Inference on Estimators defined by Mathematical Programming

Yu-Wei Hsieh†  Xiaoxia Shi‡  Matthew Shum§
USC       UW-Madison       Caltech

December 11, 2017

Abstract

We propose an inference procedure for estimators defined as optimizers of mathematical programming problems, focusing on the important cases of linear programming (LP) and convex quadratic programming (QP). Coefficients in both the objective function and the constraints of the problem may be estimated from data and hence involve sampling error. Our approach exploits the characterization of the solutions to these programming problems by complementarity conditions; by doing so, we transform the problem of doing inference on the solution of a constrained optimization problem (a non-standard inference problem) into one involving inference based on a set of inequalities with pre-estimated coefficients, which is much better understood. Our approach is valid regardless of whether the problem has a unique or multiple solutions. We evaluate of our procedure in Monte Carlo simulations and an empirical application to the classic portfolio selection problem in finance.

Keywords: Stochastic Mathematical Programming, Linear Complementarity Constraints, Moment Inequalities, Sub-Vector Inference, Portfolio Selection

JEL Classification: C10, C12, C63

*We thank Denis Chetverikov, Bulat Gafarov, Bryan Graham, Jinyong Hahn, Yuichi Kitamura, Emerson Melo, Ismael Mourifié, Andres Santos and attendees at the California Econometrics Conference 2017 for helpful comments.

†Department of Economics and Dornsife Institute for New Economic Thinking, University of Southern California. Email: yuwei.hsieh@usc.edu

‡Department of Economics, University of Wisconsin-Madison. Email: xshi@ssc.wisc.edu

§Division of Humanities and Social Sciences, Caltech. Email: mshum@caltech.edu
1 Introduction

In this paper, we consider the problem of inference on an estimator defined as the solution to a mathematical programming problem with pre-estimated coefficients. Because of the pre-estimation, these coefficients contain sampling error, and hence the mathematical programming problem is stochastic. Our focus is on the important special cases of linear programming (LP) and convex quadratic programming (QP), for which there are relevant examples in economics and finance. The difficulty with doing inference based on such estimators lies in the nondifferentiability of the estimator with respect to the data. As a result of the nondifferentiability, the estimator is not asymptotically normal, and does not allow for standard bootstrap inference.

The core of our method lies in recognizing that the necessary and sufficient optimality conditions for LP/QP can be interpreted as inequalities with pre-estimated coefficients. Typically, these optimality conditions involve Lagrange multipliers and slackness variables for constraints, and a set of linear complementarity (LC) conditions. Essentially, by focusing on these optimality conditions, we can transform the problem of doing inference on the solution of a constrained optimization problem (a non-standard inference problem) into one involving inference on a set of inequalities with pre-estimated coefficients, which is similar to moment inequality models and is much better understood. Specifically, we show that the inference on the inequalities implied by the optimality conditions of LP/QP can proceed using the computationally convenient procedures from Shi and Shum (2015).

Estimators defined by mathematical programming have a long history in econometrics, dating back to Markowitz’s (1952) classic work on optimal portfolio selection. Problems in policy evaluation, such as optimal group assignment (Graham, Imbens, and Ridder (2006), Bhattacharya (2009)) and treatment assignment (Bhattacharya and Dupas (2012)), also take the form of constrained mathematical programming problems. More recently, Chiong, Galichon, and Shum (2016) and Chiong, Hsieh, and Shum (2017) propose estimators for problems in discrete-choice analysis which also take the form of mathematical programming. Due to the absence of an inference theory, researchers often resort to bootstrap in practice; e.g., Scherer (2002). Recently, however, Fang and Santos (2016) show that canonical bootstrap is not valid if the solution is non-differentiable in the estimated coefficients. As the solution of mathematical programming is non-differentiable in general, our approach provides, to the best of our knowledge, the first valid inference method in the literature.

An important feature of our inference approach is that it remains valid in both the scenarios in which the solution to the mathematical programming problem is unique or multiple (e.g. when the solution of a LP is located on a “flat face”). In the former case, our confidence set covers the unique solution with prespecified level, while in the latter case, each point in
the solution set is covered with prespecified probability. Hence, our procedures can be used even when the researcher does not know in advance whether the solution is unique or not, which is likely in practice.\footnote{In comparison, the possibility of a “flat face” solution leads to complications in a approach based on the differentiability properties of the estimator. Shapiro (1993) studies the asymptotic properties of the solution to a mathematical programming with stochastic coefficients, but requires uniqueness of the solution, which is difficult to verify in practice (Williams (2013)).}

In the next section we review the key results from the theory of linear and quadratic programming. In section 3 and 4 we provide examples and illustrate how to conduct inference respectively. In section 5 we investigate the performance of the proposed confidence set using two LP case. For the case of QP, in section 6 we estimate Markovitz’s (1952) efficient portfolio weights and their confidence set. As far as we are aware, our analysis of the Markowitz portfolio selection problem here represents the first instance of inference for this problem based on asymptotic approximation.

2 Linear Programming and Quadratic Programming

Next we focus on the specific cases of linear programming (LP) and convex quadratic programming (QP), for which our approach is easier to understand and the results are sharpest. We will also briefly discuss more general nonlinear programming problems in conclusion. We introduce the LP and QP problems in turn.

2.1 Linear programming

We want to estimate $\theta$ defined by the following LP:

$$\theta := \arg\max \ c^T \theta \ \text{s.t.} \ A\theta \leq b$$

(1)

where $\theta \in \Theta \subset \mathbb{R}^k$, $b$ is $m \times 1$, $c$ is $k \times 1$, and $A$ is $m \times k$. Let $A$, $b$ or $c$ be estimated from data; the sample analogs are $\hat{A}$, $\hat{b}$, and $\hat{c}$. Then the parameter of interest is estimated by

$$\hat{\theta} = \arg\max \ \hat{c}^T \theta \ \text{s.t.} \ \hat{A}\theta \leq \hat{b}$$

(2)

The goal is to to derive an inference method for $\hat{\theta}$.

