From Data to Stochastic Modeling and Decision Making

Berkhout, Joost; Heidergott, Bernd; Lam, Henry; Peng, Yijie

published in
Asia-Pacific Journal of Operational Research
2019

DOI (link to publisher)
10.1142/S0217595919400128

document version
Publisher's PDF, also known as Version of record

citation for published version (APA)
Berkhout, J., Heidergott, B., Lam, H., & Peng, Y. (2019). From Data to Stochastic Modeling and Decision Making: What Can We Do Better? *Asia-Pacific Journal of Operational Research*, *36*(6), 1-20. [1940012]. https://doi.org/10.1142/S0217595919400128

Asia-Pacific Journal of Operational Research
Vol. 36, No. 6 (2019) 1940012 (20 pages)

DOI: 10.1142/S0217595919400128

From Data to Stochastic Modeling and Decision Making: What Can We Do Better?

Joost Berkhout

Department of Mathematics, Vrije Universiteit Amsterdam
1081 HV Amsterdam, Netherlands
j2.berkhout@vu.nl

Bernd Heidergott*

Department of Econometrics and Operations Research
Vrije Universiteit Amsterdam, 1081 HV Amsterdam, Netherlands
b.f.heidergott@vu.nl

Henry Lam

Department of Industrial Engineering and Operations Research
Columbia University, New York, NY 10027, USA
henry.lam@columbia.edu

Yijie Peng

Department of Industrial Engineering and Management
College of Engineering, Peking Universtiy
Beijing 100871, P. R. China
pengyijie@pku.edu.cn

Received 30 July 2018
Accepted 1 August 2019
Published 15 November 2019

In the past decades we have witnessed a paradigm-shift from scarcity of data to abundance of data. Big data and data analytics have fundamentally reshaped many areas including operations research. In this paper, we discuss how to integrate data with the model-based analysis in a controlled way. Specifically, we consider techniques to quantify input uncertainty and the decision making under input uncertainty. Numerical experiments demonstrate that different ways in decision making may lead to significantly different outcomes in a maintenance problem.

Keywords: Data analytics; Taylor series expansion; sensitivity analysis.

1. Introduction

In the past decades, we have witnessed a paradigm-shift from scarcity of data to abundance of data. As of today, the holy grail in operations research may lie in

*Corresponding author.

how to marry data science with the wealth of knowledge on models. Rather than discarding analytical models and the analysis based on the model, we advocate in this paper building a shell around theses models to allow for integrating data in a controlled way.

Here, the word "model" refers to stochastic models describing a phenomenon, called *causal model*, such as, for example, the pricing of an European option by the Black–Scholes model, the expected long-run inventory in an inventory system, or expected number of customers in queue at a service center. This is in contrast to statistical models, where a function is designed using artificial parameters to obtain a best fit of the prediction of observed dependent variable (=performance) to the independent variables, think, e.g., of the weights in a linear regression. Causal models are predominant in applied probability, operations research and systems engineering as they allow for an "what if analysis" which is an essential step in designing and optimizing a system. In this paper we focus on causal models.

An instance of a (causal) stochastic model is obtained by choosing the actual values of parameters defining the underlying dynamics of the model. In the option price example, such parameters are the strike price, time to maturity, the risk-free interest rate and the volatility. In the service center example, such parameters are the number of servers, the maximal queue length, the arrival and the service rate. While some of these parameters are controllable and thus exactly known, such as the time to maturity and the strike price in option models and the number of servers in queuing, others are not controllable and their actual value is insecure (decided by nature).

We call parameters that are not controllable and not exactly known but that do define an instance of a model *exogenous* parameters. The main issue in integrating the data and the causal model lies in calibrating exogenous parameters through the real data. Let θ denote an exogenous parameter, say, the actual (= real) volatility of a stock price process or the true arrival rate to a queue, then we let $\hat{\theta}$ denote the point estimate for true parameter θ^*, e.g., maximum likelihood estimation (MLE) and maximum *a posteriori* (MAP) in Bayesian paradigm. We call $\Delta_n = \theta^* - \hat{\theta}$ the *exogenous noise*. Under mild conditions, the exogenous noise Δ_n is asymptotically normal. The exogenous noise affects the performance of the model and the decision based on the model. This leads to the question: How to quantify the input uncertainty caused by the exogenous noise? In classical applied probability and operations research, a widely used technique is the delta method (first order Taylor expansion), which separates the analysis of an instance of model from the analysis of the exogenous noise. To implement the delta method, the sensitivity of the model with respect to the exogenous parameter needs to be estimated. For a casual model, which could be a complex stochastic model without an analytical performance, stochastic gradient estimation technique can be used to estimate the sensitivity. The infinitesimal perturbation analysis (IPA) and likelihood ratio (LR) method are two most popular single-run unbiased stochastic derivative estimation techniques (Glasserman, 1991; Rubinstein and Shapiro, 1993). The smoothed perturbation analysis (SPA) and generalized likelihood ratio (GLR) are generalizations of classic methods for addressing discontinuous sample performance (Fu and Hu, 1997; Peng *et al.*, 2018).

In practice, it is difficult to find a causal model which can accurately represent the observational data. In addition, we usually want to keep the simplicity of the model for the analytical convenience and computational efficiency. For example, the first-come-first-serve (FCFS) M/M/c queue is often used to capture the main dynamic of a call center. More specifically, the M/M/c queue assumes the arrival process is a Poisson process with a constant arrival rate, and the distribution of the service time for each server follows an exponential distribution. With the model assumption, the Erlang-C formula based on the stationary probabilities of the M/M/c queue can be used to do the staff planning problem which is to find the smallest c making the stationary probability for no customer waiting larger than certain threshold. In reality, the arrival rate usually would not be a constant, and the assumption of the FCFS M/M/c queueing model rarely holds. The customers typically arrive more frequently in some periods and less frequently in other periods, and priority queueing, batching processing, and re-entrance structures often appear in the real-world call-center. Moreover, often times we may not be able to get enough access to input data but instead have more accessibility to the output. In these cases, we would like to calibrate the simple casual model in a way that most represents the output data. The first part of this paper discusses how to calibrate an MLE θ^{out} of the causal model based on output data, building on the method in Peng *et al.* (2019). Note that although the causal model can be inaccurate, calibration via MLE has the interpretation of minimizing the Kullback-Leibler (KL) divergence between the assumed model class and the real data generating process. Furthermore, the discrepancy $\Delta_e = \theta^{\text{out}} - \theta^{\text{in}}$, where θ^{in} is the MLE based on the input data if available, can provide insights into model errors and the adequacy of the model. Lastly, for some reviews of uncertainty quantification in simulation, see Henderson (2003); Chick (2006); Barton (2012); Song *et al.* (2015); Lam (2016), and Song and Nelson (2019).

