IMS Collections
From Probability to Statistics and Back: High-Dimensional Models and Processes
Vol. 9 (2013) 47–64

DOI: 10.1214/12-IMSCOLL905

Robust tests for model selection

Lucien Birgé

Université Pierre et Marie Curie

Abstract: It was shown almost 40 years ago by Lucien Le Cam that the existence of suitable tests between Hellinger balls in the parameter set led to the construction of some sort of universal estimators for parametric statistical problems with i.i.d. observations. This idea of deriving estimators from families of robust tests was developed and substantially generalized in some of my previous work and more recently extended to Model Selection based estimation. Since the key ingredient for the design of such estimators for a given statistical framework is the construction of the relevant tests for this particular framework, it is essential to explain how to build them for as many different frameworks as possible. The purpose of this paper is to provide improved results about the existence of such tests for the problems of estimation based on independent (not necessarily i.i.d.) observations, estimation of conditional densities and of Markov transitions.

1. Introduction

The starting point of this work was a paper by Lucien Le Cam which appeared in 1973 when I was still a student and that I only read a few years later. In this fundamental paper, Le Cam addressed the problem of estimating a distribution P on the measurable space $(E, \mathcal{E})$ from an i.i.d. sample $X_1, \dots, X_n$ of this distribution. He showed that, if it is assumed that P belongs to some statistical model $\mathcal{P}$ (i.e. a set of probability measures) with a finite dimension (in a suitable sense), one can build an estimator $\widehat{P}_n(X_1, \dots, X_n) \in \mathcal{P}$ of P which converges at the rate $n^{-1/2}$ uniformly for $P \in \mathcal{P}$. More precisely, the maximal quadratic risk of $\widehat{P}_n$, $\sup_{P\in\mathcal{P}} \mathbb{E}_P[h^2(\widehat{P}_n, P)]$, is bounded by Cn^{-1} where C is a constant depending on the dimension of $\mathcal{P}$ only and h denotes the Hellinger distance. We recall that the Hellinger distance $h(P, Q)$ between two probabilities P and Q dominated by μ is given by

$$h^2(P,Q) = \frac{1}{2}\int (\sqrt{dP/d\mu} - \sqrt{dQ/d\mu})^2 \, d\mu, \tag{1.1}$$

the result being independent of the choice of the dominating measure μ. Le Cam also showed, under the same dimensionality assumptions, that if a suitable prior is given on $\mathcal{P}$, the posterior will concentrate around P at the rate $n^{-1/2}$ (for the Hellinger distance).

One justification for such a construction was the well-known fact that the traditional MLE method may not work well under the assumed dimensional restrictions (which cover most parametric problems with a compact parameter set). Many counterexamples have been known for a long time and numerous comments about the

CNRS UMR 7599 "Probabilités et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Pierre et Marie Curie, 4 Place Jussieu, F-75252 Paris Cedex 05, France, e-mail: lucien.birge@upmc.fr

AMS 2000 subject classifications: Primary 62G10, 62G35; secondary 62G05

Keywords and phrases: Robust testing, Hellinger distance, model selection, Markov chains

MLE are to be found in [30]. The situation becomes even worse for nonparametric problems although the MLE has been shown to be a good nonparametric estimator in some particular situations. Examples are provided by [2, 16–19, 32] and [33] among many others.

On the contrary, the estimator $\widehat{P}_n(X_1, \ldots, X_n)$ (based on i.i.d. observations) always exists under mild assumptions (typically Hellinger-compactness of the parameter space), even in nonparametric problems. The key argument involved in its design was the construction of tests between two convex subsets of the metric space (M, h) (where M denotes the set of all probability measures on E) the errors of which are controlled by the Hellinger distance between the two sets. More recent proofs of the existence of such tests can be found in Section 16.4 of [29] and an interesting discussion about the relationship between rates of convergence and dimensional properties of the parameter set is given in [31].

Le Cam's 1973 paper was the first of a series about the construction of estimators of a probability P belonging to a model $\mathcal{P}$ the properties of which are determined by the dimension of the model and an information index (number of observations for an i.i.d. sample, variance of the errors for Gaussian regression, etc.). In particular, in 1975, Le Cam extended his results to the case of independent but not necessarily i.i.d. observations (with some restrictions). My own work of the 80's, building on Le Cam's initial construction, but also on lower bound arguments developed by Ibragimov and Khas'minskii ([23–25] and [26]), was dedicated to extensions of Le Cam's approach in various directions: dealing with infinite-dimensional models (non-parametric problems), connecting upper and lower bounds and relating both to dimensional properties of the parameter set, improving Le Cam's results on testing for independent observations and extending them to some dependent cases ([6–8] and [10]).

At this stage, an important remark is in order. Since the construction of the estimator is based on robust tests (tests between balls) it is also robust. If the true distribution P does not belong to the assumed model $\mathcal{P}$, the estimator still exists and the quadratic Hellinger risk is only inflated by an extra "bias" term of order $\inf_{Q\in\mathcal{P}} h^2(P, Q)$ which means that the estimator can be based on approximate models as shown in [10]. Since I then worked for many years with Pascal Massart on "Model Selection", which is based on approximate models, I decided to relate this new research topic of the 90's to the general approach of Le Cam. This resulted in the construction of what I called *T-estimators* (T for "tests"). It is described in [12] which provides a general approach to Model Selection via testing with extensions to further statistical frameworks in [11] and [13].

There is nevertheless a drawback to this generality. Le Cam's approach to estimation, based on testing, is fairly abstract since it leads to a complicated construction that requires to perform too many tests to be practically implemented. This is even more true for my own extensions. However this construction has the theoretical advantage of handling cases that are more general that those one can handle with classical estimators: much greater generality but no real applicability. As we would say in French: "on ne peut avoir le beurre et l'argent du beurre"!

I already mentioned that the key argument in Le Cam's 1973 initial paper is the use of suitable tests between balls in the parameter set endowed with an appropriate metric and all subsequent works on the subject rely on the existence of such tests. This is also true for some more recent results—[14] and [15]—about the concentration of posterior distributions in non-parametric Bayesian frameworks. Each time one wants to apply the general theory to a new statistical framework, one has to demonstrate the existence of the relevant tests for this particular framework. It

has initially been proven for i.i.d. observations in [27] and an explicit construction of such tests was given in [8]. The case of independent non i.i.d. variables dates back to [28] and [7]. Gaussian sequences and bounded regression have been considered in [12], Gaussian regression with random design in [11] and Poisson processes in [13]. This new paper is devoted to provide a few examples of such tests that improve, extend, or clarify previous results on the subject.

2. A brief summary of some general results about model selection

2.1. Model based estimation

Let us consider here, to be specific, the problem of estimating a density s from an i.i.d. sample $X_1, \ldots, X_n$ when s belongs to some given set S of densities. Even if S is a compact parametric model, it is not at all obvious to design a completely general estimation method that leads to a quadratic risk of order n^{-1} for the Hellinger distance. It is well-known that the MLE is not the right solution in many situations. Moreover, from a realistic point of view, there are very few reasons why the chosen model S should actually contain the true unknown density s. A typical model is just an approximation of the truth. This has been recognized many years ago, giving birth to various types of *robust procedures*—see for instance [21]—. The problem of testing between balls and other convex sets was actually considered much earlier than 1973, not for estimating purposes but for robustness ones, in particular by Peter Huber. Milestones of the theory of robust testing are [20] and [22]. Extensions have been provided by [34], [5] and others. Recent results of [3] could also be used to derive robust tests between two probabilities based on the test statistic $T(N, t, t')$ defined in Section 2.2 of [3].

Even for the simplest case of a single model that is assumed to contain the true parameter, Le Cam's basic construction is derived from robust tests between balls in the parameter space. In the more general situation of a model S with metric dimension bounded by D (see the precise definition below) that may not contain s, the quadratic risk of a good estimator with values in S will be the sum of an approximation term (square of the distance from s to S) and a fluctuation term the size of which is roughly proportional to D/n. Since the approximation term is unknown, the design of a model which is suitable for estimating the parameter s is quite difficult. One can try to improve the procedure by introducing several models simultaneously, hoping for better approximation properties with respect to s. An optimal model within a given family is therefore one for which the sum of these two error terms (approximation and fluctuation) is minimal and "Model Selection" aims at choosing such a model from the observation of the data only. In this situation, the models typically do not contain the true parameter s (and even if one does, it is not necessarily the best one). Therefore some specific robustness properties of the estimation procedures based on the models are necessary to perform Model Selection. This is why tests between balls are at the chore of the theory for Model Selection (with applications to adaptive estimation or estimators aggregation) that I tried to build in [12]. It is actually the continuation of [6] which dealt with a single discrete model for approximating a compact parameter set and results in the construction of T-estimators that I shall now summarize.

