Dual Beta Process Priors for Latent Cluster Discovery in Chronic Obstructive Pulmonary Disease

James C. Ross
Brigham and Women's Hospital, Harvard Medical School
jross@bwh.harvard.edu

Peter J. Castaldi
Brigham and Women's Hospital, Harvard Medical School
repjc@channing.harvard.edu

Michael H. Cho
Brigham and Women's Hospital, Harvard Medical School
remhc@channing.harvard.edu

Jennifer G. Dy
Department of Electrical and Computer Engineering, Northeastern University
jdy@ece.neu.edu

ABSTRACT

Chronic obstructive pulmonary disease (COPD) is a lung disease characterized by airflow limitation usually associated with an inflammatory response to noxious particles, such as cigarette smoke. COPD is currently the third leading cause of death in the United States and is the only leading cause of death that is increasing in prevalence [15]. It also represents an enormous financial burden to society, costing tens of billions of dollars annually in the U.S. It is widely accepted by the medical community that COPD is a heterogeneous disease, with substantial evidence indicating that genetic variation contributes to varying levels of disease susceptibility. This heterogeneity makes it difficult to predict health decline and develop targeted treatments for better patient care. Although researchers have made several attempts to discover disease subtypes, results have been inconclusive, in part because standard clustering methods have not properly dealt with disease manifestations that may worsen with increased exposure. In this paper we introduce a transformative way of looking at the COPD subtyping task. Specifically, we model the relationship between risk factors (such as age and smoke exposure) and manifestations of disease severity using Gaussian Processes, which allow us to represent so-called "disease trajectories". We also posit that individuals can be associated with multiple disease types (latent clusters), which we assume are influenced by genetics. Furthermore, we predict that only subsets of the numerous disease-related quantitative features are useful for describing each latent subtype. We model these associations using two separate beta process priors, and we describe a variational inference approach to discover the most probable latent cluster assignments. Results are validated with associations to genetic markers.

KDD'14, August 24–27, 2014, New York, NY, USA.

ACM 978-1-4503-2956-9/14/08 ...$15.00.
http://dx.doi.org/10.1145/2623330.2623750.

Categories and Subject Descriptors

G.3 [**Mathematics of Computing**]: Probability and Statistics—*Correlation and regression analysis, stochastic processes, nonparametric statistics*

Keywords

Gaussian processes; beta processes; Bayesian nonparametrics; latent clusters; disease trajectories

1. INTRODUCTION

COPD is a major cause of chronic morbidity and mortality throughout the world, being the fourth leading cause of death worldwide [7]. By 2030 it is estimated that approximately 9 million people will die annually from the disease in the U.S. [15, 14]. COPD also imposes an enormous financial burden on society, with an estimated cost of $2.1 trillion in 2010 globally.

COPD is characterized by airflow limitation resulting from chronic inflammatory responses in the airways and lungs to noxious particles or gases. There are two basic components of COPD: emphysema (destruction and loss of lung tissue) and small airways disease (inflammation and thickening of airway walls). Both of these processes result in breathing difficulty. Patients often undergo high resolution computed tomography (CT) scanning, which enables the direct evaluation of the lungs and airways (Figure 1). Additionally, COPD severity is assessed in the clinic using spirometry, a technique in which patients blow into a device that measures lung function.

Tobacco smoke is the most common environmental risk factor of COPD, but it is known to be a heterogeneous disease, with genetic factors predisposing individuals to varying levels of disease severity as a function of exposure. An improved understanding of the interplay between genetics and exposure should lead to better stratification of patients for prognosis and personalization of therapies. There have been several large clinical projects to better understand this complicated disease, one of the largest being COPDGene [17] (http://www.copdgene.org). A number of authors have also applied established machine learning approachs in an effort to identify distinct disease subtypes. [2] applied principal component analysis (PCA) followed by cluster analysis using the VARCLUS procedure on eight measures of disease severity. The authors in [6] used factor analysis to select a subset of features and followed this with K-means clustering to identify population subgroups which they then

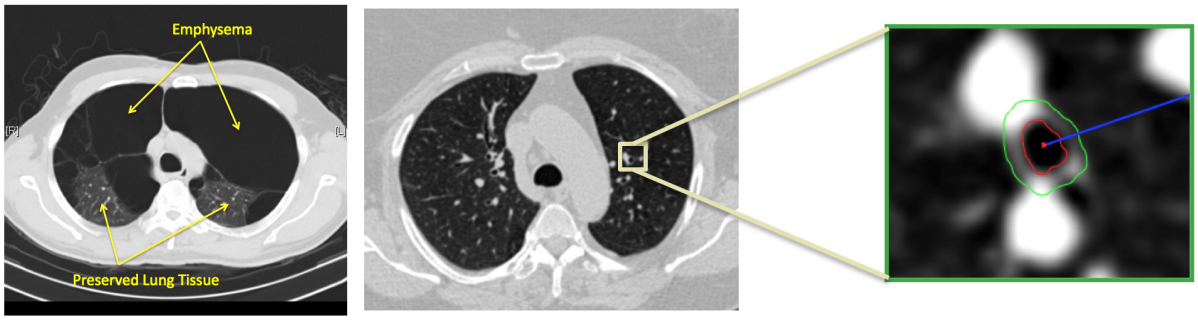

Figure 1: Left: CT scan illustrating a patient with advanced emphysema. Note that only a fraction of the lung cavity contains viable lung tissue; the remainder is empty space (emphysema). Right: the zoomed region depicts automatic detection of an airway's inner and outer walls, a step in evaluating airway thickening resulting from chronic inflammation.

investigated with genetic analysis. Recently our group performed a similar analysis in the COPDGene study cohort to identify four subgroups of subjects that associated with known genetic variants [4].

Although COPD clustering efforts to date have shed light on the disease, they have been limited by the application of approaches that do not specifically model the interplay of the available features. In particular, they pool together groups of people with different levels of disease severity, though these group differences may be partly explained by different levels of smoke exposure and age differences. Here we claim that individuals should be grouped according to biological and/or genetic similarity regardless of their level of disease severity; therefore, we seek associations of individuals to *disease trajectories* (i.e., grouping individuals based on their similarity in response to environmental and/or disease causing variables). We introduced this concept in a previous paper in which we described a clustering with constraints method using a Dirichlet process mixture of Gaussian processes in a variational Bayesian nonparametric framework [18]. This model assumes that each individual is a member of a single cluster (disease subtype), and it assumes that each feature is important for describing each of these clusters.

