IIE Transactions

Publication details, including instructions for authors and subscription information:
http://www.tandfonline.com/loi/uiie20

Quickly Assessing Contributions to Input Uncertainty

Eunhye Song[a] & Barry L. Nelson[a]

[a] Department of Industrial Engineering & Management Sciences Northwestern University Evanston, IL 60208 USA E-mail:

Accepted author version posted online: 17 Nov 2014.Published online: 17 Nov 2014.

IIE Transactions (2015) **47**, 1–17

ISSN: 0740-817X print / 1545-8830 online
DOI: 10.1080/0740817X.2014.980869

Quickly assessing contributions to input uncertainty

EUNHYE SONG and BARRY L. NELSON*

Department of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60208, USA
E-mail: nelsonb@northwestern.edu

Received February 2014 and accepted October 2014

"Input uncertainty" refers to the (often unmeasured) variability in simulation-based performance estimators that is a consequence of driving the simulation with input models that are based on real-world data. Several methods have been proposed to assess the overall effect of input uncertainty, and some also support attributing this uncertainty to the various input models. However, these methods require a lengthy sequence of diagnostic experiments. This paper provides a method to obtain an estimator of the overall variance due to input uncertainty, the relative contribution to this variance of each input distribution, and a measure of the sensitivity of overall uncertainty to increasing the real-world sample-size used to fit each distribution, all from a *single* diagnostic experiment. The approach exploits a metamodel that relates the means and variances of the input distributions to the mean response of the simulation output, and also employs bootstrapping of the real-world data to represent input-model uncertainty. Furthermore, whether and how the simulation outputs from the nominal and diagnostic experiments may be combined to obtain a better performance estimator is investigated. For the case when the analyst obtains additional real-world data, refines the input models, and runs a follow-up experiment, an analysis of whether and how the simulation outputs from all three experiments should be combined is presented. Numerical illustrations are provided.

Keywords: Stochastic simulation, input modeling, input uncertainty, output analysis

1. Introduction

There is increasing recognition of the need to quantify all sources of error in mathematical and computer models, including stochastic simulations. Every simulation language measures the statistical error due to sampling from the input models, typically via Confidence Intervals (CIs) on the performance measures. However, these CIs do not account for the possible (in fact, likely) misspecification of the input models when they are estimated from real-world data. For instance, later in this article we consider the simulation of a remote order-taking system for customers using a drive-in service at a chain of fast-food restaurants; this simulation was created to estimate a measure of customer delay. Real-world data on customer arrivals, the time it takes an agent to obtain a customer's order, and the time needed for a car to move beyond the order board are used to fit input models that drive the simulation. Because we only have a finite quantity of real-world data, these input models are imperfect representations of the actual processes. As shown in many papers (e.g., Barton (2012), Barton *et al.* (2014), Cheng and Holland (1998, 2004), Chick (2001), and Zouaoui and Wilson (2003, 2004) the error due to "input uncertainty" can overwhelm the simulation sampling error. These papers provide overall measures of input uncertainty, such as adjusted CIs or Bayesian credible intervals, where as we focus on assessing the contribution of each input model to input uncertainty as a guide to collecting more real-world data.

A predecessor of this article, Ankenman and Nelson (2012) presented a quick-and-easy diagnostic experiment to assess the overall effect of input uncertainty relative to simulation sampling variability, and a follow-up method for estimating contributions. Unfortunately, their method for identifying the input models that contribute the most to input uncertainty requires a sequence of additional diagnostic experiments; in the worst case it requires as many experiments as there are input models, and each of these experiments can be substantial. Furthermore, the variance model that underlies their diagnostic experiments has no rigorous justification.

In this article, we provide a new analysis that requires only one diagnostic experiment to assess the overall effect of input uncertainty, the relative contribution of each input distribution, and a measure of sample size sensitivity of each distribution. Using these results, the analyst can decide whether it is worth the time and expense to collect additional data and on which input processes to do so. If the analyst decides to collect additional real-world

* Corresponding author

input data to refine the input models, then yet another simulation experiment would be conducted. Thus, there are potentially three sets of simulation output data: nominal experiment, diagnostic experiment, and follow-up experiment. We study when and how these data may be combined to produce better performance estimates.

We obtain our measures of overall input uncertainty, contributions, and sample size sensitivities using the following approach:

1. Following the nominal experiment, we take repeated bootstrap samples from the real-world data and use these data to create alternative sets of input distributions representing what could have occurred with different real-world samples.
2. Using these alternative input distributions we fit a regression metamodel that relates the mean of the simulation output to the means and variances of the input distributions.
3. From the metamodel, we derive expressions for the overall variance in the simulation output due to input uncertainty, the contribution to this variance of each input distribution, and the reduction in overall variance that would result from one additional real-world sample of data to fit each input distribution.

The measures computed in step 3 can be used to guide additional real-world data collection and to heuristically adjust measures of error, such as CIs, to account for both input and simulation variability.

Ours is not the first attempt to decide from which input processes to collect more real-world data to reduce input uncertainty. Freimer and Schruben (2002) considered uncertainty in the estimated parameters of parametric input distributions (e.g., exponential with parameter λ, gamma with parameters α and β). Similar to the present article, they used bootstrapping of real-world input data to mimic the effect of different possible real-world samples and the corresponding input-parameter estimates they would yield. The basic premise of Freimer and Schruben (2002) was that sufficient real-world data on the parameters have been collected when their sampling distributions, as represented by bootstrap values, have no statistically detectable effect on the simulation output.

After what we call the nominal experiment, Ng and Chick (2001, 2006) attempted to optimally allocate a finite amount of additional effort—additional real-world input-data collection and additional replications of the simulation—to reduce overall uncertainty about the simulated system performance. Similar to our approach, they employed a regression model to relate the inputs to the outputs. Their goal was to collect additional real-world input observations and additional simulation replications to minimize the posterior variance of the simulation point estimator subject to a budget constraint.

Freimer and Schruben (2002) collected additional input data until they could establish that input uncertainty was negligible, whereas Ng and Chick (2001, 2006) optimally allocated the data-collection and simulation-replication budget to minimize overall uncertainty. Both assumed that it was possible to collect input data from any of the input distributions in whatever quantity was desired or affordable. We take the perspective that additional real-world data are often unattainable, at least for some of the input processes, and the quantity that can be obtained is more likely to be constrained by time than cost per observation; therefore, if we can get more data, we will get as much as possible. The insight we deliver starts with an overall assessment of input uncertainty, which is useful for understanding risk even if there is no follow-up experiment. When additional input data are to be collected, then our relative contributions and sensitivities provide guidance about the best targets.

This article is organized as follows. Section 2 defines the input uncertainty problem and sets up our model of it. In Section 3 we describe the sequence of experiments—nominal, diagnostic, and follow-up—focusing on the diagnostic experiment for assessing input uncertainty and the contribution of each input distribution. When and how to combine output data from these experiments is addressed in Section 4. Section 5 provides guidelines for the design of the diagnostic experiment. Section 6 summarizes results from an empirical study and an illustrative example, followed by conclusions in Section 7.

2. Problem formulation

In this section we present a definition of "input uncertainty" and introduce our model of it.

2.1. *Definition of input uncertainty*

In this article, we consider mutually independent input processes that are each independent and identically distributed (i.i.d.) random variables whose marginal distributions are unknown. We use estimated or "fitted" distributions based on real-world data as stand-ins for the unknown, true distributions. We do not consider multivariate or time-dependent input processes here.

Suppose that there are L mutually independent input processes characterized by a collection of true real-world marginal distributions $\mathbf{F}^c = \{F_1^c, F_2^c, \ldots, F_L^c\}$, where c denotes "correct." Since these distributions are unknown, we use a corresponding collection of estimated distributions $\widehat{\mathbf{F}} = \{\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_L\}$ to drive the simulation. The ℓth estimated marginal distribution, $\widehat{F}_\ell$, can be either a parametric or an empirical distribution, but in either case it is inferred from observed real-world data $X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell m_\ell} \sim$ i.i.d. F_ℓ^c, where m_ℓ indicates the number of observations for the ℓth input model. We only consider $m_\ell > 0$ and thus do not address subjectively specified distributions for which we have no data.

Given a collection of input distributions $\widehat{\mathbf{F}}$, the simulation generates performance output $Y_j(\widehat{\mathbf{F}})$ on i.i.d. replication $j = 1, 2, \ldots, n$. Ankenman and Nelson (2012) represented $Y_j(\widehat{\mathbf{F}})$ as

$$Y_j(\widehat{\mathbf{F}}) = \mathrm{E}[\,Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}] + \varepsilon_j, \tag{1}$$

where $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n$ are i.i.d. with mean zero and variance σ^2 representing the natural stochastic variability from replication to replication in the simulation. They assumed that ε_j does not depend on the input model $\widehat{\mathbf{F}}$, which is clearly an approximation and one which we also adopt. It is important to notice that $\mathrm{E}[\,Y(\widehat{\mathbf{F}})|\,\widehat{\mathbf{F}}]$ is a random variable since it is a functional of $\widehat{\mathbf{F}}$, which is estimated from real-world data. In other words, depending on the real-world observations that are used to infer $\mathbf{F}^c$, we could have different values of $\mathrm{E}[\,Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$.

The goal of our simulation is to estimate $\mathrm{E}[\,Y(\mathbf{F}^c)]$, which we typically do with the sample mean $\bar{Y}(\widehat{\mathbf{F}}) = \sum_{j=1}^{n} Y_j(\widehat{\mathbf{F}})/n$ for given $\widehat{\mathbf{F}}$. This article focuses on how to assess the relative impact of each $\widehat{F}_\ell$ on the variability of the simulation estimator $\bar{Y}(\widehat{\mathbf{F}})$. As the number of simulation replications n increases, $\bar{Y}(\widehat{\mathbf{F}})$ converges to $\mathrm{E}[\,Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$, which is not necessarily the same as $\mathrm{E}[\,Y(\mathbf{F}^c)]$. In fact, there is typically a bias in the estimator $\bar{Y}(\widehat{\mathbf{F}})$ coming from the fact that $\widehat{\mathbf{F}}$ is inferred from a finite number of observations and the simulation is a nonlinear transformation. The traditional confidence interval captures only $\mathrm{Var}[\,\bar{Y}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}] = \sigma^2/n$ for given $\widehat{\mathbf{F}}$. Therefore, we need a different approach to account for the variability of the simulation estimator depending on the input models, which we refer to as the *input uncertainty*.

Formally, the input uncertainty σ_I^2 of the simulation estimator $\bar{Y}(\widehat{\mathbf{F}})$ is defined as

$$\sigma_I^2 = \mathrm{Var}\,[\mathrm{E}[\,\bar{Y}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]], \tag{2}$$

where the Var[·] is with respect to the sampling distribution of $\widehat{\mathbf{F}}$. Therefore, $\mathrm{Var}[\,\bar{Y}(\widehat{\mathbf{F}})]$ can be decomposed as

$$\begin{aligned}\mathrm{Var}[\,\bar{Y}(\widehat{\mathbf{F}})] &= \mathrm{Var}\,[\mathrm{E}[\,\bar{Y}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]] + \mathrm{E}[\mathrm{Var}[\,\bar{Y}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]] \\ &= \sigma_I^2 + \sigma^2/n,\end{aligned} \tag{3}$$

where the first expression is general, and the second follows from the definition of σ_I^2 and the homoscedasticity assumption in Model (1).

Ankenman and Nelson (2012) introduced the ratio $\gamma = \sigma_I/(\sigma/\sqrt{n})$ as a measure of the relative significance of input uncertainty to the simulation-based estimator variability. Suppose that the analyst chooses n large enough so that the estimator variance σ^2/n is reasonably small. Then a very small γ implies that $\sigma_I^2 \ll \sigma^2/n$; i.e., the input uncertainty is insignificant. On the other hand, if γ is large, then $\sigma_I^2 \gg \sigma^2/n$; i.e., there is significant input uncertainty in the simulation estimator. In the latter case, a natural question is: *Which models contribute the most to the input uncertainty?*

Our proposed definition of the *contribution* of $\widehat{F}_\ell$ to input uncertainty is

$$\begin{aligned}\mathsf{V}_\ell(m_\ell) \equiv \mathrm{Var}\big[\mathrm{E}\big[\,Y\big(F_1^c, F_2^c, \ldots, F_{\ell-1}^c, \widehat{F}_\ell, F_{\ell+1}^c,\\ \ldots, F_L^c\big)\big|\widehat{F}_\ell\big]\big].\end{aligned} \tag{4}$$

In other words, $\mathsf{V}_\ell(m_\ell)$ is the variability in the simulation's expected value when all of the true input distributions except F_ℓ^c are known and F_ℓ^c is estimated by $\widehat{F}_\ell$. Notice that V_ℓ is a function of the sample size m_ℓ to make it clear that the contribution depends on the number of observations; the larger m_ℓ is, the smaller $\mathsf{V}_\ell(m_\ell)$ becomes as $\widehat{F}_\ell$ approaches F_ℓ^c. The *relative contribution* of the ℓth input model is

$$\frac{\mathsf{V}_\ell(m_\ell)}{\sum_{i=1}^{L} \mathsf{V}_i(m_i)}.$$

We also define the *(sample size) sensitivity* of the variance of $\bar{Y}(\widehat{\mathbf{F}})$ with respect to ℓth input model by approximating m_ℓ as real valued and taking

$$\left.\frac{\partial \mathrm{Var}[\,\bar{Y}(\widehat{\mathbf{F}})]}{\partial m'_\ell}\right|_{m'_\ell = m_\ell}, \tag{5}$$

which is the same as $\frac{\partial \sigma_I^2}{\partial m'_\ell}\big|_{m'_\ell = m_\ell}$ if we assume homoscedastic simulation variance. The sensitivity can be interpreted as a measure of how much input uncertainty can be reduced by observing one more real-world sample from the ℓth input process given that we already have m_ℓ observations.

