Regularization for Sparse Latent Gaussian Processes

Abstract

The Gaussian Process Latent Variable Model (GPLVM) is a flexible unsupervised Bayesian nonparametric modelling approach which has been applied to many learning tasks such as facial expression recognition, image reconstruction, and human pose estimation. Due to poor scaling properties of exact inference methods on GPLVMs, approximation methods based on sparse Gaussian processes (SGP) and stochastic variational inference (SVI) are necessary for inference on large data sets. One problem in SGP, especially in latent variable models, is that the distribution of inducing inputs may exhibit overdispersion which may lead to inefficient inference and poor model fit. In this paper, we first propose a regularization approach for latent sparse Gaussian processes in SVI, which balances the distribution of inducing inputs and latent inputs. We justify the use of this regularization term by proving that performing variational inference (VI) on a GPLVM with this regularization term is equivalent to perform VI on a related empirical Bayes model with a prior on inducing inputs. Second, we extend the categorical latent Gaussian process to model categorical time series and illustrate that our regularization is particularly important in this latent dynamic model with a synthetic data set. Finally, we apply our model to a real data set of stock indices and visualize the latent dynamics.

1 Introduction

A Gaussian process (GP) is a generalization of a multivariate Gaussian distribution that can be seen as a stochastic random process in the space of general continuous functions. Due to its flexibility, it is widely applied in various fields such as geostatistics [9], multitask-learning [1] and reinforcement learning [19]. Gaussian process regression and classification are deeply studied in [20].

Although the GP is flexible, its exact inference is expensive with time complexity $O(n^3)$, where n is the number of data points. This renders GP inference infeasible for large real-world datasets. Approximations based on inducing points, named sparse Gaussian process (SGP) methods, have been proposed to avoid the computational issue. The predictive process (PP/DTC) is proposed by Seeger [21] to approximate a GP by introducing inducing variables. It reduces the time complexity from $O(n^3)$ to $O(nm^2)$ where m is the number of inducing variables. [22] proposes a fully independent training conditional (FITC) approximation as one of most efficient approximation methods. Its corresponding Bayesian approach is proposed as the modified predictive process (MPP), which corrects the bias brought from the PP in [6]. Moreover, [23] proposes a partially independent training conditional (PITC) approximation and [12] proposes an expectation propagation pseudo-point approximation. In most approximation approaches, the locations of inducing points are optimized via a gradient-based optimization. From a Bayesian perspective, [8] discusses inducing input selection using MCMC sampling. On the other hand, [24] applies variational inference to SGP, marginalizing the optimal variational distribution of inducing variables. [11, 10] directly optimize the variational distribution of the inducing variables and latent inputs to gain computational benefits.

The Gaussian process latent variable model (GPLVM) [14] is proposed by Lawrence as a probabilistic dimensionality reduction method. This method extends the linear mappings from embedding space in dual probabilistic principle component analysis (DPPCA) to nonlinear mappings [14, 13]. [13] also discusses its relationship with other dimensionality reduction methods such as Multidimensional

Preprint. Under review.

Scaling [17] and Kernel PCA [2]. Due to the poor scaling property, [25] proposes Bayesian GPLVM using variational inference on a SGP in [24]. And [10] proposes stochastic variational inference on a latent SGP. Many variants of GPLVM are studied in [15, 16, 26].

The main contribution of this work is to propose a regularization approach for the latent SGP of [10], balancing the distribution of inducing inputs and latent inputs, and leading to better model prediction. Theoretically we justify the use of this regularization term by proving that performing variational inference (VI) on a GPLVM with this regularization term is equivalent to performing VI on a related empirical Bayes model with a prior on its inducing inputs. Moreover, we extend the categorical latent Gaussian process [7] to model categorical time series by incorporating a dynamical prior. We illustrate that our regularization is particularly important for this dynamic model.

The rest of this paper is organized as follows. In Section 2 we show that the distribution of inducing inputs in a SGP is important for model prediction. We propose a regularization approach for latent SGPs and justify it through a related empirical Bayesian model in Section 3. In Section 4, we define the temporal categorical latent Gaussian process (TCLGP) including its variational lower bound, predictive density and algorithm for inference of latent variables. Section 5 shows the importance of regularization in GPLVMs for high dimensional complicated data sets such as the Anuran Call data set and applies TCLGP with regularization on both a synthetic dataset and a real data set of stock indices and demonstrates the ability and necessity of regularization for a latent SGP. Finally, we summarize our work and discuss its implications in Section 6.

2 Sparse Gaussian Process

In this section, we show the importance of inducing inputs for a SGP. From a variational inference perspective, there are two efficient approaches named SGPR and SVGP. SGPR marginalizes the optimal variational distribution of inducing variables [24] while SVGP directly models and optimizes the variational distribution of inducing variables [10]. Assuming we have N observations and M inducing points, the lower bound of computational complexity in SGPR is $O(M^2N)$ while that in SVGP is $O(M^3)$. Generally, the inducing inputs are optimized by maximizing the corresponding variational bound. However, when the inducing inputs are intractable in optimization, the distribution of inducing inputs should capture the distribution of the covariate inputs for better model prediction [8]. We illustrate this on 1-D synthetic data, where we uniformly generate 100 inputs $\boldsymbol{x}$ on the unit interval $(0, 1)$. Then the corresponding observations are generated from

$$\begin{aligned} y &\sim \mathcal{N}(y|f, 0.1^2) \\ f &= \sin(2x) + 0.2\cos(22x)\,. \end{aligned}$$

We take 100 evenly spaced inputs on $(0, 1)$ as test inputs and generate corresponding outputs as their true test outputs. We use a linear combination of a Matern kernel and a linear kernel as the covariance function and take different settings of inducing points, in which inducing points are evenly distributed on an interval centered at 0.5 and of possible lengths 0.5, 1 or 2. We fix those inducing points in optimization, and the resulting predictive posterior processes are shown in Figure1. The likelihood and root mean square error are summarized in Table 1, illustrating that the model has best predictive performance when the distribution of the inducing inputs captures the distribution of actual inputs.

