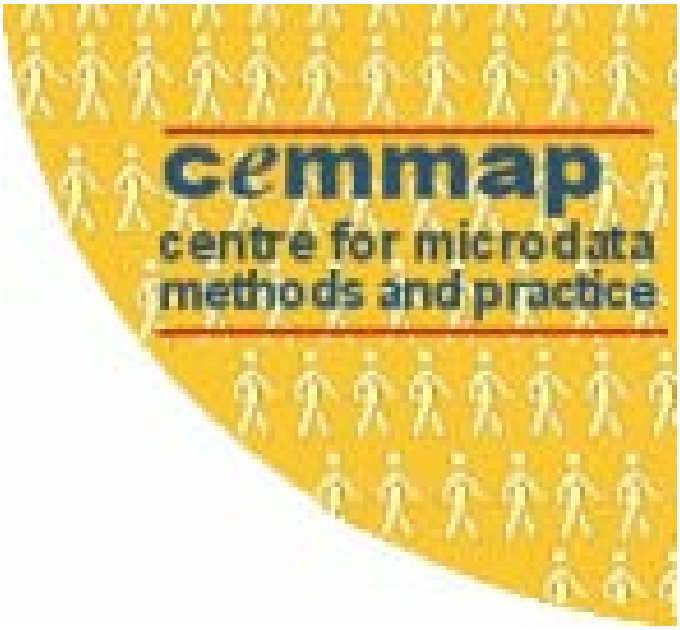

MAXIMAL UNIFORM CONVERGENCE RATES IN PARAMETRIC ESTIMATION PROBLEMS

Walter Beckert
Daniel L. McFadden

THE INSTITUTE FOR FISCAL STUDIES
DEPARTMENT OF ECONOMICS, UCL
cemmap working paper CWP06/05

Maximal Uniform Convergence Rates in Parametric Estimation Problems

Walter Beckert and Daniel L. McFadden*

Draft - June 2005

Abstract

This paper considers parametric estimation problems with i.i.d. data. It focusses on rate-efficiency, in the sense of maximal possible convergence rates of stochastically bounded estimators, as an optimality criterion, largely unexplored in parametric estimation. Under mild conditions, the Hellinger metric, defined on the space of parametric probability measures, is shown to be an essentially universally applicable tool to determine maximal possible convergence rates.

JEL Classification: C13, C16

Keywords: parametric estimators, uniform convergence, Hellinger distance, Locally Asymptotically Quadratic (LAQ) Families

Correspondence: w.beckert@bbk.ac.uk, Walter Beckert, School of Economics, Mathematics and Statistics, Birkbeck College, Malet Street, London WC1E 7HX, UK.

*The first author is at Birkbeck College, University of London, and the Institute for Fiscal Studies; the second author is Morris Cox Professor of Economics at UC Berkeley. We are indebted to Masafumi Akahira, Richard Blundell, Andrew Chesher and Hide Ichimura for helpful comments and discussions.

1 Introduction

A general aim in estimation problems is to obtain estimators that converge to the target of estimation as fast as possible, in this sense making maximal or rate efficient use of the data. The econometric literature has primarily emphasized efficiency in terms of minimal asymptotic variance-covariance matrices among estimators of a given convergence rate. This has notably been the case in the context of regular parametric, $\sqrt{n}$ convergent estimators, where the Cramér-Rao lower bound is the gold-standard for efficiency. Convergence rate efficiency is a paramount concern in non-parametric estimation problems (e.g. Stone (1980, 1982)), but has received comparatively little attention in parametric problems. This is a reflection of the fact that a very broad class of parametric problems that are sufficiently well-behaved have a best rate of $\sqrt{n}$, independently of the dimension of the parameter space or the degree of smoothness of the probability law. However, even in textbook parametric models there are exceptions, such as the scale parameter $\theta > 0$ in the uniform distribution on $[0, \theta]$. This paper provides a gold-standard for rate-efficiency in general, non-regular parametric estimation problems, i.e. an upper bound on the rates of convergence of parametric estimators.

The analysis in this paper builds on the Hellinger metric on the space of parametric densities.[1] This metric is distinguished by a number of useful properties, especially for product measures in the case of i.i.d. samples. It has been used in related work by Ibragimov and Has'minskii (1981). Provided the Hellinger distance for any two parametrizations in a given parametric family has a Hölder continuity property, their main result yields an upper bound on the uniform L^1 convergence rate of parametric estimators. This result is unsatisfactory for at least four reasons. First, it makes assumptions on the Hellinger distance on the space of densities belonging to a parametric family, rather than directly on the underlying parametric family. In particular, this assumption does not illuminate under what conditions on the parametric family the resulting rate does or does not depend on the parameter value to be estimated. Second, the notion of L^1 convergence requires parametric estimators to be integrable. This is a limitation as it does not cover a large class of estimators. Notably, estimators in locally asymptotically quadratic (LAQ) problems can typically only be shown to be stochastically bounded, when scaled appropriately (see, e.g., LeCam (1986), LeCam and Yang (2000), Hajék (1970)). Third, as a consequence of the Lipschitz assumption on the Hellinger distance, upper bounds on L^1 convergence rates turn out to be powers of n. This excludes cases in which rate-efficient estimators are

[1] Essentially all the derivations hold for more general probability measures.

known to converge at logarithmic rates (see, e.g., LeCam and Yang (2000), Prakasa Rao (1968)). Fourth, the statement of the result is tacit about identification requirements.

Hellinger distance as a metric for convergence has been considered in the context of maximum likelihood (ML) estimation. Van de Geer (1993, 2000) establishes rates of Hellinger consistency of ML estimators under entropy conditions, drawing on the theory of empirical processes (Pollard (1984, 1989)). Entropy-based rates of Hellinger consistency are not guaranteed to be optimal, however, since entropy, as a measure of the complexity of the set of densities to which the target density belongs, provides an upper bound on squared Hellinger distance and, hence, not a sharp bound on the best possible rate.[2] Moreover, the invoked entropy conditions embed a uniform envelope or dominance condition on the set of densities. This precludes non-regular cases from the analysis.

A result closely related to this paper is due to Akahira (1991) and Akahira and Takeuchi (1995). These authors show for the case of location parameters in general non-regular models that a maximum bound on the convergence rate of parametric estimators can be deduced from the absolute variation metric, which in turn can be bounded by functions of the Hellinger metric. Their result can be viewed as a special case of the main result of this paper which covers a wider class of parametric estimation problems.

Non-regular estimation problems have received increasing interest as they arise in the applied literature on auctions (Paarsch (1992)) and the literature on threshold regression models (Chan (1993), Chan and Tsay (1998), Hansen (2000), Seo and Linton (2005)). Hirano and Porter (2003) consider efficient estimation in a class of non-regular models - in the sense of their limit experiments (LeCam (1986)) not being locally asymptotically normal - which can be approximated by simpler limit models for which there exists an estimator which has the same distribution as the estimator in the original non-regular model. In the limit models considered, the data come from a distribution known up to an additive shift, an idea due to LeCam (1972). Hirano and Porter (2003) employ this idea to examine locally shifted ML estimators to achieve asymptotic efficiency.

The analysis in this paper employs arguments based on the Hellinger distance. The rate at which the distance between two parameter values converges to zero such that the Hellinger distance converges to an interior limit, henceforth referred to as the Hellinger rate, plays an central role in this analysis. The paper gives necessary and sufficient conditions under which the Hellinger rate does not depend on the parameter value to be estimated. And, under such conditions, it is shown that the Hellinger rate is an upper bound on uniform convergence rates of estimators which are stochastically bounded.

[2]See, for example, Van de Geer (2000), example 7.4.6., and Birgé and Massart (1993).

2 The Hellinger Metric

Let $\mathcal{Y}$ denote the real-valued Euclidean sample space of random variables y, and $\sigma(y)$ the Borel σ-field generated by y. Denote by $\{F(y;\theta), \theta \in \Theta\}$ the parametric family of probability measures on $\sigma(y)$, where Θ is a compact parameter space. In what follows, the scalar case $\Theta \subset \mathbb{R}$ will be considered.[3] Suppose further that $(\mathcal{Y}, \sigma(y))$ is a σ-finite measurable space, $F(y;\theta)$ is absolutely continuous with respect to Lebesgue measure, and $f(y;\theta)$ is the Radon-Nikodym derivative of $F(y;\theta)$.

