Collaborative Multi-Output Gaussian Processes for Collections of Sparse Multivariate Time Series

Steven Cheng-Xian Li **Benjamin Marlin**
College of Information & Computer Sciences
University of Massachusetts Amherst
{cxl,marlin}@cs.umass.edu

Abstract

Collaborative Multi-Output Gaussian Processes (COGPs) are a flexible tool for modeling multivariate time series. They induce correlation across outputs through the use of shared latent processes. While past work has focused on the computational challenges that result from a single multivariate time series with many observed values, this paper explores the problem of fitting the COGP model to collections of many sparse and irregularly sampled multivariate time series. This work is motivated by applications to modeling physiological data (heart rate, blood pressure, etc.) in Electronic Health Records (EHRs).

1 Introduction

Gaussian process (GP) regression is a well-known and widely-used approach for modeling temporal and spatial data [9]. The main drawback of GP models is the prohibitive cost of the required computations. To address this issue, Hensman et al. [4] recently introduced a scalable algorithm to perform GP inference based on a stochastic variational approximation [5]. Using a similar approach, Nguyen and Bonilla [7] proposed collaborative multi-output Gaussian processes (COGP) for efficiently learning multi-output GPs given a single multivariate time series with many observations. This work extends a long line of prior research on multi-output GPs [1, 2, 3, 11, 12].

In this paper, we consider the problem of learning the COGP model when the data consist of a collection of many sparse and irregularly sampled multivariate time series. This problem is motivated by the analysis of Intensive Care Unit (ICU) Electronic Health Records (EHR) data. In the ICU EHR setting, each patient is represented by an ensemble of sparse and irregularly sampled physiological time series, one per underlying physiological variable such as heart rate, blood pressure, etc. A typical ICU EHR record contains observations of physiological variables recorded at irregular intervals by clinical staff during the routine course of care. Key variables may have one to two recorded observations per hour, so the data are quite sparse. On the other hand, individual hospitals may have access to EHRs for many (thousands) of patients.

Our goal is to fit a common COGP model by leveraging the data from multiple patients. We present an extension to the COGP model and a modified variational learning algorithm that exploits the fact that we have many sparsely observed multivariate time series. We also explore the use of sparsity-inducing regularization on the factors controlling the interactions between outputs to deal with variables that are highly sparsely observed. We present predictive log likelihood results on a real ICU EHR data set.

2 Multi-Output Gaussian Processes

Consider a data set containing a collection of multivariate time series $\mathcal{D} = \{\mathcal{S}_1, \ldots, \mathcal{S}_N\}$. Each time series $\mathcal{S}_n$ consists of P channels $\mathcal{S}_n = \{(\mathbf{t}_{n1}, \mathbf{y}_{n1}), \ldots, (\mathbf{t}_{nP}, \mathbf{y}_{nP})\}$ in which $\mathbf{t}_{ni}$ is a set of time points and $\mathbf{y}_{ni}$ are the corresponding observed values. For ICU EHR data, each time series has only a small number of observations that are irregularly sampled. We extend the collaborative multi-output Gaussian processes (COGP) [7] to model correlation across different channels given a collection of multi-channel time series where each channel is sparse and irregularly sampled.

Let y_{nik} denote the kth observation of channel i from $\mathcal{S}_n$ measured at time t_{nik}, it is modeled as a noisy observation of the sum of a function h_i and a weighted combination of Q shared latent functions $g_1, \ldots, g_Q$ evaluated at t_{nik}, where each function has an independent Gaussian process (GP) prior $h_i \sim \mathcal{GP}\big(\mathbf{0}, k_i^{(h)}(\cdot,\cdot)\big)$ for $i = 1, \ldots, P$ and $g_j \sim \mathcal{GP}\big(\mathbf{0}, k_j^{(g)}(\cdot,\cdot)\big)$ for $j = 1, \ldots, Q$. Specifically,

$$p(y_{nik}) = \mathcal{N}\left(h_i(t_{nik}) + \sum_{j=1}^{Q} w_{ij} g_j(t_{nik}), \beta_i^{-1}\right)$$

Note that the hyperparameters of the covariance functions $k_i^{(h)}$ and $k_j^{(g)}$ are shared across the entire time series collection $\mathcal{D}$, and so are the weights w_{ij}. The shared Gaussian precision β_i^{-1} models the noise of the process that is shared by all of the time series in the ith channel.

