Generative Method to Discover Genetically Driven Image Biomarkers

Abstract. We present a generative probabilistic approach to discovery of disease subtypes determined by the genetic code. In many diseases, multiple types of pathology may present simultaneously in a patient, making quantification of the disease challenging. Our method seeks common co-occurring image and genetic patterns in a population as a way to model these two different data types jointly. We assume that each patient is a mixture of multiple disease subtypes and use the joint generative model of image and genetic markers to identify disease subtypes guided by known genetic influences. Our model is based on a variant of the so-called topic models that uncover the latent structure in a collection of data. We derive an efficient variational inference algorithm to extract patterns of co-occurrence and to quantify the presence of a heterogeneous disease in each patient. We evaluate the method on simulated data and illustrate its use in the context of Chronic Obstructive Pulmonary Disease (COPD) to characterize the relationship between image and genetic signatures of the COPD subtypes in a large patient cohort.

1 Introduction

We propose and demonstrate a joint model of image and genetic variations associated with a disease. Our goal is to identify disease-specific image biomarkers that are also correlated with side information, such as the genetic code or other biologically relevant indicators. Our approach targets diseases that can be thought of as a superposition of different processes, or subtypes, that are subject to genetic influences and are often present simultaneously in the same patient. Our motivation comes from a study of the Chronic Obstructive Pulmonary Disease (COPD), but the resulting model is applicable to a wide range of heterogeneous disorders.

COPD is a lung disease characterized by chronic and progressive difficulty in breathing; it is one of the leading causes of death in the United States [14]. COPD is often associated with emphysema, i.e., the destruction of lung air sacs, and an airway disease, which is caused by inflammation of the airways. In this paper, we focus on modeling emphysema based on lung CT images. Emphysema exhibits many subtypes. It is common for several subtypes to co-occur in the same lung [16]. Genetic factors play an important role in COPD [14], and it is believed that subtypes of COPD are driven by genetics [7]. We therefore aim to quantify the lung tissue heterogeneity that is associated with the genetic variations in the patient cohort.

CT imaging is used to measure the extent of COPD, and particularly of emphysema. The standard approach to quantifying emphysema is to use the volume

of sub-threshold intensities in the lung as a surrogate measure for the volume of emphysema [8]. More recently, histograms [13], texture descriptors [18], and combination of both [19] have been proposed to classify subtypes of emphysema based on training sets of CT patches labeled by clinical experts. While both the histogram and intensity features have been shown to be important for emphysema characterization, the clinical definitions of disease subtypes are based on visual assessment of CT images by clinicians and are not necessarily genetically driven. In prior studies, association between image and genetic variants was established as a separate stage of analysis and was not taken into account when extracting relevant biomarkers from images.

Most methodological innovations in joint analysis of imaging and genetics have used image data as an intermediate phenotype to enhance the discovery of relevant genetic markers in the context of neuro-degenerative diseases [5,9,21]. In the context of COPD, Castaldi*et al.* [7] used local histograms to measure distinct emphysema patters and performed genome-wide association study (GWAS) to validate their results. In contrast to prior research in imaging genetics, we use the results of genetic analysis to help us characterize image patterns associated with the disease, in effect reversing the direction of analysis for disorders with high anatomical heterogeneity and available information on genetic influences. We model imaging and genetic variations jointly, and demonstrate efficient inference of co-occurrence pattern, as indicated by our results.

In this paper, we assume that a few important genetic markers associated with the disease are available. We build a generative model that captures the commonly occurring image and genetic patterns in a population. Each subject is modeled as a sample from the population-wide collection of patterns. This abstraction of the image and genetic patterns at the population level reveals the associations between image-based and genetic subtypes and uses genetic information to guide the definition of image biomarkers for distinct disease subtypes. Our method is based on a non-parametric topic modeling [20], originally developed in machine learning for characterizing structure of documents. We build an analogy between topics contributing words to a document and disease subtypes contributing image textures and minor alleles to a patient. The closest work to our approach is by Batmanghelich *et al.* [4] who developed a topic model for global histograms of the lung intensity values. The model did not include local texture features; genotype data was not considered as part of the model. In contrast, our topic model builds on rich texture descriptors and integrates image and genetic information into a single framework. Our approach can be readily extended to include other clinical or demographic data.

