*Journal of the Korean Statistical Society (2003),* **32**: *3, pp 261–270*

# ON HELLINGER CONSISTENT DENSITY ESTIMATION[†]

THEODOROS NICOLERIS[1] AND STEPHEN G. WALKER[2]

## ABSTRACT

This paper introduces a new density estimator which is Hellinger consistent under a simple condition. A number of issues are discussed, such as extension to Kullback–Leibler consistency, robustness, the Bayes version of the estimator and the maximum likelihood case. An illustration is presented.

*AMS 2000 subject classifications.* Primary 62F12; Secondary 62G07.

*Keywords.* Density estimation, Hellinger consistency, martingale sequence, maximum likelihood estimator.

## 1. INTRODUCTION

Let $f_\theta(x) \equiv f(x;\theta)$, with $\theta \in \Theta$, be a parametric family (possibly infinite dimensional) of density functions. Independent and identically distributed samples from $f_0(x) \equiv f(x;\theta_0)$ are available; we will label these $X_1, X_2, \ldots$ The aim in this paper is to find a Hellinger consistent sequence of density estimates of $f_0$ under simple regularity conditions.

Van de Geer (1993) established sufficient conditions under which the sequence of maximum likelihood estimators yields $H(f_{\widehat{\theta}_n}, f_0) \to 0$ almost surely. Here

$$H(f, f_0) = \left\{\int \left(\sqrt{f} - \sqrt{f_0}\right)^2\right\}^{\frac{1}{2}}$$

is the Hellinger distance between $f$ and $f_0$. The condition of van de Geer (1993) involves a uniform law of large numbers result; define

$$\mathrm{G} = \left\{\sqrt{\frac{f_\theta}{f_0}} - 1 : \theta \in \Theta\right\}$$

Received March 2003; accepted April 2003.

[†]This research was financed by an Advanced Research Fellowship from the U.K. Engineering and Physical Science Research Council.

[1]Department of Statistics and Actuarial Science, University of the Aegean, 81100 Mytilene, Samos, Greece

[2]Department of Mathematical Sciences, University of Bath, Bath, BA2 7AY, U.K.

and $||g||_F = \int |g|\, dF$. If $(T, d)$ is a metric space, then the $\delta$-covering number of $T$ is labelled $N(\delta, T, d)$ and define $\mathrm{H}(\delta, T, d) = \log N(\delta, T, d)$. Van de Geer (1993) showed that if

$$\frac{1}{n}\mathrm{H}(\delta, \mathrm{G}, ||\cdot||_{F_n}) \to 0$$

in probability, then

$$\sup_{g\in \mathrm{G}} \left| \int g\, d(F_0 - F_n) \right| \to 0$$

with probability one which in turn implies that $H(f_{\widehat{\theta}_n}, f_0) \to 0$ with probability one. Here, $F_n$ is the empirical distribution function and $F_0$ the distribution function corresponding to $f_0$.

We find a necessary, but not sufficient, condition for $H(f_{\widehat{\theta}_n}, f_0) \to 0$ almost surely $(a.s.)$. We find an alternative density estimator which is Hellinger consistent when this necessary condition holds.

The density estimator, we will label it $f^n(x)$, is given by

$$f^n(x) = \frac{1}{n}\sum_{i=1}^{n} f_{\widehat{\theta}_{i-1}}(x)$$

and $\widehat{\theta}_0$ is any member of $\Theta$. It is therefore instructive to learn that there are situations when $H(f^n, f_0) \to 0$ yet $H(f_n, f_0) \not\to 0$, where $f_n = f_{\widehat{\theta}_n}$.

The practicability of our estimator is that it is possible to use the designated family of interest, that is $f(x;\theta)$, to provide a Hellinger consistent density estimator under simpler conditions than those for which the maximum likelihood density estimator is Hellinger consistent.

In Section 2, we present the result of the paper. Section 3 discusses and highlights a number of consequences of the result and an illustration is presented in Section 4.

## 2. THE RESULT

Our result is based on the following two lemmas; note that here we are not assuming the estimators are, at this point, maximum likelihood.