In order to efficiently estimate the hyperparameters mentioned above, a set of M inducing time points $\mathbf{z} = [z_1, \ldots, z_M]$ is introduced to approximate the original GP posterior for all g_j and h_i. These inducing points provide a universal reference so that we can estimate the combination weights and other hyperparameters solely on the marginal distribution. Moreover, by choosing a smaller M those GPs can be sparsified to speed up computation [4, 5].

Let $\mathbf{g}_{nij} = g_j(\mathbf{t}_{ni})$ and $\mathbf{h}_{ni} = h_i(\mathbf{t}_{ni})$ for $n = 1, \ldots, N$, $i = 1, \ldots, P$ and $j = 1, \ldots, Q$. Like other GP approximations [8], a set of inducing variables $\mathbf{u}_j$ and $\mathbf{v}_i$ are introduced such that

$$\begin{aligned}
p(\mathbf{g}_{ni}|\mathbf{u}_n) &= \prod_{j=1}^{Q} p(\mathbf{g}_{nij}|\mathbf{u}_{nj}) = \prod_{j=1}^{Q} \mathcal{N}\big(\mathbf{g}_{nij}|\boldsymbol{\mu}_{nij}^{(g)}, \mathbf{K}_{nij}^{(g)}\big) \\
p(\mathbf{u}_n) &= \prod_{j=1}^{Q} p(\mathbf{u}_{nj}) = \prod_{j=1}^{Q} \mathcal{N}\big(\mathbf{u}_{nj}|\mathbf{0}, k_j^{(g)}(\mathbf{z},\mathbf{z})\big) \\
p(\mathbf{h}_n|\mathbf{v}_n) &= \prod_{i=1}^{P} p(\mathbf{h}_{ni}|\mathbf{v}_{ni}) = \prod_{i=1}^{P} \mathcal{N}(\mathbf{h}_{ni}|\boldsymbol{\mu}_{ni}^{(h)}, \mathbf{K}_{ni}^{(h)}) \\
p(\mathbf{v}_n) &= \prod_{i=1}^{P} p(\mathbf{v}_{ni}) = \prod_{i=1}^{P} \mathcal{N}\big(\mathbf{v}_{ni}|\mathbf{0}, k_i^{(h)}(\mathbf{z},\mathbf{z})\big)
\end{aligned}$$

where $\boldsymbol{\mu}_{nij}^{(g)}$ and $\mathbf{K}_{nij}^{(g)}$ are the posterior mean and covariance defined as follows and $\boldsymbol{\mu}_{ni}^{(h)}$ and $\mathbf{K}_{ni}^{(h)}$ are defined similarly.

$$\begin{aligned}
\boldsymbol{\mu}_{nij}^{(g)} &= k_j^{(g)}(\mathbf{t}_{ni},\mathbf{z}) k_j^{(g)}(\mathbf{z},\mathbf{z})^{-1} \mathbf{u}_{nj} \\
\mathbf{K}_{nij}^{(g)} &= k_j^{(g)}(\mathbf{t}_{ni},\mathbf{t}_{ni}) - k_j^{(g)}(\mathbf{t}_{ni},\mathbf{z}) k_j^{(g)}(\mathbf{z},\mathbf{z})^{-1} k_j^{(g)}(\mathbf{z},\mathbf{t}_{ni}).
\end{aligned}$$

In this work, we use squared exponential kernels for both $k_i^{(h)}$ and $k_j^{(g)}$, that is,

$$k_i^{(h)}(x, x') = a_i \exp\big(-b_i(x - x')^2\big), \quad \text{for } a_i > 0 \text{ and } b_i > 0$$

whereas for $k_j^{(g)}$ we fix the leading coefficient $a_j = 1$ since the weights w_{ij} control the scale already.

