

APPROXIMATE GAUSSIAN PROCESS REGRESSION AND PERFORMANCE ANALYSIS USING COMPOSITE LIKELIHOOD

*Xiuming Liu** *Dave Zachariah** *Edith C.H. Ngai*†*

* Uppsala University, Uppsala, Sweden

† The University of Hong Kong, Hong Kong, China

ABSTRACT

Nonparametric regression using Gaussian Process (GP) models is a powerful but computationally demanding method. While various approximation methods have been developed to mitigate its computation complexity, few works have addressed the quality of the resulting approximations of the target posterior. In this paper we start from a general belief updating framework that can generate various approximations. We show that applying using composite likelihoods yields computationally scalable approximations for both GP learning and prediction. We then analyze the quality of the approximation in terms of averaged prediction errors as well as Kullback–Leibler (KL) divergences.

1. INTRODUCTION

Consider a general regression setting in which the output is modeled as

$$y = f(\mathbf{x}) + \varepsilon, \quad (1)$$

where $\mathbf{x}$ is an inputs and output, $f(\cdot)$ is a unknown function, and ε is the random measurement error. The goal is to model the unknown function $f(\cdot)$ and predict latent function values, based on given input-output pairs. Gaussian Process (GP) models [1] are commonly used to capture variations of the nonlinear function in a nonparametric manner. GP-based Bayesian inference provides a conceptually straight-forward means of obtaining predictions along with credibility intervals, and are therefore widely applied in machine learning, time series analysis, spatial statistics, and control systems.

However, both GP learning and inference rely on evaluating the likelihood functions with available data. In case of a large set of data, this evaluation becomes computational prohibitive due to the requirement of inverting covariance matrices, which takes complexity of $\mathcal{O}(N^3)$, where N is the size of training data. Thus GP learning and inference have been considered as a non-scalable method and only applied to applications with data sets of mild sizes.

Many research have been conducted to make GP-based Bayesian inference scalable. Examples are: low rank approximation of covariance matrices [2], sparse GPs [3], and product of experts (or Bayesian committee machines) [4]. Many methods rely on fast heuristic searches [5] for an optimal subset of data or dimensions to maximize the information gain, which has been proofed as a NP-complete problem and therefore infeasible to apply for large datasets. In this paper, we argue that approximate methods can be unified from belief updating perspective by using different loss functions. Among them, the composite likelihood method provides scalable computation and tractable impact on the posterior. We present results of performance analysis for the proposed scalable GP-based Bayesian inference, in terms of both data averaged mean square errors (MSE) and Kullback–Leibler (KL) divergence.

1.1. Problem Formulation

We consider $f(\mathbf{x})$ in (1) to be a latent variable modeled as a Gaussian process $\mathcal{GP}(\mu(\mathbf{x}), \sigma(\mathbf{x}, \mathbf{x}'))$ [1]. This model yields a prior belief distribution $p(\mathbf{z})$ over the latent variable at M test points:

$$\mathbf{z} = [f(\mathbf{x}_1^*), \dots, f(\mathbf{x}_M^*)]^\top \sim \mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\Sigma}_z). \quad (2)$$

The joint mean $\boldsymbol{\mu}_z$ and covariance matrix $\boldsymbol{\Sigma}_z$ are functions of the test points $\{\mathbf{x}_1^*, \dots, \mathbf{x}_M^*\}$. Using a training data set $\mathcal{D} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_N, y_N)\} = \{\mathbf{X}, \mathbf{y}\}$, our goal is to update the belief distribution over $\mathbf{z}$ so as to produce a prediction $\hat{\mathbf{z}}$ with a dispersion for uncertainty. In addition to the model (2), we model ε_i in each sample from (1) as $\varepsilon_i \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ (i.i.d.). This specifies a conditional data model $p(\mathbf{y}|\mathbf{z})$, that is Gaussian with mean $\boldsymbol{\mu}_{y|\mathbf{z}}$ and covariance $\boldsymbol{\Sigma}_{y|\mathbf{z}}$, from which the standard Bayesian inference framework updates the belief distribution into the posterior $p(\mathbf{z}|\mathbf{y})$ via Bayes' rule. This update requires a matrix inversion and is in general computed in $\mathcal{O}(N^3)$ [3].

