Regularized Variational Sparse Gaussian Processes

Shandian Zhe
School of Computing
University of Utah
zhe@cs.utah.edu

Abstract

Variational sparse Gaussian processes (GPs) are important GP approximate inference approaches. The key idea is to use a small set of pseudo inputs to construct a variational model evidence lower bound (ELBO). By maximizing the ELBO, we can optimize the pseudo inputs, as free variational parameters, jointly with the model parameters. The optimization, however, is highly nonlinear, nonconvex, and is easily trapped in inferior local maximums. We argue that the learning of these parameters, could be benefited from exploiting the training input information — we regularize the pseudo input estimation toward a statistical summarization of the training inputs in kernel space. To this end, we augment GPs by placing a kernelized mixture prior over the training inputs, where the mixtures components correspond to the pseudo inputs. We then derive a tight variational lower bound, which introduces an additional regularization term of the pseudo inputs and kernel parameters. We show the effectiveness of our regularized variational sparse approximation in two real regression datasets.

1 Introduction

Gaussian processes (GPs) [9] are powerful function estimators: due to their nonparametric nature, they are not restricted by any parametric functional form, and can flexibly infer various complex functions from data. However, GPs are notoriously costly for inference — it requires to compute a full covariance matrix over the training data and its inverse, which has $O(n^3)$ time and $O(n^2)$ space complexity (n is the number of training samples), making GPs infeasible for large applications.

To resolve this issue, many sparse approximate inference methods have been proposed [13, 10, 12, 11, 14, 7]. The basic strategy is to introduce a small collection of *pseudo* inputs— a compact representation of the entire training inputs. The pseudo inputs are used to break the dependencies between the training outputs, so as to prevent the calculation of the full, dense covariance matrix (and its inverse). To fulfill this goal, many methods impose simplified model assumptions. For example, FITC [13] assumes given the functional outputs of the psuedo inputs, all the training outputs are independent. Other examples include SoR [12], DTC [11], PITC [10], *etc.*

Recently [14] proposed a principled, variational sparse approximation framework — which uses a variational model evidence lower bound (ELBO) as the training objective. The pseudo inputs are treated as variational parameters and can be jointly estimated with model parameters, such as kernel parameters. The variational approximation avoids model simplification and over-estimation of the marginal likelihood, and often exhibits better quality, in recovering model parameters and posteriors [2]. The state-of-the-art large-scale GP inference approaches are all based on variational approximations, including stochastic training [4], training with MAPREDUCE [3], training with GPU on TENSORFLOW [6], and distributed asynchronous training on PARAMETERSERVER [7].

Despite the advantages of the variational methods, the estimation of the pseudo inputs — which determines the approximation quality — is by no means trivial [2]. Since the variational ELBOs

31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA.

are highly nonlinear and nonconvex, the optimization of the pseudo inputs (and kernel parameters) is easily trapped in inferior local maximums. In practice, a commonly used trick is applying k-means to obtain the training inputs' summarization, namely, the cluster centers, as the pseudo inputs' initialization [4, 5, 2, 7].

Enlightened by the often success of the k-means initialization, we attempt to further leverage the observed training inputs' information to guide/boost the learning of the pseudo inputs (and kernel parameters). We speculate that the ideal pseudo inputs should well summarize the training inputs' information in some latent (nonlinear) feature space. We tie this feature space to the one used in the GP covariance (or kernel) function. To this end, we augment the GP model by introducing a prior distribution over the inputs — we assume the inputs are sampled from a kernelized mixture distribution, and each mixture component corresponds to one pseudo input (after nonlinear feature mapping). We derive a tight variational ELBO for this augmented model, and obtain a data dependent regularization term over the pseudo inputs (and the kernel parameters), in addition to the original variational ELBO. Preliminary experiments on two real-world regression datasets show that the proposed regularized variational sparse approximation can improve the inference quality in terms of predictive performance. The new ELBO allows online and parallel inference as well.

