Learning Adaptive Forecasting Models from Irregularly Sampled Multivariate Clinical Data

Zitao Liu and **Milos Hauskrecht**
Computer Science Department, University of Pittsburgh, 210 South Bouquet St., Pittsburgh, PA, 15260 USA

Abstract

Building accurate predictive models of clinical multivariate time series is crucial for understanding of the patient condition, the dynamics of a disease, and clinical decision making. A challenging aspect of this process is that the model should be flexible and adaptive to reflect well patient-specific temporal behaviors and this also in the case when the available patient-specific data are sparse and short span. To address this problem we propose and develop an adaptive two-stage forecasting approach for modeling multivariate, irregularly sampled clinical time series of varying lengths. The proposed model (1) learns the population trend from a collection of time series for past patients; (2) captures individual-specific short-term multivariate variability; and (3) adapts by automatically adjusting its predictions based on new observations. The proposed forecasting model is evaluated on a real-world clinical time series dataset. The results demonstrate the benefits of our approach on the prediction tasks for multivariate, irregularly sampled clinical time series, and show that it can outperform both the population based and patient-specific time series prediction models in terms of prediction accuracy.

Introduction

With a wide adoption and availability of electronic health records (EHRs), the development of models of clinical multivariate time series (MTS) and tools for their analysis is becoming increasingly important for meaningful applications of EHRs in computer-based patient monitoring, adverse event detection, and improved patient management. (Bellazzi et al. 2000; Clifton et al. 2013; Lasko, Denny, and Levy 2013; Liu and Hauskrecht 2013; Liu, Wu, and Hauskrecht 2013; Schulam, Wigley, and Saria 2015; Ghassemi et al. 2015; Durichen et al. 2015).

In general, a number of models representing various time series data and their behaviors exist (Hamilton 1994). However, modeling of clinical time series data still presents numerous challenges that come from special characteristics of clinical data (Liu and Hauskrecht 2015). Briefly, clinical time series are distinguished from other time series data due to the following characteristics:

- *multiple variables*: the real-world clinical dynamics are multivariate and they often exhibit interactions and comovements among different time series.
- *irregular samples*: sequential observations are collected at different times, and the time elapsed between two consecutive observations may vary.
- *length variability*: the number of observations in each data sequence is limited and the duration they span may vary a lot from patient to patient.

The objective of this work is to study and develop models that can be used for accurate clinical time series forecasting. More specifically we are interested in developing models and methods that can predict future values of MTS for a patient given a history of past observations. This problem is rather challenging for two reasons (1) the time series of past observations made for a patient of interest may be relatively short so it may be very hard to learn a good time series model just from one patient data; (2) the patient-to-patient variability may be large so it is unclear if the population based model derived from many samples of different patients will be sufficient to support the predictions. The majority of existing approaches in the literature tackle the clinical time series forecasting problem by taking one of the two "extreme" approaches: they either build a population based model or a patient-specific model ignoring what is known about the population. In this work we seek to develop a new approach that aims to benefit from the population trend extracted from past data collection and at the same time adapt to patient-specific data, thus allowing one to make more accurate MTS predictions.

We propose and develop a new two-stage adaptive forecasting model to represent both the population and the patient-specific multivariate interactions of clinical MTS. In the first stage, we learn a population model from clinical MTS sequences from many different patients. In this paper we use and experiment with a linear dynamical system (LDS) (Kalman 1960) whose parameters are learned with the help of the EM algorithm. In the second stage, we first express the time series of past observations for a patient in terms of residuals (or differences in between predictions made by the population model and actually observed values), which reflect the patient-specific deviations from the population model. Then we use and model these deviations with a multi-task Gaussian process (MTGP) (Bonilla, Chai, and Williams 2007). In the forecasting phase, we automatically adjust the predictions from the population model based on the new patient-specific observations.

Overall this paper makes the following contributions:

- It presents a new two-stage model to represent the multivariate, irregularly sampled clinical time series, which not only represents the long-term population trend of the dynamics, but captures the multivariate interactions for each patient-specific dynamics.
- The new model is able to automatically adapt its predictions according to the newly observed data for each individual patient without retraining the population-based model.
- We evaluate our approach and its benefits on a real-world clinical MTS dataset.

The remainder of the paper is organized as follows: *Background* Section introduces the basics of the LDS, GP and MTGP models for time series modeling. Comparison to relevant research work is discussed. In *Methodology* Section, we describe the details of our two-stage adaptive forecasting model, which consist of a population model learned from the entire collection of sequences and a model of the multivariate temporal interactions based on MTGP. In *Experiment* Section, we (1) visualize the predictions made by our forecasting model; (2) show the benefits of our model over alternative approaches on a clinical data derived from the Complete Blood Count panel. We summarize our work and outline possible future extensions in *Conclusion* Section.

Background

In this section, we review the basics of three models widely used to represent time series data: the linear dynamical system (LDS), Gaussian process (GP) and multi-task Gaussian process (MTGP). After that, we discuss the differences between our model and existing approaches.