2.2. About T-estimators

In order to explain what sort of tests are needed to make the whole construction work, let us first describe our general setting. We observe some random element

$\boldsymbol{X}$ from $(\Omega, \mathcal{A})$ to $(\Xi, \mathcal{X})$ with an unknown distribution belonging to some set $\mathcal{P} = \{P_s, s \in \mathcal{S}\}$, indexed by $\mathcal{S}$, of possible distributions on $(\Xi, \mathcal{X})$, where $\mathcal{S}$ denotes a given subset of some metric space (M, d). A simple example is provided by an i.i.d. sample $\boldsymbol{X} = (X_1, \ldots, X_n)$, s denoting the unknown density of the X_i with respect to some given σ-finite dominating measure μ and M being the space $\mathbb{L}_1(\mu)$. We of course assume that the mapping $s \mapsto P_s$ is one-to-one. We denote by $\mathbb{E}_P[f(\boldsymbol{X})]$ and $\mathbb{E}_s[f(\boldsymbol{X})]$ the expectation of $f(\boldsymbol{X})$ when $\boldsymbol{X}$ has the distribution P, respectively P_s, with a similar convention for $\mathbb{P}_s$. We also write $a \vee b$ for $\max\{a, b\}$, $\mathcal{B}_d(t, z)$ for the open ball of center t and radius z in (M, d) and $|\mathcal{Q}|$ for the cardinality of the set $\mathcal{Q}$.

The construction of T-estimators requires *models*. A model $S \subset M$ is merely an approximation space for s. We shall restrict our attention here to models with *a bounded metric dimension* ([12], Definition 6 page 293). We recall that a subset S of some metric space (M, d) has *a metric dimension bounded by* D if, for every $\eta > 0$, there exists an η-net S_η for S which satisfies

$$|S_\eta \cap \mathcal{B}_d(t, z\eta)| \le \exp\bigl[Dz^2\bigr] \quad \text{for all } z \ge 2 \text{ and } t \in M. \tag{2.1}$$

Given such a model and tests between balls in M with centers in S, one can build an estimator with values in S. If one works with many models simultaneously, the procedure becomes slightly more complex but the construction is still based on the same ingredients: models and tests between balls.

Let us now be more precise about our assumptions and results. The distance $d(t, S)$ from some point $t \in M$ to some subset S of M is defined as $d(t, S) = \inf_{u \in S} d(t, u)$. We also need a proper definition of tests between the elements of M.

Definition 1. Given a random element $\boldsymbol{X}$ with values in Ξ and two distinct points t and $u \in M$, a test between t and u is a measurable function $\psi_{t,u}(\boldsymbol{X})$ with values in $\{t; u\}$, our convention being that $\psi_{t,u}(\boldsymbol{X}) = t$ means accepting t while $\psi_{t,u}(\boldsymbol{X}) = u$ means accepting u.

We shall stick to this convention throughout the paper.

The key assumption for the construction of T-estimators is the existence of some subset M_T of M and tests between the points of M_T with the following properties.

Assumption 1. *There exists a subset M_T of M and constants $a > 0$, $\kappa > 2$, $B > 0$ such that, for any pair $(t, u) \in M_T^2$ with $t \ne u$ and any $z \in \mathbb{R}$, one can find a test $\psi_{t,u}(\boldsymbol{X})$ satisfying*

$$\sup_{\{s \in M \,|\, \kappa d(s,t) \le d(t,u)\}} \mathbb{P}_s\bigl[\psi_{t,u}(\boldsymbol{X}) = u\bigr] \le B \exp\bigl[-a\bigl(d^2(t, u) + z\bigr)\bigr]; \tag{2.2}$$

$$\sup_{\{s \in M \,|\, \kappa d(s,u) \le d(t,u)\}} \mathbb{P}_s\bigl[\psi_{t,u}(\boldsymbol{X}) = t\bigr] \le B \exp\bigl[-a\bigl(d^2(t, u) - z\bigr)\bigr]. \tag{2.3}$$

The construction of T-estimators also involves a finite or countable family $\{S_m, m \in \mathcal{M}\}$ of models with respective metric dimensions bounded by D_m which are subsets of M_T. Given such a family, we fix numbers η_m for $m \in \mathcal{M}$ that satisfy the following requirements :

$$\sum_{m \in \mathcal{M}} \exp\bigl[-a\eta_m^2/21\bigr] = \Sigma < +\infty, \quad \text{and} \quad a\eta_m^2 \ge 21 D_m/5 \quad \text{for all } m \in \mathcal{M}. \tag{2.4}$$

A typical choice of the numbers η_m is as follows: select numbers Δ_m such that $\sum_{m \in \mathcal{M}} \exp[-\Delta_m] = \Sigma$ (or set a prior on $\mathcal{M}$ with a probability $\exp[-\Delta_m]$ for m so

that $\Sigma = 1$) and set

$$\eta_m = \sqrt{21a^{-1}\big((D_m/5) \vee \Delta_m\big)}. \tag{2.5}$$

Then, for each $m \in \mathcal{M}$, we chose an η_m-net S'_m for S_m that satisfies

$$|S'_m \cap \mathcal{B}_d(t, z\eta_m)| \leq \exp\big[D_m z^2\big] \quad \text{for all } z \geq 2 \text{ and } t \in M, \tag{2.6}$$

which is possible according to the definition of bounded metric dimension. Then, for each $t \in S' = \cup_{m\in\mathcal{M}} S'_m$, we set $\eta(t) = \inf\{\eta_m,\, m \in \mathcal{M} \,|\, t \in S'_m\}$. For each pair $(t, u) \in S' \times S'$ with $t \neq u$, we use Assumption 1 to design a test $\psi_{t,u}$ between t and u satisfying (2.2) and (2.3) with $z = \eta^2(u) - \eta^2(t)$.

To build the corresponding T-estimator, we set, for each $t \in S'$, $\mathcal{R}_t = \{u \in S', u \neq t \,|\, \psi_{t,u}(\boldsymbol{X}) = u\}$ and we define the random function $\mathcal{D}_{\boldsymbol{X}}$ on S' by

$$\mathcal{D}_{\boldsymbol{X}}(t) = \begin{cases} \sup\limits_{u\in\mathcal{R}_t} \big\{d(t,u)\big\} & \text{if } \mathcal{R}_t \neq \emptyset; \\ 0 & \text{if } \mathcal{R}_t = \emptyset. \end{cases} \tag{2.7}$$

A *T-estimator* is any measurable application $\widehat{s}(\boldsymbol{X})$ with values in S' which minimizes $\mathcal{D}_{\boldsymbol{X}}$. When such a minimizer does not exist, we replace it by an approximate minimizer—see [12] for details—but we shall not insist on this here, simply assuming that the minimizer exists. The properties of the T-estimator $\widehat{s}(\boldsymbol{X})$ are given by the following theorem—see Corollary 4 of [12]—.

Theorem 1. *Under the previous assumptions with $\kappa \geq 4$, the T-estimator $\widehat{s}(\boldsymbol{X})$ satisfies, for $z \geq (\kappa + 1) \inf_{m\in\mathcal{M}}[d(s, S_m) \vee \eta_m]$,*

$$\mathbb{P}_s\big[d(s, \widehat{s}) > z\big] < (B\Sigma/7) \exp\big[-(32/75)az^2\big]. \tag{2.8}$$

Consequently, for all $q \geq 1$ and $s \in M$,

$$\mathbb{E}_s\big[d^q(s, \widehat{s})\big] \leq [1 + B\Sigma\zeta_q](\kappa + 1)^q \inf_{m\in\mathcal{M}} \big\{d(s, S_m) + \eta_m\big\}^q, \tag{2.9}$$

where ζ_q denotes a constant depending on q only.

