Proceedings of the 2019 Winter Simulation Conference
N. Mustafee, K.-H.G. Bae, S. Lazarova-Molnar, M. Rabe, C. Szabo, P. Haas, and Y.-J. Son, eds.

DISTRIBUTED VARIATIONAL INFERENCE-BASED HETEROSCEDASTIC GAUSSIAN PROCESS METAMODELING

Wenjing Wang
Xi Chen

Grado Department of Industrial and Systems Engineering
Virginia Tech
1145 Perry Street
Blacksburg, VA 24061, USA

ABSTRACT

In this paper, we generalize the variational Bayesian inference-based Gaussian process (VBGP) modeling approach for handling large-scale heteroscedastic datasets. VBGP is suitable for simultaneously estimating the underlying mean and variance functions with a single simulation output available at each design point. To improve the scalability of VBGP, we consider building distributed VBGP (DVBGP) models and their hierarchical versions by partitioning a dataset and aggregating individual subset VBGP predictions based on the idea of "transductive combination of GP experts." Numerical evaluations are performed to demonstrate the performance of the DVBGP models from which some insights are derived.

1 INTRODUCTION

Since the seminal work by Sacks et al. (1989), Gaussian process (GP) metamodels (Rasmussen and Williams 2006; Santner et al. 2003) have become popular in various engineering disciplines where they are used to approximate outputs of deterministic computer experiments. With datasets becoming ever larger, research on making metamodels scalable has received increasing attention in recent years. The most notable efforts include online sparse GP models (Damianou et al. 2016; Hensman et al. 2013) that use inducing points to ease computational burden and distributed GP models that use local GP models to fit large datasets (Cao and Fleet 2014; Deisenroth and Ng 2015; Guhaniyogi and Banerjee 2018; Tresp 2000; Liu et al. 2018). However, much of the existing methods can only apply to datasets contaminated by homogeneous random errors. Erroneous conclusions may be reached when these methods are applied to analyze real-life datasets that present strong heteroscedasticity.

Heteroscedastic GP models capable of estimating both the mean and variance functions are of high relevance in practical applications (Lázaro-Gredilla and Titsias 2011; Le et al. 2005; Munoz-Gonzalez et al. 2011; Wang and Neal 2012). However, existing heteroscedastic GP models are typically lacking in scalability. One of the few most recent advances in large-scale heteroscedastic GP modeling is Liu et al. (2018), which involves the use of inducing points and can be inefficient in practice due to the need of optimizing many hyperparameters.

The primary motivation for us to study large-scale heteroscedastic metamodels is slightly different and originates from the context of stochastic simulation metamodeling. We are interested in heteroscedastic metamodels capable of effectively extracting useful information from simulation runs following a "dense and shallow" design. As an initial experimental design, a "dense and shallow" design that has many design points with only a few replications at each, has been shown to facilitate learning from an initial investment of sampling budget to better plan subsequent simulation experiments (Wang and Chen 2018). Such a design tends to produce many design points relative to the input-space dimensionality, presenting the classical computational challenge of standard GP models when handling large datasets.

978-1-7281-3283-9/19/$31.00 ©2019 IEEE

In this paper we propose to construct distributed heteroscedastic metamodels using the variational Bayesian inference-based heteroscedastic GP (VBGP) modeling approach based on the idea of "transductive combination of GP experts" (Cao and Fleet 2015; Cao 2018); existing distributed GP modeling approaches such as Bayesian committee machine (BCM), the product of Gaussian process experts (PoE), and their respective variants all fall into this category (Tresp 2000; Ng and Deisenroth 2014; Cao and Fleet 2014; Deisenroth and Ng 2015). Compared to inducing-point based methods, the benefit of this idea lies in its practicality and efficiency.

The rest of this paper is organized as follows. Section 2 briefly reviews the VBGP modeling approach. Section 3 elaborates on the distributed variational Bayesian inference-based heteroscedastic GP (DVBGP) modeling approach. Section 4 provides some numerical evaluations of the DVBGP models from which some insights are derived.

2 A REVIEW OF THE VARIATIONAL INFERENCE-BASED HETEROSCEDASTIC GAUSSIAN PROCESS MODELING APPROACH

In this section, we provide a brief review of the VBGP modeling approach which is introduced by Lázaro-Gredilla and Titsias (2011). VBGP is suitable for simultaneously estimating the underlying mean and log variance functions with a single simulation replication available at each design point.

Assume that a random simulation output at a point $\mathbf{x} \in \mathscr{X} \subset \mathbb{R}^d$ can be modeled as $y(\mathbf{x}) = f(\mathbf{x}) + \varepsilon(\mathbf{x})$, where $f(\cdot)$ is assumed to be a GP with mean zero and a covariance function $k_f(\mathbf{x},\mathbf{x}';\theta_f)$ with $\mathbf{x},\mathbf{x}' \in \mathscr{X}$, hence written as $f(\cdot) \sim \mathrm{GP}(\mu_0, k_f(\mathbf{x},\mathbf{x}';\theta_f))$; and the observation noise $\varepsilon(\mathbf{x})$ is assumed to be normally distributed with mean zero and variance $\exp(g(\mathbf{x}))$, written as $\varepsilon(\mathbf{x}) \sim \mathscr{N}(0, \exp(g(\mathbf{x})))$, where $g(\cdot) \sim \mathrm{GP}(\mu_0, k_g(\mathbf{x},\mathbf{x}';\theta_g))$. Given the choice of $k_f(\mathbf{x},\mathbf{x}';\theta_f)$ and $k_g(\mathbf{x},\mathbf{x}';\theta_g)$, a VBGP model is fully specified by the parameter vector $\theta = \left(\theta_f^\top, \theta_g^\top, \mu_0\right)^\top$, where θ_f (respectively, θ_g) denotes the hyperparameter vector for k_f (resp., k_g) and μ_0 denotes the mean hyperparameter for $g(\cdot)$. VBGP treats the underlying mean and log variance functions $f(\cdot)$ and $g(\cdot)$ as target functions to estimate.

Given that a single simulation replication is performed at each design point in the design-point set $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$, we obtain the output vector $\mathbf{y} = (y(\mathbf{x}_1), y(\mathbf{x}_2), \ldots, y(\mathbf{x}_n))^\top$. Write the dataset as $\mathscr{D} = \{\mathbf{X}, \mathbf{y}\}$. Define $\mathbf{f} = (f(\mathbf{x}_1), f(\mathbf{x}_2), \ldots, f(\mathbf{x}_n))^\top$ and $\mathbf{g} = (g(\mathbf{x}_1), g(\mathbf{x}_2), \ldots, g(\mathbf{x}_n))^\top$. It follows that the conditional distribution $(\mathbf{y}|\mathbf{f},\mathbf{g})$ is multivariate normal.

