Supplement to “Longitudinal Principal Component Analysis with an Application to Marketing Data”

Christopher Kinson, Xiwei Tang, Zhen Zuo, Annie Qu

1 Supplementary Simulation Results

1.1 Marginal time-varying curves for the estimated eigenvector loadings

We provide the marginal time-varying curves for the estimated eigenvector loadings in the simulation study in Section 4. In the following plots, each curve represents a time-varying loading of the kth ($k = 1, 2$) eigenvector, that is, $\hat{w}_{kj}(t)$ for $j = 1, \ldots, 20$.

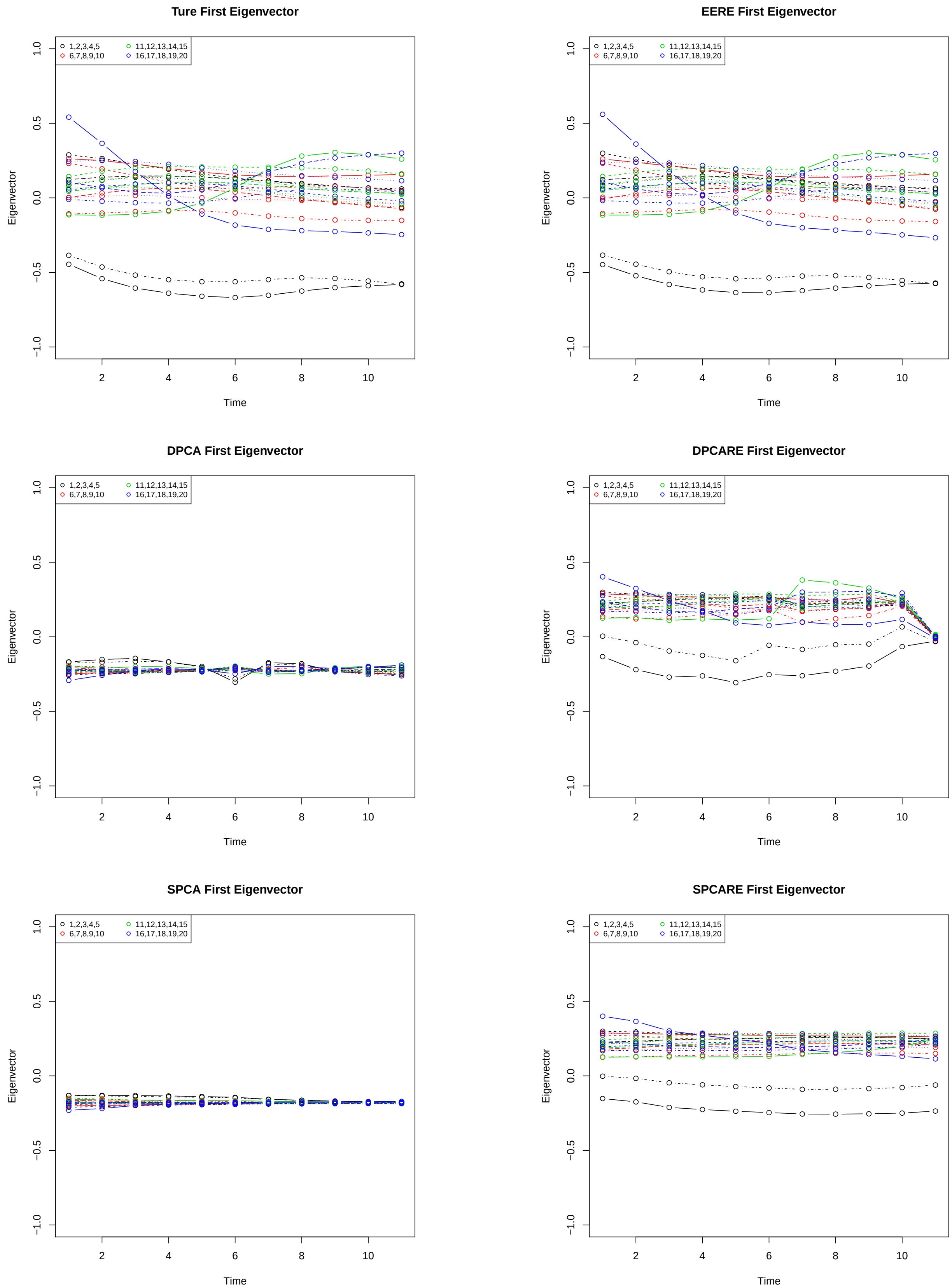

Figure 1: The estimated first eigenvector curves (each curve represents a time-varying loading of the eigenvector) of different models in the simulation study, based on 100 replications.

