Methods for Time-Dependent Adjustment of Longitudinal Trajectory Predictions (Working title)

Peter Schulam

August 6, 2015

1 Introduction

Clinical markers indicate the health status of a given organ system, and are important sources of information used to drive management and treatment decisions. Examples of clinical markers include the percent of predicted forced vital capacity (PFVC), which clinicians use to monitor the progression of interstitial lung disease in patients with scleroderma. Accurate predictions of the future trajectory of clinical markers would be a valuable tool, as clinicians could anticipate future complications and treat the disease more aggressively.

One way to predict future course is to construct a generative model of the longitudinally measured marker, estimate the model parameters, and, for a given history of observed measurements, compute the conditional distribution of future measurements under the estimated model. The generative approach to modeling and predicting longitudinally measured outcomes is attractive for a number of reasons. First, clinical data frequently has many missing measurements as observations are usually recorded at irregular time intervals. Second, computing the conditional distribution of future measurements given those observed so far under the joint model is a natural online prediction procedure in that it coherently absorbs new clinical information as it is recorded and updates forecasts. Finally, we can easily compute the conditional distribution over multiple future measurements simultaneously to predict full trajectories.

A drawback of the generative approach is that performance of the model is especially sensitive to the correctness of the assumptions underlying the formulation of the joint distribution. This weakness is particularly important to consider when attempting to incorporate other longitudinally recorded markers into our predictions. We may be able to improve our predictions by incorporating other information that is measured longitudinally. For example, other markers of lung function that are recorded when the patient undergoes a pulmonary function test may contain valuable information for detecting future decline in PFVC. In addition, if health across different organ systems is correlated, then markers measuring the status of, for example, the heart or kidneys may

also be useful for improving predictions about PFVC. Unfortunately, incorporating such markers into a joint generative model is difficult because the statistical dependencies across markers can be complicated and challenging to model. This is especially true when working in complex systemic diseases where the disease is thought to be driven by a number of genetic and environmental factors that are poorly understood.

Contributions. In this work, we consider the problem of predicting the future values of a longitudinally measured outcome using previous observations of that outcome and previous observations of other longitudinally measured outcomes. We refer to the longitudinal process for which we are making predictions as the *target* process, and we refer to the additional longitudinal outcomes used to inform our forecasts as *auxiliary* processes. We use $\boldsymbol{y} \in \mathbb{R}^{|\boldsymbol{y}|}$ to denote vectors of target process measurements observed at times $\boldsymbol{t}$ (we use $|\boldsymbol{y}|$ to denote the number of elements in the vector). Similarly, we assume that there are M auxiliary processes, and for each $m \in \{1, \ldots, M\}$ we denote the vector of observed measurements and times using $\boldsymbol{y}_m$ and $\boldsymbol{t}_m$ respectively. We assume that we are learning from a database of such observation vectors, and use the superscript i to indicate the individual to which a given vector belongs (i.e. $\boldsymbol{y}^i$ is the vector of observed target process measurements for individual i). For a given time t, we use $\boldsymbol{y}[\leq t]$ to denote the vector of observations that have been recorded prior or at time t. Let y^* denote the value of a measurement of the target longitudinal process at some time $t^* > t$, then our goal is to learn a model of the conditional distribution:

$$p(y_* \mid \boldsymbol{y}[\leq t], \boldsymbol{y}_{1:M}[\leq t]), \tag{1}$$

where we have suppressed explicit dependence on the measurement times $\boldsymbol{t}$ and $\boldsymbol{t}_{1:M}$ (and will continue to do so for the remainder of the paper). Our approach builds upon a flexible latent variable model that generalizes many of the approaches to joint analysis of multiple longitudinal outcomes. The primary contribution of our work is the formulation of an estimator for the conditional distribution in Equation 1 that is based upon the flexible joint latent variable model, but is estimated in a way that is robust to misspecification of the dependencies across longitudinal processes.

