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.

NESTED HETEROSCEDASTIC GAUSSIAN PROCESS FOR SIMULATION METAMODELING

Jin Zhao[1] and Xi Chen[1]

[1]Grado Department of Industrial System Engineering, Virginia Tech, Blacksburg, VA, USA

ABSTRACT

This paper introduces the nested heteroscedastic Gaussian process approach (NHGP) to tackle simulation metamodeling with large-scale heteroscedastic datasets. NHGP achieves scalability by aggregating sub-stochastic kriging (sub-SK) models built on disjoint subsets of a large-scale dataset, making it user-friendly for SK users. We show that the NHGP predictor possesses desirable statistical properties, including being the best linear unbiased predictor among those built by aggregating sub-SK models and being consistent. The numerical experiments demonstrate the competitive performance of NHGP.

1 INTRODUCTION

Gaussian process (GP) models have become prominent nonparametric surrogate models in various science and engineering domains (Rasmussen and Williams 2006). They offer not only accurate point estimates of true function values but also quantification of the corresponding predictive uncertainty. In the context of stochastic simulation, data are inherently subject to heteroscedasticity. For example, when simulating a queueing system, the output variance varies significantly across the input space. Stochastic kriging (SK), proposed by Ankenman et al. (2010), has demonstrated its effectiveness as a GP-based metamodeling approach, which is used for approximating the mean function implied by a stochastic simulation model (Chen and Zhou 2017; Wang and Chen 2018). However, the advent of big data and the continuous evolution of computer hardware amplify the inherent challenges faced by traditional GP models, including SK. These methods suffer from cubic complexity relative to data size, making them less feasible for applications that yield large datasets. To navigate these challenges and retain the quality of predictions, the development of scalable GP approaches has become imperative.

The development of scalable Gaussian process (GP) approaches has thrived as a result. Existing scalable approaches typically fall into two main categories: global approximations, which process the entirety of the dataset to distill essential information, and local approximations, which partition the dataset for targeted learning via local experts in sub-regions (Liu et al. 2020). Specifically, global approximations focus on sparsifying the full kernel matrix, which includes using sparse kernels and sparse approximations (Quinonero-Candela and Rasmussen 2005; Titsias 2009; Shen et al. 2006). On the other hand, local approximations aggregate predictions by local experts to improve scalability. Two major frameworks in this category are mixture of experts (MoE) and product of experts (PoE). MoE, also studied as ensemble learning, typically expresses the combination of local experts as a Gaussian mixture to enhance the overall predictive accuracy and robustness. Unlike MoE, which employs a weighted sum to combine several probability distributions associated with experts, PoE multiplies these probability distributions, hence the name. This framework avoids the weight assignment required by MoE, thereby aggregating local experts' predictions in a way that emphasizes mutual confirmation over individual contributions.

However, it has been shown that some aggregation methods in the PoE framework, including PoE, generalized PoE (gPoE), Bayesian committee machine (BCM), and robust BCM (RBCM), ignore the covariances between experts (or sub-models). As a result, they lack consistency properties (Szabó and Zanten 2019; Bachoc et al. 2022). Recently, Rullière et al. (2018) proposed nested kriging (NK) which

considers covariances between sub-models given by kriging models, leading to theoretically consistent predictions.

Like most scalable GP approaches, however, NK, does not account for the heteroscedasticity inherent in stochastic simulation outputs. A notable contribution that fills this gap with reported successful applications is the distributed variational sparse heteroscedastic Gaussian process approach (DVSHGP) proposed by Liu et al. (2021). DVSHGP offers scalable variational inference-based estimation of both the mean and variance functions, facilitated by the use of inducing points. Despite its advancements, DVSHGP aggregates local experts following the BCM formalism, rendering its potential inconsistency.

In this paper, we introduce the nested heteroscedastic Gaussian process (NHGP) approach, inspired by the principles of NK. This method is specifically developed to address the challenges of metamodeling large-scale heteroscedastic datasets. NHGP achieves scalability by aggregating sub-SK models built on disjoint subsets of a large-scale dataset, making it user-friendly for SK users. We show that the NHGP predictor possesses desirable statistical properties, including being the best linear unbiased predictor among those built by aggregating sub-SK models and being consistent.

The paper is organized as follows. Section 2 provides a brief review of SK. Section 3 elaborates on the NHGP approach. Section 4 presents numerical evaluations to demonstrate the performance of NHGP. Section 5 concludes the paper.

2 REVIEW OF STOCHASTIC KRIGING

This section provides a brief review of stochastic kriging (SK) following Ankenman et al. (2010) and Chen et al. (2012).

Given a simulation model, SK assumes that the simulation output generated on the jth replication at design point $\mathbf{x} \in \mathscr{X} \subset \mathbb{R}^d$ can be modeled as

$$\mathscr{Y}_j(\mathbf{x}) = \mathsf{Y}(\mathbf{x}) + \varepsilon_j(\mathbf{x}) = \mathbf{h}(\mathbf{x})^\top \boldsymbol{\beta} + \mathsf{M}(\mathbf{x}) + \varepsilon_j(\mathbf{x}), \quad j = 1, 2, \ldots \tag{1}$$

where $\mathsf{Y}(\cdot) = \mathbf{h}(\cdot)^\top \boldsymbol{\beta} + \mathsf{M}(\cdot)$ represents the unknown mean function to estimate, $\mathbf{h}(\cdot) = (h_1(\cdot), h_2(\cdot), \ldots, h_l(\cdot))^\top$ denotes the $l \times 1$ vector of known regression functions, and $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_l)^\top$ denotes the $l \times 1$ vector of unknown regression parameters. The simulation noise terms $\varepsilon_1(\mathbf{x}), \varepsilon_2(\mathbf{x}), \ldots$ are independent and identically distributed (i.i.d.) with mean zero and variance $\mathsf{V}(\mathbf{x})$ for $\mathbf{x} \in \mathscr{X}$. In Equation (1), $\mathsf{M}(\cdot)$ is assumed to be a random draw from a mean-zero stationary Gaussian process whose covariance function is given by $K(\tilde{\mathbf{x}}, \mathbf{x}'; \tau^2, \theta) := \mathrm{Cov}[\mathsf{M}(\tilde{\mathbf{x}}), \mathsf{M}(\mathbf{x}')]$ for any $\tilde{\mathbf{x}}, \mathbf{x}' \in \mathscr{X}$. Here $K(\cdot, \cdot; \tau^2, \theta)$ denotes the kernel function which determines the smoothness properties of $\mathsf{M}(\cdot)$, and $\tau^2 \in \mathbb{R}_+$ and $\theta \in \mathbb{R}_+^d$ respectively represent the process variance and the lengthscale parameters. To ease notation, we suppress τ^2 and θ and use $K(\cdot, \cdot)$ in the remainder of the paper.

