ABOUT THE DISCRETIZED MAXIMUM LIKELIHOOD ESTIMATOR

Lucien Birgé

About the discretized Maximum Likelihood Estimator

Lucien Birgé

Sorbonne Université - LPSM, CNRS UMR 8001
Campus Pierre et Marie Curie - case courrier 158
75252 Paris Cedex 05, France

October 9, 2018

Abstract

The most well-known and used statistical estimation procedure is probably the Maximum Likelihood method. Unfortunately, apart from rather elementary situations, the analysis of its performance requires complex probablistic tools, namely Empirical Process Theory. Examples can be found, for instance, in the books by Ibragimov and Has'minskii (1981), van der Vaart (1998), van de Geer (2000) or Massart (2007). Our purpose here is to describe a discretized version of the method, which is not so well-known, but with the advantage that one can derive its non-asymptotic performance from elementary tools. We shall explain the method, give its performances and provide several illustrative examples.

1 Introduction

To begin with, let us recall what type of problems statisticians want to study. They work within a classical probabilistic framework where one has at hand a random variable $\boldsymbol{X}$ from some abstract measurable space $(\Omega, \Xi, \mathbb{P})$ to an observational space $(\boldsymbol{E}, \mathscr{B})$ with distribution $\mathbf{P}_{\boldsymbol{X}}$ on $\boldsymbol{E}$. While probabilists want to analyze the behaviour of $\boldsymbol{X}$ from the knowledge of $\mathbf{P}_{\boldsymbol{X}}$, statisticians work in the opposite direction. They observe a realization $\boldsymbol{X}(\omega)$ of $\boldsymbol{X}$ and try to infer from it some properties of the *unknown* distribution $\mathbf{P}_{\boldsymbol{X}}$.

Obviously, the number of problems which are connected to the search of information about the probabilistic distribution of $\boldsymbol{X}$ is unlimited but we shall concentrate here on a specific one, called *estimation* and more precisely on the estimation of the true unknown distribution $\mathbf{P}_{\boldsymbol{X}}$ of $\boldsymbol{X}$. This means finding a

random variable $\widehat{\mathbf{P}}(\boldsymbol{X})$, with values in the set of all probabilities on $\boldsymbol{E}$ and such that $\widehat{\mathbf{P}}$ is close (in a suitable sense) to $\mathbf{P}_{\boldsymbol{X}}$ with probability close to 1. Unfortunately, without some prior information on $\mathbf{P}_{\boldsymbol{X}}$, it is (provably) impossible to achieve this aim. In view of solving the estimation problem, we have to make assumptions on the structure of $\mathbf{P}_{\boldsymbol{X}}$ and, to keep this presentation simple, we shall focus here on the following classical framework: $\boldsymbol{E} = E^n$, $\boldsymbol{X} = (X_1, \ldots, X_n)$ with $X_i \in E$ for $1 \leqslant i \leqslant n$ and the X_i are i.i.d. with unknown distribution P_X on E so that $\mathbf{P}_{\boldsymbol{X}} = P_X^{\otimes n}$. This is the *i.i.d. framework*. In the sequel we shall denote by $P^\star$ $(= P_X)$ the true common distribution of the X_i. We simply write $\mathbb{P}$ and $\mathbb{E}$ for probabilities and expectations of quantities depending on $\boldsymbol{X}$ under the assumption that the X_i are i.i.d. with distribution $P^\star$ on E.

2 Models and the Maximum Likelihood Estimator

2.1 Statistical models

Traditionnally, estimation of $P^\star$ has been (and still is) based on *models*. A statistical model for $P^\star$ is a given subset $\mathscr{M}$ of the set $\mathscr{P}$ of all probabilities on E. The simplest ones, namely *parametric models*, are those indexed by a subset Θ of some Euclidean space $\mathbb{R}^k$ in such a way that the application $\theta \mapsto P_\theta$ is one-to-one. Such a model will be denoted by $\mathscr{M}_\Theta = \{P_\theta,\, \theta \in \Theta\}$. Some examples are the set of Poisson distributions on $\mathbb{N}$ with unknown parameter $\theta > 0$, the normal distributions $\mathcal{N}(m, \sigma^2)$ on $\mathbb{R}$ with $\theta = (m, \sigma^2) \in \Theta \subset \mathbb{R} \times (0, +\infty)$, a translation model for which $\Theta = \mathbb{R}$ and $(dP_\theta/d\mu)(x) = p_\theta(x) = p(x - \theta)$ for some given density p with respect to the Lebesgue measure μ and the set of uniform distributions on $[0, \theta]$ with $\Theta = (0, +\infty)$. More complex models were used later. An example is that of *density estimation* with $\mathscr{M} = \{P = p \cdot \mu,\, p \in \mathcal{M}\}$ where $\mathcal{M}$ is some set of densities with respect to the Lebesgue measure μ on $\mathbb{R}^k$. One could for instance choose for $\mathcal{M}$ the Lipschitz densities on $[0, 1]^k$ or the non-increasing densities on $[0, 1]$. Since $P^\star$ is unknown, we shall have to work with different probabilities in $\mathscr{P}$ and use the notation $\mathbb{P}_Q[\boldsymbol{X} \in A]$ to indicate that the X_i have the common distribution Q, writing $\mathbb{P}_p$ as shorthand for $\mathbb{P}_{p \cdot \mu}$ and similarly $\mathbb{E}_Q$ and $\mathbb{E}_p$.

In this paper we always assume that the model $\mathscr{M}$ is dominated by some reference measure μ, which means that any $P \in \mathscr{M}$ can be written as $P = p \cdot \mu$ with density p with respect to μ, and $\mathscr{M} = \{P = p \cdot \mu,\, p \in \mathcal{M}\}$ for some set $\mathcal{M}$ of densities with respect to μ which we shall call a *density model*. The true distribution $P^\star$ of the X_i belongs to $\mathscr{M}$ so that $P^\star = P_{\theta^\star} = p_{\theta^\star} \cdot \mu$ with $\theta^\star \in \Theta$ in the parametric case and $P^\star = p^\star \cdot \mu$ with $p^\star \in \mathcal{M}$ in the general case.

2.2 The Maximum Likelihood Estimator

When the model is parametric a very popular estimator, going back to the work of Sir Ronald Fisher in the 1920's is the Maximum Likelihood Estimator (MLE for short). To introduce the method, let us assume that E is a finite or countable set, μ the counting measure on E, in which case $P_\theta = p_\theta \cdot \mu$ with $P_\theta[\{x\}] = p_\theta(x)$. Then

$$\frac{dP_\theta^{\otimes n}}{d\mu^{\otimes n}}(x_1, \ldots, x_n) = \prod_{i=1}^n p_\theta(x_i) = P_\theta^{\otimes n}[\{x_1, \ldots, x_n\}] = \mathbb{P}_{p_\theta}[X_i = x_i \text{ for } 1 \leqslant i \leqslant n].$$

As a consequence, the so-called *likelihood* of θ: $\prod_{i=1}^n p_\theta(X_i(\omega))$ can be seen as the probability, when θ is the true parameter, of the event that actually occured, namely $\{X_1(\omega), \ldots, X_n(\omega)\}$. In the Poisson model that we mentioned above, if $X_i(\omega) = k_i \in \mathbb{N}$ for $1 \leqslant i \leqslant n$, the likelihood writes

$$\prod_{i=1}^n p_\theta(k_i) = \prod_{i=1}^n e^{-\theta} \frac{\theta^{k_i}}{k_i!} = \exp\left[-n\theta + \left(\sum_{i=1}^n k_i\right) \log\theta\right] \left(\prod_{i=1}^n (k_i!)\right)^{-1}.$$

This quantity depends on θ and a natural idea is to believe that the true value $\theta^\star$ of the parameter is close to the value $\widehat{\theta}_n$ of θ that maximizes the probability of the event $\{k_1, \ldots, k_n\} = \{X_1(\omega), \ldots, X_n(\omega)\}$ that actually occured, i.e. the parameter $\widehat{\theta}_n$ that maximizes the likelihood function: $\theta \mapsto \prod_{i=1}^n p_\theta(X_i(\omega))$. In the Poisson case, one immediately derives that

$$\widehat{\theta}_n = \widehat{\theta}_n(\boldsymbol{X}) = \frac{1}{n} \sum_{i=1}^n X_i \quad \text{if} \quad \sum_{i=1}^n X_i > 0,$$

otherwise the likelihood reduces to $e^{-n\theta}$ and admits no maximum on Θ. Note that such an event occurs with a probability

$$\mathbb{P}\left[\sum_{i=1}^n X_i = 0\right] = \mathbb{P}\left[X_i = 0 \text{ for all } i\right] = \prod_{i=1}^n \mathbb{P}\left[X_i = 0\right] = \exp[-n\theta^\star],$$

which converges to 0 when n goes to infinity, although it is definitely not negligible if $n\theta^\star$ is small.

2.3 About the classical MLE

Since the work of Fisher the MLE has become an extremely popular estimator which has been widely used in all types of situations (not only parametric models and the i.i.d. framework) and has been the subject of thousands of papers. In particular, it has been proven that, under suitable assumptions, it has a nice behaviour and some optimality properties.

