Hindawi
Computational and Mathematical Methods in Medicine
Volume 2018, Article ID 7429782, 13 pages
https://doi.org/10.1155/2018/7429782

Research Article

Linearized and Kernelized Sparse Multitask Learning for Predicting Cognitive Outcomes in Alzheimer's Disease

Xiaoli Liu,[1,2] Peng Cao ,[1] Jinzhu Yang,[1,2] and Dazhe Zhao[1,2]

[1] *Computer Science and Engineering, Northeastern University, Shenyang, China*
[2] *Key Laboratory of Medical Image Computing of Ministry of Education, Northeastern University, Shenyang, China*

Correspondence should be addressed to Peng Cao; caopeng@cse.neu.edu.cn

Received 4 August 2017; Revised 18 December 2017; Accepted 26 December 2017; Published 24 January 2018

Academic Editor: Peng Li

Alzheimer's disease (AD) has been not only the substantial financial burden to the health care system but also the emotional burden to patients and their families. Predicting cognitive performance of subjects from their magnetic resonance imaging (MRI) measures and identifying relevant imaging biomarkers are important research topics in the study of Alzheimer's disease. Recently, the multitask learning (MTL) methods with sparsity-inducing norm (e.g., $\ell_{2,1}$-norm) have been widely studied to select the discriminative feature subset from MRI features by incorporating inherent correlations among multiple clinical cognitive measures. However, these previous works formulate the prediction tasks as a linear regression problem. The major limitation is that they assumed a linear relationship between the MRI features and the cognitive outcomes. Some multikernel-based MTL methods have been proposed and shown better generalization ability due to the nonlinear advantage. We quantify the power of existing linear and nonlinear MTL methods by evaluating their performance on cognitive score prediction of Alzheimer's disease. Moreover, we extend the traditional $\ell_{2,1}$-norm to a more general $\ell_q\ell_1$-norm ($q \geq 1$). Experiments on the Alzheimer's Disease Neuroimaging Initiative database showed that the nonlinear $\ell_{2,1}\ell_q$-MKMTL method not only achieved better prediction performance than the state-of-the-art competitive methods but also effectively fused the multimodality data.

1. Introduction

Alzheimer's disease (AD) is a severe neurodegenerative disorder that results in a loss of mental function due to the deterioration of brain tissue, leading directly to death [1]. It accounts for 60–70% of age related dementia, affecting an estimated 30 million individuals in 2011 and the number is projected to be over 114 million by 2050 [2]. The cause of AD is poorly understood and currently there is no cure for AD. AD has a long preclinical phase, lasting a decade or more. There is increasing research emphasis on detecting AD in the preclinical phase, before the onset of the irreversible neuron loss that characterizes the dementia phase of the disease, since therapies/treatment are most likely to be effective in this early phase. The Alzheimer's Disease Neuroimaging Initiative (ADNI, http://adni.loni.usc.edu/) has been facilitating the scientific evaluation of neuroimaging data including magnetic resonance imaging (MRI) and positron emission tomography (PET), along with other biomarkers and clinical and neuropsychological assessments for predicting the onset and progression of MCI (mild cognitive impairment) and AD. Early diagnosis of AD is key to the development, assessment, and monitoring of new treatments for AD.

Recently, rather than predicting categorical variables in the classification, various studies started to estimate continuous clinical variables from brain images. Therefore, instead of classifying a subject into binary or multiple predetermined categories or stages of the disease, regression focus is on estimating continuous values which may help to assess patient's disease progression. The most commonly used cognitive measures are Alzheimer's Disease Assessment Scale (ADAS) cognitive total score, Mini Mental State Exam (MMSE) score, and Rey Auditory Verbal Learning Test (RAVLT). Regression analyses were commonly used to predict cognitive scores from imaging measures. The relationship between commonly used cognitive measures and structural changes with MRI has been previously studied by regression models and the results demonstrated that there exists a relationship between baseline MRI features and cognitive measures [3, 4]. For example, Wan et al. proposed an elegant regression model called CORNLIN that employs a sparse Bayesian learning algorithm to predict multiple cognitive scores based on 98 structural MRI regions of interests (ROIs) for Alzheimer's disease patients. The polynomial model used in CORNLIN can detect either a nonlinear or a linear relationship between brain structure and cognitive decline [3]. Stonnington et al. adopted relevance vector regression, a sparse kernel method formulated in a Bayesian framework, to predict four sets of cognitive scores using MRI voxel based morphometry measures [4]. One of the biggest challenges in the prediction of inferring cognitive outcomes with MRI is the high dimensionality, which affects the computational performance and leads to a wrong estimation and identification of the relevant predictors. To reduce the high dimensionality and identify the relevant biomarkers, the sparse methods have attracted a great amount of research efforts in the neuroimaging field due to its sparsity-inducing property. Ye et al. applied sparse logistic regression with stability selection to ADNI data for robust feature selection [5] and successfully predicted the conversion from MCI into probable AD and identified a small subset of biosignatures.

It is known that there exist inherent correlations among multiple clinical cognitive variables of a subject. However, many works do not model dependence relation between multiple tasks and neglect the correlation between clinical tasks which is potentially useful. When the tasks are believed to be related, learning multiple related tasks jointly can improve the performance relative to learning each task separately. Multitask learning (MTL) is a statistical learning framework which aims at learning several models in a joint manner. It has been commonly used to obtain better generalization performance than learning each task individually [6, 7]. The critical issues in MTL are to identify how the tasks are related and build learning models to capture such task relatedness. The most recent studies [6, 8, 9] employed multitask learning with $\ell_{2,1}$-norm [7] regularization and aimed to select features that could predict all or most clinical scores. The $\ell_{2,1}$-norm is chosen to be the regularization. Thus, the $\ell_{2,1}$-norm regularized regression model is able to select some common features across all the tasks. However, in these learning methods, each task is traditionally performed by formulating a linear regression problem, in which the cognitive score is a linear function of the neuroimaging measures.

Kernel methods have been studied to model the cognitive scores as nonlinear functions of neuroimaging measures. Recently, many kernel-based classification or regression methods with faster optimization speed or stronger generalization performance have been proposed and investigated by theoretically analyzing and experimentally evaluating [10, 11]. Multiple kernel learning (MKL) [12], which learns the optimal kernel for a given task by a weighted, linear combination of predefined candidate kernels, has been introduced to handle the problem of kernel selection. The multiple kernel learning method not only learns an optimal combination of given base kernels but also provides a flexible framework to exploit the nonlinear relationship between MRI measures and cognitive scores.

In building the predictive model for classification or regression in AD, kernel has been widely used; therefore, it is important to extend the existing kernel-based learning methods to the case of multitask learning. In this paper, we propose two nonlinear multikernel-based multiple learning methods in [13] for building regression models, to exploit and investigate the nonlinear relationship between MRI measures and cognitive scores. Moreover, an $\ell_q\ell_1$-norm is used to extend the traditional $\ell_2\ell_1$-norm. The goal of our work is to (1) predict subjects' cognitive scores in a number of neuropsychological assessments using their MRI measures across the entire brain, (2) identify what the performance of the nonlinear method is compared with the linear $\ell_q\ell_1$-norm MTL and other MTL methods with different assumption. No previous studies have systematically and extensively examined the prediction performance by linear MTL and nonlinear MTL methods, and (3) identify what the learning capacity of the multikernel framework on fusing multimodality data is.

The rest of the paper is organized as follows. In Section 2, we provide a description of the multitask learning formulation. A linearized MTL and two multikernel-based MTL methods with $\ell_q\ell_1$-norm are provided in Section 3. In Section 4, we present the experimental results and compare the performance of linearized and kernelized MTL methods from the ADNI-1 dataset. The conclusion is drawn in Section 5.

2. Multitask Learning

Consider a multitask learning (MTL) setting with T tasks. Let p be the number of covariates, shared across all the tasks, and m be the number of samples. Let $X \in \mathbb{R}^{m\times p}$ denote the matrix of covariates, $Y \in \mathbb{R}^{m\times T}$ be the matrix of responses with each row corresponding to a sample, and $\Theta \in \mathbb{R}^{p\times T}$ denote the parameter matrix, with column $\theta_{.t} \in \mathbb{R}^{p}$ corresponding to task t, $t = 1,\dots,T$, and row $\theta_{h.} \in \mathbb{R}^{T}$ corresponding to feature h, $h = 1,\dots,p$.

The MTL formulation focuses on the following regularized loss function:

$$\min_{\Theta\in\mathbb{R}^{p\times T}} \quad F(Y, X, \Theta) + \lambda R(\Theta), \tag{1}$$

where $F(\cdot)$ denotes the loss function and $R(\cdot)$ is the regularizer. In the current context, we assume the loss to be square loss; that is,

$$F(Y, X, \Theta) = \|Y - X\Theta\|_F^2 = \sum_{i=1}^{m} \left\|\mathbf{y}_i - \mathbf{x}_i\Theta\right\|_2^2, \tag{2}$$

where $\mathbf{y}_i \in \mathbb{R}^{1\times T}$ and $\mathbf{x}_i \in \mathbb{R}^{1\times p}$ are the ith rows of Y and X, respectively, corresponding to the multitask response and covariates for the ith sample. We note that the MTL framework can be easily extended to other loss functions. Base on some prior knowledge, we then add penalty $R(\Theta)$ to encode the relatedness among tasks.

3. $\ell_q\ell_1$-Norm Regularized Linearized Multitask Learning, $\ell_q\ell_1$-MTL

The $\ell_2\ell_1$-norm was popularly used in multitask feature learning [14]. All the existing algorithms for multitask feature learning assume a linear relationship between MRI features and cognitive scores and aim to learn a common subset of features for all tasks. Since the $\ell_2\ell_1$-norm regularizer imposes the sparsity between all features and nonsparsity between tasks, the features that are discriminative for all tasks will get large weights. However, the $\ell_2\ell_1$-norm is a fixed nonadaptive penalty. To obtain an adaptive regularization and better suit different data structures, we extend the $\ell_{2,1}$-norm to a larger class of mixed norm $\ell_q\ell_1$ that can be adapted to the data. The objective function of linear $\ell_q\ell_1$-MTL is formulated:

$$\min_{\Theta} \quad \frac{1}{2}\|Y - X\Theta\|_F^2 + \lambda\|\Theta\|_{q,1}. \tag{3}$$

When $q = 1$, problem (3) reduces to the ℓ_1-regularized problem; when $q = 2$, problem (3) reduces to the $\ell_{2,1}$-regularized problem.

