Models and Algorithms for Distributionally Robust Least Squares Problems

Sanjay Mehrotra[*] and He Zhang[†]

February 12, 2011

Abstract

We present different robust frameworks using probabilistic ambiguity descriptions of the input data in the least squares problems. The three probability ambiguity descriptions are given by: (1) confidence interval over the first two moments; (2) bounds on the probability measure with moments constraints; (3) confidence interval over the probability measure by using the Kantorovich probability distance. For these three cases we derive equivalent formulations and show that the resulting optimization problem can be solved efficiently.

1 Introduction

The ordinary least squares (OLS) problem [7] is a fundamental problem with numerous applications. The OLS problem is defined as

$$\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x} - \mathbf{b}||^2 . \tag{1.1}$$

In OLS (1.1) the data $\mathbf{A}$ and $\mathbf{b}$ are considered known. In many practical situations, however, the parameters $\mathbf{A}$ and $\mathbf{b}$ have errors. Such situations arise, for example, when data are collected in physical experiments. Repeated observations under same experimental conditions do not generate the same output [11], and the error in the estimation of $\mathbf{A}$ and $\mathbf{b}$ is random. The goal of this paper is to consider robust formulations of such a problem, and propose algorithms for solving them. One possible approach is to consider the stochastic least square problem with the form:

$$\min_{\mathbf{x}} \{\mathbb{E}_P[||\mathbf{A}\mathbf{x} - \mathbf{b}||^2]\}, \tag{1.2}$$

where P is a given known distribution followed by the observation data $\mathbf{A}$ and $\mathbf{b}$. Models of this type have been considered before (for example, Rao [14]). In practice, however, we may not have full knowledge of this probability distribution. One possible way to handle

[*] Dept. of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60201

[†] Dept. of Industrial Engineering & Management Sciences, Northwestern University, Evanston, IL 60201

the data uncertainty is to use robust optimization to let $\mathbf{A}$ and $\mathbf{b}$ be in a certain range and optimize to hedge for the worst case. Ghaoui and Lebert [6] develop this idea by considering the min-max optimization problem

$$\min_{\mathbf{x}} \max_{||\boldsymbol{\xi}_{\mathbf{A}},\boldsymbol{\xi}_{\mathbf{b}}||_F \leq \rho} ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})|| \,. \tag{1.3}$$

For each given $\mathbf{x} \in \mathscr{S}_\rho := \{(\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}) | \, ||\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}||_F \leq \rho\}$, the inner problem (1.4)

$$r(\mathbf{A}, \mathbf{b}, \rho, \mathbf{x}) := \max_{||\boldsymbol{\xi}_{\mathbf{A}},\boldsymbol{\xi}_{\mathbf{b}}||_F \leq \rho} ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})|| \tag{1.4}$$

is solved to hedge the worst case over all the elements in $\mathscr{S}_\rho$. A key issue here is that (1.3) does not consider the possible probability structure over the set $||\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}||_F \leq \rho$. It only considers the worst scenario. If the error has a very small probability for the worst case, the estimates from (1.3) might be too pessimistic. Alternatively, we may require the error vector

$$\boldsymbol{\xi} = [vec(\boldsymbol{\xi}_{\mathbf{A}}); \boldsymbol{\xi}_{\mathbf{b}}] \tag{1.5}$$

to follow additional statistical properties, with partially known information on the distribution of $\boldsymbol{\xi}$ over the possible perturbation set $\mathscr{S}_\rho$. Here $vec(\cdot)$ denotes the vectorization of a given matrix. This partial information can be used to define a set of probability measures $\mathscr{P}$, which is called the probability ambiguity set. It will be more realistic to hedge the worst case over this set of probability measures instead of hedging the worst scenario over the error sample space $\mathscr{S}_\rho$. This leads to a distributionally robust least squares frameworks as

$$\min_{\mathbf{x}} \max_{P \in \mathscr{P}} \mathbb{E}_P\{||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2\}, \tag{1.6}$$

where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{b} \in \mathbb{R}^m$, $\boldsymbol{\xi} \in \mathbb{R}^{m\times(n+1)}$ and $\mathbf{x} \in \mathbb{R}^n$ is the vector of decision variables. Let $n' = mn + m$ which is the total dimension of the error $\boldsymbol{\xi}$. A general formulation of the ambiguity set $\mathscr{P}$ is defined as:

$$\mathscr{P} := \{P : \mathbb{E}_P[\mathbf{1}] = \mathbf{1}, \mathbb{E}_P[\phi_i(\boldsymbol{\xi})] = \tau_i, i = 1, \ldots, l, \mathbb{E}_P[\phi_i(\boldsymbol{\xi})] \preceq \tau_i, i = l+1, \ldots, r, \boldsymbol{\xi} \in \Omega\}, \tag{1.7}$$

where $\boldsymbol{\xi} = [vec(\boldsymbol{\xi}_{\mathbf{A}}); \boldsymbol{\xi}_{\mathbf{b}}]$. $\phi_i(\cdot)$ is a Borel measurable real function with respect to $\boldsymbol{\xi}$ for $i = 1, \ldots, r$. Here, "$\preceq$" implies both inequality $\leq$ and semidefinite condition $\preceq$, i.e. $\preceq$ means $\mathbb{E}_P[\phi_i(\boldsymbol{\xi})] \leq \tau_i$ when $\phi_i(\boldsymbol{\xi})$ is a one-dimensional function and $\preceq$ implies that $\tau_i - \mathbb{E}_P[\phi_i(\boldsymbol{\xi})]$ is positive semidefinite when $\phi_i(\boldsymbol{\xi})$ is a symmetric matrix for $\forall \boldsymbol{\xi} \in \Omega$, and more generally an ordering. The definition (1.7) is a general form. When $\phi_i(\cdot)$'s are polynomial functions, (1.6) becomes a moment-robust optimization problem [10]. Define

$$r(\mathbf{x}, \Theta) := \max_{P \in \mathscr{P}} \mathbb{E}_P\{||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2\} \tag{1.8}$$

to be the inner problem, where Θ are the parameters used to define the ambiguity set $\mathscr{P}$. The outer problem is $\min_{\mathbf{x}} r(\mathbf{x}, \Theta)$.

The idea of distributionally robust optimization is originated in Scarf [18]. A more generic distributionally robust stochastic program takes the form:

$$\min_{x\in\mathcal{X}}\{\max_{P\in\mathscr{P}}\mathbb{E}_P[h(\mathbf{x},\boldsymbol{\xi})]\}, \tag{1.9}$$

where $\mathbb{E}_P[\cdot]$ is the expectation taken with respect to the random vector $\boldsymbol{\xi}$. Distributionally robust optimization has gained significant interests recently. Dupacova [4], Bertsimas et al. [2], and Delage and Ye [3] use linear or conic constraints to describe $\mathscr{P}$ with moments. Shapiro and Ahmed [19] defined a probability ambiguity set with measure bounds and general moment constraints. Pflug and Wozabal [12] consider the probability ambiguity set defined by confidence regions over a reference probability measure. Bertsimas et al. [2] use a piece-wise linear utility with first and second moment equality constraints and show that the corresponding problem has semidefinite programming reformulations. Delage and Ye [3] give general conditions for polynomial time solvability of the generic model (1.9). Shapiro and Ahmed [19] give stochastic programming reformulations of their model.

In this paper, we use the above three ambiguity settings in the framework of the distributionally robust least square problem. In particular, the distributionally robust least squares problem is studied with (1) first and second moment constraints, (2) norm bounds with first and second moment constraints, (3) a confidence region over a reference probability measure. The first and second moment constraints case is useful because the moment estimates are the most common statistics in practice. The norm bounds constraints can be understood as the bounds over the probability density function for a continuous random variable (probability mass function for a discrete random variable). The confidence region over a reference probability measure is for the case where an empirical distribution is available. Under these three different settings of probability ambiguity, we show that the distributionally robust least squares problem can always be solved efficiently. In particular, we show that the separation problem in the convex optimization reformulation of case (1) is polynomially solvable. We give stochastic programming and finite reformulations of problems in cases (2) and (3).

2 Moment Robust Least Square

In this section, we study two moment robust least square (MRLS) models. In the first model, we assume that we know the exact first and second moment information. In the second model, confidence intervals for the first and second moments are considered.

2.1 MRLS with Exact Moment Information

Bertsimas et al. [2] consider a linear moment distributionally robust (MDR) problem with the probability ambiguity set defined by the exact information of the first and second moments. With this idea, let us consider the DRLS problem (1.6) with the probability ambiguity set $\mathscr{P}$ defined as:

$$\mathscr{P} := \{P : \mathbb{E}_P[\mathbf{1}] = \mathbf{1}, \mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}, \mathbb{E}_P[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \mathbf{Q}, ||\boldsymbol{\xi}|| \leq \rho, \mathbf{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^T \succ 0\}, \tag{2.1}$$

where ρ is the bound used to define the possible region of the perturbation vector $\boldsymbol{\xi}$. Let us define

$$\phi_0(\mathbf{x}, \boldsymbol{\xi}) := ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 . \tag{2.2}$$

For the MRLS problem with exact moment information, we make the following assumptions:

Assumption 1 *The probability ambiguity set* (2.1) *is nonempty, i.e.* $\exists$ *probability measure* ν *defined on the sample space* $\mathscr{S}_\rho$ *such that* $\mathbb{E}_P[\boldsymbol{\xi}] = \boldsymbol{\mu}, \mathbb{E}_P[\boldsymbol{\xi}\boldsymbol{\xi}^T] = \mathbf{Q}$.

