Structure Preserving Dimension Reduction with 2D Images as Predictors

Bo Zhang
IBM
Durham, NC 27703, USA
bozhang@us.ibm.com

Liwei Wang
Pharmaceutical Product Development Inc
Morrisville, NC 27560, USA
liwei.wang@ppdi.com

***Abstract*—Nearly all existing dimension reduction methods on 2D matrix-valued image predictors are unsupervised or supervised without preserving matrix structure, which can result in loss of the structure-specific relation between the response and predictors. In this paper, we propose a kernel-based solution for supervised dimension reduction which preserves the matrix structure of the reduced predictors. This approach is computationally efficient and offers a unified framework to handle image predictors. We illustrate the method using both simulations and applications.**

***Keywords*-big data; sufficient dimension reduction; central subspace; images; Kronecker product;**

I. Introduction

Among the Big Data 3Vs (volume, velocity, variety), it is the variety in data that is often ignored but actually very crucial, such as 2D matrix-valued images. Modern technologies produce, collect and store more and more such matrix-valued images. Representative examples include matrix 2D bar code, which encodes raw data in a pattern of black and white square modules; remote sensing data, which contains the accurate surface reluctance values arranged in rows and columns of the raster; Magnetoencephalography (MEG) data, which records magnetic fields over the scalp by a hundred or more sensor channels over a very short time interval; Electroencephalography (EEG) data, which measures the voltage values from multiple electrodes placed on the scalp over a short period of time.

Data analysis based on such matrix-valued images can derive meaningful insights. For example, Figure 1 exhibits characteristic EEG pattern images of an alcoholic subject (upper panel) and a non-alcoholic subject (lower panel). The vertical axis denotes the voltage value and two horizontal axes represent channel and time. It is highly suggested that the EEG pattern images are significantly different between these two subjects and hence it is of great scientific interest to perform the classification of the alcoholics and non-alcoholics via the EEG images.

While exploiting such matrix-valued images should be beneficial and essential, the reality of taking advantage of such variety is less straightforward. One key challenge is the high dimension. Sufficient dimension reduction (SDR) [1], which aims to reduce the dimension of predictors prior to any modeling efforts, is central to such high-dimensional

