Likelihood Robust Optimization for Data-driven Problems

Zizhuo Wang * Peter W. Glynn† Yinyu Ye‡

May 11, 2014

Abstract

We consider the optimal decision making problems in an uncertain environment. In particular, we consider the case in which the distribution of the input is unknown yet there is abundant historical data for it. In this paper, we propose a new type of distributionally robust optimization model which we call the likelihood robust optimization (LRO) for this class of problems. In contrast to most prior work on distributionally robust optimization which focus on certain parameters (e.g., mean, variance etc) of the input distribution, we exploit the historical data and define the accessible distribution set to contain only those distributions that make the observed data achieve a certain level of likelihood. Then we formulate the targeting problem as one of maximizing the expected value of the objective function under the worst-case distribution in that set. Our model can avoid the over-conservativeness of some prior robust approaches by ruling out unrealistic distributions while maintaining robustness of the solution for any statistically likely outcomes. We present detailed statistical analysis of our model using Bayesian statistics and empirical likelihood theory. Specifically, we prove the asymptotic behavior of our distribution set and establish the relationship between our model and other distributionally robust models. To test the performance of our model, we apply it to the newsvendor problem and the portfolio selection problem. The test results show that the solutions of our model indeed have desirable performance.

1 Introduction

The study of decision making problems under uncertain environment has been a main focus in the operations research community for decades. In such problems, one has a certain objective function to optimize, however, the objective function not only depends on the decision one makes, but also depends on some unknown parameters. Such situations are ubiquitous in practice. For example, in an inventory management problem, the inventory cost is influenced by both the inventory decisions and the random demands. Similarly, in a portfolio selection problem, the realized return is determined by both the choice of the portfolio and the random market fluctuations.

One solution method to such problems is to use the stochastic optimization. In this approach, one assumes the knowledge of the distribution of the unknown parameters and then chooses the
decision that optimizes the expected value of the objective function. If the knowledge of the
distribution is exact, then this approach is a precise characterization for a risk-neutral decision
maker. Much research has been done along this direction, we refer the readers to Shapiro et al [26]
for a comprehensive review on the topic of stochastic optimization.

However, there are several drawbacks in the stochastic optimization approach. First, although
stochastic optimization can frequently be formulated as a convex program, in order to solve it,
one often has to resort to Monte Carlo method which could be computational challenging. More
importantly, due to the limitation of knowledge, the distribution of the uncertain parameters is
rarely known in practice to a precise level. Even if enough data has been gathered in the past to
perform an accurate statistic analysis for the distribution, the analysis might be based on some
assumptions (e.g., independence of the observations, or stationary of the sequence) which are only
approximations of the reality. In addition, many decision makers in practice are not risk-neutral.
They tend to be risk-averse. A solution approach that can guard them from some bad scenarios is
of great practical interest.

Such an approach was proposed by Scarf [25] in a newsvendor context and has been studied
extensively in the past decade. It is called the distributionally robust optimization approach. In
the distributionally robust optimization approach, one considers a set of distributions of the uncer-
tain parameters and optimizes the worst-case expected value of the objective function among the
distributions in that set. Studies have been done by using different distribution sets. Most past
literature chooses the distribution set to be those with a fixed mean and variance. For example,
the earliest work by Scarf [25] shows that a closed-form solution can be obtained in the newsvendor
problem context when such a distribution set is chosen. This work is further extended by Gallego
and Moon [13]. The same form of the distribution set is also used in Calafiore and El Ghaoui [9],
Yue et al [31], Zhu et al [31] and Popescu [24] in which a linear-chance-constrained problem, a
minimax regret objective and a portfolio optimization problem are considered, respectively. Other
distribution sets beyond the mean and variance have also been proposed in the literature. For
example, Delage and Ye [12] propose a more general framework with distribution set formed by
moment constraints. A review of the recent developments can be found in Delage [11] and Shapiro
and Klegwegt [27].

Although the mean-variance distributionally robust optimization approach is intuitive and is
tractable under certain conditions, they are unsatisfactory from some aspects. First, when con-
structing the distribution set in such an approach, one only uses the moment information in the
sample data, while all the other information is ignored. This procedure may discard other impor-
tant information in the data set beyond the mean and variance. For example, a set of data drawn
from an exponential distribution with \( \lambda = 1/50 \) will have similar mean and variance as a set of data
drawn from a normal distribution with \( \mu = \sigma = 50 \). In the mean-variance distributionally robust
optimization, they will result in the same distribution set and the same decision will be chosen.
However, these two distributions have very different behaviors and the optimal decisions may be
quite different. Second, in certain cases, the worst-case distribution for any decision under the

---

\[1\] We note that there is also a vast literature on robust optimization where the worst-case parameter is chosen
for each decision made. However, the robust optimization is based on a slightly different philosophy than the
distributionally robust optimization and is usually more conservative. It can also be viewed as a special case of the
distributionally robust optimization where the distribution set only contains singleton distributions. In view of this,
we choose not to include a detailed discussions of this literature in the main text and refer the readers to Ben-Tal et
al [2] and Bertsimas et al [5] for comprehensive reviews.
mean-variance distributionally robust optimization framework is not a realistic one. For example, [25] proves that the worst-case distribution in the newsvendor context is a two-point distribution. In fact, this phenomenon exists for any problem when the objective function is concave in the unknown parameters (for a maximization problem). This raises a concern that whether the decision chosen by this approach is guarding some overly conservative scenarios (like the two-point distribution) which can not happen in practice, while performing poorly in more likely scenarios. Unfortunately, these two problems seem to be inherent in the model choice and can not be satisfactorily addressed.

In this paper, we propose another choice of the distribution set in the distributionally robust optimization framework that solves the above two drawbacks of the mean-variance approach. Instead of using the mean and variance to construct the distribution set, we choose to define it by the likelihood. More precisely, given a set of historical data, we define the distribution set to contain those distributions that make the observed data achieve a certain level of likelihood. We call this approach the likelihood robust optimization (LRO) and we study the properties of this approach and its performance in this paper.

First, we show that the likelihood robust optimization model is highly tractable. By applying the duality theory, we formulate the robust counterpart of this problem into a single convex optimization problem. In addition, we show that our model is very flexible. We can add any convex constraints (such as the moment constraints) to the distribution set while still maintaining its tractability. We use two concrete examples (a newsvendor problem and a portfolio selection problem) to illustrate the applicability of our framework.

Then we study the statistical theories behind the LRO approach by exploiting the linkage between our approach and the Bayesian statistic and empirical likelihood theory. We show that the distribution set in our approach can be viewed as a confidence region for the distributions given the set of observed data. Then we discuss how to choose the parameter in the distribution set to attain a specified confidence level, the discussion is based on the asymptotic behavior of the confidence region. Furthermore, we show a connection between the LRO approach and the previous mean-variance distributionally robust optimization approach. We show that one can also choose a set for the mean such that the distribution set has a certain confidence level under the probability measure defined by the empirical data. Our analysis shows that the LRO approach is fully data-driven, and takes advantages of the full strength of the available data while maintaining a certain level of robustness.

Finally, we test the performance of the LRO model in two problems, the newsvendor problem and the portfolio selection problem. In the newsvendor problem, we find that our approach produces similar results compared to the mean-variance distributionally robust approach when the underlying distribution is symmetric, while the solution of our approach is much better when the underlying distribution is asymmetric. In the portfolio selection problem, we show by using real historical data, our approach achieves decent returns. Furthermore, the LRO approach will naturally diversify the portfolio chosen, and produces returns with relatively low fluctuations.

