

Population Value Decomposition, a Framework for the Analysis of Image Populations

Ciprian M. CRAINICEANU, Brian S. CAFFO, Sheng LUO, Vadim M. ZIPUNNIKOV, and Naresh M. PUNJABI

Images, often stored in multidimensional arrays, are fast becoming ubiquitous in medical and public health research. Analyzing populations of images is a statistical problem that raises a host of daunting challenges. The most significant challenge is the massive size of the datasets incorporating images recorded for hundreds or thousands of subjects at multiple visits. We introduce the population value decomposition (PVD), a general method for simultaneous dimensionality reduction of large populations of massive images. We show how PVD can be seamlessly incorporated into statistical modeling, leading to a new, transparent, and rapid inferential framework. Our PVD methodology was motivated by and applied to the Sleep Heart Health Study, the largest community-based cohort study of sleep containing more than 85 billion observations on thousands of subjects at two visits. This article has supplementary material online.

KEY WORDS: Electroencephalography; Signal extraction.

1. INTRODUCTION

We start by considering the following thought experiment using data displayed in Figure 1. Inspect the plot for a minute and try to remember it as closely as possible; ignore the meaning of the data and try to answer the following question: “How many features (patterns) from this plot do you remember?” Now, consider the case when you are flipping through thousands of similar images and try to answer the slightly modified question: “How many common features from all these plots do you remember?” Regardless of who is answering either question, the answer for this dataset seems to be invariably between 3 and 25.

To mathematically represent this experiment, we introduce the population value decomposition (PVD) of a sample of matrices. In this section we focus on providing the intuition. We introduce the formal definition in Section 3. Consider a sample \mathbf{Y}_i , $i = 1, \dots, n$, of matrices of size $F \times T$, where F , T , or both are very large. Suppose that the following approximate decomposition holds:

$$\mathbf{Y}_i \simeq \mathbf{P}\mathbf{V}_i\mathbf{D}, \quad (1)$$

where \mathbf{P} and \mathbf{D} are population-specific matrices of size $F \times A$ and $B \times T$, respectively. If A or B is much smaller than F and T , then equation (1) provides a useful representation of a sample of images. Indeed, the “subject-level” features of the image are coded in the low-dimensional matrix \mathbf{V}_i , whereas the “population frame of reference” is coded in the matrices \mathbf{P} and \mathbf{D} . Im-

portant differences between PVD and the singular value decomposition (SVD) are that (a) PVD applies to a sample of images not just one image; (b) the matrices \mathbf{P} and \mathbf{D} are population-, not subject-, specific; and (c) the matrix \mathbf{V}_i is not necessarily diagonal.

With this new perspective, we can revisit Figure 1 to provide a reasonable explanation for how our vision and memory might work. First, the image can be decomposed using a partition of frequencies and time in several subintervals. A checkerboard-like partition of the image is then obtained by building the two-dimensional partitions from the one-dimensional partitions. The size of the partitions is then mentally adjusted to match the observed complexity in the image. When decomposing a sample of images, the thought process is similar, except that some adjustments are made on the fly to ensure maximum encoding of information with a minimum amount of memory. Some smoothing across subjects further improves efficiency by taking advantage of observed patterns across subjects. A mathematical representation of this process would be to consider subject-specific matrices, \mathbf{P} and \mathbf{D} , with columns and rows corresponding to the one-dimensional partitions. The matrix \mathbf{V}_i is then constructed by taking the average of the image in the induced two-dimensional subpartition. Our methods transfer this empirical reasoning into a statistical framework. This process is crucial for the following reasons:

1. Reducing massive images to a manageable set of coefficients that are comparable across subjects is of primary importance. Note that Figure 1 displays 57,000 observations, only a fraction of the total of 228,160 observations of the original uncut image. The matrix \mathbf{V}_i typically contains fewer than 100 entries.
2. Statistical inference on samples of images is typically difficult. For example, the Sleep Heart Health Study (SHHS), described in Section 2, contains one image for each of two visits for more than 3000 subjects. The total number of observations used in the analysis presented in Section 5 exceeds 450,000,000. In contrast, replacing \mathbf{Y}_i by \mathbf{V}_i reduces the dataset to 600,000 observations.

Ciprian M. Crainiceanu is Associate Professor (E-mail: crcrainic@jhsph.edu) and Brian S. Caffo is Associate Professor (E-mail: bcaffo@jhsph.edu), Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205. Sheng Luo is Assistant Professor, Division of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, 1200 Herman Pressler Dr, Houston, TX 77030 (E-mail: sheng.t.luo@uth.tmc.edu). Vadim M. Zipunnikov is Post Doctoral Fellow, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: vzipunni@jhsph.edu). Naresh M. Punjabi is Professor, Department of Epidemiology, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: punjabi@jhmi.edu). This research was supported by award R01NS060910 from the National Institute of Neurological Disorders and Stroke. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institute of Neurological Disorders and Stroke or the National Institutes of Health. The authors gratefully acknowledge the suggestions and comments of the associate editor and two anonymous reviewers. Any remaining errors are the sole responsibility of the authors.

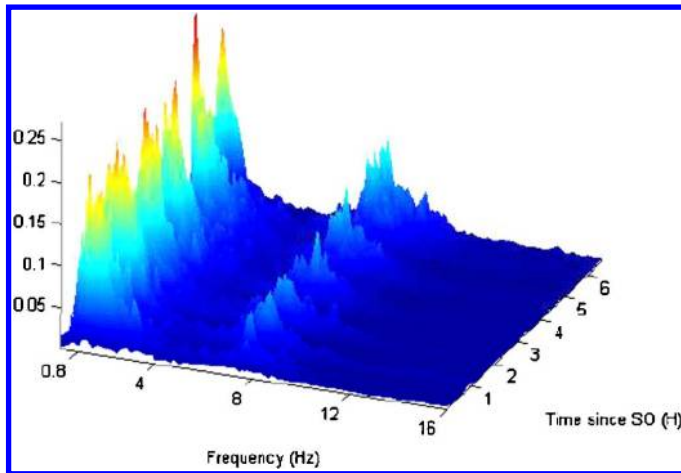


Figure 1. Frequency by time percent power for the sleep electroencephalography data for one subject. The Y -axis is time in hours since sleep onset, where each row corresponds to a 30-second interval. The X -axis is the frequency from 0.2 Hz to 16 Hz. The other frequencies were not shown because they are “quiet”; that is, the proportion of power in those frequencies is very small.

3. Obtaining the coefficient matrix \mathbf{V}_i is easy once \mathbf{P} and \mathbf{D} are known. Using the entries of \mathbf{V}_i as predictors in a regression context is then straightforward; this strategy was used by Caffo et al. (2010) for predicting the risk of Alzheimer’s disease using functional magnetic resonance imaging (fMRI).
4. Modeling of the coefficients \mathbf{V}_i can replace modeling of the images \mathbf{Y}_i . In Section 3 we show that the Karhunen–Loève (KL) decomposition (Loève 1945; Karhunen 1947) of a sample of images can be approximated by using a computationally tractable algorithm based on the coefficients \mathbf{V}_i . This avoids the intractable problem of calculating and diagonalizing very large covariance operators.

The article is organized as follows. In Section 2 we introduce the SHHS and the associated methodological challenges. In Section 3 we introduce the PVD and describe its application to the analysis of samples of images. Section 4 provides simulations, and Section 5 provides extensive results for the analysis of the SHHS dataset. Section 6 presents some unresolved methodological and applied problems.

2. THE CASE STUDY

The SHHS is a landmark study of sleep and its impacts on health outcomes. A detailed description of the SHHS has been provided by Quan et al. (1997), Crainiceanu, Staicu, and Di (2009), and Di et al. (2009). The SHHS is a multicenter cohort study that used the resources of existing epidemiologic cohorts and conducted further data collection, including measurements of sleep and breathing. Between 1995 and 1997, in-home polysomnography (PSG) data were collected from a sample of 6441 participants. A PSG is a quasi-continuous multichannel recording of physiological signals acquired during sleep that include two surface electroencephalograms (EEG). After the baseline visit, a second SHHS follow-up visit was undertaken between 1999 and 2003 that included a repeat PSG. A total of 4361 participants completed a repeat in-home PSG. The main

goals of the SHHS were to quantify the natural variability of complex measurements of sleep in a large community cohort, to identify potential biomarkers of cardiovascular and respiratory disease, and to study the association between these biomarkers and various health outcomes, including sleep apnea, cardiovascular disease, and mortality.

Our focus on sleep EEG is based on the expectation that a spectral analysis of electroneural data will provide a set of reliable, reproducible, and easily calculated biomarkers. Currently, quantification of sleep in most research settings is based on a visual-based counting process that attempts to identify brief fluctuations in the EEG (i.e., arousals) and classify time-varying electrical phenomena into discrete sleep stages. Although metrics of sleep based on visual scoring have been shown to have clinically meaningful associations, they are subject to several limitations. First, interpretation of scoring criteria and lack of experience can increase error variance in the derived measures of sleep. For example, even with the most rigorous training and certification requirements, technicians in the large multicenter SHHS were noted to have an intraclass correlation coefficient of 0.54 for scoring arousals (Whitney et al. 1998). Second, there is a paucity of definitions for classifying EEG patterns in disease states, given that the criteria were developed primarily for normal sleep. Third, many of the criteria do not have a biological basis. For example, an amplitude criterion of $75 \mu V$ is used for the identification of slow waves (Redline et al. 1998), and a shift in EEG frequency for at least 3 seconds is required for identifying an arousal. Neither of these criteria is evidence-based. Fourth, visually scored data are described with summary statistics of different sleep stages, resulting in complete loss of temporal information. Finally, visual assessment of overt changes in the EEG provides a limited view of sleep neurobiology. In the setting of sleep-disordered breathing, a disorder characterized by repetitive arousals, visual characterization of sleep structure cannot capture common EEG transients. Thus it is not surprising that previous studies have found weak correlations between conventional sleep stage distributions, arousal frequency, and clinical symptoms (Guilleminault et al. 1988; Cheshire et al. 1992; Martin et al. 1997; Kingshott et al. 1998). Power spectral analysis provides an alternate and automatic means for the studying of the dynamics of the sleep EEG, often demonstrating global trends in EEG power density during the night. Although quantitative analysis of EEG has been used in sleep medicine, its use has focused on characterizing EEG activity during sleep in disease states or in experimental conditions. A limited number of studies have undertaken analyses of the EEG throughout the entire night to delineate the role of disturbed sleep structure in cognitive performance and daytime alertness. However, most of these studies are based on samples of fewer than 50 subjects and thus are not generalizable to the general population. Finally, there are only isolated reports using quantitative techniques to characterize EEG during sleep as a function of age and sex, with the largest study consisting of only 100 subjects.

To address these problems, here we focus on the statistical modeling of the time-varying spectral representation of the subject-specific raw EEG signal. The main components of this strategy are as follows:

- C1. RAW SIGNAL \mapsto IMAGE (FFT).

- C2. FREQUENCY \times TIME IMAGE \mapsto IMAGE CHARACTERISTICS (PVD).
 C3. ANALYZE IMAGE CHARACTERISTICS (FPCA and MFPCA).

Component C1 is a well-established data transformation and compression technique at the subject level. Even though we make no methodological contributions in C1, its presentation is necessary to understand the application. The technical details of C1 are provided in Sections 2.1 and 2.2. Component C2, our main contribution, is a second level of compression at the population level. This is an essential component when images are massive, but could be eliminated when images are small. Methods for C2 are presented in Section 3. Component C3, our second contribution, generalizes multilevel functional principal component analysis (MFPCA) (Di et al. 2009) to multilevel samples of images. Technical details for C3 are presented in Sections 3.2.1 and 3.2.2.

2.1 Fourier Transformations and Local Spectra

In the SHHS, EEG sampled at a frequency of 125 Hz (125 observations per second) and an 8-hour sleep interval will contain $U = 125 \text{ Hz} \times 60'' \times 60' \times 8\text{h} = 3,600,000$ observations. A standard data-reduction step for EEG is to partition the entire time series into adjacent 5-second intervals. The 5-second intervals are further aggregated into adjacent groups of six intervals for a total time of 30 seconds. These adjacent 30-second intervals are called epochs. Thus, for an 8-hour sleep interval, the number of 5-second intervals is $U/625 = 5760$, and the number of epochs is $T = U/(625 \times 6) = 960$. In general, U and T are subject- and visit-specific, because the duration of sleep is subject- and visit-specific.

Now consider the partitioned data and let $x_{th}(n)$ denote the n th observation of the raw EEG signal, $n = 1, \dots, N = 625$, in the h th 5-second interval, $h = 1, \dots, H = 6$, of the t th 30-second epoch, $t = 1, \dots, T$. In each 5-second window, data are first centered around their mean. We continue to denote the centered data by $x_{th}(n)$. We then apply a Hann weighting window to the data, which replaces the $x_{th}(n)$ with $w(n)x_{th}(n)$, where $w(n) = 0.5 - 0.5 \cos\{2\pi n/(N - 1)\}$. To these data we apply a Fourier transform and obtain $X_{th}(k) = \sum_{n=0}^{N-1} w(n)x_{th}(n)e^{-2\pi kni/N}$ for $k = 0, \dots, N - 1$. Here $X_{th}(k)$ are the Fourier coefficients corresponding to the h th 5-second interval of the t th epoch and frequency $f = k/5$. For each frequency, $f = k/5$, and 30-second epoch, t , we calculate $P(f, t) = \frac{1}{H} \sum_{h=1}^H |X_{th}(k)|^2$ the average over the $H = 6$ 5-second intervals of the square of the Fourier coefficients. More precisely, $P(f, t) = \frac{1}{H} \sum_{h=1}^H |\sum_{n=0}^{N-1} w(n)x_{th}(n)e^{-2\pi kni/N}|^2$. Total power in a spectral window can be calculated as $\text{PS}_b(t) = \sum_{f \in D_b} P(f, t)$, where D_b denotes the spectral window (collection of frequencies) indexed by b .

In this article we focus on $P(f, t)$ and treat it as a bivariate function of frequency f (expressed in Hz) and time t (expressed in epochs). The power in a spectral window, $\text{PS}_b(t)$, was analyzed by Crainiceanu et al. (2009) and Di et al. (2009). Here we concentrate on methods that generalize the spirit of the methods of Di et al. (2009), while focusing on solutions to the much more ambitious problem of population-level analysis of images. Before describing our methods, we provide more insight into the interpretation of the frequency–time analysis.

2.2 Insight Into the Discrete Fourier Transform

First, note that the inverse Fourier transform is $w(n)x_{th}(n) = \frac{1}{N} \sum_{k=0}^{N-1} X_{th}(k)e^{2\pi kni/N}$, and the Fourier coefficients are the projections of the data on the orthonormal basis $e^{2\pi kni/N}$, $k = 0, \dots, N - 1$. Thus a larger (in absolute value) $X_{th}(k)$ corresponds to a larger contribution of the frequency $k/5$ to explaining the raw signal. Parseval's theorem provides the following equality: $\sum_{n=0}^{N-1} |w(n)x_{th}(n)|^2 = \frac{1}{N} \sum_{k=0}^{N-1} |X_{th}(k)|^2$. The left side of the equation is the total observed variance of the raw signal, and the right side provides an ANOVA-like decomposition of the variance as a sum of $|X_{th}(k)|^2$. This is the reason why $|X_{th}(k)|^2$ is interpreted as the part of variability explained by frequency $f = k/5$. In signal processing $|X_{th}(k)|^2$ is called the power of the signal in frequency $f = k/5$.

We complete our preprocessing of the data by normalizing the observed power as $Y(f, t) = P(f, t) / \sum_f P(f, t)$, which is the ‘‘proportion’’ of observed variability of the EEG signal attributable to frequency f in epoch t . In practice, for surface EEG, frequencies above 32 Hz make a negligible contribution to the total power, and we define $Y(f, t) = P(f, t) / \sum_{f \leq 32} P(f, t)$. We call $Y(f, t)$ the normalized power, and the true signal measured by $Y(f, t)$ the frequency-by-time image of the EEG time series.

Figure 1 shows a frequency-by-time plot of $Y(f, t)$ for one subject who slept for more than 6 hours. The X-axis is the frequency from 0.2 Hz to 16 Hz. The other frequencies were not shown because they are ‘‘quiet’’; that is, the proportion of power in those frequencies is very small. The Y-axis is time in hours since sleep onset, with each row corresponding to a 30-second interval. Note that a large proportion of the observed variability is in the low-frequency range, say [0.8–4.0 Hz]. This range, known as the δ -power band, is traditionally analyzed in sleep research by averaging the frequency values across all frequencies in the range. Another interesting range of frequencies is roughly between 5 and 10 Hz, with the proportion of power quickly converging to 0 beyond 12–14 Hz. The [5.0–10.0 Hz] range is not standard in EEG research. Instead, research tends to focus on the θ [4.1–8.0 Hz] and α [8.1–13.0 Hz] bands. A careful inspection of the plot will reveal that in the δ , θ , and α frequency ranges the proportion of power tends to show cycles across time. (Note the wavy pattern of the data as time progresses from sleep onset.) Although this may be less clear from Figure 1, the behavior of the δ band tends to be negatively correlated with θ and α bands. This occurs because there is a natural trade-off between slow and fast neuronal firing.

3. POPULATION VALUE DECOMPOSITION

In this section we introduce a population-level data compression that allows the coefficients of each image to be comparable and interpretable across images. If \mathbf{Y}_i , $i = 1, \dots, n$, is a sample of $F \times T$ -dimensional images, then a PVD is

$$\mathbf{Y}_i = \mathbf{P}\mathbf{V}_i\mathbf{D} + \mathbf{E}_i, \quad (2)$$

where \mathbf{P} and \mathbf{D} are population-specific matrices of size $F \times A$ and $B \times T$, \mathbf{V}_i is an $A \times B$ -dimensional matrix of subject-specific coefficients, and \mathbf{E}_i is an $F \times T$ -dimensional matrix of residuals. Many different decompositions of type (2) exist. Consider, for example, any two full-rank matrices \mathbf{P} and \mathbf{D} , where $A < F$

and $B < T$. Equation (2) can be written in vector format as follows. Denote by $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i^T)$, $\mathbf{v}_i = \text{vec}(\mathbf{V}_i^T)$, $\boldsymbol{\epsilon}_i = \text{vec}(\boldsymbol{\epsilon}_i^T)$ the column vectors obtained by stacking the row vectors of \mathbf{Y}_i , \mathbf{V}_i , and $\boldsymbol{\epsilon}_i$, respectively. If $\mathbf{X} = \mathbf{P} \otimes \mathbf{D}^T$ is the $FT \times AB$ Kronecker product of matrices \mathbf{P} and \mathbf{D} , then equation (2) becomes the following standard regression: $\mathbf{y}_i = \mathbf{X}\mathbf{v}_i + \boldsymbol{\epsilon}_i$. Thus a least squares estimator of \mathbf{v}_i is $\hat{\mathbf{v}}_i = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}_i$. This provides a simple recipe for obtaining the subject-specific scores, \mathbf{v}_i or, equivalently, \mathbf{V}_i , once the matrices \mathbf{P} and \mathbf{D} are fixed. The scores can be used in standard statistical models either for prediction or for association studies. Note that $\mathbf{X}'\mathbf{X}$ is a low-dimensional matrix that is easily inverted. Moreover, all calculations can be done on even very large images by partitioning files into subfiles and using block-matrix computations.

3.1 Default Population Value Decomposition

There are many types of PVDs, and definitions can and will change in particular applications. In this section we introduce our default procedure, which is inspired by the subject-specific SVD and by the thought experiment described in Section 1. Consider the case where the SVD can be obtained for every subject-specific image. This can be done in all applications that we are aware of, including the SHHS and fMRI studies (see Caffo et al. 2010 for an example).

For each subject, let $\mathbf{Y}_i = \mathbf{U}_i\boldsymbol{\Sigma}_i\mathbf{V}_i^T$ be the SVD of the image. If \mathbf{U}_i and \mathbf{V}_i were the same across all subjects, then the SVD would be the default PVD. However, in practice \mathbf{U}_i and \mathbf{V}_i will tend to vary from person to person. Mimicking the thought process described in Section 1, we try to find the common features across subjects among the column vectors of the \mathbf{U}_i and \mathbf{V}_i matrices.

We start by considering the $F \times L_i$ -dimensional matrix \mathbf{U}_{L_i} , consisting of the first L_i columns of the matrix \mathbf{U}_i , and the $T \times R_i$ -dimensional matrix, consisting of the first R_i columns of the matrix \mathbf{V}_i . The choices of L_i and R_i could be based on various criteria, including variance explained, signal-to-noise ratios, and practical considerations. This is not a major concern in this article.

We focus on \mathbf{U}_{L_i} ; a similar procedure is applied to \mathbf{V}_{R_i} . Consider the $F \times L$ -dimensional matrix $\mathbf{U} = [\mathbf{U}_{L_1} | \dots | \mathbf{U}_{L_n}]$, where $L = (\sum_{i=1}^n L_i)$, obtained by horizontally binding the \mathbf{U}_{L_i} matrices across subjects. The space spanned by the columns of \mathbf{U} is a subspace of \mathbb{R}^F and contains subject-specific left eigenvectors that explain most of the observed variability. Although these vectors are not identical, they will be similar if images share common features. Thus, we propose applying PCA to the matrix $\mathbf{U}\mathbf{U}^T$ to obtain the main directions of variation in the column space of \mathbf{U} . Let \mathbf{P} be the $F \times A$ -dimensional matrix formed with the first A eigenvectors of $\mathbf{U}\mathbf{U}^T$ as columns, where A is chosen to ensure that a certain percentage of variability is explained. Then the matrix \mathbf{U} is approximated via the projection equation $\mathbf{U} \approx \mathbf{P}(\mathbf{P}^T\mathbf{U})$. At the subject level, we obtain $\mathbf{U}_{L_i} \approx \mathbf{P}(\mathbf{P}^T\mathbf{U}_{L_i})$. This approximation becomes a tautological equality if $A = F$, that is, if we use the entire eigenbasis. Similar approximations can be obtained using any orthonormal basis; we prefer the eigenbasis for our default procedure, because it is parsimonious. We similarly obtain \mathbf{D}^T , a $T \times B$ -dimensional matrix of the first eigenvectors of the matrix $\mathbf{V}\mathbf{V}^T$, where $\mathbf{V} = [\mathbf{V}_{R_1} | \dots | \mathbf{V}_{R_n}]$. We have the similar

approximation $\mathbf{V} \approx \mathbf{D}(\mathbf{D}^T\mathbf{V})$. At the subject level, we obtain $\mathbf{V}_{R_i} \approx \mathbf{D}^T(\mathbf{D}\mathbf{V}_{R_i})$. We conclude that PVD is a two-step approximation process for all images that can be summarized as follows:

$$\begin{aligned} \mathbf{Y}_i &= \mathbf{U}_i\boldsymbol{\Sigma}_i\mathbf{V}_i^T \approx \mathbf{U}_{L_i}\boldsymbol{\Sigma}_{L_i,R_i}\mathbf{V}_{R_i}^T \\ &\approx \mathbf{P}\{(\mathbf{P}^T\mathbf{U}_{L_i})\boldsymbol{\Sigma}_{L_i,R_i}(\mathbf{V}_{R_i}^T\mathbf{D}^T)\}\mathbf{D}, \end{aligned} \quad (3)$$

where \mathbf{U}_{L_i} and \mathbf{V}_{R_i} are obtained by retaining the first L_i and R_i columns from the matrices \mathbf{U}_i and \mathbf{V}_i , respectively, and $\boldsymbol{\Sigma}_{L_i,R_i}$ is obtained by retaining the first L_i rows and R_i columns from the matrix $\boldsymbol{\Sigma}_i$. The first approximation of the image \mathbf{Y}_i , given in the first row in equation (3), is obtained by retaining the left and right eigenvectors that explain most of the observed variability at the subject level. The second approximation, shown in the second row in equation (3), is obtained by projecting the subject-specific left and right eigenvectors on the corresponding population-specific eigenvectors.

If we denote by $\mathbf{V}_i = (\mathbf{P}^T\mathbf{U}_{L_i})\boldsymbol{\Sigma}_{L_i,R_i}(\mathbf{V}_{R_i}^T\mathbf{D}^T)$, we then obtain the PVD equation (2). This formula shows that \mathbf{V}_i generally will not be a diagonal matrix even though $\boldsymbol{\Sigma}_{L_i,R_i}$ is. This is one of the fundamental differences between SVD and PVD. Note that all approximations can be trivially transformed into equalities. For example, choosing $L_i = F$ and $R_i = T$ will ensure equality in the first approximation, whereas choosing $A = F$ and $B = T$ will ensure equality in the second equation. From a practical perspective, these cases are not of scientific importance, because data compression would not be achieved. However, our focus is on parsimony, not on perfection of the approximation. The choices of L_i , R_i , A , and B could be based on various criteria, including variance explained, signal-to-noise ratios, and practical considerations. In this article we use thresholds for the percent variance explained.

Calculations in this section are possible because of the following matrix algebra trick. We summarize this trick, which allows calculation of SVD for very large matrices as long as one of the dimensions is not much larger than a few thousands.

Suppose that $\mathbf{Y} = \mathbf{U}\mathbf{D}\mathbf{V}^T$ is the SVD decomposition of an $F \times T$ -dimensional matrix where, say, F is very large and T is moderate. Then \mathbf{D} and \mathbf{V} can be obtained from the spectral decomposition of the $T \times T$ -dimensional matrix $\mathbf{Y}^T\mathbf{Y} = \mathbf{V}\mathbf{D}^2\mathbf{V}^T$. The \mathbf{U} matrix can then be obtained from $\mathbf{U} = \mathbf{Y}\mathbf{V}\mathbf{D}^{-1}$.