Let $h^2(\theta,\theta') = \frac{1}{2}\int_y \left(\sqrt{f(y;\theta)} - \sqrt{f(y;\theta')}\right)^2 dy$ denote the squared Hellinger distance of the parametric densities $f(y;\theta)$ and $f(y;\theta')$, $\theta,\theta' \in \Theta$. Let $H_n^2(\theta,\theta')$ denote the squared Hellinger distance of the densities, evaluated at θ and θ', respectively, of the i.i.d. sample $\{y_i, i = 1,\ldots,n\}$.

The Hellinger metric is of interest because it enjoys a number of convenient properties.

1. Let $\rho(\theta,\theta') = \int_y \sqrt{f(y;\theta)f(y;\theta')}dy$ denote the affinity between the densities $f(y;\theta)$ and $f(y;\theta')$ (see also Matusita (1955)). Then,

$$\begin{aligned}\rho(\theta,\theta') &= \int_y f(y;\theta)\exp\left(\frac{1}{2}\ln\left(\frac{f(y;\theta')}{f(y;\theta)}\right)\right)dy \\ &= E_\theta\left[\exp\left(\frac{1}{2}\ln\left(\frac{f(y;\theta')}{f(y;\theta)}\right)\right)\right],\end{aligned}$$

where $E_\theta[\cdot]$ denotes expectation with respect to $f(y;\theta)$, and it follows that

$$h^2(\theta,\theta') = 1 - \rho(\theta,\theta').$$

For i.i.d. data,

$$\begin{aligned}H_n^2(\theta,\theta') &= 1 - E_\theta\left[\exp\left(\frac{1}{2}\sum_{i=1}^n \ln\left(\frac{f(y_i;\theta')}{f(y_i;\theta)}\right)\right)\right] \\ &= 1 - \left(E_\theta\left[\exp\left(\frac{1}{2}\ln\left(\frac{f(y;\theta')}{f(y;\theta)}\right)\right)\right]\right)^n \\ &= 1 - \rho(\theta,\theta')^n.\end{aligned}$$

Hence, $H_n^2(\theta,\theta') \in [0,1]$ for any θ,θ' and any n, and the squared Hellinger distance for i.i.d. data involves a factorization of affinities.[4]

[3] The vector case can be thought of in analogous terms, provided all parameters converge at the same rate. In the vector case with different rates for each vector component, the analysis in this manuscript essentially covers the case of θ being a linear combination of these, and rates determined by the least rapidly converging subcomponent.

[4] Akahira and Takeuchi (1991) define an information measure based on Hellinger affinity, $I_n(\theta,\theta') = -8\ln\rho(\theta,\theta')^n$. This measure is interpreted as the information between the product measures of the i.i.d. sample, parameterized by θ and θ', respectively.

Notice that $\theta \neq \theta'$ implies that $\lim_n H_n^2(\theta, \theta') = 1$. Strictly speaking, this requires a notion of identification of θ. In this setup, this can be formulated as follows: $\theta_0 \in \Theta$ is identified if, for any $\theta \in \Theta$, $\rho(\theta_0, \theta) = 1$ is equivalent to $\theta = \theta_0$.

With this notion of identification, it is clear that the Hellinger distance has all the properties of a metric on the space of root densities.[5]

2. Among the most frequently used measures on the space of densities is the Kullback-Leibler divergence,

$$KL(\theta, \theta') = E_\theta \left[\ln \left(\frac{f(y;\theta)}{f(y;\theta')} \right) \right],$$

and $KL_n(\theta, \theta')$ for an i.i.d. sample obtained as a sum of such divergence measures.[6] Hellinger distance and Kullback-Leibler divergence are related by

$$H_n^2(\theta, \theta') \leq 1 - \exp\left(-\frac{1}{2} KL_n(\theta, \theta') \right).$$

Therefore, convergence of the Kullback-Leibler divergence implies convergence of the Hellinger distance, but not vice versa. Note that the Hellinger distance is always well-defined, while the Kullback-Leibler divergence may not exist. Hence, the Hellinger metric is more general and widely applicable than the Kullback-Leibler divergence.

3. Hellinger distance and convergence of estimators can also be related. Suppose that, for an estimator $\hat{\theta}_n$ of θ, $\liminf_n \sup_{\theta \in \Theta} \Pr(|\hat{\theta}_n - \theta| > \epsilon) > 0$, where $\epsilon > 0$. Then, $\limsup_n \sup_{\theta \in \Theta} \Pr(H_n^2(\theta, \hat{\theta}_n) < 1) < 1$. Conversely, $\liminf_n \sup_{\theta \in \Theta} \Pr(H_n^2(\theta, \hat{\theta}_n) < 1) = 1$ implies that $\limsup_n \sup_{\theta \in \Theta} \Pr(|\hat{\theta}_n - \theta| > \epsilon) = 0$, for any $\epsilon > 0$, i.e. $\hat{\theta}_n$ is consistent.

3 The Hellinger Rate

3.1 Theory

Let $\mathcal{Y}$ denote a real-valued (Euclidean) sample space of random variables y and $\sigma(y)$ the (Borel) σ-field generated by y. Consider a parametric family of distributions $\{F(y;\theta), \theta \in \Theta\}$, where Θ is a compact set.

Definition: A sequence $\delta_n(\theta)$, $\delta_n(\theta) > 0$ and $\delta_n(\theta) \to 0$ as $n \to \infty$, is called a *Hellinger rate* at θ if $\theta + \delta_n(\theta) t_n$, for any strictly positive, bounded sequence t_n with $t_n \to t \in (0, +\infty)$, converges to θ such that the Hellinger affinity of an i.i.d. sample $\rho(\theta, \theta +$

[5] Nonnegativity, symmetry and reflexivity are obvious, identity of indiscernibles follows from the identification definition, and the triangle inequality is the same as in the case of the L^2 norm.

[6] The Kullback-Leibler divergence is not a distance because it is not symmetric.

$\delta_n(\theta)t_n)^n$ converges to a limit $\beta(\theta,t) \in [0,1]$, continuous in t and satisfying $\beta(\theta,0) = 1$ and $\lim_{t\to+\infty} \beta(\theta,t) = 0$.

To establish existence of Hellinger rates, the following assumptions will be maintained:

A1: $y_i \sim$ i.i.d. $F(y;\theta)$, $i = 1,\ldots,n$, $\theta \in \Theta \subset \mathbb{R}$, Θ compact;

A2: $(\mathcal{Y}, \sigma(y))$ is a σ-finite measurable space, $F(y;\theta)$ is absolutely continuous with respect to Lebesgue measure a.e., and $f(y;\theta)$ is the Radon-Nikodym derivative of $F(y;\theta)$;

A3: (*identification*) for any $\theta, \theta' \in \Theta$, $\rho(\theta,\theta') = 1 \Leftrightarrow \theta = \theta'$.

A4: (*tightness*) for any statistics T_n which is a measurable map from $(\mathcal{Y}, \sigma(y))$ to $\mathbb{R}$, the sequence of probability laws $\mathcal{L}(T_n|P_{\theta,n})$ is tight on $\mathbb{R}$, where $P_{\theta,n}(\mathbf{y}) = \prod_{i=1}^n F(y_i;\theta)$;[7]

A5: (*contiguity*) $\{P_{\theta,n}, \theta \in \Theta\}$ and $\{P_{\theta+\delta_n(\theta)t_n,n}, \theta \in \Theta\}$, for $|t_n|$ bounded, $\delta_n(\theta) > 0$ for all n and $\delta_n(\theta) \to 0$ as $n \to \infty$, are contiguous;[8]

A6: (δ_n-*tail continuity*) $\{P_{\theta,n}, \theta \in \Theta\}$ is δ_n-tail continuous at $\theta_0 \in \Theta$, i.e. the L_1-norm $||P_{s_n,n} - P_{t_n,n}|| \to 0$ for all sequences $\{s_n = \delta_n^{-1}(\tilde{s}_n - \theta)\}$ and $\{t_n = \delta_n^{-1}(\tilde{t}_n - \theta)\}$ s.t. (*i*) $\delta_n^{-1}(\tilde{s}_n - \theta_0) + \delta_n^{-1}(\tilde{t}_n - \theta_0) = O(1)$ and (*ii*) $\delta_n^{-1}(\tilde{s}_n - \tilde{t}_n) = o(1)$; this holds for every $\theta_0 \in \Theta$, i.e. $\{P_{\theta,n}, \theta \in \Theta\}$ is δ_n-tail continuous.

