A Study of Data-driven Models with Challenges Arising from OR Applications

by

Ke Ren

B.S. in Physics and Applied Physics, University of Science and Technology of China, 2014

M.S. in Electrical and Computer Engineering, University of Rochester, 2016

Submitted to the Graduate Faculty of
the Swanson School of Engineering in partial fulfillment
of the requirements for the degree of

Doctor of Philosophy

University of Pittsburgh

2021

UNIVERSITY OF PITTSBURGH

SWANSON SCHOOL OF ENGINEERING

This dissertation was presented

by

Ke Ren

It was defended on

July 14 2021

and approved by

Hoda Bidkhori, Ph.D., Assistant Professor, Department of Industrial Engineering

Bo Zeng, Ph.D., Associate Professor, Department of Industrial Engineering

Jayant Rajgopal, Ph.D., Professor, Department of Industrial Engineering

Zuo-Jun Shen, Ph.D., Professor, Department of Industrial Engineering and Operations Research, University of California, Berkeley

Dissertation Advisors: Hoda Bidkhori, Ph.D., Assistant Professor, Department of Industrial Engineering,

Bo Zeng, Ph.D., Associate Professor, Department of Industrial Engineering

A Study of Data-driven Models with Challenges Arising from OR Applications

Ke Ren, PhD

University of Pittsburgh, 2021

Data-driven models have been widely adopted in solving operations research (OR) problems, especially those from real applications where the distributions of the random variables are unknown. Note that these models are mainly inspired by methods from statistics and machine learning communities. However, many features in the OR problems place additional challenges in formulating and modeling. These challenges make models that are directly borrowed from other communities less efficient or even invalid. In this dissertation, we examine four typical OR problems, where they face challenges arising from the small data, complex objective function, incomplete data, and nonstationary data, respectively.

The first problem is chance-constrained programming under a small-data regime. We propose one upper bound on the performance of the commonly used scenario approach. To address the poor performance implied by this upper bound, we propose a new model with better performance. This model demonstrates a clear physical interpretation and a simple linear/conic formulation. Moreover, it is shown to be equivalent to distributionally robust chance-constrained programming under a specific setting. The second problem is maximum weight cycle and chain packing with inhomogeneous edge existence uncertainty. We fill a major gap observed in prior studies by proposing the first scalable model to solve this problem. The proposed model is a mixed-integer linear program, which can be solved directly by a general-purpose integer programming solver. The third problem studied is distributionally robust optimization (DRO) with incomplete joint data. We develop a new DRO framework with incomplete data sets. It presents an integrated framework to jointly analyze missing data and stochastic decision-making, which enables us to derive theoretical guarantees on the performance of stochastic programming under incomplete data. Several kinds of ambiguity sets are also discussed. Finally, we examined an inventory problem with highly unpredictable nonstationary demand. The demand is considered nonstationary and assumed that future demand cannot be reliably predicted through historical features or data. Man-

agers can only make decisions based on sequentially observed demand in an online fashion. We propose methods based on the idea of distinguishing the stochasticity/randomness and demand distribution changes among the sequentially observed demand.

Keywords: data-driven decision-making, distributionally robust optimization, stochastic programming, machine learning.

Table of Contents

List of Tables

List of Figures

1.0 Introduction

Data analytics has been introduced in operations research and has achieved a great success in supply chain management [2, 56], revenue management [50, 68], and healthcare management [102, 7]. Compared to the problems in the machine learning community, which signifies the most prosperous field for data-driven models, the problems in operations research (OR) face unique challenges. These challenges prevent decision-makers from direct extending existing data-driven models from the machine learning community to OR problems. Therefore, it is of great interest and necessity to develop novel data-driven OR models to overcome these challenges. In this study, we will explore four common challenges encountered in data-driven models of OR. We develop models as well as efficient algorithms to overcome these challenges. Concrete applications from supply chain management and healthcare management are utilized to validate the proposed approaches.

We discuss challenges arising from four categories, which are challenges brought by the small data, complex objective function, incomplete data, and non-stationary data. For these topics, we study the following four specific problems. The first problem is data-driven chance constrained programming, where the number of data is very limited. The key idea of data-driven chance constrained programming is to extract relevant information of the random variable from its available data and solve the optimization problems correspondingly. Computational experience suggests that with well-chosen methods and enough observed data, data-driven chance constrained programming yields tractable optimization problems whose solutions achieve a good balance between performances and risks. However, in many real-world applications, the number of available data is very limited. For example, some real system naturally produces a small number of data because the decision-relevant events are rare but high-impact, like supply chain disruptions due to earthquakes or hurricanes. Secondly, many real-world systems are usually non-stationary, indicating that they change their underlying distributions before they generate sufficient data. This is often observed as seasonal changes in many fields like portfolio optimization, supply chain design, etc. Thirdly, decision-makers prefer to and sometimes have to take early and proactive decisions before a

large amount of data are available. All these factors have motivated a shift of thinking away from big data towards a small-data paradigm. We show that the widely used sample average approximation (SAA) and scenario approach (SA) can be very problematic when the data number is small. We propose a novel framework that greatly improves the performances.

The second problem is a data-driven kidney exchange problem, where the unknown random variable affects the topology of a network. More specifically, the unknown random variables represent whether arcs in a graph exist or not. This behavior results in the complex objective function, which prevents decision-makers from obtaining well-structured convex optimization problems. Researchers formulated this problem as a maximum weight cycle and chain packing problem in former studies. However, they all fail to obtain tractable models. The existing models depend on heuristics to enumerate all possible cases to solve this problem. We formulate a relevant non-convex optimization problem and propose a tractable mixed-integer linear programming reformulation to solve it. In addition, we propose a model that integrates both risks and the expected utilities of the matching by incorporating conditional value at risk (CVaR) into the objective function, providing a robust formulation for this problem. Subsequently, we propose an SAA based approach to solve this problem. We test our approaches on data from the United Network for Organ Sharing (UNOS) and compare them against state-of-the-art approaches. Our model provides better performance with the same running time as a leading deterministic approach.

For the third problem, we focus on one state-of-the-art methodology, i.e., distributionally robust optimization, where the available data is incomplete. The motivation comes from the fact that the state-of-the-art approaches solve the missing data problem and stochastic programming separately. This leads to heuristic approaches with no theoretical guarantees. Regardless of the rigorousness of this conventional strategy, it overlooks one critical fact of an optimization model: its optimal solution is often very sensitive to parameters. Hence, instead of depending on enormous data and sophisticated statistical methods to ensure accurate estimations, a distributionally robust optimization (DRO) framework that simultaneously tackles the missing data problem and data-driven stochastic optimization is proposed. Existing DRO methods all require the data to be complete. We propose ambiguity sets directly based on the incomplete data set and prove the statistical consistency and

finite sample guarantees of the corresponding models.

Finally, we focus on the non-stationary data. The real-world data are rarely stationary in OR applications. In this research, we study one inventory control problem, where the stochastic demand is not only non-stationary, but also cannot be reliably predicted due to some unprecedented situations. It follows different distributions in different periods with unknown transition properties. Although this problem is encountered by many companies in practice, existing literature rarely studies it due to the complexity involved. We view this problem from a data science perspective and propose new data-driven frameworks. The proposed methods include a parametric approach called Integrated-Bayesian (IB) and a non-parametric approach called separate-lasso (SL). Both methods are theoretically analyzed and empirically benchmarked against several state-of-the-art heuristics in different data environments. We show that our methods outperform existing methods by identifying weather observed changes in the daily demand are caused by stochasticity/randomness or demand season/distribution changes. Because the optimal policy is unobtainable, we defined a relaxed setting by assuming the demand knowledge in advance. Then, we derive a policy, OPT, based on the literature. The empirical results reveal that the cost of the proposed approaches is only 12% higher than that of OPT on average. Furthermore, we compare the proposed approaches with state-of-the-art heuristics and also apply them to real-world demand data from one of the world's largest e-commerce websites.

In the next chapter, we review the relevant literature. In Chapter 3, we study data-driven chance constrained programming under a small data regime. In Chapter 4, we study a data-driven kidney exchange problem. In Chapter 5, we study distributionally robust optimization with incomplete data. Finally, in Chapter 6, we study inventory management with highly unpredictable nonstationary demand.

2.0 Literature review

This research covers a variety of data-driven problems studied in OR. Below, we review their related works starting from general data-driven optimization. We then discuss two specific data-driven scenarios, i.e., chance constrained programming and optimization with incomplete data. Finally, we review two applications, including inventory management under nonstationary demand and kidney exchange under failures.

2.1 Data-driven Optimization Under Uncertainty

Data-driven optimization has gained much attention recently because real-world data distributions are usually unknown. Sample average approximation [114, 72] approximates the unknown distribution through empirical distributions. If the data samples of true unknown distributions can be generated efficiently, the optimal solution of SAA converges to the true optimal with the probability 1. Stochastic approximation (SA) [94] is a similar approach comparing to SAA and is also based on Monte Carlo sampling. It iteratively performs sub-gradient descent based on the observed samples to approach the true optimal solutions.

The min-max learning frameworks have gain popularity recently as they produce robust solutions that potentially improve out-of-sample performances. These frameworks appear in data-driven robust optimization [14] and distributionally robust optimization [37, 47] problems. In these works, the ambiguity sets are constructed from the available data based on different ways, for example, moment-based set [37], confidence-region-based set [13], metric-based set [47, 70, 107] and others [131, 33]. These ideas are also extended and explored in two-stage stochastic decision-making schemes [64, 136, 71, 32].

2.2 Optimization Under Missing Data

Missing or incomplete data has been studied widely, especially in machine learning and statistics [108, 55, 61] literature. One of the most natural options to solve the missing data problem is to discard any data that include missing values. However, this approach may lead to biased results [83] or result in the loss of information.

For incorporating incomplete data into the decision-making process, the two most common and well-developed methods are data imputation and maximum likelihood-based approaches. The data imputation approaches recover the incomplete data set by estimating missing values based on the observed data. The criteria are often developed based on the decision trees [122], support vector regression [130], neural networks [96], etc. The maximum likelihood-based approaches ideally work for parametric models with missing but relatively complete data. These techniques often aim to find a distribution that maximizes the observed-data likelihood [82] or seeks EM algorithms [38] because computational issues concerning non-convex optimization may arise [43].

The statistical approaches developed above solely focus on incomplete data sets. When these approaches are combined with stochastic optimization models, the performances are unclear. In this research, we present a new model based on distributionally robust optimization. Instead of first estimating the unknown distribution from the incomplete data set, we propose integrated models to address the missing data issue and to find optimal decisions simultaneously. This allows our model to guarantee out-of-sample performances theoretically.

2.3 Data-driven Chance Constrained Programming

There are three state-of-the-art data-driven approaches, i.e., scenario approach, sample average approximation, and distributionally robust chance constrained programming. We next briefly summarize their respective features and weakness, which substantiates our argument that new modeling and associated computational tools should be developed. The

scenario approach [24, 25, 27] and sample approximation approach [86, 98] represent the two most widely studied data-driven methods. Scenario approach requires all samples to satisfy the chance constraints, which yields simple and tractable mathematical formulations. However, it relies on sampling enough data to generate high-performance solutions, indicating that its performance may not be acceptable under a small-data case. Indeed, we derive in this paper a theoretical upper bound on the probability of the solution derived by the scenario approach that satisfies the original chance constraints for a given set of samples, the first one in the literature to the best of our knowledge. For example, when the number of samples is less than $\frac{1}{\epsilon}$ with the violation probability ϵ less than 0.1, this upper bound on the probability of the chance constraint is satisfied is less than 0.66. It, therefore, indicates that with more than one-third of the cases, the original chance constraint fails to hold, which definitely is not desired in practice. Regarding SAA, it is a method similar to scenario approach, which actually requires a subset of the samples, instead of all of them, to satisfy the chance constraint. Hence, the probability of its solution satisfying the original chance constraints is naturally upper bounded by that of scenario approach.

Compared to these two approaches, *distributionally robust chance constrained program* (DRCCP) is a recent approach in the literature. We note that it has a much better performance for small-data cases. It first constructs a parameterized ambiguity set; then, it derives a solution according to its performance in the worst-case distributions in that ambiguity set. The state-of-the-art DRCCP uses the Wasserstein balls to define ambiguity sets. By selecting appropriate radiuses of balls, the out-of-sample performance is improved compared to SAA. However, DRCCP suffers from complicated and even intractable formulations, which makes them difficult to implement in practice. Although recent progress in the literature identifies some specific settings for which tractable reformulations can be obtained, these conditions are way too restricted for many practical problems. In [134, 31], single and joint chance constraints with linear uncertainties in the right-hand or left-hand side are studied. They show that when the support of the random variable is continuous and unbounded, tractable reformulations can be obtained. Compared to these two works, [69] also considers the cases with known and discrete support.

2.4 Kidney Exchange Under Failures

We here review the basic concepts of kidney exchange problem. Kidney exchange is a centralized barter market were patients with end-stage renal disease trade willing donors in cyclic or chain-like transactions [101, 105, 1]. The aim of the kidney exchange clearinghouse is to find the "best" disjoint set of such swaps—i.e., to solve a cycle and chain packing problem. Exchanges already account for over 12% of living kidney donations in the US, and exchange programs are growing worldwide [17]—including via extensions to liver and lung [48], and even multi-organ [42], exchange. Fielded exchanges face several source of inefficiency, primarily due to pre-transplant "failure" [76]; that is, most *planned* transplants never result in transplantation due to medical or logistical incompatability [41, 57, 5, 4, 39, 59, 88, 73, 3]. In other words, the exchange program cannot be certain whether a compatible patient and donor will result in a transplant. Exchanges are often represented by directed graphs (see Section 4.2), where edges indicate potential transplants and edge weights reflect the medical or social utility of the transplant. If a planned transplant (i.e., *edge*) fails, its effects can cascade through the exchange, causing other edges to fail (see Section **??**)—thus, edge failures can severely impact the overall utility of an exchange.

2.5 Data-driven Inventory Management.

Recently, data-driven models have attracted significant attention in the field of inventory management. Here, we review these studies based on the techniques adopted for handling data. SAA is the most natural nonparametric data-driven approach [78, 77], which uses the empirical probability density function to solve the original problem by optimizing the average objective values over all available data. Adaptive learning frameworks [23, 74, 67, 117] have been extensively studied in inventory management and are mostly based on stochastic approximation frameworks. These methods include information of each piece of observed data by using gradient descent-based steps. Novel approximation methods are often combined to alleviate the assumptions or improve performance [58, 66].

Other data-driven frameworks have also been developed to solve different inventory problems. For example, the operational statistics proposed in [84] are used to solve the newsvendor problem and is a parametric approach which assumes that the demand distribution comes from a family of distributions parameterized by some unknown parameters. It finds these unknown parameters by combining the optimization and estimation steps to obtain better results than the statistics derived from separate analyses. [16] proposed forecasting-based frameworks where they recommended two data-driven frameworks comprising a regression model and a linear optimization model to set safety stock levels in the newsvendor problem by incorporating other external factors when forecasting a demand. [109] introduced the feature-based newsvendor problem and studied it from a machine learning perspective. Two approaches were proposed and compared for cases with small and big data, which derived tight generalization error bounds on the expected out-of-sample cost.

Inventory management under nonstationary demand. We classify existing studies into three categories according to the methods they adopt for handling nonstationary demand. In the first category, the nonstationary demand is explicitly defined, implying that the demand mechanism is assumed to be known to decision-makers. Studies in this category include the nonstationary stochastic lot-sizing problem [6, 133] and those conducted on inventory management under a Markov-modulated demand process [121, 112, 93]. In the second category, the mechanism of the nonstationary demand is assumed to be unknown to the decision-makers, but some well-conditioned forecasts are available. These studies include [62, 30, 9, 36]. The forecasts are either obtained through forecasting methods based on certain demand mechanisms [62] or are assumed to be given exogenously [9]. In the final category, simulation-based approaches are used to study the nonstationary demand based on a user-specified demand process or real-world data. For example, [90] used a simulation model to study the seasonal demand defined by a periodic function, whereas [36] proposed a forecast-based procedure for order-up-to-level policies and demonstrated its superiority based on simulations performed using real-world data.

3.0 Data-driven Chance-constrained Programming Under Small-data Regime

In this study, we first theoretically analyze the performances of the two most widely studied data-driven methods for chance-constrained programming, scenario approach [24], and sample approximation approach [86], under a small-data setting. The results show that the upper bound of the probability that these two methods can guarantee the original chance constraints is less than 0.66. This means that in more than one-third of the cases, the original chance constraints fail to hold. To improve their out-of-sample performances, we propose a new model with closed-form linear/conic formulations. This new model introduces a set of parameters, and it is reduced to the model given by the scenario approach and sample approximation approaches when all these parameters are set to zero. When the parameters are larger than zero, our model improves the probability that the original chance constraints are satisfied. Furthermore, when these parameters take some special forms, our model is also shown to be equivalent to *distributionally robust chance constrained programs* (DRCCPs) whose ambiguity sets are Wasserstein balls. Therefore, our model links the scenario approach, sample approximation approach, and DRCCPs. And our model provides a much simpler DRCCP formulation in this small-data case compared to the existing two tractable reformulations [134, 31].

We formulate our problem in (3.1).

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \mathrm{P}\left\{\mathbf{x}\in\mathcal{X}'(\boldsymbol{\xi})\right\}\geqslant 1-\epsilon, \\ & \mathcal{X}'(\boldsymbol{\xi})=\left\{\mathbf{x}=[\mathbf{x}_1,\mathbf{x}_2]\middle|(\mathbf{A}_k\boldsymbol{\xi}+\mathbf{a}_k)^T\mathbf{x}_1+\mathbf{d}_k^T\mathbf{x}_2\leqslant\mathbf{b}_k^T\boldsymbol{\xi},\forall k\in[K]\right\}. \end{aligned} \tag{3.1}$$

Specifically, the decision variables are $\mathbf{x}=[\mathbf{x}_1,\mathbf{x}_2]\in\mathbb{R}^m$ with $\mathbf{x}_1\in\mathbb{R}^{m_1}$ and $\mathbf{x}_2\in\mathbb{R}^{m_2}$ $(m_1+m_2=m)$ in Optimization (3.1). The random variable is assumed to be continuous and is denoted as $\boldsymbol{\xi}\in\mathbb{R}^n$, whose distribution is unknown but N independent observations, $\{\hat{\boldsymbol{\xi}}_1,\cdots,\hat{\boldsymbol{\xi}}_N\}$, are available. Parameter K is larger than or equal to 1, and we use $[K]$ to denote the set $\{1,...,K\}$. Parameters $\mathbf{A}_k\in\mathbb{R}^{m_1\times n}$, $\mathbf{a}_k\in\mathbb{R}^{m_1}$, $\mathbf{d}_k\in\mathbb{R}^{m_2}$, $\mathbf{b}_k\in\mathbb{R}^n$, and $\epsilon\in(0,1)$ are deterministic. Intuitively, (3.1) says the inequalities $(\mathbf{A}_k\boldsymbol{\xi}+\mathbf{a}_k)^T\mathbf{x}_1+\mathbf{d}_k^T\mathbf{x}_2\leqslant\mathbf{b}_k^T\boldsymbol{\xi}$

can only be violated with a probability up to ϵ. Therefore, in the rest of the study, we call ϵ the *violation probability*. We restrict the number of available observations, N, to be less than $\frac{1}{\epsilon}$ indicating a small-data regime.

Former studies on solving Optimization (3.1) can be mainly categorized into three types, i.e. scenario approach, sample approximation approach, and DRCCP, based on their methodologies. *Scenario approach* [24, 27, 26] solves the chance constraints by requiring the corresponding constraints to hold for every observed sample. That is, it solves (3.1) by replacing the chance constraint with

$$(\mathbf{A}_k\hat{\boldsymbol{\xi}}_i + \mathbf{a}_k)^T\mathbf{x}_1 + \mathbf{d}_k^T\mathbf{x}_2 \leqslant \mathbf{b}_k^T\hat{\boldsymbol{\xi}}_i, \forall k \in [K], \forall i \in [N], \tag{3.2}$$

where we use $[N]$ to denote the set $\{1, \cdots, N\}$. Under the small-data regime, the confidence level of using (3.2) can be very low. This can be evidenced from the *confidence parameter* defined in [24], where by Theorem 1 in [24], the probability that the optimal solution of scenario approach with a linear objective function satisfies the chance constraints is no smaller than $1 - \frac{m}{\epsilon(N+1)}$ (we refer to this probability as confidence level in the rest of the study). However, when $N \leqslant \frac{1}{\epsilon}$, we obtain $1 - \frac{m}{\epsilon(N+1)} \leqslant 1 - \frac{m}{2}$, which is less than zero when the decision variable dimension, m, is larger than 2. Therefore, no theoretical guarantee can be obtained using this theorem if the data size is small. In this study, we further validate this point by providing a theoretical analysis with respect to the confidence level. The result shows that with a probability of at least 0.34, the scenario approach fails to guarantee the original chance constraints when the violation probability satisfies $0 < \epsilon < 0.1$. This means in more than one-third of the cases, the scenario approach fails to guarantee the desired chance constraints. The same problem of having a low confidence level also applies for *sample approximation approach* [86, 87, 28] because sample approximation requires the constraints to hold for a subset of the observed samples. When the number of the data is less than $\frac{1}{\epsilon}$, the sample approximation approach often requires at least $(N - N\epsilon)$ data to hold, where $N - N\epsilon > N - 1$. Therefore, it also solves (3.1) based on (3.2). In order to overcome their drawbacks, we propose a new way for solving data-driven chance constraints

by replacing the chance constraint with (3.3).

$$\mathbf{b}_k^T\hat{\boldsymbol{\xi}}_i - (\mathbf{A}_k\hat{\boldsymbol{\xi}}_i + \mathbf{a}_k)^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant c_k\|\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k\|_p,\ (\|\|_p\text{: p-norm}), \forall i \in [N], \forall k \in [K], \tag{3.3}$$

where $c_k, \forall k \in [K]$ are some non-negative numbers. Formulations (3.3) are derived by requiring a neighborhood of each data point to satisfy the corresponding constraints, where details will be given in Section 3.1. Intuitively, (3.3) improves the confidence level by reducing the feasible regions given by the scenario approach and the sample approximation approach. We further prove that when c_k are chosen as follows,

$$c_k = \frac{\theta}{\epsilon}, \forall k \in [K], \tag{3.4}$$

Optimization (3.1) with chance constraints replaced by (3.3) is equivalent to *distributionally robust chance constrained programs* (DRCCPs) [70, 134, 31] whose ambiguity sets are a Wasserstein ball with radius θ, where the radius is measured through $\frac{p}{p-1}$-norm. This result not only provides one theoretical way of tuning our parameters c_k through the concentration properties of Wasserstein metric [52], but also links the scenario approach, sample approximation approach, and DRCCPs in the small-data regime, where all above approaches can be viewed as some specific choices of parameter c_k in our proposed model.

Tractable formulations of DRCCPs with Wasserstein ambiguity sets have been studied recently in [134, 31]. They both use big-M coefficients to obtain mixed-integer reformulation to solve this problem. And they both require the radius, θ, to be strictly larger than zero to ensure correctness. Compared to their formulations, our model (3.3) is big-M free and allows the parameter θ to be zero exactly. Therefore, our model brings benefits comparing to the DRCCPs. More specifically, our model is easier for analytical analyses because of its simple closed-form formulations of the feasible region. And our model is consistent with the scenario/sample approximation approach by allowing the corresponding radius of the Wasserstein ball in DRCCP to be exactly zero. Our model is more computationally robust/stable comparing to the existing DRCCP formulations based on big-M parameters.

We conclude our main results as follows.

1. We theoretically analyze the performance of the scenario/sample approximation approach for the data-driven chance constrained programming with a small-data regime. We show that the upper bound of the probability that the original chance constraints can be guaranteed is less than 0.66.
2. We propose a linear/conic program, as shown in (3.5), to solve the chance constrained programming under the small-data regime. The resulting optimizations are mathematically tractable and have simple closed-form formulations.

$$\begin{aligned} \min \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \mathbf{b}_k^T\hat{\boldsymbol{\xi}}_i - (\mathbf{A}_k\hat{\boldsymbol{\xi}}_i + \mathbf{a}_k)^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant c_k\|\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k\|_p, \forall i \in [N], \forall k \in [K]. \end{aligned} \tag{3.5}$$

   We show that this model has better out-of-sample performances than the scenario/sample approximation approach when $c_k > 0$.
3. We prove that our model (3.5) is also equivalent to distributionally robust chance constrainted programmings (DRCCPs) under a Wasserstein ball with radius θ when $c_k = \frac{\theta}{\epsilon}$. In addition, we show that our reformulations outperform the existing tractable formulations of DRCCPs in two aspects. First, our model is big-M free and has simpler closed-form formulations. Second, our formulations are consistent with scenario approach when θ is equal to 0. Two existing formulations of DRCCPs always assume $\theta > 0$ and thus have numerical problems in implementations when θ is very small.
4. We validate our approaches through computational studies. We prove that our approaches outperform the scenario approaches and sample approximation approaches under a small-data regime, and are easier and more robust in implementations compared to existing DRCCPs.

3.1 Confidence Level for Data-driven Chance Constrained Programming under Small-data Regime

In this section, we first theoretically analyze the performances of the scenario approach and the sample approximation approach under the small-data regime. Then, we propose our model for data-driven chance constrained programming based on the former analyses.

3.1.1 Assumptions and Notations

We assume N independent samples of $\boldsymbol{\xi}$ are available, and they are $\{\hat{\boldsymbol{\xi}}_1, \cdots, \hat{\boldsymbol{\xi}}_N\}$. We denote (3.6) as OPT1 for simplicity in the rest of the study. When $K = 1$, OPT1 signifies the scenario/sample approximation approach for the individual chance constrained programming. On the other hand, when $K > 1$, OPT1 represents the joint chance constrained programming.

$$
\begin{aligned}
(OPT1) \quad \min \quad & \mathbf{c}^T\mathbf{x} && \text{(3.6a)}\\
s.t. \quad & \mathbf{b}_k^T\hat{\boldsymbol{\xi}}_i - (\mathbf{A}_k\hat{\boldsymbol{\xi}}_i + \mathbf{a}_k)^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant 0, \forall i \in [N], \forall k \in [K], && \text{(3.6b)}\\
& \mathbf{x} \in \mathcal{X}. && \text{(3.6c)}
\end{aligned}
$$

For simplicity and clarity, we define *ϵ-level solution* following the definition used in [24].

Definition 1 (ϵ-level solution)**.** *Let $\mathbf{x} \in \mathcal{X}$ be a candidate solution for (3.1); the support of $\boldsymbol{\xi}$ is $\boldsymbol{\Delta}$. Then, the violation probability of $\mathbf{x}$ is defined as*

$$V(\mathbf{x}) = P\{\boldsymbol{\xi} \in \boldsymbol{\Delta} \big| \mathbf{b}_k^T\boldsymbol{\xi} - (\mathbf{A}_k\boldsymbol{\xi} + \mathbf{a}_k)^T\mathbf{x} - \mathbf{d}_k^T\mathbf{x}_2 < 0, \exists k \in [K]\}.$$

Let $\epsilon \in [0, 1]$. We say that $\mathbf{x} \in \mathcal{X}$ is an ϵ-level robustly feasible solution if $V(\mathbf{x}) \leqslant \epsilon$.

Following the above definition, the goal of chance constrained programming is to devise an algorithm that returns an ϵ-level solution. In data-driven optimization (3.6), because $\hat{\boldsymbol{\xi}}_i, \forall i \in [N]$ are randomly selected, the optimal solution, which is denoted by $\mathbf{x}^*$, is a random variable that depends on multi-sample $\hat{\boldsymbol{\xi}}_1, \cdots, \hat{\boldsymbol{\xi}}_N$. Correspondingly, we define a *confidence level* β, which equals the probability that $\mathbf{x}^*$ is an ϵ-level solution. The confidence level defined here follows the same logic as the *confidence parameter* used in [24]. In the following, we define the conficdence level β mathematically.

Definition 2 (confidence level)**.** *Suppose the optimal solution of Optimization (3.6) based on some random samples is* $\mathbf{x}^*$*. We define the confidence level of* $\mathbf{x}^*$ *as* $\beta \in [0,1]$*, which equals the probability that* $\mathbf{x}^*$ *is an* ϵ*-level solution (or equivalently, the chance constraint in (3.1) actually holds for* $\mathbf{x}^*$*), i.e.*

$$\beta = \mathbb{E}_{\mathbf{x}^*}\left\{\mathbf{1}\left[P\{\mathbf{x}^* \in \mathcal{X}'(\boldsymbol{\xi})\} \geqslant 1-\epsilon\right]\right\} = \mathbb{E}_{\mathbf{x}^*}\left\{\mathbf{1}[V(\mathbf{x}^*) < \epsilon]\right\},$$
$$\mathcal{X}'(\boldsymbol{\xi}) = \left\{\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \middle| (\mathbf{A}_k\boldsymbol{\xi} + \mathbf{a}_k)^T\mathbf{x}_1 + \mathbf{d}_k^T\mathbf{x}_2 \leqslant \mathbf{b}_k^T\boldsymbol{\xi}, \forall k \in [K]\right\}.$$

The expectation is taken with respect to the random variable $\mathbf{x}^*$*. Furthermore, because the randomness of* $\mathbf{x}^*$ *comes from the multi-sample* $\{\hat{\boldsymbol{\xi}}_1, \cdots, \hat{\boldsymbol{\xi}}_N\}$*, we can also formulate* β *as*

$$\beta = \mathbb{E}_{\hat{\boldsymbol{\xi}}_1,\cdots,\hat{\boldsymbol{\xi}}_N}\left\{\mathbf{1}\left[P\{\mathbf{x}^* \in \mathcal{X}'(\boldsymbol{\xi})\} \geqslant 1-\epsilon\right]\right\}.$$

In the following, we establish the results for the performances of the scenario approach and the sample approximation approach using the above defined confidence level β. Clearly, the confidence level of a solution $\mathbf{x}^*$ is affected by both constraints (3.6b) and (3.6c). Because we make no assumptions for the feasible set $\mathcal{X}$, the optimal solution $\mathbf{x}^*$ may be purely determined by $\mathcal{X}$ in some cases. Under such cases, the confidence level β is also purely determined by $\mathcal{X}$. Therefore, to exclude these pathological cases in our study, we make one following assumption.

Assumption 1. *The optimal solution of OPT1 is not purely determined by the objective function and the set* $\mathcal{X}$*. This means with probability* 1 *there is at least one active constraint (meaning the equality sign holds) in (3.6b) when the optimal solution* $\mathbf{x}^*$ *is achieved.*

This assumption does not limit the generality of our results much because if there is no active constraints in (3.6b), deriving β is impossible unless additional knowledge or assumptions on $\mathcal{X}$ are made. We will briefly discuss the cases where Assumption 1 does not hold at the end of this section.

3.1.2 Performances of Scenario/Sample Approximation Approach

We first introduce two results revealing the performances of scenario/sample approximation approach, i.e. OPT1, in the following Theorem 1.

Theorem 1. *The confidence level of an optimal solution in OPT1 based on N independent samples is at most $1-(1-\epsilon)^N$.*

Proof. We use $\mathbf{x}_N^*$ to denote the optimal solution based on N random samples. Recall that we define the violation probability of $\mathbf{x}_N^*$ as

$$V(\mathbf{x}_N^*) = \mathrm{P}\left\{\mathbf{x}_N^* \notin \mathcal{X}'(\boldsymbol{\xi})\right\}.$$

Then, following the definition of the confidence level we have

$$\beta = \mathrm{P}(V(\mathbf{x}_N^*) \leqslant \epsilon).$$

Each sample $\hat{\boldsymbol{\xi}}_i$ $(1 \leqslant i \leqslant N)$ brings a set of constraints, i.e.,

$$\mathbf{b}_k^T\hat{\boldsymbol{\xi}}_i + b_{k0} - (\mathbf{A}_k\hat{\boldsymbol{\xi}}_i + \mathbf{a}_k)^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant 0, \forall k \in [K]. \tag{3.7}$$

WLOG, according to Assumption 1, we assume that with probability 1, the k^*-th constraint brings active constraint, i.e.,

$$\mathbf{b}_{k^*}^T\hat{\boldsymbol{\xi}}_i + b_{k^*0} - (\mathbf{A}_{k^*}\hat{\boldsymbol{\xi}}_i + \mathbf{a}_{k^*})^T\mathbf{x}_1 - \mathbf{d}_{k^*}^T\mathbf{x}_2 = 0, \exists i \in [N]. \tag{3.8}$$

Here we want to point out that with probability one, only one sample achieves equality sign no matter how large the N is. This can be seen by the fact that the probability measure of all the ξ that satisfy the constraint 3.8 is zero (because variable ξ lose one degree of freedom), i.e.,

$$\mathrm{P}(\boldsymbol{\xi}\big|\mathbf{b}_{k^*}^T\hat{\boldsymbol{\xi}} + b_{k^*0} - (\mathbf{A}_{k^*}\hat{\boldsymbol{\xi}} + \mathbf{a}_{k^*})^T\mathbf{x}_1 - \mathbf{d}_{k^*}^T\mathbf{x}_2 = 0) = 0$$

We define $V^1(\mathbf{x}_N^*)$ as the violation probability of the constraint k^*. Because $V(\mathbf{x}_N^*)$ represents the violation probability of all constraints, we have

$$\beta = \mathrm{P}(V(\mathbf{x}_N^*) \leqslant \epsilon) \leqslant \mathrm{P}\big(V^1(\mathbf{x}_N^*) \leqslant \epsilon\big)$$

where $V^1(\mathbf{x}_N^*) = \mathrm{P}(\mathbf{b}_{k^*}^T\hat{\boldsymbol{\xi}} + b_{k^*0} - (\mathbf{A}_{k^*}\hat{\boldsymbol{\xi}} + \mathbf{a}_{k^*})^T\mathbf{x}_{1N}^* - \mathbf{d}_{k^*}^T\mathbf{x}_{2N}^* < 0)$. Next, we quantify the value of the above-defined violation probability $V^1(\mathbf{x}_N^*)$. We define $F(\alpha)$ as the distribution (cumulative distribution function) of the violation probability of $V^1(\mathbf{x}^*)$.

$$F(\alpha) = \mathrm{P}(V^1(\mathbf{x}^*) \leqslant \alpha) \tag{3.9}$$

Then, we consider the sample space generated by N samples, $\hat{\boldsymbol{\xi}}_1, \cdots, \hat{\boldsymbol{\xi}}_N$.

Because we know only one sample contains active constraint, we partition the sample space into N sets based on the index of the sample containing active constraint. We denote these sets as $\mathcal{S}_j$ $(1 \leqslant j \leqslant N)$. The probability (measure) of each $\mathcal{S}_j$ is the same and can be formulated as

$$\mathrm{P}(\mathcal{S}_j) = \int_0^1 (1-\alpha)^{N-1} F(d\alpha). \tag{3.10}$$

This is because, without loss of generality, $\mathrm{P}(\mathcal{S}_j)$ equals the probability that only the first sample contains active constraints (we denote this probability as $\mathrm{P}(\mathcal{S}_1)$).

$$\mathrm{P}(\mathcal{S}_j) = \mathrm{P}(\mathcal{S}_1) = \mathrm{P}(\text{ the first sample contains active constraints }) \tag{3.11}$$

$$= \mathrm{P}(\text{ last } N-1 \text{ samples do not violate the constraint }). \tag{3.12}$$

Additionally, if the violation probability is α. Then, for each of the last $N-1$ samples, the probability that it does not violate the constraint is $1-\alpha$.

$$\begin{aligned} &\mathrm{P}(\text{ last } N-1 \text{ samples do not violate the constraint }) \\ &= \int_0^1 (1-\alpha)^{N-1} F(d\alpha). \end{aligned}$$

Because the sample space is partitioned into N sets with equal size, we get

$$N\int_0^1 (1-\alpha)^{N-1} F(d\alpha) = 1. \tag{3.13}$$

The distribution function $F(\alpha)$ has the unique solution in (3.13) based on the results in [27], which is

$$F(\alpha) = \alpha.$$

Finally, the probability $\mathrm{P}(V^1(\mathbf{x}_N^*) \leqslant \epsilon)$ is calculated accordingly as follows.

$$\mathrm{P}(V^1(\mathbf{x}_N^*) \leqslant \epsilon) = N \int_0^{\epsilon} (1-\alpha)^{N-1} d(\alpha) = N \frac{-(1-\alpha)^N}{N} \Bigg|_0^{\epsilon} = 1 - (1-\epsilon)^N.$$

Therefore, we obtain

$$\beta \leqslant 1 - (1-\epsilon)^N.$$

□

By Theorem 1, the confidence level of the optimal solutions in scenario/sample approximation approach is upper bounded by $1-(1-\epsilon)^N$. The above results are enough for us to evaluate the performances of scenario/sample approximation approach as summarized in Proposition 1.

Proposition 1. *When* $N < \dfrac{n'}{\epsilon}$ $(n' > 0)$*, we obtain* $1-(1-\epsilon)^N < 1-(1-\epsilon)^{\frac{n'}{\epsilon}}$*. In addition,* $1-(1-\epsilon)^{\frac{n'}{\epsilon}}$ *is monotone increasing with respect to* ϵ*. When the violation probability is between* 0 *and* 0.1*, i.e.* $0 < \epsilon < 0.1$*, we have* $1-(\frac{1}{e})^{n'} < 1-(1-\epsilon)^{\frac{n'}{\epsilon}} < 1-0.349^{n'}$*. Specifically, when* $n' = 1$*, we obtain* $0.63 < 1-(1-\epsilon)^{\frac{1}{\epsilon}} < 0.66$.

Proposition 1 says the upper bound of the confidence level of the optimal solutions in scenario/sample approximation approach ranges from 0.63 and 0.66 when the violation probability is between 0 and 0.1 if $N < \dfrac{1}{\epsilon}$. Based on the definition of the confidence level, we know that under more than one-third of the cases, the scenario/sample approximation approach cannot guarantee the original chance constraints or, equivalently, obtain an ϵ-level solution. To overcome this problem, we propose a new model in the next subsection.

Finally, we briefly discuss the cases where there is no active constraints brought by the chance constraints in OPT1. If there are no active constraints, the optimal solution lies strictly inside the feasible region given by the chance constraints; therefore, the confidence level is greater than that of the cases where there are active constraints.

3.1.3 Our method

The low confidence of the optimal solutions of the scenario approach indicates that the feasible region in (3.2) is too large in most cases. To overcome this problem, we propose a new method in (3.14). Its main idea is to allow more $\boldsymbol{\xi}$ to satisfy the chance constraint by making the values around the observed data $\hat{\boldsymbol{\xi}}_i$, $i \in [N]$ to be inside the feasible region given by the optimal solution $\mathbf{x}^*$. We introduce the details in the following.

Our model is formulated in (3.14), which is denoted as OPT2 in the rest of the paper. The parameter c_i is a user-specified non-negative number.

$$
\begin{aligned}
(OPT2) \quad \min \quad & \mathbf{c}^T\mathbf{x} && \text{(3.14a)}\\
s.t. \quad & (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} && \text{(3.14b)}\\
& \geqslant c_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_p, \forall i \in [N], \forall k \in [K], && \text{(3.14c)}\\
& \mathbf{x} \in \mathcal{X}. && \text{(3.14d)}
\end{aligned}
$$

As just introduced at the beginning, the main idea of our approach is to allow the values around the observed data samples to satisfy the constraints (3.6b), i.e.,

$$(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\boldsymbol{\xi} - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \geqslant 0, \forall k \in [K],$$

simultaneously. Following this idea, we define vector $\boldsymbol{\xi}_i'$ as the difference between the observed i-th sample $\hat{\boldsymbol{\xi}}_i$ and one possible value of $\boldsymbol{\xi}$ around $\hat{\boldsymbol{\xi}}_i$. Therefore, our approach requires the following constraints to hold for all possible $\boldsymbol{\xi}_i'$.

$$(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T(\hat{\boldsymbol{\xi}}_i + \boldsymbol{\xi}_i') - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \geqslant 0, \forall i \in [N], \forall k \in [K]. \tag{3.15}$$

Clearly, the choice of $\boldsymbol{\xi}_i'$ depends on the *structure of the probability density function* and *the support* of $\boldsymbol{\xi}$, which is unknown in most general cases. Therefore, we simply restricts

$$\|\boldsymbol{\xi}_i'\|_q \leqslant c_i, \text{ where } c_i \geqslant 0.$$

Intuitively, this means we include all values inside a q-norm ball centered at $\hat{\boldsymbol{\xi}}_i$ with radius c_i, and this also means that the number of the candidate $\boldsymbol{\xi}_i'$ is infinity. However, we show that by restricting

$$(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\boldsymbol{\xi}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \geqslant c_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_p, \forall i \in [N], \forall k \in [K], \tag{3.16}$$

it is sufficient to guarantee that (3.15) holds for any $\boldsymbol{\xi}_i'$ defined above. The inequalities (3.15) are equivalent to

$$(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \geqslant (\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k)^T\boldsymbol{\xi}_i', \forall i \in [N], \forall k \in [K].$$

By Holder's inequality, we obtain

$$(\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k)^T\boldsymbol{\xi}_i' \leqslant \|\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k\|_p\|\boldsymbol{\xi}_i'\|_q, \text{ with } \frac{1}{p} + \frac{1}{q} = 1,$$

where the equality sign holds if and only if $|\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k|^p$ and $|\boldsymbol{\xi}_i'|^q$ are linearly dependent. Because the norm of $\boldsymbol{\xi}_i'$ is bounded, we have

$$(\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k)^T\boldsymbol{\xi}_i' \leqslant c_i\|\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k\|_p.$$

Therefore, by restricting (3.16) in OPT2, we proved

$$\begin{aligned}&(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\boldsymbol{\xi}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \geqslant c_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_p \\ &\geqslant (\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k)^T\boldsymbol{\xi}_i', \forall i \in [N], \forall k \in [K],\end{aligned}$$

for any $\boldsymbol{\xi}_i'$ defined above.

3.2 Relationships with DRCCPs

We explore the relationships between our model and DRCCPs. In Subsection 3.2.1, we show that OPT2 is equivalent to the DRCCPs under Wasserstein balls if $\{c_i, \forall i\}$ are chosen as the same fixed positive number when $\boldsymbol{\xi} \in \mathbb{R}^n$ and $N < \frac{1}{\epsilon}$.

3.2.1 Equivalence to DRCCP under a special case

We first present the main results for individual chance constraints, then we show it for joint chance constraints. Additionally, in this subsection, we focus on one special case where $N < \frac{1}{\epsilon}$ and $\boldsymbol{\xi} \in \mathbb{R}^n$. For clarity, we restate OPT2 here when $K = 1$.

$$
\begin{aligned}
(OPT2) \quad \min \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & (\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\boldsymbol{\xi}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} \geqslant c_i\|\mathbf{b} - \mathbf{A}^T\mathbf{x}_1\|_p, \forall i \in [N], \\
& \mathbf{x} \in \mathcal{X}.
\end{aligned}
$$

The main result is summarized in Theorem 2.

Theorem 2. *Suppose the number of observations, N, is less than $\frac{1}{\epsilon}$ ($\epsilon > 0$), and the coefficient of $\boldsymbol{\xi}$, $(\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)$, is not zero. By choosing $c_i = \frac{\theta}{\epsilon}$ ($\theta \geqslant 0$), OPT2 is equivalent to*

$$
\begin{aligned}
\min \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \min_{P \in \mathcal{B}_\theta(\hat{\mathbb{P}})} \mathbb{E}_P\left\{1[(\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\boldsymbol{\xi} - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} \geqslant 0]\right\} \geqslant 1 - \epsilon,
\end{aligned}
\tag{3.17}
$$

where the ambiguity set $\mathcal{B}_\theta(\hat{\mathbb{P}})$ is a Wasserstein ball with radius θ centered at $\hat{\mathbb{P}}$ representing the empirical distribution of $\boldsymbol{\xi}$.

Before proving Theorem 2, we first define the Wasserstein ball mathematically in Definition 3 and 4.

