Sparse Bayesian Multi-Task Learning for Predicting Cognitive Outcomes from Neuroimaging Measures in Alzheimer's Disease

Jing Wan[1,2]*, Zhilin Zhang[3]*, Jingwen Yan[1], Taiyong Li[1,4], Bhaskar D. Rao[3], Shiaofen Fang[2], Sungeun Kim[1], Shannon L. Risacher[1], Andrew J. Saykin[1], Li Shen[1,2]†, for the ADNI‡
[1]Indiana University, [2]Purdue University, [3]University of California, San Diego,
[4]Southwestern Univ. of Finance and Economics

Abstract

Alzheimer's disease (AD) is the most common form of dementia that causes progressive impairment of memory and other cognitive functions. Multivariate regression models have been studied in AD for revealing relationships between neuroimaging measures and cognitive scores to understand how structural changes in brain can influence cognitive status. Existing regression methods, however, do not explicitly model dependence relation among multiple scores derived from a single cognitive test. It has been found that such dependence can deteriorate the performance of these methods. To overcome this limitation, we propose an efficient sparse Bayesian multi-task learning algorithm, which adaptively learns and exploits the dependence to achieve improved prediction performance. The proposed algorithm is applied to a real world neuroimaging study in AD to predict cognitive performance using MRI scans. The effectiveness of the proposed algorithm is demonstrated by its superior prediction performance over multiple state-of-the-art competing methods and accurate identification of compact sets of cognition-relevant imaging biomarkers that are consistent with prior knowledge.

1. Introduction

Alzhiemer's disease (AD) is a neurodegenerative disorder characterized by progressive impairment of memory and other cognitive functions. Substantial attention has recently been given to identifying neuroimaging predictors for cognitive decline in AD in the fields of medical image analysis and pattern recognition. Regression models have been investigated to predict clinical scores from individual magnetic resonance imaging (MRI) and/or positron emission tomography (PET) scans [14, 17, 18, 24]. In [17], stepwise regression was performed in a univariate, pairwise fashion to relate each imaging measure to each cognitive score. In [14], using relevance vector regression, morphometric features of the entire brain were jointly analyzed to predict each selected clinical score. Two most recent studies [18, 24] employed multi-task learning strategies and aimed to select features that could predict all or most clinical scores, using $\ell_{2,1}$-norm coupled with ℓ_1-norm [18] and multi-task feature selection coupled with support vector machine [24]. Both methods used a simple concatenation to bundle multiple clinical scores together without learning their dependence relation.

In this study we propose a new sparse Bayesian multi-task learning method, which is built on a multivariate regression model and explicitly models the correlation structure within each row of the regression coefficient matrix. This is motivated by the fact that if a biomarker plays a role in one score of a cognitive test, then it often has more or less influence in another score of the same test. The proposed method is evaluated in an empirical study using the MRI and cognitive data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [19]. This method not only demonstrates superior performance over multiple state-of-the-art competing methods, but also identifies cognition-relevant imaging biomarkers that are consistent with prior knowledge.

2. Sparse Bayesian Multi-Task Learning

The multiple measurement vector (MMV) model, originally designed for sparse signal recovery [5], is adopted here for multivariate regression of cognitive scores $\mathbf{Y}$ on neuroimaging measures $\mathbf{\Phi}$:

$$\mathbf{Y} = \mathbf{\Phi X} + \mathbf{V}, \tag{1}$$

*Equal contribution by Jing Wan (wanjing@iupui.edu) and Zhilin Zhang (z4zhang@ucsd.edu).
†Correspondence to Li Shen (shenli@iupui.edu)
‡Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

where $\mathbf{\Phi} \in \mathbb{R}^{N\times M}$ and $\mathbf{Y} \in \mathbb{R}^{N\times L}$ are respectively the M neuroimaging measures and L cognitive scores of the N subjects, $\mathbf{V}$ is an unknown noise matrix (or called model error matrix), and $\mathbf{X} \in \mathbb{R}^{M\times L}$ is an unknown coefficient matrix. $\mathbf{X}$ is expected to have a sparse loading (i.e., only a few nonzero rows), since the brain circuitry relevant to a certain cognition task typically involves a small number of imaging markers, and these markers more or less affect all the cognitive scores under the task. Here a nonzero row is allowed to contain some zero entries.

There are many algorithms for this problem. Most of them calculate the solution by solving the following unconstrained problem (or its equivalent constrained problem)

$$\mathbf{X} = \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{X}\|_{\mathcal{F}}^2 + \lambda g_1(\mathbf{X}) \tag{2}$$

with the mixed $\ell_{q,1}$ penalty (typically, $q = 2$ or $q = \infty$) [1]

$$g_1(\mathbf{X}) \triangleq \sum_{i=1}^{M} \|\mathbf{X}_{i\cdot}\|_q, \tag{3}$$

where an ℓ_q norm is applied on each row of $\mathbf{X}$, and an ℓ_1 norm is applied on the M calculated ℓ_q norms. λ is a regularizer, which is generally tuned by cross-validation. Algorithms using this penalty include group Lasso (the variant used for the MMV model) [23], MMV based Basis Pursuit (M-BP) [4], and many domain-specific algorithms for feature extraction [13]. Note that the penalty $g_1(\mathbf{X})$ is a convex penalty. In some scenarios non-convex penalty based algorithms can yield better performance. A typical non-convex penalty is:

$$g_2(\mathbf{X}) \triangleq \sum_{i=1}^{M} \left(\|\mathbf{X}_{i\cdot}\|_2\right)^p, \quad 0 < p < 1 \tag{4}$$

The MMV based FOCal Underdetermined System Solver (M-FOCUSS) method [5] is a representative in this group.

However, instead of using these algorithms in our problem, we prefer the T-MSBL[2] algorithm [26], a recently derived variant in the family of sparse Bayesian learning (SBL) [15]. SBL is a powerful approach for regression and classification. It relies on a parameterized prior that encourages models with a few nonzero rows in $\mathbf{X}$ (*i.e.* encourages row-wise sparsity). But among various SBL algorithms we prefer T-MSBL due to the following specific reasons.

First, in our case a given imaging marker can affect multiple cognitive scores, so the coefficients in the same row of $\mathbf{X}$ are largely correlated. Recently, it has been found [26] that when such correlation is present, most existing methods have seriously degraded performance due to the ignorance of the correlation[3]. In contrast, T-MSBL can *adaptively* estimate and exploit the correlation structure in $\mathbf{X}_{i\cdot}(\forall i)$ to improve performance. Extensive experiments have verified its superior performance to most algorithms.

