Yale University

EliScholar – A Digital Platform for Scholarly Publishing at Yale

Cowles Foundation Discussion Papers | Cowles Foundation

11-1-2018

Sensitivity Analysis using Approximate Moment Condition Models

Timothy B. Armstrong

Michal Kolesár

Follow this and additional works at: https://elischolar.library.yale.edu/cowles-discussion-paper-series

 Part of the Economics Commons

Recommended Citation

Armstrong, Timothy B. and Kolesár, Michal, "Sensitivity Analysis using Approximate Moment Condition Models" (2018). *Cowles Foundation Discussion Papers*. 106.
https://elischolar.library.yale.edu/cowles-discussion-paper-series/106

This Discussion Paper is brought to you for free and open access by the Cowles Foundation at EliScholar – A Digital Platform for Scholarly Publishing at Yale. It has been accepted for inclusion in Cowles Foundation Discussion Papers by an authorized administrator of EliScholar – A Digital Platform for Scholarly Publishing at Yale. For more information, please contact elischolar@yale.edu.

SENSITIVITY ANALYSIS USING APPROXIMATE MOMENT CONDITION MODELS

By

Timothy B. Armstrong and Michal Kolesár

November 2018

Revised February 2019

COWLES FOUNDATION DISCUSSION PAPER NO. 2158R



COWLES FOUNDATION FOR RESEARCH IN ECONOMICS
YALE UNIVERSITY
Box 208281
New Haven, Connecticut 06520-8281

http://cowles.yale.edu/

Sensitivity Analysis using Approximate Moment Condition Models*

Timothy B. Armstrong† Yale University

Michal Kolesár‡ Princeton University

February 15, 2019

Abstract

We consider inference in models defined by approximate moment conditions. We show that near-optimal confidence intervals (CIs) can be formed by taking a generalized method of moments (GMM) estimator, and adding and subtracting the standard error times a critical value that takes into account the potential bias from misspecification of the moment conditions. In order to optimize performance under potential misspecification, the weighting matrix for this GMM estimator takes into account this potential bias, and therefore differs from the one that is optimal under correct specification. To formally show the near-optimality of these CIs, we develop asymptotic efficiency bounds for inference in the locally misspecified GMM setting. These bounds may be of independent interest, due to their implications for the possibility of using moment selection procedures when conducting inference in moment condition models. We apply our methods in an empirical application to automobile demand, and show that adjusting the weighting matrix can shrink the CIs by a factor of 3 or more.

*We thank Isaiah Andrews, Gary Chamberlain, Tim Christensen, Mikkel Plagborg-Møller, Jesse Shapiro, Martin Weidner and participants at several conferences and seminars for helpful comments and suggestions, and Soonwoo Kwon for research assistance. All remaining errors are our own. The research of the first author was supported by National Science Foundation Grant SES-1628939. The research of the second author was supported by National Science Foundation Grant SES-1628878.

†email: timothy.armstrong@yale.edu

‡email: mkolesar@princeton.edu

1 Introduction

Economic models are typically viewed as approximations of reality. However, conventional approaches to estimation and inference assume that a model holds exactly. In this paper, we weaken this assumption, and consider inference in a class of models characterized by moment conditions which are only required to hold in an approximate sense. The failure of the moment conditions to hold exactly may come from failure of exclusion restrictions (e.g. through omitted variable bias or because instruments enter the structural equation directly in an IV model), functional form misspecification, or other sources such as measurement error, or data contamination.

We assume that we have a model characterized by a set of population moment conditions $g(\theta)$. In the generalized method of moments (GMM) framework, for instance, $g(\theta) = E[g(w_i, \theta)]$, which can be estimated by the sample analog $\frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$, based on the sample $\{w_i\}_{i=1}^{n}$. When evaluated at the true parameter value θ_0, the population moment condition lies in a known set specified by the researcher,

$$g(\theta_0) = c/\sqrt{n}, \qquad c \in \mathcal{C}.$$

The set $\mathcal{C}$ formalizes the way in which the moment conditions may fail, and it can then be varied as a form of sensitivity analysis, with $\mathcal{C} = \{0\}$ reducing to the correctly specified case.

We focus on local misspecification: the scaling of the set by the square root of the sample size n implies that the specification error and sampling error are of the same order of magnitude. This allows us to keep the analysis tractable, yielding a simple method for inference on a structural parameter of interest, $h(\theta_0)$, rather than a pseudo-true parameter. The tractability of the local misspecification framework has made it a popular tool for sensitivity analysis in applied work, especially following the recent influential paper by Andrews et al. (2017).[1] As with any asymptotic device, our modeling of misspecification as local should not be taken to mean that we literally believe that the model would be closer to correct if we had more data. Rather, its usefulness should be judged by whether it yields accurate approximations to the finite-sample behavior of estimators and confidence intervals, which in our case requires that the set $\mathcal{C}/\sqrt{n}$ be small relative to sampling uncertainty, given the sample size at hand.

We propose a simple method for constructing asymptotically valid confidence intervals (CIs) under this setup: one takes a standard estimator, such as the GMM estimator, and adds and subtracts its standard error times a critical value that takes into account the potential

[1] For recent empirical examples using local sensitivity analysis, see Gayle and Shephard (2019), or Duflo et al. (2018).

asymptotic bias of the estimator, in addition to its variance. A key insight of this paper is that because the CIs must be widened to take into account the potential bias, the optimal weighting matrix for the correctly specified case (the inverse of the variance matrix of the moments) is generally no longer optimal under local misspecification. Rather, the optimal weighting matrix takes into account potential misspecification in the moments in addition to the variance of their estimates: it places less weight on moments that are allowed to be further from zero according the researcher's specification of the set $\mathcal{C}$. We also show that an analogous result holds for other performance criteria, such as estimation under the mean-squared error: the optimal weighting matrix again trades off the potential misspecification of the moments against their precision, although the optimal tradeoff is different.

To illustrate the practical importance of this result, we apply our methods to form misspecification-robust CIs in an empirical model of automobile demand based on Berry et al. (1995). We consider sets $\mathcal{C}$ motivated by the forms of local misspecification considered in Andrews et al. (2017), who calculate the asymptotic bias of the usual GMM estimator in this model. We find that adjusting the weighting matrix to account for potential misspecification substantially reduces the potential bias of the estimator and, as a result, leads to large efficiency improvements of the optimal CI relative to a CI based on the GMM estimator that is optimal under correct specification: it shrinks the CI by up to a factor of 3 or more in our main specifications. As a result, we obtain informative CIs in this model even under moderate amounts of misspecification.

When the set $\mathcal{C}$ is convex, the misspecification-optimal weighting and the critical value are easy to compute. In general, they can be computed by solving a convex optimization problem, which may simplify further in particular cases, yielding closed-form expressions. We show that when the set $\mathcal{C}$ is characterized by ℓ_p constraints, this leads to weightings that are analogous to penalized regression estimators, such as ridge or LASSO regression. By exploiting this analogy, we develop a simple algorithm for computing the optimal weighting under ℓ_1 and ℓ_∞ constraints that is similar to the LASSO/LAR algorithm (Efron et al., 2004; Rosset and Zhu, 2007); under ℓ_2 constraints, the optimal weighting admits a closed form.[2] To avoid having to reoptimize the objective function with respect to the new weighting matrix, one can also form the CIs by adding and subtracting this critical value from a one-step estimator (see Newey and McFadden, 1994, Section 3.4) based on any initial estimate that is $\sqrt{n}$-consistent under correct specification. This approach is particularly attractive when performing sensitivity analysis: starting with an initial GMM estimate that assumes $\mathcal{C} = \{0\}$, one can relax the moment conditions to form larger sets $\mathcal{C}$ and compute the corresponding

[2] An R package implementing our CIs under ℓ_p constraints is available at `https://github.com/kolesarm/GMMSensitivity`.

CIs. This allows one to easily assess how severely misspecified a given model has to be before a result of interest breaks down.

We show that the CIs we propose are near-optimal when the set $\mathcal{C}$ is convex and centrosymmetric ($c \in \mathcal{C}$ implies $-c \in \mathcal{C}$). To this end, we argue that the relevant "limiting experiment" for the locally misspecified GMM model is isomorphic to an approximately linear model of Sacks and Ylvisaker (1978), which falls under a general framework studied by, among others, Donoho (1994), Cai and Low (2004) and Armstrong and Kolesár (2018). We derive asymptotic efficiency bounds for CIs in the locally misspecified GMM model that formally translate bounds from the approximately linear limiting experiment to the locally misspecified GMM setting. In particular, these bounds imply that our CIs are highly efficient relative to CIs that optimize their performance at a particular value of θ_0 and $c = 0$ subject to maintaining coverage over the whole parameter space for θ and $\mathcal{C}$.

These efficiency bounds have two important implications. First, they address an important potential criticism of our CIs: the estimator used to construct the CI as well as the CI width reflect the a priori worst possible misspecification in $\mathcal{C}$ through the optimal weighting matrix and the critical value. For example, when $\mathcal{C} = \{c : \|c\| \leq M\}$ for some norm $\|\cdot\|$, the width of the CI depends on M, so that the CI will be wide even if it turns out that $\|c\|$ is in fact much smaller than M. To address this problem, one may attempt to form a CI that implicitly or explicitly estimates M, by, for example, using a statistic in a specification test such as the J statistic. One then uses the estimate to adjust the width of the resulting CI, "letting the data speak" about the amount of misspecification, rather than depending on the researcher's a priori bound M. Unfortunately, our efficiency bounds show that such a goal cannot be achieved: any CI that substantially improves upon the width of our CI when $\|c\|$ is small must *necessarily* undercover for some other $c \in \mathcal{C}$. Rather than using the data to estimate M, we therefore instead recommend reporting the results for a range of M as a form of sensitivity analysis. We illustrate this approach in our empirical application in Section 6.

Second, similar to these implications about the impossibility of using the data to estimate the *magnitude* of misspecification, our results also imply that one cannot use the data to decide *which* moments are misspecified when forming CIs. As an example, consider the case where the researcher has a set of moments that are known to be correct, along with an additional set of moments which may be misspecified. We can put this in our framework using the set $\mathcal{C} = \{0\} \times \{\tilde{c} : \|\tilde{c}\| \leq M\}$, where M is some conservative a priori bound for the misspecified moments, which may be taken to be infinite. When $M = \infty$, our CI reduces to the usual CI based on the GMM estimator that uses the correctly specified moments only. When M is smaller, our CI uses the misspecified moments and takes into account the worst-

possible misspecification by widening the CI. The weight on the misspecified moments and the width of the CI depend on $\tilde{c}$ only through the a priori bound M. One may attempt to improve upon this by using a first-stage test or estimate of $\tilde{c}$ to choose the weights. As shown by Leeb and Pötscher (2006), $\tilde{c}$ cannot be consistently estimated in this setting, and any such procedure must adjust the resulting CI for the uncertainty in the estimate if coverage is to be maintained. Nevertheless, several papers have proposed adjustments along these lines and have shown formally that the resulting CI has correct coverage, focusing on the case where $M = \infty$ (Andrews and Guggenberger, 2009; DiTraglia, 2016; McCloskey, 2017). Our results show that such CIs cannot substantially improve upon a CI that always assumes the worst possible misspecification, even when it turns out that $c = 0$. In particular, our results imply that when $M = \infty$, the usual one- and two-sided 95% CIs based on only the correctly specified moments are 100% and 84.99% efficient, respectively, uniformly over θ_0 and c, which is the same efficiency as that of the usual CI under correct specification. More generally, the scope for improvement from such procedures is severely limited whenever $\mathcal{C}$ is convex and centrosymmetric. This contrasts sharply with point estimation, for which significant improvements in the mean squared error are possible when $\|\tilde{c}\|$ is small (Liao, 2013; Cheng and Liao, 2015; DiTraglia, 2016).

Our paper is related to several strands of literature. Our efficiency results are related to those in Chamberlain (1987) for point estimation in the correctly specified setting (see also Hansen, 1985) and, more broadly, semiparametric efficiency theory in correctly specified settings (see, e.g., Chapter 25 in van der Vaart, 1998). As we discuss in Section 3.3, some of our efficiency results are novel even in the correctly specified case, and may be of independent interest. Kitamura et al. (2013) consider efficiency of point estimators satisfying certain regularity conditions when the misspecification is bounded by the Hellinger distance. As we discuss in more detail in Section 4.3, our results imply that under this form of misspecification, the optimal weighting matrix remains the same as under correct specification; both the usual GMM estimator and the estimator proposed by Kitamura et al. (2013) can thus be used to form near-optimal CIs, and both estimators have the same local asymptotic minimax properties.

Local misspecification has been used in a number of papers, which include, among others, Newey (1985), Berkowitz et al. (2012), Conley et al. (2012), Guggenberger (2012), Kitamura et al. (2013) and Bugni and Ura (2018). Andrews et al. (2017) consider this setting and note that asymptotic bias of a regular estimator can be calculated using influence function weights, which they call the sensitivity, and show how such calculations can be used for sensitivity analysis in applications (see also extensions of these ideas in Andrews et al. 2018 and Mukhin 2018). Our results imply that, if one is interested in inference, conclusions of such

sensitivity analysis may be substantially sharpened by using the misspecification-optimal weighting matrix, or, equivalently, the misspecification-optimal sensitivity. In independent work, Bonhomme and Weidner (2018) consider inference and optimal estimation under local misspecification defined relative to a reference model within a larger class of models.

The use of local neighborhoods to model misspecification has antecedents in the literature on robust statistics (see Huber and Ronchetti, 2009, and references therein). More broadly, our paper relates to the general literature on sensitivity analysis and misspecification, including, among many others, Leamer (1983), Altonji et al. (2005), Hahn and Hausman (2005), Small (2007), Nevo and Rosen (2010) and Chen et al. (2011). Finally, we note that the settings considered here, as well as most of the papers cited above, differ from the approach of redefining the parameter of interest so that a conventional estimator is asymptotically unbiased, as with the best linear predictor interpretation of the least squares estimator.

The rest of this paper is organized as follows. Section 2 presents our misspecification robust CIs and gives step-by-step instructions for computing them. Section 3 presents efficiency bounds for CIs in locally misspecified models; it can be skipped by readers interested only in implementing the methods. Section 4 discusses solutions for particular choices of the set $\mathcal{C}$. Section 5 discusses applications to particular moment condition models. Section 6 presents an empirical application. Additional results and proofs are collected in appendices and an online supplement.

2 Misspecification-robust CIs

We have a model that maps a vector of parameters $\theta \in \Theta \subseteq \mathbb{R}^{d_\theta}$ to a d_g-dimensional population moment condition $g(\theta)$ that restricts the distribution of the observed data $\{w_i\}_{i=1}^n$. We allow the moment condition model to be locally misspecified, so that at the true value θ_0, the population moment condition is not necessarily zero, but instead lies in a $\sqrt{n}$-neighborhood of 0:

$$g(\theta_0) = c/\sqrt{n}, \qquad c \in \mathcal{C}, \tag{1}$$

where $\mathcal{C} \subseteq \mathbb{R}^{d_g}$ is a known set. Because the misspecification is local, the set $\mathcal{C}$ may allow for misspecification in potentially all moment conditions; we do not require that some elements of c are zero. Our goal is to construct a CI for a scalar $h(\theta_0)$, where $h\colon \mathbb{R}^{d_\theta} \to \mathbb{R}$ is a known function. For example, if we are interested in one of the elements θ_j of θ, we would take $h(\theta) = \theta_j$. More generally, the function h will be nonlinear, as is, for example, generally the case when θ is a vector of supply or demand parameters, and $h(\theta)$ is an elasticity, or some counterfactual. Note that this setup allows (but does not require) both θ_0 and $h(\theta_0)$

to have the same interpretation as in the correctly specified case, so that our CIs may still be interpreted as CIs for the structural parameter, elasticity, or counterfactual of interest. For this interpretation, one typically needs rule out forms of misspecification that affect the mapping $\theta \mapsto h(\theta)$. While we do not formally consider cases in which this mapping itself is misspecified, such cases are covered under a mild generalization of our framework, in which h is a function of both θ and c.

To formalize the notion of asymptotic validity and efficiency of CIs, we will need to allow the true parameter value θ_0 as well as the vector c and the data generating process (and hence the map $\theta \mapsto g(\theta)$) to vary with the sample size. For clarity of exposition, we focus here on the case in which these parameters are fixed. See Section 3.1 and Appendix C for the general case. Under some forms of misspecification, such as functional form misspecification, there may be additional higher-order terms on the right-hand side of (1); our results remain unchanged if this is the case. Again, for clarity of exposition, we focus on the case in which (1) holds exactly.

We assume that the sample moment condition $\hat{g}(\theta)$, constructed using the data $\{w_i\}_{i=1}^n$, satisfies

$$\sqrt{n}(\hat{g}(\theta_0) - g(\theta_0)) \xrightarrow{d} \mathcal{N}(0, \Sigma), \tag{2}$$

where $\xrightarrow{d}$ denotes convergence in distribution as $n \to \infty$. In the GMM model, the population and sample moment conditions are given by $g(\theta) = E[g(w_i, \theta)]$ and $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n g(w_i, \theta)$, respectively, where $g(\cdot, \cdot)$ is a known function. However, to cover other minimum distance problems, we do not require that the moment conditions necessarily take this form. We further assume that the moment condition is smooth enough so that

$$\text{for any } \theta_n = \theta_0 + \mathcal{O}_P(1/\sqrt{n}), \qquad \hat{g}(\theta_n) - \hat{g}(\theta_0) = \Gamma(\theta_n - \theta_0) + o_P(1/\sqrt{n}), \tag{3}$$

where Γ is the $d_g \times d_\theta$ derivative matrix of g at θ_0. Conditions (2) and (3) are standard regularity conditions in the literature on linear and nonlinear estimating equations; see Newey and McFadden (1994) for primitive conditions. Finally, we also assume that h is continuously differentiable with the $1 \times d_\theta$ derivative matrix at θ_0 given by H.

2.1 CIs based on asymptotically linear estimators

Under correct specification, when $\mathcal{C} = \{0\}$, standard estimators $\hat{h}$ of $h(\theta)$ are asymptotically linear in $\hat{g}(\theta_0)$. This will typically extend to our locally misspecified case, so that for some vector $k \in \mathbb{R}^{d_g}$,

$$\sqrt{n}(\hat{h} - h(\theta_0)) = k'\sqrt{n}\hat{g}(\theta_0) + o_P(1) \xrightarrow{d} \mathcal{N}(k'c, k'\Sigma k), \tag{4}$$

where the convergence in distribution follows by (1) and (2). For example, in a GMM model, if we take $\hat{h} = h(\hat{\theta}_W)$ where

$$\hat{\theta}_W = \underset{\theta}{\operatorname{argmin}}\, \hat{g}(\theta)' W \hat{g}(\theta), \tag{5}$$

is the GMM estimator with weighting matrix W, (4) will hold with $k' = -H(\Gamma' W \Gamma)^{-1}\Gamma' W$ (see Newey, 1985). Because the weights k determine the local asymptotic bias of the estimator, Andrews et al. (2017) suggest referring to k as the *sensitivity* of $\hat{h}$.

Let $\hat{k}$ and $\hat{\Sigma}$ be consistent estimates of k and Σ. Then by Slutsky's theorem,

$$\frac{\sqrt{n}(\hat{h} - h(\theta_0))}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}} \xrightarrow{d} \mathcal{N}\left(\frac{k'c}{\sqrt{k'\Sigma k}}, 1\right).$$

Under correct specification, the right-hand side corresponds to a standard normal distribution, and we can form a CI with asymptotic coverage $100 \cdot (1-\alpha)\%$ as $\hat{h} \pm z_{1-\alpha/2}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}/n}$, where $z_{1-\alpha/2}$ is the $1-\alpha/2$ quantile of a $\mathcal{N}(0,1)$ distribution; this is the usual Wald CI.

When we allow for misspecification, this will no longer lead to a valid CI. However, note that the asymptotic bias $k'c/\sqrt{k'\Sigma k}$ is bounded in absolute value by $\overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}$ where $\overline{\text{bias}}_{\mathcal{C}}(k) \equiv \sup_{c\in\mathcal{C}}|k'c|$. Therefore, given c, the z-statistic in the preceding display is asymptotically $\mathcal{N}(t,1)$ where $|t| \leq \overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}$. This leads to the CI

$$\hat{h} \pm \text{cv}_\alpha\left(\frac{\overline{\text{bias}}_{\mathcal{C}}(\hat{k})}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}}\right) \cdot \sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n}, \tag{6}$$

where $\text{cv}_\alpha(\bar{t})$ is the $1-\alpha$ quantile of $|Z|$, with $Z \sim \mathcal{N}(\bar{t}, 1)$. In particular, $\text{cv}_\alpha(0) = z_{1-\alpha/2}$, so that in the correctly specified case, (6) reduces to the usual Wald CI. As we discuss in Section 3, the scaled length of this CI converges to a constant that does not depend on the local misspecification vector c. Following the terminology of Donoho (1994), we refer to (6) as an (asymptotically) fixed length confidence interval (FLCI).

To form a one-sided CI based on an estimator $\hat{h}$ with sensitivity k, one can simply subtract its maximum bias, in addition to the standard error:

$$[\hat{h} - \overline{\text{bias}}_{\mathcal{C}}(\hat{k}) - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}, \infty). \tag{7}$$

One could also form a valid two-sided CI by adding and subtracting the worst-case bias $\overline{\text{bias}}_{\mathcal{C}}(\hat{k})$ from $\hat{h}$, in addition to adding and subtracting $z_{1-\alpha/2}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}/n}$; however, since $\hat{h}$ cannot simultaneously have a large positive and a large negative bias, such CI will be conservative, and longer than the CI in (6).

2.2 Optimal CIs

We can implement an estimator with a desired sensitivity k as a one-step estimator. In particular, let $\hat{\theta}_{\text{initial}}$ be an initial $\sqrt{n}$-consistent estimator of θ_0, let $\hat{k} = k + o_P(1)$ be a consistent estimator of the desired sensitivity, and consider the one-step estimator

$$\hat{h} = h(\hat{\theta}_{\text{initial}}) + \hat{k}'\hat{g}(\hat{\theta}_{\text{initial}}).$$

A Taylor expansion then gives

$$\begin{aligned}\sqrt{n}(\hat{h} - h(\theta_0)) &= H\sqrt{n}(\hat{\theta}_{\text{initial}} - \theta_0) + \hat{k}'\sqrt{n}\hat{g}(\hat{\theta}_{\text{initial}}) + o_P(1) \\ &= (H + \hat{k}'\Gamma)\sqrt{n}(\hat{\theta}_{\text{initial}} - \theta_0) + \hat{k}'\sqrt{n}\hat{g}(\theta_0) + o_P(1),\end{aligned}$$

where the second line follows from (3). Assuming that the sensitivity is chosen so that

$$H = -k'\Gamma, \tag{8}$$

the first term converges in probability to zero, and $\hat{h}$ satisfies (4). The condition (8) ensures that the one-step estimator is asymptotically linear, and that its asymptotic distribution doesn't depend on the initial estimate $\hat{\theta}_{\text{initial}}$. Thus, we can form an asymptotically linear estimator with limiting distribution $\mathcal{N}(k'c, k'\Sigma k)$ for any k satisfying $H = -k'\Gamma$.

To derive the optimal sensitivity, observe that the asymptotic width of the CI in Equation (6) is given by

$$2 \cdot \text{cv}_\alpha\left(\overline{\text{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}\right) \cdot \sqrt{k'\Sigma k}/\sqrt{n}. \tag{9}$$

The length thus doesn't depend on the particular value of c, and it depends on θ only through Σ. Furthermore, it depends on the sensitivity only through the maximum bias, $\overline{\text{bias}}_{\mathcal{C}}(k)$, and the variance $k'\Sigma k$. Therefore, as an alternative to minimizing (9) directly over all sensitivities k, one can first minimize the variance subject to a bound $\overline{B}$ on the worst-case bias,

$$\min_k k'\Sigma k \qquad \text{s.t.} \quad H = -k'\Gamma \qquad \text{and} \quad \sup_{c\in\mathcal{C}}|k'c| \leq \overline{B}, \tag{10}$$

and then vary the bound $\overline{B}$ to find the bias-variance trade-off that leads to the shortest CI. A feasible version of the solution can be implemented as a one-step estimator with plug-in estimates of the quantities Σ, Γ and H. The length of the one-sided CI (7) is infinite by definition, so minimizing length of this CI does not make sense. For the one-sided case, we consider quantiles of excess length as the criterion for choosing a CI. We provide details in Appendix C.

As we discuss in Section 4 and Appendix A, when the set $\mathcal{C}$ is characterized by ℓ_p-constraints, then a closed-form expression for the worst-case bias $\sup_{c\in\mathcal{C}}|k'c|$ is available, and it is computationally trivial to trace out the whole solution path for (10) as a function of $\overline{B}$. More generally, the optimization problem remains tractable if the set $\mathcal{C}$ is convex. Following the usual definition, a set $\mathcal{C}$ is convex if $c, d \in \mathcal{C}$ and $\lambda \in [0,1]$ implies $\lambda c + (1-\lambda)d \in \mathcal{C}$. It follows from Low (1995) that under convexity, the optimization problem (10) can be posed as a convex optimization problem, which is easily solved numerically using convex optimizers (we explain the connection in more detail in Section 3). To describe the mapping, we also for simplicity assume that the set $\mathcal{C}$ is centrosymmetric (i.e. $c \in \mathcal{C}$ implies $-c \in \mathcal{C}$); we show how our CIs can be implemented when $\mathcal{C}$ is asymmetric, such as when one imposes sign restrictions on elements of c, in Appendix C. Given $\delta > 0$, let c_δ, θ_δ be solutions to the convex optimization problem

$$\sup_{\theta,c} H\theta \quad \text{s.t.} \quad c \in \mathcal{C}, \quad (c-\Gamma\theta)'\Sigma^{-1}(c-\Gamma\theta) \leq \delta^2/4. \tag{11}$$

Let

$$k'_\delta = k'_{\delta,\Sigma,\Gamma,H,\mathcal{C}} = \frac{-(c_\delta - \Gamma\theta_\delta)'\Sigma^{-1}}{(c_\delta - \Gamma\theta_\delta)'\Sigma^{-1}\Gamma H'/HH'}. \tag{12}$$

Then the estimator with sensitivity k_δ achieves the lowest variance among all linear estimators with bias upper-bounded by $\overline{\text{bias}}_{\mathcal{C}}(k_\delta) = -k'_\delta c_\delta$. In other words, k_δ solves the problem (10) with $\overline{B} = -k'_\delta c_\delta$. One then simply varies δ, which indexes the relative weight on variance in the tradeoff between the bias and variance, to find the tradeoff leading to the shortest CI length (9).