In the second part of this paper, we will investigate the impact of parameter uncertainty in decision-making. Our discussion here follows Liyanage and Shanthikumar (2005); Lim *et al.* (2006); Chu *et al.* (2008); Lim and Vahn (2012); Ban *et al.* (2014) on different strategies that the decision-maker can utilize the available data. In particular, we consider three (simplistic) strategies that use the perspective of taking a simple averaging on the input noise. We demonstrate them with an emergency replacement example and draw some numerical insights.

The remainder of the paper is organized as follows. Section 2 reviews techniques in quantifying uncertainty on exogenous parameters. Efficient decision making under input uncertainty using Taylor series expansion is discussed in Sec. 3. In Sec. 4, three types of decision makings under input uncertainty for a maintenance problem are proposed, and show that these three different ways lead to significantly different outcomes. Numerical experiments are shown in Sec. 5. The last section offers further discussions.

2. Quantifying Uncertainty about Exogenuous Parameters

Suppose we have a (causal) stochastic model:

$$Z_t = g(X_t; \theta),$$

where $X_t = (X_{1,t}, \ldots, X_{n,t})$, $t = 1, \ldots, T$, represent the independently and identically distributed (i.i.d.) input random variables, and the mapping $g(\cdot, \cdot)$ maps the input to the observable output Z_t, thus representing the stochastic model. In this section, we review techniques in quantifying the estimation uncertainty in θ.

2.1. *Confidence regions for exogenuous parameters*

We can quantify the estimation uncertainty of the MLE by constructing asymptotically valid confidence intervals/regions. By the asymptotic normality of the MLE for i.i.d. observations (Van der Vaart, 2000), we have

$$\sqrt{T}(\hat{\theta} - \theta^*) \xrightarrow{d} N(0, \mathcal{I}^{-1}),$$

where θ^* is the true parameter, and the Fisher information matrix is defined by

$$\mathcal{I} = \mathbb{E}\left[\left(\frac{\partial \log p(Z; \theta)}{\partial \theta}\right)^2\right]\Bigg|_{\theta=\theta^*},$$

where p is the density of the output random variable Z. Then, the ι-confidence interval of θ^* can be constructed by

$$\left[\hat{\theta} - \frac{1}{\sqrt{T}} z_{1-\iota/2} \widehat{\mathcal{I}}^{-1/2}, \hat{\theta} + \frac{1}{\sqrt{T}} z_{1-\iota/2} \widehat{\mathcal{I}}^{-1/2}\right],$$

where $z_{1-\iota/2}$ is the $(1 - \iota/2)$-quantile of the standard normal distribution, and $\widehat{\mathcal{I}}$ is an estimate of the Fisher information. Similarly, in the multivariate case where $\theta = (\theta_i)_{i=1,\ldots,d}$, we have

$$\mathcal{I} = \left(\mathbb{E}\left[\left(\frac{\partial \log p(Z; \theta)}{\partial \theta_i}\right)\left(\frac{\partial \log p(Z; \theta)}{\partial \theta_j}\right)\right]\Bigg|_{\theta=\theta^*}\right)_{i,j=1,\ldots,d}.$$

The ι-confidence region of θ^* can be constructed as

$$\{\theta \in \mathbb{R}^d : T(\theta - \hat{\theta})' \widehat{\mathcal{I}}(\theta - \hat{\theta}) \leq \chi^2_{1-\iota,d}\},$$

where $\chi^2_{1-\iota,d}$ is the $(1 - \iota)$-quantile of the χ^2-distribution with degree of freedom d.

For a complex stochastic model, its density usually would not have analytical form. So, simulation may be needed to estimate the density and its derivative in estimating the Fisher information matrix. Notice that for a continuous distribution, estimating the density and its derivative are equivalent to estimate following distribution sensitivities:

$$p(Z_t; \theta) = \frac{\partial \mathbb{E}\left[\mathbf{1}\{g(X_t; \theta) \leq z\}\right]}{\partial z}\Bigg|_{z=Z_t}, \quad \frac{\partial p(Z_t; \theta)}{\partial \theta} = \frac{\partial^2 \mathbb{E}\left[\mathbf{1}\{g(X_t; \theta) \leq z\}\right]}{\partial z \partial \theta}\Bigg|_{z=Z_t}.$$

Note that the above quantities could be challenging to estimate by IPA (due to the discontinuity introduced by the indicator function) or LR (due to the structural parameter). The GLR method circumvents these issues with a systematic smoothing mechanism (Peng *et al.*, 2019).

2.2. *Bayesian quantification of parameter uncertainty*

An unbiased estimator for the likelihood can also be applied in computing a Bayesian posterior distribution for θ, in the setting where the likelihood of Z cannot be analytically obtained. To describe, consider a Bayesian inference for θ under a prior distribution $p(\theta)$. We want to approximate the posterior distribution $p(\theta|Z_1,\ldots,Z_T) \propto L_T(\theta)p(\theta)$, where L_T is the likelihood of the observations. In the Markov chain Monte Carlo (MCMC) algorithm, the involved factor of L_T can be estimated using the unbiased density estimator by using techniques in pseudo-marginals that generate a sample of the likelihood in each iterate, e.g., Andrieu and Roberts (2009) and Doucet *et al.* (2015).

2.3. *Quantifying the impacts of uncertainty to other output performance measures*

Once the uncertainty of the calibrated input is quantified via the approximating distributions, we can also quantify the propagation of this uncertainty in the performance evaluation of other measures. Suppose that we are interested in the performance

$$v(\theta^*) = \mathbb{E}[V(Z_1(\theta^*),\ldots,Z_\ell(\theta^*))].$$

We now may quantify the uncertainty in estimating $v(\theta^*)$ by pluginng in the MLE for θ^*, see Sec. 2.1 Supposing that the simulation size in evaluating $v(\theta^*)$ is abundant, then by the delta method, the confidence interval surrounding the point estimate $v(\hat\theta)$ is

$$\left[v(\hat\theta) - \frac{1}{\sqrt{T}} z_{1-\iota/2}\sqrt{\nabla v(\hat\theta)'\widehat{\mathcal{I}}^{-1}\nabla v(\hat\theta)}, v(\hat\theta) + \frac{1}{\sqrt{T}} z_{1-\iota/2}\sqrt{\nabla v(\hat\theta)'\widehat{\mathcal{I}}^{-1}\nabla v(\hat\theta)}\right],$$

where $\nabla v(\theta) = ((\partial v/\partial\theta_i)(\theta))_{i=1,\ldots,d}$ is the gradient estimate of v (Fu, 2015).

3. Decision Making Under Input Uncertainty

Consider the following decision making problem:

$$\max_\lambda v(\lambda;\theta),$$

where θ is the parameter defining the underlying dynamics of the model v. The parameter θ is unknown, and usually needs to be estimated by the real data. We focus on a decision-maker who considers the following optimization under input uncertainty:

$$\max_\lambda \mathbb{E}[v(\lambda;\hat\theta)], \tag{1}$$

where $\hat{\theta}$ is an estimator of the parameter. The distribution generating $\mathbb{E}$ can be, for instance, the posterior distribution of $\hat{\theta}$, so that the decision-maker is optimizing the posterior profit.