Second, in our case the columns of $\mathbf{\Phi}$ is highly coherent. In some data sets used in our experiments the maximum correlation reaches 0.95. The coherent $\mathbf{\Phi}$ can result in poor performance of most algorithms. In contrast, experiments have shown that T-MSBL maintains its superior performance in this situation.

Third, T-MSBL has an effective learning rule to choose a suitable value for the regularizer λ. This relaxes the efforts of users to choose a value for λ as in many algorithms (*e.g.* M-BP and M-FOCUSS).

However, the T-MSBL is slow due to the use of the Expectation-Maximization (EM) method. In the following, we first briefly describe T-MSBL, and then propose a much faster algorithm, which is suitable for large-scale data sets. Further, we reveal its connection to some popular algorithms, including those using the mixed $\ell_{2,1}$ penalty (3) and those using kernels for regularization. This connection provides insights to the advantages of the proposed algorithm, and motivates the design of new algorithms in the future.

2.1. The T-MSBL Algorithm

In T-MSBL, each row $\mathbf{X}_{i\cdot}$ is assumed to satisfy a parameterized Gaussian distribution, given by

$$p(\mathbf{X}_{i\cdot}; \gamma_i, \mathbf{B}_i) \sim \mathcal{N}(\mathbf{0}, \gamma_i\mathbf{B}_i), \quad i = 1, \cdots, M$$

where γ_i and $\mathbf{B}_i$ are hyperparameters. γ_i is a nonnegative scalar controlling the row sparsity of $\mathbf{X}$. When $\gamma_i = 0$, the corresponding i-th row, $\mathbf{X}_{i\cdot}$, becomes zero. Due to the mechanism of automatic relevance determination [12, 15], most γ_i become zero in noiseless cases or tend to very small values in noisy cases. Generally a threshold is used to prune out these γ_i, which equivalently prunes out the corresponding rows in $\mathbf{X}$. $\mathbf{B}_i$ is an unknown positive definite matrix modeling correlation structure in $\mathbf{X}_{i\cdot}$, which is *adaptively* learned from data. It is worthy of emphasizing that the *data-adaptive* learning of the correlation structure is very important, which can effectively prevent T-MSBL from converging to local solutions in most cases [26].

To conveniently derive T-MSBL, the MMV model (1) is equivalently transformed to the block sparsity model

$$\mathbf{y} = \mathbf{D}\mathbf{x} + \mathbf{v}, \tag{5}$$

where $\mathbf{y} = \mathrm{vec}(\mathbf{Y}^T) \in \mathbb{R}^{NL\times 1}$, $\mathbf{x} = \mathrm{vec}(\mathbf{X}^T) \in \mathbb{R}^{ML\times 1}$,

[1] Throughout the paper $\mathbf{X}_{i\cdot}$ and $\mathbf{X}_{\cdot j}$ denotes the i-th row and the j-th column of $\mathbf{X}$, respectively.

[2] T-MSBL stands for Temporal MMV Sparse Bayesian Learning. Note that the concept of "Temporal" was derived from sparse signal recovery, the original application domain of T-MSBL. In our application, the temporal dimension corresponds to the dimension of multiple cognitive scores.

[3] For example, the operator $\|\mathbf{X}_{i\cdot}\|_q$ in (3) and (4) does not consider correlation structure among the entries in $\mathbf{X}_{i\cdot}$.

$\mathbf{v} = \text{vec}(\mathbf{V}^T)$, and $\mathbf{D} = \mathbf{\Phi} \otimes \mathbf{I}_L$[4]. Here $\mathbf{v}$ is assumed to be a Gaussian distribution $p(\mathbf{v}; \lambda) = \mathcal{N}(0, \lambda\mathbf{I})$. Using the Bayes rule, the posterior is $p(\mathbf{x}|\mathbf{y}; \Theta) = \mathcal{N}(\boldsymbol{\mu}, \mathbf{\Sigma})$ with

$$\begin{aligned} \boldsymbol{\mu} &= \frac{1}{\lambda}\mathbf{\Sigma}\mathbf{D}^T\mathbf{y} && (6) \\ \mathbf{\Sigma} &= (\mathbf{\Sigma}_0^{-1} + \frac{1}{\lambda}\mathbf{D}^T\mathbf{D})^{-1} && (7) \\ &= \mathbf{\Sigma}_0 - \mathbf{\Sigma}_0\mathbf{D}^T(\lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T)^{-1}\mathbf{D}\mathbf{\Sigma}_0 && (8) \end{aligned}$$

where Θ denotes the set of all the hyperparameters $\{\lambda, \gamma_i, \mathbf{B}_i, \forall i\}$, and $\mathbf{\Sigma}_0 \triangleq \text{diag}\{\gamma_1\mathbf{B}_1, \cdots, \gamma_M\mathbf{B}_M\}$[5]. Once these hyperparameters are estimated, the estimate of $\mathbf{x}$ is readily given by the mean of the posterior, *i.e.* $\boldsymbol{\mu}$. The original T-MSBL uses the EM method to estimate these hyperparameters, and thus it is slow. In the following we will derive a much faster algorithm based on MacKay's fixed-point method [11].

2.2. The New Algorithm: T-MSBL-FP

We estimate the hyperparameters in the evidence maximization framework [11, 15]. In this framework the cost function is

$$\begin{aligned} \mathcal{L}(\Theta) &\triangleq -2\log\int p(\mathbf{y}|\mathbf{x};\lambda)p(\mathbf{x};\gamma_i,\mathbf{B}_i,\forall i)d\mathbf{x} \\ &= \mathbf{y}^T(\mathbf{\Sigma}_y)^{-1}\mathbf{y} + \log|\mathbf{\Sigma}_y|, && (9) \end{aligned}$$

where $\mathbf{\Sigma}_y \triangleq \lambda\mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T$. As in T-MSBL, all the $\mathbf{B}_i(\forall i)$ are constrained to be the same matrix $\mathbf{B}$ to prevent overfitting. Thus $\mathbf{\Sigma}_0 = \mathbf{\Gamma} \otimes \mathbf{B}$ with $\mathbf{\Gamma} \triangleq \text{diag}(\boldsymbol{\gamma})$ and $\boldsymbol{\gamma} \triangleq [\gamma_1, \cdots, \gamma_M]^T$.

