Statistics and Its Interface Volume 3 (2010) 103–111

A weighted cluster kernel PCA prediction model for multi-subject brain imaging data

Ying Guo

Brain imaging data have shown great promise as a useful predictor for psychiatric conditions, cognitive functions and many other neural-related outcomes. Development of prediction models based on imaging data is challenging due to the high dimensionality of the data, noisy measurements, complex correlation structures among voxels, small sample sizes, and between-subject heterogeneity. Most existing prediction approaches apply a dimension reduction method such as PCA on whole brain images as a preprocessing step. These approaches usually do not take into account the cluster structure among voxels and between-subject differences. We propose a weighted cluster kernel PCA predictive model that addresses the challenges in brain imaging data. We first divide voxels into clusters based on neuroanatomic parcellation or data-driven methods, then extract cluster-specific principal features using kernel PCA and define the prediction model based on the principal features. Finally, we propose a weighted estimation method for the prediction model where each subject is weighted according to the percent of variance explained by the principal features. The proposed method allows assessment of relative importance of various brain regions in prediction; captures nonlinearity in feature space; and helps guard against overfitting for outlying subjects in predictive model building. We evaluate the performance of our method through simulation studies. A real fMRI data example is also used to illustrate the method.

Keywords and phrases: Kernel PCA, prediction, multi-subject data, cluster, functional magnetic resonance imaging (fMRI), weighted estimation.

1. INTRODUCTION

In vivo functional neuroimaging, such as functional magnetic resonance imaging (fMRI), has become an important tool for studying patterns of brain activity that are associated with cognitive functions, mental conditions and behavior. Traditionally, the analysis of neuroimaging data primarily focuses on identifying brain regions significantly associated with an experimental task and characterizing functional connectivity between various regions in cognitive functions. In recent years, there have been a lot of efforts devoted to predicting cognitive state, psychiatric conditions and clinical outcomes based on brain images. Pattern recognition techniques such as neural networks (Hanson et al., 2004), support vector machines (Cox and Savoy, 2003) have been used to predict what kind of cognitive state the brain is in given an acquired image.

There are several major challenges in building predictive models using brain imaging data. First, there are enormous amount of voxels in each brain image while the number of observed images is much smaller. This means predictive model based on voxel-specific measures is infeasible since the number of predictors mostly far exceeds the number of samples. Secondly, there exists correlation between the measurements across voxels in the brain. The connectivity among voxels are due to underlying structural connection or the coordinated functions. Thirdly, there is often considerable between-subject heterogeneity in brain images. Subjects may have different noise levels or brain activity patterns in their images. Most existing predictive methods first perform a PCA on the whole image, and then develop predictive models based on the extracted principal components (O'Toole et al. 2005; Mourão-Miranda et al., 2005; Kjems et al., 2002; Carlson et al., 2003; Strother et al., 2004). Limitations of these approaches include: underlying cluster structure in brain images is not taken into account resulting in less efficient summary measures of the activity in various brain regions; ignoring between-subject heterogeneity in predictive model building can cause overfitting and sensitivity to outliers.

Instead of performing the dimension reduction on the whole image, we propose to first divide voxels into clusters and then extract the principal features within each cluster. There are several advantages of taking into account the cluster structure: the brain is a very complex biological systems with enormous amount of neurons. Various brain regions are associated with different types of anatomical structures and neurological functions. Clustering brain voxels allows us to divide voxels into groups where within-cluster voxels share similar functions and activity patterns. Consequently, the principal features extracted within clusters have a better representation of the activity patterns in various regions than the principal features extracted based on all voxels. Secondly, extracting cluster-specific features also provides the possibility to consider between-cluster associations in the prediction model. It has been shown that functional connectivity between different brain regions holds potentials in

predicting neurological and clinical outcomes (Derado et al., 2009). Thirdly, extracting features within clusters also helps with computation by only considering a subset of voxels at a time, whereas the standard all-voxel approach requires handling the complete data matrix and hence has much higher memory demand.

To extract features within a cluster, we propose to use kernel PCA (Schölkopf et al., 1998). Kernel PCA is an extension of traditional PCA for nonlinear distributions using kernel methods. Instead of performing PCA directly on the original imaging space, the images are mapped into a higher-dimensional feature space in which principal components are extracted. Since PCA can be formulated in terms of the dot products of input data, kernel PCA can be performed based on the kernel matrix without the need of really performing the mapping into the feature space. The main advantages of the kernel PCA are that: 1) it reduces the computational complexity since the dimension of the kernel matrix is determined by the number of observed images rather than the number of voxels and the former is often much smaller than the latter for brain imaging data; 2) it can take into account nonlinear patterns in images which has more potential to improve the predictive power.

Prediction models are usually developed based on multi-subject imaging data. It has been found that there may be wide variability across subject's in terms of the signal-to-noise level and neural response patterns. Most existing approaches ignore between-subject differences and each subjects' data contribute equally in estimating the parameters of the predictive model. However, in the presence of outliers when certain subject's images have much higher noise level, the prediction model may be stretched to provide a better fit for the outlying subjects. We propose to include subject-specific weights in building a prediction model to make the model more robust to extreme observations from subjects whose imaging quality is questionable. Since our predictors are extracted principal features, we propose to define the subject weight parameter based on the percent of variance explained by the principal features. This approach will give more weight to subjects whose images are better represented by the extracted principal features and downweight subjects who have large amount of residual variability in their images that are not captured by the principal features.

2. METHOD

The proposed method is applicable to various types of brain imaging data including functional magnetic resonance imaging (fMRI) and positron emission tomography (PET). In this paper, we present our method for fMRI data since they are most commonly seen in imaging studies. Let $\boldsymbol{x}_{it} = [x_{it1}, \ldots, x_{itV}]'$ be an fMRI image scanned for subject i at time t where $i = 1, \ldots, N$, $t = 1, \ldots, T$ and V is the total number of voxels in the image. Let $\boldsymbol{X}_i = [\boldsymbol{x}'_{i1}, \ldots, \boldsymbol{x}'_{iT}]'$ be the $T \times V$ data matrix of subject i. Let $\boldsymbol{y}_i$ be $T \times 1$ vector of outcome of interest associated with the brain scans from subject i. For example, $\boldsymbol{y}_i$ can be the cognitive states the subject is experiencing during the imaging session. Our objective is to predict $\boldsymbol{y}_i$ based on the observed images $\boldsymbol{X}_i$.

