Joint Generative Modeling of Imaging and Genetics

Nematollah K. Batmanghelich[1], Adrian V. Dalca[1] Mert R. Sabuncu[2], and Polina Golland[1] for the ADNI*

[1] Computer Science and Artificial Intelligence Laboratory, MIT, Cambridge, MA,
[2] Martinos Center for Biomedical Imaging, Charlestown, MA
{kayhan,adalca,msabuncu,polina}@csail.mit.edu

Abstract. We propose a unified Bayesian framework for detecting genetic variants associated with a disease while exploiting image-based features as an intermediate phenotype. Traditionally, imaging genetics methods comprise two separate steps. First, image features are selected based on their relevance to the disease phenotype. Second, a set of genetic variants are identified to explain the selected features. In contrast, our method performs these tasks simultaneously to ultimately assign probabilistic measures of relevance to both genetic and imaging markers. We derive an efficient approximate inference algorithm that handles high dimensionality of imaging genetic data. We evaluate the algorithm on synthetic data and show that it outperforms traditional models. We also illustrate the application of the method on ADNI data.

Keywords: Imaging Genetics, Bayesian Models, Variational Inference, Probabilistic Graphical Model

1 Introduction

In this paper, we propose a generative probabilistic model for genetic variants associated with a disease using imaging data as an intermediate phenotype. The search for genetic variants that increase the risk of a particular disorder is one of the central challenges in medical research, and has been traditionally performed via genome-wide association studies (GWAS). Such studies examine each genetic marker and its correlation with the incidence of the disease independently of all other genetic markers in the study. However, some variants may have a weak but cumulative effect that cannot be identified by traditional GWAS analysis [12]. Imaging genetics introduces imaging-based biomarkers as a promising intermediate phenotype (i.e., endo-phenotype) between genetic variants and diagnosis. Imaging provides a rich quantitative characterization of disease and promises to aid in identifying genetic variations that are correlated with the clinical variables [1,17]. Furthermore, multivariate analysis using

* Data used in preparation of this article were obtained from the Alzheimers 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

imaging endo-phenotypes promises to stratify the population in more informative ways than the binary diagnosis. A commonly used approach in imaging genetics is to isolate image-based features affected by the disease, and then identify the relevant genetic markers that explain the observed image variations. In this work, we jointly model image-based phenotypes and clinical indicators to identify genetic variants associated with the disorder.

Imaging genetics presents numerous challenges in clinical studies due to the relatively small number of subjects and extremely high dimensionality of images (hundreds of thousands of voxels) and genetic data (millions of single nucleotide polymorphisms (SNPs)). To address the problem of high dimensionality and small sample size, earlier algorithms considered only a few imaging candidates (voxels, regions, or other biomarkers) or only a few genetic markers in the analysis [5, 15]. The reduced joint dataset is then analyzed in a univariate testing framework, where each pair of a candidate genetic variant and an imaging biomarker is tested for association via a standard statistical test. Examples include using activation maps of the prefrontal cortex to find SNPs associated with schizophrenia [15], and searching for changes of gray matter volume correlated with the Alzheimer's Disease risk factor APOE gene [5].

More recently, genome-wide voxel-wise analysis has been demonstrated using univariate methods [18]. Unfortunately, massive univariate analysis has several limitations. Due to multiple comparisons, a corrected conservative significance level is selected to limit the false positive rate, but this also dramatically reduces the power of the test. Moreover, the univariate methods are unlikely to identify weaker variants that jointly create an additive effect.

Multivariate techniques aim to overcome shortcomings of univariate analysis [9, 20]. A common approach is to use multivariate regression combined with regularization to extract a sparse set of coefficients for correlated genetic variants and image features. For example, low rank representations can be approximated via sparse reduced rank regression (sRRR) [19, 20], Partial Least Squares (PLS) [9] or Canonical Correlation Analysis (CCA) [9]. Unfortunately, these unsupervised methods do not use the subject class label (e.g., diagnosis) directly, and thus the detected genetic markers and image features are not immediately related to the disease of interest. The image features relevant to the disease are identified separately from modeling the relationship between the genetic and imaging data. For example, sRRR has been demonstrated using brain regions pre-selected for Alzheimer's disease (AD) via Linear Discriminant Analysis [19]. In contrast, we model and estimate relevant genetic variants in the context of a particular disease. Our method is applicable to any set of image biomarkers, such as anatomical regions, tissue appearance, or functional measures. We are motivated by applications to the AD and use local measures of atrophy as image features.

