BOOTSTRAP DIAGNOSTIC TESTS

Giuseppe Cavaliere[a,b], Luca Fanelli[a] and Iliyan Georgiev[a]

September 30th, 2025

Violation of the assumptions underlying classical (Gaussian) limit theory often yields unreliable statistical inference. This paper shows that the bootstrap can detect such violations by delivering simple and powerful diagnostic tests that (a) induce no pre-testing bias, (b) use the same critical values across applications, and (c) are consistent against deviations from asymptotic normality. The tests compare the conditional distribution of a bootstrap statistic with the Gaussian limit implied by valid specification and assess whether the resulting discrepancy is large enough to indicate failure of the asymptotic Gaussian approximation. The method is computationally straightforward and only requires a sample of i.i.d. draws of the bootstrap statistic. We derive sufficient conditions for the randomness in the data to mix with the randomness in the bootstrap repetitions in a way such that (a), (b) and (c) above hold. We demonstrate the practical relevance and broad applicability of bootstrap diagnostics by considering several scenarios where the asymptotic Gaussian approximation may fail, including weak instruments, non-stationarity, parameters on the boundary of the parameter space, infinite variance data and singular Jacobian in applications of the delta method. An illustration drawn from the empirical macroeconomic literature concludes.

Keywords: Bootstrap inference; Pre-testing bias; Random bootstrap measures.

[a] Department of Economics, University of Bologna, Italy.
[b] Department of Economics, University of Exeter, UK.
We thank for important comments and suggestions Aureo De Paula, Xavier D'Haultfoeuille, Samuel Engle, Patrik Guggenberger, Bruce Hansen, Jessie Li, Adam McCloskey, Simone Martinalli, Marcelo Moreira, Morten Nielsen, Guillaume Pouliot, Anders Rahbek, Jesse Shapiro, Michael Wolf, Edoardo Zanelli, seminar participants at ESSEC, LSE, Nova SBE, UCL, U Copenhagen, U Lisbon as well as participants at the Aarhus Workshop in Econometrics IV (U Aarhus), the 2025 Bristol Econometric Study Group Annual Conference, the 'Econometrics in Rio' conference (FGV Rio, 2024), the 35th $(EC)^2$ Conference (University of Amsterdam), the 5th High Voltage Econometrics workshop (Collegio Carlo Alberto), the 11th Italian Congress of Econometrics and Empirical Economics (U Palermo) and the 2025 IAAE Conference (U Torino). We also thank Filippo Nardoni for research assistance. Cavaliere and Fanelli gratefully acknowledge financial support from the Italian Ministry of University and Research (PRIN 2020 Grant 2020B2AKFW and PRIN 2022 Grant 20229PFAX5).
Correspondence to: Giuseppe Cavaliere, Department of Economics, University of Bologna; email: giuseppe.cavaliere@unibo.it.

1 INTRODUCTION

CONSIDER A STANDARDIZED ESTIMATOR $T_n := n^{1/2}(\hat{\theta}_n - \theta_0)/\hat{\sigma}_n$ based on a data sample D_n, where $\hat{\sigma}_n^2$ is an estimator of the asymptotic variance. Classical asymptotic theory is usually based on a set of assumptions guaranteeing that, in large samples, the distribution of T_n is well-approximated, to the first order, by some standard distribution, usually the normal one. That is, $T_n \xrightarrow{d} Z$, $Z \sim \mathscr{N}(0,1)$, with '$\xrightarrow{d}$' denoting convergence in distribution. When $\hat{\theta}_n$ is an extremum estimator, provided the objective function can be expanded around a (pseudo-) true value θ_0, such assumptions are usually related to (i) existence of moments, (ii) stationarity and ergodicity, (iii) non-singularity of the Hessian (or, for transformations of the original estimator, full rank of the implied Jacobian), (iv) (pseudo-) true parameter in the interior of the parameter space; see, e.g., Newey and McFadden (1994). For extremum estimators based on instrumental variables [IV], (v) assumptions on the strength of the instruments are usually required. Examples of estimators requiring assumptions such as (i)–(v) to hold are, inter alia, (quasi) ML estimators, GMM estimators, nonlinear least squares estimators and minimum distance estimators. With 'valid specification' we mean that such assumptions are met.

The detection of invalid specifications is crucial in applications. A key challenge to proper statistical tests of valid specification is that they typically induce a 'pre-testing' bias in subsequent inferences. While in some cases the pre-testing bias may be associated with conservative inference under the null hypothesis (see, e.g., de Chaisemartin and D'Haultfoeuille, 2024), tests conditional on the non-rejection of correct specification may be severely oversized, even asymptotically.

In this paper, we show the novel result that the bootstrap delivers, as a by-product, consistent (diagnostic) tests of invalid specification which do not induce any pre-testing bias into subsequent inference procedures when the null of valid specification is not rejected. That is, post-test (conditional) inference is asymptotically *exact* when conditioning is upon the bootstrap tests not rejecting the null hypothesis of valid specification.

Our approach starts from the observation that – usually under mild additional requirements – a bootstrap analog of T_n, say $T_n^* := n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n$ or $T_n^* := n^{1/2}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n^*$, will also be asymptotically normal under valid specification; i.e., $T_n^* \xrightarrow{d^*}_p Z$, with '$\xrightarrow{d^*}_p$' denoting convergence in distribution conditionally on D_n; see Appendix A for a formal definition. In contrast, under invalid specification, T_n^* may no longer be asymptotically normal; rather, it usually has a non-Gaussian, random limiting distribution; see Cavaliere and Georgiev (2020) and the references therein. For instance, randomness of the limiting bootstrap measure may arise when (i) the score contributions have infinite variance (Athreya, 1987; Knight, 1989; Cavaliere, Georgiev and Taylor, 2016), (ii) when the data are non-stationary (Basawa, Mallik, McCormick and Taylor, 1991; Cavaliere, Nielsen and Rahbek, 2015), (iii) when the Hessian or the Jacobian are (near-) rank deficient (Datta, 1995; Angelini, Cavaliere and Fanelli, 2022; see also Han and McCloskey, 2019), (iv) when the (pseudo-) true value lies near or on the boundary of the parameter space (Andrews, 2000); (v) for IV estimators, when the instruments are weak or irrelevant (see Section 2).

This different asymptotic behavior of the bootstrap distribution of T_n^* can be exploited to detect invalid specification. To see why, consider the discrepancy between the cumulative distribution function [cdf] of T_n^* conditional on D_n, say $\hat{G}_n$, and the limiting standard Gaussian cdf Φ, measured as $\hat{d}_n :=\| \hat{G}_n - \Phi \|$, where $\| \cdot \|$ is a user-chosen norm or seminorm on the space of distribution functions. The distance $\hat{d}_n$ is expected to shrink to zero at a specific rate when asymptotic normality holds, while converging to a random limit when the assumptions fail to hold. Hence, a simple test of valid specification could assess whether the realized discrepancy $\hat{d}_n$ is large enough to reject the null.

As is standard in the bootstrap literature, $\hat{d}_n$ can be treated as a known function of the data D_n, since $\hat{G}_n$ can be approximated with any desired precision via the empirical distribution function [edf] $\hat{G}^*_{n,m}$ of m simulated realizations of T_n^*, with m arbitrarily large; that is, for any fixed n, as $m \to \infty$, with probability one, $\hat{G}^*_{n,m} \to \hat{G}_n$ uniformly, and hence $\hat{d}^*_{n,m} := \|\hat{G}^*_{n,m} - \Phi\| \to \hat{d}_n$ a.s. if $\| \cdot \|$ is continuous with respect to uniform convergence.

While it could be tempting to construct diagnostic tests based on $\hat{d}_n$ (or a properly normalized version, such as $\sqrt{n}\hat{d}_n$), such tests would suffer from at least two important drawbacks. First, their large-n asymptotic properties would depend on the particular bootstrap application of interest, and would often be very difficult to derive, as they require studying higher-order asymptotic expansions of $\hat{G}_n$. Second, since $\hat{d}_n$ is a function of the data, these tests might give rise to a pre-testing bias.

We show that these drawbacks disappear if inference, rather than being based on $\hat{d}_n$, is based on its approximation $\hat{d}^*_{n,m}$, where m and n diverge *jointly* under the requirement that m cannot be too large when n is finite. This asymptotic regime differs from the standard sequential bootstrap asymptotics, where $m \to \infty$ first (such that $\hat{G}^*_{n,m} - \hat{G}_n \approx 0$) followed by $n \to \infty$ (such that $\hat{G}_n - \Phi \approx 0$); see Andrews and Buchinsky (2000). In particular, we show that when m and n diverge jointly, with m diverging at a proper rate relative to n, a test with a known asymptotic distribution and based on the bootstrap statistic $\hat{d}^*_{n,m}$ can be designed to assess specification invalidity. Moreover, this approach is computationally straightforward, as it just requires to use the set of m bootstrap repetitions to compute $\hat{d}^*_{n,m}$. Put differently, it is equivalent to the application of distance-based normality tests to the set of m bootstrap repetitions, with different normality tests corresponding to different choices of the employed (semi)norm $\| \cdot \|$. Finally, it can be performed using the same critical values in a broad range of applications, and consistently detects deviations from asymptotic Gaussianity.

The role of the rate condition on m relative to n is to ensure that, under the null hypothesis of valid specification, the test decision becomes, in the limit, stochastically independent of the original data. This fact guarantees that a bootstrap test based on $\hat{d}^*_{n,m}$ does not induce pre-testing bias in large samples, in contrast to standard pre-tests for specification (in)validity (e.g., tests of finite variance, stationarity tests, pre-tests on the instrument strengths, and so forth). Instead, under a set of relevant alternatives the test is consistent as it exploits, in the limit, the information in the data alone. In summary, according to the validity of either the null or an alternative, the data choose asymptotically between acceptance based on an independent random device or rejection with probability approaching one. Interestingly, in a recent paper, de Chaisemartin and D'Haultfœuille

(2024) show that certain specification tests[1], when used as pre-tests, lead to conservative post-test inference under the null of valid specification. We complement their result by showing that the bootstrap delivers tests of specification validity leading to *exact* post-test inference as the sample size diverges.

This diagnostic potential of the bootstrap has not been developed in the extant literature. Beran (1997) was the first to suggest examining the bootstrap distribution to diagnose bootstrap failure. Davidson (2017) proposes simulation-based diagnostics in order to determine when a given bootstrap procedure works well or not in finite samples. Bårdsen and Fanelli (2015) provide prima facie evidence that non-Gaussian bootstrap distributions may be linked to weak identification in DSGE models; see also Angelini, Cavaliere and Fanelli (2022, 2024), and Zhan (2018). Within the problem of statistical reporting in a Bayesian communication framework, Andrews and Shapiro (2025) show that the (Bayesian) bootstrap distribution can serve as a surrogate posterior, and propose comparing it with the Gaussian approximation, using distance measures such as the signed Kolmogorov metric, to assess whether the conventional report adequately conveys uncertainty. A related approach is found in Wang (2025), who proposes detecting violations of asymptotic normality for GMM and extremum estimators by comparing quasi-Bayesian posterior distributions with the Gaussian benchmark. Our contribution extends this literature by demonstrating how distances between bootstrap and Gaussian distributions can be used to construct formal specification tests that avoid pre-testing bias. In terms of econometric theory, a further novelty lies in the asymptotic regime we adopt, where both the sample size n and the number of bootstrap repetitions m pass to infinity simultaneously. This setting is rarely considered in the literature; a notable exception is Andrews and Buchinsky (2000), who exploit this joint asymptotic regime to guide the choice of m in applied work.

To demonstrate the practical relevance and broad applicability of the bootstrap in the detection of specification invalidity, we discuss its use in the five scenarios (i)–(v) mentioned above. As regards case (i) of possible weak instruments, we also present an empirical illustration where we revisit, through the lens of bootstrap diagnostics, Känzig's (2021) empirical strategy for identifying the macroeconomic effects of a structural oil supply news shock.

STRUCTURE OF THE PAPER. The paper is organized as follows. In Section 2 we introduce a running example based on instrumental variable estimation. Section 3 contains our general results. Section 4 establishes the key result that the bootstrap procedure induces no pre-test bias in large samples. Additional results and extensions are reported in Section 5, while Section 6 illustrates four further applications. An empirical example is presented in Section 7, and Section 8 concludes. Notation and definitions used throughout the paper can be found in Appendix A. Proofs are collected in Appendices B and C.

[1] De Chaisemartin and D'Haultfœuille define 'valid specification' as the scenario in which the probability distribution generating D_n is such that a target estimator is consistent and asymptotically normal. They consider pre-tests that can detect when such an estimator becomes inconsistent. This definition differs from the one employed here, where 'valid specification' means that the underlying conditions ensuring the applicability of standard asymptotic inference are satisfied in the estimated model.

2 AN EXAMPLE BASED ON INSTRUMENTAL VARIABLES

Consider the following linear IV regression with one endogenous regressor:

$$y_i = \beta x_i + u_i,\ x_i = \pi' z_i + v_i$$

where the $k \times 1$ vector of instruments z_i is non-stochastic and the errors $(u_i, v_i)'$ are i.i.d. with mean zero and variance-covariance matrix Σ; to simplify, the diagonal elements of Σ are set to $\sigma_u^2 = \sigma_v^2 = 1$ and the off-diagonal elements to $\rho_{uv} \in (0,1)$. Without loss of generality, we also set $S_{zz} = I_k$, using the generic notation $S_{ab} := n^{-1}\sum_{i=1}^n a_i b_i'$. Given a sample of n observations, the 2SLS estimator of β is $\hat{\beta}_n := S_{xx.z}^{-1} S_{xy.z} = (S_{xz}S_{zy})/(S_{xz}S_{zx})$.

Consider further a (Gaussian) parametric bootstrap where the instruments are fixed in the bootstrap world. The bootstrap data are generated as

$$y_i^* = \hat{\beta}_n x_i^* + u_i^*,\ x_i^* = \hat{\pi}_n' z_i + v_i^*$$

where $\hat{\pi}_n := S_{zx}$ is the OLS estimator from the (first-stage) regression of x_i on z_i. For simplicity, assume that Σ is known, such that $(u_i^*, v_i^*)'$ can be taken as i.i.d. $\mathscr{N}(0, \Sigma)$ conditionally on the data.

Let first $\pi \neq 0$ be fixed (i.e., the instruments be 'strong'), a case that we label a 'valid specification'. Then, under standard assumptions for the central limit theorem (CLT),

$$T_n := \sqrt{n}\frac{\hat{\beta}_n - \beta}{\omega} = \frac{\sqrt{n}\pi' S_{zu}}{(\pi'\pi)^{1/2}} + o_p(1) \xrightarrow{d} \mathscr{N}(0,1)$$

where $\omega^2 := (\pi'\pi)^{-1}$. Moreover, as $\hat{\pi}_n'\hat{\pi}_n \to_p \pi'\pi \neq 0$, conditionally on the data,

$$T_n^* := \sqrt{n}\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\hat{\omega}_n} = \frac{\sqrt{n}\hat{\pi}_n' S_{zu^*}}{(\hat{\pi}_n'\hat{\pi}_n)^{1/2}} + O_p^*(n^{-1/2}) \sim \mathscr{N}(0,1) + o_p^*(1), \tag{2.1}$$

with $\hat{\omega}_n^2 := (\hat{\pi}_n'\hat{\pi}_n)^{-1}$. Therefore $T_n^* \xrightarrow{d^*}_p \mathscr{N}(0,1)$.

Instead, suppose next that the instruments are weak as in Staiger and Stock (1997), i.e., $\pi = \lambda n^{-1/2}$ for some vector λ; see also Bound, Jaeger and Baker (1995). This is an instance of an invalid specification. Assuming that

$$\sqrt{n}\begin{pmatrix} S_{zu} \\ S_{zv} \end{pmatrix} = n^{-1/2}\sum_{i=1}^n \begin{pmatrix} u_i \\ v_i \end{pmatrix} \otimes z_i \xrightarrow{d} \zeta := (\zeta_u', \zeta_v')'$$

with ζ a zero-mean Gaussian vector with variance-covariance matrix $\Sigma \otimes I_k$, we have

$$\hat{\beta}_n - \beta = \frac{S_{xz}S_{zu}}{S_{xz}S_{zx}} = \frac{(\sqrt{n}\pi + \sqrt{n}S_{zv})'\sqrt{n}S_{zu}}{(\sqrt{n}\pi + \sqrt{n}S_{zv})'(\sqrt{n}\pi + \sqrt{n}S_{zv})} \xrightarrow{d} \xi(\lambda, \zeta) := \frac{(\lambda + \zeta_v)'\zeta_u}{(\lambda + \zeta_v)'(\lambda + \zeta_v)}. \tag{2.2}$$

Similarly, the bootstrap analog of $\hat{\beta}_n - \beta$ can be written as

$$\hat{\beta}_n^* - \hat{\beta}_n = \frac{S_{x^*z}S_{zu^*}}{S_{x^*z}S_{zx^*}} = \frac{(\sqrt{n}\hat{\pi}_n + \sqrt{n}S_{zv^*})'\sqrt{n}S_{zu^*}}{(\sqrt{n}\hat{\pi}_n + \sqrt{n}S_{zv^*})'(\sqrt{n}\hat{\pi}_n + \sqrt{n}S_{zv^*})}$$

where, conditionally on the data, $n^{1/2}(S_{zu^*}', S_{zv^*}')' \sim \zeta^*$, with $\zeta^* := (\zeta_u^{*\prime}, \zeta_v^{*\prime})'$ distributed as ζ. Since, jointly with (2.2), $\sqrt{n}\hat{\pi}_n \to_d \ell := \lambda + \zeta_v$, we find using arguments as in Cavaliere

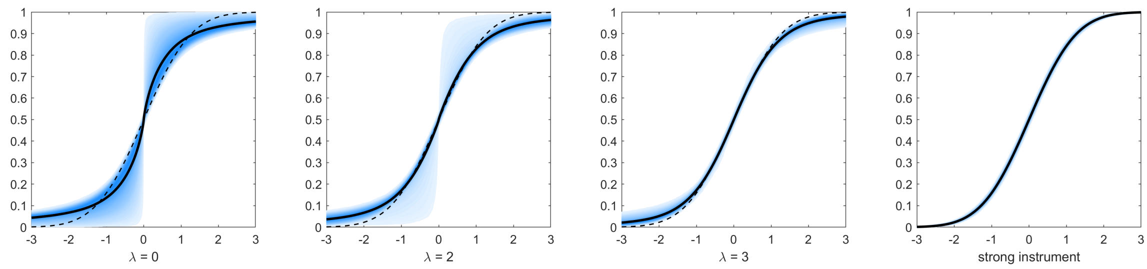

Figure 1: Fan chart of $M = 1{,}000$ i.i.d. realizations of the (conditional) cdf $\hat{G}_n$ of T_n^* under weak and strong instruments ($n = 1{,}000$).

and Georgiev (2020, proof of Theorem 4.1) that $\hat{\beta}_n^* - \hat{\beta}_n \xrightarrow{d^*}_w \xi(\ell, \zeta^*)|\ell$ and that

$$T_n^* := \sqrt{n}\frac{\hat{\beta}_n^* - \hat{\beta}_n}{\hat{\omega}_n} = \frac{\hat{\beta}_n^* - \hat{\beta}_n}{\hat{\omega}_n/\sqrt{n}} \xrightarrow{d^*}_w \frac{\xi(\ell, \zeta^*)}{\sqrt{\ell'\ell}}|\ell; \tag{2.3}$$

see Appendix A for the definition of $\xrightarrow{d^*}_w$. Hence, the bootstrap distribution has a random limit and, as $n \to \infty$, the bootstrap cdf satisfies $\hat{G}_n(x) \to_w \mathscr{G}(x) := \mathbb{P}(\xi(\ell, \zeta^*)/\sqrt{\ell'\ell} \le x|\ell)$, $x \in \mathbb{R}$, on $\mathscr{D}_{\mathbb{R}}$, where the limit differs from the Gaussian cdf.

