Supplement to "A Simple Parametric Model Selection Test"

Susanne M. Schennach*
Department of Economics, Brown University
and
Daniel Wilhelm
Department of Economics, University College London†

July 27, 2016

Abstract

This supplement provides the proofs of all results in the main text, additional results referenced in the main text, and additional simulations.

*This work was made possible in part through financial support from the National Science Foundation via grants SES-0752699 and SES-1061263/1156347, and through TeraGrid computer resources provided by the University of Texas under grant SES-070003.

†The author gratefully acknowledges financial support from a Katherine Dusak Miller Fellowship, a Wesley C. Pickard PhD Fellowship, and from the ESRC Centre for Microdata Methods and Practice at IFS (RES-589-28-0001)

Contents

1 Assumptions and Definitions from the Main Text **2**

2 Data-driven Choice of the Regularization Parameter **5**

3 Invariance of Our Test Statistic Under Permutations **10**

4 Additional Simulations **11**

5 Extensions **14**

6 Proofs **17**

7 Auxiliary Lemmas **42**

1 Assumptions and Definitions from the Main Text

Here we reproduce the assumptions from the main text for convenience because several results in subsequent sections refer to them.

Assumption 1. *For* $k = A, B$, $\sigma_k^2 > 0$, $Var_{P_0}((\ln f_k(X;\theta_k^*))^2) > 0$, *and* $Var_{P_0}(\nabla_{\theta_k} \ln f_k(X;\theta_k^*))$ *is nonsingular.*

Assumption 2. $\Theta \subset \mathbb{R}^{d_\theta}$ *is compact and* $\ln f_k(x;\cdot)$, $k = A, B$, *are twice continuously differentiable.*

Assumption 3. *(i)* $X_1, \ldots, X_n$ *is an i.i.d. sequence of random variables with common distribution* $P_0 \in \mathbf{P}$.

(ii) There is a unique $\theta^* \in int(\Theta)$ *so that* $E_{P_0} g(X;\theta^*) = 0$.

(iii) $E_{P_0}[\nabla^2_{\theta_k} \ln f_k(X;\theta_k^*)]$, $k = A, B$, *are invertible.*

Assumption 4. *(i)* $E_{P_0}[\|\nabla_{\theta_k} \ln f_k(X,\theta_k^*)\|^{2+\delta}] < \infty$ *and* $E_{P_0}[|\ln f_k(X,\theta_k^*)|^{4+\delta}] < \infty$ *for* $k = A, B$ *and some* $\delta > 0$.

(ii) There exists a function $\bar{F}_1(x)$ *such that* $E_{P_0}\bar{F}_1(X) < \infty$ *and, for* $j, k = A, B$, *for all* $\theta = (\theta_A', \theta_B')' \in \Theta$, *for all* $x \in \mathcal{X}$, *and for* $h(x;\theta)$ *being any of the functions* $\ln f_k(x;\theta_k)$, $vec(\nabla^2_{\theta_k} \ln f_k(x;\theta_k))$ *and* $\ln f_k(x;\theta_k)\nabla_{\theta_j} \ln f_j(x;\theta_j)$, *we have* $\|h(x;\theta)\| \leq \bar{F}_1(x)$.

(iii) There exists a function $\bar{F}_2(x)$ *such that* $E_{P_0}[|\bar{F}_2(X)|^{2+\delta}] < \infty$ *and* $\|\nabla_{\theta_k} \ln f_k(x;\theta_k)\| \leq \bar{F}_2(x)$ *for all* $x \in \mathcal{X}$ *and* $k = A, B$.

Assumption 5. $\hat{\varepsilon}_n$ *is a sequence of real-valued, measurable functions of* $X_1, \ldots, X_n$ *such that there exists a sequence* $\{\varepsilon_n\} \in \mathcal{E}$ *with* $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_0}(n^{-1/2})$.

Assumption 6. *Let* $\hat{\varepsilon}_n$ *be a sequence of real-valued, measurable functions of* $X_1, \ldots, X_n$ *such that, for every sequence* $\{P_n\}$ *in* $\mathcal{P}$, *there exists a sequence* $\{\varepsilon_n\} \in \mathcal{E}$ *with* $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$.

Definition 1. *For some fixed* $\delta, \kappa > 0$, $0 < \underline{M} \leq \overline{M} < \infty$, *and an increasing, continuous function* $\epsilon : (0,\infty) \to (0,\infty)$ *with* $\epsilon(0) = 0$, *let* $\mathcal{P}$ *be the set of distributions* P *on* $\mathcal{X}$ *that satisfy the following conditions for* $X \sim P$: *(i) There exists a unique* $\theta^*(P) \in \Theta$ *such that* $E_P g(X;\theta^*(P)) = 0$, *for all* $\mu > 0$, $\inf_{\theta:\|\theta-\theta^*(P)\|\geq\mu} \|E_P g(X;\theta)\| > \epsilon(\mu)$, *and* $B_\kappa(\theta^*(P)) \subseteq \Theta$, *where* $B_\kappa(\theta)$ *denotes a ball in* $\mathbb{R}^{d_\theta}$ *with radius* κ *around* θ. *(ii) There exists a function* $D(x)$ *such that* $E_P[|D(X)|^{2+\delta}] \leq \overline{M}$ *and, for all* $x \in \mathcal{X}$,

$$|\ln f_A(x;\theta_A^*(P)) - \ln f_B(x;\theta_B^*(P))| \leq D(x)\left(E_P\left[|\ln f_A(X;\theta_A^*(P)) - \ln f_B(X;\theta_B^*(P))|^2\right]\right)^{1/2}, \quad (1)$$

where $\theta^*(P) := (\theta_A^*(P)', \theta_B^*(P)')'$. *Further, we have* $E_P[|\ln f_k(X;\theta_k^*(P))|^{4+\delta}] \leq \overline{M}$ *and, similarly,* $E_P[\|\nabla_{\theta_k} \ln f_k(X;\theta_k^*(P))\|^{2+\delta}] \leq \overline{M}$ *for* $k = A, B$. *(iii) There exists a function* $\bar{F}(x)$ *such that* $E_P\bar{F}(X) \leq \overline{M}$ *and, for* $j, k = A, B$, *for all* $\theta = (\theta_A', \theta_B')' \in \Theta$, *for all* $x \in \mathcal{X}$, *and for* $h(x;\theta)$ *being any of the functions* $\ln f_k(X;\theta_k)$, $\nabla_{\theta_k} \ln f_k(X;\theta_k)$, $vec(\nabla^2_{\theta_k} \ln f_k(x;\theta_k))$ *and* $\ln f_k(x;\theta_k)\nabla_{\theta_j} \ln f_j(x;\theta_j)$, *we have* $\|h(x;\theta)\| \leq \bar{F}(x)$. *(iv) For* $k = A, B$, *we have* $\underline{M} \leq \lambda_{min}(H_k(P))$ *and* $\lambda_{max}(H_k(P)) \leq \overline{M}$, *where* $\lambda_{min}(A)$ *and* $\lambda_{max}(A)$, *respectively, denote the smallest and largest eigenvalue of a matrix* A. *Furthermore, for* $h(x;\theta)$ *being any of the functions* $\log f_k(x;\theta_k)$, $(\log f_k(x;\theta_k))^2$, *and* $\nabla_{\theta_k} \log f_k(x;\theta_k)$, $k = A, B$, $\theta := (\theta_A', \theta_B')'$, *we have* $\underline{M} \leq \lambda_{min}(Var(h(X;\theta^*(P))) \leq \lambda_{max}(Var(h(X;\theta^*(P))) \leq \overline{M}$.

Theorem 1. *If Assumptions 1–5 hold, then, under* H_0, $\tilde{t}_n \to_d N(0,1)$ *and, under* $H_A \cup H_B$, $|\tilde{t}_n| \to_p \infty$.

Theorem 2. *Suppose Assumptions 2 and 6 hold. Let* $\mathcal{P}_0$ *be the subset of distributions in* $\mathcal{P}$ *that satisfy the null hypothesis* $d^*(P) = 0$. *Then the regularized t-test of nominal level* α *is uniformly asymptotically of level* α, *viz.*

$$\lim_{n\to\infty} \sup_{P\in\mathcal{P}_0} P\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) = \alpha.$$

Theorem 3. *Suppose Assumptions 2 and 6 hold. Let* $\{P_n\} \in \mathcal{P}_\delta$ *for some localization parameter* $\delta \in \mathbb{R}$. *Denote by* $\{\varepsilon_n\} \in \mathcal{E}$ *a sequence such that* $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ *and* $\varepsilon := plim_{n\to\infty}\hat{\varepsilon}_n$ *under* P_n. *Then, under* P_n,

$$\tilde{t}_n \to_d N(\tilde{\lambda}, 1)$$

with mean

$$\tilde{\lambda} := \lim_{n\to\infty} \frac{\sqrt{n}d^*(P_n)(1+\varepsilon_n/2)}{\sqrt{(1+\varepsilon_n)\sigma^2(P_n) + \varepsilon_n^2(\sigma_A^2(P_n) + \sigma_B^2(P_n))/2}},$$

and $\sigma^2(P) = \sigma_A^2(P) - 2\sigma_{AB}(P) + \sigma_B^2(P)$.

2 Data-driven Choice of the Regularization Parameter

In this section, we provide a data-driven choice of $\hat{\varepsilon}_n$ that minimizes higher-order distortions to size and power of our test. Specifically, we balance the worse-case size distortion if the models were overlapping with the worst-case power loss if the models were not overlapping. The rationale for proceeding in this way is that, in our approach, size distortion only occurs for overlapping models while power loss only occurs when the models are not overlapping. Furthermore, in a finite sample, it may be difficult to accurately test whether the models are overlapping or not (this is the fundamental pre-testing problem we wish to avoid) and hence it is natural to consider both possibilities simultaneously. Such an approach also considerably simplifies the implementation of the method.

Fix $\alpha \in (0, 1/2)$. Let z_β denote the β-quantile of the standard normal distribution, $\phi(\cdot)$ and $\Phi(\cdot)$ the standard normal density and cumulative distribution functions, respectively.

Assumption 7. *For any $n \in \mathbb{N}$, the X_{ni} for $i = 1, \ldots, n$ are iid random variables taking value in $\mathcal{X}$ and drawn from the probability measure P_n converging weakly to some measure P_0 and each $P_n(x)$ admits a Radon-Nikodym derivative $p_n(x)$ with respect to $P_0(x)$.*

Definition 2. *We say that $g : \mathcal{X} \times \Theta \mapsto \mathbb{R}^{d_g}$ for $d_g \in \mathbb{N}$ and Θ is compact (under some metric $d_\theta(\cdot, \cdot)$) satisfies a* ***triangular array dominance condition*** *if*

1. *$g(x, \theta)$ is continuous in θ at each $(x, \theta) \in \mathcal{X} \times \Theta$;*

2. *There exists $G(x)$ such that $E_{P_0}[G(X_{0i})] < \infty$ (for X_{0i} drawn from P_0) and such that, for all $\theta \in \Theta$ and $n \in \mathbb{N}$, $\|g(x, \theta)\| p_n(x) \leq G(x)$ for all $x \in \mathcal{X}$ and for $p_n(x)$ as in Assumption 7;*

3. *There exists* $\bar{G} < \infty$ *such that* $E_{P_n}[\|g(X_{ni}, \theta)\|^4] \leq \bar{G}$ *for all* $i = 1, \ldots, n$, *all* $n \in \mathbb{N}$ *and all* $\theta \in \Theta$.

Assumption 8. $\ln f_A(x, \theta_A)$ *and* $\ln f_B(x, \theta_B)$ *satisfy a triangular array dominance condition.*

Assumption 9. $\nabla^2_{\theta_A} \ln f_A(x, \theta_A)$ *and* $\nabla^2_{\theta_B} \ln f_B(x, \theta_B)$ *satisfy a triangular array dominance condition.*

Assumption 10. $\ln f_k(x, \theta_k) \nabla_{\theta_l} \ln f_l(x, \theta_l)$ *for* $k = A, B$ *and* $l = A, B$ *satisfy a triangular array dominance condition.*

Assumption 11. $E_{P_0}[\nabla^2_{\theta_k} \ln f_k(X, \theta^*_k(P_0))]$ *and* $E_{P_0}[\nabla_{\theta_k} \ln f_k(X_{0i}, \theta^*_k(P_0)) \nabla'_{\theta_k} \ln f_k(X_{0i}, \theta^*_k(P_0))]$ *for* $k = A, B$ *are invertible.*

Assumption 12. *For some* $\delta > 0$, *we have* $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla_{\theta_A} \ln f_A(X_{ni}, \theta^*_A(P_n))\|^{4+\delta}] < \infty$ *and, similarly,* $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla_{\theta_B} \ln f_B(X_{ni}, \theta^*_B(P_n))\|^{4+\delta}] < \infty$.

Assumption 13. *For some* $\delta > 0$, *we have* $\sup_{n \in \mathbb{N}} E_{P_n}[\| \ln f_A(X_{ni}, \theta^*_A(P_n))\|^{8+\delta}] < \infty$ *and, similarly,* $\sup_{n \in \mathbb{N}} E_{P_n}[\| \ln f_B(X_{ni}, \theta^*_B(P_n))\|^{8+\delta}] < \infty$.

Assumption 14. *For some* $\delta > 0$, *we have* $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla^2_{\theta_A} \ln f_A(X_{ni}, \theta^*_A(P_n))\|^{4+\delta}] < \infty$ *and, similarly,* $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla^2_{\theta_B} \ln f_B(X_{ni}, \theta^*_B(P_n))\|^{4+\delta}] < \infty$.

Assumption 15. $\nabla^3_{\theta_A} \ln f_A(x, \theta_A)$ *and* $\nabla^3_{\theta_B} \ln f_A(x, \theta_B)$ *satisfy a triangular array dominance condition.*

Assumption 16. $\sup_{n \in \mathbb{N}} E_{P_n}[\|\nabla_{\theta_k} \ln f_k(X_{ni}, \theta^*_k(P_n)) \nabla_{\theta_l} \ln f_l(X_{ni}, \theta^*_l(P_n))\|^{4+\delta}] < \infty$ *for* $k = A, B$ *and* $l = A, B$ *for some* $\delta > 0$.

Assumption 17. $\nabla^2_k \ln f_k(x, \theta_k) \nabla_{\theta_l} \ln f_l(x, \theta_l)$ *for* $k = A, B$ *and* $l = A, B$ *satisfy a triangular array dominance condition.*

Under these moment conditions, the following theorem establishes expansions of our test's power, $P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right)$, around its asymptotic local power, $\Phi(z_{\alpha/2}+\delta/\sigma)+\Phi(z_{\alpha/2}-\delta/\sigma)$, when the models are distinct, and of our test's size, $P_0(|\tilde{t}_n| > z_{1-\alpha/2})$, around its nominal size α, when the models are equivalent.