Example 1. $E = \mathbb{R} = \Theta$ and the X_i are normal $\mathcal{N}(\theta^\star, 1)$ in which case the MLE is $\widehat{\theta}_n = n^{-1}\sum_{i=1}^n X_i$ so that $\sqrt{n}\left(\theta^\star - \widehat{\theta}_n\right)$ has distribution $\mathcal{N}(0,1)$ and the variance 1 can be shown to be optimal in a suitable sense.

Example 2. When $E = (0, +\infty)$ and the X_i are uniform on $[0, \theta]$ the MLE is the largest observation $\widehat{\theta}_n = \sup_{1\leqslant i\leqslant n} X_i < \theta$ a.s. and one can easily show that

$$\mathbb{E}_{p_\theta}\left[\left(\widehat{\theta}_n - \theta\right)^2\right] = \frac{2\theta^2}{(n+1)(n+2)}.$$

Unfortunately, the MLE sometimes does not exist as we already noticed for the case of the Poisson distribution and it is indeed impossible to derive a general theory for the MLE without strong enough assumptions, in particular boundedness of the densities in the case of a translation model.

Example 3. $E = \mathbb{R} = \Theta$ and the X_i have a density $p_\theta(x) = p(x-\theta)$ with respect to the Lebesgue measure μ. One can easily see that the MLE does not exist on the model $\mathscr{M}_{\mathbb{R}} = \{p_\theta \cdot \mu,\, \theta \in \mathbb{R}\}$ if the density p is unbounded.

Studying the MLE in a general situation not only involves strong assumptions but also sophisticated results about the suprema of empirical processes as can be seen from the books by van der Vaart (1998), van de Geer (2000) or Massart (2007). Even in the simple situation in which Θ is a compact subset of $\mathbb{R}^k$ and the mapping $\theta \to P_\theta$ has some smoothness properties, analyzing the asymptotic properties of the MLE when n goes to infinity is a difficult task as seen in papers by Le Cam (1970), Hajek (1972) and the books by Ibragimov and Has'minskii (1981) or van der Vaart (1998).

The main purpose of this paper is the presentation of a discrete version of the MLE which is not more powerful than the classical one (although it can be, in some exceptional situations) but has the advantage of being much easier to analyze. The following presentation is short and does not require the very complex tools developed, for instance, in Massart (2007) to analyze the behaviour of the classical MLE. We shall illustrate the general theorems by elementary examples in order to explain their use. We therefore hope that this unusual presentation of the MLE will be accessible to mathematicians who know nothing about empirical process theory.

3 The MLE on a finite model

On the one hand there are many examples of simple statistical models for which the MLE behaves well, at least asymptotically. On the other hand one can also find many examples of a poor behaviour of the MLE in Le Cam (1990), Birgé (2006) or Baraud, Birgé and Sart (2017), among others.

There exists nevertheless a very simple situation for which the study of the asymptotic behaviour of the MLE can be done via elementary tools and we shall now describe it.

3.1 Convergence of the MLE

Let us start by a simple, but fundamental lemma.

Lemma 1. *Given two probabilities $P = p \cdot \mu$ and $Q = q \cdot \mu$ dominated by μ, n i.i.d. observations $X_1, \dots, X_n$ with distribution P and $y \in \mathbb{R}$, the following bound holds:*

$$\mathbb{P}_P\left[\sum_{i=1}^n \log\left(\frac{q(X_i)}{p(X_i)}\right) \geqslant y\right] \leqslant \exp\left[-\frac{y}{2}\right]\left(\int \sqrt{p(x)q(x)}\, d\mu(x)\right)^n. \tag{1}$$

Proof. Applying the exponential inequality $\mathbb{P}[Y \geqslant 0] = \mathbb{E}[\mathbb{1}_{\mathbb{R}_+}(Y)] \leqslant \mathbb{E}[\exp[Y/2]]$, we derive that

$$\begin{aligned}\mathbb{P}_P\left[\sum_{i=1}^n \log\left(\frac{q(X_i)}{p(X_i)}\right) - y \geqslant 0\right] &\leqslant \mathbb{E}_P\left[\exp\left[\frac{1}{2}\left(-y + \sum_{i=1}^n \log\left(\frac{q(X_i)}{p(X_i)}\right)\right)\right]\right] \\ &= \exp\left[-\frac{y}{2}\right]\left(\mathbb{E}_P\left[\sqrt{\frac{q(X_1)}{p(X_1)}}\right]\right)^n \end{aligned} \tag{2}$$

and the conclusion follows since $\mathbb{E}_P\left[\sqrt{q(X_1)/p(X_1)}\right] = \int \sqrt{p(x)q(x)}\, d\mu(x)$. □

This bound introduces the so-called *Hellinger affinity* between P and Q.

Definition 1. *Given two probabilities P and Q on the same probability space and any measure μ such that $P \ll \mu$ and $Q \ll \mu$, the Hellinger affinity between P and Q is given by*

$$\rho(P, Q) = \int \sqrt{\frac{dP}{d\mu}\frac{dQ}{d\mu}}\, d\mu,$$

a quantity which is independent of μ.

It follows that (1) can be written as

$$\mathbb{P}_P\left[\sum_{i=1}^n \log\left(\frac{q(X_i)}{p(X_i)}\right) \geqslant y\right] \leqslant e^{-y/2}\rho^n(P, Q) = \exp\left[-\frac{y}{2} + n\log\big(\rho(P, Q)\big)\right]. \tag{3}$$

One can check (simply using Cauchy-Schwarz inequality) that $0 \leqslant \rho(P, Q) \leqslant 1$, $\rho(P, Q) = 1$ if and only if $P = Q$ and $\rho(P, Q) = 0$ if and only if there exists a set A such that $P(A) = 1 = Q(A^c)$.

Now assume that $\mathscr{M}$ is finite and $\mathcal{M} = \{p_1, \ldots, p_N\}$. Taking into account the possibility of ties (the MLE necessarily exists in this case but need not be unique), we derive from Lemma 1 that

$$\begin{aligned}\mathbb{P}\left[\widehat{p}_n \neq p^\star\right] &\leqslant \sum_{p\in\mathcal{M},\, p\neq p^\star} \mathbb{P}\left[\prod_{i=1}^n p(X_i) \geqslant \prod_{i=1}^n p^\star(X_i)\right] \\ &= \sum_{p\in\mathcal{M},\, p\neq p^\star} \mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{p(X_i)}{p^\star(X_i)}\right) \geqslant 0\right] \\ &\leqslant \sum_{p\in\mathcal{M},\, p\neq p^\star} \left(\int \sqrt{p(x)p^\star(x)}\, d\mu(x)\right)^n = \sum_{P\in\mathscr{M},\, P\neq P^\star} \rho^n(P, P^\star). \qquad (4)\end{aligned}$$

Since $P \neq P^\star$, hence $\rho(P, P^\star) < 1$, and $\mathscr{M}$ is finite, $\mathbb{P}\left[\widehat{p}_n \neq p^\star\right] \underset{n\to+\infty}{\longrightarrow} 0$ so that asymptotically the MLE will always recover the true density.

3.2 The Hellinger distance

Another important feature of the Hellinger affinity ρ, apart from (3), is its relation to a distance on the set $\mathscr{P}$ of all probabilities on E, namely the *Hellinger distance* h, the importance of which has been first emphasized by Le Cam. It is defined in the following way (independently of μ) with $p = dP/d\mu$ and $q = dQ/d\mu$:

$$h^2(P, Q) \overset{def}{=} \frac{1}{2}\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mu = 1 - \rho(P, Q) \quad \text{or} \quad h(P, Q) = \sqrt{1 - \rho(P, Q)}$$

and, since $\log\big(\rho(P, Q)\big) \leqslant \rho(P, Q) - 1 = -h^2(P, Q)$, (3) implies that

$$\mathbb{P}_P\left[\sum_{i=1}^n \log\left(\frac{q(X_i)}{p(X_i)}\right) \geqslant y\right] \leqslant \exp\left[-\frac{y}{2} - nh^2(P, Q)\right] \quad \text{for all } y \in \mathbb{R}. \qquad (5)$$

Note that h is a genuine distance since it can be viewed as an $\mathbb{L}_2$-type distance on the set of square roots of the non-negative elements of $\mathbb{L}_1(\mu)$. In the sequel and for simplicity, we shall often write $h(p, q)$ for $h(P, Q)$, considering also $\mathcal{M}$ as a metric space with distance h.

Since $\mathcal{M}$ is finite, $\delta = \inf_{p\neq p^\star} h^2(p, p^\star) > 0$. If we set $N_j = \{p \in \mathcal{M} \text{ such that } j\delta \leqslant h^2(p^\star, p) < (j+1)\delta\}$ and it follows from (4) and (5) that

$$\begin{aligned}\mathbb{P}\left[\widehat{p}_n \neq p^\star\right] &\leqslant \sum_{p\in\mathcal{M},\, p\neq p^\star} \exp\left[-nh^2(p, p^\star)\right] \leqslant \sum_{j=1}^{+\infty} N_j e^{-nj\delta} \\ &\leqslant e^{-n\delta}\left[N_1 + \sum_{j\geqslant 1} N_{j+1} e^{-jn\delta}\right], \qquad (6)\end{aligned}$$

which provides a nonasymptotic bound for the probability of error of the MLE depending on the metric structure of $\mathcal{M}$ with respect to the Hellinger distance

and, more specifically, on the number of points of $\mathcal{M}$ that belong to balls centered at $p^\star$. For n large enough, the series converges quickly and is essentially equivalent to its first term $N_1 e^{-n\delta}$.

