{x}_i(t) - \mathbf{v}(t)c_i$ is the residual part of the function vector $\mathbf{x}_i(t)$ at the time index $t$ after being projected on $\mathbf{v}(t)$, and $||\mathbf{x}_i(t) - \mathbf{v}(t)c_i||^2 = \sum_{s=1}^{S} \sum_{j=1}^{M_s} (x_{i,sj}(t) - v_{sj}(t)c_i)^2$ is the sum of squared residuals over different profiles at $t$. Finally, after integrating over $t$, $\int_0^1 ||\mathbf{x}_i(t) - \mathbf{v}(t)c_i||^2 dt$ is the overall residual values of the $i$th sample. Note that here we adopt the standard formulation of the MFPCA in Ramsay (2005), which is different from the multi-channel FPCA in Paynabar *et al.* (2016) where every profile in a sample is projected on the same eigenfunction and the PC score of the sample is a vector. The multi-channel FPCA in fact assumes different profiles exhibit similar patterns. Since this article aims to find the collective behavior or variance pattern of multiple profiles, the eigenfunction vector $\mathbf{v}(t)$ in Model (1) which includes individual components $v_{sj}(t)$ associated with each profile is a more appropriate candidate. In addition, the studied profiles (see Figure 1 for example) in

this article do not have similar patterns to those expected by the multi-channel FPCA.

In practice, the profile $x_{i,sj}(t)$ is observed at a grid of discrete time indexes, so it can be vectorized as $\mathbf{x}_{i,sj} = (x_{i,sj1}, ..., x_{i,sjT_s})^T$, whose length $T_s$ is assumed to be the same for all profiles in the $s$th stage, but might be different when $s$ varies. Then the $i$th sample data is denoted by a long vector $\mathbf{x}_i = (\mathbf{x}_{i,11}^T, ..., \mathbf{x}_{i,1M_1}^T, ..., \mathbf{x}_{i,S1}^T, ..., \mathbf{x}_{i,SM_S}^T)^T$ whose dimension is $P = \sum_{s=1}^{S}\sum_{j=1}^{M_s} T_s$. By replacing the integration in Model (1) with summation, the MFPCA is simplified to the VPCA in Grasso *et al.* (2014) and Fang *et al.* (2017), where the eigenvector is the solution to the problem:

$$\begin{aligned} \min_{\mathbf{v}} \quad & \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{v}\mathbf{v}^T\mathbf{x}_i||^2 \\ \text{s.t.} \quad & \mathbf{v}^{\mathrm{T}}\mathbf{v} = 1, \end{aligned} \tag{2}$$

and the eigenvector $\mathbf{v}$ can be divided into parts corresponding to different profiles at different stages. The eigenvectors of the remaining PCs can be similarly obtained by solving the problem (2) except that the sample $\mathbf{x}_i$ is repeatedly updated by the residual $\mathbf{x}_i - \mathbf{v}\mathbf{v}^T\mathbf{x}_i$. Hence, the following subsections only focus on the development of the first sparse eigenvector for simplicity.

### 2.2. SPCA

The obvious deficiencies of the VPCA for multivariate profiles are that the dimension $P$ is comparable to or even larger than the data size $N$ and the eigenvector $\mathbf{v}$ with all nonzero entries is hard to interpret. A seminal work that addresses these two problems by proposing an SPCA is Zou *et al.* (2006). They first proved a regression-type reformulation of the PCA, and then obtained the sparse eigenvectors by imposing the $L_1$ penalty on the regression coefficients. In the following, we adopt their method to derive a regression-type version of the VPCA in Model (2).

Specifically, let $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ be our data matrix, and then Model (2) is equivalent to a ridge regression-type optimization problem, that is

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} & ||\mathbf{X} - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T||_F^2 + \lambda||\boldsymbol{\beta}||^2 \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1, \end{aligned} \tag{3}$$

where two decision variables $\boldsymbol{\alpha}, \boldsymbol{\beta}$ are involved, $||\cdot||_F$ is the Frobenius norm, and $\lambda \geq 0$ is the tuning parameter. The first term $||\mathbf{X} - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T||_F^2 = \sum_{i=1}^{N} ||\mathbf{x}_i - \boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{x}_i||^2$ concerns the sum of squared residuals or reconstruction errors. For the second term $\lambda||\boldsymbol{\beta}||^2$, as emphasized in Zou *et al.* (2006), it is not needed if $P < N$ and $\mathbf{X}$ is a full-rank matrix. However, when $P > N$, which is the case in this article, Model (3) with $\lambda = 0$ would give no unique solution as the common regression problems in high-dimensional settings. To make the regression-type reformulation of the PCA in problem (3) determinable, $\lambda > 0$ is required. Therefore, the $L_2$ penalty here is not used to penalize the regression coefficients, but rather to enforce a unique reconstruction of the PCA, which is shown in the following theorem.

**Theorem 1.** *Let $\hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}$ be the solution of Model* (3)*. Then the following statements hold:*

(a) $\hat{\boldsymbol{\beta}} = d_1^2/(d_1^2 + \lambda)\mathbf{v}$, i.e., $\hat{\boldsymbol{\beta}} \propto \mathbf{v}$, where $d_1$ is the first singular value of $\mathbf{X}$.

(b) *As $\lambda \to \infty$, Model* (3) *is reduced to the problem as below:*

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} - \boldsymbol{\beta}||^2 - ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}||^2 \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1. \end{aligned} \tag{4}$$

The proof is given in Appendix A. After solving problem (3) and normalizing $\hat{\boldsymbol{\beta}}$, we can obtain the eigenvector of the original PCA in Model (2) as Theorem 1(a) states. Note that $\lambda$ only affects the norm of $\hat{\boldsymbol{\beta}}$, so in principle we can use an arbitrary positive $\lambda$. Letting $\lambda \to \infty$, we get another special version of Model (3). In Model (4), when $\boldsymbol{\alpha}$ is fixed, $\hat{\boldsymbol{\beta}} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$, which paves an extremely convenient way for imposing sparsity on $\boldsymbol{\beta}$, especially when $P \gg N$. Therefore, as discussed in Zou *et al.* (2006), Model (4) is customized for addressing the high-dimensional challenge, and is thus employed in this article.

Finally, the SPCA is developed by adding a penalty function for $\boldsymbol{\beta}$ in Model (4), that is

$$\begin{aligned} \min_{\boldsymbol{\alpha},\boldsymbol{\beta}} \ & ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} - \boldsymbol{\beta}||^2 - ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}||^2 + h(\boldsymbol{\beta}) \\ \text{s.t.} \quad & \boldsymbol{\alpha}^{\mathrm{T}}\boldsymbol{\alpha} = 1. \end{aligned} \tag{5}$$

In Zou *et al.* (2006), Shen and Huang (2008), and Witten *et al.* (2009), $h(\boldsymbol{\beta}) = \sum_{j=1}^{P} |\beta_j|$ which is the LASSO penalty in Tibshirani (1996) and would induce an element-wise sparsity pattern. In the next subsection, we propose our hierarchical sparse MFPCA by integrating Model (5) and the three-level hierarchical structure in the multistage multivariate profile data.

### 2.3. HSMFPCA

To exploit the three-level (i.e., stage-wise, profile-wise and element-wise) hierarchical sparsity as introduced in Section 1, similar to Zhou and Zhu (2010) and Paynabar *et al.* (2015), we perform a natural reparameterization where each entry of the $P$-dimensional vector $\boldsymbol{\beta}$ is rewritten as

$$\beta_{sjt} = \gamma_s\eta_{sj}\theta_{sjt}, s = 1, ..., S, j = 1, ..., M_s, t = 1, ..., T_s.$$

Here we constrain $\gamma_s \geq 0$ and $\eta_{sj} \geq 0$ for identification reasons, otherwise changing the signs of $\gamma_s$ and $\eta_{sj}$ together would also lead to the same value of $\beta_{sjt}$ when $\theta_{sjt}$ is fixed.

Note that the decomposition above inherently reflects our hierarchical data structure. At the first or stage level of hierarchy, $\gamma_s$ controls all entries $\beta_{sjt}$ $(j = 1, ..., M_s, t = 1, ..., T_s)$ belonging to the $s$th stage as a group, and $\gamma_s = 0$ can induce the stage-wise sparsity. At the second or profile level of hierarchy, $\eta_{sj}$ controls all entries $\beta_{sjt}$ $(t = 1, ..., T_s)$ related to the $j$th profile at the $s$th stage as a group, and $\eta_{sj} = 0$ can further induce the profile-wise sparsity when the $s$th stage is active with $\gamma_s > 0$. Finally, at the third or element level of hierarchy, $\theta_{sjt}$ allows for different values of $\beta_{sjt}$ in the same profile, and $\theta_{sjt} = 0$ can induce the element-wise sparsity when the $s$th stage and the $j$th profile are active with $\gamma_s > 0$ and $\eta_{sj} > 0$.

Then the penalty function is designed as

$$h(\boldsymbol{\beta}) = \lambda_1 \sum_{s=1}^{S} \gamma_s + \lambda_2 \sum_{s=1}^{S} \sum_{j=1}^{M_s} \eta_{sj} + \lambda_3 \sum_{s=1}^{S} \sum_{j=1}^{M_s} \sum_{t=1}^{T_s} |\theta_{sjt}|,$$

where $\lambda_1$, $\lambda_2$ and $\lambda_3$ are positive tuning parameters that control the degree of sparsity at different levels of hierarchy in $\boldsymbol{\beta}$, and their determination is discussed later in Section 2.5. Plugging this form of $h(\boldsymbol{\beta})$ into Model (5), finally we derive our HSMFPCA as below:

$$\begin{aligned} \min_{\boldsymbol{\alpha}, \{\gamma_s, \eta_{sj}, \theta_{sjt}\}_{s,j,t}} \quad & ||\mathbf{y} - \boldsymbol{\beta}||^2 - ||\mathbf{y}||^2 + \lambda_1 \sum_{s=1}^{S} \gamma_s \\ & + \lambda_2 \sum_{s=1}^{S} \sum_{j=1}^{M_s} \eta_{sj} + \lambda_3 \sum_{s=1}^{S} \sum_{j=1}^{M_s} \sum_{t=1}^{T_s} |\theta_{sjt}| \\ \text{s.t.} \quad & \mathbf{y} = \mathbf{X}^{\mathrm{T}}\mathbf{X}\boldsymbol{\alpha}, \\ & \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1, \\ & \beta_{sjt} = \gamma_s \eta_{sj} \theta_{sjt}, s = 1, ..., S, j = 1, ..., M_s, t = 1, ..., T_s, \\ & \gamma_s \geq 0, \eta_{sj} \geq 0, \end{aligned} \tag{6}$$

and the estimated sparse eigenvector is $\hat{\mathbf{v}} = \hat{\boldsymbol{\beta}}/||\hat{\boldsymbol{\beta}}||$.

**Remark 1.** Our HSMFPCA is based on Model (4) to adapt to the high-dimensional data structure, and then it achieves the three-level hierarchical sparsity by reparameterizing variables and imposing penalties. One appealing computational property of Model (6) is that the closed-form updating equations for iteratively estimating the sparse $\boldsymbol{\beta}$ can be easily obtained (see Proposition 1 in Section 2.4) as there is no multiplying matrix in front of $\boldsymbol{\beta}$ in $||\mathbf{y} - \boldsymbol{\beta}||^2$. Therefore, our HSMFPCA enjoys both high model interpretability and high computational efficiency.

**Remark 2.** Our HSMFPCA can be degenerated to two other methods with fewer levels of sparsity. If $\lambda_1 = 0$ and $\gamma_s = 1$, the Profile Sparse MFPCA (PSMFPCA) that encourages a two-level (profile-wise and element-wise) hierarchical sparsity can be obtained. Similarly, if $\lambda_1 = \lambda_2 = 0$ and $\gamma_s = \eta_{sj} = 1$, the one-level Element-wise Sparse MFPCA (ESMFPCA) is derived which is also equivalent to the original SPCA in Zou *et al.* (2006), Shen and Huang (2008), and Witten *et al.* (2009). These two alternatives are compared with our method in Section 3.

