# Computationally feasible identification-robust inference on discrete choice demand

Jaewon Lee*

Yale University

October 31, 2023

Click here for the latest version.

**Abstract**

The possibility of instruments being weak is a central debate in empirical industrial organization, particularly in the context of demand estimation. This paper applies an identification-robust confidence set proposed by Andrews (2018) in the context of the discrete choice demand model by Berry, Levinsohn, and Pakes (1995), offering a technique to enhance its computational feasibility. Given that the computational cost stems from grid searches over a large dimensional parameter space, I provide a condition under which the dimensionality of the required grid can be significantly reduced. Monte Carlo simulations show that the robust confidence set obtained by utilizing my dimension reduction technique achieves desired properties when the condition is met. Even when the potentially restrictive condition, namely homoscedasticity, is violated, this technique is able to achieve a close approximation of the confidence set that would have been obtained through a full grid search. This suggests that the technique can guide the formation of a full grid, especially when a researcher lacks prior information.

*E-mail: jaewon.lee@yale.edu. I thank my advisors, Philip Haile, Steven Berry, and Yuichi Kitamura, for their constant guidance and support. I thank Donald Andrews, Xiaohong Chen, José-Antonio Espín-Sánchez, Hortense Fong, YongKi Hong, Sid Kankanala, Kyusik Kim, Vanya Klenovskiy, Changhyun Kwak, Jonas Lieber, Katja Seim, Michael Sinkinson, Suk Joon Son, Jaehee Song, and Jintaek Song for helpful discussions, advice, and encouragement. I also thank participants of Industrial Organization Prospectus Workshop and Econometrics Prospectus Lunch at Yale University. All errors are mine.

# 1 Introduction

This paper proposes an approach to inference in BLP-style demand estimation (Berry, Levinsohn, and Pakes, 1995) when instruments may be weak. Instruments for prices and quantities are essential to identification of such models, and various types of instruments have been considered in the literature. On the other hand, there are growing concerns about instruments that are inherently weak and about the behavior of the demand estimator in the presence of weak instruments.

To the best of my knowledge, this paper is the first exploration of applying an econometric method that is robust to weak identification in the context of demand estimation. To do so, I adapt the two-step approach proposed by Andrews (2018). One challenge of this kind of approach is its computational cost when the parameters are of high dimension. My approach softens this by providing a condition under which the dimensionality of expensive grid search can be significantly reduced.

Monte Carlo simulations show that the coverage probability of the conventional non-robust confidence set is compromised under weak identification, while the robust confidence set equipped with the dimension reduction technique does attain the correct coverage probability when the condition is met. Furthermore, the condition I impose for computational ease has minimal impact on the behavior of the robust confidence set. By doing so, I provide practical guidance for adapting a recent development in econometrics to an important class of applications in empirical industrial organization.

Since Berry (1994) pointed out endogeneity caused by the unobserved product heterogeneity in demand models and provided a way to reformulate the models into generalized method of moments (GMM) problems, the BLP estimator by Berry, Levinsohn, and Pakes (1995), along with their discussion of how to choose instruments, has been a central tool in empirical industrial organization. Several other types of instruments have been proposed in response to different natures of data availability. Those instrument variables are to provide sufficient exogenous variation to shift the market shares and prices independent of the demand shocks associated with each product–market pair. An important question, of course, is whether the instrument variables are strong enough, i.e., whether they provide sufficient variation in the endogenous variables. See Berry and Haile (2021) for discussion about various types of instruments and how they are potentially weak in terms of identification.

There are several proposals to enhance such instruments to increase their identification strength, also highlighting concerns regarding potentially weak instrument variables; Reynaert and Verboven (2014) propose to construct (approximate) optimal instruments based on existing instruments such as BLP instruments, in the sense of Chamberlain (1987). Gandhi and Houde (2019) propose constructing differentiation instrument variables to avoid the weak instrument variables problem.[1]

When instrument variables are weak, GMM estimators (which include the BLP estimator) are known to not behave according to the conventional asymptotics results. In particular, estimators are not asymptotically normal, and the corresponding confidence sets do not have correct coverage probability under weak IV asymptotics.

In light of this, the econometrics literature has worked on detecting weak identification and drawing inference that is robust to weak identification. For linear IV models, the first-stage $F$-test by Stock and Yogo (2005) has become a standard procedure to detect weak identification. Anderson and Rubin (1949, AR) developed a test statistic robust to weak identification, so that a weak IV-robust confidence set can be obtained by inverting the test statistic. Keane and Neal (2023) advocate using such a robust confidence set instead of the two-stage least squares estimator.

On the other hand, developments regarding nonlinear IV models are relatively more recent. To the best of my knowledge, there does not exist a popular choice for a formal test for detecting weak identification yet, if any.[2] Regarding robust inference, Stock and Yogo (2005) extended the AR test to the general nonlinear GMM case, upon which subsequent work has been based.

This paper implements such robust inference, for the first time in the context of BLP-style models. Specifically, I employ the two-step procedure by Andrews (2018). It has an advantage in that it also provides an informal test for weak identification as the first step, while guaranteeing a pre-specified level of the resulting confidence set, under both strong and weak identification.

One hurdle in implementing an identification-robust confidence set, including that of Andrews (2018), is the expensive computational cost it incurs. Existing robust confidence sets are constructed by inverting a test statistic, for which grid search is a usual choice due to its simplicity and robustness. However, grid search often prescribes a large number of points at which the test statistic is to be evaluated. In particular, the number of grid points grows exponentially in the number of parameters, even requiring days to obtain a confidence set. Another issue with grid search is that the size and location of the grid to be formed are not clear ex ante. These issues may be increasingly relevant as more product characteristics become available and therefore the dimension of the parameter space increases due to increasing data availability.

[1] Both procedures are available in the python package PyBLP by Conlon and Gortmaker (2020).

[2] See Berry and Haile (2021) for discussion. One of the few exceptions in the BLP context is Gandhi and Houde (2019) who propose a test for independence of irrelevant alternatives (IIA) as a test of weak identification.

To overcome these challenges surrounding grid search, I provide a condition under which the dimension of the grid can be reduced. The condition consists of two assumptions, namely a just-identified demand model and homoscedasticity. Under just-identification, the robust statistic used in the two-step method reduces to a nonlinear AR-type test statistic. Using homoscedasticity, I transform the test statistic into the linear AR test statistic conditional on "nonlinear" parameters, akin to the separation between nonlinear and linear parameters in the nested fixed point algorithm for BLP. Then I apply a method by Dufour and Taamouti (2005) to obtain the analytic representation of the corresponding confidence set, which then takes a simple geometrical shape known as *quadric*, of which ellipsoids are special cases. The first step of the two-step procedure additionally requires checking whether a robust confidence set is included in the non-robust confidence set. Although the confidence sets involved are now quadrics, checking such inclusion is not a trivial task. To this end, I apply S-lemma from control theory to reduce the problem to an optimization of a single variable concave function, which permits fast numerical solution.

The technique reduces the dimension of the required grid from the number of all the parameters to the number of nonlinear parameters only. I detail the algorithm that implements the procedure, with a suggestion of how to form a grid for the remaining nonlinear parameter.

After that, I leverage the reduced computation time to conduct Monte Carlo simulation exercises. The results show that the non-robust confidence set exhibits a slight undercoverage under weak identification scenarios. In contrast, the robust two-step confidence set has coverage probability around the nominal coverage probability under both weak and strong identification. I also investigate how much the homoscedasticity assumption, which is potentially restrictive, distorts the confidence set when the true errors are actually heteroscedastic. The results show that the robust confidence set obtained using the dimension reduction technique (assuming homoscedasticity) performs even better than a confidence set that is robust against heteroscedasticity but not robust against weak identification. It also approximates well the confidence set that is robust against both weak identification and heteroscedasticity (yet requiring grid search). This suggests that the method can be useful as a good approximation to the fully robust confidence set, or as guidance for how to form a grid for obtaining the confidence set robust against both weak identification and heteroscedasticity.

The paper is organized as follows. Section 2 introduces the BLP model and describes how to construct the two-step confidence set by Andrews (2018). Section 3 discusses difficulties of applying the two-step procedure, proposes a method to reduce the dimensionality of the grid search under two main assumptions, and provides an algorithm to implement it. Section

4 conducts Monte Carlo simulations to investigate the properties of the two-step confidence set with the low-dimension grid search, under both strong and weak identification scenarios, with both homoscedastic and heteroscedastic errors.

In what follows, $\dim v$ for a vector $v$ denotes the dimension of $v$, and $\dim f$ for a vector-valued function $f$ is the dimension of the range. $\chi^2(k)$ represents the chi-squared distribution with degree of freedom $k$, and $\chi^2_{1-\alpha}(k)$ is the $1-\alpha$ quantile of $\chi^2(k)$. The smallest eigenvalue of a matrix $A$ is denoted by $\lambda_{\min}A$.

# 2 The model and the two-step confidence set

In this section, I lay down the discrete choice demand model of interest (commonly referred to as the BLP model) and describe the construction of the two-step confidence set by Andrews (2018). Along the way, the concept of weak identification and some desirable properties of the two-step procedure are briefly discussed.

The rest of this paper focuses on inference on the entire vector of the parameters, rather than its sub-vector or a known function of it.[3] Also, I consider the case where a researcher only assumes the demand side model, without restrictions on the supply side, as in Nevo (2000); Conlon and Gortmaker (2020) suggest that with correctly specified supply side and optimal instruments, the BLP estimator tends to be well-behaved in finite sample, even when the instruments are fairly weak.

