Detecting Identification Failure in Moment Condition Models

Jean-Jacques Forneron*

September 6, 2019

Abstract

This paper develops an approach to detect identification failures in a large class of moment condition models. This is achieved by introducing a *quasi-Jacobian* matrix which is asymptotically singular under higher-order local identification as well as weak/set identification; in these settings, standard asymptotics are not valid. Under (semi)-strong identification, where standard asymptotics are valid, this matrix is asymptotically equivalent to the usual Jacobian matrix. After re-scaling, it is thus asymptotically non-singular. Together, these results imply that the eigenvalues of the *quasi-Jacobian* can detect potential local and global identification failures. Furthermore, the *quasi-Jacobian* is informative about the span of the identification failure. This information permits two-step identification robust subvector inference without any *a priori* knowledge of the underlying identification structure. Monte-Carlo simulations and empirical applications illustrate the results.

JEL Classification: C11, C12, C13, C32, C36.
Keywords: GMM, robust subvector inference, higher-order/weak/set identification.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215 USA.
Email: jjmf@bu.edu, Website: http://jjforneron.com.

I would like to thank Serena Ng for discussions that initiated this project. I also greatly benefited from comments and discussions with Tim Christensen, Ivàn Fernàndez-Val, Hiro Kaido, Arthur Lewbel, Demian Pouzo, Zhongjun Qu and the participants of the BU-BC econometric workshop, the seminar participants at the Université de Montréal and University of Rochester and participants at conferences. Comments are welcome. All errors are my own.

1 Introduction

The Generalized Method of Moments (GMM) of Hansen & Singleton (1982) is a powerful estimation framework which does not require the model to be fully specified parametrically. Under regularity conditions, the estimates are consistent and asymptotically gaussian. In particular, the moment conditions should uniquely identify the finite dimensional parameters. This is very difficult to verify in practice and, as noted in Newey & McFadden (1994), is often assumed. Yet, when identification fails or nearly fails, the Central Limit Theorem provides a poor finite sample approximation for the distribution of the estimates. This has motivated a vast amount of research on tests which are robust to identification failure. As discussed in the literature review, much of this work has focused on tests for the full parameter vector. Potentially conservative confidence intervals for scalar parameters can then be built by projecting confidence sets for the full parameter vector (Dufour & Taamouti, 2005) or using a Bonferroni approach (McCloskey, 2017).

The contribution of this paper is two-fold: First, it introduces a *quasi-Jacobian* matrix which is singular under both local (first-order) and global identification failure and is informative about the coefficients involved in the failure. This is the main contribution of the paper as it provides an approach similar to Cragg & Donald (1993) and Stock & Yogo (2005) but in a non-linear setting. Second, the information from the first step allows for two-step identification robust subvector inference, akin to type I inference in Andrews & Cheng (2012) but without *a priori* knowledge of the identification structure.

To detect identification failures, this paper constructs a *quasi-Jacobian* matrix which corresponds to the best linear approximation of the sample moments function over a region of the parameters where these moments are close to zero. To find the best linear approximation, two loss functions are considered: the supremum norm measures the largest difference between the moments and its approximation while the least-squares criterion focuses on the average difference. The sup-norm approximation provides strong and intuitive results while least-squares can be easily computed by OLS using the moments as a dependent variable. Its lower computational burden and very simple implementation make the least-squares approximation the preferred approach for empirical settings.

The asymptotic behaviour of the *quasi-Jacobian* matrix, computed under these two loss functions, is studied under four identification regimes: strong, semi-strong,[1] higher-order local and weak (or set) identification. The GMM estimator is consistent and asymptotically

[1]Semi-strong identification is also known as nearly-weak identification (Antoine & Renault, 2009).

normal in the first two regimes, consistent but *not* asymptotically normal in the third and is inconsistent in the fourth. Hence, the last two regimes correspond to settings where the finite sample distribution of the estimator is poorly approximated by standard asymptotics. Under (semi)-strong identification,[2] the *quasi-Jacobian* is asymptotically equivalent to the usual Jacobian. After re-scaling, it is asymptotically non-singular. Under higher-order and weak/set set identification the *quasi-Jacobian* is asymptotically singular with eigenvalues vanishing on the span of the identification failure, i.e. directions in which identification fails.

Building on these results, this paper constructs a two-step procedure for testing linear hypotheses on the parameter $\theta \in \mathbb{R}^{d_\theta}$ of the form:

$$H_0 : R\theta = c \text{ vs. } H_1 : R\theta \neq c, \tag{1}$$

for some restriction matrix $R \in \mathbb{R}^{m \times d_\theta}$, $1 \leq m \leq d_\theta$, $c \in \mathbb{R}^m$. Assuming there is evidence of identification failure, i.e. a small eigenvalue in the *quasi-Jacobian*, the two steps can be summarized as:[3]

i. Split the parameter vector θ into two sets of parameters: one which needs to be fixed given evidence of weak, set or higher-order identification. $R\theta$ is also fixed to match the null (1). Another, for which there is no evidence of identification failure, will be treated as (semi)-strongly identified.

ii. Construct a confidence set by projection inference for $R\theta$ and the parameters fixed in i.; concentrate out the remaining parameters. The test statistic needs to be robust to identification failure. One can use the S, K or CQLR statistic of Stock & Wright (2000), Kleibergen (2005) and Andrews & Mikusheva (2016b), for instance.

Step 2 has previously been discussed in the literature.[4] The main challenge to implementing this step in practice has been in determining which nuisance parameters are (semi)-strongly identified when the others are fixed. When such decomposition is known *ex-ante*, Andrews & Cheng (2012) show how to conduct uniformly valid inference. In this paper, this knowledge is not required since the *quasi-Jacobian* is vanishing on the span of the identification failure. In practice, a cutoff is required to distinguish between matrices that are vanishing from those that are not. A rule-of-thumb, similar to Stock & Yogo (2005), is provided to construct this

[2]The term (semi)-strong will refer to cases where identification can be either strong or semi-strong.

[3]Under strong and semi-strong identification, standard inference using the Wald, QLR or LM test will be valid. Lack of evidence for weak and higher-order identification would indicate that these tests can be used.

[4]See e.g. Kleibergen (2005); Andrews & Mikusheva (2016b), among others.

cutoff when detecting weak/set as well as higher-order identification. It relies on a Nagar approximation of the size distortion under semi-strong asymptotics.

To determine which coefficients to fix, a search procedure considers a pre-determined increasing set of restrictions. The search stops once a criteria based on the quasi-Jacobian (and residual curvature for higher-order identification) indicates that the span of the identification failure is fixed. The search procedure is shown to restore point identification with probability going to 1. As a result, the two-step approach is shown to yield tests that are asymptotically valid, although not uniformly.

Finally, the quasi-Jacobian can be used to compute standard errors in the sandwich formula when sample moments are non-smooth. This may be of practical interest.

Monte-Carlo simulations illustrate the finite sample behaviour of the *quasi-Jacobian* and the two-step inference procedure. The approach is then applied to two empirical settings. The first revisits the US Euler equation: the quasi-Jacobian is flat in at least one direction confirming weakly/set identification. The root of the issue is that moments are redundant, reducing to a single moment condition. Singulary and identification robust inference is required (Andrews & Guggenberger, 2019). The second application, provided in the Supplement, confirms weak identification in quantile IV estimation of the demand for fish of Chernozhukov et al. (2007).

Structure of the Paper

After a review of the literature and an overview of the notation used in the paper, Section 2 introduces the setting, the linear approximations, precise definitions of the identification regimes considered and the main assumptions used in the paper. Section 3 derives the asymptotic behaviour of the *quasi-Jacobian* matrix. Section 4 describes the two-step inference procedures in more details including: the Algorithms used to determine which parameters to fix, the rules-of-thumb for choosing the cutoffs and the asymptotic results for the inference procedures. Section 5 provides a Monte-Carlo example to illustrate some of the results from the previous sections. An empirical example is provided in Section 6. Section 7 concludes. Appendices A and B provide the proofs for the main results of Sections 3 and 4 respectively. The Supplement consists of Appendices C, D, E, F, G and H which provide additional and preliminary results for the main text and their proofs as well as additional Monte-Carlo and Empirical results.

Related Literature

The literature on the identification of economic models is quite vast. An extensive review is given in Lewbel (2018). Within this literature, this paper mainly relates to three topics: local and global identification of finite dimensional parameters in the population, detecting identification failure in finite samples and identification robust inference.

Koopmans & Reiersol (1950) provide one of the earliest general formulations of the identification problem at the population level. To paraphrase the authors, the main problem is to determine whether the distribution of the data, assumed to be generated from a given class of models, is consistent with a unique set of structural parameters. In the likelihood setting, Fisher (1967); Rothenberg (1971) introduced sufficient conditions for local and global identification. For GMM, Komunjer (2012) introduced weaker global identification conditions.

In linear models, global identification amounts to a rank condition on the slope of the moments. This insight was used in pre-testing linear IV models for identification failure (Cragg & Donald, 1993; Stock & Yogo, 2005). Pre-tests based on the null of strong identification appear in Hahn & Hausman (2002) for linear IV and Inoue & Rossi (2011); Bravo et al. (2012) for non-linear models. Pre-testing for strong identification can be problematic for size control when the pre-test's power is low. For non-linear models, Wright (2003) uses a rank test and Antoine & Renault (2017) a distorted J-statistic to detect local identification failure. Arellano et al. (2012) develop a test for underidentification of a single coefficient.

Given the impact of (near) identification failure on standard inferences, a large body of literature has developed identification robust tests. Most consider inference for the full parameter vector.[5] Few consider the topological features of the identified set, with the notable exception of Andrews & Mikusheva (2016a). For subvector inferences, a common approach is to construct a confidence set for the full vector and project it on the dimension of interest (Dufour & Taamouti, 2005, 2007) or to use a Bonferroni correction (McCloskey, 2017). These methods might be conservative.[6] A series of papers starting with Andrews & Cheng (2012), considers uniformly valid subvector inferences in a class of model where the identification structure is known and identification strength is driven by some (semi)-strongly identified coefficients. Under higher-order identification, estimates are consistent but the delta-method is not valid; the limiting distribution is non-standard (Rotnitzky et al., 2000; Dovonon & Hall, 2018). This issue is known but much less studied than weak and set

[5] See e.g. Anderson & Rubin (1949); Stock & Wright (2000); Kleibergen (2005); Andrews & Mikusheva (2016b); Chen et al. (2018).

[6] However, as discussed in Section 4, Remark 2, when the nuisance parameters are completely unidentified projection inference may actually have exact asymptotic coverage.

identifications. Dovonon et al. (2019) study identification robust tests under second-order identification and Lee & Liao (2018) show how to conduct standard inference with known second-order identification structure.

Notation

For any matrix (or vector) A, $\|A\| = \sqrt{\sum_{i,j} A_{i,j}^2} = \sqrt{\text{trace}(AA')}$ is the Frobenius (Euclidian) norm of A. For any rectangular matrix A, the singular value $|\lambda_j(A)|$ refers to the $j-th$ eigenvalue of $(A'A)^{1/2}$. $\lambda_{\max}(A), \lambda_{\min}(A)$ refer to the largest and smallest value of $|\lambda_j(A)|$, respectively. With some abuse of notation, these singular values will be referred to as eigenvalues. For a weighting matrix $W_n(\theta)$, the norm $\|\bar{g}_n(\theta)\|_{W_n}^2$ is computed as $\bar{g}_n(\theta)'W_n(\theta)\bar{g}_n(\theta)$. For any two positive sequences a_n, b_n, $a_n \asymp b_n \Leftrightarrow \exists \underline{C}, \overline{C} > 0, \underline{C}a_n \leq b_n \leq \overline{C}a_n, \forall n \geq 1$; $a_n = o(b_n) \Leftrightarrow \forall \varepsilon > 0, \exists N > 0, \forall n \geq n, a_n \leq \varepsilon b_n$; $a_n = O(b_n) \Leftrightarrow \exists M > 0, \exists N > 0, \forall n \geq n, a_n \leq M b_n$. For X_n a sequence of random variables and a_n positive sequence, $X_n = o_p(a_n) \Leftrightarrow \forall \varepsilon > 0, \mathbb{P}(\|X_n\| \geq a_n \varepsilon) = o(1)$; $X_n = O_p(a_n) \Leftrightarrow \forall \varepsilon > 0, \exists M > 0, \exists N > 0, \forall n \geq N, \mathbb{P}(\|X_n\| > a_n M) \leq \varepsilon$.

2 Setting and Assumptions

Following Hansen & Singleton (1982), the econometrician wants to estimate the solution vector θ_0 to the system of unconditional moment equations:

$$g_n(\theta_0) \overset{def}{=} \mathbb{E}(\bar{g}_n(\theta_0)) = 0, \tag{2}$$

where $\theta_0 \in \Theta$, a compact subset of $\mathbb{R}^{d_\theta}$, $\dim(g_n) = p \geq d_\theta$. $\bar{g}_n(\theta) = 1/n \sum_{i=1}^n g(z_i, \theta)$, $(z_i)_{i=1,\ldots,n}$ is a sample of iid or stationary random variables. Throughout, it is assumed that at least one such θ_0 exists.[7] g_n is assumed to be continuously differentiable on Θ. Given the sample moments $\bar{g}_n$ and a sequence of positive definite weighting matrices $W_n(\theta)$, the GMM estimator $\hat{\theta}_n$ solves the minimization problem:

$$\hat{\theta}_n = \text{argmin}_{\theta \in \Theta} \|\bar{g}_n(\theta)\|_{W_n}^2, \text{ where } \|\bar{g}_n(\theta)\|_{W_n}^2 = \bar{g}_n(\theta)'W_n(\theta)\bar{g}_n(\theta).$$

2.1 Linear Approximations and the quasi-Jacobian Matrix

The *quasi-Jacobian* matrix $B_{n,LS/\infty}$ is defined below as the slope of a local linear approximation $\bar{g}_n$ under a given loss.

[7] This can be achieved in general by re-centering: $g_n(\theta) - g_n(\theta_0)$ where $\theta_0 = \text{argmin}_{\theta \in \Theta} \|g_n(\theta)\|_W$.

Definition 1. *(Sup-Norm and Least-Squares Approximations) Let K be a kernel function and κ_n a bandwidth. The sup-norm approximation $(A_{n,\infty}, B_{n,\infty})$ solves:*

$$(A_{n,\infty}, B_{n,\infty}) = argmin_{A,B} \sup_{\theta \in \Theta} \|A + B\theta - \bar{g}_n(\theta)\| \times \hat{K}_n(\theta), \tag{3}$$

where $\hat{K}_n(\theta) = K\left(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n\right)$. The least-squares approximation $(A_{n,LS}, B_{n,LS})$ solves:

$$(A_{n,LS}, B_{n,LS}) = argmin_{A,B} \int_\Theta \|A + B\theta - \bar{g}_n(\theta)\|^2 \times \hat{K}_n(\theta) d\theta, \tag{4}$$

where $\hat{K}_n(\theta) = K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{W_n}/\kappa_n)$. The quasi-Jacobian refers to the $B_{n,LS/\infty}$ computed using either the least-squares (LS) or sup-norm (∞) approximation.

The sup-norm approximation solves a non-smooth optimization problem and is thus more computationally demanding. However, the theory for $B_{n,\infty}$ is very intuitive and it will be quite useful to understand the relation between the *quasi-Jacobian* and identification failure. The least-squares approximation involves further topological arguments and requirements but is much more convenient to compute in practice:

$$\left(\begin{array}{cc} A_{n,LS}, & B_{n,LS} \end{array} \right)' = \left(\int_\Theta X(\theta) X(\theta)' \hat{K}_n(\theta) d\theta \right)^{-1} \int_\Theta X(\theta) \bar{g}_n(\theta)' \hat{K}_n(\theta) d\theta, X(\theta) = (1, \theta')'.$$

The two integrals can be approximated using (quasi)-Monte-Carlo (qMC) methods (Robert & Casella, 2004; Lemieux, 2009). In this paper, the Sobol sequence was used. Implementation is straightforward: qMC provides a grid for θ over which $\bar{g}_n$ and W_n are evaluated. The evaluated moments are then regressed on $(1, \theta)$ using weighted least-squares with $\hat{K}_n$ as weights. The quasi-Jacobian $B_{n,LS}$ then consists of the slope coefficients.

For OLS and IV, the approximation is exact and yields $B_{n,LS} = X'X/n$ and $Z'X/n$. The quasi-Jacobian is close to singular where the regressors are nearly multicollinear in OLS or when the instruments are not sufficiently relevant in IV. The rank of $B_{n,LS}$ is thus informative about the identification failure in these models. This extends to non-linear models.

2.2 Identification Regimes

The following describes the identification regimes considered in this paper. Their implications for the GMM estimator $\hat{\theta}_n$ are summarized in Table 1. Examples 1, 2 illustrate the definitions.

Example 1 (Non-Linear Least-Squares)**.** *Consider the non-linear regression model:*

$$y_t = \theta_1 x_{1,t} + \theta_1 \theta_2 x_{2,t} + e_t$$

Table 1: Identification Regimes and Asymptotic Properties of $\hat{\theta}_n$

Identification Regime	$\hat{\theta}_n$ consistent?	Rate of convergence	Limiting distribution
Strong	Yes	$\sqrt{n}$	Gaussian
Semi-Strong	Yes	slower than $\sqrt{n}$	Gaussian
Higher-Order	Yes	$n^{1/4}$ or slower	non-Gaussian
Weak or Set	No	-	non-Gaussian

with $x_{1,t}, x_{2,t}, e_t$ *iid with mean* 0 *and variance* 1 *such that* $\mathbb{V}(x_{1,t}, x_{2,t}, e_t) = I_3$ *and* $\theta = (\theta_1, \theta_2) \in [0,1]^2$*. The estimating moments* g_n *are:*

$$g_n(\theta) = \begin{pmatrix} \mathbb{E}(y_t x_{1,t}) - \theta_1 \\ \mathbb{E}(y_t x_{2,t}) - \theta_1\theta_2 \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ \theta_2 & \theta_{1,0} \end{pmatrix} \begin{pmatrix} \theta_1 - \theta_{1,0} \\ \theta_2 - \theta_{2,0} \end{pmatrix}.$$

Example 2 (Possibly Noninvertible MA(1) Model)**.** *Consider the model:*

$$y_t = \sigma[e_t - \vartheta e_{t-1}]$$

where e_t *is iid with mean* 0*, variance* 1 *and skewness* τ *known. The moments* $\mathbb{E}(y_t^2)$ *and* $\mathbb{E}(y_t y_{t-1})$ *only determine* $\theta \in \{(\vartheta_0, \sigma_0^2), (1/\vartheta_0, \sigma_0^2\vartheta_0^2)\}$ *when* $\vartheta_0 \in \mathbb{R}/\{-1, 0, 1\}$*. Assuming invertibility (*$|\vartheta_0| \leq 1$*) restores point identification. Gospodinov & Ng (2015) show that when* $\tau \neq 0$*, the additional moment* $\mathbb{E}(y_t^2 y_{t-1})$ *restores point identification without invertibility.*

Definition 2. *(Point Identification) The model is point identified if* $\exists \theta_0 \in int(\Theta)$ *such that* $\forall \varepsilon > 0$*,* $\exists \eta(\varepsilon) > 0$*:*

$$\inf_{\|\theta - \theta_0\| \geq \varepsilon} \|g_n(\theta)\|_W \geq \eta(\varepsilon), \forall n \geq 1. \tag{5}$$

Definition 3. *(Strong Identification) The model is strongly identified if it is point identified and* $\exists \varepsilon > 0$ *and* $\underline{C} > 0$ *such that* $\|\theta - \theta_0\| \leq \varepsilon$ *implies:*

$$\|g_n(\theta)\|_W \geq \underline{C}\|\theta - \theta_0\|, \forall n \geq 1. \tag{6}$$

Definition 3 is satisfied when the Jacobian $\partial_\theta g_n(\theta_0)$ has full rank, its smallest eigenvalue is bounded below and, $g_n(\theta) = \partial_\theta g_n(\theta_0)(\theta - \theta_0) + o(\|\theta - \theta_0\|)$ around θ_0.

Example 1 (Continued). *Computing the Jacobian of the moments at $\theta = \theta_0$ implies:*

$$\partial_\theta g_n(\theta_{0,n}) = \begin{pmatrix} 1 & 0 \\ 0 & \theta_{1,0} \end{pmatrix} \Rightarrow \begin{pmatrix} 1 & 0 \\ -\theta_2 & 1 \end{pmatrix} g_n(\theta) = \partial_\theta g_n(\theta_{0,n})(\theta - \theta_0).$$

Note that 1 is the only eigenvalue of the matrix on the left-hand side of $g_n(\theta)$ which implies that $\|g_n(\theta)\| \geq \|\partial_\theta g_n(\theta_0)(\theta - \theta_0)\|$. $|\theta_{1,0}|$ bounded away from zero implies that the eigenvalues of $\partial_\theta g_n(\theta_0)$ are bounded away from zero as well.

Example 2 (Continued). *The estimating moments are given by:*

$$g_n(\theta) = \begin{pmatrix} \mathbb{E}(y_t^2) & - & \sigma^2(1+\vartheta^2) \\ \mathbb{E}(y_t y_{t-1}) & + & \sigma^2\vartheta \\ \mathbb{E}(y_t^2 y_{t-1}) & + & \tau\sigma^3\vartheta \end{pmatrix} = \begin{pmatrix} \sigma_0^2(1+\vartheta_0^2) & - & \sigma^2(1+\vartheta^2) \\ -\sigma_0^2\vartheta_0 & + & \sigma^2\vartheta \\ -\tau\sigma_0^3\vartheta_0 & + & \tau\sigma^3\vartheta \end{pmatrix}.$$

If $\tau \neq 0$ is bounded away from 0 and $\sigma \neq 0$ fixed, $\vartheta \notin \{-1, 0, 1\}$. Point identification holds since: $\tau\sigma_0^3\vartheta_0 \neq sign(\vartheta_0)\tau\sigma_0^3\vartheta_0^2$ unless $\sigma = 0$, $\tau = 0$ or $\vartheta \in \{-1, 0, 1\}$. The eigenvalues of the Jacobian are bounded below when $|\tau|$ is bounded away from zero.

Definition 4. *(Semi-Strong Identification) The model is semi-strongly identified if it is point identified and*

i. $\exists \varepsilon > 0$, $\overline{C}, \underline{C} > 0$ *such that* $\|\theta - \theta_0\| \leq \varepsilon$ *implies:*

$$\overline{C}\|\partial_\theta g_n(\theta_0)(\theta - \theta_0)\| \geq \|g_n(\theta)\|_W \geq \underline{C}\|\partial_\theta g_n(\theta_0)(\theta - \theta_0)\|, \forall n \geq 1 \tag{7}$$

ii. $\lim_{n\to\infty} n \times \lambda_{\min}\left(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right) = +\infty$,

iii. $\|g_n(\theta_1) - g_n(\theta_2) - \partial_\theta g_n(\theta_2)(\theta_1 - \theta_2)\| = O(\|\partial_\theta g_n(\theta_2)(\theta_1 - \theta_2)\|^2)$,
if $\|\theta_1 - \theta_0\| + \|\theta_2 - \theta_0\| = o(1)$,

iv. $\left[\partial_\theta g_n(\theta_1) - \partial_\theta g_n(\theta_0)\right]\left[\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right]^{-1}\left[\partial_\theta g_n(\theta_1) - \partial_\theta g_n(\theta_0)\right]' = o(1)$, *if* $\theta_1 - \theta_0 = o(1)$

Definition 4 ii. implies that the Jacobian can be vanishing in one or several directions - but not too fast. When $\lambda_{\min}\left(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right) \leq O(n^{-1/4})$, conditions iii.-iv. also imply that the second-order term is vanishing. As a result, the moments remain approximately linear around θ_0, as in Definition 3. After re-scaling, $(\hat{\theta}_n - \theta_0)$ is asymptotically Gaussian; convergence is slower than the usual $\sqrt{n}$-rate (Antoine & Renault, 2009).

Example 1 (Continued). *Consider the drifting sequence* $\theta_{1,0,n} = c \times n^{-a}$ *with* $a \in [0, 1/2)$ *and* $c \neq 0$*:* $n \times \lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0)) = c^2 n^{1-2a} \to +\infty$ *if* $c \neq 0$ *and* $0 \leq a < 1/2$*.*

Definition 5. *(Higher-Order Local Identification) The model is locally identified at a higher order* $r \geq 2$ *if it is point identified and* $\exists \varepsilon > 0$*,* $\overline{C}_j > 0, \underline{C}_j > 0$ *for* $j = 1, \ldots, r$ *and projection matrices* $P_1, \ldots, P_r$ *satisfying* $P_r \neq 0$*,* $P_j P_{j'} = 0$ *when* $j \neq j'$ *such that* $\|\theta - \theta_0\| \leq \varepsilon$ *implies:*

$$\sum_{j=1}^{r} \overline{C}_j \|P_j(\theta - \theta_0)\|^j \geq \|g_n(\theta)\|_W \geq \sum_{j=1}^{r} \underline{C}_j \|P_j(\theta - \theta_0)\|^j, \, \forall n \geq 1. \tag{8}$$

Definition 5 implies the moments are not approximately linear around θ_0. As a result, the higher-order terms affect the limiting distribution of the estimator and $(\hat{\theta}_n - \theta_0)$ converges at a $n^{1/2r}$-rate to a non-Gaussian limiting distribution (Dovonon & Hall, 2018).

Example 2 (Continued). *Suppose that* $\kappa = 0$ *and* $\vartheta_0 = 1$*. Condition iii.a. holds since there is a unique solution and the moments are continuous. Omitting the third moment:*

$$\partial_\theta g_n(\theta_0) = \begin{pmatrix} -2\sigma_0^2 & -2 \\ \sigma_0^2 & 1 \end{pmatrix}, \partial^2_{\theta,\vartheta} g_n(\theta_0) = \begin{pmatrix} -2\sigma_0^2 & 0 \\ 1 & 0 \end{pmatrix}, \partial^2_{\theta,\sigma^2} g_n(\theta_0) = \begin{pmatrix} -2 & 0 \\ 1 & 0 \end{pmatrix}.$$

$(1, -1)$ *is the eigenvector which spans the null space of the Jacobian, and both second-order derivatives are non-singular on the span of* $(1, -1)$ *which implies second-order identification (Dovonon & Hall, 2018).*

Definition 6. *(Weak and Set Identification) The model is said to be weakly or set identified if there exists at least two* $\theta_0 \neq \theta_1$ *in the weakly identified set:*

$$\Theta_0 = \{\theta \in \Theta, \lim_{n \to \infty} \sqrt{n} \|g_n(\theta)\|_W < +\infty\}. \tag{9}$$

Definition 6 occurs when global identification fails or nearly fails. Under strong, semi-strong and higher-order identification, a robust and conservative confidence set would concentrate around a single point θ_0. Definition 6 collects all models where this does not occur. $\hat{\theta}_n$ is not consistent (Stock & Wright, 2000) and has non-standard limiting distribution.

Example 1 (Continued). *Consider the sequence* $\theta_{1,0,n} = c \times n^{-1/2}$*. Take* $\theta = (\theta_{1,0,n}, \theta_2)$*,* $\theta_2 \in [0, 1]$*, then* $\sqrt{n}\|g_n(\theta)\| \to |c| \times |\theta_2 - \theta_0| < +\infty$*. As a result* $\Theta_0 \supseteq \{\theta = (0, \theta_2), \theta_2 \in [0, 1]\}$*.*

Example 2 (Continued). *Consider the sequence* $\tau_n = c/\sqrt{n}$*, then:* $g_n(1/\vartheta_0, \sigma_0^2 \vartheta_0^2) = \left(0, \;\; 0, \;\; c\sigma_0^3/\sqrt{n}[sign(\vartheta_0)\vartheta_0^2 - \vartheta_0] \right)'$*. This implies* $\sqrt{n}\|g_n(1/\vartheta_0, \sigma_0^2\vartheta_0^2)\| \to |c\sigma_0^3[sign(\vartheta_0)\vartheta_0^2 - \vartheta_0]| < +\infty$*. As a result,* Θ_0 *is not a singleton when* $\vartheta_0 \notin \{-1, 0, 1\}$ *and* $\tau_n = O(n^{-1/2})$*.*

2.3 Main Assumptions

The following provides the main assumptions on the moments $\bar{g}_n$, weighting matrix W_n, kernel K and bandwidth κ_n to derive the results in Section 3 for $B_{n,LS/\infty}$.

Assumption 1. *(Bandwidth, Kernel)*

i. (Bandwidth) $\kappa_n > 0, \forall n \geq 1$. $\kappa_n \to 0$, $\sqrt{n}\kappa_n \to +\infty$ *and* $\sqrt{n}\kappa_n^2 \to 0$ *as* $n \to \infty$,

ii. (Compact Kernel) K *is Lipschitz-continuous on* $\mathbb{R}$ *with* $K(x) = 0$ *for* $x \in (-\infty, -1] \cap [1, +\infty)$, $K(x) > 0$ *for* $x \in (-1, 1)$,

iii. (Exponential Kernel) K *is exponential in* x, *i.e.* $\exists a \geq 1, C_1 > 0, C_2 > 0$ *such that* $K(x) = C_1 \exp(-C_2|x|^a)$, $\forall x \in \mathbb{R}$. *Define* $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a}$ *and assume* $\tilde{\kappa}_n \to 0$, $\sqrt{n}\tilde{\kappa}_n^2 \to 0$ *as* $n \to \infty$.

Condition i. ensures that the bandwidth converge to 0 at a slower than $\sqrt{n}$-rate, but faster than a $n^{1/4}$-rate. When $\kappa_n \leq O(n^{1/4})$, $B_{n,LS/\infty}$ would capture second-order non-linearities under (semi)-strong identification. When $W_n = \hat{V}_n^{-1}$, a Law of the Iterated Logarithm can be invoked to set $\kappa_n = \sqrt{2\log(\log[n])/n}$.[8] Two types of kernels K are considered. Compact kernels (condition i.), are used in both sup-norm and least-squares approximations. The Lipschitz-continuity condition simplifies the proofs, but there was almost no numerical difference with the uniform kernel $K(x) = \mathbb{1}_{x\in(-1,1)}$. Exponential kernels, e.g. $K(x) = \phi(x)$ the Gaussian density, are considered only for $B_{n,LS}$.

Assumption 2. *(Sample Moments, Weighting Matrix)*

i. (Uniform CLT, Tightness) the empirical process $\mathbb{G}_n(\theta) \overset{def}{=} \sqrt{n}\left(\bar{g}_n(\theta) - g_n(\theta)\right)$ *converges weakly to* $\mathbb{G}(\cdot)$ *a Gaussian process, as* $n \to \infty$*;* $\sup_{\theta\in\Theta} \|\mathbb{G}_n(\theta)\| = O_p(1)$,

ii. (Discoverability of Θ_0*) the weakly identified set* $\Theta_0 = \{\theta \in \Theta, \lim_{n\to\infty} \sqrt{n}\|\bar{g}_n(\theta)\|_W < +\infty\}$ *satisfies:* $\sup_{n\geq 1} \sup_{\theta\in\Theta_0} \sqrt{n}\|g_n(\theta)\|_W < +\infty$,

iii. (Stochastic Equicontinuity) $\sqrt{n}\left[\bar{g}_n(\theta_1) - \bar{g}_n(\theta_2) - (g_n(\theta_1) - g_n(\theta_2))\right] = o_p(1)$, *uniformly in* $\|\theta_1 - \theta_2\| = o(1)$,

iv. (Smoothness) g_n *is continuously differentiable on* Θ*; uniformly in* $\|\theta_1 - \theta_2\| = o(1)$, $\|g_n(\theta_1) - g_n(\theta_2) - \partial_\theta g_n(\theta_2)(\theta_1 - \theta_2)\| = O(\|\theta_1 - \theta_2\|^2)$,

[8] See Andrews & Cheng (2012) for choices of such sequences. In finite samples, one may prefer $\kappa_n = \max(q_{1-\varepsilon}, \sqrt{2\log(\log[n])/n})$ where $q_{1-\varepsilon}$ is a $1-\varepsilon$ (e.g. 0.99) quantile of a $\chi^2_{\dim(g_n)}$ distribution.

v. (Weighting Matrix) $\sup_{\theta\in\Theta}\|W_n(\theta)-W(\theta)\| = o_p(1)$, W *is Lipschitz continuous in* $\theta\in\Theta$, $\exists\underline{\lambda},\overline{\lambda}$, $0<\underline{\lambda}\leq\lambda_{\min}(W_n(\theta))\leq\lambda_{\max}(W_n(\theta))\leq\overline{\lambda}<+\infty, \forall n\geq 1$, $\theta\in\Theta$.

