Journal of the American Statistical Association

ISSN: 0162-1459 (Print) 1537-274X (Online) Journal homepage: http://amstat.tandfonline.com/loi/uasa20

MWPCR: Multiscale Weighted Principal Component Regression for High-Dimensional Prediction

Hongtu Zhu, Dan Shen, Xuewei Peng & Leo Yufeng Liu

To cite this article: Hongtu Zhu, Dan Shen, Xuewei Peng & Leo Yufeng Liu (2017) MWPCR: Multiscale Weighted Principal Component Regression for High-Dimensional Prediction, Journal of the American Statistical Association, 112:519, 1009-1021, DOI: 10.1080/01621459.2016.1261710

To link to this article: https://doi.org/10.1080/01621459.2016.1261710

JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2017, VOL. 112, NO. 519, 1009–1021, Applications and Case Studies
https://doi.org/10.1080/01621459.2016.1261710

Taylor & Francis
Taylor & Francis Group

MWPCR: Multiscale Weighted Principal Component Regression for High-Dimensional Prediction

Hongtu Zhu[a,b], Dan Shen[c], Xuewei Peng[d], and Leo Yufeng Liu[a,e] and for the Alzheimer's Disease Neuroimaging Initiative

[a]Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX; [b]Department of Biostatistics, University of North Carolina, Chapel Hill, NC; [c]Interdisciplinary Data Sciences Consortium and Department of Mathematics and Statistics, University of South Florida, Tampa, FL; [d]Texas A&M University, College Station, TX; [e]Department of Statistics and Operations Research, University of North Carolina, Chapel Hill, NC

ABSTRACT

We propose a multiscale weighted principal component regression (MWPCR) framework for the use of high-dimensional features with strong spatial features (e.g., smoothness and correlation) to predict an outcome variable, such as disease status. This development is motivated by identifying imaging biomarkers that could potentially aid detection, diagnosis, assessment of prognosis, prediction of response to treatment, and monitoring of disease status, among many others. The MWPCR can be regarded as a novel integration of principal components analysis (PCA), kernel methods, and regression models. In MWPCR, we introduce various weight matrices to prewhitten high-dimensional feature vectors, perform matrix decomposition for both dimension reduction and feature extraction, and build a prediction model by using the extracted features. Examples of such weight matrices include an importance score weight matrix for the selection of individual features at each location and a spatial weight matrix for the incorporation of the spatial pattern of feature vectors. We integrate the importance of score weights with the spatial weights to recover the low-dimensional structure of high-dimensional features. We demonstrate the utility of our methods through extensive simulations and real data analyses of the Alzheimer's disease neuroimaging initiative (ADNI) dataset. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received January 2015
Revised September 2016

KEYWORDS
Alzheimer; Feature; Principal component analysis; Regression; Spatial; Supervised

1. Introduction

The Alzheimer's Disease Neuroimaging Initiative (ADNI) study began in 2004 and is the first "Big Data" project for Alzheimer's disease (AD), which has been a groundbreaking project. It has collected imaging, genetic, clinical, and cognitive data from thousands of subjects to delineate the complex relationships among the clinical, cognitive, imaging, genetic, and biochemical biomarker characteristics of the entire spectrum of AD as the pathology evolves from normal aging (NC), to mild cognitive impairment (MCI), to dementia or AD. This article is motivated by the joint analysis of fluorodeoxyglucose positron emission tomography (FDG-PET) data and clinical and behavioral variables from $n = 196$ subjects in the ADNI study. After applying a standard preprocessing pipeline, the dimension of the processed FDG-PET images is $79 \times 95 \times 69$. We are particularly interested in addressing two questions:

- (Q1) the first one is to identify FDG-PET imaging biomarkers for classifying subjects to either AD or NC group;
- (Q2) the second one is to identify FDG-PET imaging biomarkers observed at baseline to accurately predict the change in the Alzheimer's Disease Assessment Scale-Cognitive (ADAS-Cog) test score at least 2 years later after initial assessment.

Statistically, these questions of interest can be formulated as the use of a high-dimensional vector of features (or FDG-PET), denoted as $\mathbf{x} = (\mathbf{x}_\mathbf{g} : \mathbf{g} \in \mathcal{G})$, to predict an outcome variable, denoted as $\mathbf{y}$, where $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_p\}$ is a set of locations, in which p is the total number of locations in $\mathcal{G}$. In this case, $\mathbf{x}$ is a vector of FDG-PET imaging measures on a three-dimensional (3D) lattice and $\mathbf{y}$ is either disease status in (i) or the change in the ADAS-cog score in (ii). Figure 1 shows some selected slices of the processed PET images from three randomly selected Alzheimer's Disease (AD) subjects and three randomly selected normal control (NC) subjects.

To answer questions (Q1) and (Q2), we develop a multiscale weighted principal component regression (MWPCR) framework to deal with three challenges arising from the use of high-dimensional $\mathbf{x}$ with strong spatial features (e.g., FDG-PET) to predict $\mathbf{y}$. Such challenges include (i) noisy functional data, (ii) complex spatial information, and (iii) the remarkable variability of brain structure and function across subjects. For instance, in most neuroimaging studies, the dimension of neuroimaging data (or $\mathbf{x}$) can be much larger than the number of subjects, which varies from several dozens to a few thousands. Moreover, different components of $\mathbf{x}$ may be highly correlated with each other and share some specific spatial structures (Friston 2009; Hinrichs et al. 2009; Vincent et al. 2011; Cuingnet et al. 2013).

CONTACT Hongtu Zhu htzhu@email.unc.edu Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX 77230, and University of North Carolina, Chapel Hill, NC 27599, USA.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/r/JASA.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA

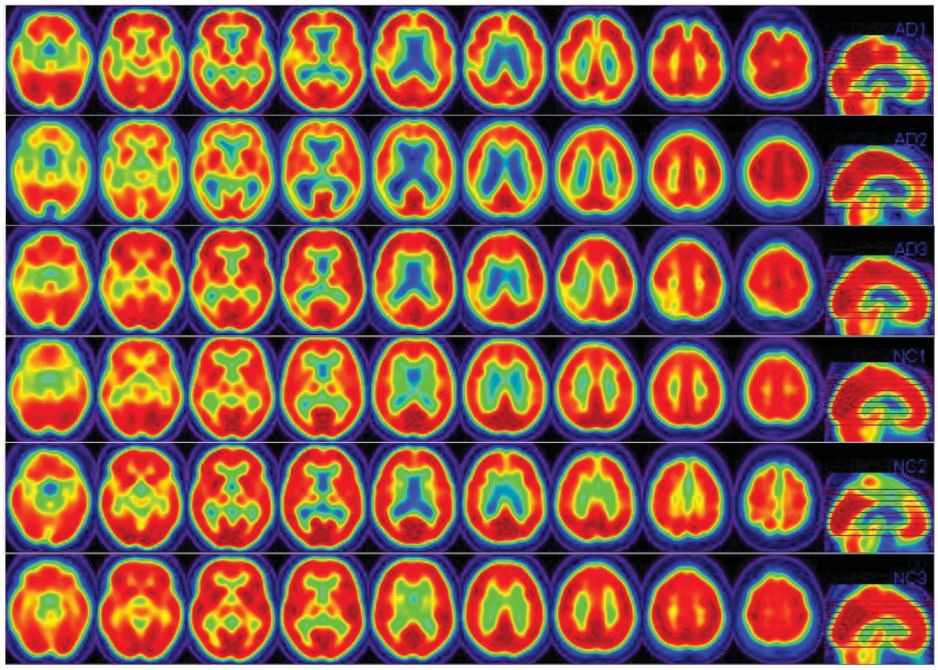

Figure 1. ADNI PET data. Each row consists of preselected two-dimensional (2D) slides obtained from a randomly selected subject. The first three rows come from three randomly selected AD subjects and the last three rows come from three randomly selected NC subjects.

Many existing supervised learning and variable selection methods (Tibshirani 1996; Bickel and Levina 2004; Fan and Fan 2008; Hastie, Tibshirani, and Friedman 2009; Clarke, Fokoue, and Zhang 2009; Bühlmann et al. 2013), however, can be sub-optimal for high-dimensional prediction problem considered here, since the effect of high-dimensional data $\mathbf{x}$ (e.g., image biomarker) on $\mathbf{y}$ is often *nonsparse* (Friston 2009; Hinrichs et al. 2009; Zhou, Li, and Zhu 2013; Li et al. 2015). First, the existing unstructured regularization methods can suffer from diverging spectra and noise accumulation in high-dimensional feature space (Bickel and Levina 2004; Fan and Fan 2008; Reiss and Ogden 2010; Bühlmann et al. 2013), whereas the structured ones (e.g., fused Lasso or Ising prior) can be computationally challenging for high-dimensional imaging predictor (Vincent et al. 2011; Cuingnet et al. 2013; Fan, Feng, and Tong 2012; Goldsmith, Huang, and Crainiceanu 2014). Alternatively, it is imperative to use some dimension reduction methods, such as principal component analysis and/or screening methods, to extract and select important "low-dimensional" features, while eliminating redundant features (Bair et al. 2006; Skocaj, Leonardis, and Bischof 2007; Fan and Fan 2008; Krishnan et al. 2011; Zhao, Ogden, and Reiss 2012). Moreover, most supervised learning methods coupled with dimension reduction methods do not account for the strong spatial features of high-dimensional imaging data as discussed above (Allen, Grosenick, and Taylor 2014; Guo, Ahn, and Zhu 2015).

A general framework of MWPCR is developed to address some of the challenges discussed above. The MWPCR provides a simple solution to the problem of interest by hierarchically and spatially extracting low-dimensional "transformed" variables from $\mathbf{x}$ to dramatically improve prediction accuracy. Compared with the existing literature (Allen, Grosenick, and Taylor 2014; Guo, Ahn, and Zhu 2015; Shen and Zhu 2015), we make several major contributions as follows:

- (i) MWPCR provides a comprehensive and powerful dimension reduction framework for integrating feature selection, smoothing, and feature extraction for continuous and discrete response variables (e.g., binary response for classification).
- (ii) We evaluate the finite sample properties of MWPCR by using both simulation studies and the analysis of ADNI data. Our numerical results reveal that MWPCR significantly outperforms many competing methods under some scenarios.
- (iii) We systematically investigate the theoretical properties of MWPCR under the high-dimensional binary classification setting. Specifically, we are able to reveal the importance of incorporating different types of weights for improving classification accuracy.
- (iv) The code for MWPCR was written in Matlab, which along with its documentation will be freely accessible from the public website *http://www.nitrc.org* and our lab website *http://odin.mdacc.tmc.edu/bigs2/*.

The article is organized as follows. In Section 2, we introduce the model setup of MWPCR. We discuss various strategies of determining global and local weights that account for an association between $\mathbf{y}$ and each individual feature $\mathbf{x}_\mathbf{g}$ across $\mathbf{g} \in \mathcal{G}$ and the spatial patterns of $\mathbf{x}$. In Section 3, simulation studies are conducted to examine the finite sample performance of MWPCR. We conduct real data analysis in Section 4 based on ADNI data to address the two questions (Q1) and (Q2) discussed above. We give some concluding remarks in Section 5. We also investigate some theoretical properties of MWPCR under the high-dimensional binary classification setting and put them in the supplementary document.

2. Multiscale Weighted Principal Component Regression

In this section, we describe data structure and then introduce the model setup and estimation method of MWPCR.

