OPERATIONS RESEARCH
Vol. 00, No. 0, Xxxxx 0000, pp. 000–000
ISSN 0030-364X | EISSN 1526-5463 | 00 | 0000 | 0001

INFORMS
DOI 10.1287/xxxx.0000.0000

Distributionally Robust Optimization under Moment Uncertainty with Application to Data-Driven Problems

Erick Delage
Department of Electrical Engineering, Stanford University, Stanford, California, USA
edelage@stanford.edu, http://www.stanford.edu/∼edelage

Yinyu Ye
Department of Management Science and Engineering, Stanford University, Stanford, California, USA
yinyu-ye@stanford.edu, http://www.stanford.edu/∼yyye

Stochastic programs can effectively describe the decision-making problem in an uncertain environment. Unfortunately, such programs are often computationally demanding to solve. In addition, their solutions can be misleading when there is ambiguity in the choice of a distribution for the random parameters. In this paper, we propose a model describing one's uncertainty in both the distribution's form (discrete, Gaussian, exponential, etc.) and moments (mean and covariance). We demonstrate that for a wide range of cost functions the associated distributionally robust stochastic program can be solved efficiently. Furthermore, by deriving new confidence regions for the mean and covariance of a random vector, we provide probabilistic arguments for using our model in problems that rely heavily on historical data. This is confirmed in a practical example of portfolio selection, where our framework leads to better performing policies on the "true" distribution underlying the daily return of assets.

Subject classifications: Programming: stochastic, Statistics: estimation, Finance: portfolio.
Area of review: Optimization.
History: Draft created February 20, 2008.

1. Introduction

Stochastic programs can effectively describe the decision-making problem in an uncertain environment. Unfortunately, the probability measures involved usually have highly specialized form; thus, solving the stochastic program can lead to real computational challenges. Even on a more practical level, only rarely does one holds enough information about the problem to commit to a specific stochastic models. In an effort to address these issues, a robust formulation for stochastic programs was proposed by Scarf in 1958 and has gain a lot of interest since then (see Scarf (1958), Shapiro and Kleywegt (2002), Calafiore and El Ghaoui (2006)). In this framework, one must define a set of probability measure that is assumed to include the true stochastic model for the problem. For example, one can consider the set of all distributions that matches a given support, mean and/or covariance. The objective of the problem is then reformulated under worse case analysis over the choice of a distribution in this set (hence the Distributionally Robust Stochastic Program):

$$\text{(DRSP)} \qquad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad \left(\max_{f_\xi \in \mathcal{D}} \mathbb{E}_\xi[h(\mathbf{x},\xi)] \right) \quad ,$$

where ξ is a vector of stochastic parameters, f_ξ is a distribution of ξ, $\mathcal{D}$ is the uncertainty set for this distribution, and $\mathcal{X}$ is a convex feasible set for the decision variable $\mathbf{x}$.

Although the DRSP optimization model has led to attractive solutions for specific problem forms, such as single item news vendor, regret minimization, linear chance-constrained and portfolio optimization problems (see Scarf (1958), Yue et al. (2006), Calafiore and El Ghaoui (2006) and Popescu (2007) respectively for details), the form still lacks encouraging computational properties for a general version of the cost function $h(\mathbf{x},\xi)$. Furthermore, the currently available methods can lead to a false sense of security as they often falsely assume exact knowledge of mean and covariance statistics for the stochastic parameters. For

instance, in many data-driven problems, one needs to build empirical point-estimates of these moments based on limited historical data. As the experiments presented in Section 5 will demonstrate, disregarding the uncertainty in these estimates can lead to taking poor decisions.

In this work, we make the following assumptions about the DRSP model.

ASSUMPTION 1. *The set $\mathcal{X}$ is convex and equipped with an oracle that can confirm feasibility of $\mathbf{x}$ or provide a separating hyperplane in polynomial time in n.*

ASSUMPTION 2. *The function $h(\mathbf{x},\xi) : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ can be represented in the form $h(\mathbf{x},\xi) = \max_{k\in\{1,\ldots,K\}} h_k(\mathbf{x},\xi)$ such that, for all k, $h_k : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is convex in $\mathbf{x}$ and concave in ξ. Hence, although $h(\mathbf{x},\xi)$ is convex in $\mathbf{x}$, it is not required to be concave in ξ. Furthermore, for each k, given a pair $(\mathbf{x},\xi)$, it is assumed that one can in polynomial time:*

1. *Evaluate the value of $h_k(\mathbf{x},\xi)$*
2. *Find a sub-gradient of $h_k(\mathbf{x},\xi)$ in $\mathbf{x}$*
3. *Find a sub-gradient of $-h_k(\mathbf{x},\xi)$ in ξ.*

ASSUMPTION 3. *One can define values $\gamma_1,\gamma_2 \geq 0$ such that the distribution f_ξ is known to lie in the following non-empty set of distributions:*

$$\mathcal{D}_1(\mathcal{S},\mu_0,\Sigma_0,\gamma_1,\gamma_2) = \left\{ f_\xi \;\middle|\; \begin{array}{l} \mathbb{P}(\xi\in\mathcal{S}) = 1 \\ (\mathbb{E}[\xi]-\mu_0)^\mathsf{T}\Sigma_0^{-1}(\mathbb{E}[\xi]-\mu_0) \leq \gamma_1 \\ \mathbb{E}[(\xi-\mu_0)(\xi-\mu_0)^\mathsf{T}] \preceq \gamma_2\Sigma_0 \end{array} \right\} ,$$

where $\preceq$ is a linear matrix inequality, $\mu_0 \in \mathbf{int}(\mathcal{S})$, $\Sigma_0 \succ 0$, and $\mathcal{S}$ is a convex set in $\mathbb{R}^m$ for which there exist an oracle that can confirm feasibility or provide a separating hyperplane in polynomial time.

Although its apparent technicality, Assumption 2 is very general as it allows $h(\mathbf{x},\xi)$ to represent many cost functions addressed by DRSPs in the past. Section 3.3 gives an overview of such examples. As for Assumption 3, we claim that, since the proposed distributional set $\mathcal{D}_1(\mathcal{S},\mu_0,\Sigma_0,\gamma_1,\gamma_2)$ accounts for moment uncertainty, it can often in practice model one's uncertainty in the distribution of ξ more accurately than previously proposed $\mathcal{D}$. We will later validate this claim by showing that Assumption 3 is implied with high probability by the knowledge that the distribution f_ξ has support on $\mathcal{S}$ and that it is the distribution that generated a set of independent samples $\{\xi_1,\xi_2,...,\xi_M\}$.

After reviewing prior work and difficulties related to solving DRSP models in Section 2, we show in Section 3 how, under the mentioned assumptions, the DRSP can be solved in polynomial time for a large range optimization problems. In fact, the structure of $\mathcal{D}(\mathcal{S},\mu_0,\Sigma_0,\gamma_1,\gamma_2)$ allows us to consider solving instances of the DRSP that were previously shown to be intractable for the usual form of $\mathcal{D}$ which assumes exact knowledge of the moments (see Example 1 of Section 3.3 for more details). In Section 4, we also use a frequentist approach to derive a new form of confidence region for the mean and covariance of a random vector which naturally lead to using $\mathcal{D}(\mathcal{S},\mu_0,\Sigma_0,\gamma_1,\gamma_2)$. These results should convince the reader that the distributional set that we present in this work is the right set to chose for distributional uncertainty in data-driven problems (*i.e.*, problems where knowledge of ξ is solely derived from historical data). We finally apply our framework to a portfolio selection problem. In Section 5, we demonstrate that, beside presenting computational advantages, in practice our model also performs best on the actual distribution that drives daily returns of popular stocks when compared to previously proposed DRSP formulations.

2. Background

In a single stage stochastic program, one is interested in finding an assignment for $\mathbf{x}\in\mathbb{R}^n$ that will minimize the expected value of a cost function given some underlying parameter uncertainty:

$$\text{(SP)} \qquad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}}\, \mathbb{E}_\xi[h(\mathbf{x},\xi)] \; .$$

Here, $\xi \in \mathbb{R}^m$ contains the stochastic parameters for which the distribution f_ξ is assumed to be known. The function $h(\mathbf{x}, \xi)$ is convex in $\mathbf{x}$ and maps a pair $(\mathbf{x}, \xi)$ to a cost (or penalty) value. Although the SP Problem is known to be convex in $\mathbf{x}$ and can even in some rare cases be reformulated as a deterministic program (*e.g.*, when $h(\mathbf{x}, \xi) = \xi^\mathsf{T}\mathbf{x}$), in order to solve the general formulation, one must often resort to Monte Carlo approximations which can be computationally challenging (see Shapiro (2000)).

Another difficulty in practice arises from the need to commit to a distribution f_ξ given only limited information about the stochastic parameters. For instance, one might only have in hand a set of independent samples $\{\xi_1, \xi_2, ..., \xi_M\}$ generated from f_ξ. In this case, it is still possible to formulate unbiased estimates of the moments of f_ξ:

$$\mathbb{E}[\xi] \approx \frac{1}{M}\sum_{i=1}^{M} \xi_i \qquad \mathbb{E}[(\xi - \mathbb{E}[\xi])(\xi - \mathbb{E}[\xi])^\mathsf{T}] \approx \frac{1}{M-1}\sum_{k=1}^{M} (\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\mathsf{T} \ .$$

For this reason, there has been a strong interest in the derivation of upper bounds on expected cost given information about mean μ, variance Σ or support $\mathcal{S}$ of the distribution f_ξ. This problem is often referred to as the Moment Problem:

$$\text{(MP)} \qquad \underset{f_\xi \in \mathcal{D}_0(\mathcal{S}, \mu, \Sigma)}{\text{maximize}} \ \mathbb{E}_\xi[h(\mathbf{x}, \xi)] \ ,$$

where $\mathcal{D}_0(\mathcal{S}, \mu, \Sigma)$ is the set of all probability measures with a given mean and covariance and with support on $\mathcal{S}$. We refer the reader to Prékopa (1995) for more details on the general form of this problem. A popular special case of this problem, which will be revisited in Section 3.3, occurs when the penalty function takes the form $h(\mathbf{x}, \xi) = \mathbb{1}\{\xi \in \mathcal{C}\}$. Solving the MP in this case leads to the formulation of interesting multivariate Chebyshev inequalities as shown in Marshall and Olkin (1960) and Bertsimas and Popescu (2005). Lately, the MP model with other forms of distributional sets has also been considered and put to practical use in robustness analysis (see *e.g.*, Barmish and Lagoa (1997)).

Although the MP model is interesting in its own right, it is often only a mean for taking optimal decisions. Therefore, in this work we will mostly be interested in solving the DRSP model. This model was first presented by Scarf (1958) in the context of an inventory management problem and since then has been referred to as *minimax stochastic programming* (*e.g.*, Dupacová (1980), Shapiro and Kleywegt (2002)), optimization with *incomplete* or *limited distribution information* (*e.g.*, Ermoliev et al. (1985)) and more recently as *distributionally robust* optimization. Its main application have focused stochastic linear programming with or without chance constraints as in Calafiore and El Ghaoui (2006) and in Chen et al. (2007). Although optimization models of a more general form have already been considered, for instance in Ermoliev et al. (1985), the field still lacks tractable solution methods for them.

In his original model, Scarf's considered a one dimensional decision variable x representing how much inventory one should hold, and ξ represented a random demand with known mean and variance. The return function had the form $h(x, \xi) = -\min\{rx - cx \ , \ r\xi - cx\}$, which actually satisfies Assumption 2. To solve this model, H. Scarf exploited the fact that the worse case distribution of demand could be chosen to be one with all its weight on two points. This idea was successfully reused in other inventory management problem model such as in Yue et al. (2006) and Zhu et al. (2006) where the objective consisted instead of minimizing the worse case regret, in absolute or relative terms, which would result from having committed to a decision once the true distribution is revealed.

More recently, a DRSP model was also proposed by Popescu (2007) to address portfolio optimization problem. Here the return function takes a more interesting shape: $h(\mathbf{x}, \xi) = -u(\xi^\mathsf{T}\mathbf{x})$ with $u(\cdot)$ including a range of useful utility functions. Again, the presented solution assumed known first and second moments of the stochastic parameters and relied on characterizing the worse case distribution of investment returns as a point distribution. Unfortunately, these examples are part of only a few special cases were the DRSP with known moments was shown to have a tractable solution. The main contribution of this paper is to provide for a range of stochastic programming models a robust yet tractable framework which takes into account distribution information that is limited with respect to both its form and moments.

3. Robust Stochastic Programming with Moment uncertainty

It is often the case in practice that one has limited information about the distribution driving the stochastic parameters which are involved in the decision-making process. For example, an investment manager can not know exactly the joint distribution of return for any available securities. Or in a different context, manufacturing decisions are rarely made knowing the distribution of future demand. We believe that in such problems, it is also rarely the case that one holds exact information about the moments of the random variables that are involved. Although the assumption of known moments has already led to interesting solutions for these problems, we will show that there is more to be gained, both on a theoretical and practical point of view, by explicitly addressing limited moments information when solving stochastic programs.

