ORIGINAL ARTICLE

Graph-guided joint prediction of class label and clinical scores for the Alzheimer's disease

Guan Yu[1] · Yufeng Liu[1,2,3] · Dinggang Shen[4,5]

Received: 23 October 2014 / Accepted: 4 October 2015

Abstract Accurate diagnosis of Alzheimer's disease and its prodromal stage, i.e., mild cognitive impairment, is very important for early treatment. Over the last decade, various machine learning methods have been proposed to predict disease status and clinical scores from brain images. It is worth noting that many features extracted from brain images are correlated significantly. In this case, feature selection combined with the additional correlation information among features can effectively improve classification/regression performance. Typically, the correlation information among features can be modeled by the connectivity of an undirected graph, where each node represents one feature and each edge indicates that the two involved features are correlated significantly. In this paper, we propose a new graph-guided multi-task learning method incorporating this undirected graph information to predict multiple response variables (i.e., class label and clinical scores) jointly. Specifically, based on the sparse undirected feature graph, we utilize a new latent group Lasso penalty to encourage the correlated features to be selected together. Furthermore, this new penalty also encourages the intrinsic correlated tasks to share a common feature subset. To validate our method, we have performed many numerical studies using simulated datasets and the Alzheimer's Disease Neuroimaging Initiative dataset. Compared with the other methods, our proposed method has very promising performance.

Keywords Alzheimer's disease · Group Lasso · Magnetic resonance imaging (MRI) · Multi-task learning · Partial correlation · Positron emission tomography (PET) · Undirected graph

Introduction

Alzheimer's disease (AD) is one of the most common forms of dementia characterized by progressive cognitive and memory deficits. It has been reported that one in every 85 persons in year 2050 will be likely affected by this disease (Brookmeyer et al. 2007). The increasing incidence of AD makes this disease a very important health issue and also huge financial burden for both patients and governments (Hebert et al. 2001; Bain et al. 2008). Thus, it is very important to develop methods for timely diagnosis of AD and its predromal stage, i.e., mild cognitive impairment (MCI). Over the last decade, many machine learning methods have been used for early diagnosis of AD and MCI based on different modalities of biomarkers, e.g., structural brain atrophy delineated by structural magnetic

✉ Dinggang Shen
dgshen@med.unc.edu

Guan Yu
guanyu@live.unc.edu

Yufeng Liu
yfliu@email.unc.edu

[1] Department of Statistics and Operations Research, The University of North Carolina at Chapel Hill, Chapel Hill, NC 27599, USA

[2] Carolina Center for Genome Sciences, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[3] Department of Biostatistics, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[4] Department of Radiology and BRIC, The University of North Carolina at Chapel Hill, Chapel Hill, NC, USA

[5] Department of Brain and Cognitive Engineering, Korea University, Seoul 02841, Republic of Korea

Published online: 17 October 2015

resonance imaging (MRI) (Du et al. 2007; McEvoy et al. 2009; Fjell et al. 2010; Yu et al. 2014), metabolic alterations characterized by fluorodeoxyglucose positron emission tomography (FDG-PET) (De Santi et al. 2001; Morris et al. 2001), and pathological amyloid depositions measured by cerebrospinal fluid (CSF) (Bouwman et al. 2007; Fjell et al. 2010). Typically, these methods learn a binary classification model from training data and use this model to predict disease status (i.e., class label) of the testing subjects.

Besides classification of disease status, accurate prediction of clinical scores such as Mini-Mental State Examination (MMSE) score and Alzheimer's Disease Assessment Scale-Cognitive Subscale (ADAS-Cog) is also important and useful since they can help evaluate the stage of AD pathology and predict future progression. Specifically, as a brief 30-point questionnaire test, MMSE is commonly used to screen for cognitive impairment. It can be used to examine a patient's arithmetic, memory and orientation (Folstein et al. 1975). As another important clinical score of AD, ADAS-Cog is a cognitive testing instrument widely used in clinical trials. It is designed to measure the severity of the most important symptoms of AD (Rosen et al. 1984). Several studies based on regression methods have been conducted to estimate MMSE and ADAS-Cog using the extracted features from MRI and FDG-PET. For example, Duchesne et al. (2005) used linear regression models, Wang et al. (2010) developed a high-dimensional kernel-based regression method, and Cheng et al. (2013) proposed a semi-supervised multi-modal relevance vector regression method. However, almost all of these regression methods model different clinical scores separately and do not use the class label information which is often available in practice.

Although the classification of disease status and the prediction of clinical scores are different tasks, there exists inherent correlation among them since the underlying pathology is the same (Fan et al. 2010; Stonnington et al. 2010). In the literature, Zhang and Shen (2012) proposed multi-modal multi-task (M3T) learning to predict both class label and clinical scores jointly. M3T formulated the estimations of class label and clinical scores as different tasks. The $l_{2,1}$ penalty was used to deliver sparse models with a common feature subset for each task. Their experimental results indicated that selecting a common feature subset for different correlated tasks could achieve better prediction of both class label and clinical scores than choosing the feature subset for each task separately. Although benefiting from using the commonality among different correlated tasks, M3T method did not incorporate the correlation information among features. Actually, many features extracted from brain images such as structural MRI are statistically correlated significantly. In this case, feature selection combined with the additional correlation information among features can improve classification/regression performance (Yang et al. 2012).

In this paper, we extract effective correlation information among features by constructing a sparse undirected feature graph. This undirected graph uses all features as nodes. Also, two features are connected by an edge in the graph if there is statistically significant partial correlation between them. In practice, we can use many existing high-dimensional precision matrix estimation methods (Friedman et al. 2008; Cai et al. 2011) to construct this undirected graph. Based on this undirected feature graph, we propose a new graph-guided multi-task learning (GGML) method to predict both class label and clinical scores simultaneously. Specifically, we utilize a new latent group Lasso penalty to encourage the significantly correlated features to be in or out of the models together. This new penalty also encourages the intrinsic correlated tasks to share a common feature subset. It is very useful for us to acquire robust and accurate feature selection. Computationally, the optimization problem for our proposed GGML method can be solved by the traditional group Lasso algorithm efficiently (Yuan and Lin 2006). Theoretically, our proposed GGML method includes M3T method as a special case. To validate our proposed GGML method, we have conducted many numerical studies using simulated datasets and the Alzheimer's Disease Neuroimaging Initiative (ADNI) (http://www.loni.ucla.edu/ADNI) dataset. Compared with the other methods, our proposed GGML method acquired very promising results.

The remainder of this paper is organized as follows. In the "Materials" section, we introduce the ADNI dataset used in this study. In the "Method" section, we show how to extract useful correlation information among features and describe our proposed new method. In "Simulation study" and "Analysis of the ADNI dataset" sections, we compare our method with the other methods by simulation study and also the analysis of the ADNI dataset. In the "Discussion" section, we discuss some possible extensions of our proposed method. Finally, we conclude this paper in the "Conclusion" section.

Materials

Data

Data used in this paper were obtained from the ADNI database. As a $ 60 million, 5-year public–private partnership, the ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the

Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations. The main goal of ADNI was to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessments could be combined to measure the progression of MCI and early AD. To that end, 800 adults with age between 55 and 90 were recruited from over 50 sites across the US and Canada. Approximately, 200 cognitively normal controls (NC) and 400 MCI individuals were followed for 3 years and 200 individuals with early AD were followed for 2 years (see http://www.adni-info.org for up-to-date information). The general inclusion/exclusion criteria of the subjects are described in Zhang and Shen (2012). In this paper, we use data from 199 subjects who have complete baseline MRI, FDG-PET, and CSF data. These 199 subjects include 50 AD subjects, 97 MCI subjects, and 52 NC subjects. The detailed demographic information about these 199 subjects is summarized in Table 1.

Data preprocessing

Imaging preprocessing was performed for MRI and PET. For MRI, the preprocessing steps include anterior commissure (AC)–posterior commissure (PC) correction, intensity inhomogeneity correction (Sled et al. 1998), skull stripping (Wang et al. 2011), cerebellum removal based on registration with atlas, spatial segmentation (Zhang et al. 2001) and registration (Shen and Davatzikos 2002). After registration, we obtained the subject-labeled image based on the Jacob template (Kabani et al. 1998) with 93 manually labeled regions of interest (ROI). For each of the 93 ROIs in the labeled MRI, we computed the volume of gray matter as a feature. For each PET image, we first aligned the PET image to its respective MRI using affine registration. Then, we got the skull-stripping image using the corresponding brain mask of MRI and computed the average intensity of every ROI in the PET image as a feature. Besides MRI and PET, we used CSF Aβ42, CSF t-tau and CSF p-tau as CSF features. For each subject, we finally obtained 93 MRI features, 93 PET features, and 3 CSF features. We also had the class label, MMSE and ADAS-Cog scores for each subject.

Methods

In this section, after introducing some notations, we will first discuss how to extract the correlation information among features. Next, in order to show how to utilize this correlation information clearly, we first introduce the graph-guided single-task learning (GGSL) method. Then, as an extension of this method, our proposed graph-guided multi-task learning method will be described.

Notation

For a set $\mathcal{A}$, we denote $|\mathcal{A}|$ as the number of elements in $\mathcal{A}$. For a matrix $\mathbf{B}$, we denote $\mathbf{B}^T$ and $\mathbf{B}^{-1}$ as the transpose and the inverse of matrix $\mathbf{B}$, respectively. We also denote $\|\mathbf{B}\|_F = \sqrt{\sum_i \sum_j \mathbf{B}_{ij}^2}$ as the Frobenius norm.

