Proceedings of the Operational Research Society Simulation Workshop 2021 (SW21)
M. Fakhimi, D. Robertson, and T. Boness, eds.

A TWO STAGE ALGORITHM FOR GUIDING DATA COLLECTION TOWARDS MINIMISING INPUT UNCERTAINTY

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

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

Prof. Susan M. Sanchez

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

ABSTRACT

In stochastic simulation the input models used to drive the simulation are often estimated by collecting data from the real-world system. This can be an expensive and time consuming process so it would therefore be useful to have some guidance on how much data to collect for each input model. Estimating the input models via data introduces a source of variance in the simulation response known as input uncertainty. In this paper we propose a two stage algorithm that guides the initial data collection procedure for a simulation experiment that has a fixed data collection budget, with the objective of minimising input uncertainty in the simulation response. Results show that the algorithm is able to allocate data in a close to optimal manner and compared to two alternative data collection approaches returns a reduced level of input uncertainty.

Keywords: Input Uncertainty, Data Collection, Budget Allocation

1 INTRODUCTION

The randomness in stochastic simulation is caused by input models which are often represented by probability distributions or processes. When the real-world processes can be observed, samples of data can be collected and used to estimate the input models. The samples of data from which to estimate the input models are finite and thus the input models will never be truly representative of reality. The uncertainty in the estimated input models is propagated through the simulation model resulting in an error in the simulation response known as input uncertainty. Input uncertainty must be quantified, along with stochastic estimation error, to measure the variability around the simulation response and ensure that decisions are made with an appropriate level of confidence; Barton (2012) illustrates the significant risk of ignoring input uncertainty. A reduction in input uncertainty can only be achieved by collecting additional observations from the real-world processes. One way this is done is by studying the contribution made to input uncertainty by each of the input models and specifying how best to allocate a budget for additional data collection amongst the input models.

Here instead of looking at additional data collection to reduce input uncertainty we introduce the idea of guiding the initial data collection process in a manner that minimises input uncertainty. We consider the case of parametric input models and by assuming some knowledge of what values the parameters may take, we develop a two stage algorithm that allocates observations amongst the input models with the objective of minimising input uncertainty. Collecting data in this way is likely to reduce input uncertainty, and hence the level of variability, in the simulation response compared to alternative approaches, thus increasing the level of insight that can be derived from experimental results. This will lessen the need for additional data collection in order to reduce input uncertainty and may also

reduce unnecessary data collection more generally, both of which are particularly beneficial when data collection is expensive and time consuming.

We discuss background literature in Section 2 and detail the existing methodology we build upon in Section 3. We present a new breakdown of the existing methodology in Section 4, which allows us to form and solve optimisation problems which minimise input uncertainty. We describe the two stage algorithm for data collection in Section 5 and illustrate the algorithm with some experiments in Section 6. We discuss some open research questions in Section 7 and then conclude in Section 8.

2 BACKGROUND

Various methods have been proposed to quantify input uncertainty, for an overview of existing techniques see Barton (2012) or Song et al. (2014). We focus on the methodology developed by Cheng and Holland (1997) for the case of parametric input distributions. Here, input model uncertainty reduces to parameter uncertainty and is modelled using a first order Taylor series expansion around the true input parameters. A recent development to this approach was made by Lin et al. (2015), who exploit the gradient estimation method of Wieland and Schmeiser (2006), to estimate input uncertainty in a single experiment. Although initially restricted to the case of stationary input distributions, further work by Morgan et al. (2016) has since extended this input uncertainty quantification method for simulation models which utilise piecewise-constant non-stationary Poisson arrival processes.

The problem of allocating resources for extra data collection was considered by Ng and Chick (2001), who use asymptotic normality properties to approximate the posterior distribution of each parameter. By considering the expected information of additional observations and propagating uncertainty through the simulation using a linear metamodel, they provide sampling plans for further data allocation which aim to reduce input uncertainty effectively. Alternatively Freimer and Schruben (2002) use an ANOVA test to detect whether a parameter has a significant effect on the expected simulation response as the parameter varies over its confidence interval. If the effect is significant then more data should be collected to narrow the confidence interval until the effect is no longer significant. Finally Song and Nelson (2015) model the expected simulation response as a function of the mean and variance of each input model, and use the sample size sensitivity of each distribution to recommend how to collect further data. These methods aim to guide data collection based on input models that have been estimated using real-world observations however our method aims to guide data collection before any real-world observations have been collected. We now describe an existing input uncertainty quantification technique that we will utilise within our approach for guiding data collection.