## 2.1 The discrete choice model and the BLP estimator

Consider a discrete choice model, where specifically, consumer $i$'s conditional indirect utility from good $j \in \{1, \ldots, J\}$ is[4]

$$u_{ijt} = x'_{jt}\beta - \alpha p_{jt} + \xi_{jt} + \mu_{ijt} + \epsilon_{ijt} \equiv \delta_{jt} + \mu_{ijt} + \epsilon_{ijt},$$

where $t = 1, \ldots, T$ denotes markets. The distribution of $\mu_{ijt}$ may be parameterized by $\gamma$. Consumer $i$'s utility depends on the characteristics $x_{jt}$, the price $p_{jt}$, the unobserved product-level heterogeneity term $\xi_{jt}$, and the individual-product-level heterogeneity terms

[3] When a researcher is interested only in a subset of parameters, deriving a confidence set for the entire vector and then taking the corresponding projection of the confidence set results in a conservative confidence set; i.e., the coverage probability is higher than the nominal coverage probability. Non-conservative inference on sub-vectors or a known non-stochastic function of the entire parameter vector is an active research area. However, in demand estimation, the objects of interest such as price elasticities or diversion ratios often involve the entire parameter vector. They are indeed functions of parameters, but those functions also depend on the data, for which case I am not aware of a result yet.

[4] I assume the number of products is the same across markets for simplicity.

$\mu_{ijt}$ and $\epsilon_{ijt}$. For example, in a random coefficient model, $\mu_{ijt} = x'_{jt}\nu_i$ with $\nu_i \sim N(0, \Sigma(\gamma))$, where $\Sigma(\gamma)$ is a positive semidefinite matrix that depends on $\gamma$. The idiosyncratic error term $\epsilon_{ijt}$ usually independently and identically follows the type I extreme value distribution. The outside option, represented by $j = 0$, is assumed to yield the utility $u_{i0t} = \epsilon_{i0t}$.

The market share is then

$$s_{jt} = \int \mathbb{1}(u_{ijt} \geq u_{ikt}\ \forall k) dF(\epsilon, \mu) = \int \frac{\exp(\delta_{jt})}{1 + \sum_{k=1}^{J} \exp(\delta_{kt})} dF(\mu)$$

where $F$ denotes the joint distribution of $\epsilon$ and $\mu$. The second equality holds under the assumption that $\epsilon_{ijt}$ follows the type I extreme value distribution.

In standard parametric discrete choice models, we can invert the market share to derive $\delta_{jt}$ from the observable variables, for a given value of $\gamma$ (Berry, 1994):[5]

$$\delta_{jt}(\gamma) \equiv \delta_{jt}(\gamma; s_t, p_t, x_t).$$

The structural error, given a parameter value $\theta = [\beta', \gamma']'$, is then

$$\xi_{jt}(\theta) = \delta_{jt}(\gamma) - x'_{jt}\beta$$

where (and hereafter) we let $x_{jt}$ and $\beta$ include $p_{jt}$ and $\alpha$ respectively, and omit the observed variables $s_t$ and $x_t$ in $\delta_{jt}(\gamma; s_t, x_t)$. As $\gamma$ enters the expression for the error nonlinearly while $\beta$ does so linearly, the parameters are called "nonlinear" and "linear" parameters respectively.

The identifying assumption is that the structural error $\xi_{jt}(\theta)$ is uncorrelated[6] with instrument variables $z_{jt}$ at (and only at) the true parameter value $\theta_0$:

$$\mathbb{E}[z_{jt}\xi_{jt}(\theta_0)] = 0.$$

Writing $g_{jt}(\theta) = z_{jt}\xi_{jt}(\theta)$, the BLP estimator $\widehat{\theta}$ (Berry, Levinsohn, and Pakes, 1995)

[5] Berry, Gandhi, and Haile (2013) provide conditions under which such an inversion is possible in a general nonparametric setting.

[6] This is usually an implication of mean-independence assumption: $\mathbb{E}[\xi_{jt}|\tilde{z}_{jt}] = 0$ almost surely, based on which $z$ is chosen as a function of instruments $\tilde{z}$.

minimizes the generalized methods of moments (GMM) criterion function:[7]

$$\widehat{\theta} = \arg\min_{\theta} \overline{g}_n(\theta)' A_n \overline{g}_n(\theta),$$

where $n = JT$ is the total number of products, $\overline{g}(\theta) = n^{-1} \sum_{jt} g_{jt}(\theta) = n^{-1} \sum_{jt} z_{jt} \xi_{jt}(\theta)$, and $A_n$ is a GMM weighting matrix that can possibly depend on $\theta$ as in the continuously updating estimator.

## 2.2 The standard confidence set

Given the GMM estimator $\widehat{\theta}$, the usual choice of confidence set is obtained by inverting the Wald statistic $W(\theta)$:

$$W(\theta) = (\widehat{\theta} - \theta)' \widehat{\Sigma}_n^{-1} (\widehat{\theta} - \theta),$$

where $\widehat{\Sigma}_n$ is a consistent estimator for the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$. The corresponding confidence set with confidence level $1 - \alpha$ is[8]

$$CS_N = \{\theta \in \Theta : W(\theta) \leq \chi^2_{1-\alpha}(\dim \theta)\},$$

where $\Theta$ is a suitable parameter space and $\chi^2_{1-\alpha}(\dim \theta)$ is the $1 - \alpha$ quantile of the $\chi^2$ distribution with degree of freedom $\dim \theta$. We shall call $CS_N$ the non-robust confidence set for the reason I explain as follows.

**Weak identification and non-robustness of $CS_N$** To discuss what weak identification means, let the expected Jacobian of the moment condition at the true parameter be defined as

$$G \equiv \mathbb{E}\left[\frac{\partial}{\partial \theta'} g_{jt}(\theta_0)\right].$$

Weak identification pertains to cases where $G$ is not full rank or $G$ is "small," whose exact meaning depends on how one models weak identification. A popular choice, for example by Staiger and Stock (1997) or Kleibergen (2005), is to have $G$ drifting to zero at the rate of $\sqrt{n}$

[7] Since the inversion of the share function to calculate $\delta_{jt}(\gamma)$ typically involves numerical integration via a Monte Carlo simulation, one needs to take into account the error from such simulation when deriving the asymptotic variance (Berry, Linton, and Pakes, 2004). In this paper, however, I assume that the number of simulation draws is sufficiently large or that the numerical integration uses other methods to achieve smaller errors (Conlon and Gortmaker, 2020), and ignore the error from numerical integration.

[8] We have two $\alpha$'s in this paper depending on the context; one is the mean price coefficient in the indirect utility function, and the other is related to the confidence level as this one.

as $n \to \infty$; e.g., $G$ depends on $n$ and $G = C/\sqrt{n}$ where $C$ is a finite matrix.[9] This modeling approach is also adopted by Andrews (2018) in the context of GMM.[10] As this paper does not pursue the derivation of new asymptotic properties of the robust confidence set, I refer readers to Andrews (2018).

In any models of weak identification, it is central to the identification status whether the Jacobian $G$ is small.[11] In the (parametric) BLP model, the Jacobian translates into

$$\left[\mathbb{E}\left[p_{jt}z_{jt}\right], -\mathbb{E}\left[z_{jt}x_{jt}'\right], \mathbb{E}\left[\left(\frac{\partial}{\partial\gamma'}\delta_{jt}(\gamma_0; s_t, p_t, x_t)\right)z_{jt}\right]\right]$$

where we write $\alpha$ and $p_{jt}$ separately from $\beta$ and $x_{jt}$ (unlike notation as above) for the sake of interpretation.

From the first part, $\mathbb{E}[p_{jt}z_{jt}]$, we see that identification is weak when the instrument variables $z_{jt}$ are not sufficiently relevant to the endogenous price $p_{jt}$. The last part is somewhat harder to interpret; we have weak identification when the instrument variables are not strongly relevant to the variability of the inverse market share function with respect to the nonlinear parameter. In a general sense, as the market share $s_t$ appears in the expression, this shows us that the instruments need to be sufficiently relevant to the market shares, in line with Berry and Haile (2014). Indeed, in a nested logit model, the last part becomes $\mathbb{E}\left[(\log s_{j|g,t})z_{jt}\right]$ where $s_{j|g,t}$ is the within-group market share of $j$ in market $t$.

It is known that, under such weak identification, the GMM estimator (and hence the BLP estimator) may not be asymptotically normal, and the corresponding standard confidence set may not have correct coverage. This motivates the use of confidence sets that are robust to weak identification, as described in the following subsection.

## 2.3 Robust confidence sets

To address incorrect coverage probability of the non-robust confidence set under weak identification, I apply the two-step procedure developed by Andrews (2018). The remainder of this

[9] A more general approach by Andrews and Guggenberger (2017) is to set a parameter space and study the asymptotic size of tests or the asymptotic coverage probability of confidence sets in a uniform sense. An advantage of this approach is that it allows for arbitrary sequences on the parameter space, including the aforementioned $\sqrt{n}$-rate sequences.

[10] However, as mentioned in that paper, the two-step procedure itself is agnostic to the modeling choice regarding weak identification, as long as its high-level assumptions are satisfied.

[11] It is worth mentioning that the boundary between strong and weak identification is often not clear-cut in finite sample, or may not be even defined when asymptotics is involved. It is not as straightforward as, for a contrived example, "the identification is weak if the matrix norm of $G$ is smaller than a certain value." Rather, when we model weak identification in terms of drifting sequences of parameters, the identification status is a feature of such *sequences*, rather than a partition of the parameter space given a fixed $T$. In this paper, however, I use the terms weak identification and strong identification loosely to avoid complication.

section reiterates Andrews (2018) on how to construct confidence sets that are robust against weak identification. Section 3 will discuss how I adapt this method in a computationally feasible way.