The purpose of the next section will be to generalize the previous results to more realistic (infinite) models $\mathscr{M}$.

4 The MLE on a discrete approximation of the model

4.1 Discretizing a model

In order to avoid the difficulties connected with the study on the MLE on a general model, we shall adopt the strategy consisting in replacing, in the metric space $(\mathscr{P}, h)$, the set $\mathscr{M} = \{p \cdot \mu,\, p \in \mathcal{M}\}$ by a finite approximation. Let us now make this precise.

Definition 2. *Let $(\mathcal{S}, \Delta)$ be a metric space and $\delta > 0$. A subset S_δ of $\mathcal{S}$ is a δ-net for $S \subset \mathcal{S}$ if, for all $s \in S$, one can find $t \in S_\delta$ such that $\Delta(s,t) \leqslant \delta$. A subset S'_δ of $\mathcal{S}$ is δ-separated if $\Delta(s,t) > \delta$ for all s, t, $s \neq t$ in S'_δ. It is a maximal δ-separated subset of S if it is δ-separated and if $S'_\delta \cup \{t\}$ is not δ-separated for all $t \in S \setminus S'_\delta$.*

Note that any maximal δ-separated subset of S is a δ-net for S and if S is compact, one can always find a finite δ-net S_δ for S, taking for S_δ the set of centers of a finite covering of S by balls with radius δ. Also observe that, to any finite δ-net $S_\delta = \{s_1, \ldots, s_N\}$ for S, one can associate a partition $(S_1, \ldots, S_N)$ of S and a mapping π from S to S_δ in such a way that $\sup_{s \in S_j} \Delta(s, s_j) \leqslant \delta$ and $\pi^{-1}(s_j) = S_j$ for each $j \in \{1, \ldots, N\}$.

In the sequel, we shall assume that one can find a finite η-net $\mathscr{M}_\eta = \{p \cdot \mu,\, p \in \mathcal{M}_\eta\}$ for $\mathscr{M}$, not necessarily a subset of $\mathscr{M}$, to approximate $\mathscr{M}$. In terms of the corresponding density models, this means that if $\mathcal{M}_\eta = \{p_1, \ldots, p_N\}$ we can find a partition $(\mathcal{M}'_1, \ldots, \mathcal{M}'_N)$ of $\mathcal{M}$ such that, for each $j \in \{1, \ldots, N\}$, $\sup_{p \in \mathcal{M}'_j} h(p, p_j) \leqslant \eta$. Defining the application π_η from $\mathcal{M}$ to $\mathcal{M}_\eta$ by $\pi_\eta(p) = p_j$ if $p \in \mathcal{M}'_j$, we get $h(p, \pi_\eta(p)) \leqslant \eta$ for all $p \in \mathcal{M}$. We finally define our estimator $\widehat{p}_n$ of $p^\star \in \mathcal{M}$ as a MLE on $\mathcal{M}_\eta$ which means that $\widehat{p}_n$ is any point in $\mathcal{M}_\eta$ which maximizes the function

$$p \to \sum_{i=1}^{n} \log\big(p(X_i)\big) \quad \text{from } \mathcal{M}_\eta \text{ to } [-\infty, +\infty) \quad \text{with } \log(0) = -\infty.$$

The case that we studied in the previous section corresponds to the case of a finite set $\mathcal{M}$, $\mathcal{M}_\eta = \mathcal{M}$ and π_η the identity function.

To mimic the proof we gave in Section 3 which was based on the inequality (5) we need to derive an analogous bound for

$$\mathbb{P}\left[\sum_{i=1}^{n}\log\left(\frac{p(X_i)}{\pi_\eta(p^\star)(X_i)}\right)\geqslant 0\right]\quad\text{when } p\in\mathcal{M}_\eta, \tag{7}$$

the difference being that we still compute the probability with respect to $P^\star$ while now the denominators involves $\pi_\eta(p^\star)$. Since, under $P^\star$, one may have $\pi_\eta(p^\star)(X_i)=0$ (which is a.s. impossible for $p^\star(X_i)$), we set $\log(a/0)=+\infty$ if $a>0$ and $\log(0/0)=0$ in (7). Unfortunately, the fact that $h(\theta^\star,\pi(\theta^\star))\leqslant\eta$ does not warrant, even if η is small, that an analogue of (5) holds as shown by the following counterexample.

Example 4. Let P_θ be the uniform distributions on $[0,\theta]$, $\theta>0$, with density $p_\theta=\theta^{-1}\mathbb{1}_{[0,\theta]}$ with respect to the Lebesgue mesure μ. Let the true distribution $P^\star$ have density $p_{\theta^\star}=(1-n^{-1})\,p_1+n^{-1}\mathbb{1}_{[100,101]}$. Then $\rho(P_1,P^\star)=\sqrt{1-n^{-1}}$ and $h^2(P_1,P^\star)$ is approximately $(2n)^{-1}$ when n is large so that P_1 provides a very good approximation of $P^\star$ for large n. With n i.i.d. observations of distribution $P^\star$, the probability that at least one of them belongs to $[100,101]$ is $1-(1-n^{-1})^n>1-e^{-1}$. It immediately follows that

$$\mathbb{P}\left[\sum_{i=1}^{n}\log\left(\frac{p_{101}(X_i)}{p_1(X_i)}\right)=+\infty\right]>1-e^{-1}$$

although $h^2(P_1,P_{101})=1-101^{-1/2}>9/10$.

The problem, in the previous example, is connected to the fact that $P^\star\not\ll\pi_\eta(p^\star)\cdot\mu$ but, with some additional effort, one can build a more sophisticated counterexample for which $P^\star\ll\pi_\eta(p^\star)\cdot\mu$. Therefore, even with this additional assumption of domination and if η is very small, the fact that $h(p^\star,\pi_\eta(p^\star))\leqslant\eta$ does not warrant that an analogue of (5) holds.

4.2 Deviation bounds for the discretized MLE

As we have just seen, getting for (7) an exponential bound similar to (5) requires that both $\mathcal{M}_\eta$ and the mapping π_η be chosen in a specific way. Such a bound will actually result from an application of the following Proposition, to be proven in Section 5.

Proposition 1. *Let $P_s=s\cdot\mu$ and $P_t=t\cdot\mu$, $P_s\ll P_t$, be two probabilities dominated by μ and such that $\int_{t>0}s^2t^{-1}\,d\mu<+\infty$. If*

$$\int_{t>0}\left(\frac{s^2}{t}\right)d\mu-1=\int_{t>0}\left(\frac{s}{t}-1\right)^2t\,d\mu\leqslant[Ah(P_s,P_t)]^2\quad\textit{with } A>0, \tag{8}$$

then $A > \sqrt{2}$ and, whatever the probability $P_u = u \cdot \mu$ and $y \in \mathbb{R}$,

$$\mathbb{P}_s\left[\sum_{i=1}^{n} \log\left(\frac{u(X_i)}{t(X_i)}\right) \geqslant y\right] \leqslant \exp\left[-nh^2(t,u)\left(1 - A\sqrt{2}\,\frac{h(P_s,P_t)}{h(P_t,P_u)}\right) - \frac{y}{2}\right]. \quad (9)$$

In particular, if $s(x) \leqslant \Delta t(x)$ for for some $\Delta > 1$ and μ-almost all x, one can take $A = \sqrt{2}\big(1 + \sqrt{\Delta}\big)$.

Remark. *If $P = p \cdot \mu$ and $Q = q \cdot \mu$ with $P \ll Q$, the quantity*

$$\chi^2(P,Q) = \int_{q>0} \left(\frac{p}{q} - 1\right)^2 d\mu - 1 = \int_{q>0} \frac{p^2}{q}\, d\mu - 1,$$

*which is independent of the dominating measure μ, is called the χ^2-*divergence *between P and Q.*

The proposition says that (8) allows to replace the analogue of (3), namely

$$\mathbb{P}_t\left[\sum_{i=1}^{n} \log\left(\frac{u(X_i)}{t(X_i)}\right) \geqslant y\right] \leqslant \exp\left[-anh^2(t,u) - \frac{y}{2}\right] \quad \text{with } a = 1$$

by (9), which is similar provided that $a = 1 - A\sqrt{2}h(P_s,P_t)/h(P_t,P_u) > 0$.

In order to control the performance of the MLE on $\mathcal{M}_\eta$ when $p^\star \in \mathcal{M}$ we shall apply the proposition to pairs (t,s) of the form $(p_j, p^\star)$ with $p^\star \in \mathcal{M}'_j$ and, since $p^\star$ is unknown, assume that (8) holds for all pairs (t,s) of the form (p_j,p) with $p \in \mathcal{M}'_j$ and $j \in \{1,\dots,N\}$. Moreover, to get precise deviation bounds between $p^\star$ and $\widehat{p}_n$, we shall also need, as for (6), a control of the number of points of $\mathcal{M}_\eta$ that belong to balls of a given radius. This leads to the following set of assumptions.