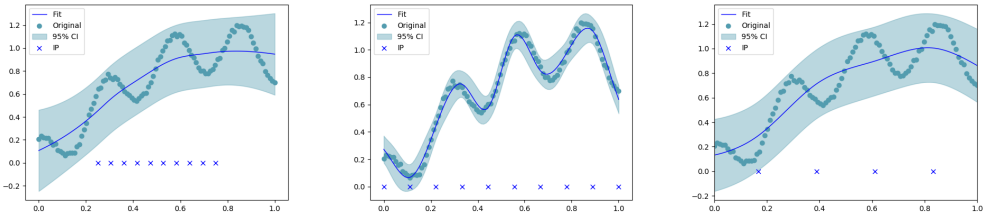

Figure 1: Stochastic variational Gaussian process on 1D synthetic data with different setting inducing inputs.

Length	0.5	1	2
ℓ	0.5226	1.7514	0.5400
RMSE	0.1424	0.0383	0.1406

Table 1: Log likelihood (ℓ) and root mean square error (RMSE) for model prediction on 1-D synthetic data with different setting of fixed inducing points.

.

3 Regularization for Latent Sparse Gaussian Processes

The Gaussian process latent variable model is a powerful dimensionality reduction approach. However, due to its poor scaling properties, the sparse Gaussian process is introduced in [25, 10].

Suppose $Y \in \mathbb{R}^{N\times D}$ is the observed data with latent variables $F \in \mathbb{R}^{N\times D}$, where N is the number of observations and D the dimension of the observations. Let the observations have corresponding latent variables $X \in \mathbb{R}^{N\times Q}$ where Q is the dimension of the latent space. Assuming independence across features, the GPLVM is

$$\begin{aligned} y_{nd}|f_{nd} &\sim \mathcal{N}(y_{nd}|f_{nd}, \sigma^2 = \beta^{-1}) \\ f_{nd} &= \mathcal{F}_d(\boldsymbol{x}_n) \\ \mathcal{F}_d &\overset{iid}{\sim} \mathrm{GP}(0, C(\boldsymbol{\theta})) \end{aligned} \tag{1}$$

with a normal prior for the latent variables X, $p(X) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|\mathbf{0}, I_Q)$ where $\boldsymbol{x}_n$ is the nth row of X. [24, 25] use a variational sparse GP formulation by introducing D separate sets of M inducing variables $U \in \mathbb{R}^{M\times D}$ evaluated at a set of inducing inputs $Z \in \mathbb{R}^{M\times Q}$. Then [25, 10] propose the same variational structure

$$q(F, U, X) = \prod_{d=1}^{D} \left(p(\boldsymbol{f}_d|\boldsymbol{u}_d, X)q(\boldsymbol{u}_d)\right) q(X)\,, \tag{2}$$

where $\boldsymbol{f}_d$ is the dth column of F and $\boldsymbol{u}_d$ is the dth column of U. Specifically, X has variational distribution $q(X) = \prod_{n=1}^{N} \mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_n, \Sigma_n)$ and U has variational distribution $q(U) = \prod_{d=1}^{D} \mathcal{N}(\boldsymbol{u}_d|\boldsymbol{m}_d, S_d)$. Then the evidence lower bound (ELBO) is

$$\mathrm{ELBO} = \sum_{d=1}^{D} E_{q(F,U,X)} p(\boldsymbol{y}_d|\boldsymbol{f}_d) - \mathrm{KL}(q(U)||p(U)) - \mathrm{KL}(q(X)||p(X))\,. \tag{3}$$

[25] derives the variational bound by marginalizing the optimal $q(U)$ based on the SGPR in [24], while [10] derives that by maximizing both parameterized $q(U)$ and $q(X)$ based on a SVGP. In the reminder of this paper, we discuss the regularization based on a SVGP, due to its computational merit.

For some larger and more complicated datasets, stochastic variational inference may fail to capture the distribution of latent inputs. We provide an example in Section 5.1. To address this concern, we next propose an innovative regularization approach.

3.1 Regularization

In order to ensure the inducing inputs capture the distribution of the latent inputs, it is necessary to propose a way to quantify the difference between the distribution of the inducing inputs and the distribution of the latent inputs, and penalize this difference in the objective function. We define the modified evidence lower bound as

$$\mathrm{MELBO} = \mathrm{ELBO} - \lambda R \tag{4}$$

where λ is a regularization weight and R is a regularization term which measures the difference between the distribution of the latent inputs X and the distribution of the inducing inputs Z. As λ increases, the optimization emphasizes more similarity in the distributions.

Specifically, we build a global model for the variational mean of X such that every $\boldsymbol{\mu}_n$ has an independent identical Gaussian distribution $p_X(\boldsymbol{\mu}_n) = \mathcal{N}(\boldsymbol{\mu}_n|\boldsymbol{\mu}_{\boldsymbol{\mu}}, \Sigma_{\boldsymbol{\mu}})$, and build another global

model for the inducing points Z such that every $\boldsymbol{z}_m$ has an independent identical distribution $p_Z(\boldsymbol{z}_m) = \mathcal{N}(\boldsymbol{z}_m|\boldsymbol{\mu}_Z, \Sigma_Z)$. Then given $\boldsymbol{\mu}$ and Z, we derive the maximum likelihood estimates $\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}, \hat{\Sigma}_{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\mu}}_Z$ and $\hat{\Sigma}_Z$ using the mean and covariance matrix of $\{\boldsymbol{\mu}_n\}$ and $\{\boldsymbol{z}_m\}$. We derive $q_X = \mathcal{N}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}})$ to summarize the global distribution of the latent inputs and derive $q_Z = \mathcal{N}(\hat{\boldsymbol{\mu}}_Z, \hat{\boldsymbol{\Sigma}}_Z)$ to summarize the global distribution of the inducing inputs Z.

We define the regularization term R by the Kullback-Leibler divergence between q_X and q_Z:

$$R = \mathrm{KL}(q_Z||q_X)\,. \tag{5}$$

In 4, λ can be chosen by cross validation or be set as the number of inducing points as a rule of thumb. As $\lambda = M$, we justify that performing VI on the GPLVM with regularization is equivalent to performing VI on a related empirical Bayes model with a prior on inducing inputs in Section 3.2.