In what follows, we represent the overall uncertainty in the distribution f_ξ as proposed in Assumption 3. Given a convex support $\mathcal{S}$ for the distribution, we assume that the uncertainty in the first and second order moments of the stochastic parameters ξ can be described by uncertainty sets centered at $\mu_0 \in \mathbf{int}(\mathcal{S})$ and positive definite matrix $\Sigma_0 \succ 0$:

$$(\mathbb{E}[\xi] - \mu_0)^\mathsf{T} \Sigma_0^{-1} (\mathbb{E}[\xi] - \mu_0) \leq \gamma_1 \tag{1a}$$
$$\mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \preceq \gamma_2 \Sigma_0 \ . \tag{1b}$$

In this formulation, the parameters $\gamma_1 \geq 0$ and $\gamma_2 > 0$ provide natural means of quantifying the size of one's confidence in his estimates of mean and covariance respectively. Constraint (1a) forces the mean of ξ to lie in an ellipsoid of radius γ_1 centered at the estimate μ_0. On the other hand, Constraint (1b) forces the second order moment matrix $\mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}]$ to lie in the intersection of two positive semi-definite cones:

$$0 \preceq \mathbb{E}[((\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \preceq \gamma_2 \Sigma_0 \ .$$

In other words, it describes how likely ξ is to be closed to μ_0 in terms of the correlations expressed in Σ_0. The distributional set $\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2)$ is necessarily non-empty since it can be shown to always contain the distribution that puts all of the weight at the point μ_0.

REMARK 1. While our proposed uncertainty model cannot be used to express arbitrarily large confidence in the second order statistics of ξ, in sections 4 and 5, we will show how in practice there are natural ways of assigning μ_0, Σ_0, γ_1 and γ_2 based on historical data and generate meaningful decisions. Of course, in some situation it might be interesting to add the following constraint on the second order moment of ξ.

$$\gamma_3 \Sigma_0 \preceq \mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \ , \tag{2}$$

where $0 \leq \gamma_3 \leq 1$. Unfortunately, this leads to important computational difficulties for the general DRSP form. Furthermore, in most applications of our model, we expect the worse case distribution to actually achieve maximum variance, thus making Constraint (2) unnecessary. For example, the portfolio optimization problem presented in Section 5 will have this characteristic because a less predictable market necessarily leads to a non-negative reduction in expected utility given that the utility function is concave.

In what follows, we will study the MP and the DRSP models under the distributional set formulated in terms of $\mathcal{D}_1(\mathcal{S}, \mu_0, \Sigma_0, \gamma_1, \gamma_2)$ which will also be referred to in short-hand notation as $\mathcal{D}_1(\gamma_1, \gamma_2)$.

3.1. The Moment Problem with Actual Moment Uncertainty

In this section, we address the Moment Problem with distributional set $\mathcal{D}_1(\gamma_1, \gamma_2)$. By Assumption 2, $h(\mathbf{x}, \xi) = \max_{k \in \{1,\ldots,K\}} h_k(\mathbf{x}, \xi)$ with each $h_k(\mathbf{x}, \xi)$ concave in ξ for all $\mathbf{x} \in \mathcal{X}$. We can now show a polynomial method for finding the optimal value of this version of the Moment Problem.

DEFINITION 1. Let $\Psi(\gamma_1, \gamma_2)$ be the optimal value of the moment problem:

$$\underset{f_\xi \ \in \ \mathcal{D}_1(\gamma_1, \gamma_2)}{\text{maximize}} \ \mathbb{E}_\xi[h(\mathbf{x}, \xi)] \ .$$

Given that one considers f_ξ to be an infinite dimensional vector indexed by $\xi \in \mathcal{S}$, such that $f_\xi(\xi) : \mathcal{S} \to \mathbb{R}^+$, and $\mu \in \mathbb{R}^m$ to be a free variable, the value $\Psi(\gamma_1, \gamma_2)$ can be reformulated as the optimal value of the infinite dimensional conic linear problem:

$$\begin{array}{ll} \underset{\mu, f_\xi}{\text{maximize}} & \mathbb{E}[h(\mathbf{x}, \xi)] \\ \text{subject to} & \mathbb{E}[1] = 1 \ , \ \mathbb{E}[\xi] = \mu \\ & \mathbb{E}[(\xi - \mu_0)(\xi - \mu_0)^\mathsf{T}] \preceq \gamma_2 \Sigma_0 \\ & \begin{bmatrix} \Sigma_0 & (\mu - \mu_0) \\ (\mu - \mu_0)^\mathsf{T} & \gamma_1 \end{bmatrix} \succeq 0 \\ & f_\xi(\xi) \geq 0 \ , \ \forall \xi \in \mathcal{S} \ . \end{array}$$

Such a problem can be referred to as our primal MP. As it is often done with the moment problem, we are about to shortcut the difficult in finding a worse case probability measure for ξ by making use of duality theory. One can show that the conditions that $\mu_0 \in \mathbf{int}(\mathcal{S})$ and $\Sigma_0 \succ 0$ are sufficient conditions for strong duality to hold in this problem. Intuitively, they ensure that the interior of the feasible set is non-empty in the topology of f_ξ. We refer the reader to Shapiro (2001) (more specifically Proposition 3.4) for a thorough discussion on duality theory in the case of general conic linear problems and general moment problems. In our case, strong duality implies that $\Psi(\gamma_1, \gamma_2)$ is also the optimal value of the dual MP:

$$\begin{array}{lll} \underset{r, \mathbf{q}, \mathbf{Q}, \mathbf{P}, \mathbf{p}, s}{\text{minimize}} & \gamma_2(\Sigma_0 \bullet \mathbf{Q}) - \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + r + (\Sigma_0 \bullet \mathbf{P}) - 2\mu_0^\mathsf{T}\mathbf{p} + \gamma_1 s & \text{(3a)} \\ \text{subject to} & \mathbf{q} + 2\mathbf{Q}\mu_0 + 2\mathbf{p} = 0 & \text{(3b)} \\ & \xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r - h_k(\mathbf{x}, \xi) \geq 0 \ , \ \forall \, \xi \in \mathcal{S}, \ k \in \{1, ..., K\} & \text{(3c)} \\ & \mathbf{Q} \succeq 0 & \text{(3d)} \\ & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & s \end{bmatrix} \succeq 0 \ , & \text{(3e)} \end{array}$$

where $(A \bullet B)$ refers to the Frobenius inner product between matrices, $\mathbf{Q}, \mathbf{P} \in \mathbb{R}^{m \times m}$ are symmetric matrices, the vectors $\mathbf{q}, \mathbf{p} \in \mathbb{R}^m$, and $r, s \in \mathbb{R}$. Note that in the dual problem, we are dealing with an infinite number of constraints indexed by $\xi \in \mathcal{S}$. Fortunately, our assumption about the structure of $h(\mathbf{x}, \xi)$ and $\mathcal{S}$ leads to the dual problem having the property that cutting planes can be generated efficiently.

We first present a lemma describing the computational difficulties of verifying if a given assignment for $\mathbf{Q}$, $\mathbf{q}$, r and $\mathbf{x}$ satisfy Constraint (3c).

LEMMA 1. *Given any fixed assignment for* $\mathbf{x}$, $\mathbf{Q}$, $\mathbf{q}$, *and* r, *such that* $\mathbf{Q} \succeq 0$, *one can find for any* $k \in \{1, 2, ..., K\}$ *in polynomial time an assignment* ξ_* *that minimizes the following problem*

$$\underset{\xi \in \mathcal{S}}{\text{minimize}} \ \ \xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r - h_k(\mathbf{x}, \xi) \ . \tag{4}$$

Proof: Because Problem (4) is convex in ξ, the result is a straightforward consequence of an important property of convex minimization problems. Theorem 8.1 in Schrader (1983) demonstrates polynomial equivalence between the separation problem and the optimization problem for general convex problems. More specifically, the author shows that convex problems can be solved in polynomial time using the ellipsoid method given that one can in time polynomial in m:

1. Given any ξ, verify feasibility.
2. Given any infeasible ξ, provide a hyperplane that separates ξ from the feasible set $\mathcal{S}$.
3. Given any feasible ξ, evaluate the objective function and generate a sub-gradient in ξ.

In Problem (4), conditions (1) and (2) are satisfied by the assumption on $\mathcal{S}$. Because $h_k(\mathbf{x}, \xi)$ satisfies Assumption 2, one can also easily conclude that the objective function meets Condition (3). It follows naturally that applying the ellipsoid method will converge to the optimal solution of Problem (4) in polynomial time. □

We are now able to derive an important result about the complexity of solving the Moment Problem equipped with $\mathcal{D}_1(\gamma_1, \gamma_2)$.

PROPOSITION 1. *The value $\Psi(\gamma_1, \gamma_2)$ can be computed in time polynomial in the dimension of ξ.*

Proof: We will compute the optimal value of this semi-infinite conic linear program by solving its equivalent dual form. Applying the Schrader conditions presented in the proof of Lemma 1 on Problem (3) will lead to showing that the problem can be solved efficiently using the ellipsoid method. Since the objective is linear, Condition (3) is necessarily met. Without loss of generality, it is sufficient to verify that constraints (3e), (3d), and (3c) meet conditions (1) and (2) since the equality Constraint (3b) can easily be removed by the change of variable $\mathbf{p} = \mathbf{q}/2 + \mathbf{Q}\mu_0$. Both constraints (3e) and (3d) are easily verified in $O(m^3)$ and $O((m+1)^3)$ respectively using eigenvalue decomposition. Moreover, if necessary, for both constraints a feasibility cut can be generated from the eigenvector corresponding to the lowest eigenvalue. Finally, when considering the k-th element of Constraint (3c), since by Lemma 1, $\xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r - h_k(\mathbf{x}, \xi)$ can be minimized over $\xi \in \mathcal{S}$ in polynomial time, the feasibility of $(\bar{\mathbf{Q}}, \bar{\mathbf{q}}, \bar{r}, \bar{\mathbf{P}}, \bar{\mathbf{p}}, \bar{s})$ for a fixed $\mathbf{x}$ depends on the optimal value of Problem (4) being greater than 0 which was shown to be solvable in polynomial time. Since Constraint (3c) contains finite set of element, Condition 1 is satisfied. In the case that one of them, say the k^*-th one, is found to be infeasible, the certificate ξ_* can be used to generate a feasibility cut:

$$(\xi_* {\xi_*}^\mathsf{T} \bullet \mathbf{Q}) + {\xi_*}^\mathsf{T}\mathbf{q} + r \geq h_{k^*}(\mathbf{x}, \xi_*)$$

We conclude that determining feasibility or finding a feasibility cut can be done in polynomial time. Therefore, $\Psi(\gamma_1, \gamma_2)$ can be computed in polynomial time too using the ellipsoid method. □

3.2. The Distributionally Robust Stochastic Program with Moment Uncertainty

We now address the more interesting DRSP model using our defined distributional set:

$$\begin{aligned} \underset{\mathbf{x}}{\text{minimize}} \quad & \left(\max_{f_\xi \in \mathcal{D}_1(\gamma_1,\gamma_2)} \mathbb{E}_\xi[h(\mathbf{x}, \xi)] \right) && \text{(5a)} \\ \text{subject to} \quad & \mathbf{x} \in \mathcal{X} \ , && \text{(5b)} \end{aligned}$$

where $\mathcal{X}$, $h(\mathbf{x}, \xi)$ and $\mathcal{D}_1(\gamma_1, \gamma_2)$ satisfy assumption 1, 2 and 3 respectively. We will show that given these assumptions, Problem (5) can actually be solved efficiently.

PROPOSITION 2. *The DRSP model equipped with $\mathcal{D}_1(\gamma_1, \gamma_2)$,* i.e., *Problem* (5)*, can be solved in time polynomial in the dimensions of $\mathbf{x}$ and ξ.*

Proof: The proof of this theorem follows similar lines as the proof for Proposition 1. We first reformulate the inner moment problem in its dual form and use the fact that min-min operations can be performed jointly:

$$\begin{aligned} \underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s}{\text{minimize}} \quad & \gamma_2(\Sigma_0 \bullet \mathbf{Q}) - \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + r + (\Sigma_0 \bullet \mathbf{P}) - 2\mu_0^\mathsf{T}\mathbf{p} + \gamma_1 s && \text{(6a)} \\ \text{subject to} \quad & \mathbf{q} + 2\mathbf{Q}\mu_0 + 2\mathbf{p} = 0 && \text{(6b)} \\ & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & r \end{bmatrix} \succeq 0 && \text{(6c)} \\ & \mathbf{Q} \succeq 0 && \text{(6d)} \\ & \xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r - h_k(\mathbf{x}, \xi) \geq 0 \ , \ \forall \xi \in \mathbb{R}^m, k \in \{1, ..., K\} && \text{(6e)} \\ & \mathbf{x} \in \mathcal{X} \ , && \text{(6f)} \end{aligned}$$

As presented earlier, the ellipsoid method will converge in polynomial time given that we can verify feasibility or provide feasibility cuts in polynomial time. The arguments that were presented in the proof of Proposition 1 still apply for constraints (6b), (6c) and (6d). However, the argument for Constraint (6e) needs to be revisited since $\mathbf{x}$ is now consider as an optimization variable. Overall feasibility of an assignment $(\bar{\mathbf{x}}, \bar{\mathbf{Q}}, \bar{\mathbf{q}}, \bar{r}, \bar{\mathbf{P}}, \bar{\mathbf{p}}, \bar{s})$ can again be verified in polynomial time because of Lemma 1 and of the fact that K is

finite. However, in the case that one of the indexed constraints, say the k^*-th one, is found to be infeasible, one can again easily compute a feasibility cut using the certificate ξ_* and a sub-gradient of $h_{k^*}(\cdot,\xi_*)$:

$$(\xi_*{\xi_*}^\mathsf{T} \bullet \mathbf{Q}) + {\xi_*}^\mathsf{T}\mathbf{q} + r - \nabla h_{k^*}(\bar{\mathbf{x}},\xi_*)^\mathsf{T}\mathbf{x} \geq h_{k^*}(\bar{\mathbf{x}},\xi_*) - \nabla h_{k^*}(\bar{\mathbf{x}},\xi_*)^\mathsf{T}\bar{\mathbf{x}} \ ,$$

where $\nabla h_k(\bar{\mathbf{x}},\xi_*)$ is a subgradient of $h_k(\cdot,\xi_*)$ at $\bar{\mathbf{x}}$. Since by Assumption 2, such a gradient can be obtained in polynomial time, we can conclude that Problem (6) can also be solved in polynomial time. □

3.3. Practical Examples

Because our framework only imposes weak conditions on $h(\mathbf{x},\xi)$, it is possible to revisit some well-known practical problems and reformulate them taking into account moment uncertainty.