2 Model

We begin by describing a *general* joint probabilistic model of the target and auxiliary longitudinal processes, which will help to motivate the approach we take to train a predictive model for the future course of the target process. First, we assume that there is a collection of latent random variables associated with each longitudinal process, which we denote using $\boldsymbol{z}$ for the target process and $\boldsymbol{z}_m$ for each of the auxiliary processes. Conditioned on the associated vector of latent variables, the elements of a measurement vector are assumed to be statistically independent of one another:

$$p(\boldsymbol{y} \mid \boldsymbol{z}) = \prod_{j=1}^{|\boldsymbol{y}|} p(y_j \mid \boldsymbol{z}). \tag{2}$$

Note that the distributions for each element of the vector may not be identical because a parameter (such as the mean) may depend on the time at which the observation is measured. This general model describes traditional random effects models, latent class models, and more elaborate hierarchical models of longitudinal data such as [4].

The challenge in moving from a model of a single longitudinal outcome to a collection of outcomes is in specifying the marginal dependencies between the latent variables. In the most general case, no assumptions are made about the way in which the marginal distribution over the latent variables factorizes. For the purposes of motivating our approach, we use this general formulation. The joint distribution over observed marker values and latent random variables can therefore be written as

$$p(\boldsymbol{y}, \boldsymbol{y}_{1:M}, \boldsymbol{z}, \boldsymbol{z}_{1:M}) = p(\boldsymbol{z}, \boldsymbol{z}_{1:M}) p(\boldsymbol{y} \mid \boldsymbol{z}) \prod_{m=1}^{M} p(\boldsymbol{y}_m \mid \boldsymbol{z}_m). \tag{3}$$

Under this joint model, we can write the target conditional distribution in Equation 1 as:

$$p(y_* \mid \boldsymbol{y}[\leq t], \boldsymbol{y}_{1:M}[\leq t]) = \sum_{\boldsymbol{z}} p(y_* \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{y}[\leq t], \boldsymbol{y}_{1:M}[\leq t]). \tag{4}$$

In this conditional distribution, all information from the previously observed measurements from both the target and auxiliary processes is conveyed through the conditional density of the target process latent variables $\boldsymbol{z}$ given the observations in $\boldsymbol{y}[\leq t]$ and $\boldsymbol{y}_{1:M}[\leq t]$. This "bottleneck" suggests a natural strategy: directly estimate a time-dependent conditional probabilistic model of the target latent variables given previously observed measurements from both the target and auxiliary longitudinal processes. We train the model to match the latent variables that best explain the complete set of target measurements without conditioning on auxiliary process measurements. In other words, for each individual i in the training data, the response we fit is

$$\boldsymbol{z}_*^i = \underset{\boldsymbol{z}}{\operatorname{argmax}} \; p(\boldsymbol{y}_i \mid \boldsymbol{z}) p(\boldsymbol{z}). \tag{5}$$

The process of formulating such a conditional model, however, is not straightforward. The main challenge is that the processes can each be measured at different schedules. Moreover, the times at which measurements are recorded within a given process may be irregularly spaced. This creates a missing data problem, and makes it impossible to rely on measurements being available at fixed times that can be used as direct inputs or to extract features for the model. Indeed, this is one of the strengths of the generative approach. Missing observations are easily marginalized and only observed measurements contribute to the conditional distribution. In the joint latent variable model, we can rewrite the conditional distribution over $\boldsymbol{z}$ in Equation 4 using Bayes' rule to see how information from the observed measurements of the auxiliary longitudinal processes are summarized

through the associated latent variables under the general model we've described above:

$$p(\boldsymbol{z} \mid \boldsymbol{y}[\leq t], \boldsymbol{y}_{1:M}[\leq t]) \propto p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) p(\boldsymbol{z} \mid \boldsymbol{y}_{1:M}[\leq t]) \tag{6}$$

$$= p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) \sum_{\boldsymbol{z}_{1:M}} p(\boldsymbol{z} \mid \boldsymbol{z}_{1:M}) p(\boldsymbol{z}_{1:M} \mid \boldsymbol{y}_{1:M}[\leq t])$$

$$\propto p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) \sum_{\boldsymbol{z}_{1:M}} \overbrace{p(\boldsymbol{z}, \boldsymbol{z}_{1:M})}^{\text{latent variable compatibility}} \underbrace{\prod_{m=1}^{M} p(\boldsymbol{y}_m[\leq t] \mid \boldsymbol{z}_m)}_{\text{info. transfer from aux.}}. \tag{7}$$