3.2 Regularization Theory

This section discusses the underlying relationship between regularization in GPLVM and an empirical Bayesian model. First, we display a related empirical Bayesian model with a prior on its inducing inputs Z and derive its variational lower bound. Then we illustrate that maximizing the MELBO is equivalent to maximizing the variational lower bound in the empirical Bayesian model.

The related empirical Bayesian model is extended from (1) and (2). We put an informative prior on the inducing inputs and propose a variational distribution on them as

$$\begin{aligned} \boldsymbol{z}_m &\sim \mathcal{N}(\boldsymbol{z}_m|\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}_{\boldsymbol{\mu}}) \\ q(\boldsymbol{z}_m) &= \mathcal{N}(\boldsymbol{z}_m|\boldsymbol{\nu}_m, \Upsilon_m) \end{aligned}$$

where $\hat{\boldsymbol{\mu}}_X, \hat{\Sigma}_X$ are mean and covariance matrix of $\boldsymbol{\mu}$. The variational joint distribution is defined as $q(F, U, X, Z) = q(Z)q(X)q(U)p(F|Z, X, U)$. Then variational lower bound is derived as

$$\log p(Y) \geq \mathrm{E}_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(Z)||p(Z)) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U))$$

We define $\hat{\boldsymbol{\mu}}_{\boldsymbol{\nu}}$ and $\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}$ as the mean and covariance matrix of $\{\boldsymbol{\nu}_m\}$ and define a distribution family for $q(Z)$ such that $\Upsilon_m = \epsilon I$ for $m = 1, \ldots, M$. We assume the covariance of $\{\boldsymbol{\nu}_m\}$ is finite, which means $|\hat{\Sigma}_{\boldsymbol{\nu}}| < K$. Then we have following three lemmas and one theorem. The proofs are provided in the supplementary materials.

Lemma 1 *When $q(\boldsymbol{z}_m) = \mathcal{N}(\boldsymbol{\nu}_m, \epsilon I)$, as $\epsilon \to 0$, $\boldsymbol{z}_m \xrightarrow{p} \nu_m$.*

Lemma 2 *The variational lower bound in the empirical Bayesian model is derived as*

$$\log p(Y) \geq E_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A + B + C$$

where $A = \frac{M}{2}(\log|\hat{\Sigma}_{\boldsymbol{\mu}}| + \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}| + Q) + \frac{1}{2}\left(\sum_{m=1}^{M}(\boldsymbol{\nu}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{\nu}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right)$, $B = \frac{M}{2}(Q\log\epsilon - \log K)$ *and* $C = \frac{2\epsilon}{M\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1})}$.

Lemma 3 *We derive the regularization term as* $M\mathrm{KL}(q_Z||q_X) = \frac{M}{2}(\log|\hat{\Sigma}_{\boldsymbol{\mu}}| + \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}| + Q) + \frac{1}{2}\left(\sum_{m=1}^{M}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right)$.

Theorem 1 *As $\epsilon \to 0$, maximizing the variational lower bound in empirical Bayesian model is equivalent to maximizing the MELBO in the GPLVM with respect to $Z, q(X)$ and $q(U)$.*

4 Temporal Categorical Latent Gaussian Process

The temporal categorical latent Gaussian process (TCLGP) is extended from the categorical latent Gaussian process in [7]. We incorporate dynamic priors [15, 4] to model categorical time series. Assume we have observations $Y \in \mathbb{Z}^{N\times D\times T}$ with time stamps $B \in \mathbb{R}^{N\times T}$. N is the number of

individuals, D is the feature size and T is the number of time stamps. If we assume that each feature is categorical with K levels, then our model is expressed as:

$$\begin{aligned} y_{ndt} &\sim \mathrm{Cat}(\mathrm{Softmax}(\boldsymbol{f}_{ndt}))\,, \\ f_{ndtk} &= \mathcal{F}_{dk}(\boldsymbol{x}_{nt})\,, \qquad u_{mdk} = \mathcal{F}_{dk}(\boldsymbol{z}_m)\,, \\ \mathcal{F}_{dk}(\cdot) &\overset{iid}{\sim} \mathrm{GP}(0, C(\boldsymbol{\theta}_d)), \qquad \boldsymbol{x}_{ntq} = \boldsymbol{v}_{nq}(B_{nt})\,, \\ v_{nq}(t) &\overset{iid}{\sim} \mathrm{GP}(C(\boldsymbol{\phi}_q))\,, \end{aligned}$$

where the categorical data have embedding inputs $X \in \mathbb{R}^{N\times T\times Q}$ on a Q-dimensional latent space. The embedding inputs are latent vectors which summarize all the characteristics of the corresponding multi-dimensional categorical data.

4.1 Inference

The evidence lower bound of the TCLGP is expressed as

$$\log p(Y) \quad \geq \quad E_{q(F,X,U)} \log p(Y|F) - \mathrm{KL}(q(X)||p(X|B)) - \mathrm{KL}(q(U)||p(U))$$

where the variational distributions of U and X are constructed using independent Gaussian distributions such as

$$\begin{aligned} q(U) &= \prod_{d=1}^{D}\prod_{k=1}^{K} \mathcal{N}(\boldsymbol{u}_{dk}|\boldsymbol{m}_{dk}, S_d)\,, \\ q(X) &= \prod_{n=1}^{N}\prod_{q=1}^{Q}\prod_{t=1}^{T} \mathcal{N}(x_{nqt}|\mu_{nqt}, \sigma^2_{nqt})\,. \end{aligned}$$

Since there is no closed form for the expectation [7], we approximate the integration using Monte Carlo integration. For a large data set, we propose a stochastic variational inference algorithm with batch learning, with implementation details in Appendix B in the supplementary material.