3.1. *Taylor series expansion with exogenous noise*

By definition, $\theta^* = \hat{\theta} + \Delta_n$, and the Taylor series expansion, we have

$$v(\lambda;\theta^*) \approx \sum_{k=0}^{K} \left.\frac{\partial^k v(\lambda;\theta)}{\partial\theta^k}\right|_{\theta=\hat{\theta}} \frac{\Delta_n^k}{k!}.$$

The Taylor series expansion above decomposes the expectation and maximization in (1):

$$\max_{\lambda} \sum_{k=0}^{K} \left.\frac{\partial^k v(\lambda;\theta)}{\partial\theta^k}\right|_{\theta=\hat{\theta}} \frac{\mathbb{E}[\Delta_n^k]}{k!}, \tag{2}$$

which significantly reduces the computational complexity. Note that in case that Δ_n is symmetrically distributed (for example, centered normal distribution), then the odd order derivatives

$$\frac{\partial^{2k+1} v(\lambda;\theta)}{\partial\theta^{2k+1}}, \quad k \in \mathbb{N},$$

have no impact on the decision making in (2), because

$$\left.\frac{\partial^{2k+1} v(\lambda;\theta)}{\partial\theta^{2k+1}}\right|_{\theta=\hat{\theta}} \mathbb{E}[\Delta_n^{2k+1}] = 0.$$

Due to asymptotic normality of the estimator such as the MLE in Sec. 2.1 and Bayesian estimation in Sec. 2.2, the asymptotic distribution of the exogenous noise Δ_n is a centered normal distribution.

For complex stochastic models, simulation might be needed to estimate the derivatives of v. In the sensitivity analysis literature, first order-derivative estimation is predominantly considered. Estimating higher-order derivatives may be technically much more difficult. For example, the sample performance of the average waiting times in the G/G/1 queueing model is continuous but its pathwise derivative is not due to the $(\max, +)$ operation, so that the classic IPA estimator fails to be unbiased for the second order derivative. Deriving an unbiased estimator of the second order derivative for the G/G/1 queueing model using the SPA is rather technical (Fu and Hu, 1993), whereas the GLR method can offer an systematic way to estimate higher order derivatives for a large scope of discontinuous sample performance.

The above Taylor series can be also developed for the multi-dimensional case. A case of particular interest is that of analyzing correlation in input data. The motivation is that while intensive input analysis may be carried out per input parameters, such as, μ and σ in an option model, typically the question whether there is dependency in the observations is much harder to answer and is typically suppressed.

Let $C(\theta_1, \theta_2)$ be the value of a financial option, with, for example, $\theta_1 = \mu$ and $\theta_2 = \sigma$ of the underlying Black–Scholes-Model. In this case $C(\theta_1, \theta_2)$ is obtainable in a closed-form solution. Compute the Taylor series of $C(\theta_1, \theta_2)$, i.e.,

$$
\begin{aligned}
&C(\theta_1 + \Delta_1, \theta_2 + \Delta_2) \\
&\quad = C(\theta_1, \theta_2) + \frac{\partial C(\theta_1, \theta_2)}{\partial \theta_1}\Delta_1 + \frac{\partial C(\theta_1, \theta_2)}{\partial \theta_2}\Delta_2 \\
&\qquad + \frac{1}{2}\left(\frac{\partial^2 C(\theta_1, \theta_2)}{\partial \theta_1^2}\Delta_1^2 + 2\frac{\partial^2 C(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\Delta_1\Delta_2 + \frac{\partial^2 C(\theta_1, \theta_2)}{\partial \theta_2^2}\Delta_2^2\right) + \cdots.
\end{aligned}
$$

The leading term for expressing interference of (possible) correlation of Δ_1 and Δ_2 on the performance outcome is

$$\frac{\partial^2 C(\theta_1, \theta_2)}{\partial \theta_1 \partial \theta_2}\Delta_1\Delta_2.$$

For the option pricing under more complicated stochastic models, e.g., jump-diffusion model and stochastic volatility model, there may be no analytical formula, so simulation is needed to estimate the derivatives.

3.2. *Taylor series expansion with epistemological insecurity*

In this section, we propose to combine input and output modeling into a single framework. Our goal is to approximate $v(\theta)$ by a Taylor series developed at $\hat{\theta}$. The key obstacle with such an approach is that we have ex ante no information on the discrepancy θ and $\hat{\theta}$. We will tackle this problem by elaborating on output modeling. Take, for example, a single server queue and assume that we have no sufficient knowledge on the arrival rate and that we want to interpret the system as an M/M/1 queue (for reasons that go beyond the paper). So, θ becomes the arrival rate. Moreover, the density of the stationary waiting time is known. Observing a sequence of waiting times of customers sufficiently apart yields a data array of samples of stationary waiting times. We can now apply MLE in Sec. 2.1 to compute the value θ^{out} which yields the best fit of the model to the observations. In addition, we can observe arrival times and compute the point estimator θ^{in} for the arrival rate. The discrepancy

$$\Delta_e = \theta^{\text{out}} - \theta^{\text{in}}$$

provides insight into the model error induced by the M/M/1 model. The discrepancy Δ_e is large if the model error is large and vice versa. Suppose that we repeat this procedure for several times, so that we get an estimator for the expected discrepancy Δ_e at θ^{in}.

The statistical information gathered on θ can be used in several ways. We can use a Taylor series expansion (IPA or GLR may be used to estimate derivatives) for $v(\lambda;\theta)$:

$$v(\lambda;\theta^{\text{out}}) \approx \sum_{k=1}^{K} \frac{\partial^k v(\lambda;\theta)}{\partial\theta^k}\bigg|_{\theta=\theta^{\text{in}}} \frac{\Delta_e^k}{k!}$$

to compute the maximal acceptable tolerance, i.e., we want to approximate $v(\theta)$, the true value, up to a precision of, say, ϵ. Via the Taylor series we then can find the largest value for δ by searching though the feasible region such that

$$\left|\sum_{k=1}^{K} \frac{\partial^k v(\lambda;\theta)}{\partial\theta^k}\bigg|_{\theta=\theta^{\text{in}}} \frac{\delta^k}{k!}\right| < \epsilon.$$

We can then perform a statistical test for the hypothesis $|\Delta_e| < |\delta|$, which yields a test whether the model insecurity is acceptable.

Parameter uncertainty might even be deliberately introduced to come up with alternative models. Think, for example, of a single server queue with non-homogeneous Poisson process as arrival stream. Even in the exponential service case this system has no closed form solution. Is it possible to interpret the queue with non-homogeneous Poisson process as arrival stream as M/M/1 queue with parameter insecurity? In other words, can we use the analytical formulas available for the M/M/1 in a randomized way to model the non-homogeneous system?

