Inference on scalar parameters in set–identified affine models

Bulat Gafarov*

Preliminary and incomplete draft.

2018-09-24

Abstract

This paper proposes both point-wise and uniform confidence intervals (CIs) for an element θ_1 of a parameter vector $\theta \in \mathbb{R}^d$ that is partially identified by affine moment equality and inequality conditions. The CIs are based on an estimator of a regularized support function of the identified set. This estimator has a Bahadur representation that provides closed form standard errors and also enables (fast) multiplier bootstrap inference. Unlike much of the existing literature, the proposed CIs can be computed as a solution to a convex optimization problem, which leads to a substantial decrease in computation time (relative to the existing uniform procedures). The proposed approach is extended to construct joint confidence sets for multiple components of θ.

This procedure can be used, for example, to compute a CI for a coefficient in a linear regression model with interval outcome without additional distributional assumptions.

Key Words: Affine moment inequalities; Bahadur representation; Delta–Method; Interval data; Partial identification; Regularization; Stochastic Programming; Subvector inference; Uniform inference.

*University of California, Davis, Department of Agricultural and Resource Economics. E-mail: bgafarov@ucdavis.edu; I am extremely grateful to Joris Pinkse and Patrik Guggenberger for their very helpful and detailed comments on this paper. I would like to thank Donald Andrews, Andres Aradillas-Lopez, Christian Bontemps, Joachim Fryberger, Ronald Gallant, Michael Gechter, Marc Henry, Keisuke Hirano, Sung Jae Jun, Nail Kashaev, Francesca Molinari, Adam Rosen, Xiaoxia Shi, Jing Tao, Alexander Torgovitsky, and Fang Zhang (in alphabetical order) for the comments and suggestions on this project. The first draft date: November 10th, 2015.

1 Introduction

Strong econometric assumptions can lead to poor estimates. Moment inequalities occasionally provide alternative estimates under weaker assumptions. Linear models with interval-valued data are a good example.[1] It is common practice to replace the income bracket data with the corresponding midpoints when estimating the returns to schooling (Trostel et al. (2002)). The conventional approach is applicable only under strong assumptions on the distribution of the residual term.[2] The affine moment inequality approach to interval-valued data proposed by Manski and Tamer (2002) can set-identify the return to schooling without such strong assumptions.

There are multiple methods that can be used to construct Confidence Sets (CS) for parameters defined by moment inequality. The pioneering procedures of Chernozhukov et al. (2007) and Andrews and Soares (2010, AS) and their subsequent refinements by Bugni et al. (2016) (BCS) and Kaido et al. (2015) (KMS) are powerful procedures that solve this inference problem in the small-dimensional case. Some applications, such as panel or semiparametric regression models with interval measured outcome variables, have a large dimension of the parameter space (the number of the regression coefficients) which poses a challenge for the existing procedures.

I propose confidence intervals (CIs) for an element θ_1 of an unknown parameter vector $\theta \in \mathbb{R}^d$ in models defined by affine moment equalities and inequalities. In the returns to schooling example, θ_1 corresponds to the returns to schooling and $\theta \in \mathbb{R}^d$ to the full vector of the regression coefficients that can include many control variables. I estimate the lower and upper extremes of the identified set for θ_1, which is an interval, using an estimator of the regularized support function. This estimator has a closed-form asymptotic Gaussian distribution which I use to construct both point–wise valid and uniform CIs for θ_1. The latter asymptotically controls the coverage probability uniformly over a class of data generating processes (DGP) (as it was pointed out in Imbens and Manski (2004), the uniformity in DGP is desirable as it controls coverage probability in finite sample properties better than point-wise CIs).

The regularized support function proposed in this paper is a solution to a convex quadratic program that minimizes the sum of θ_1 and a penalty $\mu_n \|\theta\|^2$ with $\mu_n \to 0$, subject to the sample moment restrictions. If the set of optima for $\mu = 0$ is not a singleton, this additional convex term selects the optimum with the minimal norm. The standard errors are computed using the sample variance of the weighted moment conditions at the unique optima. To correct the asymptotic bias resulting from the regularization exactly, I suggest using the argmin of the regularized program with a larger tuning parameter $\kappa_n \to 0$. If $\kappa_n/\mu_n \to \infty$ as $n \to \infty$, then the bias correction does not affect the asymptotic distribution of the estimator. To achieve a uniformly valid CI, I replace the exact correction with an upper bound on the maximum of $\mu_n \|\theta\|^2$ over the argmin set of the non-regularized program.

The proposed CIs have several attractive statistical and computational properties which make them viable in high dimensional affine moment inequality models.

Bahadur representation. The estimator of the regularized support function has a Bahadur representation that provides easy to compute asymptotic standard error and enable the multiplier bootstrap. This resampling procedure avoids the necessity of solving mathematical programs for every bootstrap draw present in the existing uniform methods.

[1] Other examples of affine moment inequalities include monotone instrumental variables (Manski and Pepper (2000),Freyberger and Horowitz (2015)) and models with missing data (Manski (2003)).

[2] Another common approach is to assume Gaussian distribution for the residuals and apply Maximum Likelihood method (Stewart (1983)).

This paper is the first to propose a closed–form estimator of the bounds on θ_1 in affine moment inequality models with asymptotic Gaussian distribution. In contrast, the estimator of the ordinary support function used in the existing literature (Beresteanu and Molinari (2008), Kaido and Santos (2014), Freyberger and Horowitz (2015, FH), Gafarov et al. (2015), among others) has non-Gaussian asymptotic distribution, which complicates inference.

Computational properties. The proposed approach requires only a fraction of the computational time of the existing pointwise and uniform procedures, in particular if θ has a large dimension. The computational cost is low since it involves only four quadratic programs, it does not require any resampling and it depends on covariance of the moment conditions at two points.

The computation time for my procedure increases only slowly in the dimension of $\theta \in \mathbb{R}^d$ and takes 1.5 sec only for $d = 15$ and 30 moment inequalities. As a result, the proposed method can address the parameters with a large dimension and a large number of moment conditions. In contrast, the existing uniform inference methods for moment inequalities proposed by AS, KMS, and BCS are based on costly non–convex optimization.

The new estimators, which are based on strictly convex programs, also avoid the problem of distinguishing between local and global optima, which is present for the existing uniform procedures even in the affine moment inequality setup. The low computational cost together with applicability of the fast multiplier bootstrap allows one to perform joint inference on the components of θ using the support function representation of a convex set.

I provide an example of an affine moment inequality model illustrating that the number of local optimal solutions in the existing uniform procedures (AS,BCS, and KMS) can grow exponentially with the dimension d. As a result, the procedures take more computational time and can produce misleadingly short CIs if the optimization routine fails to find the global optimum. It takes 630 sec to compute the CI of AS in an affine model with $d = 15$ and 30 moment inequalities which is 420 times slower than my procedure. [3] In my numerical experiments the computational time for the AS procedure increases by 30% with every additional dimension d while my procedure is barely affected by changes in the dimension.

Length comparison. The proposed uniform CIs have length properties that are not worse than these of the existing methods as suggested by Monte Carlo experiments. The proposed *uniform* CI is shorter than the projection CI of AS in every MC design considered in the paper.

Applicability. The uniform CI is applicable in a situation where the existing uniform procedures are inapplicable. I show that a linear model with an interval-valued outcome can have a moment inequality with zero variance which violates the assumptions in AS, Kaido et al. (2015, KMS) and Bugni et al. (2014, BCS).[4]

The class of DGPs over which I prove uniform coverage properties is not nested in the classes considered in BCS and KMS. I impose a rank condition on the affine constraints typical for the support function approach (Beresteanu and Molinari (2008), Kaido and Santos (2014), FH, Gafarov et al. (2015)). These conditions rule out over-identification of the solutions to the regularized

[3] I use the implementation of the AS procedure provided by KMS. This algorithm uses smooth interpolation of the critical values by the kriging method which allows one to use a Newton-type solver. This approach reduces the computational cost of the AS procedure.

[4] AS, KMS, and BCS procedures can be applied in some setups where my procedure is not applicable. I compare setups in Section 2.5.

programs. In particular, they rule out the possibility of point-identification by moment inequalities of the components of θ, which can be addressed using BCS and KMS procedures. This complication can be alleviated if one can split the full set of the inequality constraints into (possibly overlapping) subsets that meet the rank condition. Within this framework, my procedure covers θ_1 for any sequences of DGP that drift to a DGP with a moment condition orthogonal to θ_1, which is not the case in BCS.[5] As mentioned earlier, my CIs remain valid if some moment inequalities have zero variance, which violates assumptions in both KMS and BCS. I expect poor coverage of the existing procedures as the variance becomes very small (but still positive).

The number of maintained assumptions for the large sample inference in the present paper is kept to a minimum. All the assumptions are testable.

The paper is structured as follows. Section 2 describes the setup and the main result. Section 3 outlines the extension to the general subvector inference and show how one can deal with the violation of the main regularity assumptions on the moment conditions. Section 4 compares the maintained assumptions and the computational properties of the proposed procedure with the existing alternatives. Section 5 provides the results of the Monte Carlo experiments. Section 6 outlines a possible empricial application. Section 7 concludes.

Notation. I use $\triangleq$ to denote definitions. I write $\mathbb{E}_P[\cdot]$ to denote expectation with respect to a probability distribution P. I use uppercase English letters to denote random variables (scalar,vector, or matrix valued) and lower case letters to denote the corresponding realizations, W and w_i. I use $\mathbb{P}_n$ for the sample distribution. I use $f(0^+)$ for $\lim_{x\downarrow 0} f(x)$. The vector $e_j \triangleq (0, ...1, ...0)'$ is the j-th coordinate vector, where the one occurs at position j. e_j is the projector on the j-th coordinate. I use the symbol $\mathcal{J}$ for a finite set of indices $\mathcal{J} \triangleq \{i_1, ..., i_\ell\} \subset \mathbb{N}$ and $\mathbb{J} \triangleq (e_{i_1}, ..., e_{i_\ell})'$ as a coordinate projection matrix in the the corresponding Euclidean space. I use $|\mathcal{J}|$ to denote the cardinality of the set $\mathcal{J}$. The acronym u.h.c. stands for upper hemi-continuous correspondence. I will use symbol $\text{sVar}(x)$ to denote the sample variance, $\text{sVar}(x) = \frac{1}{n}\sum_{i=1}^n x_i^2 - \left(\frac{1}{n}\sum_{i=1}^n x_i\right)^2$.

2 Setup and main results

2.1 Affine moment conditions

I consider a parameter vector $\theta \in \Theta \subset \mathbb{R}^d$ where Θ is the set defined by

$$-\infty < a_\ell \leq \theta_\ell \leq b_\ell < \infty \tag{1}$$

for $\ell = 1, ..., d$. The inequalities (1) can be written as a subsystem of the following system of unconditional moment equalities/inequalities

$$\begin{cases} \mathbb{E}_P g_j(W, \theta) = 0, & j \in \mathcal{J}_k^{eq}, \\ \mathbb{E}_P g_j(W, \theta) \leq 0, & j \in \mathcal{J}_k^{ineq}, \end{cases} \tag{2}$$

where $g_j(W, \theta) \triangleq \sum_{\ell=1}^d W_{j\ell}\theta_\ell - W_{j(d+1)}$, $|\mathcal{J}_k^{eq}| = p$, $0 \leq p \leq d$, $\left|\mathcal{J}_k^{ineq}\right| = k - p \geq 2d$, $k < \infty$, the random matrix W has probability measure P with the sample space $\mathbb{R}^{k\times(d+1)}$. Correspondingly, $\left|\mathcal{J}_k^{eq} \cup \mathcal{J}_k^{ineq}\right| = k$. A solution to (2) may not be unique. Let the identified set $\Theta(P) \subset \Theta \subset$

[5] See KMS for a discussion of the assumptions in BCS

$\mathbb{R}^d$ be the set of parameter values θ that satisfy (2) for a given data generating process (DGP) parametrized by P. The stochastic programming approach described below allows me to deal with both random and deterministic (in)equalities in (2) symmetrically.

The identified set $\Theta(P)$ is a polytope or an empty set. The convexity of $\Theta(P)$ provides characterization using support functions. Such support functions, for example, can provide bounds on coordinate projections of $\Theta(P)$ or any subvectors of $\theta \in \Theta(P)$. Without loss of generality, I will consider first a special case of $\theta_1 = e_1'\theta$, the value of the first component of $\theta \in \Theta(P)$. Any other support function can be obtained by a corresponding rotation of W and θ.

Definition 1. *The marginal identified set for* θ_1 is the set

$$\mathcal{S}(P) = \{e_1'\theta | \theta \in \Theta(P)\}. \qquad \square$$

The first assumption I make rules out a possibility of specification and makes boundaries of $\mathcal{S}(P)$ well defined.

Assumption 1. $\Theta(P)$ *is nonempty for the probability measure* P.

The following example illustrates the general setup.

Example 1 (*Linear model with interval valued outcome*)**.** Consider a linear model $\mathbb{E}_P[Y|Z] = \theta'Z$, where Y is unobserved. One can only observe bounds $\underline{Y}$ and $\overline{Y}$, (random or deterministic) such that $Y \in \left[\underline{Y}, \overline{Y}\right]$ a.s. Suppose that Z, the random vector of regressors, has a finite support $S_Z = \{z_1, \ldots, z_K\} \subset \mathbb{R}^d$. In this case the model can be equivalently characterized[6] by a finite number of conditional moments:

$$\mathbb{E}_P\left[\overline{Y}|Z = z_j\right] \;\geq\; \theta'z_j \geq \mathbb{E}_P\left[\underline{Y}|Z = z_j\right],\ j = 1, \ldots, K.$$

The identified set $\Theta(P)$ is defined by the set of unconditional moment inequalities,

$$\begin{cases} \mathbb{E}_P\left[\underline{Y}1\{Z = z_j\}\right] \leq \theta'z_j\mathbb{E}_P 1\{Z = z_j\}, & j = 1, ..., K, \\ \mathbb{E}_P\left[\overline{Y}1\{Z = z_{j-K}\}\right] \geq \theta'z_j\mathbb{E}_P 1\{Z = z_{j-K}\}, & j = K+1, ..., 2K. \end{cases} \tag{3}$$

These inequalities can be converted to the form (2) with $p = 0$, $k = 2K$ and

$$W_{j\ell} \triangleq \begin{cases} -Z_\ell 1\{Z = z_j\}, & \text{for } j = 1, ..., K, \\ Z_\ell 1\{Z = z_{j-K}\}, & \text{for } j = K+1, ..., 2K, \end{cases}$$

$$W_{j(d+1)} \triangleq \begin{cases} \underline{Y}1\{Z = z_j\}, & \text{for } j = 1, ..., K, \\ -\overline{Y}1\{Z = z_{j-K}\}, & \text{for } j = K+1, ..., 2K. \end{cases}$$

One can incorporate additional information such as sign restrictions on θ in the form of linear inequalities to get a smaller identified set.