2.1 Clustering of voxels in brain images

In vivo functional neuroimaging studies have shown that different parts of the cerebral cortex are involved in different cognitive and behavioral functions. Previous work has proposed ways for parceling the brain into distinct neuroanatomic regions such as Brodmann area maps, the Freesurfer software (Fischl et al., 2004, 2002) and the macroscopic anatomical parcellation (Tzourio-Mazoyer et al., 2002). As an alternative to the approaches based on neuroanatomy, one may also use data-driven procedures such as functional clustering (Bowman and Patel, 2004a; Bowman et al., 2004b; Cordes et al., 2002) and spatial independent component analysis (Calhoun et al., 2001; Beckmann and Smith, 2005; Guo and Pagnoni, 2008). In specific, functional clustering approaches group voxels into clusters that exhibit similar patterns of observed brain activity. They define a dissimilarity metric for a pair of voxels based on the distance between voxel-specific vectors of summary statistics of brain activity. A cluster algorithm such as K-Means, fuzzy clustering or hierarchical clustering is then applied to divide voxels into clusters where within-cluster voxels share a similar pattern of activity while between-cluster voxels tend to have different patterns. Spatial independent component analysis is another data-driven procedure that decomposes the observed imaging data into a linear combination of spatio-temporal processes of functional networks with independently distributed spatial patterns. Voxels can be distributed into the identified functional networks based on their loadings on the independent components and voxels within the same functional network share similar temporal responses. Data-driven procedures define functional regions that include both adjacent and disjoint anatomical locations that share similar patterns of brain activity. Clusters based on data-driven approaches generally demonstrate more consistent within-cluster patterns than those based on neuroanatomy. The disadvantage of the data-driven approaches is that clusters were derived from a specific data set and hence are more variable across studies than clusters based on neuroanatomical parcellations.

2.2 Kernel principal component analysis (KPCA)

A well-known characteristic of neuroimaging data is its high dimensionality and enormous size. For example, a twenty-minute fMRI session with a single human subject can produce 300 to 400 three dimensional brain images each containing approximately hundreds of thousands of voxels, yielding tens of millions of data observations. Therefore, to develop predictive models based on imaging data, dimension

reduction and feature selection is often needed. In this paper, we propose to use kernel principal component analysis (KPCA) (Schölkopf et al., 1998) to extract principal patterns of brain activity in observed images. In this section, we first give a brief introduction of the KPCA.

Suppose $\boldsymbol{x}_l$ $(l = 1, \ldots, L)$ are the observed data vectors, e.g. $\boldsymbol{x}_l \in \mathbf{R}^V$ is the lth observed brain image. Principal Component Analysis (PCA) is a basis transformation to diagonalize an estimate of the covariance matrix of the original images, defined as

$$C = \frac{1}{L}\sum_{l=1}^{L} \boldsymbol{x}_l \boldsymbol{x}_l', \tag{1}$$

here, we assume the observed data vectors are centered. Eigenvectors of the covariance matrix define the basis in the transformed space and the coordinates in the eigenvectors are known as principal components.

KPCA is an extension of PCA using kernel methods. In KPCA, we first map the images nonlinearly into a high dimensional feature space F by

$$\phi : \boldsymbol{R}^V \to F, \qquad \boldsymbol{x} \mapsto \phi(\boldsymbol{x}). \tag{2}$$

For example, a second-order nonlinear mapping function ϕ maps an original image $\boldsymbol{x}$ into a feature space F which is spanned by all pairwise products between voxels of $\boldsymbol{x}$. The estimate of the covariance matrix in feature space F is then,

$$C = \frac{1}{L}\sum_{l=1}^{L} \phi(\boldsymbol{x}_l)\phi(\boldsymbol{x}_l)'. \tag{3}$$

Schölkopf et al. (1998) have shown that even if F has arbitrarily large dimensionality, for certain choices of ϕ, we can still perform PCA in feature space F by the use of kernel functions (Boser et al., 1992). In specific, they define $L \times L$ kernel matrix $\boldsymbol{K}$ by

$$\boldsymbol{K}_{l_1 l_2} = k(\boldsymbol{x}_{l_1}, \boldsymbol{x}_{l_2}), \quad l_j = 1, \ldots, L,\ j = 1, 2 \tag{4}$$

where $k(\boldsymbol{x}_{l_1}, \boldsymbol{x}_{l_2}) = \langle \phi(\boldsymbol{x}_{l_1}), \phi(\boldsymbol{x}_{l_2}) \rangle$ is the kernel function representing the dot product between two images in feature space F. Schölkopf et al. (1998) have shown that to perform PCA on the covariance matrix (3) in the feature space F is equivalent to diagonalize the kernel matrix (4). Here, we assume $\phi(\boldsymbol{x})$ is centered because Taylor and Cristianini (2004) have shown that the kernel matrix of the centered feature data can be obtained through some simple algebra operation on the original kernel matrix.

The kernel function k in (4) essentially measures the similarity between two images in the feature space. The choice of the kernel k corresponds to a specific mapping function ϕ. The most simple kernel function is the linear kernel $k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \langle \boldsymbol{x}_1, \boldsymbol{x}_2 \rangle$ which is the dot product between two images and measures the similarity between images in its original space. When a linear kernel is used, KPCA is equivalent to PCA. Nonlinear kernel functions measure the similarity between images after they are nonlinearly mapped into a higher-dimensional feature space. For example, polynomial kernel functions of degree d correspond to a map ϕ into a feature space which is spanned by all products of d entries of an input vector $\boldsymbol{x}$. One commonly chosen nonlinear kernel function is Gaussian kernel function,

$$k(\boldsymbol{x}_1, \boldsymbol{x}_2) = \exp(-||\boldsymbol{x}_1 - \boldsymbol{x}_2||^2/2\sigma^2), \tag{5}$$

where σ^2 is known as the bandwidth in Gaussian kernel and is usually chosen through cross-validation (Taylor and Cristianini, 2004). An advantage of Gaussian kernel is that it can be directly calculated from the linear kernel and thus is computationally efficient. A key advantage of the kernel methods is that evaluation of the kernel function requires much less computation than that of the explicit evaluation of the corresponding feature mapping ϕ.

