Nonparametric Mixture of Gaussian Processes with Constraints

James C. Ross[1,2] JROSS@BWH.HARVARD.EDU
Jennifer G. Dy[2] JDY@ECE.NEU.EDU
1: Brigham and Women's Hospital, Boston, MA 02115 USA
2: Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115 USA

Abstract

Motivated by the need to identify new and clinically relevant categories of lung disease, we propose a novel clustering with constraints method using a Dirichlet process mixture of Gaussian processes in a variational Bayesian nonparametric framework. We claim that individuals should be grouped according to biological and/or genetic similarity regardless of their level of disease severity; therefore, we introduce a new way of looking at subtyping/clustering by recasting it in terms of discovering associations of individuals to *disease trajectories* (i.e., grouping individuals based on their similarity in response to environmental and/or disease causing variables). The nonparametric nature of our algorithm allows for learning the unknown number of meaningful trajectories. Additionally, we acknowledge the usefulness of expert guidance by providing for their input using *must-link* and *cannot-link* constraints. These constraints are encoded with Markov random fields. We also provide an efficient variational approach for performing inference on our model.

1. Introduction

Personalized medicine holds the promise of providing individuals tailored medical care optimally suited to their needs. In recent years, there has been an explosion of clinical, biological, and genetic data, the analysis of which will hopefully bring us closer to realizing this goal. Understanding distinct mechanisms of disease – unique biological pathways and their genetic determinants – is at the core of this endeavor and is often referred to as "*disease sub-typing*".

In this paper, we are specifically motivated by the task of identifying novel and clinically relevant categories of Chronic Obstructive Pulmonary Disease (COPD), a smoking related lung disease with a significant health burden worldwide. Alpha$_1$-antitrypsin deficiency is one known form of genetic disorder leading to COPD (Silverman & Sandhaus, 2009); experts hypothesize that there are other distinct, as yet unkown, categories of this disease determined by genetic predisposition (Cho, 2010; 2012; Barker & Brightling, 2013). The challenge is to identify these subgroups given large amounts of data obtained from clinical studies. The key difficulty is grouping individuals with similar genetic make-up in spite of significantly different levels of disease severity. For example, a younger person with little exposure to smoke and relatively healthy lungs should be placed in the same category with an older, life-long smoker with advanced lung disease provided they have the same genetic or biological predisposition.

The manner in which lung health changes as a function of age and smoke exposure can be used to identify meaningful subgroups. Some people are genetically resistant to the effects of smoke exposure and have preserved lung health even after years of smoking. On the other hand, others are highly sensitive to smoke and experience rapid health decline given similar levels of exposure. This leads to the notion of "*disease trajectories*", and indeed there is an analogy to the trajectories of projectiles moving through space. We seek meaningful disease trajectories with the hypothesis that those individuals associated with the same trajectory have similar genetic predispositions to lung health decline. The problem is that we do not know how many such trajectories (disease subgroups) exist, nor do we know the functional forms of those trajectories.

The traditional way to discover unknown subgroups given data is by clustering (Jain et al., 1999). Clustering algorithms group data based on some notion

Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia, USA, 2013. JMLR: W&CP volume 28.

of similarity. Standard clustering algorithms typically define similarity in the form of some metric or a probability model. Most standard methods do not take the structure of the problem into account and treat all the features/variables in the same way; however, in our COPD sub-typing problem, we have variables such as age and smoking that are causative agents of variables that indicate lung function and disease severity. The type of grouping we are interested in discovering relates to how different groups of individuals respond to exposure. This led us to the design of a mixture of Gaussian process (GP) regression model.

There are algorithms for clustering time series data (Li & Prakash, 2011). These methods assume that each sample has a time-sampled measurement. In our case, it is not always possible to work with *longitudinal* data (data in which a given individual is studied at multiple time points); many studies are *cross-sectional*. Our approach is flexible in the sense that input variables can represent any entity that directly affects measurable lung health or disease severity, including age and smoke exposure. Our model is also able to learn component "trajectory" functions even when we only have one sample per patient. This is possible because complete clinical datasets typically have multiple representatives of the same trajectory captured at different stages of the disease process.

We use a mixture of GPs rather than the standard mixture of regression models (Grün & Leisch, 2007), because we do not know what the regression model is. A Gaussian process (Rasmussen & Williams, 2006) provides a nonparametric distribution over functions. There has been work combining GPs with mixture models (Rasmussen & Ghahramani, 2002; Meeds & Osindero, 2006; Yuan & Neubauer, 2009). These works address modeling data where there are local discontinuities. In a local region of the input space, there is a gating function that determines which GP component it is generated from. Our work addresses GP components at a global scale.