We evaluate the method on a synthetic data set that matches our clinical assumptions, demonstrating substantial benefits of using a hierarchical population model to capture common patterns of heterogeneity in the image phenotypes and in the genetic code. We also show that the genetic data as side information boosts the performance of the method compared to the baselines and a variant of our model without the genetic data. Finally, we illustrate an application of

Model Variables	
I_{sn}	image descriptor of supervoxel n in subject s
G_{sm}	genetic location of minor allele m in subject s
z^I_{sn}	subject-specific topic that generates super-voxel n in subject s $1 \leq z^I_{s,n} \leq T$
z^G_{sm}	subject-specific topic that generates minor allele m in subject s $1 \leq z^G_{s,m} \leq T$
c_{st}	population-level topic that serves as subject-specific topic t in subject s, $1 \leq c_{st} \leq K$
v	parameter vector that determines the stick-breaking proportions of topics in a population template
π_s	parameter vector that determines the stick-breaking proportions of topics in subject s
(μ_k, Σ_k)	mean and variance of image descriptors for population-level topic k
β_k	frequency of different locations in genetic signatures for population-level topic k
ω	hyper-parameters of the Beta prior for v
α	hyper-parameters for the Beta prior for π_s
η^I	hyper-parameters of the Normal-Inverse-Wishart prior for (μ_k, Σ_k)
η^G	hyper-parameters of the Dirichlet prior prior for β_k
VB Estimates	
$(\hat{\mu}_k, \hat{\Sigma}_k)$	mean and variance of image descriptors for population-level topic k
$\hat{\beta}_k$	frequency of different locations in genetic signatures for population-level topic k

Table 1: Model variables and Variational Bayes (VB) estimates used in the paper.

our method to a study of COPD and identify common emphysema subtypes associated with genetic factors implicated in COPD.

2 Model

In this section, we describe the generative model for image and genetic data based on a population-wide collection of common patterns that are instantiated in each subject. Our notation is summarized in Table 1 and the generative process is illustrated in Fig. 1.

Image and Genetic Data. We assume each subject in a study is characterized by an image and a genetic signature for the loci in the genome *previously* implicated in the disease. Based on the analogy to the "bag-of-words" representation [17], we assume that an image domain is divided for each subject into relatively homogeneous spatially contiguous regions (i.e., "supervoxels"). We let $I_{sn} \in \mathbb{R}^D$ denote the D-dimensional descriptor of supervoxel n in subject s that summarizes the intensity and texture properties of the supervoxel. The genetic data in our problem comes in a form of minor allele counts (0, 1 or 2) for a set of L loci. Our representation for genetic data is inspired by the commonly used additive model in GWAS analysis [6]. In particular, we assume that the risk of the disease increases monotonically by the minor allele count. We let $G_{sm} \in \{1, \cdots, L\}$ denote minor allele m in genetic signature of subject s. For example, suppose $L = 2$, and subject s has one and two minor alleles in locations ℓ_1 and ℓ_2 respectively. This subject is represented by a list of 3 elements $G_s = \{\ell_1, \ell_2, \ell_2\}$.

Population Model. Our population model is based on the Hierarchical Dirichlet Process (HDP) [20]. The model assumes a collection of K "topics" that are shared across subjects in the population. We let p_k^I and p_k^G denote the distributions for the image and genetic signatures, respectively, associated with topic k. Each $p_k^I = \mathcal{N}(\mu_k, \Sigma_k)$ is a Gaussian distribution that generates supervoxel descriptors I_{sn}; it is parameterized by its mean vector $\mu_k \in \mathbb{R}^D$ and covariance matrix $\Sigma_k \in \mathbb{R}^D \times \mathbb{R}^D$. Each $p_k^G = \text{Cat}(\beta_k)$ is a categorical distribution that generates minor allele locations G_{sm}; it is parameterized by its weight vector $\beta_k \in (0,1)^L$.

When sampling a new subject s, at most $T < K$ topics are drawn from the population-wide pool to determine the image and genetic signature of this subject. We let c_{st} denote the population topic selected to serve as subject-specific topic t ($1 \leq t \leq T$) in subject s. We also use $c_s = [c_{s1}, \ldots, c_{sT}]$ to refer to the entire vector of topics selected for subject s. $c_s[t] = k$ indicates that population-level topic k was selected to serve as subject-specific topic t. The subject-specific topics inherit their signature distributions from their population prototypes, but each subject is characterized by a different subset and proportions of the population-level topics represented in the subject-specific data.

As $T, K \to \infty$, this model converges to a non-parametric Hierarchical Dirichlet Process (HDP) [20]. Rather than choose specific values for T and K, HDP enables us to estimate them from the data. As part of this model, we employ the "stick-breaking" construction [20] to parameterize the categorical distribution for c_{st}:

$$c_{st} \sim \text{Cat-SB}(v), \tag{1}$$

where Cat-SB(v) is a categorical distribution whose weights are generated through the stick-breaking process from the (potentially infinite) parameter vector v whose components are in the interval $(0,1)$. Formally, if we define a random variable $x \sim \text{Cat-SB}(v)$, then

$$p(x) \triangleq v_x \prod_{i=1}^{x-1} (1 - v_i) \quad \text{for } x = 1, \ldots. \tag{2}$$

This parameterization accepts infinite alphabets. The stick-breaking construction penalizes the high numbers of topics hence encouraging parsimonious representation of data. A similar construction enables an automatic selection of the number of topics at the population level and at the subject level. We employ a truncated HDP variant that uses finite values for T and K [11]. In this setup, $v \in (0,1)^{K-1}$. In contrast to finite (fixed) models, we set K to high enough value, and the estimation procedure uses as many topics as needed but not necessarily all K topics to explain the observations.