Our model includes a common assumption of genetic studies that only a small set of SNPs is associated with any particular disease. This subset of genetic markers induces variation in certain image-based features, and a subset of these measures exhibits changes that are discriminative with respect to the dis-

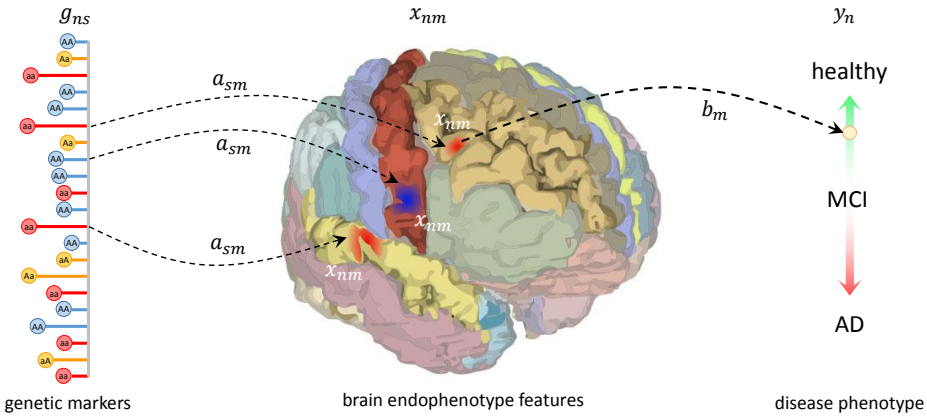

Fig. 1: A schematic illustration of the relationship between genetic, imaging and clinical measures in our model.

ease phenotype. Therefore, if a brain region is irrelevant to the target disease, it is ignored even if its measures are highly correlated with some genetic variants.

In the remainder of the paper, we define a generative model for the relationship among genetic, imaging and disease measures, derive an efficient inference algorithm to identify relevant brain regions and genetic loci, and demonstrate the method on synthetic data and the ADNI study [13]. We show that our algorithm outperforms standard univariate and regression analysis for genetic variant detection on synthetic data and yields promising results in the real clinical study.

2 Model

Our model structure is illustrated schematically in Fig.1. We are motivated by anatomical brain studies, but the analysis applies to any biomarkers derived from images.

Let y_n be the disease phenotype (0 or 1) for subject n in the study ($1 \leq n \leq N$). Let $\mathbf{x}_n$ and $\mathbf{g}_n$ be vectors of M imaging biomarkers (features) and S genetic markers (SNPs) for subject n, respectively. We capture the overall process via two coupled regression models: a logistic regression predicts class label y_n from image features $\mathbf{x}_n$; a ridge regression associates genetic variants $\mathbf{g}_n$ with image features $\mathbf{x}_n$. The graphical model in Fig.2 presents the relationships among variables of the model. All variables are summarized in Table 1. Below, we first define the relationship between imaging features and the disease phenotype and then specify the generative model for the relationship between SNPs and image features. Note that we do not model a direct link between genetic variants and disease label, but it is captured indirectly through image features.

2.1 From Imaging Features to Disease Phenotype

We adopt a Bayesian model based on logistic regression for predicting binary class label y_n from image features $\mathbf{x}_n$ [2]:

$$p(y_n|\boldsymbol{\eta}, \mathbf{x}_n) = \left[\psi(\boldsymbol{\eta}^T \mathbf{x}_n)\right]^{y_n} \left[1 - \psi(\boldsymbol{\eta}^T \mathbf{x}_n)\right]^{1-y_n}, \quad (1)$$

Model Variables	
x_{nm}	Image feature m in subject n.
g_{ns}	Genetic variant s in subject n.
y_n	Disease phenotype (class label) of subject n: 0 - healthy, 1 - diseased.
η_m	Regression coefficient for image feature m in the imaging part of the model.
$b_m \in \{0,1\}$	Indicator variable that selects image feature m.
$a_{sm} \in \{0,1\}$	Indicator variable that selects SNP s for modeling image feature m.
v_{sm}	Regression coefficient for SNP s for modeling feature m.
β	Prior probability for selecting image features.
α	Prior probability for selecting genetic variants.
σ_η^2	Variance of η_m.
σ_0^2	Variance of noise in the genetic to image regression.
Variational Variables	
ρ_m	Probability of selecting feature m.
τ_s	Probability of selecting SNP s.
ξ_n	Tightness of lower bound for the logistic function.
ν_m, ς_m	Imaging parameters for feature m.
$\vartheta = \{\mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\xi}, \boldsymbol{\nu}, \boldsymbol{\varsigma}\}$	Set of variational parameters that we optimize when fitting the model.

Table 1: Notation and variables used throughout the paper.

where $\psi(a) = \frac{1}{1+e^{-a}}$ is the logistic function and $\boldsymbol{\eta} \in \mathbb{R}^M$ are the regression coefficients that we treat as latent random variables. Similar to prior work [3], we propose to use a *spike-and-slab* prior to promote sparse solutions for the regression coefficients $\boldsymbol{\eta}$ [7, 14]:

$$p(\boldsymbol{\eta}; \beta, \sigma_\eta^2) = \prod_{m=1}^{M} \left[(1-\beta)\delta(\eta_m) + \beta \mathcal{N}(\eta_m; 0, \sigma_\eta^2)\right],$$

where $\delta(\cdot)$ is the Delta Dirac distribution concentrated at 0, parameter β controls sparsity ($0 \leq \beta \leq 1$), and $\mathcal{N}(\cdot; \mu, \sigma^2)$ is a Gaussian distribution with mean μ and variance σ^2. In a deterministic regression context, one can view the spike-and-slab prior as a combination of ℓ_0 and ℓ_2 norms for regularization. We find it convenient to introduce a latent Bernoulli random variable b_m that selects the regime for the regression coefficient η_m:

$$p(b_m) = \beta^{b_m}(1-\beta)^{1-b_m}, \qquad p(\eta_m | b_m; \sigma_\eta^2) = \begin{cases} \delta(\eta_m), & \text{if } b_m = 0, \\ \mathcal{N}(\eta_m; 0, \sigma_\eta), & \text{if } b_m = 1. \end{cases} \quad (2)$$

