Published in final edited form as:
J Am Stat Assoc. 2017 ; 112(519): 1009–1021. doi:10.1080/01621459.2016.1261710.

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

Hongtu Zhu,
Professor of Biostatistics, Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, TX, 77230, and University of North Carolina, Chapel Hill, NC, 27599

Dan Shen,
Assistant Professor in Interdisciplinary Data Sciences Consortium and Department of Mathematics and Statistics, University of South Florida, Tampa, FL 33620

Xuewei Peng, and **Leo Yufeng Liu**
Doctoral student under the supervision of Dr. Hongtu Zhu

for the Alzheimer's Disease Neuroimaging Initiative

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 score weights with the spatial weights in order 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) data set.

Keywords

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

Address for correspondence and reprints: Hongtu Zhu, Ph.D., hzhu5@mdanderson.org; Phone No: 346-814-0191.

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 in order 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 paper 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 two 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 $\mathbf{x} = (\mathbf{x}_{\mathbf{g}}: \mathbf{g} \in \mathscr{G})$, to predict an outcome variable, denoted as $\mathbf{y}$, where $\mathscr{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_p\}$ is a set of locations, in which p is the total number of locations in $\mathscr{G}$: In this case, $\mathbf{x}$ is a vector of FDG-PET imaging measures on a 3-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 3 randomly selected Alzheimer's Disease (AD) subjects and 3 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; Vincent et al., 2011; Hinrichs et al., 2009; Cuingnet et al., 2012).

Many existing supervised learning and variable selection methods (Hastie et al., 2009; Clarke et al., 2009; Fan and Fan, 2008; Bickel and Levina, 2004; Buhlmann et al., 2012; Tibshirani, 1996), 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 *non-sparse* (Li et al., 2015; Zhou et al., 2013; Friston, 2009; Hinrichs et al., 2009). First, the existing unstructured regularization methods can suffer from diverging spectra and noise accumulation in high dimensional feature space (Reiss and Ogden, 2010; Bickel and Levina, 2004; Buhlmann et al., 2012; Fan and Fan, 2008), 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., 2012; Fan et al., 2012; Goldsmith et al., 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 (Skocaj et al., 2007; Bair et al., 2006; Fan and Fan, 2008; Krishnan et al., 2011; Zhao et al., 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 et al., 2014; Guo et al., 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 **x** in order to dramatically improve prediction accuracy. Compared with the existing literature (Allen et al., 2014; Guo et al., 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 paper 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 **y** and each individual feature $\mathbf{x}_\mathbf{g}$ across $\mathbf{g} \in \mathscr{G}$ and the spatial patterns of **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 $\mathbf{x}_i = \{\mathbf{x}_{i,\mathbf{g}}: \mathbf{g} \in \mathscr{G}\}$ measured on $\mathscr{G}$ 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, $\mathscr{G} = \{\mathbf{g}_1, \ldots, \mathbf{g}_p\}$ is a set of prefixed locations, such as voxels in 3D lattices, so it is possible to define an edge set $\mathscr{S} = \{(\mathbf{g}_k, \mathbf{g}_j): \mathbf{g}_k, \mathbf{g}_j \in \mathscr{G}\}$ associated with $\mathscr{G}$. 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 $\mathscr{S}$.

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 ℓ–th scale for $\ell = 1, \ldots, L$. The low-rank model for MWPCA can be written as

$$\tilde{\mathbf{X}}^{(\ell)} = (\mathbf{X} - \mathbf{1}_n \boldsymbol{\mu}^T) Q^{(\ell)} = U^{(\ell)} D^{(\ell)} V^{(\ell)T} + \mathscr{E}^{(\ell)} = \sum_{k=1}^{K} d_k^{(\ell)} \mathbf{u}_k^{(\ell)} \mathbf{v}_k^{(\ell)T} + \mathscr{E}^{(\ell)} \quad (1)$$

for $\ell = 1, \ldots, L$, where $E(\mathbf{x}_i) = \boldsymbol{\mu}$, $K \le \min(n, p)$, and $\mathscr{E}^{(\ell)} = (\varepsilon_1^{(\ell)}, \ldots, \varepsilon_n^{(\ell)})^T$ 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 $U_C = (\mathbf{u}_{C,1} \cdots \mathbf{u}_{C,n})^T = (U^{(1)}, \ldots, U^{(L)})$. We then build a prediction model $R(\mathbf{y}_i; \boldsymbol{u}_{C,i}, \mathbf{z}_i, \boldsymbol{\theta})$ with $\mathbf{y}_i$ as response and $\boldsymbol{u}_{C,i}$ and $\mathbf{z}_i$ as covariates, where $\boldsymbol{\theta}$ is a vector of unknown (finite-dimensional or non-parametric) parameters. For instance, when $q_y = 1$, a popular prediction model is the generalized linear model given by

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

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 et al., 2009; Clarke et al., 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 (Allen et al., 2014; Skocaj et al., 2007; Huang et al., 2009). 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 (Dray and Jombart, 2011; Skocaj et al., 2007) 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) $W_I = (w_{I,\mathbf{g}}): \mathscr{G} \rightarrow R^+$ and a spatial weight matrix $W_E = (w_{E,\mathbf{g}\mathbf{g}'}): \mathscr{G} \times \mathscr{G} \rightarrow R$.
- Stage 2. At the ℓ-th 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 $R(\mathbf{y}; \boldsymbol{u}_C, \mathbf{z}, \boldsymbol{\theta})$.

