Dynamical Gaussian Process Latent Variable Model for Representation Learning from Longitudinal Data

Thanh Le
College of IST, PennState University
State College, Pennsylvania
txl252@psu.edu

Vasant Honavar
College of IST, PennState University
State College, Pennsylvania
vhonavar@ist.psu.edu

ABSTRACT

Many real-world applications involve *longitudinal data*, consisting of observations of several variables, where different subsets of variables are sampled at irregularly spaced time points. We introduce the Longitudinal Gaussian Process Latent Variable Model (L-GPLVM), a variant of the Gaussian Process Latent Variable Model, for learning compact representations of such data. L-GPLVM overcomes a key limitation of the Dynamic Gaussian Process Latent Variable Model and its variants, which rely on the assumption that the data are fully observed over all of the sampled time points. We describe an effective approach to learning the parameters of L-GPLVM from sparse observations, by coupling the dynamical model with a Multitask Gaussian Process model for sampling of the missing observations at each step of the gradient-based optimization of the variational lower bound. We further show the advantage of the Sparse Process Convolution framework to learn the latent representation of sparsely and irregularly sampled longitudinal data with minimal computational overhead relative to a standard Latent Variable Model. We demonstrated experiments with synthetic data as well as variants of MOCAP data with varying degrees of sparsity of observations that show that L-GPLVM substantially and consistently outperforms the state-of-the-art alternatives in recovering the missing observations even when the available data exhibits a high degree of sparsity. The compact representations of irregularly sampled and sparse longitudinal data can be used to perform a variety of machine learning tasks, including clustering, classification, and regression.

CCS CONCEPTS

• **Computing methodologies → Dimensionality reduction and manifold learning**; *Learning latent representations.*

KEYWORDS

Gaussian Process Latent Variable Model, Longitudinal, Dimensionality Reduction

ACM Reference Format:
Thanh Le and Vasant Honavar. 2020. Dynamical Gaussian Process Latent Variable Model for Representation Learning from Longitudinal Data. In *2020 ACM SIGIR International Conference on the Theory of Information Retrieval (FODS '20), October 19–20, 2020, Virtual Event, USA.* ACM, New York, NY, USA, 6 pages. 10.1145/3412815.3416894

1 INTRODUCTION

Longitudinal data, also called *panel data*, consist of repeated observations of several variables from a set of individuals ([12, 15, 21, 26]). In contrast to multivariate time series, wherein all of the variables are measured at regularly spaced time points, in the longitudinal setting, different subsets of variables are sampled at varried irregularly sampled time points. Such data present longitudinal counterparts of classical machine learning problems of clustering, classification, and regression. Some examples include the task of clustering experimental subjects based on longitudinal observations of performance on various cognitive tests; predicting age-related cognitive declines, e.g., in episodic memory, or predicting educational attainment of students as a function of age, demographics, educational experiences, etc., or clustering patients based on their clinical histories (captured in their Electronic Health Records). Models which accommodate longitudinal data typically address a few challenges: *sparsity:* the temporal observation are highly sparse, meaning they contain
of the observation may vary substantially.

1.1 REPRESENTATION LEARNING FROM LONGITUDINAL DATA

Given the many challenges presented by real-world longitudinal data, a problem of particular interest is the task of learning compact, low-dimensional representations or embedding of longitudinal data. Such representations make it possible to apply a wide range of existing machine learning methods for clustering, classification, and regression to longitudinal data. Against this background, this study addresses the problem of representation learning from the longitudinal data.

Related Work. Latent Variable Models (LVMs), or statistical models that relate a set of observed variables to a set of latent variables under the assumption that the observations are controlled by the latent variables, have a long history in machine learning. Of particular interest in the context of this paper are Gaussian Process Latent Variable Models (GPLVM) ([16]) which can be thought of as a combination of LVMs and Gaussian Process (GP) models ([19]). GPLVMs consist of a set of LVMs where each observed variable is the sum of the corresponding latent variables and noise. Furthermore, GPLVM assumes that the *functional variables* are generated from low dimensional variables modeled by Gaussian Processes. GPLVM represent a class of Bayesian non-parametric model whose