Here we introduce a more general framework that 1) permits both instances and features to belong to more than one cluster, 2) allows for overlapping clusters, and 3) identifies subsets of features associated with each cluster. We are again principally motivated by the concept of *disease trajectories*, but we allow for the possibility that a given individual may be suffering from multiple disease subtypes concurrently. Furthermore, it is likely the case that only a subset of features are needed to describe a particular subtype. Because the number of subtypes is unknown, we again turn to Bayesian nonparametric methods: we will describe a model that continues to use Gaussian Processes to represent trajectories and also builds on dual beta processes for instance and feature assignment to the latent subtype variables.

The rest of the paper is layed out as follows. In section 2 we provide an overview of the theory behind our model, focusing on Gaussin processes in section 2.1 and Beta processes in section 2.2. In section 3 we describe our probabilistic model; we define both the structure and the constituent probability distributions. The update equations used for variational inference are given in section 4. We demonstrate algorithm performance on both synthetic and clinical datasets in section 5 and conclude in section 6.

2. BACKGROUND

The model we propose is an instance of nonparametric overlapping subspace clustering. There are a number of related works in this context. [11] introduced the Infinite Overlapping Mixture Model (IOMM), a nonparametric clustering method that allows an unbounded number of potentially overlapping clusters. Assignments of points to (multiple) clusters is modeled using an Indian Buffet Process (IBP) and realized with a multiplicative mixture model likelihood function. This mixture model can be seen as a nonparametric generalization of the products-of-experts model introduced by [12]. While they assume an unknown number of clusters and that all features are relevant for each cluster, [8] assume that the number of clusters is known but that features are local to a particular cluster. Their Bayesian Overlapping Subspace Clustering (BOSC) model is a hierarchical generative model for matrices with potentially overlapping sub-block structures.

While [8] used a pair of Beta-Bernoulli distributions with an assumed known number of clusters as priors for the latent row and column membership vectors, we adopt nonparametric priors based on dual beta processes: one prior for the association of data instances to latent clusters, and another for the association of observed features to latent clusters. This allows our model to discover the number of latent clusters best supported by our data, the associations of instances to those clusters, and the observed features that best describe them. We continue to use a Gaussian process likelihood function as described in [18], but we extend it using a multiplicative mixture model as [11], which permits both instances and features to be assigned to multiple clusters / subtypes. Importantly, we perform inference using a variational approach which further distinguishes our method from both [8] and [11]. In the remainder of this section we describe two elements central to our model: Gaussian and beta processes.

2.1 Gaussian Processes

Gaussian Processes (GPs) have been used extensively for Bayesian nonlinear regression. We cover the key concepts here as they pertain to our framework and refer the reader to [16] for details.

Gaussian Processes can be interpreted as a nonparametric prior over functions. They have the property that given a finite sampling of the domain, the corresponding vector of function values, $\mathbf{f}$, are distributed according to a multivariate Gaussian with mean $\mathbf{m}$ (typically set to $\mathbf{0}$ in standard practice) and covariance matrix $\mathbf{K}$:

$$\mathbf{f} \sim \mathcal{N}(\mathbf{f}|\mathbf{m}, \mathbf{K}) \tag{1}$$

The elements of $\mathbf{K}$ are determined by the kernel function, k: $[\mathbf{K}]_{n,n'} = k(\mathbf{x}_n, \mathbf{x}_{n'})$. The choice of kernel function and selection of its pa-

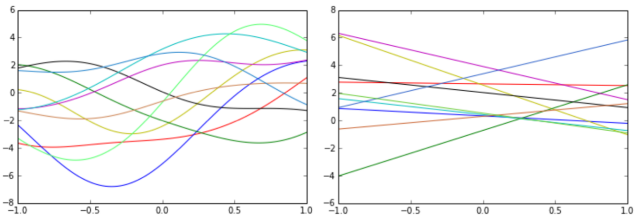

Figure 2: Ten random draws from Gaussian processes with a squared exponential kernel (left) and a linear kernel (right).

rameter values controls the behavior of the GP. One popular kernel function is the squared exponential (SE) given by

$$k\left(\mathbf{x}_n, \mathbf{x}_{n'}\right) = \theta_0 \exp\left(-\frac{\theta_1}{2}\| \mathbf{x}_n - \mathbf{x}_{n'} \|^2\right) \tag{2}$$

and another, more restrictive kernel function is the linear kernel:

$$k\left(\mathbf{x}_n, \mathbf{x}_{n'}\right) = \theta_0 + \theta_1 \sum_{d=1}^{D} (x_{n,d} - \theta_2)(x'_{n,d} - \theta_2) \tag{3}$$

where D is the vector dimension. A collection of random draws from each of these two kernels is illustrated in Figure 2.

In order to perform GP regression, we assume an observed dataset of inputs and corresponding (noisy) targets, $\mathcal{D} \equiv \{\mathbf{x}_n, y_n\}_{n=1}^{N}$, where we model the targets as $p\left(\mathbf{y}|\mathbf{f}\right) = \mathcal{N}\left(\mathbf{y}|\mathbf{f}, \sigma^2 \mathbf{I}_N\right)$. Here, σ^2 is the variance on the target variables. It can then be shown that the predicted mean and variance of target value y_* at some new input $\mathbf{x}_*$ are given by

$$\mu_* = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{y} \tag{4}$$

$$\sigma_*^2 = \sigma^2 + k_{**} - \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2 \mathbf{I}_N\right)^{-1} \mathbf{y} \tag{5}$$

where $k_{**} = k\left(\mathbf{x}_*, \mathbf{x}_*\right)$ and $\left[\mathbf{k}_*\right]_n = k\left(\mathbf{x}_n, \mathbf{x}_*\right)$.