Subject-Specific Data. To generate an image descriptor for supervoxel n in subject s, we sample random variable $z_{sn}^I \sim \text{Cat-SB}(\pi_s)$ from a categorical distribution parameterized by the vector of stick-breaking proportions $\pi_s \in (0,1)^{T-1}$. $z_{sn}^I = t$ indicates that the subject-specific topic t generates im-

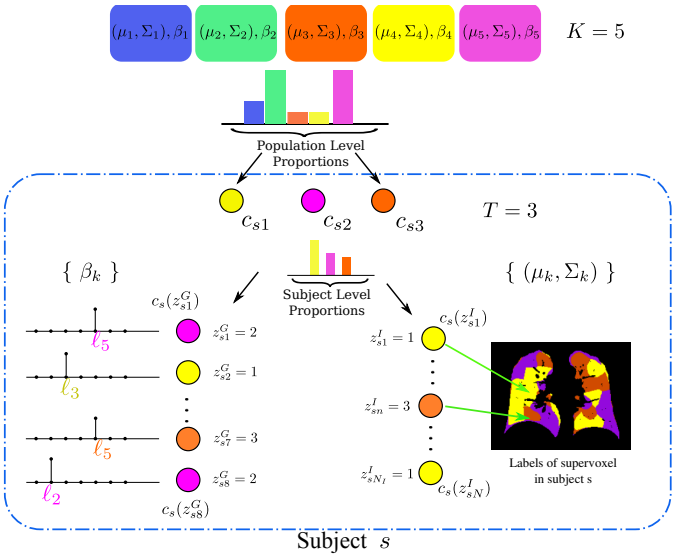


Fig. 1: The subject s draws a subset of T topics from K population-level topics. Indices of the subject-level topics are stored in $c_{s1},..,c_{sT}$ drawn from a categorical distribution. At the subject level, indices of the supervoxels $\{z^I_{sn}\}$ and location of minor alleles $\{z^G_{s,m}\}$ are drawn from subject-specific categorical distribution. Vector c_s acts as a map from subject-specific topics to the population-level topics (i.e., $c_s(z^G_{sm})$ or $c_s(z^I_{sn})$).

age descriptor I_{sn}:

$$I_{sn}|z^I_{sn},c_s \sim \mathcal{N}\left(\mu_{c_s[z^I_{sn}]},\Sigma_{c_s[z^I_{sn}]}\right). \tag{3}$$

Similarly, to generate minor allele location m in subject s, we sample random variable $z^G_{sm} \sim$ Cat-SB(π_s) and draw G_{sm} from the corresponding genetic signature of subject-specific topic z^G_{sm}:

$$G_{sm}|z^G_{sm},c_s \sim \text{Cat}\left(\beta_{c_s[z^G_{sm}]}\right). \tag{4}$$

Priors. Following the Bayesian approach, we define priors for the remaining latent variables, namely $\{v_k,\pi_{st}\}$ and the parameters of the likelihood distributions $\{\mu_k,\Sigma_k,\beta_k\}$. For the computational reasons, we choose the priors from the exponential family. Specifically, we use the Beta distribution as the prior of the parameter vectors v and π_s that determine the stick-breaking proportions at the population-wide and subject-specific levels, respectively:

$$v_k \sim \text{Beta}(1,\omega), \quad k=1,\dots,K-1, \tag{5}$$

$$\pi_{st} \sim \text{Beta}(1,\alpha), \quad t=1,\dots,T-1, \tag{6}$$

where $\omega>0$ and $\alpha>0$ are the corresponding shape parameters of the Beta distribution. For computational reasons, we also assume priors for image and genetic signature parameters that are conjugate for the corresponding likelihood distributions (3) and (4):

$$\mu_k,\Sigma_k \sim \text{NIW}(\eta^I) \qquad \text{and} \qquad \beta_k \sim \text{Dir}(\eta^G),$$

where NIW(η) is the Normal-Inverse-Wishart distribution with parameters η and Dir(η) is the Dirichlet distributions with parameters η.

