Proceedings of the 2024 Winter Simulation Conference
H. Lam, E. Azar, D. Batur, S. Gao, W. Xie, S. R. Hunter, and M. D. Rossetti, eds.

ENERGETIC VARIATIONAL GAUSSIAN PROCESS REGRESSION

Lulu Kang[1], Yuanxing Cheng[2], Yiwei Wang[3], and Chun Liu[2]

[1]Dept. of Mathematics and Statistics, University of Massachusetts, Amherst, MA, USA
[2]Dept. of Applied Mathematics, Illinois Institute of Technology, Chicago, IL, USA
[3]Dept. of Mathematics, University of California, Riverside, CA, USA

ABSTRACT

The Gaussian process (GP) regression model is a widely employed supervised learning approach. In this paper, we estimate the GP model through variational inference, particularly employing the recently introduced energetic variational inference method. Under the GP model assumptions, we derive posterior distributions for its parameters. The energetic variational inference approach bridges the Bayesian sampling and optimization and enables approximation of the posterior distributions and identification of the posterior mode. By incorporating a Gaussian prior on the mean component of the GP model, we also apply shrinkage estimation to the parameters, facilitating variable selection of the mean function. The proposed GP method outperforms some existing software packages on three benchmark examples.

1 INTRODUCTION

Uncertainty Quantification (Ghanem et al. 2017) is a highly interdisciplinary research domain involving mathematics, statistics, optimization, advanced computing technology, and various science and engineering disciplines. It provides a computational framework for quantifying input and response uncertainties and making model-based predictions and their inferences for complex science or engineering systems/processes. One key topic in uncertainty quantification is to analyze computer experimental data and build a surrogate model for the computer simulation model. Gaussian process (GP) regression model, sometimes known as "kriging", has been widely used for this purpose since the seminal paper by Sacks et al. (1989).

This paper examines GP models within a Bayesian framework, adopting the model assumptions specified by Santner et al. (2003), which include a meaningful mean function and an anisotropic covariance function for enhanced prediction flexibility. Unlike traditional Bayesian GP models that rely on Markov Chain Monte Carlo (MCMC) sampling, we introduce a variational inference method, specifically a particle-based energetic variational inference approach (EVI) from Wang, Chen, Liu, and Kang (2021), termed EVI-GP. This method, which can compute maximum a posteriori (MAP) estimates or approximate posterior distributions, offers an efficient alternative for parameter estimation and inference. Additionally, employing an l_2-regularization due to the conjugate prior of the regression coefficients facilitates sparsity in the GP mean function, optimizing model interpretation and accuracy.

The rest of the paper is organized as follows. In Section 2, we review the Gaussian process model, including its assumption and prior distributions, and derive the posterior and posterior predictive distributions. In Section 3, the preliminary background on variational inference and the particle energetic variational inference methods are briefly reviewed. The EVI-GP method is also summarized at the end of this section. Section 4 shows two main simulation examples in which different versions of EVI-GP are compared with other existing methods. The paper concludes in Section 5.

2 GAUSSIAN PROCESS REGRESSION: BAYESIAN APPROACH

2.1 Gaussian Process Assumption

Denote $\{\boldsymbol{x}_i, y_i\}_{i=1}^n$ as the n pairs of input and output data from a certain computer experiment, and $\boldsymbol{x}_i \in \Omega \subseteq \mathbb{R}^d$ are the ith experimental input values and $y_i \in \mathbb{R}$ is the corresponding output. In this paper, we only consider the case of univariate response, but the proposed EVI-GP can be applied to the multi-response GP model which involves the cross-covariance between responses (Cressie 2015).

Gaussian process regression is built on the following model assumption of the response,

$$y_i = \boldsymbol{g}(\boldsymbol{x}_i)^\top \boldsymbol{\beta} + Z(\boldsymbol{x}_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{1}$$

where $\boldsymbol{g}(\boldsymbol{x})$ is a p-dim vector of user-specified basis functions and $\boldsymbol{\beta}$ is the p-dim vector of linear coefficients corresponding the basis functions. Usually, $\boldsymbol{g}(\boldsymbol{x})$ contains the polynomial basis functions of $\boldsymbol{x}$ up to a certain order. The random noise ε_i's are independently and identically distributed following $N(0, \sigma^2)$. They are also independent of the other stochastic components of (1). We assume the GP prior on the stochastic function $Z(\boldsymbol{x})$, which is denoted as $Z(\cdot) \sim GP(0, \tau^2 K)$, i.e., $\mathbb{E}[Z(\boldsymbol{x})] = 0$ and the covariance function $\text{cov}[Z(\boldsymbol{x}_1), Z(\boldsymbol{x}_2)] = \tau^2 K(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega})$. We use the common stationary assumption of $Z(\boldsymbol{x})$, and thus the variance τ^2 is a constant. The function $K(\cdot, \cdot; \boldsymbol{\omega}) : \Omega \times \Omega \mapsto \mathbb{R}_+$ is the correlation of the stochastic process with hyperparameters $\boldsymbol{\omega}$. For it to be valid, $K(\cdot, \cdot; \boldsymbol{\omega})$ must be a symmetric positive definite kernel function. A commonly used kernel function is the Gaussian kernel defined as $K(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega}) = \exp\left[-\sum_{j=1}^d \omega_j (x_{1j} - x_{2j})^2\right]$, with $\boldsymbol{\omega} \in \mathbb{R}^d$ and $\boldsymbol{\omega}$ is a non-negative vector. The EVI-GP method can be applied in the same way for other kernel functions and non-stationary GP assumptions. The response $Y(\boldsymbol{x})$ follows a Gaussian Process with mean function $\mathbb{E}[Y(\boldsymbol{x})] = \boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\beta}$ for any $\boldsymbol{x} \in \Omega$ and covariance function

$$\text{cov}[Y(\boldsymbol{x}_1), Y(\boldsymbol{x}_2)] = \tau^2 K(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega}) + \sigma^2 \delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = \tau^2 \left[K(\boldsymbol{x}_1, \boldsymbol{x}_2; \boldsymbol{\omega}) + \eta\, \delta(\boldsymbol{x}_1, \boldsymbol{x}_2)\right], \quad \forall \boldsymbol{x}_1, \boldsymbol{x}_2 \in \Omega,$$

where $\delta(\boldsymbol{x}_1, \boldsymbol{x}_2) = 1$ if $\boldsymbol{x}_1 = \boldsymbol{x}_2$ and 0 otherwise, and $\eta = \sigma^2/\tau^2$. So η is interpreted as the noise-to-signal ratio if $\sigma^2 > 0$ or a nugget effect if $\sigma^2 = 0$ to avoid ill-conditioning of the covariance matrix (Peng and Wu 2014). The unknown parameter values of the GP model are $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega}, \tau^2, \eta)$. We are going to show how to obtain the estimation and inference of the parameters using the Bayesian framework.

2.2 GP under Bayesian Framework

We assume the following prior distributions for the parameters $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega}, \tau^2, \eta)$,

$$\begin{aligned} &\boldsymbol{\beta} \sim MVN_p(\boldsymbol{0}, v^2 \boldsymbol{R}), && \omega_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(a_\omega, b_\omega), \text{ for } i = 1, \ldots, d \\ &\tau^2 \sim \text{Inverse-}\chi^2(df_{\tau^2}), && \eta \sim \text{Gamma}(a_\eta, b_\eta). \end{aligned} \tag{2}$$

These distribution families are commonly used in the literature such as Gramacy and Lian (2012) and Hu, Gramacy, and Lian (2013). However, the choice of parameters of the prior distributions should require fine-tuning using testing data or cross-validation procedures. In some literature, parameters $\boldsymbol{\omega}$, τ^2, and η are considered to be *hyperparameters*. The conditional posterior distribution of $\boldsymbol{\beta}$ given data and $(\boldsymbol{\omega}, \tau^2, \eta)$ is a multivariate normal distribution, which is shown later.

