Proceedings of the 2016 Winter Simulation Conference
T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, eds.

STOCHASTIC SIMULATION UNDER INPUT UNCERTAINTY FOR CONTRACT-MANUFACTURER SELECTION IN PHARMACEUTICAL INDUSTRY

Alp Akcay
Tugce Martagan

Department of Industrial Engineering & Innovation Sciences
Eindhoven University of Technology
5612 AZ, Eindhoven, THE NETHERLANDS

ABSTRACT

We consider a pharmaceutical company that sources a biological product from a set of unreliable contract manufacturers. The likelihood of a manufacturer to successfully deliver the product is estimated via logistic regression as a function of the product attributes. The assignment of a product to the right contract manufacturers is of critical importance for the pharmaceutical company, and simulation-based optimization is used to identify the optimal sourcing decision. However, the input uncertainty due to the uncertain parameters of the logistic regression model often leads to poor sourcing decisions. We quantify the decrease in the expected profit due to input uncertainty as a function of the size of the historical data set, the level of dispersion in the historical data of a product attribute, and the number of products. We also introduce a sampling-based algorithm that reduces the expected decrease in the expected profit.

1 INTRODUCTION

Pharmaceutical research and development is a long and complex process where a single drug might take 10 to 15 years on average to receive approval from the Food and Drug Administration (FDA). The drug development process typically consists of preclinical studies (1 to 6 years), clinical trials (6 to 11 years) and the application/approval process (0.5 to 2 years) which is then followed by post-market surveillance (11 to 14 years). In addition to the long development lead times, the drug development process requires expensive investments and involves high risk of failure. For example, market studies indicate that the average cost of developing a single drug could reach approximately $1.2 billion (including failures) until it gets the FDA approval (Long and Works 2013). It has also been estimated that less than 12% of drugs entering the clinical trials eventually result in an approved medicine (PhRMA 2015). To hedge against the risks associated with failures, long development lead times, and high manufacturing costs, pharmaceutical companies often outsource the manufacturing of biological products (i.e., new molecules, proteins, active ingredients, etc.) used in the preclinical studies to highly specialized contract manufacturers.

In this paper, we study the sourcing decisions of large pharmaceutical companies conducting preclinical studies as part of the new drug development process. More specifically, the pharmaceutical company needs to develop some biological products and can work with different contract manufacturers. The objective of the pharmaceutical company is then to decide which biological products are sourced from which contract manufacturers. Each contract manufacturer is highly specialized and works under an engineer-to-order scheme. In the engineer-to-order scheme, each biological product is associated with a set of physical and chemical attributes, such as purity, molecular mass, size, shape, hydrophobicity, endotoxicity, etc. Once the contract manufacturer receives an order for a specific product, it needs to figure out new manufacturing procedures that would enable the final product to meet its specified set of attributes. The contract manufacturer often needs to deliver the product within a predetermined period of time requested by the pharmaceutical

company. Once the product is delivered, the pharmaceutical company proceeds with the subsequent steps scheduled in the new drug development pipeline.

If the attributes associated with the product are not satisfied within the specified period of time, then the contract manufacturer is assumed to have failed. Such failures have significant financial implications on the large pharmaceutical company, since it delays all the subsequent steps of the new drug development pipeline. However, successfully manufacturing the biological products is often challenging for the contract manufacturer since these products are manufactured as part of the preclinical studies. Therefore, the contract manufacturer has no well-established manufacturing procedures for these biologicals, and often relies on domain knowledge and expertise to successfully manufacture them. In order to avoid failures and delays in the product-development pipeline, the large pharmaceutical company is concerned about working with *reliable* contract manufacturers who have strong expertise and domain knowledge to successfully manufacture the biological products of interest. We assume that the reliability (i.e., the probability of successfully manufacturing the biological product) of each contract manufacturer is unknown to the pharmaceutical company, but it is estimated via logistic regression from the historical performance of a manufacturer for different products with specified attributes.

In practice, it is common to evaluate the performance of a sourcing decision via stochastic simulation since a computationally tractable mathematical programming formulation is not available for the contract-manufacturer selection problem in its most general form. The stochastic simulation relies on the generation of input random variables from given input models. In our setting, it is necessary to generate a binary input variable from the logistic regression model to represent whether a manufacturer is successful or not in manufacturing a product at a given set of attributes. However, the finiteness of historical data leads to uncertainty in the parameters of the logistic regression model, a problem known as *input uncertainty* in stochastic simulations; see Barton, Nelson, and Xie (2010), Barton (2012), Barton, Nelson, and Xie (2014), Xie, Nelson, and Barton (2014b), Song and Nelson (2015), and Lin, Song, and Nelson (2015) for examples of tutorials and recent work on this area. Consequently, the input uncertainty due to the uncertain parameters of the logistic regression model often leads to poor sourcing decisions if not explicitly considered in the formulation of the contract-manufacturer selection problem by the pharmaceutical company. This brings up the two research questions we address in this paper:

- What is the impact of the input uncertainty on the expected profit of the pharmaceutical company?
- How can the sourcing decisions be improved by accounting for the input uncertainty in the contract-manufacturer selection problem?

We answer the first research question by representing the uncertainty in the logistic-regression parameters via their joint posterior distribution and then quantifying the decrease in the expected profit of the pharmaceutical company as a function of the size of the historical data set and the number of biological products to be developed. In addition, we investigate how the dispersiveness of the attributes of the past products affects the impact of the input uncertainty on the expected profit. We answer the second question by introducing a sampling-based algorithm that aims to reduce the expected decrease in the expected profit caused by the input uncertainty. Our numerical experiments show that the average value of the decrease in the expected profit due to input uncertainty is reduced by up to 17% when the number of past products is small and the product attributes vary considerably across these past products.

