Inference on Estimators defined by Mathematical Programming*

Yu-Wei Hsieh[†] Xiaoxia Shi[‡] Matthew Shum[§]

June 9, 2021

Abstract

We propose an inference procedure for a class of estimators defined as the solutions to linear and convex quadratic programming problems in which the coefficients in both the objective function and the constraints of the problem are estimated from data and hence involve sampling error. We argue that the Karush-Kuhn-Tucker conditions that characterize the solutions to these programming problems can be treated as moment conditions; by doing so, we transform the problem of inference on the solution to a constrained optimization problem (which is non-standard) into one involving inference on inequalities with pre-estimated coefficients, which is better understood. Our approach is valid regardless of whether the problem has a unique solution or multiple solutions. We apply our method to various portfolio selection models, in which the confidence sets can be non-convex, lower-dimensional manifolds.

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

JEL Classification: C10, C12, C63

*We thank Denis Chetverikov, Jin-Chuan Duan, Bulat Gafarov, Bryan Graham, Jinyong Hahn, Po-Hsuan Hsu, Yuichi Kitamura, Emerson Melo, Ismael Mourifié, Andres Santos, Yixiao Sun; seminar listeners at Chinese University of Hong Kong, UC-Riverside and UC-San Diego; and attendees at the California Econometrics Conference (10/17), the CeMMAP Conference on Machine Learning and Optimisation (3/18), the Rotterdam Workshop on Machine Learning and Causal Inference (5/18), 2018 North American Summer Meeting of the Econometric Society at UC-Davis, and 2018 Taiwan Economic Research Conference for comments.

[†]Amazon.com. Email: yuweihsieh01@gmail.com. Hsieh's contributions to the paper reflect work done prior to joining Amazon.

[‡]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

We consider the problem of inference on a class of estimators defined as the solution to a convex programming problem with pre-estimated coefficients. In particular, we focus on linear programming (LP) and convex quadratic programming (QP) problems. 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 permit standard bootstrap inference.

The core of our method lies in recognizing that the necessary and sufficient optimality conditions for certain types the convex programming problems can be interpreted as a set of moment inequalities. Specifically, these optimality conditions—the Karush-Kuhn-Tucker (KKT) conditions—involve Lagrange multipliers, slackness variables, and a set of *linear complementarity* (LC) conditions. In particular, we apply the computationally convenient procedures from Shi and Shum (2015) to implement the inference on the optimality conditions of the underlying mathematical program. The distinctive structure of the KKT conditions—complementary slackness—has implications for the geometry of the resulting confidence sets: The augmented parameter space involving the model parameters and the Lagrange/slackness parameters is non-convex, potentially implying confidence sets for the model parameters which are lower-dimensional manifolds; this does not arise in typical moment inequality models.

The class of problems we consider cover several essential models in economics and finance. One notable example is Markowitz's (1952) classic optimal portfolio selection problem. 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 (2019) show that canonical bootstrap is not valid if the solution is non-differentiable in the estimated coefficients. As the solutions to mathematical programs are non-differentiable in general, our approach provides, to the best of our knowledge, the first valid inference method in the literature.[1]

[1] Although there is a literature on the asymptotic properties of the solutions to the stochastic programming problems, it is unclear how to turn these theories into a practical inference method. Shapiro (1993) studies

An essential 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. (Multiple solutions occur when the solution of a LP is located on a "flat face" of the constraint set.) In the former case, our confidence set covers the unique solution with pre-specified level, while in the latter case, each point in the solution set is covered with pre-specified 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.

We apply our method to the portfolio selection models that involve mathematical programming in finance. It is well-known that portfolio selection models, in particular the mean-variance (MV) portfolio of Markowitz (1952), are sensitive to estimation error in the input parameters; see, e.g., Michaud (1989). Despite these concerns, there is little literature on statistical inference for these models.[2] As far as we are aware, our empirical analysis of the portfolio selection models here represents the first instance of valid asymptotic inference that can handle a rich variety of constraints including the NC portfolio of DeMiguel, Garlappi, Nogales, and Uppal (2009), and the Equally-Weighted Risk Contribution portfolio of Maillard, Roncalli, and Teiletche (2010). By contrast, Jobson and Korkie (1980), Britten-Jones (1999), Okhrin and Schmid (2006), and Kan and Smith (2008) also study the sampling theory for the optimal weights. However, they only consider the MV portfolio without short-selling constraints, in which there exists a closed form solution to the programming problem; see, e.g., Merton (1972). Furthermore, this literature typically relies on assuming that the return data are normally distributed. Our method does not require this stringent assumption on the data. In Monte Carlo exercises, we show that our method performs reasonably well when the data are generated from t-distributions or follow a factor structure.

In the next section we review the key results from the theory of linear and quadratic programming, and we provide examples in Section 3. Section 4 is the heart of the paper that illustrates our inference procedure. In section 5, we provide an empirical demonstration for several portfolio selection models, and we investigate the empirical size and power

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)). King (1989) studies the generalized delta method that can be applied to the KKT conditions. As noted in his seminal paper, however, it requires fairly strong assumptions on the semi-differentiability, and hence only a simple quadratic program with deterministic constraints is considered.

[2] Researchers have attempted to address the issues of parameter uncertainty from the angle of Bayes (e.g., Garlappi, Uppal, and Wang (2007)) or robust programming (e.g., Goldfarb and Iyengar (2003)).

performances using Monte Carlo.

2 Linear Programming and Quadratic Programming

In this paper, we focus on the cases of linear programming (LP) and convex quadratic programming (QP), in which all constraints are linear. We focus on such convex problems because the constraint qualification automatically satisfied here as long as the problem is feasible (Boyd and Vandenberghe (2004), §5.2.3.). Our framework may be applied to more general convex programs, but they require case-specific examination of constraint qualifications. For expositional purposes, we will not go into details on those.

2.1 Linear programming

The goal is to conduct inference for the solution to the (possibly non-unique) optimizer of the following LP:

$$\max_{\theta\in\Theta} \quad c'\theta \quad \text{s.t.} \quad A\theta \leq b, \tag{1}$$

where Θ is a known polytope in 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 $\hat{c}$. 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 (see, e.g., Mangasarian (1969)) are

$$\begin{aligned} A\theta &\leq b && (2)\\ A'\lambda &= c && (3)\\ \lambda &\geq 0 && (4)\\ c'\theta &= b'\lambda. && (5) \end{aligned}$$

Equation (2) and (3) express, respectively, the primal and dual feasibility, where λ is interpreted as the $m \times 1$ vector of Lagrange multipliers on the inequalities (2).[3] The final equation (5) is a complementarity condition, analogous to the complementarity slackness equation in the KKT conditions.[4] 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

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

[4]Combining (3) and (5), we obtain $\lambda'(b - A\theta) = 0$, which is the usual complementary slackness condition for this problem.

sets of (in-)equalities as *primal-dual conditions*. These considerations yield the following key proposition.

Proposition 1. *Any θ solving the LP problem (1) satisfies the inequalities (2)-(5) and vice versa.*

Given this proposition, our inference procedure exploits the fact that the optimality conditions (2)-(5) are just a set of linear equalities and inequalities in the unknowns θ and λ. More broadly, by utilizing the optimality conditions (2)-(5), we can transform the challenging problem of inference on the solution set of a LP problem to inference on parameters defined by a set of linear inequalities, which is relatively transparent given the existing literature.[5] 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 (θ, λ) and projected confidence sets for θ.

2.2 Quadratic Programming

A second class of problems covered by our method is convex Quadratic Programming (QP) problems. The goal is to conduct inference on the possibly non-unique solution of the problem:

$$\begin{array}{ll} \min_{\theta\in\Theta} & c'\theta + \frac{1}{2}\theta' Q\theta \\ \text{s.t.} & A_{ineq}\theta \geq b_{ineq}, \\ & A_{eq}\theta = b_{eq}, \end{array} \tag{6}$$

where Q is positive semi-definite and Θ is a known polytope in R^k. In this case, the KKT conditions are both necessary and sufficient (see e.g., Cottle, Pang, and Stone (1992)). These conditions are, first, primal feasibility:

$$A\theta - b - s = 0, \tag{7}$$

where $A = \binom{A_{ineq}}{A_{eq}}$, $s = \binom{s_{ineq}}{\mathbf{0}}$ is the vector of slackness variables with the length of s_{ineq} equal to the number of rows in A_{ineq}; second, dual feasibility:

$$A'\lambda - c - Q\theta = 0; \tag{8}$$

[5] Indeed, characterizing the solution to a constrained optimization problem via the optimality conditions (2)-(5) 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.

and finally, the complementarity conditions:

$$\begin{aligned}\lambda' s &= 0\\ \lambda &\geq 0\\ s_{ineq} &\geq 0.\end{aligned}$$

Because both λ_i and s_i are non-negative, it follows that $\lambda' s = 0$ is equivalent to $\lambda_i s_i = 0 \quad \forall i$. Following the optimization literature, we write them collectively as

$$0 \leq \lambda_i \perp s_i \geq 0. \tag{9}$$

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 therefore have the following statement for QP:

Proposition 2. *Any θ solving the QP problem (6) satifies the inequalities (7),(8),(9) and vice versa.*

Remark. The conditions (2)-(5) in the case of LP, and conditions (7)-(9) in the case of QP are both necessary and sufficient for *global optimality*. However, there could potentially exist multiple global optima satisfying the 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 an inference theory for a quantity $(\hat{\theta})$ which is a solution to a "noisy" mathematical programming problem, where the noise arises from the sample or estimation error in both the objective function and constraints. The operations research literature has extensively studied the *robust programming* problem, which is essentially LP/QP with noisy model constraints. The goal in robust programming is to obtain a *single* solution θ which remains "optimal" in the presence of error. In contrast, 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.

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

Our paper is related to the 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 the mathematical programming context.

Our paper focuses on an inference method for the *solution* of 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).[7] Similarly, inference methods studied by Kaido, Molinari, and Stoye (2019) 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, rather than the optimum, of mathematical programming problems.

3 Examples

We next present a number of examples demonstrating the prevalence of these problems across different areas in economics.

Example 1: Optimal portfolio selection. This is perhaps the most famous QP problem in economics, and serves as our empirical application below. Suppose there are k assets, with expected return R, and covariance matrix for the return on these assets Q. Q and R are estimated from return data. The parameter of interest θ is the portfolio weight vector such that $\sum_{i=1}^{k} \theta_i = 1$.[8] The variance of the portfolio return is $\theta' Q \theta$ and $R'\theta$ is the expected return on the portfolio. Given a targeted expected return μ, Markowitz (1952) considers the optimized minimum risk, long-only portfolio which solves the following QP problem:

$$
\begin{aligned}
\min \quad & \theta' Q \theta \\
\text{s.t.} \quad & R'\theta = \mu \\
& \mathbf{1}'\theta = 1 \\
& \theta \geq 0.
\end{aligned}
\tag{10}
$$

[7] They consider inference for the value function ($\max_\theta c'\theta$) of a LP problem rather than the solution $\mathrm{argmax}_\theta c'\theta$.

[8] Negative weight means short position.

Let $\Theta_0(Q, R, \mu)$ be the argmin set. Practitioners may be interested in testing whether a given weight vector θ_0 is optimal; that is, whether $\theta_0 \in \Theta_0(Q, R, \mu)$; see, e.g., Scherer (2002). Another interesting hypothesis, considered in Britten-Jones (1999), is whether a given set of restrictions on the allocation weights affects optimality; that is, whether the optimal solution satisfies $C\theta \geq \mathbf{r}$ for a given matrix C and a given vector $\mathbf{r}$.[9] □

In many matching or treatment assignment problems, the optimal assignment strategies is defined as solutions to LP or QP problems, in which components in the objective function or constraints have been pre-estimated from observational data. Practitioners interested in program evaluation may be interested in testing whether a given assignment is optimal, or in testing whether a given set of restrictions on the assignment rule affects overall welfare. We next present two examples of this.

Example 2: Roommate assignment. Graham et al. (2006) and Bhattacharya (2009) consider the optimal grouping of pairs of individuals when complementarities or peer effects are present. Bhattacharya (2009) studies 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 γ_{ij}, for $(i, j) \in \{b, w, o\}$, denote estimates of average academic achievements (eg. GPA) for two roommates of type i and type j. The school authority may wish to optimally assign roommates to maximize the average academic achievements via the following LP problem:

$$\begin{aligned} \max_{\mu_{ij}: i,j \in \{w,b,o\}} \quad & [\mu_{bb}\gamma_{bb} + \mu_{ww}\gamma_{ww} + \mu_{oo}\gamma_{oo} + \mu_{wb}\gamma_{wb} + \mu_{wo}\gamma_{wo} + \mu_{bo}\gamma_{bo}] \\ \text{s.t.} \quad & 2\mu_{ww} + \mu_{wo} + \mu_{wb} = 2\pi_w \\ & 2\mu_{bb} + \mu_{bo} + \mu_{wb} = 2\pi_b \\ & 2\mu_{oo} + \mu_{wo} + \mu_{bo} = 2\pi_o \\ & \mu_{ij} \geq 0, \quad i, j \in \{w, b, o\}, \end{aligned} \tag{11}$$

where π_w, π_b, and π_o denote the proportion of white, black and students of other races in the population, with $\pi_w + \pi_b + \pi_o = 1$. In the above problem, the choice variables $\{\mu_{ij}\}$ denote the proportion of dorm rooms consisting of type i and type j individuals.

