Published in final edited form as:

IEEE Trans Med Imaging. 2016 August ; 35(8): 1927–1936. doi:10.1109/TMI.2016.2538289.

Sparse Multi-Response Tensor Regression for Alzheimer's Disease Study With Multivariate Clinical Assessments

Zhou Li,

Department of Statistics, North Carolina State University, Raleigh, NC 27695 USA

Heung-Il Suk,

Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea

Dinggang Shen, and

Biomedical Research Imaging Center (BRIC) and Department of Radiology, University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA, and also with the Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, South Korea

Lexin Li

Division of Biostatistics, University of California at Berkeley, Berkeley, CA 94720 USA

Zhou Li: zli15@ncsu.edu; Heung-Il Suk: hisuk@korea.ac.kr; Dinggang Shen: dgshen@med.unc.edu; Lexin Li: lexinli@berkeley.edu

Abstract

Alzheimer's disease (AD) is a progressive and irreversible neurodegenerative disorder that has recently seen serious increase in the number of affected subjects. In the last decade, neuroimaging has been shown to be a useful tool to understand AD and its prodromal stage, amnestic mild cognitive impairment (MCI). The majority of AD/MCI studies have focused on disease diagnosis, by formulating the problem as classification with a binary outcome of AD/MCI or healthy controls. There have recently emerged studies that associate image scans with continuous clinical scores that are expected to contain richer information than a binary outcome. However, very few studies aim at modeling *multiple* clinical scores *simultaneously*, even though it is commonly conceived that multivariate outcomes provide *correlated* and *complementary* information about the disease pathology. In this article, we propose a sparse multi-response tensor regression method to model multiple outcomes jointly as well as to model multiple voxels of an image jointly. The proposed method is particularly useful to both infer clinical scores and thus disease diagnosis, and to identify brain subregions that are highly relevant to the disease outcomes. We conducted experiments on the Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset, and showed that the proposed method enhances the performance and clearly outperforms the competing solutions.

Index Terms

Alzheimer's Disease; Magnetic Resonance Imaging; Multiple Responses; Region Selection; Tensor Regression

Correspondence to: Dinggang Shen, dgshen@med.unc.edu.

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

I. Introduction

Alzheimer's disease (AD), characterized by progressive impairment of cognitive and memory functions, is an irreversible neurodegenerative disorder and the leading form of dementia in elderly subjects. The number of affected subjects increases significantly every year, and is projected to be 1 in 85 by the year 2050 [1]. Amnestic mild cognitive impairment (MCI) is often a prodromal stage to AD, and individuals with MCI may convert to AD at an annual rate of as high as 15% [2]. There has been a vast body of literatures studying AD and MCI using one or more neuroimaging technologies (modalities), including magnetic resonance imaging (MRI), functional magnetic resonance imaging (fMRI), positron emission tomography (PET), diffusion tensor imaging (DTI), among many others.

The majority of AD/MCI studies have been concentrating on differentiating AD and MCI subjects from the general population, because an accurate diagnosis of AD and MCI is particularly important for timely therapy and possible delay of the disease. This can be formulated as a *classification* problem, and a variety of statistical machine learning techniques have been employed for imaging-based diagnosis. See [3]–[5] for some excellent reviews. Moreover, in addition to classifying a binary or categorical disease outcome given brain image scans, there were studies establishing associations between image activity patterns and a *continuous* clinical outcome. A variety of cognitive and memory scores have been used as the response, including the Mini-Mental State Examination (MMSE) [6]–[9], Boston Naming Testing [9], Dementia Rating Scale, Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog), and Auditory Verbal Learning Test [8].

More recently, there have emerged studies that associate image scans with *multiple* clinical outcomes [10], [11]. Our motivating data example consists of 194 subjects from the Alzheimer's Disease Neuroimaging Initiative (ADNI), among which 93 are AD patients and 101 healthy controls. For each subject, the collected data include a $32 \times 32 \times 32$ MRI scan, after proper preprocessing and downsizing, and two clinical scores. One is the MMSE, which examines orientation to time and place, immediate and delayed recall of three words, attention and calculation, language and visuo-constructional functions [12]. The other is ADAS-Cog, which is a global measure encompassing the core symptoms of AD [13]. The ADAS-Cog is usually more sensitive, but requires more than 30 minutes for participants to complete all tasks. In contrast, the administration of MMSE takes only 10–15 minutes and thus is often used for fast screening for dementia. While both can be used to measure the severity of cognitive impairment, the scores of MMSE and ADAS-Cog carry, respectively, “local” and “global” information with respect to the cognitive capability, given the fact that the examination in MMSE is conducted on more specific tasks than in ADAS-Cog. Furthermore, recent studies have shown that the two scores are correlated, as they reflect on similar cognitive aspects such as orientation and memory [8], [14]. In these regards, it is beneficiary to jointly consider these scores in AD/MCI studies. Our aim in this article is to *jointly* model multiple clinical outcomes given brain structural patterns, under the belief that the multivariate scores provide *correlated* and *complementary* information.

Whereas there is an enormous body of statistics literature on modeling multivariate predictors, there have been much fewer works on modeling multivariate responses. Some

popular multi-response solutions include partial least squares [15]–[17], canonical correlations [18], reduced-rank regressions [19]–[21], sparse regressions with various penalties [22]–[24], and sparse reduced-rank regressions [25]. All existing multi-response modeling methods universally treat the predictors as a *vector* and estimate a corresponding vector of coefficients. However, in neuroimaging analysis, the predictors take a more complex form of multi-dimensional array, a.k.a. *tensor*. Naively turning an array into a vector would result in extremely high dimensionality. For instance, a $32 \times 32 \times 32$ MRI image would imply $32^3 = 32{,}768$ parameters. Moreover, vectorizing an array would also destroy all the inherent spatial structural information of the image.

In this article, we propose a *sparse multi-response tensor regression model* to simultaneously infer multiple outcome variables and to identify brain subregions that are highly relevant to the clinical outcomes. A schematic overview of our proposed method is given in Fig. 1. The new method enjoys a number of appealing features. First, it models the multiple responses *jointly* rather than *separately*, by employing a penalty function accounting for correlated multivariate responses while inducing group sparsity. Second, the new method models brain image predictor in the form of a *tensor* rather than a *vector*. This is achieved by extending and generalizing a recent proposal of tensor predictor regression [26]. Directly modeling a tensor image predictor takes into account spatial correlations among the voxels, and is intuitively superior than the one-voxel-at-a-time modeling solution that ignores such correlations. This extension, however, is far from trivial, since [26] only considered a univariate response, and our proposal for multi-response requires a new form of penalty and a new optimization algorithm. Third, our method offers a competitive alternative to the common modeling strategy in neuroimaging literature that first groups individual voxels by predefined regions of interest (ROI) and then extracts a vector of useful features from ROIs. By contrast, our solution does not rely on any prior knowledge of ROIs but derives highly relevant features suggested directly by the data. Moreover, instead of conducting feature extraction and association modeling at two separate steps, our method simultaneously derives relevant features and builds their association with the outcomes. Last but not least, we develop a highly scalable computational algorithm that makes our method applicable to a range of massive imaging data.

II. Materials

A. Subjects

We analyzed a dataset from the ADNI database (http://adni.loni.usc.edu/). The data included 93 subjects with mild AD and 101 normal controls (NC), and each subject had both MMSE and ADAS-Cog scores recorded. The subjects were in the age between 55 and 90, with a study partner, who provided an independent evaluation of functioning. All of 194 subjects met the following general inclusion criteria: (a) NC subjects: an MMSE between 25 and 30 (inclusive), a clinical dementia rating (CDR) of 0, non-depressed, non-MCI, and non-demented; (b) mild AD subjects: an MMSE score between 18 and 27 (inclusive), a CDR of 0.5 or 1.0, and met the National Institute of Neurological and Communicative Disorders and Stroke and the Alzheimer's Disease and Related Disorders Association (NINCDS/ADRDA) criteria for probable AD. The AD and NC groups were matched in age (with the two-sample

t-test p-value = 0.68) and gender (with the two-sample proportion test p-value = 1.00). Table I presents the demographic characteristics of the subjects.

B. Preprocessing

All anatomical MRI data in this study were acquired using 1.5T scanners. The baseline MRI data were downloaded from ADNI in the neuroimaging informatics technology initiative (NIfTI) format, which had already been processed for spatial distortion correction caused by gradient nonlinearity and B1 field inhomogeneity. We further performed prevalent preprocessing procedures on all images, including Anterior Commissure-Posterior Commissure (AC-PC) correction, skull-stripping, and cerebellum removal. Specifically, for the AC-PC correction, we used MIPAV software to resample images to 256 × 256 × 256, and applied N3 algorithm [27] for non-uniform tissue intensity correction. Skull-stripping [28] was then performed, followed by cerebellum removal. We visually checked the skull-stripped images to ensure clean and dura removal. We next employed FAST of the FSL package (http://fsl.fmrib.ox.ac.uk/fsl/fslwiki/) to segment the MR images into three tissue types: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We used HAMMER [29] to spatially normalized all three tissues onto a standard space, based on a brain atlas aligned with the MNI coordinate space. Next, we generated the regional volumetric maps, called RAVENS maps, using a tissue preserving image warping method [30]. In this study, we considered only the spatially normalized GM densities (GMD), due to its relatively high relevance to AD compared to WM and CSF [31]. Finally, we downsized the GMD maps to 32 × 32 × 32 voxels. Downsizing is for estimation and computational convenience, as it would considerably reduce the number of unknown parameters and save computation time and cost. It is a tradeoff and admittedly does lose some image information; however, our results and previous studies [32] suggest that the sacrifice in prediction is relatively limited.

III. Method

A. Model

Let $\boldsymbol{Y} = (Y_1, \ldots, Y_q)^\top \in \mathbb{R}^q$ denote a vector of q responses. For our AD data, $q = 2$, and $\boldsymbol{Y} = (Y_1, Y_2)^\top$, where Y_1 = MMSE and Y_2 = ADAS-Cog. Let $\boldsymbol{X} \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ denote a D-way tensor predictor. For our AD data, $D = 3$, $p_1 = p_2 = p_3 = 32$, and $\boldsymbol{X}$ denotes MRI scan. We propose the following multivariate tensor regression model,

$$\boldsymbol{Y} = \begin{bmatrix} \langle \boldsymbol{B}_1, \boldsymbol{X} \rangle \\ \vdots \\ \langle \boldsymbol{B}_q, \boldsymbol{X} \rangle \end{bmatrix} + \boldsymbol{e}, \quad (1)$$

where $\boldsymbol{B}_j \in \mathbb{R}^{p_1 \times \cdots \times p_D}$, $j = 1, \ldots, q$, denotes the tensor coefficient that captures *association* between the tensor predictor $\boldsymbol{X}$ and the jth response Y_j. The inner product $\langle \boldsymbol{B}_j, \boldsymbol{X} \rangle$ between two tensors $\boldsymbol{B}_j$ and $\boldsymbol{X}$ is defined as

$$\langle \boldsymbol{B}_j, \boldsymbol{X} \rangle = \mathrm{vec}(\boldsymbol{B}_j)^\top \mathrm{vec}(\boldsymbol{X}) = \sum_{i_1,\ldots,i_D} \beta_{j i_1,\ldots,i_D} x_{i_1,\ldots,i_D},$$

where vec($\boldsymbol{X}$) denotes a tensor operator that stacks the entries of $\boldsymbol{X}$ into a column vector, and $\beta_{j i_1, \ldots, i_D}$, $x_{i_1, \ldots, i_D}$ denotes the $(i_1, \ldots, i_D)$th element of $\boldsymbol{B}_j$ and $\boldsymbol{X}$, respectively. $\boldsymbol{e} = (e_1, \ldots, e_q)^\top \in \mathbb{R}^q$ denotes a vector of q errors, each of which follows a normal distribution with zero mean and constant variance. Without loss of generality, we omit the intercept term in (1).