Definition 3. *For an arbitrary q-norm $\|\cdot\|_q$, define $M' = \{\mathbb{Q} : \mathbb{E}^{\mathbb{Q}}[\|\boldsymbol{\xi}\|] = \int_{\mathbb{R}^n} \|\boldsymbol{\xi}\|_q \mathbb{Q}(d\boldsymbol{\xi}) < \infty\}$. The Wasserstein distance between two distributions $\mathbb{Q}_1$, $\mathbb{Q}_2 \in M'$ is defined as*

$$
\begin{aligned}
W(\mathbb{Q}_1, \mathbb{Q}_2) := \inf\{(\int_{\mathbb{R}^{2n}} \|\boldsymbol{\xi}_1 - \boldsymbol{\xi}_2\|_q \Pi(d\boldsymbol{\xi}_1, d\boldsymbol{\xi}_2)) : \\
\Pi \text{ is a joint distribution of } \boldsymbol{\xi}_1 \text{ and } \boldsymbol{\xi}_2 \text{ with marginals } \mathbb{Q}_1 \text{ and } \mathbb{Q}_2, \text{ respectively}\}.
\end{aligned}
$$

Definition 4. *A Wasserstein ball centered at $\hat{\mathbb{P}}$ with radius θ is defined as:*

$$
\mathcal{B}_\theta(\hat{\mathbb{P}}) = \{\mathbb{H} \in M' : W(\mathbb{H}, \hat{\mathbb{P}}) \leqslant \theta\}. \tag{3.18}
$$

Optimization (3.17) represents the DRCCPs under the wasserstein distance, where its tractable reformulations have been studied in [134, 31]. Both of them obtain the big-M reformulation by assuming $\theta > 0$. In OPT2, we allow $\theta = 0$ exactly. For the ease of proof, we use the following Lemma 1 from [31].

Lemma 1. *Optimization (3.17) is equivalent to*

$$\begin{aligned} \min \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N \|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p \\ & (\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} + Mq_i \geqslant t + s_i, \forall i \in [N] \\ & M(1 - q_i) \geqslant t + s_i, \forall i \in [N] \\ & \mathbf{q} \in \{0,1\}^N, \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{3.19}$$

for $\theta > 0$, where M is a suitably large (but finite) positive constant, and $\mathbf{e}$ is a vector of all ones.

Next, we prove Theorem 2.

Proof. First of all, when $\theta = 0$, (3.17) is equivalent to scenario approximation. (Operations SAA) Therefore, it is obvious (3.17) and OPT2 are equivalent. When $\theta > 0$, by Lemma 1, (3.17) is equivalent to the following problem.

$$\begin{aligned} \min_{\mathbf{q},\mathbf{s},t,\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N \|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p, \\ & (\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} + Mq_i \geqslant t + s_i, \forall i \in [N], \\ & M(1 - q_i) \geqslant t + s_i, \forall i \in [N], \\ & \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{3.20}$$

Suppose the optimal solution for (3.20) satisfies $q_i = 0$ when $i \in I_1$ and $q_i = 1$ when $i \in I_2$ for some sets I_1 and I_2. Then, problem (3.20) is equivalent to:

$$\begin{aligned} \min_{\mathbf{s},t,\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N \|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p, \\ & (\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} \geqslant t + s_i, \forall i \in I_1, \\ & 0 \geqslant t + s_i, \forall i \in I_2, \\ & \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}. \end{aligned} \tag{3.21}$$

The Lagrange multiplier of (3.21) is

$$\begin{aligned} \min_{\mathbf{s}\leqslant 0, t\in\mathbb{R}, \mathbf{x}\in\mathcal{X}} \quad \max_{\lambda\geqslant 0, \beta_i \geqslant 0} \quad & \mathbf{c}^T\mathbf{x} - \lambda(\epsilon N t + \mathbf{e}^T\mathbf{s} - \theta N\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p) + \sum_{i\in I_2}\beta_i(t + s_i) \\ & - \sum_{i\in I_1}\beta_i\left[(\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} - t - s_i\right], \end{aligned} \tag{3.22}$$

which is equivalent to (following minimax theorem as (3.22) is linear with respect to variables $\mathbf{s}, t$ when λ, β_i are fixed and is also linear respect to λ, β_i when $\mathbf{s}, t$ are fixed):

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\lambda\geqslant 0, \beta_i\geqslant 0} \quad \min_{\mathbf{s}\leqslant 0, t\in\mathbb{R}} \quad & \mathbf{c}^T\mathbf{x} + \lambda\theta N\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p - \sum_{i\in I_1}\beta_i(\mathbf{b} - \mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i + \sum_{i\in I_1}\beta_i\mathbf{a}^T\mathbf{x}_1 \\ & + \sum_{i\in I_1}\beta_i\mathbf{d}^T\mathbf{x}_2 - \sum_{i\in I_1}\beta_i b_{k0} \\ & - \lambda\epsilon N t + \sum_{i\in I_1}\beta_i t + \sum_{i\in I_2}\beta_i t \\ & - \lambda\mathbf{e}^T\mathbf{s} + \sum_{i\in I_1}\beta_i s_i + \sum_{i\in I_2}\beta_i s_i \end{aligned} \tag{3.23}$$

The optimal solution of the original problem exists and is bounded, therefore, (3.23) is bounded meaning the coefficient of t is zero, and the coefficients of $\mathbf{s}$ are non-positive. Thus,

solving (3.23) gives

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\lambda\geqslant 0,\beta_i\geqslant 0} \quad & \mathbf{c}^T\mathbf{x} + \lambda\theta N\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p - \sum_{i\in I_1}\beta_i(\mathbf{b}-\mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i + \sum_{i\in I_1}\beta_i\mathbf{a}^T\mathbf{x}_1 + \sum_{i\in I_1}\beta_i\mathbf{d}^T\mathbf{x}_2 \\ & - \sum_{i\in I_1}\beta_i b_{k0} \\ s.t. \quad & \sum_i \beta_i = \lambda\epsilon N, \\ & \beta_i \leqslant \lambda, \quad \forall i.\end{aligned} \tag{3.24}$$

Replacing λ with β_i gives:

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\beta_i\geqslant 0} \quad & \mathbf{c}^T\mathbf{x} + \left[\frac{\sum_i\beta_i}{\epsilon N}\theta N\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p - \sum_{i\in I_1}\beta_i(\mathbf{b}-\mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i + \sum_{i\in I_1}\beta_i\mathbf{a}^T\mathbf{x}_1\right] \\ & + \sum_{i\in I_1}\beta_i\mathbf{d}^T\mathbf{x}_2 - \sum_{i\in I_1}\beta_i b_{k0} \\ s.t. \quad & \beta_i \leqslant \frac{\sum_i\beta_i}{\epsilon N}, \quad \forall i.\end{aligned} \tag{3.25}$$

Because $N \leqslant \dfrac{1}{\epsilon}$, constrains $\beta_i \leqslant \frac{\sum_i\beta_i}{\epsilon N}$ hold trivially for all $\beta_i \geqslant 0$. We drop these constraints, which gives:

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\beta_i\geqslant 0} \quad & \sum_{i\in I_2}\beta_i\frac{\theta}{\epsilon}\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p + \sum_{i\in I_1}\beta_i\left[\frac{\theta}{\epsilon}\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p - (\mathbf{b}-\mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i + \mathbf{a}^T\mathbf{x}_1\right. \\ & \left. + \mathbf{d}^T\mathbf{x}_2 - b_{k0}\right] + \mathbf{c}^T\mathbf{x}.\end{aligned} \tag{3.26}$$

Again, the original problem is feasible and bounded indicating that (3.26) is bounded. Therefore, the coefficients of β_i are all non-positive. However, we have $\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p > 0$ because we assume $\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p \neq 0$. Then the primal solution is infeasible unless $I_2 = \emptyset$. Therefore, we have $I_2 = \emptyset$. In addition, for $i \in I_1$, we restrict their coefficients to be non-positive, which gives

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & (\mathbf{b}-\mathbf{A}^T\mathbf{x}_1)^T\hat{\xi}_i - \mathbf{a}^T\mathbf{x}_1 - \mathbf{d}^T\mathbf{x}_2 + b_{k0} \geqslant \frac{\theta}{\epsilon}\|\mathbf{A}^T\mathbf{x}_1 - \mathbf{b}\|_p, \quad \forall i. \quad \square\end{aligned}$$

3.2.2 Joint chance constraint

We summarize the main results for joint chance constrained programming in Theorem 3. We also restate OPT2 when $K > 1$ here for clarity.

$$
\begin{aligned}
(OPT2) \quad \min \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\boldsymbol{\xi}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 + b_{k0} \\
& \geqslant c_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_p, \forall i \in [N], \forall k \in [K], \\
& \mathbf{x} \in \mathcal{X}.
\end{aligned}
$$

Theorem 3. *Suppose the number of observations, N, is less than $\frac{1}{\epsilon}$ ($\theta \geqslant 0$), and the coefficient of $\boldsymbol{\xi}$, $(\mathbf{A}_k^T\mathbf{x}_1 - \mathbf{b}_k)$, is not zero. When $c_i = \frac{\theta}{\epsilon}$, OPT2 is equivalent to*

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathcal{X}} \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \mathbb{E}_{P\in\mathcal{B}_\theta(\hat{\mathbb{P}})}\left\{1[\mathbf{x} \in \mathcal{X}'(\boldsymbol{\xi})]\right\} \geqslant 1 - \epsilon, \\
& \mathcal{X}'(\boldsymbol{\xi}) = \left\{\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2] \middle| (\mathbf{A}_k\boldsymbol{\xi} + \mathbf{a}_k)^T\mathbf{x}_1 + \mathbf{d}_k^T\mathbf{x}_2 \leqslant \mathbf{b}_k^T\boldsymbol{\xi} + b_{k0}, \forall k \in [K]\right\},
\end{aligned}
\tag{3.27}
$$

where the ambiguity set $\mathcal{B}_\theta(\hat{\mathbb{P}})$ is a Wasserstein ball with radius θ centered at $\hat{\mathbb{P}}$ representing the empirical distribution of $\boldsymbol{\xi}$.

The ambiguity set follows the same definition as before. The proof of Theorem 3 follows the same idea of the proof of Theorem 2, which is given in A.1.

3.3 Computational Study

In this section, we conduct numerical experiments to study our proposed model. We first compare our approach with the scenario/sample approximation approach in Section 3.3.1, where our model is shown to be greatly improving the out-of-sample performance. Then, we compare our model with the DRCCPs in Section 3.2, where our model is shown to be more robust and always gives the correct solution.

We study the following chance-constrained problem (3.28), which has applications in many real-world problems including portfolio optimization and inventory management. In portfolio optimization [60], ξ_i represents the random return of the i-th assets. The goal is to find the value at risk, x, at level ϵ for these I assets. In inventory management [116], ξ_i represents the random demand at day i. The goal is to determine an inventory level, x, such that the probability of having outstocks is less than ϵ during I days.

$$\min_{x} \quad x \tag{3.28a}$$

$$s.t. \quad \mathrm{P}(\sum_{i\in[I]} \xi_i \leqslant x) \geqslant 1-\epsilon. \tag{3.28b}$$

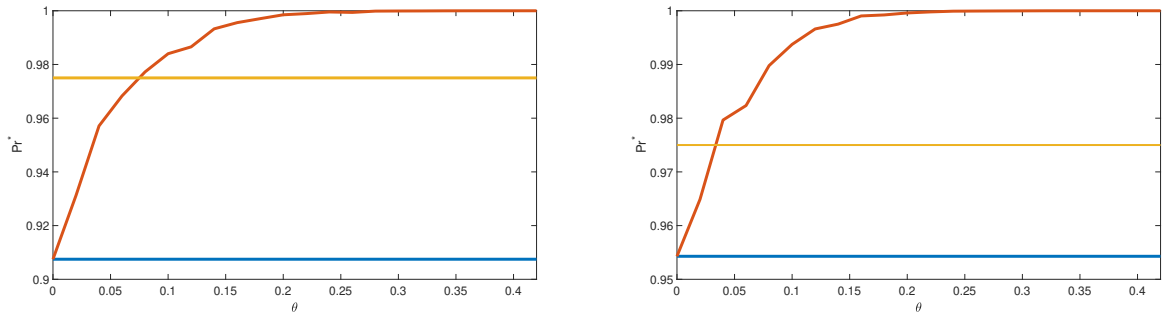
3.3.1 Benefits over scenario/sample approximation approach

We assume $I = 4$, $\epsilon = 2.5\%$. Random variables ξ_i, $i \in [I]$ are independent with each other and follow normal distribution $N(5, 5)$. We conduct two sets of experiments by simulating 10 and 20 samples for each ξ_i. Each experiment is repeated for 100 times. Scenario approach and our approach are used to solve (3.28) for comparisons. We evaluate the out-of-sample performance of the chance constraint by computing P^*, i.e.

$$\mathrm{P}^* = \mathbb{E}[1(\sum_{i\in[I]} \xi_i \leqslant \hat{x})],$$

where the expectation is taken over the true distribution of ξ_i, and $\hat{x}$ represents the data-driven solution based on different approaches. We set $c_i = \dfrac{\theta}{\epsilon}$ and vary the value of θ. The mean values of P^* are summarized in Figure 1.

Scenario/sample approximation approach achieves an average P^* that is lower than the desired value, i.e. $1-\epsilon = 0.975$, which validates our former discussions on the low confidence level of this approach under the small-data regime. Our approach improves the performance of P^* by increasing θ.



(a) 10 samples.

(b) 20 samples.

Figure 1: Blue line: scenario/sample approximation approach. Red line: our approach. Yellow line: Goal.

3.3.2 Benefits over existing DRCCP formulations

In this section, we show the existing formulations, (3.19) and (A.1), of DRCCP are numerically unstable when θ goes to zero. The M in (3.19) and (A.1) is fixed as 10^{10}. The parameters follow the same settings introduced in the last section. We assume 20 samples for each ξ_i are available. For these samples, we solve Problem (3.28) through both DRCCP and our approaches using different values of $\theta = \{0, 0.005, \cdots, 0.04\}$. We record the optimal solution and conclude the results in Figure 2.

From the results, our formulation always gives the correct solutions; however, the DRCCP formulation produces wrong solutions when $\theta \leqslant 0.025$.

3.3.2.1 Multidimensional knapsack problem We also apply our method to the distributionally robust multidimensional knapsack problem (DRMKP) [134] to further compare our approach with the DRCCP formulation with big-M coefficients. In a DRMKP, there are m items and K knapsacks. Parameters c_j represents the value of item j for all $j \in [m]$; $\boldsymbol{\xi}_k = (\xi_{k1}, \cdots, \xi_{km})$ represents the vector of random item weights in knapsack k; and $b_k > 0$ represents the capacity limit of knapsack k, for all $k \in [K]$. The decision variable $\mathbf{x}_j \in \{0, 1\}$

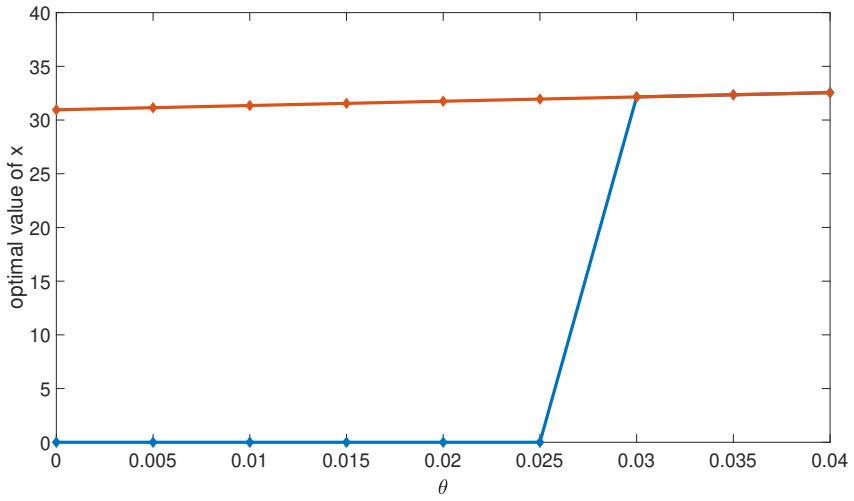


Figure 2: Blue line: DRCCP. Orange line: our approach.

represents if the jth item is picked or not. We use the Wasserstein ambiguity set with L∞-norm as distance metric. With the above notations, DRMKP is formulated in (3.29).

$$\begin{aligned} \max_{\mathbf{x}\in\mathcal{X}} \quad & \mathbf{c}^T\mathbf{x} \\ s.t. \quad & \inf_{P\in\mathcal{B}_\theta(\hat{\mathbb{P}})} \mathrm{P}\left[\boldsymbol{\xi}_k^T\mathbf{x} \leqslant b_k, \forall k \in [K]\right] \geqslant 1-\epsilon. \end{aligned} \tag{3.29}$$

We generated 10 random instances with $m = 20$ and $K = 10$. For each instance, we generated $N = 800$ empirical samples of $\boldsymbol{\xi}_k$ from a uniform distribution over a box $[1, 10]^m$. For each $j \in [m]$, we independently generated c_j from the uniform distribution on the interval $[1, 10]$, while for each $k \in [K]$, we set $b_k = 100$. We tested these 10 random instances with risk parameter $\epsilon = 0.001$. We set the big M to 1000 in this problem.

The results are summarized in Table 1 and 2. Table 1 records the time difference between BigM formulation [134, 31] and our formulation. A positive number indicates that our method takes less time. Table 2 records the optimality gap of the BigM formulation, where the optimality gap is defined in (3.30). The optimal values are obtained through our models.

$$gap = \left|\frac{Value - Opt.Val}{Opt.Val}\right|. \tag{3.30}$$

Table 1: Time differences (in seconds) between the BigM formulation and our model.

	$\theta = 10^{-5}$	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
1	3.457	0.93	1.081	1.883	-3.966	-0.884
2	-0.591	0.626	0.829	0.957	1.116	-3.319
3	1.285	1.088	1.159	-0.742	-1.537	-0.97
4	0.529	1.386	1.833	3.036	-0.854	-1.995
5	3.862	0.85	1.221	0.086	-0.632	-0.567
6	3.522	0.851	1.079	-0.578	-0.42	-1.826
7	0.845	3.308	1.426	-0.733	-0.861	-1.632
8	1.787	2.878	0.807	0.676	-0.66	-1.173
9	1.398	1.502	2.332	1.746	-0.84	0.952
10	0.35	1.705	2.056	2.915	-1.206	-1.439

Table 2: Optimality gap of the BigM formulation.

	$\theta = 10^{-5}$	10^{-6}	10^{-7}	10^{-8}	10^{-9}	10^{-10}
1	0.00%	0.00%	0.00%	0.00%	5.30%	4.70%
2	0.00%	0.00%	0.00%	0.00%	1.30%	3.10%
3	0.00%	0.00%	0.00%	4.40%	6.50%	4.20%
4	0.00%	0.00%	0.00%	0.00%	4.00%	3.40%
5	0.00%	0.00%	0.00%	3.50%	2.90%	4.80%
6	0.00%	0.00%	0.00%	5.10%	4.70%	5.30%
7	0.00%	0.00%	0.00%	4.50%	7.10%	4.20%
8	0.00%	0.00%	0.00%	0.00%	6.90%	5.30%
9	0.00%	0.00%	0.00%	0.00%	6.00%	8.90%
10	0.00%	0.00%	0.00%	0.00%	4.40%	5.60%

Regardless the computation time is seldom an issue for small-data problems, Table 1 shows that our model and the BigM formulation have similar computation times. Moreover, our approaches take less time to determine the optimal solution when $\theta \in \{10^{-5}, 10^{-6}, 10^{-7}\}$. Although the BigM formulation outperforms our model when $\theta \in \{10^{-8}, 10^{-9}, 10^{-10}\}$, the BigM formulation suffers from computational issues in these cases, which can be evidenced from Table 2.

4.0 Maximum Weight Cycle and Chain Packing with Inhomogeneous Edge Existence Uncertainty: An Application on Kidney Exchange

4.1 Introduction

Patients with end-stage renal failure often find kidney donors who are willing to donate a life-saving kidney, but who are medically incompatible with the patients. Kidney exchanges are organized barter markets that allow such incompatible patient-donor pairs to enter as a single agent—where the patient is endowed with a donor "item"—and engage in trade with other similar agents, such that all agents "give" a donor organ if and only if they receive an organ in return. In practice, organized trades occur in large cyclic or chain-like structures, with multiple agents participating in the exchange event. Planned trades can fail for a variety of reasons, such as unforeseen logistical challenges, or changes in patient or donor health. These failures cause major inefficiency in fielded exchanges, as if even one individual trade fails in a planned cycle or chain, *all or most of the resulting cycle or chain fails*. Ad-hoc, as well as optimization-based methods, have been developed to handle failure uncertainty; nevertheless, the majority of the existing methods use very simplified assumptions about failure uncertainty and/or are not scalable for real-world kidney exchanges.

Motivated by kidney exchange, we study a stochastic cycle and chain packing problem, where we aim to identify structures in a directed graph to maximize the expectation of matched edge weights. All edges are subject to failure, and the failures can have nonidentical probabilities. To the best of our knowledge, the state-of-the-art approaches are only tractable when failure probabilities are identical. We formulate a relevant non-convex optimization problem and propose a tractable mixed-integer linear programming reformulation to solve it. In addition, we propose a model that integrates both risks and the expected utilities of the matching by incorporating conditional value at risk (CVaR) into the objective function, providing a robust formulation for this problem. Subsequently, we propose a sample-average-approximation based approach to solve this problem. We fill a major gap in prior work by proposing the first *scalable* algorithm (meaning it uses a number of variables

polynomial in the input size) for maximizing expected matching weight, with *non-identical* failure probabilities. This is an important step forward, as failure probabilities are known to be inhomogeneous–some edges are inherently riskier than others [41]. We provide a mixed-integer linear program for our approach, which is compact and can be solved directly by a general-purpose integer programming solver (e.g., CPLEX, Gurobi, or SCIP).

Additionally, we propose a modified version of the kidney exchange problem which balances the *mean expected weight* with the *worst-case* weight ("risk") of an exchange with known nonidentical edge failure probabilities; we achieve this balance using a conditional value-at-risk (CVaR) objective. We are not the first to propose a CVaR approach for kidney exchange; however, previous CVaR-based approaches do not allow for arbitrary length limits on cycles and chains–which are used by all fielded exchanges. With cycle and chain length limits, the kidney exchange problem with a CVaR objective is challenging, as there is no closed-form expression for the objective function. Thus, we propose a sample-average-approximation-based method and develop an equivalent mixed-integer linear programming representation.

4.1.1 Uncertainty in Kidney Exchange

Many prior approaches address edge existence uncertainty in kidney exchange, often with the objective of maximizing *expected* matching weight, assuming all edges have identical failure probability. [39] provides a scalable formulation in this case, and [41] extends this to consider inhomogeneous edge probabilities; however the latter model can require enumeration of all feasible cycles and chains, which can be intractable for even small exchanges. Similar approaches have been proposed, but still assume that all edges have equal failure probability [4, 35]. Rather than maximizing expected edge weight, other approaches take the *risk-averse* perspective, aiming to maximize the worst-case matching weight [91, 29]; these approaches are often *too* conservative, as the worst case in kidney exchange is often arbitrarily bad (i.e., in the worst case, all planned transplants fail). [138] propose a CVaR method that endogenously balances structure length with risk; however, their model is not amenable to length caps on cycles and/or chains, a requirement in all fielded kidney exchanges.

Several other optimization-based approaches have been proposed, using recourse [5], forms of "fallback" options [88, 11, 129], and pre-match edge queries [19, 18, 92]. These methods involve additional decision stages, and are not directly comparable in our setting. Next we describe the formal model of kidney exchange and edge existence uncertainty.

4.2 Preliminaries

We represent a kidney exchange as a directed graph $G = (E, V)$ where each vertex $v_i \in V$ is an incompatible patient-donor pair, or a non-directed donor (NDD, i.e., a donor without a paired patient). Directed edges $e = (v_i, v_j)$ represent potential transplants from the donor of node i to the patient of node j; edge weights $w_e > 0$ represent the medical or social utility of each potential transplant. We assume that edge failure probabilities $p_e \in [0, 1]$ are known in advance and are not necessarily homogeneous. That is, if edge $e = (v_i, v_j)$ is matched, then with probability p_e the patient of v_j would still fail to receive a kidney from v_i's donor.

Kidney exchanges consist of two types of swaps: *cycles* consist of several patient donor pairs, while *chains* begin with an NDD and continue through one or more patient pairs [106]. The goal of the *kidney exchange clearing problem* (KEP) is often to select the set of vertex-disjoint cycles and chains in G which maximize overall edge weight. We refer to any set of vertex-disjoint cycles and chains as a *matching*. For example, let $\boldsymbol{w}$ denote the vector of weights for all cycles and chains in the graph, let $\boldsymbol{x}$ denote a vector of binary decision variables, and let $\mathcal{M}$ denote the set of feasible matchings (i.e., binary vectors $\boldsymbol{x}$ corresponding to vertex-disjoint cycles and chains); in this case the KEP is expressible as $\max_{\boldsymbol{x} \in \mathcal{M}} \boldsymbol{x}^\top \boldsymbol{w}$.

Cycles and chains are quite vulnerable to edge failure: if *any* edge in a cycle fails, then *none* of the transplants in the cycle can proceed, because at least one of the patients will be left without a compatible donor. If an edge participating in a chain fails, then none of the edges *following* that failed edge can proceed.[1] We consider modified versions of the KEP which account for edge failures, using known edge failure probabilities.

[1] We assume that chains can be *partially* executed. Some fielded exchanges cancel the entire chain if even one edge fails.

4.3 Maximizing Expected Matching Weight

We are primarily interested in maximizing the *expected* weight of a matching; indeed this is the focus of most prior work (see Section 4.1.1). We refer to this as the *stochastic* KEP. First we characterize the objective of this problem—the expected matching weight. With known edge failure probabilities, the expected weight of a cycle or chain is expressible in closed form.

Discounted weight of a cycle. The discounted weight of a k-cycle c reflects the fact that the *whole cycle* will fail if any single transplant fails. We use w_e to denote the weight of edge e in the cycle, c.

$$u(c) = \left(\sum_{e \in c} w_e\right) \left[\prod_{e \in c}(1 - p_e)\right].$$

Discounted weight of a chain. The expected weight $u(\kappa)$ of the k-chain $\kappa \equiv (v_1, ..., v_{k+1})$, where v_1 is a non-directed donor (NDD), is defined as

$$\begin{aligned} u(\kappa) = &\sum_{i=2}^{k} p_i \left(\sum_{j=1}^{i-1} w_j\right) \prod_{j=1}^{i-1}(1 - p_j) + \\ &\left(\sum_{i=1}^{k} w_i\right) \prod_{i=1}^{k}(1 - p_i). \end{aligned} \tag{4.1}$$

In the above, p_i and w_i denotes the failure probability and weight of edge (v_i, v_{i+1}), respectively. The first term above is the sum of expected weights for the chain executing exactly $i - 1 = \{1, ..., k - 1\}$ steps and then failing on the ith step. The second term is the resulting weight if the chain executes completely.

Using the above expressions, we can write the stochastic KEP as follows. With some abuse of notation, let $(\boldsymbol{C}, \boldsymbol{K}) \in \mathcal{M}$ denote a feasible matching consisting of cycles $\boldsymbol{C}$ and chains $\boldsymbol{K}$. Problem 4.2 is an equivalent formulation of the stochastic KEP.

$$\max_{(\boldsymbol{C}', \boldsymbol{K}') \in \mathcal{M}} \quad \sum_{c' \in \boldsymbol{C}} u(c') + \sum_{\kappa' \in \boldsymbol{K}} u(\kappa') \tag{4.2}$$

Next we describe our solution approach for Problem 4.2, and an equivalent compact mixed-integer linear program formulation.

4.3.1 Compact Formulation for Maximizing Expected Matching Weight

Here we present a new compact formulation to maximize the expected weight in the case of *non-identical* edge failure probabilities. *Compact* means that the counts of variables and constraints are polynomial in the size of the input. We compare the size of this model with other state-of-the-art approaches in Section 4.3.2.

In [41], the authors propose a solution approach for Problem 4.2, which enumerates all feasible cycles and chains in the graph. However the number of cycles and chains grows exponentially with the size of the graph, meaning this formulation is not compact. Further, it is intractable to even write this model in memory for large exchanges or long chain lengths.

We propose an exact, compact representation for Problem 4.2, using an equivalent expression for expected chain weight $u(\kappa)$ given in Lemma 2.

Lemma 2. *The discounted weight $u(\kappa)$ of the k-chain $\kappa = (v_1, ..., v_{k+1})$ is*

$$u(\kappa) = \sum_{i=1}^{k} w_i \prod_{j=1}^{i} (1 - p_j),$$

where w_i and p_i are the edge weight and failur probability of the i^{th} edge, (v_i, v_{i+1}), in the chain, for $i = 1, \ldots, k$.

In other words, the discounted weight of a chain can be expressed as the sum of the "discounted weights" of each *edge* in the chain, i.e. $u(\kappa) = \sum_{i=1}^{k} w'_i$, where $w'_i \equiv w_i \prod_{j=1}^{i}(1 - p_j)$, where we refer to $\prod_{j=1}^{i}(1 - p_j)$ as the *discount factor*.

The objective of Optimization (4.4) uses Lemma 2 to express the total discounted weight of all matched cycles and chains, assuming non-uniform edge failure probabilities. This is achieved using two sets of variables, o_{ek} (the discount factor of edge e at position k in a chain) and v_c (the success probability of cycle c). Optimization (4.4) uses the following parameters:

- G: kidney exchange graph, consisting of edges $e \in E$ and vertices $v \in V = P \cup N$, including patient-donor pairs P and non-directed donors (NDDs) N
- C: a set of cycles on exchange graph G
- L: chain cap (max. number of edges in a chain)
- w_e: edge weights for each edge $e \in E$

- w_c: cycle weights for each cycle $c \in C$, defined as $w_c = \sum_{e \in c} w_e$
- $\delta^-(i)$: the set of edges into vertex i
- $\delta^+(i)$: the set of edges out of vertex i
- p_e: failure probability for edge $e \in E$

Edges between an NDD $n \in N$ and a patient-donor vertex $d \in P$ may only take position 1 in a chain, while edges between two patient-donor pairs may take any position $2, \ldots, L$ in a chain. For convenience, we define the function $\mathcal{K}$ for each edge e, such that $\mathcal{K}(e)$ is the set of all possible positions that edge e may take in a chain.

$$\mathcal{K}(e) = \begin{cases} \{1\}, & e \text{ begins in } n \in N, \\ \{2, \ldots, L\}, & e \text{ begins in } d \in P. \end{cases} \tag{4.3}$$

The following decision variables are used.

- $z_c \in \{0, 1\}$: 1 if cycle c is used in the matching, and 0 otherwise
- $y_{ek} \in \{0, 1\}$: 1 if edge e is used at position k in a chain, and 0 otherwise
- $o_{ek} \in [0, 1]$: discount factor of edge e at position k in a chain

Our formulation is given in (4.4).

$$\max_{\mathbf{y},\mathbf{z},\mathbf{o}} \quad \sum_{e \in E} \sum_{k \in \mathcal{K}(e)} w_e y_{ek} o_{ek} + \sum_{c \in C} w_c z_c v_c \tag{4.4a}$$

$$s.t. \quad \{\mathbf{y}, \mathbf{z}\} \in \mathcal{X}, \tag{4.4b}$$

$$\sum_{\substack{e \in \delta^-(i) \wedge \\ k \in \mathcal{K}(e)}} o_{ek} y_{ek} \geqslant \sum_{e \in \delta^+(i)} \frac{o_{e,k+1} y_{e,k+1}}{1 - p_e},$$

$$i \in P, k \in \{1, \ldots, L-1\}, \tag{4.4c}$$

$$0 \leqslant o_{ek} \leqslant 1 - p_e, e \in E, k \in \mathcal{K}(e), \tag{4.4d}$$

$$v_c = \prod_{e \in c} 1 - p_e, c \in C, \tag{4.4e}$$

where $\mathcal{X}$ denotes the set of feasible decision variables for the PICEF formulation of kidney exchange [39], defined as

$$\mathcal{X} = \begin{cases} \sum\limits_{e \in \delta^-(i)} \sum\limits_{k \in \mathcal{K}(e)} y_{ek} + \sum\limits_{\substack{c \in C: \\ i \in c}} z_c \leqslant 1, & i \in P; \\ \sum\limits_{\substack{e \in \delta^-(i) \wedge \\ k \in \mathcal{K}(e)}} y_{ek} \geqslant \sum\limits_{e \in \delta^+(i)} y_{e,k+1}, & \begin{array}{l} i \in P \\ k \in \{1, ..., L-1\}; \end{array} \\ \sum\limits_{e \in \delta^+(i)} y_{e1} \leqslant 1, & i \in N; \\ y_{ek} \in \{0,1\}, & e \in E, k \in \mathcal{K}(e); \\ z_c \in \{0,1\}, & c \in C; \end{cases} \tag{4.5}$$

The constraints of $\mathcal{X}$ are interpreted as follows: 1) the first constraint in (4.5) requires that each patient-donor vertex $i \in P$ may only participate in one cycle or one chain; 2) the second constraint requires that each patient-donor vertex $i \in P$ can only have an outgoing edge at position $k+1$ in a chain if it has an incoming edge at position k; 3) the third constraint requires that each NDD $i \in N$ may only participate in one chain.

Constraints (4.4c), (4.4d), and (4.4e) define the discounted weight of chains and cycles. We briefly describe how the discounted weight of cycles and chains are represented in this formulation:

- For a *cycle*, the success probability is $v_c = \prod_{e \in c} 1 - p_e$. Thus the discounted weight of all cycles is expressed as $\sum_{c \in C} w_c z_c v_c$.
- For a *chain*, the discounted weight is expressed using Lemma 2. Consider the following example: suppose a k-chain consists of edges $e_1, \ldots, e_k$. Suppose that i is the *first* patient-donor pair in this chain– so e_1 is the edge *into* i, and e_2 is the edge *out of* i; that is, $e_1 \in \delta^-(i)$ and $e_2 \in \delta^+(i)$. From constraints (4.4c) we have $o_{e_1,1} \geqslant \frac{1}{1-p_{e_2}} o_{e_2,2}$ for vertex i. The sums in constraint (4.4c) contain no other terms, because $\mathcal{X}$ requires that only one edge into vertex i and one edge out of vertex i can be matched. Therefore, $(1-p_{e_2}) o_{e_1,1} \geqslant o_{e_2,2}$.

  Similarly, $(1-p_{e_{j+1}}) o_{e_j,j} \geqslant o_{e_{j+1},j+1}$ for $j = 2, \ldots, k-1$. Since Optimization (4.4) is a maximization problem, the optimal values of variables $o_{e,j}$ will satisfy $o_{e_j,j} = \prod_{i=1}^{j} (1 - p_{e_i})$, for $1 \leqslant j \leqslant k$. Accordingly, $\sum_{e \in E} \sum_{k \in \mathcal{K}(e)} w_e y_{ek} o_{ek}$ represents the total discounted weight of all chains according to Lemma 2.

4.3.2 MIP Reformulation of Optimization (4.4)

Although Optimization (4.4) exactly maximizes expected edge weight under non-identical edge failure probabilities, it is a nonconvex optimization problem. In this section, we reformulate it as a mixed-integer linear program which can be solved usings general-purpose solvers. Proposition 2 concludes our results; the main idea is to define a set of new variables O_{ek} to replace $y_{ek}o_{ek}$ in Optimization (4.4).

Proposition 2. *Optimization (4.4) is equivalent to*

$$\begin{aligned} \max_{\mathbf{y},\mathbf{z},\mathbf{O},\mathbf{o}} \quad & \sum_{e\in E}\sum_{k\in\mathcal{K}(e)} w_e O_{ek} + \sum_{c\in C} w_c z_c \Big(\prod_{e\in c} 1-p_e\Big) \\ s.t. \quad & \{\mathbf{y},\mathbf{z}\}\in\mathcal{X}, \\ & \{\mathbf{y},\mathbf{O},\mathbf{o}\}\in\mathcal{X}', \end{aligned} \tag{4.6}$$

where $\mathcal{X}$ follows the definition in (4.4), and $\mathcal{X}'$ is defined as

$$\mathcal{X}' = \left\{ \begin{array}{l} \displaystyle\sum_{e\in\delta^-(i)\wedge k\in\mathcal{K}(e)} O_{ek} \geq \sum_{e\in\delta^+(i)} \frac{O_{e,k+1}}{1-p_e}; \\ i\in P, k\in\{1,\ldots,L-1\}, \\ O_{ek}\leq y_{ek}, e\in E, k\in\mathcal{K}(e); \\ O_{ek}\leq o_{ek}, e\in E, k\in\mathcal{K}(e); \\ O_{ek}\in[0,1], e\in E, k\in\mathcal{K}(e); \\ 0\leq o_{ek}\leq 1-p_e, e\in E, k\in\mathcal{K}(e). \end{array} \right\} \tag{4.7}$$

Optimization (4.6) can be solved using standard solvers such as CPLEX and Gurobi.

4.3.2.1 Scalability We compare our model size with state-of-the-art approaches in literature. We summarize all approaches in Table 3. The size of each model (the number of variables and constraints) is expressible in terms of the chain cap L, and the number of edges ($|E|$), cycles ($|C|$), total vertices ($|V|$), NDD vertices ($|N|$), and patient-donor pair vertices $|P|$. For ease of exposition we assume $|N| = O(|V|)$ and $|P| = O(|V|)$. In Table 3, columns indicate the type of uncertainty considered in the problem (stochastic, robust (i.e., worst-case, CVaR, or none), whether or not edge failure probability is assumed to be *homogeneous*, whether or not the formulation includes a cycle and chain cap[2], and the number of variables and constraints in each formulation.

[2]Without cycle or chain cap, kidney exchange can be reduced to bipartite matching.

Table 3: Comparison of stochastic and robust approaches to kidney exchange.

Formulation	Uncertainty	Homogeneous p_e	Chains	Cycle/Chain Cap	# Vars.	# Constr.
PC-TSP [5]	None	N/A	Yes	Yes/Yes	$O(\|E\| \cdot \|V\| + \|V\|^2 + \|C\|)$	$O(\|E\| \cdot \|V\| + \|V\|^2 + \|V\| \cdot 2^{\|V\|} + \|C\|)$
PICEF [40]	Stoch.	Yes	Yes	Yes/Yes	$O(L \cdot \|E\| + \|C\|)$	$O(L \cdot \|V\| + L \cdot \|E\| + \|C\|)$
ROBUST[91]	Robust	N/A	Yes	Yes/Yes	$O(\|E\| \cdot \|V\| + \|V\|^2 + \|C\|)$	$O(\|E\| \cdot \|V\| + \|V\|^2 + \|C\|)$
SMCF-VaR/CVaR [138]	Stoch.	No	No	–/No	I assume $\|\Omega\| = 2^{\|E\|}$,$O(2^{\|E\|}(\|V\| + \|E\|))$	$O(2^{\|E\|}(\|V\| + \|E\|))$
DPS-18 [41]	Stoch.	No	Yes	Yes/Yes	$O(\|V\|^L + \|C\|)$	$O(\|V\|)$
Our model (4.6)	Stoch.	No	Yes	Yes/Yes	$O(L \cdot \|E\| + \|C\|)$	$O(L \cdot \|V\| + L \cdot \|E\| + \|C\|)$

Our size is comparable with PICEF, while accounting for non-identical failure probabilities. DPS-18 [41] considers non-identical failure probabilities at the cost of representing every single chain and cycle as a decision variable, and thus this model grows exponentially with the chain cap L; in contrast, the number of variables in our formulation is polynomial in L. Real exchanges often use a cycle cap of 3, which is sufficiently small that all cycles can be enumerated in practice–even on realistic graphs with hundreds of vertices. If exchanges grow much larger in the future (e.g., thousands of vertices), or if cycle lengths are increased substantially, we further propose a branch-and-price implementation to solve the corresponding problems brought by huge $|C|$ in Appendix B.3.

4.4 Extensions to Mean-risk Kidney Exchange Model

Next we introduce a kidney exchange model which balances both the *mean expected weight* and the *worst-case* weight ("risk") of a matching, using known non-identical edge failure probabilities. We achieve this balance using a conditional value-at-risk (CVaR) objective. This approach is motivated by the fact that the *expected* weight of a matching can be misleading when the *worst-case* outcome can be arbitrarily bad. This is especially true in kidney exchange, where a single edge failure can impact an entire cycle or chain.

4.4.1 Mean-risk Model

At a high level, the CVaR objective for kidney exchange is expressed as

$$\mu + \gamma \times \mu_\alpha,$$

where μ is the expected matching weight and μ_α is the $\alpha \times 100\%$ ($\alpha \in (0, 1]$) *worst-case* mean weight–that is, the mean matching weight in the worst $\alpha \times 100\%$ of all outcomes. The parameter γ is set by the user, and controls the trade-off between average performance and the *risk* of the solution.

For tractability and simplicity, we define $\mathbf{W}$ as an $|E|$-dimensional vector with

$$W_e = -\sum_{k \in \mathcal{K}(e)} y_{ek} - \sum_{c \in C} \mathbf{1}(e \in c) z_c, \quad \forall e \in E.$$

That is, $W_e = -1$ if edge e is used, and $W_e = 0$ otherwise. We use $\mathbf{w} \in \mathbb{R}^{|E|}$ to represent the random discounted edge weights under known edge failure probabilities. Correspondingly, $\langle \mathbf{w}, \mathbf{W} \rangle$ represents the loss (negative weight) of a matching. The $\alpha \times 100\%$ worst-case (highest) mean loss is equivalent to the CVaR objective [103] at level α. The corresponding optimization problem is expressed in (4.8), by introducing an auxiliary variable d. We use $(x)^+$ to denote $\max(0, x)$, and the expectation in (4.8) is taken over the distribution of random edge weights under the known edge failure probabilities. As before, $\mathcal{X}$ denotes the set of feasible matchings using the PICEF formulation.

$$\begin{aligned} \min_{\mathbf{y},\mathbf{z},d} \quad & \mathbb{E}(\langle \mathbf{w}, \mathbf{W} \rangle) + \gamma \left[d + \frac{1}{\alpha} \mathbb{E}\left[(\langle \mathbf{w}, \mathbf{W} \rangle - d)^+ \right] \right] \\ s.t. \quad & \{\mathbf{y}, \mathbf{z}\} \in \mathcal{X}. \end{aligned} \tag{4.8}$$

4.4.2 An SAA-based Approach for Optimization (4.8)

The main difficulty in solving Optimization (4.8) is that term $\mathbb{E}\left[(\langle \mathbf{w}, \mathbf{W} \rangle - d)^+\right]$ does not have a simple closed-form reformulation under the known edge failure probabilities. Thus, we propose an approach based on Sample Average Approximation (SAA) [5] to solve (4.8). The main idea is to first sample N realizations of edge existence according to the known edge failure probabilities; for each realization we formulate a mixed-integer linear program representing the matching weight under this realization. Finally, we combine all N models to obtain an optimization problem that is (approximately) equivalent to Optimization (4.8) based on these N realizations. Algorithm 1 gives a pseudocode description of this approach.

Algorithm 1 SAA

1: **Initialization:** N;
2: **STEP 1**:
3: Sample N edge existence realizations $\{\hat{r}_{en} \in \{0,1\}, \forall e \in E\}$, $n = 1, \ldots, N$;
4: **STEP 2**:
5: Solve Optimization (4.9).

This algorithm has only two steps: first it samples N *realizations* of edge existence from the known edge failure probabilities, where $\hat{r}_{en}$ is 1 if edge e succeeds in realization n, and 0 if it fails. These realization variables are used as input to Optimization (4.9), which uses decision variables $\hat{\mathbf{W}}_n$ to represent the *realized* edge discount factor for realization n–that is, $\hat{W}_{en}$ is 1 if edge e is matched and succeeds in realization n and 0 otherwise (see Appendix B.2 for details). Using these decision variables, the objective of Optimization (4.9) includes two terms: the mean matching weight, and the CVaR objective–both approximated using all N samples (i.e., the sample-average approximation). Thus, Optimization (4.9) represents the SAA of (4.8) under the N sampled realizations.