3.2 Functional Statistical Modeling

An immediate application of PVD is to use the entries' subject-specific matrix \mathbf{V}_i as predictors. For this purpose, we can use a range of strategies, from using one entry at a time to using groups of entries or selection or averaging algorithms based on prediction performance. The first example of such an approach is that of Caffo et al. (2010), who found empirical evidence of alternative connectivity in clinically asymptomatic subjects at risk for Alzheimer's disease compared with controls. The authors used PVD with a 5×5 -dimensional \mathbf{V}_i , boosting to identify important predictors.

Here we focus on how PVD can be used to conduct nonparametric analysis of the images themselves. Specifically, we are interested in approximating the Karhunen–Loève (KL) decomposition (Loève 1945; Karhunen 1947) of a sample of images. More precisely, if $\mathbf{y}_i = \text{vec}(\mathbf{Y}_i^T)$ is the vector obtained by stacking the rows of the matrix \mathbf{Y}_i , then we would like to obtain a de-

Table 1. Computing time (in minutes) for functional data analysis of samples of images for various number of grid points in the time and frequency domains

N_{freq}	N_{time}					
	20	40	60	80	100	120
8	0.1	0.3	0.7	1.3	2.1	3.1
16	0.3	1.4	3.0	5.7	8.7	13.2
32	1.3	5.5	12.9	19.4	32.9	49.8
64	4.7	20.5	51.8	97.3	176.0	496.5
128	21.5	100.7	467.0	681.0	1195.6	2097.1

composition of the type $\mathbf{y}_i = \sum_{k=1}^K \xi_{ik} \Phi_k + \mathbf{e}_i$, where Φ_k are the orthonormal eigenfunctions of the covariance operator, \mathbf{K}_y , of the process \mathbf{y} and ξ_{ik} are the random uncorrelated scores of subject i on eigenfunction k , and \mathbf{e}_i is an error process that could be, but typically is not, 0. A direct, or brute force, functional approach to this problem would require the calculation, diagonalization, and smoothing of $\widehat{\mathbf{K}}_y$, which is a $FT \times FT$ -dimensional matrix. This can be done relatively easily when FT is small, but it becomes computationally prohibitive as FT increases. For example, in the SHHS we could deal with data for all frequencies in the δ band ($F = 17$) and 1 hour of sleep ($T = 120$) as computational complexity increases sharply both with respect to F and T . Indeed, computational complexity is $O(F^3 T^3)$, and storage requirements are $O(F^2 T^2)$. Table 1 displays the computing time required by the direct functional approach using a personal computer with dual-core processors with 3 GHz CPU and 8 Gb RAM. Computing time increases steeply with T and F making the approach impractical when both exceed approximately 100. Thus, developing methods that accelerate the analysis is essential. The PVD offers one solution.

3.2.1 Functional Principal Component Analysis of Samples of Images. To avoid the brute force approach, we propose to first obtain the spectral decomposition of the vectors \mathbf{v}_i or, equivalently, of the corresponding matrix \mathbf{V}_i . As discussed earlier, we expect that in most applications the matrix \mathbf{V}_i will have far fewer than 500 entries; thus obtaining a decomposition for \mathbf{v}_i instead of \mathbf{y}_i is not only achievable, but very fast. The KL expansion for the \mathbf{v}_i process can be easily obtained (see, e.g., Yao, Müller, and Wang 2005; Ramsay and Silverman 2006). The expansion can be written directly in matrix format as

$$\mathbf{V}_i = \sum_{k=1}^K \xi_{ik} \phi_k + \boldsymbol{\eta}_i, \quad (4)$$

where ϕ_k are the eigenvectors of the process \mathbf{v} written as an $A \times B$ matrix, $\boldsymbol{\eta}_i$ is a noise process, and ξ_{ik} are mutually uncorrelated random coefficients. Here all vector to matrix transformations follow the same rules of the transformations $\mathbf{v}_i \leftrightarrow \mathbf{V}_i$. By left and right multiplication in equation (4) with the \mathbf{P} and \mathbf{D} matrices, respectively, we obtain the following decomposition of the sample of images:

$$\begin{aligned} \mathbf{Y}_i &= \sum_{k=1}^K \xi_{ik} \mathbf{P} \phi_k \mathbf{D} + \mathbf{P} \boldsymbol{\eta}_i \mathbf{D} + \mathbf{E}_i \\ &= \sum_{k=1}^K \xi_{ik} \Phi_k + \mathbf{e}_i, \end{aligned} \quad (5)$$

where $\Phi_k = \mathbf{P} \phi_k \mathbf{D}$ is an $F \times T$ -dimensional image, and $\mathbf{e}_i = \mathbf{P} \boldsymbol{\eta}_i \mathbf{D} + \mathbf{E}_i$ is an $F \times T$ noise process. These results provide a constructive recipe for image decomposition with the following simple steps: (a) Obtain \mathbf{P} , \mathbf{D} , and \mathbf{V}_i matrices, as described in Section 3.1; (b) obtain the eigenfunctions ϕ_k of the covariance operator of \mathbf{V}_i ; (c) obtain the scores ξ_{ik} from the mixed-effects model (4); and (d) obtain the basis for the image expansion $\Phi_k = \mathbf{P} \phi_k \mathbf{D}$. The following results provide the theoretical insights supporting this procedure.

Theorem 1. Suppose that \mathbf{P} is a matrix obtained by column binding A orthonormal eigenvectors of size $F \times 1$ and \mathbf{D} is a matrix obtained by row binding B orthonormal eigenvectors of size $1 \times T$. Then the following results hold: (a) The vector version of the eigenimages $\Phi_k = \mathbf{P} \phi_k \mathbf{D}$ are orthonormal in \mathbb{R}^{FT} , and (b) the scores ξ_{ik} are exactly the same in equations (4) and (5).

3.2.2 Multilevel Functional Principal Component Analysis of Samples of Images. There are many studies, including our own SHHS, in which images have a natural multilevel structure. This occurs, for example, when image data are clustered within the subjects or data are observed at multiple visits within the same subject. PVD provides a natural way of working with the data in this context. Suppose that \mathbf{Y}_{ij} are images observed on subject i at time j , and assume that $\mathbf{Y}_{ij} = \mathbf{P} \mathbf{V}_{ij} \mathbf{D} + \mathbf{E}_{ij}$ is the default PVD for the entire collection of images. Using the MF-PCA methodology introduced by Di et al. (2009) and further developed by Crainiceanu, Staicu, and Di (2009), we can decompose the \mathbf{V} process into subject- and subject/visit-specific components. More precisely,

$$\mathbf{V}_{ij} = \sum_{k=1}^K \xi_{ik} \phi_k^{(1)} + \sum_{l=1}^L \zeta_{ijl} \phi_l^{(2)} + \boldsymbol{\eta}_i, \quad (6)$$

where $\phi_k^{(1)}$ are mutually orthonormal subject-specific (or level 1) eigenvectors, $\phi_l^{(2)}$ are mutually orthonormal subject/visit-specific (or level 2) eigenvectors, and $\boldsymbol{\eta}_i$ is a noise process. The level 1 and 2 eigenvectors are required to be orthonormal within the level, not across levels. The subject-specific scores, ξ_{ik} , and the subject-/visit-specific scores, ζ_{ijl} , are assumed to be mutually uncorrelated random coefficients. Just as in the case of a cross-sectional sample of images, we can multiply the equation (6) with the matrix \mathbf{P} at the left and \mathbf{D} at the right. We obtain the following model for a sample of images with a multilevel structure:

$$\begin{aligned} \mathbf{Y}_{ij} &= \sum_{k=1}^K \xi_{ik} \mathbf{P} \phi_k^{(1)} \mathbf{D} + \sum_{l=1}^L \zeta_{ijl} \mathbf{P} \phi_l^{(2)} \mathbf{D} + \mathbf{P} \boldsymbol{\eta}_i \mathbf{D} + \mathbf{E}_i \\ &= \sum_{k=1}^K \xi_{ik} \Phi_k^{(1)} + \sum_{l=1}^L \zeta_{ijl} \Phi_l^{(2)} + \mathbf{e}_i, \end{aligned} \quad (7)$$

where $\Phi_k^{(1)} = \mathbf{P} \phi_k^{(1)} \mathbf{D}$ is a subject-specific $F \times T$ -dimensional image, $\Phi_l^{(2)} = \mathbf{P} \phi_l^{(2)} \mathbf{D}$ is a subject-/visit-specific $F \times T$ -dimensional image, and $\mathbf{e}_i = \mathbf{P} \boldsymbol{\eta}_i \mathbf{D} + \mathbf{E}_i$ is an $F \times T$ noise process. The following theorem shows that it is sufficient to conduct MFPCA on the simple model (6) instead of the intractable model (7).

Theorem 2. Suppose that \mathbf{P} is a matrix obtained by column binding A orthonormal eigenvectors of size $F \times 1$ and that D is a matrix obtained by row-binding B orthonormal eigenvectors of size $1 \times T$. Then the following results hold: (a) The vector version of the subject-specific eigenimages $\Phi_k^{(1)} = \mathbf{P}\phi_k^{(1)}\mathbf{D}$ are orthonormal in \mathbb{R}^{FT} ; (b) the vector version of the subject/visit-specific eigenimages $\Phi_l^{(2)} = \mathbf{P}\phi_l^{(2)}\mathbf{D}$ are orthonormal in \mathbb{R}^{FT} ; (c) the vector version of $\Phi_k^{(1)}$ and $\Phi_l^{(2)}$ are not necessarily orthogonal; and (d) the scores ξ_{ik} and ζ_{ijl} are exactly the same in equations (6) and (7).

Theorems 1 and 2 provide simple methods of obtaining ANOVA-like decompositions of very large images based on computable algorithms even for massive images, such as those obtained from brain fMRI. Proofs are provided in the Web supplement.

4. SIMULATION STUDIES

In this section, we generate the frequency-by-time image \mathbf{Y}_{ij} for subject i and visit j from the following model:

$$\mathbf{Y}_{ij}(f, t) = \sum_{k=1}^4 \xi_{ik} \phi_k^{(1)}(f, t) + \sum_{l=1}^4 \zeta_{ijl} \phi_l^{(2)}(f, t) + \epsilon_{ij}(f, t) \quad (8)$$

for $i = 1, \dots, I, j = 1, \dots, J,$

where $\xi_{ik} \sim N\{0, \lambda_k^{(1)}\}$ for $k = 1, \dots, 4$, $\zeta_{ijl} \sim N\{0, \lambda_l^{(2)}\}$ for $l = 1, \dots, 4$, $\epsilon_{ij}(f, t) \sim N(0, \sigma^2)$, $\{f = 0.2f \text{ Hz} : f = 1, \dots, F\}$, where F is the number of frequencies, and $\{t = \frac{t}{T} : m = 1, 2, \dots, T\}$, where T is the number of epochs. We consider $F = 128$ and $T = 120$ in the simulation that follows. We simulate $I = 200$ subjects (clusters) with $J = 2$ visits per subject (measurement per cluster). The true eigenvalues are $\lambda_k^{(1)} = 0.5^{k-1}$, $k = 1, 2, 3, 4$, and $\lambda_l^{(2)} = 0.5^{l-1}$, $l = 1, 2, 3, 4$. We consider multiple scenarios corresponding to different noise magnitudes: $\sigma = 0$ (no noise), $\sigma = 2$ (moderate), and $\sigma = 4$ (large). We conduct 100 simulations for each scenario. The frequency-time eigenfunctions $\phi_k^{(1)}(f, t)$ and $\phi_k^{(2)}(f, t)$ are generated from bases in frequency and time domains, as illustrated below. The bases in the frequency domain are derived from the Haar family of functions, defined as $\psi_{pq}(f) = 2^{p/2}/\sqrt{N}$ for $(q-1)/2^p \leq (f-f_{\min})/(f_{\max}-f_{\min}) < (q-0.5)/2^p$, $\psi_{pq}(f) = -2^{p/2}/\sqrt{N}$ for $(q-0.5)/2^p \leq (f-f_{\min})/(f_{\max}-f_{\min}) < q/2^p$ and $\psi_{pq}(f) = 0$ otherwise. Here N is the number of frequencies and f_{\min} and f_{\max} are the minimum and maximum frequencies under consideration, respectively. In particular, we let the level 1 eigenfunctions be $h_1^{(1)}(f) = \psi_{11}(f)$, $h_2^{(1)}(f) = \psi_{12}(f)$ and level 2 eigenfunctions be $h_1^{(2)}(f) = \psi_{21}(f)$, $h_2^{(2)}(f) = \psi_{22}(f)$. For example, if $f_{\min} = 0.2$ Hz, $f_{\max} = 1.6$ Hz, and frequency increments by 0.2 Hz, then $N = 8$. The eigenfunctions in this case are $h_1^{(1)}(f) = (0.5, 0.5, -0.5, -0.5, 0, 0, 0, 0)$, $h_2^{(1)}(f) = (0, 0, 0, 0, 0.5, 0.5, -0.5, -0.5)$ and $h_1^{(2)}(f) = (\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0, 0, 0, 0, 0, 0)$, $h_2^{(2)}(f) = (0, 0, \frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}, 0, 0, 0, 0)$. For the time domain, we consider the following two choices:

Case 1. Mutually orthogonal bases. Level 1: $g_1^{(1)}(t) = \sqrt{2} \sin(2\pi t)$, $g_2^{(1)}(t) = \sqrt{2} \cos(2\pi t)$. Level 2: $g_1^{(2)}(t) = \sqrt{2} \sin(6\pi t)$, $g_2^{(2)}(t) = \sqrt{2} \cos(6\pi t)$.

Case 2. Mutually nonorthogonal bases. Level 1: same as in Case 1. Level 2: $g_1^{(2)}(t) = 1$, $g_2^{(2)}(t) = \sqrt{3}(2t-1)$.

In the following, we present only results for Case 2; the results for Case 1 were similar. The frequency-time eigenfunctions were generated by multiplying each component of the bases in frequency and time domains, that is, $\phi_k^{(1)}(f, t) = h_{k_f}^{(1)}(f)^T g_{k_t}^{(1)}(t)$, where $k = k_f + 2(k_t - 1)$ for $k_f, k_t = 1, 2$ and $\phi_k^{(2)}(f, t) = h_{k_f}^{(2)}(f)^T g_{k_t}^{(2)}(t)$, where $l = l_f + 2(l_t - 1)$ for $l_f, l_t = 1, 2$. The first figure in the Web supplement displays simulated data from model (8) for one subject at two visits with different magnitudes of noise. The figure shows that as the magnitude of noise increases, the patterns become more difficult to delineate. For clarity, in this plot we used $F = 16$ and $T = 20$.

4.1 Eigenvalues and Eigenfunctions

Figure 2 shows estimated level 1 and 2 eigenvalues for the different magnitudes of noise using the PVD method described in Section 3. Note that the potential measurement error is not accounted for in this figure. In the case of no noise ($\sigma = 0$), the eigenvalues generally can be recovered without bias, although some small bias is present in the estimation of the first eigenvalue at level 2. The bias does not seem to increase substantially with the noise level.

Figure 3 shows estimated eigenfunctions at four randomly selected frequencies from 20 simulated datasets. The simulated data have no measurement error (i.e., $\sigma = 0$). We conclude that PVD successfully separates level 1 and 2 variation and correctly captures the shape of each individual eigenfunction.

4.2 Principal Component Scores

We estimated the principal component scores by Bayesian inference via posterior simulations using Markov chain Monte Carlo (MCMC) methods. We used the software developed by Di et al. (2009) applied to the mixed-effects model (8). Because this method uses the full model, we call it the PC-F method. Because Bayesian calculations can be slow when the dimension of \mathbf{V}_{ij} is very large, Di et al. (2009) introduced a projection model that reduces computation time by orders of magnitude. Because this uses a projection in the original mixed-effects model, we call this the PC-P method. In simulations, PC-P proved to be slightly less efficient, but much faster, than PC-F. [For a thorough introduction to Bayesian functional data analysis using WinBUGS (Spiegelhalter et al. 2003) see Crainiceanu and Goldsmith (2009).]

We use the full model PC-F and the projection model PC-P proposed by Di et al. (2009) to estimate PC scores after obtaining the estimated eigenvalues and eigenfunctions using PVD. To compare the performance of these two models, we compute the root mean squared errors (RMSEs). In each scenario, we randomly select 10 simulated datasets and estimate the PC scores using posterior means from the MCMC runs. The MCMC convergence and mixing properties are assessed by visual inspection of the chain histories of many parameters of interest. The history plots (not shown) indicate very good convergence and mixing properties. Table 2 reports the means of the RMSE, indicating that as the amount of noise increases, the RMSE also increases. A direct comparison of the RMSE

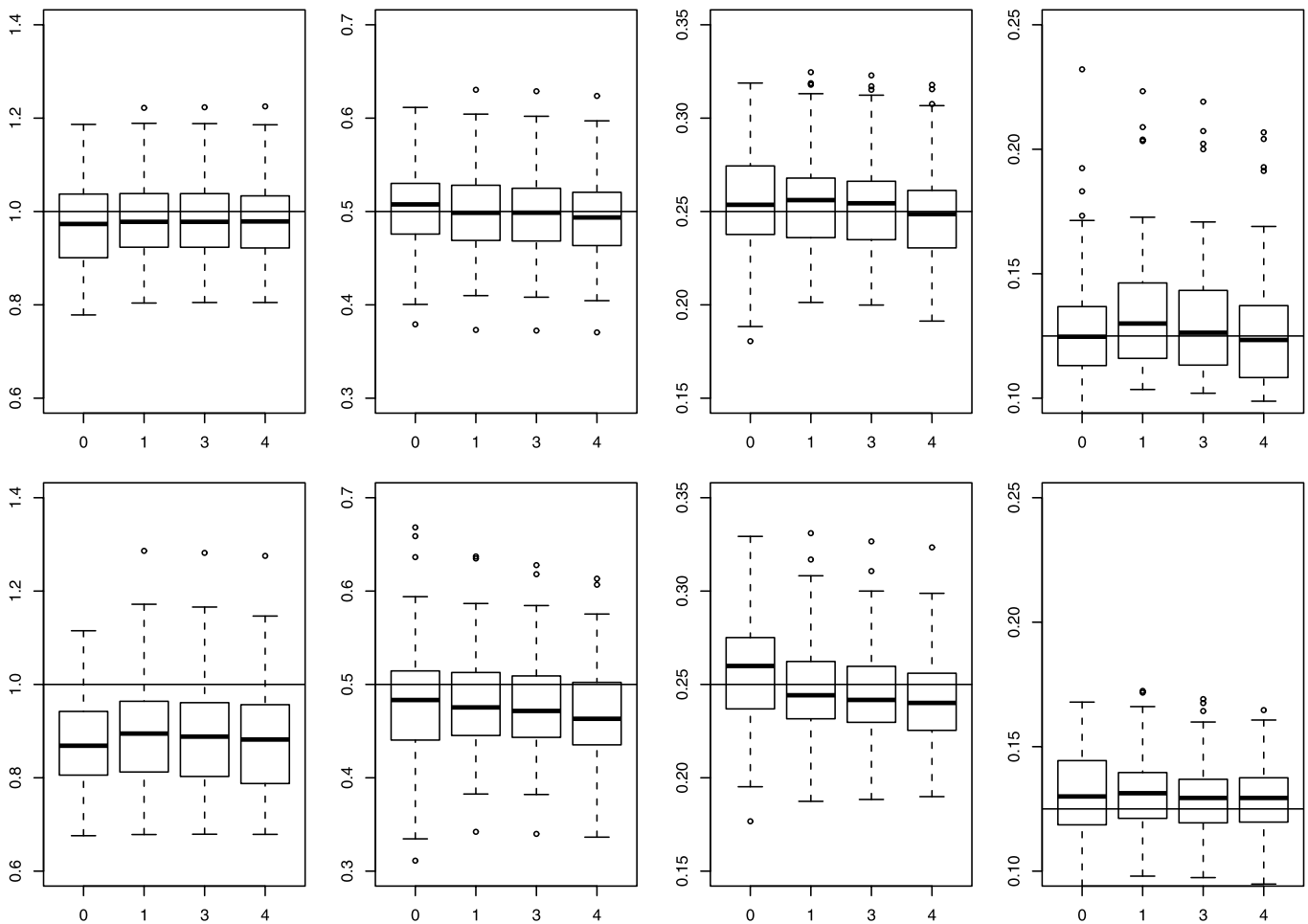


Figure 2. Boxplots of estimated eigenvalues using unsmooth MFPCA-3; the true functions are without noise and with noise. The solid gray lines are the true eigenvalues. The x -axis labels indicate the standard deviation of the noise.

with the standard deviation of the scores at the four levels (1, 0.71, 0.50, and 0.35) demonstrates that scores are well estimated, especially at level 1. Moreover, PC-F performs slightly better than PC-P in terms of RMSE; however, PC-P might still be preferred in applications where PC-F is computationally expensive.

5. APPLICATION TO THE SHHS

In Section 2 we introduced the SHHS, which collected two PSGs for thousands of subjects roughly 5 years apart. Here we focus on analyzing the frequency-by-time spectrograms for $N = 3201$ subjects at $J = 2$ visits. We analyze all frequencies from 0.2 Hz to 32 Hz in 0.2-Hz increments for a total number of $F = 160$ grid points in frequency and the first 4 hours of sleep in increments of 30 seconds, for a total number of $T = 480$ grid points in time. The total number of observations per subject per visit is $FT = 76,800$, and the total number of observations across all subjects and visits is $FTNJ = 491,673,600$. The same methods could be easily applied to fMRI studies, where one image would contain more than $V = 2,000,000$ voxels and $T = 500$ time points for a total of $VT = 1,000,000,000$ observations per image. The methods described in this article are designed to scale up well to these larger imaging studies.

For each subject i , $i = 1, \dots, I = 3201$, and visit j , $j = 1, J = 2$, we obtained \mathbf{Y}_{ij} , the $F \times T = 160 \times 480$ dimensional frequency-by-time spectrogram. We de-mean the row and column vectors of each matrix using the transformation $\mathbf{Y}_{ij} \mapsto \{\mathbf{I}_F - \mathbf{E}_F/F\}\mathbf{Y}_{ij}\{\mathbf{I}_T - \mathbf{E}_T/T\}$, where $\mathbf{I}_F, \mathbf{I}_T$ denote the identity matrices of size F and T , and \mathbf{E}_F and \mathbf{E}_T are square matrices with each entry equal to 1 of size F and T , respectively. Note that any image \mathbf{Y}_{ij} can be written as

$$\begin{aligned} \mathbf{Y}_{ij} = & \{\mathbf{I}_F - \mathbf{E}_F/F\}\mathbf{Y}_{ij}\{\mathbf{I}_T - \mathbf{E}_T/T\} \\ & + \mathbf{E}_F\mathbf{Y}_{ij}/F + \mathbf{Y}_{ij}\mathbf{E}_T/T - \mathbf{E}_F\mathbf{Y}_{ij}\mathbf{E}_T/(FT). \end{aligned}$$

The last term of the equality, $\mathbf{E}_F\mathbf{Y}_{ij}\mathbf{E}_T/(FT)$, is an $F \times T$ -dimensional matrix with all entries equal to the average of all entries in \mathbf{Y}_{ij} . The third term of the equality, $\mathbf{Y}_{ij}\mathbf{E}_T/T$, is a matrix with T identical columns equal to the row means of the matrix \mathbf{Y}_{ij} . Similarly, $\mathbf{E}_F\mathbf{Y}_{ij}/F$ is a matrix with F identical rows equal to the column means of the matrix \mathbf{Y}_{ij} . We conclude that the inherently bivariate information in the image \mathbf{Y}_{ij} is encapsulated in $\{\mathbf{I}_F - \mathbf{E}_F/F\}\mathbf{Y}_{ij}\{\mathbf{I}_T - \mathbf{E}_T/T\}$. Methods for analyzing the average of the entire image are standard. Methods for analyzing the column and row means of the image are either classical or have been developed recently (Crainiceanu, Staicu, and Di 2009; Di et al. 2009;

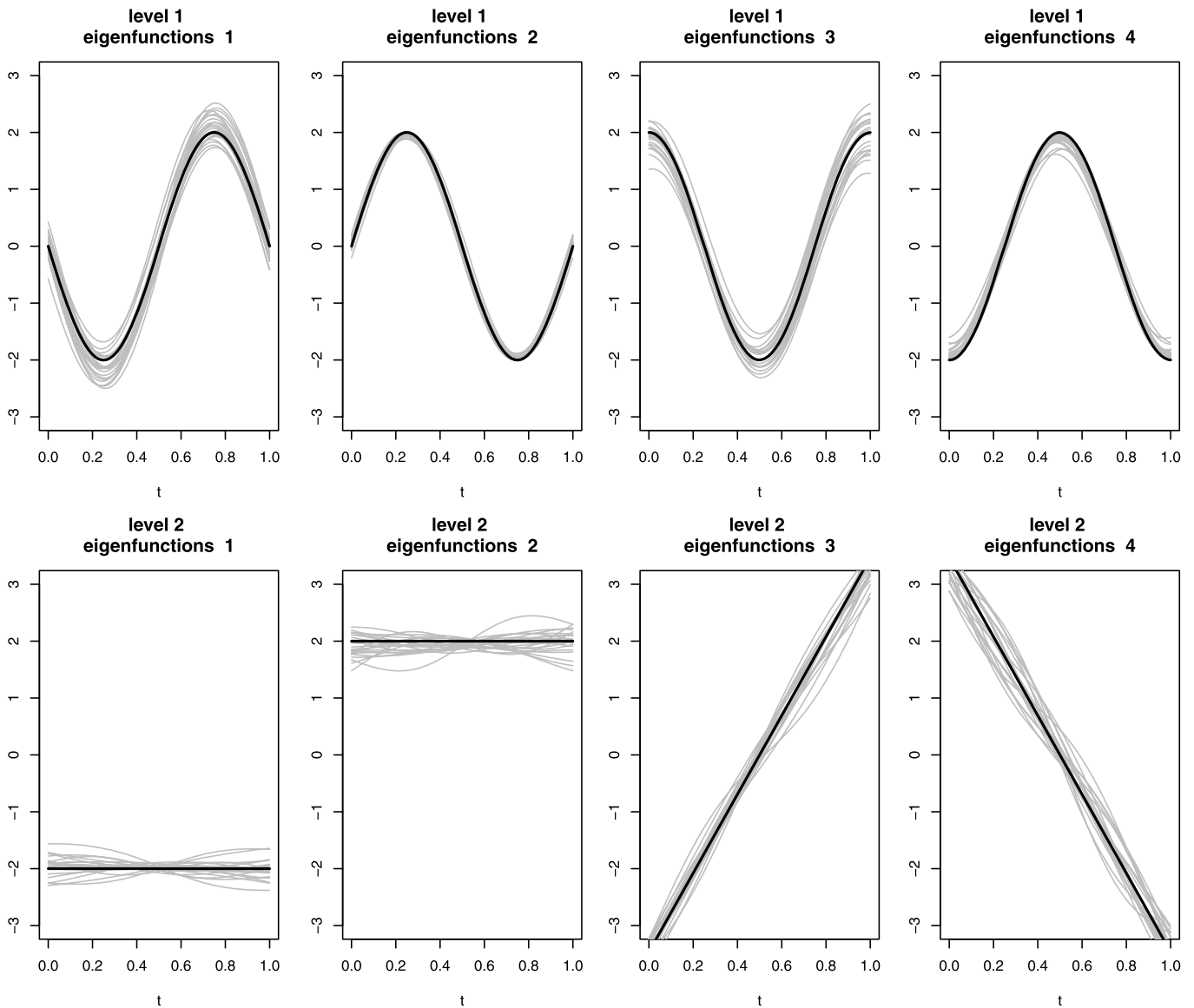


Figure 3. Estimated eigenfunctions at four randomly selected frequencies from 20 simulated datasets when the frequency–time images are observed without noise (i.e., $\sigma = 0$). The thick black lines represent true eigenfunctions at those randomly selected frequencies; the gray lines, estimated eigenfunctions.