In practice, there may be substantial sampling variation in the estimates of academic achievement γ_{ij}, perhaps due to small samples from which these estimates are obtained.[10]

[9] An equality may be expressed as two inequalities and in this way $H_0 : C\theta \geq \mathbf{r}$ covers linear equalities as well.

[10] For example, in Bhattacharya (2009), samples of only 436 male and 428 female students are used to

In this case, inference on the optimizers $\{\mu_{ij}\}$ provides policymakers with a sense of how robust the optimal assignment is to small changes in the estimates of $\{\gamma_{ij}\}$. $\square$

Example 3: 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 $x \in \mathcal{X}$ have been previously estimated (for instance, in an RCT). Bhattacharya and Dupas (2012) consider the following optimal treatment assignment problem, under a budgetary cap c on the total number of $d = 1$ treatments that can be administered:

$$\begin{aligned} \max_{p(x), x\in\mathcal{X}} \quad & \left[\textstyle\sum_{x\in\mathcal{X}} \beta_1(x)p(x) + \beta_0(x)(1-p(x))\right] f(x) \\ \text{s.t.} \quad & c = \textstyle\sum_{x\in\mathcal{X}} p(x)f(x) \\ & p(x) \geq 0, \end{aligned} \tag{12}$$

where $f(x)$, for $x \in \mathcal{X}$, denotes the fraction of individuals who have characteristics x.[11] $\square$

Example 4: 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: $\{\mathbf{s}_m, \mathbf{X}_m\}_{m=1}^M$, where $\mathbf{X}_m = \begin{pmatrix} X_m^1 \\ X_m^2 \\ \cdots \\ X_m^K \end{pmatrix}$, $k = 1, \ldots, K$ are indices for the K products and X_m^k is the (row) vector of covariates for product k in market m. Assume we are given estimated utility parameters $\hat{\beta}$ for β in $U_m^k = X_m^k \beta$.[12] Now we have a counterfactual market $M+1$ with covariates $\mathbf{X}_{M+1}$. The market shares $\mathbf{s}_{M+1}$ are not point identified, but must satisfy the cyclic monotonicity conditions taken across markets $m = 1, 2, \ldots, M, M+1$. Formally, we estimate

$$\max_{\mathbf{s}_{M+1}} s_{M+1}^k \quad s.t. \quad CM(\mathbf{s}_{M+1}; \hat{\beta}, \{\mathbf{s}_m, \mathbf{X}_m\}_{m=1}^M, \mathbf{X}_{M+1}).$$

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

$$(\mathbf{s}_m - \mathbf{s}_{m+1})(\mathbf{X}'_{m+1} - \mathbf{X}'_m)\hat{\beta} \leq 0.$$

estimate average academic outcomes conditional on race composition for each roommate pair.

[11] See Andrews, Kitagawa, and McCloskey (2019a) for an extension of this problem to consider inferences on outcomes conditional on optimal assignments.

[12] For instance, the semiparametric estimation approach in Shi, Shum, and Song (2018) could be used.

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

$$\max_{\mathbf{s}_{M+1}} \sum_{k=1,2,3} p^k_{M+1} s^k_{M+1} \quad s.t. \quad CM(\mathbf{s}_{M+1}; \hat{\beta}, \{\mathbf{s}_m, \mathbf{X}_m\}_{m=1}^M, \mathbf{X}_{M+1}),$$

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

Example 5: 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). Our methods can be used for constructing a uniform confidence band for the finite-valued unknown function g which maximizes a linear functional $c'g$ subject to shape restrictions:

$$\text{argmax}_g c'g \quad \text{s.t.} \quad \Pi' g = m; \; Sg \leq 0.$$

□

4 Inference on parameter vector θ

In this section we detail the inference procedure for LP with all-inequality constraints. (The case of QP is similar and we will discuss it in the context of an empirical application in Section 5 below.) 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 (2)-(5) as:

$$A\theta + s - b = 0, \tag{13}$$

$$A'\lambda - c = 0, \tag{14}$$

$$\lambda' s = 0, \tag{15}$$

$$\lambda \geq 0, \tag{16}$$

$$s \geq 0. \tag{17}$$

In this version of primal-dual formulation, the components of the model estimated with sampling error—(A, b, c)—enter only the equalities (13) - (14), while the LC condition (15) imposes nonlinear constraints on parameters. This specific structure enables us to apply Shi and Shum's (2015) approach, which is computationally convenient. We provide a detailed comparison of this approach versus more general approaches for moment inequality models in Section 4.3 below.[13]

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 $\text{vec}(W) = (W_{\cdot,1}', \ldots, W_{\cdot,k}')'$ where $W_{\cdot,j}$ is the jth column of W. Using this notation, we can write $g(A,b,c,\theta,\lambda,s)$ as

$$
\begin{aligned}
g(A,b,c,\theta,\lambda,s) &= \begin{pmatrix} (\theta' \otimes I_m)\text{vec}(A) + s - b \\ (I_k \otimes \lambda')\text{vec}(A) - c \end{pmatrix} \\
&= \begin{pmatrix} \theta' \otimes I_m & -I_m & 0_{m\times k} \\ I_k \otimes \lambda' & 0_{k\times m} & I_k \end{pmatrix} \begin{pmatrix} \text{vec}(A) \\ b \\ c \end{pmatrix} + \begin{pmatrix} s \\ 0_{k\times 1} \end{pmatrix}. \qquad (18)
\end{aligned}
$$

Let $G(\theta,\lambda) = \begin{pmatrix} \theta' \otimes I_m & -I_m & 0_{m\times k} \\ I_k \otimes \lambda' & 0_{k\times m} & I_k \end{pmatrix}$. Suppose that A, b, c are estimated by $\widehat{A}_n, \hat{b}_n, \hat{c}_n$, and let $\widehat{V}_n$ denote the estimated asymptotic covariance matrix for $(\text{vec}(\widehat{A}_n), \hat{b}_n, \hat{c}_n)$. Let

$$
\widehat{Q}_n(\theta,\lambda,s) = g(\widehat{A}_n,\hat{b}_n,\hat{c}_n,\theta,\lambda,s)'(G(\theta,\lambda)\widehat{V}_n G(\theta,\lambda)')^{-1} g(\widehat{A}_n,\hat{b}_n,\hat{c}_n,\theta,\lambda,s). \qquad (19)
$$

We construct a confidence set of confidence level $1-\alpha$, which we denote $CS_n^{\text{PD}}(1-\alpha)$ (PD being short for "primal-dual"), as:

$$
CS_n^{\text{PD}}(1-\alpha) = \left\{ \theta \in \Theta : \min_{\lambda \geq 0, s \geq 0 : \lambda' s = 0} n\widehat{Q}_n(\theta,\lambda,s) \leq \chi^2_{m+k}(1-\alpha) \right\}. \qquad (20)
$$

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; see, e.g., Luo, Pang, and Ralph (1996).[14] For small-scale problems, one may formulate the MPCC problem as a mixed integer nonlinear programming problem. For large-scale problems, specialized MPCC solvers, such as KNITRO or PATH, are available, which reduces the computational cost.[15]

[13] For notational simplicity we have focused on inequality constraints. If some of the inequalities are equalities, then we would simply restrict the slackness parameters for the equalities to zero.

[14] See Dong, Hsieh, and Shum (2017) for additional applications of MPCC in general moment inequality models.

[15] MPCC solvers are a bit of a black box, but they may use some regularization to smooth out constraints. This regularization may affect statistical properties and/or computational performance of these estimates. A more detailed examination of these issues would be solver-specific and require knowledge of the algorithm designs of the solvers.

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 θ_j, one can report the confidence interval $[\underline{\theta}_j(1-\alpha), \overline{\theta}_j(1-\alpha)]$, which can be obtained by solving the following problems:

$$\begin{aligned} \underline{\theta}_j(1-\alpha) &= \inf \theta_j, \quad \text{st. } \theta \in \Theta : \min_{\lambda \geq 0, s \geq 0 : \lambda' s = 0} n\widehat{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1-\alpha); \\ \overline{\theta}_j(1-\alpha) &= \sup \theta_j, \quad \text{st. } \theta \in \Theta, \min_{\lambda \geq 0, s \geq 0 : \lambda' s = 0} n\widehat{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1-\alpha). \end{aligned} \tag{21}$$

Remark: testing linear constraints. Within this framework, it is straightforward to conduct tests of whether the optimal solution satisfies a set of linear restrictions, for example:

$$H_0 : C\theta_0 \geq \mathbf{r} \text{ for a } \theta_0 \in \arg\max_{\theta \in \Theta} c'\theta \text{ s.t. } A\theta \leq b, \tag{22}$$

where C is a known matrix and $\mathbf{r}$ is a known vector. Following the previous discussion, a primal-dual test for this is

$$\varphi_n^{\mathrm{PD}}(1-\alpha) = \mathbb{1}\left\{ \min_{\theta \in \Theta, \lambda \geq 0, s \geq 0 : \lambda' s = 0, C\theta \geq \mathbf{r}} n\widehat{Q}_n(\theta, \lambda, s) \leq \chi^2_{m+k}(1-\alpha) \right\}. \tag{23}$$

■

Remark. Our approach relies on the fact that the KKT conditions can be treated as moment conditions. As illustrated by the examples above, a number of LP problems arising in economic applications enjoy this property. However, there are some exceptions of LP problems arising in economics which lie outside our framework. One prominent example of this is the classic formulation of quantile regression as a solution to a linear program; see Koenker and Bassett (1978) for details. In their formulation, each of the constraints is attached to a single data observation, so that the number of constraints increases in the sample size. The KKT conditions for this LP do not take the form of moment conditions, and hence do not fit in our framework. ■

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, and c. For clarity, we now let the data distribution P index these quantities, that is, we now write A_P, b_P, and c_P. Then the solution set of the linear programming problem can be written as

$$\Theta_0(P) = \{\theta \in R^k : \exists \lambda \geq 0, s \geq 0, \lambda' s = 0 \text{ s.t. } g(A_P, b_P, c_P, \theta, \lambda, s) = 0\}.$$

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 $\mathcal{P}_0$ such that the following assumption holds.

Assumption 1. (a) *For all $P \in \mathcal{P}_0$, $\Theta_0(P)$ is nonempty.*

(b) *For any $P \in \mathcal{P}_0$ we have, under P,*

$$\sqrt{n}\begin{pmatrix} \mathrm{vec}(\widehat{A}_n) - \mathrm{vec}(A_P) \\ \hat{b}_b - b_P \\ \hat{c}_n - c_P \end{pmatrix} \rightarrow_d N(0, V_P),$$

for some positive semi-definite matrix V_P, and $\sup_{P\in\mathcal{P}_0} \|V_P\| \leq C$ for some constant C.

(c) *For any sequence $\{P_n\}_{n\geq 1}$ such that $P_n \in \mathcal{P}_0$ for all n, and any subsequence $\{a_n\}$ of $\{n\}$ such that $V_{P_{a_n}} \rightarrow V$ for a finite matrix V, we have, under $\{P_{a_n}\}$,*

$$\sqrt{a_n}\begin{pmatrix} \mathrm{vec}(\widehat{A}_{a_n}) - \mathrm{vec}(A_{P_{a_n}}) \\ \hat{b}_{a_b} - b_{P_{a_n}} \\ \hat{c}_{a_n} - c_{P_{a_n}} \end{pmatrix} \rightarrow_d N(0, V),$$

and $\widehat{V}_{a_n} \rightarrow_p V$ as $n \rightarrow \infty$.