3 TAYLOR SERIES APPROXIMATION

Consider a simulation driven by L random processes which follow known independent parametric distributions with unknown parameters. Let the unknown but true parameters be denoted by $\boldsymbol{\theta}^c = (\theta_1^c, \ldots, \theta_p^c)$, where $p \geq L$ as some distributions may require more than one parameter. Suppose that real-world data can be collected from each input distribution and that parameters are estimated via their maximum likelihood estimators (MLEs). Let $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \ldots, \hat{\theta}_p)$ denote the MLEs of the input parameters given the observed data. In this parametric setup the simulation response can be thought of as a function of the input parameters and the output of replication j of the simulation can be denoted by

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

where $\eta(\hat{\boldsymbol{\theta}})$ is the expected value of the simulation output given the estimated parameters and ε is a random variable with mean 0 representing stochastic noise.

The goal of the simulation experiment is to estimate $\eta(\boldsymbol{\theta}^c)$, the expected value of the simulation output given the true input parameters. This can be estimated via the sample mean of the simulation output over n replications

$$\begin{aligned}\bar{Y}(\hat{\boldsymbol{\theta}}) &= \frac{1}{n}\sum_{j=1}^{n} Y_j(\hat{\boldsymbol{\theta}}), \\ &= \eta(\hat{\boldsymbol{\theta}}) + \frac{1}{n}\sum_{j=1}^{n} \varepsilon_j(\hat{\boldsymbol{\theta}}).\end{aligned}$$

As n increases the error caused by stochastic noise tends towards 0 however the impact of $\hat{\boldsymbol{\theta}}$ on the expected simulation response is not affected by the choice of n. The variance of this estimator breaks down into two distinct terms, stochastic estimation error and input uncertainty. The former arises from the random variates generated in each replication and can be easily estimated via the sample variance. The latter measures the variability in the expected output due to having estimated the input parameters, that is

$$\sigma_I^2 = \text{Var}[\eta(\hat{\boldsymbol{\theta}})].$$

Using a first-order Taylor series approximation around the true input parameters $\boldsymbol{\theta}^c$, Cheng and Holland (1997) provide the following estimate of input uncertainty

$$\sigma_I^2 \approx \nabla\eta(\boldsymbol{\theta}^c)\text{Var}(\hat{\boldsymbol{\theta}})\nabla\eta(\boldsymbol{\theta}^c)^\top,$$

where $\nabla\eta(\boldsymbol{\theta}^c)$ is the gradient of the expected value of the simulation output with respect to the input parameters $\boldsymbol{\theta}$, evaluated at $\boldsymbol{\theta}^c$. This estimate of input uncertainty depends on how sensitive the simulation output is to the input parameters and how well the input parameters have been estimated. Neither of these terms are known and so both have to be estimated. Note that this method for quantifying input uncertainty has been extended for the case of non-stationary input models by Morgan et al. (2016).

3.1 Variance Estimation

As the parameters are estimated via maximum likelihood estimation, the variance matrix can be approximated by

$$\widehat{\text{Var}}(\hat{\boldsymbol{\theta}}) = \mathbf{I}(\hat{\boldsymbol{\theta}})^{-1},$$

the inverse Fisher information matrix evaluated at the MLEs $\hat{\boldsymbol{\theta}}$. This follows since the asymptotic distribution of the MLEs is multivariate normal with covariance matrix $\mathbf{I}(\boldsymbol{\theta}^c)^{-1}$, and $\mathbf{I}(\boldsymbol{\theta}^c)^{-1}$ can be consistently estimated by $\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}$.

3.2 Gradient Estimation

As the true parameters $\boldsymbol{\theta}^c$ are unknown, Cheng and Holland (1997) approximate $\nabla\eta(\hat{\boldsymbol{\theta}})$ instead of $\nabla\eta(\boldsymbol{\theta}^c)$, however the simulation effort for the method they propose increases linearly with the number of input parameters. Lin et al. (2015) improve upon this by providing a method for estimating $\nabla\eta(\hat{\boldsymbol{\theta}})$ which is independent of the number of input parameters. This method, which extends the work of Wieland and Schmeiser (2006), requires running simulation replications using the fitted input parameters and recording the simulation output along with internal parameter estimates for each replication. Internal parameter estimates are obtained using the realisations generated from the input distributions during a simulation replication, for example inter-arrival times observed within a replication can provide an internal estimate of the arrival rate. Fitting a least-squares regression model, with the simulation output as the response variable and the internal parameter estimates as the explanatory variables, gives a regression model whose coefficients $\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}})$ provide an estimator to $\nabla\eta(\hat{\boldsymbol{\theta}})$.