FODS '20, October 19–20, 2020, Virtual Event, USA

flexible structure allows it to grow as needed to accommodate the complexity of the data. In recent years, many variants and extensions of GPLVMs have been developed. For example, The Gaussian process dynamical models (GPDM) ([27]) models the dynamics of the process using Markov transitions between states in the latent space, and variants of GPDM such as higher-order Markov dependencies in the latent space ([28]). Due to the dynamical prior, irregular time observation are treated in Markov fashion. Titsias introduced a variational inference framework for training GPLVM for Bayesian nonlinear dimensionality reduction ([24]). Damianou et al. ([9, 10]) introduced the Dynamical Gaussian Process Latent Variable Model (D-GPLVM), a natural extension of the Bayesian GPLVM that can accommodate dependencies between latent variables. Unlike GPDM, D-GPLVM uses kernels to incorporate complex (not necessarily Markov) dependencies between latent variables. It can use the Radial Basis Function (RBF) kernel to model smooth temporal dynamics, or Ornstein-Uhlbeck covariance function to model a Gauss-Markov process. As a generative model, D-GPLVM can be used to produce the observations at any desired time point. More recently, Damianou et al. ([10, 11]), building on the results of ([13, 14]), proposed general frameworks for dealing with the scenario where the inputs and outputs of the generative model are uncertain. However, this body of work assumes that the data samples are fully observed at many time points. A key limitation of the body of work summarized above is that they assume that the data is fully observed, that is the values of the observable variables are available at all observed time points. However, this assumption does not hold in many practical scenarios where only small subsets subsets of the observable variables are sampled at irregularly spaced time points.

Contributions. We propose to relax the assumption of fully observed data to develop, L-GPLVM, a variant of GPLVM for learning compact representations of longitudinal data where the observations are produced by sampling of possibly different sparse subsets of variables at irregularly spaced time points. We show how to efficiently learn the parameters of L-GPLVM from longitudinal data. We present results of experiments with synthetic as well as several variants of the Human Motion Capture (MOCAP) benchmark data that show that L-GPLVM substantially outperforms the state-of-the-art methods for representation learning from longitudinal data. The proposed approach is amenable to being extended to handle irregularly and sparsely sampled spatio-temporal data.

Organization of the Paper. The rest of the paper is organized as follows. Section 2 introduces the D-GPLVM and the variational inference framework for learning D-GPLVM. Section 3 introduces L-GPLVM for learning the representations of sparsely and irregularly sampled longitudinal data. Section 4 presents results of experiments that show that L-GPLVM significantly outperforms GPLVM ([9]), deepGP with dynamical prior ([8]) when evaluated using the reconstruction error of the longitudinal data as the performance metric.

2 PRELIMINARIES

Gaussian Processes. Gaussian Process (GP) ([23]), is a flexible Bayesian non-parametric model and is the primary building block of the Gaussian Process Latent Variable Model. In GP, we model a finite set of random functions $f = [f(x_1), ..., f(x_N)]^T$ as a joint Gaussian distribution $f \sim \mathcal{GP}(\mu, K)$ where the covariance matrix K is evaluated using choices of kernel functions.

In GP Regression, the goal is to predict the response y^* of a new input x^*, given a training set $\{(x_i, y_i)\}_{i=1}^N$ of N training samples. The response variable y_i is modeled as the function value $f(x_i)$ corrupted by noise $y_i \sim \mathcal{N}(f(x_i), \sigma^2)$. Given the joint probability of the response variables and the latent function $p(y, f) = p(y|f)p(f)$, the distribution of the latent function value f^* is a Gaussian distribution with mean and variance

$$\begin{aligned} \mu(x^*) &= k_{x^*X}(\sigma^2 I + K_{XX})^{-1} y \\ var(x^*) &= k_{x^*x^*} - k_{x^*X}(\sigma^2 I + K_{XX})^{-1} k_{Xx^*} \end{aligned} \quad (1)$$

where $k_{x^*X} = k(x^*, X)$ is the covariance between the new input x^* and the N training sample evaluated by the kernel function k.

Gaussian Process Latent Variable Model (GPLVM). GPLVM was first conceived as an approach to visualize data by reducing its dimensionality and can be seen as a non-linear extension of the Probabilistic PCA ([17]). GPLVM learns a low dimensional representation $X^{N \times Q}$ of the data matrix $Y^{N \times D}$ where $Q \ll D$. The mapping $f : X \rightarrow Y$ is a nonlinear function with Gaussian Process (GP) prior $f \sim \mathcal{GP}(0, K)$. Thus, the i^{th} sample y_i is generated as

$$y_i = f(x_i) + \epsilon \quad (2)$$

GPLVM allows the specification of a suitable prior over the latent space X. While one can utilize GPLVM without specifying a prior, that would be tantamount to maximizing the log marginal likelihood of the data with the attendant danger of over fitting ([19]). Up until ([24]), the standard approach in learning GPLVM was to find the MAP estimate of X ([16]) whilst jointly maximizing the log marginal likelihood with respect to the data and the hyperparameters. Recent work has led to several variants of GPLVM for specific applications ([19]). Of particular interest in our setting is GPLVM with a dynamical prior to model multivariate time series data ([9, 27]).