Notation

We denote clinical MTS data $\mathcal{D}$ as a collection of observed value set $\mathcal{Y}$ and the time stamp set $\mathcal{X}$ which contains observations and time stamps, i.e, $\mathcal{D} = (\mathcal{Y}, \mathcal{X})$. We denote a time series data set with N samples as $\mathcal{D} = \{\mathbf{D}^1, \mathbf{D}^2, \cdots, \mathbf{D}^N\}$ and correspondingly, we have $\mathcal{Y} = \{\mathbf{Y}^1, \mathbf{Y}^2, \cdots, \mathbf{Y}^N\}$, $\mathcal{X} = \{\mathbf{X}^1, \mathbf{X}^2, \cdots, \mathbf{X}^N\}$ and $\mathbf{D}^l =< \mathbf{Y}^l, \mathbf{X}^l >$, $l = 1, 2, \cdots, N$.

Without loss of generality, for each MTS sequence $\mathbf{D}^l$, we assume it has n dependent time series with the same length T_l. Hence, we represent $\mathbf{Y}^l$ as an $n \times T_l$ matrix. Let $\mathbf{y}^l_{i,:}$ and $\mathbf{y}^l_t$ be the *i*th row and *t*th column of $\mathbf{Y}^l$. Let y^l_{it} be the *t*th observation in the *i*th time series in $\mathbf{Y}^l$. In this work, we assume time series within each sample $\mathbf{D}^l$ are obtained at the same time stamps and $\mathbf{X}^l$ can be concisely represented as a $T_l \times 1$ vector $\mathbf{x}^l$. Let x^l_t be the *t*th time stamp in $\mathbf{x}^l$.

Let $\mathbb{E}_{\mathbf{z}}[f(\cdot)]$ denote the expected value of $f(\cdot)$ with respect to $\mathbf{z}$. For both vectors and matrices, the superscript $(\cdot)^\top$ denotes the transpose. Let $\otimes$ denote the Kronecker product. For the sake of notational brevity, we omit the explicit sample index (*l*) in the rest of *Background* section.

Linear Dynamical System

The linear dynamical system (LDS) models real-valued MTS $\{\mathbf{y}_t \in \mathbb{R}^n\}_{t=1}^T$ using hidden states $\{\mathbf{z}_t \in \mathbb{R}^d\}_{t=1}^T$:

$$\mathbf{z}_t = A\mathbf{z}_{t-1} + \boldsymbol{\epsilon}_t; \quad \mathbf{y}_t = C\mathbf{z}_t + \boldsymbol{\zeta}_t \tag{1}$$

Briefly, $\{\mathbf{z}_t\}$ is generated via the transition matrix $A \in \mathbb{R}^{d\times d}$. Observations $\{\mathbf{y}_t\}$ are generated from $\mathbf{z}_t$ via the emission matrix $C \in R^{n\times d}$ (see eq.(1)). $\{\boldsymbol{\epsilon}_t\}_{t=1}^T$ and $\{\boldsymbol{\zeta}_t\}_{t=1}^T$ are i.i.d. multivariate normal distributions with mean $\mathbf{0}$ and covariance matrices Q and R respectively. The initial state ($\mathbf{z}_1$) distribution is also multivariate normal with mean $\boldsymbol{\xi}$ and covariance matrix Ψ. The complete set of the LDS parameters is $\Omega = \{A, C, Q, R, \boldsymbol{\xi}, \Psi\}$. The LDS is arguably the most commonly used time series model for real-world engineering and financial applications, such as time series prediction (Rogers, Li, and Russell 2013; Li et al. 2009) and visual tracking (Lee, Kim, and Kweon 1995; Funk 2003).

Gaussian Process

The Gaussian process (GP) is a popular nonparametric nonlinear Bayesian model in statistical machine learning (Rasmussen 2006). In time series modeling, each GP is used to model an individual time series, which is represented by the mean function $m(x_t) = \mathbb{E}[f(x_t)]$ and the covariance function $K^G(x_t, x_{t'}) = \mathbb{E}[(f(x_t) - m(x_t))(f(x_{t'}) - m(x_{t'}))]$, where $f(x)$ is a real-valued process and x_t and $x_{t'}$ are two time stamps. The GP can be used to calculate the posterior distribution $p(f(\mathbf{x}^*)|(\mathbf{x}, \mathbf{y}_{i,:}))$ of f values for an arbitrary set of time stamps $\mathbf{x}^*$, given a set of observed values $\mathbf{y}_{i,:}$ from time series *i* at time stamps $\mathbf{x}$.

Due to the ability of exact inference, GP based model are widely used in time series regression and forecasting tasks, where time stamps are modeled as the input of GP and observations are modeled through the predicted mean function of the time series (Stegle et al. 2008; Clifton et al. 2013; Lasko, Denny, and Levy 2013; Liu and Hauskrecht 2014).