Next, we derive the posterior distributions and some conditional posterior distributions. Based on the data, the sampling distribution is

$$\boldsymbol{y}_n | \boldsymbol{\theta} \sim MVN_n\left(\boldsymbol{G}\boldsymbol{\beta}, \tau^2(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)\right),$$

where $\boldsymbol{y}_n$ is the vector of y_i's and $\boldsymbol{G}$ is a matrix of row vectors $\boldsymbol{g}(\boldsymbol{x}_i)^\top$'s. The matrix $\boldsymbol{K}_n$ is the $n \times n$ kernel matrix with entries $K_n[i, j] = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$ and is a symmetric and positive definite matrix, and $\boldsymbol{I}_n$ is an $n \times n$ identity matrix. The density function of $\boldsymbol{y}_n | \boldsymbol{\theta}$ is

$$p(\boldsymbol{y}_n | \boldsymbol{\theta}) \propto (\tau^2)^{-\frac{n}{2}} \det(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1/2} \exp\left(-\frac{1}{2\tau^2} (\boldsymbol{y}_n - \boldsymbol{G}\boldsymbol{\beta})^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} (\boldsymbol{y}_n - \boldsymbol{G}\boldsymbol{\beta})\right).$$

Following Bayes' Theorem, the joint posterior distribution of all parameters is

$$p(\boldsymbol{\theta}|\boldsymbol{y}_n) \propto p(\boldsymbol{\theta})p(\boldsymbol{y}_n|\boldsymbol{\theta}) \propto p(\boldsymbol{\beta}) \left(\prod_{j=1}^{d} p(\omega_i) \right) p(\tau^2)p(\eta)p(\boldsymbol{y}_n|\boldsymbol{\theta}).$$

The conditional posterior distribution of $\boldsymbol{\beta}$ can be easily obtained through conjugacy. It is also straightforward to obtain the posterior distribution $p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n)$. The results are summarized in Proposition 1. The posterior distribution of τ^2 can be significantly simplified if a non-informative prior is used for $\boldsymbol{\beta}$, as described in Proposition 2. Due to limited space, all proofs can be found in the appendix of Kang, Cheng, Wang, and Liu (2024).

Proposition 1 Using the prior distribution of $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\omega}, \tau^2, \eta)$ in (2), the conditional posterior distribution of $\boldsymbol{\beta}$ is $\boldsymbol{\beta}|\boldsymbol{y}_n, \boldsymbol{\omega}, \tau^2, \eta \sim MVN_p\left(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|n}\right)$, where

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}|n} = \left[\frac{1}{\tau^2} \boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{G} + \frac{1}{\nu^2} \boldsymbol{R}^{-1} \right]^{-1}, \quad \hat{\boldsymbol{\beta}}_n = \tau^{-2} \boldsymbol{\Sigma}_{\boldsymbol{\beta}|n} \left[\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \right] \boldsymbol{y}_n.$$

The marginal posterior distribution of $(\boldsymbol{\omega}, \tau^2, \eta)$ is

$$\begin{aligned} p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n) \propto & \det(\boldsymbol{\Sigma}_{\boldsymbol{\beta}|n})^{1/2} \exp\left[-\frac{1}{2} \hat{\boldsymbol{\beta}}_n^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}|n}^{-1} \hat{\boldsymbol{\beta}}_n - \frac{1}{2\tau^2} \boldsymbol{y}_n^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{y}_n \right] \\ & \times (\tau^2)^{-n/2} \det(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1/2} p(\tau^2) p(\boldsymbol{\omega}) p(\eta). \end{aligned} \tag{3}$$

Proposition 2 If using a non-informative prior distribution for $\boldsymbol{\beta}$, i.e., $p(\boldsymbol{\beta}) \propto 1$, and the same prior distributions for $(\boldsymbol{\omega}, \tau^2, \eta)$ in (2), the conditional posterior distribution of $\boldsymbol{\beta}$ is still $MVN_p\left(\hat{\boldsymbol{\beta}}_n, \boldsymbol{\Sigma}_{\boldsymbol{\beta}|n}\right)$, but the covariance and mean are

$$\boldsymbol{\Sigma}_{\boldsymbol{\beta}|n} = \tau^2 \left[\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{G} \right]^{-1}, \quad \hat{\boldsymbol{\beta}}_n = \left[\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{G} \right]^{-1} \left[\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \right] \boldsymbol{y}_n.$$

The conditional posterior distribution for τ^2 is

$$\tau^2|\boldsymbol{\omega}, \eta, \boldsymbol{y}_n \sim \text{Scaled Inverse-}\chi^2(df_{\tau^2} + n - p, \hat{\tau}^2),$$

where $\hat{\tau}^2 = (1 + s_n^2)/(df_{\tau^2} + n - p)$, $s_n^2 = \tau^{-2} \hat{\boldsymbol{\beta}}_n^\top \boldsymbol{\Sigma}_{\boldsymbol{\beta}|n}^{-1} \hat{\boldsymbol{\beta}}_n + \boldsymbol{y}_n^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{y}_n$. The marginal posterior of $(\boldsymbol{\omega}, \eta)$ is

$$p(\boldsymbol{\omega}, \eta|\boldsymbol{y}_n) \propto (\hat{\tau}^2)^{-\frac{1}{2}(df_{\tau^2} + n - p)} \det(\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{G})^{-1/2} \det(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1/2} p(\boldsymbol{\omega}) p(\eta). \tag{4}$$

If we use non-informative prior distributions for all the parameters, i.e., $p(\boldsymbol{\beta}) \propto 1$ and $p(\omega_i) \overset{\text{i.i.d.}}{\sim} \text{Uniform}[a_\omega, b_\omega]$ for $i = 1, \ldots, d$, $p(\tau^2) \propto \tau^{-2}$, and $p(\eta) \propto \text{Uniform}[a_\eta, b_\eta]$, the Bayesian framework is equivalent to the empirical Bayesian or maximum likelihood estimation method. The GP regression model estimated via this *frequentist* approach is common in both methodology research and application (Santner et al. 2003; Fang et al. 2005; Gramacy 2020). In this paper, we consider the empirical Bayesian estimation as a special case of the Bayesian GP model. The choice between the two different types of prior distributions for $\boldsymbol{\beta}$, informative or non-informative, is subject to the dimension of the input variables, the assumption on basis functions $\boldsymbol{g}(\boldsymbol{x})$, the goal of GP modeling (accurate prediction v.s. interpretation), and sometimes the application of the computer experiment. Both types have their unique advantages and shortcomings. The non-informative prior distribution for $\boldsymbol{\beta}$ reduces the computation involved in the posterior sampling for τ^2, but we would lose the l_2 regularization effect on $\boldsymbol{\beta}$ brought by the informative prior $\boldsymbol{\beta}$.