Bayesian-GPLVM. ([24, 25]) gave a full Bayesian treatment of the GPLVM, introducing a GP prior based on auxiliary inducing points ([7]) to make the resulting the variational Bayes inference problem computationally tractable. The latent variables were then variationally integrated out and a closed-form lower bound on the marginal likelihood computed. The marginal likelihood of the data $p(Y) = \int p(Y|X)p(X)dX$ is intractable because X appears nonlinear inside the covariance matrix $K_{NN} + \beta^{-1} I_N$. A variational distribution $q(X)$ is introduced to approximate the true posterior distribution $p(X|Y)$. The chosen variational distribution in the i.i.d case has a factorized Gaussian form:

$$q(X) = \prod^N \mathcal{N}(x_n|\mu_n, S_n) \quad (3)$$

This yields the Jensen's lower bound on the $log\, p(Y)$ of the form:

$$F(q) = \tilde{F}(q) - KL(q(X)|p(X)) \quad (4)$$

The negative KL divergence between the variational posterior distribution q(X) and the prior distribution $p(X)$ can be computed analytically whereas the first term can be decomposed into separate computations for each dimension:

$$\begin{aligned}\tilde{F}(q) &= q(X)logp(Y|X)dX \\ &= \sum_{d=1}^{D} \int q(X)logp(y_d|X)dX = \tilde{F}_d(q)\end{aligned} \tag{5}$$

The intractable integration of $logp(y_d|X)$ which appears in $\tilde{F}_d(q)$ can then be approximated using inducing points. For each vector f_d, a set of M inducing variables u_d is introduced. The u_d's are evaluated at a set of inducing locations given by $Z \in R^{M \times Q}$. U are simply the function samples drawn from the same conditional prior, augmenting the joint probability model:

$$p(y_d, f_d, u_d|X, Z) = p(y_d|f_d)p(f_d|u_d, X, Z)p(u_d|Z) \tag{6}$$

The likelihood $p(y_d|X)$ can be computed from the augmented model by marginalizing out (f_d, u_d) for any value of the inducing inputs Z. This allows $p(f_d|X)$ to be computed by $q(f_d, u_d) = p(f_d|u_d, X)\phi(u_d) = p(f_d|u_d)\phi(u_d)$, which is tractable. The bound for the data can then be fully specified by the Psi statistics $\Psi_0 = Tr(\langle K_{NN} \rangle_{q(X)})$, $\Psi_1 = \langle K_{NM} \rangle_{q(X)}$, $\Psi_2 = \langle K_{MN} K_{NM} \rangle_{q(X)}$ where $\langle \cdot \rangle_{q(X)}$ denotes the expectation under the variational distribution $q(X)$ ([18]). An attractive feature of this approach is its ability to automatically determine the latent dimensionality of a given data set by Automatic Relevance Determination (ARD).

Variational Gaussian Process Dynamical Systems. A dynamical prior can be imposed over the X in the GPLVM to enable modeling of dynamical system ([9, 18]). In the multivariate time series data $\{y_n, t_n\}_{n=1}^{N}$, where $y_n \in \mathbb{R}^{\mathrm{D}}$ is a d-dimensional observation at time $t_n \in \mathbb{R}^+$. The system could be summarized as follows:

$$\begin{aligned} x_q(t) &\sim \mathcal{GP}(0, k_x(t_i, t_j)), q = 1, ..., Q \\ f_d(x) &\sim \mathcal{GP}(0, k_f(x_i, x_j)), d = 1, ..., D \end{aligned} \tag{7}$$

The kernel functions k_x and k_f are parameterized by θ_x and θ_f respectively. The choice of k_x to be indefinitely differentiable function, i.e. the square exponential (RBF) allows generation of a smooth path in the latent space. Indeed, the major difference between the Bayesian GPLVM in ([24]) and D-GPLVM in ([8]) has to do with the use of dynamical prior by the latter. Consequently, the derivations of the variational lower bound of D-GPLVM is similar to that of GPLVM with one exception: the KL divergence $KL(q(X)||p(X))$ being replaced by $KL(q(X)||p(X|t))$ and X are coupled temporally leading to the factorization on $q(X)$:

$$q(X) = \prod^{Q} \mathcal{N}(x_q|\mu_q, S_q) \tag{8}$$

This results in a full-rank covariance matrix S_q with N^2 parameters. The re-parameterization trick introduced in ([20]) can then be employed to reduce the number of parameters to that of the standard Bayesian Gaussian Process. Specifically μ_q and S_q can be parameterized by the $\mu_q = K_t \bar{\mu}_q$ and $S_q = (K_t^{-1} + \Lambda_q)^{-1}$ where $\bar{\mu}_q$ and Λ_q consist of $Q \times N$ free parameters.