Suppose we have n samples and p features. Let $\mathbf{X} = (\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_p}) = (\mathbf{x_1}, \mathbf{x_2}, \ldots, \mathbf{x_n})^{\mathbf{T}}$ denote the $n \times p$ training data matrix of features, where $x_1, x_2, \ldots, x_n$ are i.i.d. samples generated from a p-dimensional multivariate distribution with mean vector $0_{p\times 1}$ and covariance matrix $\mathbf{\Sigma} = (\sigma_{ij})_{i,j=1}^p$. Also, let $\mathbf{\Omega} = (\omega_{ij})_{i,j=1}^p = \mathbf{\Sigma}^{-1}$ denote the precision matrix. Furthermore, suppose we have q response variables. Let $\mathbf{Y} = (\mathbf{Y_1}, \mathbf{Y_2}, \ldots, \mathbf{Y_q}) = (\mathbf{y_1}, \mathbf{y_2}, \ldots, \mathbf{y_n})^{\mathbf{T}}$ denote the $n \times q$ training data matrix of response variables, where the response variables can be binary (for classification) or continuous (for regression). Note that, for the ADNI dataset used in our study, we have three response variables, which are class label, MMSE score, and ADAS-Cog score. The class labels are coded as +1 and −1 for the binary classification problem considered in this paper.

Extract the correlation information among features

The correlation information is often measured by the Pearson correlation between each pair of features. We can use sample Pearson correlation coefficients to identify the statistically significant correlated features. One issue with this method is that it only estimates the marginal linear dependence between a pair of features without considering the influence of other features and common driving influences. Such issue can be overcome by using partial

Table 1 Demographic information of the 199 subjects used in this study

Characteristics	AD (50 subjects)	MCI (97 subjects)	NC (52 subjects)
Gender (F/M)	17/33	32/65	18/34
Age (mean ± SD)	75.2 ± 7.6	75.3 ± 7.0	75.1 ± 5.1
Education (mean ± SD)	14.7 ± 3.7	15.9 ± 2.9	15.8 ± 3.2
MMSE (mean ± SD)	23.7 ± 1.9	27.1 ± 1.7	29.0 ± 1.2
ADAS (mean ± SD)	18.5 ± 5.9	11.4 ± 4.4	7.36 ± 3.2

correlation which measures the linear dependence between each pair of features after eliminating the linear effect of the other features. In practice, we can compute the sample partial correlation coefficient between features i and j, denoted as $\hat{\rho}^*_{ij}$, which is defined as the sample Pearson correlation coefficient between the residuals R_i and R_j resulting from the linear regression of the feature X_i with features $\{X_k : k \neq i,j\}$ and of the feature X_j with features $\{X_k : k \neq i,j\}$, respectively. The resulting $\hat{\rho}^*_{ij}$s can be further used to identify features which are partially correlated statistically significantly.

When the number of features p is small and the sample size n is big enough (bigger than p), it is easy to get good estimates of partial correlation coefficients. In this case, many previous studies (Hampson et al. 2002; Lee et al. 2011) have used partial correlations to identify the significant correlated features. However, in the high-dimensional case with the number of features p bigger than the sample size n, the conventional methods for estimating partial correlation may result in over-fitting of the data (Ryali et al. 2012). In this case, it is difficult to get accurate estimates of partial correlation coefficients.

For our proposed method introduced in the next section, in order to incorporate the correlation information among features, instead of requiring accurate estimation of ρ^*_{ij}s, we only need to estimate which pairs of features are partially correlated, i.e., estimate the set $\mathcal{E} = \{(i,j) : i < j \text{ and } \rho^*_{ij} \neq 0\}$. It is well known that the partial correlation coefficients are proportional to the off-diagonal entries of the precision matrix $\mathbf{\Omega}$ (Meinshausen and Bühlmann 2006). Thus, estimating $\mathcal{E}$ is equivalent to estimating the set $\{(i,j) : i < j \text{ and } \omega_{\mathrm{ij}} \neq 0\}$. In this way, many existing methods (Meinshausen and Bühlmann 2006; Friedman et al. 2008; Cai et al. 2011) can be used to estimate $\mathcal{E}$ effectively.

In this paper, we will use the graphical Lasso (Friedman et al. 2008) or the neighborhood selection method (Meinshausen and Bühlmann 2006) to estimate $\mathcal{E}$ and denote its estimate as $\hat{\mathcal{E}}$. Furthermore, we represent $\hat{\mathcal{E}}$ as an undirected graph $\mathbf{G}$ with p nodes and $|\hat{\mathcal{E}}|$ edges, where each node represents one feature and each edge indicates that two involved features are partially correlated significantly. Figure 1 shows an example on how to transform the estimated precision matrix $\hat{\mathbf{\Omega}}$ into the estimated undirected graph $\mathbf{G}$. In the graph $\mathbf{G}$, features i and j are connected if and only if $\hat{\omega}_{ij} \neq 0$.

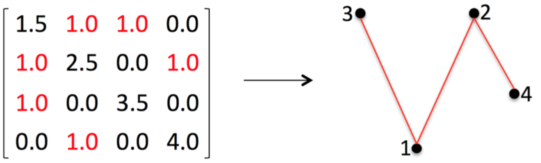


Fig. 1 Transforming the precision matrix $\hat{\mathbf{\Omega}}$ (*left*) into the undirected graph $\mathbf{G}$ (*right*). Features i and j are connected if and only if $\hat{\omega}_{ij} \neq 0$

Graph-guided single-task learning (GGSL) method

In this section, we assume that the undirected feature graph $\mathbf{G}$ has been constructed. For each $i = 1, 2, \ldots, p$, denote $\mathcal{N}_i$ as the set including the ith feature and its neighbors in the graph $\mathbf{G}$, i.e., $\mathcal{N}_i = \{j : \hat{\omega}_{ji} \neq 0\}$.

To show how to use the correlation information represented by $\mathbf{G}$, we consider the single-task learning first and then generalize this idea to multi-task learning. Without loss of generality, considering the tth task, we want to use the following linear model to predict the response variable Y_t,

$$Y_t = \mathbf{X}\mathbf{B_t} + \mathbf{\epsilon_t}, \tag{1}$$

where $B_t = (b_{1t}, b_{2t}, \ldots, b_{pt})^T \in R^p$ is the coefficient vector of interest and $\epsilon_t = (\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{nt}) \in R^n$ is the error vector with $\mathrm{E}(\epsilon_{\mathrm{st}}) = 0$ and $\mathrm{Var}(\epsilon_{\mathrm{st}}) = \sigma^2_{\mathrm{t}}$ for each $1 \leq s \leq n$.

Suppose the feature matrix $\mathbf{X}$ is independent of the error vector ϵ_t. Denote C_t as the marginal correlation vector between p features and the response variable Y_t, i.e., $C_t = \mathrm{E}(\mathbf{X^T Y_t / n}) = (\mathbf{c_{1t}}, \mathbf{c_{2t}}, \ldots, \mathbf{c_{pt}})^{\mathbf{T}} \in \mathbf{R^p}$. Then by (1), we have

$$C_t = \mathrm{E}(\mathbf{X^T Y_t / n}) = \mathrm{E}(\mathbf{X^T X B_t / n}) + \mathrm{E}(\mathbf{X^T} \mathbf{\epsilon_t} \mathbf{/ n}) = \mathbf{\Sigma B_t}. \tag{2}$$

Thus, the true coefficient vector B_t can be represented as

$$B_t = \Sigma^{-1} C_t = \Omega C_t, \tag{3}$$

where Ω shows the partial correlations among different features, and C_t reflects the marginal correlations between the features and the tth response variable Y_t.

Furthermore, the Eq. (3) can be expanded as follows:

$$\begin{aligned}
b_{1t} &= \omega_{11} c_{1t} + \omega_{12} c_{2t} + \cdots + \omega_{1i} c_{it} + \cdots + \omega_{1p} c_{pt} \\
b_{2t} &= \omega_{21} c_{1t} + \omega_{22} c_{2t} + \cdots + \omega_{2i} c_{it} + \cdots + \omega_{2p} c_{pt} \\
&\vdots \\
b_{pt} &= \omega_{p1} c_{1t} + \omega_{p2} c_{2t} + \cdots + \omega_{pi} c_{it} + \cdots + \omega_{pp} c_{pt}.
\end{aligned} \tag{4}$$

We observe that the coefficients vector $B_t = (b_{1t}, b_{2t}, \ldots, b_{pt})^T$ is the sum of p parts, where the ith part, $(\omega_{1i} c_{it}, \omega_{2i} c_{it}, \ldots, \omega_{pi} c_{it})^T$, is the ith vertical part in the right side of the above equations (4). In addition, for each i, if there is no marginal correlation between the ith feature and the response variable Y_t, i.e., $c_{it} = 0$, then the components in the ith part $(\omega_{1i} c_{it}, \omega_{2i} c_{it}, \ldots, \omega_{pi} c_{it})^T$ will be zero simultaneously due to the common factor c_{it}. Furthermore, if the ith feature and the response variable Y_t are correlated marginally, then $c_{it} \neq 0$ and the set of candidate nonzero components in the ith part is $\{j : \omega_{ji} \neq 0\}$, which can be well estimated by the set $\mathcal{N}_i$ including the ith feature and its neighbors in the estimated undirected graph $\mathbf{G}$.