Staicu, Crainiceanu, and Carroll 2010). Thus we focus on analyzing $\{\mathbf{I}_F - \mathbf{E}_F/F\}\mathbf{Y}_{ij}\{\mathbf{I}_T - \mathbf{E}_T/T\}$, and we continue to denote this $F \times T$ -dimensional matrix by \mathbf{Y}_{ij} . With this definition of \mathbf{Y}_{ij} , we proceed with the main steps of our analysis. We first obtain the subject-/visit-specific SVD $\mathbf{Y}_{ij} = \mathbf{U}_{ij}\boldsymbol{\Sigma}_{ij}\mathbf{V}_{ij}^T$. We then store the first $L_i = 10$ columns of the matrix \mathbf{U}_{ij} in the matrix $\mathbf{U}_{L_i,j}$ and construct the two matrices $\mathbf{U}_j = [\mathbf{U}_{L_1,j}], \dots, [\mathbf{U}_{L_i,j}]$ for $j = 1, 2$. Both matrices \mathbf{U}_j are $160 \times 32,010$ -dimensional, and we obtain the $160 \times 64,020$ -dimensional matrix $\mathbf{U} = [\mathbf{U}_1|\mathbf{U}_2]$ by column binding \mathbf{U}_1 and \mathbf{U}_2 . To obtain the main directions of variation in the space spanned by the column space of the matrix \mathbf{U} , we diagonalize the 160×160 -dimensional matrix $\mathbf{U}\mathbf{U}^T$. We call the eigenvectors of the matrix $\mathbf{U}\mathbf{U}^T$ population eigenfrequencies. We apply a similar construction and decomposition to the matrix $\mathbf{V}\mathbf{V}^T$, whose eigenvectors we call eigenvariates. Because $\mathbf{V}\mathbf{V}^T$ is much noisier than $\mathbf{U}\mathbf{U}^T$, we first apply

row-by-row smoothing of $\mathbf{V}\mathbf{V}^T$. Bivariate smoothing is prohibitively slow, but this approach proved to be fast.

Table 3 displays some important eigenvalues of $\mathbf{U}\mathbf{U}^T$ and $\mathbf{V}\mathbf{V}^T$, respectively. The results are reassuring and support our intuition that samples of images have many common features. Indeed, the first 13 population-level eigenfrequencies explain more than 90% of the variability of collection of first 10 subject-specific eigenfrequencies over more than 3000 subjects. Another interesting property of the population eigenfrequencies is that the most important five to seven of them explain a similar amount of variability; note the very slow decay in the associated variance components. The variance explained decays exponentially starting with component 8 and becomes practically negligible for components 15 and beyond. Returning to our thought experiment, this means that if we look at the frequency (X) dimension across subjects, we will see much consistency in terms of the shape and location of the observed sig-

Table 2. RMSEs for estimating scores using PC-F and PC-P

Method	σ	Level 1 component				Level 2 component			
		1	2	3	4	1	2	3	4
Case 2: PC-F	0	0.056	0.036	0.053	0.044	0.122	0.111	0.153	0.122
	2	0.065	0.051	0.065	0.060	0.132	0.121	0.178	0.131
	4	0.120	0.089	0.095	0.100	0.145	0.125	0.167	0.145
Case 2: PC-P	0	0.068	0.063	0.074	0.052	0.135	0.196	0.212	0.130
	2	0.079	0.087	0.087	0.060	0.138	0.227	0.258	0.150
	4	0.139	0.160	0.103	0.104	0.161	0.138	0.223	0.175

nal. This is consistent with the population data, which shows higher proportional power and variability in the δ and α power bands across subjects. Our results quantify this general observation while remaining agnostic to the classical partition of the frequency domain.

A similar story can be told about the eigenvariates, although some of the specifics differ. More precisely, the variance explained by individual eigenvariates decreases more linearly and does not exhibit any sudden drop. Moreover, the first 13 eigenvariates explain roughly 80% of the observed variability of the subject-specific eigenvariates, and 20 eigenvariates are necessary to explain 90% of the variability.

The shape of the first 10 population-level eigenfrequencies and eigenvariates are displayed in Figures 4 and 5. Figure 4 indicates that most of the variability is in a range of frequencies that roughly overlaps with the δ power band range [0.8, 4 Hz]. This should not be surprising, given that most of the observed variability is obviously in this frequency range; however, the level and type of variability that we identified in the δ power band are novel findings. For example, subjects who are positively loaded on the first eigenfrequency (top-left plot in Figure 4) will tend to have much higher percent power around frequency 0.6 Hz than around 1.2 Hz. Similarly, a subject who is positively loaded on the second eigenfrequency (top-right panel in Figure 4) will have higher percent power around frequencies 0.4 and 1.2 Hz than around 0.8 Hz. Moreover, differences between percent power in these frequencies are quite sharp. Another interesting finding is that the first five eigenfrequencies seem “dedicated” to discrepancies in the low part of the frequency range [0.2, 2 Hz]. Each of these eigenfrequencies explains roughly 10% of the eigenfrequency variability for

a combined 49% explained variability. Starting with eigenfrequency six, there is a slow but steady shift toward discrepancies at higher frequency. Moreover, higher eigenfrequencies display more detail in the 8–10 Hz range, which is well within the α power range [8.1, 13.0 Hz].

The eigenvariates shown in Figure 5 tell an equally interesting, but different, story. First, all eigenvariates indicate that differences in the time domain tend to be smooth, with very few sudden changes. An alternative interpretation would be that some transitions may occur very rapidly in time but are undetectable in the signal. A closer look at the first eigenvariate indicates that, relative to the population average, subjects who are positively loaded on this component (top-left plot) will tend to have (a) higher percent power between minutes 30 and 50; (b) slightly lower percent power between minute 70 and 80; (c) higher percent power between minutes 120 and 140, but with smaller discrepancy than that seen around minute 40; and (d) smaller percent power between minutes 180 and 210. The other eigenvariates have similarly interesting interpretations. It is noteworthy that eigenvariates become roughly sinusoidal starting with the seventh eigenvariate. There are at least two alternative explanations for this. First, it could be that there are indeed high-frequency cycles in the population. Another possible explanation is that the distances between peaks and valleys vary randomly across subjects (see Woodard, Crainiceanu, and Ruppert 2012 for an explanation of this behavior).

The eigenfrequencies and eigenvariates are interesting in themselves, but it is the Kronecker product of these bases that provides the projection basis for the actual images. Figure 6 displays some population-level basis components obtained as

Table 3. Variance and cumulated percent variance explained by population-level eigenvalues from the observed variance of eigenvalues at the subject level. The labels eigenfrequencies and eigenvariates refer to the left and right eigenvectors, respectively. Population-level eigenfrequencies are the eigenvectors in the \mathbb{R}^F -dimensional subspace spanned by the collection of the first 10 eigenfrequencies at the subject level across all subjects. Population-level eigenvariates are the eigenvectors in the \mathbb{R}^T -dimensional subspace spanned by the collection of the first 10 eigenvariates at the subject level across all subjects

	Component									
	1	5	6	7	8	9	10	11	12	13
Eigenfrequencies										
$\lambda (\times 10^{-2})$	9.95	9.58	9.38	8.80	8.19	6.73	4.45	2.37	1.92	1.69
Sum % var	25.01	49.09	58.49	67.31	75.51	82.25	86.71	89.08	90.10	92.69
Eigenvariates										
$\lambda (\times 10^{-2})$	2.10	1.21	1.07	0.89	0.74	0.63	0.55	0.49	0.42	0.37
Sum % var	30.24	47.67	54.13	59.50	63.94	67.75	71.09	74.04	76.57	78.82

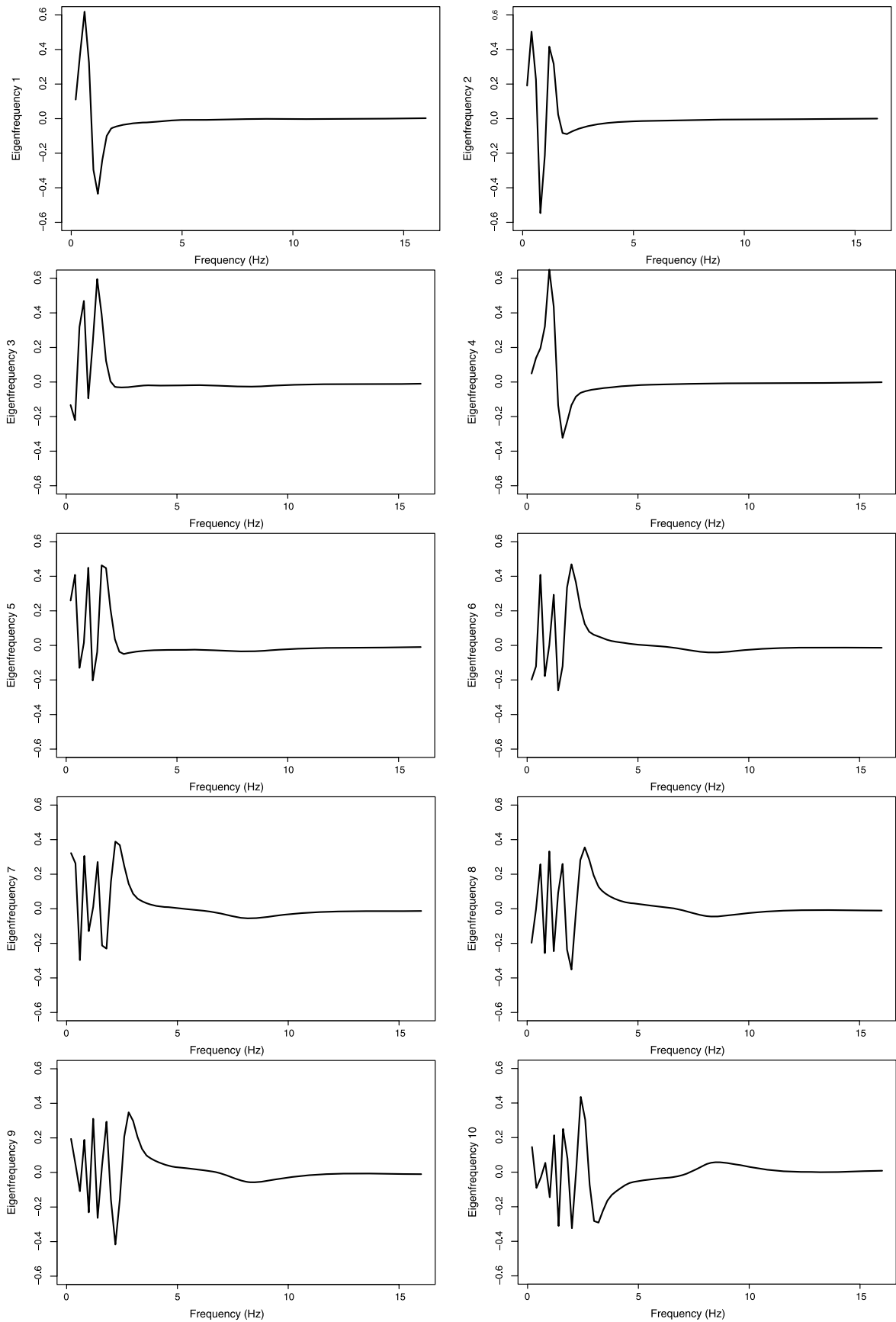


Figure 4. The first 10 population-level eigenfrequencies for the combined data from visits 1 and 2. The X-axis is frequency in Hz. Eigenfrequencies are truncated at 16 Hz for plotting purposes, but they extend to 32 Hz.

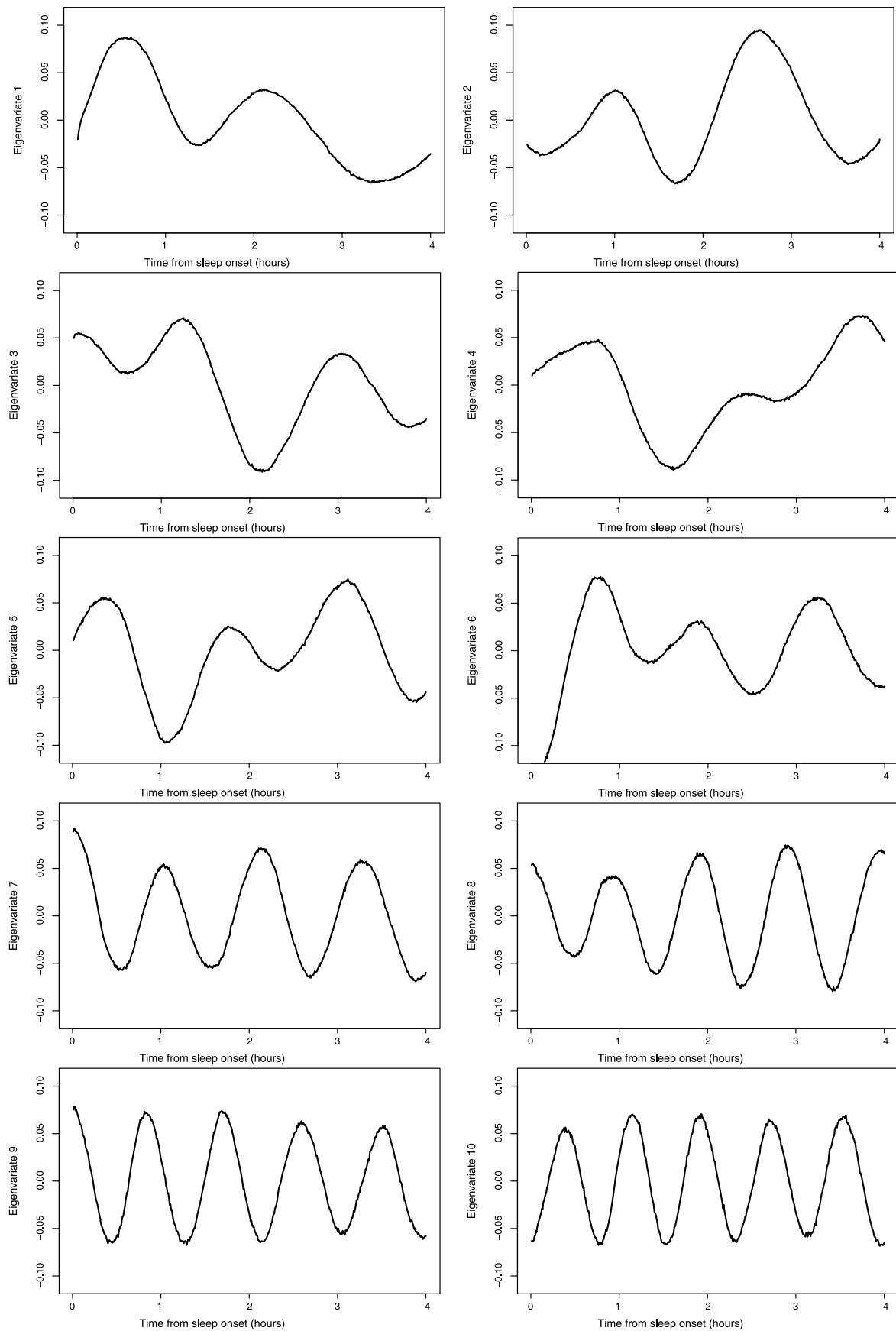


Figure 5. First 10 population-level eigenvariates for the combined data from visits 1 and 2. The X-axis represents time from sleep onset in hours.

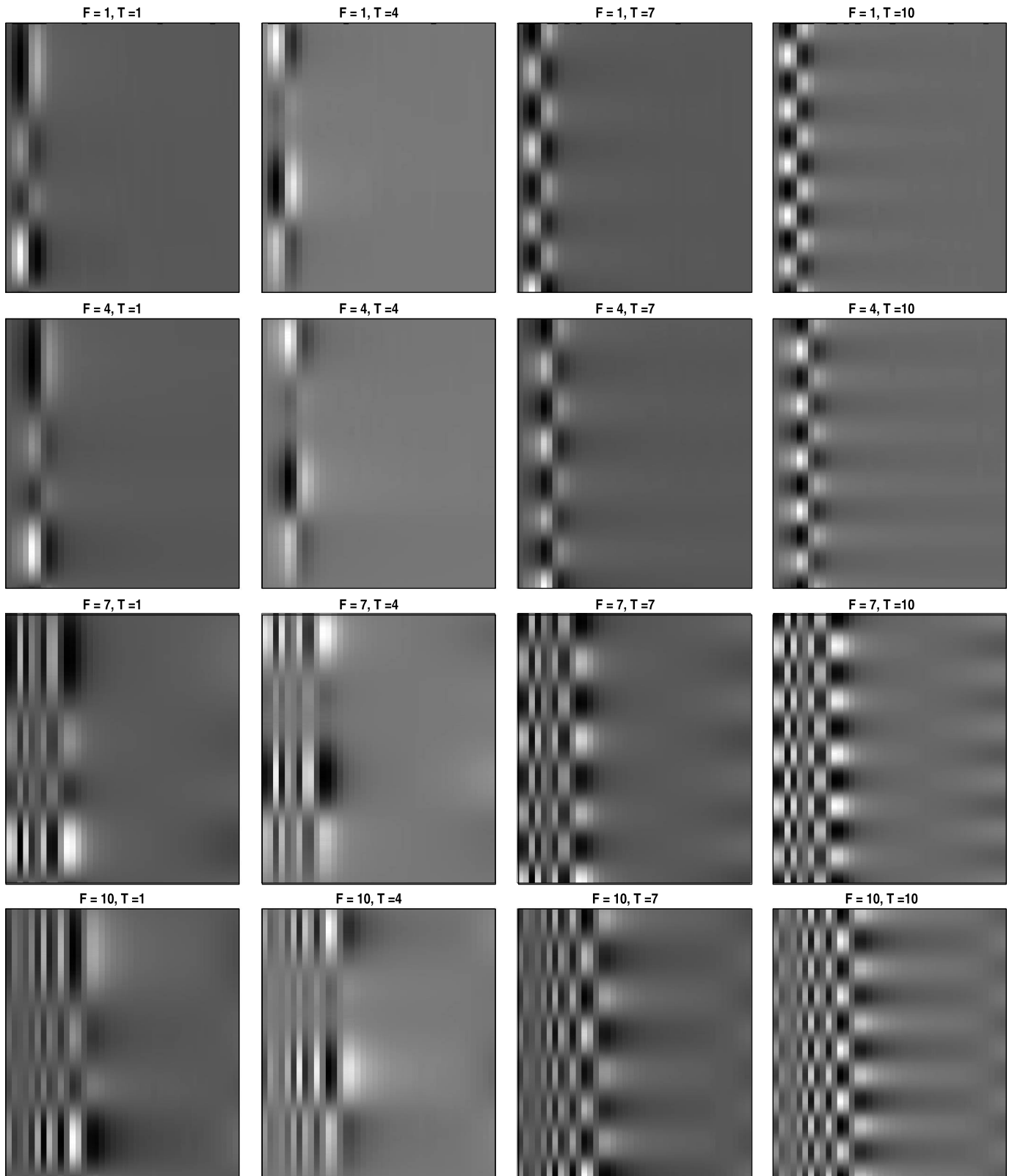


Figure 6. Some population-level basis components obtained as Kronecker products of eigenfrequencies and eigenvariates. The frequencies from 0.2 to 8 Hz are shown on the x -axis, and time from sleep onset until the end of the fourth hour is given on the y -axis. The title of each image indicates the eigenfrequency number (F) and eigenvariate number (T), as ordered by their corresponding eigenvalues. For example, $F = 1, T = 7$ indicates the basis component obtained as a Kronecker product of the first eigenfrequency and the seventh eigenvariate.

Kronecker products of eigenfrequencies and eigenvariates. We call these eigenimages. The x -axis represents the frequencies from 0.2 to 8 Hz, and the y -axis represents the time from onset of sleep until the end of the fourth hour. Images are cut at 8 Hz to focus on the more interesting part of the graph, but analyses were conducted on frequencies up to 32 Hz. The title of each image indicates the eigenfrequency number (F) and eigenvariate number (T), as ordered by their corresponding eigenvalues; for example, $F = 1, T = 7$ indicates the basis component obtained as a Kronecker product of the first eigenfrequency and the seventh eigenvariate. The checkerboard patterns seen in the right panels are due to the seventh and tenth eigenvariate, which are the sinus-like functions displayed in Figure 5.

