Published in final edited form as:
Med Image Comput Comput Assist Interv. 2011 ; 14(Pt 3): 115–123.

Identifying AD-Sensitive and Cognition-Relevant Imaging Biomarkers via Joint Classification and Regression

Hua Wang[1], **Feiping Nie**[1], **Heng Huang**[1], **Shannon Risacher**[2], **Andrew J Saykin**[2], **Li Shen**[2], and **ADNI**[*]

Hua Wang: huawangcs@gmail.com; Feiping Nie: feipingnie@gmail.com; Heng Huang: heng@uta.edu; Shannon Risacher: srisache@iupui.edu; Andrew J Saykin: asaykin@iupui.edu; Li Shen: shenli@iupui.edu

[1]Computer Science and Engineering, University of Texas at Arlington, TX

[2]Radiology and Imaging Sciences, Indiana University School of Medicine, IN

Abstract

Traditional neuroimaging studies in Alzheimer's disease (AD) typically employ independent and pairwise analyses between multimodal data, which treat imaging biomarkers, cognitive measures, and disease status as isolated units. To enhance mechanistic understanding of AD, in this paper, we conduct a new study for identifying imaging biomarkers that are associated with both cognitive measures and AD. To achieve this goal, we propose a new sparse joint classification and regression method. The imaging biomarkers identified by our method are AD-sensitive and cognition-relevant and can help reveal complex relationships among brain structure, cognition and disease status. Using the imaging and cognition data from Alzheimer's Disease Neuroimaging Initiative database, the effectiveness of the proposed method is demonstrated by clearly improved performance on predicting both cognitive scores and disease status.

1 Introduction

Neuroimaging is a powerful tool for characterizing neurodegenerative process in the progression of Alzheimer's disease (AD). Pattern classification methods have been widely employed to predict disease status using neuroimaging measures [2, 3]. Since AD is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions, regression models have been investigated to predict clinical scores from individual magnetic resonance imaging (MRI) and/or positron emission tomography (PET) scans [8, 9]. For example, in [9], stepwise regression was performed in a pairwise fashion to relate each of MRI and FDG-PET measures of eight candidate regions to each of four Rey's Auditory Verbal Learning Test (RAVLT) memory scores.

Predicting disease status and predicting memory performance, using neuroimaging data, are both important learning tasks. Prior research typically studied these tasks *separately*. One example is to first determine disease-relevant cognitive scores and then identify imaging biomarkers associated with these scores so that interesting pathways from brain structure to cognition to symptom can potentially be discovered. However, a specific cognitive function could be related to multiple imaging measures associated with different biological pathways (some of them are not related to AD). As a result, the identified imaging biomarkers are not necessarily all disease specific. To have a better understanding of the underlying mechanism

[*]Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (U01 AG024904, adni.loni.ucla.edu). HW and HH were supported by NSF-CNS 0923494, NSF-IIS 1041637, NSF-CNS 1035913. SR, AS and LS were supported in part by NIBIB R03 EB008674, NIA 1RC 2AG036535, CTSI-IUSM/CTR (RR025761), NIA P30 AG10133, and NIA R01 AG19771.

specific to AD, an interesting topic would be to only discover imaging biomarkers associated with both cognitive function and AD status.

To identify AD-sensitive and cognition-relevant imaging biomarkers, we propose a new joint classification and regression learning model to simultaneously performing two heterogeneous tasks, *i.e.*, imaging-to-disease classification and imaging-to-cognition regression. We use magnetic resonance imaging (MRI) measures as predictors and cognitive memory scores and disease status as response variables. For each individual regression or classification task, we employ a multitask learning model [1] in which tasks for predicting different memory performances (or those for predicting AD and control dummy variables in classification) are considered as homogeneous tasks. Different to LASSO and other related methods that mainly find the imaging features correlated to each individual memory score, our method selects the imaging features that tend to play an important role on influencing multiple homogenous tasks.

Our new method utilizes the sparse regularization to perform imaging biomarker selection and learn a sparse parameter matrix under a unified framework that integrates both heterogeneous and homogenous tasks. Specifically, by recognizing that the formation [6] and maintenance [4] of memory are synergically accomplished by a few brain areas, such as medial temporal lobe structures, medial and lateral parietal, as well as prefrontal cortical areas, we use the $\ell_{2,1}$-norm regularization to select features that can predict most memory scores and classify AD versus control. Empirical comparison with the existing methods demonstrates that the proposed method not only yields improved performance on predicting both cognitive scores and disease status, but also discovers a small set of AD-sensitive and cognition-relevant biomarkers in accordance with prior findings.

