Probabilistic Modeling of Imaging, Genetics and Diagnosis

Nematollah K. Batmanghelich, Adrian Dalca, Gerald Quon, Mert Sabuncu, Polina Golland, for the Alzheimer's Disease Neuroimaging Initiative*

***Abstract*—**
We propose a unified Bayesian framework for detecting genetic variants associated with disease by exploiting image-based features as an intermediate phenotype. The use of imaging data for examining genetic associations promises new directions of analysis, but currently the most widely used methods make sub-optimal use of the richness that these data types can offer. Currently, image features are most commonly selected based on their relevance to the disease phenotype. Then, in a separate step, a set of genetic variants is identified to explain the selected features. In contrast, our method performs these tasks simultaneously in order to jointly exploit information in both data types. The analysis yields probabilistic measures of clinical relevance for both imaging and genetic markers. We derive an efficient approximate inference algorithm that handles the high dimensionality of image and genetic data. We evaluate the algorithm on synthetic data and demonstrate that it outperforms traditional models. We also illustrate our method on Alzheimer's Disease Neuroimaging Initiative data.

***Index Terms*—Imaging Genetics, Bayesian Models, Variational Inference, Probabilistic Graphical Model**

I. Introduction

In this paper, we propose a probabilistic model to discover genetic variants associated with a disease using image 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). In GWAS, it is common to examine the associations of genetic variants with disease by performing a univariate analysis between the disease incidence and each genetic marker independently. However, testing one variant at a time does not fully realize the potential of GWAS because some genetic variants may have a weak but cumulative effect that is neglected by a univariate method [1], [2]. Imaging genetics introduces image-based biomarkers as a promising intermediate phenotype[1] (*i.e.,* endo-phenotype) between genetic variants and diagnosis. Given that in some pathologies, such as the Alzheimer's disease, imaging features have strong correlation with the clinical diagnosis and can offer a clearer picture of the association [5], [6], it is beneficial to exploit them to improve the associations of weak genetic markers. Furthermore, in contrast to a binary diagnosis, imaging data contains many variations caused by a disease which helps to stratify the disease population in more informative ways.

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 methods considered only a few imaging candidates (voxels, regions, or other biomarkers) or only a few genetic markers in the analysis [7], [8]. The reduced joint dataset was then analyzed in a univariate framework, where pairs of a candidate genetic variant and an imaging biomarker were tested for association via standard statistical tests. Examples include using activation maps of the prefrontal cortex to find SNPs associated with schizophrenia [8] and searching for changes in regional gray matter volumes correlated with the genetic risk of Alzheimer's disease [7], [9].

More recently, genome-wide voxel-wise analysis has been demonstrated using univariate methods [10]. However, massive univariate analysis has several limitations. Due to multiple comparisons, a conservative corrected significance level is selected to limit the false positive rate, but this correction 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 [11], [12].

A common approach is to use a multivariate regression combined with a regularization to extract a sparse set of coefficients for correlated genetic variants and image features. Various forms of relationship between imaging and genetic data along with different regularization terms have been proposed in the literature. For example, it is common to assume that image and genetic data lie in a joint hidden (latent) space. This is equivalent of enforcing different forms of low rank regularization on data: sparse reduced rank regression (sRRR) [12], [13], Partial Least Squares (PLS) [11] or Canonical Correlation Analysis (CCA) [11]. Unfortunately, these unsupervised methods do not use the clinical labels (*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 selected separately by modeling the relationship between

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

[1]The term "intermediate phenotype" or "endophenotype" is commonly used in the literature [3], [4]. It is called intermediate phenotype because in a hypothetical causal model, it falls between the genotype and disease diagnosis. The intermediate data in our case is the image feature (*e.g.,* average thickness of the cortical regions or volume of the sub-cortical areas).

image features and the phenotype of interest. For example, sRRR has been demonstrated using brain regions pre-selected for Alzheimer's disease (AD) via Linear Discriminant Analysis [13].

In contrast, we model and estimate relevant genetic variants in the context of abnormal variations that are characterized by imaging features. Our method is broadly applicable to any imaging biomarker, such as anatomical regions, tissue appearance, or functional measures. Here, we demonstrate our method in application to Alzheimer's disease, and use thickness of cortical regions and the volume of sub-cortical structures as image features.

We define a probabilistic model to encode the relationship among genetic, image and disease measures. Our model incorporates a common assumption made by genetic studies that only a small set of genetic variants is associated with any particular disease, leading to sparsity-inducing priors. The relevant 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 disease phenotype. Therefore, in our model if a brain region is irrelevant for the target disease, it is ignored even if it is strongly modulated by genetics. We also derive an efficient inference algorithm to identify relevant brain regions and genetic loci, and demonstrate the method on synthetic data and real data from the ADNI study [14]. We demonstrate that our algorithm outperforms standard univariate and regression analyses for genetic variant detection on synthetic data and yields promising results in a real clinical study. This paper extends our publication of the preliminary results [15] by deriving a novel robust inference algorithm. It also expands the empirical evaluation.

The remainder of this paper is organized as follows. In the next section, we build a graphical model that captures the relationship among image, genetic and diagnostic variables. In Section III, we propose an efficient algorithm to perform inference of the model. Derivation details are discussed in the Supplementary Material. Section IV and Section V report experimental results on simulated and real data, respectively. We conclude the paper with a discussion of the results and future directions in Section VI.

II. Method

A. Notations and Terminology

Throughout this paper, we use regular fonts (*e.g.,* x, τ) and bold fonts (*e.g.,* $\mathbf{x}$, $\boldsymbol{\tau}$) to denote scalar and vector, respectively. Some uppercase letters are reserved for the number of elements: *e.g.,* N is the number of subjects, M is the number of image regions, and S is the number of SNPs. In such cases, their lowercase counterparts are used for enumeration: *e.g.,* subject n, image region m, and SNP s. Uppercase bold letters are used to denote matrix variables (*e.g.,* $\mathbf{V} \in \mathbb{R}^{S\times M}$); in such case $\mathbf{V}_{:m}$ and $\mathbf{V}_{s:}$ denote the column m and row s of the matrix $\mathbf{V}$, respectively. We use V_{sm} to refer to the entry in the row s and column m of $\mathbf{V}$. Superscripts are used to denote iterations of the algorithm (*e.g.,* $\mathbf{b}^t$) or transpose (*e.g.,* $\mathbf{X}^T$). $\mathbb{E}[\cdot]$ and $p(\cdot)$ denote expectation and density. Table I summarizes all variables used throughout this paper.

Model Variables: Image to Disease Phenotype	
x_{nm}	Image feature m in subject n (brain endophenotype featue).
y_n	Disease phenotype (diagnosis variable / class label) of subject n: -1 - healthy, 1 - diseased.
$b_m \in \{0,1\}$	Indicator variable that selects image feature m.
f	Latent function drawn from a Gaussian Process to predict y from image feature vector $\mathbf{x}$.
β	Prior probability for selecting image features.
Model Variables: Genetics to Image	
g_{ns}	Genetic variant s in subject n.
$\boldsymbol{\omega}_m$	Regression coefficient vector for predicting image feature m using the genotype.
$a_{sm} \in \{0,1\}$	Indicator variable that selects SNP s for modeling image feature in region m.
α	Prior probability for selecting genetic variants.
$\sigma^2_{\boldsymbol{\omega}}$	Variance of an element in $\boldsymbol{\omega}_m$.
σ^2_0	Variance of noise in the genetics to image regression for the relevant regions.
Variational Variables	
ρ_m	Posterior probability of selecting feature m.
τ_s	Posterior probability of selecting SNP s.
$\boldsymbol{\nu}, \varsigma$	Mean and variance parameters of the genetics-to-image regression.

TABLE I: Notation and variables used throughout the paper.

B. Model

We are motivated by anatomical brain studies with binary phenotypes (-1 or 1), but the analysis applies to any biomarker derived from images and the constraint on the phenotype can be easily relaxed. We assume that a study contains N individuals, each with three measurements:

- disease phenotype $y \in \{-1, 1\}$ that indicates healthy *vs.* disease;
- image measurements, $\mathbf{x} \in \mathbb{R}^M$, which are usually referred to as "intermediate phenotype". In the context of AD, image features include volume or thickness measurements of M brain structures.
- genetic variants $\mathbf{g} \in \mathbb{R}^S$ at S locations along the genome;

We assume that a subset of image features is modulated by genetics and is closely related to the disease phenotype. Detecting and utilizing such imaging features can improve the detection of relevant genetic variants.

We model two types of relationships, illustrated Fig.1: 1) the association of a subset of brain regions with the diagnosis variable y, which can be quantified by the quality of the disease prediction from image features; 2) a modulation of each image feature by the genotype. A common approach is to consider these two relationships separately, selecting relevant brain structures and then performing a statistical test (*e.g.,* t-test or sparse regression) to identify the relevant genotype [13]. In contrast, we propose a model to perform these two steps jointly, via two coupled regression models:

- A sparse subset of imaging features selected by $\mathbf{b} \in \{0,1\}^M$ is related to the diagnosis variable y via a logistic regression model. For each region, we model its elements (*i.e.,* b_m) using a Bernoulli distribution (Section II-C).
- Variations in image features for region m can be explained by a sparse subset of the genotype which is selected by $\mathbf{a}_m \in \{0,1\}^S$. Similarly, we model its

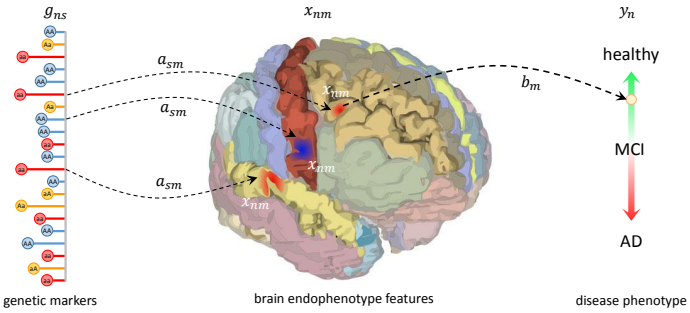


Fig. 1: A schematic illustration of the relationship between genetic, imaging and clinical measures in our model.

elements (*i.e.,* a_{sm}) via a Bernoulli distribution (Section II-D).

We treat the indicator variables $\{\mathbf{a}_m\}_{m=1}^M$ and $\mathbf{b}$ as latent. The graphical model in Fig.2 presents the relationships among all variables in the model. One can view the model shown in Fig.2 as two-layers of regression that share latent variables for the image data. Below, we first define the relationship between image features and the disease phenotype and then specify the generative model for the relationship between SNPs and image features. We do not model a direct link between genetic variants and disease label. It is captured indirectly through image features. The general idea is illustrated in Fig.1.

C. From Imaging Features to Disease Phenotype

To predict the binary class label y from a sparse set of image features $\mathbf{x}$, we use a variant of the log-odds model:

$$\log\left\{\frac{p(y=1|f,\mathbf{b},\mathbf{x})}{1-p(y=1|f,\mathbf{b},\mathbf{x})}\right\} = f\left(\mathbf{x}\odot\mathbf{b}\right), \tag{1}$$

where $\odot$ is the element-wise product, $\mathbf{b}\in\{0,1\}^M$ is the latent variable that selects relevant regions, and $f(\cdot)$ is a latent stochastic function. In effect the operation $\mathbf{x}\odot\mathbf{b}$ masks out the irrelevant features.

We assume exchangeable Bernoulli prior for $\mathbf{b}$. In other words, we model selection of each region as a biased coin flip, *i.e.,* $p(b_m) = \beta^{b_m}(1-\beta)^{1-b_m}$, where β is the prior probability of including a brain region. We use the Gaussian Process (GP) as a prior for f [16]. A Gaussian Process is a random process where any finite sample set is distributed as a multi-dimensional Gaussian distribution. GP is completely defined by its prior mean and covariance functions, *i.e.,* $f(\mathbf{x}\odot\mathbf{b})\sim\mathcal{GP}(m_{\mathbf{b}}(\mathbf{x}),k_{\mathbf{b}}(\mathbf{x},\mathbf{x}'))$, where

$$m_{\mathbf{b}}(\mathbf{x}) = \mathbb{E}[f(\mathbf{x}\odot\mathbf{b})],$$
$$k_{\mathbf{b}}(\mathbf{x},\mathbf{x}') = \mathbb{E}[(f(\mathbf{x}\odot\mathbf{b})-m_{\mathbf{b}}(\mathbf{x}\odot\mathbf{b}))^T(f(\mathbf{x}'\odot\mathbf{b})-m_{\mathbf{b}}(\mathbf{x}'))]$$

We assume $m_{\mathbf{b}}(\mathbf{x}) = 0$ since $y\in\{-1,1\}$ and we do not aim to induce a bias toward either label. The covariance function $k(\cdot,\cdot)$ is the crucial part of a GP. There are several well-known choices for $k(\cdot,\cdot)$ such as Linear $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}'$, or Squared Exponential $k(\mathbf{x},\mathbf{x}') = \exp\left(-\frac{\|\mathbf{x}-\mathbf{x}'\|_2^2}{2\sigma^2}\right)$. We use the linear kernel in this paper, setting $k(\mathbf{x},\mathbf{x}') = \mathbf{x}^T\mathbf{x}'$. The expression on the left hand side of Eq. (1) specifies the likelihood (*i.e.,* the link function). For example, a straightforward change from the logistic likelihood to a Gaussian likelihood enables modeling continuous clinical measurements (*e.g.,* cognitive scores).

