LONGITUDINAL PRINCIPAL COMPONENTS ANALYSIS FOR BINARY AND CONTINUOUS DATA

BY

CHRISTOPHER KINSON

DISSERTATION

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Statistics
in the Graduate College of the
University of Illinois at Urbana-Champaign, 2017

Urbana, Illinois

Doctoral Committee:

Professor Annie Qu, Chair
Professor Steven Culpepper
Professor John Marden
Professor Douglas Simpson

Abstract

Large-scale data or big data is an enormously popular word in the data science and statistics communities. These datasets are often collected over periods of time - at hourly and weekly rates - with the help of technological advancements in physical and cloud-based storage. The information stored is useful, especially in biomedicine, insurance, and retail, where patients and customers are crucial to business survival. In this thesis, we develop new statistical methodologies for handling two types of datasets: continuous data and binary data.

Time-varying associations among store products provide important information to capture changes in consumer shopping behavior. In the first part of this thesis, we propose a longitudinal principal component analysis (LPCA) using a random-effects eigen-decomposition, where the eigen-decomposition utilizes longitudinal information over time to model time-varying eigenvalues and eigenvectors of the corresponding covariance matrices. Our method can effectively analyze large marketing data containing sales information for selected consumer products from hundreds of stores over an 11-year time period. The proposed method leads to more accurate estimation and interpretation compared to comparable approaches, which is illustrated through finite sample simulations. We show our method's capabilities and provide an interpretation of the eigenvector estimates in an application to IRI marketing data.

In the second part of this thesis, we formulate the LPCA problem for binary data. We propose capturing the associations among the products or variables through the odds ratios, where a 2×2 contingency table contains probabilities representing the joint distribution of two binary products. The eigen-decomposition utilizes longitudinal information over time to model time-varying eigenvalues and eigenvectors of the corresponding odds ratio matrices. These odds ratio matrices measure the pairwise associations among the binary products and is more appropriate to use than the Pearson correlation coefficient. Our method illustrates an improvement in visualization and interpretation through simulation studies and an application to IRI panel data of individual customer

purchases.

Dedicated to my entire family,

for their nurturing and love,

and for being the examples in my life.

Acknowledgements

I would like to thank my advisor Professor Annie Qu for her belief in me upon arriving on campus in the summer of 2011. Her patience and knowledge guided me every step of the way. In innumerable ways, she has helped improve my understanding of statistics as a profession. I do not think I could have reached this milestone in my academic career without her.

I would also like to thank my committee members Professor Steven Culpepper, Professor John Marden and Professor Douglas Simpson for their support in my career development and suggestions for my research. I am very thankful to my collaborator Xiwei Tang for his insight and enthusiasm. He believed in this research direction at times when I struggled, and he helped me grow as a practitioner of statistics.

My thanks are given to the faculty and staff members in the Department of Statistics for providing a friendly environment. I am also grateful to my relatively large cohort of colleagues who entered in this PhD journey with me in 2011. I am very proud and elated to see them enter the next phases of their careers. I will carry that first-year spirit with me always.

I would like to thank Information Resources, Incorporated (IRI) and Mike Kruger for making the data available and for his generosity on discussing details and updates of the data set. All estimates and analyses in this paper based on Information Resources, Inc. (IRI) data are by the author(s) and not by Information Resources, Inc.

Finally, I want to thank my parents and my closest friends for their unwavering encouragement and heartfelt conversations. To my wife, I look forward to every moment with you. Thank you for being exactly who you are.

Contents

Chapter 1

Introduction

Principal component analysis (PCA) is a widely used technique for data exploration, analysis, and as a pre-processing tool in machine learning. PCA is quite useful for dimension reduction and regression, and has been applied in a broad range of fields including sports, biomedicine, marketing, imaging, and cloud computing.

Longitudinal data analysis is a field of statistics that is crucial to the development of new methodologies in clinical trials, biomedicine, and observational studies. Recently, the data applications within longitudinal data analysis have expanded to include more technological and non-medical sources. In longitudinal data, often observations are considered as repeated measurements for clusters of subjects. Due to the repeated nature, the outcomes exhibit association as correlation. This correlation presents the main challenge in statistical papers. However, developing PCA for longitudinal approaches is still gaining attraction.

In this thesis, we offer a longitudinal methodology that tackles a correlation problem through estimating the eigenvalues and eigenvectors as time-varying functions. We propose a longitudinal PCA for continuous data in Chapter 2 and for binary data in Chapter 3. In the continuous data case, we capture association through the covariance, whereas the association for the binary data case is captured through the odds ratio. In Chapter 2, we incorporate random effects into the estimation methodology, while Chapter 3 ignores the random effects. For both chapters, we estimate time-varying eigenvalues and eigenvectors as smoothing splines. We adopt a multivariate Newton-Raphson algorithm to estimate the eigenvectors, while the eigenvalues are estimated with an explicit solution from an objective function based on generalized estimating equations. The proposed methodology shows improvements over a comparable approach for both the continuous and binary data settings. The improvements in the numerical studies are substantial for both eigenvalue and eigenvector estimations. For the real data application, the proposed methodology offers a refined interpretation for the eigenvectors.

Chapter 2

Longitudinal Principal Components for Continuous Data

2.1 Introduction

Life-changing events can occur at any time, within decades or much shorter time periods. For example, technological innovation or stagnation could affect the fates of companies and their employees. In the first quarter of 2012, Kodak, a leading technological company well-known for producing photography supplies such as cameras and films, filed for bankruptcy and ceased production of digital cameras and several photography accessories. Therefore it is necessary and important to study consumers shopping behavior and marketing trends over time to help companies to avoid manufacturing products which are no longer attractive to consumers.

Principal component analysis (PCA) is often employed to handle large-dimensional multivariate data with correlations. One particular application of the PCA is in business and marketing, where sources of variation have been studied extensively (e.g., Jain et al., 1990; Fader & Lattin, 1993; Rossi & Allenby, 1993; Bradlow, 2002). In addition, the PCA is a powerful tool to provide dimension reduction in time series analysis (e.g., Brillinger, 1981; Ku et al., 1995; Peña & Yohai, 2016). However, the PCA has been adapted for longitudinal data primarily in functional data analysis, where observations are viewed as smooth functions. The functional PCA (Ramsay & Silverman, 2005; Yao et al., 2005; Hall et al., 2006) applies covariance functions at two time points for a single variable, where covariance operators are decomposed into eigenvalues and eigenfunctions. The eigenfunctions could change over a time domain. One extension of the functional PCA is to estimate principal components for observations measured over two hierarchical or longitudinal domains, such as two different levels of time (Di et al., 2009). Greven et al. (2010) make a further extension of Di et al.'s (2009) approach and propose the longitudinal functional PCA which allows for more levels of time and functional random effects to incorporate longitudinal correlation. In addition, Jiang & Wang (2010) extend the functional PCA by considering additional covariates, where a new covariance function depends on time and covariates.

The functional PCA approaches assume that only eigenfunctions change over time, while the eigenvalues remain fixed. Additionally, the covariance function is calculated for a single variable or two variables at pairs of time points. These restrictions could be limiting in our marketing data application, since our target is to capture the associations of multiple variables across all time points. For example, the associations among several products sold at a grocery store over a period of 11 years show heterogeneous variation on sales volumes from different stores. For these reasons, a new longitudinal PCA approach needs to be developed.

This paper is motivated by the IRI marketing data set, an immense collection of consumer panel data of grocery and drug store sales, pricing, and promotion strategies (Bronnenberg et al., 2008; Kruger & Pagni, 2008). The IRI data were created for marketing researchers to explore marketing trends and their impact on economics. The sales data contain weekly sales information of more than 30 product categories in over 40 regional markets in the United States.

We propose a longitudinal PCA that decomposes correlation information arising from multivariate observations over time while incorporating heterogeneity among subjects. Specifically, subject-specific random effects are implemented to capture heterogeneous variation among stores, while correlations among product sales are modeled through time-varying eigen-decomposition. In modeling the eigen-decomposition, we assume that both eigenvalues and eigenvectors are time-varying functions. The proposed method estimates these time-varying functions based on nonparametric splines, which provides more accurate estimation and interpretation of the time-varying eigenvectors and eigenvalues.

The proposed method has three advantages over existing approaches. First, incorporating random effects helps to explain the variation among different sizes of stores, thus improving the estimation of eigenvalues and eigenvectors. Moreover, the random effects can account for the variation of different products in sales from different stores. Second, modeling eigenvalues as functions of time is more sensible when the eigenvalues are likely to change over time. Additionally, modeling eigenvectors as functions of time utilizes the mechanics of functional PCA and the interpretation of PCA. In standard PCA, the eigenvectors are used to describe grouping behaviors from different product sales. In the proposed method, the time-varying eigenvectors provide longitudinal interpretation of grouping behavior of product sales over time. Third, the time-varying eigenvectors can be estimated using nonparametric splines, which are able to recover information from missing time points by utilizing neighboring data points.

To implement the proposed method, we develop an iterative algorithm which simultaneously estimates time-varying eigenvalues and time-varying eigenvectors. This algorithm resolves the identifiability issue for the eigenvectors with the use of the Gram-Schmidt orthonormalization process. In addition, we implement an Estimation-Substitution algorithm to incorporate the random-effects estimation.

2.2 Model Framework and Methodology

2.2.1 Background and Notation

Let y_{ijt} be a response measured over a time variable t for the j-th observation from the cluster i where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$, and t is in the range of $[0, 1]$. In our application, t represents the number of years, where the first recorded year of sales is 2001, and t is rescaled by the total $T = 11$ years to fall into the range of $[0, 1]$. Also in our application, i represents the i-th store, and j is the j-th product. For simplicity, a bold lowercase represents a vector, e.g., $\boldsymbol{y}_{it}$ is a vector of length J, and a bold uppercase represents a matrix, e.g., $\boldsymbol{X}$.

We provide the framework of the principal component analysis (PCA) and introduce discretized PCA in a longitudinal setting. Let $\boldsymbol{x}$ be a vector of J random variables where the covariance matrix of $\boldsymbol{x}$ is $\boldsymbol{\Sigma}$. The goal of the PCA is to maximize the overall variation of data, $\boldsymbol{X}$, using linear combinations of the observations, referred to as principal components. Let $\boldsymbol{e}_k$ be a vector of length J, and the objective function associated with the PCA be

$$\max_{\boldsymbol{e}_k} \frac{\boldsymbol{e}_k^\top \boldsymbol{\Sigma} \boldsymbol{e}_k}{\boldsymbol{e}_k^\top \boldsymbol{e}_k}, \quad k = 1, \ldots, J,$$

$$\text{s.t. } \boldsymbol{e}_k^\top \boldsymbol{e}_k = 1 \text{ and } \boldsymbol{e}_i^\top \boldsymbol{e}_k = 0, \text{ for } i = 1, \ldots, k-1.$$

The principal components are the linear combinations $\boldsymbol{e}_k^\top \boldsymbol{x}$, where the variance of the k-th principal component is $\lambda_k = \boldsymbol{e}_k^\top \boldsymbol{\Sigma} \boldsymbol{e}_k$. Following spectral decomposition, the k-th principal component's loading vectors are the eigenvectors, $\boldsymbol{e}_k$, and the corresponding variances are the eigenvalues, λ_k. The covariance matrix has the equivalent form

$$\boldsymbol{\Sigma} = \sum_{k=1}^{J} \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k^\top.$$

After maximizing the variation of the data, one often seeks to reduce the dimensionality of data. Therefore, the PCA can be used to approximate the true covariance of $\boldsymbol{x}$ by choosing a smaller number of components $K \leq J$, where the approximate matrix $\hat{\boldsymbol{\Sigma}}$ is

$$\hat{\boldsymbol{\Sigma}} = \sum_{k=1}^{K} \lambda_k \boldsymbol{e}_k \boldsymbol{e}_k^\top.$$

In the longitudinal data setting, we can perform the PCA by decomposing the covariance matrix at each time separately. We refer to this as discretized principal component analysis (DPCA). The objective function of DPCA is similar to the standard eigen-decomposition of a covariance matrix, but allows eigen-terms to be functions of time t. Let $\boldsymbol{e}_{kt}$ be a vector of length J at fixed time t, then the objective function of the DPCA can be expressed as

$$\max_{\boldsymbol{e}_{kt}} \frac{\boldsymbol{e}_{kt}^\top \boldsymbol{\Sigma}_t \boldsymbol{e}_{kt}}{\boldsymbol{e}_{kt}^\top \boldsymbol{e}_{kt}}, \quad k = 1, \ldots, J,$$

$$\text{s.t. } \boldsymbol{e}_{kt}^\top \boldsymbol{e}_{kt} = 1 \text{ and } \boldsymbol{e}_{it}^\top \boldsymbol{e}_{kt} = 0, \text{ for } i = 1, \dots, k-1.$$

Maximizing the above objective function yields the k-th principal component at fixed time t, $\boldsymbol{e}_{kt}^\top \boldsymbol{x}_t$, and the variance of the k-th principal component at fixed time t, $\lambda_{kt} = \boldsymbol{e}_{kt}^\top \boldsymbol{\Sigma}_t \boldsymbol{e}_{kt}$, where $\boldsymbol{x}_t$ is a vector with covariance $\boldsymbol{\Sigma}_t$. The approximate covariance matrix at time t is represented as

$$\hat{\boldsymbol{\Sigma}}_t = \sum_{k=1}^{K} \lambda_{kt} \boldsymbol{e}_{kt} \boldsymbol{e}_{kt}^\top, \tag{2.1}$$

where $K \leq J$.

2.2.2 Time-varying models

In this section, we represent eigenvalues and eigenvectors as continuous functions of t instead of discrete functions as in the DPCA. We assume that the time-varying eigenvalues, $\alpha(t)$, and time-varying eigenvectors $\boldsymbol{e}(t)$ can be approximated by polynomial splines, and that the time variable is in the range $[0, 1]$.

Nonparametric splines have been extensively studied for independent and longitudinal data by Anderson & Jones (1995), Huang et al. (2004), Durbán et al. (2005), Liang & Xiao (2006), and Xue & Liang (2009). Additionally, semiparametric and nonparametric approaches have been proposed for covariance and correlation matrix estimation for longitudinal data (e.g., Diggle & Verbyla, 1998; Wu & Pourahmadi, 2003; Fan et al., 2007; Sun et al., 2007; Fan & Wu, 2008; Maadooliat et al., 2013).

Let ξ be a partition of the interval $[0, 1]$ with P_N interior knots

$$\xi = \{0 = \xi_0 < \xi_1 < \cdots < \xi_{P_N} < \xi_{P_N+1} = 1\}.$$

Consequently, using ξ as knots, the polynomial splines of orders M_1 and M_2 for eigenvalues and eigenvectors are at most, $M_1 - 1$ and $M_2 - 1$ degrees of polynomial functions on intervals $[\xi_i, \xi_{i+1})$, for $i = 0, \dots, P_N - 1$, and $[\xi_{P_N}, \xi_{P_N+1}]$; and the spline functions are $M_1 - 2$ and $M_2 - 2$ continuously differentiable. Note that different sets of knots are allowed for eigenvalues and eigenvectors; however, for simplicity, we choose the same set of knots here. Then the k-th eigenvalue and the corresponding eigenvector can be modeled as

$$\alpha_k(t) \approx \sum_{l=1}^{C_N} \beta_{kl} b_{kl}(t) \, , \, e_{kj}(t) \approx \sum_{m=1}^{D_N} \nu_{kjm} g_{kjm}(t), \tag{2.2}$$

where $e_{kj}(t)$ is the j-th product's loading in the k-th eigenvector, $\{b_{kl}(t)\}_{l=1}^{C_N}$ and $\{g_{kjm}(t)\}_{m=1}^{D_N}$ are sets of spline bases with $C_N = P_N + M_1$ and $D_N = P_N + M_2$; and $\boldsymbol{\beta}_k = (\beta_{k1}, \ldots, \beta_{kC_N})^\top$ and $\boldsymbol{\nu}_{kj} = (\nu_{kj1}, \ldots, \nu_{kjD_N})^\top$ are sets of B-spline coefficients. The construction of the basis functions depends on the order of the splines and the number of knots which can be preselected by the user or selected through the Akaike Information Criterion (AIC) (Akaike, 1974) or Bayesian Information Criterion (BIC) (Schwarz, 1978). One attractive feature of polynomial splines is that they approximate a smooth function sufficiently well without requiring a large number of knots.

2.2.3 Estimation of Eigenvalues and Eigenvectors

In this section, we provide time-varying eigenvalue and eigenvector estimations under the framework of generalized estimating equations (GEE). We denote the true variance of the response as $\boldsymbol{V}_t^0 = \text{Var}(\boldsymbol{y}_{it})$, and the approximated variance matrix based on the first K eigenvalues and eigenvectors as

$$\boldsymbol{V}_t = \sum_{k=1}^{K} \alpha_k(t) \boldsymbol{e}_k(t) \boldsymbol{e}_k(t)^\top, \tag{2.3}$$

where $\alpha_k(t)$ is the k-th eigenvalue corresponding to $\boldsymbol{e}_k(t)$, the k-th eigenvector over time. The difference of the inverse of the true covariance matrix and the inverse of its approximation under the GEE framework is

$$\boldsymbol{h}_{it} = \left(\boldsymbol{V}_t^{-1} - \boldsymbol{V}_t^{0^{-1}} \right) (\boldsymbol{y}_{it} - \boldsymbol{\mu}_t), \tag{2.4}$$

where $\boldsymbol{\mu}_t$ represents the overall mean vector of responses at each time point. We include the residual term $(\boldsymbol{y}_{it} - \boldsymbol{\mu}_t)$ as a multiplier in (**??**) since it simplifies the objective function in the optimization process.

