Inference in a Class of Optimization Problems: Confidence Regions and Finite Sample Bounds on Errors in Coverage Probabilities

Joel L. Horowitz\textsuperscript{a} and Sokbae Lee\textsuperscript{b}

\textsuperscript{a}Department of Economics, Northwestern University, Evanston, IL; \textsuperscript{b}Department of Economics, Columbia University, New York, NY

\textbf{ABSTRACT}
This article describes three methods for carrying out nonasymptotic inference on partially identified parameters that are solutions to a class of optimization problems. Applications in which the optimization problems arise include estimation under shape restrictions, estimation of models of discrete games, and estimation based on grouped data. The partially identified parameters are characterized by restrictions that involve the unknown population means of observed random variables in addition to structural parameters. Inference consists of finding confidence intervals for functions of the structural parameters. Our theory provides finite-sample lower bounds on the coverage probabilities of the confidence intervals under three sets of assumptions of increasing strength. With the moderate sample sizes found in most economics applications, the bounds become tighter as the assumptions strengthen. We discuss estimation of population parameters that the bounds depend on and contrast our methods with alternative methods for obtaining confidence intervals for partially identified parameters. The results of Monte Carlo experiments and empirical examples illustrate the usefulness of our method.

\textbf{1. Introduction}
This article presents three methods for carrying out nonasymptotic inference about a function of partially identified structural parameters of an econometric model. The methods apply 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) that include discrete games (e.g., Ciliberto and Tamer 2009), 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 \( \psi \) is a vector of structural parameters, \( \mu \) 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 based on asymptotic approximations. They provide correct inference in the limit of an infinite sample size but do not provide information about the accuracy of the asymptotic approximations in finite samples. We provide three methods for obtaining finite-sample lower bounds on the coverage probability of a confidence interval for \( f(\psi) \). One method uses asymptotic approximations to obtain a confidence interval. The other two methods do not use asymptotic approximations. All the methods provide information about the accuracy of finite-sample inference.

There are several approaches to carrying out nonasymptotic inference in our framework. Sometimes a statistic with a known finite-sample distribution makes finite-sample inference possible. For example, the Clopper and Pearson (1934) confidence interval for a population probability is obtained by inverting the binomial probability distribution function. Manski (2007b) used the Clopper–Pearson interval to construct finite-sample confidence sets for counterfactual choice probabilities. Our methods apply to parameters that are not necessarily probabilities. A second existing method consists of using Hoeffding's inequality to obtain a confidence interval. Syrgkanis, Tamer, and Ziani (2018) used this inequality to construct a confidence interval for a partially identified population moment. Hoeffding's inequality requires the underlying random variable to have a known bounded support. Our methods do not require the underlying random variable to have a known or bounded support. In addition, Minsker's method depends on certain user-selected tuning parameters. There are data-based, efficient ways to choose these parameters in applications.

Our first method consists of making a normal approximation to the unknown distribution of the sample average. This method
makes certain assumptions about low-order moments of the underlying random variable but does not restrict its distribution in other ways. A variety of results provide finite-sample upper bounds on the errors made by normal approximations. The Berry-Esseen inequality for the average of a scalar random variable is a well-known example of such a bound. Bentkus (2003) provides a bound on the error of a multivariate normal approximation to the distribution of the sample average of a random vector. Other normal approximations for random vectors are given by Spokoiny and Zhilova (2015), Chernozhukov, Chetverikov, and Kato (2017), and Zhilova (2020); among others. Our first method uses a bound on the error of the multivariate normal approximation that is due to Raič (2019). Raič’s (2019) bound is a refined and tighter version of the bound of Bentkus (2003).

The bound of Chernozhukov, Chetverikov, and Kato (2017) may be tighter than that of Raič (2019) when the dimension of μ exceeds the sample size, but the bound of Raič (2019) is tighter when the dimension of μ is small compared to the sample size, which is the case we treat in this paper. In contrast to conventional asymptotic inference approaches, our first method provides a finite-sample lower bound on the coverage probability of a confidence interval for the partially identified function \( f(\psi) \).

The bound provided by our first method is loose in samples of the moderate sizes that occur in most economics applications, though not necessarily in very large samples. This is because it places only weak restrictions on the distribution of the underlying random variable, which may be far from normal. Our second method obtains a tighter bound in moderate size samples by assuming that the distributions of the components of the possibly vector-valued underlying random variable are sub-Gaussian. The sub-Gaussian assumption places stronger restrictions on the thickness of the tails of the relevant distributions than do the assumptions of the bound based on Raič’s (2019) inequality. Our third method tightens the bound obtained with our second method by assuming that if the underlying random variable is vector-valued, then its distribution is sub-Gaussian in a vector sense. This assumption is stronger than the assumption that the components of a random vector are individually sub-Gaussian. The bounds obtained with the second and third methods are identical if the underlying random variable is a scalar.

The bounds provided by all the methods depend on unknown population parameters. This dependence is unavoidable and can be removed only in special cases. The parameters of the bounds of the second and third methods can be estimated, however, which makes it possible to estimate the bounds in applications. We describe how to do this. The resulting estimated bounds are asymptotic. They do not have finite-sample validity but can provide useful, though possibly rough, indications of the magnitudes of the finite-sample bounds. We present the results of Monte Carlo experiments that illustrate the relation between the exact finite-sample bounds and the consistent estimates.

Our work is broadly related to the literature on inference in partially identified models. Tamer (2010), Canay and Shaikh (2017), Ho and Rosen (2017), and Molinari (2020) provide recent surveys. 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. Chernozhukov, Chetverikov, and Kato (2019) and Belloni, Bugni, and Chernozhukov (2018) construct confidence regions by inverting pointwise tests of a hypothesis about the (sub)vector of parameters that are partially identified by a large number of moment inequalities. The inference problem we treat is different from those in the foregoing papers in that we focus on inference about parameters that are solutions to a class of optimization problems that is different from moment inequality problems. Our methods and results do not apply to moment inequalities. Kline and Tamer (2016) describe Bayesian inference in a class of models that includes a special case of the models we treat. Two more closely related papers are Hsieh, Shi, and Shum (2017) and Shi and Shum (2015), who propose a method for asymptotic inference about estimators defined by mathematical programs. However, the class of estimation problems they treat is different from ours and overlaps with ours only under highly restrictive assumptions about both classes. Hsieh, Shi, and Shum (2017) and Shi and Shum (2015) do not provide finite-sample bounds on the errors of their asymptotic approximations.

Our work is also related to the econometrics literature on finite-sample inference. Syrgkanis, Tamer, and Zian (2018) consider finite-sample inference in auction models. Their framework and method are very different from those in this article. In a different context, 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. Their methods and the classes of models they treat are distinct from ours.

The remainder of this article is organized as follows. Section 2 describes the inferential problem we treat, our methods for obtaining a confidence interval for \( f(\psi) \), and the three methods for obtaining a finite-sample lower bound on the coverage probability of a confidence interval. Section 2 also describes two empirical studies that illustrate how the inferential problem arises in applications. Section 3 describes computational procedures for implementing our methods. Section 4 presents an empirical application of the methods. Section 5 reports the results of a Monte Carlo investigation of the numerical performance of our methods, and Section 6 gives concluding comments. The proofs of theorems are presented in Appendix A, supplementary materials. Appendices B–E, supplementary materials provide additional technical information about our methods, a description of Minsker’s (2015) method, an additional empirical application, and additional Monte Carlo results.

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 our methods for constructing confidence intervals and bounds on coverage probabilities.