2.3 Cluster KPCA on multi-subject imaging data

In this section, we propose a cluster KPCA approach to extract features in multi-subject imaging studies. These features will then be used for predicting our outcome of interest. Suppose we have grouped voxels in brain images into R clusters based on either neuroanatomical parcellation or a data-driven approach. The collection of voxels in cluster r $(r = 1, \ldots, R)$ is denoted as Ω^r with $|\Omega^r| = V^r$ and $\sum_{r=1}^{R} V^r = V$. Let $\boldsymbol{x}_{it}^r$ be the $V^r \times 1$ subimage of the rth cluster for subject i at time t. For multi-subject fMRI data, we first define a kernel matrix $\boldsymbol{K}_{ij}^r$ for cluster r for each subject pair (i, j) with $i, j \in [1, \ldots, N]$. Here, $\boldsymbol{K}_{ij}^r$ is a $T \times T$ kernel matrix where the (t_1, t_2)th element $(t_1, t_2 \in [1, \ldots, T])$ is defined as

$$\boldsymbol{K}_{ij,t_1 t_2}^r = k(\boldsymbol{x}_{it_1}^r, \boldsymbol{x}_{jt_2}^r), \tag{6}$$

where $k(\boldsymbol{x}_{it_1}^r, \boldsymbol{x}_{jt_2}^r) = \langle \phi(\boldsymbol{x}_{it_1}^r), \phi(\boldsymbol{x}_{jt_2}^r) \rangle$. Therefore, $\boldsymbol{K}_{ij}^r$ is the kernel matrix measuring similarity between images acquired from subject i and subject j. Then we define a multi-subject kernel matrix $\boldsymbol{K}^r$ for cluster r. Here, $\boldsymbol{K}^r$ is a $NT \times NT$ matrix composed of N^2 submatrices with the (i, j)th submatrix being $\boldsymbol{K}_{ij}^r$.

By diagonalizing the multi-subject kernel matrix $\boldsymbol{K}^r$, we can retrieve the eigenvectors of the sample covariance matrix of the feature data,

$$\boldsymbol{C}^r = \frac{1}{TN}\sum_{i=1}^{N}\sum_{t=1}^{T} \phi(\boldsymbol{x}_{it}^r)\phi(\boldsymbol{x}_{it}^r)'. \tag{7}$$

To show this result, let $\tilde{\boldsymbol{X}}^r$ be the feature data matrix where the rows contain the features $\phi(\boldsymbol{x}_{it}^r)$ and let $\ell = rank(\tilde{\boldsymbol{X}}^r)$. We can show $\boldsymbol{C} = \tilde{\boldsymbol{X}}^{r\prime}\tilde{\boldsymbol{X}}^r/TN$ and $\boldsymbol{K}^r = \tilde{\boldsymbol{X}}^r\tilde{\boldsymbol{X}}^{r\prime}$. Denote singular value decomposition of $\tilde{\boldsymbol{X}}^r = \boldsymbol{U}^r\boldsymbol{D}^r\boldsymbol{Q}^{r\prime}$, where $\boldsymbol{U}^r$

and $\boldsymbol{Q}^r$ are orthogonal matrices whose columns contain the left and right singular vectors, respectively and $\boldsymbol{D}^r$ is a diagonal matrix with diagonal elements $d_1^r \geq d_2^r \geq \cdots d_\ell^r \geq 0$ known as singular values. One can show that $\boldsymbol{Q}^r$ and $\boldsymbol{U}^r$ are the eigenvector matrices for $NT\boldsymbol{C}^r$ and $\boldsymbol{K}^r$, respectively with the vectors $\boldsymbol{q}_j^r$ and $\boldsymbol{u}_j^r$ known as the jth principal component direction and principal component in the feature space. Following Taylor and Cristianini (2004), we show that eigenvectors of $NT\boldsymbol{C}^r$ can be calculated from the eigenvectors of $\boldsymbol{K}^r$ as follows,

$$(8) \quad \boldsymbol{q}_j^r = d_j^{r^{-1}} \tilde{\boldsymbol{X}}^{r\prime} \boldsymbol{u}_j^r = \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{it}^{j,r} \phi(\boldsymbol{x}_{it}^r), \quad j = 1, \ldots, \ell,$$

where $\alpha_{it}^{j,r}$ is the element in the vector $\boldsymbol{\alpha}^{j,r} = d_j^{r^{-1}} \boldsymbol{u}_j^r$ that corresponds to the tth time point of subject i. Therefore, PCA of multi-subjects' subimages that are mapped to the feature space F can be performed by diagonalizing the proposed multi-subject kernel matrix $\boldsymbol{K}^r$. Equation (8) also shows that the principal component direction $\boldsymbol{q}_j^r$ lies in the space spanned by the sample feature data $\phi(\boldsymbol{x}_{it}^r), (i = 1, \ldots, N,\ t = 1, \ldots, T)$.

There are several advantages for performing PCA on the kernel matrix: 1) the dimension of the kernel matrix is the number of observed images while the dimension of the covariance matrix is equal to the number of voxels. The former is often much smaller than the latter in brain imaging studies. Therefore, KPCA can help reduce the computational complexity; 2) by using nonlinear kernels, KPCA can accommodate nonlinear features in images and hence has the potential of capturing more complicated patterns.

We propose the following regression model for predicting outcome y based on the first principal features $u_1^r\,(r = 1, \ldots, R)$ extracted through the cluster KPCA approach,

$$(9) \qquad y = \beta_0 + \sum_{r=1}^{R} \beta_r u_1^r.$$

To predict the outcome y for a new image $\boldsymbol{x}^*$, we first group the voxels into clusters $\boldsymbol{x}^{r*}$ and then calculate the cluster-specific principal features u_1^{r*} by projecting $\boldsymbol{x}^{r*}$ onto the first principal component direction defined by (8), i.e.,

$$(10) \quad u_1^{r*} = \boldsymbol{q}_1^{r\prime} \phi(\boldsymbol{x}^{r*}) = \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{it}^{1,r} \langle \phi(\boldsymbol{x}_{it}^r), \phi(\boldsymbol{x}^{r*}) \rangle$$
$$= \sum_{i=1}^{N} \sum_{t=1}^{T} \alpha_{it}^{1,r} k(\boldsymbol{x}_{it}^r, \boldsymbol{x}^{r*}).$$

The predicted outcome y^* is then obtained through the regression model (9) by plugging in u_1^{r*}. An important observation from (10) is that the projection of a new image onto the principal component directions of the feature space can be performed by evaluating the dot products between the new image and the sample images in the feature space. Note that the KPCA and subsequent projections can be performed based on the kernel matrix without actually mapping images onto the high-dimensional feature space through ϕ.

