Proceedings of the 2025 Winter Simulation Conference
E. Azar, A. Djanatliev, A. Harper, C. Kogler, V. Ramamohan, A. Anagnostou, and S. J. E. Taylor, eds.

A MODEL-FREE, PARTITION-BASED APPROACH TO ESTIMATING SOBOL' INDICES FROM EXISTING DATASETS

Jingtao Zhang[1] and Xi Chen[1]

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

ABSTRACT

This paper investigates a model-free, partition-based method for estimating Sobol' indices using existing datasets, addressing the limitations of traditional variance-based global sensitivity analysis (GSA) methods that rely on designed experiments. We provide a theoretical analysis of the bias, variance, and mean squared error (MSE) associated with the partition-based estimator, exploring the effects of the sample size of the dataset and the number of partition bins on its performance. Furthermore, we propose a data-driven approach for determining the optimal number of bins to minimize the MSE. Numerical experiments demonstrate that the proposed partition-based method outperforms state-of-the-art GSA techniques.

1 INTRODUCTION

Global sensitivity analysis (GSA) plays a crucial role in understanding and optimizing the behavior of complex systems. Among various GSA methods, Sobol' indices—variance-based measures of global sensitivity—are widely used to quantify the influence of input variables on model outputs (Sobol' 1990). These indices have been applied across diverse domains, including environmental modeling (Nossent et al. 2011), engineering design (Hübler 2020), epidemiology (Castellan et al. 2020), and climate science (Miftakhova 2021).

The pick-freeze scheme is a widely used technique for estimating Sobol' indices in GSA (Saltelli et al. 2010; Janon et al. 2014). It involves freezing one input variable (or a subset of variables) while randomly sampling the remaining inputs. By comparing the model outputs from these mixed input configurations, the approach quantifies the contribution of each input variable or group of input variables to the output variance, thereby enabling the estimation of Sobol' indices of interest. Although the pick-freeze method is known for its efficiency and robustness, it has two main limitations. First, it relies on a carefully designed experiment to perform model evaluations. While this is typically feasible in computational studies, it becomes impractical in real-world scenarios where conducting designed experiments is either infeasible or prohibitively expensive. Second, the computational cost of the pick-freeze approach scales linearly with the dimensionality of the input space to provide reliable Sobol' index estimates for all input variables. For complex systems with high-dimensional input spaces, this can lead to significant computational overhead.

Recent studies have proposed several methods for performing GSA with existing datasets, including partition-based approaches (Plischke et al. 2013; Zhai et al. 2014; Borgonovo et al. 2016), rank-based techniques (Gamboa et al. 2022; Klein and Rochet 2024), and nearest-neighbor methods (Devroye et al. 2018; Broto et al. 2020). In this work, we provide an in-depth investigation of the partition-based approach. While partition-based estimators have demonstrated empirical effectiveness in GSA with existing datasets, their theoretical properties remain underexplored. Additionally, the methodology for selecting the optimal number of partition bins remains underdeveloped, despite its recognized importance (Borgonovo et al. 2016). Zhai et al. (2014) proposed a partition selection scheme to minimize the estimator's variance, but they did not analyze the bias. Borgonovo et al. (2016) and Antoniano-Villalobos et al. (2020) numerically investigated the impact of the number of partition bins on the estimator's bias and mean squared error (MSE); however, their findings lack sufficient theoretical support. To address these gaps, we present a comprehensive study of model-free, partition-based Sobol' index estimation. We provide a theoretical analysis of how both the sample size and the number of bins affect the estimator's bias and variance. Additionally, we propose a data-driven approach for selecting the optimal number of partition bins for performing GSA on a given dataset. Our numerical experiments demonstrate that this approach consistently identifies the near-optimal number of bins and that the resulting partition-based estimator outperforms existing rank-based and nearest-neighbor methods.

The remainder of this paper is organized as follows. Section 2 reviews the definition of the Sobol' index and the partition-based approach for its estimation. Section 3 presents the theoretical analysis of the partition-based estimator. Section 4 proposes a data-driven approach for selecting the optimal number of partition bins for a given dataset. Section 5 presents numerical studies. Finally, Section 6 concludes the paper.

2 PARTITION-BASED METHOD FOR SOBOL' INDEX ESTIMATION

Consider a computational model $\mathscr{Y} = f(\mathbf{X})$, where f is a real-valued function, and $\mathbf{X} := \{X_1, X_2, \dots, X_p\} \in \mathscr{X}$ denotes the p-dimensional input vector, with $\mathscr{X} \subseteq \mathbb{R}^p$ representing the input space. Let $[p] := \{1,,2,\dots,p\}$. Given a subset $\mathbf{u} \subseteq [p]$, define $\mathbf{X_u}$ as the subvector of $\mathbf{X}$ corresponding to the indices in $\mathbf{u}$, and let $\mathbf{X_{-u}} := \mathbf{X} \setminus \mathbf{X_u}$ denote its complement. The Sobol' index of $\mathbf{X_u}$ is defined as:

$$S^{\mathbf{u}} := \frac{\mathrm{Var}\left(\mathbb{E}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right)}{\mathrm{Var}\left(\mathscr{Y}\right)} = \frac{\mathbb{E}\left(\mathbb{E}^2\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) - \left(\mathbb{E}\left(\mathscr{Y}\right)\right)^2}{\mathrm{Var}\left(\mathscr{Y}\right)} \ . \tag{1}$$

The Sobol' index $S^{\mathbf{u}}$ quantifies the contribution of $\mathbf{X_u}$ to the variance of the output $\mathscr{Y}$. The value of $S^{\mathbf{u}}$ ranges in $[0,1]$, with a larger value indicating a greater influence of $\mathbf{X_u}$ on the variability of $\mathscr{Y}$. In particular, when $\mathbf{u} = \{i\}$ and $\mathbf{X_u}$ consists of a single input X_i, S^i is known as the first-order Sobol' index of X_i, for each $i \in [p]$ (Sobol' 2001).

Let $\mathscr{D} = \{(\mathbf{X}_1, \mathscr{Y}_1), (\mathbf{X}_2, \mathscr{Y}_2), \dots, (\mathbf{X}_n, \mathscr{Y}_n)\}$ denote a given dataset of size n, where $\mathbf{X}_i = (\mathbf{X}_{\mathbf{u},i}, \mathbf{X}_{-\mathbf{u},i})$ denotes the ith input vector, $\mathbf{X}_{\mathbf{u},i}$ and $\mathbf{X}_{-\mathbf{u},i}$ denote the subset of entries in $\mathbf{X}_i$ indexed by $\mathbf{u}$ and $-\mathbf{u}$, respectively, and $\mathscr{Y}_i$ represents the corresponding model output. Our goal is to estimate the Sobol' index $S^{\mathbf{u}}$ given in (1) from $\mathscr{D}$. While estimating $\mathbb{E}(\mathscr{Y})$ and $\mathrm{Var}(\mathscr{Y})$ in (1) is relatively straightforward, estimating $\mathbb{E}\left(\mathbb{E}^2\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right)$ is more challenging. Classical methods based on the pick-freeze scheme (Saltelli et al. 2010; Janon et al. 2014) and nested simulation (Gordy and Juneja 2010) require designed experiments to run the computational model for generating outputs, which is not applicable when only existing datasets are available. To address this limitation, we consider a model-free, partition-based estimator for estimating $g := \mathbb{E}\left(\mathbb{E}^2\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right)$. We begin by noting the following identity:

$$\mathbb{E}\left(\mathbb{E}^2\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) = \mathbb{E}\left(\mathscr{Y}\mathbb{E}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) \ . \tag{2}$$