2 Sparse Model for Joint Classification and Regression

When we study either regression or classification via a multi-task learning model, given a set of input variables, (*i.e.*, features, such as imaging biomarkers), we are interested in learning a set of related models (*e.g.*, associations between image biomarkers and cognitive scores) for predicting multiple homogenous tasks (such as predicting cognitive scores). Since these homogenous tasks are typically interrelated, they share a common input space. As a result, it is desirable to learn all the models jointly rather than treating each task as an independent one. Such multi-task learning methods can help discover robust patterns, especially when significant patterns in a single task become outliers for other tasks, and potentially increase the predictive power.

To identify AD-sensitive and cognition-relevant biomarkers from imaging data, we formulate a new problem to jointly learn two heterogeneous tasks: classification and regression. We propose a new sparse model for joint classification and regression to perform multivariate regression for cognitive memory scores predictions and logistic regression for disease classification tasks simultaneously.

Notation

We write matrices and vectors as bold uppercase and lowercase letters respectively. Given a matrix $\mathbf{M} = [m_{ij}]$, we denote its i-th row as $\mathbf{m}^i$ and j-th column as $\mathbf{m}_j$. The Frobenius norm of the matrix $\mathbf{M}$ is denoted as $||\mathbf{M}||_F$, and the $\ell_{2,1}$-norm [5] of $\mathbf{M}$ is defined as

$$\left\|\mathbf{M}\right\|_{2,1} = \sum_i \sqrt{\sum_j m_{ij}^2} = \sum_i \left\|\mathbf{m}^i\right\|_2.$$

2.1 Objective of Sparse Joint Classification and Regression

First, logistic regression is used for disease classification. Given the training data $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, each data point $\mathbf{x}_i$ is associated with a label vector $\mathbf{y}^i = [y_{i1}, \ldots, y_{ic_1}] \in \mathbb{R}^{c_1}$. If $\mathbf{x}_i$ belongs to the k-th class, $y_{ik} = 1$, otherwise $y_{ik} = 0$. We write $\mathbf{Y} = [(\mathbf{y}^1)^T, \ldots, (\mathbf{y}^n)^T]^T \in \mathbb{R}^{n \times c_1}$. In traditional multi-class logistic regression, under a projection matrix $\mathbf{W} \in \mathbb{R}^{d \times c_1}$, we have

$$p(k|\mathbf{x}_i, \mathbf{W}) = \frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i}} \Rightarrow p(\mathbf{y}^i|\mathbf{x}_i, \mathbf{W}) = \prod_{k=1}^{c_1} \left(\frac{e^{\mathbf{w}_k^T \mathbf{x}_i}}{\sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i}} \right)^{y_{ik}},$$

where $p(k \mid \mathbf{x}_i, \mathbf{W})$ is the probability that $\mathbf{x}_i$ belongs to the k-th class, and $p(\mathbf{y}^i \mid \mathbf{x}_i, \mathbf{W})$ is the probability that $\mathbf{x}_i$ is associated with the given label $\mathbf{y}^i$. Therefore, the multi-class logistic loss that maximizes the Log-likelihood can be achieved by minimizing:

$$l_1(\mathbf{W}) = -\log \prod_{i=1}^{n} p(\mathbf{y}^i|\mathbf{x}_i, \mathbf{W}) = \sum_{i=1}^{n} \sum_{k=1}^{c_1} \left(y_{ik} \log \sum_{l=1}^{c_1} e^{\mathbf{w}_l^T \mathbf{x}_i} - y_{ik} \mathbf{w}_k^T \mathbf{x}_i \right). \quad (1)$$

In AD classification, we have two classes, *i.e.*, AD and health control (HC).

Second, we use multivariate least square regression to predict cognitive scores, which minimizes:

$$l_2(\mathbf{P}) = \left\| \mathbf{X}^T \mathbf{P} - \mathbf{Z} \right\|_{\mathrm{F}}^2, \quad (2)$$

where $\mathbf{X}$ is the data matrix, $\mathbf{Z} = [(\mathbf{z}^1)^T, \ldots, (\mathbf{z}^n)^T]^T \in \mathbb{R}^{n \times c_2}$ is the label matrix for the c_2 regression tasks, and $\mathbf{P} \in \mathbb{R}^{d \times c_2}$ is the projection matrix.

