Proceedings of the 2020 Winter Simulation Conference
K.-H. Bae, B. Feng, S. Kim, S. Lazarova-Molnar, Z. Zheng, T. Roeder, and R. Thiesing, eds.

ADVANCED STATISTICAL METHODS: INFERENCE, VARIABLE SELECTION, AND EXPERIMENTAL DESIGN

Ilya O. Ryzhov

Robert H. Smith School of Business
University of Maryland
7699 Mowatt Lane
College Park, MD 20742, USA

Qiong Zhang

School of Mathematical and Statistical Sciences
Clemson University
O-110 Martin Hall
Clemson, SC 29634, USA

Ye Chen

Statistical Sciences & Operations Research
Virginia Commonwealth University
1015 Floyd Avenue
Richmond, VA 23284, USA

ABSTRACT

We provide a tutorial overview of recent advances in three methodological streams of statistical literature: design of experiments, variable selection, and approximate inference. For some of these areas (such as design of experiments), their connections to simulation research have long been known and appreciated; in other cases (such as variable selection), however, these connections are only now beginning to be built. Our presentation focuses primarily on the statistical literature, aiming to show state-of-the-art thinking with regard to these problems, but we also point out possible opportunities to use these methods in new ways for both theory and applications within simulation.

1 INTRODUCTION

Statistics has long exerted a formative influence on simulation research – many foundational methodological areas in simulation, such as stochastic approximation (Robbins and Monro 1951), ranking and selection (Dudewicz and Dalal 1975), and design of experiments (Titterington 1975), were originally pioneered by statisticians. Statistics is fundamental to simulation output analysis (Glynn and Iglehart 1990), gradient estimation (Glynn 1987), metamodeling (Kleijnen 2009), uncertainty quantification (Song et al. 2014), and many other techniques of interest to the WSC community.

The authors of this tutorial all work on topics within the mainstream of simulation research. All three of us have found that, in problems that involve elements of statistics, optimization, and applied probability, the statistical aspect often turns out to be the most challenging and important. We have also found the recent statistical literature to be very useful for dealing with these challenges. With the increased prominence of big data analytics and machine learning in virtually every subfield of operations research, the importance of pure statistics for simulation research is only likely to grow.

This tutorial surveys *recent* advances in three major areas of statistics, namely 1) design of experiments, 2) variable selection, and 3) approximate inference. To an extent, this choice of topics is driven by our individual areas of expertise. However, we believe that all three areas are highly relevant to simulation

978-1-7281-9499-8/20/$31.00 ©2020 IEEE

research: thus, design of experiments is closely related to ranking and selection and simulation-based optimization (Hong and Nelson 2009); variable selection is potentially applicable to the emerging area of simulation analytics (Jiang et al. 2020); and approximate inference is broadly applicable to optimization under uncertainty when fast computation is required.

Because we focus on recent work, our presentation necessarily focuses more on statistical literature rather than simulation literature. While some of these statistical methods may be immediately applicable to existing, well-known problems in simulation, it is not necessarily our objective to show this here. In fact, in some cases, it may well be that some of these methods have not *yet* found their applications in the WSC community. We hope that these are exactly the cases where this tutorial will be the most useful, by helping to facilitate new streams of simulation research where these statistical ideas will be relevant.

In each of the main sections of the tutorial, we will aim to make connections between the material being presented and specific applications in simulation. However, we ask the reader to keep in mind that some of these may be only *potential* applications.

2 DESIGN OF EXPERIMENTS

In many branches of science and engineering, computer experiments on virtual systems are a critical tool for the study and analysis of complex physical processes. In this way, costly prototypes in the early design phase are replaced by simulations, providing considerable productivity gains. When conducting these experiments, the objectives often include 1) finding the input configuration that produces the most desirable outcome, and 2) estimating a performance measure over the input space. Usually, no explicit expression is available for the outcome as a function of the input, and running experiments can be computationally expensive, so it is necessary to efficiently allocate design points over the input space and use computationally inexpensive statistical models to emulate computer experiments.

Early developments in experimental design mostly focus on physical experiments. Wu (2015) points out that the principles of designs of computer experiments are different from designs of physical experiments, and the three principles of blocking, replication, and randomization are inessential or irrelevant to the design of computer experiments. For deterministic computer models based on partial differential equations, it is often necessary to cover the experimental region with the design points, known as the "space-filling property." Below, we review three popular directions for constructing space-filling designs from the recent literature on deterministic computer models.

Classical experimental design techniques such as factorial design and response surface methodology are well-known in the simulation community (Kelton and Barton 2003; Barton 2013). More recently, it has become popular to allocate design points sequentially; however, as parallel processors become available to practitioners, the allocation of a relatively large batch of design points may become competitive in terms of utilizing this advanced computing resource (Nelson 2016; Zhang et al. 2020). Thus, the main ideas in designing "one-shot" experiments may also be useful in stochastic simulation. In addition, stochastic simulation requires initial replications to assess the simulation estimation error as mentioned in Ankenman et al. (2010); thus, the space-filling property is still useful to set up the base design before determining the number of replications (Ankenman et al. 2010; Law 2017). There may also be other similarities between stochastic simulation and certain types of deterministic experiments. For example, Chen et al. (2013) study a stochastic simulation model with both quantitative and qualitative factors.

2.1 Optimal Space-Filling Designs

Under a fixed experimental budget n, the minimum of the distances between two points should be as large as possible to achieve better space-filling properties. Johnson et al. (1990) proposed the maximin distance design as the solution to the optimization problem

$$\max_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathscr{X}} \min_{i \neq j \in 1,\ldots,n} d(\mathbf{x}_i, \mathbf{x}_j), \tag{1}$$

where $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ is the i-th input point from the p-dimensional input space $\mathscr{X}$, and $d(\cdot,\cdot)$ is the Euclidean distance between two input points. The minimum pointwise distance of a design is called the separation distance; this is the quantity maximized by the maximin design.

Problem (1) is usually challenging to solve. Morris and Mitchell (1995) develop an alternative design criterion

$$\min_{\mathbf{x}_1,\ldots,\mathbf{x}_n \in \mathscr{X}} \left\{ \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{d^k(\mathbf{x}_i, \mathbf{x}_j)} \right\}^{1/k}, \tag{2}$$

for some $k > 0$. If k is large enough, the resulting design achieves the maximin distance in (1). However, the designs obtained from the maximin criterion are only space-filling in the entire space $\mathscr{X}$. To ensure good space-filling properties for all subspaces of $\mathscr{X}$, Joseph et al. (2015) propose the maximum projection design to minimize

$$\Phi(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \left(\sum_{i=1}^{n-1} \sum_{j=i+1}^{n} \frac{1}{\prod_{l=1}^{p} (x_{il} - x_{jl})^2} \right)^{1/p}, \tag{3}$$

which is obtained by taking the expectation of the objective in (2) with respect to the weight parameters in each dimension under a non-informative prior.

The above optimal space-filling designs are model-free, i.e., they do not rely on any statistical modeling assumptions on the output surface. Optimal space-filling designs can also be constructed based on models to improve the accuracy of the emulator for certain types of computer experiments. A widely used model for the outcome $y(\mathbf{x})$ of an experiment is the Gaussian process (GP)

$$y(\mathbf{x}) = \mu + z(\mathbf{x}),$$

where μ is a deterministic mean, and $z(\mathbf{x})$ is a mean-zero GP with variance σ^2 and correlation function $R(\mathbf{x}, \mathbf{x}')$. The correlation function value $R(\mathbf{x}, \mathbf{x}')$ decreases with the distance between $\mathbf{x}$ and $\mathbf{x}'$. Under the GP setting, space-filling designs can be constructed to minimize the expected or the maximum prediction error over the input space $\mathscr{X}$ (Sacks et al. 1989). Also, Shewry and Wynn (1987) develop the maximum entropy design obtained by maximizing the determinant of the correlation matrix in GP.