Assumption 2 *The first and second moments* $\boldsymbol{\mu}$ *and* $\mathbf{Q}$ *are finite and satisfy* $\mathbf{Q} \succ \boldsymbol{\mu}\boldsymbol{\mu}^T$, *where* $\mathbf{Q} \succ \boldsymbol{\mu}\boldsymbol{\mu}^T$ *means* $\mathbf{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^T$ *is positive definite (p.d.).*

Then for a given $\mathbf{x}$, the inner problem (1.8) of the corresponding (1.6) with probability ambiguity set $\mathscr{P}$ defined in (2.1) can be written as:

$$\begin{aligned} Z(\mathbf{x}) := \sup_{\nu \in \mathscr{P}} & \int_{\mathscr{S}_\rho} \phi_0(\mathbf{x}, \boldsymbol{\xi}) d\nu(\boldsymbol{\xi}) \\ s.t. & \int_{\mathscr{S}_\rho} \xi_i \xi_j d\nu(\boldsymbol{\xi}) = Q_{ij} \quad \forall i, j = 1, \ldots, n', \\ & \int_{\mathscr{S}_\rho} \xi_i d\nu(\boldsymbol{\xi}) = \mu_i \quad \forall i, j = 1, \ldots, n', \\ & \int_{\mathscr{S}_\rho} d\nu(\boldsymbol{\xi}) = \mathbf{1}, \end{aligned} \tag{2.3}$$

where $n' = mn + m$. Taking the dual of the inner problem (2.3), we get an equivalent formulation of the inner problem (1.8) as:

$$\begin{aligned} Z_D(\mathbf{x}) := \min_{\mathbf{Y}, \mathbf{y}, y_0} \quad & \mathbf{Q} \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + y_0 \\ s.t. \quad & \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} + \boldsymbol{\xi}^T \mathbf{y} + y_0 \geq ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 \quad \text{for } \forall \boldsymbol{\xi} \in \mathscr{S}_\rho, \end{aligned} \tag{2.4}$$

where $\mathscr{S}_\rho := \{\boldsymbol{\xi} : ||\boldsymbol{\xi}|| \leq \rho\}$.

The strong duality holds between (2.3) and (2.4) if moment vector lies in the interior of the set of feasible moment vectors, which is a direct result according to Theorem 7 in the Appendix. More detailed discussion can be found in Isii [9] and Lasserre [10]. The assumption (A2) guarantees that the covariance matrix $\mathbf{Q} - \boldsymbol{\mu}\boldsymbol{\mu}^{\boldsymbol{T}}$ is p.d. and the strong duality between (2.3) and (2.4) is satisfied. We combine (2.4) with the outer problem and have the following theorem.

Theorem 1 *The original MRLS problem* (1.6) *is equivalent to:*

$$\min_{\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0} \quad \mathbf{Q} \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + y_0 \tag{2.5a}$$

$$s.t. \quad \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} + \boldsymbol{\xi}^T \mathbf{y} + y_0 \geq ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 \quad \text{for } \forall \boldsymbol{\xi} \in \mathscr{S}_\rho. \tag{2.5b}$$

2.2 MRLS with Moment Confidence Interval

Delage and Ye [3] describe a general moment distributionally robust framework with the probability ambiguity set defined by confidence intervals over the first two moments as follows:

$$\begin{aligned}\mathscr{D} = \{\nu : \mathbb{E}_\nu[\mathbf{1}] = \mathbf{1}, &(\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu}) \le \alpha, \\ &\mathbb{E}_\nu[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T]\} \preceq \beta \mathbf{Q}, \boldsymbol{\xi} \in \mathscr{S}\}.\end{aligned} \tag{2.6}$$

They also show that under certain conditions, the moment distributionally robust optimization problem can be polynomially solved by the ellipsoid method. Their results are summarized in the next lemma.

Lemma 1 (Delage and Ye 2008) *Given an feasible* $\mathbf{x}$*, consider the moment distributionally robust problem defined as:*

$$\max_{\nu \in \mathscr{D}} \mathbb{E}_\nu[h(\mathbf{x}, \boldsymbol{\xi})], \tag{2.7}$$

where $h(\mathbf{x}, \cdot)$ *is measurable with respect to* $\forall \nu \in \mathscr{D}$ *and* $\mathbb{E}_\nu[\cdot]$ *is the expectation taken with respect to the random vector* $\boldsymbol{\xi}$*, given that it follows the probability distribution* ν *over the sample space* $\mathscr{S}$*. Suppose* $\alpha \ge 0$*,* $\beta \ge 1$*,* $\mathbf{Q} \succ 0$ *and* $h(\mathbf{x}, \boldsymbol{\xi})$ *is* ν*-integrable for all* $\nu \in \mathscr{D}$*. Then the optimal value of the inner problem* (2.7) *is equal to the optimal value of the problem:*

$$\begin{aligned}\min_{\mathbf{Y}, \mathbf{y}, y_0, t} \quad & y_0 + t \\ s.t. \quad & y_0 \ge h(\mathbf{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} - \boldsymbol{\xi}^T \mathbf{y} \quad \forall \boldsymbol{\xi} \in \mathscr{S}, \\ & t \ge (\beta \mathbf{Q} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + \sqrt{\alpha} \left|\left| \mathbf{Q}^{\frac{1}{2}} (\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) \right|\right|, \\ & \mathbf{Y} \succeq 0.\end{aligned} \tag{2.8}$$

Consider the DRLS problem (1.6) with probability ambiguity set $\mathscr{P}$ defined as:

$$\begin{aligned}\mathscr{P} = \{\nu : \mathbb{E}_\nu[\mathbf{1}] = \mathbf{1}, &(\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu}) \le \alpha, \\ &\mathbb{E}_\nu[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T]\} \preceq \beta \mathbf{Q}, \\ &\boldsymbol{\xi} \in \mathscr{S}_\rho = \{\boldsymbol{\xi} : ||\boldsymbol{\xi}|| \le \rho\}\}\end{aligned} \tag{2.9}$$

with the parameters $\alpha \ge 0$, $\beta \ge 1$. For the inner problem (1.8) with $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \rho\}$ and $\mathscr{P}$ defined as (2.9), we can apply Lemma 1 with:

$$h(\mathbf{x}, \boldsymbol{\xi}) = ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 .$$

This gives us an equivalent formulation of the inner problem (1.8) described as follows.

Theorem 2 *For a given fixed* $\mathbf{x} \in \mathbb{R}^n$*, if* $\alpha \ge 0$*,* $\beta \ge 1$*,* $\mathbf{Q} \succ 0$*. Then, the optimal value of the inner problem (1.8) must be equal to the optimal value of the problem:*

$$\begin{aligned}\min_{\mathbf{Y}, \mathbf{y}, y_0, t} \quad & (\beta \mathbf{Q} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + y_0 + \sqrt{\alpha} t \\ s.t. \quad & \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} + \boldsymbol{\xi}^T \mathbf{y} + y_0 \ge ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 \quad \forall \boldsymbol{\xi} \in \mathscr{S}_\rho, \\ & \left|\left| \mathbf{Q}^{\frac{1}{2}} (\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) \right|\right| \le t, \\ & \mathbf{Y} \succeq 0.\end{aligned} \tag{2.10}$$

Proof The first constraint $\boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi}+\boldsymbol{\xi}^T\mathbf{y}+y_0 \geq ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2$ is the same as the first constraint in (2.8). The second constraint

$$t \geq (\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}+\sqrt{\alpha}\left|\left|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y}+2\mathbf{Y}\boldsymbol{\mu})\right|\right|$$

in (2.8) can be rewriten as two constraints

$$t_1=(\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y},\ t_2\geq\sqrt{\alpha}\left|\left|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y}+2\mathbf{Y}\boldsymbol{\mu})\right|\right|,\ t=t_1+t_2.$$

Substitute $t=t_1+t_2$ and $t_1=(\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}$ in (2.8), it is equivalent to (2.10). ■

According to Theorem 2, we can combine (2.10) with the outer problem to get the equivalent formulation of $\min_{\mathbf{x}}\max_{P\in\mathscr{P}}\mathbb{E}_P\{||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2\}$ with the probability ambiguity set $\mathscr{P}$ defined in (2.9) as

$$\min_{\mathbf{x},\mathbf{Y},\mathbf{y},y_0,t}\ (\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}+y_0+\sqrt{\alpha}t \tag{2.11a}$$

$$\text{s.t.}\ \ \boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi}+\boldsymbol{\xi}^T\mathbf{y}+y_0\geq||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2\quad\forall\boldsymbol{\xi}\in\mathscr{S}_\rho, \tag{2.11b}$$

$$\left|\left|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y}+2\mathbf{Y}\boldsymbol{\mu})\right|\right|\leq t, \tag{2.11c}$$

$$\mathbf{Y}\succeq 0, \tag{2.11d}$$

2.3 Complexity of the MRLS

In this section, we will show that we can have a polynomial time solution method for solving (2.5) and (2.11). All the analysis in this section will focus on (2.11). (2.5) is similar and much simpler than (2.11). Therefore, all the following analysis can be easily applied to the complexity analysis of (2.5). Grotschel et al. [8] show that convex optimization and separation of a convex set from a point is equivalent. The result is summarized in the following lemma.

