Journal of Artificial Intelligence Research 56 (2016) 247-268 Submitted 12/15; published 06/16

Association Discovery and Diagnosis of Alzheimer's Disease with Bayesian Multiview Learning

Zenglin Xu ZLXU@UESTC.EDU.CN
Big Data Research Center
School of Computer Science & Engineering
University of Electronic Science & Technology of China
Chengdu, Sichuan, 611731 China

Shandian Zhe SZHE@PURDUE.EDU
Department of Computer Science, Purdue University
West Lafayette, IN 47906 USA

Yuan(Alan) Qi ALANQI@CS.PURDUE.EDU
Department of Computer Science & Department of Statistics
Purdue University
West Lafayette, IN 47906 USA

Peng Yu YU_PENG_PY@LILLY.COM
Eli Lilly and Company, Indianapolis, IN 46225, USA

Abstract

The analysis and diagnosis of Alzheimer's disease (AD) can be based on genetic variations, *e.g.*, single nucleotide polymorphisms (SNPs) and phenotypic traits, *e.g.*, Magnetic Resonance Imaging (MRI) features. We consider two important and related tasks: i) to select genetic and phenotypical markers for AD diagnosis and ii) to identify associations between genetic and phenotypical data. While previous studies treat these two tasks separately, they are tightly coupled because underlying associations between genetic variations and phenotypical features contain the biological basis for a disease. Here 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; in return, the disease status can guide the discovery of relationships between data sources. The sparse projection matrices not only reveal interactions between data sources but also select groups of biomarkers related to the disease. Moreover, to take advantage of the linkage disequilibrium (LD) measuring the non-random association of alleles, we incorporate a graph Laplacian type of prior in the model. To learn the model from data, we develop an efficient variational inference algorithm. Analysis on an imaging genetics dataset for the study of Alzheimer's Disease (AD) indicates that our model identifies biologically meaningful associations between genetic variations and MRI features, and achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

1. Introduction

Alzheimer's disease (AD) is the most common neurodegenerative disorder (Khachaturian, 1985). In order to predict the onset and progression of AD, NIH funded the Alzheimer's Disease Neuroimaging Initiative (ADNI) to facilitate the evaluation of genetic variations, *e.g.*, Single Nucleotide Polymorphisms (SNPs) and phenotypical traits, *e.g.*, Magnetic Reso-

nance Imaging (MRI). In addition to progression study, it is becoming important in medical studies to identify the relevant pathological genotypes and phenotypic traits, and to discover their associations. Although found in many bioinformatics applications (Consoli, Lefevre, Zivy, de Vienne, & Damerval, 2002; Hunter, 2012; Gandhi & Wood, 2010; Liu, Pearlson, Windemuth, Ruano, Perrone-Bizzozero, & Calhoun, 2009), association studies are scarce and especially in need in the AD study.

Many statistical approaches have been developed to discover associations or select features (or variables) for prediction in a high dimensional problem. For association studies, representative approaches are canonical correlation analysis (CCA) and its extensions (Harold, 1936; Bach & Jordan, 2005). These approaches have been widely used in expression quantitative trait locus (eQTL) analysis (Parkhomenko, Tritchler, & Beyene, 2007; Daniela & Tibshirani, 2009; Chen, Liu, & Carbonell, 2012). For disease diagnosis based on high dimensional biomarkers, popular approaches include lasso (Tibshirani, 1994), elastic net (Zou & Hastie, 2005), and group lasso (Yuan & Lin, 2007), and Bayesian automatic relevance determination (MacKay, 1991; Neal, 1996). Despite their wide success in many applications, these approaches are limited by the following reasons:

- Most association studies neglect the supervision from the disease status. Because many diseases, such as AD, are a direct result of genetic variations and often highly correlated to clinical traits, the disease status provides useful yet currently unutilized information for finding relationships between genetic variations and clinical traits.

- For disease diagnosis, most sparse approaches use classification models and do not consider the order of disease severity. For subjects in AD studies, there is a natural severity order from being normal to mild cognitive impairment (MCI) and then from MCI to AD. Classification models cannot capture the order in AD's severity levels.

- Most previous methods are not designed to handle heterogeneous data types. The SNPs values are discrete (and ordinal based on an additive genetic model), while the imaging features are continuous. Popular CCA or lasso-type methods simply treat both of them as continuous data and overlook the heterogeneous nature of the data.

- Most previous methods ignore or cannot utilize the valuable prior knowledge. For example, the occurrence of some combinations of alleles or genetic markers in a population are more often or less often than that would be expected from a random formation of haplotypes from alleles based on their frequencies, which is known as Linkage Disequilibrium (LD) (Falconer & Mackay, 1996). To our knowledge, this structure has not been utilized in association discovery.

To address these problems, we propose a new Bayesian approach that unifies multiview learning with sparse ordinal regression for joint association study and disease diagnosis. It can also conduct nonlinear classification over latent variables (Zhe, Xu, Qi, & Yu, 2014) and find associations by incorporating the LD information as an additional prior for the SNPs data (Zhe, Xu, Qi, & Yu, 2015) . In more detail, genetic variations and phenotypical traits are generated from common *latent* features based on separate sparse projection matrices and suitable link functions, and the common latent features are used to predict the disease status (See Section 2). To enforce sparsity in projection matrices, we assign spike and slab priors

(George & McCulloch, 1997) over them; these priors have been shown to be more effective than l_1 penalty to learn sparse projection matrices (Goodfellow, Couville, & Bengio, 2012; Mohamed et al., 2012). In order to take advantage of the linkage disequilibrium, which describes the non-random association of alleles at different loci, we employ an additional graph Laplacian type of prior for the SNPs view. The sparse projection matrices not only reveal critical interactions between the different data sources but also identify biomarkers in data relevant to disease status. Meanwhile, via its direct connection to the latent features, the disease status influences the estimation of the projection matrices so that it can guide the discovery of associations between heterogeneous data sources relevant to the disease.

To learn the model from data, we develop a variational inference approach (See Section 3). It iteratively minimizes the Kullback-Leibler divergence between a tractable approximation and exact Bayesian posterior distributions. We extend the proposed sparse multiview learning model by incorporating the linkage disequilibrium information about SNPs in Section 4. We then employ our model to the real study of AD in Section 5. The results show that our model achieves the highest prediction accuracy among all the competing methods. Furthermore, our model finds biologically meaningful predictive relationships between SNPs, MRI features, and AD status.

2. Sparse Heterogeneous Multiview Learning Models

In this section, we first present the notations and assumptions, and then present the sparse heterogeneous multiview learning model.

2.1 Notations and Assumptions

First, let us describe the data. We assume there are two heterogeneous data sources: one contains continuous data – for example, MRI features – and the other contains discrete ordinal data – for instance, 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 Poisson 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 $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_n]$, 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.

2.2 Spare Heterogeneous Multiview Learning Model

To link the 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 projection. 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 with dimension k. We denote the proposed Spare Heterogeneous Multiview Learning Model by SHML. 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.