The procedure requires some ingredients, including the non-robust confidence set described in the previous subsection, and two robust confidence sets, denoted as $CS_P$ and $CS_N$. These two sets are constructed by inverting a test statistic, which in turn is a combination of two robust test statistics called $S$ and $K$, the latter by Kleibergen (2005). In the next section and for the rest of the paper, I assume that the demand model is just-identified. I consider this assumption to be not restrictive, given that a demand model is just-identified when optimal instruments are used, and as those can be conveniently computed using packages like PyBLP (Conlon and Gortmaker, 2020). An implication of just-identification in the current context is that $K$ coincides with $S$, and hence $K$ does not need to be calculated separately from $S$. Therefore, for simplicity, I present the robust statistic in its simplified (yet still correct) form that arises under just-identification.[12]

The weak identification literature has developed methods mainly based on the $S$ statistic, which is often called the AR-type test statistic (Anderson and Rubin, 1949; Stock and Wright, 2000). The statistic is a quadratic form of the average moment function $\overline{g}_n(\theta)$:

$$S(\theta) = n\overline{g}_n(\theta)'\Omega_n(\theta)^{-1}\overline{g}_n(\theta)$$

where $\Omega_n(\theta)$ is a consistent estimator for $\mathrm{Var}(g_{jt}(\theta))$. Under the null that $\mathbb{E}[z_{jt}\xi_{jt}(\theta)] = 0$, the test statistic converges in distribution to $\chi^2(\dim g)$, regardless of the Jacobian $\mathbb{E}[\partial g_{jt}(\theta)/\partial\theta']$, i.e., even under weak identification.[13]

To construct the test statistic, a researcher first chooses a number $\zeta > 0$ (e.g., $\zeta = 0.05$) such that she is willing to take $1-\alpha-\zeta$ as the lower bound of the coverage probability under weak identification. It turns out that the larger $\zeta$ is, the more likely the two-step procedure will indicate strong identification; we see a trade-off here. If a researcher prefers that the procedure indicates strong identification (by setting a larger $\zeta$), then she takes higher risk of misclassifying the identification situation as strong when the true data generating process in fact is of weak identification. When such misclassification occurs, the coverage probability of the reported non-robust confidence set may deviate from the nominal level $1-\alpha$, more so as $\zeta$ becomes larger. Given the choice of $\zeta$, let $a = \chi^2_{1-\alpha}(\dim\theta)/\chi^2_{1-\alpha-\zeta}(\dim\theta) - 1$.

[12] See Andrews (2018) for the construction of the robust statistic under over-identification that uses $K$.

[13] As can be seen from the construction, the test statistic directly measures how much the empirical moment condition $\overline{g}_n(\theta)$ deviates from zero, instead of being based on the estimator $\widehat{\theta}$ whose asymptotic distribution depends on the Jacobian.

Now define two robust confidence sets:

$$\begin{aligned} CS_P &= \{\theta \in \Theta : S(\theta) < \chi^2_{1-\alpha}(\dim \theta)/(1+a)\} \\ CS_R &= \{\theta \in \Theta : S(\theta) \leq \chi^2_{1-\alpha}(\dim \theta)\}. \end{aligned} \tag{1}$$

We call $CS_P$ the preliminary robust confidence set, as it will be used in the first step of the procedure. As $S(\theta)$ converges in distribution to $\chi^2(\dim g)$ under the null, which is equivalent to $\chi^2(\dim \theta)$ when the model is just-identified, we see that the critical values are designed so that $CS_P$ and $CS_R$ have coverage probability of $1 - \alpha - \zeta$ and $1 - \alpha$ respectively.

### 2.4 The two-step confidence set

With the non-robust confidence set $CS_N$ and robust confidence sets $CS_P$ and $CS_R$, the two-step confidence set is constructed as follows. The first step checks whether $CS_P \subseteq CS_N$. If so, it is interpreted as an indication of strong identification, and weak identification otherwise. The idea behind this is that both robust and non-robust confidence sets $CS_N$ and $CS_R$ behave similarly to each other under strong identification. Therefore a set smaller than $CS_R$, namely $CS_P$ in this context, will tend to be included in $CS_N$ under strong identification.

In the second step, the non-robust confidence set $CS_N$ is reported if the first step indicates strong identification, and the robust confidence set $CS_R$ is reported for weak identification; i.e., the resulting confidence set is

$$CS_2 = \begin{cases} CS_N & \text{if } CS_P \subseteq CS_N \\ CS_R & \text{if } CS_P \nsubseteq CS_N. \end{cases}$$

The role of $\zeta$ becomes clearer in this context. The preliminary set $CS_P$ is constructed by shrinking $CS_R$. As the value of $\zeta$ increases, $CS_P$ becomes smaller, resulting in a reduced coverage probability of $CS_P$, specifically $1 - \alpha - \zeta$. As $CS_P$ becomes smaller, the first step is more likely to indicate strong identification, which would increase the risk of reporting the misbehaving non-robust confidence set $CS_N$ when the true data generating process is actually weakly identified. However, even when such misclassification occurs, the set inclusion relationship between $CS_N$ and $CS_P$ helps bound the extent of the misbehavior; if the first step (wrongly) indicates strong identification, then $CS_N$ contains $CS_P$ by construction. As $CS_P$ has coverage probability of $1 - \alpha - \zeta$, the coverage probability of $CS_N$ is at least $1 - \alpha - \zeta$.

The two-step method satisfies the following properties: (i) along any strongly identified sequence of parameters, the first step indicates strong identification with probability approaching one, (ii) along any weakly identified sequence of parameters, the two-step confidence set $CS_2$ has an asymptotic coverage probability of at least $1-\alpha$, and (iii) along any weakly identified sequence of parameters, $CS_2$ has an asymptotic coverage probability of at least $1-\alpha-\zeta$. For a formal statement and sufficient regularity conditions, see Theorem 1 by Andrews (2018).

The first property suggests that the first step may be considered as an "informal" test of weak identification; it gives us a consistent test for the null of weak identification, in that the rejection probability of the test approaches one when the true data generating process is of strong identification. However, it is not a test in a usual sense, in that it does not control the size of the test; we do not know the probability of falsely rejecting the null when the true data generating process is indeed of weak identification. The Monte Carlo simulations in Section 4 suggest that the first step does indicate weak identification with high probability (if not with probability one) when the instrument variables seem weak. Still, the procedure does not pre-specify the type I error.

The third property only guarantees that $CS_2$ has a coverage probability of at least $1-\alpha-\zeta$, which is due to the coverage probability of $CS_N$ under weak identification as discussed above. However, Monte Carlo simulations in Section 4 suggest that $CS_2$ tends to achieve a coverage probability around $1-\alpha$ under weak identification. Under weak identification, the first step correctly indicates weak identification with high probability, in which case $CS_2$ coincides with $CS_R$, which is designed to have a coverage probability of $1-\alpha$.

# 3 Dimension reduction in the two-step method

The two-step method introduced in the previous section usually requires grid search, which poses computational challenges. In this section I discuss how the dimensionality of the grid can be significantly reduced, in particular when (i) inverting the robust test statistic to construct $CS_P$ and $CS_R$, and (ii) checking $CS_P \subseteq CS_N$ in the first step of the procedure. The results hinge on Assumption 1 and 2 which are introduced below. Then, I detail an algorithm that utilizes the dimension reduction technique.

## 3.1 Inverting the robust test statistic

The confidence sets in the procedure—robust and non-robust ones—are obtained by inverting the corresponding tests; for example, the non-robust confidence set $CS_N$ collects all the parameter values $\theta$ (in the parameter space) at which the test statistic $W(\theta)$ does not exceed the critical value $\chi^2_{1-\alpha}(\dim\theta)$.

The non-robust confidence set $CS_N$ is easy to handle due to its simple form. The left hand side of the inequality $(\widehat{\theta}-\theta)'\widehat{\Sigma}_n^{-1}(\widehat{\theta}-\theta) \leq \chi^2_{1-\alpha}(\dim\theta)$ is quadratic in $\theta$. As a result, various properties immediately follow. These include the facts that the confidence set is an ellipsoid centered around $\widehat{\theta}$, and that a closed form solution exists when calculating projection confidence intervals.

On the other hand, inverting a test in general is a non-trivial task in empirical work, since a confidence set can take any shape and its closed form representation can be difficult to derive (if any). The most popular approach in such cases is to conduct a grid search, due its simple implementation and robustness. A researcher specifies a finite set of points in the parameter space, evaluates the test statistic at each point, and then collects all points at which the value of the statistic does not exceed the corresponding critical value.

However, grid search has two potential problems: (i) the number of grid points grows exponentially with the number of parameters, and (ii) it is not clear ex ante where to construct the grid in the large parameter space.

As an example, the demand model by Nevo (2000) involves ten linear parameters, let alone nonlinear parameters. If one wants to apply the two-step identification-robust procedure with ten parameters by constructing a coarse grid with ten points for each parameter,[14] it would require $10^{10}$, or ten billion grid points. To translate the magnitude into computation time, it takes two minutes to run the two-step procedure for a nested logit with only five parameters with ten points on each direction, in Python on a machine having 2.8 GHz CPU with 20 cores. Increasing the dimension to ten would take 100,000 times that amount of time, i.e., about 140 days. If, in addition, the grid is made slightly denser by increasing the number of points in each direction to 20, then it would take 390 years. While a better implementation, such as running the procedure in C instead, may significantly reduce the computation time, I expect the magnitude to be still large especially with large number of parameters are involved.