Motivated by the generalized method of moments (GMM), we minimize the following objective function

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\boldsymbol{h}_{it}^{\top}\boldsymbol{h}_{it}}{N} = \sum_{t=1}^{T}\frac{1}{N}\sum_{t=1}^{N}(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t)^{\top}(\boldsymbol{V}_t^{-1}-\tilde{\boldsymbol{V}}_t^{-1})^{\top}(\boldsymbol{V}_t^{-1}-\tilde{\boldsymbol{V}}_t^{-1})(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t). \tag{2.5}$$

Note that the true variance $\boldsymbol{V}_t^0$ is unknown and thus replaced with the sample covariance matrix $\tilde{\boldsymbol{V}}_t$. We minimize the following objective function with respect to both $\alpha_k(t)$ and $\boldsymbol{e}_k(t)$ simultaneously-

$$\sum_{t=1}^{T}\left(\sum_{i=1}^{N}\frac{\boldsymbol{h}_{it}^{\top}\boldsymbol{h}_{it}}{N} + \phi\sum_{i\neq j}\|\boldsymbol{e}_i(t)^{\top}\boldsymbol{e}_j(t)\|_2^2\right), \tag{2.6}$$

where ϕ is the tuning parameter for the eigenvector penalty term, where the penalty encourages orthogonality between any pairs of eigenvectors.

Using the Law of Large Numbers, (2.5) can be approximated by the expectation and trace operations as follows:

$$\begin{array}{cl}\sum_{t=1}^{T} & \mathbb{E}\left[\mathrm{Tr}\left\{(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t)^{\top}(\boldsymbol{V}_t^{-1}-\tilde{\boldsymbol{V}}_t^{-1})^2(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t)\right\}\right] \\ = & \sum_{t=1}^{T}\mathrm{Tr}\left[\mathbb{E}\left\{(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t)(\boldsymbol{y}_{it}-\boldsymbol{\mu}_t)^{\top}(\boldsymbol{V}_t^{-1}-\tilde{\boldsymbol{V}}_t^{-1})^2\right\}\right] \\ \approx & \sum_{t=1}^{T}\mathrm{Tr}\left\{\boldsymbol{V}_t^0(\boldsymbol{V}_t^{-1}-\tilde{\boldsymbol{V}}_t^{-1})^2\right\},\end{array}$$

where $\mathbb{E}$ indicates the expectation, and $\mathrm{Tr}\{\cdot\}$ indicates the trace of a matrix. For the purpose of illustration, we focus on the estimation of the eigenvalues. By using the eigenvectors from the sample covariance matrix, we approximate (2.5) in terms of the eigenvalues as

$$\sum_{t=1}^{T}\sum_{i=1}^{N}\frac{\boldsymbol{h}_{it}^{\top}\boldsymbol{h}_{it}}{N} \approx \sum_{t=1}^{T}\sum_{k=1}^{K}\alpha_{kt}^{0}\left(\frac{1}{\alpha_k(t)}-\frac{1}{\tilde{\alpha}_{kt}}\right)^2, \tag{2.7}$$

where α_{kt}^0 is the true eigenvalue, $\tilde{\alpha}_{kt}$ is the sample eigenvalue, and $\alpha_k(t)$ is the estimator of the eigenvalues ($k = 1, \ldots, K; t = 1, \ldots, T$). We can interpret (2.7) as a weighted square difference for the inverse of the true and estimated eigenvalues.

We compare the proposed objective function in (2.7) to those based on the maximum likelihood (ML) and the Frobenius norm. The objective function based on the ML is

$$\sum_{t=1}^{T} \left(\log \det(\boldsymbol{V}_t) + \text{Tr}\left\{ \boldsymbol{V}_t^{-1} \tilde{\boldsymbol{V}}_t \right\} \right),$$

where $\log \det(\cdot)$ indicates the logarithm of the determinant of a matrix. The objective function based on the Frobenius norm of the difference between the true covariance and estimated covariance matrix is

$$\sum_{t=1}^{T} \left(\text{Tr}\left\{ (\boldsymbol{V}_t - \tilde{\boldsymbol{V}}_t)(\boldsymbol{V}_t - \tilde{\boldsymbol{V}}_t)^\top \right\} \right)^{1/2}.$$

Following a similar technique of expectation and trace operations as above, we can transform the objective function of the ML to be

$$\sum_{t=1}^{T} \sum_{k=1}^{K} \left(\frac{\tilde{\alpha}_{kt}}{\alpha_k(t)} + \log(\alpha_k(t)) \right),$$

and the objective function corresponding to the Frobenius norm to be

$$\sum_{t=1}^{T} \sum_{k=1}^{K} \left(\alpha_k(t) - \tilde{\alpha}_{kt} \right)^2.$$

From our unreported simulation, the proposed norm in (2.7) produces smaller mean square errors of the estimations compared to the ones produced by the ML or Frobenius norm approaches, especially on larger eigenvalues. In addition, the convergence rate of the algorithm using (2.7) is faster than that of the ML and Frobenius.

2.2.4 Incorporation of Random Effects

Random-effects modeling (Laird & Ware, 1982; Gardiner et al., 2009) are quite useful to accommodate subject-specific effects from individuals. For the IRI marketing data application, the volumes of sales show strong variations from different types of stores and products, as mentioned in Section 2.1. Therefore, it is important to incorporate heterogeneity of the stores in our modeling.

We propose the estimated eigenanalysis with random effects (EERE) model, which assumes that the observed response can be decomposed into two parts at time t, the true response and the unobserved random effects. That is, $\boldsymbol{y}_{it}^{obs} = \boldsymbol{y}_{it} + \boldsymbol{\gamma}_i$, where $\boldsymbol{\gamma}_i$ represents a subject-specific random effect with normal distribution, $N(\mathbf{0}, \boldsymbol{D})$, and $\boldsymbol{D}$ is the variance component for the random effect $\boldsymbol{\gamma}_i$. The corresponding variance of $\boldsymbol{y}_{it}^{obs}$ is

$$\boldsymbol{V}_t^{obs} = \text{Var}(\boldsymbol{y}_{it}) + \text{Var}(\boldsymbol{\gamma}_i). \tag{2.8}$$

The advantage of incorporating random effects in the proposed EERE is that it is able to remove the heterogeneity among different stores and products, and therefore provides more accurate covariance estimation. Consequently, this leads to more accurate estimations of eigenvalues and eigenvectors. In contrast, if the true response is $\boldsymbol{y}_{it} + \boldsymbol{\gamma}_i$, then a variance of only Var$(\boldsymbol{y}_{it})$ produces a biased estimator of the covariance matrix as it ignores the covariance of the random effects $\boldsymbol{\gamma}_i$. Similarly, the DPCA in (2.1) without incorporating the random effects also produces biased covariances. Therefore, the EERE model is more effective if the heterogeneous information from the stores or products is present.

To simplify the notation, we suppress the notation on time t, and assume that the random effect $\boldsymbol{\gamma}_i$ follows a multivariate normal distribution $N(\mathbf{0}, \boldsymbol{D})$, where $\boldsymbol{\gamma}_i$ is a J-dimensional random-effect vector from the i-th store, J corresponds to the number of products, and $\boldsymbol{D}$ is the covariance matrix of the random effects. We employ an Estimation-Substitution (ES) algorithm (Elashoff & Ryan, 2004; Xu et al., 2012) by initializing the random effects $\boldsymbol{\gamma}_i$, the covariance matrix $\boldsymbol{D}$, the overall mean $\boldsymbol{\mu}_t$, and the $J \times J$ residual covariance $\boldsymbol{R}_t$. We then estimate the random effects by

$$\hat{\boldsymbol{\gamma}} = \boldsymbol{Z}\big((\mathbf{1}_T\mathbf{1}_T^\top) \otimes \boldsymbol{D} + \boldsymbol{R}\big)^{-1}\big(\mathbf{1}_T^\top \otimes \boldsymbol{D}\big)^\top, \tag{2.9}$$

where $\boldsymbol{Z} = \Bigg(\boldsymbol{y}_1^{obs} - (\mathbf{1}_N \otimes \boldsymbol{\mu}_1^\top), \boldsymbol{y}_2^{obs} - (\mathbf{1}_N \otimes \boldsymbol{\mu}_2^\top), \cdots, \boldsymbol{y}_T^{obs} - (\mathbf{1}_N \otimes \boldsymbol{\mu}_T^\top)\Bigg)$ is an $N \times JT$ matrix, $\boldsymbol{y}_t^{obs}$ is an $N \times J$ matrix of observations at each time point, $\boldsymbol{R} = \text{diag}(\boldsymbol{R}_1, \boldsymbol{R}_2, \cdots, \boldsymbol{R}_T)$ is a $JT \times JT$ block diagonal matrix, and $\mathbf{1}_T$ is a T-dimensional vector of ones for $t = 1, \ldots, T$. Here the random-effect estimation depends on the covariance of random effects, the mean of the products, and the residual covariance. This requires one to estimate each of these components iteratively until their convergence. The covariance $\boldsymbol{D}$ of random effects, the mean of the products, and the residual covariance $\boldsymbol{R}_t$ can be estimated iteratively via

$$\hat{\boldsymbol{D}}^{(l)} = \hat{\boldsymbol{D}}^{(l-1)} - \big(\mathbf{1}_T^\top \otimes \hat{\boldsymbol{D}}^{(l-1)}\big)\big((\mathbf{1}_T\mathbf{1}_T^\top) \otimes \hat{\boldsymbol{D}}^{(l-1)} + \hat{\boldsymbol{R}}^{(l-1)}\big)^{-1}\big(\mathbf{1}_T^\top \otimes \hat{\boldsymbol{D}}^{(l-1)}\big)^\top;$$

$\hat{\boldsymbol{\mu}}_t^{(l)} = \frac{1}{N}\sum_{i=1}^N(\boldsymbol{y}_{it}^{obs} - \hat{\boldsymbol{\gamma}}_i^{(l-1)})$; and $\hat{\boldsymbol{R}}_t^{(l)} = \frac{1}{N}\sum_{i=1}^N(\boldsymbol{y}_{it}^{obs} - \hat{\boldsymbol{\gamma}}_i^{(l-1)} - \hat{\boldsymbol{\mu}}_t^{(l-1)})(\boldsymbol{y}_{it}^{obs} - \hat{\boldsymbol{\gamma}}_i^{(l-1)} - \hat{\boldsymbol{\mu}}_t^{(l-1)})^T$, respectively.

Once the random effects are estimated through the iteration steps in (2.9), we redefine the objective function in (2.6) by replacing $\boldsymbol{V}_t^0$ with $\widetilde{\boldsymbol{V}}_t^* = \text{Var}(\boldsymbol{y}_{it}^{obs} - \hat{\gamma}_i)$, and define

$$\boldsymbol{h}_{it}^* = \Bigg(\boldsymbol{V}_t^{-1} - \widetilde{\boldsymbol{V}}_t^{*-1}\Bigg)(\boldsymbol{y}_{it}^{obs} - \boldsymbol{\mu}_t - \hat{\gamma}_i).$$

That is, we minimize the following objective function

$$\sum_{t=1}^T \Bigg(\sum_{i=1}^N \frac{\boldsymbol{h}_{it}^{*\top}\boldsymbol{h}_{it}^*}{N} + \phi\sum_{i \neq j} \|\boldsymbol{e}_i(t)^\top \boldsymbol{e}_k(t)\|_2^2\Bigg), \tag{2.10}$$

with respect to the eigenvalues $\alpha_k(t)$ and the eigenvectors $\boldsymbol{e}_k(t)$ simultaneously using a Newton-Raphson algorithm, as discussed in the following section.

2.3 Implementation

In this section, we provide the algorithm of the EERE model (2.8) which iterates through the Newton-Raphson and the Estimation-Substitution algorithms.

```
Algorithm: Estimated Eigenanalysis with Random Effects (EERE)
Step 1: Initialize $\boldsymbol{\gamma}_i^{(0)}$, $\boldsymbol{D}^{(0)}$, $\boldsymbol{\mu}_t^{(0)}$, and $\boldsymbol{R}^{(0)}$ ;
Step 2: Estimate the random effects $\boldsymbol{\gamma}_i^{(l)}$ using the ES algorithm in (2.9) ;
Step 3: Estimate the parameters $\boldsymbol{D}^{(l)}$, $\boldsymbol{\mu}^{(l)}$, and $\boldsymbol{R}^{(l)}$ at the $l$-th step;
Step 4: Repeat Steps 2-3 until $\|\boldsymbol{\gamma}_{\boldsymbol{i}}^{(l)} - \boldsymbol{\gamma}_{\boldsymbol{i}}^{(l-1)}\| < \epsilon_\gamma$, where $\epsilon_\gamma$ is a chosen tolerance
level.
Step 5: Set the initial values of the eigenvectors as the sample eigenvectors:
$\boldsymbol{e}_k(t)^{(0)} = \tilde{\boldsymbol{e}}_k(t)$;
Step 6: Given the current eigenvectors $\boldsymbol{e}_k(t)^{(m-1)}$,
(i) update the eigenvalues $\alpha_k(t)^{(m)}$ by minimizing the objective function in (2.10), and
(ii) update $\boldsymbol{e}_k(t)^{(m)}$ given $\alpha_k^{(m)}$ using the Newton-Raphson algorithm;
Step 7: Iterate Step 6 if
(i) $\|\alpha_k(t)^{(m)} - \alpha_k(t)^{(m-1)}\| > \epsilon_\alpha$, where $\epsilon_\alpha$ is a chosen tolerance level, or
(ii)$\|(\boldsymbol{e}_k(t)^{(m)})(\boldsymbol{e}_k(t)^{(m)})^\top - (\boldsymbol{e}_k(t)^{(m-1)})(\boldsymbol{e}_k(t)^{(m-1)})^\top\| > \epsilon_e$ , where $\epsilon_e$ is a chosen
tolerance level.
```

The Newton-Raphson algorithm for the eigenvectors estimation in Step 6 is multivariate. Specifically, at the (m)-th iteration, we update

$$\boldsymbol{e}(t)^{(m)} = \boldsymbol{e}(t)^{(m-1)} - \boldsymbol{J}_f^{-1}\Big(\boldsymbol{e}(t)^{(m-1)}\Big)\boldsymbol{f}\Big(\boldsymbol{e}(t)^{(m-1)}\Big),$$

where $\boldsymbol{e}(t)^{(m-1)} = (\boldsymbol{e}_1(t)^{(m-1)^\top}, \ldots, \boldsymbol{e}_K(t)^{(m-1)^\top})^\top$, K is the number of principal components, $\boldsymbol{f}(\cdot)$ is a vector of the first derivatives of the objective function in (2.10), and $\boldsymbol{J}_f(\cdot)$ is the Jacobian matrix of the second derivatives of (2.10). In detail, $\boldsymbol{f}$ and $\boldsymbol{J}_f$ have the following forms:

$$\boldsymbol{f}\Big(\boldsymbol{e}(t)\Big) = \frac{2}{N}\sum_{i=1}^{N} \dot{\boldsymbol{h}}_{it}^\top \boldsymbol{h}_{it} + 2\lambda\Big(\sum_{l\neq 1}(\boldsymbol{e}_1(t)^\top \boldsymbol{e}_l(t))\boldsymbol{e}_l(t), \ldots, \sum_{l\neq K}(\boldsymbol{e}_K(t)^\top \boldsymbol{e}_l(t))\boldsymbol{e}_l(t)\Big)^\top$$

and

$$\boldsymbol{J}_f\Big(\boldsymbol{e}(t)\Big) \approx \frac{2}{N}\sum_{i=1}^{N} \dot{\boldsymbol{h}}_{it}^\top \dot{\boldsymbol{h}}_{it} + 2\lambda \begin{pmatrix} \sum\limits_{l\neq 1} \boldsymbol{e}_l(t)\boldsymbol{e}_l(t)^\top & \dots & \boldsymbol{e}_1(t)^\top \boldsymbol{e}_K(t) I + \boldsymbol{e}_1(t)\boldsymbol{e}_K(t)^\top \\ \vdots & \ddots & \vdots \\ & \dots & \sum\limits_{l\neq K} \boldsymbol{e}_l(t)\boldsymbol{e}_l(t)^\top \end{pmatrix},$$

where $\dot{\boldsymbol{h}}_{it} = \frac{\partial \boldsymbol{h}_{it}}{\partial \boldsymbol{e}(t)}$, and $\boldsymbol{I}$ is a $J \times J$ identity matrix. Note that

$$\frac{1}{N}\sum_{i=1}^{N} \ddot{\boldsymbol{h}}_{it}^\top \boldsymbol{h}_{it} \to 0 \quad \text{as} \quad N \to \infty,$$

since $\mathbb{E}[\boldsymbol{h}_{it}] = \boldsymbol{0}$ and where $\ddot{\boldsymbol{h}}_{it} = \frac{\partial \dot{\boldsymbol{h}}_{it}}{\partial \boldsymbol{e}(t)}$.

