Proceedings of the 2022 Winter Simulation Conference
B. Feng, G. Pedrielli, Y. Peng, S. Shashaani, E. Song, C.G. Corlu, L.H. Lee, E.P. Chew, T. Roeder, and P. Lendermann, eds.

INPUT UNCERTAINTY QUANTIFICATION FOR QUANTILES

Drupad Parmar
Lucy E. Morgan
Andrew C. Titman
Richard A. Williams

STOR-i Centre for Doctoral Training
Lancaster University
Lancaster, LA1 4YW, UK

Susan M. Sanchez

Operations Research Department
Naval Postgraduate School
Monterey, CA 93943, USA

ABSTRACT

Input models that drive stochastic simulations are often estimated from real-world samples of data. This leads to uncertainty in the input models that propagates through to the simulation outputs. Input uncertainty typically refers to the variance of the output performance measure due to the estimated input models. Many methods exist for quantifying input uncertainty when the performance measure is the sample mean of the simulation outputs, however quantiles that are frequently used to evaluate simulation output risk cannot be incorporated into this framework. Here we adapt two input uncertainty quantification techniques for when the performance measure is a quantile of the simulation outputs rather than the sample mean. We implement the methods on two examples and show that both methods accurately estimate an analytical approximation of the true value of input uncertainty.

1 INTRODUCTION

The randomness in stochastic simulation models comes from input models that are typically represented by some probability distributions or processes. Often these input models are fit using samples of data taken from the real-world system. Since the samples are necessarily finite the fitted input models are never truly representative of reality. This introduces a source of uncertainty into the simulation model that will propagate through to the outputs. If this uncertainty is not considered in simulation output analysis then important decisions are at risk of being made with misleading levels of confidence. Input uncertainty broadly refers to the impact of input model uncertainty on simulation outputs. More specifically input uncertainty quantification aims to quantify the variance in the performance measure due to having estimated the input models. These methods often consider the performance measure to be the sample mean of the simulation outputs however alternative performance measures might be helpful, for example to learn about the distributional properties of the simulation output or to identify different features between two systems that have similar sample means.

Quantiles are useful for assessing risk. They are particularly common in financial portfolio management where they are referred to as value at risk. Quantiles can be estimated from simulation outputs using the empirical cumulative distribution function. Existing work on quantile uncertainty quantification accounts for the uncertainty in the quantile estimate due to the finite number of outputs. If the simulation model is driven by input models fitted using real-world data then there is input model uncertainty which will propagate through the model to the quantile estimate. Current methods do not account for this additional source of error.

We are interested in quantifying the effect of input model uncertainty on quantile estimates calculated from simulation outputs. That is, this work aims to quantify input uncertainty when considering a quantile of the simulation outputs, rather than the sample mean. In the following section we discuss the input uncertainty quantification problem for the mean and describe how this changes for quantiles.

2 INPUT UNCERTAINTY

Consider a simulation model driven by L independent input distributions denoted by $\boldsymbol{G} = \{G_1, \ldots, G_L\}$, each of which could be parametric or nonparametric. We represent the output of replication j as

$$Y_j(\boldsymbol{G}) = \eta(\boldsymbol{G}) + \varepsilon_j(\boldsymbol{G}),$$

where $\eta(\boldsymbol{G}) = \mathrm{E}[Y_j(\boldsymbol{G})]$ is the expected value of the simulation output random variable and $\varepsilon_j(\boldsymbol{G})$ is a random variable with mean 0 representing stochastic noise. Assume there are true input distributions denoted by $\boldsymbol{G}^0 = \{G_1^0, \ldots, G_L^0\}$. Suppose the true input distributions are unknown but real-world data can be collected from each distribution. Suppose we take a sample of m_l observations from the lth input distribution and compute a collection of fitted input distributions denoted by $\hat{\boldsymbol{G}} = (\hat{G}_1, \ldots, \hat{G}_L)$. A nominal experiment consists of running n i.i.d. simulation replications using the fitted input distributions to obtain outputs $Y_1(\hat{\boldsymbol{G}}), Y_2(\hat{\boldsymbol{G}}), \ldots, Y_n(\hat{\boldsymbol{G}})$.

2.1 Input Uncertainty Quantification for the Mean

The original input uncertainty problem assumes the goal of the experiment is to estimate $\eta(\boldsymbol{G}^0)$, the expected value of the simulation output random variable under the true input distributions. We estimate this via the sample mean of the simulation outputs

$$\bar{Y}(\hat{\boldsymbol{G}}) = \frac{1}{n}\sum_{j=1}^{n} Y_j(\hat{\boldsymbol{G}}) = \eta(\hat{\boldsymbol{G}}) + \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j(\hat{\boldsymbol{G}}).$$

The variance of this point estimator is

$$\mathrm{Var}[\bar{Y}(\hat{\boldsymbol{G}})] = \frac{1}{n}\mathrm{E}\big[\mathrm{Var}(\varepsilon_1(\hat{\boldsymbol{G}}) \mid \hat{\boldsymbol{G}})\big] + \mathrm{Var}\big[\eta(\hat{\boldsymbol{G}})\big], \tag{1}$$

where the outer expectation and variance on the right-hand side of the equation are with respect to the sampling distribution of $\hat{\boldsymbol{G}}$. The first term in Equation (1) measures the expected variability due to the stochastic noise, given the fitted input distributions. We shall define this as *stochastic uncertainty for the mean* and denote with $\sigma_{S,M}^2$

$$\sigma_{S,M}^2 = \frac{1}{n}\mathrm{E}\big[\mathrm{Var}(\varepsilon_1(\hat{\boldsymbol{G}}) \mid \hat{\boldsymbol{G}})\big]. \tag{2}$$

This can be driven towards 0 by increasing the number of replications. The second term in Equation (1) measures the variance in the system mean due to having estimated the input distributions. We shall define this as *input uncertainty for the mean* and denote with $\sigma_{I,M}^2$

$$\sigma_{I,M}^2 = \mathrm{Var}\big[\eta(\hat{\boldsymbol{G}})\big]. \tag{3}$$

This depends on the sample sizes of real-world data used to fit the input distributions, as well as the structure of $\eta(\cdot)$, which is usually unknown. The aim of input uncertainty quantification is to estimate this term.