The input uncertainty problem is related to global sensitivity analysis. Here we briefly describe similarities and differences. Suppose we have a response $y = g(\mathbf{x})$ that is a function of some parameters $\mathbf{x} = (x_1, x_2, \ldots, x_L)$. The response y might be the objective-function value in an optimization problem or the key output from a deterministic numerical simulation, for instance. However, the parameters $\mathbf{x}$ are not actually known with certainty and therefore could be modeled as a random vector $\mathbf{X}$ with *known* distribution $\mathbf{F}^c$. Loosely speaking, the goal of global sensitivity analysis is to assign to each parameter $X_1, X_2, \ldots, X_L$ a measure of impact on the random variable $Y = g(\mathbf{X})$; the focus is often on decomposing $\mathrm{Var}[\,Y]$. Many of these measures are computationally expensive to compute.

For instance, Wagner (1995) defined two global sensitivity measures for the objective function of an optimization problem with uncertain parameters. One measure of sensitivity for the ℓth parameter was based on the variance of the conditional expectation of $g(\mathbf{X})$ given all parameters except X_ℓ, whereas the other was the variance of the conditional expectation of $g(\mathbf{X})$ given only X_ℓ. Homma and Saltelli (1996) proposed a related variance-based sensitivity measure: the ratio of the variance of the total effect (main effect and all the interaction effects of the parameter of interest) to the variance of Y.

Oakley and O'Hagan (2004) introduced the idea of replacing evaluations of $g(\mathbf{x})$ with evaluations of a Gaussian

process metamodel $\widehat{g}(\mathbf{x})$ that is fit to a chosen set of parameters and outputs $(\mathbf{x}_i, g(\mathbf{x}_i))$, $i = 1, 2, \ldots, k$. Because generation of $\mathbf{X} \sim \mathbf{F}^c$ and evaluation of $\widehat{g}(\mathbf{X})$ are fast, this facilitates computing any of the global sensitivity measures described above as well as others; furthermore, the Gaussian process metamodel supports incorporating uncertainty about the true function $g(\cdot)$ into the analysis. Marrel *et al.* (2012) extended this idea to stochastic simulations in which we can only observe g with noise: $g(\mathbf{x}) + \epsilon(\mathbf{x})$. They used joint metamodels for $g(\cdot)$ and $\mathrm{Var}[\epsilon(\cdot)]$ to estimate a variance-based sensitivity measure using a functional analysis of variance (ANOVA) decomposition of g. Their setting is the closest to ours in that they do sensitivity analysis in the presence of stochastic simulation output.

Returning to deterministic outputs, Plischke *et al.* (2013) proposed density-based sensitivity measures as an alternative to variance-based measures. They evaluated the expected difference between the unconditional probability density of Y and its conditional density given $X_\ell = x_\ell$. The larger the expected difference is, the more sensitive the output is to this parameter.

From our perspective, global sensitivity tries to assess the effect of each distribution $F_1^c, F_2^c, \ldots, F_L^c$ on $Y(\mathbf{F}^c)$; that is, it decomposes the (simulation) variability represented by ε. Input uncertainty arises when $\mathbf{F}^c$ is estimated by $\widehat{\mathbf{F}}$, and an assessment tries to measure the impact of variability in $\widehat{\mathbf{F}}$ (and in this paper, each $\widehat{F}_\ell$) in the presence of simulation variability. Input uncertainty due to $\widehat{F}_\ell$ can depend on how sensitive the output is to the ℓth distribution but also on how well that distribution is estimated.

2.2. *The mean-variance effects model*

As noted earlier, $\mathrm{E}[Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$ is a random variable depending on $\widehat{\mathbf{F}}$; therefore, we can think of it as a functional of $\widehat{\mathbf{F}}$; i.e., $\mathrm{g}(\widehat{\mathbf{F}}) = \mathrm{E}[Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$. We suggest (and justify below) the following *mean-variance effects model* for $\mathrm{g}(\widehat{\mathbf{F}})$:

$$\mathrm{g}(\widehat{\mathbf{F}}) = \beta_0 + \sum_{\ell=1}^{L} \beta_\ell \mu(\widehat{F}_\ell) + \sum_{\ell=1}^{L} \nu_\ell \sigma^2(\widehat{F}_\ell), \tag{6}$$

where $\mu(\widehat{F}_\ell)$ and $\sigma^2(\widehat{F}_\ell)$ represent the mean and the variance of a random variable with distribution $\widehat{F}_\ell$, respectively, and β_ℓ and ν_ℓ are constant coefficients. The philosophy behind this model is that sensitivity of the mean simulation output to the particular realization of $\widehat{F}_\ell$ is largely captured by the realized center (mean) and spread (variance) of the distribution. This model could be extended to include higher moments such as skewness and kurtosis, or the 25th and the 75th percentiles could be chosen instead of mean and variance. However, the essence of the model is to represent the relationship between each $\widehat{F}_\ell$ and $\mathrm{E}[Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$ through some characteristic properties of $\widehat{F}_\ell$. As we show later, there are advantages to using the mean and variance when we want to estimate the contribution of each input model. This model does not include interaction terms; we expect that in many cases the main effects are more significant than the interaction effects and can capture a large part of the impact of $\widehat{F}_\ell$ on the output.

The contribution of $\widehat{F}_\ell$ can be derived by plugging this model into the definition in Equation (4). Let $\mathbf{F}_{\widehat{\ell}}^c = \{F_1^c, F_2^c, \ldots, F_{\ell-1}^c, \widehat{F}_\ell, F_{\ell+1}^c, \ldots, F_L^c\}$; i.e., the set of all true distributions except F_ℓ^c. Then $\mathsf{V}_\ell(m_\ell)$ becomes

$$\begin{aligned}
\mathsf{V}_\ell(m_\ell) &= \mathrm{Var}\left[\mathrm{E}[Y(\mathbf{F}_{\widehat{\ell}}^c)|\widehat{F}_\ell]\right] \\
&= \mathrm{Var}\Bigg[\beta_0 + \sum_{i=1, i\neq\ell}^{L} \beta_i \mu(F_i^c) + \sum_{i=1, i\neq\ell}^{L} \nu_i \sigma^2(F_i^c) \\
&\quad + \beta_\ell \mu(\widehat{F}_\ell) + \nu_\ell \sigma^2(\widehat{F}_\ell)\Bigg] \\
&= \mathrm{Var}[\beta_\ell \mu(\widehat{F}_\ell) + \nu_\ell \sigma^2(\widehat{F}_\ell)] \\
&= \beta_\ell^2 \mathrm{Var}[\mu(\widehat{F}_\ell)] + \nu_\ell^2 \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] \\
&\quad + 2\beta_\ell \nu_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell), \sigma^2(\widehat{F}_\ell)].
\end{aligned} \tag{7}$$

The third equality holds because $\mu(F_i^c)$ and $\sigma^2(F_i^c)$ are constants. The overall input uncertainty σ_I^2 can be derived by plugging Model (6) into the definition in Equation (2):

$$\begin{aligned}
\sigma_I^2 = \mathrm{Var}[g(\widehat{\mathbf{F}})] &= \sum_{\ell=1}^{L} \beta_\ell^2 \mathrm{Var}[\mu(\widehat{F}_\ell)] + \sum_{\ell=1}^{L} \nu_\ell^2 \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] \\
&\quad + \sum_{\ell=1}^{L} 2\beta_\ell \nu_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell), \sigma^2(\widehat{F}_\ell)] \\
&= \sum_{\ell=1}^{L} \left\{\beta_\ell^2 \mathrm{Var}[\mu(\widehat{F}_\ell)] + \nu_\ell^2 \mathrm{Var}[\sigma^2(\widehat{F}_\ell)]\right. \\
&\quad \left. + 2\beta_\ell \nu_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell), \sigma^2(\widehat{F}_\ell)]\right\} \\
&= \sum_{\ell=1}^{L} \mathsf{V}_\ell(m_\ell).
\end{aligned} \tag{8}$$

The second equality follows from independent sampling from the input models and the last equality follows directly from Equation (7). This result shows that under Model (6) the overall input uncertainty is the summation of individual contributions; i.e., $\sigma_I^2 = \sum_{\ell=1}^{L} \mathsf{V}_\ell(m_\ell)$. Thus, under our model the overall input uncertainty can be decomposed into the individual contributions, and the individual contributions are independent of each other. Also, under this model the sensitivity of $\widehat{F}_\ell$ becomes

$$\left.\frac{\partial \sigma_I^2}{\partial m'_\ell}\right|_{m'_\ell = m_\ell} = \left.\frac{\partial \mathsf{V}_\ell(m'_\ell)}{\partial m'_\ell}\right|_{m'_\ell = m_\ell},$$

which makes the sensitivity simply the derivative of the contribution with respect to the sample size, evaluated at the current number of samples m_ℓ.

The variance decomposition in Equation (8) coincides with a result in Cheng and Holland (1998) that was obtained by a different argument. They approximated the input uncertainty variance σ_I^2 as a function of the variances

of parameter estimators of parametric input distributions. If we let $\mathbf{F}^c(\cdot|\boldsymbol{\theta})$ be the true parametric family of input distributions, $\boldsymbol{\theta}^c$ be the collection of true parameters, and $\widehat{\boldsymbol{\theta}}$ be an estimator of $\boldsymbol{\theta}^c$, then a Taylor series expansion of $g(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}^c$ gives

$$\text{Var}[g(\widehat{\boldsymbol{\theta}})] \approx \nabla(g(\boldsymbol{\theta}^c))^T \text{Var}[\widehat{\boldsymbol{\theta}}] \nabla(g(\boldsymbol{\theta}^c)), \tag{9}$$

where $\text{Var}[\widehat{\boldsymbol{\theta}}]$ is the variance–covariance matrix of $\widehat{\boldsymbol{\theta}}$ and $\nabla(g(\boldsymbol{\theta}^c))$ is the gradient of $g(\cdot)$ at $\boldsymbol{\theta}^c$. By further assuming $\widehat{\boldsymbol{\theta}}$ is a Maximum Likelihood Estimator (MLE) for $\boldsymbol{\theta}^c$, they argued that

$$\text{Var}[g(\widehat{\boldsymbol{\theta}})] \approx \nabla(g(\boldsymbol{\theta}^c))^T \mathbf{V}[\vec{\boldsymbol{\theta}}] \nabla(g(\boldsymbol{\theta}^c)),$$

where $\mathbf{V}[\vec{\boldsymbol{\theta}}]$ is the asymptotic variance–covariance matrix of $\widehat{\boldsymbol{\theta}}$ under some regularity conditions.

To connect this to our formulation, assume that each input distribution F_ℓ^c can be parameterized by $\mu(F_\ell^c)$ and $\sigma^2(F_\ell^c)$. Then we can represent $g(\widehat{\mathbf{F}})$ as a function of the parameters $g(\widehat{\boldsymbol{\theta}})$, where $\widehat{\boldsymbol{\theta}} = \{\mu(\widehat{F}_1),\ \sigma^2(\widehat{F}_1),\ \mu(\widehat{F}_2),\ \sigma^2(\widehat{F}_2),\ \ldots,\ \mu(\widehat{F}_L),\ \sigma^2(\widehat{F}_L)\}$. Similarly, $\boldsymbol{\theta}^c = \{\mu(F_1^c), \sigma^2(F_1^c), \mu(F_2^c), \sigma^2(F_2^c), \ldots, \mu(F_L^c), \sigma^2(F_L^c)\}$ and the gradient $\nabla(g(\boldsymbol{\theta}^c))$ under Model (6) is $(\beta_1, \nu_1, \beta_2, \nu_2, \ldots, \beta_L, \nu_L)^T$. In fact, in our case the approximation in Equation (9) is an equality because Model (6) is linear in $\widehat{\boldsymbol{\theta}}$ and therefore the first-order Taylor approximation is the model itself. A feature of our approach is that we can directly compute $\text{Var}[\widehat{\boldsymbol{\theta}}]$ without making further assumptions. Since we have independence among input processes, $\text{Var}[\widehat{\boldsymbol{\theta}}]$ is a block diagonal matrix:

$$\text{Var}[\widehat{\boldsymbol{\theta}}] = \begin{bmatrix} \text{Var}[\mu(\widehat{F}_1)] & \text{Cov}[\mu(\widehat{F}_1), \sigma^2(\widehat{F}_1)] & 0 & \cdots & 0 \\ \text{Cov}[\mu(\widehat{F}_1), \sigma^2(\widehat{F}_1)] & \text{Var}[\sigma^2(\widehat{F}_1)] & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & 0 & \text{Var}[\mu(\widehat{F}_L)] & \text{Cov}[\mu(\widehat{F}_L), \sigma^2(\widehat{F}_L)] \\ 0 & \cdots & 0 & \text{Cov}[\mu(\widehat{F}_L), \sigma^2(\widehat{F}_L)] & \text{Var}[\sigma^2(\widehat{F}_L)] \end{bmatrix}.$$

Plugging $\text{Var}[\widehat{\boldsymbol{\theta}}]$ into Equation (9) gives the same expression as in Equation (8).