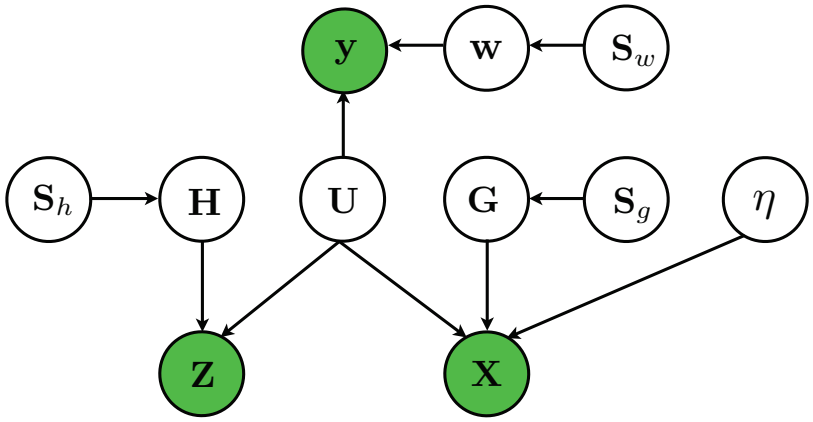


Figure 1: The graphical representation of SHML, where $\mathbf{X}$ is the continuous view, $\mathbf{Z}$ is the ordinal view, $\mathbf{y}$ are the labels.

2.2.1 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 $\eta^{-1}\mathbf{I}$ is the precision matrix of the Gaussian distribution. For the precision parameter η, we assign a conjugate prior *Gamma* prior, $p(\eta|r_1, r_2) = Gamma(\eta|r_1, r_2)$ where r_1 and r_2 are the hyperparameters and set to be 10^{-3} in our experiments.

2.2.2 Ordinal Data Distribution

For an ordinal variable $z \in \{0, 1, \ldots, R-1\}$, 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

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

where

$$p(c_{ij}|\mathbf{h}_i,\mathbf{u}_j) = \mathcal{N}(c_{ij}|\mathbf{h}_i^\top\mathbf{u}_j, 1)$$
$$p(z_{ij}|c_{ij}) = \sum_{r=0}^{2} \delta(z_{ij}=r)\delta(b_r \le c_{ij} < b_{r+1}).$$

Here $\delta(a) = 1$ if a is true and $\delta(a) = 0$ otherwise.

2.2.3 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 and

$$p(\mathbf{f}_i|\mathbf{u}_i, \mathbf{w}) = \mathcal{N}(\mathbf{f}_i|\mathbf{u}_i^\top \mathbf{w}, 1),$$
$$p(y_i|f_i) = \sum_{r=0}^{2} \delta(y_i = r)\delta(b_r \leq f_i < b_{r+1}).$$

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 $\mathbf{y}$.

2.2.4 Sparse Priors for Projection Matrices and Weights Vector

Because we want to identify a few critical interactions between different data sources, we use spike and slab prior (George & McCulloch, 1997) to sparsify the projection matrices $\mathbf{G}$ and $\mathbf{H}$. The spike and slab priors are continuous bimodal priors to model hypervariance parameters, which controls both the selection of the variable and the effective scale of choosing this variable. We apply the spike and slab prior over 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). We have the following prior over $\mathbf{G}$:

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

where

$$p(g_{ij}|s_g^{ij}) = s_g^{ij}\mathcal{N}(g_{ij}|0, \sigma_1^2) + (1 - s_g^{ij})\mathcal{N}(g_{ij}|0, \sigma_2^2),$$
$$p(s_g^{ij}|\pi_g^{ij}) = {\pi_g^{ij}}^{s_g^{ij}}(1 - \pi_g^{ij})^{1-s_g^{ij}},$$

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 $\sigma_2^2 = 1o^{-6}$). To reflect our uncertainty about $\mathbf{\Pi}_g$, we assign a *Beta* hyperprior distribution:

$$p(\mathbf{\Pi}_g|l_1, l_2) = \prod_{i=1}^{p}\prod_{j=1}^{k} Beta(\pi_g^{ij}|l_1, l_2),$$

where l_1 and l_2 are hyperparameters. We set a diffuse and non-informative hyperprior, i.e., $l_1 = l_2 = 1$ in our experiments. Similarly, $\mathbf{H}$ is sampled from

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

where $p(h_{ij}|s_h^{ij}) = s_h^{ij}\mathcal{N}(h_{ij}|0,\sigma_1^2)+(1-s_h^{ij})\mathcal{N}(h_{ij}|0,\sigma_2^2)$ and $p(s_h^{ij}|\pi_h^{ij}) = {\pi_h^{ij}}^{s_h^{ij}}(1-\pi_h^{ij})^{1-s_h^{ij}}$. $\mathbf{S}_h$ are binary selection variables and π_h^{ij} in $\mathbf{\Pi}_h$ is the probability of $s_h^{ij}=1$. We assign *Beta* hyperpriors for $\mathbf{\Pi}_h$:

$$p(\mathbf{\Pi}_h|d_1,d_2) = \prod_{i=1}^{q}\prod_{j=1}^{k} Beta(\pi_h^{ij}|d_1,d_2),$$

where d_1 and d_2 are hyperparameters. We set $d_1 = d_2 = 1$ in our experiments since we have found that they are not sensitive to the final performance. Similarly for weights vector $\mathbf{w}$,

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

where $p(w_j|s_w^j) = s_w^j\mathcal{N}(w_j|0,\sigma_1^2) + (1-s_w^j)\mathcal{N}(w_j|0,\sigma_2^2)$ and $p(s_w^j|\pi_w^j) = {\pi_w^j}^{s_w^j}(1-\pi_w^j)^{1-s_w^j}$. $\mathbf{s}_w$ are binary selection variables and π_w^j in $\boldsymbol{\pi}_w$ is the probability of $s_w^j = 1$. We assign *Beta* hyperpriors for $\boldsymbol{\pi}_w$:

$$p(\boldsymbol{\pi}_w) = \prod_{i=1}^{k} Beta(\pi_w^i|e_1,e_2),$$

where e_1 and e_2 are hyperparameters. We similarly set $e_1 = e_2 = 1$ in our experiments.

2.2.5 Joint Distribution

Based on all these specifications, the joint distribution of our model is

$$\begin{aligned} & p(\mathbf{X},\mathbf{Z},\mathbf{y},\mathbf{U},\mathbf{G},\mathbf{S}_g,\mathbf{\Pi}_g,\eta,\mathbf{C},\mathbf{H},\tilde{\mathbf{H}},\mathbf{S}_h,\mathbf{\Pi}_h,\mathbf{S}_w,\mathbf{\Pi}_w,\mathbf{f}) \\ = \; & p(\mathbf{X}|\mathbf{U},\mathbf{G},\eta)p(\mathbf{G}|S_g)p(S_g|\mathbf{\Pi}_g)p(\mathbf{\Pi}_g|l_1,l_2)p(\eta|r_1,r_2) \\ & \cdot p(\mathbf{Z},\mathbf{C}|\mathbf{U},\mathbf{H})p(\mathbf{H}|\mathbf{S}_h)p(\mathbf{S}_h|\mathbf{\Pi}_h)p(\mathbf{\Pi}_h|d_1,d_2) \\ & \cdot p(\mathbf{y}|\mathbf{f})p(\mathbf{f}|\mathbf{U},\mathbf{w})p(\mathbf{w}|\mathbf{S}_w)p(\mathbf{S}_w|\mathbf{\Pi}_w)p(\mathbf{U}). \end{aligned} \tag{1}$$

Different from Figure 1, we put the conjugate prior for $\mathbf{S}_w$ and S_g into the joint distribution. Then the next step is to estimate the distributions of the latent variables and their hyperparemeters.