2.3 Implementation and practical issues

We now summarize the construction of the optimal CIs and discuss some practical implementation issues.

For brevity, we summarize the construction of the optimal CI in terms of the optimization problem (11); if the bias-variance tradeoff (10) can be solved directly, one can use an analogous construction in terms of the sensitivity that minimizes (10) at the optimal bias bound. Given that a researcher has formalized concerns about potential misspecification by forming a set $\mathcal{C}$, the optimal misspecification-robust CI can be constructed as follows:

1. Obtain an initial estimate $\hat{\theta}_{\text{initial}}$ and estimates $\hat{H}$, $\hat{\Gamma}$ and $\hat{\Sigma}$ of H, Γ and Σ.

2. For a given δ, compute $\hat{k}_\delta = k_{\delta,\hat{\Sigma},\hat{\Gamma},\hat{H},\mathcal{C}}$ and $\overline{\text{bias}}_{\mathcal{C}}(\hat{k}_\delta)$ by solving the optimization problem

(11) with $\hat{\Sigma}$ in place of Σ, etc., as described above. Let δ^* minimize the CI length[3] $2\,\mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(\hat{k}_\delta)/\sqrt{\hat{k}_\delta'\hat{\Sigma}\hat{k}_\delta}) \cdot \sqrt{\hat{k}_\delta'\hat{\Sigma}\hat{k}_\delta}$ over δ.

3. Let $\hat{h}_\delta = h(\hat{\theta}_{\text{initial}}) + \hat{k}_\delta'\hat{g}(\hat{\theta}_{\text{initial}})$. The misspecification-robust CI is given by

$$\hat{h}_{\delta^*} \pm \hat{\chi}^*_{\delta^*}, \qquad \hat{\chi}^*_{\delta^*} = \mathrm{cv}_\alpha\left(\overline{\mathrm{bias}}_{\mathcal{C}}(\hat{k}_{\delta^*})/\sqrt{\hat{k}_{\delta^*}'\hat{\Sigma}\hat{k}_{\delta^*}}\right) \cdot \sqrt{\hat{k}_{\delta^*}'\hat{\Sigma}\hat{k}_{\delta^*}/n}, \tag{13}$$

and the optimal sensitivity is given by $\hat{k}_{\delta^*}$.

Remark 2.1. The above algorithm gives a generic procedure based on one-step estimators $\hat{h}_\delta$ that gives an asymptotically valid and optimal CI. Due to concerns about finite-sample behavior (analogous to concerns about finite sample behavior of one-step estimators in the correctly specified case), one may prefer using a different estimator that is asymptotically equivalent to $\hat{h}_\delta$. In general, one can implement an estimator with sensitivity k as a GMM or minimum distance estimator by using an appropriate weighting matrix, so that one can in particular replace $\hat{h}_\delta$ by $h(\hat{\theta}_W)$, with the weighting matrix W appropriately chosen. To give the formula for the weighting matrix, let $\Gamma_\perp$ denote a $d_g \times (d_g - d_\theta)$ matrix that's orthogonal to Γ, so that $\Gamma_\perp'\Gamma = 0$, and let $\hat{\Gamma}_\perp$ denote a consistent estimate. Let S denote a $d_g \times d_\theta$ matrix that satisfies $S'\hat{\Gamma} = -I$ and $\hat{k}_\delta = S\hat{H}'$. Then we can set $W = SW_1S' + \hat{\Gamma}_\perp W_2 \hat{\Gamma}_\perp'$ for some non-singular matrix W_1, and an arbitrary conformable matrix W_2. It can be verified by simple algebra that $\hat{\theta}_W$ will have sensitivity $k_{\delta,\Sigma,\Gamma,H,\mathcal{C}}$. We discuss this GMM implementation of the optimal sensitivity in the context of some of our specific applications in Section 5.

In many cases, a researcher may be interested in multiple sets $\mathcal{C}$, or they may not know which set $\mathcal{C}$ they are ex ante interested in. This issue may manifest itself in two different ways, as we discuss in the next two remarks.

Remark 2.2 (Known form but unknown magnitude of misspecification)**.** If a researcher has a particular form of misspecification in mind, this determines the shape of the set $\mathcal{C}$, but not necessarily the magnitude of the potential misspecification. For concreteness, suppose that we wish to examine sensitivity to the failure of the first moment while assuming correct specification of the remaining moments. In this case, one would set $\mathcal{C} = \mathcal{C}(M) = [-M, M] \times \{0\} \times \cdots \times \{0\}$. It would be desirable to use a data-driven procedure to determine M. Unfortunately, as we discuss in Section 3, our results show that this is impossible when constructing CIs: one has to specify M a priori. In light of this result, we recommend computing the optimal CI for each M and plotting it as a function of M. The resulting

[3]The critical value $\mathrm{cv}_\alpha(b)$ can easily be computed in statistical software as the square root of the $1-\alpha$ quantile of a non-central χ^2 distribution with 1 degree of freedom and non-centrality parameter b^2.

plot can be used for sensitivity analysis to see how large M needs to be before a given result breaks down. We construct such a plot in Figure 3 in the context of our empirical application in Section 6. Section 4 discusses different ways of constructing of the sets $\mathcal{C}(M)$, indexed by the magnitude of misspecification M, when misspecification affects multiple moments, and Section 5 gives suggestions for the form of this set in particular applications.

To aid with the interpretation of such graphs, it is helpful to present a discussion of which values of M one may deem plausible. This will typically depend on the economic interpretation of misspecification in the model at hand, and will require analysis on a case by case basis. We discuss plausible values of M in our application in Section 6, and we refer the reader to Conley et al. (2012) and Andrews et al. (2017) for additional examples and discussion. One can also use measures of statistical distance such as the probability of detecting that the model is misspecified to aid with interpretation of M, as suggested in Hansen and Sargent (2008) or Bonhomme and Weidner (2018).

While it is not possible to determine M automatically, it is possible to use specification tests to obtain a lower CI $[M_{\min}, \infty]$ that contains M with a prespecified probability. We develop such tests by generalizing the J-test of overidentifying restrictions in Appendix B. The lower bound for M can then be reported along with the plot of the optimal CI as a function of M.

Remark 2.3 (Multiple forms of misspecification)**.** If the researcher is unsure about the form of misspecification they are most concerned about, it is useful to consider multiple forms of misspecification to determine which forms of misspecification the results are the most sensitive to. We give such comparison in Figure 2 in the context of our empirical application in Section 6. In addition, given that the CI in Equation (13) can easily be computed for any set $\mathcal{C}$ using the initial estimate $\hat{\theta}_{\text{initial}}$ along with $\hat{\Sigma}$, $\hat{\Gamma}$ and $\hat{H}$, we recommend that researchers report the estimates $\hat{H}$, $\hat{\Gamma}$ and $\hat{\Sigma}$ along with estimates $\hat{\theta}$ of the parameter vector θ and $\hat{h} = h(\hat{\theta})$ of the object of interest (if the number of moments is large, this can be done in a supplementary appendix or as an easily accessible part of the replication code). This allows the reader to easily compute CIs under the forms and magnitude of misspecification that the reader is most concerned about.

Andrews et al. (2017) recommend reporting the sensitivity $\hat{k}$ of an estimator $\hat{h}$ along with point estimates and standard errors, as this allows the reader to estimate the local asymptotic bias $\hat{k}'c$ of the estimator under different misspecification vectors c. Given the sensitivity estimate, it is also straightforward for the reader to compute the misspecification robust CI (6) based on $\hat{h}$ and assess the effect of misspecification on inference. However, as we demonstrate in the empirical application in Section 6, adjusting the estimator so that

its sensitivity is optimal under local misspecification may lead to substantially tighter CIs. Reporting the objects $\hat{\Sigma}$, $\hat{\Gamma}$ and $\hat{H}$, as suggested by Remark 2.3, allows the reader to directly report these optimal CIs and draw potentially much sharper conclusions. Given that plug-in estimates of $\hat{k}$ typically require calculating these objects anyway (recall that for GMM estimators, the sensitivity takes the form $-\hat{H}(\hat{\Gamma}'W\hat{\Gamma})^{-1}\hat{\Gamma}'W$ for some weighting matrix W, with $W = \hat{\Sigma}^{-1}$ corresponding to the weighting that's optimal under correct specification), reporting these objects is no harder than reporting the sensitivity $\hat{k}$.

Remark 2.4 (Other performance criteria)**.** In addition to constructing a CI, one may be interested in a point estimate of $h(\theta_0)$, using mean squared error (MSE) as the criterion. The steps to forming the MSE optimal point estimate are exactly the same as above, except that, rather than minimizing CI length in Step 2, one chooses δ to minimize $\overline{\text{bias}}_{\mathcal{C}}(\hat{k}_\delta)^2 + \hat{k}_\delta'\hat{\Sigma}\hat{k}_\delta$. Similar ideas apply to other criteria, such as mean absolute deviation or quantiles of excess length of one-sided CIs (discussed in Appendix C). If δ is chosen differently in Step 2, the CI computed in Step 3 will be longer than the one computed at δ^*, but it will still have correct coverage.

3 Efficiency bounds and near optimality

The CI given in (13) has the apparent defect that the local misspecification vector c is reflected in the length of the CI only through the a priori restriction $\mathcal{C}$ imposed by the researcher. Thus, if the researcher is conservative about misspecification, the CI will be wide, even if it "turns out" that c is in fact much smaller than the a priori bounds defined by $\mathcal{C}$. Moreover, this approach requires the researcher to explicitly specify the set $\mathcal{C}$, including any tuning parameters such as the parameter M in Remark 2.2. One may therefore seek to improve upon this CI by forming a random-length CI, the length of which would depend on the data via an estimate of the magnitude of c, or estimates of the tuning parameters. Similarly, it may be restrictive to require that the CI be centered at an asymptotically linear estimator: the vector $\hat{k}$ must converge in probability to a vector that does not depend on c, which rules out, for example, using a J-test to decide which moments to use.

The main result of this section shows that, when $\mathcal{C}$ is convex and centrosymmetric, the scope for improving on the CI in (13) is nonetheless severely limited: no sequence of CIs that maintain coverage under all local misspecification vectors $c \in \mathcal{C}$ can be substantially tighter, even under correct specification. This result can be interpreted as translating results from a "limiting experiment" that is an extension of the linear regression model. We first give a heuristic derivation of this limiting experiment and explain our result in the context of this limiting experiment. We then present the formal asymptotic result, and discuss its

implications in some familiar settings. Readers who are interested only in implementing the methods, rather than efficiency results, can skip this section.

We restrict attention in this section to the GMM model, in which $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$, and we further restrict the data $\{w_i\}_{i=1}^{n}$ to be independent and identically distributed (iid). Similar to semiparametric efficiency theory in the standard, correctly specified case, this facilitates parts of the formal statements and proofs, such as the definition of the set of distributions under which coverage is required and the construction of least favorable submodels. We expect that analogous results could be obtained in other settings.

3.1 Limiting experiment

As discussed in Section 2.2, we can form estimators with asymptotic distribution $\mathcal{N}(k'c, k'\Sigma k)$ for any k satisfying (8). This suggests that the problem of constructing an asymptotically valid CI for $h(\theta)$ in the model (1) is asymptotically equivalent to the problem of constructing a CI for the parameter $H\theta$ in the approximately linear model

$$Y = -\Gamma\theta + c + \Sigma^{1/2}\varepsilon, \quad c \in \mathcal{C}, \qquad \varepsilon \sim \mathcal{N}(0, I), \tag{14}$$

where Γ, H and $\Sigma^{1/2}$ are known, and we observe Y. One can think of this model as an "approximately" linear regression model, with $-\Gamma$ playing the role of the design matrix of the (fixed) regressors, and c giving the approximation error. The analog of the asymptotically linear estimator $\hat{h}$ in (4) is the linear estimator $k'Y$. To see the analogy, note that $k'Y - H\theta$ is distributed $\mathcal{N}((-k'\Gamma - H)\theta + k'c, k'\Sigma k)$, and restricting ourselves to estimators that do not have infinite worst-case bias when θ is unrestricted gives the condition $H = -k'\Gamma$. This model dates back at least to Sacks and Ylvisaker (1978), who considered estimation in this model when $\mathcal{C}$ is a rectangular set and Σ is diagonal.

In the limiting experiment, the analog of the CI (6) is given by the linear FLCI $k'Y \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma k}) \cdot \sqrt{k'\Sigma k}$. The problem of constructing the shortest linear FLCI in the limiting experiment is a special case of a problem considered by Donoho (1994), whose results imply that the optimal CI has the form

$$k_{\delta^*}'Y \pm \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k_{\delta^*})/\sqrt{k_{\delta^*}'\hat{\Sigma}k_{\delta^*}}) \cdot \sqrt{k_{\delta^*}'\Sigma k_{\delta^*}}, \tag{15}$$

where k_δ is given by (12), and $\delta^* = \mathrm{argmin}_{\delta>0} 2\,\mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k_\delta)/\sqrt{k_\delta'\Sigma k_\delta}) \cdot \sqrt{k_\delta'\Sigma k_\delta}$ is chosen to minimize the CI length. The FLCI given in (13) is an analog of this CI, and the connection between the bias-variance optimization problem (10), and the convex optimization problem (11) in Section 2.2 follows from Low (1995).

The CI in (15) takes a familiar form in the special case in which $\mathcal{C}$ is a linear subspace of $\mathbb{R}^{d_g}$, so that for some $d_g \times d_\gamma$ full-rank matrix B with $d_\gamma \leq d_g - d_\theta$, $\mathcal{C} = \{B\gamma \colon \gamma \in \mathbb{R}^{d_\gamma}\}$. Let $B_\perp$ denote a $d_g \times (d_g - d_\gamma)$ matrix that's orthogonal to B. Then for any $\delta > 0$, $k'_\delta = k'_{LS,B}$, where

$$k'_{LS,B} = -H(\Gamma' B_\perp (B'_\perp \Sigma B_\perp)^{-1} B'_\perp \Gamma)^{-1} \Gamma' B_\perp (B'_\perp \Sigma B_\perp)^{-1} B'_\perp \tag{16}$$

is the sensitivity of the GLS estimator after pre-multiplying (14) by $B'_\perp$, (which effectively picks out the observations with zero misspecification). Since this estimator is unbiased, the CI in (15) becomes $k'_{LS,B} Y \pm z_{1-\alpha/2} \sqrt{k'_{LS,B} \Sigma k_{LS,B}}$.

Like the asymptotic FLCI (13), the CI in (15) has the potential drawback that its length is determined by the worst possible misspecification in $\mathcal{C}$. Thus, one may suspect that one could improve upon this CI substantially, particularly when $\mathcal{C}$ is large and c turns out to be close to zero. As the best-case scenario for such improvements, suppose that the researcher guesses correctly that the model is correctly specified, but to ensure validity of the CI if the guess is wrong, the researcher must still form a CI that is valid under all misspecification vectors in $\mathcal{C}$. To make the problem even easier, suppose the researcher also guesses correctly that θ is equal to a particular value θ^*. That is, consider the problem: among confidence sets with coverage at least $1 - \alpha$ for all $\theta \in \mathbb{R}^{d_\theta}$ and $c \in \mathcal{C}$, minimize expected length when $\theta = \theta^*$ and $c = 0$. We allow for confidence sets that are not intervals, in which case length is defined as Lebesgue measure (which makes such an approach even more favorable relative to the linear FLCI, the latter being constrained to be an interval). Let $\kappa_*(H, \Gamma, \Sigma, \mathcal{C})$ denote the ratio of this optimized expected length relative to the length of the FLCI in (15) (it can be shown that this ratio does not depend on θ^*).

If $\mathcal{C}$ is convex, a formula for $\kappa_*(H, \Gamma, \Sigma, \mathcal{C})$ follows from applying the general results in Corollary 3.3 in Armstrong and Kolesár (2018) to the limiting model. If $\mathcal{C}$ is also centrosymmetric, this formula is given by

$$\kappa_*(H, \Gamma, \Sigma, \mathcal{C}) = \frac{(1-\alpha) E\left[\omega(2(z_{1-\alpha} - Z)) | Z \leq z_{1-\alpha}\right]}{2 \min_\delta \mathrm{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right) \omega'(\delta)}, \tag{17}$$

where $Z \sim \mathcal{N}(0,1)$ and $\omega(\delta)$ is two times the optimized value of (11). Furthermore, we show in Theorem C.7 that the right-hand side is lower-bounded by $(z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha))/z_{1-\alpha/2}$, where $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$ for any H, Γ, Σ and $\mathcal{C}$, where $\phi(\cdot)$ denotes the standard normal density. For $\alpha = 0.05$, this universal lower bound evaluates to 71.7%. Evaluating κ_* for particular choices of H, Γ, Σ, and $\mathcal{C}$ often yields even higher efficiency.

If $\mathcal{C}$ is a linear subspace, then $\omega(\delta)$ is linear, and

$$\kappa_*(H,\Gamma,\Sigma,\mathcal{C}) = \frac{(1-\alpha)z_{1-\alpha}+\phi(z_{1-\alpha})}{z_{1-\alpha/2}} \geq \frac{z_{1-\alpha}}{z_{1-\alpha/2}}, \tag{18}$$

where the lower bound follows since $\phi(z_{1-\alpha}) \geq \alpha z_{1-\alpha}$ by the Gaussian tail bound $1-\Phi(x) \leq \phi(x)/x$ for $x > 0$. This bound corresponds to the bound derived by Pratt (1961) for the case of a univariate normal mean, and at $\alpha = 0.05$, it evaluates to 84.99%. The CI with the shortest expected length at a given θ^* is obtained by inverting uniformly most powerful tests of the null $H\theta = h_0$ against the alternative $H\theta = H\theta^*$ (which doesn't vary with the null), and these tests are given by one-sided z-tests based on $k'_{LS,B}Y$. Intuitively, the maximum gain from directing power in this way over the usual procedure is that it turns a two-sided testing problem into a one-sided problem, which is why the ratio of a one-sided to a two-sided critical value gives a lower bound. Furthermore, it follows from Joshi (1969) that the CI $k'_{LS,B}Y \pm z_{1-\alpha/2}\sqrt{k'_{LS,B}\Sigma k_{LS,B}}$ is the unique CI that achieves minimax expected length. Thus, not only is the scope for improvement at a particular θ^* bounded by (18), any CI with shorter expected length at some θ^* must necessarily perform worse elsewhere in the parameter space.

For the one-sided CI (7), the analogous CI in the limiting experiment is $[k'Y - \overline{\text{bias}}_{\mathcal{C}}(k) - z_{1-\alpha}\sqrt{k'\Sigma k}, \infty)$, and, as we discuss in Appendix C, to choose the optimal sensitivity k, one can consider optimizing a given quantile of its worst-case excess length. Since this approach is based on optimizing the worst-case quantile over $\mathcal{C}$, one may try to use a different CI in order to improve performance for small c by instead optimizing quantiles of excess length under correctly specified models (i.e. when $c = 0$). The best-case scenario for such improvements is to optimize the CI at $c = 0$ and at a particular θ^*. When $\mathcal{C}$ is convex and centrosymmetric, the results in Armstrong and Kolesár (2018) show that the scope for such improvement is severely limited in the one-sided case as well. See Appendix C for details and an analog of the efficiency bound in (17). If $\mathcal{C}$ is a linear subspace, then optimizing quantiles of worst-case excess length yields the CI $[k'_{LS,B}Y - z_{1-\alpha}\sqrt{k'_{LS,B}\Sigma k_{LS,B}}, \infty)$, independently of the quantile one is optimizing. Furthermore, the efficiency bound implies that this one-sided CI is in fact fully optimal over all quantiles of excess length and all values of θ, c in the local parameter space.

The high efficiency for the FLCI (15) in the limiting experiment (even in the case that seems most favorable for improving on this CI) suggests that the CI in (13) should be highly efficient in an asymptotic sense. Theorem 3.1, stated in the next section, uses the analogy with the approximately linear model (14) along with Le Cam-style arguments involving least favorable submodels to show that this bound indeed translates to the locally misspecified

GMM model. For one-sided CIs, we state an analogous result in Appendix C. We discuss the implications of these results in Section 3.3.

3.2 Asymptotic efficiency bound

To make precise our statements about coverage and efficiency, we need the notion of uniform (in the underlying distribution) coverage of a confidence interval. This requires additional notation, which we now introduce. Let $\mathcal{P}$ denote a set of distributions P of the data $\{w_i\}_{i=1}^n$, and let $\Theta_n \subseteq \mathbb{R}^{d_\theta}$ denote the parameter space for θ. We require coverage for all pairs $(\theta, P) \in \Theta_n \times \mathcal{P}$ such that $\sqrt{n} g_P(\theta) \in \mathcal{C}$, where the subscript P on the population moment condition makes it explicit that it depends on the distribution of the data.[4] Letting $\mathcal{S}_n = \{(\theta, P) \in \Theta_n \times \mathcal{P} : \sqrt{n} g_P(\theta) \in \mathcal{C}\}$ denote this set, the condition for coverage at confidence level $1-\alpha$ can be written

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{S}_n} P(h(\theta) \in \mathcal{I}_n) \geq 1-\alpha. \tag{19}$$

We say that a confidence set $\mathcal{I}_n$ is asymptotically valid (uniformly over $\mathcal{S}_n$) at confidence level $1-\alpha$ if this condition holds.[5]

Among two-sided CIs of the form $\hat{h} \pm \hat{\chi}$ that are asymptotically valid, we prefer CIs that achieve better expected length. To avoid issues with convergence of moments, we use truncated expected length, and define the asymptotic expected length of a two-sided CI at $P_n \in \mathcal{P}$ as $\liminf_{T\to\infty} \liminf_{n\to\infty} E_{P_n} \min\{\sqrt{n} \cdot 2\hat{\chi}, T\}$, where E_P denotes expectation under P.

We are now ready to state the main efficiency result.

Theorem 3.1. *Suppose that $\mathcal{C}$ is convex and centrosymmetric. Let $\hat{h}_{\delta^*}$ and $\hat{\chi}^*_{\delta^*}$ be formed as in Section 2.3. Suppose that Assumptions C.2, C.3, C.5, C.6 and C.7 in Appendix C hold. Suppose that the data $\{w_i\}_{i=1}^n$ are iid under all $P \in \mathcal{P}$. Let (θ^*, P_0) be correctly specified (i.e. $g_{P_0}(\theta^*) = 0$) such that $\mathcal{P}$ contains a submodel through P_0 satisfying Assumption C.1. Then:*

(i) The CI $\hat{h}_{\delta^*} \pm \hat{\chi}^*_{\delta^*}$ is asymptotically valid, and its half-length $\hat{\chi}^*_{\delta^*}$ satisfies $\sqrt{n}\hat{\chi}^*_{\delta^*} = \chi(\theta, P) + o_P(1)$ uniformly over $(\theta, P) \in \mathcal{S}_n$ where

$$\chi(\theta, P) = \min_k \mathrm{cv}_\alpha(\overline{\mathrm{bias}}_{\mathcal{C}}(k)/\sqrt{k'\Sigma_{\theta,P}k})\sqrt{k'\Sigma_{\theta,P}k}$$

[4] To be precise, we should also subscript all other quantities such as Γ and Σ by P. To prevent notational clutter, we drop this index in the main text unless it causes confusion.

[5] In general, θ_0 and $h(\theta_0)$ may be set identified for a given sample size n (although our assumptions imply that the identified set will shrink at a root-n rate). The coverage requirement (19) states that the CI must cover points in the identified set for $h(\theta)$, as in Imbens and Manski (2004); see Appendix C.

with $\overline{\text{bias}}_{\mathcal{C}}(k)$ *calculated with* $\Gamma = \Gamma_{\theta,P}$ *and* $H = H_\theta$.

(ii) For any other asymptotically valid CI $\hat{h} \pm \hat{\chi}$,

$$\frac{\liminf_{T\to\infty} \liminf_{n\to\infty} E_{P_0} \min\{\sqrt{n} \cdot 2\hat{\chi}, T\}}{2\chi(\theta^*, P_0)} \geq \kappa_*(H_{\theta^*}, \Gamma_{\theta^*,P_0}, \Sigma_{\theta^*,P_0}, \mathcal{C}),$$

where $\kappa_*(H, \Gamma, \Sigma, \mathcal{C})$ *is defined in* (17). *Furthermore, for any* H, Σ, Γ, *and* $\mathcal{C}$, κ_* *admits the universal lower bound* $(z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha))/z_{1-\alpha/2}$, *where* $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$ *and* $\phi(\cdot)$ *denotes the standard normal density.*

The proof for this theorem is given in Appendix C, which also gives an analogous result for one-sided confidence intervals. In the supplemental materials, we also give primitive conditions for the misspecified linear IV model. For the lower bound, the conditions amount to mild regularity conditions on the least favorable submodel, and in the supplemental materials, we provide a general way of constructing a submodel satisfying these conditions.

The universal lower bound on κ_* is new and may be of independent interest. For $\alpha = 0.05$, it evaluates to 71.7%. The universal lower bound is sharp in the sense that there exist Γ, Σ, H and $\mathcal{C}$ for which κ_* equals this lower bound. In particular applications, the efficiency bound κ_* can be computed at estimates of Γ, Σ and H, and often, this gives much higher efficiencies. We illustrate these bounds in the empirical application in Section 6.

3.3 Discussion

To help build intuition for the efficiency bound in Theorem 3.1, and to relate this result to the literature, we now consider some special cases. We first discuss the (standard) correctly specified case. Second, we consider the case in which some moments are known to be valid, and the misspecification in the remaining moments is unrestricted. This case may be of interest in its own right. Finally, we discuss the general case.