Theorem 4. *Fix $\alpha \in (0, 1/2)$ and suppose $\hat{\varepsilon}_n := \varepsilon_n$ is a deterministic sequence in $\mathcal{E}$. Under Assumptions 2 and 7–17, for any distribution P_0 such that $d^*(P_0) = 0$ and $\sigma^2(P_0) = 0$,*

$$P_0(|\tilde{t}_n| > z_{1-\alpha/2}) \leq \alpha + C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n + O(n^{-1/2}) + o(n^{-1/2}\varepsilon_n^{-1}\ln\ln n), \tag{2}$$

where

$$C_{SD} := 2\phi(z_{\alpha/2})\frac{\max\{|tr(H_A^{-1}V_A)|, |tr(H_B^{-1}V_B)|\}}{\sqrt{(\sigma_A^2+\sigma_B^2)/2}}.$$

For sequences of local alternatives $\{P_n\}$ satisfying $d^*(P_n) = \delta n^{-1/2}$ for any given $\delta \in \mathbb{R}\backslash\{0\}$ and $\sigma^2 := \lim_{n\to\infty}\sigma^2(P_n) > 0$,

$$P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) = \Phi\left(z_{\alpha/2}+\frac{\delta}{\sigma}\right) + \Phi\left(z_{\alpha/2}-\frac{\delta}{\sigma}\right) - C_{PL}(\delta)\varepsilon_n^2 + O\left(\varepsilon_n^3\right) + O\left(n^{-1/2}\sqrt{\ln n}\right), \tag{3}$$

where

$$C_{PL}(\delta) := \left(\phi\left(z_{\alpha/2}-\frac{\delta}{\sigma}\right) - \phi\left(z_{\alpha/2}+\frac{\delta}{\sigma}\right)\right)\frac{\delta(\sigma^2 - 2(\sigma_A^2+\sigma_B^2))}{8\sigma^3}.$$

The expansions of size and power in Theorem 4 are useful for the optimal choice of $\hat{\varepsilon}_n$ that jointly minimizes size distortion for equivalent models,

$$\begin{aligned} SD_n &:= P_0(|\tilde{t}_n| > z_{1-\alpha/2}) - \alpha \\ &= C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n + \text{remainder} \end{aligned}$$

and power loss for distinct models at alternative δ,

$$\begin{aligned} PL_n(\delta) &:= \Phi\left(z_{\alpha/2} + \frac{\delta}{\sigma}\right) + \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) - P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) \\ &= C_{PL}(\delta)\varepsilon_n^2 + \text{remainder} \end{aligned}$$

The theorem shows that size for equivalent models is decreasing in ε_n and power for distinct models is increasing in ε_n. Therefore, SD_n and $PL_n(\delta)$ converge to zero at the fastest possible rate if their respective leading terms, $\varepsilon_n^{-1}n^{-1/2}\ln\ln n$ and ε_n^2, are of the same order. This is the case when ε_n is of the order $n^{-1/6}(\ln\ln n)^{1/3}$. In fact, we can also choose the constant in front of the optimal rate $n^{-1/6}(\ln\ln n)^{1/3}$ by balancing the constants in the leading terms of SD_n and $PL_n(\delta)$. In principle, we could set $C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n$ equal to $C_{PL}(\delta)\varepsilon_n^2$ and solve for the balancing ε_n given any particular alternative δ. Alternatively, we can define a loss function over alternatives δ, e.g. weighted average power loss $WAPL_n := \varepsilon_n^2 \int C_{PL}(\delta)\omega(\delta)d\delta$ for some weighting function $\omega(\delta)$ or the worst-case power loss $WCPL_n := \varepsilon_n^2 \sup_{\delta\in\mathbb{R}\setminus\{0\}} C_{PL}(\delta)$, then set it equal to the leading term of SD_n and solve for the balancing ε_n. Weighted average power $WAPL_n$ is easy to compute for certain weight functions such as the normal density, leading to closed form solutions of the resulting optimal tuning parameter. The worst-case power $WCPL_n$ is attractive because it does not require the choice of a weighting function, but the optimization over δ typically does not lead to a closed-form solution for the resulting optimal tuning parameter. We therefore propose a simple upper bound on the worst-case power loss that does possess a closed-form solution and worked well in our simulations, viz. $C_{PL}^*\varepsilon_n^2$ where

$$C_{PL}^* := \phi\left(z_{\alpha/2} - \frac{\delta^*}{\sigma}\right)\frac{\delta^*(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{4\sigma^3}$$

with

$$\delta^* := \frac{\sigma}{2}\left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2}\right).$$

Solving $C_{SD}\varepsilon_n^{-1}n^{-1/2}\ln\ln n = C_{PL}^*\varepsilon_n^2$ then yields

$$\varepsilon_n = \left(\frac{C_{SD}}{C_{PL}^*}\right)^{1/3} n^{-1/6}(\ln\ln n)^{1/3}.$$

This tuning parameter choice balances our upper bound on power loss with the size distortion and can be implemented in practice by computing

$$\hat{\varepsilon}_n = \left(\frac{\hat{C}_{SD}}{\hat{C}_{PL}^*}\right)^{1/3} n^{-1/6}(\ln\ln n)^{1/3} \tag{4}$$

with

$$\hat{C}_{PL}^* := \phi\left(z_{\alpha/2} - \frac{\hat{\delta}^*}{\hat{\sigma}}\right)\frac{\hat{\delta}^*(\hat{\sigma}^2 - 2(\hat{\sigma}_A^2 + \hat{\sigma}_B^2))}{4\hat{\sigma}^3}$$

$$\hat{C}_{SD} := 2\phi(z_{\alpha/2})\frac{\max\{|tr(\hat{H}_A^{-1}\hat{V}_A)|, |tr(\hat{H}_B^{-1}\hat{V}_B)|\}}{\sqrt{(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)/2}}$$

$$\hat{\delta}^* := \frac{\hat{\sigma}}{2}\left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2}\right)$$

and where $\hat{H}_k$ and $\hat{V}_k$, $k = A, B$, are estimates of $H_k := H_k(P_0)$ and $V_k := V_k(P_0)$ with $V_k(P) := E_P\left[\nabla_{\theta_k}\ln f_k\left(X_i, \theta_k^*(P)\right)\left(\nabla_{\theta_k}\ln f_k\left(X_i, \theta_k^*(P)\right)\right)'\right]$, obtained by replacing expectations by sample averages.

The proposed value of $\hat{\varepsilon}_n$ in (4) can easily be computed from the data as it requires only estimates of the matrices H_k and V_k, which have to be computed for the "sandwich" variance estimator for potentially misspecified models anyway, and the sample variances $\hat{\sigma}$, $\hat{\sigma}_A^2$ and $\hat{\sigma}_B^2$.

The following corollary formalizes the above discussion.

Corollary 1. *Suppose the conditions of Theorem 4 hold and $\hat{\varepsilon}_n$ is defined as in* (4)*. Then, for any distribution P_0 satisfying the null hypothesis, i.e. $d^*(P_0) = 0$ and $\sigma^2(P_0) = 0$,*

$$SD_n \leq \left(\frac{C_{SD}^2}{C_{PL}^*}\right)^{1/3} n^{-1/3}(\ln\ln n)^{2/3} + o(n^{-1/3}(\ln\ln n)^{2/3})$$

For sequences of local alternatives $\{P_n\}$ *satisfying* $d^*(P_n) = \delta n^{-1/2}$ *for any* $\delta \in \mathbb{R} \setminus \{0\}$ *and* $\sigma^2 := \lim_{n\to\infty} \sigma^2(P_n) > 0$,

$$PL_n(\delta) \leq \left(\frac{C_{SD}^2}{C_{PL}^*}\right)^{1/3} n^{-1/3}(\ln\ln n)^{2/3} + o\left(n^{-1/3}(\ln\ln n)^{2/3}\right).$$

Moreover, $\hat{\varepsilon}_n$ *satisfies Assumption 5, and Assumption 6 with* $\mathcal{P}$ *replaced by the set of distributions satisfying the assumptions of Theorem 4.*

Remark 1. *Theorem 4 verifies that the optimal epsilon* (4) *satisfies Assumptions 5 and 6, implying that all theorems in the previous sections hold with* $\hat{\varepsilon}_n$ *replaced by the optimal expression in* (4).

3 Invariance of Our Test Statistic Under Permutations

Let $I_{even,n}$ and $I_{odd,n}$ denote the even and odd numbers in $\{1, \ldots, n\}$, respectively. Our statistic can then be written as

$$\hat{\tilde{d}} = \hat{d} + \hat{\varepsilon}_n \left(\frac{1}{n}\sum_{i\in I_{odd,n}} \ln f_A(X_i;\hat{\theta}_A) - \frac{1}{n}\sum_{i\in I_{even,n}} \ln f_B(X_i;\hat{\theta}_B)\right).$$

Consider the "permuted" statistic

$$\hat{\tilde{\tilde{d}}} := \hat{d} + \hat{\varepsilon}_n \left(\frac{1}{n}\sum_{i\in I_{1,n}} \ln f_A(X_i;\hat{\theta}_A) - \frac{1}{n}\sum_{i\in I_{2,n}} \ln f_B(X_i;\hat{\theta}_B)\right)$$

where $I_{1,n}$ and $I_{2,n}$ form some partition of $\{1, \ldots, n\}$, each containing $n/2$ elements. Let $\tilde{t}_n := \sqrt{n}\hat{\tilde{d}}/\hat{\tilde{\sigma}}$ and $\tilde{\tilde{t}}_n := \sqrt{n}\hat{\tilde{\tilde{d}}}/\hat{\tilde{\sigma}}$ be the two corresponding t-statistics, and denote by $\#A$ the number of elements in a set A.

Lemma 1. *Suppose Assumptions 2, 3, and 1–5 hold. If* $\#(I_{odd,n} \setminus I_{1,n}) = o(n)$*, then*

$$\left|\tilde{t}_n - \tilde{\tilde{t}}_n\right| = o_{P_0}(1).$$

Lemma 1 shows that not only does every partition of the sample into two groups lead to the same asymptotic distribution, but also the random difference between two test statistics based on different assignment rules is negligible in large samples. This result requires that one partition into two groups can be constructed from the other partition by $o(n)$ re-assignments of observations across groups.

Remark 2. *It is easy to see that both statistics,* $\tilde{t}_n$ *and* $\tilde{\tilde{t}}_n$*, are asymptotically* $N(0,1)$*. However, if the difference* $I_{odd,n} \setminus I_{1,n}$ *is unrestricted, then they are not asymptotically equivalent in the sense that* $|\tilde{t}_n - \tilde{\tilde{t}}_n| = o_{P_0}(1)$*. Suppose this were true, then we would have*

$$\frac{1}{2}\left(\tilde{t}_n + \tilde{\tilde{t}}_n\right) = \frac{1}{2}\left(2\tilde{t}_n + \left[\tilde{\tilde{t}}_n - \tilde{t}_n\right]\right) = \tilde{t}_n + o_{P_0}(1) \to_d N(0,1).$$

Picking $I_{1,n} := I_{even,n}$ *and* $I_{2,n} := I_{odd,n}$*, however, yields*

$$\frac{1}{2}\left(\tilde{t}_n + \tilde{\tilde{t}}_n\right) = \frac{\sqrt{n}\left(1 + \frac{\varepsilon_n}{2}\right)\hat{d}}{\hat{\tilde{\sigma}}}$$

which is not asymptotically $N(0,1)$ *when the models are equivalent. Therefore, a restriction of how* $I_{odd,n} \setminus I_{1,n}$ *depends on* n *is important. In particular, the assumption* $\#(I_{odd,n} \setminus I_{1,n}) = o(n)$ *requires* $I_{odd,n}$ *to contain less than a fixed fraction of the sample that is not in* $I_{1,n}$*.*

4 Additional Simulations

In this section, we provide additional simulations to demonstrate that our test also performs well for selecting among two misspecified, two correctly specified and nested models. Typically, one can easily establish whether models are nested or not by inspection of the

two parametric families. When they are in fact nested, the standard likelihood ratio test with a chi-square critical value is the most powerful test under well-known conditions.

Example 1 (Misspecified Normals)**.** *Let the true distribution of the random variables X_i, $i = 1, \ldots, n$, be $N(\mu, 5)$. The two parametric families to be compared are*

$$\mathcal{P}_A := \{N(\mu_A, 1) : \ \mu_A \in \Theta_A\}$$
$$\mathcal{P}_B := \{N(0, \sigma_B^2) : \ \sigma_B \in \Theta_B\}$$

The null and alternative models are generated by varying the true mean according to $\mu = \sqrt{e^{2d+4} - 5}$ with $d \in [-1, 1]$. Both models are misspecified under the null ($\mu_A^* = \sqrt{e^4 - 5}$ and $\sigma_B^* = e^2$) and the alternatives. With Θ_A not containing the origin, the two models are non-overlapping.

Example 2 (Correctly Specified Normals)**.** *Let the true distribution of the random variables X_i, $i = 1, \ldots, n$, be $N(\mu, \sigma^2)$ and the two parametric families to be compared as in the previous example. The null and alternative models are generated by varying (μ, σ^2) according to $\mu = \sqrt{e^{2d-1+\sigma^2} - \sigma^2}$ with $\sigma^2 \in [1, 5]$ and $d \in [-1, 1]$. The two models are correctly specified under the null ($\mu_A = \mu = 0$, $\sigma_B = \sigma = 1$), illustrating the case in which the two models overlap at the truth and thus are observationally equivalent under the null. Under the alternatives, they are both misspecified.*

Example 3 (Nested Regressions with one Additional Regressor)**.** *Let the random vector (Y_i, W_i, Z_i), $i = 1, \ldots, n$, satisfy the regression equation*

$$Y_i = W_i + \tau W_i Z_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 1)$$

with $W_i \sim N(3, 1)$, $Z_i \sim N(0, 1)$ and $\varepsilon_i \sim N(0, 1)$ all i.i.d. and mutually independent random variables. Consider model A,

$$Y_i = \alpha_1 + \alpha_2 W_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_A^2),$$

and model B,

$$Y_i = \beta_1 + \beta_2 W_i + \beta_3 Z_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_B^2).$$

Null and alternative models are generated by varying τ over $[0, 1.6]$. Under the null ($\tau = 0$), both models are correctly specified and model B nests model A while, under the alternatives, both are misspecified.

Example 4 (Nested Regressions with two Additional Regressors)**.** *This example is similar to the previous one, except that model B has one more regressor, viz.*

$$Y_i = \beta_1 + \beta_2 W_i + \beta_3 Z_i + \beta_4 Z_i^2 + \varepsilon_i, \qquad \varepsilon_i \sim N(0, \sigma_B^2),$$

and the alternatives are generated from within model B:

$$Y_i = W_i + \tau Z_i + \varepsilon_i, \qquad \varepsilon_i \sim N(0, 1).$$

Therefore, the two models are nested, correctly specified under the null and the larger model is correctly specified even under the alternatives. This is the standard testing situation in which the second step of Vuong's procedure is equivalent to a Neyman Pearson ("NP") test of the hypothesis $H_0 : \beta_3 = \beta_4 = 0$.

Figure 1 shows the power plots for the four examples. The lower two panels of Table 7 report the empirical rejection probabilities under the null. In both examples, compared to Vuong's and Shi's test, our test is more powerful for alternatives close to the null whereas the other two dominate for alternatives further away from the null. All three tests control size reasonably well, with Vuong's and Shi's test almost not rejecting under the null at all. All tests perform well in the examples of misspecified and the correctly specified normals. In those examples, they control size and all possess similar power curves.

Finally, we also report size-corrected versions of the power curves in the main text; see Figure 2. To produce these graphs we first simulated Example 1 and searched for the

nominal level of the tests that make the finite sample rejection probability (under the null) equal to the desired level 0.05. For example, in panel (c), the levels required by Shi's and Vuong's test to reach a finite sample rejection rate of 0.05 are 0.27 and 0.17, respectively. Such large necessary levels reflect the conservative nature of the two tests under the null. Notice that in practice achieving these improved power curves is infeasible so this is really a theoretical exercise.

5 Extensions

To simplify the presentation of our basic model selection procedure we restrict attention to a simple and stylized framework: we compare two fully specified parametric models based on the KL criterion, i.i.d. data and a t-statistic. In this section, we argue that our procedure applies much more generally and discuss some important, but mostly straightforward, extensions.

Our model selection test measures distance between the candidate models by KL distance. One could, however, consider other goodness-of-fit criteria such as in-sample or out-of-sample fit rather than KL-distance. Rivers and Vuong (2002) propose such extensions of the Vuong test which would be completely analogous in our setting. An important example would be comparing the accuracy of competing forecasts. Consider two forecasts $\{y_{(1)t}\}_{t=1}^T$ and $\{y_{(2)t}\}_{t=1}^T$ of $\{y_t\}_{t=1}^T$ and let $\{e_{(k)t}\}_{t=1}^T$, $k = 1, 2$, be the corresponding forecast errors. In an influential paper, Diebold and Mariano (1995) discuss procedures for testing the hypothesis that the two forecasts are equally accurate, viz.

$$H_0 : \; Eg(e_{(1)t}) = Eg(e_{(2)t})$$

versus the alternative that the expectations are not equal, where g is some given loss function. Diebold and Mariano (1995) consider a test statistic $\bar{d} := T^{-1/2} \sum_{t=1}^T [g(e_{(1)t}) -$

$g(e_{(2)t})]$ which is asymptotically $N(0, \sigma^2)$ under standard assumptions. Therefore, we can test H_0 by simply comparing $\bar{d}$ to a normal critical value. In this setting, we can apply our sample splitting scheme to obtain a test that is asymptotically uniformly of correct level, i.e. consider the modified statistic

$$\tilde{\bar{d}} := \frac{T^{-1/2} \sum_{t=1}^{T} \left[\omega_t(\hat{\varepsilon}_T) g(e_{(1)t}) - \omega_{t+1}(\hat{\varepsilon}_T) g(e_{(2)t})\right]}{\sqrt{(1+\hat{\varepsilon}_T)\hat{\sigma}^2 + \hat{\varepsilon}_T^2(\hat{\sigma}_1^2 + \hat{\sigma}_2^2)/2}},$$

where $\hat{\sigma}^2$, $\hat{\sigma}_1^2$, and $\hat{\sigma}_2^2$ are estimators of σ^2, and the asymptotic variances of $T^{-1/2} \sum_{t=1}^{T} g(e_{(1)t})$ and $T^{-1/2} \sum_{t=1}^{T} g(e_{(2)t})$, respectively.

A useful extension of theorems relaxes the i.i.d. assumption on the data generating process. In the case of comparing parametric likelihoods, our theory allows for conditional densities, so that time series dependence over a finite number of lags (e.g. AR(p)) can be accommodated simply by conditioning on the lagged variables. More generally, the limiting distribution of our test statistic ultimately only depends on the asymptotic normality of certain sample averages and it is clear that our results can easily be secured under a much wider range of conditions, including general stationary time series data.

Our testing procedure is based on estimating parameters from moment conditions. For simplicity of exposition we considered a Z-estimator which is simply the root of the empirical estimating equations. Clearly one could use any estimation procedure that estimates solutions to moment conditions. Our procedure requires only asymptotic normality of the resulting estimator which is readily established for a wide range of estimators (e.g. generalized method of moments (GMM), generalized empirical likelihood (GEL), minimum distance) using standard conditions available in the literature (see, for example, Hansen (1982), Newey and McFadden (1994), Newey and Smith (2004) and van der Vaart (1998)). Also, test statistics for testing $H_0 : d^* = 0$ other than the t-statistic can be used, e.g. a Wald, Lagrange Multiplier or distance metric statistic. These are first-order asymptotically

equivalent to our statistic under standard conditions.

In the present context, M-estimators are also attractive because terms can be added to the criterion function in order to penalize certain types of models. For example, one may want to avoid the selection of models with too many parameters and add a correction term that is increasing in the number of parameters in a model. See, for instance, Vuong (1989, p. 318), Sin and White (1996) and references therein for correction terms that can be interpreted through information criteria such as AIC and BIC.