D. From Genetic Variants to Imaging Features

An imaging feature m is either relevant to the disease ($b_m = 1$) or not ($b_m = 0$). In modeling the relationship between genetics and imaging, we treat these cases differently. If feature m is irrelevant ($b_m = 0$), we model the variation in the region as a Gaussian distribution centered at zero with a fixed standard deviation of one: $x_m \sim \mathcal{N}(\cdot;0,1)$. This assumption is not limiting, since we can always normalize the samples to have zero mean and unit variance. The normal distribution can be replaced by a different distribution if needed. One can view this assumption as our *null* distribution. If feature m is relevant for disease prediction ($b_m = 1$), variations in the values of this feature are explained by a sparse subset of the genetic variants $\mathbf{g}\in\mathbb{R}^S$. 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 for region m, and arrive at the second regression component of our model:

$$x_{nm} = \boldsymbol{\omega}_m^T\left(\mathbf{g}_n\odot\mathbf{a}_m\right)+\epsilon_{nm}, \tag{2}$$

where $\boldsymbol{\omega}_m$ is the vector of regression coefficients, $\epsilon_{nm}\sim\mathcal{N}(\cdot;0,\sigma_0^2)$ is the *iid* residual noise in the image feature m for subject n. Adopting Bayesian variable selection based on the spike-and-slab model [17], [18], we assume a Gaussian distribution with zero mean and variance $\sigma_0\sigma_\omega$ as a prior for the regression coefficient $\boldsymbol{\omega}_m$. This choice of parameterization facilitates derivations explained in the Section III. Similar to the indicator variable $\mathbf{b}$ that selects image features, we assume exchangeable Bernoulli distribution as a prior for $\mathbf{a}_m$:

$$p(a_{sm};\alpha) = \alpha^{a_{sm}}(1-\alpha)^{1-a_{sm}}, \tag{3}$$

where α is the prior probability of including any SNP in the model.

Combining, we obtain the likelihood of the image feature m:

$$p(x_{nm}|b_m,\mathbf{a}_m,\mathbf{g};\boldsymbol{\omega}_m,\sigma_0^2) = \begin{cases}\mathcal{N}(x_{nm};0,1), & \text{if } b_m=0,\\ \mathcal{N}(x_{nm};\boldsymbol{\omega}_m^T\left(\mathbf{g}_n\odot\mathbf{a}_m\right),\sigma_0^2), & \text{if } b_m=1.\end{cases} \tag{4}$$

The first line of Eq. (4) assumes a simple normal distribution as a null model. To handle cases where a non-disease related genetic variants affect a relevant region (*i.e.,* $b_m = 1$), we assume that the effect of the normal genetic variants along with other covariates (*e.g.,* age, gender, *etc.*) are already subtracted from the data and Eq. (4) models the normalized residual. More explicitly, we fit a regression model on all *measured* nuisance variables in a normal population. x_{nm} represents the residual of the regression which presumably regresses out all of the nuisance variables.

E. Complete model

We define $\mathcal{Z} = \{f,\mathbf{b},\mathbf{a}_1,\cdots,\mathbf{a}_m,\boldsymbol{\omega}_1,\cdots,\boldsymbol{\omega}_m\}$ to be the set of latent variables, $\mathcal{D} = \{\mathbf{X},\mathbf{y}\}$ to be the set of data variables that we model, and $\pi = \{\sigma_0^2,\sigma_\omega^2,\alpha,\beta\}$ to be the set of hyper-parameters. We use $\mathbf{y} = [y_1;\cdots;y_N]$ to denote the set of all clinical phenotypes (class labels) and $\mathbf{X}\in\mathbb{R}^{N\times M}$

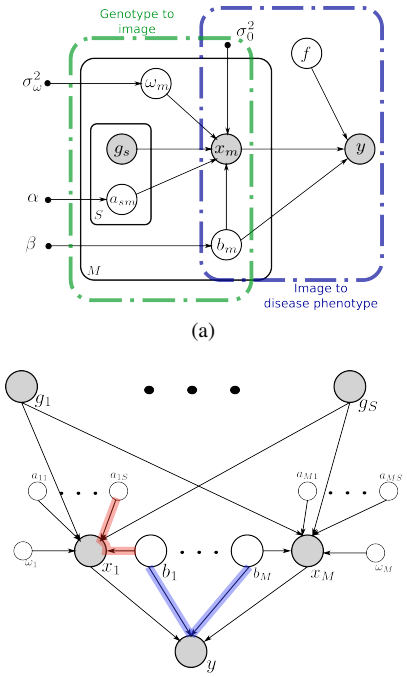


Fig. 2: (a) Graphical representation of the generative model. Hollow circles (◯) denote random variables, small solid circles (•) represent hyper-parameters, and shaded circles represent observed variables. The black plates indicate conditionally independent instantiations. More specifically, α, β, σ_ω and σ_0 are the hyper-parameters. The dashed boxes illustrate the different parts of the model. (b) Instead of plates, the repetition of the random variables are shown explicitly. To avoid the visual clutter, the hyper-parameters are not shown. The blue and the red paths show so-called v-structure dependence. It means that those variables are conditionally dependent hence the posterior values for those variables are related.

and $\mathbf{G} \in \mathbb{R}^{N\times S}$ are respectively image and genetic data of all subjects where each row is a subject and each column represents a measurement from one brain region (for $\mathbf{X}$) or genotype from all loci (for $\mathbf{G}$). Since the hyper-priors are treated slightly differently during inference, in this section we focus on the structure of the conditional probability given the hyper-parameters: $p(\mathcal{D}, \mathcal{Z}|\pi; \mathbf{G})$ (see the Supplementary Material). Combining the elements of the model in Eqs. (1)-(4), we construct the joint distribution of the hidden variables $\mathcal{Z}$ and modeled variables $\mathcal{D}$:

$$p(\mathcal{D}, \mathcal{Z}|\pi; \mathbf{G}) = p(f) \prod_{m=1}^{M} p(\boldsymbol{\omega}_m|\pi)p(b_m|\pi) \prod_{s=1}^{S} p(a_{sm}|\pi)$$
$$\times \prod_{n=1}^{N} p(y_n|f, \mathbf{b}, x_{nm}; \pi)p(x_{nm}|b_m, \mathbf{a}_m, \boldsymbol{\omega}_m; g_{ns}, \pi). \quad (5)$$

In the next section, we focus on specifying hyper-priors $p(\pi)$.

F. Hyper-priors

For clarity of presentation, Fig.2 presents the model but does not specify the priors for α, β, σ_0, and σ_ω. Here we define the prior distributions for each parameter of the model.

Prior Over Inclusion of SNPs α: We assume the conjugate prior for $\alpha \in (0, 1)$, namely a Beta distribution. The shape parameters of the Beta distribution are chosen to ensure an almost flat distribution over the entire interval $(0, 1)$ as illustrated in the experimental section.

Prior Over Variance of Residual σ_0: It is common to assume an uninformative prior distribution[2] for the variance of residuals [19]. An uninformative prior for σ_0 is proportional to $\frac{1}{\sigma_0^2}$, which can be achieved via an inverse Gamma distribution as the scale and shape priors approach zero [20], *i.e.,* $\sigma_0 \sim \text{IG}(\iota_1, \iota_2)$.

Prior Over σ_ω: Instead of directly imposing prior on σ_ω, we follow the approach of assuming a flat prior for *Proportion of Variance Explained (PVE)* in the response that consequently induces a prior on the parameter σ_ω [17], [21]. The underlying logic is that there might be a large number of SNPs with small PVE's or small number of SNPs with large PVE's; hence we assume a flat prior over PVE. Assuming that the columns of the genetic data matrix $\mathbf{G}$ are centered, the PVE of the genetic variants for image feature m is defined as follows:

$$\text{PVE}_m := \frac{\sigma_0^{-2}\frac{1}{N}\sum_{n=1}^{N}(\mathbf{G}_{n:}\boldsymbol{\omega}_m)^2}{1+\sigma_0^{-2}\frac{1}{N}\sum_{n=1}^{N}(\mathbf{G}_{n:}\boldsymbol{\omega}_m)^2}$$

A rough estimate of the expectation of PVE (*i.e.,* integrating $\boldsymbol{\omega}_m$ out) can be represented to be:

$$\widehat{\text{PVE}}_m = \frac{\alpha\frac{\sigma_\omega^2}{\sigma_0^2}\varrho}{\left(1+\alpha\frac{\sigma_\omega^2}{\sigma_0^2}\varrho\right)}, \quad (6)$$

where $\varrho = \frac{1}{N}\sum_{n=1}^{N}\sum_{s=1}^{S} g_{ns}^2$ is the sum of the sample variances of the genetic data at all S loci. We assume a uniform prior over $\widehat{\text{PVE}}$ [17]. This prior aids interpretation as it applies stronger shrinkage in models with more non-zero regression coefficients [21].

We leave the prior β for selecting image features as a non-random hyper-parameter whose effect on the final results will be studied empirically in the experimental section.

G. Joint Modeling Image and Genetics vs. Two-Step Inference

Our method jointly models imaging and genetic variations. To clarify the concept, we first explain the so-called "two-steps" method in the context of our algorithm. A two step approach (*e.g.,* [13]) first selects a subset of brain regions (columns of $\mathbf{X}$). This can be done using a univariate or multivariate approach. A univariate approach seeks a Maximum a Posterior (MAP) estimate to the following formulation accounting for each column separately:

$$\underbrace{p(b_m|\mathbf{y}, \mathbf{X}_{:,m})}_{\text{posterior}} \propto \underbrace{p(\mathbf{y}|\mathbf{X}_{:,m}, b_m)}_{\text{Likelihood term}} \underbrace{p(b_m)}_{\text{prior}}, \quad (7)$$

where b_m is an indicator variable with 1 indicating relevance, and 0 not; and $\mathbf{X}_{:,m}$ is the column m of $\mathbf{X}$, corresponding to the features from brain region m. Assuming uniform prior, most univariate methods find the most likely region by testing the likelihood term for $b_m = 1$ or $b_m = 0$, *i.e.,* $p(\mathbf{y}|\mathbf{X}_{:,m}, b_m = 1) \lessgtr p(\mathbf{y}|\mathbf{X}_{:,m}, b_m = 0)$.

[2]An uninformative prior is a prior that is not subjectively defined and can express objective information such as "the variable is positive."

Unlike univariate approaches, a multivariate method considers all variables at the same time to find MAP or posterior probability of this form:

$$p(\mathbf{b}|\mathbf{y},\mathbf{X}) \propto \underbrace{p(\mathbf{y}|\mathbf{X},\mathbf{b})}_{\text{Likelihood term}}\ \underbrace{p(\mathbf{b})}_{\text{prior}}, \tag{8}$$

where $\mathbf{b}$ is a m-dimensional binary hidden vector that denotes the relevance of the M image regions together.

Although the posterior value depends on all brain regions (simultaneously), such model does not account for the genetic variations. Our model specifically addresses this problem. The graphical model in Fig.2a implies that the posterior probability of the brain regions takes the following form:

$$p(\mathbf{b}|\mathbf{y},\mathbf{X},\mathbf{G}) \propto \underbrace{p(\mathbf{y}|\mathbf{X},\mathbf{b})}_{\text{Likelihood of imaging data}} \times \tag{9}$$
$$\times\ \underbrace{p(\mathbf{X}_{:,m}|\{\mathbf{a}_m\}_m^M;\mathbf{G})p(\mathbf{a}_m)}_{\text{Posterior of the genetic to imaging model}}\ \underbrace{p(\mathbf{b})}_{\text{prior}}$$

$$\log p(\mathbf{b}|\mathbf{y},\mathbf{X},\mathbf{G}) = \underbrace{\log p(\mathbf{y}|\mathbf{X},\mathbf{b})}_{\text{contribution of imaging data}} + \underbrace{\log p(\mathbf{b})}_{\text{prior}} +$$
$$\underbrace{\log p(\mathbf{X}_{:,m}|\{\mathbf{a}_m\}_{m=1}^M;\mathbf{G}) + \log p(\mathbf{a}_m)}_{\text{contribution of genetic data}} + \text{constant} \tag{10}$$

where the values of the posterior distribution are influenced by both diagnosis $p(\mathbf{y}|\mathbf{X},\mathbf{b})$ and genetic data $p(\mathbf{X}_{:,m}|\mathbf{a}_m;\mathbf{G}p(\mathbf{a}_m))$ simultaneously.

Another way to understand the simultaneous aspect of the model is to study the dependency structure of random variables by following the dependency paths in the graphical model. For the sake of better visualization, we have expanded the graphical model of Fig.2a to Fig.2b by removing the plates and explicitly visualizing the random variables. The so-called v-structure dependency (see [22]) between indicator variables of the brain regions (b_m's) means that given the diagnosis variable y, relevance values of different brain regions are conditionally dependent. This dependency is encoded in the posterior probability. Also there is v-structure dependency between indicator of a brain region b_m and indicator variables of the genetic loci ($\mathbf{a}_m$).

III. Inference

Our goal is to compute the posterior probability distribution $p(\mathcal{Z}|\mathcal{D};\mathbf{G},\pi)$ of the latent variables that summarize genetic and imaging influences in our model. Because of coupling of variables in the joint model, computing the posterior distribution is intractable, necessitating approximations via sampling or variational methods. Due to the amount of data and its dimensionality, we use the computationally more efficient approach of variational inference [23]. Three important quantities of the model require further explanation. These three quantities will be used later in the inference section:

1) Diagnosis Likelihood $p(\mathbf{y}|\mathbf{b};\mathbf{X},\pi)$ *:* Assuming that $\mathbf{b}$ is observed, this value is the *marginal conditional* likelihood of the diagnosis model. We use the term *marginal conditional* since it is conditioned on $\mathbf{b}$ and the f is marginalized out. For logistic likelihood Eq. (1), this value does not have a closed-form solution but can be approximated efficiently. To approximate this quantity, one can use Gaussian process classification with linear kernel and approximate the marginal likelihood. We use the expectation propagation to approximate it ([16], Section 3.6).