2 Variational Sparse Gaussian Processes

First, let us briefly review GP models and their variational sparse approximations. In this paper, we focus on GP regression. The proposed methodology is available to other GP models as well. Given a set of d-dimensional inputs $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ and outputs $\mathbf{y} = \{y_1, \ldots, y_n\}$, we aim to infer the underlying function $f : R^d \to R$. To this end, we assume the collection of all the function values $\{f(\mathbf{x})|\forall \mathbf{x}\}$ is the sample path of a Gaussian process, and hence their finite projection on $\mathbf{X}$, namely $\mathbf{f} = [f(\mathbf{x}_1), \ldots, f(\mathbf{x}_n)]$, follows a multivariate Gaussian distribution,

$$p(\mathbf{f}|\mathbf{X}) = \mathcal{N}\big(\mathbf{f}|\mathbf{0}, \mathbf{K}_{nn}\big), \tag{1}$$

where $[\mathbf{K}_{nn}]_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, and $k(\cdot, \cdot)$ is the covariance function. We can choose any semi-definite kernel function as the covariance function. For example, in this work, we focus on the ARD kernel, $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top \mathrm{diag}(\frac{1}{\boldsymbol{\eta}})(\mathbf{x}_i - \mathbf{x}_j)) + \sigma_0$ where $\{\sigma, \sigma_0, \boldsymbol{\eta}\}$ are the kernel parameters. We further assume the observed outputs $\mathbf{y}$ are the function values $\mathbf{f}$ corrupted by some random noise with the variance β^{-1}: $p(\mathbf{y}|\mathbf{f}) = \mathcal{N}(\mathbf{y}|\mathbf{f}, \beta^{-1}\mathbf{I})$. Then we can marginalize out $\mathbf{f}$ to obtain the marginal probability of $\mathbf{y}$, i.e., the model evidence,

$$p(\mathbf{y}|\mathbf{X}) = \mathcal{N}\big(\mathbf{y}|\mathbf{0}, \mathbf{K}_{nn} + \beta^{-1}\mathbf{I}\big). \tag{2}$$

GP regression, from another perspective, can be treated as linear regression after mapping the original inputs into a latent, (possibly) infinite dimensional feature space [9]. The feature mapping is determined by the covariance (or kernel) function, and usually highly nonlinear.

The inference amounts to estimating the kernel parameters and the inverse noise variance β. This is usually done by maximizing the model evidence in (2). Given these parameters, we can calculate the posterior of the functional output for a new input $\mathbf{x}^*$, which is a conditional Gaussian distribution [9].

However, maximizing (2) requires to calculate the $n \times n$ covariance matrix $\mathbf{K}_{nn} + \beta^{-1}\mathbf{I}$ and its inverse, which has $\mathcal{O}(n^3)$ time and $\mathcal{O}(n^2)$ space complexity. When the training data size n is large, the computation is obviously infeasible. To solve this problem, a variational sparse GP inference framework is proposed in [14]. Specifically, we first introduce m inducing points, $\mathbf{B} = \{\mathbf{b}_1, \ldots, \mathbf{b}_m\}$, in the same d-dimensional input space. Then we consider the GP projection jointly over the training inputs $\mathbf{X}$ and the inducing points $\mathbf{B}$ — which corresponds to another multivariate Gaussian distribution. By using Jensen's inequality, we can obtain the following lower bound of the log of the model evidence (2),

$$\log\big(p(\mathbf{y}|\mathbf{X})\big) \geq L_0(\mathbf{B}) = \log\big(\mathcal{N}(\mathbf{y}|\mathbf{0}, \beta^{-1}\mathbf{I} + \mathbf{Q}_{nn})\big) - \frac{1}{2}\beta \mathrm{tr}(\tilde{\mathbf{K}}_{nn}) \tag{3}$$

where $\mathbf{Q}_{nn} = \mathbf{K}_{nm}\mathbf{K}_{mm}^{-1}\mathbf{K}_{mn}$, $\tilde{\mathbf{K}}_{nn} = \mathbf{K}_{nn} - \mathbf{Q}_{nn}$, $\mathbf{K}_{mm}$ is the covariance matrix over the pseudo inputs $\mathbf{B}$, i.e., $[\mathbf{K}_{mm}]_{ij} = k(\mathbf{b}_i, \mathbf{b}_j)$, and $\mathbf{K}_{mn}$ is the cross covariance between $\mathbf{X}$ and $\mathbf{B}$, i.e., $[\mathbf{K}_{mn}]_{ij} = k(\mathbf{x}_i, \mathbf{b}_j)$. This lower bound is a variational bound — when $m = n$ and $\mathbf{B} = \mathbf{X}$, the equality is achieved. However, to reduce the computational cost, we have to set $m \ll n$ — so that $\mathbf{Q}_{nn}$ is low rank, and the complexity for computing $L_0(\mathbf{B})$ is reduced to $O(nm^2)$.