2.2 From Genetics Variants to Imaging Features

In modeling the relationship between genetics and imaging, we treat image features relevant for disease prediction differently from all other image features. If feature m is relevant for disease prediction (i.e., $b_m = 1$), variations in the values of this feature are explained by a sparse subset of the genetic variants $\mathbf{g}_n \in \mathbb{R}^S$. We define $\mathbf{a}_m \in \{0,1\}^S$ to be a vector of latent Bernoulli random variables that specify a subset, or *mask*, of relevant genetic markers that affect feature m, and arrive at the second regression component of our model:

$$x_{nm} = \sum_{s=1}^{S} (a_{sm} v_{sm}) g_{ns} + \epsilon_{nm} = \langle \mathbf{a}_m \odot \mathbf{v}_m, \mathbf{g}_n \rangle + \epsilon_{nm}, \quad (3)$$

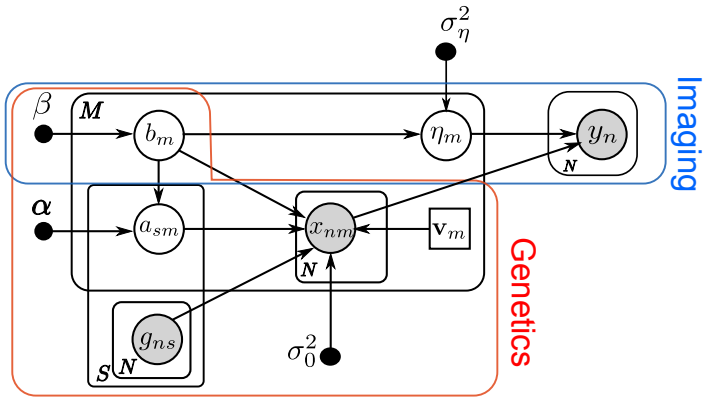

Fig. 2: Graphical representation of the generative model. Hollow circles denote random variables, solid circles represent hyper-parameters, and shaded circles represent observed variables. The rectangle containing $\mathbf{v}_m$ represents deterministic variables to be estimated. The plates indicate conditionally independent instantiations.

where $\mathbf{v}_m$ is the vector of regression coefficients, $\epsilon_{nm} \sim \mathcal{N}(\cdot; 0, \sigma_0^2)$ is the noise in the image feature m in subject n, and $\langle \cdot, \cdot \rangle$ and $\odot$ denote the inner and element-wise products, respectively. While an obvious modeling choice for regression coefficients $\{v_{sm}\}$ would be to treat them as latent random variables with a spike-and-slab prior, the large number of such variables $(S \times M)$ makes it computationally intractable. We therefore model regression coefficients $\{v_{sm}\}$ as unknown but deterministic variables.

If image feature m is irrelevant for predicting disease (i.e., $b_m = 0$), we do not model genetic contributions, and assign the probability mass uniformly between the observed feature values, i.e., $p(x) = \frac{1}{N}\delta(x - x_{nm})$. Furthermore, we set $a_{sm} = 0$ with probability 1 for all s.

Combining the two regimes, we obtain the genetic selection prior:

$$p(a_{sm}|b_m; \alpha) = \begin{cases} \delta(a_{sm}), & \text{if } b_m = 0, \\ \alpha^{a_{sm}}(1-\alpha)^{1-a_{sm}}, & \text{if } b_m = 1, \end{cases} \tag{4}$$

and the image feature likelihood:

$$p(x_{nm}|b_m, \mathbf{a}_m, \mathbf{g}_n; \mathbf{v}_m, \sigma_0^2) = \begin{cases} 1/N, & \text{if } b_m = 0, \\ \mathcal{N}(x_{nm}; \langle \mathbf{a}_m \odot \mathbf{v}_m, \mathbf{g}_n \rangle, \sigma_0^2), & \text{if } b_m = 1. \end{cases} \tag{5}$$

2.3 Complete model

We define $\mathcal{Z} = \{\boldsymbol{\eta}, \mathbf{b}, \mathbf{A}\}$ to be the set of latent variables, $\mathcal{D} = \{\mathbf{X}, \mathbf{y}\}$ to be the set of data variables that we model, and $\pi = \{\sigma_\eta^2, \sigma_0^2, \alpha, \beta\}$ to be the set of hyper-parameters. Here $\mathbf{y} = [y_1; ...; y_N]$, and $\mathbf{X} = [\mathbf{x}_1; ...; \mathbf{x}_N]$. Combining the elements of the model in Eqs. (1)-(5), we construct the joint distribution of the hidden

variables $\mathcal{Z}$ and modeled variables $\mathcal{D}$ given genetic markers $\mathbf{G} = [\mathbf{g}_1; ...; \mathbf{g}_N]$:

$$p(\mathcal{D}, \mathcal{Z}|\mathbf{G}; \mathbf{V}, \pi) = \prod_{n=1}^{N} p(y_n|\boldsymbol{\eta}, \mathbf{x}_n) \prod_{m=1}^{M} p(b_m)p(\eta_m|b_m)p(x_{nm}|b_m, \mathbf{a}_m, \mathbf{g}_n; \mathbf{v}_m) \prod_{s=1}^{S} p(a_{sm}|b_m).$$