One issue with the informative prior distribution is to choose its parameters, i.e., the constant variance v^2 and the correlation matrix $\boldsymbol{R}$. Here we recommend using a cross-validation procedure to select v^2. If the mean function $\boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\beta}$ is a polynomial function of the input variables, we specify the matrix $\boldsymbol{R}$ to be a diagonal matrix $\boldsymbol{R} = \text{diag}\{1, r, \ldots, r, r^2, \ldots, r^2, \ldots\}$, where $r \in (0,1)$ is a user-specified parameter. The power index of r is the same as the order of the corresponding polynomial term. For example, if $\boldsymbol{g}(\boldsymbol{x})^\top \boldsymbol{\beta}$ with $\boldsymbol{x} \in \mathbb{R}^2$ is a full quadratic model and contains the terms $\boldsymbol{g}(\boldsymbol{x}) = [1, x_1, x_2, x_1^2, x_2^2, x_1 x_2]^\top$, the corresponding prior correlation matrix should be specified as $\boldsymbol{R} = \text{diag}\{1, r, r, r^2, r^2, r^2\}$. In this way, the prior variance of the effect decreases exponentially as the order of effect increases, following the *effect hierarchy principle* (Hamada and Wu 1992; Wu and Hamada 2021). It states that lower-order effects are more important than higher-order effects, and the effects of the same order are equally important. The hierarchy ordering principle can reduce the size of the model and avoid including higher-order and less significant model terms. Such prior distribution was firstly proposed by Joseph (2006), and later used in Kang and Joseph (2009), Ai et al. (2009), Kang et al. (2018), Kang and Huang (2019),Kang et al. (2021), and Kang et al. (2023).

Proposition 3 Given the parameters $(\boldsymbol{\omega}, \tau^2, \eta)$, the posterior predictive distribution of $y(\boldsymbol{x})$ at any query point $\boldsymbol{x}$ is the following normal distribution.

$$y(\boldsymbol{x})|\boldsymbol{y}_n, \boldsymbol{\omega}, \tau^2, \eta \sim N(\hat{\mu}(\boldsymbol{x}), \sigma_n^2(\boldsymbol{x})),$$

where

$$\begin{aligned}
\hat{\mu}(\boldsymbol{x}) &= \boldsymbol{g}(\boldsymbol{x})^\top \hat{\boldsymbol{\beta}}_n + K(\boldsymbol{x}, \mathscr{X}_n)(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1}(\boldsymbol{y}_n - \boldsymbol{G}\hat{\boldsymbol{\beta}}_n), \\
\sigma_n^2(\boldsymbol{x}) &= \tau^2 \left\{ 1 - K(\boldsymbol{x}, \mathscr{X}_n)(\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} K(\mathscr{X}_n, \boldsymbol{x}) + \boldsymbol{c}(\boldsymbol{x})^\top \left[\boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} \boldsymbol{G} \right]^{-1} \boldsymbol{c}(\boldsymbol{x}) \right\},
\end{aligned}$$

where $\boldsymbol{c}(\boldsymbol{x}) = \boldsymbol{g}(\boldsymbol{x}) - \boldsymbol{G}^\top (\boldsymbol{K}_n + \eta \boldsymbol{I}_n)^{-1} K(\mathscr{X}_n, \boldsymbol{x})$, $K(\boldsymbol{x}, \mathscr{X}_n) = K(\mathscr{X}_n, \boldsymbol{x})^\top = [K(\boldsymbol{x}, \boldsymbol{x}_1), \ldots, K(\boldsymbol{x}, \boldsymbol{x}_n)]$. For non-informative prior, the posterior predictive distribution $y(\boldsymbol{x})|\boldsymbol{y}_n, \boldsymbol{\omega}, \eta$ is the same except that τ^2 is replaced by $\hat{\tau}^2$.

A detailed proof can be found in Santner et al. (2003) and Rasmussen and Williams (2006). Thanks to the Gaussian process assumption and the conditional conjugate prior distributions for $\boldsymbol{\beta}$ and τ^2, Proposition 1, 2, and 3 give the explicit and easy to generate conditional posterior distribution of $\boldsymbol{\beta}$ (and τ^2) and posterior predictive distribution. Therefore, how to generate samples from $p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n)$ in (3) or $p(\boldsymbol{\omega}, \eta|\boldsymbol{y}_n)$ in (4) is the bottleneck of the computation for GP models. Since $p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n)$ and $p(\boldsymbol{\omega}, \eta|\boldsymbol{y}_n)$ are not from any known distribution families, Metropolis-Hastings (MH) algorithm (Robert et al. 1999; Gelman et al. 2014), Hamiltonian Monte Carlo (HMC) (Neal 1996), or Metropolis-adjusted Langevin algorithm (MALA) can be used for sampling (Roberts and Rosenthal 1998). In this paper, we introduce readers to an alternative computational tool, namely, a variational inference approach to approximate the posterior distribution $p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n)$ or $p(\boldsymbol{\omega}, \eta|\boldsymbol{y}_n)$. More specifically, we plan to use energetic variational inference, a particle method, to generate posterior samples.

3 ENERGETIC VARIATIONAL INFERENCE GAUSSIAN PROCESS

Variational inference-based GP models have been explored in prior works such as Tran et al. (2016), Cheng and Boots (2017), and Wynne and Wild (2022). Despite sharing the variational inference idea, the proposed Energetic Variational Inference (EVI) GP differs significantly from these existing methods regarding the specific variational techniques employed. Tran et al. (2016) utilized auto-encoding, while Cheng and Boots (2017) and Wynne and Wild (2022) employed the mean-field variational inference (Blei et al. 2017).

The EVI approach presented in this paper is a newly introduced particle-based method. It offers simplicity in implementation across diverse applications without the need for training any neural networks. Notably, EVI establishes a connection between the MAP procedure and posterior sampling through a

user-specified number of particles. In contrast to the complexity of auto-encoding variational methods, the particle-based approach is much simpler, devoid of any neural network intricacies. Moreover, in comparison to mean-field methods, both particle-based and MAP-based approaches (auto-encoding falls into the MAP-based category) can exhibit enhanced accuracy, as they do not impose any parametric assumptions on a feasible family of distributions in the optimization to solve the variational problem.

3.1 Energetic Variational Inference

Due to limited space, we can only briefly review the EVI framework and explain it intuitively. Readers can refer to Wang et al. (2021) and Kang et al. (2024) for a more comprehensive review of the energetic variational approach. What is more, Wang et al. (2021) also suggested many different variants of algorithms under the EVI framework.

We first introduce the EVI using the continuous formulation. Let $\boldsymbol{\phi}_t$ be the dynamic flow map $\boldsymbol{\phi}_t : \mathbb{R}^d \to \mathbb{R}^d$ at time t that continuously transforms the d-dimensional distribution from an initial distribution toward the target one and we require the map $\boldsymbol{\phi}_t$ to be smooth and one-to-one. The functional $\mathscr{F}(\boldsymbol{\phi}_t)$ is a user-specified divergence or other machine learning objective functional. Taking the analogy of a thermodynamics system, $\mathscr{F}(\boldsymbol{\phi}_t)$ is the Helmholtz free energy. Following the First and Second Law of thermodynamics and set the kinetic energy to zero, we have