Cheng and Holland (1998) provided an approximate analysis of an exact model that requires parametric input distributions and MLEs; we provide an exact analysis of an approximate model using any form of input distribution but assuming that most of the sensitivity of the simulation response to the input distributions is captured by their means and variances.

To evaluate the contribution of each input model, we will use least-squares regression to estimate $\beta_0, \beta_1, \ldots, \beta_L, \nu_1, \nu_2, \ldots, \nu_L$. In the next section we introduce a sequence of experiments to estimate these coefficients and to evaluate the contribution and sensitivity of each input model.

3. The sequence of experiments

In this section we describe a sequence of experiments that an analyst might conduct: *nominal*, *diagnostic*, and *follow-up*. Current simulation practice is to run only the nominal experiment.

The *nominal* experiment involves the analyst collecting input data, choosing input models $\widehat{\mathbf{F}}$ to use, building the simulation model, and running n replications to obtain the point estimator $\bar{Y}(\widehat{\mathbf{F}})$ of system performance $\text{E}[Y(\mathbf{F}^c)]$. From this experiment, the analyst obtains the traditional CI for $\text{E}[Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$, which is typically not the same as a CI for $\text{E}[Y(\mathbf{F}^c)]$, as discussed earlier. The number of replications n is either chosen arbitrarily, to achieve a certain level of simulation error, or because it can be completed in the available time.

We are suggesting that this be followed by a *diagnostic* experiment to evaluate the contribution and sensitivity of each input model as well as the overall input uncertainty. The contributions can be calculated from Equation (7) given the coefficients $\beta_1, \beta_2, \ldots, \beta_L, \nu_1, \nu_2, \ldots, \nu_L$ and the variance and covariance of $\mu(\widehat{F}_\ell)$ and $\sigma^2(\widehat{F}_\ell)$. We describe a method for estimating them in the next section.

Upon completion of the diagnostic experiment, the analyst is either satisfied that input uncertainty is not substantial or is concerned that it is substantial and has a better understanding of how significant it is. In either case, the analyst has simulation results from the diagnostic experiment that could perhaps be used to improve the estimate of $\text{E}[Y(\mathbf{F}^c)]$; we study whether and how to do this as well.

When input uncertainty is substantial, the analyst may also undertake a *follow-up* experiment, which involves collecting additional real-world input data (with our sensitivities providing the most valuable targets), refining the estimated input models, and conducting another simulation with the refined input models. Conclusions could be based on this final experiment only, but we investigate whether simulation outputs from the nominal and diagnostic experiments should also be incorporated. Of course, there could be additional cycles of diagnostic and follow-up experiments as desired.

3.1. *Diagnostic experiment*

The diagnostic experiment is conducted to fit the mean-variance effects model (6) and derive our measures of input uncertainty. To estimate the unknown coefficients $\beta_0, \beta_1, \ldots, \beta_L, \nu_1, \nu_2, \ldots, \nu_L$ by least squares, we need at least $2L + 2$ design points, which means we need $2L + 2$ different $\widehat{\mathbf{F}}$s. However, this is typically impossible since we do not have more than one data set to fit $\widehat{\mathbf{F}}$; even if we did have multiple data sets they would usually be pooled to enhance the precision of $\widehat{\mathbf{F}}$. Instead, Song and Nelson (2013) suggested a bootstrap approach by treating $\widehat{\mathbf{F}}$ as the true real-world distribution $\mathbf{F}^c$ and sampling multiple times from $\widehat{\mathbf{F}}$ instead of gathering multiple real-world samples. In our context, a "bootstrap" is an i.i.d. sample $X_{\ell 1}^\star, X_{\ell 2}^\star, \ldots, X_{\ell m_\ell}^\star$ from $\widehat{F}_\ell$. We use the notation $X_{\ell j}^\star$ to denote a sample from the empirical cumulative distribution

(ecdf) or fitted parametric distribution $\widehat{F}_\ell$, as opposed to $X_{\ell j}$, which is a sample from the true real-world distribution F_ℓ^c. More generally, a $\star$ denotes a quantity defined by a bootstrap sample.

In our method a bootstrap sample $X^\star_{\ell 1}, X^\star_{\ell 2}, \ldots, X^\star_{\ell m_\ell}$ from $\widehat{F}_\ell$ is treated as a real-world sample $X_{\ell 1}, X_{\ell 2}, \ldots, X_{\ell m_\ell}$ from F_ℓ^c. From the bootstrap sample we can fit an ecdf $\widehat{F}^\star_\ell$, which plays the role of $\widehat{F}_\ell$, and calculate $\mu(\widehat{F}^\star_\ell)$ and $\sigma^2(\widehat{F}^\star_\ell)$. Repeating this for $\ell = 1, 2, \ldots, L$, a collection of ecdfs $\widehat{\mathbf{F}}^\star = \{\widehat{F}^\star_1, \widehat{F}^\star_2, \ldots, \widehat{F}^\star_L\}$ is obtained and we can run replications of the simulation with $\widehat{\mathbf{F}}^\star$ to obtain the estimator $\bar{Y}(\widehat{\mathbf{F}}^\star)$. If we repeat this process B times, we will get $\widehat{\mathbf{F}}^{\star(1)}, \widehat{\mathbf{F}}^{\star(2)}, \ldots, \widehat{\mathbf{F}}^{\star(B)}$ and the corresponding means and variances of input models, as well as the simulation estimators $\bar{Y}(\widehat{\mathbf{F}}^{\star(1)}), \bar{Y}(\widehat{\mathbf{F}}^{\star(2)}), \ldots, \bar{Y}(\widehat{\mathbf{F}}^{\star(B)})$ by which we can fit Model (6).

Why use bootstrap samples to create design points for fitting the mean-variance effects model (6) instead of a classic designed experiment? First, this is not a simple design space: it is the space of possible fitted input distributions that could result from sampling from the true distribution $\mathbf{F}^c$. Even if parametric distributions are used (which we do not require) so that the design space becomes the space of parameter values, some sets of parameters are far more likely than others, and our mean-variance model will be most effective if we get a good fit near $\mathbf{F}^c$ rather than a global fit across the space. By using bootstrap samples from $\widehat{\mathbf{F}}$ we create design points that are representative of what is likely, providing a good fit where it matters most. Furthermore, simulating at bootstrap random samples, rather than chosen design points, allows us to combine simulation results from the nominal and diagnostic experiments without introducing lack-of-fit bias; see Section 4. There are additional advantages, which we describe below.

The analogy between real-world sampling and bootstrap sampling is equivalent to assuming that the input uncertainty $\sigma_I^2 = \mathrm{Var}[g(\widehat{\mathbf{F}})]$ is approximated as

$$\mathrm{Var}[g(\widehat{\mathbf{F}})] = \mathrm{Var}[g(\widehat{\mathbf{F}})|\mathbf{F}^c] \approx \mathrm{Var}[g(\widehat{\mathbf{F}}^\star)|\widehat{\mathbf{F}}]. \tag{10}$$

Under Model (6), σ_I^2 is the sum of the contributions of all input models as in Equation (8). Therefore, in view of the approximation (10):

$$\begin{aligned}&\beta_\ell^2 \mathrm{Var}[\mu(\widehat{F}_\ell)] + \nu_\ell^2 \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] + 2\beta_\ell\nu_\ell \\ &\quad\times \mathrm{Cov}[\mu(\widehat{F}_\ell), \sigma^2(\widehat{F}_\ell)] \approx \beta_\ell^2 \mathrm{Var}[\mu(\widehat{F}^\star_\ell)|\widehat{F}_\ell] \\ &\quad+\nu_\ell^2 \mathrm{Var}[\sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell] + 2\beta_\ell\nu_\ell \mathrm{Cov}[\mu(\widehat{F}^\star_\ell), \sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell],\end{aligned}$$

which will be true if

$$\begin{aligned}\mathrm{Var}[\mu(\widehat{F}_\ell)] &\approx \mathrm{Var}[\mu(\widehat{F}^\star_\ell)|\widehat{F}_\ell] \\ \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] &\approx \mathrm{Var}[\sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell] \\ \mathrm{Cov}[\mu(\widehat{F}_\ell), \sigma^2(\widehat{F}_\ell)] &\approx \mathrm{Cov}[\mu(\widehat{F}^\star_\ell), \sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell].\end{aligned} \tag{11}$$

As the real-world sample size m_ℓ increases, this approximation is asymptotically justified under some conditions on $\widehat{F}_\ell$, as discussed in Section 3.3. A valuable advantage of the approximation (11) is that it provides expressions for the variance and covariance components in Equation (7) that we need to calculate the contributions. Since we use an empirical distribution $\widehat{F}^\star_\ell$, $\mu(\widehat{F}^\star_\ell)$ and $\sigma^2(\widehat{F}^\star_\ell)$ are simply the sample mean and the second sample central moment of $X^\star_{\ell 1}, X^\star_{\ell 2}, \ldots, X^\star_{\ell m_\ell}$, which is an i.i.d. sample from the known distribution $\widehat{F}_\ell$. Therefore, as shown in Appendix A of the online supplement, we can derive expressions for $\mathrm{Var}[\mu(\widehat{F}^\star_\ell)|\widehat{F}_\ell]$, $\mathrm{Var}[\sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell]$, and $\mathrm{Cov}[\mu(\widehat{F}^\star_\ell), \sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell]$ as

$$\begin{aligned}\mathrm{Var}[\mu(\widehat{F}^\star_\ell)|\widehat{F}_\ell] &= \frac{M_\ell^2}{m_\ell} \\ \mathrm{Var}[\sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell] &= \frac{(m_\ell-1)^2}{m_\ell^3} M_\ell^4 \\ &\quad - \frac{(m_\ell-3)(m_\ell-1)}{m_\ell^3} (M_\ell^2)^2 \\ &\approx \frac{M_\ell^4 - (M_\ell^2)^2}{m_\ell} \\ \mathrm{Cov}[\mu(\widehat{F}^\star_\ell), \sigma^2(\widehat{F}^\star_\ell)|\widehat{F}_\ell] &= \frac{(m_\ell-1)^2}{{m_\ell}^3} M_\ell^3 \approx \frac{M_\ell^3}{m_\ell},\end{aligned}$$

where M_ℓ^k is kth central moment of $\widehat{F}_\ell$. If $\widehat{F}_\ell$ is an empirical distribution, then $M_\ell^k = \sum_{i=1}^{m_\ell}(X_{\ell i} - \bar{X}_\ell)^k/m_\ell$. If $\widehat{F}_\ell$ is a parametric distribution, then we can calculate the central moments from the known representation; for instance, if $\widehat{F}_\ell$ is a gamma distribution with estimated shape parameter $\widehat{\alpha}$ and rate parameter $\widehat{\beta}$, then the second, third, and fourth moments are $\widehat{\alpha}/\widehat{\beta}^2$, $2\widehat{\alpha}/\widehat{\beta}^3$, and $3\widehat{\alpha}^2/\widehat{\beta}^4 + 6\widehat{\alpha}/\widehat{\beta}^4$, respectively.

One of the major advantages of our approach is that these variance/covariance expressions are not approximations, which improves the estimation of the contributions. Also notice that this is a very general method that is applicable to any empirical distribution and many parametric distributions; the only time we face difficulty is when $\widehat{F}_\ell$ is a parametric distribution whose parameters are in the range for which not all moments up to the fourth exist (e.g., a log-logistic distribution with shape parameter less than four). Even then, we can use our method provided that the offending distribution can be represented as a transformation of another distribution whose first four moments always exist. In this case we fit the mean-variance model to the moments of the underlying distribution. For instance, the log-logistic distribution is a transformation of a logistic distribution whose first four moments are finite, so we fit to the moments of the underlying logistic distribution. Since every distribution can be viewed as a transformation of the uniform (0, 1) distribution, our method is (at least in theory) completely general.