Multi-task Gaussian Process

The multi-task Gaussian process (MTGP) is an extension of GP which models multiple tasks (e.g., multivariate time series) simultaneously by utilizing the learned covariance between related tasks. MTGP uses K^C to model the similarities between tasks and uses K^G to capture the temporal dependence with respect to time stamps. The covariance matrix of MTGP is shown as follows:

$$K^M = K^C \otimes K^G + D \otimes I_T \tag{2}$$

where K^C is a positive semi-definite (PSD) matrix that specifies the inter-task similarities and K^C_{ij} measures the similarity between task i and task j. D is an $n \times n$ diagonal matrix in which D_{ii} is the noise variance δ^2_i for the *i*th task.

Exact inference of MTGP can be done by using the standard GP formulations and the details can be found in

(Bonilla, Chai, and Williams 2007) and its clinical application can be found in (Ghassemi et al. 2015; Durichen et al. 2015).

Related Work

The majority of existing work on clinical time series forecasting models each clinical time series separately (Marlin et al. 2012; Clifton et al. 2013; Lasko, Denny, and Levy 2013; Liu and Hauskrecht 2014; Schulam, Wigley, and Saria 2015) which does not allow one to represent dependences among the different time series. Our model deals with multivariate data and aims to capture interactions among all variables and their dynamics.

The works that try to capture MTS and dependences among its time series include (Ghassemi et al. 2015; Durichen et al. 2015). In (Ghassemi et al. 2015; Durichen et al. 2015) the authors apply MTGP to clinical MTS modeling and forecasting. However, their models that are learned from time series for just one patient are very simple and results in either constant or simple parametric mean functions. This is too restrictive to represent real clinical MTS with large variability. In addition, this approach does not take advantage of time series collected for other patients. Our approach tackles the problem in two stages and with a combination of two models: we first use an LDS to model the population trend, and then take advantage of MTGP to capture the individual-specific short-term variability. (Fox et al. 2011) utilizes the beta process to build the joint model for multiple related time series (not necessarily clinical). It requires intensive MCMC posterior computations, which is often infeasible to do in real clinical settings.

Finally, we would like to note that the majority of methods mentioned above fail to generalize to forecasting models for a collection of clinical MTS of varying lengths. They are not able to adapt their forecasts when new values are observed without retraining the model.

Methodology

In this section, we propose a two-stage model that (1) is learned from a collection of past patient time series data of varying lengths; (2) captures the patient-specific short-term multivariate interactions; (3) automatically adjusts its predictions when new observations for the target patient are obtained without retraining the population model.

Stage 1: Learning A Population Model

In the first stage, we would like to learn a population model from all available data sequences to represent the trend of the entire population. We choose an LDS model to model the population trend, which is a classical and widely used discrete-time model for real-valued sequence analysis. The LDS is Markovian and assumes the dynamic behavior of the system is captured well using a small set of real-valued hidden-state variables and linear state transitions corrupted by a Gaussian noise. It has a mathematically predictable behavior, and both exact inference and predictions for LDS models can be done efficiently.

Direct Value Interpolation In spite of the advantages of LDS models, they are restricted to discrete time domain where observations are regularly sampled. In order to apply the discrete-time LDS model over our irregularly sampled clinical data, we follow (Adorf 1995; Dezhbakhsh and Levy 1994; Åström 1969; Bellazzi et al. 1995; Kreindler and Lumsden 2006; Rehfeld et al. 2011; Liu and Hauskrecht 2014) and apply the *direct value interpolation* (DVI) technique to discretize each irregularly sampled clinical sequence and that replaces it with a regularly sampled time series data.

The DVI approach assumes that all observations are collected regularly with a pre-specified sampling frequency r. However, instead of actual readings, the values at these time points are estimated from readings at time points closest to them using various interpolation techniques. The interpolated (regular) time series, i.e., $\tilde{\mathbf{y}}^l_{i,:}$, is then used to train a discrete-time LDS model. We put a tilde sign $(\tilde{\cdot})$ over $\mathcal{Y}$, $\mathbf{Y}^l$, $\mathbf{y}^l_{i,:}$ and $\mathbf{y}^l_t$ to indicate the discretized observations. $\tilde{T}_l$ is the length of discretized sequence for patient l. The approach is illustrated in Figure 1.

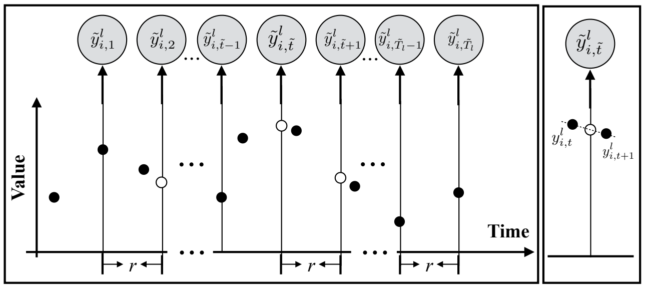
$\tilde{y}^l_{i,1}$
$\tilde{y}^l_{i,2}$
$\tilde{y}^l_{i,\tilde{t}-1}$
$\tilde{y}^l_{i,\tilde{t}}$
$\tilde{y}^l_{i,\tilde{t}+1}$
$\tilde{y}^l_{i,\tilde{T}_l-1}$
$\tilde{y}^l_{i,\tilde{T}_l}$
Value
Time
r
r
r
$\tilde{y}^l_{i,\tilde{t}}$
$y^l_{i,t}$
$y^l_{i,t+1}$

Figure 1: Transformation of an irregularly sampled time series $\mathbf{y}^l_{i,:}$ to a discrete time series $\tilde{\mathbf{y}}^l_{i,:}$ by DVI. The empty circles denote the interpolated values with no readings. The right panel illustrates the linear interpolation process.

