Proceedings of the 2013 Winter Simulation Conference
R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, eds.

LINKING STATISTICAL ESTIMATION AND DECISION MAKING THROUGH SIMULATION

Jin Fang
L. Jeff Hong

Department of Industrial Engineering and Logistics Management
The Hong Kong University of Science and Technology
Clear Water Bay, Kowloon, HONG KONG

ABSTRACT

Models that are built to help make decisions usually involve input parameters, which need to be estimated statistically using data. However, submitting these estimated parameters directly to the model may result in biased decisions because the estimated parameters are biased or the model is nonlinear. We propose a new parameter estimator called Simulation-Based Inverse Estimator (SBIE) to link the statistical estimation and decision making together. The linkage is achieved by simulating the model and adjusting the estimated parameters such that the adjusted parameters can adapt to the specific model. We prove that SBIE can provide us with consistent and unbiased decisions under some conditions and this result is supported by numerical experiments in queuing models.

1 INTRODUCTION

Models that describe system behaviors are often used in business analytics to make quantitative decisions. These models often involve parameters that need to be estimated statistically using data. Once the parameters are estimated, their estimates are plugged into the models to calculate system performances under different scenarios and to select good decisions. For instance, in call centers, staffing decisions are often made according to estimates of arrival patterns; in manufacturing companies, procurement decisions and production planning are often made based on estimates of future demand; and in financial investment firms, portfolios are often constructed according to estimates of expected returns and covariances of available financial instruments.

Estimation of parameters is a widely studied problem in the area of statistics. Many methods have been proposed to solve the problem, including maximum likelihood estimation (MLE), least square estimation, method of moments, etc. These estimators often possess some nice statistical properties, e.g., unbiasedness, minimal variance and consistency. When these estimators are plugged into the model, however, the resulted estimators of system performances may not inherit the same properties. For instance, we consider an M/M/1 queueing model where the arrival rate is estimated through a sample of interarrival times that are exponentially distributed and the service rate is set as 1. Suppose that the true value of the arrival rate is 0.9, then the average queue length of the steady state is 9 according to the model. The MLE of the arrival rate is on average 0.902 when the sample size is 500. If we plug this MLE of arrival rate into the model, we find that the the average queue length is 24.5 (i.e., with a relative bias of 172%), showing that the estimator is heavily biased. Even when we increase the sample size to 1000, the average is still 11.6 (i.e., with a relative bias of 29%). The performance estimates may further deteriorate if the arrival rate increases.

This phenomenon does not only exist in this M/M/1 example. Indeed, it is quite ubiquitous in business analytics when a model is used to make quantitative decisions. It occurs because the true values of the parameters are unknown and they are substituted by their estimators. A rigorously developed statistical

978-1-4799-2076-1/13/$31.00 ©2013 IEEE

estimator often possesses two types of properties, large-sample properties and finite-sample properties. Large-sample properties include consistency, strong consistency, asymptotic normality, etc. that describe how the estimator deviates from its true value as the sample size goes to infinity. These properties are often inherited by the model performances when they are continuous functions of the parameters. Finite-sample properties include unbiasedness, minimal variance, minimal mean square error, etc. that describe how the estimator deviates from its true value when the amount of data is limited. These properties, on the contrary, can hardly be inherited by the model performances. For instance, if the system performance is a nonlinear function of the parameters, parameter estimators that are unbiased and minimal variance may lead to an estimator of the system performance that is neither unbiased nor minimal variance. Technically, this is because the mean and variance of a function of a random variable typically do not equal to the function of the mean and variance of the random variable, i.e., $E[f(\theta)] \neq f(E[\theta])$ and $Var[f(\theta)] \neq f(Var[\theta])$. In many practical situations, however, we only have a limited amount of data and, thus, we are interested more in finite-sample properties of the system performances than in large-sample properties. For instance, in the example of call centers, one may have only a few weeks of past arrival data to make staffing decisions; in the example of manufacturing companies, one may have only a few months of past demand data to make procurement and production decisions; and in the example of financial investment firms, one may have at most a few years of past return data to make investment decisions.

The reason that models may deliver poor performance estimates is because parameter estimation and performance evaluation are considered separately. In this paper, we take a holistic view to link these two steps together and our goal is to estimate parameters so that, when the estimates are plugged into the model, the system performances are well behaved. Our approach works as follows: we first use a standard statistical method to estimate the input parameters and plug them into the model to obtain an estimate of the output performances. We call this approach a *two-step approach* and call its estimate of the output performances the *two-step estimator*. Note that, when the true values of input parameters are known, we may simulate a sample of input data to mimic our sample of observed input data and apply the two-step approach to evaluate the output performances on the simulated data. Through this way, we may build a functional relationship between the true values of the input parameters and the average output performances. We can then invert this function at the two-step estimator and use it as our estimate of the input parameters. We call this approach *simulation-based inversion estimation approach* and its estimator *simulation-based inversion estimator* (SBIE).

It is important to note that the SBIE approach is model adaptive, i.e., its estimators of input parameters rely on not only the input data but also the model. These estimators of input parameters may not be as good as estimators of more standard statistical methods, however, they lead to better estimates of model performances that we care about. In this paper we show that the performance estimators of the SBIE approach guarantee to have lower biases compared to those of the two-step approach. In particular, if the biases of the two-step approach are of the order of $n^{-\alpha}$ as in the cases of consistent estimation approaches, where α is a number greater than 0 (i.e., $\alpha = \frac{1}{2}$ in maximum likelihood estimation), the biases of the SBIE approach are of the order smaller than $n^{-2\alpha}$.

