Statistical Uncertainty Analysis for Stochastic Simulation

Wei Xie
Rensselaer Polytechnic Institute

Barry L. Nelson
Northwestern University

Russell R. Barton
The Pennsylvania State University

Abstract

When we use simulation to evaluate the performance of a stochastic system, the simulation often contains input distributions estimated from real-world data; therefore, there is both simulation and input uncertainty in the performance estimates. Ignoring either source of uncertainty underestimates the overall statistical error. Simulation uncertainty can be reduced by additional computation (e.g., more replications). Input uncertainty can be reduced by collecting more real-world data, when feasible.

This paper proposes an approach to quantify overall statistical uncertainty when the simulation is driven by independent parametric input distributions; specifically, we produce a confidence interval that accounts for both simulation and input uncertainty by using a metamodel-assisted bootstrapping approach. The input uncertainty is measured via bootstrapping, an equation-based stochastic kriging metamodel propagates the input uncertainty to the output mean, and both simulation and metamodel uncertainty are derived using properties of the metamodel. A variance decomposition is proposed to estimate the relative contribution of input to overall uncertainty; this information indicates whether the overall uncertainty can be significantly reduced through additional simulation alone. Asymptotic analysis provides theoretical support for our approach, while an empirical study demonstrates that it has good finite-sample performance.

Keywords: Input uncertainty, confidence intervals, bootstrap, stochastic kriging, simulation output analysis, metamodel

1 Introduction

Stochastic simulation is used to characterize the behavior of complex systems that are driven by random input processes. By "input process" we mean a sequence of one or more random variables with a fully specified joint distribution. In this paper we consider independent and identically distributed (i.i.d.) input processes that are themselves mutually independent, which means that the input processes can be full specified by their individual marginal distributions.

The distributions of these input processes are often estimated from real-world data. Thus, a complete statistical characterization of stochastic system performance requires quantifying both simulation and input estimation errors.[1] There are robust methods that are adequate for quantifying simulation error for many practical problems. However, the impact of input estimation error (which we call "input uncertainty") can overwhelm the simulation error, as demonstrated in Barton and Schruben (2001); ignoring it may lead to unfounded confidence in the simulation assessment of system performance, which could be the basis for critical and expensive decisions. What we call input and simulation uncertainty are also known as epistemic and aleatory uncertainty, respectively; see for instance Kleijnen (2008). Our method accounts for both sources of error.

In this paper we address problems with parametric input distributions that are mutually independent with parameters that are estimated from a finite sample of real-world data. Of course, there exist practical problems in which the input processes are not independent, and there may also be significant uncertainty about the correct parametric distribution as well as its parameter values. Nevertheless, the case of i.i.d. input processes represented by a parametric distribution is prevalent in practice, making our contribution a useful step. We build on Barton et al. (2014), which proposed a metamodel-assisted bootstrapping approach that forms a confidence interval (CI) to account for the impact of input uncertainty when estimating the system's mean performance. In that paper, bootstrap resampling of the real-world data was used to approximate the input uncertainty, while a metamodel predicted the simulation's mean response at different parameter settings corresponding to bootstrap resampled data sets.

In a metamodel-assisted bootstrapping framework there are three types of error: the input estimation error, the simulation estimation error and the error in the metamodel itself. The latter two types of error are not easily separated, so we call their combined effect "metamodel uncertainty."

Barton et al. (2014) showed that metamodel uncertainty can be ignored when the simulation budget is not tight and an appropriate type of metamodel and experiment design are used; they provided a follow-up test to insure that the metamodel fit is close enough. In this setting their method yields a valid CI. However, if the true mean response surface is complex, especially for high-dimensional problems (i.e., many input distributions), and the computational budget is tight, then the impact of metamodel uncertainty can no longer be

[1] As with any mathematical or computer model, simulations are also subject to a host of non-statistical errors in abstraction; these are not considered in this paper.

ignored without underestimating the error, which manifests itself in a CI that is too short. Computationally intensive stochastic simulations are the norm for a number of fields: Spatial stochastic simulations, e.g., of oil reservoirs, can take hours for a single run, and depend on many stochastic parameters (Bangerth et al. 2006, Wang et al. 2012). Simulations of semiconductor manufacturing (Fowler and Rose 2004) and biological systems (Ghosh et al. 2011, Kastner et al. 2002) can be similarly intensive.

This paper is a significant enhancement of Barton et al. (2014). Here we propose an approach to form an interval estimate that accounts for *both* input and metamodel uncertainty in estimating a stochastic system's mean performance. When there is little metamodel uncertainty the new method performs like Barton et al. (2014), but it does not experience a degradation in coverage when metamodel uncertainty is significant.

When the statistical error measured by our CI is too large for the estimate to be useful, then the decision maker may want to know how the error can be reduced. Our approach leads naturally to a measure of the relative contribution of input to overall uncertainty that indicates whether the error can be reduced by an additional computational investment.

The next section describes other approaches to attack the input uncertainty problem and contrasts them with our method. This is followed in Section 3 by a formal description of the problem of interest and a brief review of the metamodel-assisted bootstrapping approach in Section 4. In Section 5 we provide an algorithm to build an interval estimator accounting for both input and metamodel uncertainty, and give a method to estimate their relative contributions. We then report results from an empirical study of a difficult problem in Section 6 and conclude the paper in Section 7. All proofs are in the Appendix.

2 Background

Various approaches to account for input uncertainty have been proposed. The Bayesian methods use the posterior distributions of the inputs given the real-world data to quantify the input distribution uncertainty, and the impact on the system mean is estimated by drawing samples from these posterior distributions and running simulations at each sample point (Chick 2001; Chick and Ng 2002; Zouaoui and Wilson 2003, 2004). This could be computationally expensive when the time for each simulation run is significant because simulations need to be run at a large number of posterior sample points. In addition, for each input prior distribution we need to derive a corresponding posterior distribution which might be nonstandard and complex.

A second approach is based on direct bootstrapping; it quantifies the impact of input uncertainty using bootstrap resampling of the input data, and runs simulations at each bootstrap resample point to estimate the impact on the system mean (Barton and Schruben 1993, 2001; Barton 2007; Cheng and Holland 1997). Compared with the Bayesian approach, the direct bootstrap can be adapted to any input process without additional analysis and it is suitable for complex and nonstandard input distributions. However, similar to the Bayesian approach, this method also runs simulations at each resample point. Since the number of bootstrap resample points to construct a CI is recommended to be a few thousand, the di-

rect bootstrapping method is also computationally expensive. More subtly, since the statistic that is bootstrapped is the random output of a simulation it is not a smooth function of the input data; this violates the asymptotic validity of the bootstrap.

Differing from the previous two approaches that estimate the system mean response at each sample point by running simulations, a third approach introduces an equation-based metamodel of the mean response (see Cheng and Holland (2004) and references therein). Specifically, it assumes that the parametric families of the inputs are known, uses maximum likelihood estimators (MLEs) of the unknown parameters, and represents input-parameter uncertainty by the large-sample normal distribution of the MLEs. This uncertainty is propagated to the output mean by a linear function of the parameters that is based on a Taylor series approximation. Since the metamodel can be constructed using simulation results from a small number of runs, this method does not need substantial computational effort. However, a metamodel based on a locally linear approximation is only appropriate when there is a large quantity of real-world data so that the MLEs locate in a small neighborhood of the true parameters with high probability; it is not suitable when the underlying response surface is highly non-linear and only a modest quantity of real-world data are available. In addition, the asymptotic normal approximation for the input distribution parameters can be poor with sample sizes encountered in some applications.

The metamodel-assisted bootstrapping approach introduced by Barton et al. (2014) addresses some of the shortcomings in the prior work. Compared with Cheng and Holland (2004), the bootstrap provides a more accurate approximation of the input uncertainty than the asymptotic normal distribution of the parameter estimates in many situations (Horowitz, 2001). Further, the use of a general-form metamodel provides higher fidelity than a locally linear approximation. Compared with Bayesian and direct bootstrap approaches, the use of a metamodel reduces the impact of simulation error on the accuracy of CIs and reduces the computational effort because it does not run simulations at a large number of sampled or resampled points; instead, an equation-based metamodel is constructed based on a designed experiment at a smaller number of parameter settings. In addition, employing a metamodel makes the bootstrap statistic a smooth function of the input data so that the asymptotic validity concerns faced by the direct bootstrap method disappear. The numerical results in Barton et al. (2014) provide evidence that metamodel-assisted bootstrapping is effective and superior to competitors *when there is little metamodel uncertainty*, motivating its extension in this paper to more general and complex input-uncertainty problems in which the impact of metamodel uncertainty can no longer be ignored. The end result is a robust method for quantifying statistical uncertainty.

3 Problem Description

To make the description of the input uncertainty problem and our solution to it easier to follow we will use the queueing network in Figure 1 as an example and return to it in our empirical study in Section 6. Consider estimating the steady state expected number of customers in this network. The interarrival times follow a gamma distribution, $A \sim \text{gamma}(\alpha_A, \beta_A)$, and

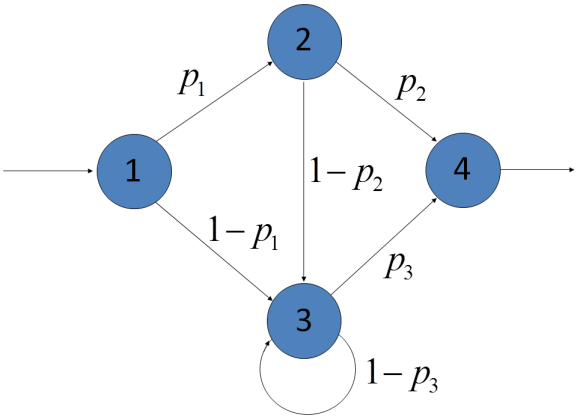

Figure 1: Queueing network example.

the service times at the ith station also follow a gamma distribution, $S_i \sim \text{gamma}(\alpha_{S_i}, \beta_{S_i})$. Customers finishing service at stations $1, 2, 3$ must make decisions about their next station. These routing decisions follow Bernoulli distributions $P_i \sim \text{Ber}(p_i), i = 1, 2, 3$. The parameters of the input distributions, α_A, β_A, $\{(\alpha_{S_i}, \beta_{S_i}), i = 1, 2, 3, 4\}$ and $\{p_i, i = 1, 2, 3\}$ are all unknown and estimated from real-world data. Notice that the inputs include both continuous and discrete distributions. Our goal is to build a CI that covers the steady-state expected number of customers in the system when the input parameters assume their true but unknown values. We assume that at these "true values" the system is in fact stable, and if we have enough real-world data (which we may not) then the simulation with estimated parameters will also be stable.

More generally, the stochastic simulation output is a function of random numbers and L independent input distributions $F \equiv \{F_1, F_2, \ldots, F_L\}$. For notation simplification, we do not explicitly include the random numbers. The output from the jth replication of a simulation with input distribution F can be written as

$$Y_j(F) = \mu(F) + \epsilon_j(F) \tag{1}$$

where $\mu(F) = \text{E}[Y_j(F)]$ denotes the unknown output mean and $\epsilon_j(F)$ represents the simulation error with mean zero. Notice that the simulation output depends on the choice of input distributions. The true "correct" input distributions, denoted by $F^c \equiv \{F_1^c, F_2^c, \ldots, F_L^c\}$, are unknown and are estimated from a finite sample of real-world data. Our goal is to quantify the impact of the statistical error by finding a $(1-\alpha)100\%$ CI $[Q_L, Q_U]$ such that

$$\Pr\{\mu(F^c) \in [Q_L, Q_U]\} = 1 - \alpha. \tag{2}$$

This is a relatively general statement of the problem which could encompass multivariate input distributions (i.e., if one or more of F_ℓ^c are multivariate) and also unknown distribution families.

However, in this paper we assume that the input distributions are univariate, the families are known, but the parameter values are not. Therefore, the input distributions F can be

completely characterized by the collection of parameters denoted by $\mathbf{x}$ with dimension d. We will define what we mean by "parameters" below. By abusing the notation, we rewrite Equation (1) as

$$Y_j(\mathbf{x}) = \mu(\mathbf{x}) + \epsilon_j(\mathbf{x}) \tag{3}$$

where $\mathbf{x} \in \Psi$ and $\Psi \equiv \{\mathbf{x} \in \Re^d$: the random variable $Y(\mathbf{x})$ is defined and $\mu(\mathbf{x})$ is finite$\}$ denotes the region of interest. The collection of true parameters is denoted by $\mathbf{x}_c$ and is assumed to be in the interior of Ψ. We also assume $\mu(\mathbf{x})$ is continuous for $\mathbf{x} \in \Psi$. Thus, our goal can be restated as finding a $(1-\alpha)100\%$ CI such that

$$\Pr\{\mu(\mathbf{x}_c) \in [Q_L, Q_U]\} = 1 - \alpha. \tag{4}$$

Let m_ℓ denote the number of i.i.d. real-world observations available from the ℓth input process $\mathbf{Z}_{\ell,m_\ell} \equiv \{Z_{\ell,1}, Z_{\ell,2}, \ldots, Z_{\ell,m_\ell}\}$ with $Z_{\ell,i} \overset{i.i.d}{\sim} F_\ell^c$, $i = 1, 2, \ldots, m_\ell$. Let $\mathbf{Z_m} = \{\mathbf{Z}_{\ell,m_\ell}, \ell = 1, 2, \ldots, L\}$ be the collection of samples from all L input distributions in F^c, where $\mathbf{m} = (m_1, m_2, \ldots, m_L)$. The real-world data are a particular realization of $\mathbf{Z_m}$, say $\mathbf{z}_\mathbf{m}^{(0)}$. Since the unknown input distributions are estimated from $\mathbf{z}_\mathbf{m}^{(0)}$, we assume the parameters are functions of $\mathbf{Z_m}$ denoted by $\mathbf{X_m} = \mathbf{X}(\mathbf{Z_m})$. Therefore, input uncertainty is fully captured by the sampling distribution of $\mu(\mathbf{X_m})$.

Since the underlying response surface $\mu(\cdot)$ is unknown, we approximate it by a metamodel fitted to a set of stochastic simulation outputs. Let $\widehat{\mu}(\cdot)$ denote the metamodel. Fitting the metamodel introduces a source of uncertainty in propagating the sampling distribution of $\mathbf{X_m}$ to the output mean: metamodel uncertainty. *The contributions of this paper are to construct an interval estimator that accounts for both input and metamodel uncertainty, and to quantify the relative impact of each contributor.*

4 Metamodel-Assisted Bootstrapping Framework

Barton et al. (2014) introduced the metamodel-assisted bootstrapping approach to quantify input uncertainty. We review it here.

The representation of the L input distributions plays an important role in the implementation of metamodel-assisted bootstrapping. Since this paper focuses on problems with independent parametric distributions having unknown parameters, F can be uniquely characterized by the corresponding collection of each distribution's parameters or, in many cases, its moments. The ℓth input distribution includes h_ℓ unknown parameters. Suppose that this h_ℓ-parameter distribution is uniquely specified by its first (finite) h_ℓ moments, which is true for the distributions that are most often used in stochastic simulation. The moments are chosen as the independent variables for the metamodel because when they are close, the corresponding distributions will be similar and therefore generate similar outputs. An extended argument for choosing moments instead of the natural distribution parameters as independent variables can be found in Barton et al. (2014). This characterization will not work for all distributions, including some so-called "heavy-tailed" distributions.

Let $\mathbf{x}_{[\ell]}$ denote an $h_\ell \times 1$ vector of the first h_ℓ moments for the ℓth input distribution and $d = \sum_{\ell=1}^{L} h_\ell$. By stacking $\mathbf{x}_{[\ell]}$ with $\ell = 1, 2, \ldots, L$ together, we have a $d \times 1$ dimensional independent variable $\mathbf{x}$. Notice that F is completely characterized by the collection of moments $\mathbf{x}$, and from here on $\mu(\cdot)$ will be a function of distribution moments organized in this way. Denote the true moments by $\mathbf{x}_c$.

For the queueing network in Figure 1, there are $L = 8$ input distributions including arrivals, services at stations $1, 2, 3, 4$ and the three routing processes. Thus, $h_1 = h_2 = h_3 = h_4 = h_5 = 2$ and $h_6 = h_7 = h_8 = 1$. The distributions for the three routing processes can be completely specified by their corresponding means. The distributions for the arrival and service processes can be uniquely characterized by the corresponding first two standardized moments: mean and standard deviation. To avoid a scaling effect, the standard deviation is used instead of the second raw moment. For the ℓth process, let τ_ℓ denote the first moment (mean) and ς_ℓ denote the standard deviation. Then $\mathbf{x} = (\tau_1, \varsigma_1, \tau_2, \varsigma_2, \tau_3, \varsigma_3, \tau_4, \varsigma_4, \tau_5, \varsigma_5, \tau_6, \tau_7, \tau_8)^\top$ with $d = 13$.

The true moments $\mathbf{x}_c$ are unknown and estimated based on a finite sample $\mathbf{Z_m}$ from F^c. As noted above, $\mathbf{X_m}$ is a $d \times 1$ dimensional moment estimator that is a function of $\mathbf{Z_m}$ written as $\mathbf{X_m} = \mathbf{X}(\mathbf{Z_m})$. Specifically, $\mathbf{X}_{\ell,m_\ell} = \mathbf{X}_\ell(\mathbf{Z}_{\ell,m_\ell})$ and $\mathbf{X}_\mathbf{m}^T = (\mathbf{X}_{1,m_1}^T, \mathbf{X}_{2,m_2}^T, \ldots, \mathbf{X}_{L,m_L}^T)$. Let $F^c_{\mathbf{X_m}}$ represent the true, unknown distribution of $\mathbf{X_m}$. Then "input uncertainty" refers to the distribution of $\mu(\mathbf{X_m})$ with $\mathbf{X_m} \sim F^c_{\mathbf{X_m}}$. Given a finite sample of real-world data $\mathbf{z}^{(0)}_\mathbf{m}$, Barton et al. (2014) used bootstrap resampling to approximate $F^c_{\mathbf{X_m}}$ and a metamodel to represent $\mu(\mathbf{x})$.

