

VARIATIONAL INFERENCE FOR INFINITE MIXTURES OF SPARSE GAUSSIAN PROCESSES THROUGH KL-CORRECTION

T. N. A. Nguyen, A. Bouzerdoum, and S. L. Phung

School of Electrical, Computer and Telecommunications Engineering,
University of Wollongong, NSW, 2522, Australia

ABSTRACT

We propose a new approximation method for Gaussian process (GP) regression based on the mixture of experts structure and variational inference. Our model is essentially an infinite mixture model in which each component is composed of a Gaussian distribution over the input space, and a Gaussian process expert over the output space. Each expert is a sparse GP model augmented with its own set of inducing points. Variational inference is made feasible by assuming that the training outputs are independent given the inducing points. In previous works on variational mixture of GP experts, the inducing points are selected through a greedy selection algorithm, which is computationally expensive. In our method, both the inducing points and hyperparameters of the experts are learned through maximizing an improved lower bound of the marginal likelihood. Experiments on benchmark datasets show the advantages of the proposed method.

Index Terms— Gaussian process, variational inference.

1. INTRODUCTION

The Gaussian process (GP) models are powerful non-parametric tools for Bayesian regression [1–5] and classification [6–10]. The main limitation of GP is the high computational cost due to the need for the inversion and storage of the kernel matrix. This unfavorable scaling hinders its application to large datasets. Many sparse GP approximation methods have been proposed to overcome this limitation [11–16]. Common to these methods is the approximation of all training data via a small set of inducing points. These methods normally work well for simple datasets. However, in large and complex datasets, the dependencies between observations cannot be well-captured by a small set of inducing points. In addition, a single GP, supported by a small set of inducing points, cannot account for the non-stationarity and locality in such datasets as argued in [17].

A class of models called mixture of GPs [17–21] has been proposed to address the above issues. In these models, a gating network divides the input space into regions within which a specific GP expert is responsible for making predictions. In this way, mixtures of GPs can naturally address the non-stationarity and locality in the datasets. In addition, the computational complexity is also reduced as the storage and inversion of a large kernel matrix are replaced by those of multiple smaller matrices. Inference for these models involves the simultaneous learning of both the experts and the gating network. Due to the complexity of the inference problem, approximate techniques are often required. Most existing mixtures of GPs [17, 19] resort to the intensive Markov chain Monte Carlo (MCMC) sampling, which can be very slow. This prohibits their application to even moderate-sized problems [21].

Recently, several variational mixtures of GP experts [20, 22] have been proposed to use variational inference [23], which is a more flexible and faster alternative to MCMC sampling. In these

methods, each GP expert has its own set of inducing points, and is described by a linear model. This linear model was first proposed by Silverman in [24] for sparse GP approximation. In [25], it was proved to be equivalent to embedding a deterministic relation between latent function values at any arbitrary points and those at the inducing points. The linear model breaks the dependency among training outputs and makes variational inference feasible. However, these variational mixtures of GP experts suffer from two limitations. First, the approximate prior given by the linear model is degenerate with only M degrees of freedom [25], where M is the number of inducing points. This degeneracy leads to unreasonable predictive distributions with over-confident predictive variances. Second, although most hyper-parameters of the model can be estimated by maximizing the variational lower-bound of the marginal likelihood, identifying the inducing points in this way becomes computationally difficult. This is due to the tight coupling between the model parameters and the inducing points. As a result, the inducing points must be fixed through a separate greedy selection algorithm.

In this paper, we propose a new variational mixture of GP experts to overcome the above limitations. In our proposed method, each GP expert accompanied by its own set of inducing points is a sparse approximation to the full GP. Instead of imposing a deterministic relation between latent variables at training points and those at inducing points like in [20, 22], our model uses an approximate conditional distribution of the latent variables assuming that they are independent given the inducing points. In this way, variational inference in our model is made feasible without suffering from the degeneracy problem mentioned above. To tackle the second limitation, we propose to learn both the hyper-parameters and the inducing points through maximizing a corrected lower bound of the true marginal likelihood, which is inspired by the KL-corrected bound proposed in [26]. The introduction of this corrected bound provides two-fold benefits. First, it improves the convergence speed of variational approximation. Second, it removes the tight dependence between the inducing points and other parameters in the original variational bound, thereby, enabling the estimation of inducing points.

It is worth noting that a mixture of GP experts proposed in [21] is also based on the same assumptions of statistical independence of the latent variables given the inducing points. However, in this model, experts are trained on and responsible for disjoint subsets of the inputs. Discontinuities and possible inaccurate predictions may arise near expert boundaries due to the lack of extrapolation between different experts. Our model uses a different philosophy where a gating network probabilistically assigns points to the experts, and thus maintains a smooth transition between different experts.

2. MODEL SPECIFICATION

Consider a dataset \mathcal{D} consisting of input data, $\mathbf{X} = ((\mathbf{x}_1)^T, \dots, (\mathbf{x}_N)^T)^T$ with input points (row vectors) $\mathbf{x}_i \in \mathcal{X} \subset \mathbb{R}^D$, and outputs

$\mathbf{y} = (y_1, \dots, y_N)^T$. Our task is to compute the predictive distribution of the output y_t at test location \mathbf{x}_t . The proposed model, depicted in Fig. 1, is essentially a Dirichlet process infinite mixture model in which each component is composed of a Gaussian distribution over the input space, and a GP expert over the output space.

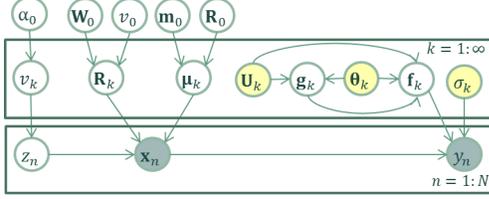


Fig. 1. Graphical representation of the proposed model.

