Sparse Bayesian Multiview Learning for Simultaneous Association Discovery and Diagnosis of Alzheimer's Disease

Shandian Zhe
Department of Computer Science
Purdue University

Zenglin Xu
School of Computer Science & Engineering,
Big Data Res. Center
University of Electronic Science
and Technology of China

Yuan Qi
Department of Computer Science
Purdue University

Peng Yu
Eli Lilly and Company

for the ADNI[*]

Abstract

In the analysis and diagnosis of many diseases, such as the Alzheimer's disease (AD), two important and related tasks are usually required: i) selecting genetic and phenotypical markers for diagnosis, and ii) identifying associations between genetic and phenotypical features. While previous studies treat these two tasks separately, they are tightly coupled due to the same underlying biological basis. To harness their potential benefits for each other, we propose a new sparse Bayesian approach to jointly carry out the two important and related tasks. In our approach, we extract common latent features from different data sources by sparse projection matrices and then use the latent features to predict disease severity levels; in return, the disease status can guide the learning of sparse projection matrices, which not only reveal interactions between data sources but also select groups of related biomarkers. In order to boost the learning of sparse projection matrices, we further incorporate graph Laplacian priors encoding the valuable linkage disequilibrium (LD) information. To efficiently estimate the model, we develop a variational inference algorithm. Analysis on an imaging genetics dataset for AD study shows that our model discovers biologically meaningful associations between single nucleotide polymorphisms (SNPs) and magnetic resonance imaging (MRI) features, and achieves significantly higher accuracy for predicting ordinal AD stages than competitive methods.

Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder (Khachaturian 1985). Given genetic variations, *e.g.*, single nucleotide polymorphisms (SNPs), and phenotypical traits, *e.g.*, magnetic resonance imaging (MRI), we want to develop noninvasive diagnosis methods for better preventative care; to understand the underlying AD pathology, we want to identify the disease related features and discover their associations.

Many approaches have been proposed to discover associations, such as canonical correlation analysis (CCA) and its extensions (Harold 1936; Bach and Jordan 2005a), or select features (or variables), such as lasso (Tibshirani 1994), elastic net (Zou and Hastie 2005), and Bayesian automatic relevance determination (MacKay 1991; Neal 1996). Despite their wide success in many applications, these approaches suffer several limitations: i) most association studies neglect the supervision from the disease status, while diseases as AD are a direct result of genetic variations and often highly correlated to clinical traits; ii) most feature selection approaches do not consider the disease severity order, while AD subjects have a natural severity order from being normal to mild cognitive impairment (MCI) and then from MCI to AD; iii) most previous approaches are not designed to handle heterogeneous data sources, *e.g.*, the SNPs values are discrete and ordinal while the imaging features are continuous; iv) most previous methods ignore the valuable prior knowledge such as Linkage Disequilibrium (LD) (Falconer and Mackay 1996) measuring the non-random association of alleles. To our knowledge, this structure has not been utilized for association discovery in medical study.

To address these limitations, we propose a new Bayesian model that unifies multiview learning with sparse ordinal regression for joint association study and disease diagnosis. Specifically, genetic variations and phenotypical traits are generated from common *latent* features based on separate sparse projection matrices and suitable link functions, and the latent features are used to predict the disease status. To encourage sparsity, we assign spike and slab priors (George and McCulloch 1997) over the projection matrices; we further employ an additional graph Laplacian prior encoding LD knowledge to boost the learning of the sparse projection matrix on the genetic variation view. The sparse projection matrices then not only reveal critical interactions between the data sources but also identify biomarkers relevant to the disease. Meanwhile, via their direct connection to the latent features, the disease status will influence the estimation of the projection matrices and guide the discovery of *disease-*

[*] 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

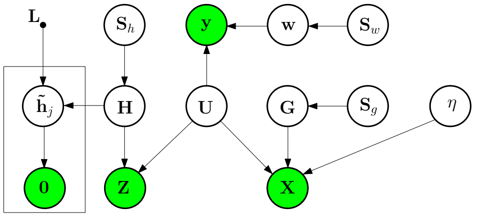


Figure 1: The graphical representation of our model, where $\mathbf{X}$ is the continuous view, $\mathbf{Z}$ is the ordinal view, $\mathbf{y}$ are the ordinal labels and $\mathbf{L}$ is the graph Laplacian generated from the LD structure.

sensitive data source associations. For efficient model estimation, we develop a variational inference approach, which iteratively minimizes the Kullback Leibler divergence between a tractable approximation and exact Bayesian posterior distributions. The results on Alzheimer's Disease Neuroimaging Initiative (ADNI) dataset show that our model achieves highest prediction accuracy among all the competing methods and finds biologically meaningful associations between SNPs and MRI features.

Methods

Notations and Assumptions

We assume there are two heterogeneous data sources: one contains continuous data – *e.g.*, MRI features – and the other contains discrete ordinal data – *e.g.*, SNPs. Note that we can easily generalize our model below to handle more views and other data types by adopting suitable link functions (*e.g.*, a Possion model for count data). Given data from n subjects, p continuous features and q discrete features, we denote the continuous data by a $p \times n$ matrix $\mathbf{X} = [\mathbf{x}_1, \dots, \mathbf{x}_n]$, the discrete ordinal data by a $q \times n$ matrix $\mathbf{Z} = [\mathbf{z}_1, \dots, \mathbf{z}_n]$ and the labels (i.e., the disease status) by a $n \times 1$ vector $\mathbf{y} = [y_1, \dots, y_n]^\top$. 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.