Another issue is regarding the location, the size, and the denseness of the grid to be chosen by a researcher. In the mixed logit model by Nevo (2000), the linear parameter estimates show a large variation in magnitude, from 0.03 to 43.04. Obviously significant part of the variability is due to different scales of the characteristics variables, and one can standardize the variables before estimation to reduce the variability. However, even so, we do not know ex ante which characteristics would have larger or smaller coefficients, nor the overall magnitude of the coefficients relative to the idiosyncratic error term $\epsilon_{ijt}$. Robust confidence sets may be even unbounded, as shown by Dufour (1997), aggravating the issue

[14] A grid with ten points in each direction is arguably very coarse; for comparison, with two parameters, Andrews (2018) chooses a grid of $201 \times 2,641$ points.

of the width choice of a grid.

Moreover, in combination with the previous point about computational burden, an inappropriate choice of the grid may result in imprecise results; if one chooses too wide a grid but at low resolution due to computational cost, then the researcher might not detect the confidence set. For example, suppose a true confidence interval (given the data) for one of the parameters is $[3.1, 3.8]$, but a researcher forms a coarse grid $\{0, 1, \ldots, 10\}$.[15] Then the grid search would report that none of the grid points belong to the confidence interval, leading the researcher to falsely conclude that the confidence set is an empty set, although the confidence interval were to deliver a rather precise information about the parameter.

With these challenges in mind, I consider two assumptions that yield an analytic representation of the robust confidence set, similar to the non-robust Wald confidence set.

**Assumption 1** (Just-identified demand model)**.** The demand model is just-identified; i.e., the number of instrument variables equals the number of parameters.

When estimating the demand side without restricting the supply side (as mentioned at the beginning of Section 2), just-identification is automatically satisfied when optimal instruments are used.[16] When there are only demand-side structural errors $\xi_{jt}$ (rather than having another set of structural errors from the supply side), the number of the optimal instrument variables coincides with the number of the parameters. Since packages such as PyBLP (Conlon and Gortmaker, 2020) provide a readily available method to approximate the optimal instruments, just-identification is not a restrictive assumption.

The next assumption is about the data generating process.

**Assumption 2** (Homoscedasticity)**.** The unobserved product-level heterogeneity $\xi_{jt}$ is homoscedastic across $j$ and $t$, i.e., $\mathbb{E}[\xi_{jt}|z_{jt}] = \mathbb{E}[\xi_{jt}]$ almost surely.

This assumption is restrictive in general. As the structural error $\xi_{jt}$ represents unobserved product-level heterogeneity, such as product quality and latent taste variation across markets for the product, the condition may not be adequate unless the products are similar in nature, both within and across markets. Many different types of violation may occur; products supplied by a particularly innovative firm may have higher variance of $\xi_{jt}$ conditional on observed characteristics. Markets with consumers that are more sensitive to product quality may have higher variance of $\xi_{jt}$.

[15] As a reference, a Stata implementation `twostepweakiv` by Sun (2018) of the two-step method for linear instrument variables sets the default number of grid points (in each dimension) as 100, 25, 11, 7, and 5 as the dimension of the grid increases from one to five.

[16] If one is fine with obtaining a conservative confidence set, then this assumption is not needed; see remarks after Proposition 3.1.

Even if homoscedasticity is considered too restrictive, my method to analytically represent the robust confidence sets as below (under the homoscedasticity assumption) can still provide guidance for constructing a relevant grid, in case one wants to apply the two-step procedure under heteroscedasticity using grid search. I discuss this later with a Monte Carlo simulation.

Before stating the analytic representation result, let us define *partial* confidence sets. For a confidence set $CS$ and a value of nonlinear parameter $\gamma$, we call

$$CS(\gamma) = \{\beta : (\beta, \gamma) \in CS\}$$

a partial confidence set (derived from $CS$ at $\gamma$).

With the two assumptions, I derive the following result, which provides a closed form representation of the partial robust confidence sets $CS_R(\gamma)$ and $CS_P(\gamma)$ given nonlinear parameter $\gamma$.

**Proposition 3.1.** *Suppose Assumptions 1 and 2 hold. For a given distortion bound $\zeta > 0$, let*

$$a = \frac{\chi^2_{1-\alpha}(\dim \theta)}{\chi^2_{1-\alpha-\zeta}(\dim \theta)} - 1 > 0.$$

*Then the partial robust confidence sets $CS_R(\gamma)$ and $CS_P(\gamma)$ can be written in the following form*

$$\{\beta : \beta' A \beta + 2b'\beta + c \leq 0\}$$

*where*

$$A = X' \left( P_Z - \frac{\mathcal{C}}{n} M_1 \right) X$$
$$b = -X' \left( P_Z - \frac{\mathcal{C}}{n} M_1 \right) \delta$$
$$c = \delta' \left( P_Z - \frac{\mathcal{C}}{n} M_1 \right) \delta$$

*with $\mathcal{C} = \chi^2_{1-\alpha}(\dim \theta)$ for $CS_R$ and $\mathcal{C} = \chi^2_{1-\alpha}(\dim \theta)/(1 + a)$ for $CS_P$,*

$$X = [x'_{1,1}, \ldots, x'_{JT}]' \in \mathbb{R}^{n \times \dim \beta}$$
$$Z = [z'_{1,1}, \ldots, z'_{JT}]' \in \mathbb{R}^{n \times \dim \theta}$$
$$\delta = [\delta_{1,1}(\gamma), \ldots, \delta_{JT}(\gamma)]' \in \mathbb{R}^n,$$

*and projection matrices $P_Z = Z(Z'Z)^{-1}Z'$ and $M_1 = I_n - \iota_n(\iota_n'\iota_n)^{-1}\iota_n'$ where $I_n$ is the $n \times n$ identity matrix and $\iota_n \in \mathbb{R}^n$ is the vector of ones.*

*Proof.* Under just-identification, the $K$ statistic by Kleibergen (2005) coincides with the $S$ statistic by Stock and Wright (2000). Therefore the robust test statistic by Andrews (2018) can be written as $(1+a)S$, yielding confidence sets as defined in (1). By homoscedasticity, the variance of the moment function is $\text{Var}(g_i(\theta)) = \sigma_\xi^2 \mathbb{E} z_{jt} z_{jt}'$, where $\sigma_\xi^2$ is the (unconditional) variance of $\xi_{jt}(\theta)$. For this we choose an estimator $\widehat{\sigma}_\xi^2(\theta) = n^{-1} \sum_{jt} (\xi_{jt}(\theta) - \overline{\xi}_n(\theta))^2$ where $\overline{\xi}_n(\theta) = n^{-1} \sum_{jt} \xi_{jt}(\theta)$. Using matrix notation,

$$\widehat{\sigma}_\xi^2(\theta) = \frac{1}{n} \xi(\theta)' M_1 \xi(\theta)$$

where $\xi(\theta) = [\xi_{1,1}, \ldots, \xi_{J,T}]' \in \mathbb{R}^n$, and also

$$\begin{aligned}
\widehat{\Sigma}_g(\theta) &= \frac{\widehat{\sigma}_\xi^2(\theta)}{n} \sum_{jt} z_{jt} z_{jt}' = \frac{\widehat{\sigma}_\xi^2(\theta)}{n} Z'Z \\
\overline{g}_n(\theta) &= \frac{1}{n} \sum_{jt} z_{jt} \xi_{jt}(\theta) = \frac{1}{n} Z'\xi(\theta).
\end{aligned}$$

Now the $S$ statistic can be written as

$$\begin{aligned}
S(\theta) &= n \overline{g}_n(\theta)' \widehat{\Sigma}_g(\theta)^{-1} \overline{g}_n(\theta) \\
&= \xi(\theta)' Z \left( \widehat{\sigma}_\xi^2(\theta) Z'Z \right)^{-1} Z' \xi(\theta) \\
&= \frac{\xi(\theta)' P_Z \xi(\theta)}{\xi(\theta)' M_1 \xi(\theta)/n} \\
&= \frac{(\delta - X\beta)' P_Z (\delta - X\beta)}{(\delta - X\beta)' M_1 (\delta - X\beta)/n}.
\end{aligned}$$

The inequality for the robust test with a critical value $\mathcal{C}$ is then $S(\theta) \leq \mathcal{C}$, which is equivalent to

$$\begin{aligned}
(\delta - X\beta)' P_Z (\delta - X\beta) &\leq \frac{\mathcal{C}}{n} (\delta - X\beta)' M_1 (\delta - X\beta), \\
(\delta - X\beta)' \left[ P_Z - \frac{\mathcal{C}}{n} M_1 \right] (\delta - X\beta) &\leq 0.
\end{aligned}$$

Observing that the left hand side is quadratic in $\beta$ given $\gamma$ (and hence given $\delta$), the result follows from arranging the terms. $\square$

The proof combines ideas by Stock and Wright (2000) and Dufour and Taamouti (2005);

it first exploits the structure of the BLP-like models, to transform the nonlinear AR-type test statistic into a linear AR test statistic (conditional on nonlinear parameters). Then it derives a quadratic representation of the confidence sets. The resulting partial confidence sets $CS_R(\gamma)$ and $CS_P(\gamma)$ are *quadrics*, i.e., the shapes defined by quadratic inequalities, whose geometrical properties are known. Dufour and Taamouti (2005) provide such properties, including conditions under which a quadric is bounded as well as the analytic solution of its projections. For later reference, I collect them (from Theorems 4.1 and 5.1–5.3 therein) in Proposition A.1 under the assumption that $A$ is nonsingular.[17]

By providing analytic representations of $CS_R(\gamma)$ and $CS_P(\gamma)$, Proposition 3.1 reduces the dimensionality of forming a grid and searching over the grid from $\dim\theta$ to $\dim\gamma$. As discussed above, this not only decreases computational burden but also gets rid of having to choose an appropriate grid over $\beta$ without prior knowledge; as Proposition A.1 shows, robust confidence sets may be unbounded under weak identification. The quadratic representation allows us to immediately check the boundedness of a partial robust confidence set. If it is bounded, then projections of the set can be easily obtained, again by invoking Proposition A.1.

