Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

DATA-DRIVEN TWO-STAGE STOCHASTIC PROGRAMMING WITH MARGINAL DATA

Ke Ren
Hoda Bidkhori

Department of Industrial Engineering
University of Pittsburgh
3700 O'Hara Street
Pittsburgh, PA 15261, USA

ABSTRACT

We present new methodologies to solve data-driven two-stage stochastic optimization when only the marginal data are available. We propose a novel data-driven distributionally robust framework that only uses the available marginal data. The proposed model is distinguished from the traditional techniques of solving missing data in that it conducts an integrated analysis of missing data and optimization problems, whereas classical methods conduct separate analyses by first recovering the missing data and then finding the optimal solutions. On the theoretical side, we show that our model produces risk-averse solutions and guarantees finite sample performance. Empirical experiments are conducted on two applications based on synthetic data and real-world data. We validate the proposed finite sample guarantee and show that the proposed approach achieves better out-of-sample performance and higher reliability than the classical data imputation-based approach.

1 INTRODUCTION

This paper proposes a risk-averse framework for data-driven two-stage stochastic optimization when only the marginal data are available. Two-stage stochastic programming (Dantzig 1955) is quite common in many real-world applications, such as network designs (Liu, Fan, and Ordóñez 2009), supply chain (Dillon, Oliveira, and Abbasi 2017), portfolio management (Gülten and Ruszczyński 2015), social networks (Wu and Küçükyavuz 2018) and health care (Denton, Miller, Balasubramanian, and Huschka 2010). Its objective is to minimize the total expected costs associated with both first-stage (here-and-now) and second-stage (wait-and-see) decisions, where the first-stage decisions have to be made before the realizations of the uncertainties in the second stage.

We denote the uncertainties of the second-stage problem as a random variable ξ. In most existing works, the distribution of ξ is assumed to be known in advance, which is often not true in practice. Recently, data-driven approaches have been proposed (Hanasusanto and Kuhn 2018; Zhao and Guan 2018; Jiang and Guan 2018; Chen, Sim, and Xiong 2020) to solve the two-stage stochastic programming. These methods work well if well-conditioned historical data for ξ are available. However, if ξ is multidimensional, the joint data of ξ are often hard to obtain in the real-world complex data environment.

For example, consider a supply chain optimization problem where ξ represents the demand scenarios at all retailers. Each dimension ξ_i represents the demand scenarios at the i-th retailer. In practice, the joint data of ξ are very few compared to the available marginal data. One other application is portfolio optimization, where missing data is a common problem (Rădulescu and Rădulescu 2013; Taylor 2006).

The existing data-driven two-stage stochastic programming models cannot address the issue of missing joint data as they rely on well-conditioned data. In this study, we propose a risk-averse, data-driven two-stage stochastic programming model that uses only the marginal data at each ξ_i. The model first

978-1-6654-3311-2/21/$31.00 ©2021 IEEE

proposes an ambiguity set $\mathscr{P}$ containing the possible joint distributions of ξ through the available marginal data and possible correlations. We then propose distributionally robust optimization (DRO) techniques to solve the original two-stage stochastic programming with respect to the worst-case joint distributions inside this ambiguity set. We further provide theoretical guarantees, tractable reformulations, and computational evaluations of the proposed methods based on synthetic and real-world data.

The proposed model uses DRO, which is a popular approach to optimization under uncertainty, with applications in reinforcement learning (Smirnova, Dohmatob, and Mary 2019), semi-supervised learning (Blanchet and Kang 2020), and classification (Abadeh, Esfahani, and Kuhn 2015). More introductions are provided in Section 3.

Our contributions. The proposed DRO model provides a new and robust way to solve data-driven two-stage stochastic optimization with marginal data. The classical approach based on data imputation (Lakshminarayan, Harp, Goldman, Samad, et al. 1996) follows a "estimate-then-optimize" paradigm, where it first imputes the missing data and then solves the optimization problem. The proposed model conducts an integrated analysis by combining the distribution estimation from missing data into the optimization problem. Our model outperforms the "estimate-then-optimize" paradigm in that it produces more robust decisions by considering the potential risks generated by the missing data estimations. The proposed model works in situations where the joint data are totally unavailable.

We evaluate the performance of the proposed model, both theoretically and numerically. We derive a probabilistic upper bound to the out-of-sample performance of the proposed model under a finite number of data. We further provide a reformulation of the proposed model, which can be applied to solve a wide range of problems efficiently. We conduct two numerical studies based on two applications: a risk-averse lot-sizing problem (Tarim and Kingsman 2004) and portfolio optimization (Alexander, Coleman, and Li 2006). In the first numerical study, we simulate 100 different joint distributions and validate the proposed theoretical upper bound. In the second numerical study, we apply our methods to 57 pairs of the training sets and test sets. These sets were obtained from real-world historical returns of exchange-traded funds (ETFs), and the US central bank (FED) rate of return from 2006 to 2016 (Boyd 2019). We justify the conclusion that the proposed model consistently yields better out-of-sample performance and higher reliability than the data imputation-based method.

2 PRELIMINARY

Assumptions and Notations Throughout this paper, we use the following notations and assumptions. A general two-stage stochastic programming can be formulated as

$$\min_{\mathbf{x}} \quad f(\mathbf{x}) + \mathbf{E}[Q(\mathbf{x}, \xi)], \tag{1}$$

where $f(\mathbf{x})$ is a first-stage cost and $Q(\mathbf{x}, \xi)$ is the optimal value of the second-stage problem depending on the random variable ξ and the first-stage decision $\mathbf{x}$. If ξ has an infinite number of possible realizations, (1) signifies a hard problem and is difficult to solve even approximately in general (Hanasusanto, Kuhn, and Wiesemann 2016). One common solution in the literature, called scenario construction (Shapiro, Dentcheva, and Ruszczyński 2014), is to select some representative scenarios. Therefore, in this paper, we assume the support of the random variable ξ is known and finite.

