Journal of Simulation (2015) 9, 249–259

www.palgrave-journals.com/jos/

Single-experiment input uncertainty

Y Lin, E Song and BL Nelson*

Northwestern University, Evanston, IL, USA

'Input uncertainty' refers to the simulation model risk caused by estimating input distributions from real-world data, and specifically the (usually unmeasured) variance in performance estimates that this introduces. We provide the first single-run method for quantifying input uncertainty, meaning that we derive our measure of input-uncertainty variance—both overall variance and the contribution to it of each input model—from the nominal experiment that the analyst would typically run using the estimated input models; other methods in the literature require additional diagnostic experiments. Application of our method is illustrated with two examples.

Journal of Simulation (2015) **9**(3), 249–259. doi:10.1057/jos.2015.2; published online 27 February 2015

Keywords: input modelling; output analysis; model risk

1. Introduction

Decisions based on models, including stochastic computer simulation models, are subject to various types of model risk. In this paper we consider the risk due to estimating ('fitting') simulation input models to finite samples of real-world data. We refer to the impact on simulation-based decisions from being uncertain about the true (perfect fidelity) input models, and specifically the additional error in the output performance estimates, as *input uncertainty*.

For example, later in the paper we describe the simulation of a computer communication network in which the mean of the exponential message-size distribution and the arrival rates for 12 different Poisson arrival processes for classes of messages are fit to network data. A simulation experiment is conducted that estimates the expected message delay in the network to a standard error of approximately $\pm 10^{-5}$ s *when uncertainty about these input parameters is ignored and only stochastic simulation variance is measured*. However, the additional error from input-parameter uncertainty is approximately $\pm 10^{-3}$ s, orders of magnitude larger than the simulation error alone. Thus, input uncertainty can be a significant model risk.

There is a large literature, both frequentist and Bayesian, on quantifying input uncertainty; see Barton (2012) and Song and Nelson (2014) for reviews. One characteristic of all of these methods is that they require additional experimentation beyond what we call the *nominal* experiment; that is, the experiment the analyst would have conducted had they made the common choice of ignoring input uncertainty. We refer to these additional experiments as *diagnostic*.

In a sequence of papers (Ankenman and Nelson, 2012; Song and Nelson, 2013, 2015) we have pursued 'quick' methods for assessing input uncertainty that are also easy to implement. While being quicker than many existing techniques, these methods nevertheless also require diagnostic experiments.

Our goal in this paper is to provide a full assessment of input uncertainty—the overall added variance and the contributions of each input model to it—using only the inputs and outputs generated by the nominal experiment and some routine calculations. *The fact that no additional experiments are required means that the barrier to practical application is low.* At its current stage of development our method can be applied to simulations with any number of parametric input models whose distribution families are known, but whose parameters are fit using maximum likelihood estimators (MLEs); empirical distributions and unspecified distribution families are currently outside the scope of our work. The method scales up easily to large numbers of input models.

Our single-experiment approach exploits the input-uncertainty approximation of Cheng and Holland (1997, 1998) and our extension of the gradient-estimation method of Wieland and Schmeiser (2006). Implementation requires collecting summary data on the simulation inputs, as well as the simulation outputs, and least-squares regression. We provide experiment results that illustrate the method's effectiveness.

The paper is organized as follows. We review Cheng and Holland's approximation next, including a small extension that allows us to attribute the added variance to individual input models. We then address estimating the terms in the approximation in Sections 3 and 4. An empirical evaluation is presented in Section 5, and conclusions are offered in Section 6.

2. Cheng and Holland's approximation

Here we review the variance approximation of Cheng and Holland (1997, 1998) that provides the foundation of our approach. While we will, in a sense, compete against their method for exploiting the approximation, their approximation

**Correspondence: BL Nelson, Department of Industrial Engineering and Management Sciences, Northwestern University, 2145 Sheridan Road, C210, Evanston, IL 60208-3119, USA.*
E-mail: nelsonb@northwestern.edu

is an important contribution to the simulation literature and central to our method.

Cheng and Holland (1997) assume that a computer simulation has p unknown input parameters associated with p or fewer independent input distributions. The number of distinct input distributions is less than or equal to p because some input models may require more than one input parameter (eg, the lognormal distribution requires the mean and variance). Further, the simulation outputs obtained from n replications can be represented as

$$Y_j(\mathbf{U}_j, \boldsymbol{\theta}) = \eta(\boldsymbol{\theta}) + e_j(\mathbf{U}_j, \boldsymbol{\theta}),\ j = 1, 2,\ \ldots\ , n,$$

where

$$\eta(\boldsymbol{\theta}) = \int Y(\mathbf{u}, \boldsymbol{\theta}) du$$

is the expected value of the simulation given parameter $\boldsymbol{\theta}$. The vector $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^{\mathrm{T}}$ represents the p input parameters, and $\mathbf{U}_1, \mathbf{U}_2, \ldots$ are the independent streams of uniform random numbers used in each replication. Where no confusion will arise, we will write simply Y_j or $Y_j(\boldsymbol{\theta})$.

Suppose that the true input parameters, denoted by $\boldsymbol{\theta}^0 = (\theta_1^0, \theta_2^0, \ldots, \theta_{\mathrm{p}}^0)^{\mathrm{T}}$, are unknown, but real-world data from the input processes are available. The goal of the simulation is to estimate $\eta(\boldsymbol{\theta}^0)$. We assume that MLEs of $\boldsymbol{\theta}^0$ are employed.[1] For example, if the ith input parameter is the rate of an exponential interarrival-time distribution and the observed real-world interarrival times are $Z_{i,1}, Z_{i,2},\ \ldots\ , Z_{i,m_i}$, then the MLE of θ_i^0 is $\widehat{\theta}_i = m_i / \left(\sum_{\ell=1}^{m_i} Z_{i,\ell}\right)$, where m_i is the number of real-world interarrival times. To simplify the presentation that follows, we assume that $m_i = m$ for all i, but this is not required.

Let $\widehat{\boldsymbol{\theta}} = (\widehat{\theta}_1, \widehat{\theta}_2,\ \ldots\ , \widehat{\theta}_p)^\top$ be the MLEs of the input parameters. The nominal performance-measure estimate is

$$\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{n}\sum_{j=1}^{n} Y_j\left(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}\right) = \frac{1}{n}\sum_{j=1}^{n}\left(\eta\left(\widehat{\boldsymbol{\theta}}\right) + e_j\left(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}\right)\right)$$

and its variance can be decomposed into two parts:

$$\mathrm{Var}\left[\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right)\right] = \mathrm{Var}_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{E}_U\left(\overline{Y} \mid \widehat{\boldsymbol{\theta}}\right)\right] + \mathrm{E}_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{Var}_U\left(\overline{Y} \mid \widehat{\boldsymbol{\theta}}\right)\right], \quad (1)$$

which represents the input-uncertainty variance and the stochastic-simulation variance, respectively. Because the MLEs $\widehat{\boldsymbol{\theta}}$ are based on real-world data, they are independent of the streams of random numbers $\mathbf{U}$. Therefore, $\mathrm{E}_{\mathbf{U}_j}[e_j(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}) \mid \widehat{\boldsymbol{\theta}}] = 0$, and the input uncertainty in Equation (1) reduces to

$$\mathrm{Var}_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{E}_U\left(\overline{Y} \mid \widehat{\boldsymbol{\theta}}\right)\right] = \mathrm{Var}\left[\eta\left(\widehat{\boldsymbol{\theta}}\right)\right].$$

Suppose that the expected value $\eta(\boldsymbol{\theta})$ as a function of $\boldsymbol{\theta}$ is twice continuously differentiable around $\boldsymbol{\theta}^0$. Then $\eta(\widehat{\boldsymbol{\theta}})$ can be expanded as a Taylor series about the true input parameter $\boldsymbol{\theta}^0$ as

$$\eta\left(\widehat{\boldsymbol{\theta}}\right) = \eta(\boldsymbol{\theta}^0) + \mathbf{g}(\boldsymbol{\theta}^0)^\top\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right) + O_p\left[\left(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0\right)^2\right]$$

where $\mathbf{g}(\boldsymbol{\theta}) = \partial\eta(\boldsymbol{\theta})/\partial\boldsymbol{\theta}$ is the gradient of the expected value with respect to the input parameters and $O_p[(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}^0)^2]$ denotes a quantity that depends on a combination of second-order terms so that it is of the same order of magnitude. Under some regularity conditions the variance of $\eta(\widehat{\theta})$ is

$$\mathrm{Var}\left[\eta\left(\widehat{\boldsymbol{\theta}}\right)\right] = \mathbf{g}(\boldsymbol{\theta}^0)^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{g}(\boldsymbol{\theta}^0) + O\left(m^{-\frac{3}{2}}\right)$$

(Cheng and Holland, 1997, 1998).

