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

SIMULATION-BASED PRODUCTION PLANNING FOR ENGINEER-TO-ORDER SYSTEMS WITH RANDOM YIELD

Alp Akcay
Tugce Martagan

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

ABSTRACT

We consider an engineer-to-order production system with unknown yield. We model the yield as a random variable which represents the percentage output obtained from one unit of production quantity. We develop a beta-regression model in which the mean value of the yield depends on the unique attributes of the engineer-to-order product. Assuming that the beta-regression parameters are unknown by the decision maker, we investigate the problem of identifying the optimal production quantity. Adopting a Bayesian approach for modeling the uncertainty in the beta-regression parameters, we use simulation to approximate the posterior distributions of these parameters. We further quantify the increase in the expected cost due to the so-called input uncertainty as a function of the size of the historical data set, the product attributes, and economic parameters. We also introduce a sampling-based algorithm that reduces the average increase in the expected cost due to input uncertainty.

1 INTRODUCTION

We consider an engineer-to-order batch manufacturing environment with random yield. In this problem setting, each order represents a custom-engineered product that is uniquely designed for a specific client. Furthermore, the client specifies some predetermined requirements on the product attributes or features. Due to the engineer-to-order nature of the product, the production process consists of a series of development runs which is then followed by a scale-up production run. During the development runs, the manufacturer conducts several product-development experiments, and collects data to estimate the production yield (i.e., the amount of products produced per batch) of the scale-up production run. The production yield is represented with a *beta-regression model* where the yield is a beta-distributed random variable with its mean value written as a function of the available product features as the explanatory variables. The objective of the manufacturer is to use the beta regression as a predictive model to estimate the yield distribution, and to identify the optimal production batch size in the scale-up production, which we refer as the *production-batch sizing problem*. However, the amount of data collected during the development experiments is limited because each experiment can be very costly and time consuming. This leads to an uncertainty in the parameters of the beta-regression model, which is also referred as the problem of *input uncertainty* in the stochastic simulation literature (Barton 2012). In this paper, we address the following research questions: (i) What is the impact of the input uncertainty on the expected cost and the production batch sizes? (ii) How can the batch-sizing decisions be improved by accounting for the input uncertainty in the production-batch sizing problem?

We answer the first research question by building a Markov chain Monte Carlo (MCMC) based algorithm that approximates the posterior distribution of the unknown beta-regression parameters. The increase in the expected cost due to the impact of the input uncertainty is then quantified as a function of the size of the historical data, number of the product features, and the economic parameters such as the production cost

and the penalty cost of the unmet demand. To address the second question, we develop a sample average approximation model that takes the posterior samples of the beta-regression parameters as inputs and aims to reduce the expected increase in the expected cost due to the input uncertainty.

An example application of our problem setting is the development of new pharmaceutical drugs. For example, consider a research and development (R&D) project conducted by a pharmaceutical company. The R&D project could require a custom-engineered component which needs to be subcontracted to a highly specialized contract biomanufacturer. In this setting, the custom-engineered product could be a special type of protein or active ingredient that needs to satisfy some predetermined features, such as, purity, hydrophobicity, etc. To produce the requested product, the contract biomanufacturer performs a series of development runs followed by the larger scale production run subject to random yield. The development run generates a limited amount of data to estimate the production outcomes. Therefore, the contract biomanufacturer faces with the production-batch sizing problem under input uncertainty.

The remainder of the paper is organized as follows. Section 2 reviews the relevant literature. Section 3 presents the details of the beta-regression for modeling the production yield and the production-batch sizing problem. Section 4 discusses the proposed solution approach. Section 5 presents our numerical analysis, and Section 6 provides the concluding remarks.

2 LITERATURE REVIEW

We categorize the related literature into two research streams: (i) the production and inventory management under yield uncertainty, and (ii) the quantification of the input uncertainty in simulation output-data analysis.

The impact of yield uncertainty on the lot sizing decisions has been extensively analyzed in the literature. A comprehensive overview of the inventory optimization models under yield uncertainty is provided by Yano and Lee (1995) and Grosfeld-Nir and Gerchak (2004). In the context of engineer-to-order production settings, Wein (1992) and Martagan, Krishnamurthy, and Maravelias (2016) consider a multistage batch manufacturing environment with random yield, and develops a Markov decision process model to optimize production decisions. Similarly, Grosfeld-Nir, Gerchak, and He (2000) build analytic models to optimize the inspection and lot sizing decisions when the production yield is distributed based on the binomial, discrete uniform or interrupted geometric distributions. In addition, several heuristics have been developed to address lot sizing decisions in serial production systems with random yield (Bollapragada and Morton 1999; Ben-Zvi and Grosfeld-Nir 2007). In the context of supply chain management, several studies consider the yield uncertainty. For example, Dada, Petruzzi, and Schwarz (2007) analyze the optimal procurement decisions in a newsvendor setting with unreliable suppliers. Also, Schmitt and Snyder (2012) develop an optimization model to simultaneously manage the yield uncertainty and supply disruptions. However, the aforementioned studies assume perfect information on the yield distribution. When there is incomplete information, Tomlin (2009) is the first to analyze a supply learning mechanism to optimize sourcing decisions. More specifically, Tomlin (2009) considers a Bayesian model of supply learning to characterize the optimal sourcing and inventory strategies. Subsequently, different learning mechanisms for yield uncertainty and the corresponding procurement decisions are analyzed in Pun and Heese (2014), Saghafian and Tomlin (2016) and Silbermayr and Minner (2016). As a contribution to this research stream, we analyze the impact of input uncertainty in an engineer-to-order batch production setting with random yield. To this end, we introduce a sampling-based algorithm to determine the optimal production batch sizes in the presence of input uncertainty.