As demonstrated in the proof, the purpose of just-identification assumption is to replace a robust statistic $K$ with $S$, since they are the same under just-identification. When the model is over-identified, one can still proceed as the procedure prescribes (i.e., by inverting $S(\theta)$ and therefore constructing $A$, $b$, and $c$ as in Proposition 3.1). The resulting two-step confidence set, however, will have a higher coverage probability compared to a confidence set that inverts the original statistic $K(\theta) + aS(\theta)$.

Note that we still need grid search over $\gamma$.[18] This resembles different roles played by the two sets of parameters in the BLP estimator; in minimization of the BLP-GMM objective function, each trial value of nonlinear parameter $\gamma$ requires a market share inversion. On the other hand, given $\gamma$, the value of $\beta$ that minimizes the objective function is analytically solved using the linearity of $\delta$ with respect to $\beta$, thereby reducing the dimensionality of nonlinear optimization from $\dim\theta$ to $\dim\gamma$.

As a non-robust confidence set is always an ellipsoid, its partial version is also an ellipsoid. I conclude this subsection by presenting a quadratic representation of the partial non-robust confidence set. Note that, in the proposition, $[\widehat{\Sigma}^{-1}]_{\beta\beta}$ is *not* the inverse of the top left $(\dim\beta \times \dim\beta)$ block of $\widehat{\Sigma}$, but the top left $(\dim\beta \times \dim\beta)$ block of the inverse of $\widehat{\Sigma}$.

**Proposition 3.2.** *Let $\widehat{\Sigma}$ be a consistent estimator of the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$*

[17] This is also a maintained assumption in Dufour and Taamouti (2005), as $A$ being singular is unlikely. See Dufour and Taamouti (2007) for singular $A$.

[18] We discuss a rule-of-thumb way to form a grid for nonlinear parameters in Section 3.3.

*where the elements of $\theta$ are ordered as $\theta = [\beta', \gamma']'$ and analogously for $\widehat{\theta}$. Let the blocks of its inverse be denoted as*

$$\widehat{\Sigma}^{-1} = \begin{bmatrix} [\widehat{\Sigma}^{-1}]_{\beta\beta} & [\widehat{\Sigma}^{-1}]_{\beta\gamma} \\ [\widehat{\Sigma}^{-1}]'_{\beta\gamma} & [\widehat{\Sigma}^{-1}]_{\gamma\gamma} \end{bmatrix}$$

*with $[\widehat{\Sigma}^{-1}]_{\beta\beta} \in \mathbb{R}^{\dim\beta \times \dim\beta}$, $[\widehat{\Sigma}^{-1}]_{\beta\gamma} \in \mathbb{R}^{\dim\beta \times \dim\gamma}$, and $[\widehat{\Sigma}^{-1}]_{\gamma\gamma} \in \mathbb{R}^{\dim\gamma \times \dim\gamma}$.*

*Then the partial non-robust Wald confidence set $CS_N(\gamma)$ can be written as $\{\beta : \beta' A \beta + 2b'\beta + c \leq 0\}$ where*

$$\begin{aligned} A &= [\widehat{\Sigma}^{-1}]_{\beta\beta} \\ b &= [\widehat{\Sigma}^{-1}]_{\beta\gamma}(\gamma - \widehat{\gamma}) - [\widehat{\Sigma}^{-1}]_{\beta\beta}\widehat{\beta} \\ c &= \widehat{\beta}'[\widehat{\Sigma}^{-1}]_{\beta\beta}\widehat{\beta} - 2\widehat{\beta}'[\widehat{\Sigma}^{-1}]_{\beta\gamma}(\gamma - \widehat{\gamma}) + (\gamma - \widehat{\gamma})'[\widehat{\Sigma}^{-1}]_{\gamma\gamma}(\gamma - \widehat{\gamma}) - \chi^2_{1-\alpha}(\dim\theta)/n. \end{aligned}$$

*Proof.* The results is a rearrangement of the inequality $(\widehat{\theta} - \theta)'\widehat{\Sigma}_n(\widehat{\theta} - \theta) \leq \chi^2_{1-\alpha}(\dim\theta)$. □

## 3.2 Checking inclusion of confidence sets

The first step of the two-step procedure requires us to check $CS_P \subseteq CS_N$. The Wald confidence set $CS_N$ is an ellipsoid. On the other hand, the robust (preliminary) confidence set $CS_P$ does not have a known shape in general. Consequently, without any structures, one would resort to a grid search over $CS_P$ to check $CS_P \subseteq CS_N$, which can be computationally expensive, or might not detect a non-inclusion if the grid is too coarse.

Proposition 3.1 provided a condition under which $CS_P(\gamma)$ takes a known form, namely ellipsoid. Then the problem becomes checking whether an ellipsoid $CS_P(\gamma)$ is included in another ellipsoid $CS_N(\gamma)$ for all $\gamma$. The following proposition provides a feasible method to check the inclusion, even when the centers of the ellipsoids do not coincide.[19]

**Proposition 3.3.** *Let $CS_P(\gamma)$ and $CS_N(\gamma)$ be two ellipsoids in $\mathbb{R}^k$ represented by*

$$\begin{aligned} CS_P(\gamma) &= \{x \in \mathbb{R}^k : x' A_P x + 2b_P' x + c_P \leq 0\} \\ CS_N(\gamma) &= \{x \in \mathbb{R}^k : x' A_N x + 2b_N' x + c_N \leq 0\}. \end{aligned}$$

[19] When two ellipsoids share the same center, then it is sufficient to check whether a matrix is positive semidefinite; suppose we have two ellipsoids represented by $E_1 = \{x : (x - x_0)' A_1 (x - x_0) \leq 1\}$ and $E_2 = \{x : (x - x_0)' A_2 (x - x_0) \leq 1\}$. Then $E_1 \subseteq E_2$ iff $A_1 - A_2$ is positive semidefinite.

*where* $A_P, A_N \in \mathbb{R}^{k\times k}$*,* $b_P, b_N \in \mathbb{R}^k$*,* $c_P, c_N \in \mathbb{R}$*. Define a function* $\phi : \mathbb{R} \to \mathbb{R}$ *by*

$$\phi(t) = \lambda_{\min}\left(t\begin{bmatrix} A_P & b_P \\ b_P' & c_P \end{bmatrix} - \begin{bmatrix} A_N & b_N \\ b_N' & c_N \end{bmatrix}\right).$$

*Then* $\phi$ *is well-defined and concave. Suppose* $CS_P(\gamma) \neq \emptyset$*. Then* $CS_P(\gamma) \subseteq CS_N(\gamma)$ *if and only if there exists* $t \geq 0$ *such that* $\phi(t) \geq 0$*.*

*Proof.* The function $\phi$ is well-defined since the matrix

$$M(t) = t\begin{bmatrix} A_P & b_P \\ b_P' & c_P \end{bmatrix} - \begin{bmatrix} A_N & b_N \\ b_N' & c_N \end{bmatrix}$$

is symmetric for all $t \in \mathbb{R}$. It is concave because it is the composition of an affine function $t \mapsto M(t)$ and a concave function $M \mapsto \lambda_{\min} M$.

S-lemma (see e.g., Boyd and Vandenberghe, 2004) shows that $CS_P(\gamma) \subseteq CS_N(\gamma)$ if and only if there exists $t \geq 0$ such that $M(t)$ is positive semidefinite. The result follows since $M(t)$ is positive semidefinite if and only if $\lambda_{\min} M(t) \geq 0$. □

Since $\phi$ is a single variable concave function, it is easy to check the condition "$\phi$ attains a non-negative value on $[0, \infty)$"; for example, first check whether $\phi(0) \geq 0$. If that is the case, then the condition holds. If not, and if $\phi$ is decreasing at 0, then the condition does not hold. Otherwise, apply gradient ascent until $\phi$ takes a positive value (in which case we conclude that the condition holds) or arrives at the global maximum. If the global maximum is non-negative, then the condition holds. If the global maximum is negative, then the condition does not hold. The only case in which the procedure fails is when $\phi(t)$ increases but does not hit zero as $t \to \infty$.[20] In practice, one can try finding the maximum of $\phi$ within a large enough bounded interval that the machine can handle, say $[0, 10^{16}]$.

Remarks on dimensionality reduction by Proposition 3.1 pertains here as well; exploiting the structure under the assumptions, checking $CS_P \subseteq CS_N$ no longer requires a grid search on $CS_P$ of dimension $\dim \theta$, but only of $\dim \gamma$.

## 3.3 The algorithm for two-step confidence set

I summarize the algorithm here using the previous results under Assumptions 1 and 2, and then discuss a few details regarding the algorithm.

[20] This may happen in the most general setting; $\lambda_{\min}\left(t\begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} - \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}\right)$ is always negative while approaching zero as $t \to \infty$. However, it is not clear whether such a case may arise in our setting, in which we have restrictions such as $A_1$ and $A_2$ must be positive definite.