Lemma 2 (Grotschel et al. 1981 and Anstreicher 1997) *Consider a convex optimization problem of the form*

$$\min_{\mathbf{z}\in\mathscr{Z}}\mathbf{c}^T\mathbf{z}$$

with linear objective and convex feasible set $\mathscr{Z}$. Given that the set of optimal solutions is nonempty, the problem can be solved to any precision ϵ in time polynomial in $\log(1/\epsilon)$ *and in the size of the problem by using the ellipsoid method or Vaidya's volumetric cutting plane method if $\mathscr{Z}$ satisfies the following two conditions:*

1. *for any $\bar{\mathbf{z}}$, it can be verified whether $\bar{\mathbf{z}}\in\mathscr{Z}$ or not in time polyhnomial in the dimension of $\mathbf{z}$.*

2. *for any infeasible $\bar{\mathbf{z}}$, a hyperplane that separates $\bar{\mathbf{z}}$ from the feasible region $\mathscr{Z}$ can be generated in time polynomial in the dimension of $\mathbf{z}$.*

According to Lemma 2, we need to find a polynomial time oracle to verify the feasibility for a given assignment $\mathbf{x}$, $\mathbf{Y}$, $\mathbf{y}$, y_0. The main difficulty is the verification of the constraints

$$\boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{y} + y_0 \geq ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2 \quad \forall \boldsymbol{\xi} \in \mathscr{S}_\rho,$$

because there are infinitely many $\boldsymbol{\xi}$'s in the sample space $\mathscr{S}_\rho$. For a given $\mathbf{Y}$, $\mathbf{y}$, y_0, consider the following optimization problem

$$g(\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0) = \max_{\boldsymbol{\xi}\in\mathscr{S}_\rho} ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2 - \boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi} - \boldsymbol{\xi}^T\mathbf{y} - y_0, \tag{2.12}$$

and we have the following proposition

Proposition 1 *For a given* $\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0$*, problem* (2.12) *is equivalent to:*

$$g(\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0) := \max_{||\boldsymbol{\xi}||\leq\rho} \boldsymbol{\xi}^T\hat{\mathbf{A}}\boldsymbol{\xi} + \hat{\mathbf{b}}\mathbf{a}, \tag{2.13}$$

where

$$\begin{aligned}&\hat{\mathbf{A}} = \mathbf{B}^T\mathbf{B} - \mathbf{Y},\ \hat{\mathbf{b}} = 2\mathbf{B}^T(\mathbf{A}\mathbf{x}-\mathbf{b}) - \mathbf{y},\\ &\mathbf{B}_{m\times m(n+1)} = \begin{pmatrix}\mathbf{X} & -\mathbf{I}_m\end{pmatrix},\ \mathbf{X} = [vec(\mathbf{e}_1\mathbf{x}^T), vec(\mathbf{e}_2\mathbf{x}^T), \cdots, vec(\mathbf{e}_m\mathbf{x}^T)]^T,\end{aligned}$$

$\mathbf{I}_m$ *is a* $m \times m$ *identity matrix and* $\mathbf{e}_i$ *is the* m *dimensional vector with* 1 *in the ith entry and* 0 *otherwise.*

Proof Since $\boldsymbol{\xi}_{\mathbf{A}}\mathbf{x} - \boldsymbol{\xi}_{\mathbf{b}} = \mathbf{B}\boldsymbol{\xi}$, where:

$$\mathbf{B}_{m\times m(n+1)} = \begin{pmatrix}\mathbf{X} & -\mathbf{I}_m\end{pmatrix},\ \mathbf{X} = [vec(\mathbf{e}_1\mathbf{x}^T), vec(\mathbf{e}_2\mathbf{x}^T), \cdots, vec(\mathbf{e}_m\mathbf{x}^T)]^T$$

(2.12) can be rewritten as:

$$\begin{aligned}g(\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0) &= \max_{\boldsymbol{\xi}\in\mathscr{S}_\rho} \boldsymbol{\xi}^T(\mathbf{B}^T\mathbf{B} - \mathbf{Y})\boldsymbol{\xi} + [2\mathbf{B}^T(\mathbf{A}\mathbf{x}-\mathbf{b}) - \mathbf{y}]^T\boldsymbol{\xi} + ||\mathbf{A}\mathbf{x}-\mathbf{b}||^2 - y_0\\ &= \max_{||\boldsymbol{\xi}||\leq\rho} \boldsymbol{\xi}^T\hat{\mathbf{A}}\boldsymbol{\xi} + \hat{\mathbf{b}}\mathbf{a}. \quad \blacksquare\end{aligned}$$

Problem (2.12) is a standard trust region subproblem which can be solved polynomially with respect to the size of $\boldsymbol{\xi}$ [15, 5]. In summary, we have the following lemma.

Lemma 3 *The optimization problem* (2.12) *can be solved in polynomial time with respect to the size of* $\boldsymbol{\xi}$*.*

Based on the above analysis, the following proposition that follows states that the equivalent formulation (2.9) of the MRLS problem (1.6) with ambiguity set defined in (2.9) can be solved in polynomial time.

Proposition 2 *The MRLS problem* (2.11) *can be solved to any precision* ϵ *in time polynomial in* $\log(1/\epsilon)$ *and the sizes of* $\mathbf{x}$ *and* $\boldsymbol{\xi}$*.*

Proof At first, the constraints of problem (2.11) describe a convex set with respect to the variables $\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0, t$, because for any $\boldsymbol{\xi} \in \mathscr{S}_\rho$, $\boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi} + \boldsymbol{\xi}^T\mathbf{y} + y_0 \geq ||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2$ is a quadratic constraint with respect to $\mathbf{x}, \mathbf{Y}, \mathbf{y}, y_0$ and it can be derived as:

$$\begin{aligned}&||(\boldsymbol{\xi}^T\boldsymbol{\xi} \bullet \mathbf{Y} + 1, \boldsymbol{\xi}^T\mathbf{y} + 1, y_0 + 1)||^2 \\ &\quad \geq 2\,||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 + (\boldsymbol{\xi}^T\boldsymbol{\xi} \bullet \mathbf{Y})^2 + (\boldsymbol{\xi}^T\mathbf{y})^2 + y_0^2 + 3,\end{aligned}$$

which is a second order cone constraint, and the constraint $\left|\left|\mathbf{Q}^{\frac{1}{2}}(\boldsymbol{\xi} + 2\mathbf{Y}\boldsymbol{\mu})\right|\right| \leq t$ is also a second order cone constraint with respect to $\mathbf{Y}, \mathbf{y}, t$. The feasible set is also nonempty because the assignment $\mathbf{x} = \mathbf{x}_0$, $\mathbf{Y} = \mathbf{I}$, $\mathbf{y} = 0$, $t = 2\left|\left|\mathbf{Q}^{\frac{1}{2}}\boldsymbol{\mu}\right|\right|$, $y_0 = \sup_{\boldsymbol{\xi} \in \mathscr{S}_\rho}\{||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 - \boldsymbol{\xi}^T\boldsymbol{\xi}\}$ is necessarily feasible. Note that such an assignment for y_0 exists because

$$||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 - \boldsymbol{\xi}^T\boldsymbol{\xi}$$

is a continuous function with respect to $\boldsymbol{\xi}$ and $\mathscr{S}_\rho$ is a compact set. For any given $\mathbf{x}$, the objective value of the inner problem (1.8) is nonnegative because $||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 \geq 0$ for $\forall \boldsymbol{\xi} \in \mathscr{S}_\rho$ and $\mathbf{x}$. The weak duality theorem tells us that problem (2.8) is bounded below by zero for any given $\mathbf{x}$ which implies that problem (2.11) is bounded below. Therefore, the optimal solution set of problem (2.11) is nonempty.

Now we need to verify the two conditions in Lemma 2 in order to show that problem (2.11) can be solved in polynomial time. In case of constraint (2.11d), feasibility can be verified in $O(n'^3)$ arithmetic operations. Moreover, a separating hyperplane can be generated, if necessary, based on the eigenvector corresponding to the lowest eigenvalue. The feasibility of constraint (2.11c) is also easily verified. Based on an infeasible assignment $(\bar{\mathbf{Y}}, \bar{\mathbf{y}}, \bar{y}_0, \bar{t})$, if $\mathbf{Q}^{\frac{1}{2}}(\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) \neq 0$, a seperating hyperplane can be constructed in polynomial time:

$$\nabla_\mathbf{Y} g(\bar{\mathbf{Y}}, \bar{\mathbf{y}}) \bullet \mathbf{Y} + \nabla_\mathbf{y} g(\bar{\mathbf{Y}}, \bar{\mathbf{y}})^T\mathbf{y} - t \leq \nabla_\mathbf{Y} g(\bar{\mathbf{Y}}, \bar{\mathbf{y}}) \bullet \bar{\mathbf{Y}} + \nabla_\mathbf{y} g(\bar{\mathbf{Y}}, \bar{\mathbf{y}})^T\bar{\mathbf{y}} - g(\bar{\mathbf{Y}}, \bar{\mathbf{y}}),$$

where $g(\mathbf{Y}, \mathbf{y}) = \left|\left|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu})\right|\right|$ and where $\nabla_\mathbf{Y} g(\mathbf{Y}, \mathbf{y})$ and $\nabla_\mathbf{y} g(\mathbf{Y}, \mathbf{y})$ are the gradients of $g(\mathbf{Y}, \mathbf{y})$ in $\mathbf{Y}$ and $\boldsymbol{\xi}$. If $\mathbf{Q}^{\frac{1}{2}}(\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) = 0$, a seperating hyperplane can be polynomially constructed as:

$$t \geq \left|\left|\mathbf{Q}^{\frac{1}{2}}(\bar{\mathbf{y}} + 2\bar{\mathbf{Y}}\boldsymbol{\mu})\right|\right|,$$

which is just $t \geq 0$. Finally, let us consider the constraint (2.11b). Lemma 3 states that the subproblem

$$\max_{\boldsymbol{\xi} \in \mathscr{S}_\rho} ||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 - \boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi} - \boldsymbol{\xi}^T\mathbf{y} - y_0$$

is solvable in polynomial time for any given assignment $(\bar{\mathbf{x}}, \bar{\mathbf{Y}}, \bar{\mathbf{y}}, \bar{y}_0, \bar{t})$. Given that the optimal value is found as r, if $r \geq 0$, one can conclude infeasibility of the constraint and generate an associated separating hyperlane using any optimal solution $\boldsymbol{\xi}_*$ as follows:

$$\begin{aligned}&(\boldsymbol{\xi}_*\boldsymbol{\xi}_*^T) \bullet \mathbf{Y} + \boldsymbol{\xi}_*^T\mathbf{y} + y_0 - 2((\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}*})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}*}))^T(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}*})\mathbf{x} \\ &\quad \geq ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}*})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}*})||^2 - 2((\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}*})\bar{\mathbf{x}} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}*}))^T(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}*}))\bar{\mathbf{x}}.\end{aligned}$$