It is essential to notice that the construction of T-estimators, as described above, involves the constant a (via the choice of the discretization parameters η_m) which has therefore to be known. But this construction does not make use of κ and B which only influence the risk bounds for T-estimators. When the constant a is partially unknown (if we only know that $a \leq a_0$) and the models have a metric structure which resembles that of Euclidean spaces (which is more restrictive than having a bounded metric dimension in the above sense), it is possible, in some cases, to modify the construction of T-estimators, replacing a by a_0. The resulting risk bounds are then multiplied by some power of $\log(a_0/a)$ and therefore suboptimal but this may be useful to deal with cases where a is unknown. We shall not insist on this extension here although it could be applied to the situations of Sections 5 and 6.

Since the statistician is absolutely free to choose suitable models, the only ingredient which is definitely necessary in order to apply Theorem 1 is Assumption 1 about the existence of robust tests between t and u that satisfy (2.2) and (2.3). For any statistical framework for which Assumption 1 holds with a proper choice of the distance d, Theorem 1 applies to any family of models S_m with a bounded metric

dimension provided that the numbers η_m have been conveniently chosen (typically satisfying (2.5)). Such families of models with nice approximating properties with respect to various sorts of functions s have been provided in the different papers devoted to T-estimators and more recently in [4] to deal with some complicated multivariate parameters. We shall now focus on the construction of tests that do possess the required properties.

3. The basic result

3.1. Preliminary considerations

Of special importance throughout the paper are the Hellinger distance and affinity between probabilities. Given two probabilities P and Q dominated by μ, the Hellinger distance h between P and Q is given by (1.1) and their Hellinger affinity ρ by

$$\rho(P,Q) = 1 - h^2(P,Q) = \int \sqrt{(dP/d\mu)(dQ/d\mu)}\, d\mu. \tag{3.1}$$

A useful tool for building tests between two probabilities P and Q is the following elementary lemma.

Lemma 1. *Let $\boldsymbol{X}$ be a random variable on some measurable space $(\Xi, \mathcal{X})$ and P, Q two probabilities on Ξ. Let ϕ be a non-negative measurable function on Ξ such that*

$$\mathbb{E}_P\big[\phi(\boldsymbol{X})\big] \le \exp a \qquad \textit{and} \qquad \mathbb{E}_Q\big[1/\phi(\boldsymbol{X})\big] \le \exp b. \tag{3.2}$$

Then, for all $z \in \mathbb{R}$,

$$\mathbb{P}_P\big[\log \phi(\boldsymbol{X}) \ge z\big] \le \exp[a - z] \qquad \textit{and} \qquad \mathbb{P}_Q\big[\log \phi(\boldsymbol{X}) \le z\big] \le \exp[b + z].$$

This lemma shows that if we are able to find a function ϕ satisfying (3.2), we immediately derive from it a test between P and Q with controlled errors, the role of z being to balance the two errors in a way chosen by the statistician. The next section will therefore be devoted to building such functions ϕ.

3.2. Fundamental inequalities

Let P_0 and P_1 ($P_0 \ne P_1$) be probabilities on Ξ and μ be some dominating measure which will not play any special role here, the results being independent of the choice of μ. We then set $v_i = \sqrt{dP_i/d\mu}$ for $i = 0, 1$ and denote by V the two-dimensional linear subspace of $\mathbb{L}_2(\mu)$ spanned by v_0 and v_1, by $\overline{V}$ the subset of V of elements of norm 1 (so that v_0 and v_1 belong to $\overline{V}$), by $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ respectively the inner product and norm in $\mathbb{L}_2(\mu)$. We define ω by

$$\rho(P_0, P_1) = \langle v_0, v_1 \rangle = \cos\omega, \qquad 0 < \omega \le \pi/2,$$

so that

$$h^2(P_0, P_1) = 1 - \cos\big(2(\omega/2)\big) = 2\sin^2(\omega/2).$$

For $\beta \in [0, 2\pi/\omega)$, we denote by v_β the element of $\overline{V}$ which is deduced from v_0 by a rotation of angle $\beta\omega$ in the direction of v_1 so that that v_β^2 is a probability density with respect to μ. It follows that

$$v_\beta = \big[\sin(\beta\omega)\, v_1 + \sin\big((1-\beta)\omega\big)\, v_0\big]/(\sin\omega) \tag{3.3}$$

and

$$\langle v_\alpha, v_\beta\rangle = \cos\bigl((\alpha-\beta)\omega\bigr). \tag{3.4}$$

We can now state the main result which improves over the previous ones in [8] (and Le Cam's results as well) since it allows to bound the error of the tests derived from ϕ_ξ via Lemma 1 whatever the true distribution R of $\boldsymbol{X}$ be.

Theorem 2. *Let P_0 and P_1 be two probabilities on Ξ such that $\rho(P_0,P_1)=\cos\omega$ with $0<\omega\le\pi/2$. For $\xi\in(0,1/2)$, let $A_\xi = 2[\sin((1-2\xi)\omega)][\sin(\xi\omega)]^{-1}$ and $\phi_\xi = v_{1-\xi}/v_\xi$ with v_ξ and $v_{1-\xi}$ given by (3.3) and $0/0=1$. Then ϕ_ξ satisfies the following property: for any random variable $\boldsymbol{X}$ with an arbitrary distribution R on Ξ,*

$$\mathbb{E}_R\bigl[\phi_\xi(\boldsymbol{X})\bigr] \le 1-(1-2\xi)^2h^2(P_0,P_1) - A_\xi\bigl[\xi^2h^2(P_0,P_1)-h^2(R,P_0)\bigr] \tag{3.5}$$

and

$$\mathbb{E}_R\left[\frac{1}{\phi_\xi(\boldsymbol{X})}\right] \le 1-(1-2\xi)^2h^2(P_0,P_1) - A_\xi\bigl[\xi^2h^2(P_0,P_1)-h^2(R,P_1)\bigr]. \tag{3.6}$$

In particular, for any distribution R,

$$\mathbb{E}_R\bigl[\phi_{1/3}(\boldsymbol{X})\bigr] \le 1-\frac{h^2(P_0,P_1)}{3}+2h^2(R,P_0)$$

and

$$\mathbb{E}_R\left[\frac{1}{\phi_{1/3}(\boldsymbol{X})}\right] \le 1-\frac{h^2(P_0,P_1)}{3}+2h^2(R,P_1).$$

3.3. Proof of Theorem 2

It relies on the following lemma.

Lemma 2. *Let P,Q and R be three probabilities with respective densities p,q,r with respect to μ. If P is absolutely continuous with respect to Q, $\sqrt{p/q}$ is bounded by λ (with the convention that $\sqrt{p/q}=1$ when $q=0$) and $\boldsymbol{X}$ is a random variable with distribution R, then*

$$\mathbb{E}_R\bigl[\sqrt{p(\boldsymbol{X})/q(\boldsymbol{X})}\bigr] \le 2\lambda h^2(R,Q)+2\rho(R,P)-\rho(P,Q). \tag{3.7}$$

Proof. The left-hand side of (3.7) can be written

$$\int r\sqrt{p/q}\,d\mu = \int \sqrt{p/q}(\sqrt{r}-\sqrt{q})^2\,d\mu + 2\int\sqrt{pr}\,d\mu - \int\sqrt{pq}\,d\mu,$$

hence the result. □

To prove Theorem 2 we set $P_\xi = v_\xi^2\cdot\mu$, $P_{1-\xi}=v_{1-\xi}^2\cdot\mu$ and we apply Lemma 2 with $P=P_{1-\xi}$ and $Q=P_\xi$ so that $\sqrt{p(\boldsymbol{X})/q(\boldsymbol{X})} = v_{1-\xi}(\boldsymbol{X})/v_\xi(\boldsymbol{X})$ since v_β is nonnegative for $0\le\beta\le1$. By definition, for all $x\in\Xi$,

$$\frac{v_{1-\xi}(x)}{v_\xi(x)} = \frac{\sin(\xi\omega)v_0(x)+\sin((1-\xi)\omega)v_1(x)}{\sin((1-\xi)\omega)v_0(x)+\sin(\xi\omega)v_1(x)} \le \frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)},$$

since $1 > (1-\xi) > \xi > 0$. This implies that $\mathbb{E}_R[v_{1-\xi}(\boldsymbol{X})/v_\xi(\boldsymbol{X})] \le K$ with

$$K = 2\frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)}h^2(R,P_\xi) + 2\rho(R,P_{1-\xi}) - \rho(P_\xi,P_{1-\xi}).$$