(d) *There exists constants $\varepsilon > 0$ and $C > 0$ such that for all $\theta \in \Theta_0(P)$ and $P \in \mathcal{P}_0$, there exists $\lambda \geq 0$, $s \geq 0$ such that $\lambda' s = 0$, $g(A_P, b_P, c_P, \theta, \lambda, s) = 0$, and the smallest eigenvalue of $G(\theta,\lambda)V_P G(\theta,\lambda)'$ is no smaller than ε, $\|\lambda\|$, and $\|\theta\| \leq C$.*

Remarks. (i) Parts (b)-(c) of the assumption are high-level assumptions that can be verified in the step where $\widehat{A}_n, \hat{b}_n$, and $\hat{c}_n$ are obtained. For example, suppose that A is a variance covariance matrix of a random vector X, and b and c are respectively the expectation of W and Z for random vectors W and Z, and we have an i.i.d. sample $(X_i', W_i', Z_i')'$. Then often we have,

$$\begin{aligned} \widehat{A}_n &= (n-1)^{-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)(X_i - \bar{X}_n)', \\ \hat{b}_n &= \bar{W}_n, \\ \hat{c}_n &= \bar{Z}_n, \end{aligned} \tag{24}$$

where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$, $\bar{W}_n = n^{-1}\sum_{i=1}^n W_i$ and $\bar{Z}_n = n^{-1}\sum_{i=1}^n Z_i$. In this case, $\mathcal{P}$ is the set of distributions for $(X_i', W_i', Z_i')'$ that we allow. Moreover,

$$V_P = E_P \left[\begin{pmatrix} vec(X_i X_i') - vec(E_P[X_i X_i']) \\ W_i - E_P W_i \\ Z_i - E_P Z_i \end{pmatrix} \begin{pmatrix} vec(X_i X_i') - vec(E_P[X_i X_i']) \\ W_i - E_P W_i \\ Z_i - E_P Z_i \end{pmatrix}' \right] \tag{25}$$

with finite-sample analog

$$\widehat{V}_n = n^{-1} \sum_{i=1}^{n} \left[\begin{pmatrix} vec\left(X_i X_i' - n^{-1} \sum_{i=1}^{n} X_i X_i'\right) \\ W_i - \bar{W}_n \\ Z_i - \bar{Z}_n \end{pmatrix} \begin{pmatrix} vec\left(X_i X_i' - n^{-1} \sum_{i=1}^{n} X_i X_i'\right) \\ W_i - \bar{W}_n \\ Z_i - \bar{Z}_n \end{pmatrix}' \right]. \tag{26}$$

Then part (b) holds by a central limit theorem as long as V_P is finite for all $P \in \mathcal{P}_0$, which in turn is implied by $E_P[\|X_i\|^4] < \infty$, $E\|W\|^2 < \infty$, and $E\|Z\|^2 < \infty$. Part (c) can be justified by appealing to CLT with the Lyapounov condition if we strengthen the moment condition to $\sup_{P \in \mathcal{P}_0} E_P\|X_i\|^{4+\delta} < \infty$, $\sup_{P \in \mathcal{P}_0} E_P\|W_i\|^{2+\delta} < \infty$, and $\sup_{P \in \mathcal{P}_0} E_P\|Z_i\|^{2+\delta} < \infty$ for some $\delta > 0$. This strengthened moment condition also ensures the uniform boundedness condition for V_P in part (b).

In addition, in this case $G(\theta, \lambda)$ has full row-rank, implying that a sufficient condition for the eigenvalues of $G(\theta, \lambda) V_P G(\theta, \lambda)'$ to be bounded away from zero is that the eigenvalues of V_P are bounded away from zero. (This is not a necessary condition as the number of rows in $G(\theta, \lambda)$ is far less than the dimension of V_P.)

(ii) Part (d) also imposes a uniform bound on the Lagrange Multiplier λ. When multiple Lagrange multiplier values satisfy the KKT conditions, this sufficient condition only requires that some of them be subject to the bound. Thus, it does not rule out writing an inequality as two equalities. On the other hand, part (d) does imply a lower bound on the norm of each row of A. This can be restrictive sometimes. See e.g. Bonnans and Shapiro (2000) and Gafarov (2016).

(iii) The assumptions do not rule out the case where the linear programming solution is non-unique. We are able to obtain uniform asymptotic coverage in this case 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 next theorem is a statement of uniform asymptotic validity.

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

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_0} \inf_{\theta\in\Theta_0(P)} \Pr{}_P(\theta \in CS_n^{\mathrm{PD}}(1-\alpha)) \geq 1-\alpha,$$

where $\Pr_P$ stands for probability under the data distribution P.

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$\}$ is nonempty. Denote the singleton by $\theta_0(P)$. Then we have for $\alpha \in (0,1)$,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}_{00}} \Pr{}_P(\theta_0(P) \in CS_n^{\mathrm{PD}}(1-\alpha)) \geq 1-\alpha.$$

On the surface, the corollary is a direct result of Theorem 1, but it gives the coverage result a desirable implication: the confidence set covers the unique solution with pre-specified 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_∞ with multiple solutions, i.e. a P_∞ 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.

4.2 Concentrating Out Nuisance Parameters

Our confidence set CS_n^{PD} uses the chi-squared critical value and thus is very easy to compute. This computational ease comes at a cost, however: the construction of the confidence set required us to introduce two sets of nuisance parameters: the Lagrange multipliers λ and slackness terms s. The confidence set $CS_n^{\mathrm{PD}}(1-\alpha)$ is a projection of a $(1-\alpha)$-confidence set for $(\theta', \lambda', s')'$ and, as such, it may over-cover, as pointed out recently in the literature; see e.g. Bugni, Canay, and Shi (2017) (BCS), Kaido et al. (2019) (KMS), and Chen, Christensen, and Tamer (2018) (CCT).[16]

[16] Alternatively, a large class of methods in the moment inequality literature, for example, Andrews and Soares (2010), Andrews and Barwick (2012), Romano and Shaikh (2010) can be applied directly to the optimality conditions (2)-(5) without introducing the slackness parameter s. The computational cost of these methods is similar to or exceeds that of computing $CS_n^{\text{PD-prof}}$, and are not guaranteed to produce confidence sets that are subsets of CS_n^{PD}.

Here we describe a simple method to profile out the nuisance parameter s, which is a direct application of Section 4.1 of Shi and Shum (2015). This method still leaves the nuisance parameter λ but it yields a valid confidence set that is always a subset (and often a strict subset) of $CS^{\text{PD}}(1-\alpha)$, at a moderate increase in computational cost. In the following section we also discuss the prospect of applying other subvector inference methods in our context.

The profiling method in Shi and Shum (2015) simulates the quantiles of the following statistic:

$$J_n(\theta,\lambda) = \min_{t+\hat{s}(\theta,\lambda)\geq 0, \lambda' t=0} \left\| G(\theta,\lambda,s)\widehat{V}^{1/2}Z + \kappa_n^{-1}n^{1/2}\left(\begin{smallmatrix} t \\ 0_{k\times 1}\end{smallmatrix}\right)' \right\|^2_{(G(\theta,\lambda,s)\widehat{V}G(\theta,\lambda,s)')^{-1}}, \quad (27)$$

where $\|b\|_W^2 = b'Wb$, and $\hat{s}(\theta,\lambda) = \arg\min_{s\geq 0:\lambda' s=0} \|s-(\widehat{A}\theta - \hat{b})\|$ and $Z \sim N(0_{m+k}, I_{m+k})$, and κ_n is a tuning parameter that satisfies $\kappa_n \to \infty$ and $\kappa_n/n^{-1/2} \to 0$. Let $cv_n(\theta,\lambda,1-\alpha)$ denote the simulated $100(1-\alpha)\%$ quantile of $J_n(\theta,\lambda)$. Then the profiled confidence set for θ is defined as

$$CS_n^{\text{PD-prof}}(1-\alpha) = \left\{\theta\in\Theta : \min_{\lambda\geq 0}\left[\min_{s\geq 0: s'\lambda=0} n\widehat{Q}_n(\theta,\lambda,s) - cv_n(\theta,\lambda,1-\alpha)\right] \leq 0\right\}. \quad (28)$$

In practice, one may report only the upper and lower bound of the confidence set of each parameter, say $[\underline{\theta}_j^{\text{prof}}(1-\alpha), \overline{\theta}_j^{\text{prof}}(1-\alpha)]$, which can be derived from the following problems:

$$\begin{aligned} \underline{\theta}_j^{\text{prof}}(1-\alpha) &= \inf \theta_j, \text{ s.t. } \theta\in\Theta : \lambda\geq 0, \min_{s\geq 0: s'\lambda=0} n\widehat{Q}_n(\theta,\lambda,s) \leq cv_n(\theta,\lambda,1-\alpha). \\ \overline{\theta}_j^{\text{prof}}(1-\alpha) &= \sup \theta_j, \text{ s.t. } \theta\in\Theta : \lambda\geq 0, \min_{s\geq 0: s'\lambda=0} n\widehat{Q}_n(\theta,\lambda,s) \leq cv_n(\theta,\lambda,1-\alpha). \end{aligned} \quad (29)$$

The only difference between $CS_n^{\text{PD-prof}}$ and CS_n^{PD} is that the latter utilizes a simulated critical value $cv_n(\theta,\lambda,1-\alpha)$ instead of the analytic chi-squared critical value. By definition, $cv_n(\theta,\lambda,1-\alpha)$ is weakly smaller.[17] Thus, $CS_n^{\text{PD-prof}}(1-\alpha) \subseteq CS_n^{\text{PD}}(1-\alpha)$, and $\underline{\theta}_j(1-\alpha) \leq \underline{\theta}_j^{\text{prof}}(1-\alpha) \leq \overline{\theta}_j^{\text{prof}}(1-\alpha) \leq \overline{\theta}_j(1-\alpha)$. The tighter critical values $cv_n(\theta,\lambda,1-\alpha)$ involve additional computational cost, and the degree of improvement that we can obtain is application specific.

4.3 Other moment inequality procedures for subvector inference

We have focused on the specialized inference procedures CS_n^{PD} and $CS_n^{\text{PD-prof}}$ in Shi and Shum (2015) here. The former is computationally very easy and the latter yields uniform

[17] To see why, note that $t = \mathbf{0}$ is feasible in the minimization problem in (27), and setting $t = 0$ makes the squared norm a χ^2_{m+k} variable. Thus, $J_n(\theta,\lambda)$ is stochastically dominated by χ^2_{m+k}, implying that $cv_n(\theta,\lambda,1-\alpha)$ weakly smaller than $\chi^2_{m+k}(1-\alpha)$.

power improvement over the former at moderately greater computational cost. Other subvector inference methods, such as BCS, KMS, and CCT have been proposed for general moment inequality models. These procedures can apply in the present context.[18] Nevertheless, they may present computational and/or statistical advantages on a case by case basis. For this reason, we briefly review the approaches of BCS, KMS, and CCT and make comparison to the extent that is possible within the scope of this paper.

Before doing so, we note several caveats. First, the comparisons will be somewhat heuristic, as we will attempt neither a theoretical nor computational comparison of coverage rates, as such exercises can be technically involved and beyond the scope of this paper. Every procedure has its pluses and minuses and the discussion here is in no way a slam-dunk comparison. Second, the alternative subvector inference approaches discussed below are applicable to the more general class of *moment inequality* models, whereas $CS_n^{\text{PD-prof}}$ has more limited applicability.

For concreteness, in the following discussion we assume that the parameter of interest is θ_1, the first element of θ. We consider, in turn, the approaches in BCS, KMS, and CCT. The focus in the comparison is in terms of computational cost in our setting. In each case, only enough details are given to facilitate an informative comparison of the computational cost.

4.3.1 Bugni, Canay and Shi (2017)

To construct a confidence interval for θ_1 using the BCS approach applied to the formulation in (2)-(5), one first constructs a profiled criterion function as the test statistic. One such test statistic is[19]

$$T_n(\theta_1) = \min_{\theta_{-1},\lambda\geq 0} n \left\| \begin{pmatrix} D_n^{-1/2}(\theta,\lambda) \begin{pmatrix} \max\{0, \hat{A}\theta - b\} \\ \hat{A}'\lambda - \hat{c} \end{pmatrix} \\ \sigma_{3,n}^{-1}(\theta,\lambda)(\hat{c}'\theta - \hat{b}'\lambda) \end{pmatrix} \right\|^2, \tag{30}$$

where $D_n(\theta,\lambda)$ is a diagonal matrix with the same diagonal elements as $G(\theta,\lambda)\widehat{V}G(\theta,\lambda)'$, and $\sigma_{3,n}^2(\theta,\lambda) = (\,\mathbf{0}_{1\times mk}\ -\lambda'\ \theta'\,)\,\widehat{V}\begin{pmatrix} \mathbf{0}_{mk\times 1} \\ -\lambda \\ \theta \end{pmatrix}$, and θ_{-1} denotes the parameter vector θ without

[18]The recent papers by Andrews, Roth, and Pakes (2019b) and Cox and Shi (2020) propose subvector inference methods for linear conditional moment inequality models. Their methods do not apply here because they require the coefficients of the nuisance parameters to either be deterministic or depend only on the conditioning variables. Unlike $CS_n^{\text{PD-prof}}$, they are not guaranteed to yield uniform power improvement upon CS_n^{PD} due to their greater difference with CS_n^{PD}.

[19]This correspond to the MMM (modified method of moment) test statistic in BCS.

the first element. Then the upper and lower bounds of the confidence interval for θ_1 are obtained by

$$\begin{aligned}\underline{\theta}_1^{BCS}(1-\alpha) &= \min \theta_1 \text{ s.t. } T_n(\theta_1) \leq cv^{BCS}(\theta_1, 1-\alpha)\\ \overline{\theta}_1^{BCS}(1-\alpha) &= \max \theta_1 \text{ s.t. } T_n(\theta_1) \leq cv^{BCS}(\theta_1, 1-\alpha)\end{aligned} \tag{31}$$

where the critical value $cv^{BCS}(\theta_1, 1-\alpha)$ is the smaller of the quantiles of two bootstrap versions of $T_n(\theta_1)$. As we can see, the computation has an inner loop where $cv^{BCS}(\theta_1, 1-\alpha)$ is computed and an outer loop, where one searches over the space of θ_1 for the boundary of the set defined by $T_n(\theta_1) \leq cv^{BCS}(\theta_1, 1-\alpha)$. The outerloop per se is not difficult as it is a one-dimensional search. However, the inner loop can be difficult because it requires solving $2B+1$ minimization problems, each of the same magnitude of difficulty as (30), where B is the number of bootstrap repetitions.[20]

In comparison, the $CS_n^{\text{PD-prof}}$ procedure (cf. Eq. (29)) also has an inner loop where the critical values $cv_n(0, \lambda, 1-\alpha)$ and the test statistic $\min_{s\geq 0: s'\lambda=0} n\hat{Q}_n(\theta, \lambda, s)$ are obtained as well as an outerloop where one search over $(\theta', \lambda')'$ for the extreme values of θ_1 (recall that $j = 1$ for the disucssion in this subsection). The inner loop is relatively easy because both the critical value and the test statistic only involve quadratic programming problems, which are computationally simpler to solve. Yet, the outer loop is costlier compared to the outer loop in BCS since it requires searching over a higher dimensional space.

