Journal of Machine Learning Research 17 (2016) 1-35 Submitted 8/15; Revised 6/16; Published 12/16

Integrative Analysis using Coupled Latent Variable Models for Individualizing Prognoses

Peter Schulam PSCHULAM@CS.JHU.EDU
Suchi Saria SSARIA@CS.JHU.EDU
Department of Computer Science
Johns Hopkins University
Baltimore, MD 21218, USA

Editor: Benjamin M. Marlin and C. David Page

Abstract

Complex chronic diseases (e.g., autism, lupus, and Parkinson's) are remarkably heterogeneous across individuals. This heterogeneity makes treatment difficult for caregivers because they cannot accurately predict the way in which the disease will progress in order to guide treatment decisions. Therefore, tools that help to predict the trajectory of these complex chronic diseases can help to improve the quality of health care. To build such tools, we can leverage clinical markers that are collected at baseline when a patient first presents and *longitudinally* over time during follow-up visits. Because complex chronic diseases are typically systemic, the longitudinal markers often track disease progression in multiple organ systems. In this paper, our goal is to predict a function of time that models the future trajectory of a single target clinical marker tracking a disease process of interest. We want to make these predictions using the histories of many related clinical markers as input. Our proposed solution tackles several key challenges. First, we can easily handle irregularly and sparsely sampled markers, which are standard in clinical data. Second, the number of parameters and the computational complexity of learning our model grows linearly in the number of marker types included in the model. This makes our approach applicable to diseases where many different markers are recorded over time. Finally, our model accounts for latent factors influencing disease expression, whereas standard regression models rely on observed features alone to explain variability. Moreover, our approach can be applied dynamically in continous-time and updates its predictions as soon as any new data is available. We apply our approach to the problem of predicting lung disease trajectories in scleroderma, a complex autoimmune disease. We show that our model improves over state-of-the-art baselines in predictive accuracy and we provide a qualitative analysis of our model's output. Finally, the variability of disease presentation in scleroderma makes clinical trial recruitment challenging. We show that a prognostic tool that integrates multiple types of routinely collected longitudinal data can be used to identify individuals at greatest risk of rapid progression and to target trial recruitment.

Keywords: gaussian processes, conditional random fields, prediction of functional targets, latent variable models, disease trajectories, precision medicine

1. Introduction

In complex chronic diseases (CCD) such as autism, lupus, and Parkinson's, the way the disease manifests may vary greatly across individuals. This makes treatment challenging

©2016 Peter Schulam and Suchi Saria.

because caregivers cannot easily predict an individual's future trajectory to guide therapy decisions. For example, in scleroderma, an autoimmune disorder, lung disease is a common cause of morbidity and mortality (Varga et al., 2012), but there are no known biomarkers or precise algorithms for stratifying individuals into groups based on similar lung disease course. A tool that can provide accurate forecasts of disease progression can help clinicians to tailor treatments to each patient based on their most likely course.

To monitor disease progression, clinicians collect many clinical markers both at baseline when an individual first visits the clinic and *longitudinally* during routine follow-up visits. Many CCDs are systemic, and so the markers are designed to monitor the disease's impact across many organ systems. In scleroderma, individuals may be affected across six organ system—the lungs, heart, skin, gastrointestinal tract, kidneys, and vasculature—to varying extents (Varga et al., 2012). Example clinical markers include PFVC (percent of predicted forced vital capacity), which is used to measure lung damage severity; TSS (total skin score), which is used to measure skin disease activity; and, PDLCO (percent of carbon monoxide diffused by the lung), used to measure vasculature health.

Our goal is to predict a function of time that models the future trajectory of a single target clinical marker tracking a disease process of interest. We want to make these predictions by leveraging baseline information and additional time-dependent clinical markers (henceforth referred to as auxiliary markers) as they are collected. This is the focal challenge of personalized medicine: integrative analysis of heterogeneous data from an individual's medical history to improve care (Collins and Varmus, 2015). So far, efforts in integrative analysis have focused on combining inferences from molecular data modalities (Rosenbloom et al., 2013). Our focus in this paper is on leveraging routinely recorded information from the electronic health record—both static and time-dependent—to make precise estimates of an individual's disease course.

A key challenge in this setting is that these data are collected during routine clinical visits and therefore they are sparse and irregularly sampled. Predicting an individual's future disease is commonly framed as a regression problem where the target clinical marker at a specific time in the future is modeled as a function of observed input features alone. These features are computed by generating summaries from the observed data (e.g., the last PFVC value or the trend in the PFVC over the last six months). However, training conditional models is less straightforward from data where varying numbers of repeated measurements are sampled per patient and across different markers. In this setting, others have focused on dynamical prediction (e.g., Rizopoulos and Ghosh 2011; Proust-Lima et al. 2014) by fitting parametric models to the longitudinal data and using the resulting model parameters as features for prediction. But existing formulations do not scale to high-dimensional problems with many auxiliary markers.

Another key challenge in predicting disease trajectories in CCDs is that differences in trajectories across individuals may be largely due to factors that are not yet known. For example, different disease pathways or biological mechanisms (e.g., genetic mutations or autoimmune markers) may be driving different subtypes of the disease (Lewis et al., 2005; Lötvall et al., 2011; Doshi-Velez et al., 2014; Saria and Goldenberg, 2015), each associated with distinct disease trajectories (Schulam et al., 2015). But, in many diseases, our knowledge of these pathways is, at best, limited. In this setting, Schulam et al. (2015) use a latent variable model to infer subtypes—subgroups with similar trajectories—using

repeated measurements of clinical marker data in the electronic health record. Schulam and Saria (2015) extend these ideas and introduce a transfer learning framework for predicting individual-specific disease trajectories that accounts for subtypes and other latent factors causing heterogeneity in disease expression. These works, however, focus on modeling single marker trajectories. We build on Schulam and Saria (2015) in this paper.

1.1 Contributions

In this paper, we describe a scalable framework for predicting a target marker trajectory (i.e. a continuous-time function) that allows us to use multiple longitudinal clinical marker histories as inputs. Our approach makes it easy to handle irregular sampling patterns across markers. Because we use a discriminative training criterion that conditions on marker histories instead of jointly modeling them, the framework is not as sensitive to misspecified dependencies across marker types. Moreover, the number of parameters and computational complexity scales linearly with the number of markers, which makes it possible to apply our approach in high-dimensional settings where many different marker types are available. Finally, our approach aligns with the dynamical nature of clinical medicine; it can be used to make predictions using continuously growing marker histories. We apply our approach to the problem of predicting lung disease trajectories in scleroderma, a complex autoimmune disease. We show that our model improves over state-of-the-art baselines in predictive accuracy and we provide a qualitative analysis of our model's output. Moreover, we demonstrate the clinical utility of our model by measuring performance on early detection of individuals who develop aggressive lung disease.

2. Related Work

Most predictive models used in medicine are cross-sectional—they use features from data measured up until the current time to predict a clinical marker or outcome at a fixed point in the future. As an example, consider the mortality prediction model by Lee et al. (2003), where logistic regression is used to integrate features into a prediction about the probability of death within 30 days for a given patient. To predict the outcome at multiple time points, it is common to fit separate models (e.g., Wang et al. 2012; Zhou et al. 2011). These models are trained to use features extracted from a fixed-size window, rather than a dynamically growing history. Moreover, they tackle heterogeneity in a limited way—any differences across individuals must be explained by observed features alone.

A common approach to dynamical prediction of trajectories is to use Markov models such as order-p autoregressive models (AR-p), HMMs, state space models, and dynamic Bayesian networks (e.g. Hassan and Nath 2005; Quinn et al. 2009; Murphy 2002). While such models naturally make dynamic predictions using the full history by forward-filtering, they typically assume discrete, regularly-spaced observation times.

To model an individual's disease trajectory using sparse and irregularly sampled clinical markers, we draw heavily from ideas in the functional data analysis (FDA) literature (see e.g., Ramsay 2006). In FDA, sequences of measurements are assumed to be samples from an underlying continuous function. A common first-step in FDA is to project the irregular observations on to a functional basis, such as B-splines, and then analyze the time series in coefficient space. However, coefficient estimates can have high variance when a time series

has too few observations, which is common in clinical data. James and Sugar (2003) address this issue by modeling the parameters of individual trajectories as random variables with a low-rank parameterization of the mean and covariance. This work is closely related to ours, and the idea of sharing statistical strength across trajectories through a structured prior over individual-specific parameters is used broadly throughout trajectory analysis to account for sparsity. Gaussian processes (GPs) are also commonly used in FDA; they offer flexible nonparametric models of trajectories but can also help to counteract sparsity by sharing kernel hyperparameters across individuals—see Roberts et al. (2013) for a recent review of GPs applied to time series data. Recent work by Liu and Hauskrecht (2014) combines the advantages of Markov models (e.g. AR processes and state space models) and Gaussian processes to make predictions of clinical laboratory test results. To account for variability in collections of functions, a number of authors have proposed variants of GPs that account for variability in the mean function (e.g. Lázaro-Gredilla et al. 2012; Shi et al. 2012) and the covariance function (e.g. Shi et al. 2005). Another related line of work in the FDA literature is function-to-function regression (e.g., Oliva et al. 2015). In most approaches to function-to-function regression (FFR) the input and output are defined on fixed domains. In contrast, our problem requires updated predictions as the clinical history continues to grow; both the input and output domains are therefore constantly changing.

Most related to our work is that by Rizopoulos (2011), where the focus is on making dynamical predictions about a time-to-event outcome (e.g. time until death) using all previously observed values of a longitudinally recorded marker. As more data is collected, they dynamically update posterior distributions over individual-specific longitudinal model parameters (as is done in FDA), which serve as time-varying features for the time-to-event prediction. Proust-Lima et al. (2014) tackles the same task but uses a mixture of trajectories to model longitudinal data. As more observations are collected, the posterior over a set of classes is updated, each of which has a distinct set of time-to-event model parameters. These are both state-of-the-art models for the task of dynamical disease trajectory prediction; we will revisit them in our experimental section where we use the approaches as baselines. To scale these models to multivariate time series, however, requires careful specification of the joint model across different markers, which can be challenging in high-dimensional settings (e.g., Dürichen et al. 2015) and may be difficult to scale. For example, Rizopoulos and Ghosh (2011) use a random effects model with a full covariance matrix to describe dependencies across markers, which scales quadratically in the number of marker types (as opposed to linearly as is the case for C-LTM). Because C-LTM is discriminatively trained (we optimize the likelihood of future target trajectories given target and auxiliary marker histories), it is less sensitive to misspecification of the dependencies across markers.

3. Coupled Latent Trajectory Model

Our goal is to predict a *continuous function* modeling the future trajectory of a *target clinical marker* (e.g. PFVC) that tracks disease progression in a specific organ. To make our predictions, we will use a collection of baseline (i.e. static) markers measured when an individual first visits the clinic, the previously observed values of the target marker, and the previously observed values of a collection of *auxiliary clinical markers* tracking related organ systems. See Figure 6a-d for example applications. In these figures, the posterior

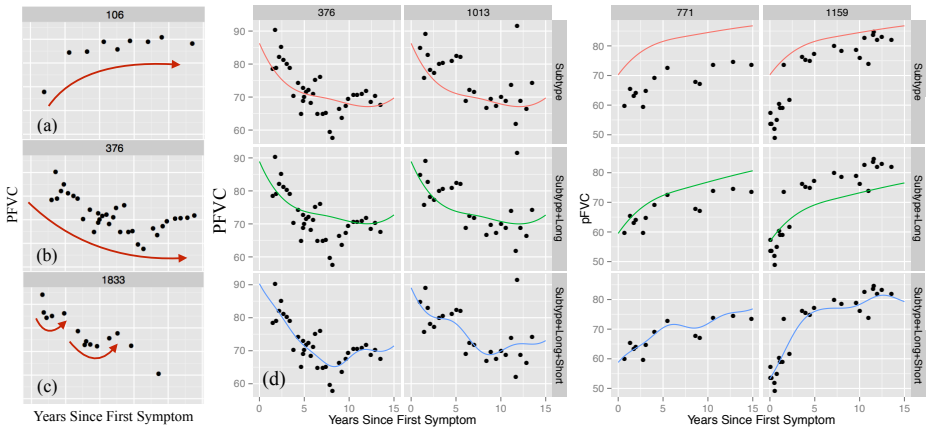


Figure 1: Plots (a-c) show example marker trajectories. Plot (d) shows four individuals with adjustments to a population and subpopulation fit (row 1). Row 2 makes an individual-specific long-term adjustment. Row 3 makes individual-specific short-term adjustments. To simplify, we only show mean functions; posterior uncertainty intervals are omitted.