The tensor coefficients $\{\boldsymbol{B}_j\}_{j=1}^q$ in (1) are the parameters of interest and require estimation given the observed data. If imposing no additional constraint, the total number of unknown parameters in $\boldsymbol{B}_j$, which equals $\prod_{d=1}^D p_d$, is prohibitive. For instance, for our AD data, there are $32^3 = 32{,}768$ parameters to estimate for each $\boldsymbol{B}_j$, $j = 1, 2$, while the sample size is only $n = 194$. To alleviate the extremely high dimensionality, [26] introduced a *low-rank* structure on the coefficient tensor $\boldsymbol{B}_j$ that substantially reduces the number of unknown parameters. Specifically, a D-way tensor $\boldsymbol{B}_j \in \mathbb{R}^{p_1 \times \cdots \times p_D}$ is said to follow a rank-R CANDECOMP/PARAFAC (CP) decomposition [33], if

$$\boldsymbol{B}_j = \sum_{r=1}^R \boldsymbol{\beta}_{j1}^{(r)} \circ \cdots \circ \boldsymbol{\beta}_{jD}^{(r)}, \quad (2)$$

where $\boldsymbol{\beta}_{jd}^{(r)} \in \mathbb{R}^{p_d}$ are all column vectors, $d = 1, \ldots, D$, $r = 1, \ldots, R$, and $\circ$ denotes an outer product among vectors. For convenience, the CP decomposition is often represented by a shorthand, $\boldsymbol{B}_j = [\![\boldsymbol{B}_{j1}, \ldots, \boldsymbol{B}_{jD}]\!]$, where $\boldsymbol{B}_{jd} = (\boldsymbol{\beta}_{jd}^{(1)}, \ldots, \boldsymbol{\beta}_{jd}^{(R)}) \in \mathbb{R}^{p_d \times R}$ for $d = 1, \ldots, D$. With this low-rank decomposition, the number of unknown parameters in $\boldsymbol{B}_j$ decreases substantially from the order of $\prod_{d=1}^D p_d$ to that of $R \times \sum_{d=1}^D p_d$. For the $32 \times 32 \times 32$ MRI image in the AD example, the dimensionality reduces from 32,768 to the order of 96 for a rank-1 model, and 288 for a rank-3 model.

Introducing this CP decomposition to $\{\boldsymbol{B}_j\}_{j=1}^q$, model (1) becomes

$$\boldsymbol{Y} = \begin{bmatrix} \langle \sum_{r=1}^R \boldsymbol{\beta}_{11}^{(r)} \circ \cdots \circ \boldsymbol{\beta}_{1D}^{(r)}, \boldsymbol{X} \rangle \\ \vdots \\ \langle \sum_{r=1}^R \boldsymbol{\beta}_{q1}^{(r)} \circ \cdots \circ \boldsymbol{\beta}_{qD}^{(r)}, \boldsymbol{X} \rangle \end{bmatrix} + \boldsymbol{e}. \quad (3)$$

This is our *base model* for multivariate responses and tensor predictors, upon which the subsequent regularization and estimation are built.

To help concretize model (3), we consider one of its special cases when $q = 2$, $D = 2$, $R = 1$. That is, there are two response variables $\boldsymbol{Y} = (Y_1, Y_2)^\top$, a matrix-valued predictor $\boldsymbol{X}$, and $\boldsymbol{B}_j$

is assumed to follow a rank-1 CP structure such that $\boldsymbol{B}_j = [\![\boldsymbol{\beta}_{j1}, \boldsymbol{\beta}_{j2}]\!] = \boldsymbol{\beta}_{j1} \circ \boldsymbol{\beta}_{j2}$, $j = 1, 2$. Then (3) reduces to

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} (\boldsymbol{\beta}_{12} \otimes \boldsymbol{\beta}_{11})^\top \mathrm{vec}(\boldsymbol{X}) \\ (\boldsymbol{\beta}_{22} \otimes \boldsymbol{\beta}_{21})^\top \mathrm{vec}(\boldsymbol{X}) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} = \begin{bmatrix} \boldsymbol{\beta}_{11}^\top \boldsymbol{X} \boldsymbol{\beta}_{12} \\ \boldsymbol{\beta}_{21}^\top \boldsymbol{X} \boldsymbol{\beta}_{22} \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix} \quad (4)$$

where $\otimes$ denotes the Kronecker product. The first equality in (4) comes from the fact that $\langle \boldsymbol{B}_j, \boldsymbol{X} \rangle = \langle (\boldsymbol{\beta}_{j2} \circ \boldsymbol{\beta}_{j1}), \mathrm{vec}(\boldsymbol{X}) \rangle$, when the matrix $\boldsymbol{B}_j$ admits a rank-1 CP decomposition (2). Furthermore, the Kronecker product equals the outer product when $\boldsymbol{\beta}_{j1}$ and $\boldsymbol{\beta}_{j2}$ are both column vectors. The second equality in (4) holds because $\boldsymbol{\beta}_{j1}^\top \boldsymbol{X} \boldsymbol{\beta}_{j2} = \mathrm{vec}(\boldsymbol{\beta}_{j1}^\top \boldsymbol{X} \boldsymbol{\beta}_{j2}) = (\boldsymbol{\beta}_{j2}^\top \otimes \boldsymbol{\beta}_{j1}^\top) \mathrm{vec}(\boldsymbol{X})$, $j = 1, 2$. Examining model (4), we see that our proposed multivariate response tensor regression model in this special case essentially postulates that the relation between the jth response variable Y_j and the matrix predictor $\boldsymbol{X}$ is in the form of left multiplying a coefficient vector $\boldsymbol{\beta}_{j1}^\top$ then right multiplying another coefficient vector $\boldsymbol{\beta}_{j2}$ with the matrix image $\boldsymbol{X}$. This relation is a natural extension of the classical linear model when $\boldsymbol{X}$ is a vector and the response-predictor relation is governed by $\boldsymbol{\beta}^\top \boldsymbol{X}$.

B. Penalized Likelihood

Imposing the CP low-rank structure on the coefficient tensor $\boldsymbol{B}_j$ substantially reduces the ultrahigh dimensionality of model (1) to a manageable level, leading to feasible estimation and prediction. However, the resulting number of unknown parameters can still be much larger than the available sample size. For instance, for our AD example, imposing a rank-3 CP structure yields $576 = 2 \times 3 \times (32 + 32 + 32)$ parameters, and the sample size is merely $n = 194$. Moreover, model (3) itself treats the components of multivariate response $Y_1, \ldots, Y_q$ *separately*, while it is commonly conceived that the multivariate outcomes, i.e., MMSE and ADAS-Cog in this work, are *correlated* and often provide *complementary* information. Regularization through penalized estimation is particularly useful to both handle the small-n-large-p challenge and to incorporate potential correlations among the response variables. Therefore, we further introduce regularization into our multivariate response tensor regression model (3), and propose penalized likelihood estimation.

Given n independent and identically distributed sample observations $\{(\boldsymbol{X}_1, \boldsymbol{Y}_1), \ldots, (\boldsymbol{X}_n, \boldsymbol{Y}_n)\}$, we propose to minimize the following objective function over $\{\boldsymbol{B}_j\}_{j=1}^q$,

$$\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q) = \boldsymbol{L}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q) + \lambda \boldsymbol{J}(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q). \quad (5)$$

In this function, the first component $\boldsymbol{L}$ is the usual negative log likelihood after imposing the CP structure, i.e.,

$$\boldsymbol{L}(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_q)=\sum_{i=1}^{n}\sum_{j=1}^{q}(Y_{ij}-\langle\boldsymbol{B}_j,\boldsymbol{X}_i\rangle)^2=\sum_{i=1}^{n}\sum_{j=1}^{q}\left(Y_{ij}-\langle\sum_{r=1}^{R}\boldsymbol{\beta}_{j1}^{(r)}\circ\cdots\circ\boldsymbol{\beta}_{jD}^{(r)},\boldsymbol{X}_i\rangle\right)^2,$$

where Y_{ij} is the jth response variable of subject i. The second component $\boldsymbol{J}$ in the objective (5) is a penalty function, which could have multiple choices. Here, we choose the *group lasso* penalty [34], which, in our context, takes the form,

$$\boldsymbol{J}(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_q)=\sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=1}^{p_d}\left(\sum_{j=1}^{q}\beta_{jdk}^{(r)\,2}\right)^{1/2}, \quad (6)$$

where $\beta_{jdk}^{(r)}$ is the kth element of $\boldsymbol{\beta}_{jd}^{(r)}$ in the CP decomposition (2). By imposing such a penalty, the individual coefficients $\beta_{jdk}^{(r)}$ that correspond to same subregion in an image but across different response variables are penalized as a *group*. As such, (6) encourages a subregion to drop out as a group if it is not associated with any of the multivariate response variables, which in effect takes into account potentially correlated responses. To further illustrate how the low-rank structure works and how the sparsity is imposed by this group penalty function, we consider a special case when $q = 2$, $D = 2$, $R = 1$, and $p_1 = p_2 = 4$. Fig. 2 shows that, for each response variable and its associated coefficient $\boldsymbol{B}_j$, $j = 1, 2$, element-wise sparsity in the column vectors $\boldsymbol{\beta}_{jd}^{(1)}$, $j = 1, 2$, $d = 1, 2$ translates into region-wise sparsity in the coefficient matrix $\boldsymbol{B}_j$. Meanwhile, the group penalty (6) encourages that the subset of elements in $\{\boldsymbol{\beta}_{1d}^{(1)},\boldsymbol{\beta}_{2d}^{(1)}\}$, $d = 1, 2$, that correspond to the same region but across different response variables would enter or drop from the model simultaneously. A similar group penalty was also used in the classical multi-response model [23]. It encourages identification of the same subregions of the coefficient images $\{\boldsymbol{B}_j\}_{j=1}^{q}$ across different responses. Meanwhile, it permits different magnitude of $\{\boldsymbol{B}_j\}_{j=1}^{q}$, i.e., strength of association, for different responses.

Algorithm 1

Algorithm for minimizing $\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q)$.

Initialize $\boldsymbol{B}_{jd} \in \mathrm{IR}^{p_d \times R}$ as a random matrix, $j = 1, \ldots, q$, $d = 1, \ldots, D$.

repeat

for $d = 1, \ldots, D$ **do**

Update $\boldsymbol{\mu}_d^{(t+1)}$, given $\boldsymbol{B}_1^{(t)},\ldots,\boldsymbol{B}_q^{(t)}$.

end for

for $j = 1, \ldots, q$ **do**

Update $\boldsymbol{B}_j^{(t+1)} = \arg\min_{\boldsymbol{B}_j} \ell(\boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_{j-1}^{(t+1)}, \boldsymbol{B}_j, \boldsymbol{B}_{j+1}^{(t)}, \ldots, \boldsymbol{B}_q^{(t)})$,

given $\boldsymbol{\mu}_1^{(t+1)}, \ldots, \boldsymbol{\mu}_D^{(t+1)}$.

end for

until $\ell(\boldsymbol{B}_1^{(t+1)}, \ldots, \boldsymbol{B}_q^{(t+1)})$ converges.