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 learnt 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 et al., 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 $\mathscr{S}$, depending on whether the corresponding correlation value exceeds a prefixed threshold or not. By choosing different thresholds, we can obtain different edge sets $\mathscr{S}$. 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 ℓ-th scale as follows. To extract important features from **x**, we construct a matrix $Q_1^{(\ell)}=\text{diag}\left(1\{w_{I,\mathbf{g}_1}\geq s_{I,\ell}\},\ldots,1\{w_{I,\mathbf{g}_p}\geq s_{I,\ell}\}\right)$, where $1\{\cdot\}$ is an indicator function and $s_{I,1} \leq \ldots \leq s_{I,L}$ are pre-specified thresholds. The use of $Q_1^{(\ell)}$ is similar to various marginal screening methods (Fan and Lv, 2008; Fan and Fan, 2008; Bair et al., 2006). By tuning the value of $s_{I,\ell}$, we can screen out 'uninformative' features at different scales.

To capture the spatial features of **x**, we may construct a spatial similarity matrix $Q_2^{(\ell)}=(|w_{E,\mathbf{g}_k\mathbf{g}_j}|1\{|w_{E,\mathbf{g}_k\mathbf{g}_j}|\geq s_{E,\ell;1},D(\mathbf{g}_k,\mathbf{g}_j)\leq s_{E,\ell;2}\})$, 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 $\{\mathbf{g}_j \in \mathscr{G}: D(\mathbf{g}_k, \mathbf{g}_j) \leq s_{E,\ell,2}\}$, 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 **x** first and then smoothing those selected features. In contrast, $Q^{(\ell)}=Q_2^{(\ell)}Q_1^{(\ell)}$ corresponds to smoothing **x** first and then extracting important features from the smoothed **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 $\tilde{\mathbf{X}}^{(\ell)}$ 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

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

subject to $\boldsymbol{u}_k^{(\ell)T}\boldsymbol{u}_k^{(\ell)}\leq 1$ and $\boldsymbol{v}_k^{(\ell)T}\boldsymbol{v}_k^{(\ell)}\leq 1$ 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 et al., 2006). For each k_0, we use the sparse method in Lee et al. (2010) to estimate $(d_{k_0}^{(\ell)},\boldsymbol{u}_{k_0}^{(\ell)},\boldsymbol{v}_{k_0}^{(\ell)})$. In this way, we can sequentially compute $(d_k^{(\ell)},\boldsymbol{u}_k^{(\ell)},\boldsymbol{v}_k^{(\ell)})$ for $k = 1, \ldots, K$.

In Stage 3, based on $\{(\mathbf{y}_i, \boldsymbol{u}_{C,i}, \mathbf{z}_i)\}_{i\geq 1}$, we use an estimation method to estimate $\boldsymbol{\theta}$ as follows:

$$\hat{\boldsymbol{\theta}}=\text{argmin}_{\boldsymbol{\theta}}\{\rho(R,\boldsymbol{\theta},\{(\mathbf{y}_i,\boldsymbol{u}_{C,i},\mathbf{z}_i)\}_{i\geq 1})+\lambda P_3(\boldsymbol{\theta})\}, \quad (4)$$

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 learnt from the training data.
- Optimize an objective function based on $R(\mathbf{y};\mathbf{u}_C^*,\mathbf{z}^*,\hat{\boldsymbol{\theta}})$ 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 $\mathbf{y}$ in order to perform feature selection according to each feature's discriminative importance. Statistically, most existing methods (Bair et al., 2006; Li et al., 2012) use a marginal model by assuming

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

where $\boldsymbol{\beta} = (\boldsymbol{\beta}(\mathbf{g}): \mathbf{g} \in \mathscr{G})$ and $\boldsymbol{\beta}(\mathbf{g})$ is introduced to quantify the association between $\mathbf{y}_i$ and $\mathbf{x}_{i,g}$ at each location $\mathbf{g} \in \mathscr{G}$. 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 et al., 2012).

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

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

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

We propose two strategies to learn the homogenous patches $\mathscr{G}_j$ in (6) by jointly modelling $(\mathbf{x}_i, \mathbf{z}_i)$ 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 $\mathscr{G}_j$ 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 $\mathscr{S}_g(h)$ 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 $h_0 = 0 \le h_1 \le \ldots \le h_S$ and $\mathscr{S}_g(h_0) = \{\mathbf{g}\} \subset \ldots \subset \mathscr{S}_g(h_S)$. To learn the homogeneous 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:

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

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

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

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 $\mathbf{g}' \in \mathscr{S}_g(h_S)$, 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 in order 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:

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

where $\mathscr{N}_g$ 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 $\mathscr{S}_g(h_0) = \{\mathbf{g}\} \subset \ldots \subset \mathscr{S}_g(h_S)$ for $h_0 = 0 \leq h_1 \leq \ldots \leq h_S$.