EXAMPLE 1. **Optimal Inequalities in Probability Theory.**

Consider the problem of finding a tight upper bound on $\mathbb{P}(\xi \in \mathcal{C})$ for a random vector ξ with known support, mean, and covariance matrix. By formulating this problem as a semi-infinite linear program:

$$\underset{f_\xi \in \mathcal{D}_0(\mathcal{S},\mu,\Sigma)}{\text{maximize}} \ \mathbb{P}(\xi \in \mathcal{C}) \ = \ \mathbb{E}[\mathbb{1}\{\xi \in \mathcal{C}\}] \ ,$$

many have studied the difficulties related to extensions of the popular Chebyshev inequalities (see Marshall and Olkin (1960), Bertsimas and Popescu (2005)). More specifically, given that $\mathcal{C}$ is a finite union of disjoint convex sets, it is demonstrated that when $\mathcal{S} = \mathbb{R}^m$, the bound can be found in polynomial time, while for restricted support such as $\mathcal{S} = \mathbb{R}^+$ the problem is NP-hard. The hardness of the problem arises already in finding a distribution that is feasible with respect to $\mathcal{D}_0(\mathbb{R}^+,\mu,\Sigma)$.

Our framework recommends relaxing the restrictions on the covariance of ξ and instead consider the distributional set $\mathcal{D}_1(\gamma_1,\gamma_2)$. This set considers any distributions on a given convex support with first and second moment lying in the uncertainty sets parameterized by γ_1 and γ_2. Already we can realize that the distribution which puts all of its weight in the mean is always feasible with respect to $\mathcal{D}_1(\gamma_1,\gamma_2)$. Furthermore, when $\mathcal{C}$ is represented as $\mathcal{C} = \bigcup_{k=1}^K \mathcal{C}_k$, with $\mathcal{C}_k$ convex, our results actually lead to a new type of Chebyshev inequality that can be evaluated in polynomial time. First, one chooses $h_0(\mathbf{x},\xi) = 0$ and $h_k(\mathbf{x},\xi) = \begin{cases} 1 & \text{, if } \xi \in \mathcal{C}_i \\ -\infty & \text{, otherwise} \end{cases}$ in order to construct a function $h(\mathbf{x},\xi)$ which satisfies Assumption 2. Then, by the equivalence:

$$\mathbb{E}_\xi[h(\mathbf{x},\xi)] = \mathbb{E}_\xi[\max_k \ h_k(\mathbf{x},\xi)] = \mathbb{E}_\xi[\mathbb{1}\{\xi \in \mathcal{C}\}] = \mathbb{P}(\xi \in \mathcal{C}) \leq \max_{f_\xi \in \mathcal{D}_1(\gamma_1,\gamma_2)} \mathbb{E}_\xi[h(\mathbf{x},\xi)] \ ,$$

it follows that for distributions in $\mathcal{D}_1(\gamma_1,\gamma_2)$, a tight Chebyshev bound can be found in polynomial time. Note that by using the form $\mathcal{D}_1(\mathcal{S},\mu,\Sigma,0,1)$ one can also provide useful approximations to the mentioned NP-hard versions of the problem with $\mathcal{D}_0$.

EXAMPLE 2. **Distributionally Robust Optimization with piecewise-linear convex costs.**

Assume that one is interested in solving the following DRSP model for a general piece-wise linear convex cost function of $\mathbf{x}$

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \ \left(\max_{f_\xi \in \mathcal{D}_1(\gamma_1,\gamma_2)} \mathbb{E}_\xi[\max_k \ \xi_k^\mathsf{T}\mathbf{x}] \right) \ ,$$

where $\xi_k \in \mathbb{R}^n$ are random vectors. By considering ξ to be a random matrix whose k-th column is the random vector ξ_k and taking $h_k(\mathbf{x},\xi) = \xi_k^\mathsf{T}\mathbf{x}$, which is linear (hence concave) in ξ, the results presented earlier allows one to conclude that the problem can be solved efficiently. Note that since any convex cost function can be approximated by a piecewise linear function, this argument could potentially be used on a wide range of stochastic programs. In particular, Section 5 will investigate further a case of portfolio optimization.

EXAMPLE 3. **Distributionally robust conditional value-at-risk.**

Conditional value-at-risk, also called mean excess loss, was introduced in the mathematical finance community as a new risk measure in decision-making. It is closely related to the more common value-at-risk measure, which for a risk tolerance level of $\vartheta \in (0,1)$ evaluates the lowest amount τ such that with probability $1-\vartheta$, the loss does not exceed τ. CVaR instead evaluates the conditional expectation of loss above the value-at-risk. In order to keep the focus of our discussion on the topic of DRSP models, we refer the reader to Rockafellar and Uryasev (2000) for technical details on this subject. CVaR has gained a lot of interest in the community because of its attractive numerical properties. For instance, Rockafellar and Uryasev (2000) demonstrated that one could evaluate the ϑ-CVaR$_\xi[c(\mathbf{x},\xi)]$ of a cost (or lost) function $c(\mathbf{x},\xi)$ with random parameters distributed according to f_ξ by solving a minimization problem of convex form:

$$\vartheta\text{-CVaR}_\xi[c(\mathbf{x},\xi)] = \min_{\lambda\in\mathbb{R}} \ \lambda + \frac{1}{\vartheta}\mathbb{E}_\xi\left[(c(\mathbf{x},\xi)-\lambda)^+\right] \ ,$$

where $(y)^+ = \max\{y\ ,\ 0\}$.

The risk measure known as CVaR still requires the decision maker to commit to a distribution f_ξ. This is a step that can be difficult to take in practice. Using the results presented earlier in this section, we can easily demonstrate how the CVaR measure can be considered in its distributionally robust form. Given that the distribution is known to lie in a distributional set $\mathcal{D}_1(\gamma_1,\gamma_2)$, the Distributionally Robust ϑ-CVaR Problem can be expressed as:

$$\text{(DR } \vartheta\text{-CVaR)} \qquad \underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \ \left(\max_{f_\xi\in\mathcal{D}_1(\gamma_1,\gamma_2)} \vartheta\text{-CVaR}_\xi[c(\mathbf{x},\xi)]\right) \ .$$

By the equivalence statement presented above, this problem can be solved in the form:

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \ \left(\max_{f_\xi\in\mathcal{D}_1(\gamma_1,\gamma_2)} \ \left(\min_{\lambda\in\mathbb{R}} \ \lambda + \frac{1}{\vartheta}\mathbb{E}_\xi\left[(c(\mathbf{x},\xi)-\lambda)^+\right]\right)\right) \ .$$

Using duality theory, one can show that the function $g(f_\xi,\lambda) = \lambda + (1/\vartheta)\mathbb{E}_\xi\left[\max\{c(\mathbf{x},\xi)-\lambda\ ,\ 0\}\right]$, which is convex in λ and concave (actually linear) in f_ξ, has the strong max-min property over the joint set $(f_\xi,\lambda) \in \mathcal{D}_1(\gamma_1,\gamma_2)\times\mathbb{R}$. Hence, changing the order of the $\max_{f_\xi}$ and $\min_\lambda$ operators leads to an equivalent formulation for the (DR ϑ-CVaR) Problem.

$$\underset{\mathbf{x}\in\mathcal{X},\lambda\in\mathbb{R}}{\text{minimize}} \ \left(\max_{f_\xi\in\mathcal{D}_1(\gamma_1,\gamma_2)} \mathbb{E}_\xi[h(\mathbf{x},\lambda,\xi)]\right) \ ,$$

where $h(\mathbf{x},\lambda,\xi) = \lambda + \frac{1}{\vartheta}(c(\mathbf{x},\xi)-\lambda)^+$. Because of the argument that

$$\begin{aligned} h(\mathbf{x},\lambda,\xi) &= \lambda + \frac{1}{\vartheta}\max\{\ 0\ ,\ c(\mathbf{x},\xi)-\lambda\ \} \\ &= \max\left\{\ \lambda\ ,\ \left(1-\frac{1}{\vartheta}\right)\lambda + \frac{1}{\vartheta}\max_k\ c_k(\mathbf{x},\xi)\ \right\} \\ &= \max\left\{\ \lambda\ ,\ \max_k\ \left(1-\frac{1}{\vartheta}\right)\lambda + \frac{1}{\vartheta}c_k(\mathbf{x},\xi)\ \right\} \ , \end{aligned}$$

it is clear that if $c(\mathbf{x},\xi)$ meets the conditions presented in Assumption 2, then necessarily $(1-\frac{1}{\vartheta})\lambda + \frac{1}{\vartheta}c_k(\mathbf{x},\xi)$ meets the same three conditions for all k. And, in a rather trivial way so does the function $c_0(\mathbf{x},\lambda,\xi) = \lambda$. Because we can show that the function $h(\mathbf{x},\lambda,\xi)$ meets Assumption 2, Proposition 2 allows us to conclude that finding an optimal $\mathbf{x}$ (and its associated λ) with respect to the worse case conditional value-at-risk obtained over the set of distributions $\mathcal{D}_1(\gamma_1,\gamma_2)$ can be done in polynomial time.

4. Data-driven Stochastic Programming

The computational results presented in the previous section rely heavily on the structure of the described distributional set $\mathcal{D}_1(\gamma_1, \gamma_2)$. This set was built to take into account moment uncertainty in the stochastic parameters. We now turn ourselves to showing that such structure can be naturally justified in the context of data-driven optimization problems. More specifically, these are problems were knowledge of the stochastic parameters is restricted to a set of samples $\{\xi_i\}_{i=1}^M = \{\xi_1, \xi_2, ..., \xi_M\}$ drawn independently from the underlying distribution f_ξ. Under such conditions, a common approach is to assume that the true moments lie in a neighborhood of there empirical estimates. In what follows, we will show how one can define a confidence region for mean and covariance statistics such that it is assured with high probability to contain the given statistics of the distribution that generated $\{\xi_i\}_{i=1}^M$. This result will in turn be used to derive a distributional set of the form $\mathcal{D}_1(\gamma_1, \gamma_2)$ and to provide probabilistic guarantees that the solution found using the MP or DRSP models is robust with respect to the true underlying distribution of the stochastic parameters ξ.

In order to simplify the derivations, we start by reformulating without loss of generality the random vector ξ in terms of a mixture of uncorrelated component ζ in order to simplify the derivations. More specifically, given the random vector $\xi \in \mathbb{R}^m$ with mean μ and covariance Σ, let us define $\zeta \in \mathbb{R}^m$ to be the normalized random vector $\zeta = \Sigma^{-1/2}(\xi - \mu)$ such that $\mathbb{E}[\zeta] = 0$ and $\mathbb{E}[\zeta\zeta^\mathsf{T}] = \mathbf{I}$. Also, let us make the following assumption about ζ:

ASSUMPTION 4. *There exists a ball of radius R that contains the entire support of the unknown distribution of ζ. Equivalently, there exist $R \geq 0$ such that*

$$\mathbb{P}\left((\xi - \mu)^T \Sigma^{-1} (\xi - \mu) \leq R^2\right) = 1 \ .$$

This assumption is made in order to use an inequality known as the "independent bounded differences inequality" and popularized by McDiarmid.

THEOREM 1. *(McDiarmid (1998)) Let $\{\xi_i\}_{i=1}^M$ be a set of independent random variables ξ_i taking values in a set $\mathcal{S}_i$ for each i. Suppose that the real-valued function $g(\xi_1, \xi_2, ..., \xi_M)$ defined on $\mathcal{S}_1 \times \mathcal{S}_2 \times ... \times \mathcal{S}_M$ satisfies*

$$|g(\xi_1, \xi_2, ..., \xi_M) - g(\xi_1', \xi_2', ..., \xi_M')| \leq c_j \tag{7}$$

whenever the vector sets $\{\xi_i\}_{i=1}^M$ and $\{\xi_i'\}_{i=1}^M$ differ only in the j-th vector. Then for any $t \geq 0$,

$$\mathbb{P}\left(g(\xi_1, \xi_2, ..., \xi_M) - \mathbb{E}[g(\xi_1, \xi_2, ..., \xi_M)] \leq -t\right) \leq \exp\left(\frac{-2t^2}{\sum_{j=1}^M c_j^2}\right) \ .$$

In practice, even when one does not have information about μ and Σ, we believe that one can often still make an educated and conservative guess about the magnitude of R. We will also revisit this issue in Section 4.3 where we derive R based on the bounded support of ξ. Note that if ξ's support is unbounded, one can also derive bounds of similar nature either by considering that ζ is bounded with high probability, or otherwise by using Markov's inequality as a foundation. However, because Markov's inequality does not require any support assumption, the bounds that are derived with it are more sensitive to the confidence level that one needs to achieve.

4.1. Uncertainty Cone Centered at Empirical Mean

A first use of the McDiarmid theorem leads to defining a conic constraint relating the true mean and true covariance of the random vector ξ to the empirical point estimate $\hat{\mu} = M^{-1}\sum_{k=1}^M \xi$. In Shawe-Taylor and Cristianini (2003), the authors used McDiarmid's theorem to demonstrate the following result.