In Step 6, the Gram-Schmidt process is applied in estimating the eigenvectors to insure their pairwise orthonormality. Namely, the Gram-Schmidt process converts linearly dependent vectors into orthonormal vectors which span the same space of the original vectors. That is, the eigenvector estimator, $\hat{\boldsymbol{e}}_k(t)$, is transformed to an orthogonal vector $\boldsymbol{e}_k^*(t)$:

$$\boldsymbol{e}_k^*(t) = \hat{\boldsymbol{e}}_k(t) - \sum_{l=1}^{k-1} \boldsymbol{e}_l^*(t)\frac{\hat{\boldsymbol{e}}_k(t)^\top \boldsymbol{e}_l^*(t)}{\boldsymbol{e}_l^*(t)^\top \boldsymbol{e}_l^*(t)},$$

where $\boldsymbol{e}_1^*(t) = \hat{\boldsymbol{e}}_1(t)$ is the time-varying first eigenvector estimator. Finally, we transform each $\boldsymbol{e}_k^*(t)$ into an orthonormal vector $\boldsymbol{e}_k^{**}(t)$ as

$$\boldsymbol{e}_k^{**}(t) = \frac{\boldsymbol{e}_k^*(t)}{\Big(\boldsymbol{e}_k^*(t)^\top \boldsymbol{e}_k^*(t)\Big)^{1/2}}.$$

2.3.1 Choosing the number of components

In this section, we provide the criterion for selecting the number of components in the eigendecomposition. Choosing the number of components remains an open problem in the PCA. One popular approach is to create a scree plot of the eigenvalues, in which the number of components is chosen at the elbow or the scree of the line drawn. This strategy is appealing since it provides a visual representation of the majority of variation to be retained from the data. For the IRI marketing data, which is collected longitudinally, analyzing the scree plots for each time point is inefficient, because different time points may result in different numbers of chosen components. Instead, we propose calculating an overall variation by averaging over the T time points.

One such numeric criterion is to choose the percentage of overall variation of the data which determines the number of components to retain; that is, to increase the number of components until the desired cumulative percentage of variation from the data is attained. The cumulative percentage of variance contributed from the first K components is

$$\frac{\sum_{k=1}^{K} \frac{1}{T} \sum_{t=1}^{T} \alpha_k(t)}{\sum_{j=1}^{J} \frac{1}{T} \sum_{t=1}^{T} \alpha_j(t)},$$

where $K \leq J$.

An alternative criterion for choosing the number of components is to implement an adaptation of the maximal eigenvalue ratio criterion proposed by Luo et al. (2009). Using the ratio of the eigenvalues

$$\max_{1 \leq k \leq J} \frac{\alpha_k(t)}{\alpha_{k+1}(t)},$$

where $t = 1, \ldots, T$, we find the largest value of this ratio among the T time points, and retain up to the k-th component corresponding to the maximum value.

2.4 Numerical Study

We evaluate the proposed methodology with simulation studies for three different settings for the eigenvalues and eigenvectors corresponding to the longitudinal PCA. In Section 2.4.1, we let the eigenvectors be fixed over time, while the eigenvalues are time-varying. In Section 2.4.2, we reverse the setting, such that the eigenvalues are fixed, but the eigenvectors are changing over time. In Section 2.4.3, we allow both eigenvalues and eigenvectors to be time-varying. For each of these settings, we perform 500 repeated simulations and compare the EERE with the discretized PCA

(DPCA).

To evaluate the estimation of the time-varying eigenvalues, we examine the plots of the average estimated eigenvalues versus time. Also, we calculate the mean absolute deviation of error (MADE) for the eigenvalues:

$$\text{MADE}_k = \sum_{t=1}^{T} \frac{1}{T} \left| \hat{\alpha}_k(t) - \alpha_{kt}^0 \right| / \text{range}(\alpha_k^0),$$

where α_{kt}^0 is the k-th true eigenvalue, $\hat{\alpha}_k(t)$ is the estimated k-th eigenvalue at time t, and $k = 1, \cdots, K$ representing the k-th component. The range$(\alpha_k^0) = \max(\alpha_{kt}^0) - \min(\alpha_{kt}^0)$, which covers the minimum and maximum among all time points.

We evaluate the time-varying eigenvector estimations through creating heatmaps corresponding to the average estimators of eigenvector loading scores over time. The heatmaps allow us to visualize group changes over time among different variables via color contrasts.

2.4.1 Simulation A: Fixed Eigenvectors, Time-varying Eigenvalues

In this section, we implement the EERE model with time-varying eigenvalues but fixed eigenvectors. We generate the first two ($K = 2$) eigenvalues as follows:

$$\alpha_{1t}^0 = \frac{1}{(a_{10} + a_{11}t + a_{12}t^2)} \text{ and } \alpha_{2t}^0 = \frac{1}{(a_{20} + a_{21}t + a_{22}t^2)}, \tag{2.11}$$

where $(a_{10}, a_{11}, a_{12})^\top = (0.19, 0.20, 0.20)^\top$, $(a_{20}, a_{21}, a_{22})^\top = (0.30, 0.50, -0.50)^\top$, and t is in the range of $[0, 1]$. The remaining $J - 2$ eigenvalues are constants around 1 at each time point.

To generate the fixed eigenvectors $\boldsymbol{e}_k$, we perform an eigen-decomposition on a $J \times J$ correlation matrix $\boldsymbol{W}$, which has diagonal entries $w_{ii} = 1$, and off-diagonal entries $w_{ij} = 0.75$, $i = 1, \cdots, J$ for $i \neq j$.

We generate the random effects $\boldsymbol{\gamma}$ from a multivariate normal distribution with mean $\mathbf{0}$ and a J-dimensional identity matrix $\boldsymbol{I}$ for the covariance matrix. The response vector $\boldsymbol{y}_t$ is generated from a multivariate normal distribution with mean $\mathbf{0}$, and a covariance matrix $\sum_{k=1}^{J} \alpha_{kt}^{0} \boldsymbol{e}_k \boldsymbol{e}_k^{\top}$ at time t.

For the EERE estimation, the eigenvectors in (2.3) are the sample eigenvectors calculated from the eigen-decomposition of the sample covariance matrix $\tilde{\boldsymbol{V}}_t$, hence the penalty term of the objective function in (2.10) can be dropped. We use the cubic spline for modeling the eigenvalues in (2.3) for the EERE model.

Table 2.1 summarizes the average MADEs for the estimators of the first and second time-varying eigenvalues when the number of products is $J = 6$ and 10. It is clear that the EERE method yields smaller averages of the MADEs compared to those obtained from the DPCA method. The improvement of the proposed EERE is more apparent when the number of products increases. Figure 2.1 demonstrates that the proposed estimators of time-varying eigenvalues are closer to the true eigenvalues on average, and there is a larger gap between the true eigenvalues and the eigenvalues estimated by the DPCA method over time. The larger gap is due to the heterogeneous variation of the data which the DPCA is not able to take into account. In contrast, the proposed EERE incorporates the random-effects estimation through integrating information from longitudinal data, which leads to more accurate estimation.

$N = 200$, $J = 6$	1^{st} Eigenvalue	2^{nd} Eigenvalue
EERE	0.29	0.18
DPCA	2.55	0.33
$N = 200$, $J = 10$	1^{st} Eigenvalue	2^{nd} Eigenvalue
EERE	0.28	0.15
DPCA	4.21	0.37

Table 2.1: Average MADEs of the first and second eigenvalues for Simulation A

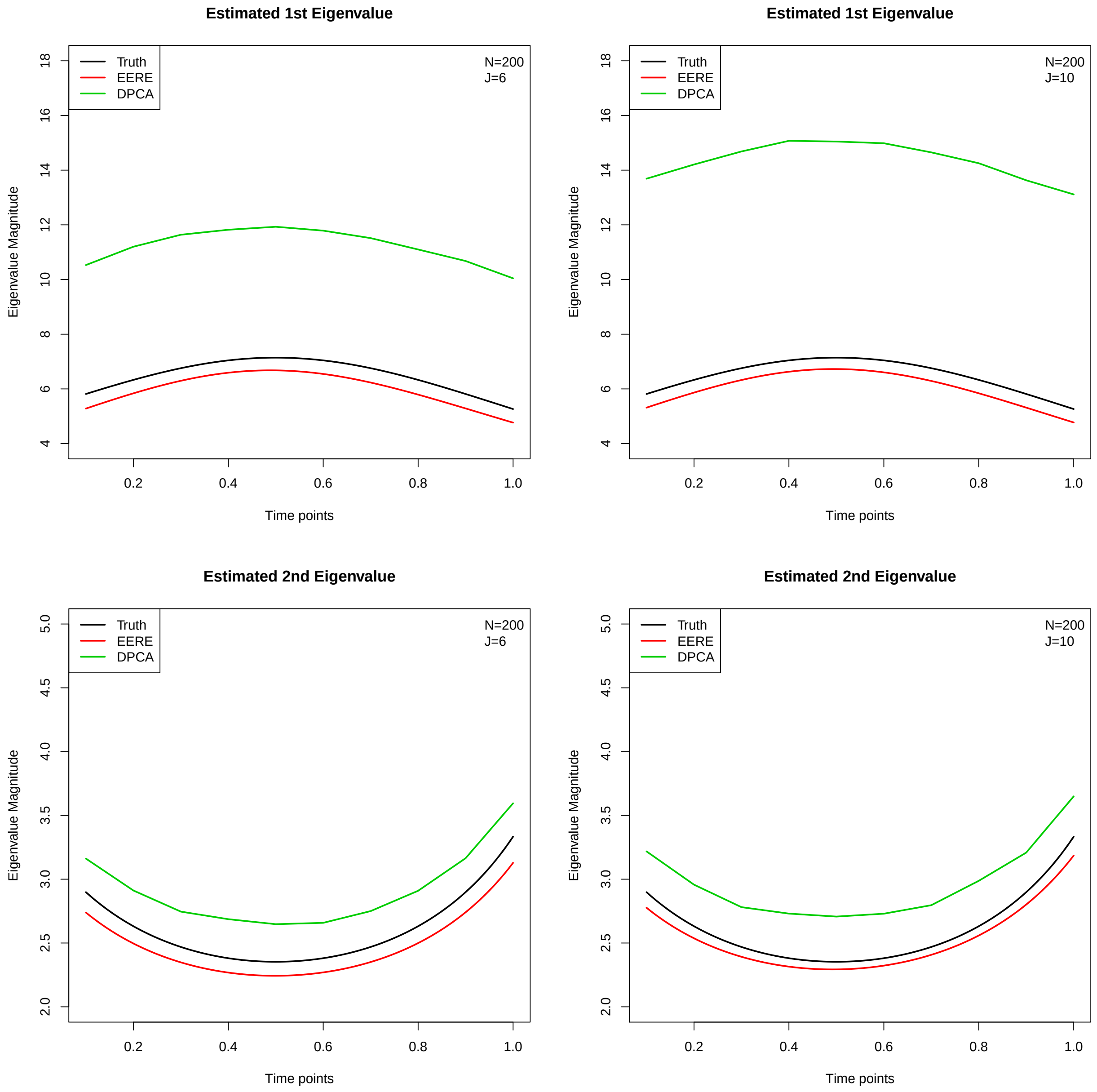


Figure 2.1: Average estimated first and second eigenvalues for Simulation A

2.4.2 Simulation B: Fixed Eigenvalues, Time-varying Eigenvectors

In this section, we employ the EERE model with time-varying eigenvectors while assuming that the eigenvalues are fixed. We generate the time-varying eigenvectors through decomposing a J-

dimensional correlation matrix $\boldsymbol{W_t}$ with

$$\boldsymbol{W}_t = \begin{pmatrix} 1 & & & \\ w_{21t} & 1 & & \\ \vdots & \vdots & \ddots & \\ w_{J1t} & w_{J2t} & \cdots & 1 \end{pmatrix}, \tag{2.12}$$

where w_{ijt} is a time-varying correlation function. We are able to induce time-varying eigenvectors that correspond to the time-varying design of $\boldsymbol{W}_t$. For example, when $J = 6$ products, we define the time-varying correlations in (2.12) by introducing the following index sets, $I^\star = \{1, \cdots, 4\}$, $J^\star = \{5\}$, and $K^\star = \{6\}$. If $i \neq j$ and $i, j \in I^\star$, then $w_{ijt} = 0.70$. If $i \in J^\star$ and $j \in I^\star$, then $w_{ijt} = (c_0 + c_1 t + c_2 t^2 + c_3 t^3)$, where $(c_0, c_1, c_2, c_3)^\top = (-0.18, 2.10, -2.20, 1.00)^\top$. If $i \in K^\star$ and $j \in I^\star$, then $w_{ijt} = 0$. If $i \in J^\star$ and $j \in K^\star$, then $w_{ijt} = (c_0 + c_1 t + c_2 t^2 + c_3 t^3)$, where $(c_0, c_1, c_2, c_3)^\top = (0.87, 0.04, 0.10, -1.00)^\top$.

The eigenvalues are generated as in (2.11) in Section 2.4.1, but are estimated through the sample eigenvalues.

The random effects $\boldsymbol{\gamma}_i$ and the responses $\boldsymbol{y}_{it}$ are generated similarly as in Section 2.4.1.

Figures 2.2 and 2.3 illustrate the average estimated eigenvectors for the first two components. The EERE produces the heatmap that most closely resembles the heatmap from the true eigenvectors in that the loading scores estimations are very close to those of the true time-varying eigenvectors. In contrast, the positive magnitudes of Product 6 of the true first eigenvector are not captured in the heatmap of the DPCA. Additionally, the DPCA cannot detect the group changing behavior over time that occurs for Product 5 in Figure 2.2.