### 2.4. Algorithm

For the problem (6), the analytical solutions of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ are intractable, but when either $\boldsymbol{\alpha}$ or $\boldsymbol{\beta}$ is fixed, the complexity is greatly reduced. Therefore, we adopt the Block Coordinate Descent (BCD) algorithm to solve this problem. Specifically, we first fix $\boldsymbol{\alpha}$ to estimate $\boldsymbol{\beta}$, and then fix $\boldsymbol{\beta}$ to estimate $\boldsymbol{\alpha}$. The closed-form updating equations are given in Propositions 1 and 2 below. These two steps are iterated until a certain convergence condition is met.

**Proposition 1.** *In Model* (6), *when* $\boldsymbol{\alpha}$ *is fixed,* $\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$ *is known, and the following hold:*

(a) *Given* $\gamma_s$ *and* $\eta_{sj}$:

$$\hat{\theta}_{sjt} = \mathbb{I}(\gamma_s \eta_{sj} > 0) \cdot \text{sign}(y_{sjt}) \cdot \left( \frac{|y_{sjt}|}{\gamma_s \eta_{sj}} - \frac{\lambda_3}{2(\gamma_s \eta_{sj})^2} \right)_+ . \tag{7}$$

(b) *Given* $\gamma_s$ *and* $\theta_{sjt}$:

$$\hat{\eta}_{sj} = \mathbb{I}(\gamma_s > 0) \cdot \mathbb{I}(\exists t, \theta_{sjt} \neq 0) \cdot \left( \sum_{t=1}^{T_s} \frac{(\gamma_s \theta_{sjt})^2}{\sum_{t=1}^{T_s} (\gamma_s \theta_{sjt})^2} \frac{y_{sjt}}{\gamma_s \theta_{sjt}} - \frac{\lambda_2}{2 \sum_{t=1}^{T_s} (\gamma_s \theta_{sjt})^2} \right)_+ . \tag{8}$$

(c) *Given* $\eta_{sj}$ *and* $\theta_{sjt}$:

$$\hat{\gamma}_s = \mathbb{I}(\exists (j,t), \eta_{sj}\theta_{sjt} \neq 0) \cdot \left( \sum_{j=1}^{M_s} \sum_{t=1}^{T_s} \frac{(\eta_{sj}\theta_{sjt})^2}{\sum_{j=1}^{M_s} \sum_{t=1}^{T_s} (\eta_{sj}\theta_{sjt})^2} \frac{y_{sjt}}{\eta_{sj}\theta_{sjt}} - \frac{\lambda_1}{2 \sum_{j=1}^{M_s} \sum_{t=1}^{T_s} (\eta_{sj}\theta_{sjt})^2} \right)_+ \tag{9}$$

**Proposition 2.** *In Model* (6), *when* $\boldsymbol{\beta}$ *is fixed,* $\hat{\boldsymbol{\alpha}} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/||\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}||$.

The proofs of Propositions 1 and 2 are given in Appendix B. In Proposition 1, when $\boldsymbol{\alpha}$ is fixed, the update of $\boldsymbol{\beta}$ can be done by an inner BCD procedure for $\{\gamma_s\}_s$, $\{\eta_{sj}\}_{s,j}$, and $\{\theta_{sjt}\}_{s,j,t}$ with the closed-form updating equations in Equation (7)-(9) for each kind of parameter. In addition, at the stage level, each $\gamma_s$ value can be updated in a parallel manner as Equation (9) implies, and the same goes for updating $\eta_{sj}$ values and $\theta_{sjt}$ values at the profile and element levels. This block parallel updating property makes the estimation of the sparse $\boldsymbol{\beta}$ scale to high dimensions easily. When $\boldsymbol{\beta}$ is fixed, Proposition 2 derives an extremely easy updating equation for $\boldsymbol{\alpha}$. To sum up, our HSMFPCA can be efficiently applied to the multistage multivariate profiles that have an inherent high-dimensional data structure.

More insights about our three-level hierarchical sparsity can be gained from Proposition 1. First, at the element level, $\hat{\theta}_{sjt}$ should be zero if $\gamma_s = 0$ or $\eta_{sj} = 0$; otherwise, its initial estimate $y_{sjt}/(\gamma_s \eta_{sj})$ is diminished by a soft thresholding. Note that the shrinkage magnitude in Equation (7) is inversely proportional to $\gamma_s^2$ and $\eta_{sj}^2$, which is intuitive, as more active stages or profiles would induce fewer penalties on their elements. Second, at the profile level in Equation (8), $\hat{\eta}_{sj} = 0$ if $\gamma_s = 0$ and $\theta_{sjt} = 0$ for all $t = 1, ..., T_s$. When some elements in this profile are active, we would have several estimates of $\eta_{sj}$ as $y_{sjt}/(\gamma_s \theta_{sjt})$, the weighted shrunken average of which is taken as the final estimate of $\eta_{sj}$. The weight allocated to each element is proportional to the magnitude of this element $\theta_{sjt}^2$. Finally, at the stage level, $\hat{\gamma}_s$ in Equation (9) is also a weighted shrunken average of all $y_{sjt}/(\eta_{sj}\theta_{sjt})$ associated with the active elements belonging to this stage.

The BCD algorithm for the HSMFPCA is detailed in Algorithm 1. In Zou *et al.* (2006), $\boldsymbol{\alpha}$ is initialized as the first eigenvector of $\mathbf{X}^T\mathbf{X}$ by using Singular-Value Decomposition (SVD), whose time complexity is as high as $O(P^3)$. Alternatively, we calculate $\mathbf{X}\mathbf{X}^T$ $(O(N^2P))$ and use the power

method ($O(N^2)$) to only obtain the first eigenvector of $\mathbf{X}\mathbf{X}^T$ denoted by $\mathbf{e}$. Then the first eigenvector of $\mathbf{X}^T\mathbf{X}$ is $\mathbf{X}^T\mathbf{e}/||\mathbf{X}^T\mathbf{e}||$ ($O(NP)$) (Bishop, 2006). The matrix $\mathbf{X}^T\mathbf{X}$ is also calculated ($O(NP^2)$) and stored in advance, as it will be used repeatedly in updating $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in Steps 2 and 3. Step 2 updates $\boldsymbol{\beta}$, the time complexity being $O(LP)$ where $L$ is the number of repeat times in updating $\{\gamma_s, \eta_{sj}, \theta_{sjt}\}_{s,j,t}$. Step 3 updates $\boldsymbol{\alpha}$ whose time complexity involved in calculating $\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$ is $O(P^2)$. The overall time complexity of Algorithm 1 is thus $O(N^2P + N^2 + NP + NP^2 + MLP + MP^2)$ where $M$ is the iteration times of Steps 2 and 3. In our context, $N \ll P$, and small values of $M$ and $L$ (e.g., 10 or 50) can already lead to an empirical convergence. Therefore, our algorithm whose time complexity is at the order of $O(P^2)$ does not induce prohibitively high computational costs.

Finally, we discuss the convergence property of Algorithm 1 in Proposition 3.

**Proposition 3.** *Algorithm* 1 *converges to a stationary point of Model* (6).

**Proof.** Denote the objective function in Model (6) before the $m$th iteration in Algorithm 1 by $f_{m-1} = f(\boldsymbol{\alpha}^{(m-1)}, \{\hat{\gamma}_s^{(m-1)}, \hat{\eta}_{sj}^{(m-1)}, \hat{\theta}_{sjt}^{(m-1)}\}_{s,j,t})$. After the inner BCD procedure in Step 2, $f(\boldsymbol{\alpha}^{(m-1)}, \{\hat{\gamma}_s^{(m)}, \hat{\eta}_{sj}^{(m)}, \hat{\theta}_{sjt}^{(m)}\}_{s,j,t}) \le f_{m-1}$. Then after Step 3, $f_m = f(\boldsymbol{\alpha}^{(m)}, \{\hat{\gamma}_s^{(m)}, \hat{\eta}_{sj}^{(m)}, \hat{\theta}_{sjt}^{(m)}\}_{s,j,t}) \le f(\boldsymbol{\alpha}^{(m-1)}, \{\hat{\gamma}_s^{(m)}, \hat{\eta}_{sj}^{(m)}, \hat{\theta}_{sjt}^{(m)}\}_{s,j,t})$. Thus, $f_m \le f_{m-1}$ and the iteration of Steps 2 and 3 decreases the objective function monotonically. Note that $||\mathbf{y}||^2 = \boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$ in Model (6) is upper bounded by the first eigenvalue $g_1$ of $\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}$, and thus our objective function $f$ is evidently lower bounded by $-g_1$. In addition, as shown in Propositions 1 and 2, $f$ has a unique minimum in $\boldsymbol{\alpha}, \{\gamma_s\}_s, \{\eta_{sj}\}_{s,j}$, and $\{\theta_{sjt}\}_{s,j,t}$ when any three kinds of these parameters are fixed. Therefore, according to Theorem 4.1 of Tseng (2001), the algorithm is guaranteed to converge to a stationary point of Model (6).

**Algorithm 1 HSMFPCA**

**Input**: Data matrix $\mathbf{X}$, tuning parameters $\lambda_1$, $\lambda_2$ and $\lambda_3$
**Output**: Sparse eigenvector $\hat{\mathbf{v}} = \hat{\boldsymbol{\beta}}/||\hat{\boldsymbol{\beta}}||$.

1: Initialization. Use the power method to obtain the first eigenvector $\mathbf{e}$ of $\mathbf{X}\mathbf{X}^T$, and let $\hat{\boldsymbol{\alpha}}^{(0)} = \mathbf{X}^T\mathbf{e}/||\mathbf{X}^T\mathbf{e}||$. Then $\mathbf{y}^{(0)} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}^{(0)}, \hat{\gamma}_s^{(0)} = 1, \hat{\eta}_{sj}^{(0)} = 1, \hat{\theta}_{sjt}^{(0)} = y_{sjt}^{(0)}$ and $\hat{\beta}_{sjt}^{(0)} = \hat{\gamma}_s^{(0)}\hat{\eta}_{sj}^{(0)}\hat{\theta}_{sjt}^{(0)}, s = 1, ..., S, j = 1, ..., M_s, t = 1, ..., T_s$.
2: Update $\hat{\boldsymbol{\beta}}$. Given $\hat{\boldsymbol{\alpha}}^{(m-1)}, \mathbf{y}^{(m-1)} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}^{(m-1)}$ is known, and then an inner BCD algorithm is used to obtain $\{\hat{\gamma}_s^{(m)}, \hat{\eta}_{sj}^{(m)}, \hat{\theta}_{sjt}^{(m)}\}_{s,j,t}$ and $\hat{\boldsymbol{\beta}}^{(m)}$.
2.1: Initially, $\hat{\gamma}_s = \hat{\gamma}_s^{(m-1)}, \hat{\eta}_{sj} = \hat{\eta}_{sj}^{(m-1)}$ and $\hat{\theta}_{sjt} = \hat{\theta}_{sjt}^{(m-1)}$.
2.2: Update $\hat{\theta}_{sjt}, \hat{\eta}_{sj}$ and $\hat{\gamma}_s$ sequentially by Proposition 1.
2.3: Repeat Step 2.2 until convergence, and finally $\hat{\gamma}_s^{(m)} = \hat{\gamma}_s, \hat{\eta}_{sj}^{(m)} = \hat{\eta}_{sj}$ and $\hat{\theta}_{sjt}^{(m)} = \hat{\theta}_{sjt}$ and $\hat{\beta}_{sjt}^{(m)} = \hat{\gamma}_s\hat{\eta}_{sj}\hat{\theta}_{sjt}$.
3: Update $\hat{\boldsymbol{\alpha}}$. Given $\hat{\boldsymbol{\beta}}^{(m)}$, update $\hat{\boldsymbol{\alpha}}$ by Proposition 2.
4: Repeat Steps 2 and 3 by letting $m = m + 1$ until $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$ converge.

### 2.5. Implementation guidelines

This subsection provides several key guidelines for the implementation of our proposed HSMFPCA.