To conveniently derive learning rules for these hyperparameters, we first simplify $\mathcal{L}(\Theta)$. First, note that

$$\begin{aligned} \mathbf{y}^T(\mathbf{\Sigma}_y)^{-1}\mathbf{y} &= \frac{1}{\lambda}\mathbf{y}^T\big[\mathbf{y} - \mathbf{D}(\lambda\mathbf{\Sigma}_0^{-1} + \mathbf{D}^T\mathbf{D})^{-1}\mathbf{D}^T\mathbf{y}\big] \\ &= \frac{1}{\lambda}\mathbf{y}^T\big[\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\big] && (10) \\ &= \frac{1}{\lambda}\big[\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^T\mathbf{D}^T\mathbf{y} - \boldsymbol{\mu}^T\mathbf{D}^T\mathbf{D}\boldsymbol{\mu}\big] \\ &= \frac{1}{\lambda}\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\|_2^2 \\ &\quad + \boldsymbol{\mu}^T\big(\mathbf{\Sigma}^{-1} - \frac{1}{\lambda}\mathbf{D}^T\mathbf{D}\big)\boldsymbol{\mu} && (11) \\ &= \frac{1}{\lambda}\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^T\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu} && (12) \end{aligned}$$

where (10) and (11) both used the equation (6), and (12) used the equation (7). Next, using the Sylvester's Determinant Theorem, we have

$$\begin{aligned} \log|\mathbf{\Sigma}_y| &= \log|\lambda\mathbf{I}_{NL}| \\ &\quad + \log|\mathbf{\Sigma}_0^{-1} + \frac{1}{\lambda}\mathbf{D}^T\mathbf{D}| + \log|\mathbf{\Sigma}_0|. && (13) \end{aligned}$$

Combining (12) and (13), the cost function becomes

$$\begin{aligned} \mathcal{L}(\Theta) &= \frac{1}{\lambda}\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\|_2^2 + \boldsymbol{\mu}^T\mathbf{\Sigma}_0^{-1}\boldsymbol{\mu} + \log|\lambda\mathbf{I}_{NL}| \\ &\quad + \log\big|\mathbf{\Sigma}_0^{-1} + \frac{1}{\lambda}\mathbf{D}^T\mathbf{D}\big| + \log|\mathbf{\Sigma}_0|. && (14) \end{aligned}$$

Now it is convenient to minimize the cost function with respect to each hyperparameter.

The derivative of $\mathcal{L}(\Theta)$ with respect to γ_i is

$$\frac{\partial\mathcal{L}}{\partial\gamma_i} = -\frac{\boldsymbol{\mu}_i^T\mathbf{B}^{-1}\boldsymbol{\mu}_i}{\gamma_i^2} - \frac{\text{Tr}\big(\mathbf{\Sigma}_i\mathbf{B}^{-1}\big)}{\gamma_i^2} + \frac{L}{\gamma_i}$$

where $\boldsymbol{\mu}_i \triangleq \boldsymbol{\mu}((i-1)L+1 \ : \ iL)$, and $\mathbf{\Sigma}_i \triangleq \mathbf{\Sigma}((i-1)L+1 \ : \ iL \ , \ (i-1)L+1 \ : \ iL)$ (using the MATLAB notations). Letting $\frac{\partial\mathcal{L}}{\partial\gamma_i} = 0$ and following MacKay's fixed-point approach [11, 15], we have

$$\gamma_i \leftarrow \frac{\boldsymbol{\mu}_i^T\mathbf{B}^{-1}\boldsymbol{\mu}_i}{L - \text{Tr}\big(\mathbf{\Sigma}_i\mathbf{B}^{-1}\big)/\gamma_i}, \quad i = 1, \cdots, M \qquad (15)$$

Similarly, we derive the learning rules for $\mathbf{B}$ and λ:

$$\begin{aligned} \mathbf{B} &\leftarrow \frac{1}{M}\sum_{i=1}^{M}\frac{\boldsymbol{\mu}_i\boldsymbol{\mu}_i^T + \mathbf{\Sigma}_i}{\gamma_i} && (16) \\ \lambda &\leftarrow \frac{\|\mathbf{y} - \mathbf{D}\boldsymbol{\mu}\|_2^2 + \lambda\big[ML - \text{Tr}(\mathbf{\Sigma}\mathbf{\Sigma}_0^{-1})\big]}{NL}. && (17) \end{aligned}$$

The learning rules (6), (7), (15), (16), and (17) comprise our algorithm. This algorithm operates in the block sparsity model (5), not the original MMV model (1). But we can simplify it using the approximation equation [26]:

$$(\lambda\mathbf{I}_{NL} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T)^{-1} \approx (\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi})^{-1} \otimes \mathbf{B}^{-1}. \qquad (18)$$

This approximation performs quite well over a broader range of conditions, especially when SNR is high or the correlation in each $\mathbf{X}_{i\cdot}$ is weak. It becomes exact when $\lambda = 0$ or $\mathbf{B} = \mathbf{I}$. Using (18) and following the simplification procedure in [26] we obtain the simplified algorithm as follows:

$$\begin{aligned} \mathbf{\Xi} &\leftarrow \big(\mathbf{\Gamma}^{-1} + \frac{1}{\lambda}\mathbf{\Phi}^T\mathbf{\Phi}\big)^{-1} \\ \mathbf{X} &\leftarrow \mathbf{\Gamma}\mathbf{\Phi}^T\big(\lambda\mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T\big)^{-1}\mathbf{Y} \\ \gamma_i &\leftarrow \frac{\mathbf{X}_{i\cdot}\mathbf{B}^{-1}\mathbf{X}_{i\cdot}^T}{L(1 - \mathbf{\Xi}_{ii}/\gamma_i)}, \quad \forall i \end{aligned}$$

[4] We denote the $L \times L$ identity matrix by $\mathbf{I}_L$. When the dimension is evident from the context, we simply use $\mathbf{I}$. $\otimes$ denotes the Kronecker product. $\text{vec}(\cdot)$ denotes the vectorization of a matrix formed by stacking its columns into a single column vector.