Remark 2.1 *The limiting randomness of the bootstrap distribution $\hat{G}_n$ under weak instruments is illustrated in Figure 1, where we report the fan chart of $M = 1,000$ i.i.d. realizations of $\hat{G}_n$ for $k = 1$, $n = 1,000$ and using a standard Gaussian parametric bootstrap with $z_i = 1$ and $\rho_{uv} = 0.9$. The randomness and non-normality in the bootstrap cdf $\hat{G}_n$, quite evident for small values of λ, ameliorates as λ increases and (up to sampling error due to finite n) disappears in the strong-instrument case where $\pi \neq 0$ is fixed.* □

Remark 2.2 *Notice that under weak instruments the bootstrap does not replicate the asymptotic distribution of the original statistic, given by $\xi(\lambda, \zeta)$ of (2.2), essentially because in the limiting bootstrap experiment λ is replaced by the random vector ℓ, where $\ell \neq \lambda$ with probability one. This result explains the inconsistency of the bootstrap in the weak IV framework, as previously documented in the literature (see, e.g., Davidson and MacKinnon, 2010), in terms of randomness of the limit bootstrap measure.* □

Remark 2.3 *The results in this section do not substantially change if z_i is random and resampled along with ε_i^* and u_i^*, and/or if $\{\varepsilon_i^*, u_i^*\}_{i=1}^n$ are i.i.d. draws from the (centered) residuals $\{\hat{\varepsilon}_i, \hat{u}_i\}_{i=1}^n$ as in standard non-parametric bootstrap designs (in the latter case, the random asymptotic distribution in (2.3) is more involved).* □

3 Testing specification validity

Assume that $\theta \in \mathbb{R}$ and consider a bootstrap statistic T_n^* that is asymptotically $\mathscr{N}(0,1)$ under valid specification. T_n^* can be, e.g., of the form $T_n^* := (\hat{\theta}_n^* - \hat{\theta}_n)/\operatorname{se}(\hat{\theta}_n^*)$ or $T_n^* := (\hat{\theta}_n^* - \hat{\theta}_n)/\operatorname{se}(\hat{\theta}_n)$ (non-studentized statistics of the form $T_n^* := \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)$ can also be considered). As we shall see in this section, a test of valid specification, which does not induce pre-testing bias, can be obtained by assessing whether the bootstrap replicates the

Gaussian asymptotic distribution. As an initial step, we quantify the discrepancy between the bootstrap cdf and the asymptotic Gaussian cdf.

3.1 PRELIMINARIES

Let $\hat{G}_n(\cdot) := \mathbb{P}^*(T_n^* \leq \cdot)$ be the cdf of T_n^*, conditional on the data D_n. We consider measuring the discrepancy between $\hat{G}_n$ and the $\mathscr{N}(0,1)$ cdf Φ by $\hat{d}_n :=\| \hat{G}_n - \Phi \|$ where $\|\cdot\| : \mathscr{D}_{\mathbb{R}} \to [0,\infty)$ is a user-chosen norm or seminorm; alternative discrepancy measures are discussed in Section 5.2. For instance, setting $\|\cdot\| = \|\cdot\|_\infty$, i.e., the sup norm on $\mathscr{D}_{\mathbb{R}}$, delivers the well known Kolmogorov-Smirnov [KS] distance

$$\hat{d}_n = \hat{d}_n^{\mathsf{KS}} := \|\hat{G}_n - \Phi\|_\infty = \sup_{u\in\mathbb{R}} |\hat{G}_n(u) - \Phi(u)|. \tag{3.1}$$

Under the bootstrap consistency hypothesis $T_n^* \xrightarrow{d^*}_p \mathscr{N}(0,1)$, by Polya's theorem,

$$\hat{d}_n = \|\hat{G}_n - \Phi\|_\infty \xrightarrow{p} 0, \tag{3.2}$$

and similarly, for any (semi)norm $\|\cdot\|$ that is continuous on $\mathscr{D}_{\mathbb{R}}$ with the uniform metric; examples are the signed KS (Andrews and Shapiro, 2025) and the Cramer-von Mises norms. However, should bootstrap consistency for the Gaussian limit fail, also (3.2) will fail to hold. In particular, for the IV example of Section 2 as well as for all the examples in Section 6, it holds that $\hat{G}_n \xrightarrow{w} \mathscr{G}$ in $\mathscr{D}_{\mathbb{R}}$, where $\mathscr{G}$ is a (random) cdf such that $\mathscr{G} \neq \Phi$ (a.s.). By the continuous mapping theorem [CMT], provided the transformation $\|\cdot\|$ is continuous on $\mathscr{D}_{\mathbb{R}}$,[2] $\hat{d}_n$ itself has a (possibly) random limit:

$$\hat{d}_n := \|\hat{G}_n - \Phi\| \xrightarrow{d} \mathscr{Y} := \|\mathscr{G} - \Phi\| \tag{3.3}$$

where $\mathscr{Y} > 0$ a.s. if $\|\cdot\|$ is a norm.[3] The different asymptotic behaviors in (3.2) and (3.3) will be exploited to develop bootstrap tests for valid specification.

As is standard, the cdf $\hat{G}_n$, as well as the discrepancy $\hat{d}_n$, can be approximated using a sample $T_{n:i}^*$, $i = 1, ..., m$, of conditionally independent copies of T_n^* (obtained by simulation):

$$\hat{G}_{n,m}^*(\cdot) := \frac{1}{m} \sum\nolimits_{i=1}^m \mathbb{I}_{\{T_{n:i}^* \leq \cdot\}}, \qquad \hat{d}_{n,m}^* := \|\hat{G}_{n,m}^* - \Phi\|, \tag{3.4}$$

with m user-chosen. Both $\hat{d}_{n,m}^*$ and $\hat{G}_{n,m}^*$, the latter being usually employed in applications of the bootstrap to compute p-values and confidence sets, are key to assess specification validity without inducing pre-testing bias.

3.2 ASYMPTOTIC REGIMES

Consider $\hat{G}_{n,m}^*$ and $\hat{d}_{n,m}^*$ of (3.4). By the Glivenko-Cantelli theorem, for any n and as $m \to \infty$, $\|\hat{G}_{n,m}^* - \hat{G}_n\|_\infty \xrightarrow{a.s.^*} 0$ (a.s.) Then, for any n, also $\hat{d}_{n,m}^* \xrightarrow{a.s.^*} \hat{d}_n$ (a.s.), provided that $\|\cdot\|$ is continuous on $\mathscr{D}_{\mathbb{R}}$ with the Skorokhod J_1 or the sup metric. Hence, for practical purposes $\hat{G}_n$ and $\hat{d}_n$ can be treated as known and asymptotic inference based on

[2] If $\mathscr{G}$ is sample-path continuous, continuity of $\|\cdot\|$ on $\mathscr{D}(\mathbb{R})$ equipped with the uniform metric suffices.

[3] For seminorms, e.g., $\hat{d}_n(A) := \sup_{u\in A\subset\mathbb{R}} |\hat{G}_n(u) - \Phi(u)|$, also $\mathbb{P}(\mathscr{Y} = 0) > 0$ may hold; for an example see the parameter-on-the-boundary case of Section 6.2, where the positivity of $\mathscr{Y}$ for $\hat{d}_n(A)$ depends on the choice of the set A.

transformations of $\hat{G}_n - \Phi$, like $\hat{d}_n$, is in principle feasible.

As anticipated in Section 1, however, it turns out that inference based on such transformations is unattractive in practice, as (i) their null asymptotic distribution as $n \to \infty$ is in general not only very hard to obtain, but also application-specific (see Section 6.2 for an example involving $\hat{d}_n$); and (ii) a problem of post-diagnostic test bias emerges.

A closer look shows that these issues arise under an implicit sequential asymptotic regime – which we label as $(m, n \to \infty)_{\text{seq}}$ using the notation in Phillips and Moon (1999) – where first $m \to \infty$ in order for $\hat{d}^*_{n,m}$ to collapse to $\hat{d}_n$, and then $n \to \infty$. Standard bootstrap asympotic theory, which would discuss $\hat{d}_n$ as $n \to \infty$ rather than the actually computed $\hat{d}^*_{n,m}$, can be interpreted as employing $(m, n \to \infty)_{\text{seq}}$ asymptotics.

We now ask whether a different asymptotic regime can be chosen such that (i) the asymptotic distributions of test statistics are invariant across a wide range of applications under the null of valid specification, (ii) no post-diagnostic bias is present under the null, and (iii) the tests are consistent against a relevant class of alternatives. Two candidate asymptotic regimes are as follows.

First, consider the sequential regime $(n, m \to \infty)_{\text{seq}}$, where $n \to \infty$ followed by $m \to \infty$. In the case of the sup norm, $\hat{d}^*_{n,m} = \|\hat{G}^*_{n,m} - \Phi\|_\infty$ can be written as $\hat{d}^*_{n,m} = \phi_m(T^*_{n:1}, ..., T^*_{n:m})$ for a continuous $\phi_m : \mathbb{R}^m \to \mathbb{R}$. Because under the null of valid specification $T^*_n \xrightarrow{d^*}_p Z \sim \mathscr{N}(0,1)$, it holds that

$$\hat{d}^*_{n,m} = \phi_m(T^*_{n:1}, ..., T^*_{n:m}) \xrightarrow{d^*}_p d^*_m := \phi_m(Z_1, ..., Z_m)$$

as $n \to \infty$, with the Z_i's being independent $\mathscr{N}(0,1)$, $i = 1, ..., m$. Now, let $m \to \infty$ after $n \to \infty$. Under the null, this yields by standard empirical process theory that, with W denoting a standard Brownian bridge, $\sqrt{m} d^*_m \xrightarrow{d} \|W(\Phi)\|_\infty$, i.e., the Kolmogorov distribution. Therefore, if $(n, m \to \infty)_{\text{seq}}$,

$$\mathscr{T}^*_{n,m} := \sqrt{m}\hat{d}^*_{n,m} = \sqrt{m}\|\hat{G}^*_{n,m} - \Phi\|_\infty \xrightarrow{d^*}_p \|W(\Phi)\|_\infty.$$

In contrast, under alternatives such that $\hat{G}_n \xrightarrow{w} \mathscr{G} \neq \Phi$ a.s., it holds that $\mathscr{T}^*_{n,m} \xrightarrow{p^*}_p \infty$ as $(n, m \to \infty)_{\text{seq}}$, yielding consistent tests.

Although the sequential 'm after n' asymptotic regime leads to tractable derivations, it provides little justification for conducting inference based on $\|W(\Phi)\|_\infty$. Indeed, in practice, one has more control on m (number of bootstrap repetitions) rather than n (sample length). Hence, we also consider an asymptotic regime where m and n diverge jointly, denoted as $(n, m \to \infty)$. Under the null and under an additional rate condition relating m to n, this regime allows to achieve asymptotic distributions that are invariant across applications (e.g., the $\|W(\Phi)\|_\infty$ limit seen previously for the sup norm), whereas under relevant alternatives the resulting tests are consistent. Focusing on the statistic $\mathscr{T}^*_{n,m} := \sqrt{m}\hat{d}^*_{n,m}$, the rate condition makes sure that $\hat{G}^*_{n,m} - \hat{G}_n$ is dominant in the derivation of $\mathscr{T}^*_{n,m}$'s asymptotic distribution under the null, whereas under alternatives of interest, $\hat{G}_n - \Phi$ is dominant in the limit.

Notice that both the $(n, m \to \infty)_{\text{seq}}$ and $(n, m \to \infty)$ regimes are intended for justifying inference employing $\hat{G}^*_{n,m}$ rather than true $\hat{G}_n$. Although this choice implies a loss of information for every fixed n, as $\hat{G}^*_{n,m}$ is used before it has converged to $\hat{G}_n$, it is the bootstrap randomness in $\hat{G}^*_{n,m} - \hat{G}_n$ that can make the asymptotic distribution of $\hat{d}^*_{n,m}$ invariant across applications and can eliminate the post-diagnostic bias. Indeed, for both regimes no post-diagnostic test bias is present asymptotically, as the proof of Theorem 4.1 below for the joint regime goes through also for the sequential 'm after n' regime.

3.3 THE DIAGNOSTIC TEST

Redefine $\mathscr{T}^*_{n,m}$ introduced above using a generic (semi)norm, i.e., $\mathscr{T}^*_{n,m} := m^{1/2}\hat{d}^*_{n,m}$, $\hat{d}^*_{n,m} := \|\hat{G}^*_{n,m} - \Phi\|$ with $\hat{G}^*_{n,m}(\cdot) := m^{-1}\sum_{i=1}^{m}\mathbb{I}_{\{T^*_{n:i}\leq\cdot\}}$ and the $T^*_{n:i}$'s being i.i.d. copies of T^*_n conditionally on the data D_n. We note the following.

First, $\mathscr{T}^*_{n,m}$ depends on the data D_n as well as on m auxiliary variates, say W^*_n, used to generate the m bootstrap draws $T^*_{n:i}$, $i = 1, \ldots, m$. For instance W^*_n, which is defined jointly with D_n on a possibly extended probability space, could be thought of as a vector of m i.i.d. $\mathscr{U}_{[0,1]}$ r.v.s, independent of D_n, such that $T^*_{n:i} = \hat{G}_n^{-1}(W^*_{n,i})$, $i = 1, \ldots, m$, with $\hat{G}_n^{-1}$ the generalized inverse of $\hat{G}_n$.

Second, $\mathscr{T}^*_{n,m}$ can be written as

$$\mathscr{T}^*_{n,m} = \mathcal{Z}^*_{n,m} + a^*_{n,m},\ \mathcal{Z}^*_{n,m} := m^{1/2}\|\hat{G}^*_{n,m} - \hat{G}_n\|, \tag{3.5}$$

with $a^*_{n,m}$ implicitly defined. Here, $\mathcal{Z}^*_{n,m}$ is a rescaled measure of the distance between the estimator $\hat{G}^*_{n,m}$ and $\hat{G}_n$, while $a^*_{n,m}$ is related to the distance between $\hat{G}_n$ and the Gaussian cdf Φ. In particular, for any n, $|a^*_{n,m}| \leq \sqrt{m}\|\hat{G}_n - \Phi\|$ by Lemma B.1 in Appendix B.

We now provide conditions such that, under the valid specification hypothesis, the asymptotic distribution of $\mathscr{T}^*_{n,m}$ can be derived and hence used to perform a proper statistical test. In particular, we consider the following assumption.

ASSUMPTION 1 *The (semi)norm $\|\cdot\|$ is continuous on $\mathscr{D}_{\mathbb{R}}$. Moreover, $\|\hat{G}_n - \Phi\| = O_p(n^{-\alpha})$ for some $\alpha > 0$.*

Assumption 1 is a condition on the rate of convergence of the bootstrap cdf to the Gaussian cdf under valid specification. It is satisfied with $\alpha = 1/2$ if $\|\cdot\| \leq C\|\cdot\|_\infty$ for some finite constant C (as is the case of, e.g., the signed KS norm or the Cramer-von Mises norm), and the bootstrap statistic has a one-term Edgeworth expansion of the form $\hat{G}_n(x) = \Phi(x) + \hat{q}_n(x)n^{-1/2} + o_p(n^{-1/2})$ uniformly in x, as is usually the case; see, e.g., Hall (1992). For some symmetric statistics, Assumption 1 can hold with $\alpha = 1$. For Gaussian parametric bootstraps, Assumption 1 may be even satisfied with arbitrary $\alpha > 0$; see, e.g., the case in Section 6.2.

The following result holds under Assumption 1.

THEOREM 3.1 *Let $T^*_n \xrightarrow{d^*}_p \mathscr{N}(0,1)$. Under Assumption 1, if $(n, m \to \infty)$ and $m/n^{2\alpha} \to 0$,*

(i) $\mathscr{T}^*_{n,m} = \mathcal{Z}^*_{n,m} + o_p(1) \xrightarrow{d^*}_p \mathscr{K} := \|W(\Phi)\|$, where W is standard Brownian bridge.