To perform parameter estimation and subsequent prediction, VBGP considers approximating the analytically intractable posterior distribution $p(\mathbf{f},\mathbf{g}|\mathbf{y})$ by a factorized approximation $q(\mathbf{f})q(\mathbf{g})$, where $q(\mathbf{f})$ and $q(\mathbf{g})$ respectively denote approximate posterior densities. Based on the Bayes theorem, VBGP further obtains

$$F(q(\mathbf{f}), q(\mathbf{g})) = \log p(\mathbf{y}) - KL(q(\mathbf{f})q(\mathbf{g})||p(\mathbf{f},\mathbf{g}|\mathbf{y})),$$

where $F(\cdot)$ provides a lower bound for the evidence $\log p(\mathbf{y})$ and the second term on the right-hand side denotes the KL divergence between the approximate and the exact posterior distributions. Notice that $\log p(\mathbf{y})$ is fixed given the dataset $\mathscr{D}$. Hence, maximizing the lower bound $F(\cdot)$ is equivalent to minimizing the KL divergence term. Now with a fixed $q(\mathbf{g})$, we can show that the optimal choice of $q(\mathbf{f})$, $q^\star(\mathbf{f}) := \mathrm{argmax}_{q(\mathbf{f})} F(q(\mathbf{f}), q(\mathbf{g}))$, is given by

$$q^\star(\mathbf{f}) = Z(q(\mathbf{g}))^{-1} p(\mathbf{f}) \exp\left(\int q(\mathbf{g}) \log p(\mathbf{y}|\mathbf{f},\mathbf{g}) d\mathbf{g}\right), \tag{1}$$

where $Z(q(\mathbf{g})) := \int \exp\left(q(\mathbf{g}) \log p(\mathbf{y}|\mathbf{f},\mathbf{g}) d\mathbf{g}\right) p(\mathbf{f}) d\mathbf{f}$ is a normalizing constant that only depends on $q(\mathbf{g})$. Hence, we have $\log p(\mathbf{y}) \geq F(q(\mathbf{g})) := F(q^\star(\mathbf{f}), q(\mathbf{g})) \geq F(q(\mathbf{f}), q(\mathbf{g}))$ for any given choice of $q(\mathbf{g})$. Now, if the choice of $q(\mathbf{g})$ is restricted to a multivariate normal family, i.e., $q(\mathbf{g}) = \mathscr{N}(\mathbf{g}|\mu, \Sigma)$, then we can write $F(\mu, \Sigma) := F(q(\mathbf{g})) = \log \mathscr{N}(\mathbf{y}|0, \mathbf{K}_f + \mathbf{R}) - 0.25\mathrm{tr}(\Sigma) - KL(\mathscr{N}(\mathbf{g}|\mu, \Sigma)||\mathscr{N}(\mathbf{g}|\mu_0\mathbf{1}, K_g))$, where $\mathbf{K}_f$ and $\mathbf{K}_g$ are $n \times n$ covariance matrices resulting from evaluating the corresponding covariance functions at $\mathbf{X}$, and $\mathbf{R}$ is the $n \times n$ diagonal matrix with diagonal elements given by $[\mathbf{R}]_{ii} = \exp([\mu]_i - [\Sigma]_{ii}/2)$, $i = 1, 2, \ldots, n$. In light of (1) and that $q(\mathbf{g}) = \mathscr{N}(g|\mu, \Sigma)$, we can show that $q^\star(\mathbf{f})$ is multivariate normally distributed as $\mathscr{N}(\mathbf{f}|\mathbf{K}_f(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{y}, \mathbf{K}_f - \mathbf{K}_f(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{K}_f)$. Hence, the task of seeking the optimal approximate posterior distribution reduces to finding μ and Σ that parameterize $q(\mathbf{g})$ to maximize $F(q^\star(\mathbf{f}), q(\mathbf{g}))$. By setting $\partial F(\mu, \Sigma)/\partial \mu = 0$ and $\partial F(\mu, \Sigma)/\partial \Sigma = 0$, we can obtain that $\mu = \mathbf{K}_g(\Lambda - 1/2\mathbf{I})\mathbf{1} + \mu_0 \mathbf{1}$, $\Sigma^{-1} = \mathbf{K}_g^{-1} + \Lambda$, where Λ is an $n \times n$ positive definite diagonal matrix. Finally, estimation of parameters in θ and Λ can be carried out by optimizing $F(\cdot)$ numerically; see Wang et al. (2019) and Wang (2019) for more details.

Regarding predicting the mean and log variance function values at prediction point $\mathbf{x}_\star \subseteq \mathscr{X}$, it can be shown that the predictive distribution for $f(\mathbf{x}_\star)$ is normal, with its mean and variance given by $\mu_f(\mathbf{x}_\star) = \mathbf{k}_f(\mathbf{X}, \mathbf{x}_\star)^\top(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{y}$ and $\sigma_f^2(\mathbf{x}_\star) = \mathbf{k}_f(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_f(\mathbf{X}, \mathbf{x}_\star)^\top(\mathbf{K}_f + \mathbf{R})^{-1}\mathbf{k}_f(\mathbf{X}, \mathbf{x}_\star)$, respectively. Similarly, the predictive distribution for $g(\mathbf{x}_\star)$ is normal, and its mean and variance are given by $\mu_g(\mathbf{x}_\star) = \mathbf{k}_g(\mathbf{X}, \mathbf{x}_\star)^\top(\Lambda - \frac{1}{2}\mathbf{I})\mathbf{1} + \mu_0$ and $\sigma_g^2(\mathbf{x}_\star) = \mathbf{k}_g(\mathbf{x}_\star, \mathbf{x}_\star) - \mathbf{k}_g(\mathbf{X}, \mathbf{x}_\star)^\top(\mathbf{K}_g + \Lambda^{-1})^{-1}\mathbf{k}_g(\mathbf{X}, \mathbf{x}_\star)$, respectively.

Wang et al. (2019) generalize VBGP to accommodate the setting where either one or multiple replications are available at each design point, and further prove that the computational complexity only depends on the number of design points used. The VBGP model reviewed in this section can be regarded as a special case of that studied by Wang et al. (2019).

3 THE DISTRIBUTED VARIATIONAL INFERENCE-BASED HETEROSCEDASTIC GAUSSIAN PROCESS MODELING APPROACH

In this section, we elaborate on the DVBVP metamodeling approach. The idea behind DVBGP is to "divide and conquer" a given dataset. Specifically, we first partition the full dataset $\mathscr{D}$ into M disjoint subsets $\mathscr{D}_i = \{\mathbf{X}^{(i)}, \mathbf{y}^{(i)}\}$, $i = 1, 2, \ldots, M$, with $\mathbf{X}^{(i)}$ and $\mathbf{y}^{(i)}$ denoting the design-point set and the corresponding set of noisy observations associated with the ith subset, respectively. We apply VBGP to approximate the mean and log variance functions for each subset in parallel on different computing units, and obtain M subset predictive distributions. The predictive results obtained from the subsets are then integrated to provide aggregated VBGP predictions such that the information of the full dataset $\mathscr{D}$ is maintained to a great extent.

3.1 Parameter Estimation

To facilitate parameter estimation for a DVBGP model, we introduce a subset independence assumption and use the following approximation of the log model evidence:

$$\log p(\mathbf{y}; \mathbf{X}, \theta) \approx \sum_{i=1}^{M} \log p(\mathbf{y}^{(i)}; \mathbf{X}^{(i)}, \theta_i) \geq \sum_{i=1}^{M} F_i$$

where θ denotes the hyperparameter vector for a VBGP model built on the full dataset $\mathscr{D}$, and θ_i and F_i respectively denote the parameter vector and the variational lower bound for the ith VBGP model built on $\mathscr{D}_i$. Hence, parameter estimation for a DVBGP model can be conveniently translated into that for M individual VBGP models, which can be carried out in parallel on M individual machines.