The remainder of the paper is organized as follows. Section 2 reviews the operations management literature on problems involving optimal procurement decisions under supply uncertainty and the literature on stochastic simulations facing input uncertainty. Section 3 presents the Bayesian logistic regression model to represent the contract-manufacturer reliability as a function of the product attributes, and formally states the contract-manufacturer selection problem in a discrete simulation-based optimization formulation. Section 4 introduces a Markov Chain Monte Carlo (MCMC) algorithm to generate samples from the joint posterior distribution of the logistic-regression parameters, and discusses the incorporation of the input

uncertainty in the contract-manufacturer selection problem. Section 5 presents our numerical analysis, and Section 6 provides concluding remarks with future research directions.

2 LITERATURE REVIEW

We group the related literature in two main research streams: Optimal procurement decisions under supply uncertainty, and the analysis of input uncertainty in stochastic simulations.

Several studies in the supply chain management literature investigate the optimal sourcing and procurement decision when suppliers are unreliable in various industry settings such as high-tech manufacturing, process industry, and retail operations. For example, Chaturvedi and Martínez-de Albéniz (2011) consider a buyer facing multiple unreliable suppliers, and analyze the optimal auction design based on the buyer's level of information about the cost and reliability of suppliers. Dada, Petruzzi, and Schwarz (2007) consider a newsvendor setting and analyze the impact of supplier reliability on procurement decisions. On the other hand, Swinney and Netessine (2009) study the long-term contracting decisions under supplier uncertainty, and analyze two-period contracting games to coordinate the supply chain in the presence of default risk. We note that the aforementioned studies implicitly assume that the suppliers' yield distributions are known with certainty, and mainly focus on analyzing the procurement decisions. In the presence of incomplete information on supplier reliability, learning the supplier uncertainty and its implications on sourcing decisions are first investigated in Tomlin (2009). More specifically, Tomlin (2009) proposes a Bayesian model of supply learning where the decision maker has a forecast of suppliers' yield distribution and updates that forecast based on its experience with suppliers. The proposed Bayesian model is then used to analyze the impact of supply learning on sourcing and inventory strategies. Subsequently, the information collection mechanisms and the corresponding sourcing and inventory decisions are further investigated by Pun and Heese (2014), Silbermayr and Minner (2016), and Saghafian and Tomlin (2016). As a contribution to this research stream, we integrate the use of stochastic simulation for supplier (i.e., contract manufacturer) selection along with estimation of supplier reliability, and analyze the impact of input uncertainty on the sourcing decisions in engineer-to-order pharmaceutical supply chains. We also note that there are several studies on the engineer-to-order systems in the pharmaceutical industry which use discrete-event simulation and stochastic optimization; see Saraph (2003), Johnston, Schruben, Yang, and Zhang (2008), and Martagan, Krishnamurthy, and Maravelias (2016). However, these studies do not consider the impact of input uncertainty in their modeling of pharmaceutical supply chains.

The problem of input uncertainty has been addressed in various ways in the literature on the design and analysis of stochastic simulation experiments: (i) the adoption of Bayesian (e.g., Chick 2001, Biller and Corlu 2011) or frequentist (e.g., Xie, Nelson, and Barton 2014a, Lin, Song, and Nelson 2015) views, and (ii) whether the sampled values of the unknown input model components are fed into the simulation directly (e.g., Ankenman and Nelson 2012, Song and Nelson 2015) or by means of a simulation metamodel (e.g., Barton, Nelson, and Xie 2014, Xie, Nelson, and Barton 2014b). We position our work as Bayesian with direct resampling from the posterior distributions of the unknown input-model parameters. More recently, Lam and Zhou (2015) and Zhou and Xie (2015) address the simulation-based optimization problems and Corlu and Biller (2015) address the ranking and selection problems under input uncertainty. In this study, we build on the 'expectation' risk formulation of simulation-based optimization introduced in Zhou and Xie (2015), which is the Bayesian counterpart of the frequentist 'expected total operating cost' concept proposed by Akcay, Biller, and Tayur (2011).

3 MODEL

We consider a large-scale pharmaceutical company which outsources the development of a set of biological products to a group of smaller-scale and specialized contract manufacturers. Let K denote the number of products to be developed and N denote the number of contract manufacturers. The product k comes with a set of discrete- or continuous-valued features (e.g., purity, molecular mass, size, shape, hydrophobicity,

endotoxicity) of size d denoted by $\mathbf{x}_k = (x_{1k}, \ldots, x_{dk}) \in \mathscr{X}$ that affect the contract-manufacturer's ability to successfully deliver the product. We let Y_{nk} be a binary variable that takes the value of 1 if the manufacturer n successfully delivers the product k with characteristics $\mathbf{x}_k$, and 0 otherwise. Let $\mathbf{Y} = (\mathbf{Y}_1, \ldots, \mathbf{Y}_N)$, where $\mathbf{Y}_n = (Y_{n1}, \ldots, Y_{nK})'$ is a collection of K independent random variables representing the success or failure of manufacturer n. In particular, $Y_{nk} = 1$ and $Y_{nk} = 0$ denote the success and failure of the manufacturer n, respectively, in developing the product k.