(ii) $p^*_{n,m} := 1 - H(\mathscr{T}^*_{n,m}) \xrightarrow{w^*}_p \mathscr{U}_{[0,1]}$ if the cdf H of $\mathscr{K}$ is continuous.

Theorem 3.1 suggests that a simple diagnostic procedure can be obtained by comparing $\mathscr{T}^*_{n,m}$ with critical values from the distribution of $\mathscr{K}$, or through the associated asymptotic

p-value $p^*_{n,m}$. Provided both m and n grow to infinity at a proper relative rate, Theorem 3.1 guarantees that the procedure is asymptotically correctly sized.

REMARK 3.1 *The logic of the proof can be easily followed in the case of the signed pointwise discrepancy* $\hat{d}_n(x) := \hat{G}_n(x) - \Phi(x)$ *for some fixed* $x \in \mathbb{R}$*. In this case,* $\mathscr{Z}^*_{n,m}$ *and* $a^*_{n,m}$ *of (3.5) are given by* $\mathscr{Z}^*_{n,m} = m^{1/2}(\hat{G}^*_{n,m}(x) - \hat{G}_n(x))$ *and* $a^*_{n,m} = m^{1/2}(\hat{G}_n(x) - \Phi(x))$*. If* $\hat{d}_n(x)$ *satisfies Assumption 1 for some* $\alpha > 0$*, it holds that* $a^*_{n,m} \to_p 0$ *as* $(n, m \to \infty)$ *with* $m/n^{2\alpha} \to 0$*. Moreover, with* $\xi^*_i := \mathbb{I}_{\{T^*_{n:i} \le x\}} - \mathbb{E}^*[\mathbb{I}_{\{T^*_{n:i} \le x\}}]$*, such that* $\mathscr{Z}^*_{n,m} = m^{-1/2} \sum_{i=1}^m \xi^*_i$*, by a standard Berry-Esseen bound we have that* $\tilde{\mathscr{Z}}^*_{n,m} := \mathbb{E}^*[\xi^{*2}_i]^{-1/2} \mathscr{Z}^*_{n,m}$ *satisfies, for some* $C < \infty$*,*

$$\| \mathbb{P}^*(\tilde{\mathscr{Z}}^*_{n,m} \le \cdot) - \Phi(\cdot) \|_\infty \le C m^{-1/2} \mathbb{E}^*[|\xi^*_i|^3] \le C m^{-1/2}$$

whenever $\mathbb{E}^*[\xi^{*2}_i] = \hat{G}_n(x)(1 - \hat{G}_n(x)) \neq 0$*, which occurs with probability approaching one as* $n \to \infty$ *since* $\mathbb{E}^*[\xi^{*2}_i] \to_p v^2(x) := \Phi(x)(1 - \Phi(x))$ *under the hypothesis that* $T^*_n \xrightarrow{d^*}_p \mathscr{N}(0,1)$ *as* $n \to \infty$*. We conclude that* $\mathscr{Z}^*_{n,m} = v(x)\tilde{\mathscr{Z}}^*_{n,m} + o^*_p(1) \xrightarrow{d^*}_p \mathscr{N}(0, v(x)) \sim W(\Phi(x))$ *as* $(n, m \to \infty)$*.* □

REMARK 3.2 *For commonly used norms, the distribution of* $\mathscr{K}$ *is well known; for example, with* $\|\cdot\| = \|\cdot\|_\infty$*,* $\mathscr{K}$ *follows the Kolmogorov distribution which has a continuous cdf. In general, critical values (as well as p-values) can be determined by Monte Carlo simulation with arbitrary accuracy.* □

The test based on $\mathscr{T}^*_{n,m}$ also has non-trivial asymptotic power against bootstrap inconsistency for the standard Gaussian distribution. This is shown next.

THEOREM 3.2 *Suppose that* $\hat{G}_n \xrightarrow{w} \mathscr{G}$ *in* $\mathscr{D}_\mathbb{R}$ *as* $n \to \infty$*. Suppose further that* $\|\cdot\|$ *is continuous on* $\mathscr{D}_\mathbb{R}$ *and* $\|\mathscr{G} - \Phi\| > 0$ *a.s. Then, for any* $c \in (0, \infty)$*, it holds that* $\mathbb{P}^*(\mathscr{T}^*_{n,m} \ge c) \xrightarrow{p} 1$ *as* $(n, m \to \infty)$*.*

An inspection of the proof of Theorem 3.2 reveals that, as expected, for large m the power of the test is determined by the realized value of $\hat{d}_n := \|\hat{G}_n - \Phi\|$, which is asymptotically distributed as the r.v. $\mathscr{Y} := \|\mathscr{G} - \Phi\| > 0$ a.s. Such realization depends on the original data D_n only, and not on the m bootstrap repetitions used to generate the bootstrap statistic. Larger outcomes of $\hat{d}_n$ correspond – ceteris paribus – to larger power.

3.4 An example based on instrumental variables (cont'd)

When $\pi \neq 0$ is fixed, such that the instruments are strong, for the parametric bootstrap described in Section 2 it holds, without additional assumptions, that $\|\hat{G}_n - \Phi\|_\infty = O_p(n^{-\alpha})$ for any $\alpha \in (0, \frac{1}{2})$; this is shown in Appendix C by using the machinery of parametric tail estimates. Hence, Assumption 1 is verified with $\alpha \in (0, \frac{1}{2})$ for the sup norm and its dominated norms, and Theorem 3.1 applies to them. Using uniform Edgeworth expansions, also $\|\hat{G}_n - \Phi\|_\infty = O_p(n^{-1/2})$ has been shown to hold for the non-parametric i.i.d. bootstrap (where u^*_i and v^*_i are resampled from the residuals $\hat{u}_i := y_i - \hat{\beta}_n x_i$ and $\hat{v}_i := x_i - \hat{\pi}'_n z_i$), under mild regularity conditions on (u_i, v_i) (essentially, existence of higher-order moments); see Moreira, Porter and Suarez (2009, Theorem 3).

(A)	KS								AD							
$n\backslash\zeta$	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00
100	4.9	5.2	5.4	5.4	5.7	6.9	8.0	10.5	5.3	5.6	5.7	5.6	5.7	6.8	7.7	9.6
200	5.2	5.4	5.6	5.6	5.6	7.0	8.7	10.4	5.6	5.7	5.7	5.6	5.8	6.7	7.5	9.8
400	5.0	5.0	5.3	5.6	6.0	6.3	7.9	10.6	5.1	5.1	5.0	5.4	5.8	6.2	7.5	9.3
800	4.9	4.9	5.4	5.6	5.7	6.6	7.8	10.2	5.1	4.7	5.2	5.4	5.7	6.2	7.4	9.7
(B)	KS								AD							
$n\backslash\zeta$	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00
100	29.6	33.7	39.2	44.6	51.8	63.7	74.3	83.3	30.7	35.6	41.9	49.1	56.5	69.1	80.1	89.2
200	37.4	43.8	50.2	56.7	62.8	76.3	86.4	93.4	40.0	46.8	55.1	62.3	68.8	82.6	91.5	97.1
400	46.1	52.8	60.9	69.0	75.9	87.5	95.2	98.4	51.0	58.1	66.7	75.3	82.1	93.0	97.9	99.6
800	54.9	63.2	71.6	79.3	85.4	94.7	98.3	99.6	60.4	69.0	77.9	85.4	91.0	97.7	99.5	100.0

Table 1: IV Regression – Empirical rejection frequencies of the bootstrap diagnostic tests based on the Kolmogorov-Smirnov (KS) and Anderson-Darling (AD) norms with $m = \lfloor n^\zeta \rfloor$, $\zeta \in [0.5, 1.0]$. Panel (A): strong instrument; Panel (B): weak instrument.

Under weak instruments, we found previously that the bootstrap cdf of T_n^* satisfies $\hat{G}_n(x) \to_w \mathscr{G}(x) := \mathbb{P}(\xi(\ell, \zeta^*)/\sqrt{\ell'\ell} \le x \mid \ell)$ on $\mathscr{D}_{\mathbb{R}}$, which is non-Gaussian with probability one. Thus, Theorem 3.2 applies in this case.

We conclude by reporting in Table 1 the (percentage) empirical rejection probabilities (ERPs) of the bootstrap test based on the sup norm (KS); for comparison, we also consider the test based on the Anderson-Darling norm (AD). The data generating process is as in Remark 2.1 with $z_t \sim$i.i.d.$\mathscr{N}(0,1)$ and the bootstrap is non-parametric (see Remark 2.3) with z_t fixed in the bootstrap world; a constant is included in estimation. The upper panel corresponds to the strong instrument case ($\pi = 1$), while the lower panel considers the weak instrument case with $\pi = \lambda n^{-1/2}$ and $\lambda = 2$. To assess how well the asymptotic theory ($n, m \to \infty$, $m/n \to 0$) approximates finite-sample behavior, we consider different choices of m, namely $m = \lfloor n^\zeta \rfloor$ with $\zeta \in \{0.50, 0.55, \dots, 0.70, 0.80, 0.90, 1.00\}$. ERPs are computed using $10,000$ Monte Carlo replications; the nominal level is 5%.

In the strong instrument case (upper panel), the ERPs remain relatively close to the nominal level, particularly for lower values of m relative to n. They increase as m grows; for KS they never exceed 11%, while for AD they can rise up to about 10%. Consistently with the theoretical expectation, under the weak instrument scenario (lower panel), the ERPs increase with m, n.

4 POST-DIAGNOSTICS INFERENCE

The bootstrap approach developed in the previous section can be used in diagnostic pre-testing of specification validity, with standard inference carried out when the diagnostic procedure does not reject. An important question is whether post-diagnostic inferences are biased by the outcome of the pre-test, in the sense that conditionally on a correct non-rejection by the pre-test, the rejection probabilities of post-diagnostic tests are affected even asymptotically. The answer is no.

The key for this result is that, under valid specification, the bootstrap statistic $\mathscr{T}_{n,m}^*$ becomes independent of the original data D_n under the joint asymptotic regime $(n, m \to \infty)$

and the rate condition of Theorem 3.1. That is, in the limit $\mathscr{T}^*_{n,m}$ only depends on the bootstrap variates used to generate $\hat{G}^*_{n,m}$ and no longer on the data D_n, thus eliminating any post-diagnostic bias.

To see why, it suffices to note that the existence of a non-random limit for the conditional law of a bootstrap statistic given the data is an asymptotic independence property. It is this very property that the conditional law of the bootstrap diagnostic statistic $\mathscr{T}^*_{n,m}$ enjoys under the conditions of Theorem 3.1. The meaning of the implied asymptotic independence is clarified in the next theorem, where $\mathscr{T}^*_{n,m}$ can stand for any bootstrap statistic and $(n, m \to \infty, R)$ means that $(n, m \to \infty)$ under a rate condition R (such as the one given in Assumption 1).

THEOREM 4.1 *Let the conditional distribution of a bootstrap statistic $\mathscr{T}^*_{n,m}$ given the data D_n converge in probability to a nonrandom distribution as $(n, m \to \infty, R)$. Then, as $(n, m \to \infty, R)$:*

(a) for measurable real functions f and continuous bounded real functions g with matching domains,

$$\sup_{\|f\|_\infty \leq 1} \left|\mathbb{E}[f(D_n)g(\mathscr{T}^*_{n,m})] - \mathbb{E}[f(D_n)]\mathbb{E}[g(\mathscr{T}^*_{n,m})]\right| \to 0;$$

(b) for non-negligible continuity sets B of $\mathscr{T}^*_{n,m}$'s limit distribution,

$$\sup_{A\in\sigma(D_n)} |\mathbb{P}(A|\mathscr{T}^*_{n,m} \in B) - \mathbb{P}(A)| \to 0,$$

where $\sigma(D_n)$ is the σ-algebra generated by the data D_n.

A non-negligible continuity set is a set with positive probability whose boundary has zero probability. For instance, if the limit distribution of $\mathscr{T}^*_{n,m}$ has a continuous cdf H, then $B = [0, t]$ is a non-negligible continuity set whenever $t > 0$ is such that $H(t) > 0$. This gives rise to the following corollary relevant for testing.

COROLLARY 4.1 *Consider any statistic $\hat{\rho}_n \in \mathbb{R}$, measurable with respect to the data D_n and with asymptotic cdf F_ρ as $n \to \infty$. Let the conditional distribution of $\mathscr{T}^*_{n,m}$ given the data converge in probability to a nonrandom distribution with a continuous cdf H as $(n, m \to \infty, R)$. Then, for every continuity point s of F_ρ any every t with $H(t) > 0$, it holds that, as $(n, m \to \infty, R)$,*

$$\sup_{s\in\mathbb{R}} |\mathbb{P}(\hat{\rho}_n \leq s|\mathscr{T}^*_{n,m} \leq t) - F_\rho(s)| \to 0.$$

Under the conditions of Theorem 3.1, including $T^*_n \xrightarrow{d^*}_p \mathscr{N}(0, 1)$, the diagnostic statistic $\mathscr{T}^*_{n,m}$ satisfies the assumptions of Corollary 4.1 with $H = \Phi$ and R being the rate condition in Assumption 1. Therefore, in large samples the finite-sample quantiles of $\hat{\rho}_n$ conditional on $\mathscr{T}^*_{n,m}$ being below (or above) a given critical value t approximate well the unconditional quantiles of $\hat{\rho}_n$'s asymptotic distribution. Hence, should the diagnostics based on $\mathscr{T}^*_{n,m}$ correctly fail to reject specification validity, inference based on $\hat{\rho}_n$ is free of pre-testing bias as $n \to \infty$.

		KS								AD							
$n\backslash\zeta$	unc.	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00	0.50	0.55	0.60	0.65	0.70	0.80	0.90	1.00
100	*4.9*	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.9	4.8	4.8	4.9
200	*5.3*	5.3	5.3	5.4	5.3	5.4	5.3	5.3	5.4	5.2	5.2	5.3	5.3	5.4	5.2	5.2	5.3
400	*4.7*	4.7	4.7	4.8	4.8	4.8	4.7	4.7	4.7	4.7	4.7	4.8	4.8	4.8	4.7	4.7	4.7
800	*5.1*	5.1	5.1	5.1	5.1	5.1	5.1	5.2	5.1	5.0	5.1	5.1	5.1	5.1	5.1	5.1	5.1

Table 2: IV Regression – Empirical rejection probabilities of the $t_{\mathsf{IV},n}$ test. Results are unconditional (in italics) and conditional on the bootstrap tests not rejecting specification validity. Bootstrap tests computed with $m = \lfloor n^{\zeta} \rfloor$, $\zeta \in [0.5, 1.0]$.

4.1 An example based on instrumental variables (cont'd)

We now briefly show Theorem 4.1 and Corollary 4.1 in action by considering the 2SLS t-test for the null $\beta = \beta_0$, denoted $t_{\mathsf{IV},n}$, in the setting of the IV example of Sections 2 and 3.4, assuming that z_t is strong (i.e., a valid specification), so that $t_{\mathsf{IV},n} \to_d \mathscr{N}(0,1)$. Indeed, an extensive literature on IV regression documents substantial bias in pre-tests of instrument strength (see, e.g., Andrews, Stock, and Sun, 2019), making it natural to ask whether the bootstrap diagnostic test introduces any such bias.

We compare two different scenarios. In the first one, the researcher runs the t-test without pre-testing valid specification. In the second one, the researcher initially (pre-)tests for valid specification using the bootstrap diagnostic test $\mathscr{T}^*_{n,m}$, and then runs the t-test only if the bootstrap test does not reject valid specification. According to Corollary 4.1 applied with $\hat{\rho}_n = t_{\mathsf{IV},n}$ and $F_\rho = \Phi$, the *unconditional* rejection probabilities of the t-test (computed without pre-testing valid specification) and the *conditional* rejection probabilities (computed conditionally on the bootstrap diagnostic test failing to reject) should coincide as $(n, m \to \infty)$ at proper relative rates, see Section 3.4. These probabilities are estimated in Table 2 using the same Monte Carlo design as in Section 3.4. The unconditional and the conditional ERPs are virtually indistinguishable, thus supporting the results in Corollary 4.1.

5 Extensions

5.1 Diagnostics reporting

An intrinsic feature of any diagnostics based on $\mathscr{T}^*_{n,m}$ is that the associated p-value, say $p^*_{n,m}$, is not D_n-measurable, as it depends also on the realization of the auxiliary bootstrap variates W^*_n used to generate $\mathscr{T}^*_{n,m}$; see Section 3.3. This differs from ordinary bootstrap inference, where the bootstrap p-values are usually regarded as measurable with respect to the data D_n, as is the case where $(m, n \to \infty)_{\text{seq}}$. Moreover, under the conditions of Theorem 3.1, the dependence of $p^*_{n,m}$ on the bootstrap variates does not vanish as $(n, m \to \infty)$, in the sense that $p^{*(1)}_{n,m} - p^{*(2)}_{n,m}$ need not go to zero for two conditionally independent copies $p^{*(1)}_{n,m}$ and $p^{*(2)}_{n,m}$ of $p^*_{n,m}$.

As a consequence, if the bootstrap sample is not given ex ante but, rather, is generated by the practitioner, then for any given data D_n different researchers may end up obtaining different p-values $p^*_{n,m}$. This may create issues in terms of reproducibility, as it is unclear which value of $p^*_{n,m}$ should be used for decision-making and reporting.[4]

A viable way to mitigate this issue is to use the convergence fact $p^*_{n,m} \xrightarrow{d^*}_p \mathscr{U}_{[0,1]}$ under specification validity, see Theorem 3.1. This convergence and Theorem 3.2 have the following corollary in terms of $\hat{\pi}_{n,m}(\eta) := \mathbb{P}^*(p^*_{n,m} \le \eta)$, $\eta \in (0,1)$.