3 Regularized Variational Sparse Gaussian Processes

The variational inference for GP regression is maximizing the ELBO, $L_0(\mathbf{B})$ in (3), to jointly optimize the pseudo inputs $\mathbf{B}$, the kernel parameters and the inverse noise variance β. However, this optimization is nontrivial because the ELBO is highly nonconvex and nonlinear. Treating the pseudo inputs $\mathbf{B}$ as free variational parameters can easily lead to unfavorable local maximums. In practice, people often use k-means to obtain the training inputs' summarization (i.e., cluster centers) as the pseudo inputs' initialization [4, 5, 2, 7], which empirically works well. This motives us to further leverage the training inputs' knowledge to guide or boost the estimation of the pseudo inputs (and the kernel parameters). Specifically, a potential problem of the k-means summarization is that its similarity measure (based on Euclidean distance) is usually very different from the one used in GPs (i.e., the nonlinear covariance functions). That means, the pseudo inputs obtained from the k-means algorithm, may not reflect the shape/distribution of the training inputs in the nonlinear feature space used by GPs, and hence could affect the approximation quality. Therefore, we argue that, rather than in the original feature space, the pseudo inputs should well summarize the training inputs in the latent feature space determined by the GP covariance function $k(\cdot,\cdot)$, with the feature mapping $\phi(\cdot)$. To bring in this assumption, we augment the GP regression model (2) with a mixture prior over each input $\mathbf{x}_i$ in the latent feature space,

$$p(z_i) = \text{Multnomial}(z_i|\frac{1}{m},\dots,\frac{1}{m}),$$
$$p(\mathbf{x}_i|z_i,\mathbf{B}) \propto \prod_{j=1}^{m}\big[\exp\big(-\frac{1}{2}\tau\cdot\|\phi(\mathbf{x}_i)-\phi(\mathbf{b}_j)\|^2\big)\big]^{\mathbb{1}(z_i=j)},$$

where z_i is the index of the component which $\mathbf{x}_i$ belongs to, $\mathbb{1}(\cdot)$ is the indicator function, and τ is the inverse variance. Note that these components are $\{\phi(\mathbf{b}_1),\dots,\phi(\mathbf{b}_m)\}$ and hence the pseudo inputs can be considered as the statistical summarization of $\mathbf{X}$ in the latent feature space. Although $\phi(\cdot)$ could be infinite dimensional, we can use the kernel trick to calculate the exponential term in each $\mathbf{p}(\mathbf{x}_i|z_i,\mathbf{B})$, which enables us to derive the variational ELBO for the augmented model:

$$\begin{aligned}\exp\big(-\frac{1}{2}\tau\cdot\|\phi(\mathbf{x}_i)-\phi(\mathbf{b}_j)\|^2\big) &= \exp\big(-\frac{1}{2}\tau\big(\phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_i)+\phi(\mathbf{x}_j)^\top\phi(\mathbf{x}_j)-2\phi(\mathbf{x}_i)^\top\phi(\mathbf{x}_j)\big)\big)\\ &= \exp\big(-\frac{1}{2}\tau\big(k(\mathbf{x}_i,\mathbf{x}_i)+k(\mathbf{x}_j,\mathbf{x}_j)-2k(\mathbf{x}_i,\mathbf{x}_j)\big)\big).\end{aligned}$$

The joint probability of the augmented model is given by

$$p(\mathbf{y},\mathbf{X},\mathbf{z}|\mathbf{B}) = p(\mathbf{z})p(\mathbf{X}|\mathbf{B},\mathbf{z})p(\mathbf{y}|\mathbf{X}) = \prod_{i=1}^{n}p(z_i)p(\mathbf{x}_i|z_i,\mathbf{B})\cdot\mathcal{N}\big(\mathbf{y}|\mathbf{0},k(\mathbf{X},\mathbf{X})+\beta^{-1}\mathbf{I}\big). \quad (4)$$

We then introduce a variational posterior $q(\mathbf{z}) = \prod_{i=1}^{n}q(z_i)$, and construct a variational lower bound of the log marginal probability,

$$\log p(\mathbf{y},\mathbf{X}) \geq L_1\big(\mathbf{B},q(\mathbf{z})\big) = \int\log\big(p(\mathbf{y},\mathbf{X},\mathbf{z})\big)q(\mathbf{z})\mathrm{d}\mathbf{z} - \int q(\mathbf{z})\log\big(q(\mathbf{z})\big)\mathrm{d}\mathbf{z}.$$