A possible limitation of the DVI data transformation is possible information loss: as we can see from Figure 1, some observations in individual time series are discarded during this discretization process. However, given that LDS is building a coarse level population model over the entire collection of data (many patients), this loss is less important. We also note that patient-specific observations are not discarded in the second stage of our approach that captures fine grained patient-specific multivariate interactions by MTGP.

EM Learning In order to learn the unified population model over the entire discretized clinical sequences, we build our model upon the probabilistic formulation of the LDS model and follow the EM learning algorithm proposed by (Ghahramani and Hinton 1996). We extend its formulation to multiple sequences setting. The log joint probability distribution of the LDS model over the entire collection of clinical sequences of varying lengths is:

$$\log\Big(p(\tilde{\mathcal{Y}},\mathcal{Z})\Big) = \sum_{l=1}^{N}\log p(\mathbf{z}_1^l) + \sum_{l=1}^{N}\sum_{t=1}^{\tilde{T}_l}\log p(\tilde{\mathbf{y}}_t^l|\mathbf{z}_t^l) + \sum_{l=1}^{N}\sum_{t=2}^{\tilde{T}_l}\log p(\mathbf{z}_t^l|\mathbf{z}_{t-1}^l) \quad (3)$$

where $\mathcal{Z} = \{\mathbf{Z}^1, \mathbf{Z}^2, \cdots, \mathbf{Z}^N\}$ and $\mathbf{Z}^l$ is the corresponding hidden state sequence of $\tilde{\mathbf{Y}}^l$.

E-Step Since the hidden state Markov chain collection $\mathcal{Z}$ defined by the LDS is unobserved, we cannot learn the LDS directly. Instead, we infer the hidden state expectations. The E-step infers a posterior distribution of latent states $\mathcal{Z}$ given the observation sequences $\tilde{\mathcal{Y}}$, $p(\mathcal{Z}|\tilde{\mathcal{Y}})$.

The E-step requires computing the expected log likelihood of the log joint probability with respect to the hidden state distribution, i.e., $\mathcal{Q} = \mathbb{E}_{\mathcal{Z}}[\log p(\mathcal{Z},\tilde{\mathcal{Y}})]$, which depends on three types of sufficient statistics $\mathbb{E}[\mathbf{z}_t^l|\tilde{\mathbf{Y}}^l]$, $\mathbb{E}[\mathbf{z}_t^l(\mathbf{z}_t^l)^\top|\tilde{\mathbf{Y}}^l]$ and $\mathbb{E}[\mathbf{z}_t^l(\mathbf{z}_{t-1}^l)^\top|\tilde{\mathbf{Y}}^l]$. Here we follow the backward algorithm in (Ghahramani and Hinton 1996) to compute them. The backward algorithm is presented in Section A2 in the supplemental material.

$$\mathcal{Q} = \sum_{l=1}^{N}\mathbb{E}_{\mathbf{Z}^l}\Big[\log p(\mathbf{z}_{:,1}^l)\Big] + \sum_{l=1}^{N}\sum_{t=1}^{\tilde{T}_l}\mathbb{E}_{\mathbf{Z}^l}\Big[\log p(\tilde{\mathbf{y}}_t^l|\mathbf{z}_t^l)\Big] + \sum_{l=1}^{N}\sum_{t=2}^{\tilde{T}_l}\mathbb{E}_{\mathbf{Z}^l}\Big[\log p(\mathbf{z}_t^l|\mathbf{z}_{t-1}^l)\Big] \quad (4)$$

M-Step In the M-step, we try to find Ω that maximizes the likelihood lower bound $\mathcal{Q}$ (eq.(4)). As we can see, $\mathcal{Q}$ function is differentiable with respect to ($\Omega = \{A, C, R, Q, \boldsymbol{\xi}, \Psi\}$). Each of these parameters is estimated similarly to (Ghahramani and Hinton 1996) by taking the corresponding derivative of the eq.(4), setting it to zero, and by solving it analytically. Due the the space limit, update rules for $A, C, R, Q, \boldsymbol{\xi}, \Psi$ are listed in Section A3 in the supplemental material.

