ESAIM: Probability and Statistics July 1997, Vol. 1, pp. 285–317

TESTING IN LOCALLY CONIC MODELS, AND APPLICATION TO MIXTURE MODELS

D. DACUNHA-CASTELLE AND É. GASSIAT

Abstract. In this paper, we address the problem of testing hypotheses using maximum likelihood statistics in non identifiable models. We derive the asymptotic distribution under very general assumptions. The key idea is a local reparameterization, depending on the underlying distribution, which is called locally conic. This method enlights how the general model induces the structure of the limiting distribution in terms of dimensionality of some derivative space. We present various applications of the theory. The main application is to mixture models. Under very general assumptions, we solve completely the problem of testing the size of the mixture using maximum likelihood statistics. We derive the asymptotic distribution of the maximum likelihood statistic ratio which takes an unexpected form.

1. Introduction

In this paper, we study the problem of hypothesis testing using maximum likelihood statistics in very general and various situations. The originating question was to solve the problem for general finite mixtures. Indeed, the problem is nor clearly neither completely solved in the literature. Partial solutions may be found for example in Berdai and Garel (1994), Ghosh and Sen (1985), Self and Liang (1987). Redner proved in Redner (1981) that the maximum likelihood estimators for finite mixtures with compact parameter space is consistent in the quotient parameter space (when quotient is taken with respect to identifiable classes). This result, though interesting, is not very tractable. Bickel and Chernoff give the asymptotic distribution of the supremum of some process which is related, following Hartigan (Hartigan (1985)), to the problem of testing a mixture of two normal distributions with same variance against a pure normal (see Bickel and Chernoff (1993)) in the *simple* mixture model, see (4.1) below in section 4.
In particular, Ghosh and Sen (Ghosh and Sen (1985)) state the asymptotic distribution of the maximum likelihood statistics for testing one population against two populations. However, their formulation requires some strong separation of the populations which is highly unsatisfactory. To be more precise and to introduce the key ideas of our solution, let us discuss briefly the simplest problem of population mixture. So let

$$g_{\pi,\gamma_1,\gamma_2} = (1-\pi) f_{\gamma_1} + \pi f_{\gamma_2} \tag{1.1}$$

URL address of the journal: **http://www.emath.fr/ps/**
Received by the journal May 23, 1995. Revised July 1996 and January 1997. Accepted for publication April 21, 1997.

be the model, where $\mathcal{F} = (f_\gamma)_{\gamma\in\Gamma}$ is a parametric regular family of densities with respect to some positive measure ν, and $\pi \in [0,1]$. The problem is to test $\exists\gamma_0,\ g_{\pi,\gamma_1,\gamma_2} = f_{\gamma_0}$ against $\forall\gamma_0,\ g_{\pi,\gamma_1,\gamma_2} \neq f_{\gamma_0}$. We now have $g_{\pi,\gamma_1,\gamma_2} = f_0$ if and only if ($\gamma_1 = 0$ and $\pi = 0$) or ($\gamma_2 = 0$ and $\pi = 1$) or ($\gamma_1 = 0$ and $\gamma_2 = 0$) . Considering the result of Redner (Redner (1981)), it would be possible to consider submodels. To find the distribution of the maximum likelihood statistic, the usual method is to make expansions around the true value of the parameter, to perform some maximization upon the identifiable parameter (in the submodel), and then to maximize the maximum upon the non identifiable parameter. Doing so, to be able to obtain a result, it is necessary to have a careful control over the remaining terms of the expansions, with respect to the complete first terms, including their coefficients depending upon the non identifiable parameter. But in our specific problem, when considering directional models, the degeneracy of Fisher information leads to the fact that the classical technic may not be performed. The remaining terms are not uniformly small with respect to the complete first terms. Indeed, consider the submodel

$$\pi f_{\gamma_1} + (1-\pi) f_{\gamma_2}$$

with $\pi \approx 0$, $\gamma_2 \approx 0$ and γ_1 free. Making an expansion with γ_1 fixed we have

$$\begin{aligned} l_n(\pi,\gamma_2) &= \pi \sum \frac{f_{\gamma_1} - f_0}{f_0}(X_i) \\ &\quad + \left(\gamma_2 \sum \frac{f_0'}{f_0}(X_i) + \frac{1}{2}\gamma_2^2 \sum \frac{f"_0}{f_0}(X_i)\right)(1-\pi) \\ &\quad - \frac{1}{2}\left(\pi \sum \frac{f_{\gamma_1} - f_0}{f_0}(X_i) + \gamma_2 \sum \frac{f_0'}{f_0}(X_i)\right)^2 + o(same). \end{aligned}$$