There are some connections between model-free and model-based designs. For example, when maximizing the determinant of the correlation matrix in a GP, the resulting maximum entropy designs (Shewry and Wynn 1987) tend to have smaller off-diagonal entries, i.e., the distance between different design points will be larger, thus meeting the maximin distance design criterion in (1). As another example, Joseph et al. (2015) point out that the model-free maximum projection design criterion in (3) can also be derived by minimizing the sum of the expected off-diagonal entries in the correlation function of GP under a non-informative prior.

2.2 Latin Hypercube-Based Space-Filling Designs

Latin hypercube sampling is another important technique to achieve space-filling. An Latin hypercube design (LHD) of n runs $\mathbf{x}_1, \cdots, \mathbf{x}_n$ with p inputs in $[0,1]$ can be constructed by

$$\mathbf{x}_i = (x_{i1}, \ldots, x_{ip}) \quad \text{with} \quad x_{ij} = \frac{\pi_j(i) - U_{ij}}{n}, \text{ for } 1 \le i \le n, 1 \le j \le p \tag{4}$$

where $\pi_1, \ldots, \pi_p$ are independent permutations of $1, \ldots, n$, and each U_{ij} is an independent $U[0,1)$ random variable independent of the values π_j. As can be seen from (4), LHDs are convenient to generate and can accommodate any number of factors. An LHD has maximum uniformity when projected onto any single dimension, known as the univariate stratification property. Figure 1 displays two LHDs with four runs and two factors. In this figure, by dividing the input space $[0,1]$ from either dimension into four equally spaced intervals, there is exactly one point located at each interval. In terms of estimating the mean performance

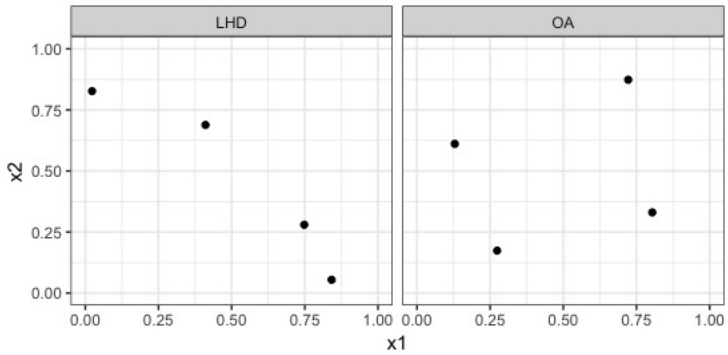

Figure 1: LHDs with four runs and two factors. Left: a regular LHD generated by (4); Right: an orthogonal array-based LHD with strength two from (Tang 1993).

of computer experiments, an LHD can achieve smaller variance than i.i.d. samples (McKay et al. 1979; Stein 1987).

However, an LHD can not guarantee stratification over higher-dimensional spaces as shown in the left panel of Figure 1. Recent work has focused on improving the space-filling property of a regular LHD generated from (4) using ideas from optimization; see, e.g., Tang (1993), Owen (1994), Joseph and Hung (2008). In particular, Morris and Mitchell (1995) search from random LHDs to optimize the optimal space-filling criterion in (2). The drawback of this type of design is that it is challenging to evaluate the optimality gap generated from heuristic optimization approaches. Also, from the algebraic perspective, Tang (1993) uses orthogonal arrays to construct Latin hypercubes. Originating from an orthogonal array (OA) of size n, p dimensions, s levels and strength t, the resulting OA-based Latin hypercube designs also stratify each t-dimensional margin, which strengthens the one-dimensional stratification property of a regular LHD (as shown in the right panel of Figure 1). The limitation is that OAs can not be constructed for any arbitrary combination of size and dimension. Therefore, the OA-based LHDs have this restriction to be applied to examples with any run size and number of input factors.

2.3 Designs With Special Structure

Optimal designs often require special structure motivated by practical considerations in the implementation of computer experiments. Such structure can occur in multi-fidelity computer experiments, sequential batched experimentation, and computer experiments with both qualitative and quantitative inputs.

Nested Latin hypercube designs (Qian 2009) are proposed for experiments conducted through multi-fidelity computer models with different levels of accuracy. A nested Latin hypercube design can contain multiple layers. As shown in Figure 2, the first layer is an LHD containing four runs, the second layer adds four runs to construct an LHD with eight runs, and the third layer adds eight additional runs to construct an LHD with 16 runs. In practice, the smaller LHD can be used to conduct experiments on the high-fidelity computer model, which is typically more time-consuming, and the larger LHD is used for the cheaper low-fidelity model. Combining the experimental outcomes, variance reduction can be achieved on the estimation of mean performance. Nested designs can also be applied to sequential batched experimentation, which adds more runs of experiments step by step. Designs with a nested structure can also be constructed based on low discrepancy sequences, such as Sobol' sequences (Haaland and Qian 2010).

Sliced LHDs (Qian 2012) are proposed for computer experiments with both qualitative and quantitative factors. A sliced LHD contains multiple slices with equal runs, with each slice forming a small LHD, and the whole design forming a large LHD. The number of slices is equal to the total number of level combinations of the qualitative factors, and each slice can be allocated to the design of quantitative factors associated with each qualitative level. This design can also be used for parallelized computer experiments,

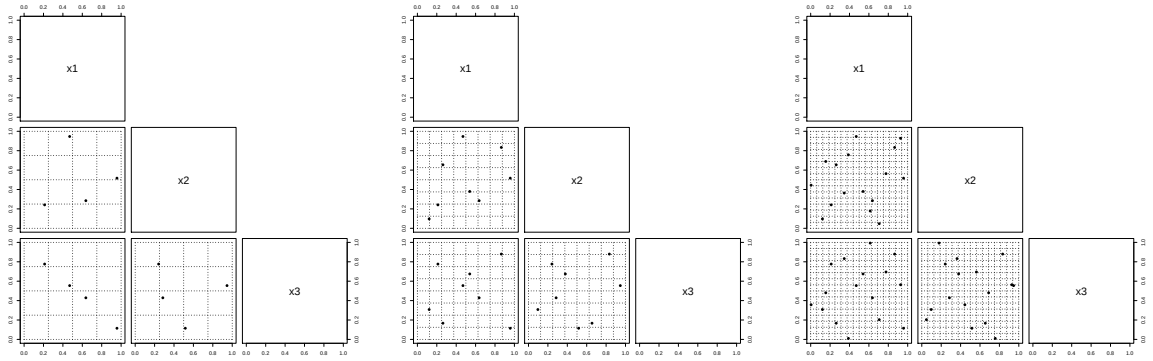

Figure 2: Two dimensional projections of design points from a nested LHD with 16 runs, three factors, and three layers. Left: the first layer is an LHD with four runs; Middle: the second layer is an LHD with eight runs; Right: the third layer is an LHD with 16 runs.

and cross-evaluation of emulators for computer models (Zhang and Qian 2013). Recent work has focused on improving the space-filling property of designs with special structure (He 2019; Joseph et al. 2020; Zhou et al. 2020).

2.4 Optimal Designs For A/B Testing

A/B testing refers to the design and analysis of an experiment to compare to treatments applied to different experimental units. Large-scale A/B testing is widely implemented at technology companies such as Facebook, LinkedIn, and Netflix, to compare different algorithms, web designs, and other online products and services. In its simplest form, the experimental design problem is to determine the proportions of test units allocated to two options A and B to reduce the uncertainty of the comparison (Shahriari et al. 2015), which is related to the optimal design literature.