2.1. Gaussian process expert and input space model

Each GP expert k is associated with a latent function $f_k(\mathbf{x}) \sim \mathcal{GP}(\mathbf{0}, \kappa_k(\mathbf{x}, \mathbf{x}'))$, where $\mathbf{x}, \mathbf{x}' \in \mathcal{X}$ and $\kappa_k(\mathbf{x}, \mathbf{x}')$ is the kernel function of the expert. The expert is also augmented with a set of M inducing points $\{\mathbf{u}_1^{(k)}, \dots, \mathbf{u}_M^{(k)}\} \subset \mathcal{X}$. They are collectively represented in matrix form by $\mathbf{U}_k = ((\mathbf{u}_1^{(k)})^T, \dots, (\mathbf{u}_M^{(k)})^T)^T$. We denote the latent function values at the training inputs and inducing points for expert k by $\mathbf{f}_k = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_N))^T$ and $\mathbf{g}_k = (f(\mathbf{u}_1^{(k)}), \dots, f(\mathbf{u}_M^{(k)}))^T$, respectively. For each observation (\mathbf{x}_n, y_n) , we have a corresponding latent variable z_n indicating to which expert it belongs. Following the standard Gaussian process regression, we assume the observed outputs are given by: $y_n = f_{z_n}(\mathbf{x}_n) + \epsilon_n$, where $\epsilon_n \sim \mathcal{N}(0, \sigma_{z_n}^2)$.

For each expert, the latent values are assumed to be independent given the inducing latent values \mathbf{g}_k , i.e.

$$p(\mathbf{f}_k | \mathbf{X}, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k) = \prod_{n=1}^N p(f_k(\mathbf{x}_n) | \mathbf{x}_n, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k), \quad (1)$$

where $\boldsymbol{\theta}_k$ is the hyperparameters for the kernel function $\kappa_k(\mathbf{x}, \mathbf{x}')$. Here, $p(f_k(\mathbf{x}_n) | \mathbf{x}_n, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k)$ is the predictive normal distribution of latent value $f_k(\mathbf{x}_n)$ given \mathbf{g}_k with mean $\bar{\mathbf{f}}_k(\mathbf{x}_n) = \mathbf{K}(\mathbf{x}_n, \mathbf{U}_k) \mathbf{K}(\mathbf{U}_k, \mathbf{U}_k)^{-1} \mathbf{g}_k$, and variance $\boldsymbol{\Lambda}_k(\mathbf{x}_n) = \kappa_k(\mathbf{x}_n, \mathbf{x}_n) - \mathbf{K}(\mathbf{x}_n, \mathbf{U}_k) \mathbf{K}(\mathbf{U}_k, \mathbf{U}_k)^{-1} \mathbf{K}(\mathbf{U}_k, \mathbf{x}_n)$:

$$p(f_k(\mathbf{x}_n) | \mathbf{x}_n, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k) = \mathcal{N}(f_k(\mathbf{x}_n) | \bar{\mathbf{f}}_k(\mathbf{x}_n), \boldsymbol{\Lambda}_k(\mathbf{x}_n)), \quad (2)$$

where $\mathbf{K}(\mathbf{A}, \mathbf{B})$ denotes the covariance matrix evaluated at all pairs of points in \mathbf{A} and \mathbf{B} . The covariance matrices involving \mathbf{U}_k are parameterized by the kernel hyperparameters $\boldsymbol{\theta}_k$ of expert k .

Notice that based on the independence assumption given in Eq. (1) and the predictive distribution in Eq. (2), each of our GP experts alone is a sparse GP model which corresponds to the fully independent training conditional (FITC) model described in [25].

The input space for each expert is modeled by a Gaussian distribution, i.e., $p(\mathbf{x} | z = k, \boldsymbol{\mu}_k, \mathbf{R}_k) = \mathcal{N}(\mathbf{x} | \boldsymbol{\mu}_k, \mathbf{R}_k^{-1})$, where \mathbf{R}_k is the precision matrix. The parameters $\boldsymbol{\mu}_k$ and \mathbf{R}_k are governed respectively by a Gaussian distribution prior and a Wishart distribution prior: $\boldsymbol{\mu}_k \sim \mathcal{N}(\boldsymbol{\mu}_0, \mathbf{R}_0^{-1})$ and $\mathbf{R}_k \sim \mathcal{W}(\mathbf{W}_0, v_0)$.

2.2. Dirichlet process mixture model

Traditional mixture models require the number of experts to be specified *a priori* for a particular dataset, which may be a difficult model selection problem. Therefore, following [17, 19, 22], we avoid this problem by using an infinite number of experts for our model. In particular, a Dirichlet process (DP) prior [27] is placed over the experts to allow the model to automatically determine the number of components. To facilitate variational inference, the *stick-breaking representation* for DP mixtures [28] is adopted as proposed in [29].

Let π_k be the mixing proportion of the mixture component k . The variable z_n can be regarded as following a multinomial distribution prior with parameters $\boldsymbol{\pi} = \{\pi_1, \dots, \pi_\infty\}$, i.e. $p(z_n = k) = \pi_k$. In stick-breaking construction, instead of sampling an infinite dimensional parameter $\boldsymbol{\pi}$ directly, the proportions π_k are given by

$$\pi_k = v_k \cdot \prod_{i=1}^{k-1} (1 - v_i), \quad (3)$$

where v_k 's are independent random variables following Beta $(1, \alpha_0)$.