The objective for joint classification and regression to identify AD-sensitive and cognition-relevant imaging biomarkers can now be formulated as follows:

$$\min J(\mathbf{V}) = l_1(\mathbf{W}) + l_2(\mathbf{P}) + \gamma \left\| \mathbf{V} \right\|_{2,1}, \quad (3)$$

where $\mathbf{V} = [\mathbf{W}\ \mathbf{P}] \in \mathbb{R}^{d \times (c_1 + c_2)}$. Thanks to the $\ell_{2,1}$-norm regularization on $\mathbf{V}$ [1], the biomarkers are identified across all tasks so that they are not only correlated to cognitive scores but also discriminative to disease status.

2.2 An Efficient Iterative Algorithm

Due to the non-smoothness of the $\ell_{2,1}$-norm term, J in Eq. (3) is hard to solve in general. Thus we derive an efficient iterative algorithm as follows.

Taking the derivatives of J w.r.t. $\mathbf{W}$ and $\mathbf{P}$, we set them to be zeros:

$$\frac{\partial J}{\partial \mathbf{W}} = \frac{\partial l_1(\mathbf{W})}{\partial \mathbf{W}} + 2\gamma \mathbf{D}\mathbf{W} = 0, \quad \frac{\partial J}{\partial \mathbf{P}} = 2\mathbf{X}\mathbf{X}^T \mathbf{P} - 2\mathbf{X}\mathbf{Z} + 2\gamma \mathbf{D}\mathbf{P} = 0, \quad (4)$$

where $\mathbf{D}$ is a diagonal matrix whose k-th diagonal element is $\frac{1}{2\left\| \mathbf{v}^k \right\|_2}$. Because $\mathbf{D}$ depends on $\mathbf{V}$, it is also an unknown variable. Following standard optimization procedures in statistical learning, we alternately optimize $\mathbf{V}$ and $\mathbf{D}$.

First, we randomly initialize $\mathbf{V} \in \mathbb{R}^{d\times(c_1+c_2)}$, upon which we calculate **D**. After obtaining **D**, we update the solution $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$ using Eq. (4). To be more precise, **P** is updated by $\mathbf{P} = (\mathbf{X}\mathbf{X}^T + \gamma\mathbf{D})^{-1}\mathbf{X}\mathbf{Z}$. Because we cannot update **W** with a closed form solution upon Eq. (4), we employ Newton's method to obtain updated **W** by solving the following problem: $\min_{\mathbf{W}} l_1(\mathbf{W}) + \gamma\,\mathbf{tr}(\mathbf{W}^T\mathbf{D}\mathbf{W})$.

Once we obtain the updated $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$, we can calculate **D**. This procedure repeats until convergence. The detailed algorithm is summarized in Algorithm 1, whose convergence is proved as following.

Lemma 1—*For any vector* $\mathbf{v}$ *and* $\mathbf{v}_0$, *we have* $\|\mathbf{v}\|_2 - \frac{\|\mathbf{v}\|_2^2}{2\|\mathbf{v}_0\|_2} \le \|\mathbf{v}_0\|_2 - \frac{\|\mathbf{v}_0\|_2^2}{2\|\mathbf{v}_0\|_2}$. *Proof is available in* [5].

Theorem 1—Algorithm 1 decreases the objective value of J in every iteration.

Proof: In each iteration, denote the updated **W** as $\tilde{\mathbf{W}}$, the updated **P** as $\tilde{\mathbf{P}}$, thus the updated **V** is $\tilde{\mathbf{V}} = [\tilde{\mathbf{W}}\ \tilde{\mathbf{P}}]$. According to step 3 of Algorithm 1, we have

$$l_1\left(\tilde{\mathbf{W}}\right) + \gamma\,\mathbf{tr}\left(\tilde{\mathbf{W}}^T\mathbf{D}\tilde{\mathbf{W}}\right) \le l_1(\mathbf{W}) + \gamma\,\mathbf{tr}(\mathbf{W}^T\mathbf{D}\mathbf{W}). \quad (5)$$