Motivated by the decompositions shown in Eq. (4), we assume that there is a latent decomposition of the coefficients vector B_t into p parts, $V^{1t}, \ldots, V^{it}, \ldots, V^{pt}$, where V^{it} is a p-dimensional latent vector representing the ith vertical part in the right side of Eq. (4). In order to incorporate the correlation information represented by the undirected graph $\mathbf{G}$, a group penalty term will be used to encourage the ith latent vector V^{it} to be zero or have nonzero components only for the indices in the set $\mathcal{N}_i$. Hence, we use the following (GGSL method to estimate B_t:

$$\min_{B_t, V^{1t}, V^{2t}, \ldots, V^{pt} \in R^p} \|Y_t - \mathbf{X}\mathbf{B_t}\|_2^2 + \lambda \sum_{\mathbf{i=1}}^{\mathbf{p}} \tau_{\mathbf{it}} \|\mathbf{V^{it}}\|_2, \tag{5}$$

subject to $B_t = \sum_{i=1}^{p} V^{it}$ and $\mathrm{supp}(\mathrm{V}^{\mathrm{it}}) \subseteq \mathcal{N}_{\mathrm{i}}$ for each $1 \leq i \leq p$, where $\mathrm{supp}(\mathrm{V}^{\mathrm{it}})$ is the index set of the nonzero components in the vector V^{it}.

In the optimization problem (5), τ_{it} is a positive weight for the ith part and tth task. Similar to the methods for adaptive Lasso (Zou 2006) and group Lasso (Yuan and Lin 2006), we can set $\tau_{it} = \frac{\sqrt{|\mathcal{N}_i|}}{|\tilde{b}_{it}|^{\gamma}}$ where γ is a positive parameter and $\tilde{b}_{it}$ is an initial estimate of b_{it}. In our experiments, we chose $\tilde{b}_{it}$ as the sample correlation coefficient between X_i and Y_t. Both the positive parameter γ and the tuning parameter λ were chosen by cross-validation. Our experimental results indicate that this method can acquire good performance in general.

Theoretically, the GGSL method is very general and covers the popular Lasso method as a special case. Specifically, if we ignore the correlation information among features, we can set the undirected graph $\mathbf{G}$ as an empty graph with no edge. In this case, if setting constant weights τ_{it}s, we can show that $\sum_{i=1}^{p} \tau_{it} \|V^{it}\|_2 \propto |B_t|_1$, and the GGSL method is the same as the Lasso method (Tibshirani 1996). In general, we can estimate a sparse undirected graph $\mathbf{G}$ for modeling the significant partial correlation information among features. The GGSL method can utilize this correlation information effectively and thus acquires good prediction performance.

Graph-guided multi-task learning (GGML) method

For the multi-task learning, we aim at estimating q response variables simultaneously. Similar to the above GGSL method, for each t, we assume that the coefficient vector B_t can be decomposed as $B_t = \sum_{i=1}^{p} V^{it}$, where each V^{it} is a p-dimensional latent vector satisfying $\mathrm{supp}(\mathrm{V}^{\mathrm{it}}) \subseteq \mathcal{N}_{\mathrm{i}}$. Furthermore, in order to make use of the intrinsic correlation among these q tasks (response variables), we also assume that the decompositions of q coefficient vectors $B_1, B_2, \ldots, B_q$ have the same pattern, i.e., $\mathrm{supp}(\mathrm{V}^{\mathrm{i1}}) = \mathrm{supp}(\mathrm{V}^{\mathrm{i2}}) = \cdots = \mathrm{supp}(\mathrm{V}^{\mathrm{iq}})$ for each $1 \leq i \leq p$. That is, for each $i = 1, 2, \ldots, p$, we assume that, if both the ith feature and its partially correlated features are useful for prediction of one response variable, they are also useful for prediction of the other response variables.

Based on the above assumption, denote $\mathbf{B} = (\mathbf{B_1}, \mathbf{B_2}, \ldots, \mathbf{B_q}) \in \mathbf{R^{p \times q}}$ and $\mathbf{V^i} = (\mathbf{V^{i1}}, \mathbf{V^{i2}}, \ldots, V^{iq}) \in R^{p \times q}$ for each $1 \leq i \leq p$, we generalize the GGSL method to the following GGML method:

$$\min_{\mathbf{B}, \mathbf{V^1}, \mathbf{V^2}, \ldots, \mathbf{V^p} \in \mathbf{R^{p \times q}} \|\mathbf{Y} - \mathbf{X}\mathbf{B}\|_{\mathbf{F}}^2 + \lambda \sum_{\mathbf{i=1}}^{\mathbf{p}} \tau_{\mathbf{i}} \|\mathbf{V^i}\|_{\mathbf{F}},} \tag{6}$$

subject to $\mathbf{B} = \sum_{\mathbf{i=1}}^{\mathbf{p}} \mathbf{V^i}$ and $\{j : \|\mathbf{V^i_{j\cdot}}\|_2 \neq \mathbf{0}\} \subseteq \mathcal{N}_{\mathbf{i}}$ for each $1 \leq i \leq p$, where $\mathbf{V^i_{j\cdot}}$ is the jth row of the matrix $\mathbf{V^i}$.

Similar to the GGSL method discussed in "Graph-guided single-task learning (GGSL) method" section, we can set the weight $\tau_i = \frac{\sqrt{|\mathcal{N}_i|}}{\max_{1 \leq t \leq q} |\tilde{b}_{it}|^{\gamma}}$. The cross-validation method can be used to choose the best γ and the best tuning parameter λ for different tasks separately. Note that the penalty term in (6) along with the additional constraints not only encourage the significantly partially correlated features to be in or out of the model jointly, but also choose a common feature subset for different tasks. Due to the use of both the correlation information among features and the intrinsic commonality among different related tasks, our proposed GGML method can acquire better prediction performance than the methods not using or only using part of these two kinds of information.

As an interesting remark, we note that the M3T method (Zhang and Shen 2012) is a special case of our proposed GGML method. In particular, when we ignore the correlation information among features, we can set the undirected graph $\mathbf{G}$ as an empty graph with no edge. In this case, if setting constant weights τ_is, we can show that $\sum_{i=1}^{p} \tau_i \|\mathbf{V}^i\|_F \propto \sum_{i=1}^{p} \|\mathbf{B}_{i\cdot}\|_2$, where $\mathbf{B}_{i\cdot}$ is the ith row of the coefficient matrix $\mathbf{B}$. Thus, our proposed GGML method is exactly the same as the M3T method using the $l_{2,1}$ penalty.

Objective function optimization

For our proposed GGML method, we need to solve the optimization problem (6). We can transform this constrained optimization problem into a simple unconstrained optimization problem by feature duplication.

Denote $\mathbf{X}_{\cdot \mathcal{N}_{\mathrm{i}}}$ as the sub-matrix of $\mathbf{X}$ with column indices in $\mathcal{N}_i$, and denote $\mathbf{V}^{\mathbf{i}}_{\mathcal{N}_{\mathrm{i}}\cdot}$ as the sub-matrix of $\mathbf{V^i}$ with row indices in $\mathcal{N}_i$. Furthermore, denote $\tilde{\mathbf{X}} = (\mathbf{X}_{\cdot\mathcal{N}_1}, \mathbf{X}_{\cdot\mathcal{N}_2}, \ldots, \mathbf{X}_{\cdot\mathcal{N}_\mathbf{p}}) \in \mathbf{R}^{\mathbf{n}\times(\sum_{\mathbf{i}=1}^{\mathbf{p}}|\mathcal{N}_\mathbf{i}|)}$ as the duplicated feature matrix and $\tilde{\mathbf{V}} = ((\mathbf{V}^{\mathbf{1}}_{\mathcal{N}_\mathbf{1}\cdot})^\mathbf{T}, (\mathbf{V}^{\mathbf{2}}_{\mathcal{N}_\mathbf{2}\cdot})^\mathbf{T}, \ldots, (\mathbf{V}^{\mathbf{p}}_{\mathcal{N}_\mathbf{p}\cdot})^\mathbf{T})^\mathbf{T}$ as the $(\sum_{i=1}^{p}|\mathcal{N}_i|) \times q$ coefficient matrix. Then, we can check that $\mathbf{XB} = \tilde{\mathbf{X}}\tilde{\mathbf{V}}$ and (6) is equivalent to the following unconstrained optimization problem:

$$\min_{\tilde{\mathbf{V}}} \|\mathbf{Y} - \tilde{\mathbf{X}}\tilde{\mathbf{V}}\|_F^2 + \lambda \sum_{i=1}^{p} \tau_i \|\mathbf{V}^i_{\mathcal{N}_i\cdot}\|_F. \tag{7}$$

The above problem (7) is a traditional group Lasso problem which can be solved efficiently by the blockwise majorization decent algorithm (Yang and Zou 2013). Denote the estimate of $\mathbf{B}$ as $\hat{\mathbf{B}}$. In the application stage, given a testing subject x^*, for the tth task, we can estimate Y_t^* by $\hat{Y}_t^* = \text{sign}(\hat{B}_t^T x^*)$ if Y_t^* is a class label and by $\hat{Y}_t^* = \hat{B}_t^T x^*$ if Y_t^* is a continuous response variable.

Simulation study

In this section, we perform numerical studies using simulated examples. For each example, we compare our proposed GGML method with (1) the Lasso method which learns different tasks separately; (2) the GGSL method which uses the correlation information among features and learns different tasks separately, and (3) M3T method which learns different tasks jointly while ignoring the correlation information among features. We implement Lasso, GGSL, and M3T methods as shown in "Objective function optimization" section to predict the response variables.

Similar to the measures used in Zhang and Shen (2012), the classification accuracy and the Pearson's correlation coefficient (CC) are also used here to evaluate the classification and regression performances, respectively. In addition, we also use the root-mean-square error (RMSE) to evaluate the regression performance.