4.1 Bootstrap Resampling

We use distribution-free bootstrapping, meaning that we resample from the empirical distribution of the data rather than a fitted parametric distribution. Under some regularity conditions, the bootstrap can provide an asymptotically consistent approximation for the sampling distribution of a moment estimator (Shao and Tu, 1995). For the general performance of the bootstrap in representing the sampling distribution of an estimator, see Horowitz (2001).

Implementation of the bootstrap in metamodel-assisted bootstrapping is as follows.

1. Draw m_ℓ samples with replacement from $\mathbf{z}^{(0)}_{\ell,m_\ell}$, denoted by $\mathbf{Z}^{(1)}_{\ell,m_\ell}$, and calculate the corresponding $h_\ell \times 1$ vector of bootstrap moment estimates denoted by $\widehat{\mathbf{X}}^{(1)}_{\ell,m_\ell} = \mathbf{X}_\ell(\mathbf{Z}^{(1)}_{\ell,m_\ell})$ for $\ell = 1, 2, \ldots, L$. Then stack the results for all L processes to obtain a $d \times 1$ vector $\widehat{\mathbf{X}}^{(1)}_\mathbf{m}$.

2. Repeat the previous step B times to generate $\widehat{\mathbf{X}}^{(b)}_\mathbf{m}, b = 1, 2, \ldots, B$.

The bootstrap resampled moments are drawn from the bootstrap distribution denoted by $\widehat{F}_{\mathbf{X_m}}(\cdot|\mathbf{z}^{(0)}_\mathbf{m})$.

The most straightforward approach to propagate input uncertainty to the output mean is direct bootstrapping. Given a total computational budget of N simulation replications, the

system mean response at each of B bootstrap resamples is estimated by the sample mean of $n = \lfloor N/B \rfloor$ replications, denoted by $\bar{Y}(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}) = \sum_{j=1}^{n} Y_j(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}), b = 1, 2, \ldots, B$. From these mean estimates a CI is formed. Direct bootstrapping consumes the total simulation budget to estimate the simulation mean responses and to form the CI. Thus, for direct bootstrapping the impact of Monte Carlo estimation error on the CI and the impact of simulation estimation error on the mean responses are not separable.

Barton et al. (2014) assume that there is enough computational budget available so that the difference between $\widehat{\mu}(\mathbf{x})$ and $\mu(\mathbf{x})$ can be ignored; i.e. $\widehat{\mu}(\mathbf{x}) \approx \mu(\mathbf{x})$. Let $\widehat{\mu}_b \equiv \widehat{\mu}(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)})$ for $b = 1, 2, \ldots, B$. Barton et al. (2014) quantified input uncertainty by a $(1-\alpha)100\%$ bootstrap percentile interval induced by the sorted responses $[Q_L^*, Q_U^*] = [\widehat{\mu}_{(\lceil \frac{\alpha}{2} B \rceil)}, \widehat{\mu}_{(\lceil (1-\frac{\alpha}{2}) B \rceil)}]$, where $\widehat{\mu}_{(i)}$ is the ith smallest value of $\widehat{\mu}_1, \widehat{\mu}_2, \ldots, \widehat{\mu}_B$. Here, the superscript "*" indicates that the input distribution is approximated with the bootstrap. This interval provides asymptotically correct coverage when all input distributions meet certain moment conditions and the metamodel is continuously differentiable with nonzero gradient in a neighborhood of $\mathbf{x}_c$ (Barton et al., 2014). Specifically, they establish the iterated limit

$$\lim_{m \to \infty} \lim_{B \to \infty} \Pr\{\mu(\mathbf{x}_c) \in [Q_L^*, Q_U^*]\} = 1 - \alpha$$

where as $m \to \infty$ we have $m_\ell / m \to 1$, for $\ell = 1, 2, \ldots, L$.

However, with a tight computational budget for building the metamodel we cannot guarantee that $\widehat{\mu}(\mathbf{x}) \approx \mu(\mathbf{x})$ holds for arbitrarily complex systems especially for problems with many input distributions. Thus, we desire an interval estimator that accounts for both input and metamodel uncertainty. Stochastic kriging (SK), introduced by Ankenman et al. (2010), facilitates this. SK is flexible. Unlike the locally linear approximation in Cheng and Holland (2004), it does not require any strong assumption about the form of the underlying true response surface $\mu(\cdot)$. Based on our previous study (Xie et al., 2010), a SK metamodel fit to a small number of simulation runs can provide good global predictions and also a characterization of metamodel uncertainty for a wide variety of examples. The characterization of metamodel uncertainty is a key contribution of the new method presented here.

Compared with direct bootstrapping, metamodel-assisted bootstrapping separates the choice of B from the budget N and reduces the influence of simulation estimation error. Instead doing simulations at B samples from the bootstrap, we run simulations at well-chosen design points and build an equation-based metamodel $\widehat{\mu}(\mathbf{x})$ to predict the mean response at different input distributions represented by bootstrap resampled moments. Notice that once we have the metamodel, we can use any B we want to control the Monte Carlo estimation error of the interval estimator, even $B > N$. Further, the metamodel can efficiently use the computational budget to reduce the uncertainty introduced when propagating the input uncertainty to the output mean.

4.2 Stochastic Kriging Metamodel

Kriging is a widely used interpolation method. Since the outputs from stochastic simulations include simulation variability that often changes significantly across the design space, SK was

introduced to distinguish the uncertainty about the response surface from the simulation error inherent in stochastic simulation output. This section provides a brief review of SK.

Suppose that the underlying true (but unknown) response surface can be thought of as a realization of a stationary Gaussian Process (GP). This fiction has been shown to provide a very useful framework for quantifying uncertainty about the unknown surface implied by a deterministic computer experiment and has been successfully employed in a wide variety of applications (see, for instance, Santner et al. (2003)). SK extends this framework to include the variability of the simulation output Y using the model

$$Y_j(\mathbf{x}) = \beta_0 + W(\mathbf{x}) + \epsilon_j(\mathbf{x}). \tag{5}$$

The independent variable $\mathbf{x}$ is interpreted as a location in space. In this paper, $\mathbf{x}$ denotes a $d \times 1$ vector of moments that uniquely characterize the input distributions. The variation in the simulation output over the $\mathbf{x}$ space is divided into extrinsic (response-surface) uncertainty $W(\mathbf{x})$ and intrinsic (simulation output) uncertainty $\epsilon_j(\mathbf{x})$. The term "intrinsic" uncertainty refers to the variability inherent in the sampling that generates stochastic simulation outputs and "extrinsic" uncertainty refers to our lack of knowledge about the response surface.

SK uses a mean-zero, second-order stationary GP $W(\mathbf{x})$ to account for the spatial dependence of the response surface. Thus, the uncertainty about the true response surface $\mu(\mathbf{x})$ is represented by a GP $M(\mathbf{x}) \equiv \beta_0 + W(\mathbf{x})$ (note that β_0 can be replaced by a more general trend term $\mathbf{f}(\mathbf{x})^\top \boldsymbol{\beta}$ without affecting our method). For many, but not all, simulation settings the output is an average of a large number of more basic outputs, so a normal approximation can be applied: $\epsilon(\mathbf{x}) \sim \mathrm{N}(0, \sigma^2_\epsilon(\mathbf{x}))$. Of course, normality of the simulation output will not always hold, but could be empirically tested if there is a concern.

In SK, the covariance between $W(\mathbf{x})$ and $W(\mathbf{x}')$ quantifies how knowledge of the surface at some design points affects the prediction of the surface. A parametric form of the spatial covariance, denoted by $\Sigma(\mathbf{x}, \mathbf{x}') = \mathrm{Cov}[W(\mathbf{x}), W(\mathbf{x}')] = \tau^2 r(\mathbf{x} - \mathbf{x}')$, is typically assumed where τ^2 denotes the variance and $r(\cdot)$ is a correlation function that depends only on the distance $\mathbf{x} - \mathbf{x}'$. Based on our previous study (Xie et al., 2010), we use the product-form Gaussian correlation function

$$r(\mathbf{x} - \mathbf{x}') = \exp\left(- \sum_{j=1}^{d} \theta_j (x_j - x'_j)^2 \right)$$

for the empirical evaluation in Section 6; however, our results do not require it. Let $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_d)$ represent the correlation parameters; for different correlation functions the dimension of $\boldsymbol{\theta}$ could change. In any event, $M(\mathbf{x})$ can be represented by a Gaussian process $M(\mathbf{x}) \sim \mathrm{GP}(\beta_0, \tau^2 r(\mathbf{x} - \mathbf{x}'))$.

To reduce the uncertainty about $\mu(\mathbf{x})$ we choose an experiment design consisting of pairs $\mathcal{D} \equiv \{(\mathbf{x}_i, n_i), i = 1, 2, \ldots, k\}$ at which to run simulations and collect observations, where $(\mathbf{x}_i, n_i)$ denotes the location and the number of replications, respectively, at the ith design point. The design that we recommend is described in more detail in the Appendix, but it is not the only design that could be effective. The simulation outputs at $\mathcal{D}$ are

$\mathbf{Y}_{\mathcal{D}} \equiv \{(Y_1(\mathbf{x}_i), Y_2(\mathbf{x}_i), \ldots, Y_{n_i}(\mathbf{x}_i)); i = 1, 2, \ldots, k\}$ and the sample mean at design point $\mathbf{x}_i$ is $\bar{Y}(\mathbf{x}_i) = \sum_{j=1}^{n_i} Y_j(\mathbf{x}_i)/n_i$. Let the sample means at all k design points be $\bar{\mathbf{Y}}_{\mathcal{D}} = (\bar{Y}(\mathbf{x}_1), \bar{Y}(\mathbf{x}_2), \ldots, \bar{Y}(\mathbf{x}_k))^T$. Since the use of common random numbers is detrimental to prediction (as opposed to optimization; see Chen et al. (2012)), the simulations at different design points are independent and the variance of $\bar{\mathbf{Y}}_{\mathcal{D}}$ is represented by a $k \times k$ diagonal matrix $C = \text{diag}\{\sigma_\epsilon^2(\mathbf{x}_1)/n_1, \sigma_\epsilon^2(\mathbf{x}_2)/n_2, \ldots, \sigma_\epsilon^2(\mathbf{x}_k)/n_k\}$.

Let Σ be the $k \times k$ spatial covariance matrix of the design points and let $\Sigma(\mathbf{x}, \cdot)$ be the $k \times 1$ spatial covariance vector between each design point and a fixed prediction point $\mathbf{x}$. If the parameters $(\tau^2, \boldsymbol{\theta}, C)$ are known, then the metamodel uncertainty can be characterized by a refined GP $M_p(\mathbf{x})$ that denotes the conditional distribution of $M(\mathbf{x})$ given all simulation outputs,

$$M_p(\mathbf{x}) \sim \text{GP}(m_p(\mathbf{x}), \sigma_p^2(\mathbf{x})) \tag{6}$$

where $m_p(\cdot)$ is the minimum mean squared error (MSE) linear unbiased predictor

$$m_p(\mathbf{x}) = \widehat{\beta}_0 + \Sigma(\mathbf{x}, \cdot)^\top (\Sigma + C)^{-1} (\bar{\mathbf{Y}}_{\mathcal{D}} - \widehat{\beta}_0 \cdot 1_{k\times 1}), \tag{7}$$

and the corresponding variance is

$$\sigma_p^2(\mathbf{x}) = \tau^2 - \Sigma(\mathbf{x}, \cdot)^\top (\Sigma + C)^{-1} \Sigma(\mathbf{x}, \cdot) + \eta^\top [1_{k\times 1}^\top (\Sigma + C)^{-1} 1_{k\times 1}]^{-1} \eta \tag{8}$$

where $\widehat{\beta}_0 = [1_{k\times 1}^\top (\Sigma + C)^{-1} 1_{k\times 1}]^{-1} 1_{k\times 1}^\top (\Sigma + C)^{-1} \bar{\mathbf{Y}}_{\mathcal{D}}$ and $\eta = 1 - 1_{k\times 1}^\top (\Sigma + C)^{-1} \Sigma(\mathbf{x}, \cdot)$ (Ankenman et al., 2010). With the parameters $(\tau^2, \boldsymbol{\theta}, C)$ known, $M_p(\mathbf{x})$ depends on the simulation outputs only through $\bar{\mathbf{Y}}_{\mathcal{D}}$. Thus, $M_p(\mathbf{x})$ is a random function having the conditional distribution of $M(\mathbf{x})$ given $\bar{\mathbf{Y}}_{\mathcal{D}}$. Notice that $\sigma_p^2(\mathbf{x})$ reflects both metamodel and simulation error, including the constant term $\widehat{\beta}_0$, with the intrinsic simulation sampling error affecting $\sigma_p^2(\mathbf{x})$ through the matrix C.

Since in reality the spatial correlation parameters τ^2 and $\boldsymbol{\theta}$ are unknown, MLEs are typically used for prediction with the log-likelihood function

$$\ell(\beta_0, \tau^2, \boldsymbol{\theta}) = -\ln[(2\pi)^{k/2}] - \frac{1}{2}\ln[|\Sigma + C|] - \frac{1}{2}(\bar{\mathbf{Y}}_{\mathcal{D}} - \beta_0 \cdot 1_{k\times 1})^\top [\Sigma + C]^{-1} (\bar{\mathbf{Y}}_{\mathcal{D}} - \beta_0 \cdot 1_{k\times 1})$$

where Σ is a function of τ^2 and $\boldsymbol{\theta}$. The sample variance is used as an estimate for the simulation variance at design points C. By plugging $(\widehat{\beta}_0, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}}, \widehat{C})$ into Equations (7) and (8) we can obtain the estimated mean $\widehat{m}_p(\mathbf{x})$ and variance $\widehat{\sigma}_p^2(\mathbf{x})$. Thus, the metamodel we use is $\widehat{\mu}(\mathbf{x}) = \widehat{m}_p(\mathbf{x})$ with marginal variance estimated by $\widehat{\sigma}_p^2(\mathbf{x})$.

Ankenman et al. (2010) demonstrate that $\widehat{m}_p(\mathbf{x})$ is still an unbiased predictor even with the plug-in estimator $\widehat{C}$, and further that the variance inflation of $\sigma_p^2(\mathbf{x})$ caused by using $\widehat{C}$ is typically small. We performed an empirical study whose results indicate that if we use an adequate experiment design, such as the one-stage space-filling design used in this paper, then the performance of metamodel-assisted bootstrapping is also not sensitive to the estimation error in $\widehat{\tau}^2$ and $\widehat{\boldsymbol{\theta}}$; see the Appendix. However, it is known that the estimator (8)

with plug-in MLEs may sometimes underestimate the prediction variance; see Den Hertog et al. (2006).

In the derivations that follow we will assume that the parameters $(\tau^2, \boldsymbol{\theta}, C)$ are known. This is necessary (and common in the kriging literature) because including the effect of parameter estimation is mathematically intractable. To apply the methods in practice (including our empirical study below), we form plug-in estimators by inserting $\widehat{\tau}^2, \widehat{\boldsymbol{\theta}}, \widehat{C}$.

5 Confidence Interval and Variance Decomposition

Our approach is to use metamodel-assisted bootstrapping to provide a CI for the true mean performance. To be robust the CI should account for both input and metamodel uncertainty. Since $m_p(\mathbf{x})$ is an unbiased predictor under the Gaussian process assumption, $\sigma_p^2(\mathbf{x}) = 0$ for all $\mathbf{x}$ would imply that there is no metamodel uncertainty due either to a finite number of design points $\mathbf{x}_i$ or finite number of replications n_i; that is, $m_p(\mathbf{x}) = \mu(\mathbf{x})$. Unfortunately, with anything short of complete information, there will always be some metamodel uncertainty; and if the budget is tight relative to the complexity of the true response surface, then the effect of metamodel uncertainty could be substantial, resulting in significant undercoverage of the confidence interval of Barton et al. (2014) as we show in Section 6. The new interval introduced here does not suffer this degradation, and therefore is robust to the amount of simulation effort that can be expended and can be recommended for general use.

The kriging literature is the foundation for our work; see for instance Santner et al. (2003). Kriging provides inference about the value of an unknown function $\mu(\cdot)$ at a fixed prediction point $\mathbf{x}_0$ where the function has not been evaluated based on values of the function at a set of design points. Kriging models uncertainty about the function as a GP $M(\cdot)$ by assuming $\mu(\cdot)$ is a realization of $M(\cdot)$. An interval constructed to cover the conditional distribution of $M(\mathbf{x}_0)$ given the values at the design points is often interpreted as a CI for $\mu(\mathbf{x}_0)$ (e.g., Picheny et al. (2010)). The success of this paradigm is not because the function of interest is actually random—it is not—but because in many problems the conditional GP appears to be a robust characterization of the remaining response-surface uncertainty.

We adopt the kriging paradigm but with two key differences: our prediction point $\mathbf{x}_c$ is also unknown and must be estimated from real-world data, and our function $\mu(\cdot)$ can only be evaluated in the presence of stochastic simulation noise. Given the simulation outputs $\bar{\mathbf{Y}}_{\mathcal{D}}$, the remaining uncertainty about $\mu(\cdot)$ is characterized by the conditional GP $M_p(\cdot)$. To account for the impact from both input and metamodel uncertainty, we construct an interval $[C_L, C_U]$ covering $M_p(\mathbf{x}_c)$ with probability $(1-\alpha)100\%$; that is,

$$\Pr\{M_p(\mathbf{x}_c) \in [C_L, C_U]\} = 1 - \alpha. \tag{9}$$

Since the conditional coverage is $1-\alpha$, the unconditional coverage of $M(\mathbf{x}_c)$ is $1-\alpha$ as well. The revised objective (9) is connected to our objective (4) through the assumption that the function $\mu(\cdot)$ is a realization of the GP $M(\cdot)$. A procedure that delivers an interval satisfying (9) will be a good approximation for a CI procedure satisfying (4) if $M_p(\cdot)$ faithfully

represents the remaining uncertainty about $\mu(\cdot)$. This is clearly an approximation because in any real problem $\mu(\cdot)$ is a fixed function, therefore we refer to $[C_L, C_U]$ as an approximation for the CI (ACI).