The right-hand side (RHS) of (2) can be estimated by $n^{-1}\sum_{i=1}^{n} \mathscr{Y}_i \widehat{f}(\mathbf{X}_{\mathbf{u},i})$, where $\widehat{f}(\mathbf{X}_{\mathbf{u},i})$ denotes an estimator for $\mathbb{E}\left(\mathscr{Y} \mid \mathbf{X_u} = \mathbf{X}_{\mathbf{u},i}\right)$. To construct $\widehat{f}(\mathbf{X}_{\mathbf{u},i})$, we partition the support of $\mathbf{X_u}$ into H non-overlapping equiprobable bins $\{B_1, B_2, \dots, B_H\}$, i.e., $\mathbf{X_u}$ falls into each bin with equal probability. Given each input vector $\mathbf{X}_i = (\mathbf{X}_{\mathbf{u},i}, \mathbf{X}_{-\mathbf{u},i})$ in the dataset $\mathscr{D}$, there exists a bin B_k such that $\mathbf{X}_{\mathbf{u},i} \in B_k$ for some $k \in [H]$. For notational convenience, for a given $i \in [n]$, let $B(\mathbf{X}_{\mathbf{u},i})$ denote the bin containing $\mathbf{X}_{\mathbf{u},i}$. Define $|B(\mathbf{X}_{\mathbf{u},i})| := \sum_{j=1}^{n} \mathbf{1}\left\{\mathbf{X}_{\mathbf{u},j} \in B(\mathbf{X}_{\mathbf{u},i})\right\}$ as the number of data points $(\mathbf{X}_j, \mathscr{Y}_j)$ whose input subvector $\mathbf{X}_{\mathbf{u},j}$ falls within the bin $B(\mathbf{X}_{\mathbf{u},i})$. We estimate $\mathbb{E}\left(\mathscr{Y} \mid \mathbf{X_u} = \mathbf{X}_{\mathbf{u},i}\right)$ by averaging the outputs whose corresponding $\mathbf{X}_{\mathbf{u},j}$'s fall within $B(\mathbf{X}_{\mathbf{u},i})$, i.e., $\widehat{f}(\mathbf{X}_{\mathbf{u},i}) = |B(\mathbf{X}_{\mathbf{u},i})|^{-1}\sum_{j:\mathbf{X}_{\mathbf{u},j} \in B(\mathbf{X}_{\mathbf{u},i})} \mathscr{Y}_j$. The partition-based estimator of g can be given by

$$\widehat{g} = \frac{1}{n}\sum_{i=1}^{n} \mathscr{Y}_i \widehat{f}(\mathbf{X}_{\mathbf{u},i}) = \frac{1}{n}\sum_{i=1}^{n} \mathscr{Y}_i |B(\mathbf{X}_{\mathbf{u},i})|^{-1} \sum_{j:\mathbf{X}_{\mathbf{u},j} \in B(\mathbf{X}_{\mathbf{u},i})} \mathscr{Y}_j \ , \tag{3}$$

and the partition-based estimator $S^{\mathbf{u}}$ follows as

$$\widehat{S}^{\mathbf{u}} = \frac{\widehat{g} - \left(n^{-1}\sum_{i=1}^{n}\mathscr{Y}_i\right)^2}{n^{-1}\sum_{i=1}^{n}\mathscr{Y}_i^2 - \left(n^{-1}\sum_{i=1}^{n}\mathscr{Y}_i\right)^2}\,. \tag{4}$$

The estimator $\widehat{S}^{\mathbf{u}}$ given in (4) addresses two key limitations of the traditional Sobol' index estimators built on the pick-freeze scheme. First, it does not require a designed experiment, making it suitable to be applied directly to a given dataset. More importantly, the sample size n for estimating all first-order Sobol' indices $S^1, S^2, \ldots, S^p$ is independent of the input-space dimensionality p. In contrast, the classical pick-freeze scheme requires a sample size that increases linearly with p to obtain reliable first-order Sobol' index estimates.

3 THEORETICAL ANALYSIS OF THE PARTITION-BASED ESTIMATOR

This section analyzes the theoretical properties of the partition-based estimator $\widehat{S}^{\mathbf{u}}$, defined in (4), for estimating first-order Sobol' indices, i.e., when $|\mathbf{u}| = 1$. We begin by analyzing the bias and variance of the estimator $\widehat{g}$ given in (3). The following assumptions are stipulated to facilitate analyses.

Assumption 1 Each component X_i of the input vector $\mathbf{X}$ is independently and uniformly distributed over $\mathscr{X}$.

Assumption 2 The variance of the model output $\mathscr{Y}$ is bounded.

Assumption 3 Let $\mathbf{x}_1$ and $\mathbf{x}_2$ be two realizations of the input vector $\mathbf{X_u}$. There exists a positive constant L such that

$$|\mathbb{E}(\mathscr{Y} \mid \mathbf{X_u} = \mathbf{x}_1) - \mathbb{E}(\mathscr{Y} \mid \mathbf{X_u} = \mathbf{x}_2)| \le L\|\mathbf{x}_1 - \mathbf{x}_2\|,$$

where $\|\cdot\|$ denotes the Euclidean norm.

Assumption 1 ensures that each bin contains the same expected number of data points. Assumption 2 guarantees the boundedness of the variance of the partition-based estimator $\widehat{g}$. Assumption 3 imposes a Lipschitz condition on the conditional mean, an assumption commonly adopted in Sobol' index inference (Klein and Rochet 2024). We are now in a position to derive the bias of $\widehat{g}$.

Proposition 1 Under Assumption 1, the bias of the partition-based estimator $\widehat{g}$ is given by

$$\text{bias}(\widehat{g}) = \frac{H}{n}\left(1-\left(1-\frac{1}{H}\right)^n\right)\cdot\mathbb{E}\left(\text{Var}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) - \frac{H}{n}\mathbb{E}\left(\frac{1}{2|B_1|}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_1}\left(\mathbb{E}(\mathscr{Y} \mid \mathbf{X}_{\mathbf{u},i}) - \mathbb{E}(\mathscr{Y} \mid \mathbf{X}_{\mathbf{u},j})\right)^2\right), \tag{5}$$

where B_1 denotes the first partition bin. Furthermore, if Assumption 3 also holds, then

$$\frac{H}{n}(1-(1-H^{-1})^n)\mathbb{E}\left(\text{Var}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) - \frac{L^2}{H^2} \le |\text{bias}(\widehat{g})| \le \frac{H}{n}(1-(1-H^{-1})^n)\mathbb{E}\left(\text{Var}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) + \frac{L^2}{H^2}\,. \tag{6}$$

The proof of Proposition 1 is deferred to Appendix A.1. We have the following remarks. First, the upper bound on bias$(\widehat{g})$ provided in (6) can be simplified to $c_3 \cdot n^{-1}H + c_4 \cdot H^{-2}$, where c_3 and c_4 are some positive constants—this form will be utilized in Section 4. Second, Proposition 1 reveals that the bias of the estimator $\widehat{g}$ depends on both the number of bins H and the sample size n. Equation (5) shows that the bias can be expressed as the difference between two nonnegative terms. The second term on the RHS of (5) reflects the average variation of the conditional expectation within an arbitrary bin (e.g., B_1), which intuitively decreases as the number of bins H increases. In contrast, the first term on the RHS increases with H. Consequently, when H is small, $\widehat{g}$ tends to be biased downward, while a larger value of H may lead to an upward bias. Notably, numerical observations made by Antoniano-Villalobos et al. (2020) can be well explained by Proposition 1.