In addition, we assume that the second-stage problem is always bounded and feasible, which is a common assumption (Shapiro and Nemirovski 2005; Zhao and Guan 2018). Without loss of generality, we assume that each ξ_i ($i = 1, \cdots, I$) can take values of $\{\xi_i(1), \cdots, \xi_i(J)\}$. In addition, we use a I-dimensional vector $\mathbf{s}$, $s_i \in \{1, \cdots, J\}$, as an index to denote one joint possible realization of ξ for simplicity. We call $\mathbf{s}$ *scenario* throughout the rest of the paper. The $s_i = j$ indicates $\xi_i = \xi_i(j)$. The set containing all indexes of the scenarios is denoted as $\mathscr{S}$. The scenario $\mathbf{s}$ happens with probability $\Pr(\mathbf{s})$, which is not known but the historical data of each $\xi_i (1 \leqslant i \leqslant I)$ are available. We denote the n-th data for ξ_i as $\hat{\xi}_{in}$, $n = 1, \cdots, N_i$. Finally, for each ξ_i, we use $p_{i,j}$ to denote the true unknown marginal probability of $\xi_i = \xi_i(j)$, or $s_i = j$ equivalently.

3 RELATED WORK

The proposed study is related to two main areas: missing data and data-driven two-stage stochastic programming. In what follows, we provide a brief review of each topic.

Missing data. Missing data is a challenging factor in machine learning and statistics (García-Laencina, Sancho-Gómez, and Figueiras-Vidal 2010; Goodfellow, Bengio, and Courville 2016) because most models in the literature rely on complete data. One of the most natural options is to discard any data that include missing values. However, this approach may lead to biased results (Little and Rubin 2019) or result in the loss of information. Another common way is data imputation, which includes (Barnard and Meng 1999; Yoon, Jordon, and Schaar 2018; Śmieja, Struski, Tabor, Zieliński, and Spurek 2018). Data imputation approaches recover the incomplete data set by estimating missing values based on the observed data.

In this study, we adopt a novel, robust approach to tackle the problems when the joint data are unavailable or very rare. Instead of recovering the joint data through data imputation, we propose an integrated model to find optimal decisions based on the information of marginal distributions. Former works have considered of using only marginal distributions for decision-making. Researchers (Gao and Kleywegt 2017) assume the marginal distributions are known. The goal is to infer the unknown joint distribution based on the available data so as to solve the corresponding stochastic programming. To this goal, they build ambiguity set to capture possible dependence structures based on the Wasserstein distance. Another work (Chen, Ma, Natarajan, Simchi-Levi, and Yan 2018) also studies a related problem, where they focus on evaluating the bound of the inner maximization problem of a DRO problem with specified marginal distributions. Although this bound is shown to be NP-hard to compute, a sufficient condition for polynomial time solvability along with instances are identified. In our problem, the marginal distribution is also unknown. We build our model directly through marginal data.

Data-driven two-stage stochastic programming. The first data-driven two-stage stochastic programming approach signifies sample average approximation (SAA), which was first studied by (Shapiro and Homem-de Mello 1998; Kleywegt, Shapiro, and Homem-de Mello 2002). SAA approximates the unknown distribution in the second stage through empirical distributions. If the data samples of future true unknown distributions can be generated efficiently, the optimal solution of SAA converges to the true optimal with probability 1. If only a small number of samples are available, the performance of the obtained solution is not good. To overcome the challenges stemming from the finite or small sample size, researchers have proposed risk-averse two-stage stochastic programming models (Hanasusanto and Kuhn 2018; Zhao and Guan 2018). We are adopting ideas from distributionally robust optimization, which optimizes the worst-case expectations with respect to all distributions $\mathbb{P}$ inside the ambiguity set $\mathscr{P}$. The ambiguity set is defined as a family of distributions containing the true unknown distribution with high confidence. Choosing some well-constructed ambiguity set (Delage and Ye 2010) guarantees the out-of-sample performance of the model.

All the approaches discussed thus far use the complete joint data of ξ. However, in this paper, we build our model from the incomplete data set (when only marginal data are available) and solve the problem by considering possible correlations between components, which has been shown to be important in stochastic optimization (Agrawal, Ding, Saberi, and Ye 2012).

4 PROBLEM FORMULATION

In this section, we develop a risk-averse two-stage stochastic programming model, in which the distribution of ξ is not known, but the historical data of each component ξ_i are available. The main model is formulated and presented in Section 4.1. In this formulation, the unknown distribution of ξ is captured by an ambiguity set $\mathscr{P}$, which is introduced in Section 4.2. We further justify the performance of this framework theoretically and prove the finite sample guarantee in Section 4.3. Finally, we provide a simple closed-form reformulation in Section 4.4.

4.1 DRO model

The proposed model is formulated in (2), where the second-stage optimal cost $Q(\mathbf{x},\xi)$ is explicitly denoted as $\min_{\mathbf{y}\in\mathscr{Y}(\mathbf{x},\mathbf{s})} h(\mathbf{x},\mathbf{y},\mathbf{s})$, with $\mathscr{Y}(\mathbf{x},\mathbf{s})$ being the feasible region for the second-stage decision $\mathbf{y}$ (although sometimes we use $\mathscr{Y}$ to denote it for convenience). We assume the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are m_1 and m_2, $\mathbf{x}\in\mathbb{R}^{m_1}$ and $\mathbf{y}\in\mathbb{R}^{m_2}$. We do not restrict the forms of $f(\mathbf{x})$ or $h(\mathbf{x},\mathbf{y},\mathbf{s})$. Furthermore, we only assume $\mathscr{X}\subseteq\mathbb{R}^{m_1}$ and $\mathscr{Y}\subseteq\mathbb{R}^{m_2}$ here. We will discuss the tractability of the proposed model under different classes of functions and sets in Section 4.4.

$$\min_{\mathbf{x}\in\mathscr{X}} \quad f(\mathbf{x}) + \max_{\{\Pr(\mathbf{s}),\forall\mathbf{s}\in\mathscr{S}\}\in\mathscr{P}} \sum_{\mathbf{s}\in S} \Pr(\mathbf{s}) [\min_{\mathbf{y}\in\mathscr{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s})] \tag{2}$$

The objective function of (2) minimizes the expectation of the second-stage cost with respect to a worst-case distribution inside the ambiguity set $\mathscr{P}$.