3.3.1 Correctly specified case

Suppose that $\mathcal{C} = \{0\}$. This is in particular a linear subspace of $\mathbb{R}^{d_g}$, with $B = 0$, and $B_\perp = I$, the $d_g \times d_g$ identity matrix. The approximately linear model (14) reduces to a standard linear regression model with known covariance matrix, so that the GLS estimator $k'_{LS,0}Y$, with $k_{LS,0}$ given in (16) (with $B = 0$), is the best unbiased linear estimator in the limiting experiment (14) by the Gauss-Markov theorem. Furthermore, this estimator minimizes the maximum mean-squared error (MSE)—it is minimax.[6] In the moment condition model,

[6]This follows, for example, by applying Proposition 8.6 in van der Vaart (1998) to the sufficient statistic $(\Gamma'\Sigma^{-1}\Gamma)^{-1}\Gamma'\Sigma^{-1}Y$.

an estimator with this sensitivity can be implemented as $h(\hat{\theta}_{\Sigma^{-1}})$, where $\hat{\theta}_{\Sigma^{-1}}$ is the GMM estimator with the optimal weighting matrix $W = \Sigma^{-1}$, defined in (5). However, since in the derivation of the limiting experiment, we have restricted attention to asymptotically linear estimators that satisfy (8), it is unclear whether this minimax optimality carries over to the moment condition model. The local asymptotic minimax bound in Chamberlain (1987) shows that it indeed does, so that $h(\hat{\theta}_{\Sigma^{-1}})$ is asymptotically minimax under the MSE criterion.

Next, consider inference. In the limiting experiment, for testing the null hypothesis $H\theta = h_0$ against the one-sided alternative $H\theta \geq h_0$, the one-sided z-statistic based on $k'_{LS,0}Y$ is uniformly most powerful (van der Vaart, 1998, Proposition 15.2). Inverting these tests yields the CI $[k'_{LS,0}Y - z_{1-\alpha}\sqrt{k'_{LS,0}\Sigma k_{LS,0}}, \infty)$. Since the underlying tests are uniformly most powerful, this CI achieves the shortest excess length, simultaneously for all quantiles and all possible values of the parameter θ. For two-sided CIs, the results described in Section 3.1 imply that the CI $h'_{LS,0}Y \pm z_{1-\alpha/2}\sqrt{k'_{LS,0}\Sigma k_{LS,0}}$ is the unique CI that achieves minimax expected length, and that this CI has efficiency $((1-\alpha)z_{1-\alpha} + \phi(z_{1-\alpha}))/z_{1-\alpha/2}$ relative to a CI that optimizes its expected length at a single value θ^* of θ when indeed $\theta = \theta^*$.

Applying Theorem 3.1 to the case $\mathcal{C} = \{0\}$ gives an asymptotic version of the two-sided efficiency bound. Furthermore, the CI in Theorem 3.1 reduces to the usual two-sided CI based on $\hat{\theta}_{\Sigma^{-1}}$. Thus, in this case, Theorem 3.1 shows that very little can be gained over the usual two-sided CI by optimizing the CI relative to a particular distribution P_0. Results in the appendix give an analogous result for one-sided CIs. In the one-sided case, this asymptotic result is essentially a version of a classic result from the semiparametric efficiency literature for one-sided tests, applied to CIs (see Chapter 25.6 in van der Vaart, 1998). In the two-sided case, the result is, to our knowledge, new.

3.3.2 Some valid and some invalid moments

Consider now the case in which the first $d_g - d_\gamma$ moments are known to be valid, with the potential misspecification for the remaining d_γ moments unrestricted. Then $\mathcal{C} = \{(0', \gamma')' : \gamma \in \mathbb{R}^{d_\gamma}\}$ corresponds to a linear subspace with B given by the last d_γ columns of the identity matrix, and $B_\perp$ given by the first $d_g - d_\gamma$ columns.

Because under this setup, the mean for the last d_γ observations is unrestricted, it follows by the same arguments as in Section 3.3.1, that the GLS estimator $k'_{LS,B}Y$ based only on the observations with no misspecification is best unbiased and minimax. This property can again be shown to carry over to the moment condition model, so that the GMM estimator $h(\hat{\theta}_{W(B)})$, with $W(B) = B'_\perp(B'_\perp \Sigma B_\perp)^{-1}B'_\perp$ is a GMM estimator that only uses the moments known to be valid, is asymptotically minimax. However, under a weighted MSE criterion,

if the weights put a sufficient mass on values of γ that are close to zero, if one does not require unbiasedness, a different estimator may be preferred, such as various shrinkage or pre-testing estimators that optimize their performance at values of γ close to zero, at the expense of worse performance for larger values of γ. In the context of the moment condition model, such estimators have been recently studied in Liao (2013), Cheng and Liao (2015), and DiTraglia (2016).

Next, consider inference. The one-sided CI based on $k'_{LS,B}Y$ achieves the shortest excess length, simultaneously for all quantiles and all possible values of the parameter θ. The two-sided CI $k'_{LS,B}Y \pm z_{1-\alpha/2}\sqrt{k'_{LS,B}\Sigma k_{LS,B}}$ is optimal in the same sense as the usual CI in Section 3.3.1: it achieves minimax expected length, and its efficiency, relative to a CI that optimizes its length at a single θ^* and $\gamma = 0$, is lower-bounded by $z_{1-\alpha}/z_{1-\alpha/2}$. Theorem 3.1 formally translates the efficiency bound from the limiting model to the GMM model, so that the usual two-sided CI based on $h(\hat{\theta}_{W(B)})$ is asymptotically efficient in the same sense as the usual CI based on $h(\hat{\theta}_{\Sigma^{-1}})$ discussed in Section 3.3.1 under correct specification. Just as with the results in Section 3.3.1, this asymptotic result is, to our knowledge, new. The one-sided analog follows from the results in Appendix C. These results stand in sharp contrast to the results for estimation, where the MSE improvement at small values of γ may be substantial.

An important consequence of these results is that asymptotically valid one-sided CIs based on shrinkage or model-selection procedures, such as one-sided versions of the CIs proposed in Andrews and Guggenberger (2009), DiTraglia (2016) or McCloskey (2017) must have *worse* excess length performance than the usual one-sided CI based on the GMM estimator $h(\hat{\theta}_{W(B)})$ that uses valid moments only. While it is possible to construct two-sided CIs that improve upon the usual CI based on $h(\hat{\theta}_{W(B)})$ at particular values of θ and γ, the scope for such improvement is smaller than the ratio of one- to two-sided critical values. Furthermore, any such improvement must come at the expense of worse performance at other points in the parameter space.[7] Therefore, in order to tighten CIs based on valid moments only, it is *necessary* to make a priori restrictions on the potential misspecification of the remaining moments.

3.3.3 General case

According to the results in Section 3.3.2, one must place a priori bounds on the amount of misspecification in order to use misspecified moments. This leads us to the general case, where we place the local misspecification vector c in some set $\mathcal{C}$ that is not necessarily

[7] Consistently with these results, in a simulation study considered in DiTraglia (2016), the post-model selection CI that he proposes is shown to be wider on average than the usual CI around a GMM estimator that uses valid moments only.

a linear subspace. One can then form a CI centered at an estimate formed from these misspecified moments using the methods in Section 2.3. In the case where $\mathcal{C}$ is convex and centrosymmetric, Theorem 3.1 shows that this CI is near optimal, in the sense that no other CI can improve upon it by more than a factor of κ_*, even in the favorable case of correct specification. Since the width of the CI is asymptotically constant under local parameter sequences $\theta_n \to \theta^*$ and sufficiently regular probability distributions $P_n \to P_0$ (for example, $P_n \to P_0$ along submodels satisfying Assumption C.1), this also shows that the CI is near optimal in a local minimax sense. In the general case, Theorem 3.1, as well as the analogous results for one-sided CIs in Appendix C are, to our knowledge, new.

In Section 4, we discuss particular examples of the set $\mathcal{C}$ that can be used in sensitivity analysis. These sets typically depend on an a priori bound M, such as when $\mathcal{C} = \mathcal{C}(M) = \{B\gamma \colon \|\gamma\| \leq M\}$ for some norm $\|\cdot\|$. Rather than choosing M a priori, one may wish to use a data-driven estimate of M, for example, by using a first-stage J test to assess plausible magnitudes of misspecification. Formally, one would seek a CI that is valid over $\mathcal{C}(\overline{M})$ while improving length when in fact $\|\gamma\| \ll \overline{M}$, where $\overline{M}$ is some initial conservative bound. When $\mathcal{C}$ is convex and centrosymmetric, Theorem 3.1 shows that the scope for such improvements is severely limited: the average length of any such CI cannot be much smaller than the CI that uses the most conservative choice $\overline{M}$, even when $c = 0$. The impossibility of choosing M based on the data is related to the impossibility of using specification tests to form an upper bound for M. On the other hand, it is possible to obtain a lower bound for M using such tests. We develop lower CIs for M in Appendix B. This test can be used as a diagnostic to check that the magnitude of M chosen by the resereach is not too small.

3.4 Extensions: asymmetric constraints and constraints on θ

In the case where the set $\mathcal{C}$ is convex but asymmetric (such as when $\mathcal{C}$ includes bounds on a norm as well as sign restrictions, or when $\mathcal{C}$ includes equality and sign restrictions, as in Moon and Schorfheide (2009)), one can still apply bounds from Armstrong and Kolesár (2018) to the limiting model described in Section 3.1. Our general asymptotic efficiency bounds in Appendix C translate these results to the locally misspecified GMM model so long as $\mathcal{C}$ is convex. Since the negative implications for efficiency improvements under correct specification use centrosymmetry of $\mathcal{C}$, introducing asymmetric restrictions, such as sign restrictions, is one possible way of getting efficiency improvements at some smaller set $\mathcal{D} \subseteq \mathcal{C}$ while maintaining coverage over $\mathcal{C}$. We derive efficiency bounds and optimal CIs for this problem in Appendix C. Interestingly, the scope for efficiency improvements can be different for one- and two-sided CIs, and can depend on the direction of the CI in this case. To get some

intuition for this, note that, in the instrumental variables model with a single instrument and single endogenous regressor, sign restrictions on the covariance of an instrument with the error term can be used to sign the direction of the bias of the instrumental variables estimator, which is useful for forming a one-sided CI only in one direction.

Finally, while we focus on restrictions on c, one can also incorporate local restrictions on θ. Our general results in Appendix C give efficiency bounds that cover this case. Similar to the discussion above, these results have implications for using prior information about θ to determine the amount of misspecification, or to shrink the width of a CI directly. In particular, while it is possible to use prior information on θ (say, an upper bound on $\|\theta\|$ for some norm $\|\cdot\|$) to shrink the width of the CI, the width of the CI and the estimator around which it is centered must depend on the a priori upper bounds on the magnitude of θ and c when this prior information takes the form of a convex, centrosymmetric set for $(\theta', c')'$. This rules out, for example, choosing the moments based on whether the resulting estimate for θ is in a plausible range.

4 Solutions for particular choices of $\mathcal{C}$

This section gives examples of sets $\mathcal{C}$ that can be used to describe a researcher's beliefs about potential misspecification. We give intuition for how this affects the optimal sensitivity k and, in cases where it is available, provide an analytic form for the optimal sensitivity. Derivations and additional details are relegated to Appendix A.

4.1 Misspecification of a single moment

If one is interested in misspecification of a particular element of $g(\theta) = (g_1(\theta), \ldots, g_{d_g}(\theta))'$, one can take $\mathcal{C}$ to allow for misspecification of only this element. For example, if the condition $g_1(\theta) = 0$ is suspected to hold only approximately, but the other conditions $g_2(\theta) = 0, \ldots, g_{d_g}(\theta) = 0$ are deemed plausible, one can use the set $\mathcal{C} = [-M, M] \times \{0\} \times \cdots \times \{0\}$ as in Remark 2.2. As discussed in Remark 2.2, the constant M can be varied to determine how sensitive a given result is to failure of the moment condition. We illustrate this approach in Section 6.

With a single misspecified moment, the worst-case bias of an asymptotically linear estimator with sensitivity k is given simply by $\overline{\text{bias}}_{\mathcal{C}}(k) = M|k_1|$, so that the Lagrangian for the bias-variance trade-off in (10) takes the form $\min_k(k'\Sigma k + \lambda M k_1^2)$ s.t. $H = -k'\Gamma$, which is minimized at $k_\lambda' = -H'(\Gamma' W_\lambda \Gamma)^{-1}\Gamma' W_\lambda$ where $W_\lambda = (\lambda M^2 e_1 e_1' + \Sigma)^{-1}$, and e_1 denotes the first unit vector. Thus, one can implement the optimal sensitivity as a GMM estimator with

weighting matrix W_{λ^*}, where λ^* minimizes $2\,\mathrm{cv}_\alpha(M|k_{\lambda,1}|/\sqrt{k_\lambda'\Sigma k_\lambda})\sqrt{k_\lambda'\Sigma k_\lambda}$ over λ.

4.2 Misspecification of multiple moments

To allow for misspecification of multiple components of $g(\theta)$, it is computationally convenient and flexible to consider sets of the form $\mathcal{C} = \{B\gamma\colon \|\gamma\| \leq M\}$, where B is a matrix with dimensions $d_g \times d_\gamma$, $\|\cdot\|$ is some norm or semi-norm, and the bound M can again be varied to determine the sensitivity of a given result. When $B = e_1$, this reduces to the previous example. Setting B to the last d_γ columns of the $d_g \times d_g$ identity matrix as in Section 3.3.2 allows for misspecification in the last d_γ moments, while maintaining that the first $d_g - d_\gamma$ moments are valid. More generally, the matrix B may incorporate standardizing the moments by their standard deviation, or it may be used to account for their correlations (see Sections 5 and 6 for examples). The choice of the norm determines how the researcher's bounds on each element of γ interact. With the ℓ_∞ norm, one places separate bounds on each element of γ, which leads to a simple interpretation: no single element of γ can be greater than M. Under an ℓ_p norm with $1 \leq p < \infty$, the bounds on each element of γ interact with each other, so that larger amounts of misspecification in one element is allowed if other elements are correctly specified.

The optimal sensitivity can be computed by casting the optimization problem (10) as a penalized regression problem. To see the connection, note that with $c = B\gamma$, one can write the approximately linear model (14) as

$$Y = -\Gamma\theta + B\gamma + \Sigma^{1/2}\varepsilon,$$

which one can think of as a regression model with correlated errors, design matrix $(-\Gamma, B)$, and coefficient vector $(\theta', \gamma')'$. With this interpretation, it is clear that if the number of regressors $d_\theta + d_\gamma$ is greater than the number of observations d_g, the constraint on the norm of γ is necessary to make the model informative. When $\|\cdot\|$ corresponds to an ℓ_p norm, the constraint on the worst-case bias in (10) becomes $M\|B'k\|_{p'} \leq \overline{B}$, where p' solves $1/p + 1/p' = 1$.

4.2.1 ℓ_2 constraints

When $\|\gamma\|$ corresponds to the Euclidean (or ℓ_2) norm, this leads to ridge regression, and the optimal sensitivity takes the form $k_\lambda' = -H(\Gamma' W_\lambda \Gamma)^{-1}\Gamma' W_\lambda$, where $W_\lambda = (\lambda M^2 BB' + \Sigma)^{-1}$, where, as in the case with a single misspecified moment, λ is the relative weight on bias when (10) is put into a Lagrangian form. The optimal sensitivity for CI construction is

then given by k_{λ^*}, where λ^* minimizes $2\,\mathrm{cv}_\alpha(M\|B'k_\lambda\|_2/\sqrt{k_\lambda'\Sigma k_\kappa})\sqrt{k_\lambda'\Sigma k_\kappa}$. If one is instead interested in estimation using MSE as the optimality criterion (see Remark 2.4), the optimal sensitivity is simply k_1, as $\lambda = 1$ is the optimal choice in this case. This sensitivity can be implemented as a GMM estimator with weighting matrix W_λ. The estimator $\hat{h}$ given in Section 2.2 is simply the one-step Newton-Raphson version of this estimator. Relative to the optimal weighting matrix Σ^{-1} under correct specification, the matrix W_λ trades off precision of the moments against their potential misspecification.

For additional intuition, observe that the weighting matrix would be optimal under correct specification if the asymptotic variance of $\hat{g}(\theta_0)$ were given by $M^2\lambda BB' + \Sigma$ instead of Σ. This form of asymptotic variance arises under a random-effects approach:[8] if one puts a prior on γ with zero mean and variance $\lambda M^2 I$, then unconditionally, the variance of the moment condition will be given by W_λ^{-1}, leading to the same optimal estimator. Thus, if one is interested in optimal estimation under the MSE criterion, the weighting matrix W_1 is optimal under both ℓ_2 constraints on γ, and under a random effects prior on γ with zero mean and variance $M^2 I$. Observe, however, that most of the mass of this random effects prior lies outside of the set $\mathcal{C}$. For example, if the prior distribution is normal, then $P(B\gamma \in \mathcal{C}) = P(\|\gamma\|_2 < M)$ equals the probability that a χ^2 random variable with d_γ degrees of freedom is smaller than 1, which is smaller than 10% for $d_\gamma \geq 4$, and smaller than 1% for $d_\gamma \geq 7$. This is because in higher dimensions, assuming that elements of γ are independent is not innocuous, as it implies that most of the mass of γ concentrates in certain regions of the parameter space, which may help with estimation (for example, by the law of large numbers, the prior will put a lot of mass in the region where the average specification error $\sum_{j=1}^{d_\gamma} \gamma_j/d_\gamma$ is small). Consequently, one needs to assume a rather high prior variance under to yield estimators that one would obtain under our approach. One can also obtain the estimator with weight matrix W_λ as the posterior mean in a Bayesian setting with a local normal prior on γ and a diffuse prior on θ.

This connection is analogous to the connection between Bayes estimators under normal priors and minimax estimators under ℓ_2 constraints in linear models (cf. Li 1982 and Section 2.8.1 of Rossi et al. 2012). However, the resulting CIs are generally different under the random effects approach from the CIs proposed in this paper, because the former approach effectively treats the misspecification bias as a source of additional variability of the moments.

[8] A random-effects approach to dealing with misspecification has been previously considered in Conley et al. (2012), in the context of the linear instrumental variables model, although the paper does not consider the implications for the form of optimal estimator.

4.2.2 ℓ_1 and ℓ_∞ constraints

When $\|\gamma\|$ corresponds to an ℓ_∞ or ℓ_1 norm, the penalized regression analogy leads to a simple algorithm for computing the optimal sensitivity similar to the LASSO/LAR algorithm (Efron et al., 2004): the solution path of the sensitivities that solve (10) as $\overline{B}$ varies is piecewise linear (see Appendix A for details). It follows from this algorithm that under ℓ_∞ constraints, if B corresponds to columns of the identity matrix (as in Section 3.3.2), as M grows, the optimal sensitivity successively drops the "least informative" moments, so that in the limit, if $d_g \leq d_\gamma + d_\theta$, the optimal sensitivity corresponds to that of an exactly identified GMM estimator based on the d_θ "most informative" moments only, where "informativeness" is given by both the variability of a given moment, and its potential misspecification. If $d_g > d_\gamma + d_\theta$, one simply drops all invalid moments in the limit, as discussed Section 3.3.2 for the case $M = \infty$.

4.3 Correct specification and Cressie-Read divergences

If $\mathcal{C} = \{0\}$, the optimal sensitivity is given by $k'_{LS,0} = -H(\Gamma'\Sigma^{-1}\Gamma)^{-1}\Gamma'\Sigma^{-1}$, which corresponds to the sensitivity of $h(\hat{\theta}_{\Sigma^{-1}})$, the GMM estimator with the "usual" optimal weighting matrix Σ^{-1}. In general, the optimal weights under misspecification will take a different form, since they take into account model misspecification allowed by $\mathcal{C}$. However, there is one interesting case where the optimal sensitivity under misspecification is the same as in the correctly specified case. Under ℓ_2 constraints with $B = \Sigma^{1/2}$, so that $\mathcal{C} = \{\Sigma^{1/2}\gamma : \|\gamma\|_2 \leq M\} = \{c\colon c'\Sigma^{-1}c \leq M^2\}$, the optimal weighting matrix derived above for ℓ_2 constraints reduces to Σ^{-1}, just as in the correctly specified case. The intuitive reason for this is that, in this case the uncertainty from potential misspecification is exactly proportional to the asymptotic sampling uncertainty in $\hat{g}(\theta)$.

For an estimator with this sensitivity, the worst-case asymptotic bias is $M\sqrt{k'_{LS,0}\Sigma k_{LS,0}}$. Thus, our CIs can be implemented as

$$h(\hat{\theta}_{\Sigma^{-1}}) \pm \mathrm{cv}_\alpha(M) \cdot \sqrt{\hat{k}'_{LS,0}\hat{\Sigma}\hat{k}_{LS,0}/n}.$$

Furthermore, we show in Appendix A that in this case, the value of (11) is given by $(\delta/2 + M)\sqrt{k'_{LS,0}\Sigma k_{LS,0}}$. Since this is affine in δ, the efficiency in (17) can be calculated explicitly. We give the expression in Appendix A.1, where we also show that the efficiency is at least as high as $\min\{\kappa^L_{*,\alpha}, 1-\alpha\}$, where $\kappa^L_{*,\alpha}$ denotes the efficiency in Equation (18) when $\mathcal{C}$ corresponds to a linear subspace. In particular, since $\kappa^L_{*,0.05} < 0.95$, it follows from Theorem 3.1 and the discussion in Section 3.3.1 that the asymptotic efficiency of the CI in

the preceding display at 95% confidence level is asymptotically at least high as the efficiency of the usual CI under correct specification.

Andrews et al. (2018) have shown that defining misspecification in terms of the magnitude of any divergence in the Cressie and Read (1984) family leads to a set $\mathcal{C}$ that asymptotically takes this form, so long as the set of probability measures under misspecification is contiguous to the probability measure corresponding to correct specification. The Cressie-Read family includes the Hellinger distance used by Kitamura et al. (2013), who consider minimax point estimation among estimators satisfying certain regularity conditions. The results above imply that any estimator with sensitivity $k_{LS,0}$ is near-optimal for CI construction. In line with these results, the estimator in Kitamura et al. (2013) has sensitivity $k_{LS,0}$. Thus, the usual GMM estimator $h(\hat{\theta}_{\Sigma^{-1}})$ and the estimator in Kitamura et al. (2013) are both near-optimal for CI construction, even if one allows for arbitrary CIs that are not necessarily centered at estimators that satisfy the regularity conditions in Kitamura et al. (2013). Also, because they have the same sensitivity, under this form of misspecification, the usual GMM estimator $h(\hat{\theta}_{\Sigma^{-1}})$ and the estimator in Kitamura et al. (2013) have the same local asymptotic minimax properties.

5 Applications

This section describes particular applications of our approach, along with suggestions for the set $\mathcal{C}$ and other implementation details appropriate to each application.

5.1 Generalized method of moments

Most of the applications we consider in this section are special cases of the generalized method of moments (GMM) framework. Here, $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^{n} g(w_i, \theta)$ and $g(\theta) = E\hat{g}(\theta) = Eg(w_i, \theta)$. Equation (2) follows from a central limit theorem, with Σ the variance matrix of $g(w_i, \theta_0)$ (or, in the case of dependent observations, the long run variance matrix). Equation (3) follows from a first order Taylor expansion along with additional arguments, as described in Newey and McFadden (1994). To estimate Σ, one can use the robust variance estimate $\frac{1}{n}\sum_{i=1}^{n} g(w_i, \hat{\theta}_{\text{initial}})g(w_i, \hat{\theta}_{\text{initial}})'$ (or, in the case of dependent observations, an autocorrelation robust version of this estimate). To estimate Γ in the case where $g(w_i, \theta)$ is smooth, one can use the derivative of the sample objective function $\frac{d}{d\theta'}\hat{g}(\theta)\big|_{\theta=\hat{\theta}_{\text{initial}}}$. When $g(w_i, \theta)$ is nonsmooth, one can use a numerical derivative with the step size decreasing at an appropriate rate with n (see Hong et al. 2015, Section 7.3 of Newey and McFadden 1994 and references therein). The derivative matrix H can be estimated with the derivative $\frac{d}{d\theta'}h(\theta)\big|_{\theta=\hat{\theta}_{\text{initial}}}$.

5.2 Instrumental variables

The single equation linear instrumental variables (IV) model is given by

$$y_i = x_i'\theta_0 + \varepsilon_i \tag{20}$$

where, in the correctly specified case, $E\varepsilon_i z_i = E(y_i - x_i'\theta_0)z_i = 0$, with z_i a d_g-vector of instruments. This is an instance of a GMM model with $g(\theta) = E(y_i - x_i'\theta)z_i$ and $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n z_i(y_i - x_i'\theta)$.

One common reason for misspecification in this model is that the instruments do not satisfy the exclusion restriction, because they appear directly in the structural equation (20), so that $\varepsilon_i = z_{Ii}'\gamma/\sqrt{n} + \eta_i$, where $E[z_i\eta_i] = 0$, and z_{Ii} corresponds to a subset I of the instruments, the validity of which one is worried about. This form of misspecification has previously been considered in a number of papers, including Hahn and Hausman (2005), Conley et al. (2012), and Andrews et al. (2017), among others. Bounding the norm of γ using some norm $\|\cdot\|$ then leads to the set

$$\mathcal{C} = \{B\gamma \colon \|\gamma\| \leq M\}, \quad \text{where} \quad B = E[z_i z_{Ii}']. \tag{21}$$

Although the matrix B is unknown, for the purposes of estimating the optimal sensitivity and constructing asymptotically valid CIs, it can be replaced by the sample analog $\hat{B} = n^{-1}\sum_{i=1}^n z_i z_{Ii}'$. This does not affect the asymptotic validity or coverage properties of the resulting CI. Under this setup, the parameter M bounds that magnitude of γ, the direct effect of the instruments on the outcome. Therefore, the appropriate choice of M will depend on the plausible magnitude of these direct effects—see, for example, Conley et al. (2012) for examples and a discussion.