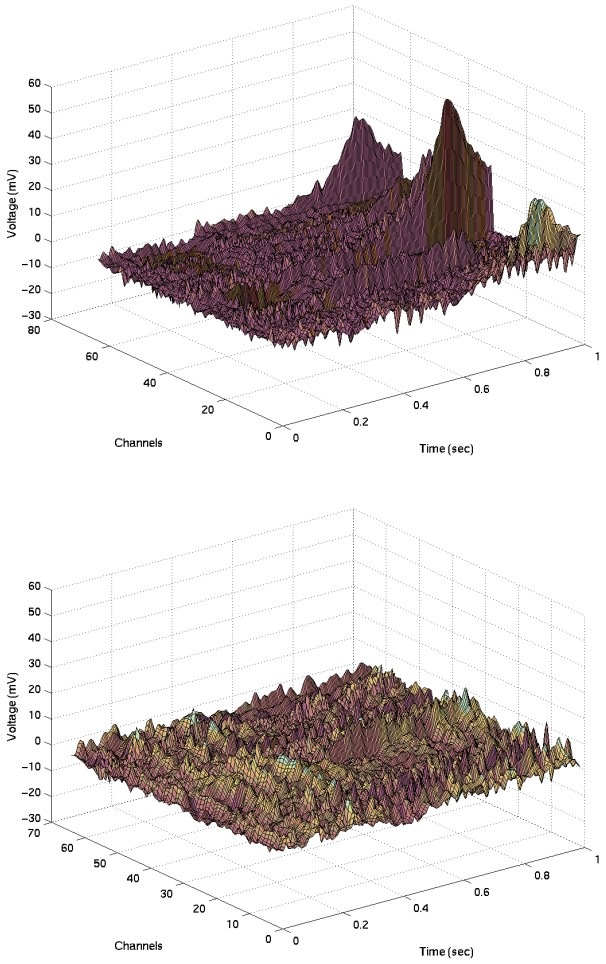

Figure 1: EEG patterns of an alcoholic and a control subject

data analysis. For classification or regression of a response Y given a p-dimensional predictor $\boldsymbol{X}$, SDR looks for a *dimension reduction subspace* $\mathcal{S}$ of $\mathbb{R}^p$ whose orthogonal basis $\boldsymbol{B} \in \mathbb{R}^{p\times d}$ satisfies that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{B}^T \boldsymbol{X}, \tag{1}$$

where $d = \dim(\mathcal{S})$ and $\perp\!\!\!\perp$ denotes independence. Under weak conditions, the intersection of all such $\mathcal{S}$ is itself a

dimension reduction subspace, in which case the intersection is called a *central subspace*. $\boldsymbol{B}^T\boldsymbol{X}$, the linear combinations of $\boldsymbol{X}$, are called the *sufficient predictors*.

There have been many popular and well studied SDR methods, which can be generally cast into three classes. The first class hinges on the inverse conditional moments of $\boldsymbol{X}|Y$ to infer about the central subspace, which employs the fact that if the conditional distribution $P(Y|\boldsymbol{X})$ concentrates on a subspace of the predictor space, then the inverse regression $E(\boldsymbol{X}|Y)$ should lie in that same subspace. Examples of this class include sliced inverse regression (SIR) [2], sliced average variance estimation (SAVE) [3], principal Hessian directions (PHD) [4], contour regression [5] and directional regression (DR) [6]. The second class relies on kernel smoothing to simultaneously estimate the basis $\boldsymbol{B}$ and the probability function $P(Y|\boldsymbol{B}^T\boldsymbol{X})$. Examples include minimum average variance estimator (MAVE) [7], constructive estimator [8], sliced regression (SR) [9]. The third class relies on the kernel method to estimate the central subspace. Examples of this class include kernel dimension reduction (KDR) [10], [11], [12] and [13].

For classification or regression of a response Y given a matrix-valued image predictor $\boldsymbol{X}$ with dimension $p\times d$, these conventional SDR approaches are achieved by reshaping $\boldsymbol{X}$ as a simple vector denoted by $\text{vec}(\boldsymbol{X})$ and then seeking for a dimension reduction subspace $\mathcal{S}$ whose orthogonal basis $\boldsymbol{B}$ satisfies that $Y \perp\!\!\!\perp \text{vec}(\boldsymbol{X})|\boldsymbol{B}^T\text{vec}(\boldsymbol{X})$. However, such dimension reduction approaches with the vectorized predictors might work poorly or even fail due to two reasons: (1) collapsing the matrix-valued predictor destroys the wealth of intrinsic structural information possessed in the matrix-valued images data, such as the correlated voltage values in the neighborhood of time points and channels in the EEG images; (2) the resulting sufficient predictors will lend themselves to interpretation inconsistent with prior structure information hidden in the original images predictors. Although many SDR methods have been proposed, relatively little attention has been paid to preserve the structure. [14] developed three inverse moment based folded SDR techniques (folded SIR, folded SAVE, and folded DR), conducting dimension reduction on the matrix-valued images predictors. However, like other inverse moment based SDR methods, all approaches rely on the elliptical assumption of the data, which may not be fulfilled in practice.

In this paper, we propose a *structure preserving dimension reduction* (SPDR) procedure using the kernel method. The goal of our method is to preserve the matrix structure information on the sufficient predictors. We achieve this goal by imposing a constraint on the central subspace, *structure preserving central subspace*.

The rest of the paper is organized as follows. In Section II, we introduce kernel method for measuring dependence. And then describe the proposed structure preserving dimension reduction in Section III. The simulation studies are presented in Section IV. In Section V, we present the applications on the motivating EEG data. Section VI concludes with a discussion.

