Proceedings of the 2018 Winter Simulation Conference
M. Rabe, A. A. Juan, N. Mustafee, A. Skoogh, S. Jain, and B. Johansson, eds.

DATA-DRIVEN RANKING AND SELECTION: HIGH-DIMENSIONAL COVARIATES AND GENERAL DEPENDENCE

Xiaocheng Li

Department of Management Science and Engineering
Stanford University
Stanford, CA 94305, USA

Xiaowei Zhang

Department of Industrial Engineering
and Decision Analytics
Hong Kong University of Science and Technology
Clear Water Bay, HONG KONG

Zeyu Zheng

Department of Industrial Engineering and Operations Research
University of California, Berkeley
Berkeley, CA 94720, USA

ABSTRACT

This paper considers the problem of ranking and selection with covariates and aims to identify a decision rule that stipulates the best alternative as a function of the observable covariates. We propose a general data-driven framework to accommodate (i) high-dimensional covariates and (ii) general (nonlinear) dependence between the mean performance of an alternative and the covariates. For both scenarios, we design new selection procedures and provide certain statistical guarantees, by leveraging the data-intensive environment and various statistical learning tools. The performances of our procedures are exhibited through simulation experiments.

1 INTRODUCTION

Ranking and selection (R&S) is concerned with making the best selection among many alternatives, whose unknown performances can be learned via sampling. Examples include selecting medicine and treatment regimes in healthcare systems, selecting online advertisements targeting Internet users, and selecting reaction/operational mechanisms for self-driving vehicles. In the past decades, the problem of R&S has been a focal point in simulation literature; see Kim and Nelson (2006). Recently, Shen et al. (2017) introduces a new R&S framework, called *ranking and selection with covariates* (R&S-C), in which the mean performance of each alternative depends on some observable random covariates. The goal of R&S-C is to identify a decision rule that stipulates the best alternative as a function of the covariates, as opposed to a single alternative that is believed to be the best universally. In this way, the covariates information for each alternative can be exploited to deliver better personalized decisions.

In this paper, we adopt such data-driven R&S framework but consider application contexts where the observable covariates are high-dimensional. This is motivated by the exponential growth of data storage in healthcare and e-commerce, in which high-dimensional covariates are becoming increasingly available. The existing selection procedures developed in Shen et al. (2017) which are based on Ordinary Least Square (OLS) may become inefficient and even inapplicable in such contexts. We propose a new selection procedure for R&S with **h**igh-dimensional **c**ovariates (R&S-HC), that effectively incorporate machine learning tools such as regularization and variable selection. Certain statistical guarantee is provided under moderate assumptions. Moreover, notice that for simplicity, the R&S-C problem in Shen et al. (2017) assumes a *linear* dependence between the mean performance of an alternative and the covariates. In the second half

of the present paper, we extend our discussion to allow a general (nonlinear) dependence between them, i.e., R&S with **g**eneral covariates **d**ependence (R&S-GD), and develop a selection procedure accordingly.

Our selection procedures for R&S-HC and R&S-GD utilize the ample data availabilities and rapidly growing computation power. These developments reflect the perspective that simulation ought to play an increasingly important role in decision making as a tool for system control; see Nelson (2016) for an extensive discussion on the subject.

The contribution of our work is three-fold. First, we demonstrate that failing to account for covariates in R&S when do they exist can lead to incorrect selection with a large probability. The problem deteriorates when the mean responses of the alternatives become closer or the dependence between response and covariates becomes stronger.

Second, we develop selection procedures to deal with high-dimensional covariates and general covariate dependence. Specifically, we view the collection of simulation samples as a training procedure to estimate the (parameters of) the dependence function. The results from machine learning theory are helpful in characterizing how many samples are needed in order to control the estimation error below certain threshold. Moreover, the estimation error can be easily linked to the probability of correction selection in R&S. Thus, by reversing the arguments, we obtain the selection procedures for R&S-HC and R&S-GD with statistical validity.

Third, we conduct experiments to show the advantage of integrating covariates in R&S and the efficiency of our procedures. On one hand, we show that a right choice of the selection procedure can greatly reduce the number of required sample; on the other hand, if one ignore the covariates or choose an over-simplified model to characterize the dependence, it may result in an unsuccessful selection procedure regardless of the number of simulation samples.

We close the introduction by briefly reviewing related work. The R&S problem was introduced in 1950's for statistical selection problems arising in medical treatment selection; see Robbins (1952) and Bechhofer (1954). We refer to the survey papers Kim and Nelson (2006) and Chen et al. (2015) for the historical development and current status of the R&S research. The selection procedures in R&S literature follow either of the two approaches: the frequentist approach (see, e.g., Rinott 1978; Kim and Nelson 2001; Hong and Nelson 2007; Fan et al. 2016) or the Bayesian approach (see, e.g., Chen 1996; Chick and Inoue 2001; Frazier et al. 2008; Chick and Frazier 2012). Our paper takes a frequentist perspective and provides stage-wise procedures.

Recently, there has been a trend of designing R&S procedures to better embrace the blessing of big data from large-scale problems. In Luo et al. (2015), the authors consider the large-scale R&S where there is a large number of alternatives. In Hunter and Nelson (2017) and Kamiński and Szufel (2018), parallel procedures are designed for R&S simulation. These two considerations reflect two typical changes brought by big data: the scale of the problem and faster computing devices. Comparatively, our work, extending Shen et al. (2017), concerns another perspective of big data: the availability of side/contextual information and the machine learning/statistical models to efficiently process this information.