For γ_1 fixed, the involved matrix $\Gamma_n(\gamma_1)$ tending to Fisher information $\Gamma(\gamma_1)$ is invertible for big n, and we have $(\hat\pi, \hat\gamma_2) \approx \Gamma_n(\gamma_1)^{-1} V_n(\gamma_1)$ with $V_n(\gamma_1)$ the score. It follows that for γ_1 fixed $\sup l_n \approx \frac{1}{2} V_n(\gamma_1)\Gamma_n(\gamma_1)^{-1} V_n(\gamma_1)$. Now, when letting γ_1 tend to 0, we obtain $(\hat\pi, \hat\gamma_2) \approx (1, \gamma_1)$. This contradicts the fact that $\pi \approx 0$, but more importantly by letting γ_1 go *slowly* to 0, the remaining terms in the expansion may be unbounded ! This shows that there is a need to separate γ_1 goes slowly to 0 and γ_1 is bounded away from 0, where $\sup l_n$ has a different behavior; this separation may not be done using γ_1 only.

Here, we propose a complete solution to this specific problem without any extra assumption on the parameters. The driving idea is to parameterize in such a way that one of the parameters is identifiable at the previously non identifiable point, so that it is possible to have asymptotic expansions in its neighborhood, and the other parameter contains all the non identifiability. We call such parameterization *locally conic*. We thus propose a new reparameterization of the mixture family space. An important property is also that all directional Fisher informations are uniformly equal to one. The first parameter can be thought around the true point as something close to a distance, the other parameter can be thought as a "direction". The first parameter is thus the only parameter that is identifiable under the null hypothesis, and the second one, around the true distribution, may be seen

as a nuisance parameter. It can not be consistently estimated. When a "direction" is fixed, the model is supposed to be regular , which, of course, does not imply the regularity of the whole model. Doing so, the key point is to assume that the closure of the derivatives in any direction at point 0 of the log-likelihood is a Donsker class of random variables, so that we prove easily that the asymptotic distribution of the maximum log-likelihood is a function of the supremum of a Gaussian process. Moreover, the first "distance" parameter converges in distribution with parametric speed $\sqrt{n}$. When testing one population against two populations, the asymptotic distribution has a term coming from "second order" unboundness, see Theorem 4.3. The simple mixture model does not lead to such extra term, compare Theorem 4.1 and Theorem 4.3.

This situation is not proper to mixture models. In this paper, we present an abstract general parameterization to find the asymptotics of maximum likelihood statistics, and its application to hypothesis testing. These general models are not identifiable in general and can be nonparametric models. We call them *locally conic models*. We develop here two major applications: mixture models, and usual parametric models. Applied to parametric models, this point of view underlines naturally the role of the geometric structure of the parameter space around the null hypothesis in the precise formulation of the limit distribution. Applied to mixture models, this leads to a theorem where it appears that an unexpected term in the limit distribution comes from the non identifiability of the model. The locally conic parameterization allows a clear understanding of what happens due to the non identifiability. Nonparametric testing (and the associated estimation of our "distance" parameter) of a probability density may be carried out using our theory for contamination (or perturbation) models. Indeed, they are an extension of the simple mixture model. Applications to ARMA processes are quickly explained and are developed in another paper (Dacunha-Castelle and Gassiat (1996)).

The organization of the paper is the following: in a first section, we set the general point of view and assumptions on the model. We explain the driving ideas. In a subsequent section, we prove an abstract result concerning the case where "classical" technic may be performed: convergence, asymptotic distribution of the maximum log-likelihood statistic and of the first "distance" parameter θ, together with asymptotic distribution of the maximum log-likelihood under contiguous sequences. We then show how these results apply to the problem of hypothesis testing, and how they apply to the classical parametric situation, with particular attention to the geometry of the parameter space. In section 4, we solve the problem for population mixtures, and in section 5 we propose further remarks and applications, in particular to nonparametric perturbation models and to ARMA models. Proofs of the main results are given in section 6.