Inserting the moment expressions into Equation (7), the contribution of $\widehat{F}_\ell$ with m_ℓ samples can be written as

$$\mathsf{V}_\ell(m_\ell) \approx \frac{1}{m_\ell}\left\{\beta_\ell^2 M_\ell^2 + \nu_\ell^2\left(M_\ell^4 - (M_\ell^2)^2\right) + 2\beta_\ell\nu_\ell M_\ell^3\right\}, \tag{12}$$

implying that the sample size sensitivity of $\widehat{F}_\ell$ is

$$\left.\frac{\partial \mathsf{V}_\ell(m'_\ell)}{\partial m'_\ell}\right|_{m'_\ell=m_\ell} = -\frac{1}{(m_\ell)^2}\left\{\beta_\ell^2 M_\ell^2 + \nu_\ell^2\left(M_\ell^4 - \left(M_\ell^2\right)^2\right) + 2\beta_\ell\nu_\ell M_\ell^3\right\} = -\frac{\mathsf{V}_\ell(m_\ell)}{m_\ell}. \tag{13}$$

Notice that the sensitivity is always negative since input uncertainty is reduced with additional real-world data. Notice also that the rank order of contributions and sensitivities of distributions do not always coincide; even if $\widehat{F}_\ell$ has the largest contribution, if m_ℓ is also large then input uncertainty may be less sensitive to $\widehat{F}_\ell$ than some other input models. In other words, an additional sample from F_ℓ^c may not make much difference to the input uncertainty variance since we already have a large sample.

The algorithm for the diagnostic experiment is as follows:

Diagnostic Experiment

1. Given the estimated input models $\widehat{\mathbf{F}}$, do the following:
2. For bootstrap sample $b = 1$ to B
   a. For input model $\ell = 1$ to L
      i. Generate $X_{\ell 1}^{\star(b)}, X_{\ell 2}^{\star(b)}, \ldots, X_{\ell m_\ell}^{\star(b)}$ by sampling m_ℓ times from $\widehat{F}_\ell$.
      ii. Let $\widehat{F}_\ell^{\star(b)}$ be the ecdf of $X_{\ell 1}^{\star(b)}, X_{\ell 2}^{\star(b)}, \ldots, X_{\ell m_\ell}^{\star(b)}$ and calculate the mean $\mu(\widehat{F}_\ell^{\star(b)})$ and variance $\sigma^2(\widehat{F}_\ell^{\star(b)})$.
   b. Using input models $\widehat{\mathbf{F}}^{\star(b)} = \{\widehat{F}_1^{\star(b)}, \widehat{F}_2^{\star(b)}, \ldots, \widehat{F}_L^{\star(b)}\}$, simulate R i.i.d. replications $Y_j(\widehat{\mathbf{F}}^{\star(b)})$, $j = 1, 2, \ldots, R$ and calculate the sample mean $\bar{Y}(\widehat{\mathbf{F}}^{\star(b)})$.
3. Fit the model

$$\bar{Y}(\widehat{\mathbf{F}}^{\star(b)}) = \beta_0 + \sum_{\ell=1}^{L}\beta_\ell\mu(\widehat{F}_\ell^{\star(b)}) + \sum_{\ell=1}^{L}\nu_\ell\sigma^2(\widehat{F}_\ell^{\star(b)}) + \bar{\varepsilon}_b \tag{14}$$

   with $\bar{Y}(\widehat{\mathbf{F}}^{\star(b)})$ from step 2b and $\mu(\widehat{F}_1^{\star(b)}), \mu(\widehat{F}_2^{\star(b)}), \ldots, \mu(\widehat{F}_L^{\star(b)})$ and $\sigma^2(\widehat{F}_1^{\star(b)}), \sigma^2(\widehat{F}_2^{\star(b)}), \ldots, \sigma^2(\widehat{F}_L^{\star(b)})$ from step 2a for $b = 1, 2, \ldots, B$ to estimate the coefficients $\beta_0, \beta_1, \ldots, \beta_L, \nu_1, \nu_2, \ldots, \nu_L$.
4. Estimate the overall input uncertainty $\widehat{\sigma}_I^2 = \sum_{\ell=1}^{L}\widehat{\mathsf{V}}_\ell(m_\ell)$ and the ratio

$$\widehat{\gamma} = \frac{\widehat{\sigma}_I}{\widehat{\sigma}/\sqrt{n}} = \frac{\sqrt{n\sum_{\ell=1}^{L}\widehat{\mathsf{V}}_\ell(m_\ell)}}{\widehat{\sigma}}.$$

5. Estimate the contribution $\widehat{\mathsf{V}}_\ell(m_\ell)$ and the sensitivity for $\ell = 1, 2, \ldots, L$.

We estimate $\widehat{\sigma}^2$ in step 4 using the sample variance from the nominal experiment, rather than using the residual mean-squared error (MSE) of the fitted model in step 3; this avoids bias due to lack of fit. Of course, n in step 4 is the number of replications used in the nominal experiment and need not be the same as R.

As discussed in Section 2.1, the ratio γ expresses input uncertainty in units of the simulation estimation error. If $\widehat{\gamma} = 0.5$, for instance, it implies that the input uncertainty is only half of the simulation estimation error, which may be acceptable depending on the type of decision that the simulation is expected to support. If $\widehat{\gamma}$ is large—e.g., $\widehat{\gamma} = 20$—then we can conclude that the simulation estimation error as measured, say, by the width of a CI, should be inflated by a factor of roughly $\sqrt{1+\widehat{\gamma}^2}$. Whether or not this is acceptable depends on the situation: If the simulation estimation error is very small—as would occur if the number of replications n is very large—then a CI that is roughly 20 times longer might have little effect on the decision that the simulation model is designed to support. We believe that it will often be the case that such an inflation is unacceptable, so the analyst may choose to collect more real-world data for the input models that have greater (more negative) sensitivities. This decision also depends on the feasibility and the cost of additional data collection.

The key insight is that $\widehat{\gamma}$ must be interpreted in light of the remaining simulation estimation error, not as an absolute number, and will be the most meaningful when n was explicitly chosen chosen to achieve a specified level of error. For instance if n was chosen (perhaps sequentially) to attain a CI whose width is no more than $\pm\Delta$, then $\widehat{\gamma}$ can be interpreted as large or small relative to $\sqrt{1+\widehat{\gamma}^2}\times\Delta$.

3.2. *Follow-up experiment*

If the analyst collects more real-world data from some or all of the input processes, then they will have an updated collection of input models with $m'_\ell > m_\ell$ for at least one $\ell \in \{1, 2, \ldots, L\}$. Using updated input models that are fit to all of the accumulated data, the analyst can run a *follow-up* experiment to obtain an estimator of $\mathrm{E}[Y(\mathbf{F}^c)]$ with reduced input uncertainty. We assume that the follow-up experiment employs at least as many simulation replications as the nominal experiment, so $n' \geq n$. If needed, this sequence of experiments can be repeated by regarding the results from the follow-up experiment in the previous sequence as a new nominal experiment.

The primary question with respect the follow-up experiment is whether we should use simulation outputs from the nominal or diagnostic experiments in the overall estimator. We address this question in a later section.

3.3. *Validity of the bootstrap approximation*

Here we consider the validity of the approximation, introduced above, where we suggested that

$$\mathrm{Var}[g(\widehat{\mathbf{F}})] = \mathrm{Var}[g(\widehat{\mathbf{F}})|\mathbf{F}^c] \approx \mathrm{Var}[g(\widehat{\mathbf{F}}^\star)|\widehat{\mathbf{F}}].$$

Given our model, this is equivalent to

$$\begin{aligned}\mathrm{Var}[\mu(\widehat{F}_\ell)] &\approx \mathrm{Var}[\mu(\widehat{F}_\ell^\star)|\widehat{F}_\ell]\\ \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] &\approx \mathrm{Var}[\sigma^2(\widehat{F}_\ell^\star)|\widehat{F}_\ell]\\ \mathrm{Cov}[\mu(\widehat{F}_\ell),\sigma^2(\widehat{F}_\ell)] &\approx \mathrm{Cov}[\mu(\widehat{F}_\ell^\star),\sigma^2(\widehat{F}_\ell^\star)|\widehat{F}_\ell].\end{aligned}$$

This assumption can be asymptotically justified under certain conditions as the sample size m_ℓ gets large; we describe some cases below. Recall that in our method the bootstrap distribution $\widehat{F}_\ell^\star$ is an ecdf. Let $\mu_\ell^k = \mathrm{E}[(X_{\ell j} - \mathrm{E}(X_{\ell j}))^k]$, the kth central moment of F_ℓ^c.

Assuming that $\widehat{F}_\ell$ is also an ecdf, and the relevant moments exist, the asymptotic variance and covariance of $\mu(\widehat{F}_\ell^\star)$ and $\sigma^2(\widehat{F}_\ell^\star)$ given $\widehat{F}_\ell$ are, with probability 1:

$$\begin{aligned}\lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\mu(\widehat{F}_\ell^\star)|\widehat{F}_\ell] &= \lim_{m_\ell\to\infty} M_\ell^2 = \mu_\ell^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\sigma^2(\widehat{F}_\ell^\star)|\widehat{F}_\ell] &= \lim_{m_\ell\to\infty} M_\ell^4 - \left(M_\ell^2\right)^2 = \mu_\ell^4 - \left(\mu_\ell^2\right)^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell^\star),\sigma^2(\widehat{F}_\ell^\star)|\widehat{F}_\ell] &= \lim_{m_\ell\to\infty} M_\ell^3 = \mu_\ell^3.\end{aligned} \tag{15}$$

Furthermore, when $\widehat{F}_\ell$ is an ecdf it is easy to show that

$$\begin{aligned}\lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\mu(\widehat{F}_\ell)] &= \mu_\ell^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] &= \mu_\ell^4 - \left(\mu_\ell^2\right)^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell),\sigma^2(\widehat{F}_\ell)] &= \mu_\ell^3.\end{aligned} \tag{16}$$

See Zhang (2007), Cho and Cho (2008), and the online supplement.

However, if $\widehat{F}_\ell$ is a parametric distribution whose parameters are estimated from the observed real-world data then neither Equation (15) nor Equation (16) is guaranteed to hold. For instance, if we fit the wrong parametric family to the data then differences can occur. A sufficient condition for both Equation (15) and (16) to hold when F_ℓ is a parametric distribution is that F_ℓ is flexible enough to match any first four moments of the data, and the distribution is fit using the Method of Moments (MMs) up to at least the fourth moment.

Even when we have the correct parametric family, there could still be differences depending on how we estimate the parameters. As mentioned above, if we use the MMs then the moments of the fitted distribution are the sample moments. And in many cases, the MLEs and MMs are asymptotically equivalent (e.g., normal). This is not always the case, however.

Suppose that F_ℓ^c is a uniform (0, 1) distribution and we have m_ℓ i.i.d. real-world observations $X_1, X_2, \ldots, X_{m_\ell}$. The MLEs are $\widehat{\alpha} = X_{(1)}$ and $\widehat{\beta} = X_{(m_\ell)}$, where $X_{(i)}$ is the ith order statistic. Then

$$\begin{aligned}\lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\mu(\widehat{F}_\ell)] &= 0 < \frac{1}{12} = \mu_\ell^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Var}[\sigma^2(\widehat{F}_\ell)] &= 0 < \frac{1}{180} = \mu_\ell^4 - (\mu_\ell^2)^2\\ \lim_{m_\ell\to\infty} m_\ell \mathrm{Cov}[\mu(\widehat{F}_\ell),\sigma^2(\widehat{F}_\ell)] &= 0 = \mu_\ell^3.\end{aligned}$$

Except for the covariance term, the asymptotic variances are smaller than those of the MMs estimators because the MLEs are asymptotically more efficient. Nevertheless, even in this case our bootstrap approximation provides a valid representation of the variability of the real-world data that could have been obtained, although not a perfect representation of the variability of the estimated parameters.

As a practical matter, we would apply our method for any sample sizes m_ℓ, $\ell = 1, 2, \ldots L$ that the analyst is comfortable using to fit distributions *if the alternative is to ignore input uncertainty*. Based on the bootstrapping literature (e.g., Hall (1992, Appendix I)), the performance of sample moment estimators, and our own experience in empirical studies, we are comfortable with $m_\ell \geq 100$.

4. Combining results from the nominal, diagnostic, and follow-up experiments

In some situations it is not feasible to gather additional real-world input data even if there is substantial input uncertainty; in others the input uncertainty is so small that there is little value in reducing it further. In either of these situations the analyst terminates the experiment at the diagnostic phase that generated $\bar{Y}(\widehat{\mathbf{F}}^{\star(b)})$, $b = 1, 2, \ldots, B$ to fit Model (6). The first question we address is whether these results can be utilized to improve the estimator $\bar{Y}(\widehat{\mathbf{F}})$ from the nominal experiment by using a weighted estimator:

$$\tilde{Y} = \alpha\, \bar{Y}(\widehat{\mathbf{F}}) + (1-\alpha)\, \bar{\bar{Y}}(\widehat{\mathbf{F}}),$$

where $\bar{\bar{Y}}(\widehat{\mathbf{F}}) = \sum_{b=1}^{B} \bar{Y}(\widehat{\mathbf{F}}^{\star(b)})/B$ and $\alpha \in [0, 1]$. This estimator only makes sense because the bootstrap distributions $\widehat{\mathbf{F}}^{\star(b)}$ are sampled directly from $\widehat{\mathbf{F}}$ and therefore indirectly from $\mathbf{F}^c$. Deterministically chosen design points would introduce an unknown and likely significant bias.