II. Kernel Method for Measuring Dependence

It has been revealed by [15] that the apparatus of positive definite kernels can be applied to measure dependence like the one in (1) with cross-covariance operators on reproducing kernel Hilbert space (RKHS). For a set Ω, a positive definite kernel k on Ω is a function of the form k: $\Omega\times\Omega \to \mathbb{R}$ such that $\sum_{i,j=1}^{n} c_i c_j k(x_i, x_j) \geq 0$ for any $x_1, \ldots, x_n \in \Omega$ and $c_1, \ldots, c_n \in \mathbb{R}$. It is known that a positive definite kernel on Ω uniquely defines a Hilbert space $\mathcal{H}$ consisting of functions on Ω such that (i) $k(\cdot, x)$ is in $\mathcal{H}$, (ii) the linear hull of $\{k(\cdot, x)|x \in \Omega\}$ is dense in $\mathcal{H}$, and (iii) for any $x \in \Omega$ and $f \in \mathcal{H}$, $\langle f, k(\cdot, x)\rangle_{\mathcal{H}} = f(x)$ (reproducing property), where $\langle\cdot,\cdot\rangle_{\mathcal{H}}$ is the inner product of $\mathcal{H}$. The Hilbert space $\mathcal{H}$ is called the RKHS associated with k.

Let $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{Y}, \mathcal{B}_{\mathcal{Y}})$ denote measurable spaces of predictor $\boldsymbol{X}$ and response Y respectively. Also, let $k_{\mathcal{X}}$ and $k_{\mathcal{Y}}$ be positive definite kernels on $\mathcal{X}$ and $\mathcal{Y}$, respectively, with respective RKHS $\mathcal{H}_X$ and $\mathcal{H}_Y$. The *cross-covariance operator* Σ_{YX} of $(\boldsymbol{X}, Y)$ is an operator from $\mathcal{H}_X$ to $\mathcal{H}_Y$ such that $\langle g, \Sigma_{YX} f\rangle_{\mathcal{H}_Y} = \text{Cov}[f(\boldsymbol{X}), g(Y)] = E_{XY}[g(Y)f(\boldsymbol{X})] - E_Y[g(Y)]E_X[f(\boldsymbol{X})]$ for every $f \in \mathcal{H}_X$ and $g \in \mathcal{H}_Y$. The *conditional cross-covariance operator* is defined as $\Sigma_{YY|X} = \Sigma_{YY} - \Sigma_{YX}\Sigma_{XX}^{-1}\Sigma_{XY}$, where the operators Σ_{XX} and Σ_{YY} are called covariance operator. The conditional cross-covariance operator has an important property: for any orthogonal basis $\boldsymbol{B}$, $\Sigma_{YY|B^TX} \geq \Sigma_{YY|X}$, where the partial order is defined in terms of the trace operator. Moreover, the equality holds if and only if equation (1) is satisfied. This gives rise to the possibility of using the trace of the conditional cross-covariance operator as a measure of the conditional dependence in (1).

One advantage of the kernel method to measure dependence is that estimation with finite data is straightforward. Given n i.i.d samples $(X_1, Y_1), \ldots, (X_n, Y_n)$, the kernels over $\boldsymbol{B}^T\boldsymbol{X}$ and Y, $\mathcal{K}_{B^TX}$ and $\mathcal{K}_Y$, are centralized with a projection matrix $\boldsymbol{H} = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1} \in \mathbb{R}^n$ is the vector whose elements are all ones. Then the conditional cross-covariance operator is estimated by

$$\hat{\Sigma}^{(n)}_{YY|B^TX} = G_Y(G_{B^TX} + n\epsilon_n I_n)^{-1}, \tag{2}$$

where $G_Y = \boldsymbol{H}\mathcal{K}_Y\boldsymbol{H}$ and $G_{B^TX} = \boldsymbol{H}\mathcal{K}_{B^TX}\boldsymbol{H}$. Here, ϵ_n is a regularizer, smoothing the kernel matrix. It should be chosen such that when $n \to +\infty$, $\epsilon_n \to 0$ and $\sqrt{n}\epsilon_n \to +\infty$ to ensure consistency.