3. Model Inference

Given the model specified in the previous section, now we present an efficient 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 $\mathbf{Z}$, the auxiliary variables $\mathbf{f}$ for generating the labels $\mathbf{y}$, the weights vector $\mathbf{w}$ for generating $\mathbf{f}$ and the corresponding selection indicators and probabilities $\mathbf{s}_w$ and $\boldsymbol{\pi}_w$. 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 posteriors based on Equation (1). Thus, we

resort to a mean-field variational approach. Specifically, we approximate the posterior distributions of $\mathbf{U}, \mathbf{H}, \mathbf{G}, \mathbf{S}_g, \mathbf{S}_h, \mathbf{\Pi}_g, \mathbf{\Pi}_h, \eta, \mathbf{w}, \mathbf{C}$ and $\mathbf{f}$ by a factorized distribution

$$Q(\boldsymbol{\theta}) = Q(\mathbf{U})Q(\mathbf{H})Q(\mathbf{G})Q(\mathbf{S}_g)Q(\mathbf{S}_h)Q(\mathbf{\Pi}_g)Q(\mathbf{\Pi}_h)Q(\eta)Q(\mathbf{w})Q(\mathbf{C})Q(\mathbf{f}) \tag{2}$$

where $\boldsymbol{\theta}$ denotes all the latent variables.

Variational inference minimizes the Kullback-Leibler (KL) divergence between the approximate and the exact posteriors

$$\min_{Q(\boldsymbol{\theta})} \mathrm{KL}\left(Q(\boldsymbol{\theta}) \| p(\boldsymbol{\theta}|\mathbf{X}, \mathbf{Z}, \mathbf{y})\right) \tag{3}$$

More specifically, using a coordinate descent algorithm, the variational approach updates one approximate distribution, e.g, $q(\mathbf{H})$, in Equation (2) at a time while having all the others fixed. The detailed updates are given in the following paragraphs.

3.1 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

$$Q(\mathbf{G}) = \prod_{i=1}^{p} \mathcal{N}(\mathbf{g}_i; \boldsymbol{\lambda}_i, \mathbf{\Omega}_i), \tag{4}$$

$$Q(\mathbf{S}_g) = \prod_{i=1}^{p}\prod_{j=1}^{k} \beta_{ij}^{s_g^{ij}} (1-\beta_{ij})^{1-s_g^{ij}}, \tag{5}$$

$$Q(\mathbf{\Pi}_g) = \prod_{i=1}^{p}\prod_{j=1}^{k} Beta(\pi_g^{ij}|\tilde{l}_1^{ij}, \tilde{l}_2^{ij}), \tag{6}$$

$$Q(\eta) = Gamma(\eta|\tilde{r}_1, \tilde{r}_2). \tag{7}$$

The mean and covariance of $\mathbf{g}_i$ are calculated as follows:

$$\mathbf{\Omega}_i = \left(\langle\eta\rangle\langle\mathbf{U}\mathbf{U}^\top\rangle + \frac{1}{\sigma_1^2}\mathrm{diag}(\langle\mathbf{s}_g^i\rangle) + \frac{1}{\sigma_2^2}\mathrm{diag}(\mathbf{1} - \langle\mathbf{s}_g^i\rangle)\right)^{-1},$$
$$\boldsymbol{\lambda}_i = \mathbf{\Omega}_i(\langle\eta\rangle\langle\mathbf{U}\rangle\tilde{\mathbf{x}}_i),$$

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 computation of parameters β_{ij} and $Q(\pi_g^{ij})$ can be found in Appendices A.

3.2 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 $\mathbf{\Pi}_h$. To make the variational distributions tractable, we update $Q(\mathbf{H})$ in a column-wise

way and re-denote $\mathbf{H} = [\tilde{h}_1, \tilde{h}_2, ...\tilde{h}_k]$, $\mathbf{S}_h = [\mathbf{s}_h^1, \mathbf{s}_h^2, ...\mathbf{s}_h^k]$ and $\mathbf{U} = [\tilde{\mathbf{u}}_1, \tilde{\mathbf{u}}_2, ...\tilde{\mathbf{u}}_k]^\top$. The variational distributions of $\mathbf{C}$ and $\mathbf{H}$ are

$$Q(\mathbf{C}) = \prod_{i=1}^{q}\prod_{j=1}^{k} Q(c_{ij}), \tag{8}$$

$$Q(c_{ij}) \propto \delta(b_{z_{ij}} \leq c_{ij} < b_{z_{ij}+1})\mathcal{N}(c_{ij}|\bar{c}_{ij}, 1), \tag{9}$$

$$Q(\mathbf{H}) = \prod_{i=1}^{k} \mathcal{N}(\bar{\mathbf{h}}_i; \boldsymbol{\gamma}_i, \boldsymbol{\Lambda}_i), \tag{10}$$

where $\bar{c}_{ij} = (\langle\mathbf{H}\rangle\langle\mathbf{U}\rangle)_{ij}$, $\boldsymbol{\Lambda}_i = \left(\langle\tilde{\mathbf{u}}_i{}^\top\tilde{\mathbf{u}}_i\rangle\mathbf{I} + \frac{1}{\sigma_1^2}\mathrm{diag}(\langle\mathbf{s}_h^i\rangle) + \frac{1}{\sigma_2^2}\mathrm{diag}(\langle\mathbf{1}-\mathbf{s}_h^i\rangle)\right)^{-1}$, $\boldsymbol{\gamma}_i = \boldsymbol{\Lambda}_i\mathbf{C}_i\langle\mathbf{u}_i\rangle$ and $\mathbf{C}_i = \mathbf{C} - \sum_{j\neq i}\boldsymbol{\gamma}_j\langle\tilde{\mathbf{u}}_j\rangle^\top$. The computation of parameters in the distributions of $\mathbf{S}_h$ and $\boldsymbol{\Pi}_h$ is given in Appendices B.

3.3 Updating Variational Distributions for Labels

For the ordinal labels $\mathbf{y}$, we update the approximation distributions of the auxiliary variables $\mathbf{f}$, the weights vector $\mathbf{w}$, the sparse selection indicators $\mathbf{s}_w$ and the selection probabilities $\boldsymbol{\pi}_w$. The variational distributions of $\mathbf{f}$ and $\mathbf{w}$ are

$$Q(\mathbf{f}) = \prod_{i=1}^{n} Q(f_i), \tag{11}$$

$$Q(f_i) \propto \delta(b_{y_i} \leq f_i < b_{y_i+1})\mathcal{N}(f_i|\bar{f}_i, \sigma_{f_i}^2), \tag{12}$$

$$Q(\mathbf{w}) = \mathcal{N}(\mathbf{w}; \mathbf{m}, \boldsymbol{\Sigma}_w), \tag{13}$$

where $\bar{f}_i = (\langle\mathbf{U}\rangle^\top\mathbf{m})_i$, $\boldsymbol{\Sigma}_w = \left(\langle\mathbf{U}\mathbf{U}^\top\rangle + \frac{1}{\sigma_1^2}\mathrm{diag}(\mathbf{s}_w) + \frac{1}{\sigma_2^2}\mathrm{diag}(1-\mathbf{s}_w)\right)^{-1}$ and $\mathbf{m} = \boldsymbol{\Sigma}_w\langle\mathbf{U}\rangle\langle f\rangle$. The computation of parameters in the variational distributions of $\mathbf{s}_w$ and π_w can be found in Appendices C.