In machine learning community, people have successfully utilized the side information/covariates to improve prediction accuracy in tasks such as image classification (Chen et al. 2012), text mining (Aggarwal et al. 2014), and recommendation system (Xu et al. 2013). This usually results in a refinement of the original problem that is absent of the covariates. Thanks to the recent advances of statistical learning theory, researchers are equipped with power tools for theoretical analysis when applying machine learning algorithms for covariates. The research on R&S-C is also related with the contextual bandit problem (Lu et al. 2010). Our R&S with high-dimensional covariates is partly motivated by the paper Bastani and Bayati (2015).

The organization of the paper is as follow. In Section 2, we provide the mathematical formulation and state the objective of R&S-C. We point out that the negligence of the covariates may lead to incorrect selection with high likelihood. In Section 3 and Section 4, we present the selection procedures for R&S-HC

and R&S-GD, respectively. We demonstrate the efficiency and effectiveness of our procedures in Section 5 through simulation experiments and conclude in Section 6.

2 MODEL SETUP

We consider a collection of K (finite) distinctive alternatives and aim to identify the best alternative through simulations. Different from the conventional R&S setting, we let the performance of each alternatives depends on $X = (X_1, ..., X_d)^\intercal$, a vector of random covariates with support $\Theta \subseteq \mathbb{R}^d$. Concretely, the performance of the k-th alternative is associated with the covariate through a function $f_k(\cdot)$, i.e.

$$Y_k(X) = f_k(X) + \varepsilon_k(X), \quad 1 \le k \le K, \tag{1}$$

where $\varepsilon_k(X) \sim N(0, \sigma_k^2)$ is the sampling error. Here, we do not make any additional restriction on the function $f_k(\cdot)$, say, it can be even a non-continuous (decision tree) function of X. The subscript k emphasizes that the dependence between response and covariate can be different across different alternatives.

The objective is to identify the best alternative with the knowledge of the covariates,

$$k^*(x) := \underset{1\le k\le K}{\operatorname{argmax}} \{E[Y_k(X)|X = x]\} = \underset{1\le k\le K}{\operatorname{argmax}} \{f_k(x)\}.$$

In such a way, our selection decision turns out to be a function of the covariates x. This further explains the meaning of data-driven R&S, i.e., let the data x drive the decision.

To quantify the result of an R&S scheme, we adopt the Indifference Zone formulation (Bechhofer 1954). Specifically, we define the correct selection (CS) event when one of the correct alternatives is selected:

$$\mathrm{CS}(x) := \left\{ f_{k^*(x)}(X) - f_{\widehat{k^*}(x)}(X) < \delta \mid X = x \right\},$$

where $k^*(x)$ and $\widehat{k^*}(x)$ are the best alternative and the selected alternative respectively. To extend the CS event for the setting of covariates, we adopt the same criteria as in the papers Ni, Ciocan, Henderson, and Hunter (2017), Shen, Hong, and Zhang (2017) and define the conditional Probability of Correct Selection (PCS) as

$$\mathrm{PCS}(x) := P\left(f_{k^*(x)}(X) - f_{\widehat{k^*}(x)}(X) < \delta \mid X = x \right).$$

And based on the conditional PCS, we define two forms of unconditional PCS:

$$\mathrm{PCS}_{\mathrm{E}} := E[\mathrm{PCS}(X)]$$

where the expectation is taken with respect to the distribution of X, and

$$\mathrm{PCS}_{\min} := \min_{x\in\Theta} \mathrm{PCS}(x).$$

Remark 1. $\mathrm{PCS}_{\min}$ provides a lower bound for $\mathrm{PCS}_{\mathrm{E}}$; in other words, $\mathrm{PCS}_{\min}$ is a more demanding criterion when evaluating R&S procedures.

2.1 The Advantage of Covariates

Consider the dependence of performance response and the covariates specified as in (1), if we ignore the covariates and adopt the conventional R&S procedures, the objective becomes to choose

$$k^* := \underset{1\le k\le K}{\operatorname{argmax}} \{E[Y_k(X)]\} = \arg \max_{1\le k\le K} \{E[f_k(X)]\}.$$

For the simplicity of discussion and without loss of generality, we assume $K = 2$ and $\mu_1 = E[f_1(X)] > \mu_2 = E[f_2(X)]$. Then a successful R&S procedure should end up selecting the first alternative, regardless of

the value of X. However, it can happen that for certain value of $X = x$, we have $f_1(x) < f_2(x) - \delta$, which means we select the significantly inferior alternative. The probability of such wrong selection is

$$P(f_1(X) < f_2(X) - \delta).$$

To get a better sense of the quantity, we consider $X \overset{\mathscr{D}}{=} N(0,\Sigma)$ and the functions

$$f_i(X) = \mu_i + \theta_i^\intercal X \overset{\mathscr{D}}{=} N(\mu_i, \theta_i^\intercal \Sigma \theta_i), \quad i = 1,2.$$

Then the probability of incorrect selection is

$$P(f_1(X) < f_2(X) - \delta) = P\left(Z < \frac{\mu_2 - \mu_1 - \delta}{\theta_1^\intercal \Sigma \theta_1 + \theta_2^\intercal \Sigma \theta_2}\right),$$

where Z is the standard normal random variable. The probability becomes larger when the gap between μ_1 and μ_2 is small and when θ_1 and θ_2 have large norms. This indicates that we will pay a high price ignoring the covariates when either the mean performances of the alternatives are similar or there is a strong dependence between the response and the covariates. To make things even worse, the probability will further scale up when there are $K > 2$ alternatives. The same analysis can be done for the case that we ignore part of the covariates or that we consider an over-simplified class of f_i's (e.g. linear instead of non-linear). In other words, the argument here warns us that the conventional objective of R&S is not the right goal to pursue with the existence of covariates. Even if we have guarantee on the probability of correct selection of the best mean performance (namely μ_1 in the above example), the procedure can lead to a high probability of selecting the inferior ones due to the negligence of the covariates.

