Proceedings of the 2017 Winter Simulation Conference
W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, eds.

IMPROVING PREDICTION FROM STOCHASTIC SIMULATION VIA MODEL DISCREPANCY LEARNING

Henry Lam
Matthew Plumlee
Xinyu Zhang

Department of Industrial and Operations Engineering
University of Michigan
1205 Beal Ave.
Ann Arbor, MI 48109, USA

ABSTRACT

Stochastic simulation is an indispensable tool in operations and management applications. However, simulation models are only approximations to reality, and typically bear discrepancies with the generating processes of real output data. We investigate a framework to statistically learn these discrepancies under the presence of data on past implemented system configurations, which allows us to improve prediction using simulation models. We focus on the case of general continuous output data that generalizes previous work. Our approach utilizes (a combination of) regression analysis and optimization formulations constrained on suitable summary statistics. We demonstrate our approach with a numerical example.

1 INTRODUCTION

Stochastic simulation is used ubiquitously in decision analytics in many operations and management applications. It describes system dynamics under alternate configurations that can be used for analytical tasks. While modelers often attempt to build models as realistic as possible, due to resource constraints and ignorance of unobserved system features, these models are at best approximations and in almost all practical cases bear discrepancies with reality. For example, in operational settings that are naturally cast as "queues", the serial structures of interarrival times (e.g., Livny et al. 1993), discretionary or strategic behaviors (e.g., balking, reneging, queue selection; Pazgal and Radas 2008, Veeraraghavan and Debo 2009), individualized and contextual behaviors of the customers and servers (e.g., Aksin et al. 2007), and structural server changes over time (e.g., Brown et al. 2005) are all difficult to capture. Inadequate reflection of these features can degrade the predictive power of the simulation model and potentially affect the reliability of decision-making.

In this paper, we investigate a general framework that builds on our previous work Plumlee and Lam (2016) to learn about the imperfectness of stochastic simulation models and subsequently use this information to improve prediction. To describe our investigation, we first introduce some notations and define the notion of model discrepancy. Imagine that we are interested in an output variable, $Y \in \mathscr{Y}$, from a stochastic system which has a true distribution $\pi(\cdot)$. The variable can represent waiting time, queue length, reward, other responses or any collections of them that are relevant to decision-making. A typical performance analysis consists of evaluating a quantity $\rho(\pi)$, which in this paper we focus on the expectation-type performance measure $E_\pi[f(Y)]$ for some function $f(\cdot) : \mathscr{Y} \to \mathbb{R}$.

When real-world observations from π are abundant, the performance analyses can be accurately approximated by replacing π with data, e.g., the empirical distribution. However, in decision analyses (e.g., optimality or feasibility tests), we are typically interested in system designs that are sparsely sampled or even never adopted before. So, instead, model builders use "simulation models" that generate outputs with distribution, say, $\tilde{\pi}(\cdot)$, as an approximation to $\pi(\cdot)$ that can be computed much more easily via simulation replications. If $\tilde{\pi}(\cdot) \neq \pi(\cdot)$, then $\rho(\tilde{\pi})$ may not equal $\rho(\pi)$ and the outcomes of the analysis become erroneous. Our interest is to combine both the observations on π and the simulation model $\tilde{\pi}$ to generate prediction for $\rho(\pi)$ that is better than using either one only.

Our approach hinges on estimating the difference between $\tilde{\pi}(\cdot)$ and $\pi(\cdot)$ and then integrating this bias estimate with the simulation model to predict $\rho(\pi)$. It requires the existence of some control variable or design point $x \in \mathscr{X}$ that represents the set of system parameters, and that real system observations are available at some x. The value of a design point can include decision variables (e.g., number of servers, routing policy) and important components of the system (e.g., arrival rate of customers). The availability of observations from different design points provides an opportunity to learn and correct for the imperfectness of $\tilde{\pi}$, including at un- or under-observed points, by data-pooling.

We consider the class of mappings $\mathscr{P} : \mathscr{X} \times \mathscr{Y} \to \mathbb{R}$, such that if $p \in \mathscr{P}$, then $p(x,\cdot)$ is a valid probability measure over $\mathscr{Y}$. We expand the definition of the true model π and the simulation model $\tilde{\pi}$ to be in $\mathscr{P}$. In general, we have the pointwise relationship

$$\pi(x,y) = \tilde{\pi}(x,y) + \delta(x,y) \tag{1}$$

where we call $\delta(\cdot,\cdot)$ the *model discrepancy*. When $\tilde{\pi} \neq \pi$, we have $\delta \neq 0$ and model discrepancy is present.

The relationship (1) can be viewed as a stochastic counterpart of a similar discrepancy notion in the literature of deterministic computer experiments (where no y in (1) is present). The latter is under the framework of model calibration that refers generally to the refinement or correction of a hypothetical simulation model through real-world data to enhance the model prediction capability. Starting with the seminal statistical article of Kennedy and O'Hagan (2001), calibration for deterministic simulation has been substantially investigated (e.g., Bayarri et al. 2012, Oakley and O'Hagan 2004, Higdon et al. 2004, Plumlee 2016). In the stochastic simulation literature, other than parameter calibration (implemented in some simulation software such as Anylogic), calibration on model structure is commonly conducted together with model validation, through iterative checking (e.g., by statistical tests) and re-building of the simulation model until a satisfactory model is obtained (Balci and Sargent 1982, Sargent 2005, Sargent 2013, Kleijnen 1995). This paper detours from these common suggestions, and utilizes the idea of inferring model discrepancies in deterministic computer experiments to improve the prediction of stochastic simulation. Compared to the established practice, the avoidance of building increasingly sophisticated models could save costs and time in refining the model, and prevent the unfortunate situation in case an ultimately satisfactory model is absent.