Cheng and Holland (1997) next address the stochastic-simulation variance in Equation (1). For one replication we can write

$$\begin{aligned}\mathrm{Var}_{\mathbf{U}_j}\left[Y_j\left(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}\right) \mid \widehat{\boldsymbol{\theta}}\right] &= \mathrm{Var}_{\mathbf{U}_j}\left[\eta\left(\widehat{\boldsymbol{\theta}}\right) + e_j\left(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}\right) \mid \widehat{\boldsymbol{\theta}}\right] \\ &= \mathrm{Var}_{\mathbf{U}_j}\left[e_j\left(\mathbf{U}_j, \widehat{\boldsymbol{\theta}}\right) \mid \widehat{\boldsymbol{\theta}}\right] = \sigma^2\left(\widehat{\boldsymbol{\theta}}\right).\end{aligned}$$

The marginal variance $\sigma^2(\widehat{\boldsymbol{\theta}})$ can also be expanded as a Taylor series about $\boldsymbol{\theta}^0$ if we suppose that $\sigma^2(\widehat{\boldsymbol{\theta}})$ is twice continuously differentiable. Thus, the stochastic uncertainty can be approximated as

$$E_{\widehat{\boldsymbol{\theta}}}\left[\mathrm{Var}_U\left(\overline{Y} \mid \widehat{\boldsymbol{\theta}}\right)\right] = \frac{\sigma^2(\boldsymbol{\theta}^0)}{n} + O\left((nm)^{-1}\right),$$

where n is the number of replications. Therefore, the total variance of the mean performance estimator in Equation (1) becomes

$$\begin{aligned}\mathrm{Var}\left[\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right)\right] &= \mathbf{g}(\boldsymbol{\theta}^0)^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{g}(\boldsymbol{\theta}^0) + \frac{\sigma^2(\boldsymbol{\theta}^0)}{n} \\ &\quad + O\left(m^{-\frac{3}{2}}\right) + O\left((nm)^{-1}\right). \quad (2)\end{aligned}$$

If m is not too small, then Cheng and Holland suggest using the first two terms of Equation (2) to approximate the overall variance (input and simulation). However, to actually employ this approximation we need estimators of the gradient $\mathbf{g}(\boldsymbol{\theta}^0)$, the variance of the input parameter estimates $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$, and the stochastic variance $\sigma^2(\boldsymbol{\theta}^0)$. We address estimation, and our contribution to it, in the next section.

Although not noted by Cheng and Holland, the variance expression (2) also yields approximate contributions of each independent input model to the input-uncertainty variance. This is valuable when input uncertainty is large, because distributions with the largest contributions are the ones from which it would be most useful to collect additional real-world data, if feasible.

To see this, let $q \leqslant p$ be the number of independent input models, and let $\boldsymbol{\theta}_j$ be the parameter vector of the jth model; $\boldsymbol{\theta}_j$ may be a scalar (eg, exponential distribution) or vector (eg, lognormal distribution). Then we can write $\mathbf{g}(\boldsymbol{\theta}^0)^\top = \left(\mathbf{g}(\boldsymbol{\theta}_1^0)^\top, \mathbf{g}(\boldsymbol{\theta}_2^0)^\top,\ \ldots\ , \mathbf{g}(\boldsymbol{\theta}_q^0)^\top\right)$, where $\mathbf{g}(\boldsymbol{\theta}_j^0)$ are the terms in the gradient $\mathbf{g}(\boldsymbol{\theta}^0)$ obtained from partial derivatives with respect

[1] MLEs are not required for Cheng and Holland's method, only that the variance-covariance matrix of the parameter estimates can be estimated; this is most easily done for MLEs.

to the parameters in $\boldsymbol{\theta}_j$. Then $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ has a block diagonal form

$$\mathrm{Var}(\widehat{\boldsymbol{\theta}}) = \begin{pmatrix} \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_1\right) & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_2\right) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_q\right) \end{pmatrix},$$

where $\mathrm{Var}(\widehat{\boldsymbol{\theta}}_j)$ is the variance-covariance matrix of $\widehat{\boldsymbol{\theta}}_j$. Therefore,

$$\mathbf{g}\left(\boldsymbol{\theta}^0\right)^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{g}\left(\boldsymbol{\theta}^0\right) = \sum_{j=1}^{q} \mathbf{g}\left(\boldsymbol{\theta}_j^0\right)^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}_j\right)\mathbf{g}\left(\boldsymbol{\theta}_j^0\right) \quad (3)$$

so that $\mathbf{g}(\boldsymbol{\theta}_j^0)^\top \mathrm{Var}(\widehat{\boldsymbol{\theta}}_j)\mathbf{g}(\boldsymbol{\theta}_j^0)$ is the variance contribution of the *j*th input model.

Remarks: Song and Nelson (2013, 2015) defined the contribution to input uncertainty of the *j*th input model to be

$$V_j \equiv \mathrm{Var}\left[\mathrm{E}\left(Y\left(\boldsymbol{\theta}_1^0, \boldsymbol{\theta}_2^0, \ldots, \boldsymbol{\theta}_{j-1}^0, \widehat{\boldsymbol{\theta}}_j, \boldsymbol{\theta}_{j+1}^0, \ldots, \boldsymbol{\theta}_q^0\right) \mid \widehat{\boldsymbol{\theta}}_j\right)\right]$$

the variance of the simulation's expected response when all of the true input parameters except $\boldsymbol{\theta}_j^0$ are known. Song and Nelson (2014) show that $\mathbf{g}(\boldsymbol{\theta}_j^0)^\top \mathrm{Var}(\widehat{\boldsymbol{\theta}}_j)\mathbf{g}(\boldsymbol{\theta}_j^0)$ can be interpreted as V_j if the first-order Taylor approximation is exact.

3. Estimating input uncertainty

Our goal now is to deploy Cheng and Holland's approximation

$$\mathrm{Var}\left[\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right)\right] \approx \mathbf{g}\left(\boldsymbol{\theta}^0\right)^\top \mathrm{Var}\left(\widehat{\boldsymbol{\theta}}\right)\mathbf{g}\left(\boldsymbol{\theta}^0\right) + \frac{\sigma^2\left(\boldsymbol{\theta}^0\right)}{n}$$

by providing estimators for each term on the right-hand side. For $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ and $\sigma^2(\boldsymbol{\theta}^0)$ we follow the suggestions of Cheng and Holland (1997, 1998):

- The variance $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$ is estimated by $\widehat{\mathrm{Var}(\widehat{\boldsymbol{\theta}})} = \mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}})$, the Fisher information matrix of the MLEs evaluated at $\widehat{\boldsymbol{\theta}}$. This is justified by the fact that the MLEs are asymptotically unbiased for $\boldsymbol{\theta}^0$ and follow a multivariate normal distribution for large $m: \widehat{\boldsymbol{\theta}} \sim N_p(\boldsymbol{\theta}^0, \mathrm{Var}(\boldsymbol{\theta}^0))$ where the variance-covariance matrix $\mathrm{Var}(\boldsymbol{\theta}^0)$ is a function of the Fisher information matrix $\mathbf{I}(\boldsymbol{\theta}^0)$. That is, $\mathrm{Var}(\boldsymbol{\theta}^0) = \mathbf{I}^{-1}(\boldsymbol{\theta}^0)$, with
$$\mathbf{I}^{-1}\left(\boldsymbol{\theta}^0\right) = \mathrm{E}\left[-\frac{\partial^2 L(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}^2}\right]\Bigg|_{\boldsymbol{\theta}=\boldsymbol{\theta}^0}$$
and $L(\boldsymbol{\theta})$ is the log-likelihood function of the input models. Therefore, an estimator of $\mathrm{Var}(\boldsymbol{\theta}^0)$ can be employed as an estimator for $\mathrm{Var}(\widehat{\boldsymbol{\theta}})$. These results are well known and expressions for $\mathrm{Var}(\boldsymbol{\theta}^0)$ are available for all standard distribution choices. The $\mathrm{Var}(\boldsymbol{\theta}^0)$ decreases as $O(m^{-1})$ in the size of the real-world data sample *m*.
- The stochastic uncertainty $\sigma^2(\boldsymbol{\theta}^0)$ is estimated by the sample variance S^2 of the *n* simulation outputs $Y_1(\mathbf{U}_1, \widehat{\boldsymbol{\theta}}), Y_2(\mathbf{U}_2, \widehat{\boldsymbol{\theta}}), \ldots, Y_n(\mathbf{U}_n, \widehat{\boldsymbol{\theta}})$, simulating at the MLEs instead of $\boldsymbol{\theta}^0$. Notice that this is the only variance that is typically estimated in simulation output analysis, which means input uncertainty is ignored and the overall error is underestimated.