In the meantime of an earlier version of this paper, Ben-Tal et al. [1] study a distributionally robust optimization model where the distribution set is defined by divergence measures. Their model contains the distribution set defined in our model as a special case. They also discussed solvability and statistical properties of their models. In this paper, we focus on the distribution set defined by the likelihood and explore further the connections to the empirical likelihood theories.

\footnote{This paper is based on the Ph.D. tutorial paper of the first author under the guidance of the second and third authors in 2009.}
In addition, we also address a scenario when the sample space is continuous, which adds insights to this class of approaches.

Two other papers that are related to this one are Iyanger [16] and Nilim and El Ghaoui [20]. In these two papers, the authors study the robust Markov Decision Process (MDP) problem in which the transition probabilities can be chosen from a certain set. They mention the likelihood set as one choice. However, they did not further explore the properties of this set nor did they try to extend it to general problems.

The remainder of this paper is organized as follows. In Section 2 we introduce our likelihood robust optimization framework and discuss some solution issues. We start with the problem with discrete state space. We then discuss the statistical properties of our approach in Section 3. In Section 4 we extend our discussions to problems with continuous state space. In Section 5, we present numerical tests of our model. And Section 6 concludes this paper.

2 Likelihood Robust Optimization Model

In this section, we formulate the likelihood robust optimization model, discuss some basic solution issues and then show its applicability to two concrete problems: the newsvendor problem and the portfolio selection problem. Throughout this and next sections, we only consider the case where the uncertainty parameters are in a discrete state space. We will extend our discussions to the continuous state space case in Section 4.

Consider an objective function $h(\mathbf{x}, \xi)$ to maximize \(^3\). $h(\mathbf{x}, \xi)$ depends on two parts, a decision variable $\mathbf{x}$ that taking values in a feasible set $D$, and a random variable $\xi$ that taking values in $\Xi = \{\xi_1, \xi_2, ..., \xi_n\}$. Assume we have observed $N$ independent samples of $\xi$, with $N_i$ occurrences of $\xi_i$. We define:

$$
\mathbb{D}(\gamma) = \left\{ \mathbf{p} = (p_1, p_2, ..., p_n) \Big| \sum_{i=1}^{n} N_i \log p_i \geq \gamma \right\}.
$$

We call $\mathbb{D}(\gamma)$ the likelihood robust distribution set with parameter $\gamma$. Note that $\mathbb{D}(\gamma)$ contains all the distributions with support in $\Xi$ such that the observed data achieves a likelihood of at least $\exp(\gamma)$. In this section, we treat $\gamma$ as a given constant. In Section 3, we are going to discuss how to choose $\gamma$ such that $\text{(1)}$ has a desirable statistical meaning. We formulate the likelihood robust optimization (LRO) problem as follows:

$$
\begin{align*}
\text{maximize}_{\mathbf{x} \in D} & \quad \min_{\mathbf{p}} \sum_{i=1}^{n} p_i h(\mathbf{x}, \xi_i) \\
\text{s.t.} & \quad \sum_{i=1}^{n} N_i \log p_i \geq \gamma \\
& \quad \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \quad \forall i = 1, ..., n.
\end{align*}
$$

In (2), we choose the decision variable $\mathbf{x}$, such that the expectation of the objective function under the worst-case distribution is maximized, where the worst-case distribution is chosen among the

\(^3\)All the discussions will be the same if one starts with a minimization problem.
distributions such that the observed data achieves a certain level of likelihood. Prior to this work, people have chosen other types of constraints of the inner problem in (2), e.g., moment constraints of the distribution. As we will show in Section 3, the likelihood constraints we define in the LRO model has a very intuitive statistical meaning.

To solve (2), we write down the Lagrangian of the inner optimization problem:

$$L(p, \lambda, \mu) = \sum_{i=1}^{n} p_i h(x, \xi_i) + \lambda \left( \gamma - \sum_{i=1}^{n} N_i \log p_i \right) + \mu \left( 1 - \sum_{i=1}^{n} p_i \right).$$

Therefore, the dual formulation of the inner problem is

$$\text{maximize}_{\lambda, \mu, y} \quad \mu + (\gamma + N - \sum_{i=1}^{n} N_i \log N_i) \cdot \lambda - N \lambda \log \lambda + \lambda \sum_{i=1}^{n} N_i \log y_i$$

s.t.

$$h(x, \xi_i) - \mu \geq y_i, \quad \forall i$$

$$\lambda \geq 0.$$

Since the inner problem of (2) is always convex, and by choosing $\gamma$ smaller than the log maximal likelihood $L^* = \sum_{i=1}^{n} N_i \log N_i / N$, there is always a strict interior feasible solution. Therefore, by Slater’s condition [8], (3) always have the same optimal value as the inner problem of (2). Then we combine (3) and the outer maximization problem together, we get that (2) is equivalent to the following:

$$\text{maximize}_{x, \lambda, \mu, y} \quad \mu + (\gamma + N - \sum_{i=1}^{n} N_i \log N_i) \cdot \lambda - N \lambda \log \lambda + \lambda \sum_{i=1}^{n} N_i \log y_i$$

s.t.

$$h(x, \xi_i) - \mu \geq y_i, \quad \forall i$$

$$x \in D, \quad \lambda \geq 0.$$

Thus we have:

**Proposition 1.** If $h(x, \xi)$ is concave in $x$, then the likelihood robust optimization problem described in (2) is a convex program and can be solved by solving (4).

Solving (4), we will get an optimal solution $(x^*, \lambda^*, \mu^*, y^*)$. To find out the worst-case distribution associated with the optimal solution, one can plug $x^*$ back into (2) and solve the inner minimization problem. However, we have the following simpler method following from the KKT conditions of this problem:

**Proposition 2.** If $(x^*, \lambda^*, \mu^*, y^*)$ is optimal for (4), then

$$p_i = \frac{\lambda^* N_i}{h(x^*, \xi_i) - \mu}$$

is the corresponding worst-case distribution.

One advantage of the LRO approach is that one can integrate the mean, variance and other information that is convex in $p$ into this model. Taking the linear case as example, the likelihood
robust optimization model can be generalized as follows:

\[
\begin{align*}
\text{maximize}_{x} \quad & \min_p \sum_{i=1}^{n} p_i h(x, \xi_i) \\
\text{s.t.} \quad & \sum_{i=1}^{n} N_i \log p_i \geq \gamma \\
& A p \geq b \\
& p_i \geq 0, \quad \forall i.
\end{align*}
\]

Note that the constraints \( A p \geq b \) can include any linear constraints on the moments of the distribution. Therefore, it could be very flexible. By applying the duality theory again, we can transform (5) into

\[
\begin{align*}
\text{maximize}_{x, \lambda, \mu, y} \quad & b^T \mu + (\gamma + N - \sum_{i=1}^{n} N_i \log N_i) \cdot \lambda - N \lambda \log \lambda + \lambda \sum_{i=1}^{n} N_i \log y_i \\
\text{s.t.} \quad & h(x, \xi_i) - a_i^T \mu \geq y_i, \quad \forall i \\
& x \in D, \mu \geq 0, \lambda \geq 0
\end{align*}
\]

which is again a convex program and readily solvable. The case of general convex constraints on \( p \) can be dealt with similarly.

In the following two subsections, we consider two applications of the LRO framework.

### 2.1 Application 1: Newsvendor Problem

In this subsection, we apply the LRO model to the newsvendor problem. In such problems, a newsvendor facing an uncertain demand has to decide how many newspapers to stock on a newsstand. If he orders too much, there will be a per unit overage cost for each copy that is left unsold; and if he orders too little, there will be a per unit underage cost for each unmet demand. The problem is to decide the optimal order quantity in order to minimize the expected cost (or equivalently, to maximize the expected profit).