The inference of $\mathbf{z}$ depends, however, on a specified mean and covariance model, which can be learned in several ways. When it is parameterized by a vector $\boldsymbol{\theta}$, the maximum

This research work was supported by The Swedish Research Council through projects under Contracts 621-2016-06079, 2017-04543, and 2018-05040.

marginal likelihood approach is a popular learning approach that aims to solve the problem

$$\hat{\theta} = \arg \max_{\theta} \int p_{\theta}(\mathbf{y}|\mathbf{z}) p_{\theta}(\mathbf{z}) d\mathbf{z}, \quad (3)$$

which requires several matrix inversions in itself. Finally, the posterior

$$p_{\theta}(\mathbf{z}|\mathbf{y}) = \frac{p_{\theta}(\mathbf{y}|\mathbf{z}) p_{\theta}(\mathbf{z})}{p_{\theta}(\mathbf{y})} \quad (4)$$

is evaluated at $\theta = \hat{\theta}$. Both (3) and (4) require a runtime on the order $\mathcal{O}(N^3)$ and a storage that scales as $\mathcal{O}(N^2)$, which renders standard GP training and inference infeasible for large N . In this work, our goal is to firstly formulate an alternative update of the belief distribution, secondly provide a scalable training and inference method and finally perform an information-theoretic analysis on the performance of approximation in terms of information loss and mean square error.

1.2. Updating Belief Distributions

The posterior can be thought of as a special case of updating the belief distribution $p(\mathbf{z})$ into a new distribution $q(\mathbf{z})$ using the data $\mathcal{D}$. A more general belief updating framework was formulated in cf. [6]. In this framework, we first define a loss function $\ell(\mathbf{y}; \mathbf{z})$ and then find the distribution $q(\mathbf{z})$ which minimizes the average loss

$$L(q(\mathbf{z})) \triangleq \int_{\mathcal{Z}} \ell(\mathbf{y}; \mathbf{z}) q(\mathbf{z}) d\mathbf{z} + D(q(\mathbf{z})||p(\mathbf{z})), \quad (5)$$

regularized by the Kullback–Leibler divergence $D(q(\mathbf{z})||p(\mathbf{z}))$ [7]. Thus the first term fits $q(\mathbf{z})$ to the data, while the second term constrains it to the prior belief distribution. The updated belief distribution is then obtained as

$$\hat{q}(\mathbf{z}) = \arg \min_{q(\mathbf{z})} L(q(\mathbf{z})). \quad (6)$$

It is readily seen that for the loss function $\ell(\mathbf{y}; \mathbf{z}) = -\ln p(\mathbf{y}|\mathbf{z})$, the minimizer of $L(q(\mathbf{z}))$ is the posterior $p(\mathbf{z}|\mathbf{y})$, cf. [6] for more details. It is also interesting to point out that the above optimal belief updating framework has the same structure as the KL divergence optimization problem in variational Bayesian inferences [8]. However, we use the similar method from a different angle: in [8], the challenge is to tackle the complicate joint distribution of high-dimensional latent variables by factorizing it into single variable factors; in this paper, we focus on the problem of high computational complexity caused by evaluation the likelihood of large data.

To circumvent the prohibitive computational requirements of this approach, we formulate a different loss function based on marginal blocks of the full data distribution $p(\mathbf{y}|\mathbf{z})$ similar to the composite likelihood approach, cf. [9]. Specifically, we choose

$$c\ell(\mathbf{y}; \mathbf{z}) = -\sum_{k=1}^K \ln p(\mathbf{y}_k|\mathbf{z}), \quad (7)$$

where the data set $\mathcal{D}$ has been divided into K segments $\mathcal{D}_k = \{\mathbf{X}_k, \mathbf{y}_k\}$, $k = 1, \dots, K$, with $N_k \ll N$ samples each. The composite likelihoods enables scalable and online processing for large data set.

Theorem 1 (Bayesian Inference with Composite Likelihoods). *By applying (7), we obtain the updated belief distribution*

$$\begin{aligned} p(\mathbf{z}|\mathbf{y}_{1:K}) &\triangleq \arg \min_{q(\mathbf{z})} \int_{\mathcal{Z}} c\ell(\mathbf{y}; \mathbf{z}) q(\mathbf{z}) d\mathbf{z} + D(q(\mathbf{z})||p(\mathbf{z})) \\ &= \frac{p(\mathbf{z}) \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{z})}{p(\mathbf{y}_{1:K})}, \end{aligned} \quad (8)$$

or in the recursive formation

$$p(\mathbf{z}|\mathbf{y}_{1:K}) = \frac{p(\mathbf{z}|\mathbf{y}_{1:K-1}) p(\mathbf{y}_K|\mathbf{z})}{p(\mathbf{y}_K)}, \quad (9)$$

where $p(\mathbf{y}_{1:K}) = \int_{\mathcal{Z}} p(\mathbf{z}) \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{z}) d\mathbf{z}$ is the marginalized distribution for all data; and $p(\mathbf{y}_K) = \int_{\mathcal{Z}} p(\mathbf{z}|\mathbf{y}_{1:K-1}) p(\mathbf{y}_K|\mathbf{z}) d\mathbf{z}$ is the marginalized distribution for segment K . For a fixed N_k , (8) can be evaluated in a runtime that is of order $\mathcal{O}(KN_k^3)$.