C. Estimation

Next, we investigate optimization of the objective function $\ell(\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q)$ in (5). We first summarize the optimization procedure in Algorithm 1, then present details for individual steps. The optimization is achieved through the variational method [25] based on the following result,

$$\min_{c, c>0} \frac{1}{2}(cx^2 + \frac{1}{c}) = |x|,$$

when $c = |x|^{-1}$, $x \neq 0$. Consequently, minimizing (5) over $\boldsymbol{B}_1, \ldots, \boldsymbol{B}_q$ is equivalent to minimizing the following objective function

$$\sum_{i=1}^{n}\sum_{j=1}^{q}(Y_{ij} - \langle \boldsymbol{B}_j, \boldsymbol{X}_i \rangle)^2 + \frac{\lambda}{2}\sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=1}^{p_d}\left(\mu_{dk}^{(r)} \|\boldsymbol{b}_{dk}^{(r)}\|^2 + \frac{1}{\mu_{dk}^{(r)}}\right), \quad (7)$$

over both $\boldsymbol{B}_j$ and $\mu_{dk}^{(r)}$, where $\boldsymbol{b}_{dk}^{(r)} = [\beta_{1dk}^{(r)}, \ldots, \beta_{qdk}^{(r)}] \in \mathbb{R}^q$. Optimization of (7) can then be achieved in an alternating fashion, updating iteratively with one set of parameters renewed and the others fixed.

Specifically, with $\boldsymbol{B}_j$ fixed, the update of $\mu_{dk}^{(r)}$ is simply

$$\mu_{dk}^{(r)} = \frac{1}{\|\boldsymbol{b}_{dk}^{(r)}\|}.$$

With $\mu_{dk}^{(r)}$ fixed, the update of $\boldsymbol{B}_j$ is achieved through a block relaxation algorithm [26]. That is, by imposing the CP structure on $\boldsymbol{B}_j = [\![\boldsymbol{B}_{j1}, \ldots, \boldsymbol{B}_{jD}]\!]$, the first part of (7) can be written as

$$\sum_{i=1}^{n}\sum_{j=1}^{q}(Y_{ij} - \langle \boldsymbol{B}_{jd}, \boldsymbol{X}_{i(d)}(\boldsymbol{B}_{jD} \odot \cdots \odot \boldsymbol{B}_{j,d+1} \odot \boldsymbol{B}_{j,d-1} \odot \cdots \odot \boldsymbol{B}_{j1}) \rangle)^2,$$

where $\boldsymbol{X}_{i(d)}$ denotes the mode-d matricization that maps tensor $\boldsymbol{X}_i$ into a $p_d \times \prod_{d' \neq d}$ matrix such that the $(i_1, \ldots, i_D)$th element of $\boldsymbol{X}_i$ maps to the $(i_d, 1 + \sum_{d' \neq d}(i_{d'} - 1) \prod_{d'' < d', d'' \neq d}$

$p_{d''}$)th element of $\boldsymbol{X}_{i(d)}$, and $\odot$ denotes the Khatri-Rao product [35]. This reformulation allows one to focus the estimation $\boldsymbol{B}_{jd}$ on while keeping all the other parameters fixed. Meanwhile, the second part of (7) can be written as

$$\frac{\lambda}{2}[\sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=1}^{p_d}\frac{1}{\mu_{dk}^{(r)}}+\sum_{j=1}^{q}\sum_{d=1}^{D}(\mathrm{vec}(\boldsymbol{B}_{jd}))^{\top}\ \mathrm{diag}\ \left(\mu_{d1}^{(1)},\ldots,\mu_{dp_d}^{(1)},\ldots,\mu_{d1}^{(R)},\ldots,\mu_{dp_d}^{(R)}\right)\ \mathrm{vec}(\boldsymbol{B}_{jd})].$$

Therefore, the objective in (7) is essentially a quadratic function of individual $\boldsymbol{B}_{jd}$ when all the other parameters are fixed. There is a closed form solution for $\boldsymbol{B}_{jd}$, such that vec($\boldsymbol{B}_{jd}$) equals

$$\left(\sum_{i=1}^{n}\tilde{\boldsymbol{X}}_{ijd}\tilde{\boldsymbol{X}}_{ijd}^{\top}+\frac{\lambda}{2}\mathrm{diag}\ \left(\mu_{d1}^{(1)},\ldots,\mu_{dp_d}^{(1)},\ldots,\mu_{d1}^{(R)},\ldots,\mu_{dp_d}^{(R)}\right)\right)^{-1}$$
$$\sum_{i=1}^{n}Y_{ij}\tilde{\boldsymbol{X}}_{ijd},$$

where

$$\tilde{\boldsymbol{X}}_{ijd}=\mathrm{vec}(\boldsymbol{X}_{i(d)}(\boldsymbol{B}_{jD}\odot\cdots\odot\boldsymbol{B}_{j,d+1}\ \odot\boldsymbol{B}_{j,d-1}\odot\cdots\odot\boldsymbol{B}_{j1}))\in\mathbb{R}^{Rp_d}.$$

We make a few remarks about the above optimization procedure. First, although the objective value decreases monotonically through iterations, the convergence to a global optimum is *not* guaranteed, since (5) involves a nonconvex optimization and there exist potentially multiple local minima. We adopt the common practice of using multiple starting values. In our setup, the stability of the algorithm with respect to initial values depends on several factors. A large sample size, a stronger signal strength, and a low-rank true image signal would all foster fast convergence and increase the chance to locate the global optimum from different initializations. In Section IV-B, we report the numerical convergence behavior of our algorithm with multiple starting values. Second, the computation of our method is fast, since both steps of iterations have closed form solutions. Actually the computational complexity for each iteration is $O(np^2q+p^3q)$ for $D\leq 2$, and $O(np^Dq)$ for $D>2$, when $p_1=\cdots=p_D=p$. In addition, since $\tilde{\boldsymbol{X}}_{ijd}$ only depends on $\boldsymbol{B}_{jd'}$, $d'\neq d$, $\{\boldsymbol{B}_{1d},\ldots,\boldsymbol{B}_{qd}\}$ can be updated simultaneously within each iteration. Again in Section IV-B, we report the computation time in simulations. Third, we note that the variational method rarely produces estimates that are exactly zero in practice. Consequently, we set a thresholding value for $\|\boldsymbol{b}_{dk}^{(r)}\|$ to achieve the desired sparsity, which is a common practice in the applications of the variational method [25]. Last but not least, in addition to the variational method, one may also consider using the alternating direction method of multipliers (ADMM) for solving the optimization of (5). We have experimented with ADMM, and found it produced similar results as the variational method, but was slower. This is partly due to that, within each block update, our problem simplifies into minimizing a quadratic function plus a group lasso penalty. The variational method can further simplify the problem by optimization over $\{\boldsymbol{B}_{1d},\ldots,\boldsymbol{B}_{qd}\}$ *separately*, whereas ADMM cannot and

has to construct relatively large matrices involving $\{\boldsymbol{B}_{1d}, \ldots, \boldsymbol{B}_{qd}\}$ jointly. For that reason, we choose the variational method as our optimization solution.

IV. Results

In this section, we first carry out Monte Carlo simulations to investigate the empirical performance of our proposed method. We then investigate its stability, convergence and computation time. Finally, we analyze the ADNI dataset to illustrate the efficacy of the new method.

A. Signal Recovery and Prediciton

We evaluate the empirical performance by two criteria: the prediction accuracy of the responses measured by root mean squared error (RMSE), and the estimation accuracy of the tensor coefficient shown by a plot. We compare our method with two alternative solutions. The first is to fit a tensor regression model for one response at a time. A comparison with this method would clearly show the gain of our method that jointly models the multivariate responses. The second is to vectorize the image predictor, ignoring all potential correlations among the image voxels, then fit a mluti-response linear regression model with a group lasso penalty [23]. This comparison would show the gain of our method that respects the image tensor structure. For abbreviation, we call our method and the two alternatives as "Multi-Resp", "Uni-Resp", and "Vectorized", respectively.

The data are simulated from model (3), with a 64 × 64 matrix predictor $\boldsymbol{X}$ whose entries independently follow a normal distribution, and q coefficient matrices $\boldsymbol{B}_j \in \mathbb{R}^{64\times 64}$, $j = 1, \ldots, q$, which take the value of 0 or 1 following specified patterns. We consider two scenarios: (i) We let all $\boldsymbol{B}_j$ follow the same pattern of "cross", "triangle", and "butterfly", respectively. (ii) We let half of $\boldsymbol{B}_j$'s take the shape of "cross", and the other half "triangle". Among the three patterns, "cross" is of an exact rank 2, while "triangle" and "butterfly" are of infinitely high rank, whereas we use a fixed rank 3 to approximate all three patterns. We then generate q response variables following (3), with the errors $\boldsymbol{e} \in \mathbb{R}^q$ following a normal distribution with mean zero and covariance $\sigma^2\boldsymbol{\Sigma}$, where $\boldsymbol{\Sigma} \in \mathbb{R}^{q\times q}$ has diagonal elements equal to 1 and off-diagonals equal to ρ. Consequently, the pairwise correlation among the q responses is governed by ρ, while σ^2 controls the relative noise level. We examine a series of values of q, σ^2 and ρ to investigate the empirical performance of different methods under varying response correlation strength and noise level. We marginally standardize the response variables, by subtracting mean and dividing by standard deviation. The models are fitted on a training set of size $n = 750$, tuned on an independent validation set, and evaluated on another independent testing set of the same size.

Table II shows RMSE of prediction of "future" responses in the testing data under 50 Monte Carlo replications, whereas Fig. 3 gives a graphical summary of the estimated coefficient signal based on one data replication under the scenario (i). Since the underlying true patterns are the same for all responses here, we only report the estimator for the first one. We present in the table the results when $q = 3, 9$, $\sigma = 10, 15, 20$, and $\rho = 0, 0.9$, respectively, but in the figure omit the results when $\rho = 0.9$, since they are visually similar to those of $\rho = 0$. We also omit in the figure the case $q = 9$ for "Uni-Resp", because it fits each response variable

separately, and thus the coefficient estimate is not affected by the number of responses q. It is clearly seen that our proposed method outperforms the univariate solution. The difference is more dominant when the signal is relatively more noisy, as one would often encounter in real imaging data, and when the number of response variables is large, as they provide more complementary information. In addition, both the multivariate and univariate tensor regression solutions have produced a much better estimation than the vectorized solution, which fails to identify any meaningful patterns.