The simulation literature involves several different approaches to capture the impact of input uncertainty on the design and analysis of stochastic simulation experiments. For example, the Bayesian approach (e.g., Chick 2001; Biller and Corlu 2011) or the frequentist approach (e.g., Xie, Nelson, and Barton 2014a; Lin, Song, and Nelson 2015) have been most commonly adopted to address the problem of input uncertainty. Studies also differ in terms of 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; Akcay and Biller 2017; Akcay and Corlu 2017) or by means of a simulation metamodel (e.g., Barton, Nelson, and Xie 2014;

Xie, Nelson, and Barton 2014b). We position our work as a Bayesian approach, and contribute to the literature of stochastic simulation under input-model uncertainty by direct resampling from the posterior distributions of the unknown parameters of a predictive model.

3 MODEL

We consider a manufacturer which performs the development experiments of an engineer-to-order product followed by its scale-up production. In the development phase, the manufacturer conducts n experiments to estimate the yield of the production. We denote the yield with Y and model it as a beta-distributed random variable, representing the percentage output obtained from one unit of production quantity. The beta distribution is a commonly used stochastic model to incorporate the random yield in many production processes (Yano and Lee 1995, Lee and Lu 2015, Inderfurth and Kiesmüller 2015). In contrary to the common notation where the beta distribution is parameterized with its two shape parameters $\alpha_1 > 0$ and $\alpha_2 > 0$, we parameterize the beta distribution with its mean $\mu \in (0,1)$ and standard deviation σ; i.e., there is a one-to-one relation between the parameters (α_1, α_2) and (μ, σ):

$$\mu = \frac{\alpha_1}{\alpha_1 + \alpha_2} \tag{1}$$

$$\sigma^2 = \frac{\alpha_1 \alpha_2}{(\alpha_1 + \alpha_2)^2(\alpha_1 + \alpha_2 + 1)}, \tag{2}$$

where it can be shown that $\sigma^2 < \mu(1-\mu)$. We let $f_Y(\cdot;\mu,\sigma)$ and $F_Y(\cdot;\mu,\sigma)$ denote the probability density function (pdf) and the cumulative distribution function (cdf) of the yield random variable Y, respectively.

In an engineer-to-order production system, it is often the case that the realized value of the production yield is influenced by certain product attributes. For instance, the percentage output of an artificial protein obtained from a production batch may depend on the temperature and size of the bioreactor used in the production as well as the unique features of the product such as hydrophobicity and endotoxicity (Akcay and Martagan 2016). Motivated by the engineer-to-order characteristics of the product, we model the mean value of the yield random variable as a function of a finite set of explanatory variables (i.e., product attributes). The so-called *beta-regression model* is widely used in situations where the variable of interest is continuous and restricted to the unit interval and is related to other variables through a regression structure Ferrari and Cribari-Neto (2004). In Section 3.1, we provide the details of the beta-regression model to capture the uncertainty in the production yield of the engineer-to-order products.

We denote the demand for the product with d and assume that the manufacturer knows the demand size before the production is initiated. The objective of the manufacturer is to choose the production batch size q such that the demand is satisfied at minimum cost under the beta-distributed yield. The economic parameters of the decision problem are given as follows. One unit of production costs c, and the manufacturer incurs a penalty cost s for each unit of unsatisfied demand, and the customer agrees to buy the amount that exceeds the prespecified demand size d at $k100\%$ discount as long as it is no more than M units.

Let $\pi(q;y)$ denote the cost incurred by the manufacturer as a function of y, which denotes the realized value of the yield random variable Y. We assume that the manufacturer is risk neutral and sets its objective as the minimization of the expected cost

$$\min_{q \geq 0} \Pi(q;\mu,\sigma) \triangleq \int_{y=0}^{1} \pi(q;y) f_Y(y;\mu,\sigma) \mathrm{d}y. \tag{3}$$

In general, the expected cost function in Equation (3) may not be written in closed form, and stochastic simulation is necessary to approximate the expected cost with its sample average $\sum_{m=1}^{M} \pi(q;y_m)/M$ after generating the realizations of the yield random variable $\{y_m, m = 1,\ldots,M\}$; we refer the reader to Kim, Pasupathy, and Henderson (2015) for a recent survey on the optimization via simulation by sample average approximation. An important practical problem is that the manufacturer cannot directly generate the

realizations of the yield random variable because the *input parameters* μ and σ are unknown. Traditionally, the input parameters are first estimated from historical data (i.e., the past realizations of the production yield). However, the finiteness of the historical data leads to an uncertainty in the estimates of the input parameters, 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 Y generated from the estimated input parameters as if they were the correct input parameters. In Section 3.2, we provide an algorithm that solves the production batch-sizing problem in Equation (3) by addressing the input-uncertainty problem caused by the unknown parameters of the beta-regression model.