2.1. Data Structure

Consider data from n independent subjects. For each subject, we observe a $q_y \times 1$ vector of discrete or continuous responses, denoted by $\mathbf{y}_i = (y_{i,1}, \ldots, y_{i,q_y})^T$, a $q_z \times 1$ vector of discrete and/or continuous clinical covariates, denoted by $\mathbf{z}_i = (z_{i,1}, \ldots, z_{i,q_z})^T$, and a $p \times 1$ vector of data $\mathbf{x}_i = \{\mathbf{x}_{i,\mathbf{g}} : \mathbf{g} \in \mathcal{G}\}$ measured on $\mathcal{G}$ for $i = 1, \ldots, n$. Let $\mathbf{X}^T = (\mathbf{x}_1 | \ldots | \mathbf{x}_n)$ be a $p \times n$ matrix. In many cases, both q_y and q_z are relatively small compared with n, whereas p is much larger than n. For instance, in many imaging studies, it is common to use high-dimensional imaging data to classify a class variable, such as disease status. In this case, q_y is as small as one, whereas p can be several millions. Moreover, $\mathcal{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_p\}$ is a set of prefixed locations, such as voxels in 3D lattices, so it is possible to define an edge set $\mathcal{S} = \{(\mathbf{g}_k, \mathbf{g}_j) : \mathbf{g}_k, \mathbf{g}_j \in \mathcal{G}\}$ associated with $\mathcal{G}$. For instance, in spatial statistics and imaging analysis, one often uses pixels and their first-order (or high-order) neighboring pixels to construct edges in $\mathcal{S}$.

2.2. Model Setup

The proposed MWPCR consists of two components: a low-rank model for multi-scale weighted PCA (MWPCA) and a prediction model. Let $Q^{(\ell)}$ be a $p \times p$ weight matrix at the ℓth scale for $\ell = 1, \ldots, L$. The low-rank model for MWPCA can be written as

$$\begin{aligned}\widetilde{\mathbf{X}}^{(\ell)} &= (\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}^T) Q^{(\ell)} = U^{(\ell)} D^{(\ell)} V^{(\ell)T} + \mathcal{E}^{(\ell)} \\ &= \sum_{k=1}^{K} d_k^{(\ell)} \mathbf{u}_k^{(\ell)} \mathbf{v}_k^{(\ell)T} + \mathcal{E}^{(\ell)} \end{aligned} \tag{1}$$

for $\ell = 1, \ldots, L$, where $E(\mathbf{x}_i) = \boldsymbol{\mu}$, $K \le \min(n, p)$, and $\mathcal{E}^{(\ell)} = (\boldsymbol{\epsilon}_1^{(\ell)}, \ldots, \boldsymbol{\epsilon}_n^{(\ell)})^T$ is an $n \times p$ matrix of measurement errors that follows a matrix-variate distribution with mean $\mathbf{0}_{n,p}$ and an arbitrary covariance matrix. Moreover, $U^{(\ell)} = (\mathbf{u}_1^{(\ell)}, \ldots, \mathbf{u}_K^{(\ell)})$, $D^{(\ell)} = \mathrm{diag}(d_1^{(\ell)}, \ldots, d_K^{(\ell)})$, and $V^{(\ell)} = (\mathbf{v}_1^{(\ell)}, \ldots, \mathbf{v}_K^{(\ell)})$ are, respectively, $n \times K$, $K \times K$, and $p \times K$ matrices such that $\mathrm{diag}(D^{(\ell)}) \ge 0$ and $U^{(\ell)T} U^{(\ell)} = V^{(\ell)T} V^{(\ell)} = I_K$, where I_K is a $K \times K$ identity matrix.

We combine all $\{U^{(\ell)}\}_{\ell \ge 1}$ from different scales into an $n \times (KL)$ matrix given by $U_C = (\mathbf{u}_{C,1} \ldots \mathbf{u}_{C,n})^T = (U^{(1)}, \ldots, U^{(L)})$. We then build a prediction model $R(\mathbf{y}_i; \mathbf{u}_{C,i}, \mathbf{z}_i, \boldsymbol{\theta})$ with $\mathbf{y}_i$ as response and $\mathbf{u}_{C,i}$ and $\mathbf{z}_i$ as covariates, where $\boldsymbol{\theta}$ is a vector of unknown (finite-dimensional or nonparametric) parameters. For instance, when $q_y = 1$, a popular prediction model is the generalized linear model given by

$$f(\mathbf{y}_i; \mathbf{u}_{C,i}, \mathbf{z}_i, \boldsymbol{\theta}) = \exp\left(\phi\{\eta_i \mathbf{y}_i - b(\eta_i)\} + s(y_i, \phi)\right), \tag{2}$$

where ϕ is a dispersion parameter and $b(\cdot)$ and $s(\cdot, \cdot)$ are known functions. Moreover, it is assumed that $\dot{b}(\eta_i) = db(\eta_i)/d\eta_i = E(\mathbf{y}_i | \mathbf{u}_{C,i}, \mathbf{z}_i)$ satisfies $h(\dot{b}(\eta_i)) = \mathbf{z}_i^T \boldsymbol{\beta}_z + \mathbf{u}_{C,i}^T \boldsymbol{\beta}_u$, where $\boldsymbol{\beta}_z$ and $\boldsymbol{\beta}_u$ are coefficient vectors associated with $\mathbf{z}_i$ and $\mathbf{u}_{C,i}$, respectively, and $h(\cdot)$ is a link function. In this case, we have $\boldsymbol{\theta} = (\phi, \boldsymbol{\beta}_z, \boldsymbol{\beta}_u)$. Our prediction model can be various parametric and nonparametric regression models for continuous and discrete responses and multivariate and univariate responses, such as survival data and classification problems (Hastie, Tibshirani, and Friedman 2009; Clarke, Fokoue, and Zhang 2009).

The key novelty of MWPCR is the use of MWPCA to extract important low-dimensional features of $\mathbf{x}$ that are predictive of $\mathbf{y}$. Our MWPCA can be regarded as a novel extension of various supervised and unsupervised dimension reduction models for matrix decomposition (Skocaj, Leonardis, and Bischof 2007; Huang, Shen, and Buja 2009; Allen, Grosenick, and Taylor 2014). Specifically, the three key features of MWPCA include the integration of importance score weights and spatial weights, a multiscale strategy for feature extraction, and its computational efficiency. In contrast, although a general duality diagram method (Skocaj, Leonardis, and Bischof 2007; Dray and Jombart 2011) explicitly incorporates two weight matrices, it only accounts for structural dependencies (e.g., smoothness) in $\mathbf{x}$.

2.3. Estimation Procedure

We introduce a three-stage algorithm for MWPCR as follows.

- Stage 1. Build an importance score vector (or function) $W_I = (w_{I,\mathbf{g}}) : \mathcal{G} \to R^+$ and a spatial weight matrix $W_E = (w_{E,\mathbf{g}\mathbf{g}'}) : \mathcal{G} \times \mathcal{G} \to R$.
- Stage 2. At the ℓth scale, use W_E and W_I to build a spatial weight matrix $Q^{(\ell)}$ and then compute the first K principal components in $U^{(\ell)}$ according to model (1). Repeat it for $\ell = 1, \ldots, L$.
- Stage 3. Build the prediction model $R(\mathbf{y}; \mathbf{u}_C, \mathbf{z}, \boldsymbol{\theta})$.

We slightly elaborate on these stages. In Stage 1, the importance scores $w_{I,\mathbf{g}}$ play an important feature screening role in MWPCR and they can be learned directly either from $\{\mathbf{x}, \mathbf{y}\}$ or other sources. Examples of $w_{I,\mathbf{g}}$ in the literature are primarily based on some statistics (e.g., Pearson correlation or distance correlation) between $\mathbf{x}_\mathbf{g}$ and $\mathbf{y}$ at each location $\mathbf{g}$ used in the sure independence screening (Bair et al. 2006; Li, Zhong, and Zhu 2012). However, most importance scores $w_{I,\mathbf{g}}$ are independently calculated at each location, so they largely ignore complex spatial structures at different locations.

In Stage 1, $W_E = (w_{E,\mathbf{g}_k\mathbf{g}_j}) \in R^{p \times p}$ can be either symmetric or asymmetric. The elements $w_{E,\mathbf{g}_k\mathbf{g}_j}$ are usually calculated by using various similarity criteria, such as Gaussian similarity from Euclidean distance, local neighborhood relationship, correlation, and prior information obtained from other data (Yan et al. 2007). Then, we can threshold W_E to create an adjacency matrix with elements of either 1 or 0, which leads to $\mathcal{S}$, depending on whether the corresponding correlation value exceeds a prefixed threshold or not. By choosing different thresholds, we can obtain different edge sets $\mathcal{S}$. In Section 2.4, we will discuss how to determine W_E and W_I, while explicitly accounting for the complex spatial structure among different locations.

In Stage 2, we construct the weight matrix $Q^{(\ell)}$ at the ℓth scale as follows. To extract important features from $\mathbf{x}$, we construct a matrix $Q_1^{(\ell)} = \mathrm{diag}(1\{w_{I,\mathbf{g}_1} \ge s_{I,\ell}\}, \ldots, 1\{w_{I,\mathbf{g}_p} \ge s_{I,\ell}\})$, where $1\{\cdot\}$ is an indicator function and $s_{I,1} \le \cdots \le s_{I,L}$ are prespecified thresholds. The use of $Q_1^{(\ell)}$ is similar to various marginal screening methods (Bair et al. 2006; Fan and Lv 2008; Fan and Fan 2008). By tuning the value of $s_{I,\ell}$, we can screen out "uninformative" features at different scales.

To capture the spatial features of $\mathbf{x}$, we may construct a spatial similarity matrix $Q_2^{(\ell)} = (|w_{E,\mathbf{g}_k\mathbf{g}_j}| 1\{|w_{E,\mathbf{g}_k\mathbf{g}_j}| \ge s_{E,\ell;1}, D(\mathbf{g}_k, \mathbf{g}_j) \le s_{E,\ell;2}\})$, where $\mathbf{s}_{E,\ell} = (s_{E,\ell;1}, s_{E,\ell;2})^T$ and $D(\mathbf{g}_k, \mathbf{g}_j)$

is a specific distance (e.g., Euclidean) between $\mathbf{g}_k$ and $\mathbf{g}_j$. The value of $s_{E,\ell;2}$ controls the number of locations in $\{\mathbf{g}_j \in \mathcal{G} : D(\mathbf{g}_k, \mathbf{g}_j) \le s_{E,\ell;2}\}$, which is a patch set at $\mathbf{g}_k$ (Taylor and Meyer 2012), whereas $s_{E,\ell;1}$ is used to shrink small $|w_{E,\mathbf{g}_k\mathbf{g}_j}|$ to zero.

Given $Q_1^{(\ell)}$ and $Q_2^{(\ell)}$, we may set $Q^{(\ell)}$ as either $Q_1^{(\ell)}Q_2^{(\ell)}$ or $Q_2^{(\ell)}Q_1^{(\ell)}$. Specifically, $Q^{(\ell)} = Q_1^{(\ell)}Q_2^{(\ell)}$ corresponds to selecting important features from $\mathbf{x}$ first and then smoothing those selected features. In contrast, $Q^{(\ell)} = Q_2^{(\ell)}Q_1^{(\ell)}$ corresponds to smoothing $\mathbf{x}$ first and then extracting important features from the smoothed $\mathbf{x}$. According to our experiences, $Q_2^{(\ell)}Q_1^{(\ell)}$ outperforms $Q_1^{(\ell)}Q_2^{(\ell)}$ in terms of prediction accuracy in many scenarios, even though the use of $Q_2^{(\ell)}Q_1^{(\ell)}$ can be computationally demanding when p is extremely large.