2.1. The Inferential Problem

Let \( \{X_i : i = 1, \ldots, n\} \) be an independent 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 \( \Sigma = \text{cov}(X). \) We assume that both exist. Let \( \psi \) be a finite-dimensional parameter and \( f(\psi) \) be a real-valued, known function. We assume throughout this section that \( f(\psi) \) is only partially identified by the sampling process, though our results also hold if \( f(\psi) \) is point identified. We seek a confidence interval for \( f(\psi) \), which we define as a data-based 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 \( g_2(\psi, \mu) = 0 \) component-wise. Define

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

subject to the component-wise constraints:

\[
g_1(\psi, \mu) \leq 0, \quad g_2(\psi, \mu) = 0, \quad \psi \in \Psi, \tag{2.2}
\]

where \( \Psi \) is a compact parameter set. Appendix B, supplementary materials extends (2.1)–(2.2) to the case in which \( g_1 \) and \( g_2 \) depend on a continuous covariate in addition to \( \mu \) and \( \psi \).

We are interested in the identification interval \( J_- \leq f(\psi) \leq J_+ \). However, this interval cannot be calculated in applications because \( \mu \) is unknown. Therefore, we estimate \( \mu \) by the sample average \( \bar{X} = n^{-1} \sum_{i=1}^{n} X_i \), and we estimate \( J_+ \) and \( J_- \) by

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

subject to the constraints

\[
g_1(\psi, m) \leq 0, g_2(\psi, m) = 0, \quad \psi \in \Psi, \tag{2.4a}
\]

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

where \( \mathcal{S} \) is a set for which \( n^{1/2}(\bar{X} - \mu) \in \mathcal{S} \) with high probability. Since \( \mu \) is unknown, we replace it with the variable of optimization \( m \) in (2.3)–(2.4a) but require \( m \) to satisfy (2.4b).

The resulting confidence interval for \( f(\psi) \) is

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

This is also a confidence interval for the identified set containing \( f(\psi) \). Section 2.3 provides three different finite-sample lower bounds on the probability that this interval contains \( f(\psi) \). That is, Section 2.3 provides three finite-sample lower bounds on

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

The three bounds correspond to increasingly strong assumptions about the distribution of \( X \) and are increasingly tight with samples of the moderate sizes found in most economics applications, though not necessarily with very large samples.

The two leading examples of \( \mathcal{S} \) in constraint (2.4b) are a box and an ellipsoid. If \( \mathcal{S} \) is a box, let \( D \) be a diagonal matrix whose diagonal elements are strictly positive. For example, \( D \) might be the diagonal elements of \( \Sigma \) if \( \Sigma \) is known or the diagonal elements of a consistent estimate, \( \hat{\Sigma} \), if \( \Sigma \) is unknown. Choose \( \kappa_b(1 - \alpha) \) so that the following holds, uniformly in \( j = 1, \ldots, p \), with probability \( 1 - \alpha \):

\[
|\bar{X}_j - \mu_j| \leq \kappa_b(1 - \alpha)(nD)_{jj}^{1/2},
\]

where the subscript \( j \) denotes the \( j \)th component of a vector or the \((j,j)\) component of a matrix. In this case, (2.4b) becomes \( p \) constraints and can be viewed as a sample analog of \( |\mathbb{E}(X_j) - \mu_j| \leq 0 \) with a relaxed constraint on \( \mu_j \) for each \( j = 1, \ldots, p \). Section 3 presents methods for choosing \( \kappa_b(1 - \alpha) \).

If \( \mathcal{S} \) is an ellipsoid, let \( \Upsilon \) denote a positive definite \( p \times p \) matrix, possibly \( \Sigma \) or \( \hat{\Sigma} \) if those matrices are nonsingular, or the identity matrix. Choose \( \kappa_b(1 - \alpha) \) so that

\[
n(\bar{X} - \mu)^\Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_b(1 - \alpha)
\]

with probability \( 1 - \alpha \). In this case, (2.4b) is a single constraint. Section 3 presents a method for choosing \( \kappa_b(1 - \alpha) \). When \( \mathcal{S} \) is difficult to estimate or is singular, we may use a sphere by choosing a critical value \( \kappa_b(1 - \alpha) \) such that

\[
n(\bar{X} - \mu)^\Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_b(1 - \alpha)
\]

with probability \( 1 - \alpha \). In general, the implementation of our methods is simpler if \( \mathcal{S} \) is indexed by a scalar critical value \( \kappa_b(1 - \alpha) \).

It is straightforward to allow the objective function \( f(\psi) \) to depend on \( \mu \). For the lower bound \( \hat{J}_- (\bar{X}) \), we introduce an auxiliary variable \( t \) that acts as an upper bound on \( f(\psi, \mu) \) and solve: min \( \psi, m, t \) subject to \( f(\psi, m) \leq t \) and (2.4). For the upper bound \( \hat{J}_+ (\bar{X}) \), we introduce a lower bound \( s \) on \( f(\psi, \mu) \) and solve: max \( \psi, m, s \) subject to \( f(\psi, m) \geq s \) and (2.4). We focus on the original form (2.3)–(2.4) in the remainder of this article because the form with the objective function \( f(\psi, \mu) \) can be rewritten in the form (2.3)–(2.4) by redefining \( f, g_1 \), and \( g_2 \).

### 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) explain how to use grouped data to estimate \( \beta \) in the following labor supply model with no income effect:

\[
h_{it} = \alpha + \beta \log w_{it} + U_{it}. \tag{2.7}
\]

In this model, \( h_{it} \) 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 \( \beta \) is identified by a relation of the form \( \beta = \beta(h_{it}, \log w_{it}) \), where \( h_{it} \) and \( \log w_{it} \) are the mean hours and log wages in year \( t \) of individuals in group \( p \). There are eight 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 \( h_{it} = \xi(w_{it}) + U_{it} \), where \( \xi \in \Xi \) is an unknown continuous function and \( \Xi \) is a function space. A nonparametric analog of \( \beta \) is the weighted average derivative

\[
\tilde{\beta} = \int \frac{\partial \xi(u)}{\partial u} w(u) du,
\]

where \( w \) is a nonnegative weight function. The average derivative \( \beta \) is not identified nonparametrically by the mean values of hours and wages for finitely many groups and time periods.
can be partially identified, however, by imposing a shape restriction such as weak monotonicity on the labor supply function \( \xi \). Assume, for example, that \( \mathbb{E}[h_{i} - \xi(w_{i})] \geq 0 \). Blundell, Duncan, and Meghir (1998) set \( \mathbb{E}[h_{i} - \xi(w_{i})] = \alpha_{q} + m_{t} \), where \( \alpha_{q} \) and \( m_{t} \), respectively, are group and time fixed effects. These are accommodated by our framework but we do not insist on this in the present discussion.

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

\[
\tilde{\beta}_{+} = \max_{\xi \in \Xi} \int \frac{\partial \xi(u)}{\partial u} w(u) \, du \quad \text{and} \quad \tilde{\beta}_{-} = \min_{\xi \in \Xi} \int \frac{\partial \xi(u)}{\partial u} w(u) \, du
\]

subject to

\[
\xi(w_{gt}) - \xi(w_{gt'}) \leq 0 \quad \text{if} \quad w_{gt} < w_{gt'}, \quad (2.8a)
\]

\[
h_{gt} - \xi(w_{gt}) = 0. \quad (2.8b)
\]

The continuous mathematical programming problem (2.8)–(2.9) can be put into the finite-dimensional framework of (2.3)–(2.4) by observing that under mild conditions on \( \Xi \), \( \xi \) can be approximated very accurately by the truncated infinite series

\[
\xi(u) \approx \sum_{j=1}^{K} \psi_{j}(\phi_{j}(u)), \quad (10.10)
\]

where the \( \psi_{j} \)'s are constant parameters, the \( \phi_{j} \)'s are basis functions for \( \Xi \), and \( K \) is a truncation point. In an estimation setting, \( K \) can be an increasing function of the sample size, though we do not undertake this extension here. The approximation error of (10.10) can be bounded. However, we assume that \( K \) is sufficiently large to make the error negligibly small. The finite-dimensional analog of (2.8)–(2.9) is

\[
\begin{align*}
J_{+} & = \max_{\psi_{j} = 1, \ldots, K} \sum_{j=1}^{K} \psi_{j} \int \frac{\partial \phi_{j}(u)}{\partial u} w(u) \, du \quad \text{and} \\
J_{-} & = \min_{\psi_{j} = 1, \ldots, K} \sum_{j=1}^{K} \psi_{j} \int \frac{\partial \phi_{j}(u)}{\partial u} w(u) \, du
\end{align*} \quad (11.01)
\]

subject to

\[
\begin{align*}
\sum_{j=1}^{K} \psi_{j} \left[ \phi_{j}(w_{gt}) - \phi_{j}(w_{gt'}) \right] & \leq 0 \quad \text{if} \quad w_{gt} < w_{gt'}, \quad (11.02a) \\
h_{gt} - \sum_{j=1}^{K} \psi_{j} \phi_{j}(w_{gt}) & = 0. \quad (11.02b)
\end{align*}
\]

\( J_{+} \) and \( J_{-} \) can be estimated, thereby obtaining \( \hat{J}_{+} \) and \( \hat{J}_{-} \), by replacing \( h_{gt} \) and \( w_{gt} \) in (11.01)–(11.02) with within-group sample averages and adding the constraint (2.4b).