The high-level conditions in Assumption 2 are quite common in GMM estimation. Condition i. allows for non-smooth or discontinuous sample moments. Conditions i. and ii. ensure that the weakly identified set Θ_0 can be conservatively estimated using $\hat{\Theta}_n = \{\theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_{\theta\in\Theta}\|\bar{g}_n(\theta)\|_{W_n}\leq\kappa_n\}$ so that all directions of the identification failure can be detected.[9] Conditions iii. is the usual stochastic equicontinuity condition. Condition iv. is only required under strong identification. Condition v. is automatically satisfied for $W_n = I_p$, the identity matrix. For the optimal weighting matrix $W_n = \hat{V}_n^{-1}$, it requires uniform consistency of $\hat{V}_n$ and additional conditions on the eigenvalues. Given the generality of the high-level assumptions, the results accommodate models where a (semi)-strongly identified nuisance parameter η is concentrated out: $\hat{\theta}_n = \text{argmin}_{\theta\in\Theta}\|\bar{g}_n(\theta,\hat{\eta}(\theta))\|_{W_n}$.

3 Asymptotic Behaviour of the Linear Approximations

This section derives the asymptotic behaviour of $(A_{n,LS/\infty}, B_{n,LS/\infty})$ under (semi)-strong identification and $B_{n,LS/\infty}$ under higher-order and weak/set identification. Table 2 summarizes the results. At the population level, the results imply (by taking $\bar{g}_n = g_n$ and $\kappa_n \searrow 0$) that the *quasi-Jacobian* is the usual Jacobian for first-order globally identified models and is singular under either local or global identification failure. This provides a simple characterization of first-order and global identification failure for GMM in the population.

The relevant metric for the sup-norm approximation is the Euclidian distance $\|\theta_0-\theta_1\|$ whereas the least-squares approximation relies on a quadratic loss. This implies another metric which relies on the measure defined below.

Definition 7. *(Quasi-Posterior $\hat{\pi}_n$) For any $\theta\in\Theta$, the quasi-posterior $\hat{\pi}_n$ associated with the sample moments $\bar{g}_n$ and the kernel K is:*

$$\hat{\pi}_n(\theta) = \frac{K(\|\bar{g}_n(\theta)/\kappa_n\|_{W_n} - \|\bar{g}_n(\hat{\theta}_n)/\kappa_n\|_{W_n})}{\int_\Theta K(\|\bar{g}_n(\tilde{\theta})/\kappa_n\|_{W_n} - \|\bar{g}_n(\hat{\theta}_n)/\kappa_n\|_{W_n})d\tilde{\theta}}.$$

The associated posterior mean $\bar{\theta}_n$ and variance Σ_n are:

$$\bar{\theta}_n = \int\theta\hat{\pi}_n(\theta)d\theta, \quad \Sigma_n = \int(\theta-\bar{\theta}_n)(\theta-\bar{\theta}_n)'\hat{\pi}_n(\theta)d\theta.$$

[9]For the exponential kernel, $\hat{\Theta}_n = \{\theta, \|\bar{g}_n(\theta)\|_{W_n}\leq\kappa_n\}$ can be used instead: $K(\|\bar{g}_n(\hat{\theta}_n)\|_{W_n})$ is a multiplicative constant in the numerator and denominator of the OLS solution of $B_{n,LS}$ and cancels out.

Table 2: Summary of the Results in Section 3

Identification Regime	Asymptotics for $\hat{\theta}_n$	Asymptotics for $B_{n,LS/\infty}$
(Semi)-Strong	Gaussian	$B_{n,LS/\infty} \simeq$ Jacobian
Higher-Order	Non-Gaussian	$B_{n,LS/\infty} v_j \asymp \text{bandwidth}^{1-1/j}$
Weak or Set	Non-Gaussian	$B_{n,LS/\infty} v \asymp \text{bandwidth}$

Note: The results in the two bottom rows hold over all directions $v = \theta_0 - \theta_1$ in which the model is weakly identified, i.e. $\theta_0, \theta_1 \in \Theta_0$ in Definition 6; and all directions v_j in which the moments are locally polynomial of order $j \geq 2$ in Definition 5.

When $K = \phi$, the Gaussian density, $\hat{\pi}_n$ corresponds to the quasi-posterior of Chernozhukov & Hong (2003), up to a prior π, for $W_n = \hat{V}_n^{-1}$ the optimal weights, and $\kappa_n = n^{-1/2}$. It is also relates to the ABC estimator when $\bar{g}_n$ are simulated moments (Marin et al., 2012).

Σ_n behaves like a sufficient statistic for identification failure under the least-squares loss. If the model is set identified, one would expect Σ_n to be non-zero in some direction; under (semi)-strong identification, a Bernstein-von Mises type result implies $\Sigma_n = o_p(1)$.

Lemma 1. *(Relationship between $B_{n,LS}$ and Σ_n)*
Suppose Assumptions 1 and 2 hold. For the exponential kernel, assume that the moments satisfy a Hölder-type condition around Θ_0: $\exists \varepsilon > 0$, $\overline{C} > 0$ and $\varsigma \in (0, 1]$ such that $\|g_n(\theta)\|_W \leq \overline{C} d(\theta, \Theta_0)^{\varsigma}$, $\forall n \geq 1$, $\forall \theta \in \Theta$ such that $d(\theta, \Theta_0) \leq \varepsilon$. Given these assumptions, the least-squares approximation, and the quasi-posterior variance satisfy:

$$0 \leq \textit{trace}\left(B_{n,LS} \Sigma_n B_{n,LS}'\right) \leq O_p(\tilde{\kappa}_n^2), \tag{10}$$

where $\tilde{\kappa}_n = \kappa_n$ for the compact kernel and $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a}$ for the exponential kernel. This implies the following inequalities:

i. $\forall j \in \{1, \ldots, d_\theta\}$, $0 \leq \lambda_j\left(B_{n,LS}' B_{n,LS}\right) \times \lambda_{d_\theta + 1 - j}\left(\Sigma_n\right) \leq O_p(\tilde{\kappa}_n^2)$ *where the eigenvalues λ_j are in increasing order. In particular:* $0 \leq \lambda_{\min}\left(B_{n,LS}' B_{n,LS}\right) \times \lambda_{\max}\left(\Sigma_n\right) \leq O_p(\tilde{\kappa}_n^2)$,

ii. let $(v_{j,n})_{j=1,\ldots,d_\theta}$ be the eigenvectors of Σ_n, suppose that for each j there exists $(r_{j,n})_{n \geq 1}$ such that $v_{j,n}' \Sigma_n v_{j,n} = O_p(r_{j,n}^2)$, then for each $j \in \{1, \ldots, d_\theta\}$: $B_{n,LS} v_{j,n} = O_p\left(\tilde{\kappa}_n / r_{j,n}\right)$,

iii. assuming the $r_{j,n}, j = 1, \ldots, d_\theta$ are in increasing order then, without loss of generality: $|\lambda_{\min}(B_{n,LS})| = O_p(\tilde{\kappa}_n / r_{d_\theta, n})$.

Lemma 1 is pivotal in deriving the results for higher-order and weak/set identification. Equation (10) implies that the behaviour of $B_{n,LS}$ is determined by the behaviour of Σ_n. This implies a similar relation for their eigenvalues. In turn, determining the rate for $\lambda_{\min}(B_{n,LS})$ is equivalent to solving for $\lambda_{\max}(\Sigma_n)$. Under the sup-norm, the diameter of the set $\{\theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_\theta \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n\}$ determines the behaviour of $\lambda_{\min}(B_{n,\infty})$. Here, $\lambda_{\max}(\Sigma_n)$ acts as the quasi-posterior measure of the diameter of this set.

For the exponential kernel, an additional condition is required where the moments must not grow too fast when θ is away from Θ_0. The posterior may not concentrate only on Θ_0 if this assumption does not hold giving non-negligible weight to other regions of Θ.

3.1 (Semi)-Strong Identification

Theorem 1. *(Approximations under (Semi)-Strong Identification) Suppose that the model is (semi)-strongly identified, Assumptions 1 and 2 hold, that the bandwidth κ_n and moments g_n are such that $\tilde{\kappa}_n^2 = o\left(\lambda_{\min}\left(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right)\right)$, where $\tilde{\kappa}_n = \kappa_n$ for the compact kernel and $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a}$ otherwise, then for $H_n = \left[\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right]^{-1/2}$, the sup-norm and least-squares approximations satisfy:*

$$A_{n,LS/\infty} = \bar{g}_n(\hat{\theta}_n) - B_{n,LS/\infty}\hat{\theta}_n + o_p(n^{-1/2}), \quad B_{n,LS/\infty}H_n = \partial_\theta g_n(\hat{\theta}_n)H_n + o_p(n^{-1/2}\tilde{\kappa}_n^{-1}),$$

where $\hat{\theta}_n = argmin_{\theta\in\Theta}\|\bar{g}_n(\theta)\|_{W_n}$. Let $\hat{\theta}_{n,LS/\infty} = -\left(B'_{n,LS/\infty}\hat{W}_n B_{n,LS/\infty}\right)^{-1} B'_{n,LS/\infty}\hat{W}_n A_{n,LS/\infty}$ be the estimator associated with the least-squares approximation and $\hat{W}_n = W_n(\hat{\theta}_n)$, then $H_n^{-1}(\hat{\theta}_{n,LS/\infty} - \hat{\theta}_n) = o_p(n^{-1/2})$. Also, $H_n^{-1}\Sigma_n H_n^{-1} = O_p(\tilde{\kappa}_n^2)$.

Under semi-strong identification, $\sqrt{n}H_n^{-1}(\hat{\theta}_n - \theta_0)$ is asymptotically Gaussian. The rate of convergence of each coefficient depends on the eigenvalues of H_n^{-1} and its eigenvectors.[10] In practice, the standard errors adjust for the rate of convergence automatically, so that standard Wald inferences are valid. The scaled convergence of $B_{n,LS/\infty}$ in Theorem 1 implies convergence of the spectral decomposition. Let $v_{j,n}$ be the jth right singular vector of $\partial_\theta g_n(\theta_0)$[11] with singular value $\lambda_{j,n}$, then $B_{n,\infty}v_{j,n} = \lambda_{j,n} \times \left[I + o_p(n^{-1/2}\kappa_n^{-1})\right] v_{j,n}$. This implies that $v_{j,n}$ is *approximately* a right singular vector for $B_{n,LS/\infty}$ with singular value $\lambda_{j,n}$.

As a result, the quasi-Jacobian matrix $B_{n,LS/\infty}$ can be used in the sandwich formula to compute standard errors. Also, the estimate $\hat{\theta}_{n,LS/\infty}$ is asymptotically equivalent to $\hat{\theta}_n$. These two results could be of practical interest.

[10] Consider the singular value decomposition $\partial_\theta g_n(\theta_0) = U_n D_n V_n'$ where D_n is the diagonal matrix of singular values. Then $\partial_\theta g_n(\theta_0)H_n = U_n I_{d_\theta} V_n'$; this implies that 1 is an singular value with multiplicity d_θ.

[11] $v_{j,n}$ is also an orthogonal eigenvector of H_n and H_n^{-1} by construction.

3.2 Higher-Order Identification

Theorem 2. *(Approximations under Higher-Order Local Identification)*
Suppose that the model is higher-order identified at an order $r \geq 2$ and that Assumptions 1 and 2 hold. For the least-squares approximation, suppose that the assumptions of Lemma 1 hold, then the sup-norm and least-squares approximations satisfy:

$$\left|\lambda_{min}(B_{n,LS/\infty})\right| = O_p(\tilde{\kappa}_n^{1-1/r}),$$

where $\tilde{\kappa}_n = \kappa_n$ for the compact kernel, $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a}$ otherwise. Also, $\forall v_j \in Span(P_j)$, $j \in \{1, \ldots, r\}$: $B_{n,LS/\infty} v_j = O_p(\tilde{\kappa}_n^{1-1/j})$, P_j are the projection matrices in Definition 5.

Theorem 2 shows that $B_{n,LS/\infty}$ becomes singular under first-order identification failure. The rate at which the eigenvalues decay depends on both the order $r \geq 2$ and the bandwidth κ_n. When moments are increasingly flat around θ_0, i.e. r is larger, then this rate becomes slower and closer to $O_p(\kappa_n)$ which corresponds to the rate under weak or set identification. Also, $B_{n,\infty}$ vanishes in the directions of the first-order identification failure: $\text{Span}(P_2, \ldots, P_r)$. Note that if the Jacobian is non-zero but the second-order term is non-negligible then rank tests on the Jacobian would have low power. Yet, standard inferences could suffer important size distortions. This is illustrated in the Monte-Carlo simulations of Appendix G.3.

3.3 Weak or Set Identification

3.3.1 Sup-Norm Approximation

Theorem 3. *(Sup-Norm Approximation under Weak/Set Identification)*
Suppose that Assumptions 1 i., ii. and 2 hold and that there exists at least two $\theta_0 \neq \theta_1$ in the weakly identified set Θ_0, then the sup-norm approximation satisfies:

$$\left|\lambda_{min}\left(B_{n,\infty}\right)\right| = O_p(\kappa_n),$$

where $\lambda_{min}\left(B_{n,\infty}\right)$ is the smallest eigenvalue of $B_{n,\infty}$. Furthermore, $B_{n,\infty}$ is vanishing on the span of the identification failure:

$$B_{n,\infty} v = O_p(\kappa_n),\ \forall v \in V = Span\left(\left\{\theta_0 - \theta_1,\ \theta_0, \theta_1 \in \Theta_0\right\}\right).$$

Remark 1. *(The Span of the Identification Failure) Note that the span of the identification failure V can be larger than what needs to be fixed to restore point identification. For instance, the set $\Theta_0 = \{\theta = (\alpha, \beta),\ \alpha = \beta^3\}$ implies $V = \mathbb{R}^2$ even though β is uniquely determined*

when α is fixed. One could fix α, compute the quasi-Jacobian with the constraint and find that identification is restored for β. The span V is not always too large: for the weakly identified set $\Theta_0 = \{\theta = (\alpha, \beta),\ \alpha^2 + \beta^2 = 1\}$ both α and β need to be fixed.

3.3.2 Least-Squares Approximation

Proposition 1. *(Least-Squares Approximation under Weak Identification) Suppose that there exists two $\theta_0 \neq \theta_1 \in \Theta_0$ such that for some $0 < \varepsilon < \|\theta_0 - \theta_1\|$, $\exists \eta > 0$:*

$$\lim_{n\to\infty} \min \left(\int_{\|\theta-\theta_0\|_2 \leq \varepsilon/3} \hat{\pi}_n(\theta) d\theta, \int_{\|\theta-\theta_1\|_2 \leq \varepsilon/3} \hat{\pi}_n(\theta) d\theta \right) \geq \eta > 0,$$

then the quasi-posterior variance satisfies $\lambda_{\max}(\Sigma_n) \geq \eta\varepsilon^2/[36 d_\theta] + o_p(1)$ and, under the conditions of Lemma 1 $|\lambda_{min}(B_{n,LS})| = O_p(\tilde{\kappa}_n)$.

Proposition 1 simply states that when there are two different θ_0, θ_1 to which the quasi-posterior gives weight asymptotically, then Σ_n is bounded below in some direction. By Lemma 1 this implies that $B_{n,LS}$ becomes singular as in Theorem 3. The following theorem provides primitive conditions to ensure that Proposition 1 holds.

Theorem 4. *(Topology of the Weakly Identified Set Θ_0 and quasi-Posterior Concentration) Suppose that the conditions in Lemma 1 hold and that one of the following is satisfied:*

i. Θ_0 has non-empty interior,

ii. Θ_0 is finite, i.e. $\Theta_0 = \cup_{j=0}^k \{\theta_j\}$ for some finite $k \geq 2$ and

  a. $\exists \varepsilon > 0$ and $\eta(\varepsilon) > 0$ such that $\inf_{\theta\in\Theta, d(\theta,\Theta_0)\geq\varepsilon} \|g_n(\theta)\|_W \geq \eta(\varepsilon)$, $\forall n \geq 1$,

  b. $\exists \underline{C} > 0, \overline{C} > 0$ and some finite $r \geq 1$ such that for ε defined in ii.a. above $\underline{C} d(\theta, \Theta_0)^r \leq \|g_n(\theta)\|_W \leq \overline{C} d(\theta, \Theta_0)^r, \forall n \geq 1, \forall \theta \in \Theta, d(\theta, \Theta_0) \leq \varepsilon$,

iii. Θ_0 is a finite union of lower-dimensional manifolds, $\Theta_0 = \cup_{j=0}^k \mathcal{S}_j$ for some finite $k \geq 1$, where $\mathcal{S}_j$ are bounded sets with $d(\mathcal{S}_j, \mathcal{S}_{j'}) > 0, j \neq j'$ such that:

  a. for each $j \in \{0, \ldots, k\}$, $\exists k_j \in \{1, \ldots, d_\theta\}$, $\exists \mathcal{U}_j \subset \mathbb{R}^{d_\theta - k_j}$ a connected and bounded subset of $\mathbb{R}^{d_\theta - k_j}$ with non-empty interior and, $\exists \varphi_j$, an invertible, continuously differentiable mapping from the open neighbourhood $\mathcal{N}(\mathcal{U}_j) \times \mathcal{N}(\{0\}) \subseteq \mathbb{R}^{d_\theta}$ to the open neighbourhood $\mathcal{N}(\mathcal{S}_j)$ such that $\vartheta \in \mathcal{U}_j \times \{0\} \Leftrightarrow \varphi_j(\vartheta) \in \mathcal{S}_j$ and $\exists \underline{\lambda}_j, \overline{\lambda}_j$ such that $\forall \vartheta \in \varphi_j^{-1}(\mathcal{N}(\mathcal{S}_j))$: $0 < \underline{\lambda}_j \leq |\lambda_{\min}(\partial_\vartheta \varphi_j(\vartheta))| \leq |\lambda_{\max}(\partial_\vartheta \varphi_j(\vartheta))| \leq \overline{\lambda}_j < +\infty$,

b. $\exists \eta > 0$ *such that* $\inf_{\theta \notin \cup_{j=0}^{k} \mathcal{N}(\mathcal{S}_j)} \|g_n(\theta)\|_W \geq \eta$, $\forall n \geq 1$,

c. $\exists \underline{C}, \overline{C} > 0$ *such that:* $\underline{C} d(\theta, \Theta_0) \leq \|g_n(\theta)\|_W \leq \overline{C} d(\theta, \Theta_0)$, $\forall n \geq 1, \forall \theta \in \cup_{j=0}^{k} \mathcal{N}(\mathcal{S}_j)$,

then the assumptions and the results of Proposition 1 hold.

Case i. is immediate. Case ii. requires that g_n behaves similarly around each point in Θ_0. This could be weakened to require only two points to have r-th order polynomial behaviour. These two points with the largest polynomial order would dominate the posterior mass.

In Case iii., Θ_0 is a union of lower-dimensional manifolds with non-empty interior in a lower-dimension space. Within each $\mathcal{S}_j$ there are two disjoint open sets with non-zero Lebesgue measure in $\mathbb{R}^{d_\theta - k_j}$. Some primitives for condition iii. a. include the constant rank theorem and the partition of unity, which ensures there exists finitely many local re-parameterizations to create a global re-parameterization φ_j (see e.g. Lee, 2012). The eigenvalue condition on φ_j ensures that a change of variable argument can be used to construct explicit bounds on the integrals in Proposition 1. Conditions iii. b. and c. ensure that Θ_0 separates well from the rest of the parameter space.

When the manifolds have different dimensions, the proof of Theorem 4 implies that for any $\mathcal{N}(\mathcal{S}_j), \mathcal{N}(\mathcal{S}_{j'})$ which do not overlap with other manifolds: $\frac{\int_{\mathcal{N}(\mathcal{S}_{j'})} \hat{\pi}_n(\theta) d\theta}{\int_{\mathcal{N}(\mathcal{S}_j)} \hat{\pi}_n(\theta) d\theta} \xrightarrow{p} 0$, if $k_j > k_{j'}$. The posterior mass is dominated by the manifold(s) with $k^\star = \max_j k_j$, i.e. the highest degree of identification failure. Lower dimensional manifolds, with $k_j < k^\star$, have posterior measure zero in the limit. A practical implication when using MCMC to estimate Θ_0 is that the Markov-Chain might get *stuck* on the larger dimensional manifold(s). Appendix C.3 illustrates the three cases and provides an example where Proposition 1 does not hold.

Corollary 1 below, re-states Theorem 4 iii. in the more familiar settings where θ can be globally and smoothly mapped into a point identified and a weakly identified coefficient.

Corollary 1. *(Weak Identification, Global Re-Parametrizations and Posterior Concentration) Suppose that there exists a global re-parametrization* $\theta = \varphi(\beta, \gamma)$ *where* $(\beta, \gamma) \in B \times \Gamma \subseteq \mathbb{R}^{d_\beta} \times \mathbb{R}^{d_\gamma}$ *such that:*

i. $\varphi : (B \times \Gamma) \to \Theta$ *is continuously differentiable and invertible on* Θ *and* $\exists \underline{\lambda}, \overline{\lambda}$ *such that* $\forall \beta, \gamma$*:* $0 < \underline{\lambda} \leq |\lambda_{\min}(\partial_{\beta,\gamma} \varphi(\beta, \gamma))| \leq |\lambda_{\max}(\partial_{\beta,\gamma} \varphi(\beta, \gamma))| \leq \overline{\lambda} < +\infty$,

ii. γ *is point identified, i.e.* $\Theta_0 = \varphi(B_0 \times \{\gamma_0\})$*,* B_0 *is bounded with non-empty interior and,* $\exists \varepsilon > 0$*,* $\eta > 0$ *such that:* $\inf_{d(\beta, B_0) + \|\gamma - \gamma_0\| \geq \varepsilon} \|g_n \circ \varphi(\beta, \gamma)\| \geq \eta$, $\forall n \geq 1$*, and,* $\exists \underline{C} > 0$*,* $\overline{C}$ *such that* $\forall (\beta, \gamma)$ *with* $(d(\beta, B_0) + \|\gamma - \gamma_0\|) \leq \varepsilon$*:* $\underline{C} \Big[d(\beta, B_0) + \|\gamma - \gamma_0\| \Big] \leq \|g_n \circ \varphi(\beta, \gamma)\| \leq \overline{C} \Big[d(\beta, B_0) + \|\gamma - \gamma_0\| \Big]$, $\forall n \geq 1$,

then the assumptions of Theorem 4 are satisfied and $\lambda_{\min}(B_{n,LS}) = O_p(\tilde{\kappa}_n)$ *where* $\tilde{\kappa}_n = \kappa_n$ *for the compact kernel and* $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a}$ *for the exponential kernel.*

4 Two-Step Subvector Inference

Given the asymptotic properties of the quasi-Jacobian, this section develops two-step inference procedures. As discussed in the introduction, the first step amounts to determining which parameters need to be fixed and the second performs robust inference given this information. The procedure is concerned with hypotheses of the form:

$$H_0 : R\theta = c \text{ vs. } H_1 : R\theta \neq c, \tag{1}$$

for a given restriction matrix $R \in \mathbb{R}^{m \times d_\theta}$ with $1 \leq m \leq d_\theta$ and $c \in \mathbb{R}^m$.

Weak/set and higher-order identification are detected with different criteria; the two are considered in separate subsections. If both weak/set and higher-order identification are a concern, one could pre-test for weak identification, determine which coefficients need to be fixed and then check for higher-order identification issues on the remaining parameters.

Definition 8. *(Nested Sequence of Restrictions) Let* $R_1 = R$ *be the restriction matrix used to test the null hypothesis (1). The pre-determined set of restriction matrices* $(R_\ell)_{1 \leq \ell \leq \mathcal{L}}$*, is given by* $R_\ell = \left(R_{\ell-1}', \tilde{R}_\ell'\right)'$, $\quad 2 \leq \ell \leq \mathcal{L}$*, where* $\tilde{R}_\ell$ *is a sequence of* $\mathcal{L} - 1$ *matrices such that* $1 \leq m = rank(R_1) < rank(R_2) < \cdots < rank(R_{\mathcal{L}}) = d_\theta$.

To determine which coefficients need to be fixed, the algorithms consider an increasing sequence of restrictions R_ℓ, as in Definition 8. For a given R_ℓ, a criteria determines, with probability going to 1, whether the unrestricted "free" parameters lie in the span of the identification failure. At each step, the search continues so long as "free" parameters lie in this span. By construction, $R_{\mathcal{L}}$ fixes all coefficients, hence there exists $\ell^\star$ such that for $\ell \geq \ell^\star$, all "free" parameters lie outside this span. By a family-wise error rate argument, with probability going to 1, the search ends at $\hat{\ell}_n \geq \ell^\star$ thereby fixing the span of the identification failure. Once the algorithm stops, confidence sets for $R\theta$ are constructed by projection inference. First, compute a confidence set for $R_{\hat{\ell}_n}\theta$:

$$CS_{1-\alpha} = \{c \in \mathbb{R}^{\text{rank}(R_{\hat{\ell}_n})}, R_{\hat{\ell}_n}\theta = c \text{ and } S_n(\hat{\theta}_{n,c}) \leq c_{1-\alpha}\}, \quad \hat{\theta}_{n,c} = \text{argmin}_{R_{\hat{\ell}_n}\theta = c} \|\bar{g}_n(\theta)\|_{W_n}.$$

Then, the first m-rows of the elements in $CS_{1-\alpha}$ yield the projected confidence set for $R\theta$. Progress towards uniform inference would require explicit bounds on the family-wise error

rate of selecting $\hat{\ell}_n < \ell^\star$ which depends on the diameter of the weakly identified set and the tail probability of the empirical process. A rigorous investigation is left to future research.

The appeal of this approach is that $B_{n,LS/\infty}$ is only computed once. This is important when $\bar{g}_n$ is costly to evaluate. Remark 1 suggests this might be conservative, however.

4.1 Weak or set identification

The first set of results deal with weakly and set identified models. A general Algorithm is introduced and its asymptotic properties are given. Subsection 4.1.2 provides a simple rule-of-thumb for choosing the cutoff for the eigenvalues of $B_{n,LS/\infty}$.

4.1.1 Algorithm and two-step inference

Algorithm 1 combines two pieces of information: the number of eigenvalues below the threshold for the full matrix $B_{n,LS/\infty}$ and the span of the "free" parameters $P^{\perp}_{R_\ell}$. The first quantity conservatively estimates the rank of the span of the identification failure V. Fixing fewer coefficients may not suffice to resolve the identification problem. The second quantity checks if the restriction R_ℓ helps with the identification problem. While the criteria suggests too few coefficients have been fixed, the algorithm continues adding restrictions.

Algorithm 1 Fixing the Span of the Identification Failure

compute $B_{n,LS/\infty}$ and $\lambda_{\min}(B_{n,LS/\infty})$
if $\lambda_{\min}(B_{n,LS/\infty}) > \underline{\lambda}_n$ **then**
  treat all d_θ parameters as point identified
else
  set $\ell = 1$
  compute $\hat{d}_V = \#\Big\{j \in \{1,\dots,d_\theta\},\ \lambda_j(B_{n,LS/\infty} \leq \underline{\lambda}_n)\Big\}$
  while $\Big(\text{rank}(R_\ell) < \hat{d}_V\Big)$ or $\Big(\lambda_{\text{rank}(P^{\perp}_{R'_\ell})}(B_{n,LS/\infty}P^{\perp}_{R'_\ell}) \leq \underline{\lambda}_n\Big)$ **do**
    $\ell = \ell + 1$
  end while
  set $\hat{\ell}_n = \ell$
  treat $R_{\hat{\ell}_n}\theta$ as weakly identified, the remaining parameters as point identified
end if

Theorem 5. *(Two-Step Weak Identification Robust Subvector Inference) For each $\ell = 1,\dots,\mathcal{L}$, suppose that $c_{\ell,0}$ is such that there exists $\theta_{0,c}$ with $\mathbb{E}(\bar{g}_n(\theta_{0,c})) = 0, R_\ell\theta_{0,\ell} = c_{\ell,0}$ for all $\ell \in \{1,\dots,\mathcal{L}\}$. Let $\hat{\theta}_{n,\ell,c} = argmin_{R_\ell\theta = c_{\ell,0}}\|\bar{g}_n(\theta)\|_{W_n}$, be the constrained estimator of*

$\theta_{0,c}$ with R_ℓ as in Definition 8. Let $S_{n,\ell}$ be a test statistic for $H_0 : R_\ell\theta_{0,c} - c_{\ell,0} = 0$ computed at $\hat{\theta}_{n,\ell,c}$ and $c_{1-\alpha,\ell}$ the corresponding critical value. Suppose the assumptions for Lemma D4 hold as well as one of the following:

1. *(semi-strong identification) the model satisfies the assumptions of Theorem 1 and $\mathbb{P}(S_{n,1} \leq c_{1-\alpha,1}) = 1 - \alpha + o(1)$,*
2. *(weak identification) the model satisfies the assumptions of Lemma D6. Let $\ell^\star$ be the smallest $\ell \in \{1, \ldots, \mathcal{L}\}$ such that rank$(P_{R'_{\ell^\star}} P_V)$ = rank(P_V), where V is the span of the identification failure, suppose that: $\inf_{\ell \in \{\ell^\star, \ldots, \mathcal{L}\}} \mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) = 1 - \alpha + o(1)$.*

Let $\hat{\ell}_n$ be the $\ell \in \{1, \ldots, \mathcal{L}\}$ selected by Algorithm 1, then:

1. *(semi-strong identification) with probability going to 1, $\hat{\ell}_n = 1$ so that $R_{\hat{\ell}_n} = R_1$ and:*

$$\mathbb{P}(S_{n,\hat{\ell}_n} \leq c_{1-\alpha,\hat{\ell}_n}) = \mathbb{P}(S_{n,1} \leq c_{1-\alpha,1}) = 1 - \alpha + o(1).$$

2. *(weak identification) with probability going to 1, $\hat{\ell}_n \geq \ell^\star$ and:*

$$\mathbb{P}(S_{n,\hat{\ell}_n} \leq c_{1-\alpha,\hat{\ell}_n}) \geq \inf_{\ell=\ell^\star,\ldots,\mathcal{L}} \mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) = 1 - \alpha + o(1).$$

Theorem 5 allows for a large class of identification-robust test statistics in the second step. The sole requirement is that the statistics yield valid inferences for the full vector $R_\ell\theta$ for each $\ell \geq \ell^\star$. Projection inference is then used to construct the confidence set for $R\theta$. Proposition C2 in Appendix C.2.2 shows that the requirement for Theorem 5 will hold when using the S-statistic and the (semi)-strongly identified parameters are estimated.

Remark 2. *(Complete Identification Failure of the Nuisance Parameters) Let $\theta = (\alpha, \beta)$, suppose that $\bar{g}_n(\alpha, \beta)$ and $\hat{V}_n(\alpha, \beta)$ do not depend on β when $\alpha = \alpha_0$ and $\mathbb{E}(\bar{g}_n(\alpha_0, \beta)) = 0$ for all β. If the model is just-identified dim$(g) = d_\alpha + d_\beta$ and $\inf_{\alpha,\beta} \|\bar{g}_n(\alpha, \beta)\|_{\hat{V}_n^{-1}} = 0$ and Assumption 2 holds, then the projected S statistic of Stock & Wright (2000) satisfies:*

$$S_n(\alpha_0) = \inf_\beta n\|\bar{g}_n(\alpha_0, \beta)\|^2_{\hat{V}_n^{-1}} = n\|\bar{g}_n(\alpha_0, \beta_0)\|^2_{\hat{V}_n^{-1}} \xrightarrow{d} \chi^2_{dim(g)}. \tag{11}$$

This implies that projection inference has exact asymptotic coverage:

$$\lim_{n\to\infty} \mathbb{P}(S_n(\alpha_0) \leq q_{\chi^2_{dim(g)}}(1 - \alpha)) = 1 - \alpha, \tag{12}$$

whereas the test is asymptotically conservative when β is strongly identified:

$$\lim_{n\to\infty} \mathbb{P}(S_n(\alpha_0) \leq q_{\chi^2_{dim(g)}}(1 - \alpha)) = \mathbb{P}(\chi^2_{dim(g)-d_\beta} \leq q_{\chi^2_{dim(g)}}(1 - \alpha)) > 1 - \alpha. \tag{13}$$

Remark 2 does not appear to have been discussed in the literature although it may be already known. A simple proof is given in the Appendix.