We answer this question by seeking $\alpha^\star$ that minimizes $\mathrm{MSE}[\tilde{Y}|\widehat{\mathbf{F}}]$, rather than minimizing $\mathrm{MSE}[\tilde{Y}]$. Minimizing $\mathrm{MSE}[\tilde{Y}]$ would make sense if we were actually able to obtain multiple real-world data sets, whereas minimizing $\mathrm{MSE}[\tilde{Y}|\widehat{\mathbf{F}}]$ acknowledges that we only have *one* real-world data set and therefore cannot improve our estimate of $\mathbf{F}^c$ beyond $\widehat{\mathbf{F}}$.

From the derivation in Appendix B of the online supplement, and under the assumption that Model (6) holds, we have

$$\mathrm{MSE}[\tilde{Y}|\widehat{\mathbf{F}}] = (1-\alpha)^2\left((\mathrm{b}^\star)^2 + \frac{\sigma_I^2}{B} + \frac{\sigma^2}{BR}\right) + \alpha^2\frac{\sigma^2}{n}, \tag{17}$$

where

$$\mathrm{b}^\star = \mathrm{Bias}[\bar{\bar{Y}}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}] = \mathrm{E}[\,Y(\widehat{\mathbf{F}}^{\star(b)})|\widehat{\mathbf{F}}] - \mathrm{E}[\,Y(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}]$$

which is the bias from bootstrapping. Therefore, the optimal $\alpha^\star$ is

$$\alpha^\star = \frac{(\mathrm{b}^\star)^2 + \sigma_I^2/B + \sigma^2/(BR)}{(\mathrm{b}^\star)^2 + \sigma_I^2/B + \sigma^2/(BR) + \sigma^2/n}. \tag{18}$$

Notice that $\alpha^\star$ is strictly less than one; hence, it is always better to pool $\bar{Y}(\widehat{\mathbf{F}})$ and $\bar{\bar{Y}}(\widehat{\mathbf{F}})$. The key term is σ^2/n that represents the simulation variance: the larger it is, the more weight is given to the diagnostic results, which are biased but reduce simulation variance. An unbiased estimator of $\mathrm{b}^\star$ is $\bar{\bar{Y}}(\widehat{\mathbf{F}}) - \bar{Y}(\widehat{\mathbf{F}})$, so every term in $\alpha^\star$ is either known or estimable from the nominal and diagnostic experiments.

Previously we suggested that the diagnostic experiment could be used to provide a heuristic adjustment to the CI or standard error of the mean of the nominal experiment by multiplying them by $\sqrt{1+\widehat{\gamma}^2}$. In Appendix C of the online supplement, we justify using $\sqrt{\widehat{\sigma}_I^2 + \widehat{\mathrm{MSE}}}$ as a plug-in approximation for the MSE of the combined estimator $\tilde{Y}$, where $\widehat{\mathrm{MSE}}$ is Equation (17) with optimal weight $\alpha^\star$ from Equation (18) but inserting estimates from the diagnostic experiment for all of the unknown quantities. As a rough CI we could multiply this by an appropriate normal quantile.

We next consider the case when all three experiments have been conducted. Now it makes sense to try to minimize the MSE of the final point estimator with respect to both input and simulation uncertainty. For this reason we discard the simulation results from the diagnostic experiment since they introduce additional bias due to bootstrapping.

For a clear distinction, let $\widehat{\mathbf{F}}_{\mathbf{m}}$ denote the collection of input models used in the nominal experiment that were estimated from $\mathbf{m} = \{m_1, m_2, \ldots, m_L\}$ real-world observations, and let $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ denote the estimator from n replications using $\widehat{\mathbf{F}}_{\mathbf{m}}$. Similarly, $\widehat{\mathbf{F}}_{\mathbf{m}'}$ denotes the collection of input models used in the follow-up experiment that were estimated from $\mathbf{m}' = \{m'_1, m'_2, \ldots, m'_L\}$ real-world observations, where $m'_\ell \geq m_\ell$ for all ℓ, and $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$ is the corresponding estimator from $n' \geq n$ replications. Notice that it is very likely that $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ and $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$ are positively correlated since the follow-up $\mathbf{m}'$ real-world observations include the nominal $\mathbf{m}$ observations.

The weighted estimator $\widehat{Y}$ is defined as

$$\widehat{Y} = \alpha\,\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}}) + (1-\alpha)\,\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'}).$$

Due to the correlation between $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ and $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$, finding $\alpha^\star$ to minimize $\mathrm{MSE}(\widehat{Y})$ is more complicated than the previous case. As shown in Appendix D of the online supplement, if we let $\mathrm{b}_{\mathbf{m}}$ and $\mathrm{b}_{\mathbf{m}'}$ denote the bias of $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ and $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$, respectively, as estimators of $\mathrm{E}[\,Y(\mathbf{F}^c)]$, and let σ_I^2 and $(\sigma'_I)^2$ denote the input uncertainty of the nominal and the follow-up experiments, respectively, then $\alpha^\star$ becomes

$$\alpha^\star = \frac{(\sigma'_I)^2 + (\sigma^2/n') + \mathrm{b}_{\mathbf{m}'}(\mathrm{b}_{\mathbf{m}'} - \mathrm{b}_{\mathbf{m}}) - \mathrm{Cov}[\,\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}}),\ \bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})]}{\sigma_I^2 + (\sigma'_I)^2 + (\sigma^2/n) + (\sigma^2/n') + (\mathrm{b}_{\mathbf{m}'} - \mathrm{b}_{\mathbf{m}})^2 - 2\mathrm{Cov}[\,\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}}),\ \bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})]}.$$

If $\alpha^\star < 0$ we use $\alpha^\star = 0$.

In general it is not easy to find a useful expression for $\mathrm{Cov}[\,\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}}),\ \bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})]$. However, if we assume all input distributions are ecdfs, then under Model (6) we can show that

$$\mathrm{Cov}[\,\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}}),\ \bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})] \approx (\sigma'_I)^2.$$

Under this condition we can also get an expression for the bias term as

$$\mathrm{b}_{\mathbf{m}} = \sum_{\ell=1}^{L} \frac{\nu_\ell \sigma^2(F_\ell^c)}{m_\ell - 1},$$

where $\sigma^2(F_\ell^c)$ represents the variance of the true distribution F_ℓ^c. These expressions are derived in Appendix D of the online supplement. Therefore, for this special case $\alpha^\star$ becomes

$$\alpha^\star \approx \frac{(\sigma^2/n') + \left\{\sum_{\ell=1}^{L} \nu_\ell \sigma^2(F_\ell^c)/(m'_\ell - 1)\right\}\left\{\sum_{\ell=1}^{L} \nu_\ell \sigma^2(F_\ell^c)\big((1/(m'_\ell - 1)) - (1/(m_\ell - 1))\big)\right\}}{(\sigma^2/n) + (\sigma^2/n') + \sum_{\ell=1}^{L}\big(\mathsf{V}_\ell(m_\ell) - \mathsf{V}_\ell(m'_\ell)\big) + \left\{\sum_{\ell=1}^{L} \nu_\ell \sigma^2(F_\ell^c)\big((1/(m'_\ell - 1)) - (1/(m_\ell - 1))\big)\right\}^2}. \tag{19}$$

From this result, we can immediately observe that as n' gets bigger, $\alpha^\star$ gets smaller, which implies that the more replications we make from the follow-up experiment, the less we value the replications from the nominal experiment, which makes sense.

For the simplest case, suppose that the analyst did not collect additional real-world input data and ran the follow-up experiment with the same input models $\widehat{\mathbf{F}}_{\mathbf{m}}$. Then $\alpha^\star = n/(n+n')$ because $m_\ell = m'_\ell$ for all ℓ. This is clearly the weight we would use to pool two estimators from n and n' replications generated by the same input models.

More generally, when $m'_\ell > m_\ell$ for at least one ℓ there is a trade-off because pooling leads to more bias as $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ tends to have a larger bias than $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$. If $\bar{Y}_n(\widehat{\mathbf{F}}_{\mathbf{m}})$ is significantly more biased than $\bar{Y}_{n'}(\widehat{\mathbf{F}}_{\mathbf{m}'})$ then it is less attractive to pool the two estimators and we suspect this is often the case. To illustrate this point, suppose that overall input uncertainty is substantial (e.g., $\gamma = 20$) and $\widehat{F}_1$ has the largest contribution among all input models while others have negligible contributions. In this case, the analyst might collect a large additional sample ($m'_1 \gg m_1$) from F_1^c to update $\widehat{F}_1$, while keeping all other input models the same as in the

nominal experiment. When this happens $\alpha^\star$ becomes

$$\alpha^\star = \frac{(\sigma^2/n') + ((\nu_1\sigma^2(F_1^c))^2/(m_1'-1))((1/(m_1'-1)) - (1/(m_1-1)))}{(\sigma^2/n) + (\sigma^2/n') + (\mathsf{V}_1(m_1) - \mathsf{V}_1(m_1')) + \{\nu_1\sigma^2(F_1^c)((1/(m_1'-1)) - (1/(m_1-1)))\}^2}. \quad (20)$$

The denominator in Equation (20) is strictly positive as $\mathsf{V}_1(m_1) > \mathsf{V}_1(m_1')$. However, the numerator can be negative if

$$\left|\frac{\sigma^2}{n'}\right| < \left|\frac{\left(\nu_1\sigma^2(F_1^c)\right)^2}{m_1'-1}\left(\frac{1}{m_1'-1} - \frac{1}{m_1-1}\right)\right|,$$

in which case $\alpha^\star = 0$. Even if the numerator is greater than zero, we suspect $\alpha^\star$ will be near zero for the following reasons.

1. The analyst may run more replications for the follow-up experiment than the nominal experiment ($n' \geq n$) as it provides a less biased (more accurate) estimator, which makes $\sigma^2/n' < \sigma^2/n$.
2. $(\sigma^2/n') + ((\nu_1\sigma^2(F_1^c))^2/(m_1'-1))((1/(m_1'-1)) - (1/(m_1-1))) < \sigma^2/n'$;
3. since $m_1' \gg m_1$, we expect $\mathsf{V}_1(m_1) - \mathsf{V}_1(m_1') \approx \mathsf{V}_1(m_1)$; and
4. since γ is large and $\widehat{F}_1$ has the biggest contribution, $\mathsf{V}_1(m_1) \gg \dfrac{\sigma^2}{n}$.

From 1 to 4;

$$\mathsf{V}_1(m_1) \gg \frac{\sigma^2}{n'} + \frac{(\nu_1\sigma^2(F_1^c))^2}{m_1'-1}\left(\frac{1}{m_1'-1} - \frac{1}{m_1-1}\right),$$

and, therefore, $\alpha^\star$ becomes small. Hence, the more we real-world data we collect for the follow-up experiment, the less benefit there is in variance reduction from pooling the two estimators, whereas the relative disadvantage from introducing additional bias increases.

As mentioned earlier, the result in Equation (19) holds under Model (6) when we assume that $\widehat{\mathbf{F}}_{\mathbf{m}}$ and $\widehat{\mathbf{F}}_{\mathbf{m}'}$ are collections of ecdfs. We believe that $\alpha^\star$ is typically near zero in this case, which implies that it is better to use the estimator from the follow-up experiment without pooling. We also suspect a similar conclusion holds when any of the input distributions are parametric.

5. Design of the diagnostic experiment

The diagnostic experiment is an essential component of our method. There are three key experiment design questions.

1. *How should we select design points?* As discussed in Section 3.1, to fit Equation (14) we have chosen bootstrap generated empirical distributions $\widehat{\mathbf{F}}_1^\star, \widehat{\mathbf{F}}_2^\star, \ldots, \widehat{\mathbf{F}}_B^\star$ as design points. This concentrates the design where we need to fit well, and it facilitates combining results from the nominal and diagnostic experiment.
2. *Should we use Common Random Numbers (CRNs) across the diagnostic simulations?* The bootstrap design points $\widehat{\mathbf{F}}^{\star(1)}, \widehat{\mathbf{F}}^{\star(2)}, \ldots, \widehat{\mathbf{F}}^{\star(B)}$ must be sampled independently from $\widehat{\mathbf{F}}$, but we can choose to use the same random numbers for the simulations conducted with each $\widehat{\mathbf{F}}^{\star(b)}$. Kleijnen (1988) and others have shown that CRN tend to reduce the variance of the slope-parameter estimators in least-squares regression, which are $\widehat{\beta}_\ell, \widehat{\nu}_\ell$, $\ell = 1, 2, \ldots, L$ in Model (6). Since the variance of $\widehat{\mathsf{V}}_\ell$ is an increasing function of the variances of $\widehat{\beta}_\ell$ and $\widehat{\nu}_\ell$, it seems clear that using CRNs is desirable.
3. *Given a budget of N simulation replications, how should it be divided between design points (B) and simulation replications per design point (R) so that RB = N?* Ankenman and Nelson (2012) showed that with their method for assessing overall input uncertainty, if N is not too small then $B \approx 10$ is optimal in terms of minimizing the expected width of the CI for γ. However, our objective is different: we focus on providing estimates of the contribution of each input model. Below we argue that $B = N$ ($R = 1$) is the best choice in terms of statistical efficiency, but $B = 2L + 3$ is best for computation. Therefore, we provide a recommendation that balances these two objectives.