Optimal designs can be developed based on the simple ordinary least-squares (OLS) regression model

$$y = \beta^\top \mathbf{x} + \varepsilon, \tag{5}$$

where y is an observed value ("response"), $\mathbf{x} \in \mathbb{R}^p$ is a vector of data (or "features") describing the observation, $\beta \in \mathbb{R}^p$ is a vector of (unknown) regression coefficients describing the effects of $\mathbf{x}$ on y, and ε is an independent zero-mean noise. Given n observations of the form $(\mathbf{x}_n, y_n)$, we estimate β by solving

$$\theta_n^{OLS} = \arg\min_{\theta \in \mathbb{R}^p} \sum_{i=1}^{n} \left(y_i - \theta^\top \mathbf{x}_i \right)^2. \tag{6}$$

One can then formulate the objective of optimal design as a function of the covariance matrix of the estimated coefficients θ_n^{OLS}. For example, the D-optimal design objective is defined as the determinant of the covariance matrix, and the D_a-optimal criterion is the determinant of the covariance matrix of a linear combination of θ_n^{OLS} (Sinha 1970).

Optimizing these criteria becomes quite challenging in the presence of additional covariates. If p is small, A and B options can simply be allocated by the stratification of each covariate, but this approach does not scale. Instead, one can use a model that treats the covariates as additive factors, as in

$$y = \beta x + \mathbf{z}^\top \gamma + \varepsilon, \tag{7}$$

where y is the outcome, $x \in \{-1, 1\}$ represents the allocation of A/B options, β is the treatment effect, $\mathbf{z}$ is a vector of covariates (including an intercept term) with coefficients γ, and ε is the error term. Under

this model setting, the accuracy of our comparison of A and B critically depends on the accuracy of our estimation of β.

One recent approach is a special case of D_a-optimal design that minimizes the variance of our estimate of β. Recently, Bhat et al. (2019) have developed offline and online mathematical programming approaches to solve this problem. The literature on causal inference approaches the problem by balancing covariates, that is, dividing the test units into two or multiple groups with similar distributions of covariates between different groups. In particular, Morgan and Rubin (2012) have proposed a rerandomization approach to reduce the Mahalanobis distance between the covariates of the two groups. This objective is equivalent to D_a optimal design under a linear model with additive treatment and covariate effect, so the covariate balancing problem can be considered in a unified framework (Kallus 2018). Zhang et al. (2020) consider optimal designs to improve the personalized decision of A and B options with application in precision medication. In the simulation literature, similar problems have been considered by, e.g., Han et al. (2016) and Shen et al. (2017), with the main distinction being that the objective optimized in these papers is economic rather than statistical; for example, one might design experiments to maximize the expected value of y, rather than minimizing a statistical criterion as in the experimental design literature.

There are also important special cases of this problem where the covariates represent network data, such as connections between users on Facebook or Twitter. Some specialized techniques have been developed for such settings. Randomized treatment allocations are proposed by Xu et al. (2015) and Basse and Airoldi (2018) in order to reduce the effects of network correlation or to reduce the error of the estimated treatment effect. Pokhilko et al. (2019) use a conditional autoregressive model to incorporate network structure, and develop a D-optimal design approach to A/B testing as an extension of the offline optimal design in Bhat et al. (2019) to the setting of social networks.

3 VARIABLE SELECTION

Let us now turn our attention to problems where the data are given, rather than designed. For illustrative purposes, we again consider the ordinary least-squares regression model (5). If this model correctly describes the relationship between $\mathbf{x}$ and y, we will have $\theta_n^{OLS} \to \beta$ (consistency of the OLS regression estimator) as $n \to \infty$ under some mild conditions (Lai and Wei 1982) on the sequence $\{x_n\}_{n=1}^{\infty}$.

There are, however, reasons not to use (6) even when the model (5) is believed to be accurate. Suppose that the number p of features is large, potentially even greater than the number n of observations. In such a case, (6) no longer has a unique solution, so it is not possible to recover the effects of individual features. At the same time, it may be that most of these effects are zero, i.e., the size of the set $\mathscr{A} = \{j \leq p : \beta_j \neq 0\}$ is very small relative to p (and smaller than n). In other words, we have a very large volume of data, but most of the data are not useful – they are simply obscuring a small number of important features.

Even if $n > p$, however, the same issue can arise. For example, pricing decisions at hotels are influenced by similar decisions at other hotels (Li et al. 2018): thus, if there are m hotels, there are $m^2 - m$ potential effects (large p). These effects can be estimated from user search histories at a travel aggregation website (large n). One can download enough data to ensure $n > p$, but any given hotel is most likely not influenced by all $m-1$ competitors; it is more likely that the main influences come from a few hotels with similar star rating or location.

Keeping uninformative features in the model will make it more difficult to obtain reliable estimates from (6). First, these features will add noise, reducing the accuracy of our estimates of the important effects. Second, with more features, it becomes more likely that many of them will be strongly correlated, i.e., that we are essentially keeping track of multiple copies of the same effect, leading to degraded model performance. Finally, with large p, there is a risk of spurious correlation, meaning that, by random chance, some of the features appear to exhibit patterns that are not actually present in the data-generating process; in other words, we are more likely to falsely identify $j \notin \mathscr{A}$ as being relevant (Fan et al. 2014).

The methodology of *variable selection* seeks to recover $\mathscr{A}$ and thereby obtain more accurate estimates of β_j for $j \in \mathscr{A}$. In the following, we will survey two different approaches to the problem: regularized

estimation methods (Section 3.1) that require assumptions such as (5), and screening methods (Section 3.2) that can be conducted before any estimation takes place. Before we proceed, we first comment on the relevance of this problem to the simulation community.

As of this writing, variable selection has only begun to enter the simulation literature through "simulation analytics" (Lin and Nelson 2016). The main distinguishing characteristic of this problem class is that the "data" used to predict the performance of a simulated system may consist of the system's entire simulated trajectory (Nelson 2016). For example, if we are simulating a complex service system over a long period of time, $\mathbf{x}$ may include the timestamp of every event that occurs, enabling a more granular analysis of the causes of long delays. At the same time, the fact that we can now store this volume of data does not mean that all of it is useful. For this reason, Jiang et al. (2020) uses variable selection (Lasso, discussed in Section 3.1) to reduce model size and improve prediction quality. We hope that the statistical tools surveyed here may be useful to researchers working on such problems.

3.1 Regularized Estimation Via the Lasso

Perhaps the best-known variable selection method is the Lasso, first introduced by Tibshirani (1996). Returning to the linear regression model from (5) we solve the modified estimation problem

$$\theta_n^{Lasso} = \arg\min_{\theta\in\mathbb{R}^p} \sum_{i=1}^n \left(y_i - \theta^\top \mathbf{x}_i\right)^2 + \lambda \sum_{j=1}^p \left|\theta_j\right|. \tag{8}$$

The second term in (8) is a penalty incurred by assigning nonzero values to the coefficients θ_j. Because the absolute value is not differentiable at zero, the Lasso penalty will tend to encourage setting $\theta_j = 0$ rather than simply reducing the magnitudes. We thus make a tradeoff between a model that is more accurate (has lower squared error) in describing the given data vs. a more compact model with fewer nonzero coefficients. Because the two terms in (8) do not have the same "units," a scaling parameter $\lambda \geq 0$ controls the relative importance of the penalty term: as λ increases, the number of nonzero coefficients will shrink.