3.4 Updating Variational Distributions for Latent Representation U

The variational distribution for $\mathbf{U}$ is given by

$$Q(\mathbf{U}) = \prod_i \mathcal{N}(\mathbf{u}_i|\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i) \tag{14}$$

where

$$\boldsymbol{\Sigma}_i = \left(\langle\mathbf{w}\mathbf{w}^\top\rangle + \langle\eta\rangle\langle\mathbf{G}^\top\mathbf{G}\rangle + \langle\mathbf{H}^\top\mathbf{H}\rangle + \mathbf{I}\right)^{-1} \tag{15}$$

$$\boldsymbol{\mu}_i = \boldsymbol{\Sigma}_i(\langle\mathbf{w}\rangle\langle f_i\rangle + \langle\eta\rangle\langle\mathbf{G}\rangle^\top\mathbf{x}_i + \langle\mathbf{H}\rangle^\top\langle\mathbf{c}_i\rangle). \tag{16}$$

The required moments are given in Appendices D.

3.5 Label Prediction

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 $\mathcal{D}_{\text{test}} = \{\mathbf{X}_{\text{test}}, \mathbf{Z}_{\text{test}}\}$. The prediction task needs the latent representation $\mathbf{U}_{\text{test}}$ for $\mathcal{D}_{\text{test}}$.

We carry out variational inference simultaneously on $\mathcal{D}_{\text{train}}$ and $\mathcal{D}_{\text{test}}$. After both $Q(\mathbf{U}_{\text{test}})$ and $Q(\mathbf{U}_{\text{train}})$ are obtained, we predict the labels for test data as follows:

$$\mathbf{f}_{\text{test}} = \langle \mathbf{U}_{\text{test}} \rangle^{\top} \mathbf{m}, \tag{17}$$

$$y^i_{\text{test}} = \sum_{r=0}^{R-1} r \cdot \delta(b_r \leq f^i_{\text{test}} < b_{r+1}), \tag{18}$$

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

4. Sparse Heterogeneous Multiview Learning Model with Linkage Disequilibrium Priors

In population genetics, lLinkage Disequilibrium (LD) refers to the non-random association of alleles at different loci, i.e., the presence of statistical associations between alleles at different loci that are different from what would be expected if alleles were independently, randomly sampled based on their individual allele frequencies (Slatkin, 2008). If there is no linkage disequilibrium between alleles at different loci they are said to be in linkage equilibrium.

Linkage Disequilibrium also appears in the SNPs, which is a measure between pairs of SNPs and can be regarded as a natural indicator for the correlation between SNPs. This information can be publicly retrieved from `www.ncbi.nlm.nih.gov/books/NBK44495/`. To incorporate such correlation as a prior in our model, we first introduce a latent $q \times k$ matrix $\tilde{\mathbf{H}}$, which is tightly linked to $\mathbf{H}$ as explained later. Each column $\tilde{\mathbf{h}}_j$ of $\tilde{\mathbf{H}}$ is regularized by the graph Laplacian of the LD structure, i.e.,

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

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 et al., 2006).

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

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

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 $\mathbf{p}(\tilde{\mathbf{H}}|\mathbf{H}) = \text{Dirac}(\tilde{\mathbf{H}} - \mathbf{H})$ where $\text{Dirac}(a) = 1$ if $a = 1$ and $\text{Dirac}(a) = 0$ if $a = 0$.

Adopting this additional information, the new graphical model is designed as shown in Fig. 2.

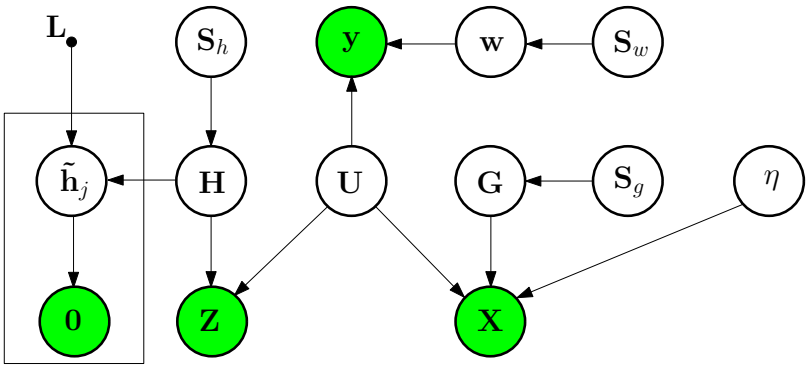


Figure 2: The graphical representation of our model, where $\mathbf{X}$ is the continuous view, $\mathbf{Z}$ is the ordinal view, $\mathbf{y}$ are the labels and $\mathbf{L}$ is the graph laplacian generated by the LD structure.

Based on all these specifications, the joint distribution of our model is

$$\begin{aligned}
&p(\mathbf{X}, \mathbf{Z}, \mathbf{y}, \mathbf{U}, \mathbf{G}, \mathbf{S}_g, \mathbf{\Pi}_g, \eta, \mathbf{C}, \mathbf{H}, \tilde{\mathbf{H}}, \mathbf{S}_h, \mathbf{\Pi}_h, \mathbf{S}_w, \mathbf{\Pi}_w, \mathbf{f}) \\
&= p(\mathbf{X}|\mathbf{U}, \mathbf{G}, \eta) p(\mathbf{G}|S_g) p(S_g|\mathbf{\Pi}_g) p(\mathbf{\Pi}_g|l_1, l_2) p(\eta|r_1, r_2) \\
&\cdot p(\mathbf{Z}, \mathbf{C}|\mathbf{U}, \mathbf{H}) p(\mathbf{H}|\mathbf{S}_h) p(\mathbf{S}_h|\mathbf{\Pi}_h) p(\mathbf{\Pi}_h|d_1, d_2) p(\tilde{\mathbf{H}}|\mathbf{H}) \\
&\cdot p(\mathbf{0}|\tilde{\mathbf{H}}, \mathbf{L}) p(\mathbf{y}|\mathbf{f}) p(\mathbf{f}|\mathbf{U}, \mathbf{w}) p(\mathbf{w}|\mathbf{S}_w) p(\mathbf{S}_w|\mathbf{\Pi}_w) p(\mathbf{U}).
\end{aligned} \tag{19}$$

The inference is almost the same with the original model described in Section 2, except the updating of the sparse projection matrix $\mathbf{H}$. Given the ordinal data $\mathbf{Z}$ and the updates of other variables, 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 $\mathbf{\Pi}_h$. The variational distributions of $\mathbf{C}$ and $\mathbf{H}$ are

$$Q(\mathbf{C}) = \prod_{i=1}^{q} \prod_{j=1}^{k} Q(c_{ij}), \tag{20}$$

$$Q(c_{ij}) \propto \delta(b_{z_{ij}} \leq c_{ij} < b_{z_{ij}+1}) \mathcal{N}(c_{ij}|\bar{c}_{ij}, 1), \tag{21}$$

$$Q(\mathbf{H}) = \prod_{i=1}^{k} \mathcal{N}(\bar{\mathbf{h}}_i; \boldsymbol{\gamma}_i, \mathbf{\Lambda}_i), \tag{22}$$

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

5. Experimental Results and Discussion

In order to examine the performance of the proposed method , we design a simulation study and a realworld study for Alzheimer's Disease.

5.1 Simulation Study

We first design a simulation study to examine the basic model, i.e., our model, in terms of (i) estimation accuracy on finding associations between the two views and (ii) prediction accuracy on the ordinal labels. Note that a similar study can be conducted on the model with LD priors.