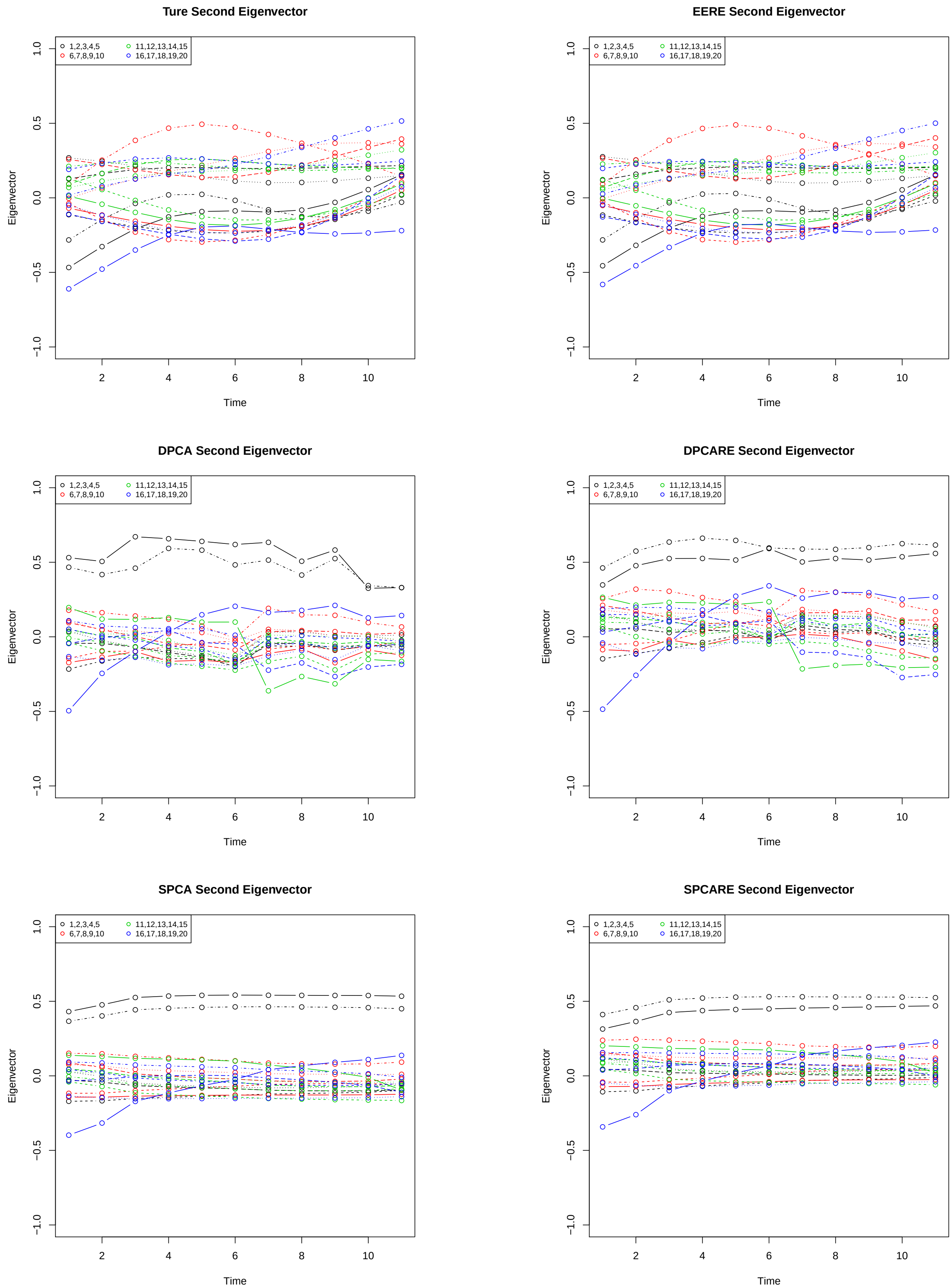

Figure 2: The estimated second eigenvector curves (each curve represents a time-varying loading of the eigenvector) of different models in the simulation study, based on 100 replications.

1.2 Time-varying Correlation Matrix

We conduct an additional simulation study to investigate a case of time-varying correlation matrix, where we obtain the time-varying eigenvectors. Specifically, we set the number of products as $J = 10$, the number of stores as $N = 100$ and the number of time points as $T = 20$. The eigenvectors are generated from the designed $J \times J$ covariance matrices $W_t(w_{pqt})$ $(t = 1, \ldots, T)$. Let $A^\star$, $B^\star$ and $C^\star$ denote three index sets: $A^\star = \{1, \cdots, 5\}$, $B^\star = \{6, 7\}$, and $C^\star = \{8, 9, 10\}$, respectively. For $m \neq n$, we set $w_{mnt} = 0.70$ if $m, n \in A^\star$, set $w_{mnt} = 0.10 + 0.07t$ if $m \in B^\star$ and $n \in A^\star$, set $w_{mnt} = 0.80$ if $m, n \in B^\star$, set $w_{mnt} = 0.10$ if $m \in C^\star$ and $n \in A^\star$, set $w_{mnt} = 0.70 - 0.07t$ if $m \in C^\star$ and $n \in B^\star$, and set $w_{mnt} = 0.80$ if $m, n \in C^\star$. Then the eigenvectors $\boldsymbol{e}_{kt}^0$'s $(k = 1, \ldots, J)$ are generated as the eigenvectors of W_t.