Let $c(\mathbf{z}; \mathbf{Y})$ denote the cost incurred by the pharmaceutical company as a result of the product-assignment (or sourcing) scheme $\mathbf{z} = (\mathbf{z}_1, \ldots, \mathbf{z}_N)$, where $\mathbf{z}_n = (z_{n1}, \ldots, z_{nK})'$ with z_{nk} equal to 1 if product k is assigned to manufacturer n, and 0 otherwise. The objective of the pharmaceutical company is to solve the following optimization problem:

$$\min_{\mathbf{z} \in \{0,1\}^{N \times K}} C(\mathbf{z}) = \mathbb{E}_{\mathbf{Y}} \left(c(\mathbf{z}; \mathbf{Y}) \right). \tag{1}$$

In its most general form, the *expected* cost function $C(\mathbf{z})$ is intractable to evaluate analytically. Therefore, stochastic simulation is used to generate random samples of the cost function $c(\mathbf{z}; \mathbf{Y}_i), i = 1, \ldots, M$, at different realizations of the product development outcomes and to perform simulation-based optimization. We refer the reader to Nelson (2010) and Hong, Nelson, and Xu (2015) for surveys on optimization via simulation over discrete decision variables. In Section 5, we perform our numerical experiments on the quantification of input uncertainty by focusing on a specific functional form of the cost function.

The distribution of $\mathbf{Y}$ is often called the *input distribution*. Traditionally, the input distribution is first estimated from historical data (i.e., the past performance of each manufacturer under various product attributes). However, the finiteness of the historical data often leads to an uncertainty in the input-distribution estimate, and the so-called *input uncertainty* is often ignored in simulation optimization (Zhou and Xie 2015); i.e., the simulation is driven by the realizations of the input random variables $\mathbf{Y}$ generated from the estimated input distribution under the assumption that it is the correct input distribution.

3.1 Logistic Regression for Modeling Contract-Manufacturer Reliability

In this section, we present the details of the logistic regression approach used by the pharmaceutical company to assess the success likelihood of each manufacturer at a specific biological product. Let $p_n(\mathbf{x}_k)$ denote the probability that the manufacturer n successfully delivers the product with feature vector $\mathbf{X} = \mathbf{x}_k$ to the pharmaceutical company; i.e., $p_n(\mathbf{x}_k) = \mathbb{P}(Y_{nk} = 1 | \mathbf{X} = \mathbf{x}_k)$. Noting that Y_{nk} is a binary variable, the estimation of $p_n(\mathbf{x}_k)$ can be turned into a two-class classification problem via logistic regression (Murphy 2012). In particular, given the feature vector $\mathbf{x}_k$, the probabilities of success and failure at the manufacturer n are given by a conditional Bernoulli distribution such that

$$\mathbb{P}(Y_{kn} = 1 | \mathbf{X} = \mathbf{x}_k) = \frac{\exp\left(\theta_{0n} + \sum_{i=1}^{d} \theta_{in} x_{ik}\right)}{1 + \exp\left(\theta_{0n} + \sum_{i=1}^{d} \theta_{dn} x_{ik}\right)}, \tag{2}$$

and

$$\mathbb{P}(Y_{kn} = 0 | \mathbf{X} = \mathbf{x}_k) = 1 - \mathbb{P}(Y_{kn} = 1 | \mathbf{X} = \mathbf{x}_k),$$

respectively, where $\theta_n = (\theta_{0n}, \theta_{1n}, \ldots, \theta_{dn})$ is the $(d+1)$-dimensional parameter vector associated with the nth contract manufacturer. More specifically, the pharmaceutical company uses the logistic regression model in (2) to assess the likelihood of the successful delivery of the product k with the feature vector $\mathbf{x}_k$ by the contract manufacturer n.

The pharmaceutical company can have product histories of different sizes from different manufacturers, leading to a heterogeneous contract-manufacturer base in terms of product attributes and manufacturer performance. We let $\mathscr{D}_n = \{(\mathbf{x}_n^t, y_n^t) : t = 1, \ldots, m_n\}$ with $\mathbf{x}_n^t = (x_{n1}^t, \ldots, x_{nd}^t)$ denote the attributes (or features) of the product t assigned by the pharmaceutical company to the contract manufacturer n; i.e., x_{ni}^t is the ith attribute of the tth product undertaken by the nth manufacturer, and y_n^t is a binary variable that takes the value of 1 if the contract manufacturer n successfully delivered product t and zero otherwise. We

let $\mathscr{D} = \{\mathscr{D}_1, \ldots, \mathscr{D}_n\}$ denote the collection of all the historical data. In Section 4.1, we present a MCMC algorithm to approximate the posterior distribution of the unknown input-model parameters $\theta = (\theta_1, \ldots, \theta_N)$ from the historical data $\mathscr{D}$.

3.2 Accounting for the Input Uncertainty in Contract-Manufacturer Selection

Notice that the parameter vector θ of the logistic regression model in (2) is unknown to the pharmaceutical company, and hence, there is an inherent uncertainty in choosing the input model from which the random samples of $\mathbf{Y}$ are generated while solving the simulation-optimization problem in (1). Traditionally, a point-estimate of θ (e.g., maximum likelihood estimate) is plugged into the logistic regression model and used as if it were equal to the the true value of θ. However, this approach ignores the input uncertainty in the simulation output data. Instead, we adopt a Bayesian approach to capture the uncertainty in the unknown parameters $\theta = (\theta_1, \ldots, \theta_N)$. In particular, we pick a prior $\pi(\theta_n)$ that represents the initial belief of the pharmaceutical company about the logistic-regression parameters $\theta_n = (\theta_{0n}, \theta_{1n}, \ldots, \theta_{dn})$ associated with the nth contract manufacturer. By Bayesian updating, the posterior distribution of θ_n can be obtained as

$$\pi(\theta_n|\mathscr{D}_n) \propto \pi(\theta_n) \prod_{t=1}^{m_n} \left(\frac{\exp\left(\theta_{0n} + \sum_{i=1}^{d} \theta_{in} x_{ni}^t\right)}{1 + \exp\left(\theta_{0n} + \sum_{i=1}^{d} \theta_{dn} x_{ni}^t\right)} \right)^{y_n^t} \left(\frac{1}{1 + \exp\left(\theta_{0n} + \sum_{i=1}^{d} \theta_{dn} x_{ni}^t\right)} \right)^{1-y_n^t}, \tag{3}$$

where the notation $\propto$ denotes equivalence up to a normalization constant. We note that the normalization constant for the posterior in this model is analytically intractable. However, even without computing the posterior distribution analytically, it is possible to generate a random sample of θ_n from its posterior distribution (see Section 4.1).