Note that any element of V can be written as θv_γ with $\theta \ge 0$ and $\gamma \in [0, 2\pi/\omega)$. Assuming, without loss of generality, that $R \ll \mu$, we may write $\sqrt{dR/d\mu} = u + \theta v_\gamma$ with $0 \le \theta \le 1$, u orthogonal to V and $\theta^2 + \|u\|^2 = 1$. It follows by (3.4) that

$$\rho(R,P_\xi) = \theta\cos\bigl((\gamma-\xi)\omega\bigr), \quad \rho(R,P_{1-\xi}) = \theta\cos\bigl((\gamma-1+\xi)\omega\bigr)$$

and $\rho(P_\xi,P_{1-\xi}) = \cos((1-2\xi)\omega)$, so that

$$\begin{aligned}
K &= 2\frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)}\bigl[1-\theta\cos\bigl((\gamma-\xi)\omega\bigr)\bigr] \\
&\quad + 2\theta\cos\bigl((\gamma-1+\xi)\omega\bigr) - \cos\bigl((1-2\xi)\omega\bigr) \\
&= 2\theta\left[\cos\bigl((\gamma-1+\xi)\omega\bigr) - \frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)}\cos\bigl((\gamma-\xi)\omega\bigr)\right] \\
&\quad + 2\frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)} - \cos\bigl((1-2\xi)\omega\bigr).
\end{aligned}$$

We now successively apply trigonometric formulas to get

$$\begin{aligned}
&\cos\bigl((\gamma-1+\xi)\omega\bigr) - \frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)}\cos\bigl((\gamma-\xi)\omega\bigr) \\
&\quad = \cos(\gamma\omega)\cos\bigl((1-\xi)\omega\bigr) + \sin(\gamma\omega)\sin\bigl((1-\xi)\omega\bigr) \\
&\qquad - \frac{\sin((1-\xi)\omega)}{\sin(\xi\omega)}\cos(\gamma\omega)\cos(\xi\omega) - \sin(\gamma\omega)\sin\bigl((1-\xi)\omega\bigr) \\
&\quad = -\frac{\cos(\gamma\omega)}{\sin(\xi\omega)}\sin\bigl((1-2\xi)\omega\bigr).
\end{aligned}$$

It follows that $K = \cos((1-2\xi)\omega) + 2[\sin(\xi\omega)]^{-1}K_1$ with

$$\begin{aligned}
K_1 &= \sin\bigl((1-\xi)\omega\bigr) - \theta\cos(\gamma\omega)\sin\bigl((1-2\xi)\omega\bigr) - \cos\bigl((1-2\xi)\omega\bigr)\sin(\xi\omega) \\
&= \sin\bigl((1-2\xi)\omega\bigr)\cos(\xi\omega) + \sin(\xi\omega)\cos\bigl((1-2\xi)\omega\bigr) \\
&\quad - \theta\cos(\gamma\omega)\sin\bigl((1-2\xi)\omega\bigr) - \cos\bigl((1-2\xi)\omega\bigr)\sin(\xi\omega) \\
&= \sin\bigl((1-2\xi)\omega\bigr)\bigl[\cos(\xi\omega) - \theta\cos(\gamma\omega)\bigr].
\end{aligned}$$

Since $\rho(P_\xi,P_0) = \cos(\xi\omega)$ and $\rho(R,P_0) = \theta\cos(\gamma\omega)$, we finally get

$$\begin{aligned}
K &= \rho(P_\xi,P_{1-\xi}) + 2\frac{\sin((1-2\xi)\omega)}{\sin(\xi\omega)}\bigl[\rho(P_\xi,P_0) - \rho(R,P_0)\bigr] \\
&= 1 - h^2(P_\xi,P_{1-\xi}) + 2\frac{\sin((1-2\xi)\omega)}{\sin(\xi\omega)}\bigl[h^2(R,P_0) - h^2(P_\xi,P_0)\bigr].
\end{aligned}$$

One then observes that

$$h^2(P_\xi,P_{1-\xi}) = 1 - \cos\bigl((1-2\xi)\omega\bigr) = 2\sin^2\bigl((1-2\xi)\omega/2\bigr),$$

$$h^2(P_\xi,P_0) = h^2(P_{1-\xi},P_1) = 2\sin^2(\xi\omega/2) \qquad \text{and} \qquad h^2(P_0,P_1) = 2\sin^2(\omega/2).$$

Since the function $x \mapsto x^{-1}\sin x$ is decreasing for $0 \le x \le \pi/2$, we deduce that $h^2(P_\xi,P_{1-\xi}) \ge (1-2\xi)^2 h^2(P_0,P_1)$ and $h^2((P_\xi,P_0) \ge \xi^2 h^2(P_0,P_1)$, which proves (3.5). Exchanging the roles of P_ξ and $P_{1-\xi}$ gives (3.6).

4. Tests for independent variables

In this section, we want to apply the previous results to the situation of independent observations, i.e. the case of $\boldsymbol{X} = (X_1, \ldots, X_n)$ for independent variables X_i with respective distributions $\overline{R}_i$, $1 \le i \le n$, on the measurable space $(E, \mathcal{E})$. This means that the distribution $R = \bigotimes_{i=1}^n \overline{R}_i$ of $\boldsymbol{X}$ belongs to the space $\mathfrak{P}$ of product probabilities on $\Xi = E^n$. We define the distance H on $\mathfrak{P}$ by

$$H^2(P,Q) = \sum_{i=1}^n h^2(\overline{P}_i, \overline{Q}_i) \quad \text{for } P = \bigotimes_{i=1}^n \overline{P}_i \quad \text{and} \quad Q = \bigotimes_{i=1}^n \overline{Q}_i.$$

4.1. Preliminary inequalities

Let us choose a value of $\xi \in (0, 1/2)$. To each pair $(\overline{P}_i, \overline{Q}_i)$ of probabilities on E with $\rho(\overline{P}_i, \overline{Q}_i) = \cos\omega_i$, we apply the results of Theorem 2 (with $P = \overline{P}_i$ and $Q = \overline{Q}_i$) in order to derive a function ϕ_i which satisfies (3.5) and (3.6). Setting $A_i = 2[\sin((1-2\xi)\omega_i)][\sin(\xi\omega_i)]^{-1}$, we get from (3.5),

$$\begin{aligned}\mathbb{E}_{\overline{R}_i}\big[\phi_i(X_i)\big] &\le 1 - (1-2\xi)^2 h^2(\overline{P}_i, \overline{Q}_i) - A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{P}_i, \overline{R}_i)\big] \\ &\le \exp\big[-(1-2\xi)^2 h^2(\overline{P}_i, \overline{Q}_i) - A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{P}_i, \overline{R}_i)\big]\big]\end{aligned}$$

and from (3.6),

$$\begin{aligned}\mathbb{E}_{\overline{R}_i}\left[\frac{1}{\phi_i(X_i)}\right] &\le 1 - (1-2\xi)^2 h^2(\overline{P}_i, \overline{Q}_i) - A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{Q}_i, \overline{R}_i)\big] \\ &\le \exp\big[-(1-2\xi)^2 h^2(\overline{P}_i, \overline{Q}_i) - A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{Q}_i, \overline{R}_i)\big]\big].\end{aligned}$$

Setting $\phi_\xi(\boldsymbol{X}) = \prod_{i=1}^n \phi_i(X_i)$, we derive that

$$\begin{aligned}&\mathbb{E}_R\big[\phi_\xi(\boldsymbol{X})\big] \\ &\quad\le \exp\left[-(1-2\xi)^2 H^2(P,Q) - \sum_{i=1}^n A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{P}_i, \overline{R}_i)\big]\right]\end{aligned} \tag{4.1}$$

and

$$\begin{aligned}&\mathbb{E}_R\left[\frac{1}{\phi_\xi(\boldsymbol{X})}\right] \\ &\quad\le \exp\left[-(1-2\xi)^2 H^2(P,Q) - \sum_{i=1}^n A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{Q}_i, \overline{R}_i)\big]\right].\end{aligned} \tag{4.2}$$

It finally follows from Lemma 1 that

$$\begin{aligned}&\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \ge z\big] \\ &\quad\le \exp\left[-z - (1-2\xi)^2 H^2(P,Q) - \sum_{i=1}^n A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{P}_i, \overline{R}_i)\big]\right]\end{aligned} \tag{4.3}$$

and

$$\begin{aligned}&\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \le z\big] \\ &\quad\le \exp\left[z - (1-2\xi)^2 H^2(P,Q) - \sum_{i=1}^n A_i\big[\xi^2 h^2(\overline{P}_i, \overline{Q}_i) - h^2(\overline{Q}_i, \overline{R}_i)\big]\right].\end{aligned} \tag{4.4}$$

These bounds should be interpreted in the following way. Given P and Q belonging to $\mathfrak{P}$, we have derived a test between them which accepts P when $\log(\phi_\xi(\boldsymbol{X})) < z$ and rejects P when $\log(\phi_\xi(\boldsymbol{X})) > z$, the decision being arbitrary in case of equality. The errors of this test are bounded according to (4.3) and (4.4) whatever the true joint distribution R of the independent observations $X_1, \dots, X_n$ be.