An efficient algorithm is based on the accelerated gradient method for solving the $\ell_q\ell_1$-regularized problem, which is applicable for all values of q larger than 1.

First, construct the following model for approximating the composite function $\mathscr{M}(\cdot)$ at the point $\Theta^{(l)}$:

$$\begin{aligned} \mathscr{M}_{L,\Theta^{(l)}}(\Theta) := {} & F\left(\Theta^{(l)}\right) + \left\langle \Theta - \Theta^{(l)}, \nabla F\left(\Theta^{(l)}\right)\right\rangle \\ & + \frac{L}{2}\left\|\Theta - \Theta^{(l)}\right\|_F^2 + R(\Theta), \end{aligned} \tag{4}$$

where $L > 0$. In the model $\mathscr{M}_{L,\Theta^{(l)}}(\Theta)$, apply the first-order Taylor expansion at the point Θ (including all terms in the square bracket) for the smooth loss function $F(\cdot)$, and directly put the nonsmooth penalty $R(\cdot)$ into the model. The regularization term $(L/2)\|\Theta - \Theta^{(l)}\|_F^2$ prevents Θ from walking far away from $\Theta^{(l)}$, and thus the model can be a good approximation to $\Phi(\Theta)$ in the neighborhood of $\Theta^{(l)}$, where $\Phi(\Theta) \equiv F(\Theta) + R(\Theta)$.

The accelerated gradient method is based on two sequences $\{\Theta^{(l)}\}$ and $\{\Gamma^{(l)}\}$ in which $\{\Theta^{(l)}\}$ is the sequence of approximate solutions and $\{\Gamma^{(l)}\}$ is the sequence of search points. The search point $\Gamma^{(l)}$ is the affine combination of $\Theta^{(l-1)}$ and $\Theta^{(l)}$ as

$$\Gamma^{(l)} = \Theta^{(l)} + \beta^{(l)}\left(\Theta^{(l)} - \Theta^{(l-1)}\right), \tag{5}$$

where $\beta^{(l)}$ is a properly chosen coefficient. The approximate solution $\Theta^{(l+1)}$ is computed as the minimizer of $\mathscr{M}_{L^{(l)},\Gamma^{(l)}}(\Theta)$:

$$\Theta^{(l+1)} = \arg\min_{\Theta} \quad \mathscr{M}_{L^{(l)},\Gamma^{(l)}}(\Theta), \tag{6}$$

where $L^{(l)}$ is determined by line search, for example, the Armijo-Goldstein rule, so that $L^{(l)}$ should be appropriate for $\Gamma^{(l)}$.

The key subroutine is (6), which can be computed as $\Theta^{(l+1)} = \pi_{1q}(\Gamma^{(l)} - \nabla F(\Gamma^{(l)})/L^{(l)}, \lambda/L^{(l)})$, where $\pi_{1q}(\cdot)$ is the $\ell_q\ell_1$-regularized Euclidean projection (EP_{1q}) problem:

$$\pi_{1q}(V, \lambda) = \arg\min_{\Theta\in\mathbb{R}^{p\times T}} \quad \frac{1}{2}\|\Theta - V\|_F^2 + \lambda\sum_{h=1}^{p}\|\theta_{h.}\|_q. \tag{7}$$

Note that the h features in (7) are independent. In [15], the method can be used for ease of different independent groups; that is, $\pi_{1q}(V,\lambda) = \arg\min_{W\in\mathbb{R}^n}(1/2)\|W - V\|_2^2 + \lambda\sum_{i=1}^{\mathscr{G}}\|w_i\|_q$, where $\mathscr{G}$ is the independent groups. In our paper, we focus on how the method deals with multitask learning problem in (7), where $\mathscr{G}$ is equal to p, and each group denotes the corresponding feature shared across the multiple tasks. Thus, the optimization in (7) decouples into a set of p independent ℓ_q-regularized Euclidean projection problems:

$$\pi_q(v_{h.}) = \arg\min_{\theta_{h.}\in\mathbb{R}^T} \quad \frac{1}{2}\|\theta_{h.} - v_{h.}\|_2^2 + \lambda\|\theta_{h.}\|_q. \tag{8}$$

Then, the optimal solution $\theta_{h.}^*$ of (8) can be gotten as follows:

$$\begin{aligned}
\text{if} \quad & \|v_{h.}\|_{\overline{q}} \le \lambda, \\
& \theta_{h.}^* = \mathbf{0}; \\
\text{else if} \quad & \|v_{h.}\|_{\overline{q}} \ge \lambda, \quad q = 1, \\
& \theta_{h.}^* = \operatorname{sgn}(v_{h.}) \odot \max(|v_{h.}| - \lambda, 0); \\
\text{else if} \quad & \|v_{h.}\|_{\overline{q}} \ge \lambda, \quad q = 2, \\
& \theta_{h.}^* = \frac{\|v_{h.}\|_2 - \lambda}{\|v_{h.}\|_2} v_{h.}; \\
\text{else if} \quad & \|v_{h.}\|_{\overline{q}} \ge \lambda, \quad q = \infty, \\
& \theta_{h.}^* = \operatorname{sgn}(v_{h.}) \odot \min(|v_{h.}|, u^*); \\
\text{else} \quad & \|v_{h.}\|_{\overline{q}} \ge \lambda, \quad 1 < q < \infty, \ q \ne 2, \\
& \theta_{h.}^* \text{ is the unique root of } \varphi_{c^*}^{v_{h.}},
\end{aligned} \tag{9}$$

where $\overline{q} = q/(q-1)$, and thus q and $\overline{q}$ satisfy the following relationship: $1/\overline{q} + 1/q = 1$, u^* is the unique root of $\zeta(u) = \sum_{h=1}^{p}\max(|v_{h.}| - u, 0) - \lambda$, and $\zeta(\cdot)$ is an auxiliary function, defined as $\zeta_c^v(\theta) = \theta + c\theta^{q-1} - v$ with $0 \le \theta \le v$; And $\varphi_c^v(\theta) = \theta + c\theta^{(q-1)} - v$, $0 < x < v$ and $c^* = \lambda\|\theta_{h.}^*\|_q^{1-q}$. Note that $\mathbf{z} = \mathbf{x} \odot \mathbf{y}$ denotes $z_i = x_i y_i$.

The algorithm $\ell_q\ell_1$-MTL is summarized in Algorithm 1.

4. Kernelized Multitask Learning

4.1. Multikernel Learning. The limitation in this traditional $\ell_{2,1}$-norm MTL model is that subjects cognitive score under a task is modeled as a linear function of his/her MRI measures. The kernel methods, for example, SVM or SVR, can model the nonlinear distribution of the data by mapping the input

```
Input: $\lambda > 0, L^{(0)} > 0, X, Y$
Output: $\Theta$.
Initialize $\Theta^{(1)} = \Theta^{(0)}, \alpha^{(-1)} = 0, \alpha^{(0)} = 1$ and $L = L^{(0)}$.
$l = 1$
repeat
    Set $\beta^{(l)} = (\alpha^{(l-2)} - 1)/\alpha^{(l-1)}, \Gamma^{(l)} = \Theta^{(l)} + \beta^{(l)}(\Theta^{(l)} - \Theta^{(l-1)})$
    Find the smalles $L = L^{(l-1)}, 2L^{(l-1)}, \ldots$ such that
            $\Phi(\Theta^{(l+1)}) \leq \mathscr{M}_{L,\Gamma^{(l)}}(\Theta^{(l+1)})$,
        where $\Theta^{(l+1)} = \arg\min_{\Theta} \mathscr{M}_{L,\Gamma^{(l)}}(\Theta)$
    $L^{(l)} = L$ and $\alpha^{(l+1)} = (1 + \sqrt{1 + 4\alpha^{(l)2}})/2$
    $l = l + 1$
until convergence criterion is satisfied
```

Algorithm 1: $\ell_q\ell_1$-MTL.

data into a nonlinear feature space by kernel embedding. In this section, we consider the case that $\ell_{2,1}$-norm regularized MTL is extended to kernel method. Let us define the kernel function $\phi_j(\mathbf{x}) : \mathbb{R}^p \rightarrow \mathbb{R}^{\hat{p}}$, which maps the data samples from an input space to a feature space (a high-dimensional Hilbert space $\mathscr{H}$), where $\hat{p}$ denotes the dimensionality of the feature space and $\mathbf{x}$ is a sample from the input space. A kernel function k is capable of attaining the inner product of two mapped datasets in $\mathscr{H}$: $k(\mathbf{x}, \mathbf{x}') = \phi(\mathbf{x}) \cdot \phi(\mathbf{x}')$ in the original space without explicitly computing the mapped data. The associated Gram matrix has entries $K(i, j) = k(\mathbf{x}, \mathbf{x}')$.

The most suitable types and parameters of the kernels for a particular task are often unknown, and the selection of the optimal kernel by exhaustive search on a predefined pool of kernels is usually time-consuming and sometimes causes overfitting. Multiple kernel learning (MKL) attempts to achieve better results by combining several base kernels instead of using only one specific kernel. MKL assumes that $\mathbf{x}_i$ can be mapped to k different Hilbert spaces, $\mathbf{x}_i \rightarrow \phi_j(\mathbf{x}_i)$, $j = 1, \ldots, k$, implicitly with k nonlinear mapping functions, and the objective of MKL is to seek the optimal kernel combination $\hat{k}(\mathbf{x}, \mathbf{x}') = \sum_{j=1}^{k} d_j k_j(x, x')$, $d_j \geq 0$, $\sum_{j=1}^{k} d_j = 1$, where $\mathbf{d}$ is the kernel weight vector. The primal objective function of multiple kernel regression model is written as follows:

$$\begin{aligned} \min_{\tilde{\theta},\xi} \quad & \frac{1}{2}\sum_{j=1}^{k}\frac{\|\tilde{\theta}_j\|_2^2}{d_j} + \frac{\lambda}{2}\sum_{i=1}^{m}\xi_i^2, \\ \text{s.t.} \quad & \sum_{j=1}^{k}\tilde{\theta}_j^T\phi_j(x_i) - y_i = \xi_i, \\ & \sum_{j=1}^{k} d_j = 1, \quad d_j \geq 0. \end{aligned} \tag{10}$$

MKL learns both the weights of the kernel combination $\mathbf{d}$ and the parameters of the regression $\tilde{\theta}$ by solving a single joint optimization problem.