Input: Baseline predictors $\vec{x}_{1:N}$

Target marker trajectories $\vec{y}_{1:N}$

Auxiliary marker trajectories $\vec{y}_{1:C,1:N}$

Stage 1: Fit Latent Trajectory Models (Model in Sec. 3.2, learning in Sec. 3.5)

Fit target marker LTM to define factor $\ell_{i,\leq t}(\cdot)$

For $c \in \{1, \ldots, C\}$

Fit auxiliary marker LTM to define factor $\ell_{c,i,\leq t}(\cdot)$

Stage 2: Fit Coupling Model (Model in Sec. 3.3 & 3.4, learning in Sec. 3.5)

For each regularization penalty λ

Fit CRF on training data:

Initialize singleton factor weights for $\phi(\cdot)$, $\ell_{i,\leq t}(\cdot)$, and $\ell_{c,i,\leq t}(\cdot)$ for $c \in \{1, \ldots, C\}$

Initialize pairwise factor weights for $\psi(\cdot, \cdot)$

Fit CRF parameters by maximizing

$$J(\gamma, \gamma_{1:C}, \theta, \theta_{1:C}, \eta_{1:C}) \triangleq \sum_{i=1}^{M} \sum_{t \in \mathcal{T}} \log p(\vec{y}_{i,>t} \mid \vec{y}_{i,\leq t}, \vec{y}_{1:C,i,\leq t}, \vec{x}_i) - \lambda(\|\gamma, \gamma_{1:C}\|_1 + \|\theta, \theta_{1:C}\|_1 + \|\eta_{1:C}\|_1)$$

Evaluate penalty on development data

Refit CRF using full training set with best λ

Figure 2: Two-stage procedure for fitting the Coupled Latent Trajectory Model (C-LTM).

distribution over the PFVC values (blue and green shaded regions) are conditioned upon baseline markers (e.g. gender and race), the observed PFVC values (black points), and auxiliary marker histories (e.g. TSS). We learn our model from a database of clinical histories of individuals, which are comprised of the individuals' baseline information and irregularly sampled trajectories of both the target and auxiliary markers. Formally, our

model will estimate the following conditional distribution (notation is described in the subsequent paragraph):

$$\mathcal{D}(i,t) \triangleq p(\mathbf{y}_i(\cdot) \mid \vec{y}_{i,\leq t}, \vec{y}_{1:C,i,\leq t}, \vec{x}_i). \tag{1}$$

Notation. For an individual i, we denote each target marker observation using y_{ij} and its measurement time using t_{ij} where $j \in \{1, \ldots, N_i\}$. We use $\vec{y}_i \in \mathbb{R}^{N_i}$ and $\vec{t}_i \in \mathbb{R}^{N_i}$ to denote all of individual i's marker values and measurement times respectively. We assume that the target marker observations are noisy observations of a latent continuous-time function (the trajectory), which we denote using $\mathbf{y}_i(\cdot)$. Each individual has baseline (static) information collected into a vector, which we denote using $\vec{x}_i$. We use C to denote the number of auxiliary marker types, N_{ci} to denote the number of observations of the c^{th} type, and use y_{cij} and t_{cij} to denote individual i's j^{th} measurement of marker type c. We use $\vec{y}_{ci} \in \mathbb{R}^{N_{ci}}$ and $\vec{t}_{ci} \in \mathbb{R}^{N_{ci}}$ to denote the vector containing all of individual i's c^{th} marker values and times respectively. We will also frequently need to refer to the vector of marker values observed up until a time t, which we denote using $\vec{y}_{i,\leq t}$ ($\vec{y}_{ci,\leq t}$ for auxiliary markers). Similarly, for markers observed after a time t, we use $\vec{y}_{i,>t}$ ($\vec{y}_{ci,>t}$ for auxiliary markers). The term $\vec{y}_{1:C,i,\leq t}$ refers to all auxiliary markers measured on individual i up until time t.

At a high-level, we will model Eq. 1 by first assuming that each clinical marker trajectory (both target and auxiliary) can be d-separated (rendered conditionally independent) of all other marker types given a marker type-specific latent variable. We denote these latent variables using z_i for the target marker and z_{ci} for auxiliary marker c, and will describe them further later in this section. Under this assumption, we can write Eq. 1 as

$$\begin{aligned}
\mathcal{D}(i,t) &= \sum_{z_i} p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i) p(z_i \mid \vec{y}_{i,\leq t}, \vec{y}_{1:C,i,\leq t}, \vec{x}_i) \\
&\propto \sum_{z_i} p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i) p(\vec{y}_{i,\leq t} \mid z_i, \vec{x}_i) p(z_i \mid \vec{y}_{1:C,i,\leq t}, \vec{x}_i) \\
&\propto \sum_{z_i} \underbrace{p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i)}_{\substack{\text{LTM predictive,} \\ \text{Section 3.2.2,} \\ \text{Eq. 24}}} \underbrace{p(\vec{y}_{i,\leq t} \mid z_i, \vec{x}_i)}_{\substack{\text{LTM likelihood,} \\ \text{Section 3.2.1,} \\ \text{Eq. 16}}} \sum_{z_{1:C,i}} \underbrace{p(z_i, z_{1:C,i} \mid \vec{x}_i)}_{\substack{\text{Coupling Model,} \\ \text{Section 3.3,} \\ \text{Eq. 25}}} \prod_{c=1}^{C} \underbrace{p(\vec{y}_{ci,\leq t} \mid z_{ci}, \vec{x}_i)}_{\substack{\text{LTM likelihood,} \\ \text{Section 3.2.1,} \\ \text{Eq. 16}}}.
\end{aligned} \tag{2}$$

We will learn this parameterization of $\mathcal{D}(i,t)$ in two stages. The models for the target and each of the auxiliary markers are learned independently during the first stage; using these, the LTM predictive and likelihood terms can be computed in Eq. 2. We treat the target and auxiliary markers as instances of the Latent Trajectory Model (LTM), which we review in Section 3.2. We emphasize, however, that any other generative model can be used if better suited to the domain. The coupling model is learned in the second stage, and is described in Section 3.3. We refer to the model created by combining these components as the Coupled Latent Trajectory Model (C-LTM), which we describe in Section 3.4. An overview of the procedure used to fit the C-LTM is shown in Figure 2.

3.1 Preliminaries

The Latent Trajectory Model (LTM) uses B-splines to model longitudinal trajectories, and our coupling model uses a conditional random field (CRF). We briefly introduce these two concepts, and point to resources where the interested reader can find additional details.

3.1.1 B-Splines

A common approach to fitting nonlinear functions of time while maintaining a linear dependence on model parameters is to use a basis expansion. Such an expansion defines some non-linear function $f(t)$ as a linear combination of other functions $\phi_1(t), \ldots, \phi_d(t)$:

$$y = f(t \mid \beta) = \sum_{i=1}^{d} \beta_d \phi_d(t) = \Phi^\top(t)\vec{\beta}, \tag{3}$$

where $\phi_1, \ldots, \phi_d$ act as bases in some vector space of nonlinear functions and $\Phi(t) \in \mathbb{R}^d$ is the vector containing the values of the p basis functions evaluated at time t. The benefit of this formulation is that the function f is linear in the model parameters β, making it relatively easy to fit complex models. B-splines are a particular family of basis functions that we can use to parameterize nonlinear functions. Others include polynomial bases and radial basis functions. However, there are two advantages to using B-splines. First, each basis function is non-zero only over a compact interval of the real line, which improves statistical stability and also allows for computational speed ups that take advantage of sparse basis matrices (Gelman et al., 2014). This is in contrast to polynomials, where each basis takes non-zero values globally. The second advantage is that the family of functions parameterized by B-splines are not infinitely differentiable (in contrast to radial basis functions) and therefore not smooth (Gelman et al., 2014). This bias is often helpful in modeling functions from the real-world that arise from non-smooth processes. Because B-splines are linear in their parameters, we can use the well-developed machinery of linear regression for learning. See Ch. 20 in Gelman et al. (2014) or Ch. 5 in Friedman et al. (2001) for further details.

Penalized B-splines. In practice, the parameters of a B-spline model are fit using a penalized least squares criterion. The penalty is typically introduced in order to control the smoothness of the fit. For data $\vec{y}$ measured at times $\vec{t}$ with corresponding basis matrix $\Phi(\vec{t}) = [\Phi(t_1), \ldots, \Phi(t_n)]^\top$, we minimize the following objective:

$$J(\vec{\beta}) = \|\vec{y} - \Phi(\vec{t})\vec{\beta}\|_2^2 + \rho \vec{\beta}^\top \Omega \vec{\beta}, \tag{4}$$

where Ω is a first-order differences matrix as described by Eilers and Marx (1996). The penalized objective is still quadratic in $\vec{\beta}$ and so can be easily minimized.

3.1.2 Conditional Random Fields

Conditional random fields (CRFs) provide a framework for modeling and learning the joint distribution of a collection of random variables conditioned on some set of observations (see e.g., Murphy 2012). The parameterization is identical to that of Markov random fields (MRF), but the factors that define the distribution can be functions of the observations (this allows the distribution to vary depending on the values of the observations). For some

output y, input x and parameters θ, the conditional probability is defined to be:

$$p(y \mid x, \theta) = \frac{1}{Z(x,\theta)} \prod_c \psi_c(y_c \mid x, \theta), \;\; Z(x,\theta) \triangleq \sum_{y'} \prod_c \psi_c(y'_c \mid x, \theta), \tag{5}$$

where $\psi_c(y_c \mid x, \theta)$ is a non-negative factor that can be interpreted as scoring the configuration of the subset of variables y_c given the observations x and parameters θ. The term $Z(x, \theta)$ is called the *partition function* and ensures that the distribution is normalized. When we can write

$$\log \psi_c(y_c \mid x, \theta) = \theta_c^\top f_c(y_c, x) \iff \psi_c(y_c \mid x, \theta) = \exp\left\{\theta_c^\top f_c(y_c, x)\right\}, \tag{6}$$

where f_c extracts some vector of features from the observations x and the target y_c, then we say that the CRF is a log-linear model. Log-linear models have a number of desirable properties, the most relevant to this work being the ease with which we can differentiate the log-likelihood with respect to model parameters. To compute the derivative with respect to θ_c (the parameters corresponding to the c^{th} factor) we have:

$$\frac{\partial \log p(y \mid x, \theta)}{\partial \theta_c} = f_c(y_c, x) - \frac{\partial \log Z(x, \theta)}{\partial \theta_c}. \tag{7}$$

To compute the partial derivative in the second term on the RHS, first note that

$$\frac{\partial Z(x,\theta)}{\partial \theta_c} = \sum_{y'} \left(\prod_{d \neq c} \psi_d(y'_d \mid x, \theta_d \right) \frac{\partial \psi_c(y'_c \mid x, \theta_c)}{\partial \theta_c} \tag{8}$$

$$= \sum_{y'} \left(\prod_{d \neq c} \psi_d(y'_d \mid x, \theta_d \right) \psi_c(y'_c \mid x, \theta_d) f_c(y'_c, x). \tag{9}$$

This implies that the partial derivative of $\log Z(x, \theta)$ is simply:

$$\frac{\partial \log Z(x,\theta)}{\partial \theta_c} = \frac{1}{Z(x,\theta)} \frac{\partial Z(x,\theta)}{\partial \theta_c} = \mathbb{E}_y \left[f_c(y_c, x) \mid x \right] \tag{10}$$

This means that the gradient of the log-likelihood with respect to a set of parameters θ_c is the difference between the observed features $f_c(y, x)$ and their expectation under the current set of parameters θ. To learn the weights, we can apply gradient-based algorithms to optimize the likelihood of a set of observed training input-output pairs. In addition, a regularizer is often added to the objective to discourage complexity or induce sparsity. We will use these ideas in the derivation of our learning algorithm. See Ch. 19 in Murphy (2012) for further details.

3.2 Latent Trajectory Model

The Latent Trajectory Model (LTM) is a probabilistic model introduced by Schulam and Saria (2015) for obtaining individualized predictions of a clinical marker trajectory in populations with diverse disease expression. LTM posits that the measured markers are

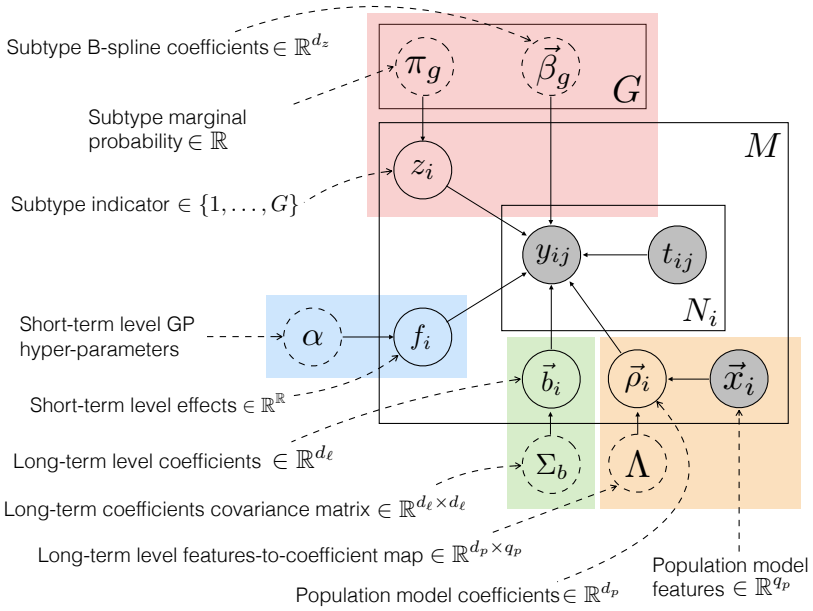


Figure 3: The LTM graphical model. Levels in the hierarchy are color-coded. Model parameters are enclosed in dashed circles. Observed variables are shaded.

noisy observations of the underlying disease activity trajectory, which is a function of both individual-specific parameters and parameters that are shared across other individuals in the population. Moreover, the LTM uses individual-specific latent factors to explain differences in trajectories across the population that are not explained by observed features alone. The LTM graphical model is shown in Figure 3. We review the LTM below using the same notation defined at the beginning of this section. In addition, we use $\Phi(t_{ij})$ to denote a column-vector containing a basis expansion of the time t_{ij} and $\Phi\left(\vec{t}_i\right) = [\Phi(t_{i1}), \ldots, \Phi(t_{iN_i})]^\top$ to denote the matrix containing the basis expansion of points in $\vec{t}_i$ in each of its rows.

The LTM models the jth marker value for individual i as a normally distributed random variable with a mean assumed to be the sum of four terms: a population component, a subpopulation component, an individual long-term component, and an individual short-term component.

$$y_{ij}|z_i, \vec{b}_i, f_i \sim \mathcal{N}\left(\underbrace{\Phi_p(t_{ij})^\top \Lambda\, \vec{x}_i}_{\text{(A) population}} + \underbrace{\Phi_z(t_{ij})^\top \vec{\beta}_{z_i}}_{\text{(B) subpopulation}} + \underbrace{\Phi_\ell(t_{ij})^\top \vec{b}_i}_{\text{(C) ind. long-term}} + \underbrace{f_i(t_{ij})}_{\text{(D) ind. short-term}}, \sigma^2\right). \tag{11}$$

The four terms in the sum serve two purposes. First, and most importantly, they allow for a number of different sources of variation — both observed and latent — to influence the observed marker value, which allows for heterogeneity both across and within individuals. Second, they share statistical strength across different subsets of observations. The population component shares strength across all observations. The subpopulation component shares strength across observations belonging to subgroups of individuals. The individual

long-term component shares strength across all observations belonging to the same individual. Finally, the individual short-term component shares information across observations belonging to the same individual that are measured at similar times. Predicting an individual's trajectory involves estimating her subtype and individual-specific parameters as new clinical data becomes available. We briefly review each of the components here for ease of presentation, but refer the interested reader to Schulam and Saria (2015) for further details.