A simulation experimental design for SK metamodeling specifies the set of design points $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ and the number of replications r_i to run the simulation model at each design point $\mathbf{x}_i \in \mathbf{X}$. Given the simulation dataset $\mathscr{D} = \{\mathbf{x}_i, \{\mathscr{Y}_j(\mathbf{x}_i)\}_{j=1}^{r_i}, i = 1, 2, \ldots, n\}$, one can obtain the $n \times 1$ vector of sample average simulation outputs $\bar{\mathscr{Y}} = \left(\bar{\mathscr{Y}}(\mathbf{x}_1), \bar{\mathscr{Y}}(\mathbf{x}_2), \ldots, \bar{\mathscr{Y}}(\mathbf{x}_n)\right)^\top$, where

$$\bar{\mathscr{Y}}(\mathbf{x}_i) = \frac{1}{r_i} \sum_{j=1}^{r_i} \mathscr{Y}_j(\mathbf{x}_i) = \mathsf{Y}(\mathbf{x}_i) + \bar{\varepsilon}(\mathbf{x}_i),$$

with $\bar{\varepsilon}(\mathbf{x}_i) = r_i^{-1} \sum_{j=1}^{r_i} \varepsilon_j(\mathbf{x}_i)$. Denote the $n \times 1$ vector of average noise terms incurred at the design points by $\bar{\varepsilon} = (\bar{\varepsilon}(\mathbf{x}_1), \bar{\varepsilon}(\mathbf{x}_2), \ldots, \bar{\varepsilon}(\mathbf{x}_n))^\top$. Let the $n \times n$ matrix $\boldsymbol{\Sigma}_\varepsilon$ denote the variance-covariance matrix of $\bar{\varepsilon}$. In this paper, we assume that common random numbers are not applied in simulation experiments. In this case, $\boldsymbol{\Sigma}_\varepsilon$ reduces to the $n \times n$ diagonal matrix $\boldsymbol{\Sigma}_\varepsilon = \mathrm{diag}\{\mathsf{V}(\mathbf{x}_1)/r_1, \mathsf{V}(\mathbf{x}_2)/r_1, \ldots, \mathsf{V}(\mathbf{x}_n)/r_n\}$.

Let the $n \times n$ matrix $K(\mathbf{X}, \mathbf{X}) = (K(\mathbf{x}_i, \mathbf{x}_j))_{1 \le i,j \le n}$ record the pairwise covariances between $\mathsf{M}(\mathbf{x}_i)$ and $\mathsf{M}(\mathbf{x}_j)$ for any $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$. Given any prediction point $\mathbf{x}_0$, let the $n \times 1$ vector $K(\mathbf{X}, \mathbf{x}_0) = (K(\mathbf{x}_0, \mathbf{x}_1), \ldots, K(\mathbf{x}_0, \mathbf{x}_n))^\top$

record the covariances between $\mathsf{M}(\mathbf{x}_0)$ and $\mathsf{M}(\mathbf{x}_j)$ for all $\mathbf{x}_j \in \mathbf{X}$. Chen et al. (2012) show that the *best linear unbiased predictor* (BLUP) of $\mathsf{Y}(\mathbf{x}_0)$ at any prediction point $\mathbf{x}_0 \in \mathscr{X}$ is given by

$$\mu(\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_0)^\top \widehat{\boldsymbol{\beta}} + K(\mathbf{X},\mathbf{x}_0)^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} \left(\bar{\mathscr{Y}} - \mathbf{H}\widehat{\boldsymbol{\beta}}\right), \tag{2}$$

where $\mathbf{H} = (\mathbf{h}(\mathbf{x}_1), \mathbf{h}(\mathbf{x}_2), \ldots, \mathbf{h}(\mathbf{x}_n))^\top$ is the $n \times l$ model matrix of full rank, and $\widehat{\boldsymbol{\beta}}$ is the generalized least squares (GLS) estimator of $\boldsymbol{\beta}$ with its closed form given by

$$\widehat{\boldsymbol{\beta}} = \left(\mathbf{H}^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} \mathbf{H}\right)^{-1} \mathbf{H}^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} \bar{\mathscr{Y}} \ . \tag{3}$$

The corresponding mean square error (MSE) of the BLUP follows as $K(\mathbf{x}_0,\mathbf{x}_0) - K(\mathbf{X},\mathbf{x}_0)^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} K(\mathbf{X},\mathbf{x}_0) + \boldsymbol{\gamma}^\top \left(\mathbf{H}^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} \mathbf{H}\right)^{-1} \boldsymbol{\gamma}$, where $\boldsymbol{\gamma} = \mathbf{h}(\mathbf{x}_0) - \mathbf{H}^\top (K(\mathbf{X},\mathbf{X}) + \boldsymbol{\Sigma}_\varepsilon)^{-1} K(\mathbf{X},\mathbf{x}_0)$.

3 NESTED HETEROSCEDASTIC GAUSSIAN PROCESS

This section details the nested heteroscedastic Gaussian process (NHGP) approach, which extends nested kriging (NK) proposed by Rullière et al. (2018) to tackle large-scale heteroscedastic datasets.

The idea underpinning NHGP is to partition the dataset into disjoint subsets, upon which sub-SK models are individually constructed. These sub-SK models are subsequently aggregated to form the NHGP predictor. Specifically, we partition the design-point set $\mathbf{X}$ into p disjoint subsets $\mathbf{X}_k = \{\mathbf{x}_i^{(k)}, i = 1,2,\ldots,n_k\}$ for $k = 1,2,\ldots,p$, such that $\mathbf{X}_k \cap \mathbf{X}_\ell = \emptyset$ for $k \neq \ell, k,\ell = 1,2,\ldots,p$ and $\cup_{k=1}^p \mathbf{X}_k = \mathbf{X}$. The subdataset corresponding to $\mathbf{X}_k$ is denoted as $\mathscr{D}_k = \{\mathbf{x}_i^{(k)}, \{\mathscr{Y}_j(\mathbf{x}_i^{(k)})\}_{j=1}^{r_i^{(k)}}, i = 1,2,\ldots,n_k\}$, where $r_i^{(k)}$ denotes the number of replications at the ith design point in $\mathbf{X}_k$, $\mathbf{x}_i^{(k)}$, $i = 1,2,\ldots,n_k$.

Given each subdataset $\mathscr{D}_k$, we apply the SK methodology and construct a sub-SK model. The resulting predictive mean at any given $\mathbf{x}_0 \in \mathscr{X}$ follows from (2) and is given by

$$\mu_k(\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_0)^\top \widehat{\boldsymbol{\beta}}_k + K(\mathbf{X}_k,\mathbf{x}_0)^\top \left(K(\mathbf{X}_K,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \left(\bar{\mathscr{Y}}_k - \mathbf{H}_k\widehat{\boldsymbol{\beta}}_k\right) = \omega_k(\mathbf{x}_0)^\top \bar{\mathscr{Y}}_k,\ k = 1,2,\ldots p, \tag{4}$$

where $\omega_k(\mathbf{x}_0)$ denotes the $n_k \times 1$ vector given by

$$\begin{aligned}
\omega_k(\mathbf{x}_0)^\top = {} & \mathbf{h}(\mathbf{x}_0)^\top \left(\mathbf{H}_k^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \mathbf{H}_k\right)^{-1} \mathbf{H}_k^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \\
& - K(\mathbf{X}_k,\mathbf{x}_0)^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \mathbf{H}_k \left(\mathbf{H}_k^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \mathbf{H}_k\right)^{-1} \mathbf{H}_k^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1} \\
& + K(\mathbf{X}_k,\mathbf{x}_0)^\top \left(K(\mathbf{X}_k,\mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1},
\end{aligned}$$

$K(\mathbf{X}_k,\mathbf{X}_k) = (K(\mathbf{x}_i^{(k)},\mathbf{x}_j^{(k)}))_{1\le i,j\le n_k}$, $K(\mathbf{X}_k,\mathbf{x}_0) = (K(\mathbf{x}_1^{(k)},\mathbf{x}_0), K(\mathbf{x}_2^{(k)},\mathbf{x}_0), \ldots, K(\mathbf{x}_{n_k}^{(k)},\mathbf{x}_0))^\top$, and $\mathbf{H}_k = (\mathbf{h}(\mathbf{x}_1^{(k)}), \mathbf{h}(\mathbf{x}_2^{(k)}), \ldots, \mathbf{h}(\mathbf{x}_{n_k}^{(k)}))^\top$ denotes the $n_k \times l$ model matrix corresponding to the kth subset. In (4), $\widehat{\boldsymbol{\beta}}_k$ denotes the $l \times 1$ GLS estimator obtained based on $\mathscr{D}_k$, $\bar{\mathscr{Y}}_k = \left(\bar{\mathscr{Y}}(\mathbf{x}_1^{(k)}), \bar{\mathscr{Y}}(\mathbf{x}_2^{(k)}), \ldots, \bar{\mathscr{Y}}(\mathbf{x}_{n_k}^{(k)})\right)^\top$ denotes the $n_k \times 1$ vector of average outputs, and $\boldsymbol{\Sigma}_\varepsilon^{(k)}$ denotes the $n_k \times n_k$ diagonal noise variance-covariance matrix corresponding to the kth subset.