[6] If S_Z is infinite, one can estimate an enlargement of $\mathcal{S}(P)$ using a finite number of unconditional moment inequalities. See Chernozhukov et al. (2007) for details. Andrews and Shi (2013) provide conditions for sharp characterization of the identified set by a finite number of unconditional moment functions. I leave the case of infinite number of moment inequalities for future extensions.

Note that this example has discrete-valued regressors which makes the existing approach of Bontemps et al. (2012) inapplicable. These authors provide a sharp characterization of the identified set and the corresponding confidence intervals in a class of linear models with interval-valued outcome if all of the regressors have continuous support. □

Since all the moment conditions are affine, the identified set $\Theta(P)$ is a polytope and the marginal identified set is an interval, $\mathcal{S}(P) = [\underline{v}(P), \bar{v}(P)]$, where

$$\underline{v}(P) = \min_{\theta \in \Theta(P)} e_1'\theta \quad \text{and} \quad \bar{v}(P) = \max_{\theta \in \Theta(P)} e_1'\theta. \tag{4}$$

It is possible that $\mathcal{S}(P)$ is a singleton. The value functions $\underline{v}(P)$ and $-\bar{v}(P)$ are the support functions for e_1 and $-e_1$, respectively. The analysis for the upper bound is analogous to that for the lower bound, so from here on I focus on the lower bound.

2.2 Regularized support function.

The Delta-method framework is a natural way to do inference on $\underline{v}(P)$. In order to use this approach we need to study the behavior of Program 4 when we replace the coefficients $\mathbb{E}_P W$ with their consistent estimators. Namely, we need to consider a (directional) derivative of $\underline{v}(P)$ with respect to $\mathbb{E}_P W$. This derivative can be obtained by the envelope theorem applied to the min/max representation for Program 4 which is valid under Assumption 1,

$$\underline{v}(P) = \min_{\theta \in \mathbb{R}^d} \max_{\lambda \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}} \{\theta_1 + \lambda' \mathbb{E}_P g(W, \theta)\}. \tag{5}$$

This representation shows that the derivative of $\underline{v}(P)$ with respect to $\mathbb{E}_P W$ depends not only on the optimal solution of Program 4 but also on the solution $\underline{\lambda}(P)$ to the corresponding dual program which is defined below. Let $\mathbb{E}_P W \triangleq (A_P, b_P)$ so that $\mathbb{E}_P g(W, \theta) = A_P \theta - b_P$. The dual program takes form

$$\begin{aligned} \underline{v}(P) = \max_{\lambda \in \mathbb{R}^p \times \mathbb{R}_+^{k-p}} & \{-\lambda' b_P\} \\ \text{s.t. } & \lambda' A_P = e_1'. \end{aligned} \tag{6}$$

If solutions to both (4) and (6) are unique, the envelope theorem suggests

$$\frac{\partial \underline{v}(P)}{\partial (A_P)_{ij}} = \underline{\lambda}_i \underline{\theta}_j \text{ and } \frac{\partial \underline{v}(P)}{\partial (b_P)_i} = -\underline{\lambda}_i. \tag{7}$$

Shapiro (1993) shows that under Assumption 1 that if we use $\frac{1}{n}\sum_{i=1}^n w_i$ as a consistent estimator for $\mathbb{E}_P W$, the value of

$$\hat{\underline{v}}_n = \min_{\theta \in \mathbb{R}^d} e_1'\theta \tag{8}$$

$$\text{s.t. } \begin{cases} \frac{1}{n}\sum_{i=1}^n g_j(w_i, \theta) = 0, & j \in \mathcal{J}_k^{eq}, \\ \frac{1}{n}\sum_{i=1}^n g_j(w_i, \theta) \leq 0, & j \in \mathcal{J}_k^{ineq}, \end{cases} \tag{9}$$

is a consistent estimator of $\underline{v}(P)$ with a non–Gaussian asymptotic distribution that depends on both $\underline{\theta}$ and $\underline{\lambda}$. These parameters are not uniquely defined in general which makes them nuisance

parameters. I suggest regularizing Program (4) to ensure a unique $\underline{\theta}$ and imposing a constraint qualification on $\mathbb{E}_P g(W,\theta)$ to ensure a unique $\underline{\lambda}$.

One can approximate the program (4) from above by its *regularized* counterpart with a unique solution,

$$\underline{v}(\mu_n, P) = \min_{\theta \in \Theta(P)} \left\{ e_1'\theta + \mu_n \|\theta\|^2 \right\}, \tag{10}$$

The argmin to (10), $\underline{\theta}(\mu_n, P)$, is a continuous and single–valued function of the tuning parameter μ_n, that shrinks to zero as the sample size grows to infinity.[7] The one–sided limit is the element with the minimal norm in to the argmin set of (4),

$$\underline{\theta}\left(0^+, P\right) \triangleq \lim_{n\to\infty} \underline{\theta}(\mu_n, P) = \operatorname*{argmin}_{\theta \in \Theta(P), \theta_1 = \underline{v}(P)} \|\theta\|.$$

If μ_n converges to zero, then

$$\underline{v}(\mu_n, \mathbb{P}_n) = \min_{\theta \in \Theta(\mathbb{P}_n)} \left\{ e_1'\theta + \mu_n \|\theta\|^2 \right\}, \tag{11}$$

is a consistent estimator of $\underline{v}(P)$ with asymptotic distribution that depends only on the distribution of $g(W, \underline{\theta}(0^+, P))$.

The constraint qualification can be formulated in form of the following two assumptions. For any $\mathcal{J} \subset \mathcal{J}_k^{ineq}$ let the projection matrix $\mathbb{J}_k^a = (e_{i_1}, ..., e_{i_\ell})'$ correspond to $\mathcal{J}_k^a \triangleq \mathcal{J}_k^{eq} \cup \mathcal{J} = \{i_1, ..., i_\ell\}$. Let $\text{eig}(\cdot)$ be the minimium eigenvalue of a matrix.

Assumption 2. *For any $\mathcal{J} \subset \mathcal{J}_k^{ineq}$ with $|\mathcal{J}| = d - p$, the following matrix has minimum eigenvalues bounded by a positive number,*

$$\text{eig}\left\{ (\mathbb{J}_k^a \mathbb{E}_P W)(\mathbb{J}_k^a \mathbb{E}_P W)' \right\} \geq \eta^2 > 0. \tag{12}$$

In geometric terms, Assumption 2 restricts angles between the gradients of any d intersecting moment restrictions to be away from zero.

Assumption 3. *There is some $\epsilon > 0$ such that for any $\mathcal{J} \subset \mathcal{J}_k^{ineq}$ with $|\mathcal{J}| \geq d + 1 - p$ and any $\theta \in \Theta(P)$,*

$$\|\mathbb{J}_k^a m(\theta, P)\| \geq \epsilon. \tag{13}$$

Together Assumptions 2-3 imply that at any point in $\theta \in \Theta(P)$ any *binding*(or *active*) moment conditions have linearly independent gradients, a condition called *Linear Independence Constraint Qualification (LICQ)* in the optimization theory.[8] LICQ is a necessary and sufficient condition for uniqueness of the Lagrange multipliers $\underline{\lambda}$.[9] Under LICQ the value of Program (10) for $\mu_n > 0$ is differentiable in $\mathbb{E}_P W$ so that its sample analog has asymptotic Gaussian distribution.[10]

Constraint qualifications are common in the literature on set identified models. In particular, BM, KS, FH and Gafarov et al. (2015) impose them in various forms.

Assumptions 2-3 are testable for any fixed η and ϵ which make them more convenient for statistical applications than LICQ. Moreover, Lemma 4 in Appendix provide an explicit bound

[7] See Lemma 5 in Appendix.

[8] See Assumption 6 and Lemma 2 in Appendix.

[9] See Wachsmuth (2013).

[10] See Lemma 7 in Appendix.

on $\underline{\lambda}$ and correspondingly on the derivative of $\underline{v}(P)$ with respect to $\mathbb{E}_P W$ that depends only on η and the diameter of Θ. This property is convenient for establishing a uniform convergence to a Gaussian limit of $\underline{v}(\mu_n, \mathbb{P}_n)$, which will be investigated below in Section 2.5.

There is a purely computational reason to impose LICQ. If LICQ is violated, Newton-type algorithms, which typically guarantee quadratic rate of convergence to a stationary point, have linear rate of convergence or do not converge at all.[11]

Assumption 3 can be violated in applications with many moment inequality conditions, in particular, if the identified set is very tight as a result.[12] In Section 3 I show that how one can use a representation of $\underline{v}(P)$ as a maximum of sub-problems that meet this requirement even if Assumption 3 is violated,.

2.3 Large sample properties of the regularized support function.

I make the following two standard assumptions for the large sample inference.

Assumption 4. $\left\{w_i \in \mathbb{R}^{k\times(d+1)} | \, i = 1, ..., n\right\}$ *is an i.i.d. sample with probability measure* P.

Assumption 5. *There exist an* $\varepsilon > 0$ *and* $\bar{M}$ *such that*

$$\mathbb{E}_P \|W\|^{2+\varepsilon} \leq \bar{M} < \infty. \tag{14}$$

Assumptions 4-5 are sufficient to guarantee a uniform law of large numbers (LLN) for first and second moments of W and a central limit theorem (CLT) for the estimated coefficients in Program (11). In particular, Assumption 5 provides an explicit bound on the CLT approximation error using the version of the theorem of Yurinskii (1978) by van der Vaart and Wellner (1996). Alternatively, one can assume LLN and CLT directly to allow for dependent data.

Let the error of the Gaussian approximation using the Levy-Prohorov metric be denoted as

$$\rho_n(P) \triangleq \pi(\sqrt{n}(\underline{v}(\mu_n, \mathbb{P}_n) - \underline{v}(\mu_n, P_n)), N\left(0, \underline{\sigma}^2(\mu_n, P)\right)).$$

The asymptotic variance here can be estimated using

$$\underline{\sigma}^2(\mu_n, \mathbb{P}_n) = \mathrm{sVar}\left(\underline{\lambda}'(\mu_n, \mathbb{P}_n) g(w_i, \underline{\theta}(\mu_n, \mathbb{P}_n))\right),$$

where $\mathrm{sVar}(\cdot)$ is the sample variance operator, $\underline{\theta}(\mu_n, \mathbb{P}_n)$ and $\underline{\lambda}(\mu_n, \mathbb{P}_n)$ are, respectively, the optimum and the vector of Lagrange multipliers of (11).

The following theorem provides consistency results for these quantities.

Theorem 1. *Consider any sequence* μ_n *such that* $\mu_n \to 0$ *and* $\mu_n\sqrt{n} \to \infty$*; class* $\mathcal{P}$*, consisting of distributions satisfying Assumptions 1-5. For any* $\epsilon > 0$ *there exist* $R > 0$ *such that for all sufficiently large* n

$$\sup_{P\in\mathcal{P}} P\left(\mu_n\sqrt{n}\,\|\underline{\theta}(\mu_n, \mathbb{P}_n) - \underline{\theta}(\mu_n, P)\| \geq R\right) \leq \epsilon, \tag{15}$$

$$\sup_{P\in\mathcal{P}} P\left(\sqrt{n}\,\|\underline{\lambda}(\mu_n, \mathbb{P}_n) - \underline{\lambda}(\mu_n, P)\| \geq R\right) \leq \epsilon, \tag{16}$$

[11] See, for example, Golishnikov and Izmailov (2006).

[12] One interesting recent example is Shi and Shum (2016).

Moreover,

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}} P\left(\left|\underline{\sigma}\left(\mu_n,\mathbb{P}_n\right)-\underline{\sigma}\left(\mu_n,P\right)\right|\geq\epsilon\right)=0, \tag{17}$$

$$\lim_{n\to\infty}\sup_{P\in\mathcal{P}}\rho_n(P)=0. \tag{18}$$

Proof. See Appendix 8.6. □

Result (18) follows from a Bahadur representation derived in Lemma 10 in Appendix,

$$\underline{v}\left(\mu_n,\mathbb{P}_n\right)=\underline{v}\left(\mu_n,P\right)+\frac{1}{n}\sum_{i=1}^{n}\underline{\lambda}\left(\mu_n,P\right)'g(w_i,\underline{\theta}\left(\mu_n,P\right))+O_p(\frac{1}{\mu_n n}). \tag{19}$$

This representation provides a justification for the convergence of the regularized estimator of the support function of $\Theta\left(P\right)$ to a Gaussian process used later in Section 3.1. Another useful implication is the applicability of the multiplier bootstrap.

2.4 Point-wise valid confidence sets.

In order to construct confidence sets we need to study how far $\underline{v}\left(\mu_n,P\right)$is from the parameter of interest, $\underline{v}\left(P\right)$. Lemma 8 shows that for any given distribution P there exist some $\bar{\mu}(P)$ such that for any $\mu<\bar{\mu}(P)$ the regularized solution is constant, $\underline{\theta}\left(\mu,P\right)=\underline{\theta}\left(0^+,P\right)$. It implies that for sufficiently large n the only source of the bias is $\mu_n\left\|\underline{\theta}\left(0^+,P\right)\right\|$ which has order of magnitude larger than $1/\sqrt{n}$. This bias can be corrected as follows,

$$\underline{\hat{v}}_n\triangleq\underline{v}\left(\mu_n,\mathbb{P}_n\right)-\mu_n\left\|\underline{\theta}\left(\kappa_n,\mathbb{P}_n\right)\right\|^2, \tag{20}$$

If κ_n converges to zero slower than μ_n, then this bias correction does not affect the asymptotic variance of $\underline{\hat{v}}_n$ and $\underline{v}\left(\mu_n,\mathbb{P}_n\right)$ have the same asymptotic variance.

Using the bias corrected estimators $\underline{\hat{v}}_n$ and their variances, I construct Delta-method confidence sets:

$$\begin{cases}\mathrm{CI}^L_{\alpha,n} &= \left[\underline{\hat{v}}_n-z_{1-\alpha}n^{-1/2}\underline{\sigma}\left(\mu_n,\mathbb{P}_n\right),\infty\right),\\ \mathrm{CI}_{\alpha,n} &= \left[\underline{\hat{v}}_n-z_{1-\alpha}n^{-1/2}\underline{\sigma}\left(\mu_n,\mathbb{P}_n\right);\hat{\bar{v}}_n+z_{1-\alpha}n^{-1/2}\bar{\sigma}\left(\mu_n;\mathbb{P}_n\right)\right],\\ \mathrm{CI}^B_{\alpha,n} &= \mathrm{CI}_{\alpha/2,n},\end{cases} \tag{21}$$

where $z_{1-\alpha}$ is $1-\alpha$ quantile of the standard Gaussian distribution. The estimators corresponding to the upper bound,$\hat{\underline{v}}_n$ and $\bar{\sigma}\left(\mu_n;\mathbb{P}_n\right)$ are defined analogously.