4.2. The i.i.d. case

When $P = \overline{P}^{\otimes n}$ and $Q = \overline{Q}^{\otimes n}$ are distributions of i.i.d. random variables, then $H^2(P,Q) = nh^2(\overline{P},\overline{Q})$ and

$$A_i = 2\big[\sin\big((1-2\xi)\omega\big)\big]\big[\sin(\xi\omega)\big]^{-1} \text{ for all } i, \quad \text{with } \rho(\overline{P},\overline{Q}) = \cos\omega.$$

The next corollary of Theorem 2 then follows straightforwardly from (4.3) and (4.4).

Corollary 1. *Let $X_1, \dots, X_n$ have the joint distribution $R = \bigotimes_{i=1}^n \overline{R}_i$ and ϕ_ξ be defined as previously indicated in Section 4.1 in order to satify (4.1) and (4.2). Let us set $A_\xi = 2[\sin((1-2\xi)\omega)][\sin(\xi\omega)]^{-1}$ with $\xi \in (0,1/2)$. Then for any real number z,*

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \ge z\big] \le \exp\big[-z - nh^2(\overline{P},\overline{Q})\big(A_\xi\xi^2 + (1-2\xi)^2\big) + A_\xi H^2(P,R)\big],$$

while

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \le z\big] \le \exp\big[z - nh^2(\overline{P},\overline{Q})\big(A_\xi\xi^2 + (1-2\xi)^2\big) + A_\xi H^2(Q,R)\big].$$

In particular, if $h(\overline{P},\overline{R}_i) \le \xi h(\overline{P},\overline{Q})$ for $1 \le i \le n$ or, more generally, $H^2(P,R) \le n\xi^2 h^2(\overline{P},\overline{Q})$, then

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \ge z\big] \le \exp\big[-z - n(1-2\xi)^2 h^2(\overline{P},\overline{Q})\big]$$

and if $h(\overline{Q},\overline{R}_i) \le \xi h(\overline{P},\overline{Q})$ for $1 \le i \le n$ or $H^2(Q,R) \le n\xi^2 h^2(\overline{P},\overline{Q})$, then

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \le z\big] \le \exp\big[z - n(1-2\xi)^2 h^2(\overline{P},\overline{Q})\big].$$

This means that if we use the statistic $\log(\phi_\xi(\boldsymbol{X}))$ to test between P and Q, we can bound the errors of the test whatever the true distribution R of the independent variables $X_1, \dots, X_n$ be. The "i.i.d. case" actually refers to the probabilities P and Q, not to the the true distribution of the variables $X_1, \dots, X_n$. The last error bounds show that the test is actually a test between two balls in the metric space $(\mathfrak{P}, H)$ with the same radius $\xi H(P,Q)$ and respective centers P and Q. When $z = 0$, both errors of this test are bounded by $\exp[-n(1-2\xi)^2 h^2(\overline{P},\overline{Q})]$. Note that this is an improvement over the initial version of [8] which could only deal with the cases of $h(\overline{P},\overline{R}_i) \le \xi h(\overline{P},\overline{Q})$ for $1 \le i \le n$ or $h(\overline{Q},\overline{R}_i) \le \xi h(\overline{P},\overline{Q})$ for $1 \le i \le n$.

4.3. The general independent case

In the case of arbitrary elements $P, Q \in \mathfrak{P}$, we get the following result.

Corollary 2. *Let $X_1, \dots, X_n$ have the joint distribution $R = \bigotimes_{i=1}^n \overline{R}_i \in \mathfrak{P}$, $0 < \xi \le 1/3$, let ϕ_ξ be defined as previously indicated in Section 4.1 in order to satify (4.1) and (4.2) and let z be an arbitrary real number. Then*

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \ge z\big]$$
$$\le \exp\left[-z - \left((1-2\xi)^2 + 2\xi^2\frac{\sin((1-2\xi)\pi/2)}{\sin(\xi\pi/2)}\right)H^2(P,Q) + 2\frac{1-2\xi}{\xi}H^2(P,R)\right]$$

and

$$\mathbb{P}_R\big[\log\big(\phi_\xi(\boldsymbol{X})\big) \le z\big] \le \exp\left[z - \left((1-2\xi)^2 + 2\xi^2\frac{\sin((1-2\xi)\pi/2)}{\sin(\xi\pi/2)}\right)H^2(P,Q) + 2\frac{1-2\xi}{\xi}H^2(Q,R)\right].$$

In particular,

$$\mathbb{P}_R\big[\log\big(\phi_{1/3}(\boldsymbol{X})\big) \ge z\big] \le \exp\left[-z - \frac{1}{3}H^2(P,Q) + 2H^2(P,R)\right] \tag{4.5}$$

and

$$\mathbb{P}_R\big[\log\big(\phi_{1/3}(\boldsymbol{X})\big) \le z\big] \le \exp\left[z - \frac{1}{3}H^2(P,Q) + 2H^2(Q,R)\right]. \tag{4.6}$$

Proof. To derive the first bounds from (4.3) and (4.4) we merely observe that $\xi \le 1-2\xi$ and the function $z \mapsto \sin((1-2\xi)z)/\sin(\xi z)$ is decreasing on $[0,\pi/2]$ so that

$$\frac{\sin((1-2\xi)\pi/2)}{\sin(\xi\pi/2)} \le \frac{A_i}{2} = \frac{\sin((1-2\xi)\omega_i)}{\sin(\xi\omega_i)} \le \frac{1-2\xi}{\xi}.$$

The case of $\xi = 1/3$ immediately follows. □

Again, this is an improvement over the treatment of [7] or [29], Section 16.4.

4.4. An application to fixed design regression

Corollary 2 typically applies to fixed-design regression as considered in Section 8 of [3]. Let us provide here another illustration with independent observations

$$X_i = s_i + \xi_i, \quad 1 \le i \le n,$$

where the errors ξ_i are i.i.d. with a common known probability distribution P and the unknown vector $s = (s_i)_{1\le i\le n} \in \mathbb{R}^n$ to be estimated belongs to a given compact subset $\mathcal{K}$ of $\mathbb{R}^n$ so that $\sup_{1\le i\le n} s_i \le K$. In the fixed-design regression case, $s_i = s(x_i)$ where s is the unknown regression function and the x_i are the design points.

We denote by P_x the distribution of $\xi_1 + x$. For many distributions P, there exist numbers a, b, α such that, for all $x, y \in [-K, K]$,

$$a|x-y| \le h^\alpha(P_x, P_y) \le b|x-y|, \quad \text{with } 0 < a \le b < +\infty, \ \ \alpha \ge 1. \tag{4.7}$$

For instance, if the translation family $\{P_x, x \in \mathbb{R}\}$ is differentiable in quadratic mean, then $\alpha = 1$ and if P is a uniform distribution, then $\alpha = 2$.

Since the X_i are independent non i.i.d. random variables, it follows from the previous sections that the natural distance between two vectors $t = (t_i)_{1\le i\le n}$ and $u = (u_i)_{1\le i\le n}$ of $\mathbb{R}^n$ is H given by $H^2(t,u) = \sum_{i=1}^n h^2(P_{t_i}, P_{u_i})$. Then, by (4.7), for $t, u \in \mathcal{K}$,

$$A^2 \sum_{i=1}^n |t_i - u_i|^{2/\alpha} \le H^2(t,u) \le B^2 \sum_{i=1}^n |t_i - u_i|^{2/\alpha} \quad \text{with} \quad A = a^{1/\alpha}, \ B = b^{1/\alpha}.$$

Therefore if

$$\mathcal{V}(t,r)=\left\{u\in\mathbb{R}^n \mid \sum_{i=1}^n |t_i-u_i|^{2/\alpha}\le r^2\right\},$$

then

$$\mathcal{V}(t,r/B)\subset\mathcal{B}_H(t,r)\subset\mathcal{V}(t,r/A). \tag{4.8}$$

Let S be a D-dimensional linear subspace of $\mathbb{R}^n$ and set, for $t\in S$, $\mathcal{V}_S(t,r)=\mathcal{V}(t,r)\cap S$. The volume of the set $\mathcal{V}_S(t,r)$ (with respect to the Lebesgue measure on S) is independent of $t\in S$ so that it is a function $v(r)$ of r only. Moreover it follows from the properties of the Lebesgue measure that $v(\lambda r)=\lambda^{\alpha D}v(r)$ for $\lambda>0$.