On the other hand, by taking the covariates into account, it avoids the above mistake. But this requires a more complicated design in the R&S procedure. Specifically, we need to propose an R&S procedure that utilizes the simulation samples efficiently to entangle the dependence between response and covariates. As a result, the structure of dependences and simulation data should be taken into account when designing simulation schemes.

3 R&S WITH HIGH-DIMENSIONAL COVARIATES

In statistics and machine learning, the model of Lasso (least absolute shrinkage and selection operator) enables one to deal with the high-dimensional data efficiently by performing regularization and variable selection at the same time; see Tibshirani (1996). Here we adopt Lasso for designing R&S procedure in the data-rich setting. In the first place, we make the following assumptions over the general framework in Section 2.

Assumption 1. *(i) The functions f_k's in (1) are linear of covariates $X \in \mathbb{R}^d$ and have sparse coefficient vectors β_k's. Specifically,*

$$f_k(X) = X^\intercal \beta_k,$$

and the number of non-zeros (cardinality) in β_k's are bounded by s_0, i.e. $\|\beta_k\|_0 \le s_0$ for all $k = 1,...,K$. (Note that our results in the paper hold for any given s_0, the choice of which depends on the application context.)

(ii) For each alternative $k = 1,...,K$ and simulation trial $l = 1,2,...$, we have

$$Y_{kl}(x) = f_k(x) + \varepsilon_{kl}$$

where the sampling error $\varepsilon_{kl} \sim N(0,\sigma_k^2)$ and ε_{kl} is independent of $\varepsilon_{k'l'}$ for any $(k,l) \neq (k',l')$. In other words, the simulation errors are independent across different alternatives and different simulation samples.

(iii) We assume the support Θ of X is bounded by an L_1 ball with radius B.

The key to an efficient R&S procedure in the high-dimensional setting is the application of Lasso in estimating the coefficients β_k's. Therefore, we first give a brief overview of the Lasso problem. Generally, the problem of Lasso regression can be posed as the following:

$$Y = \mathscr{X}\beta^0 + \varepsilon \tag{2}$$

$Y = (Y_1, ..., Y_n) \in \mathbb{R}^n$ is the vector of response, $\mathscr{X} \in \mathbb{R}^{n\times d}$ is the design matrix, and ε is a vector of $N(0,\sigma^2)$-distributed measurement errors. In the OLS regression setting, we estimate β^0 with

$$\widehat{\beta}_{\text{OLS}} := (\mathscr{X}^\intercal \mathscr{X})^{-1}\mathscr{X}^\intercal Y.$$

The blessing of high-dimensional covariates setting is a better assessment for the performance of the alternatives due to the better availability of covariates, but this can be offset in by the induced simulation costs, if the R&S scheme is not carefully designed. Conversely, when we are parsimonious in simulation budget, the number of samples n can be much smaller than the dimension d. This can result in a singular matrix of $\mathscr{X}^\intercal\mathscr{X}$, which makes the OLS estimates fail. However, Lasso provides us an elegant way to handle this situation and thus results in an R&S procedure for which the number of required simulations does not scale up with the number of covariates.

Different from OLS setting, Lasso adopts the penalized loss function

$$\widehat{\beta} = \underset{\beta}{\text{argmin}} \quad \frac{\|Y - \mathscr{X}\beta\|_2^2}{n} + \lambda\|\beta\|_1,$$

where $\|\cdot\|_2$ denotes the L_2 distance. A huge literature in statistics have demonstrated its successfulness in handling large dimensional covariates. For our R&S application, intuitively speaking, a bound for PCS entails a characterization of the Lasso estimation errors, namely, how far the $\widehat{\beta}_k$ deviates from β_k. To proceed, we follow the results in Bühlmann and van de Geer (2011) and start with a discussion of the compatibility condition.

Definition 1 (Compatibility condition) Let the (scaled) Gram matrix

$$\widehat{\Sigma} := \frac{\mathscr{X}^\intercal\mathscr{X}}{n}.$$

We say that the compatibility condition is met for the set $S_0 \subseteq S = \{1,2,...,d\}$, if for some $\phi_0 > 0$, and for all β satisfying $\|\beta_{S_0^c}\|_1 \le 3\|\beta_{S_0}\|_1$, it holds that

$$\|\beta_{S_0^c}\|_1^2 \le \left(\beta^\intercal\widehat{\Sigma}\beta\right) s_0/\phi_0^2,$$

where s_0 is given in Assumption 1. And we call ϕ_0 as the compatibility constant.

In fact, the compatibility constant ϕ_0 is an intrinsic value of matrix $\widehat{\Sigma}$. Both for fixed/random design matrix $\mathscr{X}$, one can compute restricted isometry constant (Candes et al. 2006) or restricted eigenvalue (Bickel et al. 2009) as candidates for ϕ_0. As we can see in the following proposition, ϕ_0 plays an important role in characterizing the Lasso estimation error.