- [Step (i)] Calculate $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0)$ and $\mathrm{Cov}(\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0))$ according to $\prod_{i=1}^{n} f\left(\mathbf{y}_i|\mathbf{x}_{i,g},\mathbf{z}_i;\boldsymbol{\beta}(\mathbf{g})\right)$ across all locations $\mathbf{g}$.
- [Step (ii)] Smooth $\{\hat{\boldsymbol{\beta}}(\mathbf{g}; h_0): \mathbf{g} \in \mathscr{G}\}$ to sequentially estimate $\hat{\boldsymbol{\beta}}(\mathbf{g}; h_s)$ for $s = 1, \ldots, S$ across all $\mathbf{g} \in \mathscr{G}$. Candidate methods include local polynomial, nonlocal mean, and propagation-separation, among others (Polzehl and Spokoiny, 2006; Arias-Castro et al., 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 $\mathbf{g} \in \mathscr{G}$. 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 $\{\hat{\boldsymbol{\beta}}(\mathbf{g}; h_S): \mathbf{g} \in \mathscr{G}\}$ into several homogeneous clusters (Hastie et al., 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 $Q_2^{(\ell)}{=}\sum^{-1/2}$; thus, $Q_2^{(\ell)}Q_2^{(\ell)T}{=}\sum^{-1}$ 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 non-negative weights assigned to the spatial neighboring locations of each location. Specifically, we set $w_{E,\mathbf{gg}'}$ as

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

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

The cluster-based spatial weight matrix consists of non-negative weights assigned to locations in the same homogeneous 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 non-negative weight $w_{\mathbf{gg}'}$. Thus, matrix $W = (w_{\mathbf{gg}'})$ is a weighted adjacency matrix of $\mathscr{G}$. The degree of a location $\mathbf{g} \in \mathscr{G}$ is defined as $d_{\mathbf{g}} = \Sigma_{\mathbf{g}' \in \mathscr{G}}$ $w_{\mathbf{gg}'}$ and the degree matrix WD is given by $W_D = \text{diag}(d_{g1}, \ldots, d_{gp})$.The unnormalized Laplacian matrix L of the graph $\mathscr{G}$ 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., 2012).

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 $\hat{\beta}(\mathbf{g}; h_S)$. Although the use of $\Sigma^{-1/2}$ can improve classification accuracy, 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 × 20 × 10 3D-images from a linear model given by

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

where l_i is the class label coded as either 0 or 1 and $\varepsilon_i(\mathbf{g})$ 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 × 3 × 4 region characterizes the maximum difference 1 between classes 0 and 1. In this case, we have $p = 4, 000$. Then, we set $n = 100$ with 60 images from Class 0 and the rest from Class 1.

We consider three types of noise $\varepsilon_i(\mathbf{g})$ in (11). First, $\varepsilon_i^{(1)}(\mathbf{g})$ were independently generated from a $N(0, 2^2)$ generator across all voxels. Second, $\varepsilon_i^{(2)}(\mathbf{g})=\sum_{\|\mathbf{g}'-\mathbf{g}\|_1\le 1}\varepsilon_i^{(1)}(\mathbf{g}')/m_{\mathbf{g}}$ were generated from $\varepsilon_i^{(1)}(\mathbf{g})$ 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, $\varepsilon_i^{(3)}(\mathbf{g})$ were generated according to $\varepsilon_i^{(3)}(\mathbf{g})=2\sin(\pi g_1/10)\xi_{i,1}+2\cos(\pi g_2/10)\xi_{i,2}+2\sin(\pi g_3/5)\xi_{i,3}+\varepsilon_i^1(\mathbf{g})$, where $\mathbf{g} = (g_1, g_2, g_3)^T$ and $\xi_{i,k}$ for $k = 1, 2, 3$ were independently generated from a $N(0, 1)$ 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 $\mathbf{g} \in \mathscr{G}$, we set $w_{I,\mathbf{g}} = -p \log(p(\mathbf{g}))/\{-\Sigma_{\mathbf{g}\in\mathscr{G}} \log(p(\mathbf{g}))\}$, 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 in order 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 paper. 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 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 due to the fact that 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 5 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 et al., 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 data sets.

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 in order 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 in order 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.

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 neurodegenetive disease. The incidence of AD doubles every five 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 data set to address questions (Q1) and (Q2). The ADNI PET data set 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 minutes 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 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 scatter plot 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 22p. It is probably involved with language perception and processing (Marcus et al., 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 randomly and proportionally split the data set into 2 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 in order 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 ten 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 two 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 ten 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 $\mathbf{x}_{i,\mathbf{g}}$ 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 in order to construct the low-dimensional latent variables

$\{\mathbf{u}_{k,i}\}$. In Stage 3, we fitted a linear latent variable regression given by $\mathbf{y}_i = \alpha_0 + \sum_{k=1}^{K} \alpha_k \mathbf{u}_{k,i} + \varepsilon_i$ to do prediction.

Second, we compared MWPCR and three other dimensional reduction methods including PCA, weighted PCA (WPCA) (Skocaj et al., 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 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 et al., 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 utilizing 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 Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

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. We are grateful for the many valuable suggestions from referees, associated editor, and editor.

We would like to thank the Editor, the anonymous Associate Editor, and the referees for their suggestions that have led to a much improved paper.

"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.

"Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). 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 non-profit 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."