In a practical setting, what is the next step if the interval $[C_L, C_U]$ is so wide that we are uncomfortable making decisions based on estimates with that level of error? We suggest gaining some sense of the relative contribution from each source of uncertainty as a guide toward either running more simulations or collecting more real-world input data or both. For many problems collecting additional input data is not feasible or we would have done so already; in such cases knowing that input uncertainty is substantial and cannot be reduced allows us to exercise caution in how we use the simulation results.

In this section, we first present a procedure to build an ACI that satisfies Equation (9) asymptotically. The asymptotic consistency of this interval is proved under the assumption that the true response surface is a realization of a GP with all parameters known except β_0. Next a variance decomposition is proposed to measure the relative contribution of input uncertainty to overall statistical uncertainty, and we study its asymptotic properties as well. This is a measure of input uncertainty due to *all* input distributions. A method for attributing the input uncertainty to the L distributions is provided by Song and Nelson (2013). Finally, we address problems that can arise when the system's mean performance fails to exist, or the system is undefined, for some values of the sample moments, and explain why metamodel-assisted bootstrapping tends to be tolerant of the former situation and can be adjusted for the latter.

Assumptions that are needed for the asymptotic analysis are the following:

Assumptions:

1. The ℓth input distribution is uniquely determined by its first h_ℓ moments and it has finite first $4h_\ell$ moments for $\ell = 1, 2, \ldots, L$.

2. We have i.i.d observations $Z_{\ell,1}^{(0)}, Z_{\ell,2}^{(0)}, \ldots, Z_{\ell,m_\ell}^{(0)}$ from the ℓth distribution for $\ell = 1, 2, \ldots, L$. As $m \to \infty$, we have $m_\ell/m \to c_\ell$, $\ell = 1, 2, \ldots, L$, for a constant $c_\ell > 0$.

3. The $\epsilon_j(\mathbf{x}) \overset{i.i.d.}{\sim} \mathrm{N}(0, \sigma_\epsilon^2(\mathbf{x}))$ for any $\mathbf{x}$, and $M(\mathbf{x})$ is a stationary, separable GP with a continuous correlation function satisfying

$$1 - r(\mathbf{x} - \mathbf{x}') \leq \frac{c}{|\log(\| \mathbf{x} - \mathbf{x}' \|_2)|^{1+\gamma}} \text{ for all } \| \mathbf{x} - \mathbf{x}' \|_2 \leq \delta \tag{10}$$

   for some $c > 0$, $\gamma > 0$ and $\delta < 1$, where $\| \mathbf{x} - \mathbf{x}' \|_2 = \sqrt{\sum_{j=1}^{d} (x_j - x_j')^2}$.

4. The input processes $Z_{\ell j}^{(0)}$, simulation noise $\epsilon_j(\mathbf{x})$ and GP $M(\mathbf{x})$ are mutually independent and the bootstrap process is independent of all of them.

Assumptions 1–2 give sufficient conditions for the almost sure (a.s.) consistency of bootstrap moment estimators $\widehat{\mathbf{X}}_\mathbf{m} \overset{a.s.}{\to} \mathbf{x}_c$ as $m \to \infty$ (see Lemma 1 in the Appendix). Under

Assumption 3, a GP $M(\cdot)$ with a correlation function satisfying Condition (10) has continuous sample paths almost surely (Adler 2010, Theorem 3.4.1). Condition (10) is satisfied by many correlation functions used in practice, and in particular any power exponential correlation function $r(\mathbf{x}-\mathbf{x}') = \exp\left(-\sum_{j=1}^{d}\theta_j|x_j - x_j'|^p\right)$ with $0 < p \leq 2$ and $\theta_j > 0$ (Santner et al., 2003). Assumption 4 indicates that input data are collected independently of the simulation model, and that our uncertainty about the mean response surface as represented by $M(\mathbf{x})$ is independent of the stochastic simulation noise (although both can depend on $\mathbf{x}$).

5.1 ACI Procedure

Based on a hierarchical approach, we propose the following procedure to build $(1-\alpha)100\%$ bootstrap percentile ACIs to achieve (9):

1. Given real-world data $\mathbf{z}_{\mathbf{m}}^{(0)}$, choose experiment design $\mathcal{D} = \{(\mathbf{x}_i, n_i), i = 1, 2, \ldots, k\}$ as described in the Appendix.

2. Run simulations at design points to obtain outputs $\mathbf{Y}_{\mathcal{D}}$. Compute the sample average $\bar{Y}(\mathbf{x}_i)$ and sample variance $S^2(\mathbf{x}_i)$ of the simulation outputs, $i = 1, 2, \ldots, k$. Fit the SK metamodel parameters $(\beta_0, \tau^2, \boldsymbol{\theta}, C)$ to obtain $\widehat{m}_p(\mathbf{x})$ and $\widehat{\sigma}_p^2(\mathbf{x})$ using $\left(\bar{Y}(\mathbf{x}_i), S^2(\mathbf{x}_i), \mathbf{x}_i\right)$, $i = 1, 2, \ldots, k$.

3. For $b = 1$ to B

   (a) Generate bootstrap resample $\mathbf{Z}_{\mathbf{m}}^{(b)} \overset{i.i.d.}{\sim} \mathbf{z}_{\mathbf{m}}^{(0)}$ and compute sample moments $\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}$.

   (b) Let $\widehat{\mu}_b \equiv \widehat{m}_p(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)})$.

   (c) Draw $\widehat{M}_b \sim \mathrm{N}\left(\widehat{m}_p(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}), \widehat{\sigma}_p^2(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)})\right)$.

   Next b

4. Report estimated CI and ACI, respectively,

$$\begin{aligned} \mathrm{CI}_0 &\equiv \left[\widehat{\mu}_{(\lceil B\frac{\alpha}{2}\rceil)}, \widehat{\mu}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}\right] \\ \mathrm{CI}_+ &\equiv \left[\widehat{M}_{(\lceil B\frac{\alpha}{2}\rceil)}, \widehat{M}_{(\lceil B(1-\frac{\alpha}{2})\rceil)}\right] \end{aligned}$$

   where $\widehat{\mu}_{(1)} \leq \widehat{\mu}_{(2)} \leq \cdots \leq \widehat{\mu}_{(B)}$ and $\widehat{M}_{(1)} \leq \widehat{M}_{(2)} \leq \cdots \leq \widehat{M}_{(B)}$ are the sorted values.

In this procedure, Step 1 provides an experiment design to build a SK metamodel, which is central to the metamodel-assisted bootstrapping approach. Since the input uncertainty is quantified with bootstrap resampled moments, we want the metamodel to correctly predict the responses at these points $\widehat{\mathbf{X}}_{\mathbf{m}} \sim \widehat{F}_{\mathbf{X}_{\mathbf{m}}}(\cdot|\mathbf{z}_{\mathbf{m}}^{(0)})$. Thus, the metamodel needs to be accurate and precise in a design space that covers the "most likely" bootstrap moment estimates,

which can be achieved by the experiment design proposed by Barton et al. (2014). Their design is data-driven; specifically, they first find the smallest ellipsoid denoted by E that covers the most likely bootstrap resampled moments. They then generate a space-filling design that covers E. This design methodology, which is summarized in the Appendix, yielded accurate metamodels in the examples they studied.

Based on the experiment design provided in Step 1, we run simulations and construct a metamodel in Step 2 by fitting $(\beta_0, \tau^2, \boldsymbol{\theta}, C)$. Given the metamodel, we predict the simulation's mean responses at different input settings corresponding to bootstrap resampled moments and construct interval estimators as shown in Step 3. Notice that Step 3(a) accounts for the input uncertainty and Step 3(c) accounts for the input and metamodel uncertainty. Thus, this procedure provides two types of intervals:

- CI_0, proposed in Barton et al. (2014), returns an estimate of $[Q_L, Q_U]$ in Equation (4) by assuming $\widehat{m}_p(\mathbf{x}) = \mu(\mathbf{x})$; that is, it only accounts for input uncertainty and will be in error if there is substantial metamodel uncertainty.

- CI_+ returns an estimate of $[C_L, C_U]$ in Equation (9). This ACI accounts for both input and metamodel uncertainty.

As the metamodel uncertainty decreases, CI_0 and CI_+ become closer and closer to each other. Before evaluating the finite-sample performance of CI_+ in Section 6, we establish its asymptotic consistency for objective (9).

In Theorems 1–3 that follow, we replace $\widehat{\mu}_b$ and $\widehat{M}_b$ in Steps 3(b)–(c) of the ACI procedure with

$$\mu_b \equiv m_p(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}) \text{ and } M_b \sim \text{N}\left(m_p(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}), \sigma_p^2(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)})\right)$$

recalling that $m_p(\cdot)$ and $\sigma_p^2(\cdot)$ are the marginal mean and variance of the conditional GP $M_p(\cdot)$ when $(\tau^2, \boldsymbol{\theta}, C)$ are known.

Theorem 1. *Suppose that Assumptions 1–4 hold. Then the interval* $[M_{(\lceil B\frac{\alpha}{2}\rceil)}, M_{(\lceil B(1-\frac{\alpha}{2})\rceil)}]$ *is asymptotically consistent, meaning the iterated limit*

$$\lim_{m\to\infty}\lim_{B\to\infty} \Pr\{M_{(\lceil B\alpha/2\rceil)} \le M_p(\mathbf{x}_c) \le M_{(\lceil B(1-\alpha/2)\rceil)}\} = 1-\alpha. \tag{11}$$

In brief, under the assumption that $\mu(\mathbf{x})$ is a realization of a GP, $M_p(\mathbf{x})$ characterizes the remaining metamodel uncertainty after observing $\bar{\mathbf{Y}}_{\mathcal{D}}$. And since the input uncertainty is asymptotically correctly quantified by the bootstrap moment estimator $\widehat{\mathbf{X}}_{\mathbf{m}}$, the distribution of $M_p(\widehat{\mathbf{X}}_{\mathbf{m}})$ accounts for both input and metamodel uncertainty. Theorem 1 shows that this interval satisfies objective (9) asymptotically. We are particularly interested in situations when the simulation effort is limited (Barton et al. (2014) addressed the ample budget case), so the consistency result in Theorem 1 is only with respect to the real-world data. The detailed proof is provided in the Appendix.

In practice, including our empirical evaluation in Section 6, $(\tau^2, \boldsymbol{\theta}, C)$ must be estimated, and the impact of parameter estimation (other than $\mathbf{x}_c$ and β_0) is not covered by Theorem 1. We address sensitivity to parameter estimation in the Appendix.

5.2 Variance Decomposition

CI_+ accounts for input and metamodel uncertainty. When the width of CI_+ is too large to be useful, it is important to know the relative contribution from each source. Since the total output variability is the convolution of the input uncertainty and simulation/metamodel uncertainty, it is hard to separate the effects from these sources. To estimate the relative contributions Zouaoui and Wilson (2003), Ng and Chick (2006), Ankenman and Nelson (2012) and Song and Nelson (2013) assume that the simulation noise has a constant variance. In this section, we propose a variance decomposition that does not require the homogeneity assumption.

Suppose that the parameters $(\tau^2, \boldsymbol{\theta}, C)$ are known, the metamodel uncertainty can be characterized by a GP and the simulation error follows a normal distribution. Then the metamodel uncertainty, given the simulation result $\bar{\mathbf{Y}}_{\mathcal{D}}$, is characterized by a GP $M_p(\mathbf{x}) \sim \text{N}(m_p(\mathbf{x}), \sigma_p^2(\mathbf{x}))$. Conditional on $\bar{\mathbf{Y}}_{\mathcal{D}}$, both $m_p(\mathbf{x})$ and $\sigma_p^2(\mathbf{x})$ are fixed functions. For notation simplification, all of following derivations are conditional on the simulation outputs $\bar{\mathbf{Y}}_{\mathcal{D}}$, but we will suppress the "$|\bar{\mathbf{Y}}_{\mathcal{D}}$".

The random variable $M_p(\mathbf{X_m})$ accounts for input uncertainty through the sampling distribution of $\mathbf{X_m}$ and the metamodel uncertainty through the random function $M_p(\cdot)$. To quantify the relative contribution of input and metamodel uncertainty, we decompose the total variance of $M_p(\mathbf{X_m})$ into two parts:

$$\begin{aligned} \sigma_T^2 &\equiv \text{Var}[M_p(\mathbf{X_m})] \\ &= \text{E}\{\text{Var}[M_p(\mathbf{X_m})|\mathbf{X_m}]\} + \text{Var}\{\text{E}[M_p(\mathbf{X_m})|\mathbf{X_m}]\} \\ &= \text{E}[\sigma_p^2(\mathbf{X_m})] + \text{Var}[m_p(\mathbf{X_m})]. \end{aligned} \tag{12}$$

The term $\sigma_M^2 \equiv \text{E}[\sigma_p^2(\mathbf{X_m})]$ is a measure of the metamodel uncertainty: the expected metamodel variance weighted by the density of moment estimator $\mathbf{X_m}$. This weighting makes sense because the accuracy of the metamodel in regions with higher density is more important for the estimation of system mean performance. The term $\sigma_I^2 \equiv \text{Var}[m_p(\mathbf{X_m})]$ is a measure of input uncertainty when we replace the unknown true response surface $\mu(\cdot)$ with its best linear unbiased estimate $m_p(\cdot)$.

What is the contribution of each term to ACI coverage? If the metamodel uncertainty disappears (i.e., $\sigma_p^2(\cdot) = 0$), then $\sigma_M^2 = 0$, CI_0 and CI_+ coincide and they provide asymptotically consistent coverage (Barton et al., 2014). Metamodel uncertainty is reduced by simulation effort. On the other hand, as $m \to \infty$ (more and more real-world input data), $\mathbf{X_m} \overset{a.s.}{\to} \mathbf{x}_c$ and since $m_p(\mathbf{x})$ is continuous we have $\sigma_I^2 = 0$; therefore, the width of CI_0 shrinks to zero as does coverage since there is remaining metamodel uncertainty in general. However, because CI_+ accounts for metamodel uncertainty it still provides asymptotically consistent coverage. This effect is demonstrated by the empirical study in Section 6.

Our decomposition allows us to express the total variance in Equation (12) as the sum of two variances measuring input and metamodel uncertainty: $\sigma_T^2 = \sigma_I^2 + \sigma_M^2$. In the metamodel-assisted bootstrapping framework, we can estimate each variance component as follows:

- Total variance: $\widehat{\sigma}_T^2 = \sum_{b=1}^{B}(M_b - \bar{M})^2/(B-1)$, where $\bar{M} = \sum_{b=1}^{B} M_b/B$.
- Input variance: $\widehat{\sigma}_I^2 = \sum_{b=1}^{B}(\mu_b - \bar{\mu})^2/(B-1)$, where $\bar{\mu} = \sum_{b=1}^{B} \mu_b/B$.
- Metamodel variance: $\widehat{\sigma}_M^2 = \sum_{b=1}^{B} \sigma_p^2(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)})/B$.

The ratio $\widehat{\sigma}_I/\widehat{\sigma}_T$ provides an estimate of the relative contribution from input uncertainty on CI_+. If it is close to 1, the contribution from metamodel uncertainty can be ignored. Thus, this ratio can help a decision maker determine where to put more effort: If the input variance dominates, then get more real-world data (if possible). If the metamodel variance dominates, then it can be reduced by more simulation, which can be a combination of additional design points and additional replications at existing design points. If neither dominates, then both activities are necessary to reduce CI_+ to a practically useful size.

The asymptotic properties of these variance component estimators are shown in the following theorems.

Theorem 2. *Suppose that Assumptions 1–4 hold. Then conditional on $\bar{\mathbf{Y}}_{\mathcal{D}}$ the variance component estimators $\widehat{\sigma}_M^2, \widehat{\sigma}_I^2, \widehat{\sigma}_T^2$ are consistent as $m, B \to \infty$, where as $m \to \infty$ we have $m_\ell/m \to c_\ell$, $\ell = 1, 2, \ldots, L$, for a constant $c_\ell > 0$. Specifically,*

- *As $m \to \infty$, the input uncertainty disappears:*

$$\lim_{m\to\infty} \sigma_M^2 = \sigma_p^2(\mathbf{x}_c),\ \lim_{m\to\infty} \sigma_I^2 = 0 \text{ and } \lim_{m\to\infty} \sigma_T^2 = \sigma_p^2(\mathbf{x}_c).$$

- *As $m \to \infty$ and $B \to \infty$ in an iterated limit, the variance component estimators are consistent:*

$$\begin{aligned}
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_M^2 &= \lim_{m\to\infty} \sigma_M^2 = \sigma_p^2(\mathbf{x}_c),\\
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_I^2 &= \lim_{m\to\infty} \sigma_I^2 = 0,\\
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_T^2 &= \lim_{m\to\infty} \sigma_T^2 = \sigma_p^2(\mathbf{x}_c).
\end{aligned}$$

Theorem 2 demonstrates that the variance components estimators $\widehat{\sigma}_I^2$, $\widehat{\sigma}_M^2$ and $\widehat{\sigma}_T^2$ are consistent. However, we can see that when $m \to \infty$ the input uncertainty disappears. Since $\lim_{m\to\infty} \sigma_I^2 = \lim_{m\to\infty} \lim_{B\to\infty} \widehat{\sigma}_I^2 = 0$ is not interesting, we study the consistency of scaled versions of σ_I^2 and $\widehat{\sigma}_I^2$ in Theorem 3, showing that $m\sigma_I^2$ and $m\widehat{\sigma}_I^2$ converge to the same non-zero constant.

Theorem 3. *Suppose that Assumptions 1–4 and the following additional assumptions hold:*

5. *The first three derivatives of the correlation function of the GP $M(\mathbf{x})$ exist and the third derivative is bounded; and*
6. *$m_\ell/m \to 1$ for $\ell = 1, 2, \ldots, L$.*

Then $\lim_{m\to\infty} m\sigma_I^2 = \lim_{m\to\infty}\lim_{B\to\infty} m\widehat{\sigma}_I^2 = \sigma_\mu^2$ *almost surely, where* σ_μ^2 *is a positive constant.*

Theorems 2–3 give the asymptotic properties of the variance component estimators, guaranteeing $\widehat{\sigma}_I/\widehat{\sigma}_T$ is a consistent estimator for the relative contribution of input to overall uncertainty. We will empirically evaluate its finite-sample performance in Section 6 where we form the variance component estimators by inserting $(\widehat{\tau}^2, \widehat{\boldsymbol{\theta}}, \widehat{C})$ for the unknown parameters $(\tau^2, \boldsymbol{\theta}, C)$.