Similarly, Table III and Fig. 4 provide a summary of the results for the scenario (ii). Again, the proposed method clearly outperforms the two competitors. If one assumes the resulting criteria from multiple replications are normally distributed, then the two-sample t-test would yield significant differences between "Multi-Resp" and "Uni-Resp" (with p-values less than 0.001) for all combinations of (q, ρ, σ), except for (q, ρ, σ) = (2, 0, 10) (p-value = 0.61) and (q, ρ, σ) = (2, 0.9, 10) (p-value = 0.08). Moreover, differences between "Multi-Resp" and "Vectorized" are significant for all situations (p-values = 0). It is also noteworthy that, in most multi-response literatures, the models are of a similar form as in the scenario (i). This does not necessarily imply that all the multiple response variables must admit exactly the same association with the predictors. The magnitude of those coefficients could vary. In our simulation, for simplicity, we only let the signal $\boldsymbol{B}_j$ take the value of 0 or 1. The proposed method works best under the scenario (i), but outperforms the competing solutions under the scenario (ii) as well.

B. Stability, Convergence and Computation Time

We next investigate the stability of the algorithm with respect to the regularization parameter λ and the rank R of the CP decomposition. We adopt the setting of scenario (i), the "triangle" signal, $q = 3$, $\sigma = 10$, $\rho = 0$ and $n = 750$. Fig. 5 shows the results of applying different values of λ and R. For all combinations, the method successfully identifies the signal region. However, insufficient or excessive penalty would both adversely affect the quality of the recovered signal. The rank R of the CP decomposition essentially offers a bias-variance tradeoff. A larger rank implies a more flexible model and a smaller bias, but also more unknown parameters and thus a larger variation, whereas a smaller rank implies a more parsimonious model and a smaller estimation variability, but possibly a larger bias. In reality, the true signal tensor is hardly of an exactly low-rank structure. However, given the usually limited sample size in imaging studies, a low-rank estimate often provides a reasonable *approximation* to the true tensor regression parameter, even when the true signal is of a high rank [26]. In our analysis, we usually fix the rank at $R = 3$, which offers a good balance between model complexity and estimation accuracy.

Fig. 6 shows the convergence behavior of the algorithm, as reflected by the objective value, under 100 randomly generated starting values, and the corresponding computation time. We see that, although there exist multiple local minima, the algorithm often converges to the same or similar point. The run time is recorded on a standard laptop computer with 2.9 GHz Inter i7 CPU. For instance, the median run time of fitting a rank-3 model in this example takes about 21 seconds.

C. ADNI Data Analysis

We analyze the ADNI dataset reviewed in Section II. The analysis consists of two parts: estimation of clinical scores and identification of brain subregions that are highly relevant to the clinical outcomes.

We first aim to infer the clinical scores of MMSE and ADAS-Cog given the MRI scans. Such prediction is useful for both disease diagnosis as well as understanding of disease progression. The two clinical scores are normalized (subtracted the mean and divided by standard deviation) to avoid different response scales. A 10-fold cross-validation is performed. The rank of the coefficient tensor is set as $R = 3$, and the regularization parameter λ is optimized based solely on the training set with another nested 10-fold cross-validation. We then employ the resulting model to estimate the clinical scores on the testing data. The RMSE (the smaller the better), and the Pearson correlation coefficient (the larger the better), between the predicted and the observed clinical scores on the testing data are reported in Table IV. Moreover, we show the scatter plots of the predicted versus observed scores for MMSE and ADAS-Cog in Fig. 7.

We also compare our method ("Multi-Resp") to two sets of alternative solutions. The first set consists of the univariate response solution ("Univ-Resp"), and the vectorized solution ("Vectorized"), as reported in Section IV-A. Both methods, as well our proposed method, directly model an image tensor and jointly incorporate all voxels of an image. The second set of alternatives include the multi-task method ("M3T") comprised of feature selection via a group lasso penalty and estimation via support vector regression [10], support vector regression with feature selection via lasso ("SVR + lasso"), and without any feature selection ("SVR"). It is important to note that, this family of methods do not directly handle a tensor image, but a vector of features extracted from an image. As such, we employ the Automated Anatomical Labeling (AAL) [36] to partition the image into 90 regions of interest (ROI) and then use the average intensity of each ROI as the extracted features. The results are again reported in Table IV. We see that, our solution clearly outperforms the one that models one response at a time, demonstrating the advantage of jointly modeling the multivariate responses that are correlated and complementary. We also see that, after taking the standard error into account, our new method performs essentially as well as the best solutions in the literature such as "M3T" and "SVR + lasso". On the other hand, our method works *without* requiring a specific image atlas, and thus avoid its dependence and the choice among different atlases.

In addition to the estimation of multiple clinical scores, the proposed method can simultaneously serve as a tool to select relevant brain regions. Our multivariate method is advantageous, since different clinical scores reflect the same underlying pathology while they also offer complementary information [8], [10], [14]. For this data, we use the optimal tuning parameter from the cross-validation to fit the full dataset. The voxels selected are those with non-zero regression coefficient estimates. To help visualize the selected regions, we next partition the brain into 90 ROIs based on the AAL. Then we plot the corresponding ROIs with at least 10% of its voxels selected by our method. Ordered by the percentage of selected voxels from highest, the identified regions that are relevant to the two clinical scores are: amygdala (left and right), hippocampus (left and right), parahippocampal gyrus (right

and left), olfactory cortex (left), superior temporal gyrus (left), middle temporal gyrus (left), putamen (right and left), and insula (left). These regions have been shown by numerous studies to be highly relevant to AD. See Table V for a summary of the associated literatures. Moreover, to show the path of region selection, we repeat the same procedure with a gradually increasing sparsity tuning parameter λ. We summarize the results in Fig. 8, where we map the estimate back to the original high-resolution 256 × 256 × 256 MRI image, and Table V. It is worth noting that a smaller tuning parameter would result in a larger number of selected regions and the potential problem of overfitting. Conversely, a larger tuning parameter would result in a smaller number of selected regions and potentially underfitting.

V. Discussion

We have proposed a sparse multivariate response tensor regression model in this article. Our proposed method models multiple response variables jointly, so as to exploit the correlated and complementary information possessed in multivariate responses. It also models multiple voxels in an image tensor jointly, so as to account for inherent spatial correlation in image covariates. The method is designed to simultaneously infer multiple responses and to identify brain subregions highly relevant to the outcomes. As such it is useful for both AD/MCI diagnosis, and for locating brain regions contributing to the disease. Our numerical analyses have demonstrated that the proposed method outperformed its competitors.

There are some alternative choices within our proposed model formulation. One is to consider a different penalty function than the group lasso penalty (6), and the other is an alternative tensor regression model formulation than (3). First, an alternative to the group lasso penalty (6) for multi-response regression is the L_∞ type penalty [22]. In our context, the penalty function takes the form,

$$\boldsymbol{J}(\boldsymbol{B}_1,\ldots,\boldsymbol{B}_q)=\sum_{d=1}^{D}\sum_{r=1}^{R}\sum_{k=1}^{p_d}\left(\max_j|\beta_{jdk}^{(r)}|\right). \quad (8)$$

This penalty, similar to the group lasso penalty, also induces row-wise sparsity to the regression parameters. However, it differs from the group lasso penalty in that, the L_∞ penalty selects predictors based on their *maximum* contribution to any of the response variables, whereas the group lasso penalty selects predictors based on their *joint* contribution to all of the response variables. As a result, the L_∞ penalty tends to select more variables than the group lasso penalty. The optimization with the L_∞ penalty (8) is a linearly constrained quadratic optimization problem, which can be solved by an interior-point algorithm [45].

Second, an alternative to the multi-response tensor regression model (3) is the model

$$\boldsymbol{Y}=\boldsymbol{B}_{(D+1)}\mathrm{vec}(\boldsymbol{X})+\boldsymbol{e}, \quad (9)$$

where $\boldsymbol{B}$ is a $(D+1)$-dimensional tensor that is assumed to admit a rank-R CP decomposition, and $\boldsymbol{B}_{(D+1)}$ denotes its mode-$(D+1)$ matricization. That is, $\boldsymbol{B} = [\![\boldsymbol{B}_1, \ldots, \boldsymbol{B}_{D+1}]\!]$ where $B_d = [\beta_d^{(1)}, \ldots, \beta_d^{(R)}] \in \mathbb{R}^{p_d \times R}$, for $d = 1, \ldots, D$, and $\boldsymbol{B}_{D+1} \in \mathbb{R}^{q \times R}$. To better understand this model, we again consider its special case when $q = 2$, $D = 2$, $R = 1$, i.e., two response variables with a matrix image predictor. In this case, $\boldsymbol{B}$ is a $p_1 \times p_2 \times 2$ tensor that admits a rank-1 decomposition, $\boldsymbol{B} = [\![\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3]\!] = \boldsymbol{\beta}_1 \circ \boldsymbol{\beta}_2 \circ \boldsymbol{\beta}_3$, and $\boldsymbol{\beta}_3 = [\beta_{31}, \beta_{32}]^\top$. Then model (9) becomes

$$\begin{bmatrix} Y_1 \\ Y_2 \end{bmatrix} = \begin{bmatrix} \beta_{31}(\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1)^\top \mathrm{vec}(\boldsymbol{X}) \\ \beta_{32}(\boldsymbol{\beta}_2 \otimes \boldsymbol{\beta}_1)^\top \mathrm{vec}(\boldsymbol{X}) \end{bmatrix} + \begin{bmatrix} e_1 \\ e_2 \end{bmatrix}. \quad (10)$$

A few remarks are in order for the comparison of model (10) with model (4), which is a special case of model (3) under the same setup of $q = 2$, $D = 2$, $R = 1$. While model (4) permits different coefficient vectors $\boldsymbol{\beta}_{11} \otimes \boldsymbol{\beta}_{12}$ and $\boldsymbol{\beta}_{21} \otimes \boldsymbol{\beta}_{22}$ for different response variables Y_1 and Y_2, model (10) imposes the same coefficient vectors $\boldsymbol{\beta}_1 \otimes \boldsymbol{\beta}_2$ except for allowing a different scalar β_{31} and β_{32} for different Y's. Consequently, model (4) enjoys more flexibility, while (10) requires fewer number of parameters and thus induces less estimation variability. For our problem, we note that, since the scores of MMSE and ADAS-Cog carry different levels of information with respect to the cognitive capability, it is more reasonable to assign different coefficients in predicting the two clinical scores. This is the reason we have primarily focused on model (3).

Finally, an alternative tensor decomposition, the Tucker decomposition [33], can be employed in our solution. The Tucker decomposition is more flexible than the CP decomposition, by allowing different number of factors along each mode of the tensor. However, it may introduce a larger number of unknown parameters than CP and require more parameter tunings. As such we have chosen the CP decomposition in this article.

Acknowledgments

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904). This work was supported in part by the National Science Foundation under Grant AG049371, and Grant AG042599. The work of H.-I. Suk was supported by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (NRF-2015R1C1A1A01052216). The work of D. Shen was supported by the NIH under Grant EB006733, Grant EB008374, Grant EB009634, MH100217, Grant AG041721, Grant AG049371, and Grant AG042599.

ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Abbott, AstraZeneca AB, Bayer Schering Pharma AG, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, GlaxoSmithKline, Innogenetics, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Schering-Plough, Synarc, Inc., as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the U.S. Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuron Imaging at the University of California, Los Angeles.