Simulated examples

We study three simulated examples. Each example has one classification task and two regression tasks. We set $p = 100$, $B_1 = (2, 2, \ldots, 2, 0, 0, \ldots, 0)^T$, $B_2 = B_3 = (1, 1, \ldots, 1, 0, 0, \ldots, 0)^T$, where only the first 15 elements of each B_t $(t = 1, 2, 3)$ are nonzero. For each t, the errors $\epsilon_{1t}, \epsilon_{2t}, \ldots, \epsilon_{nt} \overset{i.i.d.}{\sim} N(0, 9)$. For $s = 1, 2, \ldots, n$, the feature vector $(x_{s1}, x_{s2}, \ldots, x_{sp})^T$ is generated as follows.

Example 1 For $1 \leq j \leq 5$, $x_{sj} = z_1 + 0.4\epsilon_j^x$. For $6 \leq j \leq 10$, $x_{sj} = z_2 + 0.4\epsilon_j^x$. For $11 \leq j \leq 15$, $x_{sj} = z_3 + 0.4\epsilon_j^x$. For $16 \leq j \leq p$, $x_{sj} \overset{i.i.d}{\sim} N(0, 1)$. Here, $z_1, z_2, z_3, \epsilon_1^x, \epsilon_2^x, \ldots, \epsilon_{15}^x \overset{i.i.d}{\sim} N(0, 1)$.

Example 2 The features $(x_{s1}, x_{s2}, \ldots, x_{sp})^T \sim N(0, \mathbf{\Sigma})$ with $\sigma_{ij} = 0.5^{|i-j|}$. For this example, we have $\omega_{ii} = 1.333$, $\omega_{ij} = -0.667$ if $|i-j| = 1$ and $\omega_{ij} = 0$ if $|i-j| > 1$.

Example 3 The features $\{x_{sj} : 1 \leq j \leq 15\}$ are generated from the same model as shown in Example 1. In addition, the features $(x_{s16}, x_{s17}, \ldots, x_{sp}) \sim N(0, \tilde{\mathbf{\Omega}}^{-1})$, where $\tilde{\mathbf{\Omega}} = \mathbf{M} + \delta\mathbf{I}$. Each off-diagonal entry in $\mathbf{M}$ is generated independently and equals 0.5 with probability 0.05 or 0 with probability 0.95. The diagonal entry of $\mathbf{M}$ is 0. Here, δ is chosen such that the conditional number of $\tilde{\mathbf{\Omega}}$ is equal to $p - 15$. Finally, $\tilde{\mathbf{\Omega}}$ is standardized to have unit diagonals.

After generating each column of the response matrix $\mathbf{Y}$ by model (1), we replace the elements in the first column of $\mathbf{Y}$ by their signs (positive or negative) to simulate class labels. For all examples, we generate 40 training samples, 40 validation samples, and 400 testing samples. All the models are fitted on the training data. The validation data are used to choose the tuning parameters and the testing data are used to evaluate different methods. For each example, we repeat the simulation 30 times.

Figure 2 shows the binary maps of the true precision matrices and Fig. 3 shows the corresponding feature graphs of these three examples. All these three graphs are sparse. For Examples 1 and 3, useful features (i.e., features with nonzero regression coefficients) are only connected with useful features. For Example 2, one useful feature is connected with one useless feature. In addition, for each example, different tasks are highly correlated since they share the same useful features. It is very interesting to study whether correlation information among features represented by the feature graph and the correlation information among tasks can be incorporated to improve the prediction performance.

Simulation results

Table 2 shows the comparison of different methods using these three simulated examples. As shown in Table 2, for all these three examples, the GGSL method and GGML method acquire better performance than the Lasso method and the M3T method, respectively. This indicates that the extracted partial correlation information from features can be utilized to improve the prediction performance. In addition, the GGML method and M3T method also acquire better performance than the GGSL method and the Lasso method, respectively. It indicates that learning different correlated tasks jointly can also improve the prediction performance. For these three simulated examples, since our proposed GGML method

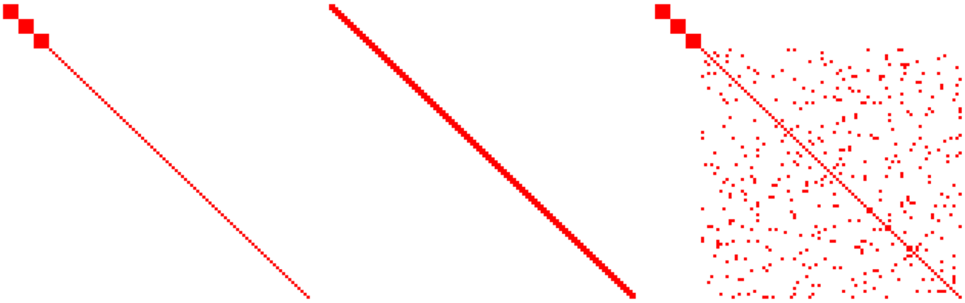


Fig. 2 Binary maps of the true precision matrices corresponding to these three simulated examples: *left* (Example 1), *middle* (Example 2), and *right* (Example 3). Each *red dot* represents a nonzero element in the precision matrix

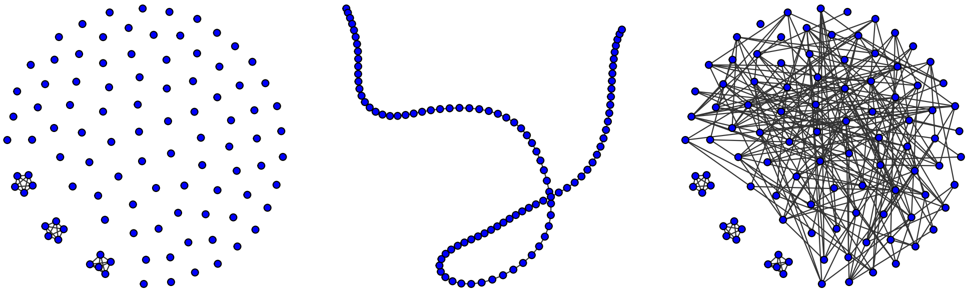


Fig. 3 True feature graphs corresponding to these three simulated examples: *left* (Example 1), *middle* (Example 2), and *right* (Example 3). Each *blue dot* indicates a feature

Table 2 Comparison of different methods using the simulated examples

Example	Method	Accuracy	CC1	CC2	RMSE1	RMSE2
1	Lasso	0.828 (0.007)	0.909 (0.004)	0.910 (0.003)	4.091 (0.070)	4.106 (0.064)
	GGSL	0.848 (0.009)	0.932 (0.003)	0.933 (0.002)	3.548 (0.062)	3.620 (0.057)
	M3T	0.840 (0.006)	0.918 (0.002)	0.917 (0.002)	3.916 (0.059)	4.005 (0.059)
	GGML	**0.872 (0.006)**	**0.938 (0.002)**	**0.936 (0.001)**	**3.402 (0.043)**	**3.488 (0.039)**
2	Lasso	0.765 (0.008)	0.781 (0.010)	0.767 (0.012)	4.567 (0.084)	4.596 (0.089)
	GGSL	0.800 (0.008)	0.823 (0.008)	0.810 (0.010)	4.134 (0.075)	4.213 (0.089)
	M3T	0.796 (0.008)	0.814 (0.008)	0.807 (0.008)	4.261 (0.075)	4.290 (0.075)
	GGML	**0.816 (0.008)**	**0.839 (0.007)**	**0.838 (0.007)**	**3.966 (0.069)**	**3.981 (0.073)**
3	Lasso	0.821 (0.005)	0.910 (0.004)	0.903 (0.005)	3.995 (0.066)	4.163 (0.096)
	GGSL	0.846 (0.008)	0.932 (0.003)	0.927 (0.004)	3.506 (0.063)	3.633 (0.084)
	M3T	0.843 (0.006)	0.918 (0.003)	0.913 (0.004)	3.907 (0.049)	3.992 (0.073)
	GGML	**0.872 (0.006)**	**0.938 (0.002)**	**0.934 (0.002)**	**3.388 (0.045)**	**3.464 (0.050)**

Bold values represent the best performance for a particular measure

CC1 (CC2) is the Pearson's correlation coefficient of the first (second) regression task; RMSE1 (RMSE2) is the root-mean-square error of the first (second) regression task. The values in the parenthesis are standard deviations

incorporates both the partial correlation information among features and the intrinsic correlation information among different related tasks, it delivers the best performance in all cases. In the next section, we will further compare these four methods using the ADNI dataset.

Analysis of the ADNI dataset

For the ADNI dataset, we estimate one class label and two clinical scores (i.e., MMSE and ADAS-Cog) using the MRI, FDG-PET and/or CSF features. Since there are two binary classification problems (AD vs. NC, and MCI vs. NC), we perform two sets of experiments. The first set of experiments uses the AD/NC dataset including only AD and NC subjects. The second set of experiments uses the MCI/NC dataset including only MCI and NC subjects. For each set of experiments, we consider four cases: (I) use only MRI features; (II) use only PET features; (III) use both MRI and PET features (denoted as MRI + PET); (IV) use all MRI, PET and CSF features (denoted as MRI + PET + CSF).

To evaluate the performance of different methods, we used the tenfold cross-validation (CV) strategy.

Specifically, the whole samples were partitioned randomly into ten subsets. Each time only nine subsets were chosen for training and the remaining one was used for testing. We repeated this process ten times with each of the ten subsets used exactly once as the testing data. Furthermore, in consideration of possible bias due to the random partition in the tenfold CV, we repeated the whole 10-CV process 30 times. In the training process, each column of the training data was normalized to have mean 0 and standard deviation 1. For all methods, we performed another inner fivefold CV on the training data to choose the tuning parameters.