4.2 Prediction

The TCLGP model has hyper-parameters $\boldsymbol{\theta}, \boldsymbol{\phi}, Z$ and variational parameters $\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \boldsymbol{m}, \boldsymbol{S}$. After model training, we get all estimates $\hat{\Theta} = (\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\phi}}, \hat{Z}, \hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\sigma}}, \hat{\boldsymbol{m}}, \hat{\boldsymbol{S}})$. We estimate the latent inputs X and the inducing variables U using their corresponding variational mean. Then given a new time stamp t^* for any individual n, we can estimate the corresponding latent input given $\hat{X}$ as

$$p(\boldsymbol{x}^*_n|\hat{X}) = \prod_{q=1}^{Q} \mathcal{N}(S_0 S_1^{-1}\hat{\boldsymbol{x}}_{n\cdot q}, C(t^*; \hat{\boldsymbol{\phi}}) - S_0 S_1^{-1} S_0^T) \tag{6}$$

where $S_0 = C(t^*, \boldsymbol{c}_n; \hat{\boldsymbol{\phi}}_q)$ and $S_1 = C(t^*, \boldsymbol{c}_n; \hat{\boldsymbol{\phi}}_q)C(\boldsymbol{c}_n; \hat{\boldsymbol{\phi}}_q)$. After taking the mean as the estimate of the latent inputs $\hat{\boldsymbol{x}}^*_n$, we estimate the corresponding outputs $\boldsymbol{f}^*_n$ by $\hat{\boldsymbol{f}}^*_n = E(\boldsymbol{f}^*_n|\hat{\boldsymbol{x}}^*_n, \hat{U})$. Specifically $\hat{f}^*_{ndk} = \hat{a}^*_{ndk}$, where $\hat{a}^*_{ndk} = \hat{\boldsymbol{v}}^{*T}_{nd}\hat{\Sigma}^{-1}_{Zd}\hat{\boldsymbol{m}}_{dk}$ and $\hat{\Sigma}_{Zd} = C(\hat{Z}; \hat{\boldsymbol{\theta}}_d)$, $\hat{\boldsymbol{v}}^*_{nd} = C(\hat{Z}, \hat{\boldsymbol{x}}^*_n; \hat{\boldsymbol{\theta}}_d)$. Therefore, the predictive distribution is estimated as $\hat{p}(\boldsymbol{y}^*_n = \boldsymbol{y}|\hat{\Theta}, t^*) = \prod_{d=1}^{D} \mathrm{Softmax}(\hat{\boldsymbol{f}}^*_n)[y_d]$.

5 Experiments

We illustrate our regularization on three datasets. First, we show that regularization is necessary in stochastic variational inference of a GPLVM using the Anuran Calls dataset. Second, we generate categorical time series data from a simple Markov model and demonstrate regularization is particularly important for the TCLGP model, even with low-dimensional data. Finally, we apply the TCLGP with regularization on a real dataset of stock indices and visualize the latent dynamics.

5.1 Anuran Calls Example

GPLVM is a powerful dimensionality reduction approach [14, 5] and it is a base model for many sophisticated models [15, 26, 4]. Thus inference for a GPLVM is important, but when dealing with

complicated data, inference can be intractable due to the non-convex optimization problem. Here we show that regularization improves inference on complicated datasets such as the Anuran Call data set. This data set is available from the UCI repository at `https://archive.ics.uci.edu/ml/datasets/Anuran+Calls+(MFCCs)`, where there are 7195 instances, and each instance has 22 attributes and belongs to one of eight Genus types. We model all instances using GPLVM and perform inference with regularization. Specifically, we set the latent dimension $Q = 5$ and use $M = 20$ inducing points in the latent SVGP model. We choose independent standard Gaussian distributions as the prior distributions of the inducing points and use the PCA approach for initialization of the inducing inputs.

Under different regularization weights λ, the latent means are displayed in Figure 2 using the two most informative dimensions in the automatic relevance determination (ARD) kernel. It suggests that as λ increase, inducing inputs are going to better capture latent inputs. When λ increase large enough, the scale inducing inputs become invariant as shown for $\lambda = 10^3$ and λ^7. After enough iterations, both ELBO and MEBLO values for different λs are shown in Figure 2. It shows that initially ELBO increases and model fit become better as λ increase due to suitable regularization. But when λ goes too large, ELBO decreases and model fit becomes worse. This is because MELBO emphasizes too much on the balance of inducing inputs and latent inputs and ignore likelihood information in ELBO. Overall, it suggests that for complicated data such as the Anuran Calls data set, suitable regularization contributes to better optimization and better model fit.

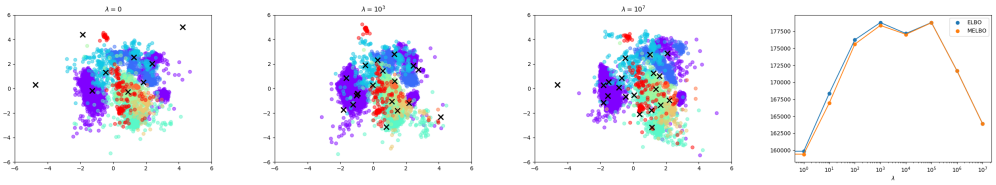

Figure 2: Latent means of GPLVM for Anuran Call data set. Left panel refers to no regularization and the other three have regularization with different weights $\lambda = 10^3, 10^7$. Different colors denote different genus types and crosses denote inducing inputs.

5.2 Synthetic Time Series

Previous experiments have already shown regularization is useful in optimization when dealing with high dimensional data. Here, we show that regularization is also crucial for model prediction, even when dealing with low dimensional data in a TCLGP model. We also let latent dimensional size Q be greater than data dimensional size to make model more flexible.

In this section, we propose a simple Markov model to generate categorical time-series. Suppose we have categorical data with two dimensions, in which the first dimension has two levels and the second dimension has three levels. Totally, we have six observable categories. The two dimensions are modeled independently. We propose transition matrices in the two dimensions as $\begin{bmatrix} 0.9 & 0.1 \\ 0.1 & 0.9 \end{bmatrix}$, and $\begin{bmatrix} 0.7, & 0.2 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.2 & 0.7 \end{bmatrix}$ respectively. We simulate initial levels uniformly and then generate time series of size 10 from the simple Markov model and set time stamps evenly distributed over an unit interval.

For each time series, we take first 9 data points for training and take the last data point for testing. We train our model using ARD kernels for all GPs, with and without regularization. We set $M = 20$ inducing points and take latent dimension size $Q = 5$. We initialize the inducing inputs with independent standard normal distributions and initialize the latent input mean with zeros. After training, the latent variables of the training data are displayed in Figure 3 using the two most informative dimensions based on ARD kernels. Specifically, for each dimensional, we select the two most informative dimensions with respect to the two smallest length-scale values in $\boldsymbol{\theta}_d$. This illustrates that the inducing inputs have bad coverage for the latent inputs without regularization. As λ increases, the coverage becomes better because of more emphasis on the balance of inducing inputs.