III. Structure Preserving Dimension Reduction with 2D Images as Predictors

In this section we propose a structure preserving dimension reduction (SPDR) procedure that aims to preserve matrix structure while conducting dimension reduction via

kernel dimension reduction. By incorporating such constraints on the central subspace, the dimension reduction problem becomes to seek for $\boldsymbol{B}_l$, the orthogonal basis of the *left dimension reduction subspace* $\mathcal{S}_l$, and $\boldsymbol{B}_r$, the orthogonal basis of the *right dimension reduction subspace* $\mathcal{S}_r$, satisfying that

$$Y \perp\!\!\!\perp \boldsymbol{X} | \boldsymbol{B}_l^T \boldsymbol{X} \boldsymbol{B}_r, \tag{3}$$

where $\boldsymbol{B}_l \in \mathbb{R}^{p \times p_l}$ and $\boldsymbol{B}_r \in \mathbb{R}^{d \times d_r}$, with $p_l \leq p$ and $d_r \leq d$.

According to [14] and the relationship of vec-operator vec and Kronecker product $\otimes$, the conditional independence (3) can be re-written as

$$Y \perp\!\!\!\perp \text{vec}(\boldsymbol{X}) | (\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T \text{vec}(\boldsymbol{X}). \tag{4}$$

Compared with the traditional independence (1), the structure preserving relation (4) requires the orthogonal basis to be structured in the form of $\boldsymbol{B}_r \otimes \boldsymbol{B}_l$, which is the Kronecker product of the orthogonal basis of left and right dimension reduction subspace. The intersection of all subspaces spanned by $\boldsymbol{B}_r \otimes \boldsymbol{B}_l$ is defined as *structure preserving central subspace*.

To minimize the conditional dependence measure in (2), while imposing the Kronecker structure on the orthogonal basis, our proposed structure preserving dimension reduction minimizes the objective function at the population level

$$\text{Trace}[G_Y (G_{(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T X} + n\epsilon_n I_n)^{-1}]. \tag{5}$$

among all $\boldsymbol{B}_l$ and $\boldsymbol{B}_r$.

The dependence measures are defined via kernels over $(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T \boldsymbol{X}$ and Y. A natural choice is a universal kernel, in particular the Gaussian kernel: $\mathcal{K}_{(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T X}(x_i, x_j) = \exp\{-||(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T x_i - (\boldsymbol{B}_r \otimes \boldsymbol{B}_l)^T x_j||^2 / \sigma_X^2\}$, and similarly for Y with a different scale σ_Y.

We propose the alternating steepest descent techniques with line search to optimize the dependence measure. The technique constrains the orthogonal basis $\boldsymbol{B}_l$ and $\boldsymbol{B}_r$ to lie on the manifold $\mathbb{S}_{p_l}^p(\mathbb{R}) = \{\boldsymbol{B}_l \in \mathbb{R}^{p \times p_l} | \boldsymbol{B}_l^T \boldsymbol{B}_r = I_{p_l}\}$ and $\mathbb{S}_{d_r}^d(\mathbb{R}) = \{\boldsymbol{B}_r \in \mathbb{R}^{d \times d_r} | \boldsymbol{B}_r^T \boldsymbol{B}_r = I_{d_r}\}$. We now describe the detailed estimation procedures for structure preserving dimension reduction at the sample level.

1. Generate the initial values of $\boldsymbol{B}_l$ and $\boldsymbol{B}_r$, say, from a sample of the $\text{N}(0, 1)$ variables;
2. Minimize the objective function (5) with respect to $\boldsymbol{B}_l$ by steepest descent with $\boldsymbol{B}_r$ fixed. The step is estimated using line search;
3. Minimize the objective function (5) with respect to $\boldsymbol{B}_r$ by steepest descent with the newly computed $\boldsymbol{B}_l$ fixed. The step is estimated using line search;
4. Repeat steps 2 and 3 until the objective function stabilizes.