3.1 Beta Regression for Modeling the Production Yield

In this section, we present the details of the beta-regression model used by the manufacturer to assess the production yield for a specific engineer-to-order product. We assume that each development experiment is independent of each other, and represent the mean value of the yield in the development experiment i as a function of the K-dimensional explanatory variables $\mathbf{x}_i = (1, x_{i1}, x_{i2}, \ldots, x_{iK})$ as follows:

$$\log\left(\frac{\mu_i}{1-\mu_i}\right) = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_K x_{iK} \tag{4}$$

where we refer $\beta = (\beta_0, \beta_1, \ldots, \beta_K) \in \mathbb{R}^{K+1}$ as the mean parameters of the beta-regression model. We follow the convention in Ferrari and Cribari-Neto (2004) and capture the dispersion of the yield random variable with a precision parameter $\phi > 0$. In particular, ϕ is defined as $\alpha_1 + \alpha_2$ for the beta-distribution with shape parameters α_1 and α_2. It follows from Equations (1)-(2) that the variance σ^2 of the beta distribution with mean μ is equal to $\mu(1-\mu)/(\phi+1)$; i.e., $\alpha_1 = \mu\phi$ and $\alpha_2 = (1-\mu)\phi$. This is why ϕ is interpreted as a precision parameter: for fixed mean μ, the larger the value of ϕ, the smaller the variance of the yield random variable.

3.2 Accounting for the Input Uncertainty in Production Batch Sizing

We note that the parameters β and ϕ of the beta regression model are unknown to the manufacturer, and hence, there is an inherent uncertainty in choosing the input model from which the random samples $\{y_m, m = 1, \ldots, M\}$ of the yield variable are generated while solving the simulation-optimization problem in Equation (3). Traditionally, a point-estimate of β and ϕ (e.g., maximum likelihood estimate) is plugged into the input model and used as if it were equal to their true values. 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 β and ϕ. In particular, we pick a prior $\pi(\beta)$ that represents the initial belief of the manufacturer about the mean parameters β in Equation (4). Likewise, we pick a prior $\pi(\gamma)$ where γ is $\log(\phi)$; i.e., γ is the transformed value of ϕ such that $\gamma \in \mathbb{R}$.

We let $\mathscr{D}_n$ denote the accumulated historical data $\{(y_i, \mathbf{x}_i) : i = 1, \ldots, n\}$ after the nth development experiment, where y_i is the realized value of the yield random variable at the explanatory variables $\mathbf{x}_i$ in the ith experiment. By Bayesian updating, the posterior distribution of β and γ can be obtained as

$$\pi(\beta, \theta | \mathscr{D}_n) \propto \prod_{i=1}^{n} \left(\frac{\Gamma(\phi)}{\Gamma(\mu_i\phi)\Gamma((1-\mu_i)\phi)} y_i^{\mu_i\phi - 1} (1-y_i)^{(1-\mu_i)\phi - 1} \right) \pi(\beta)\pi(\gamma), \tag{5}$$

with μ_i defined so that Equation (4) holds and the notation $\propto$ denoting the equivalence up to a normalization constant. 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 β and ϕ from their posterior distribution (see Section 4.1).

The posterior distribution $\pi(\beta,\theta|\mathscr{D}_n)$ represents the current belief of the manufacturer about the input-model parameters β and ϕ, and can be incorporated in the production batch-sizing problem as follows:

$$\min_{q\geq 0}\int_{\mu=0}^{1}\int_{\sigma=0}^{\sqrt{\mu(1-\mu)}}\Pi(q;\mu,\sigma)\pi(\mu,\sigma|\mathscr{D}_n)\mathrm{d}\sigma\mathrm{d}\mu. \tag{6}$$

We note that this formulation is neutral to the risk stemming from the uncertainty in the input-model parameters β and ϕ (i.e., the input uncertainty) as well as to the risk due to uncertainty in the stochastic simulation if the expected cost function $\Pi(q;\mu,\sigma)$ is estimated via sample average approximation (i.e., intrinsic simulation uncertainty). The 'expectation' formulation in (6) 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 Akcay and Martagan (2016). Zhou and Xie (2015) show that as the size of the historical input data increases, the simulation-based optimization formulation in (6) 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 beta-regression parameters β and γ in the presence of the historical data $\mathscr{D}_n$. In Section 4.2, we discuss how to solve the problem in (6) by using these posterior samples.

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

We capture the uncertainty in the beta-regression parameters β and γ via the posterior distribution $\pi(\beta,\gamma|\mathscr{D}_n)$. It is known that this cannot be done exactly since there is no convenient conjugate prior for the beta regression parameters. Therefore, we use a MCMC (Markov chain Monte Carlo) approach to approximate the posterior distribution $\pi(\beta,\gamma|\mathscr{D}_n)$. More specifically, even though we cannot compute the posterior distribution $\pi(\beta,\gamma|\mathscr{D}_n)$ analytically, we can generate random samples from this posterior distribution, and then use the posterior samples of β and γ to approximate the posterior distribution of the mean and the standard of the beta-distribution; i.e., $\pi(\mu,\sigma|\mathscr{D}_n)$ in the production batch-sizing problem in (6).

As in line with the idea of MCMC simulation, we aim to generate a sequence of realizations of β and γ whose stationary distribution is the posterior distribution $\pi(\beta,\gamma|\mathscr{D}_n)$; we refer the reader to Andrieu et al. (2003) for a survey on the MCMC algorithms. In particular, we build on the Bayesian beta-regression algorithm in Cuervo and Lopera (2015) and Cepeda-Cuervo et al. (2016) which sample the parameters β and γ from the posterior *conditional* distributions $\pi(\beta|\gamma,\mathscr{D}_n)$ and $\pi(\gamma|\beta,\mathscr{D}_n)$, respectively, in an iteratively alternating process. However, these two distributions are analytically intractable (so the possibility of Gibbs sampling is ruled out) and a Metropolis Hastings algorithm (see Murphy 2012, Section 24.3) is needed to sample β from $\pi(\beta|\gamma,\mathscr{D}_n)$ and to sample γ from $\pi(\gamma|\beta,\mathscr{D}_n)$. Thus, it is required to develop a proposal distribution (also called the kernel) for both conditional posterior distributions.