References

1. Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM. Forecasting the global burden of Alzheimer's disease. Alzheimer's Dementia. 2007; 3(3):186–191.
2. Petersen RC, et al. Mild cognitive impairment, clinical characterization and outcome. Arch. Neurol. 1999; 56(3):303–308. [PubMed: 10190820]
3. Fox MD, Greicius M. Clinical applications of resting state functional connectivity. Front. Syst. Neurosci. 2010; 4:19. [PubMed: 20592951]
4. Nordberg A, Rinne JO, Kadir A, Långström B. The use of PET in Alzheimer disease. Nature Rev. Neurol. 2010; 6(2):78–87. [PubMed: 20139997]
5. Cuingnet R, et al. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. NeuroImage. 2011; 56(2):766–781. [PubMed: 20542124]
6. Duchesne, S.; Caroli, A.; Geroldi, C.; Frisoni, GB.; Collins, DL. Medical Image Computing and Comput.-Assisted Intervention-MICCAI 2005, ser. LNCS. Vol. 3749, ch. 49. Berlin, Germany: Springer; 2005. Predicting clinical variable from MRI features: Application to MMSE in MCI; p. 392-399.
7. Duchesne S, Caroli A, Geroldi C, Collins DL, Frisoni GB. Relating one-year cognitive change in mild cognitive impairment to baseline MRI features. NeuroImage. 2009; 47(4):1363–1370. [PubMed: 19371783]
8. Stonnington CM, et al. Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. NeuroImage. 2010; 51(4):1405–1413. [PubMed: 20347044]
9. Wang Y, Fan Y, Bhatt P, Davatzikos C. High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. NeuroImage. 2010; 50(4):1519–1535. [PubMed: 20056158]
10. Zhang D, et al. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage. 2012; 59(2):895–907. [PubMed: 21992749]
11. Zhu X, Suk H-I, Shen D. A novel matrix-similarity based loss function for joint regression and classification in AD diagnosis. NeuroImage. 2014; 100:91–105. [PubMed: 24911377]
12. Folstein MF, Folstein SE, McHugh PR. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. J Psychiatric Res. 1975; 12(3):189–198.
13. Rosen WG, Mohs RC, Davis KL. A new rating scale for Alzheimer's disease. Am. J. Psychiatry. 1984
14. Fan Y, Kaufer D, Shen D. Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. Proc. IEEE Int. Symp. Biomed. Imag. From Nano to Macro. 2010:852–855.
15. Helland IS. Partial least squares regression and statistical models. Scand. J. Stat. 1990:97–114.
16. Helland IS. Maximum likelihood regression on relevant components. J. R Stat. Soc B. 1992:637–647.
17. Chun H, Keleş S. Sparse partial least squares regression for simultaneous dimension reduction and variable selection. J. R Stat. Soc. B. 2010; 72(1):3–25.
18. Zhou J, He X. Dimension reduction based on constrained canonical correlation and variable filtering. Ann. Stat. 2008; 36:1649–1668.
19. Izenman AJ. Reduced-rank regression for the multivariate linear model. J Multivariate Anal. 1975; 5(2):248–264.
20. Velu, R.; Reinsel, GC. Multivariate Reduced-Rank Regression: Theory and Applications. Vol. 136. New York: Springer; 2013.
21. Yuan M, Ekici A, Lu Z, Monteiro R. Dimension reduction and coefficient estimation in multivariate linear regression. J. R Stat. Soc. B. 2007; 69(3):329–346.
22. Turlach BA, Venables WN, Wright SJ. Simultaneous variable selection. Technometrics. 2005; 47(3):349–363.

23. Similä T, Tikka J. Input selection and shrinkage in multiresponse linear regression. Comput. Statist. Data Anal. 2007; 52(1):406–422.
24. Peng J, et al. Regularized multivariate regression for identifying master predictors with application to integrative genomics study of breast cancer. Ann. Appl. Stat. 2010; 4(1):53. [PubMed: 24489618]
25. Chen L, Huang JZ. Sparse reduced-rank regression for simultaneous dimension reduction and variable selection. J Am. Stat. Assoc. 2012; 107(500):1533–1545.
26. Zhou H, Li L, Zhu H. Tensor regression with applications in neuroimaging data analysis. J Am. Stat. Assoc. 2013; 108(502):540–552. [PubMed: 24791032]
27. Sled JG, Zijdenbos AP, Evans AC. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans. Med. Imag. 1998 Feb.17(1):87–97.
28. Wang Y, et al. Knowledge-guided robust MRI brain extraction for diverse large-scale neuroimaging studies on humans and non-human primates. PLOS ONE. 2014; 9(1):E77810. [PubMed: 24489639]
29. Shen D, Davatzikos C. HAMMER: Hierarchical attribute matching mechanism for elastic registration. IEEE Trans. Med. Imag. 2002 Nov.21(11):1421–1439.
30. Davatzikos C, Genc A, Xu D, Resnick SM. Voxel-based morphometry using the RAVENS maps: Methods and validation using simulated longitudinal atrophy. NeuroImage. 2001; 14(6):1361–1369. [PubMed: 11707092]
31. Liu M, et al. Ensemble sparse classification of Alzheimer's disease. NeuroImage. 2012; 60(2): 1106–1116. [PubMed: 22270352]
32. Liu M, Zhang D, Shen D. Hierarchical fusion of features and classifier decisions for Alzheimer's disease diagnosis. Human Brain Map. 2014; 35(4):1305–1319.
33. Kolda TG, Bader BW. Tensor decompositions and applications. SIAM Rev. 2009; 51(3):455–500.
34. Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J. R Stat. Soc. B. 2006; 68(1):49–67.
35. Rao, CR.; Mitra, SK. Generalized Inverse of Matrices and its Applications. Vol. 7. New York: Wiley; 1971.
36. Tzourio-Mazoyer N, et al. Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. NeuroImage. 2002; 15(1):273–289. [PubMed: 11771995]
37. Misra C, Fan Y, Davatzikos C. Baseline and longitudinal patterns of brain atrophy in MCI patients, and their use in prediction of short term conversion to AD: Results from ADNI. NeuroImage. 2009; 44(4):1415–1422. [PubMed: 19027862]
38. Karas G, et al. Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. NeuroImage. 2004; 23(2):708–716. [PubMed: 15488420]
39. Convit A, et al. Atrophy of the medial occipitotemporal, inferior, and middle temporal gyri in non-demented elderly predict decline to Alzheimer’s disease. Neurobiol. Aging. 2000; 21(1):19–26. [PubMed: 10794844]
40. Chetelat G, et al. Mapping gray matter loss with voxel-based morphometry in mild cognitive impairment. Neuroreport. 2002; 13(15):1939–1943. [PubMed: 12395096]
41. De Jong L, et al. Strongly reduced volumes of putamen and thalamus in Alzheimer's disease: An MRI study. Brain. 2008; 131(12):3277–3285. [PubMed: 19022861]
42. Dai Z, et al. Discriminative analysis of early Alzheimer's disease using multi-modal imaging and multi-level characterization with multi-classifier (M3). NeuroImage. 2012; 59(3):2187–2195. [PubMed: 22008370]
43. Chetelat G, Baron J-C. Early diagnosis of Alzheimer’s disease: Contribution of structural neuroimaging. NeuroImage. 2003; 18(2):525–541. [PubMed: 12595205]
44. Devanand D, et al. Olfactory deficits in patients with mild cognitive impairment predict Alzheimer's disease at follow-up. Am. J. Psychiatry. 2000; 157(9):1399–1405. [PubMed: 10964854]
45. Friedman J, et al. Pathwise coordinate optimization. Ann. Appl. Stat. 2007; 1(2):302–332.

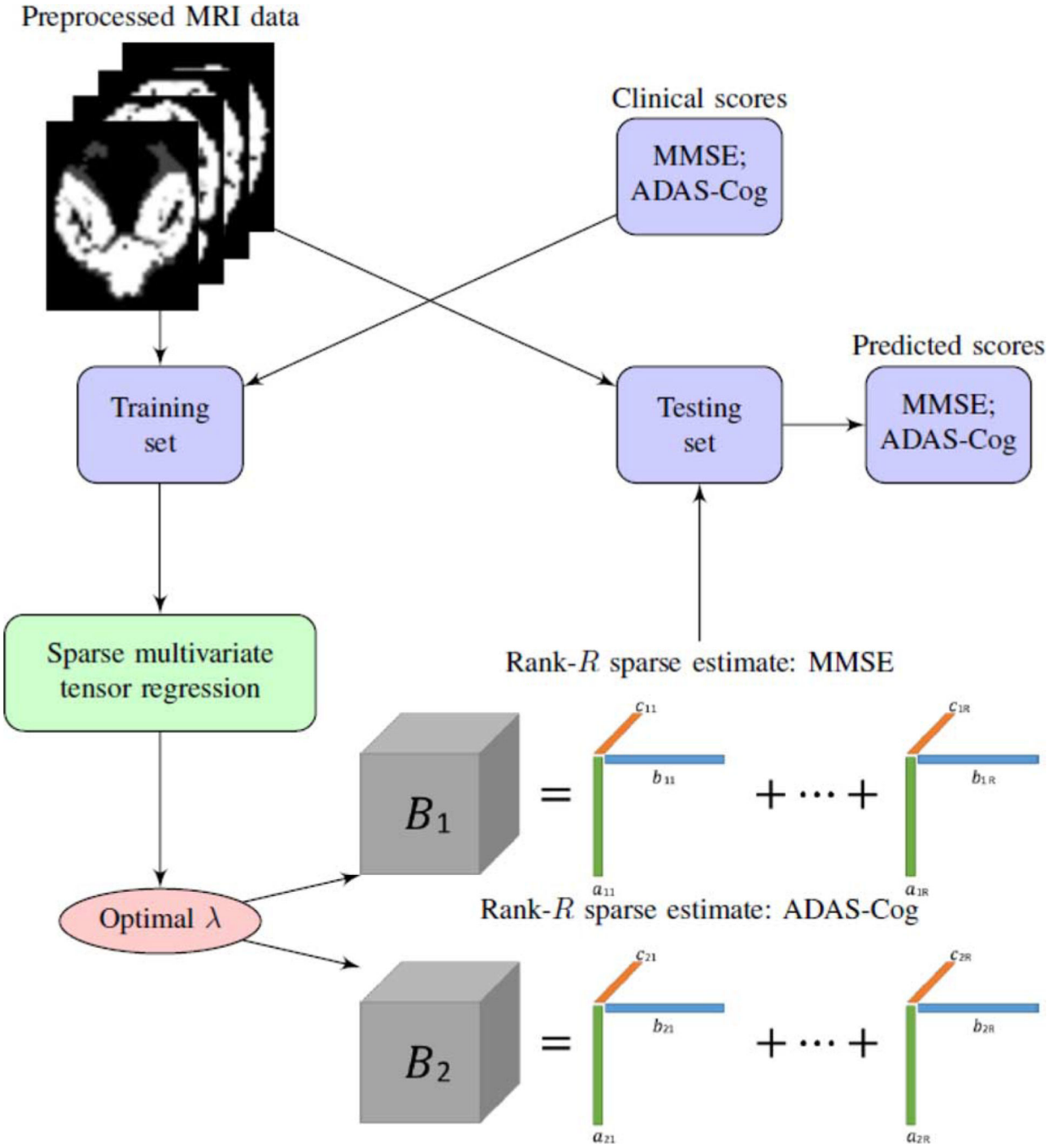

Fig. 1.

A schematic overview of the proposed sparse multi-response tensor regression with multivariate cognitive assessments.