[5] $\text{diag}\{\gamma_1\mathbf{B}_1, \cdots, \gamma_M\mathbf{B}_M\}$ indicates a block diagonal matrix with its i-th diagonal block given by $\gamma_i\mathbf{B}_i$.

$$\mathbf{B} \leftarrow \widetilde{\mathbf{B}}/\|\widetilde{\mathbf{B}}\|_{\mathcal{F}}, \quad \text{with} \quad \widetilde{\mathbf{B}} = \sum_{i=1}^{M} \frac{\mathbf{X}_{i\cdot}^T \mathbf{X}_{i\cdot}}{\gamma_i}$$

$$\lambda \leftarrow \frac{1}{NL}\|\mathbf{Y} - \mathbf{\Phi X}\|_{\mathcal{F}}^2 + \frac{\lambda}{N}\mathrm{Tr}\big[\mathbf{\Phi\Gamma\Phi}^T(\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^T)^{-1}\big]$$

where $\mathbf{\Xi}_{ii}$ is the (i,i)-th element of $\mathbf{\Xi}$. We call the algorithm **T-MSBL-FP** (i.e, T-MSBL-Fixed Point). Note that the robustness of the λ learning rule in noisy environment can be improved by setting the off-diagonal elements of $\mathbf{\Phi\Gamma\Phi}^T$ to zeros [26]. Also, the robustness of the learning rule for $\mathbf{B}$ can be improved by adopting the regularization method in [26]. The initialization values of $\gamma_i(\forall i)$, $\mathbf{B}$, and λ can be chosen 1, $\mathbf{I}$, and any guessed noise variance, respectively.

T-MSBL-FP not only is much faster than T-MSBL, but also has better prediction performance. More interestingly, from its cost function (9) we can connect it to many well-established algorithms, providing insights to our algorithm and motivations to design new algorithms. We elaborate this next.

2.3. Connection to Existing Algorithms

Our motivation of connecting T-MSBL-FP to existing algorithms is inspired by the work in [20, 25]. In [20] Wipf and Nagarajan connected the basic SBL algorithm [15] to ℓ_1 minimization algorithms in the single measurement vector model, a related but different model to the MMV model considered here. In [25] Zhang and Rao connected T-MSBL to iterative reweighted ℓ_2 algorithms. Now we connect T-MSBL-FP to the algorithms based on the $\ell_{q,1}$ penalty (3) and those employing kernel regularizers [22, 13, 21].

We consider to transform the cost function (9). Using the identity $\mathbf{y}^T(\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T)^{-1}\mathbf{y} \equiv \min_{\mathbf{x}}\big[\frac{1}{\lambda}\|\mathbf{y} - \mathbf{Dx}\|_2^2 + \mathbf{x}^T\mathbf{\Sigma}_0^{-1}\mathbf{x}\big]$, the upper-bound of the cost function (9) is

$$\mathfrak{L}(\mathbf{x}, \gamma, \mathbf{B}) = \log|\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T| + \frac{1}{\lambda}\|\mathbf{y} - \mathbf{Dx}\|_2^2 + \mathbf{x}^T\mathbf{\Sigma}_0^{-1}\mathbf{x}.$$

By first minimizing it over $\boldsymbol{\gamma}$ and $\mathbf{B}$ and then minimizing over $\mathbf{x}$, we have:

$$\mathbf{x} = \arg\min_{\mathbf{x}}\Big\{\|\mathbf{y} - \mathbf{Dx}\|_2^2 + \lambda g_{\mathrm{C}}(\mathbf{x})\Big\}, \tag{19}$$

with the penalty $g_{\mathrm{C}}(\mathbf{x})$ given by [6]

$$g_{\mathrm{C}}(\mathbf{x}) \triangleq \min_{\boldsymbol{\gamma}\succeq\mathbf{0}, \mathbf{B}\succ\mathbf{0}}\Big\{\mathbf{x}^T\mathbf{\Sigma}_0^{-1}\mathbf{x} + \log|\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T|\Big\}. \tag{20}$$

[6] $\boldsymbol{\gamma}\succeq 0$ means each element of $\boldsymbol{\gamma}$ is nonnegative. $\mathbf{B}\succ 0$ means $\mathbf{B}$ is a positive definite matrix.

We now look at the concavity of $g_{\mathrm{C}}(\mathbf{x})$. Since the function $h(\boldsymbol{\gamma}) \triangleq \log|\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T|$ is concave and non-decreasing with respect to $\boldsymbol{\gamma}\succeq\mathbf{0}$, we have

$$\log|\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T| \triangleq \min_{\mathbf{z}\succeq\mathbf{0}} \mathbf{z}^T\boldsymbol{\gamma} - h^*(\mathbf{z}) \tag{21}$$

where $h^*(\mathbf{z})$ is the concave conjugate of $h(\boldsymbol{\gamma})$ [2], and $\mathbf{z} \triangleq [z_1, \cdots, z_M]^T$. Thus using (21) we can express (20) as

$$\begin{aligned} g_{\mathrm{C}}(\mathbf{x}) &= \min_{\boldsymbol{\gamma},\mathbf{z}\succeq\mathbf{0},\mathbf{B}\succ\mathbf{0}} \mathbf{x}^T\mathbf{\Sigma}_0^{-1}\mathbf{x} + \mathbf{z}^T\boldsymbol{\gamma} - h^*(\mathbf{z}) \\ &= \min_{\boldsymbol{\gamma},\mathbf{z}\succeq\mathbf{0},\mathbf{B}\succ\mathbf{0}} \sum_i \Big(\frac{\mathbf{x}_i^T\mathbf{B}^{-1}\mathbf{x}_i}{\gamma_i} + z_i\gamma_i\Big) - h^*(\mathbf{z}) \end{aligned} \tag{22}$$

where $\mathbf{x}_i \triangleq \mathbf{x}((i-1)L+1 \ : \ iL)$, *i.e.* $\mathbf{x}_i \triangleq \mathbf{X}_{i\cdot}^T$. Minimizing (22) over γ_i for fixed $\mathbf{x}$, $\mathbf{z}$ and $\mathbf{B}$, we get

$$\gamma_i = z_i^{-\frac{1}{2}}\sqrt{\mathbf{x}_i^T\mathbf{B}^{-1}\mathbf{x}_i}, \quad \forall i \tag{23}$$

Substituting this expression into (22) leads to

$$g_{\mathrm{C}}(\mathbf{x}) = \min_{\mathbf{z}\succeq\mathbf{0},\mathbf{B}\succ\mathbf{0}} \sum_i \big(2z_i^{\frac{1}{2}}\sqrt{\mathbf{x}_i^T\mathbf{B}^{-1}\mathbf{x}_i}\big) - h^*(\mathbf{z}). \tag{24}$$

Now, from (19) and (24) we have:

$$\mathbf{x} = \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{Dx}\|_2^2 + \lambda\Big[\min_{\mathbf{z}\succeq\mathbf{0},\mathbf{B}\succ\mathbf{0}} \sum_i \big(2z_i^{\frac{1}{2}}\sqrt{\mathbf{x}_i^T\mathbf{B}^{-1}\mathbf{x}_i}\big) - h^*(\mathbf{z})\Big] \tag{25}$$

To obtain the solution $\mathbf{x}$, we need to first calculate the optimal values of $\mathbf{B}$ and z_i.