Our approach is to exploit the necessary and sufficient optimality conditions that characterize the solutions to linear programming problems, which follow from the duality theory of LP. Specifically, these optimality conditions are
Equation (3) and (4) express, respectively, primal and dual feasibility, where $\lambda$ is interpreted as the $m \times 1$ vector of Lagrange multipliers on the inequalities (3). The final equation (6) is a complementarity condition, analogous to the complementarity slackness equation in the Karush-Kuhn-Tucker (KKT) approach.

In optimization theory, these equalities and inequalities furnish the basis for primal-dual interior point methods for solving mathematical programming problems, and so in what follows we will follow this literature in referring to similar sets of (in-)equalities as primal-dual conditions. These considerations yield the following key proposition.

**Proposition 1. (LP Inference).** Inference on $\hat{\theta}$ defined as the solution to the LP problem (2) is equivalent to inference on $\hat{\theta}$ satisfying the inequalities (3)-(6) evaluated at the estimated quantities $\hat{A}, \hat{b}, \hat{c}$.

Given this proposition, our inference procedure exploits the fact that the optimality conditions (3)-(6) are just a set of linear equalities and inequalities in the unknowns $\theta$ and $\lambda$. More broadly, by utilizing the optimality conditions (3-6), we can transform the problem of inference on the solution set of a LP problem, which is difficult, to inference on parameters defined by a set of linear inequalities, which is a relatively straightforward exercise with an existing literature.

Specifically, inference on these conditions falls into the special class of inequality models considered in Shi and Shum (2015), for which computationally attractive procedures (not involving time-consuming bootstrap steps) are available for constructing joint confidence sets for $(\theta, \lambda)$, and projected confidence sets for $\theta$.

---

2Recall the dual LP problem corresponding to (1) is $\min_{\lambda \geq 0} b'\lambda$ subject to $A'\lambda = c$.

3Combining (4) and (6), we obtain $\lambda'(b - A\theta) = 0$, which is the usual complementary slackness condition for this problem.

4Indeed, characterizing the solution to a constrained optimization problem via the optimality conditions (3-6) is analogous to characterizing the solution to an unconstrained optimization problem using the first-order conditions, which underlies the usual approach for doing inference with M-estimators.
2.2 Quadratic Programming

A second class of problems covered by our method is the convex Quadratic Programming (QP) case.

\[ \begin{align*}
\min & \quad c^\prime \theta + \frac{1}{2} \theta^\prime Q \theta \\
\text{s.t.} & \quad A \theta \geq b,
\end{align*} \tag{7} \]

where \( Q \) is positive semi-definite. In this case, the KKT conditions are both necessary and sufficient (see Cottle, Pang, and Stone (1992), p. 4). These conditions are, first, primal feasibility:

\[ A \theta - b - s = 0, \tag{8} \]

where \( s \) is the vector of slackness variables; second, dual feasibility

\[ A' \lambda - c - Q \theta = 0; \tag{9} \]

and finally, the complementarity conditions

\[ \lambda^\prime s = 0 \]
\[ \lambda \geq 0 \]
\[ s \geq 0. \]

Because both \( \lambda_i \) and \( s_i \) are non-negative, it follows that \( \lambda^\prime s = 0 \) is equivalent to \( \lambda_i s_i = 0 \) \( \forall i \).

Using shorthand from the optimization literature, we write them collectively as

\[ 0 \leq \lambda_i \perp s_i \geq 0. \tag{10} \]

For inference, we consider the case where the coefficients in the QP problem, \((A, b, c, Q)\) are estimated and thus contain sampling error. Analogously to Proposition 1, we have:

**Proposition 2. (QP Inference).** Inference on \( \hat{\theta} \) defined as the solution to the QP problem (7) is equivalent to inference on \( \hat{\theta} \) satisfying the inequalities (8),(9),(10) evaluated at the estimated quantities \( \hat{A}, \hat{b}, \hat{c}, \hat{Q} \).

\(^5\)Wolak (1987) also exploits the duality theory for nonlinear programming in deriving test statistics for nonlinear parameter constraints in the linear regression model.
As in the LP case, the QP optimality conditions in Eqs. (8), (9), and (10) are linear inequalities with pre-estimated coefficients in the parameters.

**Remark.** The conditions (3)-(6) in the case of LP, and conditions (8)-(10) in the case of QP are both necessary and sufficient for *global optimality*. However, there could potentially exist multiple solutions satisfying optimality conditions.\(^6\) The inference procedures we propose in this paper are valid for both the cases of unique and multiple solutions, as discussed in Section 4 below.

### 2.3 Related literature

As far as we are aware, we are among the first to set forth inference theory for a quantity \(\hat{\theta}\) which is a solution to a “noisy” LP or QP problem, where the noise arises from the sample or estimation error in either the objective function or the constraints.\(^7\) Our approach is to exploit the optimality conditions (3-6) to show that doing inference on \(\hat{\theta}\) (defined in 2) is equivalent, as testing the inequality constraints (3-6). That is, the confidence set for \(\theta\) defined as the optimizer of (2) is equivalent to the confidence set for \(\theta\) we get by “inverting” the test of the inequalities formed from the optimality conditions of the underlying mathematical programming.

Our paper is related to work by Wolak (1987, 1989a, 1989b) on testing (in)equality constraints on parameters in linear and nonlinear econometric models. The duality in mathematical programming problems plays an important role in Wolak’s analysis, as it does in ours; however, he considers the case where the constraints are deterministic, while we focus on the case where both the coefficients in the constraints and the objective function are subject to sampling error. Guggenberger, Hahn, and Kim (2008) derive specification tests for moment inequality models by exploiting dual formulations of the constraints, but not in a mathematical programming context.

Our paper focuses on an inference method for the *solution* to a mathematical programming problem, which complements the inference methods for the *optimized criterion function* (or value function) such as Bhattacharya (2009) and Freyberger and Horowitz (2015).\(^8\)

---

\(^6\)The solution set in both LP and QP are convex. It is straightforward to establish this fact in LP. See Cottle et al. (1992) for a proof in the case of QP.