In 2012 Lázaro-Gredilla et al. (2012) introduced a mixture of Gaussian Processes to address the data association problem, which arises in multi-target tracking scenarios. As alluded to in the earlier paragraphs, this scenario is similar to the one we are interested in, with one important difference: whereas they assumed the number of trajectories is known, we do not. To address this issue, we recast their formulation in a Bayesian nonparametric framework using the stick-breaking Dirichlet Process Model (Blei & Jordan, 2006).

The added flexibility provided by the nonparametric model makes finding local minima more likely. We steer inference towards meaningful solutions by incorporating must-link and cannot-link constraints (Wagstaff & Cardie, 2000; Zhu, 2008) between data instances. This is an important feature of our model as it provides a mechanism to include expert input (doctors, biologists, geneticists, etc.). Basu et al. (2006) demonstrated the use of Hidden Markov Random Fields (HMRF) to apply such constraints for semi-supervised clustering. Orbanz & Buhmann (2008) used MRFs to impose constraints in a nonparametric setting for spatial smoothing in image segmentation; they performed inference using Gibbs sampling. Inspired by these approaches, we also use MRFs to encode must-link and cannot-link constraints, and we further demonstrate a variational approach for performing approximate inference.

In this paper, we introduce a novel variational Dirichlet process mixture of Gaussian processes that can also learn from must-link and cannot-link constraints. The contributions of this work are: 1) our model is able to learn the number of clusters (trajectories) automatically for a mixture of GPs; 2) we provide a model allowing a mixture of GPs to learn from constraints; 3) we derive a variational inference approach to clustering with contraints encoded using MRFs; and 4) we present a transformative way of looking at sub-typing COPD; instead of applying traditional clustering algorithms, we utilize our domain knowledge regarding the disease mechanism and cast it as a problem of discovering multiple "*disease trajectories*".

The rest of the paper is organized as follows. In Section 2 we give a brief overview of the theory behind our model. 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 4, and we describe the conditions under which efficient computation is possible. We demonstrate algorithm performance on both synthetic and real-world datasets in Section 5, and we conclude in Section 6.

2. Background

In this section we briefly review theory on which our model builds: Gaussian processes, Markov random fields, and Dirichlet process mixtures.

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 Rasmussen & Williams (2006) 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 $\mathbf{0}$ (arbitrary but used in standard practice) and covariance matrix $\mathbf{K}$:

$$\mathbf{f} \sim \mathcal{N}\left(\mathbf{f}|\mathbf{0}, \mathbf{K}\right) \tag{1}$$

The elements of $\mathbf{K}$ are determined by the kernel function, k: $\left[\mathbf{K}\right]_{n,n'} = k\left(\mathbf{x}_n, \mathbf{x}_{n'}\right)$. The choice of kernel function and selection of its parameter values controls the behavior of the GP. One popular kernel function (and the one used throughout our experiments) is the exponential of a quadratic form given by

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

In order to perform GP regression, we assume an observed dataset of inputs and corresponding (noisy) targets, $\mathcal{D} \equiv \{\mathbf{x}_\mathrm{n}, y_\mathrm{n}\}_{n=1}^{N}$, 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

$$\mu_* = \mathbf{k}_*^\top \left(\mathbf{K} + \sigma^2\mathbf{I}_N\right)^{-1}\mathbf{y} \tag{3}$$

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

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. Markov Random Fields

A Markov random field (MRF) is represented by an undirected graphical model in which the nodes represent variables or groups of variables and the edges indicate dependence relationships. An important property of MRFs is that a collection of variables is conditionally independent of all others in the field given the variables in their Markov blanket. The Hammersley-Clifford theorem states that the distribution, $p\left(\mathbf{Z}\right)$, over the variables in a MRF factorizes according to

$$p\left(\mathbf{Z}\right) = \frac{1}{\mathcal{Z}} \exp\left(-\sum_{\mathrm{c}\in\mathrm{C}} \mathrm{H_c}\left(\mathbf{z}_\mathrm{c}\right)\right) \tag{5}$$

where $\mathcal{Z}$ is a normalization constant called the *partition function*, C is the set of all cliques in the MRF, $\mathbf{z}_c$ are the variables in clique c, and H_c is the *energy function* over clique c (Geman & Geman, 1984; Besag, 1974). The energy function captures the desired configuration of local variables.

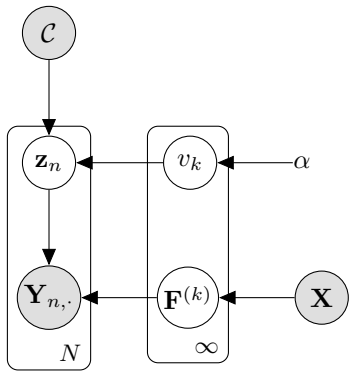

Figure 1. Probabilistic graphical model for constrained, nonparametric, Gaussian process regression.