We use the proposal distribution of β to sample where the value of β can move from the current state β^s in the the Markov chain simulation. In order to build a multivariate normal proposal distribution to determine the next move of β from its current state, we (i) transform the yield realizations $\{y_1,\ldots,y_n\}$ in the historical data into $\{\tilde{y}_1,\ldots,\tilde{y}_n\}$ such that they can be modelled as normally distributed, and (ii) assume a conditional normal prior distribution $\beta|\gamma\sim\mathscr{N}(\beta_0,\Sigma_0)$ with $(K+1)$-dimensional mean β_0 and $(K+1)\times(K+1)$ covariance matrix Σ_0. In particular, the first-order Taylor approximation of the function $\log(y_i/(1-y_i))$ around the neighborhood $\mu_i^s=1/(1+\exp(-\mathbf{x}_i^\top\beta^s))$ leads to

$$\tilde{y}_i\approx\mathbf{x}_i^\top\beta^s+\frac{y_i-\mu_i^s}{\mu_i^s(1-\mu_i^s)},\quad i=1,\ldots,n, \tag{7}$$

which in turn allows us to identify the mean and the variance of the transformed random variable as $\mathbf{x}_i^\top\beta^s$ and $1/(\mu_i^s(1-\mu_i^s)\phi^s)$, respectively, where $\phi^s=\exp(\gamma^s)$ is the precision parameter associated with the

current value of γ in the simulated Markov chain; i.e., $\tilde{y}_i \sim \mathcal{N}(\mathbf{x}_i^\top \beta^s, 1/(\mu_i^s(1-\mu_i^s)(1+\phi^s)))$. Given that the prior of β is normal and the observations $\{\tilde{y}_1,\ldots,\tilde{y}_n\}$ are also normal, the posterior distribution of β (conditional on the current values β^s and γ^s in the Markov chain) can be shown to be also normal such that $q_1(\beta|\beta^s,\gamma^s) \sim \mathcal{N}(\mathbf{v},\mathbf{V})$ where

$$\mathbf{V} = \left(\Sigma_0^{-1} + \mathbf{X}^\top \Upsilon^{-1} \mathbf{X}\right)^{-1}, \qquad \mathbf{v} = \mathbf{V}\left(\Sigma_0^{-1}\beta_0 + \mathbf{X}^\top \Upsilon^{-1} \tilde{\mathbf{Y}}\right),$$

and $\mathbf{X} = (\mathbf{x}_1,\ldots,\mathbf{x}_n)$ is $(K+1)\times n$ matrix including the historical realizations of the product features, $\tilde{\mathbf{Y}} = (\tilde{y}_1,\ldots,\tilde{y}_n)$ is the $n\times 1$ vector including the transformed values of the historical yield realizations in accordance with the approximation in (7), and Υ is a $n\times n$ diagonal matrix with diagonal entries $1/(\mu_i^s(1-\mu_i^s)\phi^s)$ for $i=1,\ldots,n$. For the derivation of the values of $\mathbf{v}$ and $\mathbf{V}$ provided above, we refer the reader to Section 7.6 in Murphy (2012). We take the distribution $q_1(\beta|\beta^s,\gamma^s)$ as the proposal distribution of β, from which we sample the next candidate value to move in the Markov chain.

To obtain posterior samples of γ, we similarly derive a transformation of the observed yield realizations so that they can be regarded as normally distributed. Let $\tau_i \triangleq (\phi/\mu_i)Y_i$ such that $\mathbb{E}(\tau_i)=\phi$. Applying a first-order Taylor approximation to the function $\log(\tau_i)$ around ϕ then leads to

$$\hat{y}_i \approx \gamma + \frac{t_i}{\phi} - 1, \quad i=1,\ldots,n. \tag{8}$$

Notice that the mean and the variance of this transformed random variable is γ^s and $(1-\mu_i^s)/(\mu_i^s(1+\phi^s))$, respectively. Assuming that the transformed variables in (8) have independent normal distributions (i.e., $\tilde{y}_i \sim \mathcal{N}(\gamma^s,(1-\mu_i^s)/(\mu_i^s(1+\phi^s)))$) and the conditional prior distribution is also normal (i.e., $\gamma|\beta \sim \mathcal{N}(\gamma_0,\sigma_{0,\gamma}^2)$), the posterior distribution of γ (conditional on the current values β^s and γ^s in the Markov chain) is also normal. We denote this distribution with $q_2(\gamma|\beta^s,\gamma^s) \sim \mathcal{N}(w,W)$, where the mean and the variance are as follows:

$$w = \frac{\frac{\gamma_0}{\sigma_{0,\gamma}^2} + \sum_{i=1}^n \frac{\mu_i^s(1+\phi^s)}{1-\mu_i^s}\hat{y}_i}{\sum_{i=1}^n \frac{\mu_i^s(1+\phi^s)}{1-\mu_i^s}} \qquad W = \left(\frac{1}{\sigma_{0,\gamma}^2} + \sum_{i=1}^n \frac{\mu_i^s(1+\phi^s)}{1-\mu_i^s}\right)^{-1}.$$

We remark that the observations (i.e., the transformed realizations $\{\hat{y}_1,\ldots,\hat{y}_n\}$ are independent normal with different means and variances due to different beta-regression covariates at each data point. If the covariates were the same in all the data points (i.e., $\mathbf{x}_1 = \ldots = \mathbf{x}_n$), then w and W reduce to the well-known posterior distribution of the unknown mean value associated with a normally distributed random variable with known variance (Gelman et al. 2014).