Let $\overrightarrow{\mu}(\mathbf{x}_0) = (\mu_1(\mathbf{x}_0), \mu_2(\mathbf{x}_0), \ldots, \mu_p(\mathbf{x}_0))^\top$ denote the $p \times 1$ vector of predictive means given by the p sub-SK models. Define the $p \times p$ variance-covariance matrix $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0) = \mathrm{Cov}[\overrightarrow{\mu}(\mathbf{x}_0), \overrightarrow{\mu}(\mathbf{x}_0)]$ and the

$p \times 1$ covariance vector $K_{\mathrm{UK}}(\mathbf{x}_0) = \mathrm{Cov}[\overrightarrow{\mu}(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_0)]$ with their specific entries given by

$$\begin{aligned}
(\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0))_{k,j} &= \mathrm{Cov}\left[\mu_k(\mathbf{x}_0), \mu_j(\mathbf{x}_0)\right] = \omega_k(\mathbf{x}_0)^\top K(\mathbf{X}_k, \mathbf{X}_j)\omega_j(\mathbf{x}_0),\ k \neq j, k, j = 1, 2, \ldots, p \\
(\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0))_{k,k} &= \mathrm{Cov}\left[\mu_k(\mathbf{x}_0), \mu_k(\mathbf{x}_0)\right] = \omega_k(\mathbf{x}_0)^\top \left(K(\mathbf{X}_k, \mathbf{X}_k) + \boldsymbol{\Sigma}_\varepsilon^{(k)}\right) \omega_k(\mathbf{x}_0),\ k = 1, 2, \ldots, p \\
(K_{\mathrm{UK}}(\mathbf{x}_0))_k &= \mathrm{Cov}\left[\mu_k(\mathbf{x}_0), \mathsf{M}(\mathbf{x}_0)\right] = \omega_k(\mathbf{x}_0)^\top K(\mathbf{X}_k, \mathbf{x}_0),\ k = 1, 2, \ldots, p.
\end{aligned} \tag{5}$$

Assuming that $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)$ is invertible, NHGP provides the following predictor of $\mathsf{Y}(\mathbf{x}_0)$ by aggregating those from all p sub-SK models:

$$\mu_{\mathrm{NHGP}}(\mathbf{x}_0) = \boldsymbol{\alpha}(\mathbf{x}_0)^\top \overrightarrow{\mu}(\mathbf{x}_0), \tag{6}$$

where

$$\begin{aligned}
\boldsymbol{\alpha}(\mathbf{x}_0)^\top = &\left(\mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \mathbf{1}_p\right)^{-1} \mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \\
&- K_{\mathrm{UK}}(\mathbf{x}_0)^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \mathbf{1}_p \left(\mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \mathbf{1}_p\right)^{-1} \mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \\
&+ K_{\mathrm{UK}}(\mathbf{x}_0)^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1},
\end{aligned} \tag{7}$$

and $\mathbf{1}_q$ denotes the $q \times 1$ vector of ones. The specific forms of $K_{\mathrm{UK}}(\mathbf{x}_0)$ and $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)$ in (5) are derived in Appendix A.

Theorem 1 below shows that $\mu_{\mathrm{NHGP}}(\mathbf{x}_0)$ given in (6) is the best linear unbiased predictor (BLUP) from combining the sub-SK model predictors. It extends Proposition 1 in Rullière et al. (2018) and Proposition 10 in Bachoc et al. (2022) for NK to the heteroscedastic GP setting. The proof is provided in Appendix B.

Theorem 1 $\mu_{\mathrm{NHGP}}(\mathbf{x}_0)$ in (6) is the BLUP of $\mathsf{Y}(\mathbf{x}_0)$ given by combining the sub-SK predictors, and the corresponding MSE of $\mu_{\mathrm{NHGP}}(\mathbf{x}_0)$ follows as

$$v_{\mathrm{NHGP}}(\mathbf{x}_0) = K(\mathbf{x}_0, \mathbf{x}_0) - 2\boldsymbol{\alpha}(\mathbf{x}_0)^\top K_{\mathrm{UK}}(\mathbf{x}_0) + \boldsymbol{\alpha}(\mathbf{x}_0)^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)\boldsymbol{\alpha}(\mathbf{x}_0).$$

Acknowledging the lack of knowledge in true noise variances, we follow Ankenman et al. (2010) and consider the sample-variance-plugged-in NHGP predictor:

$$\widehat{\mu}_{\mathrm{NHGP}}(\mathbf{x}_0) = \widehat{\boldsymbol{\alpha}}(\mathbf{x}_0)^\top \widehat{\overrightarrow{\mu}}(\mathbf{x}_0), \tag{8}$$

where all quantities with hat are obtained by replacing $\boldsymbol{\Sigma}_\varepsilon^{(k)}$ with $\widehat{\boldsymbol{\Sigma}}_\varepsilon^{(k)} = \mathrm{diag}\{\widehat{\mathsf{V}}(\mathbf{x}_1^{(k)})/r_1^{(k)}, \ldots, \widehat{\mathsf{V}}(\mathbf{x}_{n_k}^{(k)})/r_{n_k}^{(k)}\}$ in their original expressions, and $\widehat{\mathsf{V}}(\mathbf{x}_i^{(k)}) = (r_i^{(k)} - 1)^{-1} \sum_{j=1}^{r_i^{(k)}} (\mathscr{Y}_j(\mathbf{x}_i^{(k)}) - \bar{\mathscr{Y}}(\mathbf{x}_i^{(k)}))^2,\ i = 1, 2, \ldots, n_k,\ k = 1, 2, \ldots, p$.

Theorem 2 indicates that estimating noise variances via sample variances introduces no prediction bias. The proof is deferred to Appendix C.

Assumption 1 (Assumption 1 in Ankenman et al. 2010) The Gaussian process M is stationary, and $\varepsilon_1(\mathbf{x}_i), \varepsilon_2(\mathbf{x}_i), \ldots$ are i.i.d. normally distributed with mean zero and variance $\mathsf{V}(\mathbf{x}_i)$, independent of $\varepsilon_j(\mathbf{x}_h)$ for all j and $h \neq i$, and independent of M.

Theorem 2 If Assumption 1 holds, then $\mathbb{E}\left(\widehat{\mu}_{\mathrm{NHGP}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)\right) = 0$.

Recall that several aggregation methods in the PoE framework that ignore the covariance between sub-models, including PoE, gPoE, BCM, and RBCM, lack consistency. The next result ensures consistency of the NHGP predictor.