LEMMA 2. *(Shawe-Taylor and Cristianini (2003)) Let $\{\zeta_i\}_{i=1}^M$ be a set of M samples generated independently at random from ζ. Then with probability at least $(1-\delta)$ over the choice of sets $\{\zeta_i\}_{i=1}^M$, we have*

$$\left\| \frac{1}{M} \sum_{i=1}^{M} \zeta_i \right\|^2 \leq \frac{R^2}{M} \left(2 + \sqrt{2\ln(1/\delta)} \right)^2 \ .$$

We can use this result to derive a similar statement about the random vector ξ.

COROLLARY 1. *Given the true mean μ and covariance Σ of ξ and given that ξ lies on the support $(\xi-\mu)^\mathsf{T}\Sigma^{-1}(\xi-\mu) \leq R^2$ with probability one, the point estimate $\hat{\mu}$ satisfies the following constraint with probability greater than $1-\delta$:*

$$(\hat{\mu}-\mu)^\mathsf{T}\Sigma^{-1}(\hat{\mu}-\mu) \leq \beta(\delta) \ , \tag{8}$$

where $\beta(\delta) = (R^2/M)(2+\sqrt{2\ln(1/\delta)})^2$.

Proof: The generalization to a ξ with arbitrary mean and covariance matrix is quite straightforward:

$$\begin{aligned}
\mathbb{P}\left((\hat{\mu}-\mu)^\mathsf{T}\Sigma^{-1}(\hat{\mu}-\mu) \leq \beta(\delta)\right) &= \mathbb{P}\left(\left\| \Sigma^{-1/2}\left(\frac{1}{M}\sum_{i=1}^{M} \xi_i - \mu \right) \right\|^2 \leq \beta(\delta) \right) \\
&= \mathbb{P}\left(\left\| \frac{1}{M}\sum_{i=1}^{M} \Sigma^{-1/2}(\xi_i - \mu) \right\|^2 \leq \beta(\delta) \right) \\
&= \mathbb{P}\left(\left\| \sum_{i=1}^{M} \zeta_i \right\|^2 \leq \beta(\delta) \right) \geq 1-\delta \ . \qquad \square
\end{aligned}$$

Given that Σ is non-singular, the inequality of Equation (8) constrains the vector μ and matrix Σ to a convex set. This set can be represented by the following linear matrix inequality after applying the principles of Schur's complement:

$$\begin{bmatrix} \Sigma & (\hat{\mu}-\mu) \\ (\hat{\mu}-\mu)^\mathsf{T} & \beta(\delta) \end{bmatrix} \succeq 0 \ .$$

4.2. Uncertainty Cone Centered at Empirical Covariance

In order for Constraint (8) to describe a bounded set, one must be able to contain the uncertainty in Σ. While confidence regions for the covariance matrix are typically defined on a term by term basis (see for example Shawe-Taylor and Cristianini (2003)), we favor the structure imposed by a constraint of the positive semi-definite form

$$\mathbb{P}\left(c_{\min}\hat{\Sigma} \preceq \Sigma \preceq c_{\max}\hat{\Sigma} \right) \geq 1-\delta \tag{9}$$

around the point estimate of covariance matrix $\hat{\Sigma} = M^{-1}\sum_{i=1}^{M}(\xi_i-\hat{\mu})(\xi_i-\hat{\mu})^\mathsf{T}$. Note that the difficulty of this task relies heavily on the fact that one needs to derive a confidence interval for the eigenvalues of the stochastic matrix $\Sigma^{-1/2}\hat{\Sigma}\Sigma^{-1/2}$ which is an important field of study in statistics. For the case that interests us, where $M \gg m$ with m fixed, prior work usually assumes ξ is a normally distributed random vector (see Anderson (1984), Edelman (1989)). Under the Gaussian assumption, the sample covariance matrix follows the Wishart distribution, thus one can formulate the distribution of eigenvalues in a closed form expression and derive such percentile bounds. In the case where ξ takes a non-normal form, the asymptotic distribution of eigenvalues was studied by Waternaux (1976) and Fujikoshi (1980) among others. However, to the best of our knowledge, our work is the first to formulate an uncertainty sets with the characteristics presented in Equation (9) for finite sample size M.

In what follows, we present how to formulate the set of Equation (9) based on Assumption 4 about the bounded support of ζ. We start by demonstrating how, for a zero mean and uncorrelated random vector such

as ζ, a confidence region of the form presented in Equation (9) can be defined around $\hat{\mathbf{I}} = M^{-1}\sum_i \zeta_i\zeta_i^\mathsf{T}$. Next, we will assume that the mean of ξ is exactly known and will formulate it in terms of $\hat{\Sigma}(\mu) = M^{-1}\sum_{i=1}^M(\xi_i-\mu)(\xi_i-\mu)^\mathsf{T}$. We conclude this section with our main result about a confidence region for μ and Σ which relies solely on M and on support information about the random variables involved.

LEMMA 3. *Given M samples from ζ, $\{\zeta_i\}_{i=1}^M$, and an empirical estimate of the covariance $\hat{\mathbf{I}}$, then with probability greater than $1-\delta$:*

$$\frac{1}{1+\alpha(\delta/2)}\hat{\mathbf{I}} \preceq \mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}} \ , \tag{10}$$

where $\alpha(\delta/2) = (R^2/\sqrt{M})\left(\sqrt{1-m/R^4}+\sqrt{\ln(2/\delta)}\right)$, given that $M > R^4\left(\sqrt{1-m/R^4}+\sqrt{\ln(2/\delta)}\right)^2$.

Proof: The proof of this theorem relies on applying Theorem 1 twice to show that both $\frac{1}{1+\alpha(\delta/2)}\hat{\mathbf{I}} \preceq \mathbf{I}$ and $\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}}$ occur with probability $1-\delta/2$. Our statement then simply follows by the union bound. However, for the sake of conciseness, this proof will focus on deriving the upper bound since the steps we will follow can easily be adjusted for the lower bound derivation.

When applying Theorem 1 to show that $\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}}$ occurs with probability $1-\delta/2$, the main step consists of defining $g(\zeta_1,\zeta_2,...,\zeta_M) = \min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}$ and finding a lower bound for $\mathbb{E}[g(\zeta_1,\zeta_2,...,\zeta_M)]$. It will be useful to show that Constraint (7) is met when $c_j = R^2/M$ for all j.

$$|g(\zeta_1,\zeta_2,...,\zeta_M) - g(\zeta_1',\zeta_2',...,\zeta_M')| = \left|\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}'\mathbf{z}\right| \ ,$$

where $\hat{\mathbf{I}}' = \frac{1}{M}\sum_{i=1}^M \zeta_i'\zeta_i'^\mathsf{T} = \hat{\mathbf{I}} + \frac{1}{M}(\zeta_j'\zeta_j'^\mathsf{T} - \zeta_j\zeta_j^\mathsf{T})$ since $\{\zeta_i\}_{i=1}^M$ and $\{\zeta_i'\}_{i=1}^M$ only differ in the j-th vector.

Now assume that $\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} \geq \min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}'\mathbf{z}$. Then, for any $\mathbf{z}^* \in \arg\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}'\mathbf{z}$

$$\begin{aligned}
|g(\zeta_1,\zeta_2,...,\zeta_M) - g(\zeta_1',\zeta_2',...,\zeta_M'))| &= \min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \mathbf{z}^{*\mathsf{T}}\hat{\mathbf{I}}'\mathbf{z}^* \\
&\leq \mathbf{z}^{*\mathsf{T}}(\hat{\mathbf{I}}-\hat{\mathbf{I}}')\mathbf{z}^* \\
&= \mathbf{z}^{*\mathsf{T}}\frac{1}{M}(\zeta_j\zeta_j^\mathsf{T} - \zeta_j'\zeta_j'^\mathsf{T})\mathbf{z}^* \\
&= \frac{1}{M}\left((\zeta_j^\mathsf{T}\mathbf{z}^*)^2 - (\zeta_j'^\mathsf{T}\mathbf{z}^*)^2\right) \\
&\leq \frac{1\|\mathbf{z}^*\|^2\|\zeta_j\|^2}{M} \leq \frac{R^2}{M}
\end{aligned}$$

In the case that $\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} \leq \min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}'\mathbf{z}$ the same argument applies using $\mathbf{z}^* \in \arg\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}$.

As for $\mathbb{E}[g(\zeta_1,\zeta_2,...,\zeta_M)]$, the task is a bit harder. We can instead try to find an upper bound on the maximum eigenvalue of $(\mathbf{I}-\hat{\mathbf{I}})$ since

$$\mathbb{E}\left[\max_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}(\mathbf{I}-\hat{\mathbf{I}})\mathbf{z}\right] = 1 - \mathbb{E}\left[\min_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}\right] \ . \tag{11}$$

Using Jenson's inequality and basic rules of linear algebra, one can show that

$$\begin{aligned}
\left(\mathbb{E}_{\hat{\mathbf{I}}}\left[\max_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}(\mathbf{I}-\hat{\mathbf{I}})\mathbf{z}\right]\right)^2 &\leq \mathbb{E}_{\hat{\mathbf{I}}}\left[\left(\max_{\|\mathbf{z}\|=1}\ \mathbf{z}^\mathsf{T}(\mathbf{I}-\hat{\mathbf{I}})\mathbf{z}\right)^2\right] \leq \mathbb{E}_{\hat{\mathbf{I}}}\left[\sum_{i=1}^m \sigma_i^2(\mathbf{I}-\hat{\mathbf{I}})\right] = \mathbb{E}_{\hat{\mathbf{I}}}\left[\mathbf{trace}\left((\mathbf{I}-\hat{\mathbf{I}})^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{trace}\left(\left(\frac{1}{M}\sum_{k=1}^M \mathbf{I} - \zeta_i\zeta_i^\mathsf{T}\right)^2\right)\right] \\
&= \mathbf{trace}\left(\frac{1}{M^2}\sum_{k=1}^M \mathbb{E}\left[\mathbf{I} - 2\zeta_i\zeta_i^\mathsf{T} + (\zeta_i\zeta_i^\mathsf{T})^2\right]\right) \\
&= \frac{1}{M}\left(\mathbf{trace}\left(\mathbb{E}\left[(\zeta_i\zeta_i^\mathsf{T})^2\right]\right) - \mathbf{trace}\left(\mathbf{I}\right)\right) = \frac{\mathbb{E}\left[\|\zeta_i\|^4\right]-m}{M} \leq \frac{R^4-m}{M} \ ,
\end{aligned}$$

where we used the fact that ζ_i are sampled independently thus $\mathbb{E}[(\mathbf{I}-\zeta_i\zeta_i^\mathsf{T})(\mathbf{I}-\zeta_j\zeta_j^\mathsf{T})] = \mathbb{E}[\mathbf{I}-\zeta_i\zeta_i^\mathsf{T}]\mathbb{E}[\mathbf{I}-\zeta_j\zeta_j^\mathsf{T}] = 0$. By replacing this lower bound in Equation (11), we can now say that $\mathbb{E}[g(\zeta_1,\zeta_2,...,\zeta_M)] \geq 1-(R^2/\sqrt{M})\sqrt{1-m/R^4}$. More importantly, Theorem 1 allows us to confirm the proposed upper bound using the following argument. Since the statement

$$\mathbb{P}\left(\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \mathbb{E}_{\hat{\mathbf{I}}}\left[\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}\right] \leq -\epsilon\right) \leq \exp\left(\frac{-2\epsilon^2}{\sum_k (R^4/M^2)}\right) \ ,$$

implies that

$$\mathbb{P}\left(\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} - \mathbb{E}_{\hat{\mathbf{I}}}[\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}] \geq -\frac{R^2\sqrt{\ln(2/\delta)}}{\sqrt{M}}\right) \geq 1-\delta/2 \ ,$$

and since relaxing $\mathbb{E}_{\hat{\mathbf{I}}}[\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}]$ to its lower bound can only include more random events, we necessarily have that

$$\mathbb{P}\left(\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z} \geq 1 - \frac{R^2}{\sqrt{M}}\left(\sqrt{1-m/R^4}+\sqrt{\ln(2/\delta)}\right)\right) \geq 1-\delta/2 \ .$$

Thus, given that M is large enough so that $1-\alpha(\delta/2) > 0$, we conclude that

$$\mathbb{P}\left(\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}}\right) \geq 1-\delta/2 \ .$$

The task of showing that $1/(1+\alpha(\delta/2))\hat{\mathbf{I}} \preceq \mathbf{I}$ also occurs with probability $1-\delta/2$ is very similar. One simply applies Theorem 1, now defining $g(\zeta_1,\zeta_2,...,\zeta_M) = -\min_{\|\mathbf{z}\|=1} \mathbf{z}^\mathsf{T}\hat{\mathbf{I}}\mathbf{z}$, and needs to demonstrate that $\mathbb{E}[g(\zeta_1,\zeta_2,...,\zeta_M)] > -1-\alpha(\delta/2)$. The rest follows easily. □

REMARK 2. In Anderson (1984), it is shown that under the assumption that ζ is Gaussian, the eigenvalues of $\sqrt{M}(\hat{\mathbf{I}}-\mathbf{I})$ are distributed according to:

$$f(\sigma_1,\sigma_2,...,\sigma_m) = \frac{1}{Z}\exp(-\frac{1}{2}\sum_{i=1}^m \sigma_i^2)\prod_{i<j}(\sigma_i-\sigma_j) \ ,$$

where $\sigma_1 > \sigma_2 > ... > \sigma_m$ and Z is a normalizing constant. Thus in the Gaussian case, one can guarantie with probability greater than $1-\delta$ that:

$$-\frac{1}{1+\frac{r}{\sqrt{M}}}\hat{\mathbf{I}} \leq \mathbf{I} \leq \frac{1}{1-\frac{r}{\sqrt{M}}}\hat{\mathbf{I}} \ ,$$

where r is the solution to the equation $\mathbb{P}(-r \leq \sigma_1 \leq \sigma_m \leq r) = 1-\delta$ with respect to the distribution $f(\sigma_1,\sigma_2,...,\sigma_m)$. This fact leads us to believe that the bound that is presented in Lemma 3 is tight up to a constant (*i.e.*, the bound is necessarily $O(1/\sqrt{M})$).