Partial correlation among different features

In the first step of the GGSL and GGML methods, we need to extract the effective correlation information from features. Note that, only the training data matrix of features were used to estimate the sparse undirected graph **G** representing the significant partial correlation among features. Figure 4 shows the binary maps of the estimated precision matrices. Binary maps in the first two columns indicate that many features within the same modality (e.g., MRI or PET) are partially correlated statistically significantly. However, as shown by the binary maps in the third column, the partial correlation between MRI features and PET features are not statistically significantly in most cases. Furthermore, the comparison between the binary maps in the first row and the second row indicates that the partial correlation information extracted from AD/NC data is similar to that of MCI/NC data. Similar to the example shown in Fig. 1, we can transform the estimated precision matrices to some undirected graphs. The feature graphs corresponding to the estimated precision matrices are shown in Fig. 5. This graph information will be used in the GGML and GGSL methods.

Classification results

The classification accuracies of different methods are shown in Table 3. All methods deliver higher classification accuracy for the AD/NC dataset than the corresponding classification accuracy for the MCI/NC dataset. For the AD/NC dataset, when we use only MRI features or PET features, the GGSL method and GGML method acquire better classification performance than the Lasso method and the M3T method, respectively. This indicates that the extracted partial correlation information from features can be utilized to improve the classification performance. In addition, when we use both MRI and PET features or all the MRI, PET, and CSF features, since it is relatively easy to discriminate AD subjects from NC subjects in this case, all four methods acquire similar high classification accuracies.

For the MCI/NC dataset, on the one hand, the comparison between GGSL and Lasso (or GGML and M3T) indicates that using the extracted partial correlation information among features improve the classification performance significantly. On the other hand, the comparison between GGML and GGSL (or M3T and Lasso) shows that the joint classification and regression could provide better classification performance than the separate classification. Since our proposed GGML method incorporates both the partial correlation information among features and the intrinsic correlation information among different related tasks, it delivers the best classification performance.

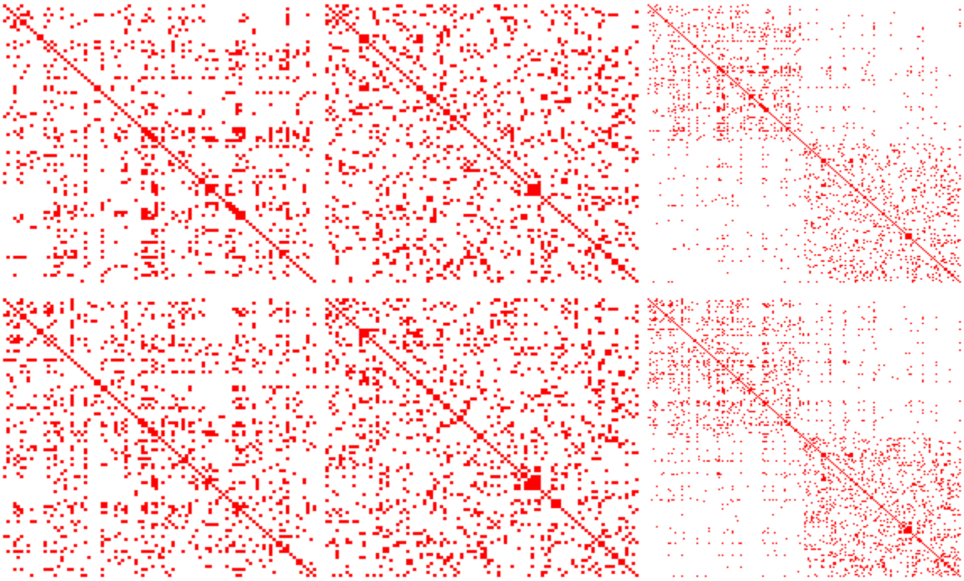


Fig. 4 Binary maps of the estimated precision matrices. *First row* uses AD/NC data; *second row* uses MCI/NC data. *First column* use only MRI features; *second column* use only PET features; *third column* use both MRI and PET features. Each *red dot* in the plot represents a nonzero element

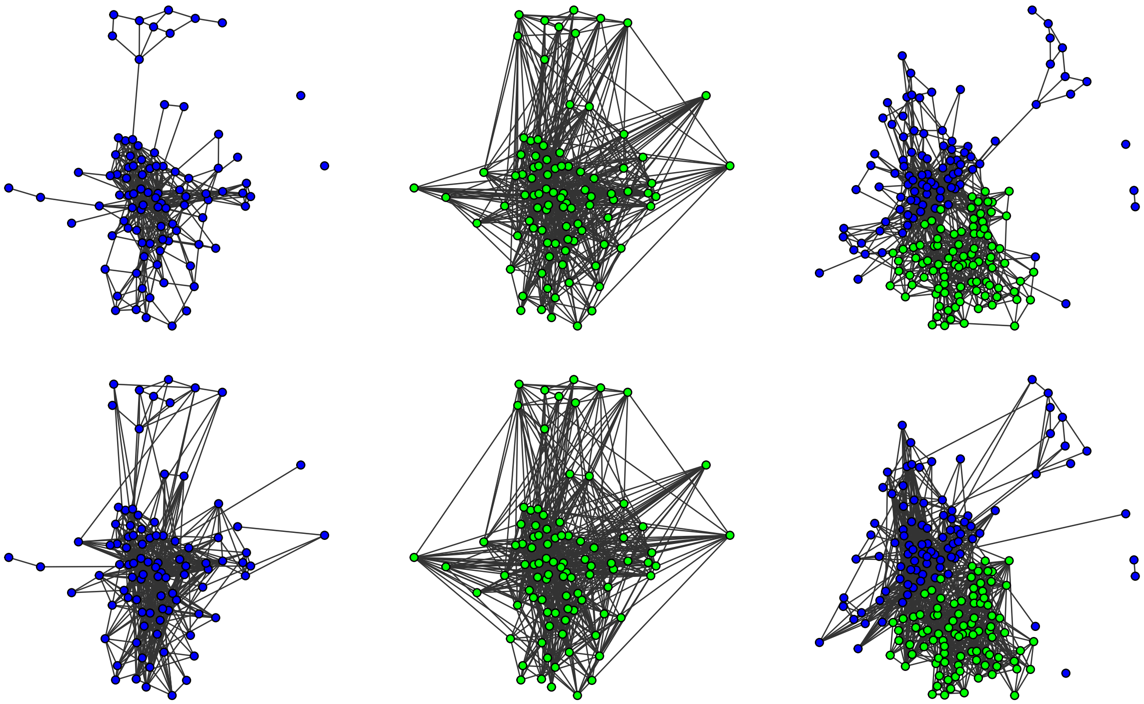


Fig. 5 Feature graphs corresponding to the estimated precision matrices. *First row* uses AD/NC data; *second row* uses MCI/NC data. *First column* use only MRI features; *second column* use only PET features; *third column* use both MRI and PET features. Each *blue dot* represents an MRI feature and each green dot represents a PET feature

Table 3 Comparison of the classification performance on the ADNI dataset

Data	Method	MRI	PET	MRI + PET	MRI + PET + CSF
AD/NC	Lasso	0.878 (0.003)	0.823 (0.003)	0.903 (0.003)	0.917 (0.003)
	GGSL	0.896 (0.003)	0.830 (0.003)	0.911 (0.002)	0.915 (0.002)
	M3T	0.884 (0.002)	0.821 (0.002)	0.914 (0.002)	0.918 (0.002)
	GGML	**0.906 (0.003)**	**0.832 (0.003)**	**0.919 (0.002)**	**0.926 (0.002)**
MCI/NC	Lasso	0.722 (0.003)	0.677 (0.003)	0.737 (0.004)	0.750 (0.004)
	GGSL	0.737 (0.004)	0.688 (0.004)	0.755 (0.005)	0.769 (0.003)
	M3T	0.738 (0.003)	0.655 (0.003)	0.775 (0.003)	0.776 (0.003)
	GGML	**0.751 (0.003)**	**0.696 (0.003)**	**0.784 (0.003)**	**0.800 (0.003)**

Bold values represent the best performance for a particular measure

The reported values are the averaged classification accuracy with standard deviation.

Regression results

For regression tasks, we need to predict both the MMSE score and the ADAS-Cog score. Tables 4 and 5 show the comparison of regression performance on the AD/NC data and the MCI/NC data, respectively. As shown in Tables 4 and 5, our proposed GGML method acquires promising performance in most cases. For example, when we use all the features to predict the MMSE score, for the AD/NC data, our proposed GGML method achieves the highest correlation coefficient 0.745 while the corresponding correlation coefficients for Lasso, GGSL, and M3T are 0.709, 0.723 and 0.724, respectively. For the MCI/NC data, GGML also has the best performance with correlation coefficient 0.382 while the corresponding correlation coefficients for Lasso, GGSL, and M3T are 0.303, 0.325 and 0.364, respectively. In addition, when we use all the features to predict the ADAS-Cog scores, for the AD/NC data, our proposed GGML method achieves the highest correlation coefficient 0.740 while the corresponding correlation coefficients for Lasso, GGSL, and M3T are 0.664, 0.719 and 0.718, respectively. For the MCI/NC data, GGML also has the best performance with correlation coefficient 0.472 while the corresponding correlation coefficients for Lasso, GGSL, and M3T are 0.336, 0.464 and 0.426, respectively.