The following result establishes the existence of Hellinger rates.

Lemma 1: *Under* **A1-A6**, *for every* $\theta \in \Theta$, *there exists a function* $\beta(\theta,t) \in (0,1)$, *continuous in* $t \in [0,+\infty]$ *such that* $\lim_n \rho(\theta, \theta + \delta_n(\theta)t_n) = \beta(\theta,t)$, *where* t_n *is positive and bounded and* $t_n \to t \in (0,\infty)$, *i.e.* $\delta_n(\theta)$ *is a Hellinger rate at* θ.

Proof: Assumptions **A1** and **A2** define the relevant probability space. Consider $P_{\tilde{s}_n,n}(\mathbf{y}) = \prod_{i=1}^n F(y_i;\theta + \delta_n(\theta)\delta_n(\theta)^{-1}(\tilde{s}_n - \theta)) = P_{\theta,s_n,n}$, where $s_n = \delta_n(\theta)^{-1}(\tilde{s}_n - \theta) = O(1)$. Suppose $|s_n| \to s \in (0,+\infty)$. Tightness (**A4**), implies that subsequences of $P_{\theta,s_n,n}$ converge weakly $P_{\theta,s_n,n} \Rightarrow P_{\theta,s}$, i.e. for any bounded and continuous function ϕ, $\int \phi dP_{\theta,s_n,n} \to \int \phi dP_{\theta,s}$.[9] Then, for two positive, convergent sequences $s_n \to s$ and $t_n = \delta(\theta)^{-1}(\tilde{t}_n - \theta) \to t$, which induce convergent subsequences of P,

$$\begin{aligned} ||P_{\theta,s} - P_{\theta,t}|| &= ||(P_{\theta,s} - P_{\theta,s_n,n}) - (P_{\theta,t} - P_{\theta,t_n,n}) + (P_{\theta,s_n,n} - P_{\theta,t_n,n})|| \\ &\leq ||P_{\theta,s} - P_{\theta,s_n,n})|| + ||P_{\theta,t} - P_{\theta,t_n,n}|| + ||P_{\theta,s_n,n} - P_{\theta,t_n,n}||. \end{aligned}$$

The first two terms go to zero because $P_{\theta,s_n,n} \Rightarrow P_{\theta,s}$ and $P_{\theta,t_n,n} \Rightarrow P_{\theta,t}$, while the third term goes to zero by **A6** since $s_n - t_n = \delta(\theta)^{-1}(\tilde{s}_n - \theta) - \delta_n^{-1}(\tilde{t}_n - \theta) \to s - t$, provided

[7] I.e. for any $\epsilon > 0$, there exist $N(\epsilon)$ and $M(\epsilon) > 0$ such that $P_{\theta,n}(|T_n| > M(\epsilon)) < \epsilon$ for all $n > N(\epsilon)$. LeCam and Yang (2000) refer to this assumption as *relative compactness*.

[8] See LeCam and Yang (2000), chapt.3; contiguity can be thought of as mutual absolute continuity for all n.

[9] See, for exampe, Durrett (1996).

$|s-t| \to 0$. This implies that the limits $P_{\theta,s}$ are continuous in s. Also, for convergent subsequences,

$$\begin{aligned}
1 &= \lim_n \int_y f(y;\theta+\delta_n(\theta)t_n)dy \\
&= \lim_n \int_y \frac{f(y;\theta+\delta_n(\theta)t_n)}{f(y;\theta)} f(y;\theta)dy \quad \text{(well-defined, by } \mathbf{A5}\text{)} \\
&= \lim_n E_\theta\left[\exp\left(\ln\left(\frac{f(y;\theta+\delta_n(\theta)t_n)}{f(y;\theta)}\right)\right)\right] \\
&\geq \lim_n \rho(\theta,\theta+\delta_n(\theta)t_n) \\
&=: \beta(\theta,t) \geq 0.
\end{aligned}$$

Continuity of $P_{\theta,t}$ implies continuity of $\beta(\theta,t)$. Identification (**A3**) implies $\beta(\theta,0)=1$. To see that $\lim_{t\to+\infty}\beta(\theta,t)=0$, take $t_n=\delta_n(\theta)^{-1}$, so that $\rho(\theta,\theta+\delta_n(\theta)t_n)=\rho(\theta,\theta+1)\in(0,1)$, and hence the conclusion follows. □

Next, it will be shown that Hellinger rates form equivalence classes. Define two sequences $\delta_n(\theta)$ and $\bar{\delta}_n(\theta)$ to be rate equivalent if $0<\liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} \leq \limsup_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} < +\infty$, and the equivalence class $\Delta(\delta_n(\theta))$, defined by $\delta_n(\theta)$, as

$$\Delta(\delta_n(\theta)) = \left\{\bar{\delta}_n(\theta)>0, \bar{\delta}_n(\theta)\to 0 : 0<\liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} \leq \limsup_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} < +\infty\right\}.$$

To facilitate notation, write $\delta_n \sim \bar{\delta}_n$ if $\bar{\delta}_n \in \Delta(\delta_n)$. This is indeed an equivalence class as it obviously satisfies reflexivity and symmetry, and transitivity holds because $0 < \liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} \leq \limsup_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} < +\infty$ and $0<\liminf_n \frac{\tilde{\delta}_n(\theta)}{\bar{\delta}_n(\theta)} \leq \limsup_n \frac{\tilde{\delta}_n(\theta)}{\bar{\delta}_n(\theta)} < +\infty$ implies that

$$\begin{aligned}
\liminf_n \frac{\tilde{\delta}_n(\theta)}{\delta_n(\theta)} &= \liminf_n \frac{\tilde{\delta}_n(\theta)}{\bar{\delta}_n(\theta)} \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} \\
&= \liminf_n \frac{\tilde{\delta}_n(\theta)}{\bar{\delta}_n(\theta)} \liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} > 0,
\end{aligned}$$

and similarly for the lim sup.

The following result establishes that Hellinger rates, assuming they exist, form an equivalence class.

Lemma 2: *If $\delta_n(\theta)$ and $\bar{\delta}_n(\theta)$ are Hellinger rates at θ, then*

$$0<\liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} \leq \limsup_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} < +\infty.$$

Proof: Suppose, to the contrary, that $\liminf_n \frac{\bar{\delta}_n(\theta)}{\delta_n(\theta)} = 0$. Then, for $t_n \to t \in (0,+\infty)$,

$$\beta(\theta,t) = \lim_n \rho(\theta,\theta+\bar{\delta}_n(\theta)t_n)^n = \lim_n \rho(\theta,\theta+\delta_n(\theta)(\bar{\delta}_n(\theta)/\delta_n(\theta))t_n)^n,$$

implying that $\lim_n \rho(\theta, \theta + \delta_n(\theta)s_n)^n = \beta(\theta, t) > 0$ for a sequence $s_n = (\bar{\delta}_n(\theta)/\delta_n(\theta))t_n$ converging to zero. This contradicts the definition of $\delta_n(\theta)$ as a Hellinger rate. An analogous argument establishes that the supposition $\limsup_n \bar{\delta}_n(\theta)/\delta_n(\theta) = +\infty$ produces a contradiction. □

To determine the Hellinger rate, Hellinger distance and/or Hellinger affinity need to be calculated. Hence, in order to characterize general properties of Hellinger rates, it seems sensible to deduce them from conditions on Hellinger distance or affinity and to check in applications whether these conditions are met. The following result provides a necessary and sufficient condition on Hellinger affinity for the Hellinger rate to be uniform on Θ.

Lemma 3: *Suppose* **A1-A6** *hold. A necessary and sufficient condition (H) for the Hellinger rate to be uniform on* Θ *is that*

$$0 < \liminf_{\tau \to 0} \frac{\rho(\theta, \theta \pm \tau)}{\rho(\alpha\theta, \alpha(\theta \pm \tau))} \leq \limsup_{\tau \to 0} \frac{\rho(\theta, \theta \pm \tau)}{\rho(\alpha\theta, \alpha(\theta \pm \tau))} < +\infty,$$

for any $\theta \in \Theta$ *and* $\alpha \in \mathbb{R} : \alpha\theta, \alpha(\theta \pm \tau) \in \Theta$.