The posterior distribution $\pi(\theta_n|\mathscr{D}_n)$ represents the current belief of the pharmaceutical company about the input-model parameters θ_n associated with the manufacturer n. There are alternatives approaches for incorporating the uncertainty around θ_n into the simulation-optimization problem in (1); see Zhou and Xie (2015). We adopt the notion of *expected total operating cost* formulation defined as

$$\min_{\mathbf{z} \in \{0,1\}^{N \times K}} \mathbb{E}_\theta \left(\mathbb{E}_{\mathbf{Y}|\theta} \left(c(\mathbf{z}; \mathbf{Y}) \right) \right), \tag{4}$$

where the expectation $\mathbb{E}_{\mathbf{Y}|\theta}$ is with respect to the distribution of $\mathbf{Y}$ at a given value of θ and the expectation $\mathbb{E}_\theta$ is with respect to the posterior distribution of θ. It is important to note that this formulation is neutral to the risk stemming from the uncertainty in the input-model parameters θ as well as the uncertainty due to the randomness in the sampled input variables $\mathbf{Y}$ (i.e., the input uncertainty and the intrinsic simulation uncertainty, respectively). The 'expactation' formulation in (4) has been used in the Bayesian input-uncertainty modeling literature in stochastic simulations; e.g., Chick (2001), Zouaoui and Wilson (2003), Zouaoui and Wilson (2004), and Biller and Corlu (2011). Zhou and Xie (2015) show that as the size of the historical input data increases, the simulation-based optimization formulation in (4) converges to the original simulation-optimization problem under the true input model.

4 SOLUTION APPROACH

In Section 4.1, we present a simulation-based algorithm to generate random samples from the posterior distribution of the logistic-regression parameters θ_n associated with the contract manufacturer n with history $\mathscr{D}_n$. The algorithm can be repeated for the other manufacturers separately. In Section 4.2, we discuss how to solve the problem in (4) under a specific functional form of the cost function.

4.1 Sampling from the Posterior Distribution of the Logistic-Regression Parameters

We capture the uncertainty in the logistic-regression parameters θ_n via its posterior distribution $\pi(\theta_n|\mathscr{D}_n)$. It is known that, unlike the linear regression case (e.g., Azoury and Miyaoka 2009), this cannot be done

exactly, since there is no convenient conjugate prior for the logistic regression. Therefore, we use a MCMC approach to approximate the posterior distribution $\pi(\theta_n|\mathscr{D}_n)$. More specifically, even though we cannot compute the posterior distribution $\pi(\theta_n|\mathscr{D}_n)$ analytically, we can generate random samples from this posterior distribution, and then use the random variates of θ_n to approximate the posterior distribution itself or to calculate the posterior mean of the logistic-regression parameter vector θ_n.

Algorithm 1 outlines how we generate a random sample from the posterior distribution of the logistic-regression parameters θ_n. The algorithm builds on the idea of MCMC sampling, which generates a sequence of realizations of θ whose stationary distribution is the posterior distribution $\pi(\theta_n|\mathscr{D}_n)$; see Andrieu, De Freitas, Doucet, and Jordan (2003) for a survey on the MCMC algorithms. In particular, we use the slice sampling algorithm in Neal (2003) that is designed to sample from a distribution with an arbitrary density function, known only up to a constant of proportionality. Notice that this is what is exactly needed for sampling the parameters of the logistic regression model from their complicated joint posterior distribution with an unknown normalization constant. Slice sampling differs from other well-known MCMC algorithms because only the scaled posterior needs to be specified – no proposal or marginal distributions are needed. In addition, relative to other MCMC techniques, slice sampling allows for larger moves, allowing us to reduce the autocorrelation in the samples of the Markov chain and thereby explore the parameter space more efficiently (DuBois, Korattikara, Welling, Smyth, et al. 2014).

Algorithm 1 Sampling from the posterior distribution of the logistic-regression parameters θ_n.

Inputs: (i) The function $\hat{\pi}(\theta_n|\mathscr{D}_n)$ proportional to the posterior density $\pi(\theta_n|\mathscr{D}_n)$; i.e., $\hat{\pi}(\theta_n|\mathscr{D}_n)$ denotes the right hand side of (3). (ii) The current point of θ_n denoted by $\theta_n^s = (\theta_{0n}^s, \theta_{1n}^s \ldots, \theta_{dn}^s)$. (iii) The width parameters $\mathbf{w} = (w_0, w_1 \ldots, w_d)$
Output: The new point $\theta_n^{s+1} = (\theta_{0n}^{s+1}, \theta_{1n}^{s+1} \ldots, \theta_{dn}^{s+1})$.
Step 1: Generate a random variate τ from the Uniform $(0,\ \hat{\pi}(\theta_n^s|\mathscr{D}_n))$ distribution.
Step 2: Randomly position the hyperrectangle $H = (L_0, R_0) \times (L_1, R_1) \times \ldots \times (L_d, R_d)$:
    for $i = 0$ to d **do**
        $U_i \leftarrow$ Uniform $(0,1)$; $L_i \leftarrow \theta_{in}^s - w_i U_i$; and $R_i \leftarrow L_i + w_i$
    end for
Step 3: Sample from H, shrinking when points are rejected:
    Repeat:
    for $i = 0$ to d **do**
        $U_i \leftarrow$ Uniform $(0,1)$; and $\theta_{in}^{s+1} \leftarrow L_i + U_i(R_i - L_i)$
    end for
    if $\tau < \hat{\pi}(\theta_n^{s+1}|\mathscr{D}_n)$ **then** Exit loop
    end if
    for $i = 0$ to d **do**
        if $\theta_{in}^{s+1} < \theta_{in}^s|\mathscr{D}_n)$ **then** $L_i \leftarrow \theta_{in}^{s+1}$
        else $R_i \leftarrow \theta_{in}^{s+1}$
        end if
    end for