Proposition 1. *Suppose that the $\widehat{\Sigma}$ is normalized such that $\widehat{\Sigma}_{jj} = 1$ for $j = 1,...,d$. Let the regularization parameter be*

$$\lambda := 4\widehat{\sigma}\sqrt{\frac{t^2 + 2\log d}{n}}$$

for a given $t > 0$, where $\widehat{\sigma}^2$ is an estimator for the noise variance σ^2. Then with probability at least $1-\alpha$, where

$$\alpha := 2\exp(-t^2/2) + P(\widehat{\sigma} \le \sigma),$$

we have that

$$\|\hat{\beta} - \beta^0\|_1 \leq 4\lambda^2 s_0/\phi_0^2.$$

Notice that in the bound, both the definition of α and λ require an estimate of the variance of error ε. A desired property of the estimate $\widehat{\sigma}$ is that it is greater than true σ with high probability. To achieve this, we use sample variance of Y as the estimator of σ as proposed in Bühlmann and van de Geer (2011). Concretely, with notations in (2), we propose the estimator

$$\widehat{\sigma}_n^2 := \frac{c}{n-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$$

where $\bar{Y} = \frac{1}{n}\sum_{i=1}^{n} Y_i$ and c as a constant to be determined. It is easy to verify that $E\widehat{\sigma}_n^2 = c\sigma_Y^2 \geq \sigma^2$ for $c \geq 1$ and

$$\frac{(n-1)\widehat{\sigma}_n^2}{c\sigma_Y^2} \overset{\mathscr{D}}{=} \chi^2(n-1).$$

Usually, people choose $c = 1$ for estimation of σ^2, but here since the goal is different - looking for an upper bound on σ. So we choose it large so as to control $P(\hat{\sigma} \leq \sigma)$.

The Lasso loss function and Proposition 1 provide a full picture of doing linear regression with sparse and high-dimensional covariates. In R&S setting, what we need to do is to conduct the same procedure for each of the alternatives. The conventional choice of the design matrix $\mathscr{X}$ is the random Gaussian matrix. To adapt it for the R&S setting, an all-one column is appended in the end. With the Proposition 1 in mind, we design the following R&S-HC procedure:

Step 0. Setup: Let $t = \sqrt{\frac{1}{2}\log\frac{6K}{\alpha}}$ and n_0 be the minimum number to guarantee the Gram matrix $\widehat{\Sigma}$ associated with the random Gaussian design matrix has a compatibility constant ϕ_0 with probability $1 - \frac{\alpha}{6K}$.

Step 1. First-stage Sampling: Generate the random design matrix $\mathscr{X}_0 \in \mathbb{R}^{N\times d}$ for some large enough N. Append all-one column to $\mathscr{X}_0$ and obtain $\mathscr{X} \in \mathbb{R}^{N\times(d+1)}$. Simulate the responses for each of the first n_1 row of the design matrix and denote them by Y_{kl}, $k = 1,...,K$, $l = 1,...,n_0$. Construct our estimation of $\widehat{\sigma}_k^2$, such that $P(\widehat{\sigma}_k^2 > \sigma_k^2) \geq 1 - \frac{\alpha}{6K}$, $k = 1,...,K$.

Step 2. Second-stage Sampling: Compute $n_k' = \max\{n_0, 128Bs_0\widehat{\sigma}_k^2(t^2 + 2\log d)/\delta\}$ for $k = 1,...,K$. If $n_1 < n_k'$, take $n_k' - n_1$ more rows from the design matrix $\mathscr{X}$ and simulate the response. Denote the design matrix for k'th alternative as $\mathscr{X}_k$ and the response as Y_k, estimate

$$\widehat{\beta}_k = \underset{\beta}{\operatorname{argmin}} \quad \frac{\|Y_k - \mathscr{X}_k\beta\|_2^2}{n} + \lambda_k\|\beta\|_1,$$

with λ_k specified as Proposition 1.

Step 3. Selection: for each covariate x, return

$$\widehat{k}^*(x) = \underset{1\leq k\leq K}{\operatorname{argmax}}\{x^\intercal\widehat{\beta}_k\}$$

as the selected one.

Theorem 1 Under Assumption 1, the R&S-HC procedure achieves $\mathrm{PCS}_{\min} \geq 1 - \alpha$.

The proof for the above statistical validity follows from a direct application of Proposition 1. The number of alternatives K appears several times in the R&S-GD procedure and this is due to our usage of Boole's inequality ("union bound") when bounding the PCS. The usage of union bound is inevitable here and this is essentially caused by the lack of distributional results on $\widehat{\beta}_k - \beta_k$ in Lasso.

4 R&S WITH GENERAL DEPENDENCE

By considering a general dependence between response and covariates, we are able to take advantage of the machine learning algorithms to model the response. Like Section 3, we need a learning scheme for estimating the dependence function f_k's. Imagine through simulations, we obtain samples (X_{kl}, Y_{kl}) $l = 1, ..., n$ for the k'th alternative. We propose to estimate f_k's through minimizing the *empirical risk*:

$$\widehat{f}_k = \operatorname*{argmin}_{f \in \mathscr{F}} \sum_{l=1}^{n} L(Y_{kl}, f(X_{kl})) \tag{3}$$

where $L(\cdot,\cdot)$ is the loss function and $\mathscr{F}$ is a class of candidate functions. For example, $\mathscr{F}$ can be linear functions, decision trees, neural networks and etc. And in this paper, we consider the L_2 loss, i.e. $L(x,y) = (x-y)^2$. A successful R&S procedure here entails a good estimation of f_k. Also, we need to know how well our estimation $\widehat{f}_k$ approximates the true dependence f_k.