Model

To link the two data sources $\mathbf{X}$ and $\mathbf{Z}$ together, we introduce common latent features $\mathbf{U} = [\mathbf{u}_1, \dots, \mathbf{u}_n]$ and assume $\mathbf{X}$ and $\mathbf{Z}$ are generated from $\mathbf{U}$ by sparse projection. The common latent feature assumption is sensible because both SNPs and MRI features are biological measurements of the same subjects. Note that $\mathbf{u}_i$ is a k dimensional latent feature for the i-th subject. In a Bayesian framework, we assign 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 distribution. 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 η is a scalar. We assign an uninformative Gamma prior over η, $p(\eta|r_1, r_2) = \mathrm{Gamma}(\eta|r_1, r_2)$, where $r_1 = r_2 = 10^{-3}$.

Ordinal data distribution. For an ordinal variable $z \in \{0, 1, \dots, R-1\}$, we introduce an auxiliary variable c and a segmentation of the number axis, $-\infty = b_0 < b_1 < \dots < b_R = \infty$. We define that $z = r$ if and only if c falls in $[b_r, b_{r+1})$. Then given a $q \times k$ projection matrix $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ...\mathbf{h}_q]^\top$ and the auxiliary matrix $\mathbf{C} = \{c_{ij}\}$, the ordinal data $\mathbf{Z}$ are generated from

$$p(\mathbf{Z}, \mathbf{C}|\mathbf{U}, \mathbf{H}) = \prod_{i=1}^{q} \prod_{j=1}^{n} p(c_{ij}|\mathbf{h}_i, \mathbf{u}_j) p(z_{ij}|c_{ij})$$

where $p(z_{ij}|c_{ij}) = \sum_{r=0}^{R-1} \delta(z_{ij} = r)\delta(b_r \leq c_{ij} < b_{r+1})$ and $p(c_{ij}|\mathbf{h}_i, \mathbf{u}_j) = \mathcal{N}(c_{ij}|\mathbf{h}_i^\top \mathbf{u}_j, 1)$. Here $\delta(a) = 1$ if a is true and $\delta(a) = 0$ otherwise. In AD study, Z take values in $\{0, 1, 2\}$ and hence $R = 3$.

Label distribution. The disease status labels $\mathbf{y}$ are ordinal variables too. To generate $\mathbf{y}$, we use the ordinal regression model based the latent representation $\mathbf{U}$,

$$p(\mathbf{y}, \mathbf{f}|\mathbf{U}, \mathbf{w}) = p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{U}, \mathbf{w}),$$

where $\mathbf{f}$ is the latent continuous values corresponding to $\mathbf{y}$, $\mathbf{w}$ is the weight vector for the latent features, $p(y_i|f_i) = \sum_{r=0}^{2} \delta(y_i = r)\delta(b_r \leq f_i < b_{r+1})$ and $p(\mathbf{f}_i|\mathbf{u}_i, \mathbf{w}) = \mathcal{N}(\mathbf{f}_i|\mathbf{u}_i^\top \mathbf{w}, 1)$. Note that $\mathbf{y}$ are linked to $\mathbf{X}$ 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 $\mathbf{y}$.

LD structure as additional priors for SNPs correlations. Linkage Disequilibrium records the occurrence of non-random combinations of alleles or genetic markers, which can be a natural indicator for the SNPs correlations. To utilize this valuable prior knowledge, we introduce a latent $q \times k$ matrix $\tilde{\mathbf{H}}$, which is tightly linked to $\mathbf{H}$. Each column $\tilde{\mathbf{h}}_j$ of $\tilde{\mathbf{H}}$ is regularized by the graph Laplacian of the LD structure, i.e.,

$$\begin{aligned} p(\tilde{\mathbf{H}}|\mathbf{L}) &= \prod\nolimits_j \mathcal{N}(\tilde{\mathbf{h}}_j|\mathbf{0}, \mathbf{L}^{-1}) = \prod\nolimits_j \mathcal{N}(\mathbf{0}|\tilde{\mathbf{h}}_j, \mathbf{L}^{-1}) \\ &= p(\mathbf{0}|\tilde{\mathbf{H}}, \mathbf{L}), \end{aligned}$$

where $\mathbf{L}$ is the graph Laplacian matrix of the LD structure. As shown above, the prior $p(\tilde{\mathbf{H}}|\mathbf{L})$ has the same form as $p(\mathbf{0}|\tilde{\mathbf{H}}, \mathbf{L})$, which can be viewed as a generative model – in other words, the observation $\mathbf{0}$ is sampled from $\tilde{\mathbf{H}}$. This view enables us to combine the generative model for graph Laplacian regularization with the sparse projection model via a principled hybrid Bayesian framework (Lasserre, Bishop, and Minka 2006).

To link the two models together, we introduce a prior over $\tilde{\mathbf{H}}$:

$$p(\tilde{\mathbf{H}}|\mathbf{H}) = \prod\nolimits_j \mathcal{N}(\tilde{\mathbf{h}}_j|\mathbf{h}_j, \lambda\mathbf{I})$$

where the variance λ controls how similar $\tilde{\mathbf{H}}$ and $\mathbf{H}$ are in our model. For simplicity, we set $\lambda = 0$ so that $\mathbf{p}(\tilde{\mathbf{H}}|\mathbf{H}) = \delta(\tilde{\mathbf{H}} - \mathbf{H})$ where $\delta(a) = 1$ if $a = 1$ and $\delta(a) = 0$ if $a = 0$.