The proof follows from the recognition that (5) is equivalent to the divergence $D(q(\mathbf{z}) || \exp(-c\ell(\mathbf{y}; \mathbf{z})) p(\mathbf{z})) \geq 0$, which attains the minimum 0 only when $q(\mathbf{z}) \propto \exp(-c\ell(\mathbf{y}; \mathbf{z})) p(\mathbf{z})$. Then the result is obtained by noting that $q(\mathbf{z})$ is a distribution that integrates and therefore integrates to unity. The recursive computation of the posterior with composite likelihood in (8) can be tackled using methods discussed in Section 2 below.

2. RECURSIVE COMPUTATION

In this section, we present the recursive computation solutions for GP learning and prediction.

2.1. Learning

The Fisher information matrix (FIM), $\mathbf{J}(\theta|\mathbf{y}_k)$, is a measure of information regarding to the unknown parameters θ contained in data $\mathbf{y}_k$. In case of Gaussian distributed data, the FIM can be obtained by the Slepian-Bangs formula. We refer readers to (B.3.3) in [10] for this formula. When it is evaluated at $\hat{\theta}_k = \arg \max_{\theta} p(\mathbf{y}_k|\theta)$, the FIM is an indication of the “sharpness” of the likelihood function around the local MLE. The large FIM provided by the data, the tighter lower variance bound (Cramér–Rao lower bound) we may obtain for the estimation [11].

Therefore, we use $\mathbf{J}(\hat{\theta}_k|\mathbf{y}_k)$ as an approximate inverse error covariance matrix for the weight of estimation $\hat{\theta}_k =$

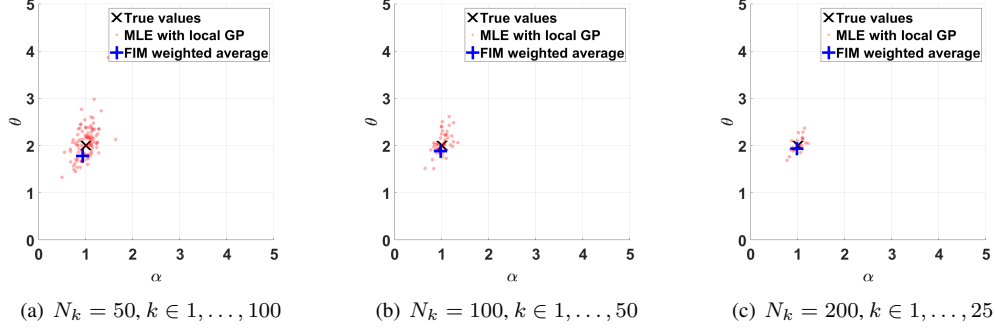


Fig. 1: FIM weighted average of MLEs based on segmented training data with different length N_k .

$\arg \max_{\theta} p(\mathbf{y}_k | \theta)$. The linear combination of the weighted MLEs $\{\hat{\theta}_k, \mathbf{J}(\hat{\theta}_k | \mathbf{y}_k) | k = 1, \dots, K\}$ is given by

$$\bar{\theta}_K = \left(\sum_{k=1}^K \mathbf{J}(\hat{\theta}_k | \mathbf{y}_k) \right)^{-1} \left(\sum_{k=1}^K \mathbf{J}(\hat{\theta}_k | \mathbf{y}_k) \hat{\theta}_k \right). \quad (10)$$

The above combination can also be implemented in an recursive manner [12]. To initialize the recursive combination, let $\mathbf{J}^{-1}(\hat{\theta}_0)$ be a diagonal matrix with very large diagonal elements to represent the high uncertainty of the prior given no data. The equations for recursive combination is $\bar{\theta}_K = \bar{\Lambda}_K^{-1} \mathbf{s}_K$, where $\bar{\Lambda}_0 = \mathbf{J}(\hat{\theta}_0)$, $\mathbf{s}_0 = \mathbf{J}(\hat{\theta}_0) \hat{\theta}_0$, and

$$\bar{\Lambda}_k = \bar{\Lambda}_{k-1} + \mathbf{J}(\hat{\theta}_k | \mathbf{y}_k), \quad (11)$$

$$\mathbf{s}_k = \mathbf{s}_{k-1} + \mathbf{J}(\hat{\theta}_k | \mathbf{y}_k) \hat{\theta}_k. \quad (12)$$

An example of recursive GP learning with a noiseless model of the zero mean function and the squared exponential (SE) covariance function ($\theta = [1, 2]^\top$) is shown in Figure 1. In the example we split the entire sequence of training data ($N = 5000$) into a number of segments. Three cases with different segmented data length ($N_k = 50, 100$, and 200) are tested. Note that MLE is an asymptotically efficient estimator. As a consequence, we observe that when N_k is too small, we are essentially combination many MLEs with low individual qualities, which would lead to an unsatisfying final result. Nevertheless, a large N_k leads to high computationally complexity. Therefore a trade-off between accuracy and computation cost must be balanced. As we have shown in the case of $N_k = 200$, the accuracy of the combined MLEs can be fulfilled while keeping the computation complexity low.