1. Obtain the BLP estimate $\widehat{\theta}$ as well as the estimate $\widehat{\Sigma}$ for the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$ (under homoscedasticity).
2. Choose the bound of distortion $\zeta$, say 0.05, and set $a = \chi^2_{1-\alpha}(\dim\theta)/\chi^2_{1-\alpha-\zeta}(\dim\theta) - 1$.
3. Form a grid $\widetilde{\Gamma}$ in the parameter space for $\gamma$.
4. Set a weak instrument indicator variable, say `weakiv`, as false.
5. For each grid point $\gamma$ do the following:
   (a) Compute $\delta(\gamma) = [\delta_{1,1}(\gamma; s_1, p_1, x_1), \ldots \delta_{J,T}(\gamma; s_T, p_T, x_T)]' \in \mathbb{R}^{J\times T}$.
   (b) Obtain $CS_R(\gamma)$ using Proposition 3.1 and store it.
   (c) If `weakiv` is true, then continue to the next value of $\gamma$.
   (d) Obtain $CS_P(\gamma)$ using Proposition 3.1.
   (e) If $CS_P(\gamma)$ is not bounded (using Proposition A.1), then set `weakiv` as true and continue to the next value of $\gamma$.
   (f) If $CS_P(\gamma) \not\subseteq CS_N(\gamma)$ (using Proposition 3.3), then set `weakiv` as true.
6. If `weakiv` is true, then report $CS_R = \{(\beta, \gamma) : \beta \in CS_R(\gamma), \gamma \in \widetilde{\Gamma}\}$ as $CS_2$. Otherwise, report $CS_N$ as $CS_2$.

**Obtaining optimal instruments** To ensure just-identification when there are more instrument variables than parameters, one can apply a convenient method provided by PyBLP to obtain the optimal instruments; before Step 1, obtain the BLP estimate, and then use it to calculate the optimal instruments.[21] Then begin with Step 1 with the newly constructed set of instruments.

**Storing confidence sets** At Step 5 (b), only the objects $A$, $b$, and $c$ (as in Proposition 3.1) for each $\gamma$ need to be stored in the memory, rather than the set of points that are in $CS_R(\gamma)$, since they determine $CS_R(\gamma)$.

**Grid search over nonlinear parameters** The algorithm still requires grid search over nonlinear parameters $\gamma$, even though grid search over $\beta$ is no longer present, which mirrors similar roles taken by nonlinear and linear parameters in BLP estimation procedure, as mentioned earlier. However, unlike the minimization task over $\gamma$ as in BLP estimation, where several methods have been studied extensively and are readily available in optimization

[21] See `ProblemResult.compute_optimal_instruments` in PyBLP.

packages, research on effective and accurate grid search over a potentially large parameter space is still growing.[22]

One rule of thumb we may consider is to form a grid *slightly larger* than the set $\{\gamma : CS_N(\gamma) \neq \emptyset\}$, i.e., the projection of $CS_N$ on the space of nonlinear parameters, which can be analytically obtained by using Proposition A.1, since (non-partial) $CS_N$ is already an ellipsoid. On that grid we can check whether $CS_P(\gamma) \subseteq CS_N(\gamma)$ for all $\gamma$; if it is violated, then say that the first step indicates weak identification (up to the choice of the denseness of the grid), and then obtain $CS_R$ while sequentially enlarging the grid for $\gamma$ as needed to enclose $CS_R$. Otherwise, if $CS_P(\gamma) \subseteq CS_N(\gamma)$ on the grid (which was taken slightly large so that it includes values of $\gamma$ at which $CS_N(\gamma)$ is empty), then say that the first step indicates strong identification, albeit with caution since there might still be a value of $\gamma$ outside the grid for which $CS_P(\gamma) \neq \emptyset$.

**Nested logit model** As in the BLP estimator for the nested logit model, the "nesting parameter," which is a nonlinear parameter, can be considered as a linear parameter when it comes to computation. Using the notation by Berry (1994), consider the nested logit model (omitting subscript $t$)

$$u_{ij} = x_j'\beta + \xi_j + \zeta_{ig} + (1 - \sigma)\epsilon_{ij}$$

where $\sigma \in (0, 1]$ is the nesting parameter and $g$ denotes the group $j$ belongs to. The model can be analytically inverted to yield

$$\log s_j - \log s_0 = x_j'\beta + \sigma \log s_{j|g} + \xi_j,$$

where $s_{j|g}$ is the within-group market share of $j$. Then the algorithm can be applied without a loop over $\gamma$ after the following renaming: $\delta_{jt} \leftarrow \log s_g - \log s_0$, $\theta \leftarrow [\beta', \sigma]'$, and $x_j \leftarrow [x_j', \log \bar{s}_{j|g}]'$.

**Application in transformation models** The dimension reduction technique can be also useful when applying the robust two-step inference to transformation models (Horowitz, 1998):

$$T(y_i; \gamma) = x_i'\beta + u_i,$$

[22]Instead of a grid search, one could follow the approach suggested by Chen, Christensen, and Tamer (2018) and use Monte Carlo draws from a quasi-posterior based on the robust test statistic.

where $y_i, u_i \in \mathbb{R}$ and $x_i \in \mathbb{R}^{\dim \beta}$, and the function $T$ is invertible and is parametrized by $\gamma$. In this case, the inverse of $T$ with respect to $y_i$, i.e., $T^{-1}(\gamma; y_i)$, takes the role of $\delta_{jt}(\gamma; s_t, p_t, x_t)$.

# 4 Monte Carlo simulations

In this section, I conduct Monte Carlo simulations to investigate the performance of the two-step procedure equipped with the dimension reduction technique. I first set the data generating process, and then conduct Monte Carlo simulations under both homoscedasticity and heteroscedasticity. We use the same notation as in Section 2.1, while omitting the market subscript $t$.

## 4.1 The data generating process

We vary the number of markets as $T = 100, 200, 500$ in the simulation exercises. In all cases, there are $J = 6$ inside goods in each market. There are 6 firms in each market, each producing one product. The indirect utility is specified as a logit model with a random coefficient on the price, as

$$u_{ij} = 1 - (3 + 0.5\nu_{ij})p_j + 1.5x_{1j} + 1.5x_{2j} + \xi_j + \epsilon_{ij},$$

where $x_{1j}$ and $x_{2j}$ are drawn independently from the standard uniform distribution, $\nu_{ij}$ follows the standard normal distribution,[23] and $\epsilon_{ij}$ follows the type-I extreme value distribution. Consequently, the true parameters are $\beta = [1, -3, 1.5, 1.5]'$, $\alpha = -3$ (which is part of $\beta$), and $\gamma = 0.5$.

The marginal cost $c_{ij}$ is determined by

$$c_j = 2x_{1j} + 2x_{2j} + \rho w_j + \omega_j,$$

where $w_j$ is a cost shifter drawn from the standard uniform distribution. For homoscedastic errors, the unobserved product heterogeneity $\xi_j$ in the previous display and the unobserved cost shifter $\omega_j$ are drawn from a bivariate normal distribution such that each has variance one, and the correlation coefficient between $\xi_j$ and $\omega_j$ is 0.9.

For heteroscedasticity, I multiply $\xi_j$ with a factor of $\sqrt{2(1 - w_j)}$. As $w_j$ is uniformly distributed on $[0, 1]$ and the original $\xi_j$ has variance one, the new $\xi_j$ has conditional variance between 0 and 2, depending on the cost shifter.

[23] In a typical empirical setting, the random component of the price coefficient is the exponential of some random variable, to ensure that the price coefficient is negative for each individual. Instead, I take advantage of the fact that $3 + 0.5\nu_{ij}$ is rarely negative; $P(3 + 0.5\nu_{ij} < 0) \approx 9.87 \times 10^{-10}$.

The parameter $\rho$ is set differently across simulations, in order to vary the degree of weak identification; decreasing $\rho$ reduces the correlation between the price and the cost shifter $w_j$, thereby leading to weak identification. This setting is also used by Conlon and Gortmaker (2020).

The endogenous market shares and prices, namely $s_j$ and $p_j$, are generated by solving the Bertrand-Nash price-setting game between firms. I use the Python package PyBLP (Conlon and Gortmaker, 2020) to conduct simulation draws and to solve for the endogenous variables.

The excluded instrument variables I use are the cost shifter $w_j$ and some simple BLP instruments, namely the sums of rival products' characteristics (within the market):

$$z_j = \left[1, x_{1j}, x_{2j}, w_j, \sum_{k \neq j} x_{1k}, \sum_{k \neq j} x_{2k}\right]'.$$

Since we have five parameters and six (included and excluded) instruments, I use PyBLP to compute the approximate optimal instrument and call the resulting vector of instruments $z_j$, thereby reducing the dimension of $z_j$ to five.