Theorem 2. *Suppose that Assumptions 1–5 hold and that in addition* $0<\alpha<1/2$ *,* μ_n *and* κ_n *are such that* $\kappa_n\to 0$*,* $\mu_n/\kappa_n\to 0$ *and* $\mu_n\sqrt{n}\to\infty$*. Moreover, suppose that* $\lim_{n\to\infty}\underline{\sigma}^2\left(\mu_n,P\right)>0$ *and* $\lim_{n\to\infty}\bar{\sigma}^2\left(\mu_n,P\right)>0$*. Then,*

$$\lim_{n\to\infty}P\left(\mathcal{S}\left(P\right)\subset CI^L_{\alpha,n}\right)=\liminf_{n\to\infty}\inf_{\theta\in\Theta(P)}P\left(\theta_1\in CI^L_{\alpha,n}\right) = 1-\alpha,$$

$$\lim_{n\to\infty}P\left(\mathcal{S}\left(P\right)\subset CI^B_{\alpha,n}\right)\geq 1-\alpha,\ \liminf_{n\to\infty}\inf_{\theta\in\Theta(P)}P\left(\theta_1\in CI^B_{\alpha,n}\right) \geq 1-\alpha.$$

If the model has no equality constraints, i.e. if $p = 0$, then

$$\liminf_{n\to\infty} \inf_{\theta\in\Theta(P)} P\left(\theta_1 \in CI_{\alpha,n}\right) \;=\; 1-\alpha. \tag{22}$$

Proof. See Appendix 8.7. □

The component θ_1 is point identified if there are equality constraints in the model that are orthogonal to e_1. If $p > 0$, I recommend the Bonferroni–type confidence set $\mathrm{CI}^B_{\alpha,n}$ which remains valid under point identification. The shorter $\mathrm{CI}_{\alpha,n}$ proposed by Imbens and Manski (2004) is valid if θ_1 is not point–identified. Under Assumption 3, $p = 0$ implies that $\mathcal{S}(P)$ is not a singleton.

The infeasible optimal choice of the tuning parameters is to set $\kappa_n = \bar{\mu}(P)$ and make μ_n go to zero at the rate that would balance the higher order bias and variance, which is $\mu_n = \kappa_n^{1/2} n^{-1/4}$. This choice is infeasible however since $\bar{\mu}(P)$ is discontinuous function of $\mathbb{E}_P W$ and hence cannot be consistently estimated. A practical choice is to let κ_n go to zero at a slow rate. A specific choice of the tuning parameters is further discussed in Section 5.

Note that if the limiting variance $\lim_{n\to\infty} \underline{\sigma}^2(\mu_n, P) = 0$, then the Gaussian limiting distribution does not provide a correct coverage probability which in this case is governed by the higher order terms. In the next subsection I will discuss how to relax this requirement.

2.5 Uniform confidence sets

Theorem 2 provides asymptotic coverage probability for a given DGP with measure P. The size of the sample required to achieve the nominal coverage of $1-\alpha$ with a given precision in this result can depend on P. In fact it is possible to construct an example where sequence of measures P_n that meet the assumptions of Theorem 2 but such that

$$\sqrt{n}\mu_n(\|\underline{\theta}(\mu_n, P_n)\| - \|\underline{\theta}(\kappa_n, P_n)\|) \to +\infty.$$

In other words, there a examples of DGP with a measure P and some $\epsilon > 0$ such that for any n it is possible to find a measure Q in a neighborhood of P with

$$Q\left(\mathcal{S}(P) \subset \mathrm{CI}^L_{\alpha,n}\right) < 1-\alpha-\epsilon.$$

In practical terms it means that the large sample theory with a fixed P may provide a poor approximation for the true coverage probability.

This feature of the confidence sets from the Theorem 2 should not come as a surprise. Parameter of interest, $\underline{v}(P)$ is a non-differentiable function of $\mathbb{E}_P W$ which in turn is a parameter of a locally asymptotically normal model. By the impossibility theorem of Hirano and Porter (2012) $\underline{v}(P)$ does not have a locally unbiased estimator.

The estimator $\hat{\underline{v}}_n$ is biased inwards of $\mathcal{S}(P)$. The inward bias reduces the asymptotic coverage probability of the sets in Theorem 2. It is possible to construct alternative estimators with outward bias.

The key idea is based on the following inequality. We can take any $\theta_n^* \in \underline{\theta}(0, P)$ and notice that

$$\underline{\theta}_1(\mu_n, P) + \mu_n \|\underline{\theta}(\mu_n, P)\|^2 \le \underline{v}(P) + \mu_n \|\theta^*\|^2. \tag{23}$$

This inequality implies

$$\underline{v}(\mu_n, P) - \mu_n \beta^2 \le \underline{v}(P) \tag{24}$$

for any $\beta^2 > \|\theta^*\|^2$. In Section 2.3 we showed that the first component on the left of the inequality (24) has a estimator with a uniform asymptotic normal distribution. One candidate for second component, β, can be consistently estimated as a norm of a basic solution, i.e. d constraint are active, with a largest norm (clearly $\underline{\theta}(0,P)$ contains at least one basic solution).

Basic solutions corresponding to an active set of constraints $\mathcal{J}$ are defined as follows,

$$\theta^{\mathcal{J}}(P) \triangleq (\mathbb{J}A_P)^{-1}(\mathbb{J}b_P).$$

We can construct a uniformly consistent super-set estimator for all basic solutions using

$$\hat{\mathcal{B}}_n \triangleq \left\{\mathcal{J} | \hat{\theta}_1^{\mathcal{J}} \leq e_1' \underline{\theta}\left(\mu_n, \mathbb{P}_n\right) + \mu_n \text{ and } \forall j \in \mathcal{J}_k^{ineq}, e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \leq \mu_n\right\}$$

as the following theorem shows.

Theorem 3. *Let $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$. Then for any $\epsilon > 0$ there exist n such that*

$$\inf_{P\in\mathcal{P}} P\{\mathcal{B}(P) \subset \hat{\mathcal{B}}_n\} \geq 1 - \epsilon.$$

Proof. See Appendix 8.8 □

Using this theorem, we can propose a new estimator,

$$\hat{\underline{v}}_n \triangleq \underline{v}\left(\mu_n, \mathbb{P}_n\right) - \mu_n \beta^2\left(\mathbb{P}_n\right), \tag{25}$$

where $\beta\left(\mathbb{P}_n\right) = \max_{\mathcal{J}\in\hat{\mathcal{B}}_n} \left\|\theta^{\hat{\mathcal{J}}}\right\|$. The asymptotic bias in this case is always non-positive. The estimator $\hat{\underline{v}}_n$ is asymptotically unbiased for any fixed P with a singleton $\underline{\theta}(0,P)$, i.e. the case with differentiable $\underline{v}\left(P\right)$.

Before I introduce the uniformly valid confidence sets, I would like to address the difficulty resulting from the degenerate Gaussian distribution. For that it is sufficient to consider a regularized estimator of the asymptotic variance,

$$\hat{\underline{\sigma}}_n \triangleq \max\{\underline{\sigma}\left(\mu_n, \mathbb{P}_n\right), \sigma_0\},$$

where σ_0 is some small positive number.

Let $\widetilde{\text{CI}}^L_{\alpha,n}$ and $\widetilde{\text{CI}}^B_{\alpha,n}$ be the confidence sets defined in (21) with $\hat{\underline{v}}_n$ and $\underline{\sigma}\left(\mu_n, \mathbb{P}_n\right)$ being replaced by $\hat{\underline{v}}_n$ and $\hat{\underline{\sigma}}_n$, correspondingly. As before, let $\mathcal{P}$ contain all measures P that satisfy Assumptions 1-5 with some uniform constants $\varepsilon, \bar{M}, \epsilon, \eta$.

Theorem 4. *Suppose that Assumption 4 holds. In addition, suppose that $0 < \alpha < 1/2$, $\mu_n \to 0$ and $\mu_n\sqrt{n} \to \infty$. Then the following results hold,*

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(\mathcal{S}\left(P\right) \subset \widetilde{CI}^L_{\alpha,n}\right) = \liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta(P)} P\left(\theta_1 \in \widetilde{CI}^L_{\alpha,n}\right) \geq 1 - \alpha,$$

$$\liminf_{n\to\infty} \inf_{P\in\mathcal{P}} P\left(\mathcal{S}\left(P\right) \subset \widetilde{CI}^B_{\alpha,n}\right) \geq 1 - \alpha, \ \liminf_{n\to\infty} \inf_{P\in\mathcal{P}} \inf_{\theta\in\Theta(P)} P\left(\theta_1 \in CI^B_{\alpha,n}\right) \geq 1 - \alpha.$$

Proof. See Appendix 8.9. □

Note that the asymptotic coverage size is exactly equal to $1 - \alpha$ in the regular cases.

3 Extensions

In this section I outline two important extensions of the main results.

3.1 Subvector inference

It is trivial to extend the analysis to

$$\underline{v}(P;a) = \min_{\theta \in \Theta(P)} a'\theta$$

for any $a \in R^d$ with $\|a\| = 1$. Indeed, Assumptions 1-5 are invariant with respect to orthogonal transformations of the coordinates, i.e. they are satisfied for the following program (with $\tilde{\theta} = U'\theta$, $\tilde{A_P} = A_P U$ and $a' = e_1'U$ for any orthogonal matrix U)

$$\underline{v}(P;a) = \min_{\theta \in \mathbb{R}^d} e_1'\tilde{\theta} \tag{26}$$

$$\text{s.t. } \begin{cases} \tilde{A_P}\tilde{\theta} = b_P, & j \in \mathcal{J}_k^{eq}, \\ \tilde{A_P}\tilde{\theta} \leq b_P, & j \in \mathcal{J}_k^{ineq}, \end{cases} \tag{27}$$

One can think about $\tilde{A_P}$ as $A_{\tilde{P}}$. The set of measures $\mathcal{P}$ from Section 2.5 includes $\tilde{P}$ corresponding to all orthogonal transformations of A_P.

The identified set $\Theta(P)$ is convex so any projectionf of it can be characterized using support functions. One can construct a joint confidence set for $\Theta(P)$ as follows. For any set of directions $\mathcal{A} \subset R^d$ take

$$\text{CS}_{\alpha,n}^{\mathcal{A}} = \{\theta | a \in \mathcal{A}, a'\theta \leq -\underline{v}(\mu_n, \mathbb{P}_n; -a) + \mu_n \beta^2(\mathbb{P}_n; a) + c_{1-\alpha} n^{-1/2} \max\{\underline{\sigma}(\mu_n, \mathbb{P}_n; -a), \sigma_0\}\},$$

where $c_{1-\alpha}$ is $1-\alpha$ quantile of the maximum of the corresponding Gaussian process that can be estimated using multiplier bootstrap.

By appropriately choosing the set of directions $\mathcal{A}$ we can construct joint confidence sets for projections of $\Theta(P)$ on any subvectors θ. If $\mathcal{A}$ has finitely many elements, $\text{CS}_{\alpha,n}^{\mathcal{A}}$ is a polygon. So we can plot it directly without performing test inversion as in the one-dimensional case.

The confidence set $\tilde{\text{CI}}_{\alpha,n}^{B}$ is a particular case of $\text{CS}_{\alpha,n}^{\mathcal{A}}$ corresponding to $\mathcal{A} = \{e_1, -e_1\}$ and the Bonferroni estimate of $c_{1-\alpha}$.

3.2 Dealing with overidentification

Another interesting extension concerns the case when Assumptions 1-3 are violated for $\Theta(P)$, but it can be represented as an intersection of sets $s = 1, ..., L$, $\Theta^s(P)$ which satisfy these assumptions. Then

$$\underline{v}(P) = \min_{\theta \in \Theta(P)} e_1'\theta \geq \max_{s=1,...,L} (\min_{\theta \in \Theta^s(P)} e_1'\theta) \geq \max_{s=1,...,L} (\underline{v}^s(\mu_n, P) - \mu_n(\beta^{(s)}(P))^2) \tag{28}$$

If we define then $\underline{v}(P) = +\infty$ in the case $\Theta(P) =$, then the bounds (28) are trivially valid too.

It is possible to make the first bound in (28) sharp if we consider all subsets of $d-p$ inequality restrictions (besides the ones that define Θ). Let $\underline{v}^{(s)}$ denote the solution to the subproblem s. Since every subproblem satisfies Assumptions 1- 3, we get

$$\sqrt{n}\begin{pmatrix} \hat{\underline{v}}^{(1)} - \underline{v}^{(1)} \\ \cdots \\ \hat{\underline{v}}^{(L)} - \underline{v}^{(L)} \end{pmatrix} \rightsquigarrow N(0, \Omega).$$

By the envelope theorem

$$\Omega_{ij} = \mathbb{E}_P[\lambda'^{(i)} g(W, \underline{\theta}^{(i)}) g(W, \underline{\theta}^{(j)})' \lambda^{(j)}]$$

Now we can use the following representation to reduce it to the original problem

$$\underline{v} = \max_{\sum_{s=1}^{L} \gamma_s = 1, \gamma_s \geq 0} \gamma_s \underline{v}^{(s)}. \tag{29}$$

The solution γ to (29) is not unique in general. Program (29) satisfies the Assumptions 1-3. In order to restore the asymptotic normality of the corresponding estimator we need to regularize it,

$$\begin{aligned} -\min_{\gamma} \quad & -\sum_{s=1}^{L} \gamma_j \underline{v}^{(j)} + \mu_n \|\gamma\|^2 \\ \text{s.t.} \quad & \sum_{s=1}^{L} \gamma_j = 1, \gamma_j \geq 0. \end{aligned} \tag{30}$$

The estimator of the value of this program also has a Bahadur representation which can be used to derive an estimator for the variance,

$$\text{sVar}(\sum_{s=1}^{L} \gamma_s \lambda'^{(s)} g(w_i, \underline{\theta}^{(s)})).$$

The bias correction like in Section 2.4 would in this case make the corresponding confidence sets longer and preserve asymptotic coverage probability of at least $1 - \alpha$ uniformly in $P \in \cap_{s=1,L} \mathcal{P}^s$.

The case with no-overidentification would result in zero bias uniformly. Case with over-identification can result in overcoverage along some DGP sequences but is not conservative for a fixed DGP if bound (28) is sharp.

4 Discussion

4.1 Scope

There are at least two recent papers, BCS and KMS, to construct confidence sets that provide explicit classes of DGP with uniform asymptotic coverage probability. Both methods are applicable in non-linear moment inequality models. They however are more restrictive in some other dimensions and there are examples of affine moment inequality models that fall outside of their scope.

Both methods are using standardized moment conditions and hence restrict variance of the moment conditions to be strictly larger that some small constant (Definition 4.2.ii in BCS and Assumption 4.1 b (iii) in KMS). Moreover, the standardized moment conditions has to be differentiable (Assumption A.3.c in BCS and 4.4.i in KMS). Both Assumptions are also present in the pioneering AS paper. The following example illustrates violation of these assumptions in an affine

models.