Given the proposal distributions $q_1(\beta|\beta^s,\gamma^s)$ and $q_2(\gamma|\beta^s,\gamma^s)$, Algorithm 1 outlines how we generate random samples from the posterior distribution of the beta-regression parameters β and γ iteratively. An important feature of this algorithm is that, when evaluating the acceptance probability of these samples, it is sufficient to know the target density up to a normalization constant. That is why we use the unnormalized posterior distribution $\pi(\cdot,\cdot|\mathscr{D}_n)$ characterized in (5) for the evalulation of the acceptance probabilities.

In our implementation of Algorithm 1, we choose the initial points β^0 and γ^0 equal to the mean values of the prior distributions of β and γ. It is also critical to verify that the Markov Chain $\{(\beta^s,\gamma^s): s=1,2,\ldots\}$ 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 $2,000$, which we consider as the end of the burn-in period; i.e., the samples during the burn-in period are discarded. We set the thinning parameter equal to 10 (i.e., we collect the samples at every 10 iterations of Algorithm 1); we find 10 is large enough to assure independence in the sampled values but at the same time small enough to achieve efficiency.

In order to illustrate the posterior distributions obtained by using Algorithm 1, we generate $n \in \{2,5,10\}$ historical data points including (i) two explanatory variables generated from Uniform(-1,1) distribution,

Algorithm 1 Sampling from the posterior distribution of the beta-regression parameters β and γ.

Inputs: (i) The current values of β and γ in the Markov chain: β^s and γ^s, and (ii) the proposal distributions $q_1(\beta|\beta^s,\gamma^s)$ and $q_2(\gamma|\beta^s,\gamma^s)$.
Output: The next values of β and γ in the Markov chain: β^{s+1} and γ^{s+1}.
Step 1: Generate a random sample β' from the proposal distribution $q_1(\cdot|\beta^s,\gamma^s)$.
Step 2: Calculate the acceptance probability $\alpha_1(\beta',\beta^s)$ and decide where the chain moves next:
    $\alpha_1(\beta',\beta^s) \leftarrow \min\left\{1, \frac{\pi(\beta',\phi^s|\mathscr{D}_n)}{\pi(\beta^s,\phi^s|\mathscr{D}_n)}\right\}$
    Generate a standard uniform random number u_1.
    $\beta^{s+1} \leftarrow \beta'$ if $u_1 < \alpha_1(\beta',\beta^s)$; and $\beta^{s+1} \leftarrow \beta^s$ if $u_1 > \alpha_1(\beta',\beta^s)$.
Step 3: Calculate the acceptance probability $\alpha_2(\gamma',\gamma^s)$ and decide where the chain moves next:
Step 4: Generate a random sample γ' from the proposal distribution $q_2(\cdot|\beta^{s+1},\gamma^s)$.
    $\alpha_2(\gamma',\gamma^s) \leftarrow \min\left\{1, \frac{\pi(\beta^{s+1},\exp(\gamma')|\mathscr{D}_n)}{\pi(\beta^{s+1},\exp(\gamma^s)|\mathscr{D}_n)}\right\}$
    Generate a standard uniform random number u_2.
    $\gamma^{s+1} \leftarrow \gamma'$ if $u_2 < \alpha_2(\gamma',\gamma^s)$; and $\gamma^{s+1} \leftarrow \gamma^s$ if $u_2 > \alpha_2(\gamma',\gamma^s)$.

(ii) the yield realizations under the true parameters $\beta^* = (0.275, -0.035, 0.025)$ and $\gamma^* = 2.125$. Under the assumption of normal priors for β and γ (i.e. $\beta \sim \mathscr{N}(\beta_0, \Sigma_0)$ where $\beta_0 = (0,0,0)$ and Σ_0 is a 3×3 identity matrix, and $\gamma \sim \mathscr{N}(0,1)$), Figure 1 illustrates the posterior distributions of the mean and the standard deviation associated with the yield random variable for an engineer-to-order product with explanatory variables $\mathbf{x}^* = (1.000, 0.870, -0.062)$. We note that the true value of the mean and standard deviation is equal to $\mu^* = 0.561$ and $\sigma^* = 0.162$ for these explanatory variables. Figure 1 shows that the prior belief of the simulation practitioner on the true value of mean is not informative in the sense that it is equally likely to take values anywhere on the unit interval $(0,1)$. Likewise, the prior information implies the true standard deviation is equally likely to take values on the interval $(0.1, \sqrt{\mu(1-\mu)}$ for any given μ value (see top-left plot in Figure 1 with $n = 0$). As the number of data points increase, we observe a concentration of the posterior distribution on the true values of the mean and the standard deviation. For instance, the mean and standard deviation is almost known with certainty after collecting $n = 10$ data points evidenced by the support of the posterior distribution of the mean and the standard deviation taking values in close proximity of the true values $\mu^* = 0.561$ and $\sigma^* = 0.162$ (see bottom-right plot in Figure 1).

