Supervised Heterogeneous Multiview Learning for Joint Association Study and Disease Diagnosis

Shandian Zhe
Department of CS
Purdue University
szhe@purdue.edu

Zenglin Xu
Department of CS
Purdue University
xu218@purdue.edu

Yuan Qi
Departments of CS and Statistics
Purdue University
alanqi@cs.purdue.edu

Peng Yu
Eli Lilly and Company
yu_peng_py@lilly.com

April 15, 2013

Abstract

Given genetic variations and various phenotypical traits, such as Magnetic Resonance Imaging (MRI) features, we consider two important and related tasks in biomedical research: i)to select genetic and phenotypical markers for disease diagnosis and ii) to identify associations between genetic and phenotypical data. These two tasks are tightly coupled because underlying associations between genetic variations and phenotypical features contain the biological basis for a disease. A variety of models have been applied for disease diagnosis or association studies (*e.g.*, eQTL analysis), separately. To unify these two tasks, we present a new sparse Bayesian approach for joint association study and disease diagnosis. In this approach, common latent features are extracted from different data sources based on sparse projection matrices and used to predict multiple disease severity levels based on Gaussian process ordinal regression; in return, the disease status is used to guide the discovery of relationships between the data sources. To learn the model from data, we develop an efficient variational expectation maximization algorithm. We apply our approach to an imaging genetics dataset for the study of Alzheimer's Disease (AD). Our method identifies biologically meaningful relationships between genetic variations, MRI features, and AD status, and achieves significantly higher accuracy for predicting ordinal AD stages than the competing methods.

1 Introduction

Recent advances in biomedical research have provided new opportunities to study diseases – for example, Alzheimer's disease (AD), the most common neurodegenerative disorder – from multiple data sources. For example, one data source contains genetic variations, such as single nucleotide polymorphisms (SNPs), which can help us understand the genetic basis of diseases. Another data source can be molecular and clinical phenotypes, such as Magnetic Resonance Imaging (MRI) data, which can reveal important phenotypic changes in patients.Finding associations between different data sources can reveal unknown biological relationships and has a wide range of applications in computational biology Consoli and

others [2002], epidemiology Hunter [2012], and imaging genetics Liu and others [2009]. In addition to the genotypes and phenotypic traits, we have valuable *labeled* information about disease stages from patient medical records. Thus we face a new data analysis setting where the objective is two-fold: i) finding associations between different data sources and ii) selecting relevant (groups of) features from all the sources to predict ordinal disease stages.

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]. These approaches have been widely used in expression quantitative trait locus (eQTL) analysis. For disease diagnosis based on high dimensional biomarkers, popular approaches include lasso Tibshirani [1994], elastic net Zou and Hastie [2005], and group lasso Yuan and Lin [2007]. Here we treat genotypes or phenotypes as predictors (i.e., biomarkers) and the disease status as the response in a linear regression or classification setting.

Despite their wide success in many applications, these approaches are limited by the following factors:

- Most association studies neglect the supervision from the disease status. 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.
- 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.

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. In the new approach, 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 based on Gaussian process ordinal regression (See Section 2). To enforce sparsity in projection matrices, we assign spike and slab priors George and McCulloch [1997] over them; these priors have been shown to be more effective than l_1 penalty to learn sparse projection matrices Goodfellow *et al.* [2012]. The sparse projection matrices not only reveal critical interactions between the different data sources but also identify *groups* of biomarkers in data relevant to disease status. 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. Hence we name this new method Supervised Heterogeneous Multiview Learning (SHML).

In Section 3, we apply SHML to an AD study. AD accounts for 60-80% of age-related dementia cases – one in eight older Americans has AD – and there is no cure for AD till now. Although AD studies have attracted a lot of attention from both academia and industry

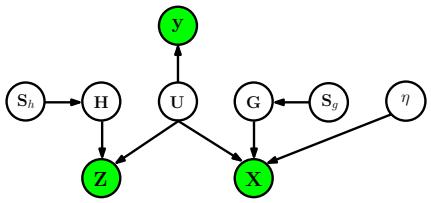

Figure 1: The graphical model of Supervised Heterogeneous Multiview Learning, where $\mathbf{X}$ is the continuous view, $\mathbf{Z}$ is the ordinal view, and $\mathbf{y}$ are the labels.

Zhou *et al.* [2012], to our best knowledge, our paper presents the first (supervised) study to uncover associations between genotypes and phenotypic traits relevant to AD. Our results on Alzheimer's Disease Neuroimaging Initiative (ADNI) data show that SHML achieves highest prediction accuracy among all the competing methods. Furthermore, SHML finds biologically meaningful predictive relationships between SNPs, MRI features, and AD status.

2 Model

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

To link two data sources $\mathbf{X}$ and $\mathbf{Z}$ together, we introduce common latent features $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_n]$ and assume $\mathbf{X}$ and $\mathbf{Z}$ are generated from $\mathbf{U}$ by sparse 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 and its dimension k is estimated by evidence maximization. In a Bayesian framework, we give a Gaussian prior over $\mathbf{U}$, $p(\mathbf{U}) = \prod_i \mathcal{N}(\mathbf{u}_i|\mathbf{0}, \mathbf{I})$, and specify the rest of the model (see Figure 1) as follows:

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

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

  where $\mathbf{G} = [\mathbf{g}_1, \mathbf{g}_2, ...\mathbf{g}_p]^\top$ is a $p \times k$ projection matrix, $\mathbf{I}$ is an identity matrix, and $\eta^{-1}\mathbf{I}$ is the precision matrix of the Gaussian distribution. We assign a Gamma prior over η, $p(\eta|r_1, r_2) = \text{Gamma}(\eta|r_1, r_2)$ where r_1 and r_2 are the hyperparameters and set to be 10^{-3} in our experiments.