Example 2 (Example 1 continued)**.** Suppose that for some support point z_0 with a positive mass the lower bound on the outcome is deterministic, i.e. $\mathbb{E}_P(\underline{Y}|Z=z_0)=y_0$ and $\mathrm{Var}_P(\underline{Y}|Z=z_0)=0$. The corresponding moment inequality condition is

$$\mathbb{E}_P g(Y,Z,\theta)=\mathbb{E}_P\left[(\underline{Y}-z_0'\theta)\,1\{Z=z_0\}\right]\leq 0. \tag{31}$$

The standard deviation of the moment condition is $|z_0'\theta-y_0|\sqrt{p_z(1-p_z)}$, where $p_z=\mathbb{E}_P 1\{z=z_0\}$. It is equal to zero for any solution of the linear equation $z_0'\theta=y_0$. These values of θ make the constraint (31) binding, which is exactly the case when this constraints determines $\underline{v}(P)$. This moment condition is also non-differentiable at these points.

$$\partial_\theta g(y_0,z_0,\theta)=\mathrm{sign}(y_0-z_0'\theta)\sqrt{p_z/(1-p_z)}.\square$$

I also avoid imposing any restrictions on the correlation matrix of the moment conditions which is present in KMS (Assumption 4.3) and the polynomial minorant condition (Assumption A.3.a in BCS, present in CHT, avoided in KMS).

I do impose an explicit assumptions that guarantee the LICQ which is not present in KMS or BCS, but I only require them to hold in subproblems, as noted in Section 3.2. I would like to note here that the M-step in KMS uses standard Newton optimization routines that can find all the stationary (KKT) points. If there is no constraint qualification, however, KKT conditions are no longer providing the nesessary conditions for the optimum, which in this case is a John-Fritz conditions.[13] As a result, E-A-M algorithm considered in KMS is not guaranteed to converge to the global optimum without some constraint qualification.

4.2 Computation properties

4.2.1 Fast convergence to a minimum

The existing uniform methods of AS, BCS and KMS are based on standardized moment condtions that are non-convex even if the original inequlaities are affine in θ. Example 2 in the previous section illustrate this feature. The estimators $\underline{\theta}(\mu_n,\mathbb{P}_n)$ is a solutions to a strictly convex quadratic programs for any affine moment inequlaity model. For convex programs the set Karush–Kuhn–Tucker (KKT) conditions[14] provide necessary and sufficient conditions for the global optimum. Moreover, convex quadratic programs can be solved using an interior point algorithms with a polynomial rate of convergence.[15] This strict convexity gives a dramatically faster rates of obtaining the optimum than the ones used in BCS and KMS. These methods are based on non-convex constraint optimization problems, which are NP-hard. Section 5 compares computational time in specific examples.

4.2.2 Uniqueness of a global optimum

KKT system for strictly convex optimization problems has a unique solution. The number of KKT points of the optimization problems in the KMS, BCS and AS procedures in affine moment

[13] See Section 5.2.2 in BS(2000)

[14] See Lemma 3 in Appendix

[15] See, for example, Ye and Tse (1989).

inequality models can be large and typically grows exponentially with the dimension d and number of inequalities k. The following example illustrates this point.

Example 3. Consider a set of moment inequalities with coefficients that have expectation

$$\mathbb{E}_P W = \begin{pmatrix} -I_d & -\iota \\ I_d & -\iota \end{pmatrix}.$$

Suppose that components of W are independent and have the same variance s^2. $\Theta(P)$ is a box $[-1,1]^d$. The standardized moment conditions take the form

$$\frac{\pm\theta_j + 1}{s\sqrt{1+\|\theta\|^2}} \leq 0,\ j = 1,\ldots,d. \tag{32}$$

The KMS procedure adds slack $c(\theta)$ to the right hand side of every standardized moment inequality. Consider , for example, $j = 1$,

$$\theta_1 \geq 1 - c(\theta)\, s\sqrt{1+\|\theta\|^2}. \tag{33}$$

The slack function $c(\theta)$ is computed using a resampling on a grid of points. Assume, for simplicity, that $c(\theta)$ is a constant, for example, provided by the Bonferroni approach. Figure 1 shows the identified set and the corresponding expansion with $c(\theta) = const$. The optimization domain of the E-A-M algoritm in KMS is similar the non-convex set on right of Figure 1. Every vertex of the $[-1,1]^d$ with $\theta_1 = -1$ corresponds to an isolated local minimum of the optimization procedure in KMS. Correspondingly, the number of local minima grows exponentially with the dimension d. For example, the number of local minima for $d = 10$ is 512. The growth in the number of local optima is even faster in models with more than 2 inequalities per coordinate. □

Multiplicity of KKT points makes the procedures of KMS, AS and BCS both computationally costly does not provide guarantees of convergence to a global optimum for large d.

4.2.3 Multiplier bootstrap

The proposed estimators of the regularized support functions have a Bahadur representation with explicit influence functions. One can use this property to justify multiplier bootstrap for inference on support of the identified set. The main advantage of this approach is that it allows one to solve the mathematical programs only once. For example, FH and KMS solve mathematical programs repeatedly for every bootstrap sample. The multiplier bootstrap approach is particularly appealing in subvector inference on more that one component as discussed in Section 3.1, since one has repeat computations for various directions a.

4.2.4 Implementation

The point-wise CIs in (21) can be computed using any Newton type optimization software that provides accurate Lagrange multipliers. I use *fmincon* function of MATLAB software. I recommend using the *'active set'* or *'SQP'* options since the *'interior point'* solver does not provides accurate Lagrange multipliers. The estimator $\hat{\beta}_n$ in the uniformly valid CIs described in Theorem 4 requires a number of operations that is proportional to the number of all combinations of d active constraints. This number grows exponentially with d so it is desirable to develop an alternative estimator of the upper bound on β that is based on linear programming.

The extended procedure outlined in Section 3.2 requires finding subproblems that make the bound (28) sharp. This can be achieved if Assumption 2 is satisfied for the original program and one considers all subproblems with exactly d moment conditions. Such an exhaustive approach may require considerable computational resources if d is large. Potentially, one could use a moment selection procedure that would restrict attention to combinations of moments inequalities that are close to be binding at point $\underline{\theta}(\mu_n, \mathbb{P}_n)$. Such an extension is beyond the scope of this paper.

5 Monte Carlo

5.1 Choice of the tuning parameters

In the Monte Carlo exercise I used the following tuning parameters $\mu_n = \hat{\mu}_1 \sqrt{\frac{\log\log n}{n}}$ and $\kappa_n = \hat{\mu}_1 \sqrt{\frac{\log n}{n}}$ with

$$\hat{\mu}_1 = \frac{\sqrt{\underline{\lambda}'(0, \mathbb{P}_n)\left(X - \frac{1}{n}\sum_{i=1}^n X\right)\left(X - \frac{1}{n}\sum_{i=1}^n X\right)'\underline{\lambda}(0, \mathbb{P}_n)}}{\max\left\{\beta(\mathbb{P}_n), 1\right\}}.$$

The choice of numerator in $\hat{\mu}_1$ is motivated by the formula for the higher order terms in the stochastic expansion of the estimator $\underline{v}(\mu_n, \mathbb{P}_n)$. The denominator makes the bias $\|\underline{\theta}(\mu_n, P_n)\|^2 - \|\underline{\theta}(\kappa_n, P_n)\|^2$, which may be non–zero for drifting DGP-2, bounded by 1. This choice gives good finite sample coverage properties for both $\mathrm{CI}_{\alpha,n}$ and $\widetilde{\mathrm{CI}}_{\alpha,n}$. The theory for optimal choice of the tuning parameter is beyond the scope of this paper.

5.2 Two-dimensional designs

I consider four different MC designs. DGP 1-3 have 4 moment inequalities with coefficients that have the following expectation:

$$\mathbb{E}_P X = \begin{pmatrix} -\cos\left(\omega_1 \frac{\pi}{2}\right) & -\sin\left(\omega_1 \frac{\pi}{2}\right) \\ \cos\left(\omega_1 \frac{\pi}{2}\right) & \sin\left(\omega_1 \frac{\pi}{2}\right) \\ -\cos\left(\omega_2 \frac{\pi}{2}\right) & -\sin\left(\omega_2 \frac{\pi}{2}\right) \\ \cos\left(\omega_2 \frac{\pi}{2}\right) & \sin\left(\omega_2 \frac{\pi}{2}\right) \end{pmatrix}, \ \mathbb{E}_P w = \begin{pmatrix} -\cos\left(\omega_1 \frac{\pi}{2}\right)\zeta_1 - \sin\left(\omega_1 \frac{\pi}{2}\right)\zeta_2 \\ \cos\left(\omega_1 \frac{\pi}{2}\right)\psi_1 + \sin\left(\omega_1 \frac{\pi}{2}\right)\psi_2 \\ -\cos\left(\omega_2 \frac{\pi}{2}\right)\psi_1 - \sin\left(\omega_2 \frac{\pi}{2}\right)\psi_2 \\ \cos\left(\omega_2 \frac{\pi}{2}\right)\zeta_1 + \sin\left(\omega_2 \frac{\pi}{2}\right)\zeta_2 \end{pmatrix}.$$

The expectations are parametrized to guarantee $\underline{v}(P) = \zeta_1, \bar{v}(P) = \psi_1$. DGP-4 has the first two inequality conditions replaced by a single moment equality condition

$$\cos\left(\omega_1 \frac{\pi}{2}\right)\theta_1 + \sin\left(\omega_1 \frac{\pi}{2}\right)\theta_2 = \cos\left(\omega_1 \frac{\pi}{2}\right)\zeta_1 + \sin\left(\omega_1 \frac{\pi}{2}\right)\zeta_2.$$

The components of $W = (X, w)$ are independent Gaussian random variables with variance σ_0^2. The parametrization used in the MC experiments is summarized in Table 1. The corresponding identified sets for DGP 1-3 are shown on Figure 2 in Appendix 8.1.

Table 2 summarizes the results of 2000 MC simulations for sample sizes $10^2, 10^3, 10^4, 10^5$. $\mathrm{CI}_{\alpha,n}$ has tendency to undercover in designs with unique optimum (DGP-1,2,4) and exceed the nominal size in the presence of multiple optima (DGP-3). Both tendencies become less prominent as sample

Table 1: Parameter values for DGP 1-4

	ω_1	ω_2	σ_0	p	$\underline{v}(P)$	$\bar{v}(P)$
DGP-1	0.25	1.5	0.1	0	-1	1
DGP-2	$n^{-1/2}$	1	0.1	0	-1	1
DGP-3	0	1	0.1	0	-1	1
DGP-4	$n^{-1/2}$	1	0.1	1	-1	$-1+2\tan\left(\frac{\pi}{2}n^{-1/2}\right)$

size grows.

$\widetilde{\text{CI}}_{\alpha,n}$ has only marginally higher coverage probability than $\text{CI}_{\alpha,n}$ in DGP-1,2,4. The exact coverage of $\widetilde{\text{CI}}_{\alpha,n}$ for these DGPs is achieved since by design $\lim_{n\to\infty}\|\underline{\theta}(\mu_n, P_n)\|^2 = \lim_{n\to\infty}\beta(P_n)$. The DGP-3, in contrast, does not meet this condition. As a result, $\widetilde{\text{CI}}_{\alpha,n}$ converges to $\mathcal{S}(P)$ at rate $\sqrt{n^{-1}\log\log n}$ which results in the coverage probability close to 100%.

5.3 Length comparison with existing approaches

One can assess the conservativeness of $\widetilde{\text{CI}}_{\alpha,n}$ by comparing the excess average length of $\widetilde{\text{CI}}_{\alpha,n}$ with the projection of the AS confidence set as the dimension of the problem grows. DGP 5 is described in Example 5 and coincides with DGP-3 if $d = 2$. I change the standard deviation with dimension by formula $\sigma_0 = 0.1/\sqrt{1+d}$ to keep the variance of the moment conditions at every vertex of $[-1,1]^d$ constant. Figure 3 shows the excess average length for $\widetilde{\text{CI}}_{\alpha,n}$ and $\text{CI}^{AS}_{\alpha,n}$ for $n = 1000$. For all experiments (except for $d = 4$), the average length of $\widetilde{\text{CI}}_{\alpha,n}$ is shorter than $\text{CI}^{AS}_{\alpha,n}$. As in DGP-3, $\widetilde{\text{CI}}_{\alpha,n}$ converges to $\mathcal{S}(P)$ at rate $\sqrt{n^{-1}\log\log n}$ while $\text{CI}^{AS}_{\alpha,n}$ has a faster rate $n^{-1/2}$. Nevertheless, Figure 3 suggests that it may be preferable to use $\widetilde{\text{CI}}_{\alpha,n}$ in samples of small and moderate size.

5.4 Time comparison

Figure 3 shows the computational time for $\widetilde{\text{CI}}_{\alpha,n}$ and the AS procedure using the implementation of KMS[16] in Example 3. The average computation time for $\widetilde{\text{CI}}_{\alpha,n}$ is almost insensitive to the dimension d. It takes 1.5 seconds to compute $\widetilde{\text{CI}}_{\alpha,n}$ for $d = 15$. The computational time for the AS procedure increases by approximately 30% with every additional dimension and takes 630 seconds to compute the CI for $d = 15$. The KMS procedure by construction is more computationally intensive than the AS procedure. The KMS procedure implemented in Matlab with precompiled solvers takes 560 sec for $d = 8$.[17] With estimated growth rate of 30% per dimension KMS procedure with precompiled code would take more than hour to compute a CI for $d = 15$.

6 Outline for an Empirical Application

This section contains a potential empirical application of the proposed procedures.

[16]The code is available on https://molinari.economics.cornell.edu/programs.html. Note that this implementation of AS procedure requires additional constraint qualification assumption, which was not made explicitly in KMS.

[17]These numbers correspond to a different DGP. The available code of KMS achieves this speed using a precompiled optimization routine for a specific model. The build-in linear programming solvers in Matlab cannot achieve this speed.

Example 4 (The returns to schooling)**.** Trostel et al. (2002) study economic returns to schooling for 28 countries using International Social Survey Programme data (ISSP), 1985–1995. They estimate a conventional Mincer (1974) model of earnings (the human capital earnings function), which has log wage determined by years of schooling, age, experience, and other explanatory variables:

$$\mathbb{E}\left[Y|Z\right] = \theta' Z, \tag{34}$$

where Y is the log of hourly wages, Z_1 is years of schooling and the other components of Z is a vector of observed exogenous explanatory variables including, where appropriate, country and year fixed effects. The component θ_1 is interpreted as the rate of returns to schooling; it is equal to the percentage change in wages due to an additional year of schooling. Their explanatory variables Z include year dummies, union status, marital status, age and age squared and, in the case of the aggregate equation, country-year dummies. Exact measures of Y are not available for some countries (including the USA); only income bracket data $\left[\underline{Y}, \overline{Y}\right]$ is available for those countries. Trostel et al. (2002) use a conventional technique to deal with this problem – they replace the interval data with the corresponding midpoints and estimate (34) using OLS. This technique is valid only under the unreasonably strong condition

$$\mathbb{E}_P\left[\left(Y - 0.5(\underline{Y} + \overline{Y})\right) Z\right] = 0. \tag{35}$$

If condition (35) is violated then the OLS estimator for the effect of schooling is inconsistent.