3.3 Contributions to Input Uncertainty

Input uncertainty can then be approximated by combining the estimates for the variance matrix and the gradient vector as follows

$$\sigma_I^2 \approx \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}})\,\mathbf{I}(\hat{\boldsymbol{\theta}})^{-1}\,\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}})^\top.$$

This approximation also provides us with an estimate of the contribution made to input uncertainty by each input distribution. Let $\boldsymbol{\theta}_l$ denote the parameter vector for input distribution l, note that this could be a scalar or a vector depending on the distribution. Since the input distributions are independent the variance matrix has a block diagonal form with elements consisting of individual variance matrices $\text{Var}[\hat{\boldsymbol{\theta}}_l]$ for each input distribution. Let $\hat{\boldsymbol{\delta}}(\boldsymbol{\theta}_l)$ denote the gradient vector for the parameters belonging to input distribution l, then

$$\sigma_I^2 \approx \sum_{l=1}^{L} \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)\mathbf{I}(\hat{\boldsymbol{\theta}}_l)^{-1}\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)^\top,$$

where the l^{th} term in the sum represents the contribution made to input uncertainty by input distribution l. This breakdown can be used to show which input distributions should be targeted for further data collection in order to reduce input uncertainty, for example see Lin et al. (2015).

4 DATA COLLECTION FOR MINIMISING INPUT UNCERTAINTY

We now present a new breakdown of the Taylor series approximation to input uncertainty which we propose as a tool for guiding data collection. The Fisher information for an i.i.d. sample of size m is simply m times the Fisher information for a single observation. Let m_l denote the number of observations used to estimate the parameters for input distribution l. The Fisher information matrix of $\hat{\boldsymbol{\theta}}_l$ is then given by

$$\mathbf{I}(\hat{\boldsymbol{\theta}}_l) = m_l \mathbf{I}_0(\hat{\boldsymbol{\theta}}_l),$$

where $\mathbf{I}_0$ represents the Fisher information of a single observation. Let $m = \sum_{l=1}^{L} m_l$ denote the total number of observations used to estimate all input distribution parameters. For each input distribution l, we can write $m_l = r_l m$, where $r_l \in (0,1)$ represents the proportion of all observations which are from input distribution l, and $\sum_{l=1}^{L} r_l = 1$. Input uncertainty can then be written as

$$\sigma_I^2 \approx \frac{1}{m} \sum_{l=1}^{L} \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)[r_l \mathbf{I}_0(\hat{\boldsymbol{\theta}}_l)]^{-1} \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)^\top. \tag{1}$$

Initially let us consider a set of specific parameter values $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)$. For these parameters the Fisher information matrix can be calculated and the gradient vector can be estimated using the method outlined above. Plugging these into (1) would give an approximation of input uncertainty at $\boldsymbol{\theta}$ in terms of the total number of observations m, and the proportions r_l in which they are allocated to each input distribution. If $\boldsymbol{\theta}$ were the set of true input parameters we could use this to guide data collection by finding the proportions in which to collect data such that input uncertainty is minimised. Note that the proportions which will minimise input uncertainty are invariant to the total number of observations.

We are therefore interested in solving the following optimisation problem

$$\{\min \sum_{l=1}^{L} \frac{a_l}{r_l} \quad \text{s.t.} \quad \sum_{l=1}^{L} r_l = 1 \quad \text{and} \quad r_l > 0 \text{ for } l = 1, \ldots, L\}, \tag{2}$$

where $a_l = \hat{\boldsymbol{\delta}}(\boldsymbol{\theta}_l)\mathbf{I}_0(\boldsymbol{\theta}_l)^{-1}\hat{\boldsymbol{\delta}}(\boldsymbol{\theta}_l)^\top$ and r_l are the proportions to be optimised. This problem can be converted to an inequality-constrained nonlinear programme and solved to optimality by studying the first-order KKT conditions proved by Karush (1939) and Kuhn and Tucker (1951). The optimal proportions are given by

$$r_l = \sqrt{\frac{a_l}{\left(\sum_{l=1}^{L} \sqrt{a_l}\right)^2}}.$$

Alternatively suppose that input parameters $\hat{\boldsymbol{\theta}}$ have been fitted via a collection of real-world data and that input uncertainty has been quantified via the Taylor series approximation. If input uncertainty is a cause for concern then it may be of interest to collect more real-world data in a manner which effectively reduces input uncertainty. This is often done by considering a sampling budget B for extra data collection. If we wish to collect data in a manner such that the overall proportions minimise input uncertainty then there is an extra consideration to be made. Given the budget for collecting extra data and the existing data collected, there is a restriction on how small each proportion can be, that is, each proportion will have a lower bound.