The final line above sheds some light on how information should be passed from auxiliary markers to the target latent variables. On the far right, we see that evidence for each of the auxiliary latent variables given the observed auxiliary measurements should come from the likelihood function. Interestingly, the parameters of these likelihood functions can be estimated by independently fitting individual joint models over $\boldsymbol{y}_m$ and $\boldsymbol{z}_m$. Second from the right, we see that there should be some form of compatibility function between the set of latent random variables. **The proposed approach will learn this compatibility function discriminatively. Such an estimate should be more robust to misspecification than if an assumed probability distribution over the latent random variables is learned using a generative criterion.**

2.1 Proposed Approach

Our strategy will be to reformulate Equation 7 as a log-linear model. This is easily done by using a log-linear parameterization for the joint model $p(\boldsymbol{z}, \boldsymbol{z}_{1:M})$. Let $g(\boldsymbol{z}, \boldsymbol{z}_{1:M}) \in \mathbb{R}^d$ denote a feature vector, and let $\boldsymbol{w} \in \mathbb{R}^d$ denote a weight vector. We assume that:

$$p_{\boldsymbol{w}}(\boldsymbol{z}, \boldsymbol{z}_{1:M}) \propto \boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M}). \tag{8}$$

Note that the normalization constant $Z(\boldsymbol{w})$ does not depend on $\boldsymbol{z}$ or $\boldsymbol{z}_{1:M}$ and so we can further simplify Equation 7 after assuming the log-linear parameterization:

$$p(\boldsymbol{z} \mid \boldsymbol{y}[\leq t], \boldsymbol{y}_{1:M}[\leq t]) \tag{9}$$

$$\propto p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) \sum_{\boldsymbol{z}_{1:M}} \frac{e^{\boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M})}}{Z(\boldsymbol{w})} \prod_{m=1}^{M} p(\boldsymbol{y}_m \mid \boldsymbol{z}_m) \tag{10}$$

$$\propto p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) \sum_{\boldsymbol{z}_{1:M}} e^{\boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M})} \prod_{m=1}^{M} p(\boldsymbol{y}_m \mid \boldsymbol{z}_m) \tag{11}$$

$$= \sum_{\boldsymbol{z}_{1:M}} \exp \left\{ \log p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) + \sum_{m=1}^{M} \log p(\boldsymbol{y}_m[\leq t] \mid \boldsymbol{z}_m) + \boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M}) \right\}. \tag{12}$$

The computational complexity of inference in the model depends how $\boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M})$ factorizes (the log-likelihood terms in the exponential are single-variable factors and so are easily handled).

The key idea of this formulation is that the *compatibility function* will be more robust to misspecification when trained using the conditional formulation above than if it were trained using a generative criterion. A separate set of parameters $\boldsymbol{w}$ will be learned for each desired prediction time t so that the empirical distribution over histories used when training the model matches the conditions under which it will be deployed.

2.2 Fitting the Model

We fit the model using penalized maximum likelihood. The model is fit for a specific time point t. The objective function is:

$$\Phi_{\text{PML}}(\boldsymbol{w}) = \frac{1}{N}\sum_{i=1}^{N} \log p_{\boldsymbol{w}}(\boldsymbol{z}_*^i \mid \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t]) - \frac{\lambda}{2}\|\boldsymbol{w}\|_2^2. \tag{13}$$

Define

$$Z(\boldsymbol{w}) = \sum_{\boldsymbol{z}=1}^{K}\sum_{\boldsymbol{z}_{1:M}} \exp\left\{\log p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) + \sum_{m=1}^{M} \log p(\boldsymbol{y}_m[\leq t] \mid \boldsymbol{z}_m) + \boldsymbol{w}^\top g(\boldsymbol{z}, \boldsymbol{z}_{1:M})\right\} \tag{14}$$