\(^7\)In the engineering literature, LP with noisy model constraints has been studied extensively under the umbrella of *robust linear programming*. The goal in robust LP is to obtain a *single* solution \(\theta\) which remains “optimal” in the presence of estimation error. In constrast, our goal is to solve the statistical inference problem of obtaining a *set of solutions* – the confidence set – that can include the true solution with pre-specified probability.

\(^8\)They consider inference for the value function \(\max_{\theta} c'\theta\) of a LP problem rather than the solution \(\arg\max_{\theta} c'\theta\).
Similarly, inference methods studied by Kaido, Molinari, and Stoye (2017) and Gafarov (2016) in moment inequality models, and by Mogstad, Santos, and Torgovitsky (2017) and Russell (2017) in treatment effect models can be viewed as inference methods for the value function to a mathematical programming problem.

In a different vein, Kline and Tamer (2016) consider Bayesian inference in a class of partially identified models in which point-identified “reduced-form” parameters can be mapped back to structural parameters of interest, an approach which can potentially also work for the problems in this paper.

3 Examples

Here we present a number of examples, which show the prevalence of these problems across different areas in economics. We begin with a classic QP problem from finance. The remainder are LP problems.

Example 1: Optimal portfolio selection. This is perhaps the most famous QP problem in economics, and will be our empirical example below. Suppose there are \( k \) assets, with expected return \( R \), and covariance matrix for the return on these assets \( Q \). In practice these two quantities are estimated from return data. \( \theta \) is the portfolio weight vector such that \( \sum_{i=1}^{k} \theta_i = 1 \).\(^9\) Clearly, \( \theta'Q\theta \) is the variance of portfolio return and \( R'\theta \) is the expected return on the portfolio. Given a targeted expected return \( \mu \), Markowitz (1952) considers the minimum risk, long-only portfolio by solving the following QP problem:

\[
\begin{align*}
\min & \quad \theta'Q\theta \\
\text{s.t.} & \quad R'\theta = \mu \\
& \quad 1'\theta = 1 \\
& \quad \theta \geq 0
\end{align*}
\tag{11}
\]

In many policy evaluation settings, we wish to compute optimal assignment strategies given preliminary estimates of payoffs from different potential assignments. We next present several examples of this.

Example 2: Two-sided matching. Let \( \mathcal{X} \) (resp. \( \mathcal{Y} \)) denote the types of men (resp. women). Given estimates of the payoff function \( \Phi(x, y) \)\(^10\) from matching a man with characteristics \( x \) to a woman with characteristics \( y \), the optimal marriage assignment (under

\(^9\)Negative weight means short position.

\(^10\)The papers by Choo and Siow (2006) or Galichon and Salanie (2015) present estimators for \( \Phi(x, y) \).
transferable utility, eg. Shapley and Shubik (1971)) takes the form of the following LP problem:

$$\max_{\mu_{xy}, \mu_{x0}, \mu_{0y}} \sum_{x \in X} \sum_{y \in Y} \mu_{xy} \Phi(x, y)$$

s.t. $$\sum_{x \in X} \mu_{xy} + \mu_{x0} = m_y \ \forall y \in Y$$

$$\sum_{y \in Y} \mu_{xy} + \mu_{0y} = n_x \ \forall x \in X$$

$$\mu_{ij} \geq 0, \quad i, j \in X \cup 0; \ j \in Y \cup 0.$$  

(12)

In the above, $$\mu_{xy}$$ denote the frequency of marriages among type $$x$$ men and type $$y$$ women, and $$\mu_{x0}$$ (resp. $$\mu_{0y}$$) denote the frequency of unmarried individuals among type $$x$$ men (resp. type $$y$$ women). $$n_x$$ and $$m_y$$ denote the total incidence of type $$x$$ men and type $$y$$ women in the population, and the constraints impose adding-up and nonnegativity restrictions on the number of couples and singles.

**Example 3: Roommate assignment.** Graham et al. (2006) and Bhattacharya (2009) consider the optimal grouping of pairs of individuals when complementarities or peer effects are present. For instance, consider the optimal assignment of roommates to college dorm-rooms, given estimates of the peer effects that roommates have on each others’ grades. Let $$b, w, o$$ denote, respectively, black students, white students, and students of other races, and let $$\gamma_{ij}$$, for $$(i, j) \in \{b, w, o\}$$, denote estimates of academic achievements (eg. GPA) for two roommates of type $$i$$ and type $$j$$. The school authority may wish to optimally assignment roommates to maximize the average academic achievements via the following LP problem:

$$\max_{\mu_{ij}} \left[ \mu_{bb}\gamma_{bb} + \mu_{ww}\gamma_{ww} + \mu_{oo}\gamma_{oo} + \mu_{wb}\gamma_{wb} + \mu_{wo}\gamma_{wo} + \mu_{bo}\gamma_{bo} \right]$$

s.t. $$2\mu_{ww} + \mu_{wo} + \mu_{wb} = 2\pi_w$$

$$2\mu_{bb} + \mu_{bo} + \mu_{wb} = 2\pi_b$$

$$2\mu_{oo} + \mu_{wo} + \mu_{bo} = 2\pi_o$$

$$\mu_{ij} \geq 0, \quad i, j \in \{w, b, o\}.$$  

(13)

where $$\pi_w$$, $$\pi_b$$, and $$\pi_o$$ denote the population incidence of white, black and students of other races, with $$\pi_w + \pi_b + \pi_o = 1$$.

