Proximal Statistics: Asymptotic Normality

David Pacini

Discussion Paper 19 / 718

September 2019

Department of Economics
University of Bristol
Priory Road Complex
Bristol BS8 1TU
United Kingdom

Proximal Statistics: Asymptotic Normality

David Pacini - University of Bristol

September 2019

Abstract. This note considers the problem of constructing an asymptotically normal statistic for the value function of a convex stochastic minimization programme, which may have more than one minimizer. It introduces the *proximal statistic* using a recursive estimator of one of the minimizers. The use of this statistic is illustrated by extending an existing selection test for point-identifying parametric models to the set-identifying case.

Keywords: Set Identification; Proximal algorithm

1. The Problem: Nonunique Global Minimizer

Consider the problem of constructing, from a random sample $\{z_i\}_{i=1}^n$, an asymptotically normal statistic for the real-valued parameter φ_o defined as

$$\varphi_o := \min_{q \in Q} F(q, P_o),$$

where Q is a known set, P_o is the unknown distribution of the random vector z_i taking values in $\mathbb{R}^L$, and $F(q, P_o) := \int f(q, z) dP_o(z)$ for a known function $q \mapsto f(q, z_i)$. The asymptotic normality requirement serves to simplify inference. This problem arises, for instance, in the context of selecting parametric statistical models (see e.g. Vuong, 1989). When $\arg\min_q F(q, P_o)$ exists and is unique, a solution is the plug-in statistic $\hat{\varphi}_n := F(\hat{q}_n, P_n)$, where $\hat{q}_n \in \arg\min_{q \in Q} F(q, P_n)$ and P_n is the empirical distribution function. Under a *Lipschitz-continuity* and an *envelope* condition on f, it is known (see e.g., Shapiro, Dentcheva, and Ruszczyinski, 2009, Theorem 5.7) that the sequence $n^{1/2}(\hat{\varphi}_n - \varphi_o)$ converges in distribution (denoted $\rightsquigarrow$) to a normal random variable $N\big(0, avar(\hat{\varphi}_n)\big)$ with mean zero and variance $avar(\hat{\varphi}_n) := E\big[[f(q_\star, z_i) - \varphi_o]^2\big]$ for $q_\star \in \arg\min_{q \in Q} F(q, P_o)$. When

$\arg\min_q F(q, P_o)$ is not unique, it is also known (see e.g., Shapiro et al., 2009, Theorem 5.7) that $n^{1/2}(\hat{\varphi}_n - \varphi_o) \rightsquigarrow \mathbb{G}_\star := \inf_{q \in Q_\star} \mathbb{G}_q$, where $Q_\star := \arg\min_{q \in Q} F(q, P_o)$ and $q \mapsto \mathbb{G}_q$ is a Gaussian process. The plug-in statistic is no longer a solution to the problem of interest because $\mathbb{G}_\star$ is not normal.

This note considers the case when $\arg\min_q F(q, P_o)$ may not be unique, $q \mapsto f(q, z_i)$ is a convex function a.e. z_i, and Q is a convex compact set. It investigates the following statistic.

Definition (Proximal Statistic). Define the proximal function

$$prox_P(v) := \arg\min_{q \in Q} F(q, P) + \frac{1}{2}\|q - v\|^2,$$

where $\|\cdot\|$ is the Euclidean norm. For $n > 8$, define $k_n := \lceil n^{1/3} \rceil$. Let $\hat{q}_{k_n}$ denote the last element in the sequence $\{\hat{q}_k\}_{k=2}^{k_n}$ defined recursively by

$$\hat{q}_{k+1} := \big(1 - k^{-1}\big) prox_n(\hat{q}_k), \text{ where } prox_n(\hat{q}_k) := prox_{P_n}(\hat{q}_k) \tag{1}$$

for an arbitrary starting value $\hat{q}_2 \in Q$. The *proximal statistic* is $\hat{\varphi}_{k_n} := F(\hat{q}_{k_n}, P_n)$. □

The proximal statistic, unlike the plug-in statistic, uses the recursive estimator $\hat{q}_{k_n}$. The recursive scheme (1), leading to $\hat{q}_{k_n}$, is a variant of the proximal algorithm.[1] The next proposition establishes sufficient conditions under which $\hat{\varphi}_{k_n}$ is asymptotically normal. The last section illustrates how this new result can assist in developing an asymptotically pivotal test for selecting between parametric set-identifying models.