2.3. Joint distribution of the model

Let $\mathbf{v}, \boldsymbol{\mu}, \mathbf{R}, \mathbf{f}, \mathbf{g}, \mathbf{U}, \boldsymbol{\theta}, \boldsymbol{\sigma}$ and \mathbf{z} be the set of all $v_k, \boldsymbol{\mu}_k, \mathbf{R}_k, \mathbf{f}_k, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k, \sigma_k$ and z_n , respectively. The hidden variables Ω in our model include the parameters $\mathbf{v}, \boldsymbol{\mu}$ and \mathbf{R} , and the latent variables \mathbf{f}, \mathbf{g} and \mathbf{z} . $\mathbf{U}, \boldsymbol{\theta}$ and $\boldsymbol{\sigma}$ constitute the hyperparameter set Θ of the GP experts that need to be learned, while $\alpha_0, \mathbf{W}_0, v_0, \mathbf{m}_0$ and \mathbf{R}_0 are the generic hyperparameters that can be easily fixed. The full joint distribution of the model $p(\mathcal{D}, \Omega | \Theta)$ is given by:

$$p(\mathcal{D}, \Omega | \Theta) = p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma}) p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta}) p(\mathbf{g} | \mathbf{U}, \boldsymbol{\theta}) p(\mathbf{X} | \mathbf{z}, \boldsymbol{\mu}, \mathbf{R}) p(\mathbf{z} | \mathbf{v}) p(\mathbf{v}) p(\boldsymbol{\mu}) p(\mathbf{R}), \quad (4)$$

where $p(\mathbf{v}) = \prod_{k=1}^{\infty} p(v_k)$, $p(\boldsymbol{\mu}) = \prod_{k=1}^{\infty} p(\boldsymbol{\mu}_k)$, $p(\mathbf{R}) = \prod_{k=1}^{\infty} p(\mathbf{R}_k)$, and other components are as follows:

$$p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma}) = \prod_{n=1}^N \prod_{k=1}^{\infty} \mathcal{N}(y_n | \mathbf{f}_k(\mathbf{x}_n), \sigma_k^2 \mathbf{I})^{[z_n == k]}, \quad (5)$$

$$p(\mathbf{g} | \mathbf{U}, \boldsymbol{\theta}) = \prod_{k=1}^{\infty} \mathcal{N}(\mathbf{g}_k | \mathbf{0}, \mathbf{K}(\mathbf{U}_k, \mathbf{U}_k)), \quad (6)$$

$$p(\mathbf{X} | \mathbf{z}, \boldsymbol{\mu}, \mathbf{R}) = \prod_{n=1}^N \prod_{k=1}^{\infty} \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_k, \mathbf{R}_k^{-1})^{[z_n == k]}, \quad (7)$$

$$p(\mathbf{z} | \mathbf{v}) = \prod_{n=1}^N p(z_n | \mathbf{v}) = \prod_{n=1}^N \prod_{k=1}^{\infty} (1 - v_k)^{[z_n > k]} v_k^{[z_n == k]}, \quad (8)$$

$$p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta}) = \prod_{k=1}^{\infty} p(\mathbf{f}_k | \mathbf{X}, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k), \quad (9)$$

with $p(\mathbf{f}_k | \mathbf{X}, \mathbf{g}_k, \mathbf{U}_k, \boldsymbol{\theta}_k)$ given in (1).

Eq. (5) follows from the standard noise model of GP regression as described above. Eq. (6) is due to the assumption of independence among the experts and applying GP prior at the inducing points for each expert, Eq. (7) follows from the Gaussian distribution over the input space of each expert, and Eq. (8) is the result of Eq. (3).

3. INFERENCE

The inference problem for our model involves estimating the posterior distribution of hidden variables, $p(\mathbf{f}, \mathbf{g}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{R} | \mathcal{D}, \Theta)$, and fixing the hyperparameters Θ . To this end, we use variational EM algorithm which iterates two steps. In the E-step, variational inference is used to optimize variational parameters of the posterior distribution for hidden variables while keeping Θ fixed. In the M-step, Θ is selected to maximize a lower-bound of the marginal likelihood.

3.1. Variational inference

In this section, we discuss the variational inference that constitutes the E-step of the algorithm. This is done by approximating the true posterior $p(\mathbf{f}, \mathbf{g}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{R} | \mathcal{D}, \Theta)$ with a tractable family of distributions $q(\mathbf{f}, \mathbf{g}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{R})$ that factorizes over $\mathbf{f}_k, \mathbf{g}_k, \mathbf{R}_k, \boldsymbol{\mu}_k, v_k$ and z_n . The infinite number of variables $\mathbf{R}_k, \boldsymbol{\mu}_k$ and v_k does not allow variational inference. To enable variational inference, we approximate the posterior Dirichlet process by a *truncated* stick-breaking representation as proposed in [29]. A value T is chosen such that $q(v_T = 1) = 1$. In this way, the mixture proportions π_k become zero for $k > T$; and the number of mixture components is limited to T . As discussed in [29], the truncation level T is not a part of the prior model specification, but a variational parameter which can be freely set without fear of over-fitting. We subsequently use the following factorized family of variational distributions $q(\mathbf{f}, \mathbf{g}, \mathbf{z}, \mathbf{v}, \boldsymbol{\mu}, \mathbf{R}) = \prod_{k=1}^T q(\mathbf{f}_k) q(\mathbf{g}_k) q(\mathbf{R}_k) q(\boldsymbol{\mu}_k) \prod_{k=1}^{T-1} q(v_k) \prod_{n=1}^N q(z_n)$. Since the variables $\mathbf{f}_k(\mathbf{x}_n)$ are independent given \mathbf{g}_k , it is reasonable to further assume: $q(\mathbf{f}_k) = \prod_{n=1}^N q(\mathbf{f}_k(\mathbf{x}_n))$. Let $\mathbb{E}_\omega(A)$ denote the expectation of A over variational distribution $q(\omega)$ of hidden variable ω , and $\mathbf{K}_u^{(k)}$ denote $\mathbf{K}(\mathbf{U}_k, \mathbf{U}_k)$. The optimal approximate distributions for the hidden variables are obtained as follows.