We will consider two methods in this paper. The first one estimates $\rho(\delta(x,\cdot)) := \rho(\pi(x,\cdot)) - \rho(\tilde{\pi}(x,\cdot))$ directly by regression against the covariate x. The second method exploits further the fact that π and $\tilde{\pi}$ are probability distributions, so that they satisfy standard conditions for valid probability measures that constrain the possible values of $\rho(\pi)$. This second method can be thought of as inferring the difference of distributions $\delta(x,\cdot) = \pi(x,\cdot) - \tilde{\pi}(x,\cdot)$ and using it to calculate $\rho(\pi)$. However, since $\delta(x,\cdot)$ can be a high- or even infinite-dimensional object (in the case of continuous output distributions) that is challenging to estimate, our second approach instead operates on a collection of summary statistics on $\delta(x,\cdot)$, and posits suitable optimization formulations to compute bounds that take into account the relation between these summary statistics induced by the probability distribution structure.

As mentioned before, this work builds on Plumlee and Lam (2016) and generalizes their investigation. Besides the difference regarding the representation of model discrepancy (additive versus multiplicative form), Plumlee and Lam (2016) focuses on finite-valued outputs in a Bayesian framework. In this paper, we are primarily interested in outputs that are continuous, which pose substantial additional challenges since they are now infinite-dimensional objects. Our aforementioned second method novelly uses summary statistics to facilitate tractable statistical estimation, but with infinite-dimensional optimization programs posited over the space of probability distributions to recover bounds incurred by the continuous distributions. With well-chosen functions to define the summary statistics, these optimization programs can be reformulated by duality as tractable finite-dimensional programs. These developments provide an implementable mechanism to take into account distributional information that is infinite-dimensional in nature.

In the rest of this paper, we first describe the two proposed methods in Sections 2 and 3. Then we provide an extensive numerical study and compare our performance with both simulation-only and data-only predictions in Section 4. We conclude this paper in Section 5.

2 LEARNING MODEL DISCREPANCY VIA REGRESSION

Consider a collection of design point values, say $\{1,\ldots,s\}$. For each value $x=i$, suppose there are i.i.d. real-world output observations $Y_{ij}\sim\pi(i,\cdot), j=1,\ldots,n_i$. The sample size n_i can be possibly zero for some i (so that no Y_{ij} exists).

We assume there is enough computational resource, so that the simulation output distribution $\tilde{\pi}$ can be exactly obtained.

Suppose we are interested in $\rho(\pi(x_0,\cdot))=E_{\pi(x_0,\cdot)}[f(Y)]$ for some x_0. Using data alone, we will output $\bar{Y}_{x_0\cdot}=(1/n_{x_0})\sum_{j=1}^{n_{x_0}}f(Y_{ij})$. Clearly, if there is no data at x_0, this approach does not apply. On the other hand, the simulation model gives $\rho(\tilde{\pi}(x_0,\cdot))=E_{\tilde{\pi}(x_0,\cdot)}[f(Y)]$ as the predicted value of $\rho(\pi(x_0,\cdot))$.

To combine both $\{Y_{ij}\}$ and $\tilde{\pi}$ in our prediction, we consider the estimation of $\rho(\delta(x_0,\cdot))$ by the regression problem

$$f(Y_{ij})-\rho(\tilde{\pi}(i,\cdot))=h(i)+\varepsilon_{ij} \tag{2}$$

where $h(\cdot)$ is an unknown function of the design point x, and $\varepsilon_{ij}, j=1,\ldots,n_i$ are mean-zero i.i.d. noises from the real data, for each i. The responses in the regression are $f(Y_{ij})-\rho(\tilde{\pi}(i,\cdot))$ whose mean values are the model discrepancies $\rho(\delta(i,\cdot))$.

Regression (2) can be conducted by a variety of tools; we use local linear regression with a standard R package in our experiments. This gives us a confidence interval (CI) for $\rho(\delta(x_0,\cdot))$, say $[L,U]$. Our prediction of $\rho(\pi(x_0,\cdot))$ is then expressed as a CI $[\rho(\tilde{\pi}(x_0,\cdot))+L,\rho(\tilde{\pi}(x_0,\cdot))+U]$. Clearly, if $[L,U]$ indeed covers the true $\rho(\delta(x_0,\cdot))$ with probability $1-\alpha$, then $[\rho(\tilde{\pi}(x_0,\cdot))+L,\rho(\tilde{\pi}(x_0,\cdot))+U]$ covers $\rho(\pi(x_0,\cdot))$ with the same probability.

The above scheme can be modified with

$$f(Y_{ij})=\beta\rho(\tilde{\pi}(i,\cdot))+h(i)+\varepsilon_{ij} \tag{3}$$

The additional unknown scaling parameter β represents a multiplicative model discrepancy. The response in (3) now has mean $\rho(\pi(i,\cdot))$. The CI for $\rho(\pi(x_0,\cdot))$ can be obtained directly from the regression (3).