Interestingly, our method can also be extended to compare models defined by moment conditions rather than parametric likelihoods. In that case, one would replace the parametric scores $E_{P_0}[\nabla_{\theta_A} \ln f_A(X;\theta_A^*)] = 0$ and $E_{P_0}[\nabla_{\theta_B} \ln f_B(X;\theta_B^*)] = 0$ by the first-order derivatives of an empirical likelihood objective function and the KL-difference between the parametric densities by the difference in the respective objective functions. Other GEL objective functions could be used as well with the small difference being that they minimize divergence measures other than KL and so one may want to adjust our third moment condition accordingly. Notice, however, that comparisons based on GMM objective functions depend on the chosen weighting matrix and can, therefore, be very misleading (Hall and Pelletier (2011)).

We propose a regularization scheme which, in the observationally equivalent case, splits consecutive observations into two subsamples. The sample could, of course, be split in other ways as well. For example, one could consider the following reweighting scheme:

$$\hat{\tilde{d}} := \frac{1}{n}\sum_{i=1}^{n}\Big((1+\varepsilon_{i,n})\ln f_A(X_i;\hat{\theta}_A) - (1-\varepsilon_{i,n})\ln f_B(X_i;\hat{\theta}_B)\Big)$$

where $\varepsilon_{i,n}$ is an i.i.d. random variable independent of the sample and with a variance that shrinks to zero with the sample size n. This type of regularization does not assign special status to any observation, but on the other hand introduces more randomness, thereby

reducing the power of the test. One could also deviate from our proposed even/odd splitting scheme and our procedure would work in the exact same way as discussed above. However, splitting into two halves is optimal in the sense that it minimizes the sum of the variances arising from the two half-samples. Furthermore, one can imagine splitting up the sample in many different ways and averaging over the resulting test statistics, but this procedure would lead to a complicated limiting distribution due to the nontrivial correlations among the individual statistics.

6 Proofs

For $\theta = (\theta_A', \theta_B')'$, let $d_i(x;\theta,\varepsilon) := \omega_i(\varepsilon)\ln f_A(x;\theta_A) - \omega_{i+1}(\varepsilon)\ln f_B(x;\theta_B)$ and abbreviate $d_i(\theta,\varepsilon) := d_i(X_{i,n};\theta,\varepsilon)$. Define $\hat{G}(\theta) := \nabla_\theta \hat{g}(\theta)$ and $G(\theta) := E_{P_0}[\nabla_\theta g(X;\theta)]$.

Lemma 2. *Suppose $\{\varepsilon_n\} \in \mathcal{E}$. Then, under any sequence P_n in $\mathcal{P}$,*

1. $$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i(\theta^*(P_n),\varepsilon_n) - (1+\varepsilon_n/2)d^*(P_n)}{\tilde{\sigma}(P_n,\varepsilon_n)} \to_d N(0,1).$$

2. $$\frac{1}{n}\sum_{i=1}^{n}\left((\ln f_k(X_{i,n};\theta_k^*(P_n)))^2 - E_{P_n}[(\ln f_k(X_{i,n};\theta_k^*(P_n)))^2]\right) = O_{P_n}(n^{-1/2}).$$

3. $\hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2})$.

Proof. For the first part, we start by showing that the following Lyapounov condition holds: for some $\delta > 0$ as $n \to \infty$,

$$\sum_{i=1}^{n/2} E_{P_n}\left[\left|\frac{\Lambda_{2i-1}(P_n) + \Lambda_{2i}(P_n) + \varepsilon_n\Lambda_{2i,2i-1}(P_n) - (2+\varepsilon_n)d^*(P_n)}{\sqrt{n}\tilde{\sigma}(P_n,\varepsilon_n)}\right|^{2+\delta}\right] \to 0, \qquad (5)$$

where $\Lambda_{i,j}(P) := \ln f_A(X_i;\theta^*(P)) - \ln f_B(X_j;\theta^*(P))$ and $\Lambda_i(P) := \Lambda_{i,i}(P)$. By the c_r-inequality,

$$\sum_{i=1}^{n/2} E_{P_n}\left[\left|\frac{\Lambda_{2i-1}(P_n)+\Lambda_{2i}(P_n)+\varepsilon_n\Lambda_{2i,2i-1}(P_n)-(2+\varepsilon_n)d^*(P_n)}{\sqrt{n}\tilde{\sigma}(P_n,\varepsilon_n)}\right|^{2+\delta}\right]$$
$$\leq \frac{2^{2+2\delta}}{n^{\delta/2}}\sum_{i=1}^{n/2} E_{P_n}\left[|Z_{2i-1,n}|^{2+\delta}+|Z_{2i,n}|^{2+\delta}+|Z_{i,n,split}|^{2+\delta}+\left|\frac{(2+\varepsilon_n)d^*(P_n)}{\tilde{\sigma}(P_n,\varepsilon_n)}\right|^{2+\delta}\right] \quad (6)$$

with $Z_{i,n} := \Lambda_i(P_n)/\tilde{\sigma}(P_n,\varepsilon_n)$ and $Z_{i,n,split} := \varepsilon_n\Lambda_{2i,2i-1}(P_n)/\tilde{\sigma}(P_n,\varepsilon_n)$. Consider the first of the four terms. If $\sigma(P_n) \geq \underline{c}$ for some $\underline{c} > 0$, then

$$\begin{aligned}
E_{P_n}\left[|Z_{i,n}|^{2+\delta}\right] &= E_{P_n}\left[\left|\frac{\ln f_A(X;\theta_A^*(P_n))-\ln f_B(X;\theta_B^*(P_n))}{\tilde{\sigma}(P_n,\varepsilon_n)}\right|^{2+\delta}\right] \\
&\leq E_{P_n}\left[\frac{|\ln f_A(X;\theta_A^*(P_n))-\ln f_B(X;\theta_B^*(P_n))|^{2+\delta}}{(1+\varepsilon_n)^{1+\delta/2}\sigma^{2+\delta}(P_n)}\right] \\
&\leq \frac{E_{P_n}\left[|D(X)|^{2+\delta}\right]\sigma^{2+\delta}(P_n)}{(1+\varepsilon_n)^{1+\delta/2}\sigma^{2+\delta}(P_n)} \\
&= (1+\varepsilon_n)^{-1-\delta/2}E_{P_n}\left[|D(X)|^{2+\delta}\right] \leq \overline{M}
\end{aligned}$$

where the first inequality follows from the fact that $\tilde{\sigma}^2(P,\varepsilon) = (1+\varepsilon)\sigma^2(P) + \varepsilon^2(\sigma_A^2(P) + \sigma_B^2(P))/2$ is larger than either $(1+\varepsilon)\sigma^2(P)$ or $\varepsilon^2(\sigma_A(P)+\sigma_B^2(P))/2$ (as $\varepsilon \geq 0$). The second inequality is implied by the dominance condition (1). Since $\overline{M}$ is independent of P_n, we have $\sup_{n\geq 1} E_{P_n}[|Z_{2i-1,n}|^{2+\delta}] \leq \overline{M}$, even if $\sigma(P_n) \to 0$ as $n \to \infty$. Therefore, the first and second expectation in (6) are finite uniformly over n.

Next, consider the third expectation in (6):

$$
\begin{aligned}
E_{P_n}\left[\left|Z_{i,n,split}\right|^{2+\delta}\right] &= E_{P_n}\left[\left|\frac{\varepsilon_n\left(\ln f_A(X_{2i};\theta_A^*(P_n)) - \ln f_B(X_{2i-1};\theta_B^*(P_n))\right)}{\tilde{\sigma}(P_n,\varepsilon_n)}\right|^{2+\delta}\right] \\
&\leq E_{P_n}\left[\left|\frac{\varepsilon_n\left(\ln f_A(X_{2i};\theta_A^*(P_n)) - \ln f_B(X_{2i-1};\theta_B^*(P_n))\right)}{\varepsilon_n\sqrt{(\sigma_A^2(P_n)+\sigma_B^2(P_n))/2}}\right|^{2+\delta}\right] \\
&= E_{P_n}\left[\left|\frac{\ln f_A(X_{2i};\theta_A^*(P_n)) - \ln f_B(X_{2i-1};\theta_B^*(P_n))}{\sqrt{(\sigma_A^2(P_n)+\sigma_B^2(P_n))/2}}\right|^{2+\delta}\right] \\
&\leq \underline{M}^{-1/2}2^{1+\delta}\left\{E_{P_n}\left[\left|\ln f_A(X_{2i};\theta_A^*(P_n))\right|^{2+\delta}\right] + E_{P_n}\left[\left|\ln f_B(X_{2i-1};\theta_B^*(P_n))\right|^{2+\delta}\right]\right\} \\
&\leq \underline{M}^{-1/2}2^{2+\delta}\overline{M}
\end{aligned}
$$

This bound is again valid uniformly over n.

Finally, by Lyapounov's Inequality, we have $(1+\varepsilon_n/2)d^*(P_n) \leq \tilde{\sigma}(P_n,\varepsilon_n)$, uniformly in n, so that the fourth expectation in (6) is also finite, uniformly in n. In conclusion, we have established (5). Lyapounov's Central Limit Theorem (e.g. Theorem 23.11 in Davidson (1994)) then implies that, under any sequence P_n in $\mathcal{P}$,

$$
\begin{aligned}
&\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{d_i(\theta^*(P_n),\varepsilon_n) - (1+\varepsilon_n/2)d^*(P_n)}{\tilde{\sigma}(P_n,\varepsilon_n)} \\
&\quad = \frac{1}{\sqrt{n}}\sum_{i=1}^{n/2}\frac{\Lambda_{2i-1}(P_n) + \Lambda_{2i}(P_n) + \varepsilon_n\Lambda_{2i,2i-1}(P_n) - (2+\varepsilon_n)d^*(P_n)}{\tilde{\sigma}(P_n,\varepsilon_n)} \to_d N(0,1).
\end{aligned}
$$

For the second part of the lemma, notice that

$$
E_P\left[\left|\frac{(\ln f_k(X;\theta_k^*(P)))^2 - E_P[(\ln f_k(X;\theta_k^*(P)))^2]}{Var_P((\ln f_k(X;\theta_k^*(P)))^2)^{1/2}}\right|^{2+\delta}\right] \leq \overline{M}\underline{M}^{-1}, \quad k = A, B, \tag{7}
$$

for all $P \in \mathcal{P}$ because $Var_P((\ln f_k(X;\theta_k^*(P)))^2)$ is bounded away from zero by the definition of $\mathcal{P}$ and because the numerator is bounded from above by $\overline{M}$. Therefore, we can apply

the Lyapounov Central Limit Theorem as in the first part of the proof and the result follows. The third part of the lemma can be proved in exactly the same fashion as the second. Q.E.D.

Lemma 3. *Let $X_{n,1}, \ldots, X_{n,n}$ be an i.i.d. sample from P_n and Assumption 2 hold. Suppose there exists a unique $\theta^*(P_n) \in \Theta$ such that $\lim_{n\to\infty} \theta^*(P_n) \in int(\Theta)$, $E_{P_n} g(X; \theta^*(P_n)) = 0$ and, for all $\kappa > 0$, there is an $\epsilon(\kappa) > 0$ such that*

$$\inf_{\theta : \|\theta - \theta^*(P_n)\| \geq \kappa} \|E_{P_n}[g(X_{n,i}; \theta)]\| > \epsilon(\kappa).$$

Further, assume the following conditions hold:

(i) $\hat{\varepsilon}_n$ is a sequence of measurable functions of $X_{n,1}, \ldots, X_{n,n}$ and there is a sequence $\{\varepsilon_n\}$ in $\mathcal{E}$ such that $|\hat{\varepsilon}_n - \varepsilon_n| = o_{P_n}(1)$.

(ii) For $h(x; \theta)$ being any of the functions $\ln f_k(x; \theta_k)$ and $\nabla \ln f_k(x; \theta_k)$, $k = A, B$, $\theta = (\theta'_A, \theta'_B)'$, we have

$$\sup_{\theta \in \Theta} \left\| \frac{1}{n} \sum_{i=1}^{n} h(X_i; \theta) - E_{P_n} h(X_i; \theta) \right\| = o_{P_n}(1).$$

Then, $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$ and $|\hat{d} - (1 + \varepsilon_n/2) d^*(P_n)| = o_{P_n}(1)$.

Proof. Let $\Psi_n(\theta) := E_{P_n}[g(X_{n,i}; \theta)]$. By assumption, for any $\kappa > 0$,

$$\inf_{\theta : \|\theta - \theta^*(P_n)\| \geq \kappa} \|\Psi_n(\theta)\| > \epsilon(\kappa) > 0.$$

The proof of $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$ therefore follows that of Theorem 5.9 in van der Vaart (1998). The second conclusion can be established as follows. A Taylor expansion around $(\theta^*(P_n), \varepsilon_n)$ yields

$$\hat{\tilde{d}} = \frac{1}{n} \sum_{i=1}^{n} d_i(\hat{\theta}, \hat{\varepsilon}_n) = \frac{1}{n} \sum_{i=1}^{n} d_i(\theta^*(P_n), \varepsilon_n) + \frac{1}{n} \sum_{i=1}^{n} \nabla_{(\varepsilon,\theta)} d_i(\bar{\theta}_n, \bar{\varepsilon}_n) \begin{pmatrix} \hat{\varepsilon}_n - \varepsilon_n \\ \hat{\theta} - \theta^*(P_n) \end{pmatrix} = 0 \quad (8)$$

where $(\bar{\theta}_n, \bar{\varepsilon}_n)$ lies on the line segment joining $(\hat{\theta}, \hat{\varepsilon}_n)$ and $(\theta^*(P_n), \varepsilon_n)$. By (ii), the triangle inequality and $\bar{\varepsilon}_n = O_{P_n}(1)$, we have $n^{-1}\sum_{i=1}^n \nabla_{(\varepsilon,\theta)} d_i(\bar{\theta}_n, \bar{\varepsilon}_n) = O_{P_n}(1)$, so that

$$\left| \hat{d} - \frac{1}{n}\sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) \right| = o_{P_n}(1) \tag{9}$$

follows from $\|\hat{\theta} - \theta^*(P_n)\| = o_{P_n}(1)$ and $|\hat{\varepsilon}_n - \varepsilon_n| = o_{P_n}(1)$. By (ii) and the triangle inequality, we also have

$$\left| \frac{1}{n}\sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) - \left(1 + \frac{\varepsilon_n}{2}\right) d^*(P_n) \right| = \left| \frac{1}{n}\sum_{i=1}^n d_i(\theta^*(P_n), \varepsilon_n) - E_{P_n} d_i(\theta^*(P_n), \varepsilon_n) \right| = o_{P_n}(1). \tag{10}$$

Together, (9) and (10) imply the second result. Q.E.D.

Lemma 4. *Let $X_{n,1}, \ldots, X_{n,n}$ be an i.i.d. sample from P_n and that the following conditions hold:*

(i) $\hat{\varepsilon}_n$ is a sequence of measurable functions of $X_{n,1}, \ldots, X_{n,n}$ such that there is a sequence $\{\varepsilon_n\}$ in $\mathcal{E}$ satisfying $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$.

(ii) For $h(x;\theta)$ being any of the functions $\ln f_k(X;\theta_k)$, $\ln f_k(x;\theta_k)\nabla \ln f_j(x;\theta_j)$, and $\nabla \ln f_k(X;\theta_k)$, $j,k = A, B$, $\theta = (\theta_A', \theta_B')'$, we have

$$\sup_{\theta\in\Theta} \left\| \frac{1}{n}\sum_{i=1}^n h(X_i;\theta) - E_{P_n} h(X_i;\theta) \right\| = o_{P_n}(1),$$

and

$$\frac{1}{n}\sum_{i=1}^n \left((\ln f_k(X_{i,n};\theta_k^*(P_n)))^2 - E_{P_n}[(\ln f_k(X_{i,n};\theta_k^*(P_n)))^2] \right) = O_{P_n}(n^{-1/2}).$$

(iii) $\|\hat{\theta} - \theta^*(P_n)\| = O_{P_n}(n^{-1/2})$,

(iv) There are constants $0 < \underline{M} \leq \overline{M} < \infty$ *such that* $\underline{M} \leq \sigma_k(P_n) \leq \overline{M}$ *for all* n *and* $k = A, B$.