- a) $q(v_k) = \text{Beta}(\alpha_{k1} + 1, \alpha_{k2} + \alpha_0)$,
where $\alpha_{k1} = \sum_{n=1}^N q(z_n = k)$ and $\alpha_{k2} = \sum_{n=1}^N q(z_n > k)$.

- b) $q(\boldsymbol{\mu}_k) = \mathcal{N}(\bar{\boldsymbol{\mu}}_k, \mathbf{V}_k)$,
 where $\mathbf{V}_k = \left(\mathbf{R}_0 + \sum_{n=1}^N q(z_n = k) \mathbb{E}_{\mathbf{R}_k}(\mathbf{R}_k)\right)^{-1}$ and
 $\bar{\boldsymbol{\mu}}_k = \mathbf{V}_k(\mathbf{R}_0 \boldsymbol{\mu}_0 + \sum_{n=1}^N q(z_n = k) \mathbb{E}_{\mathbf{R}_k}(\mathbf{R}_k) \mathbf{x}_n)$.
- c) $q(\mathbf{R}_k) = \mathcal{W}(\tilde{\mathbf{W}}_k, \tilde{v}_k)$,
 where $(\tilde{\mathbf{W}}_k)^{-1} = (\mathbf{W}_0)^{-1} + \sum_{n=1}^N q(z_n = k)$ and
 $\tilde{v}_k = \sum_{n=1}^N q(z_n = k) \mathbb{E}_{\boldsymbol{\mu}_k}[(\mathbf{x}_n - \boldsymbol{\mu}_k)(\mathbf{x}_n - \boldsymbol{\mu}_k)^T]$.
- d) z_n follows a multinomial distribution: $q(z_n = k) = r_{nk}$, where
 $r_{nk} = \rho_{nk} / \sum_{i=1}^T \rho_{ni}$ is the *responsibility* of expert k for \mathbf{x}_n and
 $\ln \rho_{nk} = \text{const} + \mathbb{E}_{v_k}(\ln v_k) + \sum_{i=1}^{k-1} \mathbb{E}_{v_i}(\ln(1 - v_i))$
 $+ (1/2)[\mathbb{E}_{\mathbf{R}_k}(\ln |\mathbf{R}_k|) - \mathbb{E}_{\{\mathbf{R}_k, \boldsymbol{\mu}_k\}}((\mathbf{x}_n - \boldsymbol{\mu}_k)^T \mathbf{R}_k (\mathbf{x}_n - \boldsymbol{\mu}_k))]$
 $+ \ln \mathcal{N}(y_n | \mathbb{E}_{\mathbf{f}_k}(\mathbf{f}_k), \sigma_k^2)$. (10)
- e) $q(\mathbf{g}_k) = \mathcal{N}(\mathbf{K}_u^{(k)} \boldsymbol{\Upsilon}_k^{-1} \mathbf{K}(\mathbf{U}_k, \mathbf{X}) \boldsymbol{\Lambda}_k(\mathbf{X})^{-1} \mathbb{E}_{\mathbf{f}_k}(\mathbf{f}_k), \mathbf{K}_u^{(k)} \boldsymbol{\Upsilon}_k^{-1} \mathbf{K}_u^{(k)})$,
 where $\boldsymbol{\Upsilon}_k = \mathbf{K}_u^{(k)} + \mathbf{K}(\mathbf{U}_k, \mathbf{X}) \boldsymbol{\Lambda}_k(\mathbf{X})^{-1} \mathbf{K}(\mathbf{X}, \mathbf{U}_k)$, and (12)
 $\boldsymbol{\Lambda}_k(\mathbf{X}) = \text{diag}([\boldsymbol{\Lambda}_k(\mathbf{x}_1), \dots, \boldsymbol{\Lambda}_k(\mathbf{x}_N)])$. (13)
- f) $q(\mathbf{f}_k(\mathbf{x}_n)) = \mathcal{N}(\mathbf{m}_k(\mathbf{x}_n), \boldsymbol{\tau}_k(\mathbf{x}_n))$ (14)
 where $\boldsymbol{\tau}_k(\mathbf{x}_n) = \{(\boldsymbol{\Lambda}_k(\mathbf{x}_n))^{-1} + \sigma^{-2} r_{nk}\}^{-1}$, and
 $\mathbf{m}_k(\mathbf{x}_n) = \boldsymbol{\tau}_k(\mathbf{x}_n) \{\boldsymbol{\Lambda}_k(\mathbf{x}_n)^{-1} \mathbf{K}(\mathbf{x}_n, \mathbf{U}_k) (\mathbf{K}_u^{(k)})^{-1} \mathbb{E}(\mathbf{g}_k) + \sigma^{-2} r_{nk} y_n\}$.

3.2. Learning the hyperparameters

Given the variational distribution $q(\Omega)$ found in E-step, the hyperparameter set Θ can be learned by maximizing the following lower bound $\mathcal{L}_1(\Theta)$ of the log marginal likelihood $\ln p(\mathcal{D}|\Theta)$:

$$\begin{aligned} \mathcal{L}_1(\Theta) &= \mathbb{E}_{\Omega} \ln p(\mathcal{D}, \Omega | \Theta) \\ &= \mathbb{E}_{\{\mathbf{f}, \mathbf{z}\}} (\ln p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma})) + \mathbb{E}_{\{\mathbf{f}, \mathbf{g}\}} (\ln p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta})) \\ &\quad + \mathbb{E}_{\mathbf{g}} (\ln p(\mathbf{g} | \mathbf{U}, \boldsymbol{\theta})) + \text{const}. \end{aligned} \quad (15)$$

The above framework is effective to learn the hyperparameters $\boldsymbol{\theta}$ and $\boldsymbol{\sigma}$. However, due to the tight coupling between \mathbf{g} and the inducing points \mathbf{U} , learning \mathbf{U} by maximizing $\mathcal{L}_1(\Theta)$ while keeping $q(\mathbf{g})$ fixed is computationally difficult as it can be vulnerable to local optima. In this paper, we propose to learn both the hyperparameters and the inducing points through maximizing a KL-corrected bound.