#### Data centering and standardization

For exposition convenience, all profile data above are assumed to have been centered and standardized. Data centering is trivial by letting $\mathbf{x}_{i,sj} = \mathbf{x}_{i,sj} - \boldsymbol{\mu}_{sj}$ where $\boldsymbol{\mu}_{sj} = \sum_{i=1}^N \mathbf{x}_{i,sj}/N$. However, to standardize profile data, much more care has to be taken, as the data points have different measurement units and scales in different profiles but share the same unit and scale in the same profile. Hence, a profile-wise standardization is preferred, i.e., $\mathbf{x}_{i,sj} = (\mathbf{x}_{i,sj} - \boldsymbol{\mu}_{sj})/\sigma_{sj}$, where $\sigma_{sj} = \sum_{i=1}^N (\mathbf{x}_{i,sj} - \boldsymbol{\mu}_{sj})^T(\mathbf{x}_{i,sj} - \boldsymbol{\mu}_{sj})/(NT_s)$, such that all data points in $\mathbf{x}_{i,sj}$ are scaled at the same order of magnitude.

#### Determining the number of PCs

The common subjective approach to choosing the number of PCs $K$ is to seek an elbow point on the scree plot or to threshold the cumulative percentage of the explained variance. However, as shown in Akemann *et al.* (2011), when $P > N$, these qualitative approaches are ineffective in separating signals from noise. A quantitative way is based on the Tracy–Widom convergence, which derives the asymptotic distribution of the largest eigenvalue when $P/N \to \rho \in (0, \infty)$ if the $\mathbf{x}_i$ are normal distributed with an identity covariance matrix (El Karoui, 2003). Therefore, as our HSMFPCA is performed sequentially, after updating $\mathbf{X}$ by $\mathbf{X} - \mathbf{X}\hat{\mathbf{v}}\hat{\mathbf{v}}^T$, we first test a hypothesis to check if the covariance matrix is an identity matrix based on the Tracy–Widom convergence with an appropriate significant level (e.g., 0.05). If rejected, we continue solving the problem (6), otherwise we stop looking for the next PC. See the supplement materials file for more information.

#### Selecting the tuning parameters

The tuning parameters can be determined by Cross Validation (CV) approaches, e.g., the 5-fold or 10-fold CV, at very high computational expense. Therefore, we use a common model selection criteria such as the Akaike Information Criterion (AIC) in this work. Furthermore, instead of seeking the optimal parameters in a three-dimension grid of $\lambda_1$, $\lambda_2$, and $\lambda_3$, similar to the non-negative garrote technique in Yuan and Lin (2006), we let $\lambda_2 \sum_{s=1}^S \sum_{j=1}^{M_s} \eta_{sj} = \lambda_3 \sum_{s=1}^S \sum_{j=1}^{M_s} T_s\eta_{sj}$ and $\lambda_1 \sum_{s=1}^S \gamma_s = \lambda_3 \sum_{s=1}^S M_sT_s\gamma_s$ in Model (6), i.e., the penalties imposed on the profiles and stages are proportional to the number of elements belonging to them. Now we only have to determine the optimal value of $\lambda_3$. The AIC of our proposed HSMFPCA is defined as

$$\mathrm{AIC}(\lambda_3) = ||\mathbf{X} - \mathbf{X}\mathbf{v}_{\lambda_3}\mathbf{v}_{\lambda_3}^T||_F^2/\sigma_\varepsilon^2 + 2 \cdot df_{\lambda_3},$$

where $\sigma_\varepsilon^2 = \mathrm{Median}(\{\sum_{i=1}^N (\mathbf{x}_{i,sjt} - \boldsymbol{\mu}_{sjt})^2/N\}_{s,j,t})$ (Ma, 2013), and $df$ counts the number of nonzero elements in $\boldsymbol{\beta}$. Generally speaking, as $\lambda_3$ increases, the residual $||\mathbf{X} -$

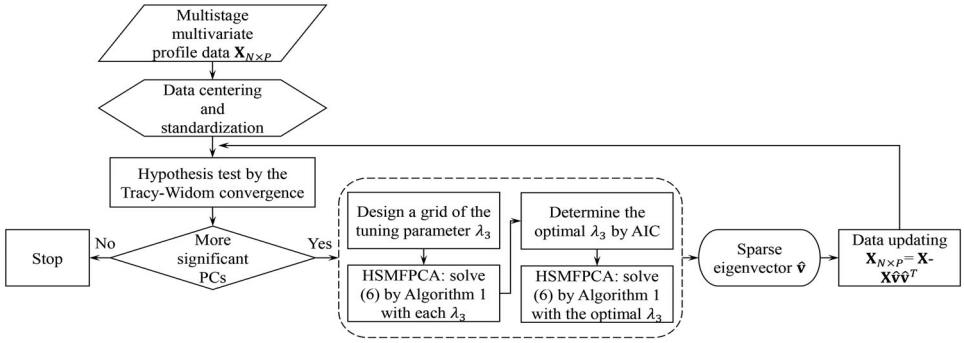


**Figure 3.** Flowchart of the implementation of our proposed HSMFPCA.

$\mathbf{X}\mathbf{v}_{\lambda_3}\mathbf{v}_{\lambda_3}^T||_F^2/\sigma_\varepsilon^2$ increases, whereas the model complexity $df_{\lambda_3}$ decreases, and thus the optimal value of $\lambda_3$ can be found. This simplified AIC($\lambda_3$) has been compared with the exact AIC($\lambda_1, \lambda_2, \lambda_3$) in our supplementary materials to verify its sufficient model selection capability.

Finally, to conclude Section 2, the implementation procedures of our proposed HSMFPCA for the multistage multivariate profile data are summarized in Figure 3. The derived sparse eigenvectors enable us to identify the active stages, profiles and elements in the variance patterns of the entire manufacturing process.

## 3. Simulations

In this section, we investigate the numerical performance of the proposed HSMFPCA through extensive simulations. Specifically, we first show the solution path of the HSMFPCA. Then, we compare the HSMFPCA with several competing methods in the estimation of the eigenvector corresponding to the first PC. Finally, the comparison is extended to the case of multiple PCs, where multiple leading eigenvectors are estimated.

### 3.1. Solution path of the HSMFPCA

This subsection presents the behavior of our estimated eigenvector $\hat{\mathbf{v}}$ from Model (6) as the tuning parameter $\lambda_3$ varies from zero to a large positive number, i.e., the entire solution path of the proposed HSMFPCA. Our multistage multivariate profile data are generated by the following factor model (Johnstone and Lu, 2009; Ma, 2013):

$$\mathbf{x}_i = \boldsymbol{\mu} + \sum_{l=1}^{L} u_{il}\mathbf{v}_l + \varepsilon_i, i = 1, ..., N, \tag{10}$$

where $\boldsymbol{\mu} = 0$ is the mean vector, $\mathbf{v}_l$ is the $l$th true eigenvector, $u_{il} \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma_l^2)$ is the realization of the $l$th random factor in the $i$th sample, $\sigma_l^2$ represents the $l$th signal size, and $\varepsilon_i \overset{\text{i.i.d}}{\sim} \mathrm{N}(0, \sigma_\varepsilon^2\mathbf{I})$ is the white noise independent of the $u_{il}$.

We first consider the single factor model with $L = 1$ in Equation (10). The true eigenvector is shown in Figure 4 where the elements belonging to different profiles and stages are divided by dashed lines. There are $S = 4$ stages, $M_s = 5$ profiles at each stage, and $T_s = 10$ elements in each profile, so the dimension of $\mathbf{x}_i$ is $P = 200$. We simulate $N = 50$ samples with $\sigma_1^2 = 10^2$ and $\sigma_\varepsilon^2 = 1^2$, and then apply our HSMFPCA to this data set. Four representative estimated eigenvectors corresponding to different $\lambda_3$ values are also plotted in Figure 4. When $\lambda_3 = 0$, our HSMFPCA is reduced to the conventional VPCA, and all of the elements in $\hat{\mathbf{v}}$ are nonzero, making the interpretation unclear. As $\lambda_3$ increases, e.g., $\lambda_3 = 3.5, 6.8, 10.0$, the insignificant elements, profiles and stages are shown to progressively shrink to zero. The best $\lambda_3$ with the smallest AIC is selected as being 6.8 in Figure 5, and the resulted $\hat{\mathbf{v}}$ is very similar to the true eigenvector with correctly selected stages and profiles. Additionally, the AIC is only increased slightly when $6.8 < \lambda_3 < 12$ in Figure 5, indicating that in practice $\lambda_3$ can be selected robustly in a wide range.

### 3.2. Estimation performance for a single PC

In this subsection, we quantify the HSMFPCA performance in eigenvector estimation. We still use the single factor model in Section 3.1 with $L = 1$, but in addition to the true eigenvector in Figure 4, we consider another five cases in Figure 6. We let $N = 50$, $\sigma_1^2 = 5^2, \sigma_\varepsilon^2 = 1^2$ and generate 100 random date sets. Table 1 summarizes the competing methods and the performance criteria. In particular, after concatenating all profile $\mathbf{x}_{i,sj}$ values in the $i$th sample into a long vector $\mathbf{x}_i$, the VPCA applies the original PCA to the data matrix $\mathbf{X} = (\mathbf{x}_1, ..., \mathbf{x}_N)^T$ (Grasso *et al.*, 2014; Fang *et al.*, 2017). The MPCA and UMPCA, however, formulate the samples of all profiles as a three-dimensional array or tensor and use the multilinear algebra to derive engenvectors. Unlike the MPCA, the UMPCA considers an additional zero-correlation constraint on PCs. We refer readers to Paynabar *et al.* (2013) and Grasso *et al.* (2014) for more technical details. The SSPCA simply employs a sequential manner to induce the three-level hierarchical sparsity in eigenvectors, i.e., the original PCA is first used to obtain an eigenvector $\hat{\mathbf{v}}$, then the group LASSO (Yuan and Lin, 2006; Friedman *et al.*, 2010) is adopted to select significant stages and profiles in $\hat{\mathbf{v}}$, and finally the LASSO (Tibshirani, 1996) is taken to screen import elements in the selected stages and profiles. The ESMFPCA and PSMFPCA are two simplified versions of our HSMFPCA as discussed in Remark 2 in Section 2.3 and hence can be implemented by using our Algorithm 1.