According to step 4 we know that

$$l_2\left(\tilde{\mathbf{P}}\right) + \gamma\,\mathbf{tr}\left(\tilde{\mathbf{P}}^T\mathbf{D}\tilde{\mathbf{P}}\right) \le l_2(\mathbf{P}) + \gamma\,\mathbf{tr}(\mathbf{P}^T\mathbf{D}\mathbf{P}). \quad (6)$$

According to the definition of **D** and Lemma 1, we have the following inequality:

$$\sum_{k=1}^{d}\|\tilde{\mathbf{v}}^k\|_2 - \sum_{k=1}^{d}\frac{\|\tilde{\mathbf{v}}^k\|_2^2}{2\|\mathbf{v}^k\|_2} \le \sum_{k=1}^{d}\|\mathbf{v}^k\|_2 - \sum_{k=1}^{d}\frac{\|\mathbf{v}^k\|_2^2}{2\|\mathbf{v}^k\|_2} \Rightarrow \gamma\sum_{k=1}^{d}\|\tilde{\mathbf{v}}^k\|_2 - \gamma\,\mathbf{tr}\left(\tilde{\mathbf{V}}^T\mathbf{D}\tilde{\mathbf{V}}\right) \le \gamma\sum_{k=1}^{d}\|\mathbf{v}^k\|_2 - \gamma\,\mathbf{tr}(\mathbf{V}^T\mathbf{D}\mathbf{V}). \quad (7)$$

Because $\mathbf{tr}(\mathbf{V}^T\mathbf{D}\mathbf{V}) = \mathbf{tr}(\mathbf{W}^T\mathbf{D}\mathbf{W}) + \mathbf{tr}(\mathbf{P}^T\mathbf{D}\mathbf{P})$, by adding Eqs. (5–7) at the both sides, we arrive at

$$l_1\left(\tilde{\mathbf{W}}\right) + l_2\left(\tilde{\mathbf{P}}\right) + \gamma\sum_{k=1}^{d}\|\tilde{\mathbf{v}}^k\|_2 \le l_1(\mathbf{W}) + l_2(\mathbf{P}) + \gamma\sum_{k=1}^{d}\|\mathbf{v}^k\|_2 \quad (8)$$

Thus, Algorithm 1 decreases the value of J in Eq. (3) in every iteration.

Because J in Eq. (3) is obviously lower-bounded by 0, Theorem 1 guarantees the convergence of Algorithm 1. In addition, because J is convex, Algorithm 1 converges at the global optimum of the problem.

3 Experimental Results

We evaluate our method by applying it to the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort. The goal is to select a compact set of AD-sensitive and cognition-relevant imaging biomarkers while maintaining high predictive power.

Data preparation

We downloaded data from the ADNI database (http://adni.loni.ucla.edu). We used baseline MRI data, from which we extracted 56 volumetric and cortical thickness values (Fig. 1) using FreeSurfer (http://surfer.nmr.mgh.harvard.edu), as described in [7]. We included memory scores from three different cognitive assessments including Mini-Mental State Exam (MMSE), Rey's Auditory Verbal Learning Test (RAVLT), and TRAILS. Details about these assessments are available in the ADNI procedure manuals (http://www.adni-info.org/Scientists/ProceduresManuals.aspx).

3.1 Biomarker Identification

The proposed method aims to identify imaging biomarkers that are associated with both disease status and cognitive scores in a joint classification and regression framework. Here we first examine the identified biomarkers. Fig. 1 shows a summarization of selected features for the three experiments (one for each type of cognitive scores) where the regression/classification weights are color-mapped for each feature and each task. Fig. 2 visualizes the cortical maps of selected features for both classification and regression in different tasks.

Fig. 1 and Fig. 2 show that a small set of MRI measures are identified, including hippocampal volume (HippVol), entorhinal cortex thickness (EntCtx), amygdala volume (AmygVol), inferior parietal gyrus thickness (InfParietal), and middle temporal gyrus thickness (MidTemporal). These are all well-known AD-relevant biomarkers. Our method also shows that these markers are jointly associated with one or more memory scores. Although we know that MRI measures, cognitive scores and diagnosis are highly correlated, the complex relationships among them remain to be discovered for a better understanding of AD mechanism. This is one major focus of our work. As shown in Fig. 1, different AD-sensitive MRI measures could be related to different cognitive tasks. The proposed sparse method for joint classification and regression enables us to sort out MRI-cognition relationships while focusing on AD-sensitive markers.