In this paper we show that the SBIE approach can also be applied to situations where consistent two-step estimators are not readily available. In many practical situations, data may be contaminated by the data collection process. For instance, demands of a product in a grocery store are often censored at the inventory level (i.e., the observed demand can never be more than the inventory level), and potential arrivals to a restaurant may be turned away by the long queue causing the interarrivals longer than what they should be. When data are contaminated, parameter estimations are often quite difficult. However, if we ignore the contamination, the resulted estimators are often inconsistent and biased. Using the SBIE approach, we can reproduce the data contamination when we simulate the input data and learn how estimates of output performances are distorted by the contamination process as well as the two-step estimation approach. We show that the SBIE approach can reduce the bias in this case from an order of one to an order smaller than $n^{-\alpha}$.

Another advantage of the SBIE approach is that it is easy to use. Once the input data can be simulated, the SBIE approach can be applied. One only needs to simulate the input data and apply the original two-step estimation process multiple times and maybe at multiple values of the parameters to find the SBIE estimator. However, the advantages of the SBIE approach come with a cost on computation, as multiple (and often a large number of) simulation runs are often needed and an optimal solution needs to be searched. Nevertheless, with today's widespread availability of computing capacity, the computational requirement of the SBIE approach is rarely an issue, and it offers a solution which may be far more inexpensive when compared to the solution of collecting more data.

Literature Review

The study of parameter uncertainties in decision models is an important research area of operations research and management science. In the stochastic simulation literature, how to account for the parameter estimation errors has always been a very important research problem. The problem is often formulated as the construction of a valid confidence interval for output performance when there are uncertainties in parameter estimation in literature; see, for instance, **?** for a review. Another common approach to handle parameter uncertainties is the robustness approach, which assumes that parameters are constrained in a set (e.g., a confidence set) and we are interested in the worst performance of the model when parameters are in the set. For instance, **?** and **?** studied optimization problems with various uncertainty sets. Similar idea has also been applied to simulation problems where the goal is to estimate the worst performance given that the parameters of the model are in an uncertainty set. For instance, **?** considered the problem of robust simulation and applied it to global warming policy makings.

The use of simulation methods in estimation is in general called simulation-based estimation (SBE) in econometrics literature; see, for instance, Chapter 12 of **?**. Its basic idea is to use a Monte Carlo method to approximate an expectation with no closed form expressions. Another SBE method is called indirect inference, introduced by **?** and **?**, using simulations performed under the initial model to correct for the asymptotic bias of input parameters estimated from an approximate model, which is used to replace the complicated and intractable initial models. These methods have been applied to many econometric models including limited dependent variable models (**?**) and dynamic panel models (**?**). Unlike the above SBE methods, which focus on precisely estimating the parameters, our method focuses on the estimations of model performance and decisions. What is closest to our work in the literature is a simulation-based estimator of contingent-claims prices proposed by **?**. However, they only considered how to reduce the bias of the maximum likelihood estimator in this application and provided little theoretical analysis beyond numerical studies. In this paper, we propose a more general method and analyse the bias correction in more details.

2 BASIC IDEAS OF SBIE

Suppose we have observed a sample of data $\mathbf{X}(\theta_0)=\{X_1,\ldots,X_n\}$, where θ_0 is an unknown parameter (or parameter vector) of a data generating process that we use to model the sample, i.e., we may generate new random samples that have the same distributions as $\mathbf{X}(\theta)$ if θ is known using the same data generating process. In a simple case, for instance, $\mathbf{X}(\theta_0)$ may be an independent and identically distributed (i.i.d.) sample of an exponentially distributed interarrival time and θ_0 is the unknown arrival rate. In a more complicated case, for instance, $\mathbf{X}(\theta)$ may be an i.i.d. sample of normally distributed demand data censored by an inventory-control scheme and θ_0 includes the unknown mean and variance of the normal distribution. Because θ_0 is unknown we need to estimate θ_0 using the observed data $\mathbf{X}(\theta_0)$. Denote the estimator by $\hat{\theta}_n$. Note that $\hat{\theta}_n$ is a function of $\mathbf{X}(\theta_0)$, i.e., $\hat{\theta}_n=\hat{\theta}_n(\mathbf{X}(\theta_0))$. Many statistical methods have been proposed to obtain $\hat{\theta}_n$ and they often focus on the properties of $\hat{\theta}_n$ with respect to θ_0. In many practical problems, however, our ultimate goal is not to estimate θ_0 but to use θ_0 in some models to estimate system performances or to reach optimal decisions. Let $p(\theta)$ denote the output performances or decisions of a model given the parameter θ and, in these problems, what we are interested in is to estimate $p(\theta_0)$. An often used approach is to use $\hat{p}_n = p(\hat{\theta}_n)$ as an estimator of $p(\theta_0)$. This is the two-step approach, because it first estimates θ_0 and then plugs the estimate to $p(\cdot)$ to estimate $p(\theta_0)$.

However, one of the critical problem of the two-step approach is that $p(\hat{\theta}_n)$ as an estimator of $p(\theta_0)$ may not possess the nice properties of $\hat{\theta}_n$ as an estimator of θ_0. For instance, when $p(\cdot)$ is a nonlinear function, $p(\hat{\theta}_n)$ may be biased even when $\hat{\theta}_n$ is unbiased. To solve the problem, an ideal approach is to estimate $p(\theta_0)$ directly. However, in many practical problems, $p(\theta_0)$ may not directly observable. Therefore, we have to use θ_0 as a bridge to estimate $p(\theta_0)$ and the estimation of θ_0 appears inevitable. Nevertheless, in this paper, we want to find an estimator of θ_0 that combines the properties of $p(\cdot)$ as well as those of the data generating process and ultimately lead to a good estimator of $p(\theta_0)$.