Let us now consider in the metric space (S,H) some maximal η-separated set S_η (which means that points in S_η are at distances larger than η). Then S_η is an η-net for S. In order to bound the metric dimension of S, we have to bound the number of points of S_η that belong to a ball $\mathcal{B}_H$ in S with radius $x\eta$, $x\ge 2$. All balls with centers in $\mathcal{B}_H\cap S_\eta$ and radius $\eta/2$ are disjointed and they are all contained in a ball of radius $(x+1/2)\eta\le 5x\eta/4$. By (4.8) this ball is itself contained in some set $\mathcal{V}_S(\cdot,5x\eta/(4A))$ while each ball of radius $\eta/2$ contains some $\mathcal{V}_S(\cdot,\eta/(2B))$. It follows that $|\mathcal{B}_H\cap S_\eta|$ is bounded by the ratio of the volumes of the corresponding $\mathcal{V}_S$ sets, i.e. $v(5x\eta/(4A))/v(\eta/(2B))=[5xB/(2A)]^{\alpha D}$. Let $C=5B/(2A)\ge 5/2$. For $x\ge 2$, $x^{-2}\log(Cx)\le(1/4)\log(2C)$, hence

$$\begin{aligned}|\mathcal{B}_H\cap S_\eta|&\le\exp\bigl[\alpha D\log(Cx)\bigr]\\&\le\exp\bigl[(\alpha D/4)\log(2C)x^2\bigr]=\exp\bigl[(\alpha D/4)\log(5B/A)x^2\bigr].\end{aligned}$$

It therefore follows from (2.1) that the metric dimension of S with respect to the distance H is bounded by $(\alpha D/4)\log(5B/A)$.

If we now consider a finite or countable family $\{S_m,m\in\mathcal{M}\}$ of linear subspaces of $\mathbb{R}^n$ with respective dimensions D_m and a family of positive numbers Δ_m such that $\sum_{m\in\mathcal{M}}\exp[-\Delta_m]\le 1$, we may apply Theorem 1 and derive from this family of models an estimator $\widehat{s}$ of s satisfying

$$\mathbb{E}_s\bigl[H^2(s,\widehat{s})\bigr]\le C'\inf_{m\in\mathcal{M}}\bigl\{H^2(s,S_m)+\Delta_m+(\alpha D_m/4)\log(5B/A)\bigr\},$$

where C' is a universal constant.

5. Conditional densities

In this section, we deal with the problem of estimating the conditional density of $Y\in F$ given $Z\in G$ with respect to some dominating σ-finite measure μ on F from n i.i.d. pairs of observations $X_i=(Y_i,Z_i)$, $1\le i\le n$. We assume that the Z_i have an unknown density f with respect to some reference σ-finite measure ν on G and we denote by $s(\cdot|z)$ the unknown conditional density of Y given $Z=z$ and by M the set of such conditional densities. As before, we consider three distributions $\overline{R}$, $\overline{P}$ and $\overline{Q}$ for X_i and the corresponding conditional densities of Y_i given $Z_i=z_i$ with respect to μ, respectively $s(y_i|z_i)$, $t(y_i|z_i)$ and $u(y_i|z_i)$. Therefore $\overline{R}$ has density $s(y|z)f(z)$ with respect to the measure $\mu\otimes\nu$ on $F\times G$ with similar results for $\overline{P}$ and $\overline{Q}$. We then define on M the distance H by

$$H^2(t,u)=\int_G h_z^2(t,u)\,d\nu(z)\quad\text{with}\quad h_z^2(t,u)=\frac{1}{2}\int_F\bigl(\sqrt{t(y|z)}-\sqrt{u(y|z)}\bigr)^2\,d\mu(y).$$

Note that H is a distance between conditional densities which is different from the Hellinger distance between the joint densities of the pair (Y_i, Z_i) since

$$h^2(\overline{P}, \overline{Q}) = \frac{1}{2}\int_{F\times G} \bigl(\sqrt{t(y|z)} - \sqrt{u(y|z)}\bigr)^2 f(z)\, d\mu(y)\, d\nu(z).$$

We shall actually make the following assumption on the density f which implies that these two distances (H and h) are equivalent.

Assumption 2. *There exist two positive constants α and β such that*

$$0 < \alpha \le f(z) \le \beta \quad \textit{for all } z \in G$$

and α is known.

Working conditionally to the value z_i of Z_i, we derive from Theorem 2 with $\xi = 1/3$ a function $\phi(Y_i, Z_i)$ which satisfies

$$\mathbb{E}_{\overline{R}_i}\bigl[\phi_i(Y_i, Z_i)|Z_i = z_i\bigr] \le 1 - \frac{1}{3}h^2_{z_i}(t,u) + 2h^2_{z_i}(s,t)$$

and

$$\mathbb{E}_{\overline{R}_i}\left[\frac{1}{\phi_i(Y_i, Z_i)}|Z_i = z_i\right] \le 1 - \frac{1}{3}h^2_{z_i}(t,u) + 2h^2_{z_i}(s,u).$$

Integrating with respect to Z_i and using Assumption 2 leads to

$$\mathbb{E}_{\overline{R}_i}\bigl[\phi_i(Y_i, Z_i)\bigr] \le 1 - \frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,t) \le \exp\left[-\frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,t)\right]$$

and

$$\mathbb{E}_{\overline{R}_i}\left[\frac{1}{\phi_i(Y_i, Z_i)}\right] \le 1 - \frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,u) \le \exp\left[-\frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,u)\right].$$

Setting $\phi(\boldsymbol{X}) = \prod_{i=1}^n \phi_i(X_i)$ leads to analogues of (4.5) and (4.6), namely

$$\mathbb{P}_R\bigl[\log\bigl(\phi(\boldsymbol{X})\bigr) \ge x\bigr] \le \exp\left[-x - \frac{n\alpha}{3}H^2(t,u) + 2n\beta H^2(s,t)\right]$$

and

$$\mathbb{P}_R\bigl[\log\bigl(\phi(\boldsymbol{X})\bigr) \le x\bigr] \le \exp\left[x - \frac{n\alpha}{3}H^2(t,u) + 2n\beta H^2(s,u)\right].$$

We can then derive a test $\psi_{t,u}$ between t and u by setting $\psi_{t,u}(\boldsymbol{X}) = t$ if $\log(\phi(\boldsymbol{X})) < x$ and $\psi_{t,u}(\boldsymbol{X}) = u$ if $\log(\phi(\boldsymbol{X})) > x$ (the case of $\log(\phi(\boldsymbol{X})) = x$ being unimportant). It then follows from the previous inequalities that this test satisfies Assumption 1 with $B = 1$, $a = n\alpha/6$ and $\kappa = \sqrt{12\beta/\alpha}$. If β is not known, so is κ but, as we already mentioned, this knowledge is not necessary for the construction of T-estimators.

6. Markov chains

We present, in this section, an improved version of some results of [7].