4.1.2 A data-driven rule-of-thumb

The following provides a data-driven approach to find a cutoff $\underline{\lambda}_n$ for Algorithm 1. The main idea is to consider (semi)-strongly identified models and local asymptotics where the Jacobian is increasingly flat. Higher-order Nagar expansions allow to approximate the resulting size-distortion and to find a cutoff on the signal to noise ratio where size distortion is greater than some pre-determined threshold (Stock & Yogo, 2005). This approach is convenient but not uniform, although one would expect it to perform reasonably well in settings where these local asymptotics approximate the identification failure well enough. A non-local approach is also considered in Appendix C.1.2.

The following relates the eigenvalues of $B_{n,LS/\infty}$ with the size distortion of a test on scalar hypotheses of the form $H_0 : v\theta = c$, for some $v \in \mathbb{R}^{d_\theta}/\{0\}$ and $c \in \mathbb{R}$, in the just-identified case.[12] Consider the asymptotic experiment:

$$\bar{g}_n(\theta) = A_{n,LS} + B_{n,LS}\theta,$$

where $A_{n,LS} + B_{n,LS}\theta_0 = Z_1$, $B_{n,LS} - \bar{B}_{n,LS} = Z_2$, $(Z_1', \mathrm{vec}(Z_2)')'$ is Gaussian and $\bar{B}_{n,LS}$ is non-stochastic and invertible. $\mathbb{E}(Z_1 Z_1') = V_1/n$, $\mathbb{E}(Z_1 Z_2) = V_{12}/n$. Furthermore, assume that $n \times \lambda_{\min}(\bar{B}_{n,LS}'\bar{B}_{n,LS}) \to +\infty$, so that the model is linear and semi-strongly identified. Using the Woodbury identity, it can be shown that: $\hat{\theta}_{n,LS} - \theta_0 = -\bar{B}_{n,LS}^{-1} Z_1 + \bar{B}_{n,LS}^{-2} Z_2 Z_1 - \bar{B}_{n,LS}^{-3}\left(I + Z_2\bar{B}_{n,LS}^{-1}\right)^{-1} Z_2^2 Z_1$. As a result, the following expansions hold for $v_{j,n} \in \mathbb{C}^{d_\theta}$ the complex eigenvector of $\bar{B}_{n,LS}$ associated the complex eigenvalue $\lambda_{j,n} \in \mathbb{C}$:

$$\begin{aligned}
\mathbb{E}(v_{j,n}'[\hat{\theta}_{n,LS} - \theta_0]) &= v_{j,n}'\bar{B}_{n,LS}^{-2} V_{21}/n + O\left(\frac{1}{n^2 \times |\lambda_{j,n}|^2}\right) \\
\mathbb{E}(|v_{j,n}'[\hat{\theta}_{n,LS} - \theta_0]|^2) &= v_{j,n}'\bar{B}_{n,LS}^{-1} V_1 \bar{B}_{n,LS}^{-1\prime}\bar{v}_{j,n}/n + O\left(\frac{1}{n^2 \times |\lambda_{j,n}|^2}\right),
\end{aligned}$$

where $\bar{v}_{j,n}$ is the complex conjugate of $v_{j,n}$. The first term approximates the higher-order bias and the second the asymptotic variance. The squared bias to variance ratio of $v_{j,n}'(\hat{\theta}_{n,LS} - \theta_0)$ can be approximated by:

$$\frac{|\text{bias}|^2}{|\text{variance}|} = \frac{1}{n \times |\lambda_{j,n}|^2}\frac{v_{j,n}' V_{21} V_{12}\bar{v}_{j,n}}{v_{j,n}' V_1 \bar{v}_{j,n}} + o\left(\frac{1}{n \times |\lambda_{j,n}|^2}\right).$$

Using Nagar semi-strong asymptotics, as in Stock & Yogo (2005),[13] the size distortion γ for a Wald test on $H_0 : v_{j,n}(\theta - \theta_0) = 0$ at the $1 - \alpha$ confidence level can be approximated by

[12] The over-identified case is discussed in the Appendix.

[13] Note that under weak identification the remainder may not be negligible.

$\gamma \simeq 1-\alpha-\mathbb{P}\left(w_n^\star \leq c_{1-\alpha}\right)$, (Rothenberg, 1984) where $w_n^\star$ follows a non-central χ_1^2 distribution with non-centrality parameter $\frac{1}{\sqrt{n}\times|\lambda_{j,n}|}\frac{\sqrt{v'_{j,n}V_{21}V_{12}\bar{v}_{j,n}}}{\sqrt{v'_{j,n}V_1\bar{v}_{j,n}}}$ and $c_{1-\alpha}$ is the $1-\alpha$ quantile of a central χ_1^2 distribution. Note that γ is increasing in the non-centrality parameter. Hence, imposing a maximum level of size distortion $\gamma \leq \bar{\gamma}_n$ can be achieved with the restriction:

$$|\lambda_{\min}(\bar{B}_{n,LS})|^2 \geq \frac{1}{n \times c(\bar{\gamma}_n)^2} \max_{\|v\|=1} \frac{v'V_{21}V_{12}v}{v'V_1v},$$

where $c(\bar{\gamma}_n)$ solves the equation $\bar{\gamma}_n = 1-\alpha-\mathbb{P}\left(\bar{w}_n^\star \leq c_{1-\alpha}\right)$, $\bar{w}_n^\star$ follows a non-central χ_1^2 distribution with non-centrality parameter $c(\bar{\gamma}_n)$. A closed-form, potentially conservative but user-friendly, upper-bound for the maximum on the non-centrality parameter can be derived from a ratio of the largest eigenvalue of $V_{21}V_{12}$ to the smallest eigenvalue of V_1. A data-driven cutoff for $\lambda_{\min}(B_{n,LS/\infty})$ is then given by:

$$\underline{\lambda}_n^2 = \frac{1}{n \times c(\bar{\gamma}_n)^2} \frac{\lambda_{\max}(V_{21}V_{12})}{\lambda_{min}(V_1)}. \tag{14}$$

In practice, the quantities V_1 and V_{21} need to be approximated to make the rule-of thumb (14) feasible. Lemma C2 suggests an approach to approximate these quantities: for now, suppose θ_0 is known, then V_1 and V_{21} can be approximated with the variance of

$$\frac{1}{n}\sum_{i=1}^{n} \text{vec}\left(\left[\int X(\theta)X(\theta)'\hat{\pi}_n(\theta)d\theta\right]^{-1}\int X(\theta)g_i(\theta)'\hat{\pi}_n(\theta)d\theta\right), \quad X(\theta) = (1,\theta')',$$

while assuming $\Sigma_n, \bar{\theta}_n$ and $\hat{\pi}_n$ are fixed. Since Z_1, Z_2 are linear transformations of this sample mean the $\hat{V}_1$ and $\hat{V}_{21}$ can be derived from the sample variance-covariance matrix for iid or the HAC estimator for time-series data. Since θ_0 is unknown in practice, one can take the least-favorable θ with $\|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n$. Under (semi)-strong asymptotics $\bar{\theta}_n$ is a consistent estimator of θ_0 so that substituting θ_0 for $\bar{\theta}_n$ is also possible.

4.2 Higher-order identification

The following deals with higher-order identified models. It is assumed throughout that weak and set identifications are not a concern here. The rule-of-thumb, in Subsection 4.2.2, is more involved than the one for weak identification because it involves the residual curvature of the moments. The results rely on projection inference as for weak/set identification.

4.2.1 Algorithm and two-step inference

The main intuition for this section is that the remainder in the first-order expansion is non-negligible under higher-order identification and has the same order of magnitude as the

variance. This implies that $\sqrt{n} \times \|R_n(\theta)\| \asymp (\sqrt{n} \times \kappa_n)$ for some of the $\theta \in \Theta$ such that $\|g_n(\theta)\|_W \leq \kappa_n$, which diverges to $+\infty$ in the directions of the first-order identification failure if $\sqrt{n}\kappa_n \to +\infty$, where $R_n(\theta) = g_n(\theta - g_n(\theta_0) - \partial_\theta g_n(\theta_0)(\theta - \theta_0)$ is the remainder in the first-order Taylor expansion. Whereas when the model is first-order identified: $\sqrt{n} \times \|R_n(\theta)\| \asymp (\sqrt{n} \times \kappa_n^2)$ for all $\theta \in \Theta$ such that $\|g_n(\theta)\|_W \leq \kappa_n$, the remainder in the first-order expansion is negligible if $\sqrt{n}\kappa_n^2 \to 0$. This difference between the two regimes suggests that $\sqrt{n}R_n(\cdot)$ is informative about local identification failure.

Algorithm 2 Fixing the Span of the Higher-Order Identification Failure

compute B_n and $\hat{R}_n(\cdot)$; approximate V_1 the asymptotic variance of $\sqrt{n} \times \bar{g}_n(\theta_0)$; set $\ell = 0$
if B_n is singular **then**
  set $\bar{h} = +\infty$
else
  compute $\bar{h}^2 = \max_{v \in \mathbb{R}^{d_\theta}, \|v\|=1} \left(\max_{\theta \in \Theta, \|\bar{g}_n(\theta)\|^2_{W_n} \leq \kappa_n^2} n \times \frac{\hat{R}_n(\theta)' B_W' v v' B_W \hat{R}_n(\theta)}{v' B_W V_1 B_W' v} \right)$
  where $B_W = -\left[B_n' W_n(\bar{\theta}_n) B_n\right]^{-1} B_n' W_n(\bar{\theta}_n)$
  compute $\gamma_0 = |1 - \alpha - \mathbb{P}(w_n^\star \leq c_{1-\alpha})|$; $w_n^\star$ follows a non-central χ_1^2 distribution with non-centrality parameter $\bar{h}$; $c_{1-\alpha}$ is the $1-\alpha$ quantile of a central χ_1^2 distribution
end if
while $\gamma_\ell > \bar{\gamma}_n$ and $\ell < \mathcal{L}$ **do**
  set $\ell = \ell + 1$
  if $B_{n,\ell} = (B_n', R_\ell')'$ is singular **then**
    set $\bar{h} = +\infty$
  else
    compute $\bar{h}^2 = \max_{v \in \mathbb{R}^{d_\theta}, \|v\|=1} \left(\max_{\theta \in \Theta, \|\bar{g}_n(\theta)\|^2_{W_n} + \|R_\ell \theta - c_\ell\|^2 \leq \kappa_n^2} n \times \frac{\hat{R}_n(\theta)' B_{W,\ell}' v v' B_{W,\ell} \hat{R}_n(\theta)}{v' B_{W,\ell} V_1 B_{W,\ell}' v} \right)$
    where $B_{W,\ell} = -\left[B_{n,\ell}' W_n(\bar{\theta}_n) B_{n,\ell}\right]^{-1} B_{n,\ell}' W_n(\bar{\theta}_n)$
    compute $\gamma_\ell = |1 - \alpha - \mathbb{P}(w_n^\star \leq c_{1-\alpha})|$
  end if
end while
set $\hat{\ell}_n = \ell$
if $\hat{\ell}_n \geq 1$ **then**
  treat $R_{\hat{\ell}_n}\theta$ as higher-order identified, the remaining parameters as first-order identified
else
  treat the full vector θ as first-order identified
end if

Algorithm 2 combines this idea with a rule-of-thumb 4.2.2 based on Nagar asymptotics. It separates (semi)-strongly from higher-order identified models under mild conditions as

shown in Appendix D.2. A level of tolerance for size distortion $\bar{\gamma}_n$ is required. The user can choose either $\bar{\gamma}_n \searrow 0$ or $\bar{\gamma}_n = \gamma > 0$ small but fixed. In the Monte-Carlo simulations, the rule-of-thumb was computed for W = identity matrix and performed well. Also, the criterion $\bar{h}^2$ in Algorithm 2 can be quickly bounded above by a more user-friendly quantity given at the end of subsection 4.2.2.

Theorem 6. *(Two-Step Higher-Order Identification Robust Subvector Inference) For each $\ell = 1, \ldots, \mathcal{L}$, suppose that $c_{\ell,0}$ is such that $R_\ell \theta_0 = c_\ell$ with $\mathbb{E}(\bar{g}_n(\theta_0)) = 0$, for all $\ell \in \{1, \ldots, \mathcal{L}\}$. Let $\hat{\theta}_{n,\ell,c} = argmin_{R_\ell \theta = c_{\ell,0}} \|\bar{g}_n(\theta)\|_{W_n}$, be the constrained estimator of θ_0 with R_ℓ as in Definition 8. Let $S_{n,\ell}$ be a test statistic for $H_0 : R_\ell \theta_0 - c_{\ell,0} = 0$ computed at $\hat{\theta}_{n,\ell,c}$ and $c_{1-\alpha,\ell}$ the corresponding critical value. Suppose one of the following holds:*

1. *(semi-strong identification) the model satisfies the assumptions of Lemma D8, $\sqrt{n}\kappa_n^2 / \lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0)) = o(\bar{\gamma}_n)$, and $\mathbb{P}(S_{n,1} \leq c_{1-\alpha,1}) = 1 - \alpha + o(1)$,*
2. *(higher-order identification) the model satisfies the assumptions of Lemma D7. Let $\ell^\star$ be the smallest $\ell \in \{1, \ldots, \mathcal{L}\}$ such that $rank(P_{R'_{\ell^\star}} P_{V_r}) = rank(P_{V_r})$, where $V_r = Span(P_2, \ldots, P_r)$ is the span of the first-order identification failure, suppose that $\inf_{\ell \in \{\ell^\star, \ldots, \mathcal{L}\}} \mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) = 1 - \alpha + o(1)$.*

Let $\hat{\ell}_n$ be the $\ell \in \{1, \ldots, \mathcal{L}\}$ selected by Algorithm 2, then:

1. *(semi-strong identification) with probability going to 1, $\hat{\ell}_n = 1$ so that $R_{\hat{\ell}_n} = R_1$ and:*

$$\mathbb{P}(S_{n,\hat{\ell}_n} \leq c_{1-\alpha,\hat{\ell}_n}) = \mathbb{P}(S_{n,1} \leq c_{1-\alpha,1}) = 1 - \alpha + o(1).$$

2. *(higher-order identification) with probability going to 1, $\hat{\ell}_n \geq \ell^\star$ and:*

$$\mathbb{P}(S_{n,\hat{\ell}_n} \leq c_{1-\alpha,\hat{\ell}_n}) \geq \inf_{\ell = \ell^\star, \ldots, \mathcal{L}} \mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) = 1 - \alpha + o(1).$$

4.2.2 A data-driven rule-of-thumb

The following provides a simple approach to approximate the size distortion due to the non-linearity in the objective function for tests on scalar hypotheses of the form $H_0 : v\theta_0 = c$, for some $v \in \mathbb{R}^{d_\theta} / \{0\}$ and $c \in \mathbb{R}$ in both just and over-identified models using semi-strong Nagar asymptotics. Consider the following asymptotic experiment:

$$\bar{g}_n(\theta) = A_n + B_n(\theta - \theta_0) + R_n(\theta - \theta_0)$$

where $\sqrt{n}A_n = Z_1 \sim \mathcal{N}(0, V_1)$, B_n and R_n are non-stochastic. Assume that $n \times \lambda_{\min}(B_n' B_n) \to +\infty$ and for any $v \in \mathbb{R}^{d_\theta}/\{0\}$, $vR_n(\theta - \theta_0) = o(\|vB_n(\theta - \theta_0)\|)$, $\|vR_n(\theta - \theta_0)\| \leq O(\|\theta - \theta_0\|^2)$. For a weighting matrix $W > 0$, the minimizer $\hat{\theta}_n$ of $\|\bar{g}_n\|_W$ solves:

$$\hat{\theta}_n - \theta_0 = -\left[B_n' W B_n\right]^{-1} B_n' W [A_n + R_n(\hat{\theta}_n - \theta_0)] = B_W A_n + B_W R_n(B_W A_n) + o_p(r_n)$$

where $B_W = -\left[B'WB\right]^{-1} B'W$ and r_n satisfies $B_W R_n(B_W A_n) = O_p(r_n)$. Let $\bar{R}_n = \mathbb{E}[R_n(B_W A_n)]$. For $v \in \mathbb{R}^{d_\theta}/\{0\}$, the Wald statistic w_n for testing $v'\theta_0 - c = 0$ is

$$w_n = n \times \left(v'[\hat{\theta}_n - \theta_0]\right)' \left(B_W V_1 B_W'\right)^{-1} \left(v'[\hat{\theta}_n - \theta_0]\right).$$

Using Nagar asymptotics for w_n (Rothenberg, 1984), $\mathbb{P}(w_n \leq c_{1-\alpha}) = \mathbb{P}(w_n^\star \leq c_{1-\alpha}) + o(r_n)$ where $w_n^\star$ follows a non-central χ_1^2 distribution with non-centrality parameter: $[n \times \bar{R}_n' B_W' v \left(v' B_W V_1 B_W' v\right)^{-1} v' B_W \bar{R}_n]^{1/2}$. A feasible upper-bound for the non-centrality parameter can be computed as:

$$\sup_{\|v\|=1,\, \theta \in \Theta,\, \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n} \left[n \times \hat{R}_n(\theta)' B_W' v \left(v' B_W V_1 B_W' v\right)^{-1} v' B_W \hat{R}_n(\theta)\right]^{1/2}$$

where $\hat{R}_n$ approximates the remainder term R_n. Plug-in estimates include $\hat{R}_n(\theta) = \bar{g}_n(\theta) - A_{n,LS/\infty} - B_{n,LS/\infty}\theta$ and $\hat{R}_n(\theta) = \bar{g}_n(\theta) - \bar{g}_n(\bar{\theta}_n) - \partial \bar{g}_n(\bar{\theta}_n)\theta$. Since $\bar{\theta}_n$ is consistent under (semi)-strong and higher-order identification, V_1 can be consistently estimated using the sample variance-covariance matrix of $\bar{g}_n$ evaluated at $\bar{\theta}_n$. For time-series data, a HAC estimator should be used. In Algorithm 2, $B_{W,\ell}$ is used instead of B_W when constraints are enforced to ensure that the non-centrality parameter does not involve a singular matrix in the denominator. This allows to remove $B_{W,\ell}$ from the optimization and directly compute:

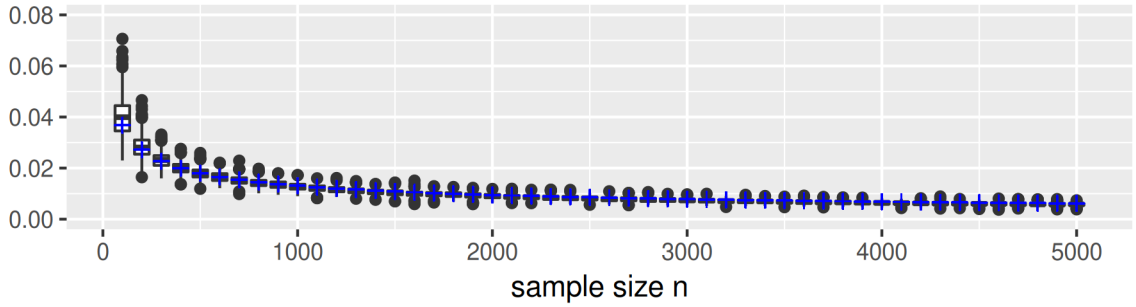
$$\sup_{\theta \in \Theta, \|\bar{g}_n(\theta)\|_{W_n} + \|R_\ell \theta - c_\ell\|^2 \leq \kappa_n^2} \frac{\|\hat{R}_n(\theta)\|_\infty^2}{\lambda_{\min}(V_1)}. \tag{15}$$

This is a more user-friendly (although potentially more conservative) upper bound for the non-centrality parameter. This quantity was used in the Monte-Carlo simulations in Appendix G.3. The other quantity used in the simulations relies on $\|B_W \hat{R}_n(\theta)\|_\infty^2$ and $\lambda_{\min}(B_W V_1 B_W')$. Both performed well in the simulations for the example considered.

5 Monte-Carlo Simulations

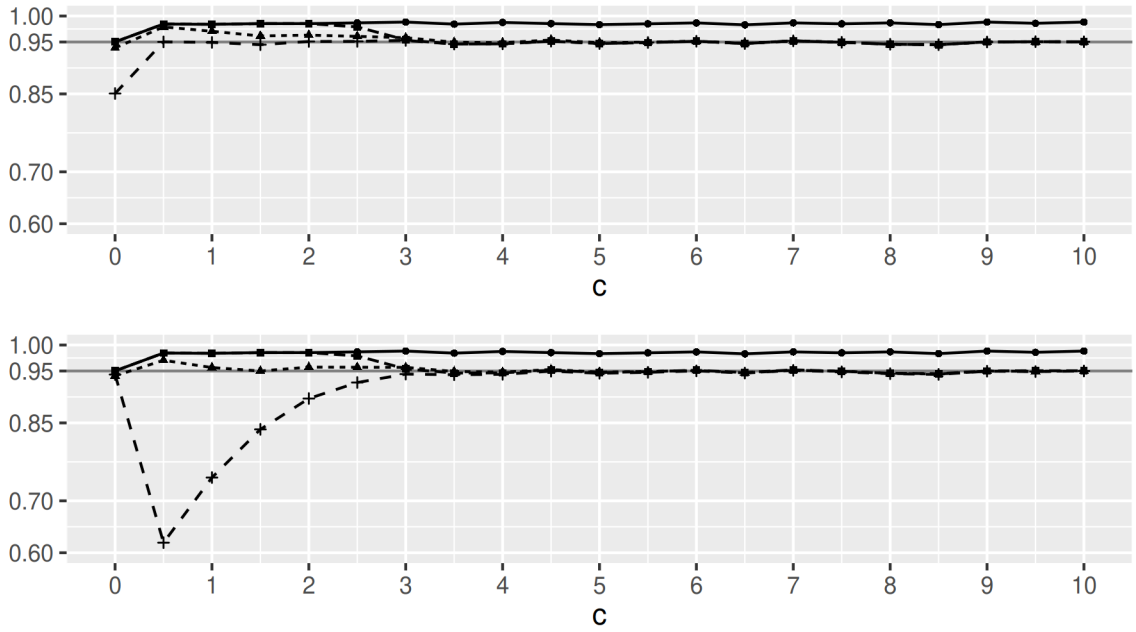
This section illustrates some of the results using Example 1. Appendices G.2 and G.3 provide additional examples. Simulations were conducted in R and C++ through Rcpp. Recall the non-linear model from Example 1: $y_i = \theta_{1,n} x_{i,1} + \theta_{1,n}\theta_2 x_{i,2} + e_i$, $(x_{i,1}, x_{i,2}, e_i) \overset{iid}{\sim} \mathcal{N}(0, I)$.

Figure 1: Distribution of $\lambda_{\min}(B_{n,LS})$ and sample size n



Note: $y_i = \theta_{1,n} x_{i,1} + \theta_{1,n}\theta_2 x_{i,2} + e_i$, $\theta_{1,n} = 2 \times n^{-1/2}$, $100 \leq n \leq 5,000$, $B = 500$ Monte-Carlo replications and $\kappa_n = \sqrt{2\log(\log[n])n^{-1/2}}$. **Legend:** Black lines - boxplot $\lambda_{\min}(B_{n,LS})$ for each n; Blue crosses - fitted rate from regressing $\lambda_{\min}(B_{n,LS})$ draws on κ_n with OLS without an intercept.

Figure 2: Coverage of the 95% Confidence Intervals



Note: $y_i = \theta_{1,n} x_{i,1} + \theta_{1,n}\theta_2 x_{i,2} + e_i$, $\theta_{1,n} = c \times n^{-1/2}$, $n = 1,000$, $B = 5,000$ replications, $\kappa_n = \sqrt{2\log(\log[n])n^{-1/2}}$. **Legend:** Anderson-Rubin (solid/dot) - projected CI; Standard (dashed/cross) - QLR (top panel) and Wald (bottom panel) CIs; Rule-of-thumb (dotted/triangle) - two-step procedure with $\underline{\lambda}_n =$ data-driven rule-of-thumb; $\sqrt{\log(n)}$ (dashed/square) - two-step procedure with $\underline{\lambda}_n = \sqrt{\log(n)}$.

When $\theta_{1,n} = 0$, θ_2 is not identified; $\theta_{1,n} \asymp n^{-1/2}$ implies weak identification. The estimating moments are $\bar{g}_n(\theta) = \frac{1}{n}\sum_{i=1}^{n}(y_i - \theta_1 x_{i,1} - \theta_1\theta_2 x_{i,2})(x_{i,1}, x_{i,2})'$. $B_{n,LS}$ was computed with

a grid of $10,000$ Sobol points. Figure 1 plots the distribution of $\lambda_{\min}(B_{n,LS})$ against the predicted rate from Theorem 1. The top panel in Figure 2 shows the coverage for standard QLR and robust inferences under weak identification. Remark 2 holds for $c = 0$ and projection inference has exact coverage. The bottom panel shows Wald inferences which display greater size distortion than QLR inferences. The two-step procedures rely on the rule-of-thumb and a pre-determined sequence. Both perform well using the QLR and the Wald statistic. Power curves are reported in Appendix G.1 to illustrate improvements over projection inference.

6 Empirical Application: US Euler Equation

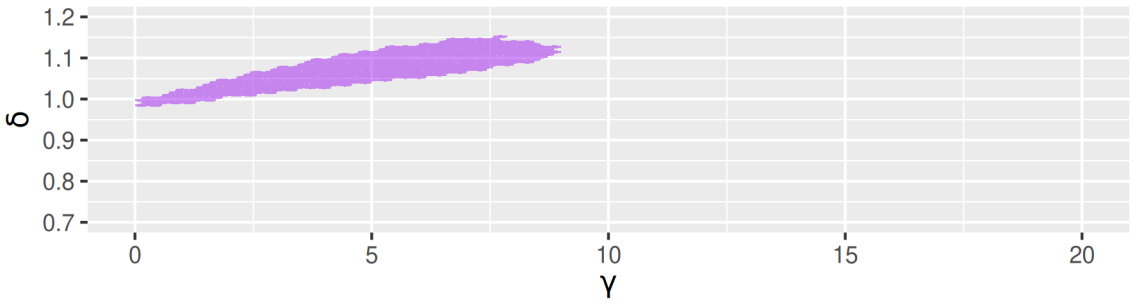
The main application considers the US Euler equation:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{t=1}^{n}\left(\left[\delta\left(\frac{C_t}{C_{t-1}}\right)^{-\gamma} R_t - 1\right] Z_t\right),$$

where C_t is US consumption, R_t the risk-free rate and $Z_t = (1, C_{t-1}/C_{t-2}, R_{t-1})$. δ corresponds to time-preference and γ relative risk-aversion. The data is from Stock & Wright (2000). After taking lags, we have $n = 103$. $W_n(\theta) = (\hat{V}_n(\theta) + 10^{-4} \cdot I)^{-1}$ is computed using *vcovHAC* in the R *sandwich* package; the regularization ensures invertibility holds. The bounds are $(\delta, \gamma) \in [0.7, 1.2] \times [0, 20]$ to match earlier replications. The grid uses 20,000 Sobol points; the results are similar with 2,000 points. Figure 3 shows the region selected by the compact kernel. The estimated $B_{n,LS}$ and its eigenvalues are:

$$B'_{n,LS} = \begin{pmatrix} 0.669 & 0.685 & 0.682 \\ -0.001 & -0.001 & 0.000 \end{pmatrix}, \quad \sqrt{n} \times \lambda(B_{n,LS}) = (11.929, 0.006).$$

Figure 3: US Euler Equation - $\hat{\Theta}_n = \{\theta \in \Theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_{\theta\in\Theta} \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n\}$



Note: $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_3)}, \sqrt{2\log(\log[n])}n^{-1/2})$; $q_{0.99}$ is the 99% quantile of a χ^2_3 distribution.

The first row in the matrix above are the coefficients for δ and the second are for γ. The eigenstructure indicates that the largest eigenvalue projects on δ and the smallest on γ. The rule-of-thumb suggests a cutoff greater than 200 because of issues with $\hat{V}_n$ discussed below. This confirms previous concerns about weak identification.

As discussed in other replications, inferences are very sensitive to tuning parameters in the HAC estimator $\hat{V}_n$ which is actually ill-conditioned because the moments are redundant (see Figure H12 in Appendix H.1). This near-singularity of the variance-covariance matrix implies that a *singularity and identification robust test* should be implemented (Andrews & Guggenberger, 2019). Since $\hat{V}_n$ is near rank one over most of $\hat{\Theta}_n$, singularity robust inference amounts to dropping two instruments and using the remaining one for inference. The intercept, $Z_t = 1$, is kept and an Anderson-Rubin statistic is inverted with χ^2_1 critical values which yields: $CI_{95\%}(\delta) = [0.98, 1.2]$; $CI_{95\%}(\gamma) = [0, 20]$. Recall that the rule-of-thumb involves $\hat{V}_n$, an ill-conditioned matrix, in the denominator. This leads to a very large cutoff when using all three moments. With a single moment condition, the cutoff is smaller, $\sqrt{n}\underline{\lambda}_n = 1.6 < 11.9 = \sqrt{n}\lambda_{\max}(B_{n,LS})$, and suggests δ is (semi)-stongly identified for γ fixed.

7 Conclusion

This paper proposes an approach to detect potential identification failure and conduct two-step robust subvector inference. It generalizes the first stage F-statistic and rank tests in linear IV to non-linear GMM. The computation is massively parallel. The recommended procedure is similar to type I inferences in Andrews & Cheng (2012), without knowing the identification structure. An important direction for future research is to extend the results to uniform type II inferences and other M-estimators.

References

Anderson, T. W. & Rubin, H. (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. & Cheng, X. (2012). Estimation and Inference With Weak, Semi-Strong, and Strong Identification. *Econometrica*, 80(5), 2153–2211.

Andrews, D. W. K. & Guggenberger, P. (2019). Identification- and Singularity-Robust Inference for Moment Condition Models. *Forthcoming in Quantitative Economics*.

Andrews, I. & Mikusheva, A. (2016a). A Geometric Approach to Nonlinear Econometric Models. *Econometrica*, 84(3), 1249–1264.

Andrews, I. & Mikusheva, A. (2016b). Conditional Inference With a Functional Nuisance Parameter. *Econometrica*, 84(4), 1571–1612.

Antoine, B. & Renault, E. (2009). Efficient GMM with nearly-weak instruments. *Econometrics Journal*, 12(1), S135–S171.

Antoine, B. & Renault, E. (2017). *Testing Identification Strength*. Discussion papers, Department of Economics, Simon Fraser University.

Arellano, M., Hansen, L. P., & Sentana, E. (2012). Underidentification? *Journal of Econometrics*, 170(2), 256–280.

Bhatia, R. (1997). *Matrix Analysis*, volume 169 of *Graduate Texts in Mathematics*. New York, NY: Springer New York.

Bravo, F., Carlos Escanciano, J., & Otsu, T. (2012). A Simple Test for Identification in GMM under Conditional Moment Restrictions. In *Badi H. Baltagi, R. Carter Hill, Whitney K. Newey, Halbert L. White (ed.) Essays in Honor of Jerry Hausman (Advances in Econometrics, Volume 29) Emerald Group Publishing Limited* (pp. 455–477).

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

Chernozhukov, V. & Hansen, C. (2005). An IV model of quantile treatment effects. *Econometrica*, 73(1), 245–261.

Chernozhukov, V., Hansen, C., & Jansson, M. (2007). Inference approaches for instrumental variable quantile regression. *Economics Letters*, 95(2), 272–277.

Chernozhukov, V. & Hong, H. (2003). An MCMC approach to classical estimation. *Journal of Econometrics*, 115(2), 293–346.

Cragg, J. G. & Donald, S. G. (1993). Testing Identifiability and Specification in Instrumental Variable Models. *Econometric Theory*, 9(2), 222–240.

Dovonon, P. & Hall, A. R. (2018). The asymptotic properties of GMM and indirect inference under second-order identification. *Journal of Econometrics*, 205(1), 76–111.

Dovonon, P., Hall, A. R., & Kleibergen, F. (2019). Inference in Second-Order Identified Models. *Forthcoming in the Journal of Econometrics*.

Dufour, J.-M. & Taamouti, M. (2005). Projection-Based Statistical Inference in Linear Structural Models with Possibly Weak Instruments. *Econometrica*, 73(4), 1351–1365.

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

133–153.

Fisher, F. M. (1967). The Identification Problem in Econometrics. *Economica*, 34(135), 344.

Gospodinov, N. & Ng, S. (2015). Minimum Distance Estimation of Possibly Noninvertible Moving Average Models. *Journal of Business & Economic Statistics*, 33(3), 403–417.

Hahn, J. & Hausman, J. (2002). A New Specification Test for the Validity of Instrumental Variables. *Econometrica*, 70(1), 163–189.

Hansen, L. P. & Singleton, K. J. (1982). Generalized Instrumental Variables Estimation of Nonlinear Rational Expectations Models. *Econometrica*, 50(5), 1269–1286.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning*. Springer Series in Statistics. New York, NY: Springer New York.