The log marginal likelihood can be written as

$$\ln p(\mathcal{D}|\Theta) = \ln \int p(\mathcal{D} | \mathbf{g}, \Theta) p(\mathbf{g} | \Theta) d\mathbf{g}. \quad (16)$$

Let $\tilde{\Omega}$ be the set of all hidden variables, except \mathbf{g} . We derive the following lower bound for $\ln p(\mathcal{D} | \mathbf{g}, \Theta)$ in terms of $q(\tilde{\Omega})$:

$$\begin{aligned} \mathcal{L}(\Theta) &= \mathbb{E}_{\tilde{\Omega}} [\ln p(\mathcal{D}, \tilde{\Omega} | \mathbf{g}, \Theta)] + \text{const} \\ &= \mathbb{E}_{\mathbf{f}, \mathbf{z}} [\ln p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma})] + \mathbb{E}_{\mathbf{f}} [\ln p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta})] + \text{const}, \end{aligned} \quad (17)$$

Replacing $p(\mathcal{D} | \mathbf{g}, \Theta)$ in Eq. (16) with $\exp(\mathcal{L}(\Theta))$ leads to the KL-corrected bound of the log marginal likelihood:

$$\begin{aligned} \mathcal{L}_2(\Theta) &= \ln \int \exp(\mathbb{E}_{\mathbf{f}} [\ln p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta})]) p(\mathbf{g} | \Theta) d\mathbf{g} \\ &\quad + \mathbb{E}_{\mathbf{f}, \mathbf{z}} [\ln p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma})] + \text{const}, \end{aligned} \quad (18)$$

The first integral of $\mathcal{L}_2(\Theta)$ can be calculated as follows:

$$\begin{aligned} &\ln \int \exp(\mathbb{E}_{\mathbf{f}} [\ln p(\mathbf{f} | \mathbf{g}, \mathbf{X}, \mathbf{U}, \boldsymbol{\theta})]) p(\mathbf{g} | \Theta) d\mathbf{g} \\ &= \sum_{k=1}^T \left\{ \ln \int \exp(\mathbb{E}_{\mathbf{f}_k} [\ln p(\mathbf{f}_k | \mathbf{g}_k, \mathbf{X}, \mathbf{U}_k, \boldsymbol{\theta}_k)]) p(\mathbf{g}_k | \mathbf{U}_k, \boldsymbol{\theta}_k) d\mathbf{g}_k \right\}, \\ &= \sum_{k=1}^T \left\{ \ln \mathcal{N}(\mathbf{m}_k(\mathbf{X}) | 0, \mathbf{Q}_k + \boldsymbol{\Lambda}_k(\mathbf{X})) - \frac{1}{2} \text{Tr}(\boldsymbol{\Lambda}_k(\mathbf{X})^{-1} \boldsymbol{\tau}_k(\mathbf{X})) \right\}, \end{aligned} \quad (19)$$

where $\boldsymbol{\tau}_k(\mathbf{X}) = \text{diag}([\boldsymbol{\tau}_k(\mathbf{x}_1), \dots, \boldsymbol{\tau}_k(\mathbf{x}_N)])$,

$$\mathbf{Q}_k = \mathbf{K}(\mathbf{X}, \mathbf{U}_k) (\mathbf{K}_u^{(k)})^{-1} \mathbf{K}(\mathbf{U}_k, \mathbf{X}); \quad (20)$$

and the following expression has been used:

$$\begin{aligned} \mathbb{E}_{\mathbf{f}_k} [\ln p(\mathbf{f}_k | \mathbf{g}_k, \mathbf{X}, \mathbf{U}_k, \boldsymbol{\theta}_k)] &= \quad (21) \\ \ln \mathcal{N}(\mathbf{m}_k(\mathbf{X}) | \mathbf{K}(\mathbf{X}, \mathbf{U}_k) (\mathbf{K}_u^{(k)})^{-1} \mathbf{g}_k, \boldsymbol{\Lambda}_k(\mathbf{X})) - \frac{1}{2} \text{Tr}(\boldsymbol{\Lambda}_k(\mathbf{X})^{-1} \boldsymbol{\tau}_k(\mathbf{X})). \end{aligned}$$

The remaining term of the bound $\mathcal{L}_2(\Theta)$ is given by:

$$\begin{aligned} \mathbb{E}_{\mathbf{f}, \mathbf{z}} [\ln p(\mathbf{y} | \mathbf{f}, \mathbf{z}, \boldsymbol{\sigma})] &= \quad (22) \\ = -\frac{1}{2} \sum_{k=1}^T \left\{ \sum_{n=1}^N r_{nk} \left\{ \ln |\sigma_k^2| + \sigma_k^{-2} ((y_n - \mathbf{m}_k(\mathbf{x}_n))^2 + \boldsymbol{\tau}_k(\mathbf{x}_n)) \right\} \right\}. \end{aligned}$$

It has been proven in [26] that the new bound $\mathcal{L}_2(\Theta)$ is an upper bound of the traditional bound $\mathcal{L}_1(\Theta)$. In fact, if we subtract $\mathcal{L}_1(\Theta)$ from $\mathcal{L}_2(\Theta)$, the result takes on the form of a Kullback-Leibler divergence, hence the name KL-corrected bound. $\mathcal{L}_2(\Theta)$ is therefore guaranteed to converge and converges even faster than $\mathcal{L}_1(\Theta)$. In addition, it does not directly depend on $q(\mathbf{g})$. Hence, both the hyperparameters and the inducing points can be learned through maximizing this bound without fear of being trapped in local optima.