4.1.1 Benefits of using Formulation (2)

Compared to the traditional approaches of solving missing data problems, model (2) is unique in that it combines the estimation step of $\{\Pr(\mathbf{s}),\forall\mathbf{s}\in\mathscr{S}\}$ and the derivation step of $\mathbf{x}\in\mathscr{X}$. Classical method relies on data imputation, which first estimates $\{\Pr(\mathbf{s}),\forall\mathbf{s}\in\mathscr{S}\}$ and then derives the optimal $\mathbf{x}$. The joint distribution inferred from model (2) will change according to $f(\mathbf{x})$ and $h(\mathbf{x},\mathbf{y},\mathbf{s})$ even if the data set remains the same. Further, by considering a worst-case distribution, specifically, the obtained optimal decision is risk-averse, which brings benefits in many real-world applications. We will illustrate this point in computational studies.

4.2 Ambiguity set $\mathscr{P}$

In this section, we propose an ambiguity set that uses only the marginal historical data. The main idea is to learn the unknown marginal distributions from the data and then consider the worst-case correlations among all components to obtain the joint distribution. The motivation of $\mathscr{P}$ is as follows. First of all, the true marginal probability $p_{i,j}$ is not known to the decision-makers. However, by the law of large numbers, the value of $p_{i,j}$ should be close to the empirical value $\hat{p}_{i,j} = \frac{\sum_{n=1}^{N_i} \mathbf{1}(\hat{\xi}_{in} = \xi_i(j))}{N_i}$, where $\mathbf{1}()$ is an indicator function. Therefore, we model the true $p_{i,j}$ by $p_{i,j} = \hat{p}_{i,j} + d_{i,j}$, where $d_{i,j}$ represents a small deviation. The ambiguity set $\mathscr{P}$ is defined as follows and we explain it line by line.

$$\mathscr{P} = \left\{ \{\Pr(\mathbf{s}),\forall\mathbf{s}\in\mathscr{S}\}: \begin{array}{l} \sum_{\{\mathbf{s}|s_i=j\}} \Pr(\mathbf{s}) = \hat{p}_{i,j} + d_{i,j}, \forall i,j, \\ \sum_{j=1}^{J} d_{i,j} = 0, \forall i, \\ \sum_{j=1}^{J} |d_{i,j}| \leqslant \tau_i, \forall i, \\ \sum_{\mathbf{s}\in\mathscr{S}} \Pr(\mathbf{s}) = 1, \\ \Pr(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s}\in\mathscr{S}. \end{array} \right\}. \tag{3}$$

The first line restricts the marginal distributions for ξ_i deviate from the empirical distributions by a small quantity. The second line ensures that the marginal distribution is valid because for any i, there is $\sum_{j=1}^{J} \hat{p}_{i,j} = 1$. The third line restricts the total magnitude of the deviations in i-th component to be less than τ_i. Later, we will show that the robustness of the model can be tuned by adjusting τ_i. The last two lines ensure that all joint distributions of ξ satisfying the corresponding constraints on marginal distributions are included in the ambiguity set.

4.3 Finite sample guarantee

In this section, we provide the finite sample guarantee (Esfahani and Kuhn 2018) of DRO frameworks for (2) by showing that the optimal objective value of (2) is a probabilistic upper bound of the objective value

in case the true distribution is known. The main results are summarized in Proposition 1 and we use the following lemma to prove it.

Lemma 1 (Bretagnolle-Huber-Carol inequality) (Bellet and Habrard 2015) If the random vector $(v_1, \cdots, v_a)$ is multinomially distributed with parameters N and $(p_1, \cdots, p_a)$, then for any $c \geqslant 0$,

$$\Pr(\sum_{i=1}^{a} |v_i - Np_i| \geqslant 2\sqrt{N}c) \leqslant 2^a e^{-2c^2}.$$

We obtain the following Proposition 1 by setting $\tau_i = \frac{2c}{\sqrt{N}}$ with some well-chosen c.

Proposition 1 Let N_i be the number of the data for ξ_i, $0 < \alpha_i < 1$ for $i = 1, \cdots, I$ and $\tau_i = \sqrt{\frac{2\ln\frac{2^J}{\alpha_i}}{N_i}}$. Then with probability at least $1 - \alpha_i$ all the true $p_{i,j}$, $j = 1, \cdots, J$, are contained in the feasible region of (2). Suppose the optimal solution of (2) is $\hat{\mathbf{x}}^*$ with objective value $\hat{O}$. With probability at least $\prod_{i=1}^{I}(1-\alpha_i)$, we have

$$f(\hat{\mathbf{x}}^*) + \sum_{s \in \mathscr{S}} \Pr^*(\mathbf{s})[\min_{\mathbf{y} \in \mathscr{Y}} h(\hat{\mathbf{x}}^*, \mathbf{y}, \mathbf{s})] \leqslant \hat{O},$$

where $\Pr^*(\mathbf{s})$ denotes the true probability for the $\mathbf{s}$-th scenario in the future.

Proof. We provide a brief proof sketch here due to the limited space. Bretagnolle-Huber-Carol inequality is equivalent to

$$\Pr(\sum_{i=1}^{a} |\frac{v_i}{N} - p_i| \geqslant \frac{2}{\sqrt{N}}c) \leqslant 2^a e^{-2c^2}.$$

We view p_i as the true distribution and $\frac{v_i}{N}$ as the empirical distribution. According to (3), the difference between the true distribution and the empirical distribution is less than τ_i. Therefore, if we set $\tau_i = \frac{2}{\sqrt{N}}c$, then $2^a e^{-2c^2}(a = J)$ represents an upper bound of the violation probability α_i. Correspondingly, solving the obtained two equalities $\tau_i = \frac{2}{\sqrt{N}}c$, $\quad 2^a e^{-2c^2} = \alpha_i$ gives us the result. □

Therefore, we can control the conservativenes of the proposed model by adjusting the values of τ_i defined in $\mathscr{P}$. For a fixed confidence level α_i, the total deviation τ_i decreases at a rate of $\frac{1}{\sqrt{N_i}}$. In general, an increase in τ_i will reduce the violation probability. Correspondingly, the true costs will have a higher probability to be bounded by the costs of the optimal solution of the proposed model. However, similar to other DRO frameworks (Delage and Ye 2010; Esfahani and Kuhn 2018), we recommend to use cross-validations to choose an appropriate τ_i in practice for better performances because this bound is often too conservative.

4.4 Worst-case reformulation

We provide a reformulation of (2) and discuss its tractability. The main idea is to replace the worst-case second-stage problem with its dual problem. The results are summarized in Theorem 1.