In analyzing the input uncertainty due to epistemological insecurity, the discrepancy Δ_e usually would not follow a centered distribution. Peng *et al.* (2019) showed that θ^{out} and θ^{in} are significantly different when the assumed model class is insufficient to describe to real data process. Unlike analyzing input certainty due to exogenous noise, the odd order derivatives have influence on the Taylor series expansion for the input uncertainty due to epistemological insecurity.

4. Cost Distribution Evaluation for Age-Based Preventive Maintenance Under Uncertainty in Lifetime Distribution

We consider a component with lifetime distribution function $F(t;s)$, where s represents the vector of parameters of this lifetime distribution. At the moment that the component fails, an emergency repair will be performed and the component will be replaced by a new one. This is called an *emergency replacement*. Furthermore, we perform a preventive maintenance action on the component when it reaches a specified age T. A preventive maintenance action is assumed to make the unit as good as new. The cost of an emergency replacement equals 1 and the cost of a preventive maintenance action equals $c < 1$.

It follows that the mean cost per unit time depends on the maintenance age T. After a preventive maintenance action or an emergency replacement, a new cycle starts. With probability $F(T)$, the unit fails before the maintenance age T is reached, implying that the cycle ends with a failure. With probability $1 - F(T)$, the cycle ends with a preventive maintenance action. The mean cost per cycle is therefore equal to $F(T;s) + (1 - F(T;s))c$. It turns out that the mean length of a cycle is equal to $\int_0^T (1 - F(x;s))dx$ and that the mean cost per unit time, as a function of the preventive maintenance age T, equals

$$\eta_{\text{age}}(T) = \frac{F(T;s) + (1 - F(T;s))c}{\int_0^T (1 - F(x;s))dx}. \tag{3}$$

If the exact value of s is known, the mean cost per unit time (3) is deterministic for any maintenance age T. However, the exact parameter value is rarely known in practice. If this is the case, the mean cost per unit time (3) is a random variable. When choosing a maintenance age, not only the expectation of this random variable is relevant, but also the shape of its probability distribution.

Special case: Uniform distribution with uniform uncertainty

For realistic lifetime distributions (such as the Weibull distribution), the mean cost per unit time in (3) cannot be analyzed algebraically. Therefore, we propose to start with the simple case of a uniform lifetime distribution on the interval $[0, s]$, i.e.,

$$F(t;s) = \begin{cases} 0, & t < 0, \\ \dfrac{t}{s}, & 0 \le t \le s, \\ 1, & t > s. \end{cases}$$

Furthermore, we will model the uncertainty in the parameter s using a uniform distribution on the interval $[1 - \alpha, 1 + \alpha]$. The value of α is a measure for the level of uncertainty with respect to the parameter s. In this simple setting, a closed-form expression for the density function of the mean cost per unit time (3) might be obtained.

If the uncertainty in the parameter s is ignored, i.e., if it is assumed that $s = 1$, the maintenance age T_{opt} that minimizes the mean cost per unit time (3) equals

$$T_{\text{opt}} = \frac{\sqrt{c(2 - c)} - c}{1 - c}.$$

4.1. *Three decision makers under input uncertainty*

In the following, we assume that the parameter s is described by a random variable S which is assumed to be uniformly distributed on the interval $[a, b]$. The decision maker knows the distribution in advance. The following three decision makers are considered, which are given game theoretical based names for notational easiness:

- **Oracle:** To this decision maker the true $s \in [a, b]$ will be revealed. So the oracle will be able to anticipate to the true value by choosing the corresponding optimal maintenance age. Observe that this decision maker is non-realistic. Let $S(\omega)$ denote the value for s for realization ω, and then the oracle finds $T_{\text{opt}}(S(\omega))$ such that

$$T_{\text{opt}}(S(\omega)) = \arg_t \min \eta_{\text{age}}(t, S(\omega)).$$

An oracle is also called *guru* in the literature.

- **Average Oracle:** This decision maker will choose the average maintenance age over all the optimal maintenance ages that the oracle chooses. With the notation above, the *Average Oracle* (Avr. Oracle) uses

$$\hat{T}_{\text{opt}} = \mathbb{E}[T_{\text{opt}}].$$

- **Traditional decision maker:** The goal of the third considered decision maker is to choose the maintenance that minimizes the expected costs. Formally,

$$\mathbf{T} = \arg_t \min \mathbb{E}[\eta_{\text{age}}(t, S(\omega))].$$

When choosing a fixed maintenance age T, i.e., in the case of the Avr. Oracle and the traditional decision maker, the expected "expected costs" (where the extra "expectation" follows from the parameter uncertainty), indicated with a hat, are

$$\hat{\eta}_{\text{age}}(T) = \frac{1}{b-a}\left\{2\ln\left|\frac{\max(\min(T,b),a))}{a}\right| + (1-c/2) \right.$$
$$\left. \times \ln\left|\frac{b-T/2}{\min(\max(T,a),b)-T/2}\right| + \frac{c}{T}(b-\min(\max(T,a),b))\right\}. \quad (4)$$

We are interested in the probability densities of the costs made by each of the decision makers. In the following three subsections, we will report the mathematical details belonging to each of the three decision makers, respectively. Afterwards, we will give some results for a numerical setting which we will discuss.

4.2. *Oracle*

In order to find the optimal maintenance age for each realization of s, defined as $T_{\text{opt}}(s)$, we have to solve

$$\frac{\partial}{\partial T}\eta_{\text{age}}(T;s) = 0,$$

where it holds for $\eta_{\text{age}}(T;s)$ that

$$\eta_{\text{age}}(T;s) = \begin{cases} \dfrac{T+(s-T)c}{T(s-T/2)} & \text{when } T < s \\ \dfrac{2}{s} & \text{when } T \geq s. \end{cases}$$

The above reduces to solving $T_{\rm opt}(s)$ out of

$$0 = (1-c)f(T_{\rm opt};s)\int_0^{T_{\rm opt}}(1-F(x;s))dx - (F(T_{\rm opt};s)$$
$$+ (1-F(T_{\rm opt};s)c))(1-F(T_{\rm opt};s)).$$

In case of uniform life-time-parameter distribution, this can be solved and results in

$$T_{\rm opt}(s) = \alpha(c)s \quad \text{where } \alpha(c) := \frac{\sqrt{c(2-c)}-c}{1-c}. \tag{5}$$

Since $\alpha(c) \in (0,1)$, it holds that

$$\eta_{\rm age}(T_{\rm opt}(s);s) = \frac{\beta(c)}{s} \quad \text{where } \beta(c) := \frac{1-c}{\left(1-\frac{c}{\sqrt{c(2-c)}}\right)\left(1-\frac{\sqrt{c(2-c)}-c}{2-2c}\right)}.$$