Letting $\theta_n^{Lasso}(\lambda)$ be the solution to (8) for a given λ value, the set

$$\widehat{\mathscr{A}_n}(\lambda) = \left\{j \leq p : \theta_{n,j}^{Lasso}(\lambda) \neq 0\right\}$$

tells us which variables have been selected by the method. Lasso performs selection and estimation simultaneously, since we also have numerical values $\theta_{n,j}^{Lasso}(\lambda)$ for any $j \in \widehat{\mathscr{A}_n}$.

Although θ_n^{Lasso} technically has higher squared error than the OLS estimator θ_n^{OLS}, in practice the Lasso estimator will perform better out of sample. In fact this is one way to optimize the choice of λ. As is commonly done in machine learning (Hastie et al. 2009), one partitions the available data into training and test sets, solves (8) using only the training data, then evaluates the sum of squared errors incurred by this estimator on the *test* data. One can repeat this process for many values of $\lambda \geq 0$ (which includes the OLS estimator as a special case) and select the value that produces the best results. As an alternative, if one does not wish to partition the data, one can also select λ to optimize an information criterion such as AIC or BIC (Zou et al. 2007). For example, under the BIC (Bayesian Information Criterion), the optimal choice of λ is given by

$$\lambda^* = \arg\min_{\lambda\geq 0} \sum_{i=1}^n \left(y_i - \left(\theta_n^{Lasso}(\lambda)\right)^\top \mathbf{x}_i\right)^2 + \left|\widehat{\mathscr{A}_n}(\lambda)\right| \log n.$$

A rich theory is available for Lasso. It can be shown that the method recovers the true set $\mathscr{A}$ of relevant features as $n \to \infty$ (Zhao and Yu 2006), even if p grows faster than n, and much of this theory carries over to countless extensions and generalizations of the Lasso concept. In general, one typically solves some variant of

$$\theta_n^{Lasso} = \arg\min_{\theta\in\mathbb{R}^p} -\sum_{i=1}^n \log L(y_i; \mathbf{x}_i, \theta) + \lambda \sum_{j=1}^p \left|\theta_j\right|, \tag{9}$$

where $L(y;\mathbf{x},\theta)$ is the likelihood of observing y given the data $\mathbf{x}$ and model parameters θ. This framework encompasses generalized linear models or GLMs (Van de Geer 2008), which satisfy

$$\mathbb{E}(y) = g\left(\beta^\top \mathbf{x}\right), \tag{10}$$

where g is some user-specified nonlinear link function (such as the logistic function, if we wish to use a logistic regression model) and β is a vector of true parameters. There are other extensions to hazard rate estimation (Gaiffas and Guilloux 2012), quantile estimation (Li et al. 2010), panel data (Ibrahim et al. 2011), nonparametric models (Huang et al. 2010) and many other statistical techniques. One can also impose more detailed constraints on the selection set $\widehat{\mathscr{A}_n}$. For example, in the group Lasso method (Meier et al. 2008), the set $\{1,...,p\}$ of features is partitioned into disjoint groups $G_1,...,G_\ell$, and (9) becomes

$$\theta_n^{Lasso} = \arg\min_{\theta\in\mathbb{R}^p} -\sum_{i=1}^n \log L(y_i;\mathbf{x}_i,\theta) + \lambda \sum_{l=1}^{\ell} \left(\sum_{j\in G_l} (\theta_j)^2\right)^{\frac{1}{2}}. \tag{11}$$

This penalty structure has the effect of selecting (or not selecting) entire groups of features, so that $j \in G_l$ is selected if and only if all other $j' \in G_l$ are. This may be desirable in some applications: for example, G_l may represent a categorical variable modeled as a set of dummy variables, and we may wish to either include every possible category in our model, or none of them.

Despite the power and versatility of the Lasso method, it is subject to several issues. First, as we have mentioned, (8) and its variants attempt to perform selection and estimation simultaneously. This seems like an advantage, but the coefficients returned by Lasso are usually biased. The bias may be corrected by refitting a model of the desired type to the selected features (Belloni and Chernozhukov 2013); for example, if we are in the setting of OLS regression, we first solve (8) and then fit a new OLS model only to the features in $\widehat{\mathscr{A}_n}$.

Second, the Lasso penalty complicates computation, as can be seen from the OLS setting where θ_n^{OLS} has a closed-form expression, but θ_n^{Lasso} does not. Fast computation of the Lasso estimator is an active area of research (Shi et al. 2010, Yang and Zou 2015), but, nonetheless, if we are working with a sufficiently complex class of models, it may not be practical to solve the Lasso problem when n and p are large. In such cases, it may be necessary to run Lasso on a small "subsample" bootstrapped from the large dataset (Kleiner et al. 2014).

Finally, the Lasso concept inherently requires us to specify a model, such as linear regression in (8) or a particular generalized linear model in (9). If the model is misspecified to begin with, it does not make much sense to ask whether or not the "true" regression coefficients are zero; furthermore, in a practical application, we may wish to defer the choice of model until after the irrelevant features have been removed. If this is a concern, one may wish to consider an alternative to Lasso.

3.2 Model-Free Selection Via Sure Independence Screening

The concept of sure independence screening (SIS) was introduced by Fan and Lv (2008). To illustrate it, let us return to the linear regression model of (5). Given two vectors $\mathbf{x},\mathbf{y} \in \mathbb{R}^n$, define

$$\hat{C}(\mathbf{x},\mathbf{y}) = \frac{1}{n-1}\sum_{i=1}^n \left(x_i - \frac{1}{n}\sum_{i'=1}^n x_{i'}\right)\left(y_i - \frac{1}{n}\sum_{i'=1}^n y_{i'}\right) \tag{12}$$

and

$$\hat{\rho}(\mathbf{x},\mathbf{y}) = \frac{\hat{C}(\mathbf{x},\mathbf{y})}{\sqrt{\hat{C}(\mathbf{x},\mathbf{x})\hat{C}(\mathbf{y},\mathbf{y})}}. \tag{13}$$

Of course, these are just the usual estimators of covariance and correlation given n independent samples from a bivariate distribution. Now, for $j = 1, ..., p$, denote by $\mathbf{x}_{\cdot,j} = (x_{1,j}, ..., x_{n,j})$ to be the vector of observed values for the jth feature only; similarly, let $\mathbf{y} = (y_1, ..., y_n)$ be the vector of observed responses. For some fixed $0 < c \leq 1$, the set

$$\widehat{\mathscr{A}_n}(c) = \left\{ j \leq p : \left| \hat{\rho}\left(\mathbf{x}_{\cdot,j}, \mathbf{y}\right) \right| \geq c \right\} \tag{14}$$

contains all the features selected by the SIS method. In words, we estimate the marginal correlation between the jth feature and the response, and remove the feature from our model if this quantity is below some pre-specified threshold c. Since correlation can be viewed as a weak measure of dependence, we are screening out any feature if the response does not (marginally) depend on it sufficiently strongly. The threshold c can be chosen in the same way as the regularization parameter λ in Lasso; unfortunately, no variable selection method is entirely tuning-free.

Once (14) has been found, we are free to fit a model of our choice to the set $\widehat{\mathscr{A}_n}(c)$ of selected features. This approach separates screening from estimation: strictly speaking, (14) does not require us to assume an OLS regression model. The main theoretical guarantee (the so-called "sure screening property") proved in the SIS literature is of the form $P\left(\mathscr{A} \subseteq \widehat{\mathscr{A}_n}\right) \to 1$ for large n, meaning that SIS is allowed to select false positives (report irrelevant features as being relevant). The idea is that the practitioner will first run SIS to reduce the size of the data (since SIS is exceptionally easy to implement), and only then select a model and possibly even conduct additional variable selection to remove the remaining false positives.