We are now interested in solving the following optimisation problem

$$\{\min \sum_{l=1}^{L} \frac{a_l}{r_l} \quad \text{s.t.} \quad \sum_{l=1}^{L} r_l = 1 \quad \text{and} \quad r_l \geq b_l \text{ for } l = 1, \ldots, L\}, \tag{3}$$

where $a_l = \hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)\mathbf{I}_0(\hat{\boldsymbol{\theta}}_l)^{-1}\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}}_l)^\top$, $b_l = m_l/(m+B)$ and r_l are the proportions to be optimised. Again this problem can be converted to an inequality-constrained nonlinear programme and solved to optimality

using first-order KKT conditions. The optimal proportions are given by finding a partition (I,J) of $\{1,\ldots,L\}$ such that

$$\text{for } l \in I:\ r_l = \sqrt{\frac{a_l}{\mu}} \geq b_l, \qquad \text{for } l \in J:\ r_l = \sqrt{\frac{a_l}{\mu - \lambda_l}} \geq b_l, \quad \lambda_l = \mu - \frac{a_l}{b_l^2} \geq 0,$$

where

$$\mu = \left(\frac{\sum_{l \in I} \sqrt{a_l}}{1 - \sum_{l \in J} b_l} \right)^2.$$

We use the solutions to these two optimisation problems to develop an algorithm that aims to guide the initial data collection process in a manner that minimises input uncertainty. Note that additional constraints could be added to either formulation to incorporate features of the data collection procedure.

5 TWO STAGE ALGORITHM FOR DATA COLLECTION

When modelling some systems, for example medical practices or manufacturing processes, collecting data to estimate the input models may be an expensive and time consuming task. In these scenarios, one may wish to consider a strategy for data collection rather than taking some arbitrary approach. Here we introduce a two stage algorithm to guide data collection. We assume that each input parameter lies in some known interval and study how data might be optimally collected to minimise input uncertainty within these intervals. In the first stage of the algorithm data is collected to hone in towards an optimal collection whilst relaying information about the true parameters values. In the second stage extra data is collected to achieve the proportions that minimise input uncertainty based on the parameter values from the first stage data collection.

Suppose there has been no data collection for a system. Although we have no data from which to estimate the input parameters we shall assume that each input parameter is known to lie in some interval, $\theta_i^c \in [l_i, u_i]$ for $i = 1,\ldots,p$, where the lower and upper bounds, l_i and u_i, are known. For example in a medical practice the number of patients arriving may be known to be between 15-20 per hour, but the exact arrival rate is unknown. The true input parameters could lie anywhere in this p dimensional space and in order to collect data in a way that minimises input uncertainty we need to understand how the optimal proportion changes for each input parameter across the space. An intuitive design that can be used to explore the input parameter space is a 2^k factorial design, which is used to study the effects of k factors each at two different levels (usually high and low) by considering every possible combination of factors and levels. Since there are p input parameters and each is known to be between a lower and upper bound, this naturally lends itself to a 2^p factorial design where each factor is an input parameter with low level l_i and high level u_i.

At each design point we can solve (2) to find the optimal proportions in which to collect data should the parameters at that design point represent the true parameters. Computing the optimal proportions at each design point will give us an idea of the behaviour of the optimal proportion for each input parameter over the specified parameter space. Rather than studying the effects of each parameter, we are instead interested in the minimum and maximum optimal proportion across the design points for each parameter. We use these to form an approximate interval for the optimal proportion for each parameter at the true parameter values. For example, suppose that $p = 2$ so the parameter vector is $\boldsymbol{\theta} = (\theta_1, \theta_2)$. A 2^p factorial design gives $2^2 = 4$ design points which enumerate every combination of factors and levels. Suppose that a 2^2 factorial design gives the optimal proportions shown in Table 1. From this we approximate that the optimal proportions for the true parameters will fall within the following intervals: $r_1 \in [0.3, 0.5]$ and $r_2 \in [0.5, 0.7]$.

Design Point i	θ_1^i	θ_2^i	r_1^i	r_2^i
1	l_1	l_2	0.5	0.5
2	u_1	l_2	0.3	0.7
3	l_1	u_2	0.4	0.6
4	u_1	u_2	0.4	0.6

Table 1: Example optimal proportions for a 2^2 factorial design

Suppose we have a budget B for collecting observations which is to be allocated amongst all of the input distributions. In stage one we aim to allocate as much of the budget as possible without ruling out the true optimal proportions which could occur anywhere within the limits of our parameter space. By allocating the budget according to the minimum optimal proportion for each parameter we can find out information regarding the true parameter values without ruling out any proportions which lie within the approximate intervals. For the example under discussion this would mean allocating 0.3 of the budget to estimating θ_1 and 0.5 of the budget to estimating θ_2, utilising 0.8 of the budget. Collecting data according to this allocation would give us information about the parameters whilst ensuring that any proportions within the intervals can still be achieved by allocating the remainder of the budget. Using the data collected in stage one we can calculate the MLEs, Fisher information matrix and estimate the gradient vector. We can then solve (3) to find the optimal proportions according to the parameter estimates gained from the first stage data collection, using the existing data to set the lower bounds. The remaining budget can then be allocated in order to achieve these proportions and guide the second stage data collection. Putting all these steps together gives us the following algorithm.