Given $Q^{(\ell)}$, we can "prewhiten" $(\mathbf{X} - \mathbf{1}_n\boldsymbol{\mu}^T)$ and calculate $\widetilde{\mathbf{X}}^{(\ell)}$ and its singular value decomposition (SVD) $(U^{(\ell)}, D^{(\ell)}, V^{(\ell)})$ in (1). In practice, a simple criterion for determining K is to include all components up to a prefixed proportion of the total variance, say 85%. For high-dimensional data, we consider a regularized PCA by iteratively solving a single-factor two-way regularized matrix factorization. Specifically, for a given K, we minimize with respect to $(U^{(\ell)}, D^{(\ell)}, V^{(\ell)})$ the following objective function given by

$$\left\|\widetilde{\mathbf{X}}^{(\ell)} - \sum_{k=1}^{K} d_k^{(\ell)}\mathbf{u}_k^{(\ell)}\mathbf{v}_k^{(\ell)T}\right\|^2 + \lambda_{\mathbf{u}}\sum_{k=1}^{K} P_1\left(d_k^{(\ell)}\mathbf{u}_k^{(\ell)}\right) + \lambda_{\mathbf{v}}\sum_{k=1}^{K} P_2\left(d_k^{(\ell)}\mathbf{v}_k^{(\ell)}\right) \tag{3}$$

subject to $\mathbf{u}_k^{(\ell)T}\mathbf{u}_k^{(\ell)} \le 1$ and $\mathbf{v}_k^{(\ell)T}\mathbf{v}_k^{(\ell)} \le 1$ for all k, where $\lambda_{\mathbf{v}}$ and $\lambda_{\mathbf{u}}$ are two tuning parameters and $P_1(\cdot)$ and $P_2(\cdot)$ are two penalty functions. We use adaptive Lasso penalties for $P_1(\cdot)$ and $P_2(\cdot)$ and then iteratively solve (3) (Aharon, Elad, and Bruckstein 2006). For each k_0, we use the sparse method in Lee et al. (2010) to estimate $(d_{k_0}^{(\ell)}, \mathbf{u}_{k_0}^{(\ell)}, \mathbf{v}_{k_0}^{(\ell)})$. In this way, we can sequentially compute $(d_k^{(\ell)}, \mathbf{u}_k^{(\ell)}, \mathbf{v}_k^{(\ell)})$ for $k = 1, \ldots, K$.

In Stage 3, based on $\{(\mathbf{y}_i, \mathbf{u}_{C,i}, \mathbf{z}_i)\}_{i\ge1}$, we use an estimation method to estimate $\boldsymbol{\theta}$ as follows:

$$\widehat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\{\rho(R, \boldsymbol{\theta}, \{(\mathbf{y}_i, \mathbf{u}_{C,i}, \mathbf{z}_i)\}_{i\ge1}) + \lambda P_3(\boldsymbol{\theta})\}, \tag{4}$$

where $\rho(\ldots)$ is a loss function, λ is a tuning parameter and $P_3(\cdot)$ is a penalty function. Given test vectors $\mathbf{x}^*$ and $\mathbf{z}^*$, we can do prediction as follows:

- Calculate $\mathbf{u}_C^* = (u^{(1)*}, \ldots, u^{(L)*})^T$ by setting $u^{(\ell)*} = (\mathbf{x}^* - \boldsymbol{\mu})^T Q^{(\ell)} V^{(\ell)}\{D^{(\ell)}\}^{-1}$, in which $\boldsymbol{\mu}$, $Q^{(\ell)}$, $V^{(\ell)}$, and $D^{(\ell)}$ are learned from the training data.
- Optimize an objective function based on $R(\mathbf{y}; \mathbf{u}_C^*, \mathbf{z}^*, \widehat{\boldsymbol{\theta}})$ to calculate an estimate of $\mathbf{y}$.

2.4. Importance Score Weights and Spatial Weights

There are two sets of weights in MWPCR, including (i) importance score weights enabling a selective treatment for individual features and (ii) spatial weights accommodating the underlying spatial dependence among features across neighboring locations. As shown in simulation studies, the use of the two sets of weights can dramatically improve prediction accuracy. Below, we propose several specific strategies to determine them.

2.4.1. Importance Score Weights

As discussed in Section 2.3, at each location $\mathbf{g}$, $w_{I,\mathbf{g}}$ is calculated based on a statistical model between $(\mathbf{x}_g, \mathbf{z})$ and $\mathbf{y}$ to perform feature selection according to each feature's discriminative importance. Statistically, most existing methods (Bair et al. 2006; Li, Zhong, and Zhu 2012) use a marginal model by assuming

$$f(\mathbf{x}_i, \mathbf{y}_i, \mathbf{z}_i) = \prod_{\mathbf{g}\in\mathcal{G}} f(\mathbf{x}_{i,\mathbf{g}}, \mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\beta}(\mathbf{g})), \tag{5}$$

where $\boldsymbol{\beta} = (\boldsymbol{\beta}(\mathbf{g}) : \mathbf{g} \in \mathcal{G})$ and $\boldsymbol{\beta}(\mathbf{g})$ is introduced to quantify the association between $\mathbf{y}_i$ and $\mathbf{x}_{i,g}$ at each location $\mathbf{g} \in \mathcal{G}$. At the $\mathbf{g}$th location, $w_{I,\mathbf{g}}$ is a statistic based on the marginal model $\prod_{i=1}^n f(\mathbf{x}_{i,\mathbf{g}}, \mathbf{y}_i, \mathbf{z}_i; \boldsymbol{\beta}(\mathbf{g}))$. A simple example is to use the Pearson correlation between each feature and class label as the importance score weight. Noninformative features (e.g., correlation less than a given threshold) can be simply discarded by setting $w_{I,\mathbf{g}} = 0$. However, those $w_{I,\mathbf{g}}$'s largely ignore complex spatial structure, such as homogenous patches defined below, across all locations (Bair et al. 2006; Li, Zhong, and Zhu 2012).

It is common to assume that $\boldsymbol{\beta}(\mathbf{g})$ across all locations are naturally clustered into G homogenous patches, denoted by $\{\mathcal{G}_j : j = 1, \ldots, G\}$, such that

$$G << p,\ \mathcal{G} = \cup_{j=1}^G \mathcal{G}_j,\ \text{and}\ \boldsymbol{\beta}(\mathbf{g})\ \text{varies smoothly in each}\ \mathcal{G}_j. \tag{6}$$

Note that a patch $\mathcal{G}_j$ consists of a set of locations that are spatially connected through edges in $\mathcal{S}$. It has been shown that algorithms based on patch information have led to state-of-the art techniques for classification and denoising (Polzehl and Spokoiny 2006; Li et al. 2011; Taylor and Meyer 2012; Arias-Castro, Salmon, and Willett 2012).

We propose two strategies to learn the homogenous patches $\mathcal{G}_j$ in (6) by jointly modeling $(\mathbf{x}_i, \mathbf{z}_i)$ and $\mathbf{y}_i$. The first strategy is to model the conditional distribution of $\mathbf{x}_i$ given $\mathbf{y}_i$ and $\mathbf{z}_i$, denoted by $f(\mathbf{x}_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$. The second strategy is to model the conditional distribution of $\mathbf{y}_i$ given $\mathbf{x}_i$ and $\mathbf{z}_i$, denoted by $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta})$. Finally, we can learn patches $\mathcal{G}_j$ from the estimated $\boldsymbol{\beta}$ and then construct importance score weights.

The first strategy is to model $f(\mathbf{x}_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$. Let $\mathcal{S}_g(h)$ be an edge set at scale h at each location $\mathbf{g}$. We consider a sequence of nested edge sets across multiscales h_s such that $h_0 = 0 \le h_1 \le \cdots \le h_S$ and $\mathcal{S}_g(h_0) = \{\mathbf{g}\} \subset \cdots \subset \mathcal{S}_g(h_S)$. To learn the homogenous patches, a general framework of multiscale adaptive regression model (MARM) developed in Li et al. (2011) is to maximize a sequence of weighted functions as follows:

$$\hat{\boldsymbol{\beta}}(\mathbf{g}; h_s) = \underset{\boldsymbol{\beta}(\mathbf{g})}{\text{argmax}} \sum_{i=1}^{n} \sum_{\mathbf{g}'\in\mathcal{S}_g(h_s)} \omega(\mathbf{g}, \mathbf{g}'; h_s) \times \log f(\mathbf{x}_{i,\mathbf{g}'}|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\beta}(\mathbf{g}))\ \text{ for }\ s = 1, \ldots, S, \tag{7}$$

where $\omega(\mathbf{g}, \mathbf{g}'; h)$ characterizes the similarity between the observations at $\mathbf{g}'$ and those at $\mathbf{g}$ with $\omega(\mathbf{g}, \mathbf{g}; h) = 1$. If $\omega(\mathbf{g}, \mathbf{g}'; h) \approx 0$, then the observations at $\mathbf{g}'$ do not provide information on $\boldsymbol{\beta}(\mathbf{g})$. Therefore, $\omega(\mathbf{g}, \mathbf{g}'; h)$ can prevent incorporation of locations, whose observations do not contain information on $\boldsymbol{\beta}(\mathbf{g})$ and preserve the edges of homogeneous regions.

Let $D_1(\mathbf{g}, \mathbf{g}')$ and $D_2(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_{s-1}), \hat{\boldsymbol{\beta}}(\mathbf{g}'; h_{s-1}))$ be, respectively, the spatial distance between locations $\mathbf{g}$ and $\mathbf{g}'$ and a similarity measure between $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_{s-1})$ and $\hat{\boldsymbol{\beta}}(\mathbf{g}'; h_{s-1})$. The

$\omega(\mathbf{g}, \mathbf{g}'; h_s)$ can be defined as

$$\omega(\mathbf{g}, \mathbf{g}'; h_s) = K_1\left(D_1(\mathbf{g}, \mathbf{g}')/h_s\right) \cdot K_2\left(D_2\left(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_{s-1}), \hat{\boldsymbol{\beta}}(\mathbf{g}'; h_{s-1})\right)/\gamma_n\right), \quad (8)$$

where $K_1(\cdot)$ and $K_2(\cdot)$ are two nonnegative kernel functions and γ_n is a bandwidth parameter that may depend on n. The weights $K_1(D_1(\mathbf{g}, \mathbf{g}')/h_S)$ give less weight to location $\mathbf{g}' \in \mathcal{S}_g(h_S)$, which is far from the location $\mathbf{g}$. The weights $K_2(u)$ downweight location $\mathbf{g}'$ with large $D_2(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S), \hat{\boldsymbol{\beta}}(\mathbf{g}'; h_S))$, which indicates a large difference between $\hat{\boldsymbol{\beta}}(\mathbf{g}'; h_S)$ and $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S)$. Moreover, by following Li et al. (2011) and Polzehl and Spokoiny (2006), we set $K_1(x) = (1-x)_+$ and $K_2(x) = \exp(-x)$. See the detailed algorithm of MARM in Li et al. (2011).

The second strategy is to model $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta})$ and the prior distribution of $\boldsymbol{\beta}$, given by $f(\boldsymbol{\beta})$. Since $\mathbf{x}_i$ is often high dimensional, it is much difficult to carry out statistical inference based on $f(\mathbf{y}_i|\mathbf{x}_i, \mathbf{z}_i, \boldsymbol{\beta})$ compared with $f(\mathbf{x}_i|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\beta})$. Moreover, our primary goal is to perform feature selection to eventually use a small subset of $\mathbf{x}_i$ to predict $\mathbf{y}_i$, while correcting for $\mathbf{z}_i$. Similar to the first strategy, we also take the marginal method and then incorporate a specific structure to estimate $\boldsymbol{\beta}$ as follows:

$$\prod_{\mathbf{g}\in\mathcal{G}}\prod_{i=1}^{n} f\left(\mathbf{y}_i|\mathbf{x}_{i,g}, \mathbf{z}_i; \boldsymbol{\beta}(\mathbf{g})\right)\left\{\prod_{\mathbf{g}\in\mathcal{G}} f\left(\boldsymbol{\beta}(\mathbf{g})|\boldsymbol{\beta}(\mathbf{g}') : \mathbf{g}' \in \mathcal{N}_g\right)\right\}, \quad (9)$$

where $\mathcal{N}_g$ is a set of the neighboring locations of location $\mathbf{g}$.