Population level. The population model predicts aspects of an individual's disease activity trajectory using *observed* baseline characteristics (e.g. gender and race), which are represented using the feature vector $\vec{x}_i$. This sub-model is shown within the orange box in Figure 3. The predicted value of the jth marker of individual i measured at time t_{ij} is shown in Eq. 11 (A), where $\Phi_p(t) \in \mathbb{R}^{d_p}$ is a basis expansion of the observation time and $\Lambda \in \mathbb{R}^{d_p \times q_p}$ is a matrix used as a linear map from an individual's covariates $\vec{x}_i$ to coefficients $\rho_i \in \mathbb{R}^{d_p}$. At this level, individuals with similar covariates will have similar coefficients.

Subpopulation level. LTM models an individual's subtype using a discrete-valued latent variable $z_i \in \{1, \ldots, G\}$, where G is the number of subtypes. z_i is a multinomially distributed random variable with parameters $\pi_{1:G} \triangleq [\pi_1, \ldots, \pi_G]$ where $\pi_g \geq 0$ and $\sum_g \pi_g = 1$. Each subtype has a unique disease activity trajectory represented using B-splines, where the number and location of the knots and the degree of the polynomial pieces are fixed prior to learning. These hyper-parameters determine a basis expansion $\Phi_z(t) \in \mathbb{R}^{d_z}$ mapping a time t to the B-spline basis function values at that time. Trajectories for each subtype are parameterized by a vector of coefficients $\vec{\beta}_g \in \mathbb{R}^{d_z}$ for $g \in \{1, \ldots, G\}$, which are learned offline. Under subtype z_i, the predicted value of marker y_{ij} measured at time t_{ij} is shown in Eq. 11 (B). This component explains differences such as those observed between the trajectories in Figures 1a and 1b.

Individual long-term level. The individual long-term component is parameterized using a linear model with basis expansion $\Phi_\ell(t) \in \mathbb{R}^{d_\ell}$ and individual-specific coefficients $\vec{b}_i \in \mathbb{R}^{d_\ell}$. This level models deviations from the population and subpopulation models using parameters that are learned dynamically as the individual's clinical history grows. An individual's coefficients are modeled as latent variables with marginal distribution $\vec{b}_i \sim \mathcal{N}(\vec{0}, \Sigma_b)$. For individual i, the predicted value of marker y_{ij} measured at time t_{ij} is shown in Eq. 11 (C). This component can explain, for example, differences in overall health due to an unobserved characteristic such as chronic smoking, which may cause atypically lower lung function than what is predicted by the population and subpopulation components. Such an adjustment is illustrated across the first and second rows of Figure 1d.

Individual short-term level. Finally, the individual short-term component f_i captures transient trends in an individual's marker sequence that do not generalize outside of a small time window. For example, an infection may cause an individual's lung function to temporarily appear more restricted than it actually is, which may cause short-term trends like those shown in Figure 1c and the third row of Figure 1d. We treat f_i as a function-valued latent variable and model it using a Gaussian process with zero-valued mean function and Ornstein-Uhlenbeck (OU) covariance function

$$K_{\text{OU}}(t_1, t_2) = a^2 \exp\left\{-\ell^{-1}|t_1 - t_2|\right\}. \tag{12}$$

The amplitude a controls the magnitude of the structured noise that we expect to see and the length-scale ℓ controls the length of time over which we expect these temporary trends

to occur. The OU kernel is ideal for modeling such deviations as it is both mean-reverting and draws from the corresponding stochastic process are only first-order continuous, which eliminates long-range dependencies between deviations (Rasmussen and Williams, 2006). Applications in other domains may require different kernel structures motivated by properties of transient deviations in the trajectories.

Accounting for treatments. Several interventions are common in scleroderma, but none have been proven to significantly alter the long-term course of the disease. For example, steroids are commonly administered, but there have been no randomized controlled trials confirming its effects on patients with scleroderma-related lung disease—see, for example, Ch. 35 in Varga et al. (2012). Immunosuppressants are also commonly used to treat scleroderma-related lung disease, but the proven effects are modest and have only been demonstrated over the course of one year (Tashkin et al., 2006). We assume that these types of transient interventions are well-modeled by the individual-specific short-term component, and so we do not explicitly model the treatment effects of steroids or immunosuppressants in our data. Others have developed methods for estimating treatment effects from observational time series (e.g., Chib and Hamilton 2002; Kleinberg and Hripcsak 2011; Brodersen et al. 2015). More recently, see Xu et al. (2016) for an application using functional data. Treatment effects can be incorporated within the trajectory likelihood in diseases where treatments are suspected to alter long term trajectory. We leave this more general case as a direction for future work.

Missing data mechanism. The LTM assumes observations of the trajectory are missing at random (MAR). This implies that we can use maximum likelihood estimation without needing to incorporate additional information about the sampling model; see Appendix B. When the data are missing not at random, assumptions about the missing data mechanism should be explicated and incorporated within the individual marker models.

In summary, the latent, individual-specific factors in the model (z_i, $\vec{b}_i$, and f_i from Eq. 11B, 11C, and 11D respectively) each contribute to describe the observed trajectory at different granularities. These are all treated as random variables and marginalized out during learning to avoid overfitting. When making predictions, we can use an individual's observed data to compute posterior distributions over these latent factors, which allows us to tailor predictions.

3.2.1 LTM Likelihood

Given parameters $\Theta = \{\Lambda, \pi_{1:G}, \vec{\beta}_{1:G}, \Sigma_b, a, \ell, \sigma^2\}$, we can compute the observed-data likelihood of a given clinical marker trajectory by marginalizing z_i, $\vec{b}_i$ and f_i out of the joint distribution defined by our model:

$$
\begin{aligned}
&p\left(\vec{y}_i \mid \vec{x}_i, \Theta\right) \\
&\quad = \sum_{z_i=1}^{G} \underbrace{p\left(z_i \mid \Theta\right)}_{\text{Multinomial prior}} \int_{\mathbb{R}^{d_\ell}} \underbrace{p\left(\vec{b}_i \mid \Theta\right)}_{\text{Normal prior}} \int_{\mathbb{R}^{N_i}} \underbrace{p\left(f_i \mid \Theta\right)}_{\text{GP prior}} \underbrace{p\left(\vec{y}_i \mid z_i, \vec{b}_i, f_i, \vec{x}_i, \vec{t}_i, \Theta\right)}_{\text{Eq. 11}} df_i \, d\vec{b}_i \qquad (13) \\
&\quad = \sum_{z_i=1}^{G} \pi_{z_i} \mathcal{N}\left(\vec{y}_i \mid \Phi_p\left(\vec{t}_i\right) \Lambda \, \vec{x}_i + \Phi_z\left(\vec{t}_i\right) \vec{\beta}_{z_i}, K\left(\vec{t}_i, \vec{t}_i\right)\right). \qquad (14)
\end{aligned}
$$

Moving from Eq. 13 to Eq. 14, we evaluate the innermost integral using the fact that the GP prior over f_i is conjugate to Eq. 11 yielding a new multivariate normal (Rasmussen and Williams, 2006). To evaluate the next integral in Eq. 13, we again have that the normal prior over $\vec{b}_i$ is conjugate to the multivariate normal obtained by marginalizing over f_i, which gives us the multivariate normal shown in Eq. 14 where the covariance function is defined as

$$K\left(t_1, t_2\right) = \Phi_\ell\left(t_1\right)^\top \Sigma_b \Phi_\ell\left(t_2\right) + K_{\text{OU}}\left(t_1, t_2\right) + \sigma^2 \mathbb{I}\left(t_1 = t_2\right). \tag{15}$$

We see that the observed-data likelihood for individual i is defined by a mixture of multivariate normals where each subtype is associated with a class in the mixture. The mixing probabilities are defined by the multinomial over subtypes. The mean of the multivariate normal is defined by the population and subpopulation models, and the covariance is defined by the individual long-term and short-term components of the model. To obtain the LTM likelihood needed in Eq. 2, we will condition Eq. 14 on subtype z_i. This gives us the following expression:

$$p(\vec{y}_i \mid z_i, \vec{x}_i) \triangleq \mathcal{N}\left(\vec{y}_i \mid \Phi_p\left(\vec{t}_i\right) \Lambda \vec{x}_i + \Phi_z\left(\vec{t}_i\right) \vec{\beta}_{z_i}, K\left(\vec{t}_i, \vec{t}_i\right)\right). \tag{16}$$

3.2.2 LTM Predictive

As presented, the LTM can be easily applied to the task of disease activity trajectory prediction. Note that the LTM provides a posterior distribution over the trajectory using baseline markers and measurements of the target marker (e.g. PFVC) as they are recorded. It does not incorporate information from other time-varying markers such as TSS and PDLCO. Suppose we have estimates of the model parameters $\Theta = \{\Lambda, \pi_{1:G}, \vec{\beta}_{1:G}, \Sigma_b, a, \ell, \sigma^2\}$, then we can predict an individual's future course by computing the posterior predictive distribution $p(\vec{y}_{i,>t} \mid \vec{y}_{i,\leq t}, \vec{x}_i)$, where $\vec{y}_{i,>t}$ denotes marker values after time t and $\vec{y}_{i,\leq t}$ denotes marker values observed prior to time t. To compute the expected marker value at time t_i^*, we evaluate the following expression:

$$\hat{\mathbf{y}}\left(t_i^*\right) = \sum_{z_i=1}^{G} \int_{\mathbb{R}^{d_\ell}} \int_{\mathbb{R}^{N_i}} \underbrace{\mathbb{E}\left[y_i^* \mid z_i, \vec{b}_i, f_i\right]}_{\text{prediction given latent vars.}} \underbrace{p\left(z_i, \vec{b}_i, f_i \mid \vec{y}_{i,\leq t}, x_{ip}, \Theta\right)}_{\text{posterior over latent vars.}} df_i \, d\vec{b}_i \tag{17}$$

$$= \mathbb{E}^*_{z_i, \vec{b}_i, f_i}\left[\Phi_p\left(t_i^*\right)^\top \Lambda \vec{x}_i + \Phi_z\left(t_i^*\right)^\top \vec{\beta}_{z_i} + \Phi_\ell\left(t_i^*\right)^\top \vec{b}_i + f_i\left(t_i^*\right)\right] \tag{18}$$

$$= \underbrace{\Phi_p\left(t_i^*\right)^\top \Lambda \vec{x}_i}_{\text{pop. prediction}} + \underbrace{\Phi_z\left(t_i^*\right)^\top \overbrace{\mathbb{E}^*_{z_i}\left[\vec{\beta}_{z_i}\right]}^{\vec{\beta}_i^*}}_{\text{subpop. prediction}} + \underbrace{\Phi_\ell\left(t_i^*\right)^\top \overbrace{\mathbb{E}^*_{\vec{b}_i}\left[\vec{b}_i\right]}^{\vec{b}_i^*}}_{\text{ind. long prediction}} + \underbrace{\overbrace{\mathbb{E}^*_{f_i}\left[f_i\left(t_i^*\right)\right]}^{f_i^*(t_i^*)}}_{\text{ind. short prediction}}, \tag{19}$$

where E^* denotes an expectation conditioned on $\vec{y}_{i,\leq t}, x_i, \Theta$. We see that the prediction takes a natural form; we compute the value of the individual's disease activity trajectory at the future time point by replacing the latent factors with their posterior expectations. Computing the population prediction is straightforward as all quantities are observed. To compute the subpopulation prediction, we need to compute the marginal posterior over z_i,

which is easily done given that the observed-data likelihood has a mixture of multivariate normals density (Eq. 14). The posterior probability of subtype g for individual i is therefore

$$\pi^*_{ig} \propto \pi_g \, \mathcal{N}\left(\vec{y}_i \mid \Phi_p\left(\vec{t}_i\right) \Lambda \, \vec{x}_i + \Phi_z\left(\vec{t}_i\right) \vec{\beta}_g, K\left(\vec{t}_i, \vec{t}_i\right)\right). \tag{20}$$

To compute the subpopulation prediction in Eq. 19 above, we simply compute the expected value of the B-spline coefficients under the posterior in Eq. 20:

$$\vec{\beta}^*_i \triangleq \left(\textstyle\sum_{z_i=1}^{G} \pi^*_{iz_i} \vec{\beta}_{z_i}\right). \tag{21}$$

The expectation required for the individual long-term prediction is:

$$\vec{b}^*_i \triangleq \left[\Sigma_b^{-1} + \Phi_\ell(\vec{t}_i)^\top K_f^{-1} \Phi_\ell(\vec{t}_i)\right]^{-1} \left[\Phi_\ell(\vec{t}_i)^\top K_f^{-1} \left(\vec{y}_i - \Phi_p^\top(\vec{t}_i) \Lambda \, \vec{x}_i - \Phi_z^\top(\vec{t}_i) \vec{\beta}^*_i\right)\right]. \tag{22}$$

Finally, the expectation required for the individual short-term prediction is:

$$\begin{aligned} f^*(t^*_i) \triangleq K_{\mathrm{OU}}(t^*_i, \vec{t}_i) \left[K_{\mathrm{OU}}(\vec{t}_i, \vec{t}_i) + \sigma^2 \boldsymbol{I}\right]^{-1} \vec{r}_i \\ \text{where } \vec{r}_i = \left(\vec{y}_i - \Phi_p^\top(\vec{t}_i) \Lambda \, \vec{x}_i - \Phi_z^\top(\vec{t}_i) \vec{\beta}^*_i - \Phi_\ell^\top(\vec{t}_i) \vec{b}^*_i\right) \end{aligned} \tag{23}$$