Theorem 1 Optimization (2) is equivalent to the following optimization problem:

$$\begin{aligned}
\min_{(\cdots)} \quad & f(\mathbf{x})+\gamma+\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{i,j}\hat{p}_{i,j}+\sum_{i=1}^{I}\rho_i\tau_i \\
s.t. \quad & h(\mathbf{x},\mathbf{y_s},\mathbf{s}) \leqslant \sum_{\{i,j|s_i=j\}}\lambda_{i,j}-\gamma, \forall \mathbf{s}\in\mathscr{S}, \\
& \mathbf{y_s}\in\mathscr{Y}(\mathbf{x},\mathbf{s}), \forall \mathbf{s}\in\mathscr{S}, \\
& \beta_i-\theta_{i,j}+\lambda_{i,j}+\pi_{i,j}=0, \forall i,j, \\
& \theta_{i,j}+\pi_{i,j}-\rho_i=0, \forall i,j, \\
& \mathbf{x}\in\mathscr{X},
\end{aligned} \tag{4}$$

where we use $(\cdots)$ to indicate $\{\mathbf{x},\lambda_{i,j},\gamma,\beta_i,\rho_i\geqslant 0,\theta_{i,j}\geqslant 0,\pi_{i,j}\geqslant 0\}$ and $\mathbf{y_s}\in\mathbb{R}^{m_2}$.

Proof. Problem (2) is equivalent to:

$$\begin{aligned}
\min_{\mathbf{x}\in\mathscr{X}} \quad & f(\mathbf{x})+\max_{\Pr(\mathbf{s}),d_{i,j},z_{i,j}}\sum_{\mathbf{s}\in\mathscr{S}}\Pr(\mathbf{s})[\min_{\mathbf{y}\in\mathscr{Y}}h(\mathbf{x},\mathbf{y},\mathbf{s})] \\
s.t. \quad & \sum_{\{\mathbf{s}|s_i=j\}}\Pr(\mathbf{s})=\hat{p}_{i,j}+d_{i,j}, \quad \forall i,j, \\
& \sum_{j=1}^{J}d_{i,j}=0, \sum_{j=1}^{J}z_{i,j}\leqslant\tau_i, \quad \forall i, \\
& z_{i,j}\geqslant d_{i,j}, z_{i,j}\geqslant -d_{i,j}, \quad \forall i,j, \\
& \sum_{\mathbf{s}\in\mathscr{S}}\Pr(\mathbf{s})=1, \Pr(\mathbf{s})\geqslant 0, \quad \forall \mathbf{s}\in\mathscr{S}.
\end{aligned} \tag{5}$$

The Lagrangian of the second-stage maximization problem is:

$$\begin{aligned}
\max_{\Pr(\mathbf{s})\geqslant 0,d_{i,j},z_{i,j}}\min_{(\cdots)} & \sum_{\mathbf{s}\in\mathscr{S}}\Pr(\mathbf{s})[\min_{\mathbf{y}\in\mathscr{Y}}h(\mathbf{x},\mathbf{y},\mathbf{s})]-\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{i,j}(\sum_{\{\mathbf{s}|s_i=j\}}\Pr(\mathbf{s})-\hat{p}_{i,j}-d_{i,j}) \\
& -\gamma(\sum_{\mathbf{s}\in S}\Pr(\mathbf{s})-1)+\sum_{i=1}^{I}\beta_i\sum_j d_{i,j}-\sum_{i=1}^{I}\rho_i(\sum_j z_{i,j}-\tau_i) \\
& +\sum_{i=1}^{I}\sum_{j=1}^{J}\theta_{i,j}(z_{i,j}-d_{i,j})+\sum_{i=1}^{I}\sum_{j=1}^{J}\pi_{i,j}(z_{i,j}+d_{i,j})
\end{aligned} \tag{6}$$

Due to the space issue, we use $(\cdots)$ to indicate $\{\lambda_{i,j},\gamma,\beta_i,\rho_i\geqslant 0,\theta_{i,j}\geqslant 0,\pi_{i,j}\geqslant 0\}$. Problem (6) is linear with respect to $\Pr(\mathbf{s})$, $\boldsymbol{d}$, and $\boldsymbol{z}$ when the Lagrangian multipliers are fixed and is also linear with respect to the Lagrangian multipliers when $\Pr(\mathbf{s}),\boldsymbol{d},\boldsymbol{z}$ are fixed. Therefore, following the minimax theorem, we can switch the min and max. Therefore, (6) is equivalent to:

$$\begin{aligned}
\min_{(\cdots)}\max_{\Pr(\mathbf{s})\geqslant 0,d_{i,j},z_{i,j}} & \sum_{\mathbf{s}\in\mathscr{S}}\Pr(\mathbf{s})[\min_{\mathbf{y}\in\mathscr{Y}}h(\mathbf{x},\mathbf{y},\mathbf{s})-\sum_{\{i,j|s_i=j\}}\lambda_{i,j}-\gamma]-\sum_{i=1}^{I}\sum_{j=1}^{J}\lambda_{i,j}(-\hat{p}_{i,j})-\gamma(-1) \\
& +\sum_{i=1}^{I}\sum_{j=1}^{J}d_{i,j}(\beta_i-\theta_{i,j}+\lambda_{i,j}+\pi_{i,j})-\sum_{i=1}^{I}\rho_i(-\tau_i)+\sum_{i=1}^{I}\sum_{j=1}^{J}z_{i,j}(\theta_{i,j}+\pi_{i,j}-\rho_k)
\end{aligned} \tag{7}$$

The primal problem is always feasible based on the assumption made in the introduction, which indicates the following constraints for (7).

$$\begin{aligned}
&\min_{\mathbf{y}\in\mathscr{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s}) - \sum_{\{i,j|s_i=j\}} \lambda_{i,j} - \gamma \leqslant 0,\\
&\beta_i - \theta_{i,j} + \lambda_{i,j} + \pi_{i,j} = 0, \forall i,j,\\
&\theta_{i,j} + \pi_{i,j} - \rho_i = 0, \forall i,j.
\end{aligned}$$