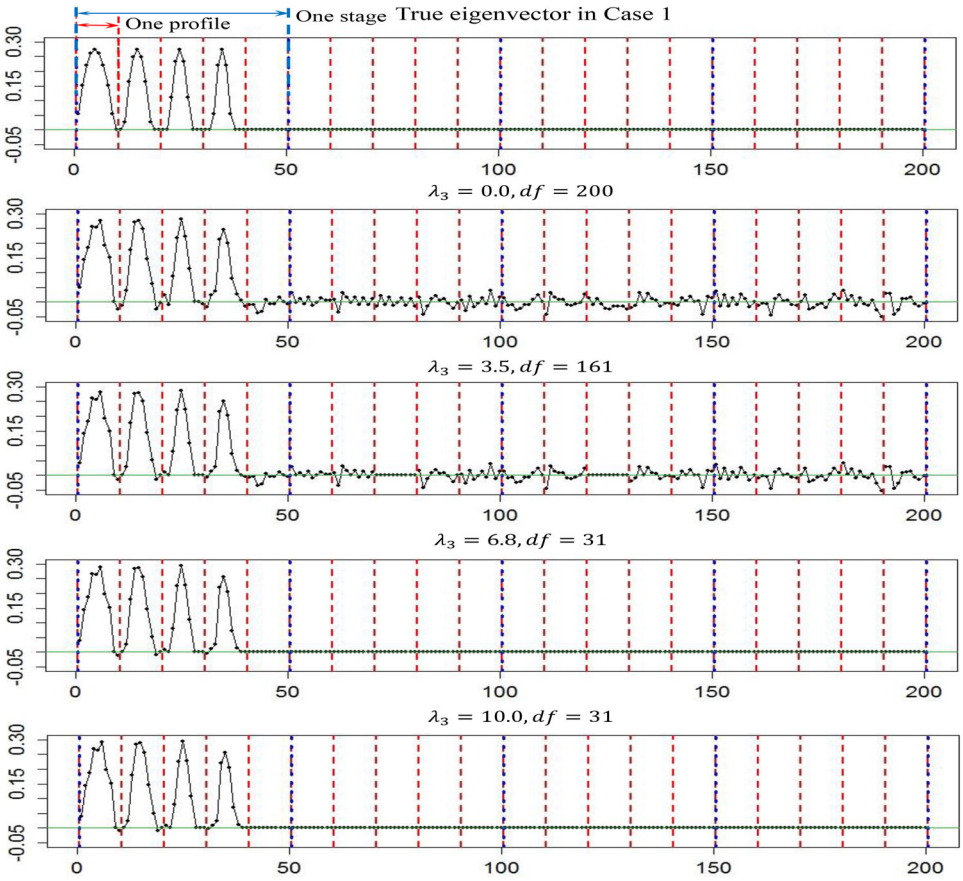


**Figure 4.** True eigenvector and the solution path of the HSMFPCA.

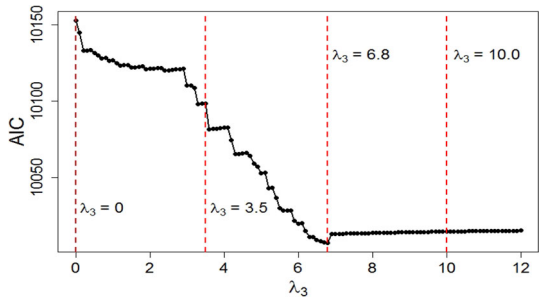


**Figure 5.** Selection of the tuning parameter $\lambda_3$ by the AIC.

The comparison results are given in Table 2 and illustrated in Figure 7. For each criterion, we also perform the paired Student's $t$-test to check the difference between the results of the best method and the counterparts in the 100 simulated data sets, and the best and comparable results are all shown in bold face in Table 2. It can be clearly seen that when the true eigenvector has the three-level hierarchical sparsity with insignificant elements, profiles and stages (cases 1 to 4), our HSMFPCA has significant superiority with higher correct rates (i.e., ZM, F1) in identifying the sparsity pattern of an eigenvector as well as smaller deviation errors (i.e., Angle, RMSE) in inferring the element values of an eigenvector. In addition, our HSMFPCA behaves very similarly to the PSMFPCA and ESMFPCA in cases 5 and 6, which only have two or one level(s) of sparsity. Compared with the VPCA which always has the largest EVs, but dense engenvectors, our HSMFPCA is found to slightly sacrifice the percent of explained variance for much more interpretable sparse results. There is little difference between the results of the MPCA and UMPCA, and neither of them perform well in our cases, as they both concern a no sparsity structure in the eigenvectors. The SSPCA seems better than the ESMFPCA since it further exploits the profile-wise and stage-wise sparsity, but it is still worse than our proposed HSMFPCA. Note that our HSMFPCA involves a minimization of the penalized reconstruction errors of PCA through iterations, whereas the SSPCA is actually a simple one-time soft-thresholding on the dense eigenvectors obtained from the original PCA. Overall, our HSMFPCA has been shown to have a better performance in estimating sparse eigenvectors, and is extremely suitable for application to multistage multivariate profile data when a three-level hierarchical sparsity exists in the variance patterns.

### 3.3. Estimation performance for multiple PCs

Next we explore the performance of our proposed HSMFPCA in estimating multiple eigenvectors by following the procedures in Figure 3. The data are simulated by the factor model in Equation (10) with $L = 3$. We set $N = 50$, $\sigma_\varepsilon^2 = 1^2$ and generate 50 random date sets. The true eigenvectors with their signal sizes are shown in Figures 8 and 9. In case 7, $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$ are all sparse, and their nonzero elements are distributed in different profiles and stages. In case 8, however, the nonzero regions in $\mathbf{v}_1, \mathbf{v}_2$, and $\mathbf{v}_3$ are

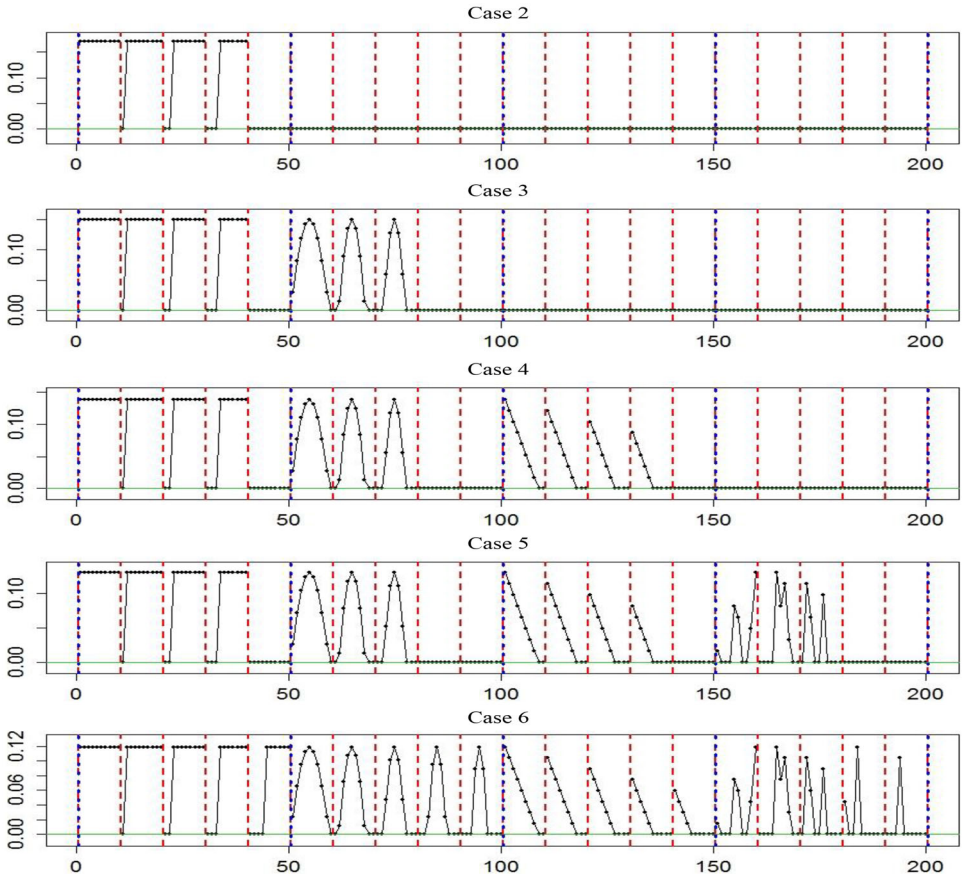


**Figure 6.** True eigenvectors in cases 2-6.

**Table 1.** Notations and definitions of the competing methods and performance criteria.

| | |
|---|---|
| Method | V: Vectorized PCA (VPCA in Grasso *et al.* (2014), Fang *et al.* (2017)).<br>M: Multilinear PCA (MPCA in Grasso *et al.* (2014)).<br>UM: Uncorrelated multilinear PCA (UMPCA in Paynabar *et al.* (2013)).<br>SS: Sequential sparse PCA (SSPCA).<br>ES: Element-wise sparse MFPCA (SPCA in Zou *et al.* (2006)).<br>PS: Profile-wise and element-wise sparse MFPCA (PSMFPCA).<br>HS: our proposed HSMFPCA. |
| Criterion | ZM: Zero-measure, ZM $= \sum_{s,j,t}(\mathbb{I}(\hat{v}_{sjt} \neq 0 \& v_{sjt} \neq 0) + \mathbb{I}(\hat{v}_{sjt} = 0 \& v_{sjt} = 0))/P$.<br>F1: F1 score, the harmonic mean of precision and recall, F1 $= 2(\text{precision}^{-1} + \text{recall}^{-1})^{-1}$, and precision $= \sum_{s,j,t}(\mathbb{I}(\hat{v}_{sjt} \neq 0 \& v_{sjt} \neq 0))/\sum_{s,j,t}\mathbb{I}(\hat{v}_{sjt} \neq 0)$,<br>recall $= \sum_{s,j,t}(\mathbb{I}(\hat{v}_{sjt} \neq 0 \& v_{sjt} \neq 0)/\sum_{s,j,t}\mathbb{I}(v_{sjt} \neq 0)$.<br>Angle: The angle between the estimated and true eigenvectors, Angle $= 2(\arccos(\lvert\hat{\mathbf{v}}^T\mathbf{v}\rvert)/\pi)$.<br>RMSE: The root mean squared error of the estimated eigenvector, RMSE $= \sqrt{\lvert\lvert\hat{\mathbf{v}} - \mathbf{v}\rvert\rvert^2/P}$.<br>EV: The ratio of the explained variance over the total variance, EV $= \text{var}(\mathbf{x}^T\hat{\mathbf{v}})/\sum_{s,j,t}\text{var}(x_{sjt})$. |

overlapping, which indicates that a profile or stage can be significant in more than one eigenvector.

Our HSMFPCA is also compared with its six counterparts in Table 1. The results regarding the estimated three leading eigenvectors $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \hat{\mathbf{v}}_3$ and the overall subspace spanned by the three estimated eigenvectors are shown in Tables 3 and 4 and Figure 10. Note that here the overall angle refers to the subspace angle, and the EVs of $\hat{\mathbf{v}}_1, \hat{\mathbf{v}}_2, \hat{\mathbf{v}}_3$ are adjusted slightly, due to the existing small correlations among the estimated sparse PCs (see Zou *et al.* (2006)). It can be seen that except in a few cells in Tables 3 and 4, our HSMFPCA is always the best one, which verifies its advantages in estimating the sparsity patterns and the element values of both the individual eigenvectors and the spanned subspace in cases 7 and 8. To sum up, it has been validated that our proposed HSMFPCA is also able to accurately discover multiple informative variance patterns with sparse structures from the multistage multivariate profile data.

As a final note, we also investigate the behavior of the proposed HSMFPCA for varying the sample size $N$ and signal size $\sigma_l^2$. The simulation results are provided in our supplementary materials file. Within our expectation, as $N$ and $\sigma_l^2$ increase, the sparsity patterns and the element values of the eigenvectors are more accurately inferred, and the variances of the performance criteria are also reduced when $N$ gets larger.

## 4. Real example

In this section, we revisit the three-stage PVD process for producing electronic panels introduced in Section 1, and