For brevity, we omit the derivation of these expectations here, but point the interested reader to Schulam and Saria (2015) and its supplementary material for the steps taken to arrive at these expressions. To obtain the LTM predictive needed in Eq. 2, we condition Eq 19 on subtype z_i. This gives us the following expression:

$$p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i) \triangleq \Phi_p(\cdot)^\top \Lambda \, \vec{x}_i + \Phi_z(\cdot)^\top \vec{\beta}_{z_i} + \Phi_\ell(\cdot)^\top \mathbb{E}^*_{\vec{b}_i}\left[\vec{b}_i\right] + \mathbb{E}^*_{f_i}\left[f_i(\cdot)\right]. \tag{24}$$

3.3 Coupling Model

The Coupled Latent Trajectory Model (C-LTM) seeks to learn and capture correlations across trajectories of different marker types. In scleroderma, for example, an individual with worse lung trajectories (e.g. the rapidly declining lung trajectory subtype) is more likely to have a severe skin disease trajectory. In the C-LTM these types of dependencies are captured by the term $p(z_i, z_{1:C,i} \mid x_i)$ shown in Eq. 2. We parameterize this distribution using a conditional random field with singleton and pairwise factors defined over z_i and $z_{1:C,i}$. Singleton factors can depend on the baseline covariates $\vec{x}_i$. Pairwise factors are defined only between the clinical marker random variables z_i and each of the auxiliary marker latent variables z_{ci}. Both are parameterized linearly. The coupling model therefore has the following form:

$$\begin{aligned} \log p(z_i, z_{1:C,i} \mid \vec{x}_i) &\propto \phi(z_i, \vec{x}_i) + \sum_{c=1}^{C} \phi(z_{ci}, \vec{x}_i) + \psi(z_i, z_{ci}) \\ &= \theta^\top f(z_i, \vec{x}_i) + \sum_{c=1}^{C} \theta_c^\top f_c(z_{ci}, \vec{x}_i) + \eta_c^\top g_c(z_i, z_{ci}). \end{aligned} \tag{25}$$

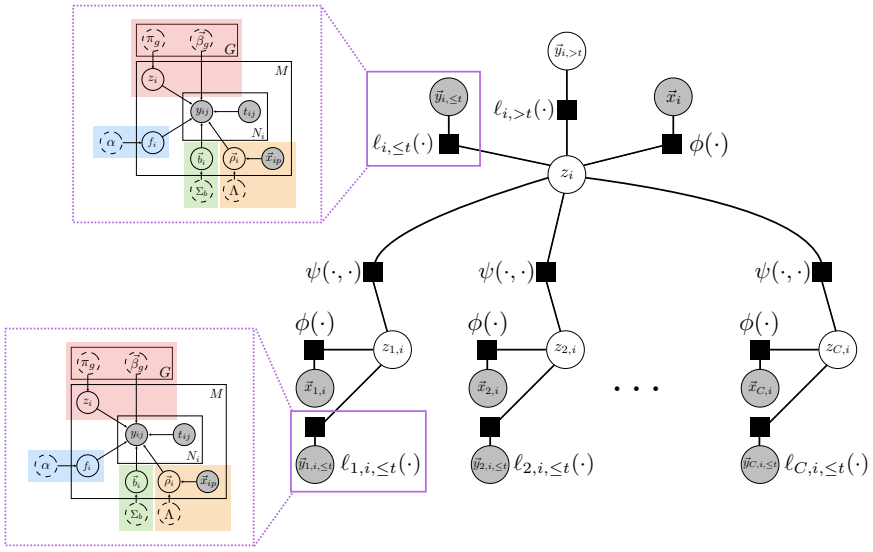


Figure 4: The factor graph of the coupled latent trajectory model. Empty nodes denote latent random variables, and shaded nodes denote observed variables. The latent trajectory model (LTM, described in Section 3.2) acts as a data-driven factor linking observed target and auxiliary marker histories into predictions.

3.4 Predicting Trajectories using the C-LTM

To predict trajectories (i.e. compute Eq. 1), we combine the LTM likelihood (Eq. 16), the LTM predictive (Eq. 24), and the coupling model (Eq. 25). Let $\ell_{i,\leq t}(z_i)$ stand as shorthand for $\log p(\vec{y}_{i,\leq t} \mid z_i, \vec{x}_i)$ and $\ell_{ci,\leq t}(z_{c,i})$ stand as shorthand for $\log p(\vec{y}_{ci,\leq t} \mid z_{ci}, \vec{x}_i)$, then we see that

$$\mathcal{D}(i,t) \propto \sum_{z_i} p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i) \sum_{z_{1:C,i}} u(z_i, z_{1:C,i} \mid \mathcal{H}(i,t)), \tag{26}$$

where we have defined $\mathcal{H}(i,t)$ to be the set of information contained in the clinical history of individual i at time t: $\{\vec{y}_{i,\leq t}, \vec{y}_{1:C,i,\leq t}, \vec{x}_i\}$, and used $u(z_i, z_{1:C,i} \mid \mathcal{H}(i,t))$ to denote the following unnormalized weight assigned to all values of the latent variables given the history:

$$u(z_i, z_{1:C,i} \mid \mathcal{H}(i,t)) \triangleq \exp\left\{ \ell_{i,\leq t}(z_i) + \theta^\top f(z_i, \vec{x}_i) + \sum_{c=1}^{C} \ell_{ci,\leq t}(z_{ci}) + \theta_c^\top f_c(z_{ci}, \vec{x}_i) + \eta_c^\top g_c(z_i, z_{ci}) \right\}, \tag{27}$$

To make $\mathcal{D}(i,t)$ a proper distribution, we normalize $u(z_i, z_{1:C,i} \mid \mathcal{H}(i,t))$ to obtain

$$p(z_i \mid \mathcal{H}(i,t)) = \frac{\sum_{z_{1:C}} u(z_i, z_{1:C} \mid \mathcal{H}(i,t))}{\sum_z \sum_{z_{1:C}} u(z, z_{1:C} \mid \mathcal{H}(i,t))} \triangleq \frac{Z'_{i,t}(z_i)}{Z_{i,t}}. \tag{28}$$

then we can write $\mathcal{D}(i,t)$ (Eq. 1) as

$$\mathcal{D}(i,t) = \sum_{z_i} p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i)\, p(z_i \mid \mathcal{H}(i,t)). \tag{29}$$

Intuitively, we see that the predictive distribution under C-LTM is simply a weighted combination of the subtype-specific predictive distributions under LTM (Eq. 24). Moreover, the distribution $p(z_i \mid \mathcal{H}(i,t))$ is the marginal distribution over z_i in a conditional random field with structure similar to the coupling model (Eq. 25) but augmented with additional singleton factors defined by the LTM likelihood functions given the marker trajectory histories. The LTM likelihood factors in Eq. 27 are added into the model unchanged, but additional parameters $\{\gamma, \gamma_{1:C}\}$ can be included to reweight those terms (a similar idea is used in Raina et al. (2003)).[1] The factor graph for this conditional random field is shown in Figure 4. Note that the weight $p(z_i \mid \mathcal{H}(i,t))$ can be efficiently computed in time linear in the number of auxiliary markers using the junction tree algorithm.

The C-LTM offers a number of advantages for predictive modeling of disease trajectories in domains where many other related marker trajectories are available. First, it allows irregularly and sparsely sampled trajectories to be neatly summarized using modularized, single-marker generative models. These can capture important latent factors and account for marker-specific measurement models and noise processes. Second, we can discriminatively use auxiliary marker trajectory histories when modeling Eq. 1 instead of specifying a joint generative model, which sidesteps the challenges associated with correctly specifying dependencies between many different marker types. Finally, the model can be used in continuous time and it dynamically updates predictions as new observations arrive.

3.5 Learning the C-LTM

We have described two components of our approach: the Latent Trajectory Model (LTM) and the coupling model. When these components are combined as shown in Section 3.4, then we obtain the C-LTM. The C-LTM has two conceptually distinct sets of parameters. The first set are those belonging to the individually trained LTMs for each marker type. To learn these, we can use the EM algorithm described in Schulam and Saria (2015). To learn the parameters for the C-LTM, we keep the single-marker model parameters fixed (e.g. those learned for the LTM), and use a standard gradient-based CRF learning algorithm (as described in Section 3.1.2) to optimize the penalized log-likelihood of example trajectory predictions. For completeness, we provide additional details for both stages in Appendix A.

3.5.1 Scalability

The EM algorithm used to learn the parameters of the LTM poses no serious challenges to scalability. The primary computational burden lies in the E-step wherein sufficient statistics from all individuals are computed and collected. This is linear in the number of patient records being analyzed, but since the inference required to compute the sufficient statistics can be performed independently for each individual given the current parameter estimates, the E-step can be easily parallelized to offset slow learning due to large numbers

1. When using a penalty, we can center the weights at 1 so that the default behavior is to leave the likelihood factors unchanged as in Eq. 27

of patient records. For any given individual, the E-step is dominated by the inversion of the $N_i \times N_i$ covariance matrix. We do not expect this to be problematic, however, because clinical markers in chronic diseases are observed at a maximum rate of 12 times per year. Moreover, such diseases occur over periods on the order of tens of years. Therefore, the number of measurements will be at most on the order of 100-200.

Learning the parameters of the CRF requires a sweep through all $M|\mathcal{T}|$ training instances in order to compute and aggregate the gradient at each iteration. The primary computational burden is computing the expected values of the features (Eq. 42), however, the tree-structured graphical model shown in Figure 4 allows the junction tree algorithm to run in time linear in the number of auxiliary markers. On a standard laptop, we are able to train the model on 772 patients (5,458 PFVC measurements) in 10-20 minutes.

Online inference for predicting a given individual's future trajectory is also computationally straightforward. The key quantities are (1) the weights $p(z_i \mid \mathcal{H}(i,t))$ in Eq. 29, which are easily computed using the junction tree algorithm in time linear in the number of auxiliary markers, and (2) the subtype-specific predictive densities $p(\mathbf{y}(\cdot) \mid z_i, \vec{y}_{i,\leq t}, \vec{x}_i)$, which have the same computational complexity as the E-step in the LTM learning algorithm.

4. Experiments

We demonstrate our approach by building a tool for predicting lung disease trajectories for individuals with scleroderma. Lung disease is currently the leading cause of death among scleroderma patients, and is notoriously difficult to treat due to the lack of accurate predictors of decline and tremendous variability across individual trajectories (Allanore et al., 2015). Clinicians use percent of predicted forced vital capacity (PFVC) to track lung severity, which is expected to drop as the disease progresses. In addition, they collect demographic information and other clinical marker values that measure the impact of disease on the different organ systems involved in scleroderma.

Data description. We train and validate our model using data from the Johns Hopkins Scleroderma Center patient registry, one of largest collections of clinical scleroderma data in the world. Demographic information is collected during the patient's first visit to the clinic. PFVC and other clinical markers are collected during routine visits thereafter. To select individuals from the registry, we used the following criteria. First, we include individuals who were seen at the clinic within two years of their earliest scleroderma-related symptom[2] ($1,186$ individuals). Second, we exclude all individuals with fewer than two PFVC measurements after first being seen by the clinic (398 individuals). Finally, we exclude individuals who received a lung transplant (16 individuals) because their natural trajectory is altered by the intervention. Transplants are rare so removing patients with transplants should not introduce significant bias. As mentioned earlier, there are no other known course-altering therapies for scleroderma.

Our final data set contains 772 individuals and a total of $5,458$ PFVC measurements tracking individuals over a period of 20 years. The first, second, and third quartiles of the total number of PFVC measurements for an individual are 3, 5, and 9 respectively. The maximum number of PFVC measurements for one individual is 63. The first, second, and third quartiles of the measurement times are 1 year, 2.8 years, and 5.9 years. The first,

2. Date of first symptom is established during the first encounter by both the patient and clinician.

second, and third quartiles of elapsed time between measurements are 0.4 years, 0.7 years, and 1.10 years. The minimum and maximum elapsed time is 0.002 years and 16.4 years respectively.

The baseline demographic information includes gender and African American race, both of which have been shown to be associated with disease severity in scleroderma (Allanore et al., 2015). Antibody data are also collected at baseline, but since these are only available for a small subset of individuals, we do not include that data here. For time-dependent predictors, we include 5 auxiliary clinical markers. Three of the auxiliary markers are similar to PFVC in that they are continuous-valued test results used to measure the health of organ systems. We include: percent of predicted forced expiratory volume in one second (PFEV1), which measures the force with which air is expelled from the lungs; percent of predicted diffusing capacity (PDLCO), which measures the efficiency of oxygen diffusion from the lungs to the bloodstream; and total skin score (TSS), which is a cumulative measure of the thickness of the skin at various points on the body. In addition, we include 2 severity scores—clinical Likert-scaled judgements of organ damage severity: Raynaud's phenomenon (RP) severity score, which measures the severity of damage to the extremities by issues related to the vasculature, and GI severity score that measures the severity of damage to the GI tract. For the interested reader, a more detailed discussion of these markers and their relationship to the disease can be found in Varga et al. (2012).

Experimental setup. For the 4 continuous-valued clinical markers (PFVC, PFEV1, PDLCO, TSS) we use the LTM and for the 2 severity scores (GI and RP) we use a simpler model that we will describe later. For the population model, we use constant functions (i.e. the basis expansion $\Phi_p(t)$ contains an intercept term whose coefficient is determined by baseline covariates). For the subpopulation B-splines, we set boundary knots at 0 and 25 years (the maximum observation time in our data set is 23 years), use two interior knots that divide the time period from 0-25 years into three equally spaced chunks, and use quadratics as the piecewise components. For the individual-specific long-term basis Φ_ℓ, we use the same basis as the population model (constant functions).