Algorithm 1: Two Stage Algorithm for Data Collection

Result: First stage data allocation $m_{\boldsymbol{\theta}_1,1}, m_{\boldsymbol{\theta}_2,1}, \ldots, m_{\boldsymbol{\theta}_L,1}$;
Second stage data allocation $m_{\boldsymbol{\theta}_1,2}, m_{\boldsymbol{\theta}_2,2}, \ldots, m_{\boldsymbol{\theta}_L,2}$;
Initialise two factorial design;
for *each design point i* **do**
  Compute $I_0(\boldsymbol{\theta}^i)$ and $\hat{\boldsymbol{\delta}}(\boldsymbol{\theta}^i)$;
  Find $r_1^i, r_2^i, \ldots, r_L^i$ by solving (2);
end
for *each input model l* **do**
  $r_{l,\min} = \min_i r_l^i$;
  $m_{\boldsymbol{\theta}_l,1} = B \times r_{l,\min}$;
end
Collect data according to first stage allocation;
Compute $\hat{\boldsymbol{\theta}}$, $I_0(\hat{\boldsymbol{\theta}})$ and $\hat{\boldsymbol{\delta}}(\hat{\boldsymbol{\theta}})$;
Find $r_1, r_2, \ldots, r_L$ by solving (3) using lower bounds $r_{1,\min}, r_{2,\min}, \ldots, r_{L,\min}$;
for *each input model l* **do**
  $m_{\boldsymbol{\theta}_l,2} = B \times r_l$;
end
Collect remaining data to achieve second stage allocation;

6 EXPERIMENTS

In this section we illustrate the algorithm on two examples. We first use an $M/M/1$ queueing model to compare the final allocation of data from the two stage algorithm with the true optimal allocation. Secondly using a more realistic simulation model we compare input uncertainty estimates given by the two stage algorithm against two commonly used approaches for data collection.

6.1 $M/M/1$ Queueing Model

To experiment with the two stage algorithm we use first use an $M/M/1$ queueing model since closed-form expressions can be found for many performance measures. We measure the mean queueing time and since we are able to derive the gradient measures analytically we can calculate the true optimal proportions in which to collect data such that input uncertainty is minimised. To evaluate the performance of the two stage algorithm we can compare the final proportions in which the data is allocated to the true optimal proportions which minimise input uncertainty.

To implement the two stage algorithm let us assume that the input parameters are known to fall within the following intervals: $\lambda^c \in [3,6]$ and $\mu^c \in [9,12]$. We run the simulation for 1000 time periods and the budget for data collection is set to $B = 1000$ observations. Within this controlled experiment we

can set true parameters and use these to generate synthetic data, here we use a Latin hypercube sample with 10 intervals to generate 10 sets of true parameters within the parameter space. At each of set of parameters we run the two stage algorithm 100 times, recording the final recommended proportions in which the data is allocated in each of the experiments. Figure 1 shows boxplots of the final proportion of data allocated to λ (r_1) for each set of input parameters. The red dots indicate the true optimal proportion for each set of parameters calculated using the analytical gradient measures. For each set of parameters the boxplot of proportions from the two stage algorithm is concentrated around the true optimal proportion, demonstrating that the two stage algorithm is able to hone in towards an optimal collection of data. We expect to see some variation around the true optimal proportions for two reasons. Firstly the final proportion of data allocated to λ is based upon the parameter estimates from the first stage data collection and these will not to be equivalent to the true parameters since we only have a finite budget. Secondly the gradient terms are estimated and hence will differ from the true gradient measures. Although variability is evident these results are promising.