Our point of departure from Cheng and Holland (1997) is in estimating the gradient $\mathrm{g}(\boldsymbol{\theta}^0)$. They suggest using finite forward differences from $\widehat{\boldsymbol{\theta}}$; notice that employing finite differences necessitate $p+1$ simulations, p more than the nominal experiment, so the burden becomes greater the more input models there are in the simulation. In addition, a finite difference must be chosen for the simulations, which requires balancing a bias-variance trade off.

To make it computationally inexpensive to estimate the impact of input uncertainty, we want a robust, easy-to-use, internal estimator of the gradient obtained from the nominal experiment. Because $\boldsymbol{\theta}^0$ is unknown, we will estimate $\mathbf{g}(\widehat{\boldsymbol{\theta}})$, rather than $\mathbf{g}(\boldsymbol{\theta}^0)$, just as Cheng and Holland do. The key observation is that $\mathbf{g}(\widehat{\boldsymbol{\theta}})$ is the gradient with respect to simulation *input-distribution parameters*, rather than some structural aspect of the simulated system. This allows us to extend and exploit a gradient-estimation method due to Wieland and Schmeiser (2006) that is particularly well suited to such parameters.

Wieland and Schmeiser's derivative estimator comes from a clever observation that we specialize to our setting. Suppose that there is a single input parameter θ, and to execute the simulation we set its value to $\widehat{\theta}$. Then, from the point of view of the simulation, $\widehat{\theta}$ is the true, nominal value of the parameter. To be concrete, suppose that θ is the mean time to failure of an exponential distribution describing the failure time of a replaceable component.

During the course of a replication, one or more random variates are generated from this input distribution with parameter $\widehat{\theta}$; in the case of the example, times until failure are generated from an exponential distribution with mean $\widehat{\theta}$ within the replication. Denote these values by $X_{j,1}, X_{j,2}, \ldots, X_{j,m_j}$ for the *j*th replication. Although seemingly unnecessary, we could use $X_{j,1}, X_{j,2}, \ldots, X_{j,m_j}$ to estimate $\widehat{\theta}$; call the estimator $\overline{\theta}_j$. In the example, $\overline{\theta}_j = \sum_{\ell=1}^{m_j} X_{j,\ell}/m_j$, the average time to failure for the component during the *j*th replication; it could be thought of as the MLE based on generated inputs from within the *j*th replication. After *n* replications of the simulation we will have observed independent and identically distributed (i.i.d.) pairs $(Y_j, \overline{\theta}_j)$, $j = 1, 2, \ldots, n$.

If the simulation output *Y* depends on the input model—as it almost certainly does—then we would expect $(Y_j, \overline{\theta}_j)$ to be dependent. Suppose their joint distribution is bivariate normal, we will describe ways to make this a good approximation below. Then

$$\mathrm{E}\left[Y\left(\widehat{\theta}\right) \mid \overline{\theta}\right] = \eta\left(\widehat{\theta}\right) + \Sigma_{Y\overline{\theta}}\Sigma_{\overline{\theta}\overline{\theta}}^{-1}\left(\overline{\theta} - \mu_{\overline{\theta}}\right) = \beta_0 + \beta_1\overline{\theta}, \quad (4)$$

where $\Sigma_{Y\overline{\theta}}$ is the covariance between Y and $\overline{\theta}$; $\Sigma_{\overline{\theta}\,\overline{\theta}}$ is the variance of $\overline{\theta}$; and $\mu_{\overline{\theta}}$ is the expected value of $\overline{\theta}$. Notice that if $\overline{\theta}$ is an unbiased estimator then $\mu_{\overline{\theta}} = \widehat{\theta}$, or if at least a consistent estimator then $\mu_{\overline{\theta}} \approx \widehat{\theta}$. Clearly the derivative of the expected response with respect to θ evaluated at $\widehat{\theta}$ is $\beta_1 = \Sigma_{Y\overline{\theta}}\Sigma_{\overline{\theta}\,\overline{\theta}}^{-1}$, which can easily be estimated via least-squares regression.

The intuition behind the method is simple: to estimate a derivative using finite differences, we run the simulation at two *fixed* settings of θ, say $\widehat{\theta}$ and $\widehat{\theta}+\delta$, and use the scaled difference in the observed responses to estimate the derivative. Wieland and Schmeiser, on the other hand, fix the value of θ, say at $\widehat{\theta}$, and then estimate the sensitivity of the response Y_j to the *realized* parameter $\overline{\theta}_j$ as it varies across n replications. In the bivariate normal case this relationship is linear, and therefore this gives the derivative at $\widehat{\theta}$.

Wieland and Schmeiser (2006) leave open the question of how to estimate a gradient with respect to a vector parameter, in particular whether to do individual regressions for each θ_i, or to do a multivariate regression. In the Appendix we establish that a multivariate regression is the correct approach. This is based on the natural extension of (4) when $(Y, \overline{\boldsymbol{\theta}}^\top)$ have a multivariate normal distribution, which implies that

$$\mathrm{E}\left[Y_j\left(\widehat{\boldsymbol{\theta}}\right) \mid \overline{\boldsymbol{\theta}}_j\right] = \eta\left(\widehat{\boldsymbol{\theta}}\right) + \Sigma_{Y\overline{\boldsymbol{\theta}}}\Sigma_{\overline{\boldsymbol{\theta}}\,\overline{\boldsymbol{\theta}}}^{-1}\left(\overline{\boldsymbol{\theta}}_j - \mu_{\overline{\boldsymbol{\theta}}}\right) = \beta_0 + \boldsymbol{\beta}_1^\top \overline{\boldsymbol{\theta}}_j, \quad (5)$$

where $\overline{\boldsymbol{\theta}}_j = (\overline{\theta}_{1j}, \overline{\theta}_{2j}, \ \ldots \ , \overline{\theta}_{pj})^\top$ are consistent estimators of $\widehat{\boldsymbol{\theta}}$ from replication j. This gives the gradient estimator $\widehat{\mathbf{g}}(\widehat{\boldsymbol{\theta}}) = \widehat{\boldsymbol{\beta}}_1$, where $\widehat{\boldsymbol{\beta}}_1$ is obtained from a least-squares regression of $Y_j(\widehat{\boldsymbol{\theta}})$ on $(1, \overline{\theta}_{1j}, \overline{\theta}_{2j}, \ \ldots \ , \overline{\theta}_{pj})$, $j = 1, 2, \ \ldots \ , n$,

When might we expect multivariate normality to hold? In many cases Y_j will itself be the average of a large number of elementary outputs within replication j, for instance the average of hundreds of individual customer waiting times during the jth replicated day. Similarly, $\overline{\boldsymbol{\theta}}_j$ will often be functions of large numbers of inputs generated within the jth replication, for instance customer interarrival and service times. In such situations approximate multivariate normality is plausible.

When this is not the case, then Wieland and Schmeiser (2006) suggest batching the replications. That is, instead of regressing Y_j on $(1, \overline{\theta}_{1j}, \overline{\theta}_{2j}, \ \ldots \ , \overline{\theta}_{pj}), j = 1, 2, \ \ldots \ , n$, regress $\overline{Y}_j(b)$ on $(1, \overline{\theta}_{1j}(b), \overline{\theta}_{2j}(b), \ldots , \overline{\theta}_{pj}(b))$, $j = 1, 2, \ldots , k$, where $k = \lfloor n/b \rfloor$, and

$$\overline{Y}_j(b) = \frac{1}{b}\sum_{\ell=(j-1)b+1}^{jb} Y_\ell$$

with $\overline{\theta}_{ij}(b)$ defined similarly. The larger b is the closer the approximation to multivariate normality should be due to the multivariate Central Limit Theorem. However, large b leaves fewer observations k for the regression, so there is a trade off that we analyse in the next section.