Proof: Uniformity of the Hellinger rate requires that

$$\rho(\theta, \theta \pm \delta_n) \sim \rho(\theta', \theta' \pm \delta_n) \text{ for any } \theta, \theta' \in \Theta.$$

Provided $\theta \neq 0$, this is equivalent to

$$\rho(\theta, \theta \pm \delta_n) \sim \rho\left(\theta\frac{\theta'}{\theta}, \theta\frac{\theta'}{\theta} \pm \delta_n\right) = \rho(\alpha\theta, \alpha\theta \pm \delta_n) \text{ for } \alpha = \tfrac{\theta'}{\theta}.$$

Since $\delta_n \sim \alpha\delta_n$, this is equivalent to

$$\rho(\theta, \theta \pm \delta_n) \sim \rho(\alpha\theta, \alpha(\theta \pm \delta_n)).$$

Since $\theta, \theta' \in \Theta$ were arbitrary, this is just condition (H) given in the claim. □

This result covers many cases of interest, in particular the case of location and scale parameters, as the following Corollary to Lemma 3 establishes, but also certain cases of shape parameters, as illustrated in the next section.

Corollary 1: *Suppose* **A1-A6** *and condition H hold, and (i)* θ *is a location parameter, or (ii)* θ, $0 < \theta < \infty$, *is a scale parameter. Then, the Hellinger rate does not depend on the value of* θ.

Proof: The case (i) is obvious, since $\rho(\theta, \theta + \tau) = \rho(\theta', \theta' + \tau) = \bar{\rho}(\tau)$ for all $\theta, \theta' \in \Theta$ and some function $\bar{\rho}(\cdot)$. Hence, $\bar{\rho}(\tau) \sim \bar{\rho}(\alpha\tau)$ for any $\alpha \in \mathbb{R}$, and since $\tau \sim \alpha\tau$, the result follows from Lemma 3.

In case (ii), let $f(\cdot)$ denote the density of the random variable Y with scale parameter 1. Then, $Z = \theta Y$, $0 < \theta < \infty$ has density $\frac{1}{\theta} f\left(\frac{z}{\theta}\right)$. Hence, for $0 < \theta, \theta' < \infty$,

$$\begin{aligned} h^2(\theta, \theta') &= 1 - \int_z \sqrt{\frac{\theta}{\theta'}} \sqrt{f(z) f\left(z \frac{\theta}{\theta'}\right)} dz \\ &= 1 - \int_z \sqrt{1 + \frac{\tau}{\theta'}} \sqrt{f(z) f\left(z\left(1 + \frac{\tau}{\theta'}\right)\right)} dz, \\ &= H\left(\frac{\tau}{\theta'}\right), \end{aligned}$$

where $\tau = |\theta - \theta'|$, and $h^2(\theta, \theta') = H\left(\frac{\tau}{\theta'}\right) \to 0$ as $\tau \to 0$. This implies that ρ is homogeneous of degree zero, so that the result follows from Lemma 3 as well. □

Remark: Notice that a Hölder continuity assumption on the Hellinger distance, as in Ibragimov and Has'minskii (1981), of the form

$$h^2(\theta, \theta + \tau) \leq E\left[K(y)\right] |\tau|^\beta, \ \beta > 0,$$

yields

$$\rho(\theta, \theta + \tau) \geq 1 - E\left[K(y)\right] |\tau|^\beta.$$

The Hölder continuity assumption is not nested by Lemma 2. The reason is that it implies that there exists a lower bound on the rate at which the Hellinger affinity $\rho(\theta, \theta + \tau)$ at any $\theta \in \Theta$ approaches 1 as $\tau \to 0$, but it does not pin down the actual rate, which may or may not depend on the value of θ.

The final result in this section shows that the Hellinger rate is invariant under transformations of the random variable that do not depend on the parameter of interest.

Lemma 4: *Suppose that* **A1-A6** *hold. Consider invertible transformations $Z = g(Y)$ of the random variable Y which do not depend on θ. Let $\Delta_Y(\delta_n(\theta))$ and $\Delta_Z(\delta_n(\theta))$ denote the Hellinger rate equivalence classes based on the random variables Y and Z, respectively. Then, $\Delta_Y(\delta_n(\theta)) = \Delta_Z(\delta_n(\theta))$ for all θ.*

Proof: Let $f_Y(y; \theta)$ denote the density of Y. Since $g(\cdot)$ is invertible, Z has density $f_Z(z; \theta) = f_Y\left(g^{-1}(z); \theta\right)\left[g'\left(g^{-1}(z)\right)\right]^{-1}$. Therefore, for any θ, if $\gamma_n(\theta) \in \Delta_Y(\delta_n(\theta))$, then

$$\begin{aligned} \frac{1}{n} &\sim \rho_Y(\theta, \theta + \gamma_n(\theta)) \\ &= \int_y \sqrt{f_Y(y; \theta) f_Y(y; \theta + \gamma_n(\theta))} dy \\ &= \int_z \sqrt{f_Y\left(g^{-1}(z); \theta\right) f_Y\left(g^{-1}(z); \theta + \gamma_n(\theta)\right)} \left[g'\left(g^{-1}(z)\right)\right]^{-1} dz \\ &= \int_z \sqrt{f_Z(z; \theta) f_Z(z; \theta + \gamma_n(\theta))} dz \\ &= \rho_Z(\theta, \theta + \gamma_n(\theta)). \end{aligned}$$

Hence, $\gamma_n(\theta) \in \Delta_Z(\delta_n(\theta))$, for any θ. A symmetric argument establishes the reverse inclusion. □

3.2 Examples

Under the conditions of Lemma 2, following the proof of Lemma 1 there exists a strictly decreasing function $\gamma(\cdot)$, such that

$$\lim_{\tau \to 0} \gamma(\tau) h^2(\theta, \theta + \tau) \text{ exists} \in (0, 1),$$

so that the Hellinger rate $\delta_n = \gamma^{-1}(n)$. The following examples illustrate this result.

Example 1: (Regular case) $\gamma(\tau) = \tau^{-2}$ and so $\delta_n = n^{-\frac{1}{2}}$.

Consider a regular Maximum Likelihood problem, with $\Theta \subset \mathbb{R}^k$. Here, the vector case poses no problem, because in the regular case, all components converge at the same rate, $\sqrt{n}$. Let $\tau \in \mathbb{R}^k$ be such that $||\tau|| \to 0$. Under regularity conditions, the log-likelihood ratio has an LAQ expansion about θ, uniformly in τ,

$$h^2(\theta, \theta + \tau) = S(\theta)\tau - \frac{1}{2}\tau' K(\theta)\tau + o_p(1),$$

where $S(\theta) = \nabla_\theta f(y; \theta)$ and $K(\theta) = \nabla_{\theta\theta} f(y; \theta)$, both having finite expectation. This implies that the Hellinger distance satisfies condition H of Lemma 2, and $\delta_n = \frac{1}{\sqrt{n}}$.

The log-likelihood ratio of an i.i.d. sample satisfies

$$\Lambda_n\left(\theta, \theta + \frac{1}{\sqrt{n}} t_n\right) = \sum_{i=1}^{n} \ln\left(\frac{f(y_i; \theta + \frac{1}{\sqrt{n}} t_n)}{f(y; \theta)}\right) = \frac{1}{\sqrt{n}} S_n(\theta) t_n - \frac{1}{2}\frac{1}{n} t_n' K_n(\theta) t_n + o_p(1),$$

where $S_n(\theta) = \sum_{i=1}^n \nabla_\theta \ln f(y_i; \theta)$ and $K_n(\theta) = \sum_{i=1}^n \nabla_{\theta\theta} \ln f(y; \theta)$. Therefore,

$$\begin{aligned} H_n^2(\theta, \theta + \frac{1}{\sqrt{n}} t_n) &= 1 - E_\theta\left[\exp\left(\frac{1}{2}\Lambda_n\left(\theta, \theta + \frac{1}{\sqrt{n}} t_n\right)\right)\right] \\ &= 1 - E_\theta\left[\exp\left(\frac{1}{2}\frac{1}{\sqrt{n}} S_n(\theta) t_n - \frac{1}{4}\frac{1}{n} t_n' K_n(\theta) t_n\right)\right] + o(1). \end{aligned}$$

The Hellinger rate ensures that the first term in the LAQ expansion converges by a Central Limit Theorem to a normal random variable, while the second term converges by a Strong Law of Large Numbers to a constant: $\frac{1}{\sqrt{n}} S_n(\theta) \xrightarrow{d} \mathcal{N}(0, \mathcal{I}(\theta))$, for $\mathcal{I}(\theta) = -E[\nabla_{\theta\theta} \ln f(y; \theta)] = E[\nabla_\theta \ln f(y; \theta) \nabla_\theta \ln f(y; \theta)']$ positive definite, and $\frac{1}{n} K_n(\theta) \to \mathcal{I}(\theta)$ a.s. Then, the Hellinger distance can be shown to converge to an interior limit, uniformly in t_n. Since the exponent in the above expression for $H_n^2(\theta, \theta + \frac{1}{\sqrt{n}} t_n)$, $\frac{1}{2}\frac{1}{\sqrt{n}} S_n(\theta) t_n -$

$\frac{1}{4}\frac{1}{n}t_n'K_n(\theta)t_n$, is nonzero with probability one, $H_n^2(\theta,\theta+\frac{1}{\sqrt{n}}t_n)>0$ for all n. Using Jensen's Inequality and the score identity $\int_y \nabla_\theta \ln f(y;\theta)f(y;\theta)dy\equiv 0$ for all θ,

$$\begin{aligned} H_n^2(\theta,\theta+\frac{1}{\sqrt{n}}t_n) &\leq 1-\exp\left(-\frac{1}{4}\frac{1}{n}t_n'E\left[K_n(\theta)\right]t_n\right) \\ &\to 1-\exp\left(-\frac{1}{4}t'\mathcal{I}(\theta)t\right)<1. \end{aligned}$$

Convergence to an interior limit follows from monotonicity and these interior bounds.[10]

Notice that frequently the expectation $E\left[\exp\left(\frac{1}{2}\frac{1}{\sqrt{n}}S_n(\theta)t_n-\frac{1}{4}\frac{1}{n}t_n'K_n(\theta)t_n\right)\right]$ cannot be computed analytically, because the expectation is taken with respect to the arbitrary regular density $f(y;\theta)$, while $\frac{1}{\sqrt{n}}S_n(\theta)$ asymptotically has a normal distribution. An analytically tractable case is $y_i \sim$ i.i.d. $\mathcal{N}(0,1)$, $i=1,\ldots,n$. In this case, standard calculations yield $H_n^2(0,\frac{1}{\sqrt{n}}t_n)=1-\exp(-t_n^2/8)$ for any n, and so $A(t)=t^2/8$.

Example 2: (Nonregular case) Suppose that $z\sim G(z)$, with density $g(z)$. Let $\psi(z,\lambda)=|z|^\lambda\text{sgn}(z)$, $\lambda\in(0,1)$, and let $y=\theta+\psi(z,1/\lambda)$. Then, $F(y;\theta,\lambda)=G(|y-\theta|^\lambda\text{sgn}(y-\theta))$ and $f(y;\theta,\lambda)=|y-\theta|^{\lambda-1}g(|y-\theta|^\lambda\text{sgn}(y-\theta))$. The density has a pole at $y=\theta$. Suppose λ is known. Then, it can be shown that $\gamma(\tau)=\tau^{-\lambda}$ and $\delta_n=n^{-\frac{1}{\lambda}}$.

Suppose λ is not known, and, w.l.o.g., $\theta=0$ and $g(z)$ the uniform density on $[-1/2,1/2]$. This example illustrates the case where λ is neither a shift nor a scale parameter, but a shape parameter. It can be shown that $h^2(\lambda,\lambda+\tau)=1-\frac{\sqrt{1+\frac{\tau}{\lambda}}}{1+\frac{\tau}{2\lambda}}\left(\frac{1}{2}\right)^{\frac{\tau}{2\lambda}}$. In this example, the Hellinger rate for λ, $\delta_{\lambda,n}$, is distinct from rate for θ, $\delta_{\theta,n}$. By virtue of Lemma 2, $\delta_{\lambda,n}$ again does not depend on the value of λ. In fact, $\delta_{\lambda,n}=\frac{1}{n}$ yields

$$\rho_\lambda\left(\lambda,\lambda+\frac{1}{n}\right)^n=\frac{\sqrt{\left(1-\frac{1}{\lambda n}\right)^n}}{1-\frac{1}{2\lambda n}}\left(\frac{1}{2}\right)^{\frac{1}{2\lambda}}\to\exp\left(-\frac{1}{2\lambda}\right)\left(\frac{1}{2}\right)^{\frac{1}{2\lambda}}\text{ as }n\to\infty.$$

Example 3: (Prakasa Rao (1968, 2003), and LeCam and Yang (2000); an example with similar features is given in Akahira (1975)) Let $\alpha>0$ be known and $f_\alpha(y)=c(\alpha)\exp(-|y|^\alpha)$, where $c(\alpha)=\left(2\Gamma\left(1/\alpha\right)\right)^{-1}$. Consider the shift parameter family $\{f(y;\theta),\theta\in\Theta\}=\{f_\alpha(y-\theta),\theta\in\mathbb{R},\alpha>0\}$. If $\alpha>\frac{1}{2}$, it can be shown that $\delta_n=n^{-\frac{1}{2}}$, while in the case of $\alpha=\frac{1}{2}$, $\gamma(\tau)=\left(\tau^2\ln\tau\right)^{-1}$, or $\delta_n=\frac{1}{\sqrt{n\ln n}}$; if $\alpha<\frac{1}{2}$, then $\delta_n=n^{-\frac{1}{1+2\alpha}}$.

4 Maximal Uniform Convergence Rates

This section derives maximal uniform convergence rates from the Hellinger rate.

[10]Rotnizky et al.(2000) show that $\text{rk}\mathcal{I}(\theta)=k-1$ implies that $\sqrt{n}$ convergence only applied to a $k-1$ dimensional subvector of θ, while the remaining component of θ converges at a slower rate.

To motivate this line of argument, two introductory examples will be useful. Both of these suggest that the random Hellinger distance $H_n^2(\theta, \hat{\theta}_n)$, when evaluated at an estimator $\hat{\theta}_n$ converging at the (inverse) Hellinger rate δ_n^{-1}, has an expectation that converges to an interior limit $\alpha(\theta) \in (0,1)$, i.e. $E[H_n^2(\theta, \hat{\theta}_n)] \to \alpha(\theta)$ as $n \to \infty$.

Example 2 (continued): Re-consider the analytically tractable special case of the regular maximum likelihood example, example 2 above, with $y_i \sim$ i.i.d. $\mathcal{N}(0,1)$, $i = 1, \ldots, n$. In this case, $H_n^2(0,\tau) = 1 - \exp(-n\tau^2/8)$. Consider the regular maximum likelihood estimator for the mean, $\bar{y}_n = \frac{1}{n}\sum_{i=1}^n y_i$. For any n, $\sqrt{n}\bar{y}_n \sim \mathcal{N}(0,1)$. Hence, replacing τ by the $\sqrt{n}$ convergent estimator $\bar{y}_n$, the random Hellinger distance satisfies

$$H_n^2(0,\bar{y}_n) = 1 - \exp(-n\bar{y}_n^2/8) \stackrel{d}{=} 1 - \exp(-x^2/8),$$

where $x \sim \mathcal{N}(0,1)$, and $\stackrel{d}{=}$ represents equality in distribution. Standard calculations yield that the expectation of the random Hellinger distance, when evaluated at this estimator converging at the (inverse) Hellinger rate,

$$E_0[H_n^2(0,\bar{y}_n)] = 1 - E_0[\exp(-x^2/8)] = 1 - 2/\sqrt{5}$$

lies in the interior of the unit interval. Note, for future reference, that the asymptotic distribution - or the exact small sample distribution in the special case of normality - is non-degenerate and does not depend on n. In fact, the realizations of the stochastically bounded limiting random variable $x \sim N(0,1)$ take the place of the limits t of the deterministic bounded sequences t_n.