2.3. Dirichlet Process Mixtures

Ferguson (1973) first introduced the Dirichlet process (DP) as a measure on measures. It is parameterized by a base measure, G_0, and a positive scaling parameter α:

$$G|\left\{G_0, \alpha\right\} \sim \mathrm{DP}\left(G_0, \alpha\right) \tag{6}$$

The notion of a Dirichlet process mixture (DPM) arises if we treat the k^{th} draw from G as a parameter of the distribution over some observation (Antoniak, 1974). DPMs can be interpreted as mixture models with an infinite number of mixture components.

More recently, Blei & Jordan (2006) described a variational inference algorithm for DPMs using the stick-breaking construction introduced by Sethuraman (1991). The stick-breaking construction represents G as

$$\pi_k\left(\mathbf{v}\right) = v_k \prod_{j=1}^{k-1}\left(1 - v_j\right) \tag{7}$$

$$G = \sum_{i=1}^{\infty} \pi_i\left(\mathbf{v}\right)\delta_{\eta_i^*} \tag{8}$$

where $\delta_{\eta_i^*}$ is the Kronecker delta, and the v_i are distributed according to a beta distribution: $v_i \sim \mathrm{Beta}\left(1, \alpha\right)$, and $\eta_i^* \sim G_0$. The use of the stick-breaking construction in our formulation will be discussed in Section 3.

3. Our Formulation

In this section we describe our formulation, including a definition of the variables in our model.

Let $\mathbf{X} = \left[\mathbf{x}_1 \cdots \mathbf{x}_Q\right]$ 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 introduce the $N \times \infty$ binary indicator matrix, $\mathbf{Z}$, to represent the association between the data instances and the latent regression functions. Following the notation in Lázaro-Gredilla et al. (2012), 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 $\left\{\mathbf{F}^{(\mathrm{k})}\right\}$.

The probabilistic graphical model describing our formulation can be seen in Figure 1. The set $\mathcal{C}$ is a collection of data instance pairs representing given must-link and cannot-link constraints. With these quantities defined, we give the joint distribution of our model:

$$p\left(\mathbf{Y}, \mathbf{Z}, \mathbf{v}, \{\mathbf{F}^{(k)}\}\right) = p\left(\{\mathbf{F}^{(k)}\} \,|\, \mathbf{X}\right) p\left(\mathbf{Y} | \{\mathbf{F}^{(k)}\}, \mathbf{Z}\right) p\left(\mathbf{Z} \,|\, \mathbf{v}, \mathcal{C}\right) p\left(\mathbf{v} \,|\alpha\right) \tag{9}$$

where

$$p\left(\{\mathbf{F}^{(k)}\} \,|\, \mathbf{X}\right) = \prod_{k=1}^{\infty} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{f}_d^{(k)} | \mathbf{0}, \mathbf{K}^{(k)}\right) \tag{10}$$

$$p\left(\mathbf{Y} | \{\mathbf{F}^{(k)}\}, \mathbf{Z}\right) = \prod_{n=1}^{N} \prod_{k=1}^{\infty} \prod_{d=1}^{D} \mathcal{N}\left(\mathbf{Y}_{n,d} | \mathbf{F}_{n,d}^{(k)}, \sigma^2\right)^{\mathbf{Z}_{n,k}} \tag{11}$$

$$p\left(\mathbf{Z} \,|\, \mathbf{v}, \mathcal{C}\right) = \frac{1}{\mathcal{Z}} \exp\left(- \sum_{(i,j) \in \mathcal{C}} \mathrm{H}\left(\mathbf{z}_i, \mathbf{z}_n\right) \right) \prod_{n=1}^{N} \prod_{k=1}^{\infty} \left(v_k \prod_{j=1}^{k\text{-}1} (1\text{-}v_j) \right)^{\mathbf{Z}_{n,k}} \tag{12}$$

$$p\left(\mathbf{v} \,|\alpha\right) = \prod_{k=1}^{\infty} \mathrm{Beta}\left(v_k | 1, \alpha\right) \tag{13}$$

Equation 10 represents the prior distribution over the infinite collection of Gaussian processes. The likelihood in our model is given in Equation 11; note that this distribution factorizes over the target dimensions but that the same Gaussian process covariance matrix for a given regressor is used for all dimensions. We also assume that the variances for each target variable dimension, σ^2, are known and constant. This is a realistic assumption for our disease sub-typing use case: devices that measure disease severity can have their measurement variance characterized. For applications where σ^2 is not known, this and other hyperparameters can be automatically learned via empirical Bayes.