3 LONGITUDINAL GAUSSIAN PROCESS LATENT VARIABLE MODEL

We proceed to first show how a simple sampling procedure of the unseen observations suffices to resolve the difficulty of learning the dynamical representation in the longitudinal setting, yielding L-GPLVM, a longitudinal variant of D-GPLVM. We will then introduce a more computationally efficient formulation for the L-GPLVM.

In the fully observed setting of GPLVM, the latent variables X are optimized so as to properly propagate the information contained in X to the observations Y. However, in the longitudinal setting, because the observations are irregular and sparse, the variational approximation underlying GPLVM becomes non-trivial. An intuitive way to overcome such difficulty is to impute the missing observations at each of the time points given possible correlation among the different variables in Y. Specifically, we model the missing observations using a Gaussian Process model. Suppose the covariance function of the dynamical model is given by a dynamical covariance K_t over X, this covariance function also implicitly specifies the covariance among observations within individual dimension of Y. Specifically, the computation of the evidence bound breaks down to the summation over the log likelihood of individual data dimensions with respect to the variational distribution $q(X)$ in (5) whose mean and covariance are parameterized by K_t in (8). Thus, the dynamical covariance function enables estimation of the sampling distribution for each unobserved variable y_d^* in the d^{th} dimension at time t^*:

$$y_d^* \sim \mathcal{GP}(\mu_{mo}^*, \Sigma_{mo}^*) \tag{9}$$

where $\mu_{mo}^*, \Sigma_{mo}^*$ are the mean and covariance of a multi-task Gaussian Process estimated over the observed outputs with the covariance function, K_t. Here, each missing observation is imputed using regression by solving a multi-task regression problem (where each task corresponds to one (unobserved) dimension of Y. Figure 1 illustrate the plate annotation of the proposed approach. we define a new objective on the composite log-likelihood over the observed data:

$$F^*(q) \geq \mathcal{L}_{mo}(y^o, K_t) + \tilde{\mathcal{F}}_*(q) - KL(q||p) \tag{10}$$

The multitask model log-likelihood $\mathcal{L}_{mo}$ is computed over the observed data using the global dynamical covariance K_t; and $\tilde{\mathcal{F}}^*(q)$ is the dynamical model log-likelihood computed over the fully sampled data $\tilde{\mathcal{F}}^*(q) = \sum_{d=1}^{D} \tilde{\mathcal{F}}_d^* = \sum_{d=1}^{D} q(X)log\ p(y_d^*, y_d^o|X)dX$. The multi-task regression model is used to adaptively and optimally impute the unobserved Y^* at each optimization step using the current best approximation of the global dynamical covariance whereas the Dynamical GPLVM attempts to optimize dynamical covariance as well as the cross-covariance structures using the sampled and observed data.

Sampling using multi-output Gaussian Process. One of the most straightforward choices to establish the sampling distribution is to use the Multi-task Learning Model (MTLM) ([5, 19]:

$$\begin{aligned} \mu_{mo}^* &= (K_f \otimes K_t(t^*, t^o))^T (\Sigma_{mo}^*)^{-1} y^o \\ \Sigma_{mo}^* &= K_f \otimes K_t + D \otimes I \end{aligned} \tag{11}$$

in which, $K_f = \Phi\Phi^T$ is the task covariance matrix to be inferred in the multi-task model. In the case where data dimensions are known

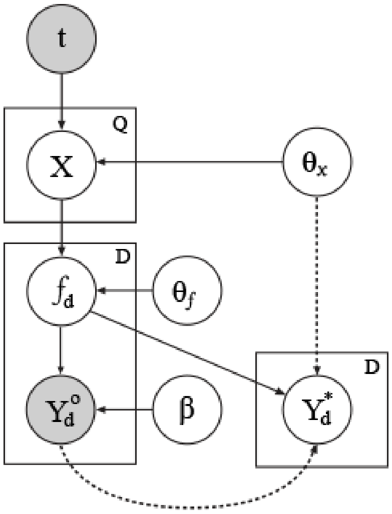

Figure 1: Overview of L-GPLVM Using The Plate Notation. The dashed line represents the utilization of dynamical model parameters to impute missing observations.