In general, any prior distribution $\pi(\theta_n)$ can be used in (3), depending on the available prior information about the reliability of the manufacturer n at certain product attributes. In our numerical experiments, we will assume a flat noninformative prior by assuming that the prior distribution of θ_{in} is uniform and independent of each other for $i = 0, \ldots, d$. The width parameters in Algorithm 1 is the collection of positive scalars that represent an interval around the current sample of the logistic-regression parameters. Algorithm 1 begins with this interval and searches for an appropriate region containing the points of the target function $\hat{\pi}(\theta_n|\mathscr{D}_n)$ that corresponds to a large enough value. In our implementation of Algorithm 1, we choose the

initial point θ_n^1 and the width parameters randomly from a specified region (see Section 5). It is also critical to verify that the Markov Chain $\{\theta_n^s : s = 1,2,\ldots\}$ simulated via Algorithm 1 converges to its stationary distribution. In our numerical experiments in Section 5, we observe in the marginal trace plots that the stationary distribution is typically achieved for s equal to $1,000$, which we assume as the end of burn-in period; i.e., the samples during the burn-in period are discarded. To reduce the serial autocorrelation in the samples, we set the thinning parameter equal to 10; i.e., we collect the samples at every 10 iterations of Algorithm 1.

4.2 Minimization of the Expected Total Operating Cost

In this section, we consider a specific functional form for the expected cost function given by

$$\sum_{k=1}^{K}\sum_{n=1}^{N} c_{nk}z_{nk} - \mathbb{E}_{\mathbf{Y}|\theta}\left[\sum_{k=1}^{K}\mathbb{I}\left(\sum_{n=1}^{N} z_{nk}Y_{nk} \geq 1\right)\pi_k\right], \tag{5}$$

where c_{ik} is the cost of assigning product k to contract manufacturer n, π_k is the reward to the pharmaceutical company if the product k is successfully developed by any of the contract manufacturer, and $\mathbb{I}(\cdot)$ is the indicator function that takes the value of 1 if the event $\cdot$ is correct and 0 otherwise. For $N = 2$, the expected cost in (5) can be written as

$$g(\mathbf{z};\mathbf{p}_1,\mathbf{p}_2) := \sum_{k=1}^{K}\sum_{n=1}^{2} c_{nk}z_{nk} - \sum_{k=1}^{K}\left(p_{1k}z_{1k} + p_{2k}z_{2k} - z_{1k}z_{2k}p_{1k}p_{2k}\right)\pi_k, \tag{6}$$

where $\mathbf{p}_1 = (p_{11},\ldots,p_{1K})$ and $\mathbf{p}_2 = (p_{21},\ldots,p_{2K})$ with p_{nk} is the likelihood of the manufacturer n in successfully delivering product k. The expected total operating cost formulation (see (4)) associated with the expected cost function in (6) can be approximated via sample average approximation as

$$\min_{\mathbf{z}\in\{0,1\}^{2\times K}} \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} g(\mathbf{z};\mathbf{p}_1^{\omega},\mathbf{p}_2^{\omega}), \tag{7}$$

where Ω is the number of random samples of the success probabilities calculated from the posterior samples of the logistic-regression parameters obtained via Algorithm 1. Since (5) can be exactly evaluated via (6) and does not require stochastic simulation, the formulation in (7) does not has the intrinsic simulation uncertainty while capturing the input uncertainty via simulation. In the remainder of the paper, we assume $N = 2$ and focus on quantifying the impact of input uncertainty in the absence of intrinsic simulation uncertainty. The same approach has been adopted by Zhou and Xie (2015) in their assessment of the input uncertainty associated with different risk formulations for simulation-based optimization problems facing input uncertainty. Notice that (7) is a quadratic binary integer program, and we solve within seconds by using the OPTI Toolbox for MATLAB developed by Currie and Wilson (2012).

5 NUMERICAL EXPERIMENTS

In this section, we consider the sourcing problem of a pharmaceutical company who works with $N = 2$ contract manufacturers on K products where $K \in \{1,2,4\}$. The objective of the pharmaceutical company is to maximize its profit through identifying the optimal sourcing scheme for assigning a subset of these K products to a subset of the contact manufacturers. The pharmaceutical company has limited historical data about the past performance of each contract manufacturer, and makes the sourcing decisions based on its assessment of each manufacturer's success probability through the logistic regression model in (2). Each of the K products represents an engineer-to-order protein manufactured using Chinese Hamster Ovary (CHO) cell culture and has the *purity* characteristic as the product attribute (i.e., $d = 1$). The purity represents the proportion of the protein amount (mg) to the total amount of protein and other byproducts (mg) in the

production batch. Therefore, the purity is a unitless measure and takes values between 0 and 1. In practice, depending on the product type, the desired purity level may range between 50% to 99.9% for preclinical studies. The contract manufacturer is considered as failed if the requested purity level cannot be achieved for a specific protein. We model the purity as the only product attribute associated with the proteins, and focus on quantifying the impact of input uncertainty associated with the logistic-regression model on the expected *profit* of the pharmaceutical company.