3 Inference

Our goal is to compute the posterior probability $p(\mathcal{Z}|\mathcal{D}; \mathbf{G}, \mathbf{V}, \pi)$ of the latent variables that summarizes genetic and imaging influences in our model. Because of coupling of variables in the joint model, computing the posterior distribution is intractable, necessitating approximation via sampling or variational methods. Due to the amount of data and its dimensionality, sampling is computationally impractical. We therefore derive a Variational Bayes approximation [2] that estimates the lower bound for the log-likelihood $p(\mathcal{D}; \mathbf{G}, \pi)$ and seeks distribution q that minimizes the cost functional:

$$F(q) = \int q(\mathcal{Z}) \ln \frac{p(\mathcal{D}, \mathcal{Z}|\mathbf{G}; \mathbf{V}, \pi)}{q(\mathcal{Z})} d\mathcal{Z}. \tag{6}$$

The optimal distribution q provides an approximation to the posterior distribution $p(\mathcal{Z}|\mathcal{D}; \mathbf{G}, \pi)$ [2]. We choose a factorization for the distribution q that captures most model assumptions and yet is computationally tractable:

$$q(\boldsymbol{\eta}, \mathbf{b}, \mathbf{A}) = \prod_{m=1}^{M} q(b_m)q(\eta_m|b_m) \prod_{s=1}^{S} q(a_{ms}|b_m), \tag{7}$$

where:

$$\begin{aligned}
&q(b_m) = \rho_m^{b_m}(1-\rho_m)^{1-b_m}, \\
&q(\eta_m|b_m) = \begin{cases} \delta(\eta_m), & \text{if } b_m = 0, \\ \mathcal{N}(\eta_m; \nu_m, \varsigma_m), & \text{if } b_m = 1, \end{cases} \\
&q(a_{sm}|b_m) = \begin{cases} \delta(a_{sm}), & \text{if } b_m = 0, \\ \tau_s^{a_{sm}}(1-\tau_s)^{1-a_{sm}}, & \text{if } b_m = 1. \end{cases}
\end{aligned} \tag{8}$$

Variational parameters ρ_m, ν_m, ς_m and τ_s of the approximating distribution q define the optimization space. In this formulation, the estimate of τ_s is interpreted as relevance of the genetic variant s. The estimate of ρ_m provides a measure of relevance for image feature m. We define $\{\boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\varsigma}\}$ to be the set of all parameters $\tau_s, \rho_m, \nu_m, \varsigma_m$.

Given the parametrization above, all terms in the cost function $F(q)$ can be optimized analytically, except for the logistic regression term $p(y_n|\boldsymbol{\eta}, \mathbf{x}_n)$. For this term, we employ the variational treatment [8] that leads to improved accuracy over Laplace approximation [2] and has been successfully used in prior work [3]. Specifically, we replace the logistic function with its lower bound:

$$\psi(x_n) \geq \psi(\xi_n) \exp\left\{\frac{1}{2}(x_n - \xi_n) + \frac{1}{2\xi_n}(\psi(\xi_n) - \frac{1}{2})(x_n^2 - \xi_n^2)\right\}, \tag{9}$$

where ξ_n controls the tightness of the lower bound for subject n and should be optimized.

We define $\vartheta = \{\mathbf{V}, \boldsymbol{\tau}, \boldsymbol{\rho}, \boldsymbol{\nu}, \boldsymbol{\varsigma}, \boldsymbol{\xi}\}$ to be the full set of parameters of distribution q, where $\mathbf{V}$ and $\boldsymbol{\xi}$ are deterministic parameters of the model, and the rest are parameters of q. Using Eqs. (7)-(9), we can maximize $F(q) = F(\vartheta)$ by updating elements of the variational parameter vector $\boldsymbol{\vartheta}$. We omit the derivations due to space constraints, but summarize the resulting updates in Appendix A.

Every update iteration reduces the cost function $F(\vartheta)$, which in turn brings q closer to the posterior distribution $p(\mathcal{Z}|\mathcal{D}, \mathbf{G}; \mathbf{V}, \pi)$.

Our imaging genetics regression bears resemblance to previously demonstrated sRRR regression [20] that considers $\mathbf{X} = \mathbf{GV}$. Our update for $\mathbf{V}$ can be viewed as a solution of a system of linear equations:

$$\mathbf{X} = \left(\mathbf{G} + (\mathbf{G^T})^{\dagger} diag\left(\frac{1-\boldsymbol{\tau}}{\boldsymbol{\tau}}\right)\right)\mathbf{V},$$

where † indicates a pseudo-inverse, and the second term $diag(\frac{1-\boldsymbol{\tau}}{\boldsymbol{\tau}})$ weighs the SNPs based on their importance. We do not impose rank or sparsity constraints on the regression coefficients matrix $\mathbf{V}$, although they can be added in a fashion similar to [20].