The newsvendor problem is one of the most fundamental problem in inventory management and has been studied for more than a century. We refer the readers to Khouja \[18\] for a comprehensive review of this problem. In the classical newsvendor problem, one assumes that the distribution of the demand is known (with distribution \( F \)). The problem can then be formulated as:

\[
\min_x \quad G_f(x) = b\mathbb{E}_F(d - x)^+ + h\mathbb{E}_F(x - d)^+.
\]

In (6), \( x \) is the order quantity, \( d \) is the random demand with a probability distribution \( F \), and \( b, h > 0 \) are the per unit underage and overage costs.

It is well-known that a closed-form solution is available for this problem. However, such a solution relies on the accurate information about the demand distribution. In practice, one rarely possesses such information. To deal with the uncertainty, various robust approaches have been proposed. The earliest study of the robust newsvendor problem is by Scarf \[25\], who derives the optimal order quantity that minimizes the worst-case expected cost under the distributions with a fixed mean and variance. He further shows that for any given order quantity, the worst-case distribution is a two-point distribution. Further discussions of his results can be found in Gallego...
and Moon [13]. Although their results are neat in theory, there are two main defects of their model in practice. First, the mean and the variance of the demand are usually not exactly known. They are calibrated from the historical data. However, such calibrations usually come with error. And the impact of such errors are not captured by their models. The second problem is that this approach tends to be overly conservative. This can be seen from the two-point worst-case distribution in their model. In practice, the true distribution can not be a two-point one. The inclusion of such distributions (especially at optimal solution) raises concerns that whether the robust solution obtained is mainly driven by such unrealistic cases while performing poorly in realistic ones. In fact, as we will see in our numerical experiment, this may indeed be the case if the true underlying distribution is not symmetric.

In practice, what is often available to the decision maker is a set of the historical demand. And we should take advantage of it. Some research has been done along this direction, in which a pure data-driven model is proposed to solve the problem. However, pure data-driven approaches tend to be less robust, since it ignores the potential deviations from the data and does not guard against it.

In the following, we propose the LRO model for the newvendor problem. Our likelihood robust model takes an intermediate path between these two approaches. On one hand, we fully make use the historical demand data when constructing the distribution set. On the other hand, we allow possible deviations from the empirical data when considering our decisions. These two features make our approach both practical and robust.

Without loss of generality, we assume that the support of all possible demands is $S = \{1, ..., n\}$, i.e., the demand takes integer values between 1 and $n$. Suppose we are given some historical demand data. We use an $n$ dimensional vector $N = (N_1, ..., N_n)$ to denote the number of times that demand equals to $i$ in the observed historical data. The total number of data is $N = \sum_{i=1}^{n} N_i$.

The LRO model for the newvendor problem is given as follows:

$$\begin{align*}
\text{minimize}_x & \quad \max p \quad \sum_{i=1}^{n} p_i \left( b(d_i - x)^+ + h(x - d_i)^+ \right) \\
\text{subject to} & \quad \sum_{i=1}^{n} N_i \log p_i \geq \gamma \\
& \quad \sum_{i=1}^{n} p_i = 1, \quad p_i \geq 0, \forall i.
\end{align*}$$

Here the outer problem chooses an optimal ordering quantity, while the inner problem finds the worst-case distribution and cost with respect to that decision. Here the first constraint makes sure that the worst-case distribution is searched among those distributions that make the observed data achieves a chosen level of likelihood $\gamma$. By applying the same techniques as in Section 2 we can write (7) as a single convex optimization problem:

$$\begin{align*}
\text{minimize}_{x, \lambda, \mu, y} & \quad \mu + (\sum_{i=1}^{n} N_i \log N_i - \sum_{i=1}^{n} N_i - \gamma) \cdot \lambda + \sum_{i=1}^{n} (N_i \lambda \log \lambda - N_i \lambda \log(y_i)) \\
\text{subject to} & \quad b(d_i - x) + y_i \leq \mu, \forall i \\
& \quad h(x - d_i) + y_i \leq \mu, \forall i \\
& \quad \lambda \geq 0, \quad y_i \geq 0.
\end{align*}$$

This problem is easily solvable and numerical examples of the solution will be shown in Section 5.
Next we extend the LRO approach to the multi-item newsvendor problem. In classical approaches, the multi-item newsvendor problem wasn’t as well-studied as the single item problem due to the “curse of dimensionality”. In this problem, the decision variable is a $K$-dimensional vector $x = (x_1, x_2, \ldots, x_K)$, where $x_i$ corresponds to the order quantity of item $i$. Without loss of generality, we assume that the demand for each item varies from 1 to $n$, i.e, the demand vector $d$ lies in the set $\Xi = \{1, \ldots, n\}^K$. Thus, there are totally $n^K$ possible values in $\Xi$. If we assign each of them a probability, the problem size seems to become exponential. However, as will be seen soon, if we apply the LRO with a polynomial data set, the problem will still be polynomial in size.

In the following, we use $\sigma$ to denote a demand profile and $\delta_\sigma$ to denote the number of historical observations we have on $\sigma$.

A simple extension of model (7) gives us:

$$\min_{x \in D} \max_{\sigma} \sum_{\sigma} p_\sigma \sum_{i=1}^{K} \left[ b(d_\sigma(i) - x_i) + h(x_i - d_\sigma(i)) \right]$$

s.t. $\sum_{\sigma} \delta_\sigma \log p_\sigma \geq \gamma$ \hspace{1cm} (8)

$$\sum_{\sigma} p_\sigma = 1, \quad p_\sigma \geq 0, \forall \sigma.$$ 

In (8), $p_\sigma$ is the probability assigned to the demand profile $\sigma$, $d_\sigma(i)$ is the demand of the $i$th item in the profile $\sigma$. Now we take the dual of the inner program and get:

$$\min_{\lambda, \mu} \mu + \left( \sum_{\sigma} \delta_\sigma \log \delta_\sigma - \sum_{\sigma} \delta_\sigma - \gamma \right) \cdot \lambda + \sum_{\sigma} (\delta_\sigma \lambda \log \lambda - \delta_\sigma \lambda \log (\mu - c_\sigma))$$

s.t. $c_\sigma = \sum_{i=1}^{K} (b(d_\sigma(i) - x_i) + h(x_i - d_\sigma(i)))$

$$\mu \geq c_\sigma, \forall \sigma$$

$$\lambda \geq 0$$

Writing out $c_\sigma$ explicitly and combining the two ‘min’ signs together, we can rewrite (8) as follows:

$$\min_{x, \lambda, \mu, y, c} \mu + \left( \sum_{\sigma} \delta_\sigma \log \delta_\sigma - \sum_{\sigma} \delta_\sigma - \gamma \right) \cdot \lambda + \sum_{\sigma} (\delta_\sigma \lambda \log \lambda - \delta_\sigma \lambda \log (y_\sigma))$$

s.t. $\mu \geq y_\sigma + \sum_{i=1}^{K} c_{\sigma, i}, \forall \sigma$

$$c_{\sigma, i} \geq b(d_\sigma(i) - x_i), \forall \sigma, i = 1, 2, \ldots, K$$

$$c_{\sigma, i} \geq h(x_i - d_\sigma(i)), \forall \sigma, i = 1, 2, \ldots, K$$

$$\lambda \geq 0, \quad y \geq 0, \quad x \in D.$$ \hspace{1cm} (9)