Inoue, A. & Rossi, B. (2011). Testing for weak identification in possibly nonlinear models. *Journal of Econometrics*, 161(2), 246–261.

Kleibergen, F. (2005). Testing Parameters in GMM Without Assuming that They Are Identified. *Econometrica*, 73(4), 1103–1123.

Komunjer, I. (2012). GLOBAL IDENTIFICATION IN NONLINEAR MODELS WITH MOMENT RESTRICTIONS. *Econometric Theory*, 28(04), 719–729.

Koopmans, T. C. & Reiersol, O. (1950). The Identification of Structural Characteristics. *The Annals of Mathematical Statistics*, 21(2), 165–181.

Lee, J. H. & Liao, Z. (2018). ON STANDARD INFERENCE FOR GMM WITH LOCAL IDENTIFICATION FAILURE OF KNOWN FORMS. *Econometric Theory*, 34(04), 790–814.

Lee, J. M. (2012). *Introduction to Smooth Manifolds*, volume 218 of *Graduate Texts in Mathematics*. New York, NY: Springer New York.

Lemieux, C. (2009). *Monte Carlo and Quasi-Monte Carlo Sampling*. Springer Series in Statistics. New York, NY: Springer New York.

Lewbel, A. (2018). The Identification Zoo - Meanings of Identification in Econometrics. *Journal of Economic Literature, forthcoming*.

Marin, J. M., Pudio, P., Robert, C., & Ryder, R. (2012). Approximate Bayesian Computation Methods. *Statistical Computations*, 22, 1167–1180.

McCloskey, A. (2017). Bonferroni-based size-correction for nonstandard testing problems. *Journal of Econometrics*, 200(1), 17–35.

Newey, W. K. & McFadden, D. (1994). Large sample estimation and hypothesis testing. *Handbook of Econometrics*, 4, 2111–2245.

Robert, C. & Casella, G. (2004). *Monte Carlo Statistical Methods*.

Rothenberg, T. J. (1971). Identification in Parametric Models. *Econometrica*, 39(3), 577.

Rothenberg, T. J. (1984). Chapter 15 Approximating the distributions of econometric estimators and test statistics.

Rotnitzky, A., Cox, D. R., Bottai, M., & Robins, J. (2000). Likelihood-Based Inference with Singular Information Matrix. *Bernoulli*, 6(2), 243.

Stock, J. H. & Wright, J. H. (2000). GMM with Weak Identification. *Econometrica*, 68(5), 1055–1096.

Stock, J. H. & Yogo, M. (2005). Testing for Weak Instruments in Linear IV Regression. In D. W. K. Andrews & J. H. Stock (Eds.), *Identification and Inference for Econometric Models* (pp. 80–108). Cambridge: Cambridge University Press.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*. Springer Series in Statistics. New York, NY: Springer New York.

Wright, J. H. (2003). DETECTING LACK OF IDENTIFICATION IN GMM. *Econometric Theory*, 19(02).

Appendix A Proofs for the Results of Section 3

The proofs are given separately for the sup-norm and least-squares approximations.

A.1 Proofs for the sup-norm approximation

Proof of Theorem 1. By compactness of the kernel K: $\sup_{\theta\in\Theta,\|\theta-\theta_0\|\geq\varepsilon} K(\|\bar{g}_n(\theta)\|_{W_n}) = 0$, with probability going to 1. This allows us to focus on θ such that $\|\theta-\theta_0\| \leq \varepsilon$. Let $H_n = (\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0))^{-1/2}$; using the re-parameterization $\theta = \theta_0 + H_n h\kappa_n$ for $h \in \mathbb{R}^{d_\theta}$ and Definition 4 i.:

$$\overline{C}\|\partial_\theta g_n(\theta_0)H_n h\| \geq \|g_n(\theta)\|_W/\kappa_n \geq \underline{C}\|\partial_\theta g_n(\theta_0)H_n h\|.$$

By definition of the Frobenius norm and H_n, we have:

$$\text{trace}\left(\partial_\theta g_n(\theta_0)H_n hh'H_n\partial_\theta g_n(\theta_0)'\right)^{1/2} = \text{trace}(hh')^{1/2} = \|h\|.$$

Uniformly in h such that $\|H_n h\kappa_n\| \leq \varepsilon$, we have:

$$\|\bar{g}_n(\theta)\|_W/\kappa_n \geq \underline{C}\|h\| - \overline{\lambda} \times \frac{\sup_{\theta\in\Theta}\|\mathbb{G}_n(\theta)\|}{\sqrt{n}\kappa_n} = \underline{C}\|h\| - o_p(1).$$

This implies that for $\|h\| \geq 3/\underline{C}$, we have: $\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \geq \|\bar{g}_n(\theta)\|_W/\kappa_n + o_p(1) \geq 3 - o_p(1) \geq 2$, with probability going to 1. Hence, $\sup_{\|h\|\geq 3/\underline{C}} K\left(\left\|\bar{g}_n\left(\theta_0 + H_n h\kappa_n\right)\right\|_{W_n}/\kappa_n\right) =$

0, also with probability going to 1. Similarly, for $\|h\| \leq 1/[2\overline{C}]$, we have:

$$\begin{aligned}\left\|\bar{g}_n\left(\theta_0 + H_n h\kappa_n\right)\right\|_{W_n}/\kappa_n &\leq \left(\left\|g_n\left(\theta_0 + H_n h\kappa_n\right)\right\|_W \times [1 + o_p(1))] + \overline{\lambda} \times \frac{\sup_{\theta\in\Theta}\|\mathbb{G}_n(\theta)\|}{\sqrt{n}}\right)/\kappa_n \\ &\leq \overline{C} \times \|h\| \times [1 + o_p(1))] + o_p(1) \leq 1/2 + o_p(1) \leq 3/4,\end{aligned}$$

with probability going to 1. This implies that: $\inf_{\|h\|\leq 1/[2\overline{C}]} K\left(\left\|\bar{g}_n\left(\theta_0 + H_n h\kappa_n\right)\right\|_{W_n}/\kappa_n\right) \geq \inf_{x\in[0,3/4]} K(x)$, which is strictly positive by assumption. For any pair (A, B) we can write, with probability going to 1:

$$\begin{aligned}\sup_{\theta\in\Theta}\|\bar{g}_n(\theta) - A - B\theta\| \times \hat{K}_n(\theta) &= \sup_{\|h\|\leq 3/[\underline{C}]}\|\bar{g}_n(\theta) - A - B\theta_0 - BH_n h\kappa_n\| \times \hat{K}_n(\theta) \\ &\geq \sup_{\|h\|\leq 1/[2\overline{C}]}\|\bar{g}_n(\theta) - A - B\theta_0 - BH_n h\kappa_n\| \times \inf_{x\in(0,3/4)} K(x)\end{aligned}$$

Note that for $\|h\|$ bounded, we have $\theta - \theta_0 = o(1)$[14] and:

$$\bar{g}_n(\theta) - \bar{g}_n(\theta_0) - \partial_\theta g_n(\theta_0)H_n h\kappa_n = o_p(n^{-1/2}) + O_p\left(\|\partial_\theta g_n(\theta_0)H_n h\kappa_n\|^2\right).$$

By construction, we also have: $O_p\left(\|\partial_\theta g_n(\theta_0)H_n h\kappa_n\|^2\right) = O_p(\kappa_n^2) = o_p(n^{-1/2})$. Altogether, if A, B are different from:

$$A_{n,\infty} = \bar{g}_n(\theta_0) - B_{n,\infty}\theta_0 + o_p(n^{-1/2}), \quad B_{n,\infty}H_n = \partial_\theta \bar{g}_n(\theta_0)H_n + o_p\left(n^{-1/2}\kappa_n^{-1}\right)$$

by more than a $o_p(n^{-1/2})$ and a $o_p(n^{-1/2}\kappa_n^{-1})$ term respectively, the sup over $\theta \in \Theta$ is greater than a $o_p(n^{-1/2})$ which is suboptimal compared to $A_{n,\infty}, B_{n,\infty}$. To get the result in terms of $\hat{\theta}_n$ instead of θ_0, note that:

$$\begin{aligned}A_{n,\infty} &= \bar{g}_n(\theta_0) - B_{n,\infty}\theta_0 + o_p(n^{-1/2}) \\ &= \bar{g}_n(\hat{\theta}_n) - B_{n,\infty}\hat{\theta}_n + [\partial_\theta g_n(\theta_0) - B_{n,\infty}]H_n H_n^{-1}[\hat{\theta}_n - \theta_0] + o_p(n^{-1/2}) \\ &= \bar{g}_n(\hat{\theta}_n) - B_{n,\infty}\hat{\theta}_n + o_p(n^{-1/2}).\end{aligned}$$

The result for $B_{n,\infty}$ follows from Definition 4 iv. and the rate of convergence of $\hat{\theta}_n$. The asymptotic equivalence between $\hat{\theta}_n$ and $\hat{\theta}_{n,\infty}$ can be shown the same way as in the least-squares proof. This concludes the proof. □

Proof of Theorem 2. Pick $v_j \in Span(P_j)$ with $\|v_j\| = 1$. Consider $h \in \mathbb{R}$ and $\theta_{j,n} = \theta_0 + (\kappa_n^{1/j}h)v_j$. Using Definition 5, we have: $\|g_n(\theta_{j,n})/\kappa_n\|_W \geq \underline{C}_j|h|$. This implies that for $|h| \leq$

[14] This is because κ_n^2 goes to zero faster than $\lambda_{\min}\left(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right)$.

$1/[2\max_j \underline{C}_j]$, we have for $0 < \varepsilon < \|K\|_\infty$: $\mathbb{P}\left(K(\|\bar{g}_n(\theta_{j,n})/\kappa_n\|_{W_n}) \geq \varepsilon\right) \to 1$, as $n \to \infty$. As before, this statement holds uniformly over $\theta_{j,n} = \theta_0 + (\kappa_n^{1/j} h)v_j$ with $|h| \leq 1/[2\max_j \underline{C}_j]$. We also have: $\|A_{n,\infty} + B_{n,\infty}\theta_0\| = O_p(\kappa_n)$ and $\|A_{n,\infty} + B_{n,\infty}\theta_{j,n}\| = O_p(\kappa_n)$. Noting that $\|\theta_{j,n} - \theta_0\| = |h|\kappa_n^{-1/j}$, these equalities yield for $0 < |h| \leq 1/[2\max_j \underline{C}_j]$:

$$\|B_{n,\infty}(\theta_{j,n} - \theta_0)\| = O_p(\kappa_n) \Rightarrow \|B_{n,\infty}v_j\| = O_p(\kappa_n) \times \kappa_n^{-1/j}|h|.$$

This implies $\|B_{n,\infty}v_j\| = O_p(\kappa_n^{1-1/j})$ and $|\lambda_{min}(B_{n,\infty})| = O_p(\kappa_n^{1-1/r})$, concluding the proof. □

Proof of Theorem 3. First note that for any $\theta \in \Theta$, the conditions on the weight matrix W in Definition 1 and the compactness of the kernel K:

$$\begin{aligned}\|\bar{g}_n(\theta)\| \times \hat{K}_n(\theta) = \|\bar{g}_n(\theta)\| \times K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n) &\leq \|\bar{g}_n(\theta)\|_{W_n} \times K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n)/\underline{\lambda} \\ &\leq \kappa_n\|K\|_\infty/\underline{\lambda} = O(\kappa_n).\end{aligned}$$

By minimization of the sup-norm criterion:

$$\sup_{\theta\in\Theta}\left(\|A_{n,\infty} + B_{n,\infty}\theta - \bar{g}_n(\theta)\| \times \hat{K}_n(\theta)\right) \leq \sup_{\theta\in\Theta}\left(\|\bar{g}_n(\theta)\| \times \hat{K}_n(\theta)\right) \leq O(\kappa_n).$$

Pick any two $\theta_0, \theta_1 \in \Theta_0$. We have, by definition of Θ_0 and Assumption 2 for $j \in \{0, 1\}$:

$$K(\|\bar{g}_n(\theta_j)\|_{W_n}/\kappa_n) = K(\|g_n(\theta_j)/\kappa_n + \mathbb{G}_n(\theta_j)n^{-1/2}/\kappa_n\|_{W_n}) = K(o_p(1)) = K(0) + o_p(1).$$

The last equality follows from the Lipschitz-continuity of K. This implies that $K(\|\bar{g}_n(\theta_0)\|_{W_n}/\kappa_n)$ is strictly positive with probability going to 1. Note that Assumption 2 implies that this result is uniform in $\theta \in \Theta_0$. Since $K > 0$ and continuous on $(-1, 1)$, there exists $\underline{K} > 0$ and $\varepsilon > 0$ such that $|x| \leq \varepsilon \Rightarrow K(x) \geq \underline{K}$ so that:

$$\begin{aligned}\mathbb{P}(\inf_{\theta\in\Theta_0} K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n) \geq \underline{K}) &\geq \mathbb{P}(\sup_{\theta\in\Theta_0} \|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \leq \varepsilon) \\ &\geq \mathbb{P}\left(\sup_{\theta\in\Theta_0} \sqrt{n}\|g_n(\theta)\| + \sup_{\theta\in\Theta} \|\mathbb{G}_n(\theta)\| \leq \sqrt{n}\kappa_n\varepsilon/\overline{\lambda}\right) \to 1,\end{aligned}$$

as $n \to \infty$ because $\sup_{\theta\in\Theta_0} \sqrt{n}\|g_n(\theta_0)\|_W$ is finite and $\sup_{\theta\in\Theta} \|\mathbb{G}_n(\theta)\| = O_p(1)$ while $\sqrt{n}\kappa_n \to +\infty$ by assumption and $0 < \underline{\lambda} \leq \lambda_{\min}(W_n) \leq \lambda_{\max}(W_n) \leq \overline{\lambda} < +\infty$. Now, by the reverse triangular inequality, we have:

$$\sup_{\theta\in\Theta} \|A_{n,\infty} + B_{n,\infty}\theta - \bar{g}_n(\theta)\|K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n) \geq (\|A_{n,\infty} + B_{n,\infty}\theta_0\| - \|\bar{g}_n(\theta_0)\|)\,[K(0) + o_p(1)].$$

Since the term on the left-hand side is a $O(\kappa_n)$ and $\|\bar{g}_n(\theta_0)\| \geq \|\bar{g}_n(\theta_0)\|_{W_n}/\overline{\lambda} = O_p(\kappa_n)$:

$$0 \leq \|A_{n,\infty} + B_{n,\infty}\theta_0\| \leq \frac{O_p(\kappa_n)}{K(0) + o_p(1)} = O_p(\kappa_n).$$

This implies that $A_{n,\infty} = -B_{n,\infty}\theta_0$ up to a $O_p(\kappa_n)$ term. Evaluating the expression at $\theta = \theta_1$:

$$\|A_{n,\infty} + B_{n,\infty}\theta_1\| = \|O_p(\kappa_n) + B_{n,\infty}(\theta_1 - \theta_0)\| \geq \|B_{n,\infty}(\theta_1 - \theta_0)\| - O_p(\kappa_n).$$

Together with the inequalities above, this implies:

$$0 \leq \|B_{n,\infty}(\theta_1 - \theta_0)\| \leq O_p(\kappa_n) + \frac{O_p(\kappa_n)}{K(0) + o_p(1)} = O_p(\kappa_n).$$

By definition of $\lambda_{min}(B_{n,\infty})$, we have:

$$0 \leq |\lambda_{min}(B_{n,\infty})|^2 \|\theta_1 - \theta_0\|^2 \leq \|(\theta_1 - \theta_0)' B'_{n,\infty} B_{n,\infty}(\theta_1 - \theta_0)\| \leq O_p(\kappa_n^2).$$

This concludes the first part of the proof. Let $V = \text{Span}\left(\{\theta_0 - \theta_1,\, \theta_0, \theta_1 \in \Theta_0\}\right)$. Take $(\theta_{1,j} - \theta_{0,j})_{j=1,\ldots,r}$, a basis of this span with $(\theta_{1,j}, \theta_{0,j}) \in \Theta_0^2$ for all j, then for any $v \in V$ there exists $(a_1, \ldots, a_r)$ such that $v = \sum_{j=1}^r a_j(\theta_{1,j} - \theta_{0,j})$. Since the derivations above were uniform in $(\theta_0, \theta_1) \in \Theta_0^2$, we have: $0 \leq \|B_{n,\infty}v\| \leq \sum_{j=1}^r |a_j| \times \|B_{n,\infty}(\theta_{1,j} - \theta_{0,j})\| = O_p(\kappa_n)$, which concludes the proof. □

A.2 Proofs for the least-squares approximation

Proof of Theorem 1. The proof is divided into several steps.

Step 1. Proof of the results concerning $\mathbf{A_{n,LS}}$:

The least-squares formula provides a closed-form for the intercept: $A_{n,LS} = \int \bar{g}_n(\theta)\hat{\pi}_n(\theta)d\theta - B_{n,LS}\int \theta\hat{\pi}_n(\theta)d\theta$. Substituting into the least-squares objective, this means that the first part of the proof amounts to showing that:

$$\int [\bar{g}_n(\theta) - \bar{g}_n(\theta_0)]\hat{\pi}_n(\theta)d\theta = o_p(n^{-1/2}), \quad H_n^{-1}\int [\theta - \hat{\theta}_n]\hat{\pi}_n(\theta)d\theta = o_p(n^{-1/2})$$

for both the compact and the exponential kernels.

Step 1.a. Compact Kernel K:

For $\|\theta - \theta_0\| \geq \varepsilon > 0$, as in the proof under the sup-norm, we have: $\hat{\pi}_n(\theta) \leq O_p(n^{-d})$ for any $d \geq 1$. Then for $\|\theta - \theta_0\| \leq \varepsilon$, as in the proof with the sup-norm, we have after the following re-parametrization $\theta = \hat{\theta}_n + H_n h\kappa_n$, with $H_n = \left(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)\right)^{-1/2}$:

$$\sup_{h \in \mathbb{R}^{d_\theta}, \|h\| \geq 4/\underline{C}} K\left(\|\bar{g}_n(\theta)\|_{W_n} - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}\right) = 0, \text{ with probability going to 1.}$$

$\theta - \hat{\theta}_n = H_n h \kappa_n = o_p(1)$ implies that, with probability going to 1:

$$\begin{aligned}\int_\Theta [\bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_n)]\hat{\pi}_n(\theta)d\theta &= \int_{\theta,\|h\|\leq 4/\underline{C}} [\bar{g}_n(\theta) - \bar{g}_n(\hat{\theta}_n)]\hat{\pi}_n(\theta)d\theta \\ &= \int_{\theta,\|h\|\leq 4/\underline{C}} \partial_\theta g_n(\theta_0)[\theta - \theta_0]\hat{\pi}_n(\theta)d\theta + o_p(n^{-1/2}) + O_p(\kappa_n^2)\end{aligned}$$

where the expansion of $\bar{g}_n(\theta)$ is derived using the stochastic equicontinuity and smoothness assumptions. To show that $\bar{\theta}_n - \hat{\theta}_n = o_p(n^{-1/2})$, expand the terms inside the kernel:

$$\begin{aligned}\bar{g}_n(\theta) &= \bar{g}_n(\hat{\theta}_n) + \partial_\theta g_n(\hat{\theta}_n) H_n h \kappa_n + o_p(n^{-1/2}) + O_p\left(\kappa_n^2\right) \\ &= \bar{g}_n(\hat{\theta}_n) + \partial_\theta g_n(\theta_0) H_n h \kappa_n + o_p(n^{-1/2}) + O_p\left(\kappa_n^2\right)\end{aligned}$$

where the last equality is due to Definition 4 iv. Since the kernel K is Lipschitz-continuous:

$$\left| K\left(\|\partial_\theta g_n(\hat{\theta}_n) H_n h\|_{\hat{W}_n}\right) - K\left(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}/\kappa_n\right) \right| = o_p(n^{-1/2}\kappa_n^{-1}) + O_p(\kappa_n).$$

This implies that:

$$\begin{aligned}H_n^{-1}[\bar{\theta}_n - \hat{\theta}_n] &= \frac{\int_{\|h\|\leq 4/\underline{C}} h\kappa_n [K\left(\|\partial_\theta g_n(\hat{\theta}_n) H_n h\|_{\hat{W}_n}\right) + o_p(n^{-1/2}\kappa_n^{-1} + O_p(\kappa_n))]}{\int_{\|h\|\leq 4/\underline{C}} [K\left(\|\partial_\theta g_n(\hat{\theta}_n) H_n h\|_{\hat{W}_n}\right) + o_p(n^{-1/2}\kappa_n^{-1} + O_p(\kappa_n))]} \\ &= 0 + o_p(n^{-1/2}) + O_p(\kappa_n^2) = o_p(n^{-1/2}),\end{aligned}$$

because $K\left(\|\partial_\theta g_n(\hat{\theta}_n) H_n h\|_{\hat{W}_n}\right)$ is symmetric in h. Now, $\int [\bar{g}_n(\theta) - \bar{g}_n(\theta_0)]\hat{\pi}_n(\theta)d\theta = o_p(n^{-1/2}) = \partial_\theta g_n(\theta_0)[\bar{\theta}_n - \hat{\theta}_n] + o_p(n^{-1/2})$ which implies the other result.

Step 1.b. Exponential Kernel K:

Using $\theta = \hat{\theta}_n + H_n h \kappa_n$, $H_n = (\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))^{-1/2}$, we have for $\|h\| \geq [d_\theta \log(n)/\underline{C}_2]^{1/a}/\underline{C}$: $\hat{K}_n(\theta) \leq C_1 \exp\left(-d_\theta \log(n) + o_p(1)\right) = O_p(n^{-d_\theta}) = o_p(n^{-1/2})$, since $d_\theta \geq 1$. Assumption 1 also implies that $\tilde{\kappa}_n = \kappa_n \log(n)^{1/a} = o(1)$ so that the stochastic equicontinuity result applies in balls centered around $\hat{\theta}_n$ with radius proportional to $\tilde{\kappa}_n$:

$$\bar{g}_n(\theta) = \bar{g}_n(\hat{\theta}_n) + \partial_\theta g_n(\hat{\theta}_n) H_n h \kappa_n + o_p(n^{-1/2}) + O_p(\|\partial_\theta g_n(\theta_0) H_n h \kappa_n\|^2).$$

Since, up to a fixed constant, we have $\|h\| \leq \log(n)^{1/a}$, the last term is a $O_p(\tilde{\kappa}_n^2) = o_p(n^{-1/2})$. As for the compact kernel: $K\left(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}/\kappa_n\right) = K\left(\|\partial_\theta g_n(\hat{\theta}_n) H_n h\|_{\hat{W}_n}\right)\Big[1 + o_p(n^{-1/2}\kappa_n^{-1}) + O_p(\kappa_n \|h\|^2)\Big]$. Note that $\partial_\theta g_n(\hat{\theta}_n) H_n = \partial_\theta g_n(\theta_0) H_n + o_p(1)$; since the first d_θ eigenvalues of $\partial_\theta g_n(\theta_0) H_n$ are bounded below by construction (this is the role of H_n), with

probability going to 1, $\partial_\theta g_n(\hat{\theta}_n)H_n$ is also non-singular. This implies that:

$$
\begin{aligned}
&H_n^{-1}[\bar{\theta}_n - \hat{\theta}_n] \\
&= \kappa_n \frac{\int_{\|h\| \leq C_K \log(n)^{1/a}} hK(\|\partial_\theta g_n(\hat{\theta}_n)H_n h\|_{\hat{W}_n}) \Big[1 + o_p(n^{-1/2}\kappa_n^{-1}) + O_p(\kappa_n \|h\|^2)\Big] dh}{\int_{\|h\| \leq C_K \log(n)^{1/a}} K(\|\partial_\theta g_n(\hat{\theta}_n)H_n h\|_{\hat{W}_n}) \Big[1 + o_p(n^{-1/2}\kappa_n^{-1}) + O_p(\kappa_n \|h\|^2)\Big] dh} + o_p(n^{-1/2}) \\
&= \frac{o_p(n^{-1/2}) + O_p(\kappa_n^2)}{\int_{\|h\| \leq C_K \log(n)^{1/a}} K(\|\partial_\theta g_n(\hat{\theta}_n)H_n h\|_{\hat{W}_n}) dh + o_p(1)} + o_p(n^{-1/2}) = o_p(n^{-1/2}).
\end{aligned}
$$

The other result can be proved the same way as for the compact kernel.

Step 2. Proof of the results concerning $\mathbf{B_{n,LS}}$:

The least-squares formula implies that:

$$
B_{n,LS} = \left(\int [\theta - \bar{\theta}_n][\theta - \bar{\theta}_n]' \hat{\pi}_n(\theta) d\theta \right)^{-1} \int [\theta - \bar{\theta}_n][\bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta}]' \hat{\pi}_n(\theta) d\theta.
$$

Using the change of variable: $\theta = \hat{\theta}_n + H_n h \kappa_n$, the first term can be re-written as:

$$
\int [\theta - \bar{\theta}_n][\theta - \bar{\theta}_n]' \hat{\pi}_n(\theta) d\theta = \int_{\|h\| \leq 4/\underline{C}} [\kappa_n^2 H_n h h' H_n + o_p(\kappa_n n^{-1/2})] \kappa_n^{d_\theta} \hat{\pi}_n(\hat{\theta}_n + H_n \kappa_n h) dh.
$$

Note that $O(n^{1/2}) \geq H_n \geq O_p(1)$ so that the $o_p(\kappa_n n^{-1/2})$ term is negligible compared to $\kappa_n^2 H_n h h' H_n$. The second term can be expanded as:

$$
\begin{aligned}
&\int [\theta - \bar{\theta}_n][\bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta}]' \hat{\pi}_n(\theta) d\theta \\
&= \int_{\|h\| \leq 4/\underline{C}} \left([\kappa_n^2 H_n h h' H_n' \partial_\theta g_n(\hat{\theta}_n)'] + [H_n o_p(\kappa_n n^{-1/2})] \right) \kappa_n^{d_\theta} \hat{\pi}_n(\hat{\theta}_n + H_n \kappa_n h) dh
\end{aligned}
$$

Together these imply that $H_n B_{n,LS}' = H_n \partial_\theta g_n(\hat{\theta}_n)' + o_p(n^{-1/2}\kappa_n^{-1})$.

Final Step, showing that: $\sqrt{n} H_n^{-1/2} \left(\hat{\theta}_n - \theta_{n,LS} \right) = o_p(1)$:

Since $\hat{\theta}_n$ is consistent, the re-parameterization $\theta_n = \theta_0 + H_n h n^{-1/2}$ implies:

$$
n \times \|\bar{g}_n(\theta_n)\|_{W_n}^2 = \left(\mathbb{G}_n(\theta_0) + \partial_\theta g_n(\theta_0) H_n h + o_p(1) \right)' \left[\hat{W}_n + o_p(1) \right] \left(\mathbb{G}_n(\theta_0) + \partial_\theta g_n(\theta_0) H_n h + o_p(1) \right).
$$

The Argmax Theorem (van der Vaart & Wellner, 1996) implies that the minimizer is $\hat{h}_n = \left(H_n \partial_\theta g_n(\theta_0)' \hat{W}_n \partial_\theta g_n(\theta_0) H_n \right)^{-1} H_n \partial_\theta g_n(\theta_0)' \hat{W}_n \left[\mathbb{G}_n(\theta_0) + o_p(1) \right]$. Since $\hat{\theta}_n = \theta_0 + H_n \hat{h}_n n^{-1/2}$:

$$
\sqrt{n} H_n^{-1} (\hat{\theta}_n - \theta_0) = \left(H_n \partial_\theta g_n(\theta_0)' \hat{W}_n \partial_\theta g_n(\theta_0) H_n \right)^{-1} H_n \partial_\theta g_n(\theta_0)' \hat{W}_n \mathbb{G}_n(\theta_0) + o_p(1).
$$

The estimator $\hat{\theta}_{n,LS}$ can be written as:

$$\hat{\theta}_{n,LS} = \hat{\theta}_n - \left(B'_{n,LS}\hat{W}_n B_{n,LS}\right)^{-1} B'_{n,LS}\hat{W}_n[\bar{g}_n(\hat{\theta}_n) + o_p(n^{-1/2})]$$

which implies that the difference with $\hat{\theta}_n$ can be written as:

$$\begin{aligned}
&\sqrt{n}H_n^{-1}(\hat{\theta}_n - \hat{\theta}_{n,LS}) \\
&= \sqrt{n}H_n^{-1}\left(B'_{n,LS}\hat{W}_n B_{n,LS}\right)^{-1} B'_{n,LS}\hat{W}_n[\bar{g}_n(\hat{\theta}_n) + o_p(n^{-1/2})] \\
&= \sqrt{n}\left(H_n\partial_\theta g_n(\theta_0)'\hat{W}_n\partial_\theta g_n(\theta_0)H_n\right)^{-1} H_n\partial_\theta g_n(\theta_0)'\hat{W}_n[\bar{g}_n(\hat{\theta}_n) + o_p(n^{-1/2})](1 + o_p(1)).
\end{aligned}$$

$\bar{g}_n(\hat{\theta}_n)$ can be expanded into:

$$\left(I - \partial_\theta g_n(\theta_0)H_n\left(H_n\partial_\theta g_n(\theta_0)'\hat{W}_n\partial_\theta g_n(\theta_0)H_n\right)^{-1} H_n\partial_\theta g_n(\theta_0)'\hat{W}_n + o_p(1)\right)\bar{g}_n(\theta_0) + o_p(n^{-1/2}).$$

The term before $\bar{g}_n(\theta_0)$ is a projection matrix orthogonal to $H_n\partial_\theta g_n(\theta_0)'\hat{W}_n$, hence $H_n\partial_\theta g_n(\theta_0)'\hat{W}_n\bar{g}_n(\hat{\theta}_n) = o_p(n^{-1/2}) \Rightarrow \sqrt{n}H_n^{-1}(\hat{\theta}_n - \hat{\theta}_{n,LS}) = o_p(1)$ which concludes the proof. □

Proof of Lemma 1. From the least-squares formula, we have $A_{n,LS} = \int \bar{g}_n(\theta)\hat{\pi}_n(\theta)d\theta - B_{n,LS}\bar{\theta}_n$. Substituting into the least-squares objective implies that $B_{n,LS}$ minimizes:

$$\int \left\| B_{n,LS}(\theta - \bar{\theta}_n) - \left[\bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta})\hat{\pi}_n(\tilde{\theta})d\tilde{\theta}\right]\right\|^2 \hat{\pi}_n(\theta)d\theta.$$

The proof proceed in 3 steps:

1. Proving that $\int \|\bar{g}_n(\theta)\|^2\hat{\pi}_n(\theta)d\theta \leq O_p(\tilde{\kappa}_n^2)$ for the compact and the exponential kernels.
2. Proving that $\int \left\| B_{n,LS}(\theta - \bar{\theta}_n)\right\|^2 \hat{\pi}_n(\theta)d\theta \leq O_p(\tilde{\kappa}_n^2)$.
3. Deducing the results from steps 1 and 2.