Stage 2: Learning Multivariate Interaction Models

A population model built from a collection of clinical data for multiple patients is crucial since each individual sequence is usually very short. The learned model from the entire population is more robust and stable. However, the prediction task is performed patient by patient and the forecasting model should also reflect and take into account the variations specific to the current patient. To address this problem, we model the patient-specific multivariate interactions by using an MTGP. More specifically, instead of simply modeling the clinical time series trends (the mean function of MTGP) by using constants or simple known parametric forms (e.g., linear functions) (Ghassemi et al. 2015; Durichen et al. 2015), we use the population model (learned in Stage 1) to reflect the time series tendency and build an MTGP on a residual signal that reflects the deviations of patients' true observations and the predictions made by the population LDS model. We define the *multivariate residual time series* as follows:

Definition 1. (MULTIVARIATE RESIDUAL TIME SERIES) For each patient l, given time series $\mathbf{Y}^l$ and its corresponding predictions $\hat{\mathbf{Y}}^l$ from model Ω, a multivariate residual time series $\mathbf{R}^l$ represents the deviations from $\mathbf{Y}^l$ to $\hat{\mathbf{Y}}^l$, i.e., $\mathbf{R}^l = \mathbf{Y}^l - \hat{\mathbf{Y}}^l$.

Notice that each residual time series $\mathbf{R}^l$ is computed by using the true observations $\mathbf{Y}^l$ (not the discretized sequence $\tilde{\mathbf{Y}}^l$), there is no information loss for each patient under the prediction task and $\mathbf{R}^l$ is irregularly sampled.

The multivariate residual time series reflect each patient's unique variations from the general population and they are distinguished patient by patient. Furthermore, clinical events usually only affect a handful of measurements within a small time window. Hence, for each patient l, we model these transient deviations nonparametrically using an MTGP. The MTGP has mean $\mathbf{0}$ and a squared exponential covariance function (eq.(2)), which is the most frequently-used example in literature (Rasmussen 2006). In eq.(2), K^G is defined as follows:

$$K^G(x_{i,t}, x_{j,t'}) = \alpha\exp\Big(-\frac{(x_{i,t} - x_{j,t'})^2}{2\beta^2}\Big) \quad (5)$$

The complete parameter set Λ in the MTGP model is $\Lambda = \{\alpha, \beta, \delta_i, K^C\}$ where $i = 1, \cdots, n$. In this work, we adopt the Cholesky decomposition and the "free-form" parameterization techniques ($K^C = LL^\top$) to learn the parameter set Λ by minimizing the negative log marginal likelihood via gradient descent (Rasmussen 2006; Ghassemi et al. 2015).

Usually the MTGP model has the computation limitation that it has $\mathcal{O}(n^3T^3)$ compared with $n \times \mathcal{O}(T^3)$ for standard GP models (T is the length of the time series). However, this limitation is not as relevant in our application setting, given that the number of clinical observations is very limited and clinical time series are usually short span.

Adaptive Prediction

In the real clinical setting, a successful forecasting model needs to be *adaptive*, that is, when newly observed values are obtained, the model should efficiently adapt to the new change and utilize new values to make better predictions. In this work, we develop a new adaptive prediction algorithm based on the Kalman filtering algorithm (Kalman 1960) that utilizes our two-stage forecasting model.

Let u denote the current patient we consider in our prediction task. $\mathbf{Y}^u$ is an $n \times T_u$ matrix which denotes the current observed values for patient *u*. Given an arbitrary future time stamp $t*$ ($t* > T_u$), the value $\hat{\mathbf{y}}_{t*}^u$ is predicted as follows:

Step 1. Compute the discretized observations $\tilde{\mathbf{Y}}^u$ by using DVI on $\mathbf{Y}^u$.

Step 2. Infer patient-specific hidden dynamics by using population model Ω and $\tilde{\mathbf{Y}}^u$. This step *adaptively* computes the patient-specific hidden state $\mathbf{Z}^u$ using patient's latest observations. Details are provided in Section A1 in the supplemental material[1].

Step 3. Make predictions by using the population model Ω and $\mathbf{Z}^u$. Note that we need to predict the value at time points closest to the target time $t*$, and after that, apply the interpolation approach to estimate the target value. The prediction made by the population model is $\hat{\mathbf{y}}^u_{t*}(\Omega)$

Step 4. Use the population model to predict patient *u*'s known observations ($\mathbf{Y}^u$) adaptively, denoting as $\hat{\mathbf{Y}}^u$. Compute the residual time series for patient *u*, i.e., $\mathbf{R}^u = \mathbf{Y}^u - \hat{\mathbf{Y}}^u$.

Step 5. Learn the MTGP model Λ^u from $\mathbf{R}^u$ to capture the patient-specific short-term variability.

Step 6. Predict patient-specific short-term variability $\hat{\mathbf{y}}^u_{t*}(\Lambda^u)$ by using Λ^u at the target time $t*$.

Step 7. Compute the final prediction $\hat{\mathbf{y}}^u_{t*}$ by combining $\hat{\mathbf{y}}^u_{t*}(\Omega)$ and $\hat{\mathbf{y}}^u_{t*}(\Lambda^u)$, i.e., $\hat{\mathbf{y}}^u_{t*} = \hat{\mathbf{y}}^u_{t*}(\Omega) + \hat{\mathbf{y}}^u_{t*}(\Lambda^u)$.

Summary

Algorithm 1 summarizes our two-stage adaptive forecasting model and its learning and prediction parts.