It appears that there are exponentially many variables in this model. An important observation is that in the objective function only those $\sigma$’s with $\delta_\sigma > 0$ takes effect. We define this set to be $\Psi$. For $\sigma$’s that are not in $\Psi$, The corresponding terms are 0. Also, for the constraints $\mu \geq c_\sigma$, although it appears to have exponentially many, for each $x$, the constraint is equivalent to

$$\mu \geq \sum_{j=1}^{K} c_j$$

$$c_j \geq b(n - x_j), \quad j = 1, 2, \ldots, K$$

$$c_j \geq h(x_j - 1), \quad j = 1, 2, \ldots, K.$$
Therefore, (9) is equivalent to
\[
\min_{\mathbf{x}, \lambda, \mu, y, c} \mu + (\sum_{\sigma \in \Psi} \delta_{\sigma} \log \delta_{\sigma} - \sum_{\sigma \in \Psi} \delta_{\sigma} - \gamma) \cdot \lambda + \sum_{\sigma \in \Psi} (\delta_{\sigma} \lambda \log \lambda - \delta_{\sigma} \lambda \log(y_{\sigma}))
\]
subject to
- \( \mu \geq y_{\sigma} + \sum_{i=1}^{K} c_{\sigma,i}, \forall \sigma \in \Psi \)
- \( c_{\sigma,i} \geq b(d_{\sigma(i)} - x_{i}), \forall \sigma \in \Psi, i = 1, 2, ..., K \)
- \( c_{\sigma,i} \geq h(x_{i} - d_{\sigma(i)}), \forall \sigma \in \Psi, i = 1, 2, ..., K \)
- \( \mu \geq \sum_{i=1}^{K} c_{j} \)
- \( c_{j} \geq b(n - x_{j}), \quad j = 1, 2, ..., K \)
- \( c_{j} \geq h(x_{j} - 1), \quad j = 1, 2, ..., K \)
- \( \lambda \geq 0, y \geq 0, \mathbf{x} \in D. \)

Denote \(|\Psi| = N\). Then (9) is a convex program with at most \(O(NK)\) variables and constraints. Thus it is easily solvable for medium-sized problems.

### 2.2 Application 2: Portfolio Selection Problem

In this subsection, we apply the LRO model to the portfolio selection problem. In the portfolio selection problem, we observe \(N\) historical daily returns of \(d\) assets \(\{\xi_{i}\}_{i=1}^{n}\) where each \(\xi_{i}\) is a \(d\)-dimensional vector. We consider a certain support \(\Psi\) of all possible returns, the choice of \(\Psi\) can be derived by using statistical models to calculate the boundaries of the possible returns. The LRO model for this problem is formulated as follows:

\[
\max_{\mathbf{x} \in D} \min_{\mathbf{p}} \sum_{i=1}^{n} p_{i} \cdot \xi_{i}^{T} \mathbf{x} + \sum_{\sigma \in \Psi} p_{\sigma} \cdot \xi_{\sigma}^{T} \mathbf{x}
\]
subject to
- \( \sum_{i=1}^{n} \log p_{i} \geq \gamma \)
- \( e^{T} \mathbf{p} = 1, \mathbf{p} \geq 0. \)

In (10), the first term in the objective function corresponds to the scenarios that have been observed in the data, while the second term corresponds to the unobserved scenarios (it will be an infinite sum if \(\Psi\) has an infinite support). The first constraint is the likelihood constraint as we introduced before. For the ease of notation, we assume that each return profile occurs only once in the data, which is legitimate if we assume that the return is continuous distributed. \(\gamma\) is the robust level we need to choose, which we will discuss in more detail in the next section. The choice of \(\Psi\) can either be discrete or continuous. In fact, as long as we can solve
\[
g(\mathbf{x}) = \min_{\xi \in \Psi} \xi^{T} \mathbf{x}
\]
effectively for any \(\mathbf{x}\), the whole problem can be formulated as a convex program with polynomial size. To see that, we take the dual of the inner problem of (10) and combine with the outside problem. We have the entire problem can be written as

\[
\max_{\mathbf{x}, \mu, \lambda} \mu + \gamma \lambda + n \lambda - n \lambda \log \lambda + \lambda \sum_{i=1}^{n} \log(\xi_{i}^{T} \mathbf{x} - \mu)
\]
subject to
- \( \mu \leq \xi_{i}^{T} \mathbf{x}, \quad \forall \sigma \in \Psi \)
- \( \lambda \geq 0, \quad \mathbf{x} \in D. \)
It appears that the first set of constraints could be of exponential size, however, if $g(x)$ defined in (11) can be solved efficiently, then this is effectively one single constraint. For example, if $\Psi = \{(r_1, r_2, ..., r_n) | r_i \in [\bar{r}_i, \bar{r}_i]\}$, then the constraint is equivalent to $\mu \leq r^T x$. If $\Psi = \{(r_1, r_2, ..., r_n) ||r - r_0||_2 \leq \eta\}$, then the constraint is equivalent to $\mu \leq r_0^T x - \eta||x||$, which is also convex.

Again, we can add any convex constraints of $p$ into (10) such as constraints on the moments of the return. Thus our model is of great flexibility. Some numerical experiments of our model for this problem will be shown in Section 5.

3 Statistical Properties of the Likelihood Robust Optimization

In this section, we study the theories behind the likelihood robust optimization model. Specifically, we focus on the likelihood robust distribution set defined in (1). We will address the following questions in the discussion:

1. What are the statistical meanings of (1)?
2. How to select a meaningful $\gamma$?
3. How does LRO relate to other types of distributionally robust optimization model?

We answer the first question in Section 3.1 by using the Bayesian statistics and empirical likelihood theory to interpret the likelihood constraints. Those interpretations clarify the statistical motivations of the LRO model. Then we answer the second question in Section 3.2 in where we perform an asymptotic analysis of the likelihood region and point out an asymptotic optimal choice of $\gamma$. We study the last question in Section 3.3 where we present a relationship between our model and the traditional mean robust optimization model.

3.1 Bayesian statistics interpretation

Consider a random variable taking values in a finite set $S$. Without loss of generality, we assume $S = \{1, ..., n\}$. Assume the underlying probability distribution of the random variable is $p = \{p_1, ..., p_n\}$. We observe historical data $\Psi = \{N_1, ..., N_n\}$ in which $N_k$ represents the number of times the random variable takes value $k$ in the data. Then the maximum likelihood estimate (MLE) of $p$ is given by

$$\bar{p}_i = \frac{N_i}{N},$$

where $N = \sum_{i=1}^n N_i$ is the total number of observations and we define $L^* = \prod_{i=1}^n \left(\frac{N_i}{N}\right)^{N_i}$ to be the maximum likelihood value.

Now we examine the set of distribution $p$ such that the likelihood of $\Psi$ under $p$ exceeds a certain threshold. We use the concepts from Bayesian statistics [14]. Instead of thinking that the data is randomly drawn from the underlying distribution, we treat them as given and define a random vector $P = (p_1, ..., p_n)$ taking values on the $n-1$-dimensional simplex $\Delta = \{p | p_1 + ... + p_n = 1, p_i \geq 0\}$ with probability density function proportional to the likelihood

$$\prod_{i=1}^n p_i^{N_i}.$$
This distribution of \( p \) is known as the Dirichlet distribution. A Dirichlet distribution with parameter \( \Psi \) (denoted by \( \text{Dir}(\Psi) \)) has the density function as follows:

\[
f(p; \Psi) = \frac{1}{B(\Psi)} \prod_{i=1}^{n} p_i^{N_i-1}
\]

with

\[
B(\Psi) = \int_{p \in \Delta} \prod_{i=1}^{n} p_i^{N_i-1} dp_1...dp_n = \frac{\prod_{i=1}^{n} \Gamma(N_i)}{\Gamma(\sum_{i=1}^{n} N_i)}
\]

where \( \Gamma(\cdot) \) is the Gamma function. The Dirichlet distribution is used to estimate the hidden parameters of a discrete probability distribution given a collection of samples. Intuitively, if the prior of a distribution is represented as \( \text{Dir}(\alpha) \), then \( \text{Dir}(\alpha + \beta) \) is the posterior distribution following a sequence of observations with histogram \( \beta \). For a detailed discussion on the Dirichlet distribution, we refer the readers to [14].