5.1.1 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 $p(y_i|f_i) = \delta(y_i = 0)\delta(f_i \leq 0) + \delta(y_i = 1)\delta(f_i > 0)$.

5.1.2 Comparative Methods

We compared our model with several state-of-the-art methods including (1) CCA (Bach & Jordan, 2005), which finds the projection direction that maximizes the correlation between two views, (2) sparse CCA (Sun, Ji, & Ye, 2011; Daniela & Tibshirani, 2009), where sparse priors are put on the CCA directions, and (3) multiple-response regression with lasso (MRLasso) (Kim, Sohn, & Xing, 2009) 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 (Archambeau & Bach, 2009) because it performed unstably in our experiments. Regarding the software implementation, we used the built-in Matlab routine for CCA and the code by (Sun et al., 2011) 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 the proposed SHML model based on the Gaussian process prior with the following ordinal or multinomial regression methods: (1) lasso for multinomial regression (Tibshirani, 1994), (2) elastic net for multinomial regression (Zou & Hastie, 2005), (3) sparse ordinal regression with the spike 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) (Chu & Ghahramani, 2005), and (7) Laplacian Support Vector Machine (LapSVM) (Melacci & Mikhail, 2011), 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 our model, 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.

5.1.3 Results

To estimate linkage (i.e., interactions) between $\mathbf{X}$ and $\mathbf{Z}$, we calculated the cross covariance matrix $\mathbf{G}\mathbf{H}^\top$. We then computed the precision and the recall based on the ground truth. The precision-recall curves are shown in Figure 3. Clearly, our method successfully recov-

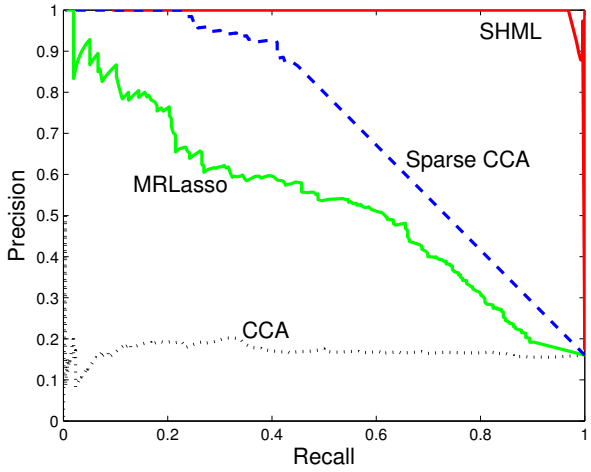
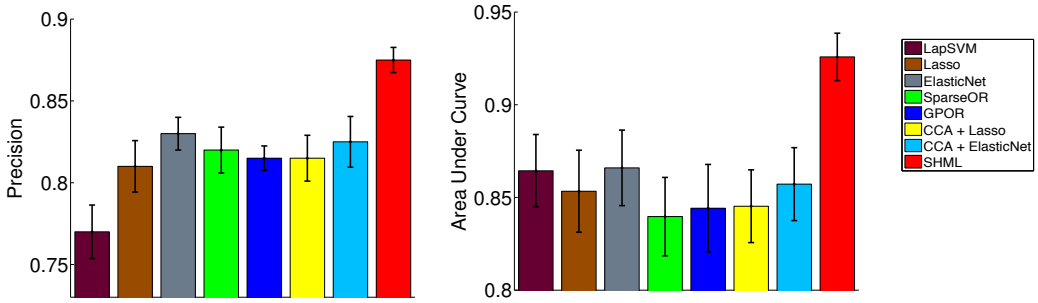


Figure 3: The precision-recall curves for association discovery.

ered 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 failing of CCA and sparse CCA may be due to the insufficient representation of all sources of data caused by using only one projection direction. The prediction accuracies on unknown $\mathbf{y}$ and their standard errors are shown in Figure 4a and the AUC and their standard errors are shown in Figure 4b. 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%.

5.2 Real-World Study on Alzheimer's Disease

Alzheimer's Disease is the most common form of dementia with about 30 million patients worldwide and payments for care are estimated to be $200 billion in 2012 (Alzheimer's



(a) Precision on simulation (b) AUC on simulation

Figure 4: The prediction results on simulated and real datasets. The results are averaged over 10 runs. The error bars represent standard errors.

Association, 2012). We conducted association analysis and diagnosis of AD based on a dataset from Alzheimer's Disease Neuroimaging Initiative(ADNI) [1]. The ADNI study is a longitudinal multisite observational study of elderly individuals with normal cognition, mild cognitive impairment, or AD. We applied the proposed method 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.

After removing missing values, the data set consists of 625 subjects including183 normal, 308 MCI and 134 AD cases, and each subject contains 924 SNPs and 328 MRI features. The selected SNPs are those top SNPs separating normal subjects from AD in ADNI. The MRI features measure the brain atrophies in different brain regions based on cortical thickness, surface areas or volumes, which are obtained from FreeSurfer software [2]. To test the diagnosis accuracy, we compared our method with the previously mentioned ordinal or multinomial regression methods. We employ the extended model with linkage disequilibrium priors, denoted as SHML-LD, to discover the associations.

We compare both SHML and SHML-LD with the state-of-the-art classification methods. And we used the 10-fold cross validation for each run to tune free parameters on the training data. To determine the dimension k for the latent features $\mathbf{U}$ in our method, we computed the variational lower bounds as an approximation to the model marginal likelihood (i.e., evidence), with various k values $\{10, 20, 40, 60\}$. We chose the value with the largest approximate evidence, which led to $k = 20$ (see Figure 5). Our experiments confirmed that with $k = 20$, our model achieved the highest prediction accuracy, demonstrating the benefit of evidence maximization.

As shown in Figure 6, 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 ($p < 0.05$).

1. `http://adni.loni.ucla.edu/`
2. `http://surfer.nmr.mgh.harvard.edu`

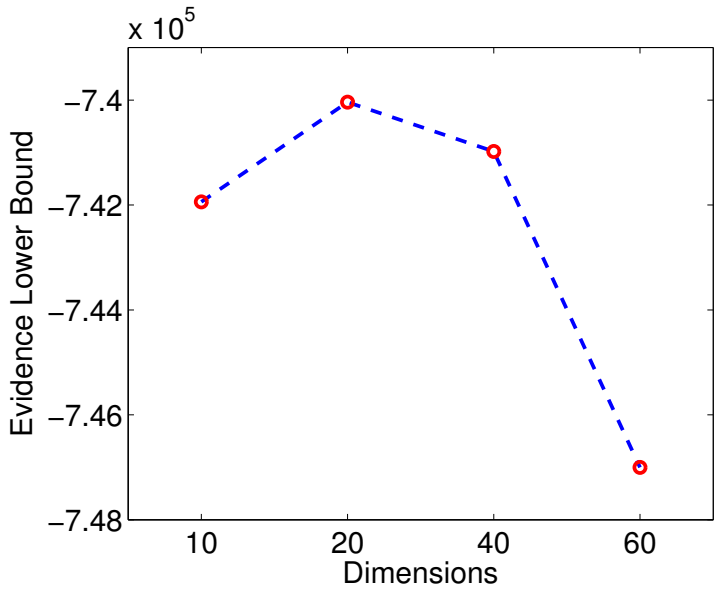


Figure 5: The variational lower bound for the model marginal likelihood.