We next investigated the smoothing effects of the population level eigenimages. The top left panel in Figure 7 displays the frequency-by-time plot of the fraction power for the same subject shown in Figure 1. The only difference is that the time interval was reduced to the first 4 hours after sleep onset. The top-right panel displays the projection of the frequency-by-time image on a basis with 45 components obtained as Kronecker products of the first 15 population-level eigenfrequencies and the first 15 population-level eigenvariates. The smooth surface provides a pleasing summary of the main features of the original data by reducing some of the observed noise. The

bottom-left plot displays a projection of the frequency-by-time image with 45 components obtained as Kronecker products of the first 15 subject-level eigenfrequencies and the first three subject-level eigenvariates. We did not include more subject-level eigenvariates because they were indistinguishable from noise. The bottom-right plot displays the difference between the projection on the subject-level basis (bottom-left panel) and the projection on the population-level basis (top-right panel). We conclude that both projections on the subject-level and the population-level bases reduce the noise in the original image and provide pleasing summaries of the main features of the data. The two summaries are not identical; the subject-level smooth is slightly closer to the original data in the δ frequency range (note the sharper peaks), whereas the population-level smooth is closer to the original data in the α frequency range (compare the number and size of peaks). Although which basis should be used at the subject level can be debated, there is no doubt that having a population-level basis with reasonable smoothing properties is an excellent tool if the final goal is statistical inference on populations of images. The current practice of taking averages over frequencies in the δ power band can be viewed as a much cruder alternative. These plots also indicate a potential challenge that was not addressed. The variability around the signal seem to be roughly proportional to the

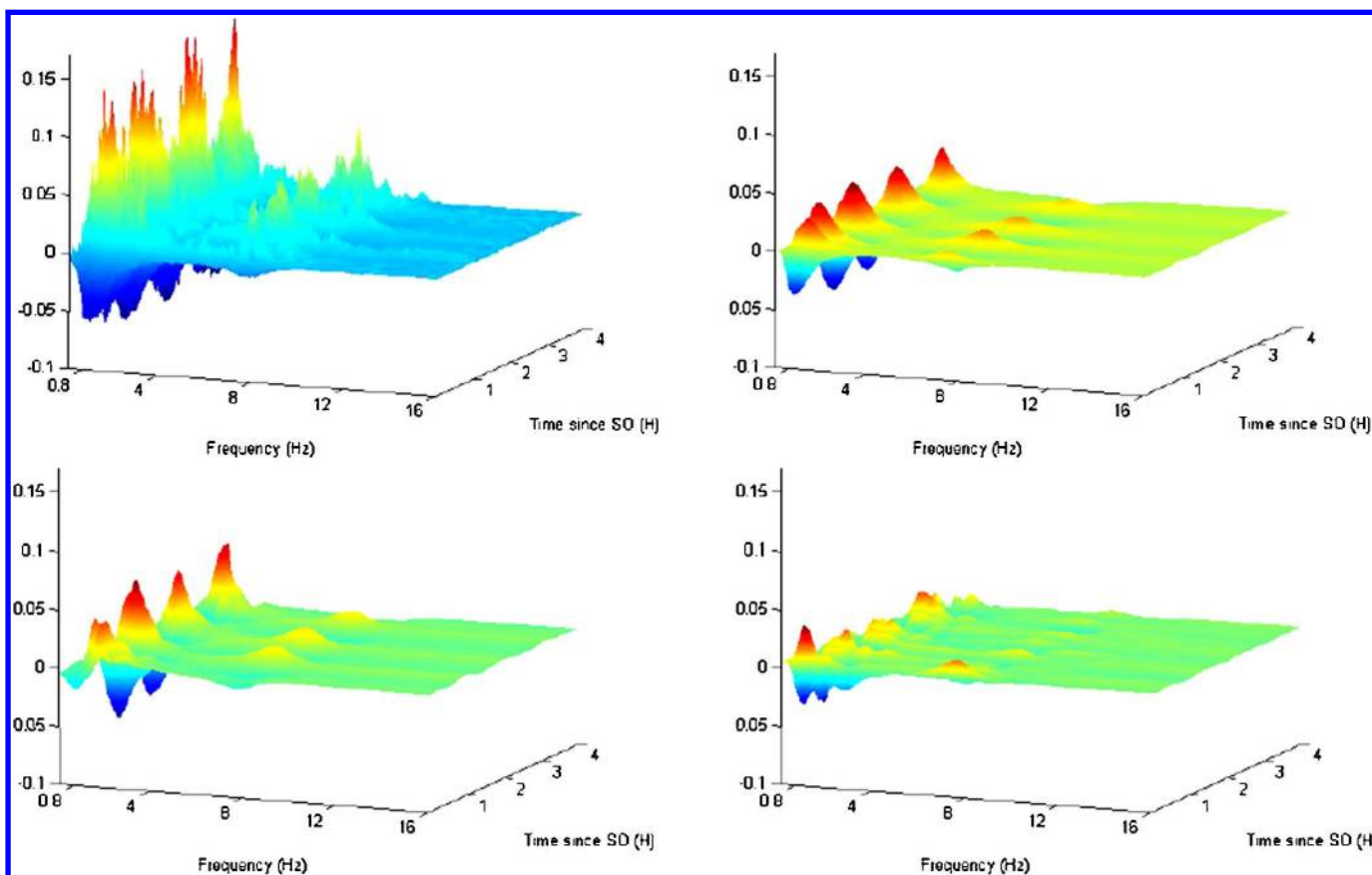


Figure 7. Image smoothing for one subject for the first 4 hours of sleep after sleep onset. Top left panel displays the normalized power up to 16 Hz, even though the analysis is based on data up to 32 Hz. Top right panel displays the smooth image obtained by projection on the first 15 eigenfrequencies and first 3 smoothed eigenvariates at the subject level; the other eigenvariates at the subject level are indistinguishable from white noise. Bottom left panel displays the smooth image obtained by projection on the first 15 eigenfrequencies and first 15 eigenvariates at the population-level (some shown in Figure 3). The bottom-right panel displays the difference between the subject-level smooth (top-right panel) and population-level smooth (bottom-left panel).

signal, a rather unexpected feature of the data that merits further investigation. This problem exceeds the scope of this article.

To analyze the clustering of images, we used a basis with 100 components obtained by taking the Kronecker product of the first 10 eigenfrequencies and first 10 eigenvariables. Examples of these components are shown in Figure 6. The subject/visit-specific coefficients were obtained by projecting the original images on this basis, which resulted in a 100-dimensional vector of coefficients. Thus we applied MFPCA (Di et al. 2009) to $I = 3201$ subjects observed at $J = 2$ visits, with each subject/visit characterized by a vector \mathbf{v}_{ij} of 100 coefficients. This took less than 10 seconds using a personal computer with a dual-core processor with 3 GHz CPU and 8 Gb RAM. We fit the model (6) from Section 3.2.2 in matrix form: $\mathbf{V}_{ij} = \sum_{k=1}^K \xi_{ik} \boldsymbol{\phi}_k^{(1)} + \sum_{l=1}^L \zeta_{ijl} \boldsymbol{\phi}_l^{(2)} + \boldsymbol{\eta}_i$, where $\xi_{ik} \sim N\{0, \lambda_k^{(1)}\}$, $\zeta_{ijl} \sim N\{0, \lambda_l^{(2)}\}$ are mutually uncorrelated. We first focused on estimating $\lambda_k^{(1)}$, $\lambda_l^{(2)}$, $\boldsymbol{\phi}_k^{(1)}$, $\boldsymbol{\phi}_l^{(2)}$, K , and L . The table in the web appendix provides the estimates for the first 10 eigenvalues indicating that the level 2 eigenvalues quantifying the visit-specific variability are roughly 100 times larger than the level 1 eigenvalues quantifying the subject-specific variability. Using the same notation as in Di et al. (2009) the proportion of variance explained by within-subject variability is $\rho_W = (\sum_{k=1}^{100} \lambda_k^{(1)}) / (\sum_{k=1}^{100} \lambda_k^{(1)} + \sum_{l=1}^{100} \lambda_l^{(2)})$. A plug-in estimator of ρ_W is $\hat{\rho}_W = 0.033$, which indicates that the between-subject variability is very small compared to the within-subject between-visit variability. In studies of δ -power (Di et al. 2009) estimated a much higher ρ_W , in the range [0.15, 0.20], depending on the particular application. Our results do not contradict these previous results, given that the subject-specific mean over all time points was removed from the bivariate spectrogram. However, they indicate that in the SHHS, most of the within-subject correlation is contained in the margins of the frequency-by-time image. The margins are the column and row means of the original bivariate plots.

The left panels in Figure 8 display the first four subject-level eigenfunctions, $\boldsymbol{\phi}_k^{(1)}$, $k = 1, \dots, K$, in the coefficient space. In matrix format, these bases are 10×10 -dimensional and are difficult to interpret; however, by premultiplying and postmultiplying them with the population-level matrices \mathbf{P} and \mathbf{D} , we obtain the eigenimages in the original space, $\boldsymbol{\Phi}_k^{(1)} = \mathbf{P} \boldsymbol{\phi}_k^{(1)} \mathbf{D}$. These eigenimages are displayed in the corresponding right panels of Figure 8. The second figure in the Web supplement provides the same results for the level 2 eigenimages.

6. DISCUSSION

Statistical analysis of populations of images when even one image cannot be loaded in the computer memory is a daunting task. Historically, data compression or signal extraction methods aim to reduce the very large images to a few indices that can be then analyzed statistically. Examples are total brain volume obtained from fMRI studies or average percent δ power in sleep EEG studies. In this article we have proposed an integrated approach to signal extraction and statistical analysis that (a) uses the information available in images efficiently, (b) is computationally fast and scalable to much larger studies, and (c) provides equivalence results between the analysis of populations of image coefficients and populations of images. We

applied our approach to the SHHS, arguably one of the largest studies analyzed statistically. Indeed, only the EEG data in the study contains more than 85 billion observations.

The most important contribution of this article is further advancing the foundation for next-generation statistical studies. We call this area the large N , large P , large J problem, where N denotes the number of subjects, P denotes the dimensionality of the problem, and J denotes the number of visits or observations within cluster. Note that the famous small N , large P problem can be obtained from our problem by setting $J = 1$ and cutting N . Our methods are designed for K -dimensional matrices, where dimensions naturally split into two different modalities (e.g., time and frequency in spectral analysis and time and space in fMRI). Because we use a two-stage SVD, our method inherits the weaknesses of the SVDm including (a) sensitivity to noise, correlation, and outliers; (b) dependence on methods for choosing the dimension of the underlying linear space; and (c) lack of invariance under nonlinear transformations of the data.

It is important to better position our work with respect to other methods used for image analysis, including PCA (Seber 1984; Christensen 2001; Jolliffe 2002), independent component analysis (ICA) (Comon 1994; Hyvärinen and Oja 2000; Hyvärinen, Karhunen, and Oja 2001) and partial least squares (Wold et al. 1984; Wold 1985; Cook 2007). In short, our method is a multistage PCA method. Indeed, the subject-level SVD of the data matrix \mathbf{Y}_i is a decomposition, $\mathbf{Y}_i = \mathbf{U}_i \boldsymbol{\Sigma}_i \mathbf{V}_i$, where (a) \mathbf{V}_i are the right eigenvectors of the matrix \mathbf{Y}_i and satisfy $\mathbf{Y}_i^T \mathbf{Y}_i = \mathbf{V}_i^T \boldsymbol{\Sigma}_i^2 \mathbf{V}_i$; (b) \mathbf{U}_i are the left eigenvectors of the matrix \mathbf{Y}_i and satisfy $\mathbf{Y}_i \mathbf{Y}_i^T = \mathbf{U}_i \boldsymbol{\Sigma}_i^2 \mathbf{U}_i^T$; and (c) $\boldsymbol{\Sigma}_i$ is a diagonal matrix containing the square roots of the eigenvalues of $\mathbf{Y}_i^T \mathbf{Y}_i$ and $\mathbf{Y}_i \mathbf{Y}_i^T$ on the main diagonal. Our proposed method is a multistage PCA method, because it extracts the first K left and right subject-specific eigenvectors, stacks them, and conducts a second-stage PCA analysis on the stacked eigenvectors. ICA is an excellent tool for decomposing variability in independent rather than uncorrelated components and works very well when signals are nonnormal. However, statistically principled ICA analysis of populations of images is still in its infancy. Group ICA (Calhoun et al. 2001; Calhoun, Liu, and Adali 2009) currently cannot be applied to, say, hundreds of fMRI images. Moreover, ICA uses PCA as a preprocessing step before conducting ICA. We are aware that the team behind the 1000 Connectome (http://www.nitrc.org/projects/fcon_1000/) has reportedly used group ICA methods for analyzing thousands of fMRIs; however, the software posted does not show how to conduct group ICA on these images. We speculate that the team pooled results from many small-group ICA analyses, which is likely computationally expensive. PVD is a simple and very fast alternative that could inform future group ICA methods. Partial least squares regression is related to principal components regression, and thus regression using SVD decompositions. We have not yet focused on the regression part of the problem and are interested in smoothing and decomposing the variability of populations of images.

A simple alternative to our two-stage SVD was suggested by the associate editor. Using the notation in equation (2), the method would sum the $\mathbf{Y}_i \mathbf{Y}_i^T$ and use the SVD of this sum to estimate \mathbf{P} , and then sum the $\mathbf{Y}_i^T \mathbf{Y}_i$ matrices and use the SVD

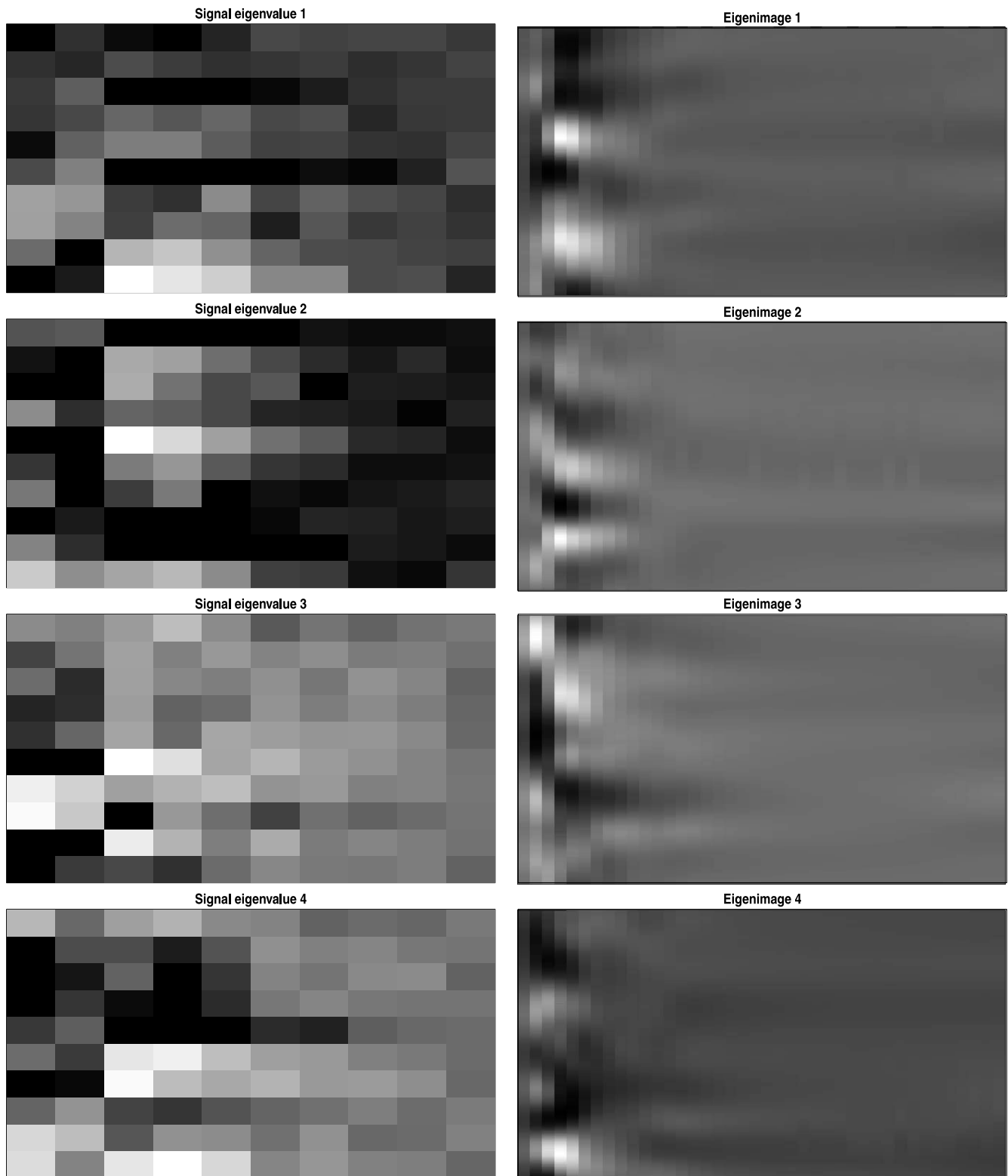


Figure 8. The left panels show the first four subject-specific eigenimages, $\phi_k^{(1)}$, of the multivariate process of image coefficients, \mathbf{V}_{ij} . The right panels show the first four subject-specific eigenimages, $\mathbf{P}\phi_k^{(1)}\mathbf{D}$, of the image process, \mathbf{Y}_{ij} . The right panels are reconstructed from the left panels using the transformation $\phi_k^{(1)} \rightarrow \mathbf{P}\phi_k^{(1)}\mathbf{D}$ from the coefficient to the image space.

of this sum to estimate \mathbf{D} . This is a very simple and compelling idea that we have also considered. It provides an excellent, and potentially faster, alternative to our default PVD procedure in the particular example that we consider here. Nonetheless, there are many reasons for using PVD. First, in many applications, one of the dimensions is very large; for example, in fMRI the number of voxels is in the millions, and calculating and diagonalizing the space-by-space covariance matrix would be out of the question. Second, our method provides the subject-specific left and right eigenfunctions and opens up new possibilities for analysis. For example, we might be interested in studying the variability of \mathbf{U}_{L_i} , the matrix containing the first L_i left eigenvectors of the data matrix \mathbf{Y}_i , around \mathbf{P} , the population-level matrix of left eigenvectors. Third, our method likely is equally as fast and requires only minimal additional coding. Fourth, both methods are reasonable ways of constructing the \mathbf{P} and \mathbf{D} matrices. Simply putting forward the PVD formula will lead to many ways of building \mathbf{P} and \mathbf{D} .

A few open problems remain that need to be addressed. First, theoretic and methodological approaches are needed to determine the cutoff dimension for the number of subject-specific eigenfrequencies and eigenvariates retained for the second stage of the analysis. Although we use the same number of eigenfrequencies and eigenvectors, it might make sense to keep a different number of bases in each dimension. Second, methods are needed to address the noise in images. The noise in the frequency-by-time plots is large, and its size probably depends on the size of the signal. SVD of images with complex noise structure remains an open area of research. Third, investigating the optimality properties, or lack thereof, of our procedure is needed and may lead to better or faster procedures. Fourth, better visualization tools need to be developed to address the data onslaught. Despite our best efforts, we believe that better ways of presenting terabytes, and soon petabytes, of data are needed. Fifth, better understanding of the geometry of images in very-high dimensional spaces is necessary.

SUPPLEMENTARY MATERIALS

Examples, plots, and proof of Theorem 1: The pdf file contains examples of simulated data used in the simulation section (page 1 of supplement), plots of the visit-specific eigenimages of the processes \mathbf{V}_{ij} and \mathbf{Y}_{ij} , respectively for the sleep EEG application (pages 2, 3 of supplement), and proof of Theorem 1 (page 4 of supplement).
(web_supplement_images.pdf)

[Received February 2010. Revised May 2011.]

REFERENCES

- Caffo, B. S., Crainiceanu, C. M., Verdusco, G., Joel, S., Mostofsky, S., Spear-Bassett, S., and Pekar, J. (2010), "Two-Stage Decompositions for the Analysis of Functional Connectivity for fMRI With Application to Alzheimer's Disease Risk," *NeuroImage*, 51, 1140–1149. [776,778]
- Calhoun, V. D., Adali, T., Pearlson, G. D., and Pekar, J. (2001), "A Method for Making Group Inferences From Functional MRI Data Using Independent Component Analysis," *Human Brain Mapping*, 14, 140–151. [788]
- Calhoun, V. D., Liu, J., and Adali, T. (2009), "A Review of Group ICA for fMRI Data and ICA for Joint Inference of Imaging, Genetic, and ERP Data," *NeuroImage*, 45, 163–172. [788]
- Cheshire, K., Engleman, H., Deary, I., Shapiro, C., and Douglas, N. J. (1992), "Factors Impairing Daytime Performance in Patients With Sleep Apnea/Hypopnea Syndrome," *Archives of Internal Medicine*, 152, 538–541. [776]
- Christensen, R. (2001), *Advanced Linear Modeling* (2nd ed.), New York: Springer. [788]
- Comon, P. (1994), "Independent Component Analysis: A New Concept?" *Signal Processing*, 36, 287–314. [788]
- Cook, D. (2007), "Fisher Lecture: Dimension Reduction in Regression," *Statistical Science*, 22, 1–26. [788]
- Crainiceanu, C. M., and Goldsmith, A. J. (2009), "Bayesian Functional Data Analysis Using WinBUGS," *Journal of Statistical Software*, 32. [780]
- Crainiceanu, C. M., Caffo, B. S., Di, C., and Punjabi, N. (2009), "Nonparametric Signal Extraction and Measurement Error in the Analysis of Electroencephalographic Activity During Sleep," *Journal of the American Statistical Association*, 104 (486), 541–555. [777]
- Crainiceanu, C. M., Staicu, A.-M., and Di, C. (2009), "Generalized Multilevel Functional Regression," *Journal of the American Statistical Association*, 104, 1550–1561. [776,779,781]
- Di, C., Crainiceanu, C. M., Caffo, B. S., and Punjabi, N. (2009), "Multilevel Functional Principal Component Analysis," *The Annals of Applied Statistics*, 3, 458–488. [776,777,779–781,788]
- Guilleminault, C., Partinen, M., Quera-Salva, M., Hayes, B., Dement, W., and Nino-Murcia, G. (1988), "Determinants of Daytime Sleepiness in Obstructive Sleep Apnea," *Chest*, 94, 32–37. [776]
- Hyvärinen, A., and Oja, E. (2000), "Independent Component Analysis: Algorithms and Application," *Neural Networks*, 13, 411–430. [788]
- Hyvärinen, A., Karhunen, J., and Oja, E. (2001), *Independent Component Analysis*, New York: Wiley. [788]
- Jolliffe, I. T. (2002), *Principal Component Analysis*, New York: Springer. [788]
- Karhunen, K. (1947), "Über lineare Methoden in der Wahrscheinlichkeitsrechnung," *Annales Academiæ Scientiarum Fennicæ, Series A1: Mathematica-Physica, Suomalainen Tiedekatemia*, 37, 3–79. [776,778]
- Kingshott, R. N., Engleman, H. M., Deary, J. J., and Douglas, N. J. (1998), "Does Arousal Frequency Predict Daytime Function," *European Respiratory Journal*, 12, 1264–1270. [776]
- Loève, M. (1945), "Fonctions Aleatoire de Second Ordre," *Comptes Rendus de l'Académie des Sciences*, 220. [776,778]
- Martin, S. E., Wraith, P. K., Deary, J. J., and Douglas, N. J. (1997), "The Effect of Nonvisible Sleep Fragmentation on Daytime Function," *American Journal of Respiratory and Critical Care Medicine*, 155, 1596–1601. [776]
- Quan, S., Howard, B., Iber, C., Kiley, J., Nieto, F., et al. (1997), "The Sleep Heart Health Study: Design, Rationale, and Methods," *Sleep*, 20, 1077–1085. [776]
- Ramsay, J., and Silverman, B. (2006), *Functional Data Analysis*, New York: Springer-Verlag. [779]
- Redline, S., Sanders, M. H., Lind, B. K., Quan, S. F., Iber, C., Gottlieb, D. J., Bonekat, W. H., Rapoport, D. M., Smith, P. L., and Kiley, J. P. (1998), "Methods for Obtaining and Analyzing Unattended Polysomnography Data for a Multicenter Study," *Sleep*, 21, 759–767. [776]
- Seber, G. A. F. (1984), *Multivariate Observations*, New York: Wiley. [788]
- Spiegelhalter, D. J., Thomas, A., Best, N. G., and Lunn, D. (2003), *WinBUGS Version 1.4 User Manual*, Cambridge: MRC Biostatistics Unit. [780]
- Staicu, A.-M., Crainiceanu, C. M., and Carroll, R. J. (2010), "Fast Methods for Spatially Correlated Multilevel Functional Data," *Biostatistics*, 11 (2), 177–194. [782]
- Whitney, C. W., Gottlieb, D. J., Redline, S., Norman, R. G., Dodge, R. R., Shahar, E., Surovec, S., and Nieto, F. J. (1998), "Reliability of Scoring Respiratory Disturbance Indices and Sleep Staging," *Sleep*, 21, 749–757. [776]
- Wold, H. (1985), "Partial Least Squares," in *Encyclopedia of Statistical Sciences*, eds. S. Kotz and N. L. Johnson, New York: Wiley. [788]
- Wold, S., Ruhe, A., Wold, H., and Dunn, W. (1984), "The Collinearity Problem in Linear Regression. The Partial Least Squares (PLS) Approach to Generalized Inverses," *Journal on Scientific and Statistical Computing*, 5, 735–743. [788]
- Woodard, D., Crainiceanu, C., and Ruppert, D. (2012), "Population Level Hierarchical Adaptive Regression Kernels," unpublished manuscript, available at <http://ecommons.cornell.edu/bitstream/1813/21991/2/WoodCraiRupp2010.pdf>. [783]
- Yao, F., Müller, H.-G., and Wang, J.-L. (2005), "Functional Data Analysis for Sparse Longitudinal Data," *Journal of the American Statistical Association*, 100, 577–590. [779]

Nicole A. LAZAR

This interesting article by Crainiceanu, Caffo, Luo, Zippunikov, and Punjabi presents a novel way of analyzing populations of images so as to pick out their main common features and provide a framework for inference. The analysis is based on sequential applications of a singular value decomposition (SVD)—first at the level of subjects, to pick out the main features of the individual data, and then at the level of groups to detect common features across the sample.

As described by the authors, the method is applicable to very large datasets, which are increasingly prevalent in all areas of science. Such datasets pose many new challenges to the statistics community, and it behooves us to meet those challenges head on. I commend Crainiceanu et al. for demonstrating one way in which existing technology can be adapted to a new purpose.

My discussion centers on the default population value decomposition (PVD) method and concentrates on two main issues: (a) scalability and (b) between- and within-group variation.

1. SCALABILITY

Often when we discuss statistical techniques—new or existing—the question of scalability arises. Typically we are concerned with how well the method “scales up,” especially considering the massive datasets that are now commonly collected and analyzed. The population value decomposition (PVD) method proposed by Crainiceanu et al. prompts the opposite question—namely, how well does it “scale down”? The PVD procedure was devised for extremely large datasets, and it is worth noting that the datasets need to be large *along every dimension*—number of subjects as well as size of the image. The limiting factor in the method is the smallest of these dimensions, because that will determine how many components can be calculated in the various SVDs (equivalently, how many eigenvalues will be nonzero) that make up the method.

This is not necessarily a drawback. Obviously, some researchers do have access to datasets that are truly large in the sense that I describe here and the authors describe in the article: very high-resolution image data collected for a very large number of subjects, for example. Crainiceanu et al. mention functional magnetic resonance imaging (fMRI) data as one possible area of application for the PVD method. However, I wonder if this is really so; fMRI data might not be massive enough, which will no doubt come as some surprise to many who work in this area! The limiting factor here, it seems to me, is the number of subjects. For the colleagues with whom I work, a “large” study is one with several dozen subjects at most; more often, a single group of experimental subjects (e.g., individuals with autism or schizophrenia) will have perhaps a dozen subjects, because these people are hard to recruit, as well as hard to image. With so few subjects, the consequence would appear to be

that only a small number of components could be retained at the individual-level SVDs. Important or interesting features might be missed as a result. Alternatively, depending on the resolution of the image, a small number of subjects might require retaining a larger number of components, thereby potentially introducing unwanted noise. Presumably, this would be picked up and discarded by the group-level decomposition, and so seems a less serious problem.