5.3 Unstable and Undefined Moments

A fundamental assumption of simulation is that the expectation $\mu(\mathbf{x}_c)$ exists. This assumption does not imply, however, that it exists for *all* possible values of $\mathbf{x}$, $\mathbf{X_m}$ or $\widehat{\mathbf{X}}_\mathbf{m}^{(b)}$ that might be realized. The prototype example is a congestion-related performance measure of a queueing system as time goes to infinity when congestion increases without bound for some values of its interarrival-time and service-time parameters. We refer to systems for which $\mu(\mathbf{x})$ is $\pm\infty$ for some values of $\mathbf{x}$ as potentially *unstable*.

Recall that $\mu(\mathbf{x}) = \mathrm{E}[Y(\mathbf{x})]$. A second problem arises when for some values of $\mathbf{x}$ the random variable $Y(\mathbf{x})$ is undefined. The prototype example is a network for which we want to estimate some start-to-finish performance measure, but the start and finish are not connected for certain values of $\mathbf{x}$. We refer to such systems as potentially *undefined*.

Below we use illustrative examples to describe what happens to metamodel-assisted bootstrapping in each case, why we expect to be robust to unstable systems, and what needs to be done for undefined systems. We assume that $\mathbf{x}_c$ is an interior point of the space Ψ for which $\mu(\mathbf{x}_c)$ is stable and $Y(\mathbf{x}_c)$ is defined so both problems disappear asymptotically ($m \to \infty$), but they may occur when we apply the metamodel-assisted bootstrapping approach to a finite sample of real-world data.

5.3.1 Unstable Moments

Consider the simulation of an $M/M/1$ queue. Let x_1 and x_2 denote the mean interarrival time and mean service time, respectively, and let $\mathbf{x} = (x_1, x_2)^\top$. The unknown mean response $\mu(\mathbf{x})$ is the steady-state expected number of customers in the system. The true values x_1^c and x_2^c are unknown and must be estimated from real-world data; $x_1^c > x_2^c$ so the system is actually stable. We denote the unstable and stable regions of $\mathbf{x}$ by $U = \{(x_1, x_2) : 0 < x_1 \le x_2\}$ and $\bar{U} = \{(x_1, x_2) : x_1 > x_2 > 0\}$, respectively, and $\mathbf{x}_c$ is an interior point of $\bar{U}$.

As described in the Appendix, we use an initial set of bootstrap resampled moments to define an ellipsoid in which to embed our experiment design to fit the metamodel, and then generate a second set at which we evaluate the metamodel to form a two-sided, equal-probability bootstrap percentile interval. The conditional probability that a bootstrap resampled moment $\widehat{\mathbf{X}}_\mathbf{m}^{(b)}$ is located in the unstable region given the real-world data is

$$P_U \equiv \Pr\left\{ \widehat{\mathbf{X}}_\mathbf{m}^{(b)} \in U \,\middle|\, \mathbf{z}_\mathbf{m}^{(0)} \right\}. \tag{13}$$

For the $M/M/1$ queue we know U so we know which bootstrap moments are in the unstable region; therefore, we could map the mean response for unstable moments (symbolically) to ∞ and only fit or use the metamodel to predict the mean response at stable moments. If P_U is large this could lead to a one-sided lower confidence interval (infinite upper limit) that would be interpreted as "the real system may be unstable;" this is an appropriate conclusion if input uncertainty is substantial.

Unfortunately, in general stochastic systems it could be difficult or impossible to determine which moments are in the unstable region either analytically or empirically (Wieland et al., 2003). Thus, in the experiment design phase we might simulate the system at an $\mathbf{x} \in U$ to fit the metamodel, and in the bootstrapping phase we might evaluate the resulting metamodel at an $\mathbf{x} \in U$ to estimate the CI. What is the effect of doing this when $P_U > 0$, possibly even large?

Suppose we start each replication of the $M/M/1$ queue with an empty and idle system. Let $\mu(\mathbf{x}, t)$ denote the true expected number of customers in the system at time $t \geq 0$. Then except for the case $x_1 = x_2$, which we ignore, it can be shown that $\mu(\mathbf{x}, t)$ satisfies the differential equation

$$\frac{d\mu(\mathbf{x}, t)}{dt} = \frac{1}{x_1} - \frac{1 - p_0(t)}{x_2}$$

where $p_0(t)$ is the probability that the system is empty at t and $p_0(0) = 1$. If $\mathbf{x} \in \bar{U}$ then $p_0(t) \to 1 - x_2/x_1$ as $t \to \infty$; however, if $\mathbf{x} \in U$ then $p_0(t) \to 0$ as $t \to \infty$. Thus, for large t,

$$\frac{d\mu(\mathbf{x}, t)}{dt} \approx \begin{cases} 0, & \text{if } \mathbf{x} \in \bar{U} \\ \frac{1}{x_1} - \frac{1}{x_2} > 0, & \text{if } \mathbf{x} \in U. \end{cases} \tag{14}$$

For any finite run length T and warm-up period $T_0 < T$ the simulation provides an unbiased estimator of

$$\bar{\mu}(\mathbf{x}, T_0, T) = \frac{1}{T - T_0} \int_{T_0}^{T} \mu(\mathbf{x}, t)\, dt. \tag{15}$$

Notice that this quantity is finite for any positive values of x_1 and x_2, whether stable or not. However, if $\mathbf{x} \in \bar{U}$ then $\bar{\mu}(\mathbf{x}, T_0, T)$ converges to $\mu(\mathbf{x})$ for large T; while if $\mathbf{x} \in U$ then $\bar{\mu}(\mathbf{x}, T_0, T)$ is increasing in T for T large enough.

The key point is this: The expected value of any simulation-based estimator will be finite, even if the true steady-state mean is not. Further, the expected value of the simulation estimator at unstable $\mathbf{x}$ *will tend to be larger than at near-by stable* $\mathbf{x}$*. This means that the simulation estimates corresponding to unstable* $\mathbf{x}$ *will tend to be the largest ones observed, but still not infinite.*

Consider the design points or bootstrap resampled moments that are in U. When $P_U < \alpha/2$ and the run length is long enough, the unstable design points used to fit the metamodel, or bootstrap moments at which it is evaluated, tend not to adversely affect either the metamodel or the interval estimate because they are in the right tail beyond the $\alpha/2$ quantile. On the other hand, when $P_U \geq \alpha/2$ the large estimates corresponding to unstable design points or bootstrap moments tend to lengthen the interval estimate beyond

what is required to cover $\mu(\mathbf{x}_c)$; this causes overcoverage rather than undercoverage. Thus, metamodel-assisted bootstrapping will often be robust to unstable moments in the sense of not being too short, but possibly too long; we demonstrate this empirically in Section 6.

5.3.2 Undefined Moments

Consider the queueing network example in Figure 1. For simplicity, suppose that the routing probabilities p_1, p_2, p_3 are the only input parameters. Let $\mathbf{x} = (x_1, x_2, x_3)^\top = (p_1, p_2, p_3)^\top$. The true parameters $\mathbf{x}_c$ are unknown and estimated by finite samples from Bernoulli distributions (1 if the customer takes a particular route, 0 otherwise). Suppose that the mean response of interest, $\mu(\mathbf{x})$, is the steady-state expected time for a customer to traverse the network, which exists and is well-defined at $\mathbf{x}_c$. Unfortunately, $Y(\mathbf{x})$ may not be defined for every possible bootstrap resampled moment $\mathbf{x}$. For instance, if $\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)} = (0, 0.665, 0)^\top$ then Stations 1 (start) and 4 (end) are disconnected and no simulation output for time to traverse the network will ever be generated. Thus, the system corresponding to this bootstrap moment is undefined.

In practical problems for which we can obtain real-world input data, we should know a priori that the system performance measure is well defined (e.g., we would not include a route unless we actually observed a customer take it). Further, it should not be difficult to detect moments for which the system output is undefined, either because we understand the system logic (as in this example) or because the simulation simply fails to run. Therefore, a reasonable solution to the problem of undefined systems is to reject (and sample again) bootstrap moments $\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}$ that imply an undefined output. This makes our assessment of uncertainty *conditional* on the system performance measure being defined, which makes sense.

6 Empirical Study

In this section we use the queueing network described in Section 3 to evaluate the performance of our metamodel-assisted bootstrapping approach. The performance measure is the steady-state expected number of customers in the system. Both interarrival and service times follow gamma distributions and the routing decisions follow Bernoulli distributions. Thus, it is a 13-dimensional problem with $L = 8$ input processes that include both continuous and discrete distributions. The true parameters of the input distributions are $\alpha_A = 1, \beta_A = 0.25$, $\alpha_{S_i} = 1$, $\beta_{S_i} = 0.2$ for $i = 1, 2, 3, 4$ and $p_1 = p_2 = 0.5, p_3 = 0.75$. These parameter values imply a tractable Jackson network with steady-state number of customers in system $\mu(\mathbf{x}_c) = 12.67$. The maximum traffic intensity at any station is 0.8.

In the experiments we assume that all parameters for all input distributions are unknown and are estimated from a finite sample of real-world data. Notice that α_A, β_A, α_{S_i}, β_{S_i} for $i = 1, 2, 3, 4$ are estimated from continuous measurements, while the routing probabilities p_1, p_2, p_3 are estimated from 0 or 1 observations that would correspond to customer routing decisions. The model with *estimated* input parameters is almost surely not a Jackson network

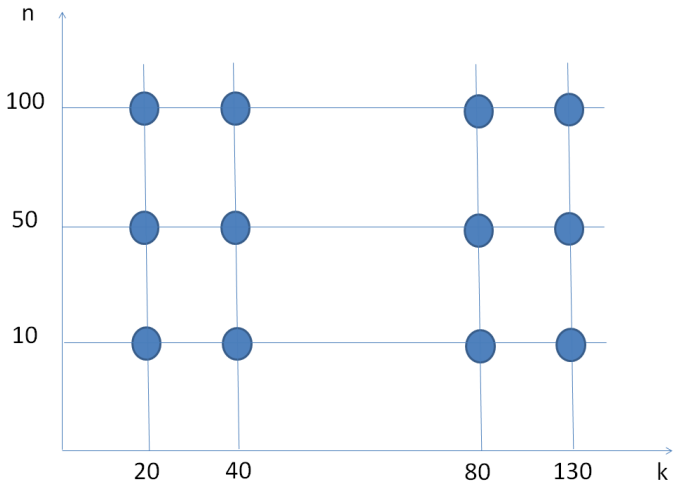

Figure 2: Experiment design.

and it could be unstable. Our measure of uncertainty is a 95% CI for $\mu(\mathbf{x}_c)$ as defined by (4) because this is the objective desired in practice.

To evaluate the robustness of the metamodel-assisted bootstrapping approach, we systematically examine the effect of the quantity of real-world data and the number of design points and replications per design point used to fit the metamodel; see Figure 2. We consider a wide range for the quantity of real-world data $m = 50, 500, 5000$, letting $m_\ell = m$ for $\ell = 1, 2, \ldots L$. The levels for the number of design points are $k = 20, 40, 80, 130$. For a 13-dimensional problem $k = 20$ is a very small design. The studies by Jones et al. (1998) and Loeppky et al. (2009) recommend that the number of design points should be 10 times the dimension of the problem for kriging; we take this as the maximum number of design points. The same number of replications are assigned to all design points and we try $n = 10, 50, 100$.

Barton et al. (2014) demonstrated that CI_0 has good performance when the impact of metamodel uncertainty is negligible. In this empirical study we focus on situations where metamodel uncertainty may be significant. However, rather than creating a problem that actually takes hours or days to run, we instead construct a problem with high metamodel uncertainty by using short run lengths for each replication: 20 time units after the warm up, which is roughly equivalent to 80 finished customers. To avoid the influence from initial bias, all simulations start loaded with the number of customers at each station being their steady-state expected values (rounded) under $\mathbf{x}_c$. Furthermore, a long warmup period of 200 time units is used. The net effect is that the point estimators of the steady-state number in the network have low bias, but may be quite variable.

To make the description of the empirical results easy to follow, we start with overall conclusions:

1. The new ACI CI_+ is robust to different levels of real-world data m, number of design points k and number of replications n.

Table 1: Percentage of unstable bootstrap resampled moments.

	$m = 50$	$m = 500$	$m = 5000$
mean of $\widehat{P}_U$	44.4%	2.3%	0
SD of $\widehat{P}_U$	31.7%	7.9%	0

2. When metamodel uncertainty is significant, CI_0 tends to have undercoverage that becomes more serious as m increases. Since CI_+ accounts for metamodel uncertainty, it does not exhibit this degradation although it sometimes has slight overcoverage.

3. Metamodel-assisted bootstrapping continues to deliver at least the nominal coverage when the probability of an unstable system P_U is large.

4. The ratio $\widehat{\sigma}_I/\widehat{\sigma}_T$ is a useful measure of the relative contribution of input to overall statistical uncertainty.

As discussed in Section 5.3.1, metamodel-assisted bootstrapping might behave differently when $P_U < \alpha/2$ vs. $P_U \geq \alpha/2$. Since P_U only depends on m and $\mathbf{x}_c$, we ran a side experiment to estimate it using

$$\widehat{P}_U = \frac{1}{B}\sum_{b=1}^{B} \text{I}\left(\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)} \in U\right), \tag{16}$$

where $\text{I}(\cdot)$ is the indicator function. The means and standard deviations (SD) of $\widehat{P}_U$ for $m = 50, 500, 5000$ were estimated based on 1000 macro-replications and are displayed in Table 1. In each macro-replication we independently generated a sample of size m of "real-world data." Then, conditional on these data, we drew $B = 2000$ bootstrap resampled moments. Finally, we calculated the estimate of $\widehat{P}_U$ using Equation (16).

As m increases the bootstrap resampled moments become more closely centered around $\mathbf{x}_c$. Thus, both the mean and SD of $\widehat{P}_U$ decrease with increasing m as shown in Table 1. When $m = 50$, P_U appears to be much larger than $\alpha/2$ so the bootstrap moments $\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}$ that correspond to the upper confidence bound are located in the unstable region U with high probability. When $m = 500$, P_U appears to be close to $\alpha/2 = 2.5\%$, while when $m = 5000$ there is little chance of getting unstable bootstrap moments.

In the following sections we describe the overall performance of CI_0 and CI_+, including the situation where $P_U > 0$, and analyze the finite-sample performance of $\widehat{\sigma}_I/\widehat{\sigma}_T$ as a measure of the relative contribution of input to overall uncertainty.

6.1 Performance of CIs

Tables 2–3 show the results for CI_0 and CI_+ when $m = 50, 500, 5000$, including the probability of covering $\mu(\mathbf{x}_c)$, and the mean and SD of the interval widths. All results are based on 1000 macro-replications.

When $m = 50$, P_U is much greater than $\alpha/2$ according to Table 1. This explains the very large CI widths in Table 2. Nevertheless, both CI_0 and CI_+ have reasonable coverage overall, an observation we explore further in Section 6.3. Notice that CI_0 does exhibit undercoverage when we use a very small experiment design of $k = 20$ points, while the coverage of CI_+ is much closer to the nominal value of 95% in this case. If we fix the number of replications n and increase the number of design points k, the coverage of CI_0 improves. For a fixed k the effect of increasing n is not as obvious.

Table 3 shows the results for $m = 500, 5000$. Compared with the results for $m = 50$, the mean and SD of the interval widths drop dramatically. The effects of k and n are easier to discern especially when $m = 5000$, which has no unstable bootstrap moments. Specifically, for a fixed quantity of real-world data m, if either the number of design points k or replications per design point n is small then CI_0 tends to have undercoverage because it fails to account for substantial metamodel uncertainty, unlike CI_+. However, because CI_+ does incorporate metamodel uncertainty it sometimes has slight overcoverage.

The most troubling observation about CI_0 is that, for fixed (n, k), as the amount of input data m increases its undercoverage becomes more serious. The diminished coverage occurs because as $m \to \infty$ the width of CI_0 shrinks to zero, which is not appropriate when there is still metamodel uncertainty. Again, CI_+ does not exhibit this degradation. As n and k increase, the coverages of CI_0 and CI_+ become closer to each other.

The behavior of CI_0 is what we would expect based on Barton et al. (2014), which introduced CI_0. Their procedure continued to add simulation effort (design points and replications) until its effect on the confidence interval was negligible. Compared to CI_0, the new interval, CI_+, is able to account for the effect of the remaining simulation estimation error. Therefore, it can work under more general situations where the simulated systems are complex and the simulation budget is tight.

6.2 Performance of $\widehat{\sigma}_I/\widehat{\sigma}_T$

Tables 2–3 also demonstrate that $\widehat{\sigma}_I/\widehat{\sigma}_T$ provides a good measure of the relative contribution of input to overall uncertainty, and behaves as it should:

- For a fixed amount of real-world data m, increasing the number of design points and replications (n, k) drives $\widehat{\sigma}_I/\widehat{\sigma}_T$ toward 1, indicating a decrease in metamodel uncertainty.

- For fixed simulation effort (n, k), increasing the amount of real-world data m decreases $\widehat{\sigma}_I/\widehat{\sigma}_T$, indicating that there is relatively less input uncertainty. Notice, however, that the relationship is not simple because as m increases the design space over which we fit the metamodel becomes smaller, so that even with the same simulation effort the absolute level of metamodel uncertainty will decrease somewhat.

- When $\widehat{\sigma}_I/\widehat{\sigma}_T$ is near 1, the behaviors (coverage and width) of CI_0 and CI_+ are similar and both have coverage close to the nominal level; this is illustrated in Figure 3. Recall

Table 2: Results for CI_0, CI_+ and $\widehat{\sigma}_I/\widehat{\sigma}_T$ when $m = 50$.