2.2 Beta Processes

The beta process can be used as a Bayesian nonparametric prior for sparse latent feature models [9], i.e. as a prior over infinite binary matrices that indicate associations between instances (rows) to a potentially infinite number of latent features (columns). Beta processes are closely related to the Indian Buffet Process (IBP) introduced by [10], which has seen wide applicability to a number of problems. One limitation of the IBP, however, is that the distribution on the number of features per object and on the total number of latent features is coupled through a single parameter. This limits the expressivity of the prior. [9] later introduced a two-parameter version of the IBP which enables the amount of sharing between instances and latent features to be controlled. This is an attractive feature, but the authors did not describe how to realize this construction in a variational inference framework (which we note has computational benefits over sampling based inference methods as used in [9]).

More recently, [3] described a two-parameter, stick-breaking construction for beta process priors specifically for variational inference approaches. Their prior can be represented in the following manner

$$p(\mathbf{Z}|\mathbf{V}, \mathbf{T}, \mathbf{d})\, p(\mathbf{d}|\gamma)\, p(\mathbf{V}|\alpha)\, p(\mathbf{T}|\mathbf{d}, \alpha) \tag{6}$$

where $\mathbf{Z}$ is the infinite binary matrix and $\mathbf{V}$, $\mathbf{T}$, and $\mathbf{d}$ are latent features in their model. γ and α are the model parameters that control the amount of latent feature sharing between instances. We refer the reader to [3] for a detailed description of this construction.

3. FORMULATION

The desiderata of the model we propose in this section are a) the ability to identify the most probable number of latent disease subtypes, b) the flexibility to assign multiple subtypes to a given data instance (patient), and c) the capability to identify the subsets of features that best describe a given subtype.

In the remainder of this section we formalize the elements of our framework. Let $\mathbf{X} = [\mathbf{x}_1 \cdots \mathbf{x}_Q]$ be the $N \times Q$ matrix of observed inputs where N is the number of instances and Q is the dimension of the inputs. Let $\mathbf{Y} = [\mathbf{y}_1 \cdots \mathbf{y}_D]$ be the $N \times D$ matrix of corresponding target values, where D represents the dimension of the target variables. We designate the set of latent functions as $\left\{ f_d^{(\mathrm{k})}(\mathbf{x}) \right\}_{k=1,d=1}^{\infty,D}$. We collect all latent functions of trajectory k in the matrix $\mathbf{F}^{(k)} = \left[\mathbf{f}_1^{(\mathrm{k})} \cdots \mathbf{f}_D^{(\mathrm{k})}\right]$, and we designate the complete set of latent functions as $\left\{\mathbf{F}^{(\mathrm{k})}\right\}$. We introduce two latent binary indicator matrices, $\mathbf{Z}^r$ and $\mathbf{Z}^c$ (where the superscripts refer to the original $N \times D$ data matrix: r referring to the rows of that matrix, and c the columns). $\mathbf{Z}^r$ is an N row by infinite column matrix where each row represents a data instance and each column represents a latent cluster (subtype). Similarly, $\mathbf{Z}^c$ is a D row by infinite column matrix corresponding to the target features. We place beta process priors on each of these binary matrices using the stick-breaking construction introduced in [3].

The probabilistic graphical model describing our formulation can be seen in Figure 3, and the corresponding joint distribution is given by

$$\begin{aligned} &p(\mathbf{Y}, \{\mathbf{F}^{(k)}\}, \mathbf{d}^r, \mathbf{V}^r, \mathbf{T}^r, \mathbf{Z}^r, \mathbf{d}^c, \mathbf{V}^c, \mathbf{T}^c, \mathbf{Z}^c) = \\ &p\left(\mathbf{Y}|\left\{\mathbf{F}^{(k)}\right\}, \mathbf{Z}^r, \mathbf{Z}^c\right) p\left(\left\{\mathbf{F}^{(k)}\right\}|\,\mathbf{X}\right) \times \\ &p(\mathbf{d}^r|\gamma_r)\, p(\mathbf{V}^r|\alpha_r)\, p(\mathbf{T}^r|\mathbf{d}^r, \alpha_r)\, p(\mathbf{Z}^r|\mathbf{V}^r, \mathbf{T}^r, \mathbf{d}^r) \times \\ &p(\mathbf{d}^c|\gamma_c)\, p(\mathbf{V}^c|\alpha_c)\, p(\mathbf{T}^c|\mathbf{d}^c, \alpha_c)\, p(\mathbf{Z}^c|\mathbf{V}^c, \mathbf{T}^c, \mathbf{d}^c) \end{aligned} \tag{7}$$

where we use two beta process priors – one for representing the association of instances to the latent clusters (indicated with the r superscript), and one for representing the association of observed features to latent clusters (indicated with the c superscript). (See section 2.2 above and [3] for details about the beta process prior).

The prior placed over the GPs representing each disease trajectory are given by

$$p\left(\left\{\mathbf{F}^{(k)}\right\}|\,\mathbf{X}\right) = \prod_{k=1}^{\infty} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{f}_d^{(k)}|\mathbf{m}_d, \mathbf{K}_d^{(k)}\right) \tag{8}$$

Note that we specify distinct kernel matrices and mean vectors for each target feature dimension d.

The likelihood term is given by

$$p\left(\mathbf{Y}_{n,d}|\left\{\mathbf{F}^{(k)}\right\}, \mathbf{Z}^r, \mathbf{Z}^c\right) = \begin{cases} \frac{1}{c(\mathbf{z})} \prod_{k=1}^{\infty} \mathcal{N}\left(\mathbf{Y}_{n,d}|\mathbf{F}_{n,d}^{(k)}, \sigma_d^2\right)^{\mathbf{z}_k}, & \text{if } \mathbf{z} \neq \mathbf{0} \\ \mathcal{N}\left(\mathbf{Y}_{n,d}|\hat{\mu}_d, \hat{\sigma}_d^2\right), & \text{Otherwise} \end{cases} \tag{9}$$

where $\mathbf{z}_k$ is the k^{th} element of the vector $\mathbf{z} \equiv \mathbf{Z}^r{}_{n,\cdot} \bigodot \mathbf{Z}^c{}_{d,\cdot}$, the element-wise (Hadamard) product between vectors $\mathbf{Z}^r{}_{n,\cdot}$ and $\mathbf{Z}^c{}_{d,\cdot}$, and $c(\mathbf{z})$ is the normalization factor. If every element of $\mathbf{z}_k$ is 0, then $\mathbf{Y}_{n,d}$ is assumed to be generated from the noise component, $\mathcal{N}\left(\mathbf{Y}_{n,d}|\hat{\mu}_d, \hat{\sigma}_d^2\right)$.