4.3.2 Kaido, Molinari, Stoye (2019)

KMS can also be applied on the formulation in (2)-(5). The upper and lower bounds of the KMS confidence interval for θ_1 are given by

$$\begin{aligned}\underline{\theta}_1^{KMS}(1-\alpha) &= \min \theta_1 \text{ s.t. } \max\left(\left|\begin{pmatrix} D_n^{-1/2}(\theta,\lambda)\begin{pmatrix}\max\{0, \hat{A}\theta - b\}\\ \hat{A}'\lambda - \hat{c}\end{pmatrix}\\ \sigma_{3,n}^{-1}(\theta,\lambda)(\hat{c}'\theta - \hat{b}'\lambda)\end{pmatrix}\right|\right) \leq cv_n^{KMS}(\theta, \lambda, 1-\alpha)\\ \overline{\theta}_1^{KMS}(1-\alpha) &= \max \theta_1 \text{ s.t. } \max\left(\left|\begin{pmatrix} D_n^{-1/2}(\theta,\lambda)\begin{pmatrix}\max\{0, \hat{A}\theta - b\}\\ \hat{A}'\lambda - \hat{c}\end{pmatrix}\\ \sigma_{3,n}^{-1}(\theta,\lambda)(\hat{c}'\theta - \hat{b}'\lambda)\end{pmatrix}\right|\right) \leq cv_n^{KMS}(\theta, \lambda, 1-\alpha),\end{aligned} \tag{32}$$

where $cv_n^{KMS}(\theta, \lambda, 1-\alpha)$ is a critical value obtained from a bootstrap procedure where in each bootstrap repetition, one solves two linear programming problems with $k+1+$

[20] One saving grace is that the $2B$ minimization problems for the bootstrap repetitions do not need to done perfectly for the resulting confidence set to be valid – not exactly finding the global minima simply yields a bigger, and hence still valid, critical value.

m constraints. The KMS procedure also involves an inner loop and an outer loop. In contrast to BCS, the KMS procedure involves, essentially, an easier inner loop and a more complicated outer loop. The inner loop, where $cv_n^{KMS}(\theta, \lambda, 1-\alpha)$ is computed, is relatively easy because there are only $2B$ linear programming problems to solve if B is the number of bootstrap repetitions. Yet the outer loop is difficult because it requires searching over the entire space of $(\theta', \lambda')'$ for an extreme point of θ_1 subject to $k + m + 1$ constraints, and the constraints involve the simulated function $cv_n^{KMS}(\theta, \lambda, 1-\alpha)$. KMS propose an approximation algorithm based on the response surface method to handle the outer loop problem more efficiently.

In comparison, the $CS_n^{\text{PD-prof}}$ is likely of comparable difficulty as KMS since the inner loop of $CS_n^{\text{PD-prof}}$ involves B quadratic programming problems with fewer variables (m vs. $m+k$) and fewer constraints ($m+1$ vs. $m+k+1$) than KMS's linear programming problems, and our outer loop is also a search over the space of $(\theta', \lambda')'$ and – indeed – KMS's response surface algorithm for handling the outer loop could be employed here as well.

4.3.3 Chen, Christensen, Tamer (2018)

CCT contains two procedures that can be applied to the present context to construct confidence intervals for the identified set of the scalar parameter θ_1. Both procedures are designed for moment equality models, and thus they can only be directly applied to the formulation (13)-(16) which involves the slackness parameters s.

CCT's Procedure 2 involves taking B MCMC (Markov chain Monte Carlo) draws of $(\theta', \lambda', s')'$ from a quasi-posterior distribution. At each point, say $(\theta^{b,\prime}, \lambda^{b,\prime}, s^{b,\prime})'$, we compute a "test statistic"

$$PL(M(\theta^b, \lambda^b, s^b)) = \sup_{\theta_1 \in M(\theta^b, \lambda^b, s^b)} \inf_{\theta_{-1} \in \Theta_{-1}(\theta_1), \lambda \geq 0, s \geq 0 : s'\lambda = 0} n\widehat{Q}_n(\theta, \lambda, s), \tag{33}$$

where $\Theta_{-1}(\theta_1) = \{\theta_{-1} :\ \text{s.t. } (\theta_1, \theta_{-1}')' \in \Theta\}$ and $M(\theta^b, \lambda^b, s^b)$ is a set of θ_1 that is observationally equivalent to $(\theta^b, \lambda^b, s^b)$ when combined with some value of $(\theta_{-1}, \lambda, s)$.[21] The critical value $cv_n^{CCT}(1-\alpha)$ is then obtained by taking the $(1-\alpha)^{th}$ quantile from the sample $\{PL(M(\theta^b, \lambda^b, s^b)) : b = 1, \ldots, B\}$. Then the upper and lower bounds of the $100(1-\alpha)\%$

[21] Obtaining $M(\theta^b, \lambda^b, s^b)$ in non-separable models as we have here can be quite challenging in practice.

CCT Procedure 2 confidence interval for θ_1 are

$$
\begin{aligned}
\underline{\theta}_1^{\text{CCT2}}(1-\alpha) &= \inf_{\theta_1 \in \Theta_1} \theta_1 \text{ s.t. } \inf_{\theta_{-1} \in \Theta_{-1}(\theta_1), \lambda \geq 0, s \geq 0 : s'\lambda = 0} n\widehat{Q}_n(\theta, \lambda, s) \leq cv_n^{CCT2}(1-\alpha) \\
\overline{\theta}_1^{\text{CCT2}}(1-\alpha) &= \sup_{\theta_1 \in \Theta_1} \theta_1 \text{ s.t. } \inf_{\theta_{-1} \in \Theta_{-1}(\theta_1), \lambda \geq 0, s \geq 0 : s'\lambda = 0} n\widehat{Q}_n(\theta, \lambda, s) \leq cv_n^{CCT2}(1-\alpha), \quad (34)
\end{aligned}
$$

where $\Theta_1 = \{\theta_1 \in R : \Theta_{-1}(\theta_1) \neq \emptyset\}$. CCT's procedure also involves an inner loop where the $cv_n^{\text{CCT2}}(1-\alpha)$ is obtained and an outer loop where we search over the space of θ_1 for an extreme point. In fact, it resembles BCS in terms of these two layers of loops: the outer loop is a one-dimensional search (thus easy) and the inner loop involves minimizing a GMM criterion function over $(\theta_{-1}, \lambda, s)$ B times. The difference is that B here is the number of MCMC draws, while for BCS, it is the number of bootstrap repetitions. As with BCS, the computational cost comparison of the $CS_n^{\text{PD-prof}}$ with CCT Procedure 2 may vary case by case and depend on implementation.

CCT's Procedure 3 first defines $(\hat{\theta}', \hat{\lambda}', \hat{s}')'$ to be a solution to the minimization problem $\min_{\theta \in \Theta, \lambda \geq 0, s \geq 0, \lambda' s = 0} \widehat{Q}_n(\theta, \lambda, s)$. Then let

$$
T_n^{\text{CCT3}}(\theta_1) = \inf_{\theta_1 \in \Theta_{-1}(\theta_1), \lambda \geq 0, s \geq 0, \lambda' s = 0} n\widehat{Q}_n(\theta, \lambda, s) - n\widehat{Q}_n(\hat{\theta}, \hat{\lambda}, \hat{s}). \quad (35)
$$

And the upper and lower bounds of the $100(1-\alpha)\%$ CCT Procedure 3 confidence interval for θ_1 are

$$
\begin{aligned}
\underline{\theta}_1^{\text{CCT3}}(1-\alpha) &= \inf_{\theta_1 \in \Theta_1} \theta_1 \text{ s.t. } T_n^{CCT3}(\theta_1) \leq \chi_1^2(1-\alpha) \\
\overline{\theta}_1^{\text{CCT3}}(1-\alpha) &= \sup_{\theta_1 \in \Theta_1} \theta_1 \text{ s.t. } T_n^{CCT3}(\theta_1) \leq \chi_1^2(1-\alpha) \quad (36)
\end{aligned}
$$

In terms of computational cost, CCT's Procedure 3 is only slightly more costly than CS_n^{PD} in that it requires the calculation of the overall minimum of $\widehat{Q}_n(\theta, \lambda, s)$. Both are much easier than $CS_n^{\text{PD-prof}}$, as well as BCS and KMS. Unlike the other procedures, however, CCT's Procedure 3 cannot be applied when the parameter of interest is not a scalar.

Remark: In a different vein, Kline and Tamer (2016) consider Bayesian inference in a class of partially identified models which delivers subvector inference automatically. Specifically, applied to our setting, their procedure involves sampling from the posterior distribution of point-identified "reduced-form" parameters (corresponding to the A, B, c). A Bayesian credible set for the subvector can be attained by computing the full parameter vector θ for each draw of (A, B, c) and only retaining the subvector of interest.

5 Application: Portfolio Selection

As an empirical illustration, we consider the classic portfolio allocation problem in finance. While this is a long-running problem in finance, inference procedures for the optimizing solutions of the problem (the optimal portfolio weights) have not been established. In practice, bootstrap-like procedures are used to assess sampling error; see, e.g., Scherer (2002). This naive approach is not valid in light of the recent results in Fang and Santos (2019). As far as we are aware, then, our procedure here constitutes the first asymptotically valid inference procedure for this model.

5.1 Models

We consider three portfolio selection models: (i) the MV portfolio without short-selling of Markowitz (1952) (MV), which is a quadratic program (QP); (ii) the (Euclidean) norm-constrained portfolio of DeMiguel et al. (2009) (NC), which is a convex quadratically constrained quadratic program (QCQP); and (iii) the equally-weighted risk contribution portfolio of Maillard et al. (2010) (ERC), which is a convex program (CP). Each model is described in turn:

1. *Mean Variance Portfolio:*
   Markowitz (1952) consider a problem of forming an optimal portfolio among k assets with weights $\{\theta_1, \ldots, \theta_k\} \equiv \theta$ that solves a convex QP defined in (10). The MV problem implies two primal feasibility conditions

$$\begin{aligned} R'\theta - \mu &= 0 \\ \mathbf{1}'\theta - 1 &= 0 \end{aligned} \tag{37}$$

   and k dual feasibility conditions:

$$\boldsymbol{\lambda}_\theta + \lambda_R R + \lambda_F \mathbf{1} - Q\theta = \mathbf{0}. \tag{38}$$

   In the above, $\boldsymbol{\lambda}_\theta$ is the vector of Lagrange multipliers of the non-negativity constraints on θ (corresponding to the restriction to long positions), and λ_R, λ_F are the Lagrange multipliers of the equality constraints of targeted return and feasible portfolio weights. There are k linear complementarity conditions: $0 \leq \boldsymbol{\lambda}_\theta \perp \theta \geq 0$. In this case, our confidence set is:

$$CS_n^{\mathrm{PD}}(1-\alpha) = \{\theta \in \Theta : \min_{\mathbf{1}'\theta - 1 = 0, 0 \leq \boldsymbol{\lambda}_\theta \perp \theta \geq 0} n\widehat{Q}_n(\theta, \lambda) \leq \chi^2_{1+k}(1-\alpha)\}. \tag{39}$$

2. *NC Portfolio:*

Besides the classic Markowitz problem above, we also consider two more recent versions of the portfolio allocation problem, which attempt to address problematic features of the Markowitz solution. DeMiguel et al. (2009) consider imposing an extra regularization restriction on the portfolio weights. Specifically, they consider the following convex QCQP problem:

$$\begin{aligned} \min \quad & \theta' Q \theta \\ \text{s.t.} \quad & \theta'\theta \leq \delta \\ & \mathbf{1}'\theta = 1. \end{aligned} \tag{40}$$

The first constraint, which is new, represents a regularization of the optimal portfolio weights away from putting full weights on any single asset. There are two primal feasibility conditions

$$\theta^{'}\theta - \delta + s \quad = 0 \tag{41}$$

$$\mathbf{1}'\theta - 1 \quad = 0, \tag{42}$$

k dual feasibility conditions

$$Q\theta + \lambda_c \theta + \lambda_F \mathbf{1} = 0, \tag{43}$$

and one complementarity constraint: $0 \leq \lambda_c \perp s \geq 0$. The primal feasibility conditions do not involve data, and hence will be treated as parameter constraints when computing the primal-dual test statistics.