The optimal value of $\mathbf{B}$ can be obtained from (20). Note that $\frac{\partial}{\partial\mathbf{B}}[\mathbf{x}^T\mathbf{\Sigma}_0^{-1}\mathbf{x} + \log|\lambda\mathbf{I} + \mathbf{D\Sigma}_0\mathbf{D}^T|] = \sum_i\big[-\mathbf{B}^{-1}\mathbf{x}_i\mathbf{x}_i^T\mathbf{B}^{-1}/\gamma_i + \gamma_i\mathbf{D}_i^T\mathbf{\Sigma}_y^{-1}\mathbf{D}_i\big]$, where $\mathbf{D}_i = \mathbf{\Phi}_i \otimes \mathbf{I}_L$ and $\mathbf{\Phi}_i$ is the i-th column of $\mathbf{\Phi}$. Setting it to zero, we have

$$\begin{aligned} \mathbf{B}^{-1}\sum_i \frac{\mathbf{x}_i\mathbf{x}_i^T}{\gamma_i}\mathbf{B}^{-1} &= \sum_i \gamma_i\mathbf{D}_i^T\mathbf{\Sigma}_y^{-1}\mathbf{D}_i \\ &= \sum_i \gamma_i(\mathbf{\Phi}_i^T \otimes \mathbf{I})\big(\lambda\mathbf{I}_{NL} + (\mathbf{\Phi\Gamma\Phi}^T)\otimes\mathbf{B}\big)^{-1}(\mathbf{\Phi}_i \otimes \mathbf{I}) \\ &\approx \sum_i \gamma_i(\mathbf{\Phi}_i^T \otimes \mathbf{I})\big[(\lambda\mathbf{I}_N + \mathbf{\Phi\Gamma\Phi}^T)^{-1} \otimes \mathbf{B}^{-1}\big] \\ &\quad \cdot(\mathbf{\Phi}_i \otimes \mathbf{I}) \\ &= \Big[\sum_i \gamma_i\mathbf{\Phi}_i^T(\lambda\mathbf{I} + \mathbf{\Phi\Gamma\Phi}^T)^{-1}\mathbf{\Phi}_i\Big]\mathbf{B}^{-1} \end{aligned} \tag{26}$$

where (26) used the approximation (18). Thus, we obtain the learning rule

$$\mathbf{B} = \frac{1}{C}\sum_{i=1}^{M}\frac{\mathbf{x}_i\mathbf{x}_i^T}{\gamma_i} = \frac{1}{C}\sum_{i=1}^{M}\frac{\mathbf{X}_{i\cdot}^T\mathbf{X}_{i\cdot}}{\gamma_i} \tag{27}$$

with $C \triangleq \sum_{i=1}^{M} \gamma_i \mathbf{\Phi}_i^T (\lambda \mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1} \mathbf{\Phi}_i$.

According to the duality property [2] in convex optimization, from the relation (21) we can directly obtain the optimal z_i as follows $z_i = \frac{\partial \log |\lambda \mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T|}{\partial \gamma_i} = \mathrm{Tr}\big[\mathbf{B}\mathbf{D}_i^T \big(\lambda \mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T\big)^{-1}\mathbf{D}_i\big]$. So,

$$\begin{aligned} z_i^{\frac{1}{2}} &= \Big(\mathrm{Tr}\big[\mathbf{B}\mathbf{D}_i^T \big(\lambda \mathbf{I} + \mathbf{D}\mathbf{\Sigma}_0\mathbf{D}^T\big)^{-1}\mathbf{D}_i\big]\Big)^{\frac{1}{2}} \\ &\approx \sqrt{L\mathbf{\Phi}_i^T (\lambda \mathbf{I} + \mathbf{\Phi}\mathbf{\Gamma}\mathbf{\Phi}^T)^{-1}\mathbf{\Phi}_i}, \end{aligned} \tag{28}$$

where we used the approximation (18) again.

Based on the above development, we see that the optimal values of $\mathbf{B}$ and z_i depend on $\mathbf{X}$ itself. Thus the whole learning procedure is an iterative algorithm. In the k-th iteration, once having used the updating rules (23) (27) and (28) to obtain $\mathbf{B}^{(k)}$ and the weight $w_i^{(k)} \triangleq 2z_i^{1/2}$, we only need to solve the following optimization problem:

$$\begin{aligned} \mathbf{x}^{(k+1)} &= \arg\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{D}\mathbf{x}\|_2^2 \\ &\quad + \lambda \sum_i w_i^{(k)} \sqrt{\mathbf{x}_i^T (\mathbf{B}^{(k)})^{-1} \mathbf{x}_i}, \end{aligned} \tag{29}$$

or equivalently,

$$\begin{aligned} \mathbf{X}^{(k+1)} &= \arg\min_{\mathbf{X}} \|\mathbf{Y} - \mathbf{\Phi}\mathbf{X}\|_{\mathcal{F}}^2 \\ &\quad + \lambda \sum_i w_i^{(k)} \sqrt{\mathbf{X}_{i\cdot} (\mathbf{B}^{(k)})^{-1} \mathbf{X}_{i\cdot}^T}. \end{aligned} \tag{30}$$

Now we draw its connection to existing algorithms. First, from (30) we can see our penalty is a correlation-aware penalty, and the correlation structure is adaptively learned from data. This is entirely different to the penalties in (3) and (4), which is blind to the correlation. Further, the matrix $\mathbf{B}^{(k)}$ in our penalty can be viewed as a data-adaptive kernel. This is different to the non-adaptive kernels used in some existing $\ell_{2,1}$-norm penalties [22], which generally need users to design kernels according to some a priori knowledge or by cross-validation. Note that the data-adaptive kernel is advantageous over the user-defined kernels, because in some applications such as our application, a priori knowledge may not be available. Also, user-designed kernels may not accurately capture the correlation structure of data.