Assumption 2. *(i) The loss function $L(\cdot,\cdot)$ takes value in $[B_l, B_u]$ on the support of response-covariate pair (Y,X).*

(ii) The covariates are drawn from a distribution p^*.

(iii) For each alternative $k = 1, ..., K$ and simulation trial $l = 1, 2, ...$, we have

$$Y_{kl}(X) = f_k(X) + \varepsilon_{kl}$$

where $X \sim p^*$ and the sampling error $\varepsilon_{kl} \sim N(0, \sigma_k^2)$; ε_{kl} is independent of $\varepsilon_{k'l'}$ for any $(k,l) \neq (k',l')$; and ε_{kl} is independent of X.

Remark 1. *(iii) is a classic assumption in statistical learning theory. It enforces the training and testing data come from the same distribution, which is the backbone to guarantee the generalization error. As a consequence, we target PCS_E here rather than $PCS_{\min}$ as in last section.*

First we introduce the notion of *Rademacher complexity* as a characterization for the complexity of a generic function class $\mathscr{C}$. Then we present the relation between Rademacher complexity and the estimation error. People usually use covering numbers or VC dimension as its upper bound. We refer the interested readers to Vapnik (1998) for more details. Also, we note that the computation of Rademachar complexity (as an upper bound) can be challenging for some base function class $\mathscr{F}$, which may result in a loose upper bound. In fact, experiments show that the sample required in practice is much smaller than the upper bound specified by the theory.

Definition 2 (Rademacher Complexity) Define the Rademacher complexity of $\mathscr{C}$ as

$$R_n(\mathscr{C}) := E\left[\sup_{f \in \mathscr{C}} \frac{1}{n} \sum_{i=1}^{n} U_i f(Z_i)\right]$$

where $Z_1, ..., Z_n$ are drawn i.i.d from p^* and $U_1, ..., U_n$ are i.i.d. uniform distribution over $\{-1, 1\}$.

Proposition 2. *Define $\mathscr{L} = \{(x,y) \to (y - f(x))^2 : f \in \mathscr{F}\}$ as the loss class, the composition of L_2 loss function and dependence function $f \in \mathscr{F}$, we have*

$$E_{p^*}[(\widehat{f}_n(X) - f(X))^2] \leq 4R_n(\mathscr{L}) + \sqrt{\frac{2\log(2/\eta)}{n}} \cdot (B_u - B_l)$$

with probability $1 - \eta$. Here the expectation is taken with respect to X and $\widehat{f}_n$ is the empirical risk minimizer for n samples as in (3).

We propose the following R&S procedure for general covariates dependence function, namely R&S-GD:

Step 0. Setup: Choose $\eta = \frac{\alpha\delta^2}{4K}$ and $n_0 = \frac{8K^2\log(2/\eta)(B_u - B_l)^2}{\delta^2}$. Specify the dependence function class $\mathscr{F}$ and thus the loss class $\mathscr{L}$. Choose n_0' such that $R_n(\mathscr{L}) < \frac{\alpha\delta^2}{8K}$ for all $n > n_0'$. Let $N = \max\{n_0, n_0'\}$.

Step 1. Sampling: Generate N samples X_{kl} from p^* for each alternative $k = 1, ..., K$, $l = 1, ..., N$, and record the responses Y_{kl} accordingly. Estimate $\widehat{f}_k$ as the empirical risk minimizer:

$$\widehat{f}_k = \underset{f \in \mathscr{F}}{\operatorname{argmin}} \sum_{l=1}^{N} (Y_{kl} - f(X_{kl}))^2$$

for $k = 1, ..., K$.

Step 2. Selection: for each covariate, return

$$\widehat{k}^*(x) = \underset{1 \le k \le K}{\operatorname{argmax}} \{\widehat{f}_k(x)\}$$

as the selected one.

Theorem 2 Under Assumption 2, the R&S-GD procedure achieves $\mathrm{PCS}_{\mathrm{E}} \ge 1 - \alpha$.

Remark 2. *The R&S-GD involves only one stage of sampling, compared with two stages sampling in Section 3. This is caused by the nature of the statistical learning procedure in (3). By minimizing (3), we estimate the dependence functions f_k's. Notice that the error bound in Proposition 2 does not change with respect to the variances σ_k^2's. The Rademacher complexity provides a "sufficient" number of samples we need to estimate the true dependence; this number can be conservative and greater than "necessary", but it is not affected by the variances. In conventional R&S procedures, the first step of sampling is usually used to estimate the variances, which becomes redundant in the GD setting here. As to the indifference zone threshold δ, it determines the basic number of samples required n_0 and also controls the Rademacher complexity $R_n(\mathscr{L})$. Noticeably, the number of samples required in R&S-GD procedure is an upper bound (more than necessary), therefore the number of samples required in practice can be smaller.*

5 SIMULATION EXPERIMENTS

In this section, we conduct two simulation experiments for our R&S procedures. Purportedly, we demonstrate that a right choice of R&S scheme can (1) efficiently reduce the numbers of simulation samples and (2) guarantee a high PCS.