4 Results

We evaluate our model on synthetic data using univariate tests and the sRRR method [20] as baseline algorithms. We also illustrate our method on the ADNI dataset, where we recover several top SNPs associated with the risk of AD.

4.1 Synthetic Data

We generate synthetic data to match a realistic scenario as much as possible. In this section, minor allele frequency (MAF) refers to the frequency of the less common allele in the population at a particular genetic location. A genetic marker (or SNP) g_{ns} is represented by the count of minor alleles at location s in subject n, i.e., $g_{ns} \in \{0, 1, 2\}$. We employ the widely used population genetics software package PLINK [16] to simulate 1,020 SNPs with a minor allele frequency uniformly sampled from an interval [0.05, 0.95], for 400 healthy subjects and 400 patients. For SNPs relevant to the disease, the heterozygote odds ratio is defined as the ratio of patients to controls with $g_{ns} = 1$, normalized by the same ratio for $g_{ns} = 0$. Similarly, one can define the homozygote odds ratio. These ratios control the disease risk in the patient population. The simulated SNPs are split into three sets:

- Set $\mathcal{G}_1$ includes 20 disease causative SNPs that affect selected areas of simulated images. The odds ratio is set to 1.125 for heterozygote SNPs, with a multiplicative homozygote risk. Other odds ratios yield similar results (we tested 1.0625 to 1.5, not shown due to space constraints).
- Set $\mathcal{G}_2$ includes 20 SNPs that are *irrelevant* to the disease (i.e., odds ratio is 1) but affect other areas in simulated images.

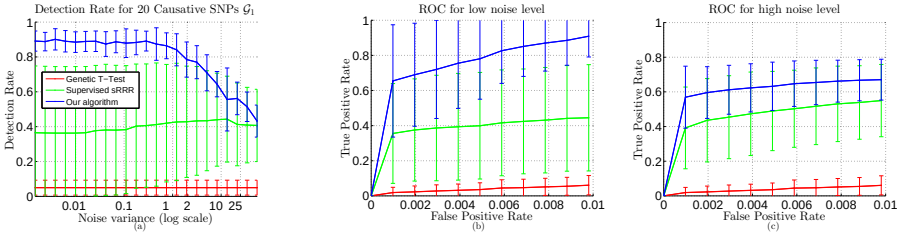

Fig. 3: Summary of results. (a) Detection rates for our algorithm (blue), the supervised sRRR pipeline (green), and the genetic t-test (red) as a function of image noise for causative SNPs in $\mathcal{G}_1$ at a false positive rate of 1%. (b,c) ROC curves for low ($\sigma^2_{noise} = 1.2$) and high ($\sigma^2_{noise} = 21.4$) noise levels are shown up to the selected false positive threshold of 1%.

- Set $\mathcal{G}_3$ includes 980 *null* SNPs that are independent of both label and images.

Based on the class labels and the genetic variants, we generate image voxels, organized in several sets:

- Voxels in set $\mathcal{I}_1$ are affected by causative SNPs ($\mathcal{G}_1$), and thus are indirectly associated with the disease. These voxels are separated into three regions. Voxel intensity in this set is correlated with genetics:

$$c^r_{nk} = \mathbf{w}^T_r \mathbf{g}^{\mathcal{G}_1}_n + \epsilon^r_{nk}, \qquad 1 \leq r \leq 3, \tag{10}$$

  where c^r_{nk} is the intensity value of voxel k in region r for subject n. The region weights $\mathbf{w}_r$ are drawn from a normal distribution $\mathcal{N}(\cdot; 0, 1)$, and ϵ^r_{kn} is Gaussian noise. Our experiments explore a range of values for the noise variance σ^2_{noise}.
- Voxels in set $\mathcal{I}_2$ are determined by non-causative SNPs $\mathcal{G}_2$, and thus are irrelevant to disease. We dedicate one region to this category:

$$c^4_{nk} = \mathbf{w}^T_4 \mathbf{g}^{\mathcal{G}_2}_n + \epsilon^4_{nk}. \tag{11}$$

- Voxels in set $\mathcal{I}_3$ are related to the disease but are not related to genetic markers, and are therefore not helpful in causative SNP detection. In fact, such features confuse the detector as they get selected as relevant to disease at the cost of features in $\mathcal{I}_1$. We generate these voxels as follows:

$$c^5_{kn} \sim \begin{cases} \mathcal{N}(0.5, 1), & \text{if } y_n = 1, \\ \mathcal{N}(-0.5, 1), & \text{if } y_n = 0. \end{cases}$$

- Voxels in set $\mathcal{I}_4$ are not relevant to either label or genetic markers. These voxels are sampled from $\mathcal{N}(0, \sigma^2_{noise})$.