We divide our data into 10 folds and use log-likelihood on the first fold for tuning hyperparameters. For PFVC, we select $G = 9$ subtypes using BIC. For the kernel hyperparameters $\Theta_1 = \{\Sigma_b, \alpha, \ell, \sigma^2\}$ we set $\Sigma_b \in \mathbb{R}$ to be 16.0, which corresponds to the variance of individual-specific intercepts. We set $\alpha = 6$, $\ell = 2$, and $\sigma^2 = 1$ using a grid search over values chosen using domain knowledge. Qualitatively, these make sense; we expect transient deviations to last around 2 years and to change PFVC by around ± 6 units. Finally, we penalize the expected log-likelihood with respect to $\vec{\beta}_{1:G}$ as in Eq. 4 and set the weight $\rho = 0.01$, which was chosen based on the clinical interpretability of the learned subtype trajectories. The remaining 9 folds were used for our cross-validation experiments. The parameters of each trajectory model are estimated independently for each fold (e.g. the B-spline coefficients of the subtype trajectories). For the severity scores, which are Likert-scaled and not continuous, we use a simple naive Bayes generative model wherein the latent "class" is an indicator of whether the individual ever reaches a high severity level (a cut-off in the severity scale determined by clinical collaborators). Severity score observations are treated as iid draws from a class-specific multinomial distribution (i.e. the likelihood for these auxiliary markers is a multinomial distribution over severity scores). Finally, we estimate the parameters of the C-LTM by maximizing the objective in Eq. 33 augmented

with an $L1$ regularizer. We optimize the objective using the Orthant-Wise Limited-memory Quasi-Newton (`OWL-QN`) algorithm (Andrew and Gao, 2007). To generate training examples for the C-LTM, we use times $\mathcal{T} = \{1, 2, 4\}$ (the first three quintiles of observation times in our data) to fit three different models. We choose time points earlier in the disease course because this is when it is most valuable to leverage all available information. In our cross-validated experimental results below, we estimate the penalty of the $L1$ regularization term in each fold by splitting a portion of the training data into a development set. We sweep the penalty from 1.0×10^{-7} to 1.0×10^{-1} and choose based on development set performance.

Baselines. As a first baseline, we fit a regression model using static predictors only (features in $\vec{x}_i$). This is to compare against typical approaches in clinical prediction which rely only on observed features to predict disease progression (e.g. Khanna et al. 2011). The regression function is as follows, where $\Phi(t)$ is a B-spline basis:

$$\hat{y}(t) \mid \vec{x}_i = \Phi(t)^\top \left(\vec{\beta}_0 + \textstyle\sum_{x_{ij} \text{ in } \vec{x}_i} x_{ij} \vec{\beta}_i + \sum_{x_{ij}, x_{ik} \text{ in pairs of } \vec{x}_i} x_{ij} x_{ik} \vec{\beta}_{ij} \right). \tag{30}$$

The following baselines reflect state-of-the-art approaches for dynamical prediction. The focus for each of these models, as discussed in the related work section, is on dynamical prediction of single marker trajectories using the marker history and static measurements collected during the first visit. The second baseline, like Rizopoulos (2011) and Shi et al. (2012), defines a single mean function parameterized in the same way as the first baseline and models individual-specific variations using a GP with the same kernel as in Equation 15 (using hyper-parameters as above). The third baseline is a mixture of B-splines, which models subpopulations that can express different trajectory shapes (as in Proust-Lima et al. (2014)).[3] Finally, we use the LTM (no coupling to auxiliary markers) as a baseline. All B-spline bases used in these baseline models are parameterized in the same way as the C-LTM (described above).

Evaluation. Prediction accuracy for all models is measured using the absolute error between the predicted and a smoothed version of the individual's observed trajectory. We make predictions after one, two, and four years of follow-up, which are summarized using averages computed in the second year of follow-up ($t \in (1, 2]$), in the third and fourth year of follow-up ($t \in (2, 4]$), fifth to eighth year of follow-up ($t \in (4, 8]$), and beyond the eighth year of follow-up ($t \in (8, 25]$)[4]. Mean absolute errors (MAE) and standard errors are estimated using 9-fold CV[5] at the level of individuals (i.e. all of an individual's data is held-out). Significance tests are computed against baselines using a paired t-test with point-wise predictions aggregated across folds.

4.1 Results

In this section, we present four sets of results. The first two are qualitative, and demonstrate the advantages of the C-LTM over the baseline models using examples. In the first

3. For the B-spline mixture, we use the subtypes discovered by LTM as the mixture classes. Without accounting for individual-specific variability explicitly, we have found that fitting a B-spline mixture using EM recovers poor classes that do not capture important trajectory shapes in the data. For additional details, see Section 3 in the supplement of Schulam and Saria (2015).
4. After the eighth year, data becomes too sparse to further divide this time span.
5. Recall that the first of 10 folds is used for hyperparameter estimates.

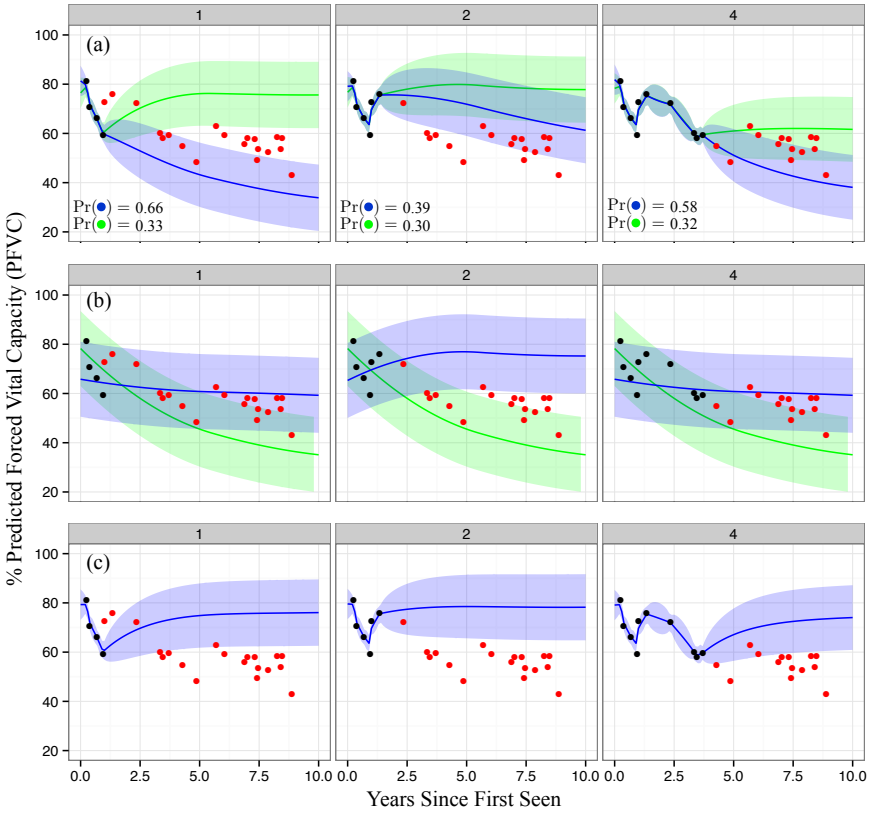


Figure 5: Examples of predictions made using 1, 2, and 4 years of data (moving across columns from left to right). Plot (a) shows dynamic predictions using C-LTM. Red markers are unobserved. Blue shows the trajectory predicted using the most likely subtype, and green shows the second most likely. Plot (b) shows dynamic predictions for the B-spline mixture baseline. Plot (c) shows the same for the B-spline + GP baseline.

qualitative analysis, we compare predictions made by C-LTM to those made by the B-spline mixture and the B-spline + GP. In the second qualitative analysis, we compare the C-LTM inferences with those from the LTM, which is a state-of-the-art single-marker model. The second two results are quantitative. The first compares predictive accuracies between the baseline models and the C-LTM. The second investigates clinical utility by using each model to predict a severity score that we use to detect individuals with aggressive lung disease.

4.1.1 Visual Comparison to Baselines

As an illustrative example to compare C-LTM with baselines, in Figures 5a, 5b, and 5c we show the dynamic predictions made using the C-LTM, the B-spline mixture, and the B-spline + GP baselines on a sample patient.[6] For each model, we show 95% posterior

6. This patient was selected as an exemplar for the types of errors commonly made by the baseline models.

intervals for the future trajectory. For the C-LTM and B-spline mixture, the most likely subtype is shown in blue and the second most likely is in green. The B-spline mixture (Figure 5b) cannot explain individual-specific sources of variation (e.g. short-term deviations from the mixture mean) and so over-reacts to the slight rise in PFVC seen in the last two observed (black) measurements in the second panel (year 2). The B-spline + GP (Figure 5c) cannot capture long-term differences in trajectory means (e.g. due to subtypes) and so pulls back to the population mean over time even after four years of data suggest a declining trajectory. On the other hand, at year 1 the C-LTM (Figure 5a) maintains the hypothesis that the individual may decline or return to stability (correctly putting most weight on the former). After 2 years of data, the temporary recovery seems to have caused confidence in the declining trajectory to fall (going from 66% to 39%), but the top-weighted hypothesis is still correct. After 4 years of data, the model again becomes confident in the declining trajectory. Clinically, this robustness to short-term changes is important. After having seen the recovery between years 1 and 2, a clinician may become less immediately concerned with the individual's future lung disease, possibly delaying immunotherapy until a rapid decline becomes more evident. Note that the B-spline mixture, on the other hand, over-reacts to the recovery and predicts that the individual will continue to recover.

4.1.2 Analysis of Example Inferences

In Figure 6a-d, we show the C-LTM's target and auxiliary marker inferences for four different patients. For the target marker (PFVC) and auxiliary markers (TSS, PDLCO, and PFEV1), we show the most likely (blue) and second most likely (green) subtype and their corresponding trajectories. For the RP and GI severity score markers, we show the most likely severity class (high versus low). The dashed lines indicate the threshold at which high and low are determined based on judgements by our clinical collaborators. For PFVC, PFEV1, and PDLCO lower values indicate more severe progression. For TSS, higher values indicate severe progression. In Figures 6e-h, we show the predictions made by LTM to visually compare against predictions made using the baseline markers and PFVC history only (i.e. that do not leverage information from auxiliary markers).

In Figure 6a, we see a 55 year-old woman who presents with mildly impaired lung function (approximately 65 PFVC), but seems to recover over the course of the first year to reach a PFVC above 75 (considered by clinicians to be relatively healthy). Using this information alone, one may suspect that she will not have future lung issues. Indeed, this is what LTM predicts as shown in Figure 6e. By examining her auxiliary markers, however, we see that the picture is less clear. In particular, PFEV1 (a clinical marker closely related to PFVC) both decreases and increases over that period. C-LTM infers a mildly declining trajectory for PFEV1. In addition, PDLCO is also noisy and overall low, which suggests that the blood is not efficiently absorbing oxygen. This can happen for a number of reasons, but active lung disease is one of them. Finally, we see that her initial skin score is quite high and C-LTM projects it to stay high for the next few years, which is associated with active lung disease. We see that C-LTM has successfully incorporated inferences about the future trends of the auxiliary markers and correctly predicts that this woman's PFVC will decline after this initial improvement.

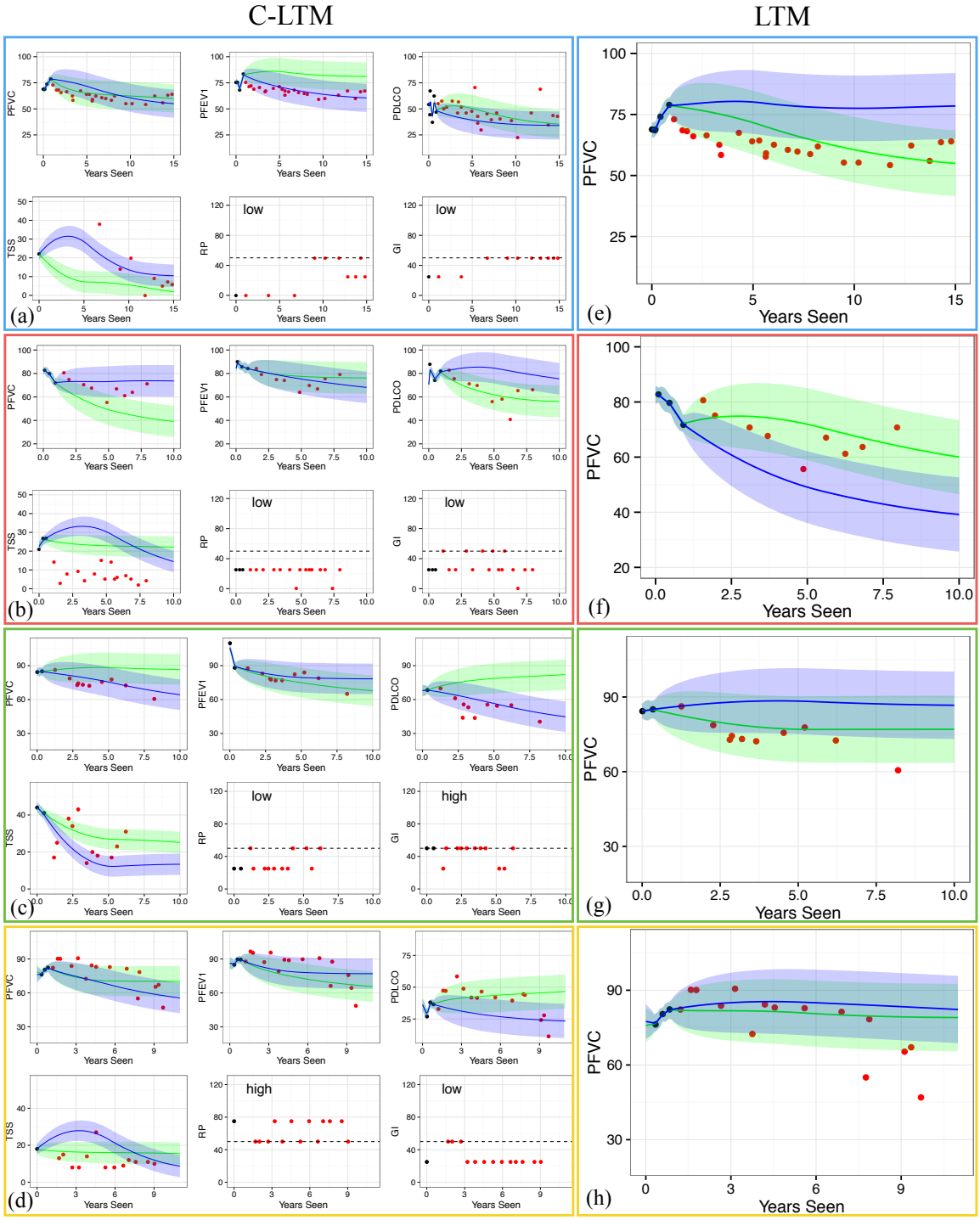


Figure 6: The predicted PFVC trajectory and the auxiliary markers are shown for two different patients. Red markers are unobserved. For the auxiliary markers TSS, PFEV1, and PDLCO we show the most likely (blue) and second most likely (green) subtype and their corresponding trajectories. For the RP and GI severity scores, we show the most likely severity class (high versus low). The dashed lines indicate the threshold at which high and low are determined clinically.