Corollary 5.1 *Under the assumptions of Theorem 3.1, $\hat{\pi}_{n,m}(\eta) \to_p \eta$ for any $\eta \in (0,1)$. In contrast, under the assumptions of Theorem 3.2, $\hat{\pi}_{n,m}(\eta) \to_p 1$.*

Hence, rather than reporting a single draw $p^*_{n,m}$, the diagnostic procedure could additionally include an informal check of whether $\hat{\pi}_{n,m}(\eta)$ substantially exceeds the user-chosen significance level η. In practice $\hat{\pi}_{n,m}(\eta)$ is not known but, as is standard, it can be approximated with any desired precision by drawing an arbitrarily large number K of i.i.d. (conditionally on the data) realizations of $p^*_{n,m}$, say $p^*_{n,m:k}$, $k = 1,\dots K$, and letting $\hat{\pi}^*_{n,m,K} := K^{-1}\sum_{k=1}^{K} \mathbb{I}_{\{p^*_{n,m:k}\le\eta\}}$.

In principle, the significance of $\hat{\pi}^*_{n,m,K}(\eta) - \eta$ could be assessed formally by means of a test, using the (pointwise) standard error $(\hat{\pi}^*_{n,m,K}(\eta)(1-\hat{\pi}^*_{n,m,K}(\eta))/K)^{1/2}$ or, with the null imposed, $(\eta\,(1-\eta)\,/K)^{1/2}$. Also a uniform test over η could be performed, using the convergence $\sqrt{K}\sup_{\eta\in[0,1]}|(\hat{\pi}^*_{n,m,K}(\eta)-\eta)| \xrightarrow{w^*}_p \sup_{[0,1]}|W|$ as $(n,m,K\to\infty)$ implied by Proposition 5.1 below. Under valid specification, however, such tests would reproduce the problem of outcome dependence on the bootstrap variates employed, here $W^*_{n,k}$ ($k = 1,\dots,K$), and results would again differ across researchers.

Proposition 5.1 *Under the assumptions of Theorem 3.1, $\sqrt{K}(\hat{\pi}^*_{n,m,K}(\cdot)-(\cdot)) \xrightarrow{w^*}_p W(\cdot)$ on $\mathscr{D}[0,1]$ as $(n,m,K\to\infty)$, where W denotes a standard Brownian bridge.*

5.2 Alternative discrepancy measures

The discussion so far has focused on the KS or uniform norm, and the norms it dominates. Apart from, e.g., the mentioned signed KS and Cramer-von Mises norms, also the seminorm $\sup_A|\cdot|$ for an interval A (e.g., $A := [1.96,\infty)$), leading to $\hat{d}_n(A) = \sup_{u\in A}|\hat{G}_n(u)-\Phi(u)|$, is dominated by the KS norm, and similarly $\hat{d}_n(x) := |\hat{G}_n(x)-\Phi(x)|$ obtained by focusing on a single point in the support (e.g., $x = 1.96$). They all satisfy the rate condition in Assumption 1 whenever the KS norm does so, and are covered by the theory.

In addition, range-based, quantile-based discrepancy measures or measures focusing on specific moments of the bootstrap distributions could be used. For instance, a moment-based discrepancy based on the third and fourth moments can be defined as

$$\hat{d}_n := \|v_n\|_\Omega := v_n'\Omega v_n,\; v_n := \int_{\mathbb{R}}(u^3,u^4)'(d\hat{G}_n(u)-d\Phi(u)) = \mathbb{E}^*(T_n^{*3},T_n^{*4}-3)'$$

with Ω a symmetric p.d. matrix. Such discrepancy measures need not be continuous on $\mathscr{D}_{\mathbb{R}}$ and a strengthening of (3.2) along the lines discussed, e.g., in Hahn and Liao (2021) may be required for their convergence to zero in probability.

[4] In fact, in some applications the set of bootstrap draws used to compute $\mathscr{T}^*_{n,m}$ is given as an input for the diagnostic procedure. This, for instance, may happen when a replication package includes the set of bootstrap repetitions used to compute a bootstrap statistic but not the code used to generate them. In such cases, unless the set of repetitions are split into subsamples, only one realization of $p^*_{n,B}$ can be computed.

A minimal justification for the use of $\|\cdot\|_\Omega$ in applications is that the ensuing $\hat{d}^*_{n,m}$ can be written as $\hat{d}^*_{n,m} = \psi_m(T^*_{n:1}, ..., T^*_{n:m})$ for a continuous $\psi_m : \mathbb{R}^m \to \mathbb{R}$. Under the null that $T^*_n \xrightarrow{d^*}_p Z \sim \mathcal{N}(0,1)$, it holds that $\hat{d}^*_{n,m} \xrightarrow{d^*}_p d^*_m := \psi_m(Z_1, ..., Z_m)$ as $n \to \infty$, with the Z_i's independent $\mathcal{N}(0,1)$, $i = 1, ..., m$. Therefore, for large n, tests could be conducted using the finite-m quantiles of $\psi_m(Z_1, ..., Z_m)$, or even its large m asymptotic quantiles. Similar considerations apply to the popular Shapiro-Wilks statistic.

5.3 ALTERNATIVE NULL HYPOTHESES

The discrepancy measures discussed so far compare the bootstrap cdf with the $\mathcal{N}(0,1)$ cdf. Suppose, however, that interest is in assessing the convergence $T^*_n \xrightarrow{d^*}_p \mathcal{N}(0,\sigma^2)$ for some unspecified $\sigma^2 > 0$. This could be done by redefining the reference statistic as $\tilde{T}^*_n := T^*_n/\hat{\sigma}_n$, $\hat{\sigma}^2_n := \mathbb{V}^*[T^*_n]$, provided $\hat{\sigma}^2_n$ is consistent[5] for σ^2, such that $\tilde{T}^*_n \xrightarrow{d^*}_p \mathcal{N}(0,1)$. Similarly, assessing the convergence $T^*_n \xrightarrow{d^*}_p \mathcal{N}(\mu,\sigma^2)$ with unknown μ and $\sigma^2 > 0$ can be done by considering $\check{T}^*_n := (T^*_n - \hat{\mu}_n)/\hat{\sigma}_n$, $\hat{\mu}_n := \mathbb{E}^*[T^*_n]$, such that $\check{T}^*_n \xrightarrow{d^*}_p \mathcal{N}(0,1)$ if $(\hat{\mu}_n, \hat{\sigma}^2_n)$ is consistent for (μ,σ^2). Here both $\hat{\mu}_n$ and $\hat{\sigma}^2_n$ can be calculated with arbitrary precision by using a sufficiently large number $M \gg m$ of bootstrap repetitions.

Because $\hat{\mu}_n$ and $\hat{\sigma}^2_n$ are functions of the data (i.e., D_n-measurable), the theory of Section 3 can be applied to $\tilde{T}^*_n$ and $\check{T}^*_n$, such that under the conditions of Theorem 3.1 their asymptotic distributions are not affected by uncertainty due to the estimation of σ^2 (and μ). Specifically, maintaining the notation $\hat{G}_n$ for the bootstrap cdf of T^*_n, the rate condition in Assumption 1 boils down to $||\hat{G}(\cdot\hat{\sigma}_n) - \Phi(\cdot)|| = O_p(n^{-\alpha})$ and $||\hat{G}(\cdot\hat{\sigma}_n + \hat{\mu}_n) - \Phi(\cdot)|| = O_p(n^{-\alpha})$ respectively for $\tilde{T}^*_n$ and $\check{T}^*_n$. The condition can be checked either directly or by using the estimates

$$
\begin{aligned}
||\hat{G}(\cdot\hat{\sigma}_n) - \Phi(\cdot)|| &\le ||\hat{G}(\cdot\sigma) - \Phi(\cdot)|| + |\hat{\sigma}^2_n - \sigma^2| O_p(1), \\
||\hat{G}(\cdot\hat{\sigma}_n + \hat{\mu}_n) - \Phi(\cdot)|| &\le ||\hat{G}(\cdot\sigma + \mu) - \Phi(\cdot)|| + (|\hat{\sigma}^2_n - \sigma^2| + |\hat{\mu}_n - \mu|) O_p(1),
\end{aligned}
$$

which hold whenever $\hat{\sigma}^2_n$ and $\hat{\mu}_n$ are consistent. Then, if $||\hat{G}(\cdot\sigma) - \Phi(\cdot)|| = O_p(n^{-\alpha})$ and $\hat{\sigma}^2_n - \sigma^2 = O_p(n^{-1/2})$, it follows that $||\hat{G}(\cdot\hat{\sigma}_n) - \Phi(\cdot)|| = O_p(n^{-\min\{\alpha,1/2\}})$, and similarly in the case of additional centering.

Finally, statistics such as $\mathcal{T}^*_{n,m} := m^{1/2}\check{d}^*_{n,m}$, $\check{d}^*_{n,m} := \|\check{G}^*_{n,m} - \Phi\|$, could also be implemented, with $\check{G}^*_{n,m}$ the edf of the standardized bootstrap sample $(T^*_{n:i} - m^*_{n,m})/s^*_{n,m}$, $i = 1, \ldots, m$, and $m^*_{n,m}$ and $s^*_{n,m}$ the sample mean and standard deviation of $T^*_{n:1}, \ldots, T^*_{n,m}$, respectively. The resulting tests, which would be similar to Lilliefors' normality test, are not covered by the theory in Section 3 because $(T^*_{n:i} - m^*_{n,m})/s^*_{n,m}$ are not conditionally i.i.d. given the data. The continuity considerations of Section 5.2 carry over, however.

6 APPLICATIONS

We introduce here some additional applications where the bootstrap can be used to detect specification invalidity. These applications, which are kept deliberately simple, focus on detecting failures of the following assumptions: (i) stationarity, (ii) parameter in the interior of the parameter space, (iii) finite variance, and (iv) non-singular Jacobian in applications

[5] The condition $\mathbb{E}^*|T^*_n|^{2+\epsilon} = O_p(1)$ for some $\epsilon > 0$ suffices.

of the delta method. For each application, we discuss the set up and conditions ensuring that the main theory results from Sections 3 and 4 hold.

6.1 NON-STATIONARITY

Assume that the data are generated by a standard autoregressive recursion:

$$y_t = \phi_0 y_{t-1} + \varepsilon_t,\ t = 1, \ldots, n$$

where $\{\varepsilon_t\}_{t=1}^{\infty}$ is an i.i.d. sequence r.v.'s with $\mathbb{E}[\varepsilon_t] = 0$ and $\sigma^2 := \mathbb{E}[\varepsilon_t^2] = 1$; y_0 is assumed to be fixed. The least squares estimator of ϕ_0 is $\hat{\phi}_n := \sum_{t=1}^n y_t y_{t-1} / \sum_{t=1}^n y_{t-1}^2$; under the stability condition $|\phi_0| < 1$, it holds that

$$T_n := \frac{\hat{\phi}_n - \phi_0}{\mathrm{se}(\hat{\phi}_n)} \xrightarrow{d} Z \sim \mathscr{N}(0,1)$$

where $\mathrm{se}(\hat{\phi}_n) := \hat{\sigma}_n (\sum_{t=1}^n y_{t-1}^2)^{-1/2}$, $\hat{\sigma}_n^2$ being the sample variance of $\hat{\varepsilon}_t := y_t - \hat{\phi}_n y_{t-1}$. A simple parametric, recursive bootstrap generates the bootstrap data as follows:

$$y_t^* = \hat{\phi}_n y_{t-1}^* + \varepsilon_t^*,\ \ \varepsilon_t^* \sim \text{i.i.d.} \mathscr{N}(0,1),\ t = 1, \ldots n,$$

initialized at $y_0^* := y_0$, with $\{\varepsilon_t^*\}_{t=1}^{\infty}$ independent of the original data. The bootstrap analog of T_n satisfies, as $n \to \infty$,

$$T_n^* := \frac{\hat{\phi}_n^* - \hat{\phi}_n}{\mathrm{se}(\hat{\phi}_n^*)} \xrightarrow{d^*}_p Z \sim \mathscr{N}(0,1),$$

$\hat{\phi}_n^*$ and $\mathrm{se}(\hat{\phi}_n^*)$ being the bootstrap analogs of $\hat{\phi}_n$ and $\mathrm{se}(\hat{\phi}_n)$, respectively; see Bose (1988).

Suppose now that $\phi_0 = 1$ or, more generally, that $\phi_0 = \phi_n = 1 + \lambda n^{-1}$, $\lambda \in \mathbb{R}$, such that the data are non-stationary. Then

$$T_n \xrightarrow{d} \xi(\lambda, B) := \Big(\int_0^1 J_\lambda(u)^2 du \Big)^{-1/2} \int_0^1 J_\lambda(u) dB(u)$$

where B is a standard Brownian motion on $\mathscr{D}_{[0,1]}$ and J_λ the Ornstein-Uhlenbeck process on $\mathscr{D}_{[0,1]}$ satisfying the stochastic differential equation $dJ = \lambda J + dB$ (see, e.g., Phillips, 1987, or Andrews and Guggenberger, 2009). An implication is that the bootstrap statistic T_n^* has a non-Gaussian, random limit:

$$T_n^* \xrightarrow{d^*}_w \xi(\ell, B^*) | \ell \tag{6.1}$$

where the random term $\ell \sim \int_0^1 J_\lambda(u) dB(u) / \int_0^1 J_\lambda(u)^2 du$ arises from the weak convergence $n(\hat{\phi}_n - \phi_0) \xrightarrow{d} \ell$ and B^* is a standard Brownian motion on $[0,1]$, independent of ℓ; see Basawa et al. (1991). In terms of cdf's, (6.1) is equivalent to the weak convergence in $\mathscr{D}_{\mathbb{R}}$

$$\hat{G}_n(x) := \mathbb{P}^*(T_n^* \le x) \xrightarrow{w} \mathscr{G}(x) := \mathbb{P}\left(\xi(\ell, B^*) \le x | \ell\right). \tag{6.2}$$

Again, $\mathscr{G} \neq \Phi$ a.s. as asymptotic normality of the bootstrap statistic fails.

VALIDITY OF THE DIAGNOSTIC PROCEDURE. Bose (1988) provides sufficient conditions for the nonparametric bootstrap based on least-squares residuals to admit an Edgeworth expansion. These primarily require that $\mathbb{E}[\varepsilon_t^8] < \infty$ and that the autoregressive characteristic roots (i.e., $1/\phi_0$) lie outside the unit circle; see his conditions (A.1)–(A.3). In particular, these conditions imply that $\| \hat{G}_n - \Phi \|_\infty = O_p(n^{-1/2})$; hence, Assumption 1 is satisfied with $\alpha = 1/2$ for the KS norm and dominated norms. Consequently, by Theorem 3.1 validity of the diagnostic procedure for the null hypothesis of stationarity requires that $m/n \to 0$ as $(n, m \to \infty)$. If, as suggested earlier in this section, the Gaussian parametric bootstrap is used instead, then (A.1) and (A.2) in Bose (1988) are automatically satisfied on the bootstrap data, irrespective of the properties of the original ε_t's, provided that $\hat{\phi}_n$, which is used to generate the bootstrap data recursively, is consistent.

Under non-stationarity ($\phi_0 = \phi_n = 1 + \lambda n^{-1}$), it holds that $\| \hat{G}_n - \Phi \|_\infty \xrightarrow{d} \mathscr{Y} := \| \mathscr{G} - \Phi \|_\infty > 0$ with $\mathscr{G}$ as in (6.2), and Theorem 3.2 guarantees that non-stationarity can be detected with probability approaching one by tests employing the KS or dominated norms.

6.2 PARAMETER ON THE BOUNDARY

As in Andrews (2000), consider an i.i.d. sample $D_n := \{y_i\}_{i=1}^n$ from a population with $\mathbb{E}[y_i] =: \theta_0 \in \Theta := [0, \infty)$, Θ being the parameter space, and $\mathbb{V}[y_i^2] = 1$. The Gaussian quasi-maximum likelihood estimator [QMLE], given by $\hat{\theta}_n := \max\{0, \bar{y}_n\}$ with $\bar{y}_n := n^{-1}\sum_{i=1}^n y_i$, is such that, when $\theta_0 \in \operatorname{int}\Theta$,

$$T_n := \sqrt{n}(\hat{\theta}_n - \theta_0) = \max\{-\sqrt{n}\theta_0, \sqrt{n}(\bar{y}_n - \theta_0)\} \xrightarrow{d} Z,\ Z \sim \mathscr{N}(0,1). \tag{6.3}$$

Consider the Gaussian parametric bootstrap MLE, $\hat{\theta}_n^* := \max\{0, \bar{y}_n^*\}$, where conditionally on the original data D_n, $\bar{y}_n^* \sim \mathscr{N}(\hat{\theta}_n, n^{-1/2})$. The bootstrap analog of T_n is $T_n^* := \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n) = \max\{-\sqrt{n}\hat{\theta}_n, \sqrt{n}(\bar{y}_n^* - \hat{\theta}_n)\}$ and satisfies

$$T_n^* | D_n \sim \max\{-\sqrt{n}\hat{\theta}_n, Z^*\},\ Z^* \sim \mathscr{N}(0,1)$$

with conditional cdf $\hat{G}_n(x) = \Phi(x)\mathbb{I}_{\{x \geq -\sqrt{n}\hat{\theta}_n\}}$. When θ_0 is an interior point of Θ, since $\sqrt{n}\hat{\theta}_n \to -\infty$ (a.s.) as $n \to \infty$, it holds that

$$T_n^* \xrightarrow{d^*}_p Z^*,\ Z^* \sim \mathscr{N}(0,1) \tag{6.4}$$

and the bootstrap replicates the standard normal distribution.

