*Proceedings of the 2023 Winter Simulation Conference*
*C. G. Corlu, S. R. Hunter, H. Lam, B. S. Onggo, J. Shortle, and B. Biller, eds.*

# ASYMPTOTIC NORMALITY OF JOINT METAMODEL-BASED SOBOL' INDEX ESTIMATORS

Jingtao Zhang
Xi Chen
Ruochen Wang

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

## ABSTRACT

This paper proposes two joint metamodel-based Sobol' index estimators and investigates their asymptotic properties. The numerical evaluation corroborates the theoretical results and highlights the impact of the combination of training sample size and Monte Carlo sample size on the estimators' performance.

## 1 INTRODUCTION

Sobol' indices are widely used in global sensitivity analysis for assessing the input parameters' impact on the model output (Sobol' 1990; Saltelli and Annoni 2010), with successful applications in epidemiological modeling, defect detection in manufacturing, pollutant transport modeling, etc. Estimating Sobol' indices can be computationally demanding, especially when the input space dimensionality is high. Many Monte Carlo (MC)-based estimators have emerged to efficiently estimate Sobol' indices; see, e.g., Tarantola et al. (2007), Saltelli et al. (2010), and Mazo (2021). However, MC-based estimators can be inefficient when the simulation model is computationally expensive to evaluate. In recent years, research on metamodel-based Sobol' index estimators has attracted increasing attention; see, e.g., Marrel et al. (2012), Janon et al. (2014), and Hart et al. (2017). Thanks to efficient metamodeling techniques and suitable experimental designs, it is possible to construct an accurate metamodel using a relatively small training sample generated by a limited number of simulation runs for the purpose of Sobol' index estimation.

While point estimates of Sobol' indices can serve as a measure of the importance of input variables, it is arguably equally important to quantify the uncertainty associated with these estimates. Non-asymptotic uncertainty quantification methods based on bootstrapping and empirical distributions have been proposed; see, e.g., Storlie et al. (2009), Marrel et al. (2009), and Janon et al. (2014). On the other hand, Janon et al. (2014) investigated the asymptotic normality of metamodel-based Sobol' index estimators for deterministic simulation models. For stochastic simulation models, Mazo (2021) conducted an asymptotic analysis for MC-based Sobol' index estimators. However, the asymptotic properties of metamodel-based estimators for a stochastic simulation model have rarely been explored. To the best of our knowledge, this work is among the first to investigate the asymptotic normality of metamodel-based Sobol' index estimators for a stochastic simulation model.

In this paper, we propose joint metamodel-based Sobol' index estimators which rely on estimation of both the mean and variance functions implied by a stochastic simulation experiment. We prove the estimators' asymptotic normality, based on which asymptotic confidence intervals can be constructed for uncertainty quantification. The rest of the paper is organized as follows. Section 2 briefly introduces Sobol' indices and proposes two joint metamodel-based Sobol' index estimators. Section 3 presents the theoretical

analysis of the proposed estimators. Section 4 presents numerical experiments for performance evaluations and verifies the theoretical results.

## 2 JOINT METAMODEL-BASED ESTIMATION

This section briefly reviews Sobol' indices for global sensitivity analysis and proposes two joint metamodel-based Sobol' index estimators. Given the input vector $X = (X_1, X_2, \ldots, X_p)^\top \in \mathcal{X} \subset \mathbb{R}^p$, we consider a stochastic simulation model of the form $\mathcal{Y} = f(X, \varepsilon)$, where $f$ is some real-valued function and $\varepsilon$ represents the uncertainty inherent to the simulation model. For any subset $u \subset \{1, 2, \ldots, p\}$, define $X_u$ as the subset of entries in $X$ with indices in $u$; e.g., for $u = \{1, 2\}$, $X_u = (X_1, X_2)$. Define $X_{-u} := X \backslash X_u$. The Sobol' index of $X_u$ quantifies the contribution of the input variable(s) in $X_u$ towards the variance of the model output $\mathcal{Y}$, based on the following functional variance decomposition: $\mathrm{Var}_{X,\varepsilon}(\mathcal{Y}) = V_\varepsilon(\mathcal{Y}) + \sum_{i=1}^{p} \sum_{|J|=i} (V_J(\mathcal{Y}) + V_{J\varepsilon}(\mathcal{Y}))$, where $V_i(\mathcal{Y}) = \mathrm{Var}_{X_i}(\mathbb{E}_{X_{-i},\varepsilon}(\mathcal{Y} \mid X_i))$, $V_\varepsilon(\mathcal{Y}) = \mathrm{Var}_\varepsilon(\mathbb{E}_X(\mathcal{Y} \mid \varepsilon))$, $V_{ij}(\mathcal{Y}) = \mathrm{Var}_{X_i,X_j}\left(\mathbb{E}_{X_{-\{i,j\}},\varepsilon}(\mathcal{Y} \mid X_i, X_j)\right) - V_i(\mathcal{Y}) - V_j(\mathcal{Y})$, $i \neq j$, $V_{i\varepsilon}(\mathcal{Y}) = \mathrm{Var}_{X_i,\varepsilon}(\mathbb{E}_{X_{-i}}(\mathcal{Y} \mid X_i, \varepsilon)) - V_i(\mathcal{Y}) - V_\varepsilon(\mathcal{Y})$, and so on. Following Mazo (2021), the Sobol' index of $X_u$ for a stochastic simulation model is defined as

$$S^{X_u} = \frac{\mathrm{Var}_{X_u}(\mathbb{E}_{X_{-u},\varepsilon}(\mathcal{Y} \mid X_u))}{\mathrm{Var}_{X,\varepsilon}(\mathcal{Y})}, \quad u \subset \{1, 2, \ldots, p\}. \tag{1}$$

The value of $S^{X_u}$ is between 0 and 1, and a larger value indicates greater importance of the input variable(s) in $X_u$ to $Y$. The Sobol' index given in (1) can be estimated via different methods, e.g., the pick-freeze scheme (Gamboa et al. 2016), which amounts to running the simulation model at inputs $(X_{u,i}, X_{-u,i})$ and $(X'_{u,i}, X_{-u,i})$ for $i \in \mathbb{N}^+$, where $X_{u,i}$ and $X_{-u,i}$ are MC samples of $X_u$ and $X_{-u}$, and $X'_{u,i}$ is an independent copy of $X_{u,i}$. When the simulation model is computationally expensive to evaluate, Janon et al. (2014), Hart et al. (2017), and Castellan et al. (2020) showed that metamodeling can facilitate efficient estimation of the Sobol' indices by only performing simulation runs to build a metamodel and evaluating the metamodel for Sobol' index estimation using a MC sample of input vectors.

Following Marrel et al. (2012), we consider joint metamodel-based Sobol' index estimation for stochastic simulation models. Denote by $Y_m(X) := \mathbb{E}_\varepsilon(\mathcal{Y} \mid X)$ the mean function and $Y_d(X) := \mathrm{Var}_\varepsilon(\mathcal{Y} \mid X)$ the variance function, which are defined on the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$. We can rewrite Equation (1) as

$$S^{X_u} = \frac{\mathrm{Var}_{X_u}(\mathbb{E}_{X_{-u}}(Y_m \mid X_u))}{\mathrm{Var}_X(Y_m) + \mathbb{E}_X(Y_d)}, \tag{2}$$