3.2 Prediction Aggregation

With the estimated parameters, we are now ready to obtain the posterior predictive distributions for f and g for each subset of data. For the purpose of aggregating the predictions given by the individual VBGP models into an ultimate prediction for the DVBGP model, several methods in the category of transductive combination of GP experts become especially handy (Tresp 2000; Ng and Deisenroth 2014; Cao and Fleet 2014; Deisenroth and Ng 2015). In this section, we focus on the generalized robust Bayesian committee machine (gRBCM) and the generalized product of GP experts (gPoE), which are considered the state-of-the-art methods in this category. Furthermore, we provide their respective hierarchical versions.

3.2.1 Generalized Robust Bayesian Committee Machine

The Bayesian committee machine (BCM) is the first well-known method in the category of transductive combination of GP models (Tresp 2000). Suppose that the target function to estimate is $m(\mathbf{x})$ with $\mathbf{x} \in \mathscr{X} \subseteq \mathbb{R}^d$. Given that the full dataset $\mathscr{D}$ is partitioned into M disjoint subsets of observations, BCM derives its aggregated predictive distribution for m by assuming that when conditioned on the prediction point $\mathbf{x}_\star$ and $m(\mathbf{x}_\star)$, the subsets of data become independent. Below we start with a brief introduction of robust Bayesian committee machine (RBCM), one of BCM's most recent variants, transform it to the generalized RBCM (gRBCM), and then build a DVBGP model based on gRBCM.

Robust Bayesian Committee Machine Sharing the same theoretical basis as BCM, an RBCM assigns each individual predictive distribution some weight β_i as a measure of reliability when forming the aggregated predictive distribution of the target function m. We have the following result for RBCM.

Proposition 1 The RBCM's aggregated predictive distribution is in the following form:

$$p(m|\mathscr{D},\mathbf{x}_\star) \propto \prod_{i=1}^{M} [p_i(m|\mathscr{D}_i,\mathbf{x}_\star)]^{\beta_i(\mathbf{x}_\star)} \, [p(m|\mathbf{x}_\star)]^{1-\sum_{i=1}^{M}\beta_i(\mathbf{x}_\star)},$$

which is a normal distribution with the predictive mean and variance at $\mathbf{x}_\star$ respectively given by $\mu_{\text{rbcm}}(\mathbf{x}_\star) = \sigma^2_{\text{rbcm}}(\mathbf{x}_\star)\sum_{i=1}^{M}\beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star)\mu_i(\mathbf{x}_\star)$ and $\sigma^2_{\text{rbcm}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M}\beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star) + (1-\sum_{i=1}^{M}\beta_i(\mathbf{x}_\star))\sigma_{\star\star}^{-2}\right)^{-1}$, where p_i denotes the predictive distribution of m given by the ith GP model and p denotes the prior distribution of m, with $\sigma^2_{\star\star}$ being the prior variance. The parameter $\beta_i(\mathbf{x}_\star) := 0.5(\log\sigma^2_{\star\star} - \log\sigma_i^2(\mathbf{x}_\star))$ denotes the weight assigned to the ith GP model, $i = 1,2,\ldots,M$.

The idea of RBCM is first proposed by Deisenroth and Ng (2015). The proof of Proposition 1 is detailed in Wang (2019). We note that BCM can be regarded as a special case of RBCM where $\beta_i(\mathbf{x}_\star) = 1$ for any $\mathbf{x}_\star$, $i = 1,2,\ldots,M$. Next we extend RBCM to gRBCM and build a DVBGP model based on it.

Generalized Robust Bayesian Committee Machine With a partition of the dataset $\mathscr{D}$ into disjoint subsets based on which individual GP models are built, some information about the structure of the entire dataset $\mathscr{D}$ is lost, but nevertheless would be captured by the cross-covariance terms in a full GP model. Aiming at replicating this behavior of a full model by a distributed version, we intend to alleviate the effects of the conditional independence assumption stipulated by BCM. One way to achieve this goal is to use a gRBCM. gRBCM is similar to RBCM, with the difference being the use of a communication set, by which we hope to account for the covariance effects between the points in $\mathscr{D}$ to some extent. The idea behind gRBCM was first discussed by Ng and Deisenroth (2014) and later formally studied by Liu et al. (2018), both of which apply it to GP modeling of large homoscedastic datasets.

Specifically, among the M disjoint subsets, gRBCM requires one of them to be the communication set (say, the Mth subset). gRBCM constructs $(M-1)$ local GP models on the $(M-1)$ subsets $\mathscr{D}_1,\mathscr{D}_2,\ldots,\mathscr{D}_{M-1}$ as well as a communication GP model on the communication set $\mathscr{D}_c$. In particular, the following assumptions on the M subsets are stipulated by gRBCM:

- The communication set $\mathscr{D}_c$ is a subset of points that are randomly selected from $\mathscr{D}$ without replacement, hence the design-point locations spread across the entire support $\mathscr{X}$ and this subset can capture the main features of the target function $m(\cdot)$. The remaining $M-1$ subsets $\mathscr{D}_1,\mathscr{D}_2,\ldots,\mathscr{D}_{M-1}$ are not required to be formed by random sampling from $\mathscr{D}$.
- Given the communication set $\mathscr{D}_c$, $\mathbf{x}_\star$, and $m(\mathbf{x}_\star)$, the independence of subsets $\mathscr{D}_i$ and $\mathscr{D}_j$ holds for $1 \le i \ne j \le M-1$.

gRBCM further combines the communication set $\mathscr{D}_c$ with each of the $(M-1)$ remaining subsets to form the augmented subsets $\mathscr{D}_{+i} := \mathscr{D}_c \cup \mathscr{D}_i$ for $i = 1,2,\ldots,M-1$, based on which enhanced local GP models with improved predictive performance are constructed. Figure 1(a) illustrates the structure of a one-layer gRBCM aggregation model.

Given the aforementioned assumptions of gRBCM and the prediction weights, we can derive the following results for gRBCM.

Proposition 2 The gRBCM's aggregated predictive distribution takes the following form:

$$p(m|\mathscr{D},\mathbf{x}_\star) \propto \prod_{i=1}^{M-1} [p(m|\mathscr{D}_{+i},\mathbf{x}_\star)]^{\beta_i(\mathbf{x}_\star)} [p(m|\mathbf{x}_\star)]^{\beta_c(\mathbf{x}_\star)},$$

which is a normal distribution with the predictive mean and variance at $\mathbf{x}_\star$ respectively given by

$$\mu_{\text{grbcm}}(\mathbf{x}_\star) = \sigma^2_{\text{grbcm}}(\mathbf{x}_\star)\left(\sum_{i=1}^{M-1}\beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star)\mu_i(\mathbf{x}_\star) + \beta_c(\mathbf{x}_\star)\sigma_c^{-2}(\mathbf{x}_\star)\mu_c(\mathbf{x}_\star)\right), \tag{2}$$

$$\sigma^2_{\text{grbcm}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M-1}\beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star) + \beta_c(\mathbf{x}_\star)\sigma_c^{-2}(\mathbf{x}_\star)\right)^{-1}, \tag{3}$$

where $\sigma_c^2(\mathbf{x}_\star)$ and $\mu_c(\mathbf{x}_\star)$ are respectively the predictive variance and mean given by the communication GP model. The weights for the enhanced local GP models are given by $\beta_1(\mathbf{x}_\star) = 1$ and $\beta_i(\mathbf{x}_\star) := 0.5(\log\sigma_c^2(\mathbf{x}_\star) - \log\sigma_i^2(\mathbf{x}_\star))$ for $i = 2,\dots,M-1$; and the weight for the communication GP model is given by $\beta_c(\mathbf{x}_\star) = 1 - \sum_{i=1}^{M-1}\beta_i(\mathbf{x}_\star)$.