Correspondingly, (7) is equivalent to:

$$\begin{aligned}
\min_{(\cdots)} \quad & \gamma + \sum_{i=1}^{I}\sum_{j=1}^{J} \lambda_{i,j}\hat{p}_{i,j} + \sum_{i=1}^{I} \rho_i \tau_i \\
s.t. \quad & \min_{\mathbf{y}\in\mathscr{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s}) - \sum_{\{i,j|s_i=j\}} \lambda_{i,j} - \gamma \leqslant 0, \forall \mathbf{s}\in\mathscr{S},\\
& \beta_i - \theta_{i,j} + \lambda_{i,j} + \pi_{i,j} = 0, \forall i,j,\\
& \theta_{i,j} + \pi_{i,j} - \rho_i = 0, \forall i,j.
\end{aligned} \tag{8}$$

Notice that the minimization in the first set of constraints appears on the smaller side of the inequality. Therefore, the minimization can be eliminated by introducing auxiliary variables $\mathbf{y_s}$.

$$\min_{\mathbf{y}\in\mathscr{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s}) \leqslant \sum_{\{i,j|s_i=j\}} \lambda_{i,j} + \gamma \Rightarrow h(\mathbf{x},\mathbf{y_s},\mathbf{s}) \leqslant \sum_{\{i,j|s_i=j\}} \lambda_{i,j} + \gamma, \mathbf{y_s}\in\mathscr{Y}(\mathbf{x},\mathbf{s}).$$

Finally, we add the first stage problem back, which completes the proof. □

4.4.1 Tractability

In a classical two-stage stochastic programming, there is $h(\mathbf{x},\mathbf{y},\mathbf{s}) = \mathbf{q}(\mathbf{s})^T\mathbf{y}$ and $\mathscr{Y} = T(\mathbf{s})\mathbf{x} + W(\mathbf{s})\mathbf{y} \leqslant \mathbf{r}(\mathbf{s})$, where $\mathbf{q}(\mathbf{s}) \in \mathbb{R}^{m_1}$, $T(\mathbf{s}) \in \mathbb{R}^{m_3\times m_1}$, $W(\mathbf{s}) \in \mathbb{R}^{m_3\times m_2}$, and $\mathbf{r}(\mathbf{s}) \in \mathbb{R}^{m_3}$. Then, Optimization (4) is a simple linear programming. In general, if $f(\mathbf{x})$ and $h(\mathbf{x},\mathbf{y},\mathbf{s})$ are convex with respect to $\mathbf{x}$ and $\mathbf{y}$, and $\mathscr{X}$, $\mathscr{Y}$ are convex sets or mixed-integer linear sets (Conforti, Cornuéjols, and Zambelli 2014). Appropriate methods can be applied to solve (4).

4.5 Applications in risk-averse lot-sizing problem

We apply the proposed model to the risk-averse lot-sizing problem. The settings used are mostly adopted from (Jiang and Guan 2018). This problem optimizes the total costs of supplying the demand ξ in T periods. Before the realization of the demand ξ, decision-makers need to decide which periods require production setup. After the realization of the demand ξ, the decision-makers schedule the production plan to meet the demand at the cheapest costs, where the production can only happen during the periods chosen before.

4.5.1 Problem setup

This problem is defined as follows. A decision-maker schedules production in a finite operational interval $\{1,\cdots,T\}$. In each period $t \in \{1,\cdots,T\}$, production amount can be made up to a capacity C_t at a setup cost c_t and a unit production cost q_t. A random demand ξ_t needs to be fulfilled at the beginning of each time period t. The demand can be satisfied by either immediate production or inventory left from previous periods, or outsourcing (buying from other suppliers). Positive inventory at the end of period t is charged with a unit holding cost h_t, while each outsourcing incurs a unit cost M_t.

The first-stage cost is the setup cost and the first-stage decision $\mathbf{x} \in \{0,1\}^T$ signifies whether or not to produce at time period t, $t = 1, \cdots, T$. The second-stage cost signifies the cost of meeting the demand, where the decisions $\mathbf{y}, \mathbf{p} \in \mathbb{R}^T$ represent the production quantity and outsourcing quantity respectively in T time periods. The random demand $\xi \in \mathbb{R}^T$ is a T-dimensional variable. Therefore, the total expected costs for lot-sizing problems are:

$$\min_{\mathbf{x}} \quad \sum_{t=1}^{T} c_t x_t + \mathbf{E}[Q(\mathbf{x}, \xi)]. \tag{9}$$

The second-stage costs depend on the random demands ξ during T periods. We define a variable $v \in \mathbb{R}^{T+1}$ to denote the inventory levels at time period t. Then, the second-stage optimal cost is formulated as

$$\begin{aligned} Q(\mathbf{x}, \xi) = \min_{\mathbf{y},\mathbf{p}} \quad & \sum_{t=1}^{T} q_t y_t(\xi) + \sum_{t=1}^{T} M_t p_t(\xi) + \sum_{t=1}^{T} h_t v_t(\xi) \\ s.t. \quad & v_{t-1}(\xi) + y_t(\xi) + p_t(\xi) \\ & = \xi_t + v_t(\xi), t = 1, \cdots, T, \\ & y_t(\xi) \leqslant C_t x_t, t = 1, \cdots, T, \\ & v_0(\xi) = v_T(\xi) = 0, \\ & y_t(\xi), p_t(\xi), v_t(\xi) \geqslant 0, \forall t. \end{aligned} \tag{10}$$

The first series of constraints in (10) characterize the relationship between the inventory levels at time t and $t-1$. The second series of constraints restrict the production quantity to be less than the production capacity. Accordingly, the risk-averse lot-sizing problem is modeled as

$$\min_{\mathbf{x}} \quad \sum_{t=1}^{T} c_t x_t + \max_{\mathbb{P} \in \mathscr{P}} \mathbf{E}_{\mathbb{P}}[Q(\mathbf{x}, \xi)],$$

where we denote $\mathbb{P}$ as the distribution of ξ.

4.5.2 Experimental settings

Suppose the optimal objective value of the proposed model under marginal data is $\hat{O}$, and the realized cost under the true joint distribution is O^*. In the experiments, we want to evaluate the finite sample guarantee $\Pr(\hat{O} > O^*)$. In addition, we are interested in determining if $\hat{O}$ can effectively approximate O^* by using only the marginal distributions. To this goal, we randomly simulate a joint distribution $\mathbb{P}$ and collect the marginal data from it. Then we solve the proposed model based on the marginal data and evaluate its performance through the distribution $\mathbb{P}$. The detailed procedure for one experiment is described as follows: 1) We randomly generate a joint distribution $\mathbb{P}$ for ξ, then we simulate 20 marginal data for each ξ_t. 2) We solve model discussed in Section 4.5.1 based on the available marginal data to obtain the optimal $\mathbf{x}$ with objective value $\hat{O}$. 3) We evaluate the out-of-sample performance of $\mathbf{x}$ by using $\mathbb{P}$ and denote the cost as O^*.