We repeatedly run the whole experiment 100 times with random initialization. ELBO, MELBO and predictive accuracy (PA) values are summarized by mean and standard deviation in Table 2. This table shows the ELBO decreases as λ increases, because of more penalty on regularization term. In contrast with previous high dimensional data, ELBO can be easier to approach maximum without regularization in low dimensional data. Although ELBO achieves the maximum without regularization in Table 2, suitable regularization ($\lambda = 50$) of balance between inducing inputs and latent inputs is still crucial for model prediction, which contributes to higher predictive accuracy. It illustrates that regularization with suitable λ contributes to better model prediction.

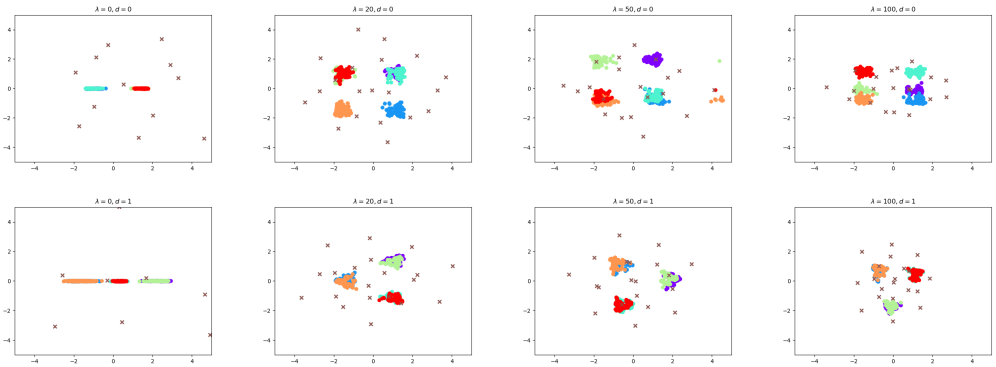

Figure 3: Latent variables of TCLGP for synthetic data. Left panel refers to no regularization and the other three have regularization with different weights $\lambda = 20, 50, 100$. Upper figures are plotted using the two most informative dimensions via $\boldsymbol{\theta}_0$ and lower figures are plotted using the two most informative dimensions via $\boldsymbol{\theta}_1$. Different colors denote distinct observable categories and crosses denote inducing inputs.

λ	0	20	50	100
ELBO	**-1543.97**(24.29)	-1593.23(18.46)	-1606.25(19.85)	-1618.57(23.81)
MELBO	-1543.97(24.29)	-1604.31(18.65)	1616.36(20.06)	-1627.14(23.98)
PA	74.67%(4.86%)	75.69%(4.65%)	**76.66%**(4.33%)	75.63%(5.01%)

Table 2: Evidence lower bound (ELBO), modified evidence lower bound (MELBO) and predictive accuracy (PA) of TCLGPs with and without regularization for synthetic categorical time series. Values are displayed by the mean (outside bracket) and standard deviation (inside bracket) of 100 repeated experiments with random initialization.

.

5.3 Stock Index

In this section, we apply the TCLGP model with regularization to a real data set of stock indices.

We focus on three indices, the SP500, the Nikkei225, and the DAX. Index records are from January 6, 1965 to December 5, 2012. We use the same data as in [18], but we take monthly stock indices and treat each year as an independent time series. Letting $\boldsymbol{y}_{nd}$ be monthly stock values for the nth year's and the dth stock indices, we have 48 annual time series, where each time series contains three stock indices for 12 months. We pre-process the data by evenly splitting the monthly return rates into five levels, via $20\%, 40\%, 60\%, 80\%$ quantiles. Levels 1-5 represent bear, slight bear, normal, slight bull and bull markets. For each time series we randomly select one month as the testing data and treat the remaining data as the training data.

5.3.1 Hyper-parameter Analysis

In this dataset, the model is sensitive to the hyper-parameters of the GP of the latent function prior. We fixed the hyper-parameters and optimized the other model parameters with regularization. We take $M = 20$ and assume the same $\boldsymbol{\phi}$ for each dimension q. We try different settings of $\boldsymbol{\phi}$ and set

$\lambda = 20$ using the rule of thumb. We evaluate model fit using the MELBO. Predictive performance is evaluated using training/testing predictive accuracy (TRPA/TPA) and training/testing mean absolute difference (TRMAD/TMAD), which are defined as follows:

$$\begin{aligned}
\text{TRPA} &= \frac{1}{N(T-1)} \sum_{n=1}^{N} \sum_{t=1}^{T-1} \mathbf{1}(\boldsymbol{y}_{n\cdot t}^{train} = \hat{\boldsymbol{y}}_{n\cdot t}^{train}), \\
\text{TRMDA} &= \frac{1}{ND(T-1)} \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T-1} |y_{ndt}^{train} - \hat{y}_{ndt}^{train}|, \\
\text{TPA} &= \frac{1}{N} \sum_{n=1}^{N} \mathbf{1}(\boldsymbol{y}_{n\cdot}^{test} = \hat{\boldsymbol{y}}_{n\cdot}^{test}), \\
\text{TMDA} &= \frac{1}{ND} \sum_{n=1}^{N} \sum_{d=1}^{D} |y_{nd}^{test} - \hat{y}_{nd}^{test}|.
\end{aligned}$$

Results can be found in Table 3.

ϕ_{σ^2}	0.5			2		
ϕ_l	0.01	0.05	0.1	0.01	0.05	0.1
MELBO	-2698.79	**-2671.42**	-2947.10	-2691.02	-2710.76	-2920.42
TRPA	0.69	0.70	0.71	**0.72**	0.71	0.71
TRMAD	0.45	**0.42**	0.49	0.46	0.48	0.50
TPA	0.21	**0.24**	0.18	0.21	0.21	0.18
TMAD	1.53	**1.45**	1.65	1.53	1.53	1.65

Table 3: Model fitting and prediction results of TCLGP with regularization under different settings of hyper-parameters. We evaluate model fitting by modified evidence lower bound (MELBO) and evaluate model prediction by training predictive accuracy (TRPA), training mean absolute difference (TRMAD), testing predictive accuracy (TRA) and testing mean absolute difference (TMAD).