2) Imaging Likelihood $p(\mathbf{X}_{:m}|b_m;\mathbf{G},\pi)$*:* A straightforward manipulation of Eq. (4) leads to:

$$\log p(\mathbf{X}_{:m}|b_m;\mathbf{G},\pi) = \sum_{n=1}^{N} \log \mathcal{N}(x_{nm};0,1) +$$
$$b_m \left(\log p(\mathbf{X}_{:m}|b_m=1;\mathbf{G},\pi) - \sum_{n=1}^{N} \log \mathcal{N}(x_{nm};0,1) \right) \tag{11}$$

where the first line corresponds to the *null* model, and $\log p(\mathbf{X}_{:m}|b_m=1;\mathbf{G},\pi)$ is the marginal conditional likelihood of the imaging features given genetics where the latent variables $\mathbf{a}_m$ and $\boldsymbol{\omega}_m$ are marginalized out. In general, the marginal likelihood does not have a closed-form but there are several methods to approximate it using Markov Chain Monte Carlo, variational approximation, and Annealed Importance Sampling (AIS) [24]. We adopt the method proposed in [17] specifically for large-scale regression with a spike-and-slab prior. The algorithm combines variational approximation with importance sampling as derived in the Supplementary Material.

3) Posterior Probability $p(\mathbf{b}|\mathcal{D};\mathbf{G},\pi)$*:* This function quantifies the posterior probability of the relevance of the brain regions given the data. $p(\mathbf{b}|\mathcal{D};\mathbf{G},\pi)$ is a function that assigns the posterior probability to all 2^M possibilities of the indicator vector $\mathbf{b}$ for M brain regions. Estimating $p(\mathbf{b}|\mathcal{D};\mathbf{G},\pi)$ is the key component to approximating the posterior distribution of the entire model. Two quantities mentioned earlier are combined in this term:

$$p(\mathbf{b}|\mathcal{D};\mathbf{G},\pi) \propto p(\mathbf{b},\mathcal{D};\mathbf{G},\pi) = \tag{12}$$
$$\underbrace{p(\mathbf{y}|\mathbf{b};\mathbf{X},\pi)}_{\text{diagnosis}}\ \underbrace{\prod_{m=1}^{M} p(\mathbf{X}_{:m}|b_m;\mathbf{G},\pi)p(b_m;\pi)}_{\text{Imaging}}$$

Computing the normalization constant entails a sum over all possible subsets of $[M] := \{1,\cdots,M\}$ which is computationally infeasible. We resort to a variational approximation to compute the posterior distribution.

A. Fixed-Form Variational Learning

A variational method approximates the posterior distribution of the latent variables in the model. It seeks a specified form of the approximating distribution q that minimizes negative of the so-called variational free energy. This quantity lower bounds logarithm of the so-called *evidence* (*i.e.,* $p(\mathbf{X})$), hence called evidence lower bound (ELBO). It can be shown that the objective is the Kullback Leibler divergence between an approximating distribution q and the joint distribution of the model. We approximate $p(\mathbf{b}|\mathcal{D};\mathbf{G},\pi)$ with a function of the following form:

$$q(\mathbf{b};\boldsymbol{\rho}) = \prod_{m=1}^{M} q(b_m;\rho_m) = \prod_{m=1}^{M} \rho_m^{b_m}(1-\rho_m)^{(1-b_m)}, \tag{13}$$

where $\boldsymbol{\rho}$ is the parameters vector of the approximate posterior distribution. To learn $\boldsymbol{\rho}$, we adopt the stochastic approximation algorithm proposed by Salimans and Knowles [25]. An important property of the framework is that it enables approximating of the posterior as long as there are efficient algorithms to sample from the assumed-form of the approximating distribution q and to evaluate the joint likelihood. These properties can be helpful for approximating distributions that are not fully factorizable. In our case, the form of the approximate posterior is fully factorizable but the framework allows for further extensions in the future. We first review this general framework.

In *structured* or *fixed-form* variational Bayes [26], the approximating distribution is chosen to be a specific member of an exponential family, namely $q(\mathbf{b};\boldsymbol{\theta}) = \exp\left(\boldsymbol{\theta}^T T(\mathbf{b}) - U\left(\boldsymbol{\theta}\right)\right)\nu\left(\mathbf{b}\right)$ where $T(\mathbf{b})$ is the sufficient statistics, $U(\boldsymbol{\theta})$ is the log partition function, $\nu(\mathbf{b})$ is the base measure and $\boldsymbol{\theta}$ are the natural parameters. To represent Eq. (13) in this form, we set

$$\begin{aligned} T(\mathbf{b}) &:= \mathbf{b}, & \nu(\mathbf{b}) &:= 1, \\ \theta_m &:= \log\frac{\rho_m}{1-\rho_m}, & U(\boldsymbol{\theta}) &:= \sum_m \log(1+\exp(\theta_m)). \end{aligned}$$

Note that $\boldsymbol{\theta}$ is the transformed version of parameter $\boldsymbol{\rho}$, introduced for notational convenience.

Variational methods find the optimal parameters by minimizing the divergence:

$$\boldsymbol{\theta}^* = \arg\min_{\boldsymbol{\theta}} \mathbb{E}_q\left[\log q_{\boldsymbol{\theta}}(\mathbf{b}) - \log p(\mathbf{b},\mathcal{D};\mathbf{G},\pi)\right]. \quad (14)$$

For notational convenience, we define $\tilde{q}_{\tilde{\boldsymbol{\theta}}} := \exp(\tilde{\mathbf{b}}^T\tilde{\boldsymbol{\theta}})$ where $\tilde{\boldsymbol{\theta}}^T = \left[\boldsymbol{\theta}^T, \theta_0\right]$ and $\tilde{\mathbf{b}}^T = \left[\mathbf{b}^T, 1\right]$. If $\theta_0 = -U(\boldsymbol{\theta})$, then $\tilde{q}$ is the normalized posterior, otherwise it is an unnormalized version [25]. Taking the gradient of the objective with respect to $\tilde{\boldsymbol{\theta}}$, we obtain:

$$\nabla_{\tilde{\boldsymbol{\theta}}}\mathbb{E}_{\tilde{q}}\left[\log\tilde{q}_{\tilde{\boldsymbol{\theta}}}\left(\mathbf{b}\right) - \log p\left(\mathbf{b},\mathcal{D}\right)\right] = \int \tilde{q}_{\tilde{\boldsymbol{\theta}}}(\mathbf{b})\left[\tilde{\mathbf{b}}\tilde{\mathbf{b}}^T\tilde{\boldsymbol{\theta}} - \tilde{\mathbf{b}}\log p(\mathbf{b},\mathcal{D})\right] d\nu.$$

By setting the equation above to zero, Salimans and Knowles [25] linked linear regression and the variational Bayes method. Namely, the optimal solution $\tilde{\boldsymbol{\theta}}$ should satisfy the linear system of equations:

$$\begin{aligned} &\mathbf{C}\tilde{\boldsymbol{\theta}} = \mathbf{g}, \\ \text{where}\quad &\mathbf{C} = \mathbb{E}_q\left[\tilde{\mathbf{b}}\tilde{\mathbf{b}}^T\right] \quad \text{and} \quad \mathbf{g} = \mathbb{E}_q\left[\tilde{\mathbf{b}}\log p\left(\mathbf{b},\mathcal{D}\right)\right] \end{aligned}$$

are estimated by weighted Monte Carlo sampling. More specifically, in iteration t of the algorithm, we sample from the current estimate of the posterior distribution, $q_{\tilde{\boldsymbol{\theta}}^t}$ parameterized by $\boldsymbol{\theta}^t$, and replace $\mathbf{C}$ and $\mathbf{g}$ with an empirical estimate. Salimans *et al.* [25] suggested to sample one instance from the q and update $\mathbf{C}$ and $\mathbf{g}$ as follows:

$$\begin{aligned} \mathbf{g}^{t+1} &= (1-w)\,\mathbf{g}^t + w\widehat{\mathbf{g}}^t, \\ \mathbf{C}^{t+1} &= (1-w)\,\mathbf{C}^t + w\widehat{\mathbf{C}}^t, \end{aligned} \quad (15)$$

where $w \in [0,1]$ is the step size and $\widehat{\mathbf{g}}^t$ and $\widehat{\mathbf{C}}^t$ are the empirical estimates of $\mathbf{g}$ and $\mathbf{C}$ using the sample $\widetilde{\mathbf{b}}^t$:

$$\begin{aligned} \widehat{\mathbf{g}}^t &= \widetilde{\mathbf{b}}^t\log p\left(\mathbf{b}^t,\mathcal{D}\right), \\ \widehat{\mathbf{C}}^t &= \widetilde{\mathbf{b}}^t\widetilde{\mathbf{b}}^{t\,T}. \end{aligned}$$

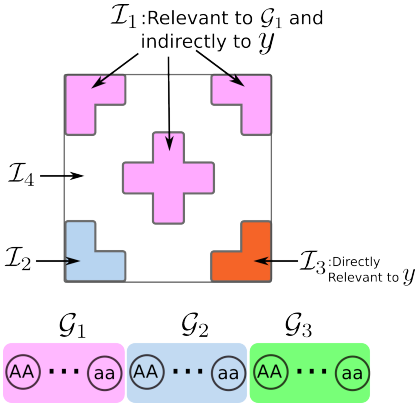


Fig. 3: Summary of the simulation setup. For both healthy subjects and diagnosed patients we split the genome into three regions, and the image into six regions of four types.

With minimal assumptions on the objective function, Nemirovski [27] showed that with a constant step size $w := \frac{1}{\sqrt{N}}$ along with averaging parameters of the last $N/2$ iterations, this procedure leads to asymptotic efficiency of the optimal learning sequence $w^t = ct^{-1}$.

For the pseudo-code of the inference algorithm and detail of derivation, please see the Supplementary Material.

IV. Simulation

We evaluate our model on synthetic data using univariate tests and the sRRR method [12] as baseline algorithms. We also illustrate our method on the ADNI dataset, where we recover several top SNPs associated with the risk of Alzheimer's Disease.

We generate synthetic data to match a realistic scenario as much as possible. Specifically, we generate a disease case-control cohort with images and genetic variants for each subject. We refer to the minor allele frequency (MAF) as 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 [28] 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 the simulated images. We use an odds ratio of 1.125 for heterozygote SNPs, with a multiplicative homozygote risk.
- Set $\mathcal{G}_2$ includes 20 SNPs that are *irrelevant* for the disease (*i.e.,* odds ratio is 1) but affect other areas in simulated images.
- Set $\mathcal{G}_3$ includes 980 *null* SNPs that are independent of both the disease label and the images.

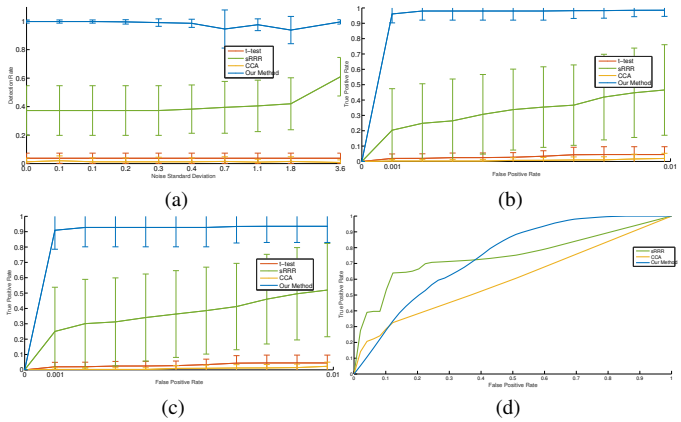


Fig. 4: Summary of the results on simulated data. (a) Detection rates for our algorithm (blue), the supervised sRRR (green), CCA (orange), and genetic t-test (red) as a function of image noise for causative SNPs in $\mathcal{G}_1$ at a false positive rate of 1%. (b,c) ROC curves for low ($\sigma^2_{\text{noise}} = 0.06$) and high ($\sigma^2_{\text{noise}} = 1.7$) noise levels respectively, up to the selected false positive threshold of 1%. The green shows the results of sRRR where any variant that has non-zero weight is considered a hit, and we vary the sparsity parameters. (d) ROC curve for the detection of relevant imaging regions for low ($\sigma^2_{\text{noise}} = 0.06$) noise level.

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 the 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:

$$c^r_{nk} = \mathbf{w}_r^T \mathbf{g}_n^{\mathcal{G}_1} + \epsilon^r_{nk}, \qquad 1 \leq r \leq 3, \tag{16}$$

  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 for disease. We dedicate one region to this category:

$$c^4_{nk} = \mathbf{w}_4^T \mathbf{g}_n^{\mathcal{G}_2} + \epsilon^4_{nk}. \tag{17}$$

- 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 for either the disease label or genetic markers. These voxels are sampled from $\mathcal{N}(0, \sigma^2_{noise})$.

A summary of the simulation setup is shown in Figure 3.

We use the synthetic data to evaluate detection of causative SNPs with our method. As a first baseline method, we perform the univariate Bonferroni corrected t-test directly between SNPs and disease labels, omitting images. As a second baseline, which we refer to as *supervised sRRR*, we perform univariate voxel filtering using disease label, followed by the sRRR multivariate regression between the surviving voxels and the genetic variants to recover relevant SNPs [12]. We compare the methods in different image noise regimes by varying the variance σ^2_{noise} in Eqs (16)- (17), and run 20 different independent simulations for each noise regime. We have also applied CCA, which can be viewed as sRRR but without sparsity regularization.