The derivative matrix $\Gamma = -Ez_i x_i'$, can be estimated as $\hat{\Gamma} = -\frac{1}{n}\sum_{i=1}^n z_i x_i'$. The asymptotic variance matrix of the moments is given by $\Sigma = E\varepsilon_i^2 z_i z_i'$, which takes the form $\Sigma_H = (E\varepsilon_i^2)(Ez_i z_i')$ under homoskedasticity (i.e. when $var(\varepsilon_i \mid z_i)$ is constant). Given an initial estimator $\hat{\theta}_{\text{initial}}$, these can be estimated using the usual plug-in formulas, $\hat{\Sigma} = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\theta}_{\text{initial}})^2 z_i z_i'$, and $\hat{\Sigma}_H = \frac{1}{n}\sum_{i=1}^n (y_i - x_i'\hat{\theta}_{\text{initial}})^2 \cdot \frac{1}{n}\sum_{i=1}^n z_i z_i'$. As the initial estimator, one can use the two-stage least squares (2SLS) estimator

$$\hat{\theta}_{\text{initital}} = \left[\left(\sum_{i=1}^n z_i x_i'\right)'\left(\sum_{i=1}^n z_i z_i'\right)^{-1}\left(\sum_{i=1}^n z_i x_i'\right)\right]^{-1}\left(\sum_{i=1}^n z_i x_i'\right)'\left(\sum_{i=1}^n z_i z_i'\right)^{-1}\sum_{i=1}^n z_i y_i.$$

When the norm in (21) corresponds to an ℓ_p norm, the optimal sensitivity can be computed using the algorithms and formulas described in Section 4.

The linearity of the moment condition leads to a particularly simple form of the optimal estimator. In particular, if the parameter of interest is also linear function of θ: $h(\theta) = H\theta$, then the one-step estimator $\hat{h}$ given in Section 2.2 does not depend on the choice of the initial estimator (except possibly in forming the desired sensitivity $\hat{k}$)

$$\begin{aligned}\hat{h} &= H\hat{\theta}_{\text{initial}} + \hat{k}'\frac{1}{n}\sum_{i=1}^{n}(y_i - x_i'\hat{\theta}_{\text{initial}})z_i = \hat{k}'\frac{1}{n}\sum_{i=1}^{n} y_i z_i + \left(H - \hat{k}'\frac{1}{n}\sum_{i=1}^{n} z_i x_i'\right)\hat{\theta}_{\text{initial}} \\ &= \hat{k}'\frac{1}{n}\sum_{i=1}^{n} y_i z_i,\end{aligned}$$

where the second line follows since the weights $\hat{k}$ satisfy $H = -\hat{k}\hat{\Gamma} = \hat{k}\frac{1}{n}\sum_{i=1}^{n} z_i x_i'$.

In the correctly specified case, the 2SLS estimator, which is the GMM estimator with weighting matrix $\hat{\Sigma}_H^{-1}$, is optimal under homoskedasticity. When homoskedasticity does not hold, the GMM estimator with weighting matrix $\hat{\Sigma}^{-1}$ is optimal. Due to concerns with finite sample performance, however, it is common to use the 2SLS estimator along with standard errors based on a robust variance estimate, even when heteroskedasticity is suspected. Mirroring this practice, one can use $\hat{\Sigma}_H$ when forming the optimal sensitivity $\hat{k}$ and worst-case bias (in Step 2 of the algorithm in Section 2.3) while using $\hat{\Sigma}$ to form the variance estimate $\hat{k}'\hat{\Sigma}\hat{k}$ in Step 3. The resulting CI will be valid under both homoskedasticity and heteroskedasticity, and will be optimal under homoskedasticity, just as with the usual CI based on 2SLS with robust standard errors in the correctly specified case.

Remark 5.1. This framework can also be used to incorporate a priori restrictions on the magnitude of coefficients on control variables in an instrumental variables regression. Suppose that we have a set of controls w_i, that appear in the structural equation (20), so that $y_i = x_i'\theta + w_i'\gamma/\sqrt{n} + \epsilon_i$, and ϵ_i is uncorrelated with w_i as well as vector of instruments $\tilde{z}_i$. If one is willing to restrict the magnitude of the coefficient vector γ, so that $\|\gamma\| \leq M$, then one can add w_i to the original vector of instruments $\tilde{z}_i$, $z_i = (\tilde{z}_i', w_i')'$. For example, if one is concerned with functional form misspecification, one can define the control variables to be higher order series terms. We then obtain the misspecified IV model with the set $\mathcal{C}$ given by (21), with $B = E[z_i w_i']$. Thus, we can interpret this model as a locally misspecified version of a model with w_i used as an excluded instrument.

Remark 5.2. Instead of bounding the coefficient vector γ, one can alternatively bound the magnitude of the direct effect $z_{Ii}'\gamma$. If all instruments are potentially invalid, $z_{Ii} = z_i$, and one sets $\mathcal{C} = \{\gamma \colon E[(z_i'\gamma)^2] \leq M\}$, then under homoscedasticity, this corresponds to the case discussed in Section 4.3, where the uncertainty from potential misspecification is exactly

proportional to the asymptotic sampling uncertainty in $\hat{g}(\theta)$. Consequently, in this case the optimal sensitivity is the same as that given by the 2SLS estimator.

5.3 Nonlinear IV

The linear IV model (20) can be generalized to a nonlinear model of the form $\varepsilon_i = \rho_i(\theta_0)$, where $E[\varepsilon_i z_i] = 0$ in the correctly specified case. As in Section 5.2, we can allow for misspecification where the instrument enters the structural equation directly, so that $\varepsilon_i = z_{Ii}'\gamma + \eta_i$ and $E[z_i\eta_i] = 0$, with z_{Ii} denoting a subset of the instruments. As in Section 5.2, bounding the norm of γ leads to a set $\mathcal{C}$ of the form given in (21). The BLP demand model in our empirical application in Section 6 takes the form of a system of nonlinear IV equations, and we consider such forms of misspecification in our empirical application.

5.4 Omitted variables bias in linear regression

Specializing to the case where $z_i = x_i$, the misspecified IV model of Section 5.2 gives a misspecified linear regression model as a special case. This can be used to assess sensitivity of regression results to issues such as omitted variables bias. In particular, consider the linear regression model

$$y_i = x_i'\theta + w_i^* + \tilde{\varepsilon}_i, \quad Ex_i\tilde{\varepsilon}_i = 0$$

where x_i and y_i are observed and w_i^* is a (possibly unobserved) omitted variable. Correlation between w_i^* and x_i will lead to omitted variables bias in the OLS regression of y_i on x_i. If w_i^* is unobserved, then we obtain our framework by making the assumption $\sqrt{n}Ew_i^*x_i \in \mathcal{C}$, for some set $\mathcal{C}$, and letting $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n x_i(y_i - x_i'\theta)$. This setup can also cover choosing between different sets of control variables. Suppose that $w_i^* = w_i'\gamma$, where w_i is a vector of observed control variables that the researcher is considering not including in the regression. If γ is unrestricted, then by the results in Section 3.3.2, the long regression of y_i on both x_i and w_i yields nearly optimal CIs. If one is willing to restrict the magnitude of γ, it is possible to tighten these CIs. In particular, we obtain the setting in Section 5.2 by setting $\hat{g}(\theta) = \frac{1}{n}\sum_{i=1}^n z_i(y_i - x_i'\theta)$, where $z_i = (x_i, w_i')$, and defining $\mathcal{C}$ as in (21), with $z_I = w_i$. The same framework can be used to incorporate selection bias by defining w_i^* to be the inverse Mills ratio term in the formula for $E[y_i \mid x_i, i \text{ observed}]$ in Heckman (1979).

5.5 Functional form misspecification

Our setup allows for misspecification in moment conditions arising from functional form misspecification. To apply our setup, one must relate this misspecification to the bounds $\mathcal{C}$ on

the moment conditions at the true parameter value. One approach to bounding functional form misspecification is to use smoothness conditions from the nonparametric statistics literature, such as bounds on derivatives (see, for example, Tsybakov, 2009, for an introduction to this literature). Since these sets are typically convex (taking a convex combination of two functions that satisfy a given bound on a given derivative gives a function that also satisfies this bound), they typically lead to convex sets $\mathcal{C}$, so that our framework can be applied.

As a simple example, consider a nonparametric IV model with discrete covariates:

$$E[y_i - m(x_i)|z_i] = 0.$$

Suppose x takes values in the finite set $\mathcal{X} = \{\tilde{x}_1, \ldots, \tilde{x}_{N_x}\}$ and z_i takes values in the finite set $\mathcal{Z} = \{\tilde{z}_1, \ldots, \tilde{z}_{N_z}\}$. This setting was considered by Freyberger and Horowitz (2015), who place only nonparametric smoothness or shape restrictions on the unknown function m. To see the connection with our setting, we note that such restrictions can be interpreted as bounds on specification error from a parametric model. If one models these restrictions as local to a parametric family, one obtains our setting. In particular, let $m(x_i) = f(x_i, \theta_0) + n^{-1/2} r(x_i)$, $r \in \mathcal{R}$, where $\mathcal{R}$ is a nonparametric smoothness class. For example, if x_i is univariate, we can let $f(x_i, \theta) = \theta_1 + \theta_2 x_i$ and define $\mathcal{R}$ to be the class of functions with $r(0) = r'(0) = r''(0)$ and second derivative bounded by some constant M. This is equivalent to placing the bound $n^{-1/2}M$ on the second derivative of $m(\cdot)$, which corresponds to a Hölder smoothness class. We can then map this to a misspecified GMM model, with the jth element of the moment function given by $g_j(x_i, y_i, \theta) = (y_i - f(x_i, \theta_0))I(z_i = \tilde{z}_j)$ and jth element of the misspecification vector c given by $Er(x_i)I(z_i = \tilde{z}_j) = \sum_{\tilde{x} \in \mathcal{X}} r(\tilde{x}) P(x_i = \tilde{x}, z_i = \tilde{z}_j)$. Stacking these equations, we see that $c = B\gamma$ where B is a matrix composed of the elements $P(x_i = \tilde{x}, z_i = \tilde{z}_j)$ and $\gamma = (r(\tilde{x}_1), \ldots, r(\tilde{x}_{N_x}))'$. As with the IV setting in Section 5.2, B is unknown, but can be replaced by a consistent estimate based on the sample analogue. So long as the set $\mathcal{R}$ is convex, we obtain convex restrictions on γ and therefore c, so that our framework applies.

This example brings up an important point about the interpretation of $h(\theta)$. If the object of interest is a functional of $m(x) = f(x, \theta_0) + n^{-1/2} r(x)$, then we will need to allow the object of interest $h(\cdot)$ to depend on the misspecification vector directly, as well as on θ. As discussed at the beginning of Section 2, this falls into a mild extension of our framework. Alternatively, under a suitable parametrization of f and r, it is often possible to define the object of interest to be function of θ alone. For example, if we are interested in the derivative $m'(x_0)$ at a particular point x_0 under a bound on the second derivative of $m(\cdot)$, we can let $f(x, \theta) = \theta_1 + \theta_2 x$ and define $\mathcal{R}$ to be the class of functions with $r(x_0) = r'(x_0) = r''(x_0) = 0$

and second derivative bounded by M. Then $m'(x_0) = \theta_2$.

5.6 Treatment effect extrapolation

Often, the average effect of a counterfactual policy on a particular subset of a population is of interest, but this effect is not identified under sufficiently weak assumptions. Rather, policy effects $\tau_1, \ldots, \tau_m$ for each of m other subsets of the population are identified, and consistent, asymptotically joint normal estimates $\hat{\tau} = (\hat{\tau}_1, \ldots, \hat{\tau}_m)$ are available. However, the researcher may have prior information about how these policy effects relate to the policy effect for the subpopulation of interest. If this information amounts to assuming that the policy effect of interest θ satisfies $(\theta - \tau_1, \ldots, \theta - \tau_m) \in \mathcal{C}/\sqrt{n}$ for some convex set $\mathcal{C}$, then this falls into our framework with $\hat{g}(\theta) = (\theta - \hat{\tau}_1, \ldots, \theta - \hat{\tau}_m)'$ and $g(\theta) = (\theta - \tau_1, \ldots, \theta - \tau_m)'$.

An example that has been of recent interest involves nonseparable models with endogeneity. Under conditions in Imbens and Angrist (1994) and Heckman and Vytlacil (2005), instrumental variables estimates with different instruments are consistent for average treatment effects for different subpopulations. A recent literature (Kowalski, 2016; Brinch et al., 2017; Mogstad et al., 2017) has focused on using assumptions on treatment effect heterogeneity to extrapolate these estimates to other populations. If these assumptions amount to placing the differences between the estimated treatment effects and the effect of interest in a known convex set that is local to zero, then our framework applies.

6 Empirical application

This section illustrates the confidence intervals developed in Section 2 in an empirical application to automobile demand based on the data and model in Berry et al. (1995, BLP hereafter). We use the version of the model as implemented by Andrews et al. (2017), who calculate the asymptotic bias of the GMM estimator with weighting matrix Σ^{-1} under local misspecification in this setting.[9]

6.1 Model description

In this model, the utility of consumer i from purchasing a vehicle j, relative to the outside option, is given by a random-coefficient logit model $U_{ij} = \sum_{k=1}^{K} x_{jk}(\beta_k + \sigma_k v_{ik}) - \alpha p_j / y_i + \xi_j + \epsilon_{ij}$, where p_j is the price of the vehicle, x_{jk} the kth observed product characteristic,

[9]The dataset for this empirical application has been downloaded from the Andrews et al. (2017) replication files, available at https://dataverse.harvard.edu/file.xhtml?persistentId=doi:10.7910/DVN/LLARSN/2KFPRA&version=1.1.

ξ_j is an unobserved product characteristic, and ϵ_{ij} is has an iid extreme value distribution. The income of consumer i is assumed to be log-normally distributed, $y_i = e^{m+\varsigma v_{i0}}$, where the mean m and the variance ς of log-income are assumed to be known and set to equal to estimates from the Current Population Survey. The unobservables $v_i = (v_{i0}, \ldots, v_{iK})$ are iid standard normal, while the distribution of the unobserved product characteristic ξ_j is unrestricted.

The marginal cost mc_j for producing vehicle j is given by $\log(mc_j) = w_j'\lambda + \omega_j$, where w_j are observable characteristics, and ω_j is an unobservable characteristic. The full vector of model parameters is given by $\theta = (\sigma', \alpha, \beta', \lambda')'$. Given this vector, and given a vector of unobservable characteristics, one can compute the market shares implied by utility maximization, which can be inverted to yield the unobservable characteristic as a function of θ, $\xi_j(\theta)$. One can similarly invert the unobserved cost component, writing it as a function of θ, $\omega_j(\theta)$, under the assumption that firms set prices to maximize profits in a Bertrand-Nash equilibrium. Given a vector z_{dj} of demand-side instruments, and a vector z_{sj} of supply-side instruments, this yields the sample moment condition

$$\hat{g}(\theta) = \frac{1}{n}\sum_{j=1}^{n}\begin{pmatrix} z_{dj}\xi_j(\theta) \\ z_{sj}\omega_j(\theta) \end{pmatrix},$$

with $g(\theta) = E[\hat{g}(\theta)]$.

The BLP data spans the period 1971 to 1990, and includes information on essentially all $n = 999$ models sold during that period (for simplicity, we have suppressed the time dimension in the description above). There are 5 observable characteristics x_j: a constant, horsepower per 10 pounds of weight (HPWt), a dummy for whether air-conditioning is standard (Air), mileage per 10 dollars (MP\$) defined as MPG over average gas price in a given year, and car size (Size), defined as length times width. The vector z_{dj} consists of x_j, plus the sum of x_j across models other than j produced by the same firm, and for rival firms. There are 6 cost variables w_j: a constant, log of HPWt, Air, log of MPG, log of Size, and a time trend. The vector z_{sj} consists of these variables, MP\$, and the sums of w_j for own-firm products other than j, and for rival firms. After excluding collinear instruments, this gives a total of $d_g = 31$ instruments, 25 of which are excluded to identify $d_\theta = 17$ model parameters. The parameter of interest is average markup, $h(\theta) = \frac{1}{n}\sum_j (p_j - mc_j(\theta))/p_j$.

One may worry that some of these instruments are invalid, because elements of z_{dj} or z_{sj} may appear directly in the utility or cost function with the coefficient on the ℓth element given by $\delta_{d\ell}\gamma_{d\ell}/\sqrt{n}$ or $\delta_{s\ell}\gamma_{s\ell}/\sqrt{n}$, respectively. Here $\delta_{d\ell}$ and $\delta_{s\ell}$ are scaling constants so that, given the sample size at hand, $\gamma_{d\ell}$ has the interpretation that the consumer willingness to pay for

one standard deviation change in the ℓth demand-side instrument $z_{dj\ell}$ is $\gamma_{d\ell}\%$ of the average 1980 car price, and changing the ℓth supply-side instrument $z_{sj\ell}$ by one standard deviation changes the marginal cost by $\gamma_{s\ell}\%$ of the average car price. Andrews et al. (2017) use this scaling in their sensitivity analysis, and they discuss economic motivation for concerns about this form of misspecification. By way of comparison, the estimates of the parameters β and λ in the utility and cost function imply that consumers are on average willing to pay between 2.2 and 10.0% of the average car price for a standard deviation change in one of the included car characteristics, and that a standard deviation change in the included cost characteristics changes the marginal cost by between 3.8 and 11.1% of the average car price. We therefore interpret specifications of the set $\mathcal{C}$ that allow for $|\gamma_{s\ell}| \approx 1$–$2$ (or $|\gamma_{d\ell}| \approx 1$–$2$) as allowing for moderate amounts of misspecification in the ℓth supply-side (or demand-side) instrument.

Given a set I of potentially invalid instruments, the sets $\mathcal{C}$ that we consider have the form

$$\mathcal{C} = \{B_I\gamma \colon \|\gamma\|_p \leq M(\#I)^{1/p}\}, \qquad B = \begin{pmatrix} \delta_d E[z_{dj} z_{dj}'] & 0 \\ 0 & \delta_s E[z_{sj} z_{sj}'] \end{pmatrix},$$

where B_I is given by the columns of B corresponding to the potentially invalid instruments, $\#I$ is the number of potentially invalid instruments, and $p \in \{1, 2, \infty\}$. The scaling by $(\#I)^{1/p}$ ensures that the vector $\gamma = M(1, \ldots, 1)'$ is always included in the set.

Andrews et al. (2017) report the sensitivity of the usual GMM estimator under this scaling for misspecification in each instrument individually. This corresponds to $c = B_I\gamma$ with I containing a single instrument and $\gamma = 1$, reported for each choice of the misspecified instrument. If one is concerned about instruments in a given set I, it is then natural to let $c = B_I\gamma$ with $\gamma = (1, \ldots, 1)$. This allows all instruments in the set I to be misspecified. The specification of $\mathcal{C}$ given above allows for this when $M = 1$, and varying M allows one to assess the sensitivity of conclusions to different amounts of misspecification. Different choices of $p \in \{1, 2, \infty\}$ correspond to different assumptions about how the bounds on misspecification interact, as discussed in Section 4.2. As we will see below, different choices of $\mathcal{C}$ lead to different sensitivities for the optimal estimator, and using the optimal sensitivity can reduce the width of the CI substantially relative to CIs based on the usual GMM estimator.

6.2 Results

We set $\hat{\theta}_{\text{initial}}$ to the GMM estimator that uses $\hat{\Sigma}_0^{-1}$ as the weighting matrix, where $\hat{\Sigma}_0$ is an initial estimate of the variance of the moment conditions. We estimate B, Γ, and H by their sample analogs.

To illustrate that using the sensitivity that is optimal under local misspecification can yield substantially tighter CIs, Figure 1 plots the confidence intervals based on the optimal

sensitivity, as well as those based on $\hat{\theta}_{\text{initial}}$ under different sets I of potentially invalid instruments and ℓ_2 constraints on γ. It is clear from the figure that using the optimal sensitivity yields substantially tighter confidence intervals, relative to simply adjusting the usual CI by using the critical value $\text{cv}_\alpha(\cdot)$ to take into account the potential bias of $h(\hat{\theta}_{\text{initial}})$, by as much as a factor of 3.4. The intuitive reason for this is that by adjusting the sensitivity of the estimator, it is possible to substantially reduce its bias at little cost in terms of an increase in variance. Thus, for example, while the CI for the average markup based on the estimate $\hat{\theta}_{\text{initial}}$ is essentially too wide to be informative when the set of potentially invalid instruments corresponds to all excluded instruments, the CI based on the optimal sensitivity, $[46.0, 66.0]\%$, is still quite tight.

As discussed in Remark 2.3, if a researcher is ex ante unsure what form of misspecification one should worry about, as a sensitivity check, it is useful to consider the effects of different forms of misspecification. In Figure 2, we plot the optimal confidence intervals for different subsets of invalid instruments, as well as for different choices of p in the ℓ_p norm that defines the set $\mathcal{C}$. When only one instrument is allowed to be invalid, the choice of norm doesn't matter. For example, as can be seen from the figure, allowing the supply-side instrument "Miles/dollar" to be invalid yields the same confidence interval under $p = 1, 2$, or ∞. Although the choice of norm matters when the number of potentially misspecified instruments is greater than one, the results are qualitatively similar. Comparing the results for different choices of the set of potentially invalid instruments suggests that allowing supply-side instruments to be invalid generally increases the average markup estimate, while allowing demand-side instruments to be invalid has the opposite effect.

As it may be ex ante unclear what magnitude of misspecification is reasonable to allow for, as discussed in Remark 2.2, it is useful to plot the optimal CI for multiple choices of M. We do this in Figure 3 for $p = 2$, and we allow all excluded instruments to be potentially invalid. One can see that while the CI is unstable for values of M smaller than about 0.4, for larger values of M, the estimate is quite stable and equal to about 50%. Even at $M = 2$, one rejects the hypothesis that the optimal markup is equal to the initial estimate $h(\hat{\theta}_{\text{initial}}) = 32.7\%$. This suggests that ignoring misspecification in the BLP model likely leads to a downward bias in the estimate of the average markup. At the same time, it is possible to obtain reasonably tight CIs for the average markup even under a moderate amount of misspecification.

The J-statistic for testing the hypothesis that all moments are correctly specified equals 404.7. Consequently, the hypothesis is rejected at the usual significance levels. Furthermore, it can be seen from Figure 2 that the CIs for "all excluded" (that allow all excluded instruments to be invalid at $M = 1$), and "all excluded demand" (that assume validity of

supply-side instruments) do not overlap. This implies that either the misspecification in the demand-side instruments must be greater than 1% of the average care price ($M = 1$), or else the supply-side instruments must also be invalid. In Appendix B, we generalize the J-test to test the hypothesis that, given a set I of potentially invalid instruments, the parameter M is at most M_0 against the alternative that it is greater. Table 1 implements this specification test, and gives the lower endpoint $M_{\min}$ of a one-sided 95% CI of the form $[M_{\min}, \infty)$ based on inverting these tests. In line with Figure 2, the test implies that if we assume that the supply-side instruments are valid, the misspecification in the demand-side instruments must be at least $M = 1.19$ for all ℓ_p norms (since, under our scaling, the volume of the set $\mathcal{C}$ decreases with p, the lower-endpoint $M_{\min}$ must be increasing in p). More generally, the results suggest that if one assumes only a subset of the instruments is invalid, the misspecification in the potentially invalid instruments must be quite large. For example, if we assume that all instruments are valid except potentially the demand-side instruments based on rival firms' product characteristics, then the misspecification in these instruments must be greater than $M = 4.08$ for all ℓ_p norms. Alternatively, if we allow all instruments to be invalid, then $M \geq 0.48$ for all ℓ_p norms.

Finally, to illustrate the implication of Theorem 3.1 that one cannot substantively improve upon the CIs that we construct, we calculate the efficiency bound κ_* for these CIs in Table 2. The table shows that the bound is at least as high as the efficiency bound for the usual CI under correct specification (given in (18) and equal to 84.99% at $\alpha = 0.05$). Thus, the asymptotic scope for improvement over the CIs reported in Figure 2 at particular values of θ and $c = 0$ is even smaller than the scope for improvement over the usual CI at particular values of θ under correct specification.

A Details of calculations

This appendix contains the details of calculations of optimal weights in Section 4.

A.1 Cressie-Read divergences

Consider the problem (11) under constraints of the form $\{c\colon c\Sigma^{-1}c \leq M^2\}$. The Lagrangian for this problem can be written as

$$2H\theta + \lambda_1(\delta^2/4 - (c - \Gamma\theta)'\Sigma^{-1}(c - \Gamma\theta)) + \lambda_2(M^2 - c'\Sigma^{-1}c).$$

(we multiply the objective function by 2 so that its optimized value equals $\omega(\delta)$). The first-order conditions are

$$H' + \lambda_1 \Gamma' \Sigma^{-1}(c - \Gamma\theta) = 0,$$
$$-\lambda_1 \Sigma^{-1}(c - \Gamma\theta) - \lambda_2 \Sigma^{-1} c = 0.$$

Solving for c in the second equation gives $c = \frac{\lambda_1}{\lambda_1+\lambda_2}\Gamma\theta$, and plugging this into the first equation gives $\theta = (\lambda_1+\lambda_2)/(\lambda_1\lambda_2)\cdot(\Gamma'\Sigma^{-1}\Gamma)^{-1}H'$. Plugging these expressions into the constraints yields $M^2 = H(\Gamma'\Sigma^{-1}\Gamma)^{-1}H'/\lambda_2^2$ and $\delta^2/4 = H(\Gamma'\Sigma^{-1}\Gamma)^{-1}H'/\lambda_1^2$. Since $H(\Gamma'\Sigma^{-1}\Gamma)^{-1}H' = k_{LS,0}'\Sigma k_{LS,0}$, solving for λ_1 and λ_2, and plugging into the expression for θ yields

$$\theta = \frac{\delta/2 + M}{\sqrt{k_{LS,0}'\Sigma k_{LS,0}}} \cdot (\Gamma'\Sigma^{-1}\Gamma)^{-1}H'.$$

Thus, $\omega(\delta) = 2H\theta = (\delta + 2M)\sqrt{k_{LS,0}'\Sigma k_{LS,0}}$, which is affine, as claimed in the text.

With this form of ω, the bound in (17) becomes

$$\kappa_*(H, \Gamma, \Sigma, \mathcal{C}) = \frac{(1-\alpha)(z_{1-\alpha} + M) + \phi(z_{1-\alpha})}{\mathrm{cv}_\alpha(M)}.$$

This efficiency equals at least $\min\{\kappa_{*,\alpha}^L, 1-\alpha\}$, where $\kappa_{*,\alpha}^L = ((1-\alpha)z_{1-\alpha} + \phi(z_{1-\alpha}))/z_{1-\alpha/2}$ denotes the efficiency given in (18) when $\mathcal{C}$ is a linear subspace. To show this, observe that $\mathrm{cv}_\alpha'(M) \leq 1$ for all $M \geq 0$. Therefore, the derivative of

$$(1-\alpha)(z_{1-\alpha} + M) + \phi(z_{1-\alpha}) - \min\{1-\alpha, \kappa_{*,\alpha}^L\}\,\mathrm{cv}_\alpha(M)$$

with respect to M, given by $1-\alpha-\min\{1-\alpha, \kappa_{*,\alpha}^L\}\,\mathrm{cv}_\alpha'(M)$, is always non-negative. Since the expression in the above display equals $(\kappa_{*,\alpha}^L - \min\{\kappa_{*,\alpha}^L, 1-\alpha\})z_{1-\alpha/2} \geq 0$ at $M = 0$, it follows that it is always non-negative. Rearranging it then yields $\kappa_*(H, \Gamma, \Sigma, \mathcal{C}) \geq \min\{\kappa_{*,\alpha}^L, 1-\alpha\}$ as claimed.