Theorem 3 Under Assumption 1, assume that the model matrices corresponding to all subsets (if they exist), $H_k, k = 1, 2, \ldots, p$, have full rank, and the design-point set $\mathbf{X}$ is recurrently dense, namely, for all $\mathbf{x}_0 \in \mathscr{X}$, $\lim_{n\to\infty} \min_{1 \leq i \leq n} \|\mathbf{x}_i - \mathbf{x}_0\| = 0$. Then the NHGP predictor $\mu_{\mathrm{NHGP}}(\mathbf{x}_0)$ is consistent, that is, $\sup_{\mathbf{x}_0 \in \mathscr{X}} v_{\mathrm{NHGP}}(\mathbf{x}_0) \to 0$ as $n \to \infty$.

We sketch the proof of Theorem 3. In light of Theorem 1, we note that, for any $\mathbf{x}_0 \in \mathscr{X}$, $v_{\text{NHGP}}(\mathbf{x}_0)$ is smaller than the MSE of any sub-SK model's predictor $\mu_k(\mathbf{x}_0)$, $k = 1,2,\ldots,p$. As $n \to \infty$, there exists a subset of design points, say, the k_0th subset $\mathbf{X}_{k_0}$, such that $n_{k_0} = |\mathbf{X}_{k_0}| \to \infty$. Hence, it suffices to show that the MSE of $\mu_{k_0}(\mathbf{x}_0)$ at prediction point $\mathbf{x}_0$, $v_{k_0}(\mathbf{x}_0) \to 0$ as $n_{k_0} \to \infty$. To establish this, consider the following two scenarios. First, the regression functions in $\mathbf{h} \neq 0$ and we need to estimate $\boldsymbol{\beta}$. Theorem 2 in Wang and Hu (2018) shows that $v_{k_0}(\mathbf{x}_0)$ decreases monotonically with n_{k_0} when $\mathbf{H}_{k_0}$ has full rank, and since $v_{k_0}(\mathbf{x}_0) \geq 0$, we have $v_{k_0}(\mathbf{x}_0) \to 0$ as $n_{k_0} \to \infty$. Second, the regression functions in $\mathbf{h} = 0$. In this case, Theorem 3 in Koepernik and Pfaff (2021) establishes that $\sup_{\mathbf{x}_0 \in \mathscr{X}} v_{\text{NHGP}}(\mathbf{x}_0) \to 0$ as $n \to \infty$ if the design-point set $\mathbf{X}$ is recurrently dense. Theorem 3 also indicates that the properties of NHGP depend on those of its constituent sub-SK models.

4 NUMERICAL EVALUATIONS

This section demonstrates the performance of NHGP through three numerical examples.

4.1 Methods in Comparison

We compare NHGP with a state-of-the-art approach, DVSHGP, proposed by Liu et al. (2021). Given a simulation dataset $\mathscr{D}$ containing B observations collected at n distinct design points in $\mathbf{X}$, both NHGP and DVSHGP require specifying the number of subgroups to divide $\mathbf{X}$ into disjoint subsets; we use p_1 for NHGP and p_2 for DVSHGP to differentiate between the two parameters used in the two approaches. Following the suggestions by Rullière et al. (2018) for NK and by Liu et al. (2021) for DVSHGP, we adopt k-means clustering (Chapter 13 of Hastie et al. 2009) for partitioning the design points into disjoint subsets. Rullière et al. (2018) provided a comprehensive discussion on the choice of p_1 for NK. Specifically, a small p_1 is suitable for low-dimensional cases ($d < 5$), while a large p_1 is preferable for high-dimensional problems. We adhere to this suggestion when implementing NHGP and make adjustments in specific examples. However, a rule-of-thumb for choosing p_2 for DVSHGP is lacking. Therefore, we select its value from a set of candidate choices that yield the best predictive performance in each example.

Regarding hyper-parameter estimation, NHGP has the lengthscale parameters in $\boldsymbol{\theta}$ and the process variance τ^2 to estimate. We adopt a two-step leave-one-out procedure that employs stochastic gradient descent detailed in Rullière et al. (2018) for their estimation. Furthermore, to facilitate a fair comparison with DVSHGP, we adopt the Gaussian kernel in our implementation and set the regression functions $h_i(\mathbf{x}) = 0$ for $i = 1,2\ldots,l$, eliminating the need to estimate $\boldsymbol{\beta}$.

The parameters for DVSHGP to estimate include the kernel parameters for GP modeling of the mean and variance functions, the B variational parameters, and the $(m+u) \times d$ inducing-point location parameters, where m (respectively u) denotes the number of inducing points used for mean (resp. variance) function estimation. Liu et al. (2021) adopts a hybrid estimation strategy that combines natural gradient descent with the Adam method to seek optimal parameter values. We follow Liu et al. (2021) and set $m = u$ in our implementation.

4.2 General Experimental Setup

We provide a brief description of the general experiment setup used in all examples. A simulation experiment is performed with a total budget of B simulation replications to expend at n distinct design points, with r_i replications allocated at design point $\mathbf{x}_i$, for $i = 1,2,\ldots,n$. We consider three budget allocation schemes: equal allocation, unequal allocation 1, and unequal allocation 2. Specifically, the equal allocation prescribes $r_i = \lceil B/n \rceil$, where $\lceil a \rceil$ denotes the smallest integer greater than or equal to a. Unequal allocation 1 sets $r_i = \left\lceil \frac{\mathsf{V}(\mathbf{x}_i)}{\sum_{j=1}^n \mathsf{V}(\mathbf{x}_j)} B \right\rceil$ and unequal allocation 2 assigns $r_i = \left\lceil \frac{\sqrt{\mathsf{V}(\mathbf{x}_i)}}{\sum_{j=1}^n \sqrt{\mathsf{V}(\mathbf{x}_j)}} B \right\rceil$.

For all examples considered, we repeat the simulation experiment for 100 independent macro-replications and calculate the empirical root mean squared error (RMSE) achieved by NHGP and DVSHGP on each

macro-replication for assessing their predictive accuracy. Specifically, a check-point set comprising n_{pred} points from $\mathscr{X}$ is generated according to the Sobol' quasi-random sequence, and the RMSE achieved on the ℓth macro-replication is given by

$$\text{RMSE}_\ell = \sqrt{\frac{1}{n_{\text{pred}}} \sum_{i=1}^{n_{\text{pred}}} (\mu_\ell(\mathbf{x}_{0,i}) - f(\mathbf{x}_{0,i}))^2}, \quad \ell = 1,2,\ldots,100, \tag{9}$$

where $\mu_\ell(\mathbf{x}_{0,i})$ denotes the predictive mean obtained by a given method at $\mathbf{x}_{0,i}$ on the ℓth macro-replication, and $f(\mathbf{x}_{0,i})$ denotes the corresponding true mean function value to estimate.