LEMMA 1. *Let $T_0, T_1, T_2, \ldots$ be a sequence of estimates of $\theta_0$ such that $T_n =$*

*$T_n(X_1,\dots,X_n)$ with $T_0$ a constant. Then*

$$\liminf_n \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \geq 1 \ a.s. \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \to 1 \ a.s.$$

$$\Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \to 0 \ a.s.$$

PROOF. Define

$$J_i = \sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}}$$

and consider the martingale sequence given by

$$S_n = \sum_{i=1}^{n}\{J_i - \ E(J_i|\mathrm{F}_{i-1})\},$$

where $\mathrm{F}_i = \sigma(X_1,\dots,X_i)$. Clearly

$$S_n = \sum_{i=1}^{n}\{J_i - 1 + h(f_{T_{i-1}}, f_0)\},$$

where $h(f, f_0) = H^2(f, f_0)/2$. It is well known, see for example Loève (1963, p. 387), that if

$$\sum_n \frac{1}{n^2}\operatorname{Var}(J_n) < \infty,$$

then $S_n/n \to 0$ *a.s.* It is easy to see that $\operatorname{Var}(J_n) \leq E(J_n^2) = 1$ and so

$$\frac{1}{n}\sum_{i=1}^{n} J_i - 1 + \frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \to 0 \ a.s.$$

Consequently,

$$\frac{1}{n}\sum_{i=1}^{n} J_i \geq 1 \ a.s. \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} J_i \to 1 \ a.s.,$$

since $h(\cdot,\cdot)$ is non-negative, and

$$\frac{1}{n}\sum_{i=1}^{n} J_i \to 1 \ a.s. \Leftrightarrow \frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \to 0 \ a.s.,$$

completing the proof. □

In the next lemma, we provide a sufficient, but not necessary, condition for

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \to 1 \ a.s.$$

LEMMA 2. *The following results hold:*

$$\liminf_n \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)} \geq 0 \ a.s. \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)} \to 0 \ a.s.$$

*and*

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)} \to 0 \ a.s. \Rightarrow \frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \to 1 \ a.s.$$

PROOF. The first result follows from

$$\mathrm{P}\left\{\prod_{i=1}^{n}\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)} > \exp(nc)\right\} < \exp(-nc)$$

for all $c > 0$. This is an application of the Markov inequality. The Borel-Cantelli lemma then establishes that

$$\limsup_n \left\{\prod_{i=1}^{n}\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}\right\}^{\frac{1}{n}} \leq 1 \ a.s.$$

and hence the result.

The second result follows since

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;T_{i-1})}{f_0(X_i)} \to 0 \ a.s.$$

implies that

$$\left\{\prod_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f_0(X_i)}}\right\}^{\frac{1}{n}} \to 1 \ a.s.$$

and an arithmetic mean is greater than or equal to a geometric mean. Hence, using Lemma 1,

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f_0(X_i)}} \to 1 \ a.s.,$$

completing the proof. □

We now have the main result of the paper.

THEOREM 1. *Let $T_0, T_1, T_2, \ldots$ be a sequence of estimates of $\theta_0$ such that $T_n = T_n(X_1, \ldots, X_n)$ with $T_0$ a constant. If*

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \to 1 \ a.s.,$$

*then*

$$H(f^n, f_0) \to 0 \ a.s.,$$

*where*

$$f^n(x) = \frac{1}{n}\sum_{i=1}^{n} f(x;T_{i-1}).$$

PROOF. Lemma 1 establishes that

$$\frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \to 0 \ a.s.$$

and, because of the convexity of $h(\cdot, f_0)$, $h(f^n, f_0) \to 0$ *a.s.*, completing the proof. □

To our knowledge, there is no previous literature on density estimators of the type

$$f^n = \frac{1}{n}\sum_{i=1}^{n} f_{T_{i-1}} \, .$$

Of course, it does resemble a kernel density estimator except that for our estimator $T_n = T_n(X_1, \ldots, X_n)$, whereas for a kernel density estimator $T_n = T_n(X_n)$. Note that we do not have the problem of establishing an arbitrary bandwidth which is always an issue with kernel density estimation. Of course, this difficulty needs to be balanced with the difficulty of the choice for $T_n$, though, where available, the maximum likelihood estimator is an obvious choice.

We should also point out that we could equally consider a sequence of density estimators $\widehat{f}_n$, rather than $T_n$, and the sequence is Hellinger consistent for $f_0$ when

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{\widehat{f}_{i-1}(X_i)}{f_0(X_i)}} \to 1 \ a.s.$$

This covers the full non-parametric case when $\theta$ is infinite dimensional and it is more convenient to consider $\widehat{f}$ converging to $f_0$ in the Hellinger sense than considering the convergence of $T_n$ to $\theta_0$ in some alternative infinite dimensional space.

Note that our condition,

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}} \to 1 \ a.s.$$

is both necessary and sufficient for

$$\frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \to 0 \ a.s.$$

Thus, it is quite clear that the condition is only a necessary one for $h(f_n, f_0) \to 0$, where $f_n = f_{T_n}$. Hence, there exist examples, though we are unable to present such examples, in which $h(f^n, f_0) \to 0$ and yet $h(f_n, f_0) \not\to 0$.

## 3. Consequences of Main Results

Here we discuss a number of points relating to the main result, and we will drop "*a.s.*" from the following.