2.2. Prediction

We now present the equations for recursive Bayesian prediction with the composite likelihood model, which are solutions for computing the posterior of each step in Theorem 1. The complete derivation of the following equations are

similar to the derivations in Särkkä's book [13]. In short, at each recursion, the Bayesian filter uses the previous posterior as the new prior and updated the posterior by combining the new data segment. Specifically, given the posterior distribution $p(\mathbf{z} | \mathbf{y}_{1:k-1}) = \mathcal{N}(\tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k-1}}, \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}})$ and a new data segment $\mathbf{y}_k$, the posterior $p(\mathbf{z} | \mathbf{y}_{1:k}) = \mathcal{N}(\tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k}}, \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k}})$ is updated as the following.

First, a prior is generated. For $k = 1$, the GP prior in (2) is used as the prior distribution for $\mathbf{z}$; for $k \geq 2$, previous posterior $p(\mathbf{z} | \mathbf{y}_{1:k-1})$ is used as the new prior.

Second, the conditional distribution of the measurement $\mathbf{y}_k$ given $\mathbf{z}$ is calculated: $p(\mathbf{y}_k | \mathbf{z}) = \mathcal{N}(\mu_{\mathbf{y}_k | \mathbf{z}}, \Sigma_{\mathbf{y}_k | \mathbf{z}})$, where

$$\mu_{\mathbf{y}_k | \mathbf{z}} = \mu_{\mathbf{y}_k} + \Sigma_{\mathbf{y}_k, \mathbf{z}} \Sigma_{\mathbf{z}, \mathbf{z}}^{-1} (\mu_{\mathbf{z} | \mathbf{y}_{1:k-1}} - \mu_{\mathbf{z}}), \quad (13)$$

$$\Sigma_{\mathbf{y}_k | \mathbf{z}} = \Sigma_{\mathbf{y}_k, \mathbf{y}_k} - \Sigma_{\mathbf{y}_k, \mathbf{z}} \Sigma_{\mathbf{z}, \mathbf{z}}^{-1} \Sigma_{\mathbf{z}, \mathbf{y}_k}. \quad (14)$$

Note that at step k , the posterior mean of $\mathbf{z}$, $\mu_{\mathbf{z} | \mathbf{y}_{1:k-1}}$, is regarded as an realization of $\mathbf{z}$.

Finally, the posterior $p(\mathbf{z} | \mathbf{y}_{1:k}) = \mathcal{N}(\tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k}}, \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k}})$ is obtained via combining the k -th data segment, where

$$\tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k}} = \tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k-1}} + \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}} \mathbf{H}_k^\top \mathbf{G}_k^{-1} [\mathbf{y}_k - \mu_{\mathbf{y}_k} - \mathbf{H}_k (\tilde{\mu}_{\mathbf{z} | \mathbf{y}_{1:k-1}} - \mu_{\mathbf{z}})], \quad (15)$$

$$\tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k}} = \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}} - \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}} \mathbf{H}_k^\top \mathbf{G}_k^{-1} \mathbf{H}_k \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}}, \quad (16)$$

where $\mathbf{H}_k \triangleq \Sigma_{\mathbf{y}_k, \mathbf{z}} \Sigma_{\mathbf{z}, \mathbf{z}}^{-1}$ and $\mathbf{G}_k \triangleq \Sigma_{\mathbf{y}_k | \mathbf{z}} + \mathbf{H}_k \tilde{\Sigma}_{\mathbf{z} | \mathbf{y}_{1:k-1}} \mathbf{H}_k^\top$.

3. PERFORMANCE ANALYSIS

After obtaining the posterior with composite likelihoods and its computation methods, a natural question rises: how good are the posteriors with composite likelihoods as approximations of GP posteriors? To answer this question, we need to define the performance metrics. The first metric is the data averaged mean square error (MSE) of the approximate posterior; the second metric is the data averaged KL divergence between the approximate and the exact posteriors. Those two

metrics define the distance between the approximate posterior and the true distribution in different but closely related ways, which will be discussed through this section.