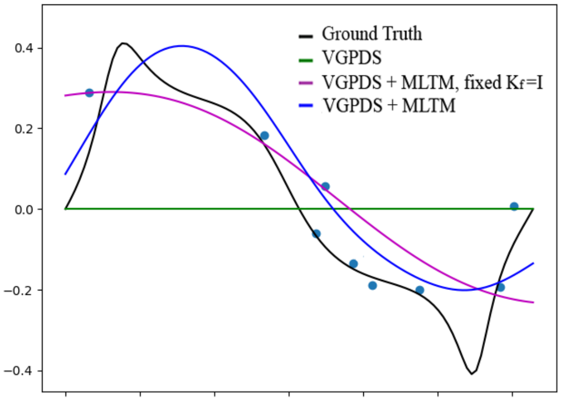

Figure 2: Visualization of Predictive Mean In A Single Dimension of A Simulated Dataset. Only 10% of the 100 simulated time points are observed (represented by the blue dots). The auxiliary MTLM model enables VGPDS to explicitly capture the dependency among the dimensions in the resulting sparse data set.

a priori to be independent, K_f can be set to the identity matrix $K_f = I_D$. Regardless of data dimension dependencies, initializing $K_f = I_D$ allows the unobserved data to first be sampled independently.

Fig. 2 demonstrates the ability of this approach to learn a representation and associated mapping to produce reasonable predictive means in a synthetic longitudinal data set. The pitfall of this simple approach is the complexity during training being dominated by the MTLM with the naive implementation of cubic complexity $O(N^3D^3)$ or a reduced complexity of $O(NDM^2P^2)$ with a couple approximations of such that $M \ll N$ and $P < D$ (For more details on the approximations and complete derivation of $\mathcal{L}_{mo}$ over the observed data, refer to ([5]). While the procedure described above is able to recover the correlation structure and impute the unobserved variables at each time point, it is not computationally efficient. We next describe a way to get around this limitation.

More efficient sampling via sparse convolved Gaussian Processes. We introduce a more efficient construction of the sampling distribution in 9 using the Sparse Convolved Gaussian Process framework. Unlike the MTLM whose multitask covariance is captured by the Kronecker product between the coregionalization matrix, K_f, and the input covariances, in the Process Convolution framework, each function $f_d(x)$, the noiseless version of y_d, can be expressed as a convolution integral between a smoothing kernel G_d and the shared latent function u ([1–3]).

$$f_d(x) = \int_{\mathcal{X}} G_d(x-z)u(z)dz \tag{12}$$

While it is possible to consider influences from multiple latent functions (u's) each with its own smoothing kernel, for simplicity of exposition we will assume the a single latent function as shown in (12). Assume $u(z)$ has the general covariance, $k(z', z)$, then the (cross-)covariances can be computed as:

$$\begin{aligned} cov[f_d(x), f_d'(x')] &= \int_{\mathcal{X}} G_d(x-z)G_d'(x'-z) \\ &\qquad k(z,z')dz'dz \\ cov[f_d(x), u(z)] &= \int_{\mathcal{X}} G_d'(x'-z)k(z,z')dz' \end{aligned} \tag{13}$$

Alvarez et al. showed that by making specific conditional independence assumptions, they arrived at efficient approximation similar in form to partially independent training conditional model (PITC) ([1, 2, 22]). Specifically, they showed that instead of drawing samples from $u(z)$, one could draw samples from its finite representation, i.e., $u = [u(z_1), ..., u(z_M)]^T$ at $Z = z_{k_{k=1}}^M$, the set of inducing vectors at which $u(z)$ is evaluated. In this study, the set of inducing input Z are M equally spaced time points in the modeling interval. Each function f_d in (12) can then be approximated by:

$$f_d(x) \approx \int_{\mathcal{X}} G_d(x-z)E[u(z)|u]dz \tag{14}$$

The likelihood for f is $p(f|u, Z, X, \theta) = \mathcal{N}(f|K_{f,u}K_{u,u}^{-1}u, K_{f,f} - K_{f,u}K_{u,u}^{-1}K_{u,f})$, where $K_{u,u}$ is evaluated using the temporal covariance $K_t(z, z')$. For this study, like in ([2] we assume the individual outputs in f are independent conditional on u. Under this assumption, the full multitask covariance matrix $K_{f,f}$ can be replaced by its low rank approximation $K_{f,u}K_{u,u}^{-1}K_{u,f}$ in all entries except in the diagonal block corresponding to K_{f_d,f_d}. Following the same line of reasoning as in the previous subsection and ([2], we can construct a sampling distribution in (9) as follows:

$$\begin{aligned} \mu_{mo}^* &= K_{f_*,u}A^{-1}K_{u,f}(D+\Sigma)^{-1}y^o \\ \Sigma_{mo}^* &= K_{f_*,f_*} - K_{f_*,u}K_{u,u}^{-1}K_{u,f_*} \\ &\quad + K_{f_*,u}A^{-1}K_{u,f_*} + \Sigma^* \end{aligned} \tag{15}$$

where $A = K_{u,u} + K_{u,f}(D + \Sigma)^{-1}K_{f,u}$, and $D = blockdiag[K_{f,f} - K_{f,u}K_{u,u}^{-1}K_{u,f}]$ is the relevant multitask likelihood over the observed output sharing the same covariance parameters as the Dynamical model:

$$\begin{aligned}\mathcal{L}_{mo} \propto -\frac{1}{2}log|K_{u,u}| - \frac{1}{2}log|A| - \frac{1}{2}tr\left[D^{-1}yy^{T}\right] \\ +\frac{1}{2}tr\left[D^{-1}K_{f,u}A^{-1}D^{-1}yy^{T}\right]\end{aligned} \tag{16}$$