Equation 12 describes the distribution over $\mathbf{Z}$ and consists of two terms: the first is a MRF that captures the pairwise constraints, and the second is a multinomial distribution with parameters drawn for a Dirichlet process using the stick-breaking construction. Equation 13 expresses the distribution over the variable, $\mathbf{v}$, used for the stick-breaking process; here α is the concentration parameter.

The energy function used in our experiments is given by

$$\mathrm{H}\left(\mathbf{z}_i, \mathbf{z}_j\right) = \begin{cases} -w_{i,j}, & <\mathbf{z}_i, \mathbf{z}_j> = 1 \text{ and } (i,j) \text{ is } ML \\ -w_{i,j}, & <\mathbf{z}_i, \mathbf{z}_j> = 0 \text{ and } (i,j) \text{ is } CL \\ 0, & \text{Otherwise} \end{cases} \tag{14}$$

where $<\mathbf{z}_i, \mathbf{z}_j>$ represents the inner product between $\mathbf{z}_i$ and $\mathbf{z}_j$, ML stands for must-link and CL for cannot-link, and $w_{i,j}$ is in the interval $[0, 1]$ with lower values expressing less confidence in the constraint and vice-versa.

While our formulation has similarities to Lázaro-Gredilla et al. (2012), we emphasize that our algorithm is both nonparametric in the number of mixture components and semi-supervised, important features for our intended application. Additionally, while Orbanz & Buhmann (2008) showed that MRFs can be incorporated with DPMs, but they performed inference using Gibbs sampling. In Section 4 we will show that variational inference can be applied provided certain conditions are satisfied by the constraints.

4. Inference

In this section we give the variational inference update equations used in 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 (Jordan et al., 1999; Jaakkola, 2001). 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: $p\left(\mathbf{Z}, \mathbf{v}, \{\mathbf{F}^{(k)}\} \,|\, \mathbf{X}, \mathbf{Y}\right)$. The posterior probability is approximated by optimizing the variational lower bound. The standard variational inference approach is to assume a factorized approximation of this distribution, in our case $p^*(\mathbf{Z})\, p^*\left(\{\mathbf{F}^{(k)}\}\right) p^*(\mathbf{v})$. In order to derive the expression for one of these factors, the expectation with respect to the other factors is considered. Derivation of the variational distributions begins with the following expressions

$$\ln p^*(\mathbf{Z}) = \mathbb{E}_{\{\mathbf{F}^{(k)}\},\mathbf{v}}\left\{\ln \mathrm{p}\left(\mathbf{Y}, \mathbf{Z}, \mathbf{v}, \{\mathbf{F}^{(k)}\}\right)\right\} + \text{const} \tag{15}$$

$$\ln p^*\left(\{\mathbf{F}^{(k)}\}\right) = \mathbb{E}_{\mathbf{Z},\mathbf{v}}\left\{\ln \mathrm{p}\left(\mathbf{Y}, \mathbf{Z}, \mathbf{v}, \{\mathbf{F}^{(k)}\}\right)\right\} + \text{const} \tag{16}$$

$$\ln p^*(\mathbf{v}) = \mathbb{E}_{\mathbf{Z},\{\mathbf{F}^{(k)}\}}\left\{\ln \mathrm{p}\left(\mathbf{Y}, \mathbf{Z}, \mathbf{v}, \{\mathbf{F}^{(k)}\}\right)\right\} + \text{const} \tag{17}$$

Given space limitations, we provide the expressions for each factor without derivation.

The variational distribution over $\{\mathbf{F}^{(k)}\}$ is given as

$$p^*\left(\{\mathbf{F}^{(k)}\}\right) = \prod_{k=1}^{K}\prod_{d=1}^{D}\mathcal{N}\left(\mathbf{f}_d^{(k)} \,|\mu^{(k)}, \mathbf{C}^{(k)}\right) \tag{18}$$

where

$$\mathbf{C}^{(k)} = \left(\mathbf{K}^{(k)^{-1}} + \mathbf{R}^{(k)}\right)^{-1} \tag{19}$$

$$\mu^{(k)} = \mathbf{C}^{(k)}\,\mathbf{R}^{(k)}\,\mathbf{y}_d \tag{20}$$

and

$$\mathbf{R}^{(k)} = \frac{1}{\sigma^2}\begin{pmatrix} \mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}_{1,k} & 0 & \cdots \\ 0 & \ddots & \\ \vdots & & \mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}_{N,k} \end{pmatrix} \tag{21}$$

Note that as in Blei & Jordan (2006), our approximate distribution truncates the stick-breaking construction, so that k ranges from 1 to K (set to 20 in all our experiments).