- **Ordinal data.** For an ordinal observation $z \in \{0, 1, \ldots,$
  $R - 1\}$ its value is decided by which region an auxiliary variable c falls in

$$-\infty = b_0 < b_1 < \ldots < b_R = \infty.$$

If c falls in $[b_r, b_{r+1})$, z is set to be r. For the AD study, the SNPs $\mathbf{Z}$ takes 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 $p(c_{ij}|\mathbf{h}_i, \mathbf{u}_j) = \mathcal{N}(c_{ij}|\mathbf{h}_i^\top \mathbf{u}_j, 1)$, and $p(z_{ij}|c_{ij}) = \sum_{r=0}^{2} \delta(z_{ij} = r)\delta(b_r \leq c_{ij} < b_{r+1})$. Here $\delta(a) = 1$ if a is true and $\delta(a) = 0$ otherwise.

- **Labels.** The disease statuses $\mathbf{y}$ are ordinal variables too. To generate $\mathbf{y}$, we use a Gaussian process ordinal regression model Chu and Ghahramani [2005] based the latent representation $\mathbf{U}$,

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

where

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

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

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

- **Sparse Priors.** Because we want to identify a few critical interactions between different data sources, we use spike and slab prior distributions George and McCulloch [1997] to sparsify the projection matrices $\mathbf{G}$ and $\mathbf{H}$. Specifically, we use a $p \times k$ matrix $\mathbf{S}_g$ to represent the selection of elements in $\mathbf{G}$: if $s_{ij} = 1$, g_{ij} is selected and follows a Gaussian prior distribution with variance σ_1^2; if $s_{ij} = 0$, g_{ij} is not selected and forced to almost zero (i.e., sampled from a Gaussian with a very small variance σ_2^2). Specifically, we have the following prior over $\mathbf{G}$:

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

where

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

Here 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. The distribution of $\mathbf{H}$ can be similarly obtained.

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

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

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 variational Bayesian Expectation Maximization (VB-EM) approach Beal [2003]. Due to the space limitation, we do not show the detailed updates.

3 Experiments

We conducted association analysis and diagnosis of AD 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 impairment, or AD. AD is the most common form of dementia with about 30 million patients worldwide and In this analysis, we used SHML to study the associations of genotypes and brain atrophy measured by MRI and to predict the subject status (normal vs MCI vs AD). Note that the labels are ordinal since the three states represent increasing severity levels of the dementia.

The dataset was downloaded from `http://adni.loni.ucla.edu/`. After removing missing data, it consists 618 subjects (183 normal, 308 MCI and 134 AD), and for each patient, there are 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.

We compared SHML with the alternative methods on accuracy of predicting whether a subject is in the normal or MCI or AD condition. We randomly split the dataset into 556 training and 62 test samples 10 times and ran all the competing methods on each partition.

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

The accuracies for predicting unknown labels $\mathbf{y}$ and their standard errors are shown in Figure 2. 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%.

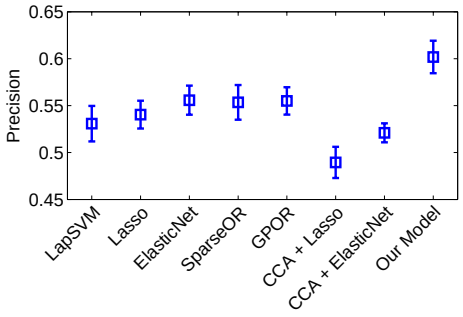

Figure 2: The prediction accuracy with standard errors on the real data.

Biclustering of the gene-MRI association reveals interesting pattern in terms of the relationship between genetic variations and brain atrophy measured by structural MRI. For example, the top ranks 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. One of the genes associated with this set of SNPs is MATP (microtubule-associated protein tau), which codes the tau gene that are associated closely with the AD.

These findings reveal strong association between MATP gene and atrophy in the memory-related brain regions. Moreover, the same set of SNPs are also highly associated with cingulate, but in an opposite direction. These results indicate an opposite effect of genotype to the cingulate region, 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.

In summary, SHML discovered synergistic predictive relationships between brain atrophy, genetic variations and the disease status.

References

M. Beal. *Variational Algorithms for Approximate Bayesian Inference*. Ph.d. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

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

L. Consoli et al. QTL analysis of proteome and transcriptome variations for dissecting the genetic architecture of complex traits in maize. *Plant Mol Biol.*, 48(5):575–581, 2002.

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

I. Goodfellow, A. Couville, and Y. Bengio. Large-scale feature learning with spike-and-slab sparse coding. In *ICML'12*. 2012.

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

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

J. Liu et al. Combining fMRI and SNP Data to Investigate Connections Between Brain Function and Genetics Using Parallel ICA. *Hum Brain Mapp.*, (1):1–30, 2009.

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

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

J. Zhou, J. Liu, V. Narayan, and J. Ye. Modeling disease progression via fused sparse group lasso. In *KDD'12*, pages 1095–1103, 2012.

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