4.2 Minimization of the Posterior Expected Cost

In this section, we present the sample average approximation counterpart of the production batch sizing problem in (6). In particular, we use the posterior samples $\{(\beta_\omega, \gamma_w) : w = 1,\ldots,\Omega\}$ obtained from Algorithm 1 to calculate the posterior samples of the mean and the standard deviation as

$$\mu_\omega = \frac{1}{1+\exp(-\beta_\omega^\top \mathbf{x})} \qquad \sigma_\omega = \sqrt{\frac{\mu_\omega(1-\mu_\omega)}{1+\exp(\gamma_\omega)}}$$

for a specified vector $\mathbf{x}$ that represents the attributes of the engineer-to-order product to be produced. Consequently, the objective of the sample average approximation problem is written as

$$\min_{q\geq 0} \frac{1}{\Omega}\sum_{\omega=1}^{\Omega} \Pi(q;\mu_\omega,\sigma_\omega) \tag{9}$$

by approximating the expected posterior cost in the production batch-sizing problem in (6) with the average of Ω number of posterior realizations of the expected cost. We refer to the optimization problem in (9) as the Joint Estimation-Optimization (JEO) policy as it considers the complete posterior distribution of

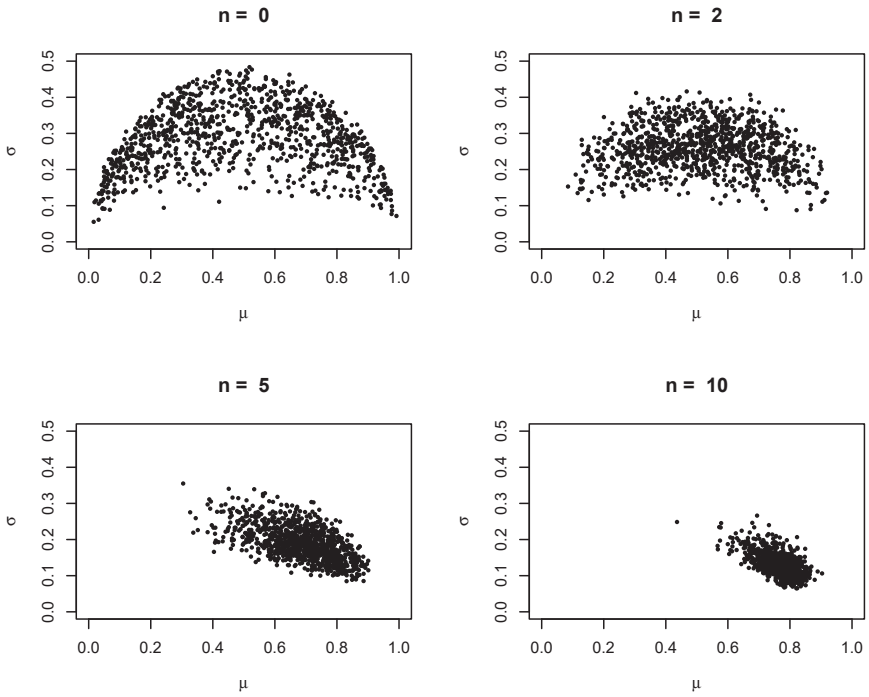

Figure 1: Posterior distribution of the beta-distribution parameters μ and σ.

the beta-regression parameters to make the batch-sizing decision. An alternative practical approach is the Point-Estimate (PE) policy which first estimates the beta-regression parameters and then plug the resulting point estimates in the optimization model. Since the expected cost function Π in (3) can be exactly evaluated and does not require stochastic simulation, the formulation in (9) does not include any intrinsic simulation uncertainty while explicitly modelling the input uncertainty via simulation. That is, in the remainder of the paper, we 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) and Akcay and Martagan (2016) for simulation-based optimization problems facing input uncertainty. Notice that (9) is a nonlinear objective function, and we solve it within seconds by using the GenSA (Generalized Simulated Annealing) R-package developed by Gubian et al. (2012). In our numerical experiments, we observe that the optimal solution converges when the value of Ω exceeds 1000, therefore, we fix the value of Ω to 1000 in the remainder of the paper.

5 NUMERICAL INSIGHTS

In this section, we run a set of numerical experiments to answer the following questions: (i) what is the impact of the penalty cost s and the production cost c on the performance of the PE and the JEO policies? (ii) what is role of the length of the development experiments n and the number of product features K in the performance of the PE and the JEO policies? (iii) when is the JEO most effective in reducing the increase in the expected cost due to the input uncertainty? We perform our numerical analysis by sampling the experimental parameters from the following sets: $n = \{5,10,15\}$, $c = \{1,2\}$, $r = 5$, $s = \{5,10,20,50\}$, $k = \{0.1,0.25,0.5\}$, $M = \{0,5,10\}$, $d = 20$. Table 1 and Table 2 summarize our results by presenting the average cost and batch-size values from the simulation replications at specified values of s,c,N and K. We choose the number of the simulation replications dynamically such that the half-width of the 95% confidence intervals for the expected cost is no more than the 1% of the reported value. We consider that the normal prior distributions for the mean-parameters and the precision parameter of the beta-regression

model have mean 0 and variance equal to 1, implying no strong prior knowledge on the mean and standard deviation; see top-left plot in Figure 1. In each simulation replication, we generate the true values of the beta-regression parameters from these prior distributions, while the historical product features and the features of the engineer-to-order product in production are generated from the Uniform$(-1,1)$ distribution.