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}(\boldsymbol{\phi}_t) = -\triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t), \tag{5}$$

where $\triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t)$ is a user-specified functional representing the rate of energy dissipation, and $\dot{\boldsymbol{\phi}}_t$ is the derivative of $\boldsymbol{\phi}_t$ with time t. So $\dot{\boldsymbol{\phi}}_t$ can be interpreted as the "velocity" of the transformation. Each variational formulation gives a natural path of decreasing the objective functional $\mathscr{F}(\boldsymbol{\phi}_t)$ toward an equilibrium. The dissipation functional should satisfy $\triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t) \geq 0$ so that $\mathscr{F}(\boldsymbol{\phi}_t)$ decreases with time. A simple yet effective specification of $\triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t)$ is a quadratic functional in terms of $\dot{\boldsymbol{\phi}}_t$,

$$\triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t) = \int_{\Omega_t} \rho_{[\boldsymbol{\phi}_t]} \|\dot{\boldsymbol{\phi}}_t\|_2^2 \mathrm{d}\boldsymbol{x}, \quad \text{and it has the variation (functional derivative)} \quad \frac{\delta \triangle(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t)}{\delta \dot{\boldsymbol{\phi}}_t} = 2\rho_{[\boldsymbol{\phi}_t]}\dot{\boldsymbol{\phi}}_t$$

where $\rho_{[\boldsymbol{\phi}_t]}$ denotes the pdf of the current distribution which is the initial distribution transformed by $\boldsymbol{\phi}_t$, Ω_t is the current support, $\|\boldsymbol{a}\|_2 = \boldsymbol{a}^\top \boldsymbol{a}$ for $\forall \boldsymbol{a} \in \mathbb{R}^d$, and δ is the variation operator.

With the specified energy-dissipation law (5), the energy variational approach derives the dynamics of the systems through two variational procedures, the Least Action Principle (LAP) and the Maximum Dissipation Principle (MDP), which leads to

$$\frac{\delta \frac{1}{2}\triangle}{\delta \dot{\boldsymbol{\phi}}_t} = -\frac{\delta \mathscr{F}}{\delta \boldsymbol{\phi}_t}, \quad \text{and} \quad \rho_{[\boldsymbol{\phi}_t]}\dot{\boldsymbol{\phi}}_t = -\frac{\delta \mathscr{F}}{\delta \boldsymbol{\phi}_t},$$

using the quadratic $\Delta(\boldsymbol{\phi}_t, \dot{\boldsymbol{\phi}}_t)$.

In general, this continuous formulation (5) is difficult to solve since the manifold of $\boldsymbol{\phi}_t$ is of infinite dimension. Naturally, there are different approaches to approximate an infinite-dimensional manifold by a finite-dimensional manifold. One such approach, as used in Wang et al. (2021), is to use *particles* (or *samples*) to approximate the $\rho_{[\boldsymbol{\phi}_t]}$ in (5) with kernel regularization, before any variational steps. It leads to a discrete version of the energy-dissipation law, i.e.,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathscr{F}_h(\{\boldsymbol{x}_i(t)\}_{i=1}^N) = -\triangle_h(\{\boldsymbol{x}_i(t)\}_{i=1}^N, \{\dot{\boldsymbol{x}}_i(t)\}_{i=1}^N). \tag{6}$$

Here $\{\boldsymbol{x}(t)\}_{i=1}^N$ is the locations of N particles at time t and $\dot{\boldsymbol{x}}_i(t)$ is the derivative of $\boldsymbol{x}_i$ with t, and thus is the velocity of the ith particle as it moves toward the target distribution. The subscript h of $\mathscr{F}$ and $\triangle$

denotes the bandwidth parameter of the kernel function used in the kernelization operation. Applying the variational steps to (6), we obtain the dynamics of decreasing $\mathscr{F}$ at the particle level,

$$\frac{\delta \frac{1}{2} \triangle_h}{\delta \dot{\boldsymbol{x}}_i(t)} = -\frac{\delta \mathscr{F}_h}{\delta \boldsymbol{x}_i}, \quad \text{for } i = 1, \ldots, N. \tag{7}$$

This leads to an ODE system of $\{\boldsymbol{x}_i(t)\}_{i=1}^N$ and can be solved by different numerical schemes, such as first order explicit and implicit Euler approaches shown in Wang et al. (2021). The solution is the particles approximating the target distribution.

In this paper, we use the KL-divergence as the energy functional as demonstrated in Wang et al. (2021),

$$D_{\mathrm{KL}}(\rho || \rho^*) = \int_\Omega \rho(\boldsymbol{x}) \log\left(\frac{\rho(\boldsymbol{x})}{\rho^*(\boldsymbol{x})}\right) \mathrm{d}\boldsymbol{x},$$

where $\rho^*(\boldsymbol{x})$ is the density function of the target distribution with support region Ω and $\rho(\boldsymbol{x})$ is to approximate $\rho^*(\boldsymbol{x})$. For EVI-GP, ρ^* is the posterior distribution of $p(\boldsymbol{\omega}, \tau^2, \eta \,|\boldsymbol{y}_n)$ or $p(\boldsymbol{\omega}, \eta \,|\boldsymbol{y}_n)$.

Using the KL-divergence, the divergence functional at time t is

$$\mathscr{F}(\boldsymbol{\phi}_t) = \int \left(\rho_{[\boldsymbol{\phi}_t]}(\boldsymbol{x}) \log \rho_{[\boldsymbol{\phi}_t]}(\boldsymbol{x}) + \rho_{[\boldsymbol{\phi}_t]}(\boldsymbol{x}) V(\boldsymbol{x})\right) \mathrm{d}\boldsymbol{x},$$

where $V(\boldsymbol{x}) = \log \rho^*(\boldsymbol{x})$, which is known up to a scaling constant. The discrete version of the energy becomes

$$\mathscr{F}_h\left(\{\boldsymbol{x}_i\}_{i=1}^N\right) = \frac{1}{N} \sum_{i=1}^N \left(\ln \left[\frac{1}{N} \sum_{j=1}^N K_h(\boldsymbol{x}_i, \boldsymbol{x}_j) \right] + V(\boldsymbol{x}_i) \right),$$

and the discrete dissipation is

$$-2\Delta_h\left(\{\boldsymbol{x}_i\}_{i=1}^N\right) = -\frac{1}{N} \sum_{i=1}^N |\dot{\boldsymbol{x}}_i(t))|^2.$$

Applying variational step to (6), we obtain (7) which is equivalent to the following nonlinear ODE system:

$$\dot{\boldsymbol{x}}_i(t) = -\left(\frac{\sum_{j=1}^N \nabla_{\boldsymbol{x}_i} K_h(\boldsymbol{x}_i, \boldsymbol{x}_j)}{\sum_{j=1}^N K_h(\boldsymbol{x}_i, \boldsymbol{x}_j)} + \sum_{k=1}^N \frac{\nabla_{\boldsymbol{x}_i} K_h(\boldsymbol{x}_k, \boldsymbol{x}_i)}{\sum_{j=1}^N K_h(\boldsymbol{x}_k, \boldsymbol{x}_j)} + \nabla_{\boldsymbol{x}_i} V(\boldsymbol{x}_i) \right), \tag{8}$$

for $i = 1, \ldots, N$. The iterative update of N particles $\{\boldsymbol{x}_i\}_{i=1}^N$ involves solving the nonlinear ODE system (8) via optimization problem (9) at the m-th iteration step

$$\{\boldsymbol{x}_i^{m+1}\}_{i=1}^N = \mathrm{argmin}_{\{\boldsymbol{x}_i\}_{i=1}^N} J_m(\{\boldsymbol{x}_i\}_{i=1}^N), \tag{9}$$

where

$$J_m(\{\boldsymbol{x}_i\}_{i=1}^N) := \frac{1}{2\tau} \sum_{i=1}^N ||\boldsymbol{x}_i - \boldsymbol{x}_i^m||^2 / N + \mathscr{F}_h(\{\boldsymbol{x}_i\}_{i=1}^N).$$

Here we use $\boldsymbol{x}_i^m$ and $\boldsymbol{x}_i^{m+1}$ to denote the i-th particle value in the m-th and $m+1$-th iteration, respectively. Wang et al. (2021) emphasized the advantages of using the Implicit-Euler solver for enhanced numerical stability in this process. We summarize the algorithm of using the implicit Euler scheme to solve the ODE system (8) into Algorithm 1. Here MaxIter is the maximum number of iterations of the outer loop. The minimization of J_m in the inner loop is solved by L-BFGS in our implementation.