### *3.1. Kullback-Leibler consistency*

Suppose we now define

$$K_i = \log\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)}$$

so that $E(K_i|\mathrm{F}_{i-1}) = -D(f_{T_{i-1}}, f_0)$. Here $D(f, f_0) = \int f_0 \log(f_0/f)$ is the Kullback–Leibler divergence from $f_0$ to $f$. Then define

$$S_n = \sum_{i=1}^{n}\{K_i + D(f_{T_{i-1}}, f_0)\}$$

which is a martingale sequence. If $\sup_n \mathrm{Var}(K_n) < \infty$, ensuring that $S_n/n \to 0$, and

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{f(X_i;T_{i-1})}{f(X_i;\theta_0)} \to 0,$$

equivalent to $n^{-1}\sum_{i=1}^{n} K_i \to 0$, then

$$\frac{1}{n}\sum_{i=1}^{n} D(f_{T_{i-1}}, f_0) \to 0$$

which implies that $D(f^n, f_0) \to 0$.

Consequently,

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i; T_{i-1})}{f(X_i; \theta_0)} \to 0$$

implies Hellinger consistency (see Lemma 2) whereas both

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i; T_{i-1})}{f(X_i; \theta_0)} \to 0$$

and $\sup_n \operatorname{Var}(K_n) < \infty$ imply Kullback–Leibler consistency.

### *3.2. Robustness*

The use of the martingale sequence introduced in Section 2 allows us to consider what happens to the Hellinger consistent density estimator when the model is wrong. A manifestation of the wrong model is typically of the type

$$\frac{1}{n}\sum_{i=1}^{n} \log \frac{f(X_i; T_{i-1})}{f(X_i; \theta_0)} \to -\delta$$

for some $\delta > 0$. It may not be this of course but we will assume this to be true. This is realistic; if $T_n \to T$, then standard law of large numbers arguments suggest that

$$\delta = \int f_0(x) \log \frac{f_0(x)}{f(x; T)}\, dx$$

and, of course, $\delta > 0$.

When this limit holds it follows that

$$\liminf_n \left\{ \prod_{i=1}^{n} \frac{f(X_i; T_{i-1})}{f(X_i; \theta_0)} \right\}^{\frac{1}{2n}} \geq \exp(-\delta/2)$$

and hence

$$\liminf_n \frac{1}{n}\sum_{i=1}^{n} J_i \geq \exp(-\delta/2).$$

It remains true that $S_n/n \to 0$ and therefore

$$\limsup_n \frac{1}{n}\sum_{i=1}^{n} h(f_{T_{i-1}}, f_0) \leq 1 - \exp(-\delta/2).$$

This implies that $\limsup_n h(f^n, f_0) \leq 1 - \exp(-\delta/2)$. So, provided the model is not too wrong, in the sense that $\delta$ is small, the estimator introduced in this paper has good robustness properties.

### *3.3. Bayes estimator*

Here we consider $f_n$ to be the predictive density for a Bayesian model. That is, we assign a prior distribution $\Pi(\theta)$ on the relevant parameter space and then construct the predictive density

$$f_n(x) = \int f(x;\theta)\, \Pi_n(d\theta).$$

Here $\Pi_n$ is the posterior distribution for $\theta$. It is shown in Walker (2003) that if the prior $\Pi$ puts positive mass on $\{\theta : D(f_\theta, f_0) < \varepsilon\}$ for all $\varepsilon > 0$, then

$$\frac{1}{n}\sum_{i=1}^{n}\sqrt{\frac{f_{i-1}(X_i)}{f_0(X_i)}} \to 1.$$

Finding priors which assign positive mass to Kullback–Leibler neighbourhoods of $f_0$ is not too difficult.

### *3.4. A symmetric estimator*

A possible drawback is that the estimator depends on an ordering of the data. Hence for any data set of size $n$, there are a possible $n!$ different estimators. For large data sets, the difference between estimators, in a Hellinger sense, will be small, as all are Hellinger consistent. It will not be an issue which estimator is chosen in this case. For small data sets, a solution is provided by averaging over all possibilities. So let $\Omega$ be the set of permutations on $\{1,\ldots,n\}$ with $|\Omega| = n!$. For each $\omega \in \Omega$, we have the estimator $f_\omega^n$ and so a symmetric estimator is provided by

$$f_\Omega^n = \frac{1}{|\Omega|}\sum_{\omega\in\Omega} f_\omega^n\,.$$

Of course, this is also a Hellinger consistent estimator of $f_0$, which can be seen by a re-working of Lemma 1 in an obvious way.