3.2 Improved Prediction Performance

Now we evaluate the performance of joint classification and regression for AD detection and cognitive score prediction using MRI data. We performed standard 5-fold cross-validation, where the parameter γ of our method in Eq. (3) was fine tuned in the range of $\{10^{-5},\ldots, 1, \ldots, 10^{5}\}$ by an internal 5-fold cross-validation in the training data of each of the 5 trials. For classification, we compared the proposed method against two baseline methods including logistic regression and support vector machine (SVM). For SVM, we implemented three different kernels including linear, polynomial and Gaussian kernels. For polynomial kernel, we searched the best results when the polynomial order varied in the range of {1, 2,…, 10}; for Gaussian kernel, we fine tuned the parameter α in the same range as that for our method and fixed parameter C as 1. For regression, we compared our method against two widely used methods including multivariate regression and ridge regression. For the latter, we fine tuned its parameter in the same range as that for our method. The results are reported in Table 1.

Table 1 shows that our method performs clearly better than both logistic regression and SVM, which are consistent with our motivations in that our method classifies participants using the information from not only MRI measures but also the reinforcement by cognitive score regression. In addition, the cognitive score regression performances of our method measured by root mean squared error (RMSE) outperform both multivariate regression and ridge regression, supporting the usefulness of joint classification and regression from another perspective. Ridge regression achieves close but slightly worse regression

performance. However, it lacks the ability to identify relevant imaging markers. All these observations demonstrate the effectiveness of the proposed method in improving the performances of both AD detection and cognitive score prediction.

Mild cognitive impairment (MCI) is thought to be the prodromal stage of AD. Including MCI in this type of analyses will be an interesting future direction to help biomarker discovery for early detection of AD. We performed an initial analyis on three-class classification for AD, MCI and HC: the accuracy of our method was 0.663 and the best of other tested methods was 0.615. Apparently this is a much harder task and warrants further thorough investigation.

4 Conclusions

We have proposed a new sparse model for joint classification and regression and applied it to the ADNI cohort for identifying AD-sensitive and cognition-relevant imaging biomarkers. Our methodological contributions are threefold: 1) proposing a new learning model, joint classification and regression learning, to identify disease-sensitive and task-relevant biomarkers for analyzing multimodal data; 2) employing structural sparsity regularization to integrate heterogenous and homogenous tasks in a unified multi-task learning framework; 3) deriving a new efficient optimization algorithm to solve our non-smooth objective function, and coupling this with rigorous theoretical analysis on global optimum convergency. Empirical comparison with the existing methods demonstrates that our method not only yields improved performance on predicting both cognitive scores and disease status using MRI data, but also discovers a small set of AD-sensitive and cognition-relevant imaging biomarkers in accordance with prior findings.

References

1. Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. NIPS 2007. 2007:41–48.
2. Fan Y, Batmanghelich N, Clark CM, Davatzikos C. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. Neuroimage. 2008; 39(4):1731–43. [PubMed: 18053747]
3. Hinrichs C, Singh V, Mukherjee L, Xu G, Chung M, Johnson S. Spatially augmented LPboosting for AD classification with evaluations on the ADNI dataset. Neuroimage. 2009; 48(1):138–49. [PubMed: 19481161]
4. Moscovitch M, Nadel L, Winocur G, Gilboa A, Rosenbaum R. The cognitive neuroscience of remote episodic, semantic and spatial memory. Curr Opin Neurobiol. 2006; 16(2):179–190. [PubMed: 16564688]
5. Nie F, Huang H, Cai X, Ding C. Efficient and robust feature selection via joint l2,1-norms minimization. NIPS 2010. 2010:1813–1821.
6. Scoville W, Milner B. Loss of recent memory after bilateral hippocampal lesions. Journal of Neurology, Neurosurgery & Psychiatry. 1957; 20(1):11.
7. Shen L, Kim S, Risacher SL, Nho K, Swaminathan S, West JD, Foroud T, Pankratz N, Moore JH, Sloan CD, Huentelman MJ, Craig DW, Dechairo BM, Potkin SG, Jack CR, Weiner MW, Saykin AJ. ADNI: Whole genome association study of brain-wide imaging phenotypes for identifying quantitative trait loci in MCI and AD: A study of the ADNI cohort. Neuroimage. 2010; 53(3):1051–1063. [PubMed: 20100581]
8. Stonnington CM, Chu C, Kloppel S, Jack CRJ, Ashburner J, Frackowiak RS. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. Neuroimage. 2010; 51(4):1405–13. [PubMed: 20347044]
9. Walhovd K, Fjell A, Dale A, McEvoy L, Brewer J, Karow D, Salmon D, Fennema-Notestine C. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. Neurobiol Aging. 2010; 31(7):1107–1121. [PubMed: 18838195]