Using $\boldsymbol{\alpha}$ to denote the Lagrange multipliers, the objective value of the dual problem of (10) can be written as follows:

$$\begin{aligned} J(\mathbf{d}) = \max_{\boldsymbol{\alpha}} \quad & -\boldsymbol{\alpha}^T\mathbf{y}_t - \frac{1}{2}\boldsymbol{\alpha}^T\widehat{\mathbf{K}}\boldsymbol{\alpha} - \frac{1}{2C}\boldsymbol{\alpha}^{*T}\boldsymbol{\alpha}, \\ \text{s.t.} \quad & \sum_{j=1}^{k} d_j = 1, \quad d_j \geq 0, \end{aligned} \tag{11}$$

where $\widehat{\mathbf{K}} = \sum_{j=1}^{k} d_j\mathbf{K}_j$ is the combined Gram matrix and K_j, $j = 1, \ldots, k$, is the given set of base kernels.

4.2. $\ell_q\ell_1$-Norm Regularized Multikernel Multitask Learning, $\ell_q\ell_1$-MKMTL.

We follow the multiple kernel learning scheme and use the $\ell_{q,1}$-norm to model the relationship between the tasks to learn a common kernel representation by imposing sparsity constraint on the kernel weight. The method, called $\ell_q\ell_1$-MKMTL, assumes that few base kernels are important for the tasks and encourages a linear combination of only few kernels and assumes few selected kernels are similar across the tasks. The formulation of $\ell_q\ell_1$-MKMTL can be expressed as follows:

$$\begin{aligned} \min_{\tilde{\theta},\xi} \quad & \frac{1}{2}\left(\sum_{j=1}^{k}\left(\sum_{t=1}^{T}\|\tilde{\theta}_{tj}\|_2^q\right)^{1/q}\right)^2 + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{i=1}^{m_t}\xi_{ti}^2, \\ \text{s.t.} \quad & \sum_{j=1}^{k}\tilde{\theta}_{tj}^T\phi_j(x_{ti}) - y_{ti} = \xi_{ti}. \end{aligned} \tag{12}$$

We now rewrite this formulation in a convenient form which can be efficiently solved using mirror-descent based algorithms. We introduce some more notations: let $\Delta_{d,r} = \{\mathbf{z} \equiv [z_1, \ldots, z_d]^T \mid \sum_{i=1}^{d} z_i^r \leq 1,\ z_i \geq 0,\ i = 1, \ldots, d\}$ and with slight abuse of notation let $\Delta_{d,1} = \Delta_d$. Next, we note the following [16].

Lemma 1. *Let $a_i \geq 0$, $i = 1, \ldots, d$ and $1 < r < \infty$. Then, for $\Delta_{d,r}$ defined as before,*

$$\min_{\eta \in \Delta_{d,r}} \sum_i \frac{a_i}{\eta_i} = \left(\sum_{i=1}^{d} a_i^{r/(r+1)} \right)^{(r+1)/r}, \tag{13}$$

and the minimum is attained at

$$\eta_i = \frac{a_i^{1/(r+1)}}{\left(\sum_{i=1}^{d} a_i^{r/(r+1)} \right)^{1/r}}, \tag{14}$$

with the convention that $a/0$ is 0 if $a = 0$ and is ∞ if $a \neq 0$.

Using the result of the lemma (with $r = 1$) and introducing variables $\mu = [\mu_1, \ldots, \mu_k]^T$, we have

$$\begin{aligned} &\left(\sum_{j=1}^{k} \left(\sum_{t=1}^{T} \left(\left\| \widetilde{\theta}_{tj} \right\|_2 \right)^q \right)^{1/q} \right)^2 \\ &\quad = \min_{\mu \in \Delta_k} \sum_{j=1}^{k} \frac{\left(\sum_{t=1}^{T} \left(\left\| \widetilde{\theta}_{tj} \right\|_2 \right)^q \right)^{2/q}}{\mu_j}. \end{aligned} \tag{15}$$

Now introducing dual variables $\nu_j = [\nu_{j1}, \ldots, \nu_{jT}]^T$, $j = 1, \ldots, k$, and using the notion of dual norm [17], we obtain

$$\left(\sum_{t=1}^{T} \left(\left\| \widetilde{\theta}_{tj} \right\|_2^2 \right)^{q/2} \right)^{2/q} = \max_{\nu_j \in \Delta_{T,\overline{q}}} \sum_{t=1}^{T} \nu_{jt} \left\| \widetilde{\theta}_{tj} \right\|_2^2, \tag{16}$$

where $\overline{q} = q/(q-2)$. With this, the objective in the $\ell_q\ell_1$-MKMTL formulation can now be written as

$$\min_{\mu \in \Delta_k} \min_{\widetilde{\theta}, \xi} \max_{\nu_j \in \Delta_{T,\overline{q}}} \frac{1}{2} \sum_{j=1}^{k} \frac{\sum_{t=1}^{T} \nu_{jt} \left\| \widetilde{\theta}_{tj} \right\|_2^2}{\mu_j} + \frac{\lambda}{2} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \xi_{ti}^2. \tag{17}$$

Using α to denote the Lagrange multipliers, this has the Lagrangian

$$\begin{aligned} \mathscr{L} = {} & \frac{1}{2} \sum_{j=1}^{k} \frac{\sum_{t=1}^{T} \nu_{jt} \left\| \widetilde{\theta}_{tj} \right\|_2^2}{\mu_j} + \frac{\lambda}{2} \sum_{t=1}^{T} \sum_{i=1}^{m_t} \xi_{ti}^2 \\ & + \sum_{t=1}^{T} \sum_{i=1}^{m_t} \alpha_{ti} \left(\sum_{j=1}^{k} \widetilde{\theta}_{tj}^T \phi_j(x_{ti}) - y_{ti} - \xi_{ti} \right). \end{aligned} \tag{18}$$

Recall our foray into Lagrange duality. We can solve the original problem by doing

$$\max_{\alpha} \min_{\widetilde{\theta}, \xi} \mathscr{L}\left(\widetilde{\theta}, \xi, \alpha \right). \tag{19}$$

To begin, we attack the inner minimization: For fixed α, we would like to solve for the minimizing $\widetilde{\theta}$ and ξ. We can do this by setting the derivatives of $\mathscr{L}$ with respect to ξ_{ti} and $\widetilde{\theta}$ to be zero. Doing this, we can find

$$\widetilde{\theta}_{tj}^* = -\alpha_t^T \left[\sum_{j=1}^{k} \frac{\mu_j \Phi_{tj}}{\nu_{jt}} \right], \tag{20a}$$

$$\xi_{ti}^* = \frac{\alpha_{ti}}{\lambda}, \tag{20b}$$

where α_t is a vector corresponding to the tth task in the $\ell_q\ell_1$-MKMTL formulation and Φ_{tj} is the data matrix with columns as $\phi_j(x_{ti})$, $i = 1, \ldots, m_t$. So, we can solve the problem by maximizing the Lagrangian (with respect to α), where we substitute the above expressions for ξ and $\widetilde{\theta}$. Thus, we have an unconstrained maximization.

$$\max_{\alpha} \sum_{t=1}^{T} \left\{ -\alpha_t^T y_t - \frac{1}{2} \alpha_t^T \left[\sum_{j=1}^{k} \frac{\mu_j K_{tj}}{\nu_{jt}} \right] \alpha_t - \frac{1}{2\lambda} \alpha_t^T \alpha_t \right\}. \tag{21}$$

Here, $\mathbf{y}_t$ is vector of scores of the tth task training data points and $\mathbf{K}_{ij}$ represents the Gram matrix of the tth task training data points with respect to the jth kernel. Equation (21) is just a quadratic in α. As such, we can find the optimum as the solution of a linear system.

Then, (17) can be written as follows:

$$\min_{\mu \in \Delta_k} \max_{\nu_j \in \Delta_{T,\overline{q}}} \max_{\alpha} \sum_{t=1}^{T} \left\{ -\alpha_t^T y_t - \frac{1}{2} \alpha_t^T \left[\sum_{j=1}^{k} \frac{\mu_j K_{tj}}{\nu_{jt}} \right] \alpha_t - \frac{1}{2\lambda} \alpha_t^T \alpha_t \right\}. \tag{22}$$

The formulation can be transformed as follows:

$$\min_{\mu \in \Delta_k} \max_{\nu_j \in \Delta_{T,\overline{q}}} \max_{\alpha} \sum_{t=1}^{T} \left\{ -\alpha_t^T y_t - \frac{1}{2} \alpha_t^T \left[\sum_{j=1}^{k} \frac{\mu_j K_{tj}}{\nu_{jt}} \right] \alpha_t - \frac{1}{2\lambda} \alpha_t^T \alpha_t \right\}. \tag{23}$$

The algorithm $\ell_q\ell_1$-MKMTL is summarized in Algorithm 2.

4.3. $\ell_{2,1}$-ℓ_q-Norm Regularized Multikernel Multitask Learning, $\ell_{2,1}\ell_q$-MKMTL.