3.1. Data Averaged MSE

In light of the Theorem 17.3.2 from Cover and Thomas's Book [7], the MSE between the scalar random variable z and $\hat{z}$ (an estimate of z , e.g., the predictive mean of the GP posterior) is bounded by the following inequality:

$$\mathbb{E}[(z - \hat{z})^2] \geq \frac{1}{2\pi e} e^{2H(z|\mathbf{y})} = \frac{1}{2\pi e} e^{2(H(z) - I(z;\mathbf{y}))}, \quad (17)$$

which is a consequence of the Fano's inequality. In other words, the MSE lower bound is determined by the entropy of z and the mutual information between z and observations $\mathbf{y}$. Under the assumption of no model uncertainties and when the predictive mean of the GP posterior is used as the estimate, the MSE lower bound is asymptotically attained because $I(z; \mathbf{y})$ is the expected information provided by the full GP model, which will be discussed later.

It is unclear yet that to what extent the composite likelihood will introduce additional MSE compare to GP. Let the approximate posterior mean $\tilde{\mu}_{z|\mathbf{y}_{1:K}}$ be an estimation of the latent variable z with data $\mathbf{y}$ divided in to K segments, and the GP posterior mean is $\mu_{z|\mathbf{y}}$. The MSE of the estimation is given by

$$\begin{aligned} \text{MSE}(\tilde{\mu}_{z|\mathbf{y}_{1:K}}) &= \mathbb{E}_{z|\mathbf{y}}[(z - \tilde{\mu}_{z|\mathbf{y}_{1:K}})^2|\mathbf{y}] \\ &= \sigma_{z|\mathbf{y}}^2 + (\mu_{z|\mathbf{y}} - \tilde{\mu}_{z|\mathbf{y}_{1:K}})^2, \end{aligned} \quad (18)$$

where $\sigma_{z|\mathbf{y}}^2$ is the GP posterior error variance. Note that in the above expression, the full and composite posterior means ($\mu_{z|\mathbf{y}}$ and $\tilde{\mu}_{z|\mathbf{y}_{1:K}}$) are dependent of a specific realization of data $\mathbf{y}$. By averaging (18) over all possible $\mathbf{y}$, we present the data averaged MSE of using the composite posterior mean as an estimator of the scalar latent variable z , when the data $\mathbf{y}$ is divided into two segments ($K = 2$).

Theorem 2 (Data averaged MSE of posteriors with composite likelihoods). *The data averaged MSE of the approximate posterior mean ($K = 2$) for the scalar latent variable z is the GP posterior variance plus a quadratic form (the penalty term). That is*

$$\mathbb{E}_{\mathbf{y}}[\text{MSE}(\tilde{\mu}_{z|\mathbf{y}_{1:2}})] = \sigma_{z|\mathbf{y}}^2 + [\alpha_1 \quad \alpha_2] \begin{bmatrix} \Sigma_{\mathbf{y}_{1,1}} & \Sigma_{\mathbf{y}_{1,2}} \\ \Sigma_{\mathbf{y}_{2,1}} & \Sigma_{\mathbf{y}_{2,2}} \end{bmatrix} \begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix}, \quad (19)$$

where $\Sigma_{\mathbf{y}_{i,j}}$ is the covariance matrix of the i -th and j -th data segments. The vector of block matrices α_1 and α_2 is given by

$$\begin{bmatrix} \alpha_1 \\ \alpha_2 \end{bmatrix} = \left(\begin{bmatrix} A & B \\ C & D \end{bmatrix} - \begin{bmatrix} A' & B' \\ C' & D' \end{bmatrix} \right)^\top \begin{bmatrix} \Sigma_{\mathbf{x}^*, \mathbf{x}_1} \\ \Sigma_{\mathbf{x}^*, \mathbf{x}_2} \end{bmatrix}, \quad (20)$$

where $\Sigma_{\mathbf{x}^*, \mathbf{x}_i}$ is the column vector of covariances between the testing data and the i -th data segment; and the two block

matrices ($[A, B; C, D]$ and $[A', B'; C', D']$) are coefficients matrices of the GP and the approximate predictions, respectively.