In Figure 6b, we see a 75 year-old white woman who presents with healthy lung function (approximately 85 PFVC), but is consistently declining over the course of the first year by nearly 15 PFVC. A clinical rule of thumb is that a drop in 10 PFVC over the course of a year warrants close monitoring for active lung disease. We see that LTM extrapolates this initial trend and predicts that this individual will continue to decline rapidly (Figure 6f). Just as in the previous example, however, the auxiliary markers paint a more complete picture of this individual. In the first few PFEV1 observations, we see that this decline is not quite as pronounced and the progression is predicted to be more mild. In PDLCO we see that oxygen is absorbed into the blood at healthy levels and also predicted to remain stable (although incorrectly in this case). Finally, C-LTM predicts that the RP and GI severity scores will remain low, which also supports the prediction that this woman will stabilize. Note that in this example C-LTM overestimates the course of PDLCO and TSS. Although the model still makes the correct prediction for PFVC in spite of this mistake, it highlights that the performance of our approach may be further improved with better auxiliary marker inferences. As research in systems biology yields new insights into modeling specific measurements more precisely, the modular architecture of C-LTM makes it possible to improve overall performance by incorporating improved versions of the target or auxiliary marker models.

In Figure 6c, we see a 76 year-old white woman that presents with healthy lung function (just under 90 PFVC), which also appears to be stable given the subsequent test result taken later that same year. The LTM predicts that this individual's most likely course is to remain stable. From the PFEV1 trajectory, however, we see that there was a large initial loss in PFEV1, which, together with the unusually high skin score (TSS) suggests that this woman's disease is active. The activity in the other organ systems allows the C-LTM to offset the stability seen in the first two PFVC measurements and correctly predict the consistently declining lung trajectory.

Finally, in Figure 6d, we see a 67 year-old African American man with mildly impaired lung function early in the disease course (around 75 PFVC) that seems to recover over the next one or two years to a healthier 85 PFVC. In Figure 6h, we see that the LTM predicts that a stable trajectory thereafter is likely. By considering other organ systems, however, we see that this man's blood-oxygen diffusion is severely limited early in the disease course (nearly 25% of the predicted DLCO). Moreover, we see that the this individual's Raynaud's phenomenon severity score is high early on and correctly predicted to remain that way. The low PDLCO and high RP severity score point to active vasculature disease, which is hypothesized to cause late deterioration in lung function. We see that C-LTM correctly uses this evidence to predict an accurate disease trajectory.

4.1.3 Predictive Accuracy

In Table 1, we report performance of the C-LTM, LTM, and the three other baseline models. First, we note that the C-LTM statistically significantly outperforms the B-spline with baseline features for all predictions. This baseline makes static predictions using baseline information only, and cannot adapt to an individual as new data becomes available. Moreover, after an initial amount of data has been collected on an individual, C-LTM statistically significantly outperforms all other models. This is not surprising. When compared to the

Model	$(1, 2]$	$(2, 4]$	$(4, 8]$	$(8, 25]$
Predictions using 1 year of data				
B-spline with Baseline Feats.	13.17 (0.43)	14.07 (0.61)	14.34 (0.65)	14.12 (1.04)
B-spline + GP	5.57 (0.24)	8.40 (0.19)	10.88 (0.42)	**11.74** (0.76)
B-spline Mixture	6.31 (0.22)	7.59 (0.36)	**9.82** (0.46)	13.77 (0.55)
LTM	5.70 (0.30)	8.02 (0.41)	11.17 (0.72)	13.93 (0.67)
C-LTM	★♣♦♠**5.12** (0.20)	★♣♦♠**6.88** (0.27)	★♣♠9.95 (0.51)	★13.70 (1.08)
Predictions using 2 years of data				
B-spline with Baseline Feats.		14.07 (0.61)	14.34 (0.65)	14.12 (1.04)
B-spline + GP		6.51 (0.19)	9.79 (0.35)	10.95 (0.68)
B-spline Mixture		6.17 (0.29)	8.34 (0.36)	12.19 (0.48)
LTM		5.74 (0.29)	8.08 (0.37)	**10.89** (0.62)
C-LTM		★♣♦♠**5.58** (0.34)	★♣**7.99** (0.61)	★♦11.27 (1.02)
Predictions using 4 years of data				
B-spline with Baseline Feats.			14.34 (0.65)	14.12 (1.04)
B-spline + GP			6.60 (0.24)	9.53 (0.56)
B-spline Mixture			6.00 (0.37)	10.11 (0.56)
LTM			**4.88** (0.28)	8.65 (0.59)
C-LTM			★♣♦5.04 (0.42)	★♣♦♠**8.07** (0.35)

Table 1: Mean absolute error of PFVC predictions for the B-spline with baseline features, the B-spline + GP, LTM, and C-LTM. Bold numbers indicate best performance across baseline models and proposed model. ★ indicates statistically significant improvement against the B-spline model with baseline features only using a paired t-test ($\alpha = 0.05$). ♣ indicates statistical significance compared against the B-spline + GP. ♦ indicates statistical significance compared against the B-spline mixture. ♠ indicates statistical significance compared against LTM.

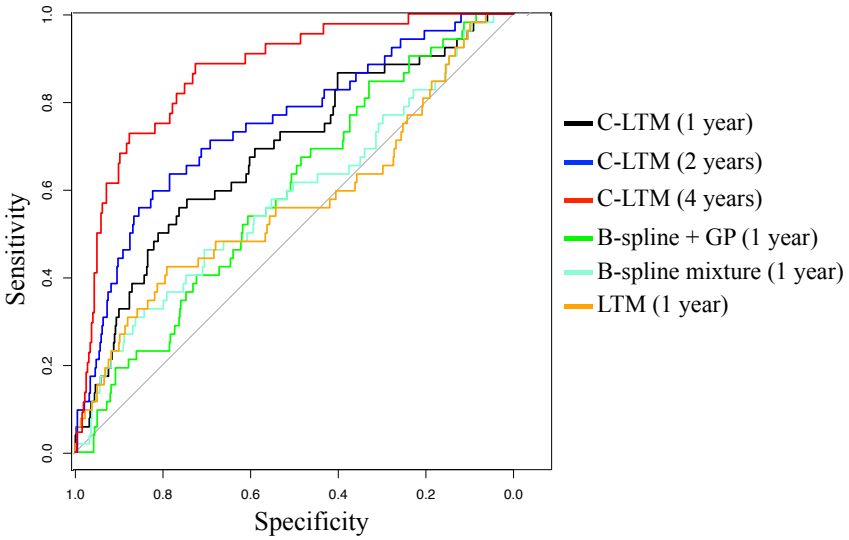
LTM, we see that C-LTM benefits from leveraging information from auxiliary markers. As more information is collected, both models are able to the individual and provide comparable predictions. The B-spline mixture is not able to personalize beyond capturing long-term trends across subpopulations, so we see that it becomes less competitive compared to both C-LTM and LTM as more data are collected. Finally, the B-spline + GP cannot capture long-term trends specific to subpopulations (as we saw in Section 4.1.1), and so we see that it does poorly when making predictions.

4.1.4 Clinical Utility

One may naturally wonder whether the observed improvements in MAE reported above translate to practical benefits in the clinic. In the examples shown in Figure 6, we have walked through cases where the model makes predictions that would seem unlikely if we were to consider PFVC alone. This suggests that the model can augment expert clinical judgement and may serve to protect against incorrect extrapolations. In this section, we further elaborate upon this intuition by studying clinical utility quantitatively. In particular, we compare how well the B-spline + GP, B-spline mixture, LTM, and C-LTM are able to detect individuals who will have rapidly declining lung function. It is notoriously difficult to predict which scleroderma patients will rapidly decline using only information from early in the disease course. In addition to improving prognoses, more accurate detection of rapidly declining lung function can help to improve the recruitment for clinical trials evaluating drugs for scleroderma-related lung disease. If we include many individuals in a study who

Model / Years of Data	1	2	4
B-spline + GP	0.59	0.63	0.74
B-spline mixture	0.58	0.63	0.76
LTM	0.57	0.71	0.84
C-LTM	**0.68**	**0.75**	**0.87**

(a) AUCs for detecting declining individuals.



(b) ROCs comparing B-spline + GP at 1 year, B-spline mixture at 1 year, LTM at 1 year, and C-LTM at years 1, 2, and 4.

Figure 7: Declining individual detection results.

are predicted to have active lung disease but do not, the results of the study are blurred because both arms of the trial may include many individuals without active lung disease.

To test how well these different models can detect individuals that will experience rapidly declining lung function, we use the predictions of future PFVC measurements to produce a score. The score is defined to be the difference between the individual's first PFVC measurement and the minimum predicted value in the future—this will be higher for individuals on whom a model predicts deteriorating lung function and lower for those predicted to be stable. To label an individual as declining, we require that they (1) have at least one observation within the first year of being seen by the clinic, (2) have 3 years between their first and last measurements, (3) have at least 4 PFVC measurements, and (4) have an initial PFVC measurement that is 20 PFVC higher than their last measurement. Requirements (2) and (3) are to ensure that the trajectory can be reliably annotated as declining or not. For each model, we make predictions at years 1, 2, and 4 and compute the score described above for each individual. We then compute the AUC for each model at each year. Table 7a displays the results of this experiment. We see that C-LTM achieves higher AUC at all years than the baseline models. Figure 7b displays the ROCs for the B-Spline + GP (green), B-Spline mixture (cyan), LTM (orange), and the proposed model (black) at year 1

and also includes the ROCs for the proposed model at years 2 (blue) and 4 (red) to visualize how performance improves as more data is added. Clinically, an AUC of 0.87 for predicting individuals with lung decline after—on average—four years of data is high and has not been shown previously.

5. Discussion

The goal of personalized (also called precision) medicine is to develop tools that help to tailor prognoses to the characteristics and unique medical history of the individual. In this paper, we describe an approach to personalized prognosis that uses an integrative analysis of multiple clinical marker histories from the individuals medical records. Our approach combines single-marker latent variable models (the LTM) with a CRF coupling model to make more accurate predictions about the future trajectory of a target clinical marker.

The coupled model (C-LTM) has several advantages. First, the marker-specific LTMs account for marker trajectory shapes using components at the population, subpopulation, individual long-term, and individual short-term levels, which simultaneously allows for heterogeneity across and within individuals, and enables statistical strength to be shared across observations at different "resolutions" of the data. Within an individual marker model, the population and subpopulation components are learned offline, while estimates of the individual-specific parameters are refined over the course of the disease as data accrues for that individual. Second, our coupling model allows us to condition both the target and auxiliary marker histories to make predictions about the future target marker trajectory. We therefore use the marker-specific latent variable models to neatly summarize and extra information from the irregularly sampled and sparse, while simultaneously sidestepping the issue of jointly modeling both the target and auxiliary markers. The conditional formulation is less sensitive to misspecified dependencies between different marker types and can also be easily scaled to diseases with a large number of auxiliary markers. Finally, our model aligns with clinical practice; predictions are dynamically updated in continuous time as new marker observations are measured. We also note that our description of the method and the experimental results focus on predicting the trajectory of a single clinical marker, but multiple latent factor regression models can be easily fit so that many markers can be simultaneously predicted. Using this extension, we only need to maintain different CRF parameters; the latent variable models are shared since they are fit independently as a precursor to learning the CRF.

There are several shortcomings of the proposed approach that are promising directions for future research. First, the model implicitly assumes that the data generating process is noninformative (i.e. missing data is missing at random (Little and Rubin, 2014)). This is appropriate for clinical markers that are routinely collected, but additional machinery would be required to model markers whose missingness is informative. For example, in some cases, additional measurements may be made due to clinical suspicion caused by factors that are not clearly document in the health record. Researchers have begun to explore more complex missing-data mechanisms for disease trajectory modeling (e.g., Lange et al. 2015; Coley et al. 2016), and it will be important to incorporate these ideas into the framework discussed here to integrate the full set of markers measured during a clinical visit. Another shortcoming is our focus on discrete latent factors of the auxiliary marker trajectories. Continuous-valued

latent factors may also be useful, but would make learning and inference in the latent factor CRF more challenging.

There are also several other immediate opportunities for improving the model. Auxiliary markers are integrated via separate marker-specific generative models. While we incorporated two different types of models—trajectory and maximum-severity based—both of which were data driven, existing and new clinical knowledge should be brought to bear to improve these models, which we expect will improve predictions of the target trajectories. Further, in this work, we focused on modeling the dependency of the target subtype on the auxiliary markers. In addition, estimates of the individual-specific long-term and short-term components may also benefit from conditioning on the auxiliary markers. Finally, the parameters for the pairwise potentials learned in our model may serve as a means for generating hypotheses about the co-evolution of organ-specific trajectories.

The ideas proposed here also open up other longer-term directions for future work. The proposed model does not account for the effects of treatment on an individual's long-term trajectory. In many chronic conditions, as is the case for scleroderma, drugs only provide short-term relief (accounted for in our model by the individual-specific adjustments). However, if treatments that alter long-term course are available and commonly prescribed, then these should be included within the model as an additional component that influences the trajectory. Learning these treatment effects from noisy electronic health record data (e.g., Xu et al. 2016) present an exciting and challenging direction for future work.