4.3 Examples

In each example, we generate simulation outputs according to the following model:

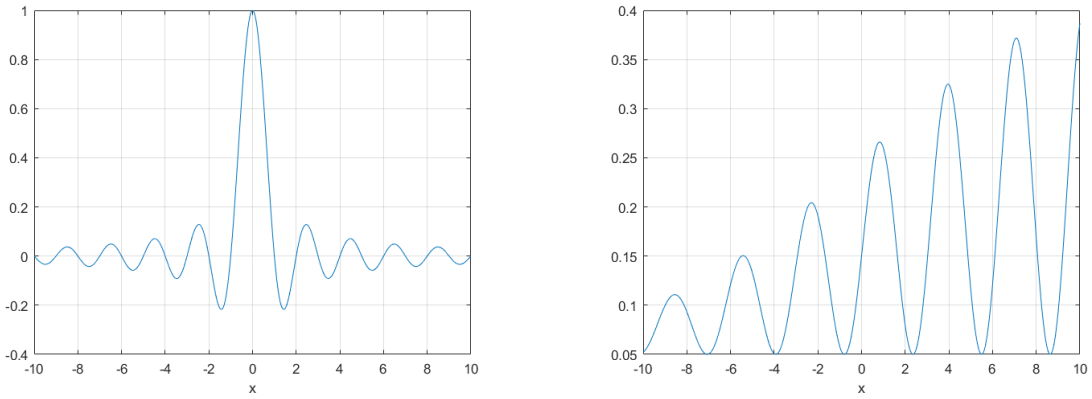
$$\mathscr{Y}_j(\mathbf{x}) = f(\mathbf{x}) + \varepsilon_j(\mathbf{x}), \quad \text{for } j = 1,2,\ldots$$

where $f(\cdot)$ denotes the true mean function of interest, and the simulation noise terms $\varepsilon_j(\mathbf{x})$'s are i.i.d. normally distributed with mean zero and variance $\mathsf{V}(\mathbf{x})$, for any $\mathbf{x} \in \mathscr{X}$. The input space $\mathscr{X}$, the mean function $f(\cdot)$, and the variance function $\mathsf{V}(\cdot)$ are to be specified separately for each example.

1-D Sinc. Consider the following 1-dimensional example, which is also studied by Liu et al. (2021). The input space is $\mathscr{X} = [-10,10]$. For any $x \in \mathscr{X}$, the mean function is given by

$$f(x) = \begin{cases} \sin(\pi x)/(\pi x) & x \neq 0 \\ 1 & x = 0 \end{cases}.$$

The noise variance function is given by $\mathsf{V}(x) = (0.05 + 0.2(1+\sin(2x))/(1+e^{-0.2x}))^2$ for $x \in \mathscr{X}$. Figure 1 (a) and (b) illustrates the mean and noise variance functions for the 1-D Sinc example.



(a) $f(\cdot)$ (b) $\mathsf{V}(\cdot)$

Figure 1: 1-D Sinc: the mean and variance functions.

2-D Branin. Consider the following two-dimensional example where the input space is $\mathscr{X} = [0,5]^2$. For any $\mathbf{x} = (x_1, x_2) \in \mathscr{X}$, the mean and variance functions are respectively given by

$$f(\mathbf{x}) = \left(x_2 - \frac{5}{4\pi^2}x_1^2 + \frac{5}{\pi}x_1 - 6\right)^2 + 10\left(1 - \frac{1}{8\pi}\right)\cos(x_1) + 10, \quad \mathsf{V}(\mathbf{x}) = 0.01|f(\mathbf{x})|.$$

2-D Sine. Consider the following two-dimensional example with the input space given by $\mathscr{X} = [-1,1]^2$. For any $\mathbf{x} = (x_1, x_2) \in \mathscr{X}$, the mean and variance functions are respectively specified as

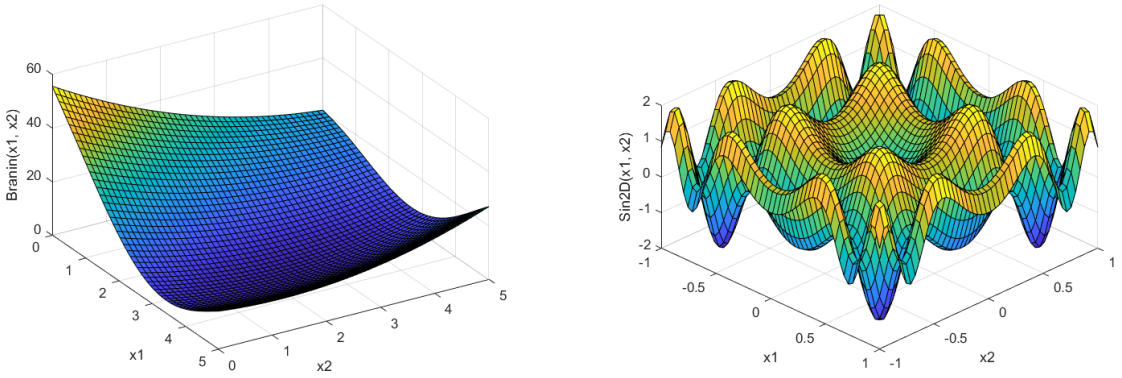
$$f(\mathbf{x}) = \sin(9x_1^2) + \sin(9x_2^2), \quad \mathsf{V}(\mathbf{x}) = 2 + \cos(\pi + (x_1 + x_2)/2).$$

The mean functions of the two 2-D examples are illustrated in Figure 2(a) and (b), showing very distinct features. We observe that the mean response surface of the Branin example changes relatively smoothly across its input space, despite the wide range of its function values. In contrast, the mean response surface of the Sine example has a narrow range of function values, yet it displays a more complex and rapidly changing landscape.

We conduct numerical evaluations according to Subsections 4.2 and 4.1, with specific parameter settings used in the three examples summarized in Table 1.

Table 1: Parameter settings for all three examples.

	B	n	n_{pred}	NHGP	DVSHGP	
				# subsets p_1	# subsets p_2	# inducing points m
1-D Sinc	5000	$\{100, 200, 500, 10^3\}$	500	5	5	10
2-D Branin	25000	$\{500, 10^3, 2500, 5000\}$	2500	3	10	30
2-D Sine	25000	$\{500, 10^3, 2500, 5000\}$	2500	5	10	20



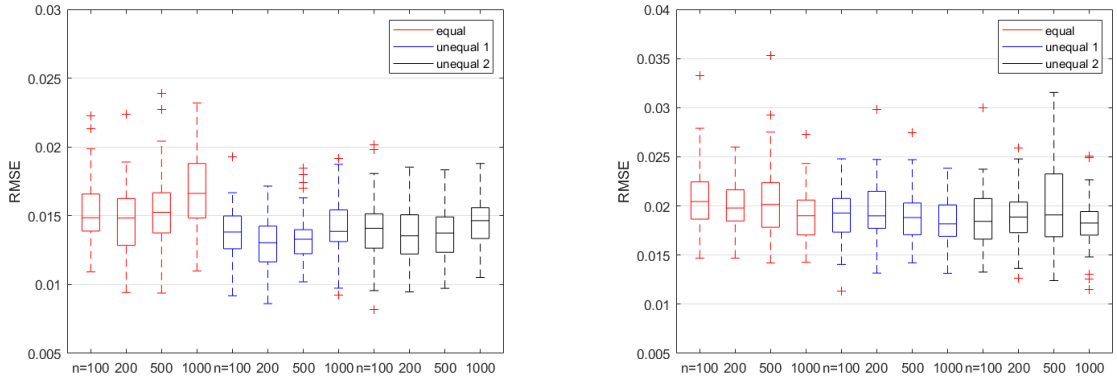
(a) $f(\cdot)$ for Branin on [0, 5]^2 (b) $f(\cdot)$ for Sine on [-1, 1]^2

Figure 2: The mean functions for the Branin and Sine examples.

4.4 Results

Figure 3(a) and (b) show the RMSEs obtained by NHGP and DVSHGP for the 1-D Sinc example across 100 macro-replications. We observe that NHGP and DVSHGP's performance is comparable, but the former outperforms the latter by delivering smaller RMSEs. Furthermore, varying the number of distinct design points n given a fixed budget B has little impact on the performance of NHGP and DVSHGP in this 1-D Sinc example. Using unequal allocation schemes yields a narrower range of RMSEs compared to using the equal allocation scheme for NHGP, an observation not as evident for DVSHGP.