Example 4: Consider the case of $y_i \sim$ i.i.d. $u[\theta - \frac{1}{2}, \theta + \frac{1}{2}]$; here, θ is again a location parameter. It is easy to show that $\bar{\rho}(\tau) = 1 - \tau$, and so $H_n^2(\theta, \theta+\tau) = 1 - (1-\tau)^n$ for all θ. Hence, $\gamma(\tau) = \frac{1}{\tau}$, and therefore $\gamma(\delta_n) = n$, or $\delta_n = \frac{1}{n}$.

Consider two estimators for θ in this example: (i) $\hat{\theta}_n = \frac{1}{2}(y_{(1)} + y_{(n)})$, where $y_{(1)}$ ($y_{(n)}$) denotes the minimum (maximum) of the sample $\{y_i, i = 1, \ldots, n\}$; and (ii) $\bar{\theta}_n = \bar{y}_n$. It can be shown[11] that

$$\begin{aligned} \mathrm{var}(\hat{\theta}_n) &= \frac{1}{2(n+1)(n+2)}, \\ \mathrm{var}(\bar{\theta}_n) &= \frac{1}{12n}, \end{aligned}$$

i.e. $\hat{\theta}_n$ converges at the Hellinger rate n, while $\bar{\theta}_n$ converges at the slower rate $\sqrt{n}$. W.l.o.g., let $\theta = 0$. Then, by a Taylor series expansion, $H_n^2(0,\tau) = n\tau + n(n-1)\tau^2 + o(\tau^2)$, and so, for $x \sim N(0,1)$,

$$E_0[H_n^2(0,\bar{\theta}_n)] = n(n-1)E_0[\bar{y}_n^2] + o(1) = (n-1)E_0[x^2] + o(1) = \frac{n-1}{12} + o(1),$$

[11] Cp., e.g., David (1970)

which diverges, locally to $\theta = 0$; since H_n^2 is uniformly bounded above by 1, this implies that the expectation converges to 1. Evaluation of the Hellinger distance at $\hat{\theta}_n$, on the other hand, yields an interior limit

$$E_0[H_n^2(0, \hat{\theta}_n)] = n(n-1)\frac{1}{2(n+1)(n+2)} \to \frac{1}{2}.$$

This suggests a proposition along the following lines: Suppose assumptions A1-A3, together with condition H, hold, and denote the Hellinger rate by δ_n; then, δ_n^{-1} is the maximal uniform convergence rate for any stochastically bounded estimator. This result, stated and proven formally below, derives from the following logic. Let $\hat{\theta}_n$ denote a δ_n^{-1} consistent estimator for $\theta_0 \in \Theta$, with $\delta_n^{-1}(\hat{\theta}_n - \theta_0) = O_p(1)$. Then, $H_n^2(\theta_0, \hat{\theta}_n) = H_n^2(\theta_0, \theta_0 + \delta_n \delta_n^{-1}(\hat{\theta}_n - \theta_0))$. Since $\delta_n^{-1}(\hat{\theta} - \theta_0) = O_p(1)$, for every $\epsilon > 0$, there exists $M(\epsilon) > 0$, decreasing in ϵ, such that, for all n, $\Pr(\delta_n^{-1}|\hat{\theta}_n - \theta_0| > M(\epsilon)) < \epsilon$. Therefore, with probability at least $1 - \epsilon$, for all n

$$H_n^2(\theta_0, \hat{\theta}_n) \leq H_n^2(\theta_0, \theta_0 + \delta_n M(\epsilon)),$$

and therefore, in the limit, with probability at least $1 - \epsilon$,

$$\lim_n H^2(\theta_0, \hat{\theta}_n) \leq 1 - \exp(-A(\theta_0, M(\epsilon))),$$

for $A(\theta_0, M(\epsilon))$ as in Lemma 1. Equivalently, for any $M > 0$, there exists $\epsilon(M) > 0$, decreasing in M, such that, for all n, with probability at least $1 - \epsilon(M)$,

$$\lim_n H_n^2(\theta_0, \hat{\theta}_n) \leq 1 - \exp(-A(\theta_0, M)).$$

Hence, the limiting distribution of $H_n^2(\theta_0, \hat{\theta}_n)$ is non-degenerate, with support $[0, 1]$. Note that the values M have the interpretation of limits of sample paths of the stochastically bounded sequence $\delta_n^{-1}(\hat{\theta}_n - \theta_0)$.

Next, suppose there exists a stochastically bounded estimator $\check{\theta}_n$ that is uniformly consistent and converges at a faster rate, say $\check{\delta}_n$, i.e. $\sup_{\theta \in \Theta} \check{\delta}_n^{-1}(\check{\theta}_n - \theta) = O_p(1)$ and $\limsup \check{\delta}_n / \delta_n = 0$. This implies that, for any bounded sequence t_n, $\limsup_n H_n^2(\theta_0, \theta_0 + \check{\delta}_n t_n) = 0$, or $\liminf_n \rho(\theta_0, \theta_0 + \check{\delta}_n t_n)^n = 1$. Suppose ρ is continuous; then, ρ is also uniformly continuous, because its arguments θ_0 and $\theta_0 + \check{\delta}_n t_n$ lie in a compact set. From the uniform continuity of ρ and the definition of δ_n, it follows that, for $\theta \in \Theta$ and $\epsilon > 0$,

$$\begin{aligned}
|\theta_0 - \theta| < \delta_n \quad &\Rightarrow \quad h^2(\theta_0, \theta) < (\gamma(\delta_n; \theta_0))^{-1} \\
\Pr(|\theta_0 - \check{\theta}_n| < \check{\delta}_n) \geq 1 - \epsilon \quad &\Rightarrow \quad \Pr\left(\limsup_n H_n^2(\theta_0, \check{\theta}_n) \leq \liminf_n 1 - \rho(\theta_0, \theta_0 + \check{\delta}_n)^n\right) \geq 1 - \epsilon \\
&\Leftrightarrow \Pr\left(\limsup_n H_n^2(\theta_0, \check{\theta}_n) \leq 0\right) \geq 1 - \epsilon \\
&\Leftrightarrow \Pr\left(\liminf_n \rho(\theta_0, \check{\theta}_n)^n \geq 1\right) \geq 1 - \epsilon.
\end{aligned}$$

It then follows from the bounded convergence theorem that $\liminf_n E[\rho(\theta_0, \check{\theta}_n)^n] = 1$. Hence, the limiting distribution of $H_n^2(\theta_0, \check{\theta}_n)$ is degenerate at 0, and the limiting distribution of $\rho(\theta_0, \check{\theta}_n)^n$ is degenerate at 1. This means that, unlike in the case of $\hat{\theta}_n$, these limiting random variables are independent of the sample paths of $\check{\delta}_n^{-1}(\check{\theta}_n - \theta_0)$. Therefore, the convergence result must hold also after linear transformations of the estimator, e.g. additive shifts. But this contradicts the hypothesized uniform $\check{\delta}_n^{-1}$ consistency of $\check{\theta}_n$.

This is formalized as

Proposition 1:[12] *Suppose* **A1-A6** *and condition H hold. If a δ_n^{-1}-consistent estimator exists, then there exists $t > 0$ such that, for every $\theta \in \Theta$,*

$$\liminf_{n\to\infty} H_n^2(\theta, \theta + t\delta_n^{-1}) > 0.$$

Proof: Assumptions A1-A3, together with condition H, imply that candidate rates exist and do not depend on θ, and that the fastest rate at which the result holds is the Hellinger rate. Let $P_\theta(B)$ denote the probability measure induced by $\prod_{i=1}^n F(y_i;\theta)$, applied to sets $B \in \prod_{i=1}^n \sigma(y_i)$, the product σ-field generates by $\{y_i, i = 1, \ldots, n\}$; the dependence on n is omitted for notational simplicity. Suppose $T_n = T(y_1, \ldots, y_n)$ is a δ_n^{-1}-consistent estimator. Then, $\exists\, \delta, L > 0$ such that, for every $\epsilon > 0$,

$$\limsup_n \sup_{|\theta-\theta_0|<\delta} P_\theta(\delta_n^{-1}|T_n - \theta| \geq L) < \epsilon.$$