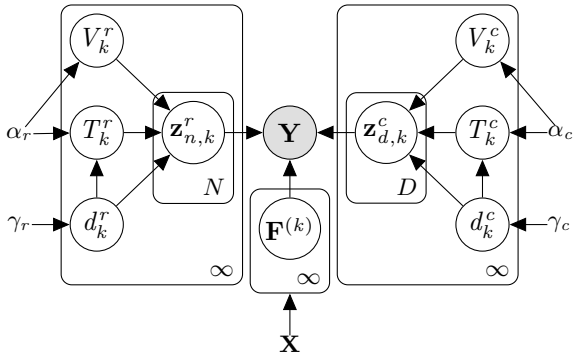

Figure 3: Probabilistic graphical model for our formulation.

4. INFERENCE

In this section we give the variational inference update equations that are specific to our model. Variational inference is a method of approximate inference that makes assumptions (typically a factorization) over the distribution of interest, and it turns an inference problem into an optimization problem [13]. Additionally, whereas approximate inference methods based on sampling (such as Monte Carlo Markov Chain) can be slow to converge, variational inference enjoys a greater computational advantage in this regard.

For our application, we are interested in the distribution over the latent variables in our model given our observations:

$$p\left(\{\mathbf{F}^{(k)}\}, \mathbf{Z}^r, \mathbf{Z}^c, \mathbf{d}^r, \mathbf{V}^r, \mathbf{T}^r, \mathbf{d}^c, \mathbf{V}^c, \mathbf{T}^c, |\, \mathbf{X}, \mathbf{Y}\right) \tag{10}$$

Direct evaluation of this posterior is intractable, so we approximate the posterior with a factorized distribution:

$$p^*\left(\{\mathbf{F}^{(k)}\}\right) p^*(\mathbf{Z}^r)\, p^*(\mathbf{Z}^c)\, p^*(\mathbf{d}^r)\, p^*(\mathbf{V}^r)\, p^*(\mathbf{T}^r) \times p^*(\mathbf{d}^c)\, p^*(\mathbf{V}^c)\, p^*(\mathbf{T}^c) \tag{11}$$

As described in [3] we have

$$p^*(\mathbf{d}_k^r) = \text{Mult}\left(\mathbf{d}_k^r | \varphi_k\right) \tag{12}$$
$$p^*(\mathbf{V}^r) = \text{Beta}\left(\mathbf{V}_k^r |\, a_k^r, b_k^r\right) \tag{13}$$
$$p^*(\mathbf{T}^r) = \text{Gam}\left(\mathbf{T}_k^r |\, u_k^r, v_k^r\right) \tag{14}$$
$$p^*\left(\mathbf{Z}_{n,k}^r\right) = \text{Bern}\left(\mathbf{Z}_{n,k}^r | \phi_{n,k}\right) \tag{15}$$

and similarly for $p^*(\mathbf{d}^r)$, $p^*(\mathbf{V}^r)$, $p^*(\mathbf{T}^r)$, and $p^*\left(\mathbf{Z}_{d,k}^c\right)$. Parameter updates for the variational distributions over $\mathbf{d}_k^r$, $\mathbf{V}^r$, and $\mathbf{T}^r$ (similarly for $\mathbf{d}_k^c$, $\mathbf{V}^c$, and $\mathbf{T}^c$) can be found in [3]. In the remainder of this section we provide updates for $p^*\left(\{\mathbf{F}^{(k)}\}\right)$, $p^*(\mathbf{Z}^r)$, and $p^*(\mathbf{Z}^c)$.

To derive update expressions for the factors in the variational distribution, we compute the expectation of the log joint distribution with respect to the other factors in the variational distribution. In the case of $p^*\left(\{\mathbf{F}^{(k)}\}\right)$ this is

$$\ln p^*\left(\{\mathbf{F}^{(k)}\}\right) = \mathbb{E}\left\{\ln p\left(\cdot\right)\right\} + \text{const} \tag{16}$$

where $p\left(\cdot\right)$ abbreviates the joint distribution given in Equation 7 and the expectation is with respect to every variable in the the variational distribution other than $\{\mathbf{F}^{(k)}\}$. This leads to the variational distribution over $\{\mathbf{F}^{(k)}\}$

$$p^*\left(\{\mathbf{F}^{(k)}\}\right) = \prod_{k=1}^{K} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{f}_d^{(k)} | \mu_d^{(k)}, \mathbf{C}_d^{(k)}\right) \tag{17}$$

where

$$\mathbf{C}_d^{(k)} = \left(\mathbf{K}_d^{(k)\text{-}1} + \mathbf{R}_d^{(k)}\right)^{-1} \tag{18}$$
$$\mu_d^{(k)} = \mathbf{C}_d^{(k)}\left(\mathbf{K}_d^{(k)\text{-}1}\mathbf{m}_d + \mathbf{R}_d^{(k)}\mathbf{y}_d\right) \tag{19}$$

and

$$\mathbf{R}_d^{(k)} = \frac{\phi_{d,k}}{\sigma_d^2}\begin{pmatrix} \xi_{1,d}\phi_{1,k} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \xi_{N,d}\phi_{N,k} \end{pmatrix} \tag{20}$$

where $\xi_{n,d}$ is the probability that instance n is associated with *any* latent feature that observed feature d can describe and is given by

$$\xi_{n,k} = 1 - \prod_{k=1}^{K}\left(1\text{-}\phi_{n,k}\phi_{d,k}\right) \tag{21}$$

Unfortunately, there is no closed form update expression for the parameters of $\mathbf{Z}^r$ and $\mathbf{Z}^c$. It can be shown that the difference between the factorized approximation and the true posterior is minimized when a quantity known as the variational lower bound is maximized. Therefore, we must turn to the variational lower bound and directly optimize those terms that depend on these parameters. We proceed by describing updates for $\mathbf{Z}^r$; updates for $\mathbf{Z}^c$ are obtained analogously. First we note that there are four terms in the lower bound that directly depend on $\mathbf{Z}^r$:

$$\sum_{n=1}^{N}\sum_{k=1}^{K} \varphi_k\left(1\right) \mathbb{E}\left\{\ln p\left(\mathbf{Z}_{n,k}^r | \mathbf{V}_k\right)\right\} + \sum_{n=1}^{N}\sum_{k=1}^{K} \varphi_k\left(r>1\right) \mathbb{E}\left\{\ln p\left(\mathbf{Z}_{n,k}^r | \mathbf{V}_k, \mathbf{T}_k\right)\right\} - \mathbb{E}\left\{\ln p^*(\mathbf{Z}^r)\right\} + \mathbb{E}\left\{\ln p\left(\mathbf{Y} | \{\mathbf{F}^{(k)}\}, \mathbf{Z}^r, \mathbf{Z}^c\right)\right\} \tag{22}$$

Variables φ_k and r appear in the stick-breaking construction of the beta process described in [3].