3. *Equally-Weighted Risk Contribution Portfolio:*

In a similar vein, Maillard et al. (2010) consider the portfolio allocation that solves the following convex programming problem

$$\begin{aligned} \min \quad & \theta' Q \theta \\ \text{s.t.} \quad & \textstyle\sum_{i=1}^{k} \log \theta_i \geq \eta \\ & \mathbf{1}'\theta = 1 \\ & \theta \geq 0. \end{aligned} \tag{44}$$

The first constraint is new, and essentially "shrinks" the optimal portfolio towards equally-weighted portfolio (where $\theta_i = \frac{1}{k}$ for all i). There are two primal feasibility conditions

$$\begin{aligned} & \textstyle\sum_{i=1}^{k} \log \theta_i - \eta - s && = 0 \\ & \mathbf{1}'\theta - 1 && = 0, \end{aligned} \tag{45}$$

k dual feasibility conditions

$$Q\theta - \lambda_c(1/\theta) - \lambda_F \mathbf{1} - \boldsymbol{\lambda}_\theta = 0, \tag{46}$$

and $k+1$ complementarity inequalities

$$0 \leq \boldsymbol{\lambda}_\theta \perp \theta \geq 0 \tag{47}$$

$$0 \leq \lambda_c \perp s \geq 0. \tag{48}$$

5.2 Empirical Results

We consider portfolio selection over three fixed income securities: AAA, AA, and BBB corporate bonds. We use the daily effective yield data from January 4, 2010 to December 31, 2017.[22] The sample return vector $\hat{R}$ and sample covariance matrix $\hat{Q}$ are:

$$\hat{R} = \begin{pmatrix} \text{AAA} \\ \text{AA} \\ \text{BBB} \end{pmatrix} = \begin{pmatrix} 2.5621 \\ 2.6405 \\ 3.9492 \end{pmatrix}; \hat{Q} = \begin{pmatrix} 0.1888 & \cdot & \cdot \\ 0.1184 & 0.1605 & \cdot \\ 0.0967 & 0.1848 & 0.2615 \end{pmatrix}. \tag{49}$$

Estimated optimal portfolio weights. To ensure an equal footing, we set the tuning parameters such that the resulting solutions $\hat{\theta}$ to the mathematical programming problems (10), (40), and (44) with estimated $\hat{R}$ and $\hat{Q}$ given above are roughly the same. Specifically, we choose $\mu = 2.8$, $\delta = 0.1 + \frac{1}{3} \approx 0.4333$, and $\eta = -3\log(3) - 1 \approx -4.2958$ for the MV, NC, and ERC portfolio, respectively.[23] The estimated θ are reported in Table 1. While numerically they are quite similar, we note that an important difference is that the estimated weights for the MV model are in the interior of the feasible solution set (namely, the non-negativity constraints are not binding). On the contrary, the estimated weights for both the NC and ERC models are on the boundary. For the NC model, the estimated $\hat{\theta}$ in Table 1 satisfies the constraint $\hat{\theta}'\hat{\theta} = \delta = 0.4333$. For the ERC model, we have $\sum_{i=1}^{3} \hat{\theta}_i = \eta = -4.3$. This will have important consequences for the confidence sets.

Confidence sets. In Figure 1, we depict the confidence set for the three different portfolio selection models. Since the confidence set is three-dimensional, we present the two-

[22] We download the data of BofA Merrill Lynch US Corporate AAA, AA, and BBB Effective Yield from Federal Reserve Bank of St. Louis. Source: https://fred.stlouisfed.org/.

[23] At these parameter values, it is easy to verify that Slater's condition is satisfied because $\theta_i = \frac{1}{3}$, for all i, belongs to the relative interior of the convex constraint set of both NC and ERC. Therefore, the KKT conditions derived before are necessary and sufficient for the global optimality.

Table 1: Estimated Portfolio Weights

	MV	NC	ERC
AAA	0.4279	0.4207	0.4035
AA	0.4247	0.5001	0.5332
BBB	0.1474	0.0792	0.0633

dimensional projections for each pair of assets separately.[24] Our method yields tight confidence sets for the estimated portfolio weights. For example, since the AAA and AA corporate bonds have similar estimated risk-return profiles, the confidence sets for all three models suggest that the AAA and AA corporate bonds are substitutable: starting from the point estimate, the confidence set includes points which involve a higher share of AAA compensated by a lower share of AA, and vice versa. On the other hand, the risk-return profile of the BBB corporate bonds is distinct from that of AAA and AA corporate bonds. As a result, the confidence sets are much tighter along the dimension of the BBB corporate bonds.

Despite the similarity of the optimal portfolio weights for all three (MV, NC, ERC) models, as reported in Table 1, the confidence sets in Fig. 1 exhibit striking differences across models. Particularly, the confidence set for the MV model exhibits a typical "elliptical" shape, while the confidence sets for the NC and ERC model weights have a non-convex "arc" shape. These differences arise from the location of the estimated portfolio weights $\hat{\theta}$ within the set of the feasible solutions; namely, whether $\hat{\theta}$ lies on the interior or boundary on the feasible solution set. The structure of the complementarity conditions in mathematical programming problems implies that small movements in the value of θ around the boundary can lead to discontinuous "jumps" in the values of the Lagrange multipliers – from zero (on the boundary) to non-zero (off the boundary) – and subsequently also to discontinuities in the test statistics. This feature is special to the mathematical programming problem under study and do not arise in typical moment inequality models.

In the case of MV portfolio, $\hat{\theta}$ lies on the relative interior of the constraint set, and the associated Lagrange multipliers $\lambda_\theta = 0$. The confidence set consists of other feasible points θ for which the test statistic is close in value to the test statistic at $\hat{\theta}$ – which will be at other points in the interior of the constraint set, leading to the ellipsoid shape.

[24] We first generate Sobol sequences from the simplex and then plot points that satisfy Eq. (20)

On the other hand, both the NC and ERC portfolio reported in Table 1 lie on the relative boundary of the constraint set. This implies that the corresponding Lagrange multipliers (λ_c) for these boundary constraints will be non-zero, implying that the test statistic will be small for other values of θ which likewise lie on the boundary. On the contrary, feasible points on the boundary will have values of $\lambda_c = 0$, leading to large changes in the test statistic, compared to the value of the test statistic at $\hat{\theta}$. Thus the resulting confidence set is thin and arc-shaped—essentially a lower-dimensional manifold tracing out a portion of the boundary of the feasible set.

In practice, investors may not need to do inference on the estimated portfolio weights $\hat{\theta}$ themselves, but rather the implied expected return $R'\hat{\theta}$ and whether this expected return exceeds some threshold value τ (which may depend on transactions costs, the status quo return, &c.). Such considerations can be written as linear constraints on the portfolio weights $R'\theta \geq \tau$, the testing of which we discussed earlier (cf. Eqs. (22), (23)).

5.3 Monte Carlo Experiments

From a practical point of view, portfolio managers may be less interested in doing inference on portfolio weights, per se, than in the "dual" problem; namely, they may want to know "how big" the change in θ would need to be in order for the change to be statistically detectable: which is a question about the *power* of the test underlying our proposed confidence set C^{PD}. The reasonably tight confidence sets obtained in the empirical application above suggest reasonably good power of our test. In the remainder of this section, we address this power question directly via a set of Monte Carlo experiments.

Design 1: Normally-Distributed Assets

The first simulation design is motivated by the daily return data used in Section 5.2. Specifically, we simulate asset return r_{it}, $i = \mathrm{AAA}, \mathrm{AA}, \mathrm{BBB}$, $t = 1, \ldots, T$, from a multivariate normal distribution with a mean vector R and a covariance matrix Q. We consider three sample sizes: $T = 100, 200, 500$. We test the null hypothesis that $H_0 : \theta_0 = \theta$ where θ_0 denotes the optimal portfolio weights, and θ are fixed values for the weights, which in this design are set equal to the estimated values given in Table 1 for the MV, NC, and ERC models, respectively. We assume that (R, Q) equals $(\hat{R}, \hat{Q})$ in Eq. (49) to generate the data under H_0.

To generate the data under the alternative hypothesis that H_0 does not hold, we shift Q as follows: In the first scenario, we shift the variance of the AAA corporate bonds away from the null data generating process ($\sigma^2_{\text{AAA}} = 0.1888$), and in the second scenario, we shift the variance of the AA corporate bonds from away from the null data generating process ($\sigma^2_{\text{AA}} = 0.1605$), holding other elements of Q and all elements of R fixed. The empirical rejection rates are computed over 500 replications, and the size α is set to 5%.

Notably, since θ_0 is a function of the parameters (R, Q) that generate the data, one cannot directly calculate the power curve over a set of predetermined mesh points of θ_0. Instead, on the left column of Figure 2 and 3, we report the empirical power curve as a function of the predetermined mesh points of σ^2_{AA} and σ^2_{AAA} that generate the return data under the alternative hypotheses.[25] On the right column, we further convert those power curves to functions of the Euclidean distance between the true θ_0 and the fixed values given in Table 1.[26] From the figures, one can find that when the sample size increases, the test power also increases. Interestingly, our test has more power when the volatility decreases.[27] We find that our test has reasonable finite sample performance across three different portfolio selection models. On the other hand, these simulations also show that our test is conservative under the null—a common problem in the subvector inference of moment inequality models.

Design 2: t-Distributed Assets

Asset return data often exhibit heavy tails. In design 2, we simulate r_{it} from a multivariate-t distribution with ten degrees of freedom. We further apply a location-scale transformation such that the resulting (population) mean and covariance matrix equal that of design 1. We summarize the results in Figure 4 and 5. We find that the test power reduces if the raw data is generated from a t-distribution. For example, in the case of the MV portfolio, the test power under $\sigma^2_{AAA} = 0.25$ and $T = 500$ is 75% if r_{it} are normally distributed (Figure 2, top left). By contrast, the test power reduces to 62% (Figure 4, top left) if r_{it} are $t(10)$ distributed.[28] This may have arisen because heavy-tailed data lead to noisier estimates of the sample covariance matrix $\hat{Q}$.

Design 3: Large-Scale Cases

[25] As some of the parameter values will result in a non-positive semidefinite covariance matrix, the left column of Figure 2 and 3 are plotted on a different grid points.

[26] Because of the asymmetric response to the positive or negative changes in volatility, we only plot the power curves associated with the positive changes.

[27] Importantly, a large shift in volatility does not necessarily imply a large shift in the portfolio weights. For example, we have found that changes in σ^2_{BBB} only lead to a negligible change in θ.

[28] If the returns are $t(5)$ distributed, the test power reduces to 30% under the same parameter setup.

Estimating the covariance matrix Q of asset return involves a large number of parameters; there are $\frac{N(N+1)}{2}$ parameters if there are N assets. Therefore, even for a moderate number of assets, the simple sample covariance matrix can perform poorly in practice; in turn, such a noisy estimate of Q will impair the reliability of the estimated portfolio weights from the portfolio allocation problem.

As a result, when many assets are considered, Jagannathan and Ma (2009) suggest estimating the covariance matrix of the assets' returns using factor models. Specifically, they consider the following one-factor model:

$$r_{it} = \alpha_i + \beta_i r_{mt} + \epsilon_{it}, \tag{50}$$

where r_{it} is the period t return of asset i, r_{mt} (common factor) is the period t return on the value-weighted portfolio of stocks traded in the market, ϵ_{it} is the idiosyncratic shock, and β_i is the factor loading. The (population) covariance matrix of $r_{it}, i = 1, \ldots, N$, is given by

$$Q = \sigma_m^2 \beta\beta' + D, \tag{51}$$

where σ_m^2 is the variance of r_{mt}, β is the column vector of factor loading β_i, and D is the diagonal matrix with variance of ϵ_{it} along its diagonal. β_i provides a convenient way to model the risk-return trade-off: the higher the β, the higher the return and also the risk. Clearly, β_i can be estimated by regressing r_{it} on r_{mt}. D_{ii}, the i-th diagonal component of D, can be estimated by the variance of residuals $\frac{1}{T-2}\sum_{t=1}^{T} \hat{\epsilon}_{it}^2$.

We consider 50 assets. Under the null hypothesis, the common factor r_{mt} follows $N(2, 1)$. The idiosyncratic shocks ϵ_{it} are i.i.d. drawn from $N(0, 0.5^2)$; therefore, D is a block-diagonal matrix. We generate 50 equally-spaced $\beta_i \in [0.5, 1.5]$, and we set the constant term $\alpha_i = 0$ for all i. We set the following tuning parameters: $\mu = 1.7$, $\delta = .006 + \frac{1}{50} \approx 0.026$, and $\eta = -50 * \log(50) - 3 \approx -198.6$ for the MV, NC, and ERC portfolio, respectively. These parameters produce positive weights for all assets that are roughly the same across three models under the null hypothesis. Our testing procedure is completely modular, and can accommodate this specification for the data-generating process of returns.