$$= \sum_{\boldsymbol{z}=1}^{K}\sum_{\boldsymbol{z}_{1:M}} e^{\ell^t(\boldsymbol{z}) + \sum_{m=1}^{M}\ell_m^t(\boldsymbol{z}_m) + \boldsymbol{w}^\top g(\boldsymbol{z},\boldsymbol{z}_{1:M})} = \sum_{\boldsymbol{z}=1}^{K} Z(\boldsymbol{z}, \boldsymbol{w}). \tag{15}$$

then the log-likelihood term for individual i is:

$$\log p_{\boldsymbol{w}}(\boldsymbol{z}_*^i \mid \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t]) = \log Z(\boldsymbol{z}_*^i, \boldsymbol{w}) - \log Z(\boldsymbol{w}). \tag{16}$$

The gradient of the log likelihood for individual i with respect to $\boldsymbol{w}$ is therefore

$$\nabla_{\boldsymbol{w}} \log p_{\boldsymbol{w}}(\boldsymbol{z}_*^i \mid \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t]) = \nabla_{\boldsymbol{w}} \log Z(\boldsymbol{z}_*^i, \boldsymbol{w}) - \log \nabla_{\boldsymbol{w}} Z(\boldsymbol{w}). \tag{17}$$

We first compute the gradient with respect to $\log Z(\boldsymbol{z}_*^i, \boldsymbol{w})$.

$$\nabla_{\boldsymbol{w}} \log Z(\boldsymbol{z}_*^i, \boldsymbol{w}) = \frac{1}{Z(\boldsymbol{z}_*^i, \boldsymbol{w})}\sum_{\boldsymbol{z}_{1:M}} e^{\ell^t(\boldsymbol{z}) + \sum_{m=1}^{M}\ell_m^t(\boldsymbol{z}_m) + \boldsymbol{w}^\top g(\boldsymbol{z},\boldsymbol{z}_{1:M})} g(\boldsymbol{z}_*^i, \boldsymbol{z}_{1:M}) \tag{18}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left[g(\boldsymbol{z}_*^i, \boldsymbol{Z}_{1:M}) \mid \boldsymbol{z}_*^i, \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t]\right]. \tag{19}$$

Similarly, the gradient with respect to $\log Z(\boldsymbol{w})$ is:

$$\nabla_{\boldsymbol{w}} \log Z(\boldsymbol{w}) = \frac{1}{Z(\boldsymbol{w})}\sum_{\boldsymbol{z}=1}^{K}\sum_{\boldsymbol{z}_{1:M}} e^{\ell^t(\boldsymbol{z}) + \sum_{m=1}^{M}\ell_m^t(\boldsymbol{z}_m) + \boldsymbol{w}^\top g(\boldsymbol{z},\boldsymbol{z}_{1:M})} g(\boldsymbol{z}, \boldsymbol{z}_{1:M}) \tag{20}$$

$$= \mathbb{E}_{\boldsymbol{w}}\left[g(\boldsymbol{Z}^i, \boldsymbol{Z}_{1:M}^i) \mid \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t]\right]. \tag{21}$$

The objective function therefore has the following gradient with respect to $\boldsymbol{w}$:

$$\nabla_{\boldsymbol{w}} \Phi_{\text{PML}}(\boldsymbol{w}) = \tag{22}$$

$$\frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{w}} \left[g(\boldsymbol{z}_*^i, \boldsymbol{Z}_{1:M}) \mid \boldsymbol{z}_*^i, \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t] \right] \tag{23}$$

$$- \frac{1}{N} \sum_{i=1}^{N} \mathbb{E}_{\boldsymbol{w}} \left[g(\boldsymbol{Z}^i, \boldsymbol{Z}_{1:M}^i) \mid \boldsymbol{y}^i[\leq t], \boldsymbol{y}_{1:M}^i[\leq t] \right] \tag{24}$$

$$- \lambda \boldsymbol{w} \tag{25}$$

We see that the key computational step in computing the gradient is computing the expectation of the feature vector under the current set of parameters. The number of terms in the sum of the expectation is exponential in the number of auxiliary markers, and so it is important that $g(\boldsymbol{z}, \boldsymbol{z}_{1:M})$ factorizes in a way that allows us to effectively use dynamic programming.