Expansion of the first term in Equation 22 gives

$$\phi_{n,k}\varphi_k\left(1\right)\left(\psi\left(a_k^r\right) \text{-} \psi\left(b_k^r\right)\right) \tag{23}$$

where $\psi\left(\cdot\right)$ is the digamma function. The second term in Equation 22 expands to

$$\phi_{n,k}\varphi_k(r>1)\left(\psi(a_k) - \psi(a_k+b_k) - \frac{u_k}{v_k} + \sum_{m=1}^{M}\Delta_k(m)\right) \tag{24}$$

where $\Delta_k(m)$ is given by

$$\frac{1}{m}\frac{\Gamma(a_k+b_k)}{\Gamma(a_k+b_k+m)}\frac{\Gamma(a_k+m)}{\Gamma(a_k)}\left(\frac{v_k}{v_k+m}\right)^{u_k} \tag{25}$$

and M is set to $1,000$ as in [3]. Expansion of the third term in 22 gives

$$-\phi_{n,k}\ln\phi_{n,k} - (1-\phi_{n,k})\ln(1-\phi_{n,k}) \tag{26}$$

Finally, the last term in 22 expands as

$$\sum_{d=1}^{D}\left(\xi_{n,d}\left[\sum_{k=1}^{K}\phi_{n,k}\phi_{d,k}\,g_{n,d}^{(k)}-h_{n,d}-\frac{1}{2}+\frac{1}{2}\ln\frac{1}{2\pi\sigma_d^2}-\Xi_{n,d}\right]\right) \tag{27}$$

where we use the following definitions

$$g_{n,d}^{(k)}\equiv\frac{1}{2}\left(\frac{\ln K}{K-1}\right)-\frac{\mathbf{Y}_{n,d}^2}{2\sigma_d^2}+\frac{\mathbf{Y}_{n,d}\mu_{n,d}^{(k)}}{\sigma_d^2} \tag{28}$$

$$h_{n,d}\equiv\frac{1}{2}\ln\left(\frac{1}{2\pi\hat{\sigma}_d^2}\right)-\frac{1}{2\hat{\sigma}_d^2}\left(\mathbf{Y}_{n,d}-\hat{\mu}_d\right)^2 \tag{29}$$

$$\Xi_{n,d}\equiv\frac{1}{2\sigma_d^2}\,\mathbb{E}\left\{\left(\sum_{k=1}^{K}\mathbf{Z}_{n,k}^r\mathbf{Z}_{d,k}^c\mathbf{F}_{n,d}^{(k)}\right)^2\right\} \tag{30}$$

and we note that the expression in Equation 30 can be evaluated in a straightforward manner by expanding the terms.

Together Equations 23, 24, 26, and 27 constitute an objective function for $\phi_{n,k}$ that we optimize using Brent's method on the interval $[0,1]$. Brent's method is a derivative-free optimization algorithm that uses inverse parabolic interpolation when possible to speed up convergence of the golden section method [1]. An analogous procedure is applied to update each $\phi_{d,k}$ using a similar objective function, except the outer sum in Equation 27 is from $n=1$ to N.

There are two terms in the updates for $\phi_{n,k}$ and $\phi_{d,k}$ that deserve special attention. The expectation in Equation 30 is an approximation the term

$$\mathbb{E}\left\{\left(\sum_{k=1}^{K}\mathbf{Z}_{n,k}^r\mathbf{Z}_{d,k}^c\mathbf{F}_{n,d}^{(k)}\right)^2\Bigg/\sum_{k=1}^{K}\mathbf{Z}_{n,k}^r\mathbf{Z}_{d,k}^c\right\} \tag{31}$$

This term emerges as a consequence of using the multiplicative mixture model, and the sum in the denominator makes this expectation intractable. However, we note that the expression we use in 30 is an upper bound of the original term given that the denominator in 31 will always be greater than or equal to 1 (recall that the likelihood given in Equation 9 is only active when $\mathbf{z}\neq 0$). Since it is the negative of Equation 31 that shows up in the lower bound, our approximation is in fact a lower bound of the original term. Hence we are gauranteed to be optimizing a lower bound of the original lower bound.

The second term of interest also comes from expanding the last term in Equation 22:

$$\mathbb{E}\left\{\ln\left(\sum_{k}^{K}\mathbf{Z}_{n,k}^r\mathbf{Z}_{d,k}^c\right)\right\} \tag{32}$$

which we again note is intractrable. However, noting that the sum will always be between 1 and K and that the natural log is a concave function, we can substitute Equation 33 with

$$\frac{\ln K}{K-1}\,\mathbb{E}\left\{\sum_{k}^{K}\mathbf{Z}_{n,k}^r\mathbf{Z}_{d,k}^c\right\}-1 \tag{33}$$

which is easy to evaluate.

5. EXPERIMENTS

In this section we describe experimental results. We begin by illustrating our approach on a synthetic dataset and then describe the application of our algorithm to data taken from the COPDGene cohort [17].

5.1 Synthetic Example

Our algorithm is designed to identify not only the number of latent clusters, but also which data instances and which observed features associate with them. We demonstrate this with a synthetic example consisting of 120 samples and four observed features. The first two observed features are designed to be redundant: they both describe the same two latent clusters, represented by two lines; half of the samples belong to one line in both features, and the other half belong to the other. The third observed feature is a noise feature. Here all samples are drawn from the same normal distribution. The fourth observed feature contains three line segments that are distinct from the groupings present in the first two observed features (each of the three line segments consists of a third of the samples from each of the two lines described by observed features one and two).