In order to calculate the cost density for the Oracle, we will use the method of transformation. To that end, we need the inverse of $\eta_{\rm age}(T;s)$ defined as $\eta_{\rm age}^{-1}(\eta)$, i.e.,

$$\eta_{\rm age}^{-1}(\eta) = \frac{\beta(c)}{\eta},$$

and its derivative to η, i.e.,

$$\frac{d}{d\eta}\eta_{\rm age}^{-1}(\eta) = \frac{-\beta(c)}{\eta^2},$$

so that by the method of transformations, it follows for the density of the oracle that

$$f_\eta^{\rm Oracle}(\eta) = f_S\left(\frac{-\beta(c)}{\eta^2}\right)\frac{\beta(c)}{\eta^2},$$

where $f_S(\cdot)$ is the uniform p.d.f. on $[a,b]$. The expected costs for the Oracle are given by

$$\int \eta f_\eta^{\rm Oracle}(\eta)d\eta = \frac{\beta(c)\ln|b/a|}{b-a}.$$

4.3. *Averaged oracle*

The Avr. Oracle averages over all maintenance ages that the Oracle possible chooses, i.e., the Avr. Oracle chooses

$$\widehat{T}_{\rm opt} = \int f_S(s)T_{\rm opt}(s)ds = \mathbb{E}[S]\cdot\alpha(c) = \frac{a+b}{2}\alpha(c),$$

where $\mathbb{E}[S]$ is the expected value for the parameter S of the life time distribution and $\alpha(c)$ is a constant in c defined in (5). The expected costs can be found by inserting $\widehat{T}_{\rm opt}$ in (4). In order to find the density function in case of a fixed maintenance age, we will again use the method of transformation (in fact, it is just a generalization of the deduction of the oracle setting). Therefore, we need

$$\eta_{\text{age}}^{-1}(T;\eta) = \begin{cases} \dfrac{T\left(1+\dfrac{1}{2}\eta T - c\right)}{\eta T - c} & \text{when } \eta \leq \dfrac{2}{T} \\ \dfrac{2}{\eta} & \text{when } \eta > \dfrac{2}{T} \end{cases} \tag{6}$$

and

$$\frac{d}{d\eta}\eta_{\text{age}}^{-1}(T;\eta) = \begin{cases} T^2 \dfrac{\dfrac{1}{2}c - 1}{(\eta T - c)^2} & \text{when } \eta \leq \dfrac{2}{T} \\ \dfrac{-2}{\eta^2} & \text{when } \eta > \dfrac{2}{T} \end{cases}$$

so that from the method of transformations, it follows that the cost density function of a fixed maintenance age T is

$$f_\eta^T(\eta) = f_S\left(\eta_{\text{age}}^{-1}(T;\eta)\right)\left|\frac{d}{d\eta}\eta_{\text{age}}^{-1}(T;\eta)\right|. \tag{7}$$

Observe that it may happen that the denominator in (6) in case $\eta \leq \frac{2}{T}$ becomes zero when $\eta T - c = 0 \Leftrightarrow \eta = c/T < 2/T$.

4.4. *Traditional decision maker*

The goal of the traditional decision maker is to minimize the expected cost. To that end, he/she has 4 options with respect to a fixed maintenance age T, i.e., $T < a$, $T = a$, $a < T < b$ or $T \geq b$. Each option simplifies (4) so that we are able to find the optimal maintenance age for each option by equating the derivative of (4) to zero. The optimal maintenance ages for the 4 possible options are, respectively:

- $\boldsymbol{T < a}$: $T = \frac{-(a+b)c+\sqrt{(a+b)^2c^2+8abc(1-c)}}{2(1-c)}$,
- $\boldsymbol{T = a}$: $T = a$,
- $\boldsymbol{a < T < b}$: $T = bc$,
- $\boldsymbol{T \geq b}$: $T = b$ (every maintenance age choice $\geq b$ is similar to $T = b$)

The traditional decision maker will evaluate all 4 options in terms of the expected costs and chooses the maintenance age that is optimal. The expected cost can again by found by evaluating (4) and for the costs density function we refer to (7).

5. Numerical Results

In this section, we consider the maintenance model for the three decision makers described above. For the instance, we take $c = 0.05$ and consider 3 situations with respect to the uncertainty:

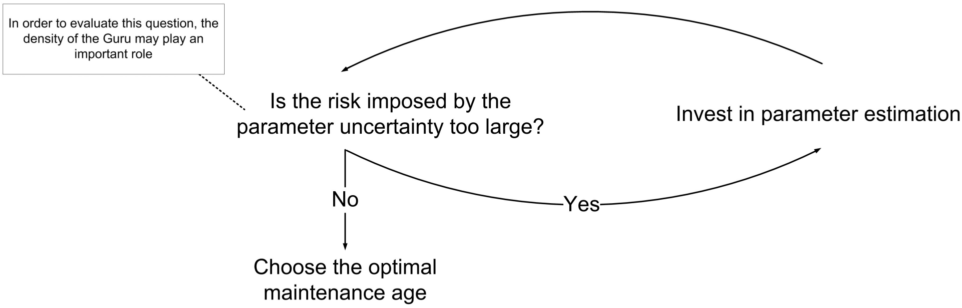

Fig. 1. The framework in parameter uncertainty where the Oracle/Guru may play an important role.

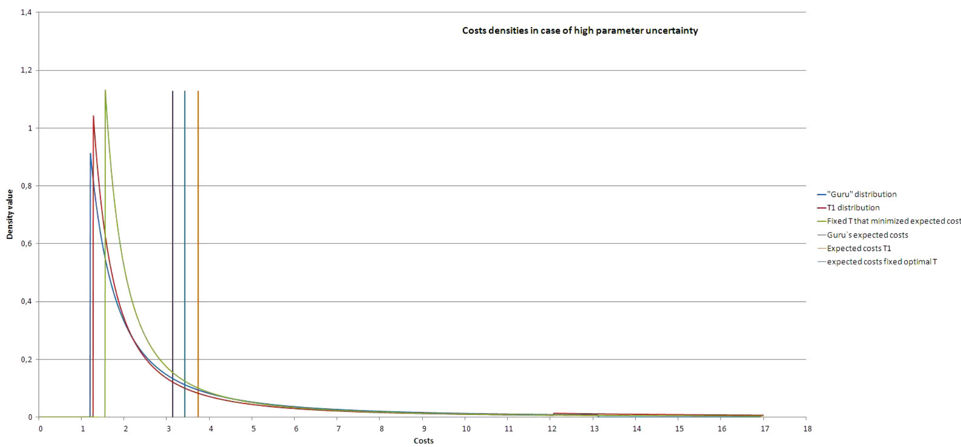

Fig. 2. The costs densities for the three decision makers in case of high uncertainty.

- High uncertainty: $a = 0.1$ and $b = 1.1$,
- Medium uncertainty: $a = 0.3$ and $b = 0.9$,
- Low uncertainty: $a = 0.5$ and $b = 0.7$.

For each uncertainty level, we consider the cost densities of the three decision makers and their cumulative distribution function (CDF). The cost densities are given in Figs. 2, 4 and 6 for the three uncertainty levels, respectively. The CDFs of the costs for the three uncertainty levels can be found in Figs. 3, 5 and 7, respectively. All figures have the same x-axis scale to ensure an easy comparison.