More specifically, we are interested in answering the following questions: (i) what fraction of time would the pharmaceutical company make sub-optimal decisions if the input uncertainty was ignored? (ii) what is the impact of the input uncertainty on the expected profit as a function of the length of the historical data? (iii) what is the impact of the dispersiveness of the impurity data (i.e., the level of spread in the past impurity requirements) and the number of proteins to be developed on the expected profit under input uncertainty? (iv) what is the value of accounting for the input uncertainty in contract manufacturer selection via the ETOC formulation in (7)?

The following parameters are used in the numerical experiments: $c_{n1} = 20$, $c_{n2} = 30$, $c_{n3} = 35$, $c_{n4} = 40$ for $n \in \{1,2\}$; $\pi_k = 50$ for $k \in \{1,2,3,4\}$. Since $d = 1$, there are two unknown parameters θ_{0n} and θ_{1n} for the logistic regression model of manufacturer n. In our numerical experiments, the purity levels in the past products vary between 0.5 and 1, and we suppose that θ_{0n} and θ_{0n} are uniformly distributed between 0 and 5, and between -5 and 0, respectively (this implies that the probability of success can take values as low as zero and as high as one for each manufacturer). We consider that the pharmaceutical company can have two types of historical data: In the *dispersed* purity data, the purity requirements of the past products assigned to a particular manufacturer vary considerably between 0.5 and 1. In the *concentrated* purity data, on the other hand, the purity requirements of the past products assigned to a particular manufacturer take similar values. In the numerical experiments, we generate the dispersed purity data by sampling all the purity data from Uniform(0.5,1). We generate the concentrated data by generating a first purity sample from Uniform(0.5,1) and then all the other purity data randomly from $\pm\%1$ of the first sample.

Table 1 summarizes the results obtained for the case of single product, $K = 1$. The rows in Table 1 correspond to the length of historical data, m_1 and m_2, associated with the first and second contract manufacturer, respectively. In Table 1, two critical performance measures are reported to assess the impact of input uncertainty on the expected profit of the pharmaceutical company: Average Difference in Expected Profit (ADEP) and Percentage of Time with Incorrect Solution (PTIS). ADEP represents the reduction in the expected profit due to input uncertainty (i.e., the difference between the optimal expected profit without input uncertainty and the expected profit associated with the sourcing decision that is "optimal" under a point estimate of the logistic regression parameters) averaged over 100 independent macro-replications. We use the posterior mean of the logistic-regression parameters as their point estimates. In each macro-replication, we randomly draw the true value of the logistic-regression parameters from their aforementioned ranges, and assume a uniform prior for each parameter. Comparing the optimal sourcing decisions without input uncertainty against the ones under input uncertainty, PTIS represents the percentage of the time the pharmaceutical company adopts a sub-optimal decision because of the input uncertainty.

Table 1: $K = 1$; expected profit without input uncertainty is 25.270.

	Dispersed purity data		Concentrated purity data	
(m_1, m_2)	ADEP	PTIS	ADEP	PTIS
(5,5)	3.168	31%	7.061	54%
(10,10)	1.330	20%	7.941	54%
(25,25)	0.759	14%	9.153	55%
(50,50)	0.305	13%	8.631	51%
(100,100)	0.111	7%	8.461	57%

Table 2: $K = 2$; expected profit without input uncertainty is 37.741.

	Dispersed purity data		Concentrated purity data	
(m_1, m_2)	ADEP	PTIS	ADEP	PTIS
(5,5)	6.492	55%	11.212	63%
(10,10)	5.063	45%	14.381	68%
(25,25)	2.225	31%	17.609	81%
(50,50)	0.951	23%	16.380	75%
(100,100)	0.627	21%	18.449	77%

Table 3: $K = 4$; expected profit without input uncertainty is 53.213.

	Dispersed purity data		Concentrated purity data	
(m_1, m_2)	ADEP	PTIS	ADEP	PTIS
(5,5)	17.280	79%	24.256	75%
(10,10)	6.837	61%	34.373	87%
(25,25)	3.698	59%	38.923	90%
(50,50)	1.752	41%	36.016	89%
(100,100)	0.657	30%	34.378	90%

For the case with dispersed purity data, Table 1 shows that the impact of input uncertainty is highest with a small number of products developed in the past, and it decreases as more products are assigned to each manufacturer. For example, PTIS decreases from 31% to 7% and ADEP decreases from 3.168 to 0.111 as the amount of historical data (m_1, m_2) gets larger. On the other hand, the same observation is not true for the concentrated purity data. Table 1 shows that there is no monotonic reduction in the ADEP and the PTIS as the amount of historical data increases. The underlying reason for this observation is the fact that high amount of concentrated purity data is not necessarily informative enough to be able to successfully estimate the supplier reliability based on the purity attribute. The concentrated purity data does not adequately explore the space of the purity attribute, and hence higher amount of data does not necessarily translate into a better inference about a supplier's reliability. When we compare the ADEP and PTIS obtained under disperse purity data and concentrated purity data, we observe that the impact of input uncertainty is lower under disperse data.