Furthermore, it follows from Equation (49) that the efficiency of one-sided CIs at $c = 0$ is given by $\kappa_*^{\mathrm{OCI},\beta} = 1$.

A.2 ℓ_p Bounds

We now consider the form of the optimal sensitivity under ℓ_p constraints of the form $\mathcal{C} = \{B\gamma\colon \|\gamma\|_p \leq M\}$. The results in Section 4 follow from these results. Observe first that by

Hölder's inequality and the fact that the inequality is sharp, the worst-case bias is given by

$$\overline{\text{bias}}_{\mathcal{C}}(k) = \sup_{\|\gamma\|_p \leq 1} M|k'B\gamma| = M\|B'k\|_{p'}.$$

where p' is the Hölder complement of p. Therefore, the optimization problem (10) is equivalent to

$$\min_k k'\Sigma k \qquad \text{s.t.} \quad H = -k'\Gamma \qquad \text{and} \quad M\|B'k\|_{p'} \leq \overline{B}. \tag{22}$$

A.2.1 $p = 2$

In this case, the Lagrangian form of (22) becomes

$$\min_k k'(\Sigma + \lambda M^2 BB')k \qquad \text{s.t.} \quad H = -k'\Gamma.$$

Observe that the Lagrange multiplier λ gives the relative weight on bias, with $\lambda = 1$ corresponding to optimizing the worst-case MSE. Optimizing this objective is isomorphic to deriving the minimum variance unbiased estimator of $H\theta$ in a regression model with design matrix $-\Gamma$ and variance $\Sigma + \lambda M^2 BB'$, so the Gauss-Markov theorem implies that the optimal weights are $k' = -H(\Gamma' W_\lambda \Gamma)^{-1}\Gamma' W_\lambda$ where $W_\lambda = [\Sigma + \lambda M^2 BB']^{-1}$. The solution for the correctly specified case follows by setting $M = 0$, and the case with a single misspecified moment follows by setting $B = e_1$.

A.2.2 $p = \infty$

Write the Lagrangian form of (22) as

$$\min_k k'\Sigma k/2 + \lambda M\|B'k\|_1 \qquad \text{s.t.} \quad H = -k'\Gamma. \tag{23}$$

It will be convenient to transform the problem so that the ℓ_1 constraint only involves d_γ elements of k. Let

$$T = \begin{pmatrix} B_\perp' \\ (B'B)^{-1}B' \end{pmatrix}, \qquad T^{-1} = \begin{pmatrix} B_\perp & B \end{pmatrix}, \tag{24}$$

where $B_\perp$ is an orthonormal matrix that's orthogonal to B. Then, since $TB = (0, I_{d_\gamma})'$, the above minimization problem is equivalent to the problem

$$\min_\kappa \kappa' S\kappa/2 + \ell \sum_{i \in I} |\kappa_i| \qquad \text{s. t.} \quad H' = -G'\kappa,$$

where $\kappa = T'^{-1}k$, $S = T\Sigma T'$, $\ell = \lambda M$, $G = T\Gamma$, and $I = \{d_g - d_\gamma, \ldots, d_g\}$ indexes the last d_γ elements of κ.

To minimize the above display and give the solution path as ℓ varies, we use arguments similar to those in Theorem 2 of Rosset and Zhu (2007). For $i \in I$, write $\kappa_i = \kappa_{+,i} - \kappa_{-,i}$, where $\kappa_{+,i} = \max\{\kappa_i, 0\}$ and $\kappa_{-,i} = -\min\{\kappa_i, 0\}$. We minimize the objective function in the preceding display over $\{\kappa_{+,i}, \kappa_{-,i}, \kappa_j \colon i \in I, j \notin I\}$ subject to the constraints $\kappa_{+,i} \geq 0$ and $\kappa_{-,i} \geq 0$. Let μ denote a vector of Lagrange multipliers on the restriction $-H' = G'\kappa$. Then the Lagrangian can be written as

$$\kappa' S\kappa/2 + \ell \sum_{i \in I} (\kappa_{+,i} + \kappa_{-,i}) + \mu'(H' + G'\kappa) - \sum_{i \in I} (\ell_{+,i}\kappa_{+,i} + \ell_{-,i}\kappa_{-,i}).$$

The first-order conditions are given by

$$e_i' S\kappa + e_i' G\mu = 0 \qquad i \in I^C, \tag{25}$$

$$e_i' S\kappa + e_i' G\mu + \ell = \ell_{+,i} \qquad i \in I, \tag{26}$$

$$-(e_i' S\kappa + e_i' G\mu) + \ell = \ell_{-,i} \qquad i \in I. \tag{27}$$

The complementary slackness conditions are given by $\ell_{+,i}\kappa_{+,i} = 0$ and $\ell_{-,i}\kappa_{-,i} = 0$ for $i \in I$, and the feasibility constraints are $\ell_{+,i} \geq 0$, $\ell_{-,i} \geq 0$ for $i \in I$ and $-H' = G'\kappa$.

Let $\mathcal{A}^C = \{i \colon i \in I, \kappa_i = 0\}$, and let $\mathcal{A} = \{i \colon i \notin \mathcal{A}^C\}$ denote the set of active constraints. Let s denote a vector of length $|\mathcal{A}|$ with elements $s_i = \operatorname{sign}(\kappa_i)$ if $i \in I$ and $s_i = 0$ otherwise.

The slackness and feasibility conditions imply that if for $i \in I$, $\kappa_i > 0$, then $\ell_{+,i} = 0$, and if $\kappa_i < 0$ or $\ell_{-,i} = 0$. It therefore follows from (26) and (27) that $e_i' S\kappa + e_i' G\mu = -\operatorname{sign}(\kappa_i)\ell = -s_i\ell$. We can combine this condition with (25) and write

$$e_i' S\kappa + e_i' G\mu = -s_i\ell, \qquad i \in \mathcal{A}. \tag{28}$$

On the other hand, if $i \in \mathcal{A}^C$, then since $\ell_{+,i}$ and $\ell_{-,i}$ are non-negative, it follows from (26) and (27) that

$$|e_i' S\kappa + e_i' G\mu| \leq \ell = |e_j' S\kappa + e_j' G\mu|, \qquad i \in \mathcal{A}^C, j \in \mathcal{A}. \tag{29}$$

Let $\kappa_{\mathcal{A}}$ denote the subset of κ corresponding to the active moments, $G_{\mathcal{A}}$ denote the corresponding rows of G, and $S_{\mathcal{A}\mathcal{A}}$ the corresponding submatrix of S. Then we can write the condition (28) together with the feasibility constraint $G'\kappa = -H'$ compactly as

$$\begin{pmatrix} 0 & G_{\mathcal{A}}' \\ G_{\mathcal{A}} & S_{\mathcal{A}\mathcal{A}} \end{pmatrix} \begin{pmatrix} \mu \\ \kappa_{\mathcal{A}} \end{pmatrix} = \begin{pmatrix} -H' \\ -s\ell \end{pmatrix}.$$

Using the block matrix inverse formula, this implies

$$\begin{aligned}\mu &= (G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}})^{-1} \left(H' - G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} s\ell\right) \\ \kappa_{\mathcal{A}} &= -S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}} \mu - S^{-1}_{\mathcal{A}\mathcal{A}} s\ell \\ &= S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}} (G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}})^{-1} \left(G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} s\ell - H'\right) - S^{-1}_{\mathcal{A}\mathcal{A}} s\ell\end{aligned}$$

Consequently, if we're in a region in where the solution path is differentiable with respect to ℓ, we have

$$\frac{\partial \kappa_{\mathcal{A}}}{\partial \ell} = S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}} (G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}})^{-1} G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} s - S^{-1}_{\mathcal{A}\mathcal{A}} s. \tag{30}$$

The differentiability of path is violated if either (a) the constraint (29) is violated for some $i \in \mathcal{A}^C$ if $\kappa(\ell)$ keeps moving in the same direction, and we add i to $\mathcal{A}$ at a point at which (29) holds with equality; or else (b) the sensitivity $\kappa_i(\ell)$ for some $i \in \mathcal{A}$ reaches zero. In this case, drop i from $\mathcal{A}$. In either case, we need to re-calculate the direction (30) using the new definition of $\mathcal{A}$.

Based on the arguments above and the fact that $\kappa(0) = -S^{-1}G(G'S^{-1}G)^{-1}H'$, we can derive the following algorithm, similar to the LAR-LASSO algorithm, to generate the path of optimal sensitivities $\kappa(\ell)$:

1. Initialize $\ell = 0$, $\mathcal{A} = \{1, \ldots, d_g\}$, $\mu = (G'S^{-1}G)^{-1}H'$, $\kappa = -S^{-1}G\mu$. Let s be a vector of length d_g with elements $s_i = \mathbb{I}\{i \in I\}\,\text{sign}(\kappa_i)$, and calculate initial directions as $\mu_\Delta = -(G'S^{-1}G)^{-1}G'S^{-1}s$, $\kappa_\Delta = -S^{-1}(G\mu_\Delta + s)$

2. While $(|\mathcal{A}| > \max\{d_g - d_\gamma, d_\theta\})$:

   (a) Set step size to $d = \min\{d_1, d_2\}$, where

   $$\begin{aligned}d_1 &= \min\{d > 0 \colon \kappa_i + d\kappa_{\Delta,i} = 0, i \in \mathcal{A} \cap \mathcal{I}\} \\ d_2 &= \min\{d > 0 \colon |e'_i(S\kappa + G\mu) + de'_i(S\kappa_\Delta + G\mu_\Delta)| = \ell + d, i \in \mathcal{A}^C\}\end{aligned}$$

   Take step of size d: $\kappa \mapsto \kappa + d\kappa_\Delta$, $\mu \mapsto \mu + d\mu_\Delta$, and $\ell \mapsto \ell + d$.

   (b) If $d = d_1$, drop $\text{argmin}(d_1)$ from $\mathcal{A}$, and if $d = d_2$, then add $\text{argmin}(d_2)$ to $\mathcal{A}$. Let s be a vector of length d_g with elements $s_i = -\mathbb{I}\{i \in I\}\,\text{sign}(e'_i S\kappa + e'_i G\mu)$, and calculate new directions as

   $$\begin{aligned}\mu_\Delta &= -(G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} G_{\mathcal{A}})^{-1} G'_{\mathcal{A}} S^{-1}_{\mathcal{A}\mathcal{A}} s_{\mathcal{A}} \\ (\kappa_\Delta)_{\mathcal{A}} &= -S^{-1}_{\mathcal{A}\mathcal{A}}(G_{\mathcal{A}}\mu_\Delta + s_{\mathcal{A}}) \\ (\kappa_\Delta)_{\mathcal{A}^C} &= 0\end{aligned}$$

The solution path $k(\lambda)$ is then obtained as $k(\lambda) = T'\kappa(\lambda M)$.

Finally, we show that in the limit $M \to \infty$, the optimal sensitivity corresponds to a method of moments estimator based on the most informative set of d_θ moments, with the remaining $d_g - d_\theta$ moments dropped. Observe that as $M \to \infty$ if B corresponds to columns of the identity matrix, minimizing (23) is equivalent to minimizing $\|k_I\|_1$ subject to $H = -k'\Gamma$. This can be written as a linear program $\min k_{I,+} + k_{I,-i}$ st $-H' = \Gamma'(k_+ - k_-)$, $k_+, k_- \geq 0$. The minimization problem is done on a d_θ-dimensional hyperplane, and solution must occur at a boundary point of the feasible set, where only d_θ variables are non-zero. So the optimal k has d_θ non-zero elements.

A.2.3 $p = 1$

The solution path can be obtained by arguments analogous to those in the preceding subsection. We give details in Appendix D in the supplemental materials.

B Specification test

One can test the null hypothesis of correct specification (i.e. the null hypothesis that $c = 0$) using the J statistic

$$J = n \min_\theta \hat{g}(\theta)'\hat{\Sigma}^{-1}\hat{g}(\theta) = n\hat{g}(\hat{\theta})'\hat{\Sigma}^{-1}\hat{g}(\hat{\theta}),$$

where $\hat{\theta} = \text{argmin}_\theta\, \hat{g}(\theta)'\hat{\Sigma}^{-1}\hat{g}(\theta)$. Alternatively, letting $\hat{\Sigma}^{-1/2}$ denote the symmetric square root of $\hat{\Sigma}^{-1}$, one can project $\hat{\Sigma}^{-1/2}\hat{g}(\tilde{\theta})$, where $\tilde{\theta}$ is some consistent estimate, onto the complement of the space spanned by $\hat{\Sigma}^{-1/2}\hat{\Gamma}$,

$$S = n\hat{g}(\tilde{\theta})'\hat{\Sigma}^{-1/2}\hat{R}\hat{\Sigma}^{-1/2}\hat{g}(\tilde{\theta}),$$

where $\hat{R} = I - \hat{\Sigma}^{-1/2}\hat{\Gamma}(\hat{\Gamma}'\hat{\Sigma}^{-1}\hat{\Gamma})^{-1}\hat{\Gamma}'\hat{\Sigma}^{-1/2}$. If the model is correctly specified, so that $c = 0$, S and J are asymptotically equivalent (Newey and McFadden, 1994, p. 2231), and distributed $\chi^2_{d_g - d_\theta}$.

Under local misspecification, the J statistic has a noncentral χ^2 distribution, with noncentrality parameter depending on c (Newey, 1985), and the asymptotic equivalence of J and S still holds. In this section, we use this observation to form a test of the null hypothesis $H_0 : c \in \mathcal{C}$. When $\mathcal{C}$ takes the form $\mathcal{C} = \mathcal{C}(M) = \{B\gamma \colon \|\gamma\| \leq M\}$ for some norm $\|\cdot\|$, inverting these tests gives a lower CI for M. We begin with a lemma deriving the asymptotic distribution of S and J under local misspecification.

Lemma B.1. *Suppose that (1), (2) and (3) hold, and that $\hat{\theta}$ and $\tilde{\theta}$ satisfy, for some K and $K'_{opt} = -(\Gamma'\Sigma^{-1}\Gamma)^{-1}\Gamma'\Sigma^{-1}$,*

$$\sqrt{n}(\hat{\theta} - \theta_0) = K'_{opt}\sqrt{n}\hat{g}(\theta_0), \qquad \text{and} \qquad \sqrt{n}(\tilde{\theta} - \theta_0) = K'\sqrt{n}\hat{g}(\theta_0).$$

Suppose that $\hat{\Sigma}$ and $\hat{\Gamma}$ are consistent estimates of Σ and Γ, and that Σ and Γ are full rank. Then $S = J + o_P(1)$ and S and J converge in distribution to a noncentral chi-square distribution with $d_g - d_\theta$ degrees of freedom and noncentrality parameter $c'\Sigma^{-1/2}R\Sigma^{-1/2}c$ where $R = I - \Sigma^{-1/2}\Gamma(\Gamma'\Sigma^{-1}\Gamma)^{-1}\Gamma\Sigma^{-1/2}$.

Proof. By (1), (2) and (3), $\sqrt{n}\hat{g}(\tilde{\theta}) = (I + \Gamma K')\Sigma^{1/2}(\Sigma^{-1/2}c + Z_n) + o_P(1)$ where $Z_n = \Sigma^{-1/2}[\sqrt{n}\hat{g}(\theta_0) - c] \xrightarrow{d} \mathcal{N}(0, I_{d_g})$, so that

$$\begin{aligned} S &= (\Sigma^{-1/2}c + Z_n)'\Sigma^{1/2}(\Sigma^{-1/2} + \Sigma^{-1/2}\Gamma K')'R(\Sigma^{-1/2} + \Sigma^{-1/2}\Gamma K')\Sigma^{1/2}(\Sigma^{-1/2}c + Z_n) + o_P(1) \\ &= (\Sigma^{-1/2}c + Z_n)'R(\Sigma^{-1/2}c + Z_n) + o_P(1) \xrightarrow{d} (\Sigma^{-1/2}c + Z)'R(\Sigma^{-1/2}c + Z) \end{aligned}$$

where $Z \sim \mathcal{N}(0, I_{d_g})$ and we use the fact that $R(I + \Sigma^{-1/2}\Gamma K'\Sigma^{1/2}) = R$. Similarly,

$$\sqrt{n}\hat{g}(\hat{\theta}) = (I - \Gamma(\Gamma'\Sigma^{-1}\Gamma)\Gamma'\Sigma^{-1})(c + \Sigma^{1/2}Z_n) + o_P(1) = \Sigma^{1/2}R(\Sigma^{-1/2}c + Z_n) + o_P(1),$$

so that $J = (\Sigma^{-1/2}c + Z_n)'R(\Sigma^{-1/2}c + Z_n) + o_P(1) = S + o_P(1)$. To prove the second claim, decompose $R = P_1P_1'$, where $P_1 \in \mathbb{R}^{d_\theta \times (d_g - d_\theta)}$ corresponds to the eigenvectors associated with non-zero eigenvalues of R. Then

$$(\Sigma^{-1/2}c + Z)'R(\Sigma^{-1/2}c + Z) = (P_1'\Sigma^{-1/2}c + P_1'Z)'(P_1'\Sigma^{-1/2}c + P_1'Z).$$

Since $P_1'Z \sim \mathcal{N}(0, I_{d_g - d_\theta})$, it follows that the random variable in the preceding display has a non-central χ^2 distribution with $d_g - d_\theta$ degrees of freedom and non-centrality parameter $c'\Sigma^{-1/2}R\Sigma^{-1/2}c$. □

Lemma B.1 can be interpreted in using the limiting experiment described in Section 3.1. In particular, the asymptotic distribution of the S and J statistics is isomorphic to the statistic $Y'\Sigma^{-1/2}R\Sigma^{-1/2}Y$ in the limiting experiment $Y = -\Gamma\theta + c + \Sigma^{1/2}\varepsilon$.

The quantiles of a non-central chi-square distribution are increasing in the noncentrality parameter (this is shown in Sun et al., 2010). Thus, to test the null hypothesis $H_0\colon c \in \mathcal{C}$, the appropriate critical value for tests based on the J or S statistic is based on a non-central

chi-squared distribution, with non-centrality parameter

$$\bar{\lambda} = \sup_{c \in \mathcal{C}} c' \Sigma^{-1/2} R \Sigma^{-1/2} c.$$

If $\mathcal{C} = \{B\gamma \colon \|\gamma\|_p \leq M\}$, then this becomes

$$\bar{\lambda} = \sup_{\|t\|_p \leq M} t' B' \Sigma^{-1/2} R \Sigma^{-1/2} B t = \sup_{\|t\|_p \leq 1} M^2 \|R \Sigma^{-1/2} B t\|_2^2 = M^2 \|A\|_{p,2}^2.$$

where the second equality uses the fact that R is idempotent, $A = R\Sigma^{-1/2}B$, and $\|A\|_{p,q} = \max_{\|x\|_p \leq 1} \|Ax\|_q$ is the (p, q) operator norm. For $p = 2$, the operator norm has a closed form, which gives $\bar{\lambda} = M \max \operatorname{eig}(B' \Sigma^{-1/2} R \Sigma^{-1/2} B)$.

C Asymptotic coverage and efficiency

This appendix contains the asymptotic coverage and efficiency results discussed in Section 3. In particular, we prove Theorem 3.1. In order to allow for stronger statements, we state upper and lower bounds separately. Theorem 3.1 then follows by combining these results. Theorem 3.1 focuses on two-sided CIs in the case where $\mathcal{C}$ is centrosymmetric, in addition to being convex. In this appendix, we also prove analogous results for one-sided CIs, and we generalize these results to the case where $\mathcal{C}$ is a convex but asymmetric set. When $\mathcal{C}$ is convex but asymmetric, the negative results about the scope for improvement when c is close to zero no longer hold. Therefore, we consider the general problem of optimizing quantiles of excess length over a set $\mathcal{D} \subseteq \mathcal{C}$, which may be a strict subset of $\mathcal{C}$.

The remainder of this appendix is organized as follows. Section C.1 presents notation and definitions, as well as an overview of the results. Section C.2 contains results on least favorable submodels as well as a two-point testing lemma used in later proofs. We then use this to obtain efficiency bounds for one-sided CIs in Section C.3, and for two-sided CIs in Section C.4. Section C.5 shows that our CIs achieve (or, for two-sided CIs, nearly achieve) these bounds. Section C.6 shows how Theorem 3.1 follows from these results, and also gives a one-sided version of this theorem. Primitive conditions for the misspecified linear IV model, as well as a general construction of a least favorable submodel satisfying the assumptions used in this section, are given in the supplemental appendix.

C.1 Setup

While our focus is on parameter spaces that place restrictions on c, we will also allow for local restrictions on θ in some results. This allows us to bound the scope for "directing power" at particular values of θ. Formally, for some parameter θ^*, we consider the local parameter space that restricts $(\sqrt{n}(\theta-\theta^*)', c')'$ to some set $\mathcal{F} \subseteq \mathbb{R}^{d_\theta+d_g}$. The unrestricted case considered throughout most of the main text corresponds to $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$ (in which case θ^* does not affect the definition of the parameter space). We also allow for additional restrictions on θ by placing it in some set Θ_n. Finally, we use $\mathcal{P}$ to denote the set of distributions P over which we require coverage.

With this notation, the set of values of θ that are consistent with the model under P (i.e. the identified set under P) is

$$\Theta_I(P) = \Theta_I(P; \mathcal{F}, \Theta_n) = \left\{\theta \in \Theta_n : \sqrt{n}((\theta-\theta^*)', g_P(\theta)')' \in \mathcal{F}\right\},$$

and the set of pairs (θ, P) over which coverage is required is given by

$$\mathcal{S}_n = \{(\theta, P) \in \Theta_n \times \mathcal{P} : \theta \in \Theta_I(P)\} = \{(\theta, P) \in \Theta_n \times \mathcal{P} : \sqrt{n}((\theta-\theta^*)', g_P(\theta)')' \in \mathcal{F}\},$$

which reduces to the definition in Section 3.1 when $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$. The coverage requirement for a CI $\mathcal{I}_n$ is then given by (19) with this definition of $\mathcal{S}_n$. To compare one-sided CIs $[\hat{c}, \infty)$, we will consider the β quantile of excess length. Rather than restricting ourselves to the minimax criterion, we consider worst-case excess length over a potentially smaller parameter space $\mathcal{G}$, which may place additional restrictions on θ and c. Let

$$q_{\beta,n}(\hat{c}; \mathcal{P}, \mathcal{G}, \Theta_n) = \sup_{P\in\mathcal{P}} \sup_{\theta\in\Theta_I(P;\mathcal{G},\Theta_n)} q_{P,\beta}(h(\theta) - \hat{c})$$

where $q_{P,\beta}$ denotes the β quantile under P. We will also consider bounds on $q_{P,\beta}(h(\theta)-\hat{c})$ at a single P, which corresponds to the optimistic case of optimizing length at a single distribution. For two-sided CIs, we will consider expected length.

Our efficiency bounds can be thought of as applying the bounds in Armstrong and Kolesár (2018) to a local asymptotic setting, which corresponds to the limiting model (14) with $\Gamma = \Gamma_{\theta^*,P_0}$, $\Sigma = \Sigma_{\theta^*,P_0}$ and $H = H_{\theta^*}$. The between class modulus of continuity for this model is

$$\begin{aligned}\omega(\delta; \mathcal{F}, \mathcal{G}, H, \Gamma, \Sigma) = \sup\ & H(s_1 - s_0) \quad \text{s.t.} \quad (s_0', c_0')' \in \mathcal{F}, (s_1', c_1')' \in \mathcal{G}, \\ & [(c_1-c_0) - \Gamma(s_1-s_0)]'\Sigma^{-1}[(c_1-c_0) - \Gamma(s_1-s_0)] \le \delta^2. \quad (31)\end{aligned}$$

We use the notation $\omega(\delta)$ and $\omega(\delta; \mathcal{F}, \mathcal{G})$ when the context is clear. In the case where $\mathcal{G} = \mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$ and $\mathcal{C}$ is centrosymmetric, the solution satisfies $s_1 = -s_0$ and $c_1 = -c_0$, which gives the same optimization problem as (11), with the objective multiplied by two (this matches the definition of $\omega(\cdot)$ used to define κ_* in the main text).

For one-sided CIs, we show that, for any CI satisfying the coverage condition (19) for a rich enough class $\mathcal{P}$, we will have

$$\liminf_{n\to\infty} \sqrt{n} q_{\beta,n}(\hat{c}; \mathcal{P}, \mathcal{G}, \Theta_n) \geq \omega(\delta_\beta; \mathcal{F}, \mathcal{G}, H, \Gamma, \Sigma) \tag{32}$$

where $\delta_\beta = z_{1-\alpha} + z_\beta$, where z_τ denotes the τ quantile of the $\mathcal{N}(0,1)$ distribution. For bounds on excess length at a single P_0 with $E_{P_0} g(w_i, \theta^*) = 0$, we obtain this bound with $\mathcal{G} = \{0\}$:

$$\liminf_{n\to\infty} \sqrt{n} q_{P_0,\beta}(h(\theta^*) - \hat{c}) \geq \omega(\delta_\beta; \mathcal{F}, \{0\}, H, \Gamma, \Sigma). \tag{33}$$

These results can be thought of as a local asymptotic version of Theorem 3.1 in Armstrong and Kolesár (2018) applied to our setting.

For two-sided CIs, we show that, if a CI $\mathcal{I}_n = \{\hat{h} \pm \hat{\chi}\}$ satisfies the coverage condition (19) for a rich enough class $\mathcal{P}$, then, for any P_0 with $E_{P_0} g(w_i, \theta^*) = 0$, expected length satisfies

$$\begin{aligned}
&\liminf_{T\to\infty} \liminf_{n\to\infty} E_{P_0} \min\{\sqrt{n} 2\hat{\chi}, T\} \\
&\geq (1-\alpha) E[\omega(z_{1-\alpha} - Z; \{0\}, \mathcal{F}, H, \Gamma, \Sigma) + \omega(z_{1-\alpha} - Z; \mathcal{F}, \{0\}, H, \Gamma, \Sigma) | Z \leq z_{1-\alpha}],
\end{aligned} \tag{34}$$

where $Z \sim \mathcal{N}(0,1)$. The above bound uses truncated expected length to avoid technical issues with convergence of moments when achieving the bound (note however that this bound immediately implies the same bound on excess length without truncation). Our results constrain the CI to take the form of an interval. We conjecture that the bound applies to arbitrary confidence sets (with length defined as Lebesgue measure) under additional regularity conditions.

Here, "rich enough" means that $\mathcal{P}$ contains a least favorable submodel. Section C.2 begins the derivation of our efficiency results by giving conditions on this submodel. In Section E.1 in the supplemental materials, we construct a submodel satisfying these conditions under mild conditions.

C.2 Least favorable submodel

Let P_0 be a distribution with $E_{P_0} g(w_i, \theta^*) = 0$ (i.e. the model holds for this data-generating process with $\theta = \theta^*$ and $c = 0$), and consider a parametric submodel P_t indexed by $t \in \mathbb{R}^{d_g}$

(i.e. the dimension of t is the same as the dimension of the values of $g(w_i, \theta)$) with P_t equal to P_0 at $t = 0$. We assume that $\{w_i\}_{i=1}^n$ are iid under P_t. Let $\pi_t(w_i)$ denote the density of a single observation with respect to its distribution under P_0, so that $E_{P_t} f(w_i) = E_{P_0} f(w_i)\pi_t(w_i)$ for any function f. We expect that the least favorable submodel for this problem will be the one that makes estimating $E_P g(W_i, \theta^*)$ most difficult. This corresponds to any subfamily with score function $g(w_i, \theta^*)$. We also place additional conditions on this submodel, given in the following assumption.

Assumption C.1. *The data are iid under P_t for all t in a neighborhood of zero, and the density $\pi_t(w_i)$ for a single observation is quadratic mean differentiable at $t = 0$ with score function $g(w_i, \theta^*)$, where $E_{P_0} g(w_i, \theta^*) = 0$. In addition, the function $(t', \theta')' \mapsto E_{p_t} g(w_i, \theta)$ is continuously differentiable at $(0', \theta^{*\prime})'$ with*

$$\left[\frac{d}{d(t', \theta')} E_{p_t} g(w_i, \theta)\right]_{t=0, \theta=\theta^*} = (\Sigma, \Gamma) \tag{35}$$

where Σ and Γ are full rank.

To understand Assumption C.1, note that Problem 12.17 in Lehmann and Romano (2005) gives the Jacobian with respect to t as Σ in the case where $g(w_i, \theta^*)$ is bounded, and the Jacobian with respect to θ is equal to Γ by definition. Assumption C.1 requires the slightly stronger condition that $E_{p_t} g(w_i, \theta)$ is continuously differentiable with respect to $(t', \theta')'$ for t close to 0 and θ close to θ^*. This is needed to apply the Implicit Function Theorem in the derivations that follow. In the supplemental materials, we give a construction of a quadratic mean differentiable family satisfying this condition, without requiring boundedness of $g(w_i, \theta^*)$ (Lemma E.1).

The bounds in Armstrong and Kolesár (2018) are obtained by bounding the power of a two-point test (simple null and simple alternative) where the null and alternative are given by the points that achieve the modulus. To obtain analogous results in our setting, we use a bound on the power of a two-point test in a least favorable submodel.

Consider sequences of local parameter values $(\theta_{0,n}', c_{0,n}')'$ and $(\theta_{1,n}', c_{1,n}')'$ where, for some s_0, c_0 s_1 and c_1,

$$\begin{aligned} \theta_{0,n} &= \theta^* + (s_0 + o(1))/\sqrt{n},\ c_{0,n} = c_0 + o(1), \\ \theta_{1,n} &= \theta^* + (s_1 + o(1))/\sqrt{n},\ c_{1,n} = c_1 + o(1). \end{aligned} \tag{36}$$

Consider a sequence of tests of $(\theta_{0,n}', c_{0,n}')'$ vs $(\theta_{1,n}', c_{1,n}')'$. Formally, for any $(\theta', c')'$, let

$$\mathcal{P}_n(\theta, c) = \left\{P \in \mathcal{P} : E_P g(w_i, \theta) = c/\sqrt{n}\right\} \tag{37}$$

be the set of probability distributions in $\mathcal{P}$ that are consistent with the parameter values $(\theta', c')'$. We derive a bound on the asymptotic minimax power of a level α test of

$$H_{0,n} : P \in \mathcal{P}_n(\theta_{0,n}, c_{0,n}) \quad \text{vs} \quad H_{1,n} : P \in \mathcal{P}_n(\theta_{1,n}, c_{1,n}), \tag{38}$$

as well as a bound on the power of a test of $H_{0,n}$ at P_0. Let Φ be the standard normal cdf and let

$$\overline{\beta}(s_0, c_0, s_1, c_1) = \Phi\left(\sqrt{[c_1 - c_0 - \Gamma(s_1 - s_0)]'\Sigma^{-1}[c_1 - c_0 - \Gamma(s_1 - s_0)]} - z_{1-\alpha}\right).$$

Lemma C.1. *Let $\mathcal{P}$ be a class of distributions that contains a family P_t that satisfies Assumption C.1. Then, for any sequence of tests ϕ_n satisfying $\limsup_n \sup_{P \in \mathcal{P}_n(\theta_{0,n}, c_{0,n})} E_P \phi_n \leq \alpha$, we have*

$$\limsup_n E_{P_0} \phi_n \leq \overline{\beta}(s_0, c_0, 0, 0) \quad \textit{and} \quad \limsup_n \inf_{P \in \mathcal{P}_n(\theta_{1,n}, c_{1,n})} E_P \phi_n \leq \overline{\beta}(s_0, c_0, s_1, c_1).$$

Lemma C.1 says that the asymptotic minimax power of any test of $H_{0,n}$ vs $H_{1,n}$ is bounded by $\overline{\beta}(s_0, c_0, s_1, c_1)$. Furthermore, if we take $s_1 = 0$ and $c_1 = 0$, then this bound is achieved at P_0. Note that, in keeping with the analogy with the linear model (14), $\overline{\beta}(s_0, c_0, s_1, c_1)$ is the power of the optimal (Neyman-Pearson) test of the simple null (s_0', c_0') vs the simple alternative (s_1', c_1') in the model (14).

Proof of Lemma C.1. The proof involves two steps. First, we use the Implicit Function Theorem to find sequences $t_{0,n}$ and $t_{1,n}$ such that $P_{t_{0,n}}$ satisfies $H_{0,n}$ and $P_{t_{1,n}}$ satisfies $H_{1,n}$. Next, we apply a standard result on testing in quadratic mean differentiable families to obtain the limiting power of the optimal test of $P_{t_{0,n}}$ vs $P_{t_{1,n}}$, which gives an upper bound on the limiting minimax power of any test of $H_{0,n}$ vs $H_{1,n}$.

Let $f(t, \theta, a) = E_{P_t} g(w_i, \theta) - a$ so that $(\theta', c')'$ is consistent with P_t iff. $f(t, \theta, c/\sqrt{n}) = 0$. Under Assumption C.1, it follows from the Implicit Function Theorem that there exists a function $r(\theta, a)$ such that, for θ in a neighborhood of θ^* and a in a neighborhood of zero,

$$E_{P_{r(\theta,a)}} g(w_i, \theta) - a = f(r(\theta, a), \theta, a) = 0.$$

Thus, letting $t_{0,n} = r(\theta_{0,n}, c_{0,n}/\sqrt{n})$ and $t_{1,n} = r(\theta_{1,n}, c_{1,n}/\sqrt{n})$, $P_{t_{0,n}}$ satisfies $H_{0,n}$ and $P_{t_{1,n}}$ satisfies $H_{1,n}$. Furthermore,

$$\left[\frac{d}{d(\theta', a')} r(\theta, a)\right]_{(\theta', a') = (\theta^*, 0)} = -\Sigma^{-1}(\Gamma, -I_{d_g})$$

so that

$$r(\theta, a) = \Sigma^{-1}a - \Sigma^{-1}\Gamma(\theta - \theta^*) + o(\|\theta - \theta^*\| + \|a\|).$$

Thus, letting $t_{0,\infty} = \Sigma^{-1}c_0 - \Sigma^{-1}\Gamma s_0$, we have

$$\begin{aligned} t_{0,n} &= r(\theta_{0,n}, c_{0,n}/\sqrt{n}) = \Sigma^{-1}c_{0,n}/\sqrt{n} - \Sigma^{-1}\Gamma(\theta_{0,n} - \theta^*) + o(\|\theta_{0,n} - \theta^*\| + \|c_{0,n}\|/\sqrt{n}) \\ &= \Sigma^{-1}c_0/\sqrt{n} - \Sigma^{-1}\Gamma s_0/\sqrt{n} + o(1/\sqrt{n}) = t_{0,\infty}/\sqrt{n} + o(1/\sqrt{n}). \end{aligned}$$

Similarly, $t_{1,n} = t_{1,\infty}/\sqrt{n} + o(1/\sqrt{n})$ where $t_{1,\infty} = \Sigma^{-1}c_1 - \Sigma^{-1}\Gamma s_1$.

Since the information matrix for this submodel evaluated at $t = 0$ is Σ, it follows from the arguments in Example 12.3.12 in Lehmann and Romano (2005), extended to the case where the null and alternative are both drifting sequences (rather than just the alternative), that the limit of the power of the Neyman-Pearson test of $P_{t_{0,n}}$ vs $P_{t_{1,n}}$ is

$$\Phi\left(\sqrt{[t_{1,\infty} - t_{0,\infty}]'\Sigma[t_{1,\infty} - t_{0,\infty}]} - z_{1-\alpha}\right) = \overline{\beta}(s_0, c_0, s_1, c_1).$$

This gives the required bound on minimax power over $H_{1,n}$. To obtain the bound on power at P_0, note that, for $\theta_{1,n} = \theta^*$ and $c_{1,n} = 0$, $t_{0,n} = 0$, the bound also corresponds to the power of a test that is optimal for $P_{t_{0,n}}$ vs P_0. □

C.3 One-sided CIs

We prove the following efficiency bound for one-sided CIs.

Theorem C.1. *Let $\mathcal{P}$ be a class of distributions that contains a submodel P_t satisfying Assumption C.1. Let $\Theta_n(C) = \{\theta : \|\theta - \theta^*\| \leq C/\sqrt{n}\}$ for some constant C, and let $\mathcal{F}$ be given. Let $[\hat{c}, \infty)$ be a sequence of CIs such that, for all C, the coverage condition (19) holds with $\Theta_n = \Theta_n(C)$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a set such that the limiting modulus ω is well-defined and continuous for all δ. Then the asymptotic lower bounds (32) and (33) hold.*

Proof. Consider a sequence of simple null and alternative values of θ and c that satisfy (36) for some s_0, c_0, s_1, c_1, with $(\sqrt{n}(\theta_{0,n} - \theta^*)', c_{0,n}')' \in \mathcal{F}$ and $(\sqrt{n}(\theta_{1,n} - \theta^*)', c_{1,n}')' \in \mathcal{G}$, for each n. Note that

$$\lim_{n\to\infty} \sqrt{n}[h(\theta_{1,n}) - h(\theta_{0,n})] = H(s_1 - s_0).$$

Consider the testing problem $H_{0,n} : P \in \mathcal{P}_n(\theta_{0,n}, c_{0,n})$ vs $H_{1,n} : P \in \mathcal{P}_n(\theta_{1,n}, c_{1,n})$ defined in (37) and (38). Suppose that

$$q_{\beta,n}(\hat{c}; \mathcal{P}, \mathcal{G}, \Theta_n) < h(\theta_{1,n}) - h(\theta_{0,n}). \tag{39}$$

Let ϕ_n denote the test that rejects when $h(\theta_{0,n}) \notin [\hat{c}, \infty)$. Since, for any $P \in \mathcal{P}_n(\theta_{1,n}, c_{1,n})$, we have $q_{P,\beta}(h(\theta_{1,n}) - \hat{c}) \leq q_{\beta,n}(\hat{c}; \mathcal{P}, \mathcal{G}, \Theta_n)$ by construction, it follows that, for all $P \in \mathcal{P}_n(\theta_{1,n}, c_{1,n})$,

$$E_P \phi_n = P(h(\theta_{1,n}) - \hat{c} < h(\theta_{1,n}) - h(\theta_{0,n})) \geq P(h(\theta_{1,n}) - \hat{c} \leq q_{P,\beta}(h(\theta_{1,n}) - \hat{c})) \geq \beta,$$

where the last step follows from properties of quantiles (Lemma 21.1 in van der Vaart, 1998). The coverage requirement (19) implies that the test ϕ_n that rejects when $h(\theta_{0,n}) \notin [\hat{c}, \infty)$ has asymptotic level α for $H_{0,n}$. Thus, by Lemma C.1, we must have $\beta \leq \overline{\beta}(s_0, c_0, s_1, c_1)$ if (39) holds infinitely often.

It follows that, if $\overline{\beta}(s_0, c_0, s_1, c_1) < \beta$, we must have

$$\liminf_{n \to \infty} \sqrt{n} q_{\beta,n}(\hat{c}; \mathcal{P}, \mathcal{G}, \Theta_n) \geq H(s_1 - s_0)$$

since otherwise, (39) would hold infinitely often. Since the sequences and limiting $(s_0', c_0') \in \mathcal{F}$ and $(s_1', c_1') \in \mathcal{G}$ were arbitrary, the above bound holds for any $(s_0', c_0') \in \mathcal{F}$ and $(s_1', c_1') \in \mathcal{G}$ with $\overline{\beta}(s_0, c_0, s_1, c_1) \leq \beta - \eta$, where $\eta > 0$ is arbitrary. The maximum of the right-hand side over s_0, c_0, s_1, c_1 in this set is equal to $\omega(\delta_{\beta-\eta}; \mathcal{F}, \mathcal{G}, H, \Gamma, \Sigma)$ by definition, so taking $\eta \to 0$ gives the result. □

C.4 Two-sided CIs

We prove the following efficiency bound for two-sided CIs.

Theorem C.2. *Suppose that, for all C, $\{\hat{h} \pm \hat{\chi}\}$ satisfies the local coverage condition (19) with $\Theta_n = \Theta_n(C) = \{\theta : \|\theta - \theta^*\| \leq C/\sqrt{n}\}$, where $\mathcal{P}$ contains a submodel P_t satisfying Assumption C.1. Suppose also that $0_{d_\theta + d_g} \in \mathcal{F}$ and a minimizer $(s_\vartheta', c_\vartheta')'$ of $(c - \Gamma s)' \Sigma^{-1} (c - \Gamma s)$ subject to $Hs = \vartheta$ and $(s', c')' \in \mathcal{F}$ exists for all $\vartheta \in \mathbb{R}$. Then the asymptotic lower bound (34) holds.*

In the case where $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$, which is the focus of the main text, a sufficient condition for the existence of the minimizer $(s_\vartheta', c_\vartheta')'$ is that $\mathcal{C}$ is compact, H is not equal to the zero vector and Γ is full rank.

Proof. For each $\vartheta \in \mathbb{R}$, let $\tilde{\theta}_{\vartheta,n} = \theta^* + s_\vartheta / \sqrt{n}$, and let $\phi_{\vartheta,n} = I(h(\tilde{\theta}_{\vartheta,n}) \notin \{\hat{h} \pm \hat{\chi}\})$ be the test that rejects when $h(\tilde{\theta}_{\vartheta,n})$ is not in the CI. When the constant C defining $\Theta_n = \Theta_n(C)$ is large enough, the asymptotic coverage condition (19) implies that $\phi_{\vartheta,n}$ is an asymptotic

level α test for $H_{0,n} : P \in \mathcal{P}_n(\tilde{\theta}_{\vartheta,n}, c_\vartheta)$ defined in (37) and (38). Thus, by Lemma C.1,

$$\limsup_{n\to\infty} E_{P_0}\phi_{\vartheta,n} \leq \Phi(\delta_\vartheta - z_{1-\alpha}) \quad \text{where} \quad \delta_\vartheta = \sqrt{(c_\vartheta - \Gamma s_\vartheta)'\Sigma^{-1}(c_\vartheta - \Gamma s_\vartheta)}. \tag{40}$$

We apply this bound to a grid of values of ϑ. Let $\mathcal{E}_n(m)$ denote the grid centered at zero with length $2m$ and meshwidth $1/m$:

$$\mathcal{E}_n(m) = \{j/m : j \in \mathbb{Z}, |j| \leq m^2\}.$$

Let

$$\widetilde{\mathcal{E}}_n(m) = \{\sqrt{n}[h(\tilde{\theta}_{\vartheta,n}) - h(\theta^*)] : \vartheta \in \mathcal{E}_n(m)\}.$$

Note that $h(\tilde{\theta}_{\vartheta,n}) = h(\theta^*) + (1 + o(1))Hs_\vartheta/\sqrt{n} = h(\theta^*) + (1 + o(1))\vartheta/\sqrt{n}$. Thus, letting $a_1, \ldots, a_{2m^2+1}$ denote the ordered elements in $\mathcal{E}_n(m)$ and $\tilde{a}_1, \ldots, \tilde{a}_{m^2+1}$ the ordered elements in $\widetilde{\mathcal{E}}_n$, we have $\tilde{a}_j \to a_j$ for each j as $n \to \infty$.

Let $\mathcal{N}(n, m)$ be the number of elements $\tilde{a}_j$ in $\widetilde{\mathcal{E}}_n$ such that $h(\theta^*) + \tilde{a}_j/\sqrt{n} = h(\tilde{\theta}_{a_j,n}) \in \{\hat{h} \pm \hat{\chi}\}$. Then

$$E_{P_0}\mathcal{N}(n, m) = \sum_{j=1}^{2m^2+1} E_{P_0} I(h(\tilde{\theta}_{a_j,n}) \in \{\hat{h} \pm \hat{\chi}\}) = \sum_{j=1}^{2m^2+1} [1 - E_{P_0}\phi_{a_j,n}].$$

It follows from (40) that (assuming the constant C that defines $\Theta_n(C)$ is large enough),

$$\liminf_{n\to\infty} E_{P_0}\mathcal{N}(n, m) \geq \sum_{j=1}^{2m^2+1} [1 - \Phi(\delta_{a_j} - z_{1-\alpha})] = \sum_{j=1}^{2m^2+1} \Phi(z_{1-\alpha} - \delta_{a_j}).$$

Note that $2\hat{\chi} \geq n^{-1/2}[\mathcal{N}(n, m) - 1] \cdot \min_{1\leq j\leq 2m^2}(\tilde{a}_{j+1} - \tilde{a}_j) = n^{-1/2}[\mathcal{N}(n, m) - 1] \cdot m^{-1} \cdot (1 + \varepsilon_n)$ where $\varepsilon_n = \min_{1\leq j\leq 2m^2}(\tilde{a}_{j+1} - \tilde{a}_j)/m^{-1} - 1$ is a nonrandom sequence converging to zero. This, combined with the above display, gives

$$\liminf_{n\to\infty} E_{P_0} \min\{2n^{1/2}\hat{\chi}, T\} \geq \left[m^{-1} \sum_{j=1}^{2m^2+1} \Phi(z_{1-\alpha} - \delta_{a_j}) - m^{-1}\right]$$

for any $T > 2m$. We have

$$m^{-1} \sum_{j=1}^{2m^2+1} \Phi(z_{1-\alpha} - \delta_{a_j}) = m^{-1} \sum_{j=1}^{2m^2+1} \int I(\delta_{a_j} \leq z_{1-\alpha} - z) d\Phi(z). \tag{41}$$

Following the proof of Theorem 3.2 in Armstrong and Kolesár (2018), note that, for $\vartheta \geq 0$,

$t \geq 0$, we have $\delta_\vartheta \leq t$ iff. $\vartheta \leq \omega(t; \{0\}, \mathcal{F})$. Indeed, note that $\omega(\delta_\vartheta; \{0\}, \mathcal{F}) \geq Hs_\vartheta = \vartheta$ by feasibility of 0 and s_ϑ, c_ϑ for this modulus problem. Since the modulus is increasing, this means that, if $\delta_\vartheta \leq t$, we must have $\vartheta \leq \omega(t; \{0\}, \mathcal{F})$. Now suppose $\vartheta \leq \omega(t; \{0\}, \mathcal{F})$. Then $Hs_{\omega(t;\{0\},\mathcal{F})} \geq \vartheta$, so, for some $\lambda \in [0,1]$, $(s'_\lambda, c'_\lambda) = \lambda(s'_{\omega(t;\{0\},\mathcal{F})}, c'_{\omega(t;\{0\},\mathcal{F})})$ satisfies $Hs_\lambda = \vartheta$, which means that $\delta_\vartheta \leq \sqrt{(c_\lambda - \Gamma s_\lambda)'\Sigma^{-1}(c_\lambda - \Gamma s_\lambda)} \leq t$ as claimed.

Thus, the part of the expression in (41) corresponding to terms in the sum with $a_j \geq 0$ is given by

$$m^{-1} \sum_{j=1}^{2m^2+1} \int I(0 \leq a_j \leq \omega(z_{1-\alpha} - z; \{0\}, \mathcal{F}))\, d\Phi(z)$$
$$\geq \int_{z \leq z_{1-\alpha}} \min\{\omega(z_{1-\alpha} - z; \{0\}, \mathcal{F}) - 1/m, m\} d\Phi(z).$$

By the Dominated Convergence Theorem, this converges to $\int_{z \leq z_{1-\alpha}} \omega(z_{1-\alpha} - z; \{0\}, \mathcal{F}) d\Phi(z)$ as $m \to \infty$. Similarly, for $\vartheta < 0$, $t \geq 0$, we have $\delta_\vartheta \leq t$ iff. $-\vartheta \leq \omega(t; \mathcal{F}, \{0\})$, so that an analogous argument shows that, for arbitrary $\varepsilon > 0$, there exists m such that $\int_{z \leq z_{1-\alpha}} \omega(z_{1-\alpha} - z; \mathcal{F}, \{0\}) d\Phi(z) - \varepsilon$ is an asymptotic lower bound for the part of the expression (41) that corresponds to terms in the sum with $a_j < 0$. Thus, for any $\varepsilon > 0$, there exist constants C and T such that, if the coverage condition (19) holds with $\Theta_n = \Theta_n(C)$,

$$\liminf_{n\to\infty} E_{P_0} \min\{n^{1/2} 2\hat{\chi}, T\} \geq \int_{z \leq z_{1-\alpha}} [\omega(z_{1-\alpha} - z; \{0\}, \mathcal{F}) + \omega(z_{1-\alpha} - z; \mathcal{F}, \{0\})] d\Phi(z) - 2\varepsilon.$$

This gives the result. □

C.5 Achieving the bound

This section gives formal results showing that the CIs proposed in the main text are asymptotically valid, and that, if the weights are chosen optimally, they achieve the efficiency bound in Theorem C.1 in the one-sided case, and nearly achieve the bound in Theorem C.2 in the two-sided case (where "nearly" means up to the sharp efficiency bound κ_* in the limiting model, given in (17), in the case where $\mathcal{C}$ is centrosymmetric).

We specialize to the case considered in the main text where we require coverage without local restrictions on θ. In the notation of Sections C.3 and C.4, this corresponds to $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$ for a convex (but possibly asymmetric) set $\mathcal{C}$.

In the main text, we focused on the case where $\mathcal{C}$ is centrosymmetric. To allow for general convex $\mathcal{C}$, we use estimators that are asymptotically affine, rather than linear. We focus on

one-step estimators, which take the form

$$\hat{h} = h(\hat{\theta}_{\text{initial}}) + \hat{k}' g(\hat{\theta}_{\text{initial}}) + \hat{a}/\sqrt{n}.$$

for some weights $\hat{k}$ and $\hat{a}$. To ensure that bias is not arbitrarily large, we continue to require the condition

$$\hat{H} = -\hat{k}'\hat{\Gamma}, \tag{42}$$

where $\hat{\Gamma}$ is an estimator of Γ satisfying conditions to be given below.

To deal with asymmetric $\mathcal{C}$, and to state results involving worst-case quantiles of excess length over different sets, it will be helpful to separately define worst-case upper and lower bias. For a set $\mathcal{C} \in \mathbb{R}^{d_g}$, let

$$\overline{\text{bias}}_{\mathcal{C}}(k, a) = \sup_{c \in \mathcal{C}} k'c + a, \quad \underline{\text{bias}}_{\mathcal{C}}(k, a) = \inf_{c \in \mathcal{C}} k'c + a$$

A one-sided asymptotic $1 - \alpha$ CI is given by $[\hat{c}, \infty)$ where

$$\begin{aligned}
\hat{c} &= \hat{h} - \overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a})/\sqrt{n} - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n} \\
&= h(\hat{\theta}_{\text{initial}}) + \hat{k}' g(\hat{\theta}_{\text{initial}}) + \hat{a}/\sqrt{n} - \overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a})/\sqrt{n} - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n} \\
&= h(\hat{\theta}_{\text{initial}}) + \hat{k}' g(\hat{\theta}_{\text{initial}}) - \overline{\text{bias}}_{\mathcal{C}}(\hat{k}, 0)/\sqrt{n} - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n},
\end{aligned}$$

and $\hat{\Sigma}$ is an estimate of Σ. Thus, the intercept term $\hat{a}$ does not matter for the one-sided CI and can be taken to be zero in this case. For two-sided CIs, however, the choice of $\hat{a}$ matters, and we assume that $\hat{a}$ is chosen so that the estimator is centered:

$$\overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a}) = \sup_{c \in \mathcal{C}} \hat{k}'c + \hat{a} = -\left(\inf_{c \in \mathcal{C}} \hat{k}'c + \hat{a}\right) = -\underline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a}). \tag{43}$$

A two-sided asymptotic $1 - \alpha$ CI is then given by $\hat{h} \pm \hat{\chi}$ where

$$\hat{\chi} = \text{cv}_\alpha\left(\overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a})/\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}\right)\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\sqrt{n}, \quad \text{where } \text{cv}_\alpha(t) \text{ is the } 1-\alpha \text{ quantile of } |\mathcal{N}(t, 1)|.$$

For both forms of CIs, we first state a result for general weights $\hat{k}$, $\hat{a}$, and then specialize to optimal weights. For the one-sided case, we consider CIs that optimize worst-case length over $(\sqrt{n}(\theta - \theta^*)', c')'$ in some set $\mathcal{G}$, subject to coverage over $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$. In principle, this allows for confidence sets that "direct power" not only at particular values of c but also at particular values of θ. However, Lemma E.2 in the supplemental materials shows that the optimal weights for this problem are the same as the optimal weights when $\mathcal{G}$ is replaced by

$\mathbb{R}^{d_\theta} \times \mathcal{D}(\mathcal{G})$, where $\mathcal{D}(\mathcal{G}) = \{c : \text{there exists } s \text{ s.t. } (s', c')' \in \mathcal{G}\}$. Thus, it is without loss of generality to consider weights that optimize worst-case excess length over $c \in \mathcal{D}$ subject to coverage over $c \in \mathcal{C}$ where $\mathcal{D} \subseteq \mathcal{C}$ is a compact convex set.