Sparse priors for projection matrices and weights vector. In order to discover critical interactions between data sources and enhance the model prediction, we use spike and slab prior (George and McCulloch 1997) to sparsify the projection matrices $\mathbf{G}$ and $\mathbf{H}$ and the weight vector $\mathbf{w}$. Specifically, we use a $p \times k$ matrix $\mathbf{S}_g$ to represent the selection of elements in $\mathbf{G}$: if $s_{ij}^g = 1$, g_{ij} is selected and follows a Gaussian prior distribution with variance σ_1^2; if $s_{ij}^g = 0$, g_{ij} is not selected and forced to almost zero (i.e., sampled from a Gaussian with a very small variance σ_2^2). Thereby, we have the following prior over $\mathbf{G}$:

$$p(\mathbf{G}|\mathbf{S}_g, \boldsymbol{\Pi}_g) = \prod_{i=1}^{p} \prod_{j=1}^{k} p(g_{ij}|s_g^{ij}) p(s_g^{ij}|\pi_g^{ij})$$

where π_g^{ij} in $\boldsymbol{\Pi}_g$ is the probability of selecting g_{ij} (i.e., $s_g^{ij} = 1$), $p(g_{ij}|s_g^{ij}) = s_g^{ij}\mathcal{N}(g_{ij}|0, \sigma_1^2) + (1 - s_g^{ij})\mathcal{N}(g_{ij}|0, \sigma_2^2)$, and $p(s_g^{ij}|\pi_g^{ij}) = {\pi_g^{ij}}^{s_g^{ij}}(1 - \pi_g^{ij})^{1-s_g^{ij}}$. Note that σ_2^2 should be close to 0 and we set $\sigma_1^2 = 1$ and $\sigma_2^2 = 10^{-6}$ in the experiment. For a full Bayesian treatment, we assign an uninformative hyperprior over $\boldsymbol{\Pi}_g$: $p(\boldsymbol{\Pi}_g|l_1, l_2) = \prod_{i=1}^{p} \prod_{j=1}^{k} \text{Beta}(\pi_g^{ij}|l_1, l_2)$ where $l_1 = l_2 = 1$. Similarly, we assign the prior for $\mathbf{H}$,

$$p(\mathbf{H}|\mathbf{S}_h, \boldsymbol{\Pi}_h) = \prod_{i=1}^{q} \prod_{j=1}^{k} p(h_{ij}|s_h^{ij}) p(s_h^{ij}|\pi_h^{ij}),$$

where $\mathbf{S}_h$ are binary selection variables, π_h^{ij} is selecting probability of h_{ij}, $p(h_{ij}|s_h^{ij}) = s_h^{ij}\mathcal{N}(h_{ij}|0, \sigma_1^2) + (1 - s_h^{ij})\mathcal{N}(h_{ij}|0, \sigma_2^2)$ and $p(s_h^{ij}|\pi_h^{ij}) = {\pi_h^{ij}}^{s_h^{ij}}(1 - \pi_h^{ij})^{1-s_h^{ij}}$. We assign uninformative Beta hyperpriors for $\boldsymbol{\Pi}_h$: $p(\boldsymbol{\Pi}_h|d_1, d_2) = \prod_{i=1}^{q} \prod_{j=1}^{k} \text{Beta}(\pi_h^{ij}|d_1, d_2)$ where $d_1 = d_2 = 1$. Finally, for weights vector $\mathbf{w}$, we have

$$p(\mathbf{w}|\mathbf{s}_w, \boldsymbol{\pi}_w) = \prod_{j=1}^{k} p(w_j|s_w^j) p(\mathbf{s}_w^j|\pi_w^j)$$

where $\mathbf{s}_w$ are binary selection variables, π_w^j is the selecting probability of w_j, $p(w_j|s_w^j) = s_w^j\mathcal{N}(w_j|0, \sigma_1^2) + (1 - s_w^j)\mathcal{N}(w_j|0, \sigma_2^2)$ and $p(s_w^j|\pi_w^j) = {\pi_w^j}^{s_w^j}(1 - \pi_w^j)^{1-s_w^j}$. We assign Beta hyperpriors for $\boldsymbol{\pi}_w$: $p(\boldsymbol{\pi}_w) = \prod_{i=1}^{k} \text{Beta}(\pi_{\text{w}}^{\text{i}}|\text{e}_1, \text{e}_2)$ where $e_1 = e_2 = 1$.

Joint distribution. Based on the aforementioned components, the joint distribution of our model is

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{Z}, \mathbf{y}, \mathbf{U}, \mathbf{G}, \mathbf{S}_g, \boldsymbol{\Pi}_g, \eta, \mathbf{C}, \mathbf{H}, \tilde{\mathbf{H}}, \mathbf{S}_h, \boldsymbol{\Pi}_h, \mathbf{S}_w, \boldsymbol{\Pi}_w, \mathbf{f}) \\
&= p(\mathbf{X}|\mathbf{U}, \mathbf{G}, \eta) p(\mathbf{G}|S_g) p(S_g|\boldsymbol{\Pi}_g) p(\boldsymbol{\Pi}_g|l_1, l_2) p(\eta|r_1, r_2) \\
&\cdot p(\mathbf{Z}, \mathbf{C}|\mathbf{U}, \mathbf{H}) p(\mathbf{H}|\mathbf{S}_h) p(\mathbf{S}_h|\boldsymbol{\Pi}_h) p(\boldsymbol{\Pi}_h|d_1, d_2) p(\tilde{\mathbf{H}}|\mathbf{H}) \\
&\cdot p(\mathbf{0}|\tilde{\mathbf{H}}, \mathbf{L}) p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{U}, \mathbf{w}) p(\mathbf{w}|\mathbf{S}_w) p(\mathbf{S}_w|\boldsymbol{\Pi}_w) p(\mathbf{U}).
\end{aligned}$$