Although the authors do not emphasize the choice of the number of individual-level components to retain (and whether this number should be the same for all subjects), I think that in real applications there are likely to be some interesting—and perhaps as-yet unforeseen—consequences of this choice. In particular, I would be interested to see how the resolution of the image data and the number of available subjects interact to determine how many components to keep, and how robust the conclusions of the PVD analysis are to the decisions that different researchers might make, given the same constraints on the size of the data.

2. WITHIN AND ACROSS GROUPS OF IMAGES

Understanding variability across subjects is important with all types of data, including image data. Yet this critical piece of the statistical puzzle is often given scant attention, perhaps because of the difficulty of quantifying image variability. Similarly, a question of great interest in many applications is variability across groups: whether, and how, groups differ. Are genes expressed differently in cancer versus noncancer subjects? Do fMRI maps for schizophrenia patients differ from those of healthy controls? And so on. Finding and statistically quantifying such differences on image data is challenging, because what constitutes a “difference” in this context is not entirely clear. Among the issues to consider: Should we look locally, at particular areas of interest within the image, or globally? (The answer to this question could be application-specific.) What is an appropriate measure of difference or distance for images? (Some answers to this question are found in the machine learning literature.) How can we determine whether a difference is statistically significant? (This requires deriving the distributions of the distance measures, which might not be straightforward.) The PVD method provides an approach to these questions of assessing within-group variability and across-group differences.

Specifically, within a group, some of the questions of interest include: How many components from the individual-level SVDs should be retained (a question that is not a point of focus in the article, but nonetheless merits some attention)? How does subject-to-subject variability manifest in the PVD analysis?

To explore some of the issues, I carried out a series of small simulations. In the first simulation, I created a 150×100 ma-

trix of $N(1, 1)$ data for each of 30 “subjects” (one image per subject). Within that matrix, I planted a patch of $N(3, 1)$ observations; the location and size of the patch varied for each subject by taking the starting row and column of the patch to be uniformly distributed on $(11, 15)$ and the ending row and column of the patch to be uniformly distributed on $(25, 30)$.

I then applied the default PVD procedure, keeping the first five or the first 10 components for each subject. Although cumulatively these explain only a small percent of the overall variability, the first two components obviously tend to be more informative than any of the others. The first thing to note is that in this simple setting, it makes no difference whether five components or 10 components per subject are retained. Figure 1 shows scatterplots of the first four components of \mathbf{P} in the PVD decomposition (similar results hold for the \mathbf{D} components), when five or 10 components are retained for each subject. As can be seen, the first two components, which are the “informative” ones that capture the horizontal and vertical nature of the patch with the higher mean, have essentially the same loadings in both cases. The third and fourth components are examples of “noninformative” components in the overall decomposition, and their values are randomly scattered, as we would expect.

The informative nature of the first two components of \mathbf{P} can be further seen in Figures 2 and 3. The former gives boxplots of the (absolute) loadings for each of the first five components of \mathbf{P} , whereas the latter shows selected scatterplots of component loadings. From Figure 2, it is apparent that the general behavior of the first two components is different from that of the next three. For example, both the first and second component loadings have much less variability, although the second component loadings also exhibit quite a few outliers. Figure 3 again demonstrates that the first two components capture essentially all of the information in the individual images.

Regarding subject-to-subject variability, we can gain some insight from Figure 4, which shows the (absolute) loadings on the first two components for each subject, together with the corresponding values from the \mathbf{P} matrix in the group decomposition. The individual loadings show a good deal of variability, especially around the location of the patch with higher mean. This is expected, given that the specific location varied from subject to subject. The first two components of the \mathbf{P} matrix follow the general trends in the individual data quite closely; not surprisingly, when we look at components beyond the first

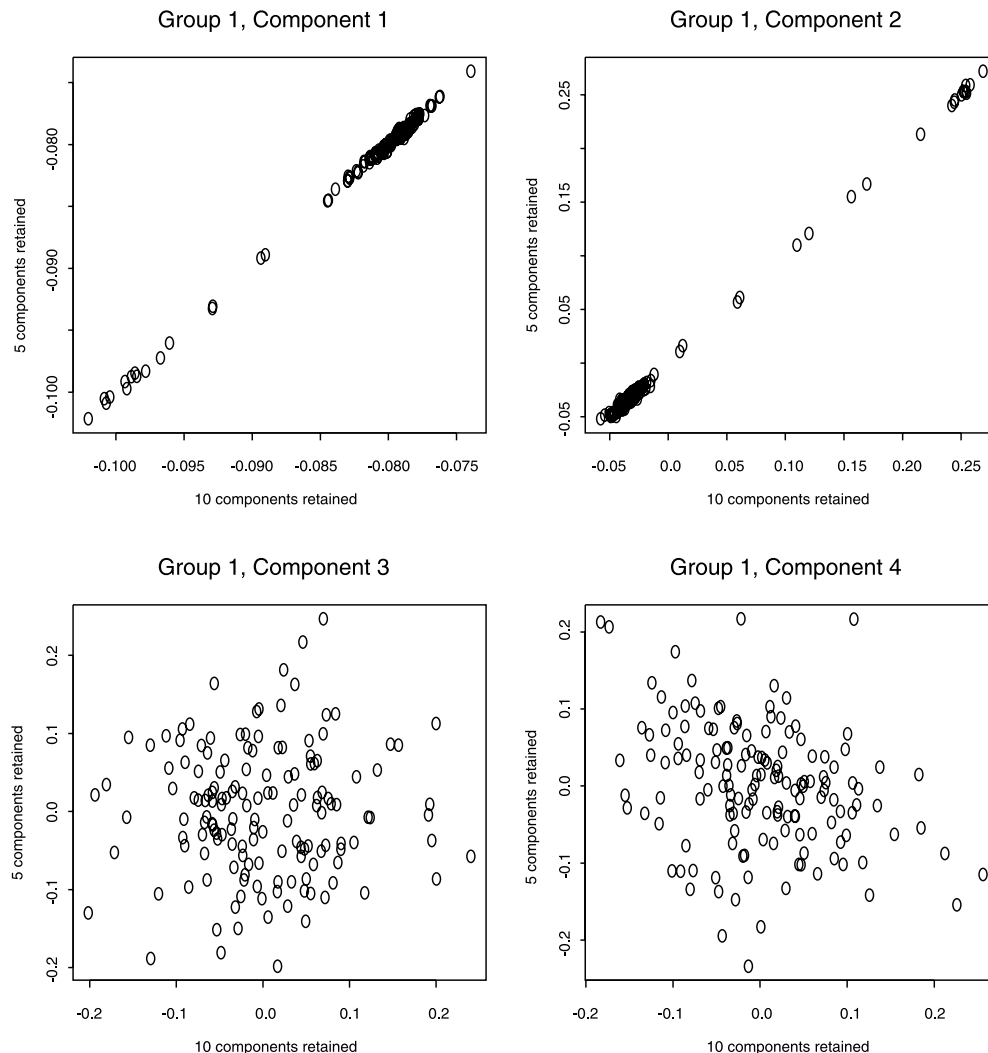


Figure 1. First four components from the PVD, retaining either 5 or 10 components from each subject. Only the first two components are expected to be informative.

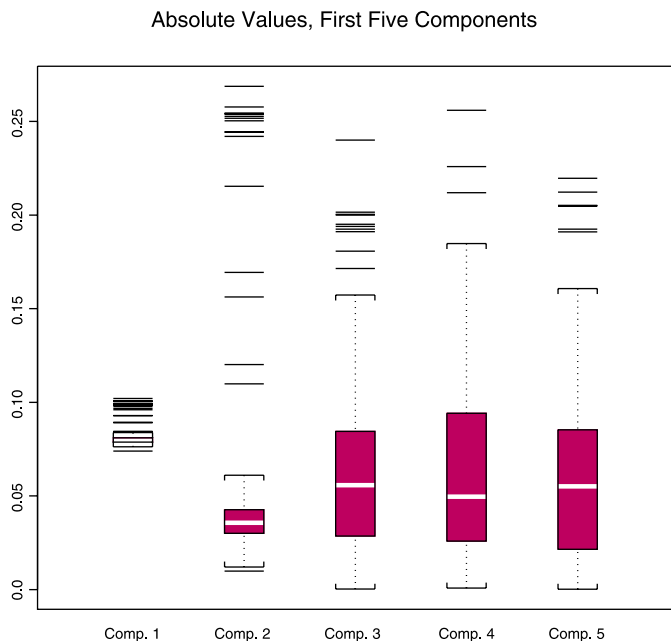


Figure 2. Boxplots of the (absolute) component loadings for the first five components. The behavior of the first two components differs markedly from that of the other three, another indication of their informative nature. The online version of this figure is in color.

two, there is little of interest to detect at either the individual or the group level. The high loadings on the first two components are between 13 and 28, roughly corresponding to the lo-

cation in the x and y directions of the planted patch for most subjects.

Although I have described the behavior of the \mathbf{P} matrix and its components, I note that very similar statements hold for the \mathbf{D} matrix and its components. In what follows, I continue to describe results of the \mathbf{P} matrix.

For the second simulation, I planted two patches of higher mean, again letting the precise locations and sizes of the patches vary within a small range for each subject. Artificial data were generated for 20 subjects, and the first 10 individual-level components were retained in each case. Now the first three components of the \mathbf{P} part of the overall decomposition contain informative content. The scatterplots and variability plots for this simulation shown in Figures 5 and 6 largely confirm the conclusions of the more detailed analysis of the “one-patch” situation.

Finally, for the third simulation, I looked at the problem of comparing two groups, and what insight PVD can lend to this more difficult, and interesting, case. As in the first simulation, for each subject in each group I planted a patch of higher mean within the $N(1, 1)$ background, with patch location and size varying within and across groups; that is, within a group, the patch location varied as before around some central values, which differed in the two groups. The first group was the same as in the first simulation. For the second group, a patch of $N(4, 1)$ intensity was planted from the starting row and column uniformly distributed on $(16, 20)$ and the ending row and column distributed uniformly on $(30, 34)$. Thus the interesting areas are close together but nonoverlapping.

I considered two approaches for detecting differences between the two groups of subjects: (a) comparing the PVD com-

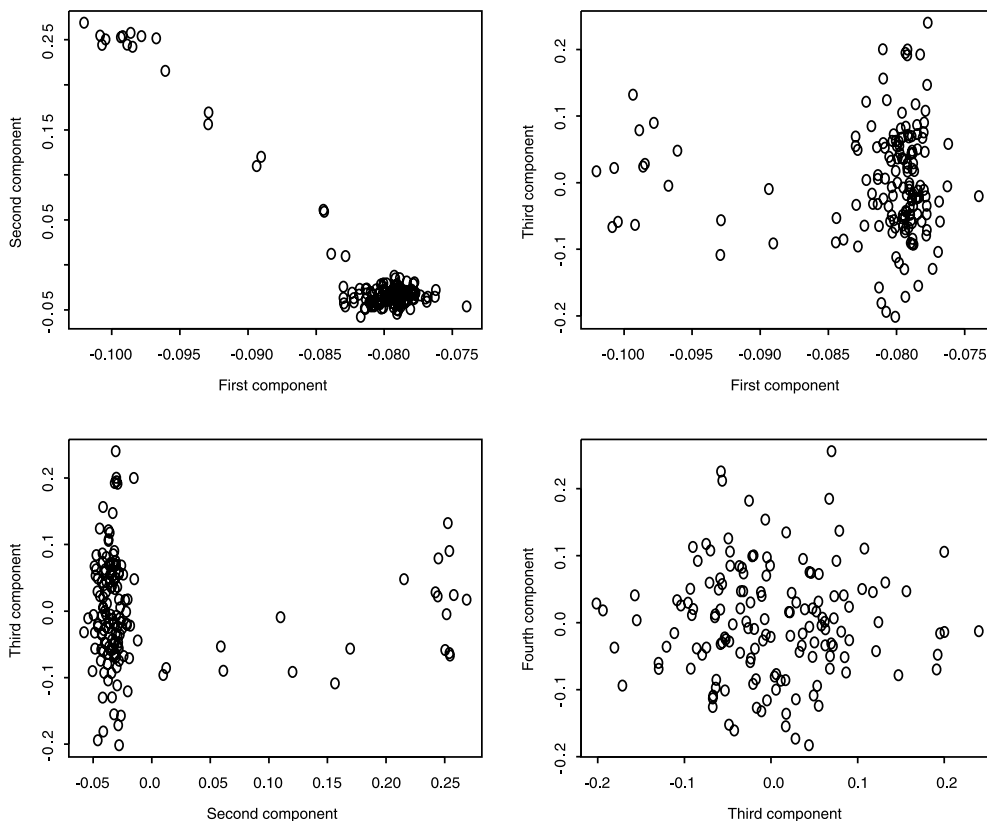


Figure 3. Scatterplots of pairs of component loadings. Noninformative components 3 and 4 show no interesting joint behavior.

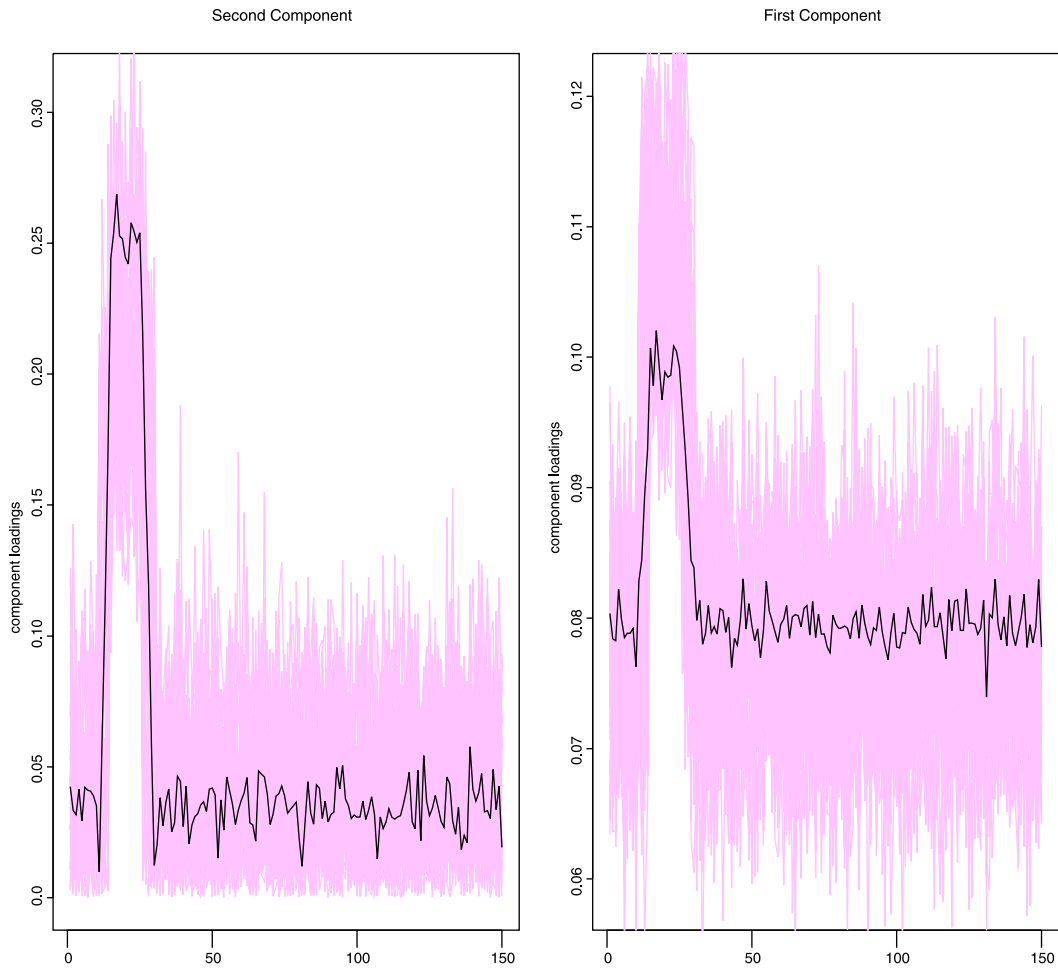


Figure 4. First two components from the individual subject SVDs, along with those from the PVD. The online version of this figure is in color.

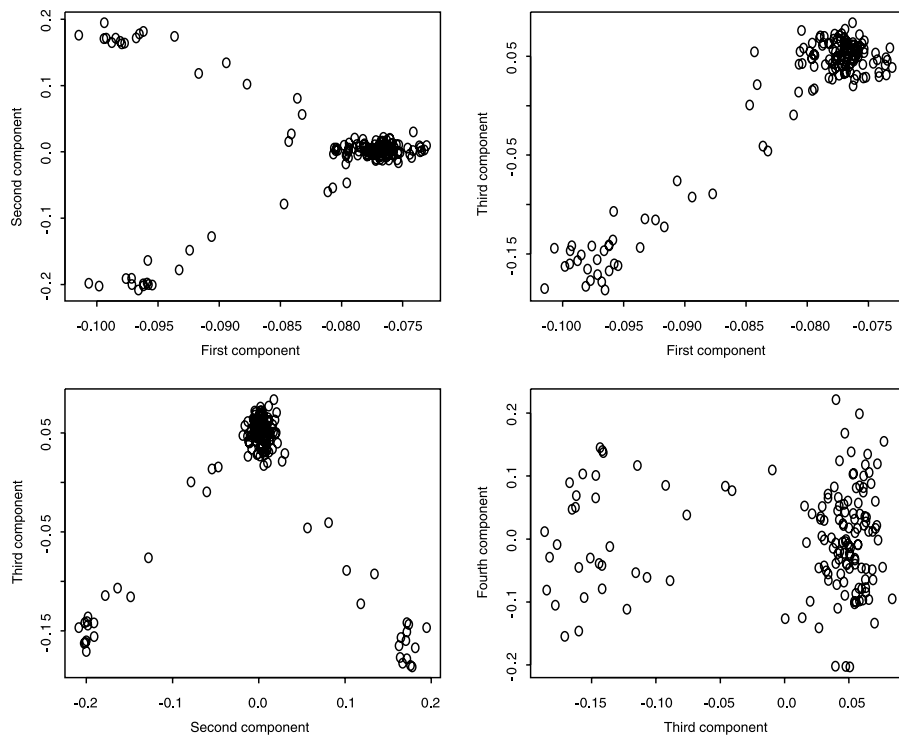


Figure 5. Scatterplots of the first four components, with two patches of higher mean planted on the background noise image. The first three components are informative.

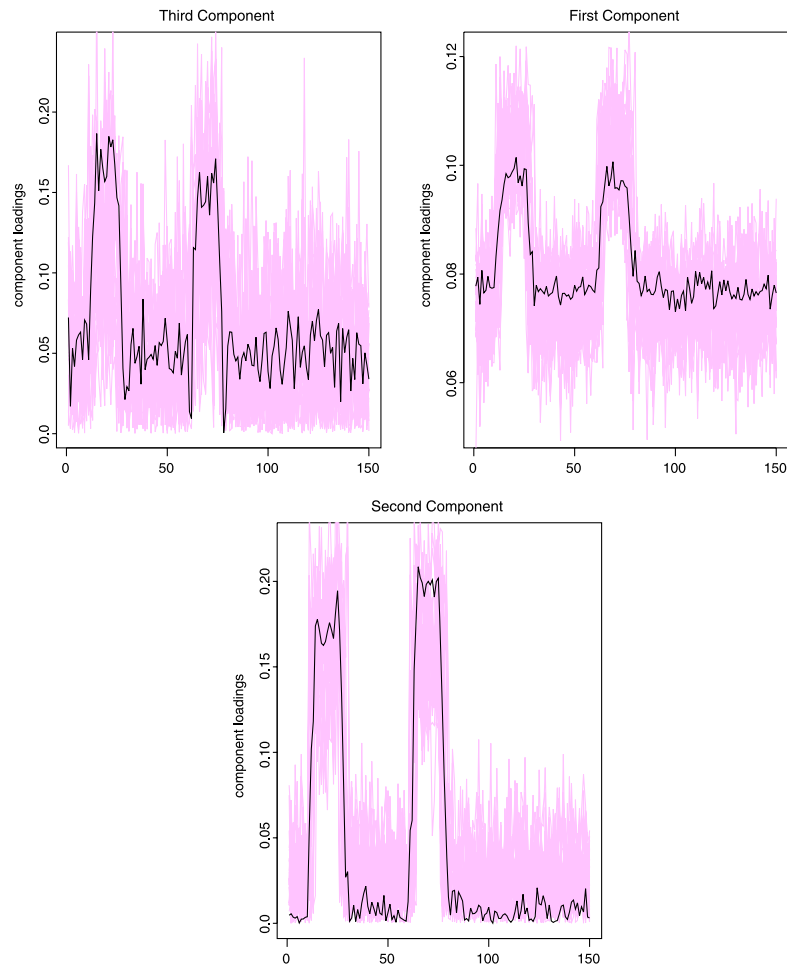


Figure 6. First three components from the individual subject SVDs, along with those from the PVD. The online version of this figure is in color.

ponents obtained from analyzing each group separately and (b) performing a “combined PVD” by retaining components from all subjects in both groups, concatenating those similarly to the single group PVD, and carrying out the PVD analysis on the combination.

Plots of the first two components in Figure 7 show the results of the combined analysis, as well as the two individual group-level analyses. As is clear from the figure, the combined analysis provides a compromise between the two group-level PVDs, and this detects the difference in the general locations of the patches. It is not clear that this would have been picked up by the combined analysis alone; rather, it is the juxtaposition of the two results from the two approaches that reveals the information of interest.

A third component turns out to be informative in the combined analysis. This is also presented in Figure 7, along with the differences between the first two components from each of the group-level PVDs. The difference between the second components from each of the groups closely matches the third component in the combined analysis.

3. CONCLUSION

Based on the three (admittedly basic) simulation studies that I carried out using the default PVD method, it is apparent that

dominant trends in the data can be detected. Simple structure in the images is captured by a small number of group-level components. In the simulation settings, the particular choice of the number of individual-level SVD components to retain is not critical; however, for more complicated and realistic image data, this choice will be more important. Group differences also can be gleaned from a combined PVD approach, and I suggest that this should be done together with group-level ordinary PVD to gain additional insight into the structure of the discrepancies.

I agree with the authors that visualization is key to helping statisticians and scientists understand the massive datasets that we now have to handle on a daily basis. Seeing the entire raw dataset on a single plot is no longer possible, and we need to find ways of determining the truly interesting geometrical directions, ideally along a low dimension so that visualization is possible. Again, PVD provides guidance here, with a number of natural graphical representations suggesting themselves. Plotting the components is a first step. Some of the other questions of interest, such as the number of informative components and the robustness of results to the number of components retained per individual, also can benefit from a graphical perspective.

I congratulate the authors for a stimulating article.

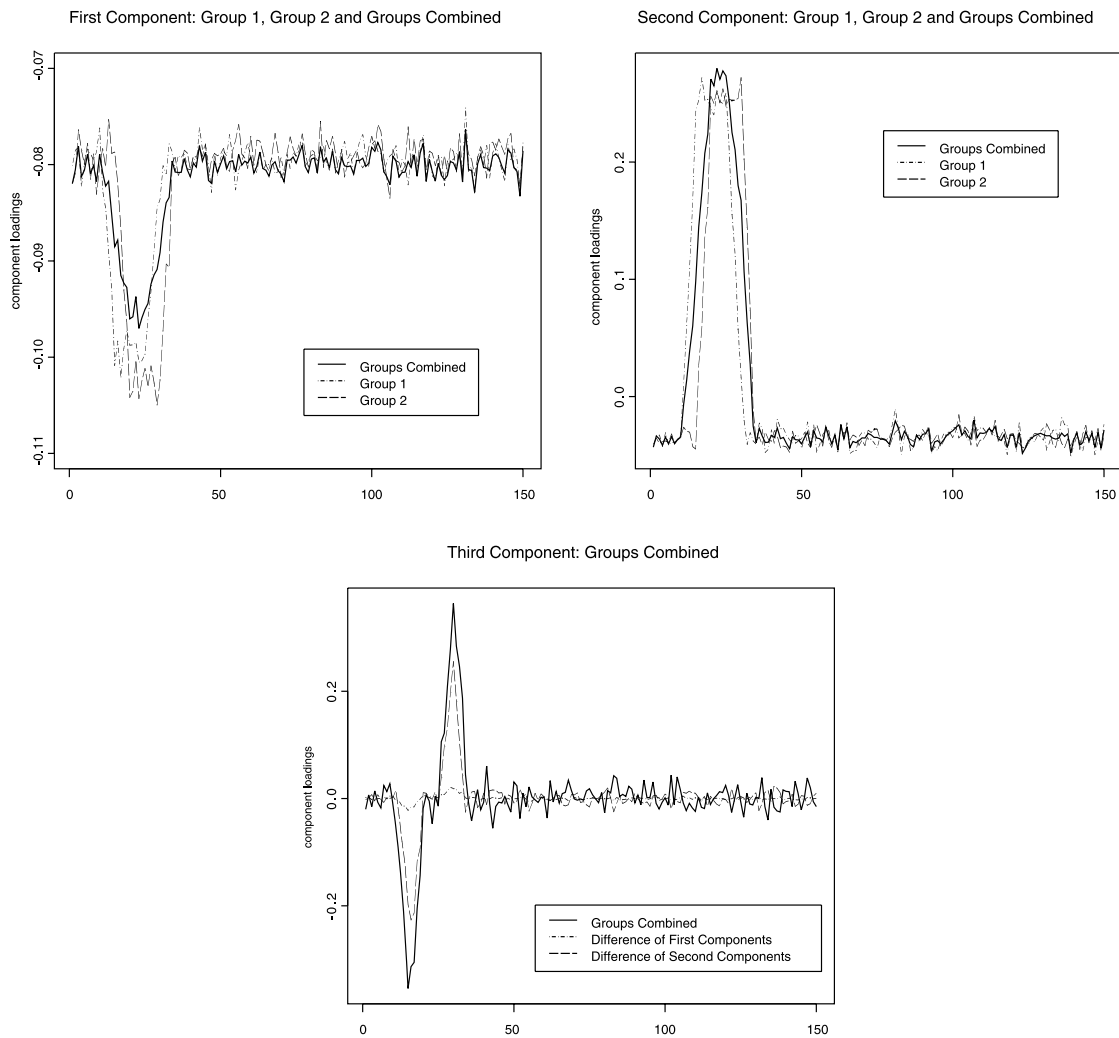


Figure 7. First three components from the combined PVD, along with corresponding components or functions of components from the group-level PVDs.