In reality, however, (14) is not quite model-free, since covariance and correlation are accurate measures of dependence only when the relationship between $\mathbf{x}$ and y is linear. For that reason, (Fan and Lv 2008) proves the sure screening property only under the assumption that the data are being generated by (5). If we use a generalized linear model, as in (10), (12)-(13) may no longer accurately identify the relevant features. In such a setting, Fan and Song (2010) proposes the following approach. For the jth feature, we solve the marginal maximum likelihood problem

$$\tilde{\theta}_j = \arg\max_{\theta_j} \sum_{i=1}^{n} \log L\left(y_i; x_{i,j}, \theta_j\right),$$

that is, we fit a GLM of our chosen class, but only to the jth feature, with no other features present. One can then decide to select the feature or screen it out based on the magnitude $\left|\tilde{\theta}_j\right|$ or perhaps the p-value returned by the GLM for this coefficient.

With the selection criterion thus redefined, the sure screening property can again be proved. Variants of the SIS concept, with different measures of marginal dependence customized to different model classes, were then developed for hazard rate estimation (Zhao and Li 2012), nonparametric models (Fan et al. 2011) and many other settings. Of particular interest to our discussion is the paper by Li et al. (2012), which combined the SIS concept with a very general measure of dependence called "distance covariance," developed by Székely et al. (2007) and Székely and Rizzo (2009). Let X and Y be scalar random variables with respective characteristic functions $\phi_X(t)$ and $\phi_Y(t)$, and let $\phi_{X,Y}(s,t)$ be their joint characteristic function. The distance covariance between X and Y is given by

$$\Delta(X,Y) = \left(\int \left| \phi_{X,Y}(s,t) - \phi_X(s)\phi_Y(t) \right|^2 \left(\pi^2 s^2 t^2\right)^{-1} ds\, dt \right)^{\frac{1}{2}}, \tag{15}$$

and, correspondingly, the distance correlation (DC) is defined as

$$\delta(X,Y) = \frac{\Delta(X,Y)}{\sqrt{\Delta(X,X)\Delta(Y,Y)}},$$

by analogy with Pearson correlation. It is shown that (15) equals zero if and only if X and Y are independent, which is *not* true for the classical covariance.

Suppose now that we have n independent samples $(\mathbf{x}_i, y_i)$ of data. Székely et al. (2007) proposed, and proved the consistency of, the estimator

$$\begin{aligned}\widehat{\Delta}(\mathbf{x},\mathbf{y}) &= \left(\widehat{S}_1 + \widehat{S}_2 - 2\widehat{S}_3\right)^{\frac{1}{2}}, \\ \widehat{\delta}(\mathbf{x},\mathbf{y}) &= \frac{\widehat{\Delta}(\mathbf{x},\mathbf{y})}{\sqrt{\widehat{\Delta}(\mathbf{x},\mathbf{x})\widehat{\Delta}(\mathbf{y},\mathbf{y})}},\end{aligned}$$

where

$$\begin{aligned}\widehat{S}_1 &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|\cdot\left|y_i - y_j\right| \\ \widehat{S}_2 &= \left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|x_i - x_j\right|\right)\cdot\left(\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\left|y_i - y_j\right|\right) \\ \widehat{S}_3 &= \frac{1}{n^3}\sum_{i=1}^{n}\sum_{j=1}^{n}\sum_{l=1}^{n}\left|x_i - x_l\right|\cdot\left|y_j - y_l\right|.\end{aligned}$$

Note that this estimator is purely data-driven and does not require any knowledge of the distribution of (X,Y) other than very general assumptions on the existence of its moments.

The SIS method proposed by Li et al. (2012) simply returns the selection set

$$\widehat{\mathscr{A}_n}(c) = \left\{j \le p : \left|\hat{\delta}\left(\mathbf{x}_{\cdot,j},\mathbf{y}\right)\right| \ge c\right\},$$

and is shown to retain the sure screening property. This version of SIS is truly model-free, as DC does not require any assumptions about the functional dependence of y on x.

This generality comes at a cost. The validity of SIS crucially depends on the degree to which the relevance of the jth feature can be captured in its marginal dependence on the response. This may not be the case; there may be complex interactions between features that make it possible to detect their relevance only when they are all considered simultaneously. Lasso would be in a better position than SIS to detect such forms of joint dependence, precisely because it includes all of the features in the penalized estimation problem. On the other hand, because SIS focuses on marginal dependence, it often runs much faster than Lasso in large applications: since the computational complexity of (9) is polynomial in both n and p, it is much easier to solve p marginal likelihood problems (or compute p DC estimators) than to solve one problem of size $n \times p$. We may note, however, that the two approaches need not be in opposition to each other, and one is always free to select a model and run Lasso after a preliminary screening step using SIS.

4 APPROXIMATE INFERENCE

Once more, let us consider the OLS regression model (5). It is well-known that (6) is solved by $\theta_n^{OLS} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$, where $\mathbf{X}$ is the matrix whose ith row is the observation $\mathbf{x}_i$. The same computation can be performed recursively. Let θ_n be the OLS estimator given the data $(\mathbf{x}_i, y_i)_{i=1}^n$, and let $\Sigma_n = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$. If a new observation $(\mathbf{x}_{n+1}, y_{n+1})$ becomes available, the recursive update

$$\Sigma_{n+1}^{-1} = \Sigma_n^{-1} + x_{n+1}x_{n+1}^\top, \tag{16}$$

$$\theta_{n+1} = \theta_n - \left(\mathbf{x}_{n+1}^\top\theta_n - y_{n+1}\right)\Sigma_{n+1}x_{n+1} \tag{17}$$

yields the correct OLS estimator for the data $(\mathbf{x}_i, y_i)_{i=1}^{n+1}$. Due to the Sherman-Morrison formula, Σ_{n+1} can be computed without explicit matrix inversion, and therefore (16)-(17) can be computed very quickly. One can use an arbitrary θ_0 and set $\Sigma_0 = \kappa \cdot I$ for small $\kappa > 0$ while preserving statistical consistency.

This is very useful in applications where statistical models are used for online decision-making. Our discussion of A/B testing in Section 2.4 was motivated by e-commerce applications where businesses or platforms seek to identify algorithms, price recommendations, and website designs. In many such problems, decisions are made in an online manner: for example, when a new visitor searches for a product, an algorithm has to calculate a pricing offer during the time that it takes to load the webpage. The ability to efficiently update a statistical model with new information becomes very important in such a setting.

Unfortunately, outside of the OLS setting, there are many useful statistical models where no clean recursive update such as (16)-(17) is available. One example is the logistic regression model, a special case of (10) where $Y \in \{0,1\}$ and

$$P(Y=1\,|\,\mathbf{x}) = \frac{1}{1+e^{-\mathbf{x}^\top \beta}}, \tag{18}$$

where $\mathbf{x}$ and β are the usual vectors of covariates and regression coefficients, respectively. In personalized pricing, this model is more relevant than OLS because the only response observed from the user is whether or not the product is purchased for the price that was offered. Typically, one fits the model by calculating the maximum likelihood estimator of β, for which there is no closed-form expression; one instead uses numerical procedures such as Newton's method. Finding a *recursive* update for the maximum likelihood estimator seems hopeless.