Fig.4 reports the performance of all four methods for an odds ratio of 1.125. To illustrate the behavior of the methods for different false positive rates, we report the receiver operating characteristic at two different noise levels. In supervised sRRR, we observe that using a standard univariate filtering p-value cutoff of 5% eliminates too many image regions and does not successfully allow for detection of genetic variants, leading to poor performance. We increased the success rate of sRRR by keeping the top 40% of regions sorted by their p-values. We found that sRRR results were robust when varying this parameter in a range around this larger percentage of regions to be included in the method. To set the detection thresholds, we fix the false positive rate to 1%. We observe 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, since there are so few relevant variants. 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 results of the CCA is consistently inferior with respect to sRRR. Given that sRRR can be viewed as CCA with sparsity constraints, this results emphasizes the importance of the sparsity regularization. The direct univariate t-tests only detect SNPs that have a very strong independent association with the disease label.

As more noise is added to the image, a two-step method starts to miss relevant regions across the image, which consequently degrades its detection rate on the genetic side. Our approach exploits other sources of information to compute the posterior probability of relevance. Namely, the $p(\mathbf{b}|\mathcal{D})$ has two terms. The second term in Eq. (10) summarizes the contribution of the genetic data which helps to compensate for the "image noise". In addition, genetics-to-image part of our model employs a powerful approach based on spike-and-slab prior. One can view spike-and-slab prior as a mixture of ℓ_0 and ℓ_2 regularization. This experiment shows that such regularization tends to perform better than ℓ_1 used in the sRRR approach. Better regularization and richer model explain the increased robustness of our approach compared to the "two-step" method.

V. Alzheimer's Disease Data

A. Data and Preliminary Evidence

Before applying our method to real data, we familiarize the reader with the data by demonstrating evidence of the association between the clinical diagnosis y, image data $\mathbf{X}$, and genotype $\mathbf{G}$ using a baseline approach.

We used clinical data from the ADNI study without focusing on a specific sub-group. ADNI is a large-scale study; the details on the study participants can be found elsewhere. The cohort includes 179 Alzheimer's patients (AD) and 198 healthy subjects (healthy) to the total of 377 subjects. We employed FreeSurfer image analysis suite[3] to process the MRI scans and produce segmentations and volume measurements for an array of regions (cortical and sub-cortical) that cover the entire brain. For details of these regions, please refer to Cortical ROIs[4], Desikan ROIs[5] in the FreeSurfer documentation. The technical details of these procedures are described in [29], [30], [31], [32], and [33].

To extract genetic variants, the standard quality control was applied to remove rare genetic variants or variants violating the Hardy-Weinberg Principle [28]. To reduce the number of SNPs considered by the algorithm, we removed SNPs that are unlikely to be associated with AD. We first imputed our genotype data to the 1000 Genomes panel using MaCH [34], then kept only SNPs whose p-value (as measured by a large-scale meta-analysis of AD [35]) was below a liberal threshold (10^{-3}), yielding 15,788 SNPs.

Fig.5 reports histograms of image features in four representative brain regions for the two cohorts of healthy and AD subjects. Two of these regions are highly relevant to the disease (entorhinal cortex and hippocampus [36]) while the other two have been less reported (putamen [37] and caudate) in the context of Alzheimer's disease. While the distribution of the cortical thickness in the left entorhinal cortex is strongly segregated across two cohorts, the right putamen and the left caudate volumes show weak or almost no statistical difference between the two populations. The entorhinal cortex is an important brain region responsible for declarative memories and memory consolidation and is implicated in early Alzheimer's disease [38].

To experiment with a classical baseline Genome-Wide Association (GWAS) methods, we fit several Generalized Linear Models (GLM) using the genotype $\mathbf{G}$ as the design matrix. In Fig.6, we used the image features from the four brain regions in Fig.5 as the response variable to the GLM. The Manhattan plot in Fig.6 shows $-\log_{10} p$-value for the genetic loci tested; the different shades of gray indicate different chromosomes. Despite the strong separation between healthy and AD in the left entorhinal cortex, no SNP passes the Bonferroni-corrected significance threshold. Nevertheless there is an indication for `APOE` variants. `APOE` is the only SNP that passes the significance level after the Bonferroni correction when the volume of the left hippocampus (Fig.6d) or clinical diagnosis y (not shown) are used as the response variables. Fig.6 therefore illustrates the limitation of classical GWAS.

B. Posterior Relevance of Brain Regions and SNPs

We applied our inference algorithm on the subset of the ADNI dataset described above. The algorithm shown integrates out the hyper-parameters through importance sampling. Only a range of hyper-parameters should be provided to the outer loop of algorithm, which translates to a weakly informative prior for the hyper-parameters. We choose the range for the hyper-parameters as follows:

σ_0 is the variance of the residual noise for the imaging features after they are explained by a subset of the genotype. For σ_0, we searched over $[0.2, 1]$. Since the imaging features are normalized to have unit variance, the variance of the residual is upper bounded by 1. We also do not expect the genetic variant to explain all the variance in the imaging feature, hence we expect a residual variance. It is common to impose a non-informative prior over σ_0 by assuming the inverse-gamma distribution for σ_0 and setting its shape parameter to a small quantity (here 0.05, Fig.2c in the Supplementary Materials).

For the variance of effect of individual SNPs σ_ω, we searched over $[0.025, 0.4]$. We do not expect a large contribution by a single SNP, but small contributions by several SNPs are possible. For this reason, the interval spans a small range. Notice that the variance of the residual, σ_0, is at most 1. In section II-F, we explained that the proportion of variance explained (PVE) can be used to impose a prior over σ_ω as suggested in [17] (Fig.2b in the Supplementary Materials).

To investigate the prior probability α of any SNP to be relevant, the range of $\log_{10}\alpha$ is set to $[-5, -3]$. For 15,788 SNPs, this is equivalent of selecting 0.1 to 16 SNPs as relevant to the endophenotype a priori. Two positive shape parameters of the beta distribution are set to 1.02 and 1 respectively which imposes almost uniform prior for the selected range of α (Fig.2a in the Supplementary Materials).

The posterior probability of the relevant SNPs (*i.e.,* $p(\mathbf{a}_m|\mathcal{D})$) is reported in Fig.7 for the brain regions examined in Fig.5 and Fig.6. The results of both approaches, *i.e.,* the proposed model and the classical approach of univariate tests, are relatively consistent. The least informative regions such as the caudate and putamen are assigned no SNPs by either methods. The hippocampus, which is known to be correlated with AD, is associated with a variant in `APOE`, a genetic

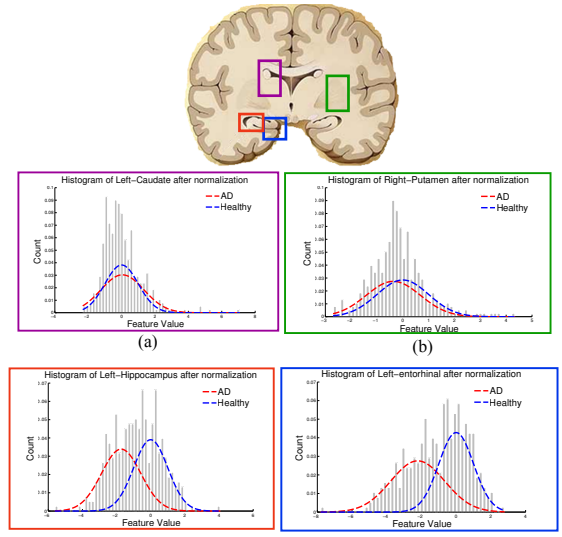


Fig. 5: Distribution of the imaging features for four different regions of the brain are shown. None or very weak differences can be seen between the groups for caudate and putamen while there are very strong differences in the volume of the left hippocampus and the average thickness of the entorhinal cortex.

[3]http://surfer.nmr.mgh.harvard.edu/
[4]https://surfer.nmr.mgh.harvard.edu/fswiki/CorticalParcellation
[5]http://freesurfer.net/fswiki/FsTutorial/AnatomicalROI

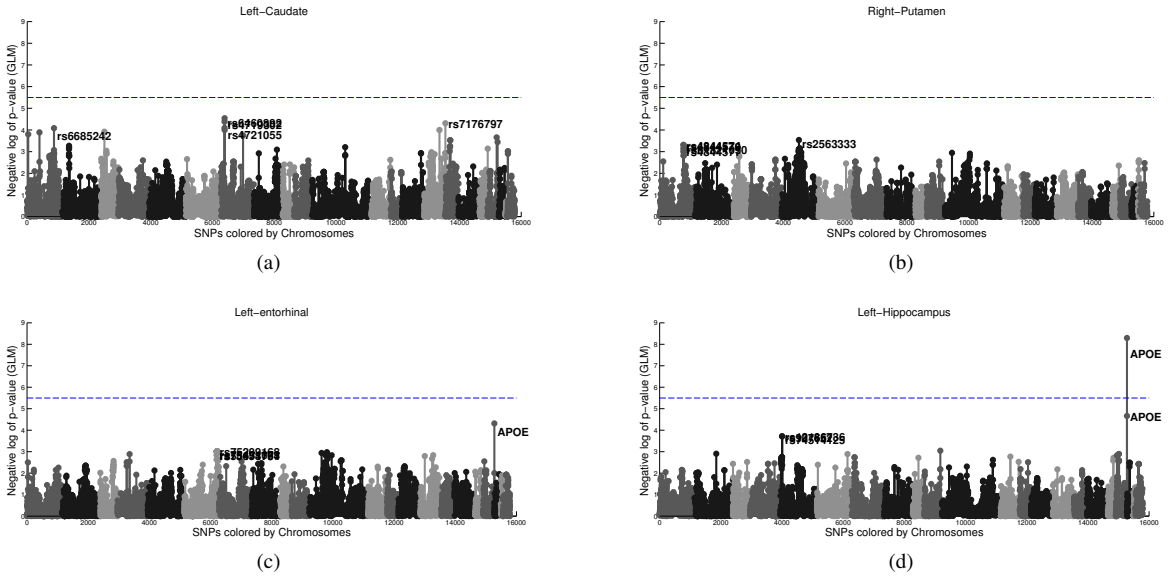


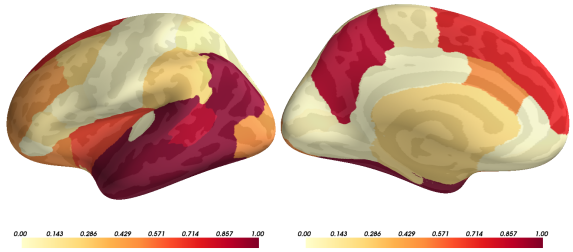
Fig. 6: Manhattan plots using different response variables in the GLM (a) volume of the left caudate (b) the volume of the right putamen (c) average cortical thickness of the left entorhinal cortex, and (d) volume of the left hippocampus. The x-axis lists the SNPs and the shades indicate different chromosomes. The y-axis reports the negative $\log_{10}$ of the p-value. The vertical line denotes the statistical significance level (0.05) after Bonferroni correction. Only APOE variants pass significance level, but only for the volume of the left hippocampus. In spite of a clear distinction between distributions of healthy and AD for the left entorhinal cortex (Fig.5), no SNP passes the significance level when using the average thickness of the left entorhinal cortex as a response variable.

marker known to be associated with Alzheimer's disease. For areas such as the entorhinal cortex, which is affected by AD [38], the classical method shows suggestive association for a variant in APOE, while for our method, APOE as well as a few others, pass the significance level.

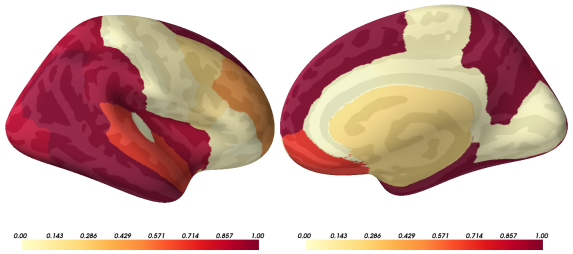
Interestingly, by computing the posterior relevance of brain regions $p(b_m = 1|\mathcal{D})$, we can go beyond the known regions of the brain affected by AD. Fig.9 reports the posterior probability of brain regions being relevant jointly for the genotype and the diagnosis. Fig.9a and Fig.9b show two hemispheres of the brain on medial and lateral views; the color indicates the posterior probability. Fig.8 represents the same results via a bar-plot. The y-axis is $p(b_m = 1|\mathcal{D})$. We sorted the regions according to the ranking produced by a classical correlation criterion (with respect to y). We observe that the classical statistical method and the results based on our model are largely consistent but our method assigns high posterior relevance to some regions that are viewed less important according to the classical test.

We emphasize that our method does not pool the genetic risk across ROIs. One can get a single set of posterior probability for all SNPs by summarizing overall association (see the Fig.10). This can be simply done by multiplying the posterior probability of the regions by the posterior probability of SNPs and summing over all brain regions that pass the $\frac{1}{2}$ threshold. Interestingly, the results are consistent with pair-wise association between genotype and diagnosis and only APOE passes the detection threshold. However, this does not mean that APOE is the only significant marker but it says that APOE is the one that almost all regions agree on due to its large effect. There is no reason to believe that genetic variants affect all regions equally. In fact, variations across locations is an interesting and worthy topic for further study.