Furthermore, the eigenvalues are generated in a similar way as in the simulation study in the paper:

$$\alpha_{1t}^0 = \left(a_{10} + a_{11}\frac{t}{T} + a_{12}(\frac{t}{T})^2\right)^{-1} \text{ and } \alpha_{2t}^0 = \left(a_{20} + a_{21}\frac{t}{T} + a_{22}(\frac{t}{T})^2\right)^{-1},$$

where $(a_{10}, a_{11}, a_{12})^\top = (0.19, 0.20, 0.20)^\top$, $(a_{20}, a_{21}, a_{22})^\top = (0.30, 0.50, -0.50)^\top$. The remaining $J - 2$ eigenvalues are constants around 1 at each time point. Then we generate random vector $\boldsymbol{y}_{it}^*$ $(i = 1, \ldots, N)$ from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{V}_t^0 = \sum_{k=1}^{10} \alpha_{kt}^0 \boldsymbol{w}_{kt}^0 (\boldsymbol{w}_{kt}^0)^T$.

Similarly, we generate the multivariate random effect $\boldsymbol{\gamma}_i$ from a multivariate normal distribution with mean $\mathbf{0}$ and an exchangeable covariance matrix $\sigma^2(\rho \mathbf{1}_{10}^T \mathbf{1}_{10} + (1 - \rho)\boldsymbol{I}_{10})$, where $\sigma^2 = 3$ and $\rho = 0.25$. Consequently, the J-dimensional response variable is generated as $\boldsymbol{y}_{it} = \boldsymbol{y}_{it}^* + \boldsymbol{\gamma}_i$, for $i = 1, \ldots, N$ and $t = 1, \ldots, T$.

The MADE values as well as the curve plots and the heatmaps are presented to evaluate the estimation of the time-varying eigenvalues and eigenvectors. Table 1 summarizes the average MADE values for the eigenvalue estimations based on 100 replications. The results show that the EERE method yields the smallest MADE values compared to other competing methods, with improvements of at least 167% and 34% for the first and second eigenvalue estimation, respectively. Figure 3 displays the estimated first and second time-varying eigenvalues obtained from various methods. The heatmaps in Figures 4 and 5 illustrate the estimated time-varying eigenvector estimators. The additional simulation results present a similar pattern to the simulation study in the paper, indicating that the smooth eigenvalue and eigenvector estimators from the EERE method are closest to the true ones.

Table 1: Average MADEs of the first and second eigenvalues for the proposed EERE, the discrete PCA (DPCA), the DPCA with random effects (DPCARE), the smooth PCA (SPCA) and the SPCA with random effects (SPCARE) in the simulation study, based on 100 replications. There are $J = 10$ products, $N = 100$ stores, and $T = 20$ time points.

MADE	1^{st} Eigenvalue	2^{nd} Eigenvalue
EERE	0.39	0.32
DPCA	4.07	0.57
DPCARE	1.21	0.56
SPCA	3.91	0.43
SPCARE	1.04	0.45

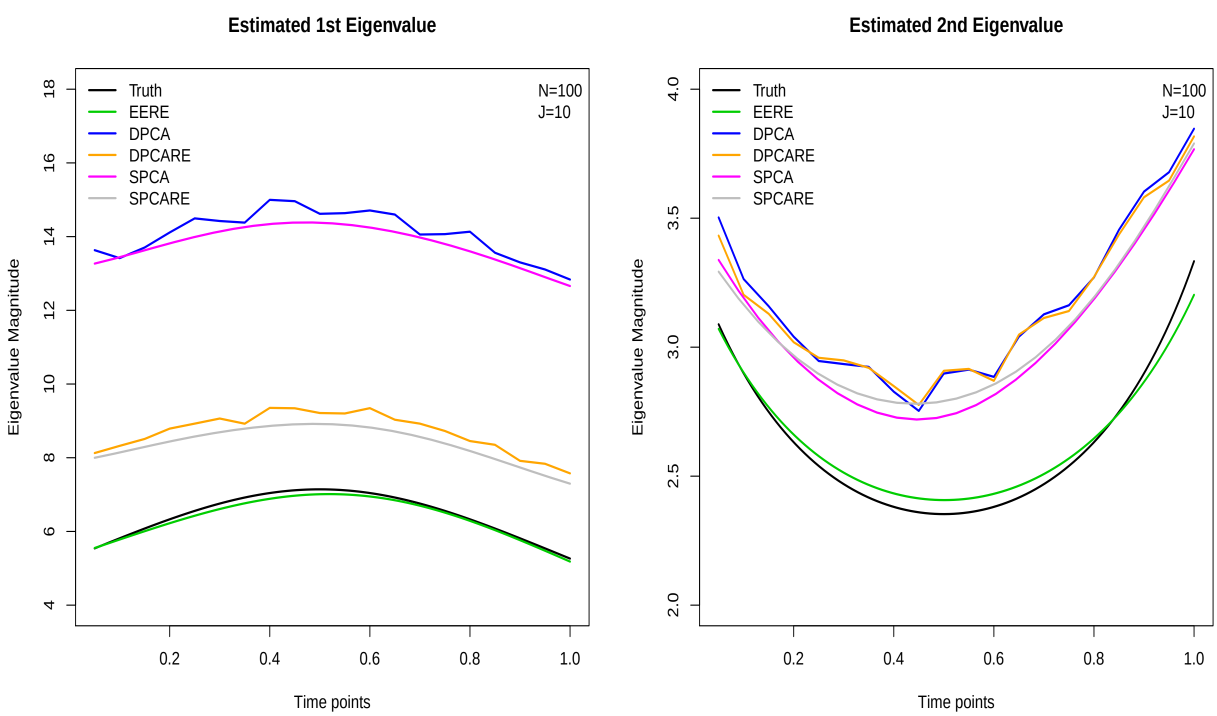

Figure 3: Estimated time-varying curve for the first and second eigenvalues in this simulation study.