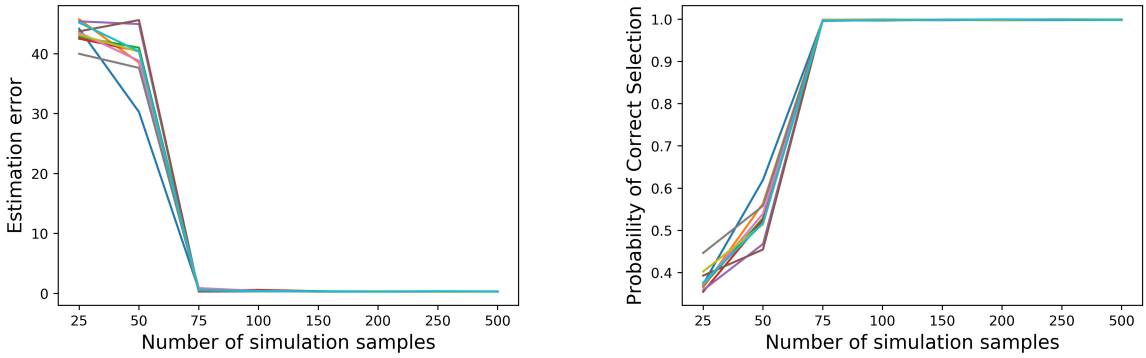
5.1 R&S-HC

In this experiment, we consider three alternatives and the responses dependent linearly on the covariates:

$$Y_k(X) = X^\intercal \beta_k + \varepsilon_k,$$

where $k = 1, 2, 3$. Let the dimension of covariate be 1000, i.e. $X \in \mathbb{R}^{1000}$ and the sparsity $\|\beta_k\|_0 = s_0 = 10$. This means each coefficient vector only has 10 non-zero entries. In R&S-HC, our procedure indeed guarantees $\mathrm{PCS}_{\mathrm{min}}$, but here to make the results comparable, we assume the entries of X is i.i.d. Gaussian normal and report the PCS_E. To make a sense of the procedure, we first conduct 10 trials of simulation, for each of which we initiate the true β_k's by randomly choosing 10 non-zero indices and assigning a random value uniformly from $\{-1, 1\}$; the simulation errors ε_k are generated i.i.d. from $N(0, 0.5)$. Then we do estimation with Lasso and make the selection decision based on our estimates $\widehat{\beta}_k$'s. The results are plotted as in Figure 1, where each trial is plotted with a different color.

From the figures, we observe that the estimation error and the wrong selection probability decrease fast to 0 with less than 100 simulation samples. Furthermore, we simulate for 5000 trials with 100 samples each, and we compute that the 95% confidence interval for $\mathrm{PCS}_{\mathrm{E}}$ is $[0.995, 0.999]$. If we adopt R&S procedure in Shen et al. (2017), we will need at least 1000 samples to implement OLS regression. Comparatively, our design leverages the sparsity (via Lasso) and effectively reduces the number of samples used. Specifically, in our design, the number of required samples grows linearly with the sparsity s_0 but logarithmically with



(a) Estimation Error (b) $\mathrm{PCS_E}$

Figure 1: Simulation Results for R&S-HC: In (a), we plot the estimation error against the number of simulation samples. The estimation error is the summation of the L_1 error for the alternatives, i.e. $\sum_{k=1}^{3} \|\widehat{\beta}_k - \beta_k\|_1$. In (b), we plot the $\mathrm{PCS_E}$ against the number of simulation samples. The $\mathrm{PCS_E}$ is obtained by randomly generating $N = 3000$ samples from p^* and computing the selection accuracy based on $\widehat{\beta}_k$'s.

the dimension of the covariates d. In fact, this sample efficiency highlights one useful scenario for R&S-HC: when there is a large number of covariates but only a few are relevant. If this is (or assumed to be) true, we can take advantage of the sparsity structure and save a great number of simulation samples.

5.2 R&S-GD

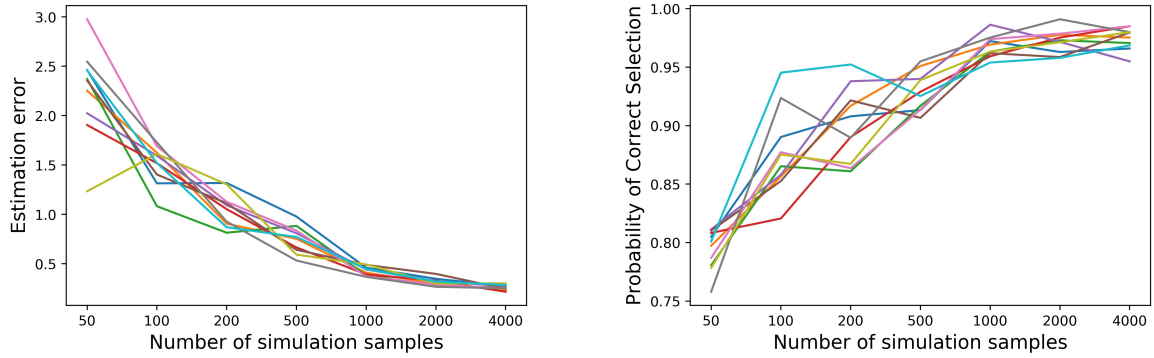
In this section, we conduct an experiment where the dependence between response and covariates is nonlinear. Similar to the previous section, we consider three alternatives:

$$Y_k(X) = f_k(X) + \varepsilon_k,$$

$k = 1,2,3$. Let the covariate $X \in \mathbb{R}^{10}$ and the entries follow i.i.d. Gaussian normal, thus specifying the p^* in Assumption 2 (b). We assign the dependence function f_k's to be random forests (Breiman (2001)) composed of three decision trees with randomly specified splits and depth. Consequently, the dependence functions are non-linear (even non-continuous). As before, the simulation errors ε_k are generated i.i.d. from $N(0,0.5)$. We obtain our estimates $\widehat{f}_k$'s by learning a random forest to minimize empirical risk (3) and then base our selection on $\widehat{f}_k$'s. We conduct 10 different random trials and plot the results as Figure 2.