Comment

Kerby SHEDDEN

The population value decomposition (PVD) proposed by Crainiceanu et al. addresses the problem of analyzing large collections of images. I share the authors' enthusiasm for this area of research, and appreciate their aim of building a straightforward, scalable method for summarizing large image collections. As the authors illustrated with sleep EEG data, the PVD can be a highly effective method for summarizing a large collection of large images, provided that the structure of the image collection is compatible with the representation used by the PVD. My comments here address the important issue of the range of data settings in which the PVD approach may be effective.

Individual images are enormously complicated forms of data. To get a sense of what we are dealing with, just consider that

images are capable of representing, to some degree of approximation, every conceivable scene in the history or future of the Earth. At the same time, it has been repeatedly noted that meaningful images are highly structured, with "real" images comprising a vanishingly small set of all possible images. The field of computer vision has made some progress on the problem of abstracting the meaning from unconstrained images (e.g., Wu, Guo, and Zhu 2008), but much work remains to be done. Given these circumstances, needless to say the problem of "summarizing" images (or collections of images) is a daunting task.

As an image summarization method, the PVD aims to extract the most important features that vary across a collection of im-

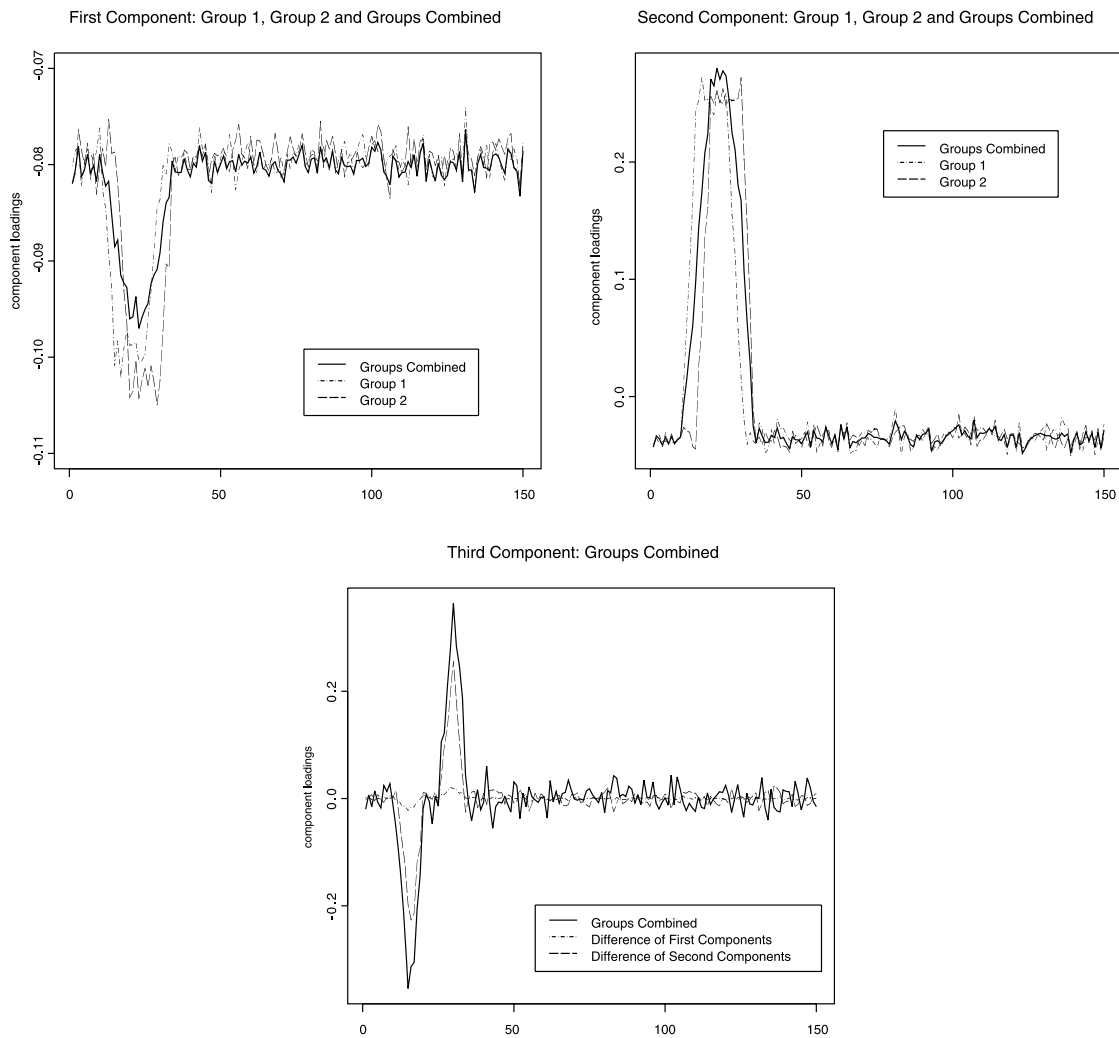


Figure 7. First three components from the combined PVD, along with corresponding components or functions of components from the group-level PVDs.

Comment

Kerby SHEDDEN

The population value decomposition (PVD) proposed by Crainiceanu et al. addresses the problem of analyzing large collections of images. I share the authors' enthusiasm for this area of research, and appreciate their aim of building a straightforward, scalable method for summarizing large image collections. As the authors illustrated with sleep EEG data, the PVD can be a highly effective method for summarizing a large collection of large images, provided that the structure of the image collection is compatible with the representation used by the PVD. My comments here address the important issue of the range of data settings in which the PVD approach may be effective.

Individual images are enormously complicated forms of data. To get a sense of what we are dealing with, just consider that

images are capable of representing, to some degree of approximation, every conceivable scene in the history or future of the Earth. At the same time, it has been repeatedly noted that meaningful images are highly structured, with "real" images comprising a vanishingly small set of all possible images. The field of computer vision has made some progress on the problem of abstracting the meaning from unconstrained images (e.g., Wu, Guo, and Zhu 2008), but much work remains to be done. Given these circumstances, needless to say the problem of "summarizing" images (or collections of images) is a daunting task.

As an image summarization method, the PVD aims to extract the most important features that vary across a collection of im-

ages. The PVD does this using a particular form of matrix factorization. As I noted earlier, the sleep EEG data application is an impressive example of how powerful the PVD approach can be. However, these EEG images have a very particular structure, because the preprocessing step using the Fourier transform neatly encodes complementary information in the two image coordinates. These are really two-dimensional “data arrays” rather than images in the usual sense. Moreover, because the rows and columns are indexed by well-defined units, it is natural to think in terms of image summaries that, like the PVD, take linear combinations of the rows and columns to reduce the dimension.

The authors allude to the possibility of using their methodology for analyzing fMRI data. I use this application to make my comments more concrete, although similar issues should arise when the PVD is applied to images collected in other settings. Common questions that arise in fMRI analysis are related to identification of active regions of the brain and “connectivity” between brain regions.

Suppose that we have two brain regions whose indicator functions are I_1 and I_2 , and the levels of activation in the two regions are viewed as random variables λ_{i1} and λ_{i2} (for subject i); that is, if J_i is the image for subject i , then $E(J_i|\lambda_{i1}, \lambda_{i2}) = \lambda_{i1}I_1 + \lambda_{i2}I_2$. This might appear to be an ideal situation for the PVD, given that the locations of activation are shared by all members of the population (as are the \mathbf{P} and \mathbf{D} matrices in the PVD), whereas the intensity and dependence of the activation levels may vary over subjects (as do the \mathbf{V}_i matrices in the PVD). We may hope to attain a two-component solution, with the two components corresponding to the two regions (if the activations represented by λ_{i1} and λ_{i2} are independent), or to two distinct contrasts of the two regions if the activations are dependent (potentially reflecting connectivity). Indeed, the PVD may work well here, but to get the most parsimonious representation, we have to be a bit lucky with the geometry of the regions. If the regions are rectangular and parallel to the image axes, then I_1 and I_2 can be represented using a single outer product, thus requiring only one left dimension ($L_i = 2$) and one right dimension ($R_i = 2$) for each region. However, other geometries are less amenable to this representation. For example, a disk of radius 50 embedded at the center of a 100×100 image requires a sum of five rank-one terms to capture 90% of its variability. The fact that the degree of compression is not invariant to rotation of the images seems to be a drawback of the method, at least when applying the PVD directly to pixel-level data.

The situation becomes even more difficult if the spatial locations of the activated areas vary among the subjects. For example, suppose that the image collection captures a wave of activation from the top to the bottom of the image. I generated a collection of 90 100×100 images, in which a 10-pixel wave of activation passes from the top to the bottom of the images. Using $L_i = R_i = 2$ (which captured 90% of the variation in the individual images), and $A = B = 30$ (not a very parsimonious representation), only 79% of the variation in the collection was captured by the PVD. The situation is analogous to the well-known difficulty of using additive combinations of basis functions to capture translation and warping behavior in collections of curves (Gervini and Gasser 2004).

It is also interesting to consider whether we should always take the “frame of reference” (\mathbf{P} and \mathbf{D}) to be universal, whereas

the “loadings” (\mathbf{V}_i) are subject-specific. What if the opposite were true? In a somewhat extreme case, we may have a complex task that the subjects approach using highly diverse mental strategies. Here we may find that different regions of the subjects’ brains are activated (requiring subject-specific \mathbf{P} and \mathbf{D} matrices), whereas the \mathbf{V}_i have low heterogeneity (reflecting common timing of the stimulus and low variability in response times). As pointed out by the authors, there may be “many types of PVDs.” However, at this point we seem to lack both theory and computational tools that would apply to the PVD in this broader sense.

My comments about fMRI analysis assume that the images are to be analyzed at the pixel level. The EEG data were transformed to the time and frequency scales. Other types of images may have their own natural two-way representations. For example, one could use a two-dimensional wavelet transform to represent an image in terms of location and scale parameters. But difficulties may yet arise. The wavelet location parameters would be most naturally indexed in two dimensions, giving three dimensions in all [direct generalization of the PVD to three-way arrays would be challenging, given the complex computational and identification issues for SVD-like decompositions of multiway arrays (De Lathauwer, De Moor, and Vandewalle 2000)]. This issue does not arise in the EEG study because the data are not spatial (at least not as presented here).

Traditionally, analysis of natural images relies heavily on a feature extraction step (e.g., Wainwright, Simoncelli, and Willsky 2001; Mikolajczyk and Schmid 2005), with the features (typically in the form of a vector) used to perform a task such as image classification. The features are often carefully constructed to respect certain geometric invariances in a problem, to resist influence from artifacts such as illumination variation or occlusion, or to account for nuisance forms of variation such as changes in pose. A notable aspect of the PVD, at least as suggested by Crainiceanu et al., is that it works directly with the pixel-level data. Although this has the advantage of being simple and direct, it remains to be seen whether there are many applications where, like the EEG example, the structure in the images can be effectively represented by the PVD or related approaches.

ADDITIONAL REFERENCES

- De Lathauwer, L., De Moor, B., and Vandewalle, J. (2000), “A Multilinear Singular Value Decomposition,” *SIAM Journal on Matrix Analysis and Applications*, 21 (4), 1253–1278. [797]
- Gervini, D., and Gasser, T. (2004), “Self-Modelling Warping Functions,” *Journal of the Royal Statistical Society, Ser. B*, 66 (4), 959–971. [797]
- Mikolajczyk, K., and Schmid, C. (2005), “A Performance Evaluation of Local Descriptors,” *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27 (10), 1615–1630. [797]
- Wainwright, M. J., Simoncelli, E. P., and Willsky, A. S. (2001), “Random Cascades on Wavelet Trees and Their Use in Analyzing and Modeling Natural Images,” *Applied and Computational Harmonic Analysis*, 11 (1), 89–123. [797]
- Wu, Y. N., Guo, C.-E., and Zhu, S.-C. (2008), “From Information Scaling of Natural Images to Regimes of Statistical Models,” *Quarterly of Applied Mathematics*, 66, 81–122. [796]

E. F. LOCK, A. B. NOBEL, and J. S. MARRON

The article by Crainiceanu et al. addresses an important and relatively undeveloped area of statistical research: the analysis of populations in which the data objects are matrices. In particular, they focus on collections of matrices that have the same row and column dimensions. Such datasets are increasingly prevalent in a number of scientific fields. Examples range from the analysis of facial data in image analysis, EEG data in neuroscience, fMRI data in medical imaging, and browsing data in the study of Internet traffic (see Table 1).

The method proposed by the authors, population value decomposition (PVD), is a useful way to simultaneously reduce the dimensionality of a collection of matrices. For matrices $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$, each of dimension $F \times T$, the PVD yields an approximation $\mathbf{Y}_i \approx \mathbf{P}\mathbf{V}_i\mathbf{D}$ for $i = 1, \dots, n$. The low-dimensional representation \mathbf{V}_i for each matrix is on a common set of coordinates determined by \mathbf{P} and \mathbf{D} . This allows for the application of standard statistical approaches, such as regression and cluster analysis, to the lower-dimensional matrices \mathbf{V}_i , rather than the \mathbf{Y}_i 's. Furthermore, inspection of the population-wide left and right loading matrices \mathbf{P} and \mathbf{D} can aid in identifying the primary modes of variation among a population of matrices.

The authors present an interesting data analysis; however, we note that a model formally equivalent to PVD has been proposed in the computer science literature under the name *two-dimensional singular value decomposition* (2DSVD) (Ding and Ye 2005; Ye 2005). This literature also provides natural additional approaches for choosing the population-wide matrices \mathbf{P} and \mathbf{D} . We discuss these approaches in Section 1.

We may regard the problem addressed in these articles by viewing the data as a three-way ($F \times n \times T$) array. In Section 2 we discuss and compare alternative approaches that treat the data structure as a three-way array. We show that two SVD-like decompositions for higher-order arrays, those of *Candecomp/Parafac* (Carroll and Chang 1970) and *Tucker* (Tucker 1966), are related to the PVD decomposition. In fact, both can be represented in the form $\mathbf{P}\mathbf{V}_i\mathbf{D}$, in which the \mathbf{V}_i matrices have a particular structure.

In Section 3 we discuss some important issues and caveats related to the application of PVD and related methods. In Section 4 we compare PVD and other methods in an application to facial image data.

1. CHOICE OF \mathbf{P} AND \mathbf{D}

The PVD article suggests determining the entries of the individual matrices \mathbf{V}_i via standard least squares regression, for a given choice of the population-wide matrices \mathbf{P} and \mathbf{D} . However, the default method for choosing \mathbf{P} and \mathbf{D} is somewhat ad

hoc, and a more principled approach would be desirable. We suggest formulating the estimation of \mathbf{P} , \mathbf{D} , and the \mathbf{V}_i matrices together as a single least squares problem. That is, for given dimensions $A < F$ and $B < T$, find $\mathbf{P}: F \times A$, $\mathbf{D}: B \times T$, and $\mathbf{V}_i: A \times B, i = 1, \dots, n$, to minimize the sum of squared residuals

$$\sum_{i=1}^n \|\mathbf{Y}_i - \mathbf{P}\mathbf{V}_i\mathbf{D}\|_F^2. \quad (1)$$

Here $\|\cdot\|_F$ defines the Frobenius norm; that is, $\|\mathbf{A}\|_F^2$ is just the sum of the squared entries of \mathbf{A} .

This approach to estimating the PVD model was previously explored by Ye (2005). Ye suggested an iterative least squares procedure that cycles among estimation of the matrices \mathbf{P} , $\mathbf{V}_1, \dots, \mathbf{V}_n$, and \mathbf{D} until convergence. Although this iterative procedure is not guaranteed to achieve the global minimum in criterion (1), Ye argued that the algorithm is insensitive to starting conditions and is generally successful at minimizing the sum of squared residuals.

An alternative approach for choosing \mathbf{P} and \mathbf{D} , termed 2DSVD by Ding and Ye (2005), makes use of the aggregated row–row and column–column covariance matrices. In particular, \mathbf{P} is determined by the first A singular vectors of the row-by-row covariance matrix $\frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i\mathbf{Y}_i'$ and \mathbf{D} determined by the first B singular vectors of the column-by-column covariance matrix $\frac{1}{n} \sum_{i=1}^n \mathbf{Y}_i'\mathbf{Y}_i$. Equivalently, \mathbf{P} can be computed as the first A left singular vectors of the aggregated matrix $[\mathbf{Y}_1 \ \mathbf{Y}_2 \ \dots \ \mathbf{Y}_n]$, and \mathbf{D} can be computed as the first B right singular vectors of $[\mathbf{Y}_1' \ \mathbf{Y}_2' \ \dots \ \mathbf{Y}_n']'$. Although both computations give the same result, the latter may be more efficient if one of the dimensions is particularly large, and computing the covariance is impractical.

The justification for the 2DSVD algorithm is that the columns of \mathbf{P} are chosen as the set of left-singular vectors that explain the most total variation across the columns of $\mathbf{Y}_1, \dots, \mathbf{Y}_n$, and the rows of \mathbf{D} are (independently) chosen as the set of right singular vectors that explain the most variation across the rows of $\mathbf{Y}_1, \dots, \mathbf{Y}_n$. Because interactions between \mathbf{P} and \mathbf{D} are not accounted for, the resulting matrices do not necessarily minimize criterion (1), but they should come close to doing so. Indeed, 2DSVD could be used to determine the initial matrices \mathbf{P}_0 and \mathbf{D}_0 for an iterative least squares procedure, such as that described earlier.

The “default” method for estimating \mathbf{P} and \mathbf{D} proposed in the PVD article is essentially a two-stage SVD. The first few singular vectors of each matrix are found separately, then another SVD of the combined singular vectors determines the global left and right singular vectors \mathbf{P} and \mathbf{D} . This method requires

Eric F. Lock is Doctoral Student (E-mail: Eric.F.Lock@gmail.com), Andrew B. Nobel is Professor (E-mail: nobel@email.unc.edu), and J. S. Marron is Professor (E-mail: marron@email.unc.edu), Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC 27514. This research was supported in part by National Institutes of Health grant R01 MH090936-01 and National Science Foundation grant DMS-0907177.

Table 1. Data examples where the objects are matrices of the same dimension

Objects	Value	Dimensions
Facial images	pixel intensity	horizontal \times vertical
EEG recordings	electrical activity	frequency \times time
fMRI scans	blood flow	voxel position \times time
Browsing histories	visits from website i to website j	websites \times websites

specifying the number of singular vectors to take for each matrix, which may be somewhat arbitrary. We feel that the 2DSVD and least squares methods described earlier are more justified, and in most cases will be computationally simpler. However, in some datasets the aggregated matrix $[\mathbf{Y}_1 \ \cdots \ \mathbf{Y}_k]$ and/or one of $\mathbf{Y}'\mathbf{Y}$, $\mathbf{Y}\mathbf{Y}'$ are too large to store in memory. In these cases, the individual-level data compression used by the two-stage SVD may be necessary.

Although we have presented various alternative methods for selecting \mathbf{P} and \mathbf{D} , we share the authors' opinion that the ideal choice of \mathbf{P} and \mathbf{D} may depend on the particular type of data, and there is no perfect method.

2. THREE-WAY METHODS

The authors apply PVD to data in which an EEG-based activity matrix of *Frequency \times Time* is available for multiple subjects. Note that these data can be framed as a three-way array: *Frequency \times Subject \times Time*. Indeed, any dataset analyzed with PVD can be considered a three-way array of dimension $F \times n \times T$. In PVD, the second mode (subject, of dimension n) is treated differently than the other two modes. In this section we consider some other SVD-like decompositions for multiway arrays that treat all three modes similarly. These methods are also appropriate for multiway arrays with more than three modes. Thus they have potential for the analysis of fMRI data that is truly multidimensional: *Length \times Width \times Height \times Time \times Subject*.

There are two standard SVD-like extensions to multiway data: the Candecomp/Parafac and Tucker decompositions. Both have been studied in the analysis of tensors for several years, but are not widely known. The survey by [Kolda and Bader \(2009\)](#) is a well-written and accessible introduction to tensor notation, the aforementioned decompositions, and related software. We briefly discuss their relationship to PVD here, but refrain from using notation that may be unfamiliar.

2.1 The Candecomp/Parafac Decomposition

The Candecomp/Parafac ([Carroll and Chang 1970](#)) decomposition extends the notion of the SVD as a sum of rank-1 approximations. We can approximate an $F \times T$ matrix \mathbf{Y} by combining the first r left singular vectors and corresponding right singular vectors; that is, the columns of $\mathbf{U}^{(1)} : F \times r$ are the first r left singular vectors of \mathbf{Y} , and the columns of $\mathbf{U}^{(2)} : T \times r$ are the first r right singular vectors of \mathbf{Y} , scaled appropriately, then

$$y_{ij} \approx \sum_{l=1}^r u_{il}^{(1)} u_{jl}^{(2)}$$

for $i = 1, \dots, F, j = 1, \dots, T$.

For a three-way array $\mathbf{Y} : F \times n \times T$, then, the Parafac decomposition yields matrices $\mathbf{U}^{(1)} : F \times r$, $\mathbf{U}^{(2)} : n \times r$, and $\mathbf{U}^{(3)} : T \times r$, so that

$$y_{ijk} \approx \sum_{l=1}^r u_{il}^{(1)} u_{jl}^{(2)} u_{kl}^{(3)}$$

for $i = 1, \dots, F, j = 1, \dots, n$, and $k = 1, \dots, T$. The matrix $\mathbf{U}^{(i)}$ serves as a low-dimensional representation for variation in the i th mode.

The three-way Parafac decomposition also can be represented in the framework of the PVD model. If $\mathbf{P} := \mathbf{U}^{(1)}$, $\mathbf{D} := \mathbf{U}^{(3)}$, and \mathbf{V}_j is a diagonal matrix whose entries are from the j th column of $\mathbf{U}^{(2)}$, $\mathbf{V}_j = \text{diag}(\mathbf{U}_j^{(2)})$, then

$$\mathbf{Y}_{\cdot j} \approx \mathbf{P}\mathbf{V}_j\mathbf{D}$$

for $j = 1, \dots, n$. Thus the three-way Parafac decomposition can be considered a PVD model in which the \mathbf{V}_j matrices are diagonal.

2.2 The Tucker Decomposition

For a standard (two-mode) SVD, combining the i th left singular vector and the j th right singular vector does not improve an approximation when $i \neq j$. No such result holds for higher-order arrays. The Tucker decomposition ([Tucker 1966](#)), then, considers all combinations from a set of basis vectors in each mode. Thus a three-way Tucker decomposition consists of matrices $\mathbf{U}^{(1)} : F \times r_1$, $\mathbf{U}^{(2)} : n \times r_2$, and $\mathbf{U}^{(3)} : T \times r_3$ and a $r_1 \times r_2 \times r_3$ tensor $\mathbf{\Lambda}$, where

$$y_{ijk} \approx \sum_{l_1=1}^{r_1} \sum_{l_2=1}^{r_2} \sum_{l_3=1}^{r_3} \lambda_{l_1 l_2 l_3} u_{il_1}^{(1)} u_{jl_2}^{(2)} u_{kl_3}^{(3)}$$

Here the ijk th entry of the tensor $\mathbf{\Lambda}$ weights the interactions among the i th column of $\mathbf{U}^{(1)}$, the j th column of $\mathbf{U}^{(2)}$, and the k th column of $\mathbf{U}^{(3)}$. The Parafac decomposition is a special case of the Tucker model where $r_1 = r_2 = r_3$ and $\lambda_{l_1 l_2 l_3} = 0$ unless $l_1 = l_2 = l_3$. Again, the matrix $\mathbf{U}^{(i)}$ serves as a low-dimensional representation for variation in the i th mode.