The linearized $\ell_q\ell_1$-MTL assumed linear

```
Input: $\lambda > 0, X, Y$
Output: $\alpha, \nu, \mu$
$n = 0$
repeat
    initiate $\mu$ and $\nu$
    for $t = 1$ to $T$ do
        With fixed $\mu$ and $\nu$, compute $\alpha_t^*$ by using an SVR solver
    end for
    optimize $\mu$ with mirror-descent algorithm
    optimize $\nu$: $-\sum_{j=1}^{k} \min_{\nu_j} \in \Delta_{T,\bar{q}} \sum_{t=1}^{T} (D_{jt}/\nu_{jt})$ where $D_{jt} = (1/2)\mu_j \alpha_t^{\mathrm{T}} K_{tj} \alpha_t$.
    $n = n + 1$
until convergence criterion is satisfied
```

Algorithm 2: $\ell_q\ell_1$-MKMTL.

relationship between the MRI features and the cognitive outcomes. Such a model is the lack of capability to capture nonlinear predictive information from the features. Although the $\ell_q\ell_1$-MKMTL builds the nonlinear relationship for the features and task by mapping to high-dimensional space, it considers that tasks to be learned share a common subset of kernel representations without capturing the interrelationships between different cognitive measures over the feature space.

To overcome the weaknesses of the previous two methods, we project the original feature vectors to a high-dimensional space using multiple nonlinear mapping functions for performing regression task in a nonlinear manner and utilize multitask learning in the multiple kernel spaces for modeling the disease's cognitive scores with a joint $\ell_{2,1}$-ℓ_q sparsity-inducing regularizers. Moreover, we construct new features as orthogonal transforms of the given features, that is, $\mathbf{L}_j\phi_j(x)$, where $\mathbf{L}_j$ is an orthogonal matrix which is to be learned. Again, low empirical risk over each task would imply minimizing the following quadratic loss: $\sum_{t=1}^{T}\sum_{i=1}^{m_t}\min(\sum_{j=1}^{k}\tilde{\theta}_{tj}^{T}\mathbf{L}_j^{T}\phi_j(x_{ti}) - y_{ti})^2$. Before describing the regularization term, we introduce some more notations: Let the entries of $\tilde{\theta}_{tj}$ be $\tilde{\theta}_{tjl}$, $l = 1, \ldots, p_j$, where p_j is the dimensionality of the feature space induced by the jth kernel. By $\tilde{\theta}_{.jl}$ we denote the vector with entries $\tilde{\theta}_{tjl}$, $t = 1, \ldots, T$. The regularization term we employ is $(\sum_{j=1}^{k}(\sum_{l=1}^{p_j}\|\tilde{\theta}_{.jl}\|_2)^q)^{2/q}$, where $q \in [1, 2]$. Different from $\ell_q\ell_1$-MKMTL, the ℓ_q-norm in $\ell_{2,1}\ell_q$-MKMTL is employed over the kernels rather than the tasks.

Mathematically, the $\ell_{2,1}\ell_q$-MKMTL formulation can be expressed as follows:

$$\begin{aligned}
&\min_{\tilde{\theta},\xi,\mathbf{L}} \quad \frac{1}{2}\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{p_j}\left\|\tilde{\theta}_{.jl}\right\|_2\right)^{q}\right)^{2/q} + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{i=1}^{m_t}\xi_{ti}^2, \\
&\text{s.t.} \quad \sum_{j=1}^{k}\tilde{\theta}_{tj}^{T}\mathbf{L}_j^{T}\phi_j(x_{ti}) - y_{ti} = \xi_{ti}, \quad \mathbf{L}_j \in O^{p_j},
\end{aligned} \tag{24}$$

where O^{p_j} represents the set of all orthogonal matrices of dimensionality p_j. In the following text, we rewrite this formulation in a form which is convenient to solve using an MD based algorithm.

Using the result of Lemma 1 and introducing new variables $\nu = [\nu_1, \ldots, \nu_k]^T$, we have

$$\left(\sum_{j=1}^{k}\left(\sum_{l=1}^{p_j}\left\|\tilde{\theta}_{.jl}\right\|_2\right)^{q}\right)^{2/q} = \min_{\nu\in\Delta_{k,\bar{q}}}\sum_{j=1}^{k}\frac{\left(\sum_{l=1}^{p_j}\left\|\tilde{\theta}_{.jl}\right\|_2\right)^2}{\nu_j}, \tag{25}$$

where $\bar{q} = q/(2-q)$. Again using the lemma and introducing new variables $\mu_j = [\mu_{j1}, \ldots, \mu_{jp_j}]^T$, $j = 1, \ldots, k$, the regularizer can be written as

$$\min_{\nu\in\Delta_{k,\bar{q}}}\min_{\mu_j\in\Delta_{p_j}}\sum_{t=1}^{T}\sum_{j=1}^{k}\sum_{l=1}^{p_j}\frac{\tilde{\theta}_{tjl}^2}{\mu_{jk}\nu_j}. \tag{26}$$

Now, we perform a change of variables: $\tilde{\theta}_{tjl}/\sqrt{\mu_{jk}\nu_j} = \bar{\theta}_{tjl}$, $l = 1, \ldots, p_j$. Using this, one can rewrite the $\ell_{2,1}\ell_q$-MKMTL formulation as

$$\begin{aligned}
&\min_{\nu,\mu_j,\mathbf{L}_j} \quad \sum_{t=1}^{T}\min_{\bar{\theta}_t,\xi_t}\frac{1}{2}\sum_{j=1}^{k}\bar{\theta}_{tj}^{T}\bar{\theta}_{tj} + \frac{\lambda}{2}\sum_{t=1}^{T}\sum_{i=1}^{m_t}\xi_{ti}^2, \\
&\text{s.t.} \quad \sum_{j=1}^{k}\bar{\theta}_{tj}^{T}\Lambda_j^{1/2}\mathbf{L}_j^{T}\phi_j(x_{ti}) - y_{ti} = \xi_{ti}, \\
&\qquad \nu \in \Delta_{k,\bar{q}}, \ \mu_j \in \Delta_{p_j}, \ \mathbf{L}_j \in O^{p_j},
\end{aligned} \tag{27}$$

where Λ_j is a diagonal matrix with entries as $\nu_j\mu_{jl}$, $l = 1, \ldots, p_j$.

Now, using α to denote the Lagrange multipliers, this has the Lagrangian of

$$\begin{aligned}
\mathscr{L} = \sum_{t=1}^{T}\Bigg(&\frac{1}{2}\sum_{j=1}^{k}\bar{\theta}_{tj}^{T}\bar{\theta}_{tj} + \frac{\lambda}{2}\sum_{i=1}^{m_t}\xi_{ti}^2 \\
&+ \sum_{i=1}^{m_t}\alpha_{ti}\left(\sum_{j=1}^{k}\bar{\theta}_{tj}^{T}\Lambda_j^{1/2}\mathbf{L}_j^{T}\phi_j(x_{ti}) - y_{ti} - \xi_{ti}\right)\Bigg).
\end{aligned} \tag{28}$$

This can be solved like $\ell_q\ell_1$-MKMTL:

$$\tilde{\theta}_{tj}^{*} = -\alpha_t^{T}\Lambda_j^{1/2}\mathbf{L}_j^{T}\Phi_{tj}, \tag{29a}$$

$$\xi_{ti}^{*} = \frac{\alpha_{ti}}{\lambda}. \tag{29b}$$

```
Input: $X, Y, \lambda > 0$
Output: $\boldsymbol{\alpha}^*$, $\overline{\mathbf{Q}}$
repeat
    optimize $\overline{\mathbf{Q}}$ with mirror-descent algorithm
    for $t = 1$ to $T$ do
        with fixed $\overline{\mathbf{Q}}$, compute $\boldsymbol{\alpha}_t^*$ by using an SVR solver
    end for
    $n = n + 1$
until convergence criterion is satisfied
```

Algorithm 3: $\ell_{2,1}$-ℓ_q-MKMTL.

Again, we substitute the above expressions for ξ and $\widetilde{\theta}$. Thus, we have the following form:

$$\begin{aligned}\min_{\nu,\mu_j,\mathbf{L}_j} \quad & \sum_{t=1}^{T} \max_{\alpha_t} -\alpha_t^T y_t - \frac{1}{2}\alpha_t^T \left(\sum_{j=1}^{k} \Phi_{tj}^T \mathbf{L}_j^T \Lambda_j \mathbf{L}_j \Phi_{tj}\right)\alpha_t \\ & - \frac{1}{2\lambda}\alpha_t^T \alpha_t \qquad (30) \\ \text{s.t.} \quad & \nu \in \Delta_{k,\bar{q}},\ \mu_j \in \Delta_{p_j},\ \mathbf{L}_j \in O^{p_j}.\end{aligned}$$

Denoting $\mathbf{L}_j^T \Lambda_j \mathbf{L}_j$ by $\overline{\mathbf{Q}}_j$ and eliminating variables ν, μ, and $\mathbf{L}$'s lead to

$$\begin{aligned}\min_{\overline{\mathbf{Q}}} \quad & \sum_{t=1}^{T} \max_{\alpha_t} -\alpha_t^T y_t - \frac{1}{2}\alpha_t^T \left(\sum_{j=1}^{k} \Phi_{tj}^T \overline{\mathbf{Q}}_j \Phi_{tj}\right)\alpha_t \\ & - \frac{1}{2\lambda}\alpha_t^T \alpha_t \qquad (31) \\ \text{s.t.} \quad & \overline{\mathbf{Q}}_j \succeq 0,\ \sum_{j=1}^{k} \left(\operatorname{tr}\left(\overline{\mathbf{Q}}_j\right)\right)^{\bar{q}} \le 1.\end{aligned}$$