Assembling all of the pieces, our estimator of the variance of $\overline{Y}(\widehat{\boldsymbol{\theta}})$ is

$$\widehat{\mathrm{Var}}\left[\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right)\right] = \widehat{\boldsymbol{\beta}}_1^\top\,\widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\theta}}\right)\widehat{\boldsymbol{\beta}}_1 + \frac{S^2}{n}$$

which also provides estimators of the variance contribution of each input model, as shown in (3). Only input and output values generated in the nominal simulation are needed to form this estimator, and the same recorded input values can be used to estimate the input-uncertainty variance for any of the simulation output measures of interest.

Remarks: Notice that multivariate normality is a sufficient condition for (5) to hold, but not necessary. The key condition is that the expected response is locally linear near $\widehat{\boldsymbol{\theta}}$ which can be true without normality and justifies the use of linear regression. Nevertheless, tests for departures from multivariate normality such as those available at http://cran.r-project.org/web/packages/ could be applied as a check, and we do so in the numerical example in Section 5.2. Gradient estimators other than finite differences or Wieland and Schmeiser's could also be employed with the Cheng and Holland approximation, including perturbation analysis, likelihood ratio method, and weak derivatives; see, for instance Fu (2015, Chapter 5). All of these methods require collecting additional information during the simulation run, and the conditions for their validity can be difficult to verify.

4. Experiment design

Multivariate normality is a sufficient condition for the linear relationship in (5) to hold, and therefore satisfying multivariate normality provides a strong justification for our gradient estimator. Batching of the input and output data can improve this approximation, but at the cost of degrees of freedom in the regression to estimate $\boldsymbol{\beta}_1$. In this section we evaluate the tradeoff and provide guidelines for batching.

We take the approach pioneered in Schmeiser (1982) by asking the batch-size question in the following way: Suppose that multivariate normality was achieved using the n original replications (ie, no batching is needed), but we nevertheless batched the data into k batches of size $b = n/k$. What is the penalty in regressing on only k observations rather than n? If there is little penalty as long as $k \geqslant k^*$, say, then there is no reason to try to work with more than k^* batches, and a good reason—to better approximate multivariate normality with larger b—to try to use around k^* batches.

Employing standard multivariate analysis, we can show that $\widehat{\beta}_1$ is unbiased, and

$$\mathrm{Var}\left[\widehat{\boldsymbol{\beta}}_1\right] = \frac{1}{k-p-2}\mathbf{C} \equiv \mathrm{Var}(k, p) \quad (6)$$

where $k = n/b$ and $\mathbf{C}$ is a $p \times p$ matrix that does not depend on the number of batches k. For convenience in presentation we are assuming b divides n evenly.

Clearly $k = n$ (batch size $b = 1$) minimizes the variance under these assumptions, but in reality the multivariate normality assumption is less likely to be true the larger k is. Suppose we increase the number of batches from k to $k' = k + 1$. Then the percentage variance reduction is

$$\begin{aligned}\Delta(k,p) &= \frac{\mathrm{Var}(k,p) - \mathrm{Var}(k',p)}{\mathrm{Var}(k,p)}\\ &= \frac{\frac{1}{k-p-2}\mathbf{C} - \frac{1}{k-p-1}\mathbf{C}}{\frac{1}{k-p-2}\mathbf{C}}\\ &= \frac{1}{k-p-1},\ k \geqslant p+3.\end{aligned}$$

When the number of input parameters p is fixed, the percentage change is simply a function of k. Figure 1 shows $\Delta(k, p)$ as k increases for various values of p.

Consider the case where $p = 1$. We see that when k changes from 4 to 5, $\mathrm{Var}(\widehat{\beta}_1)$ decreases by 50%; and $\mathrm{Var}(\widehat{\beta}_1)$ continues to decrease by 33.3% when k changes from 5 to 6, which is smaller but still significant. When k becomes large, for example $k > 22$, then $\mathrm{Var}(\widehat{\beta}_1)$ decreases by less than 5%, which implies that having one more batch leads to a much less significant reduction in $\mathrm{Var}(\widehat{\beta}_1)$ when k is already over 20. On the other hand, if k is small, for example $k < 12$, then the decrease of $\mathrm{Var}(\widehat{\beta}_1)$ would be greater than 10%. Generally, changing k from $p+3$ to $p+4$ results in the maximum decrease in $\mathrm{Var}(\widehat{\beta}_1)$. When $k > p+21$, the decrease is less than 5%. Practitioners can choose their own trade off, but based on this analysis $k^* \approx p + 21$ (but dividing n evenly) is the fewest number of batches we would want to use.

Our setting in this paper assumes n i.i.d. replications of the nominal simulation experiment. If the goal of the simulation is to estimate the steady-state performance of an ergodic simulation output process, then it is sometimes prudent to use an experiment design consisting of one (long) replication. Batching *within* this single replication is often recommended to compute measures of error for steady-state performance estimates; see, for instance, Nelson (2013, Chapter 8). In this scenario the run length plays the role of the number of replications, and our method for estimating input-uncertainty variance can be based on batch statistics, rather than replication statistics. The batching guidelines provided above are even more valuable in this setting because both normality and independence are questionable without batching.

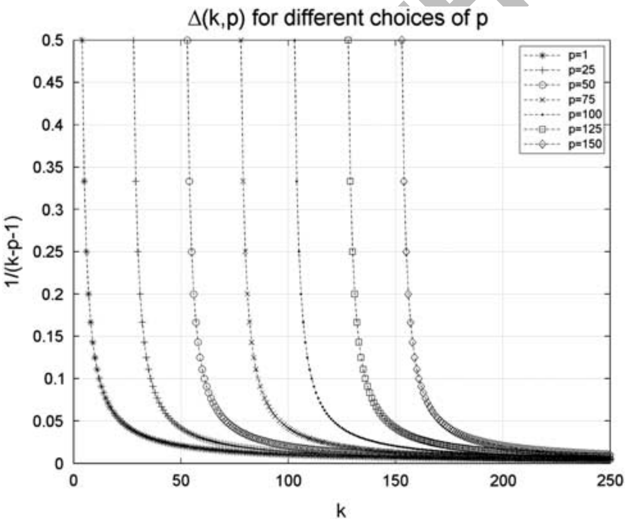

Figure 1 Effect of batching on the variance of the gradient estimator.

Remarks: The discussion in this section concerns what number of batches to use when n (or the run length for steady-state simulation) is fixed, and assumes n is large enough so that there is a number of batches (and corresponding batch size) for which the approximation of multivariate normality holds. It should not be misinterpreted to mean that only a small number of replications is needed. Larger n is always better: All else being the same, $k = 20$ batches of size $b = 100$ will be superior to $k = 20$ batches of size $b = 2$.

5. Empirical evaluation

In this section we illustrate the proposed method using two examples, comparing the results either to the true values, Cheng and Holland's method or a side experiment constructed to estimate input uncertainty.

5.1. Steady-state M/M/∞ queue

Consider an *M*/*M*/∞ queue with arrival rate λ and mean service time τ; thus the input parameters are $\boldsymbol{\theta} = (\lambda, \tau)^{\top}$. Suppose we construct a simulation experiment with output $\{Q_t(\boldsymbol{\theta});\ 0 \leqslant t \leqslant T\}$, where Q_t is the number of customers in the system at time t and T is the run length of one replication. Then $Q_t(\boldsymbol{\theta}) \Rightarrow Q(\boldsymbol{\theta})$ as $t \to \infty$, where $Q(\boldsymbol{\theta}) \sim \mathrm{Poisson}(\lambda\tau)$.

Our goal is to estimate the steady-state mean number of customers in the system, $\eta(\boldsymbol{\theta}^0) = \lambda^0\tau^0$ at the true real-world parameters λ^0 and τ^0. These parameters are estimated from observing m interarrival times and m service times which are known to be exponentially distributed, but with unknown parameters. Let $\widehat{\boldsymbol{\theta}} = (\widehat{\lambda}, \widehat{\tau})^{\top}$ be the MLEs.