Let $\mathbf{X}(\theta) = \{X_1(\theta), \ldots, X_n(\theta)\}$ be a sample of size n generated from the data generating process with parameter θ for any $\theta \in \Theta$, where Θ is the parameter space. Then, following the two-step approach, $p(\theta)$ may be estimated by $p\left(\hat{\theta}_n(\mathbf{X}(\theta))\right)$, sometimes written as $\hat{p}_n(\theta)$.

To avoid unnecessary complexity in the presentation, we assume that θ is one dimensional and leave its multi-dimensional extension as a future work. Let

$$b_n(\theta) := \mathrm{E}\left[\hat{p}_n(\theta)\right].$$

We assume that $\mathrm{E}\left[\hat{p}_n(\theta)\right]$ is well defined for all $\theta \in \Theta$. Then, $b_n(\theta)$ is a function defined on the parameter space Θ and it is called a *binding function* as in **?**. Note that $b_n(\theta)$ depends on the sample size n and $b_n(\theta) - p(\theta)$ is the bias of $\hat{p}_n(\theta)$.

Suppose that $b_n(\theta)$ is a known function and it can be inverted, i.e., $b_n^{-1}(\cdot)$ is well defined. Then, given the sample $\mathbf{X}(\theta_0)$ and the two-step estimator $\hat{p}_n(\theta_0)$, we propose a new estimator of θ_0,

$$\tilde{\theta}_n = b_n^{-1}\left(\hat{p}_n(\theta_0)\right), \tag{1}$$

assume that $\tilde{\theta}_n \in \Theta$ and call $\tilde{\theta}_n$ the *inversion estimator* of θ_0 given the performance function $p(\cdot)$. Note that $\tilde{\theta}_n$ is a quite intuitive estimator of θ_0 and we may explain the intuitions in at least the following two ways. First, $\tilde{\theta}_n$ is the root of

$$\hat{p}_n(\theta_0) = b_n(\theta), \tag{2}$$

which may be viewed as a special case of the method of moments with $b_n(\theta)$ being the mean (i.e., first moment) of the observed output performance $\hat{p}_n(\theta_0)$. Second, $\tilde{\theta}_n$ is the optimal solution of

$$\min_{\theta \in \Theta} \left[b_n(\theta) - \hat{p}_n(\theta_0)\right]^2,$$

which may be viewed as a special case of the least squared method with only one observation of output performance, i.e., $\hat{p}_n(\theta_0)$. Hence this minimization problem reduces to a root finding problem, similar to Equation (2).

To find the inversion estimator of Equation (1), we need to solve the following root-finding problem:

$$\hat{p}_n(\theta_0) = \mathrm{E}\left[p\left(\hat{\theta}_n(\mathbf{X}(\theta))\right)\right]. \tag{3}$$

If the closed-form expression of the binding function $b_n(\theta)$, i.e., $\mathrm{E}\left[p\left(\hat{\theta}_n(\mathbf{X}(\theta))\right)\right]$, is known, the root finding problem is generally a simple problem and it may be solved analytically. In most practical situations, however, the closed-form expression of the binding function $b_n(\theta)$ is unknown and it may only be estimated through running simulation experiments at the corresponding θ values. In this paper we propose to use Monte Carlo methods to solve the root finding problem.

It is worthwhile noting that, when Monte Carlo methods are used to solve the root finding problem, there is also an approximation error caused by the randomness in the Monte Carlo sample. However, this error is often negligible compared to the estimation error introduced by the randomness in the data $\mathbf{X}(\theta_0)$.

This is because Monte Carlo samples are often computationally cheap to generate and much cheaper than the actual data. Then, we can (and should) solve the root finding problem with a very large sample size to reduce the Monte Carlo error to a level that is negligible compared to the error caused by the actual data. Therefore, in this paper, we do not differentiate SBIEs and inversion estimators and treat SBIEs as inversion estimators when analyzing their statistical properties.

3 PROPERTIES OF SBIE

In the introduction we have argued that the two-step estimator $p(\hat{\theta}_n)$ may not possess the nice small-sample properties of $\hat{\theta}_n$. In particular, as illustrated by the M/M/1 example, $p(\hat{\theta}_n)$ may be heavily biased even when $\hat{\theta}_n$ is nearly unbiased. Indeed, this was one of the motivations for us to consider SBIEs. Another important motivation is that we observe that $\hat{\theta}_n$ may itself be significantly biased due to data censoring and it may leads to a two-step estimator $p(\hat{\theta}_n)$ that is also significantly biased. In this paper we propose SBIEs to address the issue of bias reduction. It is worthwhile noting that the basic idea behind SBIEs is a learning mechanism, which uses Monte Carlo simulations as a tool to learn the expected value of the two-step estimator at any θ value, i.e., the binding function $b_n(\theta)$. When θ is known (as in the learning process), $p(\theta)$ is also known. Therefore, through the learning process, we acquire the knowledge of the bias, i.e., $b_n(\theta)-p(\theta)$. Then, SBIEs use this knowledge to correct the bias in the two-step estimator when θ is unknown (i.e., when $\theta=\theta_0$). Note that Monte Carlo simulations play a critical role in the learning process. It also justifies why we call the estimator SBIE instead of only inversion estimator.

