Math. Program., Ser. A (2014) 146:123–141
DOI 10.1007/s10107-013-0681-9

FULL LENGTH PAPER

Models and algorithms for distributionally robust least squares problems

Sanjay Mehrotra · He Zhang

Received: 15 February 2011 / Accepted: 22 April 2013 / Published online: 8 May 2013

Abstract We present three different robust frameworks using probabilistic ambiguity descriptions of the data in least squares problems. These probability ambiguity descriptions are given by: (1) confidence region over the first two moments; (2) bounds on the probability measure with moments constraints; (3) the Kantorovich probability distance from a given measure. For the first case, we give an equivalent formulation and show that the optimization problem can be solved using a semidefinite optimization reformulation or polynomial time algorithms. For the second case, we derive the equivalent Lagrangian problem and show that it is a convex stochastic programming problem. We further analyze three special subcases: (i) finite support; (ii) measure bounds by a reference probability measure; (iii) measure bounds by two reference probability measures with known density functions. We show that case (i) has an equivalent semidefinite programming reformulation and the sample average approximations of case (ii) and (iii) have equivalent semidefinite programming reformulations. For ambiguity description (3), we show that the finite support case can be solved by using an equivalent second order cone programming reformulation.

Mathematics Subject Classification 93E24 Least squares and related methods · 93B35 Sensitivity (robustness) · 62G35 Robustness · 62J05 Linear regression · 62J07 Ridge regression; shrinkage estimators · 90C15 Stochastic programming · 90C22 Semidefinite programming

Research partially supported by NSF grant CMMI-1100868 and ONR grant N00014-09-10518 and N00014210051.

S. Mehrotra (✉) · H. Zhang
Department of Industrial Engineering and Management Sciences, Northwestern University, Evanston, IL 60201, USA
e-mail: mehrotra@iems.northwestern.edu

1 Introduction

The ordinary least squares (OLS) problem [9] 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 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 the same experimental conditions do not generate the same output [10], 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 reformulations and algorithms for solving them. One possible approach is to consider the stochastic least squares problem:

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

where $\mathbb{P}$ is a given known distribution of the observation data $\mathbf{A}$ and $\mathbf{b}$. Models of this type have been considered before (for example, Rao [13]). In practice, however, we may not have full knowledge of this probability distribution. One possible way to handle 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 [8] develop this idea by considering the min–max optimization problem

$$\min_{\mathbf{x}} \max_{||\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}||_F \le \rho} ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 . \tag{1.3}$$

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

$$r(\mathbf{A}, \mathbf{b}, \rho, \mathbf{x}) := \max_{||\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}||_F \le \rho} ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 \tag{1.4}$$

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 $||\boldsymbol{\xi}_{\mathbf{A}}, \boldsymbol{\xi}_{\mathbf{b}}||_F \le \rho$. It only considers the worst case 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 $\mathscr{S}_\rho$. This leads to a distributionally robust least squares (DRLS) frameworks as

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

where $\mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m, \boldsymbol{\xi} \in \mathbb{R}^{m(n+1)}$ 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 $\boldsymbol{\xi}$. In this paper we study (1.6) under the following three definitions of the ambiguity set $\mathscr{P}$.

(1) Moment robust set:

$$\begin{aligned}\mathscr{P} := \mathscr{D} = \{\nu : \mathbb{E}_\nu[\mathbf{1}] = \mathbf{1}, (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu}) \le \alpha, \\ \mathbb{E}_\nu[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T] \preceq \beta \mathbf{Q}, \\ \boldsymbol{\xi} \in \mathscr{S}_\rho = \{\boldsymbol{\xi} : ||\boldsymbol{\xi}|| \le \rho\}\}.\end{aligned} \tag{1.7}$$

(2) Moment robust with bounds on measure:

$$\begin{aligned}\mathscr{P} := \mathscr{M} = \{\nu \in \mathcal{X}_\rho : \nu(\mathscr{S}_\rho) = 1,\ \nu_1 \preceq \nu \preceq \nu_2, (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu}) \le \alpha, \\ \mathbb{E}_\nu[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T]\} \preceq \beta \mathbf{Q}\},\end{aligned} \tag{1.8}$$

where ν_1, ν_2 are two given measures defined on the sample space $\mathscr{S}_\rho$ with Borel σ-algebra $\mathcal{B}_{\mathscr{S}_\rho}$ and $\mathcal{X}_\rho$ is the space of all finite measures on $(\mathscr{S}_\rho, \mathcal{B}_{\mathscr{S}_\rho})$. Note that for two measures μ, ν defined on $(\mathscr{S}_\rho, \mathcal{B}_{\mathscr{S}_\rho})$, we write $\mu \preceq \nu$ if $\mu(C) \le \nu(C)$ for $\forall C \in \mathcal{B}_{\mathscr{S}_\rho}$.

(3) Kantorovich set:

$$\mathscr{P} := \mathscr{K} = \{\mathbb{P} : d(\mathbb{P}, \mathbb{P}^*) \le \epsilon\}, \tag{1.9}$$

where $\mathbb{P}^*$ is a given reference probability measure, d is the Kantorovich distance for probability measures and ϵ is a given constant.

Note that for case (1) and (2), α, β, $\boldsymbol{\mu}$, $\mathbf{Q}$ and ρ are given parameters. Define

$$r(\mathbf{x}, \Theta) := \max_{\mathbb{P}\in\mathscr{P}} \mathbb{E}_{\mathbb{P}} \left\{ ||(\mathbf{A} + \boldsymbol{\xi}_\mathbf{A})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_\mathbf{b})||^2 \right\} \tag{1.10}$$

to be the inner problem, where Θ represents the parameters used to define the ambiguity set $\mathscr{P}$. The outer problem is

$$\min_{\mathbf{x}} r(\mathbf{x}, \Theta). \tag{1.11}$$

The idea of distributionally robust optimization was originated in Scarf [17]. Distributionally robust optimization has gained significant interest recently. Dupacova [6], Bertsimas et al. [3], and Delage and Ye [5] used linear or conic constraints to describe $\mathscr{P}$ with moments. Shapiro and Ahmed [18] defined a probability ambiguity set with measure bounds and general moment constraints. Pflug and Wozabal [11] considered the probability ambiguity set defined by confidence regions over a reference probability measure. Bertsimas et al. [3] used a piece-wise linear utility with first and second moment equality constraints and showed that the corresponding problem has semidefinite programming reformulations. Delage and Ye [5] have given general conditions for polynomial time solvability of a generic distributionally robust model. Shapiro and Ahmed [18] gave stochastic programming reformulations of their model.

In this paper we study the distributionally robust least squares problem with the ambiguity sets $\mathscr{P}$ defined in (1.7)–(1.9). The first and second moment constraint case (1.7) is useful because the moment estimates are the most common statistics in practice. The reference measure bound constraints (1.8) 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 Kantorovich distance constraint (1.9) 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 be reformulated as a conic (semidefinite or second order cone) optimization problem. In particular, in Sect. 2, we analyze the moment robust least squares problem with ambiguity set (1.7) and prove that this problem can be solved by a cutting plane method. An equivalent semidefnite reformulation of this problem is also given. In Section 3, we study the distributionally robust least squares problem with ambiguity set (1.8) and show that the problem can be reformulated as a conic programming problem for three special cases, when sample average approximation is used to approximate the problem. In Sect. 4, we study the DRLS problem with ambiguity set (1.9) with discrete support and show that this problem has an equivalent second order cone reformulation. Section 5 has concluding remarks.

2 Moment robust least squares

In this section we study the moment robust least squares (MRLS) model with ambiguity set (2). Delage and Ye [5] describe a general moment robust framework with the probability ambiguity set defined by confidence region for the first two moments in (1.7) as follows:

$$\begin{aligned}\mathscr{D} = \{\nu : \mathbb{E}_\nu[\mathbf{1}] = \mathbf{1}, (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_\nu[\boldsymbol{\xi}] - \boldsymbol{\mu}) \leq \alpha, \\ \mathbb{E}_\nu[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T] \} \preceq \beta \mathbf{Q}, \boldsymbol{\xi} \in \mathscr{S}\}. \end{aligned} \tag{2.1}$$

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

Lemma 1 (Delage and Ye [5]) *Given an feasible* $\mathbf{x}$*, consider the moment robust problem defined as:*

$$\max_{\nu \in \mathscr{D}} \mathbb{E}_\nu[h(\mathbf{x}, \boldsymbol{\xi})], \tag{2.2}$$

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* $\alpha \geq 0, \beta \geq 1, \mathbf{Q} \succ 0$ *and* $h(\mathbf{x}, \boldsymbol{\xi})$ *is* ν*-integrable for all* $\nu \in \mathscr{D}$*. Then, the optimal value of the inner problem (2.2) is equal to the optimal value of the problem:*

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{y}, y_0, t} \quad & y_0 + t \\ s.t. \quad & y_0 \geq h(\mathbf{x}, \boldsymbol{\xi}) - \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} - \boldsymbol{\xi}^T \mathbf{y} \ \ \forall \boldsymbol{\xi} \in \mathscr{S}, \end{aligned} \tag{2.3}$$

$$t \geq (\beta \mathbf{Q} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + \sqrt{\alpha} \left|\left| \mathbf{Q}^{\frac{1}{2}} (\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) \right|\right|,$$
$$\mathbf{Y} \succeq 0.$$

Additionally, if we assume:

(i) The sample space $\mathscr{S} \subset \mathbb{R}^m$ is convex and compact, and it is equipped with an oracle that for any $\boldsymbol{\xi} \in \mathbb{R}^m$ can either confirm that $\boldsymbol{\xi} \in \mathscr{S}$ or provide a hyperplane that separates $\boldsymbol{\xi}$ from $\mathscr{S}$ in polynomial time.

(ii) The function $h(\mathbf{x}, \boldsymbol{\xi})$ is concave in $\boldsymbol{\xi}$. In addition, given $\boldsymbol{\xi}$, in polynomial time one can:

1. *evaluate the value of $h(\mathbf{x}, \boldsymbol{\xi})$;*
2. *find a supergradient of $h(\mathbf{x}, \boldsymbol{\xi})$ in $\boldsymbol{\xi}$.*

Then problem (2.3) can be solved in polynomial time using the ellipsoid method.

Now we make the following assumption.

Assumption 1 Assume the ambiguity set $\mathscr{P}$ defined as (1.7) is nonempty, i.e. $\exists$ probability measure $\mathbb{P}$ on $\mathscr{S}_\rho$ such that $(\mathbb{E}_{\mathbb{P}}[\boldsymbol{\xi}] - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\mathbb{E}_{\mathbb{P}}[\boldsymbol{\xi}] - \boldsymbol{\mu}) \leq \alpha$ and $\mathbb{E}_{\mathbb{P}}[(\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T] \} \preceq \beta \mathbf{Q}$.

In this section we assume that Assumption 1 is satisfied. Consider the DRLS problem (1.6) with probability ambiguity set $\mathscr{P}$ defined in (1.7). For the inner problem (1.10) with $\Theta = \{\boldsymbol{\mu}, \boldsymbol{\xi}, \alpha, \beta, \rho\}$ and $\mathscr{P}$ defined as (1.7), we can apply Lemma 1 by letting: $h(\mathbf{x}, \boldsymbol{\xi}) = ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2$. Consequently, we have an equivalent formulation of the inner problem (1.10) given in the following theorem.

Theorem 1 *For a given fixed $\mathbf{x} \in \mathbb{R}^n$, let $\alpha \geq 0, \beta \geq 1, \mathbf{Q} \succ 0$. Then, the optimal value of the inner problem (1.10) is equal to the optimal value of the problem:*

$$\begin{aligned} \min_{\mathbf{Y}, \mathbf{y}, y_0, t} \quad & (\beta \mathbf{Q} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y} + y_0 + \sqrt{\alpha} t \qquad (2.4) \\ s.t. \quad & \boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} + \boldsymbol{\xi}^T \mathbf{y} + y_0 \geq ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2 \quad \forall \boldsymbol{\xi} \in \mathscr{S}_\rho, \\ & \left|\left| \mathbf{Q}^{\frac{1}{2}} (\mathbf{y} + 2\mathbf{Y}\boldsymbol{\mu}) \right|\right| \leq t, \\ & \mathbf{Y} \succeq 0. \end{aligned}$$

Proof Let $h(\mathbf{x}, \boldsymbol{\xi}) = ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2$. The first constraint is $\boldsymbol{\xi}^T \mathbf{Y} \boldsymbol{\xi} + \boldsymbol{\xi}^T \mathbf{y} + y_0 \geq ||(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}})\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}})||^2$. The second constraint can be rewritten as:

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

By substituting $t = t_1 + t_2$ and $t_1 = (\beta \mathbf{Q} + \boldsymbol{\mu}\boldsymbol{\mu}^T) \bullet \mathbf{Y} + \boldsymbol{\mu}^T \mathbf{y}$ in (2.3), we obtain (2.4). □

From Theorem 1 we can combine (2.4) with the outer problem to get the equivalent formulation of (1.6) with the probability ambiguity set $\mathscr{P}$ defined in (1.7) as

$$\min_{\mathbf{x},\mathbf{Y},\mathbf{y},y_0,t} \quad (\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}+y_0+\sqrt{\alpha}t \tag{2.5a}$$

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

$$\left|\left|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y}+2\mathbf{Y}\boldsymbol{\mu})\right|\right|\leq t, \tag{2.5c}$$

$$\mathbf{Y}\succeq 0. \tag{2.5d}$$

We now show that (2.5) can be solved in polynomial time. Anstreicher [1] showed that convex optimization problems can be solved with Vaidya's volumetric cutting plane method [22] in polynomial time. This result is summarized in the following lemma.

Lemma 2 (Anstreicher [1] and Vaidya [22]). *Consider a convex optimization problem of the form*

$$\min_{\mathbf{z}\in\mathscr{Z}} \mathbf{c}^T\mathbf{z} \tag{2.6}$$

with linear objective and convex feasible set $\mathscr{Z}$. Assume that the set of optimal solutions is nonempty. Then, the problem (2.6) can be solved to any precision ϵ in time polynomial in $\log(1/\epsilon)$ and in the size of the problem by using Vaidya's volumetric cutting plane method if $\mathscr{Z}$ satisfies the following two conditions:

1. *for any $\bar{\mathbf{z}}$, it can be verified that $\bar{\mathbf{z}}\in\mathscr{Z}$ or not in polyhnomial time.*
2. *for any infeasible $\bar{\mathbf{z}}$, a hyperplane that separates $\bar{\mathbf{z}}$ from the feasible region $\mathscr{Z}$ can be generated in polynomial time.*

According to Lemma 2, we need to find a polynomial time oracle to verify the feasibility of a given $(\mathbf{x},\mathbf{Y},\mathbf{y},y_0)$ for each of the constraints of (2.5b)–(2.5d). Note that (2.5c) is a second order cone constraint and (2.5d) is a semidefinite constraint. The main difficulty is the verification of the constraint (2.5b), because there are infinitely many $\boldsymbol{\xi}$'s in the sample space $\mathscr{S}_\rho$. For a given $(\mathbf{x},\mathbf{Y},\mathbf{y},y_0)$, consider the following optimization problem

$$g(\mathbf{x},\mathbf{Y},\mathbf{y},y_0)=\max_{\boldsymbol{\xi}\in\mathscr{S}_\rho}||(\mathbf{A}+\boldsymbol{\xi}_\mathbf{A})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_\mathbf{b})||^2-\boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi}-\boldsymbol{\xi}^T\mathbf{y}-y_0. \tag{2.7}$$

The following proposition shows that (2.7) is polynomial solvable.

Proposition 1 *For a given $\mathbf{x},\mathbf{Y},\mathbf{y},y_0$, problem (2.7) is equivalent to:*

$$g(\mathbf{x},\mathbf{Y},\mathbf{y},y_0):=\max_{||\boldsymbol{\xi}||\leq\rho}\ \boldsymbol{\xi}^T\hat{\mathbf{A}}\boldsymbol{\xi}+\hat{\mathbf{b}}\boldsymbol{\xi}+\hat{c}, \tag{2.8}$$

where

$$\hat{\mathbf{A}}=\mathbf{B}^T\mathbf{B}-\mathbf{Y},\ \hat{\mathbf{b}}=2\mathbf{B}^T(\mathbf{A}\mathbf{x}-\mathbf{b})-\mathbf{y},\ \hat{c}=||\mathbf{r}||^2-y_0$$
$$\mathbf{B}=(\mathbf{X}-\mathbf{I}_m),\ \mathbf{X}=[vec(\mathbf{e}_1\mathbf{x}^T),vec(\mathbf{e}_2\mathbf{x}^T),\ldots,vec(\mathbf{e}_m\mathbf{x}^T)]^T,\ \mathbf{r}=\mathbf{A}\mathbf{x}-\mathbf{b},$$

$\mathbf{I}_m$ *is a* $m \times m$ *identity matrix,* $\mathbf{B} \in \mathbb{R}^{m \times m(n+1)}$ *and* $\mathbf{e}_i$ *is the* m *dimensional vector with* 1 *in the* i*th entry and* 0 *otherwise.*

Proof Note that $\boldsymbol{\xi}_{\mathbf{A}}\mathbf{x} - \boldsymbol{\xi}_{\mathbf{b}} = \mathbf{B}\boldsymbol{\xi}$. Now, (2.7) can be rewritten as:

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

□

Theorem 2 *The moment robust least squares problem (1.6) with ambiguity set (1.7) can be solved in polynomial time.*

Proof Problem (2.7) is a standard trust region subproblem which can be solved polynomially [7,14]. From Lemma 2, the result follows. □

The semi-infinite constraint (2.5b) can also be reformulated as a semidefinite constraint as shown in the following theorem. This result was shown by an anonymous referee.

Theorem 3 *The constraint (2.5b) is equivalent to:*

$$\begin{pmatrix} \mathbf{I} & \mathbf{B} & \mathbf{r} \\ \mathbf{B}^T & \lambda\mathbf{I} + \mathbf{Y} & \frac{1}{2}\mathbf{y} \\ \mathbf{r}^T & \frac{1}{2}\mathbf{y}^T & -\lambda\rho + y_0 \end{pmatrix} \succeq 0, \lambda \ge 0 \tag{2.9}$$

where $\mathbf{I}$ *is the identity matrix,* $\mathbf{B}$ *is defined in Proposition 1, and* $\mathbf{r} = \mathbf{A}\mathbf{x} - \mathbf{b}$.

Proof Using the notation in Proposition 1, (2.5b) is equivalent to

$$\boldsymbol{\xi}^T\hat{\mathbf{A}}\boldsymbol{\xi} + \hat{\mathbf{b}}\boldsymbol{\xi} + \mathbf{r}^T\mathbf{r} - y_0 \le 0. \tag{2.10}$$

According to the S-procedure [4, Appendix B.2], inequality (2.10) is equivalent to the condition that there exists a $\lambda \ge 0$ such that:

$$\begin{pmatrix} \lambda\mathbf{I} - \hat{\mathbf{A}} & \frac{1}{2}\hat{\mathbf{b}} \\ \frac{1}{2}\hat{\mathbf{b}}^T & -\lambda\rho - \hat{c} \end{pmatrix} \succeq 0 \tag{2.11}$$

By using the definition of $\hat{\mathbf{A}}$ and $\hat{\mathbf{b}}$ in (2.11), and applying the Schur complement we obtain (2.9). □

Now from Theorem 3, problem (2.5) is equivalent to the semidefinite programming problem:

$$\begin{aligned}\min_{\mathbf{x},\mathbf{Y},\mathbf{y},y_0,t,\lambda}\quad & (\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}+y_0+\sqrt{\alpha}t\\ \text{s.t.}\quad & \begin{pmatrix}\mathbf{I} & \mathbf{B} & \mathbf{r}\\ \mathbf{B}^T & \lambda\mathbf{I}+\mathbf{Y} & \frac{1}{2}\mathbf{y}\\ \mathbf{r}^T & \frac{1}{2}\mathbf{y}^T & -\lambda\rho+y_0\end{pmatrix}\succeq 0,\\ & \left\|\mathbf{Q}^{\frac{1}{2}}(\mathbf{y}+2\mathbf{Y}\boldsymbol{\mu})\right\|\le t,\\ & \mathbf{Y}\succeq 0,\lambda\ge 0\end{aligned}$$

where $\mathbf{B}$ and $\mathbf{r}$ are defined in Proposition 1 and Theorem 3. We observe that (2.5) only depends on the first and second moment parameters $\boldsymbol{\mu}$ and $\mathbf{Q}$. Now we discuss the dependence between the worst case distribution and the parameters $\boldsymbol{\mu}$ and $\mathbf{Q}$. For a given $\mathbf{x}$, a worst case distribution is the distribution in $\mathscr{P}$ achieving the optimal for the inner problem (1.10). Consider an ambiguity set defined as:

$$\mathscr{P}=\{\nu:\mathbb{E}_\nu[\mathbf{1}]=\mathbf{1},\mathbb{E}_\nu[\boldsymbol{\xi}]=\boldsymbol{\mu},\mathbb{E}_\nu[\boldsymbol{\xi}\boldsymbol{\xi}^T]=\boldsymbol{\mu}\boldsymbol{\mu}^T+\beta\mathbf{Q},\boldsymbol{\xi}\in\mathscr{S}_\rho\}. \quad (2.12)$$

For the moment robust least squares problem with ambiguity set (2.12), we can write the dual of the inner problem (1.10) and combine the outer problem to get an equivalent formulation as:

$$\min_{\mathbf{x},\mathbf{Y},\mathbf{y},y_0,t}\quad (\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\bullet\mathbf{Y}+\boldsymbol{\mu}^T\mathbf{y}+y_0 \quad (2.13a)$$

$$\text{s.t.}\quad \boldsymbol{\xi}^T\mathbf{Y}\boldsymbol{\xi}+\boldsymbol{\xi}^T\mathbf{y}+y_0\ge||(\mathbf{A}+\boldsymbol{\xi}_\mathbf{A})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_\mathbf{b})||^2\quad\forall\boldsymbol{\xi}\in\mathscr{S}_\rho. \quad (2.13b)$$

Since the first two moments of the distribution are given, we know that the objective function $\mathbb{E}_\nu[||(\mathbf{A}+\boldsymbol{\xi}_\mathbf{A})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_\mathbf{b})||^2]$, which is the expectation of a quadratic function of $\boldsymbol{\xi}$, is independent of the selection of ν in $\mathscr{P}$. It rewrites the problem into an unconstrained optimization problem as follows:

$$\min_{\mathbf{x}}\ trace(\mathbf{B}(\beta\mathbf{Q}+\boldsymbol{\mu}\boldsymbol{\mu}^T)\mathbf{B}^T)+2\mathbf{r}^T\mathbf{B}\boldsymbol{\mu}+||\mathbf{r}||^2. \quad (2.14)$$

Now compare this case with the moment robust least squares with moment confidence region. In (2.5), we have two additional constraints (2.5c) and (2.5d). Since $\sqrt{\alpha}t$ is not necessarily 0, we observe that the optimal objective value of (2.5) does not necessarily equal the objective of the MRLS problem where the exact information of the first moments of the error vector are given as $\boldsymbol{\mu}$ and $\boldsymbol{\mu}\boldsymbol{\mu}^T+\beta\mathbf{Q}$. It implies that the worst case distribution in the ambiguity set (1.7) does not necessarily has its first moment $\boldsymbol{\mu}$ and second moment $\boldsymbol{\mu}\boldsymbol{\mu}^T+\beta\mathbf{Q}$. For the case $\alpha=0$, we can see that constraint (2.5c) is independent of the objective (2.5a); but we still have one additional constraint (2.5d), which is not implied by the semi-infinite constraint (2.5b) or its semidefinite equivalent formulation. Intuitively, we know that the confidence region gives the problem more freedom to put more weight on the worst scenario and it does not necessarily imply that the second moment need to achieve the upper bound. In summary, we cannot solve the moment robust least squares with ambiguity set (1.7) by assuming that the worst case distribution has first moment $\boldsymbol{\mu}$ and second moment $\boldsymbol{\mu}\boldsymbol{\mu}^T+\beta\mathbf{Q}$.

Remark Now we compare the above result with the result of Delage and Ye [5]. They consider a general distributionally robust optimization problem in the form

$$\min_{\mathbf{x}\in\mathscr{X}}\{\max_{P\in\mathscr{P}}\mathbb{E}_P[h(\mathbf{x},\boldsymbol{\xi})]\} \tag{2.15}$$

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 (2.15) is that the objective of (1.6) is not concave with respect to $\boldsymbol{\xi}$. Here, we take advantage of the specific formulation of the subproblem (2.7) which can be derived as a standard trust-region subproblem, and can be solved efficiently. Alternatively, a semidefinite reformulation of the problem is given.

3 DRLS with bounds on the probability measure

In this section we consider the DRLS problem (1.6) with the ambiguity set containing bounds on the probability measure. We assume that ρ, ν_1, ν_2, $\mathbf{Q}$, $\boldsymbol{\mu}$, α and β are given. This kind of probability ambiguity set as in (1.8) was considered by Shapiro and Ahmed [18].

3.1 Ambiguity with bounds on measure and moments

With the probability ambiguity set (1.8), the inner problem (1.10) is given as

$$\begin{aligned}
\max_{\nu\in\mathcal{M}} \quad & \int_{\mathscr{S}_\rho} \phi_0(\mathbf{x},\boldsymbol{\xi})\mathrm{d}\nu(\boldsymbol{\xi}) \\
\text{s.t.} \quad & \int_{\mathscr{S}_\rho} d\nu(\boldsymbol{\xi}) = 1, \\
& \int_{\mathscr{S}_\rho} (\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T d\nu(\boldsymbol{\xi}) \preceq \beta\mathbf{Q}, \\
& \int_{\mathscr{S}_\rho} \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix} \mathrm{d}\nu(\boldsymbol{\xi}) \succeq 0,
\end{aligned} \tag{3.1}$$

where $\phi_0(\mathbf{x},\boldsymbol{\xi}) = ||(\mathbf{A}+\boldsymbol{\xi}_\mathbf{A})\mathbf{x}-(\mathbf{b}+\boldsymbol{\xi}_\mathbf{b})||^2$ and $\mathcal{M} := \{\nu : \nu_1 \preceq \nu \preceq \nu_2\}$. Let

$$\begin{aligned}
\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi}) := \; & \phi_0(\mathbf{x},\boldsymbol{\xi}) - \lambda_0 - ((\boldsymbol{\xi}-\boldsymbol{\mu})(\boldsymbol{\xi}-\boldsymbol{\mu})^T - \beta\mathbf{Q}) \bullet \boldsymbol{\Lambda}_1 \\
& + \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi}-\boldsymbol{\mu}) \\ (\boldsymbol{\xi}-\boldsymbol{\mu})^T & \alpha \end{pmatrix} \bullet \boldsymbol{\Lambda}_2.
\end{aligned} \tag{3.2}$$

and the Lagrangian of problem (3.1) be given as:

$$L(\mathbf{x}, \nu, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) := \int_{\mathscr{S}_\rho} \mathcal{L}_{\lambda_0, \Lambda_1, \Lambda_2}(\mathbf{x}, \boldsymbol{\xi}) d\nu(\boldsymbol{\xi}) + \lambda_0. \tag{3.3}$$

The Lagrangian dual of (3.1) is:

$$\begin{aligned} \min_{\lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2} \quad & \left\{ \psi(\mathbf{x}, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) := \sup_{\nu \in \mathcal{M}} L(\mathbf{x}, \nu, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) \right\} \\ \text{s.t.} \quad & \mathbf{\Lambda}_1, \mathbf{\Lambda}_2 \succeq 0. \end{aligned} \tag{3.4}$$

Proposition 2 *In the Lagrangian dual problem (3.4),*

$$\begin{aligned} \psi(\mathbf{x}, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2) := & \int_{\mathscr{S}_\rho} [\mathcal{L}_{\lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_+ d\nu_2(\boldsymbol{\xi}) \\ & - \int_{\mathscr{S}_\rho} [\mathcal{L}_{\lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2}(\mathbf{x}, \boldsymbol{\xi})]_- d\nu_1(\boldsymbol{\xi}) + \lambda_0. \end{aligned} \tag{3.5}$$

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 a given $(\mathbf{x}, \lambda_0, \Lambda_1, \Lambda_2)$, 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 set $\{\nu : \nu_1 \preceq \nu \preceq \nu_2\}$, we integrate $L(\mathbf{x}, \nu, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2)$ with respect to ν_1 over the set C_- and integrate with respect to ν_2 over the set C_+, which leads to the desired equality in (3.5). □

From the Lagrangian weak duality theory, we know that the optimal value of problem (3.1) is always less than or equal to the optimal value of its dual (3.4). The conjugate duality theory [2,15] ensures that the strong duality for (3.1) and (3.4) holds and the set of optimal solutions of the dual problem is nonempty and bounded if the following assumption holds:

Assumption 2 The optimal value of (3.1) is finite, and there exists probability measure $\nu \in \mathcal{M}$ such that

$$\begin{aligned} & \int_{\mathscr{S}_\rho} ((\boldsymbol{\xi} - \boldsymbol{\mu})(\boldsymbol{\xi} - \boldsymbol{\mu})^T - \beta \mathbf{Q}) \mathrm{d}\nu(\boldsymbol{\xi}) \prec 0, \\ & \int_{\mathscr{S}_\rho} \begin{pmatrix} \mathbf{Q} & (\boldsymbol{\xi} - \boldsymbol{\mu}) \\ (\boldsymbol{\xi} - \boldsymbol{\mu})^T & \alpha \end{pmatrix} \mathrm{d}\nu(\boldsymbol{\xi}) \succ 0, \end{aligned} \tag{3.6}$$

i.e., the set of feasible measures has a nonempty interior w.r.t. the moment constraints.

Assumption 2 ensures that there is no duality gap between (3.1) and (3.4). With Assumption 2, the original DRLS problem (1.6) with probability ambiguity set (1.8) is equivalent to:

$$\begin{aligned} \min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2} \quad & \psi(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) \\ \text{s.t.} \quad & \mathbf{\Lambda}_1,\mathbf{\Lambda}_2 \succeq 0. \end{aligned} \tag{3.7}$$

3.2 Some cases of DRLS with bounds on probability measure

In this section we will discuss three special cases. The first case assumes a finite support. The other two cases assume a continuous support sample space. The sample average approximation is used to convert them to the first finite support case for problem reformulation and convergence results.

3.2.1 Finite support case

Let us assume that the sample space defining ν_1, ν_2 and ν is finite, and $\mathscr{S} := \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$. 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 $\bar{\mathscr{P}}$ is defined as follows:

$$\begin{aligned} \mathscr{P} := \Bigg\{ & (p_1, \ldots, p_K) : \sum_{i=1}^{K} p_i = 1,\ \underline{p}_i \le p_i \le \bar{p}_i \text{ for } i = 1, \ldots, K, \boldsymbol{\tau} = \sum_{i=1}^{K} p_i \boldsymbol{\xi}^i \\ & (\boldsymbol{\tau} - \boldsymbol{\mu})^T \mathbf{Q}^{-1} (\boldsymbol{\tau} - \boldsymbol{\mu}) \le \alpha,\ \sum_{i=1}^{K} p_i (\boldsymbol{\xi}^i - \boldsymbol{\mu})(\boldsymbol{\xi}^i - \boldsymbol{\mu})^T \preceq \beta \mathbf{Q} \Bigg\}. \end{aligned} \tag{3.8}$$

Consider the DRLS problem (1.6) with probability ambiguity set (3.8). The following theorem ensures that the finite support case can be reformulated as a single conic optimization problem.

Theorem 4 *Assume that there exists a* $\mathbf{p} := (p_1, \ldots, p_K) \in \mathscr{P}$ *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* $\boldsymbol{\tau} = \sum_{i=1}^{K} p_i \boldsymbol{\xi}^i$. *The DRLS problem with ambiguity set (1.8) is equivalent to:*

$$\begin{aligned} \min_{\mathbf{x},s,\bar{s}_i,\underline{s}_i,\mathbf{u},\mathbf{X},\mathbf{z}} \quad & s + \sum_{i=1}^{K} (\bar{p}_i \bar{s}_i + \underline{p}_i \underline{s}_i) + [\sqrt{\alpha}, -(\mathbf{Q}^{-\frac{1}{2}} \boldsymbol{\mu})^T] \mathbf{z} + \beta \mathbf{Q} \bullet \mathbf{X} \\ \text{s.t.} \quad & s + \bar{s}_i + \underline{s}_i + {\boldsymbol{\xi}^i}^T \mathbf{u} + (\boldsymbol{\xi}^i - \boldsymbol{\mu})(\boldsymbol{\xi}^i - \boldsymbol{\mu})^T \bullet \mathbf{X} \ge \left\| (\mathbf{A} + \boldsymbol{\xi}^i_{\mathbf{A}}) \mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}^i_{\mathbf{b}}) \right\|^2, \\ & \qquad \textit{for } i = 1, \ldots, K, \\ & -\mathbf{u} + [\mathbf{0}, -(\mathbf{Q}^{-\frac{1}{2}})^T] \mathbf{z} = 0, \\ & \bar{s}_i \ge 0, \underline{s}_i \le 0, \quad \textit{for } i = 1, \ldots, K, \\ & \mathbf{z} \in \mathcal{K}, \mathbf{X} \succeq 0 \end{aligned} \tag{3.9}$$

where $\mathcal{K}$ is a second order cone, which is defined as $\mathcal{K} := \{\mathbf{y} := (y_1, y_2, \ldots, y_l) : y_1 \geq \sqrt{y_2^2 + \ldots + y_l^2}\}$.

Proof For a given $\mathbf{x}$, the inner problem (1.10) with ambiguity set (3.8) can be explicitly written as:

$$\begin{aligned}
\max_{\mathbf{p}:=(p_1,\ldots,p_K)} \quad & \sum_{i=1}^{K} p_i \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \qquad (3.10)\\
\text{s.t.} \quad & \sum_{i=1}^{K} p_i = 1,\\
& \underline{p}_i \leq p_i \leq \bar{p}_i,\\
& \boldsymbol{\tau} = \sum_{i=1}^{K} p_i \boldsymbol{\xi}^i,\\
& (\boldsymbol{\tau}-\boldsymbol{\mu})^T \mathbf{Q}^{-1}(\boldsymbol{\tau}-\boldsymbol{\mu}) \leq \alpha,\\
& \sum_{i=1}^{K} p_i (\boldsymbol{\xi}^i-\boldsymbol{\mu})(\boldsymbol{\xi}^i-\boldsymbol{\mu})^T \preceq \beta\mathbf{Q},\\
& \mathbf{p} := (p_1, \ldots, p_K) \in \mathbb{R}_+^K.
\end{aligned}$$

(3.10) is a linear conic programming problem for a given $\mathbf{x}$. Because the Slater's constraint qualifications are assumed, the duality theory for second order cone and semidefinite programming guarantees strong duality [21]. We rewrite constraint $(\boldsymbol{\tau}-\boldsymbol{\mu})^T\mathbf{Q}^{-1}(\boldsymbol{\tau}-\boldsymbol{\mu}) \leq \alpha$ as: $(t_0; \mathbf{t}) = (\sqrt{\alpha}; \mathbf{Q}^{\frac{1}{2}}(\boldsymbol{\tau}-\boldsymbol{\mu}))$, $(t_0; \mathbf{t}) \in \mathcal{K}$, where $\mathcal{K}$ is a second order cone. Taking the dual of (3.10) gives

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

Combining the (3.11) with the outer problem (1.11), we can get (3.9). □

3.2.2 Measure bounds by a reference probability measure

We now assume that the bounds ν_1 and ν_2 are given as:

$$\nu_1 := (1-\epsilon_1)\mathbb{P}^*, \nu_2 := (1+\epsilon_2)\mathbb{P}^*, \qquad (3.12)$$

where $\epsilon_1 \in [0, 1]$ and $\epsilon_2 \geq 0$ are given constants, and $\mathbb{P}^*$ is a given reference probability measure. We show that DRLS with ambiguity set (1.8) and ν_1, ν_2 defined in (3.12) can be reformulated as a stochastic convex optimization problem. Consequently, the sample average approximation (SAA) can be used to solve this case. The convergence of a solution of the SAA problem to that of the original problem is guaranteed by Theorem 8 in the "Appendix".

Theorem 5 *The DRLS problem (1.6) with ambiguity set (1.8) and ν_1, ν_2 defined in (3.12) is equivalent to the stochastic programming problem:*

$$\min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2 \geq 0} \int_{\mathscr{S}_\rho} ((1+\epsilon_2)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_+ - (1-\epsilon_1)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_-) d\mathbb{P}^*(\boldsymbol{\xi}) + \lambda_0. \tag{3.13}$$

Given a sample $\Omega := \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$ from the probability measure $\mathbb{P}^*$, the SAA formulation of (3.13) is equivalent to (3.9) with $\underline{p}_i = \frac{1-\epsilon_1}{K}$, $\bar{p}_i = \frac{1+\epsilon_2}{K}$ for $i = 1, \ldots, K$.

Proof Equation (3.13) is a direct result by substituting (3.12) in (3.5). Note that $(1+\epsilon_1)[\cdot]_+ - (1-\epsilon_2)[\cdot]_-$ is a convex piecewise linear increasing function, and $\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})$ is a convex function 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 (3.7) is convex in $\mathbf{x}, \lambda_0, \mathbf{\Lambda}_1, \mathbf{\Lambda}_2$. Therefore, the DRLS problem (1.6) with the ambiguity set (1.8) is reformulated as a convex stochastic programming problem with objective (3.13). We can rewrite (3.7) as:

$$\min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1 \geq 0,\mathbf{\Lambda}_2 \geq 0} \mathbb{E}_{\mathbb{P}^*}[H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi})] \tag{3.14}$$

where

$$H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi}) := (1+\epsilon_2)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_+ - (1-\epsilon_1)[\mathcal{L}_{\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})]_- + \lambda_0. \tag{3.15}$$

According to Theorem 8 in the "Appendix", the sample average approximation (SAA) techniques can be used to estimate $\mathbb{E}_{\mathbb{P}^*}[\cdot]$ in (3.14) and will guarantee the almost sure convergence of both the objective value and the optimal solution set. We refer [16,19] for details about SAA. Given a finite sample $\Omega := \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$ from distribution $\mathbb{P}^*$, the SAA of problem (3.14) can be written as:

$$\min_{\mathbf{x},\lambda_0,\mathbf{\Lambda}_1 \geq 0,\mathbf{\Lambda}_2 \geq 0} \{h(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2) := \frac{1}{K}\sum_{k=1}^{K} H(\mathbf{x},\lambda_0,\mathbf{\Lambda}_1,\mathbf{\Lambda}_2,\boldsymbol{\xi}^k)\} \tag{3.16}$$

Comparing (3.16) with (3.7), we see that (3.16) is equivalent to the DRLS with finite support case presented in Sect. 3.2.1 with support $\Omega := \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$ and measure bounds $\underline{p}_i = \frac{1-\epsilon_1}{K}$, $\bar{p}_i = \frac{1+\epsilon_2}{K}$ for $i = 1, \ldots, K$. □

3.2.3 Measure bounds with known densities

Now consider another case of DLRS by defining ν_1 and ν_2 as follows:

$$\nu_1 := (1-\epsilon_1)\underline{\mathbb{P}}^*,\ \nu_2 := (1+\epsilon_2)\bar{\mathbb{P}}^*, \tag{3.17}$$

where $\underline{\mathbb{P}}^*$ and $\bar{\mathbb{P}}^*$ are two given probability measures with known density functions $\underline{f}(\cdot)$ and $\bar{f}(\cdot)$. The condition $\nu_1 \preceq \nu_2$ gives $(1-\epsilon_1)\underline{f}(\boldsymbol{\xi}) \le (1+\epsilon_2)\bar{f}(\boldsymbol{\xi})$ for $\forall \boldsymbol{\xi} \in \mathscr{S}_\rho$. In the following theorem, we show that the DRLS problem with ambiguity set (1.8) and ν_1, ν_2 defined in (3.17) can also be solved using the SAA approach by sampling the uniform distribution. The convergence of a solution of the SAA problem to that of the original problem is guaranteed by Theorem 8 in the "Appendix".

Theorem 6 *The DRLS problem (1.6) with ambiguity set (1.8) and ν_1, ν_2 defined in (3.12) is equivalent to the stochastic convex programming problem:*

$$\min_{\mathbf{x},\lambda_0,\boldsymbol{\Lambda}_1\ge 0,\boldsymbol{\Lambda}_2\ge 0} \int_{\mathscr{S}_\rho} [(1+\epsilon_2)\bar{f}(\boldsymbol{\xi})\left[\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})\right]_+ - (1-\epsilon_1)\underline{f}(\boldsymbol{\xi})\left[\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi})\right]_-]d\boldsymbol{\xi} + \lambda_0 \tag{3.18}$$

Given a sample $\Omega := \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$ from the uniform distribution over $\mathscr{S}_\rho$, the SAA formulation of (3.18). (3.19)

$$\min_{\mathbf{x},\lambda_0,\boldsymbol{\Lambda}_1\ge 0,\boldsymbol{\Lambda}_2\ge 0} \frac{1}{K}\sum_{k=1}^{K}[(1+\epsilon_2)\bar{f}(\boldsymbol{\xi}^k)[\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi}^k)]_+ - (1-\epsilon_1)\underline{f}(\boldsymbol{\xi}^k)[\mathcal{L}_{\lambda_0,\boldsymbol{\Lambda}_1,\boldsymbol{\Lambda}_2}(\mathbf{x},\boldsymbol{\xi}^k)]_-] + \lambda_0 \tag{3.19}$$

is equivalent to (3.9) with $\underline{p}_i = \frac{1-\epsilon_1}{K}\underline{f}(\boldsymbol{\xi}^i)$, $\bar{p}_i = \frac{1+\epsilon_2}{K}\hat{f}(\boldsymbol{\xi}^i)$ for $i = 1, \ldots, K$.

Proof Formulation (3.18) is a direct result by substituting (3.17) in (3.5). 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 $\mathbf{x}$, λ_0, $\boldsymbol{\Lambda}_1$, $\boldsymbol{\Lambda}_2$. The integration in (3.18) can be considered as an expectation with the uniform distribution over the sample space $\mathscr{S}_\rho$. With an argument similar to the proof of Theorem 5, we get the desired result. □

Remark The cases in Sects. 3.2.2 and 3.2.3 are similar. There is one important difference, however. The case in Sect. 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 $\mathbb{P}^*$. The only requirement is the ability to sample from $\mathbb{P}^*$. The case in Sect. 3.2.3considers the bounds defined

by two different probability measures. But we require the knowledge of their density functions. Even though the above three cases do not cover all the possibilities, they do cover a lot of interesting situations. The first two cases are considered by Shapiro and Ahmed [18] in a more general framework. However, because of the generality, they show that SAA of the problem can be solved by the subgradient method. For our specific least squares formulation we show that the SAA of the two cases have equivalent conic programming formulations, which can be solved efficiently by existing software packages such as SeDuMi [20].

4 DRLS with C.I. by probability metric

In this section we consider the probability ambiguity set defined by the Kantorovich distance. Assume that $\mathbb{P}^*$ is a known reference probability measure and $\epsilon > 0$ is a given constant. The Kantorovich distance (or L_1 distance) [23] between two probability measures P_1 and P_2 is defined as:

$$d(\mathbb{P}_1, \mathbb{P}_2) := \sup_f \left\{ \int f(\mathbf{u}) d\mathbb{P}_1(\mathbf{u}) - \int f(\mathbf{u}) d\mathbb{P}_2(\mathbf{u}), |f(\mathbf{u}) - f(\mathbf{v})| \right. \\ \left. \leq ||\mathbf{u} - \mathbf{v}||_1 \text{ for all } \mathbf{u}, \mathbf{v} \right\}. \tag{4.1}$$

Note that every probability measure $\mathbb{P} \in \mathscr{P}$ is defined on the same sample space Ω with σ-algebra $\mathcal{F}$. Pflug and Wozabal [11] consider this type of ambiguity in portfolio optimization. They analyzed the portfolio selection problem with the ambiguity set defined by a probability confidence set with Kantorovish distance. Their problem has a linear objective and additional constraints on the ambiguity set such as the lower bound of the expected return. A successive convex programming solution method is given in [11]. Their 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 to the cutting plane method in convex optimization. Compared with their methodology, the DRLS problem with ambiguity set given by Kantorovish distance has a nice structure when the sample space is finite. This leads to a second order cone programming formulation (4.5) as shown in the following theorem.

Theorem 7 *The DRLS problem with ambiguity set (1.9) and discrete support* $\Omega = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$ *is equivalent to the conic optimization problem:*

$$\min_{\mathbf{x}, s_i, t_j, \sigma, i, j = 1, \ldots, K} \quad \sum_{j=1}^{K} p_j^* t_j + \epsilon\sigma \tag{4.2}$$

$$s.t. \quad s_i \geq \left|\left|(\mathbf{A} + \boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b} + \boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \quad \text{for } i = 1, \ldots, K,$$

$$-s_i + t_j + \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 \sigma \geq 0 \quad \text{for } i, j = 1, \ldots, K,$$

$$\sigma \geq 0.$$

where $\mathbf{p}^* := (p_1^*, \ldots, p_K^*)$ *is the reference probability measure* $\mathbb{P}^*$.

Proof According to the Kantorovich–Rubinstein theorem [12], the Kantorivich ambiguity set (1.9), where $d(\mathbb{P}_1, \mathbb{P}_2)$ is defined in (4.1), can be represented as:

$$\mathscr{P} := \{\mathbb{P}: \quad \text{there is a bivariate probability distribution } \mathbb{K}(\cdot,\cdot) \text{ such that} \tag{4.3}$$
$$\int_{\mathbf{v}} \mathbb{K}(\mathbf{u}, d\mathbf{v}) = \mathbb{P}(\mathbf{u}); \int_{\mathbf{u}} \mathbb{K}(d\mathbf{u}, \mathbf{v}) = \mathbb{P}^*(\mathbf{v}); \int_{\mathbf{u}}\int_{\mathbf{v}} \|\mathbf{u}-\mathbf{v}\|_1 \mathbb{K}(d\mathbf{u}, d\mathbf{v}) \le \epsilon\}.$$

Given the sample space $\Omega = \{\boldsymbol{\xi}^1, \ldots, \boldsymbol{\xi}^K\}$, let $\mathbf{p} := (p_1, \ldots, p_K)$ be the corresponding probability measure. The probability ambiguity set (4.3) is equivalent to:

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

Given $\mathbf{x}$, the inner problem (1.10) has the form:

$$\max_{p_i, k_{i,j}, i,j=1,\ldots,K} \sum_{i=1}^{K} p_i \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \tag{4.4}$$
$$\text{s.t.} \sum_{j=1}^{K} k_{i,j} = p_i \quad \text{for } i = 1, \ldots, K,$$
$$\sum_{i=1}^{K} k_{i,j} = p_j^* \quad \text{for } j = 1, \ldots, K,$$
$$\sum_{i=1}^{K}\sum_{j=1}^{K} \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 k_{i,j} \le \epsilon,$$
$$k_{i,j} \ge 0 \quad \text{for } i, j = 1, \ldots, K,$$
$$p_i \ge 0 \quad \text{for } i = 1, \ldots, K,$$

which is a linear programming problem. Since $\mathbf{p}^*$ is a feasible solution and the objective is bounded for a given $\mathbf{x}$, the strong duality holds. We can write the dual of (4.4) as:

$$\min_{\sigma, s_i, t_j, i,j=1,\ldots,K} \sum_{j=1}^{K} p_j^* t_j + \epsilon\sigma \tag{4.5}$$
$$\text{s.t.} \; s_i \ge \left|\left|(\mathbf{A}+\boldsymbol{\xi}_{\mathbf{A}}^i)\mathbf{x} - (\mathbf{b}+\boldsymbol{\xi}_{\mathbf{b}}^i)\right|\right|^2 \quad \text{for } i = 1, \ldots, K,$$
$$-s_i + t_j + \left|\left|\boldsymbol{\xi}^i - \boldsymbol{\xi}^j\right|\right|_1 \sigma \ge 0 \quad \text{for } i, j = 1, \ldots, K,$$
$$\sigma \ge 0,$$

by combining (4.5) with the outer problem (1.11) we get the desired formulation (4.2). □

5 Conclusions remarks

We have presented distributionally robust least squares frameworks with three different probability ambiguity sets. They include ambiguity sets defined by (i) confidence region of the first two moments, (ii) measure bounds and (iii) confidence region defined by Kantorovich distance. For the first case, we show that the equivalent semi-infinite programming formulation can be solved by a cutting plane method with oracle determined by the trust-region subproblem. We also include an observation by a referee that this problem can be reformulated as a semidefinite programming problem. The second case is shown to be equivalent to a convex stochastic programming problem. We analyze these special sub-cases, (1) finite support case, (2) measure bounds given by a reference probability measure, (3) measure bounds given by two reference probability measures with known density functions. For case (2) and (3), the SAA formulation can be reformulated as a conic programming problem. We also show that in the discrete support case the DRLS problem with ambiguity set define by Kantorovich distance can be reformulated as a second order cone programming problem. The second order cone and semidefinite formulations of the all the three cases make the DRLS problem more attractive because of their solvability using off-the-shelf software packages such as SeDuMi [20].

6 Appendix

In Sect. 3 we use SAA method to approximate the stochastic programming problems (3.13) and (3.18). The convergence results for the SAA method from [16] are summarized here. Consider a stochastic programming problem of the form

$$\min_{\mathbf{x}\in\mathcal{X}}\{f(\mathbf{x}) := \mathbb{E}_{\mathbb{P}}[F(\mathbf{x},\boldsymbol{\xi})]\}, \tag{6.1}$$

where $F(\mathbf{x},\boldsymbol{\xi})$ is a function of variables $\mathbf{x}\in\mathbb{R}^n$ and parameters $\boldsymbol{\xi}\in\mathbb{R}^d$. $\mathcal{X}\subset\mathbb{R}^n$ is a given set, and $\boldsymbol{\xi}=\boldsymbol{\xi}(\omega)$ is a random vector. The expectation in (6.1) is taken with respect to the probability distribution of $\boldsymbol{\xi}$ which is assumed to be known as $\mathbb{P}$. 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

$$\hat{f}_K(\mathbf{x}) := \frac{1}{K}\sum_{i=1}^{K} F(\mathbf{x},\xi^i). \tag{6.2}$$

The stochastic programming problem (6.1) is approximated by the optimization problem

$$\min_{\mathbf{x}\in\mathcal{X}}\left\{\hat{f}_K(\mathbf{x}) := \frac{1}{K}\sum_{i=1}^{K} F(\mathbf{x},\xi^i)\right\}. \tag{6.3}$$

Let the optimal value of (6.1) be v and it's optimal solution set be S. Let $\hat{v}_K$ and $\hat{S}_K$ be the optimal value and the set of optimal solutions of the SAA problem (6.3). 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

$$\mathbb{D}(A, B) := \sup_{x \in A} \text{dist}(x, B). \tag{6.4}$$

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 following convergence theorem ensures that a solution of SAA problem (6.2) converges to that of (6.1) as the sample size increases.

Theorem 8 [16, Chapter 6, Theorem 4.] *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 (6.1) is nonempty and bounded,* (vi) *the LLN holds pointwise. Then $\hat{v}_K \to v^*$ 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 [16,19] for details of such results. We note that the assumptions in Theorem 8 are satisfied for the models (3.13) and (3.18). Hence the convergence of the solution of SAA of these problems to a solution of the original problem is guaranteed.

References

1. Anstreicher, K.M.: On Vaidya's volumetric cutting plane method for convex programming. Math. Oper. Res. **22**(1), 63–89 (1997)
2. Bertsekas, Dimitri P.: Convex Analysis and Optimization. Athena Scientific, Belmont (2003)
3. Bertsimas, D., Doan, X.V., Natarajan, K., Teo, C.: Models for minimax stochastic linear optimization problems with risk aversion. Math. Oper. Res. **35**(3), 580–602 (2010)
4. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
5. Delage, E., Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Oper. Res. **58**(3), 595–612 (2010)
6. Dupacova, J.: The minimax approach to stochastic programming and an illustrative application. Stochastics **20**, 73–88 (1987)
7. Fortin, C., Wolkowicz, H.: The trust region subproblem and semidefinite programming. Optim. Methods Softw. **19**(1), 41–67 (2004)
8. Ghaoui, L.E., Lebret, H.: Robust solution to least-squares problem with uncertain data. SIAM J. Matrix Anal. Appl. **18**(4), 1035–1064 (1997)
9. Greene William, H.: Econometric Analysis, 6th edn. Prentice Hall, Upper Saddle River (2007)
10. Lyandres Olga, Van Duyne, Richard P., Glucksberg, Joseph T., Sanjay Mehrotra, : Prediction range estimation from noisy raman spectra with robust optimization. Analyst **135**(8), 2111–2118 (2010)
11. Pflug, G., Wozabal, D.: Ambiguity in portfolio selection. Quant. Finance **7**(4), 435–442 (2007)
12. Rachev, S.T.: Probability Metrics and the Stability of Stochastic Models. Willey Series in Probability and Mathematical Statistics. Wiley, New York (1991)

13. Calyampudi Radhakrishna Rao: The theory of least squares when the parameters are stochastic and its application to the analysis of growth curves. Biometrika **52**(3–4), 447–458 (1965)
14. Rendl, F., Wolkowicz, H.: A semidefinite framework for trust region subproblems with applications to large scale minimization. Math. Program. **77**, 273–299 (1997)
15. Rockafellar R.T.: Conjugate Duality and Optimization. CBMS-NSF Regional Conference Series in Applied Mathematics 16. SIAM, Philadelphia (1974)
16. Ruszczynski, A., Shapiro, A. (eds.): Stochastic Programming. Handbooks in Operations Research and Management Science 10. North-Holland, Amsterdam (2003)
17. Scarf, H.: A min-max Solution of an Inventory Problem. Stanford University Press, Stanford (1958)
18. Shapiro, A., Ahmed, S.: On a class of minimax stochastic programs. SIAM J. Optim. **14**(4), 1237–1249 (2004)
19. Shapiro, A., Homem de mello, T.: On the rate of convergence of optimal solutions of monte carlo approximations of stochastic programs. SIAM J. Optim. **11**(1), 70–86 (2000)
20. Sturm Jos, F.: Using SeDuMi 1.02, a MATLAB* toolbox for optimization over symmetric cones. Optim. Methods Softw. **11–12**, 625–653 (1999)
21. Sturm Jos, F.: Implementation of interior point methods for mixed semidefinite and second order cone optimization problems. Optim. Methods Softw. **17**, 1105–1154 (2002)
22. Vaidya, P.M.: A new algorithm for minimizing convex functions over convex sets. Math. Program. **73**, 291–341 (1996)
23. Vallander, S.S.: Calculation of the wasserstein distance between probability distributions on the line. Theory Probab. Appl. **18**, 784–786 (1973)