3 COMBINING REGRESSION OUTCOMES WITH OPTIMIZATION

We consider an enhancement of the method in Section 2 by introducing the auxiliary statistics $\gamma_k(\pi(x,\cdot))=E_{\pi(x,\cdot)}[g_k(Y)]$ for $k=1,\ldots,m$. One of these statistics can be taken as $\rho(\pi(x,\cdot))$, e.g., let $g_1=f$. Then, instead of using (2) only, we impose the multivariate-output regression

$$g_k(Y_{ij})-\gamma_k(\pi(i,\cdot))=h_k(i)+\varepsilon_{ijk} \tag{4}$$

where $h_k(\cdot)$ is the individual regression function for the k-th statistic. Similar to Section 2, $\varepsilon_{ijk}, j=1,\ldots,n_i$ are mean-zero i.i.d. noises, for each i, and $g_k(Y_{ij})-\gamma_k(\pi(i,\cdot))$ has mean equal to $\gamma_k(\delta(i,\cdot))$.

The idea is that estimates of $\gamma_k(\delta(x_0,\cdot)), k=1,\ldots,m$ give more information than using $\rho(\delta(x_0,\cdot))$ alone. For convenience, we denote $\gamma=(\gamma_k)_{k=1,\ldots,m}$ and $g=(g_k)_{k=1,\ldots,m}$. To incorporate the new information, we first need a confidence region (CR) for $\gamma(\delta(x_0,\cdot))$. As in Section 2, this can be done by a variety of statistical tools, and here we use a basic scheme of running individual local regression at each k, making an additional assumption that $\varepsilon_{ijk}, k=1,\ldots,m$ are independent for each i,j. For each individual regression, we can obtain the point estimate of $\gamma_k(\delta(x_0,\cdot))$, say $\hat{\gamma}_k$, and its standard error, say s_k. Under standard assumption we have $\hat{\gamma}_k \overset{approx.}{\sim} N(\gamma_k(\delta(x_0,\cdot)),\sigma_k^2)$, independent among the k's, where σ_k^2 is the sampling variance. This gives us a CR

$$\mathscr{U} = \left\{(z_1,\ldots,z_m)\in\mathbb{R}^m : \sum_{k=1}^{m}\frac{(z_k-\hat{\gamma}_k)^2}{s_k^2}\le\chi^2_{1-\alpha,m}\right\} \tag{5}$$

for $\gamma(\delta(x_0,\cdot))$, where $\chi^2_{1-\alpha,m}$ is the $(1-\alpha)$-quantile of a χ^2-distribution with degree of freedom m.

We then impose the optimization problems

$$\max\rho(q) \text{ subject to } \gamma(q)-\gamma(\tilde{\pi}(x_0,\cdot))\in\mathscr{U} \tag{6}$$

and

$$\min\rho(q) \text{ subject to } \gamma(q)-\gamma(\tilde{\pi}(x_0,\cdot))\in\mathscr{U} \tag{7}$$

where the decision variable q is a probability distribution on the response space $\mathscr{Y}$. The constraints in (6) and (7) incorporate information on the possible values of $\gamma(q)$ with probability $1-\alpha$. Given this information, the outcomes of (6) and (7) give the best upper and lower bounds on $\rho(\pi(x_0,\cdot))$. If the CR $\mathscr{U}$ contains the true $\gamma(\tilde{\pi}(x_0,\cdot))$ with probability $1-\alpha$, then (6) and (7) will form a CI for $\rho(\pi(x_0,\cdot))$ with level at least $1-\alpha$. This idea originates from the literature of data-driven distributionally robust optimization (e.g., Delage and Ye 2010, Ben-Tal et al. 2013, Goh and Sim 2010, Wiesemann et al. 2014), where $\mathscr{U}$ is often known as the uncertainty set or the ambiguity set.

Note that (6) and (7) are generalized moment problems, where the optimization contains an objective function and constraint functions that are all moments of a random variable. By using conic duality, (6) can be reformulated as the dual problem

$$\begin{array}{ll} \min_{\kappa,\nu\in\mathbb{R},\lambda\in\mathbb{R}^m} & \kappa+(\gamma(\tilde{\pi}(x_0,\cdot))+\hat{\gamma})^{\mathsf{T}}\lambda+\sqrt{\chi^2_{1-\alpha,m}}\nu \\ \text{subject to} & \kappa+g(y)^{\mathsf{T}}\lambda-f(y)\ge 0 \text{ for all } y\in\mathscr{Y} \\ & \|(s)^{\mathsf{T}}\lambda\|_2\le\nu \end{array} \tag{8}$$

where $\hat{\gamma}=(\hat{\gamma}_k)_{k=1,\ldots,m}$ and $s=(s_k)_{k=1,\ldots,m}$, and similarly for (7) (by considering $-\max\{-\rho(\pi)\}$). Strong duality holds if $\gamma(q)-\gamma(\tilde{\pi}(x_0,\cdot))$ lies in the interior of $\mathscr{U}$ and under appropriate topological assumptions on $\mathscr{Y}$ (e.g., Shapiro 2001). Without the condition, weak duality still implies that (8) provides a conservative approximation for (6).