Researchers have been working on this problem since at least Spiegelhalter and Lauritzen (1990). This paper proposed the update

$$\Sigma_{n+1}^{-1} = \Sigma_n^{-1} + v\mathbf{x}_{n+1}\mathbf{x}_{n+1}^\top, \tag{19}$$

$$\theta_{n+1} = \theta_n - \left(\frac{1}{1+e^{-\mathbf{x}_{n+1}^\top \theta_n}} - y_{n+1}\right)\Sigma_{n+1}\mathbf{x}_{n+1}, \tag{20}$$

with $v > 0$ being a fixed tunable parameter. Comparing (20) with (17), we can see that this approach essentially treats logistic regression as if it were linear regression. Equation (20) is obtained by replacing the residual term $y_{n+1} - \mathbf{x}_{n+1}^\top \theta_n$ of the linear regression with the "residual" term $y_{n+1} - \left(1+e^{-\mathbf{x}_{n+1}^\top \theta_n}\right)^{-1}$ of the logistic regression. Very similar approaches were later proposed by Jaakkola and Jordan (2000) and Qu et al. (2013); they mainly differ in the calculations used for the parameter v, which stands in for the variance of the residuals in linear regression (there being no such quantity in logistic regression).

It may seem surprising that (19)-(20) would ever work, since it appears to impose linear structure on a problem that is inherently very nonlinear, but all of the above-cited papers reported promising practical performance for this technique. More recently, Chen and Ryzhov (2020) showed that $\theta_n \to \beta$ under this method of updating θ_n. The deeper reason for why this approach works ties into the theory of stochastic approximation, which is well-known to the simulation community.

Classical stochastic approximation or SA (Pasupathy and Kim 2011) is an iterative procedure for finding roots β of the system $\nabla_\theta F(\theta) = 0$ by recursively computing $\theta_{n+1} = \theta_n + \alpha_n \nabla_\theta F(\theta_n)$, where α_n is a deterministic stepsize. Bottou (1998) applied SA to online maximum likelihood estimation, where the goal is to solve $\max_\theta \sum_{i=1}^n \log L(y_i; \mathbf{x}_i, \theta)$ for increasingly large n. One simply computes

$$\theta_{n+1} = \theta_n + \alpha_n \nabla_\theta L(y_{n+1}; \mathbf{x}_{n+1}, \theta_n), \tag{21}$$

using the gradient of the marginal likelihood of the most recent observation. As it turns out, the update in (20) can be viewed as a version of (21) for a transformed version of θ_{n+1}. Specifically, we first assume that $\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top = \mathbf{A}$, where $\mathbf{A}$ is a positive definite matrix (this is a well-known sufficient condition for the consistency of the ordinary least squares estimator). One can then interpret (20) as a version of (21) applied to the iterate $\frac{1}{v}\mathbf{A}^{-\frac{1}{2}}\theta_n$, rather than to θ_n directly. Consistency is achieved for any value of the tunable parameter v, although of course practical performance will be sensitive to this value.

This methodology can be very useful when dealing with complicated likelihood functions arising from, e.g., censored data. Suppose that we are trying to estimate a scalar quantity μ; suppose, furthermore, that

there is a sequence of i.i.d. observations $Y_n \sim \mathcal{N}\left(\mu, \sigma^2\right)$, but we only see censored binary signals of the form $B_n = 1_{\{Y_n \geq b_n\}}$ with some sequence $\{b_n\}$ of thresholds. For example, consider a medical application where b_n represents the dose of an experimental drug prescribed to a human patient, with Y_n being that individual patient's maximum tolerance for the drug, and B_n indicating the presence or absence of side effects. In such a setting, (21) has the simple form

$$\theta_{n+1} = \theta_n - \alpha_n \left(B_{n+1} \frac{1}{\sigma} \frac{\phi\left(q_{n+1}\right)}{\Phi\left(q_{n+1}\right)} - \left(1 - B_{n+1}\right) \frac{1}{\sigma} \frac{\phi\left(q_{n+1}\right)}{\Phi\left(q_{n+1}\right)} \right),$$

with $q_{n+1} = \frac{b_{n+1} - \theta_n}{\sigma}$ and ϕ, Φ being the standard normal pdf and cdf.

This method also admits another interpretation using Bayesian statistics. In Bayesian models, the unknown model parameters, such as β in logistic regression, are viewed as random variables whose distribution reflects the beliefs of the decision-maker and evolves over time as new information is acquired. By constructing a probabilistic model of, e.g., β in (18), we are able to assess the likelihood of $P(Y = 1 \,|\, \mathbf{x})$ taking on different values under the same $\mathbf{x}$. Such probabilistic forecasts can be integrated with optimization methods such as Thompson sampling (Russo and Van Roy 2014) or expected improvement (Chen and Ryzhov 2019) to make decisions that account for the uncertainty in our estimate of β.

In OLS regression, Bayesian assumptions lead to virtually no change in the update (16)-(17), but there is no convenient Bayesian update for the other examples in this section. The methods described above can be interpreted using the framework of *approximate Bayesian inference*, where the posterior distribution of belief, given a set of observations, is projected onto a desired distributional family (often normal) in order to easily interface with the aforementioned optimization procedures. In fact, such approximations, often called "variational Bayesian," have been used for many years in practical applications, the most noteworthy example being Dangauthier et al. (2007), which applied them to estimate skill levels of users in competitive online gaming. Other work along these lines includes Das and Magdon-Ismail (2009), Qu et al. (2015), and Zhang and Song (2017). The work by Chen and Ryzhov (2020) was the first to prove the statistical consistency of these various methods by interpreting them under a unified SA framework.

5 CONCLUSION

We have barely scratched the surface with regard to the opportunities for bringing ideas from statistics into simulation research. Another very promising area, which we have not been able to discuss in detail here, is the development of hypothesis tests for complex uncertain objects. For example, the recent work on robust uncertainty quantification (Lam 2016) draws on ideas from distributionally robust optimization, where confidence sets are constructed around probability distributions. Another example is the work by Plumlee and Nelson (2018), which seeks to build confidence sets for the optimal solution of a global optimization problem. We believe that there is much value in bringing a statistical perspective to these and other problems in simulation, and hope that the present tutorial will help to build interest in these topics.

REFERENCES

Ankenman, B., B. L. Nelson, and J. Staum. 2010. "Stochastic Kriging for Simulation Metamodeling". *Operations Research* 58(2):371–382.

Barton, R. R. 2013. "Designing Simulation Experiments". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 342–353: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Basse, G. W., and E. M. Airoldi. 2018. "Model-Assisted Design of Experiments in the Presence of Network-Correlated Outcomes". *Biometrika* 105(4):849–858.

Belloni, A., and V. Chernozhukov. 2013. "Least Squares After Model Selection in High-Dimensional Sparse Models". *Bernoulli* 19(2):521–547.

Bhat, N., V. F. Farias, C. C. Moallemi, and D. Sinha. 2019. "Near Optimal A-B Testing". *Management Science (to appear)*.

Bottou, L. 1998. "Online Learning and Stochastic Approximations". In *On-Line Learning in Neural Networks*, edited by D. Saad, 9–42. Cambridge: Cambridge University Press.

Chen, X., K. Wang, and F. Yang. 2013. "Stochastic Kriging with Qualitative Factors". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 790–801: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Chen, Y., and I. O. Ryzhov. 2019. "Complete Expected Improvement Converges to an Optimal Budget Allocation". *Advances in Applied Probability* 51(1):209–235.

Chen, Y., and I. O. Ryzhov. 2020. "Consistency Analysis of Sequential Learning Under Approximate Bayesian Inference". *Operations Research* 68(1):295–307.

Dangauthier, P., R. Herbrich, T. Minka, and T. Graepel. 2007. "TrueSkill Through Time: Revisiting the History of Chess". In *Advances in Neural Information Processing Systems*, edited by J. C. Platt, D. Koller, Y. Singer, and S. Roweis, Volume 20, 337–344: Red Hook, New York: Curran Associates, Inc.

Das, S., and M. Magdon-Ismail. 2009. "Adapting to a Market Shock: Optimal Sequential Market-Making". In *Advances in Neural Information Processing Systems*, edited by D. Koller, D. Schuurmans, Y. Bengio, and L. Bottou, Volume 21, 361–368. Red Hook, New York: Curran Associates, Inc.

Dudewicz, E. J., and S. R. Dalal. 1975. "Allocation of Observations in Ranking and Selection with Unequal Variances". *Sankhyā: The Indian Journal of Statistics* B37(1):28–78.

Fan, J., Y. Feng, and R. Song. 2011. "Nonparametric Independence Screening in Sparse Ultra-High-Dimensional Additive Models". *Journal of the American Statistical Association* 106(494):544–557.

Fan, J., F. Han, and H. Liu. 2014. "Challenges of Big Data Analysis". *National Science Review* 1(2):293–314.

Fan, J., and J. Lv. 2008. "Sure Independence Screening for Ultrahigh Dimensional Feature Space". *Journal of the Royal Statistical Society* B70(5):849–911.

Fan, J., and R. Song. 2010. "Sure Independence Screening in Generalized Linear Models with NP-Dimensionality". *The Annals of Statistics* 38(6):3567–3604.

Gaiffas, S., and A. Guilloux. 2012. "High-Dimensional Additive Hazards Models and the Lasso". *Electronic Journal of Statistics* 6:522–546.

Glynn, P. W. 1987. "Likelihood Ratio Gradient Estimation: An Overview". In *Proceedings of the 1987 Winter Simulation Conference*, edited by A. Thesen, H. Grant, and D. W. Kelton, 366–375: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Glynn, P. W., and D. L. Iglehart. 1990. "Simulation Output Analysis Using Standardized Time Series". *Mathematics of Operations Research* 15(1):1–16.

Haaland, B., and P. Z. G. Qian. 2010. "An Approach to Constructing Nested Space-Filling Designs for Multi-Fidelity Computer Experiments". *Statistica Sinica* 20(3):1063–1075.

Han, B., I. O. Ryzhov, and B. Defourny. 2016. "Optimal Learning in Linear Regression with Combinatorial Feature Selection". *INFORMS Journal on Computing* 28(4):721–735.

Hastie, T., R. Tibshirani, and J. Friedman. 2009. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.

He, X. 2019. "Sliced Rotated Sphere Packing Designs". *Technometrics* 61(1):66–76.

Hong, L. J., and B. L. Nelson. 2009. "A Brief Introduction To Optimization Via Simulation". In *Proceedings of the 2009 Winter Simulation Conference*, edited by M. Rosetti, R. Hill, B. Johansson, A. Dunkin, and R. Ingalls, 75–85: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Huang, J., J. L. Horowitz, and F. Wei. 2010. "Variable Selection in Nonparametric Additive Models". *The Annals of Statistics* 38(4):2282–2313.

Ibrahim, J. G., H. Zhu, R. I. Garcia, and R. Guo. 2011. "Fixed and Random Effects Selection in Mixed Effects Models". *Biometrics* 67(2):495–503.

Jaakkola, T. S., and M. I. Jordan. 2000. "Bayesian Parameter Estimation via Variational Methods". *Statistics and Computing* 10(1):25–37.

Jiang, G., L. J. Hong, and B. L. Nelson. 2020. "Online Risk Monitoring Using Offline Simulation". *INFORMS Journal on Computing* 32(2):356–375.

Johnson, M. E., L. M. Moore, and D. Ylvisaker. 1990. "Minimax and Maximin Distance Designs". *Journal of Statistical Planning and Inference* 26(2):131–148.

Joseph, V. R., E. Gul, and S. Ba. 2015. "Maximum Projection Designs for Computer Experiments". *Biometrika* 102(2):371–380.

Joseph, V. R., E. Gul, and S. Ba. 2020. "Designing Computer Experiments with Multiple Types of Factors: The MaxPro Approach". *Journal of Quality Technology (to appear)*.

Joseph, V. R., and Y. Hung. 2008. "Orthogonal-Maximin Latin Hypercube Designs". *Statistica Sinica* 18(1):171–186.

Kallus, N. 2018. "Optimal A Priori Balance in the Design of Controlled Experiments". *Journal of the Royal Statistical Society* B80(1):85–112.

Kelton, W. D., and R. R. Barton. 2003. "Experimental Design for Simulation". In *Proceedings of the 2003 Winter Simulation Conference*, edited by S. Chick, P. J. Sánchez, D. Ferrin, and D. J. Morrice, 59–65: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Kleijnen, J. P. C. 2009. "Kriging Metamodeling in Simulation: A Review". *European Journal of Operational Research* 192(3):707–716.

Kleiner, A., A. Talwalkar, P. Sarkar, and M. I. Jordan. 2014. "A Scalable Bootstrap for Massive Data". *Journal of the Royal Statistical Society* B76(4):795–816.

Lai, T. L., and C. Z. Wei. 1982. "Least Squares Estimates in Stochastic Regression Models with Applications to Identification and Control of Dynamic Systems". *The Annals of Statistics* 10(1):154–166.

Lam, H. 2016. "Advanced Tutorial: Input Uncertainty and Robust Analysis in Stochastic Simulation". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 178–192.

Law, A. M. 2017. "A Tutorial on Design of Experiments for Simulation Modeling". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 550–564.

Li, J., S. Netessine, and S. Koulayev. 2018. "Price to Compete...with Many: How to Identify Price Competition in High-Dimensional Space". *Management Science* 64(9):4118–4136.

Li, Q., R. Xi, and N. Lin. 2010. "Bayesian Regularized Quantile Regression". *Bayesian Analysis* 5(3):533–556.

Li, R., W. Zhong, and L. Zhu. 2012. "Feature Screening via Distance Correlation Learning". *Journal of the American Statistical Association* 107(499):1129–1139.

Lin, Y., and B. L. Nelson. 2016. "Simulation Analytics for Virtual Statistics via k Nearest Neighbors". In *Proceedings of the 2016 Winter Simulation Conference*, edited by T. M. K. Roeder, P. I. Frazier, R. Szechtman, E. Zhou, T. Huschka, and S. E. Chick, 448–459: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

McKay, M. D., R. J. Beckman, and W. J. Conover. 1979. "Comparison of Three Methods for Selecting Values of Input Variables in the Analysis of Output From a Computer Code". *Technometrics* 21(2):239–245.

Meier, L., S. Van de Geer, and P. Bühlmann. 2008. "The Group Lasso for Logistic Regression". *Journal of the Royal Statistical Society* B70(1):53–71.

Morgan, K. L., and D. B. Rubin. 2012. "Rerandomization to Improve Covariate Balance in Experiments". *The Annals of Statistics* 40(2):1263–1282.

Morris, M. D., and T. J. Mitchell. 1995. "Exploratory Designs for Computational Experiments". *Journal of Statistical Planning and Inference* 43(3):381–402.

Nelson, B. L. 2016. "'Some Tactical Problems in Digital Simulation' for the Next 10 Years". *Journal of Simulation* 10(1):2–11.

Owen, A. B. 1994. "Controlling Correlations in Latin Hypercube Samples". *Journal of the American Statistical Association* 89(428):1517–1522.

Pasupathy, R., and S. Kim. 2011. "The Stochastic Root-Finding Problem: Overview, Solutions, and Open Questions". *ACM Transactions on Modeling and Computer Simulation* 21(3):19:1–19:23.