The proof of Proposition 2 is given in Wang (2019). In light of this result, we can construct separate approximations to the mean and log variance functions. Specifically, for estimating the mean function f (respectively, the log variance function g), we can treat f as the target function m and apply (2) and (3) to obtain the predictive mean and variance for f (resp., g). Therefore, we have that the aggregated predictive distribution for $f(\mathbf{x}_\star)$ is normal, with its mean and variance respectively given by

$$\mu_{f,\text{grbcm}}(\mathbf{x}_\star) = \sigma^2_{f,\text{grbcm}}(\mathbf{x}_\star)\left(\sum_{i=1}^{M-1}\beta_i^f(\mathbf{x}_\star)\sigma_{f_i}^{-2}(\mathbf{x}_\star)\mu_{f_i}(\mathbf{x}_\star)\right), \tag{4}$$

$$\sigma^2_{f,\text{grbcm}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M-1}\beta_i^f(\mathbf{x}_\star)\sigma_{f_i}^{-2}(\mathbf{x}_\star) + (1-\sum_{i=1}^{M-1}\beta_i^f(\mathbf{x}_\star))\sigma_{c,f}^{-2}(\mathbf{x}_\star)\right)^{-1}, \tag{5}$$

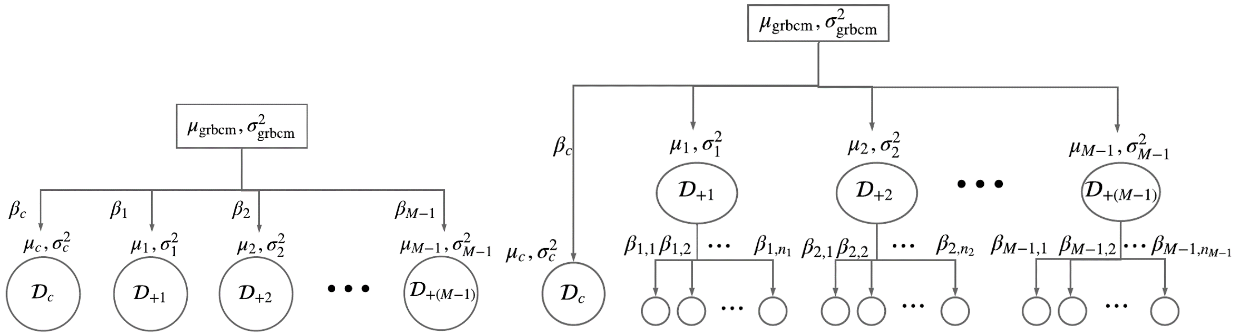
where $\sigma^2_{c,f}(\mathbf{x}_\star)$ is the predictive variance of $f(\mathbf{x}_\star)$ resulting from the communication set. Similarly, the aggregated predictive distribution for $g(\mathbf{x}_\star)$ is normal, with its mean and variance respectively given by

$$\mu_{g,\text{grbcm}}(\mathbf{x}_\star) = \sigma^2_{g,\text{grbcm}}(\mathbf{x}_\star)\left(\sum_{i=1}^{M-1}\beta_i^g(\mathbf{x}_\star)\sigma_{g_i}^{-2}(\mathbf{x}_\star)\mu_{g_i}(\mathbf{x}_\star) + (1-\sum_{i=1}^{M-1}\beta_i^g(\mathbf{x}_\star))\sigma_{c,g}^{-2}(\mathbf{x}_\star)\mu_0\right), \tag{6}$$

$$\sigma^2_{g,\text{grbcm}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M-1}\beta_i^g(\mathbf{x}_\star)\sigma_{g_i}^{-2}(\mathbf{x}_\star) + (1-\sum_{i=1}^{M-1}\beta_i^g(\mathbf{x}_\star))\sigma_{c,g}^{-2}(\mathbf{x}_\star)\right)^{-1}, \tag{7}$$

with $\sigma^2_{c,g}(\mathbf{x}_\star)$ being defined similarly as $\sigma^2_{c,f}(\mathbf{x}_\star)$.

Hierarchical Generalized Robust Bayesian Committee Machine It is well known to parallel computing researchers that passing messages among a large number of machines to a central machine is expensive, regardless of the size of the messages being passed (Scott et al. 2016). Inspired by Ng and Deisenroth (2014), below we consider a hierarchical structure of the gRBCM aggregation model built on individual GP models. This enables us to distribute the computational load over a large number of independent computational units and recursively recombine computations by independent units into a gRBCM aggregation model.



(a) A one-layer gRBCM

(b) A two-layer gRBCM

Figure 1: An illustration of possible structures of a gRBCM model: (a) a one-layer gRBCM and (b) a two-layer gRBCM.

Figure 1(b) illustrates the structure of a two-layer gRBCM aggregation model. To lighten notation, we do not make the dependence on prediction point $\mathbf{x}_\star$ explicit below. Denote the mean and variance of this model by μ_{grbcm} and σ^2_{grbcm}. We have M sub-GPs built on the subsets $\mathscr{D}_{+1}, \mathscr{D}_{+2}, \ldots, \mathscr{D}_{+(M-1)}$ and the communication set $\mathscr{D}_c$, with their corresponding pairs of predictive mean and variance given by $(\mu_1, \sigma_1^2), (\mu_2, \sigma_2^2), \ldots, (\mu_{M-1}, \sigma_{M-1}^2)$, and (μ_c, σ_c^2). We can further divide each sub-GP into n_i subsub-GPs, for each $i = 1, 2, \ldots, M-1$. Denote the mean and variance of subsub-GPs as $\mu_{i,j}$ and $\sigma^2_{i,j}$ respectively; Denote the weights of subsub-GPs as $\beta_{i,j}$, $j = 1, 2, \ldots, n_i$, $i = 1, 2, \ldots, M-1$. In addition, the weight for the communication GP, β_c, is defined similarly as for RBCM, namely, one minus the sum of all weights for the subsub-GPs.

The following result discloses the relationship between the one-layer gRBCM and a hierarchical gRBCM. It is worth noting that the predictive distribution given by aggregating sub-GPs in a single layer is equivalent to that given by aggregating subsub-GPs in a multilayer structure, regardless of the number of layers used. Therefore, in theory it does not matter whether a shallow or deep structure is chosen so long as the number of subsets of data is fixed. Nevertheless, a hierarchical gRBCM can make the computation more efficient, rendering full utilization of distributed computing units.