From the results, a few observations can be made. In case of high uncertainty, i.e., $a = 0.1$ and $b = 1.1$, the Avr. Oracle chooses a fixed maintenance age of approximately 0.17 and the traditional decision maker chooses a fixed maintenance age of approximately 0.081. It thus can be concluded that the traditional decision maker is somewhat more conservative than the Avr. Oracle by choosing a maintenance

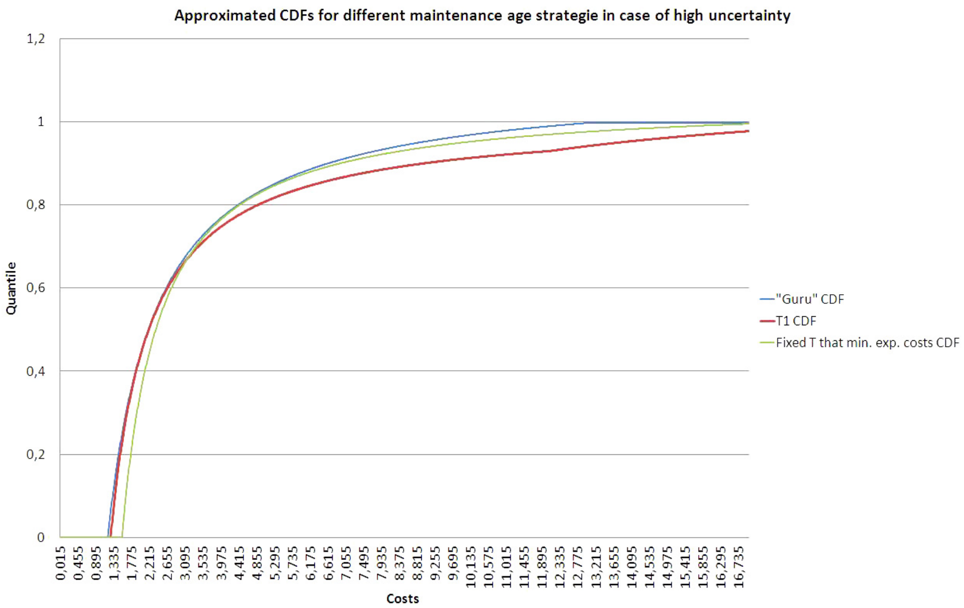

Fig. 3. The costs CDFs for the three decision makers in case of high uncertainty.

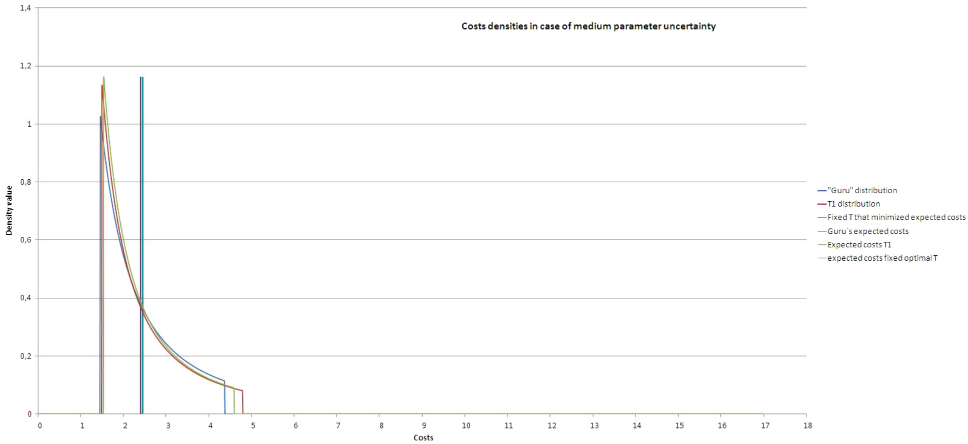

Fig. 4. The costs densities for the three decision makers in case of medium uncertainty.

age below the lower bound of all possible realizations of s. Since $c = 0.05$, the cost of an emergence replacement is relatively low, so that a cost saving strategy is to choose a relative small maintenance age. On the other hand, choosing a relatively small maintenance age in case of high parameter uncertainty can also mean that in case of a relatively large realization of s, the decision maker is performing far too many maintenance replacements per time unit. This last observation is reflected

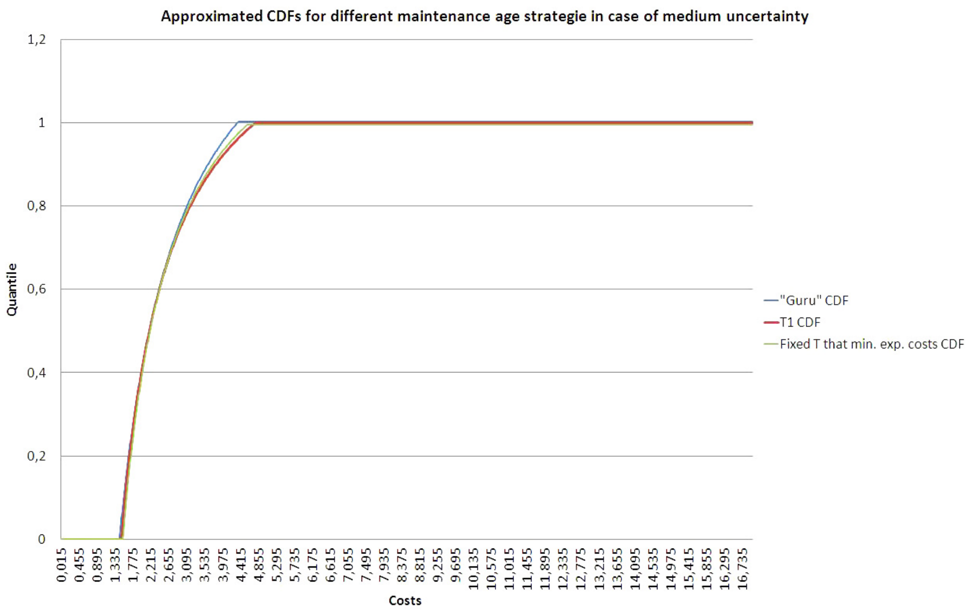

Fig. 5. The costs CDFs for the three decision makers in case of medium uncertainty.