Second, one can see (30) is an MMV-model based iterative reweighted ℓ_1 minimization algorithm [3], since its weights $w_i^{(k)}$ depends on the estimate of $\mathbf{X}$ in the previous iteration. In contrast, the framework expressed in (2)-(3) is a non-iterative-reweighted one. It is known that iterative reweighted algorithms have better performance than their non-iterative-reweighted counterparts and can provide more sparse solutions [3].

The above observations give us an intuitive, although not rigorous, explanation why our algorithm has superior performance as shown in the experiments described below. And they motivate us how to improve algorithms based on the $\ell_{q,1}$ norm and kernel regularizers, especially how to adaptively learn the correlation structure of data.

Table 1. Participant characteristics.

Category	HC	AD	p-value
Gender (M/F)	114/108	86/85	0.835
Handedness (R/L)	205/17	161/10	0.482
Baseline Age (years)	75.93 ± 5.08	75.67 ± 7.36	0.680
Education (years)	15.97 ± 2.84	14.74 ± 3.08	< 0.001

3. Experimental Results

3.1. Data Sets

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). One goal of ADNI has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org. All the healthy control (HC) and AD participants with no missing cognitive and MRI measures were included in this study. Their characteristics are summarized in Table 1.

For one baseline scan of each participant, FreeSurfer V4 was employed to automatically label cortical and subcortical tissue classes [6, 8] and to extract target region volume and cortical thickness, as well as to extract total intracranial volume (ICV). For each hemisphere, thickness measures of 34 cortical regions of interest (ROIs) and volume measures of 15 cortical and subcortical ROIs (Fig. 1) were included in this study. Three sets of baseline cognitive scores [1] were employed to test the proposed methods: Mini-Mental State Exam (MMSE), Rey Auditory Verbal Learning Test (RAVLT), and Trail Making (TRAILS). Details about these assessments are available in the ADNI procedure manuals (www.adni-info.org). Table 2(a) summarizes these cognitive scores. Using the regression weights derived from the healthy participants, all the FreeSurfer measures were adjusted for the baseline age, gender, education, handedness, and ICV, and all the cognitive measures were adjusted for the baseline age, gender, education and handedness.

3.2. Competing Methods

To show the superior performance of our algorithm, we selected several state-of-the-art or classical algorithms for comparison; each algorithm represents a group of methods using different frameworks. They are the Mixed ℓ_2/ℓ_1 Program [7], M-FOCUSS [5], Simultaneous Orthogonal Matching Pursuit (S-OMP) [16], Multi-Task Compressive Sensing (MT-CS) [10], and Ridge Regression [9]. The Mixed ℓ_2/ℓ_1 Program belongs to the group (2)-(3) with

Table 2. Comparison of cross-validation prediction performances measured by correlation coefficients

(a) Description of Cognitive Measures			(b) Cross-validation Prediction Performances						
Score Name		Description	T-MSBL-FP	T-MSBL	M-FOCUSS	Mixed ℓ_2/ℓ_1	S-OMP	RIDGE	MT-CS
MMSE		MMSE score	0.735	0.735	0.690	0.689	0.721	0.685	0.680
RAVLT	TOTAL	Total score of the first 5 trials	0.634	0.617	0.589	0.586	0.604	0.570	0.579
	T30	30min delay total # of words recalled	0.586	0.572	0.550	0.543	0.545	0.486	0.512
	RECOG	30min delay recognition score	0.561	0.559	0.526	0.501	0.539	0.504	0.509
TRAILS	TRAILSA	Trail making A score	0.467	0.450	0.391	0.380	0.400	0.312	0.344
	TRAILSB	Trail making B score	0.565	0.555	0.491	0.461	0.508	0.464	0.476
	TR(B-A)	TRAILSB-TRAILSA	0.488	0.464	0.401	0.351	0.409	0.336	0.355

$q = 2$. It is shown [7] that it has better performance than many other members in this group. M-FOCUSS represents the group using the non-convex penalty (4). In our experiment, we set $p = 0.8$ as suggested in [5]. S-OMP represents the group of greedy pursuit algorithms for the MMV model. MT-CS is an SBL algorithm, which treats the MMV model (1) as L *dependent* single measurement vector (SMV) models, *i.e.* $\mathbf{Y}_{\cdot i} = \mathbf{\Phi}\mathbf{X}_{\cdot i} + \mathbf{V}_{\cdot i}$ $(i = 1, \cdots, L)$, where every $\mathbf{X}_{\cdot i} (\forall i)$ shares a common prior. Note that this model is an alternative one to the MMV model in multi-task learning. Ridge Regression is a traditional regression approach for an SMV model. To use it in our problem, we applied it to each $\mathbf{Y}_{\cdot i} = \mathbf{\Phi}\mathbf{X}_{\cdot i} + \mathbf{V}_{\cdot i}$ $(i = 1, \cdots, L)$ independently.

3.3. Improved Performance

Regression was performed separately on each cognitive task (MMSE, RAVLT, or TRAILS) using the MRI measures as predictors, where the proposed T-MSBL-FP method and all the competing methods (T-MSBL, M-FOCUSS, Mixed ℓ_2/ℓ_1, S-OMP, RIDGE, MT-CS) were evaluated. Similar to prior studies [14, 24], in each experiment, Pearson's correlation coefficients r between the actual and predicted cognitive scores were computed to measure the prediction performance. Using a 5-fold cross-validation strategy, the testing samples across five trials were pulled together to obtain an unbiased estimate of these correlation coefficients.

Shown in Table 2(b) is the performance comparison among all seven methods. Both T-MSBL-FP and T-MSBL outperformed the other five competing algorithms in all three prediction cases. In the multi-task learning cases (i.e., RAVLT and TRAILS, where $L > 1$ for $\mathbf{Y}$ in the MMV model (1)), T-MSBL-FP outperformed T-MSBL. Besides better prediction accuracy, T-MSBL-FP also achieved significantly improved computational performance by almost one order of magnitude[7] over T-MSBL, i.e., 0.31s vs. 2.12s for MMSE, 0.11s vs. 4.68s for RAVLT, and 0.17s vs 1.36s for TRAILS.