Assuming stationary kernels, we can easily derive the optimal variational posteriors: $q^*(z_i=j)\propto\exp(\tau k(\mathbf{x}_i,\mathbf{b}_j))(1\leq i\leq n, 1\leq j\leq m)$. Now we plug $q^*(\mathbf{z}) = \prod_{i=1}^{n}q^*(z_j)$ into L_1, combine with the ELBO in (3), and finally obtain the following tight ELBO for the augmented model (4) ,

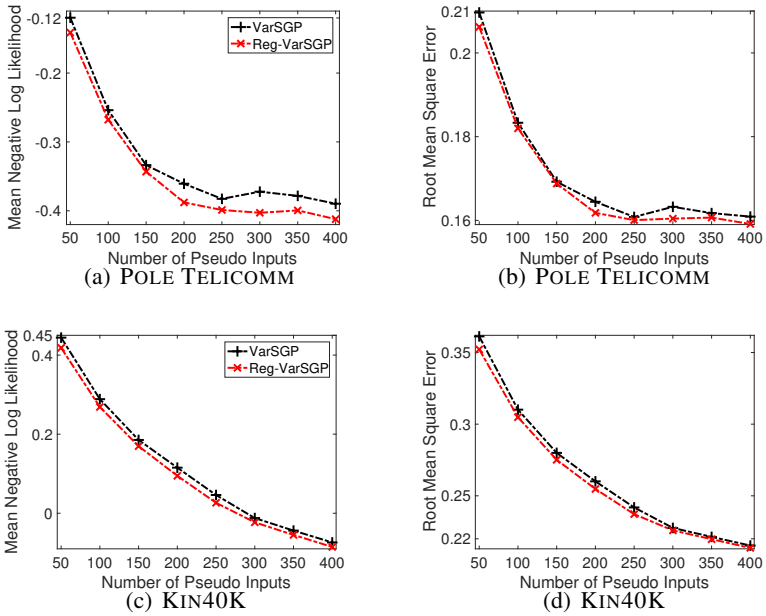
$$\log p(\mathbf{y},\mathbf{X}) \geq L_2(\mathbf{B}) = L_0(\mathbf{B}) + \tau\cdot L_r(\mathbf{B}) + \text{const} \quad (5)$$

where

$$L_r(\mathbf{B}) = \sum_{i=1}^{n}\big(-k(\mathbf{x}_i,\mathbf{x}_i) - \sum_{j=1}^{m}\theta_{ij}\big(\log(\theta_{ij}) - k(\mathbf{x}_i,\mathbf{b}_j)\big)\big), \quad (6)$$

$$\theta_{ij} = \frac{\exp\big(\tau k(\mathbf{x}_i,\mathbf{b}_j)\big)}{\sum_{t=1}^{m}\exp\big(\tau k(\mathbf{x}_i,\mathbf{b}_t)\big)}. \quad (7)$$

Our new variational approximation, $L_2(\mathbf{B})$, introduces an additional regularization term $L_r(\mathbf{B})$, which encourages the pseudo inputs to well summarize the training inputs in the kernel space. The inverse variance τ controls the regularization strength — when we set $\tau = 0$, we reduce to $L_0(\mathbf{B})$, the original variational sparse approximation. Moreover, the additive structure over each input in $L_r(\mathbf{B})$ enables an easy extension of the previous online or parallel GP inference algorithms based on $L_0(\mathbf{B})$.



(a) POLE TELICOMM (b) POLE TELICOMM

(c) KIN40K (d) KIN40K

Figure 1: Prediction accuracy *vs.* the number of pseudo inputs (a-b for POLE TELICOMM dataset, and c-d for KIN40K dataset).

4 Preliminary Results

We conducted a preliminary experiment to examine the effectiveness of the proposed regularized variational sparse GP approximation. We used two real datasets, POLE TELICOMM [8], and KIN40K [14] that have been widely used before. The POLE TELICOMM dataset consists of $10,000$ training and $5,000$ test samples, and the input dimension is 26. The KIN40K contains $10,000$ training and $30,000$ test samples, and the input dimension is 8. We compared with the original variational sparse GP approximation [14], which we denote by VarSGP. We denote our regularized variational sparse GP approximation by Reg-VarSGP. We varied the number of pseudo inputs from $\{50, 100, 150, 200, 250, 300, 350, 400\}$. We used the ARD kernel for all the evaluations, namely, $k(\mathbf{x}_i, \mathbf{x}_j) = \sigma \exp(-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^\top \mathrm{diag}(\frac{1}{\boldsymbol{\eta}})(\mathbf{x}_i - \mathbf{x}_j)) + \sigma_0$.