The full computations are omitted here due to the limited space. We present a sketch of the proof in the following. It is not difficult to show that, the coefficients matrix for GP prediction is the block inverse of the covariance matrix of data $\mathbf{y}$. For approximate predictions, let the covariance matrix between data block i and j be $\tilde{\Sigma}_{\mathbf{y}_i, \mathbf{y}_j} \triangleq \Sigma_{\mathbf{y}_i, z} \frac{1}{\sigma_z^2} \Sigma_{z, \mathbf{y}_j}$, which means the segmented observations are indirectly connected by the testing data. Therefore the approximate conditional covariance matrix of the second data block given the first one can be expressed as $\tilde{\Sigma}_{\mathbf{y}_2|\mathbf{y}_1} \triangleq \Sigma_{\mathbf{y}_2, \mathbf{y}_2} - \tilde{\Sigma}_{\mathbf{y}_2, \mathbf{y}_1} \Sigma_{\mathbf{y}_1, \mathbf{y}_1}^{-1} \tilde{\Sigma}_{\mathbf{y}_1, \mathbf{y}_2}$. By some linear algebra manipulations, the block coefficients matrix for posterior with composite likelihoods is $A' = \Sigma_{\mathbf{y}_1, \mathbf{y}_1}^{-1}$, $B' = 0$, $C = -\left(\frac{\tilde{\sigma}_{z|\mathbf{y}_1}}{\sigma_z}\right)^2 \tilde{\Sigma}_{\mathbf{y}_2|\mathbf{y}_1}^{-1} \tilde{\Sigma}_{\mathbf{y}_2, \mathbf{y}_1} \Sigma_{\mathbf{y}_1, \mathbf{y}_1}^{-1}$, and $D = \left(\frac{\tilde{\sigma}_{z|\mathbf{y}_1}}{\sigma_z}\right)^2 \tilde{\Sigma}_{\mathbf{y}_2|\mathbf{y}_1}^{-1}$, where $\tilde{\sigma}_{z|\mathbf{y}_1}^2 = \sigma_z^2 - \Sigma_{z, \mathbf{y}_1} \Sigma_{\mathbf{y}_1, \mathbf{y}_1}^{-1} \Sigma_{\mathbf{y}_1, z}$ is the posterior variance of latent variable z using the first data block $\mathbf{y}_1$.

In case of the data is truthfully generated by a GP model, Theorem 2 shows that the MSE of composite likelihood predictions will be equal ($K = 1$) or higher ($K \geq 2$) than the MSE of GP predictions. As an approximation of the GP prediction, the composite likelihood prediction effectively imposes a Markov structure by assuming that the segmented training data are conditional independence given testing data. From sparse GP's point of view, we see that the testing data serves as the inducing variable by connecting segmented training data.

3.2. Data Averaged KL Divergence

Another well-known metric for comparing probability distributions is the KL divergence [7]. The KL divergence between the GP and the approximate posteriors can be written in a closed form expression. We refer readers to (A.23) in [1]. However, this expression is again for a specific realization of data $\mathbf{y}$. As an alternative, we present the following result for a comparison between the GP and approximate posteriors with their data averaged KL divergences to the prior distribution.

Theorem 3 (The difference of data averaged KL divergences). *The difference of data average KL divergence between two posteriors is*

$$\begin{aligned} &\mathbb{E}_{\mathbf{y}}[D(p(\mathbf{z}|\mathbf{y})||p(\mathbf{z}))] - \mathbb{E}_{\mathbf{y}_{1:K}}[D(p(\mathbf{z}|\mathbf{y}_{1:K})||p(\mathbf{z}))] \\ &= I(\mathbf{z}; \mathbf{y}) - \sum_{k=1}^K I(\mathbf{z}; \mathbf{y}_k), \end{aligned} \quad (21)$$

where $I(\mathbf{z}; \mathbf{y})$ is the mutual information of the latent state $\mathbf{z}$ and all observations $\mathbf{y}$, and $\sum_{k=1}^K I(\mathbf{z}; \mathbf{y}_k)$ is the sum of mutual information of the latent state $\mathbf{z}$ and observation $\mathbf{y}_k$.

A sketch of the proof is the following. First, the data averaged KL divergence between the GP posterior and the prior is given by $\mathbb{E}_{\mathbf{y}}[D(p(\mathbf{z}|\mathbf{y})||p(\mathbf{z}))] = \int_{\mathcal{Y}} \int_{\mathcal{Z}} p(\mathbf{z}, \mathbf{y}) [\ln \frac{1}{p(\mathbf{z})} + \ln p(\mathbf{z}|\mathbf{y})] d\mathbf{z} d\mathbf{y} = I(\mathbf{z}; \mathbf{y})$. Second, note that $p(\mathbf{y}_{1:K}|\mathbf{z}) = \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{z})$, $p(\mathbf{y}_{1:K}) = \int_{\mathcal{Z}} p(\mathbf{z}) \prod_{k=1}^K p(\mathbf{y}_k|\mathbf{z}) d\mathbf{z}$, and

$$\begin{aligned} & \int_{\mathcal{Y}_j} \int_{\mathcal{Y}_k} \int_{\mathcal{Z}} p(\mathbf{y}_j|\mathbf{z}) p(\mathbf{z}, \mathbf{y}_k) \ln \frac{p(\mathbf{z}|\mathbf{y}_k)}{p(\mathbf{z})} d\mathbf{z} d\mathbf{y}_k d\mathbf{y}_j \\ &= \int_{\mathcal{Y}_k} \int_{\mathcal{Z}} p(\mathbf{z}, \mathbf{y}_k) \ln \frac{p(\mathbf{z}|\mathbf{y}_k)}{p(\mathbf{z})} d\mathbf{z} d\mathbf{y}_k, \quad \forall j \neq k. \end{aligned} \quad (22)$$