In this section we analyze the consistency and the bias-reduction property of SBIEs rigorously for the case where $\hat{\theta}_n$ is consistent as well as the case where $\hat{\theta}_n$ is inconsistent due to data censoring. We show that SBIEs are consistent no matter the $\hat{\theta}_n$ are consistent or not, and SBIEs reduce the biases of the two-step estimators in both cases. In the last of this section, we study the variance of SBIEs.

3.1 Consistency

To study the statistical properties of SBIEs, we concentrate on the one dimensional case and begin with the following assumptions.

Assumption 1 (a) There exists an open interval $\Theta_0=(a,b)$, where $a<b<\infty$ and $a,b\in\mathbb{R}$ such that θ_0 lies in Θ_0.

(b) The performance or decision $p(\cdot)$ is Lipschitz continuous and strictly monotone on Θ_0.

Next we will give some assumptions about $\hat{\theta}_n(\mathbf{X}(\theta))$. Note that the binding function is defined by $b_n(\theta):=\mathrm{E}\left[p\left(\hat{\theta}_n(\mathbf{X}(\theta))\right)\right]$, where $\mathbf{X}(\theta)$ is a sample of size n generated from the data generating process. In simulation, we can generate a sample path ω and set ω fixed, then we can simulate $\mathbf{X}(\theta,\omega)$ over different parameter θ by only changing the value θ. In this sense, $\mathbf{X}(\theta)$ is a function of θ and ω and can be denoted by $\mathbf{X}(\theta,\omega)$. Observed data can be viewed as $\mathbf{X}(\theta_0,\omega_0)$, where ω_0 is a realization of ω. In the rest of this paper, we will use $\mathbf{X}(\theta)$ and $\mathbf{X}(\theta,\omega)$ to express the same meaning for convenience.

Assumption 2 $\hat{\theta}_n(\mathbf{X}(\theta,\omega))$ is strictly monotone with respect to θ for each fixed ω.

Next, we define an uniform convergence in probability of a sequence of random variables in the following, similar as that of **?**.

Definition 1 A sequence of random variables $Y_1(\theta),\ldots,Y_n(\theta)$ converges to a function $y(\theta)$ in probability uniformly in $\theta\in\Theta$ if

$$\sup_{\theta\in\Theta}\mathbb{P}_\theta(|Y_n(\theta)-y(\theta)|>\varepsilon)\to 0$$

as n goes to infinity for every $\varepsilon>0$.

Assumption 3 The estimator $\hat{\theta}_n(\mathbf{X}(\theta))$ converges to $h(\theta)$ in probability uniformly in $\theta\in\Theta_0$ as n goes to infinity, where $h(\cdot)$ is a Lipschitz continuous and strictly monotone function.

Assumption 4 For some $r > 0$,

$$\sup_n \sup_{\theta\in\Theta_0} \mathrm{E}\left[\left|\hat{\theta}_n(\mathbf{X}(\theta))\right|^{1+r}\right] < \infty.$$

Lemma 1 (Existence and Uniqueness of $\tilde{\theta}_n$) Under Assumptions 1 and 2, and if $p(\hat{\theta}_n)$ lies in the range of $b_n(\cdot)$ on the domain Θ_0, then $\tilde{\theta}_n$ exists and is unique and lies in Θ_0.

In terms of the assumptions, $\hat{\theta}_n(\mathbf{X}(\theta,\omega))$ and $p(\cdot)$ are both strictly monotone, then $p(\hat{\theta}_n(\mathbf{X}(\theta,\omega)))$ is strictly monotone at any sample path ω. As the monotonicity exists at any sample path, the expectation of $p(\hat{\theta}_n(\mathbf{X}(\theta,\omega)))$ is strictly monotone as well, which means $b_n(\cdot)$ is monotone. The inverse function of $b_n(\cdot)$, denoted by $b_n^{-1}(\cdot)$, exists and is one-to-one mapping. Then if $p(\hat{\theta}_n)$ lies in the range of $b_n(\cdot)$, then $\tilde{\theta}_n = b_n^{-1}(p(\hat{\theta}_n))$ exists and is unique. Later we will show that $b_n(\cdot)$ converges to $p(h(\cdot))$ uniformly on Θ_0. Note that $p(\hat{\theta}_n)$ converges to $p(\theta_0)$ and θ_0 is an interior point of Θ_0, then if n is large enough, $p(\hat{\theta}_n)$ will go into the range of $b_n(\cdot)$ for sure.

Lemma 2 Under Assumptions 3 and 4, we have $\mathrm{E}(\hat{\theta}_n(\mathbf{X}(\theta)))$ converges to $h(\theta)$ uniformly on Θ_0.

Due to limit of space, all of the proof in this paper is skipped. Lemma 2 implies that $\mathrm{E}\left[p\left(\hat{\theta}_n(\mathbf{X}(\theta))\right)\right]$ converges to $p(h(\theta))$ uniformly on Θ_0. Indeed,

$$\sup_{\theta\in\Theta_0} \mathrm{E}\left|p(\hat{\theta}_n(\mathbf{X}(\theta))) - p(h(\theta))\right| \leq \sup_{\theta\in\Theta_0} \mathrm{E}\left|\Gamma(\hat{\theta}_n(\mathbf{X}(\theta)) - h(\theta))\right| \to 0$$

where the first inequality comes from the Lipschitz continuity of $p(\cdot)$. The result actually means $b_n(\theta)$ converges to $p(h(\theta))$ uniformly on Θ_0.