Step 1. Proving that $\int \|\bar{g}_n(\theta)\|^2\hat{\pi}_n(\theta)d\theta\| \leq O_p(\tilde{\kappa}_n^2)$:
Step 1.a. Compact Kernel K:
For the compact kernel, as in the sup-norm proof: $\|\bar{g}_n(\tilde{\theta})\|^2\hat{\pi}_n(\tilde{\theta}) \leq O(\kappa_n^2)\hat{\pi}_n(\tilde{\theta})$.
Step 1.b. Exponential Kernel K:
For the exponential kernel, pick $d \geq 1$ and consider a θ such that $\|\bar{g}_n(\theta)\|_{W_n} \geq \tilde{\kappa}_n \times [d/C_2]^{1/a}$:

$$\hat{K}_n(\theta) \leq C_1 \exp\left(-d[\tilde{\kappa}_n/\kappa_n]^a\right) = C_1 \exp\left(-d\log(n)\right) = C_1 n^{-d},$$

this implies for these values of θ that $\|\bar{g}_n(\theta)\|^2 \times \hat{K}_n(\theta) \leq \overline{\lambda} \times \left[\sup_{\theta \in \Theta} \|g_n(\theta)\|_W + O_p(n^{-1/2})\right]^2 \times C_1 n^{-d} = O_p(n^{-d})$. Now pick any $\theta_0 \in \Theta_0$, the Hölder-type condition implies that for any θ such that $\|\theta - \theta_0\| \leq \kappa_n^{1/\varsigma}$, we have $\|g_n(\theta)\|_W \leq \overline{C}\kappa_n$, which implies for the sample moments: $\|\bar{g}_n(\theta)\|_{W_n} \leq \overline{C}\kappa_n + O_p(n^{-1/2}) \leq (1+\overline{C})\kappa_n$, with probability going to 1. This implies that $\hat{K}_n(\theta) \geq C_1 \exp\Big(- C_2[1+\overline{C}]^a\Big)$ with probability going to 1, and:

$$\int_{\|\theta - \theta_0\| \leq \kappa_n^{1/\varsigma}} \hat{K}_n(\theta) d\theta \geq \kappa_n^{d_\theta/\varsigma} \frac{\pi^{d_\theta/2}}{\Gamma(d_\theta/2+1)} C_1 \exp\left(-C_2[1+\overline{C}]^a\right).$$

Together with the first bound, this implies that:

$$\int_{\theta, \|\bar{g}_n(\theta)\|_{W_n} \geq \tilde{\kappa}_n \times [dC_2]^{1/a}} \hat{\pi}_n(\theta) d\theta \leq \kappa_n^{d_\theta/\varsigma} n^{-d} \frac{\pi^{d_\theta/2}}{\Gamma(d_\theta/2+1)} \exp\left(-C_2[1+\overline{C}]^a\right).$$

For d large enough, $\kappa_n^{d_\theta/\varsigma} n^{-d} \leq O_p(\tilde{\kappa}_n^2)$. Putting everything together, we get:

$$\begin{aligned}
&\int_{\theta \in \Theta} \|\bar{g}_n(\theta)\|^2 \hat{\pi}_n(\theta) d\theta \leq \underline{\lambda}^{-2} \int_{\theta \in \Theta} \|\bar{g}_n(\theta)\|^2_{W_n} \hat{\pi}_n(\theta) d\theta \\
&= \underline{\lambda}^{-2} \Big[\int_{\theta, \|\bar{g}_n(\theta)\|_{W_n} \leq \tilde{\kappa}_n [d/C_2]^{1/a}} \|\bar{g}_n(\theta)\|^2_{W_n} \hat{\pi}_n(\theta) d\theta + \int_{\theta, \|\bar{g}_n(\theta)\|_{W_n} > \tilde{\kappa}_n [d/C_2]^{1/a}} \|\bar{g}_n(\theta)\|^2_{W_n} \hat{\pi}_n(\theta) d\theta \Big] \\
&\leq \underline{\lambda}^{-2} \left[\tilde{\kappa}_n^2 [d/C_2]^{2/a} + O_p(\tilde{\kappa}_n^2) \right] = O_p(\tilde{\kappa}_n^2).
\end{aligned}$$

Step 2. Proving that $\int \left\| B_{n,LS}(\theta - \bar{\theta}_n) \right\|^2 \hat{\pi}_n(\theta) d\theta \leq O_p(\tilde{\kappa}_n^2)$:
From step 1., we have:

$$\int \left\| \bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta} \right\|^2 \hat{\pi}_n(\theta) d\theta \leq 4 \int \left\| \bar{g}_n(\theta) \right\|^2 \hat{\pi}_n(\theta) d\theta \leq O_p(\tilde{\kappa}_n^2).$$

Since $B_{n,LS}$ minimizes the least-squares criterion:

$$\begin{aligned}
&\int \left\| B_{n,LS}(\theta - \bar{\theta}_n) - \left[\bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta} \right] \right\|^2 \hat{\pi}_n(\theta) d\theta \\
&\leq \int \left\| \bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta} \right\|^2 \hat{\pi}_n(\theta) d\theta \leq O_p(\tilde{\kappa}_n^2).
\end{aligned}$$

By the reverse triangle inequality:

$$\begin{aligned}
\int \left\| B_{n,LS}(\theta - \bar{\theta}_n) \right\|^2 \hat{\pi}_n(\theta) d\theta \leq &\int \left\| B_{n,LS}(\theta - \bar{\theta}_n) - \left[\bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta} \right] \right\|^2 \hat{\pi}_n(\theta) d\theta \\
&+ \int \left\| \bar{g}_n(\theta) - \int \bar{g}_n(\tilde{\theta}) \hat{\pi}_n(\tilde{\theta}) d\tilde{\theta} \right\|^2 \hat{\pi}_n(\theta) d\theta \leq O_p(\tilde{\kappa}_n^2)
\end{aligned}$$

Step 3. Deducing the results of Lemma 1:
By definition of the Frobenius norm: $\int \text{trace}\left(B_{n,LS}(\theta-\bar{\theta}_n)(\theta-\bar{\theta}_n)'B_{n,LS}'\right)\hat{\pi}_n(\theta)d\theta = \int \|B_{n,LS}(\theta-\bar{\theta}_n)\|^2\hat{\pi}_n(\theta)d\theta \leq O_p(\tilde{\kappa}_n^2)$. By linearity of the trace and integral operators:

$$\text{trace}\left(B_{n,LS}\Sigma_n B_{n,LS}'\right) = \text{trace}\left(B_{n,LS}\left[\int(\theta-\bar{\theta}_n)(\theta-\bar{\theta}_n)'\hat{\pi}_n(\theta)d\theta\right]B_{n,LS}'\right) \leq O_p(\tilde{\kappa}_n^2).$$

In turn, this implies that $\text{trace}\left(B_{n,LS}'B_{n,LS}\Sigma_n\right) \leq O_p(\tilde{\kappa}_n^2)$. Since Σ_n and $B_{n,LS}'B_{n,LS}$ are Hermitian (self-adjoint), Problem III.6.14, in Bhatia (1997)[15] implies that:

$$0 \leq \lambda_j(B_{n,LS}'B_{n,LS})\lambda_{d_\theta+1-j}(\Sigma_n) \leq \text{trace}\left(B_{n,LS}'B_{n,LS}\Sigma_n\right) = O_p(\tilde{\kappa}_n^2)$$

which, in turn, implies the desired results.

Let $v_{j,n}$ be an eigenvector of Σ_n, then $\Sigma_n^{1/2}v_{j,n} = \sqrt{\lambda_{j,n}}v_{j,n}$. Furthermore, $\|B_{n,LS}\Sigma_n^{1/2}\| = O_p(\tilde{\kappa}_n)$. Together these imply $0 \leq \sqrt{\lambda_{j,n}}\|B_{n,LS}v_{j,n}\| = \|B_{n,LS}\Sigma_n^{1/2}\| = O_p\left(\tilde{\kappa}_n\right)$. Since $\sqrt{\lambda_{j,n}} = O_p(r_{j,n})$ this implies the final result and concludes the proof. □

Proof of Theorem 2. Using Lemma 1, the proofs amounts to showing that Σ_n satisfies $P_j\Sigma_nP_j' = O_p(\tilde{\kappa}_n^{2/j})$ for each $j = 1,\ldots,r$. Before we proceed, note that:

$$\Sigma_n = \int_\Theta(\theta-\bar{\theta}_n)(\theta-\bar{\theta}_n)'\hat{\pi}_n(\theta)d\theta = (\bar{\theta}_n-\hat{\theta}_n)(\bar{\theta}_n-\hat{\theta}_n)' + \int_\Theta(\theta-\hat{\theta}_n)(\theta-\hat{\theta}_n)'\hat{\pi}_n(\theta)d\theta.$$

We need to show that $P_j(\bar{\theta}_n-\hat{\theta}_n) = O_p(\tilde{\kappa}_n^{1/j})$ and the posterior concentrates around $\hat{\theta}_n$ at a $\tilde{\kappa}_n^{1/j}$ rate in each direction P_j. This is shown separately for compact and exponential kernels.

We also need to derive the rate for $\hat{\theta}_n - \theta_0$. $\hat{\theta}_n$ is consistent by Theorem 2.1 in Newey & McFadden (1994). Re-write $\theta = \theta_0 + \sum_{j=1}^r n^{-1/[2j]}P_jh$, $h \in \mathbb{R}^{d_\theta}$; using $\sup_{\theta\in\Theta}\|\mathbb{G}_n(\theta)\|_W = O_p(1)$: $\left[\sum_{j=1}^r \overline{C}_j\|P_jh\|^j + O_p(1)\right]^2/[1+o_p(1)] \geq n\times\|\bar{g}_n(\theta_0+\sum_{j=1}^r n^{-1/2j}P_jh)\|_{W_n}^2 \geq 0$. which implies that the minimizer $\hat{h}_n$ of $\|\bar{g}_n(\theta_0+\sum_{j=1}^r n^{-1/2j}P_jh)\|_{W_n}^2$ is a $O_p(1)$. This in turn implies that $P_j(\hat{\theta}_n-\theta_0) = O_p(n^{-\frac{1}{2j}})$ for each $j \in \{1,\ldots,r\}$.

Step 1.a. Results for the Compact Kernel K:

[15]The result invoked from Bhatia (1997) is a consequence of Lidskii's theorems and states that if A, B are Hermitian then $\langle\lambda^\downarrow(A),\lambda^\uparrow(B)\rangle \leq \text{trace}\,(AB)$ where $\lambda^\downarrow,\lambda^\uparrow$ are the eigenvalues in decreasing and increasing order, respectively. Furthermore, if the matrices are positive semi-definite, then the eigenvalues of each matrix are positive so that the result implies $0 \leq \lambda_j(A)\lambda_{d_B+1-j}(B) \leq \text{trace}(AB)$.

Using the same approach as above, we can re-write $\theta = \hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h$ with $h \in \mathbb{R}^{d_\theta}$ and:

$$\begin{aligned}
\|\bar{g}_n(\theta)/\kappa_n\|_{W_n} &\geq \left[\sum_{j=1}^r \underline{C}_j \|P_j(\hat{\theta}_n - \theta_0)\kappa_n^{-1/j} + P_j h\|^j - O_p(n^{-1/2}\kappa_n^{-1})\right] / [1 + o_p(1)] \\
&= \left[\sum_{j=1}^r \underline{C}_j \|O_p(|n^{-1/2}\kappa_n^{-1}|^{1/j}) + P_j h\|^j - O_p(n^{-1/2}\kappa_n^{-1})\right] / [1 + o_p(1)] \\
&= \sum_{j=1}^r \underline{C}_j \|P_j h\|^j - o_p(1) \geq 2,
\end{aligned}$$

with probability going to 1 when $\sum_{j=1}^r \|P_j h\|^j \geq 2/\underline{C}$, $\underline{C} = \min_{j=1,\ldots,r} C_j > 0$.[16] Then we have, with probability going to 1: $\sup_{h, \sum_{j=1}^r \|P_j h\|^j \geq 2/\underline{C}} \hat{K}_n(\hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h) = 0$. This, in turn implies that, with probability going to 1:

$$\bar{\theta}_n = \int_\Theta \theta \hat{\pi}_n(\theta) d\theta = \hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j \left(\int_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} h \hat{\pi}_n(\hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h) dh \right).$$

Since the integral is taken over a bounded set, we have:

$$\| \int_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} h \hat{\pi}_n(\hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h) dh \| \leq \sup_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} \|h\| < +\infty.$$

This implies that, with probability going to 1: $P_j(\bar{\theta}_n - \hat{\theta}_n) = \kappa_n^{1/j} O_p(1)$. Since $\bar{\theta}_n - \hat{\theta}_n = \sum_{j=1}^r P_j(\bar{\theta}_n - \hat{\theta}_n)$, we have $(\bar{\theta}_n - \hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)' = \sum_{j=1}^r \kappa_n^{2/j} P_j O_p(1) P_j'$. Pre and post-multiplying this equation by P_j implies: $P_j(\bar{\theta}_n - \hat{\theta}_n)(\bar{\theta}_n - \hat{\theta}_n)' P_j' = \kappa_n^{2/j} O_p(1)$. With probability going to 1, we also have:

$$\int_\Theta (\theta - \hat{\theta}_n)(\theta - \hat{\theta}_n)' \hat{\pi}_n(\theta) d\theta = \sum_{j=1}^r \kappa_n^{2/j} P_j \left(\int_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} h h' \hat{\pi}_n(\hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h) dh \right) P_j'$$

and, using the same argument as above:

$$\| \int_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} h h' \hat{\pi}_n(\hat{\theta}_n + \sum_{j=1}^r \kappa_n^{1/j} P_j h) dh \| \leq \sup_{h, \sum_{j=1}^r \|P_j h\|^j \leq 2/\underline{C}} \|h h'\| < +\infty.$$

This implies that with probability going to 1, we have: $\int_\Theta (\theta - \hat{\theta}_n)(\theta - \hat{\theta}_n)' \hat{\pi}_n(\theta) d\theta = \sum_{j=1}^r \kappa_n^{2/j} P_j O_p(1) P_j'$. Putting everything together, we have $\Sigma_n = \sum_{j=1}^r \kappa_n^{2/j} P_j O_p(1) P_j'$ so that for any $v_j \in \text{Span}(P_j)$, $v_j \Sigma_n v_j' = O_p(\kappa_n^{2/j})$ which by Lemma 1 implies $|\lambda_{\min}(B_{n,LS})| =$

[16] If the minimum is zero, one can take the minimum over the non-zero elements instead.

$O_p(\kappa_n^{1-1/r})$ and $B_{n,LS}v_j = O_p(\kappa_n^{1-1/j})$ for any $v_j \in \text{Span}(P_j)$.

Step 1.b. Results for the Exponential Kernel K:

For the exponential kernel, consider the re-parametrization $\theta = \hat{\theta}_n + \sum_{j=1}^r \tilde{\kappa}_n^{1/j} P_j h$. This implies that, uniformly in h:

$$\|\bar{g}_n(\hat{\theta}_n + \sum_{j=1}^r \tilde{\kappa}_n^{1/j} P_j h)/\kappa_n\|_{W_n} \geq \left[\sum_{j=1}^r \underline{C}_j \|P_j h\|^j \log(n)^{1/a} - o_p(1)\right]/[1 + o_p(1))].$$

Using the monotonicity of the exponential kernel, this implies:

$$\hat{K}_n(\hat{\theta}_n + \sum_{j=1}^r \tilde{\kappa}_n^{1/j} P_j h) \leq C_1 \exp\left(-C_2\left[[\sum_{j=1}^r \underline{C}_j \|P_j h\|^j]^a \log(n) - o_p(1)\right]\right) = O_p(n^{-4})$$

uniformly in h such that $[\sum_{j=1}^r \underline{C}_j \|P_j h\|^j]^a \geq 4/C_2$. Using the same approach as for the compact kernel, we have:

$$\bar{\theta}_n = \hat{\theta}_n + \sum_{j=1}^r \tilde{\kappa}_n^{1/j} P_j \left[\int_{h, [\sum_{j=1}^r \underline{C}_j \|P_j h\|^j]^a \geq 4/(C_2)} h \hat{\pi}_n(\hat{\theta}_n + \sum_{j=1}^r \tilde{\kappa}_n^{1/j} P_j h) dh\right] + O_p(n^{-4}).$$

Since $n^{-4} = o(\kappa_n) = o(\tilde{\kappa}_n)$, this implies that $P_j(\bar{\theta}_n - \hat{\theta}_n) = O_p(\tilde{\kappa}_n^{1/j})$. Similarly: $\int_\Theta (\theta - \hat{\theta}_n)(\theta - \hat{\theta}_n)' \hat{\pi}_n(\theta) d\theta = \sum_{j=1}^r \kappa_n^{2/j} P_j O_p(1) P_j' + O_p(n^{-4})$. Putting everything together, we have $\Sigma_n = \sum_{j=1}^r \tilde{\kappa}_n^{2/j} P_j O_p(1) P_j'$ so that for any $v \in \text{Span}(P_r)$, $v\Sigma_n v' = O_p(\tilde{\kappa}_n^{2/r})$ which by Lemma 1 implies $|\lambda_{\min}(B_{n,LS})| = O_p(\tilde{\kappa}_n^{1-1/r})$ and $B_{n,LS} v_j = O_p(\tilde{\kappa}_n^{1-1/j})$ for any $v_j \in \text{Span}(P_j)$. This concludes the proof. □

Proof of Proposition 1. Let $\hat{\Pi}_n$ be the probability function associated with $\hat{\pi}_n$

$$\hat{\Pi}_n\left(\|\theta - \bar{\theta}_n\| > \varepsilon/6\right) \geq \min\left(\int_{\|\theta - \theta_0\|_2 \leq \varepsilon/3} \hat{\pi}_n(\theta) d\theta, \int_{\|\theta - \theta_1\|_2 \leq \varepsilon/3} \hat{\pi}_n(\theta) d\theta\right) \geq \eta + o_p(1)$$

since $\bar{\theta}_n$ has always a distance of at least $\varepsilon/6$ from either one of these two $\varepsilon/3$ balls. By Chebyshev's inequality: $\hat{\Pi}_n\left(\|\theta - \bar{\theta}_n\| > \varepsilon/6\right) = \hat{\Pi}_n\left(\|\theta - \bar{\theta}_n\|^2 > \varepsilon^2/36\right) \leq \frac{36}{\varepsilon^2} \times \text{trace}(\Sigma_n)$. Hence: $\frac{\eta\varepsilon^2}{36} + o_p(1) \leq \text{trace}(\Sigma_n) \leq d_\theta \lambda_{\max}(\Sigma_n)$. Since $\eta\varepsilon^2/[36 d_\theta] + o_p(1) > 0$ with probability going to 1, by Lemma 1: $\lambda_{\min}(B_{n,LS}' B_{n,LS}) = O_p(\tilde{\kappa}_n^2)$, which concludes the proof. □

Proof of Theorem 4. Each case described in the Theorem will be treated separately.

Case i. Θ_0 has non-empty interior:

By assumption, there exists $\varepsilon > 0$ and $\theta_0 \in \Theta_0$ such that $\mathcal{B}_{5\varepsilon/3}(\theta_0) \subseteq \Theta_0$. In this open ball $\mathcal{B}_{5\varepsilon/3}(\theta_0)$, we can find two θ_1, θ_2 such that $\|\theta_1 - \theta_2\| = \varepsilon$ and $\mathcal{B}_{\varepsilon/3}(\theta_1) \subseteq \Theta_0$, $\mathcal{B}_{\varepsilon/3}(\theta_2) \subseteq \Theta_0$.

Then, by definition of the weakly identified set and Assumptions 2, 1: $\sup_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\|\bar{g}_n(\theta)\|_W/\kappa_n = o_p(1)$. Since both the compact and exponential kernels are continuous: $\sup_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\hat{K}_n(\theta) = \sup_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}/\kappa_n) = K(O_p(n^{-1/2}\kappa_n^{-1})) = K(0)+o_p(1)$. Note that $K(0)>0$ by assumption. Now the integral of interest can be bounded using:

$$\begin{aligned}\frac{\text{vol}(\mathcal{B}_{\varepsilon/3}(\theta_1))K(0)+o_p(1)}{\text{vol}(\Theta)\|K\|_\infty} &\leq \frac{\int_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\hat{K}_n(\theta)d\theta}{\int_{\theta\in\Theta}\hat{K}_n(\theta)d\theta} = \int_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\hat{\pi}_n(\theta)d\theta \\ &\leq \frac{\int_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\hat{K}_n(\theta)d\theta}{\int_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_1)}\hat{K}_n(\theta)d\theta+\int_{\theta\in\mathcal{B}_{\varepsilon/3}(\theta_2)}\hat{K}_n(\theta)d\theta} = \frac{1}{2}+o_p(1)\end{aligned}$$

where $\text{vol}(\mathcal{B}_{\varepsilon/3}(\theta_1)) = \int_{\mathcal{B}_{\varepsilon/3}(\theta_1)}d\theta$. Proposition 1 is satisfied for ε and $\eta = \frac{\text{vol}(\mathcal{B}_{\varepsilon/3}(\theta_1))K(0)}{\text{vol}(\Theta)\|K\|_\infty} > 0$.

Case ii. Θ_0 finite:

Pick $\varepsilon = \min_{\theta_1,\theta_2\in\Theta_0,\theta_1\neq\theta_2}\|\theta_1-\theta_2\|$. Uniformly in $\theta\notin\cup_{j=0}^k\mathcal{B}_{\varepsilon/3}(\theta_j)$:

$$\|\bar{g}_n(\theta)\|_{W_n} \geq \left[\eta(\varepsilon/3) - \frac{\sup_{\theta\in\Theta}\|\mathbb{G}_n(\theta)\|}{\underline{\lambda}\sqrt{n}}\right]/[1+o_p(1))] = \eta(\varepsilon/3) - O_p(n^{-1/2}).$$

Case ii.a. Compact Kernel:

Assumption 1 implies that for the compact kernel $\sup_{\theta\notin\cup_{j=0}^k\mathcal{B}_{\varepsilon/3}(\theta_j)}\hat{K}_n(\theta)=0$ with probability going to 1 and for any $j\in\{0,\ldots,k\}$, $\theta\in\mathcal{B}_{\varepsilon/3}(\theta_j)$: $\underline{C}\|\theta-\theta_j\|^r/\kappa_n - o_p(1) \leq \|\bar{g}_n(\theta)\|_W/\kappa_n \leq \overline{C}\|\theta-\theta_j\|^r/\kappa_n + o_p(1)$. Using the re-parameterization $\theta=\theta_j+h\kappa_n^{1/r}$: $\underline{C}\|h\|^r - o_p(1) \leq \|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \leq \overline{C}\|h\|^r + o_p(1)$ so that $\hat{K}_n(\theta)=0$ with probability going to 1 uniformly in $\|h\|^r \geq 2/\underline{C}$ and $\hat{K}_n(\theta)\geq \inf_{x\in[0,3/4]}K(x)+o_p(1)$ uniformly in $\|h\|^r\leq 1/[2\overline{C}]$. This implies, with probability going to 1:

$$0<\kappa_n^{d_\theta/r}(\inf_{x\in[0,3/4]}K(x)+o_p(1))\int_{\|h\|^r\leq 1/[2\overline{C}]}dh \leq \int_{\mathcal{B}_{\varepsilon/3}(\theta_j)}\hat{K}_n(\theta)d\theta \leq \kappa_n^{d_\theta/r}\|K\|_\infty\int_{\|h\|^r\leq 2/\underline{C}}dh,$$

which is finite. This implies, with probability going to 1, $\forall j\in\{0,\ldots,k\}$, the desired result:

$$\int_{\mathcal{B}_{\varepsilon/3}(\theta_j)}\hat{\pi}_n(\theta)d\theta \geq \frac{\inf_{x\in[0,3/4]}K(x)}{\|K\|_\infty}\times\frac{\text{vol}(\mathcal{B}_{1/[2\overline{C}]^{1/r}}(0))}{\text{vol}(\mathcal{B}_{[2/\underline{C}]^{1/r}}(0))}+o_p(1)>0$$

Case ii.b. Exponential Kernel:

For the exponential kernel, we have: $\sup_{\theta\notin\cup_{j=0}^k\mathcal{B}_{\varepsilon/3}(\theta_j)}\hat{K}_n(\theta)\leq C_1\exp\left[-C_2[\eta(\varepsilon/3)]^a\kappa_n^{-a}+o_p(1)\right]$ which is a $o_p(\kappa_n^d)$ for any $d\geq 1$.[17] Using the re-parameterization $\theta=\theta_j+h\tilde{\kappa}_n^{-1/r}$, where $\tilde{\kappa}_n = \kappa_n\log(n)^{1/a}$ is defined in Assumption 1, we have uniformly in h: $\underline{C}\log(n)^{1/a}\|h\|^r - o_p(1)\leq$

[17] This is because $\log(\kappa_n) = o(\kappa_n^{-a})$ for any $a>0$.

$\|\bar{g}_n(\theta)\|_W/\kappa_n \leq \overline{C}\log(n)^{1/a}\|h\|^r + o_p(1)$. In turn, this implies that for $\|h\|^{ra} > d/[C_2\underline{C}^a] > 0$: $\hat{K}_n(\theta) \leq C_1 \exp\left[-C_2\underline{C}^a\|h\|^{ra}\log(n) + o_p(1)\right] = o_p(n^{-d})$ for any $d \geq 1$. The integral of interest can now be bounded above and below for $n \geq 2$:

$$\kappa_n^{d_\theta/r} C_1 \int_{\|h\|^{ra} \leq d_\theta/r \log(n)/[C_2\underline{C}^a]} \exp\left[-C_2\overline{C}^a\|h\|^{ra} + o_p(1)\right] dh + o_p(n^{-d_\theta/r}) \leq \int_{\mathcal{B}_{\varepsilon/3}(\theta_j)} \hat{K}_n(\theta) d\theta$$
$$\leq \kappa_n^{d_\theta/r} C_1 \int_{\|h\|^{ra} \leq d_\theta/r \log(n)/[C_2\underline{C}^a]} \exp\left[-C_2\underline{C}^a\|h\|^{ra} + o_p(1)\right] dh + o_p(n^{-d_\theta/r}).$$

Now note that:

$$\int_{\|h\|^{ra} \leq d_\theta/r \log(n)/[C_2\underline{C}^a]} \exp\left[-C_2\overline{C}^a\|h\|^{ra} + o_p(1)\right] dh$$
$$\geq \int_{\|h\|^{ra} \leq d_\theta/r \log(2)/[C_2\underline{C}^a]} \exp\left[-C_2\overline{C}^a\|h\|^{ra} + o_p(1)\right] dh > 0$$
$$\int_{\|h\|^{ra} \leq d_\theta/r \log(n)/[C_2\underline{C}^a]} \exp\left[-C_2\underline{C}^a\|h\|^{ra} + o_p(1)\right] dh$$
$$\leq \int_{\mathbb{R}^{d_\theta}} \exp\left[-C_2\underline{C}^a\|h\|^{ra} + o_p(1)\right] dh < +\infty.$$

This provides lower and upper bounds on the probability of the $\varepsilon/3$ ball around θ_j:

$$\frac{\int_{\|h\|^{ra} \leq d_\theta/r \log(2)/[C_2\underline{C}^a]} \exp\left[-C_2\overline{C}^a\|h\|^{ra} + o_p(1)\right] dh}{\int_{\mathbb{R}^{d_\theta}} \exp\left[-C_2\underline{C}^a\|h\|^{ra} + o_p(1)\right] dh} + o_p(1) \leq \int_{\mathcal{B}_{\varepsilon/3}(\theta_j)} \hat{\pi}_n(\theta) d\theta$$
$$\leq \frac{\int_{\mathbb{R}^{d_\theta}} \exp\left[-C_2\underline{C}^a\|h\|^{ra} + o_p(1)\right] dh}{\int_{\|h\|^{ra} \leq d_\theta/r \log(2)/[C_2\underline{C}^a]} \exp\left[-C_2\overline{C}^a\|h\|^{ra} + o_p(1)\right] dh} + o_p(1).$$

Since the first term is strictly positive with probability going to 1, this proves the result.

Case iii. $\Theta_0 = \cup_{j=1}^k \mathcal{S}_j$:

Let $\theta \notin \cup_{j=1}^k \mathcal{N}(\mathcal{S}_j)$, then condition iii.c. implies that $\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \geq \eta/\kappa_n - o_p(1) \to +\infty$ uniformly in θ. This implies that with probability going to 1: $\sup_{\theta \notin \cup_{j=1}^k \mathcal{N}(\mathcal{S}_j)} K\Big(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}/\kappa_n\Big) = \sup_{\theta \notin \cup_{j=1}^k \mathcal{N}(\mathcal{S}_j)} K\Big(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - o_p(1)\Big) = 0$ for the compact kernel. For the exponential kernel, for any $d \geq 1$: $\sup_{\theta \notin \cup_{j=1}^k \mathcal{N}(\mathcal{S}_j)} K(\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n - \|\bar{g}_n(\hat{\theta}_n)\|_{\hat{W}_n}/\kappa_n) \leq C_1 \exp\left(-C_2\left[\eta^a/\kappa_n^a - o_p(1)\right]\right) \leq o_p(n^{-d})$. Then assumption iii.c. implies that for any θ such that $d(\theta, \Theta_0) \geq 2\kappa_n/\underline{C}$, we have: $\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \geq 2 - o_p(1)$ which is greater than $3/2$ with probability going to 1 so that for the compact kernel $\hat{K}_n(\theta) = 0$ with probability going to 1 uniformly in θ with $d(\theta, \Theta_0) \geq 2\kappa_n/\underline{C}$. Similarly, for any θ such that $d(\theta, \Theta_0) \geq [d^{1/a}2\underline{C}^{-1}C_2^{-1/a}]\kappa_n \log(n)^{1/a}$, we have: $\hat{K}_n(\theta) \leq C_1 \exp\left(-[d2^a \log(n) - o_p(1)]\right) \leq o_p(n^{-d})$, for the exponential kernel. For both kernels, $\hat{K}_n(\theta) \leq o_p(n^{-d_\theta})$, with probability going to 1, uniformly in θ with $d(\theta, \Theta_0) \geq C_K\tilde{\kappa}_n$, C_K corresponds to the bounds above for each kernel.

Furthermore, for the compact kernel, $d(\theta,\Theta_0) \leq \kappa_n/[2\overline{C}]$ implies: $\|\bar{g}_n(\theta)\|_{W_n}/\kappa_n \leq 1/2 + o_p(1)$; so that, uniformly in θ with $d(\theta,\Theta_0) \leq \kappa_n/[2\overline{C}]$, we have with probability going to 1: $\hat{K}_n(\theta) \geq \inf_{x\in[0,2/3]} K(x) > 0$. Similarly, for $d(\theta,\Theta_0) \leq \kappa_n \log(n)^{1/a}/[2\overline{C}]$, for the exponential kernel: $\hat{K}_n(\theta) \geq C_1 \exp\Big(-C_2 \log(n)\Big) \times [1+o_p(1)]$. This implies the following two inequalities:

$$\int_{\theta,\, d(\theta,\mathcal{S}_j)\leq \tilde{\kappa}_n/[2\overline{C}]} \hat{K}_n(\theta)d\theta \leq \int_{\mathcal{N}(S_j)} \hat{K}_n(\theta)d\theta \leq \int_{\theta,\, d(\theta,\mathcal{S}_j)\leq C_K\tilde{\kappa}_n} \hat{K}_n(\theta)d\theta + o_p(n^{-d_\theta}).$$

The next step is to find a bounds for the terms on the left and right-hand sides using the change of variable φ_j. Before that, it is necessary to map $\tilde{\kappa}_n$ neighborhoods in the integrals with neighborhoods in the space induced by the change of variable.

By condition iii.a. $\mathcal{U}_j$ has non-empty interior and is bounded so that there exists $0 < \varepsilon_1 \leq \varepsilon_2 < \infty$ and $\nu_j \in \mathcal{S}_j$ such that: $\mathcal{B}_{\varepsilon_1}(\nu_j) \subseteq \mathcal{U}_j \subseteq \mathcal{B}_{\varepsilon_2}(\nu_j) \Rightarrow \mathcal{B}_{\varepsilon_1}(\nu_j) \times \{0\} \subseteq \mathcal{U}_j \times \{0\} = \varphi_j^{-1}(\mathcal{S}_j) \subseteq \mathcal{B}_{\varepsilon_2}(\nu_j) \times \{0\}$. Note that for any $\vartheta_1 \in \mathcal{N}(\mathcal{U}_j) \times \mathcal{N}(\{0\})$ and $\vartheta_2 \in \mathcal{U}_j \times \{0\}$, the mean-value theorem implies:

$$\|\varphi_j(\vartheta_1) - \varphi_j(\vartheta_2)\| = \|\partial_\vartheta \varphi_j(\tilde{\vartheta})(\vartheta_1 - \vartheta_2)\| \leq |\lambda_{\max}(\partial_\vartheta \varphi_j(\tilde{\vartheta}))| \times \|\vartheta_1 - \vartheta_2\| \leq \overline{\lambda}_j \times \|\vartheta_1 - \vartheta_2\|$$

This implies that $\mathcal{B}_{C_k\tilde{\kappa}_n}(\mathcal{S}_j) \subseteq \varphi_j\left(\mathcal{B}_{C_k\tilde{\kappa}_n/\overline{\lambda}_j}(\mathcal{U}_j) \times \mathcal{B}_{C_k\tilde{\kappa}_n/\overline{\lambda}_j}(\{0\})\right)$ so that:

$$\begin{aligned}\int_{\theta,\, d(\theta,\mathcal{S}_j)\leq C_K\tilde{\kappa}_n} \hat{K}_n(\theta)d\theta &\leq \int_{\theta=\varphi_j^{-1}(\vartheta),\, d(\vartheta,\mathcal{U}_j\times\{0\})\leq C_K/\overline{\lambda}_j\tilde{\kappa}_n} \hat{K}_n(\theta)d\theta \\ &\leq \int_{\theta=\varphi_j^{-1}(\vartheta),\, \vartheta\in\mathcal{B}_{2\varepsilon_2}(\nu_j)\times\mathcal{B}_{C_K/\overline{\lambda}_j\tilde{\kappa}_n}} \hat{K}_n(\theta)d\theta.\end{aligned}$$

The mean-value theorem can also be used to derive a lower bound:

$$\|\varphi_j(\vartheta_1) - \varphi_j(\vartheta_2)\| = \|\partial_\vartheta \varphi_j(\tilde{\vartheta})(\vartheta_1 - \vartheta_2)\| \geq |\lambda_{\min}(\partial_\vartheta \varphi_j(\tilde{\vartheta}))| \times \|\vartheta_1 - \vartheta_2\| \geq \underline{\lambda}_j \times \|\vartheta_1 - \vartheta_2\|$$

By definition of ε_1, for n large enough $\varphi_j(\mathcal{B}_{\varepsilon_1/2}(\nu_j)\times\mathcal{B}_{C_k\tilde{\kappa}_n/\underline{\lambda}_j}(\{0\})) \subset \varphi_j(\mathcal{B}_{\varepsilon_1}(\theta_j)\times\mathcal{B}_{C_k\tilde{\kappa}_n/\underline{\lambda}_j}(\{0\})) \subseteq \mathcal{B}_{C_k\tilde{\kappa}_n}(\mathcal{S}_j)$. This results in another inequality:

$$\int_{\theta=\varphi_j^{-1}(\vartheta),\vartheta\in\mathcal{B}_{\varepsilon_1/2}(\nu_j)\times\mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}} \hat{K}_n(\theta)d\theta \leq \int_{\theta,\, d(\theta,\mathcal{S}_j)\leq\tilde{\kappa}_n/[2\overline{C}]} \hat{K}_n(\theta)d\theta.$$

Now that the neighborhoods are defined, consider the change of variable: $\varphi_j(\vartheta) = \theta$, then the integral becomes: $\int_{\vartheta\in\mathcal{B}_{\varepsilon_1/2}\times\mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}} |\det(\partial_\vartheta\varphi_j(\vartheta))|\, \hat{K}_n \circ \varphi_j(\vartheta)d\vartheta$.