Using T-MSBL-FP, the MRI measures could predict the MMSE score the best, with a correlation coefficient $r = 0.7352$. This result is better than or competitive to a few prior MMSE prediction results: $r = 0.504$ using MRI only in [24], $r = 0.697$ using MRI, PET and CSF jointly in [24], and $r = 0.70$ using MRI in [14]. Relatively high prediction performance has also been achieved for RAVLT scores, from $r = 0.561$ to $r = 0.634$. In [14], a different, but relevant RAVLT score was predicted using MRI, with $r = 0.13$ only.

[7] Since in SBL algorithms the thresholds to prune out small γ_i affect their speed, the thresholds of the two algorithms were set to be the same (10^{-3}), making the speed comparison fair.

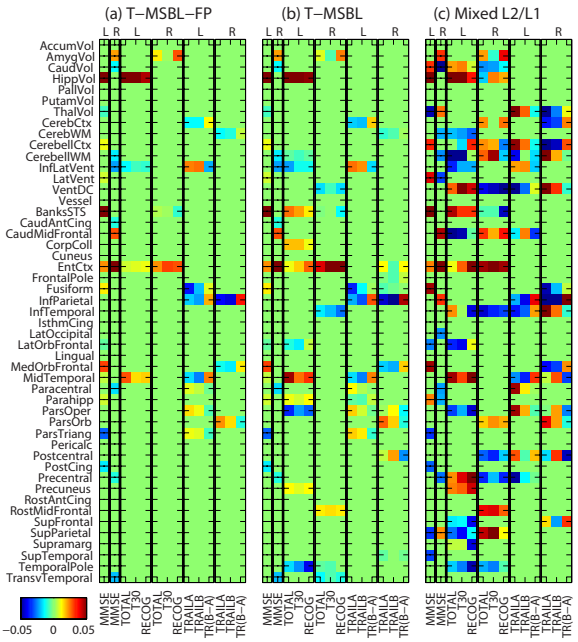

Figure 1. Heat maps of average regression weights of 5-fold cross-validation trials for (a) T-MSBL-FP, (b) T-MSBL, and (c) Mixed ℓ_2/ℓ_1. Each row corresponds to an MRI measure and each column to a cognitive score. Results for volume measures are shown in top 15 rows, and those for thickness measures in bottom 34 rows. Results for left (L) and right (R) hemispheres are shown in separate panels.

3.4. Biomarker Identification

Both T-MSBL-FP and T-MSBL are sparse models that are able to identify a compact set of relevant neuroimaging biomarkers and to explain the underlying brain structural changes related to cognitive status. Shown in Fig. 1 are

the heat maps of the regression weights (or coefficients) of the MRI measures for each cognitive score calculated by T-MSBL-FP, T-MSBL, and the Mixed ℓ_2/ℓ_1 Program. Blue indicates negative correlation, while red indicates positive correlation. The bigger the magnitude of an coefficient is, the more important its MRI measure is in predicting the corresponding cognitive score.

T-MSBL-FP clearly yielded a more sparse pattern than Mixed ℓ_2/ℓ_1 (Fig. 1), making the results easier to interpret. The pattern obtained by T-MSBL-FP was also more sparse and cleaner than those obtained by T-MSBL and other compared algorithms (not shown due to space constraint). Fig. 2 shows these regression weights mapped on the brain, where each row corresponds to one cognitive score and each column corresponds to a specific view of the brain.

The imaging biomarkers identified by T-MSBL-FP yielded promising patterns (Fig. 2) that are expected based on prior knowledge on neuroimaging and cognition. MMSE measures overall cognitive impairment; and thus its result includes important AD-relevant imaging markers such as hippocampal volume, amygdala volume, and entorhinal cortex thickness. RAVLT measures verbal learning memory; and thus its result includes regions relevant to learning and memory, such as hippocampus, entorhinal cortex, and middle temporal gyri. TRAILS measures a combination of visual, motor and executive functions; and thus its result includes regions in sensory-motor cortex (e.g., paracentral lobule), parietal lobe (relevant to visual processing), and frontal lobe (relevant to executive function).

All the above results have demonstrated that the proposed T-MSBL-FP method not only yields superior performance on prediction accuracy and computational time, but also is a powerful tool for discovering a small set of imaging biomarkers that predict cognitive performance. These results provide important information for understanding brain structural changes related to cognitive status and can potentially help characterize the progression of AD.

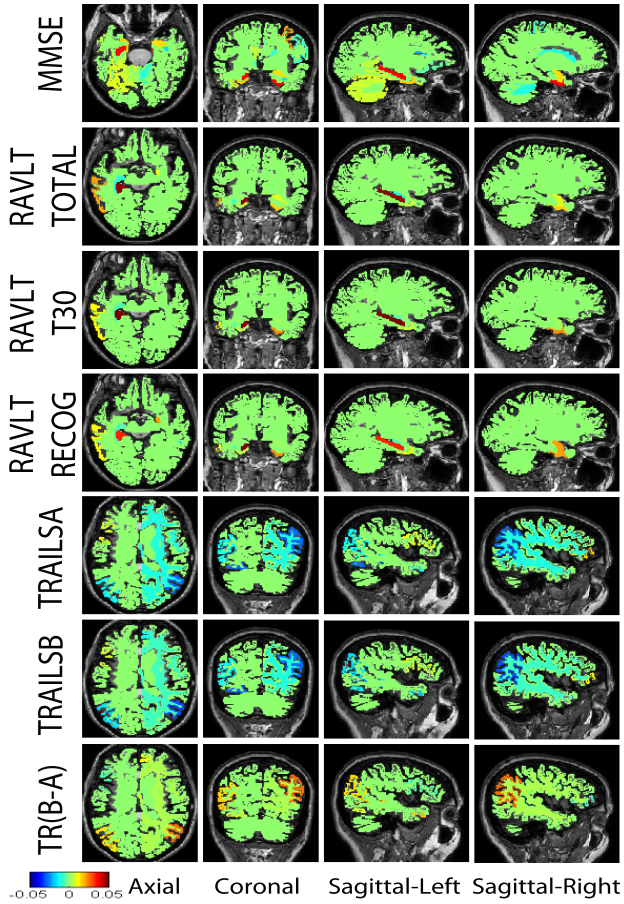

Figure 2. Regression weights (or coefficients) mapped onto brain: Each row corresponds to one cognitive score. Each column corresponds to a specific view of the brain