Table 1 assumes that there is no product feature available as an explanatory variable ($K=0$); this situation captures the situation where all the development experiments are done for one specific engineer-to order product (i.e., product features are kept constant during the development experiments). In particular, Table 1 compares the expected costs and batch sizes under the optimal policy with perfect information (Opt), the PE policy, and the JEO policy for $n=5$. In the optimal policy with perfect information, the manufacturer knows the true values of the mean and the standard deviation of the beta-distribution that represents the random yield variable. On the other hand, the PE policy uses the maximum likelihood estimation to obtain the point-estimates of the unknown beta-regression parameters and treats them as if they are the correct values.

We make two key observations from Table 1: (i) The increase in the expected cost due to the input uncertainty increases as the penalty cost s increases; furthermore, the JEO policy becomes more effective with increasing s. For instance, the average expected cost associated with the PE policy is only 1.2 units $(60.4-59.2)$ higher than the optimal policy under perfect information, and the JEO reduces the increase in the expected cost from 1.2 units to 1.1 units $(60.3-59.2)$ for the penalty cost s equal to 5 (i.e. when the penalty cost is only the loss of revenue) and for the production cost c equal to 1. On the other hand, when the penalty cost increases is equal to 50 (e.g., implied by a large loss of customer goodwill for agreed but unmet demand), the average expected cost associated with the PE policy is 19.2 units $(139.2-120.0)$ higher than the optimal policy under perfect information, and the JEO reduces the increase in the expected cost from 19.2 units to 8.9 units $(139.2-130.3)$. This can be explained by the shape of the expected cost function around the optimal batch-size quantity: If the per-unit penalty cost is larger than the revenue earned from one unit of the product, the expected cost function has a sharp dive at the optimal batch quantity. On the other hand, if the penalty cost and the revenue are close to each other, then the expected cost function is rather flat around the optimal batch quantity. The flatness implies that choosing a sub-optimal batch size due to incorrect estimation of the beta-regression parameters leads to an expected cost value that is potentially still close to the minimum expected cost under the perfect information. However, the sharp dive of the expected cost function around the optimal batch size implies that choosing a batch size that this slightly different than the optimal one leads to a substantial increase in the expected cost. (ii) The JEO policy adjusts the batch size under the PE policy in a way that balances the cost of underestimating the yield realization and the cost of production. For instance, when the penalty cost s is 50, the JEO policy increases the batch size of the PE policy (i.e., from 70.9 to 81.4) for the production cost equal to 1.

Table 1: Expected costs and batch sizes (in paranthesis) under the optimal policy with perfect information, the PE policy, and the JEO policy; $n=5$ and $K=0$.

	$c=1$			$c=2$		
s	Opt	PE	JEO	Opt	PE	JEO
5	59.2	60.4	60.3	86.1	87.9	87.7
	(37.8)	(36.4)	(32.5)	(26.3)	(20.7)	(20.7)
10	71.6	73.4	73.4	111.0	113.8	112.7
	(52.1)	(53.7)	(44.9)	(38.1)	(34.5)	(33.1)
20	80.5	84.1	82.9	142.9	145.1	144.5
	(58.7)	(56.2)	(57.7)	(49.2)	(47.8)	(45.2)
50	120.0	139.2	130.3	192.2	202.4	196.0
	(87.8)	(70.9)	(81.4)	(67.2)	(77.7)	(72.3)

Table 2: Expected costs and batch sizes (in paranthesis) under the optimal policy with perfect information, the PE policy, and the JEO policy; $c = 1$ and $s = 50$.

	$K = 0$			$K = 1$			$K = 2$		
n	Opt	PE	JEO	Opt	PE	JEO	Opt	PE	JEO
5	120.0	139.2	130.3	128.2	159.2	136.2	134.1	260.0	140.5
	(87.8)	(70.9)	(81.4)	(89.8)	(99.8)	(100.9)	(91.8)	(81.5)	(89.4)
10		128.3	124.9		142.5	134.7		160.0	141.3
		(76.0)	(83.1)		(79.3)	(90.9)		(84.4)	(93.8)
15		123.7	123.2		129.3	131.1		148.5	140.5
		(82.2)	(86.1)		(81.2)	(90.2)		(93.8)	(98.2)

However, when the production cost is 2, the cost of underestimating the yield realization is dominated by the higher production cost, and the JEO policy continues to decrease the batch size of the PE policy (i.e., from 77.7 to 72.3).

Table 2 assumes the availability of the product features as the explanatory variables for estimating the mean of the beta-distributed yield random variable; i.e., $K \in \{0,1,2\}$. We make two key observations from Table 2: (i) The increase in the expected cost due to the input uncertainty increases as the length of the historical data decreases. For instance, when there are $K = 2$ covariates, we observe an increase in the expected cost of the PE policy from 93.8 to 260.0 as the number of development experiments n goes down from 15 to 5. That is, the input uncertainty problem is the most relevant when there is limited data (e.g., often caused by expensive and time-consuming development experiments) to make an inference about the yield random variable. (ii) The JEO policy typically becomes more effective in reducing the input-uncertainty driven increase in the expected cost as the number of explanatory variables increases. For example, when n is equal to 5, we observe that the JEO policy reduces the the average expected cost of the PE policy about 11.6% (i.e., from 159.2 to 136.2) for $K = 1$ and about 46% (i.e., from 260.0 to 140.5) for $K = 2$.