Assumption 1. *The set $\mathcal{M}_\eta$ and the application π_η satisfy the following properties: there exist numbers A, b, D and a with*

$$A > \sqrt{2},\ \ b \geqslant \big(A\sqrt{2}\big) \vee 4,\ \ D \geqslant 1/2,\ \ a = 1 - Ab^{-1}\sqrt{2}\ \text{ and }\ an\eta^2 \geqslant 2D \quad (10)$$

such that

i) For each $p \in \mathcal{M}$, $h(p, \pi_\eta(p)) \leqslant \eta$.

ii) For all $j \in \{1,\dots,N\}$ and $\mathcal{M}'_j = \pi_\eta^{-1}(p_j)$,

$$\sup_{p \in \mathcal{M}'_j} \int_{p_j>0} \left(\frac{p}{p_j}\right)^2 d\mu - 1 = \sup_{p \in \mathcal{M}'_j} \int_{p_j>0} \left(\frac{p}{p_j} - 1\right)^2 p_j\, d\mu \leqslant (A\eta)^2. \quad (11)$$

iii) For all $p \in \mathcal{M}_\eta$,

$$\big|\{q \in \mathcal{M}_\eta \,|\, h(p,q) < x\eta\}\big| \leqslant \exp\big[x^2 D\big] \quad \textit{for all } x \geqslant b. \quad (12)$$

Note that it is actually enough to check (12) for $b \leqslant x \leqslant \eta^{-1}$ and that $|\mathcal{M}_\eta| = N \leqslant \exp[\eta^{-2}D]$ since, for $x > \eta^{-1}$, $\{p \in \mathcal{M}_\eta \,|\, h(p,p_0) < x\eta\} = \mathcal{M}_\eta$, the Hellinger distance being bounded by one.

In the next theorem and from now on, C will denote an arbitrary universal constant and $C(\cdot)$ a generic function of the parameters that appear as arguments of it but not on other quantities. Both C and $C(\cdot)$ may change from line to line.

Theorem 1. *Let Assumption 1 hold. Then any MLE $\widehat{p}_n$ on $\mathcal{M}_\eta$ satisfies*

$$\mathbb{P}\left[h\left(\widehat{p}_n, p^\star\right) \geqslant (z+1)\eta\right] < C_0 \exp\left[-z^2\left(an\eta^2/3\right)\right] \quad \textit{for all } z \geqslant \sqrt{3}\, b/2 \tag{13}$$

and $\mathbb{E}\left[h^2\left(\widehat{p}_n, p^\star\right)\right] \leqslant C(a,b)\,\eta^2$ *with*

$$C_0 = \sum_{k=0}^{+\infty} \exp\left[-4[(4/3)^k - 1]\right] < 1.32. \tag{14}$$

Proof. Assumption 1-(ii) and Proposition 1 imply that, if $\pi_\eta(p^\star) = p_j$, for all $p \in \mathcal{M}_\eta$,

$$\mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{p(X_i)}{p_j(X_i)}\right) \geqslant 0\right] \leqslant \exp\left[-nh^2(p,p_j)\left(1 - A\sqrt{2}\,\frac{h(p^\star,p_j)}{h(p,p_j)}\right)\right]$$

and, since $h(p^\star, p_j) \leqslant \eta$,

$$\mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{p(X_i)}{\pi_\eta(p^\star)(X_i)}\right) \geqslant 0\right] \leqslant \exp\left[-anh^2(p,\pi_\eta(p^\star))\right] \quad \text{if } h(p,p_j) \geqslant b\eta. \tag{15}$$

For $y \geqslant (3/4)(b\eta)^2$ and $k \in \mathbb{N}$, let us set

$$\mathcal{P}_k = \left\{p \in \mathcal{M}_\eta \;\middle|\; y\delta^k \leqslant h^2(\pi_\eta(p^\star), p) < y\delta^{k+1}\right\} \quad \text{with} \quad \delta = 4/3.$$

Since $y\delta^{k+1} \geqslant y\delta \geqslant (b\eta)^2$, we derive from (12) and (10) that

$$|\mathcal{P}_k| \leqslant \exp\left[y\delta^{k+1}\eta^{-2}D\right] \leqslant \exp\left[an\delta^{k+1}y/2\right]$$

and it follows from (15) that

$$\begin{aligned}
&\mathbb{P}\left[h^2\left(\widehat{p}_n, \pi_\eta(p^\star)\right) \geqslant y\right] \\
&\quad\leqslant \mathbb{P}\left[\exists p \in \mathcal{M}_\eta \;\text{ with }\; h^2(p,\pi_\eta(p^\star)) \geqslant y \;\text{ and }\; \sum_{i=1}^n \log\left(\frac{p(X_i)}{\pi_\eta(p^\star)(X_i)}\right) \geqslant 0\right] \\
&\quad\leqslant \sum_{k=0}^{+\infty} \sum_{p\in\mathcal{P}_k} \mathbb{P}\left[\sum_{i=1}^n \log\left(\frac{p(X_i)}{\pi_\eta(p^\star)(X_i)}\right) \geqslant 0\right] \leqslant \sum_{k=0}^{+\infty} \sum_{p\in\mathcal{P}_k} \exp\left[-anh^2(p,\pi_\eta(p^\star))\right] \\
&\quad\leqslant \sum_{k=0}^{+\infty} |\mathcal{P}_k| \exp\left[-an\delta^k y\right] \leqslant \sum_{k=0}^{+\infty} \exp\left[an\delta^{k+1}y/2\right] \exp\left[-an\delta^k y\right] \\
&\quad= \sum_{k=0}^{+\infty} \exp\left[-\frac{an\delta^k y}{3}\right] = \exp\left[-\frac{any}{3}\right]\left(\sum_{k=0}^{+\infty} \exp\left[-\frac{any}{3}(\delta^k - 1)\right]\right).
\end{aligned} \tag{16}$$

Now observe that (10) implies that $b \geqslant 4$ and $an\eta^2 \geqslant 1$, hence, since $y \geqslant (3/4)b^2\eta^2$, $any \geqslant (3/4)b^2an\eta^2 \geqslant 12$. Setting $z = \sqrt{y}/\eta \geqslant \sqrt{3}\,b/2$, we derive from (16) that

$$\begin{aligned}\mathbb{P}\left[h^2\left(\widehat{p}_n, \pi_\eta(p^\star)\right) \geqslant y\right] &< \exp\left[-\frac{any}{3}\right]\left(\sum_{k=0}^{+\infty}\exp\left[-4\left(\left(\frac{4}{3}\right)^k - 1\right)\right]\right)\\ &= C_0\exp\left[-(an\eta^2/3)z^2\right]\end{aligned}$$

and (13) follows from Assumption 1-(i). Finally, the bound for $\mathbb{E}\left[h^2\left(\widehat{p}_n, p^\star\right)\right]$ follows by integration since $\mathbb{E}\left[h^2\left(\widehat{p}_n, p^\star\right)\right] = \int_0^1 \mathbb{P}\left[h^2\left(\widehat{p}_n, p^\star\right) \geqslant x\right] dx$. □

Note that (13) is trivial if $\eta > \left(1 + \sqrt{3}\,b/2\right)^{-1}$ since then $(z+1)\eta > 1$, hence $\mathbb{P}\left[h\left(\widehat{p}_n, p^\star\right) \geqslant (z+1)\eta\right] = 0$, the Hellinger distance being bounded by one. Since $z + 1 \geqslant 2\sqrt{3} + 1 > 4.46$, we shall assume from now on, to avoid trivialities, that $\eta \leqslant 2/9$. Moreover, $an\eta^2/3 \geqslant 1/3$ and $z^2 \geqslant 3b^2/4$ imply that $z^2an\eta^2/3 \geqslant b^2/4$, so that (13) implies that $h\left(\widehat{p}_n, p^\star\right)$ is not larger than $\left(1 + \sqrt{3}\,b/2\right)\eta$ with a probability at least $1 - C_0\exp[-b^2/4] \geqslant 1 - C_0\exp[-4] > 0.975$. It also implies that the smaller η, the better the estimator.

4.3 How to check Assumption 1?

4.3.1 Using nets $\mathcal{M}_\eta$ which are subsets of $\mathcal{M}$

A first way of building a net $\mathcal{M}_\eta$ is to take it as a particular subset of $\mathcal{M}$ chosen in such a way that Assumption 1 holds as in the following parametric examples. In each case, the density model is $\mathcal{M} = \{p_\theta,\, \theta \in \Theta\}$ with Θ a compact subset of $\mathbb{R}^k$ and there is a simple relationship between the Euclidian distance on Θ and the Hellinger distance on $\mathcal{M}$. This relationship implies that, given η, on can find δ such that $\|\theta - \theta'\| \leqslant \delta$ implies that $h(p_\theta, p_{\theta'}) \leqslant \eta$. Therefore a δ-net $\Theta_\delta \subset \Theta$ leads to an η-net $\mathcal{M}_\eta = \{p_\theta,\, \theta \in \Theta_\delta\}$ for $\mathcal{M}$. Then, starting from a mapping $\overline{\pi}$ from Θ to Θ_δ, such that $\|\theta - \overline{\pi}(\theta)\| \leqslant \delta$, we define π_η by $\pi_\eta(p_\theta) = \overline{\pi}(\theta)$ so that Assumption 1-(i) holds. For these parametric models, the MLE on $\mathcal{M}_\eta$ is $\widehat{p}_n = p_{\widehat{\theta}_n}$ and $\widehat{\theta}_n \in \Theta$ is the MLE for the unknown parameter θ.