Theorem 1 Suppose Assumptions 1 to 4 are satisfied and $p(\hat{\theta}_n)$ lies in the range of $b_n(\cdot)$, then estimator from SBIE $\tilde{\theta}_n$ will converge to the true value θ_0, and $p(\tilde{\theta}_n)$ will converge to $p(\theta_0)$.

If the expectation of the right hand side of the equation (3) is approximated by Sample Average Approximation(SAA)(**?**), the consistency of the $\tilde{\theta}_n$ can also be guaranteed.

Corollary 2 Under the condition of Theorem 1, if $\tilde{\theta}_n$ is the unique solution of the equation

$$p(\hat{\theta}_n(\mathbf{X}(\theta_0))) = \frac{1}{K}\sum_{k=1}^{K}\left[p(\hat{\theta}_n(\mathbf{X}_k(\theta)))\right]$$

where $\mathbf{X}_k(\theta)$ is the sample simulated in the k_{th} run, then $\tilde{\theta}_n$ will converge to the true value θ_0, and $p(\tilde{\theta}_n)$ will converge to $p(\theta_0)$ in probability for any integer K.

3.2 Bias Reduction

First introduce two notations O_p and o_p for the sake of denoting the order of some remainders or errors. One writes $f(n) = O_p(g(n))$ as n goes to infinity if

$$0 < \limsup_{n\to\infty}\left|\frac{f(n)}{g(n)}\right| < \infty, \quad \textit{in probability}$$

and writes $f(n) = o_p(g(n))$ as n goes to infinity if

$$\limsup_{n\to\infty}\left|\frac{f(n)}{g(n)}\right| = 0, \quad \textit{in probability}$$

We use O and o instead if there is no randomness within $f(n)$ and $g(n)$.

Next we will introduce an expansion of $\hat{\theta}_n(\mathbf{X}(\theta))$ such that we can deduce the order of bias of $\tilde{\theta}_n$ and $p(\tilde{\theta}_n)$ and compare them with the order of bias of $\hat{\theta}_n$ and $p(\hat{\theta}_n)$.

Assumption 5 The estimator $\hat{\theta}_n(\mathbf{X}(\theta))$ admits the following Edgeworth Expansion:

$$\hat{\theta}_n(\mathbf{X}(\theta)) = h(\theta) + \frac{A(v,\theta)}{n^{\alpha}} + o_p(n^{-\alpha})$$

where $\alpha \in (0,+\infty)$, and $A(v,\theta)$ is a random term depending on a random variable v, and is infinitely differentiable with respect to θ, and has finite moments at θ_0.

The above assumption, just as **?**, is valid under some regularity conditions, mostly about sufficiently many finite moments conditions and smoothness conditions on the sampling distribution. If $\hat{\theta}_n$ has asymptotic normality property, then $\alpha = \frac{1}{2}$. Details of the expansion can be found from **?** and **?**. It is not very difficult to verify assumption 5 in real cases. For instance, if $\hat{\theta}_n(\mathbf{X}(\theta))$ is a MLE, then $\hat{\theta}_n(\mathbf{X}(\theta))$ can be expanded as

$$\hat{\theta}_n(\mathbf{X}(\theta)) = \theta + \frac{sv}{\sqrt{n}} + o_p(n^{-1/2})$$

where s is a constant that stands for the asymptotic standard deviation of $\sqrt{n}\hat{\theta}_n(\mathbf{X}(\theta))$ and v is a random variable follows standard normal distribution. In this case, $A(v,\theta) = sv$. In specific, if $\hat{\theta}_n(\mathbf{X}(\theta))$ is the MLE of rate of an exponential distribution $Exp(\theta)$, then $A(v,\theta) = \theta v$, and if $\hat{\theta}_n(\mathbf{X}(\theta))$ is the MLE of mean of an normal distribution $N(\mu,\sigma^2)$, then $A(v,\theta) = \sigma v$. In the above cases, the differentiability of $A(v,\theta)$ can also be verified. Submitting the two-stage estimator directly to the model, we get $p(\hat{\theta}_n)$. If $p(\cdot)$ is differentiable, then we can apply Taylor expansion to $p(\cdot)$ at $h(\theta_0)$

$$\begin{aligned} p(\hat{\theta}_n) &= p(h(\theta_0)) + (\hat{\theta}_n - h(\theta_0))p'(h(\theta_0)) + o_p(n^{-\alpha}) \\ &= p(h(\theta_0)) + \frac{A(v,\theta_0)}{n^{\alpha}} p'(h(\theta_0)) + o_p(n^{-\alpha}) \end{aligned}$$

If $h(\theta) \neq \theta$, then the bias of $p(\hat{\theta}_n)$ is of order $O(1)$, as the bias will not converge to 0 when n goes to infinity. Even when the two-stage estimator is consistent, which means that $h(\theta) = \theta$, the bias of $p(\hat{\theta}_n)$ is

$$\mathrm{E}\left[p(\hat{\theta}_n) - p(\theta_0)\right] = \mathrm{E}\left[A(v,\theta_0)\right] \frac{p'(\theta_0)}{n^{\alpha}} + o(n^{-\alpha})$$

When $\mathrm{E}\left[A(v,\theta_0)\right]$ and $p'(\theta_0)$ do not equal to 0, the bias of $p(\hat{\theta}_n)$ is of order $O(n^{-\alpha})$.