Various methods have been proposed to quantify input uncertainty for the mean of simulation outputs, for an overview see Song et al. (2014) or Lam (2016). There are two existing methods that we will adapt in this paper. The first, from Nelson (2013) (Section 7.2), uses bootstrapping to capture the variability of the input distributions and simulation to propagate input model uncertainty. Input uncertainty can be estimated by subtracting the stochastic uncertainty from the total variability of the bootstrapped outputs. The second method, by Cheng and Holland (1997), considers the case of parametric input distributions. Input uncertainty is modelled using a first-order Taylor series approximation and requires estimates of the parameter variance and the gradient of $\eta(\cdot)$.

2.2 Input Uncertainty Quantification for Quantiles

Suppose that instead of estimating the expected value of the simulation output random variable under the true input distributions, we estimate a quantile of the simulation output random variable under the true input distributions. For a random variable Y with a strictly increasing cumulative distribution function (CDF) F, the p-quantile, for $0 < p < 1$, is defined as the constant $\xi_p = F^{-1}(p) = \inf\{y : F(y) \geq p\}$. A common example of a quantile is the median, which is the 0.5-quantile, $\xi_{0.5}$.

Denote the simulation output random variable under the true input distributions by $Y(\boldsymbol{G}^0)$, with CDF F_0. The p-quantile of F_0, for $0 < p < 1$, is given by $\xi_p(\boldsymbol{G}^0) = F_0^{-1}(p) = \inf\{y : F_0(y) \geq p\}$. Assume that F_0 is strictly increasing, differentiable at $\xi_p(\boldsymbol{G}^0)$, and that $f(\xi_p(\boldsymbol{G}^0)) > 0$, where f is the derivative of F. Given n i.i.d. outputs we can construct an empirical CDF, which can be inverted to obtain a quantile point estimate (Serfling 2009) (Section 2.3). We can estimate F_0 via the empirical CDF $\hat{F}_n$, defined by

$$\hat{F}_n(y) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j(\hat{\boldsymbol{G}}) \leq y),$$

where $I(\cdot)$ denotes the indicator function. The p-quantile estimator from our nominal experiment is given by $\xi_{p,n}(\hat{\boldsymbol{G}}) = \hat{F}_n^{-1}(p)$. This is equivalent to $\xi_{p,n}(\hat{\boldsymbol{G}}) = Y_{(\lceil np \rceil)}(\hat{\boldsymbol{G}})$, where $Y_{(1)}(\hat{\boldsymbol{G}}) \leq Y_{(2)}(\hat{\boldsymbol{G}}) \leq \cdots \leq Y_{(n)}(\hat{\boldsymbol{G}})$ are the order statistics of the outputs and $\lceil \cdot \rceil$ represents the ceiling function.

Applying the law of total variance to the quantile estimator gives

$$\mathrm{Var}\left[\xi_{p,n}(\hat{\boldsymbol{G}})\right] = \mathrm{E}\left[\mathrm{Var}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})\right] + \mathrm{Var}\left[\mathrm{E}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})\right], \tag{4}$$

where the outer expectation and variance on the right-hand side of the equation are with respect to the sampling distribution of $\hat{\boldsymbol{G}}$. The first term in Equation (4) measures the expected variability of the quantile estimate given the fitted input distributions. We shall define this as *stochastic uncertainty for the quantile* and denote with $\sigma^2_{S,Q}$

$$\sigma^2_{S,Q} = \mathrm{E}\left[\mathrm{Var}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})\right]. \tag{5}$$

This is the uncertainty due to having estimated the quantile via simulation. This will tend towards 0 as the number of replications increases since given the fitted input distributions the simulated CDF will approximate the true CDF. The second term in Equation (4) measures the variance of the expected value of the quantile estimate given the fitted input distributions. We shall define this as *input uncertainty for the quantile* and denote with $\sigma^2_{I,Q}$

$$\sigma^2_{I,Q} = \mathrm{Var}\left[\mathrm{E}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})\right]. \tag{6}$$

This measures the uncertainty in the expectation of the quantile estimate due to having estimated the input distributions. This is complex and depends upon the sample sizes used to estimate the input distributions as well as the CDF of the simulation output random variable at the fitted input distributions, which is usually unknown. As this term is different to input uncertainty for the mean in Equation (3), we will require different methods to quantify it. Note that since the quantile estimator is asymptotically unbiased (Nakayama 2014), then for large enough n it follows that

$$\mathrm{E}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}}) \approx \xi_p(\hat{\boldsymbol{G}}), \tag{7}$$

where $\xi_p(\hat{\boldsymbol{G}})$ is the p-quantile of the simulation output random variable under the estimated input distributions.

There is little literature on input uncertainty quantification for quantiles. Zhu et al. (2020) define and provide estimators to quantiles of the mean performance measure under input uncertainty. Xie et al. (2018) develop a Bayesian framework to quantify both the stochastic uncertainty and input uncertainty of percentiles of simulation outputs. Our work aims to quantify input uncertainty for quantiles from a frequentist perspective.

3 METHODS

We have outlined the original input uncertainty problem and discussed how this changes for quantiles. We now develop two input uncertainty quantification techniques for quantiles. We consider a bootstrapping approach and a Taylor series approximation, the latter of which is restricted to the case of parametric input distributions. For each method we describe its application to the mean followed by our adaptation for quantiles. When we refer to just input uncertainty or stochastic uncertainty in this section, this will be specific to the mean or quantile depending on the subsection.