### *3.5. Maximum likelihood density estimator*

Here we let $f(\cdot;T_n) \equiv \widehat{f}_n(\cdot)$, the maximum likelihood density estimator based on a parametric family $f(\cdot;\theta)$. Recall, following Lemma 2, we are content with

$$\liminf_n \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widehat{f}_{i-1}(X_i)}{f_0(X_i)} \geq 0.$$

Now

$$\frac{1}{n}\sum_{i=1}^{n}\log\frac{\widehat{f}_{i-1}(X_i)}{f_0(X_i)} \geq \frac{1}{n}\sum_{i=1}^{n}\log\frac{\widehat{f}_{i-1}(X_i)}{\widehat{f}_n(X_i)}$$

$$= \frac{1}{n}\sum_{i=1}^{n}\sum_{k=i}^{n}\log\frac{\widehat{f}_{k-1}(X_i)}{\widehat{f}_k(X_i)}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\log\frac{L_k(\widehat{f}_{k-1})}{L_k(\widehat{f}_k)}$$

$$= \frac{1}{n}\sum_{k=1}^{n}\log\frac{L_{k-1}(\widehat{f}_{k-1})}{L_{k-1}(\widehat{f}_k)} + \frac{1}{n}\sum_{k=1}^{n}\log\frac{\widehat{f}_{k-1}(X_k)}{\widehat{f}_k(X_k)}.$$

Here $L_k(f) = \prod_{i=1}^{k} f(X_i)$. It then follows that

$$\frac{1}{n}\sum_{k=1}^{n}\log\frac{\widehat{f}_{k-1}(X_k)}{\widehat{f}_k(X_k)} \to 0$$

is sufficient for the Hellinger consistency of $f^n$.

In this case we can write

$$h(f^n, f_0) \leq \frac{1}{2n}\sum_{i=1}^{n}\log\frac{\widehat{f}_i(X_i)}{\widehat{f}_{i-1}(X_i)} + \varepsilon_n,$$

where $\varepsilon_n = S_n/n$ and $\varepsilon_n \to 0$, $E(\varepsilon_n) = 0$ and $\mathrm{Var}(\varepsilon_n) \leq 1/n$. This provides a useful upper bound for the Hellinger distance between $f^n$ and $f_0$. This allows data to be monitored as it arrives and a stopping rule determined by the size of

$$\frac{1}{2n}\sum_{i=1}^{n}\log\frac{\widehat{f}_i(X_i)}{\widehat{f}_{i-1}(X_i)},$$

which provides an upper bound for $h(f^n, f_0)$.

## 4. ILLUSTRATION

Here we will consider an example. Suppose that

$$f(x) = \exp\left\{\sum_{k=1}^{m}\tau_k\phi_k(x)\right\},$$

where $\phi_0(x) \equiv 1$ and the $\{\phi_k\}_{k=1}^m$ are a set of orthonormal functions. We assume that $f_0(x) = \exp\{\sum_{k=1}^m \tau_{k0}\phi_k(x)\}$ for some set $\{\tau_{k0}\}$. Now

$$\frac{1}{n}\sum_{i=1}^n \log\frac{f_0(X_i)}{f(X_i;\tau_{i-1})} = \frac{1}{n}\sum_{i=1}^n\sum_{k=0}^m(\tau_{k0}-\tau_{k,i-1})\,\phi_k(X_i)$$

$$\leq \sum_{k=0}^m \sqrt{\frac{1}{n}\sum_{i=1}^n(\tau_{k0}-\tau_{k,i-1})^2 \times \frac{1}{n}\sum_{i=1}^n \phi_k^2(X_i)}.$$

Consequently, we have a Hellinger consistency result for $f^n$ under the condition that the estimators $\tau_{ki}$ satisfy

$$\frac{1}{n}\sum_{i=1}^n(\tau_{ki}-\tau_{k0})^2 \to 0$$

and $\sup_k \int \phi_k^2(x)\, f_0(x)\,dx < \infty$.

For $\tau_{ki} \to \tau_{k0}$, see Crain (1974), who took the $\tau_{ki}$ to be the maximum likelihood estimators. Crain (1974) does establish $L_1$ consistency for his density estimator, however, this is under the restrictive assumption that $f_0$ has bounded support, say $[0,1]$, and also that $f_0(x) > 0$ for all $x \in [0,1]$, which is excluding the possibility that $f_0(0) = 0$ or $f_0(1) = 0$.

## Acknowledgements

The authors are grateful for the comments of two referees.

## References

Crain, B. R. (1974). "Estimation of distributions using orthogonal expansions", *The Annals of Statistics*, **2**, 454–463.

Loève, M. (1963). *Probability Theory*, 3rd ed., Van Nostrand, Toronto.

van de Geer, S. (1993). "Hellinger consistency of certain nonparametric maximum likelihood estimators", *The Annals of Statistics*, **21**, 14–44.

Walker, S. G. (2003). "On sufficient conditions for Bayesian consistency", *Biometrika*, **90**, 482–488.