In Fig.11, we investigate if regions with high posterior



(a) Left



(b) Right

Fig. 9: Posterior probability of the relevant regions (*i.e.,* $p(\mathbf{b}|\mathcal{D})$) for (a) left and (b) right hemispheres of the brain. Left and right figures in each row represent lateral and medial views respectively. The color indicates the value of the posterior probability, the hotter color, the higher the posterior.

relevance are related to AD, by examinining the importance of the features for prediction of the diagnosis. The x-axis is the number of features incrementally included in a linear classifier and y-axis is the cross-validation accuracy of the prediction of the diagnosis. Different curves denote rankings of the features according to the posterior values, correlation with diagnosis y, or random permutations (two instances). As we add more

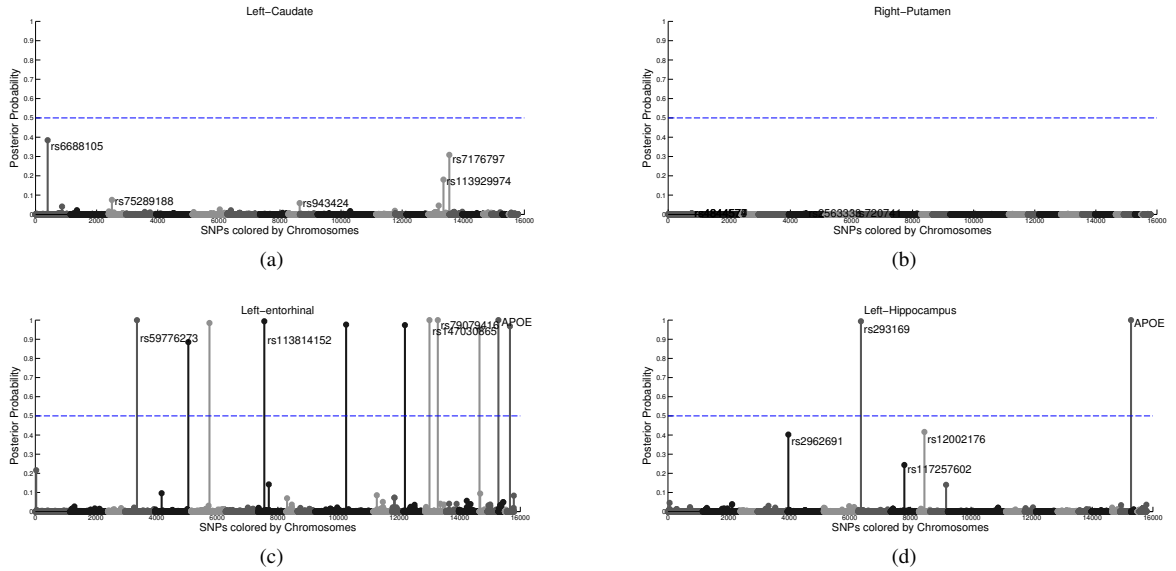


Fig. 7: Posterior relevance of the SNPs with respect to (a) volume of the left caudate, (b) right putamen, (c) average thickness of the left entorhinal cortex, and (d) volume of the left hippocampus, respectively. Compared to Fig.6. The horizontal line indicates $p = 0.5$.

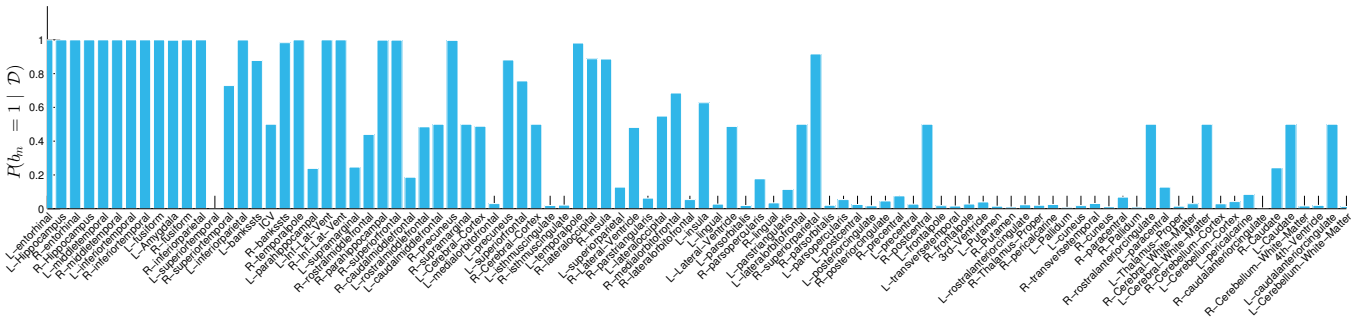


Fig. 8: The barplot of the posterior relevance for all 94 brain regions (y-axis). The regions are ordered according to the ranking produced by the two sample T-test with respect to y: We conducted a t-test to examine the difference between cases and controls for each one of these measurements and ranked them based on the t-test result.

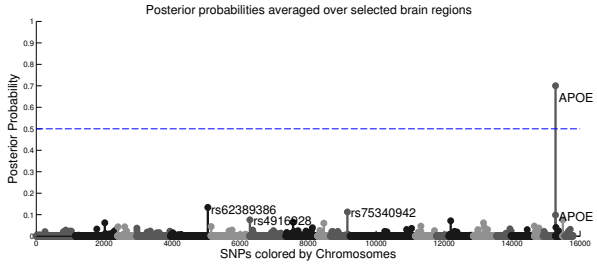


Fig. 10: Averaging regional posterior values across the selected brain regions. Only APOE is significant which means APOE is the one marker that many regions are consistently affected by.

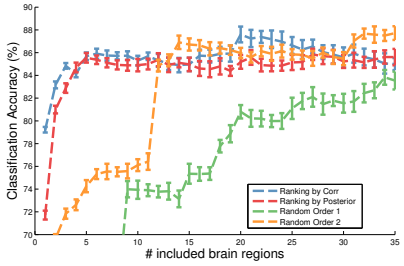
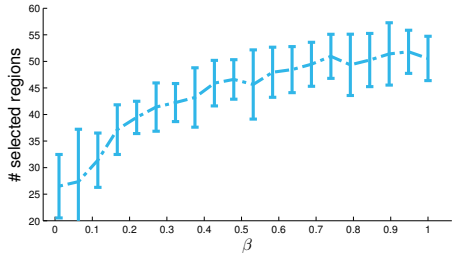


Fig. 11: Accuracy of the prediction of the disease for different number of input features ranked by correlation with disease diagnosis (blue), posterior produced by our method (red), and random ordering (orange and green). While our method and the correlation method jump quickly, it takes many more features for random ordering to match the accuracy of the informed methods.

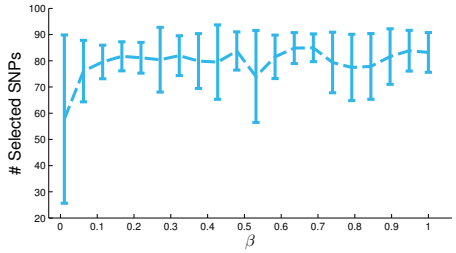
features, the accuracy of prediction increases. Our method closely follows the correlation ranking which indicates that the regions with high posterior values are closely related to the disease while the random rankings (*i.e.,* permutations) lag behind and need to include many features to finally match the accuracy of the informed methods. It is worth noting that correlation with diagnosis y only accounts for the diagnosis while the posterior values incorporate both genetic indicators and diagnosis simultaneously.

C. Sensitivity Analysis

In section II-F, we described the prior probabilities over variables α, σ_0^2, and σ_ω^2. The hyper-parameters of those variables are integrated out using importance sampling by gridding the hyper-parameters over the their corresponding intervals (see Supplementary Material). In section V-B, we explained how to choose these intervals depending on the meaning of



(a)



(b)

Fig. 12: (a) The number of selected image regions for different values of the prior β (*i.e.,* $\sum_m \left[p(b_m|\mathcal{D}) > \frac{1}{2}\right]$). (b) The total number of selected SNPs (*i.e.,* $\sum_s \left[\min_m \left[P\left(a_{sm}|\mathcal{D}\right)\right] > \frac{1}{2}\right]$).

the random variables and the data. In this section, we explore the sensitivity of the results with respect to the only remaining parameter β that specifies a prior number of relevant image regions. We change β from $\frac{1}{94}$ to $\frac{94}{94}$. For each value of β, we run the inference algorithm 20 times. Fig.12 reports the results.

We examine the number of brain regions with posterior probability higher than 0.5 computed as $\sum_m \left[p(b_m|\mathcal{D}) > \frac{1}{2}\right]$. Although this quantity increases with β, the model never chooses all regions, suggesting that some regions are not relevant regardless of the prior.

We also report the total number of selected SNPs ($\sum_s \left[p(\mathbf{a}_m|\mathcal{D}) > \frac{1}{2}\right]$) for different values of β. The curve plateaus at 80 quickly, suggesting that SNP selection is not very sensitive to the value of the prior. We can choose β in a reasonable range (depending on the application) with the least variance in Fig.12b. In all experiments of Section V-B, we set $\beta = \frac{10}{94}$ which lies in the plateaus region in Fig.12 and has low variance.

To study the behavior of the method empirically, we applied the model to the volume of left hippocampus as an intermediate phenotype (Fig.13). It shows that the number of detected SNP saturates as we include more SNPs in the model.

For 10^5 SNPs our algorithm takes about 24 hours to run. Other than computational cost, the problem with large number of SNPs is that the method starts missing APOE as the most important variant. We hypothesize it is due to small sample size and highly non-convex landscape of the objective function. Improving the stability of the method is an interesting direction of future research.

D. Biological pathway analysis

To investigate the molecular mechanisms through which these SNPs may be impacting brain morphology and AD phenotype, we mapped the 83 SNPs that were likely to target

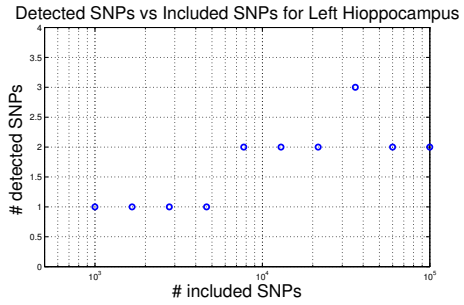


Fig. 13: The x-axis shows the number of SNPs included into the model, y-axis shows the number of selected SNPs when the volume of left hippocampus is used as the response variable in a Spike-and-Slab model.

at least one brain region to the nearest genes on the genome through the following procedure. We systematically filtered the 83 SNPs for dbSNP IDs and pruned the 83 SNPs based on linkage disequilibrium down to 77 SNPs. The pruning algorithm looks at all possible pairs of the 83 SNPs (for which their Pearson correlation is at least 0.2 from the 1000 Genomes Phase One European data [39]) and marks the SNP with lower rank for removal from the list. To determine SNP ranks, the algorithm first orders all SNPs by the number of brain regions in which their posterior is at least 0.5, then breaks ties based on the maximum posterior achieved in any brain region. We then mapped all SNPs to their nearest up and downstream protein-coding gene based on GENCODE version 10 annotations [40]. From the resulting list of 154 genes, we used Fisher's exact test to measure enrichment of our AD SNPs against 1024 known human pathways (whose size ranged from 5 to 300 genes inclusive) from the June 2011 release of the Pathway Commons database [41] (See Table I of Supplementary Materials for the list of SNPs and genes).

We found those nearest genes are significantly enriched in two biological pathways ($\alpha < 0.05$, Benjamini-Hochberg FDR), the Netrin signaling and the $\alpha4\beta1$ integrin pathways. Four genes proximal to our SNPs were direct interactors of the Netrin-1 protein complex (PITPNA, TRIO, MAP1B and DAPK1) within the Netrin signaling pathway. Netrin is a highly conserved protein involved in axon development, and is associated with negative regulation of amyloid-β production in the brains of Alzheimer's mice models [42], [43]. The amyloid-β peptide is the main component of amyloid plaques that is the hallmark of Alzheimer's Disease.

Four additional genes either formed direct complexes with, or directly interacted with, $\alpha4\beta1$ integrin, as part of the $\alpha4\beta1$ integrin signaling pathway. $\alpha4\beta1$ mediates permeation of blood barrier by leukocyte immune cells [44] and plays an important role both biologically and as a drug target in immune related diseases such as multiple sclerosis [45]. $\alpha4\beta1$ is not reported to be related to the Alzheimer's disease but it is consistent with recent work that suggests genetic variants associated with Alzheimer's disease target regulatory elements in leukocytes and other immune cells rather than brain cells [46], [47].

We also applied a separate regression and computed the residual to remove the effect of covariates (age, handedness, gender, and education). Then, we applied the algorithm on

the residual and noticed that the enrichment is not statistically significant. This suggests that the enrichment signal is weak and to correct for the effect of the covariates, they should be incorporated into Eq. (2).

VI. Discussion

In this paper, we propose a Bayesian method to identify indirect genetic associations with a diagnosis using image phenotype. Our model integrates two components: 1) selection of intermediate imaging phenotypes influenced by genetic markers and relevant to the disease and, 2) quantification of genetic associations with the disease mediated by the imaging variables. A classical strategy is to perform these two steps separately. First, an association analysis between imaging variables and disease phenotype is carried out. This step identifies imaging variables relevant to the disease status. Then, the associations between the relevant imaging markers and genotype data are probed. By performing these two tasks jointly, we avoid choosing an arbitrary threshold for feature selection.

We note that the model does not pool the genetic risk across ROIs. SNPs associated with complex diseases tend to act on cell type specific regulatory elements [48], suggesting that individual SNPs may be targeting specific cell types, and therefore brain regions. Furthermore, brain regions exhibit unique gene expression signatures [49] and epigenetic/regulatory signatures (Roadmap Epigenomics Consortium [50]), and therefore would be expected to use different sets of pathways to perform normal function.