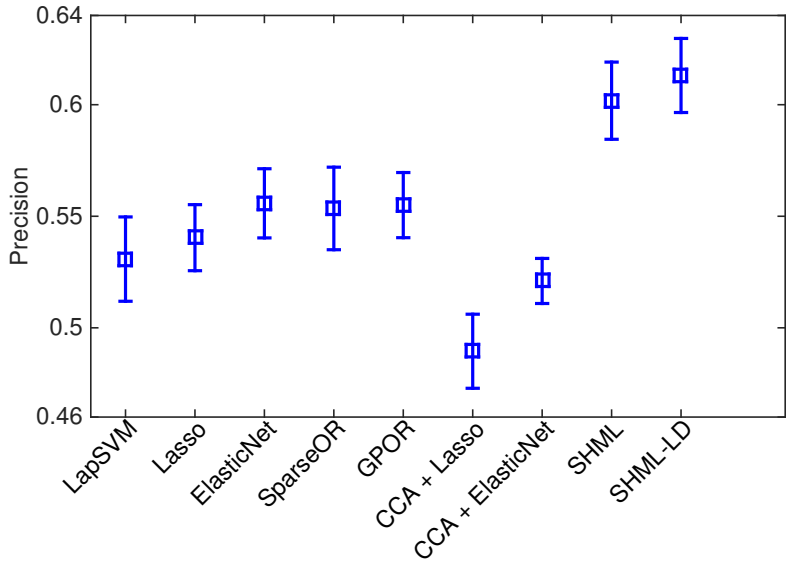
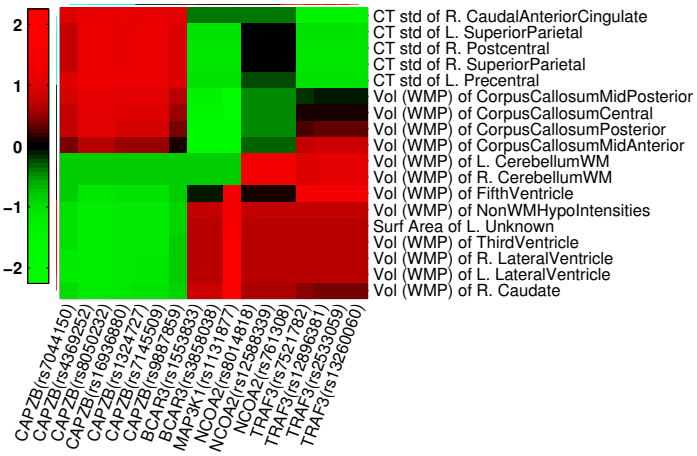
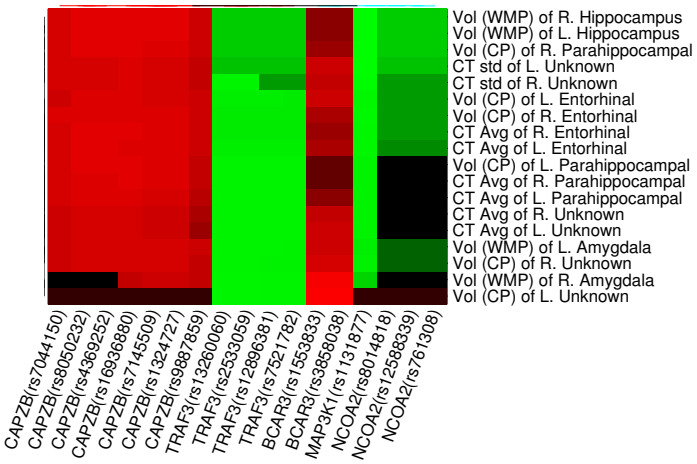


Figure 6: The prediction accuracy with standard errors on the real data.

We also examined the strongest associations discovered by our model. Firstly, 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



(a)



(b)

Figure 7: 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.

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. We also found that measurements of the same structure on the left and right sides 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.

Secondly, the analysis of associating genotype to AD prediction also 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 7, reveals interesting patterns 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, and MAP3K1 (mitogen-activated protein kinase kinase kinase 1) which have been studied more carefully in cancer research. The set of SNPs are associated with cingulate in negative directions, which is part of the limbic system and involves in emotion formation and processing. Compared with other structures such as temporal lobe, it plays a more important role in the formation of long-term memory. For example, the association between MAP3K1 and the caudate anterior cingulate cortex has been identified. Literature has shown that MAP3K1 is associated with biological processes such as apoptosis, cell cycle, chromatin binding and DNA binding[3], and cingulate cortex has been shown to be severely affected by AD (Jones et al., 2006). The strong association discovered in this work might indicate potential genetic effects in the atrophy pattern observed in this cingulate subregion.

6. Related Work

The proposed our model model is related to a broad family of probabilistic latent variable models, including probabilistic principle component analysis (Tipping & Bishop, 1999), probabilistic canonical correlation analysis (Bach & Jordan, 2005) and their extensions (Yu, Yu, Tresp, Kriegel, & Wu, 2006; Archambeau & Bach, 2009; Guan & Dy, 2009; Virtanen, Klami, & Kaski, 2011). They all learn a latent representation whose projection leads to the observed data. Recent studies on probabilistic factor analysis methods put more focus on the sparsity-inducing priors to the projection matrix. Among them, Guan and Dy (2009) used the Laplace prior, the Jeffrey's prior, and the inverse-Gaussian prior; Archambeau and Bach (2009) employed the inverse-Gamma prior; and Virtanen et al. (2011) used the Automatic Relevance Determination(ARD) prior. Despite their success, these sparsity-inducing priors have their own disadvantages – they confound the degree of sparsity with the degree of regularization on both relevant and irrelevant variables, while in practical settings there is little reason that these two types of complexity control should be so tightly bounded together. Although the inverse-Gaussian prior and the inverse-Gamma prior provide more flexibility of controlling the sparsity, they suffer from being highly sensitive to the controlling parameters and thus lead to unstable solutions. In contrast, our model adopts the spike and slab prior, which has been recently used in multi-task multiple kernel learning (Titsias & Lázaro-Gredilla, 2011), sparse coding (Goodfellow et al., 2012), and latent factor analysis (Carvalho, Chang, Lucas, Nevins, Wang, & West, 2008). Note that while our Beta priors over the selection indicators lead to simple yet effective variational updates, the hierarchical prior in the work of Carvalho et al.(2008) can better handle the selection uncertainty. Regardless what priors are assigned to the spike and slab models, they generally avoid the confounding issue by separately controlling the projection sparsity and the regularization effect over selected elements.

3. `https://portal.genego.com/`

SHML is also connected with many methods on learning from multiple sources or views (Hardoon, Leen, Kaski, & Shawe-Taylor, 2008). Multiview learning methods are often used to learn a better classifier for multi-label classification – usually in text mining and image classification domains – based on correlation structures among the training data and the labels (Yu et al., 2006; Virtanen et al., 2011; Rish, Grabarnik, Cecchi, Pereira, & Gordon, 2008). However, in medical analysis and diagnosis, we meet two separate tasks – the association discovery between genetic variations and clinical traits, and the diagnosis on patients. Our proposed SHML conducts these two tasks simultaneously: it employs the diagnosis labels to guide association discovery, while leveraging the association structures to improve the diagnosis. In particular, the diagnosis procedure in SHML leads to an ordinal regression model based on latent Gaussian process models. The latent Gaussian process treatment differentiates ours from multiview CCA models (Rupnik & Shawe-Taylor, 2010). Moreover, most multiview learning methods do not model the heterogeneous data types from different views, and simply treat them as continuous data. This simplification can degrate the predictive performance. Instead, based on a probabilistic framework , SHML uses suitable link functions to fit different types of data.