Then, for $\hat{\tilde{\sigma}}^2 = \hat{\tilde{\sigma}}^2(\hat{\theta}, \hat{\varepsilon}_n)$,

$$\left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\tilde{\sigma}}^2} - 1 \right| \to_{P_n} 0.$$

Proof. First, we establish

$$\left|\hat{\sigma}^2 - \sigma^2(P_n)\right| = O_{P_n}(n^{-1/2}) \quad \text{and} \quad \left|\hat{\sigma}_k^2 - \sigma_k^2(P_n)\right| = O_{P_n}(n^{-1/2}), k = A, B. \tag{11}$$

Notice that by a Taylor expansion around $\theta^*(P_n)$, under P_n, we have

$$\left|\hat{\sigma}^2 - \hat{\sigma}^2(\theta^*(P_n))\right| \leq \left|\nabla_\theta \hat{\sigma}^2(\bar{\theta}_n)\left(\hat{\theta} - \theta^*(P_n)\right)\right| = O_{P_n}(n^{-1/2})$$

where $\bar{\theta}_n$ lies on the line segment joining $\hat{\theta}$ and $\theta^*(P_n)$. Uniform convergence of $\ln f_k(X; \theta_k)$, $\nabla \ln f_k(X; \theta_k)$ and $\ln f_k(x; \theta_k) \nabla \ln f_j(x; \theta_j)$, $j, k = A, B$, in (ii) together with the Cauchy-Schwartz inequality imply $\|\nabla_\theta \hat{\sigma}^2(\bar{\theta}_n)\| = O_{P_n}(1)$ so that the equality above follows from the consistency requirement in (iii). Similarly, $|\hat{\sigma}_k^2 - \hat{\sigma}_k^2(\theta_k^*(P_n))| = O_{P_n}(n^{-1/2})$ for $k = A, B$. By the second part of (ii) and the Hölder inequality, $|\hat{\sigma}_k^2(\theta^*(P_n)) - \sigma_k^2(P_n)| = O_{P_n}(n^{-1/2})$ for $k = A, B$, and the desired result (11) follows.

The remainder of the proof separately treats the two cases $\sigma^2(P_n) \to \sigma_\infty^2 > 0$ and $\sigma^2(P_n) \to 0$. First, consider $\sigma^2(P_n) \to \sigma_\infty^2 > 0$. In this case, by (iv) and the definition of $\mathcal{E}$, $\tilde{\sigma}^2(P_n, \varepsilon_n)$ also converges to a finite, nonzero constant. Thus, (11) and (i) directly yield $|\hat{\tilde{\sigma}}^2 - \tilde{\sigma}^2(P_n, \varepsilon_n)| = O_{P_n}(n^{-1/2})$ so that

$$\left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\tilde{\sigma}}^2} - 1 \right| = \left| \frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\tilde{\sigma}^2(P_n, \varepsilon_n) + O_{P_n}(n^{-1/2})} - 1 \right| = o_{P_n}(1).$$

Now, consider $\sigma^2(P_n) \to 0$. We further split this case into three subcases: (a) $\sigma^2(P_n)/\varepsilon_n^2 \to 0$ which means that either $\sigma^2(P_n)$ and ε_n^2 both vanish, but $\sigma^2(P_n)$ at a faster rate, or

that $\sigma^2(P_n)$ converges to zero at an arbitrary rate while ε_n^2 stays bounded away from zero; (b) $\sigma^2(P_n)/\varepsilon_n^2 \to \infty$, i.e. $\sigma^2(P_n)$ and ε_n^2 both vanish, but ε_n^2 at a faster rate; (c) $\sigma^2(P_n)/\varepsilon_n^2 \to c \neq 0$, i.e. both vanish at the same rate.

Consider subcase (a). By Assumption (i), we have

$$\frac{\hat{\varepsilon}_n}{\varepsilon_n} = 1 + \frac{\hat{\varepsilon}_n - \varepsilon_n}{\varepsilon_n} = 1 + O_{P_n}(n^{-1/2}\varepsilon_n^{-1}) = 1 + o_{P_n}(1).$$

Similarly, by (11),

$$\frac{\hat{\sigma}^2}{\varepsilon_n^2} = \frac{\sigma^2(P_n)}{\varepsilon_n^2} + \frac{\hat{\sigma}^2 - \sigma^2(P_n)}{\varepsilon_n^2} = o(1) + O_{P_n}(n^{-1/2}\varepsilon_n^{-2}) = o_{P_n}(1).$$

Therefore,

$$\begin{aligned}\frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\tilde{\sigma}}^2} &= \frac{(1+\varepsilon_n)\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n))}{(1+\hat{\varepsilon}_n)\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)} = \frac{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + O(\frac{\sigma^2(P_n)}{\varepsilon_n^2})}{\frac{\hat{\varepsilon}_n^2}{\varepsilon_n^2 2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2) + O_{P_n}(\frac{\hat{\sigma}^2}{\varepsilon_n^2})} \\ &= \frac{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n)) + o(1)}{\frac{1}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n) + o_{P_n}(1)) + o_{P_n}(1)} = 1 + o_{P_n}(1)\end{aligned}$$

In subcase (b), we use a similar reasoning as above to show that $\hat{\varepsilon}_n^2/\sigma^2(P_n) = o_{P_n}(1)$ and $\hat{\sigma}^2/\sigma^2(P_n) = 1 + o_{P_n}(1)$. Therefore,

$$\begin{aligned}\frac{\tilde{\sigma}^2(P_n, \varepsilon_n)}{\hat{\tilde{\sigma}}^2} &= \frac{(1+\varepsilon_n)\sigma^2(P_n) + \frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n) + \sigma_B^2(P_n))}{(1+\hat{\varepsilon}_n)\hat{\sigma}^2 + \frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)} = \frac{(1+\varepsilon_n) + O(\varepsilon_n^2/\sigma^2(P_n))}{(1+\hat{\varepsilon}_n)\frac{\hat{\sigma}^2}{\sigma^2(P_n)} + O_{P_n}(\hat{\varepsilon}_n^2/\sigma^2(P_n))} \\ &= \frac{1 + o(1)}{1 + o_{P_n}(1)} = 1 + o_{P_n}(1)\end{aligned}$$

In subcase (c), we also have $\hat{\sigma}^2/\sigma^2(P_n) = 1 + o_{P_n}(1)$ and $\hat{\varepsilon}_n^2/\sigma^2(P_n) = \varepsilon_n^2/\sigma^2(P_n) + o_{P_n}(1)$

so that

$$\begin{aligned}
\frac{\tilde{\sigma}^2(P_n,\varepsilon_n)}{\hat{\tilde{\sigma}}^2} &= \frac{(1+\varepsilon_n)\sigma^2(P_n)+\frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n)+\sigma_B^2(P_n))}{(1+\hat{\varepsilon}_n)\hat{\sigma}^2+\frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2+\hat{\sigma}_B^2)} \\
&= \frac{\sigma^2(P_n)+\frac{\varepsilon_n^2}{2}(\sigma_A^2(P_n)+\sigma_B^2(P_n))+o(\varepsilon_n^2)}{\hat{\sigma}^2+\frac{\hat{\varepsilon}_n^2}{2}(\hat{\sigma}_A^2+\hat{\sigma}_B^2)+O_{P_n}(\hat{\varepsilon}_n\hat{\sigma}^2)} \\
&= \frac{1+\frac{\varepsilon_n^2}{2\sigma^2(P_n)}(\sigma_A^2(P_n)+\sigma_B^2(P_n))+o(\varepsilon_n^2/\sigma^2(P_n))}{\frac{\hat{\sigma}^2}{\sigma^2(P_n)}+\frac{\hat{\varepsilon}_n^2}{2\sigma^2(P_n)}(\sigma_A^2(P_n)+\sigma_B^2(P_n))+o_{P_n}(\varepsilon_n^2/\sigma^2(P_n))} \\
&= 1+o_{P_n}(1)
\end{aligned}$$

which uses the fact that $O_{P_n}(\hat{\varepsilon}_n\hat{\sigma}^2)=O_{P_n}(\varepsilon_n(\sigma^2(P_n)+n^{-1/2}))=o_{P_n}(\varepsilon_n^2)$. Q.E.D.

Lemma 5. *Suppose Assumption 2 holds. Let $\hat{\varepsilon}_n$ be a sequence of real-valued, measurable functions of the triangular array $X_{n,1},\ldots,X_{n,n}$, an i.i.d. sample from P_n, and $\mathcal{Q}$ be some subset of $\mathcal{P}$. Assume that, for every sequence $\{P_n\}$ in $\mathcal{Q}$, there is a sequence $\{\varepsilon_n\}\in\mathcal{E}$ with $|\hat{\varepsilon}_n-\varepsilon_n|=O_{P_n}(n^{-1/2})$. Let $\bar{\delta}\in[-\infty,+\infty]$ be such that $\sqrt{n}\,d^*(P_n)(1+\varepsilon_n/2)/\tilde{\sigma}(P_n,\varepsilon_n)\to\bar{\delta}$. Then, under any sequence $\{P_n\}$ in $\mathcal{Q}$, if $|\bar{\delta}|<\infty$,*

$$\frac{\sqrt{n}\hat{\tilde{d}}}{\hat{\tilde{\sigma}}}\to_d N(\bar{\delta},1).$$

If $|\bar{\delta}|=\infty$, then $|\sqrt{n}\hat{\tilde{d}}/\hat{\tilde{\sigma}}|\to_{P_n}\infty$.

Proof. Suppose $|\bar{\delta}|<\infty$. First, we establish two auxiliary results, viz. the orders of $\tilde{\sigma}(P_n,\varepsilon_n)^{-1}$ and $\hat{\theta}-\theta^*(P_n)$. To that end, consider two cases: (a) P_n approaches the observationally equivalent case, i.e. $\sigma(P_n)\to 0$; (b) P_n satisfies $\sigma(P_n)\to c\neq 0$. In the first case, since by part (iv) of Definition 1, $\sigma_k^2(P_n)$ is bounded away from zero and $n^{1/4}\varepsilon_n\to\infty$,

$$n\tilde{\sigma}^2(P_n,\varepsilon_n)=n(1+\varepsilon_n)\sigma^2(P_n)+n\varepsilon_n^2(\sigma_A^2(P_n)+\sigma_B^2(P_n))/2\to\infty$$

so that $\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = o(n^{1/2})$. In the second case, $\tilde{\sigma}(P_n, \varepsilon_n) \to c \neq 0$ so that $\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = O(1) = o(n^{1/2})$. In conclusion,

$$\tilde{\sigma}(P_n, \varepsilon_n)^{-1} = o(n^{1/2}). \tag{12}$$

Next, consider the order of $\hat{\theta} - \theta^*(P_n)$. A Taylor expansion with $\bar{\theta}$ on the line segment joining $\hat{\theta}$ and $\theta^*(P_n)$ yields $\hat{\theta} - \theta^*(P_n) = -\hat{G}(\bar{\theta})^{-1}\hat{g}(\theta^*(P_n))$. By Assumption 2, parts (i) and (iii) of Definition 1, and Lemma 2.4 of Newey and McFadden (1994), $\hat{G}(\theta)$ converges in probability, under P_n, uniformly over Θ. Part (i) and (iii) of Definition 1 together with Assumption 2 imply Assumption (ii) of Lemma 3, so that we can use it to obtain consistency of $\hat{\theta}$ and $\bar{\theta}$ under P_n. Therefore, letting $G_P(\theta) := E_P[\nabla_\theta g(X;\theta)]$ and $G_n := G_{P_n}(\theta^*(P_n))$, we have

$$\begin{aligned}\left\|\hat{G}(\bar{\theta}) - G_n\right\| &\leq \left\|\hat{G}(\bar{\theta}) - G_{P_n}(\bar{\theta})\right\| + \left\|G_{P_n}(\bar{\theta}) - G_n\right\| \\ &\leq \sup_{\theta\in\Theta}\left\|\hat{G}(\theta) - G_{P_n}(\theta)\right\| + o_{P_n}(1) = o_{P_n}(1).\end{aligned}$$

Furthermore, by (iv) of Definition 1, $\hat{G}(\bar{\theta})$ is invertible with probability approaching one, under P_n. By part 3. of Lemma 2, $\hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2})$, so that, in conclusion,

$$\hat{\theta} - \theta^*(P_n) = -\hat{G}(\bar{\theta})^{-1}\hat{g}(\theta^*(P_n)) = O_{P_n}(n^{-1/2}). \tag{13}$$

With the auxilliary results established, we now consider the following decomposition:

$$\frac{\sqrt{n}\hat{\tilde{d}}}{\tilde{\sigma}(P_n, \varepsilon_n)} = \frac{\sqrt{n}d^*(P_n)(1 + \frac{\hat{\varepsilon}_n}{2})}{\tilde{\sigma}(P_n, \varepsilon_n)} + \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(d_i(\hat{\theta}, \hat{\varepsilon}_n) - d^*(P_n)(1 + \frac{\hat{\varepsilon}_n}{2})\right)}{\tilde{\sigma}(P_n, \varepsilon_n)}.$$

The assumption $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ and a Taylor expansion of $d_i(\hat{\theta}, \hat{\varepsilon}_n) - d^*(P_n)(1+\hat{\varepsilon}_n/2)$

around $(\theta^*(P_n), \varepsilon_n)$ yield

$$\begin{aligned}\frac{\sqrt{n}\hat{\bar{d}}}{\tilde{\sigma}(P_n,\varepsilon_n)} &= \bar{\delta} + o_{P_n}(1) + \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(d_i(\theta^*(P_n),\varepsilon_n) - d^*(P_n)(1+\frac{\varepsilon_n}{2})\right)}{\tilde{\sigma}(P_n,\varepsilon_n)} \\ &\quad + \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta d_i(\theta^*(P_n),\varepsilon_n)(\hat{\theta}-\theta^*(P_n))}{\tilde{\sigma}(P_n,\varepsilon_n)} \\ &\quad + \frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \left(\nabla_\varepsilon d_i(\theta^*(P_n),\varepsilon_n) - \frac{1}{2}d^*(P_n)\right)(\hat{\varepsilon}_n - \varepsilon_n)}{\tilde{\sigma}(P_n,\varepsilon_n)} + R_n \end{aligned} \tag{14}$$

where, for some $(\bar{\theta}_n, \bar{\varepsilon}_n)$ on the line segment joining $(\hat{\theta}, \hat{\varepsilon}_n)$ and $(\theta^*(P_n), \varepsilon_n)$,

$$\begin{aligned}|R_n| &\le \sqrt{n}\tilde{\sigma}(P_n,\varepsilon_n)^{-1}\left\|\frac{1}{n}\sum_{i=1}^n \nabla_\theta^2 d_i(\bar{\theta}_n,\bar{\varepsilon}_n)\right\| \left\|\hat{\theta}-\theta^*(P_n)\right\|^2 \\ &\quad + \sqrt{n}\tilde{\sigma}(P_n,\varepsilon_n)^{-1}\left|\frac{1}{n}\sum_{i=1}^n \nabla_\varepsilon^2 d_i(\bar{\theta}_n,\bar{\varepsilon}_n)\right| \|\hat{\varepsilon}_n - \varepsilon_n\|^2 \\ &= \sqrt{n}o(n^{1/2})O_{P_n}(1)O_{P_n}(n^{-1}) + 0 = o_{P_n}(1).\end{aligned}$$

The first equality holds for the following reason. By Assumption 2, parts (i) and (iii) of Definition 1, and Lemma 2.4 of Newey and McFadden (1994), $\|n^{-1}\sum_{i=1}^n \nabla_\theta^2 \ln f_k(X_{n,i};\theta)\|$, $k = A, B$, converges in probability, under P_n, uniformly over Θ. By the triangle inequality and the fact that $\hat{\varepsilon}_n = O_{P_n}(1)$, and thus $\bar{\varepsilon}_n = O_{P_n}(1)$, we also have $\|n^{-1}\sum_{i=1}^n \nabla_\theta^2 d_i(\bar{\theta}_n, \bar{\varepsilon}_n)\| = O_{P_n}(1)$. (12), (13), and the assumption $|\hat{\varepsilon}_n - \varepsilon_n| = O_{P_n}(n^{-1/2})$ then imply the equality.

We now separately consider each of the remaining three terms in (14). By part 1. of Lemma 2, the first term is asymptotically $N(0,1)$ under P_n. For the second term, notice that $n^{-1}\sum_{i=1}^n \nabla_\theta d_i(\theta^*(P_n),\varepsilon_n)$ is a linear transformation of $\hat{g}(\theta^*(P_n))$ and, thus by part 3. of Lemma 2 and $\varepsilon_n = O(1)$, $O_{P_n}(n^{-1/2})$. Therefore, (12) and (13) imply

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^n \nabla_\theta d_i(\theta^*(P_n),\varepsilon_n)(\hat{\theta}-\theta^*(P_n))}{\tilde{\sigma}(P_n,\varepsilon_n)} = O_{P_n}(1)O_{P_n}(n^{-1/2})o(n^{1/2}) = o_{P_n}(1).$$

In the third term,

$$\frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\varepsilon}d_i(\theta^*(P_n),\varepsilon_n)-\frac{d^*(P_n)}{2}\right)$$
$$=\frac{1}{n}\sum_{i=1}^{n}\left(\ln f_A(X_{n,2i-1};\theta_A^*(P_n))-\ln f_B(X_{n,2i};\theta_B^*(P_n))-\frac{d^*(P_n)}{2}\right)=O_{P_n}(n^{-1/2})$$

by a similar argument as in part 1. of Lemma 2, so that

$$\frac{\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\left(\nabla_{\varepsilon}d_i(\theta^*(P_n),\varepsilon_n)-\frac{1}{2}d^*(P_n)\right)(\hat{\varepsilon}_n-\varepsilon_n)}{\tilde{\sigma}(P_n,\varepsilon_n)}=O_{P_n}(1)O_{P_n}(n^{-1/2})o(n^{1/2})=o_{P_n}(1).$$

In conclusion, $\sqrt{n}\hat{\bar{d}}/\tilde{\sigma}(P_n,\varepsilon_n)\to_d N(\bar{\delta},1)$ under P_n. The corresponding result with the estimated standard deviation, $\hat{\bar{\sigma}}$, in the denominator rather than $\tilde{\sigma}(P_n,\varepsilon_n)$ follows from Lemma 4, using (13). The case $|\bar{\delta}|=\infty$ follows from a similar argument. Q.E.D.