We ran our algorithm with $K=20$, $\sigma^2=\hat{\sigma}^2=0.2$, $\alpha^r=3.5$, and $\gamma^r=\alpha^c=\gamma^c=1.5$. We used a linear kernel for each of the Gaussian processes with $\theta_0=2.0$ and $\theta_1=1.5$. Except for the variances, no special attention was given to the parameter settings; these were the first values selected and they produced the reported results. The variances provided to the algorithm were the same values used to generate the samples.

Figure 5.1 illustrates the results. Each latent feature is color-coded, so it can be seen that the first two observed features both describe the same two latent clusters. Observed feature three was correctly identified as a noise feature, i.e. not capable of describing any of the five latent clusters present in the data set. The algorithm also found the three latent clusters described by observed feature four.

5.2 Clinical Experiments

Here we report results from an experiment performed on clinical data from the COPDGene study, a large epidemiologic and genetic study of over $10,000$ current and former smokers with and without COPD [17]. All subjects had blood collected for genetic analysis, and they completed spirometry and chest computed tomography (CT) scans, resulting in a large collection of features. The set of observed features we consider consists of seven lung function measures: functional residual capacity (FRC), pre- and post-bronchodilator forced expiratory volume in one second (FEV1), pre- and post-bronchodilator forced vital capacity (FVC), and pre- and post-bronchodilator FEV1/FVC. We also consider four CT-based measures of emphysema: total percent emphysema, the fifteenth percentile level of the intensity histogram in the lungs (Perc 15), and percent emphysema in the upper and lower lung lobes. We also include three airway disease measures: the wall area percent (WA%), percent gas trapping, and the predicted WA% of an airway with a $10mm$ perimeter (Pi10).

The predictors of our model are age, height, and pack years, which is a measure of life-long smoke exposure defined as the number of packs per day times the number of years of cigarette smoking. These constitute the inputs to the GP covariance matrices, and are meant to capture the factors that are causative of COPD severity. The use of height as a predictor of disease may seem odd, but it is known to influence lung function given that it affects lung mechanics, so we opted to include it in our model. We again choose to use the linear kernel for our experiments. This amounts to a form of nonparametric polynomial regression, and there are less computationally heavy ways to do this than with a Gaussian process framework. However, we emphasize that using GPs provide a much more flexible representation of possible disease trajectories. But in the absence of constraints between instances or observed features, we choose a more restrictive class of kernel function so as

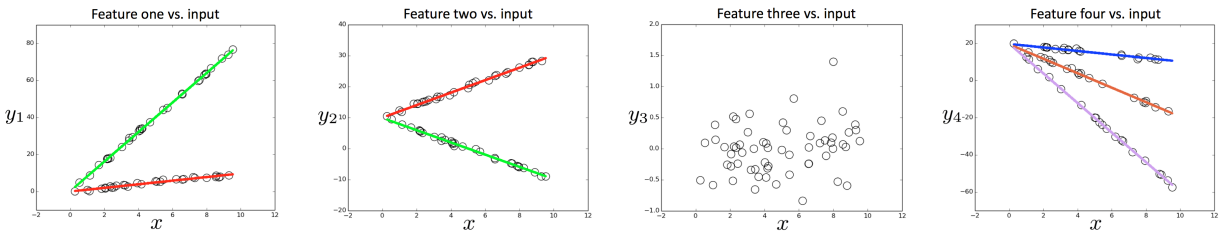

Figure 4: Synthetic example in which five latent clusters (color-coded) are discovered in a data set with four observed features. Left and left-middle: observed features y_1 and y_2 are redundant: both are useful for describing the red and green latent clusters. Right-middle: observed feature y_3 is determined to be a noise feature and is not useful for describing any of the five latent clusters. Right: observed feature y_4 describes three latent clusters, distinct from those described by observed features y_1 and y_2.

Table 1: Inputs and linear kernel parameters used for each GP in the clinical experiment

Feature	Inputs	θ_0	θ_0
Pre FEV1/FVC	Age, Pack Yrs	1.0	3.5
Post FEV1/FVC	Age, Pack Yrs	2.0	3.5
Pre FEV1	Age, Pack Yrs, Height	2.0	3.5
Post FEV1	Age, Pack Yrs, Height	2.0	3.5
Pre FVC	Age, Pack Yrs, Height	50.0	3.5
Post FVC	Age, Pack Yrs, Height	50.0	3.5
Total % Emph.	Age, Pack Yrs	50.0	3.5
Perc 15	Age, Pack Yrs	2e5	3.5
FRC	Age, Pack Yrs, Height	5.0	3.5
% Gas Trapping	Age, Pack Yrs	500.0	3.5
% Emph. Upper Lobes	Age, Pack Yrs	2.0	3.5
% Emph. Lower Lobes	Age, Pack Yrs	50.0	3.5
Pi10	Age, Height	2.0	3.5
WA%	Pack Yrs, Height	5e3	3.5

Feature	Red	Blue	Green	Magenta
Pre FEV1/FVC	1.00	1.00	1.00	1.00
Post FEV1/FVC	1.00	1.00	1.00	1.00
Pre FEV1	1.00	1.00	1.00	1.00
Post FEV1	1.00	1.00	1.00	1.00
Pre FVC	1.00	0.87	0.84	0.40
Post FVC	1.00	1.00	0.59	0.38
Total % Emph.	1.00	1.00	0.92	1.00
Perc 15	0.00	0.20	0.00	1.00
FRC	0.96	1.00	0.88	0.31
% Gas Trapping	1.00	1.00	1.00	1.00
% Emph. Upper Lobes	1.00	1.00	0.49	0.54
% Emph. Lower Lobes	1.00	1.00	0.98	0.65
Pi10	0.00	0.16	0.14	0.13
WA%	0.27	0.218	0.28	0.20

Table 2: Inputs and probability of association to each of four latent features, labeled by color for easy comparison to other figures.

to mitigate the effect of local minima during variational inference. Note that in our previous work we were able to apply more general kernel functions in the presence of data constraints [18]. We hope to extend the current approach once such constraints are available.