To minimize $|\mathrm{bias}(\widehat{g})|$, the H value should be carefully selected to balance the two competing terms on the RHS of (6). Notably, the first term contains $\mathbb{E}\left(\mathrm{Var}\left(\mathscr{Y} \mid \mathbf{X_u}\right)\right) = \mathrm{Var}\left(\mathscr{Y}\right) \cdot \left(1 - S^{\mathbf{u}}\right)$. Therefore, when the Sobol' index of $\mathbf{X_u}$, $S^{\mathbf{u}}$, is large (i.e., $\left(1-S^{\mathbf{u}}\right)$ is small), a greater value of H may help reduce the bias. Conversely, when $S^{\mathbf{u}}$ is small, the first term on the RHS of (6) dominates, favoring a smaller H. As a result, estimating Sobol' indices for different input variables using the same dataset may require different choices of H. This theoretical insight echoes the observation made by Antoniano-Villalobos et al. (2020) that tailoring H to each input variable improves estimation performance. Furthermore, as the sample size n increases, the number of bins H should also increase to reduce the bias effectively. In particular, setting $H = \mathscr{O}(n^{1/3})$ is a practical strategy for minimizing the upper bound given in (6). This choice follows the recommendation of Borgonovo et al. (2016) and Antoniano-Villalobos et al. (2020), based on heuristic and empirical grounds; our analysis offers a formal justification. Next we analyze the variance of $\widehat{g}$.

Proposition 2 Under Assumptions 1 and 2, the variance of the partition-based estimator $\widehat{g}$ can be bounded as follows:

$$\mathrm{Var}\left(\widehat{g}\right) \leq c_1 \frac{1}{n} - c_2 \frac{1}{nH} \ , \tag{7}$$

where c_1 and c_2 are some positive constants.

The proof of Proposition 2 is deferred to Appendix A.2. Proposition 2 shows that $\mathrm{Var}\left(\widehat{g}\right)$ depends on both the sample size n and the number of bins H. Given n being fixed, increasing H typically results in a greater upper bound for $\mathrm{Var}\left(\widehat{g}\right)$. This is intuitive, since a larger H value means fewer data points in each bin, increasing the variability of the partition-based estimator. As the sample size n increases, its impact dominates that of the number of bins H. To reduce the variance, it is sufficient for the sample size n to approach infinity, with no specific requirement on H. This is in stark contrast to the bias of $\widehat{g}$, which vanishes only when both H and n tend to infinity.

The next result provides an upper bound on $\mathrm{MSE}(\widehat{g})$, which directly follows from Propositions 1 and 2.

Corollary 1 Under Assumptions 1, 2, and 3, the MSE of the partition-based estimator $\widehat{g}$ can be bounded as follows:

$$\mathrm{MSE}(\widehat{g}) \leq \left(c_3 \frac{H}{n} + c_4 \frac{1}{H^2}\right)^2 + c_1 \frac{1}{n} - c_2 \frac{1}{nH} \ , \tag{8}$$

where c_1 to c_4 are some positive constants.

Corollary 1 shows that setting $H = \mathscr{O}(n^{1/3})$ yields the optimal convergence rate $\mathscr{O}(n^{-1})$ for the upper bound on $\mathrm{MSE}(\widehat{g})$. To determine the convergence rate of $\widehat{S^{\mathbf{u}}}$, we note that the remaining terms on the RHS of (4), i.e., $(n^{-1}\sum_{i=1}^{n}\mathscr{Y}_i)^2$ and $n^{-1}\sum_{i=1}^{n}\mathscr{Y}_i^2$, are both $\mathscr{O}(n^{-1})$, which do not depend on H. Building on the MSE convergence analysis for ratio estimators by Zhang (2025), we can show that $\mathrm{MSE}(\widehat{S^{\mathbf{u}}})$ achieves a convergence rate of $\mathscr{O}(n^{-1})$ when $H = \mathscr{O}(n^{1/3})$. While $H = \mathscr{O}(n^{1/3})$ serves as a general guideline, choosing H in practice is more nuanced due to its dependence on constants c_1–c_4 in (8). The next section introduces a data-driven method for estimating the optimal H.

4 A DATA-DRIVEN APPROACH FOR CHOOSING THE NUMBER OF BINS

This section presents a data-driven method for setting the number of bins H to implement the partition-based method. We leverage the bias and variance bounds established in Propositions 1 and 2, which contain the constants c_1 through c_4. Inspired by Zhang et al. (2022), who developed a method for estimating constants in optimal budget allocation for nested simulation, we propose a data-driven approach for estimating the constants in our setting.

We begin by expressing the expectation and variance of $\widehat{g}$ using (6) and (7):

$$\mathbb{E}\left(\widehat{g}\right) = c_5 + c_3 \frac{H}{n} + c_4 \frac{1}{H^2} \ , \quad \operatorname{Var}\left(\widehat{g}\right) = c_1 \frac{1}{n} - c_2 \frac{1}{nH} \ , \tag{9}$$

where c_5 is a constant serving as an estimate of $\mathbb{E}\left(g\right)$. Given a fixed value of H, $\mathbb{E}\left(\widehat{g}\right)$ and $\operatorname{Var}\left(\widehat{g}\right)$ can be estimated via bootstrapping (Zhang et al. 2022). Let $\widehat{\mu}_B(H)$ and $\widehat{\sigma}_B^2(H)$ denote their corresponding bootstrap estimators, where B denotes the bootstrap sample size. Define the set $\mathscr{H} := \{H_1, H_2, \ldots, H_m\}$, which consists of m distinct values for the number of bins H. For each $H_i \in \mathscr{H}$, we obtain the corresponding $\widehat{\mu}_B(H_i)$ and $\widehat{\sigma}_B^2(H_i)$. We then perform a least-squares regression of the estimated bootstrap means $\{\widehat{\mu}_B(H_i)\}_{i=1}^m$ on the vectors $\{(1, n^{-1}H_i, H_i^{-2})^\top\}_{i=1}^m$ to estimate the constants c_3, c_4, and c_5, denoting the estimates by $\widehat{c}_3$, $\widehat{c}_4$, and $\widehat{c}_5$. Similarly, we conduct a least-squares regression of the estimated bootstrap variances $\{\widehat{\sigma}_B^2(H_i)\}_{i=1}^m$ on the vectors $\{(n^{-1}, -(nH_i)^{-1})^\top\}_{i=1}^m$ and provide $\widehat{c}_1$ and $\widehat{c}_2$. An appropriate value of H, $\widehat{H}^*$, is the solution to the following optimization program:

$$\widehat{H}^* := \underset{H \in [2, n/2]}{\arg\min} \left(\widehat{c}_3 \frac{H}{n} + \widehat{c}_4 \frac{1}{H^2} \right)^2 + \widehat{c}_1 \frac{1}{n} - \widehat{c}_2 \frac{1}{nH} \ . \tag{10}$$