Note that (8) still has an infinite number of constraints, but for specific choices of f and g one can reduce (8) to finite-dimensional optimization problems. We focus here on the setting that $\mathscr{Y}=[0,\mathrm{UB}]\subset\mathbb{R}$ where UB is the largest possible value of Y, and $f(y)$ and $g_k(y)$ are in the form y^r where r is a rational number. This allows us to reduce (8) to finite-dimensional semidefinite programs (SDP) by generalizing the technique in Bertsimas and Popescu (2005), which consider moment equalities only and does not contain the second order cone constraint in the dual formulation (8). For example, in the case that $f(y)=y$ and $g_k(y)=y^{k/\tilde{m}}$ for $k=1,2,\cdots,m$, (8) becomes

$$\begin{array}{ll} \min_{\kappa,\nu\in\mathbb{R},\lambda\in\mathbb{R}^m} & \kappa+(\gamma(\tilde{\pi}(x_0,\cdot))+\hat{\gamma})^{\mathsf{T}}\lambda+\sqrt{\chi^2_{1-\alpha,m}}\nu \\ \text{subject to} & \kappa+\sum_{i=1}^{m}\lambda_i y^{i/\tilde{m}}-y\ge 0 \text{ for all } y\in[0,\mathrm{UB}] \\ & \|(s)^{\mathsf{T}}\lambda\|_2\le\nu \end{array} \tag{9}$$

By a change of variable $y^{1/\tilde{m}} \to y$, we can transform the algebraic constraints into positive semidefinite constraints, and get

$$\begin{array}{ll} \min_{\kappa,\nu\in\mathbb{R},\lambda\in\mathbb{R}^m,U\in\mathbb{R}^{(m+1)\times(m+1)}} & \kappa+(\gamma(\tilde{\pi}(x_0,\cdot))+\hat{\gamma})^{\mathsf{T}}\lambda+\sqrt{\chi^2_{1-\alpha,m}}\nu \\ \text{subject to} & U\succeq 0 \\ & \sum_{i,j:i+j=2l+1} u_{ij}=0, \quad l=1,\cdots,m \\ & \sum_{i,j:i+j=2l+2} u_{ij}=\sum_{r=0}^{l} y_r\binom{\tilde{m}-r}{l-r}\mathrm{UB}^{r/\tilde{m}}, \quad l=0,\cdots,m \\ & \|(s)^{\mathsf{T}}\lambda\|_2\leq\nu \end{array} \tag{10}$$

where $y_0=\kappa, y_i=\lambda_i\mathbf{1}_{i\neq\tilde{m}}+(\lambda_i-1)\mathbf{1}_{i=\tilde{m}}, [U]_{ij}=u_{ij}$. Similarly, for (7), the reformulation remains as (10) except that $y_{\tilde{m}}$ is replaced by $\lambda_{\tilde{m}}+1$.

We choose the power in $g_k(\cdot)$ as $k/\tilde{m}$ instead of integers so as to avoid a blow-up in the magnitude of the moment as k increases. Note that there are other plausible choices of statistics, e.g., quantile-type statistics in the form $g_k(y)=I(y\leq b_k)$ for some b_k, or a combination of these and the power moments, which can also be similarly converted into SDP.

As in Section 2, an alternative is to consider the regression

$$g_k(Y_{ij})=\beta_k\gamma_k(\tilde{\pi}(i,\cdot))+h_k(i)+\varepsilon_{ijk}, \tag{11}$$

with responses $g_k(Y_{ij})$ and additional scaling parameters β_k. We can form a CR $\mathscr{U}$ for $\gamma(\pi(x_0,\cdot))$ in this case as (5), where $\hat{\gamma}_k$ is now the point estimate of $\gamma_k(\pi(x_0,\cdot))$ under the regression and s_k is the corresponding standard error. Then we use the optimization problems

$$\max\rho(q) \text{ subject to } \gamma(q)\in\mathscr{U}$$

and

$$\min\rho(q) \text{ subject to } \gamma(q)\in\mathscr{U}$$

The maximization can be dualized to (8) and in the special case considered above to (10) (without the $\gamma(\tilde{\pi}(x_0,\cdot))$ term). Similar treatments for the minimization as discussed above also hold.

We note that the accuracy of our proposed approach, here and in Section 2, depends on the complexity of the simulation model and the reality, in that it determines how challenging it is to find an acceptable regression model for capturing the model discrepancy. In the next section, we will test our approach with some numerical studies on a simple queueing system.

4 NUMERICAL STUDY

We use a queueing example to test the approaches we propose in Sections 2 and 3. We first consider a simulation model and the "real" system generated as follows: The simulation model is a $M/M/x$ queue with arrival rate $\lambda=5$ and service rate $\mu=0.05$. The real system is a $M/M/x$ queue with a random arrival rate as an absolute value of a normally distributed random variable with mean 5 and standard deviation $5\cdot 0.1$, and a random service rate as an absolute value of a normally distributed random variable with mean 0.045 and standard deviation $0.045\cdot 0.1$ (Think of these as, e.g., some random daily characteristics). In addition, each customer has a probability 0.2 to abandon the queue if the waiting time is larger than an exponential random variable with mean $\frac{450}{x}$ where x is the number of servers. The output quantity of interest Y is the average waiting time among the first 10 customers.

The design point x corresponds to the number of servers, which ranges on $x=10,11,\cdots,25$. For this example, we can easily run plenty of simulation, say 10^5 replications, to get $\tilde{\pi}$ that can be considered as exactly obtained. For each $x=10,\ldots,24$, we generate 10 observations on the real system. We deliberately collect no observations for $x=25$ as a test point.

We use $f(y)=y$ and $g_k(y)=y^{k/\tilde{m}}$, so that we can use the SDP (10) when adopting the optimization-enhanced approaches. We use a confidence level $1-\alpha=95\%$ in the experiment.