Once the model has been trained, the predictive distribution for an unseen data point x_* is estimated as

$$p(y_* | \mathbf{x}_*, \Omega, \Theta) = \sum_{k=1}^T p(z_* = k | \mathbf{x}_*, \Omega) p(y_* | \mathbf{x}_*, z_* = k, \Omega, \Theta),$$

where Ω denotes the variational means of the hidden variables.

3.3. Truncating the experts

In the current inference procedure, each expert is jointly responsible for every data point, i.e. basically the responsibility $r_{nk} > 0$, for most $n \in \{1, \dots, N\}$ and $k \in \{1, \dots, T\}$. As a result, the time complexity related to the inference of each expert is $O(NM^2)$, which mainly arises from the computation of $\boldsymbol{\Upsilon}_k$ in (12), and the inversion of the matrix $\mathbf{Q}_k + \boldsymbol{\Lambda}_k$ in the M-step. The memory complexity is $O(NM)$ due to the storage of the matrix $\mathbf{K}(\mathbf{U}_k, \mathbf{X})$.

However, it can be noticed that the values of most r_{nk} are small. When r_{nk} is small, the contribution of expert k in explaining data point \mathbf{x}_n is not valuable and may otherwise add noise to the inference. To save time and memory spent on those points, we propose to truncate the range of responsibility of each expert through thresholding r_{nk} . In particular, we set r_{nk} to 0 if $r_{nk} < \beta \max\{r_{n1}, \dots, r_{nT}\}$, for a chosen value of β . After that, inference related to each expert k involves only those points \mathbf{x}_n whose corresponding r_{nk} is greater than 0. Let \mathbf{X}_k , \mathbf{y}_k and \mathbf{f}_k be the sets of inputs, outputs and latent variables at such points, respectively. Subsequently, \mathbf{X} , \mathbf{y} and \mathbf{f}_k in Eqs. (11), (12), (13), (19), (20), (21) and (22) are replaced with \mathbf{X}_k , \mathbf{y}_k and \mathbf{f}_k , respectively. Since the intersection among the sets \mathbf{X}_k becomes small, the average size of \mathbf{X}_k is approximately N/T . The time complexity related to the inference of each expert is reduced to $O(NM^2/T)$ in M-step, but remains $O(NM^2)$ in the E-step due to the calculation of r_{nk} in Eq. (10). The overall complexity for the inference is $O(TNM^2)$ in time and $O(NM/T)$ in memory. Compared to the traditional sparse GP with the same number of inducing points, i.e., $B = M \times T$, our models can possibly run T times faster and use T^2 times less memory.

4. EXPERIMENTS AND RESULTS

We present two sets of experiments with datasets of varying size to analyze different aspects of our models. The parameter settings, and the experimental setup and environment are given below.

Generic hyperparameter setting: As mentioned previously, our proposed model has a few generic hyperparameters $\{\alpha_0, \mathbf{W}_0, v_0, \mathbf{m}_0, \mathbf{R}_0\}$ that are set as follows. \mathbf{m}_0 , \mathbf{R}_0 and v_0 are set to the mean μ_x , the inverse covariance \mathbf{R}_x and the dimensionality d of the training input, respectively. \mathbf{W}_0 is set to \mathbf{R}_x/d , and α_0 is set to 10^{-2} .

Experimental setup and environment: We use the squared exponential (SE) kernel with automatic relevance determination (ARD) in all experiments. Hyperparameters are optimized using the

Table 1. Test results of the proposed method, FITC, local-FITC and FGP on the four benchmark data sets. Results are the averages over 5 trials, along with the standard deviation. The best performances are given in bold.

		<i>kin40k</i>	<i>pumadyn32nm</i>	<i>chem</i>	<i>synth8</i>	<i>sarcos</i>
SMSE	Proposed method	0.0406 (± 0.0020)	0.0457 (± 0.0006)	0.1034 (± 0.0016)	0.2507 (± 0.0035)	0.0153 (± 0.0003)
	FITC	0.0540 (± 0.0013)	0.0462 (± 0.0015)	0.1398 (± 0.0091)	0.2468 (± 0.0165)	0.0183 (± 0.0012)
	Local-FITC	0.0612 (± 0.0032)	0.0505 (± 0.0045)	0.1190 (± 0.0038)	0.2676 (± 0.0221)	0.0165 (± 0.0018)
	FGP	0.0643 (± 0.0051)	0.0568 (± 0.063)	0.1117 (± 0.0072)	0.2649 (± 0.0151)	0.0162 (± 0.0023)
MSLL	Proposed method	-2.0828 (± 0.0024)	-1.5363 (± 0.0256)	-1.9880 (± 0.0542)	-0.8689 (± 0.0264)	-2.4475 (± 0.0005)
	FITC	-2.0782 (± 0.0224)	-1.5353 (± 0.0232)	-2.1630 (± 0.0442)	-0.9049 (± 0.0383)	-2.1852 (± 0.0070)
	Local-FITC	-1.9439 (± 0.0128)	-1.4728 (± 0.0194)	-2.0134 (± 0.0448)	-0.8400 (± 0.0134)	-2.3964 (± 0.0065)
	FGP	-2.0163 (± 0.0124)	-1.4916 (± 0.0209)	-1.7348 (± 0.0352)	-0.8095 (± 0.0214)	-2.3674 (± 0.0026)
Training time (s)	Proposed method	1369 (± 33)	2380 (± 69)	6054 (± 50)	4497 (± 30)	9173 (± 631)
	FITC	5175 (± 52)	6994 (± 86)	16985 (± 312)	14528 (± 321)	51030 (± 643)
	Local-FITC	1354 (± 31)	2651 (± 57)	7438 (± 73)	3914 (± 48)	9367 (± 454)
	FGP	1012 (± 21)	2116 (± 43)	5355 (± 231)	3166 (± 127)	7726 (± 527)

conjugate gradients code in the GPML package with a maximum of 500 iterations for experiments in Section 4.1 and 1000 iterations for those in Section 4.2. The input and target in each data set are normalized to have zero mean. Each experiment is carried out on an Intel(R) Core(TM) i7-4770 3.47GHz CPU with 8GB of RAM using MATLAB R2014a.