7. Conclusions

We have presented a new Bayesian multiview learning framework to simultaneously find key associations between data sources (i.e., genetic variations and phenotypic traits) and to predict unknown ordinal labels. We have shown that the model can also employ background information, e.g., the Linkage Disequilibrium information, via an additional graph Laplacian type of prior. Our proposed approach follows a generative model: it extracts a common latent representation which encodes the structural information within all the data views, and then generates data via sparse projections. The encoding of knowledge from multiple views via the latent representation makes it possible to effectively detect the associations with high sensitivity and specificity.

Experimental results on the ADNI data indicate that our model found biologically meaningful associations between SNPs and MRI features and led to significant improvement on predicting the ordinal AD stages over the alternative classification and ordinal regression methods. Despite the drawbacks of the proposed framework in slow training speed and requirement of careful tuning parameters, it has strong modeling power due to the Bayesian nature. Although we have focused on the AD study, we expect that our model, as a powerful extension of CCA, can be applied to a wide range of applications in biomedical research – for example, eQTL analysis supervised by additional labeling information.

Acknowledgments

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 investi-

gators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how to apply/ADNI Acknowledgement List.pdf.

This work was supported by NSF IIS-0916443, IIS-1054903, CCF-0939370, NSF China (Nos. 61572111, 61433014, 61440036), a 973 project of china (No.2014CB340401), a 985 Project of UESTC (No.A1098531023601041) and a Basic Research Project of China Central University (No. ZYGX2014J058).

Zenglin Xu and Shandian Zhe have equal contributions to this article. Yuan Qi is the Principle corresponding author.

Appendix A. Parameter Update for Continuous Data

The parameter β_{ij} in $Q(s_g^{ij})$ introduced in Section 3.1 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 $\tilde{l}_1^{ij} = \beta_{ij} + l_1$ and $\tilde{l}_2^{ij} = 1 - \beta_{ij} + l_2$. The parameters of the Gamma distribution $Q(\eta)$ are updated as $\tilde{r}_1 = r_1 + \frac{np}{2}$ and $\tilde{r}_2 = r_2 + \frac{1}{2}\mathrm{tr}(\mathbf{X}\mathbf{X}^\top) - \mathrm{tr}(\langle\mathbf{G}\rangle\langle\mathbf{U}\rangle\mathbf{X}^\top) + \frac{1}{2}\mathrm{tr}(\langle\mathbf{U}\mathbf{U}^\top\rangle\langle\mathbf{G}^\top\mathbf{G}\rangle)$.

The moments required in the above distributions are calculated as $\langle\eta\rangle = \frac{\tilde{r}_1}{\tilde{r}_2}$ and

$$\begin{aligned}
\langle\log(\pi_g^{ij})\rangle &= \psi(\tilde{l}_1^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij}),\\
\langle\log(1-\pi_g^{ij})\rangle &= \psi(\tilde{l}_2^{ij}) - \psi(\tilde{l}_1^{ij} + \tilde{l}_2^{ij}),\\
\langle\mathbf{G}^\top\mathbf{G}\rangle &= \sum_{i=1}^{p}\mathbf{\Omega}_i + \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top,\\
\langle\mathbf{G}\rangle &= [\boldsymbol{\lambda}_1, \ldots, \boldsymbol{\lambda}_p]^\top,
\end{aligned} \tag{23}$$

where $\psi(x) = \frac{\mathrm{d}}{\mathrm{dx}}\ln\Gamma(x)$.

Appendix B. Parameter Update for Ordinal Data

The variational distributions of $\mathbf{S}_h$ and $\mathbf{\Pi}_h$ introduced in Section 3.2 are given by

$$Q(\mathbf{S}_h) = \prod_{i=1}^{q}\prod_{j=1}^{k}\alpha_{ij}^{s_h^{ij}}(1-\alpha_{ij})^{1-s_h^{ij}}, \tag{24}$$

$$Q(\mathbf{\Pi}_h) = \prod_{i=1}^{q}\prod_{j=1}^{k}\mathrm{Beta}(\pi_{\mathrm{h}}^{\mathrm{ij}}|\tilde{\mathrm{d}}_1^{\mathrm{ij}}, \tilde{\mathrm{d}}_2^{\mathrm{ij}}), \tag{25}$$

where $\alpha_{ij} = 1/\big(1+\exp(\langle\log(1-\pi_h^{ij})\rangle - \langle\log(\pi_h^{ij})\rangle + \frac{1}{2}\log(\frac{\sigma_1^2}{\sigma_2^2}) + \frac{1}{2}\langle h_{ij}^2\rangle(\frac{1}{\sigma_1^2} - \frac{1}{\sigma_2^2}))\big)$, $\tilde{d}_1^{ij} = \alpha_{ij} + d_1$, $\tilde{d}_2^{ij} = 1 - \alpha_{ij} + d_2$, $\langle\mathbf{s}_h^i\rangle = [\alpha_{1i}, \ldots, \alpha_{qi}]^\top$, and $\langle h_{ij}^2\rangle$ is the i-th diagonal element in $\mathbf{\Lambda}_j$.

The required moments for updating the above distributions can be calculated as follows:

$$\begin{aligned}
\langle\log(\pi_h^{ij})\rangle &= \psi(\tilde{d}_1^{ij}) - \psi(\tilde{d}_1^{ij} + \tilde{d}_2^{ij}),\\
\langle\log(1-\pi_h^{ij})\rangle &= \psi(\tilde{d}_2^{ij}) - \psi(\tilde{d}_1^{ij} + \tilde{d}_2^{ij}),\\
\langle c_{ij}\rangle &= \bar{c}_{ij} - \frac{\mathcal{N}(b_{z_{ij}+1}|\bar{c}_{ij}, 1) - \mathcal{N}(b_{z_{ij}}|\bar{c}_{ij}, 1)}{\Phi(b_{z_{ij}+1} - \bar{c}_{ij}) - \Phi(b_{z_{ij}} - \bar{c}_{ij})},
\end{aligned}$$

where $\Phi(\cdot)$ is the cumulative distribution function of a standard Gaussian distribution. Note that in Equation (26), $Q(c_{ij})$ is a truncated Gaussian and the truncation is controlled by the observed ordinal data z_{ij}.

Appendix C. Parameter Update for Labels

The variational distributions of $\mathbf{s}_w$ and π_w in Section 3.3 are given by

$$Q(\mathbf{s}_w) = \prod_{i=1}^{k} \tau_i^{\mathbf{s}_w^i}(1-\tau_i)^{1-\mathbf{s}_w^i}, \tag{26}$$

$$Q(\boldsymbol{\pi}_w) = \prod_{i=1}^{k} \mathrm{Beta}(\pi_w^i; \tilde{e}_1^i, \tilde{e}_2^i), \tag{27}$$

where $\tau_i = 1/\big(1+\exp(\langle\log(1-\pi_w^i)\rangle - \langle\log(\pi_w^i)\rangle + \frac{1}{2}\langle w_i^2\rangle(\frac{1}{\sigma_1^2}-\frac{1}{\sigma_2^2}))\big)$, $\tilde{e}_1^i = \tau_i + e_1$ and $\tilde{e}_2^i = 1-\tau_i+e_2$.