Proposition 3 The two-layer gRBCM's aggregated predictive distribution is in the following form:

$$p(m|\mathscr{D}, \mathbf{x}_\star) \propto \prod_{i=1}^{M-1} \prod_{j=1}^{n_i} \left[p_{i,j}(m|\mathscr{D}_{+i}, \mathbf{x}_\star)\right]^{\beta_{i,j}(\mathbf{x}_\star)} \left[p(m|\mathbf{x}_\star)\right]^{\beta_c(\mathbf{x}_\star)},$$

which is a normal distribution with the predictive mean and variance at $\mathbf{x}_\star$ respectively given by

$$\mu_{\text{grbcm}}(\mathbf{x}_\star) = \sigma^2_{\text{grbcm}}(\mathbf{x}_\star) \left(\sum_{i=1}^{M-1} \sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star) \sigma^{-2}_{i,j}(\mathbf{x}_\star) \mu_{i,j}(\mathbf{x}_\star) + \beta_c(\mathbf{x}_\star) \mu_c(\mathbf{x}_\star) \sigma_c^{-2}(\mathbf{x}_\star) \right),$$

$$\sigma^2_{\text{grbcm}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M-1} \sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star) \sigma^{-2}_{i,j}(\mathbf{x}_\star) + \beta_c(\mathbf{x}_\star) \sigma_c^{-2}(\mathbf{x}_\star) \right)^{-1},$$

where $\sigma_c^2(\mathbf{x}_\star)$ and $\mu_c(\mathbf{x}_\star)$ are respectively the predictive variance and mean given by the communication GP model; $\beta_{i,j}(\mathbf{x}_\star) := 0.5(\log \sigma_c^2(\mathbf{x}_\star) - \log \sigma^2_{i,j}(\mathbf{x}_\star))$ for $i = 1, \ldots, M-1, j = 1, \ldots, n_i$ denotes the weight for the (i,j)th subsub-GP model, and the weight for the communication GP model is given by $\beta_c(\mathbf{x}_\star) = 1 - \sum_{i=1}^{M-1} \sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star)$. This two-layer gRBCM gives the same predictive mean and variance as the one-layer gRBCM, whose ith subset predictive mean and variance are given by $\mu_i(\mathbf{x}_\star) := \sigma_i^2(\mathbf{x}_\star) \left(\sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star) \sigma^{-2}_{i,j}(\mathbf{x}_\star) \mu_{i,j}(\mathbf{x}_\star) \right)$ and $\sigma_i^2(\mathbf{x}_\star) := \left(\sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star) \sigma^{-2}_{i,j}(\mathbf{x}_\star) \right)^{-1}$, $i = 1, 2, \ldots, M-1$.

The proof of Proposition 3 is given in Wang (2019). In fact, we note that this result can be generalized to a hierarchical gRBCM with an arbitrary number of layers. In light of Proposition 3, a hierarchical DVBGP can be constructed similarly as the DVBGP built on a one-layer gRBCM. In particular, we mention without showing details that the corresponding predictive mean and variance for f and g in forms similar to (4)–(7) can be given subsequently.

3.2.2 Generalized Product of Gaussian Process Experts

In this subsection, we elaborate on generalized product of GP experts (gPoE), which has a slightly different theoretical basis than BCM and its variants such as RBCM and gRBCM.

First proposed by Cao and Fleet (2014), gPoE finds its roots in Heskes (1998) that provides its basis in the transduction log opinion pool framework. Assume that there are M GP models (i.e., experts), each of which produces an individual predictive distribution for the target function value $m(\mathbf{x}_\star)$, $p_i := p_i(m_\star|\mathbf{x}_\star, \mathscr{D}_i)$, at prediction point $\mathbf{x}_\star$. Also, assume that each expert has a measure of relative reliability $\beta_i(\mathbf{x}_\star) \geq 0$ at $\mathbf{x}_\star$ such that $\sum_{i=1}^M \beta_i(\mathbf{x}_\star) = 1$. gPoE intends to find its aggregated predictive distribution via solving the following optimization problem:

$$p_{\text{gpoe}}(m|\mathbf{x}_\star) = \underset{p}{\operatorname{argmin}} \sum_{i=1}^{M} \beta_i(\mathbf{x}_\star) KL(p||p_i).$$

Under the constraint that $\int p(m|\mathbf{x}_\star)dm = 1$ for all possible distributions, Heskes (1998) shows that the solution is in the form of $p_{gpoe}(m|\mathbf{x}_\star) \propto \prod_{i=1}^M [p_i(m|\mathbf{x}_\star)]^{\beta_i(\mathbf{x}_\star)}$, which is exactly a product of each expert's predictive distribution raised to power $\beta_i(\mathbf{x}_\star)$. Notice that when the weights $\beta_i(\mathbf{x}_\star)$ sum up to 1, the gPoE aggregated predictive distribution is just a weighted geometric average of individual predictive distributions.

The aforementioned framework provides a sound theoretical basis for gPoE, and it sheds light on the following properties of gPoE. First, unlike BCM and its variants, gPoE does not stipulate any form of conditional independence assumption; hence, the subsets $\mathscr{D}_i$ can be formed arbitrarily. Furthermore, the individual GP models can have different hyperparameters or even distinct covariance function specifications.

The gPoE's aggregated predictive distribution is normal, with its predictive mean and variance respectively given by

$$\mu_{\text{gpoe}}(\mathbf{x}_\star) = \sigma^2_{\text{gpoe}}(\mathbf{x}_\star) \sum_{i=1}^{M} \beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star)\mu_i(\mathbf{x}_\star), \quad \sigma^2_{\text{gpoe}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M} \beta_i(\mathbf{x}_\star)\sigma_i^{-2}(\mathbf{x}_\star)\right)^{-1}.$$

Hierarchical Generalized Product of Gaussian Process Experts Similar to hierarchical gRBCM, we can construct a hierarchical gPoE model. The following result in the same vein as Proposition 3 can be established and its proof is given in Wang (2019).

Proposition 4 The two-layer gPoE's aggregated predictive distribution is in the following form:

$$p(m|\mathscr{D}, \mathbf{x}_\star) \propto \prod_{i=1}^{M} \prod_{j=1}^{n_i} [p_{i,j}(m|\mathscr{D}_{i,j}, \mathbf{x}_\star)]^{\beta_{i,j}(\mathbf{x}_\star)},$$

which is a normal distribution with the predictive mean and variance at $\mathbf{x}_\star$ respectively given by

$$\mu_{\text{gpoe}}(\mathbf{x}_\star) = \sigma^2_{\text{gpoe}}(\mathbf{x}_\star) \sum_{i=1}^{M} \sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star)\sigma_{i,j}^{-2}(\mathbf{x}_\star)\mu_{i,j}(\mathbf{x}_\star), \quad \sigma^2_{\text{gpoe}}(\mathbf{x}_\star) = \left(\sum_{i=1}^{M} \sum_{j=1}^{n_i} \beta_{i,j}(\mathbf{x}_\star)\sigma_{i,j}^{-2}(\mathbf{x}_\star)\right)^{-1}.$$

Selection of Weights for gPoE The transduction log opinion pool framework for gPoE enables a more in-depth investigation of the selection of weights as compared to BCM and its variants.