4.1. Experiments on medium-size datasets

First we evaluate the performance of our model on a wide range of medium-size benchmark datasets including: kin40k (8 dimensions, 10000 training, 30000 test), pumadyn32nm (32 dimensions, 7168 training, 1024 test)¹, chem (15 dimensions, 31535 training, 31536 test), synth8 (8 dimensions, 30543 training, 30544 test), and sarcos (21 dimensions, 44484 training, 4449 test)². We use the same training /test splits as in [30], [21] and [31].

We compare our method with three different baselines. The first one is the FITC approximation [25]. The second baseline is local-FITC in which the training data is divided into T clusters and a FITC model is separately trained on each cluster. According to the experiment results presented in [21], local-FITC typically achieves its best performance when random clustering is used to partition the dataset as compared to using k-means or recursive projection clustering [31]. Therefore, we use local-FITC with random clustering in our experiments. The third baseline is the mixture of GP experts (FGP) proposed in [21], which uses MAP assignment for expert allocation. We set $T = 2$ experts for our method, local-FITC and FGP. To compare the performance of these baselines based on the same time complexity, we choose the number of inducing points for them as follows. Both our method and FGP use $M = 500$ inducing points per expert, i.e. 1000 inducing points in total. Local-FITC uses 700 inducing points per expert while FITC uses a total of 700 inducing points. We run each method 5 times. Each run is started with different random seeds. The results are then averaged over 5 runs to obtain the predictive performance and training time shown in Table 1. The performance is evaluated in terms of the standardized mean squared error (SMSE) and mean standardized log loss (MSLL).

First, we observe that our method, local-FITC and FGP take about the same amount of time for training given the settings described above, while FITC requires much longer time. Second, our method has lower SMSE and MSLL than the other methods in most of the datasets including kin40k, pumadyn32nm, chem and sarcos. For chem dataset, FITC produces a lower MSLL than our method does, i.e. it is more confident about its prediction. However, this confidence cannot justify its very poor accuracy (high SMSE). Our method loses to FITC in synth8 dataset but the difference in their performances is small.

¹ Available from <http://www.cs.toronto.edu/~delve/data/datasets.html>

² Available from http://homepages.inf.ed.ac.uk/ckiw/code/gpr_approx.html

4.2. Experiments on big dataset

Next we evaluate our model on the Million Song dataset [32]. We use the exact split as in [21], which comprises 10^5 songs for training and 51630 songs for testing. Each song has 90 acoustic features based on which its year of release is to be predicted. As argued in [21], when dealing with such a large dataset on a single machine, the choice regarding the sparsity and computational complexity of the prediction models should be directed by the physical memory limit of the benchmark computer. Given the memory complexity of $O(NM/T)$ of our model, we set $T = 20$ experts and $M = 300$ inducing points per expert. The performance of our method together with those of four other methods are shown in Table 2 for comparison. Results are the averages over 5 runs, along with the standard deviation. The performance is evaluated in term of SMSE and the negative log predictive density (NLPD).

In Table 2, GPSVI2000 denotes the FITC with stochastic variational inference [33], which uses $B = 2000$ inducing points. SOD2000 is the standard GP regression model in which a subset of $B = 2000$ data points is randomly sampled for training. The choice of $B = 2000$ for GPSVI and SOD is due to their memory complexity of $O(B^2)$. FGP and local-FITC are assigned the same number of inducing points and experts as our method. The performance for FGP, GPSVI2000, Local-FITC and SOD2000 were directly quoted from [21]. It can be clearly seen that our method gives the best performance among the compared methods.

Table 2. Performance of the models on the Million Song dataset. The best performances are given in bold.

METHOD	SMSE	NLPD
Proposed method	0.706 (± 0.003)	3.55 (± 0.03)
FGP	0.715 (± 0.003)	3.59 (± 0.01)
GPSVI2000	0.724 (± 0.003)	3.64 (± 0.01)
Local-FITC	0.761 (± 0.009)	3.63 (± 0.03)
SOD2000	0.794 (± 0.011)	3.69 (± 0.01)

5. CONCLUSION

We have presented a new mixture of GP experts model based on sparse GP approximation and variational inference. The proposed model nicely addresses the two issues in the existing variational mixtures of GP experts, namely the degeneracy and inducing point learning. First, our model uses a FITC-like sparse approximation for each expert, therefore, it does not suffer from the degeneracy problem that happens to the models that use linear approximation. Second, the KL-corrected bound used in our model leads to faster variational learning of both the inducing points and hyperparameters of the GP experts. Experiments on benchmark datasets showed the enhanced predictive power of our model compared to the traditional sparse GP models and many existing mixtures of GP experts.