Example 5 (Gaussian translation)**.** Let P_θ be the normal distribution $\mathcal{N}(\theta, \sigma^2 I_k)$ with $\theta \in \mathbb{R}^k$, $\sigma > 0$ and I_k the identity matrix and let p_θ be the corresponding density with respect to the Lebesgue measure. We assume that σ is known and that Θ is a compact subset of $\mathbb{R}^k$ with diameter bounded by $B\sigma$. It follows from elementary computations that

$$\rho(P_\theta, P_{\theta'}) = \exp\left[-\frac{\|\theta - \theta'\|^2}{8\sigma^2}\right] \quad \text{and} \quad h^2(p_\theta, p_{\theta'}) = 1 - \exp\left[-\frac{\|\theta - \theta'\|^2}{8\sigma^2}\right],$$

which implies, since the function $x \mapsto x^{-1}(1-e^{-x})$ is decreasing on $\mathbb{R}_+$ and $\|\theta-\theta'\| \leqslant B\sigma$, that

$$c_B \frac{\|\theta-\theta'\|^2}{8\sigma^2} \leqslant h^2(p_\theta, p_{\theta'}) \leqslant \frac{\|\theta-\theta'\|^2}{8\sigma^2} \quad \text{for all } \theta, \theta' \in \Theta. \tag{17}$$

with $c_B = 8B^{-2}\left(1-\exp\left[-B^2/8\right]\right)$. Let $\delta = 2\sqrt{2}\sigma\eta$ and Θ_δ be a maximal δ-separated subset of Θ. It is a finite δ-net for Θ and one can deduce from it, as indicated above, the corresponding set $\mathcal{M}_\eta$ and the mappings $\overline{\pi}$ and π_η. Since

$$\int \left(\frac{p_\theta^2(x)}{p_{\theta'}(x)}\right) dx = \exp\left[\frac{\|\theta-\theta'\|^2}{\sigma^2}\right] \quad \text{for all } \theta, \theta' \in \Theta,$$

it follows from (17) that $\int p_\theta^2(x) p^{-1}_{\overline{\pi}(\theta)}(x)\, dx - 1 \leqslant A^2\eta^2$ with

$$A^2 = \eta^{-2}\left(\exp\left[(\delta/\sigma)^2\right]-1\right) = \eta^{-2}\left(\exp\left[8\eta^2\right]-1\right) < 8.42 \quad \text{if} \quad \eta^2 \leqslant 1/80.$$

Then Assumption 1-(ii) holds and we may take $b = 5 > A\sqrt{2}$, in which case $\eta < \left(1+\sqrt{3}\, b/2\right)^{-1}$ and $a > 0.179$. Finally, for $\theta_0 \in \Theta_\delta$,

$$\begin{aligned} \left|\{p_\theta \in \mathcal{M}_\eta \,|\, h(p_\theta, p_{\theta_0}) < x\eta\}\right| &\leqslant \left|\left\{\theta \in \Theta_\eta \,|\, \|\theta-\theta_0\| < 2\sqrt{2/c_B}\,\sigma x\eta\right\}\right| \\ &= \left|\left\{\theta \in \Theta_\eta \,|\, \|\theta-\theta_0\| < c_B^{-1/2} x\delta\right\}\right|, \end{aligned}$$

a quantity which can be bounded using the fact that Θ_δ is a maximal δ-separated subset of Θ and allows to derive a value for D.

To illustrate this, let us assume that Θ is a Euclidean ball of radius $B\sigma/2$ in an affine subset V_j of $\mathbb{R}^k$ with dimension $j \leqslant k$. One can easily derive from volume comparisons that if Θ_δ is a δ-separated subset of V_j, the number of points of Θ_δ that belong to any ball in V_j with radius $y\delta$ is bounded by $(2y+1)^j$. It then follows from the fact that the function $x \to x^{-2}\log\left(2c_B^{-1/2}x+1\right)$ is decreasing for $x \geqslant 1$ that, for $x \geqslant b = 5$,

$$\left|\left\{\theta \in \Theta_\eta \,|\, \|\theta-\theta_0\| < c_B^{-1/2} x\delta\right\}\right| \leqslant \left(2c_B^{-1/2}x+1\right)^j \leqslant \exp\left[j\frac{\log\left(10c_B^{-1/2}+1\right)}{25}x^2\right].$$

Therefore Assumption 1-(iii) holds with $D = \left[(j/25)\log\left(10c_B^{-1/2}+1\right)\right] \bigvee (1/2)$ and (10) holds with $\eta = 3.35\sqrt{D/n}$ since $a > 0.179$. For n large enough, $\eta \leqslant 1/\sqrt{80}$ as required. Finally Theorem 1 implies that $\mathbb{E}\left[h^2\left(\widehat{p}_n, p_{\theta^\star}\right)\right] \leqslant CD/n$, hence by (17),

$$\mathbb{E}\left[\left\|\widehat{\theta}_n - \theta^\star\right\|^2\right] \leqslant C(B)\,\sigma^2 D/n.$$

Example 6 (Cauchy translation)**.** Let p be the Cauchy density with respect to the Lebesgue measure μ on $\mathbb{R}$, i.e. $p(x) = [\pi\,(1+x^2)]^{-1}$, and let $p_\theta(\cdot) = q(\cdot - \theta)$ for $\theta \in \mathbb{R}$. Then

$$\frac{p_\theta(x)}{p_{\theta'}(x)} = \frac{1+(x-\theta')^2}{1+(x-\theta)^2} \leqslant 2\left[1+(\theta'-\theta)^2\right] \quad \text{for all } \theta, \theta', x \in \mathbb{R}. \tag{18}$$

Assume that we observe n i.i.d. real variables with unknown density $p_{\theta^\star}$ with respect to μ belonging to the density model $\mathcal{M} = \{p_\theta,\, \theta \in \Theta\}$ where Θ is an interval of $\mathbb{R}$ with finite length L. It is known — see for instance Chapter 1 of Ibragimov and Has'minskii (1981) — that in this case

$$0 < m_L|\theta - \theta'| \leqslant h(p_\theta, p_{\theta'}) \leqslant (1/4)|\theta - \theta'| \quad \text{for all } \theta, \theta' \in \Theta, \tag{19}$$

with m_L depending on L only. If $\delta = 4\eta \leqslant 8/9$, $\Theta_\delta = (\delta\mathbb{Z}) \cap \Theta$ is a δ-net for Θ and it follows from (19) that $\mathcal{M}_\eta = \{p_\theta,\, \theta \in \Theta_\delta\}$ is an η-net for $\mathcal{M}$ from which we build $\overline{\pi}$ and π_η as previously explained. By (18), $p_\theta / p_{\overline{\pi}(\theta)}$ is uniformly bounded by $\Delta = 2\,[1+16\eta^2] < 4$ since $|\theta - \overline{\pi}(\theta)| \leqslant \delta$, so that Proposition 1 applies, leading to $A < 3\sqrt{2}$. It follows that Assumption 1-(ii) holds with $b = 8$, hence $a \geqslant 1/4$. Finally, if $\theta_0 \in \Theta_\eta$ and $h(\theta, \theta_0) < x\eta$, then $|\theta - \theta_0| < x m_L^{-1}\eta$ so that

$$|\{\theta \in \Theta_\eta \,|\, h(\theta,\theta_0) < x\eta\}| \leqslant \frac{x}{2m_L} \leqslant \exp\left[Dx^2\right] \quad \text{with } D = \frac{\log(4/m_L)}{64} \bigvee \frac{1}{2}$$

since $x \geqslant b = 8$. Therefore, for n large enough, Assumption 1 holds with the choice $\eta = 2\sqrt{2D/n}$ and Theorem 1 implies that $\mathbb{E}\left[h^2\left(\widehat{p}_n, p^\star\right)\right] \leqslant C(m_L)\, n^{-1}$. By (19), the same type of bound holds for $\mathbb{E}\left[(\widehat{\theta}_n - \theta^\star)^2\right]$.

Example 7 (Uniform distributions 1)**.** We observe n i.i.d. real variables with uniform distribution on $[0,\theta]$ with $\theta \in \Theta = \left[\gamma^2\overline{\theta}, \overline{\theta}\right]$, $\overline{\theta} > 0$, $\gamma < 1$. Then, if $\theta' > \theta$,

$$h^2(p_\theta, p_{\theta'}) = 1 - \sqrt{\frac{\theta}{\theta'}} \qquad \text{and} \qquad \int \left(\frac{p_\theta^2(x)}{p_{\theta'}(x)}\right) dx = \frac{\theta'}{\theta}. \tag{20}$$

To build an η-net for $\mathcal{M}$ we set $\theta_j = \overline{\theta}\,(1-\eta^2)^{2j}$ for $j \geqslant 0$, $\mathcal{M}_\eta = \{p_{\theta_j},\, \theta_j \in \Theta\}$ and $\pi_\eta(p_\theta) = p_{\theta_j}$ for $\theta \in (\theta_{j+1}, \theta_j]$. If $\theta \in (\theta_{j+1}, \theta_j]$, $h^2(p_\theta, p_{\theta_j}) \leqslant \eta^2$ and

$$\int \left(\frac{p_\theta^2(x)}{p_{\theta_j}(x)}\right) dx - 1 = \frac{\theta_j}{\theta} - 1 = \frac{1}{\left[1 - h^2(p_\theta, p_{\theta_j})\right]^2} - 1 \leqslant \frac{1}{(1-\eta^2)^2} - 1.$$

This implies that Assumption 1-(i) holds and (11) as well with

$$A = \eta^{-2}\left[\left(1-\eta^2\right)^{-2} - 1\right] < 2.16 \quad \text{since } \eta^2 \leqslant 4/81,$$

leading to the choice $b = 4$ so that $a > 1/5$.