3.1 Bootstrapping for the Mean

Here we describe the bootstrapping method from (Nelson 2013) (Section 7.2) which is used to estimate input uncertainty for the mean. Bootstrapping approximates the sampling distribution of the fitted input models. A single bootstrap consists of three parts. Firstly for each input distribution we sample with replacement m_l observations from each set of m_l initial observations. Secondly these samples are used to estimate bootstrap fitted input distributions. Thirdly the bootstrap fitted input distributions are used to run simulation replications. Suppose we use B bootstraps. We denote the bootstrap fitted input distributions by $\hat{\boldsymbol{G}}_k$ for $k = 1,\ldots,B$, and we denote the outputs from the kth bootstrap by $Y_1(\hat{\boldsymbol{G}}_k),\ldots,Y_n(\hat{\boldsymbol{G}}_k)$. This diagnostic experiment requires a total of Bn replications.

Let the mean of the outputs from the kth bootstrap be denoted by $\bar{Y}(\hat{\boldsymbol{G}}_k) = \sum_{j=1}^{n} Y_j(\hat{\boldsymbol{G}}_k)/n$, and let the mean of these means be denoted by $\bar{\bar{Y}} = \sum_{k=1}^{B} \bar{Y}(\hat{\boldsymbol{G}}_k)/B$. The total variance of the mean performance measure from the nominal experiment is estimated by the sample variance of the bootstrapped means

$$\hat{\sigma}^2_{T,M} = \frac{1}{B-1}\sum_{k=1}^{B}(\bar{Y}(\hat{\boldsymbol{G}}_k) - \bar{\bar{Y}})^2.$$

This term approximately measures both input uncertainty and stochastic uncertainty. Stochastic uncertainty is approximated by calculating the sample variance of the outputs in each bootstrap, averaging these across bootstraps, and dividing by a factor of n

$$\hat{\sigma}^2_{S,M} = \frac{1}{n}\left(\frac{1}{B}\sum_{k=1}^{B}\left(\frac{1}{(n-1)}\sum_{j=1}^{n}(Y_j(\hat{\boldsymbol{G}}_k) - \bar{Y}(\hat{\boldsymbol{G}}_k))^2\right)\right).$$

To estimate input uncertainty we subtract stochastic uncertainty from the total variance

$$\hat{\sigma}^2_{I,M} = \hat{\sigma}^2_{T,M} - \hat{\sigma}^2_{S,M}.$$

Note that this could return a negative estimate of input uncertainty, which is interpreted as meaning that the effect of input uncertainty is relatively small compared to stochastic uncertainty.

3.2 Bootstrapping for Quantiles

We now adapt the bootstrapping method for quantiles. For the kth bootstrap we can compute an empirical CDF from the outputs of the n replications

$$\hat{F}_{n,k}(y) = \frac{1}{n}\sum_{j=1}^{n} I(Y_j(\hat{\boldsymbol{G}}_k) \leq y),$$

and a quantile estimate $\xi_{p,n}(\hat{\boldsymbol{G}}_k) = \hat{F}^{-1}_{n,k}(p)$. Let $\bar{\xi}_{p,n,B} = \sum_{k=1}^{B} \xi_{p,n}(\hat{\boldsymbol{G}}_k)/B$ denote the average of the quantile estimates across bootstraps. The total variance of the quantile estimate from the nominal experiment is approximated by the sample variance of the bootstrapped quantile estimates

$$\hat{\sigma}^2_{T,Q} = \frac{1}{B-1}\sum_{k=1}^{B}(\xi_{p,n}(\hat{\boldsymbol{G}}_k) - \bar{\xi}_{p,n,B})^2.$$

This term approximately measures both input uncertainty and stochastic uncertainty. As previously, we estimate input uncertainty by subtracting an estimate of stochastic uncertainty from the total variance, however we cannot approximate stochastic uncertainty for the quantile in the same way as for the mean. Recall that stochastic uncertainty for the quantile is the expectation of the variance of the quantile estimate with respect to the sampling distribution of $\hat{\boldsymbol{G}}$. The sample variance of the simulation outputs does not provide an approximation to the variance of the quantile estimate, so we require a different method here.

There are myriad ways to approximate the variance of the quantile estimator. The quantile estimator satisfies a central limit theorem (Serfling 2009) (Section 2.3.3), however the asymptotic variance contains the density function which is typically unknown. Computing a consistent estimator of the asymptotic variance is non-trivial (Nakayama 2014). Although finite differences can be used to estimate the density (Serfling 2009) (Section 2.6.2), this requires specification of a suitable bandwidth parameter. Alternatively bootstrapping methods can be used to directly estimate the variance. The conventional unsmoothed bootstrap is shown to have high relative error (Hall and Martin 1988), which can be reduced by using a smoothed bootstrap based on a kernel density estimate (Hall et al. 1989). However this requires stronger smoothness conditions on the density and a suitable choice of smoothing bandwidth. Cheung and Lee (2005) estimate the variance of the quantile estimator using a modification of the bootstrap known as the m out of n bootstrap. Although this requires a choice for m it seems to be less crucial than the choice of the smoothing bandwidth in terms of the sensitivity and stability of the mean squared error of each estimator. Shao and Wu (1989) show that the jackknife estimator with d observations removed gives consistent and asymptotically unbiased estimates of the quantile estimator variance for suitable choices of d.

Alternatively we can approximate the variance of the quantile estimator by applying batching or sectioning (Asmussen and Glynn 2007) (Section III.5a). These methods avoid the complication of consistently estimating the density function and are less computationally intensive than bootstrapping procedures, since they only use the results from the nominal experiment. Both involve dividing the outputs into batches and taking quantile estimates from each batch. Batching utilises the variance of the batch quantile estimates, whilst sectioning replaces the sample mean in the variance calculation with the quantile estimator from all the outputs. The batching and sectioning variance divided by the number of batches provides an approximation to the quantile estimator variance.