Indeed, one can get a single set of posterior probabilities for all SNPs summarizing overall association (Fig.10). This can be simply achieved by multiplying the posterior probability of the regions by the posterior probability of SNPs and summing over all brain regions that pass a threshold of 0.5. Interestingly, the result is consistent with pair-wise associations between the genotype and diagnosis and only APOE passes the 0.5 threshold. However, this does not mean that only APOE is the significant marker but rather that APOE is the marker that almost all regions are consistently affected by.

In this paper, we assumed that genetic variants related to the disease encode variations measurable by imaging data. This assumption has some limitations. For example, if the genetic variants related to the disease do not manifest themselves on the imaging data, our method cannot detect it. Another limitations is for the genetic variants that have both normal and disease-related effects; such case is not identifiable by our model but to the best of our knowledge it is not identifiable by other approaches as well. These challenges provide fruitful directions for future work.

In this paper, we assume that genetic variants $\mathbf{G}$ have indirect associations to the disease label y. In other words, we assume that all relevant genetic associations are already captured by the image features. It is conceivable that some of the variants have a direct association, *i.e.,* their impact is not captured by the imaging features. It is possible to extend the graphical model to incorporate such effects by introducing a direct connection from $\mathbf{G}$ to y. Such a change in the graphical model renders the inference procedure more complex.

Our model ranks brain regions based on the amount of variance of imaging features explained by the genotype. The ranking of the regions gets updated according to the relevance of the brain regions to the diagnosis. The proposed procedure approximates two posterior probabilities, $p(\mathbf{b}|\mathcal{D})$ and $p(\mathbf{a}_m|\mathcal{D})$, denoting the relevance of image regions for the disease and of the genotype related to those regions, respectively.

There are two major reasons for using region-based image features: statistical and computational. Statistically, aggregate measures such as region-based image features provide more robust estimators at the expense of a coarser resolution on delineating affected brain regions. From the computational point of view, reducing the number of brain regions (fewer b_m) reduces the computational cost of Algorithms 1. Every iteration of Algorithm 1 entails solving a linear system with $O(M)$ (M is the number of brain regions) variables.

We use the language of directed graphical models to formalize our assumptions. We use Gaussian Process (GP) to model the diagnosis. The GP framework is flexible, enabling a range of functions (*i.e.,* f in the graphical model) to be used by simply changing the kernel function. To extend the method to regression (*i.e.,* continuous y), one needs to modify the likelihood function in Eq. (1) and to modify a noise model. Interestingly, for the regression case with the Gaussian noise, the marginal likelihood $\mathbb{P}(\mathbf{y}|\mathbf{b},\mathcal{D})$ has a closed-form solution and one does not need to resort to Expectation Propagation (EP) for approximation. Many noise models were investigated in [16], deriving efficient algorithms to approximate the marginal likelihood for many members of the exponential family.

The image-to-disease phenotype part of the model can be extended such that the diagnosis variable y encodes finer levels of diagnosis. For example, we can replace the logistic regression likelihood with the ordinal logistic likelihood [51] to encode discrete and ordered observations about the disease (Healthy $(j=0)<$ MCI $(j=1)<$ AD $(j=2)$):

$$p(y\leq j|\mathbf{x},\boldsymbol{\nu},\theta_1,\theta_2,\theta_3)=\frac{1}{1+\exp(\mathbf{x}^T\boldsymbol{\nu}-\theta_j)}, \quad (18)$$

where $\boldsymbol{\nu}$ and the $\theta_1<\theta_2<\theta_3$ are the parameters of the ordinal logistic regression and j encodes the ordinal stage of the Alzheimer's disease.

We model the null distribution of the image regions with a Gaussian distribution. This assumption can be easily modified by replacing the Gaussian distribution with any other distribution depending on the application. The noise model for the alternative hypothesis (*i.e.,* $b_m=1$) can also be modified. The challenge is to compute the marginal likelihood efficiently (*i.e.,* $\mathbb{P}(\mathbf{X}_{:m}|b_m=1;\mathbf{G})$). We approximate this value by the lower bound provided by the variational approximation. Our current implementation supports the Gaussian noise assumption for imaging features $\mathbf{X}_{:m}$. We leave the relaxation of this assumption to future work. We believe, at least for the most common members of the exponential family, slight modification to the variational algorithm should be possible.

The hidden random variable $\mathbf{b}$ encodes the relevant regions. Therefore, the kernel depends on $\mathbf{b}$. For example the linear kernel between two samples x_i and x_j should be defined as

$$k_{\mathbf{b}}(\mathbf{x},\mathbf{x}')=(\mathbf{x}\mathbf{b})^T(\mathbf{x}'\mathbf{b})$$

Note that $\mathbf{b}$ appears in the definition of the kernel. We chose the linear kernel because of its simplicity. Although it is possible to use a complex kernel together with a regularization, we avoided it because of two reasons. First, this would introduce extra parameter (*e.g.,* kernel width in case of Radial Basis Function). Second, the value of such parameter would depend on an unknown parameter $\mathbf{b}$. In the case of RBF, kernel width should scale with the dimensionality of the input vector. In our case, the input vector consists of the relevant regions selected by the indicator variable $\mathbf{b}$. Note that this is not the case in the classical kernel-based approaches where the prediction is the only goal but not features selection. Previously demonstrated methods for feature selection using kernel machines [52] lack a probabilistic model required by our approach. Further extension of our model to those cases is possible but beyond the scope of this paper.

In addition to minor modifications to the structure of the graphical model compared to our previous work [15], there are several major innovations introduced in this paper. First, in the image-to-phenotype part of the model, we employed the Gaussian process to model the prediction function. This modification enables us to model the complex relationship between image and clinical phenotypes. In this paper, we focused on the linear kernel to avoid over-fitting but in the presence of more samples a more sophisticated prediction function can be reliably learned. The second major contribution is in the inference algorithm. It is more stable and scalable than our earlier inference method in [15]. The flexibility of the inference algorithm enabled us to go beyond conditionally independent intermediate phenotypes. For example, we are currently pursuing the case where intermediate phenotypes are highly correlated. In this case, two intermediate phenotypes (*e.g.,* two brain regions) which are highly correlated should be viewed as approximately one phenotype. One can account for this phenomenon by modifying the prior probability of $p(\mathbf{b})$ of the selector variable $\mathbf{b}$. As long as we can sample efficiently from $p(\mathbf{b})$, the inference algorithm is computationally tractable.

Two key quantities that determine the computational complexity of the inference algorithm are the marginal likelihoods $p(\mathbf{y}|\mathbf{b}, \mathcal{D})$ and $p(\mathbf{X}_{:m}|b_m = 1; \mathbf{G})$. If no value is missing from the intermediate phenotypes, $p(\mathbf{X}_{:m}|b_m = 1; \mathbf{G})$ can be computed in parallel and stored. $p(\mathbf{y}|\mathbf{b}, \mathcal{D})$ needs to be computed for every draw of $\mathbf{b}$. We use expectation propagation (EP), which is very fast, particularly for the small sample size prevalent in imaging genetic applications (*cf.* [16] Section 3.6).

As suggested in [17], fast computation of the variational lower bound enables us to perform importance sampling and to integrate out all hyper-parameters other than the image feature selection prior β. Since we have few hyper-parameters, we only need to specify a reasonable range for each hyper-parameter. This approach also enables us to define a weakly-informed prior over the hyper-parameters. Depending on the meaning of each hyper-parameter, we defined a range that is reasonable for the application. We provide an example in Section V-B on how to choose the intervals. In Section V-C, we show that the total number of SNPs detected by the inference is not very sensitive to the specific value of β.

In Section V-B, we compared the associated SNPs to the relevant brain regions using the p-values and the posterior probabilities. Although the p-value and the posterior probability do not have the same meaning, their suggestions about the data are relatively consistent. We showed in Fig.6 and Fig.7 that the less important regions such as the putamen and the caudate do not exhibit associations in either method. Both techniques agree on the SNPs associated with left hippocampus. For the left entorhinal cortex, our method detects a few more SNPs in addition to the `APOE` variants. Furthermore, our method suggests areas to investigate further. Fig.8 showed that posterior relevance values are mostly consistent with a classical ranking results but the proposed method does not require pre-selection and considers all available data.

The results reported for the univariate approach used Bonferroni correction which is a common practice in genetic association. Bonferroni correction is a conservative multiple hypothesis correction approach in comparison to controlling false discovery. In fact, one can further analyze the results reported by our approach and apply the hypothesis testing using the image features of the detected brain regions as a response variable of a GLM and correct the results with a method of choice. Our focus has been on how to incorporate information from different sources, here diagnosis, imaging and genetics data, into one model, and not on addressing multiple hypothesis correction approaches.

VII. Conclusion

We proposed and demonstrated a unified framework for identifying genetic variants and image-based features associated with the disease. We captured the associations between imaging and disease phenotype simultaneously with the correlation from genetic variants and image features in a probabilistic model. Our model also produces spatial distribution of the genetic associations. We derive an efficient and scalable algorithm based on variational inference. We did not assume any interaction between intermediate phenotypes (*i.e.,* imaging features) but our method can be extended easily to handle such interactions. We demonstrated the benefit of simultaneously performing these two tasks (*i.e.,* finding relevant genetic and brain regions) in simulations and in a context of a real clinical study of the Alzheimer's disease.

Acknowledgments 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, Barbara J. Weedon Fellowship, and Wistron Corporation. We would like to thank the Massachusetts Green High Performance Computing Center for providing computational resources for this project.

References

[1] C. J. Hoggart, J. C. Whittaker, M. De Iorio, and D. J. Balding, "Simultaneous analysis of all snps in genome-wide and re-sequencing association studies," *PLoS Genet*, vol. 4, no. 7, p. e1000130, 2008.

[2] D. Lvovs *et al.*, "A polygenic approach to the study of polygenic diseases," *Acta Naturae*, vol. 4, no. 3, p. 59, 2012.

[3] A. Meyer-Lindenberg and D. R. Weinberger, "Intermediate phenotypes and genetic mechanisms of psychiatric disorders," *Nature Reviews Neuroscience*, vol. 7, no. 10, pp. 818–827, 2006.

[4] D. C. Glahn, P. M. Thompson, and J. Blangero, "Neuroimaging endophenotypes: strategies for finding genes influencing brain structure and function," *Human brain mapping*, vol. 28, no. 6, pp. 488–501, 2007.

[5] N. K. Batmanghelich *et al.*, "Generative-discriminative basis learning for medical imaging." *IEEE Trans Med Imaging*, vol. 31, no. 1, pp. 51–69, Jan. 2012.

[6] M. Sabuncu *et al.*, "The Relevance Voxel Machine (RVoxM): A Bayesian Method for Image-Based Prediction," in *MICCAI 2011*, ser. LNCS, T. Peters, G. Fichtinger, and A. Martel, Eds. Springer, Heidelberg, 2011, pp. 99–106.

[7] N. Filippini *et al.*, "Anatomically-distinct genetic associations of APOE epsilon4 allele load with regional cortical atrophy in Alzheimer's disease." *Neuroimage*, vol. 44, no. 3, pp. 724–728, Feb 2009.

[8] S. Potkin *et al.*, "A genome-wide association study of schizophrenia using brain activation as a quantitative phenotype." *Schizophr Bull*, vol. 35, no. 1, pp. 96–108, Jan 2009.

[9] M. R. Sabuncu, R. L. Buckner, J. W. Smoller, P. H. Lee, B. Fischl, R. A. Sperling *et al.*, "The association between a polygenic alzheimer score and cortical thickness in clinically normal subjects," *Cerebral cortex*, vol. 22, no. 11, pp. 2653–2661, 2012.

[10] J. Stein *et al.*, "Voxelwise genome-wide association study (vGWAS)." *Neuroimage*, vol. 53, no. 3, pp. 1160–1174, Nov 2010.

[11] E. Le Floch *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*, vol. 63, no. 1, pp. 11–24, Oct 2012.

[12] M. Vounou *et al.*, "Discovering genetic associations with high-dimensional neuroimaging phenotypes: A sparse reduced-rank regression approach." *Neuroimage*, vol. 53, no. 3, pp. 1147–1159, Nov 2010.

[13] ——, "Sparse reduced-rank regression detects genetic associations with voxel-wise longitudinal phenotypes in Alzheimer's disease." *Neuroimage*, vol. 60, no. 1, pp. 700–716, Mar 2012.

[14] S. Mueller *et al.*, "The alzheimer's disease neuroimaging initiative," *Neuroimaging Clinics of North America*, vol. 15, no. 4, p. 869, 2005.

[15] N. K. Batmanghelich *et al.*, "Joint modeling of imaging and genetics," in *Information Processing in Medical Imaging*. Springer, 2013, pp. 766–777.

[16] C. E. Rasmussen and C. K. I. Williams, *Gaussian Processes for Machine Learning* . The MIT Press, 2006.

[17] P. Carbonetto and M. Stephens, "Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies," *Bayesian Analysis*, vol. 7, pp. 73–108, 2012.

[18] T. J. Mitchell and J. J. Beauchamp, "Bayesian variable selection in linear regression," *Journal of the American Statistical Association*, vol. 83, no. 404, pp. 1023–1032, 1988.

[19] A. Gelman *et al.*, "Prior distributions for variance parameters in hierarchical models (comment on article by browne and draper)," *Bayesian analysis*, vol. 1, no. 3, pp. 515–534, 2006.

[20] H. Jeffreys, "An Invariant Form for the Prior Probability in Estimation Problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[21] Y. Guan, M. Stephens *et al.*, "Bayesian variable selection regression for genome-wide association studies and other large-scale problems," *The Annals of Applied Statistics*, vol. 5, no. 3, pp. 1780–1815, 2011.

[22] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

[23] C. M. Bishop, *Pattern recognition and machine learning*. New York: Springer, 2006.

[24] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 1–10, 2003.