In the LRO model, we assume a uniform prior \( \text{Dir}(e) \) on each point in the support of the data, where \( e \) is the unit vector. After observing the historical data \( \Psi \), the posterior follows \( \text{Dir}(\Psi + e) \). Note that this process can be adaptive as we observe new data.

Now we turn to examine the likelihood robust distribution set

\[
\left\{ p \left| \sum_{i=1}^{n} N_i \log p_i \geq \gamma \right. \right\}.
\]

This set represents a region in the probability distribution space. In the Bayesian statistics framework, we can compute the probability that \( p \) satisfies this constraint:

\[
P^* \left( \sum_{i=1}^{n} N_i \log p_i \geq \gamma \right) = \frac{1}{B(\Psi + e)} \int_{p \in \Delta} \prod_{i=1}^{n} p_i^{N_i-1} \cdot I \left( \sum_{i=1}^{n} N_i \log p_i \geq \gamma \right) dp_1...dp_n
\]

where \( P^* \) is the probability measure of \( p \) given the observed data and \( I(\cdot) \) is the indicator function. When deciding the likelihood robust distribution set to use, we want to choose \( \gamma^* \) such that

\[
P^* \left( \sum_{i=1}^{n} N_i \log p_i \geq \gamma^* \right) \geq 1 - \alpha
\]

for some pre-determined \( \alpha \). That is, we want to choose \( \gamma^* \) such that (12) is the \( 1 - \alpha \) confidence region of the probability parameters. And the LRO model is to choose the decision variable to maximize the worst-case objective where the worst-case distribution is chosen from the confidence region defined by the observed data.

However, in general, trying to find the exact \( \gamma^* \) that satisfies (13) is computationally challenging. In the next section, we study the asymptotic behavior of it, which will help use to find an approximate \( \gamma^* \) that satisfies (13).
3.2 Asymptotic behavior of the likelihood robust distribution set

In this section, we investigate the asymptotic behavior of the likelihood robust distribution set and give an explicit way to choose an appropriate \( \gamma \) in the LRO model. In this section, we assume that the true underlying distribution of the data is \( \bar{p} = \{\bar{p}_1, ..., \bar{p}_n\} \) with \( \bar{p}_i > 0 \). We observe \( N \) data drawn randomly from the underlying distribution with \( N_i \) observations on each outcome \( i \) (\( \sum_{i=1}^{n} N_i = N \)). We define \( \gamma_N \) to be the solution such that

\[
\mathbb{P}_N \left( \sum_{i=1}^{n} N_i \log p_i \geq \gamma_N \right) = 1 - \alpha.
\]

Here \( \alpha \) is a fixed constant and \( \mathbb{P}_N \) is the Dirichlet probability measure on \( p \) with parameters \( N_1, ..., N_n \). Clearly \( \gamma_N \) depends on \( N_i \) and thus is a random variable. We have the following asymptotic properties of \( \gamma_N \).

**Theorem 1.**

\[
\gamma_N - \sum_{i=1}^{n} N_i \log \frac{N_i}{N} \xrightarrow{p} -\frac{1}{2} \chi^2_{n-1,1-\alpha}
\]

where \( \chi^2_{d,1-\alpha} \) is the \( 1 - \alpha \) quantile of a \( \chi^2 \) distribution with \( d \) degrees of freedom.

The proof of theorem is referred to Theorem 3.1 in [22].

**Theorem 2.**

\[
\sqrt{N} \left( \frac{\gamma_N}{N} - \sum_{i=1}^{n} \bar{p}_i \log \bar{p}_i \right) \xrightarrow{d} X_0,
\]

where \( X_0 = \sum_{i=1}^{n} (1 + \log \bar{p}_i) X_i \) with \( (X_1, ..., X_n) \sim N(0, \Sigma) \) where

\[
\Sigma = \begin{pmatrix}
p_1(1 - p_1) & -p_1 p_2 & \cdots & -p_1 p_n \\
-p_1 p_2 & p_2(1 - p_2) & \cdots & -p_2 p_n \\
\vdots & \vdots & \ddots & \vdots \\
-p_1 p_n & -p_2 p_n & \cdots & p_n(1 - p_n)
\end{pmatrix}.
\]

**Proof.** Given Theorem [1] it suffices to prove that:

\[
\sqrt{N} \left( \sum_{i=1}^{n} \frac{N_i}{N} \log \frac{N_i}{N} - \sum_{i=1}^{n} \bar{p}_i \log \bar{p}_i \right) \xrightarrow{d} X_0.
\]

To show this, we note that for fixed \( N \), \( (N_1, ..., N_n) \) follows a multinomial distribution with parameters \( \bar{p}_1, ..., \bar{p}_n \). By Theorem 14.6 in [29], we have

\[
\sqrt{N} \left( \frac{N_1}{N} - \bar{p}_1, ..., \frac{N_n}{N} - \bar{p}_n \right) \xrightarrow{d} N(0, \Sigma).
\]
Define \( f(x) = x \log x \). Therefore,
\[
\sqrt{N} \left( \sum_{i=1}^{n} N_i \log \frac{N_i}{N} - \sum_{i=1}^{n} \bar{p}_i \log \bar{p}_i \right) = \sum_{i=1}^{n} f'(\eta_i) \sqrt{N} \left( \frac{N_i}{N} - \bar{p}_i \right)
\]
where \( \eta_i \) is a number between \( N_i/N \) and \( \bar{p}_i \). By the strong law of large numbers, \( \eta_i \to \bar{p}_i \) almost surely, therefore
\[
\sum_{i=1}^{n} f'(\eta_i) \sqrt{N} \left( \frac{N_i}{N} - \bar{p}_i \right) \Rightarrow_d X_0.
\]
and thus the theorem is proved. \( \square \)

Theorem 1 provides a heuristic guideline on how to choose the threshold \( \gamma \) in the LRO model. In particular, one should choose approximately
\[
\gamma^* = \sum_{i=1}^{n} N_i \log \frac{N_i}{N} - \frac{1}{2} \chi^2_{n-1,1-\alpha}
\]
in order for the likelihood robust distribution set to have a confidence level of \( 1 - \alpha \). Note that the difference between \( \gamma^* \) and \( \sum_{i=1}^{n} N_i \log (N_i/N) \) converges as \( N \) grows. This means that when the data set is large, the allowable distributions must be very close to the empirical data. This is unlike some other distributionally robust optimization approaches that construct distribution set on the mean and/or the variance. Even with the data size increases, the distribution set in those approaches may still contain distributions that are far from the empirical distribution - like a two-point distribution on the extremes of an interval. Therefore our approach does a better job in utilizing the historical data to construct the robust region.

Theorem 2 points out the behavior of \( \gamma \) with respect to the true underlying parameters. It shows that our approach is indeed statistically “consistent”.