Similar to the first strategy, we propose an adaptive smoothing algorithm to estimate $\boldsymbol{\beta}$ as follows. Consider a sequence of nested edge sets $\mathcal{S}_g(h_0) = \{\mathbf{g}\} \subset \cdots \subset \mathcal{S}_g(h_S)$ for $h_0 = 0 \le h_1 \le \cdots \le h_S$.

- Step (i). Calculate $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0)$ and $\mathrm{cov}(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0))$ according to $\prod_{i=1}^{n} f(\mathbf{y}_i|\mathbf{x}_{i,g}, \mathbf{z}_i; \boldsymbol{\beta}(\mathbf{g}))$ across all locations $\mathbf{g}$.
- Step (ii). Smooth $\{\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0) : \mathbf{g} \in \mathcal{G}\}$ to sequentially estimate $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_s)$ for $s = 1, \ldots, S$ across all $\mathbf{g} \in \mathcal{G}$. Candidate methods include local polynomial, nonlocal mean, and propagation-separation, among others (Polzehl and Spokoiny 2006; Arias-Castro, Salmon, and Willett 2012).

For both strategies, after the iteration h_S, we can obtain $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S)$ and its covariance matrix, denoted by $\mathrm{cov}(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S))$, across all $\mathbf{g} \in \mathcal{G}$. Finally, we calculate $w_{I,\mathbf{g}}$ as a function of $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S)$ and $\mathrm{cov}(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S))$, such as the Wald test and its p-value. Then, we use a clustering algorithm, such as the K-mean algorithm, to group $\{\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S) : \mathbf{g} \in \mathcal{G}\}$ into several homogenous clusters (Hastie, Tibshirani, and Friedman 2009), in which $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S)$ varies very smoothly in each cluster.

2.4.2. Spatial Weights

As discussed in Section 2.3, $w_{E,\mathbf{gg}'}$ often characterizes the degree of certain "similarity" between locations $\mathbf{g}$ and $\mathbf{g}'$. We consider three spatial weight matrices, including (i) the precision matrix, (ii) a locally spatial weight matrix, and (iii) a cluster-based spatial weight matrix as follows.

For the precision matrix, let Σ be the covariance matrix of $\mathbf{x}_i$, we can set $Q_2^{(\ell)} = \Sigma^{-1/2}$; thus, $Q_2^{(\ell)}Q_2^{(\ell)T} = \Sigma^{-1}$ is the precision matrix of $\mathbf{x}_i$. When Σ^{-1} has certain sparsity structures (e.g., factor model), various estimation methods have been developed even for extremely large p.

The locally spatial weight matrix consists of nonnegative weights assigned to the spatial neighboring locations of each location. Specifically, we set $w_{E,\mathbf{gg}'}$ as

$$w_{E,\mathbf{gg}'} = \frac{\omega(\mathbf{g}, \mathbf{g}'; h_S)\mathbf{1}\{\mathbf{g}' \in \mathcal{S}_g(h_S)\}}{\sum_{\mathbf{g}''\in\mathcal{S}_g(h_S)} \omega(\mathbf{g}, \mathbf{g}''; h_S)1\{\mathbf{g}'' \in \mathcal{S}_g(h_S)\}}, \quad (10)$$

in which $\omega(\mathbf{g}, \mathbf{g}'; h_S)$ is defined in (8). Thus, we have $w_{E,\mathbf{gg}'} = 0$ for all $\mathbf{g}' \notin \mathcal{S}_g(h_S)$ and $\sum_{\mathbf{g}'\in\mathcal{G}} w_{E,\mathbf{gg}'} = 1$.

The cluster-based spatial weight matrix consists of nonnegative weights assigned to locations in the same homogenous cluster. Specifically, we use the Laplace–Beltrami operator to construct W_E (Luxburg 2007). It is assumed that each edge between two locations $\mathbf{g}$ and $\mathbf{g}'$ carries a nonnegative weight $w_{\mathbf{gg}'}$. Thus, matrix $W = (w_{\mathbf{gg}'})$ is a weighted adjacency matrix of $\mathcal{G}$. The degree of a location $\mathbf{g} \in \mathcal{G}$ is defined as $d_{\mathbf{g}} = \sum_{\mathbf{g}'\in\mathcal{G}} w_{\mathbf{gg}'}$ and the degree matrix W_D is given by $W_D = \mathrm{diag}(d_{g_1}, \ldots, d_{g_p})$. The unnormalized Laplacian matrix L of the graph $\mathcal{G}$ is defined as $W_L = W_D - W$, which can be regarded as a discrete representation of the Laplace–Beltrami operator. Finally, we set $W_E = \exp(-0.5\ W_L/\gamma)$, where $\exp(\cdot)$ denotes the matrix exponential. In practice, when p is extremely large, it is computationally infeasible to directly use the huge $p \times p$ matrix W_E. In this case, based on the clustering results in (6), we only consider locations in each cluster and each cluster forms a connected subgraph, which leads to dramatically computational savings (Cuingnet et al. 2013).

2.4.3. Weights Selection

A critical question is how to select spatial weights and/or importance score weights for constructing $Q^{(\ell)}$ in different applications. Ideally, we may either use one of them or combine some of them together to construct $Q^{(\ell)}$. Theoretically, we have investigated the effects of applying importance score weights and different spatial weights in MWPCR on classification accuracy for high-dimensional binary classification and put them in the supplementary document. We have three key theoretical results as follows.

- The use of feature selection can substantially improve classification accuracy for high-dimensional binary classification.
- The use of spatial kernel weights and importance score weights in MWPCR can substantially improve classification accuracy even when signals are weak.
- The use of the true $\Sigma^{-1/2}$ can improve classification accuracy, where Σ is the covariance matrix of $\mathbf{x}$.

Based on these results, we suggest to first apply the locally spatial weight matrix (or the cluster-based spatial weight matrix) and then use the importance score weights based on $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S)$. Although the use of $\Sigma^{-1/2}$ can improve classification accuracy,

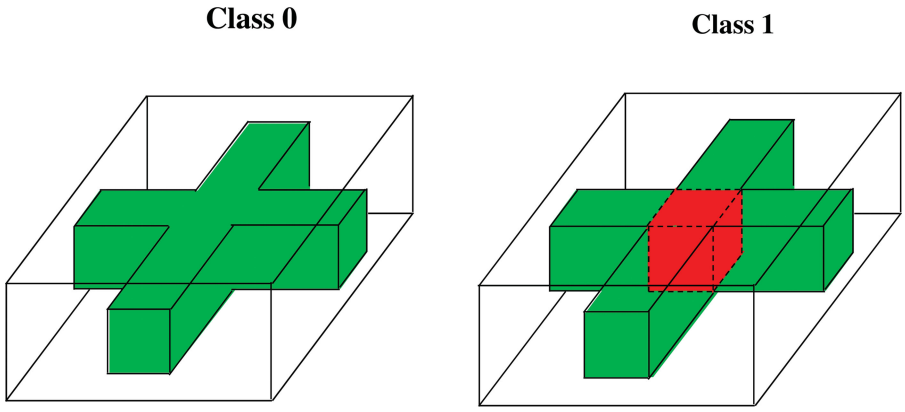

Figure 2. True mean images for the first set of simulations: Class 0 in the left panel and Class 1 in the right panel. The white, green, and red colors, respectively, correspond to 0, 1, and 2.

estimating $\Sigma^{-1/2}$ can be very challenging when p is even moderate. Thus, we avoid estimating $\Sigma^{-1/2}$ in all simulations and real data analysis.

3. Simulation Studies: Binary Outcome

We use two sets of simulation studies, including binary and continuous outcomes, to examine the finite sample performance of MWPCR under different scenarios. We demonstrate that MWPCR outperforms or at least is compatible with many state-of-the-art methods. For the sake of space, we include all simulation results for continuous outcome in the supplementary document.

We applied MWPCR to a high-dimensional binary classification problem as follows. We simulated $20 \times 20 \times 10$ 3D-images from a linear model given by

$$\mathbf{x}_{i,\mathbf{g}} = \beta_0(\mathbf{g}) + \beta_1(\mathbf{g})l_i + \epsilon_i(\mathbf{g}) \quad \text{for} \quad i = 1, \ldots, n, \qquad (11)$$

where l_i is the class label coded as either 0 or 1 and $\epsilon_i(\mathbf{g})$ are random variables with zero mean. Figure 2 presents the true mean images of class $l_i = 0$ and class $l_i = 1$, in which a red cuboid $3 \times 3 \times 4$ region characterizes the maximum difference 1 between classes 0 and 1. In this case, we have $p = 4000$. Then, we set $n = 100$ with 60 images from Class 0 and the rest from Class 1.

We consider three types of noise $\epsilon_i(\mathbf{g})$ in (11). First, $\epsilon_i^{(1)}(\mathbf{g})$ were independently generated from an $N(0, 2^2)$ generator across all voxels. Second, $\epsilon_i^{(2)}(\mathbf{g}) = \sum_{\|\mathbf{g}'-\mathbf{g}\|_1 \le 1} \epsilon_i^{(1)}(\mathbf{g}')/m_{\mathbf{g}}$ were generated from $\epsilon_i^{(1)}(\mathbf{g})$ by introducing the short range spatial correlation, where $\|\cdot\|_1$ is the L_1 norm of a vector and $m_{\mathbf{g}}$ is the number of locations in the set $\{\| \mathbf{g}' - \mathbf{g} \|_1 \le 1\}$. Third, to introduce the long range spatial correlation, $\epsilon_i^{(3)}(\mathbf{g})$ were generated according to $\epsilon_i^{(3)}(\mathbf{g}) = 2\sin(\pi g_1/10)\xi_{i,1} + 2\cos(\pi g_2/10)\xi_{i,2} + 2\sin(\pi g_3/5)\xi_{i,3} + \epsilon_i^1(\mathbf{g})$, where $\mathbf{g} = (g_1, g_2, g_3)^T$ and $\xi_{i,k}$ for $k = 1, 2, 3$ were independently generated from an $N(0, 1)$ generator. Moreover, the noise variances in all voxels of the red cuboid region equal 4, 4/6, and $4\{\sin(\pi g_1/10)^2 + \cos(\pi g_2/10)^2 + \sin(\pi g_3/5)^2\} + 4$ for Type I, II, and III noises, respectively. Therefore, among the three types of noise, Type III noise has the smallest signal-to-noise ratio and Type II noise has the largest one.

We ran the three stages of MWPCR as follows. In Stage 1, let $\{h_s = 1.2^s,\ s = 0, 1, \ldots, S = 5\}$, and for each $\mathbf{g} \in \mathcal{G}$, we set $\omega_{I,\mathbf{g}} = -p\log(p(\mathbf{g}))/\{-\sum_{\mathbf{g}\in\mathcal{G}}\log(p(\mathbf{g}))\}$, where $p(\mathbf{g})$ is the p-value of Wald test $\beta_1(\mathbf{g}) = 0$ in (11) at voxel $\mathbf{g}$. The spatial weight W_E is given by (10). We set the spatial weight W_E according to (10) and (8). Specifically, we considered three types of spatial weights W_E, including MWPCR1: only the location kernel function $K_1(.)$ in (8); MWPCR2: only the similarity kernel function $K_2(.)$ in (8); and MWPCR3: the combination of kernel functions $K_1(.)$ and $K_2(.)$ in (8). Then, we selected the bandwidth $\{h_s = 1.2^s,\ s = 0, \ldots, S = 5\}$ in these kernel functions to determine W_I and W_E. In Stage 2, we used different numbers of principal components in MWPCA to reconstruct the low-dimensional representation of simulated images. In Stage 3, we tried different classification methods, including linear regression, k-nearest neighbor (k-NN), and support vector machine (SVM), on these low-dimensional representations. Since their performances are similar to each other, we only report the results based on the linear regression throughout the article. The linear regression uses class label l_i as dependent variable and principal components as explanatory variables. An image is classified as Class 0, if its predictive value is less than 0, and as Class 1, otherwise.