Since the objective function in (10) is convex under mild conditions on $\widehat{c}_2$, $\widehat{c}_3$, and $\widehat{c}_4$, the solution is unique. We round the value of $\widehat{H}^*$ to the nearest integer and use it as the number of partition bins to build the partition-based estimator given in (4). Algorithm 1 outlines the detailed steps.

Algorithm 1 Data-driven approach for determining the number of partition bins

```
Input: dataset $\mathscr{D} = \{(\mathbf{X}_1, \mathscr{Y}_1), (\mathbf{X}_2, \mathscr{Y}_2), \ldots, (\mathbf{X}_n, \mathscr{Y}_n)\}$, set $\mathscr{H} = \{H_1, H_2, \ldots, H_m\}$, bootstrap sample size $B$
Output: the number of bins $\widehat{H}^*$
for each $H_i \in \mathscr{H}$ do
    for $b = 1$ to $B$ do
        Draw data points independently with replacement from $\mathscr{D}$ and get a bootstrap sample of size $n$, $\mathscr{D}_b^*$;
        Construct the estimator $\widehat{g}_b^*$ using $\mathscr{D}_b^*$ based on (3);
    end for
    Obtain the mean and variance estimates: $\widehat{\mu}_B(H_i) := B^{-1}\sum_{b=1}^B \widehat{g}_b^*$ and $\widehat{\sigma}_B^2(H_i) := B^{-1}\sum_{b=1}^B (\widehat{g}_b^* - \widehat{\mu}_B(H_i))^2$;
end for
Regress $\{\widehat{\mu}_B(H_i)\}_{i=1}^m$ on $\left\{(1, n^{-1}H_i, H_i^{-2})\right\}_{i=1}^m$ to obtain $\widehat{c}_3$, $\widehat{c}_4$, and $\widehat{c}_5$;
Regress $\{\widehat{\sigma}_B^2(H_i)\}_{i=1}^m$ on $\left\{(n^{-1}, -(nH_i)^{-1})\right\}_{i=1}^m$ to obtain $\widehat{c}_1$ and $\widehat{c}_2$;
Determine $\widehat{H}^*$ using (10) and round it to the nearest integer value.
```

5 NUMERICAL EVALUATIONS

This section evaluates the proposed partition-based approach. Subsection 5.1 details the numerical examples and the experimental setup. Subsection 5.2 examines the impact of the number of partition bins and the effectiveness of Algorithm 1. Subsection 5.3 compares the partition-based estimator with other Sobol' index estimation approaches using given datasets.

5.1 Numerical Examples and Experimental Settings

Ishigami function. The Ishigami function is a classical example for evaluating the performance of Sobol' index estimators (Ishigami and Homma 1990). The model is given by $\mathscr{Y} = \sin(X_1) + 7\sin(X_2)^2 + 0.1X_3^4 \sin(X_1)$, where the X_i's are independent and uniformly distributed over $[-\pi, \pi]$. We are interested in estimating the first-order Sobol' indices, whose true values are $S^1 = 0.3134$, $S^2 = 0.4424$, and $S^3 = 0$.

g-function. The p-dimensional g-function, widely used in the GSA literature (Owen 2013), is given by $\mathscr{Y} = \prod_{i=1}^{p} g_i(X_i)$, where $g_i(X_i) = (|4X_i - 2| + a_i)/(1 + a_i)$ and $a_i \geq 0$. Each X_i is independently and uniformly distributed over $[0, 1]$. In this example, we consider $p = 3$ with $a_1 = 19$, $a_2 = 9$, and $a_3 = 4$. The true first-order Sobol' indices are $S^1 = 0.0475$, $S^2 = 0.1898$, and $S^3 = 0.7594$.

For each numerical example, we use a given dataset of size $n = 500$. For each $H \in \{1, 2, \ldots, 100\}$, we conduct 1,000 independent macro-replications to estimate the MSE, bias, and variance of the partition-based estimators $\widehat{S}^1$, $\widehat{S}^2$, and $\widehat{S}^3$. The optimal number of partition bins, H^*, is defined as the value of H that minimizes the estimated MSE. We compare H^* with $\widehat{H}^*$ obtained by Algorithm 1 in Subsection 5.2 to evaluate the performance.

5.2 The Effect of H on the Performance of the Partition-based Estimator

We first investigate the impact of H on the performance of the partition-based estimator. Figure 1 shows the MSE, the squared bias, and the variance of the partition-based estimator as functions of H for the Ishigami function example. We observe that the variance exhibits only minor fluctuations as H varies, whereas the squared bias varies significantly as H changes. For the input variables with relatively large Sobol' indices (i.e., X^1 and X^2), the H^* values tend to be large. Conversely, for the input variable with small Sobol' index (i.e., X^3), the H^* value is relatively small. These observations corroborate our theoretical results presented in Propositions 1 and 2. Figure 2 presents the MSE, the squared bias, and the variance as functions of H for the g-function example, further confirming the impact of H observed in the Ishigami function example.

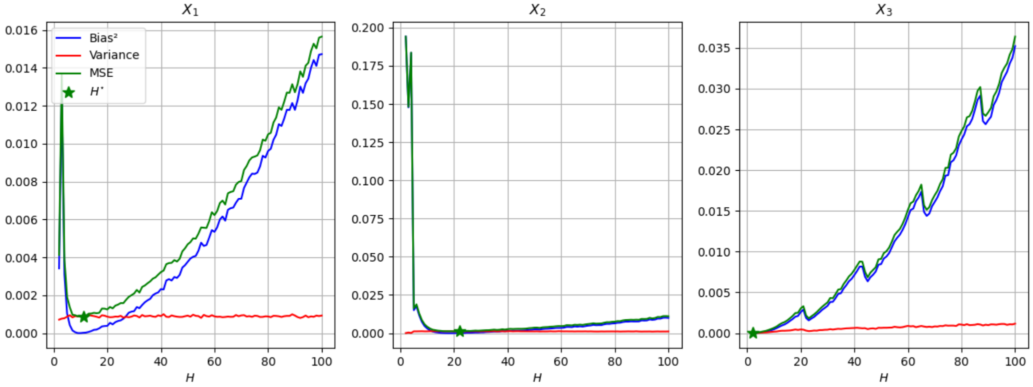

Figure 1: The Ishigami function example: MSE, squared bias, and variance of the partition-based estimator as functions of the number of bins H.