We have demonstrated our model by developing a prognostic tool for predicting lung disease trajectories in patients with scleroderma, an autoimmune disease. We showed that the proposed model makes more accurate predictions than state-of-the-art approaches. Accurate tools for prognosis can allow clinicians and patients to more actively manage their disease. While we have focused model development and evaluation on scleroderma, this work is broadly applicable to other complex diseases (Craig, 2008), many of which track disease activity using clinical scales of severity. The proposed model is most directly applicable to CCDs where heterogeneity in disease presentation is common. Examples of such diseases include lupus, multiple sclerosis, inflammatory bowel disease (IBD), chronic obstructive pulmonary disease (COPD), and asthma. Extensions of the proposed ideas, and the model, to these diseases offer an opportunity to address important open challenges in precision medicine.

Acknowledgments

We would like to thank Drs. Laura Hummers, Fredrick Wigley and Robert Wise who have provided extensive clinical guidance as well as the data set with which this study was conducted. We would also like to thank Dr. Colin Ligon for his generous support in chart reviewing patients; many of the key ideas in our work were motivated by these chart reviews. Finally, we would like to thank Zachary Barnes who helped deploy a previous iteration of this model as a tool in the clinic and shadowed clinicians while it was in use. Our work has benefited from the lessons he learned when observing the clinicians use this tool.

Appendix A. Learning the C-LTM: Details

In this section, we provide additional details on the learning algorithm for the C-LTM. Recall that this consists of two stages: (1) independently fitting the single-marker models (the LTM in our case), and (2) fitting the parameters of the coupling model. We describe both stages below.

A.1 Learning the LTM

To learn the parameters of the single-marker model $\Theta = \{\Lambda, \pi_{1:G}, \vec{\beta}_{1:G}, \Sigma_b, a, \ell, \sigma^2\}$, we maximize the observed-data log-likelihood of a training sample of M retrospectively observed trajectories (i.e. the probability of all individual's marker values $\vec{y}_i$ given measurement times $\vec{t}_i$ and features $\vec{x}_i$). Using the expression for the observed-data likelihood in Eq. 14, we have that the observed-data log-likelihood for all individuals in a training sample is

$$\mathcal{L}(\Theta) = \sum_{i=1}^{M} \log \left[\sum_{z_i=1}^{G} \pi_{z_i} \mathcal{N}\left(\vec{y}_i \mid \Phi_p\left(\vec{t}_i\right) \Lambda \vec{x}_i + \Phi_z\left(\vec{t}_i\right) \vec{\beta}_{z_i}, K\left(\vec{t}_i, \vec{t}_i\right) \right) \right]. \tag{31}$$

To maximize the observed-data log-likelihood with respect to Θ, we partition the parameters into two subsets. The first subset, $\Theta_1 = \{\Sigma_b, \alpha, \ell, \sigma^2\}$, contains values that parameterize the covariance function shown in Equation 15. As is often done when designing the kernel of a Gaussian process, we use a combination of domain knowledge to choose candidate values and model selection using observed-data log-likelihood as a criterion for choosing among candidates (Rasmussen and Williams, 2006). The second subset, $\Theta_2 = \{\Lambda, \pi_{1:G}, \vec{\beta}_{1:G}\}$, contains values that parameterize the mean of the multivariate normal distribution in Equation 14. We learn these parameters using expectation maximization (EM) to find a local maximum of the observed-data log-likelihood in Equation 31 (Dempster et al., 1977).

Expectation step. All parameters related to $\vec{b}_i$ and f_i are limited to the covariance kernel and are not optimized using EM. We therefore only need to consider the subtype indicators z_i as unobserved in the expectation step. Because z_i is discrete, its posterior is computed by normalizing the joint probability of z_i and $\vec{y}_i$. Let π^*_{ig} denote the posterior probability that individual i has subtype $g \in \{1, \ldots, G\}$, then we have

$$\pi^*_{ig} \propto \pi_g \mathcal{N}\left(\vec{y}_i \mid \Phi_p\left(\vec{t}_i\right) \Lambda \vec{x}_i + \Phi_z\left(\vec{t}_i\right) \vec{\beta}_g, K\left(\vec{t}_i, \vec{t}_i\right) \right). \tag{32}$$

Maximization step. In the maximization step, we optimize the marginal probability of the soft assignments under the multinomial model with respect to $\pi_{1:G}$. This amounts to collecting total "soft counts" computed in Eq. 32 for each subtype and renormalizing. To optimize the expected complete-data log-likelihood with respect to Λ and $\vec{\beta}_{1:G}$, we note that the mean of the multivariate normal for each individual is a linear function of these parameters. Holding Λ fixed, we can therefore solve for $\vec{\beta}_{1:G}$ in closed form and vice versa. We use a block coordinate ascent approach, alternating between solving for Λ and $\vec{\beta}_{1:G}$ until convergence. To control the smoothness of the subtypes we penalize the log-likelihood with respect to the subtype parameters $\vec{\beta}_{1:G}$ as in Eq. 4. Because the penalized expected complete-data log-likelihood is concave with respect to all parameters in Θ_2, each maximization step is guaranteed to converge. The exact computations required to maximize the expected log-likelihood can be found in Schulam and Saria (2015) and its supplement.

A.2 Learning the Coupling Model

To learn the parameters of the latent-factor CRF regression, we directly maximize the conditional probability of future target markers given previously observed target markers, previously observed auxiliary markers, and static baseline covariates on a collection of examples extracted from retrospective data. Suppose we are given records containing the target marker, auxiliary markers, and baseline covariates for M individuals. We choose a collection of times $\mathcal{T}$ that will be used to create training examples of history-future pairs. For example, we may choose $\mathcal{T} = \{1, 2\}$ because early management decisions are made using prognoses at years 1 and 2. We emphasize, however, that the model is *not* restricted to making predictions at years 1 and 2; it can make predictions at arbitrary times. The times $\mathcal{T}$ are simply used to create training instances. We also note that it is possible to train specialized models for different time periods. For example, we may train one model for making predictions in the first 2 years and another for beyond 4 years. Given the M records and times $\mathcal{T}$, we define the objective:

$$J(\gamma, \gamma_{1:C}, \theta, \theta_{1:C}, \eta_{1:C}) = \sum_{i=1}^{M} \sum_{t \in \mathcal{T}} \log p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t)) \tag{33}$$

$$= \sum_{i=1}^{M} \sum_{t \in \mathcal{T}} \log \left(\sum_{z_i} \underbrace{p(\vec{y}_{i,>t} \mid z_i)}_{\text{(A)}} \underbrace{p(z_i \mid \mathcal{H}(i,t))}_{\text{(B)}} \right), \tag{34}$$

where (A) is the subtype-specific multivariate normal likelihood in Eq. 14 and (B) is the conditional distribution over z_i shown in Eq. 28. To learn the parameters, we maximize this objective with respect to γ, $\gamma_{1:C}$, θ, $\theta_{1:C}$, and $\eta_{1:C}$ using gradient-based methods (e.g. `L-BFGS`). In our experiments, we optimize a regularized version of the objective, but for simplicity this section discusses the computations required to compute the gradient of Eq. 33 only. Consider a single summand of Eq. 33

$$\log p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t)) = \log \left(\sum_{z_i} p(\vec{y}_{i,>t} \mid z_i) p(z_i \mid \mathcal{H}(i,t)) \right). \tag{35}$$

To reiterate, the parameters of the density $p(\vec{y}_{i,>t} \mid z_i)$ are assumed to have been learned in a separate step (e.g. using the EM algorithm presented above), and so we are only concerned with estimating the parameters of the singleton and pairwise factors in the CRF: $\gamma, \gamma_{1:C}, \theta, \theta_{1:C}, \eta_{1:C}$.

Gradient of the objective. We derive the gradient for a single summand of the objective (Eq. 33), which are combined additively to form the full gradient used at each iteration. Although our model is log-linear over all latent variables z_i and $z_{1:C,i}$, Eq. 35 is not linear in the parameters because the random field does not directly estimate the conditional distribution over the future target clinical markers, but instead estimates the weights assigned to each configuration of the latent variables. We therefore have that the partial derivative of Eq. 35 with respect to any parameter θ_k is:

$$\frac{\partial \log p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t))}{\partial \theta_k} = \frac{\left(\sum_z p(\vec{y}_{i,>t} \mid z) \frac{\partial p(z \mid \mathcal{H}(i,t))}{\partial \theta_k} \right)}{p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t))}. \tag{36}$$

To complete the expression for the partial derivative, we need to compute the partial derivative of the probability of a given target marker latent variable z with respect to the parameter θ_k. We have that:

$$\frac{\partial p(z \mid \mathcal{H}(i,t))}{\partial \theta_k} = \frac{\partial}{\partial \theta_k}\frac{Z'_{i,t}(z)}{Z_{i,t}} = \frac{1}{Z_{i,t}}\frac{\partial Z'_{i,t}(z)}{\partial \theta_k} + Z'_{i,t}(z)\frac{\partial Z^{-1}_{i,t}}{\partial \theta_k}. \tag{37}$$

We can now leverage identities from the theory of log-linear models to continue with the derivation. In particular, recall that log-linear models are in the exponential family of distributions. As a consequence, we can consider the parameters $\gamma, \gamma_{1:C}, \theta, \theta_{1:C}, \eta_{1:C}$ as the *natural parameters* of the distribution. The corresponding *sufficient statistics* are therefore the factors in the log-linear model:

$$\begin{aligned} T(z, z_{1:C}, \vec{x}_i) = [&\ell_{i,\leq t}(z), \ell_{1,i,\leq t}(z_1), \ldots, \ell_{C,i,\leq t}(z_C), \\ &f^\top(z, \vec{x}_i), f_1^\top(z_1, \vec{x}_i), \ldots, f_C^\top(z_C, \vec{x}_i), \qquad (38) \\ &g_1^\top(z, z_1), \ldots, g_C^\top(z, z_C)]^\top. \qquad (39) \end{aligned}$$

An important property of exponential families is that the gradient of the log-normalizing-constant with respect to the natural parameters is simply the expected value of the sufficient statistics computed using the current value of the natural parameters. Note that both $Z'_{i,t}(z)$ and $Z_{i,t}$ are normalizing constants of exponential family distributions. In the case of $Z_{i,t}$ this is trivial to see because it is the normalizing constant of our log-linear model. In the case of $Z'_{i,t}(z)$ we see that it is the normalizing constant of a log-linear model over the auxiliary marker latent variables $z_{1:C}$ given *both* z and the clinical history $\mathcal{H}(i,t)$. We therefore have:

$$\begin{aligned} \frac{\partial \log Z'_{i,t}(z)}{\partial \theta_k} &= \mathbb{E}_{z_{1:C}}\left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t)\right] \\ &\implies \frac{\partial Z'_{i,t}(z)}{\partial \theta_k} = Z'_{i,t}(z)\mathbb{E}_{z_{1:C}}\left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t)\right], \qquad (40) \\ \frac{\partial \log Z_{i,t}}{\partial \theta_k} &= \mathbb{E}_{z, z_{1:C}}\left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t)\right] \\ &\implies \frac{\partial Z^{-1}_{i,t}}{\partial \theta_k} = -\frac{1}{Z_{i,t}}\mathbb{E}_{z, z_{1:C}}\left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t)\right], \qquad (41) \end{aligned}$$

where we have used $T(z, z_{1:C}, \vec{x}_i)_k$ to denote the feature (or sufficient statistic) corresponding to the parameter θ_k. By plugging these partial derivatives back into Eq. 37, we have

$$\begin{aligned} \frac{\partial}{\partial \theta_k}\frac{Z'_{i,t}(z)}{Z_{i,t}} &= \frac{Z'_{i,t}(z)}{Z_{i,t}}\left(\mathbb{E}_\Theta\left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t)\right] - \mathbb{E}_\Theta\left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t)\right]\right) \qquad (42) \\ &= p(z \mid \mathcal{H}(i,t))\left(\mathbb{E}_\Theta\left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t)\right] - \mathbb{E}_\Theta\left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t)\right]\right). \qquad (43) \end{aligned}$$

In words, we see that the partial derivative with respect to a parameter θ_k is the expected value of its corresponding feature given that we have observed the target marker latent variable z and clinical history $\mathcal{H}(i,t)$ minus the expected value of the feature given only the

clinical history $\mathcal{H}(i,t)$. The difference is then weighted by the probability of observing the target marker latent variable given the clinical history. By plugging this expression back into Eq. 36, we arrive at the final expression for the partial derivative of a single summand with respect to θ_k:

$$\frac{\partial \log p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t))}{\partial \theta_k} \tag{44}$$

$$= \sum_z \frac{p(\vec{y}_{i,>t} \mid z) p(z \mid \mathcal{H}(i,t))}{p(\vec{y}_{i,>t} \mid \mathcal{H}(i,t))} \left(\mathbb{E}_\Theta \left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t) \right] - \mathbb{E}_\Theta \left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t) \right] \right)$$

$$= \sum_z p(z \mid \vec{y}_{i,>t}, \mathcal{H}(i,t)) \left(\mathbb{E}_\Theta \left[T(z, z_{1:C}, \vec{x}_i)_k \mid z, \mathcal{H}(i,t) \right] - \mathbb{E}_\Theta \left[T(z, z_{1:C}, \vec{x}_i)_k \mid \mathcal{H}(i,t) \right] \right). \tag{45}$$

The partial derivative has a nice interpretation. Each summand has similar structure to the partial derivative of $p(z \mid \mathcal{H}(i,t))$ (Eq. 42), but the weight conditioned on only the clinical history has been replaced with a weight conditioned on both the clinical history *and* the future target marker trajectory. The partial derivatives of the summands of the objective in Eq. 33 are added together to obtain the partial derivative with respect to the objective. These partial derivatives are combined to form a gradient, which is easily plugged into existing first-order optimization routines. Optionally, the objective can be augmented with a regularizer to restrict the complexity of the model or to encourage a sparse solution to the learning problem. This concludes our discussion of the learning algorithm.