The expression for $p^*(\mathbf{v})$ is given by

$$p^*(\mathbf{v}) = \prod_{k=1}^{K}\mathrm{Beta}\left(v_k \middle| 1 + \sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}_{n,k},\ \alpha + \sum_{j=k+1}^{K}\sum_{n=1}^{N}\mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}_{n,j}\right) \tag{22}$$

Finally, the distribution for $p^*(\mathbf{Z})$ is given by

$$p^*(\mathbf{Z}) = \prod_{\mathrm{V}\in\mathcal{V}}\left[\frac{1}{\mathcal{Z}_{\mathrm{V}}}\exp\left(-\sum_{\substack{(i,j)\in\mathcal{C}\\ i,j\in\mathrm{V}}}\mathrm{H}(\mathbf{z}_i, \mathbf{z}_n)\right)\prod_{n\in\mathrm{V}}\prod_{k=1}^{K} r_{n,k}^{\mathbf{Z}_{n,k}}\right] \tag{23}$$

where

$$r_{n,k} = \frac{\rho_{n,k}}{\sum_{k=1}^{K}\rho_{n,k}} \tag{24}$$

$$\ln \rho_{n,k} = \sum_{d=1}^{D}\left[\ln\frac{1}{\sqrt{2\pi\sigma^2}} - \frac{1}{2\sigma^2}\left(\mathbf{Y}_{n,d}{}^2 - 2\,\mathbf{Y}_{n,d}\,\mathbb{E}_{\{\mathbf{F}^{(k)}\}}\left\{\mathbf{F}_{n,d}^{(k)}\right\} + \mathbb{E}_{\{\mathbf{F}^{(k)}\}}\left\{\mathbf{F}_{n,d}^{(k)}{}^2\right\}\right)\right] + \mathbb{E}_{\mathbf{v}}\{\ln v_k\} + \sum_{j=1}^{k\text{-}1}\mathbb{E}_{\mathbf{v}}\{\ln(1-v_j)\} \tag{25}$$

In Equation 23, $\mathcal{V}$ represents a set of sets. Each element V of $\mathcal{V}$ is a set of data indices belonging to a connected subgraph of the constraint MRF. Because the set of constraints is generally sparse, the MRF can be characterized by a collection of disconnected subgraphs. If the constraint set is dense, we can approximate the distribution by truncating the neighborhood to enforce low cardinality. It is important to note that the distribution factorizes over the resultant subgraphs. Given that each subgraph cardinality is small, it is feasible to compute the corresponding partition function, $\mathcal{Z}_{\mathrm{V}}$. This in turn enables efficient computation of $\mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}$.

As an example, consider the MRF shown in Fig. 2. Here, $\mathcal{V} = \{\{1,4\},\{2\},\{3,5,6,8\},\{7\},\{9\}\}$. Note that each subgraph cardinality is low (with a maximum of four in this example), so that their corresponding partition functions are easily computed.

Inference begins by randomly initializing the matrix $\mathbf{Z}$ such that each element is equal to or greater than zero and each row sums to one. We then iteratively update equations 18, 22, and 23 until we observe no change in $\mathbb{E}_{\mathbf{Z}}\{\mathbf{Z}\}$ or until a pre-specified number of iterations is reached.

5. Experiments

In this section we demonstrate algorithm performance on both synthetic and real-world datasets. For all our experiments, the cardinality of the constraint subgraphs was kept below 5. No special attention was

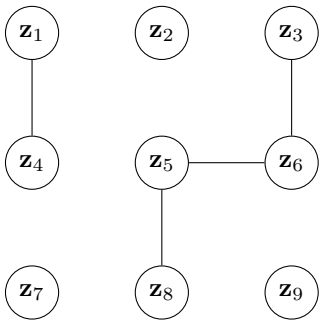

Figure 2. Example MRF illustrating disconnected subgraphs. Each graph edge represents either a must-link or cannot-link constraint.

given to the reported parameter settings for α, θ_0, or θ_1. Rather, a coarse parameter selection of reasonable values was used.

5.1. Experiments on Synthetic Data

We tested algorithm performance on two synthetic datasets. The first consists of noisy samples taken from two curves: a sinusoid and a sinusoid with a modest linear offset. The second dataset is made up of noisy samples taken from two interlaced helices in 3D. For both cases, the algorithm was run for 50 iterations. α was set to 1.0, and θ_0 was set to 1.0 in both cases. For the sinusoids experiment, $\sigma^2 = 0.02$ and $\theta_1 = 0.005$. For the helices experiment, $\sigma^2 = 0.1$ and $\theta_1 = 0.0005$. Must-link constraints were generated by randomly choosing pairs of points from a given function, preferring pairs that are spaced farther apart. Cannot-link constraints were generated by randomly choosing pairs of points from different functions, preferring pairs in regions where the functions tend to be closer to one another.