We use the synthetic data to evaluate detection of disease causative SNPs with our method. We observe that our algorithm is not sensitive to the hyperparameters, which we set as follows: $\log \frac{\beta}{1-\beta} = -1$, $\log \frac{\alpha}{1-\alpha} = -3$, $\sigma^2_\eta = 1$, and σ^2_0 to the variance of image features. As a first baseline, we perform univariate Bonferroni corrected t-tests directly between SNPs and class labels, omitting imaging. As a second baseline, which we refer to as *supervised sRRR*, we perform univariate voxel filtering using class labels, followed by sRRR mul-

tivariate regression between surviving voxels and genetic variants to recover relevant SNPs [20]. We compare the methods in different image noise regimes by varying the variance σ^2_{noise} in Eqs (10)- (11), and run 50 different independent simulations for each noise regime.

Fig.3(a) reports detection rates (TP) of disease causative SNPs in $\mathcal{G}_1$. To set the detection thresholds we fix the false positive rate to 1%. We observed similar behavior for a broad range of low false positive rates (not shown). We focus our experiments on low false positive rates because at higher rates false detections become comparable with, and ultimately overwhelm true detections. We find that for a given false positive rate, our algorithm detects significantly more disease causative SNPs in $\mathcal{G}_1$ than the baseline algorithms, and has lower standard deviation than the supervised sRRR pipeline. The direct univariate t-tests only detect SNPs that have a very strong independent association with disease label. To illustrate the behavior of the methods at different false positive rates, we report the receiver operating characteristic at two different noise levels in Fig.3(b,c). Our approach achieves a better detection than the baseline methods.

4.2 ADNI dataset

We apply our method on a subset of the Alzheimers Disease Neuroimaging Initiative (ADNI) dataset that includes T1-weighted MR images and 620,000 genetic variants for 228 AD patients and 187 normal controls (NC). All images were pre-processed and non-rigidly aligned to a common [4]. We compute the tissue density map, indicating expansion or contraction of gray matter using the determinant of the Jacobian of the deformation field. The map values in the template space are proportional to the volume of structures in the original brain scan. To reduce image dimensionality, we aggregate voxels into *supervoxels* using spatial $k-$means clustering [11] and obtain about 1700 supervoxels. We define our image features x_{nm} as the average value of the tissue density map in a supervoxel. We use a SVM classifier to asses the discriminative power of the resultant features and obtain 86% classification rate of AD versus NC, close to the state-of-the-art results [4]. We used the ENIGMA protocol to preprocess the genotype data[3]. Briefly, PLINK was used to eliminate SNPs on the basis of standard quality control criteria, e.g., low MAF (< 0.01), poor genotype calling (call rate $< 95\%$) and deviations from Hardy–Weinberg equilibrium ($P < 1\times10^6$). We then performed imputation using the Mach software[4]. Finally, we pre-selected 960 SNPs that have the strongest association with AD overlapped with SNPs reported in a prior AD-GWAS study involving over 16,000 individuals [6].

We ran our algorithm with 10 initializations, and selected the run that achieved the lowest value of the cost function. As before, we set: $\log\frac{\beta}{1-\beta} = -1$, $\log\frac{\alpha}{1-\alpha} = -3$ and $\sigma^2_\eta = 1$. We set $\sigma^2_0 = \omega \cdot \sigma^2_x$, where we sweep $\omega \in [0.1, 0.9]$ and σ^2_x is the variance of image features. Fig.4 illustrates the posterior probabilities of SNP

[3] http://enigma.loni.ucla.edu/protocols/genetics-protocols/
[4] http://www.sph.umich.edu/csg/abecasis/MaCH/index.html

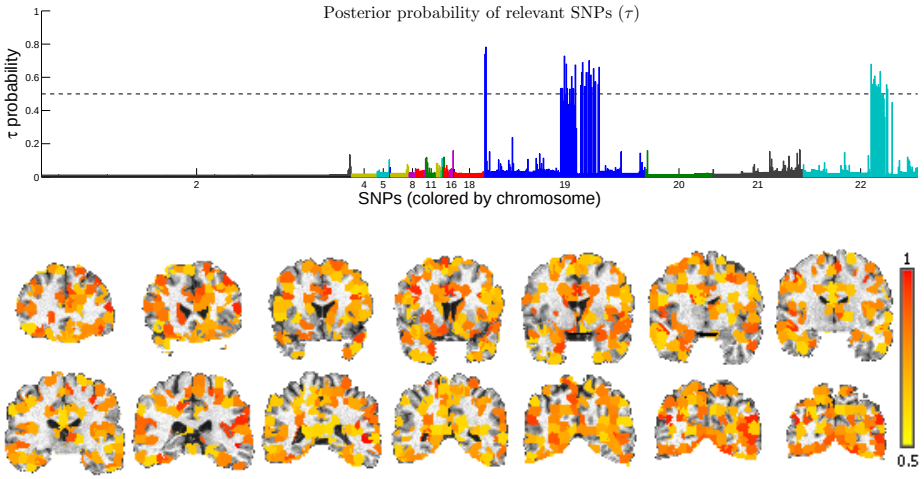

Fig. 4: Results on ADNI dataset. Top: Posterior probability τ_s (colored by chromosome), with 41 SNPs passing a $\tau = 0.5$ threshold. Bottom: Image features ($\rho_m > 0.6$) overlayed on a template MR image, with color intensities proportional to values of ρ.