Algorithm 1 EVI with Implicit Euler Scheme (EVI-Im)

Input: The target distribution $\rho^*(\boldsymbol{x})$ and a set of initial particles $\{\boldsymbol{x}_i^0\}_{i=1}^N$ drawn from a prior $\rho_0(\boldsymbol{x})$.
Output: A set of particles $\{\boldsymbol{x}_i^*\}_{i=1}^N$ approximating ρ^*.
for $m = 0$ **to** MaxIter **do**
  Solve $\{\boldsymbol{x}_i^{m+1}\}_{i=1}^N = \text{argmin}_{\{\boldsymbol{x}_i\}_{i=1}^N} J_m(\{\boldsymbol{x}_i\}_{i=1}^N)$.
  Update $\{\boldsymbol{x}_i^m\}_{i=1}^N$ by $\{\boldsymbol{x}_i^{m+1}\}_{i=1}^N$.
end for

3.2 EVI-GP

We propose two adaptations of the EVI-Im algorithm for GP model estimation and prediction. The first approach, named EVI-post for short, involves generating N particles using Algorithm 1 to approximate the posterior $p(\boldsymbol{\omega}, \tau^2, \eta|\boldsymbol{y}_n)$ when an informative normal prior distribution is adopted for $\boldsymbol{\beta}$, or $p(\boldsymbol{\omega}, \tau^2|\boldsymbol{y}_n)$ when a non-informative prior distribution is used for $\boldsymbol{\beta}$. These N particles serve as posterior samples. Conditional on their values, we can generate samples for $\boldsymbol{\beta}$ based on its conditional posterior distribution (Proposition 1) or generate samples for both $\boldsymbol{\beta}$ and τ^2 according to their conditional posterior distribution (Proposition 2). Following Proposition 3, we can generally predict $y(\boldsymbol{x})$ and confidence intervals conditional on the posterior samples.

In the second approach, we employ EVI-Im solely as an optimization tool for Maximum A Posteriori (MAP), entailing the minimization of $V(\boldsymbol{x}) = -\log \rho^*(\boldsymbol{x})$. So we call it EVI-MAP for short. This can be done by simply setting the free energy as $\mathscr{F}(\boldsymbol{x}) = V(\boldsymbol{x})$ and $N = 1$. As a result, the optimization problem (9) at the ith iteration becomes

$$\boldsymbol{x}^{m+1} = \arg\min_{\boldsymbol{x}} \frac{1}{2\tau}||\boldsymbol{x} - \boldsymbol{x}^m||^2 - \log V(\boldsymbol{x}),$$

which is the celebrated proximal point algorithm (Rockafellar 1976). Therefore, EVI is a method that connects the posterior sampling and MAP under the same general framework. Based on the MAP, we can obtain the posterior mode for $\boldsymbol{\beta}$ (or $\boldsymbol{\beta}$ and τ^2) and the mode of the prediction and the corresponding inference based on the posterior mode.

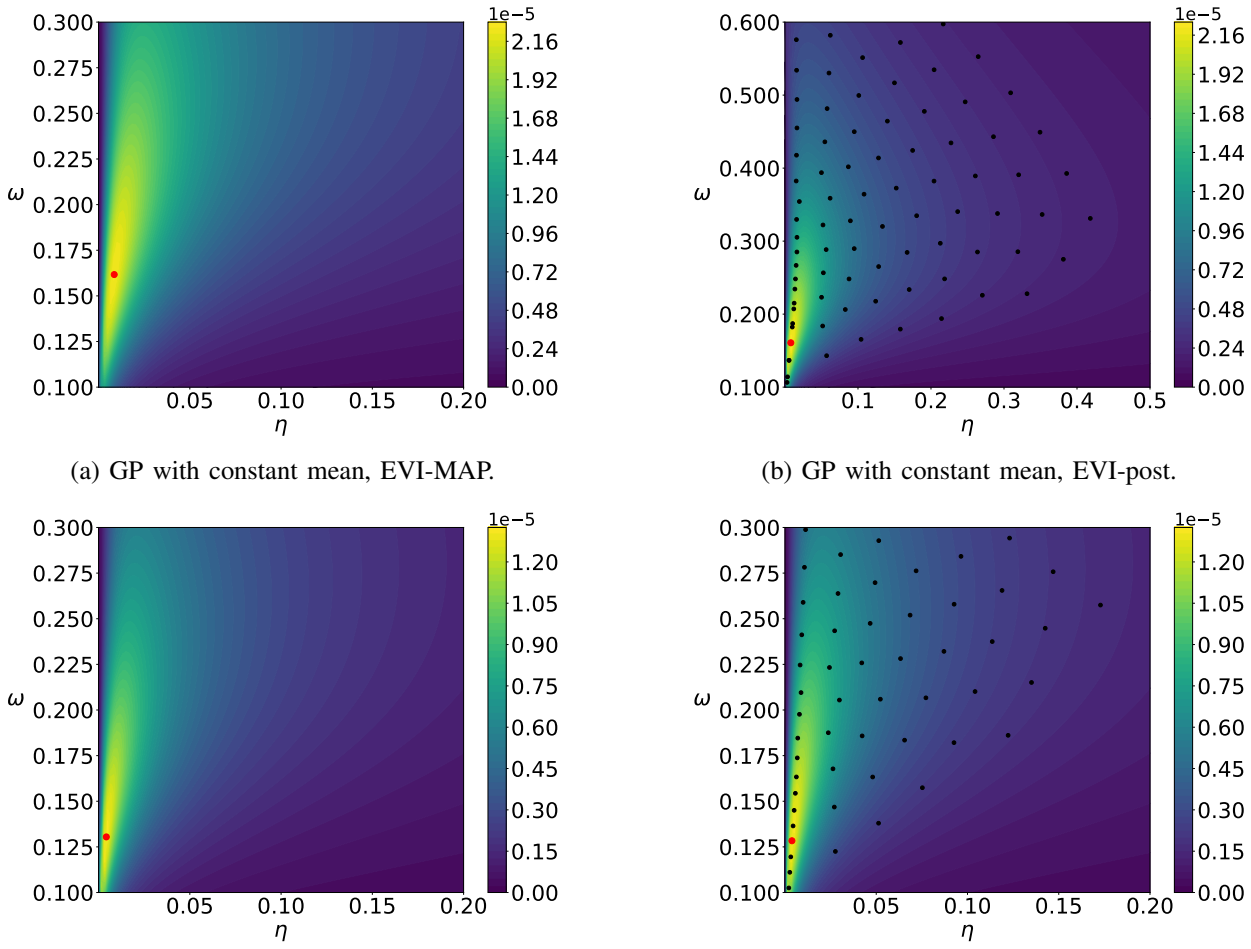
4 NUMERICAL EXAMPLES

In this section, we demonstrate the performances of EVI-GP and compare it with three commonly used GP packages in `R`, which are `gpfit` (MacDonald et al. 2015), `mlegp` (Dancik and Dorman 2008), `laGP` (Gramacy and Apley 2015; Gramacy 2016). Due to limited space, we only show the comparison via two examples. The first one is a 1-dim toy example and the second one is the Borehole example chosen from the online library built by Surjanovic and Bingham (). Readers can refer Kang et al. (2024) for another example of the OTL-circuit simulation. The codes for EVI-GP and all the examples are available on GitHub with the link https://github.com/XavierOwen/EVIGP. The EVI-GP is implemented in `Python`. The proximal point optimization in Algorithm 1 is solved by the LBFGS function of `Pytorch` library (Paszke et al. 2019). Some arguments of the EVI-GP codes are set the same for all examples, which are explained in Kang et al. (2024). These parameter settings are done through many trials and they lead to satisfactory performance in most examples we have studied.