since $\mathrm{Var}_{X_u}(\mathbb{E}_{X_{-u},\varepsilon}(\mathcal{Y} \mid X_u)) = \mathrm{Var}_{X_u}(\mathbb{E}_{X_{-u}}(Y_m \mid X_u))$ and $\mathrm{Var}_{X,\varepsilon}(\mathcal{Y}) = \mathrm{Var}_X(Y_m) + \mathbb{E}_X(Y_d)$. Consider estimating $Y_m$ and $Y_d$ by metamodels $\widetilde{Y}_{m,\mathcal{T}_N}$ and $\widetilde{Y}_{d,\mathcal{T}_N}$, where $\mathcal{T}_N$ denotes a training sample, $\{(X_i, \mathcal{Y}_i)\}_{i=1}^n$, to obtain the metamodels, $n$ denotes the training sample size, $N$ denotes the MC sample size for metamodel evaluations, and $n \to \infty$ as $N \to \infty$. As Janon et al. (2014) showed for deterministic simulation models, there exist some relationships between $n$ and $N$ to make the metamodel-based Sobol' index estimators consistent. Assume that the training sample $\mathcal{T}_N$ is defined on the probability space $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$. The metamodels $\widetilde{Y}_{m,\mathcal{T}_N}$ and $\widetilde{Y}_{d,\mathcal{T}_N}$ are defined on the product space $(\Omega_Z \times \Omega_X, \sigma(\mathcal{F}_Z \times \mathcal{F}_X), \mathbb{P}_Z \otimes \mathbb{P}_X)$. In particular, given a fixed $\omega \in \Omega_Z$, $\widetilde{Y}_{m,\mathcal{T}_N(\omega)}$ and $\widetilde{Y}_{d,\mathcal{T}_N(\omega)}$ are defined on the probability space $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$. Since $Y_m$ and $Y_d$ are unavailable, we adopt $\widetilde{Y}_{m,\mathcal{T}_N}$ and $\widetilde{Y}_{d,\mathcal{T}_N}$ for Sobol' index estimation. Define $\widetilde{Y}_{m,\mathcal{T}_N,i} := \widetilde{Y}_{m,\mathcal{T}_N}((X_{u,i}, X_{-u,i}))$, $\widetilde{Y}^X_{m,\mathcal{T}_N,i} := \widetilde{Y}_{m,\mathcal{T}_N}((X'_{u,i}, X_{-u,i}))$, $\widetilde{Y}_{d,\mathcal{T}_N,i} := \widetilde{Y}_{d,\mathcal{T}_N}((X_{u,i}, X_{-u,i}))$ , and $\widetilde{Y}^{(X)}_{m,\mathcal{T}_N,i} := \widetilde{Y}_{m,\mathcal{T}_N}((X'_{u,i}, X'_{-u,i}))$ for $i \in \mathbb{N}^+$, where $X'_{-u,i}$ is an independent copy of $X_{-u,i}$. Based on (2), we propose two estimators to be detailed next.

***The First Sobol' Index Estimator.*** The first estimator of $S^{X_u}$ in (2) is devised by noting that $\mathrm{Var}_{X_u}(\mathbb{E}_{X_{-u}}(Y_m \mid X_u)) = \mathrm{Cov}(Y_m((X_u, X_{-u})), Y_m((X'_u, X_{-u})))$, where $X'_u$ is an independent copy of $X_u$; see Lemma 2.2 of Janon et al. (2014). Based on the outputs $\{\widetilde{Y}_{m,\mathcal{T}_N,i}, \widetilde{Y}^X_{m,\mathcal{T}_N,i}, \widetilde{Y}_{d,\mathcal{T}_N,i}\}_{i=1}^N$ from evaluating the mean and variance metamodels,

the first estimator can be given as

$$\widetilde{S}^{X_u}_{\mathcal{T}_N} = \frac{\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{m,\mathcal{T}_N,i}\widetilde{Y}^X_{m,\mathcal{T}_N,i} - \left(\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{m,\mathcal{T}_N,i}\right)\left(\frac{1}{N}\sum_{i=1}^N \widetilde{Y}^X_{m,\mathcal{T}_N,i}\right)}{\frac{1}{N}\sum_{i=1}^N \widetilde{Y}^2_{m,\mathcal{T}_N,i} - \left(\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{m,\mathcal{T}_N,i}\right)^2 + \frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{d,\mathcal{T}_N,i}}. \tag{3}$$

***The Second Sobol' Index Estimator.*** There exist alternative estimators of the numerator on the right-hand side of (2); see, e.g., Table 2 in Saltelli et al. (2010). We study one of them in this paper. Define $\widetilde{Y}^{(X)}_{m,\mathcal{T}_N,i} := \widetilde{Y}_{m,\mathcal{T}_N}((X'_{u,i}, X'_{-u,i}))$. Based on the outputs $\{\widetilde{Y}_{m,\mathcal{T}_N,i}, \widetilde{Y}^X_{m,\mathcal{T}_N,i}, \widetilde{Y}^{(X)}_{m,\mathcal{T}_N,i}, \widetilde{Y}_{d,\mathcal{T}_N,i}\}_{i=1}^N$ from evaluating the metamodels, we can obtain the second estimator as

$$\widetilde{T}^{X_u}_{\mathcal{T}_N} = \frac{\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{m,\mathcal{T}_N,i}(\widetilde{Y}^X_{m,\mathcal{T}_N,i} - \widetilde{Y}^{(X)}_{m,\mathcal{T}_N,i})}{\frac{1}{N}\sum_{i=1}^N \widetilde{Y}^2_{m,\mathcal{T}_N,i} - \left(\frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{m,\mathcal{T}_N,i}\right)^2 + \frac{1}{N}\sum_{i=1}^N \widetilde{Y}_{d,\mathcal{T}_N,i}}. \tag{4}$$

Our analysis of the two metamodel-based estimators given in (3) and (4) is detailed in the next section. It is worthwhile studying the following two estimators given the knowledge of the true mean and variance functions $Y_m(\cdot)$ and $Y_d(\cdot)$:

$$S^{X_u}_N = \frac{\frac{1}{N}\sum_{i=1}^N Y_{m,i}Y^X_{m,i} - \left(\frac{1}{N}\sum_{i=1}^N Y_{m,i}\right)\left(\frac{1}{N}\sum_{i=1}^N Y^X_{m,i}\right)}{\frac{1}{N}\sum_{i=1}^N Y^2_{m,i} - \left(\frac{1}{N}\sum_{i=1}^N Y_{m,i}\right)^2 + \frac{1}{N}\sum_{i=1}^N Y_{d,i}}, \tag{5}$$

and

$$T^{X_u}_N = \frac{\frac{1}{N}\sum_{i=1}^N Y_{m,i}(Y^X_{m,i} - Y^{(X)}_{m,i})}{\frac{1}{N}\sum_{i=1}^N Y^2_{m,i} - \left(\frac{1}{N}\sum_{i=1}^N Y_{m,i}\right)^2 + \frac{1}{N}\sum_{i=1}^N Y_{d,i}}, \tag{6}$$

where the two estimators above are defined analogously to those in (3) and (4) but with the outputs from evaluating the metamodels replaced by those from evaluating the true mean and variance functions.

## 3 ASYMPTOTIC ANALYSIS

This section investigates the asymptotic normality of the estimators proposed in Section 2. We first show that this property holds true for the estimators that are based on the true mean and variance functions under a bounded moment condition. Then, we investigate when the asymptotic normality holds true if the true mean and variance functions are unavailable and are replaced by the metamodels.

### 3.1 Analysis of the Sobol' Index Estimators Using the True Mean and Variance Functions

This subsection analyzes the asymptotic normality of the Sobol' index estimators given in (5) and (6) when the true mean and variance functions, $Y_m(X)$ and $Y_d(X)$, are available. For ease of notation, we write $Y_m$ and $Y_d$ hereinafter. The analysis relies on Assumption 1 below.