Proposition 3. *Optimization (4.9) is equivalent to the SAA of (4.8) under N edge existence realizations represented by $\hat{r}_{en}$, with*

$$
\begin{aligned}
\min \quad & \frac{1}{N}\sum_{n=1}^{N}\langle \mathbf{w}, \hat{\mathbf{W}}_n\rangle + \gamma\left(d + \frac{1}{\alpha}\frac{1}{N}\sum_{n=1}^{N}\Pi_n\right) \\
s.t. \quad & \hat{W}_{en} = -\sum_{k\in\mathcal{K}(e)} O_{ekn} - \sum_{c\in C}\mathbf{1}(e\in c) z_c v_{cn}, \forall e, n, \\
& \Pi_n \geqslant 0, \forall n, \\
& \Pi_n \geqslant \langle \mathbf{w}, \hat{\mathbf{W}}_n\rangle - d, \forall n, \\
& \{\mathbf{y}, \mathbf{z}\} \in \mathcal{X}, \\
& \{\mathbf{y}, \mathbf{z}\} \in \mathcal{X}', \\
& o_{ekn} \leqslant \hat{r}_{en}, \forall e, k, n, \\
& v_{cn} = \min_{e\in c}\{\hat{r}_{en}\}, \forall c, n,
\end{aligned}
\tag{4.9}
$$

where $\mathcal{X}$ follows the definition in (4.4), and $\mathcal{X}'$ is defined as

$$
\left\{
\begin{array}{l}
\displaystyle\sum_{e\in\delta^-(i)\wedge k\in\mathcal{K}(e)} O_{ekn} \geqslant \sum_{e\in\delta^+(i)} O_{e,k+1,n}, \\
\forall i \in P, k \in \{1,\ldots,L-1\}, n \in \mathcal{N}; \\
O_{ekn} \leqslant y_{ek}, e \in E, k \in \mathcal{K}(e), n \in \mathcal{N}; \\
O_{ekn} \leqslant o_{ekn}, e \in E, k \in \mathcal{K}(e), n \in \mathcal{N}; \\
o_{ekn}, O_{ekn} \in [0,1], e \in E, k \in \mathcal{K}(e), n \in \mathcal{N}; \\
\mathcal{N} = \{1,\ldots,N\}.
\end{array}
\right\}
$$

Optimization (4.9) can be understood by viewing $\langle \mathbf{w}, \hat{\mathbf{W}}_n\rangle$ as the realized edge weight under the n-th realization with matching $\{\mathbf{y}, \mathbf{z}\}$.

5.0 A Study on Distributionally Robust Optimization with Incomplete Joint Data

Stochastic programming is a powerful framework for optimization under uncertainty. It generally assumes that a probability distribution of random variable $\boldsymbol{\xi}$ is available and seeks an optimal solution in terms of expected performance. In practice, the distribution of $\boldsymbol{\xi}$ is often unknown; consequently, data-driven approaches have been proposed to solve this problem. These data-driven methods work well if well-conditioned historical data for $\boldsymbol{\xi}$ are available. However, if $\boldsymbol{\xi}$ is multidimensional, the joint data of $\boldsymbol{\xi}$ are often hard to obtain in the real-world complex data environment due to the following issues:

- Missing data in some dimensions for $\boldsymbol{\xi}$.
- Sharing of data is limited among dimensions representing different components.
- Different sizes of data in different dimensions.

Today missing data is one of the most commonly encountered problems in practical OR problems. For example, in large-scale transportation management systems (TMSs) that are used to monitor the traffic conditions to improve the traffic congestions, the collected traffic data is far from complete. The mobility monitoring program of the Texas Transportation Institute (TTI) reports that after screening erroneous data, TMS data archives can be anywhere from 16% to 93% complete [118]. Another example is that many records in the industrial databases have fields that are not filled. A database of Honeywell studied in [75] is shown to be less than 50% complete.

The most popular method to solve the missing data problem is data imputation, where missing values are imputed prior to running optimization and other analyses on the complete data set. It is highly flexible and convenient, therefore, many variants of data-imputation are proposed, including popular Expectation-Maximization [38], mean impute [83], k-nearest neighbors [126], support vector regression [130], and random forest [122]. We refer readers to [15] for more comprehensive and detailed reviews. However, this type of method does not incorporate the missing data directly into the final decision-making problem and suffers

from two major issues for stochastic optimization. First, theoretical guarantees are hardly obtained. This is because the analysis on the missing data and the derivation of the optimal solutions of the stochastic optimization are conducted separately. Second, separated analysis or estimate-then-optimize methods are known to give sub-optimal solution in many recent studies [84, 37, 63, 49]. One simple example is the newsvendor problem in supply chain management, where the outstocking cost is often much higher than the inventory holding cost. In this type of problems, decision-makers prefer to order "more" products than "less" because of the penalty of having outstocks. However, statistical methods that overlook the decision-making problem often aim to find the "unbiased" estimations towards the unknown demand, which subsequently render sub-optimal decisions [84].

Distributionally robust optimization [37, 71, 53, 49, 137] signifies one powerful modeling paradigm that incorporates the estimation step and the optimization step. It first constructs some ambiguity sets $\mathcal{P}$ based on the available data set; optimization techniques are then proposed to solve these models with respect to the worst-case distributions within the ambiguity sets. However, to the best of our knowledge, existing DRO works have not considered any ambiguity sets based on the incomplete data points. Researchers in [137] do consider a missing data problem encountered in incomplete trajectories data. But their main goal is to reconstruct the missing location-duration path choices. And their ambiguity set is still based on the complete historical data. In this study, we aim to study the distributionally robust optimization models with ambiguity sets constructed directly based on incomplete data. This model hedges against the uncertainties brought by the missing values. We prove that the performance of the proposed DRO framework can be theoretically guaranteed. Additionally, we conduct empirical experiments to show that our DRO model outperforms the classic approaches on incomplete data sets.

The main contribution of this study is to provide a DRO framework for problems with incomplete data set. We assume only partially observed data are available, meaning that the components for each piece of data are randomly missing. Furthermore, we provide finite sample guarantees of our DRO model by introducing the observed information matrix [45] into our analysis. And we prove the statistical consistency results using the properties of maximum likelihood estimation. Tractable reformulations of the presented models and ex-

tensions to two-stage stochastic programming are presented. We also propose and comment on several kinds of possible ambiguity sets. All of them are based on incomplete data. Their properties and reformulations are discussed. Finally, we conduct computational studies to evaluate the performances of the proposed approaches compared to data-imputation-based approaches based on both synthetic and real-world data. We conclude and highlight the following contributions of this study.

1. A new DRO framework based on incomplete data is proposed. It extends the current studies on DRO by proposing ambiguity sets that are constructed directly based on the incomplete data set. W The first two kinds of ambiguity sets utilize two general metrics used in the DRO community. The last kind of ambiguity set is inspired by the special structures existing in our model. We show this ambiguity set is asymptotically optimal in the sense that it contains the true distribution with the highest probability among all ambiguity sets having the same volume.
2. The proposed work is fundamentally different from the popular data-imputation-based methods. It signifies an integrated model that solves the missing data problem and stochastic programming simultaneously instead of following an estimate-then-optimize procedure. By adopting a DRO framework, the proposed models are robust towards the uncertainties of the missing values. Therefore, it greatly improves the out-of-sample performances in applications where the optimal solutions are sensitive to the unknown parameters. We also illustrate this point through computational studies on the multi-item inventory control problem and portfolio optimization.
3. We obtain theoretical guarantees and tractable reformulations for the proposed models. More specifically, we first derive the finite sample guarantees of our model by providing a probabilistic upper bound to their out-of-sample performances. Our analyses are based on the asymptotic normality and the observed information matrix (empirical fisher information). We then prove the statistical consistency guarantee, which means the solution of our model converges to the true optimal in probability when the number of observed data goes to infinity. Finally, we show that these reformulations can be efficiently solved if the cost functions of the original stochastic program are convex, and the feasible regions are convex or mixed-integer linear sets.

5.1 Notations and Background

5.1.1 Preliminary

Throughout this study, we use the following notations and assumptions. A stochastic program can be formulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \mathbb{E}[Q(\mathbf{x},\boldsymbol{\xi})], \tag{5.1}$$

where $Q(\mathbf{x},\boldsymbol{\xi})$ represents a cost function with respect to decision $\mathbf{x}$, and $\mathcal{X}$ represents the feasible region of $\mathbf{x}$. In this study, we study the problem whose random variable $\boldsymbol{\xi}$ has known finite support. Finite discrete support sets are common and studied in many operations research problems [10, 51, 110, 135]. For example, in portfolio selection [89], decision-makers want to select a portfolio from a finite number of scenarios about possible returns to minimize some dis-utility function. Additionally, for problems with continuous random variables, methods like scenario construction [113] are often used to select some representative scenarios to simplify the problems. This also makes those random variables have known finite discrete support sets. Finally, we assume set $\mathcal{X}$ is not empty, and $Q(\mathbf{x},\boldsymbol{\xi})$ is bounded for $\mathbf{x}\in\mathcal{X},\boldsymbol{\xi}\in\Xi$ to avoid discussing trivial cases.

We assume $\boldsymbol{\xi}$ is an I-dimensional vector $\boldsymbol{\xi}\in\Xi\subset\mathbb{R}^I$, and we refer to ξ_i as the *i-th component*. If an observation of $\boldsymbol{\xi}$ is incomplete meaning that the values of some components are missing, it is very difficult to represent these missing values using the notation $\boldsymbol{\xi}$. To this goal, we define two new vectors $\mathbf{s}$ and $\mathbf{s}_{obs}$ to clearly index complete and incomplete data. We first explain the vector $\mathbf{s}$ representing the complete data of $\boldsymbol{\xi}$ as follows. Because the support of $\boldsymbol{\xi}$ is finite, we index each possible value of ξ_i with s_i which is a natural number, i.e. $s_i\in\{1,\cdots,J\}$ without loss of generality. Correspondingly, this gives an I-dimensional index vector $\mathbf{s}\in\mathbb{N}^I$, which has one-to-one correspondence with $\boldsymbol{\xi}$. Therefore, each $\mathbf{s}$ represents one *scenario* of $\boldsymbol{\xi}$. The set containing all scenarios is denoted as $\mathcal{S}$, and its cardinality is $|\mathcal{S}|$. We define another vector $\mathbf{s}_{obs}\in\{0,\mathbb{N}\}^I$ representing the incomplete data. It follows the same definition as $\mathbf{s}$ except that its components can take the value of zero. If its i-th component equals zero, i.e., $s_{obs,i}=0$, it indicates that the value of the i-th component is missing. We

want to emphasize that our model treats the zeros as categorical data. The above procedure does not mean we replace all missing values with 0.

In the rest of the study, we use $\mathbf{s}$ and $\boldsymbol{\xi}$ interchangeably. We use the following common signs for different types of convergence of random variables. 1) Converge in probability: $\xrightarrow{p}$. 2) Converge almost surely: $\xrightarrow{a.s.}$. 3) Converge in distribution: $\xrightarrow{d}$.

Terminologies in data-driven stochastic programming. We briefly review some commonly used terminologies and notations in data-driven stochastic programming.

We use $\mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\}$ to denote a joint distribution of $\mathbf{s}$, and $\mathbf{P}^* = \{\mathrm{P}^*(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\}$ represents the true unknown joint distribution of $\mathbf{s}$. We call $\mathbf{P}^*$ the *true distribution*. Then the *true optimal cost*, O^*, is defined as

$$O^* = \min_{\mathbf{x} \in \mathcal{X}} \quad \mathbb{E}_{\mathbf{P}^*}[Q(\mathbf{x}, \mathbf{s})],$$

where the expectation is taken with respect to the true distribution $\mathbf{P}^*$. Any corresponding optimal solution $\mathbf{x}^*$ is called the *true optimal solution*.

Suppose a data set of $\mathbf{s}$ contains N i.i.d. data points in total. A *data-driven solution* for (5.1) is defined as a feasible solution $\hat{\mathbf{x}}_N \in \mathcal{X}$ based on this data set. The *out-of-sample performance* of $\hat{\mathbf{x}}_N$ is defined as

$$\mathbb{E}_{\mathbf{P}^*}[Q(\hat{\mathbf{x}}_N, \mathbf{s})]. \tag{5.2}$$

However, the out-of-sample performance cannot be directly evaluated because $\mathbf{P}^*$ is unknown. Therefore, we seek an upper bound $\hat{O}_N$ to obtain the performance guarantees of the type

$$\mathrm{P}(\mathbb{E}_{\mathbf{P}^*}[Q(\hat{\mathbf{x}}_N, \mathbf{s})] \leqslant \hat{O}_N) \geqslant 1 - \alpha, \tag{5.3}$$

where $\alpha \in (0, 1)$. And we refer to $1-\alpha$ as the *reliability*, which measures the probability that the out-of-sample performance is bounded by $\hat{O}_N$. We will explicitly define $\hat{O}_N$ later. The corresponding Equation (5.3) is directly referred to as the *finite sample guarantee* throughout the rest of the study.

5.1.2 Distributionally Robust Optimization Framework with Missing Data

DRO model begins by defining an ambiguity set $\mathcal{P}$ for the unknown joint distribution based on the observed incomplete data set; $\mathcal{P}$ contains different possible joint distributions $\mathbf{P}$ that includes the true distribution $\mathbf{P}^*$ with a prescribed probability, $\mathrm{P}(\mathbf{P}^* \in \mathcal{P}) \geqslant 1-\alpha$, for $0 < \alpha < 1$. The parameter α can be used to reflect the conservatism of the model. After designing $\mathcal{P}$, a distributional robust approach finds the best solution, assuming the worst-case distribution within $\mathcal{P}$ as shown in (5.4),

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{P}\in\mathcal{P}} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}). \tag{5.4}$$

Most existing DRO frameworks construct the ambiguity set to hedge against the uncertainties brought by the finite sample size based on different types of concentration inequalities. The key difference of the DRO framework studied in this study is that the ambiguity set includes the uncertainties coming from the incomplete data set. We point out the benefits of our model comparing to the other methods of solving missing data problems as follows.

Benefits of Model (5.4). Compared to the traditional approaches of solving missing data problems, Model (5.4) is unique in that it integrates the missing data directly into the final decision-making process. This is achieved by combining the information of the incomplete data set (captured by $\mathbb{P}$) and the derivation of $\mathbf{x} \in \mathcal{X}$ into one optimization problem. Classical methods use separate analysis [46, 123], which first estimates $\mathbf{P}$ and then derives the optimal $\mathbf{x}$. It is recognized in literature [84, 63] that separate analyses ignoring the potential estimation errors can lead to sub-optimal solutions and/or heuristic methods with no performance guarantees.

In the following, we propose a method to construct the ambiguity set based on the incomplete data set; the obtained optimal decision from the incomplete data not only allows the theoretical guarantees but also improves the performance in practical problems.

5.2 Main Model

The main idea of our model is to combine the maximum likelihood estimation into the classical DRO framework (5.4). We first discuss our model under one specific ambiguity set, as will be introduced later. Then, we introduce the concepts of asymptotic normality and the observed information matrix to obtain the theoretical guarantees, which contain two main points. First, we prove the consistency result in Section 5.2.2. The consistency follows the definition in [127] meaning that the solution of our model converges in probability to the *true optimal solution* as the observed data size N goes to infinity. Second, we investigate the finite sample guarantee (5.3) in Section 5.2.3, where it shows the objective value of our model serves as a probabilistic upper bound of the out-of-sample performance. Finally, we discuss several other kinds of ambiguity sets with tractable reformulations in Section 5.2.4.

Assumption and notations. Before introducing our model, we recall the setting of the incomplete data and make one technical assumption. The incomplete data, or more intuitively the partially observed data, means that the values of the components are missing at random, and the number of observed dimensions for each data point ranges randomly from 1 to I.

Recall that vector $\mathbf{s}_{obs} \in (\{0\}, \mathbb{N})^I$ can be used to represent the partially observed data, where its i-th component $s_{obs,i} = 0$ indicates that the value of the i-th component is missing. Correspondingly, we further define an indicator vector $\boldsymbol{\phi}_{obs} \in \{0,1\}^I$ to indicate the observed dimensions of $\mathbf{s}_{obs}$ ($\phi_{obs,i} = 1$ represents the i-th dimension is observed; 0 otherwise). Finally, for a fixed $\mathbf{s}_{obs}$ we define $\mathcal{S}(\mathbf{s}_{obs})$ as a set that

$$\mathcal{S}(\mathbf{s}_{obs}) = \{\mathbf{s} \in \mathcal{S} | \mathbf{s} \odot \boldsymbol{\phi}_{obs} = \mathbf{s}_{obs}\}, \tag{5.5}$$

where $\odot$ represents the component-wise multiplication. Intuitively, set $\mathcal{S}(\mathbf{s}_{obs})$ contains all the complete data $\mathbf{s}$ that match the observed components in $\mathbf{s}_{obs}$. The n-th partially observed data is denoted as $\hat{\mathbf{s}}_{n,obs}$ ($n = 1, \cdots, N$), which is the n-th realization of $\mathbf{s}_{obs}$.

We make a common assumption on the missing data mechanism called missing at random (MAR) [83]. This assumption intuitively means that the missing probability of a missed value

is independent of this value itself. Mathematically, this means the probability of observing one $\mathbf{s}_{obs}$ given $\mathbf{s}$ is fixed for all $\mathbf{s} \in \mathcal{S}(\mathbf{s}_{obs})$ as shown below.

$$\mathrm{P}(\mathbf{s}_{obs}|\mathbf{s}) = \mathrm{P}(\mathbf{s}_{obs}|\mathbf{s}'), \quad \mathbf{s}, \mathbf{s}' \in \mathcal{S}(\mathbf{s}_{obs}) \tag{5.6}$$

If the missing probability of a component depends on the values of itself, the missing data mechanism itself has to be known or modeled explicitly to recover the missing values.

5.2.1 Ambiguity Set $\mathcal{P}'$ based on Partially Observed Data

We propose an ambiguity set $\mathcal{P}'$ centering around a nominal distribution ($\hat{\mathbf{P}}$) based on some metric $\|\dot{\|}$ as shown in (5.7).

$$\mathbb{P}' = \{\mathbf{P} \text{ is a distribution function. } : \|\mathbf{P} - \hat{\mathbf{P}}\| \leqslant \tau\}, \tau > 0. \tag{5.7}$$

Before explaining the details about (5.7), we want to point out that this nominal distribution is chosen as one optimal estimator of the maximum likelihood estimation (MLE) based on the incomplete data set. The ambiguity set $\mathcal{P}'$ has a user tunable parameter τ called *distance tolerance* to control the robustness. We will further prove the relationship between τ and out-of-sample performance. This ambiguity set enjoys several good properties. 1) The nominal distribution, or the center of the ambiguity set equivalently, represents a joint distribution that has the largest likelihood to generate the observed incomplete data set; 2) Ambiguity set $\mathcal{P}'$ is consistent with the popular metric-based/likelihood-based ambiguity sets [47, 99, 65, 49]. Because in these works, researchers define the center of the ambiguity set as the empirical distribution, which can be viewed as one optimal solution of (nonparametric) MLE [97]; 3) We will show that $\mathcal{P}'$ allows us to obtain the finite sample guarantee of Model (5.4) directly through the distance tolerance and ensure the obtained solution to converge to the true optimal as data size increases even though the data are partially observed.

Formulation of $\mathcal{P}'$ with L1 norm In what follows, we define the ambiguity set mathematically in (5.8) by using the L1 norm and will focus on this ambiguity set in the next two subsections. Other kinds of ambiguity sets will be further discussed in Section 5.2.5. The L1 norm is special in this problem because the random variables with known discrete support

often represent categorical variables instead of ordinal ones, like the portfolio selection and data of transportation routes introduced before. And L1 norm is the only meaningful norm here when the data is nominal. More details about L1 norm on this point are listed in C.1 for interested readers.

$$\mathcal{P}' = \left\{ \mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{l} \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s} \in \mathcal{S}, \\ \sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau. \end{array} \right\} \tag{5.8}$$

Set $\mathcal{P}'$ follows the general definition in (5.7). It contains all the joint distributions that are close to a nominal distribution $\hat{\mathbf{P}}$, and a *distance tolerance* τ controls the conservatism. Intuitively, we call the nominal distribution $\hat{\mathbf{P}}$ as the *center* of the ambiguity set.

We obtain the center $\hat{\mathbf{P}}$ of the ambiguity set through MLE. In Proposition 1, we show that this procedure is equivalent to solving a convex optimization.

Definition 1. *The nominal distribution $\hat{\mathbf{P}}$ is one optimal solution of the following optimization*

$$\begin{aligned} \max_{\mathbf{P}} \quad & \sum_{n=1}^{N} \ln \left[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s}) \right] \\ s.t. \quad & \mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}, \\ & \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \end{aligned} \tag{5.9}$$

where we use $\mathcal{S}(\hat{\mathbf{s}}_{n,obs})$ to denote all the $\mathbf{s} \in \mathcal{S}$ that match the observed part of the n-th data $\hat{\mathbf{s}}_{n,obs}$ as defined in (5.5).

Proof. Please refer to C.2 for detailed proofs. □

Optimization (5.9) is a maximization of a concave function under linear constraints, which can be solved through convex optimization techniques. For example, Optimization (5.9) can be solved efficiently through projected gradient descent [22] or general convex solvers like CVXOPT.

5.2.2 Consistency

In this section, we validate that the optimal solution of (5.4) under the ambiguity set (5.8) converges to the true optimal solution in probability as the data size increases to infinity. Intuitively, this result implies that the obtained optimal decision of our model is as good as a decision made under the complete information when the available data size is large enough.

Mathematically, we aim to prove that given the distance tolerance τ converging to zero, and the number of data samples, N, goes to infinity, the solution to the proposed DRO model converges in probability to the true optimal solution. We establish this desired result in Theorem 1. To prove Theorem 1, we rely on first proving Lemma 1 and Lemma 2.

Intuitively, Lemma 1 states that a unique maximum exists for (5.9) when the data size $N \to \infty$, and the corresponding solution of this unique maximum is the true joint distribution $\mathbf{P}^*$.

Lemma 1. *Define* $F(\mathbf{s}_{obs}|\mathbf{P}) = \sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s})$. *Let*

$$L(\mathbf{P}) = \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}) = \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right],$$

where the expectation is taken over the probability mass function of $\mathbf{s}_{obs}$. *Then,* $L(\mathbf{P})$ *attains its maximum uniquely at* $\mathbf{P}^*$, *where* $\mathbf{P}^*$ *represents the true joint distribution.*

Proof. Please refer to C.3 for detailed proofs. □

Based on Lemma 1, Lemma 2 further proves that the center $\hat{\mathbf{P}}$ of the ambiguity set $\mathcal{P}'$ converges in probability to the true joint distribution $\mathbf{P}^*$ as the data size increases to infinity. The main idea behind Lemma 2 is to prove the uniform law of large numbers for the objective function of (5.9), which is achieved by constructing a dominating function of it.

Lemma 2. *When the number of samples,* N, *goes to infinity, the solution* $\hat{\mathbf{P}}$ *of (5.9) converges in probability to* $\mathbf{P}^*$.

Proof. Please refer to Appendix C.4 for detailed proofs. □

Finally, we establish the consistency results by using Lemma 1 and 2. We use $\tau(N)$ to imply that the distance tolerance τ is a function of the sample size N.

Theorem 1 (Consistency)**.** *Let $\tau(N)$ satisfy*

$$\lim_{N\to\infty} \tau(N) = 0.$$

Assume $\hat{\mathbf{x}}_N$ and $\hat{O}_N$ are one optimal solution and the corresponding objective value of the proposed model, i.e.

$$\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathbf{P}\in\mathcal{P}'} \mathbb{E}[Q(\mathbf{x}, \mathbf{s})],$$

under the incomplete data set of size N. Let $\mathbf{x}^*$ and O^* be one optimal solution and the objective value of

$$\min_{\mathbf{x}\in\mathcal{X}} \mathbb{E}_{\mathbf{P}^*}[Q(\mathbf{x}, \mathbf{s})],$$

where the expectation is taken with respect to the true unknown joint distribution. Then $\hat{O}_N \xrightarrow{p} O^*$. Furthermore, if the maximizer $\mathbf{x}^*$ is unique, then $\hat{\mathbf{x}}_N \xrightarrow{p} \mathbf{x}^*$.

Proof. Please refer to C.5 for detailed proofs. □

5.2.3 Finite sample guarantee

In this section, we establish the finite sample guarantee that is defined in (5.3). We show that the optimal value of the objective function (5.4) serves as a probabilistic upper bound of the out-of-sample performance under the finite number of incomplete data. The main steps in our proofs are as follows. We first prove that the deviation between the true distribution and the nominal distribution can be approximated by a normal distribution asymptotically. Additionally, this normal distribution has mean zero and a variance-covariance matrix that depends only on the true distribution function and the missing data mechanism. Because we assume the true distribution function and the detailed missing probabilities of components are unknown, this variance-covariance matrix cannot be directly obtained. We propose to obtain it via observed information matrix [45, 85, 44], which can be viewed as a data-driven version of this variance-covariance matrix. Then, we obtain the finite sample guarantee based on the obtained asymptotic normal distribution.

Notations. We assume set

$$\mathcal{S}^+ = \{\mathbf{s} \in \mathcal{S} | \hat{\mathrm{P}}(\mathbf{s}) > 0\} = \{\mathbf{s}^{\{1\}}, \cdots, \mathbf{s}^{\{b\}}\}$$

contains b different $\mathbf{s}$. Their corresponding $\hat{\mathrm{P}}(\mathbf{s})$ are represented with a vector $\mathbf{p} = [p_1, \cdots, p_b]$; we also define $\{a_1, \cdots, a_b\}$ to be the corresponding number of observations for these b different $\mathbf{s}$ in the data set. Suppose we also observe q different incomplete data (the number of observed dimensions is less than I) $\mathbf{s}_{obs}^{\{j\}}$, $j = 1, \cdots, q$, and each $\mathbf{s}_{obs}^{\{j\}}$ appears b_j times in the data set. We define a vector $\boldsymbol{\delta}_j \in \{0,1\}^b$, $j = 1, \cdots, q$, where $\delta_{ji} = 1$ if the observed dimensions of $\mathbf{s}_{obs}^{\{j\}}$ match that of $\mathbf{s}^{\{i\}}$, i.e.

$$\mathbf{s}^{\{i\}} \odot \boldsymbol{\phi}_{obs}^{\{j\}} = \mathbf{s}_{obs}^{\{j\}}.$$

Recall that $\boldsymbol{\phi}_{obs} \in \{0,1\}^I$ was defined as the indicator vector of observed dimensions.

Proposition 1 states that for a fixed distance tolerance τ, the deviation between the true joint distribution $\mathbf{P}^*$ and the center $\hat{\mathbf{P}}$ of the ambiguity set is less than a distance tolerance τ with probability at least α, where α is a function of τ and the observed data. The main idea behind Proposition 1 is to show that the difference, i.e., $\mathbf{P}^* - \hat{\mathbf{P}}$, can be well approximated by a random variable with distribution $N(\mathbf{0}, \boldsymbol{\Sigma})$. More details are provided in the appendix.

Proposition 1. *Suppose the true joint distribution is* $\mathbf{P}^*$. *The nominal joint distribution* $\hat{\mathbf{P}}$ *satisfies* $\sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau$, *with probability at least* α, *where*

$$\alpha = \int_{[-\frac{\tau}{b}, \frac{\tau}{b}]^b} \frac{1}{\sqrt{(2\pi)^b |\boldsymbol{\Sigma}^*|}} \exp\left(-\frac{1}{2}\mathbf{x}^T \boldsymbol{\Sigma}^{*-1} \mathbf{x}\right) d\mathbf{x},$$

and $\boldsymbol{\Sigma}^* \in \mathbb{R}^{b\times b}$ *is the asymptotic variance-covariance matrix of* $\hat{\mathbf{P}} - \mathbf{P}^*$. *We obtain* $\boldsymbol{\Sigma}^*$ *empirically, which is denoted as* $\boldsymbol{\Sigma}$ *below.*

$$\boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_{b-1} & \mathbf{v}_b \\ \mathbf{v}_b^T & var(b) \end{bmatrix}$$

with $\boldsymbol{\Sigma}_{b-1} \in \mathbb{R}^{b-1\times b-1}$, $\mathbf{v}_b \in \mathbb{R}^{b-1}$. *For matrix* $\boldsymbol{\Sigma}$, *we define* $var(b) = \mathbf{1}^T \boldsymbol{\Sigma}_{b-1} \mathbf{1}$ *and* $\mathbf{v}_b = \boldsymbol{\Sigma}_{b-1}\mathbf{1}$, *where* $\mathbf{1}$ *denotes one* $(b-1)$*-dimensional column vector whose all elements equal to* 1. *In addition,*

$$\mathbf{\Sigma}_{b-1}^{-1} = diag(\frac{a_1}{p_1^2}, \cdots, \frac{a_{b-1}}{p_{b-1}^2}) + \frac{a_{b-1}}{p_{b-1}^2} \begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} \psi_{11} & \cdots & \psi_{1,b-1} \\ \vdots & \ddots & \vdots \\ \psi_{b-1,1} & \cdots & \psi_{b-1,b-1} \end{bmatrix},$$

where

$$\psi_{ik} = \sum_{j=1}^{q} \frac{b_j(\delta_{ji} - \delta_{jb})(\delta_{jk} - \delta_{jb})}{(\boldsymbol{\delta}_j^T \mathbf{p})^2}, 1 \leqslant i, k \leqslant b-1.$$

Proof. Please refer to Appendix C.6 for detailed proofs. □

Following the results of Proposition 1, we directly establish the finite sample guarantee, which is defined in (5.3), in Theorem 2 below.

Theorem 2 (Finite sample guarantee)**.** *Suppose* $\mathbf{P}^*$ *represents the true joint distribution, and* $\hat{\mathbf{x}}_N$ *and* $\hat{O}_N$ *are one optimal solution and the corresponding objective value of Model (5.4) with ambiguity set* $\mathcal{P}'$ *under* N *data. Then, we have*

$$\mathrm{P}\left[\mathbb{E}_{\mathbf{P}^*}[Q(\hat{\mathbf{x}}_N, \mathbf{s})] \leqslant \hat{O}_N\right] \geqslant \alpha,$$

where α *follows the definition in Proposition 1.*

So far we have seen that the proposed ambiguity set allows our DRO model with favorable asymptotic and finite sample guarantees when only incomplete data sets are available. In the next subsection, we show that this DRO model is also mathematically tractable for a variety number of cases.

5.2.4 Worst-case reformulation

In a classical stochastic program, the cost function is defined as

$$Q(\mathbf{x}, \mathbf{s}) = \mathbf{q}(\mathbf{s})^T \mathbf{x},$$

and the feasible region $\mathcal{X}$ is a polytope. Then, solving our model is equivalent to solving a linear program. In general, if $Q(\mathbf{x}, \mathbf{s})$ is convex with respect to $\mathbf{x}$, and $\mathcal{X}$ is a convex set or mixed-integer linear set [34], our model is mathematically tractable. We conclude the results in Proposition 2 and its proofs.

Proposition 2. *Optimization*

$$\min_{\mathbf{x} \in \mathcal{X}} \quad \max_{\{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} \in \mathcal{P}'} \sum_{\mathbf{s} \in S} \mathrm{P}(\mathbf{s}) Q(\mathbf{x}, \mathbf{s})$$

is equivalent to

$$\begin{aligned} \min_{\mathcal{B}} \quad & \gamma + e\tau + \sum_{\mathbf{s} \in \mathcal{S}} (w_{\mathbf{s}} - l_{\mathbf{s}}) \hat{\mathrm{P}}(\mathbf{s}) \\ s.t. \quad & Q(\mathbf{x}, \mathbf{s}) \leqslant \gamma + w_{\mathbf{s}} - l_{\mathbf{s}}, \forall \mathbf{s} \in \mathcal{S}, \\ & w_{\mathbf{s}} + l_{\mathbf{s}} - e = 0, \forall \mathbf{s} \in \mathcal{S}, \\ & \mathbf{x} \in \mathcal{X}, \end{aligned} \tag{5.10}$$

where $\mathcal{B} = \{\mathbf{x}, \gamma, w_{\mathbf{s}} \geqslant 0, l_{\mathbf{s}} \geqslant 0, e \geqslant 0\}$.

Proof. Please refer to C.7 for detailed proofs. □

Finally, the above results can be extended to two-stage stochastic programming cases, where related results are summarized in C.8.

5.2.5 Other ambiguity sets

In general cases, other metrics can be used in (5.7) to construct the ambiguity set. These ambiguity sets enjoy the same asymptotic result (consistency) and similar finite-sample guarantees. This is because the center of the ambiguity set is fixed to be $\hat{\mathbf{P}}$ in (5.7), which implies the difference between the center and the true joint distribution can still be well approximated by a normal random variable asymptotically. We discuss several kinds of ambiguity sets in this subsection and their reformulations.

1. **f-divergence-based ambiguity set.** In this case, we define the distance between $\mathbf{P}$ and $\hat{\mathbf{P}}$, i.e., $\|\mathbf{P}-\hat{\mathbf{P}}\|$, to be an f-divergence[81], which is a function $D_f(\mathbf{P}||\hat{\mathbf{P}})$ that measures the difference between two probability distributions. We list some f-divergence examples below in Table 4. We want to point out that the L1 norm discussed before can also be viewed as a special case of f-divergence (total variation).

$$\mathcal{P}' = \left\{ \mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{ll} \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, & \forall \mathbf{s} \in \mathcal{S}, \\ D_f(\mathbf{P}||\hat{\mathbf{P}}) \leqslant \tau. \end{array} \right\} \tag{5.11}$$

Table 4: Some f-divergence examples.

Divergence	$D_f(\mathbf{P}\|\|\hat{\mathbf{P}})$
Kullback-Leibler	$\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) \log(\frac{\mathrm{P}(\mathbf{s})}{\hat{\mathrm{P}}(\mathbf{s})})$
Burg entropy	$\sum_{\mathbf{s}\in\mathcal{S}} \hat{\mathrm{P}}(\mathbf{s}) \log(\frac{\hat{\mathrm{P}}(\mathbf{s})}{\mathrm{P}(\mathbf{s})})$
J-divergence	$\sum_{\mathbf{s}\in\mathcal{S}} [\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})] \log(\frac{\mathrm{P}(\mathbf{s})}{\hat{\mathrm{P}}(\mathbf{s})})$
χ^2-distance	$\sum_{\mathbf{s}\in\mathcal{S}} \frac{(\mathrm{P}(\mathbf{s})-\hat{\mathrm{P}}(\mathbf{s}))^2}{\mathrm{P}(\mathbf{s})}$
Hellinger distance	$\sum_{\mathbf{s}\in\mathcal{S}} (\sqrt{\mathrm{P}(\mathbf{s})} - \sqrt{\hat{\mathrm{P}}(\mathbf{s})})^2$
Total variation	$\sum_{\mathbf{s}\in\mathcal{S}} \lvert\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})\rvert$

The DRO with f-divergence-based ambiguity sets is mostly tractable, which are presented in [12]. We refer readers to it for tractable reformulations by replacing the center of the

uncertainty set discussed in their study with the nominal distribution defined in our study. Finally, we can obtain a similar finite-sample guarantee based on the results in Proposition 1. We conclude it as follows.

Corollary 1 (Finite sample guarantee.)**.** *Suppose* $\mathbf{P}^*$ *represents the true joint distribution, and* $\hat{\mathbf{x}}_N$ *and* $\hat{O}_N$ *are one optimal solution and the corresponding objective value of Model (5.4) with ambiguity set based on f-divergence* $D_f(\mathbf{P}||\hat{\mathbf{P}})$ *under* N *data. Then, we have*

$$\mathrm{P}\left[\mathbb{E}_{\mathbf{P}^*}[Q(\hat{\mathbf{x}}_N, \mathbf{s})] \leqslant \hat{O}_N\right] \geqslant \alpha.$$

And α *is defined as*

$$\alpha = \int_{\mathbf{x}\in\mathcal{X}'} \frac{1}{\sqrt{(2\pi)^b|\mathbf{\Sigma}|}} \exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right) d\mathbf{x},$$

where $\mathcal{X}' = \{\mathbf{x} = (\mathbf{P} - \hat{\mathbf{P}})_{\mathcal{S}^+} \Big| D_f(\mathbf{P}||\hat{\mathbf{P}}) \leqslant \tau\}$*. We use* $(\mathbf{P} - \hat{\mathbf{P}})_{S+}$ *to denote that we only keep the dimensions that appear in set* $\mathcal{S}^+$*. Therefore,* $\mathbf{x}$ *has* b *components/dimensions in total.*

2. **p-Wasserstein distance-based ambiguity set.** We define $\|\mathbf{P} - \hat{\mathbf{P}}\|$ as $W_p(\mathbf{P}, \hat{\mathbf{P}})$ representing the p-Wasserstein distance defined in Definition 2.

$$\mathcal{P}' = \left\{ \mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{ll} \displaystyle\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, & \forall \mathbf{s} \in \mathcal{S}, \\ W_p(\mathbf{P}, \hat{\mathbf{P}}) \leqslant \tau. \end{array} \right\} \tag{5.12}$$

Definition 2 (p-Wasserstein metric)**.** *The p-Wasserstein distance (*$p \in [1, +\infty)$*) between distribution* $\mathbf{P}$ *and* $\hat{\mathbf{P}}$ *supported on* Ξ *is defined as*

$$W_p(\mathbf{P}, \hat{\mathbf{P}}) := \inf\{(\int_{\Xi^2} d'^p(\boldsymbol{\xi}, \boldsymbol{\xi}')\Pi(d\boldsymbol{\xi}, d\boldsymbol{\xi}'))^{\frac{1}{p}} : \Pi \text{ is a joint distribution of } \boldsymbol{\xi} \text{ and } \boldsymbol{\xi}' \text{ with marginals } \mathbb{P} \text{ and } \mathbb{P}'\},$$

where d' *is a metric on* Ξ*.*

Wasserstein distance is studied widely in DRO literure [49, 54] recently. It is shown to have superior performances with tractable reformulation in a number of cases. We conclude its general reformulation for our model below.

Proposition 3 ([53]). *With p-Wasserstein distance-based ambiguity set, Model (5.4) is equivalent to*

$$\min_{\mathbf{x}\in\mathcal{X},\lambda>0} \quad \{\lambda\tau^p + \sum_{\mathbf{s}\in\mathcal{S}} \hat{P}(\mathbf{s})[\sup_{\mathbf{s}'\in\mathcal{S}}\{Q(\mathbf{x},\mathbf{s}') - \lambda d'^p(\mathbf{s}',\mathbf{s})\}]\} \tag{5.13}$$

As shown above, the tractability of using p-Wasserstein distance-based ambiguity set depends on if we can solve

$$\sup_{\mathbf{s}'\in\mathcal{S}}\{Q(\mathbf{x},\mathbf{s}') - \lambda d'^p(\mathbf{s}',\mathbf{s})\}$$

efficiently. This further depends on the structure of the support $\mathcal{S}$ and the choice of metric d', which is beyond the scope of this study.

3. **Ellipsoid ambiguity set.** We define a family of ambiguity sets called ellipsoid ambiguity sets. We make one regularization assumption here to simplify the notations used. We assume the nominal distribution $\hat{\mathbf{P}}$ contains no zero-value components. We will explain later why this does not limit the generality. In this case, we measure the distance between $\mathbf{P}$ and $\hat{\mathbf{P}}$, i.e., $\|\mathbf{P}-\hat{\mathbf{P}}\|$, through

$$(\mathbf{P}-\hat{\mathbf{P}})^T\mathbf{\Sigma}^{-1}(\mathbf{P}-\hat{\mathbf{P}}).$$

Recall that $\mathbf{\Sigma}^{-1}$ is the inverse of the observed information matrix derived in Section 5.2.3.

$$\mathcal{P}' = \left\{ \mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}\} : \begin{array}{l} \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s}\in\mathcal{S}, \\ (\mathbf{P}-\hat{\mathbf{P}})^T\mathbf{\Sigma}^{-1}(\mathbf{P}-\hat{\mathbf{P}}) \leqslant \tau. \end{array} \right\} \tag{5.14}$$

First of all, we want to point out although many existing studies have also proposed ellipsoid ambiguity sets [37, 132, 80], they mostly define it through the support by assuming the first (μ) and second moment information (Σ_0) are known or partially known.

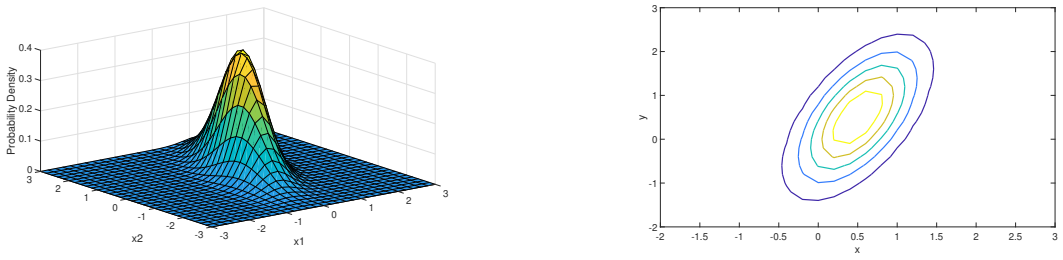
For example, one popular ambiguity set is defined to include all the distributions such that the corresponding random variable $\boldsymbol{\xi}$ satisfies

$$(\boldsymbol{\xi} - \boldsymbol{\mu})^T \Sigma_0 (\boldsymbol{\xi} - \boldsymbol{\mu}) \leqslant \tau.$$

The ellipsoid in our problem is different from them because it is defined in the probability space.

Intuitively, the ellipsoid ambiguity set defined here represents one asymptotic "optimal" ambiguity set in our model. This is because we proved the difference, $(\mathbf{P} - \hat{\mathbf{P}})$, can be well approximated by a normal random variable. The boundary of the ellipsoid ambiguity set is exactly one contour of this normal distribution. Therefore, this ambiguity set contains the true distribution with the highest probability among all the ambiguity sets that have the same volume. We illustrate the above result in a 2-dimensional example. In Figure 3a, we plot the corresponding PDF of $\hat{\mathbf{P}} - \mathbf{P}^*$, and its corresponding contours are plotted in Figure 3b. The potential benefit of this ambiguity set is that it allows small variations for components with small variance and large variations for components with large variance. Additionally, if one component in $\hat{\mathbf{P}}$ remains zero asymptotically, it means this scenario will never happen and can be safely removed from the support. Therefore, the regularization assumption we made before does not limit the generality.

Figure 3: PDF of the difference between $\mathbf{P}^*$ and $\hat{\mathbf{P}}$ and its contours.



(a) PDF (b) Contour

Finally, we also prove that DRO with this kind of ellipsoid ambiguity set is still tractable. The result is summarized in Proposition 4, which is a second-order conic programming.

Proposition 4. *The DRO with ellipsoid ambiguity set is equivalent to*

$$\begin{aligned} \min_{\mathbf{x},y,\boldsymbol{\beta},\gamma} \quad & y + \hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x}) + \hat{\mathbf{P}}^T\boldsymbol{\beta} \\ s.t. \quad & \mathbf{x} \in \mathcal{X}, \\ & \beta_{\mathbf{s}} \geqslant 0, \forall \mathbf{s} \\ & y \geqslant 0, \\ & y^2 \geqslant \tau(\mathbf{Q}(\mathbf{x}) - \boldsymbol{\gamma} + \boldsymbol{\beta})^T\boldsymbol{\Sigma}(\mathbf{Q}(\mathbf{x}) - \boldsymbol{\gamma} + \boldsymbol{\beta}), \end{aligned} \tag{5.15}$$

where we use $\boldsymbol{\gamma}$ to represent an $|S|$-dimension vector whose all elements equal γ.

Proof. Please refer to C.9 for detailed proofs. □

5.3 Computational Study

We also conduct computational studies to validate the superiority of the proposed model in some real-world applications. We first study a multi-item two-stage inventory control problem based on the synthetic data in Section 5.3.1. In this experiment, we show the improvements in the out-of-sample performances of our models compared to a data-imputation-based approach. In Section 5.3.2, we study a portfolio optimization problem based on the real-world historical returns of exchange-traded funds (ETFs) and the US central bank (FED) rate of return from 2006 to 2016 [21], where our approaches also achieve better out-of-sample performance consistently.