Here, we consider a Markov chain $X_0, X_1, \ldots, X_n$ on Ξ with stationary transition kernel $P_s(x, A) = \mathbb{P}[X_{i+1} \in A \,|\, X_i = x]$ for $i \ge 0$ indexed by some function s from

$\Xi \times \Xi$ to $\mathbb{R}_+$. More precisely, we assume the existence of some probability π on Ξ such that $P_s(x,\cdot) \ll \pi(\cdot)$ for all $x \in \Xi$ with $dP_s(x,\cdot)/d\pi = s(x,\cdot)$. The distribution of X_0 is arbitrary and unknown. We denote by $\mathcal{F}_i$ the σ-algebra generated by $X_0, X_1, \ldots, X_i$, by P_s^i the iterated kernel $P_s^i(x,A) = \mathbb{P}[X_i \in A \,|\, X_0 = x]$ and set $s^i(x,\cdot) = dP_s^i(x,\cdot)/d\pi$, so that

$$s^{i+1}(x,y) = \int s^i(x,z)s(z,y)\,d\pi(z).$$

On the set $\mathcal{S}$ of possible transition densities s, we define the semi-metric H by

$$H^2(t,u) = \frac{1}{2}\int d\pi(x)\int \left(\sqrt{t(x,y)} - \sqrt{u(x,y)}\right)^2 d\pi(y) = \int h^2(t_x,u_x)\,d\pi(x),$$

with the obvious notation $t_x(y) = t(x,y)$. In the sequel we shall assume the following:

Assumption 3. *There exist integers $k \geq 1$ and $l \geq 0$ and positive numbers α and β such that*

$$\alpha \leq \frac{1}{k}\sum_{j=1}^{k} s^{l+j}(x,y) \leq \beta \quad \textit{for all } (x,y) \in \Xi \times \Xi,$$

where k, l and α are known, but not necessarily β.

By integration with respect to $s^i(z,\cdot)\pi(\cdot)$, we derive from Assumption 3 that

$$\alpha \leq \frac{1}{k}\sum_{j=1}^{k} s^{l+i+j}(x,y) \leq \beta \quad \text{for all } i \geq 0. \tag{6.1}$$

Note that the same argument shows that Assumption 3 holds if

$$s^{l+1}(x,y) \leq \beta \qquad \text{and} \qquad \frac{1}{k}\sum_{j=1}^{k} s^j(x,y) \geq \alpha \quad \text{for all } (x,y) \in \Xi \times \Xi.$$

For given $t, u \in \mathcal{S}$ and $x \in \Xi$, we may apply Theorem 2 with $\xi = 1/3$ to the probabilities P_0 and P_1 with respective densities t_x and u_x with respect to π. This results in a function $\phi(x,\cdot)$ which satisfies, for any $s \in \mathcal{S}$ and X with density s_x,

$$\mathbb{E}_{s_x}\left[\phi(x,X)\right] \leq 1 - \frac{1}{3}h^2(t_x,u_x) + 2h^2(s_x,t_x). \tag{6.2}$$

and

$$\mathbb{E}_{s_x}\left[\frac{1}{\phi(x,X)}\right] \leq 1 - \frac{1}{3}h^2(t_x,u_x) + 2h^2(s_x,u_x). \tag{6.3}$$

The idea, in order to use Assumption 3 efficiently, is to replace the original Markov chain by a subset of it. We first split the chain into blocks of size $m = l + k + 1$ (assuming that $n \geq m$). In each such piece of size m, we ignore the $l + 1$ first elements and draw one at random among the k remaining ones. We finally keep this element together with its predecessor in each block of size m. More formally, we denote the integer part of n/m by N and we define the random integers $J_i = mi + U_i - k$, $1 \leq i \leq N$, where the U_i are i.i.d. random integers drawn uniformly

among $\{1,\dots,k\}$ and independent of the Markov chain $\boldsymbol{X}=(X_0,X_1,\dots,X_n)$. We shall only use the subset $\{X_{J_i-1},X_{J_i},1\le i\le N\}$ of $\boldsymbol{X}$ to build our tests.

Given the real number z, we define the following test function $\psi_{t,u,z}(\boldsymbol{X})$ between t and u by

$$\psi_{t,u,z}(\boldsymbol{X})=\begin{cases} t & \text{if } \sum_{i=1}^{N}\log\bigl(\phi(X_{J_i-1},X_{J_i})\bigr)<z;\\ u & \text{if } \sum_{i=1}^{N}\log\bigl(\phi(X_{J_i-1},X_{J_i})\bigr)>z,\end{cases} \tag{6.4}$$

the choice in case of equality being unimportant. The performance of this test when $\boldsymbol{X}$ is driven by the transition density s is given by the following proposition.

Proposition 1. *Let $\boldsymbol{X}=(X_0,X_1,\dots,X_n)$ be a Markov chain on Ξ with transition density $s(x,\cdot)$ with respect to π that satisfies Assumption 3 with $m=l+k+1\le n$. Let N be the integer part of n/m, t,u two elements of $\mathcal{S}$, z a real number and $\psi_{t,u,z}$ the test function defined by (6.4). Then*

$$\mathbb{P}_s\bigl[\psi_{t,u,z}(\boldsymbol{X})=u\bigr]\le\exp\Bigl[-z-\frac{N\alpha}{3}H^2(t,u)+2N\beta H^2(s,t)\Bigr] \tag{6.5}$$

and

$$\mathbb{P}_s\bigl[\psi_{t,u,z}(\boldsymbol{X})=t\bigr]\le\exp\Bigl[z-\frac{N\alpha}{3}H^2(t,u)+2N\beta H^2(s,u)\Bigr]. \tag{6.6}$$

Proof. Let us set $n'=mN$ and $\phi_i(\boldsymbol{X})=k^{-1}\sum_{j=1}^{k}\phi(X_{mi+j-k-1},X_{mi+j-k})$ for $1\le i\le N$ so that $\phi_i(\boldsymbol{X})\in\mathcal{F}_{mi}$. The independence between the J_i and $\boldsymbol{X}$ and the fact that $m(N-1)=n'-m$ imply that

$$\begin{aligned}\mathbb{E}_s\Bigl[\prod_{i=1}^{N}\phi(X_{J_i-1},X_{J_i})\Bigr]&=\mathbb{E}_s\Bigl[\prod_{i=1}^{N}\phi_i(\boldsymbol{X})\Bigr]=\mathbb{E}_s\Bigl[\mathbb{E}_s\Bigl[\prod_{i=1}^{N}\phi_i(\boldsymbol{X})\,|\,\mathcal{F}_{n'-m}\Bigr]\Bigr]\\&=\mathbb{E}_s\Bigl[\mathbb{E}_s\Bigl[\prod_{i=1}^{N-1}\phi_i(\boldsymbol{X})\Bigr]\mathbb{E}_s\bigl[\phi_N(\boldsymbol{X})\,|\,X_{n'-m}\bigr]\Bigr].\end{aligned}$$

Then

$$\begin{aligned}E&=\mathbb{E}_s\bigl[\phi_N(\boldsymbol{X})\,|\,X_{n'-m}\bigr]=\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}_s\bigl[\phi(X_{n'-k+j-1},X_{n'-k+j})\,|\,X_{n'-m}\bigr]\\&=\frac{1}{k}\sum_{j=1}^{k}\mathbb{E}_s\bigl[\mathbb{E}_s\bigl[\phi(X_{n'-k+j-1},X_{n'-k+j})\,|\,X_{n'-k+j-1}\bigr]\,|\,X_{n'-m}\bigr].\end{aligned}$$

Since, by (6.2),

$$\begin{aligned}&\mathbb{E}_s\bigl[\phi(X_{n'-k+j-1},X_{n'-k+j})\,|\,X_{n'-k+j-1}\bigr]\\&\quad\le 1-\frac{1}{3}h^2(t_{X_{n'-k+j-1}},u_{X_{n'-k+j-1}})+2h^2(s_{X_{n'-k+j-1}},t_{X_{n'-k+j-1}}),\end{aligned}$$

we derive that

$$\begin{aligned}&\mathbb{E}_s\bigl[\mathbb{E}_s\bigl[\phi(X_{n'-k+j-1},X_{n'-k+j})\,|\,X_{n'-k+j-1}\bigr]\,|\,X_{n'-m}\bigr]\\&\quad\le 1-\frac{1}{3}\mathbb{E}_s\bigl[h^2(t_{X_{n'-k+j-1}},u_{X_{n'-k+j-1}})\,|\,X_{n'-m}\bigr]\\&\qquad+2\mathbb{E}_s\bigl[h^2(s_{X_{n'-k+j-1}},t_{X_{n'-k+j-1}})\,|\,X_{n'-m}\bigr]\\&\quad=1-\frac{1}{3}\int h^2(t_x,u_x)s^{l+j}_{X_{n'-m}}(x)\,d\pi(x)+2\int h^2(s_x,t_x)s^{l+j}_{X_{n'-m}}(x)\,d\pi(x)\end{aligned}$$

and finally,

$$E \le 1 - \frac{1}{3k}\sum_{j=1}^{k}\int h^2(t_x,u_x)s^{l+j}_{X_{n'-m}}(x)\,d\pi(x) + \frac{2}{k}\sum_{j=1}^{k}\int h^2(s_x,t_x)s^{l+j}_{X_{n'-m}}(x)\,d\pi(x).$$