2. Main Result

Proposition A (Asymptotic Normality). *Suppose that $\{z_i\}_{i=1}^n$ is i.i.d. P_o and*

(A.i) There is a function $m : \mathbb{R}^L \mapsto \mathbb{R}_+$ such that $|f(q, z_i) - f(\tilde{q}, z_i)| \leq m(z_i)\|q - \tilde{q}\|$ a.e. z_i

for all $q, \tilde{q} \in Q$*;*

(A.ii) There is a function $e : \mathbb{R}^L \mapsto \mathbb{R}$*, not depending on* q*, such that* $\sup_{q \in Q} |f(q, z_i)| \leq e(z_i)$ *a.e.* z_i *and* $E[\max(1, e(z_i), m(z_i))^2]$ *is finite;*

(A.iii) $q \mapsto f(q, z_i)$ *is a proper convex function a.e.* z_i*;*

(A.iv) $Q \subset \mathbb{R}^M$ *is the closed unit ball in* $\mathbb{R}^M$*;*

(A.v) $\sup_{v \in Q} \|prox_n(v) - prox_o(v)\| = O_{P_o}(n^{-1/2})$*, where* $prox_o(v) := prox_{P_o}(v)$*.*

Then,

$$n^{1/2}(\hat{\varphi}_{k_n} - \varphi_o) \rightsquigarrow N\big(0, avar(\hat{\varphi}_{k_n})\big),$$

where $avar(\hat{\varphi}_{k_n}) := E\big[[f(q_\star, z_i) - \varphi_o]^2\big]$ *and* $q_\star \in Q_\star$ *is the minimum-norm fixed point of* $v \mapsto prox_o(v)$*.*

The proof is given below. Assumptions (A.i)-(A.iv) are, respectively, the Lipschitz-continuity, envelope, and convexity restrictions announced in the introduction. (A.v) is a rate of convergence restriction on $prox_n$. These assumptions do not restrict $Q_\star$ to be a singleton.[2] Asymptotic normality follows from the result (see Lemma 3 below) that, even when there may be multiple minimizers, $\hat{q}_{k_n}$, unlike $\hat{q}_n$, converges in probability. When $Q_\star$ is a singleton, the proximal and plug-in statistics have the same asymptotic normal distribution, c.f., Proposition A with Shapiro et al. (2009, Theorem 5.7).

Proof of Proposition A. It is sufficient to verify that

(A.1) $X_n := n^{1/2}[F(q_\star, P_n) - \varphi_o] \rightsquigarrow X := N\Big(0, E\big[[f(q_\star, z_i) - \varphi_o]^2\big]\Big)$. This is an implication of Lemma 1 below.

(A.2) For $Y_n := n^{1/2}[F(\hat{q}_{k_n}, P_n) - \varphi_o]$, one has $X_n - Y_n \underset{P_o}{\rightarrow} 0$. Lemma 2 below establishes this result using Lemma 1 and Lemmas 3 to 5.

Then, from van der Vaart (1998, Theorem 2.7(iv)), it follows that $Y_n = n^{1/2}(\hat{\varphi}_{k_n} - \varphi_o) \rightsquigarrow X$.

$\diamondsuit$

Lemma 1. *For* $\mathcal{F} := \{f(q,\cdot) : q \in Q\}$*, define* $\ell^\infty(\mathcal{F}) := \{f \in \mathcal{F} : \sup_{q\in Q} |f(q,\cdot)| < \infty\}$*. Then,* $\mathbb{G}_n f(q) := n^{1/2}[F(q,P_n) - F(q,P_o)] \rightsquigarrow \mathbb{G}f(q)$ *in the space* $\ell^\infty(\mathcal{F})$*, where* $q \mapsto \mathbb{G}f(q)$ *is a Gaussian process with zero mean and covariance function* $q, \tilde{q} \mapsto E[f(q,z_i)f(\tilde{q},z_i)] - E[f(q,z_i)]E[f(\tilde{q},z_i)]$.