Table 3 shows some examples of hyper-parameter values and that with $\phi_{\sigma^2} = 0.5$ and $\phi_l = 0.05$, the model has the best fitting and best performance on almost all measures of predictive accuracy. Therefore, we choose this setting as the optimal setting and continue on to latent process visualization.

5.3.2 Latent Processes Visualization

Here we use the optimal model from the previous section. To visualize the latent process, we denote the predictive categories given a certain latent space $\boldsymbol{x}^*$ as $\boldsymbol{y}^* \in [0, \ldots, 4]^3$. We let $\tilde{y}^* = \sum \boldsymbol{y}^* \in [0, \ldots, 12]$ to represent market status. The larger value represents that it is more likely to be a bull market. Then we plot a contour using $(\boldsymbol{x}^*, \tilde{y}^*)$ on the latent space shown in Figure 4. On the surface, light colors indicate a bull market while dark colors indicate a bear market. Furthermore, we display predictive posterior latent processes as well as estimated latent traces for both Year 2008 and Year 2009 in Figure 4. The estimated latent traces show that it always stays in darker areas in 2008 while it always stays in lighter areas in 2009. The result exactly matches the fact that a financial crisis happened in 2008 leading the US stock market to a bear market while the US economy returned to normal in 2009. Predictive sensitivity for all years is captured by the predictive posterior processes in the middle two columns in Figure 4.

6 Conclusion

Regularization for a GPLVM is necessary when dealing with complicated high dimensional data, improving global optimization in model fitting. The use of regularization is also justified by proving that performing VI on a GPLVM with this regularization is equivalent to performing VI on a related empirical Bayes model. Without cross validation for the regularization weight λ, we give a rule of thumb for the selection of regularization by setting $\lambda = M$. We propose the temporal categorical latent Gaussian process model and illustrate that our regularization is particularly important in this latent dynamic model. Finally, we illustrate its abilities on the real dataset of stock indices.

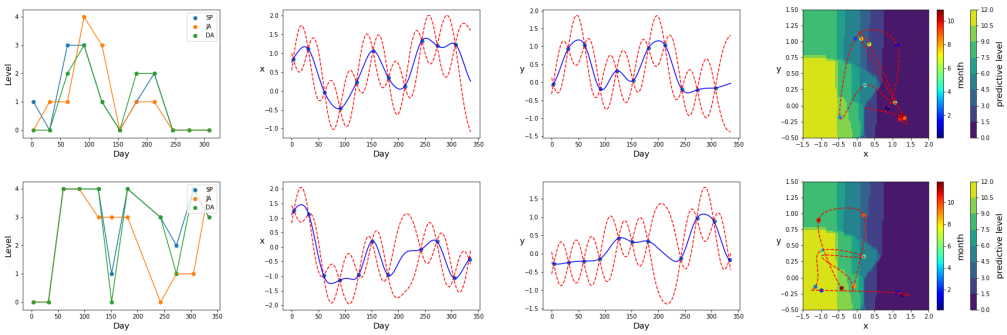

Figure 4: Categorical plot (left) and predictive posterior latent process (middle) and estimated latent tracing (right) of stock indices in Year 2008 and Year 2009.

References

[1] Banerjee, S., Gelfand, A. E., Finley, A. O., and Sang, H. (2008). Gaussian predictive process models for large spatial data sets.

[2] Bernhard, S., Alexander, S., and Klaus-Robert, M. (1998). Kernel principle component analysis. *Advances in Kernel Methods - Support Vector Learning*, pages 327–352.

[3] Bowman, S. R., Vilnis, L., Vinyals, O., Dai, A. M., Jozefowicz, R., and Bengio, S. (2015). Generating Sentences from a Continuous Space. *ArXiv e-prints*.

[4] Damianou, A. C., Titsias, M. K., and Lawrence, N. D. (2016). Variational inference for latent variables and uncertain inputs in gaussian processes. *Journal of Machine Learning Research*, 17(42):1–62.

[5] Ek, C. H., Torr, P. H. S., and Lawrence, N. D. (2007). Gaussian process latent variable models for human pose estimation. In *MLMI*.

[6] Finley, A. S., Huiyan Banerjee, S., and Gelfand, A. (2009). Improving the performance of predictive process modeling for large datasets. *Computational statistics and data analysis*.

[7] Gal, Y., Chen, Y., and Ghahramani, Z. (2015). Latent Gaussian Processes for Distribution Estimation of Multivariate Categorical Data. *ArXiv e-prints*.

[8] Guhaniyogi, R., Finley, A., Banerjee, S., and Gelfand, A. (2011). Adaptive gaussian predictive process models for large spatial datasets. *Environmetrics*.

[9] Haining, R. (1993). Statistics for spatial data. *Computers and Geosciences*, 19:615–616.

[10] Hensman, J., Fusi, N., and Lawrence, N. D. (2013). Gaussian Processes for Big Data. *ArXiv e-prints*.

[11] Hensman, J., Rattray, M., and Lawrence, N. D. (2012). Fast Variational Inference in the Conjugate Exponential Family. *ArXiv e-prints*.

[12] L., C. and Opper, M. (2002). Sparse online gaussian processes. *Neural Computation*, pages 641–669.

[13] Lawrence, N. (2005). Probabilistic non-linear principal component analysis with gaussian process latent variable models. *J. Mach. Learn. Res.*, 6:1783–1816.

[14] Lawrence, N. D. (2003). Gaussian process latent variable models for visualisation of high dimensional data. In *NIPS*.

[15] Lawrence, N. D. and Moore, A. J. (2007). Hierarchical gaussian process latent variable models. In *Proceedings of the 24th International Conference on Machine Learning*, ICML '07, pages 481–488, New York, NY, USA. ACM.

[16] Lawrence, N. D. and Quiñonero Candela, J. (2006). Local distance preservation in the gp-lvm through back constraints. In *Proceedings of the 23rd International Conference on Machine Learning*, ICML '06, pages 513–520, New York, NY, USA. ACM.