**Example 2**

Ho and Pakes (2014, HP hereinafter) use the theory of revealed preference to develop an estimator of hospital choices by individuals. HP use data on privately insured births in California. We consider a simplified version of the HP model.

Using the notation of HP, let \( p(c, h) \) denote the price an insurer is expected to pay at hospital \( h \) for a patient with medical condition \( c \). Let \( i \) index patients. Then \( c_{i} \) is the medical condition of patient \( i \), and \( p(c_{i}, h) \) is the price an insurer is expected to pay at hospital \( h \) for patient \( i \). Let \( l_{t} \) denote patient \( i \)'s location, \( h_{l} \) hospital's location, and \( d(\cdot, \cdot) \) the distance between the two locations. For hospitals \( h \neq h' \), define

\[
\Delta p(c_{i}, h, h') := p(c_{i}, h) - p(c_{i}, h') \quad \text{and} \quad \Delta d(l_{i}, l_{h}, l_{h'}) := d(l_{i}, l_{h}) - d(l_{i}, l_{h'}).
\]

That is, \( \Delta p(c_{i}, h, h') \) is the price difference between hospitals \( h \) and \( h' \) given patient condition \( c_{i} \) and \( \Delta d(l_{i}, l_{h}, l_{h'}) \) is the distance difference between hospitals \( h \) and \( h' \) given patient location \( l_{i} \).

Define the four-dimensional vector of instruments based on distance:

\[
z_{i}(h, h') := \left( \begin{array}{c}
\text{max}\left\{ \Delta d(l_{i}, l_{h}, l_{h'}), 0 \right\} \\
\text{min}\left\{ \Delta d(l_{i}, l_{h}, l_{h'}), 0 \right\} \\
\text{max}\left\{ \Delta d(l_{i}, l_{h}, l_{h'}), 0 \right\} \\
\text{min}\left\{ \Delta d(l_{i}, l_{h}, l_{h'}), 0 \right\}
\end{array} \right)
\]

Here, the instruments are based on distance measures and constructed to be positive to preserve the signs of the inequalities below in (2.13).

Let \( S(h, h', s) \) be the set of patients with severity \( s \) who chose hospital \( h \) but had hospital \( h' \) in their choice set. The identifying assumption in HP is that

\[
\mathbb{E}\left[ u_{i}(\psi, h, h') z_{i}(h, h') \mid i \in S(h, h', s), i' \in S(h', h, s) \right] \geq 0 \quad (2.13)
\]

for all \( s, h, h' \) such that \( h \neq h' \). We can rewrite (2.13) as

\[
\psi_{\mu}(h, h', s) - \mu_{d}(h, h', s) \geq 0 \quad \text{for all} \ (h, h', s) \text{ such that} \ h \neq h',
\]

where

\[
\mu_{P}(h, h', s) := \mathbb{E}\left[ z_{i}(h, h') \left\{ \Delta p(c_{i}, h, h') + \Delta p(c_{i}, h', h') \right\} \mid i \in S(h, h', s), i' \in S(h', h, s) \right],
\]

\[
\mu_{d}(h, h', s) := \mathbb{E}\left[ z_{i}(h, h') \left\{ \Delta d(l_{i}, l_{h}, l_{h'}) + \Delta d(l_{i}, l_{h'}, l_{h}) \right\} \mid i \in S(h, h', s), i' \in S(h', h, s) \right].
\]

To see the connection between our general framework and HP's inequality estimator, let \( f(\psi) = \psi, \mu = (\mu_{P}, \mu_{d}) \), and \( g_{1} \) be a collection of inequalities such that

\[
g_{1}(\psi, \mu) = \mu_{d}(h, h', s) - \psi_{\mu}(h, h', s) \leq 0 \quad \text{for all} \ (h, h', s) \text{ such that} \ h \neq h'.
\]

There is no element in \( g_{2} \) (no equality constraints here). Since each element in \( \mu \) can be estimated by a suitable sample mean, our general framework includes HP's estimator as a special case.
2.3. Analysis

This section presents our three methods for forming finite-sample lower bounds on

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

The three bounds make assumptions of differing strengths about the distribution of \( X \). The bounds are tighter in samples of moderate size with stronger assumptions. All proofs are in Appendix A, supplementary materials. We begin with the following theorem, which applies to all the methods and 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 \( \psi \). Then

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

(2.14)

Now define

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

Then \( \mathbb{E} (\tilde{Z}) = 0 \). Note that \( \Sigma = \text{cov} (Z_i) = \text{cov} (\tilde{Z}) = \text{cov} (X) \). We make the following assumption throughout the remainder of this paper.

**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) \( \Psi \) is compact. (iv) \( f (\psi) \) is bounded on \( \Psi \). (v) \( g_1 (\psi, \mu) \leq 0 \) and \( g_2 (\psi, \mu) = 0 \) for some \( \psi \).

2.3.1. Known \( \Sigma \)

Suppose for the moment that \( \Sigma \) is known. *Section 2.3.2* discusses the case in which \( \Sigma \) is unknown. If \( \Sigma \) is nonsingular, let \( \{\Sigma^{-1/2} (X_i - \mu)\}_j \) denote the \( j \)th component of \( \Sigma^{-1/2} (X_i - \mu) \).

**Method 1**

This method approximates the distribution of \( \tilde{Z} \) by a normal distribution. To do this, make the following assumption.

**Assumption 2.** (i) \( \Sigma \) is nonsingular, and its components are all finite. (ii) There is a constant \( \overline{\mu}_3 < \infty \) such that

\[ \mathbb{E} \left( \left| \Sigma^{-1/2} (X_i - \mu) \right| \right) \leq \overline{\mu}_3 \]

for all \( i = 1, \ldots, n \) and \( j = 1, \ldots, p \).

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

**Theorem 2.2.** Let Assumptions 1(i), (ii), and 2 hold. Then

\[ \mathbb{P} (\tilde{Z} \in \mathcal{S}) - \mathbb{P} (W \in \mathcal{S}) \leq \frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}}. \]

Theorem 2.2 approximates the distribution of \( \tilde{Z} \) by a multivariate normal distribution and uses a multivariate generalization of the Berry-Esséen theorem (Raić 2019) to bound the approximation error. Theorem 2.2 implies that for any \( 0 < \alpha < 1 \),

\[ \mathbb{P} \left[ n (\tilde{X} - \mu)^{1/2} (\tilde{X} - \mu) \leq \kappa_{\alpha}^2 (1 - \alpha) \right] \]

\[ \geq (1 - \alpha) - B (n, p, \overline{\mu}_3), \]

(2.15)

where \( \kappa_{\alpha}^2 (1 - \alpha) \) is the \( (1 - \alpha) \) quantile of the Chi-square distribution with \( p \) degrees of freedom and

\[ B (n, p, \overline{\mu}_3) := \frac{(42p^{1/4} + 16)p^{3/2}\overline{\mu}_3}{n^{1/2}}. \]

The term \( B (n, p, \overline{\mu}_3) \) in (2.15) is asymptotically negligible but can be large in samples of moderate size because it accommodates “worst case” distributions of \( \tilde{X} \) that may be far from normal. If \( n \) is large enough that \( B (n, p, \overline{\mu}_3) < \alpha \), then it follows from (2.15) that

\[ \mathbb{P} \left[ n (\tilde{X} - \mu)^{1/2} (\tilde{X} - \mu) \leq \kappa_{\alpha}^2 (1 - \alpha + B (n, p, \overline{\mu}_3)) \right] \geq 1 - \alpha. \]

(2.16)

It follows from Theorem 2.1 that (2.15) and (2.16) provide lower bounds on the coverage probabilities of confidence intervals for \( f (\psi) \) when \( \mathcal{S} \) is an ellipsoid.