Suppose that at prediction point $\mathbf{x}_\star$ the unknown true underlying distribution is $p_\star(m|\mathbf{x}_\star)$. To lighten notation, we do not make the dependence on prediction point $\mathbf{x}_\star$ explicit below. It is shown in Heskes (1998) that the KL divergence between $p_\star$ and the gPoE estimate p_{gpoe} can be approximated by

$$KL(p_\star||p_{gpoe}) \approx \underbrace{\sum_{i=1}^{M} \beta_i KL(p_\star||p_i)}_{A} - \underbrace{\frac{1}{4}\sum_{i=1}^{M}\sum_{j=1}^{M} \beta_i\beta_j \left(KL(p_i||p_j)+KL(p_j||p_i)\right)}_{B}$$

Reducing $KL(p_\star||p_{gpoe})$ requires to decrease A and increase B. Reducing A requires to set higher weights β_i for individual models that provide good predictions at $\mathbf{x}_\star$ (i.e., with small $KL(p_\star||p_i)$), and vice versa. Because $p_\star$ is unknown, one cannot compute the terms $KL(p_\star||p_i)$. Setting β_i as the entropy change (i.e., $\beta_i = 0.5(\log\sigma^2_{\star\star} - \log\sigma^2_i(\mathbf{x}_\star)))$ lowers weights for experts whose prediction are not significantly influenced by data (or equivalently, being uninformative for prediction at $\mathbf{x}_\star$ and hence having a high value for $KL(p_\star||p_i)$). Therefore, such a weight selection helps decrease A. Increasing B requires to increase weights for pairs of GP models that present distinct predictions at $\mathbf{x}_\star$, namely, it encourages diversified GP models. One way to incorporate B into determining the weights is to initialize the weights at $\beta_i = 0.5(\log\sigma^2_{\star\star} - \log\sigma^2_i(\mathbf{x}_\star))$ and normalize the values to sum up to 1; then take a normalized gradient step of B to encourage more diversified GP model predictions. This simple and effective updating process can be iterated a few times, but it has been reported that one iteration works sufficiently well (Cao and Fleet 2015; Cao 2018).

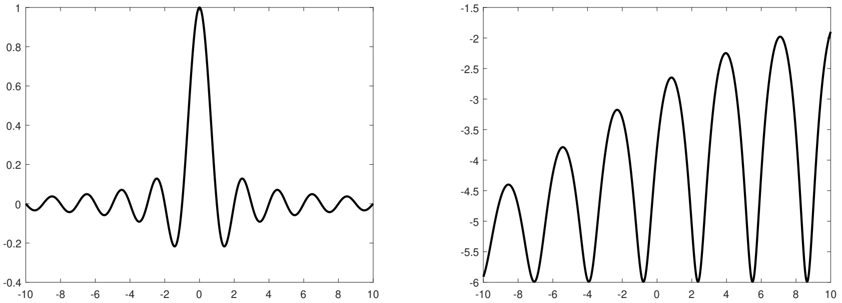
Finally, we note that a DVBGP model and its hierarchical version can be constructed based on gPoE and hierarchical gPoE. We mention without showing details that the resulting predictive mean and variance for f and g in forms similar to (4)–(7) can be obtained accordingly.

4 NUMERICAL EVALUATION

In this section, we numerically evaluate the performance of the DVBGP approach via a relatively challenging one-dimensional example. A random output $y(x)$ at $x \in \mathscr{X} = [-10, 10]$ is generated as follows:

$$y(x) = \mathrm{sinc}(x) + \varepsilon(x),$$

where the simulation noise $\varepsilon(x)$ is assumed to follow a normal distribution with mean zero and variance $\sigma^2_\varepsilon(x)$, with $\sigma_\varepsilon(x) = 0.05 + 0.2(1+\sin(2x))/(1+e^{-0.2x})$. We are interested in estimating the mean function $f(x) := \mathrm{sinc}(x)$ and the log variance function $g(x) := \log(\sigma^2_\varepsilon(x))$ with $x \in \mathscr{X}$ as shown in Figure 2 respectively.



(a) The mean function $f(\cdot)$

(b) The log variance function $g(\cdot)$

Figure 2: The mean and log variance functions to estimate in the one-dimensional example.

A simulation experiment is performed with a total budget of C observations, each being an output at a design point randomly sampled from $\mathscr{X}$. That is, we consider a single observation at each design point in this example. Given the simulation output dataset $\mathscr{D}$ comprising C observations, we examine the impact of subset partition schemes that are used to divide $\mathscr{D}$ into M disjoint subsets. Specifically, k-means partition (e.g., see Chapter 13 of Hastie et al. (2009)) and pure random partition are considered. We further investigate the effect of combining $M-1$ disjoint subsets formed by k-means with a communication set, with this set comprising $\lfloor C/M \rfloor$ observations randomly selected from $\mathscr{D}$.

We vary the total budget $C \in \{500, 5000, 10^4\}$ and the number of subsets $M \in \{5, 8, 10\}$. Under each parameter setting, the simulation experiment is repeated for $L = 25$ independent macro-replications. We are interested in comparing the predictive performance of the DVBGP models respectively built on gRBCM, gPoE, gPoE with weight selection (respectively referred to as "gRBCM," "gPoE," and "gPoE-w" for short hereinafter) with the benchmark approach—a single VBGP model (referred to as "VBGP"), in terms of the empirical root mean squared error (ERMSE) and the mean coverage probability (MCP). A set of $H = 1000$ equispaced tests points in $\mathscr{X}$ is generated for performance evaluation. The ERMSE and MCP achieved by a given method on each macro-replication are respectively calculated as follows:

$$\text{ERMSE}_\ell = \sqrt{\frac{1}{H}\sum_{i=1}^{H}\left(\mu_\ell(x_i) - m(x_i)\right)^2} \ \text{ and } \ \text{MCP}_\ell = \frac{1}{H}\sum_{i=1}^{H}\mathbf{1}\left\{m(x_i) \text{ is in the 95\% prediction interval}\right\},$$

where $\mu_\ell(\cdot)$ denotes the predictive mean value for the target function value $m(\cdot)$ obtained on the ℓth macro-replication ($\ell = 1, 2, \ldots, L$). The 95% predictive interval is constructed using the predictive mean and variance of $m(\cdot)$. Notice that $m(\cdot)$ denotes the mean function (i.e., $f(\cdot)$) and the log variance function (i.e., $g(\cdot)$) of interest. To economize on space, we only show the results obtained with $M = 5$ and 10.

Summary of results. Regarding predictive accuracy, we make the following observations from Tables 1 and 3. First, increasing the total budget C helps improve the predictive accuracies achieved by all methods considered. With a fixed budget C given, increasing the number of disjoint subsets M used by each DVBGP model leads to deteriorated predictive accuracy, for estimating both the mean and log variance functions, regardless of whether it is built on gPoE or gRBCM. Second, the DVBGP models built on gRBCM, gPoE, and gPoE-w deliver comparable predictive accuracies for estimating f and g, with gPoE and gPoE-w performing slightly better.