We first used the leave-one-out cross-validation to calculate the misclassification rates for MWPCR1, MWPCR2, MWPCR3, and a standard principal component analysis (PCA). Table 1 presents the classification results based on 5, 7, and 10 principal components. The misclassification errors for all MWPCR methods are quite stable for different numbers of principal components under different types of noise. All MWPCR methods

Table 1. Classification results for the first set of simulations: misclassification rates for MWPCR1, MWPCR2, MWPCR3, and PCA based on three numbers of principal components under three types of noise.

Noise	Number of PCs	PCA	MWPCR1	MWPCR2	MWPCR3
Type I	5	0.47	0.11	0.09	0.10
	7	0.48	0.13	0.11	0.10
	10	0.49	0.13	0.11	0.10
Type II	5	0.41	0.04	0.08	0.03
	7	0.39	0.03	0.09	0.04
	10	0.42	0.03	0.07	0.04
Type III	5	0.27	0.13	0.10	0.09
	7	0.26	0.13	0.10	0.10
	10	0.28	0.13	0.10	0.10

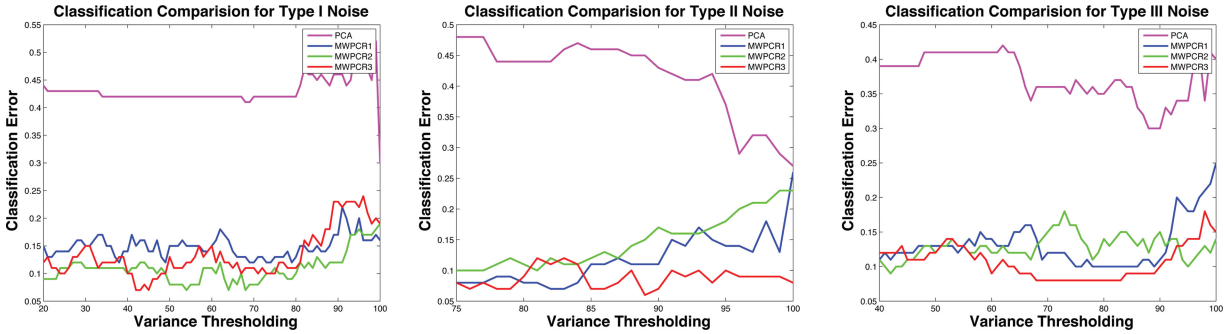

Figure 3. Classification results for the first set of simulations: classification rate curves for MWPCR1, MWPCR2, MWPCR3, and PCA based on the variance thresholding method for the three types of noise. Overall classification errors for MWPCR3 (red curve) are smaller than those of others, confirming the good performance of MWPCR3. Also MWPCR3 is quite robust to different variance thresholds. The performance of PCA is very poor and its classification error (magenta curve) is much larger than all MWPCR methods for the three types of noises.

perform relatively well for Type II noise compared with Type I and III noises, since Type II noise has the largest signal-to-noise ratio. Moreover, MWPCR3 is slightly better than MWPCR1 and MWPCR2, which may be because MWPCR3 combines both the local smooth and similarity kernels. Moreover, it seems that MWPCR3 is very robust to the long-range correlation structure of Type III noise. Compared with all MWPCR methods, PCA performs very poor, since it does not incorporate the class label information.

Second, we used the same variance thresholding to compare the three MWPCR methods with PCA. Figure 3 shows that the classification error (magenta curve) for PCA is much larger than that for all other methods. For each fixed variance threshold, the number of extracted principal components from PCA is less than that of MWPCR1, MWPCR2, and MWPCR3. Overall, MWPCR3 outperforms all other methods for all three types of noises. The variance threshold in the middle panel of Figure 3 starts from 70%, since the first principal component of PCA almost accounts for 70% of the total variance for Type II noise.

Third, we compared MWPCR3, in which five principal components were used, with eight other state-of-the-art classification methods. These eight classification methods include sparse discriminant analysis (sLDA) (Clemmensen et al. 2011), sparse partial least squares (SPLS) analysis (Chun and Keles 2010), sparse logistic regression (SLR) (Yamashita 2011), support vector machine (SVM) (Chang and Lin 2011), regularized optimal affine discriminant (ROAD) (Fan, Feng, and Tong 2012), wavelet-based multicscale PCA (WMSPCA)(Bakshi 1998), the combination of sure independence screening (SIS) (Fan et al. 2010) and principal component analysis (PCA) (SIS + PCA), and graph-constrained elastic-net (GraphNet) (Grosenick et al. 2013). We chose these classification methods due to their excellent performance in various simulated and real datasets.

Fourth, for all classification methods, we first calculated their misclassification rates by using the leave-one-out cross-validation and then generated the receiver operating characteristics (ROC) curves of all nine methods. For ROC, we used model (11) to independently generate a testing set with the same sample size and the same proportion of Class 0 to Class 1 as the training set. For each method, we applied 10-fold cross-validation to the training set to select the tuning parameter(s) and build the model based on the training set. Then, we applied the fitted model to the testing set to generate the ROC curves of all nine classification methods in Figure 4. Based on these ROC curves, we calculated their area under curve (AUC) values (Fawcett 2006).

Table 2 presents the classification results, including both misclassification rates and AUC values. Table 2 reveals that MWPCR outperforms all other classification methods, especially when the signal-to-noise ratio is low for Type I and II noises. Except WMSPCA, SIS + PCA, and MWPCR, all other classification methods are also sensitive to the presence of the long-range correlation structure in Type III noise. However, if high-dimensional features do not have strong spatial structures, then it is expected that MWPCR may perform worse than other competing classification methods.

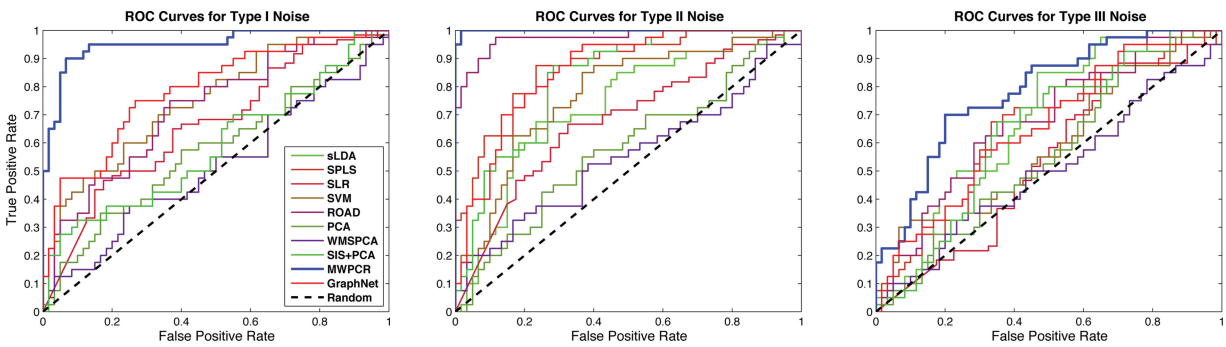

Figure 4. ROC curves of different classification methods for the three types of noise in the first set of simulations. The blue curves correspond to MWPCR and have the highest AUC value.

Table 2. Misclassification rates (MRs) and AUC values for the first set of simulations: comparison between MWPCR and eight other state-of-the-art classification methods.

Type/Measure	sLDA	SPLS	SLR	SVM	ROAD	WMSPCA	SIS + PCA	GraphNet	MWPCR
I/MR	0.28	0.43	0.45	0.38	0.36	0.20	0.33	0.32	0.10
II/MR	0.27	0.08	0.18	0.26	0.08	0.13	0.46	0.22	0.03
III/MR	0.52	0.30	0.61	0.60	0.50	0.21	0.10	0.35	0.09
I/AUC	0.59	0.59	0.66	0.75	0.72	0.52	0.59	0.79	0.95
II/AUC	0.73	0.87	0.68	0.77	0.97	0.55	0.83	0.86	0.99
III/AUC	0.68	0.65	0.54	0.59	0.68	0.50	0.64	0.66	0.78

NOTES: sLDA denotes sparse discriminant analysis; SPLS denotes sparse partial least squares; SLR denotes sparse logistic regression; SVM denotes support vector machine; ROAD denotes regularized optimal affine discriminant; WMSPCA denotes wavelet-based multiscale principal component analysis; SIS + PCA combines sure independence screening (SIS) and principal component analysis; and GraphNet denotes graph-constrained elastic-net.

4. Real Data Analysis

4.1. ADNI PET Data

Alzheimer's disease (AD) is the most common form of dementia and results in the loss of memory, thinking, and language skills. AD is an escalating national epidemic and a genetically complex, progressive, and fatal neurodegenerative disease. The incidence of AD doubles every 5 years after the age of 65 and the number of AD patients has dramatically increased recently, which has caused a heavy socioeconomic burden. AD is the sixth-leading cause of death in the United States, while there is no means to prevent, cure, or even slow its progression.

The development of MWPCR is motivated by using the baseline FDG-PET dataset to address questions (Q1) and (Q2). The ADNI PET dataset downloaded from the ADNI web site (*www.loni.usc.edu/ADNI*) consists of 196 subjects with 102 NCs and 94 AD subjects. There are three subjects, missing the gender and age information. Among all the rest of the subjects, there are 117 males whose mean age is 76.20 years with standard deviation 6.06 years and 76 females whose mean age is 75.29 years with standard deviation 6.29 years. FDG-PET images acquired 30–60 min post-injection were processed by using a standard image processing pipeline. A detailed description of PET protocols and acquisition can be found at *www.adni-info.org*. Such pipeline consists of average, spatially alignment, interpolation to a standard voxel size, intensity normalization, and smoothing to a common resolution of 8-mm full width at half maximum.