3 Inference

Given a study of S subjects with their respective image descriptors $\{I_{sn}\}$ and genetic signatures $\{G_{sm}\}$, we seek posterior distributions of the model parameters. Since exact computation of the posterior quantities is computationally intractable, we resort to an approximation. Due to the size of data and its dimensionality, sampling is computationally impractical. We therefore derive a Variational Bayes (VB) approximation [11]. For notational convenience, we define $\mathcal{D} = \{I_{sn}, G_{sm}\}_{s=1}^{S}$ to be all image and genetic data, $\mathcal{S} = \{z_{sn}^{I}, z_{s,m}^{G}, c_s, \pi_s\}_{s=1}^{S}$ be all subject-specific latent variables, and $\mathcal{P} = \{\mu_k, \Sigma_k, \beta_k, v_k\}_{k=1}^{K}$ be all population-based latent variables. We omit fixed hyper-parameters to simplify the notation. Variational Bayes inference selects an approximating distribution $q(\mathcal{S}, \mathcal{P})$ for the true posterior distribution $p(\mathcal{S}, \mathcal{P}|\mathcal{D})$ by minimizing the cost functional

$$F(q) = \mathbb{E}_q \left[\ln p(\mathcal{D}, \mathcal{S}, \mathcal{P})\right] - \mathbb{E}_q \left[\ln q(\mathcal{S}, \mathcal{P})\right] \tag{7}$$

that can be thought of as a KL divergence between the approximating distribution and the true posterior distribution. Additional details and the update rules of the iterative inference algorithm can be found in the Appendix.

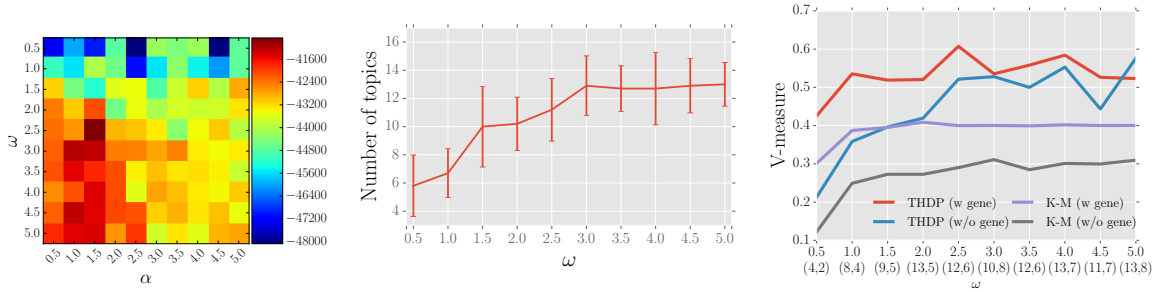
We use the parameters of the approximating distribution $q(\mathcal{S}, \mathcal{P})$ to construct estimates of the relevant model parameters. Specifically, we seek the estimates $(\hat{\mu}_k, \hat{\Sigma}_k)$ of the image descriptors and the estimates $\hat{\beta}_k$ of the associated genetic signatures for each population-level topic k. Moreover, for each subject s we estimate a distribution over the population topics for each supervoxel to visualize the spatial distributions of disease subtypes for clinical assessment.

4 Experiments

In this section, we demonstrate and evaluate the algorithm on simulated and real data. We use simulated data to study the advantages offered by the hierarchical model and investigate the effects of the side information (genetic data in our case) on the accuracy of recovering the latent topics. We also investigate the behavior of the model with respect to the hyper-parameters. Moreover, we illustrate the method on a subset of a large-scale study of lung based in CT images of COPD patients. In this experiment, we characterize co-occurring image and genetic patterns in the data.

4.1 Simulation

To evaluate the performance of the method, we sampled the data from the proposed hierarchical model. In particular, we generated image and genetic signatures for $S = 100$ subjects from 20 population-level topics while limiting the number of subject-specific topics to 5. We used Beta$(1, 8)$ and Beta$(1, 1)$ for



(a) (b) (c)

Fig. 2: Simulated data results. (a) Variational lower bound $F(q^*)$ for different values of (α, ω). (b) The number of topics discovered by the model as a function of ω averaged over α. (c) Normalized mutual information between the true and the discovered topics for our method and for k-means clustering (`K-M`) applied to pooled data. The number of discovered topics is reported in brackets under the corresponding value of ω (`w gene`,`w/o gene`).

population-level and subject-specific stick-breaking proportions that govern the relative frequencies of the topics, leading to more spread out weights at the population level than those at the subject level. We draw the imaging signature parameters for population topics from a 2-dimensional NIW distribution with a zero mean vector, identity covariance matrix, and the shape and scale parameters set to 5 and 0.5. The subject-specific image signatures ($N = 75$ for all s) are drawn from Gaussian distributions whose parameters are determined by the corresponding image parameters of the population topic. The weights of the genetic signature for each population-level topic are drawn from a Dirichlet distribution with all parameters set to one. The subject-specific genetic signatures ($M = 65$ for all s) are drawn from a categorical distributions determined by the weights of the corresponding genetic signature of the topic model.