[25] T. Salimans, D. A. Knowles *et al.*, "Fixed-form variational posterior approximation through stochastic linear regression," *Bayesian Analysis*, vol. 8, no. 4, pp. 837–882, 2013.

[26] A. Honkela *et al.*, "Approximate Riemannian conjugate gradient learning for fixed-form variational Bayes," *The Journal of Machine Learning Research*, vol. 9999, pp. 3235–3268, 2010.

[27] A. Nemirovski *et al.*, "Robust stochastic approximation approach to stochastic programming," *SIAM Journal on Optimization*, vol. 19, no. 4, pp. 1574–1609, 2009.

[28] S. Purcell *et al.*, "PLINK: a tool set for whole-genome association and population-based linkage analyses," *Am J Hum Genet*, vol. 81, no. 3, pp. 559–575, Sep. 2007.

[29] B. Fischl *et al.*, "Automatically parcellating the human cerebral cortex," *Cerebral cortex*, vol. 14, no. 1, pp. 11–22, 2004.

[30] A. M. Dale, B. Fischl, and M. I. Sereno, "Cortical surface-based analysis: I. segmentation and surface reconstruction," *Neuroimage*, vol. 9, no. 2, pp. 179–194, 1999.

[31] B. Fischl, M. I. Sereno, and A. M. Dale, "Cortical surface-based analysis: Ii: inflation, flattening, and a surface-based coordinate system," *Neuroimage*, vol. 9, no. 2, pp. 195–207, 1999.

[32] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. Van Der Kouwe, R. Killiany, D. Kennedy, S. Klaveness *et al.*, "Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain," *Neuron*, vol. 33, no. 3, pp. 341–355, 2002.

[33] B. Fischl and A. M. Dale, "Measuring the thickness of the human cerebral cortex from magnetic resonance images," *Proceedings of the National Academy of Sciences*, vol. 97, no. 20, pp. 11 050–11 055, 2000.

[34] Y. Li *et al.*, "Mach: using sequence and genotype data to estimate haplotypes and unobserved genotypes," *Genetic epidemiology*, vol. 34, no. 8, pp. 816–834, 2010.

[35] J.-C. Lambert *et al.*, "Meta-analysis of 74,046 individuals identifies 11 new susceptibility loci for alzheimer's disease," *Nature genetics*, 2013.

[36] P. J. Whitehouse, D. L. Price, R. G. Struble, A. W. Clark, J. T. Coyle, and M. R. Delon, "Alzheimer's disease and senile dementia: loss of neurons in the basal forebrain," *Science*, vol. 215, no. 4537, pp. 1237–1239, 1982.

[37] L. De Jong, K. Van der Hiele, I. Veer, J. Houwing, R. Westendorp, E. Bollen, P. De Bruin, H. Middelkoop, M. Van Buchem, and J. Van Der Grond, "Strongly reduced volumes of putamen and thalamus in alzheimer's disease: an mri study," *Brain*, vol. 131, no. 12, pp. 3277–3285, 2008.

[38] U. A. Khan *et al.*, "Molecular drivers and cortical spread of lateral entorhinal cortex dysfunction in preclinical Alzheimer's disease," *Nature neuroscience*, vol. 17, no. 2, pp. 304–311, 2014.

[39] T. . G. P. Consortium, "An integrated map of genetic variation from 1,092 human genomes," *Nature*, 2012.

[40] J. Harrow, Frankish *et al.*, "Gencode: The reference human genome annotation for the encode project," *Genome Research*, 2012.

[41] E. Cerami, B. Gross, E. Demir, I. Rodchenkov, O. Babur, N. Anwar, N. Schultz, G. Bader, and C. Sander, "Pathway commons, a web resource for biological pathway data," *Nucleic Acids Research*, 2011.

[42] N. Rama, D. Goldschneider, V. Corset, J. Lambert, L. Pays, and P. Mehlen, "Amyloid precursor protein regulates netrin-1-mediated commissural axon outgrowth," *Journal of Biological Chemistry*, vol. 287, no. 35, pp. 30 014–30 023, 2012.

[43] F. Lourenco *et al.*, "Netrin-1 interacts with amyloid precursor protein and regulates amyloid- production," *Cell Death and Differentiation*, 2009.

[44] Y. Takeshita and R. Ransohoff, "Inflammatory cell trafficking across the bloodbrain barrier: chemokine regulation and in vitro models," *Immunological Reviews*, 2012.

[45] K. A. Buzzard, S. A. Broadley, and H. Butzkueven, "What do effective treatments for multiple sclerosis tell us about the molecular mechanisms involved in pathogenesis?" *International journal of molecular sciences*, vol. 13, no. 10, pp. 12 665–12 709, 2012.

[46] E. Gjoneska, A. R. Pfenning, H. Mathys, G. Quon, A. Kundaje, L.-H. Tsai, and M. Kellis, "Conserved epigenomic signals in mice and humans reveal immune basis of alzheimer/'s disease," *Nature*, vol. 518, no. 7539, pp. 365–369, 2015.

[47] T. Raj, K. Rothamel, S. Mostafavi, C. Ye, M. N. Lee, J. M. Replogle, T. Feng, M. Lee, N. Asinovski, I. Frohlich *et al.*, "Polarization of the effects of autoimmune and neurodegenerative risk alleles in leukocytes," *Science*, vol. 344, no. 6183, pp. 519–523, 2014.

[48] M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody *et al.*, "Systematic localization of common disease-associated variation in regulatory dna," *Science*, vol. 337, no. 6099, pp. 1190–1195, 2012.

[49] M. J. Hawrylycz, E. S. Lein, A. L. Guillozet-Bongaarts, E. H. Shen, L. Ng, J. A. Miller, L. N. van de Lagemaat, K. A. Smith, A. Ebbert, Z. L. Riley *et al.*, "An anatomically comprehensive atlas of the adult human brain transcriptome," *Nature*, vol. 489, no. 7416, pp. 391–399, 2012.

[50] A. Kundaje, W. Meuleman, J. Ernst, M. Bilenky, A. Yen, A. Heravi-Moussavi, P. Kheradpour, Z. Zhang, J. Wang, M. J. Ziller *et al.*, "Integrative analysis of 111 reference human epigenomes," *Nature*, vol. 518, no. 7539, pp. 317–330, 2015.

[51] T. F. Liao, *Interpreting probability models: Logit, probit, and other generalized linear models*. Sage, 1994, no. 101.

[52] G. I. Allen, "Automatic feature selection via weighted kernels and regularization," *Journal of Computational and Graphical Statistics*, vol. 22, no. 2, pp. 284–299, 2013.

Supplementary Materials: Probabilistic Modeling of Imaging, Genetics and Diagnosis

Nematollah K. Batmanghelich, Adrian Dalca, Gerald Quon, Mert Sabuncu, Polina Golland, for the Alzheimer's Disease Neuroimaging Initiative*

I. Further Explanation about the Experiments on ADNI

Fig.1 shows the probability densities of the hyper-parameters used for the ADNI dataset.

List of the SNPs and the corresponding genes are provided in the Table I.

II. Variational Bayes to Approximate the Posteriors

The pseudo-code for our inference algorithm is shown in Algorithm 1, where $\boldsymbol{\gamma}$ is a M-dimensional vector of Bayes Factor computed as follows:

$$\log p(\mathbf{X}_{:m}|b_m;\mathbf{G},\pi) = \sum_{n=1}^{N}\log\mathcal{N}(x_{nm};0,1)+ b_m\left(\log p(\mathbf{X}_{:m}|b_m=1;\mathbf{G},\pi) - \sum_{n=1}^{N}\log\mathcal{N}(x_{nm};0,1)\right) \quad (1)$$

The computationally difficult term in the equation is the marginal likelihood, $p(\mathbf{x}_m|\mathbf{b}_m = 1; G, \pi)$. Exact computation of the marginal likelihood is computationally intractable. However, it is common to approximate it with a lower bound of the variational energy [1]–[3]. We follow the variational mean-field method proposed by Carbonetto *et al.* [4] with a slight modification to approximate $p(\mathbf{X}_{:m}|\mathbf{b}_m = 1; G, \pi)$. To be self-contained, we first briefly summarize the method in [4]:

- We descretize the hyper-parameter space of the imaging part of the model, *i.e.,* $\pi' := \{\log_{10}\alpha, \sigma_0^2, \sigma_\omega^2\} \subset \pi$ into uniform grids, namely $[\alpha(\min), \alpha(\max)] \times [\sigma_0^2(\min), \sigma_0^2(\max)] \times [\sigma_\omega^2(\min), \sigma_\omega^2(\max)]$. Let us call the grid points $\pi'(1), \cdots, \pi'(L)$, where every tuple $\pi' = (\alpha(i), \sigma_0^2(i), \sigma_\omega^2(i))$ is a set of hyper-parameter values.
- Since the space of the hyper-parameters is low dimensional, importance sampling is a simple and effective way to integrate out the hyper-priors with a reasonably small number of samples. The proposal distribution is chosen to be a uniform distribution over a sufficiently large range, *i.e.,* $\tilde{p}(\pi'(i)) = \tilde{p}(\pi'(1))$, where $\tilde{p}(\cdot)$ is the proposal distribution.
- Given a set of hyper-parameter values $\pi'(i)$, a mean-field approach is used to approximate the marginal likelihood, $p(\mathbf{x}_m|\mathbf{b}_m = 1; G, \pi')$, via the variational lower bound. Briefly, the mean-field method maximizes the following objective function:

$$\begin{aligned}
&\log p(\mathbf{x}|b=1;\mathbf{G},\pi') \\
&\geq F(\pi';\boldsymbol{\varsigma},\boldsymbol{\nu},\boldsymbol{\tau}) \equiv \mathbb{E}_q\left[\log\frac{p(\mathbf{x},\boldsymbol{\omega},\mathbf{a}|\mathbf{G},\pi')}{q(\boldsymbol{\varsigma},\boldsymbol{\nu},\boldsymbol{\tau})}\right] \\
&= -N\log(\sigma_0) - \frac{\|\mathbf{x}-\mathbf{G}\mathbf{r}\|^2}{2\sigma_0^2} \\
&-\frac{1}{2\sigma_0^2}\sum_{k=1}^{S}\left(\mathbf{G}^T\mathbf{G}\right)_{kk}\mathbb{V}ar_q[\omega_k] \\
&-\sum_{k=1}^{S}\tau_k\log\left(\frac{\tau_k}{\alpha}\right) - \sum_{k=1}^{S}(1-\tau_k)\log\left(\frac{1-\tau_k}{1-\alpha}\right) \\
&+\sum_{k=1}^{S}\frac{\tau_k}{2}\left[1+\log\left(\frac{\varsigma_k^2}{\sigma_0^2\sigma_\omega^2}\right) - \frac{\varsigma_k^2+\nu_k^2}{\sigma_0^2\sigma_\omega^2}\right]
\end{aligned} \quad (2)$$

  where $\boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$ are the parameters of the approximate posterior, q and $\mathbf{r} := \mathbb{E}[\boldsymbol{\omega}_q] = \boldsymbol{\tau} \odot \boldsymbol{\nu}$ and $\mathbb{V}ar_q[\omega_k] = \tau_k(\varsigma_k^2 + \nu_k^2) - (\tau_k\nu_k)^2$.
- Finally, the importance weights, $\zeta(i)$'s, are normalized. For each of $\boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau}$, a weighted sum over all $\pi'(i)$ is computed as an approximation to integrating out the hyper-parameters.

This procedure needs is run for every brain region. The pseudo-code of the algorithm is shown in Algorithm 2 [4].

To integrate out the hyper-priors, the main idea in [4] is to use importance sampling to compute the following integral:

$$\text{PIP}(s,m) = \int p(a_{sm}=1|\mathbf{G},\mathbf{x}_m,\pi')p(\pi'|\mathbf{G},\mathbf{x}_m)d\pi', \quad (3)$$

where $\text{PIP}(s,m)$ denotes the Posterior Inclusion Probability for SNP s and region m. Carbonetto *et al.* [4] suggest to replace it with the following importance sampling estimate

$$\text{PIP}(s,m) = \frac{\sum_{i=1}^{L} p(a_{sm}=1|\mathbf{G},\mathbf{x}_m,\pi'(i))\zeta(\pi'(i))}{\sum_{i=1}^{L}\zeta(\pi'(i))} \quad (4)$$

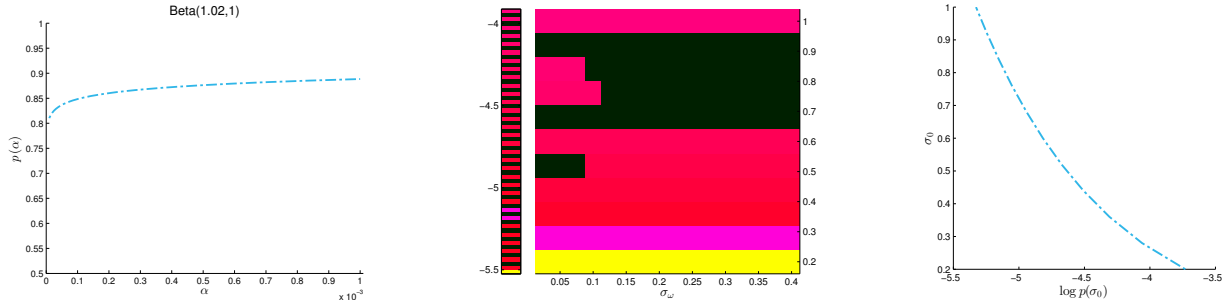
where $\zeta(\pi'(i))$ is the normalized importance weight for $\pi'(i)$. According to the importance sampling procedure,

$$\zeta(\pi') = \frac{p(\mathbf{x}_m|\mathbf{G},\pi')p(\pi')}{\tilde{p}(\pi')}, \quad (5)$$