Proof of Theorem 1. We show the result by applying Lemma 5. Let $\mathcal{Q}=\{P_0\}$ and $\bar{\delta}=0$. Part (i) of Definition 1 holds by Assumption 3(ii). Part (iii) by Assumption 4(ii). Finally, part (iv) of Definition 1 holds because of Assumptions 3(iii) and 1, and Assumption 5 implies Assumption 6. The uniform moment bounds in (ii) of Definition 1 hold because of Assumption 4(i).

It remains to show that the dominance condition (1) in (ii) of Definition 1 holds. This can be seen as follows. In the non-overlapping case, $\sigma^2>0$, (1) is implied by Assumption 4(i). In the overlapping case, the information matrix equality holds, so that $Var_{P_0}(\nabla_{\theta_k}\ln f_k(X;\theta_k^*))=-E_{P_0}[\nabla^2_{\theta_k}\ln f_k(X;\theta_k^*)]$, $k=A,B$, is invertible by Assumption 3(iii). Let λ_{min} be the minimum of the eigenvalues of both matrices and note that it must be strictly larger than zero. Then it is easy to show that (1) holds for $D(x):=\sqrt{2}\bar{F}_2(x)/\lambda_{min}$ because of Assumption 4(iii). Q.E.D.

Proof of Lemma 1. First, notice that

$$\left|\tilde{t}_n - \tilde{\tilde{t}}_n\right| = \frac{\hat{\varepsilon}_n}{\sqrt{n}\hat{\tilde{\sigma}}}\left(\sum_{i\in I_{odd,n}\setminus I_{1,n}} \ln f_A(X_i;\hat{\theta}_A) - \sum_{i\in I_{even,n}\setminus I_{2,n}} \ln f_B(X_i;\hat{\theta}_B)\right.$$

$$\left. + \sum_{i\in I_{2,n}\setminus I_{even,n}} \ln f_B(X_i;\hat{\theta}_B) - \sum_{i\in I_{1,n}\setminus I_{odd,n}} \ln f_A(X_i;\hat{\theta}_A)\right)$$

$$= \frac{\hat{\varepsilon}_n}{\sqrt{n}\hat{\tilde{\sigma}}}\left(\sum_{i\in I_{odd,n}\setminus I_{1,n}} \Big(\ln f_A(X_i;\hat{\theta}_A) + \ln f_B(X_i;\hat{\theta}_B)\Big)\right.$$

$$\left. - \sum_{i\in I_{even,n}\setminus I_{2,n}} \Big(\ln f_A(X_i;\hat{\theta}_A) + \ln f_B(X_i;\hat{\theta}_B)\Big)\right)$$

because $I_{even,n} \setminus I_{2,n} = I_{1,n} \setminus I_{odd,n}$ and $I_{odd,n} \setminus I_{1,n} = I_{2,n} \setminus I_{even,n}$. In the overlapping case,

$$\frac{\hat{\varepsilon}_n}{\hat{\tilde{\sigma}}} = \frac{1}{\sqrt{(1+\hat{\varepsilon}_n)\frac{\hat{\sigma}}{\hat{\varepsilon}_n^2} + \frac{1}{2}(\sigma_A^2 + \sigma_B^2 + o_{P_0}(1))}} = \frac{1}{\frac{1}{2}(\sigma_A^2 + \sigma_B^2)} + o_{P_0}(1) = O_{P_0}(1)$$

because $O_{P_0}(\hat{\sigma}/\hat{\varepsilon}_n^2) = O_{P_0}(n^{-1/2}/\hat{\varepsilon}_n^2) = o_{P_0}(1)$ by assumption. In the non-overlapping case, $\hat{\tilde{\sigma}} \to_{P_0} \tilde{\sigma} > 0$ and $|\hat{\varepsilon}_n| = O(1)$, so again we have $\frac{\hat{\varepsilon}_n}{\hat{\tilde{\sigma}}} = O_{P_0}(1)$. Let $a(n) := \#(I_{odd,n} \setminus I_{1,n}) = \#(I_{even,n} \setminus I_{2,n})$. Then,

$$\left|\tilde{t}_n - \tilde{\tilde{t}}_n\right| = \frac{O_{P_0}(1)\sqrt{a(n)}}{\sqrt{n}}\left(\frac{1}{\sqrt{a(n)}}\sum_{i\in I_{odd,n}\setminus I_{1,n}} \Big(\ln f_A(X_i;\hat{\theta}_A) + \ln f_B(X_i;\hat{\theta}_B)\Big)\right.$$

$$\left. - \frac{1}{\sqrt{a(n)}}\sum_{i\in I_{even,n}\setminus I_{2,n}} \Big(\ln f_A(X_i;\hat{\theta}_A) + \ln f_B(X_i;\hat{\theta}_B)\Big)\right)$$

$$= \frac{\sqrt{a(n)}}{\sqrt{n}} O_{P_0}(1) = o_{P_0}(1)$$

because $a(n)/n \to 0$ and because the standardized sums are independent and asymptotically normal with finite, nonzero variances. Q.E.D.

Proof of Theorem 2. Lemma 5, whose assumptions are satisfied by setting $\mathcal{Q} = \mathcal{P}_0$ and by Assumptions 2 and 6, implies that $\sqrt{n}\hat{\bar{d}}/\hat{\bar{\sigma}} \to_d N(\bar{\delta}, 1)$ under any sequence $\{P_n\}$ in $\mathcal{P}_0$. Using this result, the theorem follows from analogous reasoning as in the proof of Theorem 11.4.5 of Lehmann and Romano (2005). Q.E.D.

Proof of Theorem 3. The result follows directly from Lemma 5. Q.E.D.

Proof of Theorem 4. The proof proceeds by decomposing the statistic into an asymptotically normal component and non-normal remainder terms that are negligible in an almost sure sense. We first obtain some generic asymptotic expansions that hold for triangular arrays (as needed for local power calculation). These expansions, specialized to the case of sequences, are also used for size calculations.

We first observe that, by Assumptions 7 and 8, Lemma 6 in Appendix 7 implies that $n^{-1}\sum_{i=1}^n \ln f_A(X_{ni}, \theta_A)$ converges almost surely uniformly for all $\theta_A \in \Theta_A$ to $E_{P_0}[\ln f_A(X_{0i}, \theta_A)]$. This in turn implies that $\hat{\theta}_A \to_{as} \theta_A^* := \theta_A^*(P_0)$ by the usual argument for consistency of MLE, adapted for almost sure convergence. We then expand the first order condition for $\hat{\theta}_A$ as

$$0 = \frac{1}{n}\sum_{i=1}^n \nabla_{\theta_A} \ln f_A\left(X_{ni}, \hat{\theta}_A\right) = \frac{1}{n}\sum_{i=1}^n \nabla_{\theta_A} \ln f_A\left(X_{ni}, \theta_A^*\right) + \frac{1}{n}\sum_{i=1}^n \nabla^2_{\theta_A} \ln f_A\left(X_{ni}, \bar{\theta}_A\right)\left(\hat{\theta}_A - \theta_A^*\right)$$

where $\bar{\theta}_A$ is a mean value on the line segment joining $\hat{\theta}_A$ and θ_A^*. By Assumptions 7 and 9, Lemma 6 implies that $n^{-1}\sum_{i=1}^n \nabla^2_{\theta_A} \ln f_A(X_{ni}, \theta_A)$ converges uniformly to $E_{P_0}[\nabla^2_{\theta_A} \ln f_A(X_{0i}, \theta_A)]$ for all $\theta_A \in \Theta_A$. Since $n^{-1}\sum_{i=1}^n \nabla^2_{\theta_A} \ln f_A(X_{ni}, \theta_A)$ is continuous in θ_A at each n by Assumption 9 and the convergence is uniform, it follows that the limit $E_{P_0}[\nabla^2_{\theta_A} \ln f_A(X_{0i}, \theta_A)]$ is also continuous in θ_A. Since $\hat{\theta}_A \to_{as} \theta_A^*$ and thefore $\bar{\theta}_A \to_{as} \theta_A^*$, we have

$$\frac{1}{n}\sum_{i=1}^n \nabla^2_{\theta_A} \ln f_A(X_{ni}, \bar{\theta}_A) = E_{P_0}[\nabla^2_{\theta_A} \ln f_A(X_{0i}, \theta_A^*)] + o_{as}(1)$$

and it follows, under Assumption 11, that

$$\hat{\theta}_A - \theta_A^* = -\left(\left(E_{P_0}\left[\nabla^2_{\theta_A} \ln f_A\left(X_{0i}, \theta_A^*\right)\right]\right)^{-1} + o_{as}\left(1\right)\right) \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_A} \ln f_A\left(X_{ni}, \theta_A^*\right). \tag{15}$$

Let $\|M\|_F$ denote the largest eigenvalue of matrix M and θ_{kj} the j-th component of θ_k, $k = A, B$. Observe that, by Assumption 16 and dominated convergence, $V_A(P_n) \to V_A$ with $\|V_A\|_F < \infty$. Moreover V_A is invertible by Assumption 11. We can then write

$$\begin{aligned}
\limsup_{n\to\infty} \left|\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right| &= \limsup_{n\to\infty} \left|V_A(P_n)^{1/2} V_A(P_n)^{-1/2} \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right| \\
&= \left(\lim_{n\to\infty} V_A(P_n)^{1/2}\right)\left(\limsup_{n\to\infty} \left|V_A(P_n)^{-1/2} \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right|\right) \\
&= V_A^{1/2}\left(\limsup_{n\to\infty} \left|V_A(P_n)^{-1/2} \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right|\right) \\
&\leq \|V_A\|_F^{1/2} \limsup_{n\to\infty} \left|V_A(P_n)^{-1/2} \frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right|
\end{aligned}$$

The summation term in (15) then has two possible behaviors: Either

$$\limsup_{n\to\infty} \left|\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_{ni}, \theta_A^*\right)\right| \leq \|V_A\|_F^{1/2} \sqrt{2\ln n} \tag{16}$$

almost surely for the general triangular array case (by Lemma 8 under Assumption 12 and the fact that $E_{P_0}[\nabla_{\theta_{Aj}} \ln f_A(X_{ni}, \theta_A^*)] = 0$), or

$$\limsup_{n\to\infty} \left|\frac{1}{n}\sum_{i=1}^{n} \nabla_{\theta_{Aj}} \ln f_A\left(X_i, \theta_A^*\right)\right| \leq \|V_A\|_F^{1/2} \sqrt{2\ln\ln n} \tag{17}$$

almost surely when X_{ni} reduces to a sequence ($X_{ni} = X_i$ and $V_A(P_n) = V_A$), by the Law of Iterated Logarithm (LIL) (Hartman and Wintner (1941)), since Assumption 12 implies

existence of the variance. In either case, it follows that[1]

$$\hat{\theta}_A - \theta_A^* = O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) \tag{18}$$

with $s = 1$ (for arrays) or $s = 2$ (for sequences), where $\ln^{\circ s}$ represents s application(s) of the ln function. A similar result holds for $\hat{\theta}_B$.

We now consider each term in the statistic $\tilde{t}_n = (\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J)/\hat{\tilde{\sigma}}$ where

$$\hat{L}_S := n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \hat{\theta}_A\right) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B\left(X_{ni}, \hat{\theta}_B\right),$$
$$\hat{L}_J := n^{-1/2} \sum_{i=1}^{n} \left(\ln f_A\left(X_{ni}, \hat{\theta}_A\right) - \ln f_B\left(X_{ni}, \hat{\theta}_B\right)\right).$$

We derive the power and size expansions for our test when $\hat{\varepsilon}_n$ is defined by the random sequence given in (4) and $\varepsilon_n := (C_{SD}/C_{PL}(\delta))^{1/3} n^{-1/6} (\ln \ln n)^{1/3}$ (this is the setup of Corollary 1), but the special case when $\hat{\varepsilon}_n = \varepsilon_n$ is some deterministic sequence in $\mathcal{E}$ follows immediately.

Write $\hat{\varepsilon}_n \hat{L}_S = \varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n) L_S + \hat{\varepsilon}_n (R_{\theta_A} - R_{\theta_B})$ with

$$L_S := n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \theta_A^*\right) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B\left(X_{ni}, \theta_B^*\right)$$
$$R_{\theta_A} := n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \hat{\theta}_A\right) - n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \theta_A^*\right)$$
$$R_{\theta_B} := n^{-1/2} \sum_{i \text{ odd}} \ln f_B\left(X_{ni}, \hat{\theta}_B\right) - n^{-1/2} \sum_{i \text{ odd}} \ln f_B\left(X_{ni}, \theta_B^*\right)$$

[1]For some random sequence R_n and some deterministic sequence r_n, we write $R_n = O_{as}(r_n)$ if and only if there exists a finite C such that $P(\limsup_{n\to\infty} |R_n/r_n| \leq C) = 1$.

We can bound R_{θ_A} (and similarly R_{θ_B}) using an expansion to second order about $\theta_A = \theta_A^*$:

$$\begin{aligned}
R_{\theta_A} &= n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \hat{\theta}_A\right) - n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \theta_A^*\right) \\
&= n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \theta_A^*\right) + \left(\hat{\theta}_A - \theta_A^*\right)' n^{-1/2} \sum_{i \text{ even}} \nabla_{\theta_A} \ln f_A\left(X_{ni}, \theta_A^*\right) \\
&\quad + \frac{1}{2}\left(\hat{\theta}_A - \theta_A^*\right)' \left(n^{-1/2} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A\left(X_{ni}, \bar{\theta}_A\right)\right) \left(\hat{\theta}_A - \theta_A^*\right) \\
&\quad - n^{-1/2} \sum_{i \text{ even}} \ln f_A\left(X_{ni}, \theta_A^*\right) \\
&= \left(\hat{\theta}_A - \theta_A^*\right)' n^{1/2} n^{-1} \sum_{i \text{ even}} \nabla_{\theta_A} \ln f_A\left(X_{ni}, \theta_A^*\right) \\
&\quad + \frac{n^{1/2}}{4}\left(\hat{\theta}_A - \theta_A^*\right)' \left((n/2)^{-1} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A\left(X_{ni}, \bar{\theta}_A\right)\right) \left(\hat{\theta}_A - \theta_A^*\right)
\end{aligned}$$

where $\bar{\theta}_A$ is a mean value on the line segment joining $\hat{\theta}_A$ and θ_A^*. Then, we use (18) and Lemma 6 applied to $n^{-1} \sum_{i \text{ even}} \nabla_{\theta_A}^2 \ln f_A\left(X_{ni}, \bar{\theta}_A\right)$ under Assumptions 7 and 9:

$$\begin{aligned}
\|R_{\theta_A}\| &= O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) n^{1/2} O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) \\
&\quad + n^{1/2} O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right)\left(O(1) + o_{as}(1)\right) O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) \\
&= O_{as}\left(n^{-1/2} \ln^{\circ s} n\right)
\end{aligned}$$

Next, $\hat{L}_J = L_J + L_{J2A} - L_{J2B}$ where

$$\begin{aligned}
L_J &:= n^{-1/2} \sum_{i=1}^{n} \left(\ln f_A(X_{ni}, \theta_A^*) - \ln f_B(X_{ni}, \theta_B^*)\right) \\
L_{J2A} &:= n^{-1/2} \sum_{i=1}^{n} \left(\ln f_A(X_{ni}, \hat{\theta}_A) - \ln f_A(X_{ni}, \theta_A^*)\right) \\
L_{J2B} &:= n^{-1/2} \sum_{i=1}^{n} \left(\ln f_A(X_{ni}, \hat{\theta}_B) - \ln f_B(X_{ni}, \theta_B^*)\right).
\end{aligned}$$

The terms L_{J2A} and L_{J2B} can be bounded using the same techniques as for R_{θ_A} and we have:

$$|L_{J2A}| = O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right)$$

and similarly for L_{J2B}. Next, let $\sigma_S^2 := \frac{1}{2}(\sigma_A^2 + \sigma_B^2)$, $\sigma_k^2 := \sigma_k^2(P_0)$, and $\hat{\sigma}_S^2 := \frac{1}{2}(\hat{\sigma}_A^2 + \hat{\sigma}_B^2)$. We have

$$\begin{aligned}
\hat{\sigma}_A^2 &= \frac{1}{n}\sum_{i=1}^{n}\left(\ln f_A\left(X_{ni}, \hat{\theta}_A\right)\right)^2 - \left(\frac{1}{n}\sum_{i=1}^{n}\ln f_A\left(X_{ni}, \hat{\theta}_A\right)\right)^2 \\
&= \frac{1}{n}\sum_{i=1}^{n}\left(\ln f_A\left(X_{ni}, \theta_A^*\right)\right)^2 + \left(\hat{\theta}_A - \theta_A^*\right)' \frac{1}{n}\sum_{i=1}^{n}\ln f_A\left(X_{ni}, \bar{\theta}_A\right)\nabla_{\theta_A}\ln f_A\left(X_{ni}, \bar{\theta}_A\right) \\
&\quad - \left(\frac{1}{n}\sum_{i=1}^{n}\ln f_A\left(X_{ni}, \theta_A^*\right) + O_{as}\left(n^{-1}\ln^{\circ s} n\right)\right)^2 \\
&= E_{P_0}\left[\left(\ln f_A\left(X_{ni}, \theta_A^*\right)\right)^2\right] - E_{P_0}([\ln f_A\left(X_{ni}, \theta_A^*\right)])^2 + O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) \\
&\quad + O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right)\left(O\left(1\right) + o_{as}\left(1\right)\right) \\
&= \sigma_A^2 + O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right)
\end{aligned}$$

where the rate of convergence of the first term follows from Lemma 8 (under Assumption 13) while the one of the second term follows from (18) and Lemma 6 under Assumptions 7 and 10. Similarly, we have $\hat{\sigma}_B^2 = \sigma_B^2 + O_{as}(n^{-1/2}\sqrt{\ln^{\circ s} n})$ and, thus, $\hat{\sigma}_S^2 = \sigma_S^2 + O_{as}(n^{-1/2}\sqrt{\ln^{\circ s} n})$. By a similar reasoning, by Assumptions 14–17, we have $\hat{H}_k = H_k + O_{as}(n^{-1/2}\sqrt{\ln^{\circ s} n})$ and $\hat{V}_k = V_k + O_{as}(n^{-1/2}\sqrt{\ln^{\circ s} n})$ for $k = A, B$. Below, we will use $\ln\ln n = O(\ln n)$ to simplify some expressions. From the convergence of $\hat{\sigma}_S^2$, $\hat{H}_k$ and $\hat{V}_k$, it also follows that $|\hat{c}_\alpha - c_\alpha| = O_{as}(n^{-1/2}\sqrt{\ln^{\circ s} n})$ and thus

$$\hat{\varepsilon}_n = \varepsilon_n + O_{as}\left(\varepsilon_n n^{-1/2}\sqrt{\ln^{\circ s} n}\right)$$