Proof. Let $H(\epsilon,\mathcal{F},P)$ denote the cover number of the family of functions $\mathcal{F}$.[3] Under A.i and A.iv, $\mathcal{F}$ is a *type II class* in the sense of Andrews (1994, p. 2270). It follows then from Andrews (1994, Theorem 2) that $v(z_i) := max(1, e(z_i), m(z_i))$ is such that $|f(q,z_i)| \leq v(z_i)$ $\forall f \in \mathcal{F}$ and the uniform entropy integral $\int_0^1 \sup_{P\in\mathcal{D}} \left[\ln H(\epsilon(Pv^2)^{1/2}, P, \mathcal{F})\right]^{1/2} d\epsilon$ satisfies

$$\int_0^1 \sup_{P\in\mathcal{D}} \left[\ln H(\epsilon(Pv^2)^{1/2}, P, \mathcal{F})\right]^{1/2} d\epsilon \leq \infty, \tag{1.1}$$

where $\mathcal{D}$ is the set of all discretely supported probability distributions. Rewrite A.ii as

$$P_o v^2 \leq \infty. \tag{1.2}$$

Since $\mathcal{F}$ is measurable under (A.i) and (A.ii), it follows from (1.1)-(1.2), by van der Vaart (1998, Theorem 19.14), that

$$\mathcal{F} \text{ is } P_o\text{-}\textit{Donsker}. \tag{1.3}$$

Conclude by restating the definition of P_o-*Donsker* class (van der Vaart, 1998, p.269). $\triangle$

Lemma 2. $n^{1/2}F(q_\star,P_n) - n^{1/2}F(\hat{q}_{k_n},P_n) \underset{P_o}{\rightarrow} 0.$

Proof. We first verify that $q \mapsto f(q,z_i)$ is square integrable at $q_\star$:

$$\lim_{q\to q_\star} \int |f(q,z_i) - f(q_\star,z)|^2 dP_o(z) = 0. \tag{2.1}$$

For any $q \in Q$, A.i implies $|f(q, z_i) - f(q_\star, z_i)|^2 \leq m(z_i)^2 \|q - q_\star\|^2$ because $|f(q, z_i) - f(q_\star, z_i)|$ is nonnegative. Taking expectations on both sides

$$\int |f(q, z_i) - f(q_\star, z)|^2 dP_o(z) \leq \int m(z)^2 dP_o(z) \|q - q_\star\|^2.$$

Under A.ii, $\int m(z)^2 dP_o(z) < \infty$. Hence, (2.1) follows from the last display after taking limits to both sides as $q \to q_\star$.

Define $g : \ell^\infty(\mathcal{F}) \times \mathcal{F} \mapsto \mathbb{R}$ by $g(h, f) := h(f) - h(f_\star)$, where $f_\star = f(q_\star, \cdot)$. The set $\mathcal{F}$ is a semimetric space relative to the $L_2(P_o)$-metric. The function g is continuous with respect to the product semimetric at every point (h, f) such that $f \mapsto h(f)$ is continuous. Indeed, if, for any sequence $\{h_k, f_k\}_k$ in $\ell^\infty(\mathcal{F}) \times \mathcal{F}$, $\{h_k, f_k\}_k \to (h, f)$, then $h_k \to h$ uniformly and hence $h_k(f_k) = h(f_k) + o(1) \to h(f)$ if h is continuous at f. By van der Vaart (1998, Lemma 18.15), it follows from (1.3) that almost all sample paths of $\mathbb{G}$ are uniformly continuous on $\mathcal{F}$. Thus, the function h is continuous at $f_\star \in \mathcal{F}$.