We are now interested in extending Lemma 3 to general mean and covariance random vectors. Given the random event that Constraint (10) is satisfied, then:

$$\begin{aligned}
\mathbf{I} \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\mathbf{I}} &\Rightarrow \Sigma^{1/2}\mathbf{I}\Sigma^{1/2} \preceq \frac{1}{1-\alpha(\delta/2)}\Sigma^{1/2}\hat{\mathbf{I}}\Sigma^{1/2} \\
&\Rightarrow \Sigma \preceq \frac{1}{1-\alpha(\delta/2)}\frac{1}{M}\sum_{i=1}^M \Sigma^{1/2}\zeta_i\zeta_i^\mathsf{T}\Sigma^{1/2} \\
&\Rightarrow \Sigma \preceq \frac{1}{1-\alpha(\delta/2)}\frac{1}{M}\sum_{i=1}^M (\xi_i-\mu)(\xi_i-\mu)^\mathsf{T} \\
&\Rightarrow \Sigma \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\Sigma}(\mu) \ ,
\end{aligned}$$

and similarly,

$$\frac{1}{1+\alpha(\delta/2)}\hat{\mathbf{I}} \preceq \mathbf{I} \Rightarrow \frac{1}{1+\alpha(\delta/2)}\hat{\Sigma} \preceq \Sigma \ .$$

Since Constraint (10) is satisfied with probability greater than $1-\delta$, the following corollary follows easily.

COROLLARY 2. *Given that the mean of ξ, μ, is known and used to formulate the empirical estimate of the covariance, $\hat{\Sigma}(\mu)$, then with probability greater then $1-\delta$:*

$$\frac{1}{1+\alpha(\delta/2)}\hat{\Sigma}(\mu) \preceq \Sigma \preceq \frac{1}{1-\alpha(\delta/2)}\hat{\Sigma}(\mu) \ ,$$

where $\alpha(\delta/2) = (R^2/\sqrt{M})\left(\sqrt{1-m/R^4}+\sqrt{\ln(2/\delta)}\right)$ and given that M is large enough.

This statement leads to the description of a convex set that contains mean vectors and covariance matrices for which a given empirical estimate is obtained with high probability through the sampling process.

THEOREM 2. *Given M samples from ξ, $\{\xi_i\}_{i=1}^M$, and an empirical estimate of the mean $\hat{\mu}$ and covariance matrix $\hat{\Sigma}$, then with probability greater than $1-\delta$ over the choice of $\{\xi_i\}_{i=1}^M$ the following set of constraints are met:*

$$(\hat{\mu}-\mu)\Sigma^{-1}(\hat{\mu}-\mu) \leq \beta(\delta/2) \tag{12a}$$
$$\Sigma \preceq \frac{1}{1-\alpha(\delta/4)-\beta(\delta/2)}\hat{\Sigma} \tag{12b}$$
$$\Sigma \succeq \frac{1}{1-\alpha(\delta/4)}\hat{\Sigma} \ , \tag{12c}$$

where $\alpha(\delta/4) = (R^2/\sqrt{M})\left(\sqrt{1-m/R^4}+\sqrt{\ln(4/\delta)}\right)$, $\beta(\delta/2) = (R^2/M)(2+\sqrt{2\ln(2/\delta)})^2$, and given that M is large enough.

Proof: By applying Corollary 1, 2 and Lemma 3, the union bound guarantees us with probability greater than $1-\delta$ that the following constraints will be met:

$$\begin{aligned}(\hat{\mu}-\mu)\Sigma^{-1}(\hat{\mu}-\mu) &\leq \beta(\delta/2)\\ \Sigma &\preceq \frac{1}{1-\alpha(\delta/4)}\hat{\Sigma}(\mu)\\ \Sigma &\succeq \frac{1}{1+\alpha(\delta/4)}\hat{\Sigma}(\mu) \ .\end{aligned}$$

Notice that our result is not proven yet since, although the first constraint is exactly Constraint (12a), the second and third constraints actually refer to covariance estimates that uses the true mean of the distribution instead of an empirical estimate. The following steps will convince us that these conditions are sufficient for constraints (12b) and (12c) to hold.

$$\begin{aligned}(1-\alpha(\delta/4))\Sigma \preceq \hat{\Sigma}(\mu) &= \frac{1}{M}\sum_{i=1}^M (\xi_i-\mu)(\xi_i-\mu)^\mathsf{T}\\ &= \frac{1}{M}\sum_{i=1}^M (\xi_i-\hat{\mu}+\hat{\mu}-\mu)(\xi_i-\hat{\mu}+\hat{\mu}-\mu)^\mathsf{T}\\ &= \frac{1}{M}\sum_{i=1}^M (\xi_i-\hat{\mu})(\xi_i-\hat{\mu})^\mathsf{T} + (\xi_i-\hat{\mu})(\hat{\mu}-\mu)^\mathsf{T} +\\ &\qquad\qquad (\hat{\mu}-\mu)(\xi_i-\hat{\mu})^\mathsf{T} + (\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T})\\ &= \hat{\Sigma} + (\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T}\\ &\preceq \hat{\Sigma} + \beta(\delta/2)\Sigma \ ,\end{aligned}$$

where the last semi-definite inequality of the derivation can be explained using the fact that for any $\mathbf{x} \in \mathbb{R}^m$,

$$\begin{aligned}\mathbf{x}^\mathsf{T}(\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T}\mathbf{x} &= (\mathbf{x}^\mathsf{T}(\hat{\mu}-\mu))^2 = (\mathbf{x}^\mathsf{T}\Sigma^{1/2}\,\Sigma^{-1/2}(\hat{\mu}-\mu))^2 \\ &\leq \|\mathbf{x}^\mathsf{T}\Sigma^{1/2}\|^2\|\Sigma^{-1/2}(\hat{\mu}-\mu)\|^2 \leq \beta(\delta/2)\mathbf{x}^\mathsf{T}\Sigma\mathbf{x}\ .\end{aligned}$$

Thus we can conclude that Constraint (12b) is met. The same steps can be used to show that Constraint (12c) also holds for a set of events of probability $1-\delta$.

$$\begin{aligned}(1-\alpha(\delta/4))\Sigma \preceq \hat{\Sigma}(\mu) &= \frac{1}{M}\sum_{i=1}^{M}(\xi_i-\mu)(\xi_i-\mu)^\mathsf{T} \\ &= \hat{\Sigma}+(\hat{\mu}-\mu)(\hat{\mu}-\mu)^\mathsf{T} \\ &\succeq \hat{\Sigma}\end{aligned}$$

□

4.3. Bounding the Support of ζ using Empirical Data

The above derivations assumed that one is able to describe a ball containing the support of the rather fictive random vector ζ. In fact, this assumption can be replaced by an assumption on the support of the more tangible random vector ξ as is presented in the following corollary.

COROLLARY 3. *Given that the support $\mathcal{S}_\xi$ of the distribution of ξ is known, let*

$$\hat{R} = \sup_{\xi\in\mathcal{S}_\xi}\|\hat{\Sigma}^{-1/2}(\xi-\hat{\mu})\|_2$$

be an approximation of R using the available empirical data. For $\delta_1=\delta_2=1-\sqrt{1-\delta}$, given that

$$M > \max\left\{(\hat{R}^2+2)^2\left(2+\sqrt{2\ln(4/\delta_1)}\right)^2\ ,\ \frac{\left(8+\sqrt{32\ln(4/\delta)}\right)^2}{\left(\sqrt{\hat{R}+4}-\hat{R}\right)^4}\right\}\ , \tag{13}$$

then Theorem 2 applies with $\bar{\alpha}(\delta_2/4)=(\bar{R}^2/\sqrt{M})\left(\sqrt{1-m/\bar{R}^4}+\sqrt{\ln(4/\delta_2)}\right)$, $\bar{\beta}(\delta_2/2)=(\bar{R}^2/M)(2+\sqrt{2\ln(2/\delta_2)})^2$, where $\bar{R}$ is evaluated from the empirical data itself:

$$\bar{R} = \frac{\hat{R}}{\left(1-(\hat{R}^2+2)\frac{2+\sqrt{2\ln(4/\delta)}}{\sqrt{M}}\right)^{1/2}}\ .$$

Proof: Since we assumed that Σ was non-singular, the support of ξ being bounded by a ball of radius R_ξ implies that ζ is also bounded. Thus, there exists an R such that $\mathbb{P}(\|\zeta\|\leq R)=1$. Given that ζ has a bounded support, Theorem 4 guarantees us that with probability greater than $1-\delta_1$, constraints (12a), (12b) and (12c) are met. Thus

$$\begin{aligned}R &= \sup_{\zeta\in\mathcal{S}_\zeta}\|\zeta\|_2 = \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\mu)\|_2 = \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\mu+\hat{\mu}-\hat{\mu})\|_2 \\ &\leq \sup_{\xi\in\mathcal{S}_\xi}\|\Sigma^{-1/2}(\xi-\hat{\mu})\|_2+\|\Sigma^{-1/2}(\hat{\mu}-\mu)\|_2 \\ &\leq \sup_{\xi\in\mathcal{S}_\xi}\sqrt{1+\alpha(\delta_1/4)}\|\hat{\Sigma}^{-1/2}(\xi-\hat{\mu})\|_2+\sqrt{\beta(\delta_1/2)} \\ &\leq \sqrt{1+\alpha(\delta_1/4)}\hat{R}+\sqrt{\beta(\delta_1/2)} \\ &\leq \hat{R}\sqrt{1+cR^2}+cR\ ,\end{aligned}$$

where $c = (2 + \sqrt{2\ln(4/\delta_1)})/\sqrt{M}$.

A careful analysis of the function $\psi(R, \hat{R}) = \hat{R}\sqrt{1 + cR^2} + cR$ leads to the observation that if M satisfies Constraint (13) then the fact that $R \leq \psi(R, \hat{R})$ necessarily implies that $R \leq \bar{R}$. We can therefore conclude that $\mathbb{P}(R \leq \bar{R}) \geq 1 - \delta_1$.

Given the event that $R \leq \bar{R}$ occurs, since

$$\begin{aligned}\alpha(\delta_2/4) &= (R^2/\sqrt{M})\left(\sqrt{1 - m/R^4} + \sqrt{2\ln(4/\delta_2)}\right) \\ &\leq (\bar{R}^2/\sqrt{M})\left(\sqrt{1 - m/\bar{R}^4} + \sqrt{2\ln(4/\delta_2)}\right) = \bar{\alpha}(\delta_2/4)\end{aligned}$$

and since

$$\beta(\delta_2/2) = (R^2/M)(2 + \sqrt{2\ln(2/\delta_2)})^2 \leq (\bar{R}^2/M)(2 + \sqrt{2\ln(2/\delta_2)})^2 = \bar{\beta}(\delta_2/2) \ ,$$

we can conclude with a second application of Theorem 2 that with probability greater than $1 - \delta_2$ the following statements are satisfied:

$$\begin{aligned}&(\hat{\mu} - \mu)\Sigma^{-1}(\hat{\mu} - \mu) \leq \beta(\delta_2/2) \leq \bar{\beta}(\delta_2/2) \ , \\ &\Sigma \preceq \frac{1}{1 - \alpha(\delta/4) - \beta(\delta/2)}\hat{\Sigma} \preceq \frac{1}{1 - \bar{\alpha}(\delta_2/4) - \bar{\beta}(\delta_2/2)}\hat{\Sigma} \ , \\ &\Sigma \succeq \frac{1}{1 - \alpha(\delta/4)}\hat{\Sigma} \succeq \frac{1}{1 - \bar{\alpha}(\delta/4)}\hat{\Sigma} \ .\end{aligned}$$

It follows that Theorem 2 applies with $\bar{\alpha}(\delta_2/4)$ and $\bar{\beta}(\delta_2/4)$ because the probability that the event, $\mathcal{E}$, that constraints (12a), (12b) and (12c) equipped with $\bar{\alpha}(\delta_2/4)$ and $\bar{\beta}(\delta_2/4)$ are met is necessarily greater than $1 - \delta$.

$$\mathbb{P}(\mathcal{E}) \geq \mathbb{P}(\mathcal{E}|R \leq \bar{R})\mathbb{P}(R \leq \bar{R}) \geq (1 - \delta_1)(1 - \delta_2) = 1 - \delta \ . \quad \square$$

4.4. Data-driven MP and DRSP Optimization

In most real world situations where one needs to deal with uncertainty in the parameters, it might not be clear how to define an uncertainty set for the mean and covariance matrix of the random parameters ξ. It is more likely, that one only has in hand a set of independent samples, $\{\xi_i\}_{i=1}^M$, drawn from the underlying distribution and wants to solve a moment problem in order to find interesting upper bounds on a moment of the distribution, $\mathbb{E}_\xi[h(\xi)]$. Or similarly, using the set of independent samples, solve a DRSP model in a way that guarantees with high probability that the solution is robust with respect to the true underlying distribution that generated the samples. In this section, we propose using the set of samples to construct a set of distributions that has high probability of containing the distribution that actually generated the samples of ξ.

We will first use our last result to define, based on the random samples $\{\xi_i\}_{i=1}^M$, such a a set of distribution which is known to contain the distribution of ξ with high probability, given that M is sufficiently large.

DEFINITION 2. Let $\mathcal{D}_2(\mathcal{S}, \{\xi_i\}_{i=1}^M, \delta)$ be the set of distributions of ξ such that

$$\mathbb{P}(\xi \in \mathcal{S}) = 1 \tag{14a}$$

$$(\hat{\mu} - \mathbb{E}[\xi])\hat{\Sigma}^{-1}(\hat{\mu} - \mathbb{E}[\xi]) \leq \frac{\bar{\beta}(\delta_2/2)}{1 - \bar{\alpha}(\delta_2/4) - \bar{\beta}(\delta_2/2)} \tag{14b}$$

$$\mathbb{E}[(\xi - \hat{\mu})(\xi - \hat{\mu})^\intercal] \preceq \frac{1 + \bar{\beta}(\delta_2/2)}{1 - \bar{\alpha}(\delta_2/4) - \bar{\beta}(\delta_2/2)}\hat{\Sigma} \ , \tag{14c}$$

where $\hat{\mu}$ and $\hat{\Sigma}$ are the empirical estimates of the mean $\hat{\mu} = M^{-1}\sum_{i=1}^M \xi_i$ and covariance $\hat{\Sigma} = M^{-1}\sum_{i=1}^M (\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\intercal$ of ξ, and $\bar{\alpha}(\delta_2/4) = O(1/\sqrt{M})$ and $\bar{\beta}(\delta_2/2) = O(1/M)$ are constants defined in Corollary 3.

COROLLARY 4. *Given that M satisfies Constraint* (13) *and that $\{\xi_i\}_{i=1}^M$ are independent identically distributed samples from a distribution which is known to have support on $\mathcal{S}$, then with probability greater than $1-\delta$ over the choice of $\{\xi_i\}_{i=1}^M$ the distribution of ξ lies in the set $\mathcal{D}_2(\mathcal{S}, \{\xi_i\}_{i=1}^M, \delta)$.*

Proof: This result can be derived from Corollary 3. One can show that given any estimates $\hat{\mu}$ and $\hat{\Sigma}$ that satisfy both constraints (12a) and (12b) equipped with $\bar{\alpha}(\delta/4)$ and $\bar{\beta}(\delta_2/2)$, these estimates should also satisfy constraints (14b) and (14c). First,

$$(1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2))(\hat{\mu}-\mu)\hat{\Sigma}^{-1}(\hat{\mu}-\mu) \leq (\hat{\mu}-\mu)\Sigma^{-1}(\hat{\mu}-\mu) \leq \bar{\beta}(\delta_2/2)\ ,$$

where we used the fact that Constraint (12a) implies that $\mathbf{x}^\mathsf{T}\Sigma^{-1}\mathbf{x} \geq (1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2))\mathbf{x}^\mathsf{T}\hat{\Sigma}^{-1}\mathbf{x}$ for any $\mathbf{x}\in\mathbb{R}^m$. Similarly, the estimates $\hat{\mu}$ and $\hat{\Sigma}$ can be shown to satisfy Constraint (14c):

$$\begin{aligned}\frac{1}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\hat{\Sigma} \succeq \Sigma \ &= \ \mathbb{E}[\xi\xi^\mathsf{T}] - \mu\mu^\mathsf{T} \\ &\succeq \mathbb{E}[(\xi-\mu)(\xi-\mu)^\mathsf{T}] - \frac{\bar{\beta}(\delta_2/2)}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\hat{\Sigma}\ ,\end{aligned}$$

since for all $\mathbf{x}\in\mathbb{R}^m$,

$$\begin{aligned}\mathbf{x}^\mathsf{T}\mu\mu^\mathsf{T}\mathbf{x} &= (\mathbf{x}^\mathsf{T}(\mu-\hat{\mu}+\hat{\mu}))^2 = (\mathbf{x}^\mathsf{T}(\mu-\hat{\mu}))^2 + 2\mathbf{x}^\mathsf{T}(\mu-\hat{\mu})\hat{\mu}^\mathsf{T}\mathbf{x} + (\mathbf{x}^\mathsf{T}\hat{\mu})^2 \\ &= \mathbf{trace}(\mathbf{x}^\mathsf{T}\Sigma^{1/2}\Sigma^{-1/2}(\mu-\hat{\mu})(\mu-\hat{\mu})^\mathsf{T}\Sigma^{-1/2}\Sigma^{1/2}\mathbf{x}) + 2\mathbf{x}^\mathsf{T}\mu\hat{\mu}^\mathsf{T}\mathbf{x} - (\mathbf{x}^\mathsf{T}\hat{\mu})^2 \\ &\leq (\mu-\hat{\mu})^\mathsf{T}\Sigma^{-1}(\mu-\hat{\mu})\mathbf{x}^\mathsf{T}\Sigma\mathbf{x} + 2\mathbf{x}^\mathsf{T}\mu\hat{\mu}^\mathsf{T}\mathbf{x} - (\mathbf{x}^\mathsf{T}\hat{\mu})^2 \\ &\leq \mathbf{x}^\mathsf{T}\left(\frac{\bar{\beta}(\delta_2/2)}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\hat{\Sigma} + \mu\hat{\mu}^\mathsf{T} + \hat{\mu}\mu^\mathsf{T} - \hat{\mu}\hat{\mu}^\mathsf{T}\right)\mathbf{x} \\ &= \mathbf{x}^\mathsf{T}\left(\frac{\bar{\beta}(\delta_2/2)}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\hat{\Sigma} + \mathbb{E}[\xi\xi^\mathsf{T}] - \mathbb{E}[(\xi-\mu)(\xi-\mu)^\mathsf{T}]\right)\mathbf{x}\ ,\end{aligned}$$

By Corollary 3, the random variables $\hat{\mu}$ and $\hat{\Sigma}$ are guaranteed to satisfy constraints (12a) and (12b) with probability greater than $1-\delta$, therefore they must also satisfy constraints (14b) and (14c) with probability greater than $1-\delta$. □

We can now extend the results presented in sections 3 to a data-driven framework where moments of the distribution are estimated from data.

DEFINITION 3. Let $\Phi(\mathbf{x}; \mathcal{S}, \{\xi_i\}_{i=1}^M, \delta)$ be the optimal value of the MP model associated to the inner problem of Problem (5):

$$\underset{f_\xi \in \mathcal{D}_1(\mathcal{S},\hat{\mu},\hat{\Sigma},\gamma_1,\gamma_2)}{\text{maximize}} \quad \mathbb{E}_\xi[h(\mathbf{x},\xi)]\ ,$$

using the assignments:

$$\gamma_1 = \frac{\bar{\beta}(\delta_2/2)}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\ , \qquad \gamma_2 = \frac{1+\bar{\beta}(\delta_2/2)}{1-\bar{\alpha}(\delta_2/4)-\bar{\beta}(\delta_2/2)}\ . \tag{15}$$

Based on the computational argument of Proposition 2 and the probabilistic guarantees provided by Corollary 4, we present an important result for data-driven problems.

THEOREM 3. *Let δ be a risk factor and $\{\xi_i\}_{i=1}^M$ be a set of M independently and identically distributed instances of the random parameters ξ, which distribution is known to have support on a bounded set $\mathcal{S}$. One can solve in polynomial time the DRSP equipped with the distributional set $\mathcal{D}_1(\mathcal{S},\hat{\mu},\hat{\Sigma},\gamma_1,\gamma_2)$, where γ_1 and γ_2 are defined as in Equation* (15)*. Then, given that M satisfies Constraint* (13)*, one is guaranteed that the resulting optimal solution $\mathbf{x}^*$ is such that*

$$\mathbb{P}_{\{\xi\}_{i=1}^M}(\mathbb{E}_\xi[h(\mathbf{x}^*,\xi)] \leq \Phi(\mathbf{x}^*; \mathcal{S}, \{\xi_i\}_{i=1}^M, \delta)) \geq 1-\delta\ ,$$

where $\mathbb{P}(\cdot)$ is evaluated with respect to the random generation of the bound $\Phi(\mathbf{x}^*; \mathcal{S}, \{\xi_i\}_{i=1}^M, \delta)$, while the expectation is taken with respect to the underlying distribution of ξ.

Since we believe the MP model to be interesting in its own right, we find important to mention a simple consequence of the above result for data-driven moment problems when considering the special case $h(\mathbf{x}, \xi) = h(\xi)$.

COROLLARY 5. *Let δ be a risk factor and $\{\xi_i\}_{i=1}^M$ be a set of M independently and identically distributed instances of the random parameters ξ, which distribution is known to have support on a bounded set $\mathcal{S}$. One can compute in polynomial time the MP equipped with the distributional set $\mathcal{D}_1(\mathcal{S}, \hat{\mu}, \hat{\Sigma}, \gamma_1, \gamma_2)$ where γ_1, γ_2 are defined as in Equation 15. Then, given that M satisfies Constraint* (13)*, one is assured that the optimal value previously defined as $\Psi(\gamma_1, \gamma_2)$ is such that*

$$\mathbb{P}_{\{\xi_i\}_{i=1}^M}(\mathbb{E}_\xi[h(\xi)] \leq \Psi(\gamma_1, \gamma_2)) \geq 1 - \delta \ ,$$

where $\mathbb{P}(\cdot)$ is evaluated with respect to the random generation of the bound $\Psi(\gamma_1, \gamma_2)$ using the random samples $\{\xi_i\}_{i=1}^M$, while the expectation is taken over the exact underlying distribution for ξ.

5. Application to Robust Portfolio Optimization

We now turn ourselves to applying our framework to an instance of portfolio optimization. In such a problem, one is interested in maximizing his expected utility for the potential one step return of an investment portfolio. Given that n investment options are available, return can be defined as the linear function $\xi^\mathsf{T}\mathbf{x}$, where $\xi \in \mathbb{R}^n$ is a random vector of return for the different options. In the robust approach to this problem, one defines a distributional set $\mathcal{D}$ that is known to contain the distribution f_ξ and choose the portfolio which is optimal according to the following DRSP model:

$$\underset{\mathbf{x}}{\text{maximize}} \quad \min_{f_\xi \in \mathcal{D}} \ \mathbb{E}_\xi[u(\xi^\mathsf{T}\mathbf{x})] \tag{16a}$$

$$\text{subject to} \quad \sum_{i=1}^n \mathbf{x}_i = 1 \ , \ \ \mathbf{x} \geq 0 \ . \tag{16b}$$

In Popescu (2007), the author recently addressed this problem in the case where $\mathbb{E}[\xi]$ and $\mathbb{E}[\xi\xi^\mathsf{T}]$ are known exactly and one considers $\mathcal{D}$ to be the set of all distribution with such first and second moments. With these assumptions, she presents a parametric quadratic programming algorithm that is efficient for a large family of utility function $u(\cdot)$. This approach is interesting as it provides a mean of taking into account uncertainty in the form of the return distribution. Unfortunately, our experiments will show that in practice it is highly sensitive to the noise in the empirical estimation of these moments. Secondly, the algorithm also relies on solving a one dimensional non-convex mathematical program. Thus, Popescu does not provide an algorithm which is guaranteed to converge to an optimal solution in polynomial time. Although the approach that we are about to propose addresses a smaller family of utility functions, it will take into account moment uncertainty and will lead to the formulation of a semi-definite program which can be solved efficiently using interior point methods.

5.1. Portfolio Optimization with Moment Uncertainty

In order to apply our framework we need to assume that the utility function is piecewise linear concave such that $u(y) = \min_{k \in \{1,2,\ldots,K\}} \ a_k y + b_k$. This is not too constraining since in portfolio optimization the interesting utility functions are usually concave and such functions always have a finite (in K) piecewise linear approximation. We will use historical knowledge of investment returns $\{\xi_1, \xi_2, ..., \xi_M\}$ to define a distributional uncertainty set for f_ξ. This can be done either through Definition 2 or directly through the set $\mathcal{D}_1(\gamma_1, \gamma_2)$ for suitably chosen values of γ_1 and γ_2. For simplicity, in what follows we use $\mathcal{D}_1(\mathcal{S}_\xi, \hat{\mu}, \hat{\Sigma}, \gamma_1, \gamma_2)$. The parameters $\hat{\mu}$ and $\hat{\Sigma}$ are assigned as the empirical estimates of the mean $\hat{\mu} = M^{-1}\sum_{i=1}^M \xi_i$ and covariance $\hat{\Sigma} = M^{-1}\sum_{i=1}^M (\xi_i - \hat{\mu})(\xi_i - \hat{\mu})^\mathsf{T}$ of ξ respectively. On the other hand, $\mathcal{S}$ can either be $\mathbb{R}^n$ or an ellipsoidal set $\{\xi | (\xi - \xi_0)^\mathsf{T}\mathbf{\Theta}(\xi - \xi_0) \leq 1\}$ known to contain the support of ξ.[1]

Building on the results presented in Section 3, one can make the following statement about the tractability of distributionally robust portfolio problems.