First, $B = N$ is the optimal design to minimize the MSE of the combined estimator from the nominal experiment ($\bar{Y}(\widehat{\mathbf{F}})$) and the diagnostic experiment ($\bar{\bar{Y}}(\widehat{\mathbf{F}})$). This is because the conditional variance of $\bar{\bar{Y}}(\widehat{\mathbf{F}})$ can be approximated as $\mathrm{Var}[\bar{\bar{Y}}(\widehat{\mathbf{F}})|\widehat{\mathbf{F}}] \approx \sigma_I^2/B + \sigma^2/N$ under the bootstrap approximation in Equation (10); see Appendix B of the online supplement.

Second, as described in Section 3, we estimate the parameters $\beta_0, \beta_1, \ldots, \beta_L$ and $\nu_1, \nu_2, \ldots, \nu_L$ in Model (6) by regressing B simulation output estimators on the means and variances of bootstrapped ecdfs. The design matrix for the regression is

$$\begin{bmatrix} 1 & \mu(\widehat{F}_1^{\star(1)}) & \sigma^2(\widehat{F}_1^{\star(1)}) & \cdots & \mu(\widehat{F}_L^{\star(1)}) & \sigma^2(\widehat{F}_L^{\star(1)}) \\ 1 & \mu(\widehat{F}_1^{\star(2)}) & \sigma^2(\widehat{F}_1^{\star(2)}) & \cdots & \mu(\widehat{F}_L^{\star(2)}) & \sigma^2(\widehat{F}_L^{\star(2)}) \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \mu(\widehat{F}_1^{\star(B)}) & \sigma^2(\widehat{F}_1^{\star(B)}) & \cdots & \mu(\widehat{F}_L^{\star(B)}) & \sigma^2(\widehat{F}_L^{\star(B)}) \end{bmatrix}.$$

As we have $2L + 1$ parameters, we need at least $2L + 2$ unique rows in the design matrix, which implies that it is necessary to have $B \geq 2L + 2$. However, $B \geq 2L + 2$ is not always sufficient to obtain the required number of *unique* rows.

If our input models include at least one continuous parametric distribution among $\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_L$, then with prob-

ability one all rows in the design matrix will be unique because the probability of two bootstrap samples with the same sample moments is zero for a continuous distribution. On the other hand, if all input models in $\widehat{F}_1, \widehat{F}_2, \ldots, \widehat{F}_L$ are ecdfs or discrete parametric distributions, then we need to be more cautious. Bernoulli distributions, particularly when the probability of success is extreme, are the most challenging cases since ties in the sample moments are quite likely when m_ℓ is small. Thus, a larger m_ℓ gives more opportunities for unique rows. Also worth noting is that the likelihood of identical rows diminishes as the number of input models L increases, so the problem is less pronounced in simulations with many input models.

Finally, the contribution estimator $\widehat{\mathsf{V}}_\ell$ in Equation (12) is a function of β_ℓ and ν_ℓ, so a good stand-in for the properties of $\widehat{\mathsf{V}}_\ell$ are the properties of $\widehat{\beta}_\ell$ and $\widehat{\nu}_\ell$. To illustrate why $B = N$ is statistically best, we temporarily simplify Model (6) to only include the "mean effects" and no CRN:

$$Y_j(\widehat{\mathbf{F}}) = \beta_0 + \sum_{\ell=1}^{L} \beta_\ell \mu(\widehat{F}_\ell) + \varepsilon_j. \tag{21}$$

Then $\widehat{\mathsf{V}}_\ell(m_\ell) = \widehat{\beta}_\ell^2 \mathrm{Var}[\widehat{\mu}(\widehat{F}_\ell)]$. If we assume that both $\mu(\widehat{F}_\ell)$ and ε_j are normally distributed—which is plausible since $\mu(\widehat{F}_\ell)$ is asymptotically normally distributed with large sample size m_ℓ—then standard results show that

$$\mathrm{Var}[\widehat{\beta}_\ell] = \frac{1}{N}\left(\frac{B}{B-L-2}\right)\frac{\sigma^2}{\mathrm{Var}[\mu(\widehat{F}_\ell)]} \tag{22}$$

when we force $RB = N$. Notice also that

$$\mathrm{E}\,[\widehat{\beta}_\ell^2] = \beta_\ell^2 + \mathrm{Var}\,[\widehat{\beta}_\ell] = \beta_\ell^2 + \frac{1}{N}\left(\frac{B}{B-L-2}\right) \times \frac{\sigma^2}{\mathrm{Var}\,[\mu(\widehat{F}_\ell)]}. \tag{23}$$

Clearly, $B = N$ is best for minimizing variance and bias, but the marginal impact of increasing B diminishes when $BR = N$ is fixed. If we extend this analysis in the natural way to include the "variance effects" as in Model (6), then the $-L$ terms in the denominators of Equation (22) and (23) become $-2L$.

On the other hand, from a computational efficiency point of view using $B < N$ ($R > 1$) has advantages over $B = N$ ($R = 1$). There is typically a computational set-up cost for simulating a new design point (which is really a new simulation model) but very little setup required for each additional replication at a design point. If $B < N$ then there are fewer setups. Furthermore, the number of rows in the design matrix is B, implying the need to manipulate a $B \times (2L + 1)$ matrix to fit the regression model. This argues for smaller B.

We must have $B \geq 2L + 2$. When larger N is feasible, we recommend making B large enough so that the incremental decrease in the variance and bias is small, say $< \delta$. Therefore, we select the smallest B such that

$$\frac{d}{dB}\left(\frac{B}{B-2L-2}\right) > -\delta,$$

which implies selecting the smallest B such that

$$B > 2L + 2 + \sqrt{\frac{2L+2}{\delta}}$$

and $R = \lfloor N/B \rfloor$. For instance, if $L = 5$ and $\delta = 0.01$—a 1% marginal decrease—then $B \approx 12 + 35 = 47$.

6. Empirical results

This section summarizes an empirical study of the proposed method applied to two simple examples and also an illustration on a realistic problem.

In the two simple examples we apply our method and provide intuitive explanations for the results, as well as compare them to true input-model contributions as defined in Equation (4) that were estimated precisely from side experiments. These side experiments exploit the fact that the true, correct real-world distributions are known, which is obviously not the case in practice. In both examples we define F_ℓ^c for each input and then sample m_ℓ observations that we treat as the real-world data.

We first evaluate the method using a well-known Stochastic Activity Network (SAN); see Fig. 1. The goal is to estimate the mean time to complete the network. We defined a set of real-world input distributions for the activity times $\mathbf{F}^c = \{F_1^c, F_2^c, \ldots, F_5^c\}$. Experiments under different settings of sample size and mean and variance of the activity-time distributions were conducted and the variance contribution and sensitivity of each input distribution was estimated.

The second example is an $M/M/1/k$ queueing system simulation that has two input distributions: interarrival time and service time. The goal of the simulation is to estimate the mean steady-state queue length, which is known to be a highly nonlinear function of the means of the two

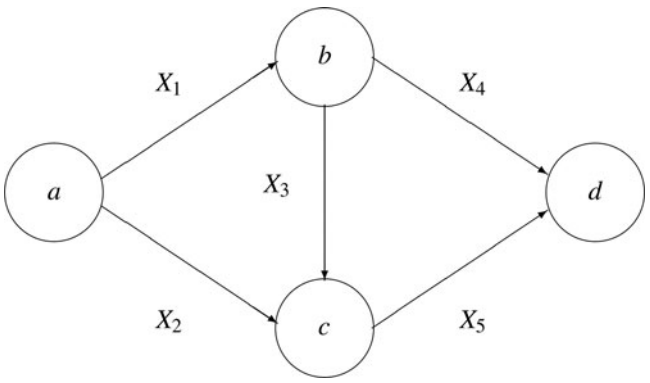
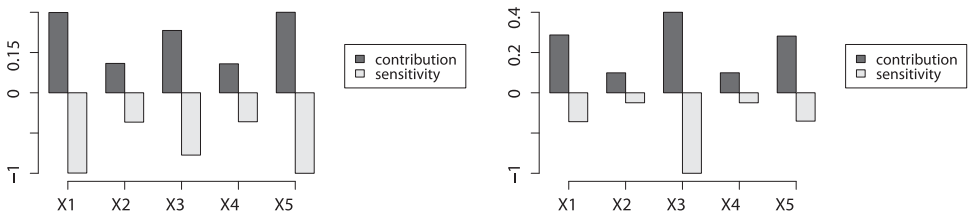


Fig. 1. A small SAN.



(i) $m_\ell = 100, \ell = 1, 2, \ldots, 5$. (ii) $m_\ell = 100, \ell = 1, 2, 4, 5$, and $m_3 = 50$.

Fig. 2. Scaled contributions (positive) and sensitivities (negative) of input models for the SAN with different activity-time sample sizes.

input distributions. We use this example to show the performance of our method when the mean response is a nonlinear function of the means and variances of the input distributions. Different settings of the means of the interarrival times and service times as well as the capacity of the queue were tested to evaluate the performance of our method.

We conclude with a more realistic example of simulating a remote order taking system to illustrate the practical application of our method.

6.1. *SAN*

For each activity time X_ℓ we have a correct real-world distribution F_ℓ^c, which we pretend is unknown. Using m_ℓ samples generated from F_ℓ^c, we fit ecdf $\widehat{F}_\ell$ for each X_ℓ. The total time to finish the project is $Y = \max\{X_1 + X_4, X_1 + X_3 + X_5, X_2 + X_5\}$ and the purpose of the simulation is to estimate the expected value of Y. In the remainder of the section we describe empirical results for contribution and sensitivity of the input distributions under two experimental settings: (i) different real-world sample sizes and (ii) different means and variances of the activities.

To evaluate how well our method measures the contribution of each input model, a series of side experiments was conducted for each experimental setting. The side experiment estimates the contribution directly from the definition in Equation (4) by exploiting the fact that we know the true distributions in this example.

Side Experiment

1. Given the true distributions $F_1^c, F_2^c \ldots, F_5^c$ for activity times $X_1, X_2, \ldots, X_5$, do the following:
2. For $b = 1$ to B
   a. Generate m_1 samples from F_1^c and fit $\widehat{F}_1^{(b)}$.
   b. Run R replications $Y_j(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c)$, $j = 1, 2, \ldots, R$ using the fitted $\widehat{F}_1^{(b)}$ and true distributions $F_2^c, F_3^c \ldots, F_5^c$ as the input models.
   c. Calculate the sample mean $\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c)$ from the replications.
3. Calculate the sample variance of $\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c)$, $b = 1, 2, \ldots, B$

$$\frac{1}{B-1}\sum_{b=1}^{B}\left\{\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c) - \bar{\bar{Y}}(\widehat{F}_1, F_2^c, \ldots, F_5^c)\right\}^2 \tag{24}$$

   to estimate the true contribution V_1 of X_1, where

$$\bar{\bar{Y}}(\widehat{F}_1, F_2^c, \ldots, F_5^c) = \frac{1}{B}\sum_{b=1}^{B}\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c).$$

4. Conduct steps 2 and 3 for each distribution F_2^c, F_3^c, F_4^c, F_5^c in turn.

A key point is that we make R large enough that

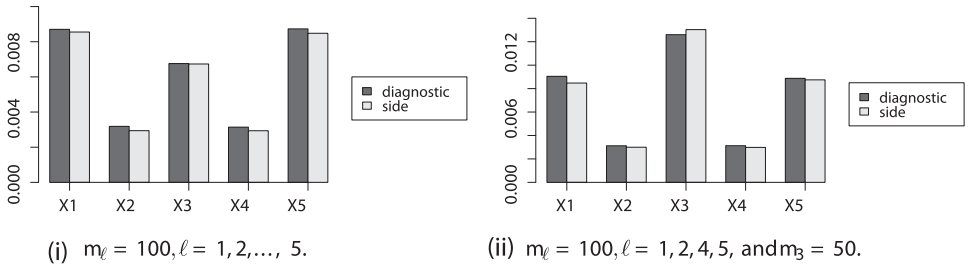
$$\mathrm{Var}\big[\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c)\big|\,\widehat{F}_1^{(b)}\big] \approx 0.$$

This implies that $\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c) \approx \mathrm{E}\big[\bar{Y}(\widehat{F}_1^{(b)}, F_2^c, \ldots, F_5^c)\big|\,\widehat{F}_1^{(b)}\big]$ and we can treat the sample variance (24) as an estimate of the variance due to different real-world samples; i.e., contribution V_1 only. We used ($B = 100$, $R = 5000$) that gave relative errors of less than 2%. Notice that B and R in these side experiments are unrelated to the choices we make for the diagnostic experiment.