The optimal weights take the form $\hat{k} = k(\delta_\beta, \hat{H}, \hat{\Gamma}, \hat{\Sigma})$ where

$$k(\delta, H, \Gamma, \Sigma)' = \frac{((c^*_{1,\delta} - c^*_{0,\delta}) - \Gamma(s^*_{1,\delta} - s^*_{0,\delta}))'\Sigma^{-1}}{((c^*_{1,\delta} - c^*_{0,\delta}) - \Gamma(s^*_{1,\delta} - s^*_{0,\delta}))'\Sigma^{-1}\Gamma H'/HH'} \tag{44}$$

and $c_{0,\delta}$, $s_{0,\delta}$, $c_{1,\delta}$, $s_{1,\delta}$ solve the between class modulus problem (31) with $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$ and $\mathcal{G} = \mathbb{R}^{d_\theta} \times \mathcal{D}$. For a two-sided CI of the form given above, the optimal weights take this form with $\mathcal{D} = \mathcal{C}$, δ minimizing $\hat{\chi}$, and with $\hat{a}$ chosen to center the CI so that (43) holds. We note that, in the case where $\mathcal{D} = \mathcal{C}$ and $\mathcal{C}$ is centrosymmetric, $s^*_{1,\delta} = s^*_{0,\delta}$ and $c^*_{1,\delta} = c^*_{0,\delta}$, and (31) reduces to two times the optimization problem (11). The weights $\hat{k}$ then take the form given in (12) in the main text, and, since $\mathcal{C}$ is centrosymmetric, $\hat{a} = 0$, which gives the two-sided CI proposed in the main text.

For our general result showing coverage for possibly suboptimal weights $\hat{k}$, $\hat{a}$, we make the following assumptions. In the following, for a set $\mathcal{A}_n$, random variables $A_{n,\theta,P}$ and $B_{n,\theta,P}$ and a sequence a_n, we say $A_{n,\theta,P} = B_{n,\theta,P} + o_P(a_n)$ uniformly over (θ, P) in $\mathcal{A}_n$ if, for all $\varepsilon > 0$, $\sup_{(\theta,P)\in\mathcal{A}_n} P(a_n^{-1}\|A_{n,\theta,P} - B_{n,\theta,P}\| > \varepsilon) \to 0$. We say $A_{n,\theta,P} = B_{n,\theta,P} + \mathcal{O}_P(a_n)$ uniformly over (θ, P) in a set $\mathcal{A}_n$ if $\lim_{C\to\infty} \limsup_{n\to\infty} \sup_{(\theta,P)\in\mathcal{A}_n} P(a_n^{-1}\|A_{n,\theta,P} - B_{n,\theta,P}\| > C) = 0$. In the following, the set $\mathcal{S}_n$ defined in Section C.1 over which coverage is required is defined with $\mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$.

Assumption C.2. *The set $\mathcal{C}$ is compact or takes the form $\widetilde{\mathcal{C}} \times \mathbb{R}^{d_{g_2}}$ where $d_{g_1} + d_{g_2} = d_g$ and $\widetilde{\mathcal{C}}$ is a compact subset of $\mathbb{R}^{d_{g_1}}$. In addition, $\hat{\theta}_{initial} - \theta = \mathcal{O}_P(1/\sqrt{n})$, $\hat{g}(\hat{\theta}_{initial}) - \hat{g}(\theta) = \Gamma_{\theta,P}(\hat{\theta}_{initial} - \theta) + o_P(1/\sqrt{n})$ and $h(\hat{\theta}_{initial}) - h(\theta) = H_\theta(\hat{\theta}_{initial} - \theta) + o_P(1/\sqrt{n})$ uniformly over $(\theta, P) \in \mathcal{S}_n$.*

Assumption C.3. *$\hat{g}(\theta) - g_P(\theta) = \mathcal{O}(1/\sqrt{n})$ uniformly over $(\theta, P) \in \mathcal{S}_n$. Furthermore, for a collection of matrices $\Sigma_{\theta,P}$ such that $k'_{\theta,P}\Sigma_{\theta,P}k_{\theta,P}$ is bounded away from zero and infinity,*

$$\sup_{t\in\mathbb{R}} \sup_{(\theta,P)\in\mathcal{S}_n} \left| P\left(\frac{\sqrt{n}k'_{\theta,P}(\hat{g}(\theta) - g_P(\theta))}{\sqrt{k'_{\theta,P}\Sigma_{\theta,P}k_{\theta,P}}} \le t \right) - \Phi(t) \right| \to 0.$$

Assumption C.4. *$\hat{k} - k_{\theta,P} = o_P(1)$ uniformly over $(\theta, P) \in \mathcal{S}_n$, and similarly for $\hat{a}$, $\hat{\Gamma}$, $\hat{H}$ and $\hat{\Sigma}$. Furthermore, $k_{\theta,P}$, $a_{\theta,P}$, $\Gamma_{\theta,P}$, H_θ and $\Sigma_{\theta,P}$ are bounded uniformly over $(\theta, P) \in \mathcal{S}_n$. In the case where $\mathcal{C} = \widetilde{\mathcal{C}} \times \mathbb{R}^{d_{g_2}}$, assume that the last d_{g_2} elements of $\hat{k}$ are zero with probability one for all $P \in \mathcal{P}$.*

Theorem C.3. *Suppose that Assumptions C.2, C.3 and C.4 hold and let $\hat{c}$ be defined above with $\hat{k}$, $\hat{\Gamma}$ and $\hat{H}$ satisfying (42). Then*

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{S}_n} P(h(\theta) \in [\hat{c}, \infty)) \geq 1-\alpha,$$

and

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{\theta\in\Theta_I(P;\mathbb{R}^{d_\theta}\times\mathcal{D},\Theta_n)} \Big\{ \sqrt{n} q_{\beta,P}(h(\theta) - \hat{c}) \\ - \Big[\overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, 0) - \underline{\text{bias}}_{\mathcal{D}}(k_{\theta,P}, 0) + (z_{1-\alpha} + z_\beta)\sqrt{k_{\theta,P}'\Sigma_{\theta,P}k_{\theta,P}} \Big] \Big\} \leq 0.$$

Proof. If $\mathcal{C} = \widetilde{\mathcal{C}} \times \mathbb{R}^{d_{g_2}}$ with $\widetilde{\mathcal{C}}$ compact, the theorem can equivalently be stated as holding with $\hat{k}$ redefined to be the vector in $\mathbb{R}^{d_{g_1}}$ that contains the first d_{g_1} elements of the original sensitivity $\hat{k}$, and with other objects redefined similarly. Therefore, it suffices to consider the case where $\mathcal{C}$ is compact.

Note that

$$\begin{aligned}
\sqrt{n}(\hat{h} - h(\theta)) &= H_\theta \sqrt{n}(\hat{\theta}_{\text{initial}} - \theta) + \hat{k}\sqrt{n}\hat{g}(\theta) + \hat{k}\sqrt{n}(\hat{g}(\hat{\theta}_{\text{initial}}) - \hat{g}(\theta)) + \hat{a} + o_P(1) \\
&= H_\theta \sqrt{n}(\hat{\theta}_{\text{initial}} - \theta) + \hat{k}\sqrt{n}(\hat{g}(\theta) - g_P(\theta)) + \hat{k}'c + \hat{k}\sqrt{n}\Gamma_{\theta,P}(\hat{\theta}_{\text{initial}} - \theta) + \hat{a} + o_P(1) \\
&= (H_\theta + k_{\theta,P}'\Gamma_{\theta,P})\sqrt{n}(\hat{\theta}_{\text{initial}} - \theta) + k_{\theta,P}'c + a_{\theta,P} + k_{\theta,P}'\sqrt{n}(\hat{g}(\theta) - g_P(\theta)) + o_P(1),
\end{aligned}$$

where $c = \sqrt{n} g_P(\theta)$ and the $o_P(1)$ terms are uniform over $(\theta, P) \in \mathcal{S}_n$ (the last equality uses the fact that $\mathcal{C}$ is compact). By Assumption C.4 and (42), $H_\theta + k_{\theta,P}'\Gamma_{\theta,P} = 0$ so this implies

$$\sqrt{n}(\hat{h} - h(\theta)) = k_{\theta,P}'c + a_{\theta,P} + k_{\theta,P}'\sqrt{n}(\hat{g}(\theta) - g_P(\theta)) + o_P(1) \tag{45}$$

uniformly over $(\theta, P) \in \mathcal{S}_n$. By compactness of $\mathcal{C}$ and Assumption C.4, we also have

$$\overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a}) = \overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, a_{\theta,P}) + o_P(1), \quad \hat{k}'\hat{\Sigma}\hat{k} = k_{\theta,P}'\Sigma_{\theta,P}k_{\theta,P} + o_P(1)$$

uniformly over $(\theta, P) \in \mathcal{S}_n$. Thus,

$$\begin{aligned}
&\sqrt{n}(\hat{c} - h(\theta)) = \sqrt{n}(\hat{h} - h(\theta)) - \overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a}) - z_{1-\alpha}\sqrt{\hat{k}'\hat{\Sigma}\hat{k}} \\
&= k_{\theta,P}'c + a_{\theta,P} + k_{\theta,P}'\sqrt{n}(\hat{g}(\theta) - g_P(\theta)) - \overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, a_{\theta,P}) - z_{1-\alpha}\sqrt{k_{\theta,P}'\Sigma_{\theta,P}k_{\theta,P}} + o_P(1)
\end{aligned}$$

uniformly over $(\theta, P) \in \mathcal{S}_n$. Since $k_{\theta,P}'c + a_{\theta,P} - \overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, a_{\theta,P}) \leq 0$ by definition, the first part of the theorem (coverage) now follows from Assumption C.3. For the last part of the

theorem, note that, using the above display and the fact that $k'_{\theta,P}c + a_{\theta,P} \geq \underline{\text{bias}}_{\mathcal{D}}(k_{\theta,P}, a_{\theta,P})$ for any (θ, P) with $c = \sqrt{n}E_P g(w_i, \theta) \in \mathcal{D}$, it follows that $\sqrt{n}(h(\theta) - c)$ is less than or equal to

$$\overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, a_{\theta,P}) - \underline{\text{bias}}_{\mathcal{D}}(k_{\theta,P}, a_{\theta,P}) + z_{1-\alpha}\sqrt{k'_{\theta,P}\Sigma_{\theta,P}k_{\theta,P}} + k'_{\theta,P}\sqrt{n}(\hat{g}(\theta) - g_P(\theta)) + o_P(1)$$

uniformly over (θ, P) with $\sqrt{n}E_P g(w_i, \theta) \in \mathcal{D}$. This, along with Assumption C.3, gives the last part of the theorem. □

Theorem C.4. *Suppose that Assumptions C.2, C.3 and C.4 hold and let $\hat{h}$ and $\hat{\chi}$ be defined above with $\hat{k}$, $\hat{a}$, $\hat{\Gamma}$ and $\hat{H}$ satisfying (42) and (43). Then*

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{S}_n} P\left(h(\theta) \in \{\hat{h} \pm \hat{\chi}\}\right) \geq 1 - \alpha.$$

In addition, we have

$$\sqrt{n}\hat{\chi} - \text{cv}_\alpha\left(\frac{\overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, a_{\theta,P})}{\sqrt{k'_{\theta,P}\Sigma_{\theta,P}k_{\theta,P}}}\right)\sqrt{k'_{\theta,P}\Sigma_{\theta,P}k_{\theta,P}} \xrightarrow{p} 0$$

uniformly over $(\theta, P) \in \mathcal{S}_n$.

Proof. As with Theorem C.3, it suffices to consider the case where $\mathcal{C}$ is compact. Let (θ_n, P_n) be a sequence in $\mathcal{S}_n$ and let $c_n = \sqrt{n}g_{P_n}(\theta_n)$. Let $b_n = k'_{\theta_n,P_n}c_n + a_{\theta_n,P_n}$, $\text{sd}_n = \sqrt{k'_{\theta_n,P_n}\Sigma_{\theta_n,P_n}k_{\theta_n,P_n}}$ and $\overline{b}_n = \overline{\text{bias}}_{\mathcal{C}}(k_{\theta_n,P_n}, a_{\theta_n,P_n})$. Note that, by (43), $\overline{\text{bias}}_{\mathcal{C}}(k_{\theta_n,P_n}, a_{\theta_n,P_n}) = -\underline{\text{bias}}_{\mathcal{C}}(k_{\theta_n,P_n}, a_{\theta_n,P_n})$ when Assumption C.4 holds. It therefore follows that $-\overline{b}_n \leq b_n \leq \overline{b}_n$.

Let $Z_n = \sqrt{n}k'_{\theta_n,P_n}(\hat{g}(\theta_n) - g_{P_n}(\theta_n))/\text{sd}_n$. Note that Z_n converges in distribution (under P_n) to a $\mathcal{N}(0,1)$ random variable by Assumption C.3. By (45),

$$\sqrt{n}(\hat{h} - h(\theta_n)) = b_n + \text{sd}_n Z_n + o_{P_n}(1).$$

Using the fact that sd_n is bounded away from zero and $\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\text{sd}_n$ converges in probability to one under P_n, it also follows that

$$\sqrt{n}(\hat{h} - h(\theta_n))/\sqrt{\hat{k}'\hat{\Sigma}\hat{k}} = b_n/\text{sd}_n + Z_n + o_{P_n}(1).$$

Also, by Assumption C.4, we have, for a large enough constant K,

$$\left|\text{cv}_\alpha\left(\frac{\overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a})}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}}\right) - \text{cv}_\alpha\left(\frac{\overline{b}_n}{\text{sd}_n}\right)\right| \leq K\left\{\left[\overline{\text{bias}}_{\mathcal{C}}(\hat{k}, \hat{a}) - \overline{b}_n\right] + \left[\sqrt{\hat{k}'\hat{\Sigma}\hat{k}} - \text{sd}_n\right]\right\} \xrightarrow{p} 0.$$

This, along with the fact that $\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}/\text{sd}_n$ converges in probability to one under P_n, gives the second part of the theorem. Furthermore, it follows from the above display that

$$P_n\left(h(\theta_n) > \hat{h} + \hat{\chi}\right) = P_n\left(\frac{\sqrt{n}\left(\hat{h} - h(\theta_n)\right)}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}} < -\text{cv}_\alpha\left(\overline{\text{bias}}_{\mathcal{C}}(\hat{k},\hat{a})/\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}\right)\right)$$
$$= P_n\left(b_n/\text{sd}_n + Z_n < -\text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right) + o_{P_n}(1)\right) = \Phi(-b_n/\text{sd}_n - \text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right)) + o(1).$$

Similarly,

$$P_n\left(h(\theta_n) < \hat{h} - \hat{\chi}\right) = P_n\left(\frac{\sqrt{n}\left(\hat{h} - h(\theta_n)\right)}{\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}} > \text{cv}_\alpha\left(\overline{\text{bias}}_{\mathcal{C}}(\hat{k},\hat{a})/\sqrt{\hat{k}'\hat{\Sigma}\hat{k}}\right)\right)$$
$$= P_n\left(b_n/\text{sd}_n + Z_n > \text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right) + o_{P_n}(1)\right) = 1 - \Phi(-b_n/\text{sd}_n + \text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right)) + o(1).$$

Thus, the probability of the CI not covering is given, up to $o(1)$, by

$$1 - \Phi(-b_n/\text{sd}_n + \text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right)) + \Phi(-b_n/\text{sd}_n - \text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right)).$$

This is the probability that the absolute value of a $\mathcal{N}(b_n/\text{sd}_n, 1)$ variable is greater than $\text{cv}_\alpha\left(\bar{b}_n/\text{sd}_n\right)$, which is less than $1-\alpha$ since $|b_n| \leq \bar{b}_n$. □

We now specialize to the case where the optimal weights are used. We make a uniform consistency assumption on $\hat{\Gamma}$, $\hat{H}$ and $\hat{\Sigma}$, as well as assumptions on the rank of H, Γ and Σ. The latter are standard regularity conditions for the correctly specified ($\mathcal{C} = \{0\}$) case.

Assumption C.5. *The estimators $\hat{\Gamma}$, $\hat{H}$ and $\hat{\Sigma}$ are full rank with probability one and satisfy $\hat{\Gamma} - \Gamma_{\theta,P} = o_P(1)$, $\hat{H} - H_\theta = o_P(1)$ and $\hat{\Sigma} - \Sigma_{\theta,P} = o_P(1)$ uniformly over $(\theta, P) \in \mathcal{S}_n$.*

Assumption C.6. *There exists a compact set $\mathcal{B}$ that contains the set $\{(H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P}) : \theta \in \Theta_n, P \in \mathcal{P}\}$ for all n, such that (i) in the case where $\mathcal{C}$ is compact, $H \neq 0$ and Γ and Σ are full rank for any $(H, \Gamma, \Sigma) \in \mathcal{B}$ or (ii) in the case where $\mathcal{C} = \widetilde{\mathcal{C}} \times \mathbb{R}^{d_{g_2}}$ with $\widetilde{\mathcal{C}}$ compact, the same holds for the sub-matrices corresponding to the first d_{g_1} moments.*

Using these assumptions, we can verify that Assumption C.4 holds with weights $k_{\theta,P}$ that achieve the efficiency bound in Theorem C.1 and nearly achieve the efficiency bound in Theorem C.2. This gives the following results.

Theorem C.5. *Suppose that Assumptions C.2, C.3, C.5 and C.6 hold and let $\hat{c}$ be defined above with $\hat{k} = k(\delta_\beta, \hat{H}, \hat{\Gamma}, \hat{\Sigma})$. Then*

$$\liminf_{n\to\infty} \inf_{(\theta,P)\in\mathcal{S}_n} P(h(\theta) \in [\hat{c}, \infty)) \geq 1 - \alpha$$

and

$$\limsup_{n\to\infty} \sup_{P\in\mathcal{P}} \sup_{\theta\in\Theta_I(P;\mathbb{R}^{d_\theta}\times\mathcal{D},\Theta_n)} \left[\sqrt{n} q_{\beta,P}(h(\theta) - \hat{c}) - \omega(\delta_\beta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \mathbb{R}^{d_\theta} \times \mathcal{D}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P})\right] \leq 0.$$

Proof. In the case where $\mathcal{C}$ is compact, it follows from Lemma E.6 in the supplemental materials, $k(\delta, H, \Gamma, \Sigma)$ is continuous on $\{\delta\} \times \mathcal{B}$. Since $\mathcal{B}$ is compact, this means that $k(\delta, H, \Gamma, \Sigma)$ is uniformly continuous. Thus, Assumption C.5 implies that $\hat{k}$ satisfies Assumption C.4 with $k_{\theta,P} = k(\delta, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P})$. Furthermore, $\hat{k}$ satisfies (42) by assumption. By properties of the modulus (Equation (24) in Armstrong and Kolesár, 2018),

$$\overline{\text{bias}}_{\mathcal{C}}(k_{\theta,P}, 0) - \underline{\text{bias}}_{\mathcal{D}}(k_{\theta,P}, 0) + (z_{1-\alpha} + z_\beta)\sqrt{k_{\theta,P}' \Sigma_{\theta,P} k_{\theta,P}}$$
$$= \omega(\delta_\beta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \mathbb{R}^{d_\theta} \times \mathcal{D}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P})$$

for this $k_{\theta,P}$. Applying Theorem C.3 gives the result.

In the case where $\mathcal{C} = \widetilde{\mathcal{C}} \times \mathbb{R}^{d_{g2}}$ with $\widetilde{\mathcal{C}}$ compact, the last d_{g_2} elements of $\hat{k}$ are equal to zero as required by Assumption C.4, and the first d_{g_1} elements are the same as the weights computed from the modulus problem with the last d_{g_2} components thrown away and H, Γ and Σ redefined to be the sub-matrices corresponding to the first d_{g_1} elements of the moments. Thus, the same arguments apply in this case. □

For two-sided CIs, we consider weights $\hat{k} = k(\delta^*(\hat{H}, \hat{\Gamma}, \hat{\Sigma}), \hat{H}, \hat{\Gamma}, \hat{\Sigma})$ given by (44) with $\mathcal{G} = \mathcal{F} = \mathbb{R}^{d_\theta} \times \mathcal{C}$, where δ^* may depend on the data through $\hat{H}$, $\hat{\Gamma}$ and $\hat{\Sigma}$. If δ^* is chosen to optimize the length of the fixed length CI, it will be given by $\delta_\chi(\hat{H}, \hat{\Gamma}, \hat{\Sigma})$ where

$$\delta_\chi(H, \Gamma, \Sigma) = \operatorname*{argmin}_\delta \text{cv}_\alpha\left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2}\right)\omega'(\delta) \tag{46}$$

where $\omega(\delta) = \omega(\delta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \mathbb{R}^{d_\theta} \times \mathcal{C}, H, \Gamma, \Sigma)$ is the single class modulus (see Section 3.4 in Armstrong and Kolesár, 2018).

We make a continuity assumption on δ^*.

Assumption C.7. *δ^* is a continuous function of its arguments on the set $\mathcal{B}$ given in Assumption C.6.*

Theorem C.6. *Suppose that Assumptions C.2, C.3, C.5, C.6 and C.7 hold and let* $\hat{h}$ *be defined above with* $\hat{k} = k(\delta^*(\hat{H}, \hat{\Gamma}, \hat{\Sigma}), \hat{H}, \hat{\Gamma}, \hat{\Sigma})$. *Then the conclusion of Theorem C.4 holds. If, in addition,* $\delta^* = \delta_\chi(\hat{H}, \hat{\Gamma}, \hat{\Sigma})$ *for* δ_χ *the CI length optimizing choice of* δ *given in (46), then the half-length* $\hat{\chi}$ *satisfies* $\sqrt{n}\hat{\chi} = \chi(\theta, P) + o_P(1)$ *uniformly over* $(\theta, P) \in \mathcal{S}_n$, *where*

$$\chi(\theta, P) = \min_\delta \mathrm{cv}_\alpha \left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2} \right) \omega'(\delta), \quad \omega(\delta) = \omega(\delta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \mathbb{R}^{d_\theta} \times \mathcal{C}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P}).$$

Proof. The result follows from using the same arguments as in the proof of Theorem C.5, along with continuity of δ^*, to verify Assumption C.4. The form of the limiting half-length for the optimal weights follows from properties of the modulus (see Section 3.4 in Armstrong and Kolesár, 2018). □

C.6 Centrosymmetric case

Theorem 3.1 in Section 3 gives a bound for two-sided CIs in the case where $\mathcal{C}$ is centrosymmetric. This follows from applying Theorems C.6 and C.2 in the centrosymmetric case. In particular, comparing the asymptotic length in Theorem C.6 to the bound in Theorem C.2 and using the fact that $\omega(\delta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \{0\}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P}) = \omega(\delta; \{0\}, \mathbb{R}^{d_\theta} \times \mathcal{C}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P}) = \frac{1}{2}\omega(2\delta; \mathbb{R}^{d_\theta} \times \mathcal{C}, \mathbb{R}^{d_\theta} \times \mathcal{C}, H_\theta, \Gamma_{\theta,P}, \Sigma_{\theta,P})$ when $\mathcal{C}$ is centrosymmetric gives the bound $\kappa_*(H_\theta, \Gamma_{\theta,P_0}, \Sigma_{\theta,P_0}, \mathcal{C})$ from the statement of Theorem 3.1. This corresponds to the bound in Corollary 3.3 of Armstrong and Kolesár (2018). The universal lower bound for κ_* follows from the following result:

Theorem C.7. *For any* H, Γ, Σ *and* $\mathcal{C}$, *the efficiency* κ_* *given in (17) is lower bounded by*

$$(z_{1-\alpha}(1-\alpha) - \tilde{z}_\alpha \Phi(\tilde{z}_\alpha) + \phi(z_{1-\alpha}) - \phi(\tilde{z}_\alpha))/z_{1-\alpha/2}$$

where $\tilde{z}_\alpha = z_{1-\alpha} - z_{1-\alpha/2}$ *and* Φ *and* ϕ *denote the standard normal cdf, and pdf respectively. The lower bound is sharp in the sense that it holds with equality if* $\omega(\delta) = K_0 \min\{\delta, 2z_{1-\alpha/2}\}$, *for some constant* K_0.