4.2. Binary Classification

The first goal is to use MWPCR to classify subjects from ADNI to either AD or NC group based on their FDG-PET images. It is associated with the second primary objective of ADNI aiming at developing new diagnostic methods for AD intervention, prevention, and treatment. We first applied MWPCR3 to ADNI and used the same setting as simulations in Section 3 except that we considered a linear model for $f(\mathbf{x}_{i,\mathbf{g}}|\mathbf{y}_i, \mathbf{z}_i, \boldsymbol{\beta}(\mathbf{g}))$, in which $\mathbf{z}_i$ includes both age and gender and $\mathbf{y}_i$ is diagnosis status (AD versus NC). We also compare MWPCR3 with nine other classification methods, including PCA and the eight state-of-the-art classification methods discussed in Section 3. For the PCA method, we applied PCA with five principal components, which

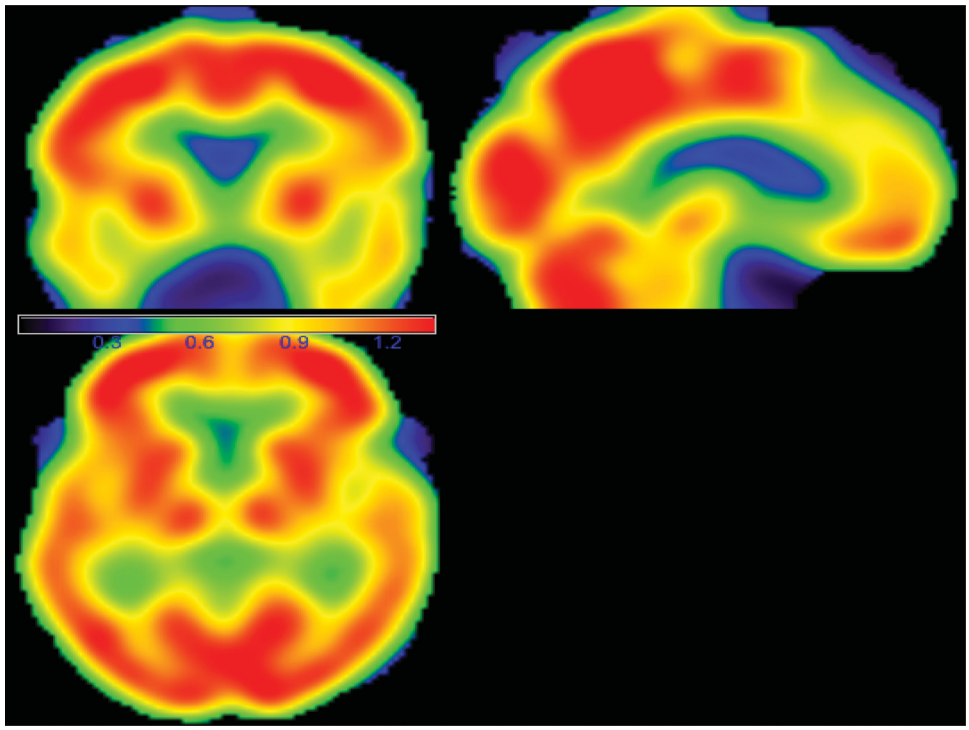

Figure 5. The images of the importance score weight matrix for the ADNI binary classification analysis. The red regions have large weight score values and contain the important classification information.

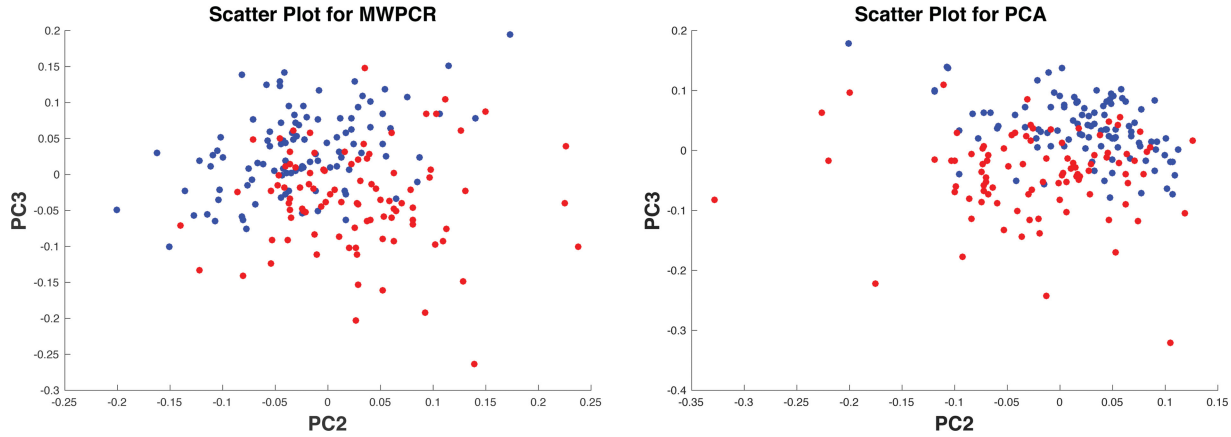

Figure 6. ADNI binary classification results: scatterplots of PC2 and PC3 scores for MWPCR (left panel) and PCA (right panel). Blue and red points in both panels correspond to NC and AD subjects, respectively.

account for around 90% of the total variance, and then used the same linear regression as MWPCR3 to perform classification analysis. Figure 5 presents three selected slices of the weight matrix W_I. The red regions, such as supramarginal gyrus right, correspond to the voxels with large importance score weights and contain the most important information for classification.

Second, for both PCA and MWPCA, we extracted their corresponding first five principal component scores and directions. Figure 6 shows the scatterplot of PC2 and PC3 scores for PCA and that for MWPCA, in which blue and red points correspond to NC and AD subjects, respectively, where PC2 and PC3 represent the second and third principal components, respectively. It seems that compared with PCA, the blue and red points are more separable for MWPCA. Furthermore, Figure 7 presents some selected slides of the principal directions corresponding to PC2 and PC3 for MWPCA. We are able to identify several key regions of interest, such as "supramarginal gyrus," "superior temporal gyrus," and "inferior frontal gyrus." For instance, the superior temporal gyrus is in the temporal lobe of the human brain and contains several important structures of the brain, including Brodmann areas 41, 42, and 22 p. It is probably involved with language perception and processing (Marcus, Mena, and Subramaniam 2014). Moreover, within the brain, the anatomical regions that show the greatest decrease in FDG uptake with aging are the bilateral superior medial frontal, motor, anterior, and middle cingulate, and bilateral parietal cortices. Among them, the superior temporal pole was found to be particularly affected.

Third, similar to Section 3, we calculated the misclassification rates of all classification methods by using the leave-one-out cross-validation and then generated their receiver operating characteristics (ROC) curves. For the ROC analysis, we

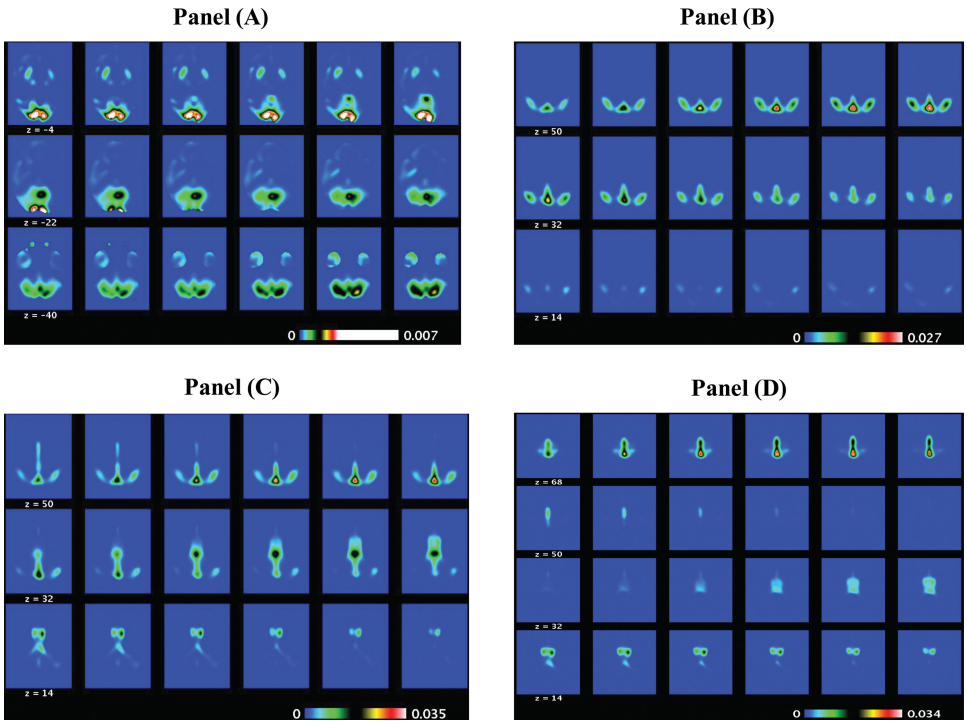

Figure 7. ADNI PET binary classification results: the selected slides of the PC2 direction image (positive elements in panel (A) and negative elements in panel (B)) and those of the PC3 direction image (positive elements in panel (C) and negative elements in panel (D)) obtained from MWPCR.

Table 3. Misclassification rates (MRs) and AUC values of different classification methods for ADNI PET data.

	sLDA	SPLS	SLR	SVM	ROAD	PCA	WMSPCA	SIS + PCA	GraphNet	MWPCR
MR	0.255	0.163	0.179	0.168	0.189	0.194	0.168	0.255	0.128	0.117
AUC	0.845	0.912	0.863	0.912	0.878	0.877	0.873	0.730	0.879	0.913

randomly and proportionally split the dataset into two parts, a training set and a testing set. For each part, the sample sizes are same (98/98). Within each part, the proportion of AD to NC remains the same. For each classification method, we used 10-fold cross-validation on the training set to select the tuning parameter(s) and build the model, and then we applied the fitted model to the testing set to calculate the relative scores. Subsequently, we generated all ROC curves and their AUC values.

Table 3 presents the classification results based on classification error and AUC, while Figure 8 presents the ROC curves of all 10 classification methods. sLDA and SIS + PCA perform much worse than all other methods. In general, SPLS, SVM, and WMSPCA are comparable with each other, but they outperform SLR and ROAD. In terms of misclassification rate, MWPCR outperforms all nine other classification methods. In contrast, in terms of AUC, MWPCR, SPLS, and SVM are compatible with each other. It may indicate that the classification accuracy can be significantly improved by incorporating spatial smoothness and correlation.

4.3. ADAS-Cog Score Prediction

The second goal is to use MWPCR to identify FDG-PET imaging biomarkers observed at baseline to accurately predict the change in the ADAS-Cog test score (or TOTAL_{11}) at least 2 years later after initial assessment. The TOTAL_{11}, which measures the cognitive performance of each subject, was calculated from the 11-item ADAS-Cog, such as Word Recall, whose details can be found in *http://adni.loni.usc.edu/data-samples/data-faq/*. Since three subjects are missing gender and age information and 10 other subjects only have the baseline TOTAL_{11}, we only use 183 subjects in this analysis.

We ran MWPCR as follows. We first fitted a linear model with the TOTAL_{11} score at the latest time point as response and the baseline TOTAL_{11} score, age, gender, time since baseline, and years of education, and then we used the residual obtained from the linear model as the response $\mathbf{y}$ and the FDG-PET image as $\mathbf{x}$. In Stage 1, we fitted a linear model for $f(\mathbf{x}_{i,\mathbf{g}}|\mathbf{y}_i, \boldsymbol{\beta}(\mathbf{g}))$, in which we dropped off $\mathbf{z}_i$. Then, W_I is calculated based on the p-value of Wald test associated with the correlation between $\mathbf{x}_{i,\mathbf{g}}$