This separating hyperplane is valid because $||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}*})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}*})||^2$ is convex with respect to $\mathbf{x}$. Then Lemma 2 can be applied and we can conclude that problem (2.11) can be solved in polynomial time. ▮

Remark. Now we compare the above result with the result of Delage and Ye [3]. They consider a general distributionally robust optimization problem in the form

$$\min_{\mathbf{x}\in\mathscr{X}}\{\max_{P\in\mathscr{P}}\mathbb{E}_P[h(\mathbf{x},\boldsymbol{\xi})]\} \tag{2.14}$$

under the assumption that $h(\mathbf{x},\boldsymbol{\xi})$ is convex with respect to $\mathbf{x}$ and concave with respect to $\boldsymbol{\xi}$. This assumption is crucial to generate separating hyperplanes, which needs a subgradient of $h(\mathbf{x},\boldsymbol{\xi})$ with respect to $\mathbf{x}$ and a supergradient of $h(\mathbf{x},\boldsymbol{\xi})$ with respect to $\boldsymbol{\xi}$. The difference between the MRLS (1.6) and (2.14) is that the objective of (1.6) is not concave with respect to $\boldsymbol{\xi}$. Here, we take the advantage of the specific formulation of the subproblem (2.12) which can be derived as a standard trust-region subproblem, and can be solved efficiently. Then we use the solution of this subproblem to generate a separating hyperplane.

3 DRLS with Bounds on the Probability Measure

In this section, we consider the DRLS (1.6) with the ambiguity set containing bounds on the probability measure. This kind of probability ambiguity is considered by Shapiro and Ahmed [19]. For a given parameter ρ, let us denote $\mathcal{X}_\rho$ to be the space of all finite measures on $(\mathscr{S}_\rho,\mathcal{B}_{\mathscr{S}_\rho})$, where $\mathcal{B}_{\mathscr{S}_\rho}$ is the Borel σ-algebra on $\mathscr{S}_\rho$. For $\forall\nu\in\mathcal{X}$, we have $\nu\succeq 0$, which means that $\nu(C)\geq 0$ for $\forall C\in\mathcal{B}_{\mathscr{S}_\rho}$. Also, for $\forall\nu_1,\nu_2\in\mathscr{S}_\rho$, we write $\nu_1\succeq\nu_2$ or $\nu_1-\nu_2\succeq 0$ if $\nu_1(C)\geq\nu_2(C)$ for $\forall C\in\mathcal{B}_{\mathscr{S}_\rho}$.

3.1 Ambiguity with both Measure Bounds and Moment Information

Given $\nu_1,\nu_2\in\mathcal{X}_\rho$ with $\nu_1\preceq\nu_2$, let us consider the probability ambiguity set:

$$\begin{aligned}\mathscr{P}:=\{\nu\in\mathcal{X}_\rho:\nu(\mathscr{S}_\rho)=1,\ \nu_1\preceq\nu\preceq\nu_2,&(\mathbb{E}_\nu[\boldsymbol{\xi}]-\boldsymbol{\mu})^T\mathbf{Q}^{-1}(\mathbb{E}_\nu[\boldsymbol{\xi}]-\boldsymbol{\mu})\leq\alpha,\\ &\mathbb{E}_\nu[(\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T]\}\preceq\beta\mathbf{Q}\}.\end{aligned} \tag{3.1}$$

With the probability ambiguity set (3.1), the inner problem (1.8) can be defined as

$$\begin{aligned}\max_{\nu\in\mathcal{M}}\ &\int_{\mathscr{S}_\rho}\phi_0(\mathbf{x},\boldsymbol{\xi})d\nu(\boldsymbol{\xi}) \qquad (3.2)\\ s.t.\ &\int_{\mathscr{S}_\rho}d\nu(\boldsymbol{\xi})=1,\\ &\int_{\mathscr{S}_\rho}(\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^Td\nu(\boldsymbol{\xi})\preceq\beta\mathbf{Q},\\ &\int_{\mathscr{S}_\rho}\begin{pmatrix}\mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu})\\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha\end{pmatrix}d\nu(\boldsymbol{\xi})\succeq 0,\end{aligned}$$

where $\phi_0(\mathbf{x},\boldsymbol{\xi}) = ||(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}})||^2$ and $\mathcal{M} := \{\nu : \nu_1 \preceq \nu \preceq \nu_2\}$. Now make the assumption that the feasible set of (3.2) is nonempty. Let us write (3.2) as:

$$\begin{aligned} \max_{\nu\in\mathcal{M}} \quad & \int_{\mathscr{S}_\rho} \phi_0(\mathbf{x},\boldsymbol{\xi})d\nu(\boldsymbol{\xi}) \\ s.t. \quad & \int_{\mathscr{S}_\rho} d\nu(\boldsymbol{\xi}) = 1, \\ & \int_{\mathscr{S}_\rho} ((\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T - \beta\mathbf{Q})d\nu(\boldsymbol{\xi}) \preceq 0, \\ & \int_{\mathscr{S}_\rho} \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix} d\nu(\boldsymbol{\xi}) \succeq 0. \end{aligned} \tag{3.3}$$

Let us define:

$$\begin{aligned} \mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi}) :=& \phi_0(\mathbf{x},\boldsymbol{\xi}) - \lambda_0 - ((\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T - \beta\mathbf{Q}) \bullet \mathbf{\Lambda}_1 \\ & + \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix} \bullet \mathbf{\Lambda}_2. \end{aligned} \tag{3.4}$$

Then the Lagrangian of problem (3.3) is:

$$L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) := \int_{\mathscr{S}_\rho} \mathcal{L}_{\lambda_0,\Lambda_1,\Lambda_2}(\mathbf{x},\boldsymbol{\xi})d\nu(\boldsymbol{\xi}) + \lambda_0 \tag{3.5}$$

and the Lagrangian dual of (3.3) is:

$$\begin{aligned} \min_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2} \quad & \{\psi(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) := \sup_{\nu\in\mathcal{M}} L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2)\} \\ s.t. \ & \mathbf{\Lambda}_1,\mathbf{\Lambda}_2 \succeq 0. \end{aligned} \tag{3.6}$$

Proposition 3 *In the Lagrangian dual problem* (3.6)*,*

$$\psi(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) := \int_{\mathscr{S}_\rho} [\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_+ d\nu_2(\boldsymbol{\xi}) - \int_{\mathscr{S}_\rho} [\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_- d\nu_1(\boldsymbol{\xi}) + \lambda_0. \tag{3.7}$$

Proof Since $L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2)$ is a measurable function with respect to ν_1 and ν_2, for given feasible $\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2$, the sets $C_- := \{\boldsymbol{\xi} : L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) < 0, \boldsymbol{\xi} \in \mathscr{S}_\rho\}$ and $C_+ := \{\boldsymbol{\xi} : L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) < 0, \boldsymbol{\xi} \in \mathscr{S}_\rho\}$ are measurable. In order to achieve the maximum over the sets $\{\nu : \nu_1 \preceq \nu \preceq \nu_2\}$, we should integrate $L(\mathbf{x},\nu,\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2)$ with respect to ν_1 over the set C_- and with respect to ν_2 over the set C_+, which leads to the desired result (3.7). ■

According to the standard Lagrangian weak duality theory, we know that the optimal value of problem (3.3) is always less than or equal to the optimal value of its dual (3.6). The conjugate duality theory [16, 1] tells us that the strong duality for (3.3) and (3.6) holds and the set of optimal solutions of the dual problem is nonempty and bounded if the following assumption holds:

Assumption 3 *The optimal value of* (3.3) *is finite, and there exists a* $\mathbf{x}$ *and a probability measure* $\nu \in \mathcal{M}$ *to* (3.3) *such that*

$$\begin{aligned}
&\int_{\mathscr{S}_\rho} ((\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T - \beta\mathbf{Q})d\nu(\boldsymbol{\xi}) \prec 0, \\
&\int_{\mathscr{S}_\rho} \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix} d\nu(\boldsymbol{\xi}) \succ 0,
\end{aligned} \tag{3.8}$$

where for two square matrics $\mathbf{A}$ *and* $\mathbf{B}$*,* $\mathbf{A} \succ \mathbf{B}$ *means* $\mathbf{A}-\mathbf{B}$ *is positive definite.*

According to the above analysis, we know that under Assumption 3, there is no duality gap between (3.3) and (3.6). With Assumption 3, the original DRLS problem (1.6) with probability ambiguity set (3.1) is equivalent to:

$$\begin{aligned}
\min_{\mathbf{x},\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2} \quad & \{\psi(\mathbf{x},\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2)\} \\
\text{s.t. } & \boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2 \succeq 0.
\end{aligned} \tag{3.9}$$

3.2 Some Cases

According the previous section, the DRLS problem (1.6) with probability ambiguity set (3.1) is equivalent to (3.9) where $\psi(\mathbf{x},\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2)$ is defined as (3.7). We assume that ν_1 and ν_2 are finite measures and $\nu_1,\nu_2 \succ 0$, where $\nu \succ 0$ means $\nu \succeq 0$ and $\exists A \in \mathcal{B}_{\mathscr{S}_\rho}$ such that $\nu(A) > 0$. In this section, we will discuss some special cases.