To generate the data under the alternative hypothesis, we add a constant $\tau \in [-0.2, 0.2]$ to β_i of the first 25 assets. One can interpret this design as if there are two sectors. The market condition affects one of them by shifting the factor loading. We consider two sample sizes: $T = 200, 500$. On the left column of Figure 6, we depict the empirical power as the function of the predetermined mesh points of τ, and the on the right column, we convert them as the functions of the Euclidean distance between the true optimal weight θ_0 implied by differ values of τ and the θ corresponding to $\tau = 0$. Our test works reasonably well under the high dimensional case. For the NC and the ERC portfolio, our test has power even for a marginal shift of the factor loading. The MV portfolio, on the other hand, would require more samples under this particular design.

Finally, in Table 2, we report the average runtime of computing the test statistic over 100 Monte Carlo repetitions.[29] The NC and the ERC portfolio scale up particular well; they both take less than 1 sec to solve. Since the NC portfolio contains only one complementarity constraint regardless of the number of assets, we speculate that the computing time may not increase even for larger problems. Computing the test statistics for the MV portfolio takes the most time; however, even for this case it still takes less than two minutes to solve on a PC. The rather light computational cost demonstrated here makes our method attractive in the real-time applications.

Table 2: Average Runtime of Computing the Test Statistic: 50 Assets

	MV	NC	ERC
runtime (in sec.)	97.72	0.23	0.42

Note: while we report the runtime for computing the test statistic, the computational cost for the confidence set can be deduced by multiplying this by the number of grid points for which the test statistic will be computed. Software: Knitro 12.0, Gurobi 8.0, AMPL 20200501, and Matlab R2017b. Hardware: Intel i7-6900K with 32 GB RAM.

6 Conclusion

We propose an inference procedure for estimators defined as the 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 linear inequalities with pre-estimated coefficients, which are amenable to the computationally simple inference procedures in Shi and Shum (2015). We provide an empirical application to the portfolio selection problem in finance; as far as we are aware, this represents the first instance of inference for this classic problem based on asymptotic approximation.

[29] We average over all mesh points τ used in Design 3 under $n = 500$.

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.[30] When the resulting inequalities, which can be arbitrarily nonlinear in the pre-estimated quantities, are moment inequalities, one can use the well-established methods in the moment inequality literature (e.g., Andrews and Soares (2010), Andrews and Barwick (2012), and Kline and Tamer (2016), among others) to construct joint confidence sets for (θ, s, λ) and then obtain the marginal confidence set for θ as projection of the joint confidence sets. Projection can lead to conservative inference, and there is a growing literature on the subvector inference (discussed in Section 4.3) which can potentially be helpful as well.

References

Andrews, D. and P. Barwick (2012): "Inference For Parameters Defined By Moment Inequalities: A Recommended Moment Selection Procedure," *Econometrica*, 80, 2805–2826.

Andrews, D. W. K. and G. Soares (2010): "Inference For Parameters Defined By Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157.

Andrews, I., T. Kitagawa, and A. McCloskey (2019a): "Inference on Winners," *Working Paper*.

Andrews, I., J. Roth, and A. Pakes (2019b): "Inference for Linear Conditional Moment Inequalities," *Working paper, Harvard University*.

Bhattacharya, D. (2009): "Inferring optimal peer assignment from experimental data," *Journal of the American Statistical Association*, 104, 486–500.

Bhattacharya, D. and P. Dupas (2012): "Inferring welfare maximizing treatment assignment under budget constraints," *Journal of Econometrics*, 167, 168–196.

Bonnans, J. and A. Shapiro (2000): *Perturbation Analysis of Optimization Problems*, Springer Science Business Media.

Boyd, S. and L. Vandenberghe (2004): *Convex Optimization*, Cambridge University Press.

Britten-Jones, M. (1999): "The Sampling Error in Estimates of Mean-Variance Efficient Portfolio Weights," *Journal of Finance*, 54, 655–671.

Bugni, F. A., I. A. Canay, and X. Shi (2017): "Inference For Subvectors And Other Functions Of Partially Identified Parameters In Moment Inequality Models," *Quantitative Economics*, 8, 1–38.

[30] For example, Martin (1985) shows that the KKT condition guarantees the global optimality for a wide class of mathematical programming problems. In general, KKT conditions may be necessary but not sufficient for the optimum in non-convex nonlinear programming problems. In this case, the inference methods in this paper may yield confidence sets for a superset of the optimizers.

Chen, X., T. M. Christensen, and E. Tamer (2018): "MCMC Confidence Sets for Identified Sets," *Econometrica*, 86, 1965–2018.

Chiong, K., Y.-W. Hsieh, and M. Shum (2017): "Counterfactual Estimation in Semiparametric Discrete Choice Models," *Working Paper*.

Chiong, K. X., A. Galichon, and M. Shum (2016): "Duality in Dynamic Discrete Choice Models," *Quantitative Economics*, 7, 83–115.

Cottle, R. W., J.-S. Pang, and R. E. Stone (1992): *The Linear Complementarity Problem*, Society for Industrial and Applied Mathematics.

Cox, G. and X. Shi (2020): "Simple Adaptive Size-Exact Test for Full-Vector and Subvector Inference in Moment Inequality Models," *Working paper, National University of Singapore.*

DeMiguel, V., L. Garlappi, F. J. Nogales, and R. Uppal (2009): "A Generalized Approach to Portfolio Optimization: Improving Performance by Constraining Portfolio Norms," *Management Science*, 798–812.

Dong, B., Y.-W. Hsieh, and M. Shum (2017): "Computing Moment Inequality Models using Constrained Optimization," *SSRN working paper #2990826.*

Fang, Z. and A. Santos (2019): "Inference on Directionally Differentiable Functions," *The Review of Economic Studies*, 86, 377–412.

Freyberger, J. and J. Horowitz (2015): "Identification and Shape Restrictions in Nonparametric Instrumental Variables Estimation." *Journal of Econometrics*, 189, 41–53.

Gafarov, B. (2016): "Inference on scalar parameters in set-identified affine models," *Working paper, UC Davis.*

Garlappi, L., R. Uppal, and T. Wang (2007): "Portfolio Selection with Parameter and Model Uncertainty: A Multi-Prior Approach," *Review of Financial Studies*, 20, 41–81.

Goldfarb, D. and G. Iyengar (2003): "Robust Portfolio Selection Problems," *Mathematics of Operations Research*, 1–38.

Graham, B., G. Imbens, and G. Ridder (2006): "Complementarity and the Optimal Allocation of Inputs," Manuscript.

Guggenberger, P., J. Hahn, and K. Kim (2008): "Specification Testing Under Moment Inequalities," *Economics Letters*, 99, 375–378.

Jagannathan, R. and T. Ma (2009): "Optimal versus Naive Diversification: How Inefficient is the Portfolio Strategy?" *The Review of Financial Studies*, 12, 937–974.

Jobson, J. D. and B. Korkie (1980): "Estimation for Markowitz Efficient Portfolios," *Journal of the American Statistical Association*, 75, 544–554.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals For Projections Of Partially Identified Parameters," *Econometrica*, 87, 1397–1432.

Kan, R. and D. R. Smith (2008): "The Distribution of the Sample Minimum-Variance Frontier," *Management Science*, 54, 1364–1380.

King, A. J. (1989): "Generalized Delta Theorems for Multivalued Mappings and Measurable Selections," *Mathematics of Operations Research*, 14, 720–736.

Kline, B. and E. Tamer (2016): "Bayesian inference in a class of partially identified models," *Quantitative Economics*, 7, 329–366.

Koenker, R. and G. Bassett (1978): "Regression Quantiles," *Econometrica*, 46, 33–50.

Luo, Z.-Q., J.-S. Pang, and D. Ralph (1996): *Mathematical Programs with Equilibrium Constraints*, Cambridge University Press.

Maillard, S., T. Roncalli, and J. Teiletche (2010): "The Properties of Equally Weighted Risk Contribution Portfolios," *The Journal of Portfolio Mangagement*, 60–70.

Mangasarian, O. (1969): *Nonlinear Programming*, Society for Industrial and Applied Mathematics.

Markowitz, H. (1952): "Portfolio Selection," *Journal of Finance*, 7, 77–91.

Martin, D. H. (1985): "The Essence of Invexity," *Journal of Optimization Theory and Applications*, 65–76.

Merton, R. C. (1972): "An Analytical Derivation of the Efficient Portfolio Frontier," *Journal of Finance and Quantitative Analysis*, 7, 1851–1872.

Michaud, R. (1989): "The Markowitz Optimization Enigma:is The Optimized Optimal?" *Financial Analysts Journal*, 45, 31–42.

Mogstad, M., A. Santos, and A. Torgovitsky (2017): "Using Instrumental Variables for Inference About Policy Relevant Treatment Effects," *NBER Working Paper 23568*.

Okhrin, Y. and W. Schmid (2006): "Distributional Properties Of Portfolio Weights," *Journal of Econometrics*, 134, 235–256.

Romano, J. P. and A. M. Shaikh (2010): "Inference For The Identified Set In Partially Identified Econometric Models," *Econometrica*, 78, 169–211.

Russell, T. (2017): "Sharp Bounds on Functionals of the Joint Distribution in the Analysis of Treatment Effects," *Working Paper, University of Toronto.*

Scherer, B. (2002): "Portfolio Resampling: Review and Critique," *Financial Analysts Journal*, 58, 98–109.

Shapiro, A. (1993): "Asymptotic Behavior of Optimal Solution in Stochastic Programming," *Mathematics of Operations Research*, 18, 829–845.

Shi, X. and M. Shum (2015): "Simple Two-stage Inference For A Class Of Partially Identified Models," *Econometric Theory*, 31, 493–520.

Shi, X., M. Shum, and W. Song (2018): "Estimating Semi-parametric Panel Multinomial Choice Models using Cyclic Monotonicity," *Econometrica*, 86, 737–761.

Williams, H. P. (2013): *Model Building in Mathematical Programming*, Wiley.

Wolak, F. A. (1987): "An exact test for multiple inequality and equality constraints in the linear regression model," *Journal of the American Statistical Association*, 82, 782–793.

——— (1989a): "Local and global testing of linear and nonlinear inequality constraints in nonlinear econometric models," *Econometric Theory*, 5, 1–35.

——— (1989b): "Testing inequality constraints in linear econometric models," *Journal of Econometrics*, 41, 205–235.

A Proof of Theorem 1

Proof. Observe that there exists a sequence $\{(P_n, \theta_n)\}$ such that $\theta_n \in \Theta_0(P_n)$ and $P_n \in \mathcal{P}_0$ such that the left-hand side of the inequality is equal to

$$\liminf_{n\to\infty} \Pr{}_{P_n}(\theta_n \in CS_n^{\mathrm{PD}}(1-\alpha)). \tag{52}$$

By the definition of $\liminf$, there exists a subsequence $\{u_n\}$ of $\{n\}$ such that the above expression is equal to

$$\lim_{n\to\infty} \Pr{}_{P_{u_n}}(\theta_{u_n} \in CS_{u_n}^{\mathrm{PD}}(1-\alpha)). \tag{53}$$

Next we show that for any subsequence of $\{u_n\}$, there exists a further subsequence $\{a_n\}$ such that

$$\lim_{n\to\infty} \Pr{}_{P_{a_n}}(\theta_{a_n} \in CS_{a_n}^{\mathrm{PD}}(1-\alpha)) \geq 1-\alpha. \tag{54}$$

This concludes the proof.