Similarly:

$$\hat{\varepsilon}_n^2 = \left(\varepsilon_n + O_{as}\left(\varepsilon_n n^{-1/2}\sqrt{\ln^{\circ s} n}\right)\right)^2 = \varepsilon_n^2 + O_{as}\left(\varepsilon_n^2 n^{-1/2}\sqrt{\ln^{\circ s} n}\right) + O\left(\varepsilon_n^2 n^{-1}\ln^{\circ s} n\right)$$
$$= \varepsilon_n^2 + O_{as}\left(\varepsilon_n^2 n^{-1/2}\sqrt{\ln^{\circ s} n}\right).$$

Next, one can handle $\hat{\sigma}_{AB}$ and, thus, $\hat{\sigma}^2$ by a similar reasoning, invoking Assumptions 7 and 10 and Lemma 6 to yield:

$$\hat{\sigma}^2 = \sigma^2 + O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right).$$

Letting $\tilde{\sigma}^2(\varepsilon_n) := \varepsilon_n^2\sigma_S^2 + (1+\varepsilon_n)\sigma^2$, we can also write

$$\begin{aligned}
\hat{\tilde{\sigma}}^2 &= \varepsilon_n^2\sigma_S^2 + (1+\varepsilon_n)\,\sigma^2 + \varepsilon_n^2\left(\hat{\sigma}_S^2 - \sigma_S^2\right) + \left(\hat{\varepsilon}_n^2 - \varepsilon_n^2\right)\hat{\sigma}_S^2 + (\hat{\varepsilon}_n - \varepsilon_n)\,\sigma^2 + (1+\hat{\varepsilon}_n)\left(\hat{\sigma}^2 - \sigma^2\right) \\
&= \tilde{\sigma}^2(\varepsilon_n) + O_{as}\left(\varepsilon_n^2 n^{-1/2}\sqrt{\ln^{\circ s} n}\right) + O_{as}\left(\varepsilon_n^2 n^{-1/2}\sqrt{\ln^{\circ s} n}\right) O_{as}(1) \\
&\quad + O_{as}\left(\varepsilon_n n^{-1/2}\sqrt{\ln^{\circ s} n}\right) O(1) + (1 + O_{as}(\varepsilon_n))\, O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right) \\
&= \tilde{\sigma}^2(\varepsilon_n) + O_{as}\left(n^{-1/2}\sqrt{\ln^{\circ s} n}\right)
\end{aligned}$$

Collecting all remainder terms for the triangular array case ($s = 1$), we have

$$\begin{aligned}
\tilde{t}_n &= \frac{\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J}{\hat{\tilde{\sigma}}} = \frac{\varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n)\, L_S + \hat{\varepsilon}_n\left(R_{\theta_A} - R_{\theta_B}\right) + L_J + L_{J2A} - L_{J2B}}{\hat{\tilde{\sigma}}} \\
&= \frac{\varepsilon_n L_S + O_{as}\left(\varepsilon_n n^{-1/2}\sqrt{\ln n}\right) O_{as}(1) + O_{as}(\varepsilon_n)\, O_{as}\left(n^{-1/2}\ln n\right)}{\tilde{\sigma}(\varepsilon_n) + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right)} \\
&\quad + \frac{L_J + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right)}{\tilde{\sigma}(\varepsilon_n) + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right)} \\
&= \frac{\varepsilon_n L_S + L_J + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right)}{\tilde{\sigma}(\varepsilon_n) + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right)} = \frac{\varepsilon_n L_S + L_J}{\tilde{\sigma}(\varepsilon_n)} + O_{as}\left(n^{-1/2}\sqrt{\ln n}\right),
\end{aligned}$$

that is, $\tilde{t}_n = t_n + \Delta t_n$ with

$$t_n := \frac{\varepsilon_n L_S + L_J}{\tilde{\sigma}(\varepsilon_n)}$$

$$|\Delta t_n| \leq \Delta \bar{t}_n \text{ a.s.}$$

for $\Delta \bar{t}_n := Bn^{-1/2}\sqrt{\ln n}$ for some constant B and where "a.s." denotes "almost surely as $n \to \infty$" , i.e., the event $|\Delta t_n| > \Delta \bar{t}_n$ has probability zero for all $n \geq n_0$ with n_0 sufficiently large.

Power expansion. We now calculate an expansion of our test's power in orders of ε_n and n. Consider the following decomposition:

$$\begin{aligned}
P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) &= 1 - P_n\left(\tilde{t}_n \leq z_{1-\alpha/2}\right) + P_n\left(\tilde{t}_n < z_{\alpha/2}\right) \\
&= \underbrace{1 - P_n\left(\tilde{t}_n \leq z_{1-\alpha/2}\right) - \left(1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta(1+\varepsilon_n/2)}{\tilde{\sigma}(\varepsilon_n)}\right)\right)}_{=:I_1} \\
&\quad + \underbrace{P_n\left(\tilde{t}_n < z_{\alpha/2}\right) - \Phi\left(z_{\alpha/2} - \frac{\delta(1+\varepsilon_n/2)}{\tilde{\sigma}(\varepsilon_n)}\right)}_{=:I_2} \\
&\quad + \underbrace{1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta(1+\varepsilon_n/2)}{\tilde{\sigma}(\varepsilon_n)}\right) - \left(1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta}{\sigma}\right)\right)}_{=:I_3} \\
&\quad + \underbrace{\Phi\left(z_{\alpha/2} - \frac{\delta(1+\varepsilon_n/2)}{\tilde{\sigma}(\varepsilon_n)}\right) - \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right)}_{=:I_4} \\
&\quad + 1 - \Phi\left(z_{1-\alpha/2} - \frac{\delta}{\sigma}\right) + \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) \qquad (19)
\end{aligned}$$

We bound each of the terms in turn. When the models are not overlapping, both L_S and L_J are asymptotically normal, since they are iid sample averages (evaluated at the true parameter values) of bounded variance quantities. Moreover, by the Berry-Esseen bound

(since Assumption 8 implies that the third moments of the log-likelihood function exist and are uniformly bounded), we have that the deviations from normality of finite sample distribution of the normalized statistic $(\varepsilon_n L_S + L_J)/\tilde{\sigma}(\varepsilon_n)$ are uniformly bounded by $Cn^{-1/2}$ for some universal constant C (this remains true for triangular arrays, since the constant is independent of the distribution among distributions sharing the same upper bound on the third moments). Fix some $\beta \in (0,1)$ and $\delta \in \mathbb{R} \setminus \{0\}$, and let $f_\beta(\varepsilon) := z_\beta - \frac{\delta(1+\varepsilon/2)}{\tilde{\sigma}(\varepsilon)}$. We then have for $n \geq n_0$,

$$
\begin{aligned}
\left|P_n\left(\tilde{t}_n \leq z_\beta\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| &= \left|P_n\left(\tilde{t}_n \leq z_\beta\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&= \left|P_n\left(t_n + \Delta t_n \leq z_\beta\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&= \left|P_n\left(t_n + \Delta t_n \leq z_\beta \mid |\Delta t_n| \leq \Delta \bar{t}_n\right) P_n\left(|\Delta t_n| \leq \Delta \bar{t}_n\right) + \right. \\
&\quad \left. + P_n\left(t_n + \Delta t_n \leq z_\beta \mid |\Delta t_n| > \Delta \bar{t}_n\right) P_n\left(|\Delta t_n| > \Delta \bar{t}_n\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&= \left|P_n\left(t_n + \Delta t_n \leq z_\beta \mid |\Delta t_n| \leq \Delta \bar{t}_n\right) \cdot 1 \right. \\
&\quad \left. + P_n\left(t_n + \Delta t_n \leq z_\beta \mid |\Delta t_n| > \Delta \bar{t}_n\right) \cdot 0 - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&= \left|P_n\left(t_n + \Delta t_n \leq z_\beta \mid |\Delta t_n| \leq \Delta \bar{t}_n\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&\leq \sup_{|u| \leq \Delta \bar{t}_n} \left|P_n\left(t_n + u \leq z_\beta\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&= \sup_{|u| \leq \Delta \bar{t}_n} \left|P_n\left(t_n \leq z_\beta - u\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| \\
&\leq \sup_{|u| \leq \Delta \bar{t}_n} \left|\Phi\left(f_\beta(\varepsilon_n) - u\right) - \Phi\left(f_\beta(\varepsilon_n)\right)\right| + Cn^{-1/2} \\
&= \sup_{|u| \leq \Delta \bar{t}_n} \phi\left(f_\beta(\varepsilon_n) + \bar{u}\right)|u| + Cn^{-1/2} \\
&\leq \sup_{|\bar{u}| \leq \Delta \bar{t}_n} \phi\left(f_\beta(\varepsilon_n) + \bar{u}\right) \Delta \bar{t}_n + Cn^{-1/2} \\
&= \phi\left(f_\beta(\varepsilon_n) + o(1)\right) \Delta \bar{t}_n + Cn^{-1/2} = O\left(\Delta \bar{t}_n\right) \qquad (20)
\end{aligned}
$$

where $\bar{u}$ is a mean value satisfying $|\bar{u}| \leq |u| \leq \Delta \bar{t}_n = o(1)$ and by continuity of $\phi(\cdot)$, we

have $\phi(z+o(1)) = \phi(z) + o(1)$. Therefore,

$$I_1 + I_2 = O(\Delta \bar{t}_n) = O(n^{-1/2}\sqrt{\ln n}).$$

Consider I_3 and I_4. First, notice that

$$f'_\beta(0) = \left.\frac{-\delta\varepsilon\left(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2)\right)}{4\tilde{\sigma}(\varepsilon)^3}\right|_{\varepsilon=0} = 0$$

so that

$$\begin{aligned}
\Phi\left(f_\beta(\varepsilon_n)\right) - \Phi\left(f_\beta(0)\right) &= \phi\left(f_\beta(\varepsilon)\right) f'_\beta(\varepsilon)\big|_{\varepsilon=0}\, \varepsilon_n \\
&\quad + \frac{1}{2}\left[\phi'\left(f_\beta(\varepsilon)\right)\left(f'_\beta(\varepsilon)\right)^2 + \phi\left(f_\beta(\varepsilon)\right) f''_\beta(\varepsilon)\right]\Big|_{\varepsilon=0} \varepsilon_n^2 + O\left(\varepsilon_n^3\right) \\
&= \frac{1}{2}\phi\left(f_\beta(\varepsilon)\right) f''_\beta(\varepsilon)\Big|_{\varepsilon=0} \varepsilon_n^2 + O\left(\varepsilon_n^3\right) \\
&= -\frac{1}{2}\phi\left(z_\beta - \frac{\delta}{\sigma}\right)\left(\frac{\delta(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{4\sigma^3}\right)\varepsilon_n^2 + O\left(\varepsilon_n^3\right)
\end{aligned}$$

Therefore, for all $\delta \in \mathbb{R} \setminus \{0\}$:

$$\begin{aligned}
I_3 + I_4 &= -\left[\Phi\left(f_{1-\alpha/2}(\varepsilon_n)\right) - \Phi\left(f_{1-\alpha/2}(0)\right)\right] + \Phi\left(f_{\alpha/2}(\varepsilon_n)\right) - \Phi\left(f_{\alpha/2}(0)\right) \\
&= -\underbrace{\left(\phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) - \phi\left(z_{\alpha/2} + \frac{\delta}{\sigma}\right)\right)\frac{\delta(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{8\sigma^3}}_{=C_{PL}(\delta)}\varepsilon_n^2 + O(\varepsilon_n^3) \qquad (21)
\end{aligned}$$

Together, (19)–(21) yield the desired expansion of power in powers of ε_n and n:

$$P_n\left(|\tilde{t}_n| > z_{1-\alpha/2}\right) = \Phi\left(z_{\alpha/2} + \frac{\delta}{\sigma}\right) + \Phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right) - C_{PL}(\delta)\varepsilon_n^2 + O\left(\varepsilon_n^3\right) + O\left(n^{-1/2}\sqrt{\ln n}\right). \qquad (22)$$

Size expansion. We now calculate the size distortion when the models are overlapping. In the overlapping case, we need to provide a more precise bound on the remainder terms of $\hat{L}_J = L_J + L_{J2A} - L_{J2B}$, because the leading term vanishes ($L_J = 0$) due to the overlap.

Drifting sequences of models are not needed for the size calculation, so the triangular array X_{ni} can be replaced by a simple iid sequence X_i drawn from P_0. Letting $\hat{g}_A := \frac{1}{n}\sum_{i=1}^{n}\left(\nabla_{\theta_A}\ln f_A\left(X_i,\theta_A^*\right)\right)$, we have

$$\begin{aligned}
L_{J2A} &= \frac{n^{1/2}}{2}\hat{g}_A'\left(H_A^{-1}+o_{as}\left(1\right)\right)\hat{g}_A \\
&= \frac{n^{1/2}}{2}\hat{g}_A' V_A^{-1/2}V_A^{1/2}\left(H_A^{-1}+o_{as}\left(1\right)\right)V_A^{1/2}V_A^{-1/2}\hat{g}_A \\
&= -\frac{n^{1/2}}{2}Z_A' V_A^{1/2}\left(-H_A^{-1}+o_{as}\left(1\right)\right)V_A^{1/2}Z_A
\end{aligned}$$

where $Z_A := V_A^{-1/2}\hat{g}_A$. The matrix $V_A^{1/2}\left(-H_A\right)^{-1}V_A^{1/2}$ is symmetric so it is diagonalizable, with eigenvalues λ_j and orthogonal eigenvectors v_j (normalized to $\|v_j\| = 1$). Moreover, the eigenvalues are all positive (because both $-H_A$ and V_A are positive-definite) and we can write $V^{1/2}\left(-H\right)^{-1}V^{1/2} = \sum_{j=1}^{\dim\theta_A} v_j\lambda_j v_j'$ and thus:

$$\begin{aligned}
\left|L_{J2A}\right| &= -L_{J2A} = \frac{n^{1/2}}{2}Z_A'\left(\sum_{j=1}^{\dim\theta_A} v_j\lambda_j v_j' + o_{as}\left(1\right)\right)Z_A. \\
&= \frac{n^{1/2}}{2}\sum_{j=1}^{\dim\theta_A} Z_A' v_j\lambda_j v_j' Z_A + o_{as}\left(1\right)\frac{n^{1/2}}{2}Z_A'Z_A \\
&= \frac{n^{1/2}}{2}\sum_{j=1}^{\dim\theta_A}\lambda_j\left(v_j'Z_A\right)^2 + o_{as}\left(1\right)\frac{n^{1/2}}{2}Z_A'Z_A
\end{aligned}$$

By construction, the covariance matrix of the $v_j'Z_A$ is the identity matrix I. We can then use the Law of the Iterated Logarithm (Hartman and Wintner (1941)) to conclude

$|v_j' Z| \leq n^{-1/2}\sqrt{2\ln\ln n}$ almost surely. We then have

$$\begin{aligned}
|L_{J2A}| &\leq \frac{n^{1/2}}{2}\sum_{j=1}^{\dim\theta_A}\lambda_j\left(n^{-1/2}\sqrt{2\ln\ln n}\right)^2 + o_{as}(1)\frac{n^{1/2}}{2}(\dim\theta_A)\left(n^{-1/2}\sqrt{2\ln\ln n}\right)^2 \\
&= \left(n^{-1/2}\ln\ln n\right)\sum_{j=1}^{\dim\theta_A}\lambda_j + o_{as}\left(n^{-1/2}\ln\ln n\right) \\
&= \left(n^{-1/2}\ln\ln n\right) tr\left(V_A^{1/2}\left(-H_A\right)^{-1}V_A^{1/2}\right) + o_{as}\left(n^{-1/2}\ln\ln n\right) \\
&= \left|tr\left(H_A^{-1}V_A\right)\right|\left(n^{-1/2}\ln\ln n\right) + o_{as}\left(n^{-1/2}\ln\ln n\right)
\end{aligned}$$

A similar reasoning holds for $|L_{J2B}|$ and since both L_{J2A} and L_{J2B} have the same sign and $L_J = 0$, we have

$$\begin{aligned}
\left|\hat{L}_J\right| &= |L_J + L_{J2A} - L_{J2B}| = |L_{J2A} - L_{J2B}| \\
&\leq \max\left\{|L_{J2A}|, |L_{J2B}|\right\} \leq \max\left\{\left|tr\left(H_A^{-1}V_A\right)\right|, \left|tr\left(H_B^{-1}V_B\right)\right|\right\} n^{-1/2}\ln\ln n \quad \text{a.s.} \\
&= \Lambda n^{-1/2}\ln\ln n,
\end{aligned}$$

where $\Lambda := \max\{|tr(H_A^{-1}V_A)|, |tr(H_B^{-1}V_B)|\}$. In the overlapping case, $\tilde{\sigma}^2(\varepsilon) = \varepsilon^2\sigma_S^2 + (1+\varepsilon)\sigma_J^2 = \varepsilon^2\sigma_S^2$ since $\sigma_J^2 = 0$. We can now compute the worst-case size distortion in $\tilde{t}_n$.