**Assumption 1** $\mathbb{E}_X\left(Y_m^4\right) < \infty$ and $\mathbb{E}_X\left(Y_d^2\right) < \infty$.

**Proposition 1** Under Assumption 1, $\sqrt{N}\left(S^{X_u}_N - S^{X_u}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0, \sigma_S^2)$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$, where

$$\sigma_S^2 = \mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X(Y_m)\right)\left(Y^X_m - \mathbb{E}_X(Y_m)\right) - S^{X_u}\left(\left(Y_m - \mathbb{E}_X(Y_m)\right)^2 + Y_d\right)\right)\left(\mathrm{Var}_{X,\varepsilon}(\mathcal{Y})\right)^{-2}. \tag{7}$$

*Proof.* Define

$$U_i := \left(\left(Y_{m,i} - \mathbb{E}_X(Y_m)\right)\left(Y^X_{m,i} - \mathbb{E}_X(Y_m)\right), Y_{m,i} - \mathbb{E}_X(Y_m), Y^X_{m,i} - \mathbb{E}_X(Y_m), \left(Y_{m,i} - \mathbb{E}_X(Y_m)\right)^2, Y_{d,i}\right), \tag{8}$$

and $\overline{U}_N := N^{-1}\sum_{i=1}^N U_i$. Define $\mu := \left(\mathrm{Cov}\left(Y_m, Y_m^X\right), 0, 0, \mathrm{Var}_X\left(Y_m\right), \mathbb{E}_X\left(Y_d\right)\right)$. We have $\sqrt{N}(\overline{U}_N - \mu) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\Gamma)$, where $\Gamma$ is the variance-covariance matrix of $U_1$. Let $\psi_S(x,y,z,a,b) = (x-yz)/(a-y^2+b)$, and $S_N^{X_u} = \psi_S(\overline{U}_N)$ according to (5). Since

$$\nabla\psi_S = \left(\frac{1}{a-y^2+b}, -\frac{z}{a-y^2+b} + \frac{2y(x-yz)}{(a-y^2+b)^2}, -\frac{y}{a-y^2+b}, -\frac{x-yz}{(a-y^2+b)^2}, -\frac{x-yz}{(a-y^2+b)^2}\right),$$

we have $\nabla\psi_S(\mu) = \left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1}, 0, 0, -\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1} S^{X_u}, -\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1} S^{X_u}\right)$. It follows that

$$\nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu) = \mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_m^X - \mathbb{E}_X\left(Y_m\right)\right) - S^{X_u}\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)^2 + Y_d\right)\right)\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^{-2}.$$

Assumption 1 ensures the validity of $\nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu)$. The proof is complete by applying the Delta method with $\sigma_S^2 = \nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu)$. □

Similarly, we can show that the asymptotic normality holds true for the estimator given in (6).

**Proposition 2** Under Assumption 1, $\sqrt{N}\left(T_N^{X_u} - S^{X_u}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma_T^2)$ in $(\Omega_X, \mathcal{F}_X, \mathbb{P}_X)$, where

$$\sigma_T^2 = \mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_m^X - Y_m^{(X)}\right) - S^{X_u}\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)^2 + Y_d\right)\right)\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^{-2}. \tag{9}$$

*Proof.* The proof is similar to that of Proposition 1. Define

$$U_i := \Big(\left(Y_{m,i} - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_{m,i}^X - \mathbb{E}_X\left(Y_m\right)\right), \left(Y_{m,i} - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_{m,i}^{(X)} - \mathbb{E}_X\left(Y_m\right)\right), \left(Y_{m,i} - \mathbb{E}_X\left(Y_m\right)\right)^2, Y_{m,i}^X - \mathbb{E}_X\left(Y_m\right), Y_{d,i}\Big),$$

and $\overline{U}_N := N^{-1}\sum_{i=1}^N U_i$. Define $\mu := \left(\mathrm{Cov}\left(Y_m, Y_m^X\right), 0, \mathrm{Var}_X\left(Y_m\right), 0, \mathbb{E}_X\left(Y_d\right)\right)$. We have $\sqrt{N}(\overline{U}_N - u) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\Gamma)$, where $\Gamma$ is the variance-covariance matrix of $U_1$. Let $\psi_S(x,y,z,a,b) = (x-y)/(z-a^2+b)$, and $T_N^{X_u} = \psi_S(\overline{U}_N)$ according to (6). Since

$$\nabla\psi_S = \left(\frac{1}{z-a^2+b}, -\frac{1}{z-a^2+b}, -\frac{x-y}{(z-a^2+b)^2}, \frac{2a(x-y)}{(z-a^2+b)^2}, -\frac{xy}{(z-a^2+b)^2}\right),$$

we have $\nabla\psi_S(u) = \left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1}, \mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1}, -\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1} S^{X_u}, 0, -\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1} S^{X_u}\right)$, and it follows that

$$\nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu) = \mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_m^X - Y_m^{(X)}\right) - S^{X_u}\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)^2 + Y_d\right)\right)\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^{-2}.$$

Assumption 1 ensures the validity of $\nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu)$. The proof is complete by applying the Delta method with $\sigma_T^2 = \nabla\psi_S(\mu)\Gamma\nabla\psi_S^\top(\mu)$. □

As the estimators given in (5) and (6) are unbiased, we are interested in which one may have a lower asymptotic variance. Define $V := \left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(Y_m^X - \mathbb{E}_X\left(Y_m\right)\right) - S^{X_u}\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)^2 + Y_d\right)$. Then we have $\sigma_S^2 = \mathrm{Var}_X\left(V\right)/\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^2$, and

$$\begin{aligned}\sigma_T^2 &= \mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(\mathbb{E}_X\left(Y_m\right) - Y_m^{(X)}\right) + V\right)\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^{-2} \\ &= \sigma_S^2 + \frac{\mathrm{Var}_X\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(\mathbb{E}_X\left(Y_m\right) - Y_m^{(X)}\right)\right) + 2\,\mathrm{Cov}\left(\left(Y_m - \mathbb{E}_X\left(Y_m\right)\right)\left(\mathbb{E}_X\left(Y_m\right) - Y_m^{(X)}\right), V\right)}{\left(\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)\right)^2}.\end{aligned} \tag{10}$$

A closer examination reveals that the sign of the second term on the right-hand side of (10) is indeterminate. In fact, the relationship between $\sigma_S^2$ and $\sigma_T^2$ is example dependent; see Section 4 for more details.

### 3.2 Analysis of the Joint Metamodel-based Sobol' Index Estimators

This subsection investigates the asymptotic normality of the two joint metamodel-based estimators $\widetilde{S}^{X_u}_{\mathcal{T}_N}$ and $\widetilde{T}^{X_u}_{\mathcal{T}_N}$, respectively given in (3) and (4), when the true mean and variance functions are unavailable.

The following decomposition is key to our analysis of the first joint metamodel-based estimator $\widetilde{S}^{X_u}_{\mathcal{T}_N}$:

$$\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N} - S^{X_u}\right) = \sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N} - \widetilde{S}^{X_u}\right) + \sqrt{N}\left(\widetilde{S}^{X_u} - S^{X_u}\right), \tag{11}$$

where $\widetilde{S}^{X_u} \coloneqq \mathrm{Var}_{X_u}\left(\mathbb{E}_{X_{-u}}\left(\widetilde{Y}_{m,\mathcal{T}_N} \mid X_u\right)\right) / \left(\mathrm{Var}_X\left(\widetilde{Y}_{m,\mathcal{T}_N}\right) + \mathbb{E}_X\left(\widetilde{Y}_{d,\mathcal{T}_N}\right)\right)$ which is similar as $S^{X_u}$ given in (2), but with the true mean and variance functions replaced by the respective metamodels built on the given data set $\mathcal{T}_N$. A decomposition similar to (11) holds true for the second estimator $\widetilde{T}^{X_u}_{\mathcal{T}_N}$ as well. We next investigate the properties of $\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N} - \widetilde{S}^{X_u}\right)$ and $\sqrt{N}\left(\widetilde{S}^{X_u} - S^{X_u}\right)$ in light of (11). Let us start with some technical conditions.

**Assumption 2** For almost every $\omega \in \Omega_Z$, $\delta_{m,\mathcal{T}_N(\omega)} \coloneqq \widetilde{Y}_{m,\mathcal{T}_N(\omega)} - Y_m \xrightarrow[N\to\infty]{\mathcal{L}^2} c$, where $c$ is some constant.

**Assumption 3** For almost every $\omega \in \Omega_Z$, $\mathbb{E}_X\left(\widetilde{Y}_{d,\mathcal{T}_N(\omega)}\right) \xrightarrow[N\to\infty]{} \mathbb{E}_X\left(Y_d\right)$.

**Assumption 3′** For almost every $\omega \in \Omega_Z$, $\delta_{d,\mathcal{T}_N(\omega)} \coloneqq \widetilde{Y}_{d,\mathcal{T}_N(\omega)} - Y_d \xrightarrow[N\to\infty]{\mathcal{L}^2} 0$.