**Table 2.** Comparison results in cases 1 to 6.

| | *V* | *M* | *UM* | *SS* | *ES* | *PS* | *HS* | *V* | *M* | *UM* | *SS* | *ES* | *PS* | *HS* |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *Case 1* | | | | | | | *Case 2* | | | | | | |
| ZM | 0.1050 | 0.1050 | 0.1050 | 0.8930 | 0.8778 | 0.9222 | **0.9309** | 0.1700 | 0.1700 | 0.1700 | 0.9088 | 0.9068 | 0.9376 | **0.9467** |
| F1 | 0.1900 | 0.1900 | 0.1900 | 0.6520 | 0.6089 | 0.7247 | **0.7490** | 0.2906 | 0.2906 | 0.2906 | 0.7913 | 0.7869 | 0.8466 | **0.8659** |
| Angle | 0.2472 | 0.1595 | 0.1606 | 0.1275 | 0.1255 | 0.1222 | **0.1180** | 0.2462 | 0.1967 | 0.1974 | 0.1515 | 0.1580 | 0.1383 | **0.1348** |
| RMSE | 0.0272 | 0.0176 | 0.0177 | 0.0141 | 0.0139 | 0.0135 | **0.0130** | 0.0272 | 0.0218 | 0.0218 | 0.0167 | 0.0175 | 0.0153 | **0.0148** |
| EV | **0.1336** | 0.1132 | 0.1132 | 0.1215 | 0.1208 | 0.1203 | 0.1198 | **0.1333** | 0.1100 | 0.1100 | 0.1226 | 0.1221 | 0.1208 | 0.1206 |
| | *Case 3* | | | | | | | *Case 4* | | | | | | |
| ZM | 0.2750 | 0.2750 | 0.2750 | 0.8917 | 0.8662 | 0.9153 | **0.9190** | 0.4100 | 0.4100 | 0.4100 | 0.8614 | 0.8239 | 0.8659 | **0.8706** |
| F1 | 0.4314 | 0.4314 | 0.4314 | 0.8287 | 0.7932 | 0.8645 | **0.8686** | 0.5816 | 0.5816 | 0.5816 | 0.8385 | 0.8005 | 0.8512 | **0.8526** |
| Angle | 0.2459 | 0.2476 | 0.2482 | 0.1841 | 0.1959 | 0.1659 | **0.1631** | 0.2454 | 0.3062 | 0.3067 | 0.2184 | 0.2281 | 0.1968 | **0.1926** |
| RMSE | 0.0270 | 0.0273 | 0.0274 | 0.0204 | 0.0217 | 0.0183 | **0.0181** | 0.0268 | 0.0334 | 0.0335 | 0.0241 | 0.0252 | 0.0216 | **0.0211** |
| EV | **0.1336** | 0.1044 | 0.1045 | 0.1240 | 0.1238 | 0.1232 | 0.1232 | **0.1335** | 0.0967 | 0.0968 | 0.1249 | 0.1250 | 0.1259 | 0.1258 |
| | *Case 5* | | | | | | | *Case 6* | | | | | | |
| ZM | 0.4700 | 0.4700 | 0.4700 | **0.8143** | 0.8126 | **0.8150** | **0.8143** | 0.5750 | 0.5750 | 0.5750 | 0.7990 | **0.8021** | 0.7963 | 0.7965 |
| F1 | 0.6395 | 0.6395 | 0.6395 | **0.8259** | 0.8115 | **0.8252** | **0.8251** | 0.7302 | 0.7302 | 0.7302 | 0.8290 | **0.8309** | **0.8300** | **0.8303** |
| Angle | 0.2462 | 0.3368 | 0.3372 | 0.2349 | 0.2426 | **0.2129** | **0.2130** | 0.2663 | 0.3669 | 0.3673 | 0.2586 | **0.2438** | **0.2437** | **0.2443** |
| RMSE | 0.0268 | 0.0367 | 0.0368 | 0.0259 | 0.0267 | **0.0233** | 0.0234 | 0.0286 | 0.0400 | 0.0400 | 0.0285 | **0.0265** | 0.0267 | 0.0270 |
| EV | **0.1333** | 0.0922 | 0.0923 | 0.1259 | 0.1256 | 0.1273 | 0.1276 | **0.1333** | 0.0875 | 0.0876 | 0.1270 | 0.1269 | 0.1303 | 0.1313 |

apply our proposed HSMFPCA to explore the variance patterns based on the profile data of process variables from all stages. In this real example, each stage involves a high-temperature and high-pressure environment, where the source material is stimulated and ejected as high-energy sputtered atoms (see Figure 1). These sputtered atoms fly to and are condensed at the surface of the raw glass material, thereby gradually forming a thin functional film. The three stages in Figure 1 share a similar process mechanism, but take different source materials to generate thin functional films with different desired properties.

Our real example data set consists of $N = 120$ samples. In each stage, five key process variables, i.e., voltage, current, pressure, air flow, and temperature, are continuously recorded every second, generating in total $M = 5 \times 3 = 15$ profiles. The durations of the three stage are programed to be 13, 49, and 13 seconds, and thus the data dimension of $\mathbf{x}_i$ is $P = 13 \times 5 + 49 \times 5 + 13 \times 5 = 375$. The original multiple profiles of the process variables from the three stages (see Figure S.4 in our supplementary materials file) are centered and standardized following the guidelines in Section 2.5, and are shown in Figure 11, where the elements of different profiles and stages are separated by dashed lines. At first glance, we observe that the variances of some elements in the pressure, air flow profiles in Stage 1 and in the voltage, current profiles in Stage 2 are much larger than those of the other elements in the other profiles and stages, indicating that there might be element-wise, profile-wise and stage-wise sparsity structures in the variance patterns of these multistage multivariate profiles data.

Our proposed HSMFPCA, which encourages a three-level hierarchical sparsity in the derived eigenvectors, is applied to this real data set. The number of significant PCs is determined as $K = 2$ by performing the hypothesis test discussed in Section 2.5. The tuning parameters are selected by the AIC. The estimated eigenvectors $\hat{\mathbf{v}}_1$ and $\hat{\mathbf{v}}_2$ are plotted in Figure 11. The first PC explains around 30% of the total data variance, and $\hat{\mathbf{v}}_1$ is very sparse with only 20 nonzero elements in the pressure and air flow profiles in Stage 1. From the manufacturing point of view, the positive correlation between the pressure and air flow profiles in $\hat{\mathbf{v}}_1$ is due to the pressure in the chamber being controlled by the air flow, and thus more air flow will produce higher pressure. The large variations at the initial and ending phases of the pressure and air flow profiles in Stage 1 might be caused by process instability during the warm-up and cool-down periods. The second PC explains about 10% of the data variance, and $\hat{\mathbf{v}}_2$ reveals that the voltage and current in Stage 2 are negatively correlated and have almost-cyclical large variations every 10 seconds during the production process in Stage 2, which is probably due to the fact that the PVD process control program will jointly adjust the voltage and current on occasion to maintain the product of them (i.e., the power) at a desired level. The third stage seems much more stable without significant variance patterns. The competing methods in Table 1 are also applied to this real example in our supplementary materials file. Our HSMFPCA is shown to be much better, generating much more sparse and interpretable results. In practice, these clear discoveries from our HSMFPCA can lead the practitioners to allocate more efforts to stabilize the gas-related process variables in Stage 1 and the electricity-related process variables in Stage 2.

## 5. Conclusion

Large-scale senor networks deployed in multistage production systems enable online sensing of all process variables at all stages. Based on the generated multistage multivariate profile data that have continuously tracked the real-time process status, this article proposes a novel PCA-based variance decomposition methodology to study the variation of the entire manufacturing process. Specifically, we integrate the conventional MFPCA with a three-level hierarchical sparsity idea to simultaneously investigate the stage-wise, profile-wise and element-wise sparsity, so that the informative key stages and process variables in each eigenvector can be clearly identified. Our HSMFPCA is developed by regression-type reformulating the PCA and reparameterizing the entries of eigenvectors, and is well equipped with an efficient optimization algorithm. Useful guidelines have been provided for practical implementations. The extensive simulations and a real example study of the PVD process have