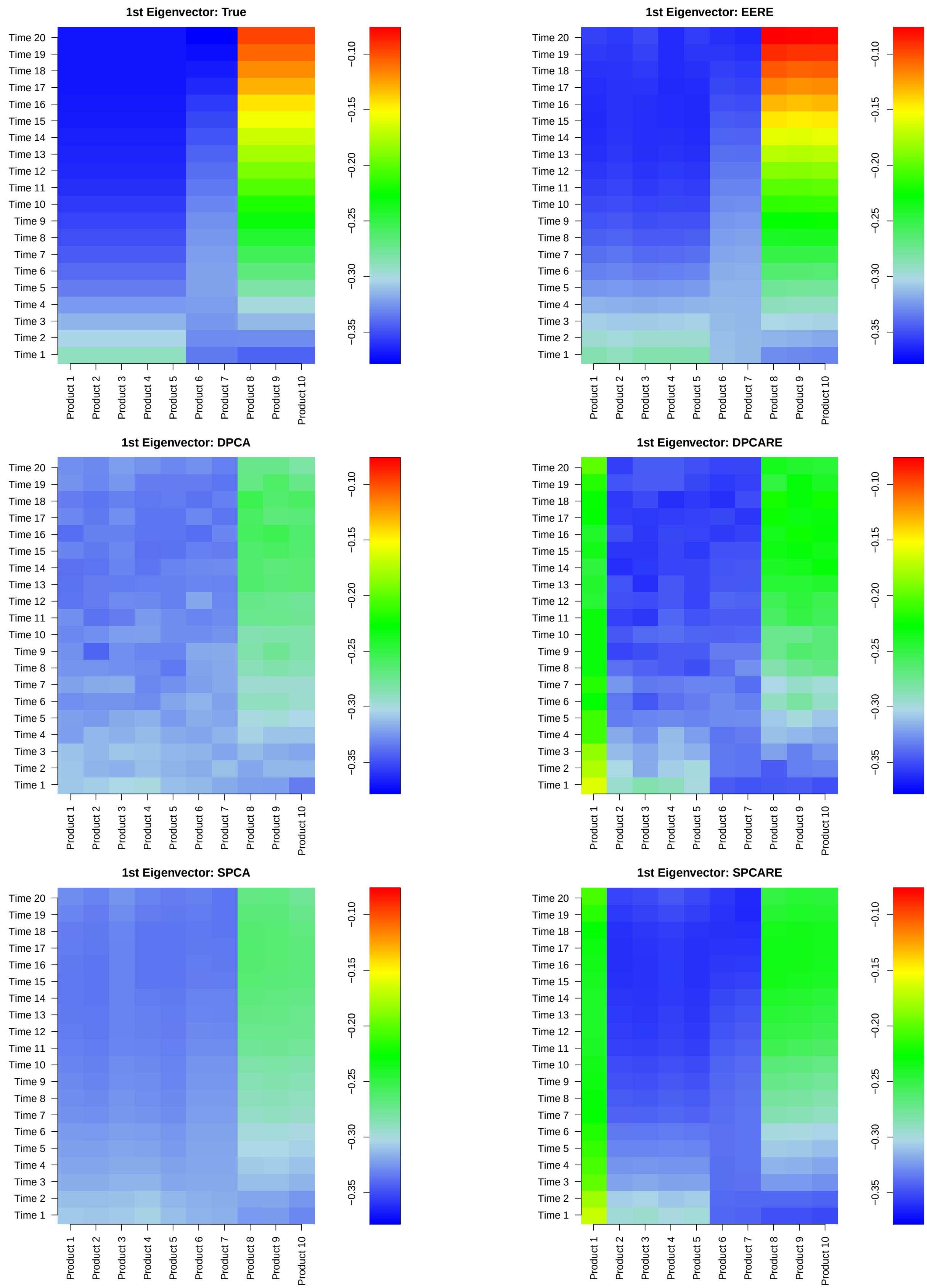

Figure 4: The heatmaps of the estimated first eigenvector over time in this simulation study.