Condition iii.a. implies that uniformly in $\vartheta \in \varphi_j(\mathcal{N}(\mathcal{S}_j))$: $0 < \underline{\lambda}_j^{d_\theta} \leq |\det(\partial_\vartheta\varphi_j(\vartheta))| \leq \overline{\lambda}_j^{d_\theta} < +\infty$. Combining this together with the bounds on $\hat{K}_n$ implies the following lower

bound; for the compact kernel, with probability going to 1:

$$\underline{\lambda}_j^{d_\theta} \times \left(\inf_{x \in [0,2/3]} K(x) \right) \int_{\mathcal{B}_{\varepsilon_1/2} \times \mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}(\{0\})} d\vartheta \leq \int_{\vartheta \in \mathcal{B}_{\varepsilon_1/2} \times \mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}} \left| \det \left(\partial_\vartheta \varphi_j(\vartheta) \right) \right| \hat{K}_n \circ \varphi_j(\vartheta) d\vartheta.$$

The left-hand side involves the volume of a d_θ-dimensional ball:

$$\int_{\mathcal{B}_{\varepsilon_1/2} \times \mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}(\{0\})} d\vartheta = \frac{\pi^{d_\theta/2}}{\Gamma(k_j/2+1)\Gamma([d_\theta - k_j]/2+1)} [\varepsilon_1/2]^{d_\theta - k_j} [\tilde{\kappa}_n/(2\overline{C}\overline{\lambda}_j)]^{k_j}$$

where Γ is Euler's gamma function. Similarly, for the exponential kernel:

$$\begin{aligned} &\underline{\lambda}_j^{d_\theta} C_1 \exp\left(-C_2 \log(n)\right) \left[1 + o_p(1)\right] \int_{\mathcal{B}_{\varepsilon_1/2} \times \mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}(\{0\})} d\vartheta \\ &\leq \int_{\vartheta \in \mathcal{B}_{\varepsilon_1/2} \times \mathcal{B}_{\tilde{\kappa}_n/[2\overline{C}\overline{\lambda}_j]}} \left| \det \left(\partial_\vartheta \varphi_j(\vartheta) \right) \right| \hat{K}_n \circ \varphi_j(\vartheta) d\vartheta. \end{aligned}$$

Similarly, a change of variable can be applied to the upper-bound:

$$\begin{aligned} \int_{\theta = \varphi_j(\vartheta),\, \vartheta \in \mathcal{B}_{2\varepsilon_2}(\nu_j) \times \mathcal{B}_{C_K/\overline{\lambda}_j \tilde{\kappa}_n}} \hat{K}_n(\theta) d\theta &= \int_{\vartheta \in \mathcal{B}_{2\varepsilon_2}(\nu_j) \times \mathcal{B}_{C_K/\overline{\lambda}_j \tilde{\kappa}_n}} \left| \det \left(\partial_\vartheta \varphi_j(\vartheta) \right) \right| \hat{K}_n \circ \varphi_j(\vartheta) d\vartheta \\ &\leq \overline{\lambda}_j^{d_\theta} \times \|K\|_\infty \times \int_{\vartheta \in \mathcal{B}_{2\varepsilon_2}(\nu_j) \times \mathcal{B}_{C_K/\overline{\lambda}_j \tilde{\kappa}_n}} d\vartheta. \end{aligned}$$

The integral on the right-hand side can also be computed analytically:

$$\int_{\vartheta \in \mathcal{B}_{2\varepsilon_2}(\nu_j) \times \mathcal{B}_{C_K/\overline{\lambda}_j \tilde{\kappa}_n}} d\vartheta = \frac{\pi^{d_\theta/2}}{\Gamma([d_\theta - k_j]/2+1)\Gamma(k_j/2+1)} [2\varepsilon_2]^{d_\theta - k_j} [\tilde{\kappa}_n C_K/\overline{\lambda}_j]^{k_j}.$$

Putting everything together and given that $n^{-d_\theta} = o(\tilde{\kappa}_n^{k_j})$, we have with probability going to 1: $\int_{\mathcal{N}(S_j)} \hat{K}_n(\theta) d\theta \asymp \tilde{\kappa}_n^{k_j}$, for each set $\mathcal{S}_j$ covering Θ_0. This implies that: $\int_{\mathcal{N}(S_j)} \hat{\pi}_n(\theta) d\theta \asymp \kappa_n^{k_j - \min_\ell k_\ell}$, which goes to 0 for all j with $k_j > \underline{k} = \min_\ell k_\ell$. Asymptotically, the sets $\mathcal{S}_j$ with the largest degree of identification failure k_j will dominate the posterior distribution.

To get the desired result, pick j such that $k_j = \underline{k}$ and the associated ε_1 described above. As before, for any $0 < \varepsilon < \varepsilon_1/2$ there exists a constant $C_\varepsilon > 0$ such that for any $\nu \in \mathcal{B}_{\varepsilon_1}(\nu_j)$, with probability going to 1:

$$C_\varepsilon \tilde{\kappa}_n^{k_j} + o_p(n^{-d}) \leq \int_{\varphi_j(\mathcal{B}_\varepsilon(\nu) \times \mathcal{B}_\varepsilon(\{0\}))} \hat{K}_n(\theta) d\theta \leq \int_{\mathcal{N}(\mathcal{S}_j)} \hat{K}_n(\theta) d\theta,$$

where the last inequality holds for ε small enough so that $\mathcal{B}_\varepsilon(\nu) \times \mathcal{B}_\varepsilon(\{0\}) \subseteq \mathcal{N}(\mathcal{U}_j) \times \mathcal{N}(\{0\})$ holds. Such $\varepsilon > 0$ exists by the definition of open sets. This implies that for any two $\nu_1 \neq \nu_2$ in $\mathcal{B}_{\varepsilon_1}(\nu_j)$, $\exists \varepsilon > 0$ small enough such that Proposition 1 holds, which concludes the proof. $\square$

Appendix B Proofs for the Results of Section 4

B.1 Weak or Set Identification

Proof of Theorem 5. The proof will treat (semi)-strong and weak identification separately.

Case 1: (semi)-strong identification.

Given the assumptions, Lemma D4 holds and $\hat{\ell}_n = 1$ with probability going to 1. Theorem 1 also holds: $\hat{\theta}_n$ and $\hat{\theta}_{n,LS/\infty}$ are consistent and asymptotically normal. The second result is then a consequence of condition 2.iii. in the Theorem.

Case 2: weak/set identification.

Under weak or set identification, Lemmas D4 and D6 imply that, with probability going to 1, $\hat{\ell}_n \geq \ell^\star$. The remaining assumptions imply that the second step, inference, is asymptotically valid which concludes the proof. □

Proof of Remark 2. Since neither $\bar{g}_n(\alpha_0, \cdot)$ nor $\hat{V}_n(\alpha_0, \cdot)$ depend on β, $S_n(\alpha_0)$ does not depend on β so that equation (11) holds. The consistency of $\hat{V}_n$ and the central limit theorem for $\bar{g}_n$ imply that equation (12) holds. Under strong identification, $S_n(\alpha_0)$ is a Likelihood-ratio statistic so that equation (13) is a standard results, see e.g. Newey & McFadden (1994). □

B.2 Higher-Order Identification

Proof of Theorem 6. The proof will treat (semi)-strong and weak identification separately.

Case 1: (semi)-strong identification.

By Lemma D8, with probability going to 1, $\bar{h}_n^2 \leq C[\sqrt{n}\kappa_n^2/\lambda_{\min}(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0))]^2) \overset{def}{=} \varepsilon_n^2$ for some $C \geq 0$. The size distortion implied by a non-central χ_1^2 distribution with non-centrality parameter ε_n is given by:

$$\begin{aligned}\int_{-\sqrt{c_{1-\alpha}}}^{\sqrt{c_{1-\alpha}}} \phi(x+\varepsilon_n)dx - (1-\alpha) &= \int_{-\sqrt{c_{1-\alpha}}}^{\sqrt{c_{1-\alpha}}} [\phi(x+\varepsilon_n) - \phi(x)]dx \\ &= \varepsilon_n \int_{-\sqrt{c_{1-\alpha}}}^{\sqrt{c_{1-\alpha}}} \phi'(x)dx + o(\varepsilon_n) = 2\varepsilon_n[\phi(\sqrt{c_{1-\alpha}}) - \phi(0)] + o(\varepsilon_n).\end{aligned}$$

This implies result a. if $\varepsilon_n = o(\bar{\gamma}_n)$.

Case 2: weak/set identification.

First, by using similar arguments as in the proof of Lemma D6 it can be shown that Lemma D7 implies that $\hat{\ell}_n \geq \ell^\star$ with probability going to 1. Given the remaining asusmptions, the rest of the proof is identical to the proof of Theorem 5. □

Supplement to "Detecting Identification Failure in Moments Condition Models"

Jean-Jacques Forneron*

September 6, 2019

This Supplemental Material consists of Appendices C, D, E, F, G and H to the main text.

*Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215.
Email: jjmf@bu.edu.

Appendix C Additional Results

C.1 Additional Rules of Thumb for Section 4.1.2

C.1.1 Over-Identified Models

Consider the over-identified linear Gaussian experiment:

$$\bar{g}_n(\theta) = A_n + B_n\theta$$

where $A_n - B_n\theta_0 = Z_1$, $B_n = \bar{B}_n + Z_2$ and $(Z_1', \text{vec}(Z_2)')'$ Gaussian, $d_\theta \leq \dim(Z_1)$. Let W be a non-stochastic positive definite matrix which does not depend on θ. Let $\mathbb{E}(Z_1 Z_1') = V_1/n$ and $\mathbb{E}(Z_2' W Z_1) = V_{21}$. Furthermore, assume that $\bar{B}_n = n^{-\delta} \times B$ with B full rank and $\delta \in [0, 1/2)$. Under the stated assumptions:

$$\hat{\theta}_n - \theta_0 = -\left(\bar{B}_n' W \bar{B}_n\right)^{-1} \bar{B}_n' W Z_1 - \left(\bar{B}_n' W \bar{B}_n\right)^{-1} Z_2' W Z_1 + O_p(n^{-3/2+3\delta}).$$

Take $v_{j,n}'$ to be a right eigenvector of $W^{1/2}\bar{B}_n$, associated with the eigenvalue $\lambda_{j,n}$ then:

$$v_{j,n}(\hat{\theta}_n - \theta_0) = -\lambda_{j,n}^{-1} W_{1/2} Z_1 - \lambda_{j,n}^{-2} Z_2' W Z_1 + O(n^{-3/2+3\delta}).$$

As in the just-identified case:

$$\frac{|bias|^2}{|variance|} = \frac{1}{n|\lambda_{j,n}|^2} \frac{v_{j,n}^\star V_{21} V_{12} v_{j,n}}{v_{j,n}^\star V_1 v_{j,n}} + o\left(\frac{1}{n|\lambda_{j,n}|^2}\right).$$

The rule-of-thumb then proceed the same way as in the just-identified case.

C.1.2 A non-local approach to the rule-of-thumb

When the model is set identified but the local identification condition holds, as in Examples 2, the rule-of-thumb above may not perform well because V_{12} will tend to be very small (Θ_0 covers a wide range but $\bar{g}_n$ is less than κ_n). A simple solution is to split Θ_0 into clusters within which the local rule-of-thumb can provide a better approximation. In practice, one can apply the *k-means* algorithm[1] to build clusters on $\{\theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_\theta \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n\}$ and apply the rule of thumb within each cluster, B_n should be re-approximated within each cluster. The largest cutoff across clusters becomes the global rule-of-thumb for Algorithm 1. This yields an approximation for size distortion within clusters.[2] A concern may be that

[1] See e.g. Hastie et al. (2009), Chapter 14.3, for an overview of cluster analysis.

[2] This approach was applied in Appendix G.2 in Monte-Carlo simulations for Example 2.

when these clusters are far from one another which leads to between-cluster size distortion. A between-cluster rule-of-thumb will have to be computed. Suppose the k-means procedure picked $\mathcal{C} \geq 2$ clusters. By construction, $\hat{\theta}_n$ belongs to one of these clusters. Denote this cluster cl_n and take θ_{cl_n}, a solution to (2) within that cluster. For $v_{j,n}$ described above, the t-statistic for testing $v'_{j,n}(\theta - \theta_{cl}) = 0$, where θ_{cl} belongs to one of the several potential clusters $cl \in \{1, \ldots, \mathcal{C}\}$, is:

$$\begin{aligned} t_n &= \sqrt{n} \times \frac{v'_{j,n}(\hat{\theta}_n - \theta_{cl})}{\sqrt{v'_{j,n} B_n V_{cl_n} B'_n \bar{v}_{j,n}}} \\ &= \sqrt{n} \times |\lambda_{j,n}|^{-1} \times \frac{v'_{j,n}(\hat{\theta}_n - \theta_{cl})}{\sqrt{v'_{j,n} V_{cl_n} \bar{v}_{j,n}}} \\ &= \underbrace{\sqrt{n} \times |\lambda_{j,n}|^{-1} \times \frac{v'_{j,n}(\hat{\theta}_n - \theta_{cl_n})}{\sqrt{v'_{j,n} V_{cl_n} \bar{v}_{j,n}}}}_{\text{within cluster}} + \underbrace{\sqrt{n} \times |\lambda_{j,n}|^{-1} \times \frac{v'_{j,n}(\theta_{cl_n} - \theta_{cl})}{\sqrt{v'_{j,n} V_{cl_n} \bar{v}_{j,n}}}}_{\text{between clusters}}. \end{aligned}$$

The last equality suggests that the between-cluster size distortion is a function of the distance between the clusters relative the sampling uncertainty. If the distance is small relative to standard errors, then size distortion is minimal. This yields the between-cluster rule-of-thumb:

$$\underline{\lambda}^2_{n,\text{between}} \geq \frac{n \times \sup_{cl,cl'} d(cl, cl')}{c(\bar{\gamma}_n)^2 \times \inf_{cl \in \{1,\ldots,\mathcal{C}\}} \sqrt{\lambda_{\min}(V_{cl})}},$$

where $d(cl, cl')$ is the distance between the two clusters cl and cl'. If there is only one cluster, this rule-of-thumb is not needed. The overall rule-of-thumb is simply the largest value of the between and within rules-of-thumb computed above. In the Monte-Carlo simulations, the following stopping rule was used to determine whether an additional cluster should be added: if the distance between centroids, with the other clusters, was greater than 1% of the diameters of the clusters then it would be added. This is a quick way to check if the two cluster are close to one another or not. Other criteria could be considered in future research.

C.2 Additional Results for Section 4.1

C.2.1 A quasi-CLT for $A_{n,LS}, B_{n,LS}$

Lemma C2. *(Addional Asymptotic Results for $A_{n,LS}, B_{n,LS}$ under Strong and Semi-Strong Identification) Suppose that the assumptions of Theorem 1 and the following hold:*

i. uniform Central Limit Theorem:

$$\sqrt{n}\begin{pmatrix} \bar{g}_n(\theta_0) \\ vec[\partial_\theta \bar{g}_n(\theta_0)] - vec[\partial_\theta g_n(\theta_0)] \end{pmatrix} \xrightarrow{d} \mathcal{N}(0, V).$$

ii. stochastic equicontinuity condition, i.e. for all $\delta_n \downarrow 0$:

$$\sup_{\|\theta_1-\theta_2\|\leq\delta_n} \|\sqrt{n}\left([\partial_\theta \bar{g}_n(\theta_1) - \partial_\theta \bar{g}_n(\theta_2)] - [\partial_\theta g_n(\theta_1) - \partial_\theta g_n(\theta_2)]\right)\| = o_p(1)$$

then $A_{n,LS} = \bar{g}_n(\theta_0) - B_{n,LS}\theta_0 + o_p(n^{-1/2})$ and $B_{n,LS}H_n = \partial_\theta \bar{g}_n(\theta_0)H_n + o_p(n^{-1/2})$ so that

$$\sqrt{n}\begin{pmatrix} A_{n,LS} + B_{n,LS}\theta_0 \\ vec(B_{n,LS} - \bar{B}_{n,LS}) \end{pmatrix} = \sqrt{n}\begin{pmatrix} \bar{g}_n(\theta_0) \\ vec[\partial_\theta \bar{g}_n(\theta_0) - \partial_\theta g_n(\theta_0)] \end{pmatrix} + o_p(1) \xrightarrow{d} \mathcal{N}(0, V)$$

where $\bar{B}_{n,LS}$ corresponds to $B_{n,LS}$ computed using the random measure $\hat{\pi}_n$:

$$\bar{B}'_{n,LS} = \Sigma_n^{-1} \int_\Theta (\theta - \bar{\theta}_n)[\int_\Theta \{g_n(\theta) - g_n(\tilde{\theta})\}\hat{\pi}_n(\tilde{\theta})d\tilde{\theta}]'\hat{\pi}_n(\theta)d\theta.$$

C.2.2 Primitive Conditions for Theorem 5

Lemma C3 (Asymptotic Distribution of the Profile S Statistic under (Semi)-Strong Identification)**.** *Suppose that $\theta = A(\beta', \gamma')'$ where A is an invertible matrix and γ is (semi)-strongly identified: the assumptions of Theorem 1 hold for γ holding $\beta = \beta_0$ fixed. Let $\hat{V}_n \xrightarrow{p} \lim_{n\to\infty} n \times var(\bar{g}_n(\beta_0, \gamma_0))$ positive definite and $\hat{\gamma}_n$ is computed using $W_n = V_n^{-1}$, then:*

$$n \times \bar{g}_n\Big(A(\beta_0', \hat{\gamma}_n')'\Big)'\hat{V}_n^{-1}\bar{g}_n\Big(A(\beta_0', \hat{\gamma}_n')'\Big) \xrightarrow{d} \chi^2_{dim(g)-dim(\gamma)}.$$

Proposition C2 (Verifying Theorem 5's Condition 2. for the S-statistic)**.** *Suppose the assumptions for Lemmas D4 and D6 hold. Suppose there exists a re-parameterization with an invertible matrix A and vectors γ_ℓ, $\ell = 1, \ldots, \mathcal{L}$ such that:*

$$\theta = A(\gamma_1', \ldots, \gamma_{\mathcal{L}}')', \quad R_\ell\theta = c \Leftrightarrow (\gamma_1', \ldots, \gamma_\ell')' = \tilde{c},$$

for $c, \tilde{c} \in \mathbb{R}^{rank(R_\ell)}$. Let $\ell^\star$, $\theta_{0,c}$ and $\hat{\theta}_{n,\ell,c}$ be defined as in Theorem 5, suppose that for each $\ell \geq \ell^\star$:

$$n \times \lambda_{\min}(\mathcal{I}_\ell A' \partial_\theta g_n(\theta_{0,c})' \partial_\theta g_n(\theta_{0,c}) A \mathcal{I}_\ell) \to +\infty,$$

as $n \to \infty$, where $\mathcal{I}_\ell = blockdiag(0_{rank(R_\ell)}, I_{d_\theta - rank(R_\ell)})$. Then for each $\ell > \ell^\star$, $(\gamma'_\ell, \ldots, \gamma'_\mathcal{L})'$ is semi-strongly identified. If furthermore, $W_n(\theta) = \hat{V}_n(\theta)^{-1}$ is a uniformly consistent estimator for $n \times var(\bar{g}_n(\theta))$, then for each $\ell \geq \ell^\star$:

$$S_{n,\ell} = \inf_{R_\ell \theta = c_\ell} n \times \|\bar{g}_n(\theta)\|^2_{\hat{V}_n^{-1}} \xrightarrow{d} \chi^2_{dim(g) - [d_\theta - rank(R_\ell)]}.$$

Let $c_{1-\alpha,\ell}$ be the $1-\alpha$ quantile of a χ^2 distribution with $dim(g) - [d_\theta - rank(R_\ell)]$ degrees of freedom, then the following holds:

$$\inf_{\ell = \ell^\star, \ldots, \mathcal{L}} \mathbb{P}(T_{n,\ell} \leq c_{1-\alpha,\ell}) = 1 - \alpha + o(1).$$

C.3 (Counter)-Examples for Proposition 1 and Theorem 4

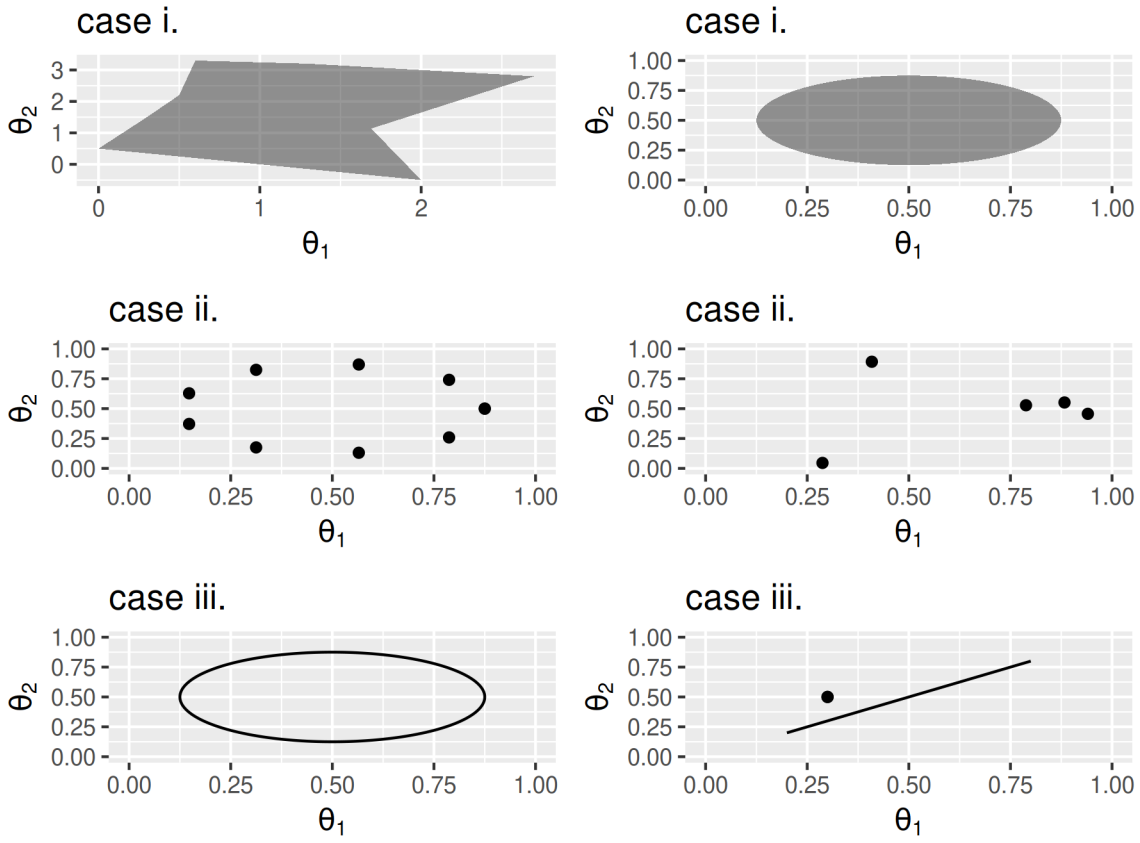
C.3.1 Examples for Theorem 4

Figure C5 shows several examples for cases i-iii considered in Theorem 4. The first row shows two sets with non-empty interior (shaded area) so that Θ_0 has non-zero measure in $\mathbb{R}^2$. In each set, it is possible to find two disjoint balls that has non-zero measure in $\mathbb{R}^2$ and thus strictly positive posterior mass asymptotically.

The second row illustrates case ii. with finite collections of singletons arranged in various patterns. The particular alignment of the points is not relevant for the result as much as the local behaviour of the moments g_n around these points.

The third row illustrates case iii. with a one dimensional manifold on the left-hand-side: the circle can be represented as $\{(\theta_1, \theta_2) = (\theta_{1,0}, \theta_{2,0}) + R \times (\cos(\vartheta), \sin(\vartheta)), \vartheta \in [0, 2\pi]\}$ for some $(\theta_{1,0}, \theta_{2,0}) \in \mathbb{R}^2, R > 0$. There is thus a one-to-one mapping between the circle and $[0, 2\pi]$ which has non-empty interior in $\mathbb{R}$. The example on the right-hand-side has $\mathcal{S}_1 =$ line and $\mathcal{S}_2 =$ point (0-dimensional manifold). In that setting it can be shown that $\hat{\pi}_n(B_\varepsilon(\mathcal{S}_2)) \xrightarrow{p} 0$ for $\varepsilon > 0$ small, fixed. The posterior is dominated by $\mathcal{S}_1$, i.e. the line. While the posterior variance is bounded below (in probability) in the direction $(1, 1)$, it converges to 0 in the direction $(1, -1)$ because S_2 has measure 0 and does not impact the posterior variance asymptotically as a result.

Figure C4: Examples of Topologies for Θ_0 in Theorem 4



C.3.2 Counter-Example: a moment function that does not satisfy the conditions for Proposition 1

Consider the function: $g_n(\theta) = \theta^6 \sin(1/\theta), \quad \theta \in [-1, 1]$. This function is twice continuously differentiable on $[-1, 1]$ with bounded second derivative. It has infinitely many zeros in $[-1, 1]$ and such that:

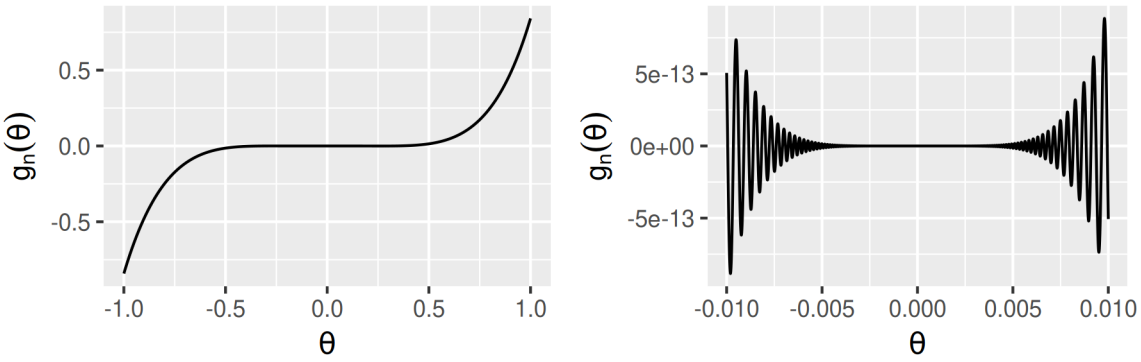
$$\#\{\theta \in [-\varepsilon, \varepsilon], g_n(\theta) = 0\} = +\infty, \quad \#\{\theta \notin [-\varepsilon, \varepsilon], g_n(\theta) = 0\} < +\infty,$$

for any $\varepsilon > 0$. This implies that: $\int_{[-\varepsilon,\varepsilon]} \hat{\pi}_n(\theta) d\theta \xrightarrow{p} 1$, for any $\varepsilon > 0$. In turn, the posterior variance can be bounded above by:

$$\Sigma_n \leq \int_{[-1,1]} \theta^2 \hat{\pi}_n(\theta) d\theta = \int_{[-\varepsilon,\varepsilon]} \theta^2 \hat{\pi}_n(\theta) d\theta + \int_{[-1,1]/[-\varepsilon,\varepsilon]} \theta^2 \hat{\pi}_n(\theta) d\theta \leq \varepsilon^2 \times [1 + o_p(1)] + o_p(1),$$

for any $\varepsilon > 0$. This implies that $\Sigma_n \xrightarrow{p} 0$ even thought the model is set identified. Note that it is possible that $\Sigma_n \xrightarrow{p} 0$ at a rate slower than κ_n^2 so that $\lambda_{\min}(B_{n,LS}) \xrightarrow{p} 0$ but at a slower than κ_n rate. Explicit rates are hard to compute analytically for this example.

Figure C5: Function g_n for which the conditions for Proposition 1 are not met



Note: Left panel: g_n plotted over $[-1, 1]$, right panel g_n plotted over $[-0.01, 0.01]$.

Appendix D Preliminary Results for Section 4

D.1 Preliminary Results for Section 4.1

Lemma D4. *(Detecting Weak/Set Identification Failures) Let $\underline{\lambda}_n = o(1)$ be a strictly positive sequence. Let $B_{n,LS/\infty}$ be either $B_{n,LS}$ or $B_{n,\infty}$.*

a. (Weak/Set Identification) Suppose that the assumptions of Theorem 3 or Lemma 1 hold so that Σ_n is a $O_p(\tilde{\kappa}_n^2)$ on the span V defined in Theorem 3 and $\tilde{\kappa}_n = o(\underline{\lambda}_n)$, then

$$\lim_{n\to\infty} \mathbb{P}(\lambda_{\min}(B_{n,LS/\infty}) < \underline{\lambda}_n) = 1,$$

b. (Semi-Strong and Strong Identification) Suppose that the assumptions of Theorem 1 hold and the Jacobian is such that $\underline{\lambda}_n^2 = o\left(\lambda_{\min}\left(\partial_\theta g_n(\theta_0)\partial_\theta g_n(\theta_0)'\right)\right)$, then

$$\lim_{n\to\infty} \mathbb{P}(\lambda_{\min}(B_{n,LS/\infty}) < \underline{\lambda}_n) = 0.$$

Lemma D4 a. is a direct implication of the results in Section 3. Lemma D4 b. suggests which sequences of (semi)-strongly identified models may lead to false positives when detecting identification failure. To illustrate, take $\kappa_n = \sqrt{2\log(\log[n])}n^{-1/2}$ and $\lambda_n =$

$\sqrt{\log(n)}n^{-1/2}$. For these sequences, models with $n^{-1/2} \ll \lambda_{\min}\left(\partial_\theta g_n(\theta_0)\right) \lesssim \sqrt{\log(n)}n^{-1/2}$ may be detected as weakly identified.[3] In practice, this range seems to be fairly small.

Lemma D5. *(Collapsing the Weakly Identified Set into a Singleton) Let* $V = Span(\{v = \theta_1 - \theta_0, (\theta_1, \theta_0) \in \Theta_0^2\})$ *where* Θ_0 *is the weakly identified set defined in equation (9). Let* P_V *be orthogonal projection matrix onto* V *and* $P_V^\perp$ *its orthogonal, if* $1 \leq rank(P_V^\perp) < d_\theta$ *then:*

a. $P_V^\perp \Theta_0 = \{P_V^\perp \theta_0\}$*, with* $\theta_0 \in \Theta_0$

b. let $(u^\star, v^\star) = (u_1^\star, \ldots, u_k^\star, v_1^\star, \ldots, v_{d_\theta - k}^\star)$ *be an orthogonal basis of* $\mathbb{R}^{d_\theta}$ *such that* $rank(P_V P_{v^\star}) = rank(P_V)$ *then for any* $c \in \mathbb{R}^{d_\theta - k}$

$$\Theta_0 \cap \{\theta \in \Theta, P_{v^\star}\theta = c\} = \{\theta_{0,c}\} \text{ or } \emptyset.$$

Lemma D6. *(Fixing the Span of Identification Failure) Suppose that the model is weakly or set identified and satisfies the assumptions of Lemma D4. Let* $\hat{\ell}_n$ *be the stopping value in Algorithm 1, then, with probability going to* 1

$$\Theta_0 \cap \{\theta \in \Theta, R_{\hat{\ell}_n}\theta = c_{\hat{\ell}_n}\}$$

is either a singleton or the empty set.