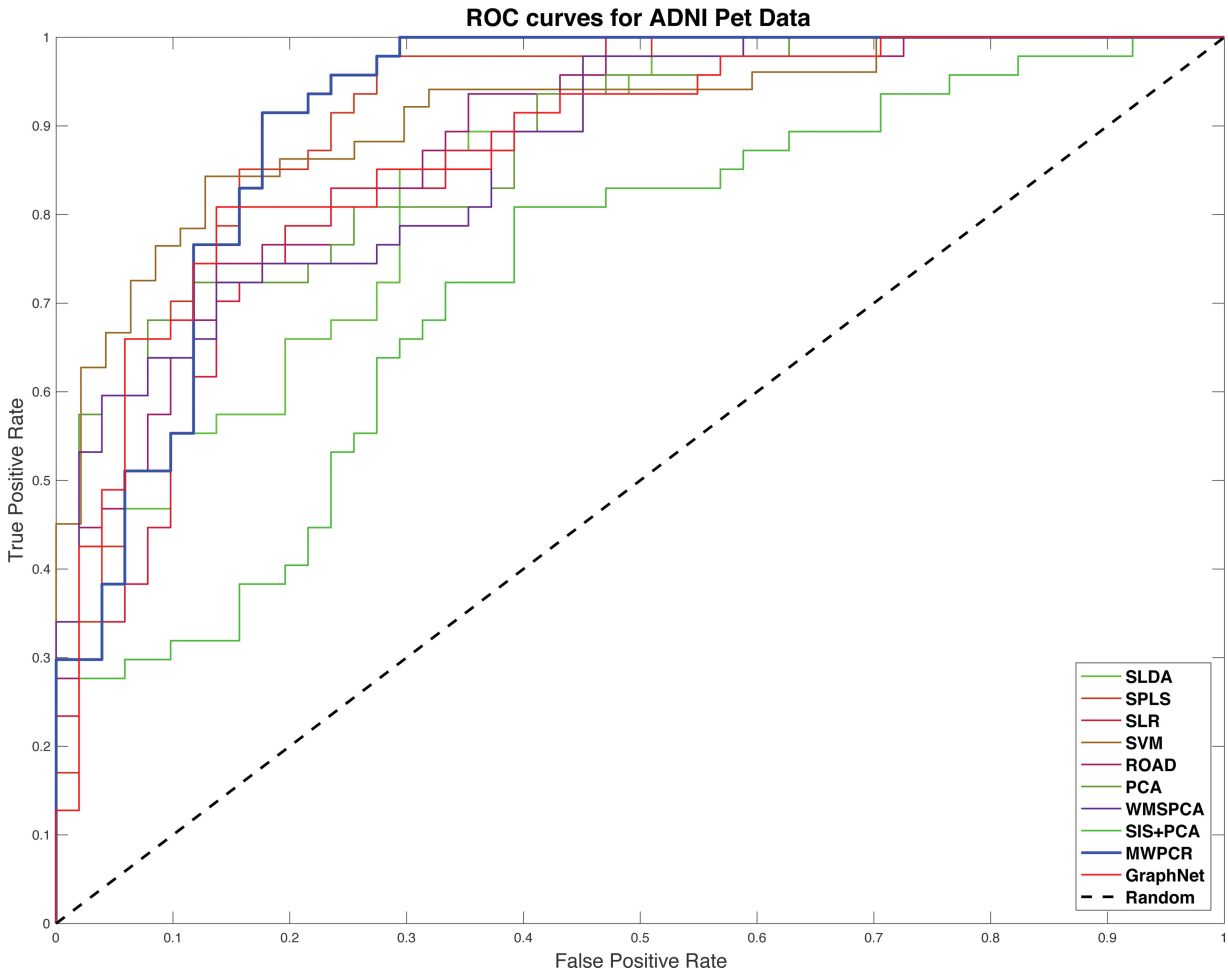

Figure 8. ADNI PET binary classification results: ROC curves of the 10 different classification methods. The blue line corresponds to MWPCR.

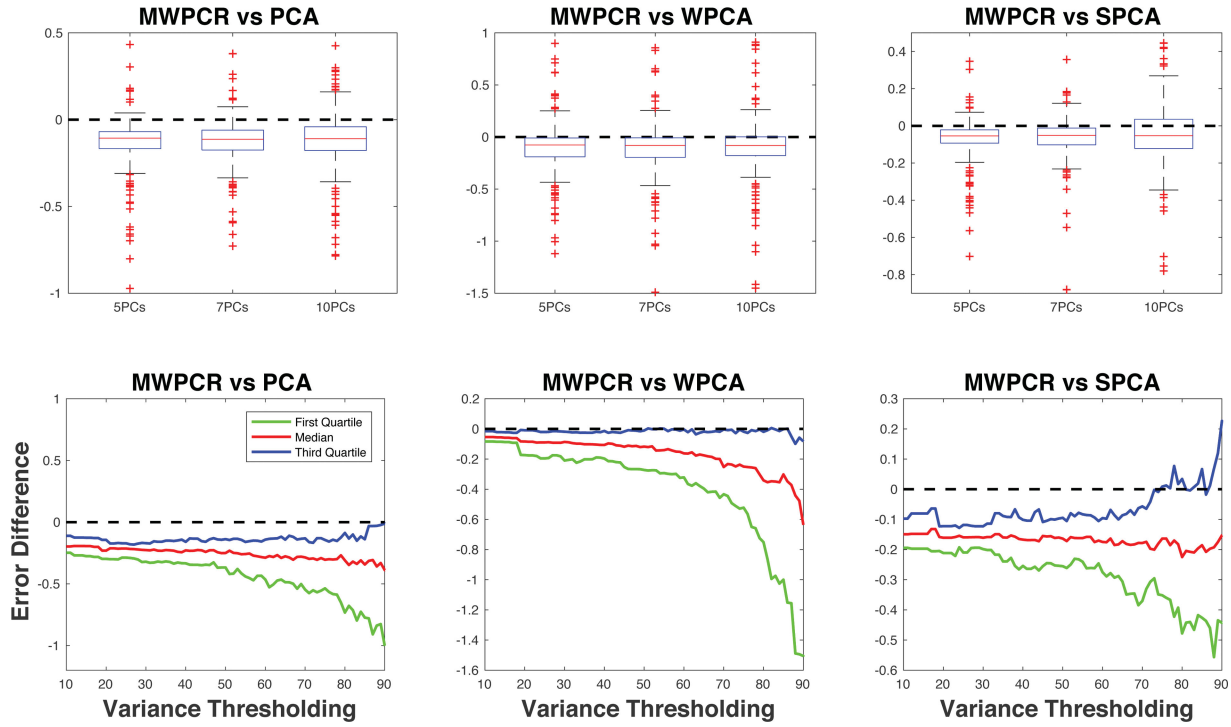

Figure 9. ADAS-Cog score prediction for ADNI PET data: comparison between MWPCR with PCA, WPCA, and SPCA. The panels in the first row show the boxplots of error differences between MWPCR and PCA (WPCA and SPCA) for different numbers of principal components. The panels in the second row show the first, second, and third quantile curves of error differences between MWPCR and PCA (WPCA and SPCA) for different variance thresholds.

and $\mathbf{y}_i$ at each voxel $\mathbf{g}$. In Stage 2, following the simulations in Section 3, we chose MWPCR3 with different numbers of principal components for MWPCA to construct the low-dimensional latent variables $\{\mathbf{u}_{k,i}\}$. In Stage 3, we fitted a linear latent variable regression given by $\mathbf{y}_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k \mathbf{u}_{k,i} + \epsilon_i$ to do prediction.

Second, we compared MWPCR and three other dimensional reduction methods including PCA, weighted PCA (WPCA) (Skocaj, Leonardis, and Bischof 2007), and supervised PCA (SPCA) (Bair et al. 2006). We used the leave-one-out cross-validation method to compute the prediction errors of all methods. Let $\hat{\mathbf{y}}_i$ be the fitted response value based on the linear

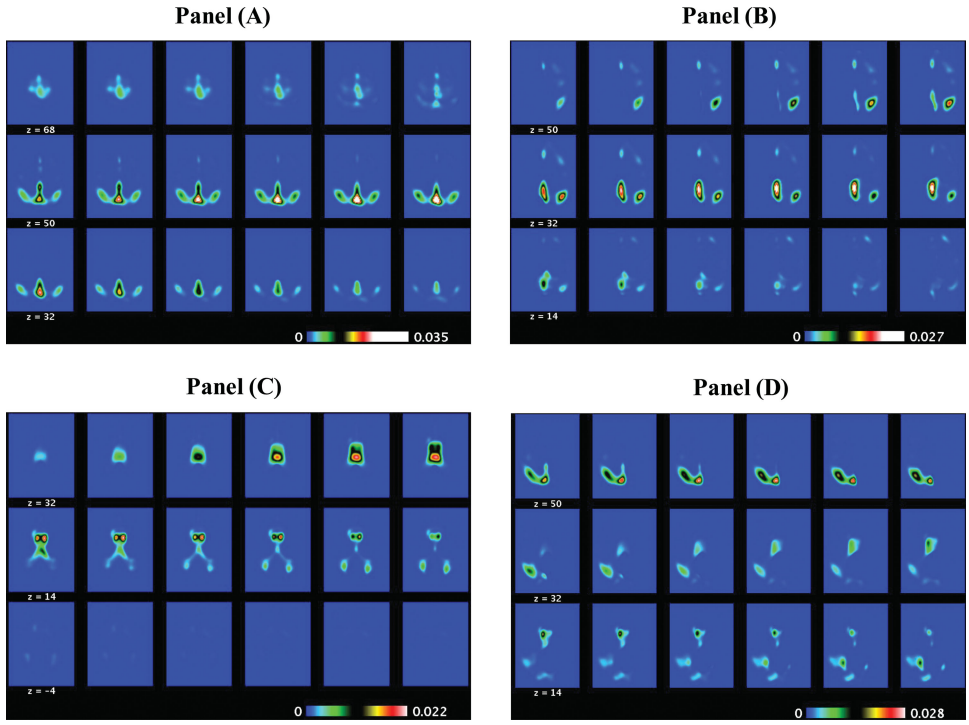

Figure 10. ADAS-Cog score prediction for ADNI PET data: the selected slides of the PC1 direction image (positive elements in panel (A) and negative elements in panel (B)) and those of the PC5 direction image (positive elements in panel (C) and negative elements in panel (D)) obtained from MWPCR.

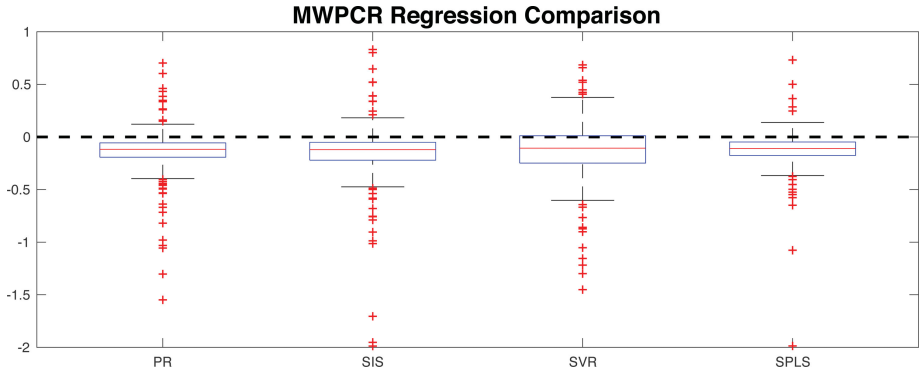

Figure 11. ADAS-Cog score prediction for ADNI PET data: comparison of MWPCR with the four other regression methods, including PR, SIS, SVR, and SPLS.

latent variable regression, we define the prediction error as $|\hat{\mathbf{y}}_i - \mathbf{y}_i|/|\mathbf{y}_i|$. Subsequently, we calculated the prediction error differences between MWPCR and all other three methods and their quantile curves across different numbers of principal components and variance thresholds. Figure 9 presents the comparison results based on the prediction error differences and their quantile curves. Both the error differences and the quantile curves are less than 0 (below the dashed line), confirming the better performance of MWPCR in predicting changes in ADAS-Cog score.

Third, for MWPCA, we extracted their corresponding first five principal component scores and directions. Figure 10 presents some selected slides of the principal directions corresponding to PC1 and PC5 for MWPCA, where PC1 and PC5 represent the first and fifth principal components, respectively. We are able to identify several key regions of interest, such as "right lateral ventricle," "right middle temporal gyrus," "right fornix," and "right middle frontal gyrus." For instance, the fornix is on the medial aspects of the cerebral hemispheres connecting the medial temporal lobes to the hypothalamus. Since the fornix serves a vital role in memory functions, it has become the subject of recent research emphasis in Alzheimer's disease (AD) and mild cognitive impairment (MCI) (Nowrangi and Rosenberg 2015).