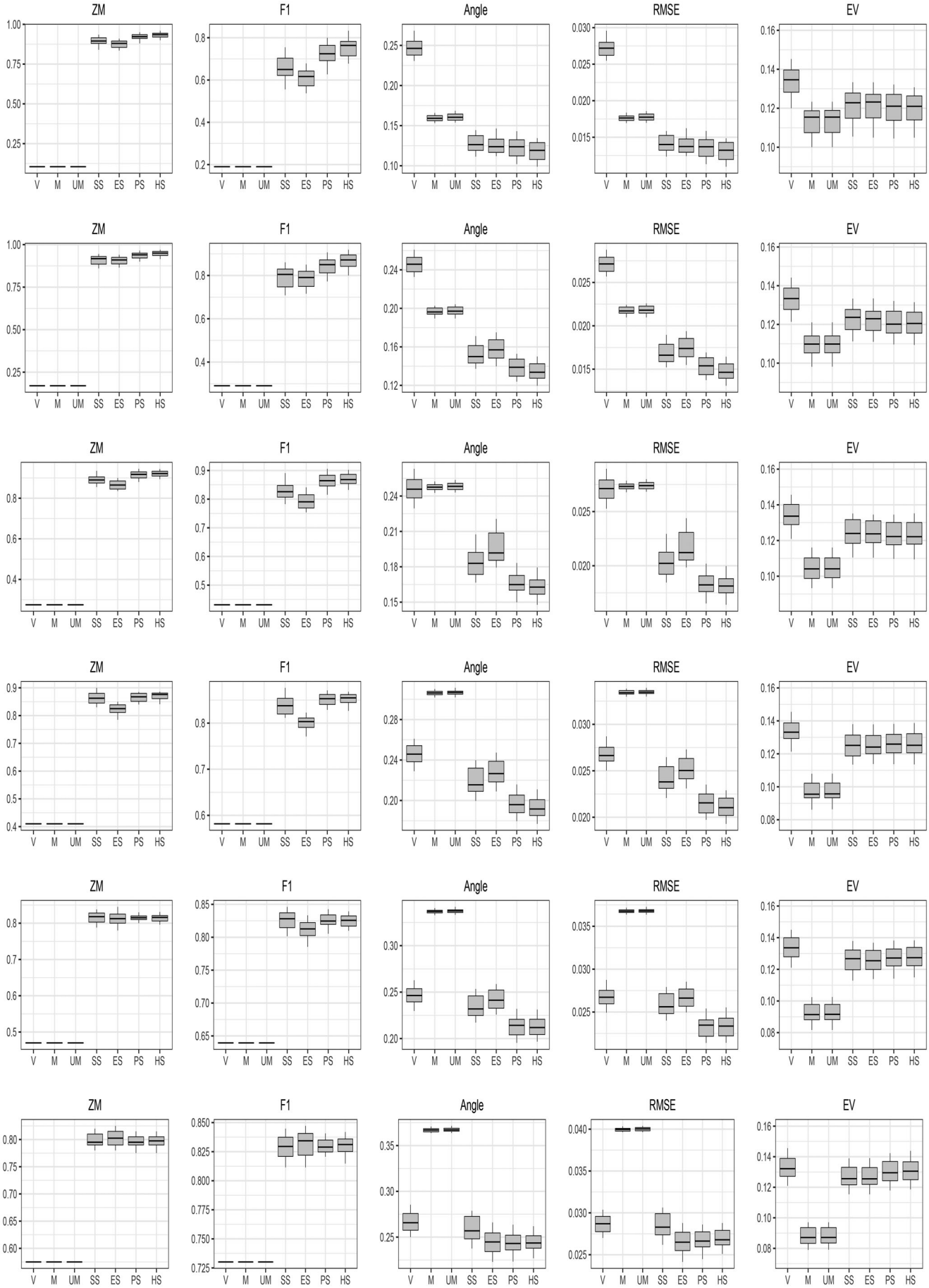


**Figure 7.** Boxplots of performance comparison. Rows 1 to 6 correspond to cases 1 to 6.

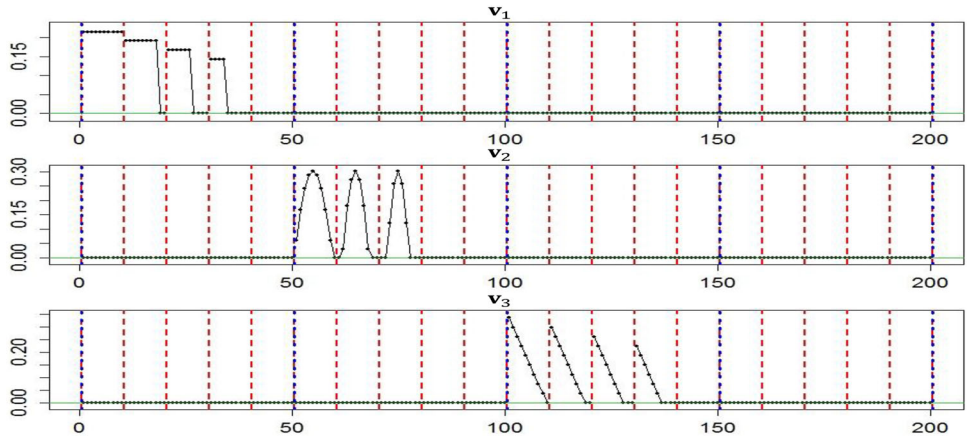


**Figure 8.** True eigenvectors in case 7: $\sigma_1^2 = 8^2, \sigma_2^2 = 5^2, \sigma_3^2 = 3^2$.

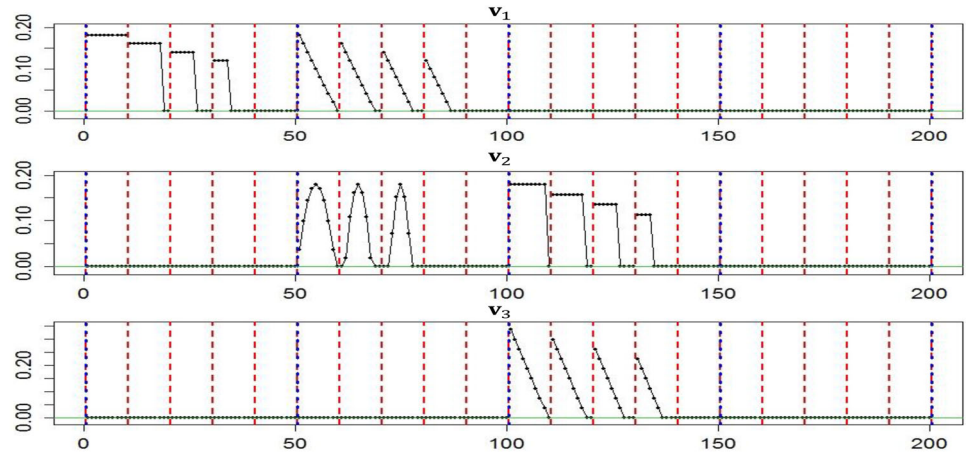


**Figure 9.** True eigenvectors in case 8: $\sigma_1^2 = 12^2, \sigma_2^2 = 8^2, \sigma_3^2 = 5^2$.

**Table 3.** Comparison results in case 7.

| | *V* | *M* | *UM* | *SS* | *ES* | *PS* | *HS* | *V* | *M* | *UM* | *SS* | *ES* | *PS* | HS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{v}_1$ | | | | | | | $\mathbf{v}_2$ | | | | | | |
| ZM | 0.1400 | 0.1400 | 0.1400 | 0.8613 | **0.8756** | 0.8673 | **0.8772** | 0.1050 | 0.1050 | 0.1050 | **0.9000** | 0.8414 | 0.8592 | 0.8725 |
| F1 | 0.2456 | 0.2456 | 0.2456 | 0.7127 | **0.7214** | 0.7000 | **0.7254** | 0.1900 | 0.1900 | 0.1900 | **0.7260** | 0.5806 | 0.6355 | 0.6630 |
| Angle | 0.1843 | 0.2653 | 0.2491 | 0.1183 | **0.1015** | 0.1156 | **0.1085** | 0.2842 | 0.7536 | 0.4118 | **0.1484** | **0.1470** | 0.1621 | **0.1458** |
| RMSE | 0.0204 | 0.0292 | 0.0275 | 0.0120 | **0.0113** | 0.0128 | **0.0111** | 0.0312 | 0.0615 | 0.0391 | **0.0153** | 0.0163 | 0.0179 | 0.0163 |
| EV | **0.2346** | 0.1930 | 0.1962 | 0.2256 | 0.2250 | 0.2253 | 0.2252 | **0.0986** | 0.0159 | 0.0592 | 0.0867 | 0.0891 | 0.0890 | 0.0888 |
| | $\mathbf{v}_3$ | | | | | | | *Overall* | | | | | | |
| ZM | 0.1500 | 0.1500 | 0.1500 | 0.8483 | 0.7853 | 0.8651 | **0.8887** | 0.1317 | 0.1317 | 0.1317 | 0.8698 | 0.8341 | 0.8638 | **0.8794** |
| F1 | 0.2609 | 0.2609 | 0.2609 | 0.6728 | 0.5425 | 0.6938 | **0.7393** | 0.2327 | 0.2327 | 0.2327 | 0.6747 | 0.5976 | 0.6635 | **0.6971** |
| Angle | 0.4253 | 0.8691 | 0.8144 | 0.2899 | 0.2885 | 0.2610 | **0.2538** | 0.1549 | 0.2399 | 0.2225 | 0.0964 | 0.0870 | 0.0890 | **0.0794** |
| RMSE | 0.0460 | 0.0719 | 0.0715 | 0.0306 | 0.0317 | 0.0285 | **0.0277** | 0.0344 | 0.0572 | 0.0507 | 0.0214 | 0.0218 | 0.0212 | **0.0191** |
| EV | **0.0459** | 0.0060 | 0.0141 | 0.0361 | 0.0382 | 0.0369 | 0.0364 | **0.3791** | 0.2150 | 0.2795 | 0.3485 | 0.3523 | 0.3512 | 0.3503 |

**Table 4.** Comparison results in case 8.

| | *V* | *M* | *UM* | *SS* | *ES* | *PS* | *HS* | *V* | *M* | *UM* | *SS* | *ES* | *PS* | HS |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | $\mathbf{v}_1$ | | | | | | | $\mathbf{v}_2$ | | | | | | |
| ZM | 0.2900 | 0.2900 | 0.2900 | 0.8003 | 0.7639 | 0.7990 | **0.8102** | 0.2400 | 0.2400 | 0.2400 | **0.7601** | 0.6783 | 0.7268 | 0.7346 |
| F1 | 0.4496 | 0.4496 | 0.4496 | 0.7461 | 0.7145 | 0.7481 | **0.7688** | 0.3871 | 0.3871 | 0.3871 | **0.6559** | 0.5957 | 0.6394 | **0.6584** |
| Angle | 0.1884 | 0.2968 | 0.2935 | 0.1559 | 0.1505 | 0.1643 | **0.1419** | 0.3266 | 0.9497 | 0.7645 | 0.2861 | 0.2865 | 0.3004 | **0.2685** |
| RMSE | 0.0208 | 0.0326 | 0.0322 | 0.0172 | **0.0166** | 0.0181 | **0.0164** | 0.0350 | 0.0610 | 0.0600 | 0.0312 | **0.0311** | 0.0323 | **0.0306** |
| EV | **0.3613** | 0.3144 | 0.3151 | 0.3565 | 0.3568 | 0.3568 | 0.3568 | **0.1788** | 0.0076 | 0.0272 | 0.1621 | 0.1749 | 0.1746 | 0.1746 |
| | $\mathbf{v}_3$ | | | | | | | *Overall* | | | | | | |
| ZM | 0.1500 | 0.1500 | 0.1500 | 0.6763 | 0.6523 | 0.6934 | **0.7104** | 0.2267 | 0.2267 | 0.2267 | **0.7456** | 0.6981 | 0.7397 | **0.7517** |
| F1 | 0.2609 | 0.2609 | 0.2609 | 0.4578 | 0.3762 | 0.4713 | **0.5012** | 0.3696 | 0.3696 | 0.3696 | 0.5816 | 0.5790 | 0.6290 | **0.6520** |
| Angle | 0.7294 | 0.9919 | 0.9376 | 0.9481 | 0.7271 | 0.7044 | **0.6992** | 0.1017 | 0.1986 | 0.2104 | 0.0755 | 0.0783 | 0.0756 | **0.0560** |
| RMSE | 0.0695 | 0.0902 | 0.0813 | 0.0945 | 0.0713 | 0.0671 | **0.0643** | 0.0470 | 0.0658 | 0.0616 | 0.0590 | 0.0464 | 0.0449 | **0.0426** |
| EV | **0.0317** | 0.0037 | 0.0082 | 0.0125 | 0.0270 | 0.0268 | 0.0265 | **0.5718** | 0.3257 | 0.3504 | 0.5311 | 0.5587 | 0.5582 | 0.5579 |

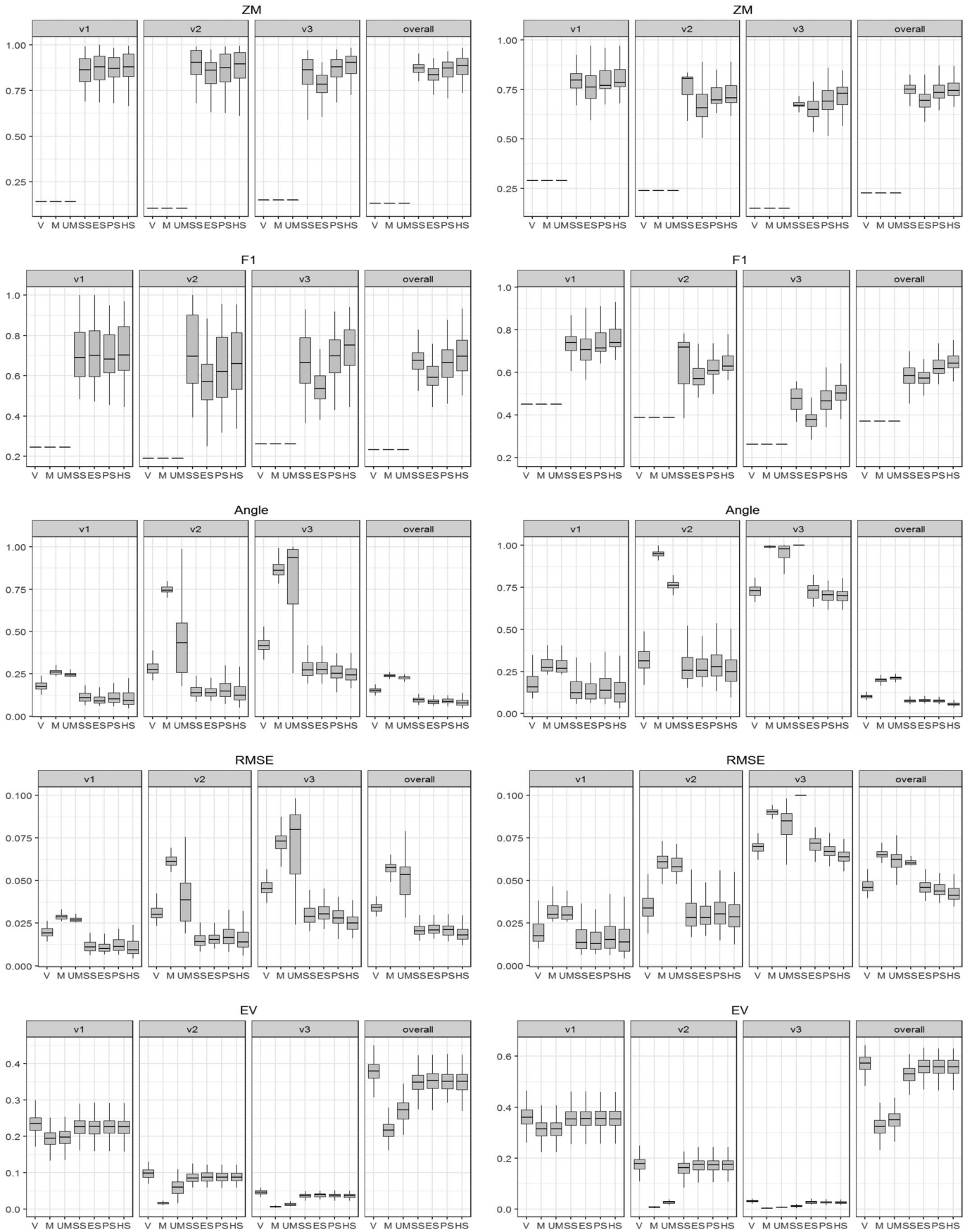


**Figure 10.** Boxplots of performance comparison. Columns 1 and 2 correspond to cases 7 and 8.

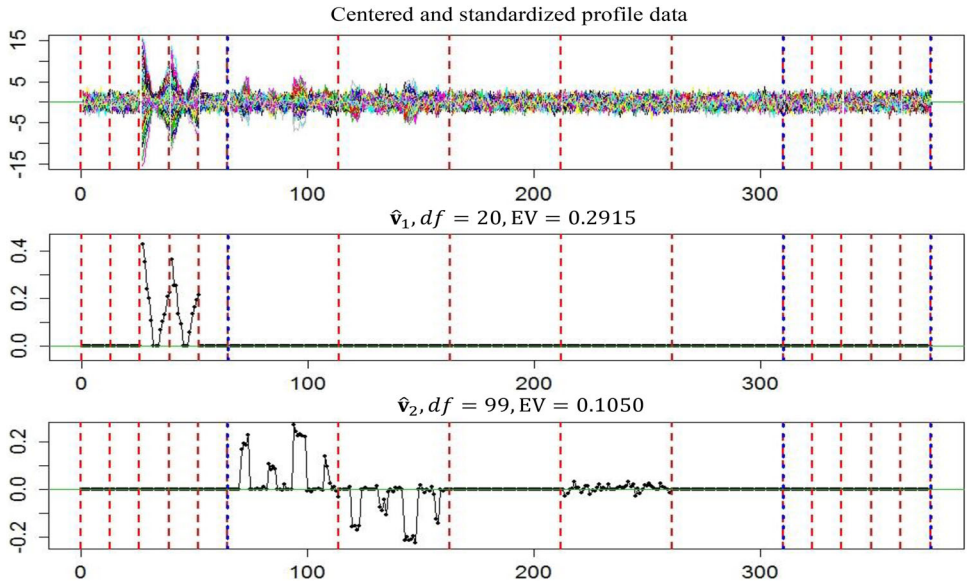


**Figure 11.** Profile data and the estimated eigenvectors in the real example.

validated the advantages of our proposed HSMFPCA in producing more parsimonious and accurate eigenvectors, which can be interpreted more clearly to provide practitioners with insightful knowledge about the process variation.