The difficulty in working with this formulation is that the explicit mappings ϕ_j's are required. We now describe a way of overcoming this problem and efficiently kernelizing the formulation (refer to [1] also). Let $\Phi_j \equiv [\Phi_{1j}, \ldots, \Phi_{Tj}]$ and the compact SVD of Φ_j be $\mathbf{U}_j \Sigma_j \mathbf{V}_j^T$. Then, we introduce a symmetric positive semidefinite $\mathbf{Q}_j$ with the same rank as that of Φ_j such that $\overline{\mathbf{Q}}_j = \mathbf{U}_j \mathbf{Q}_j \mathbf{U}_j^T$. By eliminating $\overline{\mathbf{Q}}_j$, we can rewrite the above problem using $\mathbf{Q}_j$ as

$$\begin{aligned}\min_{\mathbf{Q}} \quad & \sum_{t=1}^{T} \max_{\alpha_t} -\alpha_t^T y_t - \frac{1}{2}\alpha_t^T \left(\sum_{j=1}^{k} \mathbf{M}_{tj}^T \mathbf{Q}_j \mathbf{M}_{tj}\right)\alpha_t \\ & - \frac{1}{2\lambda}\alpha_t^T \alpha_t \qquad (32) \\ \text{s.t.} \quad & \mathbf{Q}_j \succeq 0,\ \sum_{j=1}^{k} \left(\operatorname{tr}\left(\mathbf{Q}_j\right)\right)^{\bar{q}} \le 1,\end{aligned}$$

where $\mathbf{M}_{tj} = \Sigma_j^{-1} \mathbf{V}_j^T \Phi_j^T \Phi_{tj}$. Note that calculation of $\mathbf{M}_{tj}$ does not require the kernel-induced features explicitly and hence the formulation is kernelized. It can be transformed as follows:

$$\min_{\mathbf{Q}} \quad f(\mathbf{Q}) = \sum_{t=1}^{T} -\alpha_t^T y_t - \frac{1}{2}\operatorname{tr}(\mathbf{QB}) - \frac{1}{2\lambda}\alpha_t^T \alpha_t, \qquad (33)$$

where $\mathbf{B}$ is a block diagonal matrix with entries as $\mathbf{B}_j = \sum_{t=1}^{T} \mathbf{M}_{tj} \alpha_t \alpha_t^T \mathbf{M}_{tj}^T$.

$\mathbf{Q}$ can be solved by mirror-descent. The gradient of ∇f with respect to $\mathbf{Q}$ is calculated as follows:

$$\nabla f\left(\mathbf{Q}^{(l)}\right) = -\frac{1}{2}\mathbf{B}^{(l)}, \qquad (34)$$

where $\mathbf{B}^{(l)}$ is the value obtained using optimal α_t obtained while evaluating $f(\mathbf{Q}^{(l)})$.

The algorithm $\ell_{2,1}$-ℓ_q MKMTL is summarized in Algorithm 3.

5. Experimental Results and Discussions

5.1. Experimental Setup. We use 10-fold cross valuation to evaluate our model and conduct the comparison. In each of ten trials, a 5-fold nested cross validation procedure is employed to tune the regularization parameters. Data was z-scored before applying regression methods. The range of each parameter varied from 10^{-1} to 10^{3}. The candidate kernels are as follows: six different kernel bandwidths $(2^{-2}, 2^{-1}, \ldots, 2^{3})$, polynomial kernels of degrees 1 to 3, and a linear kernel, which totally yields 10 kernels. The kernel matrices were precomputed and normalized to have unit trace. The reported results were the best results of each method with the optimal parameter. For the quantitative performance evaluation, we employed the metrics of Correlation Coefficient (CC) and Root Mean Squared Error (rMSE) between the predicted clinical scores and the target clinical scores for each regression task. Moreover, to evaluate the overall performance on all the tasks, the normalized mean squared error (nMSE) [7, 18] and weighted R-value (wR) [4] are used. The nMSE and wR are defined as follows:

$$\text{nMSE}\left(Y, \widehat{Y}\right) = \frac{\sum_{t=1}^{T}\left(\left\|Y_t - \widehat{Y}_t\right\|_2^2 / \sigma\left(Y_t\right)\right)}{\sum_{t=1}^{T} m_t}, \qquad (35)$$

$$\text{wR}\left(Y, \widehat{Y}\right) = \frac{\sum_{t=1}^{T} \operatorname{Corr}\left(Y_t, \widehat{Y}_t\right) m_t}{\sum_{t=1}^{T} m_t}, \qquad (36)$$

where Y and $\widehat{Y}$ are the ground truth cognitive scores and the predicted cognitive scores, respectively.

A smaller (higher) value of nMSE and rMSE (CC and wR) represents better regression performance. We report the mean and standard deviation based on 10 iterations of experiments on different splits of data for all comparable experiments.

In ADNI, all participants received 1.5-Tesla (T) structural MRI. The MRI features used in our experiments are based on the imaging data from the ADNI database processed by a team from UCSF (University of California at San Francisco), who performed cortical reconstruction and volumetric segmentations with the FreeSurfer image analysis suite (http://surfer.nmr.mgh.harvard.edu/) according to the atlas generated in [19]. Totally, 48 cortical regions and 44 subcortical regions are generated. For each cortical region, the cortical thickness average (TA), standard deviation of thickness (TS), surface area (SA), and cortical volume (CV) were calculated as features. For each subcortical region, subcortical volume was calculated as features. The SA of left and right hemisphere and total intracranial volume (ICV) were also included. This yielded a total of $p = 319$ MRI features extracted from cortical/subcortical ROIs in each hemisphere (including 275 cortical and 44 subcortical features). Details of the analysis procedure are available at http://adni.loni.usc.edu/methods/mri-analysis/.

Ten widely used clinical/cognitive assessment scores [3, 20, 21] were employed in this study, including Alzheimer's Disease Assessment Scale (ADAS) cognitive total score, Mini Mental State Exam (MMSE) score, Rey Auditory Verbal Learning Test (RAVLT) involving total score of the first 5 learning trials (TOTAL), Trial 6 total number of words recalled (TOT6), 30-minute delay score (T30), and 30-minute delay recognition score (RECOG), FLU involving animal total score (ANIM) and vegetable total score (VEG), and TRAILS including Trail Making test A score and B score.

5.2. Comparison with the State-of-the-Art MTL Methods. To compare the kernelized MTL with the other linearized one and illustrate how well the two multikernel-based MTL methods work by means of modeling the correlation among the tasks, we comprehensively compare our proposed methods with several popular state-of-the-art related methods. Representative comparable algorithms include

(1) Ridge [22]: $\min_{\Theta} L(X, Y, \Theta) + \lambda\|\Theta\|_F^2$

(2) Lasso [23]: $\min_{\Theta} L(X, Y, \Theta) + \lambda\|\Theta\|_1$

(3) MKL [24]: $\min_{\widetilde{\theta},\xi} (1/2)\|f\|_{\mathscr{H}}^2 + \lambda \sum_i \xi_i$, such that $y_i(f(x_i) + b) \geq 1 - \xi_i$ and $\xi_i \geq 0,\ \forall i$

(4) Robust Multitask Feature Learning (RMTL) [25]: RMTL ($\min_{\Theta} L(X, Y, \Theta) + \lambda_1\|P\|_* + \lambda_2\|S\|_{2,1}$, subject to $\Theta = P + S$), which assumes that the model Θ can be decomposed into two components: a shared feature structure P capturing task relatedness and a group-sparse structure S detecting outliers

(5) Clustered Multitask Learning (CMTL) [16]: CMTL ($\min_{\Theta,M:M^TM=I_c} L(X, Y, \Theta) + \lambda_1(\mathrm{tr}(\Theta^T\Theta) - \mathrm{tr}(M^T\Theta^T\Theta M)) + \lambda_2\mathrm{tr}(\Theta^T\Theta)$, where $M \in \mathbb{R}^{c\times k}$ is an orthogonal cluster indicator matrix and the tasks are clustered into $c < k$ clusters) incorporating a regularization term to induce clustering between tasks and then sharing information only to tasks belonging to the same cluster. In the CMTL, the number of clusters is set to 11 since the 20 tasks belong to 11 sets of cognitive functions

(6) Trace-norm regularized multitask learning (Trace) [17]: assuming that all models share a common low-dimensional subspace ($\min_{\Theta} L(X, Y, \Theta) + \lambda\|\Theta\|_*$)

(7) Sparse regularized multitask learning formulation (SRMTL) [26]: SRMTL ($\min_{\Theta} L(X, Y, \Theta) + \lambda_1\|\Theta\mathscr{Z}\|_F^2 + \lambda_2\|\Theta\|_1$, where $\mathscr{Z} \in \mathbb{R}^{T\times T}$) containing two regularization processes: (1) all tasks are regularized by their mean value, and therefore knowledge from one task can be utilized by other tasks via the mean value; (2) sparsity is enforced in the learning with ℓ_1-norm.

Experimental results are reported in Tables 1 and 2 where the best results are boldfaced. A first glance at the results shows that $\ell_{2,1}\ell_q$-MKMTL generally outperforms all the other compared methods on both metrics and across all the cognitive tasks. Additionally, a statistical analysis is performed on the results. As can be seen, our proposed method achieves statistically significant results compared to all the other methods on most of the results. These results reveal several interesting points:

(1) All the compared multitask learning methods ($\ell_q\ell_1$-MTL, $\ell_q\ell_1$-MKMTL, and $\ell_{2,1}\ell_q$-MKMTL) improve the predictive performance over the independent regression algorithms (Ridge, Lasso, and MKL). This justifies the motivation of learning multiple tasks simultaneously.

(2) The two multikernel-based MTL methods outperform the linearized $\ell_q\ell_1$-MTL in terms of nMSE, and $\ell_{2,1}\ell_q$-MKMTL outperforms the linearized $\ell_q\ell_1$-MTL in terms of wR. It indicates that the nonlinear MTL models via kernel functions can capture complex patterns between brain images and the corresponding cognitive measures.

(3) By the appropriate $\ell_{2,1}\ell_q$ regularization, the $\ell_{2,1}\ell_q$-MKMTL model enables us (1) to obtain capture nonlinear associations between MRI and cognitive outcomes, (2) to obtain the intrinsic relationships between multiple related tasks in $\mathscr{H}$, and (3) to promote the sparse kernel combinations to support the interpretability and scalability. The outcomes demonstrate that $\ell_{2,1}\ell_q$-MKMTL outperforms $\ell_q\ell_1$-MTL and $\ell_q\ell_1$-MKMTL, both of which neglect the inherently nonlinear relationship between MRI and cognitive outcomes, and the correlation among multiple related tasks in the feature space.

(4) Compared with the other multitask learning methods with different assumptions, our proposed methods belong to the multitask feature learning methods with

Table 1: Performance comparison of various methods in terms of rMSE and nMSE on 10 cross validation cognitive prediction tasks.