5.3.1 Multi-item two-stage inventory control problem

5.3.1.1 Problem formulation Multi-item two-stage inventory control problem [8] can be described in two periods. At the beginning of the second period, the decision-maker observes m product demands $\mathbf{b} \in \mathbb{R}^m$ following an unknown joint distribution $\mathbf{P}$. The demand of product i can be served by either placing an order with unit cost a_{1i} in the first period, which will be delivered at the beginning of the second period, or by placing an order with unit cost a_{2i} ($a_{2i} > a_{1i}$) at the beginning of the second period which will be delivered

immediately. The excess product units after the second period incur a unit holding cost h_i. If there is a shortfall in the available quantity, then a unit outstocking cost p_i is incurred. We denote the level of inventory for each item by $I_i(\mathbf{b})$, $i = 1, \cdots, m$. The decision-maker wishes to determine the first-stage and second-stage ordering quantities, x_{1i} and $x_{2i}(\mathbf{b})$, for all products to minimize the total ordering, backlogging and holding costs. The problem can be formulated as the following two-stage optimization problem in (5.16).

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} a_{1i}x_{1i} + \mathbb{E}\left[\sum_{i=1}^{m} a_{2i}x_{2i}(\mathbf{b}) + \sum_{i=1}^{m} \max\{-p_i I_i(\mathbf{b}), h_i I_i(\mathbf{b})\}\right] \\ s.t. \quad & I_i(\mathbf{b}) = x_{1i} + x_{2i}(\mathbf{b}) - b_i, i = 1, \cdots, m, \\ & x_{1i} \geqslant 0, i = 1, \cdots, m, \\ & x_{2i} \geqslant 0, i = 1, \cdots, m, \end{aligned} \tag{5.16}$$

Correspondingly, our model for (5.16) under the incomplete data is formulated as

$$\begin{aligned} \min \quad & \sum_{i=1}^{m} a_{1i}x_{1i} + \max_{\mathbf{P}\in\mathcal{P}} \mathbb{E}\left[\sum_{i=1}^{m} a_{2i}x_{2i}(\mathbf{b}) + \sum_{i=1}^{m} \max\{-p_i I_i(\mathbf{b}), h_i I_i(\mathbf{b})\}\right] \\ s.t. \quad & I_i(\mathbf{b}) = x_{1i} + x_{2i}(\mathbf{b}) - b_i, i = 1, \cdots, m, \\ & x_{1i} \geqslant 0, i = 1, \cdots, m, \\ & x_{2i} \geqslant 0, i = 1, \cdots, m, \end{aligned} \tag{5.17}$$

where the ambiguity set $\mathcal{P}$ is defined through (5.8). Its tractable reformulation is discussed in C.8.

5.3.1.2 Computational results We compare our models with the data-imputation-based approach. We consider $m = 3$ products whose demands are defined as

$$D_i = D_{1i} + D_0, \quad i = 1, 2, 3,$$

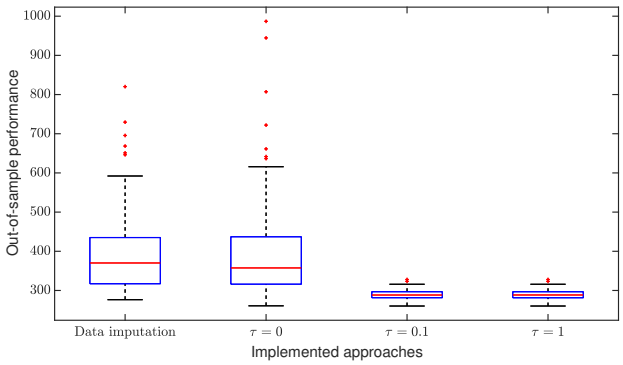
where D_0 follows a Poisson distribution, Pois(10), and D_{1i} follows a Poisson distribution Pois$(5i + 5)$. We first randomly generate 300 samples to serve as the known support in this problem. Then, we conducted 100 experiments by generating different random samples. In each experiment, we randomly generate 10 joint data and assume each component will be

missed with a probability 0.5. We set the inventory parameters as follows, $a_{i1} = 2$, $a_{2i} = 300$, $h_i = 1$, and $p_i = 240$.

To evaluate the quality of our models, we draw the boxplots of the out-of-sample performances of the proposed models under different deviation tolerance, τ. We also compare them with one data-imputation-based approach. Because the support of the random variable is assumed to be known, each incomplete data is imputed to the nearest support with respect to the L1 norm in this approach. Then, the optimal solution is obtained based on the empirical distribution of the complete data set. We conclude the results in Figure 4.

Results. With partially observed data, our model has similar performances as the data-imputation-based approach when $\tau = 0$. However, when $\tau = 0.1$ or 1, the performances are improved significantly, where the 0.75 percentile is greatly reduced and, 0.25 percentile remains stable. In general, our model achieves much lower average inventory costs.

Figure 4: Out-of-sample performance.



5.3.2 Portfolio optimization

We also benchmark the proposed approaches through one real-world data set: portfolio optimization, where the missing data is a common problem [100, 124]. In the subsequent numerical experiments, we show that our approach outperforms the data-imputation-based approach on the real-world data set. The experimental data come from the historical returns of exchange-traded funds (ETFs) and the US central bank (FED) rate of return [21]. The

data set roughly covers ten years, from January 2006 to December 2016, and includes 2520 data for each asset.

We define the model of portfolio optimization in Section 5.3.2.1. The experimental settings are presented in Section 5.3.2.2, which introduces the data set and the implemented methods. Section 5.3.2.6 concludes the results and discussions.

5.3.2.1 Mean-risk portfolio optimization model under missing data Mean-risk portfolio optimization [49] considers m assets with returns captured by a random vector $\boldsymbol{\xi} = [\xi_1, \cdots, \xi_m]$ whose joint distribution is $\mathbf{P}$. A portfolio is denoted by a vector $\mathbf{x} = [x_1, \cdots, x_m]$, where

$$\mathbf{x} \in \mathcal{X} = \{\mathbf{x} \in \mathbb{R}_+^m | \sum_{i=1}^{m} x_i = 1\}.$$

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 (5.18):

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \mathbb{E}\big[\langle -\boldsymbol{\xi}, \mathbf{x}\rangle\big] + \rho \mathrm{CVaR}_\alpha(\langle -\boldsymbol{\xi}, \mathbf{x}\rangle) \tag{5.18}$$

where $\rho > 0$.

CVaR at level α, $0 < \alpha \leqslant 1$ represents the average of the $\alpha \times 100\%$ worst portfolio losses. Replacing CVaR in (5.18) with its definition [103], we have

$$\min_{\mathbf{x}\in\mathcal{X}, d\in\mathbb{R}} \quad \mathbb{E}(\langle -\boldsymbol{\xi}, \mathbf{x}\rangle) + \rho \left[d + \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}}(\langle -\boldsymbol{\xi}, \mathbf{x}\rangle - d)^+ \right]. \tag{5.19}$$

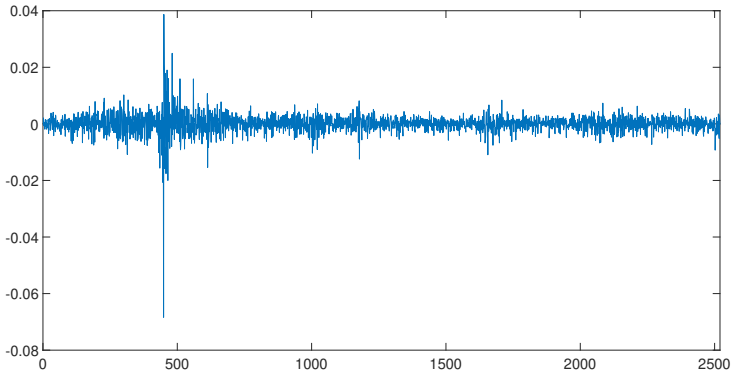
When only incomplete observations of $\boldsymbol{\xi}$ are available, we aim to solve Optimization (5.20):

$$\min_{\mathbf{x}\in\mathcal{X}, d\in\mathbb{R}} \quad \max_{\mathbf{P}\in\mathcal{P}'} \quad \mathbb{E}(\langle -\boldsymbol{\xi}, \mathbf{x}\rangle) + \rho \left[d + \frac{1}{\alpha}\mathbb{E}_{\mathbb{P}}(\langle -\boldsymbol{\xi}, \mathbf{x}\rangle - d)^+ \right]. \tag{5.20}$$

5.3.2.2 Numerical settings

5.3.2.3 Data preprocessing In the preprocessing step, we round each data to the nearest thousandth and remove the data of the first 500 days because these data follow very different patterns due to the well-known financial crisis of 2007-2008, as shown in Figure 5.

Figure 5: Daily returns of Asset #1 (iShares Core U.S. Aggregate Bond ETF).



5.3.2.4 Training set and test set We explain the way to generate the training set and test set. The training set is assumed to be known to the decision-makers, and the test set is assumed to be unknown to the decision-makers and is used to evaluate the out-of-sample performance.

- We model the support based on previous years' data. That is we approximate the known support $\mathcal{S}$ of the returns $\boldsymbol{\xi}$ 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 use the first $m = 10$ assets in the original data set for all the experiments. Therefore, the data set used in the experiments contains 10 assets with 2020 daily returns each. To model the partially observed data, we assume each dimension of the data in the training set will be missing with a fixed probability 0.5.

5.3.2.5 Implemented approaches We aim to compare Model (4.8) to a classic data-imputation-based procedure. We explain the implementation details of each approach in the following. The data-imputation-based procedure first recovers the unknown distribution using nearest neighbor imputation; then it solves Model (5.19) through the empirical distri-

bution based on the completed data set. In the first step because the support $\mathcal{S}$ is known, each incomplete data is imputed to the nearest support with respect to the L1 norm. Besides, if several supports achieve the smallest distance at the same time, one random support among them is used.

5.3.2.6 Results We conclude the main numerical results on the partially observed data in this section, where our model achieves better out-of-sample performance than the data-imputation-based procedure. We conducted 10 groups of experiments under different values of τ, where $\tau = 0.01, 0.02, \cdots, 0.1$. Each group of experiments contains 57 experiments as discussed in Section 5.3.2.2. We present the out-of-sample performance in the following.

5.3.2.7 Out-of-sample performance The out-of-sample performance is defined as the value of

$$\min_{d\in\mathbb{R}} \quad \mathbb{E}_{\hat{\mathbf{P}}}(\langle -\boldsymbol{\xi}, \hat{\mathbf{x}}\rangle) + \rho\left[d + \frac{1}{\alpha}\mathbb{E}_{\hat{\mathbf{P}}}(\langle -\boldsymbol{\xi}, \hat{\mathbf{x}}\rangle - d)^+\right], \tag{5.21}$$

where $\hat{\mathbf{P}}$ represents the empirical distribution of the test set, and $\hat{\mathbf{x}}$ represents the portfolio obtained from corresponding models based on the training set. We repeat each experiments for 10 times (random missing components) to obtain the standard deviation. The results are concluded in Figure 6.

Our model outperforms consistently by achieving lower values in (5.19) and smaller standard deviation. As the value of τ increases, the performance of our model first increases and then stays stable. And the standard deviations are also decreasing, which indicates fewer fluctuations in the costs in practice.

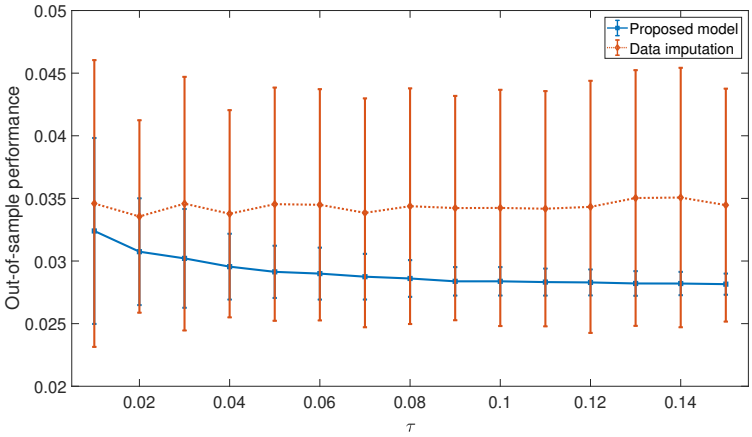


Figure 6: Out-of-sample performance ($\times 0.001$).

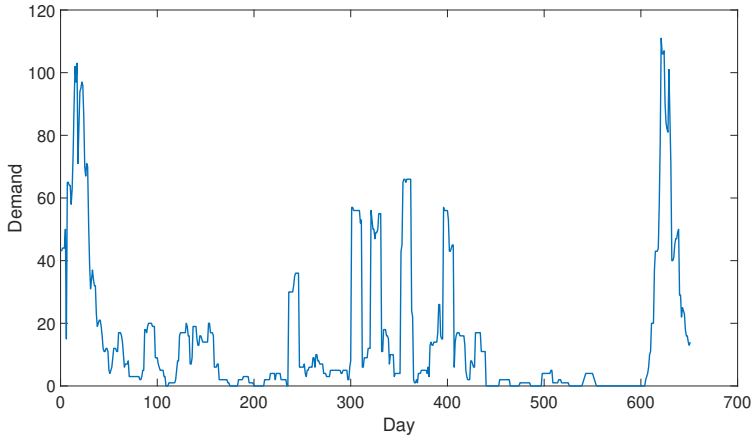
6.0 Inventory Management with Highly Unpredictable Non-stationary Demand

6.1 Introduction

Inventory management facing nonstationary demand is a fundamental problem in supply chain analyses, with classical results dating back to at least [111]. This problem has been studied with many variants, such as Markov-modulated demand problems [121, 112, 93], nonstationary stochastic lot-sizing problems [6, 20, 133], and simulation and forecast-based studies [62, 30, 9, 36]. However, these studies have largely relied on assumptions about the demand process to allow for direct or indirect predictions of future demand. More specifically, they assumed that the demand process is either known or that future demand can be effectively predicted through certain features or parameterized models. Although these assumptions can be satisfied in some practical cases, future demands are highly unpredictable in certain industrial contexts. For example, when the demand for one product is largely affected by some unprecedented factors (such as Covid-19), no historical data can be used to develop a forecasting model or to verify whether the conditions of a certain mathematical model are met. Another example is the demand patterns shown in Figure 7. It plots the demand data of a product from one of the world's largest online retailers, where the demand pattern is barely repeated. This implies that historical data provide little reference for future cases. Typically, in such challenging data environments, decision-makers in practice adjust inventory policies based on the most recent sequentially observed demand. In this study, we focus on deriving inventory policies under such challenging data environments. The only data that can be used are the sequentially observed daily demands. We examine the problems encountered by data-driven methods in the literature and propose new methods with improved performance.

The two most widely used approaches to infer future demand solely based on the observed demand are sample average approximation and exponential smoothing. SAA approximates the future demand distribution through the observed empirical distribution, whereas expo-

Figure 7: Real-world demand data for a particular product of an online retailer



nential smoothing is a weighted moving average method. The benefit of using SAA is that if the unprecedented demand data come from a single distribution, the empirical distribution used by SAA represents an optimal (nonparametric maximum likelihood) estimator for the true demand distribution. In contrast, exponential smoothing is effective at capturing demand distribution changes or nonstationarity. However, these two approaches also have some evident drawbacks. SAA is not effective when dealing with nonstationarity, and exponential smoothing suffers from overfitting. To prove this point, we apply these two methods to estimate the mean value of daily demand for the two types of demand shown in Figure 8 and 9. Figure 8 presents the daily demand generated from one distribution with nonzero variance, and Figure 9 plots that generated from two "noiseless" (the variance is zero) demand distributions.

The performances of SAA and exponential smoothing are summarized in Figure 10 and 11. In the first demand setting, we expect a good prediction to be a straight line that is close to 100. Therefore, exponential smoothing performs worse than SAA in this case. This is because exponential smoothing tends to partially "forget" a previously encountered demand. However, all historical demand data contain useful information for future predictions because they originate from the same distribution. When the smoothing factor α is 0.5, severe

Figure 8: Noisy stationary demand.

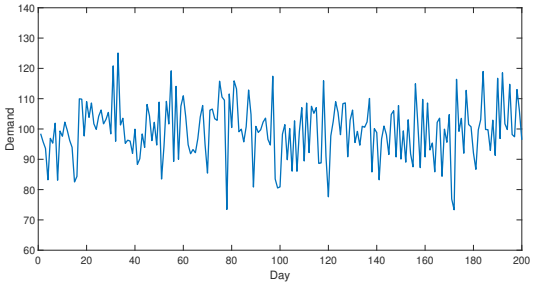
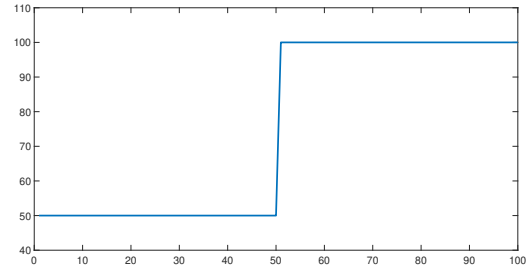


Figure 9: Noiseless nonstationary demand.



overfitting occurs. As indicated by the orange dotted line in Figure 11, the performance of SAA is considerably rendered after the demand pattern changes.

Figure 10: Noisy stationary demand.

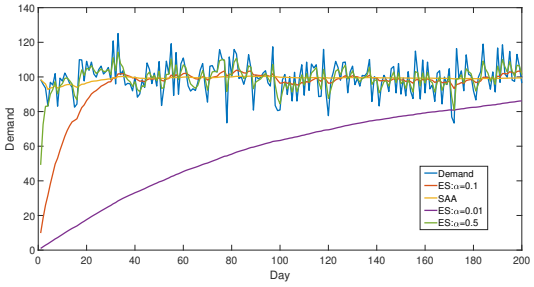
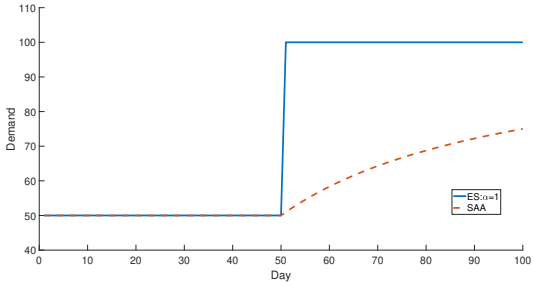


Figure 11: Noiseless nonstationary demand.



The real-world demand is a combination of Figure 8 and 9, meaning that it is stochastic/noisy (having nonzero variance) and nonstationary. Therefore, neither of the approaches can effectively capture the behaviors on the basis of sequentially observed demand data, which makes it even more difficult to obtain good inventory control policies. This paper proposes two novel methods to adjust inventory policies based on the sequentially observed demand data: integrated Bayesian (IB) approach and seperate lasso (SL) approach. However, our approaches can handle nonstationarity while behaving similarly to SAA during periods with stationary demand. This point is demonstrated in Figure 13, where SL is com-

pared with SAA and exponential smoothing. We find that the SL approach is as stable as SAA (robust to stochasticity) before the demand pattern changes. After day 50, SL follows up the demand distribution change even more rapidly than exponential smoothing and then remains more stable.

Figure 12: Real-world demand.

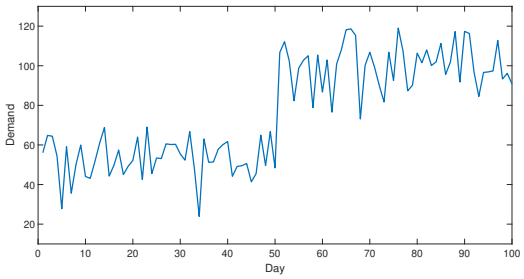
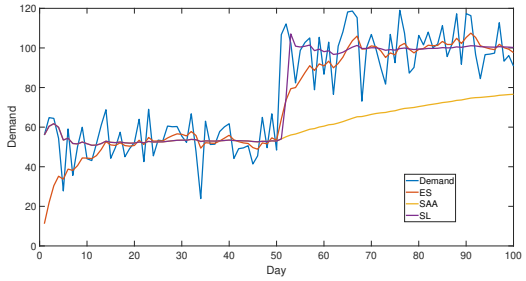


Figure 13: Comparisons.



In what follows, we define our problem settings and summarize the main results. This study's primary assumption is that we do not use historical data, or the available historical data/features cannot be used to predict future demand due to certain unprecedented situations. In particular, we assume that the demand follows different distributions for different periods, where neither distributions nor their transition properties are known. Therefore, any historical data observed long ago may belong to a demand distribution that is completely different from the present distribution. Hereafter, we refer to this setting as *highly unpredictable*.

Given the highly unpredictable data setting, we study frameworks for adjusting the inventory decisions by using sequentially observed demand data. We consider this problem in two situations: the first case represents a data environment that assumes an uncertainty set comprising all possible demand distributions (given exogenously). The second case represents a data environment in which information about the uncertainty sets is not available. We develop a parametric IB approach and a nonparametric SL approach for these two situations, respectively. While the proposed approaches can be extended to incorporate many inventory policy classes, we demonstrate the effectiveness of our approaches by focusing on the widely used (r, Q) policy. The main idea of both these methods is to distinguish be-

tween stochasticity and nonstationarity. Therefore, both methods outperform SAA when nonstationary appears and outperform exponential smoothing during stationary phases.

The IB approach assumes that an uncertainty set comprising all possible demand distributions is available and conducts an integrated analysis of demand distribution estimation and cost minimization to derive an inventory policy. At the end of each day, it obtains the inventory policy for the next day by greedily minimizing the expected inventory costs, which are evaluated based on a Bayesian analysis of the observed demand. With further theoretical analysis, we propose an easy-to-implement algorithm. The SL approach is a nonparametric approach, which does not make any assumption related to the possible distribution and uses separate analysis steps for the demand determination and policy derivation. We first formulate a lasso-based model to determine the demand data belonging to the current demand distribution at the end of each day. Then, we propose a distributionally robust optimization (DRO) model to determine the inventory policies for the next day. In addition, theoretical analyses are conducted to derive an easy-to-implement algorithm and explore its performance.

We also evaluate the performances of IB and SL approaches empirically by evaluating them through multiple sets of experiments. In this problem, the optimal inventory policies are unobtainable because a reliable forecast of future demand is not available. Therefore, we obtain baseline inventory policies (OPT) under a relaxed setting, where demand season changes and demand distributions for all demand seasons are assumed to be known at the beginning. Accordingly, we derived inventory policies (OPT) by following the methods introduced in [20]. We compare the inventory costs of IB and SL with those of OPT under nine different demand data settings, classified based on the magnitude of demand variance and length of each demand season. As shown in Table 5, both IB and SL achieve, on average, approximately 1.12 times the costs of OPT.

In the second set of experiments, we benchmark IB and SL against state-of-the-art approaches: sample average approximation (SAA), rolling horizon (RH), and exponential smoothing (ES). SAA is used to derive the inventory policies on a daily basis according to the observed empirical distributions. RH derives inventory policies according to the observed empirical distributions in some prescribed horizons. ES predicts future demand, and

accordingly, derives the inventory policies by using exponential smoothing. Under the nine previously mentioned demand data settings, IB and SL consistently outperform all these approaches (Table 5). Finally, we apply our methods to datasets obtained from one of the world's largest e-commerce websites. Although the demand data are highly nonstationary and stochastic, we observed that the proposed approaches are able to capture the hidden patterns and achieve lower inventory costs.

Table 5: Comparisons of IB and SL with OPT and other heuristics (i.e., SAA, RH, and ES). The numbers presented in this table indicated the ratio of the total inventory cost for the corresponding method and the costs of OPT for different data environments; smaller values indicate lower costs.

Data environment	1	2	3	4	5	6	7	8	9
OPT	1	1	1	1	1	1	1	1	1
IB	**1.04**	**1.2**	**1.12**	**1.05**	1.23	**1.14**	**1.04**	1.2	**1.08**
SL	1.07	1.14	1.14	1.06	**1.15**	1.15	1.05	**1.13**	1.09
SAA	1.52	1.35	1.25	1.52	1.36	1.28	1.31	1.23	1.27
RH	1.2	1.28	1.19	1.19	1.28	1.23	1.15	1.15	1.14
ES	1.1	1.27	1.14	1.11	1.28	1.16	1.08	1.18	1.17

6.2 Problem Formulation

We consider the inventory management problem for a warehouse with set-up costs, proportional holding, and penalty costs. We assume that the demand is nonstationary, implying that it follows different distributions for different time periods. We refer to one continuous period that follows the same demand distribution as one *demand season*. The managers are unaware of the demand process, number of demand seasons, starting times, and distributions of demand seasons. We assume that the demand is *highly unpredictable*, meaning that we do not have historical data, or the available historical data or features cannot be used to predict the future demand reliably. We assume that all orders are placed and will be received at the beginning of each day. Moreover, it takes L days for one order to be delivered. The ordering

cost comprises of a fixed cost for placing an order. A linear holding cost is charged for every unit carried from one day to the next, whereas a linear penalty cost is incurred for each unit of outstocking at the end of the day. Mathematically, suppose that one demand season starts on day $\mathbf{s}$. At the end of day t, the observed demand data points are $\{d_s, \cdots, d_t\}$. Our goal is to develop a model F that maps the observed data into inventory policies for day $t+1$, as shown below.

$$F(\{d_s, \cdots, d_t\}) = \langle R_{t+1}, Q_{t+1} \rangle \tag{6.1}$$

This study focuses on the widely used (R,Q) policy for demonstration (EOQ model with type-1 service level). However, the proposed approach can be extended to other policy classes. More specifically, we obtain the reorder point and order quantity as

$$Q_t = \sqrt{\frac{2K'\mu_t}{h}}, R_t = (P_t^L)^{-1}(\frac{p}{p+h}), \tag{6.2}$$

where μ_t and P_t^L are the mean values of the daily demand and the cumulative distribution function (CDF) of the L-day demand. Both are unknown and must be determined from the data.

One key point in making a good prediction in such challenging data environments is to distinguish between the randomness inside particular demand distributions and demand distribution changes. If the demand is stationary and all changes in the daily demand are caused by randomness, the function F should ideally use all the demands in $\{d_s, \cdots, d_t\}$ to derive the inventory policies. Otherwise, if the last demand distribution starts at t' ($s < t' < t$), the function F should use the demand in $\{d_{t'}, \cdots, d_t\}$ to derive the inventory policies. Therefore, our methods focus on identifying the sources of the changes observed in the demand data to improve performance.

6.3 IB Approach

In this section, we propose a heuristic IB approach to obtain the (R_t, Q_t) policy. We assume an uncertainty set of possible demand distributions (given exogenously). This approach adopts a Bayesian analysis and conducts an integrated analysis of demand distribution estimation and inventory cost minimization. Our further theoretical analyses based on Proposition 1 help us establish an easy-to-implement algorithm.

Before focusing on the main concepts of the IB approach, we first introduce a method called "dynamic rolling horizon", in which the starting day of the current horizon is always set as the first day of the current assumed/determined demand season. The time horizon is rolled over every time a new demand season is determined by our algorithm. Therefore, we assume that there are at most two demand distributions inside the current horizon. In the following section, we briefly introduce the main idea of the IB approach, through which we greedily derive the ordering policies of each day based on the demand distribution inside the uncertainty set, the inventory policy of which minimizes the current expected inventory costs. That is, at the end of day t, we base our next ordering policies for day $t+1$ on one distribution, p_{k^*}, from the uncertainty set, where k^* is the optimal solution of Optimization (6.3), whose objective function represents the expected inventory costs with respect to the policies based on distribution $p_{k'}$. Here, we explain Problem (6.3) and detail it in Section 6.3.1

$$\min_{k'=1,\cdots,K} \sum_{k=1}^{K} \mathrm{P}(B_k|A_t) C_{k'}^{k}. \tag{6.3}$$

We use $\mathrm{P}(B_k|A_t)$ to denote the probability that the current demand distribution is p_k given the observed demand data on day t. We use $C_{k'}^{k}$ to denote the expected inventory costs of basing ordering policies on demand distribution $p_{k'}$ given that the actual demand distribution is p_k, which can be estimated or analytically calculated. In (6.3), we determine $\mathrm{P}(B_k|A_t)$ through Bayesian analysis. Then, we choose the optimal distribution by solving (6.3), which is equivalent to a simple threshold test on $\mathrm{P}(B_k|A_t)$ as will be shown in Proposition 4. Finally, we obtain the ordering policies by using the determined demand distribution.

6.3.1 Main procedure

We present the details of the IB approach, described using Algorithm 2. This approach involves two main steps. Step 1 evaluates the possibilities of the current demand distribution, which can be achieved via Bayes' rule. We denote the starting day of the current time horizon as day s. At the end of day t, the demand inside the current time horizon is $\{d_s, \cdots, d_t\}$. The uncertainty set for the possible demand distributions is $P = \{p_1, p_2, \cdots, p_K\}$, and the initial demand distribution is denoted as p_{k_1}. We use $\mu(p_k)$ to represent the mean value of distribution p_k and use $F_{p_k}^{-1}(\alpha)$ $(0 < \alpha < 1)$ to indicate the α quantile of a random variable following distribution p_k. First, two events are defined as follows:

- Event A_t : The observed demand is $\{d_s, \cdots, d_t\}$.
- Event B_k : The current demand distribution is p_k.

The probability that the current demand distribution is p_k after observing $\{d_s, \cdots, d_t\}$ can be denoted as $\mathrm{P}(B_k|A_t)$. By using the dynamic rolling horizon, there exists a maximum of two demand distributions inside the current horizon. $\mathrm{P}(B_k|A_t)$ can be calculated as follows:

$$\mathrm{P}(B_k|A_t) = \frac{\mathrm{P}(A_t|B_k)\mathrm{P}(B_k)}{\sum_{j=1}^{K} \mathrm{P}(A_t|B_j)\mathrm{P}(B_j)}, \tag{6.4}$$

$$\text{where } \mathrm{P}(A_t|B_k) = \frac{1}{t-s} \sum_{j=s+1}^{t} p_{k_1}(d_s) \cdots p_{k_1}(d_{j-1}) p_k(d_j) \cdots p_k(d_t), \quad \forall k \neq k_1, \tag{6.5}$$

$$\text{and } \mathrm{P}(A_t|B_{k_1}) = p_{k_1}(d_s) \cdots p_{k_1}(d_t). \tag{6.6}$$

$\mathrm{P}(A_t|B_k)$ is the likelihood of observing $\{d_s, \cdots, d_t\}$ given that the current distribution is p_k. Eq. (6.5) provides an expression for $\mathrm{P}(A_t|B_k)$, because if the current distribution changes, $k \neq k_1$, then each possible day inside the current planning horizon has an equal probability of being the switching day as we make no assumption regarding the transition property. For the same reason, all $\mathrm{P}(B_k)$ are equal. Thus, we have

$$\mathrm{P}(B_k|A_t) = \frac{\mathrm{P}(A_t|B_k)}{\sum_{j=1}^{K} \mathrm{P}(A_t|B_j)}.$$

In Step 2, a threshold test is used to select the demand distribution and inventory policies, with the aim of minimizing the future expected costs. Optimization (6.7) is formulated

Algorithm 2 IB

Initialization: input T, $s \leftarrow 1$; $t = 2 : T$
STEP 1:
$\mathrm{P}_k = \sum_{j=s+1}^{t} p_{k_1}(d_s) \cdots p_{k_1}(d_{j-1}) p_k(d_j) \cdots p_k(d_t), \quad \forall k = 1, \cdots, K;$
$P = \max_{k \neq k_1} \dfrac{\mathrm{P}_k}{\sum_{j=1}^{K} \mathrm{P}_j};$
$k_2 = \mathrm{argmax}_{k \neq k_1} \dfrac{\mathrm{P}_k}{\sum_{j=1}^{K} \mathrm{P}_j};$
STEP 2: $P \geqslant \theta$
$s \leftarrow \mathrm{argmax}_i \dfrac{p_{k_1}(d_s) \cdots p_{k_1}(d_{i-1}) p_{k_2}(d_i) \cdots p_{k_2}(d_t)}{\sum_{j=s+1}^{n} p_{k_1}(d_s) \cdots p_{k_1}(d_{j-1}) p_{k_2}(d_j) \cdots p_{k_2}(d_t)}$;
set the current distribution p_{k_1} as p_{k_2};
set the order quantity to $\sqrt{\frac{2K'\mu(p_{k_1})}{h}}$, reorder point $r = F^{-1}_{p_{k_1}}(\frac{p}{p+h})$.

to minimize the future expected costs over the possible demand distributions in the uncertainty set. In (6.7), $C^k_{k'}$ denotes the expected costs of using a policy based on the demand distribution $p_{k'}$ while the real demand distribution is p_k

$$\min_{k'=1,\cdots,K} \sum_{k=1}^{K} \mathrm{P}(B_k|A_t) C^k_{k'}. \tag{6.7}$$

Proposition 4 states that a unique threshold value θ exists for $\mathrm{P}(B_{k_2}|A_t)$ if the policy based on p_{k_2} achieves the optimal solution for (6.7). The proofs are provided in Appendix D.1.

Proposition 4. *The policy based on demand distribution p_{k_2} achieves the optimal solution if and only if $P(B_{k_2}|A_t) \geqslant \theta$ and*

$$\theta = \frac{\sum_{k \neq k_2} P(B_k|A_t)(C^k_{k_2} - C^k_{k^*})}{C^{k_2}_{k^*} - C^{k_2}_{k_2}}, \text{ where } k^* = \arg\min_{k' \neq k_2} [P(B_{k_2}|A_t) C^{k_2}_{k'} + \sum_{k \neq k_2} C^k_{k'} P(B_k|A_t)].$$

In practice, it is not difficult to determine $C^k_{k'}$ through simulations, because the demand distribution is fixed to p_k, and the inventory policy is defined based on a fixed distribution $p_{k'}$. We also provide an example to analytically derive $C^k_{k'}$ in Appendix D.2. Once the demand distribution is determined, we derive the (R_t, Q_t) policy based on the identified distribution. Finally, the time horizon is rolled over if the inventory policy based on p_{k_2} is used. The new

time horizon starts on the day with the highest likelihood of being the starting day for the current demand distribution. Thus, the new starting day, s, is set as i^*, such that

$$\begin{aligned} &\frac{p_{k_1}(d_s)\cdots p_{k_1}(d_{i^*-1})p_{k_2}(d_{i^*})\cdots p_{k_2}(d_t)}{\sum_{j=s+1}^{n} p_{k_1}(d_s)\cdots p_{k_1}(d_{j-1})p_{k_2}(d_j)\cdots p_{k_2}(d_t)} \\ \geqslant &\frac{p_{k_1}(d_s)\cdots p_{k_1}(d_{i-1})p_{k_2}(d_i)\cdots p_{k_2}(d_t)}{\sum_{j=s+1}^{n} p_{k_1}(d_s)\cdots p_{k_1}(d_{j-1})p_{k_2}(d_j)\cdots p_{k_2}(d_t)}, \end{aligned} \tag{6.8}$$

for $i = s+1, \cdots, t$. This can be proved by defining an event, C_i, indicating that the switching day is day i. By following a procedure similar to that described in Eq. (6.4), we obtained $\mathrm{P}(C_i|A_t)$, which denotes the probability that the current demand season starts on day i, given the observed demand $\{d_s, \cdots, d_t\}$ as

$$\mathrm{P}(C_i|A_t) = \frac{p_{k_1}(d_s)\cdots p_{k_1}(d_{i-1})p_{k_2}(d_i)\cdots p_{k_2}(d_t)}{\sum_{j=s+1}^{n} p_{k_1}(d_s)\cdots p_{k_1}(d_{j-1})p_{k_2}(d_j)\cdots p_{k_2}(d_t)}.$$

In conclusion, by recursively applying the two aforementioned steps at the end of each day, the IB approach produces a sequence of (R_t, Q_t) policies. We summarize these steps in Algorithm 2, where T denotes the total number of days for this problem.

6.4 SL Approach

In this section, we propose the SL approach for cases where the uncertainty set of possible demand distributions is not available. This approach formulates a fused-lasso model and conducts a separate analysis on demand season estimation and inventory cost minimization. Our theoretical analyses based on Lemma 3 help us derive an easy-to-implement Algorithm 3.

Briefly, the SL approach first identifies the demand data that correspond to the current demand season based on the observed data. Then, it derives the inventory policies based on the identified data belonging to the current demand season. The dynamic rolling horizon heuristic is still used such that we can solve the multi-season problem by analyzing two consecutive demand seasons. Additionally, a DRO model is proposed to derive inventory policies based on the demand data to guarantee the out-of-sample performance. This is

because, in the settings of the SL approach, we do not know the exact demand distributions. Furthermore, a simple SAA framework hardly guarantees the Type-1 service level because the amount of demand data is very limited. The proposed DRO framework determines the inventory policies by considering the worst-case distribution inside an ambiguity set containing possible demand distributions, which significantly improves the out-of-sample performance. In what follows, we introduce the main steps of the SL approach, the details of which will be presented in Section 6.4.1. First, at the end of each day t, we propose Model (6.9) to identify the possible changes in demand distributions

$$\min_{\lambda_s,\cdots,\lambda_t} \frac{1}{2}\sum_{j=s}^{t}(d_j-\lambda_j)^2+\beta\sum_{j=s+1}^{t}|\lambda_j-\lambda_{j-1}|. \tag{6.9}$$

Model (6.9) is adapted based on a fused-lasso model, which has been used in offline change point detection in the machine learning community [125, 79, 128, 104]. In Model (6.9), the observed demand data in the current horizon are $\{d_s,\cdots,d_t\}$. We use λ_j to represent the means of the corresponding demand distributions on day j, $s\leqslant j\leqslant t$. We use s to denote the starting day of the current time horizon. In addition, $\beta>0$ serves as the penalty parameter controlling the trade-off between the observation noise, $\frac{1}{2}\sum_{j=s}^{t}(d_j-\lambda_j)^2$, and demand season changes, $\sum_{j=s+1}^{t}|\lambda_j-\lambda_{j-1}|$. Although fused-lasso models have been studied [104] before, it is used as an offline unsupervised approach. However, in the SL approach, we develop an online algorithm based on fused-lasso models relying on the structures of optimal solutions of consecutive demand seasons. Further analysis reveals that the performance of correctly detecting the underlying patterns of sequential demand seasons can be guaranteed in our approach. After the data of the current demand season is identified from Framework (6.9), we derive the inventory policies for the day $t+1$ based on a DRO model, which will be introduced in Section 6.4.2. In Section 6.4.3, we discuss the selection of a parameter β to guarantee the performance of the SL approach.

6.4.1 Main procedure

We present the SL approach in this section. The steps are summarized in Algorithm 3, and its details are provided in the following section. Recall that we proposed Model (6.10) to identify data of the current demand season, where d_j is the demand for day j, λ_j is the mean value of the demand distribution on day j, and β is a constant ($s \leqslant j \leqslant t$)

$$\min_{\lambda_s,\cdots,\lambda_t} \frac{1}{2}\sum_{j=s}^{t}(d_j-\lambda_j)^2+\beta\sum_{j=s+1}^{t}|\lambda_j-\lambda_{j-1}|. \tag{6.10}$$

The first term in Optimization (6.10) is the summation of the variance of the demand observations. The second term reflects the magnitude of demand season changes. Model (6.10) aims to explore the trade-off between the number of demand seasons ($\sum_{j=s+1}^{t}|\lambda_j-\lambda_{j-1}|$) and the total variance observed ($\frac{1}{2}\sum_{j=s}^{t}(d_j-\lambda_j)^2$), where the trade-off is controlled by β. We assume that $\beta>0$ is a given number here and discuss its selection in Section 6.4.3. For two consecutive demand seasons, the optimal solution in Model (6.10) has special structures as summarized in Lemma 3.

Lemma 3. *For two consecutive demand seasons, the demand data inside the current time horizon have the same demand distribution if and only if* $\lambda=\frac{\sum_{j=s}^{t}d_j}{t-s+1}$ *satisfies* $|\sum_{j=s}^{i-1}(\lambda-d_j)|\leqslant\beta,\ \forall i=s+1,\cdots,t+1$.

For the proof of this lemma, we refer our readers to Appendix D.3. Based on Lemma 3, we derive the inventory policy for day $t+1$ at the end of each day t through the following two steps. In Step 1, we decide whether a new demand distribution has started or not. First, we calculate $\lambda=\frac{\sum_{j=s}^{t}d_j}{t-s+1}$; then, we determine the largest deviation $|\sum_{j=s}^{i-1}(\lambda-d_j)|$ (x in Algorithm 3) for $i=s+1,\cdots,t+1$. In Step 2, we compare x with the threshold value β.

Algorithm 3 SL

1: **Initialization:** T,β, $s\leftarrow 1$; $t\leftarrow 1:T$
2: **STEP 1**:
3: $\lambda\leftarrow\frac{d_s+\cdots+d_t}{t-s+1}$;
4: $x\leftarrow\max_{s+1\leqslant i\leqslant t+1}|\sum_{j=s}^{i-1}(\lambda-d_j)|$;
5: **STEP 2**: $x>\beta$
6: solve Problem (6.11a) and update s, which is the first day of the current demand distribution;
7: derive inventory policies based on data $\{d_s,\cdots,d_t\}$ with Model (6.14);
8: derive inventory policies based on data $\{d_s,\cdots,d_t\}$ with Model (6.14);

If $x > \beta$, indicating that a new demand season is found. We find the data belonging to the current demand distribution through (6.11).

$$\min_{\lambda_s,\cdots,\lambda_t} \quad \frac{1}{2}\sum_{j=s}^{t}(d_j - \lambda_j)^2 \tag{6.11a}$$

$$s.t. \quad \sum_{j=s+1}^{t} ||\lambda_j - \lambda_{j-1}||_0 = 1. \tag{6.11b}$$

Constraint (6.11b) guarantees that the demand season changes only once. Problem (6.11) is easy to solve by merely enumerating all $t - s$ scenarios. That is, the optimal solution comes from the following set of solutions:

$$\lambda_s = \cdots = \lambda_{i-1} = \frac{\sum_{j=s}^{i-1} d_j}{i-s}, \quad \lambda_i = \cdots = \lambda_t = \frac{\sum_{j=i}^{t} d_j}{t-i+1}, \quad \forall i = s+1, \cdots, t.$$

Let us assume that switching on day i yields the minimal objective value. We then use data from day i to t to derive the current inventory policies according to the DRO model proposed in Section 6.4.2. We also set $s = i$ as the starting time of the current demand distribution. If no new demand distribution is detected, we add newly observed data to the dataset and derive the corresponding inventory policy based on the DRO model. At the end of the next day, we repeat the above procedures. We summarize these above steps in Algorithm 3, where the parameter T denotes the final day. In Algorithm 3, the only remaining challenge is to select a suitable threshold, β. We analyze the effect of β and provide guidelines for choosing it in Section 6.4.3.