Therefore, the average divergence between the approximate posterior and the prior is

$$\begin{aligned} & \mathbb{E}_{\mathbf{y}_{1:K}}[D(p(\mathbf{z}|\mathbf{y}_{1:K})||p(\mathbf{z}))] \\ &= \sum_{k=1}^K \int_{\mathcal{Y}_k} \int_{\mathcal{Z}} p(\mathbf{z}, \mathbf{y}_k) [\ln \frac{1}{p(\mathbf{z})} + \ln p(\mathbf{z}|\mathbf{y}_k)] d\mathbf{z} d\mathbf{y}_k \\ &= \sum_{k=1}^K I(\mathbf{z}; \mathbf{y}_k). \end{aligned} \quad (23)$$

Remark 1 (Expected information). *The data averaged KL divergence between a posterior and a prior is also referred as **the expected information** [14] about the latent variable $\mathbf{z}$, or, the average information gain about $\mathbf{z}$ provide by the data and the selected model.*

Since the data used to produce the GP and the composite posteriors are the same, the difference presented in Theorem 3 is caused by the selection of models (likelihoods). In other words, we have showed the difference of using one model (composite likelihoods) to approximate the other (exact likelihood) in terms of information about the latent variable $\mathbf{z}$. Note that the difference in (21) can be positive or negative. The sign of the difference can be interpreted as the following.

Remark 2 (Synergy and redundancy). *The difference of data average KL divergences in Theorem 3 can be positive or negative. If the difference is positive, the segmented data are said to be **synergistic**; otherwise, the segmented data is said to be **redundant** [15].*

In case of synergy, we obtained more information about the latent state by jointly considering all segmented data; in case of redundancy, there is an overlapped information in the segmented data about the latent variable. Finally, in the context of approximating GP posteriors, we interpret the sign of the difference with the following remark.

Remark 3 (Over- and underestimate of variances). *Considering the composite likelihood model as an approximation of the GP model, the synergistic segmented data lead to the overestimated variances in the approximate posterior distribution of the latent variables; and the redundant segmented data lead to the underestimated variances.*

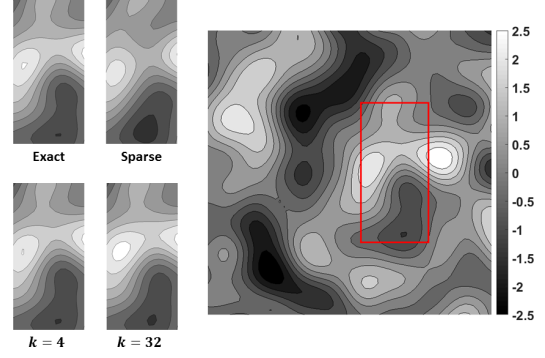


Fig. 2: An example of interpolating the missing data (within the red rectangular) of a 64×64 Gaussian random field. Three methods are applied: the exact Bayesian, the sparse method, and the recursive Bayesian with composite likelihoods ($k = 4$ and 32).

4. NUMERICAL EXAMPLES

A two-dimensional Gaussian process are sometimes referred as the Gaussian random field (GRF). Given the GRF model, a common problem is to interpolate the field at where values are missing. In this example we look at a GRF with the following definition: $y(x_1, x_2) \sim \mathcal{GP}(0, k(x_1, x_2, x'_1, x'_2))$, where $k(x_1, x_2, x'_1, x'_2) = \alpha^2 \exp[\frac{-(x_1 - x'_1)^2}{\theta_1^2} + \frac{-(x_2 - x'_2)^2}{\theta_2^2}]$. That is, the stationary GRF has the zero mean and the two-dimensional automatic relevance determination (ARD) squared exponential covariance function. The contour plot of a realization of the above GRF ($\alpha = 1, \theta_1 = 8, \theta_2 = 8$) is illustrated in Figure 2. This 64×64 GRF consists of more than four thousands “pixels”, which includes 3840 data as observations and 512 points within the red rectangular as testing data for interpolation. The predictive means given by the composite likelihood ($k = 4$) and the GP are most identical. Meanwhile, the predictive means between the sparse method and the exact method have notable differences.

A trade-off between RMSEs and runtimes is illustrated in Figure 3. Here we vary the size of data segments from $\frac{N}{4}$ to $\frac{N}{64}$ (or change the number of iterations from 4 to 64), while fixing the total size of data N ($N = 128^2$). When the data is split to more segments, the RMSE is increased due to the approximation but the complexity is mitigated. For example, when N_k is shorten from $\frac{N}{4}$ to $\frac{N}{16}$, the increased RMSE is insignificant but the runtime is greatly reduced. We also show that the different segmentation schemes has notable impact on the approximation performance. When the data is split in a random manner, the overlapped information between segments are greater than the sequentially divided segments, and therefore introduce more errors.