2.4 Estimation of the prediction model using subject-specific weights

Brain imaging data has often shown wide between-subject differences in terms of the signal-to-noise levels and activity patterns. Certain subjects' images may include non-task related variations. To obtain more robust estimates of the parameters in our prediction model, we propose to use the following weighted estimation,

$$(11) \qquad \hat{\boldsymbol{\beta}} = (\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Z})^{-1}\boldsymbol{Z}'\boldsymbol{W}\boldsymbol{Y}$$

where $\boldsymbol{Y} = [\boldsymbol{y}_1', \ldots, \boldsymbol{y}_N']'$ is the $TN \times 1$ response vector; $\boldsymbol{Z} = [\boldsymbol{Z}_1', \ldots, \boldsymbol{Z}_N']'$ where $\boldsymbol{Z}_i = [\boldsymbol{1}\, \boldsymbol{u}_{1,i}^1 \ldots \boldsymbol{u}_{1,i}^R]$, with $\boldsymbol{u}_{1,i}^r$ being $T \times 1$ vectors, includes the first principal features for subject i; $\boldsymbol{W} = \mathrm{diag}(w_1 \boldsymbol{I}_T, \ldots, w_N \boldsymbol{I}_T)$ where w_i is the subject-specific weight. We propose the following weight function,

$$(12) \qquad w_i = \frac{1}{R} \sum_{r=1}^{R} \frac{\lambda_{i1}^r}{\sum_{j=1}^{\ell} \lambda_{ij}^r},$$

where λ_{ij}^r is the jth eigenvalue from the KPCA of the rth region of the ith subject. The proposed weight parameter reflects the average percent of variance explained by the first principal features across clusters in subject i. Therefore, higher weight parameter indicates the subject's observed images are better captured by the extracted first principal features with less unexplained residual variability.

3. SIMULATION STUDIES

We performed simulation studies to evaluate the performance of the proposed method. For comparison, we considered the standard PCA regression prediction model. The simulation setup such as the noise levels were selected based on observed data from a multi-subject fMRI study of Zen Meditation. We conducted two sets of simulation studies. In each study, we generated a training data set of 10 subjects and a test data set of 5 subjects. For each subject, we generated fMRI BOLD signals obtained at 200 time points for 15,000 voxels. Among the 15,000 voxels, 12,000 voxels belong to four equal-size clusters each with cluster-specific temporal dynamics. In the first set of simulation study, the time series of the voxels in the clusters was generated as follows for each subject,

$$(13) \qquad \boldsymbol{g}_{iv}^r = \boldsymbol{g}_i^r + \boldsymbol{\epsilon}_{iv},\ r = 1, 2, 3, 4,$$

where $\boldsymbol{g}_{iv}^r$ is the ith subject's observed temporal dynamics at voxel v in region r and $\boldsymbol{g}_i^r$ is the ith subject's temporal dynamics associated with the jth cluster. We generated

Table 1. Results of simulation studies based on 200 simulation runs. Values presented are PMSE and its standard deviation in the parenthesis

	Cluster KPCA regression		PCA regression	
	weighted	unweighted	weighted	unweighted
without outliers	0.875(0.293)	0.875(0.293)	0.874(0.292)	0.874(0.292)
with outliers	1.071(0.376)	1.926(0.486)	1.072(0.375)	1.922(0.481)

$\boldsymbol{g}_i^r$ by linearly combining population cluster time courses with subject-specific variability that was simulated through a zero-mean Gaussian distribution with a standard deviation of 0.1. The population cluster time courses were selected from temporal responses of estimated independent components based on a group independent component analysis (ICA) of the fMRI study of Zen meditation (Guo and Pagnoni, 2008). In (13), $\boldsymbol{\epsilon}_{iv}$ is voxel-specific fluctuations around the cluster temporal dynamics and was generated from a Gaussian distribution with mean zero and a standard deviation of 0.23. In addition to the 12,000 voxels within the clusters, we also generated time series for 3,000 voxels that don't have consistent temporal dynamics. The temporal dynamics of these voxels were randomly selected from observed voxel time courses in the Zen Meditation study.

The outcome to predict was generated as

$$\boldsymbol{y}_i = \beta_{i1}\boldsymbol{g}_i^1 + \beta_{i2}\boldsymbol{g}_i^2 + \beta_{i3}\boldsymbol{g}_i^3 + \boldsymbol{e}_i, \tag{14}$$

where $\beta_{ir}\,(r = 1, 2, 3)$ are subject-specific parameters for region r and were simulated from $\beta_{ir} \sim N(\theta_r, 0.25)$, where $\boldsymbol{\theta} = [\theta_1, \theta_2, \theta_3]' = [1, 0.5, 2]'$ is the population parameter; $\boldsymbol{e}_i$ were zero-mean Gaussian random errors with a standard deviation of 0.25.

In the second simulation study, the 10 training subjects include 7 normal subjects and 3 outliers. The data for the normal subjects were generated in the same way as that of the first simulation study. For the outliers, the subject-specific regional temporal dynamics were contaminated with additional structured variabilities. More specifically, for each region, each subject's temporal response was generated from a new population cluster time course which is a combination of the original population cluster time course and an additional time series selected from the Zen Meditation fMRI study. We then added subject-specific variability simulated in the same way as that of the first simulation to generate $\boldsymbol{g}_i^r$ for the outliers. Each outlier's voxel-specific time series was generated according to (13) with $\boldsymbol{\epsilon}_{iv}$ simulated from a Gaussian distribution with a standard deviation of 1. Therefore, the images of the outliers are noisier than those of the normal subjects. The outcomes to predict for the outliers were generated in the same way as the normal subjects according to (14).