Since the minimization of the contrast function requires a nonlinear optimization technique, potential problems with local optima exist. There are two important tuning parameters. One is the regularization coefficient ϵ_n. The other one is the scale σ. As both of σ and ϵ_n have a similar smoothing effect, it is reasonable to fix one of them and select the other; in our study, we fixed $\epsilon_n = 0.0001$ as an arbitrary choice and varied σ. Our way to deal with varying σ is to use a continuation method in which the scale parameter σ in Gaussian kernel is gradually decreased during the iterative optimization, controlled by ρ. More concretely, we fixed a number of iterations T and calculated

$$\sigma_{(t)}^2 = \sigma_{(0)}^2 + \sigma_{(0)}^2 (T - t)(\rho - 1)/T \tag{6}$$

as the scale parameter for the t-th iteration, where the initial value of σ is set to be the median of pairwise distances of the data. In the simulations and applications, we chose 5-fold cross-validation to tune ρ.

IV. Simulations

In this section we compare the finite-sample performance of the proposed structure preserving dimension reduction (SPDR) with two conventional dimension reduction approaches with vectorized predictors: SIR [2] and DR [6], and their corresponding dimension folding methods without vectorizing matrix-valued predictors: folded dimension reduction (FDR) and folded sliced inverse regression (FSIR) [14].

To evaluate the estimation accuracy of various estimators of left and right dimension reduction, we use the vector correlation coefficient,

$$\text{VCC}_l = |\hat{\boldsymbol{B}}_l^T \boldsymbol{B}_l \boldsymbol{B}_l^T \hat{\boldsymbol{B}}_l|^{1/2} \tag{7}$$

and

$$\text{VCC}_r = |\hat{\boldsymbol{B}}_r^T \boldsymbol{B}_r \boldsymbol{B}_r^T \hat{\boldsymbol{B}}_r|^{1/2}, \tag{8}$$

where $|\cdot|$ denotes the determinant of a matrix. Note that $0 \leq \text{VCC} \leq 1$, and when $\text{VCC} = 1$, $\mathcal{S}_{Y|\boldsymbol{X}}(\hat{\boldsymbol{B}}_r \otimes \hat{\boldsymbol{B}}_l) = \mathcal{S}_{Y|\boldsymbol{X}}(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)$. Therefore, higher values of VCC imply that the two spaces are closer and, hence, the estimates are more accurate.

Also, the criterion used in [5], denoted by LZC, measures the distance between true central subspace $\mathcal{S}_{Y|\boldsymbol{X}}(\hat{\boldsymbol{B}}_r \otimes \hat{\boldsymbol{B}}_l)$ and estimated structure preserving central space $\mathcal{S}_{Y|\boldsymbol{X}}(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)$ using

$$\text{LZC} = ||\text{span}(\hat{\boldsymbol{B}}_r \otimes \hat{\boldsymbol{B}}_l) - \text{span}(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)||, \tag{9}$$

where $||\cdot||$ is a Frobenius norm of a matrix. This is a measure of discrepancy between the subspaces $\text{span}(\hat{\boldsymbol{B}}_r \otimes \hat{\boldsymbol{B}}_l)$ and $\text{span}(\boldsymbol{B}_r \otimes \boldsymbol{B}_l)$ and hence smaller values of LZC yield more accurate estimates.