References

Aharon M, Elad M, Bruckstein A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse representation. IEEE Trans on Signal Processing. 2006; 54:4311–4322.

Allen GI, Grosenick L, Taylor J. A generalized least squares matrix decomposition. Journal of the American Statistical Association. 2014; 109:145–159.

Arias-Castro E, Salmon J, Willett R. Oracle Inequalities and Minimax Rates for Nonlocal Means and Related Adaptive Kernel-Based Methods. SIAM J Imaging Sci. 2012; 5:944–992.

Bair E, Hastie T, Paul D, Tibshirani R. Prediction by supervised principal components. J Amer Statist Assoc. 2006; 101:119–137.

Bakshi BR. Multiscale PCA with application to multivariate statistical process monitoring. American Institute of Chemical Engineers AIChE Journal. 1998; 44:1596.

Basak D, Pal S, Patranabis DC. Support vector regression. Neural Information Processing-Letters and Reviews. 2007; 11:203–224.

Bickel P, Levina E. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. Bernoulli. 2004; 10:989–1010.

Buhlmann P, Rutimann P, Van de Geer S, Zhang CH. Correlated variables in regression: clustering and sparse estimation. Tech rep, ETH Zurich. 2012

Chang C-C, Lin C-J. LIBSVM: a library for support vector machines. ACM Transactions on Intelligent Systems and Technology (TIST). 2011; 2:27.

Chun H, Keles S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J Roy Statist Soc Ser B. 2010; 72:3–25.

Clarke, B., Fokoue, E., Zhang, HH. Principles and Theory for Data Mining and Machine Learning. New York: Springer Verlag; 2009.

Clemmensen L, Hastie T, Witten D, Ersbøll B. Sparse discriminant analysis. Technometrics. 2011; 53:406–413.

Cuingnet R, Glaunes JA, Chupin M, Benali H, Colliot O. ADNI. Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2012

Dray S, Jombart T. Revisiting Guerry's data: introducing spatial constraints in multivariate analysis. Annals of Applied Statistics. 2011; 5:2278–2299.

Fan J, Fan Y. High-dimensional classification using features annealed independence rules. Ann Statist. 2008; 36:2605–2637.

Fan, J., Feng, Y., Samworth, R., Wu, Y. R package version 0.6. 2010. SIS: Sure Independence Screening.

Fan J, Feng Y, Tong X. A road to classification in high dimensional space: the regularized optimal affine discriminant. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2012; 74:745–771. [PubMed: 23074363]

Fan J, Lv J. Sure independence screening for ultrahigh dimensional feature space. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2008; 70:849–911. [PubMed: 19603084]

Fawcett T. An introduction to ROC analysis. Pattern recognition letters. 2006; 27:861–874.

Friston KJ. Modalities, modes, and models in functional neuroimaging. Science. 2009; 326:399–403. [PubMed: 19833961]

Goldsmith J, Huang L, Crainiceanu CM. Smooth scalar-on-image regression via spatial Bayesian variable selection. Journal of Computational and Graphical Statistics. 2014; 23:46–64. [PubMed: 24729670]

Grosenick L, Klingenberg B, Katovich K, Knutson B, Taylor JE. Interpretable whole-brain prediction analysis with GraphNet. NeuroImage. 2013; 72:304–321. [PubMed: 23298747]

Guo R, Ahn M, Zhu H. Spatially weighted principal component analysis for imaging classification. Journal of Computational and Graphical Statistics. 2015; 24:274–296. [PubMed: 26089629]

Hastie, T., Tibshirani, R., Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2. Hoboken, New Jersey: Springer; 2009.

Hinrichs C, Singh V, Mukherjee L, Xu G, Chung MK, Johnson SC. ADNI. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. NeuroImage. 2009; 48:138–149. [PubMed: 19481161]

Huang JZ, Shen H, Buja A. The analysis of two-way functional data using two-way regularized singular value decompositions. Journal of the American Statistical Association. 2009; 104:1609–1620.

Krishnan A, Williams L, McIntosh A, Abdi H. Partial least squares (PLS) methods for neuroimaging: a tutorial and review. Neuroimage. 2011; 56:455–475. [PubMed: 20656037]

Lee M, Shen H, Huang JZ, Marron JS. Biclustering via Sparse Singular Value Decomposition. Biometrics. 2010; 66:1087–1095. [PubMed: 20163403]

Li F, Zhang T, Wang Q, Gonzalez MZ, Maresh EL, Coan JA, et al. Spatial Bayesian variable selection and grouping for high-dimensional scalar-on-image regression. The Annals of Applied Statistics. 2015; 9:687–713.

Li RZ, Zhong W, Zhu L. Feature screening via distance correlation learning. Journal of American Statistical Association. 2012; 107:1129–1139.

Li Y, Zhu H, Shen D, Lin W, Gilmore JH, Ibrahim JG. Multiscale adaptive regression models for neuroimaging data. Journal of the Royal Statistical Society: Series B. 2011; 73:559–578.

Luxburg UV. A tutorial on spectral clustering. Statistics and Computing. 2007; 17:395–416.