The rest of the parameter settings are as follows. We set $T = 20$, unit holding cost $h_t = \$1$, setup cost $c_t = \$300$, outsourcing cost $M_t = \$12$, and unit production cost $q_t = \$3$ for each $t \in \{1, \cdots, T\}$. The random demand at each time period has three scenarios, $\xi_t = 10, 20, 30$. We simulate the true joint demand distribution by randomly sampling ten joint scenarios and assigning them with equal probabilities, which gives random correlations between all components. This joint demand distribution is unknown to the proposed approach.

4.5.3 Results

We report the simulation results in this subsection. We record the probability $p = \Pr(\hat{O} > O^*)$ under different radius τ_i of ambiguity set $\mathscr{P}$ defined in (3). Because the number of marginal data is the same for

each component, we set $\tau_1 = \cdots = \tau_T = \tau$ and vary τ accordingly. Each p is calculated via 100 experiments with different joint demand distributions and marginal data. The results are summarized in Table 1. We also plot the two graphs of the values of $\hat{O}$ and O^* when $\tau = 0.1$ and 0.3 in Figure 1. The horizontal axis represents the index of demand distribution.

Table 1 shows that the robustness of the proposed model increases as τ in the ambiguity set increases. This validates the results discussed in Section 4.3. In addition, from the Figure 1, we observe the curves of $\hat{O}$ and O^* are pretty close when τ is small and deviate from each other when τ becomes large. For both cases, the shapes of the two curves are very similar. Therefore, the proposed model approximates the true costs effectively by using only the marginal data. Additionally, as shown in Table 1, when τ is very small (< 0.25), the finite sample guarantee p increases rapidly with respect to the increases in τ. However, it slows down after τ becomes large (> 0.3). Therefore, a good choice of τ will be between 0.25 and 0.3, which achieves a balance between risks and conservativeness.

Table 1: Probability $p = \Pr(\hat{O} > O^*)$ under different values of τ.

τ	0.05	0.1	0.15	0.2	0.25
p	56%	66%	74%	88%	92%
τ	0.3	0.35	0.4	0.45	
p	96%	96%	100%	100%	

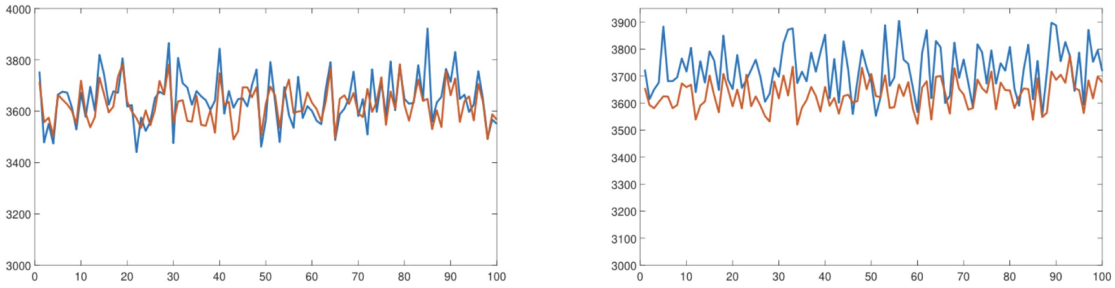

Figure 1: The red lines represent O^* and the blue lines represent $\hat{O}$. The left figure and the right figure represent the values of $\hat{O}$ and O^* for 100 different demand distributions for $\tau = 0.1$ and $\tau = 0.3$, respectively.

4.6 Applications in portfolio optimization

We demonstrate the benefits of the proposed approach by simulations on one real-world application: portfolio optimization, where missing data is a common problem (Rădulescu and Rădulescu 2013; Taylor 2006). In the subsequent numerical experiments, we show that the proposed approach outperforms the data imputation-based method on the real-world historical returns of ETFs.

4.6.1 Problem setup

Mean-risk portfolio optimization (Alexander, Coleman, and Li 2006) considers m assets with returns captured by the random vector $\xi = [\xi_1, \cdots, \xi_m]$. A portfolio is denoted by a vector $\mathbf{x} = [x_1, \cdots, x_m]$, where $\mathbf{x} \in \mathscr{X} = \{\mathbf{x} \in \mathbb{R}^m_+ | \sum_{i=1}^m x_i = 1\}$. Each x_i represents the percentage of the investment in asset i for each $1 \leqslant i \leqslant m$. The objective function aims to minimize a weighted sum of the mean of negative returns and the conditional value-at-risk as shown in (11).

$$\min_{\mathbf{x} \in \mathscr{X}} \quad \mathbf{E}_{\mathbb{P}}\left[\langle -\xi, \mathbf{x} \rangle\right] + \rho \mathrm{CVaR}_{\alpha}(\langle -\xi, \mathbf{x} \rangle) \tag{11}$$

where $\rho > 0$, and $\mathbb{P}$ represents the distribution of ξ.

Replacing the CVaR in (11) with its definition (Rockafellar, Uryasev, et al. 2000), the corresponding DRO model is equivalent to (12). It can be viewed as a two-stage problem with the first-stage cost being equal to zero.

$$\min_{\mathbf{x}\in\mathscr{X},d\in\mathbb{R}} \max_{\mathbb{P}\in\mathscr{P}} \quad \mathbf{E}_{\mathbb{P}}(\langle -\xi,\mathbf{x}\rangle)+\rho\left[d+\frac{1}{\alpha}\mathbf{E}_{\mathbb{P}}(\langle -\xi,\mathbf{x}\rangle - d)^{+}\right]. \tag{12}$$

4.6.2 Results

We first explain the way to generate the training set, which is assumed to be known to decision-makers, and the test set, which is unknown to decision-makers and is used to evaluate the out-of-sample performance, as follows. We approximate the support $\mathscr{S}$ of the returns ξ with the first 300 data. We iteratively use the data from day $[301+30\times(i-1)]$ to day $(300+30\times i)$ (approximately one month) as the training set, $1\leqslant i\leqslant 57$. The rest of the data are used as the test set. Therefore, we obtain 57 pairs of the training set and test set. We present the experiments results in the following.