We use variational inference to estimate the parameters. Following the procedure of COGP, we can derive the evidence lower bound with all $\mathbf{g}_{nij}$ and $\mathbf{h}_{ni}$ collapsed as in [4] and introduce the mean field variational distributions $q(\mathbf{u}_{nj}) = \mathcal{N}\big(\mathbf{u}_{nj}|\mathbf{m}_{nj}^{(g)}, \mathbf{S}_{nj}^{(g)}\big)$ and $q(\mathbf{v}_{ni}) = \mathcal{N}\big(\mathbf{v}_{ni}|\mathbf{m}_{ni}^{(h)}, \mathbf{S}_{ni}^{(h)}\big)$ for all n, i, j. We obtain the lower bound shown below.

$$\log p(\mathcal{D}) \geq \sum_{n=1}^{N} \Big\{ \int q(\mathbf{u}_n, \mathbf{v}_n) \mathbb{E}_{p(\mathbf{g}_n, \mathbf{h}_n | \mathbf{u}_n, \mathbf{v}_n)} \Big[\log p(\mathbf{y}_n | \mathbf{g}_n, \mathbf{h}_n) \Big] d\mathbf{u}_n d\mathbf{v}_n$$

$$- \sum_{j=1}^{Q} D_{\mathrm{KL}}(q(\mathbf{u}_{nj}) \parallel p(\mathbf{u}_{nj})) - \sum_{i=1}^{P} D_{\mathrm{KL}}(q(\mathbf{v}_{ni}) \parallel p(\mathbf{v}_{ni})) \Big\}$$

Since we are working in the scenario that the number of samples in each channel of the ICU EHR is small, instead of updating the variational parameters of $\mathbf{u}_{nj}$ and $\mathbf{v}_{ni}$ using stochastic optimization as in [7], we can estimate them analytically in the variational E-step to speed up the overall convergence. Specifically, we estimate $\mathbf{S}_{nj}^{(g)}$ and $\mathbf{S}_{ni}^{(h)}$ individually in closed form by setting the derivatives of the evidence lower bounds to zero:

$$\mathbf{S}_{nj}^{(g)*} = \left(k_j^{(g)}(\mathbf{z}, \mathbf{z})^{-1} + \sum_{i=1}^{P} \beta_i w_{ij}^2 \mathbf{A}_{nij}^{(g)\top} \mathbf{A}_{nij}^{(g)} \right)^{-1}$$

$$\mathbf{S}_{ni}^{(h)*} = \left(k_i^{(h)}(\mathbf{z}, \mathbf{z})^{-1} + \beta_i \mathbf{A}_{ni}^{(h)\top} \mathbf{A}_{ni}^{(h)} \right)^{-1}$$

where $\mathbf{A}_{nij}^{(g)} = k_j^{(g)}(\mathbf{t}_{ni}, \mathbf{z}) k_j^{(h)}(\mathbf{z}, \mathbf{z})^{-1}$ and $\mathbf{A}_{ni}^{(h)} = k_i^{(h)}(\mathbf{t}_{ni}, \mathbf{z}) k_i^{(h)}(\mathbf{z}, \mathbf{z})^{-1}$.

As for $\mathbf{m}_{nj}^{(g)}$ and $\mathbf{m}_{ni}^{(h)}$, we can estimate all of them *jointly* by solving the following linear system.