*Table 1* shows numerical values of the bound (1 - \( \alpha \)) - \( B (n, p, \overline{\mu}_3) \) and critical value \( \kappa_{\alpha}^2 (1 - \alpha + B (n, p, \overline{\mu}_3)) \) for different values of \( n \) and \( p \) at \( \alpha = 0.05 \) and \( \overline{\mu}_3 = 2 \). To have a bound close to 1 - \( \alpha \) and a finite critical value, \( n \) must be very large, especially if \( p \) is large. This is because (2.15) accommodates worst case distributions of \( \tilde{X} \). Methods 2 and 3, which are discussed next in this section, provide tighter bounds and smaller critical values when \( n \) is smaller, though Method 1 can provide a smaller critical value when \( n \) is very large and \( p \) is small. However, Method 1 is hard to use in applications even when \( n \) is large if \( \Sigma \) and \( \overline{\mu}_3 \) are unknown, because the resulting bounds depend on population parameters that are difficult to estimate. This problem is discussed in *Section 2.3.2*.

**Method 2**

Method 2 obtains bounds that are much tighter than those of Method 1 when \( n \) is smaller than in *Table 1*, and Method 2 does not require \( \Sigma \) to be invertible. This is accomplished by assuming that the distributions of the components of \( X \) are sub-Gaussian. Specifically, make the following assumption.

**Assumption 3.** Let \( \Upsilon \) denote a nonsingular, nonstochastic \( p \times p \) matrix with finite elements. For any such \( \Upsilon \); all \( i = 1, \ldots, n; \) and all \( j = 1, \ldots, p; \) there are finite constants \( \sigma_j^2 > 0 \) such that the distribution of \( \tilde{Z}_{ij} := [\Upsilon^{-1/2} (X_i - \mu)]_j \) is sub-Gaussian with variance proxy \( \sigma_j^2 \). That is, \( \mathbb{E} [\exp (\lambda \tilde{Z}_{ij})] \leq \exp (\sigma_j^2 \lambda^2 / 2) \) for all \( \lambda \in \mathbb{R}; \) \( i = 1, \ldots, n; \) and \( j = 1, \ldots, p \).

Assumption 3 requires that the distribution of \( \tilde{Z}_{ij} \) be thin-tailed. Sub-Gaussian random variables include Gaussian, Rademacher, and bounded random variables as special cases.
Table 1. Values of the bound and critical value.

| p | 1    | 2    | 3    | 4    | 5    |
|---|------|------|------|------|------|
| n = 10^3 | 0.58 (∞) | 0.00 (∞) | 0.00 (∞) | 0.00 (∞) | 0.00 (∞) |
|   | 0.83 (∞) | 0.58 (∞) | 0.21 (∞) | 0.00 (∞) | 0.00 (∞) |
| 10^6 | 0.91 (6.13) | 0.83 (∞) | 0.72 (∞) | 0.57 (∞) | 0.39 (∞) |
| 10^8 | 0.94 (4.29) | 0.91 (8.73) | 0.88 (∞) | 0.83 (∞) | 0.77 (∞) |
| 10^9 | 0.95 (3.97) | 0.94 (6.53) | 0.93 (9.21) | 0.91 (12.88) | 0.89 (∞) |

Table 2. Different critical values.

| p | κ_{SG,1}^{p}(1−α) | κ_{SG,2}^{p}(1−α) |
|---|-------------------|-------------------|
| 1 | 3.84              | 7.38              |
| 2 | 5.99              | 17.53             |
| 3 | 7.81              | 28.72             |
| 4 | 9.49              | 40.60             |
| 5 | 11.07             | 52.98             |

See, for example, Wainwright (2019). There is a tradeoff between Υ and σ^2. In particular, σ^2 may be larger if Υ is the p × p identity matrix than if Υ = Σ and Σ is nonsingular.

Define σ^2 := max_{1≤j≤p} σ_j^2. The following theorem, combined with Theorem 2.1, provides a lower bound on a confidence interval for f(ψ) when S is an ellipsoid.

**Theorem 2.3.** Let Assumptions 1(i) and 3 hold. Then, for any t > 0,

$$
P \left\{ n(\bar{X} - \mu)' \Upsilon^{-1}(\bar{X} - \mu) > t \right\} \leq 2 \sum_{j=1}^{p} \exp \left( -\frac{t}{2p\sigma_j^2} \right).
$$

In addition, for any 0 < α < 1,

$$
P \left\{ \left. \frac{n}{\sigma^2}(\bar{X} - \mu)' \Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_{SG,1}(1-α) \right\} \geq (1-α),
$$

where κ_{SG,1}(1−α) := 2p · log(2p/α).

**Theorem 2.3** and Method 2 make use of the sub-Gaussianity of $\tilde{Z}_{ij}$, whereas **Theorem 2.2** and Method 1 allow the tails of the distribution of X to be thicker than sub-Gaussian tails. The critical values of Methods 1 and 2 are compared later in this section after the description of Method 3. Estimation of σ^2 is discussed in Section 2.3.2.

**Method 3**

Method 3 makes the stronger assumption that the distribution of X is sub-Gaussian in a vector sense. Specifically, Method 3 makes the following assumption.

**Assumption 4.** Let Υ denote a nonsingular p × p matrix with finite elements. There is a finite constant σ^2 > 0 such that

$$
\mathbb{E} \left[ \exp(\lambda' \bar{Z}(\Upsilon)^{-1/2} \tilde{Z}) \right] \leq \exp(\lambda' \lambda \sigma^2/2)
$$

for all λ ∈ ℝ^p, where $\tilde{Z} = n^{-1/2} \sum_{i=1}^{n} (X_i - \mu)$.

**Assumption 4** is stronger than Assumption 3, because Assumption 4 requires the entire vector X to be sub-Gaussian. If X is multivariate normal and Υ = Σ, Assumption 4 holds with σ^2 = 1. In general, however, it is difficult to find simple conditions under which Assumption 4 is satisfied without assuming that the elements of $\tilde{Z}$ are independent of one another.

An application of Theorem 2.1 of Hsu, Kakade, and Zhang (2012) gives the following theorem which, combined with Theorem 2.2, provides a lower bound on the coverage probability of a confidence interval for f(ψ) when S is an ellipsoid.

**Theorem 2.4.** Let Assumptions 1(i) and 4 hold. Then, for any 0 < α < 1,

$$
P \left\{ \left. \frac{n}{\sigma^2}(\bar{X} - \mu)' \Upsilon^{-1}(\bar{X} - \mu) \leq \kappa_{SG,2}(1-α) \right\} \geq 1-α,
$$

where κ_{SG,2}(1−α) := 2p + 2√p · log(1/α) + 2 log(1/α).