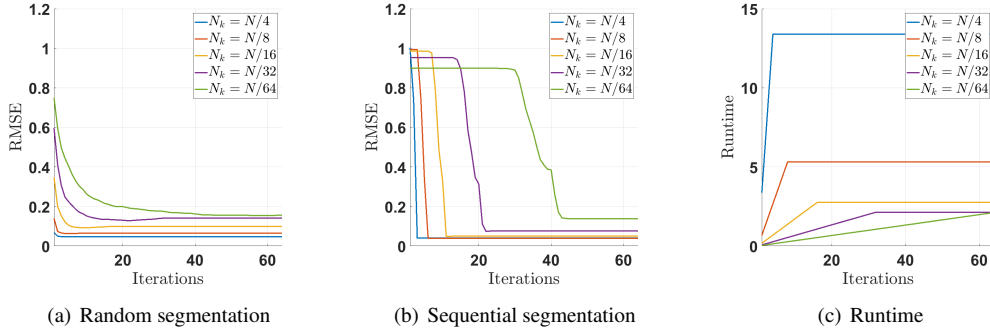


Fig. 3: The trade-off between RMSE and runtime, using composite likelihoods with different sizes of data segments. Two different segmentation schemes are applied: random and sequential. $N = 128^2$ in this simulation.

5. CONCLUSION

We have presented a method for approximating the posterior of a GP using composite likelihoods. We provided both a scalable computational method and a performance analysis of the resulting approximation. Differing from many previous works, the approximate posterior from a general belief updating framework [6]. We show that exact and approximate posteriors can be all recovered with this framework by choosing different likelihoods in the loss function. While this provides a computational gain we also quantify the quality of the approximation using averaged mean-square prediction error and KL divergences. Using the KL divergences and mutual information, we give an intuitive interpretations of how and why segmentation of data will lead to differences in posteriors.

6. REFERENCES

- [1] Carl E. Rasmussen and Christopher K. I. Williams, *Gaussian processes for machine learning*, MIT Press, Cambridge, Massachusetts, 2006.
- [2] Christopher K. I. Williams and Matthias Seeger, “Using the nyström method to speed up kernel machines,” in *Advances in Neural Information Processing Systems 13*, pp. 682–688. MIT Press, 2001.
- [3] Joaquin Quiñonero Candela and Carl Edward Rasmussen, “A unifying view of sparse approximate gaussian process regression,” *J. Mach. Learn. Res.*, vol. 6, pp. 1939–1959, Dec. 2005.
- [4] Volker Tresp, “A bayesian committee machine,” *Neural Comput.*, vol. 12, no. 11, pp. 2719–2741, Nov. 2000.
- [5] Michalis Titsias, “Variational learning of inducing variables in sparse gaussian processes,” in *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*. 16–18 Apr 2009, vol. 5, pp. 567–574, PMLR.
- [6] P. G. Bissiri, C. C. Holmes, and S. G. Walker, “A general framework for updating belief distributions,” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 78, no. 5, pp. 1103–1130, 2016.
- [7] Thomas M. Cover and Joy A. Thomas, *Elements of information theory*, Wiley, New York, 1991.
- [8] Charles W. Fox and Stephen J. Roberts, “A tutorial on variational bayesian inference,” *Artificial Intelligence Review*, vol. 38, no. 2, pp. 85–95, Aug 2012.
- [9] Cristiano Varin, Nancy Reid, and David Firth, “An overview of composite likelihood methods,” *Statistica Sinica*, vol. 21, no. 1, pp. 5–42, 2011.
- [10] Petre Stoica and Randolph L. Moses, *Introduction to spectral analysis*, Prentice Hall, Upper Saddle River, 1997.
- [11] Steven M. Kay, *Fundamentals of statistical signal processing: Vol. 1, Estimation theory*, Prentice Hall PTR, Upper Saddle River, N.J, 1993.
- [12] D. Zachariah, S. Dwivedi, P. Händel, and P. Stoica, “Scalable and passive wireless network clock synchronization in los environments,” *IEEE Transactions on Wireless Communications*, vol. 16, no. 6, pp. 3536–3546, June 2017.
- [13] Simo Särkkä, *Bayesian Filtering and Smoothing*, Cambridge University Press, New York, NY, USA, 2013.
- [14] Jose M. Bernardo, “Reference posterior distributions for bayesian inference,” *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 41, no. 2, pp. 113–147, 1979.
- [15] Adam B. Barrett, “Exploration of synergistic and redundant information sharing in static and dynamical gaussian systems,” *Phys. Rev. E*, vol. 91, pp. 052802, May 2015.