Set $f_n := f(\hat{q}_{k_n}, \cdot)$. Since $\hat{q}_{k_n} \underset{P_o}{\to} q_\star$ (Lemma 3), one has, by (2.1), that $f_n \underset{P}{\to} f_\star$ in the metric space $\mathcal{F}$. For $\mathbb{G}_n := n^{1/2}(P_n - P_o)$, by (1.3), $\mathbb{G}_n \rightsquigarrow \mathbb{G}$ in the space $\ell^\infty(\mathcal{F})$. Hence,

$$(f_n, \mathbb{G}_n) \rightsquigarrow (f_\star, \mathbb{G}) \text{ in the space } \mathcal{F} \times \ell^\infty(\mathcal{F}). \tag{2.2}$$

We have verified that g is continuous and (2.2) holds. Apply the Continuous Mapping Theorem (van der Vaart, 1998, Theorem 18.11(i)) to obtain

$$\mathbb{G}_n(f_n - f_\star) = g(\mathbb{G}_n, f_n) \rightsquigarrow g(\mathbb{G}, f_\star) = \mathbb{G}f_\star - \mathbb{G}f_\star = 0.$$

Since convergence in probability and convergence in distribution are the same for a degenerate limit (van der Vaart, 1998, Theorem 18.10(iii)), $\mathbb{G}_n(f_n - f_\star) \underset{P_o}{\to} 0$. Conclude by replacing $\mathbb{G}_n(f_n - f_\star)$ by its definition in $-\mathbb{G}_n(f_n - f_\star) = n^{1/2}F(q_\star, P_n) - n^{1/2}F(\hat{q}_{k_n}, P_n)$. $\triangle$

Lemma 3. *$\|\hat{q}_{k_n} - q_\star\| \underset{P_o}{\to} 0$, where $q_\star \in Q_\star$ is the minimum-norm fixed point of $v \mapsto prox_o(v)$.*

Proof. Use the triangle inequality to bound $\|\hat{q}_{k_n+1} - q_\star\|$ by the sum of a deterministic and a stochastic term $\|\hat{q}_{k_n+1} - q_\star\| \leq \|q_{k_n+1} - q_\star\| + \|\hat{q}_{k_n+1} - q_{k_n+1}\|$. Consider the deterministic term. By Lemma 5, $\|q_{k_n+1} - q_\star\| = o(1)$. Consider now the stochastic term. Replacing $\hat{q}_{k_n+1}$ and q_{k_n+1} recursively,

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| = \|a_{k_n} prox_n(\hat{q}_{k_n}) - a_{k_n} prox_o(q_{k_n})\|.$$

Add-and-subtract $prox_o(\hat{q}_{k_n})$ and use the triangle inequality to get

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq a_{k_n}\|prox_n(\hat{q}_{k_n}) - prox_o(\hat{q}_{k_n})\| + a_{k_n}\|prox_o(\hat{q}_{k_n}) - prox_o(q_{k_n})\|.$$

Since $v \mapsto prox_o(v)$ is nonexpansive (Lemma 4),

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq a_{k_n}\|prox_n(\hat{q}_{k_n}) - prox_o(\hat{q}_{k_n})\| + a_{k_n}\|\hat{q}_{k_n} - q_{k_n}\|.$$

By recursive substitution,

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq \frac{a_{k_n}}{1 - a_{k_n}}\|prox_n(\hat{q}_{k_n}) - prox_o(\hat{q}_{k_n})\|.$$

Since $\|prox_n(\hat{q}_{k_n}) - prox_o(\hat{q}_{k_n})\| \leq \sup_{q\in Q}\|prox_n(q) - prox_o(q)\|$,

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq \frac{a_{k_n}}{1 - a_{k_n}} \sup_{q\in Q}\|prox_n(q) - prox_o(q)\|.$$

Since we have assumed that $n^{1/2}[prox_n - prox_o]$ is asymptotically tight (see A.v), one has

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq \frac{a_{k_n}}{1 - a_{k_n}} O_{P_o}(n^{-1/2}).$$

Since $a_{k_n}/(1-a_{k_n}) = k_n - 1$ and we have assumed $k_n = o(n^{1/2})$,

$$\|\hat{q}_{k_n+1} - q_{k_n+1}\| \leq o(n^{1/2}) O_{P_o}(n^{-1/2}) \leq n^{1/2-1/2} o(1) O_{P_o}(1) \leq o_{P_o}(1).$$

Conclude then $\|\hat{q}_{k_n} - q_\star\| \leq o(1) + o_{P_o}(1) \leq o_{P_o}(1)$. △

Lemma 4. *$v \mapsto prox_o(v)$ is nonexpansive:*

$$\|prox_o(v) - prox_o(\tilde{v})\| \leq \|v - \tilde{v}\| \text{ for any } v, \tilde{v} \in Q.$$