To bound D in (12), we first observe that $h^2(p_{\gamma^2\overline{\theta}}, p_{\overline{\theta}}) = 1-\gamma$ which means that we may restrict to $x\eta \leqslant \sqrt{1-\gamma}$ in (12). Since, for $k \geqslant 1$, $h^2(p_{\theta_j}, p_{\theta_{j+k}}) = 1-(1-\eta^2)^k$, $h(p_{\theta_j}, p_{\theta_{j+k}}) < x\eta$ requires that $1-(1-\eta^2)^k < x^2\eta^2$. Equivalently $k < \log(1-x^2\eta^2)\,/\,\log(1-\eta^2)$ with $x^2\eta^2 \leqslant 1-\gamma$. Since

$$1 < \frac{-\log(1-y)}{y} < 1 + \frac{y}{2(1-y)} \leqslant 1 + \frac{1-\gamma}{2\gamma} \quad \text{for } 0 < y \leqslant 1-\gamma,$$

we conclude that $k < x^2[1+(2\gamma)^{-1}(1-\gamma)]$. It follows that, for $\theta_l \in \Theta$

$$|\{p_\theta \in \mathcal{M}_\eta \,|\, h(p_\theta, p_{\theta_l}) < x\eta\}| \leqslant 2x^2[1+(2\gamma)^{-1}(1-\gamma)],$$

and, since the function $y \to y^{-1}\exp[yD]$ is increasing for $y > D^{-1}$ and $D \geqslant 1/2$, we can take

$$D = \frac{\log\left(2b^2[1+(2\gamma)^{-1}(1-\gamma)]\right)}{b^2} = \frac{\log\left(32[1+(2\gamma)^{-1}(1-\gamma)]\right)}{16}$$

and $\eta = \sqrt{10D/n}$ provided that n is large enough. Then (13) becomes

$$\mathbb{P}\left[h\left(\widehat{p}_n, p^\star\right) \geqslant (z+1)\sqrt{10D/n}\right] < C_0 \exp\left[-\left(2D/3\right)z^2\right] \quad \text{for all } z \geqslant 2\sqrt{3}$$

and we derive from (20) that

$$\mathbb{P}\left[\sqrt{\frac{\widehat{\theta}_n \wedge \theta^\star}{\widehat{\theta}_n \vee \theta^\star}} \leqslant 1 - 10(z+1)^2\frac{D}{n}\right] < C_0 \exp\left[-\frac{2Dz^2}{3}\right] \quad \text{for all } z \geqslant 2\sqrt{3}.$$

4.3.2 Using upper approximations

In the three previous examples, $\mathcal{M}_\eta$ could be chosen as a subset of $\mathcal{M}$ but there are situations for which no finite subset of $\mathcal{M}$ can satisfy Assumption 1-(ii) and we have to build $\mathcal{M}_\eta$ as a finite set of densities with respect to μ that do not belong to $\mathcal{M}$. In such a case we proceed as follows: we build a finite partition $\{\mathcal{M}'_1, \ldots, \mathcal{M}'_N\}$ of $\mathcal{M}$ such that, for each $j \in \{1, \ldots, N\}$ there exists an element $\overline{t}_j \in \mathbb{L}_1(\mu)$ satisfying

$$\sup_{p \in \mathcal{M}'_j} p(x) \leqslant \overline{t}_j(x) \quad \text{for } \mu\text{-almost all } x, \quad \int \overline{t}_j \, d\mu \leqslant 1+\alpha \leqslant 4 \tag{21}$$

and

$$h^2(p, \overline{t}_j) \stackrel{def}{=} \frac{1}{2}\int \left(\sqrt{p} - \sqrt{\overline{t}_j}\right)^2 d\mu \leqslant \eta^2 \leqslant 1/20 \quad \text{for all } p \in \mathcal{M}'_j. \tag{22}$$

To build $\mathcal{M}_\eta$, we then use the normalized versions $t_j = \left(\int \overline{t}_j d\mu\right)^{-1}\overline{t}_j$ of the $\overline{t}_j$, setting $\mathcal{M}_\eta = \{t_1, \ldots, t_N\}$ and $\pi_\eta(p) = t_j$ for all $p \in \mathcal{M}'_j$ and, to derive Assumption 1, we use the following proposition, the proof of which will be defered to Section 5.

Proposition 2. *Let s and $\bar{t}$ be two nonnegative elements of $\mathbb{L}_1(\mu)$ with $0 \leqslant s(x) \leqslant \bar{t}(x)$ for μ-almost all $x \in E$, $\int s\,d\mu = 1$ and $\int \bar{t}\,d\mu = 1+\alpha \geqslant 4$. Let $t = (1+\alpha)^{-1}\bar{t}$. Then $h^2(s,t) \leqslant (1+\alpha)^{-1/2} h^2(s,\bar{t})$ and*

$$\int_{t>0} \left(\frac{s^2}{t}\right) d\mu - 1 \leqslant \left(1+\alpha+\sqrt{1+\alpha}\right) h^2(s,t) \leqslant \left(1+\sqrt{1+\alpha}\right) h^2(s,\bar{t}).$$

Corollary 1. *If (21) and (22) are satisfied, Assumption 1-(i) and (ii) hold with $A^2 < \left(1+\sqrt{1+\alpha}\right)^2/2$ and one can set $b = 4$ so that $a > (1/2) - (\alpha/16) \geqslant 5/16$.*

Proof. In view of (21) and (22), the proposition applies with $(s,\bar{t},t) = (,\bar{t}_j,t_j)$ for each j if $p \in \mathcal{M}'_j$. It implies that $h^2(p,t_j) \leqslant h^2(p,\bar{t}_j)$ for $p \in \mathcal{M}'_j$, hence $h(p,\pi_\eta(p)) \leqslant \eta$ and Assumption 1-(ii) with $A^2 = 1+\sqrt{1+\alpha} < 2+(\alpha/2)$. Therefore, by (21), $A^2 \leqslant 3 \wedge 2[1+(\alpha/8)]^2$ and the choice $b = 4 > A\sqrt{2}$, then leads to

$$a = 1 - \frac{A\sqrt{2}}{4} \geqslant 1 - \frac{2[1+(\alpha/8)]}{4} = \frac{1}{2} - \frac{\alpha}{16} \geqslant \frac{5}{16}. \qquad \square$$

Example 8 (Uniform distributions 2)**.** We observe $n \geqslant 11$ i.i.d. real variables with uniform distribution on $[\theta,\theta+1]$ with $\theta \in \Theta = \mathbb{R}$ and the corresponding density model, with respect to the Lebesgue measure μ, is $\{p_{\theta,\,\theta\in\mathbb{R}}\}$ with $p_\theta = \mathbb{1}_{[\theta,\theta+1]}$. Then, for $\theta < \theta'$, $h^2(p_\theta,p_{\theta'}) = (\theta'-\theta)\wedge 1$. Let us now build a suitable set $\mathcal{M}_\eta$ with $\eta^2 \leqslant 1/20$. For $j \in \mathbb{Z}$, let I_j be the interval $[2j\eta^2, 2(j+1)\eta^2)$ so that the I_j provide a partition of Θ. To each interval I_j we associate the function $\bar{t}_j = \mathbb{1}_{[2j\eta^2,2(j+1)\eta^2+1)}$ so that (21) holds with $\alpha = 2\eta^2 \leqslant 1/10$. Since $\alpha/16 \leqslant 1/160$, one can take $a = 79/160$. Moreover, for $\theta \in I_j$, $[\theta,\theta+1] \subset [2j\eta^2, 2(j+1)\eta^2+1)$, hence

$$h^2(p_\theta,\bar{t}_j) = \frac{1}{2}\int \left(\mathbb{1}_{[2j\eta^2,2(j+1)\eta^2+1)} - \mathbb{1}_{[\theta,\theta+1]}\right)^2 d\mu = \frac{2\eta^2}{2} = \eta^2$$

and (22) holds. Let $t_j = (1+\alpha)^{-1}\bar{t}_j$ be the corresponding density, $\mathcal{M}_\mathbb{Z} = \{t_j, j \in \mathbb{Z}\}$ and set $\pi_\eta(p_\theta) = t_j$ for $\theta \in I_j$. Note that t_j is supported by the interval $[2j\eta^2, 2(j+1)\eta^2+1)$. Since $p^\star = p_{\theta^\star} \in \mathcal{M}$, $\theta^\star$ belongs to some I_j which we may assume, without loss of generality, to be I_0. It follows that all X_i belong a.s. to $(\theta^\star,\theta^\star+1) \subset (0, 2\eta^2+1)$. As a consequence, the likelihood of t_j is a.s. 0 if either $2(j+1)\eta^2+1 \leqslant 0$ or $2j\eta^2 \geqslant 2\eta^2+1$, so that the MLE on $\mathcal{M}_\mathbb{Z}$ necessarily satisfies

$$\widehat{p}_n = t_j \quad \text{with} \quad -\left[1+(2\eta^2)^{-1}\right] < j < \left[1+(2\eta^2)^{-1}\right]$$

and belongs to the set $\mathcal{M}_\eta = \{t_j,\, -N < j < N\}$ with $N-1 < 1+(2\eta^2)^{-1} \leqslant N$. We can therefore consider $\widehat{p}_n$ as the MLE on $\mathcal{M}_\eta$ and we only have to check Assumption 1 on the finite set $\mathcal{M}_\eta$ to which Corollary 1 applies. It remains to check Assumption 1-(iii). For this, let us consider a closed ball $\mathcal{B}_r$ of $\mathcal{M}_\eta$ with Hellinger radius r, $1 \geqslant r \geqslant b\eta = 4\eta$. Since $t_j = (1+2\eta^2)^{-1}\,\mathbb{1}_{[2j\eta^2,2(j+1)\eta^2+1)}$, for