Estimation

Given the model, we present an efficient approach 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 selecting probabilities $\boldsymbol{\Pi}_g$ and $\boldsymbol{\Pi}_h$, the variance η, the auxiliary variables $\mathbf{C}$ for ordinal data $\mathbf{Z}$, the auxiliary variables $\mathbf{f}$ for ordinal labels $\mathbf{y}$, the weights vector $\mathbf{w}$ for prediction and the corresponding selection indicators and probabilities $\mathbf{s}_w$ and $\boldsymbol{\pi}_w$. In a Bayesian framework, this amounts to computing their posterior distributions.

However, computing the exact posteriors is infeasible because we cannot calculate the normalization constant for the posterior distributions. Therefore, we resort to a mean-field variational approach, which approximates the posterior distributions of $\{\mathbf{U}, \mathbf{H}, \mathbf{G}, \mathbf{S}_g, \mathbf{S}_h, \boldsymbol{\Pi}_g, \boldsymbol{\Pi}_h, \eta, \mathbf{w}, \mathbf{C}, \mathbf{f}\}$ by a factorized distribution $Q(\cdot) = Q(\mathbf{U})Q(\mathbf{H})Q(\mathbf{G})Q(\mathbf{S}_g)Q(\mathbf{S}_h)$ $Q(\boldsymbol{\Pi}_g)Q(\boldsymbol{\Pi}_h)Q(\eta)Q(\mathbf{w})Q(\mathbf{C})Q(\mathbf{f})$. Note that since we set $p(\tilde{\mathbf{H}}|\mathbf{H}) = \delta(\mathbf{H} - \tilde{\mathbf{H}})$, we do not need a separate distribution $Q(\tilde{\mathbf{H}})$. We minimize the Kullback-Leibler (KL) divergence between the approximate and the exact posteriors, $\text{KL}(Q||P)$ where P represents the exact joint posterior distributions. 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 sections. For brevity, the calculation of the required moments are provided in the supplementary material.

Updating variational distributions for continuous data

For 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 $\boldsymbol{\Pi}_g$ are

$$Q(\mathbf{G}) = \prod_{i=1}^{p} \mathcal{N}(\mathbf{g}_i; \boldsymbol{\lambda}_i, \boldsymbol{\Omega}_i), \tag{1}$$

$$Q(\mathbf{S}_g) = \prod_{i=1}^{p} \prod_{j=1}^{k} \beta_{ij}^{s_g^{ij}} (1 - \beta_{ij})^{1-s_g^{ij}}, \tag{2}$$

$$Q(\boldsymbol{\Pi}_g) = \prod_{i=1}^{p} \prod_{j=1}^{k} \text{Beta}(\pi_g^{ij}|\tilde{l}_1^{ij}, \tilde{l}_2^{ij}), \tag{3}$$

$$Q(\eta) = \text{Gamma}(\eta|\tilde{r}_1, \tilde{r}_2). \tag{4}$$

The parameters for $Q(\mathbf{G})$ are calculated by $\boldsymbol{\Omega}_i = \big(\langle\eta\rangle\langle\mathbf{U}\mathbf{U}^\top\rangle + \frac{1}{\sigma_1^2}\text{diag}(\langle\mathbf{s}_g^i\rangle) + \frac{1}{\sigma_2^2}\text{diag}(\mathbf{1} - \langle\mathbf{s}_g^i\rangle)\big)^{-1}$ and $\boldsymbol{\lambda}_i = \boldsymbol{\Omega}_i(\langle\eta\rangle\langle\mathbf{U}\rangle\tilde{\mathbf{x}}_i)$, where $\langle\cdot\rangle$ means the expectation over a distribution, $\tilde{\mathbf{x}}_i$ and $\mathbf{s}_g^i$ are the transpose of the i-th row in $\mathbf{X}$ and $\mathbf{S}_g$ respectively; the parameters for $Q(\mathbf{S}_g)$ are calculated by $\beta_{ij} = 1/\big(1 + \exp(\langle\log(1 - \pi_g^{ij})\rangle - \langle\log(\pi_g^{ij})\rangle + \frac{1}{2}\log(\frac{\sigma_1^2}{\sigma_2^2}) + \frac{1}{2}\langle g_{ij}^2\rangle(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))\big)$; the parameters for $Q(\boldsymbol{\Pi}_g)$ are given by $\tilde{l}_1^{ij} = \beta_{ij} + l_1$ and $\tilde{l}_2^{ij} = 1 - \beta_{ij} + l_2$; the parameters for $Q(\eta)$ are given by $\tilde{r}_1 = r_1 + \frac{np}{2}$ and $\tilde{r}_2 = r_2 + \frac{1}{2}\text{tr}(\mathbf{X}\mathbf{X}^\top) - \text{tr}(\langle\mathbf{G}\rangle\langle\mathbf{U}\rangle\mathbf{X}^\top) + \frac{1}{2}\text{tr}(\langle\mathbf{U}\mathbf{U}^\top\rangle\langle\mathbf{G}^\top\mathbf{G}\rangle)$.