This theorem and Method 3 yield a smaller critical value and confidence set than **Theorem 2.3** and Method 2 do, but they require the stronger Assumption 4. **Table 2** shows the critical values of Methods 2 and 3 and Chi-square critical values with p degrees of freedom. The critical values are all for $\alpha = 0.05$. None of the critical values depends on n. The Chi-square critical value achieves an asymptotic coverage probability of 0.95 and yields the smallest confidence set but does not ensure a finite-sample coverage probability of at least 0.95. The sub-Gaussian critical values and resulting confidence sets are larger but ensure finite-sample coverage probabilities of at least 0.95 under the regularity conditions of **Theorems 2.3** and 2.4. Assumption 3 is easier to satisfy, and its variance proxy is easier to estimate in applications, but the critical value of Method 2 increases more rapidly than the critical value of Method 3 as p gets large.

The critical values of Methods 2 and 3 in **Table 2** can be compared with those of Method 1 in **Table 1**. The critical value of Method 1 converges to the Chi-square critical value as n → ∞. The critical values of Methods 2 and 3 do not depend on n. Therefore, the critical value of Method 1 is smaller than those of Methods 2 and 3 when n is very large and p is small enough. However, the critical value of Method 1 is infinite with moderate values of n, whereas the critical values of Methods 2 and 3 are finite at all values of n.

2.3.2. Unknown Σ and sub-Gaussian Variance Proxy σ^2

In applications, Σ and the sub-Gaussian variance proxy σ^2 are unknown except in special cases. This section explains how to estimate these quantities and discusses the effect of estimation on the bounds presented in Section 2.3.1.

We begin with the bound of **Theorem 2.2** and (2.15). Let $\hat{\Sigma}$ be the following estimator of Σ:

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

Let $\Sigma^{-1}_{jk}$ denote the (j, k) component of Σ^{-1}. For each j, k = 1, . . . , p, let

$$
\Sigma_{jj} := \mathbb{E} \left[ (X_{ij} - \mu_j)^2 \right] \text{ and } 
\Gamma_{jk} := \mathbb{E} \left[ (X_{ij} - \mu_j)(X_{ik} - \mu_k) - \Sigma_{jk} \right].$$

Assumption 5. (i) There is a constant $C_{\Sigma} < \infty$ such that $\left| \Sigma^{-1}_{jk} \right| \leq C_{\Sigma}$ for each $j, k = 1, \ldots, p$. (ii) There is a finite constant $\kappa_0 \geq 1$ such that
\[
\Sigma_{jj} \leq \kappa_0 \quad \text{and} \quad \Gamma_{jk} \leq \kappa_0 \quad \text{(2.17)}
\]
for every $j, k = 1, \ldots, p$. (iii) There is a finite constant $\kappa_1$ such that
\[
\mathbb{E} \left[ (X_{ij} - \mu_j) \right] \leq \kappa_1^t - \frac{r_1}{2} \frac{\Sigma_{jj}}{J} , \quad \text{(2.18a)}
\]
\[
\mathbb{E} \left[ (X_{ij} - \mu_j)(X_{jk} - \mu_k) - \Sigma_{jk} \right] \leq \kappa_1^t - \frac{r_1}{2} \Gamma_{jk} \quad \text{(2.18b)}
\]
for every $r = 2, 3, \ldots$ and $j, k = 1, \ldots, p$.

Assumption 5 is stronger than Assumption 2. In particular, (2.18b) implies that the $(X_{ij} - \mu_j)(X_{jk} - \mu_k)$ is sub-exponential. Therefore, $X_{ij}$ is sub-Gaussian because a random variable is sub-Gaussian if and only if its square is sub-exponential (Vershynin 2018, Lemma 2.7.6). Also, the product of two sub-Gaussian variables is sub-exponential (Vershynin 2018, Lemma 2.7.7). We use Assumption 5(iii) to apply Bernstein’s inequality to the conclusion of Lemma 2.1. The following lemma gives a finite-sample bound on the error of the approximation.

Lemma 2.1. Let Assumptions 1, 2, and 5 hold and
\[
\log(2p) \leq t \leq \kappa^* n . \quad \text{(2.20)}
\]
Then,
\[
\left| \mathbb{P}(S, \tilde{\Sigma}) - \mathbb{P}(W \in S) \right| \leq w_n(t)
\]
with probability at least $1 - 2e^{-t}$.

Theorem 2.5 provides a finite-sample upper bound on the error made by approximating $\mathbb{P} \left[ \tilde{X} \in \mathcal{S} \right]$ by $\mathbb{P}(S, \tilde{\Sigma})$. Combining Theorems 2.1 and 2.5 yields

Theorem 2.6. Let Assumptions 1, 2, 5, and (2.20) hold. Then,
\[
\mathbb{P} \left[ \tilde{J} \leq J \leq f(\psi) \leq J_* \leq \tilde{J} \right] \geq \mathbb{P}(S, \tilde{\Sigma}) - \left\{ \frac{(42p^{1/4} + 16) p^{3/2} \mu_j}{n^{1/2}} + \delta_n^* \right\} . \quad \text{(2.22)}
\]

Theorem 2.6 provides a finite-sample lower bound on $\mathbb{P} \left[ \tilde{J} \leq J \leq f(\psi) \leq J_* \leq \tilde{J} \right]$ that takes account of random sampling error. It is not difficult to choose $\mathcal{S}$ so that the right-hand side of (2.22) is $1 - \alpha$ for any $0 \leq \alpha \leq 1$ if $n$ is large enough and $p$ is small enough to make the term in square brackets on right-hand side of the inequality less than $\alpha$. However, the presence of $\delta_n^*$ greatly decreases the right-hand side of (2.22) relative to what it is when $\Sigma$ is known, thereby increasing the size of the confidence region $\mathcal{S}$ for any given value of $\alpha$. In addition, the right-hand side of (2.22) depends on several population parameters that are difficult to estimate. Therefore, (2.22) is of limited use for applications. The bounds for Methods 2 and 3 with an unknown variance proxy also depend on unknown population parameters, but these are easier to estimate, as is discussed in Section 2.4. The dependence of finite-sample bounds on unknown population parameters is unavoidable except in special cases. For example, if $\gamma = 1$ in Theorem 2.3 and each element of $X - \mu$ is contained in $[-1, 1]$, then $\sigma^2_j = 1$ for all $j$, and the first inequality in Theorem 2.3 becomes the Hoeffding inequality.

We now consider Method 2, which is the sub-Gaussian case of Assumption 3. In some special cases, the sub-Gaussian variance proxies, $\sigma^2_j$ are known. For example, if $Y = I_p$ in Theorem 2.3 and each element of $X - \mu$ is contained in $[-1, 1]$, then $\sigma^2_j = 1$ for all $j$. Here, we derive bounds for the case in which $\sigma^2_j = \max_{1 \leq j \leq p} \sigma^2_j$ is unknown and must be estimated.

Then, $\delta_n^* := \min_{t} \left[ w_n(t) + 2e^{-t} \right]$ subject to (2.20).

Now combine Theorem 2.2 and Lemma 2.1 to obtain the following theorem.

Theorem 2.5. Let Assumptions 1, 2, 5, and (2.20) hold. Then,
\[
\left| \mathbb{P}(\tilde{Z} \in \mathcal{S}) - \mathbb{P}(S, \tilde{\Sigma}) \right| \leq \frac{(42p^{1/4} + 16) p^{3/2} \mu_j}{n^{1/2}} + \delta_n^* .
\]

The following lemma establishes a finite-sample probability bound on the absolute difference between the sample and population variances of $\tilde{X}_{ij}$. Recall that $\tilde{Z}_{ij} = [\gamma^{-1/2} (X_{ij} - \mu)]$ defined in Assumption 3.

Lemma 2.2. Let Assumptions 1(i) and 3 hold. Then, for any $0 < t \leq \left( 16 \sigma_j^2 \right)^2 n$,
\[
\mathbb{P} \left[ |\tilde{\Sigma}_{jj} - \mathbb{E}(\tilde{Z}_{ij}^2) | > 3 \left( \frac{t}{n} \right) + 1 + 2 \left( n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij} \right) \left( \frac{t}{n} \right)^{1/2} \right] \leq e^{-t/(512 \gamma^4)} + 2e^{-t/(2 \gamma^4)} .
\]
The next lemma establishes a link between the population variance of $Z_{ij}$ to the variance proxy $\sigma_j^2$.