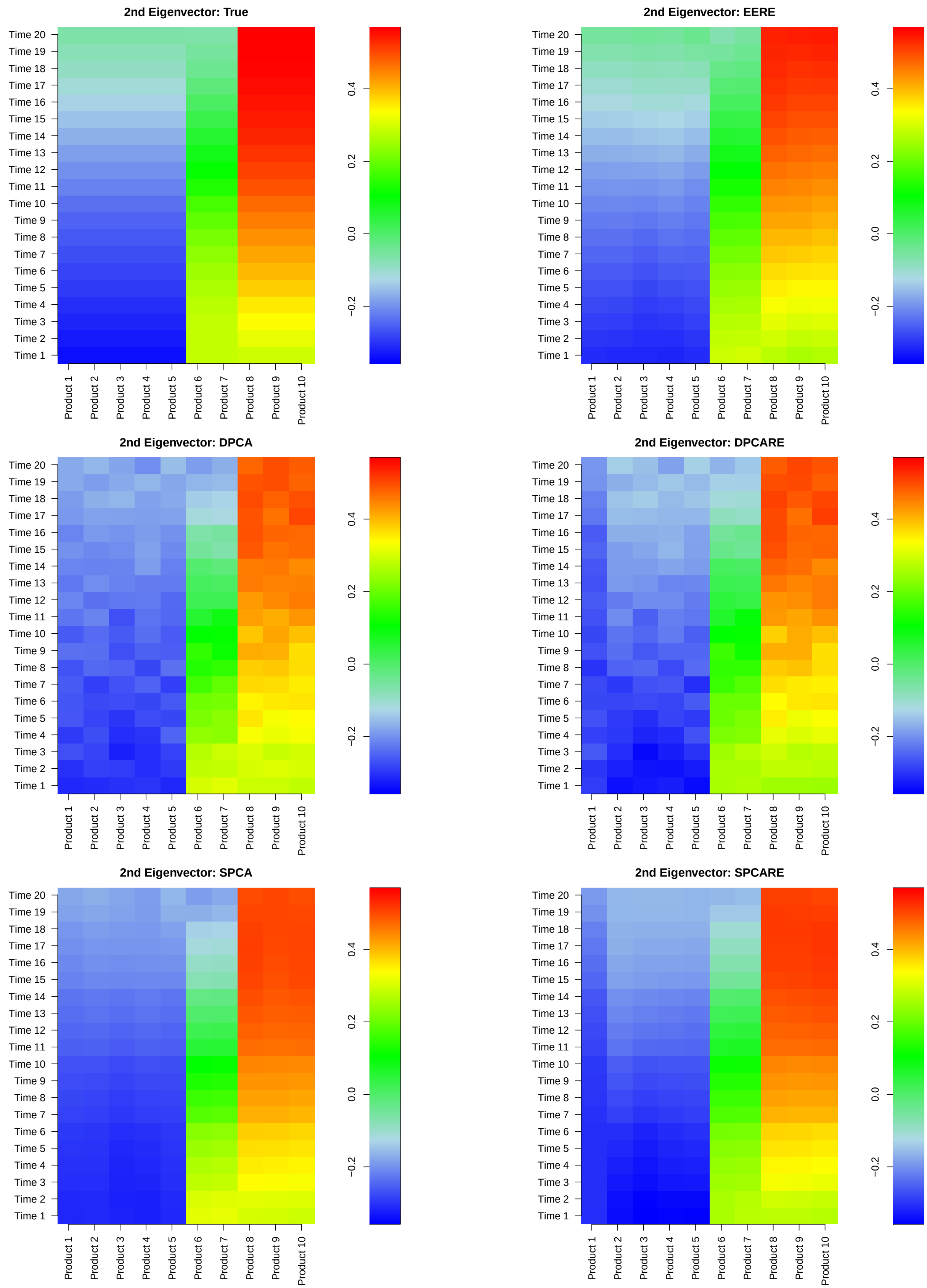

Figure 5: The heatmaps of the estimated second eigenvector over time in this simulation study.

1.3 Different Objective Functions

In this subsection, we compare the estimation results based on three different objective functions discussed in Section 2.3, the GEE-type function, the Frobenious-norm-based function and the Likelihood-based function, through a simple simulation study on the time-varying eigenvalue estimation.

In this simulation, we set the number of products as $J = 10$, the number of stores as $N = 100$ and the number of time points as $T = 20$. The first two eigenvalues are generated as follows

$$\alpha_{1t}^0 = \Big(a_{10} + a_{11}\frac{t}{T} + a_{12}(\frac{t}{T})^2\Big)^{-1} \text{ and } \alpha_{2t}^0 = \Big(a_{20} + a_{21}\frac{t}{T} + a_{22}(\frac{t}{T})^2\Big)^{-1},$$

where $(a_{10}, a_{11}, a_{12})^\top = (0.19, 0.20, 0.20)^\top$, $(a_{20}, a_{21}, a_{22})^\top = (0.30, 0.50, -0.50)^\top$. The remaining $J-2$ eigenvalues are generated from our $N(0.9, 0.1^2)$ at each time point. At each time point t, the true eigenvectors $\boldsymbol{w}_{kt}^0$'s ($k = 1, 2, \ldots, J$) are obtained from the eigenvectors of an exchangeable correlation matrix $\rho\mathbf{1}_{10}^T\mathbf{1}_{10} + (1-\rho)\boldsymbol{I}_{10}$, where ρ is uniformly generated from $[0.2, 0.7]$. Then we generate multivariate outcome $\boldsymbol{y}_{it}$ ($i = 1, \ldots, N$) from a multivariate normal distribution with mean zero and covariance matrix $\boldsymbol{V}_t^0 = \sum_{k=1}^J \alpha_{kt}^0 \boldsymbol{w}_{kt}^0 (\boldsymbol{w}_{kt}^0)^T$. Note that, in this study, we do not incorporate any random effects.