We next examine the effectiveness of Algorithm 1 in its choosing an appropriate value for the number of bins for estimating the first-order Sobol' indices. To implement Algorithm 1, we adopt $\mathscr{H} = \{10, 20, 30, 40, 50\}$ and set the bootstrap sample size to $B = 500$. To evaluate the algorithm's performance, we conduct 100 macro-replications and compare the $\widehat{H}^*$ value obtained on each macro-replication with the optimal value H^* identified in Figures 1 and 2. Figure 5 in Appendix B displays the $\widehat{H}^*$ values obtained by Algorithm 1 in both examples. In the Ishigami function example, the $\widehat{H}^*$ values used to

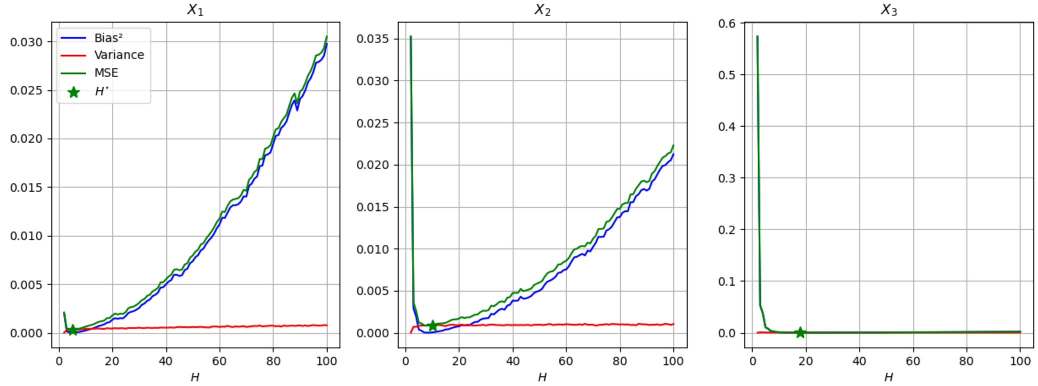

Figure 2: The g-function example: MSE, squared bias, and variance of the partition-based estimator as functions of the number of bins H.

estimate the Sobol' indices across all three dimensions yield MSE values comparable to those obtained using the H^* values. In the g-function example, although a few $\widehat{H}^*$ values selected for estimating S^1 result in significantly higher MSEs compared to H^*, the majority yield MSEs close to their respective optimal values. These results highlight the effectiveness of Algorithm 1.

5.3 Comparison of Different Sobol' Index Estimators

We evaluate the performance of the proposed partition-based estimator given in (4) by comparing it with two established approaches for estimating Sobol' indices from given datasets: the rank-based method (Gamboa et al. 2022; Klein and Rochet 2024) and the nearest-neighbor method (Devroye et al. 2018; Broto et al. 2020). Both competing methods require choosing specific parameter values: the number of neighbors for the nearest-neighbor estimator and the number of lags for the rank-based estimator. To ensure a fair comparison, we conduct 1,000 macro-replications to estimate the MSE, bias, and variance of each method, and select the parameter value that minimizes the estimated MSE for each.

Tables 1 summarizes the results obtained by the three approaches. In both examples, the partition-based estimator consistently outperforms the other two methods in estimating the first-order Sobol' indices corresponding to all input variables. Figure 3 summarizes the Sobol' index estimates in the Ishigami function example. The partition-based estimator yields lower bias than the other two methods in estimating S^1 and S^2, while maintaining comparable variance, and achieves the lowest bias and variance for S^3. Figure 4 presents the Sobol' index estimates for the g-function example, where the partition-based estimator consistently attains the lowest bias and variance across all first-order Sobol' indices.

6 CONCLUSION

This paper introduced a model-free, partition-based method with a data-driven approach for estimating Sobol' indices from a given dataset. Our theoretical analysis reveals how both the number of bins and the dataset size influence the estimator's MSE, leading to a data-driven approach for choosing an appropriate number of bins to use in practice. Numerical experiments demonstrate that the proposed approach reliably selects a near-optimal number of bins, and that the partition-based method achieves lower MSE in Sobol' index estimation compared to state-of-the-art rank-based and nearest-neighbor estimators. These results underscore the practical advantages of the proposed method for real-world sensitivity analysis.

Table 1: Comparison of the partition-based, nearest-neighbor, and rank-based estimators in the Ishigami function and g-function examples. The smallest value for each performance metric is highlighted in bold.

Sobol' index	**Method**	**Ishigami**			**g-function**		
		MSE	**Bias^2**	**Variance**	**MSE**	**Bias^2**	**Variance**
S^1	Partition	$\mathbf{7.97 \times 10^{-4}}$	$\mathbf{1.90 \times 10^{-7}}$	$\mathbf{7.97 \times 10^{-4}}$	$\mathbf{2.94 \times 10^{-4}}$	$\mathbf{2.00 \times 10^{-6}}$	$\mathbf{2.93 \times 10^{-4}}$
	Nearest-neighbor	8.93×10^{-4}	4.70×10^{-5}	8.46×10^{-4}	2.64×10^{-2}	4.00×10^{-4}	2.60×10^{-2}
	Rank	1.09×10^{-3}	6.70×10^{-5}	1.02×10^{-3}	7.72×10^{-3}	1.48×10^{-3}	6.24×10^{-3}
S^2	Partition	$\mathbf{1.09 \times 10^{-3}}$	$\mathbf{4.00 \times 10^{-8}}$	$\mathbf{1.09 \times 10^{-3}}$	$\mathbf{8.44 \times 10^{-4}}$	$\mathbf{4.00 \times 10^{-5}}$	$\mathbf{8.04 \times 10^{-4}}$
	Nearest-neighbor	1.27×10^{-3}	1.40×10^{-5}	1.25×10^{-3}	2.53×10^{-2}	1.01×10^{-3}	2.43×10^{-2}
	Rank	1.29×10^{-3}	1.06×10^{-4}	1.18×10^{-3}	1.09×10^{-2}	5.04×10^{-3}	5.85×10^{-3}
S^3	Partition	$\mathbf{1.10 \times 10^{-5}}$	$\mathbf{4.00 \times 10^{-6}}$	$\mathbf{7.00 \times 10^{-6}}$	$\mathbf{2.23 \times 10^{-4}}$	$\mathbf{7.00 \times 10^{-6}}$	$\mathbf{2.17 \times 10^{-4}}$
	Nearest-neighbor	1.89×10^{-4}	1.10×10^{-5}	1.78×10^{-4}	1.06×10^{-2}	1.57×10^{-3}	9.06×10^{-3}
	Rank	4.54×10^{-4}	8.20×10^{-5}	3.72×10^{-4}	2.15×10^{-2}	1.98×10^{-2}	1.63×10^{-3}