Whichever method is used, suppose that σ_k^2 represents the variance of the quantile estimate from the kth bootstrap. We estimate stochastic uncertainty by averaging these variance estimates across bootstraps

$$\hat{\sigma}_{S,Q}^2 = \frac{1}{B}\sum_{k=1}^{B}\sigma_k^2.$$

To estimate input uncertainty we subtract stochastic uncertainty from the total variance

$$\hat{\sigma}_{I,Q}^2 = \hat{\sigma}_{T,Q}^2 - \hat{\sigma}_{S,Q}^2.$$

Again note that this could return a negative estimate of input uncertainty which we interpret similar to previously. We now describe the Taylor series approximation for the mean.

3.3 Taylor Series Approximation for the Mean

Suppose that the L input distributions follow known parametric distributions. In this case input model uncertainty becomes input parameter uncertainty so the input models can be denoted by a set of parameters $\boldsymbol{G} = \boldsymbol{\theta} = (\theta_1,\dots,\theta_q)$, where $q \geq L$. The true input models are given by the true parameters of the distributions, denoted by $\boldsymbol{G}^0 = \boldsymbol{\theta}^0 = (\theta_1^0,\dots,\theta_q^0)$. We suppose the parameters are estimated via maximum likelihood estimators (MLEs), which we denote by $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1,\dots,\hat{\theta}_q)$.

Cheng and Holland (1997) use a Taylor series approximation to estimate input uncertainty when taking the sample mean of simulation outputs. Input uncertainty can be approximated by

$$\hat{\sigma}_{I,M}^2 = \nabla\eta(\boldsymbol{\theta}^0)\mathrm{Var}(\hat{\boldsymbol{\theta}})\nabla\eta(\boldsymbol{\theta}^0)^\top,$$

where $\nabla\eta(\boldsymbol{\theta}^0)$ is the gradient of the expected value of the simulation output with respect to the input parameters $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^0$. This approximation of input uncertainty combines the sensitivity of the expected simulation output with respect to the input parameters, with how accurately the input parameters have been estimated. To use this approximation we need to estimate both the parameter variance and the gradient of the expected value of the simulation output.

Since the input parameters are estimated via maximum likelihood, we can approximate the parameter variance by the inverse Fisher information matrix evaluated at the MLEs

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}) = I(\hat{\boldsymbol{\theta}})^{-1}.$$

This follows since the asymptotic distribution of the MLEs is multivariate normal with covariance matrix $I(\boldsymbol{\theta}^0)^{-1}$, which can be consistently estimated by $I(\hat{\boldsymbol{\theta}})^{-1}$. Lin et al. (2015) note that MLEs are not required for the Taylor series approximation method, only that the covariance matrix of the parameter estimates can be approximated, which is most easily done when using MLEs.

There are many ways to estimate the gradient of the expected value of the simulation output. Cheng and Holland (1997) describe the delta method, which employs finite forward differences and requires computational effort that increases linearly with the number of parameters. To improve upon this they develop the two-point method (Cheng and Holland 1998), which utilises the delta method but makes most simulation replications at just two settings of parameter values. Lin et al. (2015) provide a method for estimating the gradient that requires no diagnostic experiment, only the replications from the nominal experiment.

Outside the input uncertainty literature, Fu (2006) provides an overview of gradient estimation techniques. Approaches for gradient estimation are divided into two main categories, direct and indirect. Direct approaches aim to estimate the true gradient via some analysis of the underlying mechanism of the simulation model. Methods include perturbation analysis, the likelihood ratio method, and weak derivatives (also known as measure-valued differentiation). Indirect approaches are characterised by two features; they estimate an approximation to the true gradient and they only utilise evaluations of the simulation model. Methods include finite differences and simultaneous perturbations. Generally indirect gradient estimators are more widely applicable since direct estimators can involve analysis specific to the problem and may also require some changes to how the simulation model runs. However direct estimators usually give unbiased estimators and eliminate the choice of a suitable perturbation parameter.

The gradient estimate can be combined with the parameter variance estimate to employ the Taylor series approximation of input uncertainty. Lin et al. (2015) note that the approximation also provides estimates of the contribution made to input uncertainty by each input distribution. Let $\boldsymbol{\theta}_l$ denote the parameters belonging to the lth input distribution, note that this could a scalar or a vector depending on the distribution. Under the initial assumption that the input distributions are independent, it follows that

$$\nabla\eta(\boldsymbol{\theta}^0)\mathrm{Var}(\hat{\boldsymbol{\theta}})\nabla\eta(\boldsymbol{\theta}^0)^\top = \sum_{l=1}^{L} \nabla\eta(\boldsymbol{\theta}_l^0)\mathrm{Var}(\hat{\boldsymbol{\theta}}_l)\nabla\eta(\boldsymbol{\theta}_l^0)^\top,$$

where $\mathrm{Var}(\hat{\boldsymbol{\theta}}_l)$ is the covariance matrix of $\hat{\boldsymbol{\theta}}_l$. Each term in the summation represents the contribution to input uncertainty from the lth input distribution. These measures can be useful when input uncertainty is large, as we can identify which distributions it would be most beneficial to collect additional data from in order to reduce input uncertainty. Typically input distributions with the largest contributions would be the best target for additional data collection.