Marcus C, Mena E, Subramaniam RM. Brain PET in the diagnosis of Alzheimer's disease. Clinical nuclear medicine. 2014; 39:e413. [PubMed: 25199063]

Nowrangi MA, Rosenberg PB. The fornix in mild cognitive impairment and Alzheimer's disease. Frontiers in aging neuroscience. 2015; 7:1. [PubMed: 25653617]

Polzehl J, Spokoiny VG. Propagation-separation approach for local likelihood estimation. Probab Theory Relat Fields. 2006; 135:335–362.

Reiss P, Ogden R. Functional generalized linear models with images as predictors. Biometrics. 2010; 66:61–69. [PubMed: 19432766]

Shen, D., Zhu, H. International Conference on Information Processing in Medical Imaging. Springer; 2015. Spatially Weighted Principal Component Regression for High-Dimensional Prediction; p. 758-769.

Skocaj D, Leonardis A, Bischof H. Weighted and robust learning of subspace representations. Pattern Recogn. 2007; 40:1556–1569.

Taylor KM, Meyer FG. A random walk on image patches. SIAM J Imaging Sciences. 2012; 5:688–725.

Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B. 1996; 58:267–288.

Vincent M, Gramfort A, Varoquaux G, Eger E, Thirion B. Total variation regularization for fMRI-based prediction of behavior. IEEE Transactions on Medical Imaging. 2011; 30:1328–1340. [PubMed: 21317080]

Yamashita, O. Quick manual for sparse logistic regression toolbox ver1.2.1. 2011. software at http://www.cns.atr.jp/~oyamashi/SLR_WEB/

Yan S, Xu D, Zhang B, Zhang HJ, Yang Q, Lin S. Graph embedding and extensions: a general framework for dimensionality reduction. IEEE Transactions on Pattern Analysis and Machine Intelligence. 2007; 29:40–51. [PubMed: 17108382]

Zhao Y, Ogden RT, Reiss PT. Wavelet-based LASSO in functional linear regression. Journal of Computational and Graphical Statistics. 2012; 21:600–617. [PubMed: 23794794]

Zhou H, Li L, Zhu HT. Tensor regression with applications in neuroimaging data analysis. Journal of Americal Statistical Association. 2013; 108:540–552.

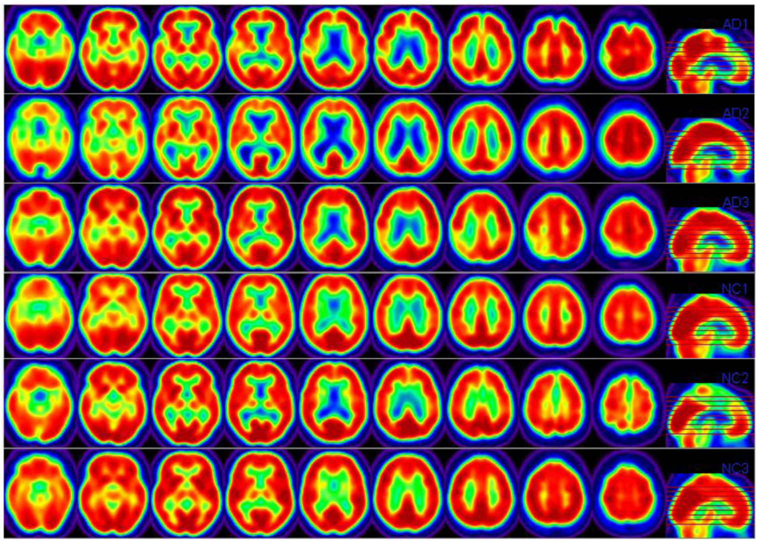
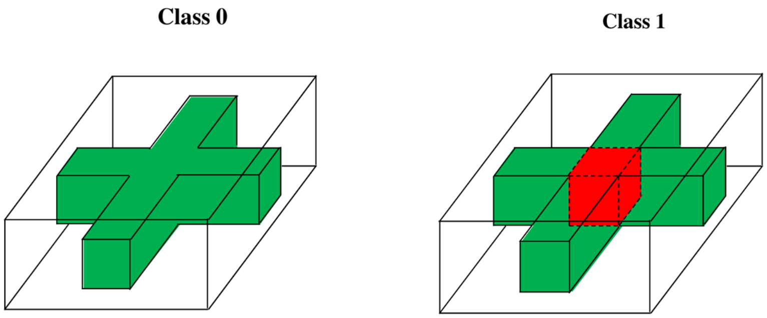


Figure 1.

ADNI PET Data. Each row consists of pre-selected 2-dimensional (2D) slides obtained from a randomly selected subject. The first three rows come from 3 randomly selected AD subjects and the last three rows come from 3 randomly selected NC subjects.



Class 0

Class 1

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.

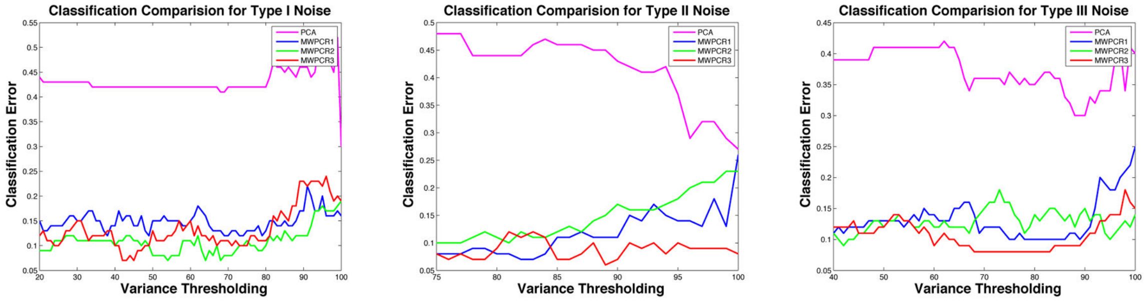


Figure 3.