The interval outcome model from Example 1 can provide estimates of the marginal identified set for returns to schooling without assumption (35). The conventional estimates based on the midpoint approach converge to one of the elements in $\mathcal{S}(P)$. All the explanatory variables are discrete, so the existing approach of Bontemps et al. (2012) is not applicable. □

7 Conclusion

This paper shows that the regularization approach provides a fast way to construct both pointwise and uniform confidence sets for θ_1 that can be shorter than those in the existing literature. Moreover, my CIs remain valid in some situations where the existing procedures cannot be used. Monte Carlo simulations show that the proposed CIs have good finite sample coverage properties. The computational benefits of the new approach are particularly prominent if the dimension of θ is large. My approach is attractive in applications like linear model with interval-valued outcome variable and a large number of regressors.

References

Andrews, D. W. and X. Shi (2013): "Inference based on conditional moment inequalities," *Econometrica*, 81, 609–666. 4

Andrews, D. W. and G. Soares (2010): "Inference for Parameters Defined by Moment Inequalities Using Generalized Moment Selection," *Econometrica*, 78, 119–157. 1

Beresteanu, A. and F. Molinari (2008): "Asymptotic properties for a class of partially identified models," *Econometrica*, 763–814. 2

Berge, C. (1963): *Topological Spaces: including a treatment of multi-valued functions, vector spaces, and convexity*, Courier Corporation. 5

Bonnans, J. F. and A. Shapiro (2000): *Perturbation Analysis of Optimization Problems*, Springer Science & Business Media. 4

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set identified linear models," *Econometrica*, 80, 1129–1155. 5, 17

Boyd, S. and L. Vandenberghe (2004): *Convex optimization*, Cambridge university press. 3

Bugni, F., I. Canay, and X. Shi (2016): "Inference for functions of partially identified parameters in moment inequality models," Tech. rep., cemmap working paper, Centre for Microdata Methods and Practice. 1, 2

Chernozhukov, V., H. Hong, and E. Tamer (2007): "Estimation and confidence regions for parameter sets in econometric models," *Econometrica*, 75, 1243–1284. 1, 4

Fischer, A. (1992): "A special Newton-type optimization method," *Optimization*, 24, 269–284. 5

Freyberger, J. and J. L. Horowitz (2015): "Identification and shape restrictions in nonparametric instrumental variables estimation," *Journal of Econometrics*, 189, 41 – 53. 1, 2

Gafarov, B., M. Meier, and J.-L. Montiel-Olea (2015): "Delta-Method inference for a class of set-identified SVARs," Tech. rep., Working paper, New York University. 2, 6

Golishnikov, M. and A. F. Izmailov (2006): "Newton-type methods for constrained optimization with nonregular constraints," *Computational Mathematics and Mathematical Physics*, 46, 1299–1319. 7

Hirano, K. and J. R. Porter (2012): "Impossibility results for nondifferentiable functionals," *Econometrica*, 1769–1790. 9

Imbens, G. W. and C. F. Manski (2004): "Confidence Intervals for Partially Identified Parameters," *Econometrica*, 72, 1845–1857. 1, 9, 10

Kaido, H., F. Molinari, and J. Stoye (2015): "Inference for projections of identified sets," *manuscript*. 1, 2

Kaido, H. and A. Santos (2014): "Asymptotically Efficient Estimation of Models Defined by Convex Moment Inequalities," *Econometrica*, 82, 387–413. 2

Manski, C. F. (2003): *Partial identification of probability distributions*, Springer Science & Business Media. 1

Manski, C. F. and J. V. Pepper (2000): "Monotone instrumental variables: With an application to the returns to schooling," *Econometrica*, 68, 997–1010. 1

Manski, C. F. and E. Tamer (2002): "Inference on regressions with interval data on a regressor or outcome," *Econometrica*, 70, 519–546. 1

Mincer, J. (1974): "Schooling, Experience, and Earnings. Human Behavior & Social Institutions No. 2." . 17

Ok, E. A. (2007): *Real analysis with economic applications*, vol. 10, Princeton University Press. 5

Shapiro, A. (1993): "Asymptotic behavior of optimal solutions in stochastic programming," *Mathematics of Operations Research*, 18, 829–845. 5

Shi, X. and M. Shum (2016): "Estimating semi-parametric panel multinomial choice models using cyclic monotonicity," *Available at SSRN 2748647*. 7

Stewart, M. B. (1983): "On least squares estimation when the dependent variable is grouped," *The Review of Economic Studies*, 50, 737–753. 1

Trostel, P., I. Walker, and P. Woolley (2002): "Estimates of the economic return to schooling for 28 countries," *Labour economics*, 9, 1–16. 1, 17

van der Vaart, A. and J. Wellner (1996): *Weak Convergence and Empirical Processes: With Applications to Statistics*, Springer Science & Business Media. 7, 8, 9

Wachsmuth, G. (2013): "On LICQ and the uniqueness of Lagrange multipliers," *Operations Research Letters*, 41, 78–80. 6

Ye, Y. and E. Tse (1989): "An extension of Karmarkar's projective algorithm for convex quadratic programming," *Mathematical programming*, 44, 157–179. 13

Yurinskii, V. V. (1978): "On the error of the Gaussian approximation for convolutions," *Theory of Probability & Its Applications*, 22, 236–247. 7

Appendix

8 Proofs

8.1 Figures

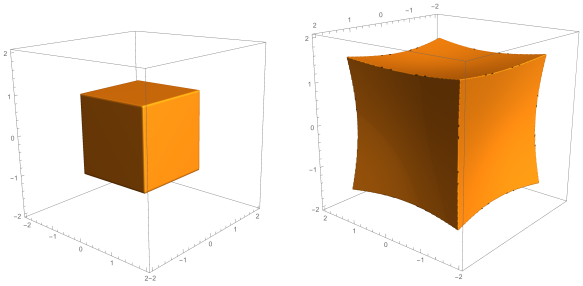


Figure 1: The identified set and the corresponding domain of KMS procedure for $d = 3$ in Example 3.

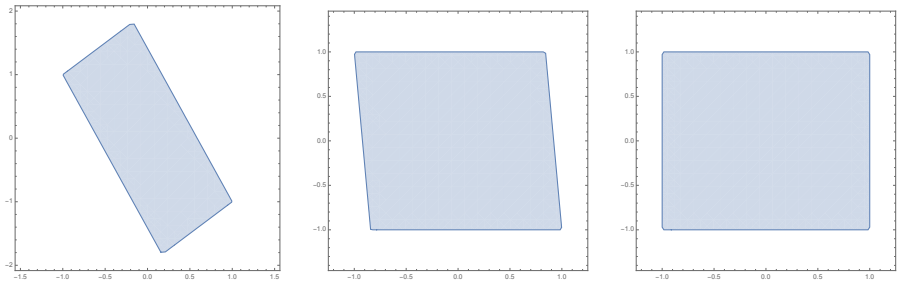
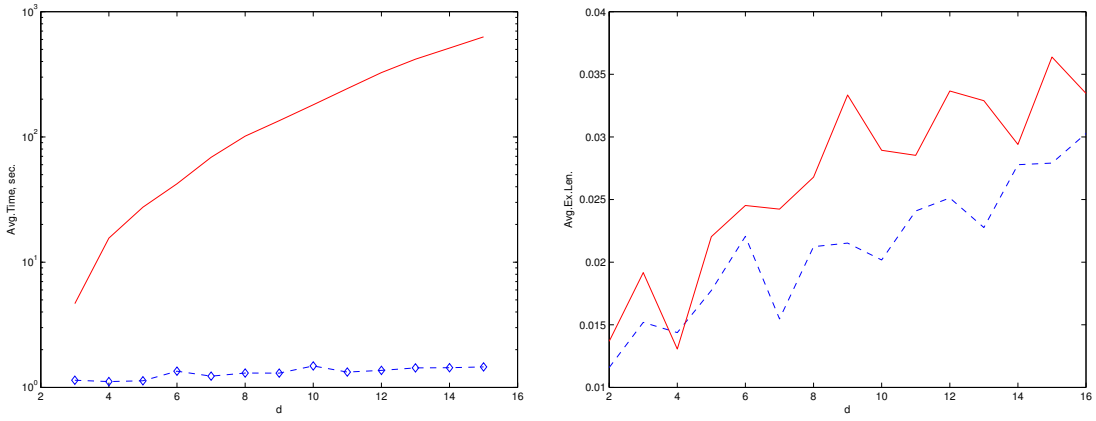


Figure 2: The identified sets for DGP 1-3 in the MC experiment

Figure 3: Average computation time and excess length for $\mathrm{CI}^{AS}_{\alpha,n}$ and $\widetilde{\mathrm{CI}}_{\alpha,n}$, $n = 1000$



Note: The solid line corresponds to $\mathrm{CI}^{AS}_{\alpha,n}$, the dashed line corresponds to $\widetilde{\mathrm{CI}}_{\alpha,n}$.

8.2 MC results

Table 2: MC results for DGP 1-4 .

		$n = 100$		$n = 1000$		$n = 10000$		$n = 100000$	
DGP	$1-\alpha$	$\mathrm{CI}_{\alpha,n}$	$\widetilde{\mathrm{CI}}_{\alpha,n}$	$\mathrm{CI}_{\alpha,n}$	$\widetilde{\mathrm{CI}}_{\alpha,n}$	$\mathrm{CI}_{\alpha,n}$	$\widetilde{\mathrm{CI}}_{\alpha,n}$	$\mathrm{CI}_{\alpha,n}$	$\widetilde{\mathrm{CI}}_{\alpha,n}$
1	0.90	0.8960	0.9890	0.8995	0.9920	0.8925	0.9925	0.9075	0.9945
	0.95	0.9460	0.9955	0.9460	0.9965	0.9445	0.9965	0.9530	0.9970
	0.975	0.9705	0.9980	0.9740	0.9980	0.9735	0.9990	0.9745	0.9995
	Avg. Ex. Len.	0.0583	0.0888	0.0186	0.0292	0.0059	0.0094	0.0019	0.0030
2	0.90	0.8820	0.8895	0.8910	0.8960	0.8950	0.8950	0.9000	0.9000
	0.95	0.9360	0.9425	0.9385	0.9390	0.9485	0.9495	0.9550	0.9550
	0.975	0.9660	0.9685	0.9655	0.9665	0.9735	0.9735	0.9775	0.9780
	Avg. Ex. Len.	0.0671	0.0678	0.0212	0.0213	0.0067	0.0068	0.0021	0.0021
3	0.90	0.9190	0.9870	0.9225	0.9890	0.9140	0.9930	0.9160	0.9935
	0.95	0.9550	0.9950	0.9595	0.9945	0.9560	0.9985	0.9515	0.9965
	0.975	0.9760	0.9985	0.9800	0.9970	0.9785	0.9990	0.9685	0.9980
	Avg. Ex. Len.	0.0630	0.0860	0.0195	0.0279	0.0061	0.0090	0.0019	0.0029
4	0.90	0.8860	0.8860	0.8980	0.9005	0.8925	0.8925	0.8975	0.8975
	0.95	0.9430	0.9430	0.9505	0.9515	0.9435	0.9470	0.9440	0.9440
	0.975	0.9685	0.9685	0.9760	0.9780	0.9720	0.9730	0.9720	0.9720
	Avg. Ex. Len.	0.0655	0.0721	0.0212	0.0220	0.0067	0.0068	0.0021	0.0021

Note: Average excess length is measured as the difference between the average $|\mathrm{CI}_{\alpha,n}|$ and $|\mathcal{S}(P)|$ for $\alpha = 0.1$.

8.3 Topological properties of optimal solutions

Consider any distribution P with support on $\mathbb{R}^{(k-2d)\times(d+1)}$ such that $(A_P, b_P) \triangleq \mathbb{E}_P W$ exist. Let $\mathcal{J}_k^a(\theta; P) \subset \{1, ..., k\}$ be the set of indices of moment equality and inequality constraints active at θ, i.e. all j s.t. $m_j(\theta, P) \triangleq \mathbb{E}_P g_j(W, \theta) = 0$. $\mathcal{J}_k^a(\theta; P)$ can be empty.

Lemma 1 (Characterization of the optimal solution)**.** *Under Assumption 1 for any $\mu \geq 0$ any minimizer θ for Program* (10) *is a solution to the corresponding Karush–Kuhn–Tucker (KKT) optimality conditions for some finite $\lambda \in \mathbb{R}^k$,*

$$\begin{cases} (e_1 + 2\mu\theta)' = -\lambda' A_P, & (36) \\ m_j(\theta, P) = 0 \qquad j \in \mathcal{J}_k^{eq}, & (37) \\ m_j(\theta, P) \leq 0, \lambda_j \geq 0, \lambda_j m_j(\theta, P) = 0 \qquad j \in \mathcal{J}_k^{ineq}. & (38) \end{cases}$$

Proof. By Assumption 1, $\Theta(P) \subset \Theta$ is non–empty and closed, so the global optima for Program (10) exist. Program (10) is convex for any $\mu \geq 0$, i.e. the objective function is convex, the constraints are affine. Assumption 1 implies Slater's condition. Since the Program (10) is convex, any global optimum $\underline{\theta}(\mu, P)$ of Program 10 satisfies (36)-(38) for some finite vector of Lagrange multipliers λ (maybe non–unique) (see p.244 in Boyd and Vandenberghe (2004)). □

If we introduce notation $\underline{\mathcal{L}}(\lambda, \theta; \mu, P) \triangleq \theta_1 + \mu \|\theta\|^2 + \lambda' m(\theta, P)$, (36) becomes

$$\partial_\theta \underline{\mathcal{L}}(\lambda, \theta; \mu, P) = 0$$

Let $\underline{\xi}(\mu, P) \triangleq (\underline{\theta}(\mu, P), \underline{\lambda}(\mu, P))$ be a set of solutions to (36)-(38). In order to have a unique solution λ Program (10) need to meet a stronger constraint qualification condition, Assumption 6 defined below.

As before, I use symbols $\mathcal{J}_k^a(\theta; P)$, $\mathcal{J}^{\mathrm{a}}(\mu, P)$ etc to denote the projectors on the coordinates with the corresponding indices.use different symbol for projection Let $\mathbb{J}_{d+1}^d \triangleq (e_1, \ldots, e_d)$.