Algorithm 1 Learning and Prediction Procedures

```
INPUT:
• Train data collection $\mathcal{D} = \{< \mathbf{Y}^l, \mathbf{x}^l >\}$, where $l = 1, \cdots, N$.
• DVI sampling frequency $r$.
• Number of hidden states in LDS $d$.
• Current observations $\mathbf{Y}^u$ for patient u who is being predicted.
• An arbitrary future time stamp $t*$ ($t* > T_u$).
PROCEDURE:
// Stage1: Learning the population model.
$\{\tilde{\mathbf{Y}}^l\} = DVI(\{\mathbf{Y}^l\}, \{\mathbf{x}^l\}, r)$.
$\Omega = LearnLDS(\{\tilde{\mathbf{Y}}^l\})$.
// Stage2: Learning the multivariate interaction model.
Compute residual time series $\mathbf{R}^u$.
$\Lambda^u = LearnMTGP(\mathbf{R}^u)$.
// Adaptive Prediction: Predicting $\hat{\mathbf{y}}^u_{t*}$ by $\Omega$ and $\Lambda^u$.
Trend prediction: $\hat{\mathbf{y}}^u_{t*}(\Omega) = PredictLDS(\Omega, t*)$.
Variability prediction: $\hat{\mathbf{y}}^u_{t*}(\Lambda^u) = PredictMTGP(\Lambda^u, t*)$.
$\hat{\mathbf{y}}^u_{t*} = \hat{\mathbf{y}}^u_{t*}(\Omega) + \hat{\mathbf{y}}^u_{t*}(\Lambda^u)$.
OUTPUT: Prediction at time stamp $t_*$: $\hat{\mathbf{y}}^u_{t*}$.
```

Experimental Evaluation

In this section we evaluate our approach on a real-world clinical dataset. We demonstrate the benefits our adaptive approach both (1) qualitatively by visualizing time series predictions made for one of the patients, and (2) quantitatively by comparing the prediction accuracy of our two-stage adaptive forecasting model to alternative approaches. We would also like to note that the hyper parameters (e.g., DVI sampling frequency r, number of hidden states in LDS d) used in our methods are selected (in all experiments) by the internal cross validation approach while optimizing models' predictive performance.

Clinical Data

We test our two-stage adaptive model on a clinical MTS data obtained from EHRs of post-surgical cardiac patients in PCP database (Hauskrecht et al. 2010; 2013). We take 500 patients from the database who had their *Complete Blood Count* (CBC) tests [2] done during their hospitalization. The MTS data consists of six individual CBC lab time series: mean corpuscular hemoglobin concentration(MCHC), mean corpuscular hemoglobin(MCH), mean corpuscular volume(MCV), mean platelet volume(MPV), red blood cell(RBC) and red cell distribution width(RDW). In the following experiments, we have randomly selected 100 patients out of 500 as a test set and used the remaining 400 patients for training the models.

Baselines

We compare our proposed approach (LDS+reMTGP) to the following methods. Some of these are widely used in both clinical pharmacology and machine learning communities:

- Mean of the entire population (P_Mean).
- Mean of each individual patient (I_Mean).
- Gaussian process regression model (GP) for each individual time series with a squared exponential covariance function (eq.(5)). (Rasmussen 2006)
- Multi-task Gaussian process model (MTGP) for MTS with a squared exponential covariance function (eq.(2)). (Ghassemi et al. 2015; Durichen et al. 2015)
- Standard LDS-based population model with adaptive prediction (LDS).
- The LDS-based population model combined with the Gaussian process regression model for each individual residual time series (LDS+reGP). It is a special (simpler) version of our model.

Evaluation Metrics

We evaluate and compare the performance of the different methods by calculating the average Mean Absolute Percentage Error (Avg-MAPE) of models' predictions. Avg MAPE measures the prediction deviation proportion in terms of true values:

$$\text{Avg-MAPE} = \frac{\sum_{l=1}^{N}\sum_{i=1}^{n}\sum_{t=1}^{T_l}|1 - \hat{y}^l_{it}/y^l_{it}|}{n\sum_{l=1}^{N}T_l} \times 100\%$$

where $|\cdot|$ denotes the absolute value; y^l_{it} and $\hat{y}^l_{it}$ are the *t*th true and predicted values from time series *i* for patient *l*.

Results

Figure 2 shows the one-step-ahead MTS predictions made by our approach (LDS+reMTGP) for one patient from our test set. The LDS population model is trained on a 400-patient clinical MTS training set. As can be seen from Figure 2, our model is able to capture very well the trend of the

[1] The supplemental material can be found at http://www.zitaoliu.com/download/aaai2016_sup.pdf.

[2] CBC panel is used as a broad screening test to check for such disorders as anemia, infection, and other diseases.