We applied various methods to predict the outcome variable $\boldsymbol{y}_i$ based on the $200 \times 15,000$ fMRI data matrix $\boldsymbol{X}_i$. We fit the predictive model using data from the 10 training subjects and then evaluated the performance by applying the model to the 5 test subjects. The prediction mean square error (PMSE) was averaged over 200 simulation runs. We compared four approaches: the proposed weighted cluster KPCA regression model, the cluster KPCA model fitted without using the subject-specific weights, the standard PCA regression where temporal principal components were extracted from the whole image for predicting the outcome, and a modified PCA regression approach which incorporates subject-specific weights. For the two cluster KPCA approaches, the cluster structures in the images were determined through K-means clustering method where the number of clusters was determined by pseudo-T^2 criterion (Bowman et al., 2004b). The first principal features extracted in each cluster were used as predictors in the regression model. Linear kernel was used in the two cluster KPCA approaches so that we can compare them to the two corresponding PCA approaches for evaluating the effects of clustering and applying subject-specific weights in prediction. For the two PCA regressions, the number of principal components were matched with the number of clusters so that the PCA regression models had the same number of predictors as the cluster KPCA regression models. We present the results of the three methods in Table 1.

From Table 1, we can see that the two weighted approaches gave the same results as their unweighted counterparts when there were no outliers. In the presence of outliers, the weighted approaches had a smaller prediction error than the unweighted methods since the outliers in the training set were weighted lower than the normal subjects when estimating the prediction model. The cluster KPCA approaches had a similar prediction error with the PCA regression model. However, when we compare the extracted principal features from both models to the true temporal dynamics in the brain regions (Figure 1), one can see that the features extracted by the cluster KPCA accurately captures the time courses of the various brain functional regions where the correlations between the true and estimated time courses were 0.938, 0.942, and 0.940 for the three relevant clusters. In comparison, the principal components of the standard PCA represented mixed signals from various regions where the correlations between the true cluster time courses and the best matched principal components were 0.687, 0.778, and 0.685 for the three relevant clusters. Consequently, it is hard to associate the extracted principal components from the standard PCA approach with specific brain regions. Therefore, though the PCA regression model may have comparable prediction accuracy as the proposed approach, it does not allow us to estimate the relative predictive power of

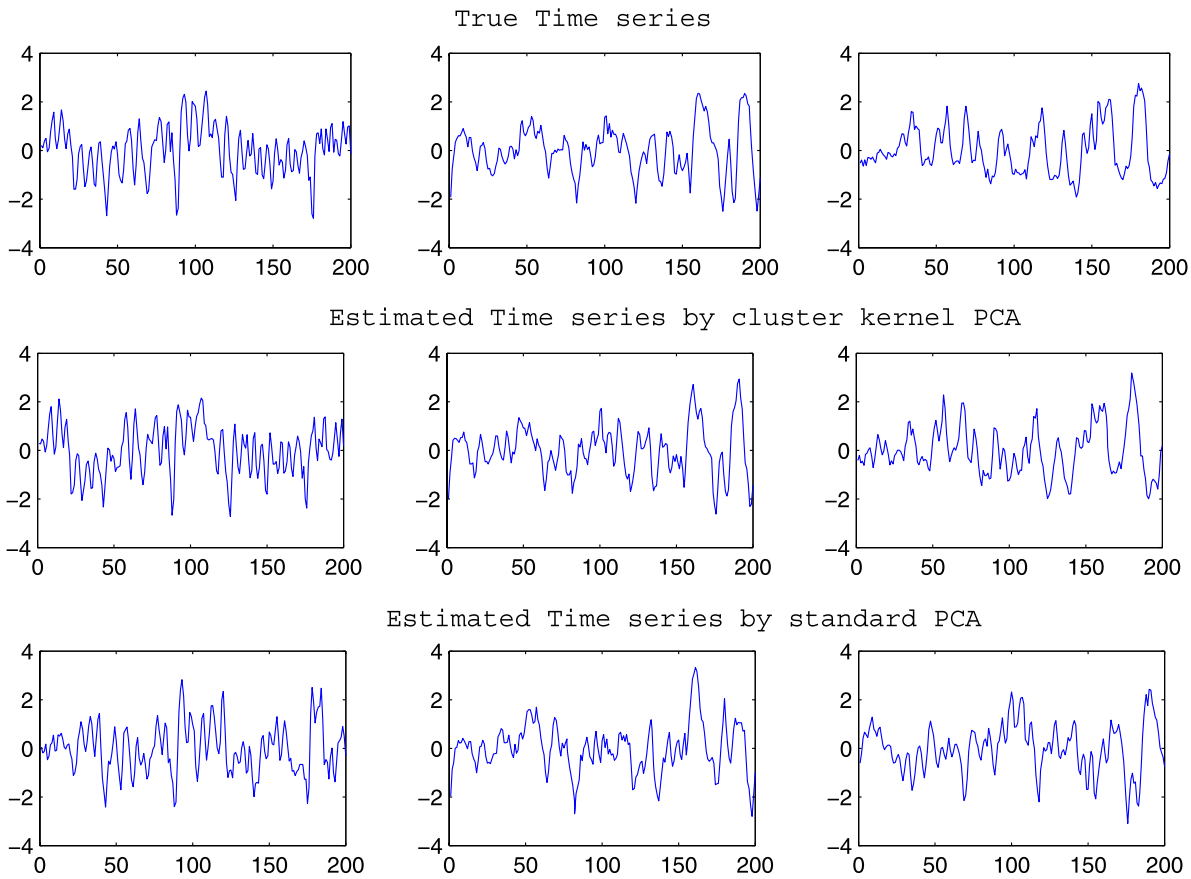

Figure 1. Comparison of the true cluster time series with the extracted principal features based on cluster KPCA and standard PCA.

various brain regions for the outcome. The proposed cluster KPCA approach also offers computational advantages over PCA regression models. Since cluster KPCA extracts features within each cluster separately, it only needs to deal with a subset of voxels rather than the whole brain. Additionally, diagonalizing the kernel matrix is computationally more efficient than diagonalizing the covariance matrix. Therefore, the cluster KPCA approach has lower memory demands and is faster than the standard PCA regression.