The point estimator from one replication is

$$Y\left(\widehat{\boldsymbol{\theta}}\right) = \frac{1}{T}\int_0^T Q_t\left(\widehat{\boldsymbol{\theta}}\right)\mathrm{d}t.$$

If the run length T is large enough, then we can show that

$$\begin{aligned}\mathrm{Var}\left[Y\left(\widehat{\boldsymbol{\theta}}\right)\right] &= \mathrm{Var}\left\{\mathrm{E}\left[Y\left(\widehat{\boldsymbol{\theta}}\right) \mid \widehat{\boldsymbol{\theta}}\right]\right\}\\ &\quad + \mathrm{E}\left\{\mathrm{Var}\left[Y\left(\widehat{\boldsymbol{\theta}}\right) \mid \widehat{\boldsymbol{\theta}}\right]\right\}\\ &\approx\ \mathrm{Var}\left(\widehat{\lambda\tau}\right) + \mathrm{E}\left(\frac{2\widehat{\lambda\tau^2}}{T}\right)\end{aligned}$$

where the term $2\widehat{\lambda}\widehat{\tau}^2$ is the asymptotic variance of the sample mean conditional on $\widehat{\lambda}$ and $\widehat{\tau}$ (Whitt, 2006). If the simulation experiment involves n replications, then the variance of the overall sample mean $\overline{Y}(\widehat{\boldsymbol{\theta}})$ is

$$\mathrm{Var}\left[\overline{Y}\left(\widehat{\boldsymbol{\theta}}\right)\right] \approx \mathrm{Var}\left(\widehat{\lambda}\widehat{\tau}\right) + \mathrm{E}\left(\frac{2\widehat{\lambda}\widehat{\tau}^2}{nT}\right).$$

Nelson (2013) gives the distributions of $\widehat{\lambda}$ and $\widehat{\tau}$, which can be used to show that

$$\begin{aligned}\mathrm{Var}\left[Y\left(\widehat{\boldsymbol{\theta}}\right)\right] &= \frac{m(2m-1)}{(m-1)^2(m-2)}\left(\lambda^0\tau^0\right)^2 \\ &\quad + \frac{2(m+1)}{n(m-1)T}\left(\lambda^0\left(\tau^0\right)^2\right) \\ &\approx \frac{2(\lambda^0\tau^0)^2}{m} + \frac{2\lambda^0(\tau^0)^2}{nT}.\end{aligned}$$

Thus, for this example we can calculate the true values of the input-uncertainty and stochastic-simulation variances given knowledge of $\boldsymbol{\theta}^0$. Moreover, since $\eta(\boldsymbol{\theta}) = \lambda\tau$, the gradient of the steady-state expected number of customers in the system with respect to $\boldsymbol{\theta} = (\lambda, \tau)^\top$ is also known: $\mathbf{g}(\boldsymbol{\theta}^0) = (\tau^0, \lambda^0)^\top$. These results allow us to evaluate the performance of our method against the truth, as well as against Cheng and Holland's method.

Our evaluation consists of the following experiment; specific parameter values are shown in Table 1. On each of R macro-replications we sample m interarrival times with rate λ^0, m service times with mean τ^0, calculate the MLEs $\widehat{\lambda}$ and $\widehat{\tau}$, and simulate n replications of length T. From this data we estimate the $\mathrm{Var}[\overline{Y}(\widehat{\boldsymbol{\theta}})]$ using our method and Cheng and Holland. For Cheng and Holland we use the finite-forward-difference method to estimate the gradient, necessitating two additional sets of simulation runs ($3n$ replications in total). The comparison between the two methods, averaged across the $R = 100$ macro-replications, and the true values of the gradients and variances are shown in the Table 2. By 'stochastic variance' we mean the variance of the sample mean without considering input uncertainty, denoted by $\sigma^2(\widehat{\boldsymbol{\theta}})/n$ in our framework; the square root of this is what is typically reported as the 'standard error' of $\overline{Y}(\widehat{\boldsymbol{\theta}})$. Accounting for input uncertainty, the standard error should be the square root of Input Variance + Stochastic Variance.

Table 1 Simulation experiment design for *M*/*M*/∞ queue

R	n	T	m	λ^0	τ^0
100	100	1000	1000	2	10

Table 2 shows that applying our method gives results comparable to Cheng and Holland with one-third the simulation effort. Notice that the stochastic-simulation variances obtained from the two methods are the same, as they should be, because our proposed method only changes the way we estimate the input-uncertainty variance. In Table 3 we gave our method the same budget as Cheng and Holland ($n = 300$ replications) to illustrate that we obtain standard errors and relative errors of the same order of magnitude (the relative error is the ratio of the standard error to mean). The finite-difference estimator of Cheng and Holland is particularly well suited for this example because the partial derivatives are, in fact, linear, so there is no bias and the size of the finite difference is less critical. This explains the somewhat smaller standard errors. Of course, our stochastic variance is smaller because we expend all 300 replications on the nominal experiment, which is an advantage. Both methods accurately estimate the gradients and variances, as seen by comparing them to the true values in the last column.

When we introduced Wieland and Schmeiser's gradient estimation method we noted the potential importance of batch statistics. In our *M*/*M*/∞ experiment the run length was $T = 1000$ time units, which allowed approximately 4000 arrivals and departures to occur during each replication. Thus, it is reasonable to assume that the sample mean response $Y(\widehat{\boldsymbol{\theta}})$ and consistent estimators of the input parameters $\overline{\boldsymbol{\theta}} = (\overline{\lambda}, \overline{\tau})^\top$ obtained from each replication already have an approximately multivariate normal distribution; this implies that batch statistics may not improve the gradient estimator for this example since $k = n$ is optimal. Nevertheless, we also ran the analysis under different numbers of batches k and batch size b, where $kb = n$ is fixed. The second column in Table 4 gives the average input-uncertainty variance over the $R = 100$ macro-replications, and the third column gives the corresponding standard error of this estimate.

Not surprisingly, batching made the input-uncertainty variance and the gradient estimates more variable as we decreased k and maintained $kb = 100$; this can be seen in the increase in the standard errors. For this example, $k = n$ is clearly the best choice, but batching did not impose too much of a penalty until k became quite small.

Table 2 Comparison between simulation results and true values for the *M*/*M*/∞ queue

	Single-run method			*Cheng & Holland's method*			*True*
	Mean	*Std. err.*	*Rel. err.*	*Mean*	*Std. err.*	*Rel. err.*	
β_λ	9.8806	0.1389	0.0141	9.8659	0.0139	0.0014	10
β_τ	1.9632	0.0238	0.0121	2.0947	0.0115	0.0055	2
Input var.	0.7732	0.0136	0.0175	0.8109	0.0046	0.0056	0.8000
Stochastic var.	0.0038	4.7×10^{-5}	0.0124	0.0038	4.7×10^{-5}	0.0124	0.0040

Table 3 Simulation results for the single-run method for the *M*/*M*/∞ queue given effort equal to Cheng and Holland ($n = 300$)

	Single-run method			*True*
	Mean	*Std. err.*	*Rel. err.*	
β_λ	9.6166	0.0765	0.0080	10
β_τ	1.9663	0.0165	0.0084	2
Input var.	0.7454	0.0085	0.0114	0.8000
Stochastic var.	0.0012	1.0×10^{-5}	0.0083	0.0013

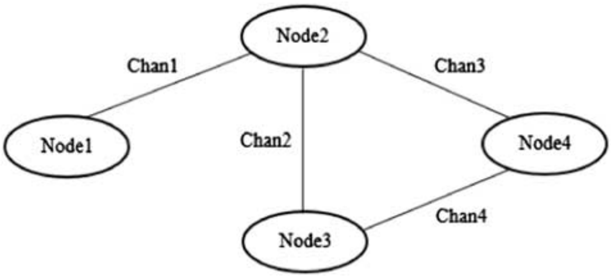

Figure 2 A computer communication network with $N = M = 4$.