Assumption 6 (Linear Independence Constraint Qualification (LICQ))**.** *The matrix $\mathbb{J}_k^a(\theta; P) A_P$ has full row rank for any $\theta \in \Theta(P)$.*

Lemma 2 (Sufficient condition for LICQ)**.** *Assumptions 2-3 imply Assumption 6.*

Proof. Assumption 3 implies that $\mathcal{J}_k^a(\theta; P)$ has at most d elements at any $\theta \in \Theta(P)$. Consider any point $\theta \in \Theta(P)$. The set $\mathcal{J}_k^{ineq}$ includes (1), so $k \geq 2d \geq d+1$. It implies that there exists a set $\mathcal{J}$ with $|\mathcal{J}| = d$ such that $\mathcal{J}_k^a(\theta; P) \subset \mathcal{J}$. By Assumption 2, $\mathrm{rk}[\mathbb{J}\mathbb{E}_P W] = d$, so $M \triangleq \mathbb{J}_k^a(\theta; P)(A_P, b_P)$ has full row rank which is equal to $|\mathcal{J}_k^a|$. By definition, $\mathbb{J}_k^a(\theta; P) A_P \theta = \mathbb{J}_k^a b_P$. It implies by the Rouché–Capelli theorem that the matrices $M_\theta \triangleq \mathbb{J}_k^a(\theta; P) A_P$ and M have the same rank. This result implies Assumption 6. □

The inverse implication does not hold in general as the following remark shows.

Remark 1. Assumption 6 implies Assumption 3 and that for any $\theta \in \Theta(P)$

$$\mathrm{rk}[\mathbb{J}_k^a(\theta; P)\, \mathbb{E}_P W] = |\mathcal{J}_k^a(\theta; P)|. \tag{39}$$

Indeed, suppose that Assumption 6 holds. It immediately implies Assumption 3. To see (39) consider any point $\theta \in \Theta(P)$ such that $|\mathcal{J}_k^a(\theta; P)| \leq d$. By the Rouché–Capelli theorem and the full row rank property of M_θ correspondingly,

$$\mathrm{rk}(M) = \mathrm{rk}(M_\theta) = |\mathcal{J}_k^a(\theta; P)|.$$

Lemma 3 (Uniqueness of the optimal solutions)**.** *Suppose that both Assumptions 1 and 6 are satisfied. Then for any $\mu \geq 0$ the set of multipliers $\underline{\lambda}(\mu, P)$ is a singleton. Moreover if $\mu > 0$, then $\underline{\xi}(\mu, P)$ is a singleton.*

Proof. By definition of $\mathbb{J}_k^a(\theta; P)$, any θ and λ satisfying (36) satisfy

$$\lambda'(\mathbb{J}_k^a(\theta; P))'\mathbb{J}_k^a(\theta; P) = \lambda'. \tag{40}$$

So (36) becomes

$$(e_1 + 2\mu\theta)' = -\gamma'\mathbb{J}_k^a(\theta; P) A_P, \tag{41}$$

where $\gamma' \triangleq \lambda'(\mathbb{J}_k^a(\theta; P))' \in \mathbb{R}^{|\mathcal{J}_k^a(\theta;P)|}$. By Assumption 6, for any $\theta \in \Theta(P)$ the matrix $A \triangleq \mathbb{J}_k^a(\theta; P) A_P$ has full rank. Hence for any θ there can be at most one $\gamma^* \in \mathbb{R}^{|\mathcal{J}_k^a(\theta;P)|}$ satisfying (41). If $e_1 + 2\mu\theta = 0$, then trivially λ is a zero vector. Otherwise it is given by

$$\gamma^* = (AA')^{-1}A'(e_1 + 2\mu\theta). \tag{42}$$

Then $(\underline{\lambda}(\mu, P))' \triangleq (\gamma^*)'\mathbb{J}_k^a(\theta; P)$ is the unique solution to (36)-(38) for any solution θ.

Now consider the case $\mu > 0$. The second order derivative matrix of $\underline{\mathcal{L}}(\lambda, \theta; \mu, P)$ with respect to θ at any solution $\underline{\xi}(\mu, P)$ is $2\mu I_d$. It is positive definite for any $\mu > 0$, so the Second Order Sufficient Condition (SOSC) is satisfied at any point. By Theorem 3.63 from Bonnans and Shapiro (2000) the second order growth condition holds at $\underline{\theta}(\mu, P)$, i.e. $\exists \varepsilon > 0$ and $c > 0$ s.t. for $\forall \theta \in \Theta(P)$ s.t. $\|\theta - \underline{\theta}(\mu, P)\| < \varepsilon$ the following inequality holds

$$\theta_1 + \mu\|\theta\|^2 \geq e_1'\underline{\theta}(\mu, P) + \mu\|\underline{\theta}(\mu, P)\|^2 + c\|\theta - \underline{\theta}(\mu, P)\|^2.$$

So the value of the objective function at $\underline{\theta}(\mu, P)$ is strictly smaller than the value at any other point in a neighborhood of $\underline{\theta}(\mu, P)$. Since for the convex program the set of global optima is convex and connected, it implies that $\underline{\theta}(\mu, P)$ is the unique global minimizer. □

Lemma 4. *Suppose that Assumptions 1-3 are satisfied and $\mu \leq 1/2$. Then*

$$\|\underline{\lambda}(\mu, P)\|^2 \leq C_\Lambda^2 \triangleq \frac{C_\Theta^3}{\eta^2} < \infty, \tag{43}$$

where $C_\Theta \triangleq (1 + \max_{\theta\in\Theta}\|\theta\|)$.

Proof. Consider any point $\theta \in \underline{\theta}(\mu, P)$ and the corresponding $\mathcal{J}_k^a(\theta; P)$. Let $A \triangleq \mathbb{J}_k^a(\theta; P) A_P$ and $b \triangleq \mathbb{J}_k^a(\theta; P) b_P$. Let $\eta_A \triangleq \text{eig}(AA')$ so that equation (42) implies

$$\|\underline{\lambda}(\mu, P)\| \leq \eta_A^{-1}\|e_1 + 2\mu\theta\|. \tag{44}$$

By the variational property of eigenvalues,

$$\eta_A^2 = \min_{v\in\mathbb{R}^\ell} \frac{v'AA'v}{v'v}. \tag{45}$$

By Assumption 2

$$\eta^2 \leq \text{eig}((A, b)(A, b)') \triangleq \min_{v\in\mathbb{R}^\ell} \frac{v'(AA' + bb')v}{v'v}.$$

Let v_A be any minimizer of the r.h.s. of (45) such that $v_A'v_A = 1$. Then

$$\eta^2 \leq v_A'(AA' + bb')v_A = (A'v_A)'(I_d + \theta\theta')(A'v_A) \tag{46}$$

where the last equality holds since by definition $b = A\theta$. Finally,

$$\frac{\eta^2}{\eta_A^2} \leq \frac{(A'v_A)'(I_d + \theta\theta')(A'v_A)}{(A'v_A)'(A'v_A)} \leq \left\|I_d + \theta\theta'\right\| \leq C_\Theta. \tag{47}$$

Result (43) then follows from (44) and (47) for any $\mu \leq 1/2$. □

Remark 2. Equation (47) provides bound for, A, a matrix with gradients of active moment conditions at any point $\theta \in \Theta$,

$$\left\| \left(AA' \right)^{-1} \right\| \leq C_\Theta \eta^{-2}. \tag{48}$$

The function $\phi(a,b) \triangleq \sqrt{a^2+b^2} + a - b$, considered in Fischer (1992), has the following property.

Proposition 1.

$$\phi(a,b) = 0 \text{ } \textit{if and only if } a \leq 0, b \geq 0, ab = 0. \tag{49}$$

It can be used to replace (38) with an equivalent equality so that the KKT system becomes a system of equations. This result can be used to establish the continuity of the solutions in μ as the following lemma shows.

Lemma 5. *Under Assumptions 1- 3 $\underline{\xi}(\mu, P)$ is u.h.c. in μ; $\underline{v}(\mu, P)$ is continuous in μ for $\mu \geq 0$.*

Proof. By Proposition 1 equation (38) is equivalent to

$$\phi\left(m_j(\theta, P), \lambda_j\right) = 0 \text{ for } j \in \mathcal{J}_k^{ineq}. \tag{50}$$

Solutions to (36),(37),(50) coincide with solutions to

$$\Psi(\theta, \lambda; \mu, P) \triangleq \left\| \partial_\theta \underline{\mathcal{L}}(\lambda, \theta; \mu, P) \right\|_2^2 + \sum_{j \in \mathcal{J}_k^{eq}} \left(m_j(\theta, P)\right)^2 + \sum_{j \in \mathcal{J}_k^{ineq}} \left(\phi\left(m_j(\theta, P), \lambda_j\right)\right)^2 = 0. \tag{51}$$

Lemmas 3-4 imply taht $\underline{\lambda}(\mu, P)$ is unique and satisfies (43) for any $\mu \in [0, 1/2]$. So the solution to (51) coincides with solutions of

$$\begin{array}{ll} \min\limits_{\theta, \lambda} & \Psi(\theta, \lambda; \mu, P) \\ \text{s.t.} & \theta \in \Theta, \lambda \in \mathbb{R}^k, \|\lambda\| \leq C_\Lambda. \end{array} \tag{52}$$

The objective function of this program is continuous in μ and the domain is a compact valued continuous correspondence in μ. By the Maximum Theorem (see Ok (2007)) $\underline{\xi}(\mu, P)$ is u.h.c. function of $\mu \geq 0$.

Function $\underline{v}(\mu, P) = e_1' \underline{\theta}(\mu, P) + \mu \left\| \underline{\theta}(\mu, P) \right\|^2$ is a composition of u.h.c. functions and hence, by Theorem VI.2.1' from Berge (1963), is u.h.c. in $\mu \in \mathbb{R}_+$. Since by definition $\underline{v}(\mu, P)$ is a single–valued function, u.h.c. implies continuity in $\mu \geq 0$ for any fixed P. □

8.4 Smoothness properties

Let $\dot{m}(\theta) \triangleq \delta X \theta - \delta y$.

$$\begin{aligned} \mathcal{J}_k^+(\mu, P) &\triangleq \left\{ j \in \mathcal{J}_k^{ineq} | \underline{\lambda}_j(\mu, P) > 0 \right\} \cup \mathcal{J}_k^{eq}, \\ \mathcal{J}_k^-(\mu, P) &\triangleq \left\{ j \in \mathcal{J}_k^{ineq} | m_j(\underline{\theta}(\mu, P), P) > 0 \right\}, \\ \mathcal{J}_k^0(\mu, P) &\triangleq \left\{ j \in \mathcal{J}_k^{ineq} | \underline{\lambda}_j(\mu, P) = 0, m_j(\underline{\theta}(\mu, P), P) = 0 \right\}, \\ \mathcal{J}_k^a(\mu, P) &\triangleq \mathcal{J}_k^0(\mu, P) \cup \mathcal{J}_k^+(\mu, P). \end{aligned}$$

For any $\delta\mu \in \mathbb{R}$ and $\delta W = (\delta X, -\delta y) \in \mathbb{R}^{k \times (d+1)}$, $t \geq 0$, $\mu > 0$ consider the program

$$\min_{\theta \in \Theta} \quad e_1' \theta + (\mu + t\delta\mu) \left\| \theta \right\|^2, \tag{53}$$

$$\text{s.t.} \quad \begin{cases} m_j(\theta, P) + t\dot{m}_j(\theta) = 0 & \text{for } j \in \mathcal{J}_k^{eq}, \\ m_j(\theta, P) + t\dot{m}_j(\theta) \leq 0 & \text{for } j \in \mathcal{J}_k^{ineq}. \end{cases}$$

Program (53) has a unique solution $\underline{\xi}_\delta(t)$ for $0 \leq t < T$ which can be represented as $\underline{\xi}_\delta(t) = \underline{\xi} + t\dot{\underline{\xi}}$. The formula for $\dot{\underline{\xi}}$ contains

$$\begin{aligned} \mathcal{J}_k^\delta(\mu, P) &\triangleq \left\{ j \in \mathcal{J}_k^0(\mu, P) \,|\dot{\underline{\lambda}}_j(\mu, P) > 0 \right\} \cup \mathcal{J}_k^a(\mu, P), \\ A_\delta(\mu, P) &\triangleq \mathbb{J}_k^\delta(\mu, P) A_P, \\ A^\dagger &\triangleq A' \left(AA'\right)^{-1}, \\ Q_\delta &\triangleq I_d - A_\delta^\dagger A_\delta. \end{aligned}$$

I suppress the argument of $A_\delta(\mu, P)$ from now on.

Lemma 6 (Local linear representation)**.** *Suppose that Assumptions 1 -3 hold. There is a neighborhood* $[0, T(\mu, \delta, P)]$ *in which Program* (53) *has a unique solution* $\underline{\xi}_\delta(t) = \underline{\xi} + t\dot{\underline{\xi}}$ *with*

$$\dot{\underline{\xi}} \triangleq - \begin{pmatrix} (2\mu)^{-1} Q_\delta & A_\delta^\dagger \\ (\mathbb{J}_k^\delta)'(A_\delta^\dagger)' & -2\mu(\mathbb{J}_k^\delta)'(A_\delta A_\delta')^{-1} \end{pmatrix} \begin{pmatrix} (\delta X)'\underline{\lambda} + 2\delta\mu\underline{\theta} \\ \mathbb{J}_k^\delta(\delta X\underline{\theta} - \delta y) \end{pmatrix}. \tag{54}$$

Proof. By Lemma 3, if $t = 0$ Program (53) has a unique solution $\underline{\xi}$. Since this solution satisfies LICQ (Assumption 6) and SOSC, it is strongly regular by Proposition 5.38 from BS(2000). The remaining argument follows the proof of Theorem 5.60 from BS(2000), which is an implicit function theorem for generalized equations at a strongly regular solution. We are going to apply it to the KKT conditions for Program (53) at the strongly regular solution $\underline{\xi}$,

$$\begin{cases} (e_1 + 2(\mu + \delta\mu t)\theta)' = -\lambda'(A_P + t\delta X), & (55) \\ m_j(\theta, P) + t\dot{m}_j(\theta) = 0 & j \in \mathcal{J}_k^{eq}, \quad (56) \\ \phi(m_j(\theta, P) + t\dot{m}_j(\theta), \lambda_j) = 0 & j \in \mathcal{J}_k^{ineq}. \quad (57) \end{cases}$$

By Theorem 5.60 from BS(2000), $\underline{\xi}_\delta(t)$ is analytic in t in some neighborhood $[0, T(\mu, \delta, P)]$, i.e. it can be represented as power series. First, let us compute the linear term. By the strong regularity, there exist a unique solution $(h, q) = \left(\dot{\underline{\theta}}, \dot{\underline{\lambda}}\right)$ to the following system of equations

$$\begin{cases} 2\mu h' I_d + q' A_P = -\underline{\lambda}'\delta X - 2\delta\mu\underline{\theta}', & (58) \\ m_j(\underline{\theta}, P) + t\dot{m}_j(\underline{\theta}) = 0 & j \in \mathcal{J}_k^+(\mu, P), \quad (59) \\ \phi(m_j(\theta, P) + t\dot{m}_j(\theta), q_j) = 0 & j \in \mathcal{J}_k^0(\mu, P), \quad (60) \\ q_j = 0 & j \in \mathcal{J}_k^-(\mu, P). \quad (61) \end{cases}$$