It then follows from Assumption 3 that

$$\begin{aligned} E &\le 1 - \frac{\alpha}{3}\int h^2(t_x,u_x)\,d\pi(x) + 2\beta\int h^2(s_x,t_x)\,d\pi(x) \\ &\le \exp\left[-\frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,t)\right], \end{aligned}$$

so that by induction,

$$\begin{aligned} \mathbb{E}_s\left[\prod_{i=1}^{N}\phi_i(\boldsymbol{X})\right] &\le \mathbb{E}_s\left[\prod_{i=1}^{N-1}\phi_i(\boldsymbol{X})\right]\exp\left[-\frac{\alpha}{3}H^2(t,u) + 2\beta H^2(s,t)\right] \\ &\le \exp\left[-\frac{N\alpha}{3}H^2(t,u) + 2N\beta H^2(s,t)\right]. \end{aligned}$$

Applying Lemma 1 leads to (6.5). The proof of (6.6) derives in the same way from (6.3). □

Proposition 1 implies that the test we have built satisfies Assumption 1 with $B = 1$, $a = N\alpha/6$ and $\kappa = \sqrt{12\beta/\alpha}$, which allows to apply Theorem 1 to the estimation of transition densities of Markov chains provided that Assumption 3 holds. This is clearly a serious restriction but we have been unable to weaken it.

7. Conclusion

It follows from the previous sections that the testing results which are valid for density estimation (with n i.i.d. observations) can be extended to the estimation of conditional densities and Markov transitions. In both cases, we have to estimate a function of two variables (not necessarily real variables), $s(y|z)$ or $s(x,y)$, with a loss function which is an adequate version of the Hellinger distance on the parameter space. The only additional assumption we need, as compared to density estimation, is the existence of upper and lower bounds on the design density (for conditional densities) or the parameter itself (for Markov transitions). This is of course a strong assumption since the value of the lower bound has to be known from the statistician. Nevertheless, under this assumption, all known results for density estimation can be translated to estimation of conditional densities and Markov transitions. For instance, using similar families of models, we get the same rates of convergences for estimating the density or the transition $s(x,y)$. We shall not insist on this point, just mentioning that many examples of applications of T-estimators which have been considered in [12, 13] and [4] can be extended to the cases of conditional densities and Markov transitions. Examples could be the estimation of Hölderian Markov transitions with anisotropic smoothness or the problems considered in [1] about conditional density estimation in the i.i.d. case. Note that Assumption (B) of [1] also requires a known lower bound (denoted $\iota(f)$ in the paper) for the design density. Many more estimation problems for conditional densities could be dealt with, at least in an abstract way, with the use of Theorem 2.

Acknowledgements

I would like to thank Yannick Baraud for many helpful discussions and wise suggestions about this paper and the correction of some mistakes.

References

[1] Akakpo, N. and Lacour, C. (2011). Inhomogeneous and anisotropic conditional density estimation from dependent data. *Electron. J. Statist.* **5** 1618–1653.
[2] Balabdaoui, F., Rufibach, K. and Wellner, J. (2009). Convergence of random processes and limit theorems in probability theory. *Ann. Statist.* **37** 1299–1331.
[3] Baraud, Y. (2011). Estimator selection with respect to Hellinger-type risks. *Probab. Theory Relat. Fields* **151** 353–401.
[4] Baraud, Y. and Birgé, L. (2011). Estimating composite functions by model selection. Available at arXiv:1102.2818.
[5] Bednarski, T. (1982). Binary experiments, minimax tests and 2-alternating capacities. *Ann. Statist.* **10** 226–232.
[6] Birgé, L. (1983a). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **65** 181–237.
[7] Birgé, L. (1983b). Robust testing for independant non identically distributed variables and Markov Chains. In *Specifying Statistical Models* 134–162. Springer-Verlag, Heidelberg.
[8] Birgé, L. (1984a). Sur un théorème de minimax et son application aux tests. *Probab. Math. Statist.* **3** 259–282.
[9] Birgé, L. (1984b). Stabilité et instabilité du risque minimax pour des variables indépendantes équidistribuées. *Ann. Inst. H. Poincaré Sect. B* **20** 201–223.
[10] Birgé, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Theory Relat. Fields* **71** 271–291.
[11] Birgé, L. (2004) Model selection for Gaussian regression with random design. *B*ernoulli **10** 1039–1051.
[12] Birgé, L. (2006). Model selection via testing: An alternative to (penalized) maximum likelihood estimators. *Ann. Inst. Henri Poincaré, Probab. et Statist.* **42** 273–325.
[13] Birgé, L. (2007). Model selection for Poisson processes. In *Asymptotics: Particles, Processes and Inverse Problems, Festschrift for Piet Groeneboom* (E. Cator, G. Jongbloed, C. Kraaikamp, R. Lopuhaä and J. Wellner, eds.). IMS Lecture Notes – Monograph Series **55** 32–64.
[14] Ghosal, S., Gosh, J. K. and van der Vaart, A. W. (2000). Convergence rates of posterior distributions. *Ann. Statist.* **28** 500–531.
[15] Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for non-i.i.d. observations. *Ann. Statist.* **35** 192–223.
[16] Groeneboom, P., Maathuis, M. and Wellner, J. (2008a). Current status data with competing risks: consistency and rates of convergence of the MLE. *Ann. Statist.* **36** 1031–1063.
[17] Groeneboom, P., Maathuis, M. and Wellner, J. (2008b). Current status data with competing risks: limiting distribution of the MLE. *Ann. Statist.* **36** 1064–1089.
[18] Groeneboom, P. and Wellner, J. (1992). *Information Bounds and Non-*

parametric Maximum Likelihood Estimation. DMV Seminar 19. Birkhauser Verlag, Basel.

[19] Huang, J. and Wellner, J. (1995). Asymptotic normality of the NPMLE of the mean for interval censored data, case I. *Statistica Neerlandica* **49** 153–163.

[20] Huber, P. J. (1965). A robust version of the probability ratio test. *Ann. Math. Statist.* **36** 1753–1758.

[21] Huber, P. J. (1981). *Robust Statistics*. John Wiley, New York.

[22] Huber, P. J. and Strassen, V. (1973). Minimax tests and the Neyman-Pearson lemma for capacities. *Ann. Statist.* **1** 251–263 and **2** 223–224.

[23] Ibragimov, I. A. and Has'minskii, R. Z. (1978). On the capacity of transmission by means of smooth signals. *Dokl. Akad. Nauk SSSR* **242** 32–35.

[24] Ibragimov, I. A. and Has'minskii, R. Z. (1980). On estimate of the density function. *Zap. Nauchn. Semin. LOMI* **98** 61–85.

[25] Ibragimov, I. A. and Has'minskii, R. Z. (1981a). On the non-parametric density estimates. *Zap. Nauchn. Semin. LOMI* **108** 73–89.

[26] Ibragimov, I. A. and Has'minskii, R. Z. (1981b). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.

[27] Le Cam, L. M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** 38–53.

[28] Le Cam, L. M. (1975). On local and global properties in the theory of asymptotic normality of experiments. *Stochastic Processes and Related Topics*, Vol. 1 (M. Puri, ed.) 13–54. Academic Press, New York.

[29] Le Cam, L. M. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.

[30] Le Cam, L. M. (1990). Maximum likelihood: an introduction. *Inter. Statist. Review* **58** 153–171.

[31] Le Cam, L. M. (1997). Metric dimension and statistical estimation. *CRM Proc. and Lecture Notes* **11** 303–311.

[32] Maathuis, M. and Wellner, J. (2008). Inconsistency of the MLE for the joint distribution of interval-censored survival times and continuous marks. *Scand. J. Statist.* **35** 83–103.

[33] Massart, P. (2007). Concentration Inequalities and Model Selection. In *Lecture on Probability Theory and Statistics, École d'Été de Probabilités de Saint-Flour XXXIII - 2003* (J. Picard, ed.). Lecture Note in Mathematics, Springer-Verlag, Berlin.

[34] Rieder, H. (1977). Least favourable pairs for special capacities. *Ann. Statist.* **5** 909–921.