3.2.1 Finite support case

Let us consider the case that the sample space is finite as $\mathscr{S} := \{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^K\}$. In this case, the measure ν_1 and ν_2 are actually two vectors which give the bounds for the probability measure. Let $\nu_1 = (\underline{p}_1,\ldots,\underline{p}_K)$ and $\nu_2 = (\bar{p}_1,\ldots,\bar{p}_K)$ with $\underline{p}_i < \bar{p}_i$ for $i = 1,\ldots,K$. The ambiguity set $\mathscr{P}$ can be defined as follows:

$$\begin{aligned}
\mathscr{P} := \{(p_1,\ldots,p_K) : \sum_{i=1}^K p_i = 1,\ \underline{p}_i \le p_i \le \bar{p}_i \text{ for } i = 1,\ldots,K, \boldsymbol{\tau} = \sum_{i=1}^K p_i\boldsymbol{\xi}^i \\
(\boldsymbol{\tau}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}(\boldsymbol{\tau}-\boldsymbol{\mu}) \le \alpha,\ \sum_{i=1}^K p_i(\boldsymbol{\xi}^i-\boldsymbol{\mu})(\boldsymbol{\xi}^i-\boldsymbol{\mu})^T\} \preceq \beta\mathbf{Q}\}
\end{aligned} \tag{3.10}$$

Consider the inner problem (1.8) with probability ambiguity set (3.10). Let us make the following assumption.

Assumption 4 *There exists* $\mathbf{p} := (p_1,\ldots,p_K) \in \mathscr{P}$ *such that* $\underline{p}_i < p_i < \bar{p}_i$*,* $(\boldsymbol{\tau}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}(\boldsymbol{\tau}-\boldsymbol{\mu}) < \alpha$*,* $\sum_{i=1}^K p_i(\boldsymbol{\xi}^i-\boldsymbol{\mu})(\boldsymbol{\xi}^i-\boldsymbol{\mu})^T \prec \beta\mathbf{Q}\}$ *with* $\boldsymbol{\tau} = \sum_{i=1}^K p_i\boldsymbol{\xi}^i$*.*

Then we have the following theorem.

Theorem 3 *Assume that Assumption 4 is satisfied. For a given* $\mathbf{x}$*, the inner problem* (3.10) *is equivalent to:*

$$\begin{aligned} \min_{s,\bar{s}_i,\underline{s}_i,\mathbf{u},\mathbf{X},\boldsymbol{\kappa}} \quad & s+\sum_{i=1}^{K}(\bar{p}_i\bar{s}_i+\underline{p}_i\underline{s}_i)+[\sqrt{\alpha},-(\mathbf{Q}^{-\frac{1}{2}}\boldsymbol{\mu})^T]\boldsymbol{\kappa}+\beta\mathbf{Q}\bullet\mathbf{X} \qquad (3.11)\\ \text{s.t. } & s+\bar{s}_i+\underline{s}_i+\boldsymbol{\xi}^{i^T}\mathbf{u}+(\boldsymbol{\xi}^i-\boldsymbol{\mu})(\boldsymbol{\xi}^i-\boldsymbol{\mu})^T\bullet\mathbf{X}\geq\left|\left|(\mathbf{A}+\boldsymbol{\xi}^i_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}^i_{\mathbf{b}})\right|\right|^2,\\ & \qquad\qquad\qquad\qquad\qquad\qquad \text{for } i=1,\dots,K,\\ & -\mathbf{u}+[\mathbf{0},-(\mathbf{Q}^{-\frac{1}{2}})^T]\boldsymbol{\kappa}=0,\\ & \bar{s}_i\geq 0,\underline{s}_i\leq 0,\text{for } i=1,\dots,K,\\ & \boldsymbol{\kappa}\in\mathcal{K},\mathbf{X}\succeq 0, \end{aligned}$$

where $\mathcal{K}$ *is a second order cone.*

Proof The inner problem (1.8) with ambiguity set (3.10) can be explicitly written as:

$$\begin{aligned} \max_{\mathbf{p}:=(p_1,\dots,p_K)} \quad & \sum_{i=1}^{K}p_i\left|\left|(\mathbf{A}+\boldsymbol{\xi}^i_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}^i_{\mathbf{b}})\right|\right|^2 \qquad (3.12)\\ \text{s.t. } & \sum_{i=1}^{K}p_i=1,\\ & \underline{p}_i\leq p_i\leq\bar{p}_i,\\ & \boldsymbol{\tau}=\sum_{i=1}^{K}p_i\boldsymbol{\xi}^i,\\ & (\boldsymbol{\tau}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}(\boldsymbol{\tau}-\boldsymbol{\mu})\leq\alpha,\\ & \sum_{i=1}^{K}p_i(\boldsymbol{\xi}^i-\boldsymbol{\mu})(\boldsymbol{\xi}^i-\boldsymbol{\mu})^T\}\preceq\beta\mathbf{Q},\\ & \mathbf{p}:=(p_1,\dots,p_K)\in\mathbb{R}^K_+. \end{aligned}$$

(3.12) is a linear conic programming problem for a given $\mathbf{x}$. According to Assumption 4, the Slater's constraint qualification is satisfied. The standard duality theory for second order cone and semidefinite programming will gurantee the strong duality. Then we can take the dual of (3.12). The dual problem has the form (3.11). For more details of duality of linear conic problem, please refer to [22]. ∎

If we combine the above result with the outer problem, we can get:

Corollary 1 *Assume that Assumption 4 is satisfied. Then the DRLS problem is equivalent* *to:*

$$
\begin{aligned}
\min_{\mathbf{x},s,\bar{s}_i,\underline{s}_i,\mathbf{u},\mathbf{X},\boldsymbol{\kappa}} \quad & s + \sum_{i=1}^{K}(\bar{p}_i\bar{s}_i + \underline{p}_i\underline{s}_i) + [\sqrt{\alpha}, -(\mathbf{Q}^{-\frac{1}{2}}\boldsymbol{\mu})^T]\boldsymbol{\kappa} + \beta\mathbf{Q}\bullet\mathbf{X} \qquad (3.13)\\
\text{s.t. } & s + \bar{s}_i + \underline{s}_i + \boldsymbol{\xi}^{i^T}\mathbf{u} + (\boldsymbol{\xi}^i - \boldsymbol{\mu})(\boldsymbol{\xi}^i - \boldsymbol{\mu})^T \bullet \mathbf{X} \geq \left|\left|(\mathbf{A}+\boldsymbol{\xi}^i_{\mathbf{A}})\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}^i_{\mathbf{b}})\right|\right|^2, \\
& \qquad \text{for } i = 1,\dots,K,\\
& -\mathbf{u} + [\mathbf{0}, -(\mathbf{Q}^{-\frac{1}{2}})^T]\boldsymbol{\kappa} = 0,\\
& \bar{s}_i \geq 0, \underline{s}_i \leq 0, \text{for } i = 1,\dots,K,\\
& \boldsymbol{\kappa} \in \mathcal{K}, \mathbf{X} \succeq 0
\end{aligned}
$$

It is now easy to see that (3.13) is a standard linear conic optimization problem.

3.2.2 Measure bounds by a reference probability measure

Let us assume that the bounds ν_1 and ν_2 are given as:

$$\nu_1 := (1-\epsilon_1)P^*, \nu_2 := (1+\epsilon_2)P^*, \qquad (3.14)$$

where $\epsilon_1 \in [0,1]$ and $\epsilon_2 \geq 0$ are given constants. Consider the dual problem (3.6) and $\psi(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2)$ can be rewritten as:

$$\psi(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) = \int_{\mathscr{S}_\rho} ((1+\epsilon_2)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_+ - (1-\epsilon_1)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_-)dP^*(\boldsymbol{\xi}) + \lambda_0. \qquad (3.15)$$

Note that the function $(1+\epsilon_1)[\cdot]_+ - (1-\epsilon_2)[\cdot]_-$ is convex piecewise linear increasing and $\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})$ is convex in $\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2$ for each given $\boldsymbol{\xi} \in \mathscr{S}_\rho$. Consequently the objective function of (3.9) is convex in $\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2$. Therefore, the orignal DRLS problem (1.6) with the ambiguity set (3.1) is reformulated as a convex stochastic programming problem with the form (3.15). Now, (3.9) can be written as:

$$
\begin{aligned}
\min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2} \quad & \mathbb{E}_{P^*}[H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi})] \qquad (3.16)\\
\text{s.t. } & \mathbf{\Lambda}_1,\mathbf{\Lambda}_2 \succeq 0,
\end{aligned}
$$

where

$$H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi}) := (1+\epsilon_2)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda},\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_+ - (1-\epsilon_1)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_- + \lambda_0. \qquad (3.17)$$