For all three objective functions, we model the time-varying covariance matrix as

$$\overline{\boldsymbol{V}}_t = \sum_{k=1}^{2} \bar{\alpha}_k(t)\tilde{\boldsymbol{w}}_{kt}\tilde{\boldsymbol{w}}_{kt}^T + \sum_{k'=3}^{J} \tilde{\alpha}_{k't}\tilde{\boldsymbol{w}}_{k't}\tilde{\boldsymbol{w}}_{k't}^T,$$

where $\tilde{\boldsymbol{w}}_{kt}$'s ($1 \le k \le J$) are sample eigenvectors, and $\tilde{\alpha}_{k't}$'s ($3 \le k' \le J$) are corresponding sample eigenvalues. Due to the full rank of $\overline{\boldsymbol{V}}_t$, we have

$$\overline{\boldsymbol{V}}_t^{-1} = \sum_{k=1}^{2} \bar{\alpha}_k^{-1}(t)\tilde{\boldsymbol{w}}_{kt}\tilde{\boldsymbol{w}}_{kt}^T + \sum_{k'=3}^{J} \tilde{\alpha}_{k't}^{-1}\tilde{\boldsymbol{w}}_{k't}\tilde{\boldsymbol{w}}_{k't}^T,$$

where $\bar{\alpha}_k^{-1}(t)$'s ($k = 1, 2$) are modeled by a polynomial regression with bases $\big[1, t/T, (t/T)^2\big]$.

The mean absolute deviation of error (MADE) and the mean squared error (MSE) are calculated to evaluate the eigenvalue estimation and the bases coefficients estimation, respectively. To compare the estimation results from different objective functions, we take the ratios of the the MADE values and the MSE values using the GEE-type estimates as the baseline (denominator). Table 2 summarizes the simulation results, which indicate that the GEE-type objective function leads to more accurate estimations regarding both time-varying eigenvalue functions and the bases coefficients, in terms of large ratio values (all greater than one) for the Likelihood-based estimates and the F-norm-based estimates.

Table 2: The ratios of the MADE values (eigenvalue estimation) and the MSE values (bases coefficients estimation) between the estimations from different objective functions, using the value of the GEE-type estimation as the baseline (denominator). The results are summarized based on 50 replications.

	Eigenvalue (MADE ratio)		Coefficients (MSE ratio)					
	First	Second	a_{10}	a_{11}	a_{12}	a_{20}	a_{21}	a_{22}
GEE-type	1	1	1	1	1	1	1	1
Likelihood	1.20	1.22	1.13	1.12	1.07	1.33	1.13	1.14
F-norm	1.29	1.44	1.25	1.15	1.21	1.47	1.33	1.32

2 Model Robustness

2.1 The L_2-penalty

We show the time-varying eigenvalue and eigenvector estimations in the simulation study in Section 4 with different values of orthogonality tuning parameter ϕ, which are $(0, 0.02, 0.03, 0.04, 0.05)$. In particular, noting that $\phi = 0$ actually corresponds to the case without the L_2-penalty. Table 3 provides the MADE values and the MCDE values for the estimated eigenvalues and eigenvectors, respectively. Figure 6 illustrates the estimated smoothing eigenvalue functions over time. The results show that the estimations of the EERE method are not sensitive to the choice of ϕ within a given range.

Table 3: The mean absolute deviation of error (MADE) for eigenvalue estimation and the mean cosine deviation error (MCDE) for eigenvector estimation, for the proposed EERE with different values of ϕ (tuning parameter associated with the L_2-penalty), based on 100 replications.

	Comp	$\phi = 0$	$\phi = 0.02$	$\phi = 0.03$	$\phi = 0.04$	$\phi = 0.05$
Eigenvalue (MADE)	First	0.39	**0.37**	0.39	0.40	0.40
	Second	0.36	**0.34**	0.36	0.36	0.37
Eigenvector (MCDE)	First	0.0356	**0.0318**	0.0325	0.0332	0.0365
	Second	0.0217	**0.0110**	0.0164	0.0162	0.0177