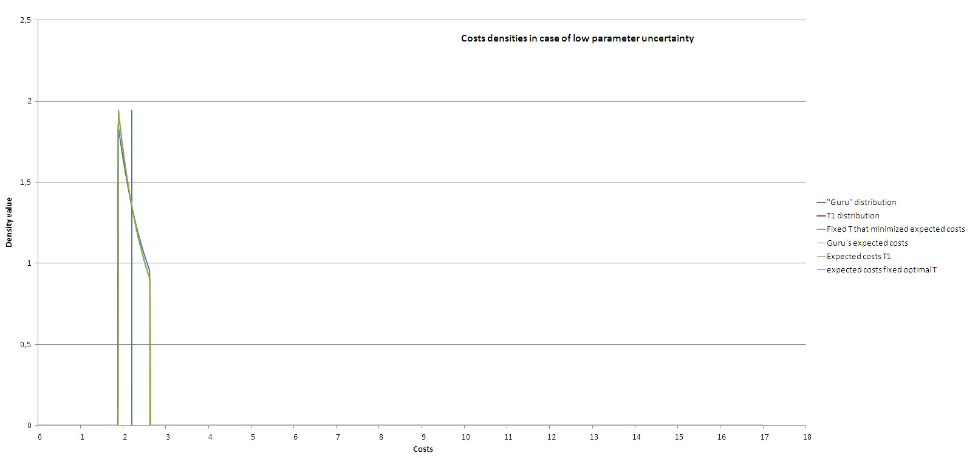

Fig. 6. The costs densities for the three decision makers in case of low uncertainty.

in the maintenance age of the Avr. Oracle, which is a bit above the lower bound $a = 0.1$, meaning that it turns out that on average it is often more optimal to choose a less conservative maintenance age. Furthermore, choosing a very conservative maintenance age means that the decision maker will not have any opportunity to get information about the failure distribution since a failure is not likely to happen often. This discussion raises the question: is there a setting where minimizing

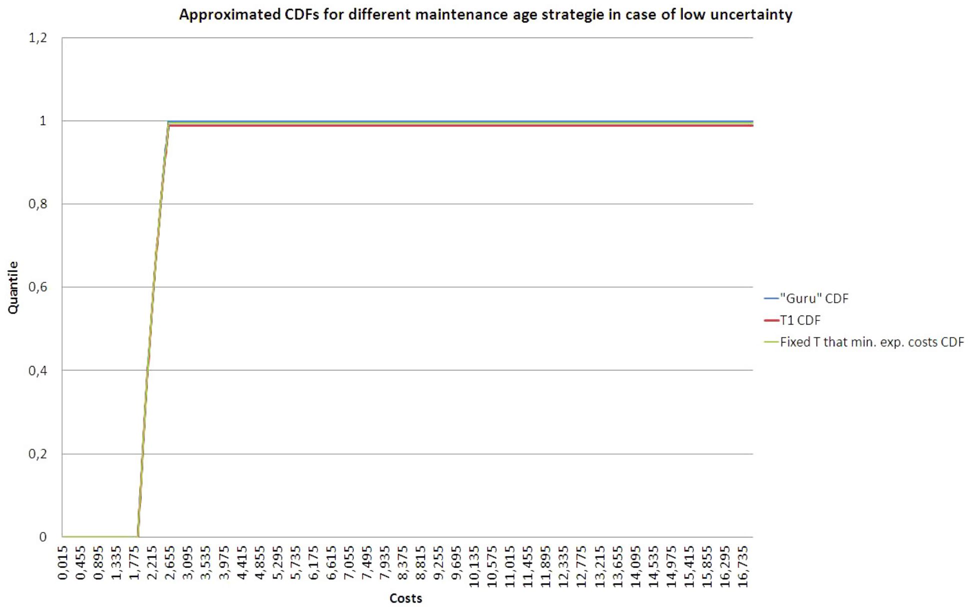

Fig. 7. The costs CDFs for the three decision makers in case of low uncertainty.

the expected costs may be too conservative in the parameter uncertainty case when the decision maker is willing to take some risk?

From the CDFs in Fig. 3, i.e., in the high uncertainty case, it can be observed that the Avr. Oracle performs with 0.7 probability better than the traditional decision maker. In the other 0.3 probability, the traditional decision maker outperforms the Avr. Oracle by its conservativeness, which secures the lower expected costs. This last effect can also be seen from the density function in Fig. 2 which shows a larger tail for the Avr. Oracle.

The figures illustrate that when the uncertainty decreases, i.e., when $b - a \to 0$, the cost densities of all three considered decision makers converge to each other.

From Fig. 3 in case of high parameter uncertainty, it follows that for the traditional decision maker, with 0.1 probability the actual cost are greater than 7, which is more than double the size of the expected costs of 3.4. Even in the medium parameter uncertainty case, it follows that with 0.1 probability the actual costs are greater than 3.9, while the expected costs that the traditional decision maker takes into consideration is 2.4, still a significant difference.

6. Discussion

In this paper, we discuss how to integrate data with the model. Techniques in quantifying input certainty are illustrated. The Taylor series expansion and sensitivity analysis for quantifying both estimation uncertainty and epistemological insecurity are proposed, which can also help efficient decision making under input uncertainty. We consider three different decision makings for a maintenance problem with an unknown parameter to be estimated.

In Sec. 5, the numerical results illustrate that parameter uncertainty plays a much bigger role in decision making as the relatively amount of attention in literature suggests. Before deciding on which maintenance age to choose the decision maker has first to pay attention whether the parameter estimation is precise enough. To this end, the decision maker has to make a trade-off between the following two:

(i) the amount of effort spend in estimating the involved parameters,
(ii) the amount of risk that is acceptable in choosing maintenance age.

The quantification of both is user-dependent and poses a difficult question on its own. Though the numerical results show that it may be beneficial to take this aspect into account. To this end, the Oracle can give an indication of the value of reducing the uncertainty. It gives information on the possible costs saving that is "out there". Comparing the Oracle's outcome with a fixed decision enables us to price the statistical insecurity. When it turns out that the Oracle is far more beneficial, this may trigger the decision maker to reduce the parameter uncertainty and thus getting closer to the Oracle's outcome. The distance in terms of the costs densities between the Oracle and the decision maker may expose the risk involved in the uncertain parameter and thus may provide an evaluation measure. This idea is captured in Fig. 1. A possibility is to map the cost savings density function for the Oracle compared to the traditional decision maker for example using convolution of random variables (or cross-correlation in terms of signal processing theory). Define

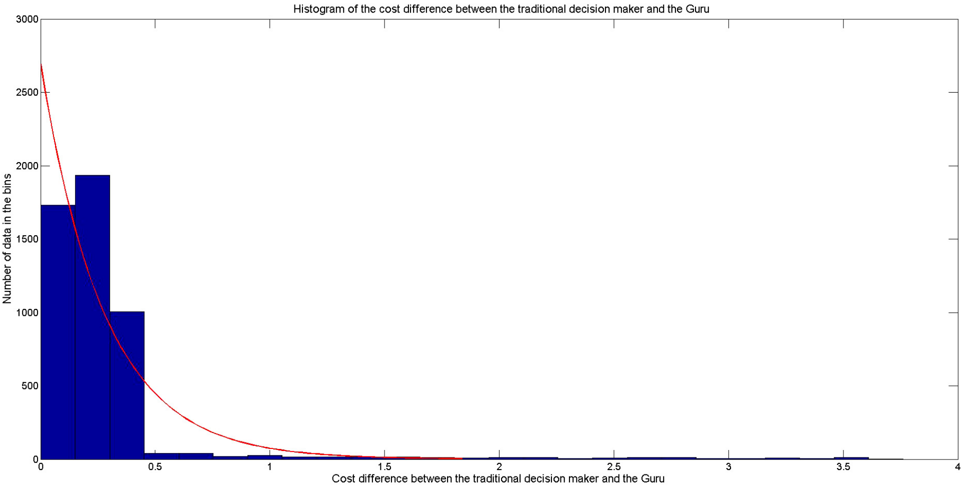

Fig. 8. Histogram of the sampled costs difference between the traditional decision maker and the Oracle.