By Assumption 1(d), there exists a sequence $\{\lambda_n, s_n\}$ such that $\lambda_n \geq 0, s_n \geq 0$, $\lambda_n' s_n = 0$, $g(A_{P_n}, b_{P_n}, c_{P_n}, \theta_n, \lambda_n, s_n) = 0$, and $\|\theta_n\|, \|\lambda_n\| \leq C$. Due to the compactness of the ball with radius C, for any subsequence of $\{n\}$, there exists a further subsequence $\{a_n\}$ such that $\theta_{a_n} \to \theta_\infty$ and $\lambda_{a_n} \to \lambda_\infty$ for some finite vectors θ_∞ and λ_∞. By Assumptions 1(b)-(c) this further subsequence can be chosen so that $V_{P_{a_n}} \to V$ and the distributional convergence in Assumption 1(c) holds. Therefore, by the delta method,

$$\sqrt{a_n} g(\widehat{A}_{a_n}, \hat{b}_{a_n}, \hat{c}_{a_n}, \theta_{a_n}, \lambda_{a_n}, s_{a_n}) \to_d N(0, G(\theta_\infty, \lambda_\infty) V G(\theta_\infty, \lambda_\infty)'). \tag{55}$$

Assumption 1(c) also implies that $\widehat{V}_{a_n} \to_p V$. Thus,

$$G(\theta_{a_n}, \lambda_{a_n}) \widehat{V}_{a_n} G(\theta_{a_n}, \lambda_{a_n})' \to_p G(\theta_\infty, \lambda_\infty) V G(\theta_\infty, \lambda_\infty)'. \tag{56}$$

Assumption 1(d) implies that the limiting variance matrix is invertible. Thus, by appealing to the continuous mapping theorem, we obtain

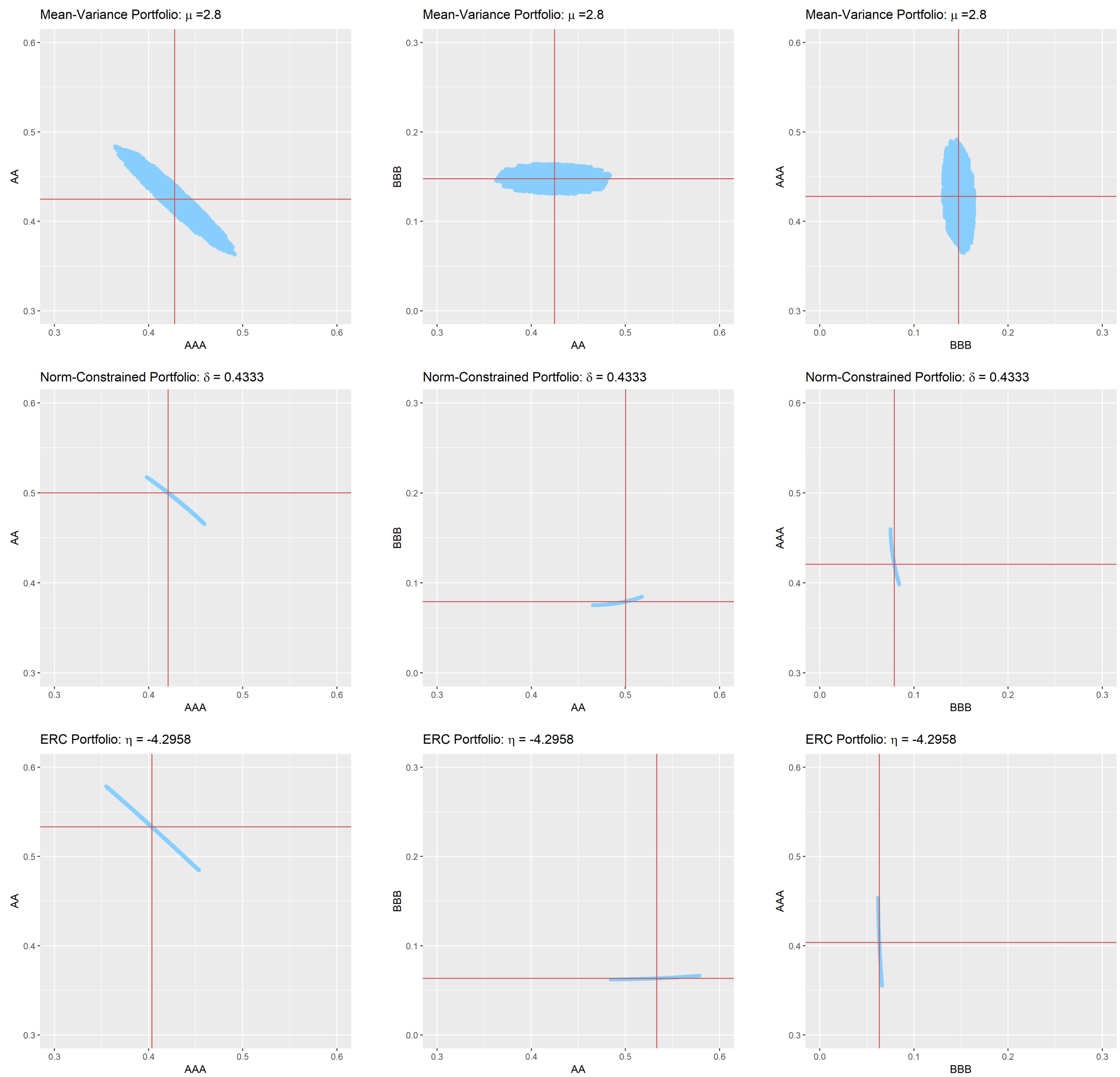
$$a_n \widehat{Q}_{a_n}(\theta_{a_n}, \lambda_{a_n}, s_{a_n}) \to_d \chi^2_{m+k}. \tag{57}$$

Therefore,

$$\begin{aligned}\Pr{}_{P_{a_n}}(\theta_{a_n} \in CS^{\mathrm{PD}}_{a_n}(1-\alpha)) &= \Pr{}_{P_{a_n}}\left(\min_{\lambda\geq 0, s\geq 0, \lambda' s=0} a_n \widehat{Q}_{a_n}(\theta_{a_n}, \lambda, s) \leq \chi^2_{m+k,1-\alpha}\right) \\ &\geq \Pr{}_{P_{a_n}}(a_n \widehat{Q}_{a_n}(\theta_{a_n}, \lambda_{a_n}, s_{a_n}) \leq \chi^2_{m+k,1-\alpha}) \\ &\to 1-\alpha. \end{aligned} \tag{58}$$

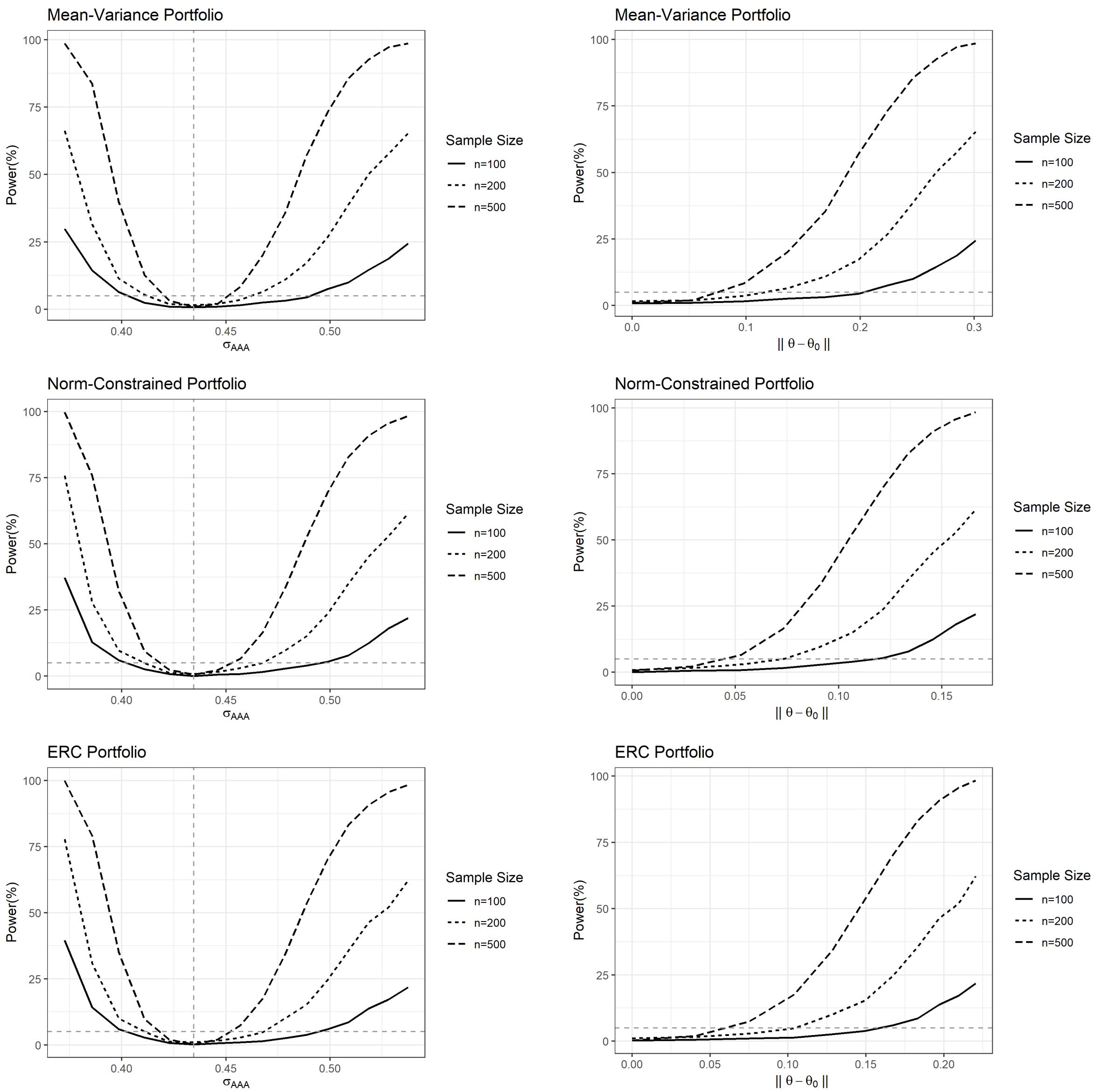
B Figures

Figure 1: 90% Confidence Set of Optimal Portfolio Weights under Different Models



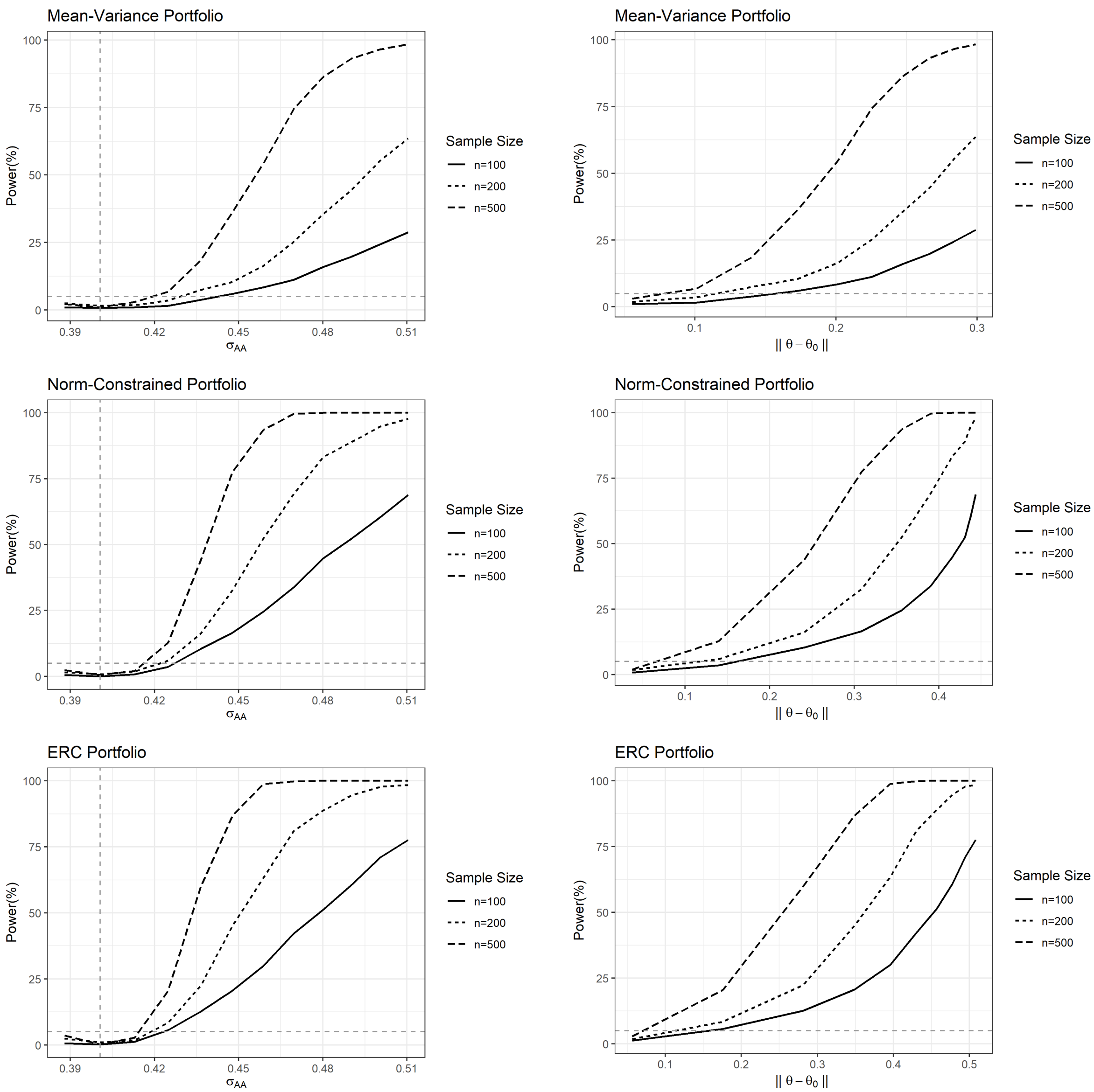
Note: The solution of the portfolio weights based on the estimated $(\hat{R}, \hat{Q})$ are located by two red lines.