This multitask Gaussian Process Model shares the set of parameters from the Dynamical model, particularly the dynamical kernel, at each step of the optimization, yielding a sampling process much more efficient $O(N^3D)$ than MTLM $O(N^3D^3)$ due to the use of the sparse approximation.

4 EXPERIMENTS AND RESULTS

We proceed to describe experiments that are designed to address the following research questions: How does the L-GPLVM (with RBF kernel) compare with with the state-of-the-art baselines, namely, the D-GPLVM ([9]), the deepGP ([8]) , Nearest Neighbor (NN), and 3 popular statistical imputations for time-series, in imputing the unobserved variables (as measured by the error of the reconstructed observations relative to the ground truth) in the longitudinal setting where observations are made at irregularly spaced time points at each of which only a small subset of the observable variables are actually observed? How does the performance of vary as a function of sparsity of the available observations? Before proceeding to discuss the experiments, we briefly describe the data sets used in our experiments.

Dataset

Synthetic Data: The synthetic data were generated by sampling 3 smooth curves using Gaussian Process unknown to the dynamical models with time as input. 15 additional dimensions were created by random linear combination of the sampled curves. The resulting curves were sampled at 100 equally spaced time points between 0 to 2π. To obtain irregularly sampled sparse data, a subset of the observations were masked by choosing a random subset of the time points independently for each dimension until the desired fraction of the observations were eliminated. Using this procedure, we generated data sets with varying degrees of data sparsity ranging from 0.1 to 0.8.

Human motion capture data We used the data from the MOCAP data set ([6]) corresponding to the walking motion of a human body represented by the positions of 59 joints. We applied the same masking procedure as the one used in the case of synthetic data to generate the data sets with varying degrees of data sparsity ranging from 0.0 to 0.6.

Experiments

We compare GPLVM based models with the identical choice of the number of latent dimensions with $Q = 3$ for the synthetic data set and $Q = 7$ for the MOCAP data set. The deepGP model has 1 intermediate layer which is over-complete, i.e., the number of nodes exceed the dimensionality of the data set (in our experiments, by a factor of 3). Each node in the intermediate layer's output is a randomly selected group of variable in Y. L-GPLVM performs imputation of unobserved data using the algorithm described in Section 3, whereas D-GPLVM and deepGP do so using mean imputation. We also included the result for three other time-series imputation methods to serve as popular baseline for comparison: NN, Last Observation Carried Forward (LOCF), Moving Window Interpolation (MW, window size=5), and Multiple Imputation by Chained Equations (MICE) ([4]).

We assessed the performance (as measured by the sum of absolute error, $SAE = \sum_{i=1}^{N} |\hat{y}_i - y_i|$, of the reconstructed data relative to the ground truth data) on the synthetic data as well as the MOCAP data with different degrees of sparsity of observations.

The results of experiments with the synthetic data, summarized in Table 1, show that L-GPLVM substantially outperforms all other methods, reconstructing the missing observations, even when presented with data with fairly high degrees of data sparsity. The experiment was repeated 3 times and the mean and standard deviation (in parentheses) of the reconstruction error are reported. In contrast, the performance of the standard D-GPLVM quickly degrades as the observations become increasingly sparse, until at data sparsity of 0.7, it fails to recover the shape of the underlying curve.

The results of experiments with the MOCAP data, summarized in Table 2, are consistent with the results obtained with the synthetic data. That is, L-GPLVM substantially outperforms all other methods, reconstructing the missing observations, even when presented with data with fairly high degrees of data sparsity, whereas the other methods fail to do so, especially as the degree of data sparsity increases.