$m = 50$	$k = 20$			$k = 40$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	91.9%	92.3%	91.5%	93.8%	94.4%	93.4%
Coverage of CI_+	93.9%	94.9%	93.7%	94.9%	95.6%	95.9%
CI_0 Width (mean)	326.4	332.4	339.5	319.1	328.6	326.5
CI_+ Width (mean)	344.1	348.8	357.1	332.3	342.3	341.2
CI_0 Width (SD)	183.1	173.6	180.7	176.4	167.6	175
CI_+ Width (SD)	188	175.7	183.8	178.2	169.2	176.1
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.963	0.965	0.964	0.973	0.973	0.971
$m = 50$	$k = 80$			$k = 130$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	94.6%	96.3%	95.4%	94.2%	95.1%	95.4%
Coverage of CI_+	95.9%	96.7%	96.1%	94.5%	96%	96.1%
CI_0 Width (mean)	312.1	314.8	322.7	322	321.86	320
CI_+ Width (mean)	322	325.7	334	330.2	331	329.4
CI_0 Width (SD)	169.7	159.1	164.7	171.5	169.3	172.3
CI_+ Width (SD)	171.2	159.4	165	172.7	169.5	172.7
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.982	0.98	0.978	0.985	0.985	0.983

that CI_0 does not account for metamodel uncertainty, and that $\widehat{\sigma}_I/\widehat{\sigma}_T \approx 1$ indicates that input uncertainty is large relative to metamodel uncertainty, which is when CI_0 will do best. Figure 3 also illustrates the general robustness of CI_+.

6.3 Robustness to Unstable Moments

Recall the observation from Table 2 that when there is a small quantity of real-world data ($m = 50$), resulting in a large probability of unstable bootstrap moments, then both CI_0 and CI_+ had large mean and SD of their widths, yet provided reasonable coverage. Examining the results, we found that most of the intervals that fail to cover the mean do so because the lower confidence bound is *above* $\mu(\mathbf{x}_c)$; this is the case for both CI_0 and CI_+. Using all 1000 macro-replications, the estimated probability that the lower confidence bound is above the mean (greater than $\mu(\mathbf{x}_c)$, meaning too large) is 4.4% for CI_0 and 3.8% for CI_+, while the estimated probability that the upper confidence bound is below the mean (less than $\mu(\mathbf{x}_c)$, meaning too small) is only 0.2% for CI_0 and 0.1% for CI_+; for two-sided equal percentile intervals we would expect these to be around 2.5%.

We conclude that even though the metamodel predicts a finite mean when it should be infinite, this will still tend to lead to overcoverage rather than undercoverage, and therefore is conservative.

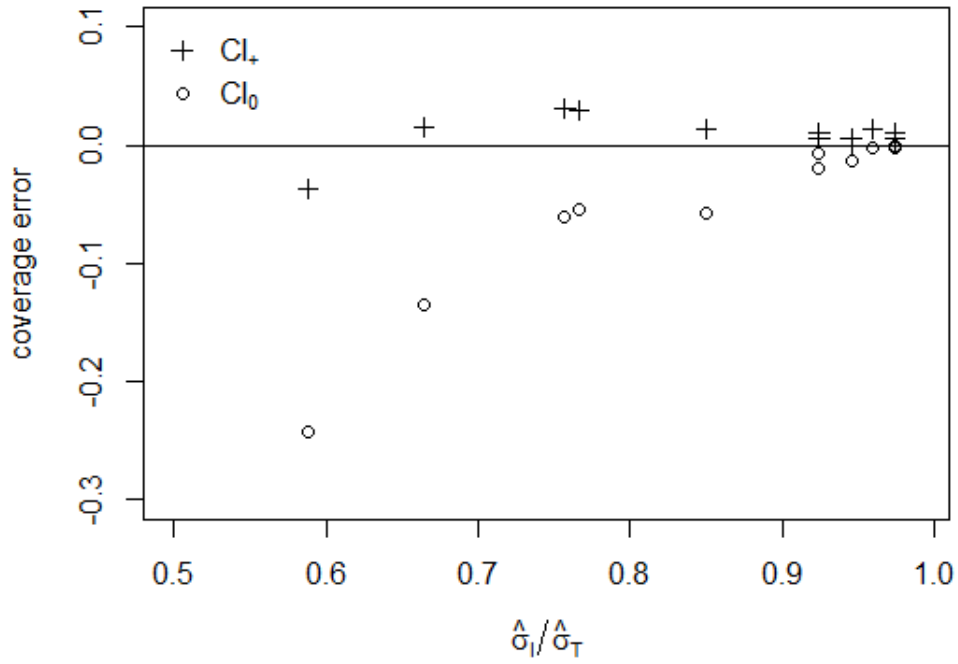

Figure 3: The coverage errors for CI_0 and CI_+ vs. $\widehat{\sigma}_I/\widehat{\sigma}_T$ when $m = 5000$ across all values of n and k.

7 Conclusions

In this paper, a metamodel-assisted bootstrapping approach is used for statistical uncertainty analysis. Input uncertainty is approximated by the bootstrap, an equation-based stochastic kriging metamodel is used to propagate the input uncertainty to the output mean, and the metamodel uncertainty is derived using properties of stochastic kriging. This approach delivers an interval estimator that accounts for *all* statistical uncertainty, both simulation and input. The asymptotic consistency of this interval is proved under the assumption that the true response surface is a realization of a Gaussian process and certain parameters are known.

An empirical study on a difficult problem demonstrates that our approach can have good finite-sample performance even when there are several input distributions (both discrete-valued and continuous-valued), a tight computational budget, and bootstrap moments corresponding to unstable systems. Thus, the new interval does not require a sequential experiment to make metamodel uncertainty negligible, as in Barton et al. (2014).

If CI_+ is too wide, then it is important to know the relative contributions from input and

metamodel uncertainty as a guide to either collecting more real-world data or doing more simulation or both. We give a measure of the relative contribution of input to overall statistical uncertainty by using a variance decomposition and analyze its asymptotic properties.

Acknowledgments

This paper is based upon work supported by the National Science Foundation under Grant Nos. CMMI-0900354 and CMMI-1068473. The authors thank Ohad Perry, the associate editor and two anonymous referees for their help in the presentation of results and certain technical issues.

References

Adler, R. J. (2010). *The Geometry of Random Fields*. Philadelphia, PA: SIAM.

Ankenman, B. E. and B. L. Nelson (2012). A quick assessment of input uncertainty. In *Proceedings of the 2012 Winter Simulation Conference*, pp. 241–250. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Ankenman, B. E., B. L. Nelson, and J. Staum (2010). Stochastic kriging for simulation metamodeling. *Operations Research 58*, 371–382.

Bangerth, W., H. Klie, M. F. Wheeler, P. L. Stoffa, and M. K. Sen (2006). On optimization algorithms for the reservoir oil well placement problem. *Computational Geosciences 10*, 303–319.

Barton, R. R. (2007). Presenting a more complete characterization of uncertainty: Can it be done? In *Proceedings of the 2007 INFORMS Simulation Society Research Workshop*, Fontainebleau. INFORMS Simulation Society.

Barton, R. R., B. L. Nelson, and W. Xie (2014). Quantifying input uncertainty via simulation confidence interval. *Informs Journal on Computing 26*, 74–87.

Barton, R. R. and L. W. Schruben (1993). Uniform and bootstrap resampling of input distributions. In *Proceedings of the 1993 Winter Simulation Conference*, pp. 503–508. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R. R. and L. W. Schruben (2001). Resampling methods for input modeling. In *Proceedings of the 2001 Winter Simulation Conference*, pp. 372–378. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Billingsley, P. (1995). *Probability and Measure*. New York: Wiley-Interscience.

Chen, X., B. E. Ankenman, and B. L. Nelson (2012). The effects of common random numbers on stochastic kriging metamodels. *ACM Transactions on Modeling and Computer Simulation 22*, 7.

Cheng, R. C. H. and W. Holland (1997). Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation 57*, 219–241.

Cheng, R. C. H. and W. Holland (2004). Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation 14*, 344–362.

Chick, S. E. (2001). Input distribution selection for simulation experiments: Accounting for input uncertainty. *Operations Research 49*, 744–758.

Chick, S. E. and S. H. Ng (2002). Joint criterion for factor identification and parameter estimation. In *Proceedings of the 2002 Winter Simulation Conference*, pp. 400–406. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Den Hertog, D., J. P. C. Kleijnen, and A. Y. D. Siem (2006). The correct Kriging variance estimated by bootstrapping. *Journal of the Operational Research Society 57*, 400–409.

Fowler, J. W. and O. Rose (2004). Grand challenges in modeling and simulation of complex manufacturing systems. *Simulation 80*, 469–476.

Ghosh, S., Y. Matsuoka, Y. Asai, K.-Y. Hsin, and H. Kitano (2011). Software for systems biology: From tools to integrated platforms. *Nature Reviews Genetics 12*, 821–832.

Horowitz, J. L. (2001). *The Bootstrap*, Volume 5 of *Handbook of Econometrics*. Oxford, UK: North Holland.

Jones, D., M. Schonlau, and W. Welch (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization 13*, 455–492.

Kastner, J., J. Solomon, and S. Fraser (2002). Modeling a hox gene network in silico using a stochastic simulation algorithm. *Developmental Biology 246*, 122–131.

Kleijnen, J. P. C. (2008). *Design and Analysis of Simulation Experiements*. New York: Springer.

Lehmann, E. and G. Casella (1998). *Theory of Point Estimation*. New York: Springer-Verlag.

Loeppky, J. L., J. Sachs, and W. J. Welch (2009). Choosing the sample size of a computer experiment: A practical guide. *Technometrics 51*, 366–376.

Ng, S. H. and S. E. Chick (2006). Reducing parameter uncertainty for stochastic systems. *ACM Transactions on Modeling and Computer Simulation 16*, 26–51.

Picheny, V., D. Ginsbourger, O. Roustant, R. T. Haftka, and N. Kim (2010). Adaptive designs of experiments for accurate approximation of a target region. *Journal of Mechanical Design 132*, 071008.

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

Serfling, R. J. (2002). *Approximation Theorems of Mathematical Statistics*. New York: Wiley.

Severini, T. (2005). *Elements of Distribution Theory*. New York: Cambridge University Press.

Shao, J. and D. Tu (1995). *The Jackknife and Bootstrap*. New York: Springer.

Song, E. and B. L. Nelson (2013). A quicker assessment of input uncertainty. In *Proceedings of the 2013 Winter Simulation Conference*, pp. 474–485. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Sun, H. and M. Farooq (2002). Note on the generation of random points uniformly distributed in hyper-ellipsoids. In *Proceedings of the Fifth International Conference on Information Fusion*, pp. 489–496.

Van Der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge, UK: Cambridge University Press.

Wang, H., D. E. Ciaurri, L. J. Durlofsky, and A. Cominelli (2012). Optimal well placement under uncertainty using a retrospective optimization framework. *SPE Journal 17*, 112–121.

Wieland, J. R., R. Pasupathy, and B. W. Schmeiser (2003). Queueing-network stability: Simulation-based checking. In *Proceedings of the 2003 Winter Simulation Conference*, pp. 520–527. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Xie, W., B. L. Nelson, and J. Staum (2010). The influence of correlation functions on stochastic kriging metamodels. In *Proceedings of the 2010 Winter Simulation Conference*, pp. 1067–1078. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zouaoui, F. and J. R. Wilson (2003). Accounting for parameter uncertainty in simulation input modeling. *IIE Transactions 35*, 781–792.

Zouaoui, F. and J. R. Wilson (2004). Accounting for input-model and input-parameter uncertainties in simulation. *IIE Transactions 36*, 1135–1151.

Table 3: Results for CI_0, CI_+ and $\widehat{\sigma}_I/\widehat{\sigma}_T$ when $m = 500$ and $m = 5000$.

$m = 500$	$k = 20$			$k = 40$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	90.5%	94.6%	95.1%	94.9%	96.7%	96.4%
Coverage of CI_+	95.7%	97.7%	97.8%	96.6%	98.3%	97.8%
CI_0 Width (mean)	24.8	28.1	29.4	27.1	28.5	28.7
CI_+ Width (mean)	28.9	30.8	32.2	29.6	30.3	30.5
CI_0 Width (SD)	19.9	19.4	20.6	19.1	19.2	19.9
CI_+ Width (SD)	20.6	20.4	21.7	19.7	19.9	20.6
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.88	0.932	0.933	0.932	0.957	0.958
$m = 500$	$k = 80$			$k = 130$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	96.5%	97.5%	95.8%	95.4%	96.5%	95.9%
Coverage of CI_+	98%	98.3%	97.3%	97.5%	97.1%	96.9%
CI_0 Width (mean)	26.3	28	28.7	26.4	27.9	27.6
CI_+ Width (mean)	28	29	29.7	27.9	28.6	28.2
CI_0 Width (SD)	17.4	18	19.3	18.8	19.6	19.3
CI_+ Width (SD)	17.7	18.4	19.6	18.9	19.9	19.5
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.952	0.977	0.978	0.957	0.984	0.987
$m = 5000$	$k = 20$			$k = 40$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	70.7%	89.2%	93.1%	81.5%	94.3%	94.8%
Coverage of CI_+	91.3%	96.3%	95.6%	96.5%	96.1%	96.3%
CI_0 Width (mean)	3.29	3.97	4.14	3.93	4.23	4.3
CI_+ Width (mean)	5.85	4.8	4.56	6.08	4.64	4.52
CI_0 Width (SD)	1.89	1.2	1	1.64	0.87	0.83
CI_+ Width (SD)	2.12	1.13	1	1.52	0.89	0.85
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.588	0.85	0.924	0.664	0.924	0.959
$m = 5000$	$k = 80$			$k = 130$		
	$n = 10$	$n = 50$	$n = 100$	$n = 10$	$n = 50$	$n = 100$
Coverage of CI_0	88.9%	93.6%	94.9%	89.5%	93.7%	94.8%
Coverage of CI_+	98.1%	95%	96%	98%	95.6%	95.5%
CI_0 Width (mean)	4.54	4.29	4.29	4.52	4.35	4.32
CI_+ Width (mean)	6.1	4.56	4.42	5.98	4.64	4.45
CI_0 Width (SD)	1.37	0.85	0.77	1.28	0.9	0.79
CI_+ Width (SD)	1.27	0.85	0.78	1.13	0.87	0.77
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.757	0.946	0.974	0.766	0.945	0.974

8 Appendix (Intended for an Online Companion)

In this appendix we prove Theorems 1–3 and provide a brief description of the experiment design used to build stochastic kriging metamodels.

To be self-contained, we first state some definitions, lemmas and theorems that are used in the proofs. Let $\stackrel{D}{\to}$ denote convergence in distribution.

- **Borel-Cantelli Lemma** (Billingsley, 1995): For events $A_1, A_2, \ldots$, if $\sum_{n=1}^{\infty} \Pr(A_n)$ converges, then
$$\Pr\left(\limsup_n A_n\right) = 0$$
where
$$\limsup_n A_n = \cap_{n=1}^{\infty} \cup_{k=n}^{\infty} A_k$$
is the set of outcomes that occur infinitely many times.

- **Lemma 2.11** (Van Der Vaart, 1998): Suppose that $\mathbf{X}_n \stackrel{D}{\to} \mathbf{X}$ for a random vector $\mathbf{X}$ with a continuous distribution function. Then the distribution function of $\mathbf{X}_n$ converges uniformly to that of $\mathbf{X}$: $\| F_{\mathbf{X}_n} - F_{\mathbf{X}} \|_\infty \to 0$, where $\| h \|_\infty$ is the sup-norm of h on $\Re$, $\| h \|_\infty = \sup_t |h(t)|$.

- **Portmanteau Lemma** (Van Der Vaart, 1998): For any random vectors $\mathbf{X}_n$ and $\mathbf{X}$ the following statements are equivalent.

  1. $\mathbf{X}_n \stackrel{D}{\to} \mathbf{X}$.
  2. $\mathrm{E}[f(\mathbf{X}_n)] \to \mathrm{E}[f(\mathbf{X})]$ for all bounded, continuous functions f.

- **Theorem 2.3** (Van Der Vaart, 1998): Let $g : \Re^k \to \Re^m$ be continuous at every point in a set $\mathcal{C}$ such that $\Pr\{X \in \mathcal{C}\} = 1$. Then

  1. If $X_n \stackrel{D}{\longrightarrow} X$ then $g(X_n) \stackrel{D}{\longrightarrow} g(X)$.
  2. If $X_n \stackrel{P}{\longrightarrow} X$ then $g(X_n) \stackrel{P}{\longrightarrow} g(X)$.
  3. If $X_n \stackrel{a.s.}{\longrightarrow} X$ then $g(X_n) \stackrel{a.s.}{\longrightarrow} g(X)$.

  In the proofs when we refer to the "continuous mapping theorem" we will mean Theorem 2.3.

- **Glivenko-Cantelli Theorem** (Van Der Vaart, 1998): If $X_1, X_2, \ldots, X_n$ are i.i.d. random variables with distribution function F and F_n is the empirical cdf of $X_1, X_2, \ldots, X_n$, then $\| F_n - F \|_\infty \stackrel{a.s.}{\longrightarrow} 0$ as $n \to \infty$.

- **Lemma 21.2** (Van Der Vaart, 1998): For cdf F, define the inverse cdf to be

$$F^{-1}(p) = \inf\{t\colon F(t) \geq p\}.$$

Then a sequence of cdfs $F_n(t) \to F(t)$ for every t where F is continuous if and only if $F_n^{-1}(p) \to F^{-1}(p)$ for every p where F^{-1} is continuous.