### 3.3 Relation to other types of distributionally robust optimization

In this section, we show how one can relate other types of distributionally robust optimization model to our LRO model through techniques in the empirical likelihood theory. Given observed data \( X_1, \ldots, X_n \) with the empirical distribution:
\[
F_N(t) = \frac{1}{N} \sum_{i=1}^{N} I(X_i \leq t),
\]
we define the likelihood ratio of any distribution \( F \) by
\[
R(F) = \frac{L(F)}{L(F_N)}.
\]
Now suppose we are interested in a certain parameter (or certain parameters) \( \theta = T(F) \) of the distribution. we can define the profile likelihood ratio function \( [21] \) as follows:
\[
R_T(\theta) = \sup_{F} \{ R(F) | T(F) = \theta \}. \tag{14}
\]
Here common choices of $T$ include the moments or quantiles of a distribution. In distributionally robust optimization, one selects a region $\Theta$ and maximizes the worst-case objective value for $\theta \in \Theta$. Using the profile likelihood ratio function in [14], we can define $\Theta$ to be of the form

$$\Theta = \{ \theta | R_T(\theta) \geq r_0 \}.$$ 

When $\theta$ is the mean of the distribution, the next theorem helps to determine $r_0$ such that $\Theta$ is a certain confidence interval for the true mean.

**Theorem 3.** [Theorem 2.2 in [21]] Let $X_1, \ldots, X_n$ be independent random variables with common distribution $F_0$. Let $\mu_0 = \mathbb{E}[X_1]$, and suppose that $0 < \text{Var}(X_1) < \infty$. Then $-2\log(R(\mu_0))$ converges in distribution to $\chi^2(1)$ as $n \to \infty$.

Therefore, in order for the set $\{ \theta | R(\theta) \geq r_0 \}$ to achieve $1 - \alpha$ confidence level (under the Dirichlet distribution induced by the observed data), one should approximately choose the boundary $\theta$'s such that

$$R(\theta) = e^{-\frac{1}{2} \chi^2_{1,1-\alpha}}.$$ 

To compute $R(\theta)$ for each $\theta$, we solve the following optimization problem:

$$z = \max \sum_i N_i \log p_i \quad \text{s.t.} \quad \sum_i p_i = 1 \quad \sum_i d_i p_i = \theta \quad p \geq 0$$

and let $R(\theta) = e^{z/L^*}$. Then for each threshold $\alpha$, we can find the corresponding boundaries of $\theta$ by implementing bisections on $\theta$. By this way, we can construct a mean robust optimization whose uncertainty set has a certain confidence level under the Dirichlet distribution defined by the observed data.

4 Continuous State Space Case

In the previous sections, we assume that the probability space is discrete. In the following, we extend our discussions to continuous state space case.

It is tempting to directly extend the previous definition of the likelihood region to the continuous case by using the probability density function rather than the probability mass function. However, with finite historical data, the corresponding constraint will be defined only on finitely many points of the probability density functions which effectively does not make any restrictions to the distribution at all. Thus we have to take a different approach. In this section, we propose an approach that defines the robust distribution set by constructing a band on the cumulative distribution function (CDF). We show that such an approach results in a tractable robust counterpart and with proper choice of the band, the formulation is statistically meaningful.
Given a set of observations $X_1,...,X_n$ drawn i.i.d. from an underlying distribution $X$. A band on the CDF with support $\{X_i\}$ and lower and upper bounds $\{L_i\}_{i=1}^n$ and $\{U_i\}_{i=1}^n$ is defined by

$$\{F(\cdot)| L_i \leq F(X_i) \leq U_i, i = 1, ..., n\}.$$

Now we briefly discuss one example of such bands that are statistical meaningful. We define the Kolmogorov-Smirnov bands as:

$$\left\{F(\cdot) \bigg| \frac{i}{n} - D_{n,1-\alpha} \leq F(X_i) \leq \frac{i}{n} + D_{n,1-\alpha}, i = 1, ..., n \right\}.$$ 

This bands come from the Kolmogorov-Smirnov goodness-of-fit test with the statistics

$$D_n = \sup_x \left| \tilde{F}(x) - F(x) \right|$$

where $\tilde{F}$ is the empirical distribution of $X_1,...,X_n$. By choosing $D_{n,1-\alpha}$ such that $P(D_n \leq D_{n,1-\alpha}) \leq 1 - \alpha$, the band covers the true CDF $1 - \alpha$ fraction of times. This method can be modified to construct a weighted Kolmogorov-Smirnov band where a different set of $D_{n,1-\alpha}$’s are used at each point based on different $\tilde{F}(x)$ values. We refer the readers to [19] for related discussions.

Once we have obtained such bands, we can write the corresponding robust program as:

$$\max_{\mathbf{x} \in D} \min_F \mathbb{E}_F[h(\mathbf{x},\xi)] \quad \text{(15)}$$

s.t. $L_i \leq F(X_i) \leq U_i, \quad i = 1, ..., n.$

$F(\cdot)$ is a CDF

By writing the CDF as the integral of the probability density function (PDF), we can write (15) as a semi-infinite programming:

$$\max_{\mathbf{x} \in D} \min_f \int_{-\infty}^{\infty} h(\mathbf{x},\xi)f(\xi) d\xi$$

s.t. $L_i \leq \int_{-\infty}^{X_i} f(\xi) d\xi \leq U_i, \quad i = 1, ..., n$

$$\int_{-\infty}^{\infty} f(\xi) d\xi = 1, f(\xi) \geq 0.$$ 

By using the duality theorem, we can write the dual of the inner program as:

$$\max_{y,z,\lambda} \sum_{i=1}^n y_i U_i - \sum_{i=1}^n z_i L_i + \lambda$$

s.t. $h(\mathbf{x},\xi) - \sum_{i=1}^k y_i + \sum_{i=1}^k z_i - \lambda \geq 0, \quad \forall \xi \in [X_{k-1},X_k], \quad k = 1, ..., n + 1 \quad (16)$

where we define $X_0 = -\infty$ and $X_{n+1} = \infty$. If $h(\mathbf{x},\xi)$ is convex in $\xi$, then the constraints in (16) can be reduced to:

$$h(\mathbf{x},X_{k-1}) \geq \sum_{i=1}^k y_i - \sum_{i=1}^k z_i + \lambda, \quad \forall k$$

$$h(\mathbf{x},X_k) \geq \sum_{i=1}^k y_i - \sum_{i=1}^k z_i + \lambda, \quad \forall k.$$
Then combined with the outer problem, we obtain the robust counterpart of (15):

$$\max \sum_{i=1}^{n} y_i U_i - \sum_{i=1}^{n} z_i L_i + \lambda$$

subject to

$$h(x, X_{k-1}) \geq \sum_{i=1}^{k} y_i - \sum_{i=1}^{k} z_i + \lambda, \quad \forall k$$

$$h(x, X_k) \geq \sum_{i=1}^{k} y_i - \sum_{i=1}^{k} z_i + \lambda, \quad \forall k$$

$$y_i \geq 0, z_i \geq 0, \lambda \geq 0$$

$$x \in D$$

(17)

If $h(x, \xi)$ is concave in $x$, then (17) will be a convex program with finite variables and constraints, and thus can be solved easily.

5 Numerical Results

In this section, we perform numerical tests using our LRO model. The tests are performed on both the newsvendor problem (see Section 5.1) and the portfolio selection problem (see Section 5.2).

5.1 Newsvendor problem

In this section, we apply our LRO model to the newsvendor problem introduced in Section 2.1. We show the features of the solutions of the LRO model and compare them to the solutions of other methods, in particular, the distributionally robust approach with fixed mean and variance. In the following, we consider a newsvendor problem with unit production cost of $5, unit selling price of $6, and unit salvage value of $4. After proper transformation, the problem is equivalent to problem (6) with underage and overage costs $b = h = 1$. We consider two underlying demand distributions. The first one is a normal distribution with mean $\mu = 50$ and standard deviation $\sigma = 50$. The second one is an exponential distribution with $\lambda = 1/50$. We assume that both demand distributions are truncated at 0 and 200.