Theorem 3 If $\hat{\theta}_n$ is inconsistent, then under the condition of Theorem 1 and Assumption 5 and $p(\cdot)$ is differentiable, we have

$$\mathrm{E}\left[p(\tilde{\theta}_n) - p(\theta_0)\right] = o(n^{-\alpha})$$

Note that in the above theorem, we consider the case that the two-step estimator is inconsistent. Later, we will show that when the two-step estimator is consistent itself, which means $h(\theta) = \theta$, we have a stronger result about the bias order. We first need expand $\hat{\theta}_n$ to a higher order.

Assumption 6 When $\hat{\theta}_n$ is consistent, it admits the following Edgeworth Expansion:

$$\hat{\theta}_n(\mathbf{X}(\theta)) = \theta + \frac{A(v,\theta)}{n^{\alpha}} + \frac{B(v,\theta)}{n^{2\alpha}} + o_p(n^{-2\alpha})$$

Then by the same argument as that in theorem 3, we can show that $\tilde{\theta}_n$ has the following forms:

$$\tilde{\theta}_n = \theta_0 + \frac{A^*}{n^{\alpha}} + \frac{B^*}{n^{2\alpha}} + o_p(n^{-\alpha})$$

Proposition 1 If $\hat{\theta}_n$ is consistent, then under the condition of Theorem 1 and Assumption 6, and $p(\cdot)$ is twice differentiable, then $p(\tilde{\theta}_n)$ admits the following expansion:

$$p(\tilde{\theta}) = p(\theta_0) + \frac{1}{n^{\alpha}}A^* p'(\theta_0) + \frac{1}{n^{2\alpha}}\left(B^* p'(\theta_0) + \frac{1}{2}(A^*)^2 p''(\theta_0)\right) + o_p(n^{-2\alpha})$$

where the coefficients of order $n^{-\alpha}$ and $n^{-2\alpha}$ satisfy

$$\begin{aligned}
&A^* p'(\theta_0) = A(v_0,\theta_0)p'(\theta_0) - \mathrm{E}_v\left[A(v,\theta_0)p'(\theta_0)\right]\\
&B^* p'(\theta_0) + \frac{1}{2}(A^*)^2 p''(\theta_0) = -\mathrm{E}_v\left[\frac{\partial A(v,\theta_0)}{\partial \theta}p'(\theta_0)A^*\right] - \mathrm{E}_v\left[A(v,\theta_0)p''(\theta_0)A^*\right]\\
&+B(v_0,\theta_0)p'(\theta_0) + \frac{1}{2}A(v_0,\theta_0)^2 p''(\theta_0) - \mathrm{E}_v\left[B(v,\theta_0)p'(\theta_0)\right] - \frac{1}{2}\mathrm{E}_v\left[A(v,\theta_0)^2 p''(\theta_0)\right]
\end{aligned}$$

where v_0, v are i.i.d. random variables and E_v means the expectation is taken with respect to v. Note A^* contains randomness, while A^* is independent of v.

Proposition 2 The coefficient of the term of order $n^{-\alpha}$ and $n^{-2\alpha}$ of $p(\tilde{\theta}_n)$ satisfy:

$$\mathrm{E}[A^* p'(\theta_0)] = 0$$

If $p'(\theta_0) \neq 0$, then

$$\mathrm{E}\left[B^* p'(\theta_0) + \frac{1}{2}(A^*)^2 p''(\theta_0)\right] = 0$$

Theorem 4 Under the condition of Propositions 1 and 2, we have

$$E\left[p(\tilde{\theta}_n) - p(\theta_0)\right] = o(n^{-2\alpha})$$

Theorem 4 is a direct result of proposition 1 and 2.

Corollary 5 Under the condition of Theorem 4, if p is convex and increasing or concave and decreasing, then $\tilde{\theta}_n$ is biased low in $n^{-2\alpha}$ order and if p is concave and increasing or convex and decreasing, then $\tilde{\theta}_n$ is biased high in $n^{-2\alpha}$ order.

3.3 Variance

Generally, there exists a tradeoff between bias and variance. It is hard to improve both bias and variance simultaneously. For example, Jackknife can reduce the bias order while variance will increase (**?**). For SBIE, though bias is reduced, the variance may not increase. The analysis is quite similar as that of **?**

Combining Assumptions 1 and 2 and $b_n(\cdot)$ is invertible, we have $\tilde{\theta}_n = b_n^{-1}(p(\hat{\theta}_n))$. Denote $p(\hat{\theta}_n)$ by $\hat{p}_n$. Then

$$Var\left[p(\tilde{\theta}_n)\right] = Var\left[p(b_n^{-1}(\hat{p}_n))\right]$$

We can apply mean value theorem to the right hand side of the above formula. Denote $p(h(\theta_0))$ by $g(\theta_0)$.

$$p(b_n^{-1}(\hat{p}_n)) = p(b_n^{-1}(g(\theta_0))) + \frac{p'(b_n^{-1}(\bar{p}_n))}{b_n'(b_n^{-1}(\bar{p}_n))}(\hat{p}_n - g(\theta_0))$$

where $\bar{p}_n$ lies between $\hat{p}_n$ and $g(\theta_0)$. As n goes to infinity, $\hat{p}_n$ converges to $g(\theta_0)$ and therefore $\bar{p}_n$ converges to $g(\theta_0)$. Moreover, $b_n^{-1}(\cdot)$ converges to $g^{-1}(\cdot)$ uniformly. We can approximate $p(b_n^{-1}(\hat{p}_n))$ by

$$p(b_n^{-1}(\hat{p}_n)) \approx p(b_n^{-1}(g(\theta_0))) + \frac{p'(\theta_0)}{b_n'(\theta_0)}(\hat{p}_n - g(\theta_0))$$