We note that Assumptions 2 and 3′ can be fulfilled by well-known metamodeling techniques under mild conditions. For instance, Kohler et al. (2003) showed that kernel smoothing is strongly universally consistent under some assumptions, i.e., $\mathbb{E}_X(|\widetilde{Y}_{m,\mathcal{T}_N} - Y_m|^2) \xrightarrow[N\to\infty]{a.s.} 0$ and $\mathbb{E}_X(|\widetilde{Y}_{d,\mathcal{T}_N} - Y_d|^2) \xrightarrow[N\to\infty]{a.s.} 0$ with respect to $(\Omega_Z, \mathcal{F}_Z, \mathbb{P}_Z)$. Rather than examining a particular metamodeling technique (e.g., kernel smoothing, Gaussian process), we focus on the asymptotic analysis of the metamodel-based estimators under some sufficient conditions on the metamodels. We first examine $\widetilde{S}^{X_u}$ defined in (11).

**Proposition 3** Under Assumptions 1 to 3, $\widetilde{S}^{X_u} \xrightarrow[N\to\infty]{} S^{X_u}$ for almost every $\omega \in \Omega_Z$.

*Proof.* Under Assumption 2 and by the Cauchy–Schwarz inequality, we have

$$\left|\mathbb{E}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right) - \mathbb{E}_X\left(Y_m + c\right)\right| \le \left(\mathbb{E}_X\left(\left|\delta_{m,\mathcal{T}_N(\omega)} - c\right|^2\right)\right)^{\frac{1}{2}} \xrightarrow[N\to\infty]{} 0.$$

It follows that $\mathbb{E}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right) \xrightarrow[N\to\infty]{} \mathbb{E}_X\left(Y_m\right) + c$. On the other hand, by the continuity of the $L_2$ norm,

$$\left|\left(\mathbb{E}_X\left(\widetilde{Y}^2_{m,\mathcal{T}_N(\omega)}\right)\right)^{\frac{1}{2}} - \left(\mathbb{E}_X\left((Y_m + c)^2\right)\right)^{\frac{1}{2}}\right| \le \left(\mathbb{E}_X\left(\left(\delta_{m,\mathcal{T}_N(\omega)} - c\right)^2\right)\right)^{\frac{1}{2}} \xrightarrow[N\to\infty]{} 0,$$

which implies that $\mathbb{E}_X\left(\widetilde{Y}^2_{m,\mathcal{T}_N(\omega)}\right) \xrightarrow[N\to\infty]{} \mathbb{E}_X\left((Y_m + c)^2\right)$. Hence,

$$\mathrm{Var}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right) = \mathbb{E}_X\left(\widetilde{Y}^2_{m,\mathcal{T}_N(\omega)}\right) - \mathbb{E}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right)^2 \xrightarrow[N\to\infty]{} \mathbb{E}_X\left((Y_m + c)^2\right) - \mathbb{E}_X\left((Y_m + c)\right)^2 = \mathrm{Var}_X\left(Y_m\right).$$

Also, we have

$$\mathbb{E}_{X_u}\left(\left|\mathbb{E}_{X_{-u}}\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}-Y_m-c\mid X_u\right)\right|^2\right)=\int_{X_u}\left|\mathbb{E}_{X_{-u}}\left(\delta_{m,\mathcal{T}_N(\omega)}-c\mid X_u\right)\right|^2 d\mu_{X_u}$$
$$\leq\int_{X_u}\mathbb{E}_{X_{-u}}\left(\left|\delta_{m,\mathcal{T}_N(\omega)}-c\right|^2\mid X_u\right)d\mu_{X_u}=\mathbb{E}_X\left(\left|\delta_{m,\mathcal{T}_N(\omega)}-c\right|^2\right)\xrightarrow[N\to\infty]{}0,$$

which yields

$$\mathbb{E}_{X_{-u}}\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\mid X_u\right)\xrightarrow[N\to\infty]{\mathcal{L}^2}\mathbb{E}_{X_{-u}}\left(Y_m\mid X_u\right)+c, \tag{12}$$

and hence,

$$\mathrm{Var}_{X_u}\left(\mathbb{E}_{X_{-u}}\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\mid X_u\right)\right)\xrightarrow[N\to\infty]{}\mathrm{Var}_{X_u}\left(\mathbb{E}_{X_{-u}}\left(Y_m\mid X_u\right)\right).$$

Since $\mathrm{Var}_X\left(Y_m\right)+\mathbb{E}_X\left(Y_d\right)>0$, according to Assumption 3 and the quotient law for convergent sequences, for almost every $\omega\in\Omega_Z$, $\widetilde{S}^{X_u}=\mathrm{Var}_{X_u}\left(\mathbb{E}_{X_{-u}}\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\mid X_u\right)\right)/\left(\mathrm{Var}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(\widetilde{Y}_{d,\mathcal{T}_N(\omega)}\right)\right)$ converges to $S^{X_u}=\mathrm{Var}_{X_u}\left(\mathbb{E}_{X_{-u}}\left(Y_m\mid X_u\right)\right)/\left(\mathrm{Var}_X\left(Y_m\right)+\mathbb{E}_X\left(Y_d\right)\right)$ as $N\to\infty$. □

To analyze the term $\sqrt{N}(\widetilde{S}^{X_u}_{\mathcal{T}_N}-\widetilde{S}^{X_u})$ (respectively $\sqrt{N}(\widetilde{T}^{X_u}_{\mathcal{T}_N}-\widetilde{S}^{X_u})$) related to the first (resp. second) metamodel-based estimator in the decomposition shown in (11), we stipulate the following assumption.

**Assumption 4** There exist $s_1,s_2>0$ and $C>0$ such that for almost every $\omega\in\Omega_Z$, $\mathbb{E}_X\left(|\widetilde{Y}_{m,\mathcal{T}_N(\omega)}|^{4+s_1}\right)<C$ and $\mathbb{E}_X\left(|\widetilde{Y}_{d,\mathcal{T}_N(\omega)}|^{2+s_2}\right)<C$, $\forall N\in\mathbb{N}^+$.

**Proposition 4** Under Assumptions 2, 3′, and 4, for almost every $\omega\in\Omega_Z$, $\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)}-\widetilde{S}^{X_u}\right)\xrightarrow[N\to\infty]{d}\mathcal{N}(0,\sigma_S^2)$, where $\sigma_S^2$ is defined in (7).

*Proof.* Define

$$\widetilde{U}_{\mathcal{T}_N(\omega),i}:=\left(\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega),i}-\mathbb{E}_X\left(\widetilde{Y}_m\right)\right)\left(\widetilde{Y}^X_{m,\mathcal{T}_N(\omega),i}-\mathbb{E}_X\left(\widetilde{Y}_m\right)\right),\widetilde{Y}_{m,\mathcal{T}_N(\omega),i}-\mathbb{E}_X\left(\widetilde{Y}_m\right),\right.$$
$$\left.\widetilde{Y}^X_{m,\mathcal{T}_N(\omega),i}-\mathbb{E}_X\left(\widetilde{Y}_m\right),\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega),i}-\mathbb{E}_X\left(\widetilde{Y}_m\right)\right)^2,\widetilde{Y}_{d,\mathcal{T}_N(\omega),i}\right),$$

and $\overline{\widetilde{U}}_{\mathcal{T}_N(\omega)}:=N^{-1}\sum_{i=1}^N\widetilde{U}_{\mathcal{T}_N(\omega),i}$. By Assumption 4, there exist $s'>0$ and $C'>0$ such that for almost every $\omega\in\Omega_Z$, $\mathbb{E}_X\left(\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^{2+s'}\right)<C'$, $\forall N\in\mathbb{N}^+$. Then, we have

$$\mathbb{E}_X\left(\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^2\mathbf{1}\left\{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|>\varepsilon\sqrt{N}\right\}\right)\xrightarrow[N\to\infty]{}0,\forall\varepsilon>0,$$

and

$$\mathbb{E}_X\left(\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^2\mathbf{1}\left\{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|>\varepsilon\sqrt{N}\right\}\right)=\mathbb{E}_X\left(\frac{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^{2+s'}}{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^{s'}}\mathbf{1}\left\{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|>\varepsilon\sqrt{N}\right\}\right)\leq\frac{C'}{\varepsilon^{s'}N^{s'/2}},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Therefore, for each $i$, $\left\{\left\|\widetilde{U}_{\mathcal{T}_N(\omega),i}\right\|^2\right\}_{N\geq1}$ is uniformly integrable. By Assumptions 2 and 3′, $\widetilde{U}_{\mathcal{T}_N(\omega),i}\xrightarrow[N\to\infty]{\mathbb{P}}U_i$ (recall (8)), hence the same convergence holds true in $L^2$. As a result, the covariance matrices of $\widetilde{U}_{\mathcal{T}_N(\omega),i}$ converge to $\Gamma$. The rest of the proof follows by applying the Delta method as shown in the proof of Proposition 1. □

Similarly, we have the following result for $\sqrt{N}\left(\widetilde{T}^{X_u}_{\mathcal{T}_N} - \widetilde{S}^{X_u}\right)$ in (11). The proof of Proposition 5 is in the same vein as that of Proposition 4 and is omitted for the sake of brevity.