We used the same initialization for both VarSGP and Reg-VarSGP. For the pseudo inputs, we ran k-means++ [1] to obtain the cluster centers as their initial values; for the kernel parameters, we initialized $\sigma = 1$, $\sigma_0 = 1e-6$, and each element in $\boldsymbol{\eta}$ to be square of the median of the pair-wise distances between the initial pseudo inputs. For our method, Reg-VarSGP, we set the regularization strength $\tau = 1$ for POLE TELICOMM dataset and $\tau = 0.1$ for KIN40K dataset.

We used L-BFGS to maximize the variational ELBOs. We chose the MinFunc package implemented by Mark Schmidt (`https://www.cs.ubc.ca/~schmidtm/Software/minFunc.html`). The maximum number of iterations is set to 100.

We report the Mean Negative Log Likelihood (MNLL) and Root Mean Square Error (RMSE) of the test samples, as shown in Figure 1. The MNLL is calculated based on the posterior distributions of the test outputs, and RMSE is calculated based on their posterior means. As we can see, in all the cases, our regularized variational sparse approximation, Reg-VarSGP, exhibits improved predictive performance. Although the improvement is not big, we still see the GP inference is benefited from the proposed regularizer which encodes the training inputs' knowledge (see (5)).

In the future research, we plan to explore two directions. First, we plan to develop a stochastic variational inference algorithm based on our regularized sparse GP approximation. Then we can examine our method on much larger datasets, say, millions of training samples, and exploit much more inputs' information; we would like to see if our method's performance can thereby improve the original variational sparse GP by a large margin. Second, we plan to use the same framework as in Section 3, namely, augmenting GP models with a prior, $p(\mathbf{X}|\mathbf{B})$, to develop different regularizers and examine their performance.

References

[1] David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 1027–1035. Society for Industrial and Applied Mathematics, 2007.

[2] Matthias Bauer, Mark van der Wilk, and Carl Edward Rasmussen. Understanding probabilistic sparse Gaussian process approximations. In *Advances in Neural Information Processing Systems 29*, pages 1525–1533, 2016.

[3] Yarin Gal, Mark van der Wilk, and Carl Rasmussen. Distributed variational inference in sparse Gaussian process regression and latent variable models. In *Advances in Neural Information Processing Systems 27*, pages 3257–3265, 2014.

[4] James Hensman, Nicolo Fusi, and Neil D Lawrence. Gaussian processes for big data. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, 2013.

[5] James Hensman, Alexander G de G Matthews, and Zoubin Ghahramani. Scalable variational gaussian process classification. 2015.

[6] Alexander G de G Matthews, Mark van der Wilk, Tom Nickson, Keisuke Fujii, Alexis Boukouvalas, Pablo León-Villagrá, Zoubin Ghahramani, and James Hensman. GPflow: A Gaussian process library using TensorFlow. *Journal of Machine Learning Research*, 18(40):1–6, 2017.

[7] Hao Peng, Shandian Zhe, Xiao Zhang, and Yuan Qi. Asynchronous distributed variational Gaussian process for regression. In *International Conference on Machine Learning*, pages 2788–2797, 2017.

[8] Joaquin QuiÃ±onero-Candela, Carl Edward Rasmussen, AnÃbal R Figueiras-Vidal, et al. Sparse spectrum Gaussian process regression. *Journal of Machine Learning Research*, 11(Jun):1865–1881, 2010.

[9] Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning*. MIT press Cambridge, 2006.

[10] Anton Schwaighofer and Volker Tresp. Transductive and inductive methods for approximate Gaussian process regression. In *Advances in Neural Information Processing Systems 15*, pages 953–960. MIT Press, 2003.

[11] Matthias Seeger, Christopher Williams, and Neil Lawrence. Fast forward selection to speed up sparse Gaussian process regression. In *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*, 2003.

[12] Alexander J. Smola and Peter L. Bartlett. Sparse greedy Gaussian process regression. In *Advances in Neural Information Processing Systems 13*. MIT Press, 2001.

[13] Edward Snelson and Zoubin Ghahramani. Sparse Gaussian processes using pseudo-inputs. In *Advances in Neural Information Processing Systems*, pages 1257–1264, 2005.

[14] Michalis K Titsias. Variational learning of inducing variables in sparse Gaussian processes. In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, pages 567–574, 2009.