Appendix B. Missing Data for Continous-Time Trajectories

Trajectories in continuous-time can be thought of as random *functions* $F(\cdot)$ (Gaussian processes are an example of a family of distributions over random functions). Although the function specifies infinitely many values, to learn continuous-time models we maximize the probability of a finite set of observations (or a penalized version of this objective). In *observational* health care data, we need to be careful that we do not bias our likelihood-based learners by unduly ignoring the dependence between the finite set of times at which we observe the trajectory and the trajectory's values at those times. For example, if the trajectory is more likely to be sampled when its value is low, then our model will learn that trajectories with high values are less likely than they actually are.

The aim of this section is to posit a set of assumptions about continuous trajectory observation times that are (1) substantively reasonable, and (2) justify the use of standard likelihood-based learning. At a high-level, we assume that trajectory observation times are functions of the previous observation times and the values of the trajectory sampled at those times. These assumptions are more formally encoded in the graphical model shown in Figure 8, which expresses dependencies for an individual with three trajectory observations. In the figure, $F(\cdot)$ denotes the full trajectory, $\{T_1, T_2, T_3\}$ are random variables denoting the times at which the trajectory is sampled, and $\{Y_1^\star, Y_2^\star, Y_3^\star\}$ are the observed data. The conditional probability distribution of any $Y_i^\star$ given the trajectory and associated observation time is simply:

$$p(Y_i^\star = y_i^\star \mid T_i = t_i, F = f) = \mathbb{I}(f(t_i) = y_i^\star). \tag{46}$$

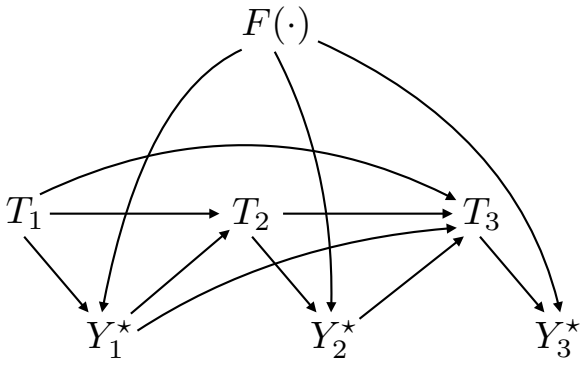


Figure 8: Example missing data mechanism in continuous-time.

These assumptions are reasonable in many healthcare settings. For example, in an ICU where a patient is constantly under supervision, we can reasonably assume that clinical marker measurements are made at times that depend on the previous observations (e.g. the individual is thought to be at risk and so measurements are taken more frequently) and on previous observation times (e.g. a measurement has not been recorded in a while, so we should collect a new one). In the outpatient setting, an individual with a particular disease that is being actively managed by a physician will have follow-up visits scheduled either routinely or more frequently if the physician is especially concerned. On the other hand, modeling the progression of a disease such as the flu using information from a general practitioner's office may not satisfy our assumption because individual's with less severe manifestations are less likely to visit.

Conditioned on these assumptions about the dependencies between the trajectory, observation times, and observed values, we want to justify likelihood-based learning. Suppose we have a trajectory model with parameters Θ that allows us to compute the probability of any finite set of trajectory values. For example, we can compute $p_\Theta(F(t_1) = y_1^\star, F(t_2) = y_2^\star, F(t_3) = y_3^\star)$. The observed data, however, are the observation times and sampled values: $\{T_{1:n}, Y_{1:n}^\star\}$. Proper likelihood-based learning requires that we maximize:

$$p(T_{1:n} = t_{1:n}, Y_{1:n}^\star = y_{1:n}^\star). \tag{47}$$

However, this expression is determined by both the observation time mechanism and the trajectory model. Our goal is to show that this can be factored into two terms: one that depends on the observed data and the observation time mechanism parameters, and the other that depends on the sampled trajectory values and the trajectory model parameters Θ. To do this, we first see that Equation 47 can be written as

$$\int p(F = f) p(T_{1:n} = t_{1:n}, Y_{1:n}^\star = y_{1:n}^\star \mid F = f) dF. \tag{48}$$

The integrand in Equation 48 can be now be factored further to obtain

$$p(F = f) \prod_{i=1}^{n} p(T_i = t_i \mid \mathcal{H}_i) p(Y_i^\star = y_i^\star \mid T_i = t_i, F = f), \tag{49}$$

where $\mathcal{H}_i$ is defined to be the previous $i-1$ observation times and sampled trajectory values. Note that the first term in the product of Equation 49 can be pulled out of the integral, allowing us to write Equation 48 as

$$\left[\prod_{i=1}^{n} p(T_i = t_i \mid \mathcal{H}_i)\right]\left[\int p(F = f)\prod_{i=1}^{n} p(Y_i^\star = y_i^\star \mid T_i = t_i, F = f)dF\right]. \tag{50}$$

The left-hand factor above depends only on the observation time mechanism and the observed data. Moreover, the right-hand factor depends only on the trajectory model and the sampled trajectory values, which we now show:

$$\begin{aligned}
&\int p(F = f)\prod_{i=1}^{n} p(Y_i^\star = y_i^\star \mid T_i = t_i, F = f)dF \\
&\qquad = \int p(F = f)\prod_{i=1}^{n} \mathbb{I}(f(t_i) = y_i^\star)dF \\
&\qquad = \int p(F = f)\mathbb{I}(f(t_1) = y_1^\star, \ldots, f(t_n) = y_n^\star)dF \\
&\qquad = p_\Theta(f(t_1) = y_1^\star, \ldots, f(t_n) = y_n^\star).
\end{aligned} \tag{51}$$

We therefore see that, given our observation time mechanism assumptions, maximizing the likelihood of the sampled trajectory values under our trajectory model is equivalent to maximizing the "proper" likelihood in Equation 47 with respect to the model parameters Θ. This result aligns with Theorems 7.1 and 8.1 found in Rubin's original paper on missing data (Rubin, 1976).

References

Y. Allanore, R. Simms, O. Distler, M. Trojanowska, J. Pope, C.P. Denton, and J. Varga. Systemic sclerosis. *Nature Reviews Disease Primers*, 2015.

G. Andrew and J. Gao. Scalable training of l1-regularized log-linear models. In *International Conference on Machine Learning (ICML)*, 2007.

K.H. Brodersen, F. Gallusser, J. Koehler, N. Remy, and S.L. Scott. Inferring causal impact using bayesian structural time-series models. *The Annals of Applied Statistics*, 9(1): 247–274, 2015.

S. Chib and B.H. Hamilton. Semiparametric bayes analysis of longitudinal data treatment models. *Journal of Econometrics*, 110(1):67–89, 2002.

R.Y. Coley, A.J. Fisher, M. Mamawala, H.B. Carter, K.J. Pienta, and S.L. Zeger. A bayesian hierarchical model for prediction of latent health states from multiple data sources with application to active surveillance of prostate cancer. *Biometrics*, 2016.

F.S. Collins and H. Varmus. A new initiative on precision medicine. *New England Journal of Medicine*, 372(9):793–795, 2015.

J. Craig. Complex diseases: Research and applications. *Nature Education*, 1(1):184, 2008.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (methodological)*, pages 1–38, 1977.

F. Doshi-Velez, Y. Ge, and I. Kohane. Comorbidity clusters in autism spectrum disorders: an electronic health record time-series analysis. *Pediatrics*, 133(1):e54–e63, 2014.

R. Dürichen, M.A.F. Pimentel, L. Clifton, A. Schweikard, and D.A. Clifton. Multitask gaussian processes for multivariate physiological time-series analysis. *Biomedical Engineering, IEEE Transactions on*, 62(1):314–322, 2015.

P.H.C Eilers and B.D. Marx. Flexible smoothing with b-splines and penalties. *Statistical Science*, pages 89–102, 1996.

J. Friedman, T. Hastie, and R. Tibshirani. *The Elements of Statistical Learning*. Springer, 2001.

A. Gelman, J.B. Carlin, H.S. Stern, and D.B. Rubin. *Bayesian Data Analysis*. Taylor & Francis, 2014.

M.R Hassan and B. Nath. Stock market forecasting using hidden markov model: a new approach. In *Intelligent Systems Design and Applications, 5th International Conference on*, pages 192–196. IEEE, 2005.

G.M. James and C.A. Sugar. Clustering for sparsely sampled functional data. *Journal of the American Statistical Association*, 98(462):397–408, 2003.

D. Khanna, C.H Tseng, N. Farmani, V. Steen, D.E. Furst, P.J. Clements, M.D. Roth, J. Goldin, R. Elashoff, J.R. Seibold, R. Saggar, and D.P. Tashkin. Clinical course of lung physiology in patients with scleroderma and interstitial lung disease: analysis of the scleroderma lung study placebo group. *Arthritis & Rheumatism*, 63(10):3078–3085, 2011.

S. Kleinberg and G. Hripcsak. A review of causal inference for biomedical informatics. *Journal of Biomedical Informatics*, 44(6):1102–1112, 2011.

J.M. Lange, R.A. Hubbard, L.Y.T Inoue, and V.N. Minin. A joint model for multistate disease processes and random informative observation times, with applications to electronic medical records data. *Biometrics*, 71(1):90–101, 2015.

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

D.S. Lee, P.C. Austin, J.L. Rouleau, P.P. Liu, D. Naimark, and J.V. Tu. Predicting mortality among patients hospitalized for heart failure: derivation and validation of a clinical model. *Journal of the American Medical Association*, 290(19):2581–2587, 2003.

S.J.G. Lewis, T. Foltynie, A.D. Blackwell, T.W. Robbins, A.M. Owen, and R.A. Barker. Heterogeneity of parkinsons disease in the early clinical stages using a data driven approach. *Journal of Neurology, Neurosurgery & Psychiatry*, 76(3):343–348, 2005.

R.J.A. Little and D.B. Rubin. *Statistical Analysis with Missing Data*. John Wiley & Sons, 2014.

Z. Liu and M. Hauskrecht. Clinical time series prediction: Toward a hierarchical dynamical system framework. *Artificial Intelligence in Medicine*, 2014.

J. Lötvall, C.A. Akdis, L.B. Bacharier, L. Bjermer, T.B. Casale, A. Custovic, R.F. Lemanske, A.J. Wardlaw, S.E. Wenzel, and P.A. Greenberger. Asthma endotypes: a new approach to classification of disease entities within the asthma syndrome. *Journal of Allergy and Clinical Immunology*, 127(2):355–360, 2011.

K.P. Murphy. *Dynamic bayesian networks: representation, inference and learning*. PhD thesis, University of California, Berkeley, 2002.

K.P. Murphy. *Machine Learning: A Probabilistic Perspective*. MIT press, 2012.

J.B. Oliva, W. Neiswanger, B. Póczos, E.P. Xing, H. Trac, S. Ho, and J.G. Schneider. Fast function to function regression. In *Conference on Artificial Intelligence and Statistics (AISTATS)*, 2015.

C. Proust-Lima, M. Séne, J.M.G Taylor, and H. Jacqmin-Gadda. Joint latent class models for longitudinal and time-to-event data: A review. *Statistical Methods in Medical Research*, 23(1):74–90, 2014.

J.A. Quinn, C.K. Williams, and N. McIntosh. Factorial switching linear dynamical systems applied to physiological condition monitoring. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(9):1537–1551, 2009.

R. Raina, Y. Shen, A. Mccallum, and A.Y. Ng. Classification with hybrid generative/discriminative models. In *Advances in Neural Information Processing Systems (NIPS)*, 2003.

J.O. Ramsay. *Functional Data Analysis*. Wiley Online Library, 2006.

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

D. Rizopoulos. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*, 67(3):819–829, 2011.

D. Rizopoulos and P. Ghosh. A bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Statistics in Medicine*, 30(12):1366–1380, 2011.

S. Roberts, M. Osborne, M. Ebden, S. Reece, N. Gibson, and S. Aigrain. Gaussian processes for time-series modelling. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 371(1984):20110550, 2013.

K.R. Rosenbloom, C.A. Sloan, V.S. Malladi, T.R. Dreszer, K. Learned, V.M. Kirkup, M.C. Wong, M. Maddren, R. Fang, S.G. Heitner, B.T. Lee, G.P. Barber, R.A. Harte, M. Diekhans, J.C. Long, S.P. Wilder, A.S. Zweig, D. Karolchik, R.M. Kuhn, D. Haussler, and W.J. Kent. Encode data in the UCSC genome browser: year 5 update. *Nucleic acids research*, 41(D1):D56–D63, 2013.

D.B. Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

S. Saria and A. Goldenberg. Subtyping: What Is It and Its Role in Precision Medicine. *IEEE Intelligent Systems*, 30, 2015.

P.F. Schulam and S. Saria. A framework for individualizing predictions of disease trajectories by exploiting multi-resolution structure. In *Advances in Neural Information Processing Systems (NIPS)*, pages 748–756, 2015.

P.F. Schulam, F. Wigley, and S. Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Conference on Artificial Intelligence (AAAI)*, 2015.

J.Q. Shi, R. Murray-Smith, and D.M. Titterington. Hierarchical gaussian process mixtures for regression. *Statistics and Computing*, 15(1):31–41, 2005.

J.Q. Shi, B. Wang, E.J. Will, and R.M. West. Mixed-effects gaussian process functional regression models with application to dose–response curve prediction. *Statistics in Medicine*, 31(26):3165–3177, 2012.

D.P. Tashkin et al. Cyclophosphamide versus placebo in scleroderma lung disease. *New England Journal of Medicine*, 354(25):2655–2666, 2006.

J. Varga, C.P. Denton, and F.M. Wigley. *Scleroderma: From Pathogenesis to Comprehensive Management*. Springer Science & Business Media, 2012.

H. Wang, F. Nie, H. Huang, J. Yan, S. Kim, S. Risacher, A. Saykin, and L. Shen. High-order multi-task feature learning to identify longitudinal phenotypic markers for alzheimer's disease progression prediction. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1277–1285, 2012.

Y. Xu, Y. Xu, and S. Saria. A bayesian nonparametic approach for estimating individualized treatment-response curves. *arXiv preprint arXiv:1608.05182*, 2016.

J. Zhou, L. Yuan, J. Liu, and J. Ye. A multi-task learning formulation for predicting disease progression. In *International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 814–822, 2011.