Hyper-parameters ω and α control the model size, i.e., the number of topics at the population level and the subject level respectively. Of the two, the population-level parameter ω has a broader influence on how well the model explains subject-specific data. We sweep a range $(0.5, 5.0)$ for both parameters. Fig.2a reports the value of the lower bound $F(q^*)$ for each pair of the parameter setting which we use for model selection. We observe that the algorithm's performance depends smoothly on the parameter values. In subsequent experiments, we set α to the optimal values based on $F(q^*)$ and study the behavior of the model for a range of values of ω. Fig.2b reports the number of population topics estimated by the model as a function of ω. Not surprisingly, the model size grows with ω, but is quite stable for a wide range of values of ω.

To evaluate the effects of the hierarchical model and of joint modeling of image and genetic information, we compare our approach (with/out genetic data) to a k-means algorithm applied to the pooled data from all subjects. We apply the baseline k-means clustering to image data only, and also to the data set of

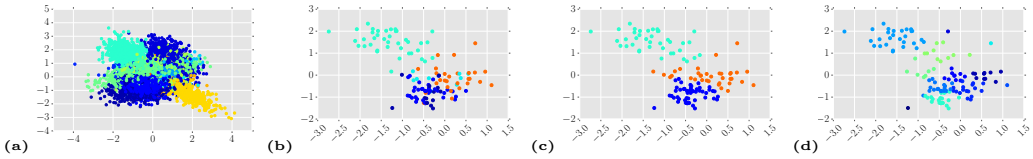
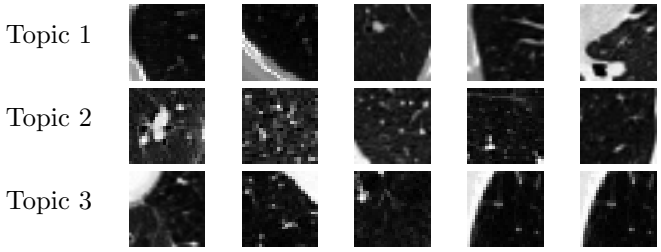


Fig. 3: Example simulated image data using 2D features. (a) Features from all subjects pooled into one set. Colors correspond to true topics, unavailable to the algorithm. (b) Image features for a single subject in a set. (c) Topics recovered by our algorithm (with genetic data) for the same subject based on the whole data set. (d) Topics recovered by k-means clustering applied to the pooled data in (a).

image signatures of all supervoxels concatenated with the entire genetic signature of the same subject. Fig.2c compares our method with the two k-means variants using a standard measure of normalized mutual information (v-measure) [15] between the true and discovered topics. The measure varies between 0 and 1; 1 corresponds to the perfect match. While adding genetic information to image features boosts the performance of our method and clustering on pooled data, our hierarchical model outperforms both baseline methods substantially for a wide range of values of ω. The difference between two variants of our method (with/out genetic) illustrates the value of the side information to improve the performance. Fig.3 illustrates this point on an example from our simulations for one setting of the parameters.



SNP	$\hat{\beta}_k$
rs11	0.0854
rs19	0.0802
rs13	0.0774
rs10	0.0763
rs12	0.0736
rs20	0.0735

SNP	$\hat{\beta}_k$
rs11	0.0869
rs19	0.0797
rs10	0.0771
rs13	0.0741
rs12	0.0731
rs15	0.0710

SNP	$\hat{\beta}_k$
rs11	0.0845
rs10	0.0774
rs19	0.0757
rs12	0.0731
rs13	0.0699
rs20	0.0697

Fig. 4: Representative patches of three topics (each row) randomly drawn from patients with different severity of emphysema. The Topic 1 is the most frequent topic across all subjects presenting along the edges of the lung. The tables below report the top six SNPs for each topic with their estimated relative weights. We observe that the genetic signatures vary across topics.

4.2 COPD Study

We apply the method to CT images of lung in 1790 subjects from a large-scale COPD study. After automatic segmentation of the lung [2], we compute intensity histograms and Rotation-Invariant Histogram of Gradient [12] for every voxel. This dense 28-dimensional descriptor quantifies the local texture of every voxel and is rotationally invariant. We chose to work with this particular texture descriptor based on our previous experience with classification of lung image patches, where combining intensity histogram and texture descriptors substantially improved classification accuracy, suggesting that the texture features offer significant complementary information for sub-typing[1]. We employ a modified version of super-voxelization method [1] to subdivide the lungs into coherent, spatially contiguous regions. For each supervoxel, the intensity histogram is represented with 10 bins within the CT intensity range relevant to emphysema. Finally, we concatenate the intensity and the texture descriptors and pool all features vectors of all subjects and apply Principal Component Analysis to reduce the dimensionality to 15, i.e., $I_{sn} \in \mathbb{R}^{15}$. Moreover, we compiled a list of SNPs that are highly correlated with the lung function based on GWAS for different clinical phenotypes, such as respirometry measurements [3]. The list is compiled based a relatively stringent p-value[2]. Based on our experience with the simulated data and the expected number of disease subtypes, we set $K = 20$ and $T = 10$. Furthermore, we set $\alpha = 1$, $\omega = 5$ and set uninformative priors for the image and genetic signature parameters.