This article exploits sparse eigenvectors. Another issue related to the profile data is smoothness. Our method can be extended to utilize the smoothness structures in eigenvectors by replacing the LASSO penalty imposed on the element values of profiles by other tailored ones, e.g., the fused LASSO in Tibshirani *et al.* (2005). This will be formally studied in a separate future work. Our proposed HSMFPCA can also be taken as an ingredient of an online monitoring framework in Phase II for multivariate profile data as in Zhang *et al.* (2018a, 2018b), where the scores of a new sample with respect to the first few sparse eigenvectors and the residuals are plotted in the control charts for process surveillance. When the quality of the final product is available, modeling the relationship between the multistage multivariate profile data and the categorical or numerical quality characteristics is also a very promising research direction. The three-level hierarchical sparsity can be employed to address the high-dimensional challenge and enhance the model interpretability, its exact performance in this classification or regression context deserving our future efforts.

## Acknowledgments

The authors greatly acknowledge the valuable comments provided by the editor and three referees that have resulted in great improvements of this article.

## Funding

This work was supported by the National Natural Science Foundation of China under Grant 71931006, 71602155; Hong Kong RGC General Research Funds under Grant 16201718 and 16203917; Hong Kong Innovation and Technology Fund Postdoctoral Hub Program under Grant PiH/246/18; and Hong Kong Innovation and Technology Fund Project under Grant ITS/184/18FX.

## Notes on contributors

*Kai Wang* is currently an Assistant Professor in the Department of Industrial Engineering, School of Management, at the Xi'an Jiaotong University, Xi'an, China. He received his Ph.D. in Industrial Engineering and Logistics Management in 2018 from the HKUST, Hong Kong, and his bachelor's degree in Industrial Engineering in 2014 from Xi'an Jiaotong University, Shaanxi, China. His research focuses on industrial big data analytics, machine learning and transfer learning, statistical process control and monitoring.

*Fugee Tsung* is a Chair Professor in the Department of Industrial Engineering and Decision Analytics (IEDA), Director of the Quality and Data Analytics Lab (QLab), at the Hong Kong University of Science and Technology (HKUST), Hong Kong, China. He is a Fellow of the American Society for Quality, Fellow of the American Statistical Association, Academician of the International Academy for Quality, and Fellow of the Hong Kong Institution of Engineers. He received both his M.Sc. and Ph.D. from the University of Michigan, Ann Arbor, and his B.Sc. from the National Taiwan University. His research interests include quality analytics in advanced manufacturing and service processes, industrial big data and statistical process control, monitoring, and diagnosis.

## References

Akemann, G., Baik, J. and Di Francesco, P. (2011) *The Oxford Handbook of Random Matrix Theory*, Oxford University Press, Oxford, UK.

Allen, G.I. (2013) Sparse and functional principal components analysis. Cornell University, Ithaca, New York. *arXiv preprint arXiv: 1309.2895*.

Bishop, C.M. (2006) *Pattern Recognition and Machine Learning*, Springer, New York, NY.

Cadima, J. and Jolliffe, I.T. (1995) Loading and correlations in the interpretation of principle compenents. *Journal of Applied Statistics*, **22**, 203–214.

Chang, S.I. and Yadama, S. (2010) Statistical process control for monitoring non-linear profiles using wavelet filtering and B-spline approximation. *International Journal of Production Research*, **48**, 1049–1068.

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

Chicken, E., Pignatiello, J.J., Jr and Simpson, J.R. (2009) Statistical process monitoring of nonlinear profiles using wavelets. *Journal of Quality Technology*, **41**, 198–212.
Colosimo, B.M. and Pacella, M. (2007) On the use of principal component analysis to identify systematic patterns in roundness profiles. *Quality and Reliability Engineering International*, **23**, 707–725.
d'Aspremont, A., Bach, F. and Ghaoui, L.E. (2008) Optimal solutions for sparse principal component analysis. *Journal of Machine Learning Research*, **9**, 1269–1294.
Ding, Y., Zeng, L. and Zhou, S. (2006) Phase I analysis for monitoring nonlinear profiles in manufacturing processes. *Journal of Quality Technology*, **38**, 199–216.
El Karoui, N. (2003) On the largest eigenvalue of Wishart matrices with identity covariance when n, p and p/n $\to \infty$. arXiv preprint math/0309355.
Fang, X., Gebraeel, N.Z. and Paynabar, K. (2017) Scalable prognostic models for large-scale condition monitoring applications. *IISE Transactions*, **49**, 698–710.
Friedman, J., Hastie, T. and Tibshirani, R. (2010) A note on the group lasso and a sparse group lasso. arXiv preprint arXiv:1001.0736.
Grasso, M., Colosimo, B.M. and Pacella, M. (2014) Profile monitoring via sensor fusion: The use of PCA methods for multi-channel data. *International Journal of Production Research*, **52**, 6110–6135.
Johnstone, I.M. and Lu, A.Y. (2009) On consistency and sparsity for principal components analysis in high dimensions. *Journal of the American Statistical Association*, **104**, 682–693.
Kang, L. and Albin, S.L. (2000) On-line monitoring when the process yields a linear profile. *Journal of Quality Technology*, **32**, 418–426.
Khan, Z., Shafait, F. and Mian, A. (2015) Joint group sparse PCA for compressed hyperspectral imaging. *IEEE Transactions on Image Processing : A Publication of the IEEE Signal Processing Society*, **24**, 4934–4942.
Linn, R.J., Au, E. and Tsung, F. (2002) Process capability improvement for multistage processes. *Quality Engineering*, **15**, 281–292.
Liu, K. and Shi, J. (2013) Objective-oriented optimal sensor allocation strategy for process monitoring and diagnosis by multivariate analysis in a Bayesian network. *IIE Transactions*, **45**, 630–643.
Ma, Z. (2013) Sparse principal component analysis and iterative thresholding. *The Annals of Statistics*, **41**, 772–801.
Noorossana, R., Saghaei, A. and Amiri, A. (2011) *Statistical Analysis of Profile Monitoring*, John Wiley & Sons, Hoboken, NJ.
Paynabar, K., Jin, J. and Pacella, M. (2013) Monitoring and diagnosis of multichannel nonlinear profile variations using uncorrelated multilinear principal component analysis. *IIE Transactions*, **45**, 1235–1247.
Paynabar, K., Jin, J. and Reed, M.P. (2015) Informative sensor and feature selection via hierarchical nonnegative garrote. *Technometrics*, **57**, 514–523.
Paynabar, K., Zou, C. and Qiu, P. (2016) A change-point approach for Phase-I analysis in multivariate profile monitoring and diagnosis. *Technometrics*, **58**, 191–204.
Qiu, P., Zou, C. and Wang, Z. (2010) Nonparametric profile monitoring by mixed effects modeling. *Technometrics*, **52**, 265–277.
Ramsay, J. (2005) *Functional Data Analysis*, Springer, New York, NY.
Shang, Y., Zi, X., Tsung, F. and He, Z. (2014) LASSO-based diagnosis scheme for multistage processes with binary data. *Computers & Industrial Engineering*, **72**, 198–205.
Shen, H. and Huang, J.Z. (2008) Sparse principal component analysis via regularized low rank matrix approximation. *Journal of Multivariate Analysis*, **99**, 1015–1034.
Shu, L.J. and Tsung, F. (2003) On multistage statistical process control. *Journal of the Chinese Institute of Industrial Engineers*, **20**, 1–8.
Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013) A sparse-group lasso. *Journal of Computational and Graphical Statistics*, **22**, 231–245.
Tibshirani, R., Saunders, M., Rosset, S., Zhu, J., and Knight, K. (2005) Sparsity and smoothness via the fused lasso. *Journal of The Royal Statistical Society Series B-Statistical Methodology*, **67**(1), 91–108.
Tibshirani, R. (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, **58**, 267–288.
Tseng, P. (2001) Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of Optimization Theory and Applications*, **109**, 475–494.
Wang, Y., Mei, Y. and Paynabar, K. (2018) Thresholded multivariate principal component analysis for phase I multichannel profile monitoring. *Technometrics*, **60**, 360–372.
Witten, D.M., Tibshirani, R. and Hastie, T. (2009) A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics (Oxford, England)*, **10**, 515–534.
Xiang, L. and Tsung, F. (2008) Statistical monitoring of multi-stage processes based on engineering models. *IIE Transactions*, **40**, 957–970.
Yu, G., Zou, C. and Wang, Z. (2012) Outlier detection in functional observations with applications to profile monitoring. *Technometrics*, **54**, 308–318.
Yuan, M. and Lin, Y. (2006) Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, **68**, 49–67.
Zhang, C., Yan, H., Lee, S. and Shi, J. (2018a) Multiple profiles sensor-based monitoring and anomaly detection. *Journal of Quality Technology*, **50**, 344–362.
Zhang, C., Yan, H., Lee, S. and Shi, J. (2018b) Weakly correlated profile monitoring based on sparse multi-channel functional principal component analysis. *IISE Transactions*, **50**, 878–891.
Zhou, N. and Zhu, J. (2010) Group variable selection via a hierarchical lasso and its oracle property. arXiv preprint arXiv:1006.2871.
Zou, C., Tsung, F. and Wang, Z. (2008) Monitoring profiles based on nonparametric regression methods. *Technometrics*, **50**, 512–526.
Zou, H., Hastie, T. and Tibshirani, R. (2006) Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, **15**, 265–286.

# Appendices

## Appendix A

The proof of Theorem 1 can be done by following the work in Zou *et al.* (2006), which is detailed in this appendix to make this article self-contained.

*Proof of Part (a).* First, let $\mathbf{A}_\perp$ be any orthonormal $P \times (P-1)$ matrix such that $[\boldsymbol{\alpha}, \mathbf{A}_\perp]$ is $P \times P$ orthonormal. Then $\mathrm{tr}(\mathbf{X}\mathbf{X}^T) = \mathrm{tr}(\mathbf{X}[\boldsymbol{\alpha}, \mathbf{A}_\perp][\boldsymbol{\alpha}, \mathbf{A}_\perp]^T\mathbf{X}^T) = \mathrm{tr}(\mathbf{X}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{X}^T) + \mathrm{tr}(\mathbf{X}\mathbf{A}_\perp\mathbf{A}_\perp^T\mathbf{X}^T)$. So

$$\begin{aligned}
||\mathbf{X} - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T||_F^2 &= \mathrm{tr}((\mathbf{X} - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T)(\mathbf{X} - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T)^T) \\
&= \mathrm{tr}(\mathbf{X}\mathbf{X}^T - \mathbf{X}\boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{X}^T - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T\mathbf{X}^T + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T) \\
&= \mathrm{tr}(\mathbf{X}\mathbf{A}_\perp\mathbf{A}_\perp^T\mathbf{X}^T) + \mathrm{tr}(\mathbf{X}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{X}^T - \mathbf{X}\boldsymbol{\alpha}\boldsymbol{\beta}^T\mathbf{X}^T \\
&\quad - \mathbf{X}\boldsymbol{\beta}\boldsymbol{\alpha}^T\mathbf{X}^T + \mathbf{X}\boldsymbol{\beta}\boldsymbol{\beta}^T\mathbf{X}^T) \\
&= ||\mathbf{X}\mathbf{A}_\perp||_F^2 + ||\mathbf{X}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}||^2.
\end{aligned}$$

When $\boldsymbol{\alpha}$ is fixed in Model (3), $\hat{\boldsymbol{\beta}} = \mathrm{argmin}_{\boldsymbol{\beta}}||\mathbf{X}\boldsymbol{\alpha} - \mathbf{X}\boldsymbol{\beta}||^2 + \lambda||\boldsymbol{\beta}||^2 = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$. By substituting $\hat{\boldsymbol{\beta}}$ in the objective function of Model (3), we get

$$\begin{aligned}
||\mathbf{X} - \mathbf{X}\hat{\boldsymbol{\beta}}\boldsymbol{\alpha}^T||_F^2 + \lambda||\hat{\boldsymbol{\beta}}||^2 &= ||\mathbf{X}\mathbf{A}_\perp||_F^2 + ||\mathbf{X}\boldsymbol{\alpha} - \mathbf{X}\hat{\boldsymbol{\beta}}||^2 + \lambda||\hat{\boldsymbol{\beta}}||^2 \\
&= \mathrm{tr}(\mathbf{X}\mathbf{A}_\perp\mathbf{A}_\perp^T\mathbf{X}^T) + \mathrm{tr}(\mathbf{X}\boldsymbol{\alpha}\boldsymbol{\alpha}^T\mathbf{X}^T) \\
&\quad - \boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} \\
&= \mathrm{tr}(\mathbf{X}\mathbf{X}^T) - \boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha},
\end{aligned}$$

and then $\hat{\boldsymbol{\alpha}} = \mathrm{argmax}_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$, subject to $\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$.