For each underlying distribution, we perform the following procedures for the LRO approach:

1. Generate 1000 historical data drawn from the underlying distribution, and record the number of times the demand equals to $i$ by $N_i$.

2. Construct the likelihood robust distribution set $D(\gamma) = \{ p \mid \sum_{i=1}^{n} N_i \log p_i \geq \gamma \}$. To choose a proper $\gamma$, we use our asymptotic result in Theorem 1. We choose $\gamma$ such that $D(\gamma)$ covers the true distribution with probability 0.95, where the probability is under the Dirichlet measure with parameters $N_0 + 1, \ldots, N_n + 1$. By Theorem 1, $\gamma$ should be approximately $ML - \chi^2_{100,0.95}$, where $ML = \sum_{i=1}^{n} N_i \log \frac{N_i}{N}$.

3. Solve the LRO model (7) with the $\gamma$ chosen in Step 2.

Using the same group of sample data, we also test the following approaches:

1. LRO with fixed mean and variance: Denote the sample mean by $\hat{\mu}$ and the sample variance by $\hat{\sigma}^2$. We add two constraints $\sum_{i=1}^{n} p_i d_i = \hat{\mu}$ and $\sum_{i=1}^{n} p_i d_i^2 = \hat{\mu}^2 + \hat{\sigma}^2$ to the inner problem of (7), that is, we only allow those distributions that have the same mean and variance as the sample mean and variance. As we discussed in (5), this would still be a convex problem and one can obtain the optimal solution easily. We denote this approach by LRO$(\hat{\mu}, \hat{\sigma}^2)$. 

16
2. The distributionally robust approach with fixed mean and variance only: This is the approach proposed by Scarf [25]. In this approach, one minimizes the worst-case expected cost where the worst-case distribution is chosen from all the distributions with a fixed mean (equal to the sample mean) and a fixed variance (equal to the sample variance). We call this approach the Scarf approach. In [25], it is shown that the optimal decision for the Scarf approach can be expressed in a closed-form:

\[ q_{\text{scarf}}^* = \hat{\mu} + \frac{\hat{\sigma}}{2} \left( \sqrt{\frac{b}{h}} - \sqrt{\frac{h}{b}} \right). \]

3. The empirical distribution approach: We solve the optimal solution using the empirical distribution as the true distribution.

4. We solve the optimal solution using the true underlying distribution.

|                | Solution (q*) | \(E_{F_{true}}[h(q^*, \xi)]\) | \(h_{\text{LRO}}(q^*)\) |
|----------------|---------------|-------------------------------|---------------------------|
| LRO            | 56            | 35.97 (+0.84%)                | 49.32 (+0.00%)            |
| LRO(\(\hat{\mu}, \hat{\sigma}^2\)) | 54            | 35.80 (+0.37%)                | 49.47 (+0.30%)            |
| Scarf          | 53            | 35.74 (+0.20%)                | 49.63 (+0.63%)            |
| Empirical      | 50            | 35.79 (+0.34%)                | 50.03 (+1.40%)            |
| Underlying     | 51            | 35.67 (+0.00%)                | 49.87 (+1.11%)            |

Table 1: Results when the underlying distribution is normal

|                | Solution (q*) | \(E_{F_{true}}[h(q^*, \xi)]\) | \(h_{\text{LRO}}(q^*)\) |
|----------------|---------------|-------------------------------|---------------------------|
| LRO            | 46            | 35.72 (+3.51%)                | 53.71 (+0.00%)            |
| LRO(\(\hat{\mu}, \hat{\sigma}^2\)) | 41            | 34.99 (+1.39%)                | 54.45 (+1.38%)            |
| Scarf          | 50            | 36.64 (+6.17%)                | 54.27 (+1.04%)            |
| Empirical      | 37            | 34.55 (+0.16%)                | 55.75 (+3.80%)            |
| Underlying     | 35            | 34.51 (+0.00%)                | 56.20 (+4.64%)            |

Table 2: Results when the underlying distribution is exponential

Our test results are shown in Table 1 and 2. In Table 1 and 2, the second column is the optimal decision computed under each model. The third column shows the expected cost of each decision under the true underlying distribution, which is truncated \(N(50, 2500)\) in Table 1 and truncated \(\text{Exp}(1/50)\) in Table 2. The last column shows the objective value of each decision under the LRO model, that is the worst-case expected objective value when the distribution could be chosen in \(\mathbb{D}(\gamma)\). The numbers in the parenthesis in both tables are the differences between the current solution and the optimal solution under the measure specified in the corresponding column.

In Table 1, one can see that when the underlying demand distribution is a normal distribution, the solution of each model doesn’t differ much from each other. This is mainly because the symmetric property of the normal distribution, which pushes the solution to the middle in either
model. Nevertheless, the worst-case distributions in each case are quite different. We plot them in Figure 1.

In Figure 1, we see that the worst-case distribution in the Scarf’s model is a two-point distribution (with positive mass at 3 and 103). However, a two-point distribution is not a realistic one, which means that the Scarf’s approach might be guarding some over-conservative scenarios. In contrast, the worst-case distributions when we use LRO and LRO($\hat{\mu}, \hat{\sigma}^2$) are much closer to the empirical data and look much more plausible. In particular, the LRO with mean and variance constraint results in a worst-case distribution closest to the empirical data among these three.

The situation is a little bit different in the second case when the underlying distribution is an exponential distribution. In this case, the Scarf solution is significantly worse than the LRO solutions. This is because it doesn’t use the information that the data is skewed. In fact, it still only takes the mean and variance as the input. In contrast, the LRO and LRO($\hat{\mu}, \hat{\sigma}^2$) adapt to the asymmetry of the data and only considers the distributions that are close to the empirical distribution.

In both cases, we observe that the LRO($\hat{\mu}, \hat{\sigma}^2$) seems to work better. Indeed, we find that adding some constraints on $p$ regarding the mean/variance of the distribution usually helps the performance. This is because that it helps to further concentrating the set of the distributions that we take into consideration to those have similar shape as the empirical distribution. Lastly, we find the empirical distribution works well if we use the true distribution to evaluate the performance. However, it is not robust to potential deviations to the underlying distribution. As shown in the last column in Table 1 and 2, when we allow the underlying distribution to change a little bit (within the 95% confidence range), the performance of the solution obtained by using the empirical distribution might be away from the optimal.

5.2 Portfolio selection problem

In this section, we apply the LRO model to the portfolio selection problem introduced in Section 2.2. The experimental setup is as follows:

- We gather the historical data of 30 assets from the SP500 index during the time period from 2001 to 2004. In each experiment, we choose 4 assets to focus on and the decision is to
construct a portfolio using these 4 assets for each day during this period. And we use the past 30 days data as the observed data in the LRO approach.