5 SUMMARY AND DISCUSSION

In this paper, we have introduced L-GPLVM, a variant of GPLVM for representation learning from high dimensional longitudinal observations, where different subsets of the variables are sampled at irregularly spaced time points. L-GPLVM overcomes a key limitation of D-GPLVM and related models which rely on the assumption that the data are fully observed over all of the sampled time points, an assumption that is often violated the longitudinal setting. We have also described an effective approach to learning the parameters of L-GPLVM from sparse observations, by coupling the dynamical model with a Multitask Gaussian Process model for sampling of the missing observations at each step of the gradient-based optimization of the variational lower bound. We have further shown how to take advantage of the Sparse Process Convolution framework ([1]) to learn the latent representation of sparsely and irregularly sampled longitudinal data with minimal computational overhead relative to a standard LVM. We have presented results of experiments with synthetic data as well as MOCAP data that show that L-GPLVM substantially and consistently outperforms the state-of-the-art alternatives in recovering the missing observations even when the available data exhibits a high degree of sparsity. The compact representations of irregularly sampled and sparse longitudinal data can be used to perform a variety of machine learning tasks, including clustering, classification, and regression.

Some promising directions for further research include extensions of L-GPLVM to irregularly sampled, sparse, spatio-temporal data, dynamic processes over networks (topological longitudinal

Table 1: Reconstruction Absolute Sum of Error of The Different Models Learned From The Irregularly And Sparsely Sampled Longitudinal Data.

Sparsity	0.1	0.3	0.5	0.6	0.7	0.8
D-GPLVM	21.5(1.6)	48.2(1.4)	75.4(1.5)	87.2(2.2)	128(0.4)	128(0.4)
deepGP	17.6(1.1)	33.0(2.3)	57.4(1.3)	67.4(0.9)	85.9(0.8)	104(3.8)
NN	22.4(0.9)	29.9(4.6)	62.0(6.0)	65.2(5.6)	80.1(6.4)	107(1.1)
LOCF	38.1(2.1)	66.3(1.3)	94.7(1.9)	105(2.3)	127(1.0)	129(0.2)
MW	31.6(1.2)	56.9(2.0)	87.2(1.9)	101.4(3.8)	128(0.4)	128(0.4)
MICE	24.9(1.4)	31.7 (1.8)	68.8(2.1)	81.8(1.6)	117(2.3)	130(10)
L-GPLVM	**14.2(0.7)**	**19.4(2.8)**	**28.8(1.2)**	**32.4(2.5)**	**36.1(0.7)**	**53.4(1.9)**

Table 2: Reconstruction Error Using Only Time As Input In The Temporal Models

Sparsity	0.0	0.2	0.4	0.6
D-GPLVM	84.2	252	332	332
deepGP	**82.3**	143	201	320
L-GPLVM	84.1	**84.9**	**135**	**221**

data) and to settings where the observations are influenced by transitions between latent states (e.g., as in the case of longitudinal electronic health records data which reflect observations resulting from transitions between different latent states, e.g., healthy and unhealthy). Additional applications of interest include modeling of environmental data, climate data, etc.

Acknowledgements

This work was funded in part by grants from the NIH NCATS through the grant UL1 TR002014 and by the NSF through the grants 1640834, and 1636795, the Penn State Center for Big Data Analytics and Discovery Informatics, the Edward Frymoyer Endowed Professorship in Information Sciences and Technology at Pennsylvania State University and the Sudha Murty Distinguished Visiting Chair in Neurocomputing and Data Science funded by the Pratiksha Trust at the Indian Institute of Science (both held by Vasant Honavar). The content is solely the responsibility of the authors and does not necessarily represent the official views of the sponsors.

REFERENCES

[1] M. Álvarez and N. D. Lawrence. 2009. Sparse Convolved Gaussian Processes for Multi-output Regression. *Advances in Neural Information Processing Systems 21* (2009), 57–64.

[2] Mauricio A Alvarez and Neil D. Lawrence. 2011. Computationally efficient convolved multiple output gaussian processes. *The Journal of Machine Learning Research* 12 (2011), 1459–1500.

[3] Mauricio A. Alvarez, David Luengo, Michalis K. Titsias, and Neil D. Lawrence. 2011. Efficient multioutput Gaussian processes through variational inducing kernels. *International Conference on Artificial Intelligence and Statistics* x (2011), 25–32.

[4] Melissa J Azur, Elizabeth A Stuart, Constantine Frangakis, and Philip J Leaf. 2012. MICE - What is it, and how does it work. 20, 1 (2012), 40–49. https://doi.org/10.1002/mpr.329.Multiple

[5] Edwin V. Bonilla, Kian Ming Chai, and Christopher Williams. 2008. Multi-task Gaussian Process Prediction. *Advances in Neural Information Processing Systems* 20, October (2008), 153–160. https://doi.org/10.1017/CBO9781107415324.004

[6] CMU. 2001. MOCAP database.

[7] Lehel Csató, Manfred Opper, Lehel Csató, and Manfred Opper. 2001. Sparse representation for Gaussian process models. *Advances in Neural Information Processing Systems* (2001), 444–450.