It is obvious that $\hat{\boldsymbol{\alpha}}$ should be the first eigenvector of $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}$. By using SVD, we have $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^T$, and $\mathbf{X}^T\mathbf{X}(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X} = \mathbf{V}\mathbf{D}^2(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}^2\mathbf{V}^T$. It is now proved that $\hat{\boldsymbol{\alpha}}$ is the the first column of $\mathbf{V}$, which is the also the first eigenvector $\mathbf{v}$

in Model (2), and $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{X}\hat{\boldsymbol{\alpha}} = \mathbf{V}(\mathbf{D}^2 + \lambda\mathbf{I})^{-1}\mathbf{D}^2\mathbf{V}^T\mathbf{v} = d_1^2/(d_1^2 + \lambda)\mathbf{v} \propto \mathbf{v}$, where $d_1$ is the first diagonal element of $\mathbf{D}$.

*Proof of Part (b)* . Consider a new problem

$$\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\} = \text{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} ||\mathbf{X} - \mathbf{X}\frac{\boldsymbol{\beta}}{1+\lambda}\boldsymbol{\alpha}^T||_F^2 + \lambda||\frac{\boldsymbol{\beta}}{1+\lambda}||^2 \quad \text{s.t.}\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1.$$

It can be seen that $\boldsymbol{\alpha}^* = \hat{\boldsymbol{\alpha}} = \mathbf{v}$ and $\boldsymbol{\beta}^* = (1+\lambda)\hat{\boldsymbol{\beta}} \propto \mathbf{v}$, and thus this problem is equivalent to Model (3). Note that

$$\begin{aligned}
&||\mathbf{X} - \mathbf{X}\frac{\boldsymbol{\beta}}{1+\lambda}\boldsymbol{\alpha}^T||_F^2 + \lambda||\frac{\boldsymbol{\beta}}{1+\lambda}||^2 \\
&= \text{tr}\left(\mathbf{X}\mathbf{X}^T - \mathbf{X}\boldsymbol{\alpha}\frac{\boldsymbol{\beta}^T}{1+\lambda}\mathbf{X}^T - \mathbf{X}\frac{\boldsymbol{\beta}}{1+\lambda}\boldsymbol{\alpha}^T\mathbf{X}^T + \mathbf{X}\frac{\boldsymbol{\beta}\boldsymbol{\beta}^T}{(1+\lambda)^2}\mathbf{X}^T\right) \\
&\quad + \frac{\lambda}{(1+\lambda)^2}\text{tr}(\boldsymbol{\beta}^T\boldsymbol{\beta}) \\
&= \text{tr}(\mathbf{X}\mathbf{X}^T) + \frac{1}{1+\lambda}(-2\text{tr}(\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) + \frac{1}{1+\lambda}\text{tr}(\boldsymbol{\beta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) \\
&\quad + \frac{\lambda}{1+\lambda}\text{tr}(\boldsymbol{\beta}^T\boldsymbol{\beta})) \\
&= \text{tr}(\mathbf{X}\mathbf{X}^T) + \frac{1}{1+\lambda}\left(-2\text{tr}(\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}) + \text{tr}\left(\boldsymbol{\beta}^T\frac{\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}}{1+\lambda}\boldsymbol{\beta}\right)\right),
\end{aligned}$$

which implies

$$\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\} = \text{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} -2\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\frac{\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I}}{1+\lambda}\boldsymbol{\beta} \quad \text{s.t.}\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1.$$

As $\lambda \to \infty, (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})/(1+\lambda) \to \mathbf{I}$ and $-2\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} + \boldsymbol{\beta}^T\boldsymbol{\beta} = (\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} - \boldsymbol{\beta})^T(\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} - \boldsymbol{\beta}) - \boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$, so the above problem is further equivalent to

$$\{\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*\} = \text{argmin}_{\boldsymbol{\alpha},\boldsymbol{\beta}} ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha} - \boldsymbol{\beta}||^2 - ||\mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}||^2 \quad \text{s.t.}\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1. \qquad \square$$

## Appendix B

*Proof of Proposition 1.* When $\boldsymbol{\alpha}$ is fixed, $\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}$ is known, and Model (6) is reduced to

$$\begin{aligned}
&\min_{\{\gamma_s, \eta_{sj}, \theta_{sjt}\}_{s,j,t}} \quad ||\mathbf{y} - \boldsymbol{\beta}||^2 + \lambda_1\sum_{s=1}^{S}\gamma_s + \lambda_2\sum_{s=1}^{S}\sum_{j=1}^{M_s}\eta_{sj} + \lambda_3\sum_{s=1}^{S}\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}|\theta_{sjt}| \\
&\text{s.t.} \quad \beta_{sjt} = \gamma_s\eta_{sj}\theta_{sjt}, s = 1, ..., S, \text{j} = 1, ..., M_s, t = 1, ..., T_s, \\
&\gamma_s \geq 0, \eta_{sj} \geq 0.
\end{aligned}$$

First, given $\gamma_s$ and $\eta_{sj}$, $\hat{\theta}_{sjt} = \text{argmin}_{\theta_{sjt}}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt})^2 + \lambda_3|\theta_{sjt}|$. It is clear that if $\gamma_s\eta_{sj} = 0, \hat{\theta}_{sjt} = 0$ due to $\lambda_3 > 0$. Otherwise, by letting the subgradient being zero, i.e., $-2\gamma_s\eta_{sj}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt}) + \lambda_3 g = 0 \ (g \in [-1, 1])$, we have

$$\hat{\theta}_{sjt} = \text{sign}\left(\frac{y_{sjt}}{\gamma_s\eta_{sj}}\right) \cdot \left(\frac{|y_{sjt}|}{\gamma_s\eta_{sj}} - \frac{\lambda_3}{2(\gamma_s\eta_{sj})^2}\right)_+.$$

Thus, the updating equation of $\theta_{sjt}$ can be summarized as

$$\hat{\theta}_{sjt} = \mathbb{I}(\gamma_s\eta_{sj} > 0) \cdot \text{sign}(y_{sjt}) \cdot \left(\frac{|y_{sjt}|}{\gamma_s\eta_{sj}} - \frac{\lambda_3}{2(\gamma_s\eta_{sj})^2}\right)_+.$$

Second, if $\gamma_s$ and $\theta_{sjt}$ are given, $\hat{\eta}_{sj} = \text{argmin}_{\eta_{sj}}\sum_{t=1}^{T_s}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt})^2 + \lambda_2\eta_{sj}, \text{s.t.}\eta_{sj} \geq 0$. If $\gamma_s = 0$ or all $\theta_{sjt} = 0, \hat{\eta}_{sj} = 0$ as $\lambda_2 > 0$. Otherwise, as the simplified problem is convex, the Karush-Kuhn-Tucker (KKT) sufficient conditions can be given as

$$\eta_{sj} \geq 0, \tau \geq 0, \eta_{sj}\tau = 0, \ -2\sum_{t=1}^{T_s}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt})\gamma_s\theta_{sjt} + \lambda_2 - \tau = 0,$$

where $\tau$ is the Lagrange multiplier. Solving these equations gives $\hat{\eta}_{sj} = (\sum_{t=1}^{T_s} y_{sjt}\gamma_s\theta_{sjt} - \lambda_2/2)_+ / \sum_{t=1}^{T_s}(\gamma_s\theta_{sjt})^2$. To sum up, the updating equation of $\theta_{sjt}$ is

$$\begin{aligned}
\hat{\eta}_{sj} = &\mathbb{I}(\gamma_s > 0) \cdot \mathbb{I}(\exists t, \theta_{sjt} \neq 0) \\
&\cdot \left(\sum_{t=1}^{T_s}\frac{(\gamma_s\theta_{sjt})^2}{\sum_{t=1}^{T_s}(\gamma_s\theta_{sjt})^2}\frac{y_{sjt}}{\gamma_s\theta_{sjt}} - \frac{\lambda_2}{2\sum_{t=1}^{T_s}(\gamma_s\theta_{sjt})^2}\right)_+.
\end{aligned}$$

Finally, given $\eta_{sj}$ and $\theta_{sjt}$, $\hat{\gamma}_s = \text{argmin}_{\gamma_s}\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt})^2 + \lambda_1\gamma_s, \text{s.t.}\gamma_s \geq 0$. Obviously, if all $\eta_{sj}\theta_{sjt} = 0, \hat{\gamma}_s = 0$ as $\lambda_1 > 0$. Otherwise, we also solve this convex problem based on the KKT sufficient conditions as

$$\gamma_s \geq 0, \tau \geq 0, \gamma_s\tau = 0, \ -2\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}(y_{sjt} - \gamma_s\eta_{sj}\theta_{sjt})\eta_{sj}\theta_{sjt} + \lambda_1 - \tau = 0,$$

which yields $\hat{\gamma}_s = (\sum_{j=1}^{M_s}\sum_{t=1}^{T_s} y_{sjt}\eta_{sj}\theta_{sjt} - \lambda_1/2)_+ / \sum_{j=1}^{M_s}\sum_{t=1}^{T_s}(\eta_{sj}\theta_{sjt})^2$. Thus, the closed form updating equation of $\gamma_s$ is

$$\begin{aligned}
\hat{\gamma}_s = &\mathbb{I}(\exists(j,t), \eta_{sj}\theta_{sjt} \neq 0) \\
&\cdot \left(\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}\frac{(\eta_{sj}\theta_{sjt})^2}{\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}(\eta_{sj}\theta_{sjt})^2}\frac{y_{sjt}}{\eta_{sj}\theta_{sjt}} - \frac{\lambda_1}{2\sum_{j=1}^{M_s}\sum_{t=1}^{T_s}(\eta_{sj}\theta_{sjt})^2}\right)_+.
\end{aligned}$$

$\square$

*Proof of Proposition 2.* When $\boldsymbol{\beta}$ is fixed, the problem (6) is reduced to

$$\min_{\boldsymbol{\alpha}} \quad ||\mathbf{y} - \boldsymbol{\beta}||^2 - ||\mathbf{y}||^2 \quad \text{s.t.}\mathbf{y} = \mathbf{X}^T\mathbf{X}\boldsymbol{\alpha}, \boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1,$$

so $\hat{\boldsymbol{\alpha}} = \text{argmax}_{\boldsymbol{\alpha}}\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}, \text{s.t.}\boldsymbol{\alpha}^T\boldsymbol{\alpha} = 1$. By the Cauchy–Schwartz inequality, $\boldsymbol{\alpha}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\beta} \leq ||\boldsymbol{\alpha}|| \cdot ||\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}||$, and the equality holds when $\boldsymbol{\alpha} \propto \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}$. Therefore, the updating equation of $\boldsymbol{\alpha}$ is $\hat{\boldsymbol{\alpha}} = \mathbf{X}^T\mathbf{X}\boldsymbol{\beta}/||\mathbf{X}^T\mathbf{X}\boldsymbol{\beta}||$. $\square$