In each example, we use the same pair of training and testing datasets for all the methods under comparison. The designs for both datasets are generated via `maximinLHS` procedure from the `lhs` package in `R` (Carnell 2022). We run 100 simulations. In each simulation, we compute the standardized Root Mean Square Prediction Error (RMSPE) on the test data set, which is defined as follows

$$\text{standardized RMSPE} = \left(\sqrt{\frac{1}{n_{\text{test}}}\sum_i (\hat{y}_i - y_i)^2}\right) \Big/ \text{standard deviation of test}(\boldsymbol{y}),$$

where $\hat{y}_i$ is the predicted value at the test point $\boldsymbol{x}_i$ and y_i is the corresponding true value. Box plots of the 100 standardized RMSPEs of all methods are shown for comparison.



(a) GP with constant mean, EVI-MAP.

(b) GP with constant mean, EVI-post.

(c) GP with linear mean, EVI-MAP.

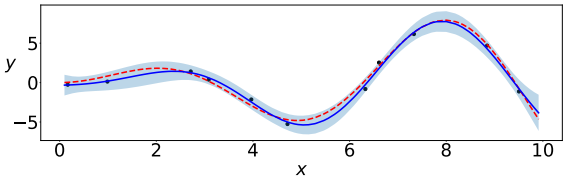
(d) GP with linear mean, EVI-post.

Figure 1: One-Dim Toy Example: in Figure 1a–1d the contours are plotted according to $p(\omega,\eta|\boldsymbol{y}_n)$ evaluated on mesh points without the normalizing constant, the largest red points are the modes of the posterior distribution of (ω,η) returned by EVI-MAP approach, and the black dots in Figure 1b and 1d are the particles returned by EVI-post approach.

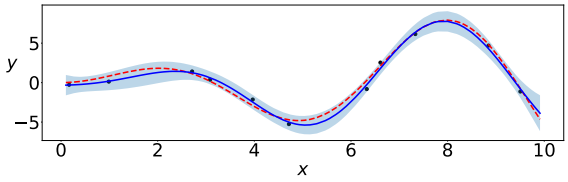
4.1 One-Dim Toy Example

In the one-dim example, the data are generated from the test function $y(x) = x\sin(x) + \varepsilon$ for $x \in [0,10]$ and $\sigma^2 = 0.5^2$. The size of the training and test data sets are $n = 11$ and $m = 100$, respectively.

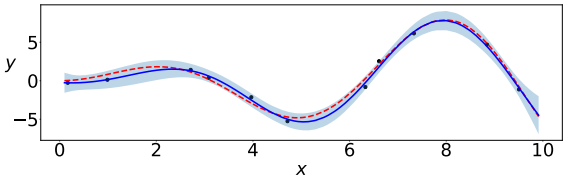
The parameters of the prior distributions of the EVI-GP are $a_\omega = a_\eta = 1$, $b_\omega = b_\eta = 0.5$, and $df_{\tau^2} = 0$ which is equivalent to $p(\tau^2) \propto 1/\tau^2$. We consider two possible mean functions for the GP model, the constant mean $\mu(x) = \beta_0$ and the linear mean $\mu(x) = \beta_0 + \beta_1 x$. There is no need for parameter regularization for these simple mean functions, and thus we use non-informative prior for $\boldsymbol{\beta}$. We use the EVI-post to approximate $p(\omega,\eta|\boldsymbol{y}_n)$ and set the number of particles $N = 100$, kernel bandwidth $h = 0.02$ and stepsize $\tau = 1$ in the EVI procedure. The initial particles are sampled from the uniform distribution in $[0,0.1]\times[0.1,0.4]$. Figure 1 and 2 show the posterior modes and particles returned by EVI-MAP and EVI-post, as well as the



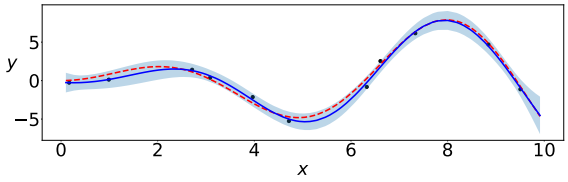
(a) GP Model with constant mean by EVI-MAP.



(b) GP Model with constant mean fitted by EVI-post.



(c) GP Model with linear mean fitted by EVI-MAP.



(d) GP Model with linear mean fitted by EVI-post.

Figure 2: The Gaussian process prediction by four different approaches. Black dots are training data, the red dashed curve is the true test function without noise $y(x) = x\sin(x)$, the blue curve is the predicted curve, and the light blue area shows the 95% predictive confidence interval.

prediction and predictive confidence interval returned by both methods. From Figure 1, we can see that the particles generated from EVI-post well approximate the target distribution $p(\omega, \eta | \boldsymbol{y}_n)$ represented by the contours. The posterior modes are accurately identified by EVI-MAP.

We compare the EVI-GP with three `R` packages. For the `gpfit` package, we set the nugget threshold to be $[0, 25]$, corresponding to η in our method, and the correlation is the Gaussian kernel function. For the `mlegp` package, we set the argument `constMean` to be 0 for the constant mean model and 1 for the linear mean model. The settings of the optimization-related arguments in `mlegp` are carefully chosen for the best performance possible. They are elaborated in Kang et al. (2024). Figure 3 compares the box plots of the RMSPEs from 100 simulations. The prediction accuracy of both versions of the EVI-GP performs almost equally well and significantly outperforms the three `R` packages.

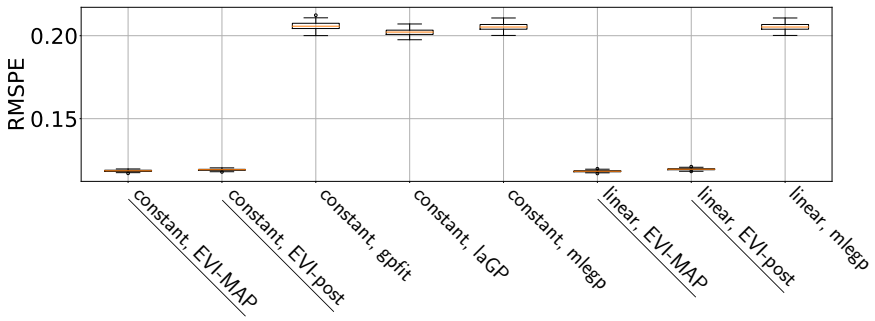


Figure 3: Box plots of the standardized RMSPEs of the toy example with different mean models (constant or linear) and different methods (EVI-MAP, EVI-post, gpfit, laGP, mlegp). The labels of the two EVI-GP methods are underlined.