The method summarizes the population into seven population-level topics. The number of topics per patient varies from one to four. Fig.4 visualizes the top three topics in the CT space of three different subjects in the study, together with top six minor alleles in the genetic signature of each topic. We observe that the genetic signatures (relative weights or rankings) vary across topics, suggesting variable genetic patterns that give rise to different image properties. To visualize the spatial distribution of the topics, we first identify a subject with the maximum number of topics detected. We then compute the membership value of the supervoxel in the population-level topics (i.e., $\sum_t \phi^I_{sn}(t)\xi_{st}$). Fig.5 illustrates such spatial maps for the three topics reported in Fig.4. We observe that the spatial pattern varies across topics, with localized topic 1 and widely diffuse topic 2.

5 Discussion and Conclusions

We proposed and demonstrated a generative model based on the truncated Hierarchical Dirichlet Process to identify common image and genetic patterns in a population (i.e., population-level topics). The underlying assumption of our model is that every subject is a superposition of few topics. Our main contribution is to model side information jointly with imaging data. We demonstrated

[1] While histogram and texture alone yield around 45% and 46% 10-fold cross-validation accuracy rates for six classes, the combination produced 61% accuracy confirming that the texture feature we used has a significant additional information for sub-typing

[2] For the anonymous review stage, we masked the exact SNPs locations.

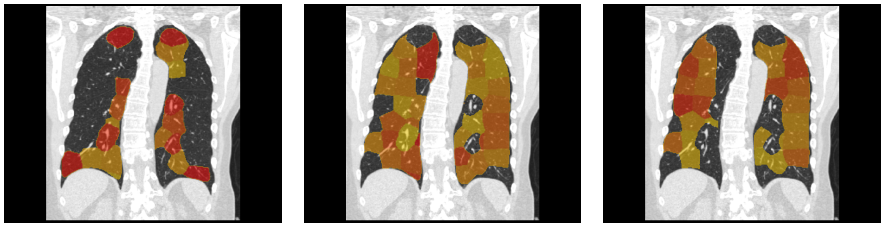


Fig. 5: Spatial distribution of the top three topics from Fig.4 in one subject. The color indicates the posterior probability. The warmer the color, the higher the probability.

the method on synthesized data and reported preliminary results for the COPD study.

Once population-wide template of image and genetic variability has been constructed, it enables us to answer many interesting questions about the heterogeneity of the disease in the population and in individual subjects. In particular, investigating the variability of topic representation in different subjects and using subject-specific topic proportions, promise to provide a handle on how the disease varies in a population and suggest numerous interesting directions for future work.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: Slic superpixels compared to state-of-the-art superpixel methods. Pattern Analysis and Machine Intelligence, IEEE Transactions on 34(11), 2274–2282 (2012)
2. Authors, A.: Anonymous title. Anonymous Journal X(NN), pp–pp (20YY)
3. Authors, A.: Anonymous title of the genetic study. Anonymous Journal X(NN), pp–pp (20YY)
4. Batmanghelich, K.N., Cho, M., San Jose, R., Golland, P.: Spherical topic models for imaging phenotype discovery in genetic studies. In: Bayesian and grAphical Models for Biomedical Imaging, pp. 107–117. Springer (2014)
5. Batmanghelich, N.K., Dalca, A.V., Sabuncu, M.R., Golland, P.: Joint modeling of imaging and genetics. In: Information Processing in Medical Imaging. pp. 766–777. Springer (2013)
6. Bush, W.S., Moore, J.H.: Genome-wide association studies. PLoS computational biology 8(12), e1002822 (2012)
7. Castaldi, P.J., et al: Genome-wide association identifies regulatory loci associated with distinct local histogram emphysema patterns. American journal of respiratory and critical care medicine 190(4), 399–409 (2014)
8. Castaldi, P.J., San José Estépar, R., Mendoza, C.S., Hersh, C.P., Laird, N., Crapo, J.D., Lynch, D.A., Silverman, E.K., Washko, G.R.: Distinct quantitative computed tomography emphysema patterns are associated with physiology and function in smokers. American journal of respiratory and critical care medicine 188(9), 1083–1090 (2013)
9. Filippini, N., Rao, A., Wetten, S., Gibson, R.A., Borrie, M., Guzman, D., Kertesz, A., Loy-English, I., Williams, J., Nichols, T., et al.: Anatomically-distinct genetic