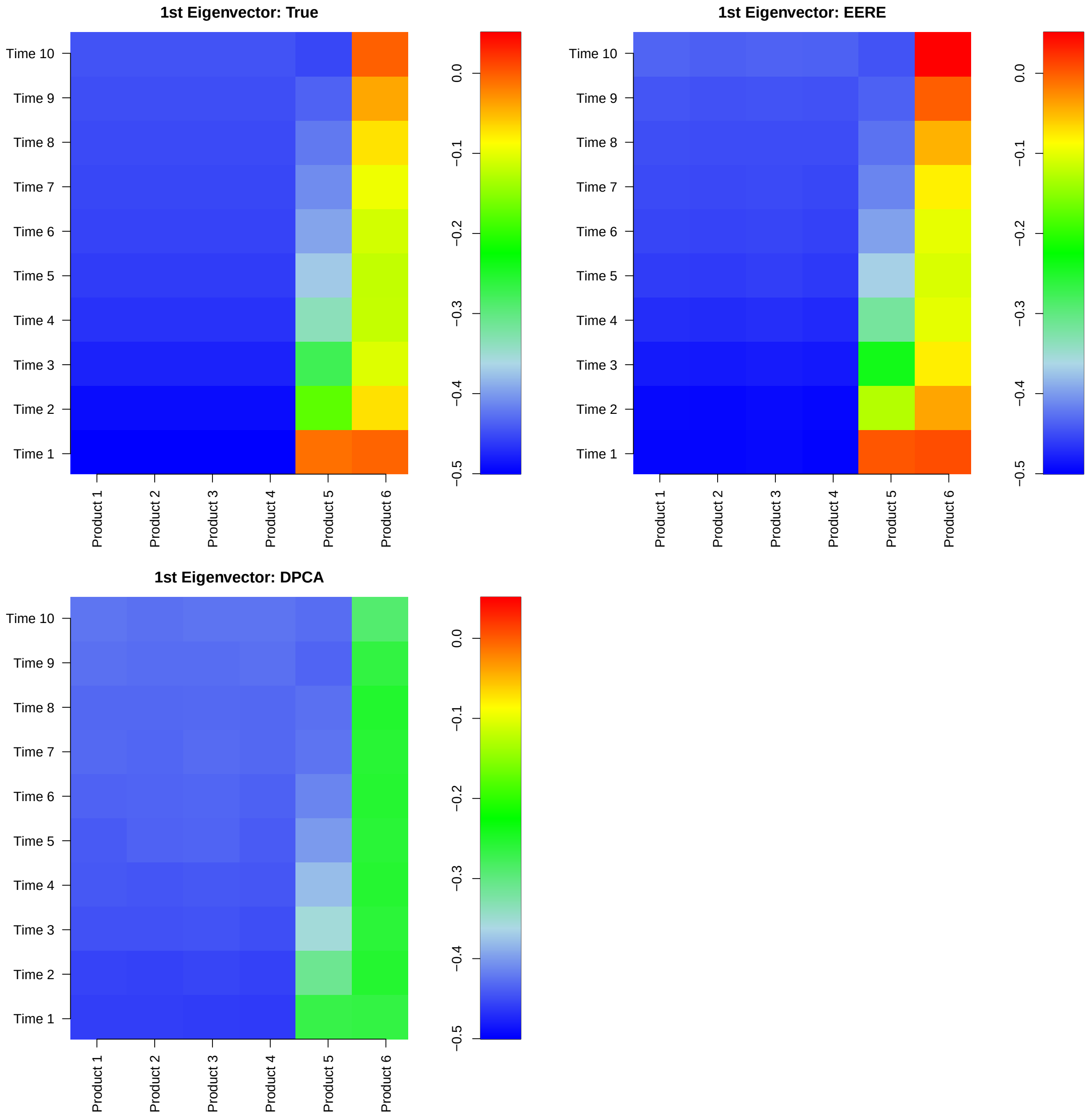
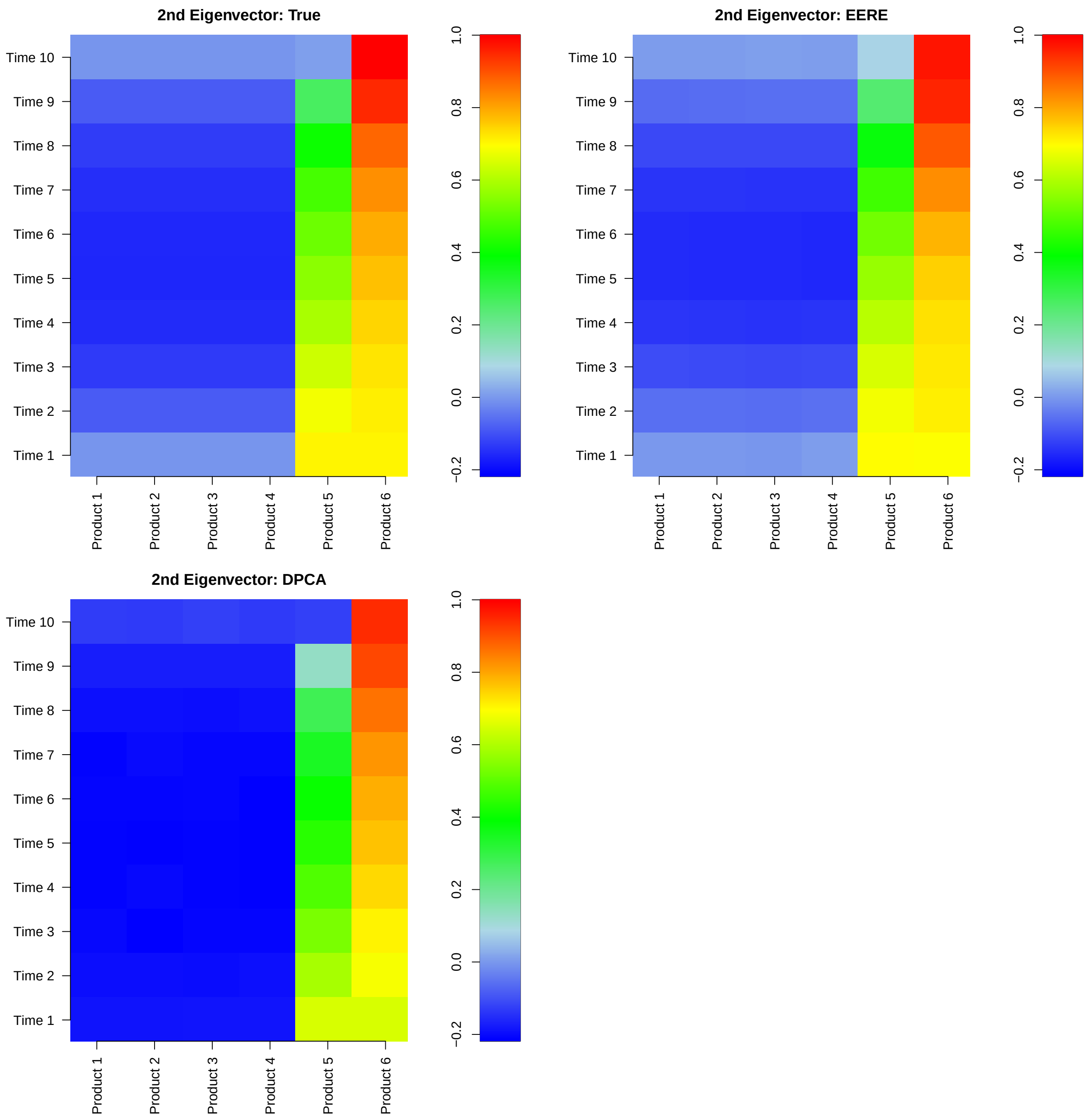


Figure 2.2: The heatmaps of the first eigenvector over time for Simulation B

For the second eigenvector estimator, Figure 2.3 shows that the heatmap of the true eigenvector is similar to the heatmap produced by the EERE. The DPCA approach cannot clearly detect a gradual change of loading score patterns over time for Products 1 to 4.



2nd Eigenvector: True

2nd Eigenvector: EERE

2nd Eigenvector: DPCA

Figure 2.3: The heatmaps of the second eigenvector over time for Simulation B

2.4.3 Simulation C: Time-varying Eigenvalues, Time-varying Eigenvectors

In this section, we carry out the EERE model assuming that both eigenvalues and eigenvectors are time-varying. We combine the data generation from Simulations A and B. In this setting, the eigenvectors are generated with $J = 10$ products belonging to the following index sets: $I^\star = \{1, \cdots, 5\}$, $J^\star = \{6, 7\}$, and $K^\star = \{8, 9, 10\}$. If $i \neq j$ and $i, j \in I^\star$, then $w_{ijt} = 0.70$. If $i \neq j$

and $i \in J^\star$ and $j \in I^\star$, then $w_{ijt} = 0.10 + 0.07t$. If $i \neq j$ and $i, j \in J^\star$, then $w_{ijt} = 0.8$. If $i \neq j$ and $i \in K^\star$ and $j \in I^\star$, then $w_{ijt} = 0.10$. If $i \neq j$ and $i \in K^\star$ and $j \in J^\star$, then $w_{ijt} = 0.70 - 0.07t$. If $i \neq j$ and $i, j \in K^\star$, then $w_{ijt} = 0.80$. The random effects $\boldsymbol{\gamma}_i$ and the responses $\boldsymbol{y}_{it}$ are generated in the same way as in Section 2.4.1.

Table 2.2 summarizes the average MADE values for the first and second eigenvalue estimators when $N = 200$ stores, $J = 10$ products, and $T = 10$ time points which are in the range $[0, 1]$. It shows that the EERE method yields smaller average of MADEs compared to DPCA, and the efficiency improvement from the EERE method is more substantial for the first eigenvalue. Figure 2.4 displays the average first and second time-varying eigenvalue estimators obtained from the EERE and DPCA. In Figure 2.4, we notice that the time-varying eigenvalue estimators produced by the EERE are closer to the true eigenvalue curves, while the eigenvalue estimators from the DPCA are far from the true eigenvalue curves.

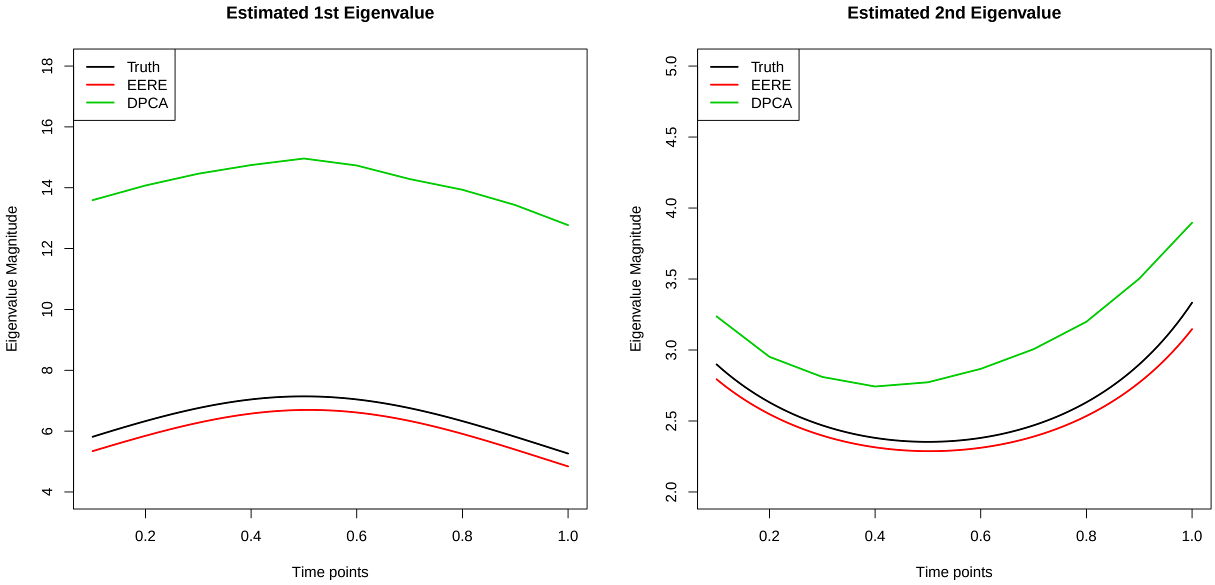
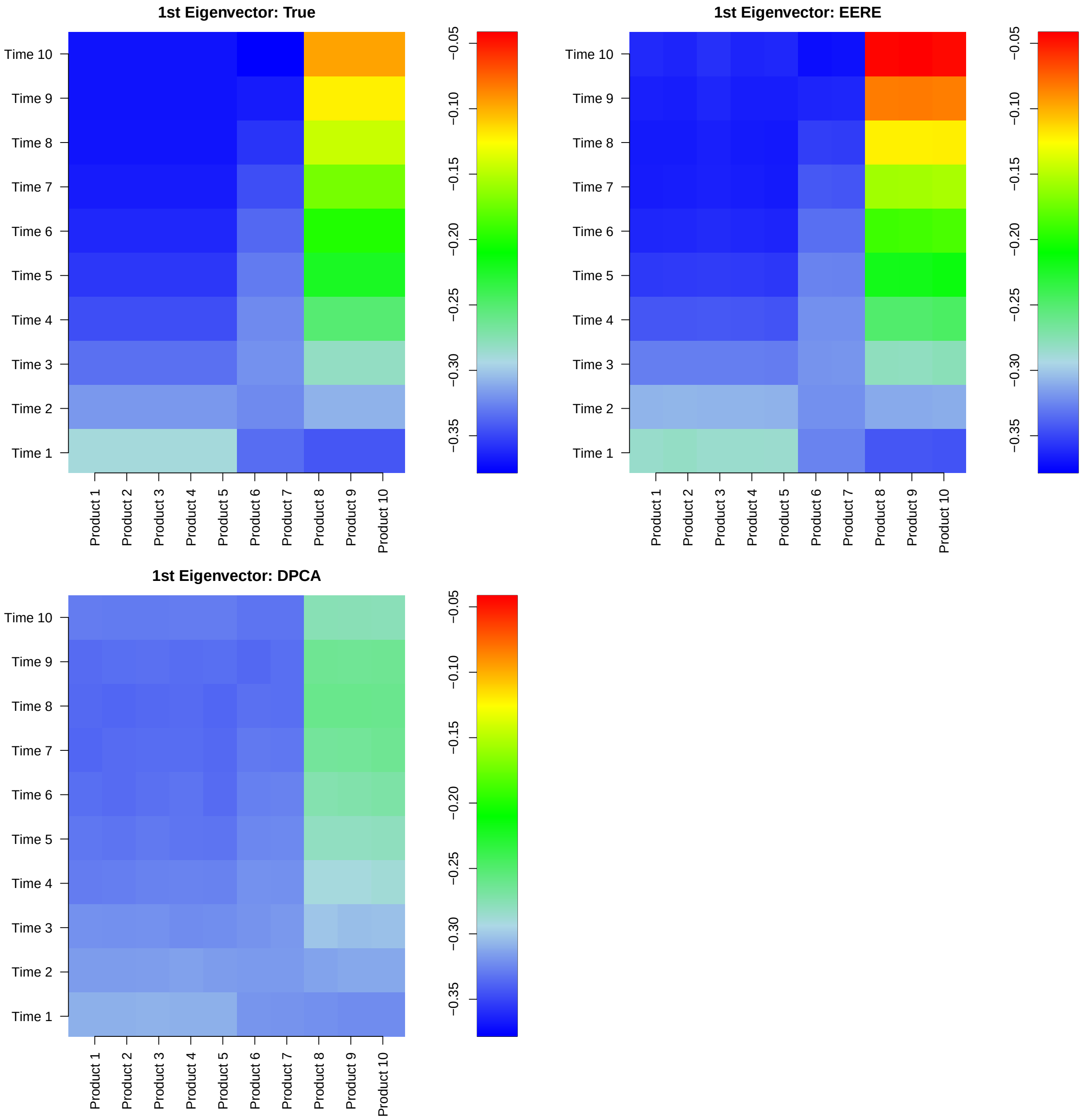


Figure 2.4: Average estimated first and second eigenvalues for Simulation C

$N = 200$, $J = 10$	1^{st} Eigenvalue	2^{nd} Eigenvalue
EERE	0.44	0.35
DPCA	4.08	0.48

Table 2.2: Average MADEs of the first and second eigenvalues for Simulation C

Likewise, the heatmaps of Figure 2.5 show that the average first eigenvector estimator from the EERE is not much different from the true first eigenvector. On the other hand, the DPCA method, which fails to account for the variation of the stores, results in estimates that give comparatively more negative loading scores. Therefore, it is not able to capture the group changes over time for Products 8 to 10 at later time points. For the second eigenvector estimation, Figure 2.6 indicates that the heatmap of the true eigenvector is quite close to that of the EERE method, while the DPCA is incapable of detecting the dynamic change of patterns for Products 1 to 7 over time.



1st Eigenvector: True

1st Eigenvector: EERE

1st Eigenvector: DPCA

Figure 2.5: The heatmaps of the first eigenvector over time for Simulation C

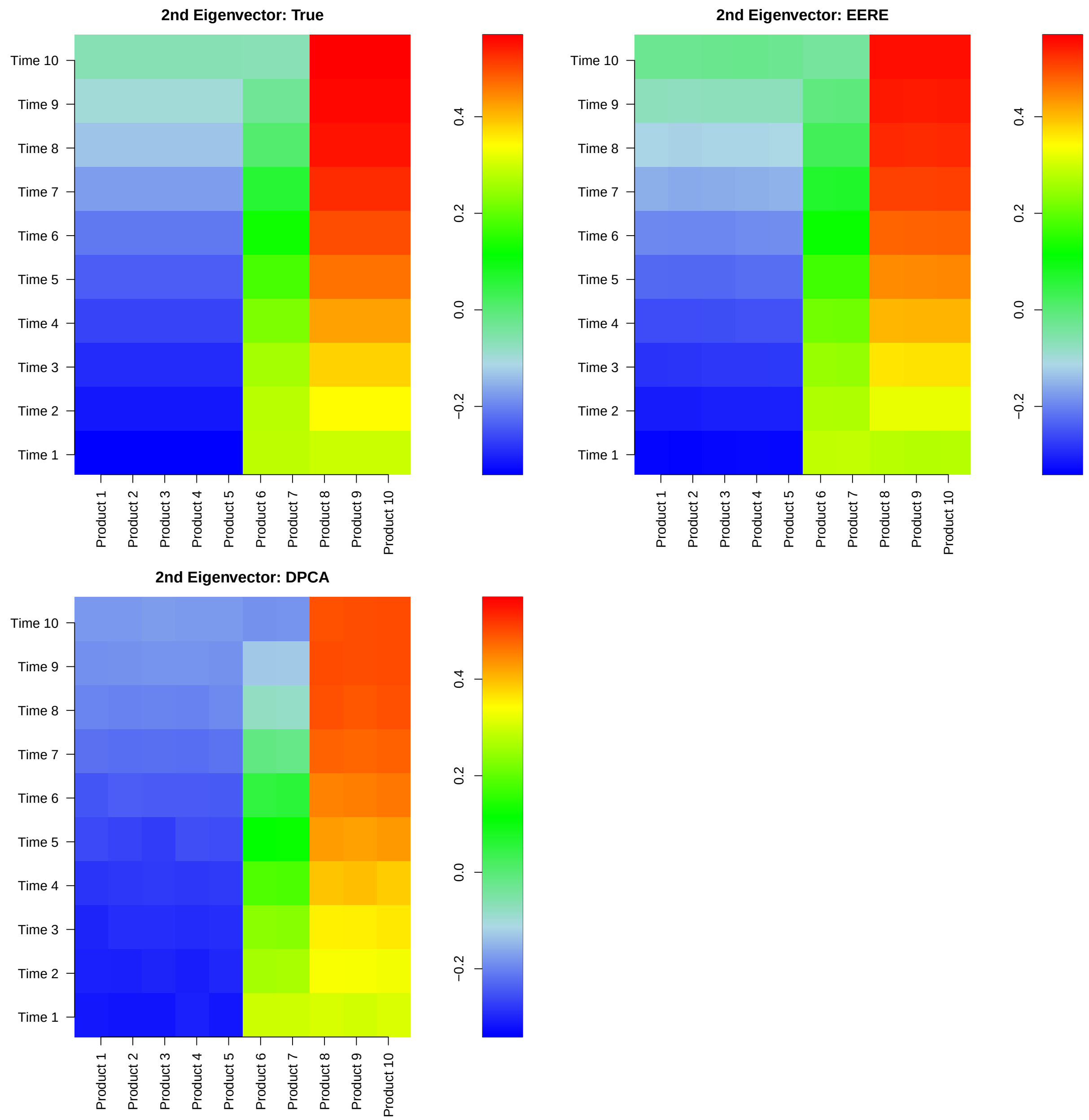


Figure 2.6: The heatmaps of the second eigenvector over time for Simulation C

2.5 Real Data Application: IRI Grocery Store Sales Data

In this section, we focus on grocery store sales spanning the eleven-year time period, and apply the proposed method to the IRI marketing data set, which consists of sales units of 552 grocery stores and 20 products collected over the years 2001 - 2011. Among the stores, 30% are located in the South, 28% in the West, 24% the Northeast, and 18% in the Midwest, and fewer than 1% of the stores do not belong to any chain. The product categories include beer, razor blades, carbonated beverages, cigarettes, cold cereals, deodorants, diapers, frozen dinners, frozen pizzas, hot dogs, household cleaners, laundry detergent, milk, mustards and ketchups, peanut butters, photography supplies, salty snacks, shampoos, soup, and toothbrushes. The twenty products represent a broad spectrum of consumer packaged goods with varying amounts of sales over time. Among the products, milk has the largest volume of sales across time, and photography supplies have the smallest volume of sales over time.