Table 2 and Table 3 report the impact of input uncertainty on ADEP and PTIS when there are $K = 2$ and $K = 4$ products, respectively. Managerial insights from Table 2 and Table 3 align with the ones from Table 1. For example, we observe that assigning a manufacturer more products of disperse purity attributes alleviates the impact of input uncertainty on the expected profit as evidenced by the decreasing ADEP and PTIS values. Table 2 and Table 3 also show that the negative impact of the input uncertainty on the expected profit typically becomes more evident as the number of products K increases. For example, the percentage reduction in the expected profit due to input uncertainty (i.e., ADEP/expected profit without input uncertainty) increases from $3.168/25.270 = 12.5\%$ to $6.492/37.741 = 17.2\%$ and to $17.280/53.213 = 32.47\%$ as K increases from 1 to 2 and to 4, respectively. This observation can be explained by the fact that the number of possible sourcing decisions increases as the amount of products increases, and the uncertainty in the logistic-regression parameters is more likely to cause a sub-optimal sourcing decision.

Table 4 reports the ADEP values associated with the sourcing decisions reached by the ETOC formulation in (7) for the dispersed purity data (i.e., minimization of ETOC corresponds to minimizing the negative profit in this setting). We observe that the sourcing decisions associated with the ETOC formulation may reduce ADEP up to $(17.280 - 14.337)/17.280 = 17\%$ when the amount of historical data $(m_1, m_2) = (5,5)$.

Table 4: The value of accounting for the input uncertainty.

K	(m_1, m_2)	ADEP	% of scenarios with higher EP	% of scenarios with less EP
4	(5,5)	14.337	21%	5%
	(10,10)	6.595	9%	8%
	(25,25)	3.647	7%	5%
2	(5,5)	5.396	12%	5%
	(10,10)	4.789	3%	2%
	(25,25)	2.084	1%	2%

Table 4 also reports the percentage of macro-replications in which the Expected Profit (EP) at the solution of the ETOC formulation is higher (and less) than the expected profit at the solution associated with treating the point estimate of the logistic-regression parameters as their true values. We note that the value of the ETOC formulation is the most apparent for small amount of historical data, and it decreases as the amount of historical data increases.

6 CONCLUSION

We study the sourcing decisions of a pharmaceutical company who works with a set of contract manufacturers for product development. In this setting, the products are engineer-to-order biologicals (i.e., proteins, active ingredients, etc.) having some predetermined set of attributes. The pharmaceutical company has limited historical data about the past performance of the contract manufacturers, and makes inferences about the success likelihood of the contract manufacturers through a logistic regression model. However, the pharmaceutical company faces with the problem of input uncertainty due to unknown parameters of the logistic regression model, affecting the performance of the resulting sourcing decisions. In this paper, we represent the uncertainty in the logistic-regression parameters by approximating their joint posterior distribution via a MCMC algorithm, and then investigate the impact of input uncertainty on the sourcing decisions and the expected profit of the pharmaceutical company. In addition, we introduce a sampling-based algorithm that aims to reduce the negative impact of input uncertainty on the expected profit of the pharmaceutical company.

Managerial insights generated from the numerical analysis can be summarized as follows: First, we observe that ignoring the input uncertainty often leads to sub-optimal decisions and could financially hurt the pharmaceutical company. For example, ignoring the input uncertainty in the numerical experiments resulted in sub-optimal decisions for 7% to 90% of the time. We find that assigning more products to a particular manufacturer leads to a decrease in both the ADEP and the PTIS if the purity requirements of the products vary considerably. However, this is not necessarily the case if the purity requirements of the products are not significantly different. Secondly, we observe that the negative impact of the input uncertainty on the expected profit becomes more evident as the number of products, K, increases. That is, it becomes more critical to account for the input uncertainty as the product portfolio becomes larger. Lastly, we observe that the average value of the decrease in the expected profit due to the input uncertainty is reduced up to 11% through accounting for the input uncertainty in our numerical experiments. As future work, the proposed model can be extended to capture the risk averseness of the pharmaceutical company using the formulations in Zhou and Xie (2015). It is also possible to extend the proposed model as a stochastic game in order to identify the optimal pricing decisions that coordinate the supply chain.

REFERENCES

Akcay, A., B. Biller, and S. Tayur. 2011. "Improved Inventory Targets in the Presence of Limited Historical Demand Data". *Manufacturing & Service Operations Management* 13 (3): 297–309.

Andrieu, C., N. De Freitas, A. Doucet, and M. I. Jordan. 2003. "An Introduction to MCMC for Machine Learning". *Machine Learning* 50 (1-2): 5–43.

Ankenman, B., and B. L. Nelson. 2012. "A Quick Assessment of Input Uncertainty". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. Uhrmacher, 241–250. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Azoury, K. S., and J. Miyaoka. 2009. "Optimal Policies and Approximations for a Bayesian Linear Regression Inventory Model". *Management Science* 55 (5): 813–826.

Barton, R. 2012. "Tutorial: Input Uncertainty in Output Analysis". In *Proceedings of the 2012 Winter Simulation Conference*, edited by C. Laroque, J. Himmelspach, R. Pasupathy, O. Rose, and A. M. Uhrmacher, 67–78. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Barton, R., B. Nelson, and W. Xie. 2010. "A Framework for Input Uncertainty Analysis". In *Proceedings of the 2010 Winter Simulation Conference*, edited by B. Johansson, S. Jain, J. Montoya-Torres, J. Hugan, and E. Yücesan, 1189–1198. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

Barton, R., B. Nelson, and W. Xie. 2014. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26 (1): 74–87.

Biller, B., and C. Corlu. 2011. "Accounting for Parameter Uncertainty in Large-Scale Stochastic Simulations with Correlated Inputs". *Operations Research* 59 (3): 661–673.

Chaturvedi, A., and V. Martínez-de Albéniz. 2011. "Optimal Procurement Design in the Presence of Supply Risk". *Manufacturing & Service Operations Management* 13 (2): 227–243.

Chick, S. 2001. "Input Distribution Selection for Simulation Experiments: Accounting for Input Uncertainty". *Operations Research* 49 (5): 744–758.