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.

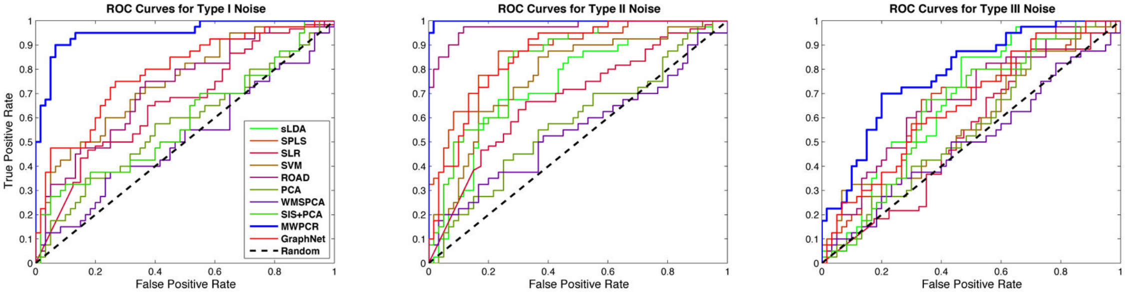


Figure 4.

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.

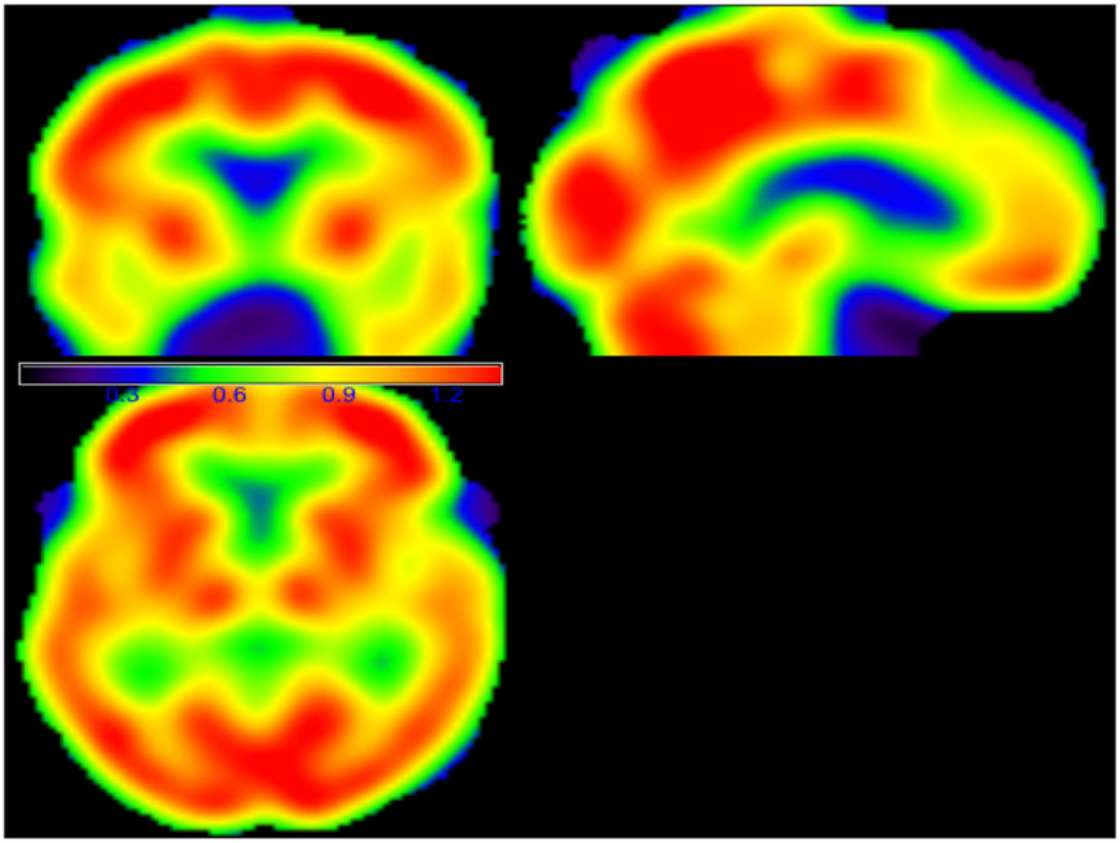


Figure 5.

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.