$$\left(\mathbf{S}_{nj}^{(g)*} \right)^{-1} \mathbf{m}_{nj}^{(g)} = \sum_{i=1}^{P} \beta_i w_{ij} \mathbf{A}_{nij}^{(g)\top} \left(\mathbf{y}_{ni} - \mathbf{A}_{ni}^{(h)} \mathbf{m}_{ni}^{(h)} - \sum_{k \neq j} w_{ik} \mathbf{A}_{nik}^{(g)} \mathbf{m}_{nk}^{(g)} \right), \text{ for all } j$$

$$\left(\mathbf{S}_{ni}^{(h)*} \right)^{-1} \mathbf{m}_{ni}^{(h)} = \beta_i w_{ij} \mathbf{A}_{ni}^{(h)\top} \left(\mathbf{y}_{ni} - \sum_{j=1}^{Q} w_{ij} \mathbf{A}_{nij}^{(g)} \mathbf{m}_{nj}^{(g)} \right), \text{ for all } i$$

3 Experiment and Results

We evaluate the performance of our extension of the multi-output GP model (COGP) using predictive likelihood on held out data. Our experiments are based on a pediatric ICU EHR data set collected at the Children's Hospital of Los Angeles. The data contain sparse and irregularly sampled time series for 13 standard physiological variables. The data set we use for these experiments contains a collection of 1000 patient records. We extract the samples from the first 24 hours in each episode. The average number of observations per day varies between 7 and 50 for these variables with considerable variation between patients.

We compare the predictive performance on the held-out data points using the COGP with different regularization schemes. We also compare to a baseline method that models each channel as an independent GP (INDEP-GP). We randomly split the 1000 episodes into 500 for training and test on the remaining half. For each channel, we hold out the middle one-third of the observations of each episode to evaluate the predictive distribution on the held-out time points, so that inference has to account for information from other channels due to the lack of reference in the neighborhood. This involves estimating $\mathbf{m}_j^{(g)}, \mathbf{S}_j^{(g)}, \mathbf{m}_i^{(h)}, \mathbf{S}_i^{(h)}$ for each test case given the shared hyperparameters that

Table 1: Held-out log-likelihood comparison

method	average log-likelihood	Q	regularization parameter
COGP-COL	-0.674 (± 0.045)	3	$\tau = 0.2$
COGP-ROW	-0.688 (± 0.036)	3	$\lambda = 2.0$
COGP-IND	-0.747 (± 0.043)	5	$\lambda = 0.8$
COGP	-0.748 (± 0.042)	3	–
INDEP-GP	-2.653 (± 0.171)	–	–

Table 2: Average log-likelihood on each channel

channel	COGP-COL	COGP	INDEP-GP	# test in use	avg length
SpO2	-0.92 (±0.06)	-1.22 (±0.10)	-4.51 (±0.29)	4871	27.3
HR	-0.92 (±0.10)	-1.19 (±0.14)	-5.84 (±0.45)	4879	27.3
RR	-0.04 (±0.01)	0.01 (±0.01)	-0.77 (±0.16)	2410	12.4
sBP	-0.53 (±0.10)	-0.53 (±0.10)	-2.46 (±0.34)	1646	9.2
dBP	-0.88 (±0.02)	-1.37 (±0.04)	-0.78 (±0.15)	956	4.9
EtCO2	0.04 (±0.01)	0.09 (±0.01)	-0.97 (±0.17)	2553	13.2
Temp	-0.82 (±0.36)	-0.80 (±0.37)	-0.25 (±0.19)	508	3.1
TGCS	-0.58 (±0.08)	-0.58 (±0.08)	-4.48 (±0.29)	5682	32.2
CRR	-0.71 (±0.05)	-0.75 (±0.07)	-0.85 (±0.19)	418	2.3
UO	-1.13 (±0.08)	-1.14 (±0.08)	-6.32 (±0.29)	5708	32.3
FiO2	-1.75 (±0.28)	-1.76 (±0.29)	-5.57 (±0.89)	5949	33.7
Gluc	-0.05 (±0.01)	-0.01 (±0.01)	-0.57 (±0.06)	2359	12.2
pH	-0.48 (±0.05)	-0.47 (±0.05)	-1.11 (±0.14)	1997	10.3

have been trained. Note that we discard episodes that have less than 3 observations in the given channel. As the sampling density varies a lot across channels, the number of test cases in use to evaluate predictive performance for each channels can be considerably different. Therefore, we compute the average log likelihood on each channel and report the average over all 13 average log likelihoods as the evaluation metric.