Corlu, C. G., and B. Biller. 2015. "Subset Selection for Simulations Accounting for Input Uncertainty". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 437–446. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Currie, J., and D. I. Wilson. 2012, 8–11 January. "OPTI: Lowering the Barrier Between Open Source Optimizers and the Industrial MATLAB User". In *Foundations of Computer-Aided Process Operations*, edited by N. Sahinidis and J. Pinto. Savannah, Georgia, USA.

Dada, M., N. C. Petruzzi, and L. B. Schwarz. 2007. "A Newsvendor's Procurement Problem when Suppliers are Unreliable". *Manufacturing & Service Operations Management* 9 (1): 9–32.

DuBois, C., A. Korattikara, M. Welling, P. Smyth et al. 2014. "Approximate Slice Sampling for Bayesian Posterior Inference". In *JMLR Workshop and Conference Proceedings*, Number 33, 185–193.

Hong, L. J., B. L. Nelson, and J. Xu. 2015. "Discrete Optimization via Simulation". In *Handbook of Simulation Optimization*, 9–44. Springer.

Johnston, L., L. Schruben, A. Yang, and D. Zhang. 2008. "Establishing the Credibility of a Biotech Simulation Model". In *Proceedings of the 2008 Winter Simulation Conference*, edited by S. J. Mason, R. R. Hill, L. Mönch, O. Rose, T. Jefferson, and J. W. Fowler, 822–826. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Lam, H., and E. Zhou. 2015. "Quantifying Uncertainty in Sample Average Approximation". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 3846–3857. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

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

Long, G. and Works, J. 2013. "Innovation in the Biopharmaceutical Pipeline: A Multidimensional View". Analysis Group, Accessed on 04.16.2016.

Martagan, T., A. Krishnamurthy, and C. T. Maravelias. 2016. "Optimal Condition-Based Harvesting Policies for Biomanufacturing Operations with Failure Risks". *IIE Transactions* 48 (5): 440–461.

Murphy, K. P. 2012. *Machine Learning: a Probabilistic Perspective*. MIT press.

Neal, R. M. 2003. "Slice Sampling". *Annals of Statistics*:705–741.
Nelson, B. L. 2010. "Optimization via Simulation over Discrete Decision Variables". *Tutorials in Operations Research* 7:193–207.
PhRMA 2015. "2015 Biopharmaceutical Research Industry Profile". Pharmaceutical Research and Manufacturers of America. Washington, DC: PhRMA; April 2015.
Pun, H., and H. S. Heese. 2014. "Outsourcing to Suppliers with Unknown Capabilities". *European Journal of Operational Research* 234 (1): 108–118.
Saghafian, S., and B. Tomlin. 2016. "The Newsvendor under Demand Ambiguity: Combining Data with Moment and Tail Information". *Operations Research*: To Appear.
Saraph, P. V. 2003. "Manufacturing Analysis and Control: Shared Resource Capacity Analysis in Biotech Manufacturing". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sànchez, D. Ferrin, and D. J. Morrice, 1247–1250. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
Silbermayr, L., and S. Minner. 2016. "Dual Sourcing Under Disruption Risk and Cost Improvement Through Learning". *European Journal of Operational Research* 250 (1): 226–238.
Song, E., and B. L. Nelson. 2015. "Quickly Assessing Contributions to Input Uncertainty". *IIE Transactions* 47 (9): 893–909.
Swinney, R., and S. Netessine. 2009. "Long-Term Contracts Under the Threat of Supplier Default". *Manufacturing & Service Operations Management* 11 (1): 109–127.
Tomlin, B. 2009. "Impact of Supply Learning When Suppliers are Unreliable". *Manufacturing & Service Operations Management* 11 (2): 192–209.
Xie, W., B. Nelson, and R. Barton. 2014a. "Statistical Uncertainty Analysis for Stochastic Simulation". Working paper, Departmant of Industrial Engineering and Management Sciences, Northwestern University.
Xie, W., B. L. Nelson, and R. R. Barton. 2014b. "A Bayesian Framework for Quantifying Uncertainty in Stochastic Simulation". *Operations Research* 62 (6): 1439–1452.
Zhou, E., and W. Xie. 2015. "Simulation Optimization When Facing Input Uncertainty". In *Proceedings of the 2015 Winter Simulation Conference*, edited by L. Yilmaz, W. K. V. Chan, I. Moon, T. M. K. Roeder, C. Macal, and M. D. Rossetti, 3714–3724. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.
Zouaoui, F., and J. Wilson. 2003. "Accounting for Parameter Uncertainty in Simulation Input Modeling". *IIE Transactions* 35 (9): 781–792.
Zouaoui, F., and J. Wilson. 2004. "Accounting for Input-Model and Input-Parameter Uncertainties in Simulation". *IIE Transactions* 36 (11): 1135–1151.

AUTHOR BIOGRAPHIES

ALP AKCAY is an Assistant Professor in the Department of Industrial Engineering and Innovation Sciences at Eindhoven University of Technology. He holds a Ph.D. in Operations Management from the Tepper School of Business at Carnegie Mellon University. His research interests include sequential decision-making under uncertainty, simulation-based optimization, and data-driven operations management, focusing on applications in supply chain planning and condition-based maintenance. His email address is a.e.akcay@tue.nl.

TUGCE MARTAGAN is an Assistant Professor in the Department of Industrial Engineering and Innovation Sciences at Eindhoven University of Technology. She holds a Ph.D. in Industrial Engineering from the University of Wisconsin-Madison. Her research interests include stochastic modeling and optimization, Markov decision processes, and queuing theory with applications in engineer-to-order systems, biomanufacturing operations, and pharmaceutical supply chains. Her email address is t.g.martagan@tue.nl.