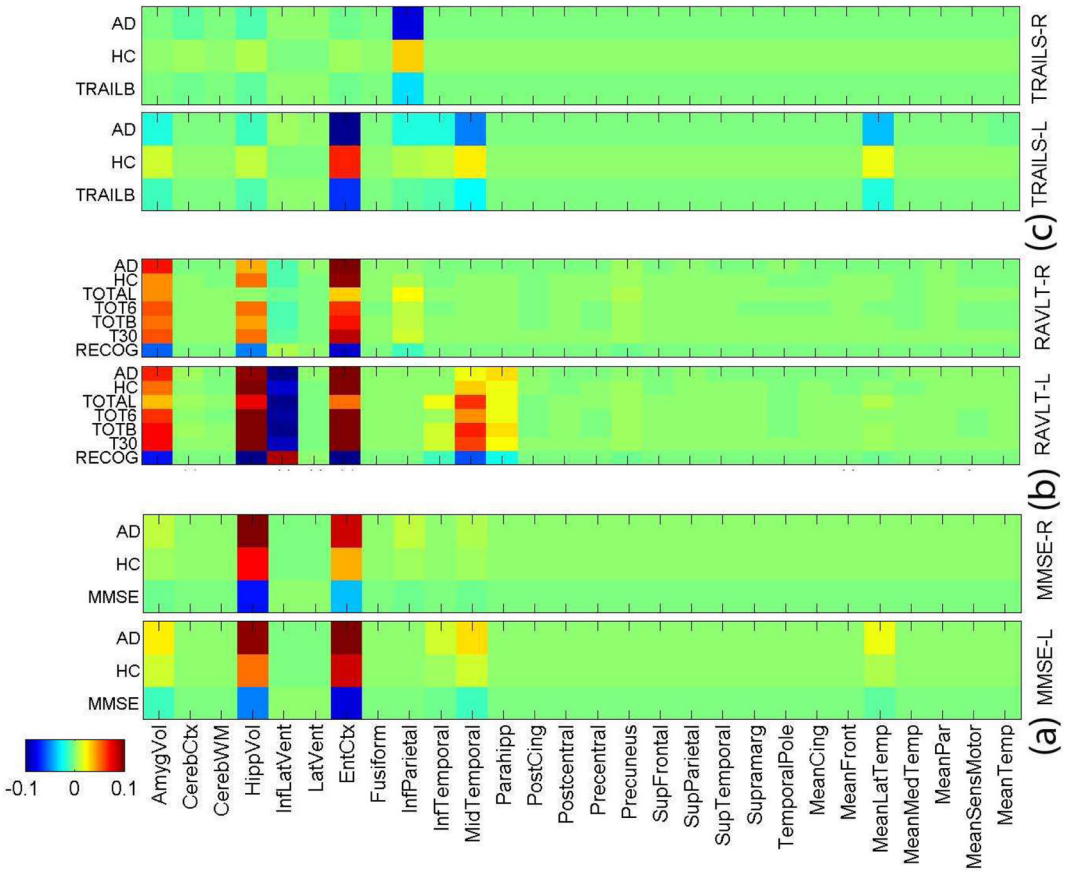

Fig. 1.

Weight maps of the joint classification and regression tasks. One binary classification task for AD and HC. Three different groups of cognitive scores for regression: (a) MMSE score, (b) RAVLT score, (c) TRAILS score. "−L" indicates the FreeSurfer biomarkers at the left side, and "−R" indicates those at the right side.