The inputs and kernel parameters for each of the observed features are given in Table 1. In order to guide the selection of inputs and kernel parameters, we performed standard multi-variate linear regression on each of the observed features using each of the three inputs as predictors. For a given observed feature, we only used as kernel inputs those predictors that had a significant association in the linear regression model. The slope intercept generated by the model informed the selection of the θ_0 parameter, which controls the intercept range over which a draw from the GP is likely to come from. For all fourteen observed features, we noticed only modest slope values for each of the regression coefficients. Therefore, we selected a θ_1 value of 3.5 for all kernel functions; we observed empirically from repeated draws of GPs using this value that slopes tend to be modest.

We chose σ_d^2 for each of the observed features by taking one tenth of the variance of the residuals from the multi-variate regression stage. This was an ad hoc selection, but the measurement variances for each of the observed features is as yet unknown. The selected values were chosen to be significantly lower than the residual variances in order to explore subgroups within the data. By comparison, simply using the residual variances would likely have caused the algorithm to simply recapitulate the multi-variate regression result, although we did not attempt this experiment. The noise means and variances, $\hat{\mu}_d$ and $\hat{\sigma_d^2}$, were simply chosen to be the mean and variance of each observed feature.

For the beta process priors, we set $K = 20$, $\alpha^r = 3.5$, and $\gamma^r = \alpha^c = \gamma^c = 1.5$ as in the synthetic experiments. We considered the first $1,000$ subjects to have enrolled in the COPDGene study, keeping only those that have complete data for all inputs and observed features, resulting in a collection of 851 subjects. We deployed our algorithm on our institution's computer cluster, and ran 200 jobs in parallel, executing 15 iterations for each job. Using this dataset with these parameters, each job took approximately 10 hours to complete. For each job we recorded the final variational lower bound value, and report results for the job that gave the largest lower bound.

Our algorithm identified four latent clusters in this data set, and assigned approximately half of the samples to the noise model. For each of the observed features, we give the probability of each being associated with each of the latent clusters in Table 2. We see that both Pi10 and WA% poorly describe the latent clusters; that is to say, those observed features are better described by the noise model. This is not unexpected, given that they both rely on direct measurements of airway wall thickness on CT images. Given that airways are small structures and difficult to measure accurately, these measurements are known to be noisy. On the other hand, the lung function measurements tend to do a much better job at describing the latent clusters.

Figure 5 provides scatter plots for several of the observed features, where we have again color-coded the latent clusters. We

include a plot of Pi10, one of the variables that poorly describes the latent clusters, to illustrate the lack of structure in the data. It is also interesting to note that Perc 15 is only useful for describing two of these latent clusters. Summarizing these data we can identify two groups of individuals who have preserved lung function and lung tissue despite increasing age and smoke exposure (red and blue groups). The magenta and green groups appear to represent those individuals that are more susceptible to COPD: they show increased levels of emphysema and gas trapping as well as lower levels of lung function given increasing age and smoke exposure, with the magenta group being more susceptible than the green group.

Finally, we evaluate our clustering results by performing genetic association analysis using several single nucleotide polymorphisms (SNPs) known to associate with COPD [5]. As a comparison, we apply the same methodology we used in [4]: K-means clustering (with $K = 4$) on a feature set determined by factor analysis. We note however that our previous analysis was conducted on the entire COPDGene cohort, and here we only consider the 851 subjects analyzed by our algorithm. Table 3 shows the resulting odds ratios and associated confidence intervals. The odds ratio represents the effect size of the genetic variant; it is the odds of having the variant in one subtype divided by the odds of having the variant in another subtype – values further from 1.0 indicate greater effect of the genetic variant on membership in that subtype. Odds ratios are computed with respect to the healthiest group in each clustering result (the "red" group in the case of our algorithm). In the genome-wide analyses of COPD, effect sizes for the previously described genetic variants were approximately less than or equal to 1.4, suggesting that our approach can identify a more genetically susceptible subgroup.

6. CONCLUSION

We have introduced a nonparametric overlapping subspace clustering algorithm that relies on dual beta process priors in order to identify latent cluster structure. Our model finds associations between data instances and latent clusters, allowing for a given instance to belong to multiple clusters. Additionally, our algorithm identifies the observed features that best describe the latent clusters and thus serves as a form of feature selection. The likelihood term in our model is specifically chosen to address the challenges of disease subtype discovery in COPD: by using Gaussian processes to represent the dependence between observed features that measure levels of disease severity and inputs that are causative agents of disease progression, we can flexibly represent so-called "disease trajectories". We believe our contribution represents a step forward towards a better understanding of a complicated disease that will hopefully lead to better patient care.

We have made an implementation of our algorithm available on GitHub here, and the COPDGene data can be obtained from dbGaP here.

7. ACKNOWLEDGMENTS

This work was supported by NSF grant IIS-0915910, and by the US National Heart, Lung, and Blood Institute grants R01HL089856, R01HL089897 (COPDGene), K08HL102265, K08HL097029 and R01HL113264. The COPDGene study is also supported by the COPD Foundation through contributions made to an Industry Advisory Board comprised of AstraZeneca, Boehringer Ingelheim, Novartis, Pfizer, Siemens, and Sunovion.

References

[1] R. P. Brent. *Algorithms for minimization without derivatives*. Courier Dover Publications, 2013.

[2] P. R. Burgel, J. Paillasseur, D. Caillaud, I. Tillie-Leblond, P. Chanez, R. Escamilla, T. Perez, P. Carré, N. Roche, et al. Clinical copd phenotypes: a novel approach using principal component and cluster analyses. *European Respiratory Journal*, 36(3):531–539, 2010.

[3] L. Carin, D. M. Blei, and J. W. Paisley. Variational inference for stick-breaking beta process priors. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 889–896, 2011.

[4] P. J. Castaldi, J. Dy, J. Ross, Y. Chang, G. R. Washko, D. Curran-Everett, A. Williams, D. A. Lynch, B. J. Make, J. D. Crapo, et al. Cluster analysis in the copdgene study identifies subtypes of smokers with distinct patterns of airway disease and emphysema. *Thorax*, 69(5):416–423, 2014.