- **Theorem 13.1** (Severini, 2005): Let $\mathbf{X}_1, \mathbf{X}_2, \ldots$ denote a sequence of d-dimensional random vectors such that, for some vector $\boldsymbol{\mu}$,

$$\sqrt{n}(\mathbf{X}_n - \boldsymbol{\mu}) \xrightarrow{D} \mathrm{N}(\mathbf{0}_{d\times 1}, \Sigma) \text{ as } n \to \infty,$$

where Σ is a $d \times d$ positive definite matrix with $|\Sigma| < \infty$. Let $g\colon \Re^d \to \Re^k$ denote a continuously differentiable function and let $\nabla g(\mathbf{x})$ denote the $d \times k$ matrix of partial derivatives of g with respect to $\mathbf{x}$. Then

$$\sqrt{n}(g(\mathbf{X}_n) - g(\boldsymbol{\mu})) \xrightarrow{D} \mathrm{N}(\mathbf{0}_{k\times 1}, \nabla g(\boldsymbol{\mu})^\top \Sigma \nabla g(\boldsymbol{\mu})) \text{ as } n \to \infty.$$

- **Theorem 3.8** (Shao and Tu, 1995): Let $\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m$ denote d-dimensional i.i.d. random vectors and $\bar{\mathbf{X}}_m = m^{-1}\sum_{i=1}^m \mathbf{X}_i$. Let $\bar{\mathbf{X}}_m^* = m^{-1}\sum_{i=1} \mathbf{X}_i^*$ where $\{\mathbf{X}_1^*, \mathbf{X}_2^*, \ldots, \mathbf{X}_m^*\}$ are randomly and independently drawn with replacement from $\{\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_m\}$. Let $g\colon \Re^d \to \Re^k$ denote a continuously differentiable function and $\nabla g(\mathbf{x})$ denote the $d \times k$ matrix of partial derivatives of g with respect to $\mathbf{x}$. Let $T_m = g(\bar{\mathbf{X}}_m)$ and denote the bootstrap variance estimator for T_m by $v_m^* = \mathrm{Var}_*[g(\bar{\mathbf{X}}_m^*)]$. Suppose that $\mathrm{E}[\mathbf{X}_1^\top \mathbf{X}_1] < \infty$ and $\nabla g(\boldsymbol{\mu}) \neq \mathbf{0}_{d\times k}$ where $\boldsymbol{\mu} = \mathrm{E}[\mathbf{X}_1]$. Suppose further that

$$\max_{i_1,\ldots,i_m} |T_m(\mathbf{X}_{i_1}, \ldots, \mathbf{X}_{i_m}) - T_m|/\tau_m \xrightarrow{a.s.} 0, \tag{17}$$

where the maximum is taken over all integers $i_1, \ldots, i_m$ satisfying $1 \leq i_1 \leq \cdots \leq i_m \leq m$, and $\{\tau_m\}$ is a sequence of positive numbers satisfying $\liminf_m \tau_m > 0$ and $\tau_m = O(e^{m^q})$ with a $q \in (0, 1/2)$. Then v_m^* is strongly consistent, i.e., $v_m^*/\sigma_m^2 \xrightarrow{a.s.} 1$, where $\sigma_m^2 = m^{-1}\nabla g(\boldsymbol{\mu})^\top \Sigma \nabla g(\boldsymbol{\mu})$ and $\Sigma = \mathrm{Var}(\mathbf{X}_1)$.

- **Theorem 1.1** (Lehmann and Casella (1998), Chapter 6): Let $X_1, X_2, \ldots, X_m$ be i.i.d. with $\mathrm{E}(X_1) = \mu$, $\mathrm{Var}(X_1) = \sigma^2$, and finite fourth moment, and suppose h is a function of a real variable whose first four derivatives $h'(x), h''(x), h^{(3)}(x)$ and $h^{(4)}(x)$ exist for all $x \in I$, where I is an interval with $\Pr(X_1 \in I) = 1$. Furthermore, suppose that $|h^{(4)}(x)| \leq M$ for all $x \in I$, for some $M < \infty$. Then

$$\mathrm{E}[h(\bar{X})] = h(\mu) + \frac{\sigma^2}{2m} h''(\mu) + \mathcal{R}_m.$$

If, in addition, the fourth derivative of h^2 is also bounded, then

$$\mathrm{Var}[h(\bar{X})] = \frac{\sigma^2}{m}[h'(\mu)]^2 + \mathcal{R}_m.$$

In both cases the remainder $\mathcal{R}_m$ is $O(1/m^2)$.

- **Multivariate Taylor Formula** (Serfling (2002), page 44): Let the function g defined on $\Re^d$ posses continuous partial derivatives of order n at each point of an open set $S \subset \Re^d$. Let $\mathbf{x} \in S$. For each point $\mathbf{y}$, $\mathbf{y} \neq \mathbf{x}$, such that the line segment $L(\mathbf{x}, \mathbf{y})$ joining $\mathbf{x}$ and $\mathbf{y}$ lies in S, there exists a point $\mathbf{z}$ in the interior of $L(\mathbf{x}, \mathbf{y})$ such that

$$g(\mathbf{y}) = g(\mathbf{x}) + \sum_{k=1}^{n-1} \frac{1}{k!} \sum_{i_1=1}^{d} \cdots \sum_{i_k=1}^{d} \left. \frac{\partial^k g(t_1, \ldots, t_d)}{\partial t_{i_1} \cdots \partial t_{i_k}} \right|_{\mathbf{t}=\mathbf{x}} \cdot \prod_{j=1}^{k} (y_{i_j} - x_{i_j})$$
$$+ \frac{1}{n!} \sum_{i_1=1}^{d} \cdots \sum_{i_n=1}^{d} \left. \frac{\partial^n g(t_1, \ldots, t_d)}{\partial t_{i_1} \cdots \partial t_{i_n}} \right|_{\mathbf{t}=\mathbf{z}} \cdot \prod_{j=1}^{n} (y_{i_j} - x_{i_j}).$$

8.1 Asymptotic Consistency of $\mathbf{CI}_+$

To prove Theorem 1, we first establish three supporting lemmas.

Lemma 1. *Suppose that Assumptions 1–2 hold. Then the bootstrap resampled moments converge almost surely to the true moments* $\widehat{\mathbf{X}}_m \overset{a.s.}{\to} \mathbf{x}_c$ *as* $m \to \infty$.

Proof: Since all of the input processes are independent, we establish the result for one input distribution F^c without loss of generality. We prove the result for $\mathbf{x}_c$ being the generic hth-order moment, $\alpha_h \equiv \mathrm{E}(Z^h) < \infty$, for $Z \sim F^c$.

The hth-order bootstrap resampled moment is

$$\widehat{X}_m = \frac{1}{m} \sum_{j=1}^{m} (Z^{(j;m)})^h \text{ with } Z^{(j;m)} \overset{i.i.d}{\sim} \mathbf{Z}_m^{(0)} \tag{18}$$

where "$Z^{(j;m)} \sim \mathbf{Z}_m^{(0)}$" denotes the jth independent sample with replacement from $\mathbf{Z}_m^{(0)}$. We use the Chebychev Inequality and the Borel-Cantelli Lemma to prove the result.

By the Chebychev Inequality, for every $\epsilon > 0$, we have

$$\Pr\left\{ |\widehat{X}_m - \alpha_h| > \epsilon \right\} \leq \frac{\mathrm{E}\left[(\widehat{X}_m - \alpha_h)^4 \right]}{\epsilon^4}. \tag{19}$$

Notice that

$$\mathrm{E}\left[(\widehat{X}_m - \alpha_h)^4 \right] = \mathrm{E}\left[\widehat{X}_m^4 \right] - 4\alpha_h \mathrm{E}\left[\widehat{X}_m^3 \right] + 6\alpha_h^2 \mathrm{E}\left[\widehat{X}_m^2 \right] - 4\alpha_h^3 \mathrm{E}\left[\widehat{X}_m \right] + \alpha_h^4. \tag{20}$$

We will analyze each term in Equation (20). First, we show that any ith bootstrap resampled moment, denoted as $\widehat{\alpha}_i$, is unbiased,

$$\begin{aligned} \mathrm{E}\left[\widehat{\alpha}_i \right] &\equiv \mathrm{E}\left[\frac{1}{m} \sum_{j=1}^{m} (Z^{(j;m)})^i \right] \\ &= \mathrm{E}\left[\mathrm{E}\left[(Z^{(j;m)})^i | Z_1^{(0)}, \ldots, Z_m^{(0)} \right] \right] \\ &= \mathrm{E}\left[\frac{1}{m} \sum_{j=1}^{m} (Z_j^{(0)})^i \right] \\ &= \alpha_i. \end{aligned} \tag{21}$$

Thus, $\mathrm{E}[\widehat{X}_m] = \alpha_h$. Notice that

$$\begin{aligned}
\mathrm{E}\left[\widehat{X}_m^2\right] &= \mathrm{E}\left[\left(\frac{1}{m}\sum_{j=1}^{m}(Z^{(j;m)})^h\right)^2\right] \\
&= \frac{1}{m^2}\mathrm{E}\left[\left(\sum_{j=1}^{m}(Z^{(j;m)})^h\right)^2\right] \\
&= \frac{1}{m^2}\mathrm{E}\left[\sum_{j=1}^{m}(Z^{(j;m)})^{2h} + \sum_{i\neq j}(Z^{(i;m)})^h(Z^{(j;m)})^h\right] \\
&= \frac{1}{m^2}\left(m\alpha_{2h} + m(m-1)\mathrm{E}\left[\mathrm{E}[(Z^{(i;m)})^h|Z_1^{(0)},\ldots,Z_m^{(0)}]\cdot\mathrm{E}[(Z^{(j;m)})^h|Z_1^{(0)},\ldots,Z_m^{(0)}]\right]\right) \\
&= \frac{1}{m^2}\left(m\alpha_{2h} + m(m-1)\mathrm{E}\left[\left(\frac{1}{m}\sum_{i=1}^{m}(Z_i^{(0)})^h\right)^2\right]\right) \\
&= \frac{1}{m^2}\left(m\alpha_{2h} + \frac{m(m-1)}{m^2}\mathrm{E}\left[\sum_{i=1}^{m}(Z_i^{(0)})^{2h} + \sum_{i\neq j}(Z_i^{(0)})^h(Z_j^{(0)})^h\right]\right) \\
&= \frac{1}{m^2}\left(m\alpha_{2h} + \frac{m(m-1)}{m^2}(m\alpha_{2h} + m(m-1)\alpha_h^2)\right) \\
&= \frac{1}{m^2}[(2m-1)\alpha_{2h} + (m-1)^2\alpha_h^2] \\
&= \frac{2}{m}\alpha_{2h} + \left(1-\frac{2}{m}\right)\alpha_h^2 + O(m^{-2})
\end{aligned}$$

where $O(m^{-2})$ means terms at most order $1/m^2$. Similar derivations show that

$$\begin{aligned}
\mathrm{E}\left[\widehat{X}_m^3\right] &= \frac{1}{m^4}\Big([m(4m-3)+(m-1)(m-2)]\alpha_{3h} \\
&\qquad + [3m(m-1)^2 + 3(m-1)^2(m-2)]\alpha_h\alpha_{2h} + (m-1)^2(m-2)^2\alpha_h^3\Big) \\
&= \frac{6}{m}\alpha_h\alpha_{2h} + \left(1-\frac{6}{m}\right)\alpha_h^3 + O(m^{-2})
\end{aligned}$$

and

$$\mathrm{E}\left[\widehat{X}_m^4\right] = \frac{12}{m}\alpha_h^2\alpha_{2h} + \left(1-\frac{12}{m}\right)\alpha_h^4 + O(m^{-2}).$$

Thus,

$$\begin{aligned}
\mathrm{E}\left[(\widehat{X}_m - \alpha_h)^4\right] &= \mathrm{E}[\widehat{X}_m^4] - 4\alpha_h \mathrm{E}[\widehat{X}_m^3] + 6\alpha_h^2 \mathrm{E}[\widehat{X}_m^2] - 4\alpha_h^3 \mathrm{E}[\widehat{X}_m] + \alpha_h^4 \qquad (22)\\
&= \frac{12}{m}\alpha_h^2\alpha_{2h} + \left(1 - \frac{12}{m}\right)\alpha_h^4 - 4\alpha_h\left[\frac{6}{m}\alpha_h\alpha_{2h} + \left(1 - \frac{6}{m}\right)\alpha_h^3\right]\\
&\quad + 6\alpha_h^2\left[\frac{2}{m}\alpha_{2h} + \left(1 - \frac{2}{m}\right)\alpha_h^2\right] - 3\alpha_h^4 + O(m^{-2})\\
&= 0 + O(m^{-2})
\end{aligned}$$

because all of the $O(m^{-1})$ terms cancel. Therefore, combining Equations (19), (20) and (22), we have

$$\sum_{m=1}^{\infty} \Pr\{|\widehat{X}_m - \alpha_h| > \epsilon\} \le \sum_{m=1}^{\infty} \frac{c}{m^2\epsilon^4} < \infty$$

where c is some finite constant. Thus, if $\alpha_{4h} < \infty$, then $\widehat{X}_m \overset{a.s.}{\to} \alpha_h$ by the first Borel-Cantelli Lemma in Section 4 of Billingsley (1995).

Since Assumption 2 guarantees $m_\ell \to \infty$ for each moment associated with the ℓth input distribution, we can generalize the almost sure convergence to a vector of moments by applying the converging together lemma. Therefore, we have $\widehat{\mathbf{X}}_{\mathbf{m}} \overset{a.s.}{\to} \mathbf{x}_c$. □

Remark: The independent variables in our stochastic kriging metamodel consist of central moments and standardized central moments. Since standardized moments are continuous functions of raw moments, we can use the continuous mapping theorem to obtain corresponding almost sure convergence of the standardized moments.

Given a fixed and finite number of design points $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$, let $\mathbf{M} = (M(\mathbf{x}_1), M(\mathbf{x}_2), \ldots, M(\mathbf{x}_k))^\top$. The simulation error at design point $\mathbf{x}_i$ is $\epsilon(\mathbf{x}_i)$, so let $\bar{\epsilon}(\mathbf{x}_i) = \sum_{j=1}^{n_i} \epsilon(\mathbf{x}_i)/n_i$ for $i = 1, 2, \ldots, k$ denote the average. Therefore, the sample means of simulation outputs at all design points can be represented as $\bar{\mathbf{Y}}_{\mathcal{D}} = \mathbf{M} + \bar{\epsilon}$, where $\bar{\epsilon} = (\bar{\epsilon}(\mathbf{x}_1), \bar{\epsilon}(\mathbf{x}_2), \ldots, \bar{\epsilon}(\mathbf{x}_k))^\top$. Finally, let $M_p(\cdot)$ be a GP having the conditional distribution of $M(\cdot)$ given $\bar{\mathbf{Y}}_{\mathcal{D}}$.

Lemma 2. *Suppose Assumptions 3–4 hold. Then $M_p(\cdot)$ has continuous sample paths almost surely.*

Proof: Let (Ω_M, P_M) be the underlying probability space for the GP $M(\cdot)$, and $(\Omega_\epsilon, P_\epsilon)$ be the underlying probability space for $\bar{\epsilon}$. Notice that $(\Omega_\epsilon, P_\epsilon)$ depends on the particular design points $\mathbf{x}_1, \mathbf{x}_k, \ldots, \mathbf{x}_k$ and corresponding numbers of replications $n_1, n_2, \ldots, n_k$ which we consider fixed and given, while (Ω_M, P_M) does not.

Let $\omega_M \in \Omega_M$ be an elementary outcome and $M(\cdot, \omega_M)$ the resulting random function. For notational convenience, let $\mathbf{M}(\omega_M) = (M(\mathbf{x}_1, \omega_M), M(\mathbf{x}_2, \omega_M), \ldots, M(\mathbf{x}_k, \omega_M))^\top$

the random function evaluated at $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$. Similarly, $\bar{\epsilon} = \bar{\epsilon}(\omega_\epsilon)$ for elementary outcome $\omega_\epsilon \in \Omega_\epsilon$. Notice that under Assumption 3, $\bar{\epsilon}(\omega_\epsilon)$ has a multivariate normal distribution.

Theorem 3.4.1 of Adler (2010) asserts that there is a P_M-measurable set $\Omega_M^c \subset \Omega_M$ such that $\Pr\{\omega_M \in \Omega_M^c\} = P_M(\Omega_M^c) = 1$, and for every $\omega_M \in \Omega_M^c$ the function $M(\cdot, \omega_M)$ is continuous.

The random variable $\bar{\mathbf{Y}}_{\mathcal{D}}$ maps $\Omega = \Omega_M \times \Omega_\epsilon \to \Re^k$ as $\bar{\mathbf{Y}}_{\mathcal{D}}(\omega) = \mathbf{M}(\omega_M) + \bar{\epsilon}(\omega_\epsilon)$ for $\omega = (\omega_M, \omega_\epsilon) \in \Omega$ with probability measure $P = P_M \cdot P_\epsilon$ since they are independent. Our goal is to prove that

$$\Pr\{\omega_M \in \bar{\Omega}_M^c | \bar{\mathbf{Y}}_D\} = 0 \tag{23}$$

almost surely.

We know that $0 \leq \Pr\{\omega_M \in \bar{\Omega}_M^c | \bar{\mathbf{Y}}_D\} \leq 1$ with probability 1. But also

$$0 = \Pr\{\omega_M \in \bar{\Omega}_M^c\} = \mathrm{E}\left[\Pr\{\omega_M \in \bar{\Omega}_M^c | \bar{\mathbf{Y}}_D\}\right].$$

Therefore, (23) must hold. □

Lemma 3. *Suppose that Assumptions 1–4 hold. Then* $M_p(\widehat{\mathbf{X}}_\mathbf{m}) \xrightarrow{a.s.} M_p(\mathbf{x}_c)$ *as* $m \to \infty$.