[17] Mardia, K. V., Kent, J. T., and Bibby, J. M. (1979). Multivariate analysis.

[18] Nicolau, J. (2014). A new model for multivariate markov chains. *Scandinavian Journal of Statistics*, 41(4):1124–1135.

[19] Rasmussen, C. and Kuss, M. (2004). Gaussian processes in reinforcement learning. In *Advances in Neural Information Processing Systems 16*, pages 751–759, Cambridge, MA, USA. Max-Planck-Gesellschaft, MIT Press.

[20] Rasmussen, C. E. and Williams, C. K. I. (2005). *Gaussian Processes for Machine Learning (Adaptive Computation and Machine Learning)*. The MIT Press.

[21] Seeger, M. (2003). Pac-bayesian generalisation error bounds for gaussian process classification. *J. Mach. Learn. Res.*, 3:233–269.

[22] Snelson, E. and Ghahramani, Z. (2006). Sparse gaussian processes using pseudo-inputs. In Weiss, Y., Schölkopf, B., and Platt, J. C., editors, *Advances in Neural Information Processing Systems 18*, pages 1257–1264. MIT Press.

[23] Snelson, E. and Ghahramani, Z. (2007). Local and global sparse gaussian process approximations. In Meila, M. and Shen, X., editors, *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics*, volume 2 of *Proceedings of Machine Learning Research*, pages 524–531, San Juan, Puerto Rico. PMLR.

[24] Titsias, M. (2009). Variational learning of inducing variables in sparse gaussian processes. In van Dyk, D. and Welling, M., editors, *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, volume 5 of *Proceedings of Machine Learning Research*, pages 567–574, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA. PMLR.

[25] Titsias, M. and Lawrence, N. D. (2010). Bayesian gaussian process latent variable model. In Teh, Y. W. and Titterington, M., editors, *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, volume 9 of *Proceedings of Machine Learning Research*, pages 844–851, Chia Laguna Resort, Sardinia, Italy. PMLR.

[26] Urtasun, R. and Darrell, T. (2007). Discriminative gaussian process latent variable model for classification. In *ICML*.

The source code for re-running all the experiments detailed here and all data are available from `https://github.com/Corleno/RGPLVM`.

A Regularization Theorems

Lemma A.1 *When* $q(\boldsymbol{z}_m) = \mathcal{N}(\boldsymbol{\nu}_m, \epsilon I)$*, as* $\epsilon \to 0$*,* $\boldsymbol{z}_m \xrightarrow{p} \boldsymbol{\nu}_m$*.*

Proof A.1 *Since*

$$\begin{aligned}\lim_{\epsilon\to 0} p(|\boldsymbol{z}_m - \boldsymbol{\nu}_m| > \epsilon_0) &= \lim_{\epsilon\to 0} p(|\frac{\boldsymbol{z}_m - \boldsymbol{\nu}_m}{\epsilon}| > \frac{\epsilon_0}{\epsilon}) \\ &= 2\lim_{\epsilon\to 0}(1 - \Phi(\frac{\epsilon_0}{\epsilon}))^Q \\ &= 0\,,\end{aligned}$$

we conclude that $\boldsymbol{z}_m \xrightarrow{p} \boldsymbol{\nu}_m$*.*

Lemma A.2 *The variational lower bound in the empirical Bayesian model is derived as*

$$\log p(Y) \geq E_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A + B + C$$

where $A = \frac{M}{2}(\log|\hat{\Sigma}_{\boldsymbol{\mu}}| + \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}| + Q) + \frac{1}{2}\left(\sum_{m=1}^{M}(\boldsymbol{\nu}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{\nu}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right)$*,* $B = \frac{M}{2}(Q\log\epsilon - \log K)$ *and* $C = \frac{2\epsilon}{M\mathrm{tr}(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1})}$*.*

Proof A.2

$$\begin{aligned}\log p(\boldsymbol{Y}) &\geq E_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(\boldsymbol{X})||p(\boldsymbol{X})) - \mathrm{KL}(q(\boldsymbol{U})||p(\boldsymbol{U})) - \mathrm{KL}(q(\boldsymbol{Z})||p(\boldsymbol{Z})) \\ &= E_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(\boldsymbol{X})||p(\boldsymbol{X})) - \mathrm{KL}(q(\boldsymbol{U})||p(\boldsymbol{U})) - A \\ &\quad + \frac{M}{2}(Q\log\epsilon - \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{\nu}}|) + C \\ &\geq E_{q(F,U,X,Z)} \log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A + B + C\end{aligned}$$

Lemma A.3 *We derive the regularization term as* $M\mathrm{KL}(q_Z||q_X) = \frac{M}{2}(\log|\hat{\Sigma}_{\boldsymbol{\mu}}| + \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}| + Q) + \frac{1}{2}\left(\sum_{m=1}^{M}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right)$*.*

Proof A.3

$$
\begin{aligned}
\mathrm{KL}(q_Z||q_X) &= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \mathrm{tr}(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}\hat{\Sigma}_Z) + (\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}((\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)(\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)^T + \hat{\Sigma}_Z)\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(M(\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)(\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}} - \hat{\boldsymbol{\mu}}_Z)^T + \sum_{m=1}^{M}(\boldsymbol{u}_m - \hat{\boldsymbol{\mu}}_Z)(\boldsymbol{u}_m - \hat{\boldsymbol{\mu}}_Z)^T\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(M\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T - M\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_Z^T - M\hat{\boldsymbol{\mu}}_Z\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T + M\hat{\boldsymbol{\mu}}_Z\hat{\boldsymbol{\mu}}_Z^T\right.\right. \\
&\quad \left.\left. + \sum_{m=1}^{M}\boldsymbol{u}_m\boldsymbol{u}_m^T - (\sum_{m=1}^{M}\boldsymbol{z}_m)\hat{\boldsymbol{\mu}}_Z^T - \hat{\boldsymbol{\mu}}_Z(\sum_{m=1}^{M}\boldsymbol{z}_m)^T + M\hat{\boldsymbol{\mu}}_Z\hat{\boldsymbol{\mu}}_Z^T\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\sum_{m=1}^{M}\boldsymbol{z}_m\boldsymbol{z}_m^T - M\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_Z^T - M\hat{\boldsymbol{\mu}}_Z\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T + M\hat{\boldsymbol{\mu}}_X\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T)\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\sum_{m=1}^{M}\boldsymbol{z}_m\boldsymbol{z}_m^T - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}(\sum_{m=1}^{M}\boldsymbol{z}_m)^T - (\sum_{m=1}^{M}\boldsymbol{z}_m)\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T + M\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T)\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\sum_{m=1}^{M}\boldsymbol{z}_m\boldsymbol{z}_m^T - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}(\sum_{m=1}^{M}\boldsymbol{u}_m)^T - (\sum_{m=1}^{M}\boldsymbol{z}_m)\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T + M\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}\hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}}^T)\right)\right] \\
&= \frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + \frac{1}{M}\mathrm{tr}\left(\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\sum_{m=1}^{M}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T)\right)\right] \\
&= \frac{1}{M}\sum_{m=1}^{M}\frac{1}{2}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + (\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right].
\end{aligned}
$$