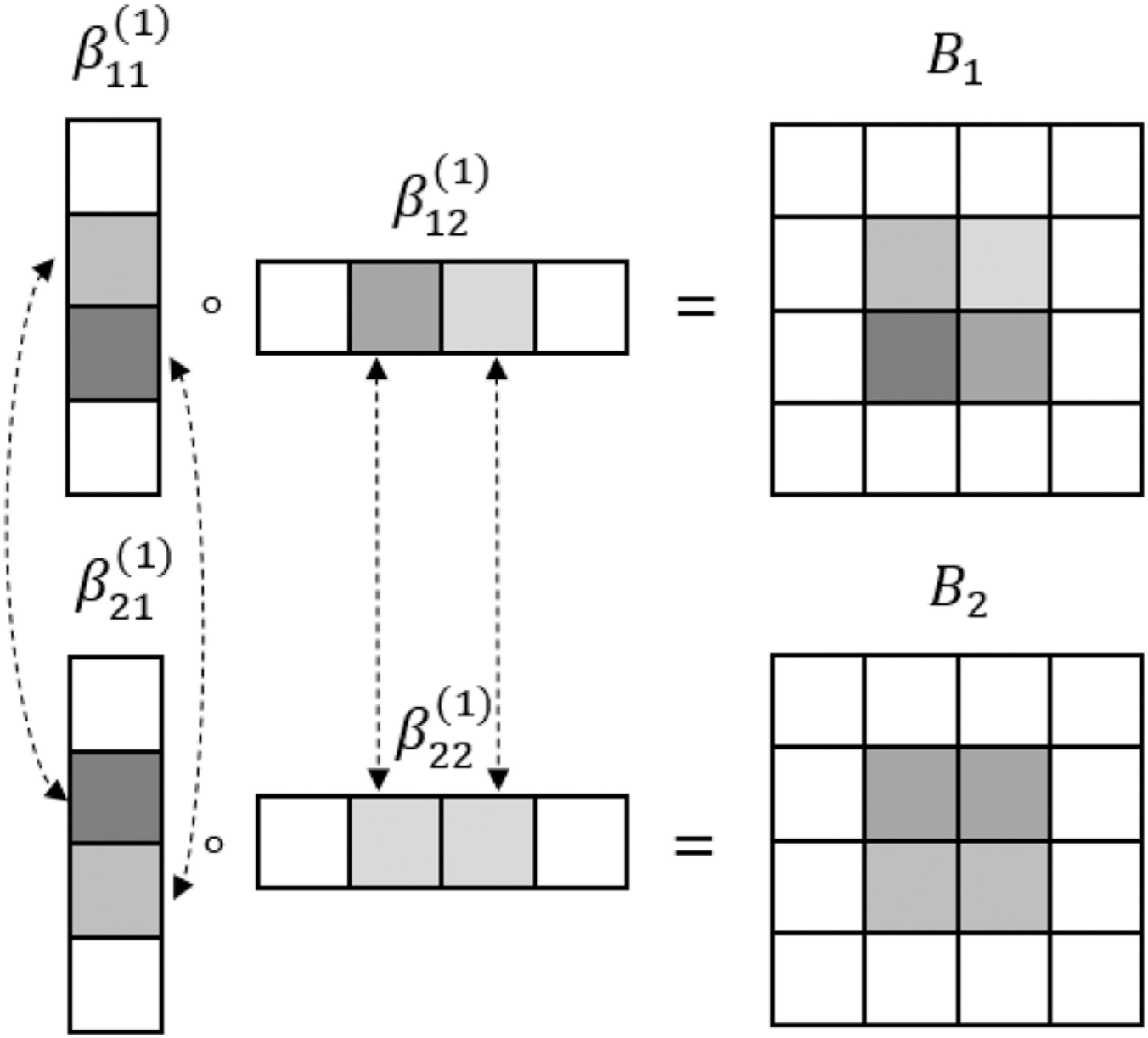

Fig. 2.

An illustration of the low-rank and sparse estimation of the coefficient signal when $q = 2$, $D = 2$, $R = 1$ and $p_1 = p_2 = 4$. ∘ denotes the outer product. Dotted lines connect the elements corresponding to the same region but across different responses, which are encouraged to enter or drop from the model simultaneously. Different colors denote different strength of association.

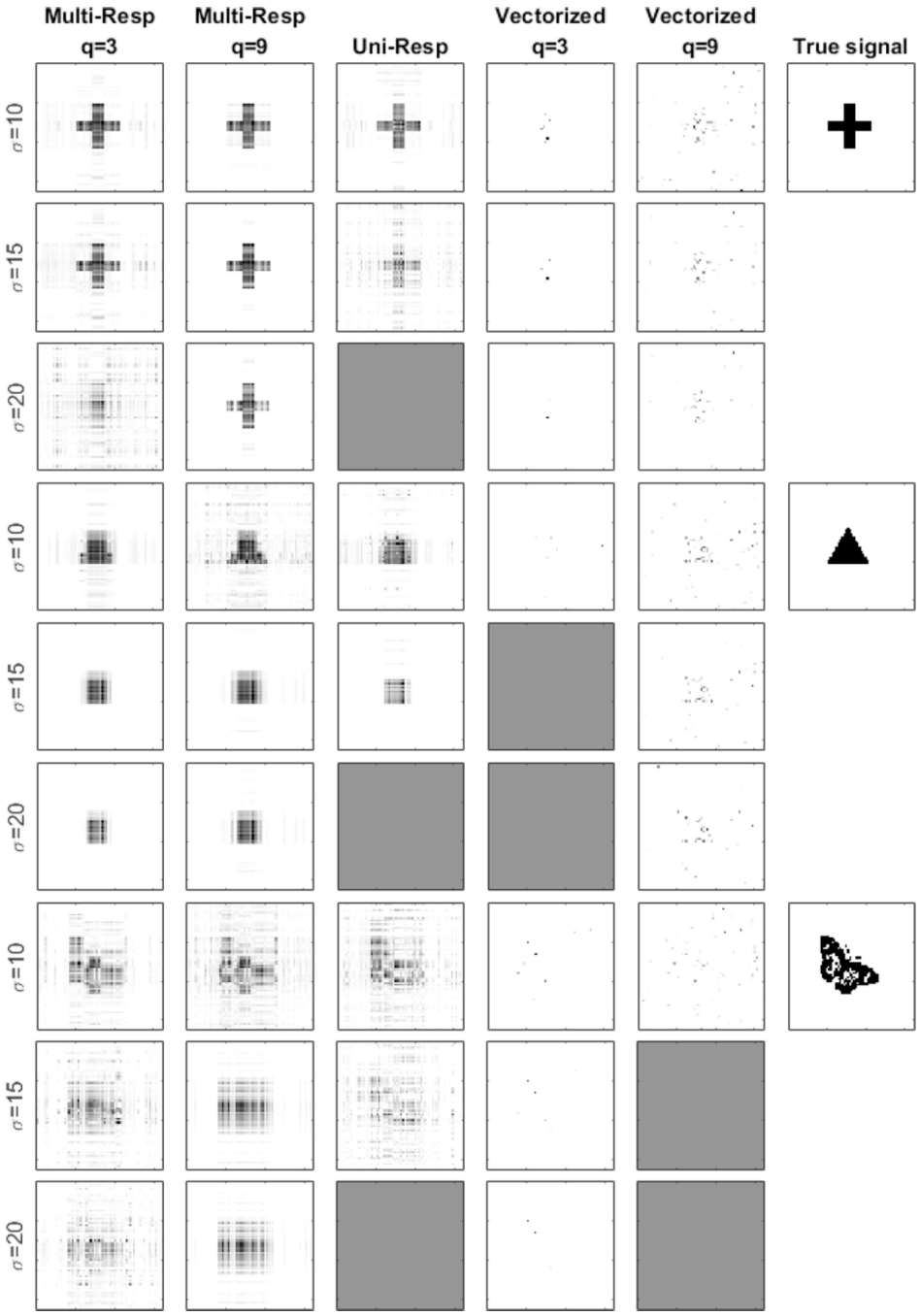

Fig. 3.

Estimated coefficient images under varying noise level (σ = 10, 15, 20) and number of response variables (q = 3, 9). The coefficient patterns are the same for all responses.

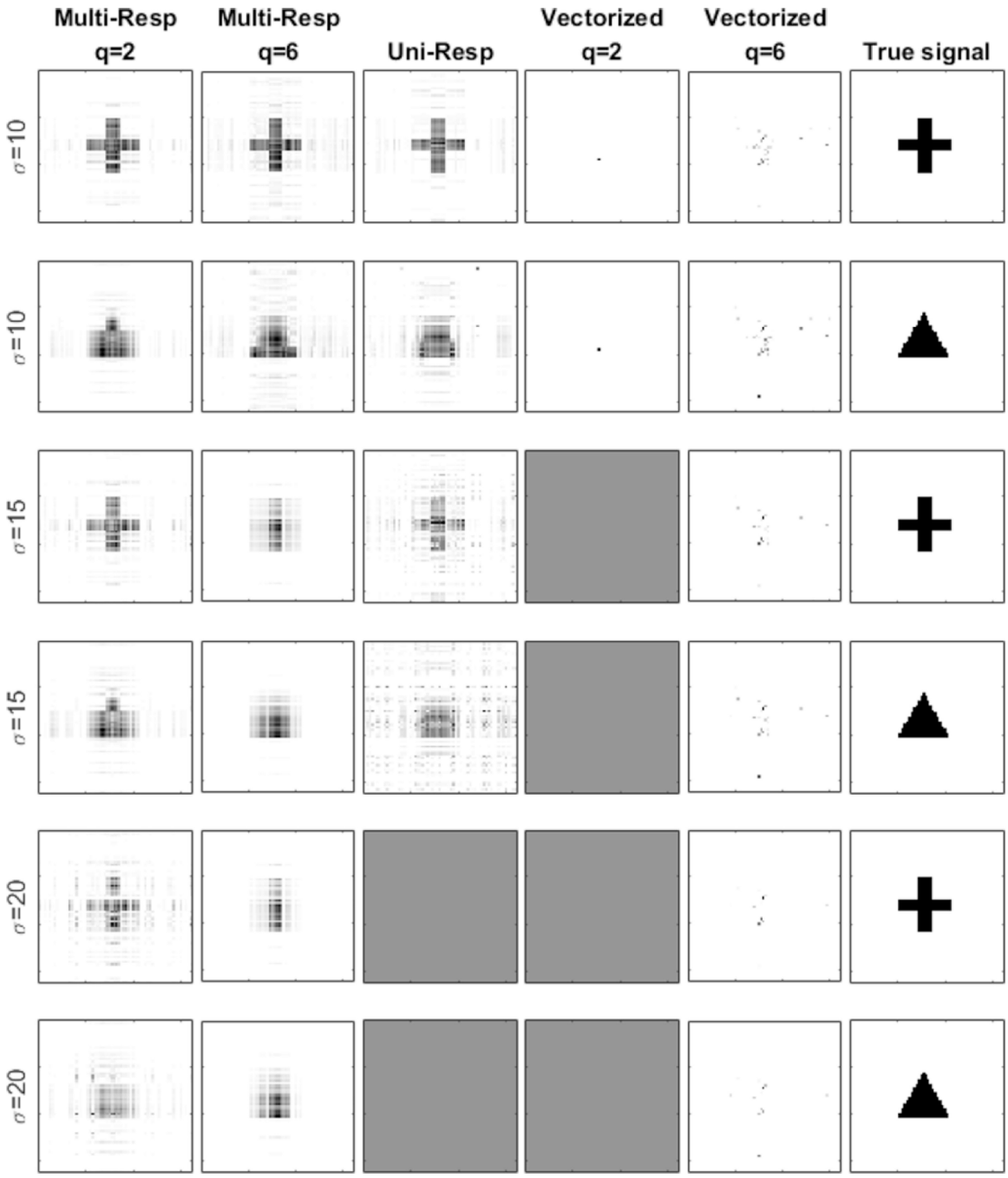

Fig. 4.