We consider five methods:

- Data-only method (Method D) only utilizes real data Y_{ij}. For $i=1,\ldots,24$, the CI is

$$\left[\bar{Y}_i - z_{1-\alpha/2}\frac{\hat{\sigma}_i}{\sqrt{10}}, \bar{Y}_i + z_{1-\alpha/2}\frac{\hat{\sigma}_i}{\sqrt{10}}\right],$$

  where $\bar{Y}_i=\frac{\sum_{j=1}^{10} Y_{ij}}{10}$, $\hat{\sigma}_i=\sqrt{\frac{\sum_{i=1}^{10}(Y_{ij}-\bar{Y}_i)^2}{10-1}}$, and $z_{1-\alpha/2}$ is the $(1-\alpha/2)^{th}$ quantile of a standard normal distribution. For $x=25$, there is no data so Method D does not apply.
- Regression-only method 1 (Method R1) is based on the regression problem (2) in Section 2.
- Optimization-enhanced method 1 (Method O1) is based on optimizing over the CR obtained from (4) in Section 3.
- Regression-only method 2 (Method R2) is based on (3) in Section 2.
- Optimization-enhanced method 2 (Method O2) is based on optimizing over the CR obtained from (11) in Section 3.
- Simulation-only method (Method S) uses the (accurate) point estimate of our simulation model.

Each method except Method S outputs a lower bound and a upper bound value for each experiment, thereby forming a $(1-\alpha)$-level confidence interval. For Methods R1, O1, R2, and O2, we use the R function "npreg" in the "np" package to make local linear regressions.

We first investigate the coverage probabilities of the CI generated by the methods. Table 1 shows the estimated coverage probabilities from repetition of 100 experiments with $g_k(y)=y^{k/10}, k=1,\cdots,10$. For each entry in the table, we use $a\pm b$ to denote the CI of the estimated coverage probability from the 100 experiments, i.e., a is the average number of success in covering the true value and b is the half-width of the associated CI for the coverage probability.

For Method D, the coverage probability ranges from 0.816 to 0.999. It is roughly around the theoretical coverage probability 95% but there are mild fluctuations due to the small number of repetitions. That being said, both regression and optimization methods obtain more fluctuated values as the number of servers varies. For example, the coverage probability ranges from 0.372 to 0.999 for Method R1 and ranges from 0.493 to 1 for Method O1. Therefore, the data-only method shows a more reliable performance overall. Even though in some cases (e.g., when $x=20$) the correction methods (i.e., Methods R1, O1, R2, O2) all obtain reasonable coverages, the data-only method has uniformly better performances and is the best in other cases (e.g., $x=10$). However, if one considers the case $x=25$ (i.e., the extreme case where there is no data), then Method D does not apply, but the correction methods still give reasonable coverage (though the performances vary). That the correction methods tend to have poorer coverage probabilities could be due to the crudeness of the regression models, whose assumptions and post-hoc analyses have not been carefully conducted in this experiment. In other words, the regression models used in this example could deviate from the true behavior of $\rho(\delta(x,\cdot))$ or $\rho(\pi(x,\cdot))$.

Table 1: Coverage probabilities for different methods

Number of servers	Method D	Method R1	Method O1	Method R2	Method O2
10	0.88±0.064	0.47± 0.098	0.59±0.097	0.57±0.098	0.66±0.093
11	0.91±0.056	0.59± 0.097	0.63±0.095	0.67±0.093	0.77±0.083
12	0.96±0.039	0.67± 0.093	0.78±0.082	0.81±0.077	0.88±0.064
13	0.88±0.064	0.77± 0.083	0.83±0.074	0.90±0.059	0.93±0.050
14	0.89±0.062	0.76± 0.084	0.79±0.080	0.90±0.059	0.90±0.059
15	0.91±0.056	0.79± 0.080	0.78±0.082	0.98±0.028	0.96±0.039
16	0.93±0.050	0.85± 0.070	0.91±0.056	0.94±0.047	0.96±0.039
17	0.88±0.064	0.91± 0.056	0.95±0.043	0.98±0.028	0.97±0.034
18	0.93±0.050	0.93± 0.050	0.91±0.056	0.95±0.043	0.93±0.050
19	0.92±0.053	0.95± 0.043	0.97±0.034	0.90±0.059	0.94±0.047
20	0.92±0.053	0.96± 0.039	0.96±0.039	0.99±0.020	0.97±0.034
21	0.95±0.043	0.96± 0.039	0.97±0.034	0.95±0.043	0.94±0.047
22	0.95±0.043	0.90± 0.059	0.89±0.062	0.99±0.020	0.95±0.043
23	0.91±0.056	0.89± 0.062	0.88±0.064	0.99±0.020	0.95±0.043
24	0.95±0.043	0.89± 0.062	0.92±0.053	0.87±0.066	0.87±0.066
25	-	0.81± 0.077	0.70±0.090	0.87±0.066	0.73±0.087

Besides coverage probability, we are interested in gaining some understanding on how conservative the methods are. This can be measured by the difference between the lower and upper confidence bounds for each method. We call these differences the prediction gaps. Table 2 shows the statistics of this gap for each of the methods, among the 100 experiments with $g_k(y) = y^{k/10}, k = 1, \cdots, 10$. For each entry in the table except the ones in the last column, we use $a \pm b$ to denote the CI of the prediction gap from the 100 experiments (i.e., a is the average gap, and b is the CI half-width). The last column shows the absolute difference between the point estimate of the simulation model and the truth, which can be viewed as the magnitude of the model error made in using the simulation.