Regarding coverage ability, the following observations can be made from Tables 2 and 4. First, with a fixed total budget C, the coverage ability of all DVBGP models deteriorates as the number of subsets M increases, for estimating both the mean and log variance functions. Second, the DVBGP models built on gPoE and gPoE-w outperform their counterpart built on gRBCM, especially for estimating the log variance function g. Interestingly, the coverage abilities of all methods when estimating g seem to decrease with the total budget C.

Regarding subset partition schemes, we observe that using a communication subset combined with k-means partition helps improve predictive accuracy and coverage ability achieved by the DVBGP models built on gRBCM, gPoE, and gPoE-w, especially when estimating the log variance function g. In fact, pure random partition seems to lead to the worst predictive performance of all DVBGP models.

Last but not least, we note that the predictive accuracy and coverage ability achieved by a single VBGP model built on the entire dataset and by the DVBGP models are comparable when estimating the mean function. Nevertheless, a single VBGP model outperforms the DVBGP models when estimating the log variance function. Hence, the loss of information due to dataset partition is more severe in this case which deserves further investigation.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation under Grants No. CMMI-1846663 and III-1849300.

Table 1: Summary of the average ERMSEs with corresponding standard errors (in parentheses) when estimating f with $C = 500$, 5000 and 10^4.

	Subset formation		k-means with comm		k-means w/t comm		random	
C	VBGP	DVBGP	M=5	M=10	M=5	M=10	M=5	M=10
500	0.035 (0.001)	(g)RBCM	0.050 (0.002)	0.061 (0.004)	0.040 (0.001)	0.044 (0.001)	0.144 (0.006)	0.052 (0.002)
		gPoE	0.044 (0.001)	0.060 (0.003)	0.039 (0.001)	0.044 (0.001)	0.051 (0.002)	0.064 (0.004)
		gPoE-w	0.046 (0.002)	0.052 (0.003)	0.040 (0.001)	0.052 (0.002)	0.053 (0.002)	0.076 (0.005)
5000	0.0117 (4e-4)	(g)RBCM	0.016 (7e-4)	0.019 (7e-4)	0.013 (4e-4)	0.015 (4e-4)	0.018 (0.003)	0.020 (0.003)
		gPoE	0.015 (6e-4)	0.019 (7e-4)	0.013 (4e-4)	0.014 (3e-4)	0.019 (0.004)	0.021 (0.003)
		gPoE-w	0.014 (6e-4)	0.015 (6-04)	0.013 (4e-4)	0.015 (3e-4)	0.020 (0.004)	0.023 (0.003)
10^4	0.009 (3e-4)	(g)RBCM	0.012 (5e-4)	0.014 (5e-4)	0.030 (8e-4)	0.011 (4e-4)	0.011 (8e-4)	0.015 (0.002)
		gPoE	0.012 (5e-4)	0.014 (5e-4)	0.030 (6e-4)	0.011 (4e-4)	0.011 (8e-4)	0.016 (0.002)
		gPoE-w	0.010 (4e-4)	0.011 (3e-4)	0.010 (3e-4)	0.011 (4e-4)	0.011 (8e-4)	0.017 (0.002)

Table 2: Summary of the average MCPs with corresponding standard errors (in parentheses) when estimating f with $C = 500$, 5000 and 10^4.

	Subset formation		k-means with comm		k-means w/t comm		random	
C	VBGP	DVBGP	M=5	M=10	M=5	M=10	M=5	M=10
500	0.94 (0.05)	(g)RBCM	0.69 (0.09)	0.67 (0.09)	0.95 (0.04)	0.86 (0.07)	0.88 (0.06)	0.87 (0.07)
		gPoE	0.97 (0.03)	0.96 (0.04)	0.95 (0.04)	0.95 (0.04)	0.96 (0.04)	0.98 (0.02)
		gPoE-w	0.96 (0.04)	0.97 (0.03)	0.99 (0.01)	0.99 (0.02)	0.96 (0.04)	0.97 (0.03)
5000	0.96 (0.04)	(g)RBCM	0.5 (0.1)	0.5 (0.1)	0.97 (0.03)	0.97 (0.04)	0.70 (0.09)	0.73 (0.09)
		gPoE	0.97 (0.03)	0.99 (0.02)	0.97 (0.03)	0.97 (0.03)	0.86 (0.07)	0.94 (0.05)
		gPoE-w	0.98 (0.03)	0.99 (0.02)	0.999 (0.005)	1.000 (0.002)	0.86 (0.07)	0.94 (0.05)
10^4	0.95 (0.04)	(g)RBCM	0.5 (0.1)	0.5 (0.1)	0.84 (0.07)	0.82 (0.08)	0.72 (0.09)	0.74 (0.09)
		gPoE	0.96 (0.04)	0.99 (0.02)	0.98 (0.02)	0.97 (0.03)	0.94 (0.04)	0.95 (0.04)
		gPoE-w	0.97 (0.03)	0.99 (0.02)	0.998 (0.009)	1 (0)	0.94 (0.05)	0.95 (0.04)

Table 3: Summary of the average ERMSEs with corresponding standard errors (in parentheses) when estimating g with $C = 500$, 5000 and 10^4.

	Subset formation		k-means with comm		k-means w/t comm		random	
C	VBGP	DVBGP	M=5	M=10	M=5	M=10	M=5	M=10
500	0.351 (0.007)	(g)RBCM	0.68 (0.03)	0.85 (0.03)	0.97 (0.04)	1.24 (0.07)	1.8 (0.1)	1.97 (0.09)
		gPoE	0.69 (0.03)	0.93 (0.03)	0.95 (0.03)	1.09 (0.03)	1.43 (0.04)	1.6 (0.1)
		gPoE-w	0.75 (0.03)	0.96 (0.02)	0.98 (0.02)	1.15 (0.03)	1.45 (0.05)	1.7 (0.1)
5000	0.155 (0.009)	(g)RBCM	0.33 (0.01)	0.53 (0.02)	0.77 (0.01)	1.13 (0.03)	2.4 (0.4)	2.1 (0.3)
		gPoE	0.33 (0.01)	0.56 (0.02)	0.78 (0.01)	1.15 (0.03)	2.2 (0.4)	2.0 (0.3)
		gPoE-w	0.323 (0.008)	0.56 (0.02)	0.79 (0.01)	1.15 (0.02)	2.1 (0.4)	2.0 (0.3)
10^4	0.123 (0.007)	(g)RBCM	0.38 (0.01)	0.56 (0.02)	0.941 (0.008)	1.21 (0.03)	1.9 (0.2)	2.1 (0.3)
		gPoE	0.372 (0.009)	0.58 (0.02)	0.943 (0.008)	1.21 (0.03)	1.7 (0.2)	2.1 (0.3)
		gPoE-w	0.38 (0.01)	0.57 (0.02)	0.81 (0.01)	1.21 (0.03)	1.6 (0.2)	2.2 (0.3)

Table 4: Summary of the average MCPs with corresponding standard errors (in parentheses) obtained when estimating g with $C = 500$, 5000 and 10^4.