Take variance gives:

$$Var\left[p(b_n^{-1}(\hat{p}_n))\right] \approx \left(\frac{p'(\theta_0)}{b_n'(\theta_0)}\right)^2 Var[\hat{p}_n]$$

provided that $p(\cdot)$ and $b_n(\cdot)$ are differentiable and variance of $\hat{p}_n$ exists. Thus if $\left(\frac{p'(\theta_0)}{b_n'(\theta_0)}\right)^2 > 1$, then the variance of SBIE may increase, compared with the referenced estimator. While if$\left(\frac{p'(\theta_0)}{b_n'(\theta_0)}\right)^2 < 1$, the variance of SBIE may decrease.

4 QUEUEING MODEL

We continue the M/M/1 queue example introduced in the introduction section, to see how SBIE will work. Assume the arrival process of the queuing system is a time homogenous poisson process. The input parameter of the queuing model is the arrival rate θ_0, lying in an bounded set Θ_0. We can infer the parameter value in terms of observed inter arrival times $\mathbf{X}_0 = \{X_1, X_2, \cdots, X_n\}$. Then the Maximum Likelihood Estimator(MLE) of θ can be calculated as

$$\hat{\theta}_n = \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^{-1}$$

To make the example simple and clear, assume we already have some information about the system. The service rate is assumed to be exponential with rate $\mu = 1$ and the true value θ_0 is known to be smaller than μ. The objective performance is the average queue length of the steady state , which is $p(\theta) = \frac{\theta}{\mu-\theta}$ and $p(\hat{\theta}_n)$ is the MLE of the objective. It is possible that the estimate of θ_0 is greater than or equal to μ, then the system will blow up. In particular, $\mathrm{E}\left[p(\hat{\theta}_n)\right]$ does not exist. To deal with this problem, we set a large buffer size J to approximate the infinite buffer size. That means, if $\hat{\theta}_n$ is greater than or equal to $\mu = 1$, we let the estimated average queue length be exactly the buffer size J. Set the true value $\theta_0 = 0.8$ and $J = 999$. The binding function can be calculated in terms of Sample Average Approximation as the following:

$$b_n(\theta) = \frac{1}{K}\sum_{k=1}^{K} p\left(min(0.999, \hat{\theta}_n(\mathbf{X}_k(\theta)))\right)$$

where $\mathbf{X}_k(\theta)$ is the sample generated in the k_{th} run. Figure 1 plots the binding function $b_n(\theta)$ and the original function $p(\theta)$. Since p is convex, we can see that $b_n(\theta)$ is always above $p(\theta)$. The expectation of performance behavior is actually $b_n(\theta_0)$, which is biased high. The vertical distance between $b_n(\theta)$ and $p(\theta)$ is actually the bias of MLE when the true value is at θ. The SBE $\tilde{\theta}_n$ will be smaller than $\hat{\theta}_n$, as illustrated in Figure 1. Thus $p(\tilde{\theta}_n) < p(\hat{\theta}_n)$. Intuitively, the bias of $P(\hat{\theta}_n)$ has sort of been erased.

In Table 1, the sample size is 500, true values are $\theta_0 = 0.8$ and $p(\theta_0) = 4$. MLE and Jackknife(JK) methods are applied to estimate θ. MLE is used as the two-step estimator in the SBIE approach. Numerical results in the table show that all of the estimates of θ_0 from different methods are good. However, when we consider estimates of $p(\theta_0)$, MLE of $p(\theta_0)$ is 4.2829, which is biased high for about 7%. JK method can estimate θ_0 very well, while estimates of $p(\theta_0)$ does not obtain much improvement. On the contrary, SBIE of θ_0 has the largest bias, compared with other two methods, while SBIE can significantly reduce the bias of performance. $p(\tilde{\theta}_n) = 3.9575$ is quite close to the true value and gets only about 1% bias. What's more, the variance of $p(\theta_0)$ is also reduced.

We explain how SBIE works in the above example. Nevertheless, in the above example, we have to set an upper bound manually, since the expectation of the performance actually does not exist. In the real

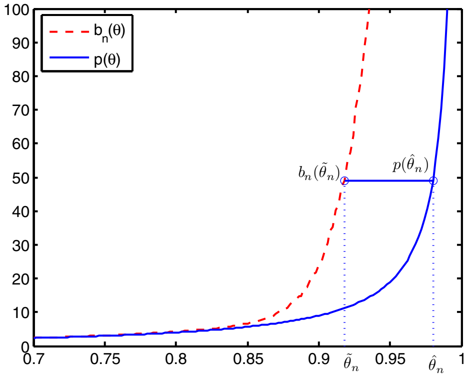

Figure 1: Mechanism of SBIE in M/M/1 queue example

Method	$Mean(\theta)$	$Std(\theta)$	$Se(\theta)$	$Mean(p(\theta))$	$Std(p(\theta))$	$Se(p(\theta))$
MLE	0.8024	0.0374	0.0375	4.2829	1.2624	1.2931
JK	0.8008	0.0374	0.0373	4.2351	1.2311	1.2527
SBIE	0.7926	0.0335	0.0342	3.9575	0.8648	0.8654