Proof: Under Assumption 3, the GP $M(\cdot)$ has continuous sample paths almost surely; applying Lemma 2, $M_p(\cdot)$ also has continuous sample paths almost surely. Under Assumptions 1–2, $\widehat{\mathbf{X}}_\mathbf{m} \xrightarrow{a.s.} \mathbf{x}_c$ as $m \to \infty$ by Lemma 1. And $M_p(\cdot)$ and $\widehat{\mathbf{X}}_\mathbf{m}$ are independent. The result follows by applying the continuous mapping theorem. □

Theorem 1. *Suppose that Assumptions 1–4 hold. Then the interval* $[M_{(\lceil B\frac{\alpha}{2} \rceil)}, M_{(\lceil B(1-\frac{\alpha}{2}) \rceil)}]$ *is asymptotically consistent, meaning*

$$\lim_{m\to\infty} \lim_{B\to\infty} \Pr\{M_{(\lceil B\alpha/2 \rceil)} \leq M_p(\mathbf{x}_c) \leq M_{(\lceil B(1-\alpha/2) \rceil)}\} = 1 - \alpha. \tag{24}$$

Proof: Define $K_\mathbf{m}(t) \equiv \Pr\left\{M_p(\widehat{\mathbf{X}}_\mathbf{m}) \leq t\right\}$. Notice that the distribution $K_\mathbf{m}(t)$ depends on both the distributions of $M_p(\cdot)$ and $\widehat{\mathbf{X}}_\mathbf{m}$. Specifically,

$$\begin{aligned} K_\mathbf{m}(t) &= \int \Pr\left\{M_p(\mathbf{x}) \leq t | \widehat{\mathbf{X}}_\mathbf{m} = \mathbf{x}\right\} d\widehat{F}_{\mathbf{X}_\mathbf{m}}(\mathbf{x}|\mathbf{z}_\mathbf{m}^{(0)}) \\ &= \int \Phi\left(\frac{t - m_p(\mathbf{x})}{\sigma_p(\mathbf{x})}\right) d\widehat{F}_{\mathbf{X}_\mathbf{m}}(\mathbf{x}|\mathbf{z}_\mathbf{m}^{(0)}). \end{aligned}$$

Thus, $K_\mathbf{m}(t)$ is a continuous distribution almost surely. Let $\widehat{K}_\mathbf{m}$ be the empirical cdf of $M_1, M_2, \ldots, M_B$, which are i.i.d. from $K_\mathbf{m}(t)$. Notice that $\widehat{K}_\mathbf{m}^{-1}(\gamma) = M_{(\lceil B\gamma \rceil)}$ for $\gamma = \alpha/2$ and $1 - \alpha/2$.

By the Glivenko-Cantelli Theorem (Van Der Vaart, 1998), $||K_\mathbf{m} - \widehat{K}_\mathbf{m}||_\infty \xrightarrow{a.s.} 0$ as $B \to \infty$. Therefore, by Lemma 21.2 of Van Der Vaart (1998),

$$|M_{(\lceil B\gamma \rceil)} - K_\mathbf{m}^{-1}(\gamma)| \xrightarrow{a.s.} 0$$

as $B \to \infty$ for $\gamma = \alpha/2, 1-\alpha/2$. As a result,

$$\lim_{B\to\infty} \Pr\{M_{(\lceil B\alpha/2\rceil)} \le M_p(\mathbf{x}_c) \le M_{(\lceil B(1-\alpha/2)\rceil)}\} = \Pr\{K_{\mathbf{m}}^{-1}(\alpha/2) \le M_p(\mathbf{x}_c) \le K_{\mathbf{m}}^{-1}(1-\alpha/2)\}.$$

Therefore, Equation (24) becomes

$$\lim_{m\to\infty} \Pr\{K_{\mathbf{m}}^{-1}(\alpha/2) \le M_p(\mathbf{x}_c) \le K_{\mathbf{m}}^{-1}(1-\alpha/2)\} = 1-\alpha. \tag{25}$$

To show Equation (25), we only need to show that

$$\lim_{m\to\infty} \Pr\{K_{\mathbf{m}}^{-1}(\alpha/2) > M_p(\mathbf{x}_c)\} = \alpha/2$$

because the proof of the upper bound is similar.

Since, conditional on $\bar{\mathbf{Y}}_{\mathcal{D}}$, $M_p(\mathbf{x}_c) \sim \mathrm{N}(m_p(\mathbf{x}_c), \sigma_p^2(\mathbf{x}_c))$, the cdf $H(t) \equiv \Pr\{M_p(\mathbf{x}_c) \le t\}$ is continuous. By Lemma 3 and Lemma 2.11 in Van Der Vaart (1998),

$$\begin{aligned} &\sup_t |\Pr\{M_p(\widehat{\mathbf{X}}_{\mathbf{m}}) \le t\} - \Pr\{M_p(\mathbf{x}_c) \le t\}| \\ &= \| K_{\mathbf{m}} - H \|_\infty \to 0 \text{ as } m \to \infty. \end{aligned}$$

Therefore,

$$\begin{aligned} &\Pr\{K_{\mathbf{m}}^{-1}(\alpha/2) > M_p(\mathbf{x}_c)\} \\ &= \Pr\{\alpha/2 \ge K_{\mathbf{m}}(M_p(\mathbf{x}_c))\} \\ &= \Pr\{\alpha/2 \ge H(M_p(\mathbf{x}_c))\} + o(1) \qquad (26) \\ &= \Pr\{M_p(\mathbf{x}_c) \le H^{-1}(\alpha/2)\} + o(1) \\ &= \alpha/2 + o(1). \end{aligned}$$

Equation (26) is obtained because

$$|K_{\mathbf{m}}(M_p(\mathbf{x}_c)) - H(M_p(\mathbf{x}_c))| \le \| K_{\mathbf{m}} - H \|_\infty \to 0 \text{ as } m \to \infty.$$

Thus, we have

$$\begin{aligned} &\lim_{m\to\infty} \Pr\{K_{\mathbf{m}}^{-1}(\alpha/2) \le M_p(\mathbf{x}_c) \le K_{\mathbf{m}}^{-1}(1-\alpha/2)\} \\ &= \lim_{m\to\infty} \Pr\{M_p(\mathbf{x}_c) \le K_{\mathbf{m}}^{-1}(1-\alpha/2)\} - \lim_{m\to\infty} \Pr\{M_p(\mathbf{x}_c) < K_{\mathbf{m}}^{-1}(\alpha/2)\} \\ &= (1-\alpha/2) - \alpha/2 \\ &= 1-\alpha. \end{aligned}$$

□

8.2 Asymptotic Analysis of Variance Component Estimators

Theorem 2. *Suppose that Assumptions 1–4 hold. Then the variance component estimators* $\widehat{\sigma}^2_M, \widehat{\sigma}^2_I, \widehat{\sigma}^2_T$ *are consistent as* $m, B \to \infty$.

Proof: When a GP $M(\cdot)$ has a continuous correlation function with all parameters finite, the SK predictor

$$m_p(\mathbf{x}) = \widehat{\beta}_0 + \tau^2 R(\mathbf{x})^\top [\Sigma + C]^{-1} (\bar{\mathbf{Y}}_{\mathcal{D}} - \widehat{\beta}_0 \cdot 1_{k\times 1}), \tag{27}$$

and corresponding variance

$$\sigma^2_p(\mathbf{x}) = \tau^2 - \tau^4 R(\mathbf{x})^\top [\Sigma + C]^{-1} R(\mathbf{x}) + \eta^\top [1^\top_{k\times 1} (\Sigma + C)^{-1} 1_{k\times 1}]^{-1} \eta$$

where $R(\mathbf{x})^\top = (r(\mathbf{x} - \mathbf{x}_1), r(\mathbf{x} - \mathbf{x}_2), \ldots, r(\mathbf{x} - \mathbf{x}_k))$ and $\eta = 1 - 1^\top_{k\times 1}(\Sigma + C)^{-1}\tau^2 R(\mathbf{x})$, are continuous and bounded functions of $\mathbf{x}$.

By the Strong Law of Large Numbers, the raw moment estimator $\mathbf{X_m} \overset{a.s.}{\to} \mathbf{x}_c$ as $m \to \infty$ under Assumptions 1–2. This almost sure convergence can be extended to central moments and standardized central moments by the continuous mapping theorem. By applying the Portmanteau Lemma in Van Der Vaart (1998), we have

$$\lim_{m\to\infty} \sigma^2_M = \lim_{m\to\infty} \int \sigma^2_p(\mathbf{x})\, dF^c_{\mathbf{X_m}}(\mathbf{x}) = \lim_{m\to\infty} \mathrm{E}[\sigma^2_p(\mathbf{X_m})] = \sigma^2_p(\mathbf{x}_c),$$

and

$$\begin{aligned}
\lim_{m\to\infty} \sigma^2_I &= \lim_{m\to\infty} \int (m_p(\mathbf{x}) - \mu_0)^2\, dF^c_{\mathbf{X_m}}(\mathbf{x}) \\
&= \lim_{m\to\infty} \mathrm{E}\left[\big(m_p(\mathbf{X_m}) - \mathrm{E}[m_p(\mathbf{X_m})]\big)^2 \right] \\
&= \big(m_p(\mathbf{x}_c) - m_p(\mathbf{x}_c)\big)^2 = 0
\end{aligned}$$

where $\mu_0 = \int\int \nu\, dF(\nu|\mathbf{x})\, dF^c_{\mathbf{X_m}}(\mathbf{x}) = \int m_p(\mathbf{x})\, dF^c_{\mathbf{X_m}}(\mathbf{x})$.

Recall that $F(\nu|\mathbf{x})$ is a normal distribution $\mathrm{N}(m_p(\mathbf{x}), \sigma^2_p(\mathbf{x}))$. Let $g(\mathbf{x}) \equiv \int (\nu - \mu_0)^2\, dF(\nu|\mathbf{x})$. Then

$$\begin{aligned}
\lim_{m\to\infty} \sigma^2_T &= \lim_{m\to\infty} \int\int (\nu - \mu_0)^2\, dF(\nu|\mathbf{x})\, dF^c_{\mathbf{X_m}}(\mathbf{x}) \\
&= \lim_{m\to\infty} \int g(\mathbf{x})\, dF^c_{\mathbf{X_m}}(\mathbf{x}).
\end{aligned}$$

However,

$$\begin{aligned}
g(\mathbf{x}) &= \int (\nu - m_p(\mathbf{x}) + m_p(\mathbf{x}) - \mu_0)^2\, dF(\nu|\mathbf{x}) \\
&= \int (\nu - m_p(\mathbf{x}))^2\, dF(\nu|\mathbf{x}) + (m_p(\mathbf{x}) - \mu_0)^2 + (m_p(\mathbf{x}) - \mu_0) \int (\nu - m_p(\mathbf{x}))\, dF(\nu|\mathbf{x}) \\
&= \sigma^2_p(\mathbf{x}) + (m_p(\mathbf{x}) - \mu_0)^2 + 0.
\end{aligned}$$

Since $m_p(\mathbf{x})$ and $\sigma_p^2(\mathbf{x})$ are continuous and bounded functions, so is $g(\mathbf{x})$. Therefore,

$$\lim_{m\to\infty} \sigma_T^2 = \lim_{m\to\infty} \mathrm{E}[g(\mathbf{X_m})] = g(\mathbf{x}_c) = \sigma_p^2(\mathbf{x}_c)$$

by applying the Portmanteau Lemma.

Next, we will show consistency of the variance estimators. By Lemma 1, $\widehat{\mathbf{X}}_\mathbf{m} \overset{a.s.}{\to} \mathbf{x}_c$. For the metamodel uncertainty estimator, we have

$$\begin{aligned}
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_M^2 &= \lim_{m\to\infty}\lim_{B\to\infty} \frac{1}{B}\sum_{b=1}^{B} \sigma_p^2(\widehat{\mathbf{X}}_\mathbf{m}^{(b)}) \\
&= \lim_{m\to\infty} \mathrm{E}\Big[\sigma_p^2(\widehat{\mathbf{X}}_\mathbf{m})|\mathbf{Z}_\mathbf{m}^{(0)}\Big] \\
&= \sigma_p^2(\mathbf{x}_c).
\end{aligned}$$

The last step follows by applying the Portmanteau Lemma.

For the input uncertainty estimator, we have

$$\begin{aligned}
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_I^2 &= \lim_{m\to\infty}\lim_{B\to\infty} \frac{B}{B-1}\left[\frac{1}{B}\sum_{b=1}^{B} m_p^2(\widehat{\mathbf{X}}_\mathbf{m}^{(b)}) - \bar{\mu}^2\right] \\
&= \lim_{m\to\infty}\left(\mathrm{E}\Big[m_p^2(\widehat{\mathbf{X}}_\mathbf{m})|\mathbf{Z}_\mathbf{m}^{(0)}\Big] - \mathrm{E}^2\Big[m_p(\widehat{\mathbf{X}}_\mathbf{m})|\mathbf{Z}_\mathbf{m}^{(0)}\Big]\right) \\
&= m_p^2(\mathbf{x}_c) - m_p^2(\mathbf{x}_c) = 0.
\end{aligned}$$

The last step follows by applying Lemma 1 and the Portmanteau Lemma.

For the total variance estimator, we have

$$\begin{aligned}
\lim_{m\to\infty}\lim_{B\to\infty} \widehat{\sigma}_T^2 &= \lim_{m\to\infty}\lim_{B\to\infty} \frac{B}{B-1}\left(\frac{1}{B}\sum_{b=1}^{B} M_b^2 - \bar{M}^2\right) \\
&= \lim_{m\to\infty} \mathrm{E}\big[M_p^2(\widehat{\mathbf{X}}_\mathbf{m})|\mathbf{Z}_\mathbf{m}^{(0)}\big] - \lim_{m\to\infty} \mathrm{E}^2\big[M_p(\widehat{\mathbf{X}}_\mathbf{m})|\mathbf{Z}_\mathbf{m}^{(0)}\big] \\
&= \mathrm{E}[M_p^2(\mathbf{x}_c)] - \mathrm{E}^2[M_p(\mathbf{x}_c)] \qquad (28) \\
&= m_p^2(\mathbf{x}_c) + \sigma_p^2(\mathbf{x}_c) - m_p^2(\mathbf{x}_c) \\
&= \sigma_p^2(\mathbf{x}_c).
\end{aligned}$$

By Lemma 3, $M_p(\widehat{\mathbf{X}}_\mathbf{m}) \overset{a.s.}{\to} M_p(\mathbf{x}_c)$ as $m \to \infty$. Then Step (28) follows by applying Portmanteau Lemma. □

Theorem 3. *Suppose that Assumptions 1–4 and the following additional assumptions hold:*

5. *The first three derivatives of the correlation function of the GP $M(\mathbf{x})$ exist and the third derivative is bounded; and*

6. $m_\ell/m \to 1$ *for* $\ell = 1, 2, \ldots, L$.

Then $\lim_{m\to\infty} m\sigma_I^2 = \lim_{m\to\infty} \lim_{B\to\infty} m\widehat{\sigma}_I^2 = \sigma_\mu^2$ *almost surely, where* σ_μ^2 *is a positive constant.*

Proof: Under Assumptions 1–2, and applying the multivariate central limit theorem, we have as $m \to \infty$,

$$\sqrt{m}(\mathbf{X}_m - \mathbf{x}_c) \xrightarrow{D} \mathrm{N}(\mathbf{0}_{d\times 1}, \Lambda)$$

where Λ denotes the $d \times d$ positive definite asymptotic variance-covariance matrix of $\mathbf{X_m}$.

When a GP $M(\mathbf{x})$ has a continuous correlation function with all parameters finite, the SK predictor

$$m_p(\mathbf{x}) = \widehat{\beta}_0 + \tau^2 R(\mathbf{x})^\top [\Sigma + C]^{-1}(\bar{\mathbf{Y}}_\mathcal{D} - \widehat{\beta}_0 \cdot 1_{k\times 1}), \tag{29}$$

given the simulation sample mean $\bar{\mathbf{Y}}_\mathcal{D}$, is continuous and bounded. Under Assumption 5, the gradient $\nabla m_p(\mathbf{x})$ exists and is continuous. We will show that $\nabla m_p(\mathbf{x}) \neq \mathbf{0}_{d\times 1}$ almost surely. By taking the derivative of $m_p(\mathbf{x})$ in Equation (29), we have

$$\frac{\partial m_p(\mathbf{x})}{\partial x_j} = \underbrace{\frac{\partial R(\mathbf{x})^\top}{\partial x_j}\tau^2[\Sigma + C]^{-1}}_{\mathbf{A}}(\bar{\mathbf{Y}}_\mathcal{D} - \widehat{\beta}_0 \cdot 1_{k\times 1}). \tag{30}$$

Since $\partial R(\mathbf{x})^\top/\partial x_j = (\partial r(\mathbf{x} - \mathbf{x}_1)/\partial x_j, \partial r(\mathbf{x} - \mathbf{x}_2)/\partial x_j, \ldots, \partial r(\mathbf{x} - \mathbf{x}_k)/\partial x_j) \neq \mathbf{0}_{1\times k}$ and $\tau^2[\Sigma + C]^{-1}$ is positive definite, $\mathbf{A}$ is a non-zero constant vector. Under Assumption 3, $\mathbf{A}(\bar{\mathbf{Y}}_\mathcal{D} - \widehat{\beta}_0 \cdot 1_{k\times 1})$ is a normal random variable that is equal to 0 with probability 0. Thus, $\nabla m_p(\mathbf{x}) \neq \mathbf{0}_{d\times 1}$ almost surely. Applying Theorem 13.1 in Severini (2005), we have

$$\sqrt{m}(m_p(\mathbf{X}_m) - m_p(\mathbf{x}_c)) \xrightarrow{D} \mathrm{N}(0, \sigma_\mu^2)$$

where $\sigma_\mu^2 = \nabla m_p(\mathbf{x}_c)^\top \Lambda \nabla m_p(\mathbf{x}_c) > 0$. This establishes the constant.

Since $m_p(\cdot)$ is continuous and bounded, there always exists a finite $M_1 > 0$ such that $|m_p(\mathbf{x})| < M_1$ for all $\mathbf{x} \in \Re^d$. Therefore, $\max_{\mathbf{x}\in\Re^d} |m_p(\mathbf{x}) - m_p(\mathbf{X_m})| < 2M_1$. Let $\tau_m = \mathrm{e}^{m^{1/4}}$. Since $2M_1/\tau_m \to 0$ as $m \to \infty$, Condition (17) of Theorem 3.8 of Shao and Tu (1995) holds. Thus, the bootstrap variance estimator is strongly consistent: $\lim_{m\to\infty} \lim_{B\to\infty} m\widehat{\sigma}_I^2 = \sigma_\mu^2$ almost surely.