6 CONCLUSION

We study the production batch-sizing decisions of a manufacturer who develops and manufactures an engineer-to-order product with a predetermined set of attributes. The manufacturer has limited historical data about the past yield realizations under different product attributes, and makes inferences about the distribution of the yield distribution through a beta regression model. The manufacturer faces the problem of input uncertainty due to unknown parameters of the beta regression model, affecting the performance of the resulting production batch-sizing decisions. In this paper, we represent the uncertainty in the logistic-regression parameters by approximating their joint posterior distribution via an MCMC algorithm, and then investigate the impact of input uncertainty on the batch-sizing decisions and the expected cost of the manufacturer. We further introduce a sample average approximation of the batch-sizing problem to reduce the impact of input uncertainty on the expected cost of the manufacturer. Future research directions include the production and batch sizing decisions in the context of multiple products and limited capacity.

ACKNOWLEDGMENTS

The second author is grateful to the generous support Marie Sklodowska-Curie Fellowship by the European Commission under the Horizon 2020 program.

REFERENCES

Akcay, A., and B. Biller. 2017. "Input Uncertainty in Stochastic Simulations in the Presence of Dependent Discrete Input Variables". *Journal of Simulation*: to appear.

Akcay, A., and C. G. Corlu. 2017. "Simulation of Inventory Systems with Unknown Input Models: A Data-Driven Approach". *International Journal of Production Research*: to appear.

Akcay, A., and T. Martagan. 2016. "Stochastic Simulation under Input Uncertainty for Contract-Manufacturer Selection in Pharmaceutical Industry". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, and E. Zhou, 2292–2303. Piscataway, New Jersey: Institute of Electrical and Electronics Engineers.

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.

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. 2014. "Quantifying Input Uncertainty via Simulation Confidence Intervals". *INFORMS Journal on Computing* 26 (1): 74–87.

Ben-Zvi, T., and A. Grosfeld-Nir. 2007. "Serial Production Systems with Random Yield and Rigid Demand: A Heuristic". *Operations Research Letters* 35 (2): 235–244.

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

Bollapragada, S., and T. E. Morton. 1999. "Myopic Heuristics for the Random Yield Problem". *Operations Research* 47 (5): 713–722.

Cepeda-Cuervo, E., D. Jaimes, M. Marín, and J. Rojas. 2016. "Bayesian Beta Regression with Bayesianbetareg R-package". *Computational Statistics* 31 (1): 165–187.

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

Cuervo, E. C., and L. G. Lopera. 2015. "Bayesian Beta Regression Models with Joint Mean and Dispersion Modeling". *Monte Carlo Methods and Applications* 21 (1): 49–58.

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.

Ferrari, S., and F. Cribari-Neto. 2004. "Beta Regression for Modelling Rates and Proportions". *Journal of Applied Statistics* 31 (7): 799–815.

Gelman, A., J. B. Carlin, H. S. Stern, and D. B. Rubin. 2014. *Bayesian Data Analysis*, Volume 2. Chapman & Hall/CRC Boca Raton, FL, USA.

Grosfeld-Nir, A., and Y. Gerchak. 2004. "Multiple Lotsizing in Production to Order with Random Yields: Review of Recent Advances". *Annals of Operations Research* 126 (1): 43–69.

Grosfeld-Nir, A., Y. Gerchak, and Q.-M. He. 2000. "Manufacturing to Order with Random Yield and Costly Inspection". *Operations Research* 48 (5): 761–767.

Gubian, S., Y. Xiang, B. Suomela, and J. Hoeng. 2012. "GenSA: R Functions for Generalized Simulated Annealing". *R Package Version* 1 (2): 70.

Inderfurth, K., and G. P. Kiesmüller. 2015. "Exact and Heuristic Linear-Inflation Policies for an Inventory Model with Random Yield and Arbitrary Lead Times". *European Journal of Operational Research* 245 (1): 109–120.

Kim, S., R. Pasupathy, and S. G. Henderson. 2015. "A Guide to Sample Average Approximation". In *Handbook of Simulation Optimization*, 207–243. New York: Springer.

Lee, C.-Y., and T. Lu. 2015. "Inventory Competition with Yield Reliability Improvement". *Naval Research Logistics (NRL)* 62 (2): 107–126.

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

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*. Cambridge: MIT press.

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.

Schmitt, A. J., and L. V. Snyder. 2012. "Infinite-horizon Models for Inventory Control under Yield Uncertainty and Disruptions". *Computers & Operations Research* 39 (4): 850–862.

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.

Tomlin, B. 2009. "Impact of Supply Learning When Suppliers are Unreliable". *Manufacturing & Service Operations Management* 11 (2): 192–209.

Wein, A. S. 1992. "Random Yield, Rework and Scrap in a Multistage Batch Manufacturing Environment". *Operations Research* 40 (3): 551–563.

Xie, W., B. Nelson, and R. Barton. 2014a. "Statistical Uncertainty Analysis for Stochastic Simulation". Working paper, Department 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.

Yano, C. A., and H. L. Lee. 1995. "Lot Sizing with Random Yields: A Review". *Operations Research* 43 (2): 311–334.

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 School of Industrial Engineering 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 School of Industrial Engineering 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 with applications in engineer-to-order systems. She is the recipient of the first prize in 2016 POMS Applied Research Challenge, and the honorable mention in 2016 INFORMS George B. Dantzig Dissertation Award. Her email address is t.g.martagan@tue.nl.