**Example 4: Optimal treatment assignment under budget constraint.** Consider a binary ($$d \in \{0, 1\}$$) treatment, where the average treatment effects $$\beta_d(x)$$ for each treatment $$d = 0, 1$$ on individuals with characteristics $$x \in X$$ have been previously estimated (for instance, in an RCT). Bhattacharya and Dupas (2012) consider the following optimal treatment assignment, under a budgetary cap $$c$$ on the total number of $$d = 1$$ treatments
that can be administered:

\[
\max_{p(x), x \in \mathcal{X}} \left[ \sum_{x \in \mathcal{X}} \beta_1(x)p(x) + \beta_0(x)(1 - p(x)) \right] f(x)
\]

\[
s.t. \quad c = \sum_{x \in \mathcal{X}} p(x)f(x) \quad p(x) \geq 0
\]  

(14)

where \( f(x) \), for \( x \in \mathcal{X} \), denotes the fraction of individuals who have characteristics \( x \).

\[\square\]  

**Example 5: Market share prediction in semiparametric discrete choice models.**  
We wish to predict market shares in a semiparametric multinomial choice demand model.  
Following the treatment in Chiong et al. (2017), we observe market shares and covariates across \( M \) markets: \( \{s_m, X_m\}_{m=1}^M \). Assume we are given estimated utility parameters: \( U_m^k = \beta X_m^k \). Now we have a counterfactual market \( M+1 \) with covariates \( X_{M+1} \). The market shares \( s_{M+1} \) are not point identified, but must satisfy the cyclic monotonicity conditions taken across markets \( m = 1, 2, \ldots, M, M+1 \). Formally we estimate

\[
\max_{s_{M+1}} \sum_{k=1,2,3} p_{M+1}^k s_{M+1}^k \quad s.t. \quad CM(s_{M+1}^1; \hat{\beta}, \{s_m, X_m\}_{m=1}^M, X_{M+1}).
\]

CM denotes the linear inequalities arising from cyclic monotonicity. For instance, if we consider only length-2 cycles, then they are, for all \( m \in \{1, 2, \ldots, M\} \):

\[
(s_m - s_{M+1})(X_M^1 - X_m^1) \hat{\beta} \leq 0.
\]

We may be interested in other quantities. For instance, for a multiproduct firm which produces goods (say) 1,2,3, the highest counterfactual revenue is

\[
\max_{s_{M+1}} \sum_{k=1,2,3} p_{M+1}^k s_{M+1}^k \quad s.t. \quad CM(s_{M+1}; \hat{\beta}, \{s_m, X_m\}_{m=1}^M, X_{M+1})
\]

and the market shares of (say) good 2 among the set of revenue-maximizing market shares would be the argmax of this problem.  

\[\square\]

**Example 6: bounds on nonparametric regression function subject to shape restrictions.**  
Following Freyberger and Horowitz (2015), consider a nonparametric regression model \( Y = g(X) + U \) with \( \mathbb{E}[U|W = w] = 0 \ \forall w \); here \( Y \) is an outcome of interest, \( X \) is a possibly endogenous regressor and \( W \) is an instrument (and both \( X \) and \( W \) are finite-valued). We may wish to derive bounds on values of the finite-valued unknown function \( g \) which maximize a linear functional \( c'g \) subject to shape restrictions:

\[
\argmax_g c'g \quad s.t. \quad \Pi' g = m; \ S g \leq 0.
\]

\[\square\]

\[\text{11}\] For instance, the semiparametric estimation approach in Shi, Shum, and Song (2017) could be used.
4 Inference on parameter vector $\theta$

In this section we detail the inference procedure for LP with all-inequality constraints. In order to apply the computationally simple procedure of Shi and Shum (2015), we first introduce the $m \times 1$ vector $s$ of nonnegative slackness parameters. Then we can rewrite the primal-dual feasibility and linear complementarity conditions (3-6) as:

\begin{align*}
A\theta + s - b &= 0 \quad \text{(15)} \\
A'\lambda - c &= 0 \quad \text{(16)} \\
\lambda' s &= 0 \quad \text{(17)} \\
\lambda &\geq 0 \quad \text{(18)} \\
s &\geq 0 \quad \text{(19)}
\end{align*}

In this version of primal-dual formulation, the components of the model estimated with sampling error – $(A, b, c)$ – enter only the equalities (15 - 16), while the LC condition (17) serves as the nonlinear constraints on parameters. For equality constraints, there is no complementary slackness and non-negative constraint for $\lambda$. Eq. (15) and Eq. (16) are modified as

\begin{align*}
A\theta - b &= 0 \\
A'\lambda - c &= 0 
\end{align*}

Let $g(A, b, c, \theta, \lambda, s) = \begin{pmatrix} A\theta + s - b \\ A'\lambda - c \end{pmatrix}$. For any $m \times k$ matrix $W$, let $\text{vec}(W) = (W'_{1,1}, \ldots, W'_{1,k})'$ where $W_{i,j}$ is the $j$th column of $W$. Given another matrix $X$, the Kronecker product $W \otimes X$, is defined

\[ W \otimes X = \begin{pmatrix} w_{11}X, \ldots, w_{1k}X \\ \vdots \quad \vdots \\ w_{m1}X, \ldots, w_{mk}X \end{pmatrix}. \]

Using this notation, we can write $g(A, b, c, \theta, \lambda, s)$ as

---

12 The case of QP is similar and is discussed in Section 5 below.

13 Recently, Chen, Christensen, O’Hara, and Tamer (2016) also introduce slackness parameters to transform moment inequalities to equalities in their MCMC-based estimator for partially identified models.
\[
\begin{align*}
g(A, b, c, \theta, \lambda, s) &= \left( (\theta' \otimes I_m) \text{vec}(A) + s - \hat{b} \right) \\
&\quad \left( I_k \otimes \lambda' \right) \text{vec}(A) - c \\
&= \left( \begin{array}{ccc}
\theta' & -I_m & \mu_{m \times k} \\
I_k & 0_{k \times m} & I_k
\end{array} \right) \left( \begin{array}{c}
\text{vec}(A) \\
b \\
c
\end{array} \right) + \left( \begin{array}{c}
s \\
0_{k \times 1}
\end{array} \right).
\end{align*}
\]