Let $t > 2L$. Then, $\exists n_0$ such that for $n > n_0$, $\delta_n^{-1} > \delta_{n_0}^{-1} > t/\delta$ and

$$\sup_{|\theta-\theta_0|<t\delta_n} P_\theta(\delta_n^{-1}|T_n - \theta| \geq L) < \epsilon,$$

and

$$\limsup_n P_{\theta+t\delta_n}(\delta_n^{-1}|T_n - \theta - t\delta_n| \geq L) < \epsilon$$
$$\limsup_n P_\theta(\delta_n^{-1}|T_n - \theta| \geq L) < \epsilon.$$

Note that, since $t > 2L$, $\delta_n^{-1}|T_n - \theta| < L$ implies $\delta_n^{-1}|T_n - \theta - t\delta_n| \geq t - L$, and hence,

$$P_{\theta+t\delta_n}(\delta_n^{-1}|T_n - \theta| < L) \leq P_{\theta+t\delta_n}(\delta_n^{-1}|T_n - \theta - t\delta_n| \geq t - L) < \epsilon.$$

Therefore, $\liminf_n P_{\theta+t\delta_n}(\delta_n^{-1}|T_n - \theta| \geq L) \geq 1 - \epsilon$. Since by the Cauchy-Schwartz inequality,

$$H_n^2(\theta, \theta') \geq \frac{1}{2}(P_\theta(B) + P_{\theta'}(B)) - (P_\theta(B)P_{\theta'}(B))^{\frac{1}{2}} \text{ for any } B \in \prod_{i=1}^n \sigma(y_i),$$

[12]The form of the proposition and its proof closely parallel Akahira (1975), but it covers a considerably wider class of parametric estimation problems.

it follows that

$$\begin{aligned}\liminf_n H_n^2(\theta,\theta+t\delta_n) &\geq \liminf_n \left\{\frac{1}{2}P_\theta(\delta_n^{-1}|T_n-\theta|\geq L)+\frac{1}{2}P_{\theta+t\delta_n}(\delta_n^{-1}|T_n-\theta|\geq L)\right.\\ &\quad \left.-\left(P_\theta(\delta_n^{-1}|T_n-\theta|\geq L)P_{\theta+t\delta_n}(\delta_n^{-1}|T_n-\theta|\geq L)\right)^{\frac{1}{2}}\right\}\\ &\geq \frac{1}{2}(1-\epsilon)-\sqrt{\epsilon}.\end{aligned}$$

Since $\epsilon > 0$ was arbitrary, the result follows. □

Comment: The result of the proposition is perhaps best interpreted by its contrapositive: At a rate $\check{\delta}_n$ faster than the Hellinger rate δ_n, $\liminf_n H_n^2(\theta,\theta+t\check{\delta}_n) = 0$, and hence Proposition 1 leads to the conclusion that no uniformly $\check{\delta}_n^{-1}$-consistent estimator can exist. Hence, the Hellinger rate is the fastest possible uniform convergence.

Remark 1: Akahira (1991) and Akahira and Takeuchi (1995) provide a related result for the special case of location parameter families. For $\mathbf{y} = (y_1,\ldots,y_n)'$, they use the absolute variation metric (L^1 norm)[13]

$$d_n(\theta,\theta') = \int_y |f(\mathbf{y};\theta) - f(\mathbf{y};\theta')|dy,$$

and show that, if a δ_n^{-1} consistent estimator exists, then, for each $\theta\in\Theta$ and every $\epsilon > 0$, there exists a positive number t_0 such that, for any $t \geq t_0$,

$$\liminf_{n\to\infty} d_n(\theta,\theta-t\delta_n) \geq 2-\epsilon.$$

Akahira and Takeuchi (1995) show (Lemma 3.5.1) that, for any $\theta,\theta'\in\Theta$,

$$2H_n^2(\theta,\theta') \leq d_n(\theta,\theta') \leq 2\sqrt{2H_n^2(\theta,\theta')},$$

which implies that

$$\frac{1}{8}d_n^2(\theta,\theta') \leq H_n^2(\theta,\theta') \leq \frac{1}{2}d_n(\theta,\theta').$$

Hence, convergence in the Hellinger metric is equivalent to convergence in the absolute variation metric.

Remark 2: View $m_n = \prod_{i=1}^n f(y_i;\theta_0)$ and $\hat{m}_n(\omega_n) = \prod_{i=1}^n f(y_i;\hat{\theta}_n(\omega_n))$, $n = 1,\ldots,$ as elements of an infinite direct product measure space $\{(\Omega_n,\mathcal{B}_n,m_n), n = 1,\ldots\}$, for $\mathcal{B}_n$ a Borel σ-field of subsets of the set Ω_n, $\omega_n\in\Omega_n$, $\omega = \{\omega_n, n = 1,\ldots\}\in\Omega = \prod_{1=i}^\infty \Omega_n$, where ω denotes a given state of the world, with coordinates ω_n. Let $\rho(m_n,\hat{m}_n(\omega_n))$ denote the Hellinger affinity of these product measures, given ω_n. Each state of the world $\omega\in\Omega$ induces a sample path $\{\hat{\theta}_n(\omega_n), n = 1,\ldots\}$ and a sequence of measures $\{\hat{m}_n(\omega_n), n =$

[13]See also Hoeffding and Wolfowitz (1958) for a discussion of the properties of this metric.

$1, \ldots\}$. Applying a result by Kakutani (1948) shows that, if m_n and $\hat{m}_n(\omega_n)$ are mutually absolutely continuous (or: equivalent) measures , then the infinite product measures $m = \lim_n m_n$ and $\hat{m}(\omega) = \lim_n \hat{m}_n(\omega_n)$ are either mutually absolutely continuous or orthogonal, depending on whether $\lim_n \prod_{i=1}^n \rho(m_n, \hat{m}_n)$ is > 0 or $= 0$. In light of Kakutani's result, the above proposition can be interpreted as follows. Consider estimators $\hat{\theta}_n$ such that $\Pr(\Omega(M)) = 0$, where $\Omega(M) = \{\omega \in \Omega : \sup_{\theta\in\Theta} \delta_n^{-1}|\hat{\theta}_n(\omega) - \theta| > M \text{ for all } n,\ M > 0\}$, i.e. stochastically bounded, uniformly δ_n^{-1} consistent estimators. Then, estimators whose convergence rate δ_n^{-1} is the Hellinger rate generate sample paths such that the induced infinite product measures are mutually absolutely continuous (i.e. equivalent).

Proposition 1 covers in particular the class of locally asymptotically quadratic (LAQ) problems (see, e.g., LeCam and Yang (2000), sec. 5.2, for a definition). Let $S_n(\theta_0)$ and $K_n(\theta_0)$ denote the first and second term in the LAQ expansion of the log-likelihood ratio $\Lambda_n(\theta_0, \theta)$ about θ_0, $\theta, \theta_0 \in \Theta$. Define the infeasible LAQ estimator $\hat{\theta}_n$ of θ_0 by $\hat{\theta}_n = \theta_0 + [K_n(\theta_0)]^{-1} S_n(\theta_0)$. Then, the proposition above has the following corollary:

Corollary 2: *Let $\{f(x;\theta), \theta \in \Theta\}$, Θ compact, be an LAQ family of densities, with $\delta_n > 0$ and $\delta_n \to 0$ as $n \to \infty$ such that $\delta_n S_n(\theta) = O_p(1)$ and $\delta_n^2 K_n(\theta)$ positive definite almost surely.*[14] *Then, the infeasible LAQ estimator $\hat{\theta}_n = \theta_0 + [K_n(\theta_0)]^{-1} S_n(\theta_0)$ of θ_0 satisfies (i) $|\hat{\theta}_n - \theta_0| = o_p(1)$, (ii) $\delta_n^{-1}(\hat{\theta}_n - \theta_0) = O_p(1)$, and (iii) $\hat{\theta}_n$ converges at the maximal rate (i.e. δ_n is the Hellinger rate).*