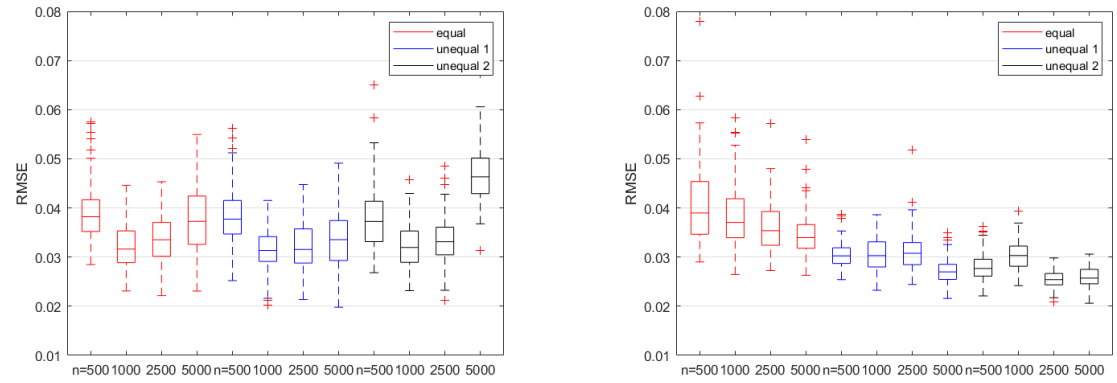
Figure 4(a) and (b) summarize the RMSEs obtained by NHGP and DVSHGP for the 2-D Branin example. It is observed that while NHGP and DVSHGP exhibit comparable performance, the latter slightly



(a) NHGP (b) DVSHGP

Figure 3: 1-D Sinc: RMSEs obtained by NHGP and DVSHGP with varying numbers of design points under three budget allocation rules.

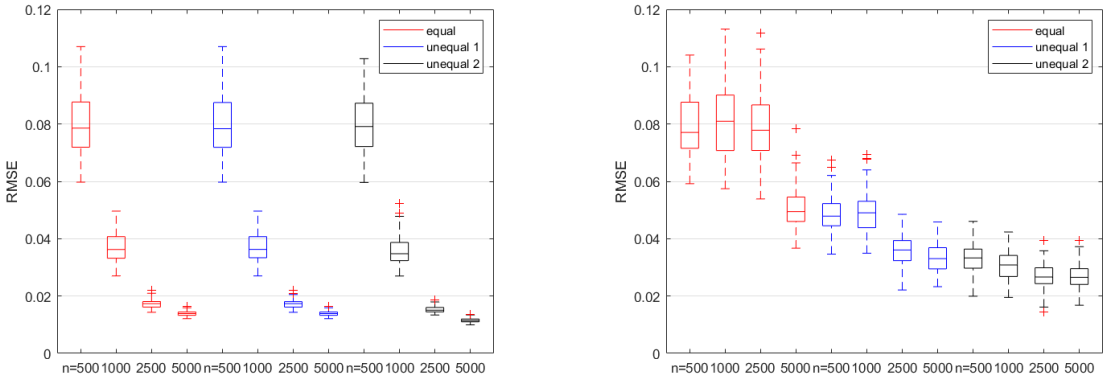
outperforms the former. Notably, NHGP and DVSHGP demonstrate different behaviors under varying budget allocation schemes and as the number of design points increases given a fixed budget. Specifically, NHGP's RMSEs remain comparable across different budget allocation schemes, whereas the unequal allocation schemes assist DVSHGP in achieving lower RMSEs compared to the equal allocation scheme. Furthermore, under a fixed budget B, the RMSEs obtained by DVSHGP decrease with an increasing number of design points n, whereas for NHGP, the RMSEs first decrease and then increase with n. This divergence in behaviors of NHGP and DVSHGP can be explained as follows: NHGP uses output data at the sample average level. Given a fixed budget B to expend, NHGP faces a clear trade-off between increasing the number of distinct design points n and increasing the simulation efforts allocated at each design point to achieve high predictive accuracy while tackling the impact of strong heteroscedasticity in this example. In contrast, DVSHGP exploits information from all B outputs at an individual observation level, resulting in a different behavior given a fixed budget B to expend.



(a) NHGP (b) DVSHGP

Figure 4: 2-D Branin: RMSEs obtained by NHGP and DVSHGP with varying numbers of design points under three budget allocation rules.

Finally, the RMSEs obtained by NHGP and DVSHGP for the 2-D Sine example are shown in Figure 5(a) and (b). We have the following observations. First, NHGP outperforms DVSHGP by yielding much smaller RMSEs, demonstrating NHGP's promising capability in approximating mean functions with intricate, complex patterns. Second, the RMSEs obtained under the two unequal allocation schemes are lower than those obtained under the equal budget allocation scheme for both NHGP and DVSHGP. However, NHGP's RMSEs remain comparable across different budget allocation schemes, whereas the benefit of using unequal allocation schemes is more pronounced for DVSHGP. Furthermore, the RMSEs obtained by both NHGP and DVSHGP decrease with the number of design points n, given a fixed budget B to expend. This observation for NHGP contrasts with that made in the 2-D Branin example; recall from Figure 2 that the mean response surface of this 2-D Sine example is more complex. Therefore, to achieve high predictive accuracy with NHGP given a fixed budget B, it is more effective to prioritize using a greater number of distinct design points to capture the rapidly changing landscape rather than allocating more simulation efforts at each design point to address the impact of heteroscedasticity in this example.



(a) NHGP (b) DVSHGP

Figure 5: 2-D Sine: RMSEs obtained by NHGP and DVSHGP with varying numbers of design points under three budget allocation rules.

Our preliminary numerical evaluations in this section provide the following insights. DVSHGP's primary advantage lies in its use of all outputs at an individual observation level, permitting a wider range of simulation experimental designs given a fixed budget to expend. However, the performance of DVSHGP relies on reliably estimating a collection of parameters, which becomes increasingly challenging as the number of inducing points grows, especially when the total budget B and the input space dimensionality d become large. In contrast, the proposed NHGP builds on sub-SK models and has only a few parameters to estimate. NHGP demonstrates competitive, robust performance under various experimental settings, particularly with respect to approximating complex mean functions.

5 CONCLUSION

In this paper, we proposed a scalable heteroscedastic simulation metamodeling approach called NHGP. NHGP achieves scalability by aggregating sub-SK models built on disjoint subsets of a large-scale dataset and is shown to provide the best linear unbiased predictor among those built by aggregating sub-SK models. The weights obtained by NHGP for its predictor consider all pairwise covariances between the sub-SK models, ensuring the consistency of the NHGP predictor and avoiding pitfalls due to the restrictive covariance-free aggregation adopted by some state-of-the-art methods in the PoE framework, including DVSHGP. Our preliminary numerical evaluations demonstrate that NHGP compares favorably with DVSHGP.

Looking ahead, we identify several avenues for further research. First, extending NHGP to accommodate settings where only a single replication is available at some design points by exploring alternative methods for noise variance estimation. Second, leveraging parallel computing for NHGP parameter estimation and prediction represents a critical next step in advancing NHGP's applicability and efficiency.

ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation CAREER [CMMI-1846663].

A FORMS OF $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)$ AND $K_{\mathrm{UK}}(\mathbf{x}_0)$

Recall that

$$\begin{aligned}\omega_k(\mathbf{x}_0)^\top := \mathbf{h}(\mathbf{x}_0)^\top &\left(\mathbf{H}_k^\top \left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1}\mathbf{H}_k\right)^{-1}\mathbf{H}_k^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1}\\ &- K(\mathbf{X}_k,\mathbf{x}_0)^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1}\mathbf{H}_k\left(\mathbf{H}_k^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1}\mathbf{H}_k\right)^{-1}\mathbf{H}_k^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1}\\ &+ K(\mathbf{X}_k,\mathbf{x}_0)^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)^{-1},\end{aligned}$$

and we can write $\mu_k(\mathbf{x}_0)=\omega_k(\mathbf{x}_0)^\top\bar{\mathscr{Y}}(\mathbf{X}_k)$. Hence for $\forall j\neq k, j,k\in\{1,2,...,p\}$,

$$\begin{aligned}\mathrm{Cov}\left[\mu_k(\mathbf{x}_0),\mu_j(\mathbf{x}_0)\right] &= \mathrm{Cov}\left[\omega_k(\mathbf{x}_0)^\top\bar{\mathscr{Y}}(\mathbf{X}_k),\omega_j(\mathbf{x}_0)^\top\bar{\mathscr{Y}}(\mathbf{X}_j)\right] = \omega_k(\mathbf{x}_0)^\top\mathrm{Cov}[\bar{\mathscr{Y}}(\mathbf{X}_k),\bar{\mathscr{Y}}(\mathbf{X}_j)]\omega_j(\mathbf{x}_0)^\top\\ &= \omega_k(\mathbf{x}_0)^\top\mathrm{Cov}[\mathsf{Y}(\mathbf{X}_k)+\bar{\varepsilon}(\mathbf{X}_k),\mathsf{Y}(\mathbf{X}_j)+\bar{\varepsilon}(\mathbf{X}_j)]\omega_j(\mathbf{x}_0)^\top = \omega_k(\mathbf{x}_0)^\top K(\mathbf{X}_k,\mathbf{X}_j)\omega_j(\mathbf{x}_0)^\top.\end{aligned}$$

Similarly, we can obtain that for $k=1,2,\ldots,p$,

$$\begin{aligned}\mathrm{Cov}\left[\mu_k(\mathbf{x}_0),\mu_k(\mathbf{x}_0)\right] &= \omega_k(\mathbf{x}_0)^\top\left(K(\mathbf{X}_k,\mathbf{X}_k)+\boldsymbol{\Sigma}_\varepsilon^{(k)}\right)\omega_k(\mathbf{x}_0),\\ \mathrm{Cov}\left[\mu_k(\mathbf{x}_0),\mathsf{M}(\mathbf{x}_0)\right] &= \omega_k(\mathbf{x}_0)^\top K(\mathbf{X}_k,\mathbf{x}_0).\end{aligned}$$

The forms of $K_{\mathrm{UK}}(\mathbf{x}_0)=\mathrm{Cov}[\overrightarrow{\mu}(\mathbf{x}_0),\overrightarrow{\mu}(\mathbf{x}_0)]$ and $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)=\mathrm{Cov}[\overrightarrow{\mu}(\mathbf{x}_0),\mathsf{M}(\mathbf{x}_0)]$ then follow.

B PROOF OF THEOREM 1

Proof. To find the BLUP predictor of $\mathsf{Y}(\mathbf{x}_0)=\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta}+\mathsf{M}(\mathbf{x}_0)$ among all linear predictors in the form of $\lambda_0+\boldsymbol{\lambda}^\top\overrightarrow{\mu}(\mathbf{x}_0)$, we formulate and solve the following minimization problem:

$$\begin{aligned}&\min\ \mathbb{E}\left[\left(\mathsf{Y}(\mathbf{x}_0)-\lambda_0-\boldsymbol{\lambda}^\top\overrightarrow{\mu}(\mathbf{x}_0)\right)^2\right]\\ &\text{s.t. }\mathbb{E}\left[\mathsf{Y}(\mathbf{x}_0)-\lambda_0-\boldsymbol{\lambda}^\top\overrightarrow{\mu}(\mathbf{x}_0)\right]=0,\end{aligned}\tag{10}$$

where λ_0 is a scalar and $\boldsymbol{\lambda}$ is a $p\times 1$ vector.

The constraint is equivalent to $\lambda_0=0$ and $\boldsymbol{\lambda}^\top(\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta})\mathbf{1}_p=\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta}$ for all $\boldsymbol{\beta}$, where the second equation holds since $\mathbb{E}[\mu_k(\mathbf{x}_0)]=\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta}$ for $k=1,2,\ldots,p$, hence $\boldsymbol{\lambda}^\top\mathbf{1}_p=1$. Following similar steps given in Section 1.5 of Stein (2012), we can show that for $\boldsymbol{\lambda}^\top\overrightarrow{\mu}(\mathbf{x}_0)$ to be a BLUP of $\mathsf{Y}(\mathbf{x}_0)$, there must be a scalar η such that the following condition holds:

$$\begin{pmatrix}\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0) & \mathbf{1}_p\\ \mathbf{1}_p^\top & 0\end{pmatrix}\begin{pmatrix}\boldsymbol{\lambda}\\ \eta\end{pmatrix}=\begin{pmatrix}K_{\mathrm{UK}}(\mathbf{x}_0)\\ 1\end{pmatrix}.$$

Since $\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)$ is assumed invertible, we have

$$\boldsymbol{\lambda}^* = \left(\mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} - \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \mathbf{1}_p \left(\mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \mathbf{1}_p \right)^{-1} \mathbf{1}_p^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} \right) K_{\mathrm{UK}}(\mathbf{x}_0).$$

Substituting $\boldsymbol{\lambda}^*$ into $\boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0)$, we obtain the BLUP predictor $\mu_{\mathrm{NHGP}}(\mathbf{x}_0) = \boldsymbol{\alpha}(\mathbf{x}_0)^\top \overrightarrow{\mu}(\mathbf{x}_0)$, where $\boldsymbol{\alpha}(\mathbf{x}_0)^\top$ is as defined in (7). □

Remark. Notice that if $\boldsymbol{\beta}$ is known or $\mathbf{h} = \mathbf{0}$, the unbiasedness constraint is not required. Therefore, we are seeking an MSE-optimal predictor via solving the following unconstrained optimization problem:

$$\min_{\lambda_0, \boldsymbol{\lambda}} \ \mathbb{E}\left[\left(\mathsf{Y}(\mathbf{x}_0) - \lambda_0 - \boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0) \right)^2 \right].$$

Below we consider the case where $\boldsymbol{\beta}$ is known; the case where $\mathbf{h} = \mathbf{0}$ is similar. We can get

$$\begin{aligned}
&\mathbb{E}\left[\left(\mathsf{Y}(\mathbf{x}_0) - \lambda_0 - \boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0) \right)^2 \right] \\
&= \mathbb{E}\left[\left(\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \lambda_0 - \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right)^2 \right] + \mathbb{E}\left[\left(\mathsf{M}(\mathbf{x}_0) - \boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0) + \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right)^2 \right] \\
&\quad + 2\mathbb{E}\left[\left(\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \lambda_0 - \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right) \left(\mathsf{M}(\mathbf{x}_0) - \boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0) + \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right) \right] \\
&= \left(\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \lambda_0 - \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right)^2 - \mathbb{E}\left[\left(\mathsf{M}(\mathbf{x}_0) - \boldsymbol{\lambda}^\top \overrightarrow{\mu}(\mathbf{x}_0) + \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right)^2 \right] \\
&= \left(\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \lambda_0 - \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \right)^2 + K(\mathbf{x}_0, \mathbf{x}_0) - 2\boldsymbol{\lambda}^\top K_{\mathrm{UK}}(\mathbf{x}_0) + \boldsymbol{\lambda}^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0) \boldsymbol{\lambda}.
\end{aligned}$$