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.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.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

SNP	Gene	SNP	Gene	SNP	Gene	SNP	Gene
rs10106827	NECAB1,TMEM55A	rs12476069	COL6A3,MLPH	rs9393059	FOXQ1,HUS1B	rs71327107	EP300,RBX1
rs10487075	C7orf62,STEAP2-AS1	rs12535226	EGFR,LANCL2	rs2906657	PILRA,ZCWPW1	rs74322721	DDHD1,FERMT2
rs10504488	EYA1,XKR9	rs12778247	ANXA8,ZNF488	rs293168	NDUFA4,NXPH1	rs75340942	HABP2,TCF7L2
rs10812555	C9orf11,LINC00032	rs12997264	ATG16L1,INPP5D	rs34380708	KIAA0317,LTBP2	rs7536931	CR1,CR1L
rs111863968	ATG16L1,INPP5D	rs13040601	CBLN4,DOK5	rs3764648	ABCA7,HMHA1	rs76222305	LRRTM1,SUCLG1
rs113814152	CHRNA2,PTK2B	rs13138250	FGFRL1,IDUA	rs3779632	CHRNA2,PTK2B	rs76448372	AMICA1,SCN2B
rs114773661	CRBN,SUMF1	rs13314819	BBX,CCDC54	rs4133300	KCNJ3,NR4A2	rs76822114	GC,SLC4A4
rs114956101	KCNK17,KCNK5	rs145767144	MAF,WWOX	rs4916928	.,FAM20C	rs76978231	CSMD1,MCPH1
rs115815527	ASB5,SPCS3	rs146373627	ECHDC3,USP6NL	rs56034708	CDC7,TGFBR3	rs77271157	CLU,EPHX2
rs11662059	ACAA2,LIPG	rs146643250	DNAH5,TRIO	rs57677986	ADAM10,FAM63B	rs77287774	ZEB1-AS1,ZNF438
rs117119586	MTDH,TSPYL5	rs147030865	ATP8B4,SLC27A2	rs59776273	CORIN,NFXL1	rs7812465	PLEKHF2,TP53INP1
rs117281307	CTTNBP2,NAA38	rs16849237	RHOU,TMEM78	rs6020063	B4GALT5,SLC9A8	rs78180796	PTPRM,RAB12
rs117547283	POM121L1P,PRAME	rs17108960	CBX5,SMUG1	rs622354	OR10G7,VWA5A	rs79079416	CSNK1G1,KIAA0101
rs117655211	ATP8B4,DTWD1	rs17781348	GAK,TMEM175	rs62389386	CLK4,COL23A1	rs792806	CA10,KIF2B
rs1178036	GNRH2,PTPRA	rs1806522	C3orf27,RPN1	rs6571632	EGLN3,SPTSSA	rs8030340	RSL24D1,UNC13C
rs117984432	ANKRD11,SPG7	rs1834554	MS4A4A,MS4A4E	rs6685242	CD46,CR1L	rs8707	MAP3K12,PCBP2
rs118091716	GRM3,SEMA3D	rs1912718	ATOH1,GRID2	rs6934812	CD2AP,TNFRSF21	rs2824734	LINC00320,TMPRSS15
rs118192075	IMPACT,OSBPL1A	rs2048330	LINC00210,RRP15	rs6949677	C7orf70,CYTH3	rs28592859	HLA-DQA1,HLA-DQB1
rs11875667	CETN1,COLEC12	rs2048330	LINC00210,RRP15	rs7027316	IFNE,MTAP	$\epsilon 3/\epsilon 4$	APOE
rs12002176	C9orf170,DAPK1	rs2136987	CCKAR,RBPJ	rs7068614	ECHDC3,USP6NL		
rs12137076	GNG4,LYST	rs2701623	DTX1,RASAL1	rs7087150	CCNY,GJD4		
rs12198405	CD2AP,TNFRSF21	rs79914380	LMO3,MGST1	rs7129687	EED,PICALM		

TABLE II: Detected SNPs and the corresponding genes.



(a) (b) (c)

Fig. 1: Prior probability density for different hyper-parameters: (a) Density of α for prior distribution of Beta$(1.02, 1)$. The prior assigns almost uniform weight to all values. (b) log of PVE density as a function of σ_ω (x-axis) and σ_0 (y-axis). PVE functions as a prior for σ_ω given a value of σ_0. (c) log of the density of σ_0.

where the numerator is proportional to $p(\pi'|\mathbf{x}_m, \mathbf{G})$ and the denominator is the proposal distribution, a uniform distribution in our experiments. Since the marginal likelihood $p(\mathbf{x}_m|\mathbf{G}, \pi')$ cannot be computed directly, it is approximated by the highest lower bound:

$$\log p(\mathbf{x}_m|\mathbf{G}, \pi') \geq F(\pi'; \boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau}). \tag{6}$$

To approximate the $p(\mathbf{x}_m|\mathbf{G})$, we can apply the following procedure:

$$\begin{aligned} p(\mathbf{x}_m|\mathbf{G}) &= \int p(\mathbf{x}_m, \pi'|\mathbf{G}) dp(\pi'|\mathbf{G}) = \mathbb{E}_{\pi'|\mathbf{G}}\left[p(\mathbf{x}_m, \pi'|\mathbf{G})\right] \\ &\geq \mathbb{E}_{\pi'|\mathbf{G}}\left[e^{F(\pi'; \boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau})}\right] \\ &\geq \exp\left[\mathbb{E}_{\pi'|\mathbf{G}}\left[F(\pi'; \boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau})\right]\right], \end{aligned} \tag{7}$$

where the last line in Eq. (7) follows from the convexity of exponential function.

Similar to Eq. (4), the idea is to replace the expectation with the importance sampling approximation:

$$\mathbb{E}_{\pi'|\mathbf{G}}\left[F(\pi'; \boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau})\right] \approx \frac{\sum_{i=1}^{L} F(\pi'; \boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \zeta(\pi'(i))}{\sum_{i=1}^{L} \zeta(\pi'(i))}. \tag{8}$$

REFERENCES

[1] M. Beal and Z. Ghahramani, "The variational Bayesian EM algorithm for incomplete data: with application to scoring graphical model structures," *Bayesian Statistics*, vol. 7, pp. 1–10, 2003.

[2] R. B. Grosse, "Model selection in compositional spaces," Ph.D. dissertation, Massachusetts Institute of Technology, 2014.

[3] C. Ji, H. Shen, and M. West, "Bounded approximations for marginal likelihoods," 2010.

[4] P. Carbonetto and M. Stephens, "Scalable Variational Inference for Bayesian Variable Selection in Regression, and its Accuracy in Genetic Association Studies," *Bayesian Analysis*, vol. 7, pp. 73–108, 2012.

Algorithm 1: Variational Learning to Approximate Posterior Relevance of Brain Regions

Parameters: (a) prior for the regions $\boldsymbol{\alpha}$, (b) number of iterations: T

Data: (a) Diagnosis $\mathbf{y}$, (b) Imaging Data $\mathbf{X}$, (c) Genotype $\mathbf{G}$

Output: Parameters of the posterior distribution: $\boldsymbol{\theta}$

Approximate Bayes Factors ($\boldsymbol{\gamma}$)

  for $m \leftarrow 1$ ***to*** M **do**

    Set γ_m to an approximation of Eq. (1) (see Algorithm 2)

for $i \leftarrow 1$ ***to*** T **do**

  Draw a set from the current estimate of the posterior distribution: $\mathbf{b}^t \sim q_{\tilde{\boldsymbol{\theta}}^t}$;

  Approximate the marginal conditional likelihood $p(\mathbf{y}|\mathbf{b}^t; \mathbf{X}, \pi)$;

  Set $\hat{\mathbf{g}}^t = \tilde{\mathbf{b}}^t \left(\log p(\mathbf{y}|\mathbf{b}^t) + \boldsymbol{\gamma}^T \mathbf{b}^t + |\mathbf{b}^t| \log \alpha \right)$;

  Set $\hat{\mathbf{C}}^t = \tilde{\mathbf{b}}^t (\tilde{\mathbf{b}}^t)^T$;

  Set $\mathbf{g}^{t+1} = (1-w)\mathbf{g}^t + w\tilde{\mathbf{g}}^t$;

  Set $\mathbf{C}^{t+1} = (1-w)\mathbf{C}^t + w\tilde{\mathbf{C}}^t$;

  Solve $\mathbf{C}^{t+1}\tilde{\boldsymbol{\theta}}_{t+1} = \mathbf{g}^{t+1}$;

  if $t > N/2$ **then**

    Set $\bar{\mathbf{g}} = \bar{\mathbf{g}} + \hat{\mathbf{g}}^t$;

    Set $\bar{\mathbf{C}} = \bar{\mathbf{C}} + \hat{\mathbf{C}}^t$;

return *the solution of a linear system of equations* $\bar{\mathbf{C}}\boldsymbol{\theta} = \bar{\mathbf{g}}$;

Algorithm 2: Variational Learning to Approximate $p(\mathbf{x}|b=1;\mathbf{G})$

Data: (a) Imaging features for one region $\mathbf{x}$, (b) Genotype $\mathbf{G}$

Parameters: Set of hyper-parameters, $\pi'(1), \cdots, \pi'(L)$

Output

  (a) Approximate marginal likelihood $\hat{\gamma} \approx \log p(\mathbf{x}|b=1;\mathbf{G})$;

  (b) Variational estimate of posterior inclusion probability $\hat{\tau}_s := \mathbb{E}_q[a_s|\mathbf{x}, b=1;\mathbf{G}], \quad \forall 1 \leq s \leq S$;

  (c) Variational estimate of posterior variance $(\hat{\boldsymbol{\varsigma}} := \mathbb{E}_q[\boldsymbol{\omega}|\mathbf{x}, b=1;\mathbf{G}])$;

  (d) Variational estimate of posterior variance $(\hat{\boldsymbol{\nu}} := \mathbb{V}ar_q[\boldsymbol{\omega}|\mathbf{x}, b=1;\mathbf{G}])$;

for $i \leftarrow 1$ ***to*** L **do**

  Initialize $(\boldsymbol{\varsigma}_{\text{Init}}, \boldsymbol{\nu}_{\text{Init}}, \boldsymbol{\tau}_{\text{Init}})$ randomly ;

  $(\boldsymbol{\varsigma}(i), \boldsymbol{\nu}(i), \boldsymbol{\tau}(i), Z(i)) \leftarrow$ **Mean-Field**$(\mathbf{G}, \mathbf{x}, \pi'(i))$;

Set $(\boldsymbol{\varsigma}_{\text{Init}}, \boldsymbol{\nu}_{\text{Init}}, \boldsymbol{\tau}_{\text{Init}})$ to the parameters associated with highest Z ;

for $i \leftarrow 1$ ***to*** L **do**

  $(\boldsymbol{\varsigma}(i), \boldsymbol{\nu}(i), \boldsymbol{\tau}(i), Z(i)) \leftarrow$ **Mean-Field**$(\mathbf{G}, \mathbf{x}, \pi'(i))$;

  Compute importance weight $\zeta(i) \leftarrow Z(i)p(\pi'(i))/\tilde{p}(\pi'(i))$;

Normalize importance weights: $\zeta(i) \leftarrow \zeta(i)/(\sum_i \zeta(i))$;

Average Over Hyper-parameters

  $\hat{\boldsymbol{\nu}} \leftarrow \sum_{i=1}^{L} \zeta(i)\boldsymbol{\nu}(i)$;

  $\hat{\boldsymbol{\varsigma}} \leftarrow \sum_{i=1}^{L} \zeta(i)\boldsymbol{\varsigma}(i)$;

  $\hat{\boldsymbol{\tau}} \leftarrow \sum_{i=1}^{L} \zeta(i)\boldsymbol{\tau}(i)$;

  $\hat{\gamma} \leftarrow \sum_{i=1}^{L} \zeta(i) \left(\log Z(i)\right)$;

return $(\hat{\boldsymbol{\nu}}, \hat{\boldsymbol{\varsigma}}, \hat{\boldsymbol{\tau}}, \hat{\gamma})$;

Mean-Field Subroutine

  Input: (a) Genotype $(\mathbf{G})$, (b) Response $(\mathbf{y})$, (c) Hyper-parameters (π')

  $(\boldsymbol{\varsigma}, \boldsymbol{\nu}, \boldsymbol{\tau}) \leftarrow (\boldsymbol{\varsigma}_{\text{Init}}, \boldsymbol{\nu}_{\text{Init}}, \boldsymbol{\tau}_{\text{Init}})$;

  repeat

    Choose $s \in \{1, \cdots, S\}$;

    $\varsigma_s^2 \leftarrow \sigma_0^{-2} \left((\mathbf{G}^T\mathbf{G})_{ss} + 1/\sigma_\omega^2 \right)$;

    $\nu_s \leftarrow \varsigma_s^2 \sigma_0^{-2} \left((\mathbf{G}^T\mathbf{y})_s - \sum_{j \neq s} (\mathbf{G}^T\mathbf{G})_{js} \tau_j \nu_j \right)$;

    $\frac{\tau_s}{1-\tau_s} \leftarrow \frac{\alpha}{1-\alpha} \times \frac{\varsigma_s}{\sigma_0 \sigma_\omega} \times \exp(\frac{1}{2}\nu_s^2/\varsigma_s^2)$

  until *Convergence*;

  Set $\log Z$ to the approximate lower bound by Eq. (2) ;

  return $\boldsymbol{\varsigma}$, $\boldsymbol{\nu}$, $\boldsymbol{\tau}$, Z ;