For multi-output GPs, we consider three schemes to regularize the combination weight matrix $\mathbf{w} \in \mathbb{R}^{P\times Q}$. First, we apply ℓ_1 regularization on each entry of $\mathbf{w}$ by imposing the constraint $\|\mathbf{w}\|_1 < \tau$ (COGP-IND). We also consider regularizing $\mathbf{w}$ using group lasso by adding an extra term $\lambda \sum_G \sqrt{\sum_{(i,j)\in G} w_{ij}^2}$ to the negative evidence lower bound where G is a set of indices of $\mathbf{w}$ that forms a group, where $\lambda > 0$ controls the strength of the regularization. We consider taking each column as a group (COGP-COL) and taking each row as a group (COGP-ROW). In the experiment, we use a projected quasi-Newton algorithm [10] to optimize the regularized evidence lower bound. We also compare to COGP without regularization (COGP).

We test on different values of Q as well as parameters τ, λ for each regularization scheme. In the interest of space, we only show the best results of each method. Table 1 shows the best average held-out log-likelihood using different methods. We consider $Q \in \{3, 5, 8, 10\}$. The results show that smaller numbers of latent GPs results in better performance. Importantly, COGP significantly outperforms the independent baseline model. Table 1 also shows that regularization on columns gives the best result, although there is no column being zeroed out completely.

Table 2 shows the average log-likelihood of each channel. We can see that COGP outperforms INDEP-GP in all cases except for two of the sparser channels (dBP and Temp). With sparsity inducing regularization, COGP-COL is able to significantly improve the results for dBP while having a mild (positive or negative) effect on other channels.

4 Conclusion and Future Directions

In this work, we extend the collaborative multi-output GPs to learn correlations across different outputs based on a collection of multivariate sparse and irregularly-sampled time series. This is an important step toward follow-up machine learning tasks such as time series classification or clustering. Our work can be integrated with, for example, the expected Gaussian kernel [6] to perform various machine learning tasks while making use of the more accurate modeling provided by COGPs.

References

[1] Mauricio A Alvarez, Lorenzo Rosasco, and Neil D Lawrence. Kernels for vector-valued functions: A review. *arXiv preprint arXiv:1106.6251*, 2011.

[2] Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2007.

[3] Phillip Boyle and Marcus Frean. Dependent gaussian processes. In *Advances in neural information processing systems*, pages 217–224, 2004.

[4] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Conference on Uncertainty in Artificial Intellegence*, pages 282–290. auai.org, 2013.

[5] Matthew D Hoffman, David M Blei, Chong Wang, and John Paisley. Stochastic variational inference. *The Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

[6] Steven Cheng-Xian Li and Benjamin Marlin. Classification of sparse and irregularly sampled time series with mixtures of expected gaussian kernels and random features. In *Conference on Uncertainty in Artificial Intellegence*, 2015.

[7] Trung V Nguyen and Edwin V Bonilla. Collaborative multi-output gaussian processes. UAI, 2014.

[8] Joaquin Quiñonero-Candela and Carl Edward Rasmussen. A unifying view of sparse approximate gaussian process regression. *The Journal of Machine Learning Research*, 6:1939–1959, 2005.

[9] C.E. Rasmussen and C. Williams. *Gaussian processes for machine learning*. 2006.

[10] Mark W Schmidt, Ewout Berg, Michael P Friedlander, and Kevin P Murphy. Optimizing costly functions with simple constraints: A limited-memory projected quasi-newton algorithm. In *International Conference on Artificial Intelligence and Statistics*, page None, 2009.

[11] Yee-Whye Teh, Matthias Seeger, and Michael Jordan. Semiparametric latent factor models. In *Artificial Intelligence and Statistics 10*, number EPFL-CONF-161317, 2005.

[12] Andrew Wilson, Zoubin Ghahramani, and David A Knowles. Gaussian process regression networks. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, pages 599–606, 2012.