Method	ADAS	MMSE	RAVLT			
			TOTAL	TOT6	T30	RECOG
Ridge	7.556 ± 0.294	2.656 ± 0.134	11.41 ± 0.498	3.907 ± 0.236	4.052 ± 0.224	4.331 ± 0.294
Lasso	6.846 ± 0.361	2.216 ± 0.098	10.02 ± 0.548	3.320 ± 0.195	3.443 ± 0.177	3.639 ± 0.213
MKL	6.893 ± 0.528	2.214 ± 0.106	9.911 ± 0.695	3.424 ± 0.296	3.570 ± 0.340	3.745 ± 0.237
Robust MTL	7.651 ± 0.442	3.326 ± 0.266	11.02 ± 0.590	3.574 ± 0.235	3.704 ± 0.171	3.858 ± 0.310
CMTL	7.642 ± 0.373	3.083 ± 0.461	11.56 ± 0.510	3.907 ± 0.260	4.038 ± 0.244	4.381 ± 0.226
Trace	8.180 ± 0.605	6.113 ± 2.038	13.09 ± 3.128	3.782 ± 0.491	3.906 ± 0.431	4.520 ± 0.859
SRMTL	6.882 ± 0.325	2.331 ± 0.271	9.961 ± 0.561	3.320 ± 0.152	3.445 ± 0.116	3.639 ± 0.261
$\ell_q\ell_1$-MTL	6.772 ± 0.312	2.206 ± 0.081	9.606 ± 0.448	3.344 ± 0.154	3.440 ± 0.151	3.644 ± 0.247
$\ell_q\ell_1$-MKMTL	6.825 ± 0.455	2.417 ± 0.197	9.699 ± 0.505	3.396 ± 0.188	3.495 ± 0.144	3.653 ± 0.243
$\ell_{2,1}\ell_q$-MKMTL	6.806 ± 0.447	2.185 ± 0.106	9.628 ± 0.510	3.331 ± 0.196	3.467 ± 0.172	3.627 ± 0.199

Method	FLU		TRAILS		nMSE
	ANIM	VEG	A	B	
Ridge	6.521 ± 0.418	4.322 ± 0.178	27.18 ± 1.702	83.72 ± 5.713	16.44 ± 1.725
Lasso	5.352 ± 0.447	3.701 ± 0.093	23.75 ± 1.398	71.23 ± 2.812	12.05 ± 0.758
MKL	5.342 ± 0.510	3.761 ± 0.137	24.71 ± 1.781	78.09 ± 6.916	13.56 ± 1.133
Robust MTL	5.946 ± 0.398	3.988 ± 0.083	27.78 ± 1.922	90.12 ± 7.098	17.68 ± 2.303
CMTL	6.608 ± 0.561	4.398 ± 0.284	27.46 ± 1.980	83.66 ± 5.418	16.67 ± 1.912
Trace	6.743 ± 1.425	4.672 ± 0.778	28.82 ± 3.278	89.68 ± 7.838	20.23 ± 5.215
SRMTL	5.327 ± 0.334	3.713 ± 0.088	25.09 ± 1.421	80.00 ± 4.637	14.01 ± 1.169
$\ell_q\ell_1$-MTL	5.298 ± 0.439	3.704 ± 0.096	23.42 ± 1.110	71.32 ± 2.945	11.92 ± 0.969
$\ell_q\ell_1$-MKMTL	5.304 ± 0.350	3.676 ± 0.094	23.09 ± 1.438	70.28 ± 0.898	11.72 ± 0.222
$\ell_{2,1}\ell_q$-MKMTL	5.232 ± 0.434	3.675 ± 0.157	23.13 ± 1.473	69.82 ± 1.236	**11.56 ± 0.602**

Table 2: Performance comparison of various methods in terms of CC and wR on 10 cross validation cognitive prediction tasks.

Method	ADAS	MMSE	RAVLT			
			TOTAL	TOT6	T30	RECOG
Ridge	0.603 ± 0.031	0.407 ± 0.040	0.401 ± 0.084	0.361 ± 0.092	0.377 ± 0.096	0.261 ± 0.080
Lasso	0.655 ± 0.036	0.540 ± 0.046	0.493 ± 0.084	0.507 ± 0.100	0.523 ± 0.106	0.416 ± 0.087
MKL	0.658 ± 0.030	0.544 ± 0.052	0.502 ± 0.066	0.476 ± 0.095	0.506 ± 0.105	0.391 ± 0.072
Robust MTL	0.587 ± 0.022	0.338 ± 0.084	0.423 ± 0.090	0.432 ± 0.096	0.444 ± 0.094	0.354 ± 0.105
CMTL	0.603 ± 0.025	0.381 ± 0.042	0.397 ± 0.072	0.362 ± 0.090	0.381 ± 0.099	0.260 ± 0.068
Trace	0.548 ± 0.039	0.144 ± 0.091	0.342 ± 0.172	0.395 ± 0.159	0.402 ± 0.142	0.253 ± 0.130
SRMTL	0.655 ± 0.034	0.525 ± 0.058	0.492 ± 0.079	0.505 ± 0.097	0.523 ± 0.103	0.413 ± 0.092
$\ell_q\ell_1$-MTL	0.662 ± 0.043	0.532 ± 0.056	0.532 ± 0.082	0.492 ± 0.109	0.522 ± 0.105	0.404 ± 0.091
$\ell_q\ell_1$-MKMTL	0.661 ± 0.034	0.460 ± 0.099	0.519 ± 0.072	0.470 ± 0.089	0.494 ± 0.094	0.412 ± 0.090
$\ell_{2,1}\ell_q$-MKMTL	0.660 ± 0.035	0.547 ± 0.045	0.529 ± 0.079	0.500 ± 0.095	0.508 ± 0.094	0.421 ± 0.075

Method	FLU		TRAILS		wR
	ANIM	VEG	A	B	
Ridge	0.185 ± 0.090	0.396 ± 0.073	0.291 ± 0.097	0.330 ± 0.110	0.361 ± 0.041
Lasso	0.365 ± 0.096	0.506 ± 0.059	0.363 ± 0.041	0.467 ± 0.096	0.484 ± 0.049
MKL	0.375 ± 0.071	0.496 ± 0.067	0.374 ± 0.056	0.457 ± 0.060	0.478 ± 0.046
Robust MTL	0.253 ± 0.096	0.443 ± 0.057	0.282 ± 0.113	0.292 ± 0.123	0.385 ± 0.038
CMTL	0.180 ± 0.089	0.390 ± 0.071	0.287 ± 0.116	0.335 ± 0.112	0.358 ± 0.036
Trace	0.212 ± 0.143	0.331 ± 0.112	0.270 ± 0.112	0.290 ± 0.122	0.319 ± 0.083
SRMTL	0.362 ± 0.093	0.503 ± 0.064	0.340 ± 0.063	0.361 ± 0.095	0.468 ± 0.045
$\ell_q\ell_1$-MTL	0.379 ± 0.076	0.501 ± 0.063	0.399 ± 0.060	0.467 ± 0.098	0.489 ± 0.050
$\ell_q\ell_1$-MKMTL	0.381 ± 0.080	0.521 ± 0.067	0.421 ± 0.064	0.481 ± 0.076	0.482 ± 0.047
$\ell_{2,1}\ell_q$-MKMTL	0.409 ± 0.073	0.516 ± 0.065	0.417 ± 0.067	0.490 ± 0.087	**0.500 ± 0.043**

Table 3: Performance comparison of various methods with fusing multiple modalities data in terms of rMSE and nMSE on 10 cross validation cognitive prediction tasks.

Method	ADAS	MMSE	FLU	TRAILS	
			ANIM	A	B
$\ell_q\ell_1$-MTL-MRI	6.494 ± 1.029	1.964 ± 0.306	4.911 ± 0.256	16.39 ± 2.906	55.82 ± 7.689
$\ell_q\ell_1$-MTL-PET	6.941 ± 1.244	2.118 ± 0.298	5.192 ± 0.145	16.56 ± 3.533	56.88 ± 9.447
$\ell_q\ell_1$-MTL-MP	6.219 ± 1.037	2.067 ± 0.293	4.928 ± 0.260	16.09 ± 2.768	53.70 ± 7.144
$\ell_q\ell_1$-MTL-ALL	6.174 ± 0.978	2.062 ± 0.272	4.789 ± 0.206	15.97 ± 2.785	53.37 ± 7.243
$\ell_q\ell_1$-MKMTL-MRI	6.369 ± 0.941	2.074 ± 0.291	4.993 ± 0.235	16.18 ± 3.089	55.95 ± 9.479
$\ell_q\ell_1$-MKMTL-PET	6.812 ± 1.155	2.060 ± 0.364	5.151 ± 0.227	16.61 ± 3.588	57.85 ± 11.24
$\ell_q\ell_1$-MKMTL-MP	6.112 ± 0.886	2.005 ± 0.258	4.966 ± 0.269	16.13 ± 2.988	54.13 ± 9.450
$\ell_q\ell_1$-MKMTL-ALL	5.960 ± 0.834	1.959 ± 0.256	4.821 ± 0.224	16.00 ± 3.062	53.48 ± 9.592
$\ell_{2,1}\ell_q$-MKMTL-MRI	6.425 ± 0.951	1.951 ± 0.308	4.886 ± 0.264	16.11 ± 2.939	54.96 ± 7.499
$\ell_{2,1}\ell_q$-MKMTL-PET	6.783 ± 1.059	2.058 ± 0.323	5.107 ± 0.258	16.52 ± 3.515	55.51 ± 9.568
$\ell_{2,1}\ell_q$-MKMTL-MP	6.086 ± 0.987	1.917 ± 0.299	4.855 ± 0.249	15.95 ± 2.996	52.44 ± 8.074
$\ell_{2,1}\ell_q$-MKMTL-ALL	6.034 ± 0.978	1.905 ± 0.294	4.809 ± 0.244	15.88 ± 3.028	52.20 ± 8.120