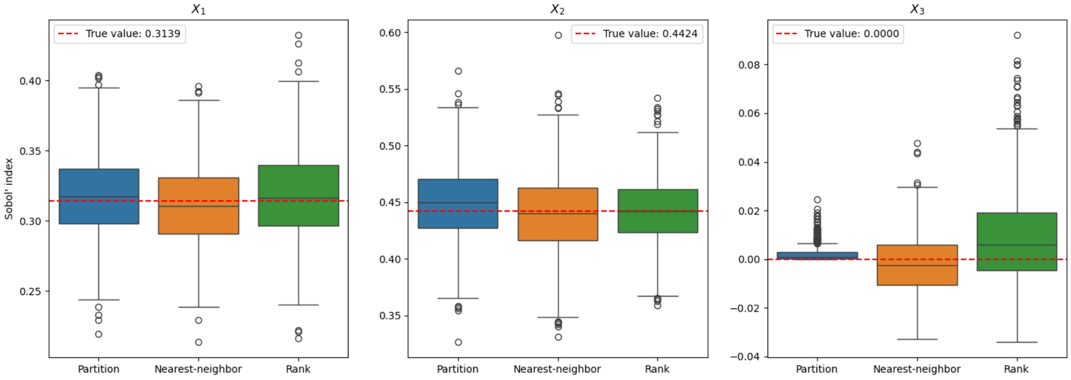

Figure 3: The Ishigami function example: summary of Sobol' index estimates obtained from 1,000 macro-replications using the partition-based, nearest-neighbor, and rank-based estimators.

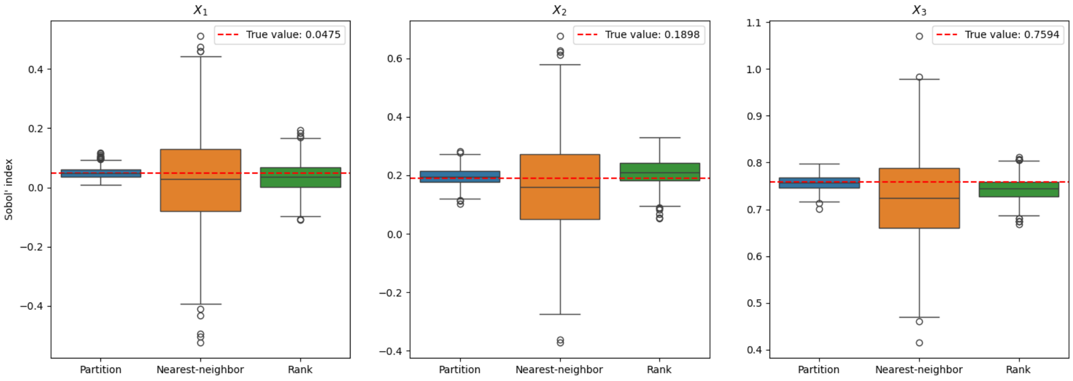

Figure 4: The g-function example: summary of Sobol' index estimates obtained from 1,000 macro-replications using the partition-based, nearest-neighbor, and rank-based estimators.

ACKNOWLEDGMENTS

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

A PROOF IN SECTION 3

A.1 Proof of Proposition 1

Proof. We first write the bias of $\widehat{g}$ given in (3) as follows:

$$\begin{aligned}
&\mathbb{E}\left(\widehat{g}-\mathbb{E}\left(\mathscr{Y}\mathbb{E}\left(\mathscr{Y}\mid\mathbf{X}_{\mathbf{u}}\right)\right)\right)=\mathbb{E}\left(\frac{1}{n}\sum_{i=1}^{n}\mathscr{Y}_i|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\sum_{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})}\mathscr{Y}_j-\frac{1}{n}\sum_{i=1}^{n}\mathscr{Y}_i\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)\\
=&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\mathscr{Y}_i^2\right)-\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\mathbb{E}\left(|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)\\
&+\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\mathbb{E}\left(|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\sum_{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i}),j\neq i}(\mathscr{Y}_j-\mathscr{Y}_i)\mid\mathbf{X}_{\mathbf{u},i}\right)\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)\\
=&\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\operatorname{Var}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)}{|B(\mathbf{X}_{\mathbf{u},i})|}\right)}_{(i)}+\underbrace{\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\mathscr{Y}_i}{|B(\mathbf{X}_{\mathbf{u},i})|}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}\mathscr{Y}_j-\frac{\mathscr{Y}_i}{|B(\mathbf{X}_{\mathbf{u},i})|}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)}_{(ii)}.
\end{aligned}$$

For term (i), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{1}{|B(\mathbf{X}_{\mathbf{u},i})|}\operatorname{Var}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)=\mathbb{E}\left(\frac{1}{|B(\mathbf{X}_{\mathbf{u},1})|}\operatorname{Var}\left(\mathscr{Y}_1\mid\mathbf{X}_{\mathbf{u},1}\right)\right)\\
=&\ \mathbb{E}\left(\mathbb{E}\left(\frac{1}{n_B}\operatorname{Var}\left(\mathscr{Y}_1\mid\mathbf{X}_{\mathbf{u},1}\right)\ \middle|\ |B(\mathbf{X}_{\mathbf{u},1})|=n_B\right)\right)=\sum_{n_B=1}^{n}\frac{1}{n_B}\mathbb{E}\left(\operatorname{Var}\left(\mathscr{Y}_1\mid\mathbf{X}_{\mathbf{u},1}\right)\right)\cdot\mathbb{P}(|B(\mathbf{X}_{\mathbf{u},1})|=n_B)\\
=&\ \sum_{n_B=1}^{n}\frac{1}{n_B}\binom{n-1}{n_B-1}H^{1-n_B}(1-H^{-1})^{n-n_B}\mathbb{E}\left(\operatorname{Var}\left(\mathscr{Y}_1\mid\mathbf{X}_{\mathbf{u},1}\right)\right)=\frac{H}{n}\left(1-(1-H^{-1})^n\right)\mathbb{E}\left(\operatorname{Var}\left(\mathscr{Y}_1\mid\mathbf{X}_{\mathbf{u},1}\right)\right)\ ,
\end{aligned}$$

where the second to last equality follows from the fact that $(|B(\mathbf{X}_{\mathbf{u},i})|-1)$ follows a Binomial distribution $\text{Binomial}(n-1,H^{-1})$.