3 Connections to Previous Methods

We have explored two approaches that fit within the conceptual framework of estimating the left hand side of Equation 6. In the first approach, we additively adjust the single-marker model posterior over the target latent variables $\boldsymbol{z}$:

$$p(\boldsymbol{z} \mid \boldsymbol{y}[\leq t]) \propto p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z}) p(\boldsymbol{z}). \tag{26}$$

In this approach, we assume that the latent variable $\boldsymbol{z}$ is a single categorical outcome taking values in $\{1, \ldots, K\}$. We parameterize the adjustment by first casting the single-marker posterior in Equation 26 as a multinomial logistic regression model. In multinomial logistic regression, we predict a probability distribution over K outcomes by transforming a set of scores $\{s_1, \ldots, s_K\}$ using the softmax function. the estimated probability of outcome $\boldsymbol{z} = k$ is then:

$$p(\boldsymbol{z} = k) = \frac{e^{s_k}}{\sum_{k'=1}^{K} e^{s_{k'}}}. \tag{27}$$

The single-marker posterior can be reformulated in terms of these scores. We have that

$$p(\boldsymbol{z} = k \mid \boldsymbol{y}[\leq t]) = \frac{p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z} = k) p(\boldsymbol{z} = k)}{\sum_{k'=1}^{K} p(\boldsymbol{y}[\leq t] \mid \boldsymbol{z} = k') p(\boldsymbol{z} = k')} \tag{28}$$

$$= \frac{e^{\log p(\boldsymbol{y}[\leq t] | \boldsymbol{z} = k) + \log p(\boldsymbol{z} = k)}}{\sum_{k'=1}^{K} e^{\log p(\boldsymbol{y}[\leq t] | \boldsymbol{z} = k') + \log p(\boldsymbol{z} = k')}}. \tag{29}$$

We therefore have that

$$s_{ik} = \log p(\boldsymbol{y}^i[\leq t] \mid \boldsymbol{z}^i = k) + \log p(\boldsymbol{z}^i = k). \tag{30}$$

Assuming that the single-marker generative model on the right hand side of Equation 26 is correct and given only the previously observed measurements of the target process $\boldsymbol{y}^i[\leq t]$, choosing the

subtype with the maximum score (i.e. $\operatorname{argmax}_k s_{ik}$) gives the Bayes' optimal decision rule for 0-1 loss. Once we condition on the observed measurements from the auxiliary processes, however, these scores are no longer optimal. One strategy for accounting for the new evidence is to adjust these scores.

Method 1. In Method 1 we adjust the scores additively

$$s'_{ik} = s_{ik} + \sum_{m=1}^{M} \boldsymbol{\theta}_m^\top \mathbb{E}[\phi(k, \boldsymbol{Z}_m^i) \mid \boldsymbol{y}_m^i]. \tag{31}$$

These scores comprise a multinomial log-linear model over target latent variables. There are two key elements to note here. First, only pairwise interactions between the target latent variable and each auxiliary latent variable are included in the model. Second, we use the expected value of the feature function $\phi(k, \boldsymbol{Z}_m^i)$ taken with respect to the posterior given $\boldsymbol{y}_m^i$ using the auxiliary marker's marginal model (that is, the joint distribution over measurements and latent variables trained in isolation). Note that the expectation is assumed to be independent of the target latent variable $\boldsymbol{z}^i$ and all other auxiliary random variables $\boldsymbol{z}_{m'}^i$ for $m' \neq m$.

Method 2. In Method 2, we train M separate conditional models using scores identical to the components of the sum in Equation 31. We combine the single-marker posterior (Eq. 26) and the M additional conditional models in a mixture to produce the adjusted posterior probabilities over target latent variables. The scores for each of the M separate conditional models are parameterized as:

$${}_m s'_{ik} = \boldsymbol{\theta}_m^\top \mathbb{E}[\phi(k, \boldsymbol{Z}_m^i) \mid \boldsymbol{y}_m^i]. \tag{32}$$