	Subset formation		k-means with comm		k-means w/t comm		random	
C	VBGP	DVBGP	M=5	M=10	M=5	M=10	M=5	M=10
500	0.93 (0.05)	(g)RBCM	0.6 (0.1)	0.5 (0.1)	0.5 (0.1)	0.32 (0.09)	0.08 (0.05)	0.05 (0.04)
		gPoE	0.80 (0.08)	0.6 (0.1)	0.52 (0.1)	0.31 (0.09)	0.08 (0.05)	0.10 (0.06)
		gPoE-w	0.74 (0.08)	0.6 (0.1)	0.6 (0.1)	0.4 (0.1)	0.076 (0.05)	0.083 (0.06)
5000	0.87 (0.07)	(g)RBCM	0.31 (0.09)	0.17 (0.08)	0.26 (0.09)	0.19 (0.08)	0.04 (0.04)	0.06 (0.05)
		gPoE	0.69 (0.09)	0.5 (0.1)	0.25 (0.09)	0.18 (0.08)	0.11 (0.06)	0.17 (0.08)
		gPoE-w	0.68 (0.09)	0.5 (0.1)	0.4 (0.1)	0.29 (0.09)	0.11 (0.06)	0.17 (0.08)
10^4	0.86 (0.07)	(g)RBCM	0.19 (0.08)	0.10 (0.06)	0.28 (0.09)	0.08 (0.05)	0.02 (0.03)	0.05 (0.04)
		gPoE	0.5 (0.1)	0.3 (0.1)	0.3 (0.1)	0.13 (0.07)	0.06 (0.05)	0.18 (0.08)
		gPoE-w	0.5 (0.1)	0.33 (0.09)	0.30 (0.09)	0.23 (0.08)	0.07 (0.05)	0.18 (0.08)

REFERENCES

Cao, Y. 2018. *Scaling Gaussian Processes*. Ph. D. thesis, Department of Computer Science, University of Toronto, Toronto, Canada.

Cao, Y., and D. J. Fleet. 2014. "Generalized Product of Experts for Automatic and Principled Fusion of Gaussian Process Predictions". In *Modern Nonparametrics 3: Automating the Learning Pipeline Workshop*, 1–5. Montreal, Canada: Conference on Neural Information Processing Systems.

Cao, Y., and D. J. Fleet. 2015. "Transductive Log Opinion Pool of Gaussian Process Experts". In *Nonparametric Methods for Large Scale Representation Learning Workshop*, 1–5. Montreal, Canada: Conference on Neural Information Processing Systems.

Damianou, A. C., M. K. Tisias, and N. D. Lawrence. 2016. "Variational Inference for Latent Variables and Uncertain Inputs in Gaussian Processes". *The Journal of Machine Learning Research* 17(1):1425–1486.

Deisenroth, M. P., and J. Ng. 2015. "Distributed Gaussian Processes". In *Proceedings of the 2015 International Conference on Machine Learning*, Volume 37, 1481–1490. July 6^{th}-11^{th}, Lille, France.

Guhaniyogi, R., and S. Banerjee. 2018. "Meta-Kriging: Scalable Bayesian Modeling and Inference for Massive Spatial Datasets". *Technometrics* 60(4):430–444.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning*. 2nd ed. New York: Springer.

Hensman, J., N. Fusi, and N. D. Lawrence. 2013. "Gaussian Process for Big Data". In *Proceedings of the 29th Conference on Uncertainty in Artificial Intelligence*, 282–290. July 12^{th}-14^{th}, Bellevue, Washington, USA.

Heskes, T. 1998. "Selecting Weighting Factors in Logarithmic Opinion Pools". In *Advances in Neural Information Processing Systems*, 266–272. Cambridge, Massachusetts: The MIT Press.

Lázaro-Gredilla, M., and M. Titsias. 2011. "Variational Heteroscedastic Gaussian Process Regression". In *Proceedings of the 28th International Conference on Machine Learning*, 841–848. June 28^{th}-July 2^{nd}, Bellevue, Washington, USA.

Le, Q. V., A. J. Smola, and S. Canu. 2005. "Heteroscedastic Gaussian Process Regression". In *Proceedings of the 22nd International Conference on Machine Learning*, 489–496. August 7^{th}-11^{th}, Bonn, Germany.

Liu, H., J. Cai, Y. Wang, and Y. Ong. 2018. "Generalized Robust Bayesian Committee Machine for Large-Scale Gaussian Process Regression". In *Proceedings of the 35th International Conference on Machine Learning*, Volume 80, 3131–3140. July 10^{th}-15^{th}, Stockholm, Sweden.

Liu, H., Y. Ong, and J. Cai. 2018. "Large-scale Heteroscedastic Regression via Gaussian Process". arXiv:1811.01179v1.

Munoz-Gonzalez, L., M. Lázaro-Gredilla, and A. R. Figueiras-Vidal. 2011. "Heteroscedastic Gaussian Process Regression Using Expectation Propagation". In *2011 IEEE International Workshop on Machine Learning for Signal Processing*, 1–6.

Ng, J., and M. P. Deisenroth. 2014. "Hierarchical Mixture-of-Experts Model for Large-Scale Gaussian Process Regression". arXiv:1412.3078v1.

Rasmussen, C. E., and C. K. I. Williams. 2006. *Gaussian Processes for Machine Learning*. Massachusetts: the MIT Press.

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.

Santner, T. J., B. J. Williams, and W. I. Notz. 2003. *The Design and Analysis of Computer Experiments*. New York: Springer.

Scott, S. L., A. W. Blocker, F. V. Bonassi, H. A. Chipman, E. I. George, and R. McCulloch. 2016. "Bayes and Big Data: The Consensus Monte Carlo Algorithm". *Journal of Management Science and Engineering Management* 11(2):78–88.

Tresp, V. 2000. "A Bayesian Committee Machine". *Neural Computation* 12:2719–2741.

Wang, C., and R. M. Neal. 2012. "Gaussian Process Regression with Heteroscedastic or Non-Gaussian Residuals". *arXiv preprint arXiv:1212.6246*.

Wang, W. 2019. *A Dual Metamodeling Perspective for Design and Analysis of Stochastic Simulation Experiments*. Ph. D. thesis, Grado Department of Industrial and Systems Engineering, Virginia Tech, Blacksburg, Virginia, USA.

Wang, W., N. Chen, X. Chen, and L. Yang. 2019. "A Variational Inference-Based Heteroscedastic Gaussian Process Approach for Simulation Metamodeling". *ACM Transactions on Modeling and Computer Simulation* 29(1):6.

Wang, W., and X. Chen. 2018. "An Adaptive Two-Stage Dual Metamodeling Approach for Stochastic Simulation Experiments". *IISE Transactions* 50(9):820–836.

AUTHOR BIOGRAPHIES

WENJING WANG is a Ph.D. candidate in the Grado Department of Industrial and Systems Engineering at Virginia Tech. Her current research interests include stochastic metamodeling and simulation optimization. Her email address is wenjing@vt.edu.

XI CHEN is an Assistant Professor in the Grado Department of Industrial and Systems Engineering at Virginia Tech. Her research interests include stochastic modeling and simulation methodology, applied probability and statistics, and computer experiment design and analysis. Her email address is xchen6@vt.edu and her website is https://sites.google.com/vt.edu/xi-chen-ise/home.