**Lemma 2.3.** Let Assumptions 1(i) and 3 hold. Then, the population variance $\mathbb{E}[Z_{ij}^2]$ is a lower bound on the variance proxy $\sigma_j^2$.

Now define

$$\hat{\sigma}^2(t) := \max_{j=1,\ldots,p} \left\{ \tilde{\Sigma}_{jj} + 3 \left( \frac{t}{n} \right) + \left( 1 + 2 \right) \left( \frac{n^{-1} \sum_{i=1}^{n} \tilde{X}_{ij} \right) \left( \frac{t}{n} \right)^{1/2} \right\},$$

which we use to estimate $\sigma^2 = \max_{1 \leq j \leq p} \sigma_j^2$. The following theorem holds for the sub-Gaussian case of Assumption 3 (Method 2).

**Theorem 2.7.** Let Assumptions 1(i), (iii), (iv), (v), and 3 hold. Let the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ be given by

$$n(\bar{X} - \mu)^\top \Sigma^{-1}(\bar{X} - \mu) \leq 2\hat{\sigma}^2(t) \cdot p \cdot \log(2p/\alpha),$$

where $\Sigma$ is nonrandom and $\hat{\sigma}^2(t)$ is given in (2.23). Then, for any $0 < t \leq (16\alpha^2)^2 n$,

$$P \left[ \tilde{J}_{-} < \bar{X} \leq \tilde{J}_{+} \right] \geq 1 - \alpha - \frac{p}{\sqrt{n}} \left( e^{-t/(512\sigma^2 t)} + 2e^{-t/(2\sigma^2 t)} \right).$$

A smaller choice of $t$ in (2.23) makes the confidence set in (2.24) tighter but results in a lower probability bound (2.25). The right-hand side of (2.25) can be close to zero and depends on the unknown parameters $\sigma_j^2$. Therefore, the bound (2.25), like the bound (2.22), is of limited use for applications. As noted in the discussion of (2.22), dependence of finite-sample bounds on unknown population quantities is unavoidable except in special cases. Section 2.4 describes a practical approach to dealing with this problem that can be used in applications.

A result similar to Theorem 2.7 can be obtained for Method 3. We do not undertake this here, however, because the resulting bound analogous to (2.25) is loose, depends on unknown population parameters, and is of limited use for applications. Instead, in Section 2.4, we describe a way of dealing with unknown population parameters in Methods 2 and 3 that can be used in applications.

### 2.4. Dealing with Unknown Population Parameters in Methods 2 and 3 in Applications

Except in special cases, it is not possible to obtain finite-sample inequalities for Methods 2 and 3 that do not depend on unknown population parameters. It is possible, however, to estimate lower bounds on these parameters consistently. It follows from Lemma 2.3 that $\tilde{\Sigma}$ and $\max_{1 \leq j \leq p} \tilde{\Sigma}_{jj}$, respectively, are consistent estimates of lower bounds on the variance proxies $\sigma_j^2$ and $\sigma^2$ of Method 2. The differences between the lower bounds and the variance proxies are often small. Arguments like those used to prove Lemma 2.3 show that the largest eigenvalue of the covariance matrix of $Z_{ij}$ ($j = 1, \ldots, p$) is a lower bound on the variance proxy of Method 3. This can be estimated consistently by the largest eigenvalue of the sample covariance matrix.

Standard methods can be used to obtain asymptotic confidence intervals for the variance proxies of Method 2. These methods do not provide information about differences between true and nominal coverage probabilities in finite samples, but they provide practical indications of the magnitudes of the Method 2 bounds that can be implemented in applications.

Obtaining a useful asymptotic confidence interval for the largest eigenvalue of the covariance matrix of $Z_{ij}$ is difficult. A wide interval whose true coverage probability is likely to be much greater than the nominal probability can be obtained from the Frobenius norm of the difference between the estimated and true covariance matrices.

Table 4 in Section 5 compares values of $\tilde{J}_{+}$ obtained using estimates and population values of the variance proxies of Methods 2 and 3.

### 3. Computational Algorithms

Recall that our general framework is to obtain the bound $\min_{\psi \in \Psi} f(\psi), \max_{\psi \in \Psi} f(\psi)$ subject to $g_1(\psi, m) \leq 0, g_2(\psi, m) = 0, \psi \in \Psi, m \in \mathcal{S}$. In many examples we consider, $f(\psi)$ is linear in $\psi$. For example, $\psi$ is the vector of all the parameters in an econometric model and $f(\psi)$ is just one element of $\psi$ or a linear combination of elements of $\psi$.

The restrictions $g_1(\psi, \mu) \leq 0$ include shape restrictions among the elements of $\psi$. Equality restrictions are imposed via $g_2(\psi, \mu) = 0$. The easiest case is that $g_j(\psi, \mu)$ is linear in $\psi$, $\mu$ for each $j = 1, 2$. In some of examples we consider, $g_j(\psi, \mu)$ is linear in $\psi$, fixing $\mu$ fixed, and linear in $\mu$, keeping $\psi$ fixed, but not linear in $\psi$ and $\mu$ 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 $\psi$ and one of elements of $\mu$. In practice, $\Psi$ can always be chosen large enough that the constraint $\psi \in \Psi$ is not binding additionally and can be ignored. For example, suppose that $\psi$ is a probability and the constraints in $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ impose a restriction on $\psi$ such as $[a, b]$ for $0 \leq a < b \leq 1$. Then, it is not necessary to impose $\psi \in \Psi = [0, 1]$ additionally.

Recall that the leading cases of $\mathcal{S}$ include an ellipsoid and a box. For brevity, we focus on the scenario that normal distributions are used in obtaining $\mathcal{S}$. When $\mathcal{S}$ is a box, the critical value $\kappa_{\chi^2}(1 - \alpha)$ can be easily simulated from the $N(0, \tilde{\Sigma})$ and the restriction $n^{1/2}(X - \mu) \in \mathcal{S}$ can be written as linear constraints. When $\mathcal{S}$ is an ellipsoid, the critical value $\kappa_{\chi^2}(1 - \alpha)$ can be obtained from the $\chi^2(d_{\mu})$ distribution, where $d_{\mu}$ is the dimension of $\mu$. Then, the restriction $n^{1/2}(\bar{X} - \mu) \in \mathcal{S}$ can be written as

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

This is a convex quadratic constraint in $\mu$.

When some of the constraints $g_1(\psi, \mu) \leq 0$ and $g_2(\psi, \mu) = 0$ are bilinear, the resulting feasible region may not be convex. To deal with the bilinear constraints, we solve optimization problems using mixed integer programming (MIP) with Gurobi in R. By virtue of the developments in MIP solvers.
and fast computing environments, MIP has become increasingly used in recent applications. For example, Bertsimas, King, and Mazumder (2016) adopted an MIO approach for obtaining \( \ell_0 \)-constrained estimators in high-dimensional regression models and Reguant (2016) used mixed integer linear programming for computing counterfactual outcomes in game theoretical models.

4. An Empirical Application

Angrist and Evans (1998) use data from the 1980 and 1990 U.S. census to estimate a model of the relation between the number of weeks per year a woman works and the number of children she has. A simplified but nonparametric version of their model is

\[
Y = \phi(D) + U; \quad E[U|Z] = 0, \tag{4.1}
\]

where \( Y \) is the number of weeks a woman works in a year; \( \phi \) is an unknown function; and \( D = 0, 1, 2 \) according to whether a woman has 2, 3, or 4 or more children. \( D \) is endogenous. \( Z \) is a binary instrument for \( D \) equal to 1 if the first two children are of the same sex and 0 otherwise. We obtain bounds on the number of children she has increases from two to three. We use data consisting of 394,840 observations \((Y, D, Z)\) from the 1980 U.S. census (Ruggles et al. 2021).