According to Theorem 8 in the Appendix, the sample average approximation (SAA) techniques can be used to estimate $\mathbb{E}_{P^*}[\cdot]$ in (3.16) and will guarantee the almost surely convergence of both the objective value and the optimal solution set. We refer [17, 20] for more details about SAA. Given a finite sample $\Omega := \{\boldsymbol{\xi}^1,\dots,\boldsymbol{\xi}^K\}$, the SAA of the problem (3.16) can be written as:

$$
\begin{aligned}
\min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2} \quad & \{h(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) := \frac{1}{K}\sum_{k=1}^{K} H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi}^k)\} \qquad (3.18)\\
s.t. \; & \mathbf{\Lambda}_1,\mathbf{\Lambda}_2 \succeq 0.
\end{aligned}
$$

We have the following theorem:

Theorem 4 *The SAA formulation* (3.18) *is equivalent to the linear conic programming problem:*

$$
\begin{aligned}
\min_{\mathbf{x},s_1^k,s_2^k,\lambda_0,\boldsymbol{\Lambda_1},\boldsymbol{\Lambda_2}} \ & \frac{1}{K}\sum_{k=1}^{K}[(1+\epsilon_2)s_2^k-(1-\epsilon_1)]s_1^k]+\lambda_0 \qquad (3.19)\\
\text{s.t. } & \| \, (s_2^k+1, s_1^k-1, \lambda_0+1, ((\boldsymbol{\xi}^k-\boldsymbol{\mu})(\boldsymbol{\xi}^k-\boldsymbol{\mu})^T-\beta\mathbf{Q})\bullet\boldsymbol{\Lambda}_1+1,\\
& \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix}\bullet\boldsymbol{\Lambda}_2+1) \, \|^2\\
& \geq \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^k)\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2+(s_2^k)^2+(s_1^k)^2+\lambda_0^2\\
& \quad +[((\boldsymbol{\xi}^k-\boldsymbol{\mu})(\boldsymbol{\xi}^k-\boldsymbol{\mu})^T-\beta\mathbf{Q})\bullet\boldsymbol{\Lambda}_1]^2+[\begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix}\bullet\boldsymbol{\Lambda}_2]^2+5,\\
& \quad \text{for } k=1,\ldots,K,\\
& s_1^k, s_2^k \geq 0 \ \text{ for } k=1,\ldots,K,\\
& \boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2 \succeq 0.
\end{aligned}
$$

Proof Consider the optimization problem:

$$
\begin{aligned}
\min_{\mathbf{x},s_1^k,s_2^k,\lambda_0,\boldsymbol{\Lambda_1},\boldsymbol{\Lambda_2}} \ & \frac{1}{K}\sum_{k=1}^{K}[(1+\epsilon_2)s_2^k-(1-\epsilon_1)]s_1^k]+\lambda_0 \qquad (3.20)\\
\text{s.t. } & s_2^k-s_1^k \geq \mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi}^k) \ \text{ for } k=1,\ldots,K,\\
& s_1^k, s_2^k \geq 0 \ \text{ for } k=1,\ldots,K,\\
& \boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2 \succeq 0,
\end{aligned}
$$

where $\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}$ is defined in (3.4). We first claim that the optimal solution $(\hat{\mathbf{s}}_1,\hat{\mathbf{s}}_2,\hat{\mathbf{x}},\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2)$ of (3.20) satisfies

$$\hat{s}_2^k=[\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)]_+, \hat{s}_1^k=[\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)]_-, \text{for } \forall k\in\{1,\ldots,K\},$$

where $\hat{\mathbf{s}}_i=(\hat{s}_i^1,\ldots,\hat{s}_i^K)$ for $i=1,2$. Assume $\exists k\in\{1,\ldots,K\}$ such that $\hat{s}_2^k-\hat{s}_1^k>\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)$, then assume $\hat{s}_2^k-\hat{s}_1^k-\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)=\delta^k>0$. Then, the solution $(\hat{\mathbf{s}}_1,\hat{\mathbf{s}}_2-\delta^k\mathbf{e}^k,\hat{\mathbf{x}},\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2)$ with $\mathbf{e}^k$ to be the K-dimensional vector with 1 in the kth entry and 0 otherwise is a feasible solution which gives a smaller objective value. Therefore, $\hat{s}_2^k-\hat{s}_1^k=\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)$. This implies

$$\hat{s}_2^k\geq[\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)]_+, \hat{s}_1^k\geq[\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)]_-.$$

Now assume $\exists k\in\{1,\ldots,K\}$ such that $\hat{s}_2^k-[\mathcal{L}_{\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2}(\hat{\mathbf{x}},\boldsymbol{\xi}^k)]_+=\zeta^k>0$, then it is easy to check that the solution $(\hat{\mathbf{s}}_1-\zeta^k\mathbf{e}^k,\hat{\mathbf{s}}_2-\zeta^k\mathbf{e}^k,\hat{\mathbf{x}},\hat{\lambda}_0,\hat{\boldsymbol{\Lambda}}_1,\hat{\boldsymbol{\Lambda}}_2)$ will give us a smaller objective value, which verifies the claim and proves that (3.20) is equivalent to (3.18).

On the other hand, the first constraint of (3.20) can be rewritten as:

$$s_2^k - s_1^k \geq \left|\left|(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}}^k)\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 - \lambda_0 - ((\boldsymbol{\xi}^k - \boldsymbol{\mu})(\boldsymbol{\xi}^k - \boldsymbol{\mu})^T - \beta\mathbf{Q}) \bullet \boldsymbol{\Lambda}_1$$
$$+ \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi} - \boldsymbol{\mu}) \\ (\boldsymbol{\xi} - \boldsymbol{\mu})^T & \alpha \end{pmatrix} \bullet \boldsymbol{\Lambda}_2 \ \text{ for } \forall k = 1, \ldots, K$$

which is equivalent to the first conic constraint in (3.19). Therefore, the formulation (3.20) is equivalent to the linear conic programming problem (3.19). ▮

3.2.3 Measure bounds given by two probability measures

In Section 3.2.2, we analyze the DRLS problem with ambiguity set (3.1) and measure bounds defined with one given probability measure. In this section, we will consider another important case by defining ν_1 and ν_2 as follows:

$$\nu_1 := (1 - \epsilon_1)\underline{P}^*, \ \nu_2 := (1 + \epsilon_2)\bar{P}^*, \tag{3.21}$$

where $\underline{P}^*$ and $\bar{P}^*$ are two given probability measure with known density functions $\underline{f}(\cdot)$ and $\bar{f}(\cdot)$. The condition $\nu_1 \preceq \nu_2$ tells us $(1 - \epsilon_1)\underline{f}(\boldsymbol{\xi}) \leq (1 + \epsilon_2)\bar{f}(\boldsymbol{\xi})$ for $\forall \boldsymbol{\xi} \in \mathscr{S}_\rho$. Then the problem (3.9) can be written as:

$$\begin{aligned} \min_{\mathbf{x}, \lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2} \ & \int_{\mathscr{S}_\rho} [(1 + \epsilon_2)\bar{f}(\boldsymbol{\xi})[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_+ - (1 - \epsilon_1)\underline{f}(\boldsymbol{\xi})[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_-]d\boldsymbol{\xi} + \lambda_0 \\ & \text{s.t. } \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \succeq 0. \end{aligned} \tag{3.22}$$

Since for $\forall \boldsymbol{\xi} \in \mathscr{S}_\rho$, the function $(1 + \epsilon_2)\bar{f}(\boldsymbol{\xi})[\cdot]_+ - (1 - \epsilon_1)\underline{f}(\boldsymbol{\xi})[\cdot]_-$ is a convex increasing function, we know that

$$(1 + \epsilon_2)\bar{f}(\boldsymbol{\xi})[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_+ - (1 - \epsilon_1)\underline{f}(\boldsymbol{\xi})[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_-$$

is a convex function with respect to the decision variables $\mathbf{x}, \lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2$.

The integration in (3.22) can be considered as an expectation with the uniform distribution over the sample space $\mathscr{S}_\rho$. With an argument similar to that in Section 3.2.2, we can use the SAA formulation (3.23) to solve (3.22) as:

$$\begin{aligned} \min_{\mathbf{x}, \lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2} \ & \frac{1}{K}\sum_{k=1}^{K}[(1 + \epsilon_2)\bar{f}(\boldsymbol{\xi}^k)[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi}^k)]_+ - (1 - \epsilon_1)\underline{f}(\boldsymbol{\xi}^k)[\mathcal{L}_{\lambda_0, \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi}^k)]_-] + \lambda_0 \\ & \text{s.t. } \boldsymbol{\Lambda}_1, \boldsymbol{\Lambda}_2 \succeq 0, \end{aligned} \tag{3.23}$$

where $\{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\} \subset \mathscr{S}_\rho$ is a given sample. We have the following theorem.