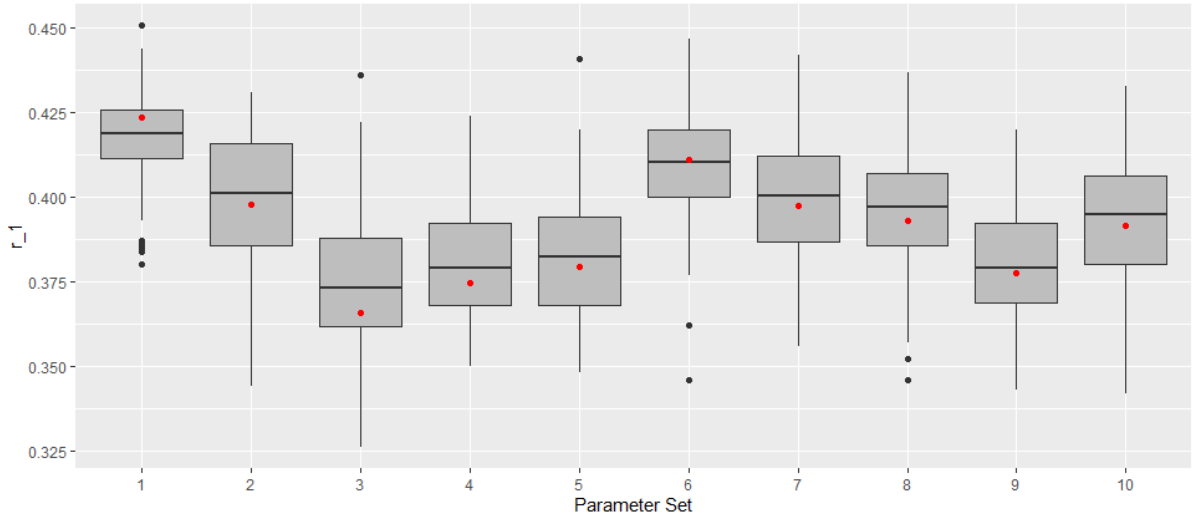

Figure 1: Boxplots showing 100 final proportions for λ given by the two stage algorithm compared to the true optimal proportion at 10 different sets of parameters

6.2 Network Queueing Model

We now consider a more realistic simulation model, a network queueing model consisting of three consecutive multi-server queues. Entities arriving at the system join the queue at node 1. After receiving service at node 1 an entity may leave the system or join the queue at node 2, and similarly after receiving service at node 2 an entity may leave the system or join the queue at node 3. After service at node 3 an entity departs the system. This model could represent a medical centre for example and may be used to solve problems relating to capacity planning and resource allocation. Systems such as this are referred to as operational models of healthcare units in Brailsford (2007).

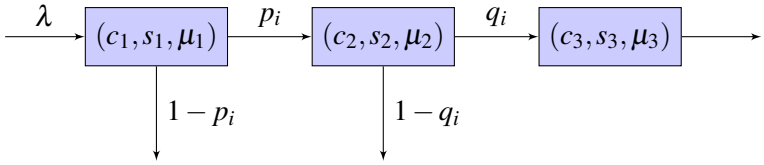

Figure 2: A graphical representation of the network queueing model

The network queueing model is set up as follows. At nodes 1, 2 and 3 there are c_1, c_2, and c_3 servers respectively each of which have a shifted exponential service distribution with parameters (s_1, μ_1), (s_2, μ_2) and (s_3, μ_3). Arrivals to the system follow a stationary Poisson process with rate λ. To represent different demographics of the population the arrivals are split into three different types; 50%

of arrivals are of type A, 30% are type B, and 20% are type C. Each type refers to how likely an entity is to travel through the system and so each is defined by a set of probabilities (p_i, q_i) representing the probability of continuing to node 2 and node 3 respectively. Finally let us suppose the performance measure of interest is the average queueing time weighted by type. For simplicity, we shall assume the the shift parameter of each service distribution and the routing probabilities for each type are known. The unknown input model parameters that require estimation are therefore the arrival rate λ and the three service rates μ_1, μ_2, μ_3. To implement the two stage algorithm we assume that the input parameters are known to fall within the following intervals: $\lambda^c \in [12, 16]$, $\mu_1^c \in [18, 22]$, $\mu_2^c \in [8, 12]$, and $\mu_3^c \in [6, 10]$. The remaining information about the system is known and is as follows: there are two servers at each node $(c_1, c_2, c_3) = (2, 2, 2)$, the shift parameters for the service distributions are $(s_1, s_2, s_3) = (0.05, 0.05, 0.1)$, and the routing probabilities for types A, B, and C are $(p_A, q_A) = (0.4, 0.4)$, $(p_B, q_B) = (0.7, 0.7)$, and $(p_C, q_C) = (0.9, 0.9)$ respectively.

We now evaluate the performance of the two stage algorithm against other data collection approaches by comparing the input uncertainty passed to the simulation response. One alternative we consider is the equal observations approach, where the same amount of data is collected to estimate each input model. This may be the case in a simple service system where arrivals and services are recorded consecutively. To implement this approach we can simply split the budget equally amongst the input models and generate observations from each true input distribution. The second alternative we consider is the timed observation approach, where data is collected by observing the true system over some set period of time. An example of this can be found in Griffiths et al. (2005) where an intensive care unit model was developed using data taken over the course of a year. By using the true parameters to run our simulation model for some chosen period of time we can imitate collecting data from a timed observation of the real-world system.