6.4.2 Distributionally robust optimization framework for deriving (R,Q) policies

Our setting in the SL approach corresponds to a nonparametric case, where we only have data of the current demand season but do not know the exact forms of the demand distributions. The most common way to derive the inventory policies is to use the empirical distribution, which is often referred to as SAA, as shown in (6.12). In (6.12), we use d^L

to denote the random variable of the L-day demand (demand during the lead time). Its empirical PDF is denoted as $\hat{p}^L$, and the empirical CDF is denoted as $\hat{\mathbb{P}}^L$.

$$\begin{aligned} \min_{Q,R} \quad & \mathbb{E}_{\hat{p}^L}[\frac{K'd^L}{QL} + h\frac{Q}{2} - hd^L)] + hR \\ s.t. \quad & \hat{\mathbb{P}}^L(R) \geqslant \alpha. \end{aligned} \tag{6.12}$$

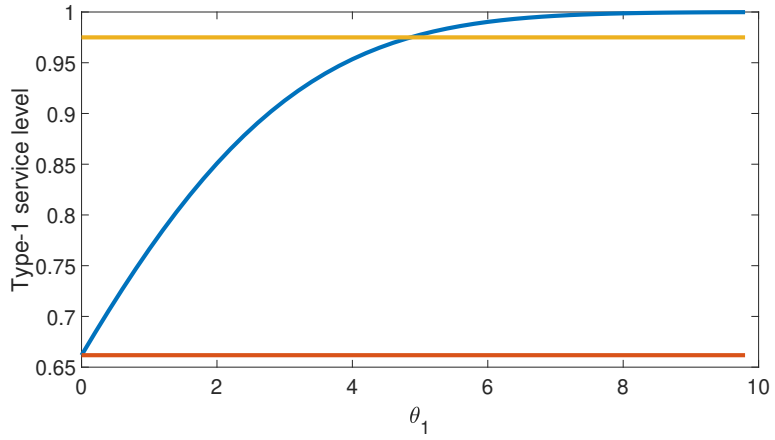
However, for our problem, the size of the available dataset is usually very small because it requires L days to obtain one L-day demand. Decision-makers may apply methods like bootstrapping to estimate the empirical distributions. Under these case, the estimation error involved is large, and the out-of-sample performance of SAA is poor [119]. Therefore, the Type-1 service level is hard to guarantee. To overcome this drawback, we propose a model based on distributionally robust optimization (DRO) which is formulated as follows. In (6.13), the PDF of the L-day demand is denoted as p^L and its corresponding CDF is denoted as P^L

$$\begin{aligned} \min_{Q,R} \max_{p^L \in \mathcal{B}_{\theta_1}} \quad & \mathbb{E}_{p^L}[\frac{K'd^L}{QL} + h\frac{Q}{2} - hd^L)] + hR \\ s.t. \quad & \min_{p^L \in \mathcal{B}_{\theta_1}(\hat{p}^L)} P^L(R) \geqslant \alpha. \end{aligned} \tag{6.13}$$

Optimization (6.13) considers the p^L residing in an ambiguity set $\mathcal{B}_{\theta_1}(\hat{p}^L)$ that achieves the worst-case Type-1 service level. The constraint restricts this level to be larger than α. The ambiguity set is controlled by a parameter θ_1 and the obtained empirical distribution $\hat{p}^L$. Model (6.13) improves the out-of-sample performance of the Type-1 service level because the ambiguity set $\mathcal{B}_{\theta_1}(\hat{p}^L)$ models the estimation errors contained in the empirical distribution. We refer to an example in our computational experiments here to briefly illustrate this point. Details on the experiments can be found in Appendix D.5. In Figure 14, the Type-1 service level is improved by DRO (blue) with appropriate values of θ_1 compared with the results of SAA (red). The yellow line indicates the target Type-1 service level of 97.5%.

Detailed Discussion: We choose to construct ambiguity sets based on the Wasserstein distance in this study due to two reasons: First, from the modeling perspective, this ambiguity set contains infinitely many distributions that are close to the empirical demand distributions in that the Wasserstein distance equals to the lowest cost of transporting the

Figure 14: Average Type-1 service level with 2 available demand data.



probability mass from one distribution to the other. It not only allows the weights of the empirical distribution to be adjusted but also allows the probability mass itself to be adjusted. Moreover, it guarantees the out-of-sample performance theoretically [49]. Second, from the computational perspective, Optimization (6.13) has closed-form reformulations that can be solved efficiently, as we will show later. The details of the proposed model are as follows. We begin with the definition of the Wasserstein distance and the corresponding Wasserstein balls used in our model.

Definition 5. *For any closed set* $\Xi \subset \mathbb{R}^{n+m}$*, define*

$$M^1(\Xi) = \{\mathbb{Q} : \mathbb{E}^{\mathbb{Q}}[\|\xi\|_1] = \int_{\Xi} \|\xi\|_1 \mathbb{Q}(d\xi) < \infty\}.$$

The 1*-Wasserstein distance between two distributions* $\mathbb{Q}_1$*,* $\mathbb{Q}_2 \in M^1(\Xi)$ *is defined as*

$$W_1(\mathbb{Q}_1, \mathbb{Q}_2) := \inf\{\int_{\Xi^2} \|\xi_1 - \xi_2\|_1 \Pi(d\xi_1, d\xi_2) :$$

$$\Pi \text{ is a joint distribution of } \xi_1 \text{ and } \xi_2 \text{ with marginals } \mathbb{Q}_1 \text{ and } \mathbb{Q}_2, \text{ respectively}\}.$$

Definition 6. *We use* $\hat{\mathbb{P}}$ *to denote the empirical demand distribution. A Wasserstein ball centerred at* $\hat{\mathbb{P}}$ *with radius* θ *under* 1*-Wasserstein distance is defined as:*

$$\mathcal{B}_\theta(\hat{\mathbb{P}}) = \{\mathbb{H}^{(L)} \in M^1(\Xi) : W_1(\mathbb{H}^{(L)}, \hat{\mathbb{P}}) \leqslant \theta\}.$$

Following the results of [31], we reformulate (6.13) and summarize the results in Proposition 5. The proofs are provided in Appendix D.4.

Proposition 5. *Optimization (6.13) is equivalent to*

$$\begin{aligned}
\min_{Q,R,g_k,t,q_k} \quad & \frac{K'(\hat{\mu}+\theta_1)}{Q} + h\frac{Q}{2} + h(R - L\hat{\mu} - L\theta_1) \\
s.t. \quad & (1-\alpha)N't + \mathbf{e}^T\mathbf{g} \geqslant \theta_1 N', \\
& R - \hat{D}_k + Mq_k \geqslant t + g_k, \quad \forall k = 1, \cdots, N' \\
& M(1-q_k) \geqslant t + g_k, \quad \forall k = 1, \cdots, N' \\
& g_k \leqslant 0, t \in \mathbb{R}, q_k \in \{0,1\}, \quad \forall k = 1, \cdots, N',
\end{aligned} \tag{6.14}$$

where we use $\hat{D}_k$ to denote k-th observed L-day demand and $\hat{\mu}$ to denote the empirical mean of the daily demand. Additionally, we suppose that there exists a total of N observed daily demands for the current demand season, as well as N' observed L-day demands. Clearly, $N' = \lfloor N/L \rfloor$. Here, M is a large but bounded number.

Optimization (6.14) can be solved efficiently. First, variable Q has closed-form solutions:

$$Q = \sqrt{\frac{2K'(\hat{\mu}+\theta_1)}{h}}. \tag{6.15}$$

Second, after Q is determined, the rest of (6.14) is a mixed-integer linear programming. The parameter θ_1 is tunable by users. We recommend selecting it via cross-validations in practice, and we also conducted computational studies for the choice of θ_1 in Appendix D.6. Additional empirical experiments with respect to the working inventory costs of the proposed DRO model can be found in Appendix D.5.

6.4.3 Choosing the threshold value β

The threshold value, β, plays an essential role in the SL approach. In this section, we describe the theoretical analysis of β, which is based on two types of errors. In the first error type, i.e., Type A error, the demand distribution is retained, but the algorithm detects a change in that distribution. In the second error type, i.e., Type B error, the demand distribution changes but the algorithm fails to detect that change. As these two error types cover all possible situations, they can be used to evaluate the efficacy of our approaches.

Next, we describe the theoretical analysis based on error Types A and B errors through a two-demand season scenario, in which the demand season only changes once. The multiseason cases can be viewed as a sequence of two demand season scenarios. We assume normal distributions for demand distributions as an example. The other distributions employ similar analysis steps. We used the following notations and assumptions. For convenience, we suppose that the current demand season starts from day 1 ($s = 1$). We denote the current demand distribution as p_1 and the second distribution as p_2. The variables, λ_1 and λ_2 denote the mean and σ_1 and σ_2 denote the standard deviations of p_1 and p_2, respectively. We use d_t to represent the demand data observed on day t. Variable z_α is the critical α value of the standard normal distribution. Proposition 6 summarizes the main results.

Proposition 6. *Suppose the demand for one warehouse has remained in the current demand distribution for n days. If β satisfies $\beta > z_{\frac{2\alpha-2+n(n-1)}{n(n-1)}} \frac{\sqrt{n}\sigma_1}{2}$, then with a probability of at least $2\alpha - 1$, Algorithm 3 will not cause any Type A errors.*

Suppose the demand for one warehouse has entered the second demand distribution for m days at the end of day $m+n$. Let F denote the CDF of the standard normal distribution. If β satisfies:

$$F\left(\frac{\beta - \frac{mn}{m+n}(\lambda_1 - \lambda_2)}{\sqrt{\frac{nm(m\sigma_1^2+n\sigma_2^2)}{(m+n)^2}}}\right) - F\left(\frac{-\beta - \frac{mn}{m+n}(\lambda_1 - \lambda_2)}{\sqrt{\frac{nm(m\sigma_1^2+n\sigma_2^2)}{(m+n)^2}}}\right) < 1 - \alpha',$$

then with a probability of at least α', Algorithm 3 will not cause any Type B errors after day $m+n$.

We use the following lemma in our analysis.

Lemma 4 (Fréchet-Hoeffding bounds)**.** *Let us assume that $G_1, \cdots, G_d$ are marginal distribution functions, and G denotes any joint distribution function with those given marginals; then, for all $\mathbf{x} \in \mathcal{R}^{\lceil}$, $\mathbf{x} = (x_1, \cdots, x_d)$,*

$\sum_{i=1}^{d} G_i(x_i) + (1-d)^+ \leqslant G(\mathbf{x}) \leqslant \min(G_1(x_1), \cdots, G_d(x_d)$.

Proof. Proof.

- **Type A error:** We define x_i^k as $x_i^k = \sum_{t=1}^{i}(d_t - \lambda(k))$, where $\lambda(k) = \dfrac{\sum_{t=1}^{k} d_t}{k}$. Then the fact that there is no Type A error until day n indicates $|x_i^k| \leqslant \beta$ for $1 \leqslant i < k$ and $1 < k \leqslant n$. We do not include the cases where $i = k$ since by definition $x_k^k = 0$. Thus, not having Type A error until day n with a probability at least $2\alpha - 1$ is equivalent to

$$\mathrm{P}(|x_i^k| \leqslant \beta, \forall 1 < k \leqslant n, 1 \leqslant i < k) > 2\alpha - 1.$$

First, we prove: $x_i^k \sim \mathcal{N}(0, i\sigma_1^2 - \dfrac{i^2\sigma_1^2}{k})$ for $1 \leqslant i \leqslant k$. Because all d_t $(1 \leqslant t \leqslant k)$ are the $i.i.d$ samples from distribution $\mathcal{N}(\lambda_1, \sigma_1^2)$ and $\lambda(k)$ follows the normal distribution $\mathcal{N}(\lambda_1, \dfrac{\sigma_1^2}{k})$. Thus, x_i^k still follows a normal distribution. We calculate the mean and variance of x_i^k in the following manner:

$$\mathbb{E}(x_i^k) = i\lambda_1 - i\lambda_1 = 0;$$

$$\mathrm{Var}(x_i^k) = \mathbb{E}((x_i^k)^2) - \mathbb{E}^2(x_i^k) = \underbrace{\mathbb{E}(i^2\lambda(k)^2 + \sum_{t=1}^{i}\sum_{j=1}^{i} d_t d_j - \sum_{t=1}^{i} 2i\lambda(k)d_t)}_{\mathbb{E}((x_i^k)^2)}$$

$$\underbrace{- [i^2\mathbb{E}^2(\lambda(k)) + \sum_{t=1}^{i}\sum_{j=1}^{i}\mathbb{E}(d_t)\mathbb{E}(d_j) - \sum_{t=1}^{i} 2i\mathbb{E}(\lambda(k))\mathbb{E}(d_t)]}_{\mathbb{E}^2(x_i^k)}$$

$$= \left[i^2\mathbb{E}(\lambda(k)^2) - i^2\mathbb{E}^2(\lambda(k))\right] + \left[\sum_{t=1}^{i}\sum_{j=1}^{i}\mathbb{E}(d_t d_j) - \sum_{t=1}^{i}\sum_{j=1}^{i}\mathbb{E}(d_t)\mathbb{E}(d_j)\right]$$

$$- \left[\sum_{t=1}^{i} 2i\mathbb{E}(\lambda(k)d_t) - \sum_{t=1}^{i} 2i\mathbb{E}(\lambda(k))\mathbb{E}(d_t)]\right]$$

$$= \frac{i^2\sigma_1^2}{k} + \sum_{t=1}^{i}[\mathbb{E}(d_t^2) - \mathbb{E}^2(d_t)] - 2i\sum_{t=1}^{i}\left[\mathbb{E}(\frac{\sum_{j=1}^{k} d_j}{k}d_t) - \mathbb{E}(\frac{\sum_{j=1}^{k} d_j}{k})\mathbb{E}(d_t)\right]$$

$$= \frac{i^2\sigma_1^2}{k} + i\sigma_1^2 - 2i\sum_{t=1}^{i}\left[\mathbb{E}(\frac{d_t^2}{k}) - \frac{\mathbb{E}^2(d_t)}{k}\right] = \frac{i^2\sigma_1^2}{k} + i\sigma_1^2 - \frac{2i^2\sigma_1^2}{k} = i\sigma_1^2 - \frac{i^2\sigma_1^2}{k}.$$

Thus, we showed $x_i^k \sim \mathcal{N}(0, i\sigma_1^2 - \frac{i^2\sigma_1^2}{k})$. In addition, for the variance we have $i\sigma_1^2 - \frac{i^2\sigma_1^2}{k} \leq \frac{k\sigma_1^2}{4}$, because the variance of x_i^k is a quadratic function with respect to i and the biggest variance is accomplished when $i = \frac{k}{2}$. Therefore, we have: $\mathrm{Var}(x_i^k) = i\sigma_1^2 - \frac{i^2\sigma_1^2}{k} \leqslant \frac{k\sigma_1^2}{4} \leqslant \frac{n\sigma_1^2}{4}$. Next, we prove:

$$\mathrm{P}(|x_i^k| \leqslant \beta, \forall 1 < k \leqslant n, 1 \leqslant i < k) \geqslant [1 + (n-1)nF(\frac{2\beta}{\sqrt{n}\sigma_1}) - (n-1)n]^+ \qquad (6.16)$$

and $F(x) = \mathrm{P}[y < x]$, where y follows a standard normal distribution.

We further define $F_{|x_i^k|}$ as $F_{|x_i^k|}(x) = \mathrm{P}[|x_i^k| < x]$ and $F_{x_i^k}(x)$ as $F_{x_i^k}(x) = \mathrm{P}[x_i^k < x]$. We apply the Fréchet-Hoeffding Bounds on the left side of Formula (6.16):

$$\mathrm{P}(|x_i^k| \leqslant \beta, \forall 1 < k \leqslant n, 1 \leqslant i < k) \geqslant [1 - \frac{n(n-1)}{2} + \sum_{k=2}^{n}\sum_{i=1}^{k-1} F_{|x_i^k|}(\beta)]^+ \qquad (6.17)$$

Since $x_i^k \sim \mathcal{N}(0, i\sigma_1^2 - \frac{i^2\sigma_1^2}{k})$, we have:

$$F_{|x_i^k|}(\beta) = F_{x_i^k}(\beta) - F_{x_i^k}(-\beta) = 2F_{x_i^k}(\beta) - 1 = 2F(\frac{\beta}{\sqrt{i\sigma_1^2 - \frac{i^2\sigma_1^2}{k}}}) - 1 \geqslant 2F(\frac{\beta}{\sqrt{\frac{n\sigma_1^2}{4}}}) - 1. \tag{6.18}$$

By substituting Formula (6.18) into Formula (6.17), we proved the following:

$$\mathrm{P}(|x_i^k| \leqslant \beta, \forall 1 < k \leqslant n, 1 \leqslant i < k) \geqslant [1 + (n-1)nF(\frac{2\beta}{\sqrt{n}\sigma_1}) - (n-1)n]^+.$$

Thus, in order to make sure $\mathrm{P}(|x_i^k| \leqslant \beta, \forall 1 < k \leqslant n, 1 \leqslant i < k) > 2\alpha - 1$, it suffices to have:

$$1 + n(n-1)F(\frac{2\beta}{\sqrt{n}\sigma_1}) - n(n-1) > 2\alpha - 1.$$

Solving above inequality gives: $\beta > z_{\frac{2\alpha-2+n(n-1)}{n(n-1)}} \frac{\sqrt{n}\sigma_1}{2}$.

- **Type B error:** Let us suppose that we are at day $n+m$ and have entered the second demand distribution for m days. This is indicative of the fact that the first demand distribution lasts for n days. If we detect a distribution change at day $n+m$ with a probability of at least α', the Type B error is guaranteed to be less than $1-\alpha'$ at day $m+n$. Because

$$\mathrm{P}(\max(|x_t|) > \beta) > \mathrm{P}(|x_n| > \beta), \quad \forall 1 \leqslant t \leqslant m+n.$$

To make sure $\mathrm{P}(\max(|x_t|) > \beta) > \alpha'$, we select β so that

$$\mathrm{P}(|x_n| > \beta) > \alpha'. \tag{6.19}$$

The distribution of the variable x_n can be calculated in the following manner: Recall $x_n = \sum_{t=1}^{n}(d_t - \lambda)$, and $\lambda = \frac{\sum_{t=1}^{m+n} d_t}{m+n}$,

$$x_n = \sum_{t=1}^{n} d_t - \frac{n}{m+n}\sum_{t=1}^{m+n} d_t = \frac{m}{m+n}\sum_{t=1}^{n} d_t - \frac{n}{m+n}\sum_{t=n+1}^{n+m} d_t.$$

Since $\{d_1, \cdots, d_n\}$ come from the distribution $\mathcal{N}(\lambda_1, \sigma_1^2)$ and $\{d_{n+1}, \cdots, d_{m+n}\}$ come from the distribution $\mathcal{N}(\lambda_2, \sigma_2^2)$. Therefore, variable x_n follows the distribution $\mathcal{N}(\frac{mn}{m+n}(\lambda_1 - \lambda_2), \frac{nm(m\sigma_1^2 + n\sigma_2^2)}{(m+n)^2})$. Formula (6.19) is equivalent to:

$$\mathrm{P}(|x_n| > \beta) = \mathrm{P}(x_n < -\beta) + \mathrm{P}(x_n > \beta) = 1 - \mathrm{P}(x_n < \beta) + \mathrm{P}(x_n < -\beta) > \alpha'.$$

This leads to $F(\frac{\beta - \frac{mn}{m+n}(\lambda_1 - \lambda_2)}{\sqrt{\frac{nm(m\sigma_1^2 + n\sigma_2^2)}{(m+n)^2}}}) - F(\frac{-\beta - \frac{mn}{m+n}(\lambda_1 - \lambda_2)}{\sqrt{\frac{nm(m\sigma_1^2 + n\sigma_2^2)}{(m+n)^2}}}) < 1 - \alpha'$.

As discussed previously, β allows a trade-off between the noise and bias. A larger value of β reduces the Type A error, whereas a smaller value of β reduces the Type B error. Therefore, β can be tuned accordingly.

6.5 Computational Studies

In this section, we present the computational studies conducted to empirically evaluate the IB and SL approaches. We benchmark these approaches against state-of-the-art approaches. In addition, the optimal policy for our data-driven model is considered unknown. Therefore, we demonstrate the effectiveness of our approaches by comparing them with a policy derived under a relaxed setting, where the distribution information and changes in the demand distributions are known in advance.

6.5.1 Benchmarking the Proposed Approach

We benchmark the IB and SL approaches against three approaches: SAA, RH, and ES. SAA is widely used in data-driven inventory management [78, 77], where the optimal (R, Q) policies are derived according to the observed empirical distribution. RH is a modification of SAA under nonstationary demand [93], where the inventory policy at the end of each day is updated according to the empirical distribution observed in some prescribed horizons. To determine these horizons in our experiments, the average length of the demand seasons is assumed to be known. ES is a powerful demand forecasting method that has achieved

considerable success in practice. In this study, we adopted a single exponential smoothing method because our demand model does not comprise trends or seasonalities. In this approach, we predicted mean μ_t, and variance during the lead time σ_t^2 according to the methods introduced in [120]. Then, we obtained order quantity Q_t and reorder point R_t according to $Q_t = \sqrt{\frac{2K\mu_t}{h}}$ and $R_t = L\mu_t + z_{\frac{p}{p+h}}\sqrt{\sigma}$, where $z_{\frac{p}{p+h}}$ represents the standard normal deviate such that $\mathrm{P}(Z < z_{\frac{p}{p+h}}) = \frac{p}{p+h}$.

We compared our approaches with OPT by assuming that the demand distributions and their changes are known a priori. OPT obtains inventory policies according to the methods introduced in [20]. This method precomputes the (s, S) policies for all possible stationary demand distributions and tabulates the results. Then, it solves the nonstationary problem through a stationary problem by averaging the demand means over an estimate of the expected time between two orders. The corresponding optimal (s, S) is obtained through interpolations from the (s, S) values in the table.

We compare the performances of IB, SL, SAA, RH, and ES under both (R, Q) and (s, S) policies. For convenience, we calculate all (s, S) policies in this study based on (R, Q) policy approximations: $s = R$, $S = R+Q$. Moreover, we consider nine different demand data environments based on the length of the demand seasons and the magnitude of demand variance. The details are provided in Section 6.5.1.1. We show that 1) the proposed approaches, IB and SL, achieve reasonable average performances compared to OPT; they achieve approximately 1.12 times the costs of OPT on average; 2) IB and SL outperform SAA, RH, and ES consistently; 3) IB outperforms SL when the ratio of the difference in means to the variance is large for two consecutive demand seasons; otherwise, SL performs better than IB.

6.5.1.1 Experimental settings We considered nine different demand settings to completely evaluate all approaches. The various demand settings are classified based on two features: 1) the length of the demand season and 2) the variance of the demand distributions. More specifically, we defined three types of demand season lengths: 1) long season, where each season randomly contains $100 - 140$ days, 2) short season, where each season randomly contains $10 - 30$ days, and 3) varying seasons, where each season randomly contains $10 - 100$ days. We also defined three types of demand variances: 1) high variance, in

which the variance of the demand equals two times the mean; 2) medium variance, where the variance of the demand equals the mean; and 3) low variance, where the variance of the demand equals half the mean. In addition, in our experiments, we assume that there exist 10 demand seasons for short and varying season settings and 5 demand seasons for the long season setting.

We randomly selected the demand distributions of different demand seasons from a set of normal distributions with mean $\{10, 30, 50, 70, 90\}$. The inventory parameters are set as follows: holding cost rate $h = 1$, stockout cost rate $p = 50$, fixed ordering cost $K = 10000$, and lead time $L = 7$ days. For RH, the length of each horizon is equal to the average length of the demand seasons (which are assumed to be known for RH) under the corresponding demand settings. All parameters used in each model are tuned based on nine separate datasets under the aforementioned settings. More specifically, we set the smoothing parameter to 0.2 for ES. We assume that SL does not know the distributions and set the threshold value β as $3\sqrt{t}\sqrt{\mu}$ on day t based on Proposition 6 in all experiments, where μ represents the empirical mean of the current time horizon. To make a fair comparison with SAA, we set the parameter θ_1 in the proposed DRO model to 0 in all experiments. In addition, the empirical distribution of the demand during the lead time is estimated by obtaining n data points based on the first nL observed daily demand points (n is chosen as large as possible). When the number of the observed daily demands is less than L, sampling with replacement (bootstrapping) is used to obtain one demand data during the lead time. Note that both the choices of θ_1 and the empirical distributions in our methods are not good. However, we show that the proposed methods still perform better overall because they can distinguish between the demand season changes and the changes caused by randomness. This means that, in practice, with better choices of θ_1 and better estimations of the empirical distribution, our methods can yield even better results than those reported in this study.

Furthermore, we define a term called *cost ratio*,

$$ratio(*) = \frac{Cost(*)}{Cost(\text{OPT})}, \tag{6.20}$$

where $*$ represents any of the methods introduced previously. We use $Cost()$ to represent the total inventory costs of using the corresponding methods and use the OPT costs as the baseline. Thus, we eliminate the variations in costs caused by different demand settings.

6.5.1.2 Comparisons among IB, SL, and OPT First, we compare the performances of the proposed approaches, IB and SL, with OPT. For each method, we simulate a non-stationary demand according to the nine demand settings described above. We conduct 100 experiments under each demand setting and then calculate the average inventory costs incurred during these 900 experiments for all approaches. The detailed results are presented under the (R, Q) policy in Table 6 and the (s, S) policy in Table 7. Although the proposed methods assume that the demand seasons are unknown and highly unpredictable, they still achieve approximately 1.12 times the cost of OPT on average. When the length of the demand season is long, IB achieves only 1.04 times the cost of OPT on average.

Table 6: Results based on (R, Q) policy.

data environment	IB	SL	OPT
long season & low variance	1.04	1.07	1
short season & low variance	1.20	1.14	1
varying season & low variance	1.12	1.14	1
long season & medium variance	1.05	1.06	1
short season & medium variance	1.23	1.15	1
varying season & medium variance	1.14	1.15	1
long season & high variance	1.04	1.05	1
short season & high variance	1.20	1.13	1
varying season & high variance	1.08	1.09	1
Average	1.12	1.11	1

Table 7: Results based on (s, S) policy.

data environment	IB	SL	OPT
long season & low variance	1.06	1.06	1
short season & low variance	1.29	1.16	1
varying season & low variance	1.09	1.09	1
long season & medium variance	1.04	1.06	1
short season & medium variance	1.29	1.17	1
varying season & medium variance	1.11	1.11	1
long season & high variance	1.04	1.06	1
short season & high variance	1.16	1.14	1
varying season & high variance	1.09	1.08	1
Average	1.13	1.10	1

6.5.1.3 Comparisons among IB, SL, SAA, RH, and ES In this section, we compare the performances of the proposed approaches, IB and SL, with SAA, RH, and ES that were introduced at the beginning of Section 6.5.1. For each method, we simulate nonstationary demand according to the nine demand settings, and under each demand setting, we conduct 100 experiments. Finally, we calculate the average inventory costs incurred for these 900

experiments for all approaches. The results are presented with the (R, Q) policy in Table 8 and the (s, S) policy in Table 9. We use the same metric (6.20) for all approaches. As indicated by the results, the proposed methods outperform SAA, RH, and ES consistently under all demand settings.

Table 8: Results based on (R, Q) policy.

	IB	SL	SAA	RH	ES
long season & low variance	1.04	1.07	1.52	1.20	1.10
short season & low variance	1.20	1.14	1.35	1.28	1.27
varying season & low variance	1.12	1.14	1.25	1.19	1.14
long season & medium variance	1.05	1.06	1.52	1.19	1.11
short season & medium variance	1.23	1.15	1.36	1.28	1.28
varying season & medium variance	1.14	1.15	1.28	1.23	1.16
long season & high variance	1.04	1.05	1.31	1.15	1.08
short season & high variance	1.20	1.13	1.23	1.15	1.18
varying season & high variance	1.08	1.09	1.27	1.14	1.17
Average	1.12	1.11	1.34	1.20	1.17

Table 9: Results based on (s, S) policy.

	IB	SL	SAA	RH	ES
long season & low variance	1.06	1.06	1.23	1.13	1.10
short season & low variance	1.29	1.16	1.33	1.26	1.23
varying season & low variance	1.09	1.09	1.23	1.14	1.16
long season & medium variance	1.04	1.06	1.23	1.11	1.09
short season & medium variance	1.29	1.17	1.34	1.27	1.24
varying season & medium variance	1.11	1.11	1.26	1.15	1.17
long season & high variance	1.04	1.06	1.20	1.12	1.08
short season & high variance	1.16	1.14	1.23	1.14	1.21
varying season & high variance	1.09	1.08	1.26	1.11	1.15
Average	1.13	1.10	1.26	1.16	1.16

6.5.1.4 A comparison between IB and SL for different data environments In this section, we conduct a detailed comparison between IB and SL to further investigate their performances in different data environments. We demonstrate that IB is preferred when decision-makers know the uncertainty set of demand distributions and the *bias-variance ratio* is large for two consecutive demand seasons. Otherwise, SL is recommended. Mathematically, suppose that the first and second demand seasons have mean values of μ_1, μ_2 and standard deviations of σ_1, σ_2, while the bias-variance ratio is defined as $\dfrac{|\mu_1 - \mu_2|}{\sigma_1}$.

To eliminate the effects of different inventory parameters, we consider a scenario containing only two consecutive seasons and compare error Types A and B (defined in Section 6.4.3). In all our experiments, Type A errors are measured as the failure rate; that is, the value of the Type A error equals the number of experiments that have Type A errors divided by the total number of experiments. Hence, it ranges from 0 to 1. Type B errors are measured based on the average time required to detect a change in the distribution (i.e., the delay in detecting the switching date in cases where no Type A error occurs). Therefore, the range for Type B errors can be any positive number. Thus, to determine the best approach, we

first ensure that the Type A error is close to 0. Then we consider the Type B error to be as small as possible.

Comparison: We compare the performances of the proposed algorithms in cases where the uncertainty set is available when the bias-variance ratio is 3 or 6. In each setting, we randomly generate 10 datasets by assuming that the first demand distribution is $\mathcal{N}(10^4, 10^4)$ and the second demand distribution is $\mathcal{N}(10300, 10300)$ (for a bias-variance ratio of 3) or $\mathcal{N}(10600, 10600)$ (for a bias-variance ratio of 6). Each demand distribution lasts 50 days. We assume that the uncertainty set of distributions used in the IB approach is:

$$P = \{\mathcal{N}(9400, 9400), \mathcal{N}(9700, 9700), \mathcal{N}(10000, 10000), \mathcal{N}(10300, 10300), \mathcal{N}(10600, 10600)\}$$

(for a bias-variance ratio of 3) or

$$\{\mathcal{N}(9400, 9400), \mathcal{N}(10000, 10000), \mathcal{N}(10600, 10600), \mathcal{N}(11200, 11200), \mathcal{N}(11800, 11800)\}$$

(for a bias-variance ratio of 6). In the SL approach, the threshold value is equal to $3\sqrt{t}\sqrt{\mu}$ on day t, where μ is the mean value for the data in the current demand season (Proposition 6). The results are listed in Tables 10 and 11.

Table 10: Bias-variance ratio of 3.

	Type A error	Type B error
IB	0.6	1.5
SL	0	5.5

Table 11: Bias-variance ratio of 6.

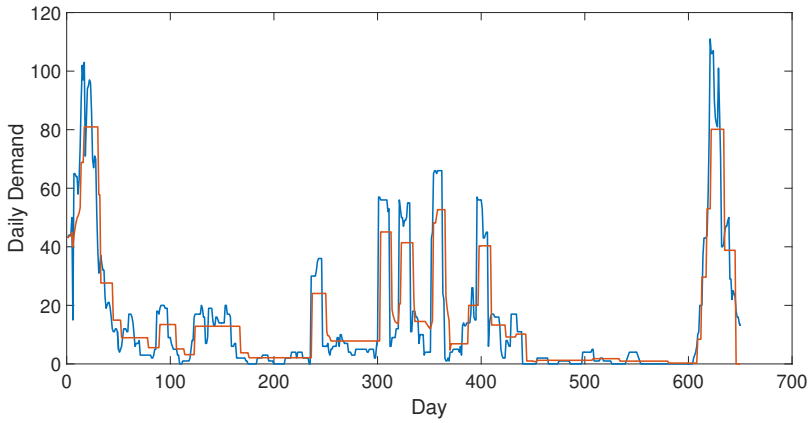
	Type A error	Type B error
IB	0.1	1
SL	0	2.7

The SL approach outperforms the IB approach when the bias-variance ratio is 3. Although the IB approach detects the demand season changes much more rapidly, with an average of 1.5 days, the Type A error is as high as 0.6, which indicates that it mistakenly treats noise as demand season changes in more than half the datasets. The SL approach does not cause any Type A error and achieves a reasonable Type B error of 5.5 days. When the bias-variance ratio is large, the chances of making Type A errors decrease significantly for the IB approach, which detects the demand season changes instantly after day 1. The SL approach takes an average of 2.7 days to identify the changes.

6.5.2 Real-world datasets from one of the world's largest e-commerce websites

In this section, we illustrate the proposed approach on datasets obtained from one of the world's largest e-commerce websites. First, we intuitively show that our method can successfully find the hidden patterns when the daily demand is highly noisy and nonstationary. This dataset comprises 650 demand data points for a specific product. We have no prior knowledge of the possible demand distributions. Thus, we use the SL approach based on Proposition 6, where the threshold values are still equal to $3\sqrt{t}\sqrt{\mu}$ on day t with μ being the mean value for the data in the current detected demand season. Figure 15 demonstrates the demand season (red line) dynamically detected by our approach as well as the demand data (blue line) that belongs to it. The red line captures the primary trend of the blue line. Thus, although the daily demand is highly unpredictable, our method successfully identifies the hidden patterns.

Figure 15: Detected demand seasons.



Second, we evaluate the performance by comparing the total inventory management costs. We adopt the same setting as mentioned in the previous section. The dataset contains 15 products. We can only compare SL with SAA and ES because we do not know the true demand seasons. The results are presented in Table 12. According to the results, our methods and ES outperform SAA. This is because the real-world demand in the dataset is highly nonstationary. Our method also outperforms ES because of its robustness to the noise during each demand season.

Table 12: The average costs for 15 products.

	SL	SAA	ES
Costs (R, Q)	71528	219310	75726
Costs (s, S)	74317	102500	76577

7.0 Summary

In this dissertation, we address four common issues related to data-driven models in OR problem.

In Chapter 3, we derive one tight upper bound of the performance of the scenario approach for the chance constrained programming for a fixed number of data. Then, we propose a linear/conic program to solve the chance constrained programming under the small-data regime. The resulting optimizations have simple closed-form formulations and improve the performances. Additionally, our model (3.5) is shown to be equivalent to DRCCPs under specific settings.

In Chapter 4, We fill a major gap in prior work by proposing the first *scalable* algorithm (meaning it uses a number of variables polynomial in the input size) for maximizing expected matching weight, with *non-identical* failure probabilities. This is an important step forward, as failure probabilities are known to be inhomogeneous–some edges are inherently riskier than others. We provide a mixed-integer linear program for our approach, which is compact and can be solved directly by a general-purpose integer programming solver (e.g., CPLEX, Gurobi, or SCIP).

In Chapter 5, we develop a new DRO framework based on incomplete data sets. The proposed models have two major contributions. First, it provides theoretical guarantees for the stochastic programming under incomplete data set, whereas currently used estimate-then-optimize procedures do not. It represents an integrated analysis of missing data and stochastic optimization, which is fundamentally different from the most popular data-imputation approaches. Second, it extends the study of distributionally robust optimization by introducing ambiguity sets directly based on the partially observed data or incomplete data set. Several kinds of ambiguity sets with their reformulations are discussed.

In Chapter 6, we examined a practical data-driven inventory problem. The demand is considered nonstationary and stochastic. The demand process is *highly unpredictable* meaning future demand can not be reliably predicted through historical features and data. Two solution approaches were presented: (a) the IB approach is proposed for the case in which

an uncertainty set of possible demand distributions is known, and (b) the SL approach is proposed when the uncertainty set is unknown. Both approaches were theoretically analyzed and empirically benchmarked against state-of-the-art heuristics. Real-world data were utilized to verify both approaches. While this study focused on a particular inventory management problem, the frameworks can be used for other dynamic decision-making problems facing nonstationary data that are revealed over time.

Appendix A

Appendix for Chapter 3

A.1 Proof of Theorem 3

Again, we base our proof on one established Lemma 5 in [31] for simplicity.

Lemma 5. *Optimization (3.27) is equivalent to*

$$\begin{aligned}
\min \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N \\
& p_i + Mq_i \geqslant t + s_i, \forall i \in [N] \\
& (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_*, \forall i \in [N], k \in [K] \\
& M(1-q_i) \geqslant t + s_i, \forall i \in [N] \\
& \mathbf{q} \in \{0,1\}^N, \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}, \mathbf{p} \in \mathbb{R}^N,
\end{aligned} \tag{A.1}$$

for $\theta > 0$, where M is a suitably large (but finite) positive constant, and $\mathbf{e}$ is a vector of all ones.

Proof. First of all, when $\theta = 0$, (3.27) is equivalent to the sample average approximation. Therefore, (3.27) and OPT2 are equivalent.

When $\theta > 0$, by Lemma 5, it is equivalent to solve Problem (A.2).

$$\begin{aligned}
\min_{\mathbf{s},t,\mathbf{q},\mathbf{p},\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N, \\
& p_i + Mq_i \geqslant t + s_i, \forall i \in [N], \\
& (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_*, \forall i \in [N], k \in [K], \\
& M(1-q_i) \geqslant t + s_i, \forall i \in [N], \\
& \mathbf{q} \in \{0,1\}^N, \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}.
\end{aligned} \tag{A.2}$$

Suppose the optimal solution for (A.2) satisfies $q_i = 0$ for $i \in I_1$ and $q_i = 1$ for $i \in I_2$. The problem (A.2) can be reformulated to:

$$\begin{aligned}
\min_{\mathbf{s},t,\mathbf{p},\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \epsilon N t + \mathbf{e}^T\mathbf{s} \geqslant \theta N, \\
& (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 \geqslant p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_*, \forall i \in [N], k \in [K], \\
& 0 \geqslant t + s_i, \forall i \in I_2, \\
& p_i \geqslant t + s_i, \forall i \in I_1, \\
& \mathbf{s} \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}.
\end{aligned} \tag{A.3}$$

The Lagrange multiplier of (A.3) equals

$$\begin{aligned}
\min_{\mathbf{s}\leqslant 0,t,\mathbf{p},\mathbf{x}} \quad \max_{\lambda\geqslant 0,\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} \quad & \mathbf{c}^T\mathbf{x} - \lambda(\epsilon N t + \mathbf{e}^T\mathbf{s} - \theta N) - \sum_{i\in I_1}\beta_i(p_i - t - s_i) - \sum_{i\in I_2}\beta_i(-t - s_i) \\
& - \sum_{i\in[N],k\in[K]}\gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 - p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_*\right] \\
s.t. \quad & \mathbf{x} \in \mathcal{X}.
\end{aligned} \tag{A.4}$$

Optimization (A.4) is linear with respect to variables $\mathbf{s}, t, \mathbf{p}$ when $\lambda, \beta_i, \gamma_{i,k}$ are fixed, and is also linear respect to $\lambda, \beta_i, \gamma_{i,k}$ when $\mathbf{s}, t, \mathbf{p}$ are fixed. Therefore, we can switch the min and max based on the minimax theory as shown below.

$$\begin{aligned}
\min_{\mathbf{x}} \max_{\lambda\geqslant 0,\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} \quad \min_{\mathbf{s}\leqslant 0,t,\mathbf{p},} \quad & \mathbf{c}^T\mathbf{x} - \lambda(\epsilon N t + \mathbf{e}^T\mathbf{s} - \theta N) - \sum_{i\in I_1}\beta_i(p_i - t - s_i) - \sum_{i\in I_2}\beta_i(-t - s_i) \\
& - \sum_{i\in[N],k\in[K]}\gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2 - p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_*\right] \\
s.t. \quad & \mathbf{x} \in \mathcal{X}
\end{aligned} \tag{A.5}$$

The original problem is feasible indicating (A.5) is bounded, which means the coefficients of t and $\mathbf{p}$ are zero, and the coefficients of $\mathbf{s}$ are less than zero. Therefore, we first rearrange

each term to find the coefficients:

$$\begin{aligned}\min_{\mathbf{x}} \quad \max_{\lambda\geqslant 0,\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} \quad \min_{\mathbf{s}\leqslant 0,t,\mathbf{p},} \quad & \mathbf{c}^T\mathbf{x} + t(-\lambda\epsilon N + \sum_{i\in I_1}\beta_i + \sum_{i\in I_2}\beta_i) \\ & + \sum_{i\in I_1}(\beta_i-\lambda)s_i + \sum_{i\in I_2}(\beta_i-\lambda)s_i \\ & - \sum_{i\in I_1}\beta_i p_i + \sum_{i\in[N],k\in[K]}\gamma_{i,k}p_i\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* \\ & + \lambda\theta N - \sum_{i\in[N],k\in[K]}\gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2\right] \\ s.t. \quad & \mathbf{x}\in\mathcal{X}.\end{aligned} \tag{A.6}$$

Secondly, we place the corresponding constraints to these coefficients:

$$\begin{aligned} & -\lambda\epsilon N + \sum_{i\in I_1}\beta_i + \sum_{i\in I_2}\beta_i = 0 \Rightarrow \sum_{i\in[N]}\beta_i = \lambda\epsilon N, \\ & \beta_i - \lambda \leqslant 0, \quad \forall i\in[N], \\ & -\beta_i + \sum_{m\in[M]}\gamma_{i,m}\|\mathbf{b}_m - \mathbf{A}_m^T\mathbf{x}_1\|_* = 0, \quad \forall i\in I_1, \\ & \sum_{m\in[M]}\gamma_{i,m}\|\mathbf{b}_m - \mathbf{A}_m^T\mathbf{x}_1\|_* = 0, \quad \forall i\in I_2. \end{aligned}$$

Therefore, (A.5) is equivalent to

$$\begin{aligned}\min_{\mathbf{x}} \quad \max_{\lambda\geqslant 0,\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} \quad & \mathbf{c}^T\mathbf{x} + \lambda\theta N - \sum_{i\in[N],k\in[K]}\gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2\right] \\ s.t. \quad & \sum_{i\in[N]}\beta_i = \lambda\epsilon N, \\ & \beta_i - \lambda \leqslant 0, \quad \forall i\in[N], \\ & -\beta_i + \sum_{k\in[K]}\gamma_{i,k}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* = 0, \quad \forall i\in I_1, \\ & \sum_{k\in[K]}\gamma_{i,k}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* = 0, \quad \forall i\in I_2.\end{aligned}$$

We replace λ with $\lambda = \frac{\sum_{i\in[I]}\beta_i}{\epsilon N}$. In addition, because $\beta_i > 0$ and we assume $N \leqslant \frac{1}{\epsilon}$, we have

$$\beta_i \leqslant \sum_{i\in[N]} \beta_i = \lambda\epsilon N \leqslant \lambda\frac{N}{N} = \lambda$$

for all $i \in [N]$. Therefore, we remove the redundant constraints and obtain

$$\begin{aligned}
\min_{\mathbf{x}} \quad \max_{\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} \quad & \mathbf{c}^T\mathbf{x} + \frac{\sum_{i\in[N]}\beta_i}{\epsilon N}\theta N - \sum_{i\in[N],k\in[K]} \gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2\right] \\
s.t. \quad & -\beta_i + \sum_{k\in[K]} \gamma_{i,k}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* = 0, \quad \forall i \in I_1, \\
& \sum_{k\in[K]} \gamma_{i,k}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* = 0, \quad \forall i \in I_2,
\end{aligned}$$

which is equivalent to

$$\begin{aligned}
\min_{\mathbf{x}} \max_{\beta_i\geqslant 0,\gamma_{i,k}\geqslant 0} & \sum_{i\in I_1}\left\{\frac{\theta}{\epsilon}\sum_{k\in[K]}\gamma_{i,k}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* - \sum_{k\in[K]}\gamma_{i,k}\left[(\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2\right]\right\} \\
& + \sum_{i\in I_2}\frac{\theta}{\epsilon}\beta_i + \mathbf{c}^T\mathbf{x}.
\end{aligned} \tag{A.7}$$

Therefore, to maintain the feasibility, we obtain $I_2 = \emptyset$, and the coefficients of $\gamma_{i,k}$ are less than zero in (A.7).

$$\begin{aligned}
\min_{\mathbf{x}\in\mathcal{X}} \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \frac{\theta}{\epsilon}\|\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1\|_* \leqslant (\mathbf{b}_k - \mathbf{A}_k^T\mathbf{x}_1)^T\hat{\boldsymbol{\xi}}_i - \mathbf{a}_k^T\mathbf{x}_1 - \mathbf{d}_k^T\mathbf{x}_2, \forall k \in [K], i \in [N]. \quad \square
\end{aligned} \tag{A.8}$$

Appendix B

Appendix for Chapter 4

B.1 Proof of Lemma 2

Proof. The expected discounted weight of a chain with k edges is expressed as

$$u(k) = \sum_{i=2}^{k} p_i (\sum_{j=1}^{i-1} w_j) \prod_{j=1}^{i-1} (1 - p_j) + (\sum_{i=1}^{k} w_i) \prod_{i=1}^{k} (1 - p_i).$$

The coefficient on weight w_i (the i^{th} edge in the chain), for any $1 \leq i \leq k$, is expressed as $\prod_{j=1}^{i}(1 - p_j)$. Thus,

$$u(k) = \sum_{i=1}^{k} w_i \prod_{j=1}^{i} (1 - p_j).$$

□

B.2 Optimization (4.8) under one realization of edge existence

Before showing the equivalence between the SAA of (4.8) and (4.9), we first obtain the objective value of (4.8) under one fixed realization of edge existence. The objective value of (4.8) is obtained in (B.1), where we assume the fixed realization is $r_e \in \{0, 1\}$, $e \in E$, where 1 means the edge exists, and 0 otherwise.