It is interesting to note that for the MCI/NC dataset, the PET and CSF data seem to be not useful for the prediction

Table 4 Comparison of the regression performance on the AD/NC dataset

Response	Method	MRI	PET	MRI + PET	MRI + PET + CSF
MMSE	Lasso	0.601 (0.005)	0.601 (0.004)	0.688 (0.003)	0.709 (0.003)
	GGSL	0.656 (0.003)	**0.611 (0.003)**	0.698 (0.003)	0.723 (0.003)
	M3T	0.651 (0.004)	0.585 (0.003)	0.693 (0.002)	0.724 (0.002)
	GGML	**0.671 (0.002)**	0.598 (0.003)	**0.712 (0.002)**	**0.745 (0.002)**
ADAS-Cog	Lasso	0.695 (0.003)	0.611 (0.004)	0.652 (0.004)	0.664 (0.004)
	GGSL	0.703 (0.002)	0.632 (0.004)	0.708 (0.003)	0.719 (0.002)
	M3T	0.703 (0.002)	0.635 (0.003)	0.709 (0.003)	0.718 (0.002)
	GGML	**0.705 (0.002)**	**0.644 (0.003)**	**0.721 (0.002)**	**0.740 (0.002)**

Bold values represent the best performance for a particular measure

The reported values are the averaged correlation coefficient with standard deviation.

Table 5 Comparison of the regression performance on the MCI/NC dataset

Response	Method	MRI	PET	MRI + PET	MRI + PET + CSF
MMSE	Lasso	0.326 (0.006)	0.168 (0.010)	0.303 (0.007)	0.303 (0.007)
	GGSL	0.313 (0.007)	0.181 (0.004)	0.323 (0.005)	0.325 (0.005)
	M3T	0.382 (0.004)	0.182 (0.007)	0.379 (0.004)	0.364 (0.004)
	GGML	**0.394 (0.004)**	**0.213 (0.005)**	**0.392 (0.005)**	**0.382 (0.004)**
ADAS-Cog	Lasso	0.355 (0.006)	0.427 (0.006)	0.343 (0.006)	0.336 (0.006)
	GGSL	0.378 (0.005)	0.451 (0.005)	0.462 (0.004)	0.464 (0.003)
	M3T	0.354 (0.004)	0.406 (0.006)	0.429 (0.003)	0.426 (0.003)
	GGML	**0.391 (0.004)**	**0.469 (0.005)**	**0.462 (0.003)**	**0.472 (0.003)**

Bold values represent the best performance for a particular measure.

The reported values are the averaged correlation coefficient with standard deviation.

of MMSE score. All four methods acquire poor prediction of the MMSE scores when only the PET data are used. In addition, compared with the cases only using MRI data, both M3T and GGML methods acquire worse performance when the additional PET/CSF data are used. Similar to the previous discussion about classification performance, the comparison between GGSL and Lasso (or GGML and M3T) indicates that using the extracted partial correlation information among features improves the prediction of MMSE and ADAS-Cog scores significantly. In addition, the comparison between GGML and GGSL (or M3T and Lasso) shows that joint classification and regression could deliver better prediction performance than the separate regression of MMSE (or ADAS-Cog) on the features. Since our GGML method incorporates both the partial correlation information among features and the intrinsic correlation information among different tasks, it delivers the best prediction of the MMSE and ADAS-Cog scores.

Most discriminative brain regions

In this subsection, we investigate the most discriminative brain regions for the diagnosis of disease status and the prediction of the MMSE and ADAS-Cog scores. For each method, we repeated the whole 10-CV process 30 times and acquired 300 different models using different training datasets. Figure 6 shows the selection frequency of each of 93 ROIs for the AD/NC classification task using only MRI features, where the selection frequency for each ROI is defined as

$$\text{Frequency} = 100 \times \frac{\text{The times of being selected in the 300 models}}{300}.$$

For each method, some ROIs are always selected while some ROIs are seldom selected. Compared with Lasso and M3T, the GGSL and GGML methods tend to select more ROIs since they use the feature graph information and encourage the significantly partially correlated features to be selected jointly. According to the selection frequency, we compare the top ten selected ROIs of different methods for different tasks. Tables 6, 7 and 8 show the indices of the top ten selected ROIs of the four methods for different tasks (classification or regression), different datasets (AD/NC or MCI/NC) and different modalities (MRI or PET). Table 9 contains the full names of the ROIs.

As shown in Tables 6, 7 and 8, for different tasks, the top ten selected ROIs of the single-task learning methods such as Lasso and GGSL are different while the top ten selected ROIs of the multi-task learning methods such as M3T and GGML are the same. We can also observe that the top ten selected ROIs for the cases using MRI

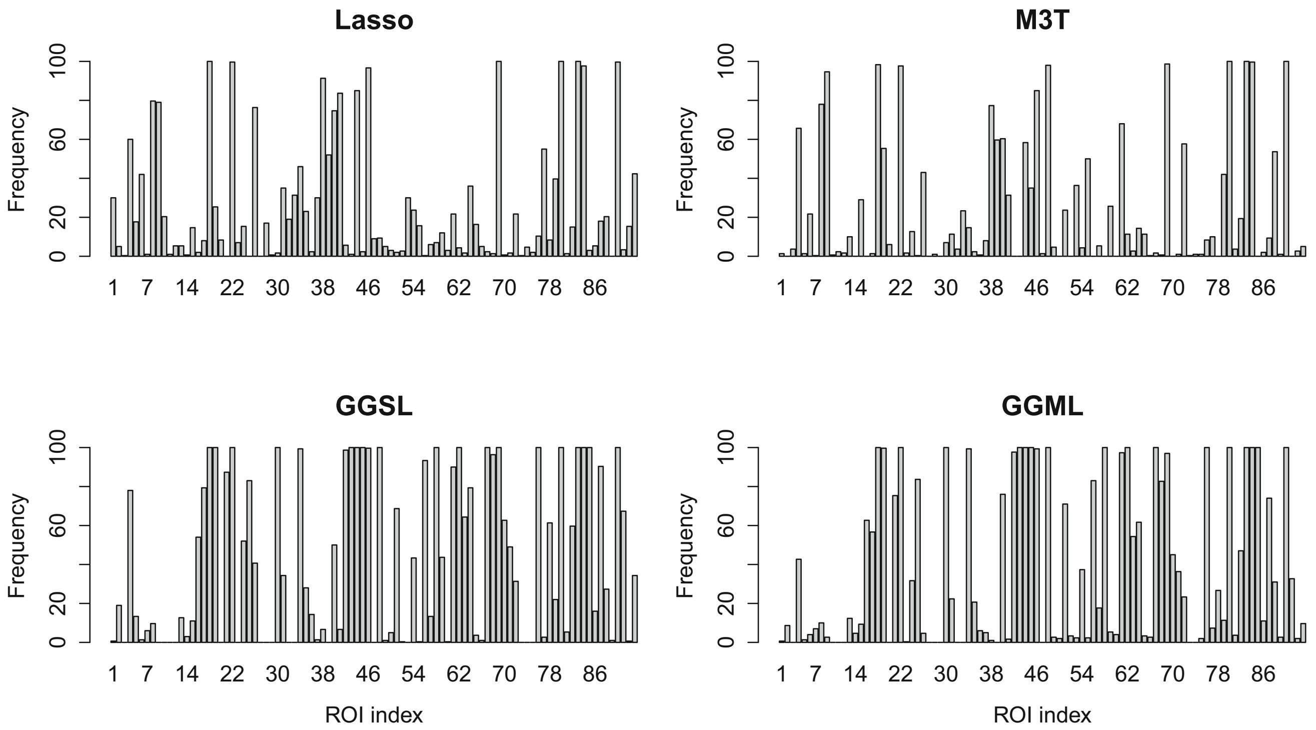


Fig. 6 Selection frequency of 93 ROIs for the AD/NC classification task using only MRI features

Table 6 Comparison of the top ten selected ROIs for the classification task

	MRI	PET
AD/NC		
Lasso	18, 22, 38, 44, 46, 69, 80, 83, 84, 90	12, 18, 23, 26, 41, 68, 69, 73, 81, 87
GGSL	18, 22, 30, 44, 58, 69, 80, 83, 84, 90	12, 18, 26, 35, 41, 68, 69, 73, 79, 87
M3T	9, 18, 22, 46, 48, 69, 80, 83, 84, 90	12, 23, 26, 35, 62, 68, 69, 73, 81, 87
GGML	18, 22, 30, 44, 48, 67, 80, 83, 84, 90	7, 12, 23, 26, 35, 62, 68, 69, 73, 87
MCI/NC		
Lasso	17, 28, 40, 48, 63, 64, 69, 83, 86, 92	2, 37, 39, 41, 54, 55, 63, 68, 81, 87
GGSL	17, 22, 30, 40, 46, 64, 69, 76, 83, 92	11, 12, 23, 26, 28, 29, 38, 40, 41, 87
M3T	17, 40, 46, 48, 53, 63, 64, 69, 83, 86	12, 35, 41, 62, 64, 68, 73, 79, 81, 87
GGML	22, 40, 45, 46, 61, 64, 69, 76, 83, 86	11, 12, 26, 29, 38, 40, 41, 47, 79, 87

Table 7 Comparison of the top ten selected ROIs for the regression task (MMSE)

	MRI	PET
AD/NC		
Lasso	9, 15, 18, 19, 22, 40, 80, 83, 84, 90	12, 18, 23, 26, 62, 63, 68, 69, 73, 79
GGSL	19, 22, 48, 58, 62, 67, 80, 83, 84, 85	7, 12, 23, 26, 35, 41, 62, 68, 69, 73
M3T	9, 18, 22, 46, 48, 69, 80, 83, 84, 90	12, 23, 26, 35, 62, 68, 69, 73, 81, 87
GGML	18, 22, 30, 44, 48, 67, 80, 83, 84, 90	7, 12, 23, 26, 35, 62, 68, 69, 73, 87
MCI/NC		
Lasso	17, 33, 40, 44, 48, 53, 62, 64, 69, 86	4, 23, 24, 33, 41, 61, 62, 68, 84, 87
GGSL	22, 45, 46, 48, 61, 64, 69, 76, 83, 86	11, 12, 23, 26, 28, 29, 38, 40, 41, 87
M3T	17, 40, 46, 48, 53, 63, 64, 69, 83, 86	12, 35, 41, 62, 64, 68, 73, 79, 81, 87
GGML	22, 40, 45, 46, 61, 64, 69, 76, 83, 86	11, 12, 26, 29, 38, 40, 41, 47, 79, 87