Table 4 Simulation results using batch statistics

(k, b)	*Input var.*	*Std. err.*	β_λ	*Std. err.* (β_λ)	β_τ	*Std. err.* (β_τ)
(100, 1)	0.7723	0.0136	9.8806	0.1389	1.9632	0.0238
(50, 2)	0.7627	0.0214	9.7069	0.2116	1.9192	0.0393
(20, 5)	0.7927	0.0352	9.9716	0.3498	1.7805	0.0661
(10, 10)	0.9332	0.0538	9.8759	0.4897	1.9372	0.0963

In this *M*/*M*/∞ example, the gradients and input uncertainty estimated from the proposed method are close to the true values. Moreover, the recorded simulation time for Cheng and Holland's method was more than twice the time of the proposed method if we give them the same budget for the nominal experiment. This is only a two-dimensional problem, so if the dimension p increases, the simulation effort required by Cheng and Holland's method will increase linearly with respect to p, but the proposed method will still only need essentially the same amount of effort because all the gradient estimators can be obtained in a single set of replications.

5.2. Computer communication network

The second example is a computer communication network first considered by Kleinrock (1976): it is a system with N message processing centres connected by M channels that can be described by an undirected graph with N nodes (centres) and M edges (channels). The M channels have a capacity of C_ℓ bits/second and a length l_ℓ for the ℓth channel, which implies that there are channel queueing and transmission delays. External messages arriving to node h that are to be transmitted to node k follow a Poisson arrival process with rate γ_{hk} messages/second. The messages have lengths described by an exponential distribution with mean δ bits. The ℓth node takes K_ℓ seconds to process one message, and the nodes have unlimited storage capacity. Each message type is transmitted on a fixed path. All messages transmit through channels with a velocity v, so the propagation time P_ℓ required for the lth channel is $P_\ell = l_\ell / v$. Thus, a message with δ bits will occupy the lth channel for $P_\ell + \delta / C_\ell$ seconds.

Cheng and Holland (1997) used a small-scale version of this communication network to illustrate their method; it had four nodes connected by four links, and is shown in Figure 2. We use the same model to estimate the expected delay for messages transmitted between the nodes via the links and to apply our proposed method to analyse the impact of input uncertainty, both overall and the contributions of each input model. However, unlike the *M*/*M*/∞ example, we do not compare the proposed method to Cheng and Holland's results because the true input-uncertainty variance and parameter gradients are not available; thus, there is no way to know if Cheng and Holland's results are correct. Instead, we conduct an extensive side experiment using a random-effects model to estimate the input-uncertainty variance as a basis for evaluating our method.

The parameters of the communication network are shown in Table 5. In this example there are 13 input parameters including the true mean message length $\delta^0 = 300$ bits, and the 12 true message arrival rates γ^0_{hk} messages/second, $h, k = 1, 2, 3, 4, h \neq k$, shown in Table 6. Therefore, the true input parameter vector is $\boldsymbol{\theta}^0 = (\delta^0, \gamma^0_{12}, \gamma^0_{13}, \dots, \gamma^0_{34})^\top$. To simplify the experiment we set the number of real-world observations from each distribution to a common value of m.

To execute an experiment, we first generate a sample of 'real-world' input data from these distributions and estimate $\boldsymbol{\theta}^0$ using the MLEs $\widehat{\boldsymbol{\theta}}$. The simulation is run with $\widehat{\boldsymbol{\theta}}$. To estimate the input parameter uncertainty using our method, we observe within each of n replications the average transmission delay $Y(\boldsymbol{\theta})$, along with the sample-average message length, $\overline{\delta}$, and the sample message arrival rates, $\overline{\gamma}_{hk}$, which we obtain from the average interarrival times for each type of message as $\overline{\gamma}_{hk} = 1/\overline{X}_{hk}$. Here $\overline{X}_{hk}$ is the average of all interarrival times of messages arriving at node h that are destined for node k during the replication. Thus, the following simulation input and output data are recorded from the nominal experiment:

$$\left(Y_1\left(\widehat{\boldsymbol{\theta}}\right), \overline{\delta}_1, \frac{1}{\overline{X}_{12,1}}, \dots, \frac{1}{\overline{X}_{hk,1}}, \dots, \frac{1}{\overline{X}_{43,1}}\right)^\top,$$

$$\left(Y_2\left(\widehat{\boldsymbol{\theta}}\right), \overline{\delta}_2, \frac{1}{\overline{X}_{12,2}}, \dots, \frac{1}{\overline{X}_{hk,2}}, \dots, \frac{1}{\overline{X}_{43,2}}\right)^\top,$$

$$\vdots$$

$$\left(Y_n\left(\widehat{\boldsymbol{\theta}}\right), \overline{\delta}_n, \frac{1}{\overline{X}_{12,n}}, \dots, \frac{1}{\overline{X}_{hk,n}}, \dots, \frac{1}{\overline{X}_{43,n}}\right)^\top,$$

Table 5 Parameters of the small-scale communication network

K_ℓ *(s)*	C_ℓ *(bits/s)*	l_ℓ *(miles)*	*v (miles/s)*
0.001	275 000	$\ell \times 100$	150 000

Table 6 True arrival rates of the external messages from node h to node k, γ_{hk}^0

Starting node h	*Ending node k*			
	1	2	3	4
1	—	60	40	50
2	80	—	65	20
3	100	22	—	26
4	40	50	60	—

Table 7 Comparison of the single-run method and the random-effects model for the communications network with $m=50$

Single-run method	*Input variance*	*Stochastic variance*
Average	$\mathbf{2.5672 \times 10^{-6}}$	$\mathbf{1.7881 \times 10^{-10}}$
Standard error	2.5692×10^{-7}	1.8908×10^{-11}
Random-effects method	$\mathbf{1.7685 \times 10^{-6}}$	$\mathbf{1.7881 \times 10^{-10}}$
99% confidence interval	$[1.6 \times 10^{-6}, 2.0 \times 10^{-6}]$	

Note: The primary values to compare are highlighted in boldface type.

where

$$\left(Y_j\left(\widehat{\boldsymbol{\theta}}\right), \bar{\delta}_j, \bar{\gamma}_{12,j} = \frac{1}{\bar{X}_{12,j}}, \ \ldots \ , \bar{\gamma}_{hk,j} = \frac{1}{\bar{X}_{hk,j}}, \ \ldots \ , \bar{\gamma}_{43,j} = \frac{1}{\bar{X}_{43,j}}\right)^\top$$

are the sample average delay and the consistent estimators of $\widehat{\boldsymbol{\theta}} = (\widehat{\delta}, \widehat{\gamma}_{12}, \ \ldots \ , \widehat{\gamma}_{\ell k}, \ \ldots \ , \widehat{\gamma}_{43})^\top$ from the jth replication. With these data we perform linear regression using the first-order model

$$Y_j\left(\widehat{\boldsymbol{\theta}}\right) = \beta_0 + \beta_\delta \bar{\delta}_j + \sum_{h,k=1;h \neq k}^{4} \beta_{\gamma_{hk}} \bar{\gamma}_{hk,j} + e_j.$$

The estimated regression coefficient $\widehat{\boldsymbol{\beta}}_1 = (\widehat{\beta}_\delta, \widehat{\beta}_{\gamma_{12}}, \ \ldots \ , \widehat{\beta}_{\gamma_{43}})^\top$ is the gradient estimator of the input parameters. The input-uncertainty variance is then approximated by

$$\widehat{\mathrm{Var}}\left[\eta\left(\widehat{\boldsymbol{\theta}}\right)\right] \approx \widehat{\boldsymbol{\beta}}_1^\top \widehat{\mathrm{Var}}\left(\widehat{\boldsymbol{\theta}}\right) \widehat{\boldsymbol{\beta}}_1 = \widehat{\beta}_\delta^2 \frac{\widehat{\delta}^2}{m} + \sum_{h,k=1,h \neq k}^{4} \widehat{\beta}_{\gamma_{hk}}^2 \frac{\widehat{\gamma}_{hk}^2}{m}. \quad (7)$$

Recall that m is the sample size of real-world data from each input process, and $\widehat{\mathrm{Var}}(\widehat{\boldsymbol{\theta}})$ is calculated from the Fisher information matrix $\mathbf{I}^{-1}(\widehat{\boldsymbol{\theta}})$.