Updating variational distributions for ordinal data

For ordinal data $\mathbf{Z}$, we update the approximate distributions of the projection matrix $\mathbf{H}$, the auxiliary variables $\mathbf{C}$, the selection indicators $\mathbf{S}_h$ and the selecting probabilities $\mathbf{\Pi}_h$. To make the variational distributions tractable, we factorize $Q(\mathbf{H})$ in a column-wise way. This is different from $Q(\mathbf{G})$ which are factorized in a row-wise way. Specifically, we denote the i-th column of $\mathbf{H}$ and $\mathbf{S}_h$ by $\tilde{\mathbf{h}}_i$ and $\mathbf{s}_h^i$, the i-th row of $\mathbf{U}$ by $\tilde{\mathbf{u}_i}^\top$, and calculate the variational distributions of $\mathbf{C}$ and $\mathbf{H}$ by

$$Q(\mathbf{C}) = \prod_{i=1}^{q} \prod_{j=1}^{k} Q(c_{ij}), \tag{5}$$

$$Q(c_{ij}) \propto \delta(b_{z_{ij}} \leq c_{ij} < b_{z_{ij}+1}) \mathcal{N}(c_{ij}|\bar{c}_{ij}, 1), \tag{6}$$

$$Q(\mathbf{H}) = \prod_{i=1}^{k} \mathcal{N}(\tilde{\mathbf{h}}_i; \boldsymbol{\gamma}_i, \mathbf{\Lambda}_i), \tag{7}$$

where $\bar{c}_{ij} = (\langle \mathbf{H} \rangle \langle \mathbf{U} \rangle)_{ij}$, $\mathbf{\Lambda}_i = \big(\langle \tilde{\mathbf{u}_i}^\top \tilde{\mathbf{u}_i} \rangle \mathbf{I} + \mathbf{L} + \frac{1}{\sigma_1^2} \mathrm{diag}(\langle \mathbf{s}_h^i \rangle) + \frac{1}{\sigma_2^2} \mathrm{diag}(\langle \mathbf{1} - \mathbf{s}_h^i \rangle)\big)^{-1}$, and $\boldsymbol{\gamma}_i = \mathbf{\Lambda}_i (\langle \mathbf{C} \rangle - \sum_{j \neq i} \boldsymbol{\gamma}_j \langle \tilde{\mathbf{u}}_j \rangle) \langle \mathbf{u}_i \rangle$.

The variational distributions of $\mathbf{S}_h$ and $\mathbf{\Pi}_h$ are

$$Q(\mathbf{S}_h) = \prod_{i=1}^{q} \prod_{j=1}^{k} \alpha_{ij}^{s_h^{ij}} (1 - \alpha_{ij})^{1 - s_h^{ij}}, \tag{8}$$

$$Q(\mathbf{\Pi}_h) = \prod_{i=1}^{q} \prod_{j=1}^{k} \mathrm{Beta}(\pi_\mathrm{h}^\mathrm{ij} | \tilde{\mathrm{d}}_1^\mathrm{ij}, \tilde{\mathrm{d}}_2^\mathrm{ij}), \tag{9}$$

where $\alpha_{ij} = 1 / \big(1 + \exp(\langle \log(1 - \pi_h^{ij}) \rangle - \langle \log(\pi_h^{ij}) \rangle + \frac{1}{2} \log(\frac{\sigma_1^2}{\sigma_2^2}) + \frac{1}{2} \langle h_{ij}^2 \rangle (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))\big)$, and $\tilde{d}_1^{ij} = \alpha_{ij} + d_1$, $\tilde{d}_2^{ij} = 1 - \alpha_{ij} + d_2$.

Note that in Equation (6), $Q(c_{ij})$ is a truncated Gaussian and the truncation is controlled by the observed ordinal data z_{ij}.

Updating variational distributions for labels

For ordinal labels $\mathbf{y}$, we calculate the approximate distributions of the auxiliary variables $\mathbf{f}$, the weights vector $\mathbf{w}$, the selection indicators $\mathbf{s}_w$ and the selecting probabilities $\boldsymbol{\pi}_w$ by

$$Q(\mathbf{f}) = \prod_{i=1}^{n} Q(f_i), \tag{10}$$

$$Q(f_i) \propto \delta(b_{y_i} \leq f_i < b_{y_i+1}) \mathcal{N}(f_i | \bar{f}_i, \sigma_{f_i}^2), \tag{11}$$

$$Q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{\Sigma}_w), \tag{12}$$

$$Q(\mathbf{s}_w) = \prod_{i=1}^{k} \tau_i^{\mathbf{s}_w^i} (1 - \tau_i)^{1 - \mathbf{s}_w^i}, \tag{13}$$

$$Q(\boldsymbol{\pi}_w) = \prod_{i=1}^{k} \mathrm{Beta}(\pi_w^i; \tilde{e}_1^i, \tilde{e}_2^i), \tag{14}$$

where $\bar{f}_i = (\langle \mathbf{U} \rangle^\top \mathbf{m})_i$, $\mathbf{\Sigma}_w = \big(\langle \mathbf{U} \mathbf{U}^\top \rangle + \frac{1}{\sigma_1^2} \mathrm{diag}(\langle \mathbf{s}_w \rangle) + \frac{1}{\sigma_2^2} \mathrm{diag}(\langle \mathbf{1} - \mathbf{s}_w \rangle)\big)^{-1}$, $\mathbf{m} = \mathbf{\Sigma}_w \langle \mathbf{U} \rangle \langle f \rangle$, $\tau_i = 1 / \big(1 + \exp(\langle \log(1 - \pi_w^i) \rangle - \langle \log(\pi_w^i) \rangle + \frac{1}{2} \langle w_i^2 \rangle (\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))\big)$, $\tilde{e}_1^i = \tau_i + e_1$ and $\tilde{e}_2^i = 1 - \tau_i + e_2$.

Note that $Q(f_i)$ is also a truncated Gaussian and the truncated region is decided by the ordinal label y_i. 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.

Updating variational distributions for latent feature U

The approximate distribution for $\mathbf{U}$ is

$$Q(\mathbf{U}) = \prod_{i} \mathcal{N}(\mathbf{u}_i | \boldsymbol{\mu}_i, \mathbf{\Sigma}_i) \tag{15}$$

where $\mathbf{\Sigma}_i = \big(\langle \mathbf{w}\mathbf{w}^\top \rangle + \langle \eta \rangle \langle \mathbf{G}^\top \mathbf{G} \rangle + \langle \mathbf{H}^\top \mathbf{H} \rangle + \mathbf{I}\big)^{-1}$ and $\boldsymbol{\mu}_i = \mathbf{\Sigma}_i (\langle \mathbf{w} \rangle \langle f_i \rangle + \langle \eta \rangle \langle \mathbf{G} \rangle^\top \mathbf{x}_i + \langle \mathbf{H} \rangle^\top \langle \mathbf{c}_i \rangle)$.

Prediction

Let us denote the training data by $\mathcal{D}_\mathrm{train} = \{\mathbf{X}_\mathrm{train}, \mathbf{Z}_\mathrm{train}, \mathbf{y}_\mathrm{train}\}$ and the test data by $\mathcal{D}_\mathrm{test} = \{\mathbf{X}_\mathrm{test}, \mathbf{Z}_\mathrm{test}\}$. We jointly carry out variational inference on $\mathcal{D}_\mathrm{train}$ and $\mathcal{D}_\mathrm{test}$. After the latent features for $\mathcal{D}_\mathrm{test}$ are obtained (i.e., $Q(U_\mathrm{test})$), we predict the labels by

$$\mathbf{f}_\mathrm{test} = \langle \mathbf{U}_\mathrm{test} \rangle^\top \mathbf{m}, \tag{16}$$

$$y_\mathrm{test}^i = \sum_{r=0}^{R-1} r \cdot \delta(b_r \leq f_\mathrm{test}^i < b_{r+1}), \tag{17}$$

where y_test^i is the prediction for i-th test sample.

Experimental Results and Discussion

We conducted association analysis and AD diagnosis based on a dataset from Alzheimer's Disease Neuroimaging Initiative (ADNI). The ADNI study is a longitudinal multisite observational study of elderly individuals with normal cognition, mild cognitive impairmen (MCI) or AD. We applied our model to study the associations between genotypes and brain atrophy measured by MRI and to predict the subject status (normal vs MCI vs AD). Note that the statuses are ordinal since they represent increasing severity levels.

The dataset was downloaded from http://adni.loni.ucla.edu/. After removing missing values, it consists of 618 subjects (183 normal, 308 MCI and 134 AD) and each subject contains 924 SNPs (selected as the top SNPs to separate normal subjects from AD in ADNI) and 328 MRI features (measuring the brain atrophies in different brain regions based on cortical thickness, surface area or volume using FreeSurfer software). Moreover, the LD structure was retrieved from www.ncbi.nlm.nih.gov/books/NBK44495/.

To evaluate the diagnosis accuracy, we compared with the following ordinal or multinomial regression methods: (1) lasso for multinomial regression (Tibshirani 1994), (2) elastic net for multinomial regression (Zou and Hastie 2005), (3) sparse ordinal regression with the spike and slab prior, (4) CCA + lasso, for which we first ran CCA to obtain the projected data and then applied lasso for prediction, (5) CCA + elastic net, which is the same as CCA + lasso except that we applied elastic net for prediction, (6) Gaussian Process Ordinal Regression (GPOR) (Chu and Ghahramani

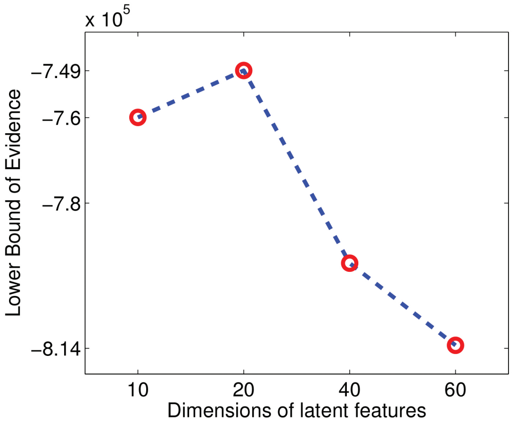


Figure 2: The variational lower bound for the model marginal likelihood.

2005), which employs Gaussian processes to learn the latent functions for ordinal regression, and (7) Laplacian Support Vector Machine (LapSVM) (Melacci and Mikhail 2011), a semi-supervised SVM classification scheme. We used the Glmnet package by (Friedman, Hastie, and Tibshirani 2010) for lasso and elastic net, the GPOR package by (Chu and Ghahramani 2005), and the LapSVM package by (Melacci and Mikhail 2011). For all the methods, we used 10-fold cross validation (i.e., each fold we have 556 training and 62 test samples) to tune free parameters, *e.g.*, the kernel form and parameters for GPOR and LapSVM. Note that all the alternative methods stack $\mathbf{X}$ and $\mathbf{Z}$ together into a whole data matrix and ignore their heterogeneous nature.

To determine the dimension k for the latent features $\mathbf{U}$ in our method, we calculated the variational lower bound as an approximation to the model evidence, with various k values $\{10, 20, 40, 60\}$. The variational lower bound can be easily calculated based on what we have presented in the model estimation section. We chose the value with the largest approximate evidence, which led to $k = 20$ (see Figure 2).

Our experiments confirmed that with $k = 20$, our model achieved the highest prediction accuracy, demonstrating the benefit of evidence maximization.

As shown in Figure 3, 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%. The two-sample t test shows our model outperforms the alternative methods significantly ($p < 0.05$).

We also examined the strongest associations discovered by our model. First of all, the ranking of MRI features in terms of prediction power for the three different disease populations (normal, MCI and AD) demonstrate that most of the top ranked features are based on the cortical thickness measurement. On the other hand, the features based on volume and surface area estimation are less predictive. Particularly, thickness measurements of middle temporal lobe, precuneus, and fusiform were found to be most predictive compared with other brain regions. These findings are consistent with the memory-related function in these regions and findings in the literature for their prediction power of AD (Whitwell et al. 2007; Risacher et al. 2009; Teipel et al. 2013). We also found that measurements of the same structure on the left and right side have similar weights, indicating that the algorithm can automatically select correlated features in groups, since no asymmetrical relationship has been found for the brain regions involved in AD (Kusbeci et al. 2009).

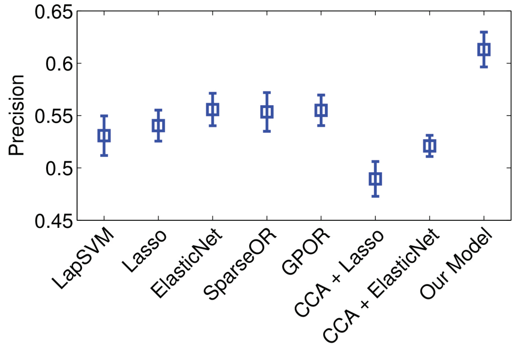


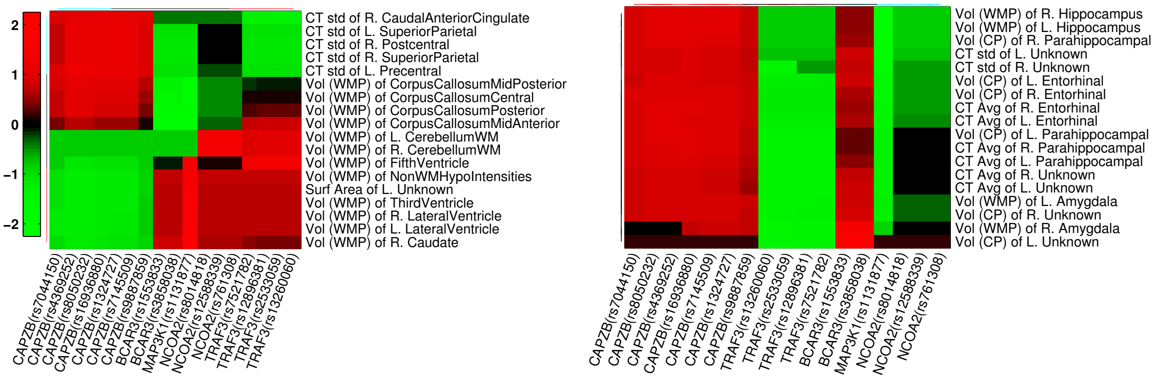
Figure 3: Prediction accuracy with standard errors.

Second, the analysis of associating genotype to AD prediction generated interesting results. Similar to the MRI features, SNPs that are in the vicinity of each other are often selected together, indicating the group selection characteristics of the algorithm. For example, the top ranked SNPs are associated with a few genes including CAPZB (F-actin-capping protein subunit beta), NCOA2 (The nuclear receptor coactivator 2) and BCAR3(Breast cancer anti-estrogen resistance protein 3).

At last, biclustering of the gene-MRI associations, as shown in Figure 4 reveal interesting pattern in terms of the relationship between genetic variations and brain atrophy measured by structural MRI. For example, the top ranked SNPs are associated with a few genes including BCAR3 (Breast cancer anti-estrogen resistance protein 3) and NCOA2, which have been studied more carefully in cancer research (Stephens et al. 2012). The set of SNPs are associated with cingulate in negative direction, which is part of the limbic system and involve in emotion formation and processing, compared with other structures such as temporal lobe, which plays a more important role in the formation of long-term memory.