The MSE-optimal weights follow as $\lambda_0^* = \mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \boldsymbol{\lambda}^\top \mathbf{1}_p (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta})$ and $\boldsymbol{\lambda}^* = \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} K_{\mathrm{UK}}(\mathbf{x}_0)$. With these weights, the MSE-optimal predictor (when $\boldsymbol{\beta}$ is known) follows as

$$\mu_{\mathrm{NHGP}}(\mathbf{x}_0) = \mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} + K_{\mathrm{UK}}(\mathbf{x}_0)^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} (\overrightarrow{\mu}(\mathbf{x}_0) - (\mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta}) \mathbf{1}_p),$$

and the corresponding MSE is given by

$$v_{\mathrm{NHGP}}(\mathbf{x}_0) = K(\mathbf{x}_0, \mathbf{x}_0) - K_{\mathrm{UK}}(\mathbf{x}_0)^\top \mathbb{K}_{\mathrm{UK}}(\mathbf{x}_0)^{-1} K_{\mathrm{UK}}(\mathbf{x}_0).$$

C PROOF OF THEOREM 2

Proof. The proof follows the same line of argument as given in the proof of Theorem 1 in Ankenman et al. (2010). First, we show that for any fixed positive definite covariance matrices $(\boldsymbol{\Sigma}_\varepsilon^{(k)})'$, $k = 1, 2, \ldots, p$, the predictor $\widehat{\mu}'_{\mathrm{NHGP}}(\mathbf{x}_0) = \widehat{\boldsymbol{\alpha}}'(\mathbf{x}_0)^\top \widehat{\overrightarrow{\mu}}'(\mathbf{x}_0)$ is unbiased, where $\widehat{\boldsymbol{\alpha}}'(\mathbf{x}_0)^\top$ and $\widehat{\overrightarrow{\mu}}'(\mathbf{x}_0)'$ are obtained by replacing $\boldsymbol{\Sigma}_\varepsilon^{(k)}$ in the forms of $\boldsymbol{\alpha}$ and $\overrightarrow{\mu}(\mathbf{x}_0)$ with $(\boldsymbol{\Sigma}_\varepsilon^{(k)})'$. This follows immediately from $\mathbb{E}(\widehat{\mu}'_{\mathrm{NHGP}}(\mathbf{x}_0) - \mathsf{Y}(\mathbf{x}_0)) = \mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} - \mathbf{h}(\mathbf{x}_0)^\top \boldsymbol{\beta} = 0$.

Next, for $k = 1, 2, \ldots, p$, notice that the sample variance at design point $\mathbf{x}_i^{(k)}$ in the kth subset $\mathbf{X}_k$ follows as

$$\widehat{\mathsf{V}}(\mathbf{x}_i^{(k)}) = \frac{1}{r_i^{(k)} - 1} \sum_{j=1}^{r_i^{(k)}} \left(\mathscr{Y}_j(\mathbf{x}_i^{(k)}) - \bar{\mathscr{Y}}(\mathbf{x}_i^{(k)}) \right)^2 = \frac{1}{r_i^{(k)} - 1} \sum_{j=1}^{r_i^{(k)}} \left(\varepsilon_j(\mathbf{x}_i^{(k)}) - \bar{\varepsilon}(\mathbf{x}_i^{(k)}) \right)^2, \ i = 1, 2, \ldots, n_k,$$

where $\bar{\varepsilon}_k$ is the average noise vector incurred at design points in $\mathbf{X}_k$. Under Assumption 1, $\widehat{\mathsf{V}}(\mathbf{x}_i^{(k)})$ is independent of $\bar{\mathscr{Y}}_k$ by the properties of the multivariate normal distribution (recall that M is also independent of $\varepsilon_j(\mathbf{x}_i^{(k)})$). Then, it follows that

$$\begin{aligned}\mathbb{E}\left(\widehat{\mu}_{\mathrm{NHGP}}(\mathbf{x}_0)-\mathsf{Y}(\mathbf{x}_0)\right) &= \mathbb{E}\left[\mathbb{E}\left(\widehat{\mu}_{\mathrm{NHGP}}(\mathbf{x}_0)-\mathsf{Y}(\mathbf{x}_0)\mid \widehat{\boldsymbol{\Sigma}}_\varepsilon^{(1)},\widehat{\boldsymbol{\Sigma}}_\varepsilon^{(2)},\ldots,\widehat{\boldsymbol{\Sigma}}_\varepsilon^{(p)}\right)\right]\\ &= \mathbb{E}\left(\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta}-\mathbf{h}(\mathbf{x}_0)^\top\boldsymbol{\beta}\right)=0.\end{aligned}$$

□

REFERENCES

Ankenman, B. E., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58:371–382.

Bachoc, F., N. Durrande, D. Rullière, and C. Chevalier. 2022. "Properties and Comparison of Some Kriging Sub-model Aggregation Methods". *Mathematical Geosciences* 54:941–977.

Chen, X., B. E. Ankenman, and B. L. Nelson. 2012. "The Effects of Common Random Numbers on Stochastic Kriging Metamodels". *ACM Transactions on Modeling and Computer Simulation* 22:1–20.

Chen, X. and Q. Zhou. 2017. "Sequential Design Strategies for Mean Response Surface Metamodeling via Stochastic Kriging with Adaptive Exploration and Exploitation". *European Journal of Operational Research* 262:575–585.

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

Koepernik, P. and F. Pfaff. 2021. "Consistency of Gaussian Process Regression in Metric Spaces". *Journal of Machine Learning Research* 22:1–27.

Liu, H., Y. Ong, X. Shen, and J. Cai. 2020. "When Gaussian Process Meets Big Data: A Review of Scalable GPs". *IEEE Transactions on Neural Networks and Learning Systems* 31:4405–4423.

Liu, H., Y. S. Ong, and J. Cai. 2021. "Large-scale Heteroscedastic Regression via Gaussian Process". *IEEE Transactions on Neural Networks and Learning Systems* 32:708–721.

Quinonero-Candela, J. and C. E. Rasmussen. 2005. "A Unifying View of Sparse Approximate Gaussian Process Regression". *The Journal of Machine Learning Research* 6:1939–1959.

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

Rullière, D., N. Durrande, F. Bachoc, and C. Chevalier. 2018. "Nested Kriging Predictions for Datasets with A Large Number of Observations". *Statistics and Computing* 28:849–867.

Shen, Y., A. Ng, and M. Seeger. 2006. "Fast Gaussian Process Regression Using KD-Trees". In *Advances in Neural Information Processing Systems 18*, 1225–1232. Cambridge, MA, USA: MIT Press.

Stein, M. L. 2012. *Interpolation of Spatial Data: Some Theory for Kriging*. Springer Science & Business Media.

Szabó, B. and H. V. Zanten. 2019. "An Asymptotic Analysis of Distributed Nonparametric Methods". *Journal of Machine Learning Research* 20:1–30.

Titsias, M. 2009. "Variational Learning of Inducing Variables in Sparse Gaussian Processes". In *Proceedings of the Twelfth International Conference on Artificial Intelligence and Statistics*, Volume 5, 567–574.

Wang, B. and J. Hu. 2018. "Some Monotonicity Results for Stochastic Kriging Metamodels in Sequential Settings". *INFORMS Journal on Computing* 30:278–294.

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

AUTHOR BIOGRAPHIES

JIN ZHAO is a Ph.D. candidate in the Grado Department of Industrial and Systems Engineering at Virginia Tech. His current research interests include stochastic simulation modeling of large-scale and high-dimensional data and inference. His email address is zjin20@vt.edu.

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