We summarize our findings in the following simulation example:

$$Y = \begin{cases} 1 & \text{if } 0.1(\boldsymbol{B}_r^T \boldsymbol{X} \boldsymbol{B}_l)^2 > 1, \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Table I: Comparison among Dimension Reduction Methods for Matrix-valued Predictors

n	Method	$\boldsymbol{B}_r(VCC)$	$\boldsymbol{B}_l(VCC)$	$\boldsymbol{B}(LZC)$
100	SPDR	0.987	0.992	0.047
	FDR	0.652	0.653	1.106
	DR			2.591
	FSIR	0.426	0.345	1.896
	SIR			2.578
200	SPDR	0.967	0.968	0.091
	FDR	0.879	0.875	0.392
	DR			1.651
	FSIR	0.411	0.407	1.889
	SIR			2.498
300	SPDR	0.999	0.999	0.007
	FDR	0.957	0.905	0.148
	DR			0,798
	FSIR	0.386	0.363	1.887
	SIR			2.487
500	SPDR	0.990	0.993	0.025
	FDR	0.961	0.965	0.104
	DR			0,504
	FSIR	0.442	0.359	1.873
	SIR			2.476

We set $\boldsymbol{B}_r = (1, -1, 1, 0, 0)^T, \boldsymbol{B}_l = (-1, 1, -1, 0, 0)^T$ and generate the orthogonal basis of the true central subspace as $\boldsymbol{B} = \boldsymbol{B}_r \otimes \boldsymbol{B}_l$. Then we generate the matrix-valued predictors $\boldsymbol{X}$ of size 5×5. We vary the sample size $n = 100, 200, 300, 500$.

We evaluate the performance of each method from two aspects:

1. the left central subspace estimation using VCC defined in (7) and the right central subspace estimation using VCC defined in (8); Note this is not applied to the two conventional approaches SIR and DR since they don't estimate left and right central subspace;
2. the whole central subspace estimation using LZC defined in (9).

Table I reports the means of VCC and LZC, as calculated from $M = 100$ simulated samples for each n. The standard errors of theses means are all less than 0.05 so that we skip reporting them.

We have the following observations. First, we observe big improvements by the dimension reduction methods without vectorizing the matrix-valued predictors. This is because the structured orthogonal basis $\boldsymbol{B}_r \otimes \boldsymbol{B}_l$ has only 10 parameters while the unstructured orthogonal basis $\boldsymbol{B}$ has 25 parameters. Secondly, our SPDR outperforms FDR and FSIR in all scenarios. This is because FDR and FSIR rely on the elliptical assumption of the data, which is not fulfilled in the simulated data.

V. Applications

We now apply the proposed SPDR method to analyze the motivating EEG images data, which is from a research study to examine EEG correlates of genetic predisposition to alcoholism. The study involved two groups of subjects: an alcoholic group of 77 subjects and a control group of 45 subjects. Each subject was exposed to either one stimulus or two stimuli. During an exposure, the voltage values were measured from 64 channels of electrodes and for 256 time points. The 64 electrodes are placed at different locations on the subject's scalp. The stimuli were pictures chosen from a picture set. When two pictures were shown, they were displayed in either a matched condition, where two pictures were identical, or a unmatched condition, where they were different. Each subject had 120 trials under these three conditions: single stimulus, two matched stimuli and two unmatched stimuli. The primary interest was to study the association between alcoholism and the pattern of voltage values over times and channels.

To keep matters simple, in this paper, we used only part of the data set: we included only the single stimulus condition and, for each subject, we took the average of all the trials under that condition. That is, the portion of the data we used consists of $(\boldsymbol{X}_1, Y_1), ..., (\boldsymbol{X}_{122}, Y_{122})$, where $\boldsymbol{X}_i$ is a 256×64 matrix with each entry representing the mean voltage value of subject i at a combination of a time point and a channel, averaged over all trials under the single stimulus condition, and Y_i is a binary random variable indicating whether the ith subject is alcoholic ($Y_i = 1$) or nonalcoholic ($Y_i = 0$). To visualize and illustrate the results in a meaningful way, we set $p_l = d_r = 2$.