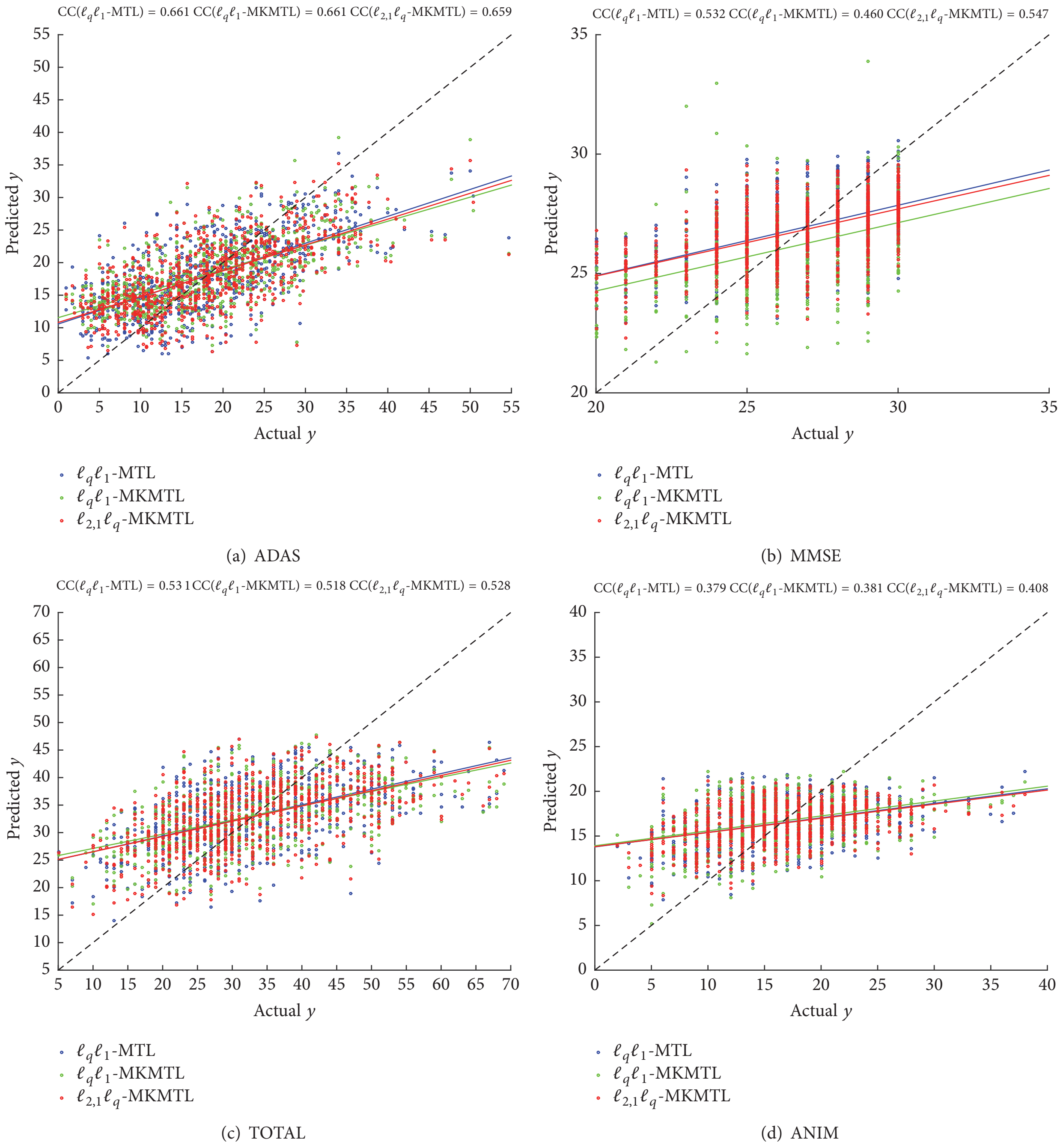
Method	RAVLT				nMSE
	TOTAL	TOT6	T30	RECOG	
$\ell_q\ell_1$-MTL-MRI	10.18 ± 0.640	3.538 ± 0.147	3.735 ± 0.199	3.169 ± 0.306	10.24 ± 0.735
$\ell_q\ell_1$-MTL-PET	10.41 ± 0.441	3.627 ± 0.140	3.796 ± 0.176	3.258 ± 0.360	10.72 ± 1.163
$\ell_q\ell_1$-MTL-MP	10.01 ± 0.556	3.501 ± 0.149	3.693 ± 0.196	3.164 ± 0.314	9.710 ± 0.627
$\ell_q\ell_1$-MTL-ALL	9.755 ± 0.575	3.450 ± 0.151	3.643 ± 0.200	3.172 ± 0.313	9.525 ± 0.608
$\ell_q\ell_1$-MKMTL-MRI	10.09 ± 0.605	3.532 ± 0.081	3.731 ± 0.253	3.203 ± 0.304	10.21 ± 1.019
$\ell_q\ell_1$-MKMTL-PET	10.30 ± 0.436	3.592 ± 0.145	3.754 ± 0.231	3.200 ± 0.357	10.82 ± 1.455
$\ell_q\ell_1$-MKMTL-MP	9.787 ± 0.375	3.471 ± 0.089	3.664 ± 0.199	3.159 ± 0.302	9.713 ± 0.968
$\ell_q\ell_1$-MKMTL-ALL	9.350 ± 0.460	3.402 ± 0.030	3.604 ± 0.221	3.196 ± 0.291	9.410 ± 0.985
$\ell_{2,1}\ell_q$-MKMTL-MRI	9.984 ± 0.525	3.477 ± 0.130	3.678 ± 0.204	3.143 ± 0.314	9.937 ± 0.753
$\ell_{2,1}\ell_q$-MKMTL-PET	10.19 ± 0.410	3.565 ± 0.146	3.745 ± 0.212	3.191 ± 0.351	10.31 ± 1.105
$\ell_{2,1}\ell_q$-MKMTL-MP	9.727 ± 0.467	3.397 ± 0.136	3.593 ± 0.162	3.112 ± 0.323	9.282 ± 0.869
$\ell_{2,1}\ell_q$-MKMTL-ALL	9.561 ± 0.442	3.361 ± 0.124	3.556 ± 0.170	3.104 ± 0.327	**9.160 ± 0.860**

sparsity-inducing norms, having an advantage over the other comparative multitask learning methods. Since not all the brain regions are associated with AD, many of the features are irrelevant and redundant. Sparse based MTL methods are appropriate for the task of predicting cognitive measures and better than the non-sparse-based MTL methods.

We also show the scatter plots of actual values versus predicted values for the score of ADAS, MMSE, TOTAL, and ANIM on testing data in Figure 1.

5.3. Multimodalities Fusion. To estimate the effect of combining multimodality image data with the linearized and kernelized MTL methods and provide a more comprehensive comparison of the results from the comparable MTL models, we further perform some experiments, and they are (1) using only MRI modality, (2) using only PET modality, (3) combining two modalities: PET and MRI (MP), and (4) combining three modalities: PET, MRI, and demographic information including age, gender, years of education, and ApoE genotyping (MPD). Different from the above experiments, the samples from ADNI-2 are used instead of ADNI-1, since the amount of the patients with PET is sufficient. From the ADNI-2, we obtained all the patients with both MRI and PET, totally 756 samples. The PET imaging data are from the ADNI database processed by the UC Berkeley team, who use a native-space MRI scan for each subject that is segmented and parcellated with FreeSurfer to generate a summary cortical and subcortical ROI, and they coregister each florbetapir scan to the corresponding MRI and calculate the mean florbetapir uptake within the cortical and reference regions. The procedure of image processing is described in http://adni.loni.usc.edu/updated-florbetapir-av-45-pet-analysis-results/. In the $\ell_q\ell_1$-MKMTL and $\ell_{2,1}\ell_q$-MKMTL, ten different kennel functions described in the first experiment are used for each modality. To show the advantage of the kernel-based methods, we compare them with linear $\ell_q\ell_1$-MTL method, which concatenated the multiple modalities features into a long vector features.

The prediction performance results are shown in Tables 3 and 4. From the results, it is clear that the methods with



(a) ADAS

(b) MMSE

(c) TOTAL

(d) ANIM

Figure 1: Scatter plots of actual versus predicted values of cognitive scores on each fold testing data using three comparable MTL methods based on MRI features.

multimodality outperform the methods using one single modality of data. This validates our assumption that the complementary information among different modalities is helpful for cognitive function prediction. Regardless of two or three modalities, $\ell_{2,1}\ell_q$-MKMTL achieved better performances than the linear based multitask learning for the most cases, the same as for the single modality learning task above.

6. Conclusion

Many multitask learning methods with sparsity-inducing regularization for modeling AD cognitive outcomes have been proposed in the past decades. However, the current formulations remain restricted to the linear models and cannot capture the relationship between the MRI features and cognitive outcomes. To address these shortcomings, we applied two multikernel multitask learning methods with a joint sparsity-inducing regularization to model the more complicated but more flexible relationship between MRI features and cognitive outcomes and demonstrated their effectiveness compared with linearized multitask learning methods by applying them to the ADNI data for predicting cognitive outcomes from MRI scans. Extensive experiments

Table 4: Performance comparison of various methods with fusing multiple modalities data in terms of CC and wR on 10 cross validation cognitive prediction tasks.