4.2 Borehole Function

In this subsection, we test the EVI-GP with the famous Borehole function. The definition of the Borehole function and detailed variable definition is included in Kang et al. (2024).

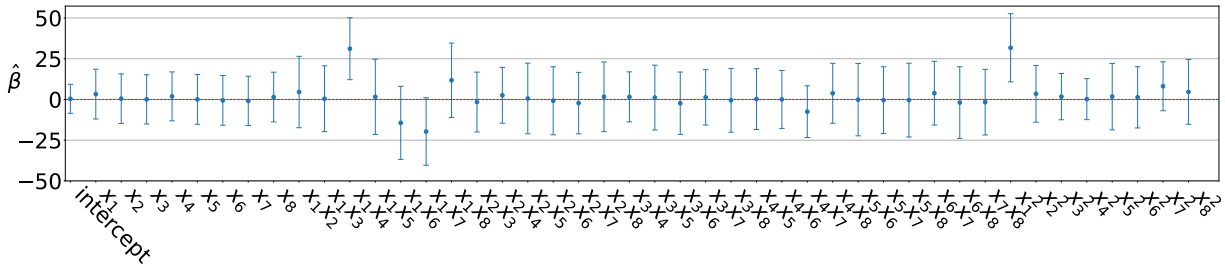


Figure 4: The 95% posterior confidence interval of $\boldsymbol{\beta}$ for the full quadratic mean model.

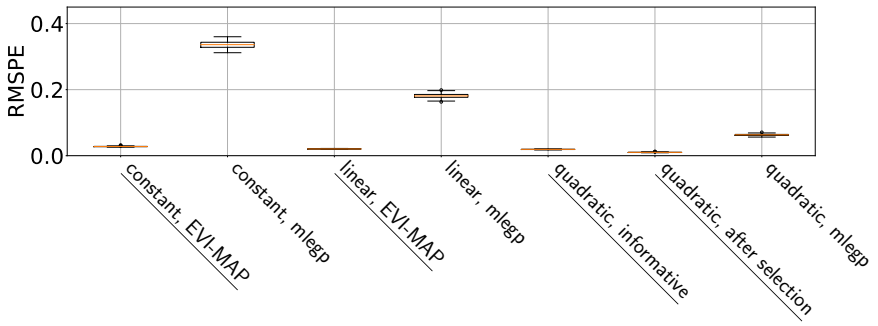


Figure 5: Box plots of standardized RMSPE's of the Borehole Example with different mean models (constant, linear, quadratic, and reduced model) and different methods (EVI-MAP, mlegp). The labels of the EVI-GP methods are underlined.

We use a training dataset of size 200 and 100 testing datasets of size 100. The noise variance is $\sigma^2 = 0.02^2$. For both EVI-post and EVI-MAP, we set $h = 0.001$ and $\tau = 0.1$. For the EVI-post approach, we use $N = 100$ particles, and the initial particles of $(\boldsymbol{\omega}, \eta)$ are sampled uniformly in $[0, 0.1]^7$. The two versions of EVI-GP perform very similarly, so we only return the result of EVI-MAP. The initial mean model of the GP is assumed as a quadratic function of the input variables, including the 2-way interactions. Regarding the variance of the prior distribution of $\boldsymbol{\beta}$, $\nu = 4.55$ was the result of 5-fold cross-validation when the biggest mean model was used. After variable selection, the 5-fold cross-validation is conducted again to find the optimal ν to fit the finalized GP model. Coincidentally, the optimal ν is also 4.55. The initial mean model of the GP is assumed as a quadratic function of the input variables, including the 2-way interactions. Based on the 95% posterior confidence interval in Figure 4 of the $\boldsymbol{\beta}$, we select intercept, x_1, x_4, and $x_1 x_4$ as the significant terms to be kept for the final model. Similarly, we compare the EVI-GP with the `mlegp` package, and the standardized RMSPEs of the 100 simulations are shown in Figure 5. Again, EVI-GP significantly outperforms the `R` package. More importantly, variable selection proves to be essential for the Borehole example, as the final GP model with the reduced mean model (labeled by "quadratic, after selection") returns the smallest standardized RMSPE.

5 CONCLUSION

In this paper, we review the conventional Gaussian process regression model under the Bayesian framework. More importantly, we propose a new variational inference approach, called Energetic Variational Inference, as an alternative to traditional MCMC approaches to estimate and make inferences for the GP regression model. Through comparing with some commonly used `R` packages, the new EVI-GP performs better in terms of prediction accuracy. Although not completely revealed in this paper, the true potential of variational inference lies in transforming the MCMC sampling problem into an optimization problem. As a result, it can be used to solve a more complicated Bayesian framework. For instance, we can adapt the

GP regression and classification by adding fairness constraints on the parameters such that it can meet the ethical requirements in many social and economic contexts. In this case, variational inference can easily solve the constrained optimization whereas it is very challenging to do MCMC sampling with fairness constraints. We are pursuing in this direction in the future work.

ACKNOWLEDGMENTS

The authors' work is partially supported by NSF Grant # 2153029 and # 1916467.

REFERENCES

Ai, M., L. Kang, and V. R. Joseph. 2009. "Bayesian Optimal Blocking of Factorial Designs". *Journal of Statistical Planning and Inference* 139(9):3319–3328 https://doi.org/10.1016/j.jspi.2009.03.008.

Blei, D. M., A. Kucukelbir, and J. D. McAuliffe. 2017. "Variational Inference: a Review for Statisticians". *Journal of the American Statistical Association* 112(518):859–877 https://doi.org/10.1080/01621459.2017.1285773.

Carnell, R. 2022. *lhs: Latin Hypercube Samples*. R package version 1.1.5.

Cheng, C.-A. and B. Boots. 2017. "Variational Inference for Gaussian Process Models with Linear Complexity". In *Proceedings of the 31st International Conference on Neural Information Processing Systems (NIPS 2017)*, edited by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, Volume 30.

Cressie, N. 2015. *Statistics for Spatial Data*. 1st ed. Hoboken, NJ, USA: John Wiley & Sons https://doi.org/10.1002/9781119115151.

Dancik, G. M. and K. S. Dorman. 2008. "mlegp: Statistical Analysis for Computer Models of Biological Systems using R". *Bioinformatics* 24(17):1966–1967 https://doi.org/10.1093/bioinformatics/btn329.

Fang, K.-T., R. Li, and A. Sudjianto. 2005. *Design and Modeling for Computer Experiments*. 1st ed. New York: CRC press https://doi.org/10.1201/9781420034899.

Gelman, A., J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari and D. B. Rubin. 2014. *Bayesian Data Analysis*. 3rd ed. New York: Chapman and Hall/CRC https://doi.org/10.1201/b16018.

Ghanem, R., D. Higdon, and H. Owhadi. 2017. *Handbook of Uncertainty Quantification*. Cham, Switzerland: Springer https://doi.org/10.1007/978-3-319-12385-1.

Gramacy, R. B. 2016. "laGP: Large-Scale Spatial Modeling via Local Approximate Gaussian Processes in R". *Journal of Statistical Software* 72(1):1–46 https://doi.org/10.18637/jss.v072.i01.

Gramacy, R. B. 2020. *Surrogates: Gaussian Process Modeling, Design and Optimization for the Applied Sciences*. Boca Raton, Florida: Chapman Hall/CRC https://doi.org/10.1201/9780367815493. http://bobby.gramacy.com/surrogates/.