6.1.1. *SAN with different activity-time sample sizes*

In this experiment, all activity times have exponential real-world distributions with mean one. Thus, the path $X_1 + X_3 + X_5$ is likely to be the longest. To see the effect of different real-world sample sizes on $\widehat{\mathsf{V}}_\ell$, we conducted simulations for two cases: (i) $m_\ell = 100, \ell = 1, 2, \ldots, 5$; and (ii) $m_\ell = 100, \ell = 1, 2, 4, 5$, and $m_3 = 50$. We used ($B = 50$, $R = 200$) for the diagnostic experiment.

Figure 2 displays the simulation results when averaged over 1000 macro replications to provide a relative error of less than 3%. Plotted are the relative contributions $\widehat{\mathsf{V}}_\ell(m_\ell)/\sum_{i=1}^{L}\widehat{\mathsf{V}}_i(m_i)$, and the sensitivities scaled so that the greatest sensitivity has value -1. In case (i), X_1 and X_5 have the largest contributions and are more sensitive to additional real-world data since these two activity times are involved in two-out-of-three paths on the network. On



(i) $m_\ell = 100, \ell = 1, 2, \ldots, 5$. (ii) $m_\ell = 100, \ell = 1, 2, 4, 5$, and $m_3 = 50$.

Fig. 3. Comparison of estimated input-model contributions and the true contributions for the SAN with different activity-time sample sizes.

the other hand, X_2 and X_4 have the smallest contributions and sensitivities since they are only involved in one path that is likely to be dominated by $X_1 + X_3 + X_5$. The higher contribution/sensitivity of X_3 compared with X_2 and X_4 can be explained by the same reasoning. This trend changes slightly when m_3 decreases to 50 in case (ii): since there are fewer real-world samples for X_3 in this case, X_3 shows the largest contribution and sensitivity. This is a good illustration that input uncertainty is a combination of how much the distribution itself matters in the simulation and how well it has been estimated.

Figure 3 compares the estimated contributions from the diagnostic experiments and the side experiments for cases (i) and (ii). Notice that the contributions are not scaled in this graph. Clearly, the estimated contributions from the two approaches are close, which implies our diagnostic experiment successfully estimated the true contributions.

6.1.2. *SAN with different activity-time distributions*

In this section, gamma distributions are used as the true distributions for the activity times. Two sets of experiments were conducted to investigate the effect of different mean and variance values on the contribution of X_3. In the first set of experiments, we fixed the variance of all activity times to 0.5 and set X_1, X_2, X_4, X_5 to have mean one and X_3 to have mean 10. In the second set, the means of all activities are fixed to 10, but X_1, X_2, X_4, X_5 have variance one, whereas X_3 has variance five. In all cases, $m_\ell = 100$ real-world samples were collected for each activity time to fit ecdf $\widehat{F}_\ell$. For the diagnostic experiment we used ($B = 50$, $R = 200$) and the results were averaged over 1000 macroreplications.

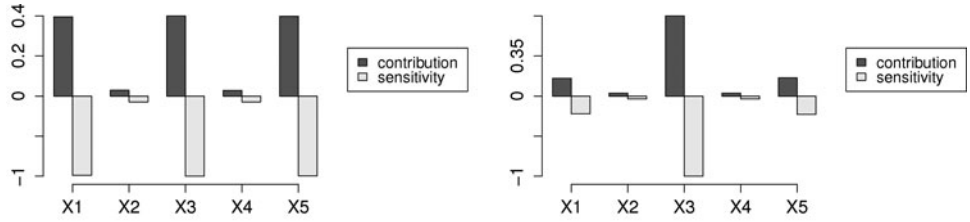
Figure 4 shows the experimental results. Observe that in the first set the trend in contribution/sensitivity is not much different from case (i) in Section 6.1.1. This might not seem intuitive since one might think that the large mean value of X_3 would increase its contribution. However, a large mean for X_3 only makes the path $X_1 + X_3 + X_5$ more likely to be dominant and, therefore, the input uncertainty due to X_1 and X_5 still has a significant impact on output variability since they are also included in other paths. In the second set, X_3 has the largest contribution and is more sensitive to additional data collection, which can be explained by its relatively large variance.

Figure 5 presents the estimated input contributions from the side experiments and compares them to the results from our method. As in Section 6.1.1, the two estimated values are close in both cases (i) and (ii).

6.2. *M/M/1/k queueing system*

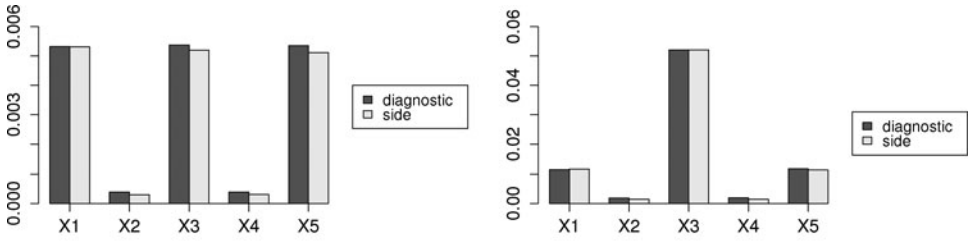
In this section, we apply our method to a queueing system with finite capacity k where the interarrival times and services times are i.i.d. exponential random variables with mean θ_1 and θ_2, respectively. Our measure of interest is the mean of the number of customers in the system at steady-state, Y. Given θ_1 and θ_2, it is known that the steady-state number of customers in system n follows a truncated geometric distribution with probability

$$\mathrm{P}_n = (\theta_2/\theta_1)^n \frac{1 - \theta_2/\theta_1}{1 - (\theta_2/\theta_1)^{k+1}}, \quad \text{for} \quad n = 0, 1, \ldots, k, \quad (25)$$



(i) $X_\ell \sim \text{Gamma}(2,2), \ell = 1,2,4,5$ and $X_3 \sim \text{Gamma}(200,20)$. (ii) $X_\ell \sim \text{Gamma}(100,10), \ell = 1,2,4,5$ and $X_3 \sim \text{Gamma}(20,2)$.

Fig. 4. Scaled contributions (positive) and sensitivities (negative) of input models for the SAN with different activity-time distributions.



(i) $X_\ell \sim \text{Gamma}(2,2), \ell = 1,2,4,5$ and $X_3 \sim \text{Gamma}(200,20)$.

(ii) $X_\ell \sim \text{Gamma}(100,10), \ell = 1,2,4,5$ and $X_3 \sim \text{Gamma}(20,2)$.

Fig. 5. Comparison of estimated input-model contributions and the true contributions for the SAN with different activity-time distributions.

provided $\theta_1 \neq \theta_2$. Therefore, the expected number of customers in the system is

$$\mathrm{E}[Y|\theta_1,\theta_2] = \frac{\theta_2/\theta_1}{1-\theta_2/\theta_1} - \frac{(k+1)(\theta_2/\theta_1)^{k+1}}{1-(\theta_2/\theta_1)^{k+1}}, \qquad (26)$$

which is clearly a nonlinear function of θ_1 and θ_2. Assume for the moment that the expression in Equation (26) is unknown and we have to estimate it by simulation, where the true parameters θ_1^c and θ_2^c are estimated from the observed interarrival times and the service times, respectively. Then we can use our method to analyze the input uncertainty in the simulation estimator and evaluate the contributions of the two input models. However, our method fits Model (6) to the simulation response. In other words, it assumes $\mathrm{E}[Y|\widehat{\theta}_1,\widehat{\theta}_2]$ to be a linear function of $\widehat{\theta}_1, \widehat{\theta}_1^2, \widehat{\theta}_2$, and $\widehat{\theta}_2^2$. Therefore, the estimation quality of contributions depends on how well Model (6) approximates $\mathrm{E}[Y|\widehat{\theta}_1,\widehat{\theta}_2]$ near θ_1^c and θ_2^c.

To investigate the performance of our method, the estimated contributions from diagnostic experiments at different settings of θ_1^c, θ_2^c, and k are compared to the estimated contributions from the side experiments. As in the SAN example in Section 6.1, we exploit the fact that θ_1^c and θ_2^c are known. In all cases the sample sizes for the real-world interarrival times and services times are $m_1 = 200$ and $m_2 = 100$, respectively. Table 1 provides the estimated contributions from the diagnostic experiments and the side experiments. For the diagnostic experiment, we used $B = 40$, $R = 100$. The results are averaged over 100 macro-replications and the standard errors are provided in parentheses.

In all settings, the estimated contributions are of the same order of magnitude (by rounding to the first decimal point, if necessary) with the results from the side experiments. Also, the order of importance is preserved; i.e., the contribution of the service time is higher than that of the interarrival time in all cases in both diagnostic and side experiments. The first two cases show the results when the queue is congested. Therefore, the system is likely to be filled up to its limit size k. In fact, in both $k = 1$ and $k = 100$ cases (26) is quite flat near $(\theta_1^c = 0.2, \theta_2^c = 1)$. Therefore, Model (6) captures the shape of Equation (26) well, which results in good estimation quality. In Cases 3 and 4, the queue is lightly loaded and therefore the expected queue length in this case is less than k. Especially when $k = 100$, the system effectively has no capacity limit and behaves similar to a corresponding $M/M/1$ queueing system. When $k = 1$, Equation (26) shows smooth curvature near $(\theta_1^c = 1, \theta_2^c = 0.5)$, which can be easily captured by Model (6). When $k = 100$, Equation (26) is flat near $(\theta_1^c = 1, \theta_2^c = 0.5)$. Therefore, in both cases the diagnostic experiments approximate the contributions well. When $\theta_1^c = 1$ and $\theta_2^c = 0.9$, our method performs better in Case 5 than in Case 6. This difference is because the traffic intensity θ_2^c/θ_1^c is close to one in this case. In Case 5, the size of the queue is limited by $k = 1$; therefore, even if the estimated traffic intensity $\widehat{\theta}_2/\widehat{\theta}_1$ is greater than one, the queue length is still less than one. In Case 6, however, if $\widehat{\theta}_2/\widehat{\theta}_1 > 1$, the simulated average queue length becomes close to $k = 100$, whereas it is much less than 100 when $\widehat{\theta}_2/\widehat{\theta}_1 < 1$. Therefore, as small change in $\widehat{\theta}_1$ and $\widehat{\theta}_2$ can cause the mean response to change dramatically and Model (6) approximates the surface relatively poorly in Case 6. However, our method still estimates the contributions of the input distributions to the right order of magnitude and the order of importance is also correct in this case compared to the results from the side experiment.

From this example, we can confirm that even if the mean response is a highly nonlinear function of input distributions' means and variances, our method works reasonably well. Notice that our goal is not to approximate the mean response surface globally but rather to fit Model (6) near the true means and variances of the input distributions. Therefore, even when the mean response is nonlinear globally, if it is linear in terms of means and variances in the neighborhood of interest, our method works well.

6.3. *Illustration: remote order-taking system*

In this section we apply our method to a more realistic simulation to illustrate the sequence of experiments proposed in Section 3. All experiments presented in this section were conducted by using Simio (www.simio.com) and functionality that has been added to Simio that performs the diag-

Table 1. Estimated contributions of interarrival-time and service-time distributions from diagnostic experiments and side experiments at different settings of θ_1^c, θ_2^c, and k

				Interarrival time		*Service time*	
Case	θ_1^c	θ_2^c	k	*Diagnostic*	*Side*	*Diagnostic*	*Side*
1	0.2	1	1	$1.55 \times 10^{-4}(1.53 \times 10^{-5})$	9.46×10^{-5}	$2.72 \times 10^{-4}(2.15 \times 10^{-5})$	1.95×10^{-4}
2	0.2	1	100	$7.33 \times 10^{-3}(5.81 \times 10^{-5})$	4.88×10^{-4}	$1.18 \times 10^{-3}(7.07 \times 10^{-5})$	1.06×10^{-3}
3	1	0.5	1	$3.31 \times 10^{-4}(2.57 \times 10^{-5})$	2.49×10^{-4}	$5.97 \times 10^{-4}(4.07 \times 10^{-5})$	4.84×10^{-4}
4	1	0.5	100	$2.95 \times 10^{-2}(2.62 \times 10^{-3})$	2.25×10^{-2}	$5.51 \times 10^{-2}(4.57 \times 10^{-3})$	4.71×10^{-2}
5	1	0.9	1	$4.33 \times 10^{-4}(3.33 \times 10^{-5})$	3.11×10^{-4}	$7.03 \times 10^{-4}(3.98 \times 10^{-5})$	6.25×10^{-4}
6	1	0.9	100	$1.92 \times 10^{2}(1.50 \times 10)$	3.31×10^{2}	$3.83 \times 10^{2}(2.90 \times 10)$	5.77×10^{2}

nostic experiment and displays our contribution measures. The problem, which is taken from Nelson (2013), is to evaluate replacing the current drive-through order windows for a chain of fast-food restaurants with the equivalent of a call center. The current design for each store has two fully staffed windows, one for order taking and another for food delivery. The proposal is to replace the first window with a remote order-taking service in which agents communicate with customers through the electronic order board and then relay the order to the store. The fast-food chain has high standards for customer service and requires that the average waiting time for a customer to be greeted by an agent once they reach the order board be less than 1.5 seconds. The proof-of-concept simulation uses data from seven stores to investigate whether the call center can meet this standard with substantially fewer agents than the number of stores served.