Plumlee, M., and B. L. Nelson. 2018. "Plausible Optima". In *Proceedings of the 2018 Winter Simulation Conference*, edited by M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, 1981–1992.

Pokhilko, V., Q. Zhang, L. Kang, and D. P. Mays. 2019. "D-Optimal Design for Network A/B Testing". *Journal of Statistical Theory and Practice* 13(4). Article 61.

Qian, P. Z. 2009. "Nested Latin Hypercube Designs". *Biometrika* 96(4):957–970.

Qian, P. Z. 2012. "Sliced Latin Hypercube Designs". *Journal of the American Statistical Association* 107(497):393–399.

Qu, H., I. O. Ryzhov, and M. C. Fu. 2013. "Learning Logistic Demand Curves in Business-to-Business Pricing". In *Proceedings of the 2013 Winter Simulation Conference*, edited by R. Pasupathy, S.-H. Kim, A. Tolk, R. Hill, and M. E. Kuhl, 29–40: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Qu, H., I. O. Ryzhov, M. C. Fu, and Z. Ding. 2015. "Sequential Selection with Unknown Correlation Structures". *Operations Research* 63(4):931–948.

Robbins, H., and S. Monro. 1951. "A Stochastic Approximation Method". *The Annals of Mathematical Statistics* 22(3):400–407.

Russo, D., and B. Van Roy. 2014. "Learning to Optimize via Posterior Sampling". *Mathematics of Operations Research* 39(4):1221–1243.

Sacks, J., W. J. Welch, T. J. Mitchell, and H. P. Wynn. 1989. "Design and Analysis of Computer Experiments". *Statistical Science* 4(4):409–423.

Shahriari, B., K. Swersky, Z. Wang, R. P. Adams, and N. De Freitas. 2015. "Taking the Human Out of the Loop: A Review of Bayesian Optimization". *Proceedings of the IEEE* 104(1):148–175.

Shen, H., L. J. Hong, and X. Zhang. 2017. "Ranking and Selection with Covariates". In *Proceedings of the 2017 Winter Simulation Conference*, edited by W. K. V. Chan, A. D'Ambrogio, G. Zacharewicz, N. Mustafee, G. Wainer, and E. Page, 2137–2148: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Shewry, M. C., and H. P. Wynn. 1987. "Maximum Entropy Sampling". *Journal of Applied Statistics* 14(2):165–170.

Shi, J., W. Yin, S. Osher, and P. Sajda. 2010. "A Fast Hybrid Algorithm for Large-scale l1-Regularized Logistic Regression". *Journal of Machine Learning Research* 11:713–741.

Sinha, B. K. 1970. "On the Optimality of Some Designs". *Calcutta Statistical Association Bulletin* 19(1):1–22.

Song, E., B. L. Nelson, and C. D. Pegden. 2014. "Advanced Tutorial: Input Uncertainty Quantification". In *Proceedings of the 2014 Winter Simulation Conference*, edited by A. Tolk, S. Y. Diallo, I. O. Ryzhov, L. Yilmaz, S. Buckley, and J. A. Miller, 162–176: Piscataway, New Jersey: Institute of Electrical and Electronics Engineers, Inc.

Spiegelhalter, D. J., and S. L. Lauritzen. 1990. "Sequential Updating of Conditional Probabilities on Directed Graphical Structures". *Networks* 20(5):579–605.

Stein, M. 1987. "Large Sample Properties of Simulations Using Latin Hypercube Sampling". *Technometrics* 29(2):143–151.

Székely, G. J., and M. L. Rizzo. 2009. "Brownian Distance Covariance". *The Annals of Applied Statistics* 3(4):1233–1303.

Székely, G. J., M. L. Rizzo, and N. K. Bakirov. 2007. "Measuring and Testing Dependence by Correlation of Distances". *The Annals of Statistics* 35(6):2769–2794.

Tang, B. 1993. "Orthogonal Array-Based Latin Hypercubes". *Journal of the American Statistical Association* 88(424):1392–1397.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society* B58(1):267–288.

Titterington, D. M. 1975. "Optimal Design: Some Geometrical Aspects of D-optimality". *Biometrika* 62(2):313–320.

Van de Geer, S. A. 2008. "High-Dimensional Generalized Linear Models and the Lasso". *The Annals of Statistics* 36(2):614–645.

Wu, C. J. 2015. "Post-Fisherian Experimentation: From Physical to Virtual". *Journal of the American Statistical Association* 110(510):612–620.

Xu, Y., N. Chen, A. Fernandez, O. Sinno, and A. Bhasin. 2015. "From Infrastructure to Culture: A/B Testing Challenges in Large Scale Social Networks". In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2227–2236: New York, New York: Association for Computing Machinery.

Yang, Y., and H. Zou. 2015. "A Fast Unified Algorithm for Solving Group-Lasso Penalize Learning Problems". *Statistics and Computing* 25(6):1129–1141.

Zhang, Q., A. Khademi, and Y. Song. 2020. "Robust Optimal Design of Two-Armed Trials with Side Information". *arXiv preprint arXiv:2002.01095*.

Zhang, Q., and P. Z. Qian. 2013. "Designs for Crossvalidating Approximation Models". *Biometrika* 100(4):997–1004.

Zhang, Q., and Y. Song. 2017. "Moment-Matching-Based Conjugacy Approximation for Bayesian Ranking and Selection". *ACM Transactions on Modeling and Computer Simulation* 27(4):26:1–26:23.

Zhang, Q., B. Wang, and W. Xie. 2020. "A Pooled Quantile Estimator for Parallel Simulations". *Journal of Simulation (to appear)*.

Zhao, P., and B. Yu. 2006. "On Model Selection Consistency of Lasso". *Journal of Machine Learning Research* 7:2541–2563.

Zhao, S. D., and Y. Li. 2012. "Principled Sure Independence Screening for Cox Models with Ultra-High-Dimensional Covariates". *Journal of Multivariate Analysis* 105(1):397–411.

Zhou, W., J.-F. Yang, and M.-Q. Liu. 2020. "Optimal Maximin L2-Distance Latin Hypercube Designs". *Journal of Statistical Planning and Inference* 207:113–122.

Zou, H., T. Hastie, and R. Tibshirani. 2007. "On the 'Degrees of Freedom' of the Lasso". *The Annals of Statistics* 35(5):2173–2192.

AUTHOR BIOGRAPHIES

ILYA O. RYZHOV is an Associate Professor of Operations Management in the Decision, Operations and Information Technologies department of the Robert H. Smith School of Business, with a joint appointment in the Institute for Systems Research, all at the University of Maryland. His research primarily focuses on simulation optimization and statistical learning, with applications in business analytics, revenue management, and nonprofit/humanitarian operations. He is a coauthor of the book *Optimal Learning* (Wiley, 2012). His work was recognized in WSC's Best Theoretical Paper Award competition on three separate occasions (winner in 2012, finalist in 2009 and 2016), and he received I-SIM's Outstanding Publication Award in 2017. His email address is iryzhov@rhsmith.umd.edu.

QIONG ZHANG is an Assistant Professor of Statistics at Clemson University. She holds a Ph.D. in statistics from the University of Wisconsin-Madison. Her research interests include computer experiments, uncertainty quantification and spatial and spatial-temporal modeling. She is a member of ASA and INFORMS. Her email address is qiongz@clemson.edu.

YE CHEN is an Assistant Professor of Statistical Sciences and Operations Research at Virginia Commonwealth University. He received a Ph.D. in Statistics from the University of Maryland in 2018. His research interests include applied probability, statistical learning, and stochastic optimization. He was a finalist in the Best Theoretical Paper Award competition at the 2016 Winter Simulation Conference. His email address is ychen24@vcu.edu.