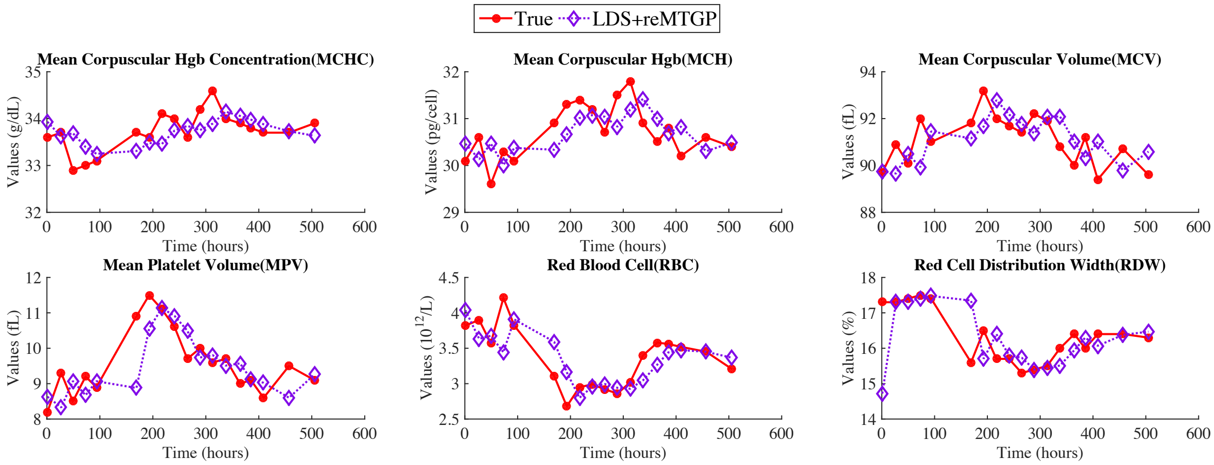

Figure 2: Clinical MTS predictions for one patient. The population based LDS model is trained on 400 patient sequences.

true MTS dynamics. More specifically, LDS+reMTGP can quickly adapt to sudden changes in the true signal and its short term variability (ups and downs), as can be observed, for example, in MPV, RBC, RDW subgraphs of Figure 2.

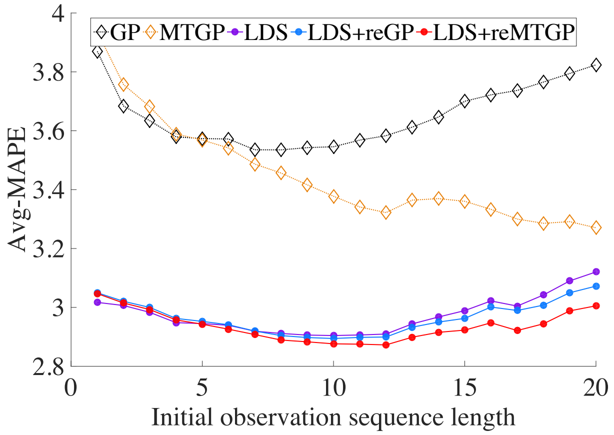

Figure 3: Avg-MAPE results with different initial observation lengths.

Figure 3 compares our new method *LDS+reMTGP* in terms of Avg-MAPE to various state-of-the-art approaches listed in *Baselines* subsection. Due to the poor performance of the *P_Mean* and *I_Mean* methods, we don't visualize it in Figure 3; however, all numerical results are included and listed in the supplemental material. Since our two-stage approach has to fit the parameters of the GP or MTGP models of the residual time series it may experience some initial period in which it is not stable and may lead to suboptimal predictions. To reflect this, Figure 3 shows the Avg-MAPE performance of all methods when they start to predict with a fixed delay corresponding to the different number of initial observations (initial observation sequence length). For example, when the initial observation sequence length is set to 4 the Avg-MAPE reflects the errors of all one-step-ahead predictions the method makes when starting from four initial observations for the target patient (that is, when all predictions the model can make for sequences of $0, 1, 2, 3$ initial observations are ignored). The results show that the population-based LDS model is the best performer when very little is known about the target patient and when patient's observation sequences are short. However, *LDS+reGP* and *LDS+reMTGP*) methods outperform the LDS rather quickly and become superior when more than five initial observations for the target patient become available and are considered. In contrast to the LDS, pure patient-specific models (GP and MTGP) that ignore any population data adapt very slowly and do not reach the performance of LDS or our methods even for the initial observation sequence of length 20. Finally, a simple population-based method (*P_Mean*) and a simple patient-specific method (*I_Mean*) lag behind (see supplemental material for the results) and perform much worse than more advanced time series prediction models.

Conclusion

In this paper, we presented a new two-stage adaptive forecasting model for irregularly sampled multivariate clinical time series data. In contrast to the traditional time-series forecasting models, our model learns from both the population data (time series for other patients) and the target patient data (time series of past observations for the target patient). Our experimental results demonstrate that our model can outperform after a short adaptation period other prediction models and approaches. In the future, we plan to study the different ways of combining/switching among the different prediction models in order to improve their overall prediction accuracy.

Acknowledgment

The work presented in this paper was supported by grant R01GM088224 from the NIH. Its content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH.

References

Adorf, H.-M. 1995. Interpolation of irregularly sampled data series—a survey. In *Astronomical Data Analysis Software and Systems IV*, volume 77, 460.

Åström, K. J. 1969. On the choice of sampling rates in parametric identification of time series. *Information Sciences* 1(3):273–278.