THEOREM 4. *Given that $u(\cdot)$ is piecewise linear, finding an optimal solution $\mathbf{x} \in \mathbb{R}^n$ to the distributionally robust portfolio Problem* (16) *equipped with the set of distributions $\mathcal{D}_1(\mathcal{S}_\xi, \hat{\mu}, \hat{\Sigma}, \gamma_1, \gamma_2)$ can be done in $O(n^{6.5})$.*

Proof: We first reformulate the objective of Problem (16) in its minimization form :

$$\underset{\mathbf{x}\in\mathcal{X}}{\text{minimize}} \quad \left(\max_{f_\xi \,\in\, \mathcal{D}_1(\mathcal{S}_\xi,\hat{\mu},\hat{\Sigma},\gamma_1,\gamma_2)} \mathbb{E}_\xi[\max_k \; -a_k\xi^\mathsf{T}\mathbf{x} - b_k] \right) \quad .$$

After confirming that $h(\mathbf{x},\xi) = \max_k \; -a_k\xi^\mathsf{T}\mathbf{x} - b_k$ satisfies the conditions in Assumption 2, a straightforward application of Proposition 2 already confirms that Problem (16) can be solved in polynomial time. In order to get a more precise computational bound, one needs to take a closer look at the dual formulation of Problem (16):

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s}{\text{minimize}} \quad & \gamma_2(\Sigma_0 \bullet \mathbf{Q}) - \mu_0^\mathsf{T}\mathbf{Q}\mu_0 + r + (\Sigma_0 \bullet \mathbf{P}) - 2\mu_0^\mathsf{T}\mathbf{p} + \gamma_1 s && \text{(17a)}\\
\text{subject to} \quad & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & s \end{bmatrix} \geq 0 \;\; , \;\; \mathbf{p} = -\mathbf{q}/2 - \mathbf{Q}\hat{\mu} \; , && \text{(17b)}\\
& \xi^\mathsf{T}\mathbf{Q}\xi + \xi^\mathsf{T}\mathbf{q} + r \geq -a_k\xi^\mathsf{T}\mathbf{x} - b_k \; , \; \forall \xi \in \mathcal{S}_\xi, k \in \{1,2,...,K\} && \text{(17c)}\\
& \sum_{i=1}^n \mathbf{x}_i = 1 \;\; , \;\; \mathbf{x}_i \geq 0 \;\; , \; \forall i \;\; . && \text{(17d)}
\end{aligned}$$

Given that $\mathcal{S}_\xi = \mathbb{R}^n$, one can use Schur's complement to replace Constraint (17c) by an equivalent linear matrix inequality.

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s}{\text{minimize}} \quad & \gamma_2(\hat{\Sigma} \bullet \mathbf{Q}) - \hat{\mu}^\mathsf{T}\mathbf{Q}\hat{\mu} + r + (\hat{\Sigma} \bullet \mathbf{P}) - 2\hat{\mu}^\mathsf{T}\mathbf{p} + \gamma_1 s\\
\text{subject to} \quad & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & s \end{bmatrix} \geq 0 \;\; , \;\; \mathbf{p} = -\mathbf{q}/2 - \mathbf{Q}\hat{\mu}\\
& \begin{bmatrix} \mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2 \\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k \end{bmatrix} \geq 0 \; , \; \forall k\\
& \sum_{i=1}^n \mathbf{x}_i = 1 \;\; , \;\; \mathbf{x}_i \geq 0 \;\; , \; \forall i \;\; .
\end{aligned}$$

While if $\mathcal{S}_\xi$ is an ellipsoid, the S-lemma can be used to replace Constraint (17c)

$$\begin{bmatrix} \xi \\ 1 \end{bmatrix}^\mathsf{T} \begin{bmatrix} \mathbf{\Theta} & -\mathbf{\Theta}\xi_0 \\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1 \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix} \leq 0 \rightarrow \begin{bmatrix} \xi \\ 1 \end{bmatrix}^\mathsf{T} \begin{bmatrix} \mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2 \\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k \end{bmatrix} \begin{bmatrix} \xi \\ 1 \end{bmatrix} \geq 0 \;\; ,$$

with an equivalent constraint:

$$\begin{bmatrix} \mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2 \\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k \end{bmatrix} \geq -\tau_k \begin{bmatrix} \mathbf{\Theta} & -\mathbf{\Theta}\xi_0 \\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1 \end{bmatrix} \;\; , \;\; \tau_k \geq 0 \;\; ,$$

where τ_k, $k \in \{1,...,K\}$, are extra slack variables. The problem can therefore also be reformulated as a semi-definite program:

$$\begin{aligned}
\underset{\mathbf{x},\mathbf{Q},\mathbf{q},r,\mathbf{P},\mathbf{p},s,\tau}{\text{minimize}} \quad & \gamma_2(\hat{\Sigma} \bullet \mathbf{Q}) - \hat{\mu}^\mathsf{T}\mathbf{Q}\hat{\mu} + r + (\hat{\Sigma} \bullet \mathbf{P}) - 2\hat{\mu}^\mathsf{T}\mathbf{p} + \gamma_1 s\\
\text{subject to} \quad & \begin{bmatrix} \mathbf{P} & \mathbf{p} \\ \mathbf{p}^\mathsf{T} & s \end{bmatrix} \geq 0 \;\; , \;\; \mathbf{p} = -\mathbf{q}/2 - \mathbf{Q}\hat{\mu}\\
& \begin{bmatrix} \mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2 \\ \mathbf{q}^\mathsf{T}/2 + a_k\mathbf{x}^\mathsf{T}/2 & r + b_k \end{bmatrix} \geq -\tau_k \begin{bmatrix} \mathbf{\Theta} & -\mathbf{\Theta}\xi_0 \\ -\xi_0^\mathsf{T}\mathbf{\Theta} & \xi_0^\mathsf{T}\mathbf{\Theta}\xi_0 - 1 \end{bmatrix} \; , \; \forall k\\
& \tau_k \geq 0 \;\; \forall k\\
& \mathbf{Q} \geq 0\\
& \sum_{i=1}^n \mathbf{x}_i = 1 \;\; , \;\; \mathbf{x}_i \geq 0 \;\; , \; \forall i \;\; .
\end{aligned}$$

In both cases, the optimization problem that needs to be solved is a semi-definite program. It is well known that an interior point algorithm can be used to solve an SDP of the form

$$\begin{array}{ll} \underset{\mathbf{x}\in\mathbb{R}^{\tilde{n}}}{\text{minimize}} & c^{\mathsf{T}}\mathbf{x} \\ \text{subject to} & A_i(\mathbf{x}) \succeq 0 \ \ \forall i = 1,2,...,\tilde{K} \end{array}$$

in $O\left(\left(\sum_i^{\tilde{K}} \tilde{m}_i\right)^{0.5}\left(\tilde{n}^2\sum_i^{\tilde{K}}\tilde{m}_i^2 + \tilde{n}\sum_i^{\tilde{K}}\tilde{m}_i^3\right)\right)$, where $\tilde{m}_i$ stands for the dimension of the positive semi-definite cone (*i.e.*, $A_i(\mathbf{x}) \in \mathbb{R}^{\tilde{m}_i\times\tilde{m}_i}$) (see Nesterov and Nemirovski (1994)). In both SDP that interests us here, one can show that $\tilde{n} \leq n^2+4n+2+K$ and that the problem can be solved in $O(K^{3.5}n^{6.5})$ operations, with K being the number of pieces in the utility function $u(\cdot)$. We conclude that the portfolio optimization problem can be solved in $O(n^{6.5})$. $\square$

REMARK 3. The computational complexity presented here is based on general theory for solving semi-definite programs. Based on an implementation that uses SeDuMi (Sturm (1999)), we actually observed empirically that complexity grows in the order of $O(n^5)$ for dense problems. In practice, one may also be able to exploit structure in problems where subsets of assets are known to behave independently from each other.

5.2. A Case where the Worse Distribution has Largest Covariance Matrix

When presenting our distributionally robust framework, we generally argued in Remark 1 that lower bounds on covariance were uninteresting. We are now interested in presenting a more rigorous argument for this modeling decision. Actually, in the case of a portfolio optimization problem with piecewise concave utility function, we can show that the lower bound on the covariance matrix is irrelevant. The proof of the following proposition also provides valuable insight on the structure of portfolio optimization problems.

PROPOSITION 3. *The robust portfolio optimization problem with piecewise linear concave utility and infinite support constraint on the distribution is an instance of distributionally robust optimization where the covariance constraint is tight for the worse case distribution.*

Proof: Consider the inner problem of our robust portfolio optimization with unconstrained support for the distribution:

$$\max_{f_\xi \,\in\, \mathcal{D}_1(\mathbb{R}^m,\hat{\mu},\hat{\Sigma},0,\gamma_2)} \mathbb{E}_\xi[\max_k \ -a_k\xi^{\mathsf{T}}\mathbf{x} - b_k] \ . \tag{18}$$

For simplicity of our derivations, we consider that there is no uncertainty in the mean of the distribution (*i.e.*, $\gamma_1 = 0$). The dual of this problem can be formulated as:

$$\begin{array}{ll} \underset{\mathbf{Q},\mathbf{q},r}{\text{minimize}} & (\hat{\Sigma}\bullet\mathbf{Q}) + \hat{\mu}^{\mathsf{T}}\mathbf{Q}\hat{\mu} + \hat{\mu}^{\mathsf{T}}\mathbf{q} + r \\ \text{subject to} & \begin{bmatrix} \mathbf{Q} & \mathbf{q}/2 + a_k\mathbf{x}/2 \\ \mathbf{q}^{\mathsf{T}}/2 + a_k\mathbf{x}^{\mathsf{T}}/2 & r + b_k \end{bmatrix} \geq 0 \ , \ \forall k \ . \\ & \mathbf{Q} \geq 0 \end{array}$$

Applying duality theory a second time leads to formulating a slightly different version of the primal problem which by strong duality achieves the same optimum.

$$\underset{\{(\Lambda_k,\lambda_k,\nu_k)\}_{k=1}^K}{\text{maximize}} \quad \sum_{k=1}^K a_k\mathbf{x}^{\mathsf{T}}\lambda_k + \nu_k b_k \tag{19a}$$

$$\text{subject to} \quad \sum_{k=1}^K \Lambda_k \preceq \gamma_2\hat{\Sigma} + \hat{\mu}\hat{\mu}^{\mathsf{T}} \tag{19b}$$

$$\sum_{k=1}^{K} \lambda_k = \hat{\mu} \ , \ \sum_{k=1}^{K} \nu_k = 1 \tag{19c}$$

$$\begin{bmatrix} \Lambda_k & \lambda_k \\ \lambda_k^T & \nu_k \end{bmatrix} \succeq 0 \quad \forall\, k \in \{1,2,...,K\} \ . \tag{19d}$$

We can show that there always exists an optimal solution such that Constraint (19b) is satisfied with equality. Given an optimal assignment $X^* = \{(\Lambda_k^*, \lambda_k^*, \nu_k^*)\}_{k=1}^K$ such that $\Delta = \gamma_2\hat{\Sigma} + \hat{\mu}\hat{\mu}^T - \sum_{k=1}^K \Lambda_k^* \succeq 0$, consider an alternate solution $X' = \{(\Lambda_k', \lambda_k', \nu_k')\}_{k=1}^K$ which is exactly the same as the original solution X^* except for $\Lambda_1' = \Lambda_1^* + \Delta$. Obviously the two solutions achieve the same objective values since the variables $\{(\lambda_k, \nu_k)\}_{k=1}^K$ were not modified. If we can show that X' is also feasible then it is necessarily optimal. The only feasibility constraint that seriously needs to be verified is the following:

$$\begin{bmatrix} \Lambda_1' & \lambda_1' \\ \lambda_1'^T & \nu_1' \end{bmatrix} = \begin{bmatrix} \Lambda_1^* & \lambda_1^* \\ \lambda_1^{*T} & \nu_1^* \end{bmatrix} + \begin{bmatrix} \Delta & 0 \\ 0 & 0 \end{bmatrix} \succeq 0 \ ,$$

and is necessarily satisfied since by definition X^* is feasible and that by construction Δ is positive semi-definite. It is therefore the case that there exists a solution X^* that is optimal with respect to Problem (19) and satisfies Constraint (19b) with equality. Furthermore, one is assured that the sum $\sum_{k=1}^K a_k \mathbf{x}^T \lambda_k^* + \nu_k^* b_k$ is equal to the optimal value of Problem (18).

After assuming without loss of generality that all $\nu_k > 0$, let us now construct K random vectors $(\zeta_1, \zeta_2, ..., \zeta_K)$ that satisfy the following conditions:

$$\mathbb{E}[\zeta_k] = \frac{1}{\nu_k}\lambda_k^* \ , \ \mathbb{E}[\zeta_k\zeta_k^T] = \frac{1}{\nu_k}\Lambda_k^* \ .$$

Note that since X^* satisfies Constraint (19d), we are assured that:

$$\begin{aligned}
\mathbb{E}[\zeta_k\zeta_k^T] - \mathbb{E}[\zeta_k]\mathbb{E}[\zeta_k]^T &= \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix}^T \begin{bmatrix} \mathbb{E}[\zeta_k\zeta_k^T] & \mathbb{E}[\zeta_k] \\ \mathbb{E}[\zeta_k]^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix} \\
&= \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix}^T \begin{bmatrix} \frac{1}{\nu_k}\Lambda_k^* & \frac{1}{\nu_k}\lambda_k^* \\ \frac{1}{\nu_k}\lambda_k^{*T} & 1 \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix} \\
&= \frac{1}{\nu_k^*} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix}^T \begin{bmatrix} \Lambda_k^* & \lambda_k^* \\ \lambda_k^{*T} & \nu_k^* \end{bmatrix} \begin{bmatrix} \mathbf{I} \\ -\mathbb{E}[\zeta_k]^T \end{bmatrix} \ \succeq \ 0 \ .
\end{aligned}$$

Hence, the random vectors $(\zeta_1, \zeta_2, ..., \zeta_K)$ exists. For instance, if $\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])(\zeta_k - \mathbb{E}[\zeta_k])^T] \succ 0$, then ζ_k can take the form of a multivariate Gaussian distribution with such mean and covariance. Otherwise, one can construct a lower dimensional random vector; for instance, if $\mathbb{E}[(\zeta_k - \mathbb{E}[\zeta_k])(\zeta_k - \mathbb{E}[\zeta_k])^T] = 0$ then the random vector should be a deterministic point at $\mathbb{E}[\zeta_k]$.