relevance $\boldsymbol{\tau}$, averaged over the swept parameters. We list the top SNPs in Table 2. The top variants are APOE-$\epsilon 4$ and APOE-$\epsilon 3$, which are strongly correlated with AD [6]. We also detect variants on APOC1, TOMM40 and PVRL among our top hits, all of which are on chromosome 19 and have been frequently reported [6]. Similarly, several chromosome 22 variants are identified [10]. Fig.4 illustrates the average posterior probability of feature relevance $\boldsymbol{\rho}$. We find high probability in hippocampus and temporal lobe, which have been frequently reported to undergo significant shrinkage in AD [4], and are associated with memory.

rank	τ_s	SNP (Gene)	chr
1	0.78	APOE-$\epsilon 4$	19
2	0.74	APOE-$\epsilon 3$	19
3	0.73	rs283812 (PVRL2)	19
4	0.70	rs5117 (APOC1)	19
5	0.69	rs75627662	19

rank	τ_s	SNP (Gene)	chr
6	0.68	rs6857 (PVRL2)	19
7	0.68	rs75843224	22
8	0.67	rs59007384 (TOMM40)	19
9	0.66	rs66626994 (APOC1P1)	19
10	0.65	rs12721051 (APOC1)	19

Table 2: Summary of selected SNPs with the highest posterior probability τ_s.

5 Conclusion

We proposed and demonstrated a unified framework for identifying genetic variants and image-based features associated with the disease. We capture the associations between imaging and disease phenotype simultaneously with the correlation from genetic variants and image features in a probabilistic model.

We derive an algorithm that iteratively refines the relevant variants using disease phenotype and imaging features. The algorithm also isolates representative features that are discriminative with respect to the disease and are modulated by the genetic variants. We demonstrated the benefit of simultaneously performing these two tasks in simulations and in a context of a real clinical study.

Acknowledgements This work was supported by NIH NIBIB NAMIC U54-EB005149, NIH NCRR NAC P41-RR13218 and NIH NIBIB NAC P41-EB-015902, NIH K25 NIBIB 1K25EB013649-01, AHAF pilot research grant in Alzheimer's disease A2012333, NSERC CGS-D and Barbara J. Weedon Fellowship.

References

1. N. K. Batmanghelich, B. Taskar, and C. Davatzikos. Generative-discriminative basis learning for medical imaging. *IEEE Trans Med Imaging*, 31(1):51–69, January 2012.
2. C M. Bishop. *Pattern recognition and machine learning*. Springer, New York, 2006.
3. P Carbonetto and M Stephens. Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies. *Bayesian Analysis*, 7:73–108, 2012.
4. Y Fan, N Batmanghelich, C.M. Clark, C. Davatzikos, and ADNI. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *Neuroimage*, 39(4):1731–1743, Feb 2008.
5. N. Filippini, A. Rao, S. Wetten, R. A. Gibson, et al. Anatomically-distinct genetic associations of APOE epsilon4 allele load with regional cortical atrophy in Alzheimer's disease. *Neuroimage*, 44(3):724–728, Feb 2009.
6. D. Harold, R. Abraham, P. Hollingworth, R. Sims, et al. Genome-wide association study identifies variants at clu and picalm associated with alzheimer's disease. *Nat Genet*, 41(10):1088–1093, Oct 2009.
7. J. M. Hernandez-Laborto and D. Hernandezi-Lobato. Convergent Expectation Propagation in Linear Models with Spike-and-Slab Priors . December 2011.
8. T.S. Jaakkola and M.I. Jordan. Bayesian Paramater Estimation via Variational Methods. *Statistics and Computing*, (10):25–37, 2000.
9. E. Le Floch, V. Guillemot, V. Frouin, P. Pinel, et al. Significant correlation between a set of genetic polymorphisms and a functional brain network revealed by feature selection and sparse Partial Least Squares. *Neuroimage*, 63(1):11–24, Oct 2012.
10. J.H. Lee, R. Cheng, N. Graff-Radford, T. Foroud, et al. Analyses of the national institute on aging late-onset alzheimer's disease family study: implication of additional loci. *Archives of neurology*, 65(11):1518, 2008.
11. A. Lucchi, K. Smith, R. Achanta, G. Knott, and P. Fua. Supervoxel-based segmentation of mitochondria in em image stacks with learned shape features. *IEEE Trans Med Imaging*, 31(2):474–486, Feb 2012.
12. D. Lvovs, OO Favorova, and AV Favorov. A polygenic approach to the study of polygenic diseases. *Acta Naturae*, 4(3):59, 2012.
13. S.G. Mueller, M.W. Weiner, L.J. Thal, R.C. Petersen, et al. The alzheimer's disease neuroimaging initiative. *Neuroimaging Clinics of North America*, 15(4):869, 2005.
14. R. B. O'Hara and M. J. Sillanpää. A Review of Bayesian Variable Selection Methods: What, How and Which,. *Bayesian Analisis*, 4(1):85–118, 2009.
15. S.G. Potkin, J.A Turner, G. Guffanti, A. Lakatos, and other. A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype. *Schizophr Bull*, 35(1):96–108, Jan 2009.