4. AN FMRI DATA EXAMPLE

We illustrate the proposed method using data from an fMRI study of Zen meditation. In this study, twelve subjects who have practiced Zen meditation for years were recruited. During the fMRI scanning session, word and phonologically and orthographically matched nonword items were presented visually on a screen in a pseudo-random temporal order. And subjects were asked to respond whether each displayed item was "a real English word" via pressing a buttonbox using left hand fingers. During the intervals between the stimuli, subjects were instructed to stay in a resting state by using the awareness of their breathing throughout the session as a reference point. A series of 520 functional MRI images were acquired during the experiment with a 3.0 Tesla Siemens Magnetom Trio scanner. Each of these fMRI images contains $53 \times 63 \times 46$ voxels. The images were corrected for slice acquisition timing differences and subject movement, registered and spatially normalized to the MNI standard brain space.

We applied the proposed method to predict the visual semantic and nonsemantic stimuli level that a subject was exposed to by using the observed fMRI images during the session. The stimuli level was calculated as the task time series convolved with the HRF function. In previous analysis (Guo and Pagnoni, 2008), we identified 14 functional networks in the Zen meditation data based on the group ICA. Based on the ICA results, we defined 14 clusters by assigning voxels to the independent component on which they had the highest loading. We then applied the cluster KPCA to the 14 clusters. In each cluster, we extracted the first principal feature for predicting the stimuli level. Single subject KPCA was performed and the proportion of variance explained by the first principal components in each of the 14 regions was obtained to determine subject weights. We then constructed a prediction model for the stimuli level using the extracted first principal features from the 14 regions.

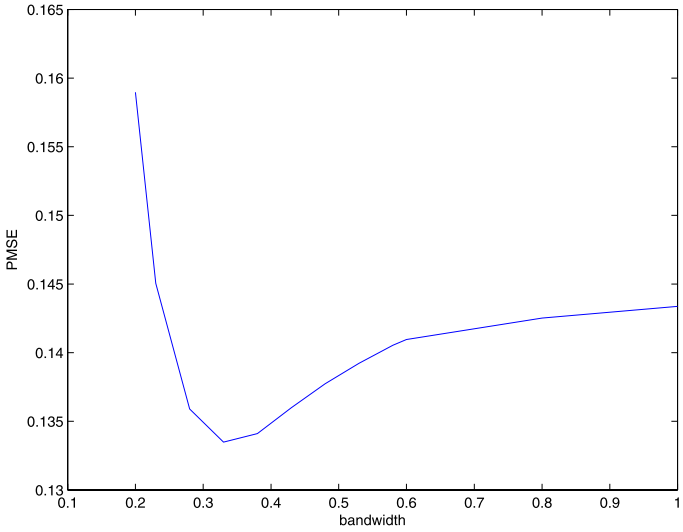

Figure 2. Choosing bandwidth for the Gaussian kernel using cross-validation in meditation data.

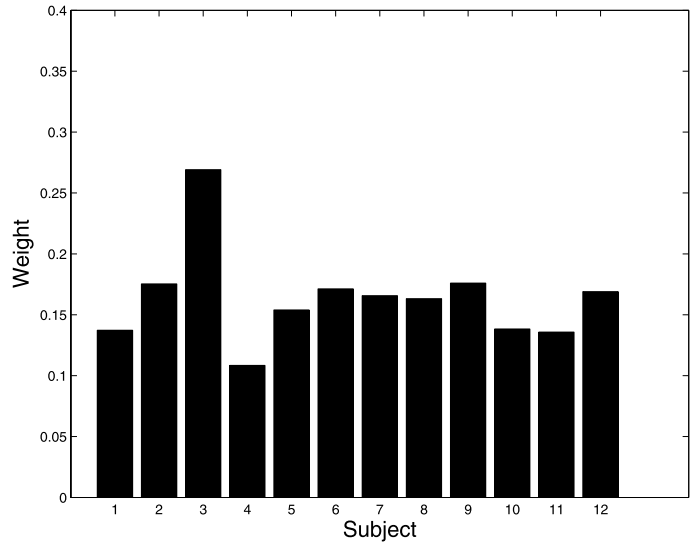

Figure 3. Subject weights for cluster KPCA regression model in meditation data.

Table 2. Comparison of results based on various prediction models for the meditation study

Method	PMSE
Cluster KPCA (linear kernel) without subject weight	0.145
Cluster KPCA (linear kernel) with subject weight	0.143
Cluster KPCA (Gaussian kernel) without subject weight	0.131
Cluster KPCA (Gaussian kernel) with subject weight	0.130
PCA regression (with the first PC)	0.164
PCA regression (with 14 PCs)	0.142

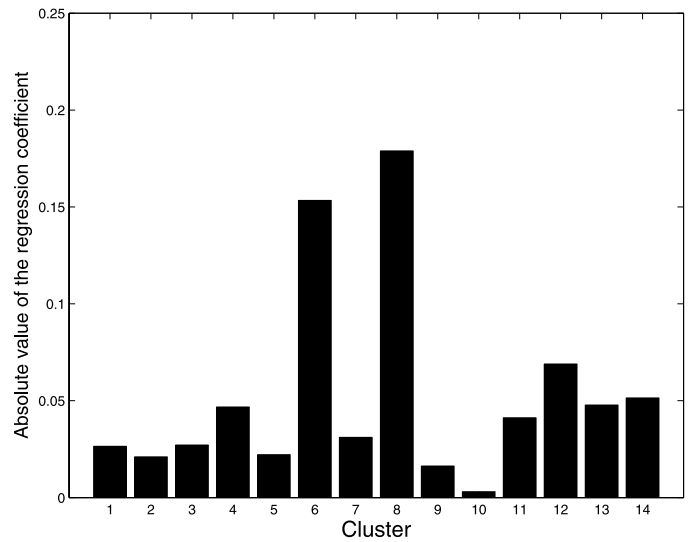

Figure 4. Estimated regression parameters from the cluster KPCA model for the 14 clusters in meditation data.