Table 8 Comparison of the top ten selected ROIs for the regression task (ADAS-Cog)

	MRI	PET
AD/NC		
Lasso	9, 18, 46, 48, 61, 62, 80, 83, 84, 90	12, 23, 26, 30, 35, 62, 73, 76, 81, 92
GGSL	18, 30, 48, 58, 62, 67, 80, 83, 84, 85	7, 12, 23, 26, 30, 35, 62, 69, 73, 92
M3T	9, 18, 22, 46, 48, 69, 80, 83, 84, 90	12, 23, 26, 35, 62, 68, 69, 73, 81, 87
GGML	18, 22, 30, 44, 48, 67, 80, 83, 84, 90	7, 12, 23, 26, 35, 62, 68, 69, 73, 87
MCI/NC		
Lasso	10, 17, 18, 38, 45, 46, 69, 72, 83, 87	10, 12, 14, 19, 35, 39, 41, 62, 64, 88
GGSL	17, 45, 46, 61, 62, 69, 72, 76, 83, 87	11, 12, 28, 29, 35, 38, 41, 71, 79, 87
M3T	17, 40, 46, 48, 53, 63, 64, 69, 83, 86	12, 35, 41, 62, 64, 68, 73, 79, 81, 87
GGML	22, 40, 45, 46, 61, 64, 69, 76, 83, 86	11, 12, 26, 29, 38, 40, 41, 47, 79, 87

Table 9 Names of the selected ROIs in this study

ROI index	ROI name	ROI index	ROI Name
2	Middle frontal gyrus right	47	Middle occipital gyrus right
4	Insula right	48	Middle temporal gyrus left
7	Cingulate region right	53	Postcentral gyrus left
9	Medial frontal gyrus left	54	Inferior frontal gyrus right
10	Superior frontal gyrus right	55	Precentral gyrus left
11	Globus pallidus right	58	Perirhinal cortex right
12	Globus pallidus left	61	Perirhinal cortex left
14	Inferior frontal gyrus left	62	Inferior temporal gyrus left
15	Putamen right	63	Temporal pole left
17	Parahippocampal gyrus left	64	Entorhinal cortex left
18	Angular gyrus right	67	Lateral occipitotemporal gyrus right
19	Temporal pole right	68	Entorhinal cortex right
22	Uncus right	69	Hippocampal formation left
23	Cingulate region left	71	Parietal lobe WM right
24	Fornix left	72	Insula left
26	Precuneus right	73	Postcentral gyrus right
28	Cerebral peduncle left	76	Amygdala left
29	Cerebral peduncle right	79	Anterior limb of internal capsule right
30	Hippocampal formation right	80	Middle temporal gyrus right
33	Caudate nucleus left	81	Occipital pole right
35	Anterior limb of internal capsule left	83	Amygdala right
37	Middle frontal gyrus left	84	Inferior temporal gyrus right
38	Superior parietal lobule left	85	Superior temporal gyrus right
39	Caudate nucleus right	86	Middle occipital gyrus left
40	Cuneus left	87	Angular gyrus left
41	Precuneus left	88	Medial occipitotemporal gyrus right
44	Supramarginal gyrus right	90	Lateral occipitotemporal gyrus left
45	Superior temporal gyrus left	92	Occipital pole left
46	Uncus left		

features are not very similar to the top ten selected ROIs for the cases using PET features. One possible reason is that MRI features and PET features provide complementary information for the diagnosis of AD. However, for each case, the top ten selected ROIs of the four methods are similar. For example, for the AD/NC classification task using MRI features, Table 6 indicates that the ROIs with indices 18, 80, 83, 84, and 90 are frequently selected by all four methods. It is interesting to point out that both GGML and M3T methods also select the 48th ROI frequently for the AD/NC classification task while this ROI is not one of the top ten selected ROIs of Lasso and GGSL for this task. However, as shown in Table 8, the 48th ROI is frequently selected by Lasso and GGSL for the regression task (ADAS-Cog) using AD/NC data. This indicates that the multi-task learning methods such as GGML and M3T incorporate the clinical score information for the classification task. On the other hand, as shown in Table 8, both GGML and M3T methods select the 22th ROI frequently for the regression task (ADAS-Cog) using AD/NC data while this ROI is not one of the top ten selected ROIs of Lasso and GGSL for this task. However, as shown in Table 6, the 22th ROI is frequently selected by Lasso and GGSL for the classification task (AD vs NC). This indicates that the multi-task learning methods such as GGML and M3T incorporate the class label information for the regression task.

Furthermore, as shown in Tables 6, 7 and 8, for the study using AD/NC data and MRI features, the common top ten selected ROIs of Lasso for different tasks are the ROIs with indices 18, 80, 83, 84 and 90. The common top ten selected ROIs of the GGSL method for different tasks are the ROIs with indices 58, 80, 83, and 84. Most of these ROIs are the top ten selected ROIs of our proposed GGML method. In Figs. 7 and 8, we visualize the top ten selected ROIs of our proposed GGML method when different datasets (AD/NC or MCI/NC) and different modalities (MRI or PET) are used. Most of the selected regions, e.g., uncus right (22), hippocampal formation right (30), uncus left (46), middle temporal gyrus left (48), hippocampus formation left (69), middle temporal gyrus right (80) and amygdale right (83), are known to be highly correlated with AD and MCI by many studies using group comparison methods (Jack et al. 1999; Misra et al. 2009; Zhang and Shen 2012).

Discussion

In this section, we first discuss some issues about constructing the undirected feature graph **G**. Then, some possible extensions of our proposed method will be discussed.

Construction of the undirected feature graph G

Before performing our proposed GGML method, we need to construct an undirected feature graph **G** representing the significant correlation information among features. In "Extract the correlation information among features" section, we proposed to use the graphical Lasso method to construct this graph. For some datasets, the constructed graph **G** may include many edges corresponding to weak or even wrong partial correlation due to bad estimation of the precision matrix. In this case, by thresholding of the estimated precision matrix, we can construct a sparse undirected graph for representing only the most reliable partial correlation.

Furthermore, besides partial correlation information among features, we can also combine other useful information (e.g., some prior information about features) to construct this graph **G**. Our proposed GGML method can be used for any given undirected feature graph **G** representing the relationships among different features.

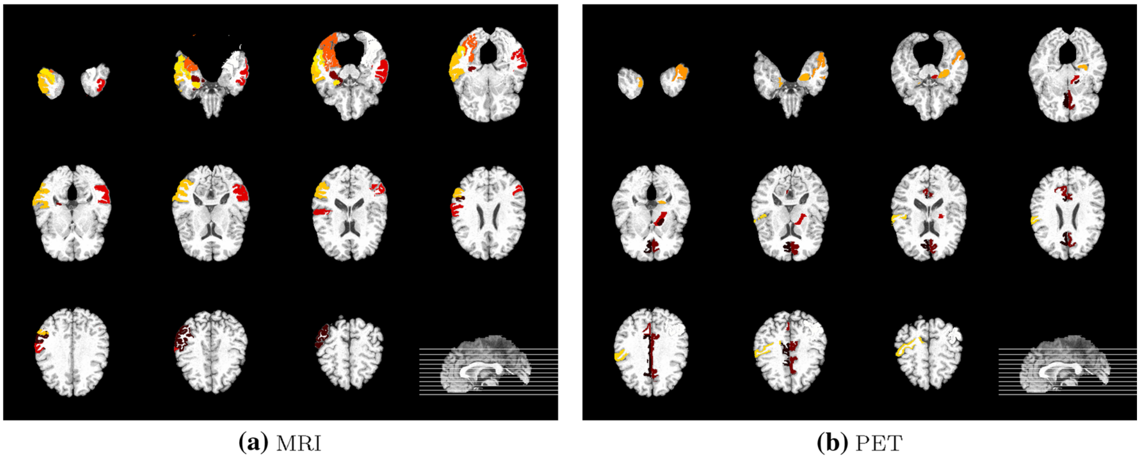
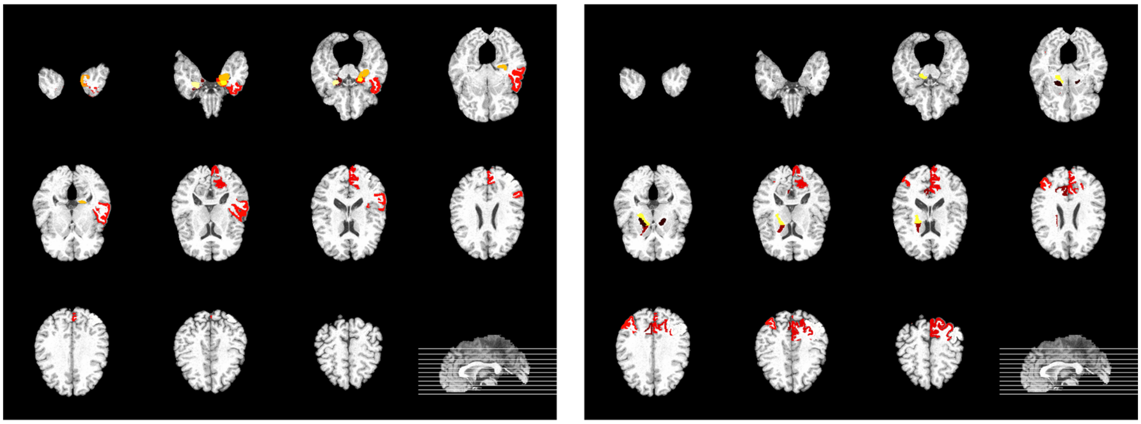

(a) MRI **(b)** PET

Fig. 7 Top ten most discriminative brain regions selected by GGML method using AD/NC dataset



(a) MRI

(b) PET

Fig. 8 Top ten most discriminative brain regions selected by GGML method using MCI/NC dataset

Use of the structure information among different subjects

Our proposed GGML method utilizes both the correlation information among features and the intrinsic correlation information among different response variables. Actually, we can also generalize GGML method to incorporate the structure information among different subjects. Similar to the locality preserving projection (LPP) method (He and Niyogi 2004), we can model the structure information among different training subjects as another sparse undirected graph **S**. Here, **S** has n nodes and each node represents one subject. The connectivity of the graph **S** can be defined by the k nearest neighbors, i.e., subjects x_s and x_l are connected by an edge if x_s is among the k nearest neighbors of x_l, or x_l is among the k nearest neighbors of x_s. In order to use the structure information among different training subjects represented by **S**, we can preserve the neighborhood structure of subjects, i.e., encouraging the predicted response variables $\hat{y}_s = \mathbf{B}^T x_s$ and $\hat{y}_l = \mathbf{B}^T x_l$ to be close if the sth and the lth subjects are connected in the undirected graph **S**.