Table 1: Comparison of different approaches while $\theta_0 = 0.8$

world, a more practical queue system could be M/M/N+J, which has N servers and finite buffer size J. The finite buffer size can ensure that the system will not blow up. The objective performance is chosen to be the average queue length,which can be denoted by

$$p(\theta) = \frac{\sum_{i=1}^{J} i \left(\frac{\theta/\mu}{N}\right)^{i} \frac{(\theta/\mu)^N}{N!}}{\sum_{i=0}^{N-1} \frac{(\theta/\mu)^i}{i!} + \frac{(\theta/\mu)^N}{N!} \sum_{i=0}^{J} \left(\frac{\theta/\mu}{N}\right)^{i}}$$

where θ is the arrival rate that need to be estimated. In this numerical experiment, we choose service rate $\mu = 1$ server number $N = 1$ and buffer size $J = 100$. We draw the original function and binding function, as illustrated in Figure 2. While the sample size n is small, the binding function lies away from the original function, and the bias is large and the correction will be large as well, which means difference between $\hat{\theta}_n$ and $\tilde{\theta}_n$ will be large; while n increases, the binding function will be closer to the original function, and bias is decreased and smaller correction of the input estimation is needed.

The true value of the arrival rate is $\theta_0 = 0.9$. Under this setting, the true value of average queue length should be $p(\theta_0) = 8.0978$. To discover the relationship between sample size and bias, Table 2 reports the results of MLE and SBIE methods when sample size n increases and the results are plotted in Figure 3. SBIEs are calculated by using two approaches, SAA and Stochastic Approximation (SA)(**?**). Figure 3 shows both SAA and SA approaches work well in reducing the bias. When the sample size is 800, the bias of MLE is about 19%, SBIEs already get very close to the true value.

5 CONCLUSION

In this paper, we propose an easily implementable approach that builds a bridge between parameter estimations and models. The proposed SBIE provides us with an unbiased estimator of $p(\theta_0)$. The bias caused by either nonlinearity of the model $p(\cdot)$ or the inconsistency of the parameter estimator $\hat{\theta}_n$ can be reduced to some orders related to the sample size n. The variance of SBIE can be approximated and it shows that sacrifice of variance is not necessary while reducing the bias. Nevertheless, we only discuss the

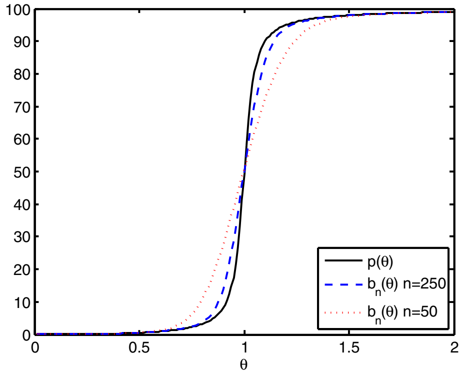

Figure 2: Binding functions with respect to different *n*

Table 2: Comparison of MLE and SBIE with *n* increasing

Sample size		500	650	800	1000	1200	1500	2000	2500	3000
mean(p)	MLE	10.8584	10.2342	9.6358	9.1548	8.9632	8.8428	8.6928	8.4547	8.4111
	SA	8.7868	8.5816	8.2046	8.2343	8.1577	8.1463	8.1851	8.2186	8.1547
	SAA	8.6708	8.4018	8.1381	8.0674	8.0110	8.0820	8.1377	8.0390	8.0706
std(p)	MLE	8.8739	6.9092	5.3275	4.3139	3.7476	3.1348	2.5770	2.1201	1.8593
	SA	8.3551	6.6253	4.5740	3.9356	3.2168	2.8920	2.4591	2.1636	1.9235
	SAA	8.3080	5.9533	4.3770	3.4436	3.0466	2.5766	2.2170	1.8666	1.6748

case when the parameter is one dimensional and the model is one dimensional and monotone, though the SBIE approach is not subject to this simple case. The future work includes discussing multidimensional cases, building confidence intervals for the estimator and so on.

ACKNOWLEDGMENTS

The authors of this paper are supported partially by the Hong Kong Research Grants Council under projects GRF613011 and GRF613012

AUTHOR BIOGRAPHIES

L. JEFF Hong is a professor in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology (HKUST). His research interests include Monte

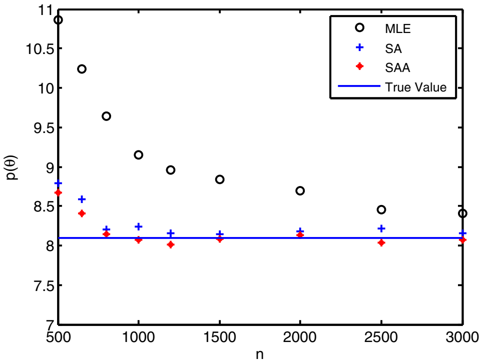

Figure 3: Bias reduction of SBIE in M/M/1+J queue example

Carlo method, financial engineering and risk management, and stochastic optimization. He is currently an associate editor for *Operations Research*, *Naval Research Logistics* and *ACM Transactions on Modeling and Computer Simulation*. His email address is hongl@ust.hk.

JIN FANG is a Ph.D. candidate in the Department of Industrial Engineering and Logistics Management at the Hong Kong University of Science and Technology. Her research interest include stochastic modeling, simulation based methods and parametric estimation. Her email address is jfang@ust.hk.