For the meditation data, we considered the linear kernel and Gaussian kernel. The bandwidth parameter in the Gaussian kernel function was chosen based on 4-fold cross-validation in the training data set (Figure 2). We considered both weighted and unweighted estimation of the cluster KPCA prediction model. The standard PCA regression model was also fitted in comparison to the proposed method. We evaluated the prediction performance of various approaches via Leave-One-Out cross validation. The comparison of results is presented in Table 2.

From Table 2, we can see Gaussian kernel outperformed the linear kernel in prediction which indicates that there is an advantage in considering nonlinear patterns in the data. The prediction error of weighted cluster KPCA regression was pretty close to that of the unweighted cluster KPCA regression. This is because the subject weight was fairly comparable across the subjects without obvious outliers (see Figure 3). To understand the roles of various brain regions in the predictive model, we compared the regression coefficients across different clusters (Figure 4) and identified two clusters that showed much higher predictive power. Figure 5 shows the spatial locations of the two clusters. The first cluster included the supplementary motor area (SMA), the hand region of the right sensorimotor cortex contralateral to the (left) hand that was pressing the button box, and the visual cortex. The first cluster is clearly functionally related to the performance of the button-press task in response to the visual stimuli. The second cluster represented a fronto-parietal system including the bilateral intraparietal sulcus and the supplementary eye fields, which is consistent with the general architecture of attentional function (Corbetta and Shulman, 2002).

When compared to the standard PCA regression model (Table 2), the cluster KPCA regression is significantly better than the PCA regression model based on the first principal component extracted across all voxels. When we increased the number of principal components to the number of clusters, the performance of the PCA became similar to that

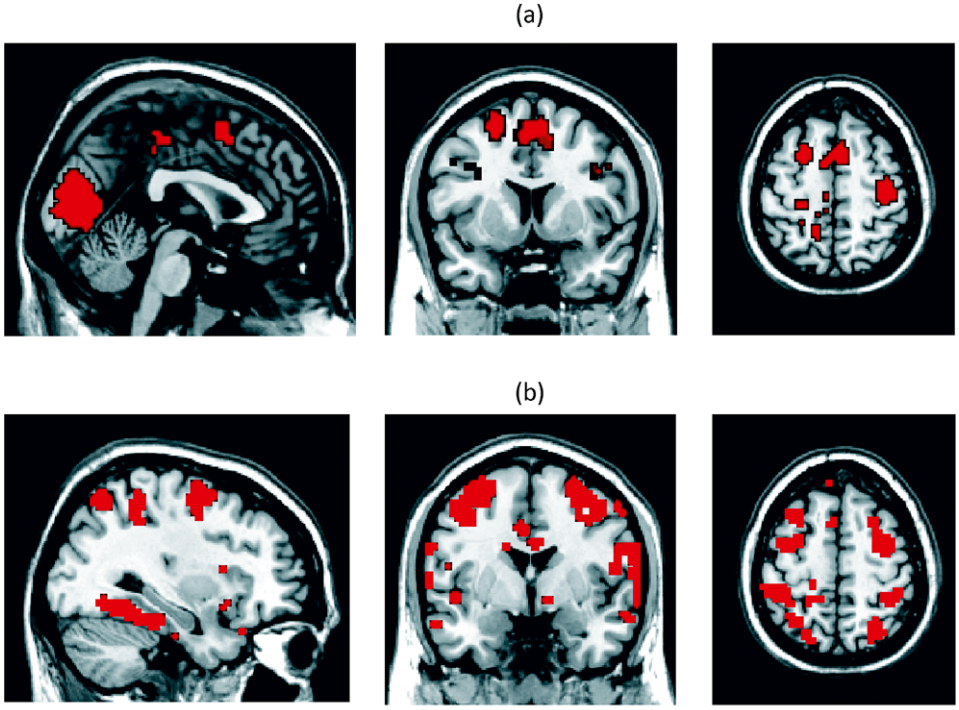

Figure 5. Spatial locations of Cluster 6 (a) and 8 (b) in meditation data.

of the cluster KPCA with the linear kernel. However, the principal components predictors in PCA model are hard to interpret since they were not associated with specific brain regions.

5. DISCUSSION

In this paper, we present a weighted cluster KPCA approach for predicting outcomes such as cognitive states based on multi-subject brain imaging data. Development of prediction models based on imaging data is a challenging topic due to the enormous number of voxels and complex correlation structures in the high dimensional data. Common approaches are to perform PCA on the whole brain image for dimension reduction and then build the predictive model on PCA basis without taking into account of the spatial structure of the brain (O'Toole et al. 2005; Mourão-Miranda et al., 2005; Kjems et al., 2002; Carlson et al., 2003; Strother et al., 2004). A few approaches considered regions of interests in the brain but summarized the ROI activity only through simple statistics. For example, Mitchell et al. (2004) used the mean activity of the selected ROI voxels to create a "supervoxel" for each ROI and built predictive models based on these supervoxels. In our method, we propose to first group voxels into functional or neuroanatomical regions and then use kernel PCA to extract features from these regions as predictors. In comparison to the existing approaches, the advantages of our method are that we retain the spatial structure of the brain in our data reduction step which allows us to evaluate the relative importance of various brain regions in the prediction; furthermore, kernel PCA is a computationally efficient feature extraction tool that provides better reflection of the principle pattern of activity within the brain regions and also allows us to consider higher dimensional patterns. Another important distinction of our method from previous works is that we propose to use subject-specific weights in the development of a predictive model. Our simulation studies have shown that it can significantly improve the prediction accuracy in the presence of outliers.

Our proposed method uses the kernel PCA method for feature extraction. In recent years, kernel based techniques have shown promising results in fMRI analysis (Hardoon and Manevitz, 2005; Ni et al. 2008). The advantages of kernel methods include reduction of computational complexity and consideration of high dimensional features in the brain images. For example, if one uses the second-order polynomial kernel, the feature space contains all pairwise products between voxels and thereby takes into account the second order association between voxels in PCA. The disadvantages of kernel methods are that: it is difficult to interpret and visualize the extracted features in the input space when nonlinear kernels are used; secondly, the choice of the kernel function is still an open question in the kernel methods literature.