In (B.1), we use two sets of variables $o_{ek} \in \{0, 1\}$ and $v_c \in \{0, 1\}$, which indicate the validity of chains and cycles, respectively.

- For any cycle c, c is only valid ($v_c = 1$) if all edges in c exist. Therefore, we restrict $v_c = \min_{e \in c}\{r_e\}$.

- For any chain, an edge e at position k is only valid ($o_{ek} = 1$) if 1) this edge exists ($r_e = 1$) and 2) the prior edges in this chain are valid too. Therefore, we use $o_{ek} \leqslant r_e$ to guarantee this edge e exists. The constraint (B.1c) serves the goal to guarantee the prior edges are valid. To see this point, we consider the following example. Suppose edge $e_1 \in \delta^-(i)$ and $e_2 \in \delta^+(i)$. Both edges are selected in one chain with $y_{e_1,k} = 1$, $y_{e_2,k+1} = 1$. If edge e fails ($r_{e_1} = 0$), then $o_{e_1,k} = 0$ restricting $o_{e_2,k+1}$ to be zero too. Therefore, all the edges after position k in this chain will be invalid.

$$
\begin{aligned}
\min_{\mathbf{y},\mathbf{z},\mathbf{o},\mathbf{v},d} \quad & \left(-\sum_{e\in E}\sum_{k\in\mathcal{K}(e)} w_e y_{ek} o_{ek} - \sum_{c\in C} w_c z_c v_c\right) \\
& + \gamma\left[d + \frac{1}{\alpha}(-\sum_{e\in E}\sum_{k\in\mathcal{K}(e)} w_e y_{ek} o_{ek} - \sum_{c\in C} w_c z_c v_c - d)^+\right] && \text{(B.1a)} \\
s.t. \quad & \{\mathbf{y},\mathbf{z}\} \in \mathcal{X}, && \text{(B.1b)} \\
& \sum_{e\in\delta^-(i),k\in\mathcal{K}(e)} o_{ek} y_{ek} \geq \sum_{e\in\delta^+(i)} o_{e,k+1} y_{e,k+1}, \\
& i \in P, k \in \{1,\ldots,L-1\}, && \text{(B.1c)} \\
& o_{ek} \leqslant r_e, e \in E, k \in \mathcal{K}(e), && \text{(B.1d)} \\
& v_c = \min_{e\in c}\{r_e\}, c \in C; && \text{(B.1e)} \\
& o_{ek} \in [0,1], e \in E, k \in \mathcal{K}(e). && \text{(B.1f)}
\end{aligned}
$$

Optimization (B.1) has a tractable reformulation as shown in Proposition 7.

Proposition 7. *Optimization (B.1) is equivalent to*

$$
\begin{aligned}
\min_{\mathbf{y},\mathbf{z},\mathbf{o},\mathbf{v},d} \quad & \langle \mathbf{w}, \hat{\mathbf{W}}\rangle + \gamma\left[d + \frac{1}{\alpha}(\langle \mathbf{w}, \hat{\mathbf{W}}\rangle - d)^+\right] \\
s.t. \quad & \{\mathbf{y},\mathbf{z}\} \in \mathcal{X}, \mathcal{X}' \\
& \hat{W}_e = -\sum_{k\in\mathcal{K}(e)} O_{e,k} - \sum_{c\in C} \mathbf{1}(e\in c) z_c v_c, \forall e, && \text{(B.2)} \\
& o_{e,k} \leqslant r_e, \forall e, k, \\
& v_c = \min_{e\in c}\{r_e\}, \forall c.
\end{aligned}
$$

where $\mathcal{X}'$ is defined as

$$\mathcal{X}' = \left\{ \begin{array}{l} \sum_{e\in\delta^-(i)\wedge k\in\mathcal{K}(e)} O_{e,k} \geqslant \sum_{e\in\delta^+(i)} O_{e,k+1}, \\ i \in P, k \in \{1,\ldots,L-1\}; \\ O_{e,k} \leqslant y_{e,k}, e \in E, k \in \mathcal{K}(e); \\ O_{e,k} \leqslant o_{e,k}, e \in E, k \in \mathcal{K}(e); \\ o_{e,k}, O_{e,k} \in [0,1], e \in E, k \in \mathcal{K}(e). \end{array} \right\}$$

By comparing Optimization (B.2) and (4.9), it is easy to see that the objective value of (4.9) equals the average realized weights of N realizations. Therefore, we get the conclusion that Optimization (4.9) is equivalent to the SAA of Optimization (4.8).

B.3 A branch and price implementation

In this section, we present a method for scaling our model to graphs with high cycle capacities. Theoretically, the number of cycles of length at most M is $O(|P|^M)$, making explicit representation and enumeration of all cycles infeasible for large enough instances. To solve this problem, we propose a branch and price algorithm, which uses column generation to incrementally consider the possible cycles in a graph. Similar ideas in other kidney exchange problems have also been explored [57, 39]. We show that our formulation with non-identical failure probabilities also scales well with large cycle numbers.

The detailed procedure is introduced as follows; for convenience, we use a vector $\boldsymbol{X}$ to denote the solution $\boldsymbol{X} = [\mathbf{y}, \mathbf{z}]$. First, we define a set $\mathcal{X}_f$ that indicates the fixed components in the solution $\boldsymbol{X}$. For example, $\mathcal{X}_f = \{X_i = 0, X_j = 1\}$ means the i-th and j-th components in $\boldsymbol{X}$ are fixed to 0 and 1, respectively. Our algorithm begins with $\mathcal{X}_f = \emptyset$. Next, an LP relaxation (B.3) based on a (random) subset of cycles C' ($C' \subset C$) is solved.

$$\max_{\mathbf{y},\mathbf{z},\mathbf{O},\mathbf{o}} \quad \sum_{e\in E}\sum_{k\in\mathcal{K}(e)} w_e O_{ek} + \sum_{c\in C'} w_c z_c \left(\prod_{e\in c} 1-p_e\right) \tag{B.3a}$$

$$s.t. \quad \sum_{e\in\delta^-(i)}\sum_{k\in\mathcal{K}(e)} y_{ek} + \sum_{c\in C': i\in c} z_c \leqslant 1, i\in P, \tag{B.3b}$$

$$\sum_{e\in\delta^-(i)\wedge k\in\mathcal{K}(e)} y_{ek} \geqslant \sum_{e\in\delta^+(i)} y_{e,k+1}, \tag{B.3c}$$

$$i\in P, k\in\{1,\ldots,L-1\}, \tag{B.3d}$$

$$\sum_{e\in\delta^+(i)} y_{e1} \leqslant 1, i\in N, \tag{B.3e}$$

$$y_{ek}\in[0,1], e\in E, k\in\mathcal{K}(e), \tag{B.3f}$$

$$z_c\in[0,1], c\in C', \tag{B.3g}$$

$$\sum_{e\in\delta^-(i)\wedge k\in\mathcal{K}(e)} O_{ek} \geqslant \sum_{e\in\delta^+(i)} \frac{O_{e,k+1}}{1-p_e}, \tag{B.3h}$$

$$i\in P, k\in\{1,\ldots,L-1\}, \tag{B.3i}$$

$$O_{ek}\leqslant y_{ek}, e\in E, k\in\mathcal{K}(e), \tag{B.3j}$$

$$O_{ek}\leqslant o_{ek}, e\in E, k\in\mathcal{K}(e), \tag{B.3k}$$

$$O_{ek}\in[0,1], e\in E, k\in\mathcal{K}(e), \tag{B.3l}$$

$$0\leqslant o_{ek}\leqslant 1-p_e, e\in E, k\in\mathcal{K}(e). \tag{B.3m}$$

The following step is to find positive price cycles: cycles that have the potential to improve the objective value if included in the model. The price of a cycle c is defined as $\left[w_c\prod_{e\in c}(1-p_e)-\sum_{i\in c}\lambda_i\right]$, where λ_i are the dual values corresponding to the constraints (B.3b). While there exist any positive price cycles, optimality of the reduced LP has not yet been proved. This can be evidenced from Proposition 8, which can be proved through the strong duality of linear programming.

Proposition 8. *Suppose the dual variables corresponding to the constraints (B.3b) are λ_i, $i\in P$. Then the optimal λ_i, $i\in P$, satisfy*

$$w_c\prod_{e\in c}(1-p_e)-\sum_{i\in c}\lambda_i\leqslant 0.$$

Therefore, we incrementally add (one or more) cycles that have positive prices, i.e. $w_c \prod_{e \in c}(1 - p_e) - \sum_{i \in c} \lambda_i > 0$ into C' until no positive price cycles exist in C. Afterwards, if the optimal solutions of the relaxed LP, i.e. (B.3), are integral, then they are the desired optimal solutions. Otherwise, branching occurs by following the standard branch-and-bound tree search. For example, suppose the i-th component of $\boldsymbol{X}$ is fractional, then we fix $X_i = 0$ or $X_i = 1$. We record these fixed components in set $\mathcal{X}_f$ and repeat above procedures with the new set $\mathcal{X}_f$. We conclude the above discussions in the following Algorithm 4. By running BranchAndPrice(G, $\emptyset$), we obtain the optimal solution.

Algorithm 4 BranchAndPrice(G, $\mathcal{X}_f$)

Generate a subset $C' \subset C$;
Solve LP relaxation (B.3) based on C' and fixed components in $\mathcal{X}_f$;
while $\max_{c \in C} w_c \prod_{e \in c}(1 - p_e) - \sum_{i \in c} \lambda_i > 0$ **do**
    Add $c^* = \text{argmax}_{c \in C} w_c \prod_{e \in c}(1 - p_e) - \sum_{i \in c}$ to C';
end while
$\boldsymbol{X} = [\mathbf{y}, \mathbf{z}] \leftarrow$ solve LP relaxation (B.3) based on C' and fixed components in $\mathcal{X}_f$;
if $\boldsymbol{X}$ is fractional **then**
    Find fractional binary variable $X_i \in \boldsymbol{X}$ closest to 0.5;
    $\boldsymbol{X}_1 = \text{BranchAndPrice}(G, \mathcal{X}_f \cup X_i = 0)$;
    $\boldsymbol{X}_2 = \text{BranchAndPrice}(G, \mathcal{X}_f \cup X_i = 1)$;
    Return $\boldsymbol{X}_1$ or $\boldsymbol{X}_2$ that gives larger objective values in the original Problem (4.6).
else
    Return $\boldsymbol{X}$.
end if

Appendix C

Appendix for Chapter 5

C.1 L1 norm for Nominal data

In this section, we show that the L1 norm is a natural result in our ambiguity set when the data are nominal rather than ordinal. We achieve this goal by revealing relationships between our ambiguity set and the Wasserstein balls in the probability space. The Wasserstein metric has attracted widespread attention in machine learning and optimization recently because of its nice properties to capture the similarities between distributions.

We begin with an introduction of the Wasserstein distance.

Definition 3 (Wasserstein metric)**.** *The Wasserstein distance between distribution* $\mathbb{P}$ *and* $\mathbb{P}'$ *supported on* Ξ *is defined as*

$$W(\mathbb{P},\mathbb{P}') := \inf\{(\int_{\Xi^2} d'(\boldsymbol{\xi},\boldsymbol{\xi}')\Pi(d\boldsymbol{\xi},d\boldsymbol{\xi}')) : \Pi \textit{ is a joint distribution of } \boldsymbol{\xi} \textit{ and } \boldsymbol{\xi}' \textit{ with marginals } \mathbb{P} \textit{ and } \mathbb{P}'\},$$

where d' *is a metric on* Ξ*.*

Correspondingly, the definition of a Wasserstein ball is as follows.

Definition 4. *A Wasserstein ball centered at* $\hat{\mathbb{P}}$ *with radius* θ *is defined as:*

$$\mathcal{B}_\theta(\hat{\mathbb{P}}) = \{\mathbb{P} : W(\mathbb{P},\hat{\mathbb{P}}) \leqslant \theta, \mathbb{P}(\Xi) = 1\}.$$

C.1.0.1 Metric d' for nominal data. The metric d' in Definition 3 measures the "costs" of moving unit mass from $\boldsymbol{\xi}$ to $\boldsymbol{\xi}'$. When the data are nominal, the metric d' is naturally defined as

$$d'(\boldsymbol{\xi}, \boldsymbol{\xi}') = \begin{cases} 0 & \text{if } \boldsymbol{\xi} = \boldsymbol{\xi}' \\ 1 & \text{if } \boldsymbol{\xi} \neq \boldsymbol{\xi}' \end{cases}, \tag{C.1}$$

without considering scaling. The following Proposition 5 characterizes the Wasserstein ball equipped with the metric defined in (C.1) around the nominal distributions.

$$\mathcal{P}' = \left\{ \mathbf{P} = \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{l} \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s} \in \mathcal{S}, \\ \sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau. \end{array} \right\} \tag{C.2}$$

By comparing $\mathcal{Q}$ to the ambiguity set $\mathcal{P}'$, the equivalence is clearly seen, which validates that the L1 norm is a meaningful and natural metric for measuring the deviations in nominal data. (Recall that we use $\boldsymbol{\xi}$ and $\mathbf{s}$ interchangeably.)

Proposition 5. *Suppose one nominal distribution is $\hat{\mathbf{P}}(\mathbf{s})$. Then a Wasserstein ball with radius τ around this nominal distribution is equivalent to (C.3) if the metric defined in (C.1) is used.*

$$\mathcal{Q} = \left\{ \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} \} : \begin{array}{l} \mathrm{P}(\mathbf{s}) = \hat{\mathrm{P}}(\mathbf{s}) + d(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}, \\ 0 \leqslant \hat{\mathrm{P}}(\mathbf{s}) + d(\mathbf{s}) \leqslant 1, \forall \mathbf{s} \in \mathcal{S}, \\ \sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) = 0, \\ \sum_{\mathbf{s}\in\mathcal{S}} |d(\mathbf{s})| \leqslant \tau. \end{array} \right\}, \tag{C.3}$$

Proof. The Wasserstein distance defined in Definition 4 can be formulated as an optimal transportation problem. This indicates that the Wasserstein distance is equivalent to the optimal cost of transporting the probability mass of one distribution $\hat{\mathbf{P}}$ to another distribution $\mathbf{P}$, where the unit transportation cost is determined by a metric d'.

Suppose the cost of transporting the unit probability mass from $\mathbf{s}$ to $\mathbf{s}'$ is $d'(\mathbf{s}, \mathbf{s}')$, which is defined as $d'(\mathbf{s}, \mathbf{s}') = 0$ if $\mathbf{s} = \mathbf{s}'$, otherwise, $d'(\mathbf{s}, \mathbf{s}') = 1$. Then the Wasserstein distance between a distribution $\mathbf{P}$ and $\hat{\mathbf{P}}$, $W(\mathbf{P}, \hat{\mathbf{P}})$, is equivalent to the objective value of the following

optimization problem (C.4a). We use $\hat{\mathbf{P}}$ to denote the initial probability mass. We denote $\mathbf{P}$ as the probability mass after the transportation. In addition, we define $m_{\mathbf{s},\mathbf{s}'}$ as the amount of the probability mass transferred from $\mathbf{s}$ to $\mathbf{s}'$.

$$\min \quad \sum_{\mathbf{s}\in\mathcal{S}}\sum_{\mathbf{s}'\in\mathcal{S}} d'(\mathbf{s},\mathbf{s}')m_{\mathbf{s},\mathbf{s}'} \tag{C.4a}$$

$$s.t. \quad \mathrm{P}(\mathbf{s}) = \hat{\mathrm{P}}(\mathbf{s}) - \sum_{\mathbf{s}'} m_{\mathbf{s},\mathbf{s}'} + \sum_{\mathbf{s}'} m_{\mathbf{s}',\mathbf{s}}, \forall \mathbf{s},\mathbf{s}' \in \mathcal{S}, \tag{C.4b}$$

$$0 \leqslant \mathrm{P}(\mathbf{s}) \leqslant 1, \forall \mathbf{s} \in \mathcal{S}, \tag{C.4c}$$

$$m_{\mathbf{s},\mathbf{s}'} \geqslant 0, \forall \mathbf{s},\mathbf{s}' \in \mathcal{S}. \tag{C.4d}$$

The objective function calculates the total cost of the transportation plan defined by $m_{\mathbf{s},\mathbf{s}'}$. The Constraint (C.4b) restricts the probability mass to be $\mathrm{P}(\mathbf{s})$ after the transportation.

Next, we prove the above optimization is equivalent to:

$$\min \quad \sum_{\mathbf{s}\in\mathcal{S}} |d(\mathbf{s})| \tag{C.5a}$$

$$s.t. \quad \mathrm{P}(\mathbf{s}) = \hat{\mathrm{P}}(\mathbf{s}) + d(\mathbf{s}), \forall \mathbf{s}, \tag{C.5b}$$

$$0 \leqslant \mathrm{P}(\mathbf{s}) \leqslant 1, \forall \mathbf{s} \in \mathcal{S}, \tag{C.5c}$$

$$\sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) = 0. \tag{C.5d}$$

First, by setting $d(\mathbf{s}) = -\sum_{\mathbf{s}\in\mathcal{S}} m_{\mathbf{s},\mathbf{s}'} + \sum_{\mathbf{s}'\in\mathcal{S}} m_{\mathbf{s}',\mathbf{s}}$, it is easy to verify:

$$\sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) = -\sum_{\mathbf{s}\in\mathcal{S}}\sum_{\mathbf{s}'\in\mathcal{S}} m_{\mathbf{s},\mathbf{s}'} + \sum_{\mathbf{s}\in\mathcal{S}}\sum_{\mathbf{s}'\in\mathcal{S}} m_{\mathbf{s}',\mathbf{s}} = 0.$$

Second, the optimal solutions in the above optimization has the following properties. If $d(\mathbf{s}) = -\sum_{\mathbf{s}'\in\mathcal{S}} m_{\mathbf{s},\mathbf{s}'} + \sum_{\mathbf{s}'\in\mathcal{S}} m_{\mathbf{s}',\mathbf{s}} > 0$ for $\mathbf{s}$, it is optimal to set $m_{\mathbf{s},\mathbf{s}'} = 0$ for all $\mathbf{s}'$ in order to minimize the objective function (C.4a). Correspondingly, if $d(\mathbf{s}) < 0$ for $\mathbf{s}$, it is optimal to set $m(\mathbf{s}',\mathbf{s}) = 0$ for all $\mathbf{s}' \in \mathcal{S}$. In addition, $m_{\mathbf{s},\mathbf{s}'} = 0$ always holds if $\mathbf{s} = \mathbf{s}'$ in order to minimize the objective function. Without loss of generality, let $\mathcal{S}_+$ be the set of $\mathbf{s}$ where $d(\mathbf{s}) > 0$ and $\mathcal{S}_-$ be the set of $\mathbf{s}$ where $d(\mathbf{s}) < 0$. Then

$$\sum_{\mathbf{s}\in\mathcal{S}} |d(\mathbf{s})| = \sum_{\mathbf{s}\in\mathcal{S}_+}\sum_{\mathbf{s}',\mathbf{s}'\neq\mathbf{s}} m_{\mathbf{s}',\mathbf{s}} + \sum_{\mathbf{s}\in\mathcal{S}_-}\sum_{\mathbf{s}',\mathbf{s}'\neq\mathbf{s}} m_{\mathbf{s},\mathbf{s}'}$$

In addition, the optimal form of $m_{\mathbf{s},\mathbf{s}'}$ discussed above also indicates

$$\sum_{\mathbf{s}\in\mathcal{S}}\sum_{\mathbf{s}'\in\mathcal{S}} d'(\mathbf{s},\mathbf{s}')m_{\mathbf{s},\mathbf{s}'} = \sum_{\mathbf{s}\in\mathcal{S}_+}\sum_{\mathbf{s}',\mathbf{s}'\neq\mathbf{s}} m_{\mathbf{s}',\mathbf{s}} + \sum_{\mathbf{s}\in\mathcal{S}_-}\sum_{\mathbf{s}',\mathbf{s}'\neq\mathbf{s}} m_{\mathbf{s},\mathbf{s}'} = \sum_{\mathbf{s}\in\mathcal{S}} |d(\mathbf{s})|$$

by following the definition of metric d'. Therefore, Optimization (C.4a) and (C.5a) are equivalent.

Therefore, the Wasserstein distance $W(\mathbf{P},\hat{\mathbf{P}})$ equals to the objective value of Optimization (C.5a). Correspondingly, a Wasserstein ball $W(\mathbf{P},\hat{\mathbf{P}}) \leqslant \tau$ is equivalent to:

$$\mathcal{Q} = \left\{ \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}\}\} : \begin{array}{l} \mathrm{P}(\mathbf{s}) = \hat{\mathrm{P}}(\mathbf{s}) + d(\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}, \\ 0 \leqslant \hat{\mathrm{P}}(\mathbf{s}) + d(\mathbf{s}) \leqslant 1, \forall \mathbf{s}\in\mathcal{S}, \\ \sum_{\mathbf{s}\in\mathcal{S}} d(\mathbf{s}) = 0, \\ \sum_{\mathbf{s}\in\mathcal{S}} |d(\mathbf{s})| \leqslant \tau. \end{array} \right\}.$$

□

C.2 Proof of Proposition 1

Proof. In the following, we define a function $\theta_{mis}(\mathbf{s},\mathbf{s}_{obs}) : \mathcal{S}\times\mathcal{S}_{obs} \to [0,1]$, which represents the probability of observing $\mathbf{s}_{obs}$ for a given scenario $\mathbf{s}$. Here, we define $\mathcal{S}_{obs}$ as a set containing all possible $\mathbf{s}_{obs}$. Therefore, the function θ_{mis} characterizes the missing data mechanism. According to the MAR assumption, $\theta_{mis}(\mathbf{s},\mathbf{s}_{obs})$ is a fixed value for one $\mathbf{s}_{obs}$ and all $\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})$, i.e.

$$\theta_{mis}(\mathbf{s},\mathbf{s}_{obs}) = \theta_{mis}(\mathbf{s}',\mathbf{s}_{obs}), \forall \mathbf{s},\mathbf{s}'\in\mathcal{S}(\mathbf{s}_{obs}),$$

because the missing probability does not depend on the missed values. In addition,

$$\theta_{mis}(\mathbf{s},\mathbf{s}_{obs}) = 0, \forall \mathbf{s}\notin\mathcal{S}(\mathbf{s}_{obs}),$$

because only $\mathbf{s}$ that matches the observed part of $\mathbf{s}_{obs}$ are able to produce $\mathbf{s}_{obs}$ under the missing mechanism θ_{mis}. We use $f(\mathbf{s}_{obs}|\theta_{mis},\mathbf{P})$ to denote the probability mass function (PMF) of observing $\mathbf{s}_{obs}$ given the joint distribution $\mathbf{P}$ and a missing data mechanism defined by θ_{mis}. Recall that $\hat{\mathbf{s}}_{n,obs}$ represents the n-th observed data, or the n-th realization of $\mathbf{s}_{obs}$.

Using the above notations, MLE is formulated as

$$\max_{\mathrm{P}(\mathbf{s})\geqslant 0} \quad \sum_{n=1}^{N} \ln f(\hat{\mathbf{s}}_{n,obs}|\theta_{mis},\mathbf{P}) \tag{C.6a}$$

$$= \max_{\mathrm{P}(\mathbf{s})\geqslant 0} \quad \sum_{n=1}^{N} \ln \left\{ \sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s})\theta_{mis}(\mathbf{s},\hat{\mathbf{s}}_{n,obs}) \right\} \tag{C.6b}$$

$$= \max_{\mathrm{P}(\mathbf{s})\geqslant 0} \quad \sum_{n=1}^{N} \ln \left\{ \Big[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s}) \Big] \theta_{mis}(\mathbf{s},\hat{\mathbf{s}}_{n,obs}) \right\} \tag{C.6c}$$

$$= \max_{\mathrm{P}(\mathbf{s})\geqslant 0} \quad \sum_{n=1}^{N} \ln \left[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s}) \right] + \sum_{n=1}^{N} \ln[\theta_{mis}(\mathbf{s},\hat{\mathbf{s}}_{n,obs})] \tag{C.6d}$$

Equality (C.6b) follows the fact that only $\mathbf{s}$ that match the observed part of the n-th data $\hat{\mathbf{s}}_{n,obs}$, i.e. $\mathbf{s} \in \mathcal{S}(\hat{\mathbf{s}}_{n,obs})$, are able to produce $\hat{\mathbf{s}}_{n,obs}$ under the missing data mechanism θ_{mis}. The Equality (C.6c) follows the fact that $\theta_{mis}(\mathbf{s},\hat{\mathbf{s}}_{n,obs})$ is a fixed value for all $\mathbf{s}$ in $\mathcal{S}(\hat{\mathbf{s}}_{n,obs})$ by the definition of MAR. Therefore, solving (C.6) is equivalent to solving

$$\begin{aligned} \max_{\mathrm{P}(\mathbf{s})} \quad & \sum_{n=1}^{N} \ln \left[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s}) \right] \\ s.t. \quad & \mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}, \\ & \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1. \end{aligned}$$

□

C.3 Proof of Lemma 1

Proof. In the following, we still use θ_{mis} to denote the missing data mechanism (see Appendix C.2). We use $f(\mathbf{s}_{obs}|\theta_{mis},\mathbf{P})$ to denote the PMF of observing $\mathbf{s}_{obs}$ given the joint distribution $\mathbf{P}$ and a missing data mechanism defined by θ_{mis}, which can be represented by

$$f(\mathbf{s}_{obs}|\theta_{mis},\mathbf{P}) = \sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s})\theta_{mis}(\mathbf{s},\mathbf{s}_{obs}). \tag{C.7}$$

Recall that $L[\mathbf{P}] = \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right]$, where the expectation is taken over the probability mass function of $\mathbf{s}_{obs}$, and $F(\mathbf{s}_{obs}|\mathbf{P}) = \sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s})$.

$$
\begin{aligned}
L[\mathbf{P}] - L[\mathbf{P}^*] &= \mathbb{E}\left[\ln \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)}\right] && \text{(C.8a)}\\
&\leqslant \mathbb{E}[\frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} - 1] = \mathbb{E}[\frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)}] - 1 && \text{(C.8b)}\\
&= \sum_{\mathbf{s}_{obs}} \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} f(\mathbf{s}_{obs}|\theta_{mis}, \mathbf{P}^*) - 1 && \text{(C.8c)}\\
&= \sum_{\mathbf{s}_{obs}} \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} \left[\sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}^*(\mathbf{s})\right] \theta_{mis}(\mathbf{s}, \mathbf{s}_{obs}) - 1 && \text{(C.8d)}\\
&= \sum_{\mathbf{s}_{obs}} \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} F(\mathbf{s}_{obs}|\mathbf{P}^*)\theta_{mis}(\mathbf{s}, \mathbf{s}_{obs}) - 1 && \text{(C.8e)}\\
&= \sum_{\mathbf{s}_{obs}} F(\mathbf{s}_{obs}|\mathbf{P})\theta_{mis}(\mathbf{s}, \mathbf{s}_{obs}) - 1 && \text{(C.8f)}\\
&= \sum_{\mathbf{s}_{obs}} \sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s})\theta_{mis}(\mathbf{s}, \mathbf{s}_{obs}) - 1 && \text{(C.8g)}\\
&= \sum_{\mathbf{s}_{obs}} f(\mathbf{s}_{obs}|\theta_{mis}, \mathbf{P}) - 1 = 1 - 1 = 0. && \text{(C.8h)}
\end{aligned}
$$

The Inequality (C.8b) holds because $\ln x \leqslant x - 1$, $\forall x > 0$. The Equality (C.8d) follows the definition of $f(\mathbf{s}_{obs}|\theta_{mis}, \mathbf{P}^*)$, and $\theta_{mis}(\mathbf{s}, \mathbf{s}_{obs})$ is fixed for all $\mathbf{s} \in \mathcal{S}(\mathbf{s}_{obs})$ according to MAR. The last step (C.8h) also again follows the definition of $f(\mathbf{s}_{obs}|\theta_{mis}, \mathbf{P})$.

The inequality $\ln \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} \leqslant \frac{F(\mathbf{s}_{obs}|\mathbf{P})}{F(\mathbf{s}_{obs}|\mathbf{P}^*)} - 1$ is an equality if and only if

$$F(\mathbf{s}_{obs}|\mathbf{P}) = F(\mathbf{s}_{obs}|\mathbf{P}^*)$$

holds almost surely, which only happens when $\mathbf{P} = \mathbf{P}^*$. This is because under the definition of the partially observed data, the dimension of the observed part in $\mathbf{s}_{obs}$ ranges from 1 to I. Then, $\mathbf{P} = \mathbf{P}^*$ holds trivially when the dimension of the observed part equal to I. In conclusion, $L(\mathbf{P})$ attains its maximum uniquely at $\mathbf{P}^*$.

□

C.4 Proof of Lemma 2

Proof. In this proof, we aim to show that the center $\hat{\mathbf{P}}$ of the ambiguity set converges in probability to the true joint distribution $\mathbf{P}^*$, i.e. $\hat{\mathbf{P}} \xrightarrow{p} \mathbf{P}^*$, where we recall their definitions in (C.9) and (C.10) with feasible set $\mathcal{Q}$ (C.11). All expectations in this proof are taken with respect to the probability mass function of $\mathbf{s}_{obs}$.

$$\mathbf{P}^* = \text{argmax}_{\mathbf{P}\in\mathcal{Q}} \quad \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right], \tag{C.9}$$

$$\hat{\mathbf{P}} = \text{argmax}_{\mathbf{P}\in\mathcal{Q}} \quad \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}), \tag{C.10}$$

$$\text{where } \mathcal{Q} = \{\mathbf{P}|\mathrm{P}(\mathbf{s}) \geqslant 0, \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \forall \mathbf{s}\}. \tag{C.11}$$

To this goal, for a given ϵ with

$$0 < \epsilon \leqslant \min_{\mathbf{s}\in\mathcal{S}^+}[\mathrm{P}^*(\mathbf{s})], \text{ where } \mathcal{S}^+ = \{\mathbf{s} \in \mathcal{S}|\mathrm{P}^*(\mathbf{s}) > 0\}, \tag{C.12}$$

we define another feasible set $\mathcal{Q}(\epsilon)$ and two corresponding joint distributions $\mathbf{P}^*(\epsilon)$ and $\hat{\mathbf{P}}(\epsilon)$ as shown below. The feasible set $\mathcal{Q}(\epsilon)$ adds additional constraints $\sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s}) \geqslant \epsilon$ to $\mathcal{Q}$. Obviously, $\mathcal{Q}(\epsilon) = \mathcal{Q}$ when ϵ equals 0.

$$\mathbf{P}^*(\epsilon) = \text{argmax}_{\mathbf{P}\in\mathcal{Q}(\epsilon)} \quad \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right], \tag{C.13}$$

$$\hat{\mathbf{P}}(\epsilon) = \text{argmax}_{\mathbf{P}\in\mathcal{Q}(\epsilon)} \quad \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}), \tag{C.14}$$

$$\text{where } \mathcal{Q}(\epsilon) = \{\mathbf{P}|\mathrm{P}(\mathbf{s}) \geqslant 0, \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s}) \geqslant \epsilon, \forall \mathbf{s}\}. \tag{C.15}$$

In the rest of the paper, ϵ follows the definition in (C.12). Therefore, we aim to first prove $\hat{\mathbf{P}}(\epsilon) \xrightarrow{a.s.} \mathbf{P}^*(\epsilon)$ for an arbitrary small ϵ. Then, we prove $\mathbf{P}^*(\epsilon) = \mathbf{P}^*$ and $\lim_{\epsilon\to 0} \hat{\mathbf{P}}(\epsilon) = \hat{\mathbf{P}}$. Finally, the above established results indicate $\hat{\mathbf{P}} \xrightarrow{a.s.} \mathbf{P}^*$, which proves the desired $\hat{\mathbf{P}} \xrightarrow{p} \mathbf{P}^*$. We present the details below.

- We first prove $\hat{\mathbf{P}}(\epsilon) \xrightarrow{a.s.} \mathbf{P}^*(\epsilon)$. We use the following Lemma.

  Lemma 3 ([95])**.** *If*

1. *A set Ω is compact.*
2. *A function $\tilde{f}(\mathbf{x}, \boldsymbol{\omega})$ is continuous at each $\boldsymbol{\omega} \in \Omega$ for almost all $\mathbf{x}$, and measurable function of $\mathbf{x}$ at each $\boldsymbol{\omega}$.*
3. *There exists a dominating function $d(\mathbf{x})$ such that $\mathbb{E}[d(\mathbf{x})] < \infty$, and*

$$|\tilde{f}(\mathbf{x}, \boldsymbol{\omega})| \leqslant d(\mathbf{x}), \forall \boldsymbol{\omega} \in \Omega.$$

Then, $\mathbb{E}[\tilde{f}(\mathbf{x}, \boldsymbol{\omega})]$ is continuous in $\boldsymbol{\omega}$, and

$$\sup_{\boldsymbol{\omega} \in \Omega} |\frac{1}{N} \sum_{n=1}^{N} \tilde{f}(\mathbf{x}_n, \boldsymbol{\omega}) - \mathbb{E}[\tilde{f}(\mathbf{x}, \boldsymbol{\omega})]| \xrightarrow{a.s.} 0.$$

We let the $\tilde{f} = \ln F(\mathbf{s}_{obs}|\mathbf{P})$ and $\Omega = \mathbb{Q}(\epsilon)$ in the above Lemma, and we verify the conditions as follows.

1. First, it is clear that set $\mathcal{Q}(\epsilon)$ is compact.
2. Second, by definition,

$$\ln F(\mathbf{s}_{obs}|\mathbf{P}) = \ln \sum_{\mathbf{s} \in \mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s}).$$

   Using the following properties of the continuous function, 1) sum of continuous functions is continuous, and 2) function composition of two continuous functions is continuous, we obtain the conclusion that $\ln F(\mathbf{s}_{obs}|\mathbf{P})$ is continuous with respect to $\mathbf{P}$ for every $\mathbf{s}_{obs}$.
3. Following the definition of $\mathcal{Q}(\epsilon)$, we have:

$$\ln \epsilon \leqslant \ln \sum_{\mathbf{s} \in \mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}(\mathbf{s}) = \ln F(\mathbf{s}_{obs}|\mathbf{P}) \leqslant 0.$$

   This proves that there exists a dominating function: $\ln \frac{1}{\epsilon}$, e.g. $|\ln F(\mathbf{s}_{obs}|\mathbf{P})| \leqslant \ln \frac{1}{\epsilon}$.

Therefore, we establish the uniform convergence results according to the above lemma.

$$\sup_{\mathbf{P}\in\mathcal{Q}(\epsilon)} \left| \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}) - \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right] \right| \xrightarrow{a.s.} 0.$$

This indicates that $\frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P})$ converges almost sure to $\mathbb{E}[\ln F(\mathbf{s}_{obs}|\mathbf{P})]$ *uniformly* in set $\mathcal{Q}(\epsilon)$. Therefore, the maximizer of $\frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P})$ converges almost sure to the maximizer of $\mathbb{E}[\ln F(\mathbf{s}_{obs}|\mathbf{P})]$, which is

$$\hat{\mathbf{P}}(\epsilon) \xrightarrow{a.s.} \mathbf{P}^*(\epsilon).$$

- Second, we show $\mathbf{P}^*(\epsilon) = \mathbf{P}^*$.

  In Lemma 1, the unique maximizer $\mathbf{P}^*$ is shown to be the true joint distribution. By the definition of ϵ, i.e. $0 < \epsilon \leqslant \min_{\mathbf{s}\in\mathcal{S}^+} \mathrm{P}^*(\mathbf{s})$, inequality $\sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}^*(\mathbf{s}) \geqslant \epsilon$ holds trivially for $\mathbf{P}^*$ as shown below.

$$\sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}^*(\mathbf{s}) \geqslant \min_{\mathbf{s}\in\mathcal{S}^+} \mathrm{P}^*(\mathbf{s}) \geqslant \epsilon.$$

  Therefore, adding constraints $\sum_{\mathbf{s}\in\mathcal{S}(\mathbf{s}_{obs})} \mathrm{P}^*(\mathbf{s}) \geqslant \epsilon$ to $\mathcal{Q}$ in (C.9) does not change the optimal solution, which indicates that $\mathbf{P}^*$ is also the maximizer of $\max_{\mathbf{P}\in\mathcal{Q}(\epsilon)} \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right]$.

$$\mathbf{P}^* = \mathrm{argmax}_{\mathbf{P}\in\mathcal{Q}(\epsilon)} \quad \mathbb{E}\left[\ln F(\mathbf{s}_{obs}|\mathbf{P})\right]$$

  This completes our proof for $\mathbf{P}^*(\epsilon) = \mathbf{P}^*$.

- Third, we show $\lim_{\epsilon\to 0} \hat{\mathbf{P}}(\epsilon) = \hat{\mathbf{P}}$. This result can be simply verified because $\lim_{\epsilon\to 0} \mathcal{Q}(\epsilon) = \mathcal{Q}$.

Finally, we establish the desired result. Because ϵ is defined as $0 < \epsilon \leqslant \min_{\mathbf{s}\in\mathcal{S}^+} \mathrm{P}^*(\mathbf{s})$, the result

$$\hat{\mathbf{P}}(\epsilon) \xrightarrow{a.s.} \mathbf{P}^*(\epsilon)$$

holds for an arbitrarily small $\epsilon > 0$. Because $\mathbf{P}^*(\epsilon) = \mathbf{P}^*$, we obtain

$$\hat{\mathbf{P}}(\epsilon) \xrightarrow{a.s.} \mathbf{P}^*.$$

In addition, $\lim_{\epsilon\to 0} \hat{\mathbf{P}}(\epsilon) = \hat{\mathbf{P}}$. Therefore, we obtain

$$\hat{\mathbf{P}} \xrightarrow{a.s.} \mathbf{P}^*,$$

indicating $\hat{\mathbf{P}} \xrightarrow{p} \mathbf{P}^*$. □

□

C.5 Proof of Theorem 1

Proof. We define $O_{DRO}(\mathbf{x}, N)$ as the objective value of

$$\max_{\mathbf{P}\in\mathcal{P}'} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}) \tag{C.16}$$

under N data of $\mathbf{s}$. Without loss of generality, we denote the optimal solution of (C.16) as $\hat{\mathbf{P}}^* = \{\hat{\mathrm{P}}^*(\mathbf{s}), \forall \mathbf{s}\}$, i.e.

$$\max_{\mathbf{P}\in\mathcal{P}'} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}) = \sum_{\mathbf{s}\in S} \hat{\mathrm{P}}^*(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}).$$

We define

$$O(\mathbf{x}) = \sum_{\mathbf{s}\in S} \mathrm{P}^*(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}), \tag{C.17}$$

where $\mathrm{P}^*(\mathbf{s})$ is the true joint distribution. We first prove

$$\sup_{\mathbf{x}\in\mathcal{X}} \quad |O_{DRO}(\mathbf{x}, N) - O(\mathbf{x})| \xrightarrow{p} 0. \tag{C.18}$$

Because $Q(\mathbf{x}, \mathbf{s})$ is bounded for $\mathbf{x} \in \mathcal{X}$, we assume $|Q(\mathbf{x}, \mathbf{s})| \leqslant C$ without loss of generality. Then, we obtain

$$\sup_{\mathbf{x}\in\mathcal{X}} \quad |O_{DRO}(\mathbf{x}, N) - O(\mathbf{x})| \tag{C.19a}$$

$$= \sup_{\mathbf{x}\in\mathcal{X}} \quad |\sum_{\mathbf{s}\in S} \hat{\mathrm{P}}^*(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}) - \sum_{\mathbf{s}\in S} \mathrm{P}^*(\mathbf{s}) Q(\mathbf{x}, \mathbf{s})| \tag{C.19b}$$

$$= \sup_{\mathbf{x}\in\mathcal{X}} \quad \big|\sum_{\mathbf{s}\in S} [\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}^*(\mathbf{s})] Q(\mathbf{x}, \mathbf{s})\big| \tag{C.19c}$$

$$\leqslant C \sum_{\mathbf{s}\in S} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}^*(\mathbf{s})| \tag{C.19d}$$

$$= C \sum_{\mathbf{s}\in S} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s}) + \hat{\mathrm{P}}(\mathbf{s}) - \hat{\mathrm{P}}^*(\mathbf{s})| \tag{C.19e}$$

$$\leqslant C \sum_{\mathbf{s}\in S} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| + C \sum_{\mathbf{s}\in S} |\hat{\mathrm{P}}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \tag{C.19f}$$

$$\leqslant C \sum_{\mathbf{s}\in S} |\hat{\mathrm{P}}(\mathbf{s}) - \mathrm{P}^*(\mathbf{s})| + C\tau(N). \tag{C.19g}$$

Inequalities (C.19d) and (C.19f) follow the fact that $|a + b| \leqslant |a| + |b|$. Inequality (C.19g) holds because $\{\hat{\mathrm{P}}^*(\mathbf{s}), \forall \mathbf{s}\} \in \mathcal{P}'$, where

$$\mathcal{P}' = \left\{ \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{ll} \sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau(N), & \forall \mathbf{s} \in \mathcal{S}, \\ \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, & \\ \mathrm{P}(\mathbf{s}) \geqslant 0, & \forall \mathbf{s} \in \mathcal{S}. \end{array} \right\} \tag{C.20}$$

The term $C\tau(N)$ in (C.19g) goes to zero by the definition of $\tau(N)$, i.e. $\lim_{N\to\infty} \tau(N) = 0$. The other term satisfies

$$\sum_{\mathbf{s}\in S} |\hat{\mathrm{P}}(\mathbf{s}) - \mathrm{P}^*(\mathbf{s})| \xrightarrow{p} 0$$

because of Lemma 2. Therefore,

$$\sup_{\mathbf{x}\in\mathcal{X}} \quad |O_{DRO}(\mathbf{x}, N) - O(\mathbf{x})| \xrightarrow{p} 0, \tag{C.21}$$

which intuitively means that $O_{DRO}(\mathbf{x}, N)$ converges to $Q(\mathbf{x})$ uniformly in $\mathcal{X}$ when N goes to infinity.

Because $\hat{O}_N$ and O^* are the minimum of $O_{DRO}(\mathbf{x}, N)$ and $O(\mathbf{x})$, respectively. The uniform convergence (C.21) directly indicates $\hat{O}_N \xrightarrow{p} O^*$. We also present one detailed proof of $\hat{O}_N \xrightarrow{p} O^*$ in the following.