Figure 2: Power Curve: Volatility Shift in AAA Corporate Bonds; Normally Distributed Data



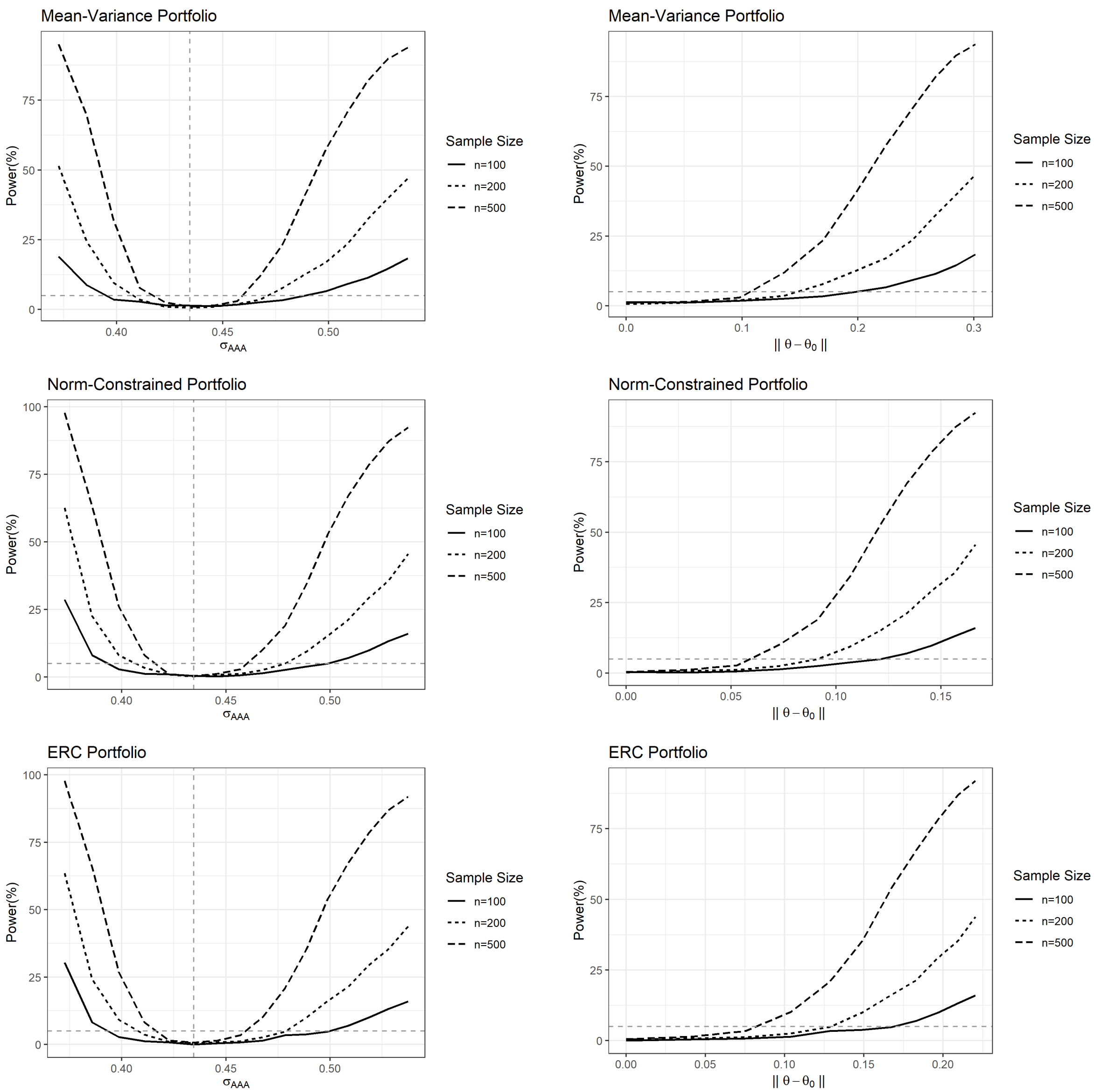
Note: The horizontal line locates the nominal size $\alpha = 0.05$, and the vertical line locates the value of σ_{AAA} that implies the null hypothesis. Left column: power as the function of the predetermined mesh points of σ_{AAA}. Right column: power as the function of the Euclidean distance between the true optimal portfolio weight θ_0 (that changes with σ_{AAA}) and the fixed values of θ, which are set to the estimated values in Table 1.

Figure 3: Power Curve: Volatility Shift in AA Corporate Bonds; Normally Distributed Data



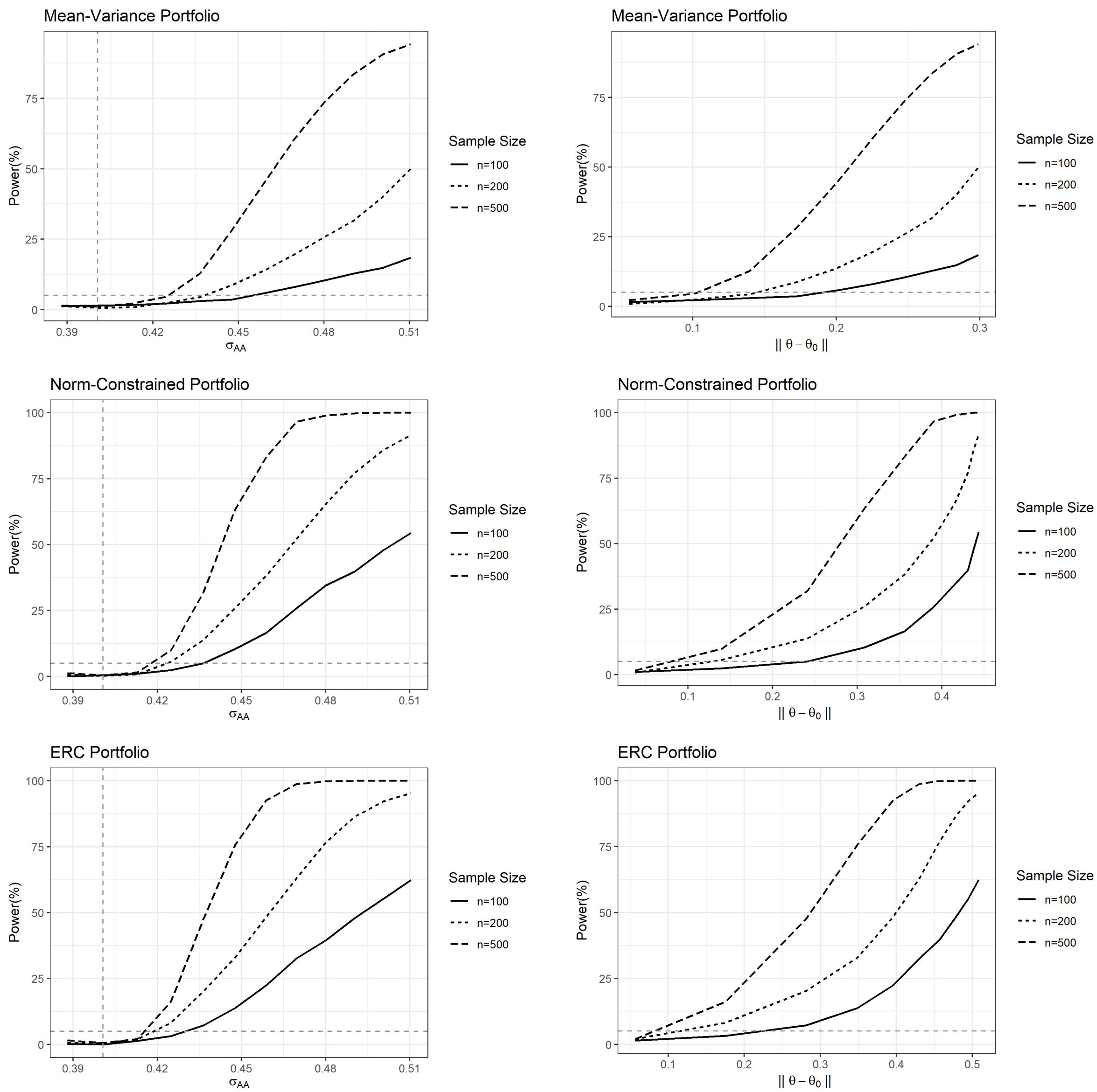
Note: The horizontal line locates the nominal size $\alpha = 0.05$, and the vertical line locates the value of σ_{AA} that implies the null hypothesis. Left column: power as the function of the predetermined mesh points of σ_{AA}. Right column: power as the function of the Euclidean distance between the true optimal portfolio weight θ_0 (that changes with σ_{AA}) and the fixed values of θ, which are set to the estimated values in Table 1.

Figure 4: Power Curve: Volatility Shift in AAA Corporate Bonds; t Distributed Data



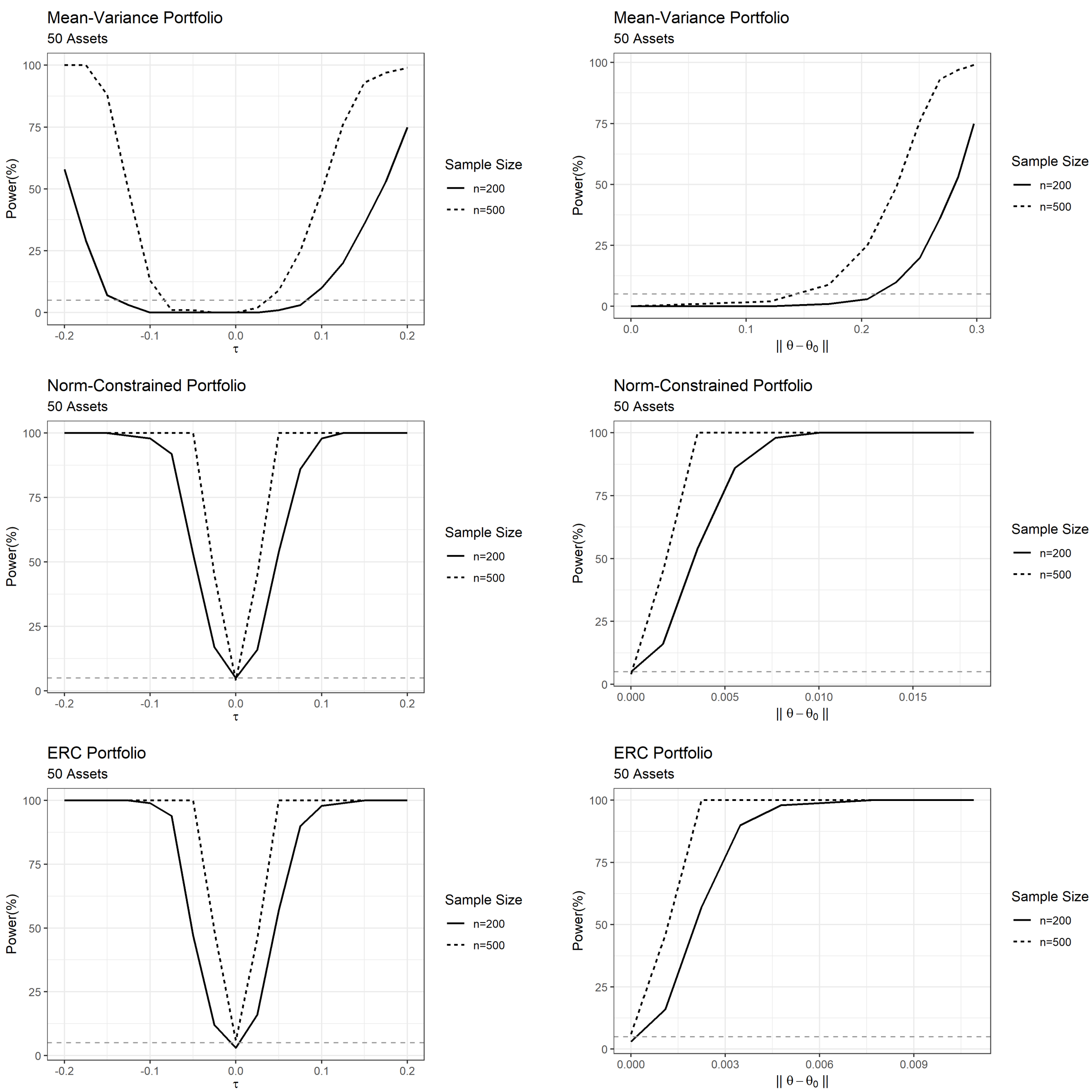
Note: The horizontal line locates the nominal size $\alpha = 0.05$, and the vertical line locates the value of σ_{AAA} that implies the null hypothesis. Left column: power as the function of the predetermined mesh points of σ_{AAA}. Right column: power as the function of the Euclidean distance between the true optimal portfolio weight θ_0 (that changes with σ_{AAA}) and the fixed values of θ, which are set to the estimated values in Table 1.

Figure 5: Power Curve: Volatility Shift in AA Corporate Bonds; t Distributed Data



*Note:*The horizontal line locates the nominal size $\alpha = 0.05$, and the vertical line locates the value of σ_{AA} that implies the null hypothesis. Left column: power as the function of the predetermined mesh points of σ_{AA}. Right column: power as the function of the Euclidean distance between the true optimal portfolio weight θ_0 (that changes with σ_{AA}) and the fixed values of θ, which are set to the estimated values in Table 1.

Figure 6: Power Curve: Factor Loading Shift



Note: The horizontal line locates the nominal size $\alpha = 0.05$. Left column: power as the function of the predetermined mesh points of τ that shifts the first 25 assets' factor loadings; $\tau = 0$ implies the null hypothesis. Right column: power as the function of the Euclidean distance between the true optimal weights θ_0 (that changes with τ) and the value of θ under the null hypothesis ($\tau = 0$).