We used normalized mutual information (*NMI*) (Strehl & Ghosh, 2003) to investigate algorithm performance for a number of different constraints. Letting A represent the cluster assignments determined by the algorithm and B represent the ground-truth cluster assignments, the *NMI* is given by $NMI = \frac{H(A)-H(A|B)}{\sqrt{H(A)H(B)}}$, where $H(\cdot)$ is the entropy. Higher *NMI* values mean that the clustering results are more similar to ground-truth; the criterion reaches its maximum value of one when there is perfect agreement.

Figure 3 gives the results of the synthetic experiments. Each entry in the rightmost plot of this figure represents the average *NMI* score across fifty, randomly initialized runs. There is a clear increase in performance

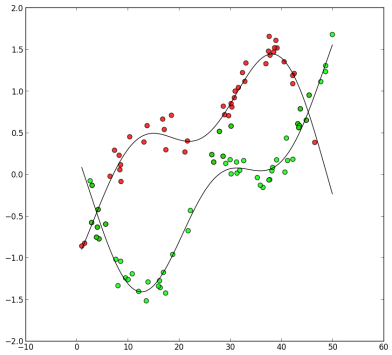

Figure 4. Illustration of an algorithm output for the unconstrained case. While the curves provide a reasonable explanation of the data, it may not be the solution of interest.

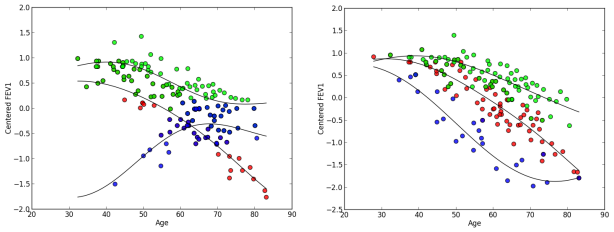

Figure 5. Left: example of learned regressors during training for the unconstrained cases. Right: learned regressors using constraints. Plots are taken from different folds.

with added constraints in both cases. The center plots illustrate the regression curves found by the algorithm, and the data instances are color coded according to their association to each curve.

Without constraints, the algorithm has a greater tendency to converge on solutions that may not be of interest. This is illustrated in Figure 4. While the solution shown does a reasonably good job of explaining the data, this particular solution might not be "optimal". By adding constraints, the optimization landscape is modified to one more favorable for finding interesting solutions.

5.2. Experiments on Real-World Data

As stated in the introduction, the motivation for our model stems from the need to identify clinically meaningful subtypes of lung disease. Here we show results on data from the Normative Aging Study (NAS) (Bell, 1972), a longitudinal study designed to investigate the role of aging on various health issues, including lung function. The complete dataset includes a large number of features; here we focus on the effects of age on a widely used measure of lung function, FEV1 (forced

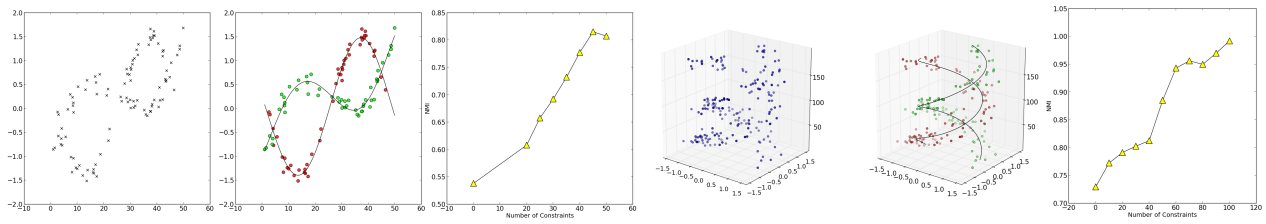

Figure 3. Illustration of algorithm performance on two synthetic datasets. Left: original data. Middle: a correct result found by the algorithm, color-coded by association to regression curves. Right: Average *NMI* score as a function of totaled ML and CL constraints used.

expiratory volume in one second). We randomly chose a subset of forty subjects such that each subject was represented at a minimum of five time points and everyone had approximately the same height.

Since our goal is to identify regression curves that associate subjects according to genetics, longitudinal studies like NAS provide a good arena in which to test and implicitly provide constraints: all data instances belonging to a given subject are must-linked together (i.e., there are "built-in" constraints).