Proof: Since the log-likelihood permits an LAQ expansion at θ_0, uniformly in t,

$$\Lambda_n(\theta_0, \theta_0 + \delta_n t_n) = \delta_n S_n(\theta_0) t_n - \frac{1}{2}\delta_n^2 t_n' K_n(\theta_0) t_n + o_p(1),$$

which is stochastically bounded as a consequence of the contiguity assumption (as part of the LAQ property), the quadratic has a unique maximizer

$$\hat{t}_n = \arg\max_s \Lambda_n(\theta_0, \theta_0 + \delta_n s) = [\delta_n^2 K_n(\theta_0)]^{-1} \delta_n S_n(\theta_0) = O_p(1),$$

as a direct consequence of the LAQ property: The inverse is finite almost surely and the second term is stochastically bounded. Hence, $\delta_n^{-1}(\hat{\theta}_n - \theta_0) = \hat{t}_n$ implies that (ii) follows immediately. Then, (i) follows from $\delta_n^2 S_n(\theta_0) = o_p(1)$. To show (iii), it suffices to show that the squared Hellinger distance of an i.i.d. sample at θ and $\theta + \delta_n t_n$, $H_n^2(\theta, \theta + \delta_n t_n)$, $\theta \in \Theta$ and t_n any bounded sequence, converges to an interior limit; the result then follows from the preceding proposition. To demonstrate the interior limit, notice first that

$$\begin{aligned} H_n^2(\theta, \theta + \delta_n t_n) &= 1 - E_\theta\left[\exp\left(\frac{1}{2}\Lambda_n(\theta, \theta + \delta_n t_n)\right)\right] \\ &= 1 - E_\theta\left[\exp\left(\frac{1}{2}\delta_n S_n(\theta_0) t_n - \frac{1}{4}\delta_n^2 t_n' K_n(\theta_0) t_n + o_p(1)\right)\right]. \end{aligned}$$

[14] It seems implicit in the definition of LAQ that δ_n does not depend on θ.

By contiguity, the exponent is stochastically bounded, hence $H_n^2(\theta_0, \theta_0 + \delta_n t_n)$ is bounded away from 1. To see that $H_n^2(\theta_0, \theta_0 + \delta_n t_n)$ is bounded away from 0, notice that for any sequence $\{t_n, n \geq 1\}$ for which $\Lambda_n(\theta_0, \theta_0 + \delta_n t_n) = 0$ a.s., the sequence $\{-t_n, n \geq 1\}$ yields $\Pr(\Lambda_n(\theta_0, \theta_0 - \delta_n t_n) \neq 0) > 0$. Therefore,

$$\lim_n E_\theta \left[\exp\left(\frac{1}{2}\delta_n S_n(\theta_0) t_n - \frac{1}{4}\delta_n^2 t_n' K_n(\theta_0) t_n + o_p(1) \right) \right] \in (0, 1),$$

and the result follows.

In the LAQ case, the result can also be obtained by a direct argument. Let $\check{\delta}_n > 0$ be such that $\check{\delta}_n \to \infty$ and $\limsup_n \check{\delta}_n / \delta_n = 0$. Then,

$$\check{\delta}_n^{-1}(\hat{\theta}_n - \theta_0) = \check{\delta}_n^{-1}\delta_n \left[\delta_n^2 K_n(\theta_0)\right]^{-1} \delta_n S_n(\theta_0) = O_p(\check{\delta}_n^{-1}\delta_n),$$

and hence diverges. □

5 Conclusions

This paper considers rate efficiency in parametric estimation as a criterion to judge the quality of estimators, next to other efficiency criteria, such as e.g. the Cramér Rao bound, within a give class estimators converging at a specific rate, e.g. $\sqrt{n}$. It addresses the question what maximal convergence rates parametric estimators can achieve in parametric estimation problems with i.i.d. data. The Hellinger metric is proposed as a very convenient tool to identify the Hellinger rate as a benchmark or gold standard for rate-efficiency.

This work deals only with scalar parameters of interest, or with parameter vectors whose components converge at the same rate. Future work might deal with cases like Example 3, in which different components of a parameter vector converge at different rates, and the rates of convergence of one depend on the other; and with the case of dependent data, where convergence rates may depend on the value of the parameter of interest.[15]

References

[1] Akahira, M. (1975): "Asymptotic Theory for Estimation of Location in Non-regular Cases, I: Order of Convergence of Consistent Estimators", *Stat. Appl. Res., JUSE*, **22(1)**, 8-26

[15] An example is, for instance, the case of the parameter of an autoregressive process. In the unit root case, estimators converge at rate T, while otherwise they converge at rate $\sqrt{T}$, where T is the sample size.

[2] Akahira, M. (1991): "The amount of information and the bound for the order of consistency for a location parameter family of densitites", *Symposia Gaussiana*, Conf. B, Mammitzsch and Schneeweiss, eds.; Berlin: Gryuter & Co.

[3] Akahira, M. and K. Takeuchi (1991): "A definition of information amount applicable to non-regular cases", *J. Comput. Inform.*, **2**, 71-92

[4] Akahira, M. and K. Takeuchi (1995): *Non-Regular Statistical Estimation*, New York: Spinger Verlag

[5] Birgé, L. and P. Massart (1993): "Rates of Convergence for Minimum Contrast Estimators", *Probability Theory and Related Fields*, **97**, 113-150

[6] Chan, K.S. (1993): "Consistency and Limiting Distribution of the Least Squares Estimator of a Threshold Autoregression Model", *The Annals of Statistics*, **21**, 520-533

[7] Chan, K.S. and R.S. Tsay (1998): "Limiting properties of the least squares estimator of a continuous threshold autoregressive model", *Biometrika*, **85(2)**, 413-426

[8] David, H.A. (1997): *Order Statistics*, 2nd edition, New York: Wiley

[9] Durrett, R. (1996): *Probability: theory and examples*, 2nd ed., Belmont, CA: Duxbury Press

[10] Hajek, J. (1970): "A characterization of the limiting distributions of regular estimates", *Z. Wahrsch. Verw. Gebiete*, **14**, 323-330

[11] Hansen, B. (2000): "Sample splitting and threshold estimation", *Econometrica*, **68**, 575-603

[12] Hirano, K. and J.R. Porter (2003): "Efficiency in Asymptotic Shift Experiments", mimeo, University of Miami and Harvard University

[13] Hoeffding, W. and J. Wolfowitz (1958): "Distinguishability of sets of distributions", *Ann. Math. Statist.*, **3**, 700-718

[14] Ibragimov, I.A. and R.Z. Has'minskii (1981): *Statistical Estimation*, New York: Springer

[15] Kakutani, S. (1948): "On Equivalence of Infinite Product Measures", *The Annals of Mathematics*, **49**, 214-224

[16] LeCam, L.M. (1986): *Asymptotic Methods in Statistical Decision Theory*, New York: Springer

[17] LeCam, L.M. and G.L. Yang (2000): *Asymptotics in Statistics - Some Basic Concepts*, New York: Springer

[18] Matusita, K. (1955): "Decision rules based on the distance for problems of fit, two samples and estimation", *Annals of Mathematical Statistics*, **26**, 631-640

[19] Pollard, D. (1984): *Convergence of Stochastic Processes*, New York: Springer

[20] Pollard, D. (1989): "Asymptotics via Empirical Processes", *Statistical Science*, **4(4)**, 341-354

[21] Prakasa Rao, B.L.S. (1968): "Estimation of the location of the cusp of a continuous density", *The Annals of Mathematical Statistics*, **39**, 76-87

[22] Prakasa Rao, B.L.S. (2003): "Estimation of Cusp in Nonregular Nonlinear Regression Models", mimeo, Indian Statistical Institute

[23] Rotnizky, A., Cox, D.R., Bottai, M. and J. Robins (2000): "Likelihood-based inference with singular information matrix", *Bernoulli*, **6(2)**, 243-284

[24] Seo, M. and O. Linton (2005): "A Smoothed Least Squares Estimator For The Threshold Regression Model", mimeo, London School of Economics

[25] Stone, C.J. (1980): "Optimal Rates of Convergence for Nonparametric Estimators", *The Annals of Statistics*, **8(6)**, 1348-1360

[26] Stone, C.J. (1982): "Optimal Global Rates of Convergence for Nonparametric Regression", *The Annals of Statistics*, **10(4)**, 1040-1053

[27] Van de Geer, S. (1993): "Hellinger-Consistency of Certain Nonparametric Maximum Likelihood Estimators", *The Annals of Statistics*, **21(1)**, 14-44

[28] Van de Geer, S. (2000): *Empirical Processes in M-Estimation*, Cambridge: Cambridge University Press