Table 2: Prediction gaps for different methods

Number of servers	Method D	Method R1	Method O1	Method R2	Method O2	Method S
10	7.488 ± 0.355	1.820 ± 0.263	2.035 ± 0.242	2.660 ± 0.205	2.943 ± 0.171	0.413
11	6.005 ± 0.346	1.561 ± 0.190	1.670 ± 0.155	2.631 ± 0.176	2.884 ± 0.161	0.348
12	5.527 ± 0.275	1.325 ± 0.150	1.486 ± 0.127	2.322 ± 0.125	2.489 ± 0.094	0.236
13	4.751 ± 0.249	1.315 ± 0.140	1.467 ± 0.133	2.127 ± 0.085	2.360 ± 0.077	0.191
14	4.490 ± 0.220	1.334 ± 0.134	1.371 ± 0.103	1.979 ± 0.040	2.188 ± 0.050	0.137
15	3.878 ± 0.208	1.172 ± 0.087	1.247 ± 0.083	1.890 ± 0.031	2.035 ± 0.041	0.041
16	3.582 ± 0.177	1.143 ± 0.087	1.184 ± 0.086	1.797 ± 0.030	1.933 ± 0.030	0.027
17	2.987 ± 0.144	1.038 ± 0.053	1.060 ± 0.049	1.713 ± 0.037	1.808 ± 0.031	0.018
18	2.791 ± 0.138	1.000 ± 0.057	0.981 ± 0.050	1.627 ± 0.051	1.679 ± 0.038	0.059
19	2.533 ± 0.125	0.981 ± 0.046	0.973 ± 0.047	1.527 ± 0.065	1.557 ± 0.046	0.093
20	2.274 ± 0.108	0.958 ± 0.043	0.945 ± 0.045	1.579 ± 0.067	1.470 ± 0.047	0.101
21	2.112 ± 0.098	0.906 ± 0.026	0.886 ± 0.029	1.489 ± 0.075	1.383 ± 0.054	0.140
22	1.831 ± 0.086	0.893 ± 0.025	0.849 ± 0.028	1.452 ± 0.081	1.306 ± 0.045	0.136
23	1.673 ± 0.088	0.885 ± 0.024	0.832 ± 0.030	1.522 ± 0.082	1.272 ± 0.048	0.143
24	1.501 ± 0.077	0.898 ± 0.019	0.806 ± 0.026	1.462 ± 0.084	1.165 ± 0.050	0.164
25	1.422 ± 0.067	0.999 ± 0.084	0.871 ± 0.078	1.514 ± 0.080	1.157 ± 0.067	0.162

Table 2 shows that the prediction gaps using correction methods are significantly shorter than that using data-only method for every x. For example, at $x = 18$, the prediction gap ranges from 2.653 to 2.929 for Method D, but ranges only from 0.943 to 1.057 for Method R1 and from 0.937 to 1.037 for Method O1. Moreover, at some particular x, both high coverage probability and small prediction gap are obtained simultaneously for correction methods. For instance, the coverage probabilities at $x = 18$ for both data-only and correction methods are roughly around 95% and the probabilities only differ by $5\% \sim 10\%$ for many choice of x such as $15, 16, 23, 24$. This shows that, while the coverage probabilities are more reliable for using only data, their predictions are generally much more conservative. The correction methods use information on the simulation model to improve the conservativeness.

Considering the last column in Table 2, we see that the simulation model errors are in overall smaller than the prediction gaps in this example. However, simulation method in this example is, in a sense, always erroneous as the Monte Carlo errors of an inaccurate model are washed away with abundant simulation runs, leaving point estimates that are different from the truth. We note that different simulation model will give different model error, and there is no direct guarantee whether the model error as we define will be smaller or larger than the prediction gaps from the correction methods.

Next we repeat the experiment, but using a smaller number, but higher orders, of moments $g_k(y) = y^{k/2}, k = 1, \cdots, 3$. Tables 3 and 4 give the coverage probabilities and prediction gaps respectively under this new scheme. Table 3 shows that coverage probabilities for regression methods and optimization methods are comparable to each other for every x. For example, for $x = 16$, the coverage probabilities differ by only roughly 0.02 between regression methods and optimization methods. Under the new scheme, the prediction gaps obtained from optimization methods in overall are shorter than that from regression methods. For instance, in Table 4, we have prediction gap ranging from 1.106 to 1.34 for Method R1 and from 1.047 to 1.237 for Method O1 at $x = 14$. Also, the prediction gap ranges from 1.609 to 1.731 for Method R2 and ranges from 1.417 to 1.539 for Method O2 at $x = 19$. Besides the shorter gaps, the optimization methods give prediction gaps with smaller variability, e.g., the CI length of the prediction gap is shorter for almost every x for Method O2 than that for Method R2.

Table 3: Coverage probabilities for different methods under a different set of moments

Number of servers	Method D	Method R1	Method O1	Method R2	Method O2
10	0.89 ± 0.06	0.39 ± 0.10	0.41 ± 0.10	0.54 ± 0.10	0.47 ± 0.10
11	0.93 ± 0.05	0.59 ± 0.10	0.59 ± 0.10	0.76 ± 0.08	0.71 ± 0.09
12	0.87 ± 0.07	0.70 ± 0.09	0.66 ± 0.09	0.87 ± 0.07	0.86 ± 0.07
13	0.91 ± 0.06	0.79 ± 0.08	0.75 ± 0.09	0.92 ± 0.05	0.90 ± 0.06
14	0.97 ± 0.03	0.82 ± 0.08	0.79 ± 0.08	0.94 ± 0.05	0.86 ± 0.07
15	0.90 ± 0.06	0.82 ± 0.08	0.79 ± 0.08	0.95 ± 0.04	0.93 ± 0.05
16	0.92 ± 0.05	0.88 ± 0.06	0.86 ± 0.07	0.97 ± 0.03	0.95 ± 0.04
17	0.92 ± 0.05	0.89 ± 0.06	0.85 ± 0.07	0.95 ± 0.04	0.94 ± 0.05
18	0.94 ± 0.05	0.99 ± 0.02	0.99 ± 0.02	0.95 ± 0.04	0.92 ± 0.05
19	0.90 ± 0.06	0.98 ± 0.03	0.98 ± 0.03	0.99 ± 0.02	0.97 ± 0.03
20	0.84 ± 0.07	0.93 ± 0.05	0.94 ± 0.05	0.99 ± 0.02	0.98 ± 0.03
21	0.93 ± 0.05	0.97 ± 0.03	0.97 ± 0.03	0.95 ± 0.04	0.94 ± 0.05
22	0.89 ± 0.06	0.93 ± 0.05	0.93 ± 0.05	0.98 ± 0.03	0.97 ± 0.03
23	0.87 ± 0.07	0.93 ± 0.05	0.93 ± 0.05	0.98 ± 0.03	0.97 ± 0.03
24	0.94 ± 0.05	0.83 ± 0.07	0.87 ± 0.07	0.95 ± 0.04	0.95 ± 0.04
25	–	0.74 ± 0.09	0.68 ± 0.01	0.77 ± 0.08	0.78 ± 0.01

Table 4: Prediction gaps for different methods under a different set of moments

Number of servers	Method D	Method R1	Method O1	Method R2	Method O2	Method S
10	7.165 ± 0.390	1.753 ± 0.238	1.581 ± 0.190	2.583 ± 0.176	2.377 ± 0.149	0.413
11	6.156 ± 0.293	1.498 ± 0.187	1.345 ± 0.132	2.314 ± 0.150	2.079 ± 0.124	0.348
12	5.544 ± 0.285	1.358 ± 0.157	1.288 ± 0.129	2.304 ± 0.115	2.074 ± 0.097	0.236
13	4.850 ± 0.260	1.295 ± 0.122	1.202 ± 0.098	2.146 ± 0.084	1.950 ± 0.068	0.191
14	4.468 ± 0.208	1.223 ± 0.117	1.142 ± 0.095	2.048 ± 0.056	1.812 ± 0.041	0.137
15	3.729 ± 0.162	1.229 ± 0.097	1.150 ± 0.077	1.885 ± 0.026	1.665 ± 0.026	0.041
16	3.637 ± 0.176	1.104 ± 0.082	1.041 ± 0.064	1.803 ± 0.030	1.583 ± 0.034	0.027
17	3.145 ± 0.184	1.082 ± 0.079	1.018 ± 0.060	1.754 ± 0.031	1.585 ± 0.039	0.018
18	2.920 ± 0.138	1.048 ± 0.066	0.999 ± 0.050	1.659 ± 0.047	1.492 ± 0.051	0.059
19	2.563 ± 0.126	0.977 ± 0.044	0.941 ± 0.035	1.670 ± 0.061	1.478 ± 0.061	0.093
20	2.244 ± 0.120	0.962 ± 0.039	0.922 ± 0.029	1.523 ± 0.065	1.362 ± 0.064	0.101
21	2.053 ± 0.102	0.917 ± 0.031	0.904 ± 0.026	1.557 ± 0.076	1.383 ± 0.071	0.140
22	1.879 ± 0.089	0.889 ± 0.025	0.875 ± 0.024	1.526 ± 0.083	1.373 ± 0.075	0.136
23	1.690 ± 0.077	0.867 ± 0.019	0.873 ± 0.025	1.523 ± 0.082	1.376 ± 0.073	0.143
24	1.565 ± 0.072	0.909 ± 0.026	0.897 ± 0.028	1.532 ± 0.078	1.378 ± 0.071	0.164
25	1.388 ± 0.073	1.017 ± 0.066	0.966 ± 0.005	1.519 ± 0.080	1.345 ± 0.008	0.162

In comparing different regression models (2) and (3), we see that correction methods based on (3) generally give higher coverage probabilities compared with those based on (2). For example, at $x = 12$, Method R1 obtains coverage probability 0.67 ± 0.093 and Method R2 obtains 0.81 ± 0.077 in Table 1, and 0.70 ± 0.09 and 0.87 ± 0.07 respectively in Table 3. Likewise, for optimization methods, Method O1 obtains coverage probability 0.78 ± 0.082 and Method O2 obtains 0.77 ± 0.083 in Table 1, and 0.66 ± 0.09 and 0.86 ± 0.07 respectively in Table 3. However, higher coverage probabilities seem to come with the price of wider prediction gaps. For instance, Method R1 gives prediction gaps 1.325 ± 0.015 and Method R2 gives 2.322 ± 0.0125 in Table 1, and 1.358 ± 0.157 and 2.304 ± 0.115 respectively in Table 3. Likewise, for optimization methods, Method O1 gives prediction gaps 1.486 ± 0.127 and Method O2 obtains 2.489 ± 0.094 in Table 1, and 1.288 ± 0.129 and 2.074 ± 0.097 respectively in Table 3. From these observations, there does not seem to be a concrete conclusion which regression model is better. One who prefers a higher coverage probability at the cost of a conservative prediction gap may plausibly choose the second regression model, and vice versa.