This longitudinal sales data presents some interesting features, but is also challenging to analyze due to the sheer size, variability, and time-varying nature of the data set. The variation among stores can be due to several extrinsic factors, such as geographic location and store size, or due to intrinsic factors, such as popularity and reputation of the stores. Figure 2.7 illustrates the average number of units sold at large, medium, and small stores for two products, beer and peanut butter. We notice that smaller stores tend to have a decreasing trend in sales of beer and peanut butter, while larger stores have an overall increase. Figure 2.8 shows that a selected group of eight products' average sales present quite different patterns over time. In particular, there is an obvious decrease in average sales for photography supplies.

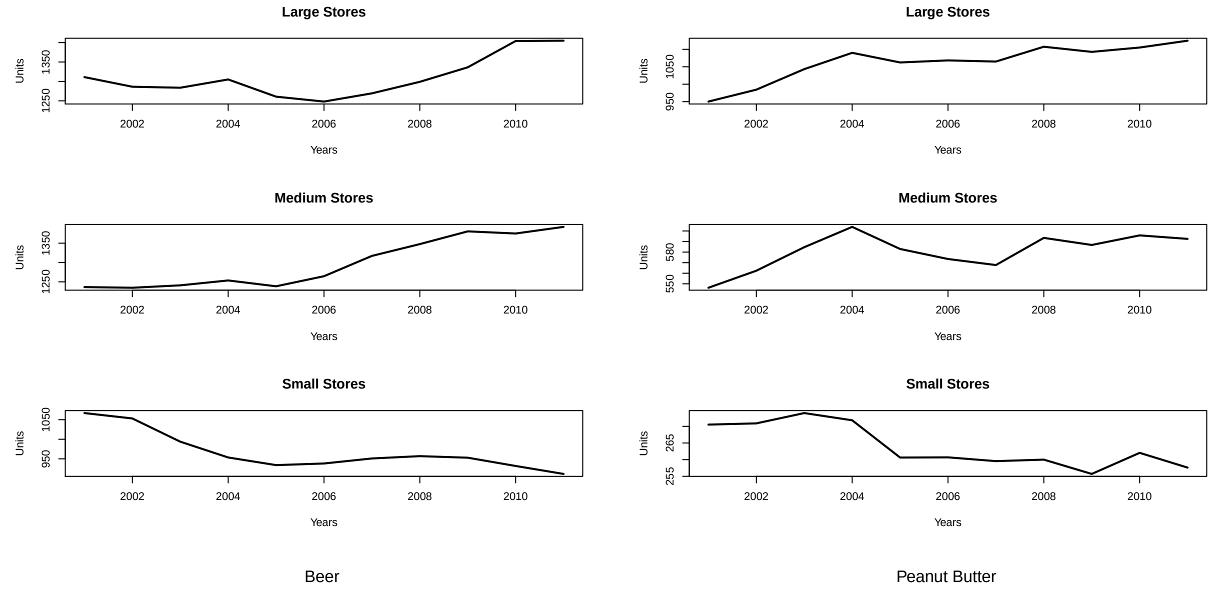


Figure 2.7: Average product sales for three store sizes for beer and peanut butter products.

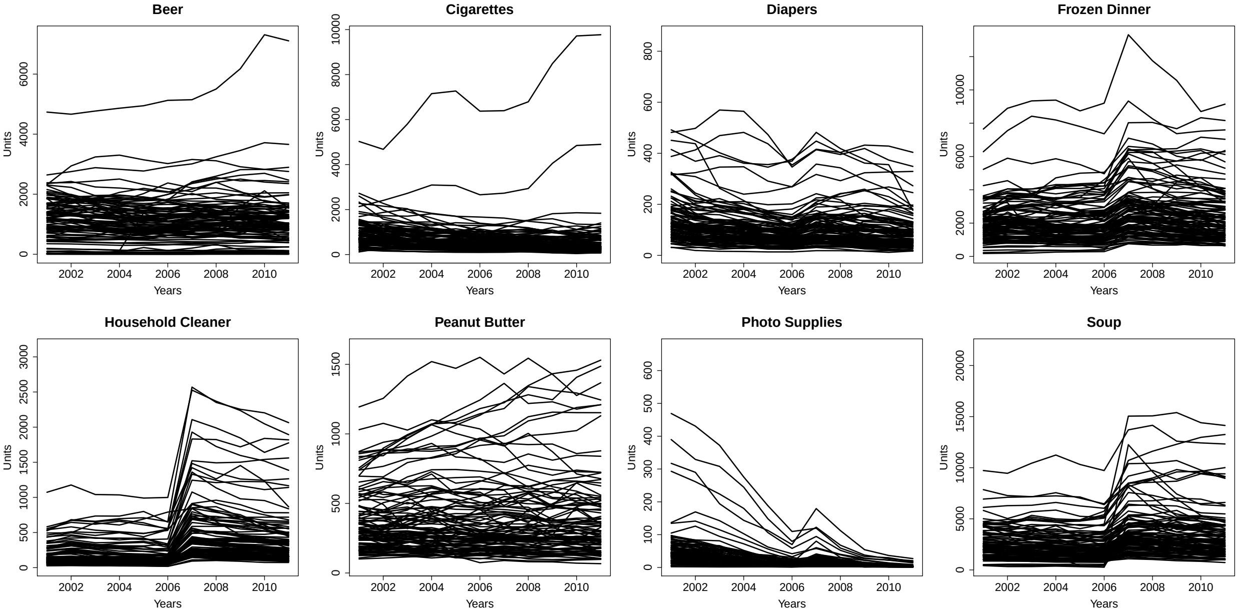


Figure 2.8: Average product sales from 100 stores over time among eight products.

Since the number of products is quite large, it is also essential to investigate correlations among product sales to capture the associations among them over time. This aids marketing researchers who want to better understand consumer shopping behavior and marketing trends. Figure 2.9 illustrates the heatmaps for the correlation matrices among several selected products. We note that the magnitude among the pairwise correlations for photography supplies and other products changes over time, and the change is more obvious in later years. This phenomenon reflects the fact that consumers are likely to change their purchasing habits as technology progresses over time. Incorporating time-varying information plays an important role in analyzing this type of data.

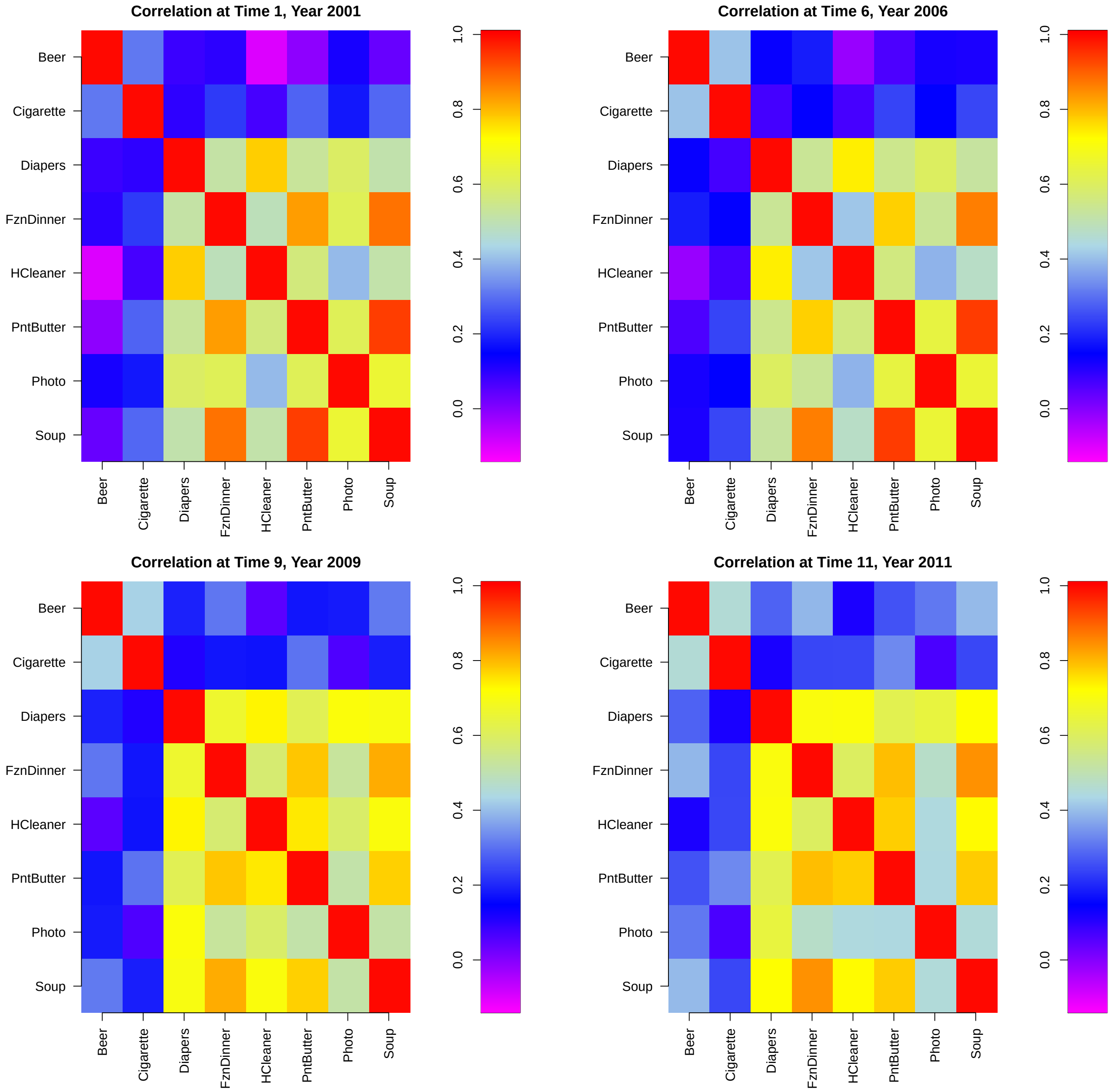


Figure 2.9: Correlations for eight products at the years 2001, 2006, 2009, and 2011.

We implement the proposed EERE method and compare it to the DPCA approach. Figures 2.10

and 2.11 show the first and second time-varying eigenvector heatmaps, respectively. Based on the these heatmaps, the DPCA has an overall averaging behavior for the first eigenvector and a grouping behavior for the second eigenvector. The two groups indicated in the DPCA second eigenvector form a contrast of products that are ingestible versus non-ingestible. The negative magnitudes of the loading scores in the second eigenvector for most years correspond to the following products: beer, carbonated beverages, cigarettes, cold cereals, frozen dinners, frozen pizzas, hot dogs, milks, mustards and ketchups, peanut butters, salty snacks, and soup. These products are those which consumers take into their bodies via swallowing or inhaling, hence "ingestible." The remaining non-ingestible products include razor blades, deodorants, diapers, household cleaners, laundry detergent, photography supplies, shampoo, and toothbrushes. The non-ingestible products are the ones obviously not taken.

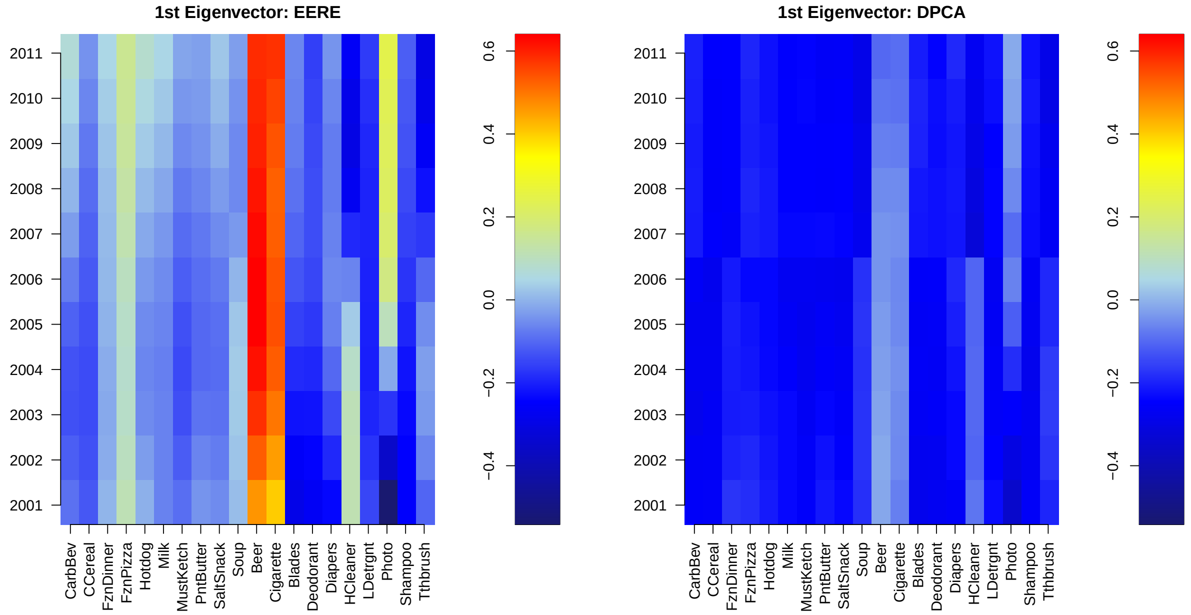


Figure 2.10: Heatmaps of the first eigenvector of the IRI marketing data containing 20 products

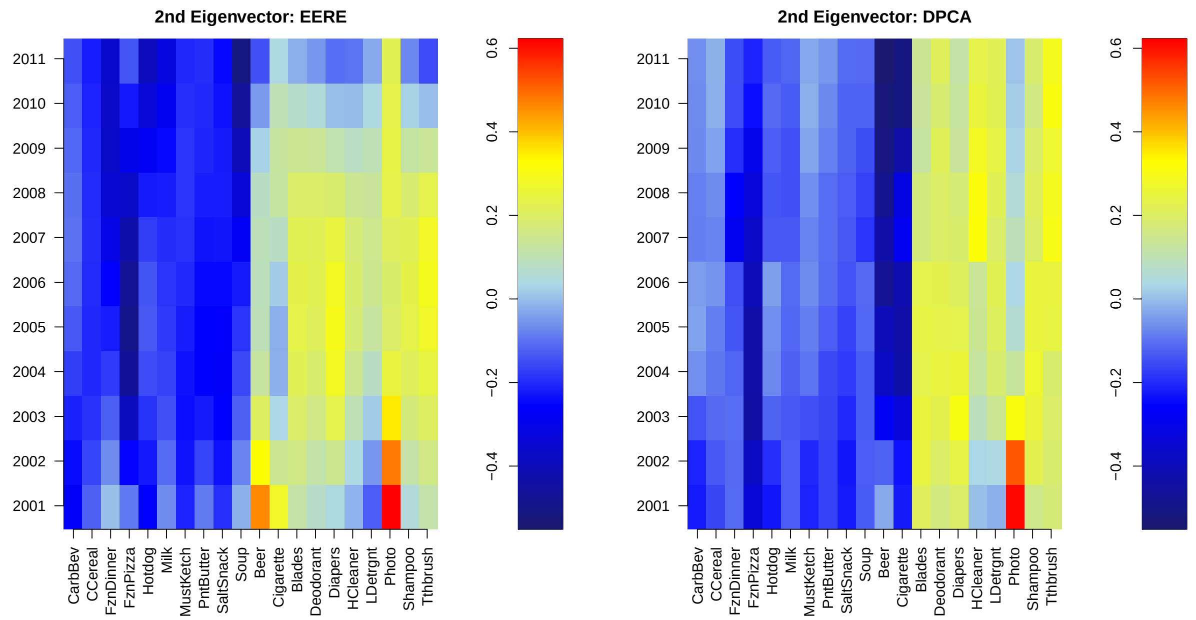


Figure 2.11: Heatmaps of the second eigenvector of the IRI marketing data containing 20 products

The EERE heatmap in Figure 2.10 does not provide an overall averaging behavior like the DPCA. On the contrary, it indicates that the beer and cigarette products should be grouped together in a different category than the remaining products. Another interpretation of the first eigenvector from the EERE is the distinction between products for general consumers versus products for age-restricted consumers. Beer and cigarettes are age-restricted products for which consumers must be at least ages 21 and 18 years old, respectively, to consume and purchase.