We prove by contradiction. Suppose $\hat{O}_N$ does not converge to O^* in probability, which is equivalent to

$$\exists \epsilon > 0, \lim_{N\to\infty} \mathrm{P}(|\hat{O}_N - O^*| > \epsilon) \neq 0$$

Without loss of generality, we assume there exists $\epsilon' > 0$ such that $\lim_{N\to\infty} \mathrm{P}(\hat{O}_N - O^* < -\epsilon') = q$ with $q > 0$ and the corresponding minimizers of $O_{DRO}(\mathbf{x}, N)$ and $O(\mathbf{x})$ are $\hat{\mathbf{x}}_N$ and $\mathbf{x}^*$, which can be formulated as

$$\lim_{N\to\infty} \mathrm{P}(\hat{O}_N - O^* < -\epsilon') = \lim_{N\to\infty} \mathrm{P}(O_{DRO}(\hat{\mathbf{x}}_N, N) - O(\mathbf{x}^*) < -\epsilon') = q.$$

Because $\mathbf{x}^*$ is the minimizer of $O(\mathbf{x})$, we have $O(\hat{\mathbf{x}}_N) \geqslant O(\mathbf{x}^*)$, which indicates

$$O_{DRO}(\hat{\mathbf{x}}_N, N) - O(\hat{\mathbf{x}}_N) \leqslant O_{DRO}(\hat{\mathbf{x}}_N, N) - O(\mathbf{x}^*).$$

Therefore,

$$\lim_{N\to\infty} \mathrm{P}\left(O_{DRO}(\hat{\mathbf{x}}_N, N) - O(\hat{\mathbf{x}}_N) < -\epsilon'\right) \geqslant q. \tag{C.22}$$

However, (C.22) contradicts the uniform convergence result (C.21), which is

$$\sup_{\mathbf{x}\in\mathcal{X}} |O_{DRO}(\mathbf{x}, N) - O(\mathbf{x})| \xrightarrow{p} 0.$$

Therefore, we proved $\hat{O}_N$ converges to O^* in probability, i.e.

$$\hat{O}_N \xrightarrow{p} O^*. \tag{C.23}$$

In addition, if the maximizer $\mathbf{x}^*$ of O^* is unique, we obatin $\hat{\mathbf{x}}_N \xrightarrow{p} \mathbf{x}^*$ directly from (C.23). □ □

C.6 Proof of Proposition 1

Proof. In this proof, we prove the results that $[0, \tau)$ is an asymptotic $(1-\alpha)$ confidence interval for $\sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})|$, where α can be determined through the observed data. In the rest of this proofs, all expectations are taken over the probability mass function of $\mathbf{s}_{obs}$.

Without loss of generality, we assume $\mathrm{P}^*(\mathbf{s}) > 0$ for $\forall \mathbf{s} \in \mathcal{S}$ in this proof for convenience. This assumption is valid because if $\mathrm{P}^*(\mathbf{s}) = 0$ for some $\mathbf{s}$ meaning this scenario $\mathbf{s}$ never happens, then $\hat{\mathrm{P}}(\mathbf{s})$ equal to 0 trivially in Optimization (C.6) when the sample size is large. Therefore, it will not affect our discussions for the asymptotic behaviors of $\sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})|$.

In order to determine α for a given $\tau > 0$, we first prove that the difference between the center of the ambiguity set and the true unknown joint distribution, i.e. $\hat{\mathbf{P}} - \mathbf{P}^*$, converges in distribution (weakly) to a normal distribution. Recall the former definitions

$$L_N(\mathbf{P}) = \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}) = \frac{1}{N}\sum_{n=1}^{N} \ln[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s})], \tag{C.24}$$

$$L(\mathbf{P}) = \lim_{N\to\infty} \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}) = \mathbb{E}[\ln F(\mathbf{s}_{obs}|\mathbf{P})]. \tag{C.25}$$

Recall that the joint distribution $\mathbf{P} \in [0,1]^{|\mathcal{S}|}$ has $|\mathcal{S}|$ dimensions, and the nominal distribution $\hat{\mathbf{P}}$ is defined as the optimal solution of:

$$\max \quad L_N(\mathbf{P}) \tag{C.26a}$$

$$s.t. \quad \mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}, \tag{C.26b}$$

$$\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1. \tag{C.26c}$$

We replace one $\mathrm{P}(\mathbf{s}^*)$ with

$$\mathrm{P}(\mathbf{s}^*) = 1 - \sum_{\mathbf{s}\in\mathcal{S},\mathbf{s}\neq\mathbf{s}^*} \mathrm{P}(\mathbf{s}) \tag{C.27}$$

in $L_N(\mathbf{P})$, where $\mathbf{s}^*$ is an arbitrary element in $\mathcal{S}$. Originally, $L_N(\mathbf{P})$ is a function of $|\mathcal{S}|$ variables. After replacing $\mathrm{P}(\mathbf{s}^*)$ (C.27), we denote it as $L_N(\mathbf{P}_0)$ for clarification, which is a function of $|\mathcal{S}|-1$ variables. Same procedure is applied to $L(\mathbf{P})$, and we denote the resulting

$L(\mathbf{P})$ as $L(\mathbf{P}_0)$. We also define $\hat{\mathbf{P}}_0$ and $\mathbf{P}_0^*$ as the vectors of $\hat{\mathbf{P}}$ and $\mathbf{P}^*$ excluding $\hat{\mathrm{P}}(\mathbf{s}^*)$ and $\mathrm{P}^*(\mathbf{s}^*)$, respectively. In this way, we eliminate the constraint of $\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1$ in (C.26a).

When N is large, constraints $\{\mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}\}$ are redundant in Optimization (C.26a); this is because terms $\{\ln \mathrm{P}(\mathbf{s}), \forall \mathbf{s}\}$ appear in $L_N(\mathbf{P})$, where $\ln \mathrm{P}(\mathbf{s})$ goes to $-\infty$ when $\mathrm{P}(\mathbf{s})$ approaches 0. Therefore, we can also eliminate the constraints of $\{\mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}\}$ in (C.26a).

Following the above discussions, we define ∇ as a standard Laplace operator with dimension $|\mathcal{S}| - 1$. Because $\hat{\mathbf{P}}_0$ is the maximizer of $L_N(\mathbf{P}_0)$, we have $\nabla L_N(\hat{\mathbf{P}}_0) = 0$. We expand the function $\nabla L_N(\hat{\mathbf{P}}_0)$ by the Taylor series at point $\mathbf{P}_0^*$, which is valid because L_N is a smooth function having derivatives of all orders everywhere in its domain according to (C.24). Therefore, we have

$$\nabla L_N(\hat{\mathbf{P}}_0) = \nabla L_N(\mathbf{P}_0^*) + \nabla^2 L_N(\mathbf{P}_0')(\hat{\mathbf{P}}_0 - \mathbf{P}_0^*) = 0, \tag{C.28}$$

with one $\mathbf{P}_0' \in [\mathbf{P}_0^*, \hat{\mathbf{P}}_0]$. Because $L_N(\mathbf{P}_0')$ contains $|\mathcal{S}| - 1$ independent variables, Matrix $\nabla^2 L_N(\mathbf{P}_0')$ has full rank and is invertible. From (C.28) we get

$$\hat{\mathbf{P}}_0 - \mathbf{P}_0^* = -[\nabla^2 L_N(\mathbf{P}_0')]^{-1} \nabla L_N(\mathbf{P}_0^*) \text{ and } \sqrt{N}(\hat{\mathbf{P}}_0 - \mathbf{P}_0^*) = -[\nabla^2 L_N(\mathbf{P}_0')]^{-1} \sqrt{N} \nabla L_N(\mathbf{P}_0^*). \tag{C.29}$$

1. We first check the term $\sqrt{N}\nabla L_N(\mathbf{P}_0^*)$. In Lemma 2, we proved $\mathbf{P}_0^*$ is the maximizer of $L(\mathbf{P}_0)$, which is equivalent to $\nabla L(\mathbf{P}_0^*) = 0$ (same Laplace operator). Accordingly, we reformulate the numerator of (C.29) as

$$\begin{aligned}
&\sqrt{N}\nabla L_N(\mathbf{P}_0^*) = \sqrt{N}(\nabla L_N(\mathbf{P}_0^*) - 0) = \sqrt{N}\left(\nabla L_N(\mathbf{P}_0^*) - \nabla L(\mathbf{P}_0^*)\right) \\
&\overset{(*)}{=} \sqrt{N}\left[\nabla \frac{1}{N}\sum_{n=1}^{N} \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0^*) - \nabla \mathbb{E}[\ln F(\mathbf{s}_{obs}|\mathbf{P}_0^*)]\right] \\
&= \sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N} \nabla \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0^*) - \mathbb{E}[\nabla \ln F(\mathbf{s}_{obs}|\mathbf{P}_0^*)]\right].
\end{aligned}$$

   Above equality (*) holds by definitions (C.24) and (C.25). In the last step, we can interchange the integration ($\mathbb{E}$) and differentiation (∇) because the support of $\mathbf{s}_{obs}$ is finite.

By using Central Limit Theorem and defining $\boldsymbol{\phi}_{\mathbf{P}_0^*} = \nabla \ln F(\mathbf{s}_{obs}|\mathbf{P}_0^*)$ and $\boldsymbol{\phi}^n_{\mathbf{P}^*_-} = \nabla \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0^*)$ for convenience, we conclude

$$\sqrt{N}\nabla L_N(\mathbf{P}_0^*) = \sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N}\nabla \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0^*) - \mathbb{E}[\nabla \ln F(\mathbf{s}_{obs}|\mathbf{P}_0^*)]\right] \tag{C.30}$$

$$= \sqrt{N}\left[\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}^n_{\mathbf{P}_0^*} - \mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*})\right] \xrightarrow{d} \mathcal{N}(\mathbf{0}, \mathrm{Cov}(\boldsymbol{\phi}_{\mathbf{P}_0^*})). \tag{C.31}$$

Recall that $\nabla L(\mathbf{P}_0^*) = 0$ is equivalent to $\mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}) = 0$ by definition (C.25) and, the covariance matrix of a random variable $\mathbf{x}$ is defined as $\mathrm{Cov}(\mathbf{x}) = \mathbb{E}(\mathbf{x}\mathbf{x}^T) - \mathbb{E}^2(\mathbf{x})$. Therefore, we obtain

$$\mathcal{N}(\mathbf{0}, \mathrm{Cov}(\boldsymbol{\phi}_{\mathbf{P}_0^*})) = \mathcal{N}(\mathbf{0}, \mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T) - \mathbb{E}^2(\boldsymbol{\phi}_{\mathbf{P}_0^*})) = \mathcal{N}(\mathbf{0}, \mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)). \tag{C.32}$$

2. We next check $\nabla^2 L_N(\mathbf{P}_0')$. By the definition of L_N and interchanging the summation and differentiation, we obtain

$$\nabla^2 L_N(\mathbf{P}_0') = \nabla^2 \frac{1}{N}\sum_{n=1}^{N}\ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0') = \frac{1}{N}\sum_{n=1}^{N}\nabla^2 \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0').$$

When sample size $N \to \infty$, we proved $\hat{\mathbf{P}}_0 \xrightarrow{p} \mathbf{P}_0^*$ in Lemma 2, which indicates $\mathbf{P}_0' \xrightarrow{p} \mathbf{P}_0^*$ because $\mathbf{P}_0' \in [\mathbf{P}_0^*, \hat{\mathbf{P}}_0]$. Therefore, the law of large number indicates

$$\frac{1}{N}\sum_{n=1}^{N}\nabla^2 \ln F(\hat{\mathbf{s}}_{n,obs}|\mathbf{P}_0') \xrightarrow{p} \mathbb{E}\left[\nabla^2 \ln F(\mathbf{s}_{obs}|\mathbf{P}_0^*)\right] = \mathbb{E}(\nabla\boldsymbol{\phi}_{\mathbf{P}_0^*}). \tag{C.33}$$

In the next, we characterize the asymptotic behaviors of (C.29) based on (C.31) and (C.33) according to Slutsky's Lemma:

Lemma 4 (Slutsky's Lemma)**.** *Let $\{X_n\}$, $\{Y_n\}$ be sequences of vector random variables. If X_n converges in probability to a constant matrix $\mathbf{C} \in \mathbb{R}^{m_1 \times m_2}$; and Y_n converges in distribution to a random variable $Y \in \mathbb{R}^{m_2}$, then $X_n Y_n \xrightarrow{d} \mathbf{C}Y$.*

Following Lemma 4 and (C.29), we obtain

$$\sqrt{N}(\hat{\mathbf{P}}_0 - \mathbf{P}_0^*) = -[\nabla^2 L_N(\mathbf{P}_0')]^{-1}\sqrt{N}\nabla L_N(\mathbf{P}_0^*) \xrightarrow{d} [\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*})]^{-1}\mathcal{N}(\mathbf{0}, \mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)),$$
$$\text{where } [\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*})]^{-1}\mathcal{N}(\mathbf{0}, \mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)) = \mathcal{N}\left(\mathbf{0}, [\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*})]^{-1}\mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)[\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)]^{-1}\right).$$

The second line follows basic algebra. Equivalently, the above formula is equivalent to

$$\hat{\mathbf{P}}_0 - \mathbf{P}_0^* - \mathcal{N}\left(\mathbf{0}, \frac{1}{N}[\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*})]^{-1}\mathbb{E}(\boldsymbol{\phi}_{\mathbf{P}_0^*}\boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)[\mathbb{E}(\nabla \boldsymbol{\phi}_{\mathbf{P}_0^*}{}^T)]^{-1}\right) \xrightarrow{d} 0,$$

which tells that the difference between $(\hat{\mathbf{P}}_0 - \mathbf{P}_0^*)$ and a normal distribution converges in distribution to zero. We denote the asymptotic variance-covariance matrix as $\boldsymbol{\Sigma}_0^*$ for convenience.

$$(\hat{\mathbf{P}}_0 - \mathbf{P}_0^*) - \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}_0^*) \xrightarrow{d} 0$$

Because $\mathrm{P}(\mathbf{s}^*)$ is fulled determined by the rest of the $\mathrm{P}(\mathbf{s})$, i.e.,

$$\mathrm{P}(\mathbf{s}^*) = 1 - \sum_{\mathbf{s}\in\mathcal{S}, \mathbf{s}\neq\mathbf{s}^*} \mathrm{P}(\mathbf{s}),$$

and the summation of normal random variables still follows normal distribution. Therefore, we have

$$(\hat{\mathbf{P}} - \mathbf{P}^*) - \mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma}^*) \xrightarrow{d} 0. \tag{C.34}$$

The new variance-covariance matrix $\boldsymbol{\Sigma}^*$ can be determined according to $\boldsymbol{\Sigma}_0^*$ and $\mathrm{P}(\mathbf{s}^*) = 1 - \sum_{\mathbf{s}\in\mathcal{S}, \mathbf{s}\neq\mathbf{s}^*} \mathrm{P}(\mathbf{s})$. Without loss of generality, we let the last dimension of $\hat{\mathbf{P}}$ denote the probability corresponding to $\mathbf{s}^*$ for convenience. Therefore, the new asymptotic variance-covariance matrix, $\boldsymbol{\Sigma}^*$, can be explicitly written as

$$\boldsymbol{\Sigma}^* = \begin{bmatrix} \boldsymbol{\Sigma}_0^* & \mathbf{v}_{\mathbf{s}^*} \\ \mathbf{v}_{\mathbf{s}^*}^T & var(\mathbf{s}^*) \end{bmatrix},$$

where $var(\mathbf{s}^*) = \mathbf{1}^T\boldsymbol{\Sigma}_0^*\mathbf{1}$ and $\mathbf{v}_{\mathbf{s}^*} = \boldsymbol{\Sigma}_0^*\mathbf{1}$. We use $\mathbf{1}$ to denote a $(|\mathcal{S}| - 1)$-dimensional column vector whose all elements equal to 1.

Therefore, we obtain a lower bound of the probability that the true joint distribution $\mathbf{P}^*$ deviates from $\hat{\mathbf{P}}$ by a distance tolerance τ as follows.

$$\begin{aligned}
&\mathrm{P}(\sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau) \geqslant \mathrm{P}(|\mathrm{P}^*(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \frac{\tau}{|\mathcal{S}|}, \forall \mathbf{s} \in \mathcal{S}) \\
&= \int_{[-\frac{\tau}{|\mathcal{S}|}, \frac{\tau}{|\mathcal{S}|}]^{|\mathcal{S}|}} (2\pi)^{-\frac{|\mathcal{S}|}{2}} \det(\mathbf{\Sigma}^*)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\mathbf{x}^T \mathbf{\Sigma}^{*-1}\mathbf{x}\right) d\mathbf{x} = \alpha.
\end{aligned}$$

This asymptotic variance-covariance matrix $\mathbf{\Sigma}^*$ can be obtained empirically based on the observed information matrix [45, 44]. More specifically, the asymptotic variance-covariance matrix equals the inverse of the observed information matrix, which intuitively can be viewed as a sample-based version of $\mathbf{\Sigma}^*$. We present the details as follows.

Recall the notations: we assume the set $\mathcal{S}^+ = \{\mathbf{s} \in \mathcal{S} | \hat{\mathrm{P}}(\mathbf{s}) > 0\}$ contains b different $\mathbf{s}$, i.e. $|\mathcal{S}^+| = b$. And we denote them as $\mathbf{s}^{\{k\}}$, $k = 1, \cdots, b$. Their corresponding $\hat{\mathrm{P}}(\mathbf{s})$ are denoted with a vector $\mathbf{p} = [p_1, \cdots, p_b]$. In addition, we define a_i as the number of observations of $\mathbf{s}^{\{i\}}$ for $i = 1, \cdots, b$. Suppose we observe q different incomplete data (the number of observed dimensions is less than I) $\mathbf{s}_{obs}^{\{l\}}$, $l = 1, \cdots, q$, and each $\mathbf{s}_{obs}^{\{l\}}$ appears b_l number of times in the data set. We define a vector $\boldsymbol{\delta}_j \in \{0, 1\}^b$, $j = 1, \cdots, q$, where $\delta_{ji} = 1$ if the observed $\mathbf{s}_{obs}^{\{j\}}$ matches the i-th $\mathbf{s}^{\{i\}}$, i.e. $\mathbf{s}^{\{i\}} \odot \boldsymbol{\phi}_{obs}^{\{j\}} = \mathbf{s}_{obs}^{\{j\}}$. (Here $\boldsymbol{\phi}_{obs}^{\{j\}} \in \{0, 1\}^I$ is defined as the corresponding indicator vector (5.5) of observed dimensions in $\mathbf{s}_{obs}^{\{j\}}$.)

Then the likelihood function

$$l = \sum_{n=1}^{N} \ln \left[\sum_{\mathbf{s}\in\mathcal{S}(\hat{\mathbf{s}}_{n,obs})} \mathrm{P}(\mathbf{s}) \right]$$

is equivalent to

$$l = \sum_{i=1}^{b} a_i \ln p_i + \sum_{j=1}^{q} b_j \ln(\boldsymbol{\delta}_j^T \mathbf{p}), \quad \sum_{i=1}^{b} p_i = 1 \tag{C.35}$$

following the above definitions. Intuitively, the first term in (C.35) represents the log-likelihood of the observed complete data and the second term represents the log-likelihood of the incomplete data. We replace p_b with $p_b = 1 - \sum_{i=1}^{b-1} p_i$, and define the corresponding Laplace operator ∇ with dimension $b - 1$.

Therefore, the observed information matrix of (C.35) has dimension $(b-1)\times(b-1)$ and is defined as $\mathbf{\Sigma}_{b-1}^{-1} = -\nabla\nabla^T l$, which can be explicitely reformulated as

$$\mathbf{\Sigma}_{b-1}^{-1} = -\nabla\nabla^T l = diag(\frac{a_1}{p_1^2},\cdots,\frac{a_{b-1}}{p_{b-1}^2}) + \frac{a_{b-1}}{p_{b-1}^2}\begin{bmatrix} 1 & \cdots & 1 \\ \vdots & \ddots & \vdots \\ 1 & \cdots & 1 \end{bmatrix} + \begin{bmatrix} \psi_{11} & \cdots & \psi_{1,b-1} \\ \vdots & \ddots & \vdots \\ \psi_{b-1,1} & \cdots & \psi_{b-1,b-1} \end{bmatrix}$$

with

$$\psi_{ik} = \sum_{j=1}^{q} \frac{b_j(\delta_{ji}-\delta_{jb})(\delta_{jk}-\delta_{jb})}{(\boldsymbol{\delta}_j^T\mathbf{p})^2}.$$

Following the same procedure introduced for $\mathbf{\Sigma}^*$, the $b\times b$-dimensional covariance matrix $\mathbf{\Sigma}\in\mathbb{R}^{b\times b}$ is calulcated according to $p_b = 1-\sum_{i=1}^{b-1} p_i$. That is

$$\mathbf{\Sigma} = \begin{bmatrix} \mathbf{\Sigma}_{b-1} & \mathbf{v}_b \\ \mathbf{v}_b^T & var(b) \end{bmatrix},$$

where $var(b) = \mathbf{1}^T\mathbf{\Sigma}_{b-1}\mathbf{1}$ and $\mathbf{v}_b = \mathbf{\Sigma}_{b-1}\mathbf{1}$. We use $\mathbf{1}$ to denote a $(b-1)$-dimensional column vector whose all elements equal to 1.

Therefore, we find the value α accordingly.

$$\alpha = \int_{[-\frac{\tau}{b},\frac{\tau}{b}]^b} (2\pi)^{-\frac{b}{2}}\det(\mathbf{\Sigma})^{-\frac{1}{2}}\exp\left(-\frac{1}{2}\mathbf{x}^T\mathbf{\Sigma}^{-1}\mathbf{x}\right)d\mathbf{x}. \quad \square$$

$\square$

C.7 Proof of Proposition 2.

Proof. Recall that the proposed model is formulated as (C.36)

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\{\mathrm{P}(\mathbf{s}),\forall \mathbf{s}\in\mathcal{S}\}\in\mathcal{P}'} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s})Q(\mathbf{x},\mathbf{s}), \tag{C.36}$$

with ambiguity set

$$\mathcal{P}' = \left\{ \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}\}: \begin{array}{l} \sum_{\mathbf{s}\in\mathcal{S}} |\mathrm{P}(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s})| \leqslant \tau, \\ \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s}\in\mathcal{S}. \end{array} \right\} \tag{C.37}$$

Therefore, we formulate (C.36) explicitly in (C.38).

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \quad & \max_{\mathrm{P}(\mathbf{s}),r(\mathbf{s})} \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s})Q(\mathbf{x},\mathbf{s}) \\ s.t. \quad & \tau - \sum_{\mathbf{s}\in\mathcal{S}} r(\mathbf{s}) \geqslant 0, \\ & r(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s}) + \mathrm{P}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}\in\mathcal{S}, \\ & r(\mathbf{s}) - \mathrm{P}(\mathbf{s}) + \hat{\mathrm{P}}(\mathbf{s}) \geqslant 0, \forall \mathbf{s}\in\mathcal{S}, \\ & \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ & \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s}\in\mathcal{S}. \end{aligned} \tag{C.38}$$

The Lagrangian of (C.36) is

$$\begin{aligned} \min_{\mathbf{x}\in\mathcal{X}} \max_{\mathrm{P}(\mathbf{s})\geqslant 0, r(\mathbf{s})} \min_{\mathcal{B}'} & \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s})Q(\mathbf{x},\mathbf{s}) - \gamma(\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) - 1) \\ & + \sum_{\mathbf{s}\in\mathcal{S}} w_{\mathbf{s}}(r(\mathbf{s}) - \mathrm{P}(\mathbf{s}) + \hat{\mathrm{P}}(\mathbf{s})) \\ & + \sum_{\mathbf{s}\in\mathcal{S}} l_{\mathbf{s}}(r(\mathbf{s}) - \hat{\mathrm{P}}(\mathbf{s}) + \mathrm{P}(\mathbf{s})) + e(\tau - \sum_{\mathbf{s}\in\mathcal{S}} r(\mathbf{s})), \end{aligned} \tag{C.39}$$

where we use $\mathcal{B}'$ to denote $\{\gamma, w_{\mathbf{s}} \geqslant 0, l_{\mathbf{s}} \geqslant 0, e \geqslant 0\}$ due to the space issue. Following the minimax theorem, Problem (C.39) is equivalent to:

$$\begin{aligned}\min_{\mathbf{x}\in\mathcal{X}} \min_{\mathcal{B}'} \max_{\mathrm{P}(\mathbf{s})\geqslant 0, r(\mathbf{s})} & \sum_{s\in\mathcal{S}} \mathrm{P}(\mathbf{s})[Q(\mathbf{x},\mathbf{s}) - \gamma - w_{\mathbf{s}} + l_{\mathbf{s}}] \\ & - \gamma(-1) + \sum_{\mathbf{s}\in\mathcal{S}} r(\mathbf{s})(w_{\mathbf{s}} + l_{\mathbf{s}} - e) \\ & + e\tau + \sum_{\mathbf{s}\in\mathcal{S}} (w_{\mathbf{s}} - l_{\mathbf{s}})\hat{\mathrm{P}}(\mathbf{s}),\end{aligned} \tag{C.40}$$

Solving the maximization problem directly in (C.40) gives

$$\begin{aligned}\min_{\mathcal{B}} \quad & \gamma + e\tau + \sum_{\mathbf{s}\in\mathcal{S}} (w_{\mathbf{s}} - l_{\mathbf{s}})\hat{\mathrm{P}}(\mathbf{s}) \\ s.t. \quad & Q(\mathbf{x},\mathbf{s}) - \gamma - w_{\mathbf{s}} + l_{\mathbf{s}} \leqslant 0, \forall \mathbf{s} \in \mathcal{S} \\ & w_{\mathbf{s}} + l_{\mathbf{s}} - e = 0, \forall \mathbf{s} \in \mathcal{S}.\end{aligned} \tag{C.41}$$

where $\mathcal{B} = \{\mathbf{x} \in \mathcal{X}, \gamma, w_{\mathbf{s}} \geqslant 0, l_{\mathbf{s}} \geqslant 0, e \geqslant 0\}$. □

C.8 Extensions to two-stage stochastic programming

We extend the results obtained so far to the two-stage stochastic programming.

$$\min_{\mathbf{x}\in\mathcal{X}} \quad f(\mathbf{x}) + \mathbb{E}[Q(\mathbf{x},\mathbf{s})]. \tag{C.42}$$

We make one commonly used assumption [115, 136] to guarantee that the second-stage problem is bounded and always feasible for a given first-stage decision; that is the two-stage stochastic programming has a relatively complete recourse and is bounded. We denote the second-stage optimal cost $Q(\mathbf{x},\mathbf{s})$ explicitly as $\min_{\mathbf{y}\in\mathcal{Y}(\mathbf{x},\mathbf{s})} h(\mathbf{x},\mathbf{y},\mathbf{s})$, with $\mathcal{Y}(\mathbf{x},\mathbf{s})$ being the feasible region for the second-stage decision $\mathbf{y}$ (sometimes we use $\mathcal{Y}$ to denote it for convenience). Then, the corresponding model under the incomplete data is (C.43). We assume the dimensions of $\mathbf{x}$ and $\mathbf{y}$ are m_1 and m_2, $\mathbf{x} \in \mathbb{R}^{m_1}$ and $\mathbf{y} \in \mathbb{R}^{m_2}$.

$$\min_{\mathbf{x}\in\mathcal{X}} \quad f(\mathbf{x}) + \max_{\{\mathrm{P}(\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}\}\in\mathcal{P}} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s})[\min_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s})] \tag{C.43}$$

The objective function of (C.43) minimizes the expectation of the second-stage cost with respect to a worst-case distribution inside the ambiguity set $\mathcal{P}$.

Model (C.43) uses ambiguity set (5.8). We provide one reformulation in Corollary 2 according to Theorem 2 and discuss its tractability.

Corollary 2. *Optimization*

$$\min_{\mathbf{x}\in\mathcal{X}} \quad f(\mathbf{x}) + \max_{\{\mathrm{P}(\mathbf{s}),\forall \mathbf{s}\in\mathcal{S}\}\in\mathcal{P}'} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s})[\min_{\mathbf{y}\in\mathcal{Y}} h(\mathbf{x},\mathbf{y},\mathbf{s})]$$

is equivalent to

$$\begin{aligned} \min_{\mathcal{B}'} \quad & f(\mathbf{x}) + \gamma + e\tau + \sum_{\mathbf{s}\in\mathcal{S}} (w_{\mathbf{s}} - l_{\mathbf{s}})\hat{\mathrm{P}}(\mathbf{s}) \\ s.t. \quad & h(\mathbf{x},\mathbf{y}_{\mathbf{s}},\mathbf{s}) \leqslant \gamma + w_{\mathbf{s}} - l_{\mathbf{s}}, \forall \mathbf{s}\in\mathcal{S}, \\ & \mathbf{y}_{\mathbf{s}} \in \mathcal{Y}(\mathbf{x},\mathbf{s}), \forall \mathbf{s}\in\mathcal{S}, \\ & w_{\mathbf{s}} + l_{\mathbf{s}} - e = 0, \forall \mathbf{s}, \\ & \mathbf{x}\in\mathcal{X}, \end{aligned} \tag{C.44}$$

where $\mathcal{B}' = \{\mathbf{x}, \gamma, w_{\mathbf{s}} \geqslant 0, l_{\mathbf{s}} \geqslant 0, e \geqslant 0\}$.

Tractability. In a classical two-stage stochastic program, the second-stage cost is defined as $h(\mathbf{x},\mathbf{y},\mathbf{s}) = \mathbf{q}(\mathbf{s})^T\mathbf{y}$, and $\mathcal{Y} = \{\mathbf{y}|T(\mathbf{s})\mathbf{x} + W(\mathbf{s})\mathbf{y} \leqslant \mathbf{r}(\mathbf{s})\}$, 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 (C.44) is a single-stage linear program if $\mathcal{X}$ only consists of linear constraints. 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 $\mathcal{X}$, $\mathcal{Y}$ are convex sets or mixed-integer linear sets [34], (C.44) is mathematically tractable.

C.9 Proof of Proposition 4.

Proof. Recall that the proposed model is formulated as

$$\min_{\mathbf{x}\in\mathcal{X}} \quad \max_{\{\mathrm{P}(\mathbf{s}),\forall \mathbf{s}\in\mathcal{S}\}\in\mathcal{P}'} \sum_{\mathbf{s}\in S} \mathrm{P}(\mathbf{s})Q(\mathbf{x},\mathbf{s}), \tag{C.45}$$

with ambiguity set

$$\mathcal{P}' = \left\{ \{\mathrm{P}(\mathbf{s}), \forall \mathbf{s} \in \mathcal{S}\} : \begin{array}{l} (\mathbf{P}-\hat{\mathbf{P}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{P}-\hat{\mathbf{P}}) \leqslant \tau, \\ \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) = 1, \\ \mathrm{P}(\mathbf{s}) \geqslant 0, \quad \forall \mathbf{s} \in \mathcal{S}. \end{array} \right\} \tag{C.46}$$

The Lagrangian is

$$\begin{aligned} &\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathrm{P}(\mathbf{s})} \min_{\boldsymbol{\beta}\geqslant 0, \lambda \geqslant 0, \gamma} \quad \sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) Q(\mathbf{x}, \mathbf{s}) - \gamma (\sum_{\mathbf{s}\in\mathcal{S}} \mathrm{P}(\mathbf{s}) - 1) \\ &\qquad\qquad + \lambda[\tau - (\mathbf{P}-\hat{\mathbf{P}})^T \boldsymbol{\Sigma}^{-1} (\mathbf{P}-\hat{\mathbf{P}})] + \mathbf{P}^T \boldsymbol{\beta} \\ =&\min_{\mathbf{x}\in\mathcal{X}} \max_{\mathrm{P}(\mathbf{s})} \min_{\boldsymbol{\beta}\geqslant 0, \lambda \geqslant 0, \gamma} \quad \lambda\tau + \gamma + \mathbf{P}^T \mathbf{Q}(\mathbf{x}) - \mathbf{P}^T \boldsymbol{\gamma} - \lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1} \mathbf{P} + 2\lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{P}} \\ &\qquad\qquad - \lambda \hat{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{P}} + \mathbf{P}^T \boldsymbol{\beta}, \end{aligned} \tag{C.47}$$

where we define $\boldsymbol{\gamma}$ as an $|S|$-dimension vector whose all elements equal to γ for convenience.

The objective function is concave with respect to $\mathbf{P}$ for fixed $\{\boldsymbol{\beta}, \lambda, \gamma\}$, and is convex with respect to $\{\boldsymbol{\beta}, \lambda, \gamma\}$ for fixed $\mathbf{P}$. Therefore, following the minimax theorem, it is equivalent to (Σ^{-1} is symmetric by definition)

$$\begin{aligned} &\min_{\mathbf{x}\in\mathcal{X}} \min_{\boldsymbol{\beta}\geqslant 0, \lambda \geqslant 0, \gamma} \max_{\mathrm{P}(\mathbf{s})} \quad \lambda\tau + \gamma + \mathbf{P}^T \mathbf{Q}(\mathbf{x}) - \mathbf{P}^T \boldsymbol{\gamma} - \lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1} \mathbf{P} + 2\lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{P}} \\ &\qquad\qquad - \lambda \hat{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} \hat{\mathbf{P}} + \mathbf{P}^T \boldsymbol{\beta} \end{aligned} \tag{C.48}$$

Solving the inner maximization problem, we obtain

$$\begin{aligned} &\mathbf{Q}^T(\mathbf{x}) - \boldsymbol{\gamma}^T - \lambda[\mathbf{P}^T \boldsymbol{\Sigma}^{-1} + \mathbf{P}^T (\boldsymbol{\Sigma}^{-1})^T] + 2\lambda \hat{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} + \boldsymbol{\beta}^T = 0, \\ \Rightarrow &\mathbf{Q}^T(\mathbf{x}) - \boldsymbol{\gamma}^T - 2\lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1} + 2\lambda \hat{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} + \boldsymbol{\beta}^T = 0, \\ \Rightarrow &\mathbf{Q}^T(\mathbf{x}) - \boldsymbol{\gamma}^T + 2\lambda \hat{\mathbf{P}}^T \boldsymbol{\Sigma}^{-1} + \boldsymbol{\beta}^T = 2\lambda \mathbf{P}^T \boldsymbol{\Sigma}^{-1}, \\ \Rightarrow &\mathbf{Q}^T(\mathbf{x}) \boldsymbol{\Sigma} - \boldsymbol{\gamma}^T \boldsymbol{\Sigma} + 2\lambda \hat{\mathbf{P}}^T + \boldsymbol{\beta}^T \boldsymbol{\Sigma} = 2\lambda \mathbf{P}^T, \end{aligned}$$

Therefore, we obtain

$$
\begin{aligned}
\min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\boldsymbol{\gamma}} \max_{\mathrm{P}(\mathbf{s})} \quad & \lambda\tau+\gamma+\mathbf{P}^T\mathbf{Q}(\mathbf{x})-\mathbf{P}^T\boldsymbol{\gamma}-\lambda\mathbf{P}^T\boldsymbol{\Sigma}^{-1}\mathbf{P}+2\lambda\mathbf{P}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
& -\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\mathbf{P}^T\boldsymbol{\beta} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \frac{1}{2\lambda}\left[\mathbf{Q}^T(\mathbf{x})\boldsymbol{\Sigma}-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}+2\lambda\hat{\mathbf{P}}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma}\right](\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+2\lambda\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\boldsymbol{\beta}) \\
& -\frac{1}{2}(\mathbf{Q}^T(\mathbf{x})\boldsymbol{\Sigma}-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}+2\lambda\hat{\mathbf{P}}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma})\boldsymbol{\Sigma}^{-1}\frac{1}{2\lambda}\left[\boldsymbol{\Sigma}\mathbf{Q}(\mathbf{x})-\boldsymbol{\Sigma}\boldsymbol{\gamma}+2\lambda\hat{\mathbf{P}}+\boldsymbol{\Sigma}\boldsymbol{\beta}\right] \\
& \lambda\tau+\gamma-\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \frac{1}{2\lambda}\left[\mathbf{Q}^T(\mathbf{x})\boldsymbol{\Sigma}-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}+2\lambda\hat{\mathbf{P}}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma}\right](\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+2\lambda\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\boldsymbol{\beta}) \\
& -\frac{1}{4\lambda}(\mathbf{Q}^T(\mathbf{x})\boldsymbol{\Sigma}-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}+2\lambda\hat{\mathbf{P}}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma})\left[\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+2\lambda\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\boldsymbol{\beta}\right] \\
& \lambda\tau+\gamma-\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \frac{1}{4\lambda}\left[\mathbf{Q}^T(\mathbf{x})\boldsymbol{\Sigma}-\boldsymbol{\gamma}^T\boldsymbol{\Sigma}+2\lambda\hat{\mathbf{P}}^T+\boldsymbol{\beta}^T\boldsymbol{\Sigma}\right](\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+2\lambda\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\boldsymbol{\beta}) \\
& \lambda\tau+\gamma-\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \frac{1}{4\lambda}\left(\mathbf{Q}^T(\mathbf{x})-\boldsymbol{\gamma}^T+2\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}+\boldsymbol{\beta}^T\right)\boldsymbol{\Sigma}\left(\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+2\lambda\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}+\boldsymbol{\beta}\right) \\
& \lambda\tau+\gamma-\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \lambda\tau+\gamma+\frac{1}{4\lambda}\left(\mathbf{Q}^T(\mathbf{x})-\boldsymbol{\gamma}^T+\boldsymbol{\beta}^T\right)\boldsymbol{\Sigma}\left(\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+\boldsymbol{\beta}\right)+\hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x}) \\
& -\hat{\mathbf{P}}^T\boldsymbol{\gamma}+\hat{\mathbf{P}}^T\boldsymbol{\beta}+\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}}-\lambda\hat{\mathbf{P}}^T\boldsymbol{\Sigma}^{-1}\hat{\mathbf{P}} \\
\Rightarrow \min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad & \lambda\tau+\gamma+\frac{1}{4\lambda}\left(\mathbf{Q}^T(\mathbf{x})-\boldsymbol{\gamma}^T+\boldsymbol{\beta}^T\right)\boldsymbol{\Sigma}\left(\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+\boldsymbol{\beta}\right)+\hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x}) \\
& -\hat{\mathbf{P}}^T\boldsymbol{\gamma}+\hat{\mathbf{P}}^T\boldsymbol{\beta}
\end{aligned}
\tag{C.49}
$$

Define $\mathbf{V}=\mathbf{Q}(\mathbf{x})-\boldsymbol{\gamma}+\boldsymbol{\beta}$, the above optimization is equivalent to

$$
\min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\lambda\geqslant 0,\gamma} \quad \lambda\tau+\gamma+\frac{1}{4\lambda}\mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V}+\hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x})-\hat{\mathbf{P}}^T\boldsymbol{\gamma}+\hat{\mathbf{P}}^T\boldsymbol{\beta}
$$

Because Σ is a positive definite matrix, $\mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V}>0$ if $\mathbf{V}\neq\mathbf{0}$. Therefore, the optimal λ equals to

$$
\lambda=\frac{1}{2\sqrt{\tau}}\sqrt{\mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V}}.
$$

Plugging in this optimal λ gives

$$\min_{\mathbf{x}\in\mathcal{X},\boldsymbol{\beta}\geqslant 0,\gamma} \quad \gamma + \sqrt{\tau}\sqrt{\mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V}} + \hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x}) - \hat{\mathbf{P}}^T\boldsymbol{\gamma} + \hat{\mathbf{P}}^T\boldsymbol{\beta}. \tag{C.50}$$

In addition, $\hat{\mathbf{P}}^T\boldsymbol{\gamma} = \gamma$ because $\hat{\mathbf{P}}$ is a distribution function.

In conclusion, the original optimization is equivalent to

$$\begin{aligned} \min_{\mathbf{x},y,\boldsymbol{\beta},\gamma} \quad & y + \hat{\mathbf{P}}^T\mathbf{Q}(\mathbf{x}) + \hat{\mathbf{P}}^T\boldsymbol{\beta} \\ s.t. \quad & \mathbf{x} \in \mathcal{X}, \\ & \beta_s \geqslant 0, \forall \mathbf{s} \\ & y^2 \geqslant \tau(\mathbf{V}^T\boldsymbol{\Sigma}\mathbf{V}), \\ & y \geqslant 0, \\ & \mathbf{V} = \mathbf{Q}(\mathbf{x}) - \boldsymbol{\gamma} + \boldsymbol{\beta}, \end{aligned} \tag{C.51}$$

where we use $\boldsymbol{\gamma}$ to represent an $|S|$-dimension vector whose all elements equal to γ. The reformulation is a second-order conic programming.

Appendix D

Appendix for Chapter 6

D.1 Proof of Proposition 4 in Section 6.3.1

Proposition 4. *The policy based on demand distribution p_{k_2} achieves the optimal solution if and only if $P(B_{k_2}|A_t) \geqslant \theta$ and*

$$\theta = \frac{\sum_{k \neq k_2} P(B_k|A_t)(C^k_{k_2} - C^k_{k^*})}{C^{k_2}_{k^*} - C^{k_2}_{k_2}}, \text{ where } k^* = \arg\min_{k' \neq k_2}[P(B_{k_2}|A_t)C^{k_2}_{k'} + \sum_{k \neq k_2} C^k_{k'} P(B_k|A_t)].$$

Proof. Proof. Because the policy based on p_{k_2} achieves the lowest cost, the following inequality holds:

$$\mathrm{P}(B_{k_2}|A_t)C^{k_2}_{k_2} + \sum_{k \neq k_2} C^k_{k_2}\mathrm{P}(B_k|A_t) \leqslant \min_{k' \neq k_2}[\mathrm{P}(B_{k_2}|A_t)C^{k_2}_{k'} + \sum_{k \neq k_2} C^k_{k'}\mathrm{P}(B_k|A_t)]. \tag{D.1}$$

Suppose the right side of Eq. (D.1) achieves the minimum at k^*, then Eq. (D.1) is equivalent to

$$\mathrm{P}(B_{k_2}|A_t) \geqslant \frac{\sum_{k \neq k_2} \mathrm{P}(B_k|A_t)(C^k_{k_2} - C^k_{k^*})}{C^{k_2}_{k^*} - C^{k_2}_{k_2}}.$$

By defining θ as the value of the right side of the above inequality, the policy based on p_{k_2} will be chosen if and only if

$$\mathrm{P}(B_{k_2}|A_t) \geqslant \theta.$$

□

D.2 Derivation of $C_{k'}^k$

Lemma 6. *Suppose that we derive the inventory policy according to a demand distribution $p_{k'} = Poisson(\lambda)$ and the actual demand distribution is $p_k = Poisson(\hat{\lambda})$. If $\lambda < \hat{\lambda}$, then the expected cost of this policy is:*

$$C_{k'}^k = \sqrt{\frac{K'h}{2}}(\frac{\hat{\lambda}}{\sqrt{\lambda}} + \sqrt{\lambda}) + hz_\alpha\sqrt{L}\sqrt{\lambda} + (p-h)L(\hat{\lambda}-\lambda) + pz_\alpha\sqrt{L}(\sqrt{\hat{\lambda}} - \sqrt{\lambda}).$$

If $\lambda > \hat{\lambda}$, the expected cost of this policy is:

$$C_{k'}^k = \sqrt{\frac{K'h}{2}}(\frac{\hat{\lambda}}{\sqrt{\lambda}} + \sqrt{\lambda}) + hz_\alpha\sqrt{L}\sqrt{\lambda} + hL(\lambda - \hat{\lambda}),$$

where the fixed cost of placing orders is denoted as K'. The holding cost of each item (per day) is h. The lead time is L and the Type-1 service level is α. The penalty per item (per day) for violating the Type-1 service level α is p.

Proof. Proof. We approximate the Poisson distribution with a mean λ through a normal distribution $\mathcal{N}(\lambda, \lambda)$. The warehouse begins to order when the inventory level hits r and receives these products after L days. The reorder point r is: $r = L\lambda + z_\alpha\sqrt{L\lambda}$ and the order quantity Q equals to: $Q = \sqrt{\frac{2K'\lambda}{h}}$.