$k \in \mathbb{N}$,

$$h^2\left(t_j, t_{j+k}\right) = \frac{1}{2\left(1+2\eta^2\right)} \int \left(\mathbb{1}_{[2j\eta^2, 2(j+1)\eta^2+1)} - \mathbb{1}_{[2(j+k)\eta^2, 2(j+k+1)\eta^2+1)}\right)^2 d\mu$$
$$= \frac{(4k\eta^2) \wedge (4\eta^2+2)}{2\left(1+2\eta^2\right)} = \frac{2k\eta^2}{1+2\eta^2} \bigwedge 1.$$

If $r = 1$, $\mathcal{B}_r$ contains at most $2N - 1 < 3 + \eta^{-2} \leqslant (23/20)\eta^{-2}$ points. If $r < 1$, it contains at most

$$1 + 2\frac{r^2\left(1+2\eta^2\right)}{2\eta^2} < 1 + \frac{11}{10}\frac{r^2}{\eta^2} \leqslant \frac{93}{80}\frac{r^2}{\eta^2}$$

points since $2\eta^2 \leqslant 1/10$ and $r^2 \geqslant 16\eta^2$. The result still holds if $r = 1$ in view of the bound on $2N - 1$. Since $(93/80)x^2 < \exp\left[x^2/2\right]$ for $x \geqslant 4$, we conclude that (iii) is satisfied with $D = 1/2$ and (10) holds with $a = 79/160$ provided that $\eta^2 = 160/(79n)$, which is compatible with the condition $\eta^2 \leqslant 1/20$ for $n \geqslant 41$. It finally follows from Theorem 1 that

$$\mathbb{P}\left[h\left(\widehat{p}_n, p^\star\right) \geqslant 1.43(z+1)n^{-1/2}\right] < C_0 \exp\left[-z^2/3\right] \quad \text{for all } z \geqslant 2\sqrt{3}.$$

Example 9 (Approximation with respect to the sup norm - General)**.** Let μ be a probability on E, $\mathcal{M}$ a set of densities with respect to μ and assume that $\sqrt{\mathcal{M}} \stackrel{def}{=} \left\{\sqrt{p},\, p \in \mathcal{M}\right\}$ is a totally bounded subset of $\mathbb{L}_\infty(\mu)$. Let $(\mathcal{B}_1, \ldots, \mathcal{B}_{N_\eta})$ be a finite covering of $\sqrt{\mathcal{M}}$ (with respect to the $\mathbb{L}_\infty(\mu)$-distance) by balls of radius $\delta = \eta/\sqrt{2}$, $0 < \eta \leqslant 1/\sqrt{20}$, with respective centers $\sqrt{p_j}$, $1 \leqslant j \leqslant N_\eta$. It follows that, for all p such that $\sqrt{p} \in \mathcal{B}_j$,

$$\bar{t}_j(x) = \left(\sqrt{p_j(x)} + \delta\right)^2 \geqslant p(x) \geqslant \left(\sqrt{p_j(x)} - \delta\right)^2 \quad \mu\text{-a.s.},$$

therefore

$$\left|\sqrt{\bar{t}_j}(x) - \sqrt{p}(x)\right| \leqslant 2\delta \quad \mu\text{-a.s.}, \quad \text{hence} \quad h^2\left(\bar{t}_j, p\right) \leqslant 2\delta^2 = \eta^2 \leqslant 1/20$$

and (22) holds. Moreover, by Jensen's inequality,

$$\int \bar{t}_j = \int \left(\sqrt{p}_j + \delta\right)^2 d\mu = 1 + \delta^2 + 2\delta \int \sqrt{p}_j \leqslant 1 + \left(\eta^2/2\right) + \eta\sqrt{2} < 1.35$$

which implies (21) with $\alpha = 0.35$. We then define t_j and $\mathcal{M}_\eta = \{t_1, \ldots, t_{N_\eta}\}$ as indicated before and choose π_η in such a way that $\pi_\eta^{-1}(t_j) \subset \{p \,|\, \sqrt{p} \in \mathcal{B}_j\}$. Corollary 1 applies leading to Assumption 1-(i) and (ii) with $a > (1/2) - (\alpha/16) > 0.478$. Finally, whatever $p_0 \in \mathcal{M}_\eta$, $|\{p \in \mathcal{M}_\eta \,|\, h(p, p_0) < x\eta\}| \leqslant N_\eta$ and (12) holds with $D = (\log N_\eta)/16$. Then Theorem 1 applies if (10) holds which is the case provided that

$$0.478 n\eta^2 \geqslant (\log N_\eta)/8 \qquad \text{or} \qquad n\eta^2 \geqslant 0.262 \log N_\eta.$$

Such an inequality is satisfied for n large enough and, given n, the optimal value of η is the smallest possible one which clearly depends on the relationship between η and N_η.

Example 10 (Approximation with respect to the sup norm - Smooth densities)**.** Let $E = [0,1]$, μ be the Lebesgue measure on $[0,1]$, w a modulus of continuity on E and $\mathcal{M}$ the set of all densities p with respect to μ satisfying

$$\left|\sqrt{p(y)} - \sqrt{p(x)}\right| \leqslant w(y-x) \quad \text{for all } x < y, \;\; x, y \in E.$$

Let $m_0 = \inf\left\{m \in \mathbb{N},\, m \geqslant 2 \;\middle|\; w(m^{-1}) \leqslant 40^{-1/2}\right\}$, $m \geqslant m_0$ and consider a partition $(I_1, \ldots, I_m)$ of $[0,1]$ into m intervals with the same length $l = m^{-1}$ so that $w(l) \leqslant 40^{-1/2}$. Set $\eta^2 = 2w^2(l) \leqslant 1/20$. For $p \in \mathcal{M}$ and $1 \leqslant j \leqslant m$, we denote by k_j the integer such that $(k_j - 1)w(l) < \sup_{x \in I_j} \sqrt{p(x)} \leqslant k_j w(l)$, so that $(k_j - 2)w(l) < \sqrt{p(x)} \leqslant k_j w(l)$ for all $x \in I_j$ since the variation of $\sqrt{p}$ on I_j is bounded by $w(l)$. We finally set

$$\mathbf{k}(p) = (k_1, \ldots, k_m) \qquad \text{and} \qquad \bar{t}_{\mathbf{k}(p)} = \left[\left(\sum_{j=1}^{m} k_j \mathbb{1}_{I_j}\right) w(l)\right]^2. \tag{23}$$

Then

$$\sqrt{\bar{t}_{\mathbf{k}(p)}(x)} - \sqrt{2}\eta < \sqrt{p(x)} \leqslant \sqrt{\bar{t}_{\mathbf{k}(p)}(x)} \quad \text{for all } x \in [0,1] \tag{24}$$

so that

$$h^2\left(\bar{t}_{\mathbf{k}(p)}, p\right) = \frac{1}{2}\int \left(\sqrt{\bar{t}_{\mathbf{k}(p)}} - \sqrt{p}\right)^2 d\mu \leqslant \eta^2.$$

and, by Jensen's inequality,

$$1 \leqslant \int \bar{t}_{\mathbf{k}(p)}\, d\mu \leqslant \int \left(\sqrt{p} + \sqrt{2}\eta\right)^2 d\mu = 1 + 2\eta^2 + 2\sqrt{2}\eta \int \sqrt{p} d\mu < 7/4.$$

Performing this procedure for all $p \in \mathcal{M}$ leads to a set $\mathcal{M}_\eta = \{t_{\mathbf{k}},\, \mathbf{k} \in \mathbf{K}\}$ with $\mathbf{K} \subset \mathbb{N}^m$ and such that each $p \in \mathcal{M}$ can be approximated by some $\bar{t}_{\mathbf{k}(p)}$, $\mathbf{k}(p) \in \mathbf{K}$ satisfying (24). This results in the function π_η given by $\pi_\eta(p) = t_{\mathbf{k}(p)}$ and we may apply Corollary 1 with $\alpha = 3/4$ so that Assumptions 1-(i) and (ii) hold and $a > 29/64$.