Suppose now that θ_0 is on the boundary ($\theta_0 = 0$) or 'close' to it ($\theta_0 = \lambda n^{-1/2}$, $\lambda \in [0, \infty)$). In this case, (6.3) no longer holds; instead

$$T_n = \max\{-\sqrt{n}\theta_0, \sqrt{n}(\bar{y}_n - \theta_0)\} \xrightarrow{d} \xi(\lambda) := \max\{-\lambda, Z\}$$

which differs from a normal random variable. Similarly, and in contrast to (6.4), the bootstrap cdf satisfies $\hat{G}_n(\cdot) = \Phi(\cdot)\mathbb{I}_{\{\cdot \geq -\sqrt{n}\hat{\theta}_n\}} \to_w \Phi(\cdot)\mathbb{I}_{\{\cdot \geq -\ell\}}$ in $\mathscr{D}_{\mathbb{R}}$, which is established using the convergence fact $\sqrt{n}\hat{\theta}_n = \sqrt{n}(\hat{\theta}_n - \theta_0) + \sqrt{n}\theta_0 \xrightarrow{d} \xi(\lambda) + \lambda =: \ell$. This is equivalent to the weak convergence in distribution

$$T_n^* \xrightarrow{d^*}_w \xi^*(\ell) | \ell$$

where $\xi^*(\ell) := \max\{-\ell, Z^*\}$, $Z^* \sim \mathscr{N}(0,1)$ independent of ℓ. Since ℓ is finite a.s., the conditional cdf of T_n^* is non-Gaussian in the limit, with probability one.

VALIDITY OF THE DIAGNOSTIC PROCEDURE. As seen above, conditionally on the data the parametric bootstrap statistic T_n^* has conditional cdf $\hat{G}_n(x) = \Phi(x)\mathbb{I}_{\{x \geq -\sqrt{n}\hat{\theta}_n\}}$, and the associated KS distance $\hat{d}_n := \| \hat{G}_n - \Phi \|_\infty$ satisfies

$$\hat{d}_n = \sup_{x\in\mathbb{R}} |\Phi(x)\mathbb{I}_{\{x \geq -\sqrt{n}\hat{\theta}_n\}} - \Phi(x)| = \sup_{x\in\mathbb{R}} |\Phi(x)\mathbb{I}_{\{x \leq -\sqrt{n}\hat{\theta}_n\}}| = \Phi(-\sqrt{n}\hat{\theta}_n).$$

Under the null that θ_0 is a fixed interior point ($\theta_0 > 0$), $\hat{d}_n \xrightarrow{p} 0$ exponentially fast, and the KS norm satisfies Assumption 1 for any $\alpha > 0$. This implies that any power growth rate of m in terms of n is allowed in Theorem 3.1 for tests employing the KS distance.

In contrast, when θ_0 is on (or near) the boundary, $\hat{d}_n$ satisfies

$$\hat{d}_n = \Phi(-\sqrt{n}\hat{\theta}_n) \xrightarrow{d} \mathscr{Y} = \Phi(-\ell),$$

with ℓ as previously defined. Thus, $\hat{d}^*_{n,m}$ diverges at the rate of $\sqrt{m}$ as $(n, m \to \infty)$.

6.3 INFINITE VARIANCE

As in Section 6.2, assume that $D_n := \{y_i\}_{i=1}^n$ where the y_i's are i.i.d. with $\theta_0 := \mathbb{E}[y_i]$ and $\sigma^2 := \mathbb{V}[y_i^2] \in (0, \infty)$. This assumption implies that the sample mean $\hat{\theta}_n := n^{-1}\sum_{i=1}^n y_i$ satisfies $T_n := \sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma \xrightarrow{d} Z$, $Z \sim \mathscr{N}(0,1)$, as $n \to \infty$. With $\{y_i^*\}_{i=1}^n$ denoting a sample of n draws, independent conditionally on D_n, from the edf of $\{y_i\}_{i=1}^n$, the i.i.d. bootstrap counterpart of T_n is $T_n^* := \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n$, where $\hat{\theta}_n^* := n^{-1}\sum_{i=1}^n y_i^*$, $\hat{\sigma}_n^2 := n^{-1}\sum_{i=1}^n (y_i - \hat{\theta}_n)^2$. As is known, as $n \to \infty$ (e.g., Singh, 1981), $T_n^* \xrightarrow{d^*}_p Z$, $Z \sim \mathscr{N}(0,1)$; in terms of cdfs, $\hat{G}_n(x) := \mathbb{P}^*(T_n^* \leq x) \xrightarrow{p} \Phi(x)$.

Consider now the case where $\mathbb{E}[y_t^2] = \infty$. Specifically, assume that y_t is in the domain of attraction of a symmetric stable law with tail index $\nu \in (1,2)$, denoted as $\mathscr{S}(\nu)$. In this case the CLT fails to hold; instead, for some diverging real sequence $\{a_n\}$,

$$a_n n^{1/2} T_n = a_n n(\hat{\theta}_n - \theta_0) \xrightarrow{d} \mathscr{S}(\nu);$$

see, e.g., Feller (1971). Note that $\mathscr{S}(\nu)$ can be written as $\mathscr{S}(\nu) \sim \sum_{k=1}^\infty \delta_k Z_k$, where the δ_k's are i.i.d. Rademacher and $Z_k^{1/\nu} := \sum_{i=1}^k E_i$ with $\{E_i\}_{i=1}^\infty$ an i.i.d. sequence of exponential r.v.'s with $\mathbb{E}[E_i] = 1$; see Lepage, Woodroofe and Zinn (1981). Athreya (1987) and Knight (1989) show that it this case also the i.i.d. bootstrap counterpart of T_n is not asymptotically Gaussian; precisely, the bootstrap measure is random in the limit:

$$T_n^* := a_n n(\hat{\theta}_n^* - \hat{\theta}_n) \xrightarrow{d^*}_w \frac{\sum_{k=1}^\infty \delta_k Z_k (M_k^* - 1)}{(\sum_{k=1}^\infty Z_k^2)^{1/2}} \Big| \{\delta_k, Z_k\}, \tag{6.5}$$

where the δ_k's and Z_k's are as previously defined. The M_k^*'s, which are i.i.d. $\mathscr{P}(1)$ r.v.'s, independent of $\{\delta_k, Z_k\}$, induce the randomness in the limit bootstrap measure; see Theorem 2 in Knight (1989). If a wild bootstrap with Rademacher multipliers is used instead, that is, $\hat{\theta}_n^* := \hat{\theta}_n + n^{-1}\sum_{i=1}^n (y_i - \hat{\theta}_n)w_i^*$, with the w_i^*'s being i.i.d. Rademacher random variables conditionally on D_n, then $T_n^* \xrightarrow{d^*}_w \sum_{k=1}^\infty \delta_k^* Z_k (\sum_{k=1}^\infty Z_k^2)^{-1/2} \big| \{Z_k\}$, see Cavaliere, Georgiev and Taylor (2013, 2016). For both bootstrap schemes, the asymptotic bootstrap measure is random and a.s. non-Gaussian.

VALIDITY OF THE DIAGNOSTIC PROCEDURE. Consider first the case of valid specification where the y_t's have finite variance. As seen above, $\hat{G}_n(x) := \mathbb{P}^*(T_n^* \leq x) \xrightarrow{p} \Phi(x)$. The rate of the previous convergence, required as an input in Assumption 1, can be established under slightly more than finite second moments.

Specifically, if $\mathbb{E}[|y_t|^{2\kappa}] < \infty$ for some $\kappa > 1$, then by the Berry-Esseen bound

$$\|\hat{G}_n - \Phi\|_\infty \leq \frac{C}{n^{1/2}} \mathbb{E}^*[|y_i^* - \hat{\theta}_n|^3] = \frac{C}{n^{1/2}} \frac{1}{n} \sum_{i=1}^n |y_i - \hat{\theta}_n|^3$$

For $\kappa \geq \frac{3}{2}$ (such that $\mathbb{E}|y_i|^3 < \infty$), the r.h.s. of the previous equation is $O_{a.s.}(n^{-1/2})$, while for $\kappa \in (1, 3/2)$ it is $o_{a.s.}(n^{-\frac{3(\kappa-1)}{2\kappa}})$ by the Marcinkiewicz-Zigmund strong law. Hence, the KS norm satisfies Assumption 1 with $\alpha = \min\{\frac{3(\kappa-1)}{2\kappa}, \frac{1}{2}\}$ and Theorem 3.1 applies. Agnostically, one may select m as growing at a logarithmic rate, e.g., $m = \ln(n)$, which would satisfy the requirement in Theorem 3.1 for any $\kappa > 1$.

Consider now the case where $\mathbb{E}[y_i^2] = \infty$. Then, T_n^* has the random limit given in (6.5); by Theorem 3 in Knight (1989), its random limiting cdf, $\mathscr{G}$ say, is a.s. sample-path continuous. Hence, (6.5) implies that $\|\hat{G}_n - \Phi\|_\infty \to_d \mathscr{Y} := \|\mathscr{G} - \Phi\|_\infty > 0$ a.s.; Theorem 3.2 applies and the diagnostic test based on the KS norm rejects with probability approaching one.

6.4 NEAR-SINGULAR JACOBIAN

Suppose that an estimator $\hat{\theta}_n$ of an unknown parameter θ_0 satisfies

$$Z_n := \sqrt{n}(\hat{\theta}_n - \theta_0)/\sigma \xrightarrow{d} Z \sim \mathscr{N}(0,1), \tag{6.6}$$

and a bootstrap analog $Z_n^* := \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n$ is available such that $Z_n^* \xrightarrow{d^*}_p Z^* \sim \mathscr{N}(0,1)$ jointly with (6.6), with Z^* independent of Z and $\hat{\sigma}_n^2 \xrightarrow{p} \tilde{\sigma}^2 > 0$ not necessarily equal to σ^2.

Consider inference on $\tau_0 := g(\theta_0)$, where g is twice continuously differentiable. With $\hat{\tau}_n := g(\hat{\theta}_n)$, let $T_n := \sqrt{n}(\hat{\tau}_n - \tau_0)$. By a second-order expansion around θ_0,

$$T_n = \dot{g} Z_n + \tfrac{1}{2} g''(\bar{\theta}_n) n^{-1/2} \sigma^2 Z_n^2$$

where $\bar{\theta}_n$ is on the line segment between θ_0 and $\hat{\theta}_n$, and $\dot{g} := g'(\theta_0)$. Provided $\dot{g} \neq 0$, it holds that $T_n \xrightarrow{d} \sigma \dot{g} Z$.

Now, consider the bootstrap analog of T_n. With $\hat{\tau}_n^* := g(\hat{\theta}_n^*)$ and $\hat{\tau}_n := g(\hat{\theta}_n)$, we have

$$T_n^* := \sqrt{n}(\hat{\tau}_n^* - \hat{\tau}_n) = g'(\hat{\theta}_n)\hat{\sigma}_n Z_n^* + \tfrac{1}{2} g''(\bar{\theta}_n^*) n^{-1/2} \hat{\sigma}_n^2 Z_n^{*2}$$

for some $\bar{\theta}_n^*$ between $\hat{\theta}_n$ and $\hat{\theta}_n^*$. As for T_n, if $\dot{g} \neq 0$ then $T_n^* \xrightarrow{d^*}_p \tilde{\sigma} \dot{g} Z$; hence, T_n^* is asymptotically normal.

Suppose instead that $\dot{g} := g'(\theta_0) = \lambda n^{-1/2}$, $\lambda \in [0, \infty)$, such that the Jacobian is singular ($\lambda = 0$) or nearly singular ($\lambda \in (0, \infty)$). Then $T_n = o_p(1)$ while, provided $\ddot{g} := g''(\theta_0) \neq 0$, $\sqrt{n} T_n$ has a chi-square-type limit distribution:

$$\sqrt{n} T_n = \lambda \sigma Z_n + \tfrac{1}{2} g''(\hat{\theta}_n^*) \sigma^2 Z_n^2 \xrightarrow{d} \lambda \sigma Z + \tfrac{1}{2} \ddot{g} \sigma^2 Z^2.$$

Similarly, using the convergence fact

$$\sqrt{n}g'(\hat{\theta}_n) = \sqrt{n}g'(\theta_0) + g''(\theta_0)\sqrt{n}(\hat{\theta}_n - \theta_0) + o_p(1) \xrightarrow{d} \ell := \lambda + \ddot{g}\sigma Z,$$

which holds jointly with the convergence of Z_n^*, it follows that T_n^* satisfies

$$\sqrt{n}T_n^* = n(\hat{\tau}_n^* - \hat{\tau}_n) = \sqrt{n}g'(\hat{\theta}_n)\hat{\sigma}_n Z_n^* + \tfrac{1}{2}g''(\bar{\theta}_n^*)\hat{\sigma}_n^2 Z_n^{*2} \xrightarrow{d^*}_w \ell\tilde{\sigma}Z^* + \tfrac{1}{2}\ddot{g}\tilde{\sigma}^2 Z^{*2}\big|\,\ell \tag{6.7}$$

where $Z^* \sim \mathscr{N}(0,1)$ is independent of ℓ. The right hand side of (6.7) defines a random, non-Gaussian distribution, which we denote by $\mathscr{G}$.

VALIDITY OF THE DIAGNOSTIC PROCEDURE. Assume first that the Jacobian is non-singular; i.e., $\ddot{g} \neq 0$. If Z_n^* admits a standard Edgeworth expansion, such that its conditional cdf, say $\hat{F}_n$, satisfies $\| \hat{F}_n - \Phi \|_\infty = O_p(n^{-1/2})$, then also $\hat{G}_n$, the conditional cdf of T_n^*, satisfies $\| \hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot) - \Phi(\cdot) \|_\infty = O_p(n^{-1/2})$; see Appendix C. As $\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot)$ is the edf of the bootstrap t-ratio $\tilde{T}_n^* := T_n^*/(g'(\hat{\theta}_n)\hat{\sigma}_n)$, which is well-defined with probability approaching one, it follows that Assumption 1 is verified with $\alpha = 1/2$ for $\tilde{T}_n^*$ and the KS norm, and Theorem 3.1 applies for diagnostics based on $\tilde{T}_n^*$.

In contrast, when the Jacobian is near zero, $\sqrt{n}g'(\hat{\theta}_n) \xrightarrow{d} \ell$ and (6.7) hold jointly. As a result, it is shown in Appendix C that

$$\hat{d}_n := \|\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot) - \Phi\|_\infty \to 1 - \mathscr{G}(0)\mathbb{I}_{\{\ell\geq 0\}} > 0 \text{ a.s.}$$

Further, the conclusions of Theorem 3.2 are valid for $\tilde{T}_n^*$-based diagnostics employing the KS norm; see again Appendix C.

7 AN EMPIRICAL ILLUSTRATION

To illustrate the usefulness and potential of our bootstrap diagnostic approach, in this section we consider an example from the empirical macroeconomic framework. We focus on the strategy employed by Känzig (2021) to identify a structural oil supply news shock within a structural VAR identified with an external instrument (SVAR-IV), see Stock and Watson (2018). The novelty of Känzig's (2021) analysis lies in his construction of an external instrument, z_t, for the oil supply news shock, $\varepsilon_{\text{oil},t}$, by extending the high-frequency (HF) approach, originally introduced by Gertler and Karadi (2015), outside the monetary policy framework. Specifically, in Kanzig's (2021) baseline model, the instrument z_t is the first principal component of six time series which reflect variations in oil price futures with different maturities around OPEC production announcements.

Känzig's (2021) model includes six variables ($g = 6$): real oil prices, world oil production, world oil inventories, world industrial production, US industrial production, and the US consumer price index (CPI). These variables, observed monthly over the period 1974M1–2017M12 ($n = 528$), are collected in the $g \times 1$ vector y_t and modeled through a VAR system with $p = 12$ lags. Interest is in the dynamic responses of y_{t+h} at horizon $h \in \{0, 1, \ldots\}$ to an oil supply news shock of magnitude $\varepsilon_{\text{oil},t} := \mathsf{x}$, measured as

$$\mathsf{IRF}(h) := \mathbb{E}[y_{t+h} \mid \varepsilon_{\text{oil},t} = \mathsf{x}, \varepsilon_{2,t} = 0, \mathcal{I}_{t-1}] - \mathbb{E}[y_{t+h} \mid \varepsilon_t = 0, \mathcal{I}_{t-1}] = (R'\mathcal{C}_\Pi^h R)b\mathsf{x} \tag{7.1}$$

where $\varepsilon_t := (\varepsilon_{\text{oil},t}, \varepsilon_{2,t}')'$, $\varepsilon_{2,t}$ being the vector of latent non-target (not of interest) shocks, $\mathcal{I}_{t-1}$ is the information set at time $t-1$, $\mathcal{C}_\Pi$ the VAR companion matrix, R a selection matrix such that $R'R = I_g$, and b a $g \times 1$ vector of structural parameters capturing the instantaneous response of the variables to the shock (note that $\mathsf{IRF}(0) = b\mathsf{x}$).

The dynamic causal effects in (7.1) are identified and estimated using z_t as instrument for $\varepsilon_{\text{oil},t}$. Under relevance and exogeneity of z_t, i.e., $\mathbb{E}[z_t\varepsilon_{\text{oil},t}] = \phi \neq 0$ and $\mathbb{E}[z_t\varepsilon_{2,t}'] = 0$,

$$\mathbb{E}[u_t z_t] = \phi b,\ b = (b_1, b_2')'; \tag{7.2}$$

here, b is partitioned conformably with the vector of structural shocks. The moment condition (7.2) is the key ingredient for the estimation of (7.1). With $u_t := (u_{1,t}, u_{2,t}')'$ (and $y_t := (y_{1,t}, y_{2,t}')'$) also partitioned conformably, (7.2) implies the representation:

$$\beta := b_2/b_1 = \mathbb{E}[u_{2,t}z_t]/\mathbb{E}[u_{1,t}z_t] \tag{7.3}$$

where β captures the (relative) instantaneous responses of the variables in $y_{2,t}$ to a supply news shock of magnitude $\varepsilon_{\text{oil},t} = \mathsf{x} = b_1^{-1}$, i.e., such that the instantaneous response of $y_{1,t}$ to $\varepsilon_{\text{oil},t}$ equals 1 (the so-called 'unit effect' normalization). The IV estimator of β is given by $\hat{\beta} := S_{\hat{u}_1 z}^{-1} S_{\hat{u}_2 z}$, where $\hat{u}_t := (\hat{u}_{1,t}, \hat{u}_{2,t}')'$, $t = 1, ..., n$, are the VAR residuals. For $\mathsf{x} = b_1^{-1}$, the IRFs (7.1) are estimated by replacing β with $\hat{\beta}$ and the VAR companion matrix $\mathcal{C}_\Pi$ with the corresponding OLS estimate.