Proof. Since $\mathrm{cv}_\alpha(b) \leq b + z_{1-\alpha/2}$, the denominator in (17) is upper-bounded by

$$\min_\delta 2\,\mathrm{cv}_\alpha \left(\frac{\omega(\delta)}{2\omega'(\delta)} - \frac{\delta}{2} \right) \omega'(\delta) \leq 2\,\mathrm{cv}_\alpha \left(\frac{\omega(2z_{1-\alpha/2})}{2\omega'(2z_{1-\alpha/2})} - z_{1-\alpha/2} \right) \omega'(2z_{1-\alpha/2}) \leq \omega(2z_{1-\alpha/2}). \quad (47)$$

On the other hand, the numerator in (17) can be decomposed as

$$(1-\alpha)E\left[\omega(2(z_{1-\alpha}-Z))\mid Z\le z_{1-\alpha}\right]=E\left[\omega(2(z_{1-\alpha}-Z))\,\mathbb{I}\{Z\le z_{1-\alpha}-z_{1-\alpha/2}\}\right]\\+E\left[\omega(2(z_{1-\alpha}-Z))\,\mathbb{I}\{z_{1-\alpha}-z_{1-\alpha/2}\le Z\le z_{1-\alpha}\}\right].$$

Since the modulus $\omega(\delta)$ is non-decreasing, the first summand is lower-bounded by

$$E\left[\omega(2z_{1-\alpha/2})\,\mathbb{I}\{Z\le z_{1-\alpha}-z_{1-\alpha/2}\}\right]=\omega(2z_{1-\alpha/2})\Phi(z_{1-\alpha}-z_{1-\alpha/2}).$$

Since the modulus $\omega(\delta)$ is concave, $\omega(2(z_{1-\alpha}-Z))\ge(z_{1-\alpha}-Z)/z_{1-\alpha/2}\cdot\omega(2z_{1-\alpha/2})$, so that the second summand is lower-bounded by

$$\frac{\omega(2z_{1-\alpha/2})}{z_{1-\alpha/2}}E\left[(z_{1-\alpha}-Z)\,\mathbb{I}\{z_{1-\alpha}-z_{1-\alpha/2}\le Z\le z_{1-\alpha}\}\right]\\=\frac{\omega(2z_{1-\alpha/2})}{z_{1-\alpha/2}}\left(z_{1-\alpha}(1-\alpha-\Phi(z_{1-\alpha}-z_{1-\alpha/2}))+\phi(z_{1-\alpha})-\phi(z_{1-\alpha}-z_{1-\alpha/2})\right),$$

where the equality follows by the formula for the expectation of a truncated normal random variable. Combining the two preceding displays then yields

$$(1-\alpha)E\left[\omega(2(z_{1-\alpha}-Z))\mid Z\le z_{1-\alpha}\right]\\\ge\omega(2z_{1-\alpha/2})\frac{z_{1-\alpha}(1-\alpha)-\tilde{z}_\alpha\Phi(\tilde{z}_\alpha)+\phi(z_{1-\alpha})-\phi(\tilde{z}_\alpha)}{z_{1-\alpha/2}},\quad(48)$$

where $\tilde{z}_\alpha=z_{1-\alpha}-z_{1-\alpha/2}$. Combining this with the bound in (47) then yields the result. The sharpness of the bound for the case $\omega(\delta)=K_0\min\{\delta,2z_{1-\alpha/2}\}$ follows from by noting that in this case, both (47) and (48) hold as equalities. □

For the one-sided case, we obtain the following bound.

Theorem C.8. *Consider the setting of Theorem C.5, with $\mathcal{C}$ centrosymmetric. Then the weights $\hat{k}=\hat{k}(\delta_\beta,\hat{H},\hat{\Gamma},\hat{\Sigma})$ with $\mathcal{D}=\mathcal{C}$ are identical to the weights $\hat{k}(\delta_{\tilde{\beta}},\hat{H},\hat{\Gamma},\hat{\Sigma})$ computed with $\mathcal{D}=\{0\}$, but with $\tilde{\beta}=\Phi((z_\beta-z_{1-\alpha})/2)$. Furthermore, letting $\hat{c}_{\text{minimax}}$ denote the lower endpoint of the CI computed with these weights ($\hat{k}(\delta_\beta,\hat{H},\hat{\Gamma},\hat{\Sigma})$ with $\mathcal{D}=\mathcal{C}$), we have*

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P;\mathbb{R}^{d_\theta}\times\{0\},\Theta_n)}\left\{\sqrt{n}q_{\beta,P}(h(\theta)-\hat{c}_{\text{minimax}})-\frac{1}{2}\left[\omega_{\theta,P}(\delta_\beta)+\delta_\beta\omega'_{\theta,P}(\delta_\beta)\right]\right\}\le 0$$

where $\omega_{\theta,P}(\delta)=\omega(\delta;\mathbb{R}^{d_\theta}\times\mathcal{C},\mathbb{R}^{d_\theta}\times\mathcal{C},H_\theta,\Gamma_{\theta,P},\Sigma_{\theta,P})$. For $\hat{c}$ computed instead with $\mathcal{D}=\{0\}$,

we obtain

$$\limsup_{n\to\infty}\sup_{P\in\mathcal{P}}\sup_{\theta\in\Theta_I(P;\mathbb{R}^{d_\theta}\times\{0\},\Theta_n)}\left\{\sqrt{n}q_{\beta,P}(h(\theta)-\hat{c})-\frac{1}{2}\omega_{\theta,P}(2\delta_\beta)\right\}\leq 0.$$

Proof. The first statement follows from Corollary 3.2 in Armstrong and Kolesár (2018). The second statement follows from applying Theorem C.3 as in the proof of Theorem C.5, noting that $\underline{\text{bias}}_{\{0\}}(k_{\theta,P},0)=0$, and using arguments from the proof of Corollary 3.2 in Armstrong and Kolesár (2018). The last statement follows from Theorem C.5 and the fact that $\omega(\delta;\mathbb{R}^{d_\theta}\times\mathcal{C},\mathbb{R}^{d_\theta}\times\{0\},H_\theta,\Gamma_{\theta,P},\Sigma_{\theta,P})=\frac{1}{2}\omega(2\delta;\mathbb{R}^{d_\theta}\times\mathcal{C},\mathbb{R}^{d_\theta}\times\mathcal{C},H_\theta,\Gamma_{\theta,P},\Sigma_{\theta,P})$. □

Thus, directing power toward the correctly specified case yields the same one-sided CI once one changes the quantile over which one optimizes excess length. If one does attempt to direct power, the scope for doing so is bounded by a factor of

$$\kappa_*^{\text{OCI},\beta}(H_\theta,\Gamma_{\theta,P_0},\Sigma_{\theta,P_0},\mathcal{C})=\frac{\omega_{\theta,P}(2\delta_\beta)}{\omega_{\theta,P}(\delta_\beta)+\delta_\beta\omega'_{\theta,P}(\delta_\beta)}. \tag{49}$$

This gives a bound for the one-sided case analogous to the bound κ_* in (17) for two-sided CIs.

A consistent estimate of these bounds can be obtained by plugging in $\omega(\delta;\mathbb{R}^{d_\theta}\times\mathcal{C},\mathbb{R}^{d_\theta}\times\mathcal{C},\hat{H},\hat{\Gamma},\hat{\Sigma})$ for $\omega_{\theta,P}(\delta)=\omega(\delta;\mathbb{R}^{d_\theta}\times\mathcal{C},\mathbb{R}^{d_\theta}\times\mathcal{C},H_\theta,\Gamma_{\theta,P},\Sigma_{\theta,P})$. Table 2 reports estimates of this bound under different forms of misspecification in the empirical application in Section 6.

References

Altonji, J. G., Elder, T. E., and Taber, C. R. (2005). Selection on observed and unobserved variables: Assessing the effectiveness of catholic schools. *Journal of Political Economy*, 113(1):151–184.

Andrews, D. W. K. and Guggenberger, P. (2009). Hybrid and size-corrected subsampling methods. *Econometrica*, 77(3):721–762.

Andrews, I., Gentzkow, M., and Shapiro, J. M. (2017). Measuring the sensitivity of parameter estimates to sample statistics. *Quarterly Journal of Economics*, 132(4):1553–1592.

Andrews, I., Gentzkow, M., and Shapiro, J. M. (2018). On the informativeness of descriptive statistics for structural estimates. Working paper, Harvard University.

Armstrong, T. B. and Kolesár, M. (2018). Optimal inference in a class of regression models. *Econometrica*, 86(2):655–683.

Berkowitz, D., Caner, M., and Fang, Y. (2012). The validity of instruments revisited. *Journal of Econometrics*, 166(2):255–266.

Berry, S. T., Levinsohn, J., and Pakes, A. (1995). Automobile prices in market equilibrium. *Econometrica*, 63(4):841–890.

Bonhomme, S. and Weidner, M. (2018). Minimizing Sensitivity to Model Misspecification. arXiv: 1807.02161.

Brinch, C. N., Mogstad, M., and Wiswall, M. (2017). Beyond LATE with a Discrete Instrument. *Journal of Political Economy*, 125(4):985–1039.

Bugni, F. A. and Ura, T. (2018). Inference in dynamic discrete choice problems under local misspecification. arXiv: 1603.07987.

Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *The Annals of Statistics*, 32(5):1805–1840.

Chamberlain, G. (1987). Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334.

Chen, X., Tamer, E. T., and Torgovitsky, A. (2011). Sensitivity analysis in semiparametric likelihood models. SSRN working paper ID 1963746.

Cheng, X. and Liao, Z. (2015). Select the valid and relevant moments: An information-based LASSO for GMM with many moments. *Journal of Econometrics*, 186(2):443–464.

Conley, T. G., Hansen, C. B., and Rossi, P. E. (2012). Plausibly exogenous. *The Review of Economics and Statistics*, 94(1):260–272.

Cressie, N. and Read, T. R. C. (1984). Multinomial goodness-of-fit tests. *Journal of the Royal Statistical Society. Series B (Methodological)*, 46(3):440–464.

DiTraglia, F. J. (2016). Using invalid instruments on purpose: Focused moment selection and averaging for GMM. *Journal of Econometrics*, 195(2):187–208.

Donoho, D. L. (1994). Statistical estimation and optimal recovery. *The Annals of Statistics*, 22(1):238–270.

Duflo, E., Greenstone, M., Pande, R., and Ryan, N. (2018). The value of regulatory discretion: Estimates from environmental inspections in india. *Econometrica*, 86(6):2123–2160.

Efron, B., Hastie, T., Johnstone, I. M., and Tibshirani, R. J. (2004). Least angle regression. *The Annals of Statistics*, 32(2):407–451.

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

Gayle, G.-L. and Shephard, A. (2019). Optimal taxation, marriage, home production, and family labor supply. *Econometrica*, 87(1):291–326.

Guggenberger, P. (2012). On the Asymptotic Size Distortion of Tests When Instruments Locally Violate the Exogeneity Assumption. *Econometric Theory*, 28(2):387–421.

Hahn, J. and Hausman, J. A. (2005). IV estimation with valid and invalid instruments. *Annales d'Économie et de Statistique*, (79/80):25–57.

Hansen, L. P. (1985). A method for calculating bounds on the asymptotic covariance matrices of generalized method of moments estimators. *Journal of Econometrics*, 30(1-2):203–238.

Hansen, L. P. and Sargent, T. J. (2008). *Robustness*. Princeton University Press, Princeton, NJ.

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica*, 47(1):153–161.

Heckman, J. J. and Vytlacil, E. (2005). Structural Equations, Treatment Effects, and Econometric Policy Evaluation1. *Econometrica*, 73(3):669–738.

Hong, H., Mahajan, A., and Nekipelov, D. (2015). Extremum estimation and numerical derivatives. *Journal of Econometrics*, 188(1):250–263.

Huber, P. J. and Ronchetti, E. M. (2009). *Robust Statistics*. John Wiley & Sons, Hoboken, NJ, second edition.

Imbens, G. W. and Angrist, J. D. (1994). Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475.

Imbens, G. W. and Manski, C. F. (2004). Confidence intervals for partially identified parameters. *Econometrica*, 72(6):1845–1857.

Joshi, V. M. (1969). Admissibility of the usual confidence sets for the mean of a univariate or bivariate normal population. *The Annals of Mathematical Statistics*, 40(3):1042–1067.

Kitamura, Y., Otsu, T., and Evdokimov, K. (2013). Robustness, infinitesimal neighborhoods, and moment restrictions. *Econometrica*, 81(3):1185–1201.

Kowalski, A. E. (2016). Doing more when you're running LATE: Applying marginal treatment effect methods to examine treatment effect heterogeneity in experiments. Working Paper 22363, National Bureau of Economic Research.

Leamer, E. E. (1983). Let's take the con out of econometrics. *The American Economic Review*, 73(1):31–43.

Leeb, H. and Pötscher, B. M. (2006). Can one estimate the conditional distribution of post-model-selection estimators? *The Annals of Statistics*, 34(5):2554–2591.

Lehmann, E. L. and Romano, J. P. (2005). *Testing statistical hypotheses*. Springer, New York, third edition.

Li, K.-C. (1982). Minimaxity of the Method of Regularization of Stochastic Processes. *The Annals of Statistics*, 10(3):937–942.

Liao, Z. (2013). Adaptive GMM shrinkage estimation with consistent moment selection. *Econometric Theory*, 29(05):857–904.

Low, M. G. (1995). Bias-variance tradeoffs in functional estimation problems. *The Annals of Statistics*, 23(3):824–835.

McCloskey, A. (2017). Asymptotically uniform tests after consistent model selection in the linear regression model. Working paper, Brown University.

Mogstad, M., Santos, A., and Torgovitsky, A. (2017). Using instrumental variables for inference about policy relevant treatment effects. Working Paper 23568, National Bureau of Economic Research. DOI: 10.3386/w23568.

Moon, H. R. and Schorfheide, F. (2009). Estimation with overidentifying inequality moment conditions. *Journal of Econometrics*, 153(2):136–154.

Mukhin, Y. (2018). Sensitivity of regular estimators. arXiv: 1805.08883.

Nevo, A. and Rosen, A. M. (2010). Identification with imperfect instruments. *Review of Economics and Statistics*, 94(3):659–671.

Newey, W. K. (1985). Generalized method of moments specification testing. *Journal of Econometrics*, 29(3):229–256.

Newey, W. K. and McFadden, D. L. (1994). Large sample estimation and hypothesis testing. In McFadden, D. L. and Engle, R. F., editors, *Handbook of Econometrics*, volume 4, chapter 36, pages 2111–2245. Elsevier.

Pratt, J. W. (1961). Length of confidence intervals. *Journal of the American Statistical Association*, 56(295):549–567.

Rosset, S. and Zhu, J. (2007). Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3):1012–1030.

Rossi, P. E., Allenby, G. M., and McCulloch, R. (2012). *Bayesian Statistics and Marketing*. John Wiley & Sons.

Sacks, J. and Ylvisaker, D. (1978). Linear estimation for approximately linear models. *The Annals of Statistics*, 6(5):1122–1137.

Small, D. S. (2007). Sensitivity analysis for instrumental variables regression with overidentifying restrictions. *Journal of the American Statistical Association*, 102(479):1049–1058.

Sun, Y., Baricz, Á., and Zhou, S. (2010). On the monotonicity, log-concavity, and tight bounds of the generalized marcum and nuttall Q-functions. *IEEE Transactions on Information Theory*, 56(3):1166–1186.

Tsybakov, A. B. (2009). *Introduction to nonparametric estimation*. Springer, New York, NY.

van der Vaart, A. W. (1998). *Asymptotic Statistics*. Cambridge University Press, New York, NY.

Table 1: J-test of overidentifying restrictions in the application to Berry et al. (1995) under different forms of ℓ_p misspecification.

Instrument set	$p = 1$	$p = 2$	$p = \infty$
D/F: # cars	9.77	9.77	9.77
S/F: # cars	15.32	15.32	15.32
Supply: Miles/dollar	17.00	17.00	17.00
All D/F	2.38	2.59	2.59
All D/R	4.08	5.22	5.40
All S/F	2.04	2.61	2.62
All S/R	2.47	4.16	6.99
All excluded demand	1.19	1.72	1.88
All excluded supply	1.02	1.64	1.78
All excluded	0.48	1.08	2.54

Notes: The table gives the minimum value of M such that the test of overidentifying restrictions has p-value equal to 0.05.

Table 2: Efficiency bounds (in %) for one and two-sided 95% confidence intervals at $c = 0$ under ℓ_p misspecification in the application to Berry et al. (1995).

	Two-sided			One-sided		
Instrument set	$p = 1$	$p = 2$	$p = \infty$	$p = 1$	$p = 2$	$p = \infty$
D/F: # cars	85.9	85.9	85.9	100.0	100.0	100.0
S/F: # cars	90.1	90.1	90.1	99.8	99.8	99.8
Supply: Miles/dollar	85.0	85.0	85.0	100.0	100.0	100.0
All D/F	85.4	85.5	85.7	100.0	100.0	100.0
All D/R	94.3	94.8	95.3	95.3	93.9	95.3
All S/F	88.0	88.6	89.1	99.9	99.7	99.7
All S/R	89.5	89.4	89.2	99.1	98.5	99.5
All excluded demand	95.0	95.4	96.4	97.7	95.0	97.3
All excluded supply	89.8	90.3	90.1	98.8	98.2	99.6
All excluded	96.3	97.0	97.5	99.0	99.5	98.2

Notes: For two-sided confidence intervals, the table calculates the ratio of the expected length of a 95% confidence interval that minimizes its length at $c = 0$ relative to the length of the FLCI in (15), given in (17). For one-sided confidence intervals, the table calculates an analogous bound, given in Appendix C.6, when the confidence interval optimizes the 80% quantile of excess length.

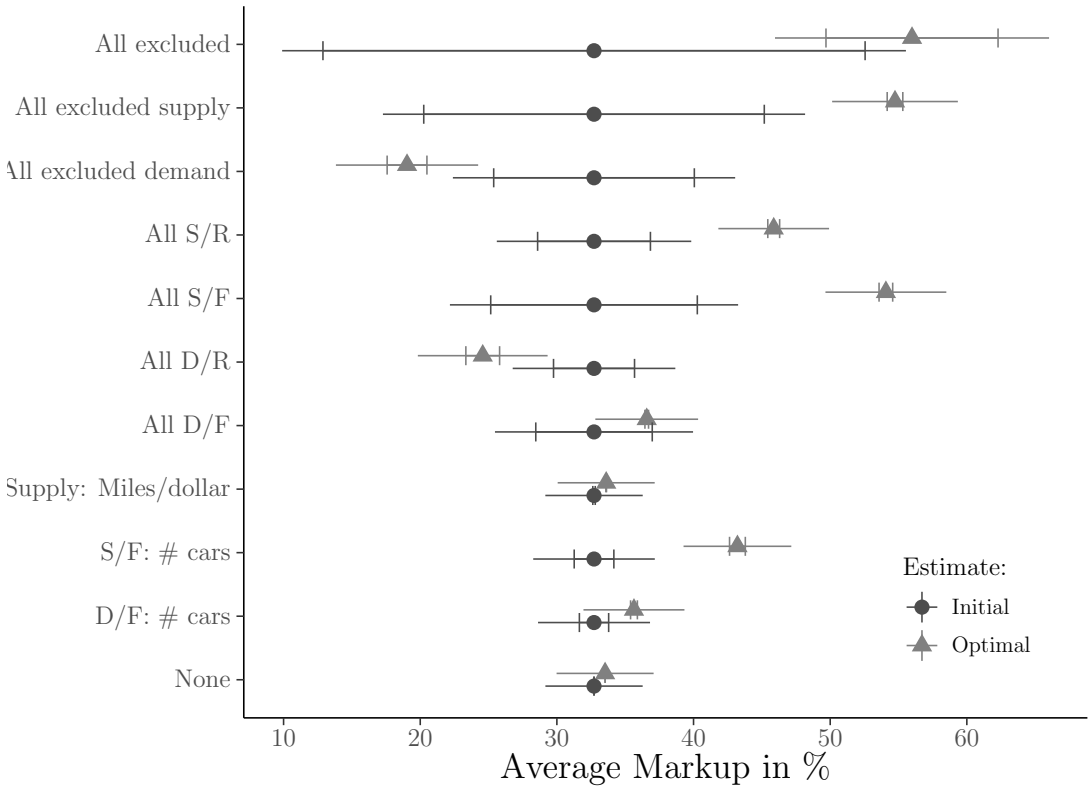


Figure 1: Confidence intervals under ℓ_2 misspecification and $M = 1$ in the application to Berry et al. (1995).

Vertical lines correspond to the estimate ± the worst case bias, and horizontal lines correspond to 95% confidence intervals. Different rows correspond to assuming that different subsets of instruments are potentially invalid. "None": correct specification. "D/F" "S/F": Demand-side instrument based on characteristics of other cars produced by the same firm. "S/F": Supply-side instrument based on characteristics of other cars produced by the same firm. "D/R": Demand-side instrument based on characteristics of cars produced by rivals. "S/R": Supply-side instrument based on characteristics of cars produced by rivals. "All excluded": All excluded instruments are potentially invalid.

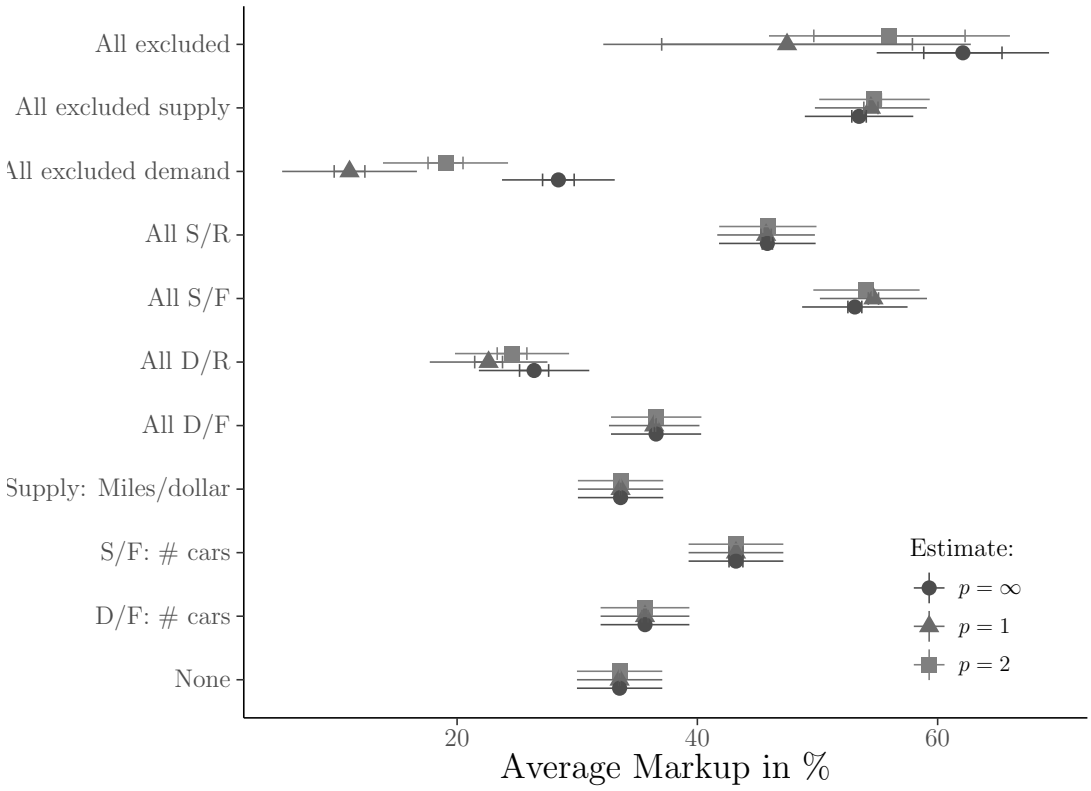


Figure 2: Optimal Confidence intervals under ℓ_1, ℓ_2, and ℓ_∞ misspecification and $M = 1$ in the application to Berry et al. (1995).

Vertical lines correspond to the estimate $\pm$ the worst case bias, and horizontal lines correspond to 95% confidence intervals. Different rows correspond to assuming that different subsets of instruments are potentially invalid. "None": correct specification. "D/F" "S/F": Demand-side instrument based on characteristics of other cars produced by the same firm. "S/F": Supply-side instrument based on characteristics of other cars produced by the same firm. "D/R": Demand-side instrument based on characteristics of cars produced by rivals. "S/R": Supply-side instrument based on characteristics of cars produced by rivals. "All excluded": All excluded instruments are potentially invalid.

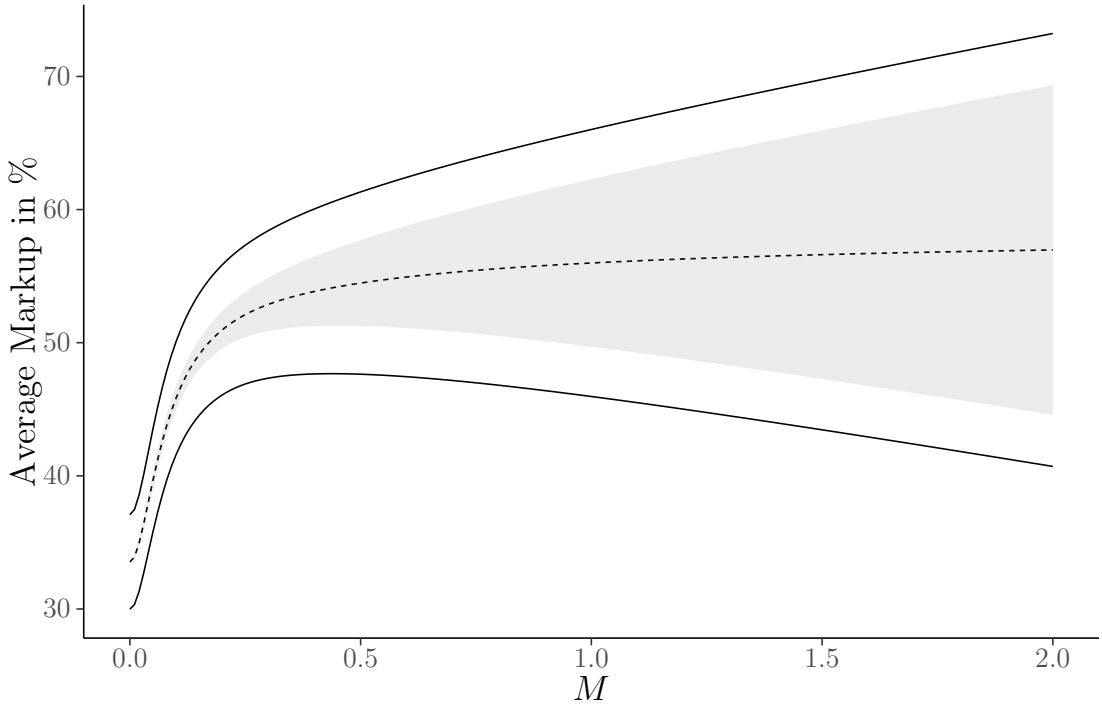


Figure 3: Optimal confidence intervals under ℓ_2 misspecification the application to Berry et al. (1995) as a function of misspecification parameter M, when all excluded instruments are allowed to be potentially invalid.

Dotted line corresponds to point estimate, shaded region denotes the estimate $\pm$ its worst-case bias, and a 95% confidence band is denoted by solid lines.