Let \( G(\theta, \lambda, s) = \left( \begin{array}{ccc}
\theta' & -I_m & 0_{m \times k} \\
I_k & 0_{k \times m} & I_k
\end{array} \right) \). Suppose that \( A, b, c \) are estimated by \( \hat{A}, \hat{b}, \hat{c} \). Assume that
\[
\sqrt{n} \left( \begin{array}{c}
\text{vec}(\hat{A}) - \text{vec}(A) \\
\hat{b} - b \\
\hat{c} - c
\end{array} \right) \rightarrow_d N(0, V).
\]

Let \( V \) be estimated by \( \hat{V} \). Let
\[
\hat{Q}_n(\theta, \lambda, s) = g(\hat{A}, \hat{b}, \hat{c}, \theta, \lambda, s)'(G(\theta, \lambda, s)\hat{V}G(\theta, \lambda, s))^{-1}g(\hat{A}, \hat{b}, \hat{c}, \theta, \lambda, s)
\]

Accordingly, we construct a confidence set of confidence level \( 1 - \alpha \), which we denote \( CS_n^{PD}(1 - \alpha) \) (PD being short for “primal-dual”), as:
\[
CS_n^{PD}(1 - \alpha) = \{ \theta \in \Theta : \min_{\lambda \geq 0, s \geq 0, \lambda s = 0} n\hat{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1 - \alpha) \}.
\]

Computing the profile test statistic itself only involves a GMM objective function of linear moments, subject to LC constraints. This falls into the class of Mathematical Programming with Complementarity Constraints (MPCC) problems which are well-understood computationally.\(^\text{14}\) Therefore our method is user-friendly and is not computationally demanding.\(^\text{15}\)

In practice, it is usually convenient to report the upper and lower bound of the confidence set of each parameter. For example, for the parameter \( \theta_j \), one can report the confidence interval \([\hat{\theta}_j(1 - \alpha), \bar{\theta}_j(1 - \alpha)]\), which can be obtained by solving the following problems:

\(^{14}\)MPCC problems can be easily specified in the KNITRO interface for MATLAB; see https://www.artelys.com/tools/knitro_doc/2_userGuide/complementarity.html.

\(^{15}\)See Dong, Hsieh, and Shum (2017) for additional applications of MPCC in general moment inequality models.
\[ \theta_j(1 - \alpha) = \inf_{\theta} \theta \in \Theta : \theta \geq 0, \min_{\lambda \geq 0, s \geq 0, \lambda' s = 0} n \tilde{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1 - \alpha); \]

\[ \bar{\theta}_j(1 - \alpha) = \sup_{\theta} \theta \in \Theta : \theta \geq 0, \min_{\lambda \geq 0, s \geq 0, \lambda' s = 0} n \tilde{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1 - \alpha). \]  

(26)

4.1 Uniform coverage of confidence sets

Now we show the uniform asymptotic validity of our confidence sets. Note that the data enters our inference problem only through \( A, b, c, \) and \( V \). For clarity, we now let the data distribution \( P \) index these quantities, that is, we now write \( A_P, b_P, c_P, \) and \( V_P \). Then the solution set of the linear programming problem can be written as

\[ \Theta_0(P) = \{ \theta \in \Theta : \exists \lambda \geq 0, s \geq 0, \lambda' s = 0 \text{ s.t. } g(A_P, b_P, c_P, \theta, \lambda, s) = 0 \}. \]

This set is a singleton when the linear programming problem has a unique solution, but contains multiple values otherwise. Our confidence set is uniformly asymptotically valid within the set of data distributions \( P_0 \) such that the following assumption holds.

**Assumption 1.**

(a) For all \( P \in P_0, \Theta_0(P) \) is nonempty.

(b) For any sequence \( \{P_n\}_{n \geq 1} \) such that \( P_n \in P_0 \) for all \( n \), and \( V_{P_n} \to V \) for some positive semi-definite matrix \( V \), we have under \( \{P_n\}_{n \geq 1} \),

\[ \sqrt{n} \begin{pmatrix} \text{vec}(\tilde{A}) - \text{vec}(A_{P_n}) \\ \hat{b} - b_{P_n} \\ \hat{c} - c_{P_n} \end{pmatrix} \to_d N(0, V). \]

(c) There exists constants \( \varepsilon > 0 \) and \( C > 0 \) such that for all \( \theta \in \Theta_0(P), \lambda \geq 0, s \geq 0, \lambda' s = 0, g(A_P, b_P, c_P, \theta, \lambda, s) = 0, \) and all \( P \in P_0, \) we have that the smallest eigenvalue of \( G(\theta, \lambda, s)' V_P G(\theta, \lambda, s) \) is no smaller than \( \varepsilon \) and \( \|G(\theta, \lambda, s)\| \leq C. \)

(d) There exists a constant \( C > 0 \) such that \( \|V_P\| \leq C \) for all \( P \in P_0. \)

**Remarks.** (i) Part (b) of the assumption is a high-level assumptions that can be verified in the step where \( \tilde{A}, \hat{b}, \) and \( \hat{c} \) are obtained, and typically is implied by a central limit theorem for triangular arrays. Part (c) bounds the eigenvalue of the GMM weight matrix from below, and also bounds \( \theta \) and \( \lambda \) from above. These can be evaluated in practice based on estimators of the relevant quantities.

(ii) The assumptions do not rule out the case where the linear programming solution occurs on a flat face of the constraint set. We are able to obtain uniform asymptotic coverage.
without ruling out flat faces because our confidence set is a projection of an Anderson-Rubin type confidence set of the same confidence level for the full vector of unknown parameters. Similar confidence sets are proposed in, for example, Andrews and Soares (2010).

The theorem below states the uniform asymptotic validity.

**Theorem 1.** Suppose that Assumption 1 holds. Then we have for \( \alpha \in (0,1) \),

\[
\lim_{n \to \infty} \inf_{P \in \mathcal{P}_0} \inf_{\theta \in \Theta_0(P)} \Pr_P(\theta \in CS_{n}^{PD}(1 - \alpha)) \geq 1 - \alpha,
\]

where \( \Pr_P \) stands for probability under the data distribution \( P_n \).