Figure 2.11 reveals that the proposed EERE groups products differently from the DPCA. Particularly, the EERE makes a distinction between foods and non-alcoholic beverages versus the remaining products, where the foods and non-alcoholic beverages group includes carbonated beverages, cold cereal, frozen dinners, frozen pizzas, hot dogs, milk, peanut butter, salty snacks, and soup. The remaining product group includes beer, cigarettes, razor blades, deodorants, diapers, household cleaners, laundry detergent, photography supplies, shampoo, and toothbrushes. Incorporating the random effects improves the estimation efficiency of the second eigenvector in that the contrast between groups is more specific. Although beer and cigarettes are commonly categorized as digestible products, they are quite different from foods.

In the heatmaps of Figure 2.10, the EERE displays photography supplies shifting from a moderately negative value to a positive value as the years progress from 2001-2011. The household cleaner category shows a change in magnitude from positive to negative values. The time-changing behavior is almost unnoticeable for the DPCA heatmap of Figure 2.10. Notice that the sale of photography supplies in grocery stores experienced a major decline nationwide after 2005, which is when camera phones and digital cameras gained popularity.

The strength of the EERE method lies in the interpretation of essential features and associations among products over time, which can be captured in the time-varying eigenvectors. As a result, we can determine grouping behavior through the signs and magnitudes of the eigenvectors. We show that our method, which incorporates random effects, leads to better overall estimation of time-varying eigenvectors compared to the DPCA. Further, we show the advantage of the proposed method in interpreting the time-varying eigenvectors. The eigenvalues act as weights of the principal directions of the time-varying eigenvectors, but do not enhance the interpretation on grouping products for the marketing data. In summary, the EERE method, which incorporates random effects from stores, provides a more informative and precise grouping strategy for products compared to methods that fail to consider random effects.

2.6 Discussion

In this chapter, we propose the EERE methodology to incorporate random effects in modeling time-varying eigenvalues and eigenvectors for the longitudinal PCA. The proposed longitudinal PCA performs the decomposition of covariance matrices over time, and takes store variability into account through modeling the heterogeneous effects of product sales. This leads to improved interpretation of the eigenvectors, which could be extremely useful in clustering grocery products that consumers purchase.

Our simulation studies indicate that the proposed method has lower mean absolute deviation of errors for the time-varying eigenvalue estimation, and that the proposed time-varying eigenvector estimators match the true eigenvectors more closely compared to the DPCA method. The algorithm can be extended for handling a larger number of products, and the nonparametric functions can be implemented beyond the quadratic or cubic terms for time-varying eigenvalue approximations.

In addition, the analysis of the IRI marketing data provides an illustration of insight on how statistics and data analytics can play a role in business decision-making. Specifically, how we can effectively utilize large marketing data over time to capture changes in consumer shopping behavior longitudinally, and how we can extract intrinsic information about the associations among products more accurately.

This work tackles a real data problem involving the complex correlated nature of observations over time. While there is no definite rule for selecting the number of principal components, we acknowledge that the proposed method relies on an ad-hoc optimum number of components based on the cumulative percentage of total variation averaged over time. In addition, further research might be needed in incorporating random effects from multiple sources, such as time-varying random effects to account for temporal correlation, and block-wise random effects to capture spatial locations of stores.

Chapter 3

Longitudinal Principal Components for Binary Data

3.1 Introduction

Binary data abound in biomedical studies about patients and their behaviors, as well as in images about those patients. The simple dichotomization of traits and various measures of health can be extended to areas outside of biomedical practice, specifically to retail. Time-varying associations among consumers and the items they purchase provide important information to capture changes in consumer shopping behavior. Consumer-level demographics, which are often found in panel data, can illuminate the various types of consumers that exist. Some demographic information included in the IRI panel data are household income, age, race, marital status, education level, and type of residence. Although demographics change infrequently over time, purchasing behavior can change dramatically, especially for different types of product categories. When the purchases can be dichotomized into "did purchase" or "did not purchase" categories, analysis can become strenuous even though binary outcomes may appear simple. Thus, it is important to consider methods that capture product associations among consumers as time changes. Among all products included in the data set, pairs of binary variables can be stored in a 2×2 contingency table.

Among various measures of association for 2×2 contingency tables, the odds ratio or cross-product ratio has advantageous properties and lends itself to interpretation. The odds for the i-th row of the table can be generally written as π_{i1}/π_{i2}, where the numerator represents a success probability and the denominator represents a failure. Thus, the odds ratio of the two rows is

$$\tau = \frac{\pi_{11}/\pi_{12}}{\pi_{21}/\pi_{22}} = \frac{\pi_{11}\pi_{22}}{\pi_{12}\pi_{21}}.$$

The odds ratio represents the ratio of the odds of success for row 1 versus the odds of success for row 2. The odds ratio takes values in the interval $(0, \infty)$, while $\tau = 1$ represents independence of two variables in a contingency table. Pairs of variables may exhibit strong association when $\tau << 1$ or when $\tau >> 1$. Another property of the odds ratio is the invariance property, which indicates that interchanging the row variable and the column variable in the table does not change the value of θ(Bishop et al., 2007; Agresti, 2002). When considering departures from dependence, the Pearson correlation is not fully related to the odds ratio, and thus may not always be appropriate for comparison (Mosteller, 1968; Bishop et al., 2007). Thus, we use the odds ratio as the measure of variation that we want to maximize in principal components analysis.

Odds ratios have been used as measures of association for binary data. In Lipsitz et al. (1991), odds ratios are used for longitudinal observations, where the 2×2 contingency tables were generated from responses at pairs of time points. Ultimately, their goal was to estimate correlation parameter α using generalized estimating equations (GEE). In Carey et al. (1993), they propose a method of alternating logistic regressions to overcome cluster size limitations in estimating α. Also, pairwise odds ratios are used as associations among multivariate binary data. Both papers extend the work of Liang & Zeger (1986) and Prentice (1988) by modeling both mean and association parameters α and β through GEEs.

There are several methods developed for principal component analysis of binary data. PCA for binary data has been formulated with a likelihood approach for Bernoulli distributed random variables (Collins et al., 2002; Schein et al., 2003; Leeuw, 2006; Lee et al., 2010). In those papers, the authors generalize their modeling because the Bernoulli distribution belongs to the exponential family. Alternative modeling approaches have also been considered. PCA is modeled with hidden variables such that the binary variables are generated via a nonlinear threshold (Lee &

Sompolinsky, 1999). Nonlinear PCA (Gifi, 1990) has been developed for non-numerical data, including binary, ordinal, and categorical data. This methodology is comparable to standard PCA and achieves the same type of dimension reduction when the data are numeric. PCA is built as a saturated model in Landgraf & Lee (2015), while only solving for principal component loadings or eigenvectors. Although the spectrum of available approaches for PCA is broad, the existing approaches do not extend readily to longitudinal binary data. In addition, the above methods do not account for variation over time, while estimating both eigenvectors and eigenvalues. For these reasons, a longitudinal PCA approach needs to be developed to account for longitudinal binary data.

This research is motivated by the IRI marketing and panel data sets. In the previous chapter, we discuss the IRI marketing data set. In this chapter, we apply our longitudinal PCA methodology to the IRI panel data set, which contains weekly purchases of consumers called panelists, along with 20 demographic variables from two of the regional markets - Eau Claire, Wisconsin and Pittsfield, Massachusetts. The panel data is available for all 31 products over the 11-year time period. This data requires processing because all panelists do not purchase all products, and the time component allows for irregular trajectories of panelists' shopping behavior. See Kruger & Pagni (2008) for more details.

We propose a longitudinal PCA for binary data that decomposes association information arising from multivariate observations over time. Due to nature of binary data, the association information is captured by the odds ratios of pairs of variables at each time point. Associations among product sales are modeled through time-varying eigen-decomposition. This method is an extension of the methodology in the previous chapter for the binary data case. The techniques and models required for the extension are complicated due to the challenging binary data structure.

3.2 Model Framework and Methodology

3.2.1 The Logit Model and Odds Ratio Matrix

Let y_{ijt} be a binary response measured over time t for the j-th observation from the cluster i where $i = 1, 2, \ldots, N$, $j = 1, 2, \ldots, J$, and t is in the range of $[0, 1]$. For this outcome, $y_{ijt} = 1$ if the i-th individual for the j-th variable makes a purchase at week t, and $y_{ijt} = 0$ otherwise. We model $E(y_{ijt})$ since the binary responses are multivariate. The mean for the response vector $E(y_{ijt}) = \pi_{ijt} = \text{pr}(y_{ijt} = 1)$ contains the marginal probabilities. The logit link function $\log(\frac{\pi_{ijt}}{1-\pi_{ijt}}) = \mu_{jt}$.

Because we are studying the binary data setting, the association among the variables will be captured by the odds ratio. Let $\text{OR}(\boldsymbol{y}_{it})$ be a function that computes the odds ratio of the response vector, then the EE model is

$$\boldsymbol{O}_t^0 = \text{OR}(\boldsymbol{y}_{it}). \tag{3.1}$$

The $\boldsymbol{O}_t^0$ is a $J \times J$ matrix of true pairwise odds ratios and is generally unknown. We can replace $\boldsymbol{O}_t^0$ with a carefully designed matrix of sample pairwise odds ratios $\tilde{\boldsymbol{O}}_t$. The matrix $\tilde{\boldsymbol{O}}_t = (\theta)_{jkt}$ is designed to attain the properties of symmetry as a result of the invariance property.The sample pairwise odds ratios for binary variables j and k at time t is

$$\theta_{jkt} = \frac{n_{jkt}(1 - n_{jt} - n_{kt} + n_{jkt})}{(n_{jt} - n_{jkt})(n_{kt} - n_{jkt})},$$

where $n_{jkt} = \sum_{i=1}^{N}(y_{ijt} + y_{ikt})/N$ and $n_{jt} = \sum_{i=1}^{N} y_{ijt}/N$, where $j = 1, \ldots, J$ and $k = 1, \ldots, J$. These θ_{jkt} make up the off-diagonal elements of $\tilde{\boldsymbol{O}}_t$, since the pair of variables $j \neq k$, and $\theta_{jkt} = \theta_{kjt}$ by the invariance property of the odds ratio. Hence for arbitrary diagonal elements, $\tilde{\boldsymbol{O}}_t = \tilde{\boldsymbol{O}}_t^\top$. The odds ratio matrix has two shortcomings. First, it needs to be adapted for the cases when a cell in the 2 × 2 table equals 0, because this can lead to $\theta_{jkt} = \infty$. Second, the diagonal of $\tilde{\boldsymbol{O}}_t$ must be defined to attain positive semi-definiteness. A covariance matrix attains positive semi-definiteness by definition, and since we want to capture associations, it is fruitful for our measure of association for binary variables to also attain this property. Additionally, the eigenvalues that we estimate are not assumed to be negative-valued. To address these two shortcomings, we transform the odds ratios via the Yule's Q function ,

$$q_{jkt} = \frac{\theta_{jkt} - 1}{\theta_{jkt} + 1},$$

where the new range of q_{jkt} is $(-1, 1)$. This function is monotonic, maintains the same properties as θ_{jkt}, and has a useful interpretation that parallels the interpretation Pearson correlation. Two variables have strong association if $q_{jkt} = \pm 1$. A pair of variables are deemed independent when $q_{jkt} = 0$. These q_{jkt} values are off-diagonal elements of $\tilde{\boldsymbol{Q}}_t$.The diagonal elements are $q_{jjt} = 1$, and help attain the positive semi-definite property. Yule's Q as a measure of association also has useful asymptotic properties. See Bishop et al. (2007) for more details.

With the aforementioned properties and the measures of association, the behavior of $\tilde{\boldsymbol{Q}}_t$ is similar to a covariance matrix and correlation matrix, and we can decompose it into a linear combination of eigenvalues and eigenvectors that are time-varying. The time varying models for the eigenvalues and eigenvectors in of Chapter 2 still hold for the binary data setting.

3.2.2 Estimation of Eigenvalues and Eigenvectors

In this section, we extend the framework of generalized estimating equations (GEE) to handle the binary data case capturing association among variables via the odds ratio. The matrix of true pairwise odds ratios shall be approximated by 2.3, however we note that this matrix is informing us about the association instead of the variance. Now, the difference $\boldsymbol{h}_{it}$ in 2.4 replaces $\tilde{\boldsymbol{V}}_t$ with the odds ratio matrix as

$$\boldsymbol{h}_{it} = \left(\boldsymbol{V}_t^{-1} - \boldsymbol{O}_t^{0^{-1}} \right) (\boldsymbol{y}_{it} - \boldsymbol{\mu}_t), \tag{3.2}$$

where $\boldsymbol{\mu}_t$ represents the marginal probabilities at each time point. We minimize the following objective function with respect to both $\alpha_k(t)$ and $\boldsymbol{e}_k(t)$

$$\sum_{t=1}^{T} \left(\sum_{i=1}^{N} \frac{\boldsymbol{h}_{it}^\top \boldsymbol{h}_{it}}{N} + \phi \sum_{i \neq j} \| \boldsymbol{e}_i(t)^\top \boldsymbol{e}_j(t) \|_2^2 \right), \tag{3.3}$$

We have presented the EE methodology, which does not incorporate random effects, but does include time-varying eigenvalue and eigenvector estimation techniques. The work on random effects incorporation for binary data is ongoing.

3.3 Implementation

In this section, we provide the algorithm of the EE model 3.1 which iterates through the Newton-Raphson and the Estimation-Substitution algorithms. To minimize the objective function in 3.3, we replace $\boldsymbol{O}_t^{0^{-1}}$ by its sample version $\tilde{\boldsymbol{O}}_t^{-1}$ mentioned above.

```
Algorithm: Estimated Eigenanalysis (EE)
Step 1: Set the initial values of the eigenvectors as the sample eigenvectors:
$\boldsymbol{e}_k(t)^{(0)} = \tilde{\boldsymbol{e}}_k(t)$;
Step 2: Given the current eigenvectors $\boldsymbol{e}_k(t)^{(m-1)}$,
(i) update the eigenvalues $\alpha_k(t)^{(m)}$ by minimizing the objective function in (3.3), and
(ii) update $\boldsymbol{e}_k(t)^{(m)}$ given $\alpha_k^{(m)}$ using the Newton-Raphson algorithm;
Step 3: Iterate Step 2 if
(i) $\|\alpha_k(t)^{(m)} - \alpha_k(t)^{(m-1)}\| > \epsilon_\alpha$, where $\epsilon_\alpha$ is a chosen tolerance level, or
(ii)$\|(\boldsymbol{e}_k(t)^{(m)})(\boldsymbol{e}_k(t)^{(m)})^\top - (\boldsymbol{e}_k(t)^{(m-1)})(\boldsymbol{e}_k(t)^{(m-1)})^\top\| > \epsilon_e$ , where $\epsilon_e$ is a chosen tolerance level.
```

In Step 2, the Newton-Raphson algorithm for the eigenvectors estimation takes place, followed by the Gram-Schmidt orthonormalization process, just as in the previous chapter. We use the same criterion for selecting the number of components to retain, where the eigenvalues $\alpha_k(t)$ contribute to the proportion of association in the data.

3.4 Numerical Study

In this section, we provide a simulation study to investigate the numerical performance of the proposed method for one setting. We compare the eigenvalues and eigenvector results from the proposed method with the discretized PCA (DPCA). In this setting, a discretized PCA has two separate forms, DPCA which is an eigen-decomposition of a correlation matrix at each time point, and DPCAT, which is an eigen-decomposition of a tetrachoric correlation matrix at each time point. The tetrachoric correlation is a justified comparable measure of association for binary data. Often in the literature, it has been used for binary data arising from 2×2 tables (Goodman, 1981; Becker & Clogg, 1988; Bonett, 2007). As mentioned in the numerical study of the previous chapter, we evaluate the eigenvalue estimates through plots and the mean absolute deviation of errors, while the eigenvector estimates are evaluated through heatmaps.