Related Work

Our model is closely related to probabilistic factor analysis methods which try to learn a latent representation whose projection leads to the observed data (Tipping and Bishop 1999; Bach and Jordan 2005b; Guan and Dy 2009; Yu et al. 2006; Archambeau and Bach 2009; Virtanen, Klami, and Kaski 2011). In addition to learning latent representation, our model uses spike and slab prior to learn sparse projection matrices in order to select features in different data sources and find their critical associations. The spike and slab prior avoids confounding the degree of sparsity with the degree of



(a)

(b)

Figure 4: 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.

regularization and has shown to outperform l_1 regularization in an unsupervised setting (Mohamed, Heller, and Ghahramani 2012). Our model is also connected with learning from multiple sources or multiview learning methods (Hardoon et al. 2008), many of which try to learn a better classifier for multi-label classification based on the correlation structure among the training data and the labels (Yu et al. 2006; Virtanen, Klami, and Kaski 2011). However, our work conducts two tasks—the association discovery and ordinal label prediction—simultaneously to benefit each other. Finally, our work can be considered as an extension of the work (Zhe et al. 2013) by incorporating the LD structure knowledge and using sparse ordinal regression to model disease severity level.

Conclusions

We presented a new Bayesian multiview learning model for joint associations discovery and disease prediction in AD study. We expect that our model can be further applied to a wide range of applications in biomedical research – *e.g.*, eQTL analysis supervised by additional labeling information.

Acknowledgement

This work was supported by NSF IIS-0916443, IIS-1054903, CCF-0939370 and a Project 985 grant from University of Electronic Science and Technology (No. A1098531023601041).

References

Archambeau, C., and Bach, F. 2009. Sparse probabilistic projections. In *Advances in Neural Information Processing Systems 21*. 73–80.

Bach, F., and Jordan, M. 2005a. A probabilistic interpretation of canonical correlation analysis. Technical report, UC Berkeley.

Bach, F., and Jordan, M. 2005b. A probabilistic interpretation of canonical correlation analysis. Technical report, UC Berkley.

Chu, W., and Ghahramani, Z. 2005. Gaussian processes for ordinal regression. *Journal of Machine Learning Research* 6:1019–1041.

Falconer, D., and Mackay, T. 1996. *Introduction to Quantitative Genetics (4th ed.)*. Addison Wesley Longman.

Friedman, J.; Hastie, T.; and Tibshirani, R. 2010. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software* 33(1):1–22.

George, E., and McCulloch, R. 1997. Approaches for bayesian variable selection. *Statistica Sinica* 7(2):339–373.

Guan, Y., and Dy, J. 2009. Sparse probabilistic principal component analysis. *Journal of Machine Learning Research - Proceedings Track* 5:185–192.

Hardoon, D.; Leen, G.; Kaski, S.; and Shawe-Taylor, J., eds. 2008. *NIPS Workshop on Learning from Multiple Sources*.

Harold, H. 1936. Relations between two sets of variates. *Biometrika* 28:321–377.

Khachaturian, S. 1985. Diagnosis of Alzheimer's disease. *Archives of Neurology* 42(11):1097–1105.

Kusbeci, O. Y.; Bas, O.; Gocmen-Mas, N.; Karabekir, H. S.; Yucel, A.; Ertekin, T.; and Yazici, A. C. 2009. Evaluation of cerebellar asymmetry in Alzheimer's disease: a stereological study. *Dement Geriatr Cogn Disord* 28(1):1–5.

Lasserre, J.; Bishop, C. M.; and Minka, T. P. 2006. Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, 87–94.

MacKay, D. 1991. Bayesian interpolation. *Neural Computation* 4:415–447.

Melacci, S., and Mikhail, B. 2011. Laplacian support vector machines trained in the primal. *Journal of Machine Learning Research* 12:1149–1184.

Mohamed, S.; Heller, K.; and Ghahramani, Z. 2012. Bayesian and L1 approaches for sparse unsupervised learning. In *ICML'02*.

Neal, R. 1996. *Bayesian Learning for Neural Networks*.

Risacher, S.; Saykin, A.; West, J.; Firpi, H.; and McDonald, B. 2009. Baseline MRI predictors of conversion from MCI to probable AD in the ADNI cohort. *Curr. Alzheimer Res.* 6:347–361.

Stephens, P.; Tarpey, P.; Davies, H.; Van, Loo, P.; Greenman, C.; Wedge, D.; et al. 2012. The landscape of cancer genes and mutational processes in breast cancer. *Nature* 486(7403):400–404.

Teipel, S.; Grothe, M.; Lista, S.; Toschi, N.; Garaci, F.; and Hampel, H. 2013. Relevance of magnetic resonance imaging for early detection and diagnosis of Alzheimer disease. *Medical Clinics of North America* 97(3):399–424.

Tibshirani, R. 1994. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B* 58:267–288.

Tipping, M., and Bishop, C. 1999. Probabilistic principal component analysis. *Journal of The Royal Statistical Society Series B-statistical Methodology* 61:611–622.

Virtanen, S.; Klami, A.; and Kaski, S. 2011. Bayesian CCA via group sparsity. In *ICML'11*, 457–464.

Whitwell, J.; Przybelski, S.; Weigand, S.; et al. 2007. 3D maps from multiple MRI illustrate changing atrophy patterns as subjects progress from mild cognitive impairment to Alzheimer's disease. *Brain* 130(7):1777–1786.

Yu, S.; Yu, K.; Tresp, V.; Kriegel, H.; and Wu, M. 2006. Supervised probabilistic principal component analysis. In *KDD'06*, 464–473.

Zhe, S.; Xu, Z.; Qi, Y.; and Yu, P. 2013. Joint association discovery and diagnosis of alzheimer's disease by supervised heterogeneous multiview learning. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, volume 19, 300–311. World Scientific.

Zou, H., and Hastie, T. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67:301–320.