Let us now bound $|\mathbf{K}|$. For this we observe that all the $\bar{t}_{\mathbf{k}}$ with $\mathbf{k} \in \mathbf{K}$ necessarily share the following properties. Since p is a continuous density, it takes the value 1 in some interval I_j which implies that, on this I_j, $1 \leqslant \sqrt{\bar{t}_{\mathbf{k}(p)}} \leqslant 1 + w(l)$. If $(k_0 - 1)w(l) < 1 \leqslant k_0 w(l)$, then $\sqrt{\bar{t}_{\mathbf{k}(p)}}$ equals either $k_0 w(l)$ or $(k_0 + 1)w(l)$ on I_j. Moreover, since the variation of $\sqrt{p}$ on each I_j is bounded by $w(l)$, we have $k_{j+1} = k_j + \gamma w(l)$ with $\gamma = -1, 0$ or 1 in (23). This means that

$$|\mathbf{K}| \leqslant 2m3^{m-1} = \exp[(m-1)\log 3 + \log(2m)] < \exp[3m/2],$$

hence (12) holds with $D = 3m/32$ since $b = 4$. Finally, Assumption 1 holds provided that

$$(29n/32)w^2(m^{-1}) \geqslant 3m/16 \qquad \text{or equivalently} \qquad nw^2(m^{-1}) \geqslant 6m/29, \tag{25}$$

which can always be realized for some $m \geqslant m_0$ as soon as $n \geqslant (6/29)m_0 w^{-2}(m_0^{-1})$. Since we want η to be as small as possible, we choose for m the largest integer which satisfies the previous inequality.

Let us now turn to a concrete illustration, assuming that $w(x) = Lx^\beta$ for $0 \leqslant x \leqslant 1$ with $L > 0$ and $0 < \beta \leqslant 1$, which corresponds to Hölderian smoothness for $\sqrt{p}$. Then $w(m^{-1}) = Lm^{-\beta}$ and (25) amounts to $m \leqslant \left[(29/6)nL^2\right]^{1/(2\beta+1)}$. Assuming that n is large enough to warrant that $m_0 \leqslant \left[(29/6)nL^2\right]^{1/(2\beta+1)}$ we should choose m such that

$$\left[(29/6)nL^2\right]^{1/(2\beta+1)} - 1 < m \leqslant \left[(29/6)nL^2\right]^{1/(2\beta+1)}.$$

The right-hand side is at least 2 since $m \geqslant 2$, hence $m > (1/2)\left[(29/6)nL^2\right]^{1/(2\beta+1)}$ and

$$\begin{aligned}\eta = \sqrt{2}w(m^{-1}) &\leqslant \sqrt{2}Lm^{-\beta} \\ &\leqslant 2^{\beta+(1/2)}L\left[(29/6)nL^2\right]^{-\beta/(2\beta+1)} = C(\beta)L^{1/(2\beta+1)}n^{-\beta/(2\beta+1)},\end{aligned}$$

which implies that

$$\mathbb{E}\left[h^2\left(\widehat{p}_n, p^\star\right)\right] \leqslant C(\beta)L^{2/(2\beta+1)}n^{-2\beta/(2\beta+1)}.$$

Moreover, since all elements of $\mathcal{M} \cup \mathcal{M}_\eta$ are uniformly bounded by a constant depending only on L and β,

$$\begin{aligned}\left(\widehat{p}_n(x) - p^\star(x)\right)^2 &= \left(\sqrt{\widehat{p}_n(x)} - \sqrt{p^\star(x)}\right)^2\left(\sqrt{\widehat{p}_n(x)} + \sqrt{p^\star(x)}\right)^2 \\ &\leqslant C(L,\beta)\left(\sqrt{\widehat{p}_n(x)} - \sqrt{p^\star(x)}\right)^2,\end{aligned}$$

therefore, if $\|\cdot\|_2$ denotes the norm in $\mathbb{L}_2(\mu)$,

$$\mathbb{E}\left[\|\widehat{p}_n - p^\star\|_2^2\right] \leqslant C(L,\beta)n^{-2\beta/(2\beta+1)} \quad \text{for all } p^\star \in \mathcal{M}.$$

5 Additional proofs

Proof of Proposition 1 First observe that, since $s = 0$ μ-a.s. when $t = 0$,

$$\begin{aligned}\int_{t>0}\left(\frac{s}{t} - 1\right)^2 t\, d\mu &= \int_{t>0}\left(\sqrt{s} - \sqrt{t}\right)^2 \frac{\left(\sqrt{s} + \sqrt{t}\right)^2}{t}\, d\mu \\ &> \int_{t>0}\left(\sqrt{s} - \sqrt{t}\right)^2 d\mu = 2h^2(s,t),\end{aligned} \tag{26}$$

hence $A^2 > 2$. If we set $\gamma(x) = [s(x)/t(x)] - 1$ for $x \in E$, then $s = t(1+\gamma)$, $\int t\gamma\, d\mu = 0$ and

$$\mathbb{E}_s\left[\sqrt{u(X)/t(X)}\right] = \int\sqrt{tu}\, d\mu + \int\sqrt{tu}\,\gamma\, d\mu = 1 - h^2(t,u) + \int\sqrt{tu}\,\gamma\, d\mu. \tag{27}$$

Since $\int t\gamma\, d\mu = 0$ we derive from Cauchy-Schwarz inequality and (11) that

$$\begin{aligned}\int \sqrt{tu}\,\gamma\, d\mu &= \int \sqrt{tu}\,\gamma\, d\mu - \int t\gamma\, d\mu = \int \gamma\sqrt{t}\left(\sqrt{u}-\sqrt{t}\right) d\mu \\ &\leqslant \Big[\int \gamma^2 t\, d\mu \int \left(\sqrt{u}-\sqrt{t}\right)^2 d\mu\Big]^{1/2} \leqslant h(t,u)\sqrt{2\int_{t>0}\left(st^{-1}-1\right)^2 t\, d\mu}\end{aligned}$$

and (27) becomes

$$\begin{aligned}\mathbb{E}_s\left[\sqrt{\frac{u(X)}{t(X)}}\right] &\leqslant 1 - h^2(t,u) + h(t,u)\left[\sqrt{2}\,Ah(s,t)\right] \\ &\leqslant \exp\left[-h^2(t,u)\left(1-\sqrt{2}\,A\frac{h(s,t)}{h(t,u)}\right)\right]\end{aligned}$$

and (9) follows from (2). Finally, if $s(x) \leqslant \Delta t(x)$ μ-a.s., by (26),

$$\int_{t>0}\left(\frac{s}{t}-1\right)^2 t\, d\mu \leqslant \left(1+\sqrt{\Delta}\right)^2 \int_{t>0}\left(\sqrt{s}-\sqrt{t}\right)^2 d\mu = 2\left(1+\sqrt{\Delta}\right)^2 h^2(s,t)$$

which leads to $A^2 = 2\left(1+\sqrt{\Delta}\right)^2$.

Proof of Proposition 2 Let us set $\gamma = (\bar{t}/s) - 1 \geqslant 0$. Then $\int \gamma s\, d\mu = \alpha$ and $t = (1+\gamma)s/(1+\alpha)$. It follows that

$$h^2\left(s,\bar{t}\right) = \frac{1}{2}\int\left(\sqrt{1+\gamma}-1\right)^2 s\, d\mu = 1 + \frac{\alpha}{2} - \int\sqrt{1+\gamma}\, s\, d\mu \tag{28}$$

and

$$h^2(s,t) = 1-\rho(s,t) = \int\left(1-\sqrt{\frac{1+\gamma}{1+\alpha}}\right) s\, d\mu = \frac{\sqrt{1+\alpha} - \int\sqrt{1+\gamma}\, s\, d\mu}{\sqrt{1+\alpha}},$$

therefore, by (28),

$$\sqrt{1+\alpha}\, h^2(s,t) = \sqrt{1+\alpha} - \int\sqrt{1+\gamma}\, s\, d\mu \leqslant h^2\left(s,\bar{t}\right). \tag{29}$$

Moreover,

$$\int_{t>0}\frac{s^2}{t}d\mu - 1 = \int\left(\frac{1+\alpha}{1+\gamma}-1\right) s\, d\mu.$$

Since $\gamma \geqslant 0$ and the function $x \to (x-1)/\left(1-x^{-1/2}\right) = x+\sqrt{x}$ is increasing for $x \geqslant 0$,

$$\frac{[(1+\alpha)/(1+\gamma)]-1}{1-[(1+\alpha)/(1+\gamma)]^{-1/2}} \leqslant \frac{1+\alpha-1}{1-(1+\alpha)^{-1/2}} = 1+\alpha+\sqrt{1+\alpha}$$

and it follows from (29) that

$$\int_{t>0} \frac{s^2}{t} d\mu - 1 \leqslant \left(1+\alpha+\sqrt{1+\alpha}\right) \int \left(1 - \sqrt{\frac{1+\gamma}{1+\alpha}}\right) s \, d\mu$$
$$= \left(1+\alpha+\sqrt{1+\alpha}\right) h^2(s,t) \leqslant \left(1+\sqrt{1+\alpha}\right) h^2(s,\bar{t}).$$

Remerciements Je tiens à remercier tout particulièrement Yannick Baraud pour son soutien et ses commentaires critiques et suggestions à propos d'une première version de ce texte.

References

Baraud, Y., Birgé, L., and Sart, M. (2017). A new method for estimation and model selection: ρ-estimation. *Invent. Math.*, 207(2):425–517.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.*, 42(3):273–325.

Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability (Univ. California, Berkeley, Calif., 1970/1971), Vol. I: Theory of statistics*, pages 175–194. Univ. California Press, Berkeley, Calif.

Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical Estimation. Asymptotic Theory*, volume 16. Springer-Verlag, New York.

Le Cam, L. (1970). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Ann. Math. Statist.*, 41:802–828.

Le Cam, L. (1990). Maximum likelihood: An introduction. *Inter. Statist. Review*, 58(2):153–171.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003.

van de Geer, S. A. (2000). *Applications of empirical process theory*, volume 6 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.

van der Vaart, A. W. (1998). *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge.