cemmap

centre for microdata methods and practice

Non-asymptotic inference in a class of optimization problems - July 2019

Joel L. Horowitz
Sokbae Lee

The Institute for Fiscal Studies
Department of Economics,
UCL

cemmap working paper CWP23/19

E·S·R·C
ECONOMIC
& SOCIAL
RESEARCH
COUNCIL

An ESRC Research Centre

NON-ASYMPTOTIC INFERENCE IN A CLASS OF OPTIMIZATION PROBLEMS

JOEL L. HOROWITZ[1] AND SOKBAE LEE[2] [3]

Abstract. This paper describes a method for carrying out non-asymptotic inference on partially identified parameters that are solutions to a class of optimization problems. The optimization problems arise in applications in which grouped data are used for estimation of a model's structural parameters. The parameters are characterized by restrictions that involve the population means of observed random variables in addition to the structural parameters of interest. Inference consists of finding confidence intervals for the structural parameters. Our method is non-asymptotic in the sense that it provides a finite-sample bound on the difference between the true and nominal probabilities with which a confidence interval contains the true but unknown value of a parameter. We contrast our method with an alternative non-asymptotic method based on the median-of-means estimator of Minsker (2015). The results of Monte Carlo experiments and an empirical example illustrate the usefulness of our method.

Keywords: partial identification, normal approximation, finite-sample bounds

JEL Classification: C12, C61.

[1] Department of Economics, Northwestern University, Evanston, IL 60208, USA.
[2] Centre for Microdata Methods and Practice, Institute for Fiscal Studies, 7 Ridgmount Street, London, WC1E 7AE, UK.
[3] Department of Economics, Columbia University, New York, NY 10027, USA.
E-mail addresses: joel-horowitz@northwestern.edu, sl3841@columbia.edu.

Date: July 3, 2019.
An earlier version of the paper was presented at the conference "Incomplete Models" on November 10, 2018 at Northwestern University. We would like to thank the conference participants for helpful comments. This work is in part supported by the European Research Council for financial support (ERC-2014-CoG-646917-ROMIA) and the UK Economic and Social Research Council for research grant (ES/P008909/1) to the CeMMAP.

1. Introduction

We present a method for carrying out non-asymptotic inference about a partially identified function of structural parameters of an econometric model. Our method applies to models that impose shape restrictions (e.g., Freyberger and Horowitz, 2015; Horowitz and Lee, 2017), a variety of partially identified models (e.g., Manski, 2007a; Tamer, 2010), and models in which a continuous function is inferred from the average values of variables in a finite number of discrete groups (e.g., Blundell, Duncan, and Meghir, 1998; Kline and Tartari, 2016). The specific inference problem consists of finding upper and lower bounds on the partially identified function $f(\psi)$ under the restrictions $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$, where ψ is a vector of structural parameters; μ is a vector of unknown population means of observable random variables; f is a known, real-valued function; and g_1 and g_2 are known possibly vector-valued functions. The inequality $g_1(\psi, \mu) \leq 0$ holds component-wise.

Most existing methods for inference in our framework are asymptotic. They provide correct inference in the limit $n \to \infty$ but do not provide information about the accuracy of finite-sample inference. Our method is non-asymptotic in the sense that it provides a finite-sample bound on the difference between the true and nominal coverage probabilities of a confidence interval for $f(\psi)$. In contrast to methods that provide only asymptotic inference, our results provide information about the accuracy of finite-sample inference. Canay and Shaikh (2017) and Ho and Rosen (2017) survey asymptotic inference in partially identified models. Chen, Christensen, and Tamer (2018) describe a Monte Carlo method for carrying out asymptotic inference for a class of models that includes our framework. Bugni, Canay, and Shi (2017) and Kaido, Molinari, and Stoye (2019) develop asymptotic inference methods for subvectors of partially identified parameters in moment inequality models. Hsieh, Shi, and Shum (2017) propose a method for asymptotic inference about estimators defined by mathematical programs. In contemporaneous work, Syrgkanis, Tamer, and Ziani (2018) consider finite-sample inference in auction models. Their framework and method are very different from those in this paper. In other settings that are also very different from ours, Chernozhukov, Hansen, and Jansson (2009) and Rosen and Ura (2019) propose finite-sample inference for quantile regression models and for the maximum score estimand, respectively.

There are several approaches to carrying out non-asymptotic inference (as defined in the previous paragraph) in our framework. In some cases, a statistic with a known finite-sample distribution makes finite-sample inference possible. For example, the

Clopper–Pearson (1934) confidence interval for a population probability is obtained by inverting the binomial probability distribution function. We use the Clopper-Pearson confidence interval in the empirical example presented in Section 5 of this paper. Manski (2007b) used the Clopper-Pearson interval to construct finite-sample confidence sets for counterfactual choice probabilities. A second method consists of using a finite-sample concentration inequality to obtain a confidence interval. This method is useful for applications only if the inequality provides a bound that does not depend on unknown population parameters. Hoeffding's inequality for the mean of a scalar random variable with known bounded support provides such a bound. Syrgkanis, Tamer, and Ziani (2018) used Hoeffding's inequality to construct a confidence interval for a partially identified population moment. Hoeffding's inequality gives confidence intervals that are wider than the intervals provided by the method of this paper and cannot be used if the (bounded) support of the underlying random variable is unknown. The generalization of Hoeffding's inequality to sub-Gaussian random variables requires information about a certain moment of the distribution of the underlying random variable that is typically unavailable in applications. Minsker (2015) developed a confidence set for a vector of population means using a method called "median of means." This method depends on certain tuning parameters. There are no data-based ways to choose these parameters in applications. Section 4 of this paper presents the results of Monte Carlo experiments comparing the widths of confidence intervals obtained by using Minsker's (2015) method and our method.

A third approach, which we use here, consists of making a normal approximation to the unknown distribution of a sample average. A variety of results provide finite-sample upper bounds on the errors made by normal approximations. The Berry-Esséen inequality for the average of a scalar random variable is a well-known example of such a bound. Bentkus (2003) provides a bound for the error of a multivariate normal approximation to the distribution of the sample average of a random vector. Other normal approximations are given by Spokoiny and Zhilova (2015) and Chernozhukov, Chetverikov, and Kato (2017); among others. The method described in this paper uses the normal approximation of Bentkus (2003); which does not require boundedness of the random variables involved; treats random vectors; and yields tighter bounds in our setting than do the methods of Spokoiny and Zhilova (2015) and Chernozhukov, Chetverikov, and Kato (2017). In contrast to conventional asymptotic inference approaches, our method provides a finite-sample bound on the difference between the true and nominal coverage probabilities of a confidence interval for the partially identified function $f(\psi)$.

The remainder of this paper is organized as follows. Section 2 presents our method for obtaining confidence intervals and describes two empirical studies that illustrate how the inferential problem the method addresses arises in applications. Section 3 describes computational procedures for implementing our method. Section 4 reports the results of a Monte Carlo investigation of the numerical performance of our method, and Section 5 presents an empirical application of the method. Section 6 gives concluding comments. Appendix A presents the proofs of theorems. Appendix B provides additional details on our computational procedures. Appendix C describes Minsker's (2015) median of means method.

2. The Method

Section 2.1 presents an informal description of inferential problem we address. Section 2.2 gives two examples of empirical applications in which the inferential problem arises. Section 2.3 provides a formal description of the method for constructing confidence intervals. Section 2.4 treats the possibility that g_1 and g_2 depend on a continuous covariate in addition to (ψ, μ).

2.1. The Inferential Problem.

Let $\{X_i : i = 1, \ldots, n\}$ be a random sample from the distribution of the random vector $X \in \mathbb{R}^p$ for some finite $p \geq 1$. Define $\mu = \mathbb{E}(X)$ and $\Sigma = \mathrm{cov}(X)$. Let ψ be a finite-dimensional parameter and $f(\psi)$ be a real-valued, known function. We assume throughout this section that f is only partially identified by the sampling process, though our results also hold if f is point identified. We seek a confidence interval for $f(\psi)$, which we define as an interval that contains $f(\psi)$ with probability exceeding a known value. Let $g_1(\psi, \mu)$ and $g_2(\psi, \mu)$ be possibly vector valued known functions satisfying $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$. Define

$$J_+ := \max_{\psi} f(\psi) \text{ and } J_- := \min_{\psi} f(\psi) \tag{2.1}$$

subject to the component-wise constraints:

$$g_1(\psi, \mu) \leq 0, \tag{2.2a}$$

$$g_2(\psi, \mu) = 0, \tag{2.2b}$$

$$\psi \in \Psi, \tag{2.2c}$$

where Ψ is a compact parameter set.

The tight identification region for $f(\psi)$ in this setting is $J_- \leq f(\psi) \leq J_+$. However, this interval cannot be calculated in applications because μ is unknown. Therefore, we estimate μ by the sample average $\bar{X} = n^{-1} \sum_{i=1}^{n} X_i$, and we estimate J_+ and J_- by

$$\hat{J}_+(\bar{X}) := \max_{\psi,m} f(\psi) \text{ and } \hat{J}_-(\bar{X}) := \min_{\psi,m} f(\psi) \tag{2.3}$$

subject to

$$g_1(\psi, m) \leq 0, \tag{2.4a}$$

$$g_2(\psi, m) = 0, \tag{2.4b}$$

$$\psi \in \Psi, \tag{2.4c}$$

and

$$n^{1/2}(\bar{X} - m) \in \mathcal{S}, \tag{2.4d}$$

where $\mathcal{S}$ is a set, specified in Section 2.3, for which $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ with high probability. Since μ is unknown, we replace it with a variable of optimization in (2.3)–(2.4) but restrict that variable to $\mathcal{S}$. The resulting confidence interval for $f(\psi)$ is

$$\hat{J}_-(\bar{X}) \leq f(\psi) \leq \hat{J}_+(\bar{X}). \tag{2.5}$$

Section 2.3 provides a finite-sample lower bound on the probability that this interval contains $f(\psi)$. That is, Section 2.3 provides a finite-sample lower bound on

$$\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right]. \tag{2.6}$$

2.2. **Examples of Empirical Applications.**

Example 1. Blundell, Duncan, and Meghir (1998) use grouped data to estimate labor supply effects of tax reforms in the United Kingdom. To motivate our setup, we consider a simple model with which Blundell, Duncan, and Meghir (1998) describe how to use grouped data to estimate β in the labor supply model with no income effect:

$$h_{it} = \alpha + \beta \ln w_{it} + U_{it}, \tag{2.7}$$

where h_{it} and w_{it}, respectively, are hours of work and the post-tax hourly wage rate of individual i in year t, and U_{it} is an unobserved random variable that satisfies certain conditions. The parameter β is identified by a relation of the form

$$\beta = \beta(h_{gt}, lw_{gt}),$$

where h_{gt} and lw_{gt} are the mean hours and log wages in year t of individuals in group g. There are 8 groups defined by four year-of-birth cohorts and level of education. The data span the period 1978-1992.

A nonparametric version of (2.7) is

$$h_{it} = f(w_{it}) + U_{it}, \tag{2.8}$$

where $f \in \mathcal{F}$ is an unknown continuous function and $\mathcal{F}$ is a function space. A nonparametric analog of β is the weighted average derivative

$$\tilde{\beta} = \int \frac{\partial f(u)}{\partial u} w(u) du,$$

where w is a non-negative weight function. The average derivative $\tilde{\beta}$ is not identified non-parametrically by the mean values of hours and wages for finitely many groups and time periods. It can be partially identified, however, by imposing a shape restriction such as weak monotonicity on the labor supply function g. Assume, for example, that $\mathbb{E}[h_{it} - f(w_{it})|g,t] = 0$. (Blundell, Duncan, and Meghir (1998) set $\mathbb{E}[h_{it} - f(w_{it})|g,t] = a_g + m_t$, where a_g and m_t, respectively, are group and time fixed effects. These are accommodated by our framework but we do not do this in the present discussion.)

The identification interval for $\tilde{\beta}$ is $\tilde{\beta}_- \leq \tilde{\beta} \leq \tilde{\beta}_+$, where

$$\tilde{\beta}_+ = \max_{f \in \mathcal{F}} \int \frac{\partial f(u)}{\partial u} w(u) du \quad \text{and} \quad \tilde{\beta}_- = \min_{f \in \mathcal{F}} \int \frac{\partial f(u)}{\partial u} w(u) du \tag{2.9}$$

subject to

$$f(w_{gt}) - f(w_{g't'}) \leq 0 \;\; \text{if } w_{gt} < w_{g't'}, \tag{2.10a}$$

$$h_{gt} - f(w_{gt}) = 0. \tag{2.10b}$$

The continuous mathematical programming problem (2.9)-(2.10) can be put into the finite-dimensional framework of (2.3)-(2.4) by observing that under mild conditions on $\mathcal{F}$, f can be approximated very accurately by the truncated infinite series

$$f(u) \approx \sum_{j=1}^{J} \psi_j \phi_j(u), \tag{2.11}$$

where the ψ_j's are constant parameters, the ϕ_j's are basis functions for $\mathcal{F}$, and J is a truncation point. In an estimation setting, J can be an increasing function of the sample size, though we do not undertake this extension here. The approximation error of (2.11) can be bounded. Here, however, we assume that J is sufficiently large to make the error negligibly small. The finite-dimensional analog of (2.9)-(2.10) is

(2.12)

$$J_+ = \max_{\psi_j : j=1,\ldots,J} \sum_{j=1}^{J} \psi_j \int \frac{\partial \phi_j(u)}{\partial u} w(u) du \ \text{ and } \ J_- = \min_{\psi_j : j=1,\ldots,J} \sum_{j=1}^{J} \psi_j \int \frac{\partial \phi_j(u)}{\partial u} w(u) du$$

subject to

$$\sum_{j=1}^{J} \psi_j \left[\phi_j(w_{gt}) - \phi_j(w_{g't'})\right] \leq 0 \ \text{ if } w_{gt} < w_{g't'}, \tag{2.13a}$$

$$h_{gt} - \sum_{j=1}^{J} \psi_j \phi_j(w_{gt}) = 0. \tag{2.13b}$$

J_+ and J_- can be estimated, thereby obtaining $\hat{J}_+$ and $\hat{J}_-$, by replacing h_{gt} and w_{gt} in (2.12)-(2.13) with within-group sample averages and adding the constraint (2.4d).

Example 2. Kline and Tartari (2016, KT hereafter) studied the impact of Connecticut's Jobs First (JF) welfare reform experiment on women's labor supply and welfare participation decisions. KT compared behavior under the JF and federal Aid to Families with Dependent Children (AFDC) regimes. The parameters of interest in KT are the probabilities with which a woman makes certain choices. The choice set under each regime (JF and AFDC) is denoted by $\{0n, 1n, 2n, 0r, 1r, 1u, 2u\}$, where 0 denotes no earnings, 1 denotes earnings below the poverty line, 2 denotes earnings above the poverty line, n denotes non-participation in welfare, r denotes welfare participation with truthful reporting of earnings, and u denotes welfare participation with under-reporting of earnings. Let π_{s^A,s^J} denote the probability of a woman's choosing alternative s^J under JF conditional on her choosing alternative s^A under AFDC. The possible choice probabilities and parameters of interest in KT are

$$\pi_{s^A,s^J} = \left[\pi_{0n,1r}, \pi_{0r,0n}, \pi_{2n,1r}, \pi_{0r,2n}, \pi_{0r,1r}, \pi_{0r,1n}, \pi_{1n,1r}, \pi_{0r,2u}, \pi_{2u,1r}\right]'.$$

The observable choices are welfare participation status and reported earnings under JF and AFDC. The population probabilities of observable choices are

$$\boldsymbol{p}^t := \left[p^t_{0n}, p^t_{1n}, p^t_{2n}, p^t_{0p}, p^t_{1p}, p^t_{2p}\right]'$$

for $t = A$ or J and the subscript p denotes welfare participation. These probabilities do not identify π_{s^A,s^J}.

To illustrate inference about the partially identified parameter π_{s^A,s^J} in our framework, consider the lower bound on $\pi_{2n,1r}$. By inequality (15) of KT,

$$\pi_{2n,1r} \geq \max\left\{0, \frac{p^A_{2n} - p^J_{2n}}{p^A_{2n}}\right\}.$$

Therefore,

$$p^A_{2n} - p^J_{2n} - p^A_{2n}\pi_{2n,1r} \leq 0.$$

Random sampling error is due to estimation of p^A_{2n} and p^J_{2n}, which are population moments. Let $\hat{p}^A_{2n}$ and $\hat{p}^J_{2n}$ be estimates of these moments. Then the estimated lower bound on $\pi_{2n,1r}$ is

$$\hat{J}_- = \min_{\psi, m_A, m_J} \psi \tag{2.14}$$

subject to

$$m_A - m_J - m_A\psi \geq 0, \tag{2.15a}$$

$$\psi \geq 0, \tag{2.15b}$$

$$\psi \in \Psi, \tag{2.15c}$$

$$[n^{1/2}(\hat{p}^A_{2n} - m_A), n^{1/2}(\hat{p}^J_{2n} - m_J)] \in \mathcal{S}. \tag{2.15d}$$

This example is continued in the empirical application of Section 5, where we specify the set $\mathcal{S}$ and find $\hat{J}_-$ satisfying $\mathbb{P}(\pi_{2n,1r} \geq \hat{J}_-) \geq 0.95$. Since ψ is a probability, we can set $\Psi = [0, 1]$ in this example.

2.3. **Analysis.**

This section presents a finite-sample lower bound on

$$\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right].$$

All proofs are in Appendix A. We begin with the following theorem, which forms the basis of our approach.

Theorem 2.1. *Assume that $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ for some ψ. Then*

$$\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right] \geq \mathbb{P}\left[n^{1/2}(\bar{X} - \mu) \in \mathcal{S}\right]. \tag{2.16}$$

Now define

$$Z_i := X_i - \mu \quad \text{and} \quad \bar{Z} := n^{-1/2}\sum_{i=1}^{n} Z_i = n^{-1/2}\sum_{i=1}^{n}(X_i - \mu).$$

Then $\mathbb{E}(\bar{Z}) = 0$. Define $\Sigma := \text{cov}(Z_i) = \text{cov}(\bar{Z}) = \text{cov}(X)$. If Σ is non-singular, let $[\Sigma^{-1/2}(X_i - \mu)]_j$ denote the j'th component of $\Sigma^{-1/2}(X_i - \mu)$ and Σ^{-1}_{jk} denote the (j, k) component of Σ^{-1}. Let μ_j denote the j'th component of μ and X_{ij} denote the j'th component of X_i. Make the following assumptions.

Assumption 1. *(i) $\{X_i : i = 1, \ldots, n\}$ is an independent random sample from the distribution of X. (ii) $\mathcal{S}$ is compact and convex. (iii) Ψ is compact. (iv) $f(\psi)$ is bounded on Ψ.*

Assumption 2. *(i) Σ is non-singular, and its components are all finite. (ii) There is a constant $\overline{\mu}_3 < \infty$ such that $\mathbb{E}\left(\left|[\Sigma^{-1/2}(X_i - \mu)]_j\right|^3\right) \leq \overline{\mu}_3$ for all $i = 1, \ldots, n$ and $j = 1, \ldots, p$. (iii) There is a constant $C_\Sigma < \infty$ such that $\left|\Sigma^{-1}_{jk}\right| \leq C_\Sigma$ for each $j, k = 1, \ldots, p$.*

Assumption 3. *There is a finite constant κ_1 such that*

$$\begin{aligned} \mathbb{E}\left[(X_{ij} - \mu_j)(X_{ik} - \mu_k)\right] &\leq \kappa_1, \\ \mathbb{E}\left[|X_{ij} - \mu_j|^r\right] &\leq \kappa_1^{r-1} r!, \\ \mathbb{E}\left[|(X_{ij} - \mu_j)(X_{ik} - \mu_k) - \Sigma_{jk}|^r\right] &\leq \kappa_1^{r-1} r! \end{aligned} \tag{2.17}$$

for every $r = 3, 4, 5, \ldots$ and $j, k = 1, \ldots, p$.

Assumption 3 requires the distribution of X to be thin-tailed. The assumption is satisfied, for example, if the distribution of X is sub-exponential.

Suppose for the moment that Σ is known. Define the independent random p-vectors $W_i \sim N(0, \Sigma)$ $(i = 1, \ldots, n)$ and $\bar{W} := n^{-1/2} \sum_{i=1}^n W_i \sim N(0, \Sigma)$. The multivariate generalization of the Lindeberg-Levy central limit theorem shows that $\bar{Z}$ is asymptotically distributed as $N(0, \Sigma)$, so the distribution of $\bar{Z}$ can be approximated by that of $\bar{W}$. The following lemma bounds the error of this approximation.

Lemma 2.1. *Let Assumptions 1, 2(i), and 2(ii) hold. Then*

$$\left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right| \leq \frac{400 p^{7/4} \overline{\mu}_3}{n^{1/2}}.$$

In applications, Σ is unknown. Let $\widehat{\Sigma}$ be the following estimator of Σ:

$$\widehat{\Sigma} := n^{-1} \sum_{i=1}^n X_i X_i' - \bar{X} \bar{X}'.$$

Define the random vector $\widehat{\bar{W}} \sim N(0, \widehat{\Sigma})$ and the function

$$r(t) := \left(\frac{6\kappa_1 t}{n}\right)^{1/2}$$

for $t > 0$. Approximate the distribution of $\bar{W}$ by the distribution of $\widehat{\bar{W}}$ with $\widehat{\Sigma}$ treated as a non-stochastic matrix. Define

$$w_n(t) := C_\Sigma p^3 2^{p+1} r(t).$$

The following lemma gives a finite-sample bound on the error of the approximation.

Lemma 2.2. *Let Assumptions 1-3 hold. Treat $\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})$ as if $\widehat{\Sigma}$ were a non-stochastic matrix. If $r(t) \geq 1$, then*

$$\left|\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right| \leq w_n(t)$$

with probability at least $1 - 4p^2 e^{-t}$.

The conclusion of Lemma 2.2 holds, that is,

$$\left|\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right| \leq w_n(t)$$

only if $\widehat{\Sigma}$ satisfies certain conditions that are stated in the proof of the lemma in Appendix A. These conditions are satisfied with probability at least $1 - 4p^2e^{-t}$, not with certainty.

Now combine Lemmas 2.1 and 2.2 to obtain the following theorem.

Theorem 2.2. *Let Assumptions 1-3 hold. Treat $\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})$ as if $\widehat{\Sigma}$ were a non-stochastic matrix. If $r(t) \geq 1$, then*

$$\left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})\right| \leq \frac{400 p^{7/4} \overline{\mu}_3}{n^{1/2}} + w_n(t) + 4p^2 e^{-t}.$$

Theorem 2.2 provides a finite-sample upper bound on the error made by approximating $\mathbb{P}\left[n^{1/2}(\bar{X} - \mu) \in \mathcal{S}\right]$ by $\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})$ with $\widehat{\Sigma}$ treated as a non-stochastic matrix. Combining Theorem 2.1 and 2.2 yields

Theorem 2.3. *Let Assumptions 1-3 hold and that $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ for some ψ. Treat $\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})$ as if $\widehat{\Sigma}$ were a non-stochastic matrix. If $r(t) \geq 1$, then*

$$\begin{aligned} &\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right] \\ &\geq \mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \left\{\frac{400 p^{7/4} \overline{\mu}_3}{n^{1/2}} + w_n(t) + 4p^2 e^{-t}\right\}. \end{aligned} \tag{2.18}$$

Theorem 2.3 provides a finite-sample lower bound on $\mathbb{P}\left[\hat{J}_-(\bar{X}) \leq J_- \leq f(\psi) \leq J_+ \leq \hat{J}_+(\bar{X})\right]$. Theorems 2.2 and 2.3 are the main results of this paper.

Like other large deviation bounds in statistics and the Berry-Esséen bound, the bounds in Theorems 2.2 and 2.3 can be loose unless n is large because they accommodate worst-case distributions of the observed variables. The numerical performance of our method in less extreme cases is illustrated in Section 4.

2.4. **Continuous Covariates.** In this section, we consider the case in which g_1 and g_2 depend on a continuous covariate ν in addition to (ψ, μ). This situation occurs, for example, in applications where some observed variables are group averages and others are continuously distributed characteristics of individuals. If ν is discrete, the results of Section 2.3 apply after replacing problem (2.3)-(2.4) with (2.21)-(2.22) below. When there is a continuous covariate, ν, (2.3)-(2.4) become

$$\hat{J}_+(\bar{X}) := \max_{\psi,m} f(\psi) \quad \text{and} \quad \hat{J}_-(\bar{X}) := \min_{\psi,m} f(\psi) \tag{2.19}$$

subject to

$$g_1(\psi, m, \nu) \leq 0 \text{ for every } \nu, \tag{2.20a}$$

$$g_2(\psi, m, \nu) = 0 \text{ for every } \nu, \tag{2.20b}$$

$$\psi \in \Psi, \tag{2.20c}$$

$$n^{1/2}(\bar{X} - m) \in \mathcal{S}. \tag{2.20d}$$

Thus, there is a continuum of constraints. We form a discrete approximation to (2.20a)-(2.20b) by restricting ν to a discrete grid of points. Let L denote the number of grid points. We give conditions under which the optimal values of the objective functions of the discretized version of (2.19)-(2.20) converge to $\hat{J}_+(\bar{X})$ and $\hat{J}_-(\bar{X})$ as $L \to \infty$. To minimize the notational complexity of the following discussion we assume that ν is a scalar. The generalization to a vector is straightforward. We also assume that ν is contained in a compact set which, without further loss of generality, we take to be $[0, 1]$.

To obtain the grid approximation, let $0 = x_0 < x_1 < x_2 < \ldots < x_L = 1$ be a grid of equally spaced points in $[0, 1]$. The distance between grid points is $1/(L-1)$. Approximate problem (2.19)-(2.20) by

$$\tilde{J}_+(\bar{X}) := \max_{\psi,m} f(\psi) \quad \text{and} \quad \tilde{J}_-(\bar{X}) := \min_{\psi,m} f(\psi) \tag{2.21}$$

subject to the constraints:

$$g_1(\psi, m, \nu_\ell) \leq 0;\ \ell = 1, \ldots, L, \tag{2.22a}$$

$$g_2(\psi, m, \nu_\ell) = 0;\ \ell = 1, \ldots, L, \tag{2.22b}$$

$$\psi \in \Psi, \tag{2.22c}$$

and

$$n^{1/2}(\bar{X} - m) \in \mathcal{S}. \tag{2.22d}$$

We then have

Theorem 2.4. *Assume that f is continuous, $\nu \in [0,1]$, and m in (2.22) is contained in a compact set $\mathcal{M}$. Moreover,*

$$|g_j(\psi, m; x) - g_j(\psi, m; x_\ell)| \leq C\,|x - x_\ell|$$
$$|g_j(\psi, m; x) - g_j(\psi, m; x_{\ell+1})| \leq C\,|x - x_{\ell+1}|$$

for $j = 1$ or 2, some $C < \infty$, and all $\psi \in \Psi$, all $m \in \mathcal{M}$, all $x \in [x_\ell, x_{\ell+1}] \in [0,1]$. Then

$$\lim_{L\to\infty} \tilde{J}_+ = \hat{J}_+ \quad \textit{and} \quad \lim_{L\to\infty} \tilde{J}_- = \hat{J}_-. \tag{2.23}$$

Theorem 2.4 implies that under weak smoothness assumptions, a sufficiently dense grid provides an arbitrarily accurate approximation to the continuously constrained optimization problem (2.19)-(2.20).

3. Computational Algorithms

Recall that our general framework is to obtain the bound

$$[\min_{\psi,m} f(\psi), \max_{\psi,m} f(\psi)]$$

subject to

$$g_1(\psi, m) \leq 0, g_2(\psi, m) = 0, \psi \in \Psi, \text{ and } n^{1/2}(\bar{X} - m) \in \mathcal{S}.$$

3.1. **Objective function $f(\psi)$.** In many examples, $f(\psi)$ is linear in ψ. For example, ψ is the vector of all the parameters in an econometric model and $f(\psi)$ is just one element of ψ or a linear combination of elements of ψ.

3.2. **Restrictions $g_1(\psi, \mu) \leq 0$, $g_2(\psi, \mu) = 0$, and $\psi \in \Psi$.** The restrictions $g_1(\psi, \mu) \leq 0$ include shape restrictions among the elements of ψ. Equality restrictions are imposed via $g_2(\psi, \mu) = 0$. The easiest case is that $g_j(\psi, \mu)$ is linear in (ψ, μ) for each

$j = 1, 2$. In some of examples we consider, $g_j(\psi, \mu)$ is linear in ψ, holding μ fixed, and linear in μ, keeping ψ fixed, but not linear in (ψ, μ) jointly. This corresponds to the case of bilinear constraints. For example, $g_j(\psi, \mu)$ may depend on the product between one of elements of ψ and one of elements of μ. In practice, Ψ can always be chosen large enough that the constraint $\psi \in \Psi$ is not binding and can be ignored.

3.3. **Restrictions $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$.** There are two leading cases of $\mathcal{S}$: an ellipsoid and a box. We start with the case that $\mathcal{S}$ is a box (that is, the Cartesian product of intervals). Let $\widehat{D}$ denote the diagonal matrix consisting of diagonal elements of $\widehat{\Sigma}$. Choose $\kappa(1-\alpha)$ such that

$$\sqrt{n} \max \left\{ \left| \widehat{D}_j^{-1/2} (\bar{X}_j - \mu_j) \right| : j = 1, \ldots, 2J \right\} \leq \kappa(1-\alpha)$$

with probability $1 - \alpha$. Here, the subscript j denotes the j-th element of a vector or the (j,j) element of a diagonal matrix. Note that when $\mathcal{S}$ is a box, the critical value can be easily simulated from the $N(0, \widehat{\Sigma})$ and the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as linear constraints.

Consider now the case that $\mathcal{S}$ is an ellipsoid. Choose $\kappa(1-\alpha)$ such that

$$n(\bar{X} - \mu)' \widehat{\Sigma}^{-1} (\bar{X} - \mu) \leq \kappa(1-\alpha)$$

with probability $1 - \alpha$.

When $\mathcal{S}$ is an ellipsoid, we consider two types of critical values. First, the critical value $\kappa(1-\alpha)$ can be obtained from the $\chi^2(J)$ distribution, where J is the dimension of μ. Second, it can be obtained via the bootstrap. We consider the $(1-\alpha)$ quantile of the bootstrap statistic

$$n(\bar{X}^* - \bar{X})' \left[\widehat{\Sigma}^* \right]^{-1} (\bar{X}^* - \bar{X}),$$

where $\bar{X}^*$ and $\widehat{\Sigma}^*$ are computed for each bootstrap sample. For both critical values, the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as

$$\mu' \widehat{\Sigma}^{-1} \mu - 2\mu' \widehat{\Sigma}^{-1} \bar{X} \leq n^{-1} \kappa(1-\alpha) - \bar{X}' \widehat{\Sigma}^{-1} \bar{X}.$$

This is a convex quadratic constraint in μ.

3.4. **Mathematical programming for leading cases.** Table 1 gives the scheme of mathematical programming we use for leading cases of $f(\psi)$, $g_1(\psi, \mu) \leq 0$, $g_2(\psi, \mu) = 0$, and $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$. In the table, LP, QP and QCP refer to linear programming, quadratic programming, and quadratically constrained programming, respectively. MILP, MIQP and MIQCP correspond to mixed integer linear programming, mixed integer quadratic programming, and mixed integer quadratically constrained programming, respectively.

Table 1. Class of Optimization Problems

Case	$f(\psi)$	$g_1(\psi,\mu) \leq 0$ $g_2(\psi,\mu) = 0$	$n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$	Programming
1	linear	linear	box	LP
2	linear	linear	ellipsoid	QCP
3	quadratic	linear	box	QP
4	quadratic	linear	ellipsoid	QCP
5	linear	bilinear	box	MILP/LP
6	linear	bilinear	ellipsoid	MIQCP/QCP
7	quadratic	bilinear	box	MIQP/QP
8	quadratic	bilinear	ellipsoid	MIQCP/QCP

When some of the constraints $g_1(\psi,\mu) \leq 0$ and $g_2(\psi,\mu) = 0$ are bilinear, the resulting problem may not be convex. To deal with non-convexity, we rely on a sequence of convex relaxations to obtain an outer bound for $f(\psi)$ and use a set of restricted inner bounds. When the union of restricted inner bounds matches the best outer bound by convex relaxations, we obtain the exact solution to the problem. Even if they do not match exactly, the best outer and inner bounds will give an approximate solution to the problem. The convex relaxations for bilinear constraints are implemented using mixed integer optimization (MIO). In Case 5, MILP/LP refers to the use of MILP for the outer bound and that of LP for the inner bound. Cases 6-8 are similar. Appendix B gives a detailed description of dealing with bilinear constraints. By virtue of the developments in MIO solvers and fast computing environments, the MIO has become increasingly used in recent applications. For example, Bertsimas, King, and Mazumder (2016) adopted an MIO approach for obtaining ℓ_0-constrained estimators in high-dimensional regression models and Reguant (2016) used mixed integer linear programming for computing counterfactual outcomes in game theoretic models.

4. Monte Carlo Experiments

4.1. **Identification Problem.** Suppose that

$$Y_i^* = h(Z_i) + e_i, \tag{4.1}$$

where $h : \mathbb{R} \mapsto \mathbb{R}$ is an unknown function and the error term e_i satisfies $E[e_i|Z_i] = 0$ almost surely. Assume that for each individual i, we do not observe Y_i^*, but only the interval data $[L_i, U_i]$ such that $Y_i^* \in [L_i, U_i]$ along with Z_i. Here, L_i and U_i are random variables.

Assume that the support of Z_i is finite, that is, $\mathcal{Z} \in \{z_1, \ldots, z_J\}$. Denote the values of $h(\cdot)$ on $\mathcal{Z}$ by $\{\psi_1, \ldots, \psi_J\}$. That is, $h(u) = \sum_{j=1}^{J} \psi_j 1(u = z_j)$ for $u \in \mathcal{Z}$.

Suppose that the object of interest is the value of $\psi^* \equiv h(z^*)$, where z^* is not in the support of Z_i but $z_{j-1} < z^* < z_j$ for some j. This type of extrapolation problem is given as a motivating example in Manski (2007a, pp. 4-5).

To partially identify ψ^*, assume that $h(\cdot)$ is monotone non-decreasing. Specifically, we impose the monotonicity on $\mathcal{Z} \cup \{z^*\}$. That is, $h(z_1) \leq h(z_2)$ whenever $z_1 \leq z_2$ for any $z_1, z_2 \in \mathcal{Z} \cup \{z^*\}$. In addition, we have the following inequality constraints:

$$E[L_i | Z_i = z_j] \leq \psi_j \leq E[U_i | Z_i = z_j] \tag{4.2}$$

for any $1 \leq j \leq J$. Note that (4.2) alone does not provide a bounded interval containing ψ^* since z^* is not in $\mathcal{Z}$. The monotonicity assumption combined with (4.2) provides an informative bound on ψ^*.

To write the optimization problem in our canonical form, let μ denote the population moments of the following $\bar{X}$:

$$n\bar{X} = \begin{pmatrix} \sum_{i=1}^{n} L_i 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} L_i 1(Z_i = z_J) \\ \sum_{i=1}^{n} U_i 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} U_i 1(Z_i = z_J) \\ \sum_{i=1}^{n} 1(Z_i = z_1) \\ \vdots \\ \sum_{i=1}^{n} 1(Z_i = z_J) \end{pmatrix}.$$

Then, we can rewrite the constraints (4.2) in a bilinear form:

$$E[L_i 1(Z_i = z_j)] \leq \psi_j E[1(Z_i = z_j)] \leq E[U_i 1(Z_i = z_j)] \tag{4.3}$$

for any $1 \leq j \leq J$. To deal with the bilinear constraints, we rely on a method called piecewise McCormick relaxation, which is given in Appendix B.

4.2. **Results of a Monte Carlo Experiment.** Suppose that (4.1) holds with $h(z) = 2z$, the covariate Z_i is uniformly distributed on

$$\mathcal{Z} = \{-3/2, -1, -1/2, 1/2, 1, 3/2\},$$

and $e_i \sim \text{Unif}[-1/2, 1/2]$. The interval data are generated from $L_i = Y_i^* + V_i 1(V_i < 0)$ and $U_i = Y_i^* + V_i 1(V_i \geq 0)$, where $V_i \sim N(0,1)$. Here, V_i and U_i are independent of each other. The parameter of interest is $\psi^* = h(0)$. Note that zero is not included in $\mathcal{Z}$. The monotonicity constraint is imposed as

$$\begin{pmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 0 \\ -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & -1 \end{pmatrix} \begin{pmatrix} h(0) \\ h(-3/2) \\ h(-1) \\ h(-1/2) \\ h(1/2) \\ h(1) \\ h(3/2) \end{pmatrix} \leq \mathbf{0}_{6\times 1}.$$

The simulation design here is similar to that of Bontemps, Magnac, and Maurin (2012) except that the support of X_i is discrete and the linearity of $h(\cdot)$ is not used in estimation. The sample size was 200, 500, 1000 and 2000. There were 100 repetitions for each Monte Carlo experiment. The identified set containing $h(0)$ is $-1.40 \leq h(0) \leq 1.40$. The reported coverage probability is the frequency that the estimated lower bound is smaller than or equal to the true lower bound (-1.40) and the estimated upper bound is greater than or equal to the true upper bound (1.40). The nominal coverage probability was 0.95.

We consider both cases that $\mathcal{S}$ is a box and an ellipsoid. We take $\Psi = \{\psi_j \in [-5, 5] \text{ for each } j = 1, \ldots, J\}$. The outer bounds were computed with piecewise linear relaxations with $K = 10$ that is described in Appendix B. To describe how to obtain inner bounds, first partition μ and $\bar{X}$ into $\mu \equiv (\mu_1, \mu_2)$ and $\bar{X} \equiv (\bar{X}_1, \bar{X}_2)$, where $\mu_1 \equiv (E[L_i 1(Z_i = z_1)], \ldots, E[L_i 1(Z_i = z_J)], E[U_i 1(Z_i = z_1)], \ldots, E[U_i 1(Z_i = z_J)])$, $\mu_2 \equiv (E[1(Z_i = z_1)], \ldots, E[1(Z_i = z_J)])$, and $\bar{X}_1$ and $\bar{X}_2$ are corresponding sample moments. In other words, only components of μ_2 appear as bilinear terms in (4.3) and those of μ_1 are linearly separable. Then, to obtain a lower bound, we fix μ_2 at its feasible value and optimize with respect to (ψ, μ_1). When $\mathcal{S}$ is a box, it is straightforward to obtain a lower bound since the feasible value of μ_1 does not depend on that of μ_2. When $\mathcal{S}$ is an ellipsoid, recall that the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as

$$\mu' \widehat{\Sigma}^{-1} \mu - 2\mu' \widehat{\Sigma}^{-1} \bar{X} \leq n^{-1} \kappa(1-\alpha) - \bar{X}' \widehat{\Sigma}^{-1} \bar{X}. \tag{4.4}$$

Following μ and $\bar{X}$, partition $\widehat{\Sigma}^{-1}$ into blocks such that $\widehat{\Sigma}^{-1} \equiv \{\widehat{\Sigma}^{-1}_{(k,\ell)}, k = 1, 2, \ell = 1, 2\}$. Now rewrite (4.4) as

$$\begin{aligned} &\mu_1' \widehat{\Sigma}^{-1}_{(1,1)} \mu_1 + 2\mu_1' \left[\widehat{\Sigma}^{-1}_{(1,2)} \mu_2 - \widehat{\Sigma}^{-1}_{(1,1)} \bar{X}_1 - \widehat{\Sigma}^{-1}_{(1,2)} \bar{X}_2 \right] \\ &\leq n^{-1} \kappa(1-\alpha) - \bar{X}' \widehat{\Sigma}^{-1} \bar{X} - \mu_2' \widehat{\Sigma}^{-1}_{(2,2)} \mu_2 + 2\mu_2' \left[\widehat{\Sigma}^{-1}_{(2,1)} \bar{X}_1 + \widehat{\Sigma}^{-1}_{(2,2)} \bar{X}_2 \right]. \end{aligned} \tag{4.5}$$

Given μ_2, this is a convex, quadratic constraint. First, we generate a random grid of μ_2 using the box version of $\mathcal{S}$. Then we optimize with respect to (ψ, μ_1) under the restrictions $g(\psi, \mu) \leq 0$ and (4.5). In both cases, the inner bounds were obtained with a random grid search with $G = 1000$. As an alternative to the outer and inner bounds, we also consider the bounds when μ_2 is fixed at $\bar{X}_2$. These correspond to the bounds when the observed covariates $Z_1, \ldots, Z_n$ are regarded as non-stochastic. These bounds will be tighter than those constructed under the random design. In each Monte Carlo repetition when $\mathcal{S}$ is a box, the number of simulations to draw $N(0, \widehat{\Sigma})$ was 1,000. The χ^2 critical value is used when $\mathcal{S}$ is an ellipsoid.

Table 2 presents the simulation results. First, we comment on the results with a box. When $n = 200$, there are minor discrepancies between the outer and inner bounds. For all other large sample sizes, averages of the bounds are identical. As the sample size increases, the length of the estimated bounds decreases rapidly. We now look at the results with an ellipsoid. The estimated bounds are much tighter with an ellipsoid than with a box. There are more noticeable differences between the average values of the outer and inner bounds when $\mathcal{S}$ is an ellipsoid. However, these differences shrink as the sample size gets larger. The bounds with fixed Z_i's are much tighter if n is smaller or if a box is used for $\mathcal{S}$. The empirical coverage probabilities are all larger than the nominal probability of 0.95. This is consistent with Theorem 2.3, which provides a lower bound on the coverage probability, not a point probability.

4.3. **Comparison with Minsker's Method.** In this subsection, we provide the results of a small Monte Carlo experiment that is designed to compare our main proposal with Minsker's method described in Appendix C. We make two changes in the experimental design of Section 4.2. First, in addition to the standard normal distribution, the errors V_i are generated from the t-distribution with degrees of freedom equal to 3 to consider a fat-tailed distribution. Note that the fat-tailed distribution is adopted here since the median-of-means approach is robust to outliers. Second, we set $n = 10000$ or 20000 because Minsker's method requires a relatively large sample size. The estimated bounds obtained with Minsker's method were uninformative

Table 2. Results of Monte Carlo Experiments

Type of $\mathcal{S}$	Type of bounds	sample size	Avg. of estimated lower bound	Avg. of estimated upper bound	Coverage probability
Box	Outer bounds ($K = 10$)	200	-3.87	3.78	1
		500	-2.59	2.54	1
		1000	-2.14	2.14	1
		2000	-1.89	1.89	1
Box	Inner bounds ($G = 1000$)	200	-3.89	3.75	1
		500	-2.59	2.54	1
		1000	-2.14	2.14	1
		2000	-1.89	1.89	1
Box	Fixing $Z_1, \ldots, Z_n$	200	-2.07	2.12	1
		500	-1.83	1.84	1
		1000	-1.71	1.69	1
		2000	-1.62	1.61	1
Ellipsoid	Outer bounds ($K = 10$)	200	-2.34	2.35	1
		500	-1.90	1.91	1
		1000	-1.78	1.76	1
		2000	-1.67	1.66	1
Ellipsoid	Inner bounds ($G = 1000$)	200	-2.25	2.19	1
		500	-1.84	1.81	1
		1000	-1.68	1.68	1
		2000	-1.59	1.59	1
Ellipsoid	Fixing $Z_1, \ldots, Z_n$	200	-1.92	1.93	0.99
		500	-1.72	1.73	1
		1000	-1.64	1.62	1
		2000	-1.57	1.56	1

Note: "Box" and "Ellipsoid" are our proposed methods with a box S and an ellipsoidal S, respectively.

when $n = 200, 500, 1000, 2000$. Specifically, they were $[-5, 5]$ for all of these sample sizes. We consider coordinate-wise medians for the median of means since μ is relatively low-dimensional. For simplicity, we consider only the outer bounds with 10 repetitions for each experiment.

Table 3 reports the experimental results. The true lower and upper bounds are -1.40 and 1.40 for the $N(0, 1)$ errors and -1.37 and 1.37 for the $t(3)$ errors, respectively. The estimated bounds for Minsker's method are much wider than the bounds estimated by our proposed method, although they shrink from $n = 10000$ to

Table 3. Comparison with Minsker's Method

Distribution of V_i	Type of $\mathcal{S}$	Type of bounds	sample size	Avg. of estimated lower bound	Avg. of estimated upper bound	Coverage probability
$N(0, 1)$	Ellipsoid	Outer	10000	-1.52	1.53	1
		bounds	20000	-1.49	1.48	1
$N(0, 1)$	Minsker	Outer	10000	-3.87	3.86	1
		bounds	20000	-3.20	3.22	1
$t(3)$	Ellipsoid	Outer	10000	-1.71	1.73	1
		bounds	20000	-1.67	1.67	1
$t(3)$	Minsker	Outer	10000	-4.19	4.17	1
		bounds	20000	-3.48	3.46	1

Note: "Ellipsoid" is our proposed method with an ellipsoidal S.

$n = 20000$. Moreover, Minsker's method does not produce a better result for the t-distribution. Even with the t-distribution, Minsker's method gives wider bounds than our method does. The bounds from the ellipsoid, which is our proposed method, provides much tighter bounds but they also seem conservative, as noted in the previous section.

5. An Empirical Example

This section provides an empirical example based on the study of KT that is described in Section 2.2. Specifically, we use the information in Table 4 of KT to obtain the set $\mathcal{S}$ in (2.15d) and the lower endpoint of a 95% confidence bound for $\pi_{2n,1r}$. We consider only the lower endpoint because KT found the upper endpoint to be 1 and, therefore, uninformative.

KT used the JF welfare reform experimental data and pooled all person-quarter observations in the seven quarters following randomization of participants. They treated each person-quarter observation as a potentially separate decision, allowing time-varying behaviors. Because assignment of individuals to the JF treatment and AFDC control groups was random, we assume that observations in each regime are independent of the observations in the other. We further assume that observations within the JF and AFDC regimes are independently and identically distributed (iid). The set $\mathcal{S}$ can be expressed as a confidence region for p_{2n}^A and p_{2n}^J. We used the

Clopper–Pearson (1934) procedure to construct the rectangular 95% confidence region

$$L_A \leq p_{2n}^A \leq U_A,$$
$$L_J \leq p_{2n}^J \leq U_J,$$

where L_A, U_A, L_J, and U_J are random lower and upper bounds chosen so that

$$\mathbb{P}\left(L_A \leq p_{2n}^A \leq U_A; L_J \leq p_{2n}^J \leq U_J\right) \geq 0.95.$$

The sample sizes in KT are $n_A = 16,268$ and $n_J = 16,226$ for the AFDC and JF regimes, respectively. The estimated values of p_{2n}^A and p_{2n}^J are $\hat{p}_{2n}^A = 0.099$ and $\hat{p}_{2n}^J = 0.068$. The resulting confidence region for p_{2n}^A and p_{2n}^J is $0.092 \leq p_{2n}^A \leq 0.106$ and $0.062 \leq p_{2n}^J \leq 0.074$. Solving (2.14)-(2.15) yielded 0.195 as the lower endpoint of a 95% confidence region for $\pi_{2n,1r}$. KT obtained a lower endpoint of 0.198, which is similar to ours. However, there are important differences between our method and that of KT. KT used a block bootstrap procedure that resamples a woman's entire profile of choices for the first seven quarters after randomization, whereas we used the non-asymptotic inference method for iid data. The method of KT relies on asymptotic arguments but allows serial dependence of the observations of the same woman. Our method is valid for any sample size but does not take account of any serial dependence. It turns out that the conservative nature of our inference method is offset by our assumption of independence, thereby yielding a confidence interval of approximately the same size as that of KT.

6. Conclusions

This paper has described a method for carrying out non-asymptotic inference on partially identified parameters that are solutions to a class of optimization problems. These problems arise, for example, in applications in which grouped data are used for estimation of a model's structural parameters. Inference consists of finding confidence intervals for the structural parameters. The method is non-asymptotic in the sense that it provides a finite-sample bound on the difference between the true and nominal probabilities with which a confidence interval contains the true but unknown value of a parameter. The paper has described computational algorithms for implementing the method. The results of Monte Carlo experiments and an empirical example illustrate the method's usefulness.

Appendix A. Proofs of Theorems

Proof of Theorem 2.1. Suppose $n^{1/2}(\bar{X}-\mu)$ is in $\mathcal{S}$. Any feasible solution of (2.3)–(2.4) is also a feasible solution of (2.1)–(2.2). Therefore, the feasible region of (2.1)–(2.2) contains the feasible region of (2.3)–(2.4). Consequently,

$$\hat{J}_-(\bar{X}) \le J_- \le J_+ \le \hat{J}_+(\bar{X}),$$

which in turn proves (2.16). □

Proof of Lemma 2.1. Define the random p-vector $\bar{V} := \Sigma^{-1/2}\bar{Z}$. Then $\mathbb{E}(\bar{V}) = 0$ and $\mathrm{cov}(\bar{V}) = I_{p\times p}$. Define the set

$$\mathcal{S}_\Sigma := \left\{\xi : \xi = \Sigma^{-1/2}\zeta; \zeta \in \mathcal{S}\right\}. \tag{A.1}$$

Then $\mathcal{S}_\Sigma$ is a convex set. Define the random vectors $U_i \sim N(0, I_{p\times p})$ and $\bar{U} := n^{1/2}\sum_{i=1}^n U_i$. It follows from Assumption 2(ii) and the generalized Minkowski inequality that

$$\mathbb{E}\left[\left(\sum_{j=1}^p [\Sigma^{-1/2}(X_i-\mu)]_j^2\right)^{3/2}\right] \le p^{3/2}\overline{\mu}_3.$$

In addition, it follows from Theorem 1.1 of Bentkus (2003) that

$$\left|\mathbb{P}(\bar{V} \in \mathcal{S}_\Sigma) - \mathbb{P}(\bar{U} \in \mathcal{S}_\Sigma)\right| \le \frac{400p^{7/4}\overline{\mu}_3}{n^{1/2}},$$

which proves the lemma. □

Proof of Lemma 2.2. Define

$$\Delta_n := \sup_{\mathcal{S}} \left|\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right|.$$

Now

$$\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S}) = \mathbb{P}(\Sigma^{-1/2}\widehat{\bar{W}} \in \mathcal{S}_\Sigma) - \mathbb{P}(\xi \in \mathcal{S}_\Sigma),$$

where $\xi \sim N(0, I_{p\times p})$ and $\mathcal{S}_\Sigma$ is defined in (A.1). Therefore,

$$\begin{aligned}\Delta_n &= \sup_{\mathcal{S}} \left|\mathbb{P}(\Sigma^{-1/2}\widehat{\bar{W}} \in \mathcal{S}_\Sigma) - \mathbb{P}(\xi \in \mathcal{S}_\Sigma)\right| \\ &\le \mathrm{TV}\left[N(0, I_{p\times p}), N(0, \Sigma^{-1}\widehat{\Sigma})\right],\end{aligned}$$

where $\mathrm{TV}(P_1, P_2)$ is the total variation distance between distributions P_1 and P_2. By Example 2.3 of Dasgupta (2008),

$$\mathrm{TV}\left[N(0, I_{p\times p}), N(0, \Sigma^{-1}\widehat{\Sigma})\right] \le p2^{p+1}\left\|\Sigma^{-1}\widehat{\Sigma} - I_{p\times p}\right\|_{\mathrm{F}}$$

where for any matrix A,

$$\|A\|_{\mathrm{F}}^2 := \sum_{j=1}^{p} \sum_{j=1}^{p} a_{jk}^2.$$

Define $\omega := \widehat{\Sigma} - \Sigma$. Then,

$$\Sigma^{-1}\widehat{\Sigma} - I_{p\times p} = \Sigma^{-1}(\widehat{\Sigma} - \Sigma) = \Sigma^{-1}\omega,$$

$$\left|(\Sigma^{-1}\omega)_{jk}\right| \leq \sum_{\ell=1}^{p} \left|\Sigma_{j\ell}^{-1}\omega_{\ell k}\right| \leq C_\Sigma \sum_{\ell=1}^{p} |\omega_{\ell k}|$$

and

$$\left\|\Sigma^{-1}\widehat{\Sigma} - I_{p\times p}\right\|_{\mathrm{F}} \leq C_\Sigma p^{1/2} \left[\sum_{k=1}^{p}\left(\sum_{\ell=1}^{p} |\omega_{k\ell}|\right)^2\right]^{1/2}.$$

To obtain the conclusion of the lemma, it remains to show that $|\omega_{jk}| \leq r(t)$ with probability at least $1 - 4p^2e^{-t}$. We prove this claim below.

Write

$$\omega = n^{-1}\sum_{i=1}^{n} [(X_i - \mu)(X_i - \mu)' - \Sigma] - (\bar{X} - \mu)(\bar{X} - \mu)'.$$

By Bernstein's inequality,

$$\mathbb{P}\left[\left|\bar{X}_j - \mu_j\right| \geq r(t)\right] \leq 2\exp\left(-\frac{nr(t)^2}{2\kappa_1[2 + r(t)]}\right)$$

for each $j = 1, \ldots, p$ and

$$\mathbb{P}\left\{n^{-1}\left|\sum_{i=1}^{n} [(X_{ij} - \mu)(X_{ik} - \mu)' - \Sigma_{jk}]\right| \geq r(t)\right] \leq 2\exp\left(-\frac{nr(t)^2}{2\kappa_1[2 + r(t)]}\right)$$

for each (j, k) with $j, k = 1, \ldots, p$. Therefore, if $r(t) \geq 1$,

$$\mathbb{P}\left[\left|\bar{X}_j - \mu_j\right| \geq r(t)\right] \leq 2\exp\left(-\frac{nr(t)^2}{6\kappa_1}\right)$$

for each $j = 1, \ldots, p$ and

$$\mathbb{P}\left\{n^{-1}\left|\sum_{i=1}^{n} [(X_{ij} - \mu)(X_{ik} - \mu)' - \Sigma_{jk}]\right| \geq r(t)\right] \leq 2\exp\left(-\frac{nr(t)^2}{6\kappa_1}\right)$$

for each (j, k) with $j, k = 1, \ldots, p$. However,

$$r(t)^2 = \frac{6\kappa_1 t}{n}.$$

Therefore,

$$\mathbb{P}\left[\left|\bar{X}_j - \mu_j\right| \geq r(t)\right] \leq 2e^{-t}$$

for each $j = 1, \dots, p$ and

$$\mathbb{P}\left\{ n^{-1} \left| \sum_{i=1}^{n} [(X_{ij} - \mu)(X_{ik} - \mu)' - \Sigma_{jk}] \right| \geq r(t) \right] \leq 2e^{-t}$$

for each (j, k) with $j, k = 1, \dots, p$. Thus,

$$\mathbb{P}\left[\max_{j,k} |\omega_{jk}| \leq r(t) \right] \geq 1 - 4p^2 e^{-t},$$

which proves the claim. □

Proof of Theorem 2.2. Write

$$\begin{aligned}\left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(\widehat{\bar{W}} \in \mathcal{S})\right| &= \left| \left[\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right] - \left[\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right] \right| \\ &\leq \left|\mathbb{P}(\bar{Z} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right| + \left|\mathbb{P}(\widehat{\bar{W}} \in \mathcal{S}) - \mathbb{P}(\bar{W} \in \mathcal{S})\right|.\end{aligned}$$

Thus, the theorem follows immediately by combining Lemmas 2.1 and 2.2. □

Proof of Theorem 2.3. Combining Theorem 2.1 and 2.2 yields Theorem 2.3. □

Proof of Theorem 2.4. We focus on the maximization problem since the minimization problem can be analyzed analogously.

Let (ψ_L, μ_L) denote the optimal solution to the maximization version of (2.21)-(2.22). Define $g(\psi, \mu; \nu) = [g_1(\psi, \mu; \nu), g_2(\psi, \mu; \nu), -g_2(\psi, \mu; \nu)]$, so that $g(\psi, \mu; \nu) \leq 0$ componentwise. Define $\ell(\nu) := \arg\min_{\ell} |\nu - \nu_\ell|$. Then

$$\sup_{\nu \in [0,1]} |g(\psi, \mu; \nu) - g(\psi, \mu; \ell(\nu))| \leq C/(L-1)$$

and $g(\psi, \mu; \nu_\ell) \leq 0$ implies that

$$g(\psi, \mu; \nu) \leq C/(L-1)$$

componentwise uniformly over $\nu \in [0, 1]$. Therefore, (ψ_L, μ_L) is a feasible solution to

$$J_{+}^{*} := \max_{\psi,\mu} f(\psi) \tag{A.2}$$

subject to the new constraint:

$$g(\psi, \mu; \nu) \leq C/(L-1) \text{ for all } \nu \in [0,1],\ \nu = \text{rational, and } \ n^{1/2}(\bar{X} - \mu) \in \mathcal{S}.$$

Consequently, $J_+^* \geq \tilde{J}_+ \geq J_+$, where $J_+ = \max f(\psi)$ subject to $g_1(\psi, \mu, x) \leq 0$, $g_2(\psi, \mu, x) = 0$, and $\psi \in \Psi$. Define

$$\Pi := \Big\{ \xi \geq 1, \eta : \text{there is } (\psi, \mu) \text{ such that } n^{1/2}(\bar{X} - \mu) \in \mathcal{S},\ f(\psi) \leq \eta, \\ \text{and } g(\varphi, \mu; \nu) \leq C/\xi \text{ for all } \nu \in [0,1] \Big\}.$$

Note that Π is a closed set. Therefore, by Proposition 3.3 of Jeyakumar and Wolkowicz (1990), $J_+^* \to J_+$ as $L \to \infty$ if the constraints are restricted to rational values of $\nu \in [0,1]$. It follows from continuity of g as a function of ν that the constraints hold for all $\nu \in [0,1]$. □

Appendix B. Details about Computation with Bilinear Constraints

To explain how to deal with the constraints in a bilinear form, suppose that we have a cross product term $\mu_j \psi_\ell$ in $g(\psi, \mu) \leq 0$ for some j and ℓ, where $\mu = (\mu_1, \ldots, \mu_J)'$ and $\psi = (\psi_1, \ldots, \psi_L)'$.

The existence of the bilinear term $\mu_j \psi_\ell$ can make the corresponding optimization problem non-convex. As mentioned in the main text, we rely on a sequence of convex relaxations to obtain an outer bound for $f(\psi)$. Specifically, we use piecewise-linear relaxations that are called piecewise McCormick relaxation in the operation research and engineering literature.

There exist a number of different formulations for piecewise McCormick relaxations. For instance, Gounaris, Misener, and Floudas (2009) applied 15 different formulations. We follow the formulation called 'nf4l' in Gounaris, Misener, and Floudas (2009). This formulation was one of recommended formulations in Gounaris, Misener, and Floudas (2009). To simplify the notation, we will drop dependence on the subscripts and write $\mu_j \psi_\ell$ as $\mu\psi$. In practice, one has to apply piecewise McCormick relaxation to each bilinear term.

For any two positive terms $a \in [0, \overline{a}]$ and $b \in [0, \overline{b}]$, McCormick relaxation of $c \equiv ab$ consists of the following four inequalities:

$$c \geq 0,\ c \geq a\overline{b} + \overline{a}b - \overline{a}\overline{b},\ c \leq a\overline{b},\ c \leq \overline{a}b, \tag{B.1}$$

which is known as the tightest possible convex relaxation.

To explain how to apply McCormick relaxation to $\mu\psi$, we introduce a new variable φ and replace $\mu\psi$ with φ. Then instead of imposing the bilinear constraint that $\varphi = \mu\psi$, we relax this in a piecewise fashion.

Suppose that ψ belongs to a known interval $[\underline{\psi}, \overline{\psi}]$. Assume that $\mu \in [\underline{\mu}, \overline{\mu}]$ with known end points. In practice, they can be deduced from $\mathcal{S}$ since $\mathcal{S}$ will be imposed simultaneously.

We now partition the space $[\underline{\psi}, \overline{\psi}]$ for ψ by a grid of $(K+1)$ points $\{m_k : k = 0, \dots, K, m_0 = \underline{\psi}, m_K = \overline{\psi}\}$. Define λ_k to be a set of binary variables such that

$$\lambda_k = \begin{cases} 1 & \text{if } m_{k-1} \leq \psi \leq m_k \\ 0 & \text{otherwise} \end{cases}$$

for $k = 1, \dots, K$. Since we would like to ensure that ψ belongs to only one of segments $[m_{k-1}, m_k]$, we impose the summing up constraint such that

$$\sum_{k=1}^{K} \lambda_k = 1. \tag{B.2}$$

To reflect that $[\underline{\psi}, \overline{\psi}]$ is partitioned as described above, we introduce a set of continuous variables δ_k, $k = 1, \dots, K$, where $0 \leq \delta_k \leq (m_k - m_{k-1})$. Then we impose the following set of restrictions

$$\begin{aligned} \psi &= \sum_{k=1}^{K} \{m_{k-1}\lambda_k + \delta_k\}, \\ 0 &\leq \delta_k \leq (m_k - m_{k-1})\lambda_k \ \forall k. \end{aligned} \tag{B.3}$$

It can be seen that $\delta_k = 0$ if $\lambda_k = 0$ and $\delta_k = \psi - m_{k-1}$ for the index k such that $\lambda_k = 1$. For μ, we also introduce a set of continuous variables η_k, $k = 1, \dots, K$, where $0 \leq \eta_k \leq (\overline{\mu} - \underline{\mu})$. Impose the following restrictions

$$\begin{aligned} \mu &= \underline{\mu} + \sum_{k=1}^{K} \eta_k, \\ 0 &\leq \eta_k \leq (\overline{\mu} - \underline{\mu})\lambda_k \ \forall k. \end{aligned} \tag{B.4}$$

As before, $\eta_k = 0$ if $\lambda_k = 0$ and $\eta_k = \mu - \underline{\mu}$ for the index k such that $\lambda_k = 1$.

Using newly defined variables δ_k and η_k, we now write

$$\varphi = \underline{\mu}\psi + \sum_{k=1}^{K} m_{k-1}\eta_k + \sum_{k=1}^{K} \delta_k \eta_k. \tag{B.5}$$

The first two terms on the right-hand side of (B.5) are linear in ψ and η_k; whereas, the third term involves K bilinear terms of $\delta_k \eta_k$. Applying McCormick relaxation

(B.1) to $\delta_k\eta_k$ gives four inequalities for each k:

$$\begin{aligned}
&\delta_k\eta_k \geq 0,\\
&\delta_k\eta_k \geq (m_k - m_{k-1})\eta_k + (\mu - \underline{\mu})\delta_k - (\mu - \underline{\mu})(m_k - m_{k-1}),\\
&\delta_k\eta_k \leq (\mu - \underline{\mu})\delta_k,\\
&\delta_k\eta_k \leq (m_k - m_{k-1})\eta_k.
\end{aligned} \tag{B.6}$$

Instead of introducing a k-specific variable for each $\delta_k\eta_k$, define a single continuous variable Δ, where $0 \leq \Delta \leq \max_{k=1,\dots,K}(m_k - m_{k-1})(\overline{\mu} - \underline{\mu})$. Then rewrite (B.5) as

$$\varphi = \underline{\mu}\psi + \sum_{k=1}^{K} m_{k-1}\eta_k + \Delta \tag{B.7}$$

and aggregate equations in (B.6) over k to yield the following restrictions

$$\begin{aligned}
&\Delta \geq \sum_{k=1}^{K}(m_k - m_{k-1})\eta_k + (\overline{\mu} - \underline{\mu})\left(\sum_{k=1}^{K}\left[\delta_k - (m_k - m_{k-1})\lambda_k\right]\right),\\
&\Delta \leq (\overline{\mu} - \underline{\mu})\sum_{k=1}^{K}\delta_k,\\
&\Delta \leq \sum_{k=1}^{K}(m_k - m_{k-1})\eta_k.
\end{aligned} \tag{B.8}$$

In summary, the formulation of piecewise McCormick relaxation consists of (B.2), (B.3), (B.4), (B.7), and (B.8). The variables of optimization are μ, ψ, φ, Δ, $\lambda_k \in \{0, 1\}$, $\delta_k \in [0, (m_k - m_{k-1})]$, $\eta_k \in [0, (\overline{\mu} - \underline{\mu})]$, where $k = 1, \dots, K$. The total number of variables for optimization has increased from 2 to $4+3K$, but a bilinear constraint is relaxed to mixed integer linear constraints. A modern optimization solver (e.g. Gurobi) can handle efficiently mixed integer linear constraints.

We now describe how to construct inner bounds. Recall that $(\psi, \mu) \in [\underline{\psi}, \overline{\psi}] \times [\underline{\mu}, \overline{\mu}]$. When the bilinear term $\mu\psi$ exists in the optimization problem and we fix μ at one of values on its feasible set, the corresponding constrained optimization problem becomes convex but sup-optimal. Hence, solving the constrained optimization problem yields an inner bound. To obtain a tighter inner bound, we can create a grid of points for possible values of μ with size G and solve a constrained problem at each value of the grid. Taking the union of all these inner bounds gives a tight inner bound.

Note that K and G are tuning parameters to choose in implementation. To implement the method described above, we can start with small K and G and increase K

and G gradually up to the point that the set difference between the resulting outer and inner bounds is negligible up to some tolerance level. Even if the algorithm does not converge in a fixed time, we can compute the gap between the outer and inner bounds. This optimality gap is useful for evaluating the quality of the solution.

We state the proposed algorithm as follows.

Algorithm 1: Algorithm for outer and inner bounds

```
1. Select the type of $\mathcal{S}$ and choose tuning parameters $K$ and $G$.
2. Obtain the outer bounds by solving $[\min_{\psi,\mu} f(\psi), \max_{\psi,\mu} f(\psi)]$ subject to
   (B.9)    $g(\psi,\mu) \leq 0, \psi \in \Psi, \text{ and } n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$,
   while replacing each incidence of a bilinear term with the formulation of $K$-piecewise McCormick relaxation consisting of (B.2), (B.3), (B.4), (B.7), and (B.8).
3. Construct a $G$-dimensional grid for components of $\mu$, say $\mu_2$, appearing in the problem as bilinear terms. Obtain the lower bounds by solving $[\min_{\psi,\mu} f(\psi), \max_{\psi,\mu} f(\psi)]$ subject to (B.9), while fixing $\mu_2$ at a fixed value of the grid points. Take the union of all $G$ inner bounds to construct the best inner bounds.
4. If the gap between outer and inner bounds is small, terminate. If not, increase $K$ and $G$ to see whether the gap can decrease further. Repeat the last step only fixed number of times.
5. Report the resulting outer and inner bounds.
```

Appendix C. Minsker's (2015) Median of Means Method

In this appendix, we carry out non-asymptotic inference based on Minsker (2015). In particular, we consider two versions of the median of means: the one based on geometric median and the other using coordinate-wise medians. Lugosi and Mendelson (2019) propose a different version of the median of means estimator that has theoretically better properties but is more difficult to compute.

First, for the case of geometric median, let $\alpha_* := 7/18$ and $p_* := 0.1$. Define

$$\psi(\alpha_*; p_*) = (1-\alpha_*) \log \frac{1-\alpha_*}{1-p_*} + \alpha_* \log \frac{\alpha_*}{p_*}.$$

Let $0 < \delta < 1$ be the level of the confidence set and set

$$k := \left\lfloor \frac{\log(1/\delta)}{\psi(\alpha_*; p_*)} \right\rfloor + 1. \tag{C.1}$$

Assume that δ is small enough that $k \leq n/2$. Divide the sample $X_1, \ldots, X_n$ into k disjoint groups $G_1, \ldots, G_k$ of size $\lfloor \frac{n}{k} \rfloor$ each, and define

$$\hat{\mu}_j := \frac{1}{|G_j|} \sum_{i \in G_j} X_i, \ j = 1, \ldots, k,$$
$$\hat{\mu} := \text{G.med}(\hat{\mu}_1, \ldots, \hat{\mu}_k),$$

where G.med refers to the geometric median. See Minsker (2015) and references therein for details on the geometric median. The intuition behind $\hat{\mu}$ is that it is a robust measure of the population mean vector μ since each subsample mean vector $\hat{\mu}_j$ is an unbiased estimator for μ and the aggregation method via the geometric median is robust to outliners. Because of this feature, it turns out that the finite sample bound for the Euclidean norm distance between $\hat{\mu}$ and μ depends only on $\text{tr}(\Sigma)$, but not on the higher moments (see Corollary 4.1 of Minsker, 2015). This is the main selling point of the median of means since the finite sample probability bound for the usual sample mean assumes the existence of a higher moment (e.g. the third absolute moment in Bentkus (2003) and Lemma 2.1 in Section 2.3).

Second, Minsker (2015) also considered using coordinate-wise medians instead of using the geometric median. In this case, let $\alpha_* = 1/2$ and $p_* = 0.12$. Then k is redefined via (C.1). Let $\hat{\mu}_*$ denote the vector of coordinate-wise medians.

To estimate $\text{tr}(\Sigma)$, Minsker (2015) proposed the following:

$$\hat{T}_j := \frac{1}{|G_j|} \sum_{i \in G_j} \|X_i - \hat{\mu}_j\|^2, \ j = 1, \ldots, k,$$
$$\hat{T} := \text{med}(\hat{T}_1, \ldots, \hat{T}_k).$$

where $\|a\|$ is the Euclidean norm of a vector a. Let $B(h, r)$ denote the ball of radius r centered at h and let

$$r_n := 11\sqrt{2}\sqrt{\hat{T}\frac{\log(1.4/\delta)}{n}},$$
$$r_{n,*} := 4.4\sqrt{2}\sqrt{\hat{T}\frac{\log(1.6d_\mu/\delta)}{n - 2.4\log(1.6d_\mu/\delta)}},$$

where d_μ is the dimension of μ.

Lemma C.1 (Minsker (2015))**.** *Assume that*

$$15.2\sqrt{\frac{\mathbb{E}\|X - \mu\|^4 - (tr(\Sigma))^2}{(tr(\Sigma))^2}} \leq \left(\frac{1}{2} - 178\frac{\log(1.4/\delta)}{n}\right)\sqrt{\frac{n}{\log(1.4/\delta)}}. \tag{C.2}$$

Then

$$\mathbb{P}\left[\mu \in B(\hat{\mu}, r_n)\right] \geq 1 - 2\delta \quad \textit{and} \quad \mathbb{P}\left[\mu \in B(\hat{\mu}_*, r_{n,*})\right] \geq 1 - 2\delta. \tag{C.3}$$

Proof of Lemma C.1. The result on the geometric median is the exactly the same as Corollary 4.2 of Minsker (2015). The case for the vector of coordinate-wise medians follows from combining equation (4.4) in Minsker (2015) with Proposition 4.1 of Minsker (2015). □

Lemma C.1 indicates that $\mathcal{S}$ in our setup can be chosen as

$$(\hat{\mu} - \mu)'(\hat{\mu} - \mu) \leq r_n^2,$$

or

$$(\hat{\mu}_* - \mu)'(\hat{\mu}_* - \mu) \leq r_{n,*}^2,$$

either of which gives the bound with probability at least $1 - 2\delta$. The former produces a tighter bound than the latter only when the dimension of μ is sufficiently high. Note that (C.2) requires the existence of fourth moments due to the fact that $\mathrm{tr}(\Sigma)$ is estimated by the median of means as well. The inequality in (C.2) is a relatively mild condition when n is large. In Section 4, we provide a numerical comparison between our main proposal and Minsker's method.

References

Bentkus, V. (2003): "On the dependence of the Berry–Esseen bound on dimension," *Journal of Statistical Planning and Inference*, 113(2), 385–402.

Bertsimas, D., A. King, and R. Mazumder (2016): "Best subset selection via a modern optimization lens," *Annals of Statistics*, 44(2), 813–852.

Blundell, R., A. Duncan, and C. Meghir (1998): "Estimating Labor Supply Responses Using Tax Reforms," *Econometrica*, 66(4), 827–861.

Bontemps, C., T. Magnac, and E. Maurin (2012): "Set Identified Linear Models," *Econometrica*, 80(3), 1129–1155.

Bugni, F. A., I. A. Canay, and X. Shi (2017): "Inference for subvectors and other functions of partially identified parameters in moment inequality models," *Quantitative Economics*, 8(1), 1–38.

Canay, I. A., and A. M. Shaikh (2017): "Practical and Theoretical Advances in Inference for Partially Identified Models," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and

L. Samuelson, vol. 2 of *Econometric Society Monographs*, pp. 271–306. Cambridge University Press.

Chen, X., T. M. Christensen, and E. Tamer (2018): "Monte Carlo Confidence Sets for Identified Sets," *Econometrica*, 86(6), 1965–2018.

Chernozhukov, V., D. Chetverikov, and K. Kato (2017): "Central limit theorems and bootstrap in high dimensions," *Annals of Probability*, 45(4), 2309–2352.

Chernozhukov, V., C. Hansen, and M. Jansson (2009): "Finite sample inference for quantile regression models," *Journal of Econometrics*, 152(2), 93–103.

Clopper, C. J., and E. S. Pearson (1934): "The Use of Confidence or Fiducial Limits Illustrated in the Case of the Binomial," *Biometrika*, 26(4), 404–413.

Dasgupta, A. (2008): *Asymptotic Theory of Statistics and Probability.* Springer, New York.

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

Gounaris, C. E., R. Misener, and C. A. Floudas (2009): "Computational comparison of piecewise–linear relaxations for pooling problems," *Industrial & Engineering Chemistry Research*, 48(12), 5742–5766.

Ho, K., and A. M. Rosen (2017): "Partial Identification in Applied Research: Benefits and Challenges," in *Advances in Economics and Econometrics: Eleventh World Congress*, ed. by B. Honoré, A. Pakes, M. Piazzesi, and L. Samuelson, vol. 2 of *Econometric Society Monographs*, pp. 307–359. Cambridge University Press.

Horowitz, J. L., and S. Lee (2017): "Nonparametric estimation and inference under shape restrictions," *Journal of Econometrics*, 201(1), 108–126.

Hsieh, Y.-W., X. Shi, and M. Shum (2017): "Inference on Estimators defined by Mathematical Programming," arXiv:1709.09115.

Jeyakumar, V., and H. Wolkowicz (1990): "Zero duality gaps in infinite-dimensional programming," *Journal of Optimization Theory and Applications*, 67(1), 87–108.

Kaido, H., F. Molinari, and J. Stoye (2019): "Confidence Intervals for Projections of Partially Identified Parameters," *Econometrica*, forthcoming.

Kline, P., and M. Tartari (2016): "Bounding the Labor Supply Responses to a Randomized Welfare Experiment: A Revealed Preference Approach," *American Economic Review*, 106(4), 972–1014.

Lugosi, G., and S. Mendelson (2019): "Sub-Gaussian estimators of the mean of a random vector," *Annals of Statistics*, 47(2), 783–794.

Manski, C. F. (2007a): *Identification for Prediction and Decision.* Harvard University Press, Cambridge, Massachusetts.

——— (2007b): "Partial Identification of Counterfactual Choice Probabilities," *International Economic Review*, 48(4), 1393–1410.

Minsker, S. (2015): "Geometric median and robust estimation in Banach spaces," *Bernoulli*, 21(4), 2308–2335.

Reguant, M. (2016): "Bounding Outcomes in Counterfactual Analysis," Northwestern University Working Paper.

Rosen, A. M., and T. Ura (2019): "Finite Sample Inference for the Maximum Score Estimand," arXiv:1903.01511 [econ.EM].

Spokoiny, V., and M. Zhilova (2015): "Bootstrap confidence sets under model misspecification," *Annals of Statistics*, 43(6), 2653–2675.

Syrgkanis, V., E. Tamer, and J. Ziani (2018): "Inference on Auctions with Weak Assumptions on Information," arXiv:1710.03830 [econ.EM].

Tamer, E. (2010): "Partial Identification in Econometrics," *Annual Review of Economics*, 2(1), 167–195.