It is known that lung function decline (as measured by a decrease in FEV1) is a natural part of the aging process, even in healthy individuals. However, some individuals are thought to experience a more rapid decline while others a more modest decline. This effect is thought to be even more pronounced as a function of smoke exposure. For our initial analysis we focus on age as the input variable. (Accurately capturing exposure to smoke is nontrivial and will be the focus of our future work). We investigate our algorithm's performance by performing a five-fold cross validation. For each fold the test set consists of a randomly selected time point for each individual, and the training set consists of the remaining data. No data point is repeated as a test instance across the different folds. For each of the five training sessions, we learn regression curves both with and without constraints, identifying solutions with the lowest variational bound in each case.

We want to identify curves that are genetically/biologically meaningful despite various levels of measured lung function, so we desire solutions such that the data points for an individual are associated to the same curve during the training phase. We report the percentage of times this occurs for each of the five folds, both with and without constraints.

We are also interested in the predictive power of the learned regressors. For each instance in the test set we identify the curve most often associated to that individual in the training set and use that regressor to predict the FEV1 value associated with the test instance; we do this for both the constrained and unconstrained cases. Additionally, we compare our predictions to those made by the currently accepted prediction equation used in clinical practice (Hankinson, 1999) given by

$$\text{FEV1} = 0.5536 - 0.01303 \times age - \\ 0.000172 \times age^2 + 0.00011607 \times height^2 \qquad (26)$$

We ran 50 iterations for all experiments and set $\sigma^2 = 0.0225$, $\alpha = 1.0$, $\theta_0 = 1.0$, and $\theta_1 = 0.002$. The results are summarized in Table 1.

We also highlight examples of learned regressors for both the constrained and unconstrained case in Figure 5. The constrained case depicts trends that agree well with clinical expectation, while the unconstrained case shows an unexpected increase in lung health for one of the sub-populations, clearly contrary to what is known about lung physiology and the aging process.

Although our algorithm was designed specifically for application to lung disease sub-typing, our last experiment shows that it is potentially useful for related tracking scenarios. We demonstrate this by considering a video-sequence taken from the EU CAVIAR

Figure 6. Detected individuals in a frame from EU CAVIAR video sequence.

Table 1. Five-fold cross-validation results on clinical data taken from the Normative Aging Study. The first three columns show the average mean squared error between actual and predicted FEV1 measures using the standard clinical prediction equation ("Clin. Pred.") and predictions using our algorithm with ("Alg. Pred. (Constr.)") and without ("Alg. Pred. (Unconstr.)") constraints. The last two columns show the matching percentages for the constrained and unconstrained cases; see text for details.

Fold	Clin. Pred.	Alg. Pred. (Constr.)	Alg. Pred. (Unconstr.)	Match Perc. (Constr.)	Match Perc. (Unconstr.)
1	0.57	**0.24**	0.38	**0.88**	0.57
2	0.36	**0.06**	0.18	**0.84**	0.63
3	0.58	**0.20**	0.37	**0.84**	0.60
4	0.46	**0.28**	0.40	**0.84**	0.56
5	0.49	**0.24**	0.31	**0.83**	0.59

dataset[1]. This is a human-labeled benchmark sequence featuring four individuals walking through a scene. Each of the 1,164 data instances used here consist of the sequence's frame number, and the target values are the centroids of each detected bounding-box. Each of the four individuals in the scene is assigned a unique ID in the available ground-truth, and we use that information to impose ML and CL constraints on our algorithm. We again run for 50 iterations and set $\sigma^2 = 2.0$, $\alpha = 0.03$, $\theta_0 = 100.0$, and $\theta_1 = 0.0005$. Results are shown in Figure 7.

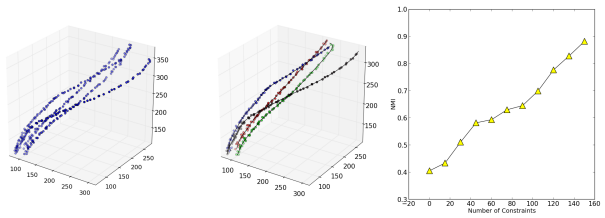

Figure 7. Algorithm performance on the EU CAVIAR video sequence. Left: original data. Middle: a correct result found by the algorithm, color-coded by association to regression curves. Right: Average normalized mutual information *NMI* score as a function of totaled ML and CL constraints used.

For all experiments described in this section, our algorithm was able to identify meaningful results both in terms of the number of regressors as well as their functional forms. As the number of constraints increases, the results converged on are more likely to represent the solution of interest. The flexibility to automatically identify both the number of regressors and their forms while honoring valuable expert input are the key advantages of our approach.