3.4 Taylor Series Approximation for Quantiles

We now adapt the Taylor series approximation for quantiles. Applying Equations (6) and (7) in the context of known parametric input distributions, for large enough n it follows that

$$\sigma_{I,Q}^2 \approx \mathrm{Var}\left[\xi_p(\hat{\boldsymbol{\theta}})\right],$$

where $\xi_p(\hat{\boldsymbol{\theta}})$ is the p-quantile of the simulation output random variable at the estimated parameters. Subsequently under the same regularity conditions as stated in Cheng and Holland (1997), we have that

$$\hat{\sigma}^2_{I,Q} = \nabla\xi_p(\boldsymbol{\theta}^0)\mathrm{Var}(\hat{\boldsymbol{\theta}})\nabla\xi_p(\boldsymbol{\theta}^0)^\top,$$

where $\nabla\xi_p(\boldsymbol{\theta}^0)$ is the gradient of the p-quantile with respect to the input parameters $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^0$. This approximation of input uncertainty combines the sensitivity of the quantile with respect to the input parameters, with how accurately the input parameters have been estimated. To use this approximation we need to estimate both the parameter variance and the gradient of the p-quantile. As in the mean case, we can estimate the parameter variance using the inverse Fisher information matrix, however the gradient estimation requires a bit more thought.

A point of difference between the mean and quantiles is the number of estimates we obtain. For the mean each of the n outputs provides an estimate to the expectation of the simulation output random variable. However for quantiles we only obtain a single estimate from a set of n outputs. Consequently the gradient estimation methods described by Cheng and Holland (1997), Cheng and Holland (1998), and Lin et al. (2015), which can be used for the mean are not applicable for quantiles. Neither are the direct gradient estimators described in Fu (2006), since they are specifically derived for performances measures that are expectations. The indirect gradient estimators described in Fu (2006) however can be applied when the performance measure is a quantile.

There is a small amount of fairly recent literature on direct gradient estimators for quantiles. Hong (2009) proposes a consistent estimator by combining infinitesimal perturbation analysis with batching. Alternatively Heidergott and Volk-Makarewicz (2009) present a quantile gradient estimate based on measure-valued differentiation. Liu and Hong (2009) describe a kernel estimator that is consistent and more efficient than Hong (2009). Fu et al. (2009) use conditional Monte Carlo to derive a consistent estimator that does not require batching. More recently Lei et al. (2018) applied a generalised likelihood ratio method to develop an estimator that also does not require batching. Since these methods all fall under the category of direct gradient estimation they typically require some problem-specific analysis and consequently they are not applicable to a broad range of simulation models.

Note we do not advocate for any particular method to be used to estimate the quantile gradient, but in the experiments that follow we use the symmetric difference estimator described in Fu (2006) (Section 3.1). Once we have an estimate for the quantile gradient we can combine this with the parameter variance estimate to employ the Taylor series approximation of input uncertainty. Again this naturally yields the approximate contribution to input uncertainty by each input distribution, where each term in the summation represents the contribution to input uncertainty from the lth input distribution

$$\nabla\xi_p(\boldsymbol{\theta}^0)\mathrm{Var}(\hat{\boldsymbol{\theta}})\nabla\xi_p(\boldsymbol{\theta}^0)^\top = \sum_{l=1}^{L}\nabla\xi_p(\boldsymbol{\theta}^0_l)\mathrm{Var}(\hat{\boldsymbol{\theta}}_l)\nabla\xi_p(\boldsymbol{\theta}^0_l)^\top.$$

4 EXPERIMENTS

We now implement both input uncertainty methods for quantiles on two different examples. In the remainder of this paper, when we refer to just input uncertainty or stochastic uncertainty, this will be for the quantile. Firstly we derive an analytical example. This allows us to illustrate the problem of input uncertainty and compare input uncertainty estimates produced by each method against an approximation of the true value. We then use a stochastic activity network that utilises more input distributions than the analytical example. This allows us to consider more interesting results for the estimated contributions to input uncertainty.

4.1 Analytical Example

To illustrate the problem of input uncertainty we shall create an analytical example. This is contrived to enable input uncertainty to be derived analytically and is not meant to represent a realistic simulation problem.

To create an analytical example we need to be able to compute both $\mathrm{E}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})$ and $\mathrm{Var}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})$. If we know the CDF of the simulation output random variable and can derive the inverse CDF then we can write $\xi_p(\hat{\boldsymbol{G}})$ explicitly. Since the quantile estimator is asymptotically unbiased this will give us an approximation to $\mathrm{E}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}})$. We can approximate the variance of the quantile estimator given the fitted input models using the well-known asymptotic distribution of the sample quantile (Nelson 2013) (Section 7.1). Using the asymptotic variance it follows that

$$\mathrm{Var}(\xi_{p,n}(\hat{\boldsymbol{G}})|\hat{\boldsymbol{G}}) \approx \frac{p(1-p)}{nf(\xi_p(\hat{\boldsymbol{G}}))^2}, \tag{8}$$

where $f(\xi_p(\hat{\boldsymbol{G}}))$ is the probability density function evaluated at the p-quantile under the fitted input models.

Suppose a simulation model has two input models which are known to follow exponential distributions with unknown parameters. The input models are defined by two parameters $\boldsymbol{G} = (\mu, \beta)$ both of which are to be estimated from real-world observations. In this case input model uncertainty can be thought of as input parameter uncertainty. Suppose we observe $m_1 = m_2 = m$, i.i.d. observations from each distribution. Observations $x_1, \ldots, x_m$ are used to estimate μ and observations $z_1, \ldots, z_m$ are used to estimate β. The fitted input models are then given by the estimated parameters $\hat{\boldsymbol{G}} = (\hat{\mu}, \hat{\beta})$. The parameters can be estimated by their MLEs

$$\hat{\mu} = \left(\frac{1}{m}\sum_{i=1}^{m} x_i\right)^{-1}, \qquad \hat{\beta} = \left(\frac{1}{m}\sum_{i=1}^{m} z_i\right)^{-1}.$$