associations of apoe e4 allele load with regional cortical atrophy in alzheimer's disease. Neuroimage 44(3), 724–728 (2009)
10. Guan, Y., Dy, J.G., Niu, D., Ghahramani, Z.: Variational inference for nonparametric multiple clustering. In: MultiClust Workshop, KDD-2010 (2010)
11. Hoffman, M.D., Blei, D.M., Wang, C., Paisley, J.: Stochastic variational inference. The Journal of Machine Learning Research 14(1), 1303–1347 (2013)
12. Liu, K., Skibbe, H., Schmidt, T., Blein, T., Palme, K., Brox, T., Ronneberger, O.: Rotation-invariant hog descriptors using fourier analysis in polar and spherical coordinates. International Journal of Computer Vision 106(3), 342–364 (2014)
13. Mendoza, C.S., Washko, G.R., Ross, J.C., Diaz, A., Lynch, D.A., Crapo, J.D., Silverman, E.K., Acha, B., Serrano, C., Estepar, R.S.J.: Emphysema quantification in a multi-scanner hrct cohort using local intensity distributions. In: Biomedical Imaging (ISBI), 2012 9th IEEE International Symposium on. pp. 474–477. IEEE (2012)
14. Regan, E.A., Hokanson, J.E., Murphy, J.R., Make, B., Lynch, D.A., Beaty, T.H., Curran-Everett, D., Silverman, E.K., Crapo, J.D.: Genetic epidemiology of copd (copdgene) study design. COPD: Journal of Chronic Obstructive Pulmonary Disease 7(1), 32–43 (2011)
15. Rosenberg, A., Hirschberg, J.: V-measure: A conditional entropy-based external cluster evaluation measure. In: EMNLP-CoNLL. vol. 7, pp. 410–420. Citeseer (2007)
16. Satoh, K., Kobayashi, T., Misao, T., Hitani, Y., Yamamoto, Y., Nishiyama, Y., Ohkawa, M.: Ct assessment of subtypes of pulmonary emphysema in smokers. CHEST Journal 120(3), 725–729 (2001)
17. Sivic, J., Zisserman, A.: Efficient visual search of videos cast as text retrieval. Pattern Analysis and Machine Intelligence, IEEE Transactions on 31(4), 591–606 (2009)
18. Song, Y., Cai, W., Zhou, Y., Feng, D.D.: Feature-based image patch approximation for lung tissue classification. IEEE Trans. Med. Imaging 32(4), 797–808 (2013)
19. Sorensen, L., Shaker, S.B., De Bruijne, M.: Quantitative analysis of pulmonary emphysema using local binary patterns. Medical Imaging, IEEE Transactions on 29(2), 559–569 (2010)
20. Teh, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M.: Hierarchical dirichlet processes. Journal of the american statistical association 101(476) (2006)
21. Vounou, M., Nichols, T.E., Montana, G.: Discovering genetic associations with high-dimensional neuroimaging phenotypes: a sparse reduced-rank regression approach. Neuroimage 53(3), 1147–1159 (2010)

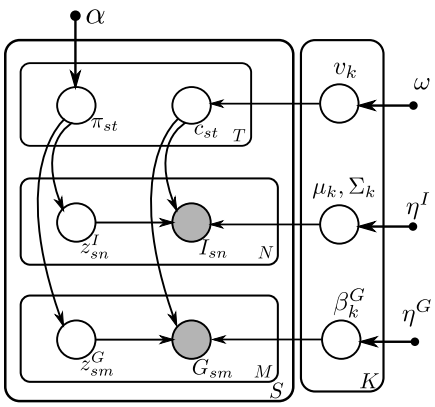
Appendix: Variational Bayes Inference Procedure

Combining all components of the model defined in Section 2, we construct the joint distribution of all variables in the model:

$$p(\mathcal{D},\mathcal{S},\mathcal{P}) = \underbrace{\prod_{k=1}^{K} p(\mu_k,\Sigma_k;\eta^I)\, p(\beta_k;\eta^G)\, p(v_k;\omega)}_{\text{population-level topics}} \times$$

$$\prod_{s=1}^{S}\prod_{t=1}^{T} \underbrace{p(c_{st}|v_k)p(\pi_{st};\alpha)}_{\text{topics for subject } s} \prod_{n=1}^{N} \underbrace{p(z^I_{sn}|\pi_{st})}_{\text{image topic}} \underbrace{p(I_{sn}|z^I_{sn},c_{st},\{\mu_k,\Sigma_k\})}_{\text{image likelihood}} \prod_{m=1}^{M} \underbrace{p(z^G_{sm}|\pi_{st})}_{\text{genetic topic}} \underbrace{p(G_{sm}|z^G_{sm},c_{st},\beta_k)}_{\text{genetic likelihood}},$$

where N and M are the number of supervoxels and minor alleles, respectively, identified for subject s.



Algorithm 1 Variational Bayes update rules.