Theorem 5 *The SAA formulation* (3.23) *is equivalent to the linear conic programming problem:*

$$
\begin{aligned}
\min_{\mathbf{x},s_1^k,s_2^k,\lambda_0,\mathbf{\Lambda_1},\mathbf{\Lambda_2}} \quad & \frac{1}{K}\sum_{k=1}^{K}[(1+\epsilon_2)\bar{f}(\boldsymbol{\xi}^k)s_2^k-(1-\epsilon_1)\underline{f}(\boldsymbol{\xi}^k)s_1^k]+\lambda_0 \qquad (3.24)\\
\text{s.t.} \quad & \|\ (s_2^k+1, s_1^k-1, \lambda_0+1, ((\boldsymbol{\xi}^k-\boldsymbol{\mu})(\boldsymbol{\xi}^k-\boldsymbol{\mu})^T-\beta\mathbf{Q})\bullet\mathbf{\Lambda}_1+1,\\
& \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix}\bullet\mathbf{\Lambda}_2+1)\ \|^2\\
& \geq \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^k)\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2+(s_2^k)^2+(s_1^k)^2+\lambda_0^2\\
& \quad +[((\boldsymbol{\xi}^k-\boldsymbol{\mu})(\boldsymbol{\xi}^k-\boldsymbol{\mu})^T-\beta\mathbf{Q})\bullet\mathbf{\Lambda}_1]^2+[\begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix}\bullet\mathbf{\Lambda}_2]^2+5,\\
& \quad \text{for } k=1,\ldots,K,\\
& s_1^k, s_2^k \geq 0 \ \text{ for } k=1,\ldots,K,\\
& \mathbf{\Lambda}_1, \mathbf{\Lambda}_2 \succeq 0.
\end{aligned}
$$

Proof The proof is the same as the proof of Theorem 4. ▮

Remark. The two cases in Section 3.2.2 and Section 3.2.3 are similiar. There is one important difference, however. The case in Section 3.2.2 is considering the measure bounds defined by one reference probability measure. Here we do not require the knowledge of the density function of the reference probability measure P^*. The only requirement is the ability to sample from P^*. The case in Section 3.2.3 considers the bounds defined by two different probability measures. But we require the specific formulations of their density functions. Even though the above three cases do not cover all the possibilities, they do cover a lot of interesting cases. The first two cases are considered by Shapiro and Ahmed [19] in a more general framework. However, because of the generality, they show that the problem can be solved by the subgradient method. According to our specific least squares formulation, we further show that the two cases are actually have equivalent linear conic programming formulations, which can be solved efficiently by some existing software packages, such as SeDuMi [21].

4 DRLS with C.I. by Probability Metric

In this section we consider a more general probability ambiguity set with the form

$$
\mathscr{P} := \{P : d(P, P^*) \leq \epsilon\}, \qquad (4.1)
$$

where P^* is a reference probability measure, d is some distance for probability measures and $\epsilon > 0$ is a given constant. Note that every probability measure $P \in \mathscr{P}$ is defined on the same sample space Ω with σ-algebra $\mathcal{F}$. Pflug and Wozabal [12] consider this type of ambiguity in portfolio optimization. They point out some different ambiguity sets by choosing different distance functions d. In this paper, we will choose the Kantorovich distance (or L_1 distance) [23]. The Kantorovich distance between two probability measures P_1 and P_2 is defined as:

$$d(P_1,P_2) := \sup_f \left\{ \int f(\mathbf{u})dP_1(\mathbf{u}) - \int f(\mathbf{u})dP_2(\mathbf{u}), |f(\mathbf{u})-f(\mathbf{v})| \leq ||\mathbf{u}-\mathbf{v}||_1 \text{ for all } \mathbf{u},\mathbf{v} \right\}.$$

According to the Kantorovich-Rubinstein theorem [13], the Kantorivich ambiguity set (4.1) can be represented as:

$$\begin{aligned} \mathscr{P} := \{P : \text{there is a bivariate probability } K(\cdot,\cdot) \text{ such that } \int_{\mathbf{v}} K(\mathbf{u},d\mathbf{v}) = P(\mathbf{u}); \qquad (4.2) \\ \int_{\mathbf{u}} K(d\mathbf{u},\mathbf{v}) = P^*(\mathbf{v}); \ \int_{\mathbf{u}}\int_{\mathbf{v}} ||\mathbf{u}-\mathbf{v}||_1 K(d\mathbf{u},d\mathbf{v}) \leq \epsilon\}. \end{aligned}$$

In our case, we assume that the sample space is finite, i.e $\Omega := \{\boldsymbol{\xi}^1,\ldots,\boldsymbol{\xi}^N\}$. Then a probability measure is give by a N-dimensional vector $\mathbf{p} := (p_1,\ldots,p_N)$. Given the reference probability measure P^* as $\mathbf{p}^* := (p_1^*,\ldots,p_N^*)$, the probability ambiguity set (4.2) is equivalent to:

$$\mathscr{P} := \{\mathbf{p} := (p_1,\ldots,p_N) : p_i = \sum_{j=1}^N k_{i,j}, \sum_{i=1}^N k_{i,j} = p_j^*, k_{i,j} \geq 0, \sum_{i=1}^N\sum_{j=1}^N \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 k_{i,j} \leq \epsilon\}.$$

The DRLS problem with ambiguity set (4.1) is now written as:

$$\begin{aligned} \min_{\mathbf{x}} \max_{p_i,k_{i,j},i,j=1,\ldots,N} & \sum_{i=1}^N p_i \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \qquad (4.3) \\ \text{s.t. } & p_i = \sum_{j=1}^N k_{i,j} \quad \text{for } i=1,\ldots,N, \\ & \sum_{i=1}^N k_{i,j} = p_j^* \quad \text{for } j=1,\ldots,N, \\ & \sum_{i=1}^N\sum_{j=1}^N \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 k_{i,j} \leq \epsilon, \\ & p_i \geq 0 \quad \text{for } i=1,\ldots,N, \\ & k_{i,j} \geq 0 \quad \text{for } i,j=1,\ldots,N. \end{aligned}$$

We have the following theorem

Theorem 6 *The DRLS problem with ambiguity set is equivalent to the linear conic optimization problem:*

$$\begin{aligned} \min_{\mathbf{x},s_i,t_j,\sigma,i,j=1,\ldots,N} & \sum_{j=1}^N p_j^* t_j + \epsilon\sigma \qquad (4.4) \\ \text{s.t. } & s_i \geq \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \quad \text{for } i=1,\ldots,N, \\ & -s_i + t_j + \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 \sigma \geq 0 \quad \text{for } i,j=1,\ldots,N, \\ & \sigma \geq 0. \end{aligned}$$

Proof Given an $\mathbf{x}$, the inner problem has the form:

$$\begin{aligned}
\max_{p_i,k_{i,j},i,j=1,\dots,N} \; & \sum_{i=1}^{N} p_i \left|\left|(\mathbf{A}+\boldsymbol{\xi}^i_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}^i_{\mathbf{b}})\right|\right|^2 \qquad (4.5)\\
\text{s.t. } & p_i = \sum_{j=1}^{N} k_{i,j} \quad \text{for } i=1,\dots,N,\\
& \sum_{i=1}^{N} k_{i,j} = p^*_j \quad \text{for } j=1,\dots,N,\\
& \sum_{i=1}^{N}\sum_{j=1}^{N} \left|\left|\boldsymbol{\xi}^i-\boldsymbol{\xi}^j\right|\right|_1 k_{i,j} \le \epsilon,\\
& k_{i,j} \ge 0 \quad \text{for } i,j=1,\dots,N,\\
& p_i \ge 0 \quad \text{for } i=1,\dots,N,
\end{aligned}$$

which is a standard linear optimization problem. Since $\mathbf{p}^*$ is a feasible solution and the objective is bounded for one given $\mathbf{x}$, the strong duality holds. We can write the dual of (4.5) as:

$$\begin{aligned}
\min_{s_i,t_j,\sigma,i,j=1,\dots,N} \; & \sum_{j=1}^{N} p^*_j t_j + \epsilon\sigma \qquad (4.6)\\
\text{s.t. } & s_i \ge \left|\left|(\mathbf{A}+\boldsymbol{\xi}^i_{\mathbf{A}})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}^i_{\mathbf{b}})\right|\right|^2 \quad \text{for } i=1,\dots,N,\\
& -s_i + t_j + \left|\left|\boldsymbol{\xi}^i-\boldsymbol{\xi}^j\right|\right|_1 \quad \text{for } i,j=1,\dots,N,\\
& \sigma \ge 0,
\end{aligned}$$

Combine (4.6) with the outer problem and we can get the desired equivalent formulation (4.4). ∎

Remark. In [12], Pflug and Wozabal analyze the portfolio selection problem with the ambiguity set defined by a probability confidence set with Kantorovish distance. Because of the specific property of portfolio selection problem, their problem has a linear objective. At the same time, they add additional constraints upon the ambiguity, i.e. the lower bound of the expected return, which implies a certain level of generality of the constraints. The solution method for their problem is called successive convex programming (SCP). The basic idea is to start from a simple ambiguity set, find probability measures which violate the constraints and add those probability measures to the ambiguity set. This idea is similar as the cutting plane method in convex optimization. Compared with their methodology, the DRLS problem with ambiguity set defined by confidence set with Kantorovish distance has a nice structure, which leads to an exact equivalent linear conic formulation (4.6).

5 Conclusions Remarks

We have presented a distributionally robust least squares framework with three different probability ambiguity sets. They include ambiguity sets defined by i) confidence intervals of the first two moments, ii) moment confidence intervals with measure bounds and iii) confidence interval defined by using Kantorovich distance. For the first case, we show that the equivalent semi-infinite programming formulation can be solved efficiently by using ellipsoid method or Vaidya's volumetric cutting plane method with oracle determined by a trust-region subproblem. The second case is shown to be equivalent as a convex stochastic programming problem. Especially, the SAA formulation can be proved to have a linear conic equivalent formulation. For the third case, we consider the finite support case and show that the DRLS problem is equivalent to a linear conic programming problem. The linear conic equivalent formulations of the latter two cases make the DRLS problem more attractive because it is easy to apply in practice, i.e. using SeDuMi [21] to solve the problem.

References

[1] D. P. Bertsekas, *Convex Analysis and Optimization*, Athena Scientific, 2003.

[2] D. Bertsimas, X. V. Doan, K. Natarajan, and C. Teo, *Models for minimax stochastic linear optimization problems with risk aversion*, Mathematics of Operations Research, 35 (2010), pp. 580–602.

[3] E. Delage and Y. Ye, *Distributionally robust optimization under moment uncertainty with application to data-driven problems*, Operations Research, published online (2008).

[4] J. Dupacova, *The minimax approach to stochastic programming and an illustrative application*, Stochastics, 20 (1987), pp. 73–88.

[5] C. Fortin and H. Wolkowicz, *The trust region subproblem and semidefinite programming*, Optimization Methods and Software, 19 (2004), pp. 41–67.

[6] L. E. Ghaoui and H. Lebret, *Robust solution to least-squares problem with uncertain data*, SIAM Journal on Matrix Analysis and Applications, 18 (1997), pp. 1035–1064.

[7] W. H. Greene, *Econometric Analysis*, Prentice Hall, 6th edition ed., 2007.

[8] L. L. Grotschel, M. and A. Schrijver, *The ellipsoid method and its consequences in combinatorial optimization*, Combinatorica, 1 (1981), pp. 169–197.

[9] K. Isii, *On the sharpness of chebyshev-type inequalities*, Annals of the Institute of Statistical Mathematics, 14 (1963), pp. 185–197.

[10] J. Lasserre, *Moments, Positive Polynomials and Their Applications*, Imperial College Press, London, 2010.

[11] O. Lyandres, R. P. Van Duyne, J. T. Walsh, M. R. Glucksberg, and S. Mehrotra, *Prediction range estimation from noisy raman spectra with robust optimization*, Analyst, 135 (2010), pp. 2111–2118.

[12] G. Pflug and D. Wozabal, *Ambiguity in portfolio selection*, Quantitative Finance, 7 (2007), pp. 435–442.

[13] S. T. Rachev, *Probability metrics and the stability of stochastic models*, Willey Series in Probability and Mathematical Statistics, Wiley, 1991.

[14] C. R. Rao, *The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves*, Biometrika, 52 (1965), pp. 447–458.

[15] F. Rendl and H. Wolkowicz, *A semidefinite framework for trust region subproblems with applications to large scale minimization*, Mathematical Programming, 77 (1997), pp. 273–299.

[16] R. Rockafellar, *Conjugate Duality and Optimization*, CBMS-NSF Regional Conference Series in Applied Mathematics 16, SIAM, Philadelphia, 1974.

[17] A. Ruszczynski and A. Shapiro, eds., *Stochastic Programming*, Handbooks in Operations Research and Management Science 10, North-Holland, Amsterdam, 2003.

[18] H. Scarf, *A min-max solution of an inventory problem, K. S. Arrow*, Stanford University Press, Stanford, CA, Stanford, CA, 1958, pp. 201–209.

[19] A. Shapiro and S. Ahmed, *On a class of minimax stochastic programs*, SIAM Journal of Optimization, 14 (2004), pp. 1237–1249.

[20] A. Shapiro and T. H. de mello, *On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs*, SIAM Journal on Optimization, 11 (2000), pp. 70–86.

[21] J. F. Sturm, *Using SeDuMi 1.02, a MATLAB* toolbox for optimization over symmetric cones*, Optim. Methods Softw., 11-12 (1999), pp. 625–653.

[22] ——, *Implementation of interior point methods for mixed semidefinite and second order cone optimization problems*, Optm. Methods Softw., 17 (2002), pp. 1105–1154.

[23] S. S. Vallander, *Calculation of the wasserstein distance between probability distributions on the line*, Theory of Probability and Its Applications, 18 (1973), pp. 784–786.

6 Appendix

In the appendix, some known theorems are summarized. Consider the distributional optimization problem in the form:

$$\begin{aligned} Z_{primal} := \max_{\nu\in\mathscr{B}} & \int_{\mathscr{S}} \phi_0(\mathbf{q})d\nu(\mathbf{q}) \\ \text{s.t.} & \int_{\mathscr{S}} \phi_i(\mathbf{q})d\nu(\mathbf{q}) = b_i, \text{for } \forall i = 1, \dots, M, \end{aligned} \tag{6.1}$$

where $\mathscr{B}$ is the set consisting all nonnegative measures on $\mathscr{S}$ with respect to each of which $\phi_0, \dots, \phi_M$ are measurable and integrable.

Theorem 7 (Isii 1963) *Let* $\mathcal{M} := \{(m_1, \ldots, m_M) | m_i = \int_{\mathscr{S}} \phi_i(\mathbf{q}) d\nu(\mathbf{q}), \nu \in \mathscr{B}\}$. *If* $(b_1, \ldots, b_M)$ *is an interior point of* $\mathcal{M}$*, then we have:*

$$\begin{aligned} Z_{primal} = Z_{dual} := \inf_{\boldsymbol{\lambda}} & \sum_{i=1}^{M} b_i \lambda_i \\ s.t. & \sum_{i=1}^{M} \lambda_i \phi_i(\mathbf{q}) \geq \phi_0(\mathbf{q}), \quad \forall \mathbf{q} \in \mathscr{S} \end{aligned} \tag{6.2}$$

In Section 3, we use SAA method to approximately solve the stochastic programming problem. The convergence results for the SAA method from [17] are summarized here. Consider the stochastic programming problems of the form

$$\min_{\mathbf{x} \in \mathcal{X}} \{f(\mathbf{x}) := \mathbb{E}_P[F(\mathbf{x}, \boldsymbol{\xi})]\}, \tag{6.3}$$

where $F(\mathbf{x}, \boldsymbol{\xi})$ is a function of two vector variables $\mathbf{x} \in \mathbb{R}^n$ and $\boldsymbol{\xi} \in \mathbb{R}^d$, $\mathcal{X} \subset \mathbb{R}^n$ is a given set and $\boldsymbol{\xi} = \boldsymbol{\xi}(\omega)$ is a random vector. The expectation in (6.3) is taken with respect to the probability distribution of $\boldsymbol{\xi}$ which is assumed to be known as P. Denote by $\Xi \subset \mathbb{R}^d$ the support of the probability distribution of $\boldsymbol{\xi}$, that is, Ξ is the smallest closed set in $\mathbb{R}^d$ such that the probability of the event $\boldsymbol{\xi} \in \mathbb{R}^d \setminus \Xi$ is zero. Also denote by $\mathbb{P}(A)$ the probability of an event A. With the generated sample $\xi^1, \ldots, \xi^K$, we associate the sample average function

$$\hat{f}_K(\mathbf{x}) := \frac{1}{K} \sum_{i=1}^{K} F(\mathbf{x}, \xi^i). \tag{6.4}$$

The original stochastic programming problem (6.3) is approximated by the optimization problem

$$\min_{\mathbf{x} \in \mathcal{X}} \{\hat{f}_K(\mathbf{x}) := \frac{1}{K} \sum_{i=1}^{K} F(\mathbf{x}, \xi^i)\}. \tag{6.5}$$

Before giving the theorem of the statistical properties of the SAA formulation, we need to define some notations. Let the optimal value of the original problem (6.3) be ν and it's optimal solution set be S. Let $\hat{\nu}_K$ and $\hat{S}_K$ be the optimal value and the set of optimal solutions of the SAA problem (6.5). For sets $A, B \subset \mathbf{R}^n$, denote $\text{dist}(x, A) := \inf_{x' \in A} ||x - x'||$ to be the distance from $x \in \mathbb{X}^n$ to A, and

$$\mathbb{D}(A, B) := \sup_{x \in A} \text{dist}(x, B) \tag{6.6}$$

Also, define the function $(\mathbf{x}, \xi) \mapsto F(\mathbf{x}, \xi)$ to be a random lower semicontinuous function if the associated epigraphical multifunction $\xi \mapsto \text{epi}F(\cdot, \xi)$ is closed valued and measurable. We say that the Law of Large Numbers (LLN) holds, for $\hat{f}_K(\mathbf{x})$, pointwise if $\hat{f}_K(\mathbf{x})$ converges w.p.1 to $f(\mathbf{x})$, as $K \to \infty$, for any fixed $\mathbf{x} \in \mathbb{R}^n$. The convergence theorem is as follows:

Theorem 8 *Suppose that:* (i) *the integrand function* F *is random lower semicontinuous,* (ii) *for almost every* $\xi \in \Xi$ *the function* $F(\cdot, \xi)$ *is convex,* (iii) *the set* $\mathcal{X}$ *is closed and convex,* *(iv) the expected value function f is lower semicontinuous and there exists a point $\hat{\mathbf{x}} \in \mathcal{X}$ such that $f(\mathbf{x}) \leq +\infty$ for all x in a neighborhood of $\hat{\mathbf{x}}$,* (v) *the set S of optimal solutions of the original problem* (6.3), (vi) *the LLN holds pointwise. Then $\hat{\nu}_K \to \nu^*$ and $\mathbb{D}(\hat{S}_K, S) \to 0$ w.p.1 as $K \to \infty$.*

There are also results about the exponential convergence rate of the SAA method. Please refer to [17, 20] for details of such results.