We have compared the proposed cluster KPCA regression model with the standard PCA regression. Though the two may have similar predictive power when we increase the number of principal components used in the standard PCA regression, it is hard to identify the spatial sources of the

principal components of PCA since it doesn't take into account the cluster structure in the brain in the extraction of components. Furthermore, certain outcomes of interest may be related to the interconnection between different brain regions. Our proposed approach allows us to consider connectivity between various brain regions and hence can improve the performance of the predictive model. For more comprehensive evaluation of the performance of the proposed method in comparison to the standard PCA regression, one can conduct simulation studies where the outcome to predict is generated from more sophisticated models, e.g. predictive models that not only involve within-region brain activity but also depend on between-region correlations.

In this paper, we present our method for predicting a scan-specific continuous variable. The proposed procedure can be readily extended to predicting other types of outcomes by associating with the appropriate outcome model. For example, to predict categorical outcomes such as a subject's disease status, we can apply a similar procedure where the outcome is predicted by the extracted principal features through the logistic regression model and the subjects' weights can be incorporated in the iteratively weighted least square estimation of the logistic model.

ACKNOWLEDGMENTS

This work was supported by a URC grant funded by Emory University Research Committee and NIMH grant R01-MH079251. The author thanks Dr. Giuseppe Pagnoni for providing the Zen meditation data. We also thank the associate editor and the two referees for valuable comments and suggestions.

Received 2 August 2009

REFERENCES

Beckmann, C.F. and Smith, S.M. (2005). Tensorial extensions of independent component analysis for multisubject FMRI analysis, *NeuroImage* **25**, 294–311.

Boser, B. E., Guyon, I. M. and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Fifth Annual Workshop on COLT*, Pittsburgh. ACM.

Bowman, F. D. and Patel, R. (2004a). Identifying Spatial Relationships in Neural Processing Using a Multiple Classification Approach, *NeuroImage* **23**, 260–268.

Bowman, F. D., Patel, R. and Lu, C. (2004b). Methods for Detecting Functional Classifications in Neuroimaging Data, *Human Brain Mapping*, **23**, 109–119.

Calhoun, V.D., Adali, T., Pearlson, G.D. and Pekar, J.J. (2001). A method for making group inferences from functional MRI data using independent component analysis, *Human Brain Mapping* **14**, 140–151.

Carlson, T.A., Schrater, P. and He, S. (2003). Patterns of activity in the categorical representations of objects. *J. Cogn. Neurosci* **15(5)**, 704–717.

Cordes. D., Haughton, V., Carew, J.D., Arfanakis, K. and Maravilla, K. (2002). Hierarchical clustering to measure connectivity in fMRI resting-state data. *Magn Reson Imag* **20**, 305–317.

Corbetta, M. and Shulman, G.L. (2002). Control of goal-directed and stimulus-driven attention in the brain, *Nature Reviews Neuroscience* **3**, 201–215.

Cox, D.D. and Savoy RL (2003). Functional Magnetic Resonance Imaging (fMRI) "Brain Reading": Detecting and Classifying Distributed Patterns of fMRI Activity in Human Visual Cortex. *NeuroImage*, **19**: 261–270.

Derado, G., Chen, S., Guo, Y., Mayberg, H., and Bowman, F. D. (2009). Classification Methods for Identifying the Neural Characteristics of Antidepressant Treatment. Poster presentation at the 15th annual meeting of the Organization for Human Brain Mapping, San Francisco , CA.

Fischl, B., Salat, D., Busa, E., Dieterich, M., Haselgrove, C., Van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., Montillo, A., Makris, N, Rosen, B. and Dale A. (2002). Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron.*, **33**, 341–355.

Fischl, B., Van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D., Busa, E., Seidman, L., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B. and Dale A. (2004). Automatically parcellating the human cerebral cortex. *Cerebral Cortex*, **14**, 11–22.

Guo, Y. and Pagnoni, G. (2008). A unified framework for group independent component analysis for multi-subject fMRI data. *NeuroImage*, **42**, 1078–1093.

Hanson, S. J., Matsuka, T. and Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: is there a "face" area? *NeuroImage*, **23**, 156–166.

Hardoon, D. and Manevitz, L.M. (2005). fMRI Analysis via One-class Machine Learning Techniques, *19th International Joint Conference on Artificial Intelligence*, Edinburgh, Scotland, UK.

Kjems, U., Hansen, L., Anderson, J., Frutiger, S., Muley, S., Sidtis, J., Rottenberg, D. and Strother, S. C. (2002). The quantitative evaluation of functional neuroimaging experiments: Mutual information learning curves. *NeuroImage*, **15**, 772–786.

Mitchell, T., Hutchinson, R., Niculescu, R. S., Pereira, F. and Wang, X. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, **57**(1–2), 145–175.

Mourão-Miranda, J., Bokde, A. L. W., Born, C., Hampel, H. and Stetter, M. (2005). Classifying brain states and determining the discriminating activation patterns: Support Vector Machine on functional MRI data. *Neuroimage*, **28**, 980–995.

Ni, Y., Chu, C., Saunders, C.J. and Ashburner, J. (2008). Kernel Methods for fMRI Pattern Prediction. *IEEE World Congress on Computational Intelligence*, 692–697.

O'Toole, A. J., Jiang, F., Abdi, H. and Haxby, J. V. (2005). Partially distributed representations of objects and faces in ventral temporal cortex. *J Cogn Neurosci*, **17**, 580–590.

Schölkopf, B., Smola, A. J. and Müller, K.-R. (1998). Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, **10**, 1299–1319.

Strother, S., LaConte, S., Hansen, L.K., Anderson, J., Zhang, J., Pulapura, S. and Rottenberg, D. (2004). Optimizing the fMRI data-processing pipeline using prediction and reproducibility performance metrics: I. A preliminary group analysis. *NeuroImage*, **23** (Suppl. 1), s196–s207.

Taylor, J.S. and Cristianini, N. (2004). Kernel Methods for Pattern Analysis. Cambridge University Press.

Tzourio-Mazoyer N., Landeau B., Papathanassiou D., Crivello F., Etard O., Delcroix N., Mazoyer, B. and Joliot, M. (2002) Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage*, **15**, 273–289.

Ying Guo
Department of Biostatistics and Bioinformatics
Rollins School of Public Health of Emory University
1518 Clifton RD NE, Atlanta, GA, 30322, USA
E-mail address: yguo2@emory.edu