**Proposition 5** Under Assumptions 2, 3′, and 4, for almost every $\omega \in \Omega_{\mathbf{Z}}$, $\sqrt{N}\left(\widetilde{T}^{X_u}_{\mathcal{T}_N(\omega)} - \widetilde{S}^{X_u}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma_T^2)$, where $\sigma_T^2$ is defined in (9).

Finally, Theorem 1 gives a set of sufficient conditions for establishing the asymptotic normality of the two metamodel-based estimators given in (3) and (4). In addition to Assumptions 2, 3′, and 4, we note that the convergence rates of the metamodels to the true mean and variance functions also play an important role.

**Theorem 1** Define

$$C_{\delta,\mathcal{T}_N(\omega)} := 2\left(\mathrm{Var}_X\left(Y_m\right)\right)^{\frac{1}{2}}\left(\mathrm{Corr}\left(Y_m,\delta_{m,\mathcal{T}_N(\omega)}\right) - \frac{\mathrm{Var}_X\left(Y_m\right)}{\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)}\cdot\mathrm{Corr}\left(Y_m,Y_m^X\right)\cdot\mathrm{Corr}\left(Y_m,\delta_{m,\mathcal{T}_N(\omega)}\right)\right)$$
$$+\left(\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)\right)^{\frac{1}{2}}\left(\mathrm{Corr}\left(\delta_{m,\mathcal{T}_N(\omega)},\delta^X_{m,\mathcal{T}_N(\omega)}\right) - \frac{\mathrm{Var}_X\left(Y_m\right)}{\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)}\cdot\mathrm{Corr}\left(Y_m,Y_m^X\right)\right),$$

for almost every $\omega\in\Omega_Z$, where $\mathrm{Corr}\left(A,B\right) = \mathrm{Cov}\left(A,B\right)/\left(\mathrm{Var}\left(A\right)\mathrm{Var}\left(B\right)\right)^{1/2}$, given any $L^2$ random variables $A$ and $B$ of nonzero variance, and $\delta^X_{m,\mathcal{T}_N(\omega)} := \widetilde{Y}^X_{m,\mathcal{T}_N(\omega)} - Y^X_m$.

Assume that $C_{\delta,\mathcal{T}_N(\omega)}$ does not converge to 0 as $N\to\infty$ and Assumptions 2, 3′ and 4 are fulfilled.

1. Suppose $\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right) = o\left(N^{-1}\right)$ and $\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right) = o\left(N^{-1}\right)$ for almost every $\omega\in\Omega_Z$, then for $\forall x\in\mathbb{R}$,

$$\mathbb{P}_Z\left(\omega\in\Omega_Z : \limsup_{N\to\infty}\left|\mathbb{P}_X\left(\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)} - \widetilde{S}^{X_u}\right)/\sigma_S \le x\right) - \Phi(x)\right| > \varepsilon\right) = 0, \forall\varepsilon>0, \tag{13}$$

   where $\Phi(x)$ is the standard normal cumulative density function.
2. Suppose $C_{\delta,\mathcal{T}_N(\omega)}$ converges to a constant $C\neq 0$ and there exists $\gamma\in\mathbb{R}$ so that $\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right) = (CN)^{-1}\gamma + o\left(N^{-1}\right)$ and $\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right) = o\left(N^{-1}\right)$ for almost every $\omega\in\Omega_Z$, then there is a constant $\gamma'$ such that for $\forall x\in\mathbb{R}$,

$$\mathbb{P}_Z\left(\omega\in\Omega_Z : \limsup_{N\to\infty}\left|\mathbb{P}_X\left(\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)} - \widetilde{S}^{X_u} - \gamma'\right)/\sigma_S \le x\right) - \Phi(x)\right| > \varepsilon\right) = 0, \forall\varepsilon>0. \tag{14}$$

The convergence in (13) and (14) holds true for the second metamodel-based estimator $\widetilde{T}^{X_u}_{\mathcal{T}_N(\omega)}$, with $\sigma_S$ replaced by $\sigma_T$.

*Proof.* We focus on the proof regarding the first metamodel-based estimator $\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)}$, as the proof regarding $\widetilde{T}^{X_u}_{\mathcal{T}_N(\omega)}$ can be given in the same vein and hence is omitted. Recall the decomposition in (11). Since $\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)} - \widetilde{S}^{X_u}\right) \xrightarrow[N\to\infty]{d} \mathcal{N}(0,\sigma_S^2)$ in $(\Omega_X,\mathcal{F}_X,\mathbb{P}_X)$ for a fixed $\omega\in\Omega_Z$ according to Proposition 4, if $\sqrt{N}\left(\widetilde{S}^{X_u} - S^{X_u}\right)$ goes to some constant $\kappa$, then for the fixed $\omega\in\Omega_Z$, $\sqrt{N}\left(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)} - S^{X_u}\right) \xrightarrow[N\to\infty]{d}$

$\mathcal{N}(\kappa,\sigma_S^2)$ in $(\Omega_X,\mathcal{F}_X,\mathbb{P}_X)$. Regarding $\sqrt{N}\left(\widetilde{S}^{X_u}-S^{X_u}\right)$, we have

$$\begin{aligned}\widetilde{S}^{X_u}-S^{X_u} &= \frac{\mathrm{Cov}\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)},\widetilde{Y}^X_{m,\mathcal{T}_N(\omega)}\right)}{\mathrm{Var}_X\left(\widetilde{Y}_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(\widetilde{Y}_{d,\mathcal{T}_N(\omega)}\right)}-\frac{\mathrm{Cov}\left(Y_m,Y_m^X\right)}{\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)}\\ &= \frac{\mathrm{Cov}\left(Y_m,Y_m^X\right)+2\,\mathrm{Cov}\left(Y_m,\delta^X_{m,N}\right)+\mathrm{Cov}\left(\delta_{m,\mathcal{T}_N(\omega)},\delta^X_{m,\mathcal{T}_N(\omega)}\right)}{\mathrm{Var}_X\left(Y_m\right)+2\,\mathrm{Cov}\left(Y_m,\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(Y_d\right)+\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)}-\frac{\mathrm{Cov}\left(Y_m,Y_m^X\right)}{\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)}\\ &= \frac{\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)^{\frac{1}{2}}C_{\delta,\mathcal{T}_N(\omega)}-\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1}\mathrm{Cov}\left(Y_m,Y_m^X\right)\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)}{\mathrm{Var}_X\left(Y_m\right)+2\,\mathrm{Cov}\left(Y_m,\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(Y_d\right)+\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)}. \end{aligned} \tag{15}$$