[1] http://homepages.inf.ed.ac.uk/rbf/CAVIAR/

6. Conclusion

We have introduced a nonparametric, mixture of Gaussian process regression framework that uses must-link and cannot-link constraints to identify solutions of interest. Our motivation for building this model is to assist with lung disease sub-type idenfication; we have provided a new way of looking at this problem by recasting it in terms of discovering associations of individuals to disease trajectories, and we have demonstrated the efficacy of our approach on real-world clinical data. In the process of designing an appropriate learning model for solving this clinical problem, we have developed a novel Dirichlet process mixture of Gaussian processes with constraints. It is applicable to other applications requiring clustering/data association to trajectories or nonparametric functions. We have also successfully shown its effectiveness on synthetic and tracking data.

Acknowledgments

This work was supported by US NIH grants R01 HL089856 and R01 HL089897. We thank Michael H. Cho and Peter J. Castaldi for their guidance on clinical aspects of our model, and we thank Augusto Litonjua for supplying Normative Aging Study data.

References

Antoniak, Charles E. Mixtures of dirichlet processes with applications to bayesian nonparametric problems. *The annals of statistics*, pp. 1152–1174, 1974.

Barker, Bethan L and Brightling, Christopher E. Phenotyping the heterogeneity of chronic obstructive pulmonary disease. *Clinical Science*, 124(6):371–387, 2013.

Basu, S., Bilenko, M., Banerjee, A., and Mooney, R.J. Probabilistic semi-supervised clustering with constraints. *Semi-supervised learning*, pp. 71–98, 2006.

Bell, Benjamin et al. The normative aging study: an interdisciplinary and longitudinal study of health and aging. *The International Journal of Aging and Human Development*, 3(1):5–17, 1972.

Besag, Julian. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

Blei, David M and Jordan, Michael I. Variational inference for dirichlet process mixtures. *Bayesian Analysis*, 1(1):121–143, 2006.

Cho, Michael H et al. Variants in fam13a are associated with chronic obstructive pulmonary disease. *Nature genetics*, 42(3):200–202, 2010.

Cho, Michael H et al. A genome-wide association study of copd identifies a susceptibility locus on chromosome 19q13. *Human molecular genetics*, 21(4):947–957, 2012.

Ferguson, Thomas S. A bayesian analysis of some nonparametric problems. *The annals of statistics*, pp. 209–230, 1973.

Geman, Stuart and Geman, Donald. Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.

Grün, Bettina and Leisch, Friedrich. Fitting finite mixtures of generalized linear regressions in r. *Computational Statistics & Data Analysis*, 51(11):5247–5252, 2007.

Hankinson, John L et al. Spirometric reference values from a sample of the general us population. *American journal of respiratory and critical care medicine*, 159(1):179–187, 1999.

Jaakkola, Tommi S. 10 tutorial on variational approximation methods. *Advanced mean field methods: theory and practice*, pp. 129, 2001.

Jain, Anil K, Murty, M Narasimha, and Flynn, Patrick J. Data clustering: a review. *ACM computing surveys (CSUR)*, 31(3):264–323, 1999.

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

Lázaro-Gredilla, Miguel, Vaerenbergh, Steven Van, and Lawrence, Neil D. Overlapping mixtures of gaussian processes for the data association problem. *Pattern Recognition*, 45(4):1386–1395, 2012.

Li, Lei and Prakash, B Aditya. Time series clustering: Complex is simpler! In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pp. 185–192, 2011.

Meeds, Edward and Osindero, Simon. An alternative infinite mixture of gaussian process experts. *Advances in Neural Information Processing Systems*, 18:883, 2006.

Orbanz, P. and Buhmann, J.M. Nonparametric bayesian image segmentation. *International Journal of Computer Vision*, 77(1):25–45, 2008.

Rasmussen, C.E. and Ghahramani, Z. Infinite mixtures of gaussian process experts. *Advances in neural information processing systems*, 2:881–888, 2002.

Rasmussen, C.E. and Williams, C.K.I. *Gaussian processes for machine learning*, volume 1. MIT press Cambridge, MA, 2006.

Sethuraman, Jayaram. A constructive definition of dirichlet priors. Technical report, DTIC Document, 1991.

Silverman, Edwin K and Sandhaus, Robert A. Alpha1-antitrypsin deficiency. *New England Journal of Medicine*, 360(26):2749–2757, 2009.

Strehl, Alexander and Ghosh, Joydeep. Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *The Journal of Machine Learning Research*, 3:583–617, 2003.

Wagstaff, Kiri and Cardie, Claire. Clustering with instance-level constraints. In *Proceedings of the Seventeenth International Conference on Machine Learning*, pp. 1103–1110, 2000.

Yuan, Chao and Neubauer, Claus. Variational mixture of gaussian process experts. *Advances in Neural Information Processing Systems*, 21:1897–1904, 2009.

Zhu, Xiaojin. Semi-supervised learning literature survey. 2008.