Proof. (A.iii) implies that $q \mapsto F(q, P_o)$ is a proper convex function. Conclude then, from Moreau (1965, *Proposition 5.b.*), that $v \mapsto prox_o(v)$ is nonexpansive. △

Lemma 5. *Let $a_{k_n} := 1 - k_n^{-1}$. Define $q_{k_n+1} := a_{k_n} prox_o(q_{k_n})$ for an arbitrary starting point $q \in Q$. q_{k_n} converges to $q_\star \in Q_\star$ for $q_\star$ the minimum-norm fixed point of $v \mapsto prox_o(v)$:*

$$\|q_{k_n} - q_\star\| = o(1).$$

Proof. Since $prox_o : Q \mapsto Q$ is nonexpansive (Lemma 4) and Q is the closed unit ball in a Hilbert space (see A.iv), $a_{k_n} := 1 - k_n^{-1} = 1 - \lceil n^{-1/3} \rceil$ is *acceptable* in the sense of Halpern (1967, Corollary p. 961), viz. $\|q_{k_n} - q_\star\| = o(1)$, where $q_\star$ is the fixed point of $v \mapsto prox_o(v)$ with the smallest norm. Since the fixed points of $v \mapsto prox_o(v)$ belong to $Q_\star$, one has $q_\star \in \arg\min_{q \in Q} F(q, P_o)$. △

If $q_\star$ also belongs to $Q_{\star\star} := \arg\min_{q \in Q_\star} E\big[[f(q_\star, z_i)^2\big]$, $\hat{\varphi}_{k_n}$ has minimum asymptotic variance. A sufficient condition for $\hat{\varphi}_{k_n}$ having this property is

Corollary. *If $q \mapsto E[f(q, z_i)^2]$ is convex, then $avar(\hat{\varphi}_{k_n}) = \min_{q_\star \in Q_\star} E\big[[f(q_\star, z_i) - \varphi_o]^2\big]$.*

One could also construct a minimum asymptotic variance statistic by using the iteration $q_{k+1} = (1-a_k)\hat{q}_{\star\star} + a_k prox_n(\hat{q}_k)$ for any consistent estimator $\hat{q}_{\star\star}$ of $q_{\star\star} \in \arg\min_{q \in Q_\star} E[f(q, z_i)^2]$.

3. Illustration: A Model Selection Test under Loss of Point-Identification

This Section illustrates the proximal statistic in the context of extending Vuong (1989) selection test from non-nested point-identifying models to the set-identifying case. Let p_o denote the density associated to P_o. For modeling p_o, consider the families of parametric density functions, from now so-called the models, $\mathcal{G} := \{z \mapsto g(z, \theta) : \theta \in \Theta \subset \mathbb{R}^{\dim(\theta)}\}$ and $\mathcal{H} := \{z \mapsto h(z, \gamma) : \gamma \in \Gamma \subset \mathbb{R}^{\dim(\gamma)}\}$. The functions $z \mapsto g(z, \theta)$ and $z \mapsto h(z, \gamma)$ are known up to the parameters θ and γ, respectively. The aim is to choose the model that is 'closest' to p_o. Consider the Kullback-Liebler information criterion defined as

$$KLIC_o(\mathcal{G}) := \int \ln p_o(z) dP_o(z) - \min_{\theta \in \Theta} G(\theta, P_o),$$

where $G(\theta, P_o) := \int -\ln g(z, \theta) dP_o(z)$. A similar definition follows for $KLIC_o(\mathcal{H})$. $KLIC_o(\mathcal{G})$ is nonnegative and is equal zero if and only if $p_o(z_i) = g(z_i, \theta_\star)$ a.e. z_i for $\theta_\star \in \arg\min G(\theta, P_o)$. Define $\rho_o := KLIC_o(\mathcal{H}) - KLIC_o(\mathcal{G}) = \min_{\theta \in \Theta} G(\theta, P_o) - \min_{\gamma \in \Gamma} H(\gamma, P_o)$. Consider the following hypotheses and definitions:

$H_0 : \rho_o = 0$, meaning that $\mathcal{G}$ and $\mathcal{H}$ are *equivalent*.

$H_{\mathcal{G}} : \rho_o > 0$, meaning that $\mathcal{G}$ is *better* than $\mathcal{H}$.

$H_{\mathcal{H}} : \rho_o < 0$, meaning that $\mathcal{G}$ is *worse* than $\mathcal{H}$.

These definitions do not require that either model is point-identifying the model's parameter (i.e., there may be $\theta_o \neq \tilde{\theta}$ such that $g(z_i, \theta_o) = g(z_i, \tilde{\theta}) = p_o(z_i)$ a.e. z_i).

When $\arg\max_\theta G(\theta, P_o)$ and $\arg\max_\gamma H(\gamma, P_o)$ are unique, it is known (see Vuong, 1989, Theorem 5.1) that, if the models are non-nested, the $n^{1/2}$-scaled version of the plug-in statistic $\rho_n := \min_{\theta \in \Theta} G(\theta, P_n) - \min_{\gamma \in \Gamma} H(\gamma, P_n)$ is, under H_0, asymptotically normal and, under

$H_{\mathcal{G}}$ (res. $H_{\mathcal{H}}$), diverges to $+\infty(-\infty)$. When $\arg\max_\theta G(\theta, P_o)$ and/or $\arg\max_\gamma H(\gamma, P_o)$ are not unique, the asymptotic distribution of $n^{1/2}\rho_n$, under H_0, is the difference between the infima of two Gaussian processes.[4] This asymptotic distribution is not normal. To construct an asymptotic normal statistic, let $\hat{\varphi}_{g,k_n} := G(\hat{\theta}_{k_n}, P_n)$ denote the proximal statistic for $\varphi_{go} := \min_{\theta\in\Theta} G(\theta, P_o)$. Let $avar_n(\hat{\varphi}_{g,k_n})$ denote the plug-in estimator for the asymptotic variance $avar(\hat{\varphi}_{g,k_n}) := E\big[(\ln g(z_i, \theta_\star))^2\big] - E\big[\ln g(z_i, \theta_\star)\big]^2$. Similarly, define $\hat{\varphi}_{h,k_n}$, $\hat{\gamma}_{k_n}$, $avar_n(\hat{\varphi}_{h,k_n})$, and $acov_n(\hat{\varphi}_{g,k_n}, \hat{\varphi}_{h,k_n})$. Define the test statistic $\hat{\rho}_{k_n} := \hat{\varphi}_{g,k_n} - \hat{\varphi}_{h,k_n}$ and the standard deviation estimator $\hat{\omega}_n := [avar_n(\hat{\varphi}_{g,k_n}) + avar_n(\hat{\varphi}_{h,k_n}) - 2acov_n(\hat{\varphi}_{g,k_n}, \hat{\varphi}_{h,k_n})]^{1/2}$.

Proposition B (Model Selection Test for Strictly Non-Nested Models). *Suppose that $\{z_i\}_{i=1}^n$ is i.i.d. P_o and*

(B.i) There exists $m : \mathbb{R}^L \mapsto \mathbb{R}$ such that $|\ln g(z_i, \theta) - \ln g(z_i, \tilde{\theta})| \leq m(z_i)\|\theta - \tilde{\theta}\|$ a.e. z_i;

(B.ii) There exists $e : \mathbb{R}^L \mapsto \mathbb{R}$ such that $\sup_{\theta\in\Theta} |g(z_i, \theta)| \leq e(z_i)$ a.e. z_i and $E[\max(1, e(z_i), m(z_i))^4] < \infty$;

(B.iii) $\theta \mapsto \ln g(z_i, \theta)$ is a proper concave function a.e. z_i;

(B.iv) Θ is a compact convex set;

(B.v) $\sup_{v\in\Theta} \|prox_{gn}(v) - prox_{go}(v)\| = O_{P_o}(n^{-1/2})$, where $prox_{go}(v) := \arg\min_{\theta\in\Theta} G(\theta, P_o) + 1/2\|\theta - v\|^2$;

(B.vi) If Θ and $g(z_i, \cdot)$ are, respectively, replaced by Γ and $h(z_i, \cdot)$, (B.i) to (B.v) hold;

(B.vii) $\mathcal{G} \cap \mathcal{H} = \emptyset$.