the cost savings density by $\gamma(z)$, and then using the convolution theorem, it holds that

$$\gamma(z) = \int_{-\infty}^{\infty} f_\eta^{\text{Oracle}}(x) f_\eta^T(x+z)dx. \tag{8}$$

The cost savings density gives insight in the possible gains that the Oracle can obtain because of its advantage of knowing the true outcome of s. That is why the Oracle may provide information whether the decision maker could be satisfied or not by giving insight what more parameter certainty can possibly yield. In Fig. 8, we created a histogram for the costs differences based on sampling realizations of s in case of high uncertainty. It shows that the costs difference can be relatively large.

Lastly, the numerical results led to the suggestion that there might be a setting where minimizing the expected costs may be too conservative in case of parameter uncertainty when the decision maker is willing to take some risk. In other words, rather than finding the maintenance age that minimizes the expected cost, we find the maintenance age that yields the best approximation of the cost density of the Oracle depending on the user's preferences. In conclusion, practice may require an user-dependent interplay between optimization and parameter estimation.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China (NSFC) under Grant 91846301, 71790615, 71571048, 71720107003, and 71690232, by the National Science Foundation (NSF) under Grants CMMI-1542020 and CAREER CMMI-1653339.

References

Andrieu, C and GO Roberts (2009). The pseudo-marginal approach for efficient monte carlo computations. *Annals of Statistics*, 37(2), 697–725.

Ban, GY, NE Karoui and AEB Lim (2014). Machine learning and portfolio optimization. *Mathematics*, 64(3).

Barton, RR (2012). Input uncertainty in outout analysis. in *Proceedings of the Winter Simulation Conference*, p. 6.

Chick, SE (2006). Bayesian ideas and discrete event simulation: Why, what and how. in *Proceedings of the Winter Simulation Conference*, pp. 96–106.

Chu, LY, JG Shanthikumar and ZJM Shen (2008). Solving operational statistics via a bayesian analysis. *Operations Research Letters*, 36(1), 110–116.

Doucet, A, MK Pitt, G Deligiannidis and R Kohn (2015). Efficient implementation of markov chain monte carlo when using an unbiased likelihood estimator. *Biometrika*, 102(2), 295–313.

Fu, M and JQ Hu (1997). *Conditional Monte Carlo: Gradient Estimation and Optimization Applications*. Kluwer Academic Publishers, Boston.

Fu, MC (2015). Stochastic gradient estimation, in Chapter 5 in Handbooks of Simulation Optimization, Fu, MC. (ed.) Springer, pp. 105–147.

Fu, MC and JQ Hu (1993). Second derivative sample path estimators for the GI/G/m queue. *Management Science*, 39(3), 359–383.

Glasserman, P (1991). *Gradient Estimation via Perturbation Analysis*. Kluwer Academic Publishers, Boston.

Henderson, SG (2003). Input model uncertainty: Why do we care and what should we do about it? in *Proceedings of the Winter Simulation Conference*, IEEE Computer Society, pp. 90–100.

Lam, H (2016). Input model uncertainty: Why do we care and what should we do about it? in *Proceedings of the Winter Simulation Conference*, IEEE Computer Society.

Lim, AEB, JG Shanthikumar and ZJM Shen (2006). Model uncertainty, robust optimization and learning. *Mathematics of Operations Research*.

Lim, AEB and GY Vahn (2012). Robust portfolio choice with learning in the framework of regret: Single-period case. *Management Science*, 58(9), 1732–1746.

Liyanage, LH and JG Shanthikumar (2005). A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4), 341–348.

Peng, Y, MC Fu, B Heidergott and H Lam (2019). Maximum likelihood estimation by monte carlo simulation: Towards data-driven stochastic modeling. *Operations Research*.

Peng, Y, MC Fu, JQ Hu and B Heidergott (2018). A new unbiased stochastic derivative estimator for discontinuous sample performances with structural parameters. *Operations Research*, 66(2), 487–499.

Rubinstein, RY and A Shapiro (1993). *Discrete Event Systems: Sensitivity Analysis and Stochastic Optimization by the Score Function Method*. Wiley, New York.

Song, E and BL Nelson (2019). Input–output uncertainty comparisons for discrete optimization via simulation. *Operations Research*, 67(2), 562–576.

Song, E, BL Nelson and CD Pegden (2015). Advanced tutorial: Input uncertainty quantification. *Proceedings of the Winter Simulation Conference*, pp. 162–176.

Van der Vaart, AW (2000). *Asymptotic Statistics*, Cambridge University Press.

Biography

Joost Berkhout is an Assistant Professor in the Analytics and Optimization department at the Vrije Universiteit Amsterdam. He has a PhD degree in Operations Research and his research interests include robustness of mathematical models, production scheduling optimization and Markov chain theory with a focus on applications such as social networks analysis and Google's PageRank.

Bernd Heidergott is the Professor of Stochastic Optimization at the Department of Econometrics and Operations Research at the Vrije Universiteit Amsterdam, the Netherlands. He received his PhD degree from the University of Hamburg, Germany, in 1996, and held postdoc positions at various universities before joining the Vrije Universiteit. He served for more than 5 years as programme director of the BSc and MSc Econometrics and Operations Research. Bernd is research fellow of the Tinbergen Institute and board member of the Amsterdam Business Research Institute. His research interests are optimization and control of discrete event systems, perturbation analysis, Markov chains, max-plus algebra, and social networks.

Henry Lam is an Associate Professor in the Department of Industrial Engineering and Operations Research in Columbia University. He received his PhD in statistics from Harvard University in 2011, and was on the faculty of Boston University and the University of Michigan before joining Columbia in 2017. Henry's research interests include Monte Carlo simulation, risk and uncertainty quantification, and optimization under uncertainty. His work has been recognized by several venues such as the NSF Career Award (2017) and the INFORMS JFIG Competition Second Prize (2016). He serves on the editorial boards of Operations Research, INFORMS Journal on Computing and Stochastic Models.

Yijie Peng is an Assistant Professor in the Department of Industrial Engineering and Management at Peking University. He received his PhD degree from Fudan University in 2014, and was a research fellow at various universities before joining Peking University. His research interests include stochastic modeling and analysis, simulation optimization, machine learning, data analytics, and healthcare.