Collecting the order of all remainders, we have,

$$
\begin{aligned}
\tilde{t}_n &= \frac{\hat{\varepsilon}_n \hat{L}_S + \hat{L}_J}{\hat{\tilde{\sigma}}} = \frac{\varepsilon_n L_S + (\hat{\varepsilon}_n - \varepsilon_n) L_S + \hat{\varepsilon}_n (R_{\theta_A} - R_{\theta_B}) + L_J + L_{J2A} - L_{J2B}}{\tilde{\sigma}(\varepsilon_n) + O_{as}\left(n^{-1/2}\sqrt{\ln\ln n}\right)} \\
&= \frac{\varepsilon_n L_S + O_{as}\left(\varepsilon_n n^{-1/2}\sqrt{\ln\ln n}\right) O_{as}(1) + O_{as}(\varepsilon_n) O_{as}\left(n^{-1/2}\ln\ln n\right)}{\varepsilon_n \sigma_S + O_{as}\left(n^{-1/2}\sqrt{\ln\ln n}\right)} \\
&\quad + \frac{L_{J2A} - L_{J2B}}{\varepsilon_n \sigma_S + O_{as}\left(n^{-1/2}\sqrt{\ln\ln n}\right)} \\
&= \frac{\varepsilon_n L_S + (L_{J2A} - L_{J2B}) + O_{as}\left(\varepsilon_n n^{-1/2}\ln\ln n\right)}{\varepsilon_n \sigma_S + O_{as}\left(n^{-1/2}\sqrt{\ln\ln n}\right)} \\
&= \frac{L_S}{\sigma_S} \frac{\varepsilon_n + (L_{J2A} - L_{J2B})/L_S + O_{as}\left(\varepsilon_n n^{-1/2}\ln\ln n\right)}{\varepsilon_n + O_{as}\left(n^{-1/2}\sqrt{\ln\ln n}\right)} \\
&= \frac{L_S}{\sigma_S} \frac{1 + (L_{J2A} - L_{J2B})/(\varepsilon_n L_S) + O_{as}\left(n^{-1/2}\ln\ln n\right)}{1 + O_{as}\left(n^{-1/2}\left(\sqrt{\ln\ln n}\right)/\varepsilon_n\right)} \\
&= \left(\frac{L_S}{\sigma_S} + \frac{L_{J2A} - L_{J2B}}{\varepsilon_n \sigma_S} + O_{as}\left(n^{-1/2}\ln\ln n\right)\right) \times \frac{1}{\left(1 + O_{as}\left(n^{-1/2}\left(\sqrt{\ln\ln n}\right)/\varepsilon_n\right)\right)} \\
&= \frac{L_S}{\sigma_S} + \frac{L_{J2A} - L_{J2B}}{\varepsilon_n \sigma_S} + O_{as}\left(n^{-1/2}\ln\ln n\right) \\
&= \frac{L_S}{\sigma_S} + \Delta t_n
\end{aligned}
$$

where $\Delta t_n := (L_{J2A} - L_{J2B})/(\varepsilon_n \sigma_S) + O_{as}(n^{-1/2}\ln\ln n)$. We can bound Δt_n as follows, substituting in ε_n:

$$
\begin{aligned}
|\Delta t_n| &= \frac{|L_{J2A} - L_{J2B}|}{\varepsilon_n \sigma_S} + O_{as}\left(n^{-1/2}\ln\ln n\right) \\
&\leq \frac{\Lambda n^{-1/2}\ln\ln n}{\varepsilon_n \sigma_S} + O_{as}\left(n^{-1/2}\ln\ln n\right) \quad \text{a.s.}
\end{aligned}
$$

Notice that the size of the test can be decomposed as

$$\begin{aligned} P_0(|\tilde{t}_n| > z_{1-\alpha/2}) &= \alpha + P_0(\tilde{t}_n > z_{1-\alpha/2}) - (1 - \Phi(z_{1-\alpha/2})) + P_0(\tilde{t}_n < z_{\alpha/2}) - \Phi(z_{\alpha/2}) \\ &= \alpha + \Phi(z_{1-\alpha/2}) - P_0(\tilde{t}_n \leq z_{1-\alpha/2}) + P_0(\tilde{t}_n < z_{\alpha/2}) - \Phi(z_{\alpha/2}) \end{aligned} \tag{23}$$

By a derivation similar to that in (20), we have

$$\begin{aligned} \left|P_0(\tilde{t}_n \leq z_\beta) - \Phi(z_\beta)\right| &\leq \sup_{|u| \leq \Delta\bar{t}_n} \left|P_0\left(t_n + u \leq z_\beta\right) - \Phi\left(z_\beta\right)\right| \\ &\leq \sup_{|u| \leq \Delta\bar{t}_n} \left|\Phi\left(z_\beta - u\right) - \Phi\left(z_\beta\right)\right| + Cn^{-1/2} \\ &= \sup_{|u| \leq \Delta\bar{t}_n} \phi\left(z_\beta + \bar{u}\right)|u| + Cn^{-1/2} \\ &\leq \sup_{|\bar{u}| \leq \Delta\bar{t}_n} \phi\left(z_\beta + \bar{u}\right)\Delta\bar{t}_n + Cn^{-1/2} \\ &= \phi\left(z_\beta + o(1)\right)\Delta\bar{t}_n + Cn^{-1/2} \\ &= \phi\left(z_\beta\right)\Delta\bar{t}_n + Cn^{-1/2} + o(\Delta\bar{t}_n) \end{aligned} \tag{24}$$

Therefore, (23), (24), and the expression for $\Delta\bar{t}_n$ yield the expansion of size in terms of orders of ε_n and n:

$$\begin{aligned} P_0(|\tilde{t}_n| > z_{1-\alpha/2}) &\leq \alpha + \left[\phi\left(z_{1-\alpha/2}\right) + \phi(z_{\alpha/2})\right]\Delta\bar{t}_n + Cn^{-1/2} + o(\Delta\bar{t}_n) \\ &\leq \alpha + C_{SD}\frac{n^{-1/2}\ln\ln n}{\varepsilon_n} + O(n^{-1/2}) + o(n^{-1/2}\varepsilon_n^{-1}\ln\ln n) \end{aligned} \tag{25}$$

where $C_{SD} := 2\phi(z_{\alpha/2})\Lambda/\sigma_S$. Q.E.D.

Proof of Corollary 1. The expansions in Theorem 4 were established under the more general conditions of this corollary in which $\hat{\varepsilon}_n$ is a random sequence defined by (4).

We first show $0 \leq C_{PL}(\delta) \leq C_{PL}^*$ for all $\delta \in \mathbb{R}$. It is easy to see that $C_{PL}(\delta) \geq 0$ for all $\delta \in \mathbb{R}$ with equality if and only if $\delta = 0$. Solving $g_1'(\delta) = 0$ with

$$g_1(\delta) := \phi\left(z_{\alpha/2} - \frac{\delta}{\sigma}\right)\frac{\delta(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{8\sigma^3}$$

for δ and computing the second derivative of g_1, shows that there are two solutions, one being the global maximizer of $g_{1,\beta}$,

$$\delta^* := \frac{\sigma}{2}\left(z_{\alpha/2} - \sqrt{4 + z_{\alpha/2}^2}\right).$$

Similarly, one can show that

$$g_2(\delta) := -\phi\left(z_{\alpha/2} + \frac{\delta}{\sigma}\right)\frac{\delta(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{8\sigma^3}$$

has a global maximizer at $-\delta^*$. Therefore, for all $\delta \in \mathbb{R}$,

$$0 \leq C_{PL}(\delta) \leq g_1(\delta^*) + g_2(-\delta^*) = 2\phi\left(z_{\alpha/2} - \frac{\delta^*}{\sigma}\right)\frac{\delta^*(\sigma^2 - 2(\sigma_A^2 + \sigma_B^2))}{8\sigma^3} = C_{PL}^*$$

Second, it is immediate to see that the first-order term of power loss and size distortion are equal,

$$C_{PL}^*\varepsilon_n^2 = C_{SD}\frac{n^{-1/2}\ln\ln n}{\varepsilon_n},$$

when

$$\varepsilon_n = \left(\frac{C_{SD}}{C_{PL}^*}\right)^{1/3} n^{-1/6}(\ln\ln n)^{1/3}$$

which directly implies the expansions in the statement of the corollary.

Finally, we observe that ε_n is in $\mathcal{E}$ by construction and since we have shown in the proof of Theorem 4 that $\hat{\varepsilon}_n = \varepsilon_n + O_{as}(\varepsilon_n n^{-1/2}\sqrt{\ln^{\circ s} n})$, we automatically have $\hat{\varepsilon}_n - \varepsilon_n = O_p(n^{-1/2})$, for either sequences ($s = 2$) or triangular arrays ($s = 1$), and it follows that $\hat{\varepsilon}_n$ satisfies Assumptions 5 and 6. Q.E.D.

7 Auxiliary Lemmas

The following Lemma provides a uniform strong law of large numbers for triangular arrays. It is stated for scalars, but can also be used, element by element, for vectors valued $g(x, \theta)$.

Lemma 6. *For $n \in \mathbb{N}$, let X_{ni} for $i = 1, \ldots, n$ be iid random variables taking value in $\mathbb{R}^{d_X}$ and drawn from the probability measure P_n. Assume that the measures P_n converge weakly to some measure P_0 and that each $P_n(x)$ admits a Radon-Nikodym derivative $p_n(x)$ with respect to $P_0(x)$. For Θ compact (under some metric $d_\theta(\cdot,\cdot)$), let $g : \mathbb{R}^{d_X} \times \Theta \mapsto \mathbb{R}$ be continuous in x at each $\theta \in \Theta$. Assume further that there exists $G(x)$ such that $E_{P_0}[G(X_{0i})] < \infty$ (for X_{0i} drawn from P_0) and such that, for all $\theta \in \Theta$ and $n \in \mathbb{N}$,*

$$|g(x,\theta)|\, p_n(x) \leq G(x)$$

and that there exists $\bar{G} < \infty$ such that $E_{P_n}[|g(X_{ni},\theta)|^4] \leq \bar{G}$ for all $i = 1, \ldots, n$, all $n \in \mathbb{N}$ and all $\theta \in \Theta$. Then,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g(X_{ni},\theta) - g(\theta) \right| \overset{as}{\to} 0$$

for $g(\theta) := E_{P_0}[g(X_{0i},\theta)]$, where X_{0i} is drawn from P_0.

Proof. This proof parallels the one of Lemma 1 in Tauchen (1985), but adapted for triangular arrays. Define

$$u(x,\theta,d) = \sup_{\tilde{\theta}: d_\theta\left(\tilde{\theta},\theta\right) \leq d} \left| g\left(x,\tilde{\theta}\right) - g(x,\theta) \right|.$$

By almost sure continuity of $g(x,\theta)$, $\lim_{d\to 0} u(x,\theta,d) = 0$ almost surely, for a given θ. Also observe that, by P_n converging weakly to P_0, we must have that $p_n(x) \to 1$ pointwise for all x in a set of probability 1 under P_0. To study the convergence of $E_{P_n}[u(X,\theta,d)]$ as $d \to 0$ and $n \to \infty$, we employ dominated convergence. We have

$$E_{P_n}[u(X,\theta,d)] = \int u(x,\theta,d)\, dP_n(x) = \int u(x,\theta,d)\, p_n(x)\, dP_0(x)$$

where

$$|u(x,\theta,d)\, p_n(x)| \leq \sup_{d_\infty\left(\tilde{\theta},\theta\right) \leq d} \left| g\left(x,\tilde{\theta}\right) \right| p_n(x) + |g(x,\theta)|\, p_n(x) \leq G(x) + G(x) = 2G(x),$$

where $\int G(x)dP_0(x) < \infty$. Thus, for a given $\varepsilon > 0$, there exists $\bar{d}(\theta)$ and $\bar{N}(\theta, \varepsilon)$ such that $E_{P_n}[u(X_{ni}, \theta, d)] \leq \varepsilon$ whenever $d \leq \bar{d}(\theta)$ and $n \geq \bar{N}(\theta, \varepsilon)$. By a similar reasoning, $|g(\tilde{\theta}) - g(\theta)| \leq \varepsilon$ whenever $d(\tilde{\theta}, \theta) \leq \bar{d}(\theta)$. Let $B(\theta)$ be the open ball of radius $\bar{d}(\theta)$ about θ. By compactness of Θ, there exists a finite covering $B_k = B(\theta_k)$, $k = 1, \ldots, K$. Let $d_k = \bar{d}(\theta_k)$ and $\mu_k = E[u(X, \theta_k, d_k)]$ and write, for $\theta \in B_k$,

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n} g\left(X_{ni}, \theta\right) - g\left(\theta\right)\right| &\leq \left|\frac{1}{n}\sum_{i=1}^{n} g\left(X_{ni}, \theta\right) - g\left(X_{ni}, \theta_k\right)\right| + \left|\frac{1}{n}\sum_{i=1}^{n} g\left(X_{ni}, \theta_k\right) - E_{P_0}\left[g\left(X_{0i}, \theta_k\right)\right]\right| \\
&\quad + \left|E_{P_0}\left[g\left(X_{0i}, \theta_k\right)\right] - g\left(\theta\right)\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^{n} u\left(X_{ni}, \theta_k, d_k\right) - \mu_k\right| + \mu_k + \left|\frac{1}{n}\sum_{i=1}^{n} g\left(X_{ni}, \theta_k\right) - E_{P_0}\left[g\left(X_{0i}, \theta_k\right)\right]\right| \\
&\quad + \left|g\left(\theta_k\right) - g\left(\theta\right)\right| \\
&:= R_1 + \mu_k + R_2 + \left|g\left(\theta_k\right) - g\left(\theta\right)\right|
\end{aligned}$$

By construction, $\mu_k \leq \varepsilon$ and $|g(\theta_k) - g(\theta)| \leq \varepsilon$ for all $n \geq \bar{N}(\theta_k, \varepsilon)$. To apply a strong law of large number for triangular arrays (Lemma 7) for R_1 and R_2 above, we need to calculate fourth moments of the summands. We have

$$\begin{aligned}
E\left[\left|g\left(X_{ni}, \theta_k\right) - E_{P_0}\left[g\left(X_{0i}, \theta_k\right)\right]\right|^4\right] &\leq 8\left(E\left[\left|g\left(X_{ni}, \theta_k\right)\right|^4\right] + \left|E_{P_0}\left[g\left(X_{0i}, \theta_k\right)\right]\right|^4\right) \\
&\leq 16E\left[\left|g\left(X_{ni}, \theta\right)\right|^4\right] \leq 16\bar{G}
\end{aligned}$$

by the C_r and Jensen's inequalities and by the uniform boundedness of the fourth moment assumption. Similarly,

$$E\left[\left|u\left(X_{ni}, \theta_k, d_k\right)\right|^4\right] = E\left[\left|\sup_{\tilde{\theta}: d_\theta\left(\tilde{\theta}, \theta_k\right) \leq d_k} \left|g\left(X_{ni}, \tilde{\theta}\right) - g\left(X_{ni}, \theta\right)\right|\right|^4\right] = E\left[\left|g\left(X_{ni}, \theta^*\right) - g\left(X_{ni}, \theta\right)\right|^4\right]$$

for some θ^*, by compactness of (the closure of) $B(\theta_k)$. By the C_r inequality, we have $E[|g(x, \theta^*) - g(x, \theta)|^4] \leq 16\bar{G}$. Hence, we can apply Lemma 7 to conclude that there exists

$N_k(\varepsilon)$ such that $R_1 \leq \varepsilon$ and $R_2 \leq \varepsilon$ almost surely for all $n \geq N_k(\varepsilon)$. Thus,

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} g\left(X_{ni}, \theta\right) - g\left(\theta\right) \right| \leq 4\varepsilon$$

for $n \geq \max_k \max\{N_k(\varepsilon), \bar{N}(\theta_k, \varepsilon)\}$ almost surely. Since ε was arbitrary, the conclusion follows. Q.E.D.

The following lemma is a strong law of large number for triangular arrays.