Then, under H_0, $n^{1/2}\hat{\rho}_{k_n}/\hat{\omega}_n \rightsquigarrow N(0, 1)$; under $H_{\mathcal{G}}$, $n^{1/2}\hat{\rho}_{k_n}/\hat{\omega}_n \underset{P_o}{\rightarrow} +\infty$; and under $H_{\mathcal{H}}$, $n^{1/2}\hat{\rho}_{k_n}/\hat{\omega}_n \underset{P_o}{\rightarrow} -\infty$.

Proposition B extends to set-identifying models a result in Vuong (1989, Theorem 5.1). It provides an asymptotically pivotal selection test for the models. One chooses a critical value c from the standard normal distribution for some significance level. If the realized value v of the statistic $n^{1/2}\hat{\rho}_{k_n}/\hat{\omega}_n$ is higher than c, then one rejects the null hypothesis that the models are equivalent in favor of $\mathcal{G}$. If v is smaller than $-c$, then one rejects the null hypothesis that the models are equivalent in favor of $\mathcal{H}$. Finally, if the absolute value of v is smaller than c, one cannot discriminate between the two models given the data. When both models are point-identifying, this test is asymptotically equivalent to the Vuong test.[5]

We decompose the proof of Proposition B in three Lemmas.

Lemma 6. *$n^{1/2}(\hat{\varphi}_{gk_n} - \varphi_{go}) \rightsquigarrow N(0, avar(\hat{\varphi}_{gk_n}))$ and $n^{1/2}(\hat{\varphi}_{hk_n} - \varphi_{ho}) \rightsquigarrow N(0, avar(\hat{\varphi}_{hk_n}))$.*

Proof. Under (B.i)-(B.v), we are justified to set $q = \theta$, $Q = \Theta$, $f(q, z_i) = -\ln g(z_i, \theta)$, $F(q, P_o) = G(\theta, P_o)$, etc. It follows then from Proposition A that $n^{1/2}[\hat{\varphi}_{g,k_n} - \varphi_{g,k_n}] \rightsquigarrow N(0, avar(\hat{\varphi}_{gk_n}))$. A similar reasoning yields $n^{1/2}(\hat{\varphi}_{hk_n} - \varphi_{ho}) \rightsquigarrow N(0, avar(\hat{\varphi}_{hk_n}))$. △

Lemma 7. *$\hat{\omega}_n \underset{P_o}{\rightarrow} \omega_o$.*

Proof. Define $avar_n(\theta) := n^{-1}\sum_{i=1}^n \ln g(z_i, \theta)^2 - \left[n^{-1}\sum_{i=1}^n \ln g(z_i, \theta)\right]^2$ and $avar(\theta) := E[\ln g(z_i, \theta)^2] - E[\ln g(z_i, \theta)]^2$. By the triangle inequality with probability approaching one

$$|avar_n(\hat{\varphi}_{k_n}) - avar(\hat{\varphi}_{k_n})| \leq |avar_n(\hat{\theta}_{k_n}) - avar(\hat{\theta}_{k_n})| + |avar(\hat{\theta}_{k_n}) - avar(\theta_\star)|.$$

Consider the first term in the right hand side of this inequality. From B.i, B.ii and the i.i.d. assumption, $\sup_\theta |avar_n(\theta) - avar(\theta)| = o_{P_o}(1)$. Hence, $|avar_n(\hat{\theta}_{k_n}) - avar(\hat{\theta}_{k_n})| = o_{P_o}(1)$. Consider now the second term. From (B.i), $\theta \mapsto avar(\theta)$ is continuous. Since $\hat{\theta}_{k_n} \underset{P_o}{\rightarrow} \theta_\star$, by the Continuous Mapping Theorem, $|avar(\hat{\theta}_{k_n}) - avar(\theta_\star)| = o_{P_o}(1)$. It follows then that $avar_n(\hat{\varphi}_{k_n}) \underset{P_o}{\rightarrow} avar(\hat{\varphi}_{k_n})$. A similar result follows for $avar_n(\hat{\varphi}_{hk_n})$ and $acov_n(\hat{\varphi}_{gk_n}, \hat{\varphi}_{hk_n})$. Then, by the Continuous Mapping Theorem, $\hat{\omega}_n \underset{P_o}{\rightarrow} \omega_o$. △