3.4.1 Simulation: Time-varying Eigenvalues, Time-varying Eigenvectors

In this section, we carry out the EE model assuming that both eigenvalues and eigenvectors are time-varying. Just as mentioned in the numerical study of the previous chapter, we generate the time-varying eigenvalues as 2.11, where $(a_{10}, a_{11}, a_{12})^\top = (0.19, 0.20, 0.20)^\top$, $(a_{20}, a_{21}, a_{22})^\top = (0.30, 0.50, -0.50)^\top$, and t is in the range of $[0, 1]$. The remaining $J - 2$ eigenvalues are constants around 1 at each time point.

The time-varying eigenvectors are generated by decomposing $\boldsymbol{W}_t$ in are generated with $J = 10$ variables belonging to the following index sets: $I^\star = \{1, \cdots, 5\}, J^\star = \{6, 7\}$, and $K^\star = \{8, 9, 10\}$. If $i \neq j$ and $i, j \in I^\star$, then $w_{ijt} = 0.70$. If $i \neq j$ and $i \in J^\star$ and $j \in I^\star$, then $w_{ijt} = 0.10 + 0.07t$. If $i \neq j$ and $i, j \in J^\star$, then $w_{ijt} = 0.8$. If $i \neq j$ and $i \in K^\star$ and $j \in I^\star$, then $w_{ijt} = 0.10$. If $i \neq j$ and $i \in K^\star$ and $j \in J^\star$, then $w_{ijt} = 0.70 - 0.07t$. If $i \neq j$ and $i, j \in K^\star$, then $w_{ijt} = 0.80$. The resulting eigenvectors from the eigendecomposition of $\boldsymbol{W}_t$ are represented by $\boldsymbol{e}_{kt}$, which is a $J = 10$-dimensional vector for each time t and variable k. This generation setup is the same as Simulation B of the previous chapter.

The responses $\boldsymbol{y}_{it}$ are generated as random multivariate binary outcomes with specified marginal probabilities $\pi_j = 0.5$ for the 10 variables and correlations specified as standardized matrix $\boldsymbol{O}_t^0 = \sum_{k=1}^{10} \alpha_{kt}^0 \boldsymbol{e}_{kt} \boldsymbol{e}_{kt}^\top$. Note that α_{1t}^0 and α_{2t}^0 have a quadratic representation. The binary outcomes are created using the R package **bindata** based on the work of Leisch et al. (1998).

Table 3.1 summarizes the average MADE values for the first and second eigenvalue estimators when $N = 500$ stores, $J = 10$ products, and $T = 10$ time points which are in the range $[0, 1]$. It shows that the EE method yields MADEs on average that are much lower than the average MADEs for DPCA and DPCAT for the first and second eigenvalues. Figure 3.1 displays the average first and second time-varying eigenvalue estimators obtained from the EE, DPCA, and DPCAT. In Figure 3.1, the EE outperforms the two discretized approaches visually with similar shape and curvature for the first and second components. The DPCA and DPCAT have quite different results and indicate, at least in these numerical studies, that the DPCAT may be more appropriate for binary data than the DPCA, which is purely an eigen-decomposition of a correlation matrix.

$N = 500$, $J = 10$	1^{st} Eigenvalue	2^{nd} Eigenvalue
EERE	0.460	0.842
DPCA	1.472	1.117
DPCAT	0.784	0.899

Table 3.1: Average MADEs of the first and second eigenvalues

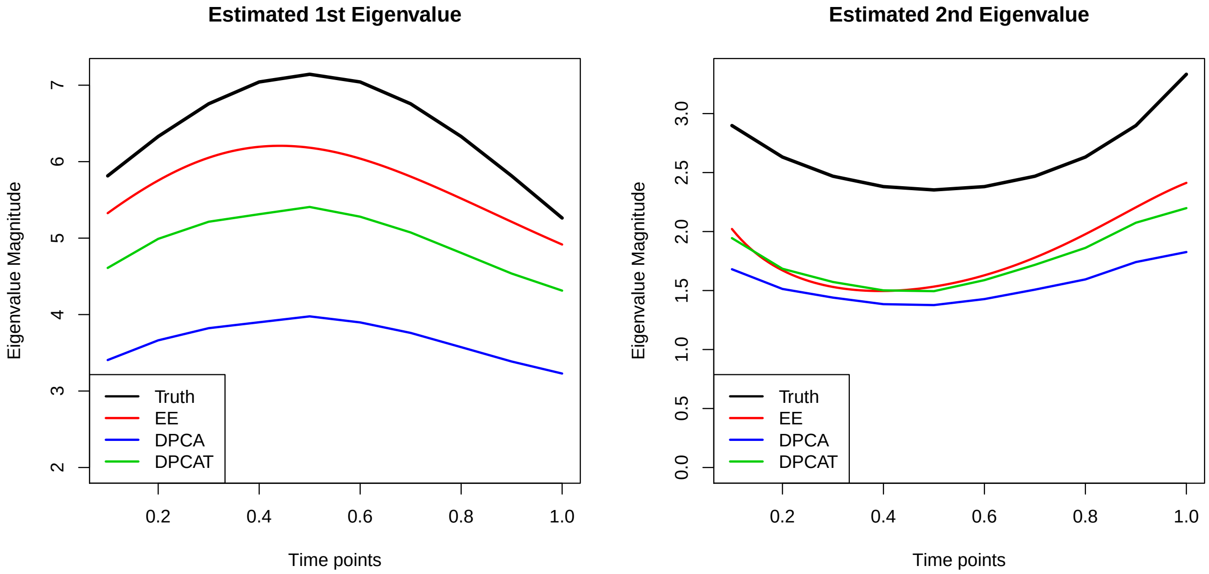
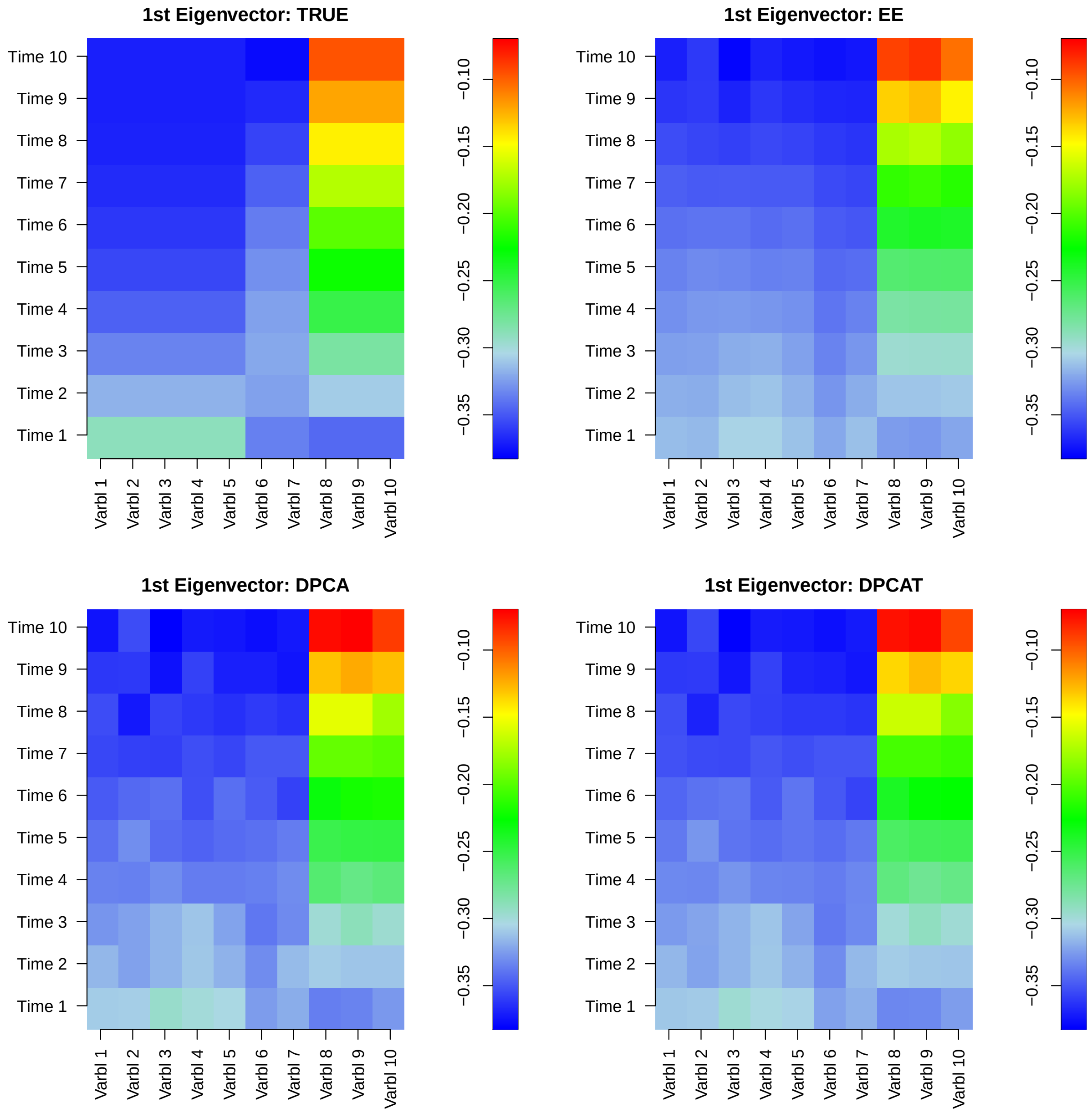


Figure 3.1: Average estimated first and second eigenvalues

The heatmaps of Figure 3.2 show that the average first eigenvector estimator from the three EE, DPCA, and DPCAT are not much different from the true first eigenvector. Similarly, Figure 3.3 indicates that the heatmap of the true eigenvector is quite close to that of the EE, as well as the DPCA and DPCAT methods. The biggest difference between the EE method and the discretized approaches is that EE incorporates smoothness, and enhances its visual appeal.



1st Eigenvector: TRUE

1st Eigenvector: EE

1st Eigenvector: DPCA

1st Eigenvector: DPCAT

Figure 3.2: The heatmaps of the first eigenvector over time

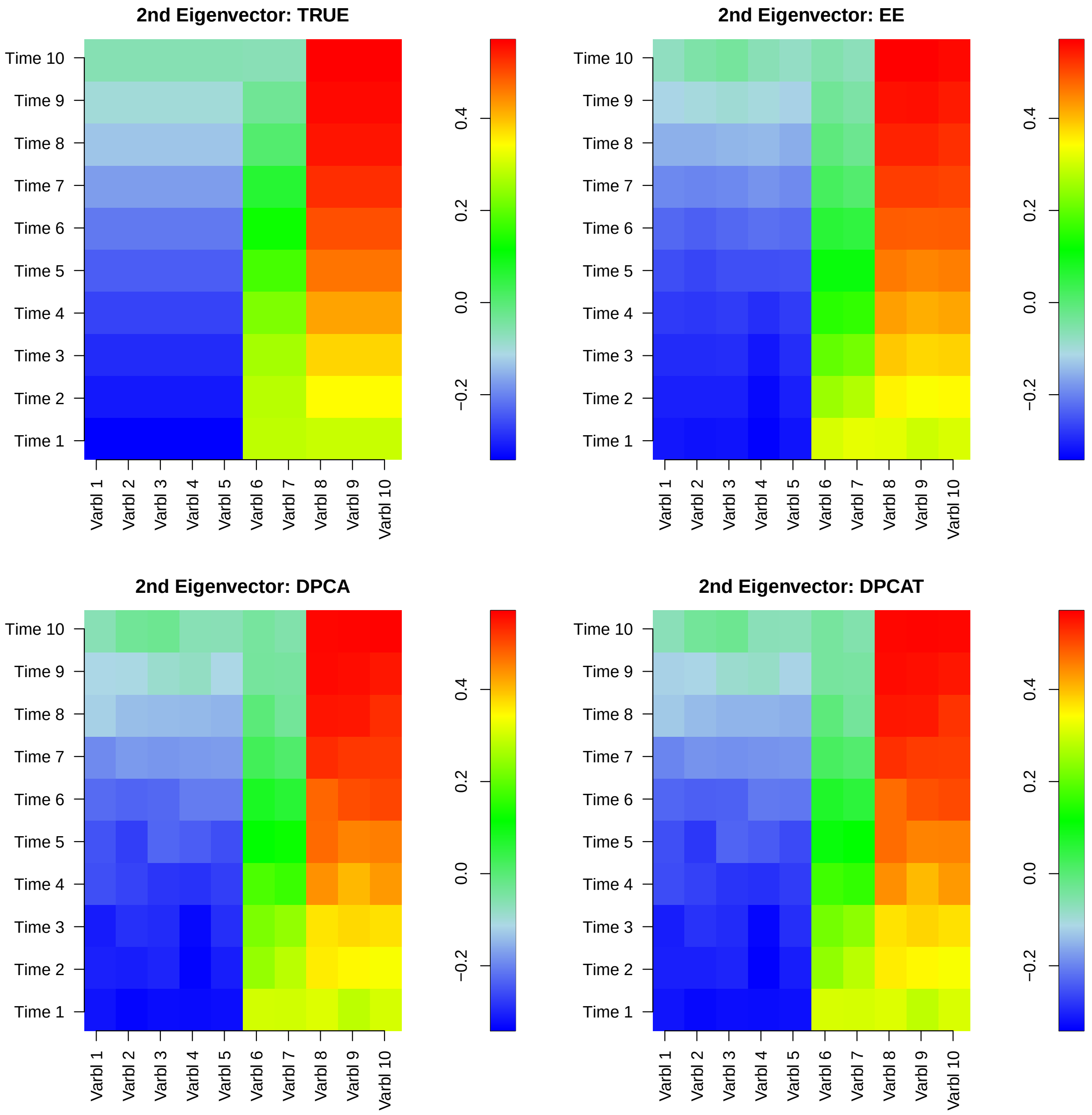


Figure 3.3: The heatmaps of the second eigenvector over time

3.5 Real Data Application: IRI Panel Purchases Data

In this section, we apply our EE methodology to the monthly consumers' purchasing behavior in the year 2007. This dataset consists of binary observations where 1 indicates that a panelist made a purchase that month, and 0 indicates they did not make a purchase. In terms of marketing information, the panelists all shop in the Pittsfield, Massachusetts market or region. There are 332 consumers and 16 products in this subset of data. There are several interesting demographics that we know about these particular panelists. The incomes of panelists cover a wide range of salaries, where 78% if panelists earn at least $35,000. Among families of four, only 7% of panelists' income are less than $25,000. A huge majority, 86%, of these consumers are homeowners in 2007, while 14% are renters . Among these panelists in 2007, roughly 11% are unemployed, while 12% are retired, the remaining 77% are employed in some fashion. About 18% of panelists report that they have at least graduated from college. The majority of panelists, 72%, report having at least graduated from high school but not graduated college. Among all panelists, 77% of panelists are married, 96% are white, and 42% have a reported family size of at least four people.

Additionally, we have insights about the product categories that these consumers purchase. The product categories include carbonated beverages, coffees, cold cereals, deodorants, frozen dinners, frozen pizzas, hot dogs, household cleaners, laundry detergent, paper towels, salty snacks, shampoos, soup, spaghetti sauces, toilet tissues, and toothpastes. The 16 products represent a broad spectrum of consumer packaged goods with varying amounts of sales over time. Among the products, salty snacks have the largest proportion of purchases across time, and deodorants have the smallest proportion of purchases across time.

We implement the proposed EE method and compare it to the DPCA and DPCAT approaches. Because we want to gain better interpretations of the insights, the eigenvectors are most useful. In the results that follow, we only show the eigenvector estimation results in the form of heatmaps. Figures 3.4 and 3.5 show the first and second time-varying eigenvector heatmaps, respectively. The eigenvector heatmaps of the DPCA and DPCAT display very jagged color changes. The first eigenvector heatmap of DPCA and DPCAT shows that all 16 products have negative loadings. In this sense, the behavior is like an overall average. The second eigenvector heatmaps of both DPCA and DPCAT contain two groups. The group of products with loadings of mostly blue hue are: deodorant, toothpaste, hot dogs, toilet tissue, and paper towels. The group of products with loadings of mostly orange hue are: carbonated beverages, household cleaners, salty snacks, frozen pizzas, and frozen dinners. These groupings are difficult to see from the heatmaps, but can be seen through assessing the numerical matrix of the estimates second eigenvector.