16. S Purcell, B Neale, K Todd-Brown, L Thomas, et al. PLINK: a tool set for whole-genome association and population-based linkage analyses. *Am J Hum Genet*, 81(3):559–575, September 2007.
17. M.R. Sabuncu and K. Van Leemput. The Relevance Voxel Machine (RVoxM): A Bayesian Method for Image-Based Prediction. In T. Peters, G. Fichtinger, and A. Martel, editors, *MICCAI 2011*, LNCS, pages 99–106. Springer, Heidelberg, 2011.
18. J.L Stein, X. Hua, S. Lee, A.J. Ho, et al. Voxelwise genome-wide association study (vGWAS). *Neuroimage*, 53(3):1160–1174, Nov 2010.
19. M. Vounou, E. Janousova, R. Wolz, J. L Stein, et al. Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease. *Neuroimage*, 60(1):700–716, Mar 2012.
20. M. Vounou, T.E. Nichols, G. Montana, and ADNI. Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach. *Neuroimage*, 53(3):1147–1159, Nov 2010.

Appendix A

We define $\mathbf{X} \in \mathbb{R}^{N\times M}$ to be a matrix of all image features (each row is a subject), $\mathcal{J}(x,y) := (1-x)\log(\frac{1-x}{1-y}) + x\log(\frac{x}{y})$, and use $diag(\cdot)$ to transforms a vector into a diagonal square matrix or the diagonal of a square matrix into a vector. $\mathcal{E}_m = \langle\cdot\rangle_{q|b_m=1}$ denotes expectation with respect to q conditioned on $b_m = 1$ of the genetics-to-image regression. We define $\mathbf{Q} = \mathbf{G}^T\mathbf{G}$, and $\mathbf{D} = diag(diag(\mathbf{Q}) \odot \frac{1-\boldsymbol{\tau}}{\boldsymbol{\tau}})$.

Parameters of the genetic part of the model are updated as follows:

$$\mathcal{E}_m := \sum_n \langle x_{nm} - \mathbf{g}_n^T(\mathbf{v}_m \odot \mathbf{a}_m)\rangle_{q|b_m=1} = \tag{12a}$$
$$(\mathbf{x}^m)^T\mathbf{x}^m + \mathbf{v}_m^T(\mathbf{Q} \odot (\boldsymbol{\tau}\boldsymbol{\tau}^T - diag(\boldsymbol{\tau}^2 - \boldsymbol{\tau})))\mathbf{v}_m - 2(\mathbf{x}^m)^T\mathbf{G}diag(\boldsymbol{\tau})\mathbf{v}_m,$$
$$\mathbf{v}_m = diag(\boldsymbol{\tau})\mathbf{D}^{-1}\mathbf{G}^T[\mathbf{U}_O((I + \Sigma_O)^{-1})\mathbf{U}_O^T]\mathbf{x}^m, \tag{12b}$$
$$\log\frac{1-\tau_s}{\tau_s} = \log\frac{1-\beta_s}{\beta_s} + \frac{2}{\sum_m \rho_m}\sum_{m=1}^{M}\frac{\rho_m}{2\sigma_0}\frac{\partial\mathcal{E}_m}{\partial\tau_s}. \tag{12c}$$

$\mathbf{U}_O\Sigma_O\mathbf{U}_O^T$ is the Singular Value Decomposition of $\mathbf{G}\mathbf{D}^{-1}\mathbf{G}^T$, whose complexity $\mathcal{O}(N^3)$ is not expensive for a modest number of subjects N. $\mathbf{x}^m$ denotes column m of matrix $\mathbf{X}$. In Eq.(12c), the posterior log-odds ratio is updated by adding the prior log-odd ratio and a weighted sum of the derivatives of the regression error terms for all m with respect to τ_s. Moreover, we obtain

$$\xi_n^2 = \mathbf{x}_n^T(diag(\boldsymbol{\nu}^2 + \boldsymbol{\varsigma}^2))\mathbf{x}_n, \tag{13a}$$
$$1/(\varsigma_m)^2 = (\mathbf{X}^T\mathbf{X})_{mm} + 1/\sigma_\mu^2, \tag{13b}$$
$$\nu_m = (\varsigma_m)^2((\mathbf{X}^T\hat{y})_m - \sum_{j\neq m}(\mathbf{X}^T\mathbf{X})_{jm}\rho_j\nu_j), \tag{13c}$$
$$\log\frac{1-\rho_m}{\rho_m} = \log\frac{1-\alpha}{\alpha} + \log\frac{\sigma_\mu}{\varsigma_m} + \sum_{s=1}^{S}\mathcal{J}(\tau_s,\beta_s) - \frac{1}{2}(\frac{\nu_m}{\varsigma_m})^2 + \frac{\mathcal{E}_m}{2\sigma_0^2} + \log\sigma_0. \tag{13d}$$

Eq.(13b)-(13c) update the mean and standard deviations of the normal distributions in the approximate posterior. Eq.(13d) updates posterior probability of the relevance of region m.