Method	ADAS	MMSE	FLU	TRAILS	
			ANIM	A	B
$\ell_q\ell_1$-MTL-MRI	0.670 ± 0.091	0.539 ± 0.117	0.481 ± 0.112	0.417 ± 0.115	0.525 ± 0.073
$\ell_q\ell_1$-MTL-PET	0.619 ± 0.058	0.482 ± 0.087	0.395 ± 0.105	0.385 ± 0.120	0.501 ± 0.060
$\ell_q\ell_1$-MTL-MP	0.700 ± 0.070	0.549 ± 0.108	0.486 ± 0.119	0.437 ± 0.119	0.567 ± 0.070
$\ell_q\ell_1$-MTL-ALL	0.705 ± 0.067	0.560 ± 0.096	0.527 ± 0.102	0.450 ± 0.115	0.575 ± 0.064
$\ell_q\ell_1$-MKMTL-MRI	0.677 ± 0.093	0.512 ± 0.113	0.464 ± 0.095	0.411 ± 0.113	0.529 ± 0.094
$\ell_q\ell_1$-MKMTL-PET	0.634 ± 0.056	0.493 ± 0.100	0.410 ± 0.133	0.375 ± 0.090	0.478 ± 0.061
$\ell_q\ell_1$-MKMTL-MP	0.710 ± 0.060	0.537 ± 0.106	0.472 ± 0.111	0.426 ± 0.105	0.566 ± 0.081
$\ell_q\ell_1$-MKMTL-ALL	0.727 ± 0.062	0.551 ± 0.112	0.512 ± 0.097	0.444 ± 0.099	0.582 ± 0.065
$\ell_{2,1}\ell_q$-MKMTL-MRI	0.673 ± 0.096	0.548 ± 0.124	0.491 ± 0.095	0.422 ± 0.135	0.528 ± 0.102
$\ell_{2,1}\ell_q$-MKMTL-PET	0.631 ± 0.057	0.488 ± 0.108	0.418 ± 0.119	0.386 ± 0.095	0.524 ± 0.065
$\ell_{2,1}\ell_q$-MKMTL-MP	0.714 ± 0.067	0.566 ± 0.107	0.499 ± 0.094	0.437 ± 0.122	0.583 ± 0.077
$\ell_{2,1}\ell_q$-MKMTL-ALL	0.721 ± 0.064	0.574 ± 0.105	0.512 ± 0.094	0.445 ± 0.120	0.589 ± 0.073

Method	RAVLT				wR
	TOTAL	TOT6	T30	RECOG	
$\ell_q\ell_1$-MTL-MRI	0.576 ± 0.077	0.536 ± 0.085	0.516 ± 0.041	0.444 ± 0.079	0.523 ± 0.082
$\ell_q\ell_1$-MTL-PET	0.548 ± 0.103	0.497 ± 0.124	0.490 ± 0.092	0.409 ± 0.098	0.481 ± 0.081
$\ell_q\ell_1$-MTL-MP	0.593 ± 0.079	0.547 ± 0.086	0.529 ± 0.038	0.450 ± 0.075	0.540 ± 0.077
$\ell_q\ell_1$-MTL-ALL	0.618 ± 0.072	0.563 ± 0.077	0.546 ± 0.027	0.446 ± 0.085	0.554 ± 0.069
$\ell_q\ell_1$-MKMTL-MRI	0.585 ± 0.069	0.533 ± 0.093	0.511 ± 0.044	0.434 ± 0.077	0.517 ± 0.079
$\ell_q\ell_1$-MKMTL-PET	0.559 ± 0.110	0.508 ± 0.111	0.503 ± 0.085	0.432 ± 0.081	0.488 ± 0.075
$\ell_q\ell_1$-MKMTL-MP	0.617 ± 0.080	0.561 ± 0.100	0.541 ± 0.057	0.462 ± 0.079	0.543 ± 0.075
$\ell_q\ell_1$-MKMTL-ALL	0.654 ± 0.071	0.577 ± 0.082	0.560 ± 0.038	0.444 ± 0.087	0.561 ± 0.068
$\ell_{2,1}\ell_q$-MKMTL-MRI	0.594 ± 0.070	0.554 ± 0.080	0.536 ± 0.033	0.459 ± 0.071	0.534 ± 0.082
$\ell_{2,1}\ell_q$-MKMTL-PET	0.563 ± 0.104	0.510 ± 0.111	0.501 ± 0.081	0.436 ± 0.095	0.495 ± 0.072
$\ell_{2,1}\ell_q$-MKMTL-MP	0.621 ± 0.075	0.582 ± 0.083	0.564 ± 0.046	0.475 ± 0.073	0.560 ± 0.071
$\ell_{2,1}\ell_q$-MKMTL-ALL	0.637 ± 0.068	0.593 ± 0.077	0.575 ± 0.041	0.479 ± 0.081	**0.570 ± 0.067**

on ADNI dataset illustrate that the multikernel multitask learning method not only yields superior performance on regression performance but also is a powerful tool for fusing multimodalities data.

Conflicts of Interest

The authors declare that they have no conflicts of interest regarding the publication of this paper.

Acknowledgments

This research was supported by the National Natural Science Foundation of China (no. 61502091), the Fundamental Research Funds for the Central Universities (no. N161604001 and no. N150408001), and the National Science Foundation for Distinguished Young Scholars of China under Grants no. 71325002 and no. 61225012.

References

[1] Z. S. Khachaturian, “Diagnosis of Alzheimer’s disease,” *JAMA Neurology*, vol. 42, no. 11, pp. 1097–1105, 1985.

[2] A. Wimo, B. Winblad, H. Aguero-Torres, and E. von Strauss, “The magnitude of dementia occurrence in the world,” *Alzheimer Disease & Associated Disorders*, vol. 17, no. 2, pp. 63–67, 2003.

[3] J. Wan, Z. Zhang, B. D. Rao et al., “Identifying the neuroanatomical basis of cognitive impairment in Alzheimer’s disease by correlation-and nonlinearity-aware sparse bayesian learning,” *IEEE Transactions on Medical Imaging*, vol. 33, no. 7, pp. 1475–1487, 2014.

[4] C. M. Stonnington, C. Chu, S. Klöppel, C. R. Jack, J. Ashburner, and R. S. J. Frackowiak, “Predicting clinical scores from magnetic resonance scans in Alzheimer’s disease,” *NeuroImage*, vol. 51, no. 4, pp. 1405–1413, 2010.

[5] J. Ye, M. Farnum, E. Yang et al., “Sparse learning and stability selection for predicting MCI to AD conversion using baseline ADNI data,” *BMC Neurology*, vol. 12, article no. 46, no. 1, 2012.

[6] Y. Zhang and D.-Y. Yeung, “A convex formulation for learning task relationships in multi-task learning,” in *Proceedings of the 26th Conference on Uncertainty in Artificial Intelligence, UAI 2010*, pp. 733–742, Catalina Island, California, USA, July 2010.

[7] A. Argyriou, T. Evgeniou, and M. Pontil, “Convex multi-task feature learning,” *Machine Learning*, vol. 73, no. 3, pp. 243–272, 2008.

[8] H. Wang, F. Nie, H. Huang et al., “Sparse multi-task regression and feature selection to identify brain imaging predictors for memory performance,” in *Proceedings of the 2011 IEEE International Conference on Computer Vision, ICCV 2011*, pp. 557–562, Barcelona, Spain, November 2011.

[9] H. Wang, F. Nie, H. Huang et al., “High-Order Multi-Task Feature Learning to Identify Longitudinal Phenotypic Markers for Alzheimers Disease Progression Prediction,” in *in Advances in Neural Information Processing Systems (NIPS)*, 2012.

[10] B. Gu and V. S. Sheng, “A robust regularization path algorithm for ν-support vector classification,” *IEEE Transactions on Neural Networks and Learning Systems*, 2016.

[11] B. Gu, V. S. Sheng, K. Y. Tay, W. Romano, and S. Li, “Incremental support vector learning for ordinal regression,” *IEEE Transactions on Neural Networks and Learning Systems*, vol. 26, no. 7, pp. 1403–1416, 2015.

[12] M. Gönen and E. Alpaydın, “Multiple kernel learning algorithms,” *Journal of Machine Learning Research*, vol. 12, pp. 2211–2268, 2011.

[13] P. Jawanpuria and J. S. Nath, “Multi-task multiple kernel learning,” in *Proceedings of the 11th SIAM International Conference on Data Mining, SDM 2011*, pp. 828–838, Mesa, Arizona, USA, April 2011.

[14] A. Argyriou, T. Evgeniou, and M. Pontil, “Multi-task feature learning,” in *in: Advances in neural information processing systems*, pp. 41–48, 2007.

[15] J. Liu and J. Ye, “Efficient l1/lq norm regularization,” Arxiv preprint arXiv 1009.

[16] J. Zhou, J. Chen, and J. Ye, “Clustered multi-task learning via alternating structure optimization in,” *Advances in Neural Information Processing Systems*, pp. 702–710, 2011.

[17] S. Ji and J. Ye, “An accelerated gradient method for trace norm minimization,” in *Proceedings of the 26th International Conference On Machine Learning (ICML ’09)*, pp. 457–464, ACM, June 2009.

[18] J. Zhou, J. Liu, V. A. Narayan, and J. Ye, “Modeling disease progression via multi-task learning,” *NeuroImage*, vol. 78, pp. 233–248, 2013.

[19] R. S. Desikan, F. Ségonne, B. Fischl et al., “An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest,” *NeuroImage*, vol. 31, no. 3, pp. 968–980, 2006.

[20] X. Liu, P. Cao, D. Zhao, and A. Banerjee, “Multi-task Spare Group Lasso for Characterizing AlzheimerΓs Disease in,” *5th Workshop on Data Mining for Medicine and Healthcare*, p. 49, 2016.

[21] J. Wan, Z. Zhang, J. Yan et al., “Sparse Bayesian multi-task learning for predicting cognitive outcomes from neuroimaging measures in Alzheimer’s disease,” in *Proceedings of the 2012 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2012*, pp. 940–947, USA, June 2012.

[22] N. R. Draper and H. Smith, *Applied Regression Analysis*, John Wiley & Sons, New York, NY, USA, 3rd edition, 1981.

[23] J. Liu and J. Ye, “Efficient Euclidean projections in linear time,” in *Proceedings of the 26th International Conference On Machine Learning, ICML 2009*, pp. 657–664, can, June 2009.

[24] A. Rakotomamonjy, F. R. Bach, S. Canu, and Y. Grandvalet, “Simple MKL,” *Journal of Machine Learning Research*, vol. 9, pp. 2491–2521, 2008.

[25] J. Chen, J. Zhou, and J. Ye, “Integrating low-rank and group-sparse structures for robust multi-task learning,” in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD’11*, pp. 42–50, San Diego, Calif, USA, August 2011.

[26] J. Zhou, “Multi-task learning in crisis event classification,” Tech. Rep., http://www. public. asu. edu/jzhou29.