This unique solution defines the set $\mathcal{J}_k^\delta$. System (58)-(61) can be represented in a matrix form:

$$\begin{pmatrix} 2\mu I_d & A_\delta' \\ A_\delta & 0 \end{pmatrix} \begin{pmatrix} \dot{\underline{\theta}} \\ \mathbb{J}_k^\delta \dot{\underline{\lambda}} \end{pmatrix} = - \begin{pmatrix} (\delta X)'\underline{\lambda} + 2\delta\mu\underline{\theta} \\ \mathbb{J}_k^\delta(\delta X\underline{\theta} - \delta y) \end{pmatrix}, \tag{62}$$

In addition to that, $\dot{\underline{\lambda}} = (\mathbb{J}_k^\delta)'\mathbb{J}_k^\delta\dot{\underline{\lambda}}$. One can check by direct computation that

$$\begin{pmatrix} 2\mu I_d & A_\delta' \\ A_\delta & 0 \end{pmatrix}^{-1} = \begin{pmatrix} (2\mu)^{-1} Q_\delta & A_\delta^\dagger \\ (A_\delta^\dagger)' & -2\mu(A_\delta A_\delta')^{-1} \end{pmatrix}.$$

Since the higher order derivatives of every constraint function and the objective function of Program (53) with respect to t are zero, the higher order directional derivatives of $\underline{\xi}_\delta$ are equal to zero at $t = 0$. Thus the power series expansion of $\underline{\xi}_\delta$ has only constant and linear terms. □

Lemma 7 (Uniform error bounds)**.** *Suppose that Assumptions 1-3 hold and $\delta\mu = 0$. Then for any $\mu \in (0, 1/2]$ the solution of Program (53) satisfies*

$$\|\underline{\theta}_\delta(t) - \underline{\theta}\| \leq \frac{\sqrt{2C_\Theta^3}}{\eta} \|\delta W\| \frac{t}{\mu} \tag{63}$$

$$\|\underline{\lambda}_\delta(t) - \underline{\lambda}\| \leq \frac{\|\mathbb{E}_P W\| C_\Theta^4 + \eta^2 C_\Theta^2}{\eta^4} \|\delta W\| t \tag{64}$$

$$\left|\underline{v}_\delta(t) - \underline{v} - \underline{\lambda}'(t\dot{m}(\underline{\theta}))\right| \leq \frac{2C_\Theta^3}{\eta^2} \|\delta W\|^2 \frac{t^2}{\mu} \tag{65}$$

Proof. By Lemma 6 the value function $\underline{v}_\delta(t) \triangleq e_1'\underline{\theta}_\delta(t) + \mu \|\underline{\theta}_\delta(t)\|^2$ can be represented as

$$\underline{v}_\delta(t) = \underline{v} + t\left(e_1 + 2\mu\underline{\theta}\right)'\dot{\underline{\theta}} + \mu t^2 \left\|\dot{\underline{\theta}}\right\|^2. \tag{66}$$

First, consider the second term. Since by definition $\mathcal{J}_k^+ \subseteq \mathcal{J}_k^\delta$, we have $\underline{\lambda}' = \underline{\lambda}'(\mathbb{J}_k^\delta)'\mathbb{J}_k^\delta$. Correspondingly, $\underline{\lambda}' A_P = \underline{\lambda}'(\mathbb{J}_k^\delta)' A_\delta$. By Lemma 3, $(e_1 + 2\mu\underline{\theta})' = -\underline{\lambda}' A_P$. So

$$\left(e_1 + 2\mu\underline{\theta}\right)' Q_\delta = \underline{\lambda}'(\mathbb{J}_k^\delta)'(A_\delta Q_\delta) = 0, \tag{67}$$

$$\left(e_1 + 2\mu\underline{\theta}\right)' A_\delta^\dagger = \underline{\lambda}'(\mathbb{J}_k^\delta)'(A_\delta A_\delta^\dagger) = \underline{\lambda}'(\mathbb{J}_k^\delta)'. \tag{68}$$

Equations (67) and (68) imply that

$$\left(e_1 + 2\mu\underline{\theta}\right)'\dot{\underline{\theta}} = \underline{\lambda}'\dot{m}(\underline{\theta}). \tag{69}$$

Second, by Lemma 4 and Remark 2 for any $\mu \leq 1/2$,

$$\left\|\left(A_\delta A_\delta'\right)^{-1}\right\| \leq C_\Theta \eta^{-2} \text{ and } \|\underline{\lambda}\|^2 \leq C_\Theta^3 \eta^{-2}. \tag{70}$$

Then by the triangular inequality and inequalities (70)

$$\left\|\dot{\underline{\theta}}\right\|^2 = \frac{1}{(2\mu)^2}\left(\underline{\lambda}'\delta X\right) Q_\delta \left(\underline{\lambda}'\delta X\right)' + \dot{m}(\underline{\theta})'(\mathbb{J}_k^\delta)'\left(A_\delta A_\delta'\right)^{-1}\mathbb{J}_k^\delta \dot{m}(\underline{\theta}) \tag{71}$$

$$\leq \|\delta W\|^2 \frac{C_\Theta^3}{\eta^2}\left(\frac{1}{\mu^2} + 1\right), \tag{72}$$

which implies (63). Equation (64) can be proven similarly,

$$\left\|\dot{\underline{\lambda}}\right\| = \left\|(\mathbb{J}_k^\delta)'(A_\delta^\dagger)'\left(\underline{\lambda}'\delta X\right) - 2\mu(\mathbb{J}_k^\delta)'\left(A_\delta A_\delta'\right)^{-1}\mathbb{J}_k^\delta \dot{m}(\underline{\theta})\right\| \leq \tag{73}$$

$$\leq \frac{\|\mathbb{E}_P W\| C_\Theta^4 + \eta^2 C_\Theta^2}{\eta^4} \|\delta W\| \tag{74}$$

Finally, the bound in (65) follows from equations (63), (66), and (69). □

Lemma 8. *Suppose that Assumptions 1-3 hold. There exist some $\bar{\mu}(P) > 0$ such that for any $\mu < \bar{\mu}(P)$ the solution to Program (10), $\underline{\theta}(\mu, P)$ is constant.*

Proof. Take $\delta W = 0$, $\delta\mu = 1$ and any $\mu_0 > 0$ in a neighborhood of 0. By Lemma 6 we get

$$\dot{\underline{\theta}}(\mu_0, P) = -\frac{1}{\mu_0} Q_\delta(\mu_0, P)\, \underline{\theta}(\mu_0, P). \tag{75}$$

Substitute difference in eq.(36) (Lemma 1) between $\mu = 0$ and μ_0 for $\underline{\theta}$ in (75),

$$\dot{\underline{\theta}}(\mu_0, P) = (2\mu_0^2)^{-1} Q_\delta(\mu_0, P)\, A_P'\left(\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)\right). \tag{76}$$

Now we need to show that $\dot{\underline{\theta}}(\mu_0, P) = 0$. To establish this result, we need to study the behavior of the set of inequalities with positive Lagrange multipliers.

Consider any $j \in \mathcal{J}_k^{ineq}$. We know by Lemma 5 that $\underline{\lambda}_j(\mu, P)$ is continuous in μ. If $\underline{\lambda}_j(0, P) > 0$ then by continuity $\underline{\lambda}_j(\mu, P) > 0$ in some neighborhood $(0, \bar{\mu}_j(P)]$. If $\underline{\lambda}_j(0, P) = 0$ set $\bar{\mu}_j = 1$. Take $\bar{\mu}(P) \triangleq \min_{j \in \mathcal{J}_k^{ineq}} \bar{\mu}_j(P)$. WLOG suppose that $\mu_0 \in [0, \bar{\mu}(P)]$, so we get the inclusion

$$\mathcal{J}_k^+(0, P) \subseteq \mathcal{J}_k^+(\mu_0, P). \tag{77}$$

Equation (40) implies

$$\mathcal{J}_k^+(\mu_0, P) \subseteq \mathcal{J}_k^\delta(\mu_0, P). \tag{78}$$

By definition of the index matrices, inclusions (77) and (78) imply that

$$\underline{\lambda}(0, P) = (\mathbb{J}_k^\delta(\mu_0, P))' \mathbb{J}_k^\delta(\mu_0, P)\, \underline{\lambda}(0, P), \tag{79}$$

$$\underline{\lambda}(\mu_0, P) = (\mathbb{J}_k^\delta(\mu_0, P))' \mathbb{J}_k^\delta(\mu_0, P)\, \underline{\lambda}(\mu_0, P), \tag{80}$$

so

$$A_P'\left(\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)\right) = A_\delta'(\mu_0, P)\left(\mathbb{J}_k^\delta(\mu_0, P)\right)\left(\underline{\lambda}(\mu_0, P) - \underline{\lambda}(0, P)\right). \tag{81}$$

Since by definition $Q_\delta(\mu_0, P) A_\delta(\mu_0, P) = 0$, equation (76) implies $\dot{\underline{\theta}}(\mu_0, P) = 0$. By Lemma 5 the single valued function $\underline{\theta}(\mu, P)$ is continuous for $\mu > 0$. So the r.h.s. directional derivative being equal to zero implies that $\underline{\theta}(\mu_0, P) = \underline{\theta}(\bar{\mu}(P), P)$ for any $\mu_0 \in (0, \bar{\mu}(P)]$. □

Remark 3. Equation (36) from Lemma 1 with $\mu = 0$ and $\mu = \mu_0$ also implies

$$\underline{\lambda}(\mu_0, P) = \underline{\lambda}(0, P) - 2\mu_0 \underline{\theta}'(\bar{\mu}(P), P)\, A_\delta^\dagger(\mu_0, P)\, \mathbb{J}_k^\delta(\mu_0, P).$$

This implies that $\underline{\lambda}(\mu, P)$ is Lipschitz at $\mu = 0$. By Lemma 4 the Lipschitz constant can be taken equal to $2C_\Lambda$. So for any $j \in \mathcal{J}_k^{ineq}$ with $\underline{\lambda}_j(0, P) > 0$, we can take $\bar{\mu}_j(P) = \underline{\lambda}_j(0, P)/2C_\Lambda$. On a top of that, Lemma 6 implies that $\underline{\lambda}(\mu, P)$ is Lipschitz in μ with the same constant for any $\mu \in [0, 1/2]$

8.5 Estimators

Proposition 2. *Suppose that* $\mathbb{E}_P |\xi|^{1+\epsilon} \leq \infty$ *for some* $\epsilon > 0$. *Then for any* $r > 0$

$$\mathbb{E}_P[|\xi| I\{|\xi| \geq r\}] \leq \mathbb{E}_P |\xi|^{1+\epsilon} / r^\epsilon.$$

Proof. The result follows from the monotonicity of integrals. □

$$G_n(P) \triangleq \sqrt{n}\left(\operatorname{vec}\left(\frac{1}{n}\sum_{i=1}^n w_i\right) - \operatorname{vec}(\mathbb{E}_P W)\right).$$

Let $\pi(P, Q)$ denote the Prohorov distance between probability laws on P and Q , which induces the weak topology (see p. 456 in invan der Vaart and Wellner (1996)).

Lemma 9. *Consider $\mathcal{P}$, a class of distributions satisfying Assumptions 4-5, and any $\epsilon > 0$. Then*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(\left\| \frac{1}{n}\sum_{i=1}^{n} w_i - \mathbb{E}_P W \right\| \geq \epsilon\right) = 0, \tag{82}$$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} P\left(\left\| \frac{1}{n}\sum_{i=1}^{n} w_i \otimes w_i - \mathbb{E}_P[W\otimes W] \right\| \geq \epsilon\right) = 0, \tag{83}$$

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}} \pi(G_n(P), N(0, \Omega_P)) = 0, \tag{84}$$

where $\Omega_P = Cov_P(\mathrm{vec}(W))$.

Proof. Consider any combination of indices for any combination of indices r, ℓ, j, m. Assumption 5 together with Schwarz and Jensen's inequalities implies,

$$\mathbb{E}_P |W_{r,\ell} W_{j,m}|^{1+\varepsilon/2} \leq \left(\mathbb{E}_P |W_{r,\ell}|^{2+2\varepsilon} \mathbb{E}_P |W_{j,m}|^{2+2\varepsilon}\right)^{1/2} \leq \bar{M}. \tag{85}$$

So the random variables $|W_{r,\ell}|$ and $|W_{r,\ell}W_{j,m}|$ have correspondingly finite $1+\varepsilon/2$ and $2+\varepsilon$ moments. The bound (85) on the moments is independent of $P \in \mathcal{P}$, so these random variables are uniformly integrable on $\mathcal{P}$ by Proposition 2. The limits (82) and (83) follow immediately from Proposition A.5.1 in van der Vaart and Wellner (1996). The result (84) follows from Proposition A.5.2 in the same book. □

Lemma 10. *Suppose that Assumptions 1-5 holds for P. Then*

$$\underline{v}(\mu_n, \mathbb{P}_n) = \underline{v}(\mu_n, P) + \frac{1}{n}\sum_{i=1}^{n} \underline{\lambda}(\mu_n, P)' g(w_i, \underline{\theta}(\mu_n, P)) + R_n, \tag{86}$$

where $|R_n| \leq \|G_n(P)\|^2 \frac{2C_\Theta^3}{\eta^2}\frac{1}{\mu_n n}$ with probability approaching 1.

Proof. Consider $t = 1/\sqrt{n}$ and $\dot{m}(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(w_i,\theta) - m(\theta, P))$. The result follows from Lemma 7 for a sufficiently large n with probability approaching 1.

□

8.6 Proof or Theorem 1

$$\begin{aligned}
\mathbf{\Sigma}(\theta) &\triangleq \mathbb{E}_P\left[g(W,\theta)g(W,\theta)'\right] - m(\theta,P)m(\theta,P)', \\
\hat{\mathbf{\Sigma}}_n(\theta) &\triangleq \frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta)g(w_i,\theta)' - \frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta)\frac{1}{n}\sum_{i=1}^{n} g(w_i,\theta)'.
\end{aligned}$$

Let

$$\rho_n(P) \triangleq \pi(\sqrt{n}(\underline{v}(\mu_n, \mathbb{P}_n) - \underline{v}(\mu_n, P_n)), N\left(0, \underline{\sigma}^2(\mu_n, P)\right))$$

Proof. Consider $t = 1/\sqrt{n}$ and $\dot{m}(\theta) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n}(g(w_i,\theta) - m(\theta,P))$. The results (15) and (16) follow from Lemmas 7 and 9.