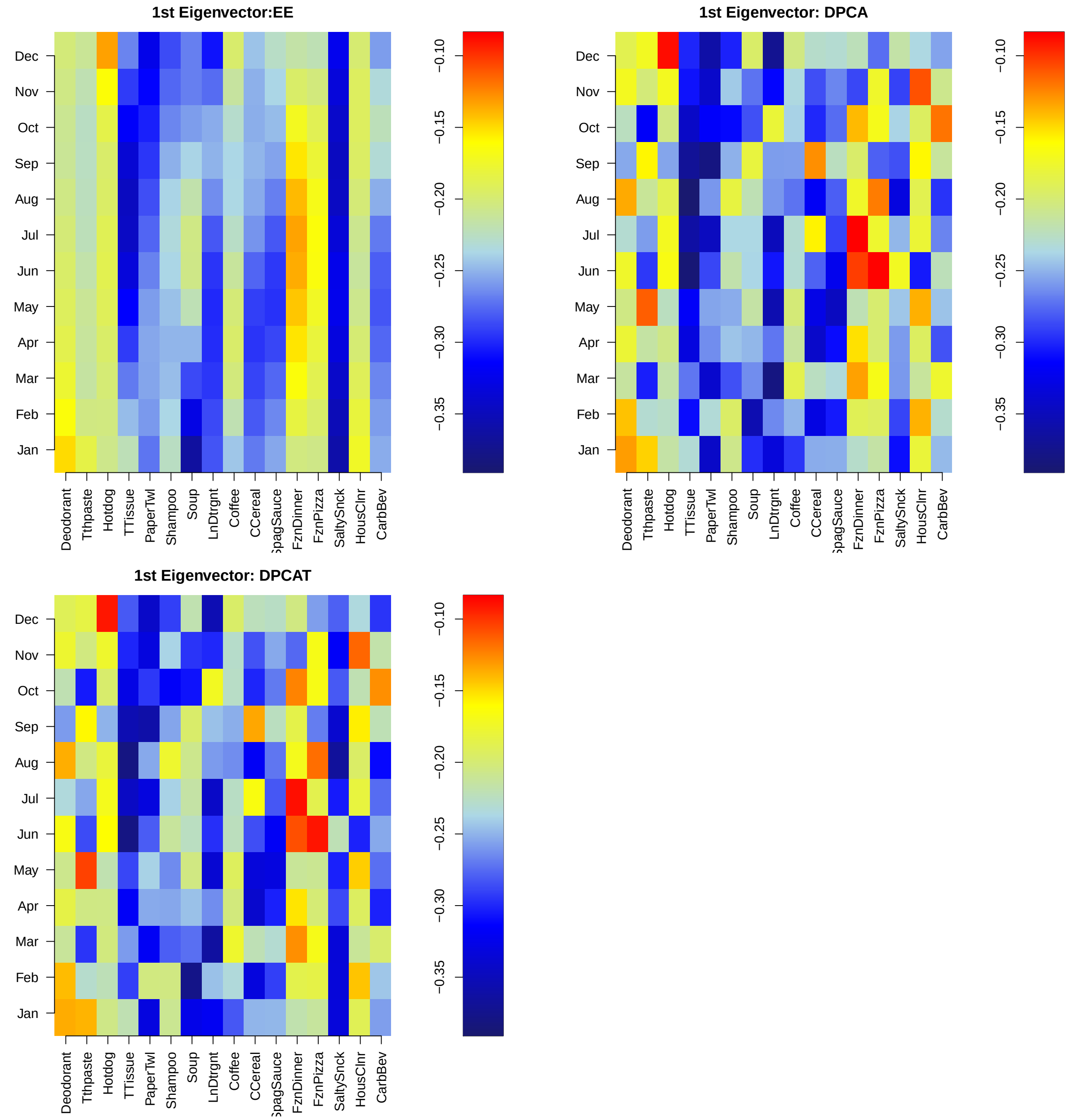
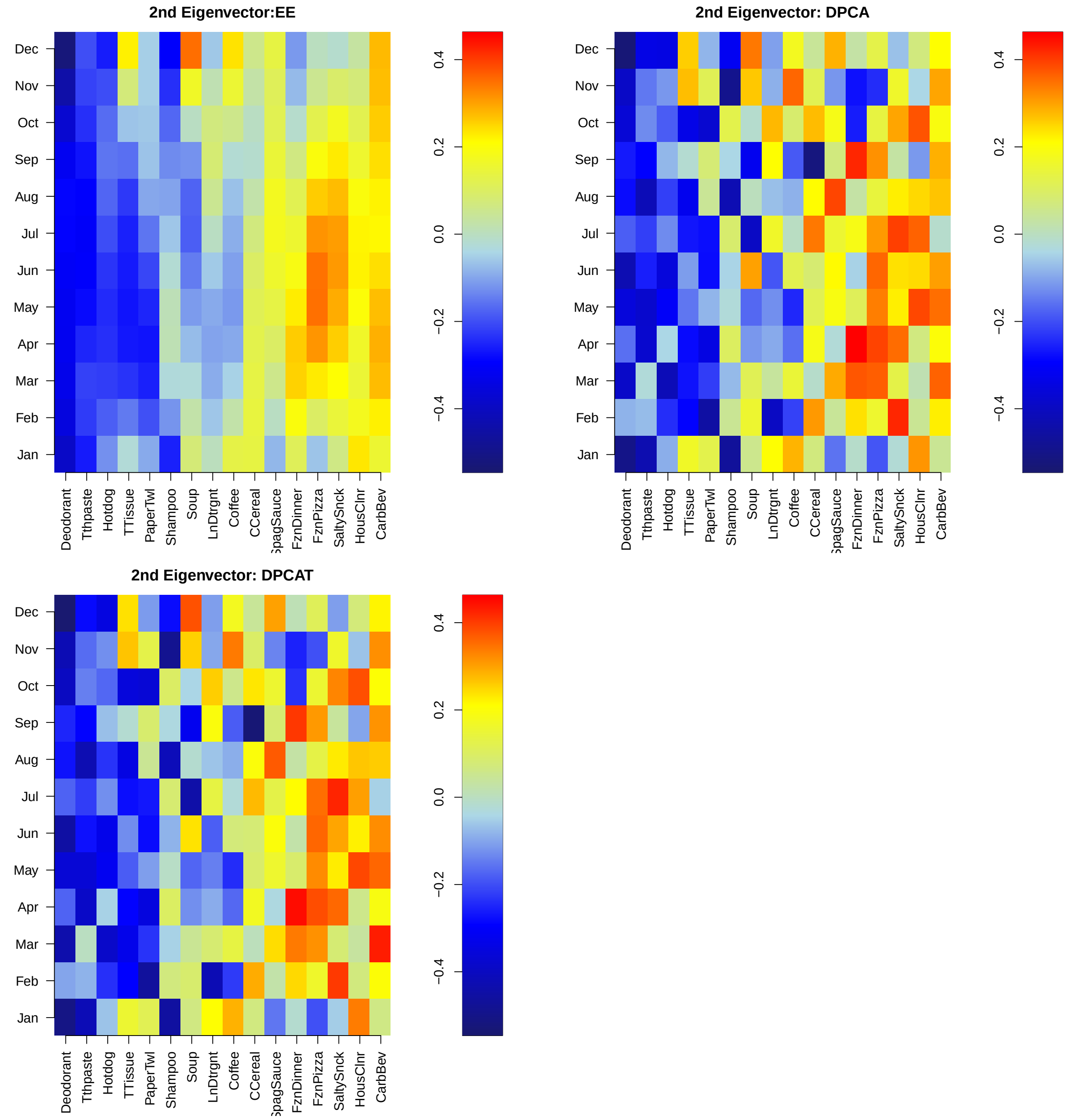


Figure 3.4: Heatmaps of the first eigenvector of the IRI panel data containing 16 products


2nd Eigenvector:EE

2nd Eigenvector: DPCA

2nd Eigenvector: DPCAT

Figure 3.5: Heatmaps of the second eigenvector of the IRI panel data containing 16 products

The EE heatmap in Figure 3.4 displays very smooth first eigenvector loadings. The second eigenvector heatmap of the EE estimates shows product groupings similar to the DPCA and DPCAT heatmaps. However, the EE heatmap attains smoothness and visual appeal that allows the product groups to be seen more clearly. The advantage of the EE heatmap is of visualization, which aids in interpretation, and this advantage is due in large part to the time-varying models of the eigenvectors. The discretized approaches fail to borrow neighboring information from the time points.

3.6 Discussion

In this chapter, we consider the longitudinal principal components analysis for binary data. It is an extension for handling a delicate and complicated data structure of which is ever-present in biomedical studies, imaging, and in the field of marketing and retail. The challenges of this work are multifaceted. Generating binary data with a particular association structure becomes increasingly difficult when that structure changes over time and when the number of variables is large. Then, the estimation approach is quite different from standard PCA and the measure of association, the odds ratio, is being used, which is not a standard approach in this longitudinal dimension reduction research area. To capture associations among the binary variables, we design a transformed odds ratio matrix with properties shared by a covariance-correlation matrix.

Our numerical studies indicate that the proposed method outperforms both discretized PCA approaches for eigenvalue estimation for the first and second components. The EE methodology also performs as well as a discretized approaches for eigenvector estimation, but enhances visualization through smoothness of the time-varying models of the eigenvectors. The real data application exemplifies that our proposed method can enhance visualization and in turn, interpretation.

The proposed methodology has not addressed the heterogeneity that may be present in the panelists through random effects, although the real data for the year 2007 has little demographic diversity. The random effects incorporation is quite challenging in the binary data setting, because extraction of the effects is not straight-forward. Additionally, the structure of binary data impose limitations on how much of the heterogeneous information is transferred from the marginal and joint probabilities. Care must be taken to ensure the marginal and joint probabilities correspond with the association structure, especially when randomness is introduced to the marginal probabilities. One future investigation to consider is the asymptotic properties of the eigenvalues.

References

Agresti, A. (2002). *Categorical Data Analysis*. Wiley, 2nd ed.

Akaike, H. (1974). A New look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*, 716–723.

Anderson, S. J., & Jones, R. H. (1995). Smoothing splines for longitudinal data. *Statistics in Medicine*, *14*, 1235–1248.

Becker, M. P., & Clogg, C. C. (1988). A Note on approximating correlations from odds ratios. *Sociological Methods and Research*, *16*, 407–424.

Bishop, Y. M., Fienberg, S. E., & Holland, P. W. (2007). *Discrete Multivariate Analysis*. Springer.

Bonett, D. G. (2007). Transforming odds ratios into correlations for meta-analytic research. *The American Psychologist*, *62*, 254–255.

Bradlow, E. T. (2002). Exploring repeated measures data sets for key features using principal components analysis. *International Journal of Research in Marketing*, *19*, 167–179.

Brillinger, D. R. (1981). *Time Series Data Analysis and Theory*. Holden-Day.

Bronnenberg, B. J., Kruger, M. W., & Mela, C. F. (2008). The IRI marketing data set. *Marketing Science*, *27*, 745–748.

Carey, V., Zeger, S. L., & Diggle, P. (1993). Modelling multivariate binary data with alternating logistic regressions. *Biometrika*, *80*, 517–526.

Collins, M., Dasgupta, S., & Schapire, R. E. (2002). A Generalization of principal component analysis to the exponential family. In T. Dietterich, S. Becker, & Z. Ghahramani (Eds.) *Advances in Neural Information Processing Systems 14*, (pp. 617–624). MIT Press.

Di, C. Z., Crainiceanu, C. M., Caffo, B. S., & Punjabi, N. M. (2009). Multilevel functional principal component analysis. *The Annals of Applied Statistics*, *3*, 458–488.

Diggle, P. T., & Verbyla, A. P. (1998). Nonparametric estimation of covariance structure in longitudinal data. *Biometrics*, *54*, 401–415.

Durbán, M., Harezlak, J., Wand, M. P., & Carroll, R. J. (2005). Simple fitting of subject-specific curves for longitudinal data. *Statistics in Medicine*, *24*, 1153–1167.

Elashoff, M., & Ryan, L. (2004). An EM algorithm for estimating equations. *Journal of Computational and Graphical Statistics*, *13*(1), 48–65.

Fader, P. S., & Lattin, J. M. (1993). Accounting for heterogeneity and nonstationarity in a cross-sectional model of consumer purchase behavior. *Marketing Science*, *12*, 304–317.

Fan, J., Huang, T., & Li, R. (2007). Analysis of longitudinal data with semiparametric estimation of covariance function. *Journal of the American Statistical Association*, *102*, 632–641.

Fan, J., & Wu, Y. (2008). Semiparametric estimation of covariance matrixes for longitudinal data. *Journal of the American Statistical Association*, *103*, 1520–1533.

Gardiner, J. C., Luo, Z., & Roman, L. A. (2009). Fixed effects, random effects, and gee: What are the differences? *Statistics in Medicine*, *28*, 221–239.

Gifi, A. (1990). *Nonlinear Multivariate Analysis*. Wiley.

Goodman, L. A. (1981). Association models and the bivariate normal for contingency tables with ordered categories. *Biometrika*, *68*, 347–355.

Greven, S., Crainiceanu, C. M., Caffo, B. S., & Reich, D. (2010). Longitudinal functional principal component analysis. *Electronic Journal of Statistics*, *4*, 1022–1054.

Hall, P., Müller, H. G., & Wang, J. L. (2006). Properties of principal component methods for functional and longitudinal data analysis. *The Annals of Statistics*, *34*, 1493–1517.

Huang, J. Z., Wu, C. O., & Zhou, L. (2004). Polynomial spline estimation and inference for varying coefficient models with longitudinal data. *Statistica Sinica*, *14*, 763–788.

Jain, D., Bass, F. M., & Chen, Y. M. (1990). Estimation of latent class models with heterogeneous choice probabilities: An application to market structuring. *Journal of Marketing Research*, *27*, 94–101.

Jiang, C. R., & Wang, J. L. (2010). Covariate adjusted functional principal component analysis. *The Annals of Statistics*, *38*, 1194–1226.

Kruger, M. W., & Pagni, D. (2008). *IRI Academic Data Set Description*. Information Resources Incorporated, 2.2 ed.

Ku, W., Storer, R. H., & Georgakis, C. (1995). Disturbance detection and isolation by dynamic principal component analysis. *Chemometrics and Intelligent Laboratory Systems*, *30*, 179–196.

Laird, N. M., & Ware, J. H. (1982). Random-effects models for longitudinal data. *Biometrics*, *38*, 963–974.

Landgraf, A. J., & Lee, Y. (2015). Dimensionality reduction for binary data through the projection of natural parameters. Tech. Rep. 890, The Ohio State University.

Lee, D., & Sompolinsky, H. (1999). Learning a continuous hidden variable model for binary data. In M. Kearns, S. Solla, & D. Cohn (Eds.) *Advances in Neural Information Processing Systems 11*, (pp. 515–521). MIT Press.

Lee, S., Huang, J. Z., & Hu, J. (2010). Sparse logistic principal component analysis for binary data. *The Annals of Applied Statistics*, *4*, 1579–1601.

Leeuw, J. D. (2006). Principal component analysis of binary data by iterated singular value decomposition. *Computational Statistics and Data Analysis*, *50*, 21–39.

Leisch, F., Weingessel, A., & Hornik, K. (1998). On the generation of correlated artificial binary data. Tech. Rep. 13, WU Vienna University of Economics and Business.

Liang, H., & Xiao, Y. (2006). Penalized splines for longitudinal data with an application in AIDS studies. *Journal of Modern Applied Statistical Methods*, *5*, 130–139.

Liang, K. Y., & Zeger, S. L. (1986). Longitudinal data analysis using generalized linear models. *Biometrika*, *73*, 13–22.

Lipsitz, S. R., Laird, N. M., & Harrington, D. P. (1991). Generalized estimating equations for correlated binary data: using the odds ratio as a measure of association. *Biometrika*, *78*, 153–160.

Luo, R., Wang, H., & Tsai, C. L. (2009). Contour projected dimension reduction. *The Annals of Statistics*, *37*, 3743–3778.

Maadooliat, M., Pourahmadi, M., & Huang, J. Z. (2013). Robust estimation of the correlation matrix of longitudinal data. *Statistics and Computing*, *23*, 17–28.

Mosteller, F. (1968). Association and estimation in contingency tables. *Journal of the American Statistical Association*, *63*, 1–28.

Peña, D., & Yohai, V. J. (2016). Generalized dynamic principal components. *Journal of the American Statistical Association*, *111*, 1121–1131.

Prentice, R. L. (1988). Correlated binary regression with covariates specific to each binary observation. *Biometrics*, *44*, 1033–1048.

Ramsay, J. O., & Silverman, B. W. (2005). *Functional Data Analysis*. New York: Springer, 2nd ed.

Rossi, P. E., & Allenby, G. M. (1993). A Bayesian approach to estimating household parameters. *Journal of Marketing Research*, *30*, 171–182.

Schein, A., Saul, L., & Ungar, L. (2003). A Generalized linear model for principal component analysis of binary data. In C. M. Bishop, & B. J. Frey (Eds.) *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, (pp. 14–21).

Schwarz, G. (1978). Estimating the dimension of a model. *The Annals of Statistics*, *6*, 461–464.

Sun, Y., Zhang, W., & Tong, H. (2007). Estimation of the covariance matrix of random effects in longitudinal studies. *The Annals of Statistics*, *35*, 2795–2814.

Wu, W. B., & Pourahmadi, M. (2003). Nonparametric estimation of large covariance matrices of longitudinal data. *Biometrika*, *90*, 831–844.

Xu, R., Faig, W., & Natarajan, L. (2012). The analysis of multivariate longitudinal data. In *The Statistical Analysis of Multi-Outcome Data Workshop*.

Xue, L., & Liang, H. (2009). Polynomial spline estimation for a generalized additive coefficient model. *Scandinavian Journal of Statistics, Theory and Applications*, *37*, 26–46.

Yao, F., Müller, H. G., & Wang, J. L. (2005). Functional data analysis for sparse longitudinal data. *Journal of the American Statistical Association*, *100*, 577–590.