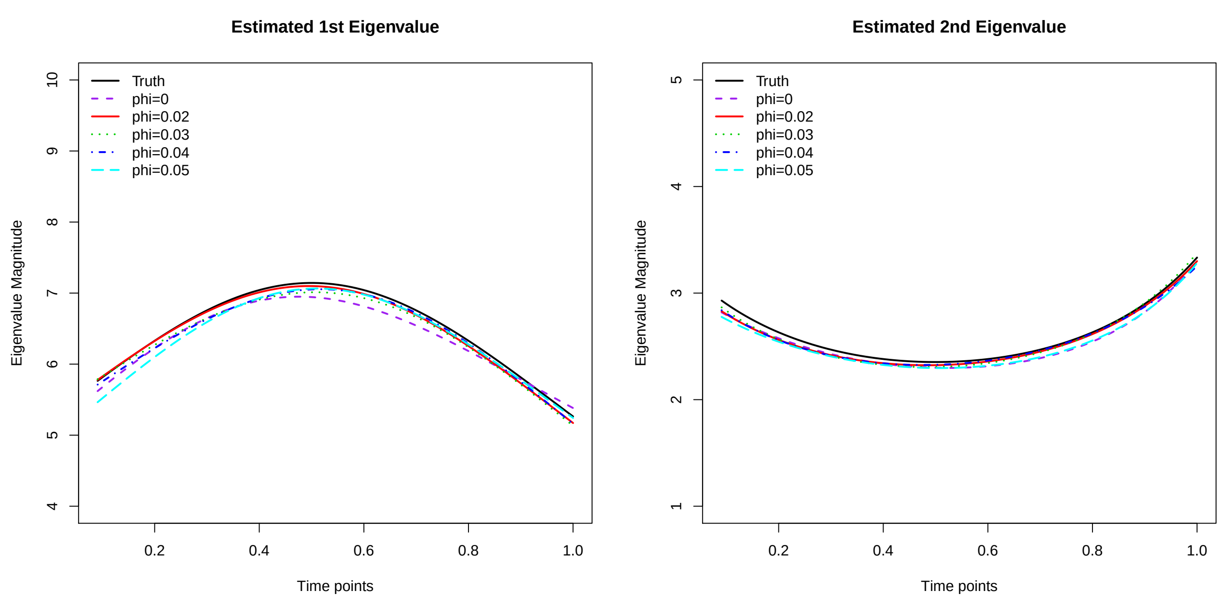

Figure 6: The estimated eigenvalue curves of the EERE model with different values of ϕ under the simulation setting specified in Section 4.

2.2 Knots Selection

In this simulation study, we show the estimation results of the EERE method with respect to four different numbers of knots in the spline construction: $P_{N1} = 10$, $P_{N2} = 7$, $P_{N3} = 5$, and $P_{N4} = 4$. The knots are evenly spaced in all cases. Figure 7 and 8 provide the heatmaps of the estimated first and second eigenvectors, respectively, which indicate that the EERE method is quite robust to different choices of knots' numbers.