Under the assumption that $C_{\delta,\mathcal{T}_N(\omega)}$ converges to some $C\neq 0$, the denominator of the right-hand side of (15) follows as

$$\begin{aligned}&\mathrm{Var}_X\left(Y_m\right)+2\,\mathrm{Cov}\left(Y_m,\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(Y_d\right)+\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)\\ \leq&\,\mathrm{Var}_X\left(Y_m\right)+\mathbb{E}_X\left(Y_d\right)+2\left(\mathrm{Var}_X\left(Y_m\right)\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)\right)^{\frac{1}{2}}+\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)+\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)\\ =&\,\mathrm{Var}_X\left(Y_m\right)+\mathbb{E}_X\left(Y_d\right)+o(1).\end{aligned}$$

Hence, Equation (15) can be written as

$$\widetilde{S}^{X_u}-S^{X_u}=\frac{C_{\delta,\mathcal{T}_N(\omega)}\,\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)^{\frac{1}{2}}-\mathrm{Var}_{X,\varepsilon}\left(\mathcal{Y}\right)^{-1}\mathrm{Cov}\left(Y_m,Y_m^X\right)\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)}{\mathrm{Var}_X\left(Y_m\right)+\mathbb{E}_X\left(Y_d\right)+o(1)}.$$

If $\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)=o(N^{-1})$ and $\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)=o\left(N^{-1}\right)$, we have $\sqrt{N}\,\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)^{\frac{1}{2}}=o\left(1\right)$ and $\sqrt{N}\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)=o\left(N^{-1/2}\right)$, thus $\sqrt{N}(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)}-S^{X_u})=o\left(1\right)$ and for almost every $\omega\in\Omega_{\mathbf{Z}}$, $\sqrt{N}(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)}-S^{X_u})\xrightarrow[N\to\infty]{d}\mathcal{N}(0,\sigma_S^2)$. Hence, the convergence takes place in the product space $(\Omega_Z\times\Omega_X,\sigma(\mathcal{F}_Z\times\mathcal{F}_X),\mathbb{P}_Z\otimes\mathbb{P}_X)$, which leads to (13).

If $\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)=(CN)^{-1}\gamma+o(N^{-1})$ and $\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)=o\left(N^{-1}\right)$, we have $\sqrt{N}\,\mathrm{Var}_X\left(\delta_{m,\mathcal{T}_N(\omega)}\right)^{1/2}\xrightarrow[N\to\infty]{}\sqrt{\gamma/C}$ and $\sqrt{N}\mathbb{E}_X\left(\delta_{d,\mathcal{T}_N(\omega)}\right)\xrightarrow[N\to\infty]{}0$, thus there exists $\gamma'\in\mathbb{R}$ such that for almost every $\omega\in\Omega_Z$, $\sqrt{N}(\widetilde{S}^{X_u}-S^{X_u})\xrightarrow[N\to\infty]{}\gamma'$, and $\sqrt{N}(\widetilde{S}^{X_u}_{\mathcal{T}_N(\omega)}-S^{X_u})\xrightarrow[N\to\infty]{d}\mathcal{N}(\gamma',\sigma_S^2)$. The resulting convergence takes place in the product space $(\Omega_Z\times\Omega_X,\sigma(\mathcal{F}_Z\times\mathcal{F}_X),\mathbb{P}_Z\otimes\mathbb{P}_X)$, which results in (14). The proof is complete. □

## 4 NUMERICAL EVALUATION

In this section, we numerically evaluate the efficiency of the Sobol' index estimators given in (3) to (6) and verify the theoretical results. We consider the Ishigami function which is a classical example for evaluating global sensitivity analysis approaches (Ishigami and Homma 1990; Marrel et al. 2012):

$$Y=f\left(X_1,X_2,X_3\right)=\sin\left(X_1\right)+7\sin\left(X_2\right)^2+0.1X_3^4\sin\left(X_1\right), \tag{16}$$

where $X_i$'s are independent and uniformly distributed in $[-\pi,\pi]$, $i=1,2,3$. To make model (16) a stochastic one, we treat $X_1$ and $X_2$ as the input variables and $X_3$ as the random variable that incurs stochastic noise. The true mean and variance functions follow from (16) as

$$Y_m(x_1,x_2)=\left(1+\frac{\pi^4}{50}\right)\sin(x_1)+(\sin(x_2))^2,\ Y_d(x_1,x_2)=\pi^8\left(\frac{1}{900}-\frac{1}{2500}\right)(\sin(x_1))^2,\ \text{for } x_i\in[-\pi,\pi].$$

We are interested in obtaining point estimates and confidence interval estimates for the first-order Sobol' indices of $X_1$ and $X_2$. The true values are available in this case, which are respectively $S^{X_1} = 0.3139$ and $S^{X_2} = 0.4424$.

***Experimental settings.*** Obtaining the joint metamodel-based Sobol' index estimators given in (3) and (4) requires metamodel construction and MC sampling to evaluate the metamodels. Regarding the metamodel construction, we adopt the following iterative procedure proposed by Marrel et al. (2012) for estimating the mean and variance functions. Specifically, we first generate a training sample of size $n$ via Latin hypercube sampling (LHS) and build a standard Gaussian process (GP) model (denoted as $GP_{m,1}$) for estimating the mean function. Next, taking the squared residuals based on $GP_{m,1}$, we construct a metamodel (denoted as $\hat{V}_1$) for approximating the variance function. We then construct a heteroscedastic GP model (denoted as $GP_{m,2}$) for the mean function estimation, with the noise variances being estimated by $\hat{V}_1$. Finally, the squared residuals are calculated based on $GP_{m,2}$ to construct another metamodel (denoted as $\hat{V}_2$) for the ultimate variance function estimation. We adopt the resulting $GP_{m,2}$ and $\hat{V}_2$ as $\widetilde{Y}_{m,\mathcal{T}_N}$ and $\widetilde{Y}_{d,\mathcal{T}_N}$ in the subsequent Sobol' index estimation. We consider two variants of this iterative procedure in terms of constructing the variance metamodels $\hat{V}_1$ and $\hat{V}_2$, via either GP modeling (referred to as "Variant 1") or kernel smoothing (referred to as "Variant 2"). Variant 2 seems to enhance numerical stability in implementation. For MC evaluations, we use LHS to draw a sample of size $N$ for model evaluations. Similar designs such as Sobol' quasi-random sequences are also suggested by Saltelli et al. (2010) for MC-based Sobol' index estimation.

***Performance metrics.*** To assess the accuracy of the metamodels $\widetilde{Y}_{m,\mathcal{T}_N}$ and $\widetilde{Y}_{d,\mathcal{T}_N}$, we use the predictivity coefficient $Q_2$ (Marrel et al. 2012). For a given metamodel $\widetilde{Y}$, $Q_2(\widetilde{Y}) := 1 - \sum_{i=1}^N (Y_i - \widetilde{Y}_i)^2 / \sum_{i=1}^N \left(N^{-1}\sum_{i=1}^N Y_i - Y_i\right)^2$, where recall that $N$ denotes the MC sample size, $\widetilde{Y}_i$ denotes the metamodel prediction at input $X_i$, and $Y_i$ is the corresponding true function value. The closer $Q_2(\widetilde{Y})$ to 1, the higher the accuracy of the metamodel $\widetilde{Y}$.