4. Conclusion

We have proposed a new sparse Bayesian multi-task learning algorithm, T-MSBL-FP, and demonstrated its effectiveness by applying it to the ADNI cohort for predicting cognitive outcomes from MRI scans. The proposed T-MSBL-FP method adaptively learns and exploits the correlation structure within each coefficient row in the multiple measurement vector model, which improves its performance. Its computational cost has been improved by one order of magnitude over its predecessor T-MSBL, making it possible to be used in applications with large-scale data sets. We have also revealed its connection to existing algorithms such as those based on $\ell_{2,1}$-norm and kernel regularization, which demonstrates that our algorithm can be viewed as an iterative reweighted $\ell_{2,1}$ algorithm using a data-adaptive kernel, providing motivation to design new algorithms.

In its application to the ADNI cohort, compared to multiple state-of-the-art algorithms, T-MSBL-FP not only demonstrated superior prediction performances over the competing methods, but also identified compact sets of cognition-relevant imaging biomarkers. These imaging biomarkers can predict multiple cognitive scores simultaneously and have a potential to play an important role in determining cognitive functions and characterizing AD progression. The identified biomarkers are consistent with the prior knowledge in existing literatures. All the results have clearly demonstrated the effectiveness of T-MSBL-FP. Potential future directions include (1) extension of T-MSBL-FP to multi-model imaging data (e.g. PET, fMRI) to predict cognitive performance, (2) extension of T-MSBL-FP to exploit more complex correlation structure inherent in data and among data sets, and (3) improving the $\ell_{q,1}$-norm based algorithms such that they can also exploit correlation structure while still maintaining their fast speed.

Acknowledgements

This research was supported by NSF IIS-1117335, CCF-0830612, NIH UL1 RR025761, U01 AG024904, NIA RC2 AG036535, NIA R01 AG19771, NIA P30 AG10133-18S1.

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; Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Amorfix Life Sciences Ltd.; AstraZeneca; Bayer HealthCare; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Disease Cooperative Study at the University of California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles. This research was also supported by NIH grants P30 AG010129, K01 AG030514, and the Dana Foundation.

References

[1] P. S. Aisen, R. C. Petersen, et al. Clinical core of the alzheimer's disease neuroimaging initiative: progress and plans. *Alzheimers Dement*, 6(3):239–46, 2010. 5

[2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004. 4, 5

[3] E. J. Candes et al. Enhancing sparsity by reweighted ℓ_1 minimization. *J Fourier Anal Appl*, 14:877–905, 2008. 5

[4] J. Chen and X. Huo. Theoretical results on sparse representations of multiple-measurement vectors. *IEEE Trans. on Signal Processing*, 54(12):4634 –4643, dec. 2006. 2

[5] S. F. Cotter, B. D. Rao, et al. Sparse solutions to linear inverse problems with multiple measurement vectors. *IEEE Trans. Signal Process.*, 53(7):2477–2488, 2005. 1, 2, 5, 6

[6] A. Dale, B. Fischl, and M. Sereno. Cortical surface-based analysis. i. segmentation and surface reconstruction. *Neuroimage*, 9(2):179–94, 1999. 5

[7] Y. Eldar and M. Mishali. Robust recovery of signals from a structured union of subspaces. *IEEE Transactions on Information Theory*, 55(11):5302–5316, 2009. 5, 6

[8] B. Fischl, M. Sereno, and A. Dale. Cortical surface-based analysis. ii: Inflation, flattening, and a surface-based coordinate system. *Neuroimage*, 9(2):195–207, 1999. 5

[9] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2 edition, 2009. 5

[10] S. Ji, D. Dunson, and L. Carin. Multi-task compressive sensing. *IEEE Trans. Signal Processing*, 57(1):92–106, 2009. 5

[11] D. J. C. MacKay. Bayesian interpolation. *Neural Computation*, 4(3):415–447, 1992. 3

[12] R. M. Neal. *Bayesian learning for neural networks*. Springer, 1996. 2

[13] F. Nie, H. Huang, X. Cai, and C. Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS 23*, pages 1813–1821, 2010. 2, 4

[14] C. M. Stonnington, C. Chu, et al. Predicting clinical scores from magnetic resonance scans in alzheimer's disease. *Neuroimage*, 51(4):1405–13, 2010. 1, 6

[15] M. E. Tipping. Sparse Bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001. 2, 3, 4

[16] J. Tropp, A. Gilbert, and M. Strauss. Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing*, 86(3):572–588, 2006. 5

[17] K. Walhovd, A. Fjell, et al. Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol Aging*, 31(7):1107–1121, 2010. 1

[18] H. Wang et al. A new sparse multi-task regression and feature selection method to identify brain imaging predictors for memory performance. *ICCV 2011*, pages 557–562. 1

[19] M. W. Weiner, P. S. Aisen, et al. The alzheimer's disease neuroimaging initiative: progress report and future plans. *Alzheimers Dement*, 6(3):202–11 e7, 2010. 1

[20] D. Wipf and S. Nagarajan. A new view of automatic relevance determination. *NIPS 20*, pages 1625–1632, 2008. 4

[21] Z. Xiang, Y. Xi, U. Hasson, and P. Ramadge. Boosting with spatial regularization. *NIPS 22*, pages 2107–2115, 2009. 4

[22] Y. Yang et al. Tag localization with spatial correlations and joint group sparsity. In *CVPR 2011*, pages 881–888. 4, 5

[23] M. Yuan et al. Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. B*, 68, 2006. 2

[24] D. Zhang and D. Shen. Multi-modal multi-task learning for joint prediction of multiple regression and classification variables in alzheimer's disease. *Neuroimage*, 2011. 1, 6

[25] Z. Zhang and B. D. Rao. Iterative reweighted algorithms for sparse signal recovery with temporally correlated source vectors. In *ICASSP 2011*, pages 3932 – 3935. 4

[26] Z. Zhang and B. D. Rao. Sparse signal recovery with temporally correlated source vectors using sparse Bayesian learning. *IEEE J. Sel. Topics Signal Process.*, 5(5):912–926, 2011. 2, 3, 4