D.2 Preliminary Results for Section 4.2

Lemma D7 (Behaviour of $\bar{h}^2$ under Higher-Order Identification)**.** *Suppose that the model is higher-order identified and satisfies the assumptions of Theorem 2. Furthermore, assume that* $\bar{g}_n$ *is continuously differentiable around* θ_0*. Let* $A_n = \bar{g}_n(\hat{\theta}_n) - B_n\hat{\theta}_n$*,* $B_n = \partial_\theta \bar{g}_n(\hat{\theta}_n)$ *and* $\hat{R}_n(\theta) = \bar{g}_n(\theta) - A_n - B_n\theta$*, where* $\hat{\theta}_n$ *is a GMM estimator of* θ_0 *for some weighting matrix* $W_n = W + O_p(n^{-1/2})$*,* W *positive definite. Suppose that* $\underline{\lambda}_V + O_p(n^{-1/2}) \leq \lambda_{\min}(V_1) \leq \lambda_{\max}(V_1) \leq \overline{\lambda}_V + O_p(n^{-1/2})$ *for some* $0 < \underline{\lambda}_V \leq \overline{\lambda}_V < +\infty$*. Suppose that there exists* $\underline{\tilde{C}}_j \geq 0$*,* $j = 1, \ldots, r$*, with strict inequality when* $\underline{C}_j > 0$ *in Definition 5 such that for any* θ_1, θ_2 *with* $\|g_n(\theta_j)\|_W = O(\kappa_n)$*,* $j \in \{1, 2\}$*:*

$$\|\bar{g}_n(\theta_1) - \bar{g}_n(\theta_2) - \partial_\theta \bar{g}_n(\theta_2)(\theta_1 - \theta_2)\| \geq \sum_{j=2}^{r} \underline{\tilde{C}}_j \|P_j(\theta_1 - \theta_2)\|^j + O_p(n^{-1/2}),$$

where the P_j *are the same as in Definition 5, for each* $\ell \in \{1, \ldots, \mathcal{L}\}$*. Suppose that* $R_\ell \theta_0 = c_\ell$ *holds.*

[3]The relation $a_n \lesssim b_n$ implies that a_n is bounded by b_n modulo a constant: $\exists C > 0$, $a_n \leq C b_n$ for two sequences $(a_n)_{n\geq 1}$ and $(b_n)_{n\geq 1}$.

Let $V_r = Span(P_2, \ldots, P_r)$ be the span of the first-order identification failure. Let $\bar{h}^2$ be computed using the procedure in Algorithm 2. If $rank(P_{V_r} P_{\mathcal{R}_\ell}) < rank(P_{V_r})$ then $\bar{h}^2 \to +\infty$, as $n \to \infty$, with probability going to 1.

Lemma D7 shows that the criterion in Algorithm 2 diverges under higher-order identification. This implies that higher-order identification can be detected even when the sequence of tolerance $\bar{\gamma}_n$ is fixed, i.e. $\bar{\gamma}_n = \gamma$. The moments $\bar{g}_n$ are assumed to be smooth to simplify the Algorithm and the proofs. It is also assumed that the first-order identification failure occurs at θ_0 as well as in shrinking neighborhoods of θ_0. This will allow to substitute the residual curvature at the unknown θ_0 with a plug-in estimate.

Lemma D8 (Behaviour of $\bar{h}^2$ under (Semi)-Strong Identification)**.** *Suppose that the model is (semi)-strongly identified and satisfies the assumptions of Theorem 1. Furthermore, assume that $\bar{g}_n$ is continuously differentiable around θ_0 and $\partial_\theta \bar{g}_n(\theta)$ is non-singular in a neighborhood of θ_0. Let $A_n = \bar{g}_n(\hat{\theta}_n) - B_n \hat{\theta}_n$, $B_n = \partial_\theta \bar{g}_n(\hat{\theta}_n)$ and $\hat{R}_n(\theta) = \bar{g}_n(\theta) - A_n - B_n \theta$, where $\hat{\theta}_n$ is a GMM estimator of θ_0 for some weighting matrix $W_n = W + O_p(n^{-1/2})$, W positive definite. Suppose that $\hat{R}_n(\theta) \leq \bar{C}_n \|\theta - \hat{\theta}_n\|^2$ in a neigborhood of θ_0 for some $\bar{C}_n = O_p(1)$. Furthermore suppose that κ_n is such that $\sqrt{n}\kappa_n^2 = o\left(\lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))\right)$, then:*

$$\bar{h}_n^2 \leq O_p([\sqrt{n}\kappa_n^2 / \lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))]^2) = o_p(1).$$

Lemma D8 shows which (semi)-strongly identified models can be detected as such with high probability. Suppose $\bar{\gamma}_n = \gamma$ fixed and $\kappa_n = \sqrt{2\log(\log[n])} n^{-1/2}$, then semi-strongly identified models with $\lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0)) \lesssim \sqrt{2\log(\log[n])} n^{-1/2}$ may be subject to false positives. To put this into perspective, when the second derivative is non-zero, the local expansion is non-linear as soon as $\lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0)) \asymp n^{-1/2}$. This implies that the range of rates between $n^{-1/2}$ and $\sqrt{2\log(\log[n])} n^{-1/2}$ is subject to false positives. As in the case of detecting weak identification, this is a fairly narrow range.[4]

Appendix E Proofs for the Results of Appendix C

E.1 A quasi-CLT for $A_{n,LS}, B_{n,LS}$

Proof of Lemma C2. First recall that Theorem 1 implies:

$$A_{n,LS} = \bar{g}_n(\hat{\theta}_n) - B_{n,LS}\hat{\theta}_n + o_p(n^{-1/2}), \quad B_{n,LS} H_n = \partial_\theta g_n(\hat{\theta}_n) H_n + o_p(n^{-1/2}\kappa_n^{-1}).$$

[4] In terms of $\lambda_{\min}(\partial_\theta g_n(\theta_0))$ it corresponds to the $n^{-1/4}$ to $[\log(\log[n])]^{1/4} n^{-1/4}$ range.

The proof is divided into two parts, the first provides results for $A_{n,LS}$ and the second part derives the result for $B_{n,LS}$.

Step 1. Re-expressing $A_{n,LS}$ in terms of θ_0:

First, $A_{n,LS}$ can be expressed in terms of θ_0 rather than $\hat{\theta}_n$ up to a $o_p(n^{-1/2})$ term:

$$\begin{aligned} A_{n,LS} &= \bar{g}_n(\hat{\theta}_n) - B_{n,LS}\hat{\theta}_n + o_p(n^{-1/2}) \\ &= \bar{g}_n(\theta_0) - B_{n,LS}\theta_0 + \bar{g}_n(\hat{\theta}_n) - \bar{g}_n(\theta_0) - B_{n,LS}(\hat{\theta}_n - \theta_0) + o_p(n^{-1/2}) \end{aligned}$$

Note that $[B_{n,LS} - \partial_\theta g_n(\hat{\theta}_n)]H_n H_n^{-1}(\hat{\theta}_n - \theta_0) = o_p(n^{-1/2}\kappa_n^{-1})O_p(n^{-1/2}) = o_p(n^{-1/2})$ which implies:

$$B_{n,LS}(\hat{\theta}_n - \theta_0) = \partial_\theta g_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) + o_p(n^{-1/2}).$$

Substituting this into $A_{n,LS}$ together with condition iii. in Definition 4 yields:

$$\begin{aligned} A_{n,LS} &= \bar{g}_n(\theta_0) - B_{n,LS}\theta_0 + [\bar{g}_n(\hat{\theta}_n) - \bar{g}_n(\theta_0) - \partial_\theta g_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)] + o_p(n^{-1/2}) \\ &= \bar{g}_n(\theta_0) - B_{n,LS}\theta_0 + O(\|\partial_\theta g_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)\|^2) + o_p(n^{-1/2}). \end{aligned}$$

Using condition iv. in Definition 4, the last term can be re-written as:

$$\begin{aligned} \partial_\theta g_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0) &= \underbrace{\partial_\theta g_n(\theta_0)H_n}_{=O(1)}\underbrace{H_n^{-1}(\hat{\theta}_n - \theta_0)}_{=O_p(n^{-1/2})} + \underbrace{[\partial_\theta g_n(\hat{\theta}_n) - \partial_\theta g_n(\theta_0)]H_n}_{=o_p(1)}\underbrace{H_n^{-1}(\hat{\theta}_n - \theta_0)}_{=O_p(n^{-1/2})} \\ &= O_p(n^{-1/2}). \end{aligned}$$

As a result, $\|\partial_\theta g_n(\hat{\theta}_n)(\hat{\theta}_n - \theta_0)\|^2 = o_p(n^{-1/2})$, and $A_{n,LS} = \bar{g}_n(\theta_0) - B_{n,LS}\theta_0 + o_p(n^{-1/2})$.

Step 2. Expressing $B_{n,LS}$ as a function of $\partial_\theta \bar{g}_n$:

Using the least-squares formula, $B_{n,LS}$ can be expressed as:

$$\begin{aligned} B'_{n,LS} &= \Sigma_n^{-1}\int_\Theta (\theta - \bar{\theta}_n)[\int_\Theta \{g_n(\theta) - g_n(\tilde{\theta})\}\hat{\pi}_n(\tilde{\theta})d\tilde{\theta}]'\hat{\pi}_n(\theta)d\theta \\ &= \Sigma_n^{-1}\int_\Theta (\theta - \bar{\theta}_n)[\int_\Theta \left\{\{\bar{g}_n(\theta) - \bar{g}_n(\tilde{\theta})\} - \{g_n(\theta) - g_n(\tilde{\theta})\}\right\}\hat{\pi}_n(\tilde{\theta})d\tilde{\theta}]'\hat{\pi}_n(\theta)d\theta \\ &= \Sigma_n^{-1}\int_\Theta (\theta - \bar{\theta}_n)[\int_\Theta \left\{\partial_\theta \bar{g}_n(\check{\theta}) - \partial_\theta g_n(\check{\theta})\right\}\{\theta - \tilde{\theta}\}\hat{\pi}_n(\tilde{\theta})d\tilde{\theta}]'\hat{\pi}_n(\theta)d\theta \end{aligned}$$

for some intermediate values $\check{\theta}$ such that $\|\theta - \check{\theta}\| \leq \|\theta - \tilde{\theta}\|$ for each pair $\theta, \tilde{\theta}$. Using the arguments from the proofs of Theorem 1, for any $d \geq 1$ there exists a $C_K > 0$ such that,

with probability going to 1:

$$\begin{aligned}
&\int_\Theta (\theta-\bar\theta_n)[\int_\Theta \left\{\partial_\theta \bar g_n(\check\theta)-\partial_\theta g_n(\check\theta)\right\}\{\theta-\tilde\theta\}\hat\pi_n(\tilde\theta)d\tilde\theta]'\hat\pi_n(\theta)d\theta + o_p(n^{-d})\\
&= \int_{\|H_n^{-1}(\theta-\theta_0)\|\le C_K\tilde\kappa_n} (\theta-\bar\theta_n)[\int_{\|H_n^{-1}(\theta-\theta_0)\|\le C_K\tilde\kappa_n} \left\{\partial_\theta \bar g_n(\check\theta)-\partial_\theta g_n(\check\theta)\right\}\{\theta-\tilde\theta\}\hat\pi_n(\tilde\theta)d\tilde\theta]'\hat\pi_n(\theta)d\theta\\
&\quad + o_p(n^{-d})\\
&= \left(\int_\Theta (\theta-\bar\theta_n)(\theta-\bar\theta_n)'\hat\pi_n(\tilde\theta)d\tilde\theta]'\hat\pi_n(\theta)d\theta\right)\left(\partial_\theta \bar g_n(\theta_0)-\partial_\theta g_n(\theta_0)+o_p(n^{-1/2})\right)+o_p(n^{-d})\\
&= \Sigma_n\left(\partial_\theta \bar g_n(\theta_0)-\partial_\theta g_n(\theta_0)+o_p(n^{-1/2})\right)+o_p(n^{-d})
\end{aligned}$$

where the $o_p(n^{-1/2})$ term is due to the stochastic equicontinuity assumption for $\partial_\theta \bar g_n(\theta)$. Pick d large enough such that we have $\Sigma_n^{-1}o_p(n^{-d}) = o_p(n^{-1/2})$ and then:

$$B_{n,LS} - \bar B_{n,LS} = \partial_\theta \bar g_n(\theta_0) - \partial_\theta g_n(\theta_0) + o_p(n^{-1/2}).$$

This implies that $A_{n,LS}$ and $B_{n,LS}$ satisfy the following:

$$\sqrt{n}\begin{pmatrix} A_{n,LS}+B_{n,LS}\theta_0 \\ \mathrm{vec}(B_{n,LS}-\bar B_{n,LS})\end{pmatrix} = \sqrt{n}\begin{pmatrix} \bar g_n(\theta_0) \\ \mathrm{vec}\Big(\partial_\theta \bar g_n(\theta_0)-\partial_\theta g_n(\theta_0)\Big)\end{pmatrix} + o_p(n^{-1/2}) \xrightarrow{d} \mathcal{N}(0,V),$$

which concludes the proof. □

E.2 Primitive Conditions for Theorem 5

Proof of Lemma C3. To simplify notation, the proof will consider $\theta = (\alpha,\gamma)$ and H_n will be defined using derivatives of γ only in the below. By a re-parameterization, we have $\hat\gamma_{n,GMM} = \gamma_0 + H_n\hat h_n/\sqrt{n}$ yields via the Argmax Theorem (van der Vaart & Wellner, 1996):

$$\begin{aligned}
\hat h_n &= \mathrm{argmin}_h\Big(\mathbb{G}_n(\beta_0,\gamma_0)+\partial_\gamma g_n(\beta_0,\gamma_0)H_n h\Big)'\hat V_n^{-1}\Big(\mathbb{G}_n(\beta_0,\gamma_0)+\partial_\gamma g_n(\beta_0,\gamma_0)H_n h\Big)+o_p(1)\\
&= -\Big(H_n\partial_\gamma g_n(\beta_0,\gamma_0)'\hat V_n^{-1}H_n\partial_\gamma g_n(\beta_0,\gamma_0)\Big)^{-1}H_n\partial_\gamma g_n(\beta_0,\gamma_0)'\hat V_n^{-1}\mathbb{G}_n(\beta_0,\gamma_0)+o_p(1).
\end{aligned}$$

Consider an estimator $\hat\gamma_n$ satisfying $H_n^{-1}(\hat\gamma_n-\hat\gamma_{n,GMM}) = o_p(n^{-1/2})$, then:

$$\bar g_n(\beta_0,\hat\gamma_n) = (I-P_n)\,\bar g_n(\alpha_0,\gamma_0)+o_p(n^{-1/2}).$$

where $P_n = \partial_\gamma g_n(\alpha_0,\gamma_0)H_n\left(H_n\partial_\gamma g_n(\beta_0,\gamma_0)'\hat V_n^{-1}H_n\partial_\gamma g_n(\beta_0,\gamma_0)\right)^{-1}H_n\partial_\gamma g_n(\beta_0,\gamma_0)'\hat V_n^{-1}$. Using usual arguments, P_n is a projection matrix and:

$$S_n(\beta_0) = n\times \bar g_n(\beta_0,\hat\gamma_n)'\hat V_n^{-1}\bar g_n(\beta_0,\hat\gamma_n)\xrightarrow{d}\chi^2_{\dim(g)-\dim(\gamma)}.$$

This concludes the proof. □

Proof of Proposition C2. The first result is a consequence of Lemma C3. For the second results, note that for each $\ell = \ell^\star, \dots, \mathcal{L}$:

$$\left|\mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) - (1-\alpha)\right| = o(1).$$

Since this holds for finitely many ℓ, this implies that:

$$\sup_{\ell=\ell^\star,\dots,\mathcal{L}} \left|\mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) - (1-\alpha)\right| = o(1).$$

Note that:

$$\left|\inf_{\ell=\ell^\star,\dots,\mathcal{L}} \mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) - (1-\alpha)\right| \leq \sup_{\ell=j^\star,\dots,\mathcal{L}} \left|\mathbb{P}(S_{n,\ell} \leq c_{1-\alpha,\ell}) - (1-\alpha)\right| = o(1).$$

This concludes the proof. □

Appendix F Proofs for the Results of Appendix D

F.1 Weak or Set Identification

Proof. Proof of Lemma D4

a. Weak/set identification:

Under the stated assumptions, either Theorem 3 or Lemma 1 hold. These results indicate that $\lambda_{\min}(B_{n,LS/\infty}) = O_p(\tilde{\kappa}_n) = o_p(\underline{\lambda}_n)$ which implies the result.

b. (Semi)-strong identification:

Under the stated assumptions, $B_{n,LS/\infty}$ is such that:

$$[B_{n,LS/\infty} - \partial_\theta g_n(\theta_0)]H_n = o_p(n^{-1/2}\tilde{\kappa}_n^{-1})$$

Let $\|\cdot\|_\star$ be the spectral norm, i.e. the absolute value of the largest eigenvalue, which is well defined for real symmetric matrices. We have:

$$H_n^{-1}[B'_{n,LS/\infty}B_{n,LS/\infty}]^{-1}H_n^{-1} = H_n^{-1}[\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)]^{-1}H_n^{-1} + o_p(n^{-1/2}\tilde{\kappa}_n^{-1}).$$

Note that by definition of H_n, we have $\lambda(H_n\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0)H_n) = (1,\dots,1)$, i.e. all the eigenvalues are equal to 1. Using Problem III.6.14 in Bhatia (1997), this implies:

$$\lambda_{\min}(H_n^{-2}) \times \lambda_{\max}([B'_{n,LS/\infty}B_{n,LS/\infty}]^{-1}) \leq d_\theta + o_p(n^{-1/2}\tilde{\kappa}_n^{-1}).$$

Given that $\lambda_{\max}([B'_{n,LS/\infty}B_{n,LS/\infty}]^{-1}) = [\lambda_{\min}(B'_{n,LS/\infty}B_{n,LS/\infty})]^{-1}$ and $\lambda_{\min}(H_n^{-2}) = \lambda_{\max}(H_n^2) = [\lambda_{\min}(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0))]^{-1}$, this implies:

$$\frac{\lambda_{\min}(\partial_\theta g_n(\theta_0)'\partial_\theta g_n(\theta_0))}{d_\theta + o_p(n^{-1/2}\tilde{\kappa}_n^{-1})} \leq \lambda_{\min}(B'_{n,LS/\infty}B_{n,LS/\infty}).$$

Since by assumption $\underline{\lambda}_n^2 = o\left(\lambda_{\min}\left(\partial_\theta g_n(\theta_0)\partial_\theta g_n(\theta_0)'\right)\right)$, the above inequality implies that

$$\underline{\lambda}_n^2 = o_p\left(\lambda_{\min}(B'_{n,LS/\infty}B_{n,LS/\infty})\right),$$

which concludes the proof. □

Proof of Lemma D5. The first result is an immediate implication of the definition of V as the span of the identification failure, by definition it's orthogonal span a linear subset where the values of $\theta \in \Theta_0$ are unique. For the second result, suppose that the intersection is non-empty, pick $\theta \in \Theta_0 \cap \{\theta \in \Theta, P_{v^\star}\theta = c\}$. The following holds:

$$\begin{pmatrix} P_V^\perp \\ P_{v^\star} \end{pmatrix} \theta = \begin{pmatrix} P_V^\perp \theta_0 \\ c \end{pmatrix}.$$

The two terms on the right-hand side are unique since c is fixed and $P_V^\perp\Theta_0$ is a singleton. By construction $(P_V^\perp, P_V)$ is invertible, $(P_V^\perp, P_{v^\star})$ also has full rank since $\text{rank}(P_V P_{v^\star}) = \text{rank}(P_V)$ i.e. $P_{v^\star}$ preserves the span of P_V. This concludes the proof. □

Proof of Lemma D6. First, recall the definition of the identification failure: $V = \text{Span}(\theta_0 - \theta_1, \theta_0, \theta_1 \in \Theta_0)$, where Θ_0 is the weakly identified set of Definition 6. By definition of V and of projection matrices for any two $\theta_0, \theta_1 \in \Theta_0$: $P_V^\perp\theta_0 = P_V^\perp\theta_1$. This means that $P_V^\perp\theta_0$ is unique in Θ_0.

Pick $\ell \in \{1, \ldots, \mathcal{L}\}$, $\text{rank}(P_V P_{R_\ell}^\perp) < \text{rank}(P_V)$ implies that there exists a pair $\theta_0 \neq \theta_1$ in Θ_0 such that:

$$\begin{pmatrix} P_V^\perp \\ P_{R_\ell} \end{pmatrix} (\theta_0 - \theta_1) = 0$$

since the matrix on the left-hand side does not have full rank. This implies that: $\Theta_0 \cap \{\theta, R_\ell\theta = R_\ell\theta_0\}$ is neither empty nor a singleton since it contains $\theta_1 \neq \theta_0$ as well as θ_0. By Lemma D5, $\text{rank}(P_V P_{R_\ell}^\perp) = \text{rank}(P_V)$ implies that such sets are either empty or a singleton. By definition of $\ell^\star$, this is the case for any $\ell \in \{\ell^\star, \ldots, \mathcal{L}\}$.

Lemma D6 can thus be re-stated as $\hat{\ell}_n \geq \ell^\star$ with probability going to 1. Take $\ell < \ell^\star$, there exists a basis $v_1, \ldots, v_{\text{rank}(R_\ell)}$ such that $P_{R_\ell}^\perp v_j = 0$ for all $j \in \{1, \ldots, \text{rank}(R_\ell)\}$. This

implies that 0 is an eigenvalue of $B_{n,LS/\infty}P_{R_\ell}^\perp$ with multiplicity (at least) $\text{rank}(R_\ell)$. Now, we only have to find one more eigenvalue which is less than $\underline{\lambda}_n$ with probability going to 1 to show that $\ell \neq \hat{\ell}_n$ with probability going to 1.

Since $\text{rank}(P_V P_{R_\ell}) < \text{rank}(P_V)$, there exists $v^\star \neq 0$ such that $v^\star \in V$, $P_V^\perp v^\star = 0$, $P_{R_\ell} v^\star = 0$ and $P_{R_\ell}^\perp v^\star \neq 0$. By construction, the family $(v_1, \ldots, v_{\text{rank}(R_\ell)}, v^\star)$ has rank equal to $[\text{rank}(R_\ell) + 1]$. Also, $v^\star$ is such that $P_{R_\ell}^\perp v^\star = v^\star$ so that $B_{n,LS/\infty} v^\star = O_p(\kappa_n)$. This implies that $|\lambda_{d_\theta - \text{rank}(R_\ell)}(B_{n,LS/\infty} P_{R_\ell}^\perp)| = O_p(\kappa_n)$ which is strictly less than $\underline{\lambda_n}$ with probability going to 1. This implies that $B_{n,LS/\infty} P_{R_\ell}^\perp$ has, with probability going to 1, at least $[\text{rank}(R_\ell) + 1]$ eigenvalues which are strictly less than $\underline{\lambda}_n$ so that the smallest $[\text{rank}(R_\ell) + 1]$ eigenvalues: $\lambda_{d_\theta}(B_{n,LS/\infty} P_{R_\ell}^\perp), \ldots, \lambda_{d_\theta - \text{rank}(R_\ell)}(B_{n,LS/\infty} P_{R_\ell}^\perp)$ are $o_p(\underline{\lambda}_n)$. This implies that $\ell \neq \hat{\ell}_n$ with probability going to 1.

Now, to show that $\hat{\ell}_n \geq \ell^\star$ with probability going to 1, consider the family-wise probability over $\ell \in \{1, \ldots, \ell^\star - 1\}$ for the event $\{\lambda_{d_\theta - \text{rank}(R_\ell)}(B_{n,LS/\infty} P_{R_\ell}^\perp) < \underline{\lambda_n}\}$:

$$\begin{aligned}
\mathbb{P}\left(\hat{\ell}_n < \ell^\star\right) &= \mathbb{P}\left(\max_{\ell=1,\ldots,\ell^\star-1} \left[\lambda_{d_\theta - \text{rank}(R_\ell)}(B_{n,LS/\infty} P_{R_\ell}^\perp)\right] > \underline{\lambda}_n\right) \\
&\leq \sum_{\ell=1}^{\ell^\star - 1} \mathbb{P}\left(\lambda_{d_\theta - \text{rank}(R_\ell)}(B_{n,LS/\infty} P_{R_\ell}^\perp) > \underline{\lambda}_n\right) \to 0,
\end{aligned}$$

as $n \to \infty$ since the sum is finite and with element going to 0. This concludes the proof. □

F.2 Higher-Order Identification

Proof of Lemma D7. If $B_{n,\ell} = (B_n', R_\ell')'$ is singular then $\bar{h}^2 = +\infty$. Suppose it is not singular, Definition 5 and the stated assumptions imply that:

$$\begin{aligned}
&\|R_\ell(\theta - \theta_0)\| + \underline{\lambda}^{-1} \sum_{j=1}^{r} \underline{C}_j \|P_j(\theta - \theta_0)\|^j + O_p(n^{-1/2}) \leq \|R_\ell(\theta - \theta_0)\| + \|\bar{g}_n(\theta)\|_{W_n} \\
\leq &\|R_\ell(\theta - \theta_0)\| + \overline{\lambda} \sum_{j=1}^{r} \overline{C}_j \|P_j(\theta - \theta_0)\|^j + O_p(n^{-1/2}).
\end{aligned}$$

By assumption $R_\ell \theta_0 = c_\ell$. Given that $\text{rank}(P_{V_r} P_{R_\ell}) < \text{rank}(P_{V_r})$, there exists $j_\ell \geq 2$, $v_{j_\ell} \in \text{Span}(P_{j_\ell})$ non-zero with $\|v_{j_\ell}\| = 1$ such that $P_{R_\ell} v_{j_\ell} = 0$ and for $h \in \mathbb{R}$ not too large:

$$\|\bar{g}_n(\theta_0 + h\kappa_n^{1/j_\ell} v_{j_\ell})\|_{W_n} + \|R_\ell(\theta_0 + h\kappa_n^{1/j_\ell} v_{j_\ell}) - c_\ell\| \leq \overline{\lambda} \times \overline{C}_{j_\ell} |h|^{j_\ell} \kappa_n + O_p(n^{-1/2}) \leq 3/4 \times \kappa_n$$

with probability going to 1 when $|h| \leq 1/[2\underline{C}_j \overline{\lambda}]^{1/j_\ell}$. This implies that $|h| = 1/[2\underline{C}_j \overline{\lambda}]^{1/j_\ell}$ is in the maximization set of Algorithm 2 with probability going to 1. The following shows

that the criterion will diverge for this choice of h. By definition of $\hat{R}_n$ and the assumptions on the behaviour of $\bar{g}_n$ around θ_0:

$$\|\hat{R}_n(\theta_0 + h\kappa_n^{1/j_\ell} v_{j_\ell})\| \geq \underline{\tilde{C}}_{j_\ell} |h|^{j_\ell} \times \kappa_n + O_p(n^{-1/2}).$$

Now note that for any θ, we have:

$$\sup_{\|v\|=1} \frac{\hat{R}_n(\theta)' \tilde{B}'_{W,\ell} v v' \tilde{B}_{W,\ell} \hat{R}_n(\theta)}{v' \tilde{B}_{W,\ell} V_1 \tilde{B}'_{W,\ell} v} = \sup_{\|\tilde{v}\|=1} \frac{\hat{R}_n(\theta)' \tilde{v} \tilde{v}' \hat{R}_n(\theta)}{\tilde{v}' V_1 \tilde{v}}.$$

This holds because (B'_n, R'_ℓ) is non-singular by assumption so that $B_{W,\ell}$ is also non-singular. Given the bounds on the eigenvalues of V_1, the following inequality holds:

$$\frac{\hat{R}_n(\theta)' \tilde{v} \tilde{v}' \hat{R}_n(\theta)}{\overline{\lambda}_V \tilde{v}' \tilde{v}} + O_p(n^{-1/2}) \leq \frac{\hat{R}_n(\theta)' \tilde{v} \tilde{v}' \hat{R}_n(\theta)}{\tilde{v}' V_1 \tilde{v}} \leq \frac{\hat{R}_n(\theta)' \tilde{v} \tilde{v}' \hat{R}_n(\theta)}{\underline{\lambda}_V \tilde{v}' \tilde{v}} + O_p(n^{-1/2}).$$

Since $\tilde{v}$ covers the full unit circle, there exists a $\tilde{v} \neq 0$ over the optimizing set such that $\hat{R}_n(\theta)' \tilde{v}' \tilde{v} \hat{R}_n(\theta) = \|\hat{R}_n(\theta)\|_\infty^2$ (the sup-norm). This, in turn, implies the following inequality:

$$\|\hat{R}_n(\theta)\|_\infty^2 / \overline{\lambda}_V + O_p(n^{-1/2}) \leq \sup_{\|\tilde{v}\|=1} \frac{\hat{R}_n(\theta)' \tilde{v} \tilde{v}' \hat{R}_n(\theta)}{\tilde{v}' V_1 \tilde{v}}.$$

Also note that $\|\hat{R}_n(\theta)\|_\infty^2 \geq \|\hat{R}_n(\theta)\|^2 / \dim(g_n)^2$ by equivalence between norms. Combining this with the previous result, we have with probability going to 1:

$$\bar{h}_n^2 \geq n \times \|\hat{R}_n(\theta_0 + h\kappa_n^{1/j_\ell} v_{j_\ell}))\|_\infty^2 / \overline{\lambda}_V \times (1 + o_p(1))^2 \geq \frac{\underline{\tilde{C}}_{j_\ell} |h|^{j_\ell}}{\overline{\lambda}_V \times \dim(g_n)^2} \times [\sqrt{n} \times \kappa_n]^2 \times (1 + o_p(1)).$$

Assumption 1 implies that $\sqrt{n} \times \kappa_n \to +\infty$ which implies that $\bar{h}^2 \to +\infty$ with probability going to 1. This concludes the proof. □

Proof of Lemma D8. From Definition 4 and Assumptions 1, 2, for $\|\theta - \theta_0\| \leq \varepsilon$:

$$\begin{aligned}
\|\bar{g}_n(\theta)\|_{W_n} &\geq \underline{C} \|\partial_\theta g_n(\theta_0)(\theta - \theta_0)\| \times (1 + o_p(1)) - O_p(n^{-1/2}) \\
&\geq \underline{C} \sqrt{\lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))} \times \|\theta - \theta_0\| \times (1 + o_p(1)) - O_p(n^{-1/2}).
\end{aligned}$$

Hence, with probability going to 1, $\|\theta - \theta_0\| \geq 2\kappa_n / [\underline{C} \sqrt{\lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))}]$ implies that $\|\bar{g}_n(\theta)\|_{W_n} \geq 3/2\kappa_n > \kappa_n$. Also, $\|\theta - \theta_0\| > \varepsilon$ implies $\|\bar{g}_n(\theta)\|_{W_n} \geq 3/2\kappa_n$ by point identification. Then, under the stated assumptions:

$$\begin{aligned}
\sup_{\theta, \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n} \|\hat{R}_n(\theta)\| &\leq \sup_{\theta, \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n} \bar{C}_n \|\theta - \hat{\theta}_n\|^2 \leq 4 \times \bar{C}_n \sup_{\theta, \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n} \|\theta - \theta_0\|^2 \\
&\leq 16 \times \bar{C}_n / \underline{C} \times \kappa_n^2 / \lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0)).
\end{aligned}$$

Since $\partial_\theta \bar{g}_n(\hat{\theta}_n)$ is non-singular, using a similar argument as in Lemma D7 implies: $\bar{h}_n^2 \leq O_p([\sqrt{n}\kappa_n^2 / \lambda_{\min}(\partial_\theta g_n(\theta_0)' \partial_\theta g_n(\theta_0))]^2) = o_p(1)$, which concludes the proof. □

Appendix G Additional Monte-Carlo Simulations

G.1 Example 1. Non-Linear Least Squares

Figure G6 compares three power curves for testing the null hypothesis $\theta_1 = [c+\delta]/\sqrt{n}$. The true value is $\theta_{1,0} = c/\sqrt{n}$, δ is the Pittman drift coefficient. The tests considered are an oracle t-statistic based on a simple re-parameterization ($y_i = \theta_1 x_{1,i} + \tilde{\theta}_2 x_{2,i} + e_i$), a projected Anderson-Rubin confidence set and the two-step approach with the local rule-of-thumb. For $c = 0$, the two-step approach is nearly identical to AR inferences; the oracle has higher power. For $c = 1$, AR inferences have non-monotonic power and are more powerful around $\delta + c = 0$,[5] the oracle has higher power elsewhere; the two-step approach is in-between the two. For $c = 3$, the two-step approach nearly coincides with the maximum of the AR and oracle's power curves suggesting improvements over both the robust and the reduced-form approach.