2. Locally conic parameterization

$\mathcal{G}$ is a set of probability densities g in $L_1(\nu)$, where ν is a positive measure on $\mathbb{R}^k$. Most often in the sequel we refer to the situation where we observe

a sample $(X_1, \ldots, X_n)$ of i.i.d. random vectors with common distribution the underlying probability $g_0\nu$.

The fundamental assumption on the model is the following: we assume that there exists a parameterization of $\mathcal{G}$ through two parameters θ and β: $(\theta, \beta) \in [0, M] \times \mathcal{B}$, M is a positive real number, $\mathcal{B}$ is a precompact set in a Polish space with metric d.

$$\mathcal{G} = \{g_{(\theta,\beta)}, (\theta, \beta) \in \mathcal{T}\}$$

where $\mathcal{T} \subset [0, M] \times \mathcal{B}$ is endowed with the product topology of $\mathbb{R}$ and $\mathcal{B}$. $\overline{\mathcal{T}}$ is the (compact) closure of $\mathcal{T}$. The parameterization satisfies the following assumptions:

(A1) It is possible to extend the application $(\theta, \beta) \to g_{(\theta,\beta)}$ to a map from $\overline{\mathcal{T}}$ to $\mathcal{G}$ such that $(\theta, \beta) \to g_{(\theta,\beta)}(x)$ is continuous ν a.s., and:

$$g_{(\theta,\beta)} = g_0 \iff \theta = 0.$$

For any β, let

$$\theta_\beta = \sup\{t \geq 0 \ : \ [0, t] \times \{\beta\} \subset \mathcal{T}\}.$$

We say that a model is *locally conic* if the local parameterization verifies:

$$\text{(A2)} \qquad \forall \beta \in \mathcal{B}, \theta_\beta > 0.$$

This assumption says that it is impossible to find accumulation sequences of parameter leading to $\theta = 0$ with directions β where the submodel $(g_{(\theta,\beta)}\nu$, $(\theta, \beta) \in \mathcal{T})_\theta$ (where β is fixed) is not defined in a right neighborhood of 0.

The driving ideas are the following. First, to be able to expand the likelihood, we need a point around which to make the expansion. In other words, we need a parameter θ which can be consistently estimated. This is the reason of the locally conic parameterization such that (A1) and (A2) hold. Second, we have to make an expansion till the remaining terms may be *uniformly* bounded, so that the maximization may also be performed on the parameter β. In the parametric situation and for the simple mixture, an expansion till order 2 will be enough, see the next section. In the mixture model, this is not possible any more, as we shall explain further. However, the locally conic parameterization allows to see exactly what happens and to find the solution.

The first point which holds for both applications is the uniform convergence of the estimator of θ, we show it now. The log-likelihood is:

$$l_n(\theta, \beta) = \sum_{i=1}^{n} \log g_{(\theta,\beta)}(X_i).$$

Define the maximum likelihood estimator $(\hat{\theta}, \hat{\beta})$ to be any maximizer of l_n over $\overline{\mathcal{T}}$, which exists, thanks to (A1). As usual we shall need:

(AC) There exists a function h in $L_1(g_0\nu)$ such that: $\forall g \in \mathcal{G}$, $|\log g| \leq h$ ν-a.e.

The following Theorem states the convergence of $\hat{\theta}$:

THEOREM 2.1. *Under assumptions (A1) (A2) and (AC), $\hat{\theta}$ converges in probability to 0 as n tends to infinity.*

Notice that $\hat{\beta}$ may or may not converge, see the examples developed in subsequent sections: for mixture models, $\hat{\beta}$ does not converge in general, and for regular parametric models, $\hat{\beta}$ converges in probability.

To study the maximum likelihood statistic, we shall use Taylor expansions. The first term in the expansion is the empirical process of the first derivative of the density. The uniformity of the convergence in the central limit theorem will be the second key point in the paper. Define H the Hilbert space $L_2(g_0 \cdot \nu)$, and

$$\mathcal{D} = \{\frac{g'_{(0,\beta)}}{g_0},\ \beta \in \mathcal{B}\}.$$

(AD) $\mathcal{D}$ is a Donsker class. Let ξ_d be a Gaussian process on $\mathcal{D}$ with covariance the usual scalar product in H and with continuous sample paths w.r.t. the intrinsic variance metric.

Donsker classes are defined in Van der Vaart and Wellner (1996). Roughly speaking, a Donsker class is a set of functions for which the empirical distributions (with i.i.d. variables) verify a uniform central limit theorem, with limit distribution a Gaussian process.

(AN) We assume that the following normalization condition holds:

$$\forall d \in \overline{\mathcal{D}}, \|d\|_H = 1.$$

So that $\mathcal{D}$ is a subset of the unit sphere in H. $\overline{\mathcal{D}}$ is then a compact subset of this unit sphere in H, since Donsker classes are necessarily precompact.

COMMENTS ON THE ASSUMPTIONS.

- The parameterization depends upon the underlying distribution g_0. For instance, in case of simple mixtures such as (4.1), we set

$$g_{\pi,\gamma} = f_0(1 + \theta \cdot \beta),\ \ \theta = \|\frac{g_{\pi,\gamma} - f_0}{f_0}\|_H,\ \ \beta = \frac{g_{\pi,\gamma} - f_0}{f_0} \cdot \|\frac{g_{\pi,\gamma} - f_0}{f_0}\|_H^{-1}.$$

  and in case of parametric models (g_γ), where $g_0 = g_{\gamma_0}$:

$$g_\gamma = g_{\gamma_0 + \theta\beta},\ \ \theta^2 = (\gamma - \gamma_0)^T \cdot I(\gamma_0) \cdot (\gamma - \gamma_0),\ \ \beta = \frac{\gamma - \gamma_0}{\theta}$$

  where $I(.)$ is the Fisher information of the model.
- Sufficient conditions for a set to be a Donsker class of functions are given in Van der Vaart and Wellner (1996). A sufficient condition for $\mathcal{D}$ to be a Donsker class is that the L_2-entropy with bracketing is integrable, see Ossiander (1987).
- Sufficient conditions for a Gaussian process to have continuous sample paths are given in Dudley (1967). A sufficient condition is that the L_2-entropy is integrable. The existence of a continuous Gaussian process is automatic if the class is Donsker.
- The assumptions imply that $(\mathcal{D})^2$ is a Glivenko-Cantelli class in probability.
- The parameterization may be non identifiable. The only identifiability is that of θ at point g_0.
- When β lies on $\overline{\mathcal{B}}$, it is possible that the only possible θ is $\theta = 0$: this comes from the originating non identifiability. To expand the likelihood, the parameter set may not be taken as $\overline{T}$.

3. Asymptotic results: simple case

3.1. General results

Assume:

(A3) For all β in $\mathcal{B}$, $g_{\theta,\beta}$ is twice continuously differentiable with respect to θ ν a.e., with right continuous derivatives at point $\theta = 0$. Denote by $g'_{\theta,\beta}$ and $g"_{\theta,\beta}$ the derivatives (which are right derivatives at $\theta = 0$). These derivatives are ν a.s. continuous over $\overline{\mathcal{T}}$ (with respect to (θ, β)).

To have uniformly small remainder terms in the Taylor expansion till order 2, we introduce:

(A4) There exists real functions l and m such that:

$$\forall(\theta,\beta) \in \overline{\mathcal{T}},\ \mid \frac{g'_{(\theta,\beta)}}{g_{(\theta,\beta)}} \mid \leq l \text{ and } \mid \frac{g"_{(\theta,\beta)}}{g_{(\theta,\beta)}} \mid \leq m$$

with

$$E_{g_0\nu}[l]^2 < +\infty \text{ and } E_{g_0\nu} m < +\infty.$$

Notice that (A3), (A4) state the regularity of the model parameterized only with $\theta \geq 0$ when β is fixed, which is defined on a small right-neighborhood of 0, $[0, \theta_\beta]$, thanks to (A2).

Define T_n as the maximum likelihood statistic: $T_n = l_n(\hat{\theta}, \hat{\beta})$. We have the following asymptotic result:

Theorem 3.1. *Assume (A1),(A2), (A3), (AD), (AN), (AC), (A4) hold. Then, under $g_0.\nu$, $T_n - l_n(0)$ converges in distribution to the following variable:*

$$\frac{1}{2} \cdot \sup_{d\in\mathcal{D}} (\xi_d)^2 \cdot 1_{\xi_d \geq 0}.$$

Remark 3.2. Depending on the structure of the Gaussian process on $\mathcal{D}$, the indicator function may disappear in the limit variable. For the classical parametric case, it depends on the geometrical structure around θ_0, see Section 3.1.

The following theorem states the asymptotic distribution of $\hat{\theta}$:

Theorem 3.3. *$\sqrt{n} \cdot \hat{\theta}$ converges in distribution as n tends to infinity to*

$$\sup_{d\in\mathcal{D}} (\xi_d) \cdot 1_{\xi_d \geq 0}$$

where ξ_d is the Gaussian process on $\mathcal{D}$ with covariance the usual scalar product in H.

It is possible to check the asymptotic limit distribution of the log-likelihood statistic for each direction under alternative contiguous distributions as usual. The following result was proved by Ghosh and Sen in Ghosh and Sen (1985) for mixtures of two populations under their strong separation assumption.

Theorem 3.4. *Assume the underlying distribution is $g_{(\theta_n,\beta_0)} \cdot \nu$, where $\theta_n = c/\sqrt{n}$, c a positive real number, $\beta_0 \in \mathcal{B}$. For any β in $\mathcal{B}$, define $V_n(\beta) =$*

$\sup_{\theta\,:\,(\theta,\beta)\in\mathcal{T}} l_n(\theta,\beta)$. *Then,* $V_n(\beta) - l_n(0)$ *converges in distribution to:*

$$\frac{1}{2}\cdot(\xi_d + c\langle d, d_0\rangle_H)^2\cdot 1_{\xi_d + c\langle d,d_0\rangle_H \geq 0}$$

where $d = \frac{g'_{(0,\beta)}}{g_0}$ *and* $d_0 = \frac{g'_{(0,\beta_0)}}{g_0}$.

Of course, this is not completely satisfactory since this result holds only directionally. We prove the following result:

THEOREM 3.5. *Assume (A1),(A2), (A3), (AD), (AN), (AC), (A4) hold. Under the distribution* $g_{(\theta_n,\beta_0)}\cdot\nu$, *where* $\theta_n = c/\sqrt{n}$, *c a positive real number,* $\beta_0 \in \mathcal{B}$, *the maximum likelihood statistic* $T_n - l_n(0)$ *converges in distribution to:*

$$\frac{1}{2}\cdot\sup_{d\in\mathcal{D}}(\xi_d + c\langle d, d_0\rangle_H)^2\cdot 1_{\xi_d + c\langle d,d_0\rangle_H \geq 0}$$

where $d_0 = \frac{g'_{(0,\beta_0)}}{g_0}$.

3.2. Application to hypothesis testing

As was underlined before, the locally conic parameterization depends on the unknown true density. However, the maximum likelihood statistic does not depend on the parameterization, it only depends on the family $\mathcal{G}$, whatever be its description. Moreover, when subtracting two maximum likelihood statistics over different models, the difference makes the terms $l_n(0)$ disappear. It is then clear that the previous results allow to test hypotheses using maximum likelihood statistics with asymptotically known level in the following way. Define $\mathcal{T}_0$ and $\mathcal{T}_1$ to be sets of parameters such that the models $\mathcal{G}_0 = \{g_{(\theta,\beta)},\ (\theta,\beta)\in\mathcal{T}_0\}$ and $\mathcal{G}_1 = \{g_{(\theta,\beta)},\ (\theta,\beta)\in\mathcal{T}_1\}$ verify all assumptions of section 2.1., $\mathcal{T}_0 \subset \mathcal{T}_1$. Define:

$$T_n(i) = \sup_{(\theta,\beta)\in\mathcal{T}_i} l_n(\theta,\beta) \quad i = 1,2.$$

We have

THEOREM 3.6. *Suppose* $g_0 \in \mathcal{G}_0$. *The asymptotic level of the test of* H_0 : $(\theta,\beta)\in\mathcal{T}_0$ *against* $H_1 : (\theta,\beta)\in\mathcal{T}_1 - \mathcal{T}_0$ *with critical region*

$$T_n(1) - T_n(0) \geq C_\alpha$$

is:

$$\alpha := P\Big(\frac{1}{2}\cdot\sup_{d\in\overline{\mathcal{D}}_1}(\xi_d)^2\cdot 1_{\xi_d\geq 0} - \frac{1}{2}\cdot\sup_{d\in\overline{\mathcal{D}}_0}(\xi_d)^2\cdot 1_{\xi_d\geq 0} \geq C_\alpha\Big)$$

with obvious notations.