By the triangular inequality,

$$\left|\hat{\underline{\sigma}}^2_{\mu_n} - \underline{\sigma}^2_{\mu_n}\right| = \left|\hat{\underline{\lambda}}'_{\mu_n}\hat{\mathbf{\Sigma}}_n(\hat{\underline{\theta}}_{\mu_n})\hat{\underline{\lambda}}_{\mu_n} - \underline{\lambda}'_{\mu_n}\mathbf{\Sigma}(\underline{\theta}_{\mu_n})\underline{\lambda}_{\mu_n}\right| \leq \tag{87}$$

$$\left|\hat{\underline{\lambda}}'_{\mu_n}\hat{\mathbf{\Sigma}}_n(\hat{\underline{\theta}}_{\mu_n})\hat{\underline{\lambda}}_{\mu_n} - \underline{\lambda}'_{\mu_n}\hat{\mathbf{\Sigma}}_n(\underline{\theta}_{\mu_n})\underline{\lambda}_{\mu_n}\right| + \left|\underline{\lambda}'_{\mu_n}\hat{\mathbf{\Sigma}}_n(\underline{\theta}_{\mu_n})\underline{\lambda}_{\mu_n} - \underline{\lambda}'_{\mu_n}\mathbf{\Sigma}(\underline{\theta}_{\mu_n})\underline{\lambda}_{\mu_n}\right| \tag{88}$$

Together with (15), (16) and Lemmas 4 and 9 it implies (17).

Equation (18) follows from Lemma 9, Lemma 10, and Slutsky's theorem. □

8.7 Proof of Theorem 2

Proof. Step 1. Consider

$$\underline{\zeta}_n \triangleq \sqrt{n}\frac{\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \mu_n \left\|\underline{\theta}\left(\kappa_n, \mathbb{P}_n\right)\right\|^2 - \underline{v}\left(P\right)}{\underline{\sigma}\left(\mu, \mathbb{P}_n\right)} \tag{89}$$

Since $\mu_n/\kappa_n \to 0$ and $\mu_n \to 0$, for all n large enough, such that $\mu_n \leq \kappa_n \leq \bar{\mu}\left(P\right)$, by Lemma 8 we get

$$\underline{\theta}\left(\kappa_n, P\right) = \underline{\theta}\left(\mu_n, P\right) = \underline{\theta}\left(0+, P\right), \tag{90}$$

which implies $\underline{\theta}_1\left(\mu_n, P\right) = \underline{v}\left(P\right)$.

Using Lemma **??** and applying the delta method we get

$$\mu_n\sqrt{n}\left(\left\|\underline{\theta}\left(\kappa_n, \mathbb{P}_n\right)\right\|^2 - \left\|\underline{\theta}\left(\kappa_n, P\right)\right\|^2\right) = \frac{\mu_n}{\kappa_n}O_p(1) = o_p(1). \tag{91}$$

By Lemma 5, $\underline{\theta}\left(\mu, P\right)$ and $\underline{\lambda}\left(\mu, P\right)$ are continuous for $\mu > 0$. The matrix function $\mathbf{\Sigma}\left(\theta, P\right)$ is continuous in θ and thus $\underline{\sigma}\left(\mu, P\right)$ is continuous in μ for $\mu > 0$. So the limit $\lim_{n\to\infty}\underline{\sigma}^2\left(\mu_n, P\right)$ exists and belongs to the set $\underline{\sigma}^2\left(0, P\right)$ which by assumptions implies $\lim_{n\to\infty}\underline{\sigma}^2\left(\mu_n, P\right) \geq \sigma_0^2$. Lemma **??** together with (90) and (91) imply by Slutsky's theorem that $\underline{\zeta}_n$ converges in distribution to $N\left(0, 1\right)$.

Step 2. Consider the one–sided confidence interval $\mathrm{CI}^L_{\alpha,n}$.

$$\begin{aligned}
& \lim_{n\to\infty} \quad P\left\{\mathcal{S}\left(P\right) \subset \mathrm{CI}^L_{\alpha,n}\right\} \\
= & \lim_{n\to\infty} \quad P\left\{\underline{v}\left(P\right) \geq \underline{v}\left(\mu_n, \mathbb{P}_n\right) - \mu_n\left\|\underline{\theta}\left(\kappa_n, \mathbb{P}_n\right)\right\|^2 - \underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) z_{1-\alpha} n^{-1/2}\right\} \\
= & \lim_{n\to\infty} \quad P\left\{\underline{\zeta}_n \leq z_{1-\alpha}\right\} \\
= & \Phi\left(z_{1-\alpha}\right) \quad = 1-\alpha.
\end{aligned}$$

The proof for $\mathrm{CI}^R_{\alpha,n}$ is analogous. Proof for $\mathrm{CI}^*_{\alpha/2,n}$ follows immediately from the Bonferroni inequality.

Suppose that $p = 0$. The following argument follows refer to specific argumentthe proof of Imbens and Manski (2004). By Lemma 2, $\underline{v}\left(0, P\right) < -\bar{v}\left(0, P\right)$. So

$$\begin{aligned}
& \lim_{n\to\infty}\min_{\theta\in\Theta(P)} P\left(\theta \in \mathrm{CI}^*_{\alpha,n}\right) = \\
\min\left\{\lim_{n\to\infty} P\left\{\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \mu_n\left\|\underline{\theta}\left(\kappa_n, \mathbb{P}_n\right)\right\|^2 - \underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) z_{1-\alpha} n^{-1/2} \leq \underline{v}\left(P\right)\right\}, 1, \ \ldots\right. \\
\left.\lim_{n\to\infty} P\left\{-\bar{v}\left(\mu_n; \mathbb{P}_n\right) + \mu_n\left\|\bar{\theta}\left(\kappa_n, \mathbb{P}_n\right)\right\|^2 + z_{1-\alpha}\bar{\sigma}\left(\mu_n; \mathbb{P}_n\right) n^{-1/2} \geq \bar{v}\left(P\right)\right\}\right\} = \\
\min\left\{\lim_{n\to\infty} P\left\{\mathcal{S}\left(P\right) \subset \mathrm{CI}^L_{\alpha,n}\right\}, 1, \lim_{n\to\infty} P\left\{\mathcal{S}\left(P\right) \subset \mathrm{CI}^R_{\alpha,n}\right\}\right\} = \qquad (92) \\
\min\left\{1-\alpha, 1, 1-\alpha\right\} = 1-\alpha. \qquad (93)
\end{aligned}$$

To understand the second equation, consider the following argument. Suppose that $\theta \in \Theta\left(P\right)$ is such that $\underline{v}\left(P\right) < \theta_1 < \bar{v}\left(P\right)$. Then θ_1 will be covered with probability 1 since the $\mathrm{CI}^L_{\alpha,n}$ and $\mathrm{CI}^R_{\alpha,n}$ and upper bounds of $\mathrm{CI}^B_{\alpha,n}$ cover correspondingly $\underline{v}\left(P\right)$ and $\bar{v}\left(P\right)$. □

8.8 Proof of Theorem 3

Let the set of the basic solutions be defined as follows,

$$\mathcal{B}(P) \triangleq \{\mathcal{J}_k^a(\theta;P) \,|\theta \in \underline{\theta}(0,P), |\mathcal{J}_k^a(\theta;P)| = d\}$$

Let $\theta^{\mathcal{J}}(P) \triangleq (\mathbb{J}A_P)^{-1}(\mathbb{J}b_P)$ and

$$\hat{\mathcal{B}}_n \triangleq \left\{\mathcal{J} | \hat{\theta}_1^{\mathcal{J}} \leq \underline{\hat{\theta}}_1^{\mu_n} + \mu_n \text{ and } \forall j \in \mathcal{J}_k^{ineq}, e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \leq \mu_n\right\}$$

Proof. We need to show that

$$\inf_{\mathcal{J}\in\mathcal{B}(P);P\in\mathcal{P}} P\Big\{ \bigcap_{j\in\mathcal{J}_k^{ineq}} \{e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \leq \mu_n\} \cap \{\hat{\theta}_1^{\mathcal{J}} \leq \underline{\hat{\theta}}_1^{\mu_n} + \mu_n\}\Big\} \geq 1 - \epsilon$$

By Boole's inequality,

$$\inf_{\mathcal{J}\in\mathcal{B}(P);P\in\mathcal{P}} P\Big\{ \bigcap_{j\in\mathcal{J}_k^{ineq}} \{e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \leq \mu_n\} \cap \{\hat{\theta}_1^{\mathcal{J}} \leq \underline{\hat{\theta}}_1^{\mu_n} + \mu_n\}\Big\} \geq$$

$$1 - \sup_{\mathcal{J}\in\mathcal{B}(P);P\in\mathcal{P}} P\{\hat{\theta}_1^{\mathcal{J}} \geq \underline{\hat{\theta}}_1^{\mu_n} + \mu_n\} - \sum_{j\in\mathcal{J}_k^{ineq}} \sup_{\mathcal{J}\in\mathcal{B}(P);P\in\mathcal{P}} P\{e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \geq \mu_n\}$$

By definition, $\theta_1^{\mathcal{J}} = \underline{\theta}_1(0,P)$. By Theorem 1

$$\underline{\hat{\theta}}_1^{\mu_n} = \underline{\theta}_1(\mu_n,P) + \mu_n(\|\underline{\theta}(\mu_n,P)\|^2 - \left\|\underline{\hat{\theta}}^{\mu_n}\right\|^2) + O_{\mathcal{P}}(\frac{1}{\sqrt{n}})$$

By definition,

$$\underline{\theta}_1(\mu_n,P) \geq \underline{\theta}_1(0,P)$$

By Theorem 1,

$$\mu_n(\|\underline{\theta}(\mu_n,P)\|^2 - \left\|\underline{\hat{\theta}}^{\mu_n}\right\|^2) = O_{\mathcal{P}}(\frac{1}{\sqrt{n}}).$$

As a corollary of Lemma 9 we have $\hat{\theta}_1^{\mathcal{J}} = \theta_1^{\mathcal{J}} + O_{\mathcal{P}}(\frac{1}{\sqrt{n}})$
$P\{\hat{\theta}_1^{\mathcal{J}} \geq \underline{\hat{\theta}}_1^{\mu_n} + \mu_n\} \leq \sup P\{O_{\mathcal{P}}(\frac{1}{\mu_n\sqrt{n}}) - O_{\mathcal{P}}(\frac{1}{\mu_n\sqrt{n}}) \geq 1\} \leq \frac{\epsilon}{1+\left|\mathcal{J}_k^{ineq}\right|}$.
Step 2
Since $\theta^{\mathcal{J}}$ is a solution, it is a feasible point, for any $j \in \mathcal{J}_k^{ineq}$

$$e_j'(A_P\theta^{\mathcal{J}} - b_P) \leq .0$$

$$P\{e_j'(\hat{A}\hat{\theta}^{\mathcal{J}} - \hat{b}) \geq \mu_n\} \leq \sup P\{O_{\mathcal{P}}(\frac{1}{\mu_n\sqrt{n}}) \geq 1\} \leq \frac{\epsilon}{1 + \left|\mathcal{J}_k^{ineq}\right|}.$$

Since $\mathcal{B}(P)$ is a finite set, the statement of the theorem is true.

□

8.9 Proof of Theorem 4

Proof. The proof is analogous for all CI. Consider, for example, $\tilde{\text{CI}}_{\alpha,n}^L$. Consider an arbitrary convergent measure $P \in \mathcal{P}$.

Consider any $\delta > 0$ such that $z_{1-\alpha}\sigma^0 > 3\delta > 0$. Then by Theorem 3 and Lemma **??** correspondingly there exist $n(\delta, \epsilon)$ such that for any $n > n(\delta, \epsilon)$

$$\inf_{P\in\mathcal{P},\mathcal{J}\in\mathcal{B}(P)} P\{\hat{\beta}_n - \left\|\hat{\theta}^{\mathcal{J}}\right\|^2 \geq 0\} \geq 1-\epsilon, \tag{94}$$

$$\inf_{P\in\mathcal{P},\forall\mathcal{J}} P\{\mu_n\sqrt{n}\left|\left\|\hat{\theta}^{\mathcal{J}}\right\|^2 - \left\|\theta^{\mathcal{J}}\right\|^2\right| \leq \delta\} \geq 1-\epsilon, \tag{95}$$

$$\inf_{P\in\mathcal{P}} P\{z_{1-\alpha}\left|\underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) - \underline{\sigma}\left(\mu_n, P\right)\right| \leq \delta\} \geq 1-\epsilon, \tag{96}$$

$$\sup_{P\in\mathcal{P}} \rho_n(P) \leq \delta \tag{97}$$

By definition $\underline{v}\left(P_n\right) = \theta_1^{\mathcal{J}}$. Take any $\mathcal{J} \in \mathcal{B}(P)$. Since $\theta^{\mathcal{J}}$ is a feasible point,

$$\underline{\theta}_1(\mu_n, P) + \mu_n \left\|\underline{\theta}\left(\mu_n, P\right)\right\|^2 \leq \underline{v}\left(P\right) + \mu_n \left\|\theta^{\mathcal{J}}\right\|^2. \tag{98}$$

So

$$\begin{aligned}
P\left\{\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \mu_n\hat{\beta}_n - \underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) z_{1-\alpha} n^{-1/2} \leq \underline{v}\left(P\right)\right\} \geq \\
P\left\{\sqrt{n}(\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \underline{v}\left(\mu_n, P\right)) - \mu_n\sqrt{n}(\hat{\beta}_n - \left\|\theta^{\mathcal{J}}\right\|^2) \leq \underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) z_{1-\alpha}\right\} \geq \\
P\left\{\sqrt{n}(\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \underline{v}\left(\mu_n, P\right)) - \mu_n\sqrt{n}(\left\|\hat{\theta}^{\mathcal{J}}\right\|^2 - \left\|\theta^{\mathcal{J}}\right\|^2) \leq \underline{\sigma}\left(\mu_n, \mathbb{P}_n\right) z_{1-\alpha}\right\}(1-\epsilon) \geq \\
P\left\{\sqrt{n}(\underline{v}\left(\mu_n, \mathbb{P}_n\right) - \underline{v}\left(\mu_n, P\right)) \leq \underline{\sigma}\left(\mu_n, P\right) z_{1-\alpha} - 2\delta\right\}(1-\epsilon)^3 \geq \\
(\Phi(z_{1-\alpha} - \frac{3\delta}{\sigma^0}) - \delta)(1-\epsilon)^3
\end{aligned}$$

Since P is arbitrary , for any $n > n(\delta, \epsilon)$

$$\inf_{P\in\mathcal{P}} \min_{\theta\in\Theta(P)} P\left(\theta_1 \in \tilde{\mathrm{CI}}^L_{\alpha,n}\right) \geq (\Phi(z_{1-\alpha} - \frac{3\delta}{\sigma^0}) - \delta)(1-\epsilon)^3. \tag{99}$$

Hence,

$$\liminf_{n\to\infty} \inf_{P_n\in\mathcal{P}} \min_{\theta\in\Theta(P_n)} P_n\left(\theta_1 \in \tilde{\mathrm{CI}}^L_{\alpha,n}\right) \geq (1-\alpha). \tag{100}$$

□