Lemma 7. *Let Y_{ni} be a triangular array ($n \in \mathbb{N}$, $i = 1, \ldots, n$) of random variables, iid across $i = 1, \ldots, n$. If, for all $n \in \mathbb{N}$, $i = 1, \ldots, n$, $E\left[Y_{ni}\right] = 0$ and $E\left[\left|Y_{ni}\right|^4\right] \leq \bar{Y} < \infty$, then $n^{-1} \sum_{i=1}^{n} Y_{ni} \overset{as}{\to} 0$.*

Proof. The principle of this proof is borrowed from Example 5.41 in Romano and Siegel (1986). Note that

$$P\left[\left|\frac{1}{n} \sum_{i=1}^{n} Y_{ni}\right| \geq \varepsilon\right] \leq \frac{E\left[\left|\frac{1}{n} \sum_{i=1}^{n} Y_{ni}\right|^4\right]}{\varepsilon^4}$$

where

$$\begin{aligned}
E\left[\left(\frac{1}{n} \sum_{i=1}^{n} Y_{ni}\right)^4\right] &= n^{-4} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} \sum_{i_3=1}^{n} \sum_{i_4=1}^{n} E\left[Y_{ni_1} Y_{ni2} Y_{ni_3} Y_{ni_4}\right] \\
&= n^{-4} \sum_{i_1=1}^{n} \sum_{i_2=1}^{n} E\left[\left|Y_{ni_1}\right|^2 \left|Y_{ni2}\right|^2\right] + n^{-4} \sum_{i_1=1}^{n} E\left[\left|Y_{ni_1}\right|^4\right] \\
&= n^{-2} E\left[\left|Y_{ni}\right|^2\right] E\left[\left|Y_{ni}\right|^2\right] + n^{-3} E\left[\left|Y_{ni}\right|^4\right] \\
&\leq n^{-2} E\left[\left|Y_{ni}\right|^4\right]^{1/2} \left(E\left[\left|Y_{ni}\right|^4\right]\right)^{1/2} + n^{-3} E\left[\left|Y_{ni}\right|^4\right] \\
&\leq n^{-2} \bar{Y} + n^{-3} \bar{Y}.
\end{aligned}$$

Hence,

$$\sum_{n=1}^{\infty} P\left[\left|\frac{1}{n} \sum_{i=1}^{n} Y_{ni}\right| \geq \varepsilon\right] \leq \bar{Y} \sum_{n=1}^{\infty} n^{-2} + \bar{Y} \sum_{n=1}^{\infty} n^{-3} < \infty$$

and, by the Borel-Cantelli Lemma, the event $|n^{-1}\sum_{i=1}^{n} Y_{ni}| \geq \varepsilon$ occurs finitely often almost surely for any $\varepsilon > 0$, i.e. $n^{-1}\sum_{i=1}^{n} Y_{ni} \overset{as}{\to} 0$. Q.E.D.

The following provides a law of the "iterated" logarithm for triangular arrays.

Lemma 8. *Let Y_{ni} be a triangular array ($n \in \mathbb{N}$, $i = 1, \ldots, n$) of random variables, iid across $i = 1, \ldots, n$. If, for all $n \in \mathbb{N}$, $i = 1, \ldots, n$, $E[Y_{ni}] = 0$, $E[Y_{ni}^2] > 0$ and $E\left[|Y_{ni}|^{4+\delta}\right] \leq \bar{Y} < \infty$, then*

$$P\left[\limsup_{n\to\infty} \frac{|\sum_{i=1}^{n} Y_{ni}|}{\sqrt{2E[Y_{ni}^2]\, n \ln n}} \to 1\right] = 1. \tag{26}$$

Proof. We use Theorem 1 in Rubin and Sethuraman (1965), in the special case of iid variables across the i dimension, noting that our assumptions imply their Assumptions (7), (8), (9) and (11) for their N set to n and their constants q and c set to $q = 4 + \delta$ and $c^2 = 2 + \varepsilon$ for any $\varepsilon < \delta$. Their Theorem 1 then shows that

$$s_n := P\left[\left|\sum_{i=1}^{n} Y_{ni}\right| > c\sqrt{E[Y_{ni}^2]\, n \ln n}\right] = (1 + o(1))\, \frac{n^{-c^2/2}}{c\sqrt{2\pi \ln n}},$$

which can be used with the Borel-Cantelli Lemma. Indeed, the s_n for $c^2 = 2 + \varepsilon$ are such that $\sum_{n=2}^{\infty} s_n < \infty$ for any $\varepsilon > 0$ since

$$\sum_{n=2}^{\infty} \frac{n^{-1} n^{-\varepsilon/2}}{\left(\sqrt{2+\varepsilon}\right)\sqrt{2\pi \ln n}} \leq C \sum_{n=2}^{\infty} n^{-1-\varepsilon/2} < \infty$$

for some universal constant C and for any $\varepsilon > 0$. It follows that the event

$$\left\{n^{-1}\sum_{i=1}^{n} Y_{ni} > \sqrt{2+\varepsilon} E\left[Y_{ni}^2\right] n^{-1/2}\sqrt{\ln n}\right\}$$

occurs only finitely often for any $\varepsilon > 0$ arbitrarily close to 0. By a similar reasoning, $\sum_{n=2}^{\infty} s_n \to \infty$ for $\varepsilon < 0$ and that event occurs infinitely often for any $\varepsilon < 0$ arbitrarily

close to 0 and the conclusion (26) follows. (See also Theorem 3 in Hu and Weber (1992) for a similar use of this inequality, in a context where independence across n is also assumed, although it is not needed for the application of Theorem 1 in Rubin and Sethuraman (1965).) Q.E.D.

	our test						
n	no reg	$\varepsilon_n = 0.5$	$\varepsilon_n = 1$	optimal	Vuong	Shi	NP
			bivariate normal location				
100	0.000	0.041	0.045	0.037	0.000	0.000	
200	0.000	0.046	0.045	0.039	0.000	0.000	
500	0.000	0.039	0.037	0.038	0.000	0.000	
			misspecified normals				
100	$0.062^{\dagger}$	0.073	0.076	0.070	0.062	0.048	
200	$0.062^{\dagger}$	0.053	0.059	0.058	0.062	0.045	
500	$0.059^{\dagger}$	0.062	0.062	0.063	0.059	0.043	
			correctly specified normals				
100	0.003	0.035	0.039	0.026	0.003	0.000	
200	0.000	0.043	0.045	0.038	0.000	0.000	
500	0.000	0.036	0.034	0.035	0.000	0.000	
			nested regressions with one additional regressor				
100	0.001	0.039	0.044	0.042	0.001	0.000	
200	0.000	0.047	0.052	0.050	0.000	0.000	
500	0.000	0.056	0.056	0.056	0.000	0.000	
			nested regressions with two additional regressors				
100	0.008	0.049	0.050	0.048	0.006	0.000	0.063
200	0.003	0.049	0.049	0.048	0.002	0.001	0.054
500	0.002	0.059	0.058	0.059	0.002	0.000	0.045

Table 1: Null rejection probabilities (nominal size 0.05) of our, Vuong's, Shi's, and the Neyman Pearson ('NP') test for the different examples and different sample sizes ('n'). 'no reg', '$\hat{\varepsilon}_n = 0.5$', '$\hat{\varepsilon}_n = 1$', and 'optimal' refer to our test using $\hat{\varepsilon}_n = 0$, $\hat{\varepsilon}_n = 0.5$, $\hat{\varepsilon}_n = 1$, and the optimal epsilon defined in (4). "†" denotes the cases in which the unregularized "no reg" t-statistic is asymptotically N(0,1).

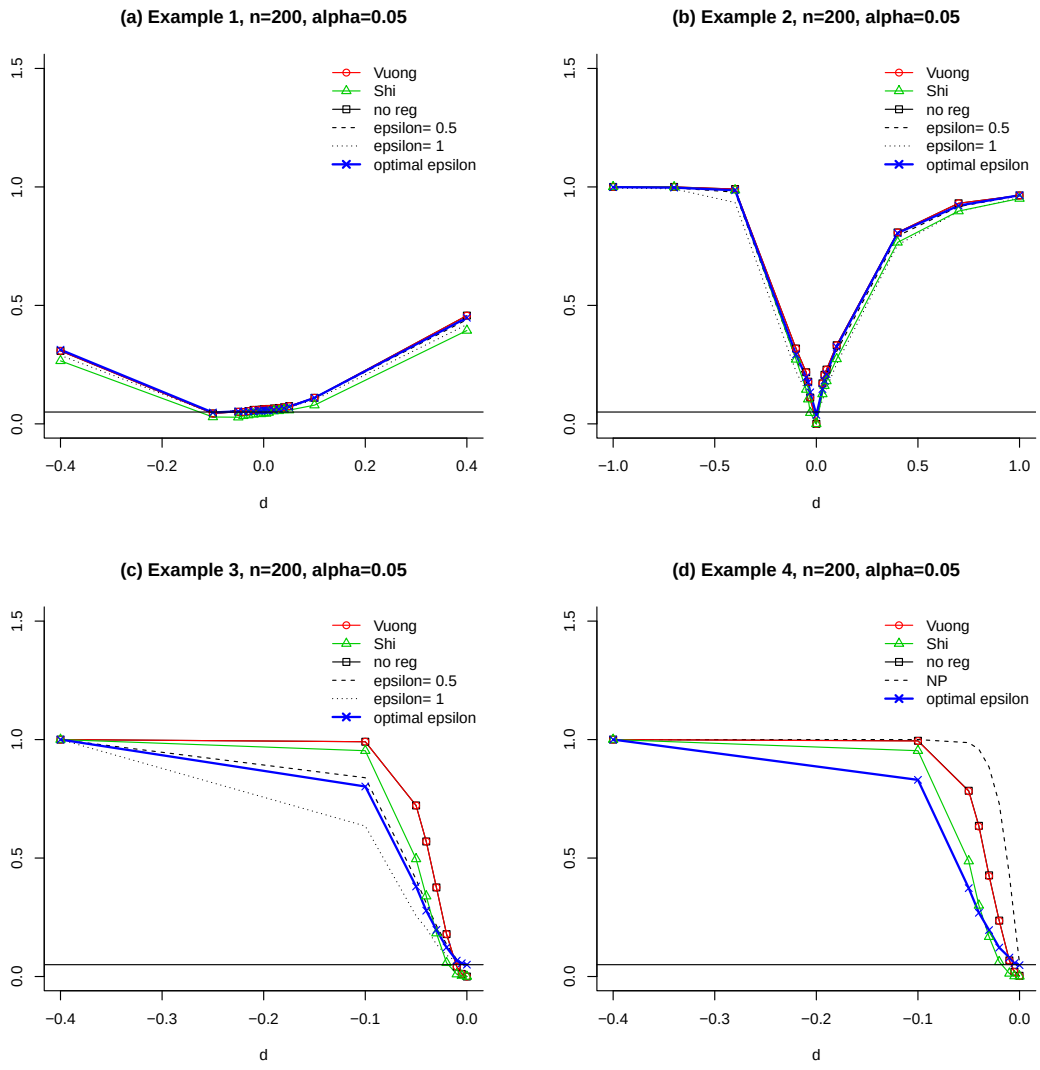

Figure 1: Rejection frequencies of Vuong's, Shi's, and our test. 'NP' refers to the Neyman-Pearson likelihood ratio test, and 'no reg' and 'optimal epsilon' to our test using $\hat{\varepsilon}_n = 0$ and the optimal epsilon in (4), respectively. On all graphs, the nominal level is marked by a black horizontal line.

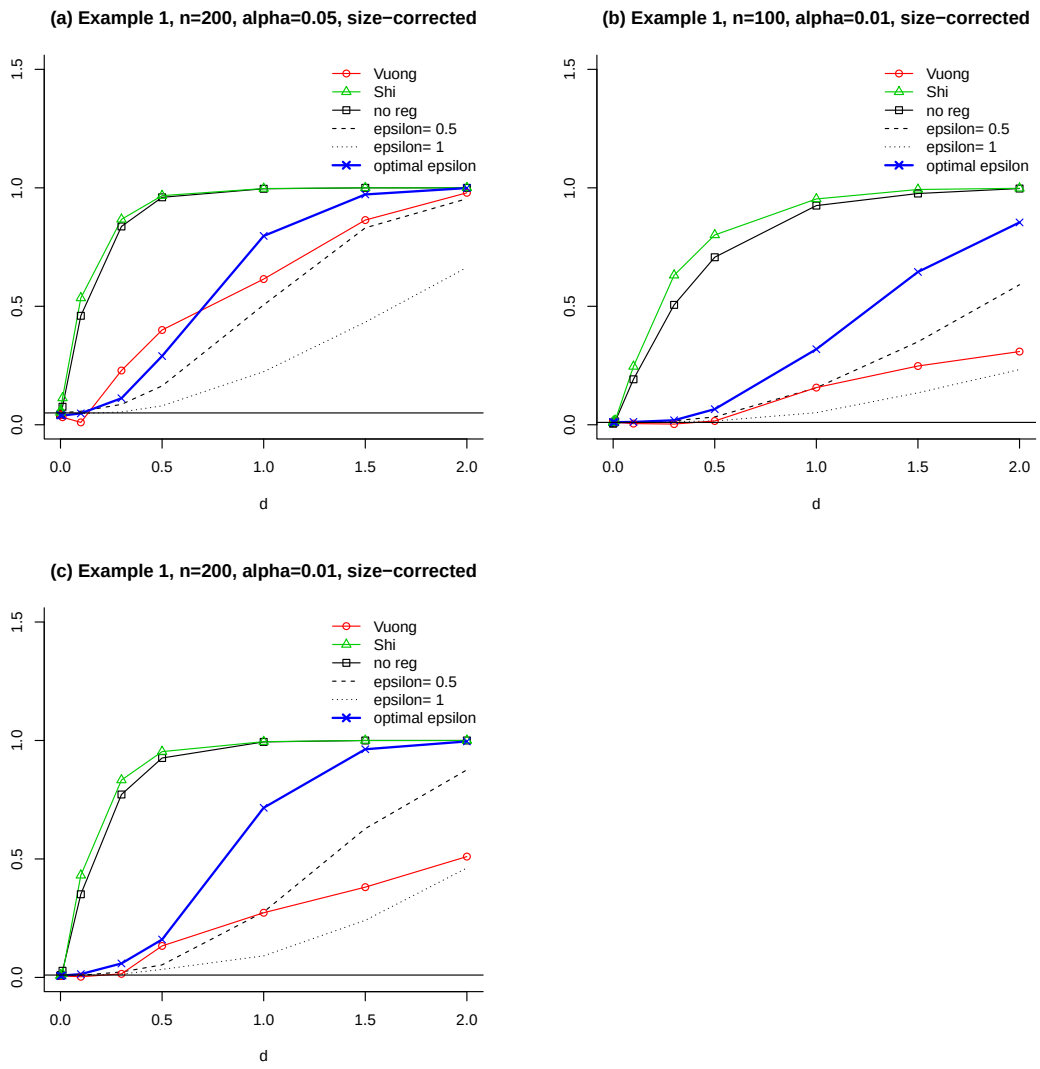

Figure 2: Size-corrected comparison of the rejection frequencies of the different tests considered. For Example 1, panels (a)-(c) report power curves for different confidence levels α and sample sizes n as function of the alternative model, indexed by d. On all graphs, the nominal level is marked by a black horizontal line.

References

Davidson, J. (1994): *Stochastic Limit Theory*. Oxford University Press.

Diebold, F. X., and R. S. Mariano (1995): "Comparing Predictive Accuracy," *Journal of Business & Economic Statistics*, 13(3), 253–263.

Hall, A. R., and D. Pelletier (2011): "Nonnested Testing in Models Estimated Via Generalized Method of Moments," *Econometric Theory*, 27(02), 443–456.

Hansen, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029–1054.

Hartman, P., and A. Wintner (1941): "On the Law of the Iterated Logarithm," *American Journal of Mathematics*, 63(1), 169–176.

Hu, T.-C., and N. C. Weber (1992): "On the rate of convergence in the strong law of Large numbers for arrays," *Bulletin of the Australian Mathematical Society*, 45, 479–482.

Lehmann, E. L., and J. P. Romano (2005): *Testing Statistical Hypotheses*. Springer, New York.

Newey, W. K., and D. McFadden (1994): "Large Sample Estimation and Hypothesis Testing," in *Handbook of Econometrics*, ed. by R. F. Engle, and D. L. McFadden, vol. IV, pp. 2111–2245. Elsevier Science B.V.

Newey, W. K., and R. J. Smith (2004): "Higher Order Properties of GMM and Generalized Empirical Likelihood Estimators," *Econometrica*, 72(1), 219–255.

Rivers, D., and Q. H. Vuong (2002): "Model Selection Tests for Nonlinear Dynamic Models," *Econometrics Journal*, 5, 1–39.

Romano, J. P., and A. F. Siegel (1986): *Counterexamples in Probability And Statistics*. CRC Press, New York.

Rubin, H., and J. Sethuraman (1965): "Probabilities of Moderate Deviations," *Sankhya A*, 27, 325–346.

Sin, C.-Y., and H. White (1996): "Information Criteria for Selecting Possibly Misspecified Parametric Models," *Journal of Econometrics*, 71(1-2), 207–225.

Tauchen, G. (1985): "Diagnostic Testing And Evaluation Of Maximum Likelihood Models," *Journal of Econometrics*, 30, 415–443.

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

Vuong, Q. H. (1989): "Likelihood Ratio Tests for Model Selection and Non-nested Hypotheses," *Econometrica*, 57(2), 307–333.