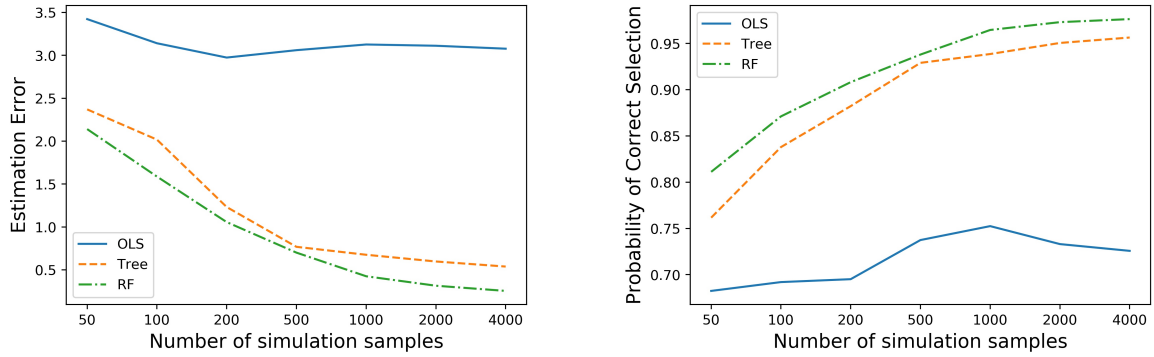
Different from the linear setting, much more simulation samples is required to have good prediction and selection accuracies. There is no surprise because we are searching the true dependence f_k's from a larger class of functions than the linear class. To counter the effect of the generality of dependence, we need more samples to identify the true functions and to make the correct selections.

Additionally, we do another experiment to see the importance of choosing a general enough function class. With the same setting as above, we consider two different function classes, say, $\mathscr{F}$ in (3), to be the class of linear functions or decision trees. Notice that the true function as a random forest function is not included in the two classes, this will result in an unsuccessful selection no matter how the procedure is designed or how many number of samples are collected. A linear model or a decision tree model here is not sufficient to capture the complicated dependence between response and covariates. As we can see in Figure 3, it results in the gap between the blue/orange curve and the green curve. This warns us the danger of choosing an over-simple model, like the linear model here, which will cause a large probability of selecting the non-optimal alternative.



(a) Out-of-sample MSEs (b) PCS_{E}

Figure 2: Simulation Results for R&S-GD: In (a), we plot the out-of-sample Mean-squared Errors (MSEs) against the number of simulation samples. The MSE's are computed from the prediction error of $\widehat{f}_k$ on $N = 3000$ new samples from p^*. In (b), we plot the PCS_{E} against the number of simulation samples. The selection decision is made upon the estimate $\widehat{f}_k$'s.



(a) Out-of-sample MSEs (b) PCS_{E}

Figure 3: Comparison of Different Function Classes: Given the true dependence being a random forest function, we compare learning the true from linear functions (OLS), Decision Tree (Tree) and Random Forest (RF). The MSE and PCS_{E} are defined in the same way as Figure 2. The curve is averaged from 100 simulation trials.

In short, if the response and covariates have a general nonlinear dependence, it costs more simulation samples to achieve a reliable R&S scheme. However, the additional samples are indispensable because adopting an over-simple function class to model the dependence pays a high price of making wrong selections.

6 CONCLUSION

In this paper, we extend the framework of R&S-C with two novel R&S procedures: R&S-HC and R&S-GD. These two procedures enable the full generality of how we model the dependence between responses and covariates in R&S. Statistical guarantees are established with the help of results from high-dimensional statistics and statistical learning theory. Simulation experiments demonstrate the efficiency of our procedures

and illustrate the importance of choosing the right scheme as well as the proper class of dependence functions.

REFERENCES

Aggarwal, C. C., Y. Zhao, and S. Y. Philip. 2014. "On the Use of Side Information for Mining Text Data". *IEEE Transactions on Knowledge and Data Engineering* 26(6):1415–1429.

Bastani, H., and M. Bayati. 2015. "Online Decision-Making with High-Dimensional Covariates". *Working Paper*.

Bechhofer, R. 1954. "A Single-Sample Multiple Decision Procedure for Ranking Means of Normal Populations with Known Variances". *The Annals of Mathematical Statistics* 25(1):16–39.

Bickel, P. J., Y. Ritov, and A. B. Tsybakov. 2009. "Simultaneous Analysis of Lasso and Dantzig Selector". *The Annals of Statistics* 37(4):1705–1732.

Breiman, L. 2001. "Random Forests". *Machine Learning* 45(1):5–32.

Bühlmann, P., and S. van de Geer. 2011. *Statistics for High-Dimensional Data: Methods, Theory and Applications*. Berlin Heidelberg: Springer-Verlag.

Candes, E. J., J. K. Romberg, and T. Tao. 2006. "Stable Signal Recovery from Incomplete and Inaccurate Measurements". *Communications on Pure and Applied Mathematics* 59(8):1207–1223.