We assume that \( \phi \) is monotone nonincreasing and focus on the parameter \( f(\psi) = [\phi(0) - \phi(1), (\phi(0), \phi(1), \phi(2))]' \). The population mean vector is \( \mu = [\text{vec}(\Pi), \text{vec}(\nu)] \), where

\[
\Pi = \begin{pmatrix} P_{DZ}(0,0) & P_{DZ}(1,0) & P_{DZ}(2,0) \\ P_{DZ}(0,1) & P_{DZ}(1,1) & P_{DZ}(2,1) \end{pmatrix},
\]

\[
\nu = \begin{pmatrix} E[Y|Z = 0] \\ E[Y|Z = 1] \end{pmatrix},
\]

\[p_{DZ}(d,z) := \Pr(D = d, Z = z), \text{ and } I(\cdot) \text{ is the indicator function. The dependent variable } Y \text{ is contained in the interval [0,52]. In the analysis described below, we divide } Y \text{ by } 52, \text{ so it is contained in the interval [0,1], and denote the empirical analog of } \mu \text{ by } \bar{X}. \text{ The inequality constraints are}
\]

\[
g_1(\psi, \mu) = \begin{pmatrix} -1 & 1 & 0 \\ 0 & -1 & 1 \end{pmatrix} \begin{pmatrix} \phi(0) \\ \phi(1) \\ \phi(2) \end{pmatrix} \leq 0. \tag{4.2}
\]

The equality constraints are

\[
g_2(\psi, \mu) = \sum_{d=0}^{24} \sum_{z=0}^{52} p_{DZ}(d,z) \psi - \mu = 0. \tag{4.3}
\]

The instrumental variable constraints in (4.3) are bilinear in the sense that \( \Pi \psi \) contains 6 bilinear terms.

We estimate the following 8 identification or 95% confidence intervals for \( f(\psi) \).

1. (**Sample**) The identification interval given by (2.1)–(2.2) under the assumption that the sample analogs of the components of \( \mu \) equal the population values. The resulting estimated identification interval is a consistent point estimate of the population identification interval but does not take account of random sampling errors in the estimate of \( \mu \).

2. (**Box**) The 95% confidence interval given by (2.3)–(2.4) with \( S \) a box, the \( D_j \)'s equal to the diagonal elements of the sample covariance matrix \( \hat{\Sigma} \), and \( k_p(1 - \alpha) = 2.66 \), which is obtained via the method given in Section 3. This estimate treats \( \hat{\Sigma} \) as if it were the population covariance matrix.

3. (**Chi-Square**) The 95% confidence interval given by (2.3)–(2.4) with \( S \) an ellipsoid, \( \Psi = \hat{\Sigma} \), and \( k_p(1 - \alpha) = 14.08 \), which is the 0.95 quantile of the Chi-square distribution with 7 degrees of freedom (the number of distinct components of \( \mu \) taking account of the constraint that \( \sum_{d=0}^{24} \sum_{z=0}^{52} p_{DZ}(d,z) = 1 \)). This estimate, like the previous one, treats \( \hat{\Sigma} \) as if it were the population covariance matrix.

4. (**Method 2 (\( \Psi = \hat{\Sigma}, \sigma^2 = 1 \))**) The 95% confidence interval obtained from Method 2 with \( S \) an ellipsoid, \( \Psi = \hat{\Sigma} \), \( \sigma^2 = 1 \), and \( k_{SG_1}(0.95) = 78.89 \). This estimate also treats \( \Psi = \hat{\Sigma} \) as if it were nonrandom. The choices of \( \Psi \) and \( \sigma^2 \) are motivated by consistency of \( \hat{\Sigma} \) for \( \Sigma \).

5. (**Method 2 (\( \Psi = I_7, \sigma^2 = 1 \))**) The 95% confidence interval obtained from Method 2 with \( S \) an ellipsoid, \( \Psi = I_7 \), \( \sigma^2 = 1 \), and \( k_{SG_1}(0.95) = 78.89 \). This estimate does not use \( \hat{\Sigma} \). This choices of \( \Psi \) and \( \sigma^2 \) ensure finite sample validity because each element of \( \Psi \) is contained in the interval \([-1,1]\).

6. (**Method 2 (\( \Psi = I_7, \sigma^2 = \hat{\sigma}^2 (\log n) \))**) The 95% confidence interval obtained from Method 2 with \( S \) an ellipsoid, \( \Psi = I_7 \), \( \sigma^2 = \hat{\sigma}^2 (\log n) \). This estimate does not use \( \hat{\Sigma} \). This choices of \( \Psi \) and \( \sigma^2 \) are motivated by consistency of \( \hat{\Sigma} \) for \( \Sigma \).

7. (**Method 3**) The 95% confidence interval obtained from Method 3 with \( S \) an ellipsoid, \( \Psi = \hat{\Sigma} \), \( \sigma^2 = 1 \), and \( k_{SG_2}(0.95) = 22.15 \). This estimate treats \( \Psi = \hat{\Sigma} \) as if it were not random. The choices of \( \Psi \) and \( \sigma^2 \) are motivated by consistency of \( \hat{\Sigma} \) for \( \Sigma \).

8. (**Minsker**) The 95% confidence interval based on Minsker (2015). Specifically, it is obtained by replacing \( n^{1/2}(\bar{X} - \mu) \in S \) with \( (\hat{\mu}_x - \mu) \) \( \leq \hat{r}_{n,x} \), where \( \hat{\mu}_x \) is a median-of-means estimator of \( \mu \) based on coordinate-wise medians and \( \hat{r}_{n,x} \) is the critical value that ensures finite sample coverage. See Appendix C, supplementary materials for details.

We do not consider a confidence interval based on (2.16) because \( B(n, p, \bar{\pi}_3) \approx 2.49 \bar{\pi}_3 \) (here, \( n = 394,840 \) and \( p = 7 \)) and as a result, it is extremely unlikely that \( B(n, p, \bar{\pi}_3) < 0.05 \), although \( \bar{\pi}_3 \) is unknown. Computation was carried out on a MacBookPro laptop with an Apple M1 chip and 16 GB of memory.

The results, including computing times, are shown in Table 3. The intervals are in weeks, not weeks divided by 52. The estimates obtained by methods 1–8 above are labeled in typewriter font: Sample; Box; Chi-Square; three types of Method 2; Method 3; and Minsker; respectively. As expected, the Sample interval is narrower than the Box and Chi-Square intervals, and the Chi-Square interval is narrower than the Box interval. The Method 2 \( (\Psi = \hat{\Sigma}, \sigma^2 = 1) \) and Method 3 intervals are wider than the Chi-Square interval. The Method 2 \( (\Psi = \hat{\Sigma}, \sigma^2 = 1) \) interval is wider than the Method 3 interval, which is based on a stronger...
assumption about the distribution of the observed random variables. The foregoing methods are motivated by consistency of the estimates of the population parameters they depend on. They do not take account of random sampling error in the estimates. The Method 2 (\(Y = \Psi_2, \sigma^2 = 1\)) and Minsker intervals ensure finite-sample coverage probabilities of 0.95 but are wider than the other intervals. The Method 2 (\(Y = \Psi_2, \sigma^2 = \hat{\sigma}^2(\log n)\)) interval provides a tighter upper bound of 26.73 by estimating the variance proxy instead of relying on the bounded support of \(X - \mu\). Only the Sample estimate yields an informative lower bound on \(\phi(0) - \phi(1)\). Excluding the Sample bounds, which do not take account of random sampling error in the estimate of \(\mu\), the upper bounds indicate that an increase in the number of children a woman has from 2 to 3 reduces her annual employment at most by approximately 11–37 weeks, depending on the estimation method. All of the computing times are small.