Next, we will show $\lim_{m\to\infty} m\sigma_I^2 = \sigma_\mu^2$ by proving a multi-variate version of Theorem 1.1 in Lehmann and Casella (1998), Chapter 6. Let $L(\mathbf{X_m}, \mathbf{x}_c)$ denote the line segment joining $\mathbf{X_m}$ and $\mathbf{x}_c$. By the Multivariate Taylor Formula (Serfling, 2002),

$$m_p(\mathbf{X_m}) = m_p(\mathbf{x}_c) + \nabla m_p(\mathbf{x}_c)^\top(\mathbf{X_m} - \mathbf{x}_c) + \frac{1}{2}(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla^2 m_p(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c) + \mathcal{R}(\mathbf{X_m}, \mathbf{x}_c).$$

The remainder term

$$\mathcal{R}(\mathbf{X_m}, \mathbf{x}_c) = \frac{1}{3!}\sum_{i_1=1}^{d}\sum_{i_2=1}^{d}\sum_{i_3=1}^{d} \frac{\partial^3 m_p(x_1, \ldots, x_d)}{\partial x_{i_1}\partial x_{i_2}\partial x_{i_3}}\bigg|_{\mathbf{x}=\mathbf{z}} \prod_{j=1}^{3}(X_{\mathbf{m},i_j} - x_{c,i_j})$$

where $\mathbf{z}$ denotes a value in the interior of $L(\mathbf{X_m}, \mathbf{x}_c)$, and $X_{\mathbf{m},i}$ and $x_{c,i}$ denote the ith components of the vectors $\mathbf{X_m}$ and $\mathbf{x}_c$. By taking the expectation over both sides, we have

$$\mathrm{E}[m_p(\mathbf{X_m})] = m_p(\mathbf{x}_c) + \frac{1}{2}\mathrm{E}\left[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla^2 m_p(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)\right] + \mathrm{E}[\mathcal{R}(\mathbf{X_m}, \mathbf{x}_c)] \tag{31}$$

where ∇^2 is the Hessian operator.

We will show that the second and third terms on the RHS of Equation (31) are $O(m^{-1})$ and $O(m^{-2})$, respectively, under Assumption 5. Since all of the input processes are independent, we establish the result for one input distribution F^c without loss of generality.

We prove the result for $\mathbf{x}_c$ being the generic hth-order moment, $x_{c,h} = \mathrm{E}(Z_1^h) < \infty$, and $X_{m,h} = m^{-1}\sum_{j=1}^m Z_j^h$ for $Z_j \overset{\text{iid}}{\sim} F^c$.

Let $C_{ij} \equiv \frac{1}{2}[\nabla^2 m_p(\mathbf{x}_c)]_{i,j}$. We first consider components of the second term on the RHS of Equation (31).

$$\begin{aligned}
\mathrm{E}[C_{ij}(X_{m,i} - x_{c,i})(X_{m,j} - x_{c,j})] &= C_{ij}\mathrm{E}\left[\frac{1}{m}\sum_{k_1=1}^m (Z_{k_1}^i - x_{c,i}) \cdot \frac{1}{m}\sum_{k_2=1}^m (Z_{k_2}^j - x_{c,j})\right] \\
&= \frac{C_{ij}}{m^2}\mathrm{E}\left[\sum_{k=1}^m (Z_k^i - x_{c,i})(Z_k^j - x_{c,j}) + \sum_{k_1 \neq k_2} (Z_{k_1}^i - x_{c,i})(Z_{k_2}^j - x_{c,j})\right] \\
&= \frac{C_{ij}}{m^2}\mathrm{E}\left[\sum_{k=1}^m (Z_k^i - x_{c,i})(Z_k^j - x_{c,j}) + 0\right] = O\left(\frac{1}{m}\right).
\end{aligned}$$

The last two steps follow because the Z_i are i.i.d. and Assumption 5 holds. Thus, the second term on the RHS of Equation (31) is

$$\frac{1}{2}\mathrm{E}[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla^2 m_p(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)] = \sum_{i=1}^d \sum_{j=1}^d \mathrm{E}[C_{ij}(X_{m,i} - x_{c,i})(X_{m,j} - x_{c,j})] = O\left(\frac{1}{m}\right).$$

Similarly, for the components of the third term of the RHS of Equation (31), we have

$$\begin{aligned}
&D_{ijk}\mathrm{E}[(X_{m,i} - x_{c,i})(X_{m,j} - x_{c,j})(X_{m,k} - x_{c,k})] \\
&= D_{ijk}\mathrm{E}\left[\frac{1}{m^3}\sum_{k_1=1}^m (Z_{k_1}^i - x_{c,i}) \cdot \sum_{k_2=1}^m (Z_{k_2}^j - x_{c,j}) \cdot \sum_{k_3=1}^m (Z_{k_3}^k - x_{c,k})\right] \\
&= \frac{D_{ijk}}{m^3}\mathrm{E}\left[\sum_{k_1=1}^m (Z_{k_1}^i - x_{c,i})(Z_{k_1}^j - x_{c,j})(Z_{k_1}^k - x_{c,k}) + 0\right] = O\left(\frac{1}{m^2}\right).
\end{aligned}$$

where

$$D_{ijk} \equiv \frac{1}{3!}\left.\frac{\partial^3 m_p(x_1, \ldots, x_d)}{\partial x_i \partial x_j \partial x_k}\right|_{\mathbf{x}=\mathbf{z}}.$$

Again, the last two steps follow because the Z_i are i.i.d. and Assumption 5 holds. Thus, the third term in Equation (31) is

$$\mathrm{E}[\mathcal{R}(\mathbf{X_m}, \mathbf{x}_c)] = \sum_{i=1}^{d}\sum_{j=1}^{d}\sum_{k=1}^{d} D_{ijk}\mathrm{E}[(X_{m,i} - x_{c,i})(X_{m,j} - x_{c,j})(X_{m,k} - x_{c,k})] = O\left(\frac{1}{m^2}\right).$$

Squaring both sides of Equation (31), we have

$$[\mathrm{E}(m_p(\mathbf{X_m}))]^2 = m_p^2(\mathbf{x}_c) + m_p(\mathbf{x}_c)\mathrm{E}[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla^2 m_p(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)] + O\left(\frac{1}{m^2}\right). \quad (32)$$

By repeating the same derivation that results in Equation (31) but using $m_p^2(\cdot)$ instead of $m_p(\cdot)$, we obtain

$$\begin{aligned}
\mathrm{E}[m_p^2(\mathbf{X_m})] &= m_p^2(\mathbf{x}_c) + \frac{1}{2}\mathrm{E}[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla^2 m_p^2(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)] + O\left(\frac{1}{m^2}\right) \\
&= m_p^2(\mathbf{x}_c) + \mathrm{E}\Big[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla m_p(\mathbf{x}_c)\nabla m_p(\mathbf{x}_c)^\top (\mathbf{X_m} - \mathbf{x}_c) \\
&\quad + (\mathbf{X_m} - \mathbf{x}_c)^\top m_p(\mathbf{x}_c)\nabla^2 m_p(\mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)\Big] + O\left(\frac{1}{m^2}\right). \quad (33)
\end{aligned}$$

Then,

$$\begin{aligned}
\mathrm{Var}[m_p(\mathbf{X_m})] &= \mathrm{E}[m_p^2(\mathbf{X_m})] - \Big(\mathrm{E}[m_p(\mathbf{X_m})]\Big)^2 \\
&= \mathrm{E}[(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla m_p(\mathbf{x}_c)\nabla m_p(\mathbf{x}_c)^\top (\mathbf{X_m} - \mathbf{x}_c)] + O\left(\frac{1}{m^2}\right) \\
&= \mathrm{E}[\nabla m_p(\mathbf{x}_c)^\top (\mathbf{X_m} - \mathbf{x}_c)(\mathbf{X_m} - \mathbf{x}_c)^\top \nabla m_p(\mathbf{x}_c)] + O\left(\frac{1}{m^2}\right) \quad (34) \\
&= \frac{1}{m}\nabla m_p(\mathbf{x}_c)^\top \Lambda \nabla m_p(\mathbf{x}_c) + O\left(\frac{1}{m^2}\right).
\end{aligned}$$

Step (34) follows because $\nabla m_p(\mathbf{x}_c)^\top (\mathbf{X_m} - \mathbf{x}_c)$ is a scalar. Thus, we have $\lim_{m\to\infty} m\sigma_I^2 = \sigma_\mu^2$. □

Remark: The independent variables in our stochastic kriging metamodel consist of central moments and standardized central moments, rather than raw moments. However, Theorem 3 can easily be extended to central and standardized central moments as follows.

Since standardized moments are continuous functions of raw moments, denoted generically as $g(\cdot)$, we can consider the composite function $(m_p \circ g)(\cdot)$ and follow steps analogous to those in the proof of Theorem 3. Up to the third derivatives we have

$$\begin{aligned}
(m_p \circ g)'(t) &= m_p'(g(t))g'(t) \\
(m_p \circ g)''(t) &= m_p''(g(t))[g'(t)]^2 + m_p'(g(t))g''(t) \\
(m_p \circ g)^{(3)}(t) &= m_p^{(3)}(g(t))[g'(t)]^3 + 2m_p''(g(t))g'(t)g''(t) + m_p''(g(t))g'(t)g''(t) + m_p'(g(t))g^{(3)}(t).
\end{aligned}$$

Let u denote the mean, u_i' denote the ith order raw moment and u_i denote the ith order central moment. Then the first three central moments can be expressed as functions of raw moments as follows:

$$\begin{aligned} u_1 &= u, \\ u_2 &= u_2' - u^2 \\ u_3 &= u_3' - 3uu_2' + 2u^3. \end{aligned}$$

The first three standardized central moments are $u_1, \sqrt{u_2}$ and $u_3/u_2^{3/2}$. For a non-degenerate distribution, the second central moment is positive and bounded away from 0. Thus, the first three derivatives $g', g'', g^{(3)}$ exist and are finite.

8.3 Experiment Design

To fit SK metamodels we recommend the experiment design developed in Barton et al. (2014) which demonstrated robust performance over a number of test examples. In this section, we briefly review the basic methodology; for detailed information please refer to Barton et al. (2014).

The experiment design is not specified a priori; instead the design space, denoted by $\mathcal{D}$, depends on the real-world data $\mathbf{z}_{\mathbf{m}}^{(0)}$ that will eventually be resampled. In this way the design is adaptive.

At a high level, this is the approach: Generate a large number of bootstrap samples from the real-world data $\mathbf{z}_{\mathbf{m}}^{(0)}$ and compute the corresponding sample moments. Find a regular region that encompasses a large fraction of this sample; this will be the design space. Generate additional bootstrap samples to test that the regular region does indeed cover the desired fraction of the feasible space of sample moments, and refine if necessary. Once satisfied, embed a space-filling design in the regular region. These design points correspond to input distribution moments at which to run simulation experiments to fit the SK metamodel. We provide some more details below.

Suppose we are interested in a $(1-\alpha)100\%$ CI; we set $\alpha = 0.05$ in our empirical study. We want the experiment design to lead to a metamodel that is accurate for moments $\mathbf{x}$ that are the most likely bootstrap moment vectors generated from $\mathbf{z}_{\mathbf{m}}^{(0)}$; by "likely" we mean, for instance, covering $q = 99\% > (1-\alpha)100\% = 95\%$ of the feasible bootstrap moments.

To this end we find an ellipsoid that will contain an independent bootstrap moment vector obtained by random sampling from $\mathbf{z}_{\mathbf{m}}^{(0)}$ with probability at least q. We then generate a space-filling experiment design inside this ellipsoid. The procedure for constructing the design is as follows:

1. Generate B_0 bootstrap resamples from $\mathbf{z}_{\mathbf{m}}^{(0)}$ and compute the corresponding sample moments to generate a set of sample moments $D_T = \{\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}, b = 1, 2, \ldots, B_0\}$.

2. Find the smallest ellipsoid E such that it contains the fraction q of the data in D_T when the ellipsoid's center and shape are the sample mean and covariance matrix, respectively, of the elements of D_T.

3. Perform a hypothesis test where the null hypothesis is that a bootstrap moment will be contained in this ellipsoid with probability at least q. This requires computing the number of bootstrap moment resamples, denoted by B_1, and the constant c that defines the rejection region to attain the desired Type I error and power for the test.

4. Generate B_1 additional independent bootstrap resamples from $\mathbf{z}_{\mathbf{m}}^{(0)}$ and compute the moments $\widehat{\mathbf{X}}_{\mathbf{m}}^{(b)}, b = B_0+1, B_0+2, \ldots, B_0+B_1$. If more than c of these B_1 resamples are contained in the ellipsoid, then accept the current E as the design space. Otherwise, add these bootstrap resamples to D_T, let $B_0 \leftarrow B_0 + B_1$ and go to Step 2 to update the ellipsoid.

5. Generate k space-filling design points in the ellipsoid E. To place design points into this space, we employ an algorithm due to Sun and Farooq (2002), §3.2.1, for generating points uniformly distributed in an ellipsoid. The algorithm first generates the polar coordinates of a point uniformly distributed in a hypersphere, then transforms it to Cartesian coordinates, and finally transforms it again to a point uniformly distributed in an ellipsoid. The advantage of this approach is that each element of the initial polar coordinates are independently distributed, allowing them to be generated coordinate by coordinate via their inverse cumulative distribution function. Rather than use randomly chosen points, however, we begin with a Latin hypercube sample on $(0,1)^d$.

6. Assign $n = N/k$ replications to each design point, where N denotes total computational budget. Together the transformed Latin hypercube design points and the number of replications n define the experiment design $\mathcal{D}$.

In our experiments we set Type I error of the hypothesis test to 0.005 and its power to 0.95 when the true probability is $q = 0.97$.

8.4 Sensitivity of Inference to SK Parameter Estimation Error

Since the parameters $(\beta_0, \tau^2, \boldsymbol{\theta}, C)$ are unknown, we use estimators $(\widehat{\beta}_0, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}}, \widehat{C})$ to form a SK metamodel. However, the properties of SK, and in particular Theorems 1–3, have only been established when at least $(\tau^2, \boldsymbol{\theta}, C)$ are known. Nevertheless, kriging and SK have been observed to provide robust inference without accounting for parameter-estimation error provided we employ an adequate experiment design. Here we report a small-scale empirical study that examines parameter sensitivity for our particular problem: forming an ACI for $\mu(\mathbf{x}_c)$ and assessing the relative contribution of input uncertainty.

When we use the plug-in estimator $\widehat{C}$, we get an unbiased SK predictor $\widehat{m}_p(\mathbf{x})$ and small variance inflation, based on results in Ankenman et al. (2010). Therefore, we focus on sensitivity to the parameters $\boldsymbol{\phi} = (\beta_0, \tau^2, \boldsymbol{\theta})$. These parameters are estimated by maximum likelihood using the log-likelihood function

$$\ell(\boldsymbol{\phi}) = -\frac{k}{2}\ln(2\pi) - \frac{1}{2}\ln[|\Sigma + C|] - \frac{1}{2}(\bar{\mathbf{Y}}_{\mathcal{D}} - \beta_0 \cdot 1_{k\times 1})^\top (\Sigma + C)^{-1}(\bar{\mathbf{Y}}_{\mathcal{D}} - \beta_0 \cdot 1_{k\times 1})$$

Table 4: Sensitivity of the ACI to the estimation of $(\widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$.

$k = 40$, $n = 50$	Case 1			Case 2		
	$m = 50$	$m = 500$	$m = 5000$	$m = 50$	$m = 500$	$m = 5000$
Coverage of CI_+	95.6%	97.9%	96.4%	95.5%	98.1%	95.5%
CI_+ Width (mean)	340	29.4	4.6	340	29.3	4.4
CI_+ Width (SD)	171	18.8	0.93	170	18.8	0.91
$\widehat{\sigma}_I/\widehat{\sigma}_T$	0.972	0.960	0.925	0.971	0.963	0.922

where Σ is a function of τ^2 and $\boldsymbol{\theta}$. The only random variable in the log-likelihood is $\mathbf{Y}_\mathcal{D}$, and the estimation uncertainty of the MLE $(\widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$ is a complex function of the sampling distribution of $\mathbf{Y}_\mathcal{D}$ (the contribution to uncertainty due to $\widehat{\beta}_0$ is tractable).

To study the sensitivity of our ACI to the estimation error of $(\widehat{\beta}_0, \widehat{\tau}^2, \widehat{\boldsymbol{\theta}})$, we again use the queueing network example in Section 6. In each macro-replication, we generate m real-world observations from each input model, find k design points and run simulations to obtain $\bar{\mathbf{Y}}_\mathcal{D}$ and $\widehat{C}$, now denoted by $\bar{\mathbf{Y}}_1$ and $\widehat{C}_1$. Then we compare the performance of our method under two settings:

Case 1: Use $\bar{\mathbf{Y}}_1$ and $\widehat{C}_1$ to compute the MLEs for $(\widehat{\beta}_{01}, \widehat{\tau}_1^2, \widehat{\boldsymbol{\theta}}_1)$, use them to build the SK metamodel and construct CI_+ as before (Section 5.1).

Case 2: Using the same experiment design, draw another independent sample of simulation outputs to obtain $\bar{\mathbf{Y}}_\mathcal{D}$ and $\widehat{C}$, denoted by $\bar{\mathbf{Y}}_2$ and $\widehat{C}_2$, and obtain the corresponding MLEs $(\widehat{\beta}_{02}, \widehat{\tau}_2^2, \widehat{\boldsymbol{\theta}}_2)$. Use these estimates along with $\bar{\mathbf{Y}}_1$ and $\widehat{C}_1$ to build the metamodel and again construct CI_+.

Notice that in Case 2 we are obtaining GP parameter estimates from a sample of data that is independent of the data that forms the metamodel.

Table 4 shows the coverage and contribution results based on 1000 macro-replications. The nearly identical performance of Cases 1 and 2 demonstrates that our procedure is not sensitive to SK parameter estimation error if we employ the one-stage space-filling design used in the paper.