Bellazzi, R.; Siviero, C.; Stefanelli, M.; and De Nicolao, G. 1995. Adaptive controllers for intelligent monitoring. *Artificial Intelligence in Medicine* 7(6):515–540.

Bellazzi, R.; Larizza, C.; Magni, P.; Montani, S.; and Stefanelli, M. 2000. Intelligent analysis of clinical time series: an application in the diabetes mellitus domain. *Artificial intelligence in medicine* 20(1):37–57.

Bonilla, E. V.; Chai, K. M.; and Williams, C. 2007. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, 153–160.

Clifton, L.; Clifton, D. A.; Pimentel, M.; Watkinson, P. J.; Tarassenko, L.; et al. 2013. Gaussian processes for personalized e-health monitoring with wearable sensors. *Biomedical Engineering, IEEE Transactions on* 60(1):193–197.

Dezhbakhsh, H., and Levy, D. 1994. Periodic properties of interpolated time series. *Economics Letters* 44(3):221–228.

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

Fox, E. B.; Sudderth, E. B.; Jordan, M. I.; and Willsky, A. S. 2011. Joint modeling of multiple related time series via the beta process. *arXiv preprint arXiv:1111.4226*.

Funk, N. 2003. A study of the kalman filter applied to visual tracking. *University of Alberta, Project for CMPUT* 652:6.

Ghahramani, Z., and Hinton, G. E. 1996. Parameter estimation for linear dynamical systems. Technical report, Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science.

Ghassemi, M.; Pimentel, M. A.; Naumann, T.; Brennan, T.; Clifton, D. A.; Szolovits, P.; and Feng, M. 2015. A multivariate timeseries modeling approach to severity of illness assessment and forecasting in icu with sparse, heterogeneous clinical data. In *Proc. Twenty-Ninth AAAI Conf. on Artificial Intelligence*.

Hamilton, J. D. 1994. *Time series analysis*, volume 2. Princeton university press Princeton.

Hauskrecht, M.; Valko, M.; Batal, I.; Clermont, G.; Visweswaran, S.; and Cooper, G. F. 2010. Conditional outlier detection for clinical alerting. In *AMIA Annual Symposium*, 286–290.

Hauskrecht, M.; Batal, I.; Valko, M.; Visweswaran, S.; Cooper, G. F.; and Clermont, G. 2013. Outlier detection for patient monitoring and alerting. *Journal of Biomedical Informatics* 46(1):47–55.

Kalman, R. E. 1960. A new approach to linear filtering and prediction problems. *Journal of Fluids Engineering* 82(1):35–45.

Kreindler, D., and Lumsden, C. 2006. The effects of the irregular sample and missing data in time series analysis. *Nonlinear Dynamics, Psychology, and Life Sciences* 10(2):187–214.

Lasko, T. A.; Denny, J. C.; and Levy, M. A. 2013. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one* 8(6):e66341.

Lee, J. W.; Kim, M. S.; and Kweon, I. S. 1995. A kalman filter based visual tracking algorithm for an object moving in 3d. In *Intelligent Robots and Systems 95.'Human Robot Interaction and Cooperative Robots', Proceedings. 1995 IEEE/RSJ International Conference on*, volume 1, 342–347. IEEE.

Li, L.; McCann, J.; Pollard, N. S.; and Faloutsos, C. 2009. Dynammo: Mining and summarization of coevolving sequences with missing values. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, 507–516. ACM.

Liu, Z., and Hauskrecht, M. 2013. Clinical time series prediction with a hierarchical dynamical system. In *Artificial Intelligence in Medicine*. Springer. 227–237.

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

Liu, Z., and Hauskrecht, M. 2015. A regularized linear dynamical system framework for multivariate time series analysis. In *The 29th AAAI Conference on Artificial Intelligence*.

Liu, Z.; Wu, L.; and Hauskrecht, M. 2013. Modeling clinical time series using gaussian process sequences. In *SIAM International Conference on Data Mining (SDM)*, 623–631.

Marlin, B. M.; Kale, D. C.; Khemani, R. G.; and Wetzel, R. C. 2012. Unsupervised pattern discovery in electronic health care data using probabilistic clustering models. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, 389–398. ACM.

Murphy, K. P. 2002. *Dynamic bayesian networks: representation, inference and learning*. Ph.D. Dissertation, University of California, Berkeley.

Rasmussen, C. E. 2006. Gaussian processes for machine learning.

Rehfeld, K.; Marwan, N.; Heitzig, J.; and Kurths, J. 2011. Comparison of correlation analysis techniques for irregularly sampled time series. *Nonlinear Processes in Geophysics* 18(3):389–404.

Rogers, M.; Li, L.; and Russell, S. J. 2013. Multilinear dynamical systems for tensor time series. In *Advances in Neural Information Processing Systems*, 2634–2642.

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

Stegle, O.; Fallert, S. V.; MacKay, D. J.; and Brage, S. 2008. Gaussian process robust regression for noisy heart rate data. *Biomedical Engineering, IEEE Transactions on* 55(9):2143–2151.