Lemma 8. *(B.vii) implies $\omega_o > 0$.*

Proof. It suffices to verify that $\omega_o = 0$ iff $g(z_i, \theta_\star) = h(z_i, \gamma_\star)$ for any $\theta_\star \in \arg\min_{\theta\in\Theta} G(\theta, P_o)$, $\gamma_\star \in \arg\min_{\gamma\in\Gamma} H(\gamma, P_o)$. Fix $\theta_\star$ and $\gamma_\star$. From the definition of ω_o, we have $\omega_o = 0$ iff there exists a constant ϵ such that $g(z_i, \theta_\star) = \epsilon h(z_i, \gamma_\star)$ a.e. z_i. Since $z \mapsto g(z, \theta_\star)$ and $z \mapsto h(z, \gamma_\star)$

are density functions, they integrate to one. It follows then, by integrating both sides of $g(z, \theta_\star) = \epsilon h(z, \gamma_\star)$ with respect to z, that $\epsilon = 1$. $\triangle$

Proof of Proposition B. $n^{1/2}(\hat{\varphi}_{gk_n} - \varphi_{go}) - n^{1/2}(\hat{\varphi}_{hk_n} - \varphi_{ho}) = n^{1/2}\hat{\rho}_{k_n} - n^{1/2}\rho_o$. Under H_0, $n^{1/2}\rho_o = 0$. Lemma 8 justifies to use Slutzky's Lemma (van der Vaart, 1998, Lemma 2.8 (iii)) to combine Lemmas 6 and 7 to obtain $n^{1/2}\hat{\rho}_{k_n}/\hat{\omega}_n \rightsquigarrow N(0,1)$. The claim for $n^{1/2}\hat{\rho}_{k_n}/\omega_n$ under $H_{\mathcal{G}}$ and $H_{\mathcal{H}}$ follows similarly. $\diamondsuit$

Endnotes

[1]For an exposition on the proximal algorithm, see e.g., Polson, Scott and Willard (2015).

[2]Assumption (A.iv) can be relaxed, at the cost of loosing conciseness in the exposition, to Q being a closed convex subset of $\mathbb{R}^M$ and proving Lemma 5 below by verifying the conditions in Bauschke and Combettes (2017, Theorem 30.1). Possible extensions to Proposition A include studying: (a) the conditions under which the convergence in distribution also holds uniformly; the properties of the proximal statistic when (b) $q \mapsto f(q, z_i)$ is strongly amenable; (c) $1/2\|q - \hat{q}_k\|^2$ is replaced by another Bregman divergence; (d) Q is defined by moment inequality restrictions. These extensions are out of the scope of this note.

[3]For a definition, see Andrews (1994 p. 2268) or van der Vaart (1998, p. 274).

[4]This follows from applying Shapiro et al. (2009, Theorem 5.7)

[5] The following extensions to Proposition B are out of the scope of this note. First, the asymptotic approximation in Proposition B is pointwise in P_o. The development of a uniform asymptotic approximation is needed. Second, one could compare more than two models using multiple testing methods. Third, one could apply Proposition A to a test based on a goodness-of-fit criteria other than the KLIC.

References

Andrews, D. (1994): "Empirical Processes in Econometrics", *Handbook of Econometrics*, Vol. 4, Elsevier, Amsterdam.

Bauschke, H. and P. Combettes (2017): *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*, Springer.

Halpern, B. (1967) "Fixed Points of Nonexpanding Maps", *Bulletin of the American Mathematical Society.*

Moreau, J.-J. (1965): "Proximite et Dualite dans un Espace Hilbertien", *Bulletin de la Societe Mathematique de France*.

Polson, N., J. Scott and B. Willard (2015): "Proximal Algorithms in Statistics and Machine Learning", *Statistical Science*.

Shapiro, A., D. Dentcheva, and A. Ruszczynski (2009): *Lecture Notes on Stochastic Programming: Modeling and Theory*, SIAM, Philadelphia.

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

Vuong, Q. (1989): "Likelihood Ratio Tests for Model Selection and Non-Nested Hypotheses", *Econometrica*.