Gramacy, R. B. and D. W. Apley. 2015. "Local Gaussian Process Approximation for Large Computer Experiments". *Journal of Computational and Graphical Statistics* 24(2):561–578 https://doi.org/10.1080/10618600.2014.914442.

Gramacy, R. B. and H. Lian. 2012. "Gaussian Process Single-Index Models as Emulators for Computer Experiments". *Technometrics* 54(1):30–41 https://doi.org/10.1080/00401706.2012.650527.

Hamada, M. and C. J. Wu. 1992. "Analysis of Designed Experiments with Complex Aliasing". *Journal of Quality Technology* 24(3):130–137 https://doi.org/10.1080/00224065.1992.11979383.

Hu, Y., R. B. Gramacy, and H. Lian. 2013. "Bayesian Quantile Regression for Single-Index Models". *Statistics and Computing* 23:437–454.

Joseph, V. R. 2006. "A Bayesian Approach to the Design and Analysis of Fractionated Experiments". *Technometrics* 48(2):219–229 https://doi.org/10.1198/004017005000000652.

Kang, L., Y. Cheng, Y. Wang, and C. Liu. 2024. "Energetic Variational Gaussian Process Regression for Computer Experiments". *arXiv* preprint arXiv:1412.698.

Kang, L., X. Deng, and R. Jin. 2023. "Bayesian D-Optimal Design of Experiments with Quantitative and Qualitative Responses". *The New England Journal of Statistics in Data Science* 1(3):371–385 https://doi.org/10.51387/23-NEJSDS30.

Kang, L. and X. Huang. 2019. "Bayesian A-Optimal Design of Experiment with Quantitative and Qualitative Responses". *Journal of Statistical Theory and Practice* 13:1–23 https://doi.org/10.1007/s42519-019-0063-6.

Kang, L. and V. R. Joseph. 2009. "Bayesian Optimal Single Arrays for Robust Parameter Design". *Technometrics* 51(3):250–261 https://doi.org/10.1198/tech.2009.08057.

Kang, L., X. Kang, X. Deng, and R. Jin. 2018. "A Bayesian Hierarchical Model for Quantitative and Qualitative Responses". *Journal of Quality Technology* 50(3):290–308 https://doi.org/10.1080/00224065.2018.1489042.

Kang, X., S. Ranganathan, L. Kang, J. Gohlke and X. Deng. 2021. "Bayesian Auxiliary Variable Model for Birth Records Data with Qualitative and Quantitative Responses". *Journal of Statistical Computation and Simulation* 91(16):3283–3303 https://doi.org/10.1080/00949655.2021.1926459.

MacDonald, B., P. Ranjan, and H. Chipman. 2015. "GPfit: An R Package for Fitting a Gaussian Process Model to Deterministic Simulator Outputs". *Journal of Statistical Software* 64(12):1–23 https://doi.org/10.18637/jss.v064.i12.

Neal, R. M. 1996. *Bayesian Learning for Neural Networks*, Chapter Monte Carlo Implementation, 55–98. New York, NY: Springer New York https://doi.org/10.1007/978-1-4612-0745-0_3.

Paszke, A., S. Gross, F. Massa, A. Lerer and J. Bradbury *et al*. 2019. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Volume 32.

Peng, C.-Y. and C. F. J. Wu. 2014. "On the Choice of Nugget in Kriging Modeling for Deterministic Computer Experiments". *Journal of Computational and Graphical Statistics* 23(1):151–168 https://doi.org/10.1080/10618600.2012.738961.

Rasmussen, C. E. and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. 1st ed. Adaptive Computation and Machine Learning. Cambridge, Massachusett, USA: MIT Press https://doi.org/10.7551/mitpress/3206.001.0001.

Robert, C. P., G. Casella, and G. Casella. 1999. *Monte Carlo Statistical Methods*. 2nd ed, Volume 2. Berlin, Germany: Springer https://doi.org/10.1007/978-1-4757-4145-2.

Roberts, G. O. and J. S. Rosenthal. 1998. "Optimal Scaling of Discrete Approximations to Langevin Diffusions". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60(1):255–268 https://doi.org/10.1111/1467-9868.00123.

Rockafellar, R. T. 1976. "Monotone Operators and the Proximal Point Algorithm". *SIAM Journal on Control and Optimization* 14(5):877–898 https://doi.org/10.1137/0314056.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. "Design and Analysis of Computer Experiments". *Statistical Science* 4(4):409 – 423 https://doi.org/10.1214/ss/1177012413.

Santner, T. J., B. J. Williams, W. I. Notz, and B. J. Williams. 2003. *The Design and Analysis of Computer Experiments*. 2nd ed, Volume 1. New York, NY, USA: Springer https://doi.org/10.1007/978-1-4939-8847-1.

Surjanovic, S. and Bingham, D. Virtual Library of Simulation Experiments: Test Functions and Datasets. http://www.sfu.ca/~ssurjano, accessed 18th September.

Tran, D., R. Ranganath, and D. M. Blei. 2016. "The Variational Gaussian Process". In *Proceedings of the 4th International Conference on Learning Representations (ICLR 2016)*, May 2nd–4th. Caribe Hilton, San Juan, Puerto Rico.

Wang, Y., J. Chen, C. Liu, and L. Kang. 2021. "Particle-Based Energetic Variational Inference". *Statistics and Computing* 31(3):1–17 https://doi.org/10.1007/s11222-021-10009-7.

Wu, C. J. and M. S. Hamada. 2021. *Experiments: Planning, Analysis, and Optimization*. Hoboken, NJ, USA: John Wiley & Sons, Inc https://doi.org/10.1002/9781119470007.

Wynne, G. and V. Wild. 2022, 28–30 Mar. "Variational Gaussian Processes: A Functional Analysis View". In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, edited by G. Camps-Valls, F. J. R. Ruiz, and I. Valera, Volume 151 of *Proceedings of Machine Learning Research*, 4955–4971: PMLR.

AUTHOR BIOGRAPHIES

LULU KANG is an Associate Professor in the Department of Mathematics and Statistics at University of Massachusetts Amherst. Her research interests include Statistical Learning/Machine Learning, Statistical Design of Experiments, Uncertainty Quantification, Bayesian Statistical Modeling, and Optimization. She serves as the an associate editor for *Technometrics* and *SIAM/ASA Journal on Uncertainty Quantification*. Her email address is lulukang@umass.edu and her website is https://sites.google.com/umass.edu/lulukang/home.

YUANXING CHENG is a Ph.D student in the Department of Applied Mathematics at Illinois Institute of Technology in Chicago, IL. Advised by Dr. Lulu Kang and Dr. Chun Liu, he has worked on the topic of uncertainty quantification. His email address is ycheng46@hawk.iit.edu.

YIWEI WANG is an Assistant Professor in the Department of Mathematics at the University of California, Riverside. His research interests include mathematical modeling and scientific computing with applications in physics, material science, biology, and data science. His email address is yiwei.wang@ucr.edu and his website is https://profiles.ucr.edu/app/home/profile/yiweiw.

CHUN LIU is a Professor and Chair of the Department of Applied Mathematics at Illinois Institute of Technology, Chicago, IL. His research interests center around partial differential equations, calculus of variations, and their applications in complex fluids. His e-mail address is cliu124@iit.edu and his website is https://www.iit.edu/directory/people/chun-liu.