G.2 Example 2. Possibly Non-Invertible MA Model

The second example is the MA(1) model:

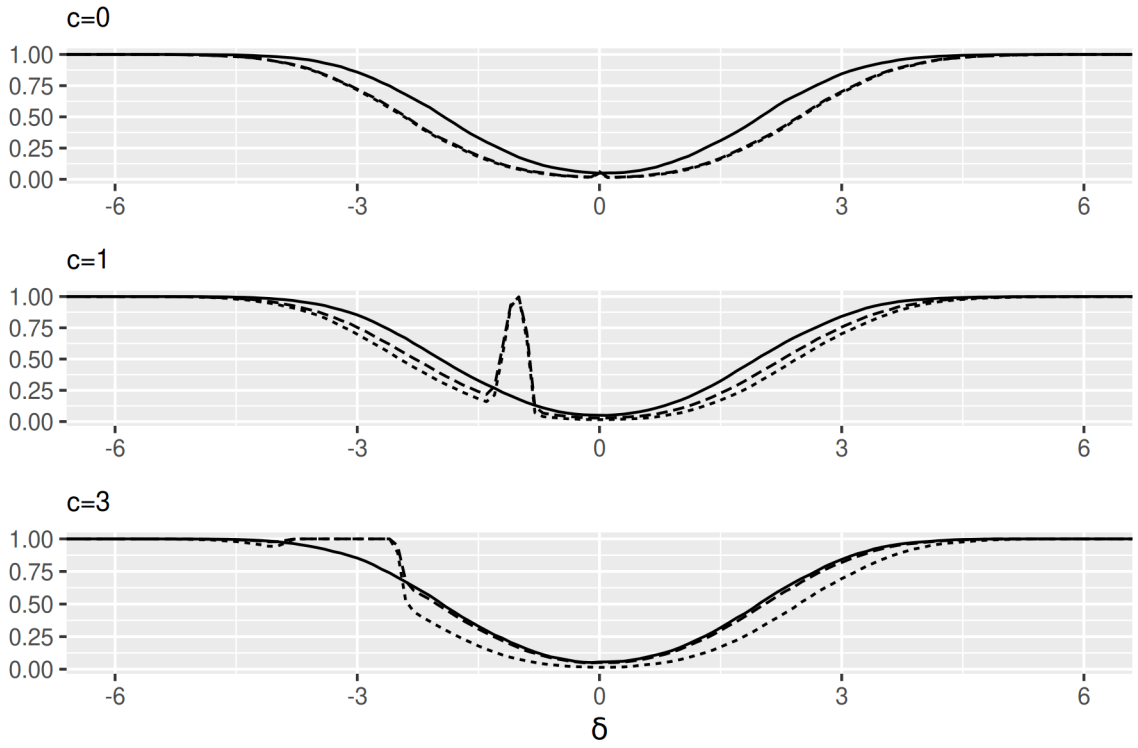
$$y_t = \sigma[e_t - \vartheta e_{t-1}],$$

with e_t iid distributed from a Generalized Extreme Value distribution with mean 0, variance 1 and skewness τ. For $\tau = 0$, (ϑ, σ) is not identified and weakly identified for $\tau_n \asymp n^{-1/2}$. The estimating moments for the parameters $\theta = (\vartheta, \sigma, \tau)$ are:

$$\bar{g}_n(\theta) = \frac{1}{n}\sum_{t=2}^{n}\left(y_t^2 - [1+\vartheta^2]\sigma^2, y_t y_{t-1} + \vartheta\sigma^2, y_t^3 + [1-\vartheta^3]\sigma^3\tau, y_t^2 y_{t-1} + \vartheta\sigma^3\tau, y_t y_{t-1}^2 + \theta\sigma^3\tau\right)'.$$

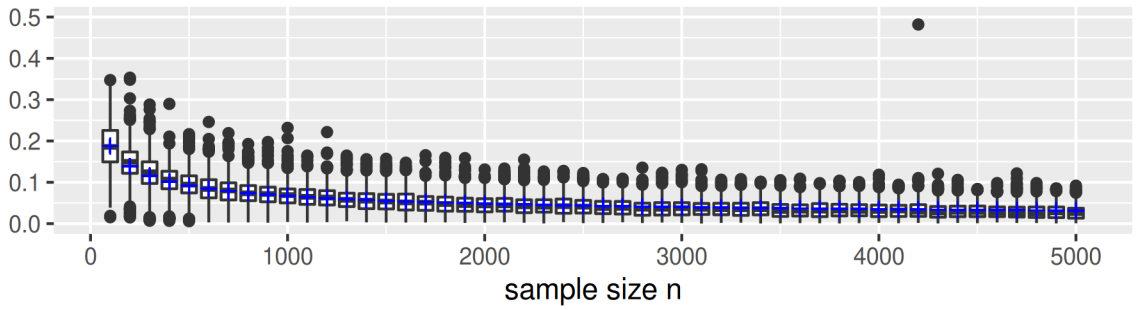
[5] $\delta + c = 0 \Rightarrow \tilde{\theta}_2 = 0$ under the null, the oracle is based on a reduced form specification and does not use that information. Hence, it has lower power against this alternative.

Figure G6: Power Comparison: Oracle, Projection and Two-Step Inferences



Note: $y_i = \theta_{1,n} x_{i,1} + \theta_{1,n}\theta_2 x_{i,2} + e_i$, $\theta_{1,n} = c \times n^{-1/2}$, for $c \in \{0, 1, 3\}$ (top/middle/bottom panel), $n = 1,000$, $B = 5,000$ Monte-Carlo replications and $\kappa_n = \sqrt{2\log(\log[n])} n^{-1/2}$. **Legend:** Oracle (solid) - OLS based reduced-form inference, the reduced form model is $y_i = \theta_{1,n} x_{i,1} + \tilde{\theta}_2 x_{i,2} + e_i$; Anderson-Rubin (dashed) - projection CI; Rule-of-Thumb (dotted) - QLR based two-step CI with data-driven rule-of-thumb. y-axis: rejection rate when testing $\theta_1 = \theta_{1,n} + \delta \times n^{-1/2}$. x-axis: Pittman coefficient δ.

Figure G7: Distribution of $\lambda_{\min}(B_{n,LS})$ and sample size n



Note: $y_t = \sigma[e_t - \vartheta e_{t-1}]$, $e_t \sim (0,1), \mathbb{E}(e_t^3) = 2 \times n^{-1/2}$, $(\vartheta_0, \sigma_0) = (0.5, 1)$, $100 \leq n \leq 5{,}000$, $B = 500$ Monte-Carlo replications and $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_4)}, \sqrt{2\log(\log[n])n^{-1/2}})$. **Legend:** Black lines - boxplot of the distribution of $\lambda_{\min}(B_{n,LS})$ for each n; Blue crosses - fitted rate from regressing the Monte-Carlo $\lambda_{\min}(B_{n,LS})$ draws on κ_n by OLS with no intercept.

Figure G7 plots the distribution of $\lambda_{\min}(B_{n,LS})$ against the predicted rate from Theorem 1. There is a value of $\lambda_{\min}(B_{n,LS})$ which is much larger ($\simeq 0.5$) than the rest of the distribution. A closer investigation into this draw reveals that the set $\hat{\Theta}_n = \{\theta, \|\bar{g}_n\|_{W_n} - \inf_{\theta\in\Theta} \|\bar{g}_n\|_{W_n} \leq \kappa_n\}$ (which is a conservative identification robust confidence set) is centered around a single point. This implies that, for this occurrence robust and standard inferences would be similar.

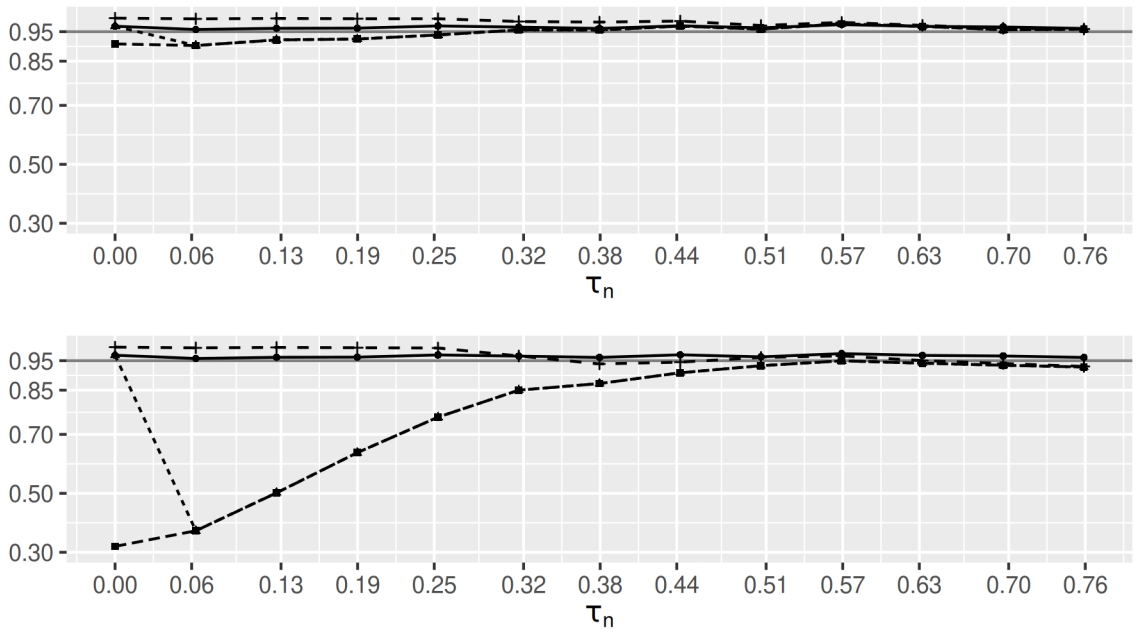
The top panel of Figure G8 shows the coverage of the two-step, Anderson-Rubin and QLR 95% confidence sets. The projection-based confidence sets are computed using: $\{\vartheta, n \times \inf_{\sigma,\tau} \|\bar{g}_n(\vartheta,\sigma,\tau)\|^2_{\hat{V}_n^{-1}} \leq c_{1-\alpha}\}$. The projected Anderson-Rubin confidence set assumes that the econometrician knows σ and τ are point identified when ϑ is fixed. The critical value is the 95% quantile of a χ^2_2 distribution.

The true value $\vartheta_0 = 2$ lies outside the unit circle. For this design, the unconstrained estimator $\hat{\vartheta}_n$ is biased towards $\vartheta = 0.5$, which is inside the unit circle. This leads to some size distortion for QLR/Wald inferences. The two-step procedure with the non-local rule-of-thumb is not too significantly size distorted.

The two-step approach uses the rule-of-thumb and $\underline{\lambda_n} = \sqrt{\log[n]}$ as cutoffs. The sequential search fixes ϑ, as implied by H_0, then σ and finally τ - with critical values corresponding the 95% quantile of a χ^2_2, χ^2_3 and χ^2_4 distribution respectively. When $\lambda_{\min}(B_{n,LS}) > \underline{\lambda}_n$, the tests switches to a QLR/Wald statistic with critical values corresponding the 95% quantile of a χ^2_1 distribution. The rule-of-thumb implemented here relies on the between and within cluster rule-of-thumb described in Appendix C.1.2. First, the set $\hat{\Theta}_n = \{\theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_{\theta\in\Theta} \|\bar{g}_n\|_{W_n} \leq \kappa_n\}$ is split into clusters using the k-means algorithm when the estimated

clusters display enough separation.[6] The results rely on $B_{n,LS}$ computed on the whole parameter space. However, H_0 fixes the cluster and using a $B_{n,LS}$ computed under H_0 could lead to a less conservative decision rule and power improvements. This was not investigated in the simulations.

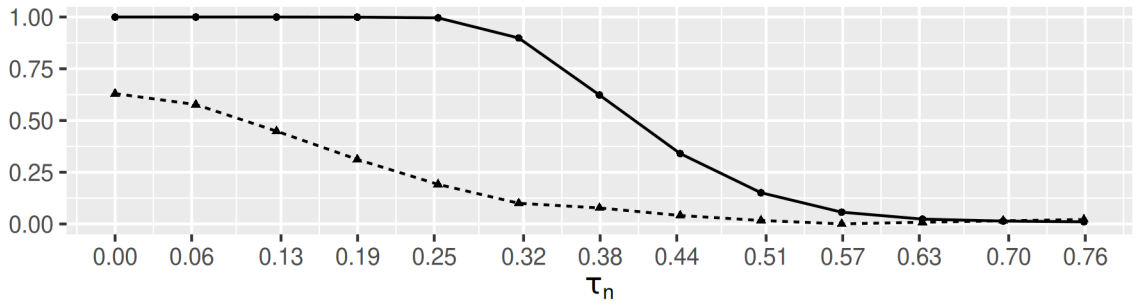
Figure G8: Coverage of the 95% Confidence Intervals



Note: $y_t = \sigma[e_t - \vartheta e_{t-1}]$, $e_t \sim (0,1), \mathbb{E}(e_t^3) = \tau_n = c \times n^{-1/2}$, $c \in [0, 24]$, $(\vartheta_0, \sigma_0) = (2, 1)$, $n = 1,000$, $B = 2,000$ Monte-Carlo replications and $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_4)}, \sqrt{2\log(\log[n])n^{-1/2}})$. **Legend:** Anderson-Rubin (solid/dot) - projection CI; Standard (dashed/square) - QLR (top panel)/Wald (bottom panel) CI; Two-step (dashed/cross) - two-step procedure with $\underline{\lambda}_n$ = data-driven rule-of-thumb; $\sqrt{\log(n)}$ (dotted/triangle) - two-step procedure with $\underline{\lambda}_n = \sqrt{\log(n)}$.

[6]For these simulations, the criterion to pick the number of clusters relied on a ratio of the distance between the centroids of the clusters to the diameter of these clusters. Other approaches may be considered.

Figure G9: Detection of Identification Failure and Size Distortion of the Wald test



Note: $y_t = \sigma[e_t - \vartheta e_{t-1}]$, $e_t \sim (0,1), \mathbb{E}(e_t^3) = \tau_n = c \times n^{-1/2}$, $c \in [0, 24]$, $(\vartheta_0, \sigma_0) = (2, 1)$, $n = 1,000$, $B = 5,000$ Monte-Carlo replications and $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_4)}, \sqrt{2\log(\log[n])n^{-1/2}})$. **Legend:** Size distortion (dotted/triangle) - size distortion of a 95% Wald confidence interval; Two-step (dashed/square) - detection rate for identification failure using the rule-of-thumb.

G.3 Example 3. Second-Order Identified Non-Linear Least-Squares

The third example considers the simple non-linear least-squares model:

$$y_i = \theta_1 x_{i,1} + \theta_{2,n}(\theta_{2,n} - \theta_1)^2 x_{i,2} + e_i, \quad x_{i,1}, x_{i,2} \sim \mathcal{N}(0, I_2)$$

where $\theta_{2,n}$ is bounded away from zero. This model can be estimated using the following moment conditions:

$$g_n(\theta) = \mathbb{E}\left(y_i(x_{i,1}, x_{i,2})'\right) - \left(\theta_1,\, \theta_2(\theta_2 - \theta_1)^2\right)'.$$

This model is second-order identified. Consider the re-parameterization $(\vartheta_1, \vartheta_2) = (\theta_1, \theta_2 - \theta_1)$: $g_n(\vartheta) = g_n(\vartheta_0) - (\vartheta_1,\, [\vartheta_1 + \vartheta_2]\vartheta_2^2)'$. For ϑ_1 bounded away from zero and $\vartheta_{2,n} = c \times n^{-1/4}$ the second-order term is non-negligible in the Taylor expansion of g_n around the true $(\vartheta_1, \vartheta_{2,n})$. When $c = 0$, $\partial_\theta g_n(\theta_0)$ is singular so that the higher-order identification failure problem is summarized by the rank of the Jacobian. However, when $c \neq 0$ but small, the Jacobian is non singular and the information from the eigenvalues may be misleading. Even though the Jacobian has full rank, the moments are not approximately linear around θ_0 which results in non-standard asymptotics.

This is illustrated in Figure G10 which plots the distribution of the $\lambda_{\min}(B_{n,LS})$, $\lambda_{\min}(\partial_\theta g_n(\theta_0))$ and $\lambda_{\min}(\partial_\theta g_n(\hat{\theta}_n))$. The Figure shows a disconnect between the quasi-Jacobian and the Jacobian matrices under higher-order identification.

To further illustrate this disconnect, consider a sample simulated with $n = 1,000$, $c = 0.4$. The estimated quasi-Jacobian and Jacobians are:

$$B_{n,LS} = \begin{pmatrix} 0.01 & -1.00 \\ 0.02 & 0.01 \end{pmatrix}, \partial_\theta \bar{g}_n(\hat{\theta}_n) = \begin{pmatrix} 1.08 & -1.15 \\ -0.96 & -0.04 \end{pmatrix}, \partial_\theta \bar{g}_n(\theta_0) = \begin{pmatrix} -0.46 & -1.02 \\ 0.42 & 0.02 \end{pmatrix}.$$

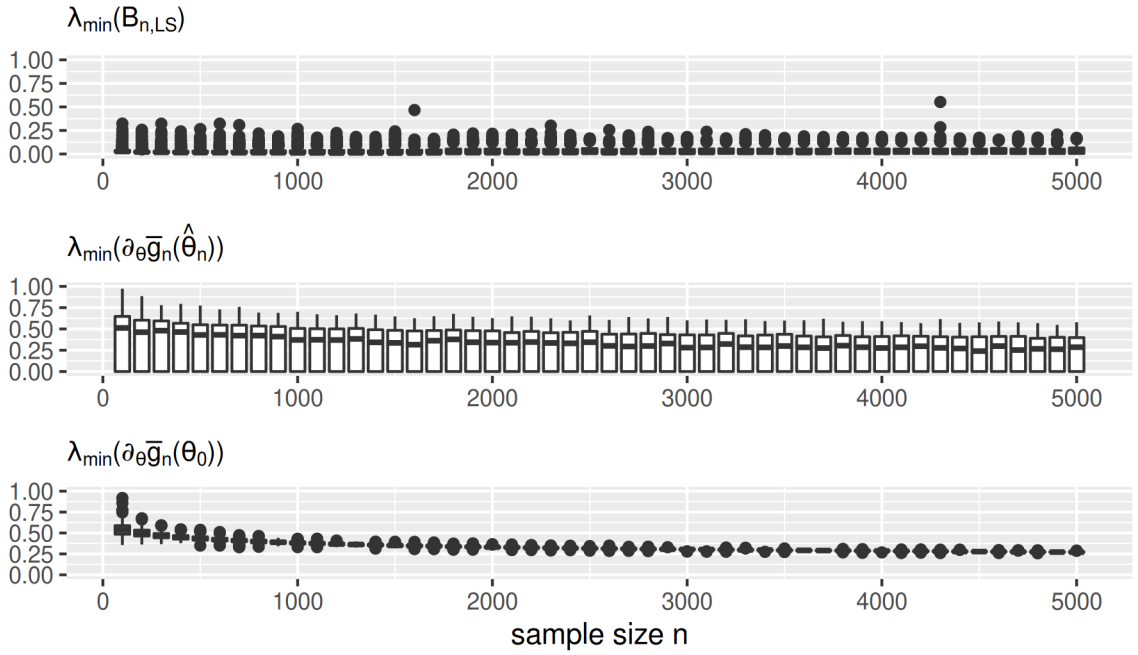
In line with Theorem 2, the quasi-Jacobian is close to being singular while the Jacobians are not. Now setting $c = 2$, which is closer to (semi)-strong identification yields:

$$B_{n,LS} = \begin{pmatrix} -2.33 & 2.14 \\ -1.08 & 0.08 \end{pmatrix}, \partial_\theta \bar{g}_n(\hat{\theta}_n) = \begin{pmatrix} -2.33 & 2.14 \\ -1.08 & 0.08 \end{pmatrix}, \partial_\theta \bar{g}_n(\theta_0) = \begin{pmatrix} -2.15 & 1.99 \\ -1.08 & 0.07 \end{pmatrix}.$$

This is in line with the predictions of Theorem 1. Finally, for $c = 0$, these matrices become:

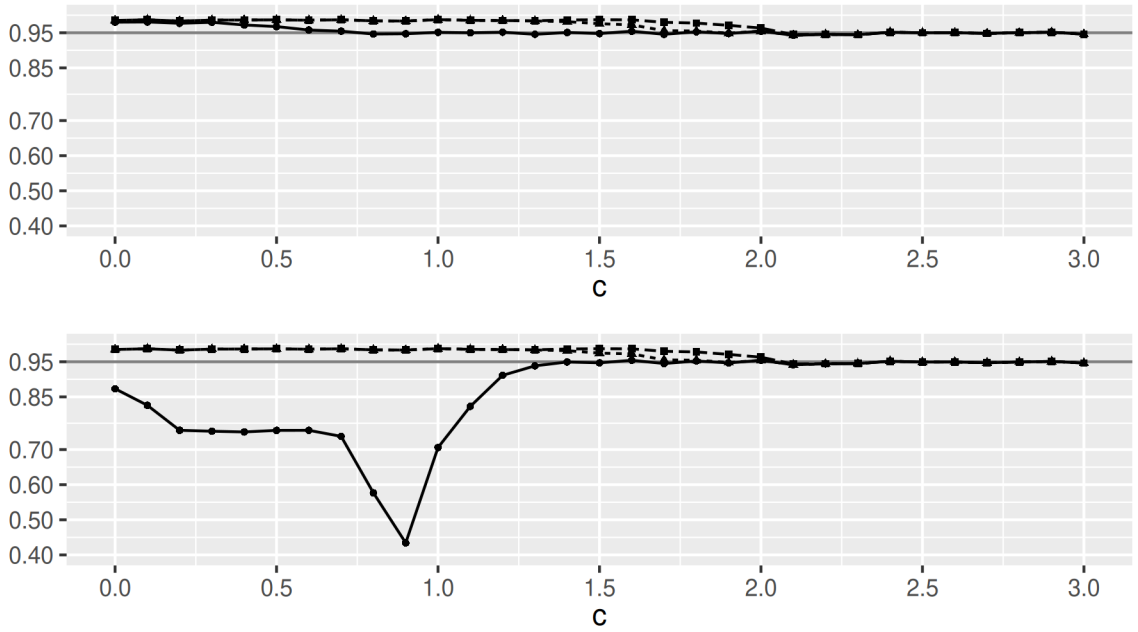
$$B_{n,LS} = \begin{pmatrix} -0.15 & -0.04 \\ -1.00 & -0.00 \end{pmatrix}, \partial_\theta \bar{g}_n(\hat{\theta}_n) = \begin{pmatrix} 0.98 & -1.04 \\ -0.96 & -0.04 \end{pmatrix}, \partial_\theta \bar{g}_n(\theta_0) = \begin{pmatrix} -0.04 & 0.00 \\ -1.00 & 0.00 \end{pmatrix}.$$

Figure G10: Distribution of $\lambda_{\min}(B_{n,LS})$, $\lambda_{\min}(\partial_\theta \bar{g}_n(\hat{\theta}_n))$, $\lambda_{\min}(\partial_\theta \bar{g}_n(\theta_0))$ and sample size n



Note: $y_i = \theta_1 x_{i,1} + \theta_{2,n}(\theta_{2,n} - \theta_1)^2 x_{i,2} + e_i$, $e_i, x_{i,1}, x_{i,2} \sim \mathcal{N}(0, I_3)$, $(\theta_1, \theta_{2,n}) = (3,\ 3 + 0.4 \times n^{-1/4})$, $100 \leq n \leq 5,000$. $B = 500$ Monte-Carlo replications and $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_2)}, \sqrt{2\log(\log[n])}n^{-1/2})$.

Figure G11: Coverage of the 95% Confidence Intervals



Note: $y_i = \theta_1 x_{i,1} + \theta_{2,n}(\theta_{2,n} - \theta_1)^2 x_{i,2} + e_i$, $e_i, x_{i,1}, x_{i,2} \sim \mathcal{N}(0, I_3)$, $(\theta_1, \theta_{2,n}) = (3, 3 + c \times n^{-1/4})$, $c \in [0, 3]$, $n = 1,000$, $B = 2,000$ Monte-Carlo replications and $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_2)}, \sqrt{2\log(\log[n])}n^{-1/2})$.
Legend: Standard (solid/dot) - Wald-based confidence interval; Two-step - Rule 1 (dotted/triangle) - two-step procedure rule-of-thumb described in Algorithm 2 without B_n in $\bar{h}^2$; Two-step - Rule 2 (dashed/square) - two-step procedure rule-of-thumb described in Algorithm 2 with $B_n = B_{n,LS}$ in $\bar{h}^2$.

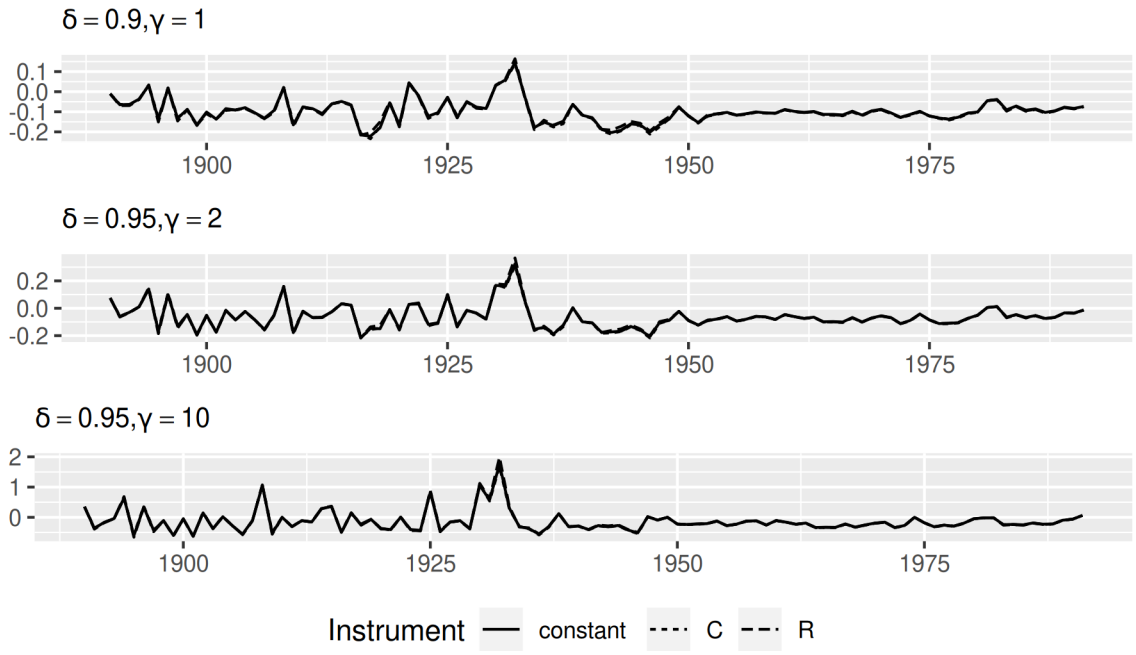
The Jacobians are informative when local identification fails exactly ($c = 0$), but can be misleading in intermediate cases where the first-order term is not singular yet higher-order terms matter.

The top panel in Figure G11 illustrates projection and standard inferences as well as the two-step procedure from Algorithm 2. The subvector hypothesis considered here is: $H_0 : \theta_2 = \theta_{2,n}$, with $\theta_{2,n} = \theta_1 + c \times n^{-1/4} = 3 + 0.4 \times n^{-1/4}$. The criterion $\bar{h}^2$ in the algorithm is computed in two ways, the first ignores the B_n in the formula (since $B_n' v' = u'$ for some vector u when B_n is non-singular) and the second uses the formula in the algorithm but with $B_w = B_{n,LS}$. To simplify the Monte-Carlo exercise, the procedure switches between Anderson-Rubin based full projection inference and a QLR/Wald test. Some power improvements could be made by checking if θ_1 is semi-strongly identified when θ_2 is fixed as in the other examples.

Appendix H Additional Empirical Results

H.1 US Euler Equation

Figure H12: US Euler Equation - Time-Series Plot of the Moments $g_t(\delta, \gamma)$



Note: annual series $g_t(\delta, \gamma) = [\delta(\frac{C_t}{C_{t-1}})^{-\gamma} R_t - 1] Z_{j,t}$ with $Z_{j,t} = 1$, $C_{t-1}/C_{t-2}, R_{t-1}$ for $j \in \{1, 2, 3\}$ respectively and a fixed value of $\theta = (\delta, \gamma)$.

H.2 Quantile IV

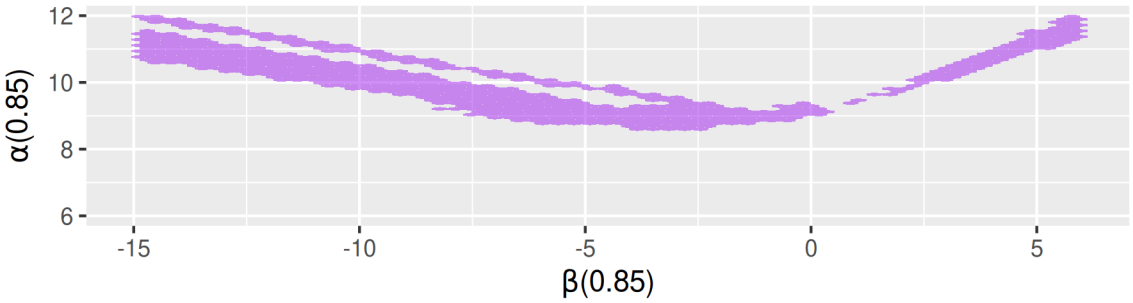
This additional empirical application deals with the quantile IV model of Chernozhukov & Hansen (2005) using the Fish data of Chernozhukov et al. (2007). For a given quantile τ, the estimating moments for fish demand are:

$$\bar{g}_n(\theta(\tau), \tau) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(\log(q_i) - [\alpha(\tau) + \beta(\tau) \log(p_i)] \leq -\tau)} \times (1, z_i')',$$

where q_i is the quantity of fish sold, p_i is the endogenous price of fish and, z_i is the vector of exogenous instruments - an indicator for stormy weather and another for mixed weather conditions. The following will focus on a specific quantile: $\tau = 0.85$. Since the model is non-smooth and non-linear it is not possible to check the relevance of the instrument

using a first-stage F-statistic. The bounds used to compute the integral in $B_{n,LS}$ are:[7] $\theta(\tau) = (\alpha(\tau), \beta(\tau)) \in [6, 12] \times [-15, 6]$. The sample consists of $n = 111$ observations. Figure

Figure H13: Demand for Fish - $\hat{\Theta}_n = \{\theta \in \Theta, \|\bar{g}_n(\theta)\|_{W_n} - \inf_{\theta \in \Theta} \|\bar{g}_n(\theta)\|_{W_n} \leq \kappa_n\}$ for $\tau = 0.85$



Note: region $\hat{\Theta}_n$ computed for $\kappa_n = \max(\sqrt{q_{0.99}(\chi^2_3)}, \sqrt{2\log(\log[n])}n^{-1/2})$ where $q_{0.99}$ is the 99% quantile of a χ^2_3 distribution. $W_n = \hat{V}_n^{-1}$. $\alpha(0.85)$ = intercept, $\beta(0.85)$ = slope.

H13 shows the region $\hat{\Theta}_n$ selected by the compact kernel to compute the integrals. It suggests either set or weak identification. The eigenvalues are (0.027, 0.005) which is quite small. The cutoff $\underline{\lambda}_n = 0.15$ is greater than both eigenvalues. The 95% level robust confidence set for $\beta(0.85)$ is $[-14.81, -11.29] \cup [-2.88, 0.03] \cup [5.55, 5.83]$.

[7] The grid was constructed using $20,000$ points from the Sobol sequence.