For term (ii), we have

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\frac{\mathscr{Y}_i}{|B(\mathbf{X}_{\mathbf{u},i})|}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}\mathscr{Y}_j-\frac{\mathscr{Y}_i}{|B(\mathbf{X}_{\mathbf{u},i})|}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\right)\\
=&\ \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(\mathbb{E}\left(\mathscr{Y}_i\mid\mathbf{X}_{\mathbf{u},i}\right)\cdot\mathbb{E}\left(|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}(\mathscr{Y}_j-\mathscr{Y}_i)\mid\mathbf{X}_{\mathbf{u},i}\right)\right)
\end{aligned}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\left(|B(\mathbf{X}_{\mathbf{u},i})|^{-1}\sum_{\substack{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})\\ j\neq i}}\mathbb{E}\left(\mathscr{Y}_i \mid \mathbf{X}_{\mathbf{u},i}\right)\cdot\left(\mathbb{E}\left(\mathscr{Y}_j \mid \mathbf{X}_{\mathbf{u},i}\right)-\mathbb{E}\left(\mathscr{Y}_i \mid \mathbf{X}_{\mathbf{u},i}\right)\right)\right)$$

$$= -\frac{1}{n}\mathbb{E}\left(\sum_{k=1}^{H}(2|B_k|)^{-1}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_k}\left(\mathbb{E}\left(\mathscr{Y}_i \mid \mathbf{X}_{\mathbf{u},i}\right)-\mathbb{E}\left(\mathscr{Y}_j \mid \mathbf{X}_{\mathbf{u},j}\right)\right)^2\right) ,$$

Hence, we have

$$|\mathrm{bias}(\widehat{g})| = \frac{H}{n}\left(1-(1-H^{-1})^n\right)\mathbb{E}\left(\mathrm{Var}\left(\mathscr{Y}_1 \mid \mathbf{X}_{\mathbf{u},1}\right)\right)-\frac{H}{n}\mathbb{E}\left((2|B_1|)^{-1}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_1}\left(\mathbb{E}\left(\mathscr{Y}_i \mid \mathbf{X}_{\mathbf{u},i}\right)-\mathbb{E}\left(\mathscr{Y}_j \mid \mathbf{X}_{\mathbf{u},j}\right)\right)^2\right) .$$

If Assumption 3 holds, we further have

$$\begin{aligned}
&\frac{H}{n}\mathbb{E}\left((2|B_1|)^{-1}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_1}\left(\mathbb{E}\left(\mathscr{Y}_i \mid \mathbf{X}_{\mathbf{u},i}\right)-\mathbb{E}\left(\mathscr{Y}_j \mid \mathbf{X}_{\mathbf{u},j}\right)\right)^2\right)\\
\leq&\frac{H}{n}\mathbb{E}\left((2|B_1|)^{-1}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_1}L^2\left(\mathbf{X}_{\mathbf{u},i}-\mathbf{X}_{\mathbf{u},j}\right)^2\right)\\
\leq&\frac{H}{n}\mathbb{E}\left((2|B_1|)^{-1}\sum_{(i,j):\mathbf{X}_{\mathbf{u},i},\mathbf{X}_{\mathbf{u},j}\in B_1}\frac{L^2}{H^2}\right)\leq\frac{L^2}{H^2} ,
\end{aligned}$$

which yields (6).

□

A.2 Proof of Proposition 2

Proof. We first write the estimator $\widehat{g}$ given in (3) as follows:

$$\widehat{g} = \frac{1}{n}\sum_{i=1}^{n}\mathscr{Y}_i\frac{1}{|B(\mathbf{X}_{\mathbf{u},i})|}\sum_{j:\mathbf{X}_{\mathbf{u},j}\in B(\mathbf{X}_{\mathbf{u},i})}\mathscr{Y}_j = \frac{1}{n}\sum_{k=1}^{H}\sum_{i:\mathbf{X}_{\mathbf{u},i}\in B_k}\mathscr{Y}_i\frac{1}{n_k}\sum_{j:\mathbf{X}_{\mathbf{u},j}\in B_k}\mathscr{Y}_j = \frac{1}{n}\sum_{k=1}^{H}\frac{S_k^2}{n_k} ,$$

where B_k denotes the kth bin, $n_k := |B_k|$ is the number of data points in the kth bin, and $S_k := \sum_{i:\mathbf{X}_{\mathbf{u},i}\in B_k}\mathscr{Y}_i$.

Notice that $\mathrm{Var}\left(\widehat{g}\right) = n^{-2}\sum_{k=1}^{H}\mathrm{Var}\left(S_k^2/n_k\right)$. We begin by analyzing $\mathrm{Var}\left(S_k^2/n_k\right)$ for $k\in[H]$. Specifically, for a fixed k, we have the following decomposition:

$$\mathrm{Var}\left(\frac{S_k^2}{n_k}\right) = \mathrm{Var}\left(\mathbb{E}\left(\left.\frac{S_k^2}{n_k}\right| n_k\right)\right)+\mathbb{E}\left(\mathrm{Var}\left(\left.\frac{S_k^2}{n_k}\right| n_k\right)\right) .$$

On the one hand, for $i\geq 1$,

$$\begin{aligned}
\mathbb{E}\left(\left.\frac{S_k^2}{n_k}\right| n_k=i\right) =& n_k\cdot\mathbb{E}\left(\left.\frac{S_k^2}{n_k^2}\right| n_k=i\right) = n_k\cdot\mathrm{Var}\left(\left.\frac{S_k}{n_k}\right| n_k=i\right)+n_k\cdot\left(\mathbb{E}\left(\left.\frac{S_k}{n_k}\right| n_k=n\right)\right)^2\\
=& n_k\left(\frac{\sigma_k^2}{n_k}+\mu_k^2\right) = \sigma_k^2+n_k\mu_k^2 ,
\end{aligned}$$

where μ_k and σ_k^2 are the mean and variance of the outputs $\mathscr{Y}$ whose corresponding $\mathbf{X}_{\mathbf{u}}$ values fall within B_k, respectively. On the other hand, $\mathrm{Var}\left(n_k^{-1}S_k^2\middle|\, n_k=i\right)=n_k^2\,\mathrm{Var}\left(n_k^{-2}S_k^2\middle|\, n_k=i\right)$. The Taylor expansion yields $n_k^{-2}S_k^2=\mu_k^2+2\mu_k\left(n_k^{-1}S_k-\mu_k\right)+R(S_k)$, where $R(S_k)$ denotes the remainder satisfying $R(S_k)=\mathscr{O}(n_k^{-2})$. It follows that $n_k^2\,\mathrm{Var}\left(n_k^{-2}S_k^2\middle|\, n_k=n\right)\le 8n_k\mu_k^2\sigma_k^2+2\,\mathrm{Var}\left(R(S_k)\right)\le c\cdot n_k\mu_k^2\sigma_k^2$, where c is some positive constant, and the last inequality follows from $R(S_k)=\mathscr{O}(n_k^{-2})$. Hence, we have

$$\begin{aligned}\mathrm{Var}\left(\frac{S_k^2}{n_k}\right)&\le \mathrm{Var}\left(\sigma_k^2+n_k\mu_k^2\right)+\mathbb{E}\left(c\cdot n_K\mu_k^2\sigma_k^2\right)\\&=\mu_k^4\,\mathrm{Var}\left(n_k\right)+c\cdot\mu_k^2\sigma_k^2\mathbb{E}\left(n_k\right)=\mu_k^4np_k(1-p_k)+c\mu_k^2\sigma_k^2np_k\\&=\mu_k^4n\cdot\frac{1}{H}\left(1-\frac{1}{H}\right)+c\cdot\mu_k^2\sigma_k^2\cdot n\cdot\frac{1}{H}=\mu_k^4\left(\frac{n}{H}-\frac{n}{H^2}\right)+c\cdot\mu_k^2\sigma_k^2\cdot\left(\frac{n}{H}\right)\ .\end{aligned}$$