Conclusion

In summary, we propose a new graph-guided multi-task learning method to incorporate the correlation information among features and the intrinsic correlation information among different tasks. To use the correlation information among features, our proposed GGML method encourages the partially correlated features to be in or out of the model jointly. Furthermore, in order to acquire more robust and accurate feature selection, our proposed GGML method encourages different tasks to share a common useful feature subset. Theoretically, our proposed GGML method is very general and includes the M3T method as a special case. The experimental results on the simulated examples and the ADNI dataset also show the advantage of the proposed GGML method over the existing methods.

Acknowledgments This work was supported in part by NIH Grants AG041721, EB006733, EB008374, EB009634, NSF DMS-1407241 and NIH/NCI Grant R01 CA-149569. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) National Institutes of Health Grant U01 AG024904. 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, Amorfix, Bayer Schering Pharma AG, Bioclinica Inc., Biogen Idec, Bristol-Myers Squibb, Eisai Global Clinical Development, Elan Corporation, Genentech, GE Healthcare, Innogenetics, IXICO, Janssen Alzheimer Immunotherapy, Johnson and Johnson, Eli Lilly and Co., Medpace, Inc., Merck and Co., Inc., Meso Scale Diagnostic, & LLC, Novartis AG, Pfizer Inc, F. Hoffman-La Roche, Servier, Synarc, Inc., and Takeda Pharmaceuticals, as well as non-profit partners the Alzheimer's Association and Alzheimer's Drug Discovery Foundation, with participation from the US Food and Drug Administration. Private sector contributions to ADNI are facilitated by the Foundation for the National Institutes of Health (http://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 Neuro Imaging at the University of California, Los Angeles.

References

Bain LJ, Jedrziewski K, Morrison-Bogorad M, Albert M, Cotman C, Hendrie H, Trojanowski JQ (2008) Healthy brain aging: a meeting report from the Sylvan M. Cohen annual retreat of the University of Pennsylvania Institute on Aging. Alzheimer's Dement J Alzheimer's Assoc 4:443–446

Bouwman FH, van der Flier WM, Schoonenboom NS, van Elk EJ, Kok A, Rijmen F, Blankenstein MA, Scheltens P (2007) Longitudinal changes of csf biomarkers in memory clinic patients. Neurology 69(10):1006–1011

Brookmeyer R, Johnson E, Ziegler-Graham K, Arrighi HM (2007) Forecasting the global burden of Alzheimer's disease. Alzheimer's Dement 3:186–191

Cai T, Liu W, Luo X (2011) A constrained l_1 minimization approach to sparse precision matrix estimation. J Am Stat Assoc 106(494):594–607

Cheng B, Zhang D, Chen S, Kaufer DI, Shen D (2013) Semi-supervised multimodal relevance vector regression improves cognitive performance estimation from imaging and biological biomarkers. Neuroinformatics 11(3):339–353

De Santi S, de Leon MJ, Rusinek H, Convit A, Tarshish CY, Roche A, Tsui WH, Kandil E, Boppana M, Daisley K et al (2001) Hippocampal formation glucose metabolism and volume losses in MCI and AD. Neurobiol Aging 22(4):529–539

Du AT, Schuff N, Kramer JH, Rosen HJ, Gorno-Tempini ML, Rankin K, Miller BL, Weiner MW (2007) Different regional patterns of cortical thinning in Alzheimer's disease and frontotemporal dementia. Brain 130:1159–1166

Duchesne S, Caroli A, Geroldi C, Frisoni GB, Collins DL (2005) Predicting clinical variable from MRI features: application to MMSE in MCI. In: Medical image computing and computer-assisted intervention—MICCAI 2005, Springer, pp 392–399

Fan Y, Kaufer D, Shen D (2010) Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. In: 2010 IEEE international symposium on biomedical imaging: from nano to macro, IEEE, pp 852–855

Fjell AM, Walhovd KB, Fennema-Notestine C, McEvoy LK, Hagler DJ, Holland D, Brewer JB, Dale AM (2010) CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. J Neurosci 30:2088–2101

Folstein MF, Folstein SE, McHugh PR (1975) Mini-mental state: a practical method for grading the cognitive state of patients for the clinician. J Psychiatr Res 12(3):189–198

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9:432–441

Hampson M, Peterson BS, Skudlarski P, Gatenby JC, Gore JC (2002) Detection of functional connectivity using temporal correlations in MR images. Hum Brain Mapp 15(4):247–262

He X, Niyogi P (2004) Locality preserving projections. In: Thrun S, Saul LK (eds) Neural information processing systems, vol 16. MIT Press, Cambridge, p 153

Hebert LE, Beckett LA, Scherr PA, Evans DA (2001) Annual incidence of Alzheimer disease in the United States projected to the years 2000 through 2050. Alzheimer Dis Assoc Disord 15:169–173

Jack C, Petersen RC, Xu YC, O'Brien PC, Smith GE, Ivnik RJ, Boeve BF, Waring SC, Tangalos EG, Kokmen E (1999) Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. Neurology 52(7):1397–1397

Kabani N, MacDonald D, Holmes C, Evans A (1998) A 3d atlas of the human brain. NeuroImage 7:S717

Lee H, Lee DS, Kang H, Kim BN, Chung MK (2011) Sparse brain network recovery under compressed sensing. IEEE Trans Med Imaging 30(5):1154–1165

McEvoy LK, Fennema-Notestine C, Roddey JC Jr, DJH, Holland D, Karow DS, Pung CJ, Brewer JB, Dale AM (2009) Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. Radiology 251:195–205

Meinshausen N, Bühlmann P (2006) High-dimensional graphs and variable selection with the lasso. Ann Stat 34(3):1436–1462

Misra C, Fan Y, Davatzikos C (2009) 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 44(4):1415–1422

Morris JC, Storandt M, Miller JP, McKeel DW, Price JL, Rubin EH, Berg L (2001) Mild cognitive impairment represents early-stage Alzheimer disease. Arch Neurol 58(3):397

Rosen WG, Mohs RC, Davis KL (1984) A new rating scale for alzheimer's disease. Am J Psychiatry 141(11):1356–1364

Ryali S, Chen T, Supekar K, Menon V (2012) Estimation of functional connectivity in FMRI data using stability selection-based sparse partial correlation with elastic net penalty. Neuroimage 59(4):3852–3861

Shen D, Davatzikos C (2002) Hammer: hierarchical attribute matching mechanism for elastic registration. IEEE Trans Med Imaging 21(11):1421–1439

Sled JG, Zijdenbos AP, Evans AC (1998) A nonparametric method for automatic correction of intensity nonuniformity in MRI data. IEEE Trans Med Imaging 17(1):87–97

Stonnington CM, Chu C, Klöppel S, Jack CR Jr, Ashburner J, Frackowiak RS (2010) Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. Neuroimage 51(4):1405–1413

Tibshirani R (1996) Regression shrinkage and selection via the lasso. J R Stat Soc Ser B 58(1):267–288

Wang Y, Fan Y, Bhatt P, Davatzikos C (2010) High-dimensional pattern regression using machine learning: from medical images to continuous clinical variables. NeuroImage 50(4):1519–1535

Wang Y, Nie J, Yap PT, Shi F, Guo L, Shen D (2011) Robust deformable-surface-based skull-stripping for large-scale studies. Med Image Comput Comput Assist Interv 6893:635–642

Yang S, Yuan L, Lai YC, Shen X, Wonka P, Ye J (2012) Feature grouping and selection over an undirected graph. In: Proceedings of the 18th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 922–930

Yang Y, Zou H (2013) gglasso: group lasso penalized learning using a unified BMD algorithm. http://CRAN.R-project.org/package=gglasso, r package version 1.1

Yu G, Liu Y, Thung KH, Shen D (2014) Multi-task linear programming discriminant analysis for the identification of progressive MCI individuals. PloS One 9(5):e96458

Yuan M, Lin Y (2006) Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B 68:49–67

Zhang D, Shen D (2012) Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in Alzheimer's disease. NeuroImage 59(2):895–907

Zhang Y, Brady M, Smith S (2001) Segmentation of brain MR images through a hidden markov random field model and the expectation-maximization algorithm. IEEE Trans Med Imaging 20(1):45–57

Zou H (2006) The adaptive lasso and its oracle properties. J Am Stat Assoc 101(476):1418–1429