Finally, we compare MWPCR with four other high-dimensional regression methods including penalized regression (PR) (Tibshirani 1996), sure independence screening (SIS) regression (Fan and Lv 2008), support vector regression (SVR) (Basak, Pal, and Patranabis 2007), and SPLS (Chun and Keles 2010). Figure 11 shows the boxplots of the prediction error differences between MWPCR and all the other regression methods, indicating that MWPCR outperforms all other regression methods.

5. Discussion

We have developed a general MWPCR framework for the use of high-dimensional data on graph to predict a low-dimensional response. MWPCR enables an efficient and selective treatment of individual features, accommodates the complex dependence among features, and has the ability of using the underlying spatial pattern possessed by image data. MWPCR integrates feature selection, smoothing, and feature extraction in a single framework. In the simulation studies and real data analyses, MWPCR shows substantial improvement over many state-of-the-art methods for high-dimensional problems. Moreover, both theoretically and numerically, we have demonstrated the importance of using both importance score weights and spatial weights in prediction problems.

Supplementary Materials

The theoretical properties are discussed in supplementary material. The MATLAB implementation of the proposed method is provided as well.

Acknowledgments

The authors thank the editor, the anonymous associate editor, and the referees for their suggestions that have led to a much improved article.

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (*adni.loni.usc.edu*). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: *http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf*. The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and nonprofit organizations, as a $60 million, 5-year publicprivate partnership. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). Determination of sensitive and specific markers of very early AD progression is intended to aid researchers and clinicians to develop new treatments and monitor their effectiveness, as well as lessen the time and cost of clinical trials. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many coinvestigators from a broad range of academic institutions and private corporations, and subjects have been recruited from over 50 sites across the U.S. and Canada. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The follow-up duration of each group is specified in the protocols for ADNI-1, ADNI-2, and ADNI-GO. Subjects originally recruited for ADNI-1 and ADNI-GO had the option to be followed in ADNI-2. For up-to-date information, see *www.adni-info.org*.

Funding

Dr. Zhu's work was partially supported by NIH grants MH086633, NSF grants SES-1357666 and DMS-1407655, and a grant from Cancer Prevention Research Institute of Texas. This material was based upon work partially supported by the NSF grant DMS-1127914 to the Statistical and

Applied Mathematical Science Institute. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Aharon, M., Elad, M., and Bruckstein, A. (2006), "K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation," *IEEE Transactions on Signal Processing*, 54, 4311–4322. [1012]

Allen, G. I., Grosenick, L., and Taylor, J. (2014), "A Generalized Least Squares Matrix Decomposition," *Journal of the American Statistical Association*, 109, 145–159. [1010,1011]

Arias-Castro, E., Salmon, J., and Willett, R. (2012), "Oracle Inequalities and Minimax Rates for Nonlocal Means and Related Adaptive Kernel-Based Methods," *SIAM Journal on Imaging Sciences*, 5, 944–992. [1012]

Bair, E., Hastie, T., Paul, D., and Tibshirani, R. (2006), "Prediction by Supervised Principal Components," *Journal of the American Statistical Association*, 101, 119–137. [1010,1011,1012,1019]

Bakshi, B. R. (1998), "Multiscale PCA With Application to Multivariate Statistical Process Monitoring," *American Institute of Chemical Engineers. AIChE Journal*, 44, 1596–1610. [1015]

Basak, D., Pal, S., and Patranabis, D. C. (2007), "Support Vector Regression," *Neural Information Processing-Letters and Reviews*, 11, 203–224. [1020]

Bickel, P., and Levina, E. (2004), "Some Theory for Fisher's Linear Discriminant Function, 'Naive Bayes', and Some Alternatives When There are Many More Variables Than Observations," *Bernoulli*, 10, 989–1010. [1010]

Bühlmann, P., Rutimann, P., Van de Geer, S., and Zhang, C. H. (2013), "Correlated Variables in Regression: Clustering and Sparse Estimation," *Journal of Statistical Planning and Inference*, 143, 1835–1858. [1010]

Chang, C.-C., and Lin, C.-J. (2011), "LIBSVM: A Library forSupport Vector Machines," *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2, 27. Available at *http://www.csie.ntu.edu.tw/∼cjlin/libsvm*. [1015]

Chun, H., and Keles, S. (2010), "Sparse Partial Least Squares Regression for Simultaneous Dimension Reduction and Variable Selection," *Journal of the Royal Statistical Society*, Series B, 72, 3–25. [1015,1020]

Clarke, B., Fokoue, E., and Zhang, H. H. (2009), *Principles and Theory for Data Mining and Machine Learning*, New York: Springer Verlag. [1010,1011]

Clemmensen, L., Hastie, T., Witten, D., and Ersbøll, B. (2011), "Sparse Discriminant Analysis," *Technometrics*, 53, 406–413. [1015]

Cuingnet, R., Glaunes, J. A., Chupin, M., Benali, H., Colliot, O., and ADNI. (2013), "Spatial and Anatomical Regularization of SVM: A General Framework for Neuroimaging Data," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35, 682–696. [1009,1010,1013]

Dray, S., and Jombart, T. (2011), "Revisiting Guerry's Data: Introducing Spatial Constraints in Multivariate Analysis," *Annals of Applied Statistics*, 5, 2278–2299. [1011]

Fan, J., and Fan, Y. (2008), "High-Dimensional Classification Using Features Annealed Independence Rules," *Annals of Statistics*, 36, 2605–2637. [1010,1011]

Fan, J., Feng, Y., Samworth, R., and Wu, Y. (2010), *SIS: Sure Independence Screening*, R Package Version 0.6. [1015]

Fan, J., Feng, Y., and Tong, X. (2012), "A Road to Classification in High Dimensional Space: The Regularized Optimal Affine Discriminant," *Journal of the Royal Statistical Society*, Series B, 74, 745–771. [1010,1015]

Fan, J., and Lv, J. (2008), "Sure Independence Screening for Ultrahigh Dimensional Feature Space," *Journal of the Royal Statistical Society*, Series B, 70, 849–911. [1011,1020]

Fawcett, T. (2006), "An Introduction to ROC Analysis," *Pattern Recognition Letters*, 27, 861–874. [1015]

Friston, K. J. (2009), "Modalities, Modes, and Models in Functional Neuroimaging," *Science*, 326, 399–403. [1009,1010]

Goldsmith, J., Huang, L., and Crainiceanu, C. M. (2014), "Smooth Scalar-on-Image Regression via Spatial Bayesian Variable Selection," *Journal of Computational and Graphical Statistics*, 23, 46–64. [1010]

Grosenick, L., Klingenberg, B., Katovich, K., Knutson, B., and Taylor, J. E. (2013), "Interpretable Whole-Brain Prediction Analysis With GraphNet," *NeuroImage*, 72, 304–321. [1015]

Guo, R., Ahn, M., and Zhu, H. (2015), "Spatially Weighted Principal Component Analysis for Imaging Classification," *Journal of Computational and Graphical Statistics*, 24, 274–296. [1010]

Hastie, T., Tibshirani, R., and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (2nd ed.), Hoboken, NJ: Springer. [1010,1011,1013]

Hinrichs, C., Singh, V., Mukherjee, L., Xu, G., Chung, M. K., Johnson, S. C., and ADNI, (2009), "Spatially Augmented LPboosting for AD Classification With Evaluations on the ADNI Dataset," *NeuroImage*, 48, 138–149. [1009,1010]

Huang, J. Z., Shen, H., and Buja, A. (2009), "The Analysis of Two-Way Functional Data Using Two-Way Regularized Singular Value Decompositions," *Journal of the American Statistical Association*, 104, 1609–1620. [1011]

Krishnan, A., Williams, L., McIntosh, A., and Abdi, H. (2011), "Partial Least Squares (PLS) Methods For Neuroimaging: A Tutorial and Review," *Neuroimage*, 56, 455–475. [1010]

Lee, M., Shen, H., Huang, J. Z., and Marron, J. S. (2010), "Biclustering via Sparse Singular Value Decomposition," *Biometrics*, 66, 1087–1095. [1012]

Li, F., Zhang, T., Wang, Q., Gonzalez, M. Z., Maresh, E. L., and Coan, J. A. (2015), "Spatial Bayesian Variable Selection and Grouping for High-Dimensional Scalar-on-Image Regression," *The Annals of Applied Statistics*, 9, 687–713. [1010]

Li, R. Z., Zhong, W., and Zhu, L. (2012), "Feature Screening via Distance Correlation Learning," *Journal of American Statistical Association*, 107, 1129–1139. [1011,1012]

Li, Y., Zhu, H., Shen, D., Lin, W., Gilmore, J. H., and Ibrahim, J. G. (2011), "Multiscale Adaptive Regression Models for Neuroimaging Data," *Journal of the Royal Statistical Society*, Series B, 73, 559–578. [1012,1013]

Luxburg, U. V. (2007), "A Tutorial on Spectral Clustering," *Statistics and Computing*, 17, 395–416. [1013]

Marcus, C., Mena, E., and Subramaniam, R. M. (2014), "Brain PET in the Diagnosis of Alzheimer's Disease," *Clinical Nuclear Medicine*, 39, e413–e426. [1017]

Nowrangi, M. A., and Rosenberg, P. B. (2015), "The Fornix in Mild Cognitive Impairment and Alzheimer's Disease," *Frontiers in Aging Neuroscience*, 7, 1. Available at *https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4301006/* [1020]

Polzehl, J., and Spokoiny, V. G. (2006), "Propagation-Separation Approach for Local Likelihood Estimation," *Probability Theory and Related Fields*, 135, 335–362. [1012,1013]

Reiss, P., and Ogden, R. (2010), "Functional Generalized Linear Models With Images as Predictors," *Biometrics*, 66, 61–69. [1010]

Shen, D., and Zhu, H. (2015), "Spatially Weighted Principal Component Regression for High-Dimensional Prediction," in *International Conference on Information Processing in Medical Imaging*, Springer, pp. 758–769. [1010]

Skocaj, D., Leonardis, A., and Bischof, H. (2007), "Weighted and Robust Learning of Subspace Representations," *Pattern Recognition*, 40, 1556–1569. [1010,1011,1019]

Taylor, K. M., and Meyer, F. G. (2012), "A Random Walk on Image Patches," *SIAM Journal on Imaging Sciences*, 5, 688–725. [1012]

Tibshirani, R. (1996), "Regression Shrinkage and Selection via the Lasso," *Journal of the Royal Statistical Society*, Series B, 58, 267–288. [1010,1020]

Vincent, M., Gramfort, A., Varoquaux, G., Eger, E., and Thirion, B. (2011), "Total Variation Regularization for fMRI-Based Prediction of Behavior," *IEEE Transactions on Medical Imaging*, 30, 1328–1340. [1009,1010]

Yamashita, O. (2011), "Quick Manual for Sparse Logistic Regression Toolbox ver1.2.1." Available at *http://www.cns.atr.jp/oyamashi/SLR_WEB/*. [1015]

Yan, S., Xu, D., Zhang, B., Zhang, H. J., Yang, Q., and Lin, S. (2007), "Graph Embedding and Extensions: A General Framework for Dimensionality Reduction," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29, 40–51. [1011]

Zhao, Y., Ogden, R. T., and Reiss, P. T. (2012), "Wavelet-Based LASSO in Functional Linear Regression," *Journal of Computational and Graphical Statistics*, 21, 600–617. [1010]

Zhou, H., Li, L., and Zhu, H. T. (2013), "Tensor Regression With Applications in Neuroimaging Data Analysis," *Journal of American Statistical Association*, 108, 540–552. [1010]