Therefore, $M\mathrm{KL}(q_Z||q_X) = \frac{1}{2}\sum_{m=1}^{M}\left[\log\frac{|\hat{\Sigma}_{\boldsymbol{\mu}}|}{|\hat{\Sigma}_Z|} - Q + (\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right] = \frac{M}{2}(\log|\hat{\Sigma}_{\boldsymbol{\mu}}| + \log|\hat{\boldsymbol{\Sigma}}_{\boldsymbol{z}}| + Q) + \frac{1}{2}\left(\sum_{m=1}^{M}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})^T\hat{\Sigma}_{\boldsymbol{\mu}}^{-1}(\boldsymbol{z}_m - \hat{\boldsymbol{\mu}}_{\boldsymbol{\mu}})\right)$.

Theorem A.1 *As $\epsilon \to 0$, maximizing the variational lower bound in empirical Bayesian model is equivalent to maximizing the MELBO in the GPLVM with respect to $Z, q(X)$ and $q(U)$.*

Proof A.4 *In the empirical Bayesian model, variational parameters are $\boldsymbol{\mu}, \boldsymbol{\Sigma}, \boldsymbol{m}, \boldsymbol{s}, \boldsymbol{\nu}$. We denote all parameters as Θ. We have $\lim_{\epsilon\to 0} C = 0$ and*

$$
\lim_{\epsilon\to 0} E_{q(F,U,X,Z)}\log p(Y|F) = E_{q(F,U,X|Z=\boldsymbol{\nu})}\log p(Y|F) \tag{7}
$$

Then according to Lemma 2, we derive that

$$
\begin{aligned}
&\arg\max_{\Theta}\lim_{\epsilon\to 0} E_{q(F,U,X,Z)}\log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A + B + C \\
=\ &\arg\max_{\Theta}\lim_{\epsilon\to 0} E_{q(F,U,X,Z)}\log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A \\
=\ &\arg\max_{\Theta} E_{q(F,U,X|Z=\boldsymbol{\nu})}\log p(Y|F) - \mathrm{KL}(q(X)||p(X)) - \mathrm{KL}(q(U)||p(U)) - A
\end{aligned}
$$

When we replace $\boldsymbol{\nu}$ as Z, due to Lemma 3, this optimization is equivalent to maximize $\mathrm{ELBO} - M\mathrm{KL}(q_Z||q_X)$ *which is the exactly MELBO. Finally due to Lemma 1, the $q(Z)$ in empirical Bayesian model will converges to the same optimized Z in the GPLVM with regularization.*

B Stochastic Variational Inference Algorithm

This section displays a general stochastic variational inference algorithm for large datasets. In this framework, we set the number of training epochs N_{train} and evenly divide the whole dataset into

N_{batch} clusters. Each cluster includes the observations $\boldsymbol{Y}_i$ and their corresponding time stamp data B_i and their corresponding hyper-parameters of embedding inputs, $\boldsymbol{m}_i$ for the mean and $\boldsymbol{s}_i$ for the standard deviation. In the context of the TCLGP, the model parameters include $\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and the model inputs include both observable data $\boldsymbol{Y}_i, B_i$ and latent hyper-parameters $\boldsymbol{m}_i, \boldsymbol{s}_i$. Suppose the MELBO(g) is rewritten as

$$\begin{aligned}
\text{MELBO(g)} &= \text{ELBO(g)} - \lambda \text{R} \\
&= gR_0 + R_1 - \lambda \text{R}\,, \\
R_0 &= -\text{KL}(q(\boldsymbol{X})||p(\boldsymbol{X}|B)) - \text{KL}(q(\boldsymbol{U})||p(\boldsymbol{U}))\,, \\
R_1 &= \int q(\boldsymbol{X})q(\boldsymbol{U})p(\boldsymbol{F}|\boldsymbol{X},\boldsymbol{U}) \log \left(\sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{t=1}^{T} p(\boldsymbol{y}_{ndt}|\boldsymbol{f}_{ndt}) \right) d\boldsymbol{X}\boldsymbol{U}\boldsymbol{F}\,,
\end{aligned}$$

where R_0 and R_1 are the regularization term and the reconstruction term in the ELBO separately and R is a regularization term related to inducing inputs, and $g \in [0, 1])$ is the annealing factor referred in [3]. The annealing increase factor is denoted as Δg. Then inference algorithm is displayed as follows:

Set $g = 0$;
for $i = 1$ **to** N_{train} **do**
  for $j = 1$ **to** N_{batch} **do**
    Assign both observable data $\boldsymbol{Y}_i, B_i$ and latent data $\boldsymbol{m}_i, \boldsymbol{s}_i$ to the TCLGP model;
    Update model parameters $\boldsymbol{\theta}, \boldsymbol{Z}, \boldsymbol{\mu}, \boldsymbol{\Sigma}$ and latent hyper-parameters $\boldsymbol{m}_i, \boldsymbol{s}_i$ through maximizing the MELBO(g) using a stochastic gradient descend method.;
  end
  $g = \min(g + \Delta g, 1)$;
end

Algorithm 1: Stochastic variational inference algorithm for large datasets.