Within our experiment we wish to compare input uncertainty estimates given by the three approaches when using the same budget. Since the timed observation approach has no fixed number of observations we run this first for 250 time periods. The total number of observations gathered from this approach is then used as the budget for the two stage algorithm and for the equal observations approach. We generate 5 random sets of true parameters so we can compare the approaches across the parameter space when the optimal proportions of the true parameters vary. For each of set of parameters we run the three approaches 100 times. Input uncertainty estimates for each approach are estimated using the same amount of simulation effort and are recorded in the boxplots in Figure 3.

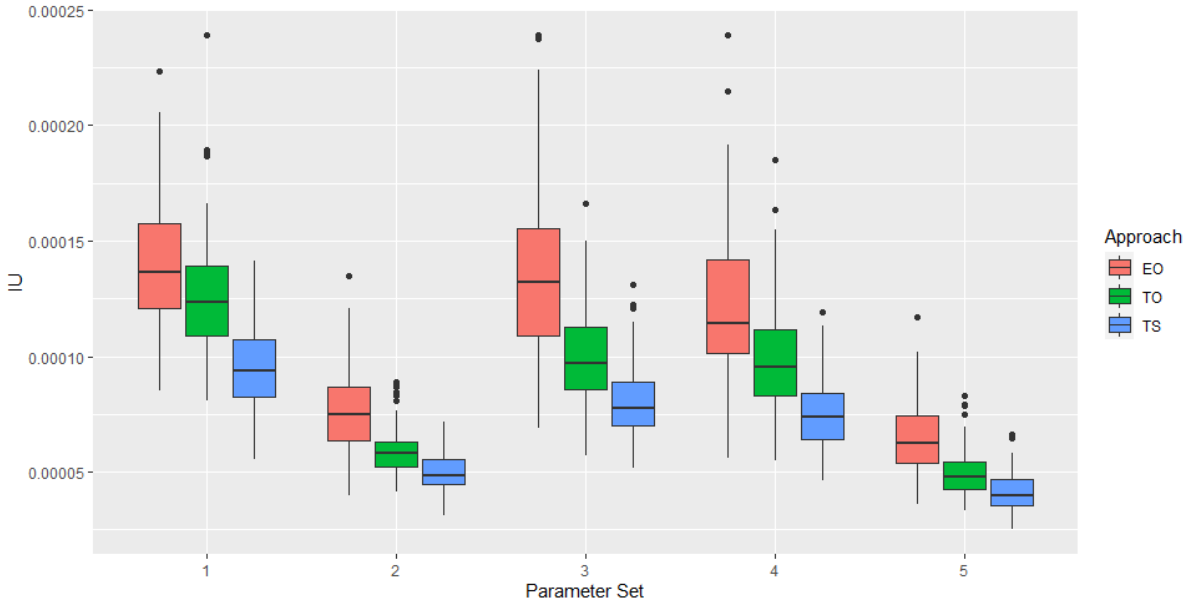

Figure 3: Boxplots comparing 100 estimates of input uncertainty given by the two stage algorithm (TS) compared to equal observations (EO) and timed observation (TO) at 5 different sets of parameters

For each parameter set the two stage algorithm reduced the mean input uncertainty between 31.44% and 40.81% compared to the equal observation approach, and between 16.05% and 24.23% compared

to the timed observation approach. For each set of parameters the two stage algorithm allocates more observations to the arrival rate compared to the other two approaches and by doing so reduces input uncertainty despite using the same overall number of observations. The two stage algorithm shows that by using some knowledge of what values the input parameters might take we can collect data for our input models in a manner that effectively reduces the input uncertainty of our performance measure. A caveat here is that both the equal observation and timed observation approach require no prior knowledge of input parameter values and can be completed in a single collection.

7 FUTURE RESEARCH

The two stage algorithm assumes that each input parameter lies within some known interval however in reality it is unlikely that such an interval is known with complete confidence. Although experts and practitioners may be able provide estimates for such intervals these will only be approximations and therefore we cannot be certain that the true parameters will lie within the intervals. It is worth noting however that the two stage algorithm works regardless of the values that the true parameters take. The second stage allocation minimises input uncertainty for the MLEs calculated from the first stage collection regardless of whether the MLE for each parameter lies in its interval or not. The minimisation however is constrained by the first stage data allocation which is based upon the optimal proportions found from the two factorial design over the parameter space. The issue with true parameters taking values outside their specified interval is that the optimal proportions for these parameters without any constraints may not be achievable as the first stage data collection could rule them out. Preliminary results show that the two stage algorithm still provides a reduction in input uncertainty when parameters lie up to half an intervals width outside their interval, however further research is required here.