Similarly to the IV regression framework of Section 6.3, Montiel Olea, Stock and Watson (2020) show that if z_t is a weak instrument, the estimator of $\mathsf{IRF}(h)$ is inconsistent and asymptotically non-Gaussian. As is typical in the literature, Känzig (2021) pre-tests the strength of the instrument z_t using an F-test from a first-stage regression of $\hat{u}_{1,t}$ on z_t. The reported 'regular' F-statistic in his Table 1 is 22.7, which is above the homoskedastic threshold of 10.3 (Stock and Yogo, 2005). In addition, Känzig (2021) reports a 'robust' F-statistic of 10.6. Based on this evidence, which Känzig (2021) considers supportive of z_t being a strong instrument, the IRFs are estimated as described above, and their uncertainty is assessed using 68% and 90% moving block bootstrap confidence intervals (Jentsch and Lunsford, 2019, 2022), based on 10,000 bootstrap replications. These confidence intervals, shown in Känzig's Figure 3, appear broadly in line with the uncertainty commonly encountered by applied macroeconomists.

By exploiting Känzig's bootstrap computations, we are able to reassess specification validity of his SVAR-IV model through the lens of bootstrap diagnostics. In this case, assuming correct specification of the VAR model, the null hypothesis of valid specification corresponds to the instrument z_t being sufficiently strong for the target shock $\varepsilon_{\text{oil},t}$, a condition that ensures that IRFs are estimated consistently and standard asymptotic inference is valid.

Using Känzig's (2021) Matlab code, we first estimate the 5 bootstrap distributions $\hat{G}_n^{(j)}(\cdot) := \mathbb{P}^*(\hat{\beta}_j^* - \hat{\beta}_j \leq \cdot)$, $j = 1, \ldots, 5$, using 10,000 moving block bootstrap replications of the vector $\hat{\beta}^*$ (properly standardized, see Section 5.3). The $\hat{G}_n^{(j)}$'s, reported in the upper panel of Figure 2, are substantially non-Gaussian. For each $j = 1, \ldots, 5$, we run $K = 500$ tests based on independent samples of $m \in \{10, 20\}$ bootstrap replications; m is

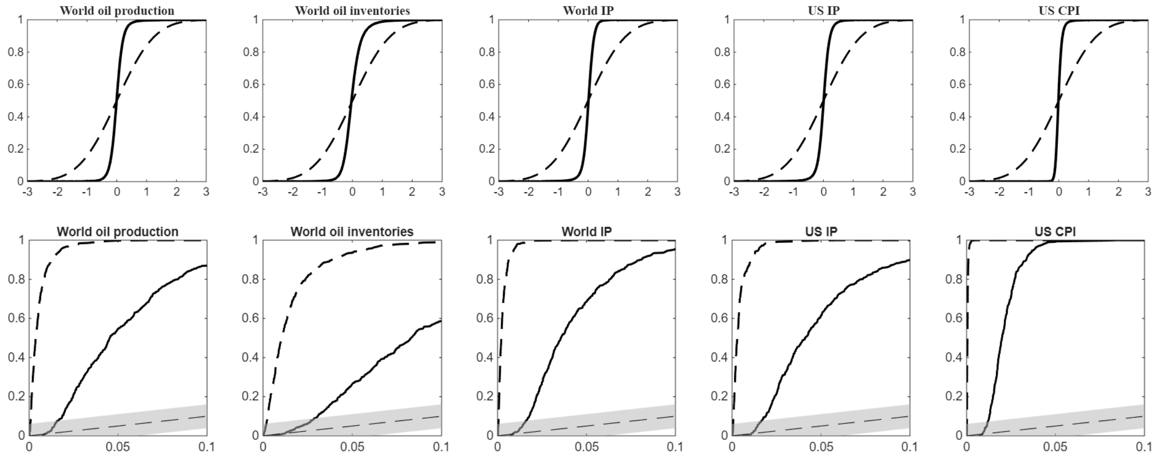

Figure 2: Bootstrap diagnostics. Upper panel: bootstrap (conditional) cdfs for the on-impact IRFs (solid lines) against the Gaussian distribution (dashed lines). Lower panel: average fractions of rejections ($\hat{\pi}^*_{n,m,K}(\eta)$) for nominal levels $\eta \in (0, 0.10)$, $m = 10$ (solid lines) and $m = 20$ (dashed lines).

deliberately small with respect to n. The lower panel of Figure 2 reports, for significance levels $\eta \in (0, 0.1)$, the average rejection rates $\hat{\pi}^*_{n,m,K}(\eta) := K^{-1} \sum_{k=1}^{K} \mathbb{I}_{\{p^*_{n,m:k} \leq \eta\}}$, where $p^*_{n,m:k}$, $k = 1, \ldots, K$ are the p-values associated to the K tests; see Section 5.1. Note that due to use of the moving block bootstrap, these results are robust to the presence of conditional heteroskedasticity of unknown form in the data (Jentsch and Lunsford, 2022).

It can be observed that for all components of β, the null hypothesis is strongly rejected at any conventional significance level. Larger values of m reinforce this conclusion. This result[6] likely indicates that the SVAR-IV estimation is based on a weak instrument. Interestingly, neither Känzig's (2021) 68% and 90% bootstrap confidence intervals nor the F-tests fully capture this phenomenon, which the bootstrap diagnostic test detects effectively.

8 Concluding Remarks

This paper shows that the bootstrap delivers, as a by-product, a versatile diagnostic procedure for detecting invalid specifications, i.e., violations of the assumptions underlying standard asymptotic theory. The procedure, which is based on measures of the discrepancy between the conditional distribution of a bootstrap statistic and the Gaussian distribution, provides a flexible and computationally straightforward alternative to traditional misspecification tests. A key advantage of this approach is that it does not induce pre-testing bias, thereby improving the reliability of post-test statistical inference. That is, under the null of valid specification, inference conditional on not rejecting the bootstrap tests is asymptotically *exact*.

A further major feature of this procedure is its flexibility. On the one hand, it can be tailored to test the validity of a specific assumption, such as stationarity of the data or relevance of a set of instrumental variables, hence allowing to construct tests specifically designed to have power against alternatives of interest. On the other hand, it can function as a more general diagnostic test for bootstrap consistency; that is, to test whether the

[6] Results are robust to (i) the choice of the horizon h, (ii) the choice of the norm and (iii) the standardization of the bootstrap estimator.

distribution of a given bootstrap statistic is close to a theoretical Gaussian limit. This dual capability makes the method broadly applicable, whether testing for specific irregularities or assessing the validity of bootstrap procedures in general.

Our paper also contributes to the literature on bootstrap inference. The usual approach in bootstrap theory is to analyze the asymptotic properties of a bootstrap cdf, say $\hat{G}_n$, as the sample size n diverges. In practice, however, $\hat{G}_n$ is estimated using the edf $\hat{G}^*_{n,m}$ of m (conditionally i.i.d.) realization of the bootstrap statistic. Therefore, standard bootstrap theory is implicitly based on letting $m \to \infty$ first (such that the estimation error $\hat{G}^*_{n,m} - \hat{G}_n$ disappears), followed by $n \to \infty$. Our paper complements the standard approach by exploring the case where $n, m \to \infty$ jointly, rather than sequentially. We show that, under suitable conditions $\hat{G}^*_{n,m}$ has an asymptotic distribution which is known under a wide range of applications and, crucially, becomes asymptotically independent of the original data. This asymptotic independence result is key to avoid that the bootstrap test of valid specification distorts post-test inference.

The results in this paper can be extended in several directions. For instance, we have not discussed the role of the choice of the norm in determining the finite sample behavior and the power properties of the proposed tests. Second, an open question is how to efficiently choose m in practice, such that an optimal balance of size and power is achieved. Third, all the examples considered here assume that the reference statistical model is of low dimension. It is of interest to establish whether our approach could be used to detect validity of the Gaussian asymptotic approximations in high-dimensional cases. Fourth, to what extent the properties of the bootstrap highlighted in this paper are shared by other methods such as the jackknife (see, e.g., Hansen, 2025), permutation tests (Young, 2019), subsampling (Politis, Romano and Wolf, 1999) or (quasi) Bayesian methods (as in Wang, 2025) are open questions. These are left to future research.

APPENDIX

A NOTATION AND DEFINITIONS

Throughout this paper, the notation $\sim$ indicates equality in distribution. We write '$x := y$' and '$y =: x$' to mean that x is defined by y. The standard Gaussian cdf is denoted by Φ; $\mathscr{U}_{[0,1]}$ and $\mathscr{P}(\lambda)$ are the uniform distribution on $[0,1]$ and the Poisson distribution with mean λ, respectively. $\mathbb{I}_{\{\cdot\}}$ is the indicator function. The space of càdlàg functions $\mathbb{R} \to \mathbb{R}$ (equipped with its Skorokhod J_1-topology, unless otherwise stated; see Kallenberg, 1997, Appendix A2) is denoted by $\mathscr{D}_{\mathbb{R}}$; similarly, $\mathscr{D}_{[0,1]}$ denotes the space of càdlàg functions $[0,1] \to \mathbb{R}$. For matrices a, b, c with n rows, $S_{ab} := a'b/n$ and $S_{ab.c} := S_{ab} - S_{ac}S_{cc}^{-1}S_{cb}$, assuming that S_{cc} has full rank.

For a bootstrap sequence, say Y_n^*, we use $Y_n^* \xrightarrow{p^*}_p 0$ or $Y_n^* = o_{p^*}(1)$ to mean that, for any $\epsilon > 0$, $\mathbb{P}^*(|Y_n^*| > \epsilon) \to_p 0$, where $\mathbb{P}^*$ denotes the probability measure conditionally on the original data D_n. We also write $Y_n^* = O_{p^*}(1)$ to mean that $\mathbb{P}^*(|Y_n^*| > M) \to_p 0$ for some large enough M. Expectations and variance under $\mathbb{P}^*$ are denoted with $\mathbb{E}^*$ and $\mathbb{V}^*$, respectively. We use $Y_n^* \xrightarrow{d^*}_p \xi$ to mean that $\mathbb{E}^*[g(Y_n^*)|D_n] \xrightarrow{p} \mathbb{E}[g(\xi)]$ for all bounded continuous functions $g : \mathbb{R} \to \mathbb{R}$. In terms of cdfs, $\hat{G}_n(u) := \mathbb{P}^*(Y_n^* \le u) \to_p G(u) := \mathbb{P}(\xi \le u)$ at all continuity points $u \in \mathbb{R}$ of G.

We refer to the fact that $\mathbb{E}[g(Y_n^*)|D_n] \xrightarrow{w} \mathbb{E}[g(\xi)|\mathcal{D}]$ for all bounded continuous functions $g : \mathbb{R} \to \mathbb{R}$ (with (Y_n^*, D_n) and $(\xi, \mathcal{D})$ random elements of metric spaces $\mathbb{R} \times \mathcal{S}_{D_n}$ and $\mathbb{R} \times \mathcal{S}_{\mathcal{D}}$, respectively) as 'weak convergence in distribution', denoted as '$\xrightarrow{d^*}_w$'. For the special case of scalar random variables Y_n^* and ξ, if the conditional distribution $\xi|\mathcal{D}$ is diffuse (non-atomic), weak convergence in distribution is equivalent to the weak convergence $\hat{G}_n(\cdot) := \mathbb{P}^*(Y_n^* \le \cdot) \xrightarrow{w} \mathscr{G}(\cdot) := \mathbb{P}(\xi \le \cdot|\mathcal{D})$ in $\mathscr{D}_{\mathbb{R}}$, where $\mathscr{G}$ is a random distribution function. For multivariate generalizations, see Cavaliere and Georgiev (2020, Appendix A). Finally, if G is a (random) cdf, G^{-1} denotes its right-continuous generalized inverse.

B PROOFS OF THE MAIN RESULTS

In the proof of Theorem 3.1, we make use of the following lemma, which provides a bound on $a_{n,m}^* := \mathscr{T}_{n,m}^* - \mathscr{Z}_{n,m}^* = m^{1/2}\|\hat{G}_{n,m}^* - \Phi\| - m^{1/2}\|\hat{G}_{n,m}^* - \hat{G}_n\|$ of (3.5). Its proof follows directly from the triangle inequality for (semi)norms.

LEMMA B.1 *With $a_{n,m}^*$ defined in (3.5), it holds that $|a_{n,m}^*| \le \sqrt{m}\|\hat{G}_n - \Phi\|$.*

PROOF OF THEOREM 3.1. By Lemma B.1 we have that $|a_{n,m}^*| \le \sqrt{m}\|\hat{G}_n - \Phi\| = O_p(m^{1/2}n^{-\alpha})$ under Assumption 1. The limit of $\mathscr{Z}_{n,m}^*$ can be found, e.g., as in Bickel and Freedman (1981), Theorem 4.1. Specifically, let $\psi_m(F)$ denote the law of $\sqrt{m}(\hat{F}_m(\cdot) - F(\cdot))$, where $\hat{F}_m$ is the edf of m independent r.v.s with common law F; then, in the space of probability measures on $\mathscr{D}_{\mathbb{R}}$ equipped with the weak topology, $\psi_m(F)$ tends to the law of $W(F)$ as $m \to \infty$ by a standard invariance principle, where W is a standard Brownian bridge. Conditionally on D_n, $\sqrt{m}(\hat{G}_{n,m}^*(\cdot) - \hat{G}_n(\cdot))$ has law $\psi_m(\hat{G}_n)$; since $\|\hat{G}_n - \Phi\|_\infty \to_p 0$, by Proposition 4.1 of Bickel and Freedman (1981) it also holds that $\psi_m(\hat{G}_n)$ becomes arbitrarily close (in probability) to $\psi_m(\Phi)$ as $(n, m \to \infty)$, where $\psi_m(\Phi)$ converges (weakly) to the law of $W(\Phi)$ as $m \to \infty$. Thus, the law of $\sqrt{m}(\hat{G}_{n,m}^*(\cdot) - \hat{G}_n(\cdot))$ conditionally on D_n converges weakly in probability to the law of $W(\Phi)$ as $(n, m \to \infty)$. By the CMT it follows that, conditionally on the data, $\|\sqrt{m}(\hat{G}_{n,m}^*(\cdot) - \hat{G}_n(\cdot))\| \xrightarrow{w}_p \|W(\Phi)\|$ as $(n, m \to \infty)$. The requirement that $m^{1/2}n^{-\alpha} \to 0$ completes the proof. □

PROOF OF THEOREM 3.2. By Skorokhod coupling, consider a probability space where distributional copies of $\hat{G}_n$ and $\mathscr{G}$ are defined (for simplicity, we denote also these copies by $\hat{G}_n$ and $\mathscr{G}$), such that $\hat{G}_n \xrightarrow{a.s.} \mathscr{G}$ in $\mathscr{D}_{\mathbb{R}}$ as $n \to \infty$. Fix an outcome ω in the probability-one event $\{\hat{G}_n \to \mathscr{G}\} \cap \{\|\mathscr{G} - \Phi\| > 0\}$. For this outcome there exists a sequence of increasing continuous bijections $\lambda_n : [0,1] \to [0,1]$ such that $\|\hat{G}_n(\omega) \circ \lambda_n - \mathscr{G}\|_\infty \to 0$ and $\|\lambda_n - \mathrm{id}_{[0,1]}\|_\infty \to 0$ as $n \to \infty$.

Consider further a product extension of the coupling probability space such that on the added factor space m i.i.d. draws from $\hat{G}_n(\omega)$ are defined, with associated empirical process $\hat{G}_{n,m}^*(\omega)$, and also m i.i.d. draws of $\mathscr{G}(\omega)$ are defined, with associated empirical process $\mathscr{G}_m(\omega)$ $(m = 1, 2, \ldots)$. Finally, let $\mathbb{P}^*$ denote the probability measure on the added factor space.

As the triangle inequality yields $\|\hat{G}^*_{n,m}(\omega) - \Phi\| \geq \|\hat{G}_n(\omega) - \Phi\| - \|\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega)\|$, it holds that

$$\begin{aligned}\mathbb{P}^*(\mathscr{T}^*_{n,m}(\omega) \geq c) &:= \mathbb{P}^*(\|\hat{G}^*_{n,m}(\omega) - \Phi\| \geq cm^{-1/2}) \\ &\geq \mathbb{P}^*(\|\hat{G}_n(\omega) - \Phi\| - \|\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega)\| \geq cm^{-1/2}) \\ &= \mathbb{P}^*(\|\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega)\| \leq \|\hat{G}_n(\omega) - \Phi\| - cm^{-1/2}).\end{aligned}$$

Here, first, the rhs of the inequality satisfies $\|\hat{G}_n(\omega) - \Phi\| - cm^{-1/2} \to \|\mathscr{G}(\omega) - \Phi\| =: \mathscr{Y}(\omega) > 0$ as $(n, m \to \infty)$, by the continuity of $\|\cdot\|$. Second, the lhs satisfies $\|\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega)\| \xrightarrow{\mathbb{P}^*} 0$ as $(n, m \to \infty)$ by the following argument. On the one hand, as $(n, m \to \infty)$, $\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega) \xrightarrow{\mathbb{P}^*} 0$ in $\mathscr{D}_{\mathbb{R}}$ if and only if $\hat{G}^*_{n,m}(\omega) \circ \lambda_n - \hat{G}_n(\omega) \circ \lambda_n \xrightarrow{\mathbb{P}^*} 0$ in $\mathscr{D}_{\mathbb{R}}$. On the other hand, regarding the latter, it is checked directly that $\hat{G}^*_{n,m}(\omega) \circ \lambda_n$ is an empirical process for the cdf $\hat{G}_n(\omega) \circ \lambda_n$. By Proposition 4.1 of Bickel and Freedman (1981), the convergence $\|\hat{G}_n(\omega) \circ \lambda_n - \mathscr{G}(\omega)\|_\infty \to 0$ as $n \to \infty$ implies that the law of $\sqrt{m}(\hat{G}^*_{n,m}(\omega) \circ \lambda_n - \hat{G}_n(\omega) \circ \lambda_n)$, as a probability measure on $\mathscr{D}_{\mathbb{R}}$, becomes arbitrarily close to the law of $\sqrt{m}(\mathscr{G}_m(\omega) - \mathscr{G}(\omega))$ as $(n, m \to \infty)$. As the latter law converges to the law of $W(\mathscr{G}(\omega))$ as $m \to \infty$ by a standard invariance principle, so does the law of $\sqrt{m}(\hat{G}^*_{n,m}(\omega) \circ \lambda_n - \hat{G}_n(\omega) \circ \lambda_n)$ as $(n, m \to \infty)$. By the CMT, this implies that $\|\hat{G}^*_{n,m}(\omega) \circ \lambda_n - \hat{G}_n(\omega) \circ \lambda_n\|_\infty = O_{\mathbb{P}^*}(m^{-1/2})$ as $(n, m \to \infty)$. Hence, $\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega) \xrightarrow{\mathbb{P}^*} 0$ in $\mathscr{D}_{\mathbb{R}}$ as $(n, m \to \infty)$. By the continuity of $\|\cdot\|$, it follows that $\|\hat{G}^*_{n,m}(\omega) - \hat{G}_n(\omega)\| \xrightarrow{\mathbb{P}^*} 0$ as $(n, m \to \infty)$, as asserted previously. Therefore,

$$\liminf_{(n,m\to\infty)} \mathbb{P}^*(\mathscr{T}^*_{n,m}(\omega) \geq c) \geq 1.$$

Since ω was chosen in an almost certain event, it can be concluded that $\mathbb{P}^*(\mathscr{T}^*_{n,m} \geq c) \to 1$ a.s. on the coupling probability space. Therefore, on the original probability space, $\mathbb{P}^*(\mathscr{T}^*_{n,m} \geq c) \xrightarrow{p} 1$. □

Proof of Theorem 4.1. (a) Say $\mathscr{T}^*_{n,m} \xrightarrow{w}_p \mathscr{T}^*_\infty$ as $(n, m \to \infty)$. By the law of iterated expectations, it holds that

$$\begin{aligned}\left|\mathbb{E}[f(D_n)g(\mathscr{T}^*_{n,m})] - \mathbb{E}[f(D_n)]\mathbb{E}[g(\mathscr{T}^*_{n,m})]\right| &= \left|\mathbb{E}\{f(D_n)[\mathbb{E}^*(g(\mathscr{T}^*_{n,m})) - \mathbb{E}(g(\mathscr{T}^*_{n,m}))]\}\right| \\ &\leq \|f\|_\infty \mathbb{E}\left|\mathbb{E}^*(g(\mathscr{T}^*_{n,m})) - \mathbb{E}(g(\mathscr{T}^*_{n,m}))\right|.\end{aligned}$$

It follows that as $(n, m \to \infty)$

$$\begin{aligned}&\sup_{\|f\|_\infty \leq 1} \left|\mathbb{E}[f(D_n)g(\mathscr{T}^*_{n,m})] - \mathbb{E}[f(D_n)]\mathbb{E}[g(\mathscr{T}^*_{n,m})]\right| \\ &\qquad \leq \mathbb{E}\left|\mathbb{E}^*[g(\mathscr{T}^*_{n,m})] - \mathbb{E}[g(\mathscr{T}^*_\infty)]\right| + |\mathbb{E}[g(\mathscr{T}^*_{n,m})] - \mathbb{E}[g(\mathscr{T}^*_\infty)]| \to 0\end{aligned}$$

because $\mathbb{E}^*[g(\mathscr{T}^*_{n,m})] \xrightarrow{p} \mathbb{E}[g(\mathscr{T}^*_\infty)]$ and the dominated convergence theorem applies.
(b) It is sufficient to establish that

$$\sup_{A_n \in \sigma(D_n)} |\mathbb{P}(A_n \cap \{\mathscr{T}^*_{n,m} \in B\}) - \mathbb{P}(A_n)\mathbb{P}(\mathscr{T}^*_{n,m} \in B)| \to 0 \tag{B.1}$$

as $(n, m \to \infty)$. This can be done using part (a) and approximations of indicator functions with continuous functions, or directly as done below. Then, as $\mathbb{P}^*(\mathscr{T}^*_{n,m} \in B) \xrightarrow{p} \mathbb{P}(\mathscr{T}^*_\infty \in$

$B) =: \mathbb{P}_\infty(B)$ for the $\mathbb{P}_\infty$-continuity set B, also $\mathbb{P}(\mathscr{T}^*_{n,m} \in B) \to \mathbb{P}_\infty(B) > 0$ by dominated convergence. Thus, division by $\mathbb{P}(\mathscr{T}^*_{n,m} \in B)$, which is bounded away from zero under the assumption that $\mathbb{P}_\infty(B) > 0$, completes the proof.

We turn to (B.1). By the law of iterated expectations (and the fact that $\mathbb{P}(X \in \mathcal{E}) = \mathbb{E}(\mathbb{I}_{\{X\in\mathcal{E}\}})$), it holds that

$$\begin{aligned}&\left|\mathbb{P}(A_n \cap \{\mathscr{T}^*_{n,m} \in B\}) - \mathbb{P}(A_n)\mathbb{P}(\mathscr{T}^*_{n,m} \in B)\right| \\ &\quad= \left|\mathbb{E}[\mathbb{E}^*(\mathbb{I}_{A_n}\mathbb{I}_{\{\mathscr{T}^*_{n,m}\in B\}})] - \mathbb{E}[\mathbb{I}_{A_n}\mathbb{E}(\mathbb{I}_{\{\mathscr{T}^*_{n,m}\in B\}})]\right| \\ &\quad= \left|\mathbb{E}\{\mathbb{I}_{A_n}[\mathbb{E}^*(\mathbb{I}_{\{\mathscr{T}^*_{n,m}\in B\}}) - \mathbb{E}(\mathbb{I}_{\{\mathscr{T}^*_{n,m}\in B\}})]\}\right| \\ &\quad\le \mathbb{E}\left|\mathbb{P}^*(\mathscr{T}^*_{n,m} \in B) - \mathbb{P}(\mathscr{T}^*_{n,m} \in B)\right|.\end{aligned}$$

It follows that

$$\begin{aligned}&\sup_{A_n\in\sigma(D_n)} \left|\mathbb{P}(A_n \cap \{\mathscr{T}^*_{n,m} \in B\}) - \mathbb{P}(A_n)\mathbb{P}(\mathscr{T}^*_{n,m} \in B)\right| \\ &\quad\le \mathbb{E}\left|\mathbb{P}^*(\mathscr{T}^*_{n,m} \in B) - \mathbb{P}_\infty(B)\right| + \left|\mathbb{P}(\mathscr{T}^*_{n,m} \in B) - \mathbb{P}_\infty(B)\right| \to 0\end{aligned}$$

as $(n, m \to \infty)$, because $\mathbb{P}^*(\mathscr{T}^*_{n,m} \in B) \xrightarrow{p} \mathbb{P}_\infty(B)$ and dominated convergence applies. □

Proof of Proposition 5.1. First note that $\hat{\pi}^*_{n,m,K}(\cdot)$ with $\cdot \in \mathbb{R}$ is an empirical process. Then, the proof is analogous to that of Theorem 14.3 in Billingsley (1999). Specifically, first note that, if $\xi^*_{n,m:k}$, $k = 1, \ldots K$, are i.i.d. uniform conditionally on the data, and $\hat{\xi}^*_{n,m}$ is their empirical process, then the result $\sqrt{K}(\hat{\xi}^*_{n,m,K}(\cdot) - (\cdot)) \xrightarrow{w^*}_p W(\cdot)$ on $\mathscr{D}[0,1]$ as $(n, m, K \to \infty)$ is standard. Second, the representation $\sqrt{K}(\hat{\pi}^*_{n,m,K}(\cdot) - (\cdot)) = \sqrt{K}(\hat{\xi}^*_{n,m,K}(\Pi^*_{n,m}(\cdot)) - (\cdot))$ can be used, where $\Pi^*_{n,m}$ is the cdf of $p^*_{n,m}$ conditionally on the data. That is, $\sqrt{K}(\hat{\pi}^*_{n,m,K}(\cdot)-(\cdot))$ can be represented as the composition $\sqrt{K}(\hat{\xi}^*_{n,m,K}(\cdot)-(\cdot))$ after $\Pi^*_{n,m}(\cdot)$. Since the limiting law of $\sqrt{K}(\hat{\xi}^*_{n,m,K}(\cdot) - (\cdot))$ is that of the continuous process W and since $\Pi^*_{n,m}(\cdot)$ converges to the identity function in probability, by the CMT $\sqrt{K}(\hat{\pi}^*_{n,m,K}(\cdot) - (\cdot))$ converges to the composition $W(\cdot)$ after $id(\cdot)$, which is $W(\cdot)$. □

C Additional proofs and results

Results in Section 3.4. Write

$$T^*_n = \frac{\hat{\omega}_n^{-1}}{(\hat{\pi}_n + S_{zv^*})'(\hat{\pi}_n + S_{zv^*})}(\sqrt{n\hat{\omega}_n}\hat{\pi}'_n S_{zu^*} + \sqrt{n\hat{\omega}_n}S'_{zv^*}S_{zu^*}).$$

Here the term $\sqrt{n\hat{\omega}_n}\hat{\pi}'_n S_{zu^*}$ is $\mathscr{N}(0,1)$ conditionally on D_n. This fact and the inequality

$$\begin{aligned}|\mathbb{P}(\zeta(\xi+\eta) \le x) - \mathbb{P}(\xi \le x)| &\le \mathbb{P}(\tfrac{x}{1-a\,\mathrm{sgn}(x)} + b) - \mathbb{P}(\tfrac{x}{1+a\,\mathrm{sgn}(x)} - b) \\ &\quad + \mathbb{P}(|\zeta - 1| \ge a) + \mathbb{P}(|\eta| \ge b),\end{aligned}$$

which holds for any r.v.'s ξ, η, ζ and any $a \in (0,1)$ and $b > 0$, yield

$$\begin{aligned}|\hat{G}_n(x) - \Phi(x)| &\le \Phi(\tfrac{x}{1-a\,\mathrm{sgn}(x)} + b) - \Phi(\tfrac{x}{1+a\,\mathrm{sgn}(x)} - b) \\ &\quad + \mathbb{P}^*\left(\left|\frac{\hat{\omega}_n^{-1}}{(\hat{\pi}_n + S_{zv^*})'(\hat{\pi}_n + S_{zv^*})} - 1\right| \ge a\right) + \mathbb{P}^*(\sqrt{n\hat{\omega}_n}|S'_{zv^*}S_{zu^*}| \ge b).\end{aligned} \tag{C.1}$$

Let $a \in (0, \frac{1}{2})$ in the following. Then, by the mean-value theorem and the boundedness of $|x|\,\Phi'(x)$ on $\mathbb{R}$, it holds that

$$\Phi\left(\tfrac{x}{1-a\operatorname{sgn}(x)} + b\right) - \Phi\left(\tfrac{x}{1+a\operatorname{sgn}(x)} - b\right) \leq C\,(a+b)$$

for some universal constant $C > 0$. As $(\hat{\pi}_n + S_{zv^*})'\,(\hat{\pi}_n + S_{zv^*}) \xrightarrow{p} \pi'\pi > 0$, the event in the first tail probability in (C.1) equals, with probability approaching one, the complement of the event $\left\{-a(1+a)^{-1} < \hat{\omega}_n(2\hat{\pi}_n' S_{zv^*} + S_{zv^*}' S_{zv^*}) < a(1-a)^{-1}\right\}$, such that the tail probability of interest does not exceed

$$\begin{aligned}\mathbb{P}^*\left(\hat{\omega}_n|2\hat{\pi}_n' S_{zv^*} + S_{zv^*}' S_{zv^*}| \geq 2a\right) &\leq \mathbb{P}^*(|2\hat{\pi}_n' S_{zv^*}| \geq a\hat{\omega}_n^{-1}) + \mathbb{P}^*\left(S_{zv^*}' S_{zv^*} \geq a\hat{\omega}_n^{-1}\right) \\ &= \bar{\Phi}(\sqrt{n\hat{\omega}_n^{-1}}\tfrac{a}{2}) + \bar{\Psi}_k(n\hat{\omega}_n^{-1}a),\end{aligned}$$

the last line because $\hat{\pi}_n' S_{zv^*} \sim N(0, n^{-1}\hat{\omega}_n^{-1})$ and $nS_{zv^*}' S_{zv^*} \sim \chi^2(k)$; here $\bar{\Phi} = 1 - \Phi$ and $\bar{\Psi}_k$ denote resp. the complementary cdfs of the $N(0,1)$ and the $\chi^2(k)$ distributions. To discuss the second tail probability in (C.1), by using the decomposition $v_i^* = \rho_{uv}u_i^* + \sqrt{1-\rho_{uv}^2}\varepsilon_i^*$ with $(u_i^*, \varepsilon_i^*) \sim$i.i.d.$N(0, I_2)$, it is found that $S_{zv^*}' S_{zu^*} = \rho_{uv} S_{zu^*}' S_{zu^*} + \sqrt{1-\rho_{uv}^2} S_{z\varepsilon^*}' S_{zu^*}$ and

$$\begin{aligned}\mathbb{P}^*\left(\sqrt{n\hat{\omega}_n}|S_{zv^*}' S_{zu^*}| \geq b\right) &\leq \mathbb{P}^*\left(\sqrt{n\hat{\omega}_n}|\rho_{uv}|S_{zu^*}' S_{zu^*} \geq \tfrac{b}{2}\right) \\ &\quad + \mathbb{P}^*\left(\sqrt{n\hat{\omega}_n(1-\rho_{uv}^2)}|S_{z\varepsilon^*}' S_{zu^*}| \geq \tfrac{b}{2}\right) \\ &\leq \bar{\Psi}_k\left(\sqrt{n\hat{\omega}_n^{-1}}\tfrac{b}{2|\rho_{uv}|}\right) + k\bar{\Pi}\left(\sqrt{n\hat{\omega}_n^{-1}(1-\rho_{uv}^2)^{-1}}\tfrac{b}{2k}\right),\end{aligned}$$

the last line because $nS_{zu^*}' S_{zu^*} \sim \chi^2(k)$ and $nS_{z\varepsilon^*}' S_{zu^*} \sim \xi'\eta$ with $(\xi', \eta')' \sim N(0, I_{2k})$; here $\bar{\Pi}$ stands for the complementary cdf of $\xi_1\eta_1$. Recapitulating, with probability approaching one, it holds that

$$\begin{aligned}\|\hat{G}_n - \Phi\|_\infty \leq C(a+b) + \bar{\Phi}\left(\sqrt{n\hat{\omega}_n^{-1}}\tfrac{a}{2}\right) + \bar{\Psi}_k(n\hat{\omega}_n^{-1}a) + \\ + \bar{\Psi}_k\left(\sqrt{n\hat{\omega}_n^{-1}}\tfrac{b}{2|\rho_{uv}|}\right) - k\bar{\Pi}\left(\sqrt{n\hat{\omega}_n^{-1}(1-\rho_{uv}^2)^{-1}}\tfrac{b}{2k}\right)\end{aligned}$$

for any $a \in (0, \frac{1}{2})$ and $b > 0$. The rate of decay of $\bar{\Phi}$, $\bar{\Psi}_k$ and $\bar{\Pi}$ is well known:

$$\bar{\Phi}(x) \sim Cx^{-1}e^{-x^2/2}, \bar{\Psi}_k(x) \sim Cx^{k/2}e^{-x/2}, \bar{\Pi}(x) \sim x^{-1/2}e^{-x}$$

as $x \to \infty$, each one with its own C; see eq. (2) in Leipus, Siaulys, Dirma and Zove (2023) for the case of $\bar{\Pi}$. By choosing $a = b = \epsilon n^{-1/2}\ln n$ with $\epsilon > 0$ as small as convenient, it is seen that $\|\hat{G}_n - \Phi\|_\infty = O_p(n^{-\alpha})$ for any $\alpha \in (0, \frac{1}{2})$. □

Results in Section 6.4. Let first $\dot{g} \neq 0$ be fixed. Say that a sequence of possibly conditional cdf's F_n is square-root consistent if $\|F_n - \Phi\|_\infty = O_p(n^{-1/2})$. Interest is in establishing the fact that $\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot)$ is square-root consistent under the assumption that the conditional cdf's $\hat{F}_n$ of $Z_n^* := \sqrt{n}(\hat{\theta}_n^* - \hat{\theta}_n)/\hat{\sigma}_n$ do so.