Suppose that these parameters drive the simulation model for n i.i.d. replications and the distribution of the simulation output random variable is given by $Y \sim \mathrm{Gumbel}(\hat{\mu}, \hat{\beta})$. Using Equations (5) and (8), and the probability density function of the Gumbel distribution, stochastic uncertainty for the quantile is approximately

$$\begin{aligned} \sigma_{S,Q}^2 &\approx \mathrm{E}\left[p(1-p)n^{-1}\left(e^{-(z+e^{-z})}\hat{\beta}^{-1}\right)^{-2}\right], \\ &\approx p(1-p)n^{-1}\mathrm{E}\left[\left(e^{-(z+e^{-z})}\hat{\beta}^{-1}\right)^{-2}\right], \end{aligned} \tag{9}$$

where $z = (\xi_p(\hat{\mu}, \hat{\beta}) - \hat{\mu})/\hat{\beta}$. Although we cannot compute this expectation analytically it can be approximated via numerical integration. The expectation term in Equation (9) will not depend upon the number of outputs n and therefore stochastic uncertainty will tend towards 0 as n increases. Using Equations (6) and (7), and the inverse CDF of the Gumbel distribution, input uncertainty for the quantile is approximately

$$\begin{aligned} \sigma_{I,Q}^2 &\approx \mathrm{Var}\left[\hat{\mu} - \hat{\beta}\ln(-\ln(p))\right], \\ &\approx \frac{m^2\mu^2}{(m-1)^2(m-2)} + \left(\ln(-\ln(p))\right)^2 \frac{m^2\beta^2}{(m-1)^2(m-2)}. \end{aligned} \tag{10}$$

This follows since if observations $a_1, \ldots, a_m$ are i.i.d. from an exponential distribution with rate θ and $X = \sum_{i=1}^{m} a_i$, then $1/X \sim \text{Inv-Gamma}(m, \theta)$ and hence $\mathrm{Var}[1/X] = \theta^2/((m-1)^2(m-2))$. Equation (10) does not depend on the number of outputs n, but does depend on the number of observations m used to fit the input parameters.

To illustrate the importance of input uncertainty quantification we will consider the following experiment. Let $\mu = 2$, $\beta = 3$ and $m = 250$. Suppose we use a nominal experiment of $n = 10000$ replications from which we estimate the 0.95-quantile. If input uncertainty is not considered then we approximate the variance of our quantile estimate by applying any of the methods described in Section 3.2. Suppose we use sectioning. This involves dividing the outputs into batches and calculating the sum of squared errors between the batch quantile estimates and the overall quantile estimate. The variance is then given by the sum of squared errors divided by the number of batches. We run 1000 macro replications of our nominal experiment and each time we compute the variance of the quantile estimate using sectioning with 20 batches (Asmussen and Glynn (2007) suggest choosing 30 batches or fewer). The average variance across the macro replications is approximately 0.01812.

The variance of our quantile estimate from the nominal experiment should be given by the sum of stochastic uncertainty and input uncertainty. Using Equations (9) and (10) these are approximately

$$\sigma^2_{S,Q} \approx 0.01794, \qquad \sigma^2_{I,Q} \approx 0.3390,$$

where we have used 1×10^5 samples of $(\hat{\mu}, \hat{\beta})$ to estimate the expectation term in Equation (9). Sectioning provides an approximation of stochastic uncertainty, but does not capture input uncertainty. Input uncertainty is almost 19 times larger, so ignoring it could have serious practical consequences. Although we can derive an analytical approximation of input uncertainty in this particular example, for most realistic simulation problems this is not the case. This motivates the need for methods to quantify input uncertainty.

We use this analytical example to test the accuracy of our two methods. If a nominal experiment uses n replications and we estimate input uncertainty via B bootstraps, then this diagnostic experiment requires a total of Bn replications. For a fair comparison between the bootstrapping method and the Taylor series approximation method we use the same number of total replications to estimate input uncertainty for each method. Since estimating the parameter variance requires no replications, we utilise Bn replications to estimate the quantile gradient.

We keep $\mu = 2, \beta = 3, n = 10000$, and compare estimates of input uncertainty for quantiles $p = (0.8, 0.95)$ and input sample sizes $m = (250, 1000)$. For the bootstrapping method we use $B = 10000$ bootstraps and apply sectioning with 10 batches. For the Taylor series approximation we estimate the quantile gradient using the symmetric difference gradient estimator described in Fu (2006) (Section 3.1), with $c = (0.1, 0.1)$. This requires simulation runs at 4 sets of parameters, so for each set we use 2.5×10^7 replications. The results from each approach, averaged across 1000 macro replications are shown in Table 1, along with the analytical approximation of input uncertainty computed using Equation (10).

Table 1: Comparing input uncertainty estimates for the analytical example.

m	Method	$p = 0.8$		$p = 0.95$	
		Mean	Std. Error	Mean	Std. Error
250	Bootstrapping	9.967×10^{-2}	1.564×10^{-2}	3.432×10^{-1}	6.002×10^{-2}
	Taylor Series Approximation	9.853×10^{-2}	1.081×10^{-2}	3.389×10^{-1}	4.176×10^{-2}
	Analytical Approximation	9.856×10^{-2}	-	3.390×10^{-1}	-
1000	Bootstrapping	2.433×10^{-2}	1.853×10^{-3}	8.356×10^{-2}	7.074×10^{-3}
	Taylor Series Approximation	2.433×10^{-2}	1.321×10^{-3}	8.361×10^{-2}	5.068×10^{-3}
	Analytical Approximation	2.435×10^{-2}	-	8.373×10^{-2}	-

For $m = 250$ the Taylor series approximation estimates have a more accurate mean and a smaller standard error than the bootstrapping estimates for both values of p. For $m = 1000$ the bootstrapping estimates and the Taylor series approximation estimates return similarly accurate means for both values of p, although the Taylor series approximation estimates have a smaller standard error. These results show us that both methods are accurately estimating the approximate true value.