Another area which requires further investigation relates to the width of the parameter intervals. Although very wide intervals are more likely to contain the true parameter, in queueing style simulation models they may lead to design points in which the system doesn't reach steady state. In general the design points will represent extreme scenarios in which certain parameters may require barely any data collection and hence have very small optimal proportions. Consequently the minimum optimal proportion for some parameters may be very small which can lead to a small first stage data allocation.

We currently impose no constraint on how large the first stage data allocation needs to be for each parameter. If the minimum optimal proportion for a parameter is small, or the budget multiplied by the minimum optimal proportion is small, then the first stage data allocation will suggest collecting very few observations for that parameter which is likely to lead to an inaccurate parameter estimate. If this is the case then the proportions calculated using the first stage data collection, which are the target proportions, will minimise input uncertainty at a point in the parameter space which may be far away from where the true parameters lie. Consequently the proportions in which the data is collected may differ greatly from the optimal proportions at the true parameters. A possible solution is to introduce a minimum allocation level, meaning that at least a minimum number of observations for each parameter must be allocated in the first stage in order to obtain reasonable parameter estimates.

8 CONCLUSION

In this paper we introduced the novel idea of allocating an initial budget for data collection in a manner that minimises input uncertainty. In particular we have developed an algorithm that by collecting data in two different stages aims to hone in on an optimal allocation of data across the input models. Using an $M/M/1$ queueing model we have demonstrated that the algorithm achieves an allocation of data that is close to the true optimal allocation. On a more realistic simulation model we have shown that the two stage algorithm results in a reduced level of input uncertainty compared to two other viable approaches for data collection. Further experimentation needs to be conducted to gain a deeper understanding of the performance when parameters lie outside their specified intervals, as well as when the parameter intervals are extremely wide.

ACKNOWLEDGEMENTS

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

Barton, R. R. 2012. "Tutorial: Input uncertainty in outout analysis". In *Proceedings of the 2012 Winter Simulation Conference*. IEEE.

Brailsford, S. C. 2007. "Tutorial: Advances and challenges in healthcare simulation modeling". In *Proceedings of the 2007 Winter Simulation Conference*. IEEE.

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.

Freimer, M., and L. Schruben. 2002. "Collecting data and estimating parameters for input distributions". In *Proceedings of the 2002 Winter Simulation Conference*. IEEE.

Griffiths, J. D., N. Price-Lloyd, M. Smithies, and J. E. Williams. 2005. "Modelling the requirement for supplementary nurses in an intensive care unit". *Journal of the Operational Research Society* 56 (2): 126–133.

Karush, W. 1939. "Minima of functions of several variables with inequalities as side constraints". *M. Sc. Dissertation. Dept. of Mathematics, Univ. of Chicago*.

Kuhn, H. W., and A. W. Tucker. 1951. "Nonlinear Programming". In *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*: University of California Press, Berkeley.

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

Morgan, L. E., A. C. Titman, D. J. Worthington, and B. L. Nelson. 2016. "Input uncertainty quantification for simulation models with piecewise-constant non-stationary poisson arrival processes". In *Proceedings of the 2016 Winter Simulation Conference*. IEEE.

Ng, S. H., and S. E. Chick. 2001. "Reducing input parameter uncertainty for simulations". In *Proceedings of the 2001 Winter Simulation Conference*. IEEE.

Song, E., and B. L. Nelson. 2015. "Quickly assessing contributions to input uncertainty". *IIE Transactions* 47 (9): 893–909.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced tutorial: Input uncertainty quantification". In *Proceedings of the 2014 Winter Simulation Conference*. IEEE.

Wieland, J. R., and B. W. Schmeiser. 2006. "Stochastic gradient estimation using a single design point". In *Proceedings of the 2006 Winter Simulation Conference*. IEEE.

AUTHOR BIOGRAPHIES

DRUPAD PARMAR is a PhD student at the Statistics and Operational Research Centre for Doctoral Training in Partnership with Industry at Lancaster University. His research interests are stochastic simulation and input uncertainty quantification. His email address is d.parmar1@lancaster.ac.uk.

LUCY E. MORGAN is a Lecturer in Simulation and Stochastic Modelling in the Department of Management Science at Lancaster University. Her research interests are input uncertainty in simulation models and arrival process modelling. Her e-mail address is l.e.morgan@lancaster.ac.uk.

ANDREW C. TITMAN is a Senior Lecturer 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. 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.

SUSAN M. SANCHEZ is a Professor of Operations Research at the Naval Postgraduate School with a joint appointment in the Graduate School of Business & Public Policy. Her interests include the design and analysis of large-scale simulation experiments, robust design, and applied statistics, with application to military operations, manufacturing, and health care. Her email address is ssanchez@nps.edu.