5 CONCLUSION

We investigate an approach based on a combination of regression and optimization to learn and correct for the model discrepancy of a simulation model with real-world observations. Our approach targets at continuous output variables that are not directly handleable by previous work due to their infinite-dimensional nature. One version of our approach regresses the objective function of interest on the observed model discrepancies, or the observations themselves, against the design points. Another version further regresses the moments of the model discrepancies or observations and uses a moment optimization to compute bounds that account for the distributional nature of the outputs. We present the regression details and how to reformulate and subsequently solve the optimization in the second version via duality and semidefinite programming. Our experimental results show that the data-only approach could give better coverage probability in the availability of data, but could be more conservative in terms of wider confidence interval length than our correction methods. Moreover, the data-only approach is only applicable for historically observed design points but trivially breaks down in the new design points. Our results also show that our correction methods improve pure simulation models that are misspecified, in that the simulation-only estimates are systematically different from the truth while our correction methods generate interval estimates that cover or are closer to it.

ACKNOWLEDGMENTS

We gratefully acknowledge support from the National Science Foundation under grants CMMI-1542020, CMMI-1523453 and CAREER CMMI-1653339.

REFERENCES

Aksin, Z., M. Armony, and V. Mehrotra. 2007. "The modern call center: A multi-disciplinary perspective on operations management research". *Production and Operations Management* 16 (6): 665–688.

Balci, O., and R. G. Sargent. 1982. "Some examples of simulation model validation using hypothesis testing". In *Proceedings of the 14th conference on Winter Simulation-Volume 2*, 621–629. Winter Simulation Conference.

Bayarri, M. J., J. O. Berger, R. Paulo, J. Sacks, J. A. Cafeo, J. Cavendish, C.-H. Lin, and J. Tu. 2012. "A framework for validation of computer models". *Technometrics*.

Ben-Tal, A., D. Den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. 2013. "Robust solutions of optimization problems affected by uncertain probabilities". *Management Science* 59 (2): 341–357.

Bertsimas, D., and I. Popescu. 2005. "Optimal inequalities in probability theory: A convex optimization approach". *SIAM Journal on Optimization* 15 (3): 780–804.

Brown, L., N. Gans, A. Mandelbaum, A. Sakov, H. Shen, S. Zeltyn, and L. Zhao. 2005. "Statistical analysis of a telephone call center: A queueing-science perspective". *Journal of the American statistical association* 100 (469): 36–50.

Delage, E., and Y. Ye. 2010. "Distributionally robust optimization under moment uncertainty with application to data-driven problems". *Operations research* 58 (3): 595–612.

Goh, J., and M. Sim. 2010. "Distributionally robust optimization and its tractable approximations". *Operations research* 58 (4-part-1): 902–917.

Higdon, D., M. Kennedy, J. C. Cavendish, J. A. Cafeo, and R. D. Ryne. 2004. "Combining field data and computer simulations for calibration and prediction". *SIAM Journal on Scientific Computing* 26 (2): 448–466.

Kennedy, M. C., and A. O'Hagan. 2001. "Bayesian calibration of computer models". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 63 (3): 425–464.

Kleijnen, J. P. 1995. "Verification and validation of simulation models". *European journal of operational research* 82 (1): 145–162.

Livny, M., B. Melamed, and A. K. Tsiolis. 1993. "The impact of autocorrelation on queuing systems". *Management Science* 39 (3): 322–339.

Oakley, J. E., and A. O'Hagan. 2004. "Probabilistic sensitivity analysis of complex models: A Bayesian approach". *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 66 (3): 751–769.

Pazgal, A. I., and S. Radas. 2008. "Comparison of customer balking and reneging behavior to queueing theory predictions: An experimental study". *Computers & Operations Research* 35 (8): 2537–2548.

Plumlee, M. 2016. "Bayesian calibration of inexact computer models". *Journal of the American Statistical Association* (just-accepted).

Plumlee, M., and H. Lam. 2016. "Learning stochastic model discrepancy". In *Winter Simulation Conference (WSC), 2016*, 413–424. IEEE.

Sargent, G. R. 2013. "Verification and validation of simulation models". *Journal of Simulation* 7 (1): 12–24.

Sargent, R. G. 2005. "Verification and validation of simulation models". In *Proceedings of the 37th conference on Winter simulation*, 130–143. winter simulation conference.

Shapiro, A. 2001. "On duality theory of conic linear problems". In *Semi-infinite programming*, 135–165. Springer.

Veeraraghavan, S., and L. Debo. 2009. “Joining Longer Queues: Information Externalities in Queue Choice”. *Manufacturing & Service Operations Management* 11 (4): 543–562.

Wiesemann, W., D. Kuhn, and M. Sim. 2014. “Distributionally robust convex optimization”. *Operations Research* 62 (6): 1358–1376.

AUTHOR BIOGRAPHIES

HENRY LAM is an Assistant Professor in the Department of Industrial and Operations Engineering at the University of Michigan (UM), Ann Arbor. He received his Ph.D. degree in statistics at Harvard University, and was an Assistant Professor in the Department of Mathematics and Statistics at Boston University before joining UM. His research focuses on stochastic simulation, risk analysis, and simulation optimization. His email address is khlam@umich.edu.

MATTHEW PLUMLEE is an Assistant Professor at the Department of Industrial and Operations Engineering at the University of Michigan. He received his Ph.D. degree from the School of Industrial and Systems Engineering at Georgia Tech. His interests lie in the interface of data and modeling, specifically in methods for experimentation and uncertainty quantification for complex systems. His email address is mplumlee@umich.edu.

XINYU ZHANG is a Ph.D. student in the Department of Industrial and Operations Engineering at the University of Michigan (UM), Ann Arbor. He obtained his bachelor’s degree in UM, majoring in physics and applied mathematics. His research interests include stochastic optimization and extreme event analysis. His email address is xinyuzh@umich.edu.