The three-way Tucker decomposition also can be given in the PVD framework, where the matrices \mathbf{V}_j have a particular factorized form. If $\mathbf{P} := \mathbf{U}^{(1)}$ and $\mathbf{D} := \mathbf{U}^{(3)}$, then

$$\mathbf{Y}_{\cdot j} \approx \mathbf{P}\mathbf{V}_j\mathbf{D},$$

where $\mathbf{V}_j : r_1 \times r_3$ is

$$\mathbf{V}_{\cdot j} := \sum_{l_2=1}^{r_2} u_{jl_2}^{(2)} \mathbf{\Lambda}_{\cdot l_2 \cdot}$$

Intuitively, here each \mathbf{V}_j can be considered a weighted combination of basis matrices $\mathbf{\Lambda}_{\cdot 1}, \dots, \mathbf{\Lambda}_{\cdot r_2}$, where the weights specific to the j th individual are given by the j th row of $\mathbf{U}^{(2)}$.

3. POTENTIAL ISSUES

There are important caveats in the application of PVD and related methods, such as those discussed in the previous section. We briefly discuss four common issues that must be considered before describing an application of PVD.

3.1 Registration

The PVD approach requires that the coordinates of the matrices $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ be aligned; that is, the (i, j) entry of the matrix \mathbf{Y}_1 must correspond to the (i, j) entry of the matrix \mathbf{Y}_2 , and so on. This is a common issue in the analysis of image populations, where a slight shift or rotation of perspective can cause difficulties when integrating information across the images. Here registration methods that transform a collection of images to the same coordinate system can be useful. For an overview of image registration methods, see the survey by [Zitova and Flusser \(2003\)](#).

3.2 Scaling

Direct application of PVD also may be problematic if the matrices $\mathbf{Y}_1, \mathbf{Y}_2, \dots, \mathbf{Y}_n$ differ in scale. For example, estimation of \mathbf{P} and \mathbf{D} can be unduly influenced by a select few matrices with a large amount of variability. A potential solution is to scale each matrix to have the same total variation, that is, sum of squares. This approach was suggested by, for example, [Lock et al. \(2011\)](#) as a preprocessing step for integrating across multiple data matrices (possibly of different dimension) available for the same set of objects.

To remove baseline differences between matrices, it is helpful to center the data by subtracting the overall mean from each matrix. To control total variability, one can then divide by the standard deviation of the matrix entries; that is, letting \bar{y}_i be the mean and s_i be the standard deviation of the entries of \mathbf{Y}_i , define

$$\mathbf{Y}_i^{\text{scaled}} = \frac{\mathbf{Y}_i - \bar{y}_i}{s_i}.$$

The matrices $\mathbf{Y}_i^{\text{scaled}}$ then have the same total sum of squared entries.

We note, however, that the choice of a normalization procedure should depend on the type of data and goals of the analysis.

3.3 Dimensional Compatibility

Recall that for a rank $r + 1$ SVD approximation, the first r right singular vectors, left singular vectors, and singular values remain the same. If the PVD model is estimated via minimizing the sum of squared residuals, then there is no such dimensional compatibility; that is, if either dimension A or B is changed, then all entries of the estimated matrices \mathbf{P} , \mathbf{D} , and each \mathbf{V}_i may change. This is an important caveat when interpreting the columns and rows of \mathbf{P} and \mathbf{D} . In many cases, changing A or B slightly might not lead to a dramatic change in the entries of \mathbf{P} , \mathbf{D} , and \mathbf{V}_i , but the stability of these estimates are worth considering. The Tucker and Parafac models described in Section 2 also are not necessarily compatible on different dimensions.

3.4 Choice of A and B

In light of the foregoing comments, the choices of A and B in the PVD approximation can be particularly important. In any case, the choices of A and B may be somewhat arbitrary in practice, and a principled approach to choosing these dimensions is desired. We do not give a specific approach here, but note that certain ideas may be borrowed from related work. One potential criterion is a cross-validation–based estimate of the reconstruction error, similar to that used to determine the number of principal components by [Wold \(1978\)](#). Another potential approach is permutation testing, similar to the rank selection procedure described by [Lock et al. \(2011\)](#). Yet another potential approach may be motivated by random matrix theory (see [Shabalyn and Nobel 2010](#)).

4. APPLICATION: FACIAL IMAGES

As an example, we apply PVD and related methods to the Database of Faces procured by AT&T Laboratories Cambridge. This is a publicly available database of $n = 400$ total gray-scale images for 40 individuals (10 per individual). Each image \mathbf{Y}_i , $i = 1, \dots, n$, is 92×112 in size. All subjects are in an upright, frontal position, but facial characteristics (e.g., smiling, not smiling; glasses, no glasses) vary in each image. We apply four factorization models to these data and compare the results. We apply the PVD model in which \mathbf{P} and \mathbf{D} are estimated by iteratively minimizing the sum of squared residuals, as in Section 1. We apply the Parafac and Tucker models, also estimated by least squares using the N-way MATLAB toolbox ([Andersson and Rasmus 2000](#)). We also try a SVD of the vectorized data; for each \mathbf{Y}_i , the rows are stacked to form a vector of length $92 \times 112 = 10,304$, and an SVD is applied to the resulting $10,304 \times 400$ matrix.

We compare these factorized approximations in terms of data compression for this example; that is, we consider the sum of squared residuals versus the total number of degrees of freedom (free parameters) needed for each model. For example, a PVD approximation with $A = B = 5$ requires $\mathbf{P}: 92 \times 5$, $\mathbf{D}: 5 \times 112$, and $\mathbf{V}_i: 5 \times 5$, $i = 1, \dots, 400$, or $92 \times 5 + 5 \times 112 + 5 \times 5 \times 400 = 11,020$ free parameters.

Figure 1(A) displays the sum of squared residuals for each model as the number of free parameters increases. In this analysis, for simplicity, we restrict $A = B$ for the PVD model and $r_1 = r_2 = r_3$ for the Tucker model. Relaxing these restrictions could give these methods additional power. The SVD of the vectorized data is by far the worst-performing method by this measure, whereas the other three methods are relatively comparable. This indicates that there are advantages to exploiting the two-dimensional nature of these images, rather than simply vectorizing them.

The resulting approximations for three of the facial images are shown in Figure 1(B). Here each method uses approximately 70,000 degrees of freedom, whereas the original data had $112 \times 92 \times 400 = 4,121,600$ total pixel values. The approximations resulting from an SVD of the vectorized data bear little resemblance to the original images. The other three methods are fairly comparable, although one could argue that the Parafac approximations give the best visual impression.

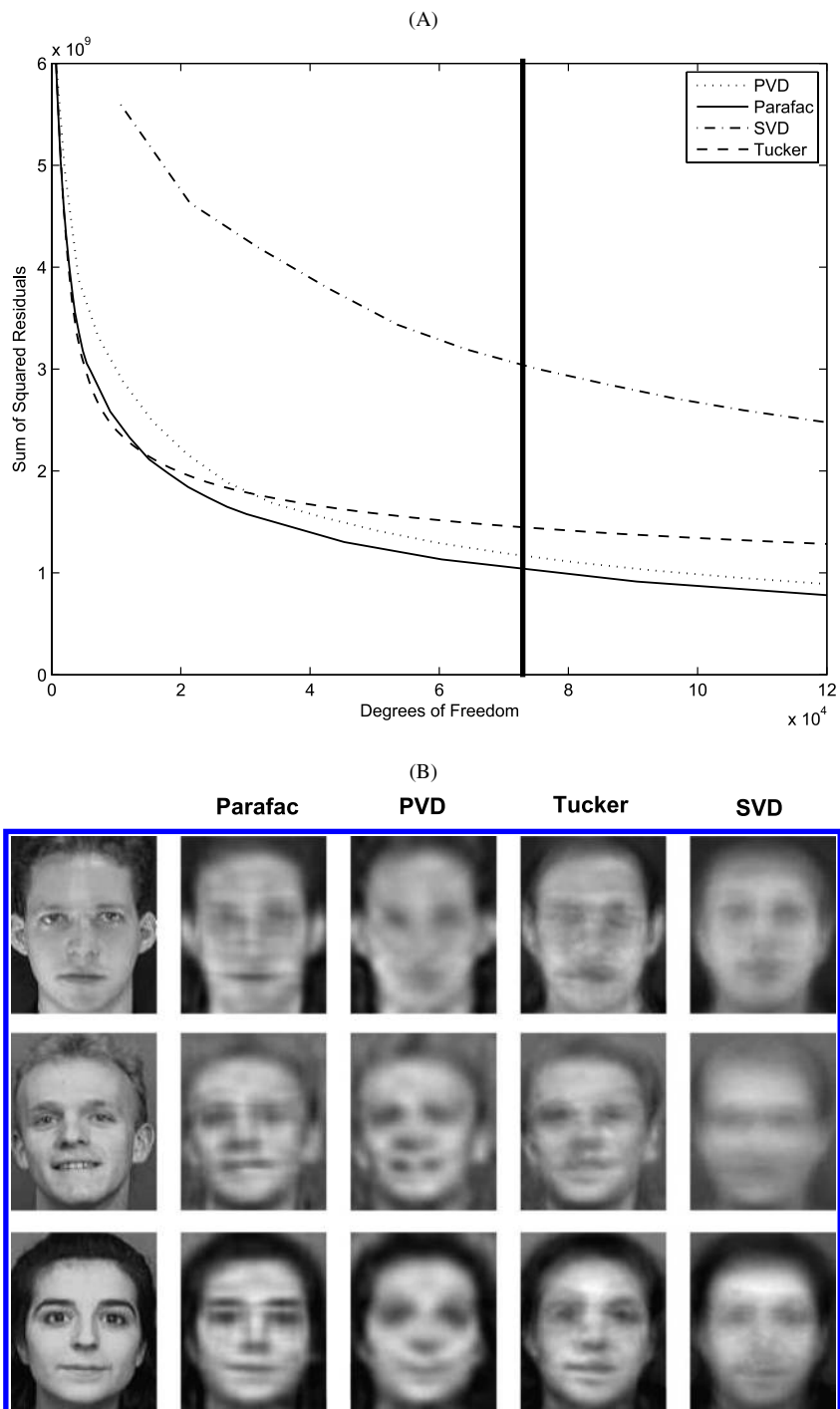


Figure 1. Application of PVD, Tucker, Parafac, and SVD factorizations to facial image data. (A) The sum of squared residuals versus the degrees of freedom used to fit the model for each method. (B) Three facial images (at left), and their reconstructions using the four methods. Each reconstruction uses similar degrees of freedom, close to the vertical line in (A). The Parafac approximation shown uses 72,480 ($r = 120$) degrees of freedom, PVD uses 70,252 ($A = B = 13$), Tucker uses 73,001 ($r_1 = r_2 = r_3 = 37$), and SVD uses 74,928 ($r = 7$).

All of the factorization methods compared here can be used to reduce the dimensionality and provide insight into the primary modes of variation among a collection of matrices. The relative success of these factorization methods will depend on the structure and dimensions of any given dataset. Here we have focused exclusively on data compression. There are other important considerations, such as which method provides the best interpretation for a given application.

ADDITIONAL REFERENCES

Andersson, C. A., and Rasmus, B. (2000), "The N-Way Toolbox for MATLAB," *Chemometrics and Intelligent Laboratory Systems*, 52, 1–4. [800]

Carroll, J. D., and Chang, J. (1970), "Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of "Eckart–Young" Decomposition," *Psychometrika*, 35, 283–391. [798,799]

Ding, C., and Ye, J. (2005), "Two-Dimensional Singular Value Decomposition (2DSVD) for 2D Maps and Images," in *Proceedings of SIAM International*

- Conference on Data Mining (SDM'05)*, Philadelphia, PA: SIAM, pp. 32–43. [798]
- Kolda, T. G., and Bader, B. W. (2009), "Tensor Decompositions and Applications," *SIAM Review*, 51, 455–500. [799]
- Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2011), "Joint and Individual Variation Explained (JIVE) for the Integrated Analysis of Multiple Datatypes," available at [arXiv:1102.4110](https://arxiv.org/abs/1102.4110). [800]
- Shabalin, A. A., and Nobel, A. B. (2010), "Reconstruction of a Low-Rank Matrix in the Presence of Gaussian Noise," available at [arXiv:1007.4148](https://arxiv.org/abs/1007.4148). [800]
- Tucker, L. R. (1966), "Some Mathematical Notes on Three-Mode Factor Analysis," *Psychometrika*, 31, 279–311. [798,799]
- Wold, S. (1978), "Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models," *Technometrics*, 20, 397–405. [800]
- Ye, J. (2005), "Generalized Low Rank Approximations of Matrices," *Machine Learning*, 61, 167–191. [798]
- Zitova, B., and Flusser, J. (2003), "Image Registration Methods: A Survey," *Image and Vision Computing*, 21, 977–1000. [800]

Comment

Ying Nian Wu

The population value decomposition method proposed in this article is an interesting advance in analyzing massive high-dimensional data. I am impressed by the simplicity of the model and the associated computational algorithm. Its application in the Sleep Heart Health Study demonstrates the usefulness of the proposed methodology.

The proposed computational algorithm is based on subject-specific singular value decompositions. Is it possible to find a more rigorous algorithm that minimizes some objective function?

The proposed model assumes the same \mathbf{P} and \mathbf{D} for the whole population. In a population consisting of multiple clusters, it is possible that different clusters may have different \mathbf{P} and \mathbf{D} . Is it possible to extend the model and algorithm to address this issue?

As the authors point out, the proposed method can be considered a multistage principal component analysis (PCA). As such, it shares the limitations of PCA, such as the inability to capture the non-Gaussian and nonlinear properties in the data. Although the proposed method appears to be very sensible for SHHS data, it might not be adequate for other types of image data, such as natural scene images.

As to dimension reduction, it is worthwhile to mention the work of Olshausen and Field (1996) on sparse coding that goes beyond PCA or factor analysis. For PCA, one finds a small number of orthogonal basis vectors that capture most of the variations in the data. In sparse coding, however, one finds a large dictionary of basis vectors that are not necessarily orthogonal to one another, so that each observed signal can be represented by a small number of basis vectors selected from the dictionary, but different signals may be represented by different sets of selected basis vectors.

Specifically, Olshausen and Field (1996) considered the modeling of natural image patches (e.g., 12×12 images, so the signal is 144 dimensional vector). Let $\{\mathbf{I}_m, m = 1, \dots, M\}$ be the set of M image patches represented by the following linear model:

$$\mathbf{I}_m = \sum_{k=1}^K c_{m,k} \mathbf{B}_k + \epsilon_m, \quad (1)$$

where each \mathbf{B}_k is a basis vector of the same dimensionality as \mathbf{I}_m and $c_{m,k}$ is the coefficient. In the language of linear regression, \mathbf{I}_m is the response vector and $(\mathbf{B}_k, k = 1, \dots, K)$ are the regressors or predictors. It is often assumed that the number of regressors K is greater than the dimensionality of the response vector (called the " $p > n$ " problem in regression). Meanwhile, it is also assumed that $(c_{m,k}, k = 1, \dots, K)$ is sparse, in that for each \mathbf{I}_m , only a small number of $c_{m,k}$ are nonzero (or significantly different from 0). Given the dictionary of regressors $(\mathbf{B}_k, k = 1, \dots, K)$, inferring $(c_{m,k}, k = 1, \dots, K)$ is a variable selection problem. But here the twist is that the regressors $(\mathbf{B}_k, k = 1, \dots, K)$ are unknown and are to be learned from the training data $\{\mathbf{I}_m, m = 1, \dots, M\}$. Interestingly, by enforcing sparsity on $(c_{m,k}, k = 1, \dots, K)$, the $(\mathbf{B}_k, k = 1, \dots, K)$ learned from natural image patches are localized, oriented, and elongated wavelets. This provides a statistical justification for the use of wavelets in representing natural images.

The sparsity of $(c_{m,k}, k = 1, \dots, K)$ leads to dimension reduction of \mathbf{I}_m . However, unlike PCA, the dimension reduction in sparse coding is adaptive or subject-specific, because the sets of nonzero $c_{m,k}$ can be different for different m . This is much more flexible than PCA. It is also related to the aforementioned clustering issue, where different clusters may lie in different low-dimensional subspaces.

Recently (Wu et al. 2010), we attempted to model such clusters. In our approach we first assume that the basis vectors are already learned or designed, and so there is a dictionary of localized, oriented, and elongated wavelets $\{\mathbf{B}_{x,s,\alpha}\}$, indexed or attributed by location x , scale s , and orientation α . Each $\mathbf{B}_{x,s,\alpha}$ is like a stroke for sketching the image. We then model each cluster by

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} \mathbf{B}_{x_i, s_i, \alpha_i} + \epsilon_m, \quad (2)$$

where $(\mathbf{B}_{x_i, s_i, \alpha_i}, i = 1, \dots, n)$ is the set of a small number n of basis vectors selected from the dictionary for representing the cluster. $(\mathbf{B}_{x_i, s_i, \alpha_i}, i = 1, \dots, n)$ is like a template with n strokes. We allow small perturbations $(\Delta x_{m,i}, \Delta \alpha_{m,i}, i = 1, \dots, n)$ in locations and orientations, so that the template is

- Conference on Data Mining (SDM'05)*, Philadelphia, PA: SIAM, pp. 32–43. [798]
- Kolda, T. G., and Bader, B. W. (2009), “Tensor Decompositions and Applications,” *SIAM Review*, 51, 455–500. [799]
- Lock, E. F., Hoadley, K. A., Marron, J. S., and Nobel, A. B. (2011), “Joint and Individual Variation Explained (JIVE) for the Integrated Analysis of Multiple Datatypes,” available at [arXiv:1102.4110](https://arxiv.org/abs/1102.4110). [800]
- Shabalin, A. A., and Nobel, A. B. (2010), “Reconstruction of a Low-Rank Matrix in the Presence of Gaussian Noise,” available at [arXiv:1007.4148](https://arxiv.org/abs/1007.4148). [800]
- Tucker, L. R. (1966), “Some Mathematical Notes on Three-Mode Factor Analysis,” *Psychometrika*, 31, 279–311. [798,799]
- Wold, S. (1978), “Cross-Validatory Estimation of the Number of Components in Factor and Principal Components Models,” *Technometrics*, 20, 397–405. [800]
- Ye, J. (2005), “Generalized Low Rank Approximations of Matrices,” *Machine Learning*, 61, 167–191. [798]
- Zitova, B., and Flusser, J. (2003), “Image Registration Methods: A Survey,” *Image and Vision Computing*, 21, 977–1000. [800]

Comment

Ying Nian Wu

The population value decomposition method proposed in this article is an interesting advance in analyzing massive high-dimensional data. I am impressed by the simplicity of the model and the associated computational algorithm. Its application in the Sleep Heart Health Study demonstrates the usefulness of the proposed methodology.

The proposed computational algorithm is based on subject-specific singular value decompositions. Is it possible to find a more rigorous algorithm that minimizes some objective function?

The proposed model assumes the same \mathbf{P} and \mathbf{D} for the whole population. In a population consisting of multiple clusters, it is possible that different clusters may have different \mathbf{P} and \mathbf{D} . Is it possible to extend the model and algorithm to address this issue?

As the authors point out, the proposed method can be considered a multistage principal component analysis (PCA). As such, it shares the limitations of PCA, such as the inability to capture the non-Gaussian and nonlinear properties in the data. Although the proposed method appears to be very sensible for SHHS data, it might not be adequate for other types of image data, such as natural scene images.

As to dimension reduction, it is worthwhile to mention the work of Olshausen and Field (1996) on sparse coding that goes beyond PCA or factor analysis. For PCA, one finds a small number of orthogonal basis vectors that capture most of the variations in the data. In sparse coding, however, one finds a large dictionary of basis vectors that are not necessarily orthogonal to one another, so that each observed signal can be represented by a small number of basis vectors selected from the dictionary, but different signals may be represented by different sets of selected basis vectors.

Specifically, Olshausen and Field (1996) considered the modeling of natural image patches (e.g., 12×12 images, so the signal is 144 dimensional vector). Let $\{\mathbf{I}_m, m = 1, \dots, M\}$ be the set of M image patches represented by the following linear model:

$$\mathbf{I}_m = \sum_{k=1}^K c_{m,k} \mathbf{B}_k + \epsilon_m, \quad (1)$$

where each \mathbf{B}_k is a basis vector of the same dimensionality as \mathbf{I}_m and $c_{m,k}$ is the coefficient. In the language of linear regression, \mathbf{I}_m is the response vector and $(\mathbf{B}_k, k = 1, \dots, K)$ are the regressors or predictors. It is often assumed that the number of regressors K is greater than the dimensionality of the response vector (called the “ $p > n$ ” problem in regression). Meanwhile, it is also assumed that $(c_{m,k}, k = 1, \dots, K)$ is sparse, in that for each \mathbf{I}_m , only a small number of $c_{m,k}$ are nonzero (or significantly different from 0). Given the dictionary of regressors $(\mathbf{B}_k, k = 1, \dots, K)$, inferring $(c_{m,k}, k = 1, \dots, K)$ is a variable selection problem. But here the twist is that the regressors $(\mathbf{B}_k, k = 1, \dots, K)$ are unknown and are to be learned from the training data $\{\mathbf{I}_m, m = 1, \dots, M\}$. Interestingly, by enforcing sparsity on $(c_{m,k}, k = 1, \dots, K)$, the $(\mathbf{B}_k, k = 1, \dots, K)$ learned from natural image patches are localized, oriented, and elongated wavelets. This provides a statistical justification for the use of wavelets in representing natural images.

The sparsity of $(c_{m,k}, k = 1, \dots, K)$ leads to dimension reduction of \mathbf{I}_m . However, unlike PCA, the dimension reduction in sparse coding is adaptive or subject-specific, because the sets of nonzero $c_{m,k}$ can be different for different m . This is much more flexible than PCA. It is also related to the aforementioned clustering issue, where different clusters may lie in different low-dimensional subspaces.

Recently (Wu et al. 2010), we attempted to model such clusters. In our approach we first assume that the basis vectors are already learned or designed, and so there is a dictionary of localized, oriented, and elongated wavelets $\{\mathbf{B}_{x,s,\alpha}\}$, indexed or attributed by location x , scale s , and orientation α . Each $\mathbf{B}_{x,s,\alpha}$ is like a stroke for sketching the image. We then model each cluster by

$$\mathbf{I}_m = \sum_{i=1}^n c_{m,i} \mathbf{B}_{x_i, s_i, \alpha_i} + \epsilon_m, \quad (2)$$

where $(\mathbf{B}_{x_i, s_i, \alpha_i}, i = 1, \dots, n)$ is the set of a small number n of basis vectors selected from the dictionary for representing the cluster. $(\mathbf{B}_{x_i, s_i, \alpha_i}, i = 1, \dots, n)$ is like a template with n strokes. We allow small perturbations $(\Delta x_{m,i}, \Delta \alpha_{m,i}, i = 1, \dots, n)$ in locations and orientations, so that the template is



Figure 1. Templates learned from images of animal faces by model-based clustering. Each template consists of a set of wavelet basis elements, each of which is illustrated by a bar. Number of training images, 320; image height and width, 120×120 pixels; number of clusters, 4; number of selected wavelet elements, 60.

deformable. Different clusters are represented by different templates (B_{x_i, s, α_i} , $i = 1, \dots, n$). We assume that the scale s is fixed.

We have done some preliminary experiments on finding such clusters. Figure 1 displays four templates obtained from 320 images (120×120 pixels) of animal faces by model-based clustering, where in each template (B_{x_i, s, α_i} , $i = 1, \dots, n = 60$), B_{x_i, s, α_i} is illustrated by a bar at the location x_i , scale s , and orientation α_i .

It remains unclear whether or not the clustering experiments could be scaled up to learn thousands of templates or partial templates from image patches of natural scenes or various object categories. The templates of those clusters may become the

“visual words” for representing images, leading to sparser representations of natural images than wavelets.

I would also like to mention the recent work of Hinton and Salakhutdinov (2006) on dimension reduction based on the so-called “auto-encoder” network, which is a multilayer neural network with a low-dimensional central layer for reconstructing the high-dimensional observed signal. The connection weights of this network are pretrained by learning a restricted Boltzmann machine layer by layer. This autoencoder network appears to be able to capture some structures that elude PCA.

The aforementioned dimension reduction methods might not be applicable to the data that the authors deal with. I bring them up mainly to expand the discussion of existing tools for unsupervised learning. I would like to end my discussion by applauding what the authors have achieved in this interesting article.

ADDITIONAL REFERENCES

- Hinton, G. E., and Salakhutdinov, R. R. (2006), “Reducing the Dimensionality of Data With Neural Networks,” *Science*, 313, 504–507. [803]
- Olshausen, B. A., and Field, D. J. (1996), “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images,” *Nature*, 381, 607–609. [802]
- Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010), “Learning Active Basis Model for Object Detection and Recognition,” *International Journal of Computer Vision*, 90, 198–235. [802]

Rejoinder

Ciprian M. CRAINICEANU, Brian S. CAFFO, Sheng LUO, Vadim M. ZIPUNNIKOV, and Naresh M. PUNJABI

We would like to thank our colleagues who have commented on our proposed methodology. We found their comments extremely thoughtful, with many ideas that are both useful and stimulating. The quality of these comments and the amount of time spent writing them goes well beyond what one would expect for a standard discussion. Here we provide our reactions to these comments.