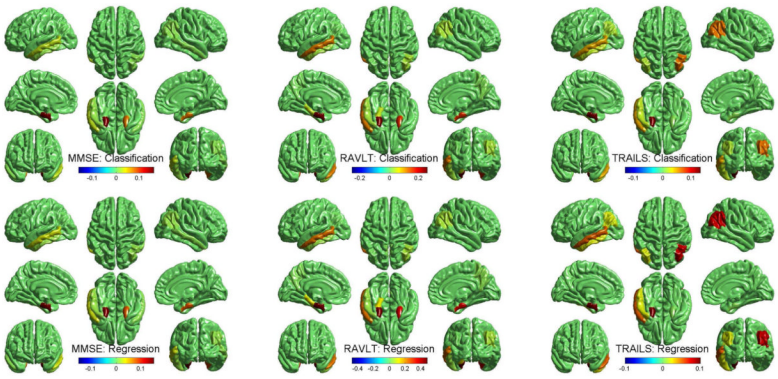

Fig. 2.
Cortical map of selected features for cognitive score prediction using FreeSurfer measures in the three joint classification and regression tasks.

Table 1

Comparison of classification and regression performance.

				Our method		Classification accuracy		RMSE (mean ± std)	
Memory score	**# subjects**	**# AD**	**# HC**	**Classification accuracy**	**Regression RMSE**	**Logistic regression**	**SVM**	**Multivariate regression**	**Ridge regression**
MMSE	378	175	203	0.881	0.034 ±0.002	0.832	0.783 (linear kernel)	0.041 ± 0.003	0.039 ± 0.004
RAVLT	371	172	199	0.884	0.019 ±0.001		0.839 (Polynomial kernel)	0.028 ± 0.002	0.024 ± 0.003
TRAILS	369	166	203	0.864	0.043 ±0.002		0.796 (Gausssian kernel)	0.049 ± 0.003	0.046 ± 0.003

Med Image Comput Comput Assist Interv. Author manuscript; available in PMC 2011 October 25.

Algorithm 1

An efficient algorithm to solve Eq. (3).

Input: $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, $\mathbf{Y} = [\mathbf{y}_1, \ldots, \mathbf{y}_n]^T \in \mathbb{R}^{n \times c_1}$, and $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]^T \in \mathbb{R}^{n \times c_2}$.

1. Initialize $\mathbf{W} \in \mathbb{R}^{d \times c_1}$, $\mathbf{P} \in \mathbb{R}^{d \times c_2}$, and let $\mathbf{V} = [\mathbf{W}\ \mathbf{P}] \in \mathbb{R}^{d \times (c_1 + c_2)}$;

while *not converge* **do**

2. Calculate the diagonal matrix $\mathbf{D}$, of which the k-th element is $\frac{1}{2\|\mathbf{v}^k\|_2}$;

3. Update $\mathbf{w}$ by $\mathbf{w} - \mathbf{B}^{-1}\mathbf{a}$, where $(d \times (p-1) + u)$-th element of $\mathbf{a} \in \mathbb{R}^{dc_1 \times 1}$ is $\frac{\partial\left[l_1(\mathbf{W}) + \gamma\,\mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)\right]}{\partial \mathbf{W}_{up}}$ for $1 \le u \le d, 1 \le p \le c_1$, the $(d \times (p-1) + u, d \times (q-1) + v)$-th element of $\mathbf{B} \in \mathbb{R}^{dc_1 \times dc_1}$ is $\frac{\partial\left[l_1(\mathbf{W}) + \gamma\,\mathbf{tr}\left(\mathbf{W}^T\mathbf{D}\mathbf{W}\right)\right]}{\partial \mathbf{W}_{up}\partial \mathbf{W}_{vq}}$ for $1 \le u, v \le d$ and $1 \le p, q \le c_1$.

Construct the updated $\mathbf{W} \in \mathbb{R}^{d \times c_1}$ by the updated vector $\mathbf{w} \in \mathbb{R}^{dc_1}$, where the (u, p)-th element of $\mathbf{W}$ is the $(d \times (p-1) + u)$-th element of $\mathbf{w}$;

4. Update $\mathbf{P}$ by $\mathbf{P} = \left(\mathbf{X}\mathbf{X}^T + \gamma\mathbf{D}\right)^{-1}\mathbf{X}\mathbf{Z}$;

5. Update $\mathbf{V}$ by $\mathbf{V} = [\mathbf{W}\ \mathbf{P}]$;

end

Output: $\mathbf{W} \in \mathbb{R}^{d \times c_1}$ and $\mathbf{P} \in \mathbb{R}^{d \times c_2}$.