Estimated coefficient images under varying noise level (σ = 10, 15, 20) and number of response variables (q = 2, 6). The coefficient patterns are different for different responses. One half adopts “cross”, while the other half “triangle”.

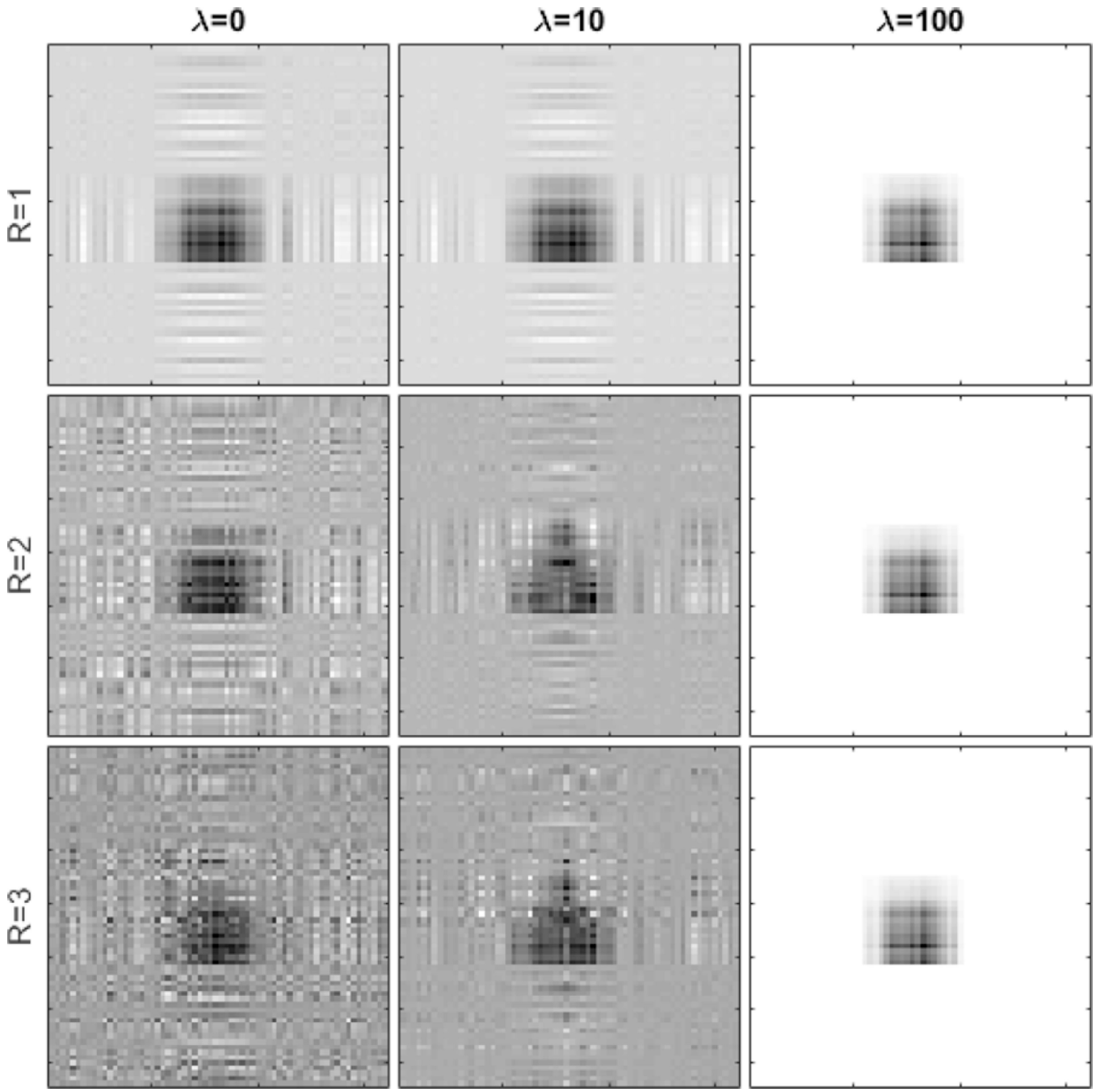

Fig. 5.

Algorithm stability with respect to the varying regularization parameter λ = 0, 10, 100 and varying rank R = 1, 2, 3.

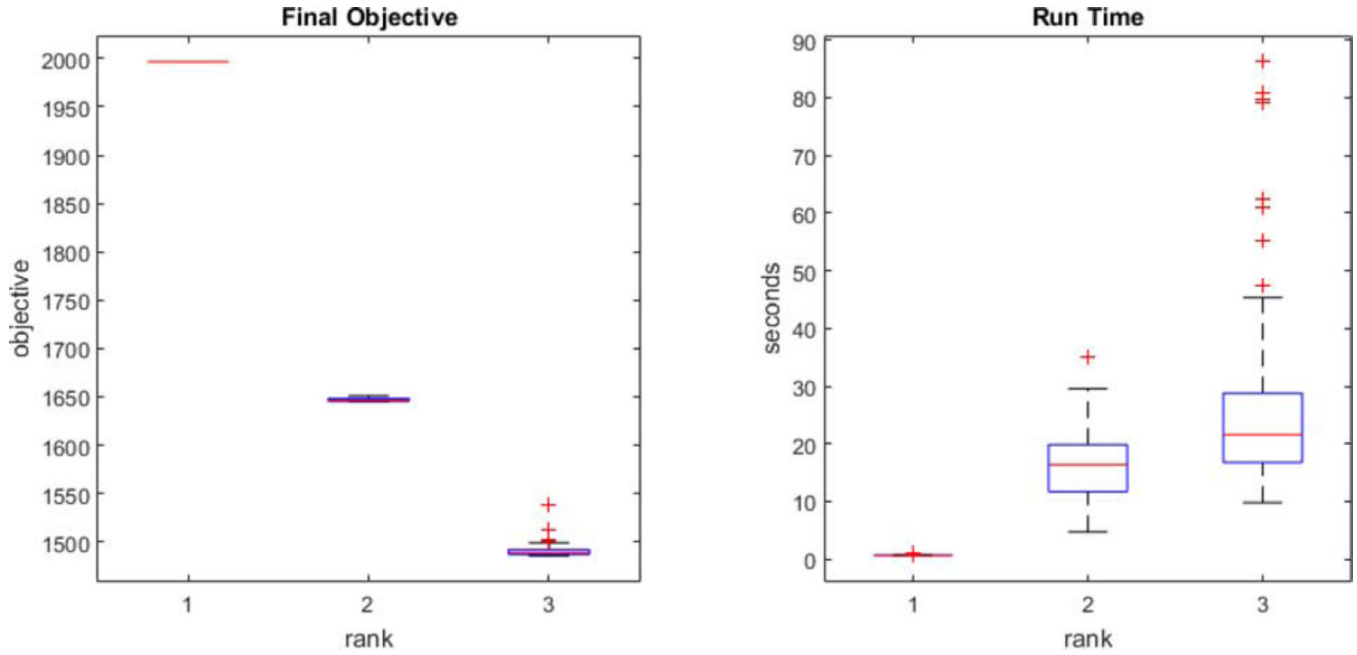

Fig. 6.

Convergence behavior with 100 randomly generated starting values and the corresponding run time.

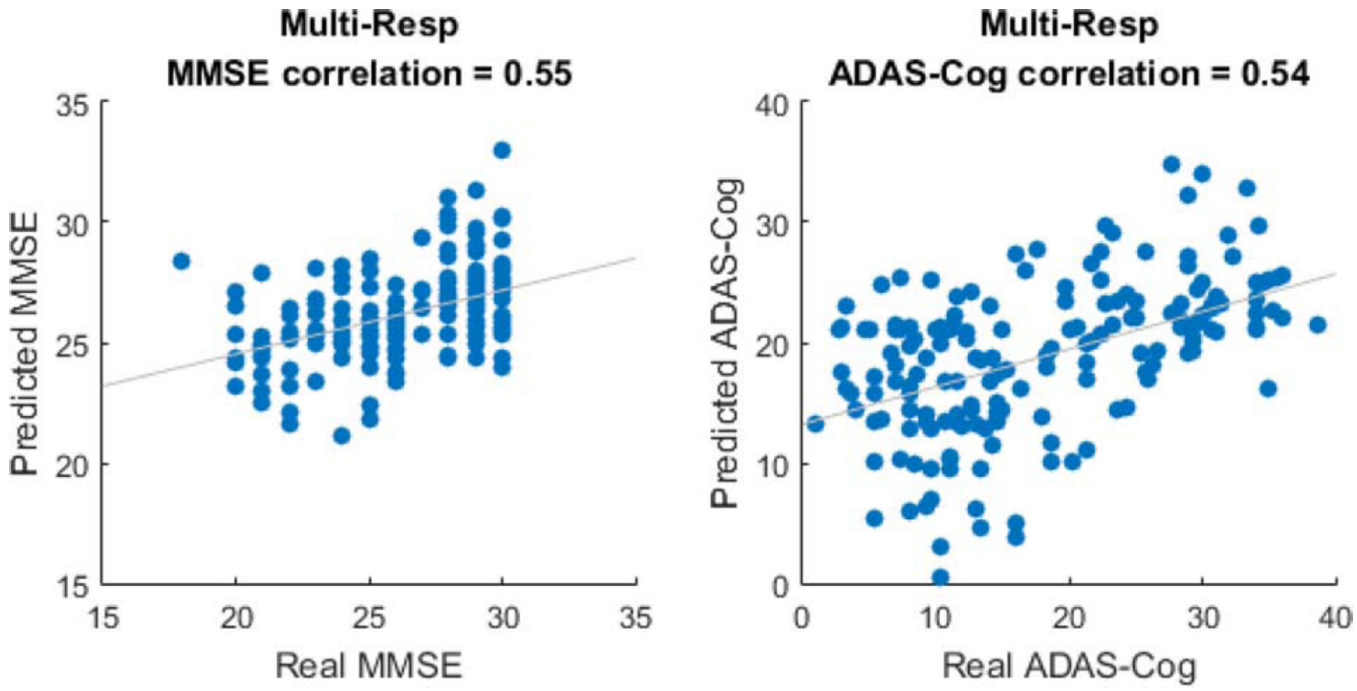

Fig. 7.
Scatter plots of the predicted MMSE and ADAS-Cog versus the observed scores.