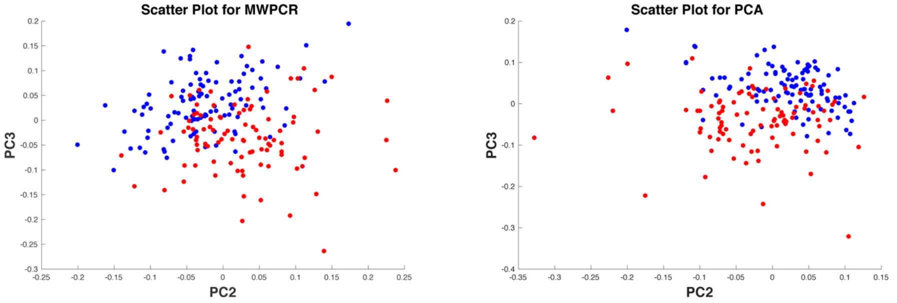


Figure 6.

ADNI binary classification results: scatter plots 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.

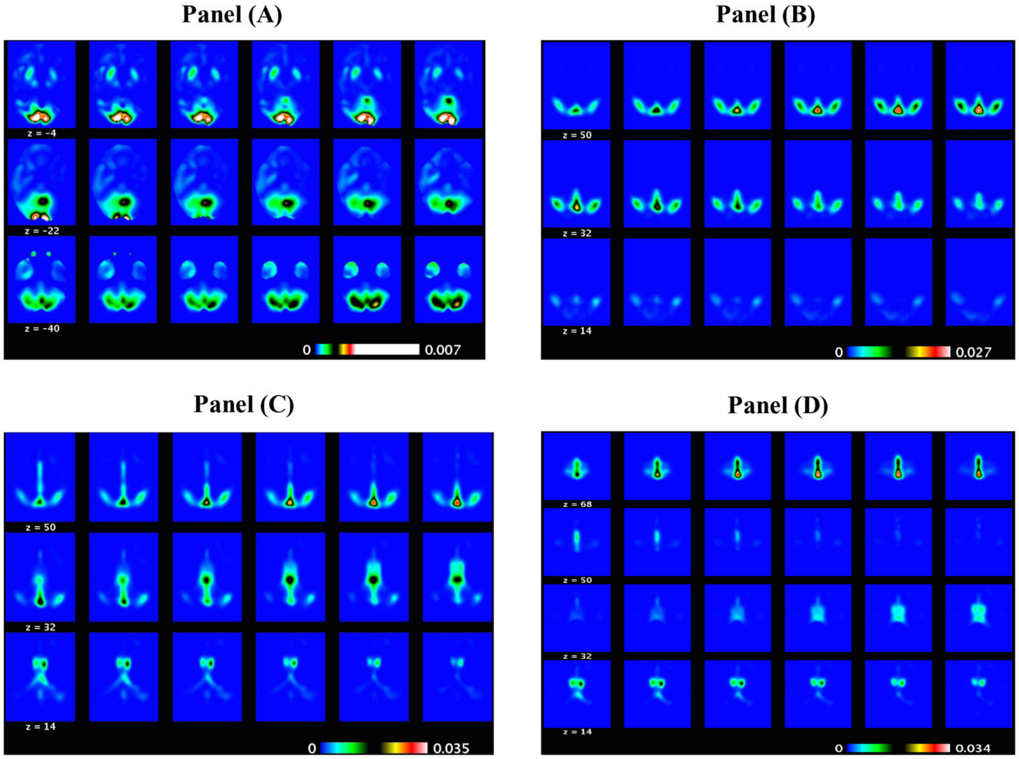


Figure 7.

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.

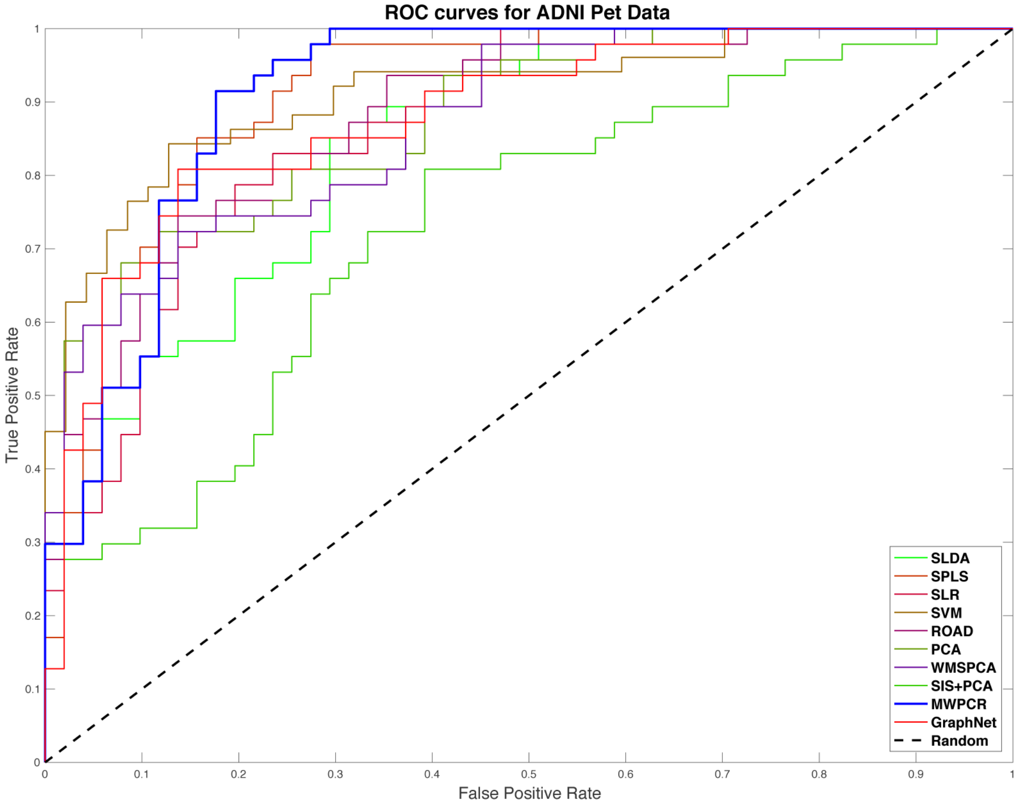


Figure 8.