Note that if the uniformity holds over \( \mathcal{P}_0 \), it also holds over any subset of \( \mathcal{P}_0 \). Thus, we immediately have the following corollary.

**Corollary 1.** Suppose that Assumption 1 holds, and \( \mathcal{P}_{00} = \{ P \in \mathcal{P}_0 : \Theta_0(P) \) is a singleton\}. Denote the singleton by \( \theta_0(P) \). Then we have for \( \alpha \in (0,1) \),

\[
\lim_{n \to \infty} \inf_{P \in \mathcal{P}_{00}} \Pr_P(\theta_0(P) \in CS_{n}^{PD}(1 - \alpha)) \geq 1 - \alpha.
\]

On the surface, the corollary is a trivial implication of Theorem 1, but it gives the coverage result a desirable implication: the confidence set covers the unique solution with prespecified probability asymptotically, uniformly over the set of data distributions under which the solution is unique. The set of data distributions under which the solution is unique is not closed in typical topology on the set of probability measures (e.g., the total variation topology). That implies that a sequence of \( P \) with unique solution can converge to a \( P_\infty \) with multiple solutions, i.e. a \( P_\infty \) under which the solution occurs on a flat face of the constraint set. The corollary shows that the uniform coverage does not break down along such sequences.

5 Monte Carlo Simulations

To investigate the performance of the Primal-Dual confidence set, we consider two simulation examples for the LP case.

5.1 Simulation 1

Consider the following population LP problem (1) with

\[
A = \begin{pmatrix} 1 & 2 \\ 1 & -1 \end{pmatrix}, \quad b = \begin{pmatrix} 4 \\ 1 \end{pmatrix}, \quad c = \begin{pmatrix} 3 \\ 2 \end{pmatrix}.
\]  

(27)
We further impose the solution $\theta$ is non-negative. We use these numbers as population means and generate normal random numbers with variance 1, and then compute the corresponding sample means $\hat{A}, \hat{b}, \hat{c}$. The solution of the population LP problem is $\theta = (2, 1)$. We report the empirical coverage rate of $CS^{PD}_{n}(1 - \alpha)$ (“Primal-Dual”) defined in Eq. (25) in Table 1. we can see that the empirical coverage rate is greater than the pre-specified confidence level. When the sample size increases, our confidence set also becomes more conservative. This over-conservative finding is consistent with the simulation results in Shi and Shum (2015), which is a general issue in sub-vector inference in inequality models including moment inequality models.

5.2 Simulation 2: Intersection Bounds

To further understand whether the over-conservative problem would lead to severe power loss, we conduct another simulation in which there exists alternative inference methods which are not conservative. Suppose we want to estimate $\theta = \max \{E X_1, E X_2\}$, which is a very simple example of intersection bounds. Fang and Santos (2016) show that non-conservative inference can be carried out by resampling that takes into account the directional derivative. Alternatively, this intersection bound can be written as the solution of the following LP problem.

$$\min_{\theta} \quad s.t. \quad \theta \geq E X_1, \theta \geq E X_2.$$ 

For this example we introduce two slackness parameters $s = [s_1, s_2]'$ and also two Lagrange multiplier $\lambda = [\lambda_1, \lambda_2]$, and the sample moment conditions are:

$$\frac{1}{N} \sum_i X_{1i} - \theta + s_1 = 0$$

$$\frac{1}{N} \sum_i X_{2i} - \theta + s_2 = 0$$

$$\lambda_1 + \lambda_2 - 1 = 0$$

$$0 \leq \lambda_1 \perp s_1 \geq 0$$

$$0 \leq \lambda_2 \perp s_2 \geq 0$$

Because here we have non-stochastic $A$ and $b$, only the first two equations (primal feasibility) are treated as moment conditions defined in (22). The dual feasibility and LC conditions are treated as constraints when computing the profiled test statistic in (25):

---

\^16See, eg., Chernozhukov, Lee, and Rosen (2013)

\^17For compatibility as stated in Eq. (1), we have $A = [-1 - 1]'$, $c = -1$, and $b = [-E X_1, -E X_2]'$. 

\[
\min_{\lambda_1 + \lambda_2 = 1, \lambda_i \geq 0, s_i \geq 0: \lambda_is_i = 0} n \hat{Q}_n(\theta, \lambda, s)
\]

The asymptotic variance $\hat{V}$ defined in (23) is the sample covariance matrix of $(X_1, X_2)$, while and Jacobian $G(\theta, \lambda, s)$ is the identity matrix. We report the coverage probability under different combinations of DGP and sample size in Table 2. In general, our method produces 98% coverage rate, which is slightly higher than the desired 95% coverage, while Fang and Santos’s (2016) method produces correct empirical coverage. However, when we plot these two sets of confidence intervals in Figure 1, it is apparent that our confidence intervals (labelled “Primal-Dual”) are only slightly longer, suggesting that the distortion from conservatism of our approach does not seem too severe for this example.

6 Empirical Application: Portfolio Selection

As an empirical application, we consider the portfolio allocation problem in finance (Markowitz (1952)), which was discussed as Example 1 above. Using our approach, we compute the confidence set for the optimal weights.

While the optimal portfolio allocation problem is one of the classic problems in finance and is utilized ubiquitously in real-world investment operations, the distributional theory for the model has not received much attention. Jobson and Korkie (1980) derive the sampling theory for the optimal weights under the unrealistic assumptions of allowing for short positions and imposing normality on the return data. For the typical case, in which short positions are ruled out, practitioners often rely on bootstrap inference (see Scherer (2002)). However, canonical bootstrap is not valid in this setting, due to nondifferentiability of the solutions (see Fang and Santos (2016)). Our method is valid in this general case as it can accommodate different types of constraints commonly encountered in practice, as well as weaker requirements for the underlying DGP.