When $\hat{\lambda} < \lambda$, the expected inventory level just before the ordered products arriving is $L(\lambda - \hat{\lambda}) + z_\alpha\sqrt{L\lambda}$. After the products arrive, the inventory level becomes $Q + L(\lambda - \hat{\lambda}) + z_\alpha\sqrt{L\lambda}$. After some time, the same procedure continues when the inventory level touches r again. Therefore, the cost of this model can be viewed as an EOQ model with a safety stock $L(\lambda - \hat{\lambda}) + z_\alpha\sqrt{L\lambda}$. The time between orders is $T = \frac{Q}{\hat{\lambda}} = \frac{1}{\hat{\lambda}}\sqrt{\frac{2K'\lambda}{h}}$. So the average cost for $\hat{\lambda} < \lambda$ is

$$\begin{aligned} Cost &= \frac{K'}{T} + h\frac{Q}{2} + hL(\lambda - \hat{\lambda}) + hz_\alpha\sqrt{L\lambda} \\ &= \sqrt{\frac{K'h}{2}}(\frac{\hat{\lambda}}{\sqrt{\lambda}} + \sqrt{\lambda}) + hz_\alpha\sqrt{L\lambda} + hL(\lambda - \hat{\lambda}). \end{aligned}$$

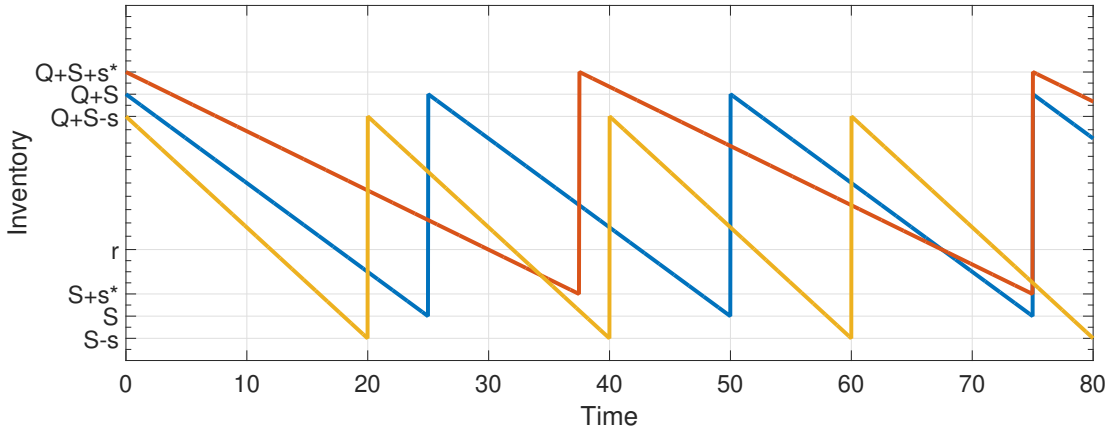
When $\lambda < \hat{\lambda}$, the warehouse still places an order of quantity Q when the inventory level hits r. During the lead time L days, the extra expected number of products is $L(\hat{\lambda} - \lambda)$. The

required safety stock is $z_\alpha\sqrt{L\hat{\lambda}}$. However, the realized safety stock is $z_\alpha\sqrt{L\lambda} - L(\hat{\lambda}-\lambda)$. Therefore, the penalty cost is $p[z_\alpha\sqrt{L\hat{\lambda}} - z_\alpha\sqrt{L\lambda} + L(\hat{\lambda}-\lambda)]$. In addition, the time between orders is $T = \frac{Q}{\hat{\lambda}} = \frac{1}{\hat{\lambda}}\sqrt{\frac{2K'\lambda}{h}}$, so the average inventory cost is:

$$\begin{aligned} Cost &= \frac{K'}{T} + h\frac{Q}{2} + h[z_\alpha\sqrt{L\lambda} - (\hat{\lambda}-\lambda)L] + p[z_\alpha\sqrt{L}(\sqrt{\hat{\lambda}}-\sqrt{\lambda}) + L(\hat{\lambda}-\lambda)] \\ &= \sqrt{\frac{K'h}{2}}(\frac{\hat{\lambda}}{\sqrt{\lambda}} + \sqrt{\lambda}) + hz_\alpha\sqrt{L\lambda} + (p-h)L(\hat{\lambda}-\lambda) + pz_\alpha\sqrt{L}(\sqrt{\hat{\lambda}}-\sqrt{\lambda}). \end{aligned}$$

The following Figure 16 illustrates the expected inventory level of this warehouse. Regardless of the starting point, the expected inventory level after one cycle is exactly the same as shown in Figure 16 because the ordering point r is fixed.

Figure 16: Expected inventory level of three cases. The blue one is $\lambda = \hat{\lambda}$. The yellow one is $\lambda < \hat{\lambda}$. The red one is $\lambda > \hat{\lambda}$. We use $S = z_\alpha\sqrt{L\lambda}$ to denote the safety stock. Q is the order quantity and r is the reorder point.



D.3 Proofs of Lemma 3

In this section, we illustrate the optimal conditions of Problem (6.10) in Section 6.4.

Lemma 3. *For two consecutive demand seasons, the demand data inside the current time horizon have the same demand distribution if and only if* $\lambda = \dfrac{\sum_{j=s}^{t} d_j}{t-s+1}$ *satisfies* $|\sum_{j=s}^{i-1}(\lambda - d_j)| \leqslant \beta,\ \forall i = s+1, \cdots, t+1.$

Proof. *Proof.* Recall that the Problem (6.10) is

$$\min_{\lambda_s, \cdots, \lambda_t} \frac{1}{2}\sum_{j=s}^{t}(d_j - \lambda_j)^2 + \beta \sum_{j=s+1}^{t} |\lambda_j - \lambda_{j-1}|.$$

We find the KKT conditions of Problem (6.10). First, the above problem can be reformulated as follows:

$$\begin{aligned} \min_{\lambda_s, \cdots, \lambda_t, m_{s+1}, \cdots, m_t} \quad & \frac{1}{2}\sum_{j=s}^{t}(d_j - \lambda_j)^2 + \beta \sum_{j=s+1}^{t} |m_j| \\ s.t. \quad & m_j = \lambda_j - \lambda_{j-1}, \quad \forall s+1 \leqslant j \leqslant t. \end{aligned} \tag{D.2}$$

By using the Lagrangian multiplier, we obtain:

$$\frac{1}{2}\sum_{j=s}^{t}(d_j - \lambda_j)^2 + \beta \sum_{j=s+1}^{t} |m_j| + \sum_{j=s+1}^{t} x_j(-m_j + \lambda_j - \lambda_{j-1}). \tag{D.3}$$

- By setting the first differentials of Formula (D.3) with respect to λ_j as zero, we obtain:

$$x_{s+1} = \lambda_s - d_s, \tag{D.4}$$

$$x_j - x_{j+1} = d_j - \lambda_j, \quad \forall j \in \{s+1, \cdots, t-1\}, \tag{D.5}$$

$$x_t = d_t - \lambda_t. \tag{D.6}$$

We can clearly see that the dual variables x_j, $s+1 \leqslant j \leqslant t$ are equivalent to

$$x_j = \sum_{i=s}^{j-1}(\lambda_i - d_i).$$

It must be noted that by Formula (D.6) we have: $x_t = \sum_{j=s}^{t-1}(\lambda_j - d_j) = d_t - \lambda_t$, which is equivalent to $\sum_{j=s}^{t}(\lambda_j - d_j) = 0$. Therefore, we can naturally define one more dual variable $x_{t+1} = \sum_{j=s}^{t}(\lambda_j - d_j)$ that is always zero.

- By setting the first differential (sub-gradient) of Formula (D.3) with respect to m_j as zero, we obtain the following optimality conditions:

$$Sgn(m_j)\beta - x_j = 0, \quad \forall s+1 \leqslant j \leqslant t, \tag{D.7}$$

$$\text{where } Sgn(x) = \begin{cases} 1, x > 0 \\ [-1, 1], x = 0 \\ -1, x < 0 \end{cases} . \tag{D.8}$$

Therefore, if $m_j = \lambda_j - \lambda_{j-1} \neq 0$, x_j must equal β or $-\beta$. Otherwise, there is no constraints on x_j. Thus, we can write it as: $(|x_j| - \beta)(\lambda_j - \lambda_{j-1}) = 0$. In conclusion, we have the KKT conditions:

$$\begin{aligned} & x_{t+1} = 0, \\ & |x_j| \leqslant \beta, \quad s+1 \leqslant j \leqslant t+1, \\ & (|x_j| - \beta)(\lambda_j - \lambda_{j-1}) = 0, \quad s+1 \leqslant j \leqslant t+1, \\ & x_j = \sum_{i=s}^{j-1} (d_i - \lambda_i), \quad s+1 \leqslant j \leqslant t. \end{aligned}$$

Thus, if $\lambda = \dfrac{\sum_{i=s}^{t} d_i}{t-s+1}$ satisfies $|x_j| = |\sum_{i=s}^{j-1}(\lambda - d_i)| \leqslant \beta, \quad \forall j = s+1, \cdots, t+1$, all KKT conditions are met, which indicates that this is an optimal solution.

On the other hand, if a demand distribution can achieve the optimal solution, we have $\lambda = \dfrac{\sum_{i=s}^{t} d_i}{t-s+1}$ because of $x_{t+1} = 0$. Additionally, we require $|x_j| = |\sum_{i=s}^{j-1}(\lambda - d_i)| \leqslant \beta$ for $j = s+1, \cdots, t+1$, since $|x_j| \leqslant \beta$.

D.4 Proof of Proposition 5 in Section 6.4.2

Proposition 5. *Optimization (6.13) is equivalent to*

$$\begin{aligned}
\min_{Q,R,g_k,t,q_k} \quad & \frac{K'(\hat{\mu}+\theta_1)}{Q} + h\frac{Q}{2} + h(R - L\hat{\mu} - L\theta_1) \\
s.t. \quad & (1-\alpha)N't + \mathbf{e}^T\mathbf{g} \geqslant \theta_1 N', \\
& R - \hat{D}_k + Mq_k \geqslant t + g_k, \quad \forall k = 1, \cdots, N' \\
& M(1-q_k) \geqslant t + g_k, \quad \forall k = 1, \cdots, N' \\
& g_k \leqslant 0, t \in \mathbb{R}, q_k \in \{0,1\}, \quad \forall k = 1, \cdots, N',
\end{aligned} \tag{6.14}$$

where we use $\hat{D}_k$ to denote k-th observed L-day demand and $\hat{\mu}$ to denote the empirical mean of the daily demand. Additionally, we suppose that there exists a total of N observed daily demands for the current demand season, as well as N' observed L-day demands. Clearly, $N' = \lfloor N/L \rfloor$. Here, M is a large but bounded number.

We use the following Lemma for the proof.

Lemma 7 ([31])**.** *Consider an individual chance constrained program with a set $S(\mathbf{x}) = \{\xi \in \mathbb{R}^k | (\mathbf{A}\xi + \mathbf{a})^T\mathbf{x} < \mathbf{b}^T\xi\}$ and with $\mathcal{X}$ being compact:*

$$\begin{aligned}
\min \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \mathbf{x} \in \hat{\mathcal{X}} = \{\mathbf{x} \in \mathcal{X} | \sup_{P \in \mathcal{B}_\theta(\hat{P})} \mathbb{P}[\hat{\xi} \notin S(\mathbf{x})] \leqslant \epsilon\}
\end{aligned}$$

It is equivalent to the following mixed integer conic program.

$$\begin{aligned}
\min_{\mathbf{q},\mathbf{s},t,\mathbf{x}} \quad & \mathbf{c}^T\mathbf{x} \\
s.t. \quad & \epsilon Nt + \mathbf{e}^T\mathbf{s} \geqslant \theta N \|\mathbf{b} - \mathbf{A}^T\mathbf{x}\|_\infty, \\
& (\mathbf{b} - \mathbf{A}^T\mathbf{x})^T\hat{\xi}_k - \mathbf{a}^T\mathbf{x} + Mq_k \geqslant t + s_k, \forall k \\
& M(1-q_k) \geqslant t + s_k, \forall k \\
& q_k \in \{0,1\}^N, s_k \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}, \forall k
\end{aligned}$$

where M is some sufficiently large but bounded variable.

Proof. Proof.

The constraint $\min_{p^L \in \mathcal{B}_{\theta_1}(\hat{p}^L)} P^L(R) \geqslant \alpha$ is equivalent to

$$\max_{p^L \in \mathcal{B}_{\theta_1}(\hat{p}^L)} \mathrm{P}(R \leqslant D) \leqslant 1 - \alpha,$$

where D represents the random variable whose CDF is P^L. By viewing ξ as D, ϵ as $1-\alpha$, and $\mathbf{x}$ as R in Lemma 7, we can set $\mathbf{A} = 0$, $\mathbf{a} = -1$, and $\mathbf{b} = -1$ correspondingly. Therefore, the constraint is equivalent to

$$\begin{aligned}
&(1-\alpha)N't + \mathbf{e}^T\mathbf{g} \geqslant \theta_1 N', \\
&-\hat{D}_k - (-1)R + Mq_k \geqslant t + g_k, \forall k, \\
&M(1-q_k) \geqslant t + g_k, \forall k, \\
&q_k \in \{0,1\}^N, g_k \leqslant 0, t \in \mathbb{R}, \mathbf{x} \in \mathcal{X}, \forall k.
\end{aligned}$$

The objective function is directly obtained by the definition of the Wasserstein ball. By the definition, the values of the observed demand can change at most $|\theta_1|$ in average. And the worst-case distribution for the constraint part is achieved when observed demand become larger (because higher reorder points represent higher cost). Therefore, the worst-case distribution has mean value of $\hat{\mu} + \theta_1$. In conclusion, we obtain the final formulation as shown in (6.14).

D.5 Working inventory costs for (R_t, Q_t) policies obtained by the DRO model

This section presents the empirical experiments conducted to study the working inventory costs of the proposed DRO model. To evaluate the out-of-sample expected working inventory cost, we define a variable AD as the average difference between the out-of-sample expected working inventory costs of our methods and the theoretically derived optimal costs. In (D.9), we use λ_t to denote the true mean of the demand distribution on day t and $\hat{Q}_t$ to denote the order quantity determined by the proposed approaches. The values of λ_t are

unknown in our approaches but are known when calculating the theoretical optimal order quantities Q_t.

$$\begin{aligned} AD &= \frac{1}{T}\sum_{t=1}^{T}(\frac{K\lambda_t}{\hat{Q}_t} + h\frac{\hat{Q}_t}{2}) - \frac{1}{T}\min_{Q_t>0}\sum_{t=1}^{T}(\frac{K\lambda_t}{Q_t} + h\frac{Q_t}{2}) \\ &= \frac{1}{T}\sum_{t=1}^{T}(\frac{K\lambda_t}{\hat{Q}_t} + h\frac{\hat{Q}_t}{2}) - \frac{1}{T}\sum_{t=1}^{T}\sqrt{2K\lambda_t h}. \end{aligned} \tag{D.9}$$

During the experiments, the demand is assumed to be revealed sequentially, and we use the proposed approaches to derive their inventory policies. We then calculate and plot the value of AD.

Experiment setting: The time horizon contains four demand seasons. The demand distributions for these four sequential demand seasons are $\mathcal{N}(10000, 10000)$, $\mathcal{N}(10500, 10500)$, $\mathcal{N}(11000, 11000)$, and $\mathcal{N}(10000, 10000)$. Each demand season lasted for N days, and we varied N ($N = 30, 35, \cdots, 100$) to generate 15 scenarios. We conducted 50 experiments for both approaches under each scenario. The IB approach derives the inventory policies according to the uncertainty set, which is set as

$$\{\mathcal{N}(9000, 9000), \mathcal{N}(9500, 9500), \mathcal{N}(10000, 10000), \mathcal{N}(10500, 10500), \mathcal{N}(11000, 11000)\}$$

during the experiments. The SL approach derives inventory policies using the DRO framework., which can be tuned by adjusting θ_1 to guarantee the Type-1 service level. We notice that an appropriate θ_1 causes very little change in AD, which is expected because the working inventory costs are generally not sensitive to order quantities. Therefore, we simply set θ_1 as zero to evaluate the working inventory costs. The parameters for the warehouse are set to ensure that the lead time is less than each order period: holding cost rate $h = 2 \times 10^{-5}$, fixed cost of placing orders $K = 8$ and lead time $L = 3$.

The results are shown in Figure 17. Each graph plots the average values of AD with respect to the length of each demand season. First, we conclude that the value of AD is sublinear with respect to the length of demand seasons when the number of demand seasons is fixed. This property is highly desirable, as it indicates that all methods approach the optimal value as the number of demand data increases. Second, when the demand season lasts for a long period of time, the average values of AD are relatively smaller than the daily cost, *which indicates that the performance of the derived policies is very close to the optimal*

performance. For example, the largest AD is approximately 4×10^{-5} when each demand season lasts for 100 days, whereas the theoretically derived optimal expected daily working inventory cost is $\sqrt{2Kh\lambda}$ ($\lambda = 10000$). Therefore, the ratio of AD to the theoretically derived optimal daily cost, $\frac{AD}{\sqrt{2K\lambda h}}$, is approximately 2×10^{-5}.

Figure 17: The average difference with respect to the number of days in each demand season. (4 demand seasons)

Figure 18: Integrated-Bayesian

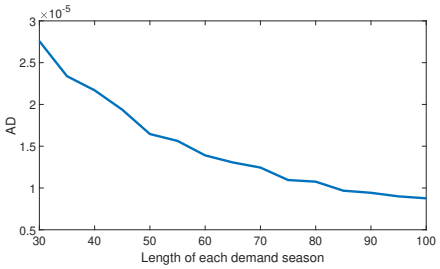
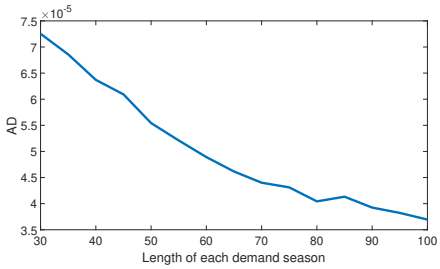


Figure 19: Separate-lasso



D.6 Type-1 service level in DRO model

This section validates our claim that during each demand season, the proposed DRO model improves the out-of-sample Type-1 service level compared to SAA. The Type-1 service level is affected by choice of the θ_1. Therefore, we computed the average out-of-sample Type-1 service level with respect to different values of θ_1 under different demand observation numbers.

Experiment setting: The daily demand distribution is set to $\mathcal{N}(10000, 10000)$, and the lead time is $L = 3$. We vary the number of observations (2, 5, 10, 20) of the demand during the lead time. For each scenario, 100 experiments are conducted, each of which is tested under different values of θ_1. The average Type-1 service levels for the 100 experiments are recorded.

Figure 20: Average Type-1 service level. Yellow line: target service level. Blue line: DRO. Red line: SAA.

Figure 21: 2 observations

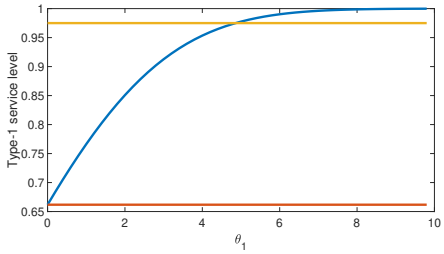


Figure 22: 5 observations

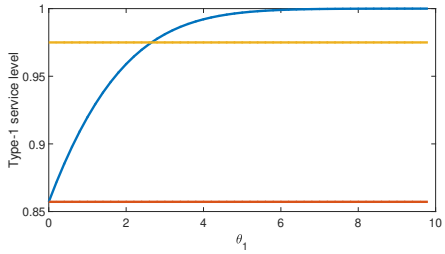


Figure 23: 10 observations

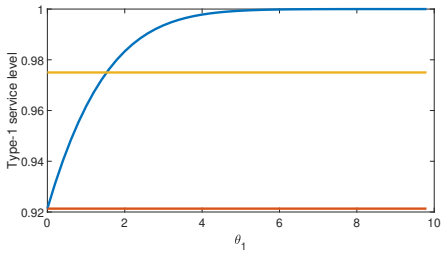
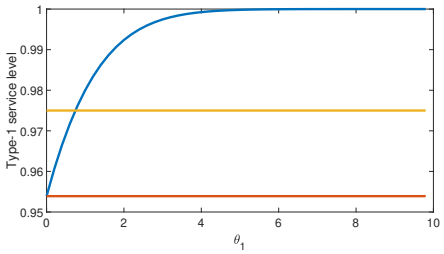


Figure 24: 20 observations



The results are summarized in Figure 20, which indicates that the proposed DRO framework improves the Type-1 service level compared to SAA. Overall, SAA yields a Type-1 service level that is lower than the target ones. This becomes a severe issue when the number of demand observations is small. For example, if there are only 2 observations, the average Type-1 service level achieved by SAA is only approximately 66%. The DRO model improves the Type-1 service level by increasing θ_1, and the optimal value of θ_1 (the intersection of the blue and yellow lines) decreases with respect to the increase in the number of observations. In conclusion, the proposed DRO model provides more robust and better solutions to the reorder points for (R, Q) policies especially at the beginning of each new demand season detected by our approaches.

Bibliography

[1] David Abraham, Avrim Blum, and Tuomas Sandholm. Clearing algorithms for barter exchange markets: Enabling nationwide kidney exchanges. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, pages 295–304, 2007.

[2] Jason Acimovic and Stephen C Graves. Making better fulfillment decisions on the fly in an online retail environment. *Manufacturing & Service Operations Management*, 17(1):34–51, 2015.

[3] Nikhil Agarwal, Itai Ashlagi, Eduardo Azevedo, Clayton R Featherstone, and Ömer Karaduman. Market failure in kidney exchange. *American Economic Review*, 109(11):4026–70, 2019.

[4] Filipe Alvelos, Xenia Klimentova, Abdur Rais, and Ana Viana. A compact formulation for maximizing the expected number of transplants in kidney exchange programs. In *Journal of Physics: Conference Series*, volume 616. IOP Publishing, 2015.

[5] Ross Anderson, Itai Ashlagi, David Gamarnik, and Alvin E Roth. Finding long chains in kidney exchange using the traveling salesman problem. *Proceedings of the National Academy of Sciences*, 112(3):663–668, 2015.

[6] Ronald G Askin. A procedure for production lot sizing with probabilistic dynamic demand. *Aiie Transactions*, 13(2):132–137, 1981.

[7] Anil Aswani, Zuo-Jun Max Shen, and Auyon Siddiq. Data-driven incentive design in the medicare shared savings program. *Operations Research*, 67(4):1002–1026, 2019.

[8] Sven Axsäter, Christoph Schneeweiss, and Edward Silver. *Multi-stage production planning and inventory control*, volume 266. Springer Science & Business Media, 2012.

[9] Mohamed Zied Babai and Yves Dallery. Dynamic versus static control policies in single stage production-inventory systems. *International Journal of Production Research*, 47(2):415–433, 2009.

[10] Gulay Barbarosovglu and Yasemin Arda. A two-stage stochastic programming framework for transportation planning in disaster response. *Journal of the operational research society*, 55(1):43–53, 2004.

[11] Valentin Bartier, Bart Smeulders, Yves Crama, and Frits CR Spieksma. Recourse in kidney exchange programs, 2019. Working paper.

[12] Aharon Ben-Tal, Dick Den Hertog, Anja De Waegenaere, Bertrand Melenberg, and Gijs Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[13] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Robust sample average approximation. *Mathematical Programming*, pages 1–66, 2017.

[14] Dimitris Bertsimas, Vishal Gupta, and Nathan Kallus. Data-driven robust optimization. *Mathematical Programming*, 167(2):235–292, 2018.

[15] Dimitris Bertsimas, Colin Pawlowski, and Ying Daisy Zhuo. From predictive methods to missing data imputation: an optimization approach. *The Journal of Machine Learning Research*, 18(1):7133–7171, 2017.

[16] Anna-Lena Beutel and Stefan Minner. Safety stock planning under causal demand forecasting. *International Journal of Production Economics*, 140(2):637–645, 2012.

[17] Péter Biró, Bernadette Haase-Kromwijk, Tommy Andersson, Eyjólfur Ingi Ásgeirsson, Tatiana Baltesová, Ioannis Boletis, Catarina Bolotinha, Gregor Bond, Georg Böhmig, Lisa Burnapp, et al. Building kidney exchange programmes in europe—an overview of exchange practice and activities. *Transplantation*, 103(7):1514, 2019.

[18] Avrim Blum, John P. Dickerson, Nika Haghtalab, Ariel D. Procaccia, Tuomas Sandholm, and Ankit Sharma. Ignorance is almost bliss: Near-optimal stochastic matching with few queries. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, pages 325–342, 2015.

[19] Avrim Blum, Anupam Gupta, Ariel D. Procaccia, and Ankit Sharma. Harnessing the power of two crossmatches. In *Proceedings of the ACM Conference on Electronic Commerce (EC)*, pages 123–140, 2013.

[20] Srinivas Bollapragada and Thomas E Morton. A simple heuristic for computing nonstationary (s, s) policies. *Operations Research*, 47(4):576–584, 1999.

[21] Stephen Boyd. Data for finance and portfolio optimization, 2019.

[22] Stephen Boyd and Lieven Vandenberghe. *Convex optimization*. Cambridge university press, 2004.

[23] Apostolos N Burnetas and Craig E Smith. Adaptive ordering and pricing for perishable products. *Operations Research*, 48(3):436–443, 2000.

[24] Giuseppe Calafiore and Marco C Campi. Uncertain convex programs: randomized solutions and confidence levels. *Mathematical Programming*, 102(1):25–46, 2005.

[25] Giuseppe C Calafiore and Marco C Campi. The scenario approach to robust control design. *IEEE Transactions on automatic control*, 51(5):742–753, 2006.

[26] Giuseppe Carlo Calafiore. Random convex programs. *SIAM Journal on Optimization*, 20(6):3427–3464, 2010.

[27] Marco C Campi and Simone Garatti. The exact feasibility of randomized solutions of uncertain convex programs. *SIAM Journal on Optimization*, 19(3):1211–1230, 2008.

[28] Marco C Campi and Simone Garatti. A sampling-and-discarding approach to chance-constrained optimization: feasibility and optimality. *Journal of Optimization Theory and Applications*, 148(2):257–280, 2011.

[29] Margarida Carvalho, Xenia Klimentova, Kristiaan Glorie, Ana Viana, and Miguel Constantino. Robust models for the kidney exchange problem. *INFORMS Journal on Computing*, 2020. To appear.

[30] Frank Chen, Zvi Drezner, Jennifer K Ryan, and David Simchi-Levi. Quantifying the bullwhip effect in a simple supply chain: The impact of forecasting, lead times, and information. *Management science*, 46(3):436–443, 2000.

[31] Zhi Chen, Daniel Kuhn, and Wolfram Wiesemann. Data-driven chance constrained programs over wasserstein balls. *arXiv preprint arXiv:1809.00210*, 2018.

[32] Zhi Chen, Melvyn Sim, and Peng Xiong. Robust stochastic optimization. *History*, 2019.

[33] Zhi Chen, Melvyn Sim, and Huan Xu. Distributionally robust optimization with infinitely constrained ambiguity sets. *Operations Research*, 67(5):1328–1344, 2019.

[34] Michele Conforti, Gérard Cornuéjols, and Giacomo Zambelli. *Integer programming*, volume 271. Springer, 2014.

[35] Miguel Constantino, Xenia Klimentova, Ana Viana, and Abdur Rais. New insights on integer-programming models for the kidney exchange problem. *European Journal of Operational Research*, 231(1):57–68, 2013.

[36] Eduardo de Oliveira Pacheco, Salvatore Cannella, Ricardo Lüders, and Ana Paula Barbosa-Povoa. Order-up-to-level policy update procedure for a supply chain subject to market demand uncertainty. *Computers & Industrial Engineering*, 113:347–355, 2017.

[37] Erick Delage and Yinyu Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations research*, 58(3):595–612, 2010.

[38] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[39] John P. Dickerson, David Manlove, Benjamin Plaut, Tuomas Sandholm, and James Trimble. Position-indexed formulations for kidney exchange. In *Proceedings of the ACM Conference on Economics and Computation (EC)*, 2016.

[40] John P. Dickerson, David Manlove, Benjamin Plaut, Tuomas Sandholm, and James Trimble. Position-indexed formulations for kidney exchange. *CoRR*, abs/1606.01623, 2016.

[41] John P. Dickerson, Ariel D. Procaccia, and Tuomas Sandholm. Failure-aware kidney exchange. *Management Science*, 2018. To appear; earlier version appeared at EC-13.

[42] John P. Dickerson and Tuomas Sandholm. Multi-organ exchange. *Journal of Artificial Intelligence Research*, 60:639–679, 2017.

[43] Wei Ding and Peter X-K Song. Em algorithm in gaussian copula with missing data. *Computational Statistics & Data Analysis*, 101:1–11, 2016.

[44] Fanghu Dong and Guosheng Yin. Maximum likelihood estimation for incomplete multinomial data via the weaver algorithm. *Statistics and Computing*, 28(5):1095–1117, 2018.

[45] Bradley Efron and David V Hinkley. Assessing the accuracy of the maximum likelihood estimator: Observed versus expected fisher information. *Biometrika*, 65(3):457–483, 1978.

[46] Craig K Enders and Deborah L Bandalos. The relative performance of full information maximum likelihood estimation for missing data in structural equation models. *Structural equation modeling*, 8(3):430–457, 2001.

[47] E Erdougan and Garud Iyengar. Ambiguous chance constrained problems and robust optimization. *Mathematical Programming*, 107(1-2):37–61, 2006.

[48] Haluk Ergin, Tayfun Sönmez, and M Utku Ünver. Multi-donor organ exchange, 2017. Working paper.

[49] Peyman Mohajerin Esfahani and Daniel Kuhn. Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. *Mathematical Programming*, 171(1-2):115–166, 2018.

[50] Vivek F Farias, Srikanth Jagabathula, and Devavrat Shah. A nonparametric approach to modeling choice with limited data. *Management science*, 59(2):305–322, 2013.

[51] Yonghan Feng and Sarah M Ryan. Scenario construction and reduction applied to stochastic power generation expansion planning. *Computers & Operations Research*, 40(1):9–23, 2013.

[52] Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability Theory and Related Fields*, 162(3-4):707–738, 2015.

[53] Rui Gao, Xi Chen, and Anton J Kleywegt. Wasserstein distributional robustness and regularization in statistical learning. *arXiv preprint arXiv:1712.06050*, 2017.

[54] Rui Gao and Anton J Kleywegt. Distributionally robust stochastic optimization with wasserstein distance. *arXiv preprint arXiv:1604.02199*, 2016.

[55] Pedro J García-Laencina, José-Luis Sancho-Gómez, and Aníbal R Figueiras-Vidal. Pattern classification with missing data: a review. *Neural Computing and Applications*, 19(2):263–282, 2010.

[56] Chloe Kim Glaeser, Marshall Fisher, and Xuanming Su. Optimal retail location: Empirical methodology and application to practice: Finalist–2017 m&som practice-based research competition. *Manufacturing & Service Operations Management*, 21(1):86–102, 2019.

[57] Kristiaan M. Glorie, J. Joris van de Klundert, and Albert P. M. Wagelmans. Kidney exchange with long chains: An efficient pricing algorithm for clearing barter exchanges with branch-and-price. *Manufacturing & Service Operations Management (MSOM)*, 16(4):498–512, 2014.

[58] Gregory A Godfrey and Warren B Powell. An adaptive, distribution-free algorithm for the newsvendor problem with censored demands, with applications to inventory and distribution. *Management Science*, 47(8):1101–1112, 2001.

[59] Noam Goldberg and Michael Poss. Maximum probabilistic all-or-nothing paths. *European Journal of Operational Research*, 2019.

[60] Donald Goldfarb and Garud Iyengar. Robust portfolio selection problems. *Mathematics of operations research*, 28(1):1–38, 2003.

[61] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[62] Stephen C Graves. A single-item inventory model for a nonstationary demand process. *Manufacturing & Service Operations Management*, 1(1):50–61, 1999.

[63] Vishal Gupta and Paat Rusmevichientong. Small-data, large-scale linear optimization with uncertain objectives. *Management Science*, 67(1):220–241, 2021.

[64] Grani A Hanasusanto and Daniel Kuhn. Conic programming reformulations of two-stage distributionally robust linear programs over wasserstein balls. *Operations Research*, 66(3):849–869, 2018.

[65] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[66] Woonghee Tim Huh, Retsef Levi, Paat Rusmevichientong, and James B Orlin. Adaptive data-driven inventory control with censored demand based on kaplan-meier estimator. *Operations Research*, 59(4):929–941, 2011.

[67] Woonghee Tim Huh and Paat Rusmevichientong. A nonparametric asymptotic analysis of inventory planning with censored demand. *Mathematics of Operations Research*, 34(1):103–123, 2009.

[68] Adel Javanmard and Hamid Nazerzadeh. Dynamic pricing in high-dimensions. *arXiv preprint arXiv:1609.07574*, 2016.

[69] Ran Ji and Miguel A Lejeune. Data-driven distributionally robust chance-constrained optimization with wasserstein metric. *Journal of Global Optimization*, 79(4):779–811, 2021.

[70] Ruiwei Jiang and Yongpei Guan. Data-driven chance constrained stochastic program. *Mathematical Programming*, 158(1-2):291–327, 2016.

[71] Ruiwei Jiang and Yongpei Guan. Risk-averse two-stage stochastic program with distributional ambiguity. *Operations Research*, 66(5):1390–1405, 2018.

[72] Anton J Kleywegt, Alexander Shapiro, and Tito Homem-de Mello. The sample average approximation method for stochastic discrete optimization. *SIAM Journal on Optimization*, 12(2):479–502, 2002.

[73] Xenia Klimentova, João Pedro Pedroso, and Ana Viana. Maximising expectation of the number of transplants in kidney exchange programmes. *Computers & Operations Research*, 73:1–11, 2016.

[74] Sumit Kunnumkal and Huseyin Topaloglu. Using stochastic approximation methods to compute optimal base-stock levels in inventory control problems. *Operations Research*, 56(3):646–664, 2008.

[75] Kamakshi Lakshminarayan, Steven A Harp, and Tariq Samad. Imputation of missing data in industrial databases. *Applied intelligence*, 11(3):259–275, 1999.

[76] Ruthanne Leishman. Challenges in match offer acceptance in the OPTN kidney paired donation pilot program. Presentation at the INFORMS Annual Meeting, 2019. Head of UNOS (US-wide kidney paired donation program).

[77] Retsef Levi, Georgia Perakis, and Joline Uichanco. The data-driven newsvendor problem: new bounds and insights. *Operations Research*, 63(6):1294–1306, 2015.

[78] Retsef Levi, Robin O Roundy, and David B Shmoys. Provably near-optimal sampling-based policies for stochastic inventory control models. *Mathematics of Operations Research*, 32(4):821–839, 2007.

[79] Céline Levy-leduc and Zaïd Harchaoui. Catching change-points with lasso. In *Advances in Neural Information Processing Systems*, pages 617–624, 2008.

[80] Bowen Li, Johanna L Mathieu, and Ruiwei Jiang. Distributionally robust chance constrained optimal power flow assuming log-concave distributions. In *2018 Power Systems Computation Conference (PSCC)*, pages 1–7. IEEE, 2018.

[81] Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information theory*, 37(1):145–151, 1991.

[82] Roderick JA Little and Donald B Rubin. On jointly estimating parameters and missing data by maximizing the complete-data likelihood. *The American Statistician*, 37(3):218–220, 1983.

[83] Roderick JA Little and Donald B Rubin. *Statistical analysis with missing data*, volume 793. John Wiley & Sons, 2019.

[84] Liwan H Liyanage and J G Shanthikumar. A practical inventory control policy using operational statistics. *Operations Research Letters*, 33(4):341–348, 2005.

[85] Thomas A Louis. Finding the observed information matrix when using the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 44(2):226–233, 1982.

[86] James Luedtke and Shabbir Ahmed. A sample approximation approach for optimization with probabilistic constraints. *SIAM Journal on Optimization*, 19(2):674–699, 2008.

[87] James Luedtke, Shabbir Ahmed, and George L Nemhauser. An integer programming approach for linear programs with probabilistic constraints. *Mathematical programming*, 122(2):247–272, 2010.

[88] David Manlove and Gregg O'Malley. Paired and altruistic kidney donation in the UK: Algorithms and experimentation. *ACM Journal of Experimental Algorithmics*, 19(1), 2015.

[89] Harry M Markowitz. Portfolio selection. *Wily, New York*, 1978.

[90] Stig-Arne Mattsson. Inventory control in environments with seasonal demand. *Operations Management Research*, 3(3-4):138–145, 2010.

[91] Duncan McElfresh, Hoda Bidkhori, and John P. Dickerson. Scalable robust kidney exchange. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2019.

[92] Duncan C McElfresh, Michael Curry, Tuomas Sandholm, and John P Dickerson. Improving policy-constrained kidney exchange via pre-screening. *arXiv preprint*, 2020.

[93] Walid W Nasr and Ibrahim J Elshar. Continuous inventory control with stochastic and non-stationary markovian demand. *European Journal of Operational Research*, 270(1):198–217, 2018.

[94] Arkadi Nemirovski, Anatoli Juditsky, Guanghui Lan, and Alexander Shapiro. Robust stochastic approximation approach to stochastic programming. *SIAM Journal on optimization*, 19(4):1574–1609, 2009.

[95] Whitney K Newey and Daniel McFadden. Large sample estimation and hypothesis testing. *Handbook of econometrics*, 4:2111–2245, 1994.

[96] Kancherla Jonah Nishanth and Vadlamani Ravi. Probabilistic neural network based categorical data imputation. *Neurocomputing*, 218:17–25, 2016.

[97] Art B Owen. *Empirical likelihood*. Chapman and Hall/CRC, 2001.

[98] Bernardo K Pagnoncelli, Shabbir Ahmed, and Alexander Shapiro. Sample average approximation method for chance constrained programming: theory and applications. *Journal of optimization theory and applications*, 142(2):399–416, 2009.

[99] Georg Pflug and David Wozabal. Ambiguity in portfolio selection. *Quantitative Finance*, 7(4):435–442, 2007.

[100] Constanta Radulescu. Mean-variance models with missing data. *Studies in Informatics and Control*, 22:299–206, 12 2013.

[101] F. T. Rapaport. The case for a living emotionally related international kidney donor exchange registry. *Transplantation Proceedings*, 18:5–9, 1986.

[102] Sandeep Rath, Kumar Rajaram, and Aman Mahajan. Integrated anesthesiologist and room scheduling for surgeries: Methodology and application. *Operations Research*, 65(6):1460–1478, 2017.

[103] R Tyrrell Rockafellar, Stanislav Uryasev, et al. Optimization of conditional value-at-risk. *Journal of risk*, 2:21–42, 2000.

[104] Cristian R Rojas and Bo Wahlberg. On change point detection using the fused lasso method. *arXiv preprint arXiv:1401.5408*, 2014.

[105] Alvin Roth, Tayfun Sönmez, and Utku Ünver. Kidney exchange. *Quarterly Journal of Economics*, 119(2):457–488, 2004.

[106] Alvin Roth, Tayfun Sönmez, and Utku Ünver. A kidney exchange clearinghouse in New England. *American Economic Review*, 95(2):376–380, 2005.

[107] Johannes O Royset and Roger J-B Wets. Variational theory for optimization under stochastic ambiguity. *SIAM Journal on Optimization*, 27(2):1118–1149, 2017.

[108] Donald B Rubin. Inference and missing data. *Biometrika*, 63(3):581–592, 1976.

[109] Cynthia Rudin and Gah-Yi Vahn. The big data newsvendor: Practical insights from machine learning. *Forthcoming in Operations Research*, 2018. https://ssrn.com/abstract=2559116.

[110] Napat Rujeerapaiboon, Kilian Schindler, Daniel Kuhn, and Wolfram Wiesemann. Scenario reduction revisited: Fundamental limits and guarantees. *Mathematical Programming*, pages 1–36, 2018.

[111] Herbert Scarf. The optimality of (S, s) policies in the dynamic inventory problem. *Proc. of the First Stanford Symposium on Mathematical Methods in the Social Science*, 1959.

[112] Suresh P Sethi and Feng Cheng. Optimality of (s, s) policies in inventory models with markovian demand. *Operations Research*, 45(6):931–939, 1997.

[113] Alexander Shapiro, Darinka Dentcheva, and Andrzej Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

[114] Alexander Shapiro and Tito Homem-de Mello. A simulation-based approach to two-stage stochastic programming with recourse. *Mathematical Programming*, 81(3):301–325, 1998.

[115] Alexander Shapiro and Arkadi Nemirovski. On complexity of stochastic programming problems. In *Continuous optimization*, pages 111–146. Springer, 2005.

[116] Zuo-Jun Max Shen, Collette Coullard, and Mark S Daskin. A joint location-inventory model. *Transportation science*, 37(1):40–55, 2003.

[117] C Shi, W Chen, and I Duenyas. Nonparametric data-driven algorithms for multi-product inventory systems. *Operations Research*, 64(2):362–370, 2016.

[118] Brian L Smith, William T Scherer, and James H Conklin. Exploring imputation techniques for missing data in transportation management systems. *Transportation Research Record*, 1836(1):132–142, 2003.

[119] James E Smith and Robert L Winkler. The optimizer's curse: Skepticism and post-decision surprise in decision analysis. *Management Science*, 52(3):311–322, 2006.

[120] Ralph D Snyder, Anne B Koehler, Rob J Hyndman, and J Keith Ord. Exponential smoothing models: Means and variances for lead-time demand. *European Journal of Operational Research*, 158(2):444–455, 2004.

[121] Jing-Sheng Song and Paul Zipkin. Inventory control in a fluctuating demand environment. *Operations Research*, 41(2):351–370, 1993.

[122] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2012.

[123] Jonathan AC Sterne, Ian R White, John B Carlin, Michael Spratt, Patrick Royston, Michael G Kenward, Angela M Wood, and James R Carpenter. Multiple imputation

for missing data in epidemiological and clinical research: potential and pitfalls. *Bmj*, 338:b2393, 2009.

[124] BOB Taylor. Developing portfolio optimization models. *The MathWorks News & Notes*, pages 30–32, 2006.

[125] Robert Tibshirani, Michael Saunders, Saharon Rosset, Ji Zhu, and Keith Knight. Sparsity and smoothness via the fused lasso. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(1):91–108, 2005.

[126] Olga Troyanskaya, Michael Cantor, Gavin Sherlock, Pat Brown, Trevor Hastie, Robert Tibshirani, David Botstein, and Russ B Altman. Missing value estimation methods for dna microarrays. *Bioinformatics*, 17(6):520–525, 2001.

[127] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[128] Jean-Philippe Vert and Kevin Bleakley. Fast detection of multiple change-points shared by many signals using group lars. In *Advances in neural information processing systems*, pages 2343–2351, 2010.

[129] Wen Wang, Mathieu Bray, Peter XK Song, and John D Kalbfleisch. An efficient algorithm to enumerate sets with fallbacks in a kidney paired donation program. *Operations Research for Health Care*, 20:45–55, 2019.

[130] Xian Wang, Ao Li, Zhaohui Jiang, and Huanqing Feng. Missing value estimation for dna microarray gene expression data by support vector regression imputation and orthogonal coding scheme. *BMC bioinformatics*, 7(1):32, 2006.

[131] Zizhuo Wang, Peter W Glynn, and Yinyu Ye. Likelihood robust optimization for data-driven problems. *Computational Management Science*, 13(2):241–261, 2016.

[132] Wolfram Wiesemann, Daniel Kuhn, and Melvyn Sim. Distributionally robust convex optimization. *Operations Research*, 62(6):1358–1376, 2014.

[133] Mengyuan Xiang, Roberto Rossi, Belen Martin-Barragan, and S Armagan Tarim. Computing non-stationary (s, s) policies using mixed integer linear programming. *European Journal of Operational Research*, 271(2):490–500, 2018.

[134] Weijun Xie. On distributionally robust chance constrained programs with wasserstein distance. *arXiv preprint arXiv:1806.07418*, 2018.

[135] Zhe Zhang, Shabbir Ahmed, and Guanghui Lan. Efficient algorithms for distributionally robust stochastic optimization with discrete scenario support. *arXiv preprint arXiv:1909.11216*, 2019.

[136] Chaoyue Zhao and Yongpei Guan. Data-driven risk-averse stochastic optimization with wasserstein metric. *Operations Research Letters*, 46(2):262–267, 2018.

[137] Shuaidong Zhao and Kuilin Zhang. A distributionally robust optimization approach to reconstructing missing locations and paths using high-frequency trajectory data. *Transportation Research Part C: Emerging Technologies*, 102:316–335, 2019.

[138] Qipeng P Zheng, Siqian Shen, and Yuhui Shi. Loss-constrained minimum cost flow under arc failure uncertainty with applications in risk-aware kidney exchange. *IIE Transactions*, 47(9):961–977, 2015.