[5] M. H. Cho, P. J. Castaldi, E. S. Wan, M. Siedlinski, C. P. Hersh, D. L. Demeo, B. E. Himes, J. S. Sylvia, B. J. Klanderman, J. P. Ziniti, et al. A genome-wide association study of copd identifies a susceptibility locus on chromosome 19q13. *Human molecular genetics*, 21(4):947–957, 2012.

[6] M. H. Cho, G. R. Washko, T. J. Hoffmann, G. J. Criner, E. A. Hoffman, F. J. Martinez, N. Laird, J. J. Reilly, and E. K. Silverman. Cluster analysis in severe emphysema subjects using phenotype and genotype data: an exploratory investigation. *Respir Res*, 11:30, 2010.

[7] M. Decramer, W. Janssens, and M. Miravitlles. Chronic obstructive pulmonary disease. *The Lancet Respiratory Medicine*, 379(9823), 2012.

[8] Q. Fu and A. Banerjee. Bayesian overlapping subspace clustering. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 776–781. IEEE, 2009.

[9] Z. Ghahramani, T. L. Griffiths, and P. Sollich. Bayesian nonparametric latent feature models. 2007.

[10] T. Griffiths and Z. Ghahramani. Infinite latent feature models and the indian buffet process. 2005.

[11] K. A. Heller and Z. Ghahramani. A nonparametric bayesian approach to modeling overlapping clusters. In *International Conference on Artificial Intelligence and Statistics*, pages 187–194, 2007.

[12] G. E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural computation*, 14(8):1771–1800, 2002.

[13] M. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. *Machine learning*, 37(2):183–233, 1999.

[14] C. D. Mathers and D. Loncar. Projections of global mortality and burden of disease from 2002 to 2030. *PLoS Med*, 3(11):e442, 11 2006.

[15] S. L. Murphy, J. Xu, and K. D. Kochanek. Deaths: final data for 2010. *National vital statistics reports*, 61(4), 2013.

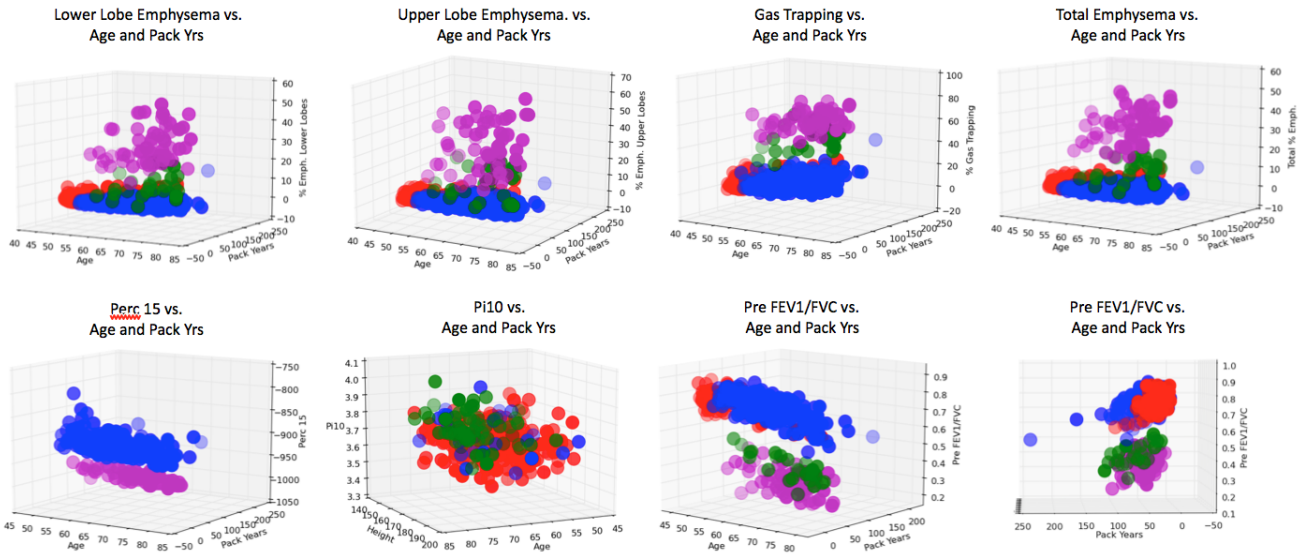

Figure 5: Scatter plots of COPDGene subjects color-coded by association to discovered latent clusters. The red and blue groups correspond to healthier individuals while the green and magenta groups represent those more susceptible to disease. The scatter plot of Pi10 illustrates a noisy feature. The Pre FEV1/FVC plot in the lower right is rotated to better illustrate that each group exhibits a decline in lung function with increasing smoke exposure.

Table 3: Odds ratios and associated confidence intervals (in parentheses) for three SNPs known to associate with COPD. All reported values are statistically significant at the $p < 0.05$ confidence level; a dash indicates that no significant association was found. Our dual beta-process, Gaussian process algorithm is abbreviated DBP-GP. For the K-means result we abbreviate the two cluster groups found to have significant genetic associations as C1 and C2. For the DBP-GP groups, we indicate cluster groups using the color-coding scheme described above. Next to each cluster identifier is the number of samples in that group.

	K-means		DBP-GP		
SNP	C1 (197)	C2 (328)	Blue (206)	Green (26)	Magenta (60)
rs13180	-	1.44 (1.26-1.64)	-	0.46 (0.33 - 0.65)	0.63 (0.50 - 0.80)
rs8034191	-	0.74 (0.65 - 0.84)	1.76 (1.43 - 2.17)	3.19 (2.18 - 4.66)	1.72 (1.33 - 2.23)
rs7937	1.38 (1.20 - 1.59)	-	-	-	-

[16] C. Rasmussen and C. Williams. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

[17] E. A. Regan, J. E. Hokanson, J. R. Murphy, B. Make, D. A. Lynch, T. H. Beaty, D. Curran-Everett, E. K. Silverman, and J. D. Crapo. Genetic epidemiology of copd (copdgene) study design. *COPD: Journal of Chronic Obstructive Pulmonary Disease*, 7(1):32–43, 2011.

[18] J. Ross and J. Dy. Nonparametric mixture of gaussian processes with constraints. In *Proceedings of the 30th International Conference on Machine Learning (ICML-13)*, pages 1346–1354, 2013.