As in the previous example, we made R independent macro-replications of the entire experiment. On each macro-replication we generate a sample of m 'real-world' data from the specified distributions, obtain the corresponding MLEs $\widehat{\boldsymbol{\theta}}$, run n simulation replications of length T time units, apply our method, and average the results across the macro-replications. We used $R=1000$, $m=50$, $n=100$ and $T=50$.

To evaluate our method, we conducted a side experiment using the same settings and applying a random-effects model to estimate the input-uncertainty variance, as described in Nelson (2013, §7.2). As above, we generate R samples of size m of real-world data, yielding MLEs $\widehat{\boldsymbol{\theta}}_i, i = 1, 2, \ \ldots \ R$. For each MLE, we obtain n independent replications of length T from the simulation, representing the output as

$$Y_{ij} = \eta\left(\widehat{\boldsymbol{\theta}}_i\right) + \varepsilon_{ij}, \ i = 1, 2, \ \ldots \ , R \text{ and } j = 1, 2, \ \ldots \ , n. \quad (8)$$

Recall that $\eta(\boldsymbol{\theta})$ is the true mean simulation response given parameter setting $\boldsymbol{\theta}$. We let ε_{ij} be the random error term from the jth replication using MLE $\widehat{\boldsymbol{\theta}}_i$. Equation (8) is a random-effects model, and under this model the input-uncertainty variance is $\sigma_I^2 = \mathrm{Var}[\eta(\widehat{\boldsymbol{\theta}})]$; see Nelson (2013). Using standard results for random-effects models, σ_I^2 can be estimated from the difference between an estimator of the total variance of the simulation output $\widehat{\sigma}_T^2$ and the simulation variance $\widehat{\sigma}_S^2$ as follows:

$$\widehat{\sigma}_I^2 = \frac{\widehat{\sigma}_T^2 - \widehat{\sigma}_S^2}{n},$$

where

$$\widehat{\sigma}_T^2 = \frac{n}{R-1} \sum_{i=1}^{R} \left(\bar{Y}_{i\cdot} - \bar{Y}_{\cdot\cdot}\right)^2 \text{ and}$$

$$\widehat{\sigma}_S^2 = \frac{1}{R(n-1)} \sum_{i=1}^{R} \sum_{j=1}^{n} \left(Y_{ij} - \bar{Y}_{i\cdot}\right)^2.$$

Here $\bar{Y}_{i\cdot} = \sum_{j=1}^{n} Y_{ij}/n$ and $\bar{Y}_{\cdot\cdot} = \sum_{i=1}^{R} \sum_{j=1}^{n} Y_{ij}/(Rn)$. Clearly both $\widehat{\sigma}_T^2$ and $\widehat{\sigma}_S^2$ can be calculated from the simulation results. We expect that the side experiment's estimate is more accurate than our method because it uses multiple real-world samples of input data and multiple sets of simulation runs to obtain a single estimate of σ_I^2, while our method works with a single real-world sample and a single set of simulation runs.

We estimated the input-uncertainty variance using the two methods; the results are displayed in Table 7. Compared with the *M*/*M*/∞ example, that had only two input parameters, the communication network is a much higher-dimensional problem. We see that the two sources of uncertainty obtained from our method are comparable with those from the random-effects model, which implies that the our method is practicable in this 13-dimensional problem. Notice that input-uncertainty variance is two orders of magnitude larger than stochastic simulation variance. If the simulation user had considered only stochastic error, then they would have reported a standard error of $\sqrt{1.7881 \times 10^{-10}} \approx 10^{-5}$ s. However, the overall standard error is $\sqrt{2.5672\times10^{-6}+1.7881\times10^{-10}}\approx10^{-3}$ s, which is significantly greater error in this application.

Why do we not obtain a perfect match between the estimated input-uncertainty variance using our method and the random-effects model? While the random-effects model makes fewer approximations than our method, it does assume that the simulation variance is independent of the input parameters, which is not true in this example. However, the more likely explanation is the small real-world sample size of $m=50$. The validity of Cheng and Holland's variance approximation, which is based on a Taylor series expansion, is for m large. However, even with such a small sample we obtained an input uncertainty variance of around 2×10^{-6}, which is in the same ball park as the random-effects model and close enough to provide the analyst with a useful assessment of the model risks due to input uncertainty.

To test our hypothesis we ran the experiment again but with $m=500$ observations from each real-world distribution; the results are displayed in Table 8. Clearly with the larger, but still modest sample size we are obtaining a very accurate assessment of input-uncertainty variance from a single run.

Finally, as a check on the underlying linearity assumption for gradient estimation we applied two standard tests for multivariate normality to the $n=100$ vectors of observations from all 1000 macroreplications for each quantity of real-world data m; the results are displayed in Table 9. If the data were actually multivariate normal we would expect 50 of the 1000 tests to reject the hypothesis anyway. In the table we report the fraction rejected along with a 95% confidence interval for the true rejection rate. Notice that the empirical rejection rate was only slightly higher than 0.05, suggesting reasonable conformance to multivariate normality. The R package MVN at http://cran.r-project.org/web/packages/ was used to execute the tests.

Realizing now that input-uncertainty variance is substantial, we can estimate each distribution's contribution to it using the breakdown in (7). For instance, the input-uncertainty variance contributed by having to estimate the mean message length δ^0 is $\widehat{\beta}^2_\delta\widehat{\delta}^2/m$, and the input-uncertainty variance contributed by the message-arrival distribution with rate γ^0_{hk} is $\widehat{\beta}^2_{\gamma_{hk}}\widehat{\gamma}^2_{hk}/m$. Figure 3 displays the relative contributions (normalized so that they sum to 1) averaged over the 1000 macro-replications for the $m=50$ case. The relative errors for all the contribution estimates vary from 0.27 to 5.23%.

From the figure we see that the length of transmitted messages had the most significant contribution of approximately 82% of the total input-uncertainty variance. The transmission delay is defined as P_ℓ + (message length)/C_ℓ, where both P_ℓ and C_ℓ are constants in this example, so once a message starts to be transmitted the variability in the transmission delay is entirely due to the message length. Recall that $m=50$ observations were collected to estimate the true mean message length, δ^0. If further data collection was possible, then Figure 3 indicates that the greatest reduction in uncertainty would be from observing the lengths of additional messages.

The other sources of input uncertainty are the estimated channel arrival rates which affect the queueing at each node. Among the 12 types of messages, the one with the largest arrival rate γ^0_{31} is the second leading contributor to input-uncertainty variance of approximately 9%. The remaining 9% of input-uncertainty variance was contributed by all the other input models.

The advantage of our method in terms of simulation effort saving in this example, relative to Cheng and Holland, is significant: to apply Cheng and Holland's method we would have to execute 13 additional diagnostic experiments as well as selecting 13 finite difference values. By contrast, applying the proposed method gave the overall and individual contributions to input-uncertainty variance from just the nominal experiment. More importantly, the ease of use of the proposed method makes it applicable for real-world simulation problems. Cheng and Holland (1998, 2004) also describe a two-point

Table 8 Comparison of the single-run method and the random-effects model for the communications network with $m=500$

Single-run method	*Input variance*	*Stochastic variance*
Average	$\mathbf{1.1488\times10^{-7}}$	$\mathbf{7.1301\times10^{-11}}$
Standard error	1.5088×10^{-9}	7.7480×10^{-13}
Random-effects method	$\mathbf{1.1265\times10^{-7}}$	$\mathbf{7.1301\times10^{-11}}$
99% confidence interval	$[1.0\times10^{-7}, 1.3\times10^{-7}]$	

Note: The primary values to compare are highlighted in boldface type.