Hence, $\mathrm{Var}\left(\widehat{g}\right)\le n^{-1}\sum_{k=1}^{H}\mu_k^4\left(H^{-1}-H^{-2}\right)+c\cdot n^{-1}\sum_{k=1}^{H}\mu_k^2\sigma_k^2\cdot H^{-1}\le c_1n^{-1}-c_2(nH)^{-1}$, where c_1 and c_2 are some positive constants. □

B NUMBER OF PARTITION BINS OBTAINED FROM ALGORITHM 1

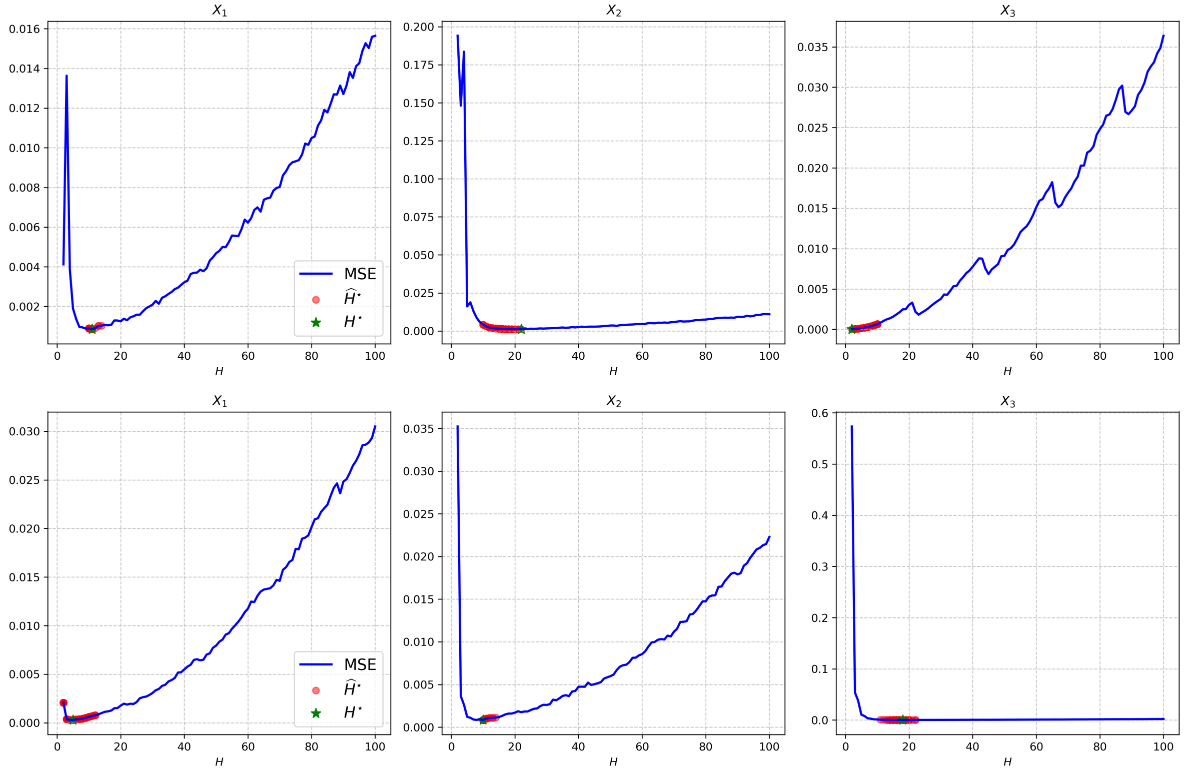

Figure 5: Comparisons of the $\widehat{H}^*$ values obtained by Algorithm 1 in the 100 macro-replications and the H^* value for each input variable in the Ishigami function example (top) and the g-function example (bottom).

REFERENCES

Antoniano-Villalobos, I., E. Borgonovo, and X. Lu. 2020. "Nonparametric Estimation of Probabilistic Sensitivity Measures". *Statistics and Computing* 30(2):447–467.

Borgonovo, E., G. B. Hazen, and E. Plischke. 2016. “A Common Rationale for Global Sensitivity Measures and Their Estimation”. *Risk Analysis* 36(10):1871–1895.

Broto, B., F. Bachoc, and M. Depecker. 2020. “Variance Reduction for Estimation of Shapley Effects and Adaptation to Unknown Input Distribution”. *SIAM/ASA Journal on Uncertainty Quantification* 8(2):693–716.

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.

Devroye, L., L. Györfi, G. Lugosi, and H. Walk. 2018. “A Nearest Neighbor Estimate of the Residual Variance”. *Electronic Journal of Statistics* 12(1):1752 – 1778.

Gamboa, F., P. Gremaud, T. Klein, and A. Lagnoux. 2022. “Global Sensitivity Analysis: A Novel Generation of Mighty Estimators Based on Rank Statistics”. *Bernoulli* 28(4).

Gordy, M. B., and S. Juneja. 2010. “Nested Simulation in Portfolio Risk Measurement”. *Management Science* 56(10):1833–1848.

Hübler, C. 2020. “Global Sensitivity Analysis for Medium-Dimensional Structural Engineering Problems Using Stochastic Collocation”. *Reliability Engineering & System Safety* 195:106749.

Ishigami, T., and T. Homma. 1990. “An Importance Quantification Technique in Uncertainty Analysis for Computer Models”. In *[1990] Proceedings. First International Symposium on Uncertainty Modeling and Analysis*, 398–403: Institute of Electrical and Electronics Engineers.

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.

Klein, T., and P. Rochet. 2024. “Efficiency of the Averaged Rank-Based Estimator for First Order Sobol Index Inference”. *Statistics & Probability Letters* 207:110015.

Miftakhova, A. 2021. “Global Sensitivity Analysis for Optimal Climate Policies: Finding What Truly Matters”. *Economic Modelling* 105:105653.

Nossent, J., P. Elsen, and W. Bauwens. 2011. “Sobol’ sensitivity Analysis of a Complex Environmental Model”. *Environmental Modelling & Software* 26(12):1515–1525.

Owen, A. B. 2013. “Better Estimation of Small Sobol’ Sensitivity Indices”. *ACM Transactions on Modeling and Computer Simulation* 23(2):1–17.

Plischke, E., E. Borgonovo, and C. L. Smith. 2013. “Global Sensitivity Measures from Given Data”. *European Journal of Operational Research* 226(3):536–550.

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. 2001. “Global Sensitivity Indices for Nonlinear Mathematical Models and Their Monte Carlo Estimates”. *Mathematics and Computers in Simulation* 55(1):271–280.

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

Zhai, Q., J. Yang, and Y. Zhao. 2014. “Space-partition Method for the Variance-based Sensitivity Analysis: Optimal Partition Scheme and Comparative Study”. *Reliability Engineering & System Safety* 131:66–82.

Zhang, J. 2025. *Advances in Sobol’ Index Estimation: Joint Metamodeling, Multilevel Monte Carlo Metamodeling, and Nested Simulation Techniques*. Ph.D. thesis, Virginia Tech.

Zhang, K., G. Liu, and S. Wang. 2022. “Technical Note—Bootstrap-based Budget Allocation for Nested Simulation”. *Operations Research* 70(2):1128–1142.

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, global sensitivity analysis, 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.