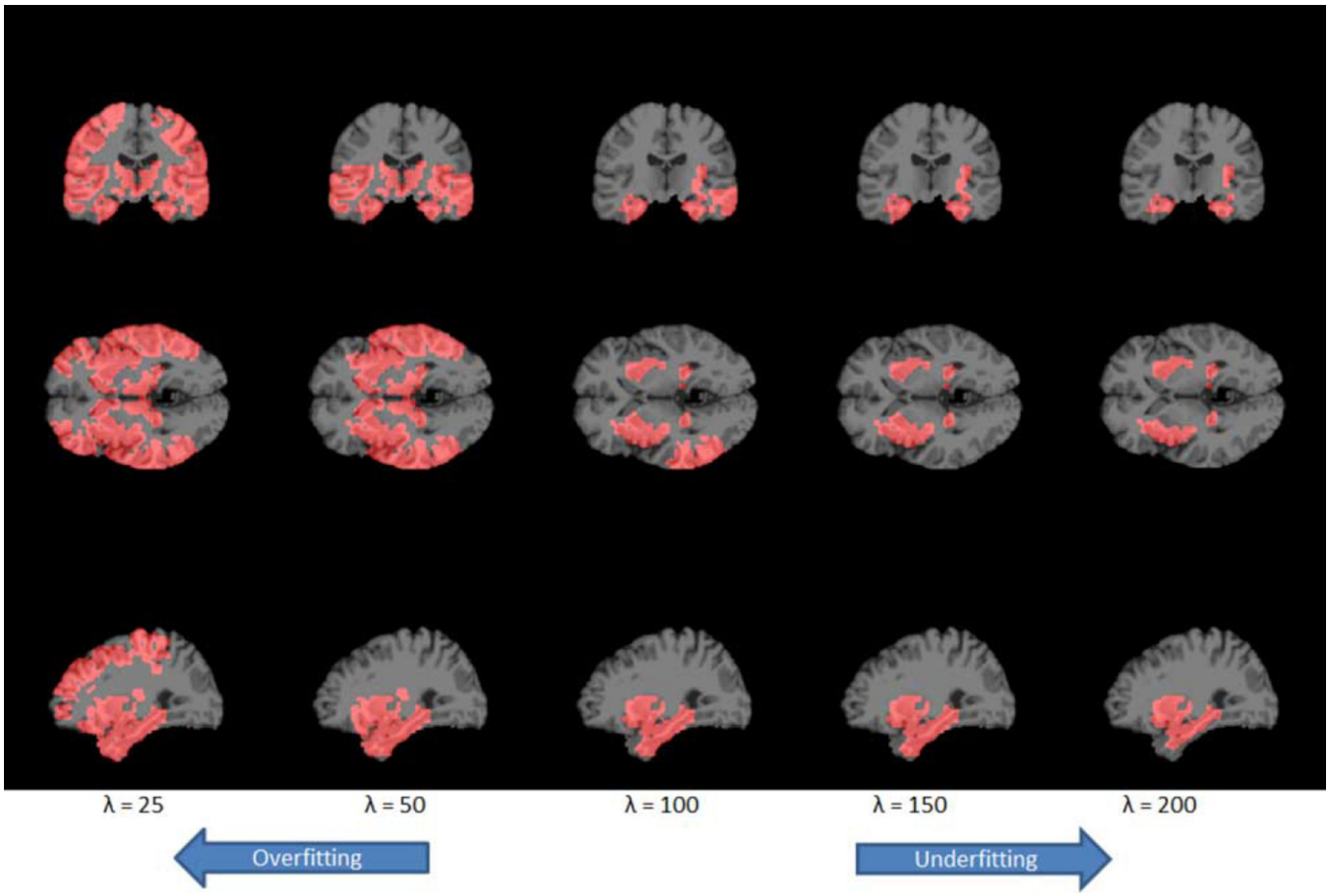

Fig. 8.

Regions (red part) selected by the sparse multivariate tensor regression model that are relevant to the Alzheimer's disease. The optimal tuning parameter based on cross-validation is $\lambda = 100$.

TABLE I

Demographic and Clinical Information of the Subjects.

Group	AD (*n* = 93)	NC (*n* = 101)
Female/Male	36/57	39/62
Age (Mean ± SD)	75.5 ± 7.4	75.9 ± 4.8
MMSE (Mean ± SD)	23.5 ± 2.1	28.9 ± 1.1
ADAS-Cog (Mean ± SD)	27.7 ± 10.7	10.4 ± 4.2

SD: Standard Deviation

TABLE II

RMSE of Prediction Under Varying Noise Level (σ = 10, 15, 20), Number of Response Variables (q = 3, 9), and Response Pairwise Correlation (ρ = 0, 0.9). Reported are the Mean RMSE and Standard Error (in Parenthesis) Based on 50 Data Replications

Signal	Parameter			Root mean squared error		
	q	ρ	σ	Multi-Resp	Uni-Resp	Vectorized
Cross	3	0	10	0.599 (0.003)	0.614 (0.003)	0.993 (0.003)
			15	0.766 (0.003)	0.804 (0.003)	0.996 (0.003)
			20	0.895 (0.005)	0.961 (0.004)	0.998 (0.002)
		0.9	10	0.619 (0.005)	0.613 (0.003)	0.994 (0.003)
			15	0.798 (0.006)	0.808 (0.005)	0.998 (0.003)
			20	0.966 (0.006)	0.976 (0.005)	1.000 (0.003)
	9	0	10	0.615 (0.005)	0.613 (0.002)	0.974 (0.002)
			15	0.767 (0.005)	0.810 (0.002)	0.986 (0.002)
			20	0.875 (0.004)	0.981 (0.003)	0.993 (0.002)
		0.9	10	0.632 (0.008)	0.612 (0.003)	0.986 (0.003)
			15	0.794 (0.006)	0.808 (0.005)	1.000 (0.003)
			20	0.904 (0.005)	0.977 (0.005)	1.001 (0.003)
Triangle	3	0	10	0.668 (0.002)	0.696 (0.002)	0.996 (0.003)
			15	0.847 (0.003)	0.871 (0.005)	1.000 (0.002)
			20	0.956 (0.006)	0.981 (0.004)	1.000 (0.002)
		0.9	10	0.696 (0.004)	0.697 (0.004)	0.997 (0.003)
			15	0.855 (0.005)	0.868 (0.006)	1.001 (0.003)
			20	0.966 (0.007)	0.979 (0.006)	1.003 (0.003)
	9	0	10	0.678 (0.002)	0.696 (0.002)	0.978 (0.003)
			15	0.805 (0.002)	0.873 (0.004)	0.988 (0.002)
			20	0.875 (0.001)	0.982 (0.003)	0.995 (0.002)
		0.9	10	0.702 (0.003)	0.696 (0.003)	0.994 (0.003)

IEEE Trans Med Imaging. Author manuscript; available in PMC 2016 December 13.

Signal	Parameter			Root mean squared error		
	q	ρ	σ	Multi-Resp	Uni-Resp	Vectorized
Butterfly	3	0	15	0.817 (0.004)	0.868 (0.005)	1.009 (0.003)
			20	0.895 (0.004)	0.980 (0.006)	1.003 (0.003)
		0.9	10	0.766 (0.005)	0.801 (0.004)	0.996 (0.003)
			15	0.860 (0.003)	0.921 (0.003)	0.998 (0.002)
			20	0.949 (0.004)	0.998 (0.002)	0.999 (0.002)
	9	0	10	0.797 (0.005)	0.799 (0.004)	0.996 (0.003)
			15	0.925 (0.005)	0.922 (0.004)	0.998 (0.002)
			20	1.014 (0.006)	0.999 (0.003)	0.999 (0.002)
		0.9	10	0.766 (0.004)	0.799 (0.003)	0.990 (0.003)
			15	0.861 (0.002)	0.923 (0.003)	1.000 (0.002)
			20	0.909 (0.002)	1.000 (0.002)	1.000 (0.002)
	0.9	0	10	0.811 (0.004)	0.798 (0.004)	0.999 (0.002)
			15	0.996 (0.005)	0.922 (0.004)	1.000 (0.002)
			20	0.944 (0.004)	1.000 (0.003)	1.000 (0.002)

TABLE III

Root Mean Squared Error of Prediction Under Varying Noise Level (σ = 10, 15, 20), Number of Response Variables (q = 2, 6), and Response Pairwise Correlation (ρ = 0, 0.9). Reported are the Mean RMSE and Standard Error (in Parenthesis) Based on 50 Data Replications

Signals	Parameter			Root mean squared error		
	q	ρ	σ	Multi-Resp	Uni-Resp	Vectorized
Cross & Triangle	2	0	10	0.424 (0.004)	0.427 (0.003)	0.999 (0.005)
			15	0.655 (0.005)	0.711 (0.007)	1.000 (0.005)
			20	0.837 (0.006)	0.963 (0.008)	1.002 (0.005)
		0.9	10	0.421 (0.004)	0.431 (0.004)	1.002 (0.005)
			15	0.673 (0.006)	0.713 (0.007)	1.005 (0.005)
			20	0.886 (0.009)	0.960 (0.011)	1.007 (0.006)
Cross & Triangle	6	0	10	0.411 (0.002)	0.429 (0.002)	0.976 (0.004)
			15	0.682 (0.003)	0.720 (0.004)	0.989 (0.004)
			20	0.805 (0.003)	0.967 (0.005)	0.995 (0.003)
		0.9	10	0.450 (0.004)	0.429 (0.004)	0.982 (0.005)
			15	0.700 (0.006)	0.777 (0.009)	0.999 (0.005)
			20	0.833 (0.007)	0.966 (0.009)	1.009 (0.006)

TABLE IV

Estimation of the two Clinical Scores by Various Methods. Reported are the Average and Standard Error (in Parenthesis) of the Root Mean Square Error and the Pearson Correlation Coefficient Between the Predicted and the Observed Scores Based on 10-Fold Cross-Validation

Method	Correlation coefficient		Root-mean-square error	
	MMSE	ADAS-Cog	MMSE	ADAS-Cog
Multi-Resp	0.55 (0.03)	0.54 (0.04)	2.83 (0.17)	10.13 (0.53)
Uni-Resp	0.27 (0.08)	0.28 (0.10)	3.18 (0.11)	11.12 (0.57)
Vectorized	0.55 (0.05)	0.49 (0.07)	2.83 (0.12)	10.31 (0.57)
M3T	0.56 (0.05)	0.58 (0.06)	2.78 (0.13)	9.94 (0.73)
SVR + lasso	0.58 (0.06)	0.54 (0.06)	2.73 (0.16)	10.14 (0.68)
SVR	0.42 (0.06)	0.49 (0.06)	3.58 (0.20)	11.67(0.52)

TABLE V

AAL Regions (Colored Cells) Selected by the Sparse Multivariate Tensor Regression Model With Varying Tuning Parameters, Along With Their Support in the Literature. 13 Additional Regions Which are Only Selected When $\lambda = 50$ are not Shown Here in the Interest of Space.

AAL Region	$\lambda = 50$	$\lambda = 100$	$\lambda = 150$	$\lambda = 200$	Literature Support
AMYGD_L					[37], [38]
AMYGD_R					
HIPPO_L					[37], [39], [40]
HIPPO_R					
NL_L					[41], [42]
NL_R					
IN_L					[37], [38]
IN_R					
PARA_HIPPO_L					[39], [43]
PARA_HIPPO_R					
COB_L					[44]
T1A_L					[39]
T1A_R					
T2_L					[39]
T2_R					

Abbreviations: AMYGD = Amygdala; HIPPO = Hippocampus; NL = Lenticular Nucleus, Putamen; IN = Insula; PARA_HIPPO = Parahippocampal Gyrus; COB = Olfactory Cortex; T1A = Temporal Pole: Superior Temporal Gyrus; T2 = Middle Temporal Gyrus