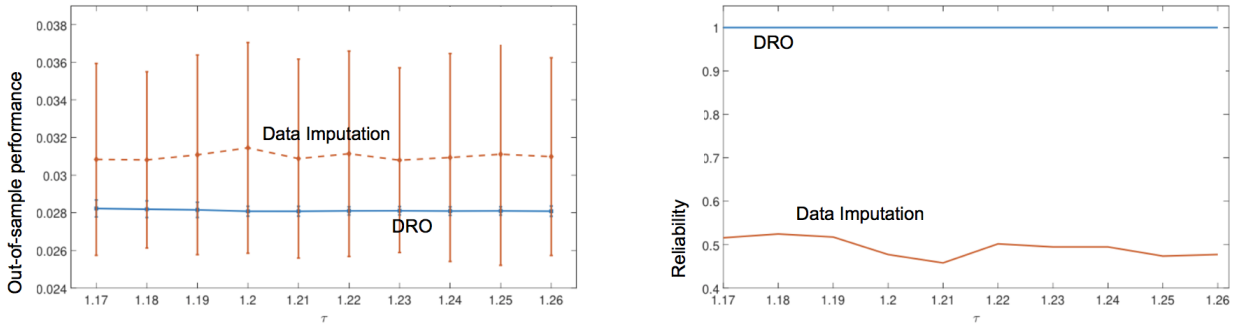

Figure 2: Out-of-sample performance and reliability under different values of τ.

Table 2: Out-of-sample performance ($\times 0.001$) under different values of τ.

τ	DRO	Data imputation
1.17	28.23 ± 0.45	30.84 ± 5.10
1.18	28.19 ± 0.44	30.81 ± 4.69
1.19	28.16 ± 0.41	31.08 ± 5.31
1.20	28.08 ± 0.26	31.45 ± 5.60
1.21	28.08 ± 0.26	31.14 ± 5.29
1.22	28.10 ± 0.22	30.80 ± 5.46
1.23	28.11 ± 0.22	30.94 ± 4.91
1.24	28.09 ± 0.23	31.11 ± 5.53
1.25	28.10 ± 0.22	30.99 ± 5.90
1.26	28.09 ± 0.27	30.88 ± 5.25

Table 3: Reliability $\Pr(\hat{O}>O^{*})$ under different values of τ.

τ	DRO	Data imputation
1.17	100%	51.58%
1.18	100%	52.46%
1.19	100%	51.75%
1.20	100%	47.72%
1.21	100%	45.79%
1.22	100%	50.18%
1.23	100%	49.47%
1.24	100%	49.47%
1.25	100%	47.37%
1.26	100%	47.72%

We use the first $m=5$ assets during the experiments. Therefore, the data set used in the experiments contains 5 assets with 2020 daily returns each. To model the marginal data, we assume each data in the training set has only one dimension observed, and we restrict each dimension has the same size of observed data for convenience. The data imputation approach uses nearest neighbor imputation first to recover the incomplete data set; then it solves model (11) through the empirical distribution to obtain the final decisions. In the first step, because the support of $\mathscr{S}$ is known, we impute each incomplete data to the nearest support with respect to L1 norm. Besides, random support is used if several supports achieve the smallest distance at the same time. We conclude the main numerical results based on marginal data in this section, where

the proposed DRO model achieves better out-of-sample performance and higher reliability than the data imputation approach.

We conducted 10 groups of experiments under different values of τ. We set all τ_t as the same value τ because the sizes of the marginal data at each dimension are the same. Notice that the feasible region can be empty for the ambiguity set defined in (3) when the sample size and τ are both extremely small. This is because the observed data are highly missed making the nominal marginal distribution highly biased. Therefore, we test the τ starting from the smallest τ that makes the problem feasible during the experiments. Each group of experiments contains 57 experiments as discussed in Section 4.5.2.

We repeat each experiments for 10 times to obtain the standard deviation. The results are concluded in Table 2 and Figure 2. The proposed DRO model outperforms the data imputation approach consistently by achieving lower values in (12) and smaller standard deviations. The standard deviations of the DRO model slightly decrease as the increase of τ. The reliability is defined as the probability $\Pr(\hat{O} > O^*)$. We conclude results in Table 3 and Figure 2. The proposed DRO model achieves much higher reliability than the data imputation approach. Because we consider the worst-case correlation for the marginal distributions, the reliability of the DRO model stays at 100% in this problem.

5 CONCLUSION

In this paper, we develop a novel integrated approach to solve data-driven stochastic optimization when only marginal data are available. We first construct an ambiguity set of possible joint distributions from the available marginal data. Then we formulate our model based on a distributionally robust optimization framework. Numerical experiments are conducted based on both synthetic data and real-world data. The proposed approach achieves promising results as it demonstrates that the risk-averse property of the proposed model can be tuned by adjusting the parameters defined in the ambiguity set and produces better out-of-sample performance than the classical estimate-then-optimize approach.

REFERENCES

Abadeh, S. S., P. M. M. Esfahani, and D. Kuhn. 2015. "Distributionally Robust Logistic Regression". In *Advances in Neural Information Processing Systems*, 1576–1584.

Agrawal, S., Y. Ding, A. Saberi, and Y. Ye. 2012. "Price of Correlations in Stochastic Optimization". *Operations Research* 60(1):150–162.

Alexander, S., T. F. Coleman, and Y. Li. 2006. "Minimizing CVaR and VaR for a Portfolio of Derivatives". *Journal of Banking & Finance* 30(2):583–605.

Barnard, J., and X.-L. Meng. 1999. "Applications of Multiple Imputation in Medical Studies: from AIDS to NHANES". *Statistical Methods in Medical Research* 8(1):17–36.

Bellet, A., and A. Habrard. 2015. "Robustness and Generalization for Metric Learning". *Neurocomputing* 151:259–267.

Blanchet, J., and Y. Kang. 2020. "Semi-supervised Learning Based on Distributionally Robust Optimization". *Data Analysis and Applications 3: Computational, Classification, Financial, Statistical and Stochastic Methods* 5:1–33.

Boyd, Stephen 2019. "Data for Finance and Portfolio Optimization". http://stanford.edu/class/ee103/portfolio.html.