5. Monte Carlo Experiments

This section presents the results of Monte Carlo experiments that investigate the widths and coverage probabilities of nominal 95% confidence intervals for \(f(\psi)\) that are obtained by using the methods described in Section 2. For reasons explained in Section 5.1, we concentrate on the upper confidence limit \(\hat{\psi}_+\) and do not investigate \(\hat{\psi}_-\). The experiments are designed to mimic the empirical application of Section 4.

5.1. Design

Data were generated by simulation from model (4.1) as follows. Using the notation of Section 4, set

\[
\psi = [\phi(0), \phi(1), \phi(2)] = \frac{1}{52}(23, 18, 16) \quad \text{and} \quad \Pi = \begin{pmatrix} 0.31 & 0.13 & 0.05 \\ 0.29 & 0.16 & 0.06 \end{pmatrix}.
\]

The values of \(\psi\) are similar to those used in Freyberger and Horowitz (2015), who considered larger support of \(D\), and the values of \(\Pi\) are the same as the sample probabilities of \(\Pi\) in the empirical example of Section 4. Simulated larger values of \((D, Z)\) were drawn from \(\Pi\). The outcome variable \(Y\) was generated from

\[
Y = \mathbb{E}[\phi(D)|Z] + \frac{V}{52}; \quad V \sim N(0, 1).
\]

The parameter of interest is \(f(\psi) = 52[\phi(0) - \phi(1)] = 5\), which is not point-identified. We impose that each element of \((\psi, \mu)\) is nonnegative and impose the monotonicity constraint (4.2). The population value of \([\hat{\psi}_-, \hat{\psi}_+]\) is \([0.381, 5.438]\). Thus, the data are not very informative about \(\hat{\psi}_-\). Consequently, we focus on \(\hat{\psi}_+\), the estimate of \(\hat{\psi}_+\), in the experiments and do not investigate \(\hat{\psi}_-\).

We carried out experiments with \(S\) an ellipsoid and \(S\) a box. We report the averages and empirical coverage probabilities of \(\hat{\psi}_+\) obtained by using eight methods described in Section 4. The nominal coverage probability was 95% and there were 500 Monte Carlo replications per experiment. In experiments in which \(S\) is a box, \(\kappa_{\hat{\psi}}(1 - \alpha)\) was computed using \(10^6\) random draws from \(N(0, \hat{\Sigma})\).

5.2. Results

Figure 1 shows only 100 (out of 500) realized values for each different estimated bound by sample size \(n \in \{10^4, 10^5, 10^6\}\). Each symbol corresponds to one Monte Carlo realization, and the population value \(\hat{\psi}_+\) is shown as a horizontal line. In the figure, both \(x\) and \(y\) axes are shown in the log scale with base 10. Table 4 summarizes the results of the Monte Carlo experiments. The quantities in parentheses in the table are the values of \(\hat{\psi}_+\) for the Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) and Method 3 methods obtained with the population values of the sub-Gaussian variance proxies, not estimates. In the examples treated in these experiments, inference based on the true variance proxies and on consistent estimates of the proxies would be almost identical.

The Sample bounds converge to the true upper bound as \(n\) increases; however, they are not suitable for inference since sampling errors are ignored. All bounds get smaller as \(n\) gets large, but the Chi-Square, Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)), and Method 3 methods provide much tighter bounds than the other methods. The values of \(\hat{\psi}_+\) for the Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) and Method 3 methods obtained with estimated and true values of the sub-Gaussian variance proxies are nearly equal. In every case except the Sample bounds, the estimated bounds are larger than \(\hat{\psi}_+\), resulting in the 100% empirical coverage probabilities. Overall, the simulation results are consistent with those of the empirical application in Section 4.

Table 4 shows that the bounds on \(\hat{\psi}_+\) vary greatly, depending on the method used and assumptions made about population parameters. In an application, a researcher might compute the

| Table 3. Angrist and Evans (1998) revisited. |
| --- |
| \(S\) | Bounds | Computing Time (sec) |
| Sample | [1.14, 6.36] | 0.08 |
| Box | [0.17, 0.06] | 11 |
| Chi-Square | [0.10, 0.85] | 12 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) | [0.17, 0.19] | 15 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) | [0.36, 0.39] | 5 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = \hat{\sigma}^2(\log n)\)) | [0.26, 0.73] | 12 |
| Method 3 | [0.12, 0.01] | 13 |
| Minsker | [0.37, 0.12] | 6 |

| Table 4. Results of Monte Carlo experiments. |
| --- |
| \(n = 10^4\) | \(10^5\) | \(10^6\) |
| Averages of \(\hat{\psi}_+\) |
| Sample | 5.60 | 5.43 | 5.45 |
| Box | 35.61 | 21.12 | 10.70 |
| Chi-Square | 17.92 | 6.84 | 5.84 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) | 36.88 | 10.51 | 6.44 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) | 36.84 | 10.51 | 6.44 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = 1\)) | 49.75 | 38.73 | 32.37 |
| Method 2 (\(Y = \hat{\Sigma}, \sigma^2 = \hat{\sigma}^2(\log n)\)) | 41.78 | 36.09 | 19.07 |
| Method 3 | 25.98 | 7.30 | 5.94 |
| Method 3 with true \(\Sigma\) | (25.43) | | |
| Minsker | 52.62 | 39.47 | 35.21 |
| Proportions that \(\hat{\psi}_+ \geq \hat{\psi}_+ = 5.438\) |
| Sample | 0.49 | 0.48 | 0.51 |
| All other methods | 1.00 | 1.00 | 1.00 |
bounds using all the methods and assumptions that are relevant to the application. The researcher can then decide which bounds to use or report, depending on his/her beliefs about the accuracy of the asymptotic approximations s/he is making and how much risk of inaccuracy s/he is willing to accept.

To check sensitivity to sub-Gaussian assumptions, we carried out additional Monte Carlo experiments by replacing $V$ in (5.1) with $V \sim (\chi^2(5) - 5)/\sqrt{10}$. In other words, we use a standardized Chi-square random variable, which is not sub-Gaussian but sub-exponential. The results of the additional Monte Carlo experiments are similar to those reported here with the standard normal $V$. See Appendix E, supplementary materials for details.

6. Conclusions

This article has described a method for carrying out inference on partially identified parameters that are solutions to a class of optimization problems. The parameters arise, for example, in applications in which grouped data are used for estimation of a model’s structural parameters. Inference consists of obtaining confidence intervals for the partially identified parameters. The article has presented three methods for obtaining finite-sample lower bounds on coverage probabilities of the confidence intervals. The methods correspond to three sets of assumptions of increasing strength about the underlying random variable. With the moderate sample sizes found in most economics applications, the bounds become tighter as the assumptions strengthen. The article has also described a computational algorithm for implementing the methods. The results of Monte Carlo experiments and an empirical example illustrate the methods’ usefulness. The article has focused on the case in which a vector of first moments is the only unknown population parameter in the population version of the optimization problem. It might be useful to extend our formulation to the case in which the population optimization problem contains other population parameters. It is an open question how to obtain finite-sample lower bounds on coverage probabilities in such a case.

Supplementary Materials

The supplementary materials consist of five appendices. Appendix A presents the proofs of theorems. Appendix B provides additional technical information about our methods. Appendix C describes Minsker’s method. Appendix D provides an additional empirical example, and Appendix E provides additional Monte Carlo results.

Acknowledgments

An earlier version of this article was presented at the conference “Incomplete Models,” which took place at Northwestern University in November 2018. We thank the conference participants, the associate editor, and two anonymous referees for helpful comments. Part of Joel L. Horowitz’s research was carried out while he was a visitor to the Department of Economics at University College London.

Funding

This research was supported in part by the European Research Council (grant no. ERC-2014-CoG-646917-ROMIA) and the UK Economic and Social Research Council (grant no. ES/P008909/1 to the CeMMAP).

References

Angrist, J. D., and Evans, W. N. (1998), “Children and their Parents’ Labor Supply: Evidence from Exogenous Variation in Family Size,” American Economic Review, 88, 450–477. [935,936]