There are nine sets of real-world data (actually created by us): customer interarrival times from the seven stores; order-taking times; and the time for a car to pull away from the order board and the next one in line (if there is one) to pull up. Interarrival times were collected during the busiest 3-hour period over 10 days. The numbers of observations of order-taking time and car-moving time available were 150 and 70, respectively. With these data, parametric distributions were fitted for the initial experiment (the fitted distributions were exponentials, Weibulls, and gammas, and we did not exploit knowing the true distribution families). To evaluate busy-period performance, steady-state simulations were conducted using a replication-deletion experiment design employing $n = 1000$ replications. With four agents at the call center, the 95% confidence interval for mean customer waiting time in the nominal experiment was 0.99 ± 0.04 seconds, which clearly excludes 1.5 seconds.

To analyze the impact of input uncertainty, we conducted a diagnostic experiment with ($B = 80$, $R = 125$). The design was chosen using the guidelines presented in Section 5. Below we report results from the simulation with four agents, which we concluded to be adequate based on the nominal experiment.

The diagnostic experiment gave $\widehat{\gamma} = 17.48$, which implies that input uncertainty is significantly larger than simulation uncertainty. If we adjust the CI to account for input uncertainty, the 95% CI length should be more like $0.04 \times \sqrt{1 + 17.48^2} \approx 0.63$. This is a rough adjustment, and we do not claim that it is a CI in the formal sense. Clearly, the adjusted interval 0.99 ± 0.63 includes the critical value 1.5. Therefore, to reduce the input uncertainty we are interested in which distributions are the largest contributors.

Table 2 displays the estimated scaled contribution and sensitivity of each input model, as well as its real-world sample size and its sample mean (in seconds). The input model for order-taking time makes the largest contribution to input uncertainty, and it is the most sensitive to collecting additional data. This makes sense: every customer requires an order-taking time, and the real-world sample size for this input is relatively small. If it is possible to collect additional input data, then this is the distribution for which we would achieve the most benefit.

Notice that the rank order of contribution and sensitivity do not always coincide. For instance, the interarrival times of stores 6 and 7 have the same contribution; however, the sensitivity of the former is greater because it has a smaller sample size, so that additional real-world samples for store 6 can reduce the variability in the simulation result by a larger amount.

Table 2. Scaled contribution and sensitivity results for the remote order-taking system

ℓ	*Input data*	m_ℓ	*Average*	*Contribution*	*Sensitivity*
1	Interarrival 1	351	307.0	0.014	−0.007
2	Interarrival 2	153	701.3	0.011	−0.012
3	Interarrival 3	421	256.1	0.021	−0.008
4	Interarrival 4	261	413.0	0.008	−0.005
5	Interarrival 5	342	308.9	0.018	−0.008
6	Interarrival 6	354	304.4	0.006	−0.003
7	Interarrival 7	472	228.5	0.006	−0.002
8	**Order taking**	**150**	84.8	**0.914**	**−1.000**
9	Moving	70	5.4	0.003	−0.006

Boldface type indicates the input model with the largest contribution and sensitivity.

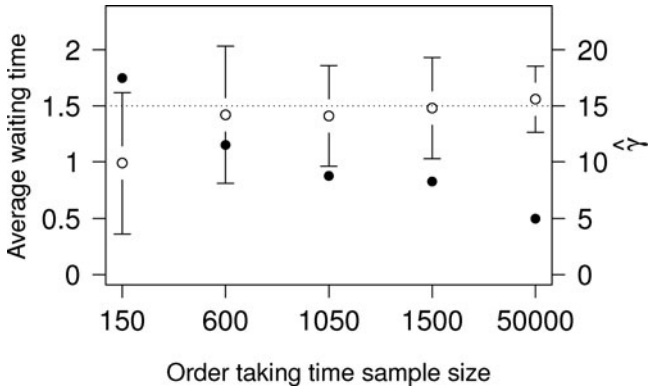


Fig. 6. The adjusted CI for expected customer waiting time along with $\widehat{\gamma}$ (indicated by •) for different real-world sample sizes for the order-taking time.

Since we concluded that the order-taking time has the largest contribution, we collected more real-world data to perform a follow-up experiment (the "real-world" order-taking time in this example is an exponential distribution with mean 90 seconds). Figure 6 shows the trend in $\widehat{\gamma}$, indicated by •, and the roughly adjusted CI for expected waiting time, as a function of the total amount of real-world data on order taking. These are CIs from the nominal experiments inflated by $\sqrt{1+\widehat{\gamma}^2}$, not CIs using combined nominal and diagnostic experiment results (we discuss one such CI below). The number of simulation replications remains $n = 1000$.

We observe that $\widehat{\gamma}$ tends to decrease as we collect more order-taking data. However, even with 50 000 observations of the order-taking time, $\widehat{\gamma}$ does not become zero because there is still residual input uncertainty from the other distributions. The contribution and the sensitivity of the order-taking time also decrease as the sample size grows. At $m = 50\,000$, the relative contribution of the order taking time is 4.1%, which is smaller than any of the other distributions except the moving time. We also observe that the average customer waiting time increases as we collect more data. This implies that the first 150 real-world observations did not provide a particularly good approximation of the true distribution. With more real-world data, the adjusted CI length decreases and it becomes more clear that the critical value (1.5 seconds) has not been achieved. Therefore, we need to consider increasing the number of agents from four to five, which we would not have considered based on the nominal experiment alone.

In this illustration we chose to collect more real-world data on the order-taking time to mitigate input uncertainty. Had that not been possible, then we could have improved our point estimate from the nominal experiment by using the combined estimator $\tilde{Y} = \alpha^\star \bar{Y}(\widehat{\mathbf{F}}) + (1-\alpha^\star)\bar{\bar{Y}}(\widehat{\mathbf{F}})$, where

$$\alpha^\star = \frac{(\mathrm{b}^\star)^2 + \sigma_I^2/B + \sigma^2/BR}{(\mathrm{b}^\star)^2 + \sigma_I^2/B + \sigma^2/BR + \sigma^2/n}.$$

To estimate $\alpha^\star$ we used $\widehat{\sigma}_I^2$ from the diagnostic experiment, $\widehat{\sigma}^2$ from the nominal experiment, and $\widehat{b}^\star = \bar{\bar{Y}}(\widehat{\mathbf{F}}) - \bar{Y}(\widehat{\mathbf{F}})$ to estimate the bias. This gave $\widehat{\alpha} = 0.88$, putting most of the weight on the nominal experiment. In this case the adjustment was so small that $\tilde{Y}$ is the same as $\bar{Y}(\widehat{\mathbf{F}})$ to two decimal places. It is important to note that the estimated bias from bootstrapping matters: had we ignored it (set $\mathrm{b}^\star = 0$) then $\widehat{\alpha} = 0.80$ putting somewhat more weight on the results from the diagnostic experiment. The heuristically adjusted CI, which uses data from the nominal and diagnostic experiment, has halfwidth $\pm 1.96\sqrt{\widehat{\sigma}_I^2 + \widehat{\mathrm{MSE}}} = \pm 0.45$ seconds, which is narrower than if we had used the nominal data alone.

7. Conclusions

In this article we presented a method to quantify the overall impact of input-model uncertainty on simulation-based performance estimates, to identify which input models make the largest contribution to this uncertainty, and to identify the input data sources from which additional real-world observations would lead to the greatest reduction in input uncertainty. Our approach builds on Ankenman and Nelson (2012) but obtains the contribution and sensitivity results from the same experiment that they used just to measure the overall input uncertainty.

Our philosophy is to try to obtain a lot of useful information without substantial additional effort or sophisticated computations. In fact, all we require is the capability to sample from the ecdf or fitted parametric distribution of the real-world input data—which is necessary for doing simulation—and least-squares regression to fit the mean-variance model. As a proof of concept, our diagnostic experiment and uncertainty measures have been implemented as a standard feature in Simio.

To achieve all of this we modeled the relationship between the input distributions and the simulation's output through the means and variances of these distributions, without interactions. We know that higher moments can matter, and there certainly could be interactions. However, our goal is to rank the input distributions as to their contributions to input uncertainty, rather than to precisely estimate each contribution. For this goal, a first-order model that characterizes a distribution by its mean and variance will often suffice, as illustrated by our experiments.

Open questions remain about the total budget N that should be expended on the diagnostic experiment to obtain reliable results. We suspect that no blanket recommendation is possible, but instead N would have to be discovered adaptively. Important extensions include assessing the contributions of multivariate (e.g., age, health status, and income of customers) and time-dependent (e.g., arrivals to a call center) input processes.

Acknowledgments

We thank Bruce Ankenman, Russell Barton, Dennis Pegden, and Dave Sturrock for helpful discussions and suggestions. Portions of this article were previously published in the *Proceedings of the 2013 Winter Simulation Conference* as Song and Nelson (2013).

Funding

This research was partially supported by National Science Foundation Grant CMMI-1068473 and GOALI sponsor Simio LLC.

Supplemental Material

Supplemental data for this article can be accessed on the publisher's website.

References

Ankenman, B.E. and Nelson, B.L. (2012) A quick assessment of input uncertainty, in *Proceedings of the 2012 Winter Simulation Conference*, Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O. and Uhrmacher A.M. (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 241–250.

Barton, R.R. (2012) Tutorial: Input uncertainty in output analysis, in *Proceedings of the 2012 Winter Simulation Conference*, Laroque, C., Himmelspach, J., Pasupathy, R., Rose, O., and Uhrmacher A.M. (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 67–78.

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

Cheng, R.C.H. and Holland, W. (1998) Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation*, **60**, 183–205.

Cheng, R.C.H. and Holland, W. (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.

Cho, E. and Cho, M.J. (2008) Variance of sample variance, in *Proceedings of the 2008 Joint Statistical Meetings, Section on Survey Research Methods*, American Statistical Association, Washington, DC, pp. 1291–1293.

Freimer, M. and Schruben, L.W. (2002) Collecting data and estimating parameters for input distributions, in *Proceedings of the 2002 Winter Simulation Conference*, Yücesan, E., Chen, C., Snowdon, J.L. and Charnes, J.M. (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 392–399.

Hall. P. (1992) *The Bootstrap and Edgeworth Expansion*, Springer, New York, NY.

Homma, T. and Saltelli, A. (1996) Importance measures in global sensitivity analysis of model output. *Reliability Engineering and System Safety*, **52**, 1–17.

Kleijnen, J.P.C. (1988) Analyzing simulation experiments with common random numbers. *Management Science*, **34**, 65–74.

Marrel, A., Iooss, B., Da Veiga, S. and Ribatet, M. (2012) Global sensitivity analysis of stochastic computer models with joint metamodels. *Statistics and Computing*, **22**, 833–847.

Nelson, B.L. (2013) *Foundations and Methods of Stochastic Simulation: A First Course*, Springer, New York, NY.

Ng, S.H. and Chick, S.E. (2001) Reducing input parameter uncertainty for simulations, in *Proceedings of the 2001 Winter Simulation Conference*, Peters, B.A., Smith, J.S., Medeiros, D.J. and Rohrer, M.W. (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, 364–371.

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

Oakley, J.E. and O'Hagan, A. (2004) Probabilistic sensitivity analysis of complex models: a Bayesian approach. *Journal of the Royal Statistical Society*, **66**, 751–769.

Plischke, E., Borgonovo, E. and Smith, C.L. (2013) Global sensitivity measures from given data. *European Journal of Operational Research*, **226**, 536–550.

Song, E. and Nelson, B.L. (2013) A quicker assessment of input uncertainty, in *Proceedings of the 2013 Winter Simulation Conference*, Pasupathy, R., Kim, S.-H., Tolk, A., Hill, R. and Kuhl, M.E. (eds), Institute of Electrical and Electronics Engineers, Piscataway, NJ, pp. 474–485.

Wagner, H.M. (1995) Global sensitivity analysis. *Operations Research*, **43**, 948–969.

Zhang, L. (2007) Sample mean and sample variance: their covariance and their (in)dependence. *The American Statistician*, **61**, 159–160.

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

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

Biographies

Eunhye Song is a Ph.D. candidate in the Department of Industrial Engineering and Management Sciences at Northwestern University. Her research interests are input uncertainty quantification, design of simulation experiments, and simulation optimization. Her current research is focused on simulation optimization under input uncertainty.

Barry L. Nelson is the Walter P. Murphy Professor of the Department of Industrial Engineering and Management Sciences at Northwestern University. He is a Fellow of INFORMS and IIE. His research centers on the design and analysis of computer simulation experiments on models of stochastic systems, and he is the author of *Foundations and Methods of Stochastic Simulation: A First Course*.