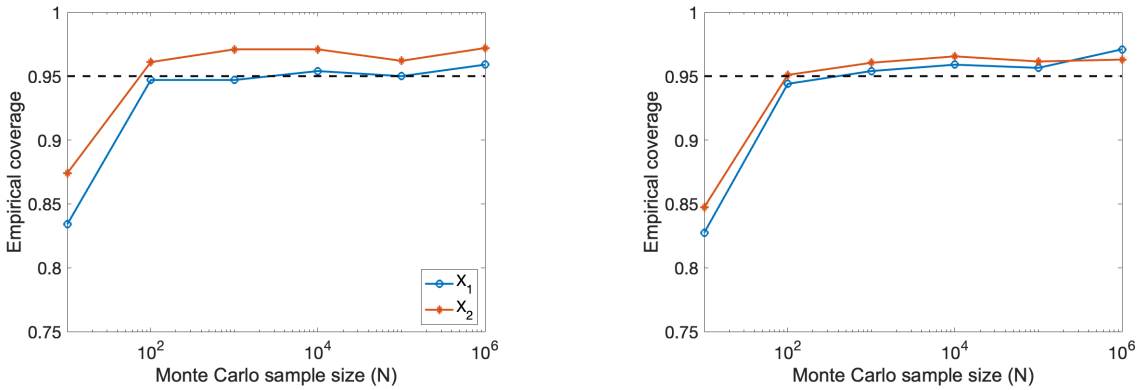
For performance evaluation, we perform $R$ independent macro-replications. The root mean squared error (RMSE), standard deviation, and bias of Sobol' index estimators given in (3) and (4) are calculated across the macro-replications. Specifically, RMSE $:= \sqrt{R^{-1}\sum_{i=r}^R \left(\widetilde{S}_r^{X_u} - S^{X_u}\right)^2}$, where $\widetilde{S}_r^{X_u}$ is a given estimator of $S^{X_u}$ obtained on the $r$th macro-replication. To assess the asymptotic normality, we examine the empirical coverage of the asymptotic confidence interval (referred to as CI) with the target level set to 0.95. The confidence interval $CI_r$ based on the two estimators on the $r$th macro-replication is given by $\widetilde{S}_{\mathcal{T}_N,r}^{X_u}$ (resp. $\widetilde{T}_{\mathcal{T}_N,r}^{X_u}$) $\pm 1.96\widetilde{\sigma}_{S,r}$ (resp. $\widetilde{\sigma}_{T,r}$) $/\sqrt{N}$, where $\widetilde{S}_{\mathcal{T}_N,r}^{X_u}$ and $\widetilde{T}_{\mathcal{T}_N,r}^{X_u}$ denote the joint metamodel-based estimators, and $\widetilde{\sigma}_{S,r}$ and $\widetilde{\sigma}_{T,r}$ are the MC-based estimators of $\sigma_S$ and $\sigma_T$ obtained on the $r$th macro-replication; to ease notation, we write $\widetilde{\sigma}_S$ and $\widetilde{\sigma}_T$ hereinafter. The empirical coverage is calculated as $R^{-1}\sum_{r=1}^R \mathbf{1}\left\{S^{X_u} \in CI_r\right\}$.

***Results.*** We first examine the two estimators constructed using the true mean and variance functions, $S_N^{X_u}$ and $T_N^{X_u}$, given in (5) and (6). The results are obtained based on $R = 2000$ macro-replications. Figure 1 shows the empirical coverage of the CIs built based on Propositions 1 and 2. We see that as the MC sample size $N$ increases, for $S_N^{X_u}$ and $T_N^{X_u}$ $(u = 1,2)$, the empirical coverage meets and slightly overshoots the target level. The results corroborate Propositions 1 and 2. Table 1 displays the CI widths (rescaled by $\sqrt{N}$) and the RMSE of $S_N^{X_u}$ and $T_N^{X_u}$ for $u = 1,2$. Since the rescaled CI widths are equal to $3.92\widetilde{\sigma}_S$ (resp. $3.92\widetilde{\sigma}_T$), we see from Table 1 that the Sobol' index of $X_1$, estimated by the second estimator $T_N^{X_1}$, has a smaller variance, while the first estimator $S_N^{X_2}$ is better at estimating the index of $X_2$. We have the same observation regarding the RMSEs of the two estimators. Therefore, neither estimator dominates the other, and one can adopt different estimators for estimating Sobol' indices of different input variables for higher statistical accuracy.

For evaluating the joint metamodel-based Sobol' index estimators, we perform $R = 100$ macro-replications. Table 2 presents the average $Q_2$ values across the macro-replications and the RMSEs of

Table 1: The confidence interval widths (rescaled by $\sqrt{N}$) and the RMSEs of $S_N^{X_u}$ and $T_N^{X_u}$ given in (5) and (6) obtained with an MC sample size $N = 10^6$.

| | $S_N^{X_1}$ | $S_N^{X_2}$ | $T_N^{X_1}$ | $T_N^{X_2}$ |
|---|---|---|---|---|
| Rescaled $CI$ width | 2.39 | 2.15 | 2.06 | 2.47 |
| RMSE ($\times 10^{-4}$) | 6.1 | 5.5 | 4.8 | 5.8 |




(a) $S_N^{X_u}$ (b) $T_N^{X_u}$

Figure 1: The empirical coverage of the asymptotic CIs for $S_N^{X_u}$ and $T_N^{X_u}$ as a function of the MC sample size $N$.

the first joint metamodel-based Sobol' index estimator $\widetilde{S}_{\mathcal{T}_N}^{X_1}$ obtained using the metamodels constructed via the two variants of the iterative fitting procedure. We observe that the variance metamodel obtained via Variant 1 has higher accuracy than that obtained via Variant 2, and the accuracy of the mean metamodels obtained via the two variants is comparable. Furthermore, the RMSEs of the resulting joint metamodel-based Sobol' index estimators are also comparable. This indicates that the mean metamodel is likely to play a dominating role in the joint metamodel-based Sobol' index estimation. We note that, as the training sample size $n$ increases, the heteroscedastic GP model produced by Variant 1 for the mean function estimation becomes numerically unstable, rendering the Sobol' index estimation unreliable. Hence, we focus on Variant 2 in the rest of the study.

Table 2: Comparisons of $Q_2(\widetilde{Y}_{m,\mathcal{T}_N})$, $Q_2(\widetilde{Y}_{d,\mathcal{T}_N})$, and RMSE of $\widetilde{S}_{\mathcal{T}_N}^{X_1}$ produced by the two variants of the iterative fitting procedure.

| $N$ | $n$ | $Q_2(\widetilde{Y}_{m,\mathcal{T}_N})$ | | $Q_2(\widetilde{Y}_{d,\mathcal{T}_N})$ | | RMSE of $\widetilde{S}_{\mathcal{T}_N}^{X_1}$ | |
|---|---|---|---|---|---|---|---|
| | | Variant 1 | Variant 2 | Variant 1 | Variant 2 | Variant 1 | Variant 2 |
| 50 | 1000 | 0.99 | 0.99 | 0.88 | 0.47 | 0.082 | 0.087 |
| 100 | 1000 | 0.99 | 0.99 | 0.90 | 0.48 | 0.061 | 0.058 |
| 500 | 1000 | 0.99 | 0.99 | 0.89 | 0.48 | 0.033 | 0.027 |
| 1000 | 1000 | 0.99 | 0.99 | 0.90 | 0.49 | 0.026 | 0.021 |

Table 3 shows the point estimation accuracy and the variability of the first metamodel-based estimator $\widetilde{S}_{\mathcal{T}_N}^{X_u}$ given in (3) for $u = 1, 2$. Since the results of the second estimator $\widetilde{T}_{\mathcal{T}_N}^{X_u}$ lead to similar conclusions, we omit them for the sake of brevity. We see from Table 3 that, for a fixed training sample size $n$, increasing $N$ does not always lower the bias, as the accuracy of the metamodels is constrained by the given training sample. In contrast, the RMSE and the standard deviation of the joint metamodel-based estimator decrease

with $N$ in most cases. While one may be inclined to increase $N$ to improve the point estimation accuracy, we highlight that, once $N$ becomes very large, the law of diminishing marginal returns emerges. To reduce the RMSE to a given level, a slight increase in the training sample size $n$ can be more computationally efficient. We do not go into the details to economize on space.

