JOINT ASSOCIATION DISCOVERY AND DIAGNOSIS OF ALZHEIMER'S DISEASE BY SUPERVISED HETEROGENEOUS MULTIVIEW LEARNING

SHANDIAN ZHE[1], ZENGLIN XU[1], YUAN QI[1,2], PENG YU[3], FOR THE ADNI*

[1]*Department of Computer Science, Purdue University,*
[2]*Department of Statistics, Purdue University,*
West Lafayette, IN 47907, USA
E-mail: {szhe,xu218,alanqi}@purdue.edu

[3]*Eli Lilly and Company, Indianapolis, IN 46225, USA*
E-mail: yu_peng_py@lilly.com

A key step for Alzheimer's disease (AD) study is to identify associations between genetic variations and intermediate phenotypes (*e.g.*, brain structures). At the same time, it is crucial to develop a noninvasive means for AD diagnosis. Although these two tasks—association discovery and disease diagnosis—have been treated separately by a variety of approaches, they are tightly coupled due to their common biological basis. We hypothesize that the two tasks can potentially benefit each other by a joint analysis, because (i) the association study discovers correlated biomarkers from different data sources, which may help improve diagnosis accuracy, and (ii) the disease status may help identify *disease-sensitive* associations between genetic variations and MRI features. Based on this hypothesis, we present a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels based on Gaussian process ordinal regression; in return, the disease status is used to guide the discovery of relationships between the data sources. The sparse projection matrices not only reveal the associations but also select groups of biomarkers related to AD. To learn the model from data, we develop an efficient variational expectation maximization algorithm. Simulation results demonstrate that our approach achieves higher accuracy in both predicting ordinal labels and discovering associations between data sources than alternative methods. We apply our approach to an imaging genetics dataset of AD. Our joint analysis approach not only identifies meaningful and interesting associations between genetic variations, brain structures, and AD status, but also achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

Keywords: disease diagnosis, Alzheimer's disease, genetic variations, brain structures, multiview learning, ordinal regression.

1. Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder associated with aging. Although it accounts for 60-80% of age-related dementia cases, currently there is no cure for AD and its underlying mechanism remain elusive. To study AD mechanism, a crucial step is to identify associations between genetic variations and intermediate phenotypes (*e.g.*, endophenotypical traits). In other words, we want to discover cross linkages between genetic risk factors based on

*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

genomic data—such as single nucleotide polymorphisms (SNPs)—and indicative intermediate phenotypes—such as cortical thickness of different brain regions (based on magnetic resonance imaging (MRI)). This identification can help us locate a subset of polymorphisms which may have functional consequences on brain structures. Although GWAS studies have been applied to AD studies,[1,2] the association study between genetic variations and multiple intermediate phenotypes is still relatively scarce for AD. A similar task arises for expression quantitative trait locus (eQTL) analysis, where canonical correlation analysis (CCA) and its extensions[3–6] have been widely applied. Meanwhile, it has become increasingly important to develop a noninvasive means for AD diagnosis based on various biomarkers, including both genetic variations and MRI features. Because many of these biomarkers are irrelevant to the diagnosis, sparse models are needed to identify the relevant ones. For disease diagnosis, popular sparse models include lasso,[7] elastic net,[8] and automatic relevance determination.[9] Here we treat genotypes or intermediate phenotypes as biomarkers and the disease status as the response in a linear regression or classification setting. Non-zero regression or classification weights in our estimation indicate relevant biomarkers for the disease.[10,11]

Although these two tasks—association discovery and disease diagnosis—have been addressed separately in the previous works, they are closely related—due to the their common underlying biological basis—and can potentially benefit each other by a joint analysis. To harness the natural synergy between the two tasks, we propose a new Bayesian approach that integrates multiview learning for association discovery with sparse ordinal regression for disease diagnosis. In the new approach, genetic variations and phenotypical traits are generated from common *latent* features based on separate sparse projection matrices and the common latent features are used to predict the disease status based on Gaussian process ordinal regression (See Section 2). To enforce sparsity in projection matrices, we assign spike and slab priors[12] over them; these priors have been shown to be more effective than l_1 penalty to learn sparse projection matrices.[13,14] The sparse projection matrices not only reveal critical interactions between the different data sources but also identify *groups* of biomarkers in data relevant to disease status. Finding groups of biomarkers can avoid over-sparsification (*i.e.*, selecting one instead of multiple correlated features), thus boosting the accuracy for disease diagnosis. It can also help provide a better biological understanding because these groups may form biologically meaningful units (*e.g.*, pathways). Meanwhile, via its direct connection to the latent features, the disease status influences the estimation of the projection matrices. Hence we name this new method Supervised Heterogeneous Multiview Learning (SHML). In addition to enjoying the benefit of integrating the related tasks, two features of our model distinguish it from previous approaches:

- There is a severity order for AD, from being normal to mild cognitive impairment (MCI) and then to AD; and our ordinal regression component captures the AD severity order. Alternative sparse models, by contrast, use classification or regression likelihoods and do not consider the order of disease severity.
- The data are heterogeneous: SNPs values are discrete (or ordinal) and the imaging features are continuous. While popular CCA-type methods treat both of them as continuous data, our model captures the heterogeneous nature of the data.

To learn the model from data, we develop a variational Bayesian expectation maximization (VB-EM) approach (See Section 3). Maximizing this estimate enables us to automatically choose a suitable dimension for the latent features in a principled Bayesian framework.

In Section 4, we test our approach SHML on both synthetic and real datasets. On synthetic data, SHML achieves both higher estimation accuracy in recovering true associations between different views and higher prediction accuracy than alternative state-of-the-art methods. We then apply SHML to an AD study. SHML achieved highest prediction accuracy among all competing methods and yielded biologically meaningful relationships between genetic variations, brain atrophy, and the disease status.

2. Model

First, let us describe the data. We assume there are two heterogeneous data sources: one contains continuous data – for example, MRI features – and one discrete ordinal data – for instance, SNPs. Given data from n subjects, p continuous features and q discrete features, we denote the continuous data by a $p \times n$ matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, the discrete ordinal data by a $q \times n$ matrix $\mathbf{Z} = [\mathbf{z}_1, \ldots, \mathbf{z}_n]$, and the labels (*i.e.*, the disease status) by a $n \times 1$ vector $\mathbf{y} = [y_1, \ldots, y_n]^\top$. For the AD study, we let $y_i = 0, 1,$ and 2 if the i-th subject is in the normal, MCI or AD condition, respectively.

To link two data sources $\mathbf{X}$ and $\mathbf{Z}$ together, we introduce common latent features $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ and assume $\mathbf{X}$ and $\mathbf{Z}$ are generated from $\mathbf{U}$ by sparse projections. The common latent feature assumption is sensible for association studies because both SNPs and MRI features are biological measurements of the same subjects. Note that $\mathbf{u}_i$ is the latent feature for the i-th subject and its dimension k is estimated by evidence maximization. In a Bayesian framework, we give a Gaussian prior over $\mathbf{U}$, $p(\mathbf{U}) = \prod_i \mathcal{N}(\mathbf{u}_i|\mathbf{0}, \mathbf{I})$, and specify the rest of the model (see Figure 1) as follows:

Continuous data. Given $\mathbf{U}$, $\mathbf{X}$ is generated from

$$p(\mathbf{X}|\mathbf{U}, \mathbf{G}, \eta) = \prod_{i=1}^{n} \mathcal{N}(\mathbf{x}_i|\mathbf{G}\mathbf{u}_i, \eta^{-1}\mathbf{I})$$

where $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, ...\mathbf{g}_p]^\top$ is a $p \times k$ projection matrix, $\mathbf{I}$ is an identity matrix, and $\eta^{-1}\mathbf{I}$ is the precision matrix of the Gaussian distribution. For η, we assign an uninformative diffuse Gamma prior, $p(\eta|r_1, r_2) = \text{Gamma}(\eta|r_1, r_2)$ with $r_1 = r_2 = 10^{-3}$.

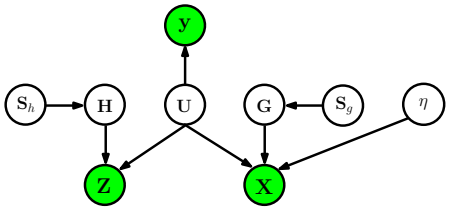


Fig. 1. The probabilistic graphical model of SHML, where $\mathbf{X}$ is the continuous view, $\mathbf{Z}$ is the ordinal view, and $\mathbf{y}$ are the labels.

Ordinal data. For an ordinal observation $z \in \{0, 1, \ldots, R-1\}$, its value is decided by which region an auxiliary variable c falls in $-\infty = b_0 < b_1 < \ldots < b_R = \infty$. If c falls in $[b_r, b_{r+1})$, z is set to be r. For the AD study, the SNPs $\mathbf{Z}$ take values in $\{0, 1, 2\}$ and therefore $R = 3$. Given a $q \times k$ projection matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ...\mathbf{h}_q]^\top$, the auxiliary variables $\mathbf{C} = \{c_{ij}\}$ and the ordinal data $\mathbf{Z}$ are generated from

$$p(\mathbf{Z}, \mathbf{C}|\mathbf{U}, \mathbf{H}) = \prod_{i=1}^{q} \prod_{j=1}^{n} \mathcal{N}(c_{ij}|\mathbf{h}_i^\top \mathbf{u}_j, 1) \sum_{r=0}^{2} \delta(z_{ij} = r)\delta(b_r \leq c_{ij} < b_{r+1})$$

where $\delta(a) = 1$ if a is true and $\delta(a) = 0$ otherwise, and $[b_0, \ldots, b_3]$ are set to $[-\infty, -1, 1, \infty]$.

Labels. For ordinal labels $\mathbf{y}$, we use a Gaussian process ordinal regression model[15] based the latent representation $\mathbf{U}$,

$$p(\mathbf{y}|\mathbf{U}) = \mathcal{N}(\mathbf{f}|\mathbf{0}, \mathbf{K}) \prod_{i=1}^{n} \sum_{r=0}^{2} \delta(y_i = r)\delta(b_r \leq f_i < b_{r+1})$$

where $[b_0, \ldots, b_3]$ are set to $[-\infty, -1, 1, \infty]$, and $K_{ij} = k(\mathbf{u}_i, \mathbf{u}_j)$ is the cross-covariance between $\mathbf{u}_i$ and $\mathbf{u}_j$. We can choose k from a rich family of kernel functions such as linear, polynomial, and Gaussian kernels to model relationships between the labels $\mathbf{y}$ and the latent features $\mathbf{U}$.

Note that the labels $\mathbf{y}$ are linked to the data $\mathbf{X}$ and $\mathbf{Z}$ via the latent features $\mathbf{U}$ and the projection matrices $\mathbf{H}$ and $\mathbf{G}$. Due to the sparsity in $\mathbf{H}$ and $\mathbf{G}$, only a few groups of variables in $\mathbf{X}$ and $\mathbf{Z}$ are selected to predict $\mathbf{y}$. Note that each of group is linked to a feature in $\mathbf{U}$.

Sparse Priors. Because we want to identify a few critical interactions between different data sources, we use spike and slab prior distributions[12] to sparsify the projection matrices $\mathbf{G}$ and $\mathbf{H}$. Specifically, we use a $p \times k$ matrix $\mathbf{S}_g$ to represent the selection of elements in $\mathbf{G}$: if $s_{ij} = 1$, g_{ij} is selected and follows a Gaussian prior distribution with variance σ_1^2; if $s_{ij} = 0$, g_{ij} is not selected and forced to almost zero (*i.e.*, sampled from a Gaussian with a very small variance σ_2^2). Specifically, we have the following prior over $\mathbf{G}$:

$$p(\mathbf{G}|\mathbf{S}_g, \mathbf{\Pi}_g) = \prod_{i=1}^{p} \prod_{j=1}^{k} {\pi_g^{ij}}^{s_g^{ij}} (1 - \pi_g^{ij})^{1 - s_g^{ij}} \left(s_g^{ij} \mathcal{N}(g_{ij}|0, \sigma_1^2) + (1 - s_g^{ij}) \mathcal{N}(g_{ij}|0, \sigma_2^2) \right)$$

where π_g^{ij} in $\mathbf{\Pi}_g$ is the probability of $s_g^{ij} = 1$, and $\sigma_1^2 \gg \sigma_2^2$ (in our experiment, we set $\sigma_1^2 = 1$ and $\sigma_2^2 = 10^{-6}$). Without any prior preference over the selecting probabilities, we assign uniform priors, $p(\mathbf{\Pi}_g) = 1$. Similarly, $\mathbf{H}$ is sampled from

$$p(\mathbf{H}|\mathbf{S}_h, \mathbf{\Pi}_h) = \prod_{i=1}^{q} \prod_{j=1}^{k} {\pi_h^{ij}}^{s_h^{ij}} (1 - \pi_h^{ij})^{1 - s_h^{ij}} \left(s_h^{ij} \mathcal{N}(h_{ij}|0, \sigma_1^2) + (1 - s_h^{ij}) \mathcal{N}(h_{ij}|0, \sigma_2^2) \right)$$

where $\mathbf{S}_h$ are binary selection variables and π_h^{ij} in $\mathbf{\Pi}_h$ is the probability of $s_h^{ij} = 1$. Again, we assign uninformative uniform priors over $\mathbf{\Pi}_h$: $p(\mathbf{\Pi}_h) = 1$.

Finally, the joint distribution of our model, SHML, is simply the product of all the prior distributions and the conditional density distributions.

3. Algorithm

3.1. *Estimating latent variables*

Given the model specified in the previous section, now we present an efficient, principled method to estimate the latent features $\mathbf{U}$, the projection matrices $\mathbf{H}$ and $\mathbf{G}$, the selection indicators $\mathbf{S}_g$ and $\mathbf{S}_h$, the selection probabilities $\mathbf{\Pi}_g$ and $\mathbf{\Pi}_h$, the variance η, the auxiliary variables $\mathbf{C}$ for generating ordinal data $\mathbf{Z}$, and the auxiliary variables $\mathbf{f}$ for generating the labels $\mathbf{y}$. In a Bayesian framework, this estimation task amounts to computing their posterior distributions. However, computing the exact posteriors turns out to be infeasible since we cannot calculate the normalization constant of the exact posterior distribution. Thus, we resort

to a variational Bayesian Expectation Maximization (VB-EM) approach. More specifically, in the E step, we approximate the posterior distributions of $\mathbf{H}, \mathbf{G}, \mathbf{S}_g, \mathbf{S}_h, \mathbf{\Pi}_g, \mathbf{\Pi}_h, \eta, \mathbf{C}$ and $\mathbf{f}$ by a factorized distribution $Q(\mathbf{H})Q(\mathbf{G})Q(\mathbf{S}_g)Q(\mathbf{S}_h)Q(\mathbf{\Pi}_g)Q(\mathbf{\Pi}_h)Q(\eta)Q(\mathbf{C})Q(\mathbf{f})$; and in the M step, based on the approximate distributions, we optimize the latent features $\mathbf{U}$.

To obtain the variational approximation, we minimize the Kullback-Leibler (KL) divergence between the approximate and the exact posteriors. To this end, we use coordinate descent; we update an approximate distribution, say, $Q(\mathbf{H})$, while fixing the other approximate distributions, and iteratively refine all the approximate distributions. The detailed updates are given in the following paragraphs.

3.1.1. *Updating variational distributions for continuous data*

For the continuous data $\mathbf{X}$, the approximate distributions of the projection matrix $\mathbf{G}$, the noise variance η, the selection indicators $\mathbf{S}_g$ and the selection probabilities $\mathbf{\Pi}_g$ are

$$Q(\mathbf{G}) = \prod_{i=1}^{p} \mathcal{N}(\mathbf{g}_i; \boldsymbol{\lambda}_i, \mathbf{\Omega}_i) \qquad Q(\eta) = \text{Gamma}(\eta|\tilde{r}_1, \tilde{r}_2), \tag{1}$$

$$Q(\mathbf{S}_g) = \prod_{i=1}^{p}\prod_{j=1}^{k} \beta_{ij}^{s_g^{ij}}(1-\beta_{ij})^{1-s_g^{ij}} \qquad Q(\mathbf{\Pi}_g) = \prod_{i=1}^{p}\prod_{j=1}^{k} \text{Beta}(\pi_g^{ij}|\tilde{l}_1^{ij}, \tilde{l}_2^{ij}). \tag{2}$$

The mean and covariance of $\mathbf{g}_i$ are calculated as $\mathbf{\Omega}_i = \big(\langle\eta\rangle\mathbf{U}\mathbf{U}^\top + \frac{1}{\sigma_1^2}\text{diag}(\langle\mathbf{s}_g^i\rangle) + \frac{1}{\sigma_2^2}\text{diag}(\mathbf{1} - \langle\mathbf{s}_g^i\rangle)\big)^{-1}$ and $\boldsymbol{\lambda}_i = \mathbf{\Omega}_i(\langle\eta\rangle\mathbf{U}\tilde{\mathbf{x}}_i)$, where $\langle\cdot\rangle$ means expectation over a distribution, $\tilde{\mathbf{x}}_i$ and $\mathbf{s}_g^i$ are the transpose of the i-th rows of $\mathbf{X}$ and $\mathbf{S}_g$, $\langle\mathbf{s}_g^i\rangle = [\beta_{i1}, \ldots, \beta_{ik}]^\top$, and $\langle g_{ij}^2\rangle$ is the j-th diagonal element in $\mathbf{\Omega}_i$. The parameters of the Gamma distribution $Q(\eta)$ are updated as $\tilde{r}_1 = r_1 + \frac{np}{2}$ and $\tilde{r}_2 = r_2 + \frac{1}{2}\text{tr}(\mathbf{X}\mathbf{X}^\top) - \text{tr}(\langle\mathbf{G}\rangle\mathbf{U}\mathbf{X}^\top) + \frac{1}{2}\text{tr}(\mathbf{U}\mathbf{U}^\top\langle\mathbf{G}^\top\mathbf{G}\rangle)$. The parameter β_{ij} in $Q(s_g^{ij})$ is calculated as $\beta_{ij} = 1/\big(1 + \exp(\langle\log(1-\pi_g^{ij})\rangle - \langle\log(\pi_g^{ij})\rangle + \frac{1}{2}\log(\frac{\sigma_1^2}{\sigma_2^2}) + \frac{1}{2}\langle g_{ij}^2\rangle(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))\big)$. The parameters of the Beta distribution $Q(\pi_g^{ij})$ is given by $\tilde{l}_1^{ij} = \beta_{ij} + 1$ and $\tilde{l}_2^{ij} = 2 - \beta_{ij}$.

The moments required in the above distributions are calculated as $\langle\eta\rangle = \frac{\tilde{r}_1}{\tilde{r}_2}$, $\langle\mathbf{G}\rangle = [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p]^\top$, $\langle\mathbf{G}^\top\mathbf{G}\rangle = \sum_{i=1}^{p}\mathbf{\Omega}_i + \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top$, $\langle\log(\pi_g^{ij})\rangle = \psi(\tilde{l}_1^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij})$ and $\langle\log(1-\pi_g^{ij})\rangle = \psi(\tilde{l}_2^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij})$, where $\psi(x) = \frac{\mathrm{d}}{\mathrm{dx}}\ln\Gamma(x)$.

3.1.2. *Updating variational distributions for ordinal data*

For the ordinal data $\mathbf{Z}$, we update the approximate distributions of the projection matrix $\mathbf{H}$, the auxiliary variables $\mathbf{C}$, the sparse selection indicators $\mathbf{S}_h$ and the selection probabilities $\mathbf{\Pi}_h$. Specifically, the variational distributions of $\mathbf{C}$, $\mathbf{H}$, $\mathbf{S}_h$ and $\mathbf{\Pi}_h$ are

$$Q(\mathbf{C}) \propto \prod_{i=1}^{q}\prod_{j=1}^{k} \delta(b_{z_{ij}} \le c_{ij} < b_{z_{ij}+1})\mathcal{N}(c_{ij}|\bar{c}_{ij}, 1) \qquad Q(\mathbf{H}) = \prod_{i=1}^{q}\mathcal{N}(\mathbf{h}_i; \boldsymbol{\gamma}_i, \mathbf{\Lambda}_i), \tag{3}$$

$$Q(\mathbf{S}_h) = \prod_{i=1}^{q}\prod_{j=1}^{k} \alpha_{ij}^{s_h^{ij}}(1-\alpha_{ij})^{1-s_h^{ij}} \qquad Q(\mathbf{\Pi}_h) = \prod_{i=1}^{q}\prod_{j=1}^{k} \text{Beta}(\pi_{\text{h}}^{\text{ij}}|\tilde{\text{d}}_1^{\text{ij}}, \tilde{\text{d}}_2^{\text{ij}}), \tag{4}$$

where $\bar{c}_{ij} = \boldsymbol{\gamma}_i^\top\mathbf{u}_j$, $\mathbf{\Lambda}_i = \big(\mathbf{U}\mathbf{U}^\top + \frac{1}{\sigma_1^2}\text{diag}(\langle\mathbf{s}_h^i\rangle) + \frac{1}{\sigma_2^2}\text{diag}(\langle\mathbf{1} - \mathbf{s}_h^i\rangle)\big)^{-1}$, $\boldsymbol{\gamma}_i = \mathbf{\Lambda}_i(\mathbf{U}\langle\tilde{\mathbf{c}}_i\rangle)$ where $\tilde{\mathbf{c}}_i$ is the transpose of the i-th row of $\mathbf{C}$, $\alpha_{ij} = 1/\big(1 + \exp(\langle\log(1-\pi_h^{ij})\rangle - \langle\log(\pi_h^{ij})\rangle + \frac{1}{2}\log(\frac{\sigma_1^2}{\sigma_2^2}) +$

$\frac{1}{2}\langle h_{ij}^2\rangle(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}))\big)$, $\tilde{d}_1^{ij}=\alpha_{ij}+1$, $\tilde{d}_2^{ij}=2-\alpha_{ij}$, $\langle \mathbf{s}_h^i\rangle=[\alpha_{i1},\ldots,\alpha_{ik}]^\top$, and $\langle h_{ij}^2\rangle$ is the j-th diagonal element in $\mathbf{\Lambda}_i$.

The required moments for updating the above distributions can be calculated as $\langle\log(\pi_h^{ij})\rangle=\psi(\tilde{d}_1^{ij})-\psi(\tilde{d}_1^{ij}+\tilde{d}_2^{ij})$, $\langle\log(1-\pi_h^{ij})\rangle=\psi(\tilde{d}_2^{ij})-\psi(\tilde{d}_1^{ij}+\tilde{d}_2^{ij})$, $\langle\tilde{c}_i\rangle=[\langle c_{i1}\rangle,\ldots,\langle c_{in}\rangle]^\top$ and $\langle c_{ij}\rangle=\bar{c}_{ij}-\big(\mathcal{N}(b_{z_{ij}+1}|\bar{c}_{ij},1)-\mathcal{N}(b_{z_{ij}}|\bar{c}_{ij},1)\big)/\big(\Phi(b_{z_{ij}+1}-\bar{c}_{ij})-\Phi(b_{z_{ij}}-\bar{c}_{ij})\big)$, where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution. Note that in Equation (3), $Q(\mathbf{C})$ is the product of truncated Gaussian distributions and the truncation is controlled by the observed ordinal data $\mathbf{Z}$.

3.1.3. *Updating variational distributions for labels*

We update the variational distribution of the auxiliary variables $\mathbf{f}$ as follows:

$$Q(\mathbf{f})\propto\prod_{i=1}^{n}\delta(b_{y_i}\le f_i<b_{y_i+1})\mathcal{N}(f_i|\bar{f}_i,\sigma_{f_i}^2) \tag{5}$$

where $\bar{f}_i=\mathbf{K}_{i,\neg i}\mathbf{K}_{\neg i,\neg i}^{-1}\langle\mathbf{f}_{\neg i}\rangle$ and $\sigma_{f_i}^2=\mathbf{K}_{i,i}-\mathbf{K}_{i,\neg i}\mathbf{K}_{\neg i,\neg i}^{-1}\mathbf{K}_{\neg i,i}$. $\mathbf{K}_{i,\neg i}$ is the covariance between $\mathbf{u}_i$ and $\mathbf{U}_{\neg i}$, $\mathbf{K}_{\neg i,\neg i}$ is the covariance on $\mathbf{U}_{\neg i}$ $(\mathbf{U}_{\neg i}=[\mathbf{u}_1,\cdots\mathbf{u}_{i-1},\mathbf{u}_{i+1},\cdots\mathbf{u}_n])$, $\langle\mathbf{f}_{\neg i}\rangle=[\langle f_1\rangle,\cdots,\langle f_{i-1}\rangle,\langle f_{i+1}\rangle,\cdots,\langle f_n\rangle]^\top$, and each $\langle f_i\rangle$ is $\langle f_i\rangle=\bar{f}_i-\sigma_{f_i}^2\cdot\big(\mathcal{N}(b_{y_i+1}|\bar{f}_i,\sigma_{f_i}^2)-\mathcal{N}(b_{y_i}|\bar{f}_i,\sigma_{f_i}^2)\big)/\big(\Phi(\frac{b_{y_i+1}-\bar{f}_i}{\sigma_{f_i}})-\Phi(\frac{b_{y_i}-\bar{f}_i}{\sigma_{f_i}})\big)$. Note that $Q(\mathbf{f})$ is also the product of truncated Gaussian distributions and the truncated region is decided by the ordinal label $\mathbf{y}$. In this way, the supervised information from $\mathbf{y}$ is incorporated into estimation of $\mathbf{f}$ and then estimation of the other quantities by the recursive updates.

3.1.4. *Optimizing the latent representation* $\mathbf{U}$

After the expectations of the other variables are calculated, we optimize $\mathbf{U}$ by maximizing the following variational lower bound

$$\begin{aligned}F(\mathbf{U})=&-\frac{1}{2}\mathrm{tr}(\mathbf{U}\mathbf{U}^\top)+\langle\eta\rangle\mathrm{tr}(\mathrm{X}^\top\langle\mathbf{G}\rangle\mathbf{U})-\frac{1}{2}\mathrm{tr}(\langle\mathbf{H}^\top\mathbf{H}\rangle\mathbf{U}\mathbf{U}^\top)-\frac{1}{2}\log|\mathbf{K}|-\frac{1}{2}\mathrm{tr}(\langle\mathbf{f}\mathbf{f}^\top\rangle\mathbf{K}^{-1})\\&-\frac{\langle\eta\rangle}{2}\mathrm{tr}(\langle\mathbf{G}^\top\mathbf{G}\rangle\mathbf{U}\mathbf{U}^\top)+\mathrm{tr}(\langle\mathbf{C}\rangle^\top\langle\mathbf{H}\rangle\mathbf{U})+\text{constant},\end{aligned} \tag{6}$$

where $\langle\mathbf{H}\rangle=[\mathbf{h}_1,\ldots,\mathbf{h}_q]^\top$, $\langle\mathbf{H}^\top\mathbf{H}\rangle=\sum_{i=1}^p\mathbf{\Lambda}_i+\boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top$, $\langle\mathbf{f}\mathbf{f}^\top\rangle=\langle\mathbf{f}\rangle\langle\mathbf{f}\rangle^\top-\mathrm{diag}(\langle\mathbf{f}\rangle^2)+\mathrm{diag}(\langle\mathbf{f}^2\rangle)$, $\langle\mathrm{f_i^2}\rangle=\langle\mathrm{f_i}\rangle^2+\sigma_{\mathrm{f_i}}^2+\sigma_{\mathrm{f_i}}^2\cdot\big((\mathrm{b_{y_i}}-\langle\mathrm{f_i}\rangle)\mathcal{N}(\mathrm{b_{y_i}}|\langle\mathrm{f_i}\rangle,\sigma_{\mathrm{f_i}}^2)\big)/\big(\Phi(\frac{\mathrm{b_{y_i+1}}-\langle\mathrm{f_i}\rangle}{\sigma_{\mathrm{f_i}}})-\Phi(\frac{\mathrm{b_{y_i}}-\langle\mathrm{f_i}\rangle}{\sigma_{\mathrm{f_i}}})\big)-\sigma_{\mathrm{f_i}}^2\cdot\big((\mathrm{b_{y_i+1}}-\langle\mathrm{f_i}\rangle)\mathcal{N}(\mathrm{b_{y_i+1}}|\langle\mathrm{f_i}\rangle,\sigma_{\mathrm{f_i}}^2)\big)/\big(\Phi(\frac{\mathrm{b_{y_i+1}}-\langle\mathrm{f_i}\rangle}{\sigma_{\mathrm{f_i}}})-\Phi(\frac{\mathrm{b_{y_i}}-\langle\mathrm{f_i}\rangle}{\sigma_{\mathrm{f_i}}})\big)$, and the constant means a value independent of $\mathbf{U}$ so that it is irrelevant for optimizing $\mathbf{U}$. Note that we can optimize the dimension k by maximizing the full variational lower bound of our model, which involves other quantities as well, such as $\langle\mathbf{H}\rangle$ and $\langle\mathbf{G}\rangle$. To save space, we do not present the long equation for the full lower bound (which can be easily derived based on what we have presented). We use the L-BFGS algorithm to maximize the cost function F over $\mathbf{U}$. The gradient of $\mathbf{U}$ is given by

$$\frac{\partial F}{\partial\mathbf{U}}=\langle\eta\rangle\langle\mathbf{G}\rangle^\top\mathbf{X}+\langle\mathbf{H}\rangle^\top\langle\mathbf{C}\rangle-\big(\mathbf{I}+\langle\eta\rangle\langle\mathbf{G}^\top\mathbf{G}\rangle+\langle\mathbf{H}^\top\mathbf{H}\rangle\big)\mathbf{U}-\frac{1}{2}\big(\mathbf{K}^{-1}-\frac{1}{2}\mathbf{K}^{-1}\langle\mathbf{f}\mathbf{f}^\top\rangle\mathbf{K}^{-1}\big)\frac{\partial\mathbf{K}}{\partial\mathbf{U}}. \tag{7}$$

Note that $\frac{\partial\mathbf{K}}{\partial\mathbf{U}}$ depends on the form of the kernel function $k(\mathbf{u}_i,\mathbf{u}_j)$.

Computational complexity. Based on the previous equations, we can show that the total computational complexity of our algorithm is $O(\max(n^3, (p+q)nk^2))$—it is either cubic in the number of samples n or linear in the number of the features.

3.2. *Predicting disease status*

Let us denote the training data as $\mathcal{D}_{\text{train}} = \{\mathbf{X}_{\text{train}}, \mathbf{Z}_{\text{train}}, \mathbf{y}_{\text{train}}\}$ and the test data as $\mathcal{D}_{\text{test}} = \{\mathbf{X}_{\text{test}}, \mathbf{Z}_{\text{test}}\}$. To obtain the latent representation $\mathbf{U}_{\text{train}}$ and $\mathbf{U}_{\text{test}}$ for prediction, we carry out variational EM simultaneously on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. The benefit is that the variational EM learning procedure can utilize both the training and test data. Note that there are no updates for ordinal label part on $\mathbf{D}_{\text{test}}$ and the terms regarding ordinal labels should also be removed from Equation (6) and (7). After both $\mathbf{U}_{\text{test}}$ and $\mathbf{U}_{\text{train}}$ are obtained from the M-step, we predict the labels for test data as follows:

$$\mathbf{f}_{\text{test}} = \mathbf{K}\big(\mathbf{U}_{\text{test}}, \mathbf{U}_{\text{train}}\big)\mathbf{K}^{-1}\big(\mathbf{U}_{\text{train}}, \mathbf{U}_{\text{train}}\big)\langle \mathbf{f}_{\text{train}} \rangle \qquad y^i_{\text{test}} = \sum_{r=0}^{R-1} r \cdot \delta(b_r \leq f^i_{\text{test}} < b_{r+1}),$$

where y^i_{test} is the prediction for i-th test sample.

4. Experiments

4.1. *Simulation Study*

We first design a simulation study to examine SHML in terms of (i) estimation accuracy on finding associations between the two views and (ii) prediction accuracy on the ordinal labels.

Simulation data. To generate the ground truth, we set $n = 200$ (200 instances), $p = q = 40$, and $k = 5$. We designed $\mathbf{G}$, the 40×5 projection matrix for the continuous data $\mathbf{X}$, to be a block diagonal matrix; each column of $\mathbf{G}$ had 8 elements being ones and the rest of them were zeros, ensuring each row with only one nonzero element. We designed $\mathbf{H}$, the 40×5 projection matrix for the ordinal data $\mathbf{Z}$, to be a block diagonal matrix; each of the first four columns of $\mathbf{H}$ had 10 elements being ones and the rest of them were zeros, and the fifth column contained only zeros. We randomly generated the latent representations $\mathbf{U} \in \mathbb{R}^{k \times n}$ with each column $\mathbf{u}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$. To generate $\mathbf{Z}$, we first sampled the auxiliary variables $\mathbf{C}$ with each column $\mathbf{c}_i \sim \mathcal{N}(\mathbf{H}\mathbf{u}_i, 1)$, and then decided the value of each element z_{ij} by the region c_{ij} fell in—in other words, $z_{ij} = \sum_{r=0}^{2} r\delta(b_r \leq c_{ij} < b_{r+1})$. Similarly, to generate $\mathbf{y}$, we sampled the auxiliary variables $\mathbf{f}$ from $\mathcal{N}(0, \mathbf{U}^\top\mathbf{U} + \mathbf{I})$ and then each y_i was generated by $p(y_i|f_i) = \delta(y_i = 0)\delta(f_i \leq 0) + \delta(y_i = 1)\delta(f_i > 0)$.

Comparative methods. We compared SHML with several state-of-the-art methods including (1) CCA,[4] which finds the projection directions that maximize the correlation between two views, (2) sparse CCA,[6,18] where sparse priors are put on the CCA directions, and (3) multiple-response regression with lasso (MRLasso)[19] where each column of the second view ($\mathbf{Z}$) is regarded as the output of the first view ($\mathbf{X}$). We did not include results from the sparse probabilistic projection approach[20] because it performed unstably in our experiments. Regarding the software implementation, we used the built-in Matlab Matlab routine for CCA and the code by[18] for sparse CCA. We implemented MRLasso based on the Glmnet package (`cran.r-project.org/web/packages/glmnet/index.html`).

To test prediction accuracy, we compared our method with the following ordinal or multinomial regression methods: (1) lasso for multinomial regression,[7] (2) elastic net for multinomial regression,[8] (3) sparse ordinal regression with the splike and slab prior, (4) CCA + lasso, for which we first ran CCA to obtain the latent features $\mathbf{H}$ and then applied lasso to predict $\mathbf{y}$, (5) CCA + elastic net, for which we first ran CCA to obtain the projection matrices and then applied elastic net on the projected data, (6) Gaussian Process Ordinal Regression (GPOR),[15] and (7) Laplacian Support Vector Machine (LapSVM),[21] a semi-supervised SVM classification method. We used the published code for lasso, elastic net, GPOR and LapSVM. For all the methods, we used 10-fold cross validation on the training data for each run to choose the kernel form (Gaussian or linear or Polynomials) and its parameters (the kernel width or polynomial orders) for SHML, GPOR, and LapSVM.

Because alternative methods cannot learn the dimension automatically for simple comparison, we provided the dimension of the latent representation to all the methods we tested in our simulations. We partitioned the data into 10 subsets and used 9 of them for training and 1 subset for testing; we repeated the procedure 10 times to generate the averaged test results.

Results.To estimate linkage (*i.e.*, interactions) between $\mathbf{X}$ and $\mathbf{Z}$, we calculated the cross covariance matrix $\mathbf{GH}^\top$. We then computed the precision and the recall based on the ground truth. The precision-recall curves are shown in Figure 2. Clearly, our method successfully recovered almost all the links and significantly outperformed all the competing methods. This improvement may come from i) the use of the spike and slab priors, which not only remove irrelevant elements in the projection matrices but also avoid over-penalizing the active association structures (the Laplace prior used in sparse CCA does over penalize the relevant ones) and ii) more importantly, the supervision from the labels $\mathbf{y}$, which is probably the biggest difference between ours and the other methods for the association study. The prediction accuracies on unknown $\mathbf{y}$ and their standard errors are shown in Figure 3a and the AUC and their standard errors are shown in Figure 3b. Our proposed SHML model achieves significant improvement over all the other methods. It reduces the prediction error of elastic net (which ranks the second best) by 25%, and reduces the error of LapSVM by 48%.

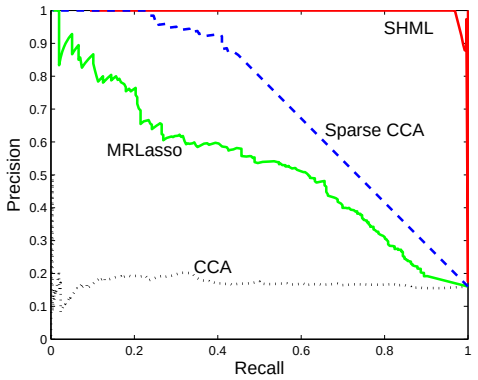
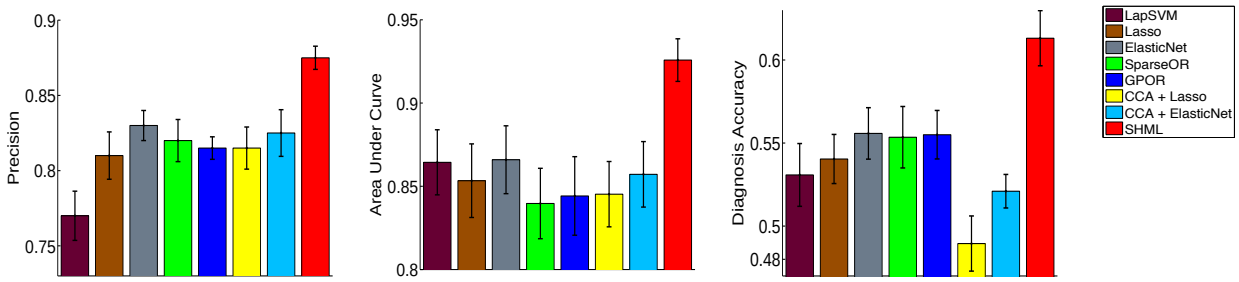


Fig. 2. The precision-recall curves for association discovery.

4.2. *AD Study*

We conducted joint association analysis and AD diagnosis based on the Alzheimer's Disease Neuroimaging Initiative 1 (ADNI 1) dataset. The ADNI study is a longitudinal multisite observational study of elderly individuals with normal cognition, mild cognitive impairment, or AD. Specifically, we used SHML to study the associations of genotypes and brain atrophy measured by MRI and to predict the disease status (normal vs MCI vs AD). Note that the labels are ordinal since the three states represent increasing severity levels of AD.



(a) Precision on simulation (b) AUC on simulation (c) Precision on ADNI dataset

Fig. 3. The prediction results on simulated and real datasets. The results are averaged over 10 runs. The error bars represent standard errors. For the real ADNI dataset, we predict the ordinal disease status, Normal, MCI and AD.

Genetic and phenotypic data used in this study were obtained from the ADNI database (http://www.loni.ucla.edu/ADNI). Genomic DNA samples of 818 ADNI 1 subjects were analyzed on the Human610-Quad BeadChip according to the manufacturer's protocols. After quality control, a list of 512,788 SNPs was used in an initial GWAS analysis associating them with the disease trait (AD vs. normal subjects). As a result, the top 1000 SNPs were preselected for analysis in this study. For structural MRI, we used image analysis results from UCSF based on the Freesurfer software package (http://surfer.nmr.mgh.harvard.edu); the resulting imaging data includes volumetric, cortical thickness and surface area measurements for a variety cortical and subcortical regions. After removing missing data, the final dataset consists of 618 subjects (183 normal, 308 MCI and 134 AD), and 924 SNPs and 328 MRI features measuring the brain atrophies for each subject at baseline.

We compared SHML with the alternative methods on accuracy of predicting whether a subject is in the normal or MCI or AD condition. We randomly split the dataset into 556 training and 62 test samples 10 times and ran all the competing methods on each partition. We used the 10-fold cross validation for each run to tune free parameters on the training data. In SHML, in order to determine k, the dimension of $\mathbf{U}$, we computed the variational lower bound as an approximation to the model marginal likelihood with various k values $\{10, 20, 40, 60\}$. We chose the value with the largest approximate evidence, which led to $k = 20$ (see Figure 4). Our experiments confirmed that, with $k = 20$, SHML achieved highest prediction accuracy, demonstrating the benefit of evidence maximization.

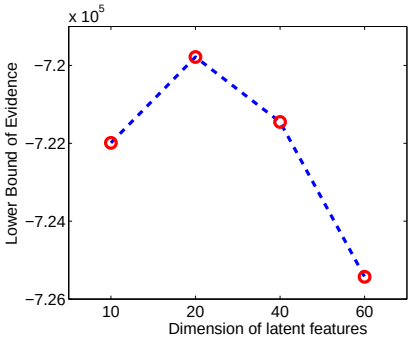


Fig. 4. The variational lower bound of the marginal likelihood (i.e., evidence).

The accuracies for predicting unknown labels $\mathbf{y}$ and their standard errors are shown in Figure 3c. Our method achieved the highest prediction accuracy, higher than that of the second best method, GP ordinal Regression, by 10% and than that of the worst method, CCA+lasso, by 22%.

We also examined the strongest associations discovered by SHML based on the whole dataset. First of all, the ranking of MRI features in terms of their prediction power of different disease stages (normal, MCI and AD) demonstrates that most of the top ranked

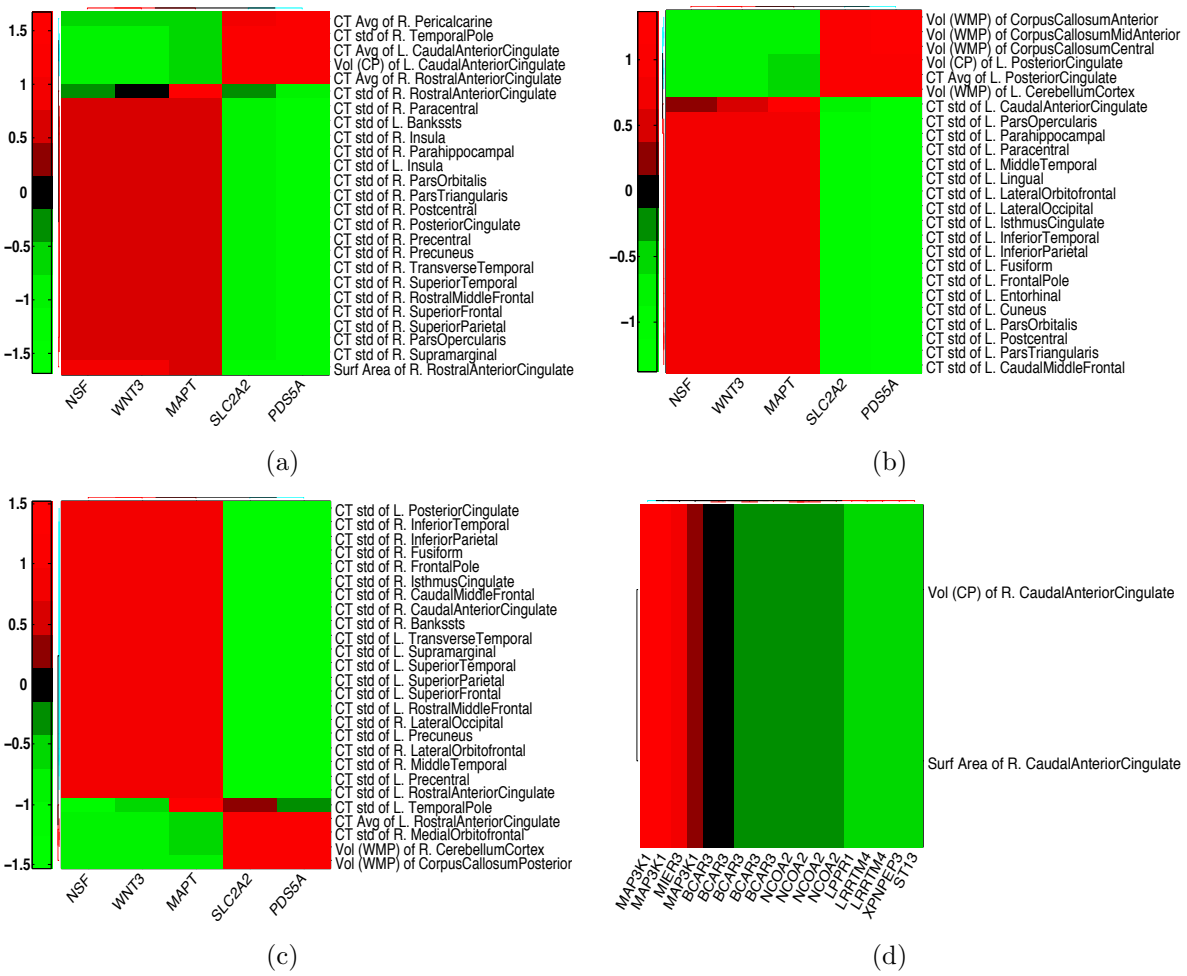


Fig. 5. The estimated associations between MRI features and SNPs. In each sub-figure, the MRI features are listed on the right and the SNP names are given at the bottom.

features are the cortical thickness measurements, followed by the volume of white matter, volume of gray matter in cortical regions, and the cortical surface area measurements. These results are consistent with the literature for demonstrating that the cortical thickness measurement is potentially a more sensitive measurement of the brain atrophy for Alzheimer's dementia.[22,23] Particularly, thickness measurements of frontal lobe, middle temporal lobe, and precuneus were found to be most predictive compared with other brain regions. These findings are consistent with their atrophy pattern and prediction power of AD found in the literature[23–27]. We also found that measurements of the same structure on the left and right hemisphere have similar weights (See Table 1); this is again consistent with the related literature—no asymmetrical relationship has been found for the brain regions involved in AD.[28]

Table 1. The weights of the average cortical thickness of ROI on the left and right hemispheres.

ROI	weight	
	left	right
Superior Frontal	1.37	1.35
Middle Temporal	1.33	1.37
Precuneus	1.33	1.36
Inferior Parietal	1.29	1.34
Inferior Temporal	1.32	1.29
Caudal Middle Frontal	1.32	1.31
Rostral Middle Frontal	1.31	1.30

Secondly, the analysis of associating genotypes to AD also generated interesting results. Similar to the MRI features, SNPs that are in the vicinity of each other are selected together due to the *group-selection* characteristics of our algorithm. The top ranked SNPs are associated with a few genes including PSMC1P12 (proteasome 26S subunit, ATPase), NCOA2 (The nuclear receptor coactivator 2), and WDR52 (WD repeat domain 52). These genes have been associated with diseases such as breast neoplasms, carcinoma, and endometrial neoplasms.[29]

At last, biclustering of the genotype-MRI association, as shown in Figure 5, revealed interesting patterns in terms of the relationship between genetic variations and brain atrophy in association with AD. For example, the highest ranked association was found between genes such as MAP3K1 (mitogen-activated protein kinase kinase kinase 1) and MIER3 (mesoderm induction early response 1, family member 3) with the caudate anterior cingulate cortex. MAP3K1 and MIER3 genes are associated with biological process such as apoptosis, cell cycle, chromatin binding and DNA binding (https://portal.genego.com/), and cingulate cortex has been shown to be severely affected by AD[30]. The strong association discovered in this work might indicate potential genetic effect in the atrophy pattern observed in this cingulate subregion. Additionally, SNPs in MAPT (microtubule-associated protein tau) gene were also found to have association with brain atrophy in a variety of cortical regions including frontal, cingulate and temperate lobes. The hyperphosphorylation of tau protein, which is a product of MAPT, can result in the self-assembly of tangles that are involved in the pathogenesis of AD. Therefore, the genetic variation of MAPT has been associated with increased risk of AD[31–35]. The association between MATP gene and brain atrophies found in this analysis is consistent with the gray matter loss observed in MATP genetic variant carrier in recent studies.[36]

In summary, SHML discovered the synergistic predictive relationships between brain atrophy, genetic variations and the disease status, and achieved higher prediction accuracy than the alternative methods.

5. Conclusions

We have presented, SHML, a new Bayesian supervised multiview learning algorithm for AD study. By integrating association discovery with disease diagnosis, it improves performance for both tasks. Although we have focused on the AD study in this paper, we expect that SHML can be applied to a wide range of applications in biomedical research—for example, eQTL analysis supervised by additional labeling information. As to the future work, we plan to incorporate additional biological or side information into our model to improve its quality. In particular, linkage disequilibrium structures encode important correlation information between SNPs. Our current model uses *independent*, uniform priors over the selection probabilities of SNPs, which ignore the correlation between SNPs (note that the posterior distribution of the model does capture some correlation between genetic variations based on the data likelihood). To overcome this limitation, we plan to use graph Laplacian matrices to encode linkage disequilibrium structures and use these matrices in our prior distributions. We have explored a similar strategy to incorporate biological pathway constraints for biomarker selection and obtained improved performance over the models that do not use the pathway information.[37] We expect a similar improvement can be obtained by incorporating LD structures into SHML.

Acknowledgments

This work was supported by NSF IIS-0916443, NSF IIS-1054903, and the Center for Science of Information (CSoI), an NSF Science and Technology Center, under grant agreement CCF-0939370. Data used in the work were obtained from the ADNI database. ADNI funding information is available at http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_DSP_Policy.pdf

References

1. D. Harold *et al.*, *Nat. Genet.* **41**, 1088 (2009).
2. M. Vounou *et al.*, *Neuroimage* **60**, 700 (2012).
3. H. Harold, *Biometrika* **28**, 321 (1936).
4. F. Bach and M. Jordan, *A probabilistic interpretation of canonical correlation analysis*, tech. rep., UC Berkeley (2005).
5. E. Parkhomenko, D. Tritchler and J. Beyene, *BMC Proc.* **1 Suppl 1**, p. S119 (2007).
6. M. Daniela and R. Tibshirani, *Stat Appl Genet Mol Biol.* **383** (2009).
7. R. Tibshirani, *Journal of the Royal Statistical Society, Series B* **58**, 267 (1994).
8. H. Zou and T. Hastie, *Journal of the Royal Statistical Society, Series B* **67**, 301 (2005).
9. D. MacKay, *Neural Computation* **4**, 415 (1991).
10. P. Yu, R. A. Dean *et al.*, *J. Alzheimers Dis.* **32**, 373 (2012).
11. L. Shen, Y. Qi *et al.*, *Med Image Comput Comput Assist Interv.* **13**, 611 (2010).
12. E. George and R. McCulloch, *Statistica Sinica* **7**, 339 (1997).
13. I. Goodfellow *et al.*, Large-scale feature learning with spike-and-slab sparse coding, in *ICML*, 2012
14. S. Mohamed *et al.*, Bayesian and L1 approaches for sparse unsupervised learning, in *ICML*, 2012.
15. W. Chu and Z. Ghahramani, *Journal of Machine Learning Research* **6**, 1019 (2005).
16. J. Zhou *et al.*, Modeling disease progression via fused sparse group lasso, in *KDD'12*, 2012.
17. L. Yuan *et al.*, Multi-source learning for joint analysis of incomplete multi-modality neuroimaging data, in *KDD'12*, 2012.
18. L. Sun *et al.*, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **33**, 194 (2011).
19. S. Kim, K. Sohn and E. Xing, *Bioinformaics* **25**, 204 (2009).
20. C. Archambeau and F. Bach, Sparse probabilistic projections, in *NIPS'09*, 2009.
21. S. Melacci and B. Mikhail, *Journal of Machine Learning Research* **12**, 1149 (2011).
22. J. Lerch, J. Pruessner, A. Zijdenbos *et al.*, *Neurobiol Aging* **1**, 23 (2008).
23. S. Teipel *et al.*, *Medical Clinics of North America* **97**, 399 (2013).
24. J. Whitwell, S. Przybelski, S. Weigand *et al.*, *Brain* **130**, 1777 (2007).
25. S. Risacher, A. Saykin, J. West, H. Firpi and B. McDonald, *Curr. Alzheimer Res.* **6**, 347 (2009).
26. S. Galluzzi, C. Geroldi *et al.*, *J. Neurol.* **14**, 2004 (2010).
27. J. Whitwell, H. Wiste *et al.*, *Arch. Neurol.* **69**, 614 (May 2012).
28. O. Y. Kusbeci *et al.*, *Dement Geriatr Cogn Disord* **28**, 1 (2009).
29. P. J. Stephens, P. S. Tarpey *et al.*, *Nature* **486**, 400 (Jun 2012).
30. B. F. Jones *et al.*, *Cereb. Cortex* **16**, 1701 (Dec 2006).
31. M. J. Bullido *et al.*, *Neurosci. Lett.* **278**, 49 (Jan 2000).
32. H. Tanahashi, T. Asada and T. Tabira, *Neuroreport* **15**, 175 (Jan 2004).
33. T. Feulner, S. Laws *et al.*, *Mol. Psychiatry* **15**, 756 (Jan 2010).
34. E. Di Maria, S. Cammarata *et al.*, *J. Alzheimers Dis.* **19**, 909 (2010).
35. L. Samaranch, S. Cervantes *et al.*, *J. Alzheimers Dis.* **22**, 1065 (2010).
36. J. Goñi *et al.*, *J. Alzheimers Dis.* **33**, 1009 (2013).
37. S. Zhe *et al.*, *Bioinformatics* **29**, 1987 (2013).