In this case, we see that the expectation depends only on $\boldsymbol{y}_m^i$ and so the model makes the implicit assumption that $\boldsymbol{z}_m^i$ is independent of $\boldsymbol{z}^i$ given $\boldsymbol{y}_m^i$. In contrast to Method 1, however, it does not make the assumption that $\boldsymbol{z}_m^i$ is independent of $\boldsymbol{y}_{m'}^i$ for $m' \neq m$. This assumption is not made because the M models are trained independently (that is, not conditioned on the measurements from other auxiliary markers). The probability of the target latent variable given $\boldsymbol{y}_m^i$ is then:

$$p(\boldsymbol{z}^i = k \mid \boldsymbol{y}_m^i) = \frac{e^{{}_m s'_{ik}}}{\sum_{k'=1}^{M} e^{{}_m s'_{ik}}}. \tag{33}$$

The M models are trained independently. Given the parameters for each of the models, the weights $\boldsymbol{\pi}$ are fit to maximize the likelihood of the MAP configuration of the latent variables are fit using maximum likelihood estimate, which produces the final estimate:

$$p(\boldsymbol{z}^i = k \mid \boldsymbol{y}^i, \boldsymbol{y}_{1:M}^i) = \underbrace{\boldsymbol{\pi}[1] p(\boldsymbol{z}^i = k \mid \boldsymbol{y}^i)}_{\text{original posterior}} + \sum_{m=1}^{M} \underbrace{\boldsymbol{\pi}[m+1] p(\boldsymbol{z}^i = k \mid \boldsymbol{y}_m^i)}_{\text{independent auxiliary predictions}}. \tag{34}$$

4 Related Work

The dominant approach to modeling longitudinal data is the mixed effects or hierarchical regression framework [1]. In a mixed effects model, a population regression model specifies the average trajectory. An individual trajectory is marginally centered around the population, but each individual is associated with a collection of latent random variables that parameterize an individual-specific adjustment to the population mean. Marginally, this induces correlation between observations from the same individual. Conditioned on the values of the latent variables, this produces an individual-specific mean trajectory.

Recently, there has been increased interest in jointly modeling multiple longitudinal outcomes (see e.g. [5, 2] for recent reviews). A common approach to jointly modeling longitudinal outcomes builds on the single-outcome mixed effects model. Each longitudinal outcome has a population mean trajectory and individuals have latent variables for each outcome that adjust the population trajectory. The dependence across outcomes is then determined by a joint distribution over the latent variables from each longitudinal process. This approach, while flexible, leads to complex models that are difficult to estimate. One can simplify the model by carefully designing the dependency structure across markers, but this requires a detailed understanding of the underlying phenomenon, which may not always be available. Another approach is to use latent class models (e.g. [3]), which can reduce the number of parameters while still maintaining a relatively flexible dependency structure. **When we are interested only in using other longitudinally measured outcomes to improve predictions about a target outcome (e.g. PFVC), we suspect that assumptions about cross-outcome dependencies can be minimized because we are not seeking to describe the full generative process.** This note sketches a conceptual argument for two of methods for training online adjustments of a generative model to improve predictive accuracy of future longitudinal trajectories.

References

[1] Peter Diggle, Patrick Heagerty, Kung-Yee Liang, and Scott Zeger. *Analysis of longitudinal data*. Oxford University Press, 2002.

[2] S Fieuws, Geert Verbeke, and G Molenberghs. Random-effects models for multivariate repeated measures. *Statistical methods in medical research*, 16(5):387–397, October 2007.

[3] Hein Putter, Tineke Vos, Hanneke de Haes, and Hans van Houwelingen. Joint analysis of multiple longitudinal outcomes: application of a latent class model. *Statistics in medicine*, 27(29):6228–6249, 20 December 2008.

[4] Peter Schulam, Fredrick Wigley, and Suchi Saria. Clustering longitudinal clinical marker trajectories from electronic health data: Applications to phenotyping and endotype discovery. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*. people.ee.duke.edu, 2015.

[5] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. The analysis of multivariate longitudinal data: a review. *Statistical methods in medical research*, 23(1):42–59, February 2014.

[6] Hongxia Yang, Fan Li, Enrique F Schisterman, Sunni L Mumford, and David Dunson. Bayesian inference on dependence in multivariate longitudinal data. 14 August 2012.