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

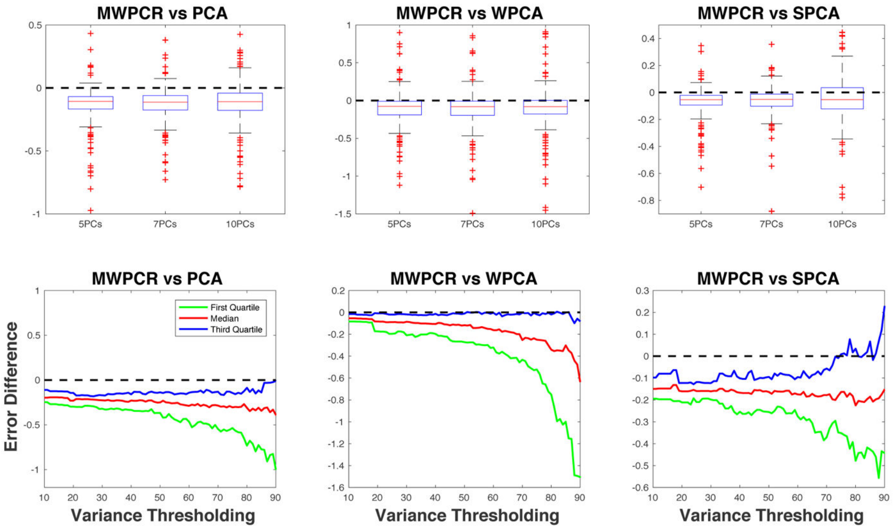


Figure 9.

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.

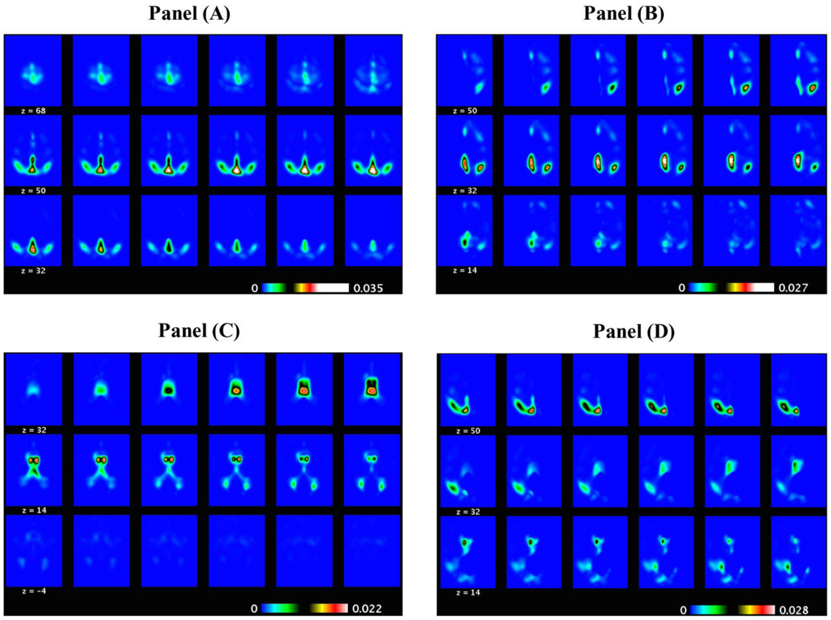


Figure 10.

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.

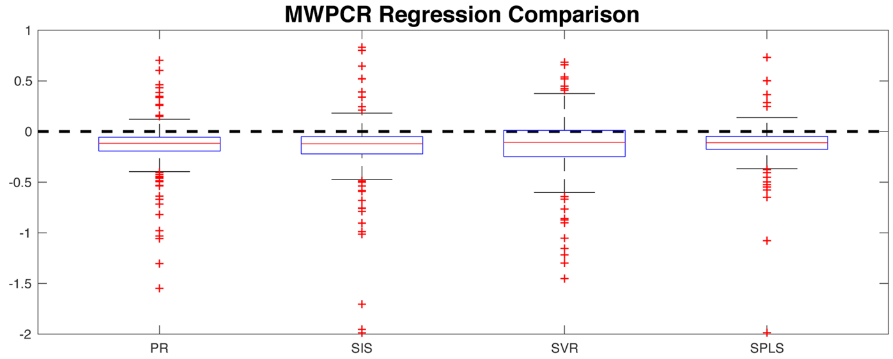


Figure 11.

ADAS-Co**g** score prediction for ADNI PET Data: comparison of MWPCR with the four other regression methods, including PR, SIS, SVR and SPLS.

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

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. 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.

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

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