1. RESPONSE TO COMMENTS BY N. LAZAR

1.1 Scaling-Down the Method

We agree with Lazar’s comments about the need for more research to understand the properties of PVD in the context of a small number of subjects. A good approach might be to think about the problem in terms of signal-to-noise ratio and signal complexity (effective size) rather than number of subjects (n)

versus the number of parameters (apparent size) of the problem (p). Consider the following, arguably exotic, example. Assume that we compare a group of diseased subjects with a group of healthy controls using brain imaging. Images are either 0 or 1 for all voxels of each subject-specific image; that is, every subject’s information is perfectly well summarized by any one of its voxels. In this case it would be easy to estimate the one component, even with only 10–12 subjects per group, and most asymptotic results would hold. However, in realistic settings with 10 subjects per group, one could hope to estimate two, maybe three, components for large signal-to-noise ratios. For the specific case of fMRI studies with a small number of subjects, we think that plots of subject-specific versus population eigenvariates and eigenimages would provide valuable insights into the overall variability and signal-to-noise ratio. Comparisons of within-group versus between-group variability of eigenvariates would be especially instructive and relatively straightforward. For eigenimages, plotting would be more difficult, although we are currently working on better visualization tools. In all these settings, honest visualization remains a key component of the analysis, because misleading color plots are still incredibly simple to make.

We also agree that the standard number of subjects in fMRI studies is small. However, like several useful technologies with

Ciprian M. Crainiceanu is Associate Professor (E-mail: ccrainic@jhsph.edu) and Brian S. Caffo is Associate Professor (E-mail: bcaffo@jhsph.edu), Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205. Sheng Luo is Assistant Professor, Division of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, 1200 Herman Pressler Dr, Houston, TX 77030 (E-mail: sheng.t.luo@uth.tmc.edu). Vadim M. Zipunnikov is Post Doctoral Fellow, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: vzipumi@jhsph.edu). Naresh M. Punjabi is Professor, Department of Epidemiology, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: punjabi@jhmi.edu).



Figure 1. Templates learned from images of animal faces by model-based clustering. Each template consists of a set of wavelet basis elements, each of which is illustrated by a bar. Number of training images, 320; image height and width, 120×120 pixels; number of clusters, 4; number of selected wavelet elements, 60.

deformable. Different clusters are represented by different templates (B_{x_i, s, α_i} , $i = 1, \dots, n$). We assume that the scale s is fixed.

We have done some preliminary experiments on finding such clusters. Figure 1 displays four templates obtained from 320 images (120×120 pixels) of animal faces by model-based clustering, where in each template (B_{x_i, s, α_i} , $i = 1, \dots, n = 60$), B_{x_i, s, α_i} is illustrated by a bar at the location x_i , scale s , and orientation α_i .

It remains unclear whether or not the clustering experiments could be scaled up to learn thousands of templates or partial templates from image patches of natural scenes or various object categories. The templates of those clusters may become the

“visual words” for representing images, leading to sparser representations of natural images than wavelets.

I would also like to mention the recent work of Hinton and Salakhutdinov (2006) on dimension reduction based on the so-called “auto-encoder” network, which is a multilayer neural network with a low-dimensional central layer for reconstructing the high-dimensional observed signal. The connection weights of this network are pretrained by learning a restricted Boltzmann machine layer by layer. This autoencoder network appears to be able to capture some structures that elude PCA.

The aforementioned dimension reduction methods might not be applicable to the data that the authors deal with. I bring them up mainly to expand the discussion of existing tools for unsupervised learning. I would like to end my discussion by applauding what the authors have achieved in this interesting article.

ADDITIONAL REFERENCES

- Hinton, G. E., and Salakhutdinov, R. R. (2006), “Reducing the Dimensionality of Data With Neural Networks,” *Science*, 313, 504–507. [803]
- Olshausen, B. A., and Field, D. J. (1996), “Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images,” *Nature*, 381, 607–609. [802]
- Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010), “Learning Active Basis Model for Object Detection and Recognition,” *International Journal of Computer Vision*, 90, 198–235. [802]

Rejoinder

Ciprian M. CRAINICEANU, Brian S. CAFFO, Sheng LUO, Vadim M. ZIPUNNIKOV, and Naresh M. PUNJABI

We would like to thank our colleagues who have commented on our proposed methodology. We found their comments extremely thoughtful, with many ideas that are both useful and stimulating. The quality of these comments and the amount of time spent writing them goes well beyond what one would expect for a standard discussion. Here we provide our reactions to these comments.

1. RESPONSE TO COMMENTS BY N. LAZAR

1.1 Scaling-Down the Method

We agree with Lazar’s comments about the need for more research to understand the properties of PVD in the context of a small number of subjects. A good approach might be to think about the problem in terms of signal-to-noise ratio and signal complexity (effective size) rather than number of subjects (n)

versus the number of parameters (apparent size) of the problem (p). Consider the following, arguably exotic, example. Assume that we compare a group of diseased subjects with a group of healthy controls using brain imaging. Images are either 0 or 1 for all voxels of each subject-specific image; that is, every subject’s information is perfectly well summarized by any one of its voxels. In this case it would be easy to estimate the one component, even with only 10–12 subjects per group, and most asymptotic results would hold. However, in realistic settings with 10 subjects per group, one could hope to estimate two, maybe three, components for large signal-to-noise ratios. For the specific case of fMRI studies with a small number of subjects, we think that plots of subject-specific versus population eigenvariates and eigenimages would provide valuable insights into the overall variability and signal-to-noise ratio. Comparisons of within-group versus between-group variability of eigenvariates would be especially instructive and relatively straightforward. For eigenimages, plotting would be more difficult, although we are currently working on better visualization tools. In all these settings, honest visualization remains a key component of the analysis, because misleading color plots are still incredibly simple to make.

We also agree that the standard number of subjects in fMRI studies is small. However, like several useful technologies with

Ciprian M. Crainiceanu is Associate Professor (E-mail: ccrainic@jhsph.edu) and Brian S. Caffo is Associate Professor (E-mail: bcaffo@jhsph.edu), Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205. Sheng Luo is Assistant Professor, Division of Biostatistics, School of Public Health, University of Texas Health Science Center at Houston, 1200 Herman Pressler Dr, Houston, TX 77030 (E-mail: sheng.t.luo@uth.tmc.edu). Vadim M. Zipunnikov is Post Doctoral Fellow, Department of Biostatistics, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: vzipumi@jhsph.edu). Naresh M. Punjabi is Professor, Department of Epidemiology, Johns Hopkins University, 615 N. Wolfe St., Baltimore, MD 21205 (E-mail: punjabi@jhmi.edu).

their roots in basic science and clinical research, fMRI studies are moving into public health and population-based studies. For example, we are involved in studies with roughly 200 subjects scanned on three occasions to study Alzheimer's disease (Caffo et al. 2010). Moreover, the NITRC website <http://www.nitrc.org/> includes more than 1000 scans and basic covariate data, whereas the NITRC ADHD 200 dataset contains more than 700 scans. We are also aware of several large cohort studies planning on, or in the process of, collecting functional imaging data. Thus we explored PVD research with an eye toward usage in fMRI as it becomes increasingly popular in public health research.

1.2 Within- and Between-Group Variability

Our second major comment is related to within- and between-group variability. This very insightful comment by Lazar is backed by a series of simulations that seem to indicate the versatility of PVD. The suggestion of gleaning group differences using a "combined PVD" approach in addition to the "group-level ordinary PVD" is especially useful.

Our own research has been heavily focused on the statistically principled analysis of within- and between-subject variability when the observed data at the subject level are ultra-high-dimensional. Whereas mixed one-way ANOVA models are easy to understand and fit for scalar random variables, careful treatment is required when data are random functions or images. Indeed, for data of the type $Y_{ij}(t)$, Di et al. (2009) proposed multilevel functional principal component analysis (MFPCA) to decompose variability. More precisely, the model for demeaned data is $Y_{ij}(t) = X_i(t) + U_{ij}(t) + \epsilon_{ij}(t)$. The covariance operators of the subject-specific process, $X_i(t)$, and the subject/visit-specific process, $U_{ij}(\cdot)$, are obtained directly from covariance operators of the observed process, $Y_{ij}(\cdot)$, under standard assumptions. However, when the dimension of the observed process, $Y_{ij}(t)$, is high, the covariance operators cannot be stored, calculated, or diagonalized. This problem was solved by Caffo et al. (2012) by using a singular value decomposition (SVD) of the observed data matrix and by carefully identifying the MFPCA model components from the SVD. In this article we propose an alternative decomposition for the case where the functional data is in matrix format. More precisely, if \mathbf{Y}_{ij} is the $F \times T$ -dimensional matrix, then we find a decomposition of the type $\mathbf{Y}_{ij} = \mathbf{P}\mathbf{V}_{ij}\mathbf{D} + \mathbf{E}_{ij}$, where the dimensions of \mathbf{V}_{ij} is small. We propose an MFPCA decomposition of \mathbf{V}_{ij} using ideas of Di et al. (2009) and show that this implies a similar decomposition on the original data.

1.3 Dimensionality

In response to the concerns about the choice of the number of components expressed both by Lazar and by Lock, Nobel, and Marron, we would like to discuss two important points. Assume, for simplicity, that observed data is of the type $Y_i(t)$, $t = 1, \dots, T$, $i = 1, \dots, I$, that K_Y is the covariance operator of the process $Y_i(t)$, and that $\phi_k(t)$, $k \geq 0$, are the eigenfunctions of K_Y . Then $Y_i(t)$ admits the following decomposition:

$$Y_i(t) = \mu(t) + \sum_{k \geq 1} \xi_{ik} \phi_k(t) + \epsilon_i(t),$$

where ξ_{ik} are mutually uncorrelated mean 0, variance λ_k random variables and $\epsilon_i(t)$ is white noise with variance σ_ϵ^2 . Here $\lambda_1 \geq \lambda_2 \geq \dots$ are the eigenvalues of the covariance operator K_Y . In this context, choosing the dimension of the covariance operator can be viewed as a model selection problem. Indeed, in the nested sequence of models

$$M_K : Y_i(t) = \mu(t) + \sum_{k=1}^K \xi_{ik} \phi_k(t) + \epsilon_i(t),$$

testing M_K versus M_{K+1} is equivalent to testing $H_{0,K} : \lambda_{K+1} = 0$ versus $H_{A,K} : \lambda_{K+1} > 0$. This is a test for a zero-variance component that could be addressed using restricted likelihood ratio tests for a zero variance component (Crainiceanu and Ruppert 2004; Greven et al. 2008).

Second, we consider another constructive approach that avoids the problem of choosing the dimension. Consider the case when we have two population-level matrices, \mathbf{P}_1 and \mathbf{D}_1 , and we are unsure whether they are sufficient to capture the observed variability in the data matrices. The residuals $\tilde{\mathbf{Y}}_i = \mathbf{Y}_i - \mathbf{P}_1 \mathbf{V}_{i,1} \mathbf{D}_1$ could then be analyzed to detect residual structure not captured by \mathbf{P}_1 and \mathbf{D}_1 . This may suggest another duo, \mathbf{P}_2 and \mathbf{D}_2 , that can be used to model $\tilde{\mathbf{Y}}_i$. Thus it would be relatively easy to test the null model $\mathbf{Y}_i = \mathbf{P}_1 \mathbf{V}_{i,1} \mathbf{D}_1 + \mathbf{E}_i$ against the alternative model $\mathbf{Y}_i = \mathbf{P}_1 \mathbf{V}_{i,1} \mathbf{D}_1 + \mathbf{P}_2 \mathbf{V}_{i,2} \mathbf{D}_2 + \mathbf{E}_i$. This approach also suggests a generalization of PVD to what could be labeled "additive PVD"

$$\mathbf{Y}_i = \sum_{k=1}^K \mathbf{P}_k \mathbf{V}_{i,k} \mathbf{D}_k + \mathbf{E}_i.$$

Even though this model is more complex, it should be relatively easy to fit using back-fitting if the matrices \mathbf{P}_k and \mathbf{D}_k are known.

2. RESPONSE TO COMMENTS BY K. SHEDDEN

We thank Shedden for raising two very important points. Our first comment centers around the question of when PVD is a good approach and when it might fail. One of the major points is related to the fact that the PVD essentially uses a Kronecker product of left and right population eigenvalues to represent a population of images. Shedden is concerned that such a representation might not be parsimonious, and that too many right and left eigenvectors may be necessary to capture the variability. We agree that capturing, say, a disk in a two-dimensional image using Kronecker products of marginal bases indeed may be wasteful.

However, for the analysis of two-dimensional images in which both axes have the same interpretation, we probably would favor a different approach. Specifically, any two-dimensional image can be unfolded in one long vector. A basic principal component analysis on the population of unfolded vectors should have no problem recovering a disk, or other structures that are difficult to approximate by Kronecker products. In fact, in our fMRI application (Caffo et al. 2010), subject-level data are represented as $V \times T$ matrices, where V is the number of voxels in the brain and T is the length of the time series. Thus the three-dimensional brain image is unfolded into a long vector, and the corresponding eigenimages can be very

complex when folded back in the original three-dimensional image.

Another very important idea discussed by Shedden is the possibility of an alternative model, $\mathbf{Y}_i = \mathbf{P}_i \mathbf{V} \mathbf{D}_i + \mathbf{E}_i$. In this model, the left and right eigenvectors are subject-specific and the matrix \mathbf{V} is population-specific. We find this idea intriguing and worth pursuing, given the proper type of application. We are currently investigating alternative decompositions of the type $\mathbf{Y}_i = \mathbf{P}_i \mathbf{V} \mathbf{D}_i + \mathbf{E}_i$ and $\mathbf{Y}_i = \mathbf{P} \mathbf{V} \mathbf{D}_i + \mathbf{E}_i$, where only the left eigenvectors or the right eigenvectors are population-specific. This would be especially useful in cases when the subject-specific information in the right side of the equation is low-dimensional.

3. RESPONSE TO COMMENTS BY E. F. LOCK, A. B. NOBEL, AND J. S. MARRON

3.1 Generalized Low-Rank Approximations of Matrices (GLRAM)

LNM brought to our attention some recent developments in the computer science literature that are directly related to PVD. Specifically, Ye (2004) introduced generalized low-rank approximations of matrices (GLRAM) as a method of compressing a sample of two-dimensional $F \times T$ images \mathbf{Y}_i , $i = 1, \dots, I$. (Here we use our notation for consistency.) In particular, Ye (2004) was interested in finding an $F \times A$ matrix \mathbf{P} and a $B \times T$ matrix \mathbf{D} such that $\mathbf{P} \mathbf{P}' = \mathbf{I}_A$ and $\mathbf{D} \mathbf{D}' = \mathbf{I}_B$, as well as $A \times B$ -dimensional matrices \mathbf{V}_i^{PD} that minimize $\sum_{i=1}^I \|\mathbf{Y}_i - \mathbf{P} \mathbf{V}_i^{AB} \mathbf{D}\|_2$. An important result is that the optimal \mathbf{V}_i^{AB} are given by $\mathbf{P}' \mathbf{Y}_i \mathbf{D}'$ and the problem is to find the optimal \mathbf{P} and \mathbf{D} . *To the best of our knowledge, there is no closed-form solution to this problem.* Therefore, Ye (2004) suggested an iterative algorithm for calculating the GLRAM. However, the convergence properties of this algorithm are still under investigation, given that the solution might depend on the starting value (Lu, Liu, and An 2006; Liu et al. 2010). Surprisingly, there seems to be little discussion regarding the fact that the GLRAM objective function can be rewritten as $\max_{\mathbf{P}, \mathbf{D}} \text{trace}(\mathbf{P} \mathbf{P}' [\sum_{i=1}^I \mathbf{Y}_i \mathbf{D}' \mathbf{D} \mathbf{Y}_i'])$. Thus if a solution exists, then it could be unique only up to orthogonal rotation of matrix \mathbf{P} columns and matrix \mathbf{D} rows.

It is fair to say that GLRAM is under intense methodological research, with methods for obtaining optimal solutions still under investigation. For example, Liang and Shi (2005) and Liang, Zhang, and Shi (2007) proposed an analytical solution to GLRAM, but Inoue and Urahama (2006) and Hu, Lv, and Zhang (2008) showed, using counterexamples, that the solution is not optimal. In addition, Liu et al. (2010) studied theoretical properties of GLRAM that address and explain several experimental phenomena, and established a close relationship between the GLRAM of images and the SVD of vectorized images. In particular, they showed that the objective functions of the two procedures are similar, but the former imposes additional orthogonal constraints, resulting in greater reconstruction error.

LNM discuss an approximation proposed by Ding and Ye (2005) termed 2DSVD, in which the matrix \mathbf{P} is the matrix of the first A eigenvectors of the row-by-row covariance matrix $\frac{1}{n} \sum_i \mathbf{Y}_i \mathbf{Y}_i'$ and \mathbf{D} determined by the first B singular vectors of the column-by-column covariance matrix $\frac{1}{n} \sum_i \mathbf{Y}_i' \mathbf{Y}_i$. This is a reasonable and fast approach that we also discuss in our article, but it has serious problems with scalability. Indeed, if one of

the dimensions of the image is very large, then the corresponding covariance operator cannot be calculated or diagonalized. After we pointed out this problem, LNM offered a potential solution and suggested calculating \mathbf{P} as the first A left singular vectors of the aggregated matrix $[\mathbf{Y}_1, \dots, \mathbf{Y}_n]$ and \mathbf{D} as the first B right singular vectors of $[\mathbf{Y}_1', \dots, \mathbf{Y}_n']'$.

Although this is an important step forward from the proposed 2DSVD procedure, it still has problems in cases when one of the dimensions is very large. Indeed, the method would require calculation of the SVD of a $F \times (nT)$ -dimensional matrix and a $T \times (nF)$ -dimensional matrix. Assuming that one dimension is large, say F , then calculating the SVD would require that nT be of manageable size. This can hold in many applications, but it is not scalable to cases where n is moderate. For example, in the fMRI example (Caffo et al. 2010), F is approximately 50,000, $T = 500$ and $n = 300$, making $nT = 150,000$. In general, calculating the SVD of such a matrix poses a very serious computational challenge. Our default PVD procedure avoids this problem. The main differences between our PVD approach and GLRAM can be summarized as follows:

1. The PVD model $\mathbf{Y}_i = \mathbf{P} \mathbf{V}_i \mathbf{D} + \mathbf{E}_i$ provides a general statistical framework for handling data stored in two-dimensional matrices. In particular:
 - (a) Matrices \mathbf{P} and \mathbf{D} are not required to be orthonormal and could contain wavelet, Fourier, or spline bases.
 - (b) The entries of \mathbf{E}_i are not required to be iid $N(0, \sigma_\epsilon^2)$. Thus symmetric homoscedastic distributions, such as the t or double exponential, and heteroscedastic distributions also could be allowed.
 - (c) PVD is focused on estimating and modeling \mathbf{V}_i and analyzing its predictive properties with respect to outcomes.
2. The GLRAM is an optimization procedure that minimizes the criterion $\sum_{i=1}^I \|\mathbf{Y}_i - \mathbf{P} \mathbf{V}_i^{AB} \mathbf{D}\|_2$ conditional on orthonormality assumptions on \mathbf{P} and \mathbf{D} .
 - (a) There is currently no known closed-form solution to the GLRAM optimization problem.
 - (b) The 2DSVD approximation provides an approximation to the GLRAM solution that seems to work well in practice. The 2DSVD solution does not scale up well in the case when one of the dimensions is high-dimensional and there are a moderate number of subjects.
 - (c) Default PVD can be viewed as another approximation to the GLRAM problem in the case when the entries of \mathbf{E}_i are assumed to be independent, normal, and homoscedastic.
3. The default PVD procedure produces a set of subject-specific matrices \mathbf{P}_i and \mathbf{D}_i . Here we provide two potential uses of the subject-specific matrices of eigenvectors.
 - (a) Compare \mathbf{P}_i and \mathbf{D}_i with the population-specific eigenvectors \mathbf{P} and \mathbf{D} . This could provide important additional insights or suggest potential modeling problems.
 - (b) Provide useful diagnostic and influence statistics based on $\|\mathbf{P}_i - \mathbf{P}\|_2$ and $\|\mathbf{D}_i - \mathbf{D}\|_2$.

3.2 Three-Way Methods

We thank LNM for the insightful discussions of Candecomp/Parafac (Carroll and Chang 1970) and Tucker (Tucker 1966) decompositions. It is very valuable to bring these powerful decompositions to the attention of the statistical community exactly when the number of studies collecting multidimensional arrays is increasing dramatically. We also found the observation that both of these decompositions can be viewed as PVD very insightful.

3.3 Potential Issues

LNM also discuss several issues that must be carefully considered when using PVD: registration, scaling, dimension compatibility, and the choice of dimension. We agree with these points and can attest to their importance. In particular, there seems to be some empirical evidence that choosing $A = B$ tends to work better in practice. Some theoretical justification and some relevant references have been given by Liu et al. (2010). In addition, we would like to point out our discussion of choosing the dimension of the projection space, as well as the “additive PVD” idea described in our response to Lazar’s comments.

4. RESPONSE TO COMMENTS BY Y. N. WU

Wu’s comments are in the context of clustering problems using natural images, where the application is to predict the animal type from images of animal faces. This is a very interesting and hard problem, where a direct application of the default PVD procedure to the two-dimensional image matrices likely would not be competitive with the methods of Wu et al. (2010).

However, we think that principal component approaches and PVD, applied to transformed image data, could provide a good competitor. An initial idea would be to unfold the animal face images into vectors and apply principal component analysis clustering on these vectors. A second idea would be to construct the subsequent new matrix of data from the original image data. Each row corresponds to a pixel in the image, together with all of its neighboring pixels in a particular fixed-size neighborhood. This will form a matrix of size $P \times H$, where P is the number of pixels in the original image and H is the size of the neighborhood. In the example provided by Wu, $P = 14,400$. If only nearest neighbors are considered, then $H = 8$; if only the nearest neighbors and their neighbors are considered, then $H = 24$; and so on. If \mathbf{Y}_i is the $P \times H$ -dimensional transformation of the original image, we then can apply the PVD procedure and obtain $\mathbf{Y}_i = \mathbf{P}\mathbf{V}_i\mathbf{D} + \mathbf{E}_i$. The image-specific \mathbf{V}_i could be used to train prediction algorithms, at least in principle. Whether this would be competitive remains to be verified, but the PVD algorithm would be easy to implement and extremely fast even on this highly augmented dataset.

Wu also asked whether different, cluster-specific matrices \mathbf{P} and \mathbf{D} could be derived. If clusters are known, then cluster-specific matrices can be easily obtained by applying the PVD algorithm to each individual cluster. However, the method has been developed to represent populations of large data matrices in a fixed basis obtained as the Kronecker product between left and right marginal bases. Thus we would think about clustering with respect to the subject-specific information encoded in the matrix \mathbf{V}_i , where the various entries of the \mathbf{P} and \mathbf{D} matrices could be the various directions of clustering.

ADDITIONAL REFERENCES

- Zipunnikov, V., Caffo, B. S., Davatzikos, C., Schwartz, B., and Crainiceanu, C. M. (2012), “Multilevel Functional Principal Component Analysis for High Dimensional Data,” *Journal of Computational and Graphical Statistics*, to appear. [804]
- Carroll, J. D., and Chang, J. (1970), “Analysis of Individual Differences in Multidimensional Scaling via an N-Way Generalization of “Eckart–Young” Decomposition,” *Psychometrika*, 35, 283–391. [806]
- Crainiceanu, C. M., and Ruppert, D. (2004), “Likelihood Ratio Tests in Linear Mixed Models With One Variance Component,” *Journal of the Royal Statistical Society, Ser. B*, 66 (1), 165–185. [804]
- Ding, C., and Ye, J. (2005), “Two-Dimensional Singular Value Decomposition (2DSVD) for 2D Maps and Images,” in *Proceedings of SIAM International Conference on Data Mining (SDM’05)*, Philadelphia, PA: SIAM, pp. 32–43. [805]
- Greven, S., Crainiceanu, C. M., Küchenhoff, H., and Peters, A. (2008), “Restricted Likelihood Ratio Testing for Zero Variance Components in Linear Mixed Models,” *Journal of Computational and Graphical Statistics*, 17 (4), 870–891. [804]
- Hu, Y., Lv, H., and Zhang, X. (2008), “Comments on an Analytical Algorithm for Generalized Low-Rank Approximations of Matrices,” *Pattern Recognition*, 41, 2133–2135. [805]
- Inoue, K., and Urahama, K. (2006), “Equivalence of Non-Iterative Algorithms for Simultaneous Low Rank Approximations of Matrices,” in *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Washington, DC: IEEE Computer Society, pp. 154–159. [805]
- Liang, Z., and Shi, P. (2005) “An Analytical Algorithm for Generalized Low Rank Approximations of Matrices,” *Pattern Recognition*, 38, 2213–2216. [805]
- Liang, Z., Zhang, D., and Shi, P. (2007), “The Theoretical Analysis of GLRAM and Its Applications,” *Pattern Recognition*, 40, 1032–1041. [805]
- Liu, J., Chen, S., Zhou, Z.-H., and Tan, X. (2010), “Generalized Low Rank Approximations of Matrices Revisited,” *IEEE Transactions on Neural Networks*, 21 (4), 621–632. [805,806]
- Lu, C., Liu, W., and An, S. (2006), “Revisit to the Problem of Generalized Low Rank Approximation of Matrices,” in *ICIC 2006. Lecture Notes in Control and Information Sciences*, Vol. 345, Berlin: Springer-Verlag, pp. 450–460. [805]
- Tucker, L. R. (1966), “Some Mathematical Notes on Three-Mode Factor Analysis,” *Psychometrika*, 31, 279–311. [806]
- Wu, Y. N., Si, Z., Gong, H., and Zhu, S. C. (2010), “Learning Active Basis Model for Object Detection and Recognition,” *International Journal of Computer Vision*, 90, 198–235. [806]
- Ye, J. (2004), “Generalized Low Rank Approximations of Matrices,” in *Proceedings of the Twenty-First International Conference on Machine Learning*, New York: ACM, pp. 887–894. [805]