1: $\tilde{\pi}^{(1)}_{st} = 1 + \sum_{n=1}^{N} \phi^I_{sn}(t) + \sum_{m=1}^{M} \phi^G_{sm}(t)$

2: $\tilde{\pi}^{(2)}_{st} = \alpha + \sum_{n=1}^{N} \sum_{j=t+1}^{T} \phi^I_{sn}(j) + \sum_{m=1}^{M} \sum_{j=t+1}^{T} \phi^G_{sm}(j)$

3: $\xi^k_{st} \propto \exp\{\mathbb{E}[\log SB_k(V)] + \sum_{m=1}^{M} \phi^G_{sm}(t)\mathbb{E}[\log \beta^G_{k,G_{sm}}] + \sum_{n=1}^{N} \phi^I_{sn}(t)\mathbb{E}[\log \mathbb{P}(I_{sn}|\{\mu_k, \Sigma_k\}, c_{st}, z^I_{sn})]\}$

4: $\phi^I_{sn}(t) \propto \sum_{k=1}^{K} \xi^k_{st}\mathbb{E}[\log \mathbb{P}(I_{sn}|\{\mu_k, \Sigma_k\}, c_{st}, z^I_{sn})] + \exp\{\mathbb{E}[\log SB_t(\pi_s)]\}$

5: $\phi^G_{sn}(t) \propto \sum_{k=1}^{K} \xi^k_{si}\mathbb{E}[\log \beta^G_{k,G_{sn}}] + \exp\{\mathbb{E}[\log SB_t(\pi_s)]\}$

6: Update $\tilde{\eta}^I_k$ based on NIW update equations (see [10])

7: $\tilde{\eta}^G_{kv} = \eta^G + \sum_{s=1}^{S} \sum_{i=T}^{T} \xi^k_{st} \sum_{n=1}^{M} \phi^G_{sn}(t) G_{sn}$

8: $\tilde{\omega}^{(1)}_k = 1 + \sum_{s=1}^{S} \sum_{i=1}^{T} \xi^k_{si}$

9: $\tilde{\omega}^{(2)}_k = \omega + \sum_{s=1}^{S} \sum_{i=1}^{T} \sum_{\ell=k+1}^{K} \xi^\ell_{si}$

Fig. 6: Left: Graphical model that represents the joint distribution. The open gray and white circles correspond to the observed and the latent random variables, respectively. The full circles represent fixed hyper-parameters. Superscript `I` and `G` denote image and genetic parts of the model respectively. Right: Update rules for the variational parameters.

We choose a factorization for the distribution q that captures most model assumptions and yet is computationally tractable:

$$q(\mathcal{S}, \mathcal{P}) = \underbrace{\prod_{k=1}^{K} \mathrm{NIW}(\mu_k, \Sigma_k; \tilde{\eta}^I_k)\, \mathrm{Dir}(\beta_k; \tilde{\eta}^G_k)\, \mathrm{Beta}(v_k; \tilde{\omega}_k)}_{\text{population-level topics}} \times$$

$$\prod_{s=1}^{S} \prod_{t=1}^{T} \underbrace{\mathrm{Cat}(c_{st}; \xi_{st})\, \mathrm{Beta}(\pi_{st}; \tilde{\alpha}_{st})}_{\text{topics for subject } s} \prod_{n=1}^{N} \underbrace{\mathrm{Cat}(z^I_{sn}; \phi^I_{sn})}_{\text{image topic}} \prod_{m=1}^{M} \underbrace{\mathrm{Cat}(z^G_{sm}; \phi^G_{sm})}_{\text{genetic topic}},$$

where we choose an appropriate approximating distribution for each latent variable and use ˜ to denote parameters of the approximating distributions. The optimization is defined in the space of the variational parameters $\{\tilde{\eta}^I, \tilde{\eta}^G, \tilde{\omega}, \xi, \tilde{\alpha}, \phi^I, \phi^G\}$. We omit the derivation of the updates due to space constraints; Algorithm 1 provides pseudocode for the resulting updates. We run the algorithm five times starting from different random initializations and report the result with the highest lower bound $F(q)$.

Once the algorithm converges, we estimate the population-level quantities of interest as means of the corresponding approximating distributions:

$$\hat{\mu}_k = \mathbb{E}\left[\mu_k|\mathcal{D}\right] \approx \mathbb{E}_q\left[\mu_k; \tilde{\eta}^I_k\right], \quad \hat{\Sigma}_k = \mathbb{E}\left[\Sigma_k|\mathcal{D}\right] \approx \mathbb{E}_q\left[\Sigma^I_k; \tilde{\eta}^I_k\right], \quad \hat{\beta}_k = \mathbb{E}\left[\beta_k|\mathcal{D}\right] \approx \mathbb{E}_q\left[\beta^G_k; \tilde{\eta}^G_k\right].$$

Each expectation above can be easily evaluated from the parameters of the corresponding distribution. In addition, we construct spatial maps that display the posterior probability of each population topic for each supervoxel in a particular subject s to visually evaluate the disease structure in that subject.