## 4.2 Confidence sets under homoscedastic errors

Table 1: Coverage of confidence sets under homoscedasticity

| | $T = 100$ | | | $T = 200$ | | | $T = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | 1 | 3 | 5 | 1 | 3 | 5 | 1 | 3 | 5 |
| $\mathrm{corr}(p_j, w_j)$ | 0.217 | 0.558 | 0.747 | 0.216 | 0.548 | 0.745 | 0.216 | 0.556 | 0.745 |
| Weak IV | 1.000 | 1.000 | 0.944 | 1.000 | 1.000 | 0.996 | 1.000 | 1.000 | 0.138 |
| Coverage of $CS_N$ | 0.820 | 0.864 | 0.891 | 0.844 | 0.888 | 0.891 | 0.865 | 0.880 | 0.909 |
| Coverage of $CS_R$ | 0.900 | 0.900 | 0.894 | 0.911 | 0.905 | 0.890 | 0.918 | 0.881 | 0.910 |
| Coverage of $CS_2$ | 0.900 | 0.900 | 0.894 | 0.911 | 0.905 | 0.890 | 0.918 | 0.881 | 0.907 |
| Length of $CS_{\alpha,N}$ | 1.990 | 1.105 | 0.669 | 1.441 | 0.892 | 0.606 | 0.924 | 0.523 | 0.355 |
| Length of $CS_{\alpha,R}$ | 2.840 | 1.226 | 0.719 | 1.665 | 0.921 | 0.619 | 0.973 | 0.550 | 0.359 |
| Length of $CS_{\gamma,N}$ | 0.679 | 0.291 | 0.139 | 0.512 | 0.251 | 0.139 | 0.322 | 0.149 | 0.079 |
| Length of $CS_{\gamma,R}$ | 1.010 | 0.355 | 0.153 | 0.655 | 0.265 | 0.143 | 0.357 | 0.152 | 0.080 |

*Notes:* $\mathrm{corr}(p_j, w_j)$ is the average correlation between the price and the cost shifter, across simulation draws. Weak IV denotes the sample probability that the first step indicates weak identification. Coverage of $CS$ denotes the sample probability of $\theta \in CS$, i.e., the coverage probability. Length of $CS_\alpha$ and Length of $CS_\gamma$ are the average lengths of the projection confidence intervals for $\alpha$ and $\gamma$, respectively. Each column is obtained using 1,000 simulation draws.

Table 1 tabulates simulated behaviors of confidence sets $CS_N$, $CS_R$, and $CS_2$, when the structural errors are homoscedastic. The nominal coverage rate is set to 0.90, and the coverage distortion bound $\zeta$ is set to 0.10. Overall, the robust confidence set $CS_R$ has coverage probabilities around the nominal coverage probability, as expected, for all $\rho$ considered in the exercise. The robust confidence set behaves well even for a fairly small $T$. On the other hand, the non-robust confidence set $CS_N$ shows slight under-coverage for low values of $\rho$, more so as $T$ gets smaller.

The first step tends to indicate strong identification when $T = 500$ and $\rho = 5$, and weak identification otherwise. One thing to note is that the procedure indicates strong identification only when $T$ is large, even at the same level of $\rho$ (and hence with the same correlation between $p_j$ and $w_j$). This seems to be due to $CS_N$ being less stable with small $T$, which in turn affects whether the first-step set inclusion holds.

As for the two-step confidence set $CS_2$, it performs well in all simulation cases, attaining the coverage probability around 0.90. This simulation result is favorable, especially considering the theoretically guaranteed lower bound of 0.80. This happens because the first step correctly forces $CS_2$ to discard $CS_N$ and use $CS_R$ instead when $CS_N$ performs poorly.

If correct coverage probability is the only concern, then one could have just used $CS_R$ to begin with, as it already attains the correct coverage probability. However, one benefit of using the two-step procedure is that it gives an indicator of weak identification, analogously to the $F$-test in the linear instrument variables models. Another potential benefit can be found in the lower part of the table. The lengths of projections of $CS_N$ are smaller than those of $CS_R$ in all cases. Due to this, one might find it desirable to use $CS_N$ when the first step indicates that it is “safe” to use $CS_N$.

Nonetheless, the results suggest that there is little cost to using $CS_R$ (i.e., a slightly larger confidence set under strong identification), while the benefit is large (i.e., the correct coverage probability as compared to that of $CS_N$ under weak identification). Therefore one might as well consider using $CS_R$ without the two-step procedure, in spirit of Keane and Neal (2023). Even in that case, the dimension reduction technique provided in Proposition 3.1 remains useful.

### 4.3 Confidence sets under heteroscedastic errors

Table 2 tabulates simulated behaviors of confidence sets $CS_N$, $CS_R$, $CS_2$, and $CS_H$. Here, $CS_H$ is the Wald confidence set obtained by inverting the Wald test statistic, like $CS_N$, but using the heteroscedasticity-robust variance estimator; that is, $CS_H$ is robust against heteroscedasticity but not robust against weak identification.[24]

[24] A confidence set that is robust against both heteroscedasticity and weak identification would be an ideal benchmark. However, I do not simulate it due to its extensive computational cost. See Figure 1 for such

Table 2: Coverage of confidence sets under heteroscedasticity

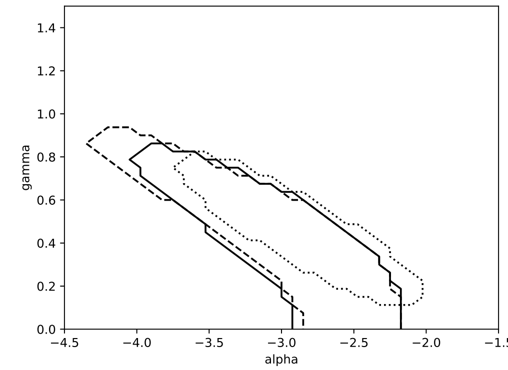
| | $T = 100$ | | | $T = 200$ | | | $T = 500$ | | |
|---|---|---|---|---|---|---|---|---|---|
| $\rho$ | 1 | 3 | 5 | 1 | 3 | 5 | 1 | 3 | 5 |
| $\mathrm{corr}(p_j, w_j)$ | 0.218 | 0.555 | 0.745 | 0.219 | 0.557 | 0.745 | 0.217 | 0.556 | 0.745 |
| Weak IV | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 1.000 | 0.272 |
| Coverage of $CS_N$ | 0.832 | 0.854 | 0.858 | 0.813 | 0.869 | 0.883 | 0.833 | 0.872 | 0.883 |
| Coverage of $CS_R$ | 0.914 | 0.885 | 0.870 | 0.887 | 0.895 | 0.885 | 0.887 | 0.888 | 0.885 |
| Coverage of $CS_2$ | 0.914 | 0.885 | 0.870 | 0.887 | 0.895 | 0.885 | 0.887 | 0.888 | 0.885 |
| Coverage of $CS_H$ | 0.748 | 0.822 | 0.860 | 0.756 | 0.847 | 0.886 | 0.815 | 0.873 | 0.880 |
| Length of $CS_{\alpha,N}$ | 2.148 | 1.285 | 0.887 | 1.488 | 0.878 | 0.600 | 0.920 | 0.533 | 0.348 |
| Length of $CS_{\alpha,R}$ | 2.484 | 1.363 | 0.924 | 1.655 | 0.905 | 0.614 | 0.964 | 0.540 | 0.353 |
| Length of $CS_{\alpha,2}$ | 2.484 | 1.363 | 0.924 | 1.655 | 0.905 | 0.614 | 0.964 | 0.540 | 0.348 |
| Length of $CS_{\alpha,H}$ | 2.174 | 1.315 | 0.904 | 1.512 | 0.896 | 0.614 | 0.935 | 0.546 | 0.355 |
| Length of $CS_{\gamma,N}$ | 0.740 | 0.370 | 0.208 | 0.531 | 0.249 | 0.138 | 0.324 | 0.147 | 0.077 |
| Length of $CS_{\gamma,R}$ | 0.862 | 0.416 | 0.219 | 0.646 | 0.261 | 0.142 | 0.354 | 0.149 | 0.079 |
| Length of $CS_{\gamma,2}$ | 0.862 | 0.416 | 0.219 | 0.646 | 0.261 | 0.142 | 0.354 | 0.149 | 0.077 |
| Length of $CS_{\gamma,H}$ | 0.739 | 0.367 | 0.204 | 0.533 | 0.246 | 0.136 | 0.325 | 0.146 | 0.076 |

*Notes:* See Table 1 for a description of terms "corr", "coverage", and "length." "Weak IV" is the sample probability that the first step indicates weak identification, *under the assumption that the errors are homoscedastic*. Likewise, $CS_N$, $CS_R$, and $CS_2$ are obtained based on the assumption that the errors are homoscedastic. $CS_H$ is the Wald confidence set that is robust against heteroscedasticity while not robust against weak identification.

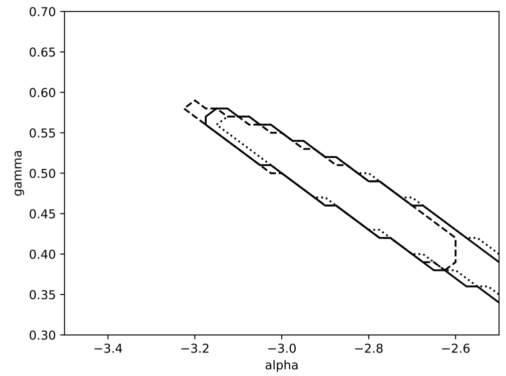
Even though the true data is generated under heteroscedasticity, the weak-identification robust (but not robust against heteroscedasticity) confidence set $CS_R$ does not suffer a severe distortion in coverage probability, although slight under-coverage is present. $CS_N$, which is not robust against both heteroscedasticity and weak identification, shows slight under-coverage, more so for smaller $T$ and lower $\rho$. Heteroscedasticity-robust (but not weak-identification robust) confidence set $CS_H$ exhibits severe under-coverage for smaller $T$ and lower $\rho$; although $CS_H$ is supposed to correct $CS_N$ to be robust against heteroscedasticity, its performance appears worse than $CS_N$ under weak identification. This might need further research to understand the cause of the particularly poor performance of heteroscedasticity-robust estimator under weak identification. Still, the simulation result hints that addressing weak identification might be more important than addressing heteroscedasticity when there is a concern of potential weak identification.