Table 3: The RMSE, the standard deviation (std), and the bias of $\widetilde{S}^{X_u}_{\mathcal{T}_N}$ under different combinations of $(N,n)$.

| $N$ | $n$ | $\widetilde{S}^{X_1}_{\mathcal{T}_N}$ | | | $\widetilde{S}^{X_2}_{\mathcal{T}_N}$ | | |
|---|---|---|---|---|---|---|---|
| | | RMSE | std | bias | RMSE | std | bias |
| 150 | 300 | 0.052 | 0.052 | 0.002 | 0.055 | 0.052 | 0.018 |
| 500 | 300 | 0.039 | 0.039 | 0.004 | 0.046 | 0.042 | 0.018 |
| 800 | 300 | 0.036 | 0.034 | 0.002 | 0.041 | 0.040 | 0.010 |
| 250 | 500 | 0.040 | 0.040 | 0.003 | 0.046 | 0.045 | 0.010 |
| 500 | 500 | 0.035 | 0.034 | 0.007 | 0.036 | 0.033 | 0.015 |
| 800 | 500 | 0.028 | 0.028 | 0.004 | 0.035 | 0.034 | 0.009 |
| 500 | 1000 | 0.027 | 0.027 | 0.002 | 0.030 | 0.027 | 0.012 |
| 800 | 1000 | 0.028 | 0.028 | 0.003 | 0.030 | 0.028 | 0.010 |
| 1000 | 1000 | 0.021 | 0.021 | 0.000 | 0.023 | 0.023 | 0.004 |
| 800 | 2000 | 0.021 | 0.021 | 0.000 | 0.023 | 0.022 | 0.003 |
| 1000 | 2000 | 0.021 | 0.021 | 0.002 | 0.023 | 0.022 | 0.007 |
| 1200 | 2000 | 0.020 | 0.020 | 0.002 | 0.018 | 0.017 | 0.006 |

Finally, we investigate the empirical coverage of the asymptotic confidence intervals obtained based on Theorem 1. Table 4 shows the combinations of $(N,n)$ that help achieve the target coverage level. We see that, for the Sobol' index of $X_1$, both asymptotic CIs based on estimators $\widetilde{S}^{X_1}_{\mathcal{T}_N}$ and $\widetilde{T}^{X_1}_{\mathcal{T}_N}$ achieve the target coverage level using $N \approx 0.5n$. For the Sobol' index of $X_2$, the asymptotic CI based on $\widetilde{S}^{X_2}_{\mathcal{T}_N}$ achieves the target coverage level using $N \approx 0.2n$ while that based on $\widetilde{T}^{X_2}_{\mathcal{T}_N}$ achieves the target level using $N \approx 0.1n$. Without showing details, we mention some important observations made throughout the experiment. First, as $n$ and $N$ increase, the empirical coverage of the asymptotic CIs approaches the target level, which corroborates Theorem 1. Second, increasing or decreasing $N$ does not always improve the empirical coverage, and the relationship between $n$ and $N$ is crucial to make the empirical coverage meets the target level. Our results echo those of Janon et al. (2014), who demonstrated that different metamodeling techniques require different combinations of $(N,n)$ for the asymptotic CIs constructed to reach a prescribed target coverage level.

Table 4: The empirical coverage of the asymptotic CIs based on Thoerem 1 under different combinations of $(N,n)$. The values in parentheses in Column "$N$" are those that lead to the corresponding empirical coverages closer to the target level at 0.95.

| $n$ | $\widetilde{S}^{X_1}_{\mathcal{T}_N}$ | | $\widetilde{S}^{X_2}_{\mathcal{T}_N}$ | | $\widetilde{T}^{X_1}_{\mathcal{T}_N}$ | | $\widetilde{T}^{X_2}_{\mathcal{T}_N}$ | |
|---|---|---|---|---|---|---|---|---|
| | $N$ | coverage | $N$ | coverage | $N$ | coverage | $N$ | coverage |
| 300 | 150 | 0.96 | 60 | 0.91 | 150 | 0.92 | 30 | 0.94 |
| 500 | 250 | 0.95 | 100 | 0.96 | 250 (100) | 0.9 (0.93) | 50 | 0.94 |
| 1000 | 500 | 0.97 | 200 | 0.97 | 500 | 0.94 | 100 | 0.95 |
| 1500 | 750 (650) | 0.89 (0.94) | 300 | 0.98 | 750 | 0.96 | 150 | 0.91 |
| 2000 | 1000 | 0.94 | 400 | 0.93 | 1000 | 0.93 | 200 | 0.96 |

In summary, while increasing the MC sample size $N$ for obtaining the joint metamodel-based estimators is computationally convenient, our results suggest that setting $N$ to an extremely large value given a fixed training sample is ineffective for improving the point estimation accuracy and the empirical coverage of the asymptotic confidence intervals.

## ACKNOWLEDGMENTS

This paper is based upon work supported by the National Science Foundation [IIS-1849300] and NSF CAREER [CMMI-1846663].

## REFERENCES

Castellan, G., A. Cousien, and V. C. Tran. 2020. "Non-parametric Adaptive Estimation of Order 1 Sobol' Indices in Stochastic Models, with an Application to Epidemiology". *Electronic Journal of Statistics* 14(1):50–81.

Gamboa, F., A. Janon, T. Klein, A. Lagnoux, and C. Prieur. 2016. "Statistical Inference for Sobol Pick-freeze Monte Carlo Method". *Statistics* 50(4):881–902.

Hart, J. L., A. Alexanderian, and P. A. Gremaud. 2017. "Efficient Computation of Sobol' Indices for Stochastic Models". *SIAM Journal on Scientific Computing* 39(4):A1514–A1530.

Ishigami, T., and T. Homma. 1990. "An Importance Quantification Technique in Uncertainty Analysis for Computer Models". In *The Proceedings of the First International Symposium on Uncertainty Modeling and Analysis*, 398–403.

Janon, A., T. Klein, A. Lagnoux, M. Nodet, and C. Prieur. 2014. "Asymptotic Normality and Efficiency of Two Sobol Index Estimators". *ESAIM: Probability and Statistics* 18:342–364.

Janon, A., M. Nodet, and C. Prieur. 2014. "Uncertainties Assessment in Global Sensitivity Indices Estimation from Metamodels". *International Journal for Uncertainty Quantification* 4(1):21–36.

Kohler, M., A. Krzyżak, and H. Walk. 2003. "Strong Consistency of Automatic Kernel Regression Estimates". *Annals of the Institute of Statistical Mathematics* 55(2):287–308.

Marrel, A., B. Iooss, S. Da Veiga, and M. Ribatet. 2012. "Global Sensitivity Analysis of Stochastic Computer Models with Joint Metamodels". *Statistics and Computing* 22(3):833–847.

Marrel, A., B. Iooss, B. Laurent, and O. Roustant. 2009. "Calculations of Sobol Indices for the Gaussian Process Metamodel". *Reliability Engineering & System Safety* 94(3):742–751.

Mazo, G. 2021. "A Trade-off between Explorations and Repetitions for Estimators of Two Global Sensitivity Indices in Stochastic Models Induced by Probability Measures". *SIAM/ASA Journal on Uncertainty Quantification* 9(4):1673–1713.

Saltelli, A., and P. Annoni. 2010. "How to Avoid a Perfunctory Sensitivity Analysis". *Environmental Modelling & Software* 25(12):1508–1517.

Saltelli, A., P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. 2010. "Variance Based Sensitivity Analysis of Model Output. Design and Estimator for the Total Sensitivity Index". *Computer Physics Communications* 181(2):259–270.

Sobol', I. M. 1990. "On Sensitivity Estimation for Nonlinear Mathematical Models". *Matematicheskoe Modelirovanie* 2(1):112–118.

Storlie, C. B., L. P. Swiler, J. C. Helton, and C. J. Sallaberry. 2009. "Implementation and Evaluation of Nonparametric Regression Procedures for Sensitivity Analysis of Computationally Demanding Models". *Reliability Engineering & System Safety* 94(11):1735–1763.

Tarantola, S., D. Gatelli, S. Kucherenko, W. Mauntz et al. 2007. "Estimating the Approximation Error When Fixing Unessential Factors in Global Sensitivity Analysis". *Reliability Engineering & System Safety* 92(7):957–960.

## AUTHOR BIOGRAPHIES

**JINGTAO ZHANG** is a Ph.D. candidate in the Grado Department of Industrial and Systems Engineering at Virginia Tech. His research interests include design and analysis of stochastic simulation experiments and simulation optimization. His email address is jingtaozhang@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 xchen6@vt.edu and her web page is https://sites.google.com/vt.edu/xi-chen-ise/home.

**RUOCHEN WANG** is a Ph.D. student in the Grado Department of Industrial and Systems Engineering at Virginia Tech. His research interests include queueing theory, sensitivity analysis, and optimization in healthcare. His email address is rcwangise@vt.edu.