Let $\tilde{k}$ be an independent multinomial with parameters $(\nu_1, \nu_2, ..., \nu_K)$, such that $\mathbb{P}(\tilde{k} = i) = \nu_i$, and use it to construct the random vector $\xi = \zeta_{\tilde{k}}$. Since X^* satisfies constraints (19b) and (19c) tightly, one can show that the distribution function of ξ^* lies in $\mathcal{D}(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, \gamma_2)$ and has largest covariance.

$$\begin{aligned}
\mathbb{E}[\xi^*] &= \sum_{k=1}^{K} \mathbb{E}[\zeta_k|\tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^{K} \frac{1}{\nu_k}\lambda_k\nu_k = \hat{\mu} \\
\mathbb{E}[\xi^*\xi^{*T}] &= \sum_{k=1}^{K} \mathbb{E}[\zeta_k\zeta_k^T|\tilde{k} = k]\mathbb{P}(\tilde{k} = l) = \sum_{k=1}^{K} \frac{1}{\nu_k}\Lambda_k\nu_k = \gamma_2\hat{\Sigma} + \hat{\mu}\hat{\mu}^T
\end{aligned}$$

Moreover, when used as a candidate worse case distribution in Problem (18) it actually achieves the maximum since we can show it must be greater or equal to it.

$$\mathbb{E}\left[\max_l \ -a_l\mathbf{x}^T\xi^* - b_l\right] = \sum_{k=1}^{K} \mathbb{E}\left[\max_l \ -a_l\mathbf{x}^T\zeta_{\tilde{k}} - b_l \middle| \tilde{k} = k\right] \mathbb{P}(\tilde{k} = k)$$

$$
\begin{aligned}
&\geq \sum_{k=1}^{K} \mathbb{E}[-a_k \mathbf{x}^\mathsf{T} \zeta_k - b_k] \mathbb{P}(\tilde{k} = k) \\
&= \sum_{k=1}^{K} -a_k \mathbf{x}^\mathsf{T} \lambda_k - b_k \nu_k \\
&= \max_{f_\xi \in \mathcal{D}_1(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, \gamma_2)} \mathbb{E}_\xi[\max_k \ -a_k \mathbf{x}^\mathsf{T} \xi - b_k]
\end{aligned}
$$

We conclude that we just constructed a worse case distribution that does have largest covariance. □

REMARK 4. An interesting consequence of Proposition 3 is that in the framework considered in Popescu (2007), if the utility function is piecewise concave, one can find the optimal portfolio in polynomial time using our semi-definite programming formulation with the distributional set $\mathcal{D}_1(\mathbb{R}^m, \hat{\mu}, \hat{\Sigma}, 0, 1)$. We believe our semi-definite program formulation to be more tractable than the original algorithm. However, it is true that our framework does not provide a polynomial time algorithm for the larger range of utility functions considered in Popescu's work.

5.3. Experiments

We evaluate our portfolio optimization method on stock market investments. We use a historical dataset of 30 assets over a horizon of 15 years (1992-2007), obtained from the Yahoo! Finance.[2] Each experiment consists of randomly choosing 4 assets, and building a dynamic portfolio with these assets through the years 2001-2007. At any given day of the experiment, the algorithms are limited to 30 days of the most recent history to assign the portfolio. All methods assume that in this period the samples are independent and identically distributed. Note that 30 samples of data is not much to generate good empirical estimates of the mean and covariance of returns; however, the use of a larger history causes the i.i.d. assumption to be somewhat unrealistic.

In implementing our method, the distributional set is formulated as $\mathcal{D}_1(\mathbb{R}^4, \hat{\mu}, \hat{\Sigma}, 1.35, 8.32)$, where $\hat{\mu}$ and $\hat{\Sigma}$ are the empirical estimates of the mean and covariance of ξ respectively. The values for γ_1 and γ_2 are chosen based on a simple statistical analysis of moment estimation during the years 1997-2001.[3] We compare our approach to the one proposed by Popescu (2007), where the mean and covariance of the distribution f_ξ is assumed to be equal to the empirical estimates over the 30 days history. The method is also compared to a naive approximation of the stochastic program in which the selected portfolio is the one that maximizes the average utility over the last 30 days samples. We believe that the statistics obtained over the set of 300 experiments demonstrate how much there is to gain in terms of performance and risk reduction by considering an optimization model that accounts for both distribution and moment uncertainty.

Method	Single Day		2001-2004		2004-2007	
	Avg. utility	1-perc.	Avg. yearly return	10-perc.	Avg. yearly return	10-perc.
DRPO model	1.000	0.983	0.944	0.846	1.1017	1.025
Popescu's DRPO model	1.000	0.975	0.700	0.334	1.047	0.9364
SP model	1.000	0.973	0.908	0.694	1.045	0.923

First, from the analysis of the daily returns generated by each method, one observes that they achieve comparable average daily utility. However, the DRSP model stands out as being more reliable. For instance, the lower 1%-percentile of the utility distribution is 0.8% higher then the two competing methods. Also, this difference in reliability becomes more obvious when considering the respective long term performances. Figure 1 presents the average evolution of wealth on a six years period when managing a portfolio of 4 assets on a daily basis with either of the three methods. The performances over the years 2001-2004 are presented separately from the performances over the years 2004-2007 in order to measure how they are affected by different level of economic growth. The figures also indicate periodically the 10% and 90% percentile of the

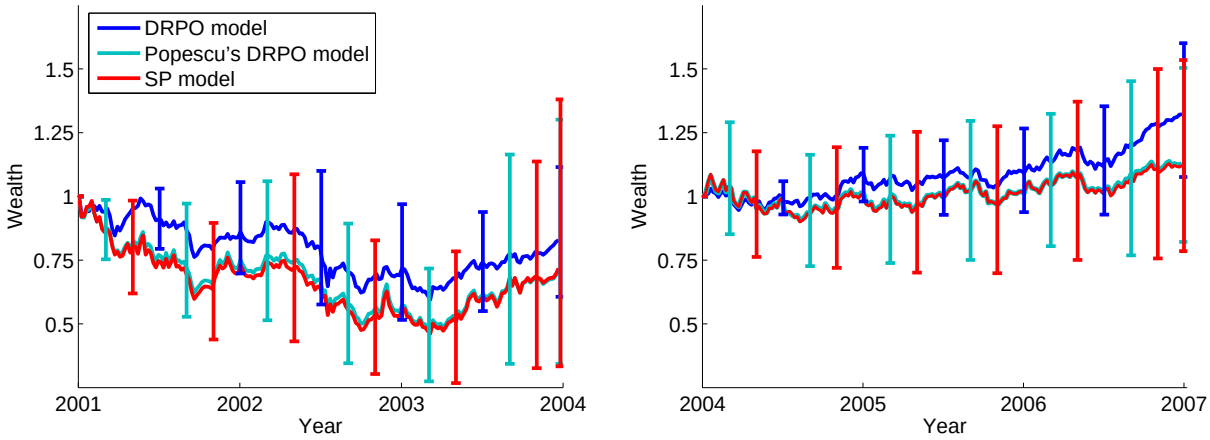

Figure 1 Comparison of wealth evolution in 300 experiments conducted over the years 2001-2007 using three different portfolio optimization models. For each model, the figures indicate periodically the 10% and 90% percentile of the wealth distribution in the set of experiments.

wealth distribution over the set of 300 experiments. The statistics of the long term experiments demonstrate empirically that our method significantly outperforms the two more naive ones in terms of average return and risks during both the years of economic growth and the years of decline. More specifically, the DRPO model outperformed the Popescu's model in terms of total return cumulated over the period 2001-2007 in 79.2% of our experiments (total set of 300 experiments). Also, it performed on average at least 1.67 times better than any competing models. Note that these experiments are purely illustrative of the strengths and weaknesses of the different models. For instance, the returns obtained in each experiment does not take into account transaction fees. The data is also biased by the fact that the assets involved in our experiments were known to be major assets in their category in January 2007. On the other hand, the return were also negatively biased by the fact that in each experiment the models were managing a portfolio of only four correlated assets. Overall we believe that these biases were affecting all methods in a similar manner.

Notes

[1]One should also verify that $\hat{\mu} \in \mathbf{int}(\mathcal{S})$ and that $\hat{\Sigma} \succ 0$ in order to meet the technical conditions required through the application of duality theory.

[2]The list of assets that is used in our experiments was inspired by Goldfarb and Iyengar (2003). More specifically, the 30 assets are: AAR Corp., Boeing Corp., Lockheed Martin, United Technologies, Intel Corp., Hitachi, Texas Instruments, Dell Computer Corp., Palm Inc., Hewlett Packard, IBM Corp., Sun Microsystems, Bristol-Myers-Squibb, Applera Corp.-Celera Group, Eli Lilly and Co., Merck and Co., Avery Denison Corp., Du Pont, Dow Chemical, Eastman Chemical Co., AT&T, Nokia, Motorola, Ariba, Commerce One Inc., Microsoft, Oracle, Akamai, Cisco Systems, Northern Telecom, Duke Energy Company, Exelon Corp., Pinnacle West, FMC Corp., General Electric, 'Honeywell, Ingersoll Rand.

[3]More specifically, given that one chooses 4 stocks randomly and samples a random period of 60 days between 1997 and 2001, the values for γ_1 and γ_2 are chosen such that when using the first 30 days of the period to center $\mathcal{D}(\gamma_1, \gamma_2)$, the distributional set contains, with 99% probability, distributions with moments equal to the moments estimated from the last 30 days of the period.

Acknowledgments

The authors acknowledge the Fonds Québécois de la recherche sur la nature et les technologies and Boeing for their financial support and thank Amir Dembo and Benjamin Armbruster for helpful discussions.

References

Anderson, T. W. 1984. *An Introduction to Multivariate Analysis*. John Wiley & Sons, New York, NY, USA.

Barmish, B. R., C. M. Lagoa. 1997. The uniform distribution: A rigorous justification for its use in robustness analysis. *Mathematics of Control Signals and Systems* **10**(3) 203–222.

Bertsimas, D., I. Popescu. 2005. Optimal inequalities in probability theory: A convex optimization approach. *SIAM Journal on Optimization* **15**(3) 780–804.

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

Chen, X., M. Sim, P. Su. 2007. A robust optimization perspective on stochastic programming. *Operations Research* **55**(6) 1058–1071.

Dupacová, J. 1980. On minimax decision rule in stochastic linear programming. A. Prekopa, ed., *Studies on Mathematical Programming*. Akademiai Kiado, 47–60.

Edelman, A. 1989. Eigenvalues and condition numbers of random matrices. Ph.D. thesis, MIT, Boston, MA, USA.

Ermoliev, Y., A. Gaivoronski, C. Nedeva. 1985. Stochastic optimization problems with partially known distribution functions. *Journal on Control and Optimization* **23** 696–716.

Fujikoshi, Y. 1980. Asymptotic expansions for the distributions of the sample roots under nonnormality. *Biometrika* **67**(1) 45–51.

Goldfarb, D., G. Iyengar. 2003. Robust portfolio selection problems. *Mathematics of Operations Research* **28**(1) 1–38.

Marshall, A., I. Olkin. 1960. Multivariate Chebyshev inequalities. *Annals of Mathematical Statistics* **31** 1001–1024.

McDiarmid, C. 1998. Concentration. M. Habib, C. McDiarmid, J. Ramirez-Alfonsin, B. Reed, eds., *Probabilistic Methods for Algorithmic Discrete Mathematics*. Springer, 195–248.

Nesterov, Y., A. Nemirovski. 1994. *Interior-point polynomial methods in convex programming*. SIAM, Philadelphia, PA, USA.

Popescu, I. 2007. Robust mean-covariance solutions for stochastic optimization. *Operations Research* **55**(1) 98–112.

Prékopa, A. 1995. *Stochastic Programming*. Kluwer Academic Publishers.

Rockafellar, R.T., S. Uryasev. 2000. Optimization of conditional value-at-risk. *Journal of Risk* **2**(3) 21–41.

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

Schrader, R. 1983. The ellipsoid method and its implications. *OR Spectrum* **5**(1) 1–13.

Shapiro, A. 2000. Stochastic programming by monte carlo simulation methods. Stochastic Programming E-Print Series.

Shapiro, A. 2001. On duality theory of conic linear problems. M. A. Goberna, M. A. López, eds., *Semi-Infinite Programming: Recent Advances*. Kluwer Academic Publishers, 135–165.

Shapiro, A., A.J. Kleywegt. 2002. Minimax analysis of stochastic problems. *Optimization Methods and Software* **17** 523–542.

Shawe-Taylor, J., N. Cristianini. 2003. Estimating the moments of a random vector with applications. J. Siemons, ed., *Proceedings of GRETSI 2003 Conference*. Cambridge University Press, 47–52.

Sturm, J.F. 1999. Using SeDuMi 1.02, a MATLAB toolbox for optimization over symmetric cones. *Optimization Methods and Software* **11–12** 625–653.

Waternaux, C. 1976. Asymptotic distribution of the sample roots for the nonnormal population. *Biometrika* **63**(3) 639–645.

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

Zhu, Z., J. Zhang, Y. Ye. 2006. Newsvendor optimization with limited distribution information. Technical Report, Stanford University.