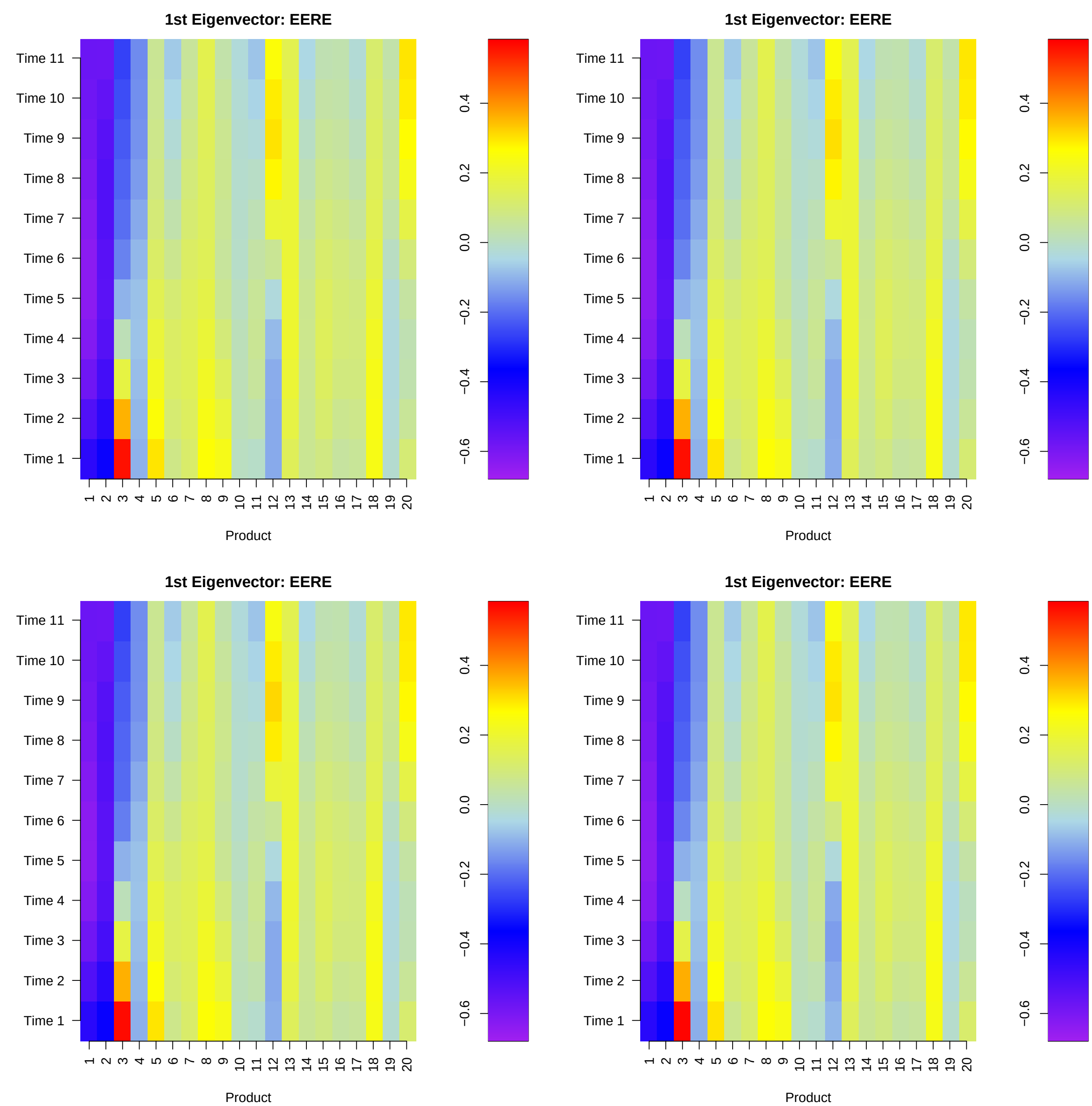

Figure 7: The heatmaps of the estimated first eigenvector from the EERE method with different numbers of knots: top-left($P_{N1} = 10$), top-right($P_{N2} = 7$), bottom-left($P_{N3} = 4$), bottom-right($P_{N4} = 3$).

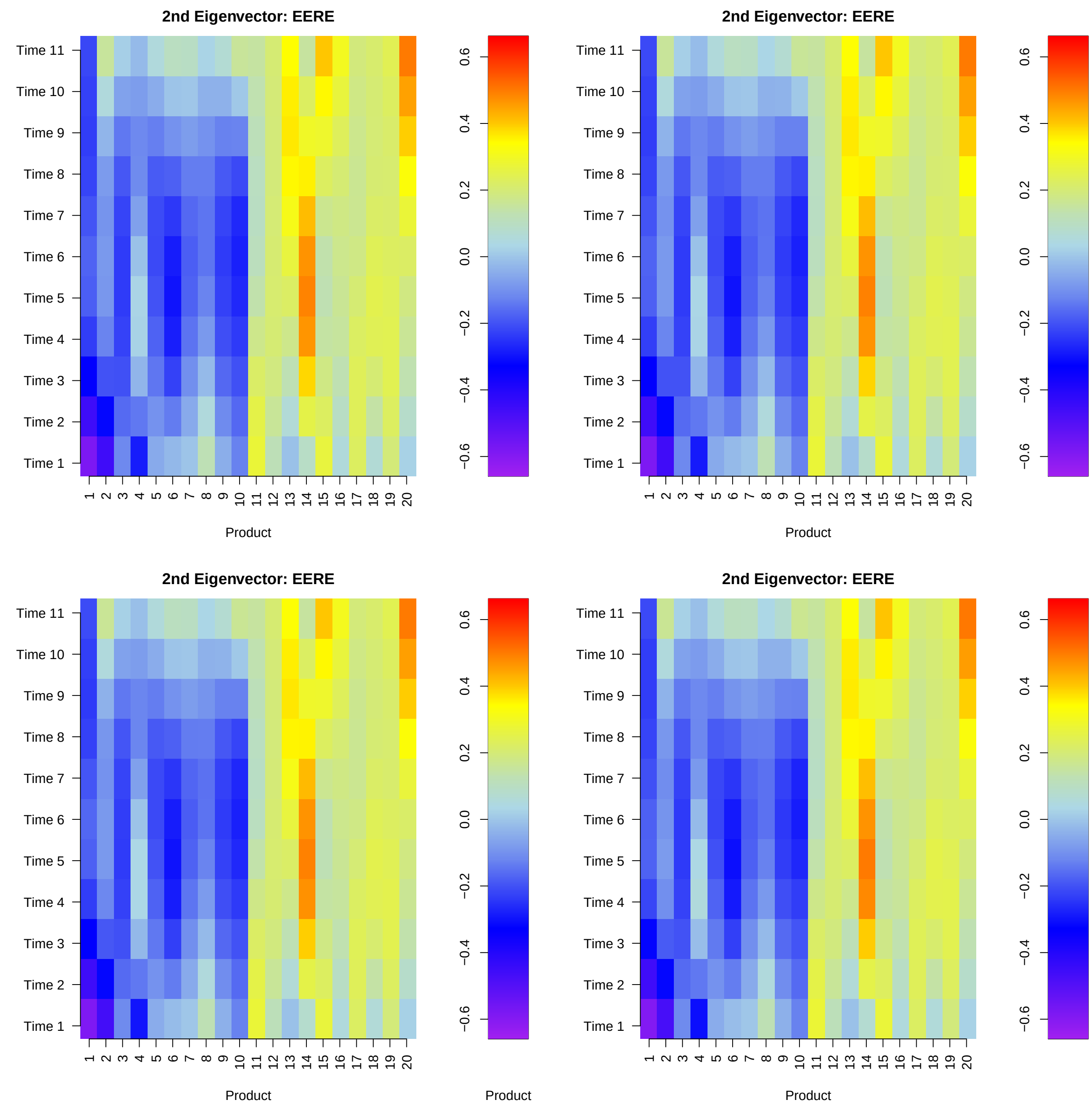

Figure 8: The heatmaps of the average estimated second eigenvector from the EERE method with different numbers of knots: top-left($P_{N1} = 10$), top-right($P_{N2} = 7$), bottom-left($P_{N3} = 4$), bottom-right($P_{N4} = 3$).