[8] Andreas C. Damianou and Neil D. Lawrence. 2013. Deep Gaussian Processes. *Artificial Intelligence and Statistics (AISTATS)* (2013), 207–215. https://doi.org/10.1002/nme.1296

[9] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. 2011. Variational Gaussian Process Dynamical Systems. *Advances in Neural Information Processing Systems* (2011), 2510–2518. http://arxiv.org/abs/1107.4985

[10] Andreas C. Damianou, Michalis K. Titsias, and Neil D. Lawrence. 2014. Variational Inference for Uncertainty on the Inputs of Gaussian Process Models. (2014).

[11] Andreas C Damianou, Michalis K Titsias, and Neil D Lawrence. 2016. Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes. *Journal of Machine Learning Research* (2016).

[12] Garrett M Fitzmaurice, Nan M Laird, and James H Ware. 2012. *Applied Longitudinal Analysis.* Wiley. 752 pages.

[13] Agathe Girard. 2004. Approximate methods for propagation of uncertainty with Gaussian process models. *Ph.D. Thesis* October (2004).

[14] Agathe Girard, Carl Edward Rasmussen, Joaquin Quinonero Candela, and Roderick Murray-Smith. 2003. Gaussian process priors with uncertain inputs-application to multiple-step ahead time series forecasting. *Advances in neural information processing systems* (2003), 545–552.

[15] Donald Hedeker and Robert D Gibbons. 2006. *Longitudinal Data Analysis.* Wiley. 360 pages.

[16] Neil Lawrence. 2005. Probabilistic non-linear Principal Component Analysis with Gaussian Process Latent Variable Models. *Journal ofMachine Learning Research* 6 (2005), 1783–1816.

[17] Neil D. Lawrence. 2004. Gaussian Process Latent Variable Models for Visualisation of High Dimensional Data. *NIPS* (2004), 329–336.

[18] Neil D. Lawrence and Andrew J. Moore. 2007. Hierarchical Gaussian Process Latent Variable Models. *International Conference on Machine Learning* (2007), 481–488. https://doi.org/10.1145/1273496.1273557 arXiv:0402594v3 [cond-mat]

[19] Ping Li and Songcan Chen. 2016. A review on Gaussian Process Latent Variable Models. *CAAI Transactions on Intelligence Technology* 1, 4 (2016), 366–376. https://doi.org/10.1016/j.trit.2016.11.004

[20] Manfred Opper and C??dric Archambeau. 2009. The variational gaussian approximation revisited. *Neural Computation* 21, 3 (2009), 786–792. https://doi.org/10.1162/neco.2008.08-07-592

[21] Eleanor M. Pullenayegum and Lily S.H. Lim. 2016. Longitudinal data subject to irregular observation: A review of methods with a focus on visit processes, assumptions, and study design. *Statistical Methods in Medical Research* 25, 6 (2016), 2992–3014. https://doi.org/10.1177/0962280214536537

[22] Joaquin Quiñonero-candela, Carl Edward Rasmussen, and Ralf Herbrich. 2005. A unifying view of sparse approximate Gaussian process regression. *Journal of Machine Learning Research* 6 (2005), 1935–1959. https://doi.org/10.1163/016918609X12529286896877

[23] C E Rasmussen and C K I Williams. 2006. *Gaussian Process for Machine Learning.* https://doi.org/10.1094/PHYTO-96-0876

[24] Michalis Titsias and Neil Lawrence. 2010. Bayesian Gaussian Process Latent Variable Model. *Artificial Intelligence* 9 (2010), 844–851. https://doi.org/10.1162/089976699300016331

[25] Michalis K Titsias. 2009. Variational Learning of Inducing Variables in Sparse Gaussian Processes. *Artificial Intelligence and Statistics (AISTATS)* 5 (2009), 567–574.

[26] Geert Verbeke, Steffen Fieuws, Geert Molenberghs, and Marie Davidian. 2014. the analysis of multivariate longitudinal data: A review. *Statistical Methods in Medical Research* 23, 1 (2014), 42–59. https://doi.org/10.1177/0962280214539863

[27] Jack Wang, David Fleet, and Aaron Hertzmann. 2006. Gaussian process dynamical models. *Advances in Neural Information Processing Systems* (2006), 1441–1448. https://doi.org/10.1109/TPAMI.2007.1167

[28] Jing Zhao and Shiliang Sun. 2016. High-Order Gaussian Process Dynamical Models for Traffic Flow Prediction. *IEEE Transactions on Intelligent Transportation Systems* 17, 7 (2016), 2014–2019. https://doi.org/10.1109/TITS.2016.2515105