Recall that $T_n^* = g'(\hat{\theta}_n)\hat{\sigma}_n Z_n^* + \frac{1}{2}g''(\hat{\theta}_n)n^{-1/2}\hat{\sigma}_n^2 Z_n^{*2}$. Notice that the conditional cdf's of Z_n^* and $-Z_n^*$ either both are, or both are not square-root consistent. As square-root consistency of for the conditional cdf's of Z_n^* is the only property of Z_n^* used in this proof, it follows that Z_n^* and $-Z_n^*$ are exchangeable for the purposes of the proof. Hence, possibly upon replacing Z_n^* and $g'(\hat{\theta}_n)$ by respectively $Z_n^*(1-2\mathbb{I}_{\{g'(\hat{\theta}_n)<0\}})$ and $g'(\hat{\theta}_n)(1-2\mathbb{I}_{\{g'(\hat{\theta}_n)<0\}})$ in the expression for T_n^*, it is legitimate to discuss T_n^* under the assumption $g'(\hat{\theta}_n) \geq 0$. Under this assumption, and given that $\mathbb{P}(g'(\theta_n) = 0) \to 0$, it holds that

$$\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n x) = \mathbb{P}^*(Z_n^* + \tfrac{1}{2}n^{-1/2}\gamma_n Z_n^{*2} \leq x)$$

with probability approaching one, where $\gamma_n := g''(\hat{\theta}_n)\hat{\sigma}_n / g'(\hat{\theta}_n)$. Next, the conditional cdf's of $\pm(Z_n^* + \frac{1}{2}n^{-1/2}\gamma_n Z_n^{*2})$ either both are, or both are not square-root consistent, and since Z_n^* and $-Z_n^*$ were said to be exchangeable for the purposes of the proof, there is no loss of generality in imposing also $\gamma_n \geq 0$. With the signs imposed, it holds that

$$\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n x) = \hat{F}_n(x)\mathbb{I}_{\{\gamma_n=0\}} + \mathbb{P}^*(\hat{q}_-(x) \leq Z_n^* \leq \hat{q}_+(x))\mathbb{I}_{A_n}$$

with probability approaching one, where $A_n := \{x \geq -\sqrt{n}\gamma_n^{-1}/2, \gamma_n \neq 0\}$ and

$$\hat{q}_\pm(x) := n^{1/2}\gamma_n^{-1}(\pm\sqrt{1 + 2n^{-1/2}x\gamma_n} - 1).$$

From the assumption that the conditional cdf's $\hat{F}_n$ of Z_n^* are square-root consistent, it is further found that

$$\begin{aligned}\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n x) &= \hat{F}_n(x)\mathbb{I}_{\{\gamma_n=0\}} + (\hat{F}_n(\hat{q}_+(x)) - \hat{F}_n(\hat{q}_-(x)))\mathbb{I}_{A_n} \\ &= \Phi(x)\mathbb{I}_{\{\gamma_n=0\}} + (\Phi(\hat{q}_+(x)) - \Phi(\hat{q}_-(x)))\,\mathbb{I}_{A_n} + O_p(n^{-1/2})\end{aligned}$$

uniformly in x. Turn now to

$$\|\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot) - \Phi\|_\infty \leq \sup_{x\in\mathbb{R}} |[\hat{G}(g'(\hat{\theta}_n)\hat{\sigma}_n x) - \Phi(x)]\mathbb{I}_{A_n}| + \sup_{x\in\mathbb{R}} |\Phi(x)\mathbb{I}_{A_n^c}\mathbb{I}_{\{\gamma_n\neq 0\}}|.$$

As $\hat{\theta}_n \xrightarrow{p} \theta_0$ and $g''(\theta_n)$ (thus, γ_n) is bounded, it follows that $\mathbb{P}(A_n^c \cap \{\gamma_n \neq 0\}) \to 0$ and the previous upper bound equals, with probability approaching one,

$$\begin{aligned}\sup_{x\in\mathbb{R}} |[\hat{G}(g'(\hat{\theta}_n)\hat{\sigma}_n x) - \Phi(x)]\mathbb{I}_{A_n}| &= \sup_{x\in\mathbb{R}} |[\Phi(\hat{q}_+(x)) - \Phi(\hat{q}_-(x)) - \Phi(x)]\mathbb{I}_{A_n}| \\ &\leq \sup_{x\in\mathbb{R}} |\Phi(\hat{q}_-(x))\mathbb{I}_{A_n}| + \sup_{4\gamma_n|x|\leq n^{1/2}} |[\Phi(\hat{q}_+(x)) - \Phi(x)]\mathbb{I}_{A_n}| \\ &\quad + \sup_{4\gamma_n|x|>n^{1/2}} |[\Phi(\hat{q}_+(x)) - \Phi(x)]\mathbb{I}_{A_n}|.\end{aligned}$$

Here, by the monotonicity of Φ,

$$\sup_{x\in\mathbb{R}} |\Phi(\hat{q}_-(x))\mathbb{I}_{A_n}| \leq \mathbb{I}_{A_n}\Phi(-n^{1/2}\gamma_n^{-1})| = O_p(n^{-1/2})$$

because $\gamma_n = O_p(1)$. Further, for outcomes in A_n, $|\hat{q}_+(x) - x| \leq 2n^{-1/2}\gamma_n x^2$ by simple algebra. Hence, by the mean-value theorem,

$$|\Phi(\hat{q}_+(x)) - \Phi(x)| \leq 2n^{-1/2}\gamma_n x^2\Phi'\left(x + \zeta_n[\hat{q}_+(x) - x]\right)$$

for some $\zeta_n \in [0,1]$, such that

$$\sup_{4\gamma_n|x|\le n^{1/2}} |[\Phi(\hat{q}_+(x)) - \Phi(x)]\mathbb{I}_{A_n}| \le 2n^{-1/2}\gamma_n \sup_{x\in\mathbb{R}} \left[x^2 \max_{2|y|\le|x|} \Phi'(x+y)\right]$$
$$= 2n^{-1/2}\gamma_n \sup_{x\ge 0}\left[x^2\Phi'\left(\tfrac{x}{2}\right)\right] = O_p(n^{-1/2})$$

by using the exponential functional form of Φ'. Finally, by the monotonicity of Φ,

$$\sup_{4\gamma_n|x|>n^{1/2}} |[\Phi(\hat{q}_+(x)) - \Phi(x)]\mathbb{I}_{A_n}| \le \mathbb{I}_{\{\gamma_n\ne 0\}}[\Phi(-\tfrac{1}{4}n^{1/2}\gamma_n^{-1}) + \Phi(-\hat{q}_+(-\tfrac{1}{4}n^{1/2}\gamma_n^{-1}))]$$
$$\le \mathbb{I}_{\{\gamma_n\ne 0\}}[\Phi(-\tfrac{1}{4}n^{1/2}\gamma_n^{-1}) + \Phi(-n^{1/2}\gamma_n^{-1})] = O_p(n^{-1/2}).$$

By combining the previous results, $\|\hat{G}_n(g'(\hat{\theta}_n)\hat{\sigma}_n\cdot) - \Phi\|_\infty = O_p(n^{-1/2})$.

Next, let instead $g'(\theta_0) = \lambda n^{1/2}$. Then $(\sqrt{n}g'(\hat{\theta}_n), \hat{G}_n) \xrightarrow{w} (\ell, \mathscr{G})$ on $\mathbb{R}\times\mathscr{D}_\mathbb{R}$ as shown in Section 6.4. Further, introducing $\hat{\varsigma}_n := g'(\hat{\theta}_n)\hat{\sigma}_n$, it holds that

$$L_n \le \|\hat{G}_n(\hat{\varsigma}_n\cdot) - \Phi\|_\infty \le U_n$$

with

$$L_n := |\hat{G}_n(n\hat{\varsigma}_n) - \Phi(n)|\mathbb{I}_{\{\hat{\varsigma}_n<0\}}$$
$$+ \max\{|\hat{G}_n(\sqrt[3]{n}\hat{\varsigma}_n) - \Phi(\sqrt[3]{n})|, |\hat{G}_n(-\sqrt[3]{n}\hat{\varsigma}_n) - \Phi(-\sqrt[3]{n})\}\mathbb{I}_{\{\hat{\varsigma}_n\ge 0\}}$$

obtained by evaluating $\hat{G}_n(\hat{\varsigma}_n\cdot) - \Phi$ at n and $\pm\sqrt[3]{n}$, and

$$U_n := \mathbb{I}_{\{\hat{\varsigma}_n<0\}} + \max\{\hat{G}_n(0), 1-\hat{G}_n(0)\}\mathbb{I}_{\{\hat{\varsigma}_n\ge 0\}}$$

obtained by considerations of monotonicity. As the sample paths of $\mathscr{G}$ are a.s. continuous and $\mathbb{P}(\ell = 0) = 0$, it follows from the CMT that

$$(U_n, U_n - L_n) \xrightarrow{d} (\mathbb{I}_{\{\ell<0\}} + \max\{\mathscr{G}(0), 1-\mathscr{G}(0)\}\mathbb{I}_{\{\ell\ge 0\}}, 0) = (1 - \mathscr{G}(0)\mathbb{I}_{\{\ell\ge 0\}}, 0)$$

because $\mathscr{G}(0) < \frac{1}{2}$ a.s. From here the limit of $\|\hat{G}_n(\hat{\varsigma}_n\cdot) - \Phi\|_\infty$ is obtained.

To see that the consistency conclusion of Theorem 3.2 remains valid, some modifications are due in its proof. First, ω is taken as an outcome in the almost certain event $\{\hat{G}_n \to \mathscr{G}\} \cap \{\mathscr{G}(0)\mathbb{I}_{\{\ell\ge 0\}} \ne 1\}$. Second,

$$\mathbb{P}^*(\mathscr{T}^*_{n,m}(\omega) \ge c) := \mathbb{P}^*(\|\hat{G}^*_{n,m}(\omega, \hat{\varsigma}_n(\omega)\cdot) - \Phi\|_\infty \ge cm^{-1/2})$$
$$\ge \mathbb{P}^*(\|\hat{G}^*_{n,m}(\omega, \hat{\varsigma}_n(\omega)\cdot) - \hat{G}_n(\omega, \hat{\varsigma}_n(\omega)\cdot)\|_\infty \le \|\hat{G}_n(\omega, \hat{\varsigma}_n(\omega)\cdot) - \Phi\|_\infty - cm^{-1/2}).$$

Here

$$\|\hat{G}^*_{n,m}(\omega, \hat{\varsigma}_n(\omega)\cdot) - \hat{G}_n(\omega, \hat{\varsigma}_n(\omega)\cdot)\|_\infty = |\hat{G}^*_{n,m}(\omega, 0) - \hat{G}_n(\omega, 0)|\mathbb{I}_{\{\hat{\varsigma}_n(\omega)=0\}}$$
$$+ \|\hat{G}^*_{n,m}(\omega, \cdot) - \hat{G}_n(\omega, \cdot)\|_\infty\mathbb{I}_{\{\hat{\varsigma}_n(\omega)\ne 0\}}$$
$$\le \|\hat{G}^*_{n,m}(\omega, \cdot) - \hat{G}_n(\omega, \cdot)\|_\infty \xrightarrow{\mathbb{P}^*} 0$$

as $(n, m\to\infty)$, by the proof of Theorem 3.2, whereas $\|\hat{G}_n(\omega, \hat{\varsigma}_n(\omega)\cdot) - \Phi\|_\infty - cm^{-1/2} \to 1 - \mathscr{G}(\omega, 0)\mathbb{I}_{\{\ell(\omega)\ge 0\}} > 0$ as $(n, m\to\infty)$. Hence, the conclusion. □

REFERENCES

Andrews, D.W.K. (2000). Inconsistency of the bootstrap when a parameter is on the boundary of the parameter space. *Econometrica*, 68, 399–405.

Andrews, D.W.K. and M. Buchinsky (2000). A three-step method for choosing the number of bootstrap repetitions. *Econometrica* 67, 23–51.

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

Andrews, I. and J. Shapiro (2025). Communicating scientific uncertainty via approximate posteriors. NBER Working Paper 32038.

Andrews, I., J.H. Stock and L. Sun (2019). Weak instruments in instrumental variable regression: theory and practice. *Annual Review of Economics* 11, 727–753.

Angelini, G., G. Cavaliere and L. Fanelli (2022). Bootstrap inference and diagnostics in state space models: with applications to dynamic macro models. *Journal of Applied Econometrics* 37, 3–22.

Angelini, G., G. Cavaliere and L. Fanelli (2024). An identification and testing strategy for proxy-SVARs with weak proxies. *Journal of Econometrics* 238, 105604.

Athreya, K.B. (1987). Bootstrap of the mean in the infinite variance case, *Annals of Statistics* 15, 724–731.

Bårdsen, G. and L. Fanelli (2015). Frequentist evaluation of small DSGE models. *Journal of Business & Economic Statistics* 33, 307–322.

Basawa, I.V., A.K. Mallik, W.P. McCormick, J.H. Reeves, and R.L. Taylor (1991). Bootstrapping unstable first-order autoregressive processes, *Annals of Statistics* 19, 1098–1101.

Beran, R. (1997). Diagnosing bootstrap success, *Annals of the Institute of Statistical Mathematics* 49, 1–24.

Bickel, P.J. and D.A. Freedman (1981). Some asymptotic theory for the bootstrap. *Annals of Statistics* 9, 1196–1217.

Billingsley, P. (1999). *Convergence of probability measures.* 2nd edition, Wiley: NY.

Bound, J., A. Jaeger and R.M. Baker (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association* 90, 443–450.

Bose, A. (1988). Edgeworth correction by bootstrap in autoregressions. *Annals of Statistics* 16, 1709–1722.

Cavaliere, G., and I. Georgiev (2020). Inference under random limit bootstrap measures. *Econometrica* 88, 2547–2574.

Cavaliere, G., I. Georgiev and A.M.R. Taylor (2013). Wild bootstrap of the mean in the infinite variance case. *Econometric Reviews* 32, 204–219.

Cavaliere, G., I. Georgiev and A.M.R. Taylor (2016). Sieve-based inference for infinite-variance linear processes. *Annals of Statistics* 44, 1467–1494.

Cavaliere, G., H.B. Nielsen and A. Rahbek (2015). Bootstrap testing of hypotheses on co-integration relations in vector autoregressive models. *Econometrica*, 83, 813–831.

Datta, S. (1995). On a modified bootstrap for certain asymptotically non-normal statistics. *Statistics and Probability Letters* 24, 91–98.

Davidson, R. (2017). Diagnostics for the bootstrap and fast double bootstrap. *Econometric Reviews* 36, 1021–1038.

Davidson, R. and J.G. Mackinnon (2010). Wild bootstrap tests for IV regression. *Journal of Business & Economic Statistics* 28, 128–144.

de Chaisemartin, C. and X. D'Haultfoeuille (2024). Under the null of valid specification, pre-tests cannot make post-test inference liberal. arXiv:2407.03725; R&R *American Economic Review.*

Feller, W. (1971). *An introduction to probability theory and its applications*, Vol. 2, 2nd edition, Wiley: NY.

Gertler, M. and P. Karadi (2015). Monetary policy surprices, credit costs, and economic activity. *American Economic Journal: Macroeconomics* 7, 44-76.

Hall, P. (1992). *The Bootstrap and Edgeworth Expansion*, Springer-Verlag: Berlin.

Hahn, J. and Z. Liao (2021). Bootstrap standard error estimates and inference. *Econometrica* 89, 1963–1977.

Han, S. and A. McCloskey (2019). Estimation and inference with a (nearly) singular Jacobian. *Quantitative Economics* 10, 1019–1068.

Hansen, B.E. (2025). Jackknife Standard Errors for Clustered Regression. Under revision at the *Review of Economic Studies.*

Jentsch, C. and K.C. Lunsford (2019). The dynamic effects of personal and corporate income tax changes in the United States: Comment. *American Economic Review* 109, 2655–2678.

Jentsch, C. and K.C. Lunsford (2022). Asymptotic valid bootstrap inference for Proxy SVARs. *Journal of Business and Economic Statistics* 40, 1876–1891.

Känzig, D.R. (2021). The macroeconomic effects of oil supply news: evidence from OPEC announcements. *American Economic Review* 111, 1092-1125.

Kallenberg, O. (1997). *Foundations of modern probability.* Springer: NY.

Knight, K. (1989). On the bootstrap of the sample mean in the infinite variance case. *Annals of Statistics* 17, 1168–1175.

Leipus R., J. Siaulys, M. Dirma and R. Zove (2023). On the distribution-tail behaviour of the product of normal random variables. *Journal of Inequalities and Applications*, 32

Lepage, R., M. Woodroofe and J. Zinn (1981). Convergence to a stable distribution via order statistics. *Annals of Probability* 9, 624–632.

Montiel Olea, J.L., J.H. Stock, and M.W. Watson (2021). Inference in SVARs identified with an external instrument. *Journal of Econometrics* 225, 74-87.

Moreira, M.J., J.R. Porter and G.A. Suarez (2009). Bootstrap validity for the score test when instruments may be weak. *Journal of Econometrics* 149, 52–64.

Newey, W.K. and D. McFadden (1994). Large sample estimation and hypothesis testing. In R.F. Engle and D.L. McFadden (eds.), *Handbook of Econometrics* IV, Chap. 36.

Phillips, P.C.B. (1987). Towards a unified asymptotic theory for autoregression. *Biometrika* 74, 535–547.

Phillips, P.C.B. and H.R. Moon (1999). Linear regression limit theory for nonstationary panel data. *Econometrica* 67, 1057–1111.

Politis, D.N., J.P. Romano, and M. Wolf (1999). *Subsampling*. Springer: NY.

Singh, K. (1981). On the asymptotic accuracy of Efron's bootstrap. *Annals of Statistics* 9, 1187–1195.

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

Stock, J.H., and M.W. Watson (2018). Identification and estimation of dynamic causal effects in macroeconomics using external instruments. *Economic Journal* 128, 917–948.

Stock, J.H. and M. Yogo (2005). Testing for weak instruments in linear IV regression. In *Identification and Inference for Econometric Models: Essays in Honor of Thomas Rothenberg*, eds. D.W.K. Andrews, J.H. Stock, pp. 80–108. Cambridge, UK: Cambridge University Press.

Wang, A. (2025). A general diagnostic of the normal approximation in GMM models. *Econometrics Journal* 28, 261–275.

Young, A. (2019). Channeling Fisher: randomization tests and the statistical insignificance of seemingly significant experimental results. *Quarterly Journal of Economics* 134, 557–598.

Zhan, Z. (2018). Detecting weak identification by bootstrap. Accessed on July 1, 2025. https://sites.google.com/site/zhaoguozhan/