Table 9 Empirical type I error from multivariate normality tests applied to 1000 macroreplications of the communications network with nominal level 0.05

m	*Henze-Zirkler*	*Royston's H*
50	0.053 ± 0.014	0.098 ± 0.018
500	0.078 ± 0.016	0.099 ± 0.019

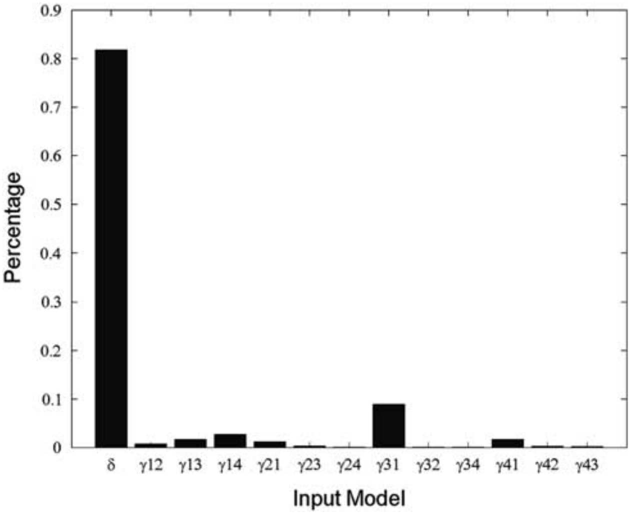

Figure 3 Normalized contributions of individual input models.

method, but it provides conservative (upper bound) estimates of input-uncertainty variance and does not allow estimating the contributions of the individual input distributions.

6. Conclusions

Input uncertainty is a type of model risk that is present in nearly all stochastic simulations. In this paper we presented a method to quantify that risk, and to distribute it among each of the input models, for situations in which the input models are parametric distributions whose parameters are estimated from real-world data. Our key contribution is that we facilitate applying the approximation of Cheng and Holland (1997, 1998, 2004) within the nominal experiment that would be performed even if input uncertainty were ignored; thus no additional diagnostic experiments are required, as they are in all other methods of which we are aware.

Application of our method requires only that generated values of the inputs, as well as the outputs, be collected during the simulation run for use in a least-squares regression. While we focused on parameter estimates that are MLEs, any method for fitting the input distributions that provides consistent estimators of the parameters, and also allows estimation of the variance-covariance matrix of the parameter estimates, will suffice.

Acknowledgements—This paper is based upon work supported by the National Science Foundation under Grant No. CMMI-1068473.

References

Ankenman BE and Nelson BL (2012). A quick assessment of input uncertainty. In: Laroque C, Himmelspach J, Pasupathy R, Rose O and Uhrmacher AM (eds). *Proceedings of the 2012 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc.: Piscataway, NJ, pp 241–250.

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

Cheng RCH and Holland W (1997). Sensitivity of computer simulation experiments to errors in input data. *Journal of Statistical Computation and Simulation* **57**(1): 219–241.

Cheng RCH and Holland W (1998). Two-point methods for assessing variability in simulation output. *Journal of Statistical Computation and Simulation* **60**(3): 183–205.

Cheng RCH and Holland W (2004). Calculation of confidence intervals for simulation output. *ACM Transactions on Modeling and Computer Simulation* **14**(4): 344–362.

Fu MC (ed) (2015). Stochastic gradient estimation. In: *Handbook of Simulation Optimization*. Chapter 5, Springer: New York.

Kleinrock L (1976). *Queueing Systems: II Computer Applications*. John Wiley: New York.

Nelson BL (2013). *Foundations and Methods of Stochastic Simulation: A First Course*. Springer: New York.

Schmeiser B (1982). Batch size effects in the analysis of simulation output. *Operations Research* **30**(3): 556–568.

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

Song E and Nelson BL (2014). Advanced tutorial: Input uncertainty quantification. In: Tolk A, Diallo SY, Ryzhov IO, Yilmaz L, Buckley S and Miller JA (eds). *Proceedings of the 2014 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc.: Piscataway, NJ, pp 162–176.

Song E and Nelson BL (2015). Quickly assessing contributions to input uncertainty. *IIE Transactions*. forthcoming.

Whitt W (2006). Analysis for design. In: Henderson SG and Nelson BL (eds). *Handbooks in Operations Research and Management Science: Simulation*. Chapter 13, North-Holland: New York.

Wieland JR and Schmeiser BW (2006). Stochastic gradient estimation using a single design point. In: Perrone LF, Wieland FP, Liu J, Lawson BG, Nicol DM and Fujimoto RM. *Proceedings of the 2006 Winter Simulation Conference*, Institute of Electrical and Electronics Engineers, Inc.: Piscataway, NJ, pp 390–397.

Appendix

Multivariate gradient estimator

Wieland and Schmeiser (2006) provide a single-design-point derivative estimator for stochastic simulation based on least-squares regression and justified using bivariate normality. When a p-variate gradient is needed they suggest that either p individual regressions, or a single multivariate regression, might be performed, but do not conclude which to use. We argue that the correct answer is multivariate.

In our setting,

$$\mathbf{g}\left(\widehat{\boldsymbol{\theta}}\right) = \left(\frac{\partial\eta(\boldsymbol{\theta})}{\partial\theta_1}, \frac{\partial\eta(\boldsymbol{\theta})}{\partial\theta_2}, \ \ldots\ , \frac{\partial\eta(\boldsymbol{\theta})}{\partial\theta_p}\right)\Bigg|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}.}$$

We use $\widehat{\boldsymbol{\theta}}$ as the point at which we want to evaluate the gradient for consistency with the paper, but the argument below could be for any arbitrary fixed value of $\boldsymbol{\theta}$.

Recall that

$$\left.\frac{\partial\eta(\boldsymbol{\theta})}{\partial\theta_1}\right|_{\boldsymbol{\theta}=\widehat{\boldsymbol{\theta}}} = \lim_{h\to 0}\frac{\eta\left(\widehat{\theta}_1+h,\widehat{\theta}_2,\ \ldots\ ,\widehat{\theta}_p\right)-\eta\left(\widehat{\theta}_1,\widehat{\theta}_2,\ \ldots\ ,\widehat{\theta}_p\right)}{h}.$$

In words, the partial derivative is the rate of change of the response with respect to θ_1 when all of the coordinates are fixed at $\widehat{\boldsymbol{\theta}}$. Thus, the value of the partial derivative of interest depends on the particular values of all of the parameters. This is why $p+1$ simulations are required to estimate the gradient using finite differences.

The method of Wieland and Schmeiser (2006) estimates the derivative by regressing the output on the realized, and therefore random, values of $\overline{\theta}_1$ in each replication, say $\overline{\theta}_{1j}$, $j=1,2,\ldots,n$. When we have multiple parameters, then the realized values of $(\overline{\theta}_{1j},\overline{\theta}_{2j},\ \ldots\ ,\overline{\theta}_{pj})$ are all random and all in motion simultaneously. If they are independent, as would occur if we had p independent, univariate input models each with one parameter, then estimating each term in the gradient by a single-variable regression is appropriate because the effects of the other parameters are averaged out. However, if any of the parameter estimators are dependent, as would occur in most multi-parameter distributions using MLE, then the realized values move together and the gradient estimator is incorrect.

To illustrate this effect, suppose that there are only $p=2$ parameters for a single distribution, and the realized values from the *j*th replication have joint distribution

$$\begin{pmatrix} Y_j\left(\widehat{\boldsymbol{\theta}}\right) \\ \overline{\theta}_{1j} \\ \overline{\theta}_{2j} \end{pmatrix} \sim \mathrm{N}\left[\begin{pmatrix} \eta\left(\widehat{\boldsymbol{\theta}}\right) \\ \widehat{\theta}_{1j} \\ \widehat{\theta}_{2j} \end{pmatrix}, \begin{pmatrix} \sigma_Y^2 & \sigma_{Y1} & \sigma_{Y2} \\ \sigma_{Y1} & \sigma_1^2 & \sigma_{12} \\ \sigma_{Y2} & \sigma_{12} & \sigma_2^2 \end{pmatrix}\right]. \qquad (9)$$

Then the coefficient of the θ_1 term in the conditional expected value (5) is

$$\beta_1 = \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2\sigma_2^2 - \sigma_{12}^2}\left(\frac{\sigma_{Y1}}{\sigma_1^2} - \sigma_{Y2}\left(\frac{\sigma_{12}}{\sigma_1^2\sigma_2^2}\right)\right).$$

If the parameter estimators are uncorrelated, then $\sigma_{12}=0$ and $\beta_1 = \sigma_{Y1}/\sigma_1^2$, the coefficient we would obtain if we considered $(Y_j(\widehat{\boldsymbol{\theta}}), \overline{\theta}_{1j})$ as an isolated bivariate pair. The multivariate regression correctly accounts for the joint effect of the two parameters, providing the rate of change of $\eta(\widehat{\boldsymbol{\theta}})$ with respect to θ_1 with θ_2 fixed.

Received 12 September 2014;
accepted 20 January 2015 after two revisions