6. REFERENCES

- [1] A. O’Hagan and J. Kingman, “Curve fitting and optimal design for prediction,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 40, no. 1, pp. 1–42, 1978.
- [2] P. Boyle and M. Frean, “Dependent Gaussian processes,” *Advances in Neural Information Processing Systems*, vol. 17, pp. 217–224, 2005.
- [3] P. W. Goldberg, C. K. Williams, and C. M. Bishop, “Regression with input-dependent noise: A Gaussian process treatment,” *Advances in Neural Information Processing Systems*, vol. 10, pp. 493–499, 1998.
- [4] R. B. Gramacy, H. K. Lee, and W. G. Macready, “Parameter space exploration with Gaussian process trees,” *Proc. 21st Intern. Conf. on Machine Learning*, pp. 353–360, Banff, Canada, Jul. 4–8, 2004.
- [5] P. Sollich and C. Williams, “Using the equivalent kernel to understand Gaussian process regression,” *Advances in Neural Information Processing Systems*, vol. 17, pp. 1313–1320, 2005.
- [6] M. Kuss and C. E. Rasmussen, “Assessing approximate inference for binary Gaussian process classification,” *Journal of Machine Learning Research*, vol. 6, pp. 1679–1704, 2005.
- [7] Y. Altun, T. Hofmann, and A. J. Smola, “Gaussian process classification for segmenting and annotating sequences,” *Proc. 21st Intern. Conf. on Machine Learning*, pp. 4–11, Banff, Canada, Jul. 4–8, 2004.
- [8] A. Kapoor, K. Grauman, R. Urtasun, and T. Darrell, “Active learning with Gaussian processes for object categorization,” *Proc. 11th Intern. Conf. on Computer Vision*, pp. 1–8, Rio de Janeiro, Brazil, Oct. 14–20, 2007.
- [9] H. Nickisch and C. E. Rasmussen, “Approximations for binary Gaussian process classification,” *Journal of Machine Learning Research*, vol. 9, no. 10, pp. 2035–2078, 2008.
- [10] T. P. Centeno and N. D. Lawrence, “Optimising kernel parameters and regularisation coefficients for non-linear discriminant analysis,” *Journal of Machine Learning Research*, vol. 7, pp. 455–491, 2006.
- [11] M. Seeger, C. Williams, and N. Lawrence, “Fast forward selection to speed up sparse Gaussian process regression,” *Proc. 9th Intern. Workshop on Artificial Intelligence and Statistics*, pp. 3–6, Key West, FL, USA, Jan. 3–6, 2003.
- [12] N. Lawrence, M. Seeger, and R. Herbrich, “Fast sparse Gaussian process methods: The informative vector machine,” *Advances in Neural Information Processing Systems*, vol. 15, pp. 625–632, 2003.
- [13] A. J. Smola and P. L. Bartlett, “Sparse greedy Gaussian process regression,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 619–625, 2001.
- [14] M. Seeger, “Bayesian Gaussian process models: PAC-Bayesian generalisation error bounds and sparse approximations,” Ph.D. dissertation, University of Edinburgh, 2003.
- [15] L. Csató, “Gaussian processes: iterative sparse approximations,” Ph.D. dissertation, Aston University, 2002.
- [16] V. Tresp, “A Bayesian committee machine,” *Neural Computation*, vol. 12, no. 11, pp. 2719–2741, 2000.
- [17] C. E. Rasmussen and Z. Ghahramani, “Infinite mixtures of Gaussian process experts,” *Advances in Neural Information Processing Systems*, vol. 14, pp. 881–888, 2002.
- [18] V. Tresp, “Mixtures of Gaussian processes,” *Advances in Neural Information Processing Systems*, vol. 13, pp. 654–660, 2000.
- [19] E. Meeds and S. Osindero, “An alternative infinite mixture of Gaussian process experts,” *Advances in Neural Information Processing Systems*, vol. 18, pp. 883–890, 2006.
- [20] C. Yuan and C. Neubauer, “Variational mixture of Gaussian process experts,” *Advances in Neural Information Processing Systems*, vol. 21, pp. 1897–1904, 2009.
- [21] T. Nguyen and E. Bonilla, “Fast allocation of Gaussian process experts,” *Proc. 31st Intern. Conf. on Machine Learning*, pp. 145–153, Beijing, China, Jun. 21–26, 2014.
- [22] S. Sun and X. Xu, “Variational inference for infinite mixtures of Gaussian processes with applications to traffic flow prediction,” *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 2, pp. 466–475, 2011.
- [23] C. M. Bishop, *Pattern Recognition and Machine Learning*. Verlag, New York: Springer, 2006.
- [24] B. W. Silverman, “Some aspects of the spline smoothing approach to non-parametric regression curve fitting,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 47, no. 1, pp. 1–52, 1985.
- [25] J. Quiñonero-Candela and C. E. Rasmussen, “A unifying view of sparse approximate Gaussian process regression,” *Journal of Machine Learning Research*, vol. 6, pp. 1939–1959, 2005.
- [26] N. J. King and N. D. Lawrence, “Fast variational inference for Gaussian process models through KL-correction,” *Proc. 17th European Conference on Machine Learning*, pp. 270–281, Berlin, Germany, Sep. 18–22, 2006.
- [27] T. S. Ferguson, “A Bayesian analysis of some nonparametric problems,” *The Annals of Statistics*, vol. 1, no. 2, pp. 209–230, 1973.
- [28] J. Sethuraman, “A constructive definition of Dirichlet priors,” *Statistica Sinica*, vol. 4, pp. 639–650, 1994.
- [29] D. M. Blei and M. I. Jordan, “Variational inference for Dirichlet process mixtures,” *Bayesian analysis*, vol. 1, no. 1, pp. 121–143, 2006.
- [30] C. K. Williams and C. E. Rasmussen, *Gaussian Processes for Machine Learning*. Cambridge, Massachusetts: the MIT Press, 2006.
- [31] K. Chalupka, C. K. Williams, and I. Murray, “A framework for evaluating approximation methods for Gaussian process regression,” *Journal of Machine Learning Research*, vol. 14, no. 1, pp. 333–350, 2013.
- [32] T. Bertin-Mahieux, D. P. Ellis, B. Whitman, and P. Lamere, “The million song dataset,” *Proc. 12th Intern. Conf. on Music Information Retrieval*, pp. 591–596, Miami, FL, USA, Oct. 24–28, 2011.
- [33] J. Hensman, N. Fusi, and N. D. Lawrence, “Gaussian processes for big data,” *Proc. 29th Conf. on Uncertainty in Artificial Intelligence*, pp. 282–290, Bellevue, WA, USA, Jul. 11–15, 2013.