Figure 1 depicts, for each of three scenarios, a simulation draw of each of three confidence sets: $CS_2$ under the homoscedasticity assumption (solid line), $CS_2$ without the homoscedas-

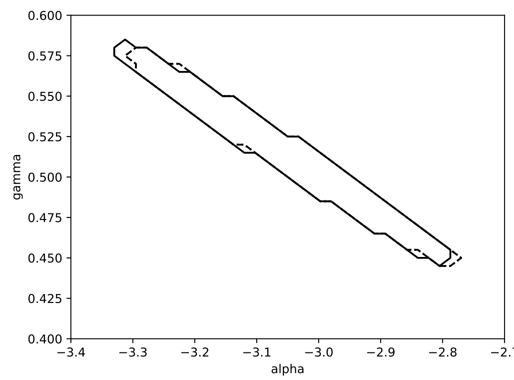
confidence set plotted based on a single simulation draw.




(a) $T = 100, \rho = 1$




(b) $T = 200, \rho = 3$




(c) $T = 500, \rho = 5$

Figure 1: Confidence sets under heteroscedasticity

*Notes:* In each panel, a confidence set with a solid line is $CS_2$ obtained using the dimension reduction technique, under the assumption that errors are homoscedastic. A confidence set with a dashed line is $CS_2$ obtained by grid search over $\theta$, without assuming that errors are homoscedastic. A confidence set with a dotted line is $CS_H$, i.e., the confidence set that is robust against heteroscedasticity while not robust against weak identification. All the confidence sets are projected on the $\alpha$–$\gamma$ plane. In panel (c), the dotted line and the dashed line overlap. The panels have different scales.

ticity assumption (dashed line), and $CS_H$ (dotted line). Let us call the first two confidence sets $CS_2^{\text{hom}}$ and $CS_2^{\text{het}}$ respectively. Among the three confidence sets, $CS_2^{\text{het}}$ is designed to be robust against both heteroscedasticity and weak identification, and therefore serves as a benchmark.

The confidence sets almost overlap under strong identification ($T = 500, \rho = 5$). On the other hand, under weak identification ($T = 100, \rho = 1$), the $CS_2^{\text{hom}}$ and $CS_H$ deviate from $CS_2^{\text{het}}$. Between $CS_2^{\text{hom}}$ and $CS_H$, the latter seems particularly worse in capturing the overall shape of $CS_2^{\text{het}}$, since $CS_H$ is bound to be an ellipsoid in every direction.[25] On the other hand, $CS_2^{\text{hom}}$, which takes advantage of dimension reduction, seems to approximate the shape and the location of $CS_2^{\text{het}}$, which requires a grid search over the entire dimension and hence can be computationally prohibitive when the dimensionality is large and the grid is fine. This visual inspection suggests two potential uses for the analytic representation technique: (i) as an approximation for the fully robust confidence set $CS_2^{\text{het}}$, or (ii) as guidance on the location and the breadth of the grid for computing $CS_2^{\text{het}}$.

# 5 Conclusion

In this paper, in light of concerns about instrument variables being weak in estimating discrete choice demand models (BLP-style models), I adapt a recent econometric method that is robust to weak identification, in particular the two-step procedure by Andrews (2018). As the procedure involves computationally intensive grid searches especially as the number of parameters increases, I propose a computationally feasible method by reducing the dimensionality of the grid searches under some condition. This is done by deriving an analytic representation of the robust confidence sets in the space of linear parameters and by providing a fast method to check inclusion between two ellipsoids. As a result, the dimensionality of the required grid reduces from the total number of parameters to just the number of nonlinear parameters.

I conduct Monte Carlo simulations to check the performance of the two-step procedure equipped with my computationally feasible method. The coverage probability of the two-step confidence set is around the pre-specified level under both strong and weak identification, while the non-robust confidence set shows slight under-coverage when instruments are weak. To assess the impact of the homoscedasticity assumption I impose for dimension reduction, I compare several confidence sets while setting the true structural errors to be heteroscedastic. The simulation results show that the impact of the assumption is minimal, suggesting that

[25] On the other hand, $CS_2^{\text{hom}}$ is an ellipsoid only after fixing $\gamma$. In the Figure 1, this shape restriction presents itself in that the projection confidence set for $\alpha$ is an interval for each $\gamma$.

the dimension reduction technique can be used as a fast approximation to a full robust inference or as guidance for how to form a grid that can be used in the full robust inference, even under heteroscedasticity.

# References


Anderson, T. W. and H. Rubin (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of mathematical statistics 20*(1), 46–63.

Andrews, D. W. and P. Guggenberger (2017). Asymptotic size of kleibergen's lm and conditional lr tests for moment condition models. *Econometric Theory 33*(5), 1046–1080.

Andrews, I. (2018). Valid two-step identification-robust confidence sets for gmm. *Review of Economics and Statistics 100*(2), 337–348.

Berry, S., A. Gandhi, and P. Haile (2013). Connected substitutes and invertibility of demand. *Econometrica 81*(5), 2087–2111.

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

Berry, S., O. B. Linton, and A. Pakes (2004). Limit theorems for estimating the parameters of differentiated product demand systems. *The Review of Economic Studies 71*(3), 613–654.

Berry, S. T. (1994). Estimating discrete-choice models of product differentiation. *The RAND Journal of Economics*, 242–262.

Berry, S. T. and P. A. Haile (2014). Identification in differentiated products markets using market level data. *Econometrica 82*(5), 1749–1797.

Berry, S. T. and P. A. Haile (2021). Foundations of demand estimation. In *Handbook of industrial organization*, Volume 4, pp. 1–62. Elsevier.

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

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

Chen, X., T. M. Christensen, and E. Tamer (2018). Monte carlo confidence sets for identified sets. *Econometrica 86*(6), 1965–2018.

Conlon, C. and J. Gortmaker (2020). Best practices for differentiated products demand estimation with pyblp. *The RAND Journal of Economics 51*(4), 1108–1161.

Dufour, J.-M. (1997). Some impossibility theorems in econometrics with applications to structural and dynamic models. *Econometrica: Journal of the Econometric Society*, 1365–1387.

Dufour, J.-M. and M. Taamouti (2005). Projection-based statistical inference in linear structural models with possibly weak instruments. *Econometrica 73*(4), 1351–1365.

Dufour, J.-M. and M. Taamouti (2007). Further results on projection-based inference in iv regressions with weak, collinear or missing instruments. *Journal of Econometrics 139*(1), 133–153.

Gandhi, A. and J.-F. Houde (2019). Measuring substitution patterns in differentiated-products industries. *NBER Working paper* (w26375).

Horowitz, J. L. (1998). *Transformation Models*, pp. 141–178. New York, NY: Springer New York.

Keane, M. and T. Neal (2023). Instrument strength in iv estimation and inference: A guide to theory and practice. *Journal of Econometrics*.

Kleibergen, F. (2005). Testing parameters in gmm without assuming that they are identified. *Econometrica 73*(4), 1103–1123.

Nevo, A. (2000). Mergers with differentiated products: The case of the ready-to-eat cereal industry. *The RAND Journal of Economics*, 395–421.

Reynaert, M. and F. Verboven (2014). Improving the performance of random coefficients demand models: The role of optimal instruments. *Journal of Econometrics 179*(1), 83–98.

Staiger, D. and J. H. Stock (1997). Instrumental variables regression with weak instruments. *Econometrica 65*(3), 557–586.

Stock, J. and M. Yogo (2005). *Testing for Weak Instruments in Linear IV Regression*, pp. 80–108. New York: Cambridge University Press.

Stock, J. H. and J. H. Wright (2000). Gmm with weak identification. *Econometrica 68*(5), 1055–1096.

Sun, L. (2018). Implementing valid two-step identification-robust confidence sets for linear instrumental-variables models. *The Stata Journal 18*(4), 803–825.

# Appendix

# A Properties of quadric sets

**Proposition A.1** (Dufour and Taamouti, 2005)**.** *Suppose* $\mathcal{C} = \{\beta : \beta' A \beta + 2b'\beta + c \leq 0\}$ *with nonsingular* $A$*.*

1. *(Boundedness check)* $\mathcal{C}$ *is bounded iff* $A$ *is positive definite.*

2. *(Projection confidence interval) Let* $w \in \mathbb{R}^{\dim \beta} \setminus \{0\}$*, and* $\mathcal{C}_w$ *be the projection of* $\mathcal{C}$*:*

$$\mathcal{C}_w = \{w'\beta : \beta \in \mathcal{C}\}.$$

*Let* $d = b' A^{-1} b - c$ *and* $\widetilde{\beta} = -A^{-1} b$*.*

(a) *Case 1: if* $A$ *is positive definite and*

- $d \geq 0$*, then*

$$\mathcal{C}_w = \left[ w'\widetilde{\beta} - \sqrt{d(w'A^{-1}w)}, w'\widetilde{\beta} + \sqrt{d(w'A^{-1}w)} \right].$$

- $d < 0$*, then* $\mathcal{C}_w$ *is empty.*

(b) *Case 2: if* $A$ *has exactly one negative eigenvalue and*

- $w'A^{-1}w < 0$ *and* $d < 0$*, then*

$$\mathcal{C}_w = \left( -\infty, w'\widetilde{\beta} - \sqrt{d(w'A^{-1}w)} \right] \cup \left[ w'\widetilde{\beta} + \sqrt{d(w'A^{-1}w)}, \infty \right).$$

- $w'A^{-1}w = 0$ *and* $d < 0$*, then* $\mathcal{C}_w = \mathbb{R} \setminus \{w'\widetilde{\beta}\}$*.*
- *otherwise, then* $\mathcal{C}_w = \mathbb{R}$*.*

(c) *Case 3: if* $A$ *has more than one negative eigenvalues, then* $\mathcal{C}_w = \mathbb{R}$*.*