Chen, C.-H. 1996. "A Lower Bound for the Correct Subset-Selection Probability and Its Application to Discrete-event System Simulations". *IEEE Transactions on Automatic Control* 41(8):1227–1231.

Chen, C.-H., S. E. Chick, L. H. Lee, and N. A. Pujowidianto. 2015. "Ranking and Selection: Efficient Simulation Budget Allocation". In *Handbook of Simulation Optimization*, edited by M. C. Fu, Chapter 3, 45–80. New York: Springer-Verlag.

Chen, Q., Z. Song, Y. Hua, Z. Huang, and S. Yan. 2012. "Hierarchical Matching with Side Information for Image Classification". In *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR 2012)*, 3426–3433. Red Hook, New York: Curran Associates Inc.

Chick, S. E., and P. Frazier. 2012. "Sequential Sampling with Economics of Selection Procedures". *Management Science* 58(3):550–569.

Chick, S. E., and K. Inoue. 2001. "New Two-Stage and Sequential Procedures for Selecting the Best Simulated System". *Operations Research* 49(5):732–743.

Fan, W., L. J. Hong, and B. L. Nelson. 2016. "Indifference-Zone-Free Selection of the Best". *Operations Research* 64(6):1499–1514.

Frazier, P., W. Powell, and S. Dayanik. 2008. "A Knowledge-Gradient Policy for Sequential Information Collection". *SIAM Journal on Control and Optimization* 47(5):2410–2439.

Hong, L. J., and B. L. Nelson. 2007. "Selecting the Best System when Systems Are Revealed Sequentially". *IIE Transactions* 39(7):723–734.

Hunter, S. R., and B. L. Nelson. 2017. "Parallel Ranking and Selection". In *Advances in Modeling and Simulation: Seminal Research from 50 Years of Winter Simulation Conferences*, edited by A. Tolk et al., 249–275. Basel, Switzerland: Springer International Publishing AG.

Kamiński, B., and P. Szufel. 2018. "On parallel policies for ranking and selection problems". *Journal of Applied Statistics* 45(9):1690–1713.

Kim, S.-H., and B. L. Nelson. 2001. "A Fully Sequential Procedure for Indifference-Zone Selection in Simulation". *ACM Transactions on Modeling and Computer Simulation* 11(3):251–273.

Kim, S.-H., and B. L. Nelson. 2006. "Selecting the Best System". In *Handbooks in Operations Research and Management Science*, edited by S. G. Henderson and B. L. Nelson, Volume 13, 501–534. Elsevier.

Lu, T., D. Pál, and M. Pál. 2010. "Contextual Multi-Armed Bandits". In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, edited by Y. W. Teh et al., Volume 9, 485–492: PMLR.

Luo, J., L. J. Hong, B. L. Nelson, and Y. Wu. 2015. "Fully Sequential Procedures for Large-Scale Ranking-and-Selection Problems in Parallel Computing Environments". *Operations Research* 63(5):1177–1194.

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

Ni, E. C., D. F. Ciocan, S. G. Henderson, and S. R. Hunter. 2017. "Efficient Ranking and Selection in Parallel Computing Environments". *Operations Research* 65(3):821–836.

Rinott, Y. 1978. "On Two-Stage Selection Procedures and Related Probability-Inequalities". *Communications in Statistics-Theory and Methods* 7(8):799–811.

Robbins, H. 1952. "Some Aspects of the Sequential Design of Experiments". *Bulletin of the American Mathematical Society* 58(5):527–535.

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 et al., 2137–2148. Piscataway, New Jersey: IEEE.

Tibshirani, R. 1996. "Regression Shrinkage and Selection via the Lasso". *Journal of the Royal Statistical Society. Series B (Methodological)* 58(1):267–288.

Vapnik, V. N. 1998. *Statistical Learning Theory*. New York: John Wiley & Sons, Inc.

Xu, M., R. Jin, and Z.-H. Zhou. 2013. "Speedup Matrix Completion with Side Information: Application to Multi-Label Learning". In *Proceedings of the 26th International Conference on Neural Information Processing Systems (NIPS 2013)*, edited by C. J. C. Burges et al., 2301–2309. Red Hook, New York: Curran Associates Inc.

ACKOWNLEDGEMENT

We thank the three anonymous reviewers for their helpful comments.

AUTHOR BIOGRAPHIES

XIAOCHENG LI is currently a Ph.D. candidate in Management Science and Engineering at Stanford University, specializing in Operations Research. He holds a B.Sc. in Mathematics from Peking University. His research interests include stochastic modeling, data-driven decision making, and machine learning applications. His email address is chengli1@stanford.edu.

XIAOWEI ZHANG received his Ph.D. in Management Science and Engineering and M.S. in Financial Mathematics, both from Stanford University. He is currently Assistant Professor in the Department of Industrial Engineering and Decision Analytics at the Hong Kong University of Science and Technology. His research interests include stochastic simulation, statistical learning, and decision analytics. His email address is xiaoweiz@ust.hk.

ZEYU ZHENG is an assistant professor in the Department of Industrial Engineering and Operations Research at the University of California, Berkeley. He received his Ph.D. in Management Science and Engineering, Ph.D. minor in Statistics, and M.A. in Economics, from Stanford University, and B.S. in Mathematics from Peking University. His has research interests in simulation, data analytics, stochastic modeling, machine learning, and fintech. His email address is me@zhengzeyu.com.