As shown in Section 2.2, the LRO model for the portfolio selection problem can be written as:

\[
\begin{align*}
\text{maximize} & \quad x, \mu, \lambda, \mu + \gamma \lambda + n \lambda - n \lambda \log \lambda + \lambda \sum_{i=1}^{n} \log(\xi_i^T x - \mu) \\
\text{s.t.} & \quad \mu \leq \xi_\sigma^T x, \quad \forall \sigma \in \Psi \\
& \quad \lambda \geq 0, \quad x \in D.
\end{align*}
\]

Here \(\Psi\) is the support of the return vectors and \(\gamma\) is the likelihood level we have to choose. In our experiments, we choose \(\Psi = [\ell_1, \bar{\ell}_1] \times \ldots \times [\ell_k, \bar{\ell}_k]\), where \(\ell_i\) and \(\bar{\ell}_i\) are the lowest and highest returns of asset \(i\) in the historical data respectively. As discussed in Section 2.2, although there seems to be infinitely many constraints in (18), it can be represented by polynomial many of them. For \(\gamma\), we still use the asymptotic result (Theorem 1) to choose it. In particular, we choose the degree of freedom \(d\) equals to the number of historical observations. Although this is a heuristic choice, we find it works well in the numerical test.

To test our result, we compare our approach to a naive approach in which at each day, the stock with the highest past 30 days return is chosen to be the sole stock in the portfolio (we call it SS approach, which stands for Single Stock). We acknowledge that there are many other systematic ways that one can choose the portfolio and our study doesn’t mean to show that our approach is the best way to choose portfolio (for the portfolio selection problem, it is hard to argue that one approach is better than another one, without intensive and extensive tests, which is beyond the scope of this paper). Instead, we aim to illustrate that the LRO approach gives a decent performance with some desired features.

We first present some overall results. We did 100 experiments (each with 4 different stocks randomly chosen from the 30 to form the portfolio for each day) and the tested period covers 721 days. Among the 100 experiments, LRO performs better than the SS approach 63 times. And the average gain over the SS approach is about 2%. Again, we comment that this is not to say that our approach must be a better way to choose portfolio, but it does mean that the performance is quite decent.

Now we study the features of the solution given by the LRO model. First we examine how many stocks are chosen to construct the portfolio in the LRO model. We find that 52% of the time, it chooses a single stock to form the portfolio. And it chooses two stocks, three stocks and all the four stocks in its solution for 39%, 8% and 1% of the times. On the contrary, the SS approach always chooses a single stock at each time period. Therefore, we find that LRO approach implicitly achieves diversification, which is a feature that is usually desired in such problems.

We also look at the fluctuations of the returns of both approaches. We plot the daily returns of LRO and SS of one experiment in Figure 2. We can see that the LRO approach has much less variations than the SS approach (in the data shown in Figure 2, the standard deviation of the returns of LRO is 0.023, while the standard deviation of the returns of SS is 0.033). Indeed, in all the experiments, the standard deviations of the LRO approach is smaller than that of the SS approach. Less fluctuations combined with higher overall return proves that LRO is a decent model for the portfolio selection problem.
6 Conclusion

In this paper, we propose a new distributionally robust framework, the likelihood robust optimization model. The model maximizes the worst-case expected value of a certain objective function where the worst-case distribution is chosen from the set of distributions that make the observed data achieve a certain level of likelihood. The proposed model is easily solvable and has strong statistic meanings. It avoids the over-conservatism of some other robust models while protecting the decisions from reasonable deviations from the empirical data. We discuss two applications of our model. The numerical results show that our model might be appealing in applications.

7 Acknowledgement

The authors thank Erick Delage, Dongdong Ge, and Zhisu Zhu for valuable insights and discussions.

References

[1] A. Ben-Tal, D. den Hertog, A. De Waegenaere, B. Melenberg, and G. Rennen. Robust solutions of optimization problems affected by uncertain probabilities. *Management Science*, 59(2):341–357, 2013.

[2] A. Ben-Tal, L. El Ghaoui, and A. Nemirovski. *Robust Optimization*. Princeton Series in Applied Mathematics, 2009.

[3] A. Ben-Tal and A. Nemirovski. Robust solutions of uncertain linear programs. *Operations Research Letters*, 25(1):1–13, 1999.

[4] D. Bertsimas and A. Thiele. Robust and data-driven optimization: Modern decision-making under uncertainty. In *Tutorial on Operations Research*. INFORMS, 2006.
[5] D. Bertsimas, D. Brown, and C. Caramanis. Theory and applications of robust optimization. *SIAM Review*, 53(3):464–501.

[6] D. Bertsimas and M. Sim. The price of robustness. *Operations Research*, 52(1):35–53, 2004.

[7] D. Bertsimas and A. Thiele. A robust optimization approach to inventory theory. *Operations Research*, 54(1):150–168, 2006.

[8] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.

[9] G. Calafiore and L. El Ghaoui. On distributionally robust chance-constrained linear programs. *Optimization Theory and Applications*, 130(1):1–22, 2006.

[10] X. Chen, M. Sim, and P. Sun. A robust optimization perspective on stochastic programming. *Operations Research*, 55(6):1058–1071, 2007.

[11] E. Delage. *Distributionally Robust Optimization in context of Data-Driven Problem*. PhD thesis, 2009.

[12] E. Delage and Y. Ye. Distributionally robust optimization under moment uncertainty with application to data-driven problems. *Operations Research*, 58(3):595–612, 2008.

[13] G. Gallego and I. Moon. The distribution free newsboy problem: Review and extension. *The Journal of the Operational Research Society*, 44(8):825–834, 1993.

[14] A. Gelman, J. Carlin, H. S. Stern, and D. B. Rubin. *Bayesian Data Analysis*. Chapman Hall Press, 1995.

[15] H. Scarf. Bayes solutions of the statistical inventory problem. *Annal of Mathematical Statistics*, 30(2):490–508, 1959.

[16] G. Iyengar. Robust dynamic programming. *Mathematics of Operations Research*, 30(2):257–280, 2005.

[17] N. Johnson, S. Kotz, and N. Balakrishnan. *Continuous Univariate Distributions, Vol. 1*. Wiley Series in Probability and Statistics, 1994.

[18] M. Khouja. The single period newsvendor problem: Literature review and suggestions for future research. *Omega*, 27:537–553, 1999.

[19] D. Mason and J. Schuenemyer. A modified Kolmogorov-Smirnov test sensitive to tail alternatives. *Annals of Statistics*, 11(3):933–946, 1983.

[20] A. Nilim and L. El Ghaoui. Robust control of markov decision processes with uncertain transition matrices. *Operations Research*, 53(5):780–798, 2005.

[21] A. Owen. *Empirical Likelihood*. Chapman Hall Press, 2001.

[22] L. Pardo. *Statistical Inference Based on Divergence Measures*. Chapman Hall Press, 2005.

[23] G. Perakis and G. Roels. Regret in the newsvendor model with partial information. *Operations Research*, 56(1):188–203, 2008.
[24] I. Popescu. Robust mean-covariance solutions for stochastic optimization. *Operations Research*, 55(1):98–112, 2007.

[25] H. Scarf. A min-max solution of an inventory problem. *Studies in The Mathematical Theory of Inventory and Production*, pages 201–209, 1958.

[26] A. Shapiro, D. Dentcheva, and A. Ruszczynski. *Lectures on Stochastic Programming: Modeling and Theory*. MPS-SIAM Series on Optimization, 2009.

[27] A. Shapiro and A. J. Kleywegt. Minimax analysis of stochastic programs. *Optimization Methods and Software*, 17:523–542, 2002.

[28] A. M-S. So, J. Zhang, and Y. Ye. Stochastic combinatorial optimization with controllable risk aversion level. *Mathematics of Operations Research*, 34(3):522–537, 2009.

[29] L. Wassaman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics, 2009.

[30] J. Yue, B. Chen, and M. Wang. Expected value of distribution information for the newsvendor problem. *Operations Research*, 54(6):1128–1136, 2006.

[31] Z. Zhu, J. Zhang, and Y. Ye. Newsvendor optimization with limited distribution information. Technical report, 2006.