Firstly, we implemented a matrix version of principal components analysis preprocessing by removing noisy information. Let $\boldsymbol{V} = (v_1, ..., v_{64})$, where $v_1, ..., v_{64}$ are the first 64 eigenvectors of the matrix $\mathrm{E}_n(\boldsymbol{X} - \bar{\boldsymbol{X}})(\boldsymbol{X} - \bar{\boldsymbol{X}})^T$. And let $\boldsymbol{W} = (w_1, ..., w_{16})^T$, where $w1, ..., w_{16}$ are the first 16 eigenvectors of the matrix $\mathrm{E}_n(\boldsymbol{X} - \bar{\boldsymbol{X}})^T(\boldsymbol{X} - \bar{\boldsymbol{X}})$. Then the preprocessed predictors used in the dimension reduction will be $\boldsymbol{V}^T\boldsymbol{X}_i\boldsymbol{W} \in \mathbb{R}^{64\times 16}$ for each subject i.

Secondly, we conducted sufficient dimension reduction using SPDR on the whole 122 subjects data set. Figure 2 shows the scatterplot matrix of the four sufficient predictors obtained by SPDR in the 2×2 matrix

$$\boldsymbol{B}_r^T\boldsymbol{X}\boldsymbol{B}_l = \begin{bmatrix} X11 & X12 \\ X21 & X22 \end{bmatrix}. \tag{11}$$

In the scatterplots, the EEG images data for alcoholic subjects (red $+$) show more variation than those for control subjects (blue $\circ$), which indicates that the EEG images for the alcoholic subjects are more similar than the EEG images for the control subjects.

Thirdly, we applied the following leave-one-out procedure to predict a subject's alcoholic status using the subject's EEG image data.

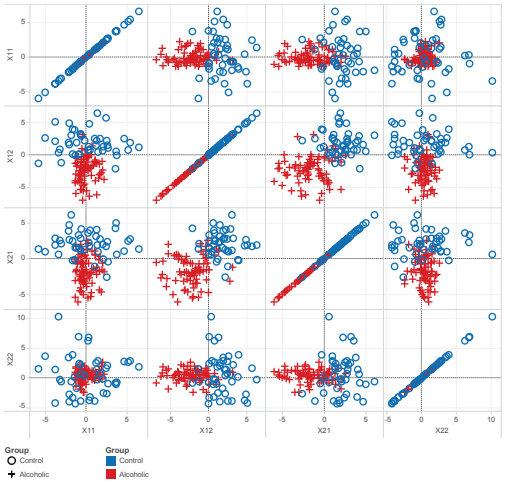

Figure 2: Scatterplot matrices for the four reduced predictors estimated by SPDR.

1) (Leave-one-out) Set aside the ith subject from all 122 subjects, reorder the remaining 121 subjects and divide them sequentially into K = 5 roughly equal-sized folds of roughly 24 subjects each;
2) (Parameter-tuning) Choose the scale parameter ρ from 5,10,20,30,40;
3) (Five-fold-cross-validation) Hold out one of folds, leverage only the data from the other 4 folds to estimate $\boldsymbol{B}_l$ and $\boldsymbol{B}_r$ using SPDR, predict the class labels for the withheld fold using logisitic regression with sufficient predictor $\boldsymbol{B}_r^T \boldsymbol{X} \boldsymbol{B}_l$, and obtain the misclassification error;
4) Repeat step 3 for each of the folds and average to estimate the validation error;
5) Repeat steps 2 and 3 for each of the scale parameter value, choose the optimal tuning parameter corresponding to the minimum validation error, fit the training data (121 subjects) with tuned parameter using SPDR to estimate $\boldsymbol{B}_l$ and $\boldsymbol{B}_r$, predict the class label for the withheld subject;
6) Repeat steps 1,2 and 3 for each subject.