4.2 Stochastic Activity Network

Additionally we run experiments using the stochastic activity network described in Dong and Nakayama (2014) and Nelson (2013) (Section 3.4). A stochastic activity network models the completion time of a project using a group of activities with precedence constraints and random durations. The model consists of $L = 5$ random processes, each of which models the duration of an activity. Each follows an independent exponential distribution and therefore we have $q = 5$ parameters, denoted by $\boldsymbol{\theta} = (\theta_1, \theta_2, \theta_3, \theta_4, \theta_5)$. The activities form 3 paths in the network, denoted by $P_1 = \{1,2\}$, $P_2 = \{1,3,5\}$ and $P_3 = \{4,5\}$. The simulation output is the completion time of the project. This is measured by the longest path in the network. Using A_j to denote the duration of the jth activity, for $1 < j < 5$, the simulation output is given by $Y = \max\{A_1 + A_2, A_1 + A_3 + A_5, A_4 + A_5\}$.

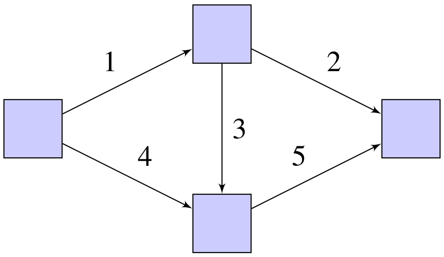

Figure 1: A graphical representation of the stochastic activity network.

Suppose the true parameters are given by $\boldsymbol{\theta}^0 = (1,1,1,1,1)$ and each parameter is estimated using the same number of observations, that is $m_l = m$, for $1 < l < 5$. Again our nominal experiment consists of $n = 10000$ replications and we compare estimates of input uncertainty for $p = (0.8, 0.95)$ and $m = (250, 1000)$. For the bootstrapping method we use $B = 10000$ bootstraps and apply sectioning with $b = 10$ batches. For the Taylor series approximation we use the symmetric difference gradient estimator with $c_l = 0.1$, for $1 < l < 5$. This requires simulation runs at 10 sets of parameters, so for each set we use 1×10^7 replications. The results from each approach, averaged across 1000 macro replications are shown in Table 2.

Table 2: Comparing input uncertainty estimates for the stochastic activity network.

m	Method	$p = 0.8$		$p = 0.95$	
		Mean	Std. Error	Mean	Std. Error
250	Bootstrapping	2.044×10^{-2}	2.047×10^{-3}	4.375×10^{-2}	4.924×10^{-3}
	Taylor Series Approximation	2.015×10^{-2}	1.549×10^{-3}	4.296×10^{-2}	3.934×10^{-3}
1000	Bootstrapping	4.994×10^{-3}	2.690×10^{-4}	1.060×10^{-2}	6.592×10^{-4}
	Taylor Series Approximation	4.971×10^{-3}	1.891×10^{-4}	1.055×10^{-2}	4.831×10^{-4}

For 3 of the 4 combinations of m and p, the mean estimates of input uncertainty from the bootstrapping and Taylor series approximation match to 2 decimal places. Although we cannot approximate the true values of input uncertainty it is reassuring that both methods are returning similar mean estimates. Across all 4 combinations we see that the Taylor series approximation has a smaller standard error than bootstrapping. Although we see this smaller standard error across both experiments there are too many variables to conclude whether we would expect to see this generally.

Using the results from the Taylor series approximation we can also look at the contributions made to input uncertainty by each input distribution. We calculate the average normalised contribution to input uncertainty by each input distribution across the 1000 macro replications. Table 3 shows the results for $p = (0.8, 0.95)$ when $m = 1000$.

Table 3: Average normalised contributions to input uncertainty for the stochastic activity network.

p	θ_1	θ_2	θ_3	θ_4	θ_5
0.8	32.0%	6.4%	23.4%	6.4%	31.8%
0.95	33.8%	4.1%	24.5%	4.1%	33.5%

Firstly note that for each quantile the normalised contributions to input uncertainty are approximately the same for parameters θ_1 and θ_5, and also for θ_2 and θ_4. We would expect to see this due to the symmetry of the stochastic activity network. We could switch the labels of these activities and the simulation model would remain the same. Secondly for both quantiles we see that θ_1 and θ_5 make the largest contributions. We would expect to see this since all activities are identically distributed and both these parameters represent activities that feature in 2 out of the 3 paths in the network. Both also feature in the path with the highest number of activities, which is likely to be the longest path. The second largest contributions to input uncertainty for both quantiles comes from θ_3, whilst θ_2 and θ_4 return the smallest contributions. Moving from the 0.8-quantile to the 0.95-quantile the contributions made by θ_1, θ_3 and θ_5 increase, whilst the contributions made by θ_2 and θ_4 decrease. Parameters θ_1 and θ_5 should be targeted for additional data collection since these make the largest normalised contribution for both quantiles.

5 CONCLUSION

In this work we considered input uncertainty quantification for quantile performance measures of simulation outputs. This allows us to identify a source of uncertainty in quantile estimates that may previously have been ignored, enabling simulation practitioners to make better-informed decisions.

We focused on the case where input models follow independent distributions and input modelling is done from a frequentist perspective. We adapt two methods of quantifying input uncertainty for the mean, a bootstrapping approach and a Taylor series approximation. The latter is only appropriate for parametric input distributions. We applied both methods to an analytical example which shows they accurately estimate an analytical approximation of the true value of input uncertainty. We also applied both methods to a stochastic activity network where they returned similar mean estimates of input uncertainty.

In the future, we should consider how to construct asymptotically valid confidence intervals for the quantile estimator, that account for both stochastic uncertainty and input uncertainty. This will help with the interpretation of input uncertainty for quantiles. We could also consider how other input uncertainty quantification techniques for the mean, which may offer benefits over both methods used here, could be adapted for quantiles. We should also investigate how input uncertainty estimates are impacted when using a smaller number of replications, which would violate the asymptotic relationship in Equation (7).

ACKNOWLEDGMENTS

This paper is based on work completed while Drupad Parmar was part of the EPSRC funded STOR-i Centre for Doctoral Training (EP/L015692/1).

REFERENCES

Asmussen, S., and P. W. Glynn. 2007. *Stochastic Simulation: Algorithms and Analysis*. Springer Science & Business Media.

Cheng, R. C., and W. Holland. 1997. “Sensitivity of Computer Simulation Experiments to Errors in Input Data”. *Journal of Statistical Computation and Simulation* 57(1-4):219–241.

Cheng, R. C., and W. Holland. 1998. “Two-point Methods for Assessing Variability in Simulation Output”. *Journal of Statistical Computation Simulation* 60(3):183–205.

Cheung, K., and S. Lee. 2005. “Variance Estimation for Sample Quantiles Using the *m* out of *n* Bootstrap”. *Annals of the Institute of Statistical Mathematics* 57(2):279–290.

Dong, H., and M. K. Nakayama. 2014. “Constructing Confidence Intervals for a Quantile Using Batching and Sectioning when Applying Latin Hypercube Sampling”. In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, and S. J. B. . J. A. Miller, 640–651. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Fu, M. C. 2006. “Gradient Estimation”. *Handbooks in Operations Research and Management Science* 13(1):575–616.

Fu, M. C., L. J. Hong, and J.-Q. Hu. 2009. “Conditional Monte Carlo Estimation of Quantile Sensitivities”. *Management Science* 55(12):2019–2027.

Hall, P., T. J. DiCiccio, and J. P. Romano. 1989. “On Smoothing and the Bootstrap”. *The Annals of Statistics* 17(2):692–704.

Hall, P., and M. A. Martin. 1988. "Exact Convergence Rate of Bootstrap Quantile Variance Estimator". *Probability Theory and Related Fields* 80(2):261–268.

Heidergott, B., and W. Volk-Makarewicz. 2009. "Quantile Sensitivity Estimation". In *International Conference on Network Control and Optimization*. November 23rd-25th, Eindhoven, The Netherlands, 16-29.

Hong, L. J. 2009. "Estimating Quantile Sensitivities". *Operations Research* 57(1):118–130.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lei, L., Y. Peng, M. C. Fu, and J.-Q. Hu. 2018. "Applications of Generalized Likelihood Ratio Method to Distribution Sensitivities and Steady-State Simulation". *Discrete Event Dynamic Systems* 28(1):109–125.

Lin, Y., E. Song, and B. L. Nelson. 2015. "Single-Experiment Input Uncertainty". *Journal of Simulation* 9(3):249–259.

Liu, G., and L. J. Hong. 2009. "Kernel Estimation of Quantile Sensitivities". *Naval Research Logistics (NRL)* 56(6):511–525.

Nakayama, M. K. 2014. "Confidence Intervals for Quantiles Using Sectioning when Applying Variance-Reduction Techniques". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 24(4):1–21.

Nelson, B. 2013. *Foundations and Methods of Stochastic Simulation: A First Course*. Springer Science & Business Media.

Serfling, R. J. 2009. *Approximation Theorems of Mathematical Statistics*, Volume 162. John Wiley & Sons.

Shao, J., and C. J. Wu. 1989. "A General Theory for Jackknife Variance Estimation". *The Annals of Statistics* 17(3):1176–1197.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. J. Buckley, and J. A. Miller, 162–176. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Xie, W., B. Wang, and Q. Zhang. 2018. "Metamodel-Assisted Risk Analysis for Stochastic Simulation with Input Uncertainty". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1766–1777. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Zhu, H., T. Liu, and E. Zhou. 2020. "Risk Quantification in Stochastic Simulation Under Input Uncertainty". *ACM Transactions on Modeling and Computer Simulation (TOMACS)* 30(1):1–24.

AUTHOR BIOGRAPHIES

DRUPAD PARMAR is a Ph.D. student at the Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry at Lancaster University. His research is focused on input uncertainty quantification in stochastic simulation models. His email address is d.parmar1@lancaster.ac.uk. His website is https://www.lancaster.ac.uk/∼parmard1/.

LUCY E. MORGAN is an AI and Optimisation Research Specialist in Applied Research at BT and a Visiting Researcher in the Department of Management Science at Lancaster University. Her research interests are input uncertainty quantification, arrival process modelling and simulation analytics. Her e-mail address is l.e.morgan@lancaster.ac.uk.

SUSAN M. SANCHEZ is a Distinguished Professor of Operations Research at the Naval Postgraduate School, and Co-Director of the Simulation Experiments & Efficient Design (SEED) Center for Data Farming. She also holds a joint appointment in the Graduate School of Defense Management. She has a B.S. in Industrial & Operations Engineering from the University of Michigan, and a Ph.D. in Operations Research from Cornell. She has been an active member of the simulation community for many years, and has been recognized as a Titan of Simulation and an INFORMS Fellow. Her email address is ssanchez@nps.edu. Her web page is http://faculty.nps.edu/smsanche/.

ANDREW C. TITMAN received his Ph.D. from the University of Cambridge and currently is a Professor in Statistics in the Department of Mathematics and Statistics at Lancaster University. His research interests include survival and event history analysis and latent variable modelling, with applications in biostatistics and health economics. His e-mail address is a.titman@lancaster.ac.uk.

RICHARD A. WILLIAMS is a Senior Lecturer in Management Science at the Department of Management Science, Lancaster University. He has a Ph.D. in Computer Science from the University of York. His research focuses on complex systems science, with particular emphasis on cybernetics and agent-based modelling and simulation to further our understanding of complex dynamical social systems. His email address is r.williams4@lancaster.ac.uk.