The required moments for updating the above distributions can be calculated as follows:

$$\begin{aligned}
\langle\log(\pi_w^i)\rangle &= \psi(\tilde{e}_1^i) - \psi(\tilde{e}_1^i+\tilde{e}_2^i),\\
\langle\log(1-\pi_w^i)\rangle &= \psi(\tilde{e}_2^i) - \psi(\tilde{e}_1^i+\tilde{e}_2^i),\\
\langle f_i\rangle &= \bar{f}_i - \frac{\mathcal{N}(b_{y_i+1}|\bar{f}_i,1) - \mathcal{N}(b_{y_i}|\bar{f}_i,1)}{\Phi(b_{y_i+1}-\bar{f}_i) - \Phi(b_{y_i}-\bar{f}_i)}.
\end{aligned}$$

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.

Appendix D. Parameter Update for Latent Representation U

The required moments $\langle\mathbf{w}\mathbf{w}^\top\rangle$, $\langle\mathbf{G}^\top\mathbf{G}\rangle$ and $\langle\mathbf{H}^\top\mathbf{H}\rangle$ introduced in Section 3.4 are calculated as $\langle\mathbf{w}\mathbf{w}^\top\rangle = \boldsymbol{\Sigma}_w + \mathbf{m}\mathbf{m}^\top$, $\langle\mathbf{G}^\top\mathbf{G}\rangle = \sum_{i=1}^{p}\boldsymbol{\Omega}_i + \boldsymbol{\lambda}_i\boldsymbol{\lambda}_i^\top$ and

$$(\langle\mathbf{H}^\top\mathbf{H}\rangle)_{ij} = \begin{cases} \mathrm{trace}(\boldsymbol{\Lambda}_i + \boldsymbol{\gamma}_i\boldsymbol{\gamma}_i^\top) & i=j \\ \boldsymbol{\gamma}_i^\top\boldsymbol{\gamma}_j & i\neq j \end{cases}.$$

The other required moments have already been listed in the previous sections. The moments regarding $\mathbf{U}$ required in the updates of other variational distributions are $\langle\mathbf{U}\rangle = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, ...\boldsymbol{\mu}_n]$, $\langle\mathbf{U}\mathbf{U}^\top\rangle = \sum_{i=1}^{n}\boldsymbol{\mu}_i\boldsymbol{\mu}_i^\top + \boldsymbol{\Sigma}_i$, $\langle\tilde{\mathbf{u}}_i\rangle = [\mu_1^i, \mu_2^i, ...\mu_n^i]^\top$ and $\langle\tilde{\mathbf{u}}_i^\top\tilde{\mathbf{u}}_i\rangle = \sum_{j=1}^{n}(\mu_j^i)^2 + (\boldsymbol{\Sigma}_j)_{ii}$.

References

Alzheimer's Association (2012). 2012 facts and figures alzheimer's disease facts and figures. Tech. rep..

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

Bach, F., & Jordan, M. (2005). A probabilistic interpretation of canonical correlation analysis. Tech. rep., UC Berkeley.

Carvalho, C., Chang, J., Lucas, J., Nevins, J., Wang, Q., & West, M. (2008). High-dimensional sparse factor modeling: applications in gene expression genomics. *Journal of the American Statistical Association*, *103*(484), 1438–1456.

Chen, X., Liu, H., & Carbonell, J. (2012). Structured sparse canonical correlation analysis.. In *AISTATS'12*, Vol. 22, pp. 199–207.

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

Consoli, L., Lefevre, A., Zivy, M., de Vienne, D., & Damerval, C. (2002). QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Mol Biol.*, *48*(5), 575–581.

Daniela, M., & Tibshirani, R. (2009). Extensions of sparse canonical correlation analysis, with applications to genomic data. *Stat Appl Genet Mol Biol.*, *383*(1).

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

Gandhi, S., & Wood, N. (2010). Genome-wide association studies: the key to unlocking neurodegeneration?. *Nature Neuroscience*, *13*, 789–794.

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

Goodfellow, I., Couville, A., & Bengio, Y. (2012). Large-scale feature learning with spike-and-slab sparse coding. In *Proceedings of International Conference on Machine Learning*.

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

Hardoon, D., Leen, G., Kaski, S., & 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.

Hunter, D. (2012). Lessons from genome-wide association studies for epidemiology. *Epidemiology*, *23*(3), 363–367.

Jones, B. F., et al. (2006). Differential regional atrophy of the cingulate gyrus in Alzheimer disease: a volumetric MRI study. *Cereb. Cortex*, *16*(12), 1701–1708.

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

Kim, S., Sohn, K., & Xing, E. (2009). A multivariate regression approach to association analysis of a quantitative trait network. *Bioinformaics*, *25*(12), 204–212.

Lasserre, J., et al. (2006). Principled hybrids of generative and discriminative models. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 1, pp. 87–94.

Liu, J., Pearlson, G., Windemuth, A., Ruano, G., Perrone-Bizzozero, N., & Calhoun, V. (2009). Combining fMRI and SNP data to investigate connections between brain function and genetics using parallel ICA. *Hum Brain Mapp*, *30*(1), 241–255.

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

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

Mohamed, S., et al. (2012). Bayesian and L1 approaches for sparse unsupervised learning. In *Proceedings of International Conference on Machine Learning*.

Neal, R. M. (1996). *Bayesian Learning for Neural Networks*. Springer-Verlag New York, Inc.

Parkhomenko, E., Tritchler, D., & Beyene, J. (2007). Genome-wide sparse canonical correlation of gene expression with genotypes. *BMC Proc.*

Rish, I., Grabarnik, G., Cecchi, G., Pereira, F., & Gordon, G. (2008). Closed-form supervised dimensionality reduction with generalized linear models. In *Proceedings of International Conference on Machine Learning'08*, pp. 832–839.

Rupnik, J., & Shawe-Taylor, J. (2010). Multi-view canonical correlation analysis. In *Proceedings of SIG Conference on Knowledge Discovery and Mining'10*.

Slatkin, M. (2008). Linkage disequilibrium — understanding the evolutionary past and mapping the medical future. In *Nature Reviews Genetics*, Vol. 6, pp. 477–485.

Sun, L., Ji, S., & Ye, J. (2011). Canonical correlation analysis for multi-label classification: A least squares formulation, extensions and analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *33*(1), 194–200.

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

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

Titsias, M., & Lázaro-Gredilla, M. (2011). Spike and slab variational inference for multitask and multiple kernel learning. In *Advances in Neural Information Processing Systems'11*, pp. 2339–2347.

Virtanen, S., Klami, A., & Kaski, S. (2011). Bayesian CCA via group sparsity. In *Proceedings of International Conference on Machine Learning'11*, pp. 457–464.

Yu, S., Yu, K., Tresp, V., Kriegel, H., & Wu, M. (2006). Supervised probabilistic principal component analysis. In *Proceedings of SIG Conference on Knowledge Discovery and Mining'06*, pp. 464–473.

Yuan, M., & Lin, Y. (2007). Model selection and estimation in regression with grouped variables.. *Journal of the Royal Statistical Society, Series B*, *68*(1), 49–67.

Zhe, S., Xu, Z., Qi, Y., & Yu, P. (2014). Supervised heterogeneous multiview learning for joint association study and disease diagnosis. *Pacific Symposium on Biocomputing*, *19*.

Zhe, S., Xu, Z., Qi, Y., & Yu, P. (2015). Sparse bayesian multiview learning for simultaneous association discovery and diagnosis of alzheimer's disease. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pp. 1966–1972.

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