Following the above procedure, we correctly predicted 96 out of the 122 subjects. In addition to such good classification accuracy, one key advantage of using SPDR is its potential to find linear combinations of interesting channel and time regions associated alcoholism. The matrix-valued sufficient predictors $\boldsymbol{B}_r^T \boldsymbol{X} \boldsymbol{B}_l$ is helpful to gain further insights into the relationship between EEG patterns and alcoholism. Specifically, the right dimension reduction subspace spanned by $\boldsymbol{B}_r$ measures the weights of channels that are more useful in classifying alcoholism status, whereas the left dimension reduction subspace spanned by $\boldsymbol{B}_l$ detects useful patterns of how voltage varies over time in these useful channels. This cannot be achieved by black box classifiers or conventional dimension reduction methods. Note that if we use a larger portion of the full data set, a higher prediction accuracy could be achieved.

VI. Conclusion

Motivated by big variety data such as 2D matrix-valued images, we propose a structure preserving version of sufficient dimension reduction. Preserving image structure into dimension reduction is pivotal in practice, because not only it can improve the interpretability of sufficient predictors, but also increase the estimation accuracy. There are more variety given by the predictors could be handled by structure preserving dimension reduction, such as 3D array-valued predictors. We're currently pursuing this extension and will report the results elsewhere.

Acknowledgment

The authors would like to thank Professor Lexin Li from University of California, Berkeley and Professor Hua Zhou from University of California, Los Angeles for their constructive and inspiring comments and suggestions on our research on dimension reduction with images as predictors.

References

[1] R. D. Cook, "Regression graphics: Ideas for studying regressions through graphics," *New York: Wiley*, vol. 1, 1998.

[2] K.-C. Li, "Sliced inverse regression for dimension reduction," *Journal of the American Statistical Association*, vol. 86, pp. 316–327, 1991.

[3] R. D. Cook and S. Weisberg, "Discussion of li," *Journal of the American Statistical Association*, vol. 86, pp. 328–332, 1991.

[4] K.-C. Li, "On principal hessian directions for data visualization and dimension reduction: Another application of stein's lemma," *Journal of the American Statistical Association*, vol. 87, pp. 1025–1039, 1992.

[5] B. Li, H. Zha, and F. Chiaromonte, "Contour regression: a general approach to dimension reduction," *The Annals of Statistics*, vol. 33, no. 4, pp. 1580–1616, 2005.

[6] B. Li and S. Wang, "On directional regression for dimension reduction," *Journal of the American Statistical Association*, vol. 102, pp. 997–1008, September 2007.

[7] Y. Xia, H. Tong, W. K. Li, and L.-X. Zhu, "An adaptive estimation of dimension reduction space," *Journal of The Royal Statistical Society Series B*, vol. 64, no. 3, pp. 363–410, 2002.

[8] Y. Xia, "A constructive approach to the estimation of dimension reduction directions," *The Annals of Statistics*, vol. 35, pp. 2654–2690, 2007.

[9] H. Wang and Y. Xia, "Sliced regression for dimension reduction." *J. Am. Stat. Assoc.*, vol. 103, no. 482, pp. 811–821, 2008.

[10] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Kernel dimension reduction in regression," *The Annals of Statistics*, vol. 37, no. 4, pp. 1871–1905, 2009.

[11] Q. Wu, S. Mukherjee, and F. Liang, "Localized sliced inverse regression." in *NIPS*, 2008, pp. 1785–1792.

[12] H. Zhu and L. Li, "Biological pathway selection through nonlinear dimension reduction," *Biostatistics*, vol. 12, pp. 429–444, 2011.

[13] B. Li, A. Artemiou, and L. Li, "Principal support vector machines for linear and nonlinear sufficient dimension reduction," *Annals of Statistics*, vol. 39, no. 6, pp. 3182–3210, 2011.

[14] B. Li, M. K. Kim, and N. Altman, "On dimension folding of matrix- or array-valued statistical objects," *Annals of Statistics*, vol. 38, pp. 1094–1121, 2010.

[15] K. Fukumizu, F. R. Bach, and M. I. Jordan, "Dimensionality reduction for supervised learning with reproducing kernel Hilbert spaces," *J. Mach. Learn. Res.*, vol. 5, pp. 73–99 (electronic), 2004.