Chen, L., W. Ma, K. Natarajan, D. Simchi-Levi, and Z. Yan. 2018. "Distributionally Robust Linear and Discrete Optimization with Marginals". *Available at SSRN 3159473*.

Chen, Z., M. Sim, and P. Xiong. 2020. "Robust Stochastic Optimization Made Easy with RSOME". *Management Science* 66(8):3329–3339.

Conforti, M., G. Cornuéjols, and G. Zambelli. 2014. *Integer Programming*, Volume 271. Springer.

Dantzig, G. 1955. "Linear Programming under Uncertainty". *Management Science* 1(3-4):197–281.

Delage, E., and Y. Ye. 2010. "Distributionally Robust Optimization Under Moment Uncertainty with Application to Data-Driven Problems". *Operations Research* 58(3):595–612.

Denton, B. T., A. J. Miller, H. J. Balasubramanian, and T. R. Huschka. 2010. "Optimal Allocation of Surgery Blocks to Operating Rooms Under Uncertainty". *Operations Research* 58(4-part-1):802–816.

Dillon, M., F. Oliveira, and B. Abbasi. 2017. "A Two-Stage Stochastic Programming Model for Inventory Management in the Blood Supply Chain". *International Journal of Production Economics* 187:27–41.

Esfahani, P. M., and D. Kuhn. 2018. "Data-Driven Distributionally Robust Optimization Using the Wasserstein Metric: Performance Guarantees and Tractable Reformulations". *Mathematical Programming* 171(1-2):115–166.

Gao, R., and A. J. Kleywegt. 2017. "Data-Driven Robust Optimization with Known Marginal Distributions". *Working paper*.

García-Laencina, P. J., J.-L. Sancho-Gómez, and A. R. Figueiras-Vidal. 2010. "Pattern Classification with Missing Data: a Review". *Neural Computing and Applications* 19(2):263–282.

Goodfellow, I., Y. Bengio, and A. Courville. 2016. *Deep Learning*. MIT Press.

Gülten, S., and A. Ruszczyński. 2015. "Two-Stage Portfolio Optimization with Higher-Order Conditional Measures of Risk". *Annals of Operations Research* 229(1):409–427.

Hanasusanto, G. A., and D. Kuhn. 2018. "Conic Programming Reformulations of Two-Stage Distributionally Robust Linear Programs over Wasserstein Balls". *Operations Research* 66(3):849–869.

Hanasusanto, G. A., D. Kuhn, and W. Wiesemann. 2016. "A Comment on "Computational Complexity of Stochastic Programming Problems"". *Mathematical Programming* 159(1-2):557–569.

Jiang, R., and Y. Guan. 2018. "Risk-Averse Two-Stage Stochastic Program with Distributional Ambiguity". *Operations Research* 66(5):1390–1405.

Kleywegt, A. J., A. Shapiro, and T. Homem-de Mello. 2002. "The Sample Average Approximation Method for Stochastic Discrete Optimization". *SIAM Journal on Optimization* 12(2):479–502.

Lakshminarayan, K., S. A. Harp, R. P. Goldman, T. Samad et al. 1996. "Imputation of Missing Data Using Machine Learning Techniques.". In *KDD*, Volume 96.

Little, R. J., and D. B. Rubin. 2019. *Statistical Analysis with Missing Data*, Volume 793. Wiley.

Liu, C., Y. Fan, and F. Ordóñez. 2009. "A Two-Stage Stochastic Programming Model for Transportation Network Protection". *Computers & Operations Research* 36(5):1582–1590.

Rădulescu, M., and C. Z. Rădulescu. 2013. "Mean-Variance Models with Missing Data". *Studies in Informatics and Control* 22(4):299–306.

Rockafellar, R. T., S. Uryasev et al. 2000. "Optimization of Conditional Value-at-Risk". *Journal of Risk* 2:21–42.

Shapiro, A., D. Dentcheva, and A. Ruszczyński. 2014. *Lectures on Stochastic Programming: Modeling and Theory*. SIAM.

Shapiro, A., and T. Homem-de Mello. 1998. "A Simulation-Based Approach to Two-Stage Stochastic Programming with Recourse". *Mathematical Programming* 81(3):301–325.

Shapiro, A., and A. Nemirovski. 2005. "On Complexity of Stochastic Programming Problems". In *Continuous optimization*, 111–146. Springer.

Śmieja, M., Ł. Struski, J. Tabor, B. Zieliński, and P. Spurek. 2018. "Processing of Missing Data by Neural Networks". In *Advances in Neural Information Processing Systems*, 2719–2729.

Smirnova, E., E. Dohmatob, and J. Mary. 2019. "Distributionally Robust Reinforcement Learning". *arXiv preprint arXiv:1902.08708*.

Tarim, S. A., and B. G. Kingsman. 2004. "The Stochastic Dynamic Production/Inventory Lot-Sizing Problem with Service-Level Constraints". *International Journal of Production Economics* 88(1):105–119.

Taylor, B. 2006. "Developing Portfolio Optimization Models". *The MathWorks News & Notes*:30–32.

Wu, H.-H., and S. Küçükyavuz. 2018. "A Two-Stage Stochastic Programming Approach for Influence Maximization in Social Networks". *Computational Optimization and Applications* 69(3):563–595.

Yoon, J., J. Jordon, and M. Schaar. 2018. "Gain: Missing Data Imputation Using Generative Adversarial Nets". In *International Conference on Machine Learning*, 5689–5698. PMLR.

Zhao, C., and Y. Guan. 2018. "Data-Driven Risk-Averse Stochastic Optimization with Wasserstein Metric". *Operations Research Letters* 46(2):262–267.

AUTHOR BIOGRAPHIES

KE REN is a Ph.D. candidate at the Department of Industrial Engineering at the University of Pittsburgh. He earned his Master's degree in Electrical and Electronics Engineering from University of Rochester. His research interests include machine learning, optimization, artificial intelligence, and supply chain management. His email address is KER102@pitt.edu.

HODA BIDKHORI is an assistant professor at the Department of Industrial Engineering at the University of Pittsburgh. She earned her Ph.D. in Applied Mathematics from the Massachusetts Institute of Technology (MIT), where she subsequently spent several years as a postdoctoral researcher and lecturer in Operations Research and Statistics. Her current research focuses on the theory and applications of data-driven decision-making. Her e-mail address is bidkhori@pitt.edu. Her website is https://sites.google.com/view/hodabidkhori/.