Specifically, we consider a problem of forming an optimal portfolio among $k$ assets with weights $\{\theta_1, \ldots, \theta_k\} \equiv \theta$ in order to minimize portfolio risk

\[
\min_{\theta} \theta' Q \theta
\]

subject to two primal feasibility conditions

\[
R' \theta - \mu = 0 \quad \quad 1' \theta - 1 = 0
\]  \hspace{1cm} (29)

\footnote{For exposition purpose only the first 50 intervals from the first design in Table 2 are displayed.}
and $k$ dual feasibility conditions:

$$\lambda_\theta + \lambda_R R + \lambda_F 1 - Q\theta = 0.$$  \hfill (30)

In the above, $\lambda_\theta$ is the vector of Lagrange multipliers of the non-negativity constraints on $\theta$ (corresponding to the restriction to long positions), and $\lambda_R, \lambda_F$ are the Lagrange multipliers of the equality constraints of targeted return and feasible portfolio weights. There are $k$ linear complementarity conditions: $0 \leq \lambda_\theta \perp \theta \geq 0$. Because the portfolio weight constraint $1'\theta - 1 = 0$ does not involve estimated quantities, we exclude it from the moment conditions and incorporate it directly in the definition of the test statistic (see Eq. (31) below). The remaining primal-dual conditions can then be expressed in terms of the following moment conditions

$$g(\hat{Q}, \hat{R}, \mu; \theta, \lambda) = \begin{pmatrix} \hat{R}' & 0' & 0 \\ -\hat{Q} & I_{k \times k} & \hat{R} & 1 \end{pmatrix} \cdot \begin{pmatrix} \theta \\ \lambda_\theta \\ \lambda_R \\ \lambda_F \end{pmatrix} - \begin{pmatrix} \mu \\ 0 \end{pmatrix}$$

$$= \begin{pmatrix} 0_{1 \times k^2} & \theta' \\ -\theta' \otimes I_{k \times k} & \lambda_R I_{k \times k} \end{pmatrix} \cdot \begin{pmatrix} \text{vec}(\hat{Q}) \\ \hat{R} \end{pmatrix} + \begin{pmatrix} -\mu \\ \lambda_\theta + \lambda_F 1 \end{pmatrix}$$

The portfolio feasibility constraint $1'\theta - 1 = 0$ is instead treated as a parameter constraint when computing the test statistic:

$$CS_n^{PD}(1 - \alpha) = \{ \theta \in \Theta : \min_{1'\theta - 1 = 0, 0 \leq \lambda_\theta \perp \theta \geq 0} n\hat{Q}_n(\theta, \lambda) \leq \chi^2_{1+k}(1 - \alpha) \}. \hfill (31)$$

We consider portfolio selection over three fixed income securities: 10-year Treasury Bill, AAA corporate bond and BBB corporate bond. We calculate the annualized returns (in percentage terms) and corresponding covariance matrix for these assets using the daily data from January 4, 2010 to July 31, 2017:\footnote{We use 10-Year Treasury Constant Maturity Rate, and BofA Merrill Lynch US Corporate AAA and BBB Effective Yield, downloaded from Federal Reserve Bank of St. Louis, \url{https://fred.stlouisfed.org/}.}

$$\hat{R} = \begin{pmatrix} \text{T-Bill} \\ \text{AAA} \\ \text{BBB} \end{pmatrix} = \begin{pmatrix} 2.2550 \\ 2.5137 \\ 3.9256 \end{pmatrix}, \hat{Q} = \begin{pmatrix} 0.5976 & \cdot & \cdot \\ 0.2336 & 0.2674 & \cdot \\ 0.2758 & 0.2285 & 0.4488 \end{pmatrix}. \hfill (32)$$

In Figure 2, we depict the confidence set under different target return $\mu$. Since the confidence set is three-dimensional, we present the two-dimensional projections for each pair of assets separately, for three different values of the target return $\mu \in \{2.3\%, 2.5\%, 3.0\%\}$. 
Despite the conservative tendency of our approach, in this case it does yield tight estimates. The upper panel of Figure 2 suggests that, when the target return is low ($\mu = 2.3\%$), the confidence set is a set of linear combinations among T-Bill and AAA Corporate bonds; the portfolio weight in BBB Corporate bonds is always zero and they are excluded from the optimal portfolio. As the required return increases, however, to $\mu = 2.5\%$ and $\mu = 3\%$, the confidence sets depicted in the lower panels of Figure 2 takes on an “elliptical” shape, indicating that the portfolios in the confidence set typically involve non-zero amount of all three assets.

Clearly, proper inference on the optimal portfolio weights is an important input for making investment decisions. When new data on asset returns become available, the portfolio weights based on the updated returns can differ from old weights, thus raising the question whether an investor should adjust his portfolio using the new weights (and incur the transaction costs from doing so), or keep it at the old weights. This decision depends on whether the difference between them is statistically significant given the sampling noise in estimating the returns. To address this question, we could use our inference procedures to compute the confidence set for the new weights. If this confidence set includes the old weights, then the new and old weights are not statistically distinguishable and the investor may not need to adjust her portfolio.

7 Conclusion

We propose an inference procedure for estimators defined as optimizers of stochastic versions linear and quadratic programming problems with pre-estimated coefficients in the objective function or constraints. The Karush-Kuhn-Tucker conditions which characterize the optimum are re-interpreted as inequalities with pre-estimated coefficients. We provide an empirical application to the portfolio selection problem in finance; as far as we are ware, this represents the first instance of inference for this classic problem based on asymptotic approximation.

More broadly, since KKT conditions can be applied in nonlinear programming problems with suitable constraint qualification conditions, our inference approach might also work in those more general contexts.\(^{20}\) When the resulting inequalities are moment inequalities, one can use the well-established methods in the moment inequality literature (e.g. Andrews and Soares (2010), and Andrews and Barwick. (2012), among others) to construct joint confidence sets for $(\theta, s, \lambda)$ and then obtain the marginal confidence set for $\theta$ as projection of the joint confidence sets. For methods that focuses on marginal confidence sets for $\theta$ (which usually yield tighter inference than the simple projection method above), one could

\(^{20}\)Martin (1985) established KKT sufficiency for a wide class of mathematical programming problems.
use more elaborate methods such as Kaido et al. (2017) or Bugni, Canay, and Shi (2017).

References


A Proof of Theorem 1

Proof. For notational simplicity define the set
\[ \Delta_0(P) = \{ (\theta', \lambda', s') : \lambda \geq 0, s \geq 0, \lambda's = 0, g(A_P, b_P, c_P, \theta, \lambda, s) = 0 \}. \]

To show the theorem, it suffices to show that for any sequence \( \{(P_n, \theta_n)\}_{n \geq 1} \) such that \( P_n \in \mathcal{P}_0 \) and \( \theta_n \in \Theta_0(P_n) \) for all \( n \), we have
\[ \liminf_{n \to \infty} \Pr_{\theta_n} (\theta_n \in CS_{n}^{PD}(1 - \alpha)) \geq 1 - \alpha. \] (33)

By the definition of \( CS_{n}^{PD}(1 - \alpha) \), it suffices to show that there exists \( (\lambda_n, s_n) \) such that \( (\theta'_n, \lambda'_n, s'_n) \in \Delta_0(P) \) such that
\[ \liminf_{n \to \infty} \Pr_{\theta_n} (n \hat{Q}_n(\theta_n, \lambda_n, s_n) \leq \chi^2_{m+k}(1 - \alpha)) \geq 1 - \alpha. \] (34)

By the definition of infimum, there exists a subsequence of \( \{a_n\}_{n \geq 1} \) of \( \{n\}_{n \geq 1} \) such that the left hand-side of (34) equals
\[ \lim_{n \to \infty} \Pr_{\theta_n} (n \hat{Q}_n(\theta_n, \lambda_n, s_n) \leq \chi^2_{m+k}(1 - \alpha)). \] (35)

Due to the completeness of the Euclidean space and Assumption 1(c)-(d), there exists a subsequence \( \{u_n\}_{n \geq 1} \) of \( \{a_n\}_{n \geq 1} \) such that \( V_{P_{u_n}} \to V \) as \( n \to \infty \) for some positive semi-definite matrix \( V \), and \( G(\theta_{u_n}, \lambda_{u_n}, s_{u_n}) \to G \) for some \( m + k \) by \( mk + m + k \) matrix \( G \). Also by Assumption 1(c), \( GV_{G'} \) is invertible, and by Assumption 1(b), under the sequence \( \{P_{u_n}\}_{n \geq 1} \),
\[ \sqrt{u_n} \begin{pmatrix} \text{vec}(\hat{A}) - \text{vec}(A_{P_{u_n}}) \\ \hat{b} - b_{P_{u_n}} \\ \hat{c} - c_{P_{u_n}} \end{pmatrix} \to_d N(0, V). \]

Therefore, under the sequence of data distributions \( \{P_{u_n}\}_{n \to \infty} \),
\[ \sqrt{u_n} g(\hat{A}, \hat{b}, \hat{c}, \theta_{u_n}, \lambda_{u_n}, s_{u_n}) \to_d N(0, GV_{G'}). \]

That implies that \( u_n \hat{Q}_{u_n}(\theta_{u_n}, \lambda_{u_n}, s_{u_n}) \to_d \chi^2_{m+k} \).

Consequently, the limit in (35) is 1 - \( \alpha \), and hence (34) holds, proving the theorem.
### Tables and Figures

Table 1: Simulation 1: Empirical Coverage Rate of $CS^{PD}$ ("Primal-Dual") at 95% Confidence Level.

<table>
<thead>
<tr>
<th>sample size</th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>true parameter = $(2, 1)^a$</td>
<td>0.986</td>
<td>0.99</td>
<td>0.996</td>
</tr>
</tbody>
</table>

$a$1000 Monte Carlo Repetitions

Table 2: Simulation 2: Empirical Coverage Rate of $CS^{PD}$ ("Primal-Dual") at 95% Confidence Level.

<table>
<thead>
<tr>
<th>sample size</th>
<th>n=100</th>
<th>n=200</th>
<th>n=500</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\theta = \max{\mathbb{E}X_1, \mathbb{E}X_2}^a$</td>
<td>Method</td>
<td></td>
<td></td>
</tr>
<tr>
<td>design 1$^b$</td>
<td>Primal-Dual</td>
<td>0.983</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>Fang-Santos</td>
<td>0.952</td>
<td>0.955</td>
</tr>
<tr>
<td>design 2$^c$</td>
<td>Primal-Dual</td>
<td>0.983</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>Fang-Santos</td>
<td>0.952</td>
<td>0.952</td>
</tr>
<tr>
<td>design 3$^d$</td>
<td>Primal-Dual</td>
<td>0.983</td>
<td>0.982</td>
</tr>
<tr>
<td></td>
<td>Fang-Santos</td>
<td>0.953</td>
<td>0.954</td>
</tr>
</tbody>
</table>

$^a(X_1, X_2) \sim \mathcal{N}(\mu, \Sigma)$. 1000 Monte Carlo Repetitions.

$^b\mu = (5, 3), (\sigma_1^2, \sigma_2^2) = (1, 1), \sigma_{12} = 0$

$^c\mu = (5, 3), (\sigma_1^2, \sigma_2^2) = (1, 3), \sigma_{12} = 0$

$^d\mu = (5, 3), (\sigma_1^2, \sigma_2^2) = (1, 3), \sigma_{12} = 1.5$
Figure 1: Comparing Confidence Intervals in Simulation 2

Note: 95% confidence level. The tuning parameter $\kappa_n$ in the resampling CI of Fang and Santos (2016) is set to $\frac{1}{\log(n)}$, and number of bootstrap samples = 10,000.
Figure 2: 90% Confidence Set of Efficient Portfolio Weights under Different Target Return $\mu$

Note: The solution of portfolio selection (11) based on the estimated $(\hat{R}, \hat{Q})$ is located by two red lines.