The proof follows that of Theorem 3.1 for the distribution of $(T_n(1) - l_n(0), T_n(0) - l_n(0))$ where the true distribution g_0 lies in $\mathcal{G}_0$.
If the true distribution is a fixed g_1 not in $\mathcal{G}_0$, the asymptotic power of the test is obviously one. If the true distribution is $g_{(\theta_n,\beta_0)}\cdot\nu$ as in Theorem 3.4, the asymptotic power is:

$$P\Big(\frac{1}{2}\cdot\sup_{d\in\overline{\mathcal{D}}_1}(\xi_d + c\langle d,d_0\rangle_H)^2\cdot 1_{\xi_d + c\langle d,d_0\rangle_H\geq 0} - \frac{1}{2}\cdot\sup_{d\in\overline{\mathcal{D}}_0}(\xi_d + c\langle d,d_0\rangle_H)^2\cdot 1_{\xi_d + c\langle d,d_0\rangle_H\geq 0} \geq C_\alpha\Big)$$

where $d_0 = g'_{(0,\beta_0)}/g_0$.

REMARK 3.7.

- To compute the asymptotic distribution of $T_n(1) - T_n(0)$, notice that the processes involved are correlated and that $\overline{\mathcal{D}}_0 \subset \overline{\mathcal{D}}_1$.
- The limit distribution may depend on g_0 or may be free of g_0. This depends on the spaces $\mathcal{B}$ and $\overline{\mathcal{D}}$. Indeed, if $\mathcal{B}$ does not depend on g_0, the distribution of the supremum over $\overline{\mathcal{D}}$ of the square of the Gaussian process may be free of g_0. This is the case for parametric testing where the parameter to be tested is in the interior of the parameter set, see section 3.1.
- Analytic derivations of the distributions of the supremum of the Gaussian process as involved in the Theorems are difficult problems. In a recent work, Azais and Wschebor (Azais and Wschebor (1995)) give an explicit formula for computing the distribution of the supremum of a random process in various situations. A recent text of introduction in the topics of continuity and extrema for Gaussian processes, together with references, is the one of Adler (Adler (1990)). Also, in similar contexts, Beran and Millar (Beran and Millar (1987)) have proposed stochastic procedures using bootstrapping to find the estimated level of confidence sets when the asymptotic distribution is too intractable. Similar ideas could be used here.
- Though the assumption (A4) does not hold for mixtures, we shall derive an asymptotic distribution for the maximum likelihood statistic which will be also some function of the maximum of the Gaussian process indexed by $\mathcal{D}$. Application to hypothesis testing follows obviously the same lines.

3.3. Application to parametric models

Let $\mathcal{G} = \{g_\gamma,\ \gamma \in \Gamma\}$ be an identifiable parametric model where Γ is a compact subset of $\mathbb{R}^p$. We make the following geometrical assumption on Γ:

(RP1) For all γ in Γ and u in $\mathbb{R}^p$ define:

$$T(\gamma, u) = \{t \in \mathbb{R},\ \gamma + t.u \in \Gamma\},$$

$$U(\gamma) = \{u \in \mathbb{R}^p,\ T(\gamma, u) \text{ contains a right-neighborhood of } 0\ , [0, u_\gamma[\}.$$

Then,

$$\forall \gamma \in \Gamma,\ \exists c > 0,\ \forall u,\ (\exists t \in T(\gamma, u))\ and\ t < c \Longrightarrow u \in U(\gamma)).$$

Moreover, for all γ in Γ, $U(\gamma)$ spans $\mathbb{R}^p$.

Let $g_0 = g_{\gamma^0}$. Assume the model is *locally regular* in the following way:

(RP2) The application $t \to g_{\gamma^0 + t.u}$ is twice continuously differentiable for $t > 0$, $g_0\nu$ a.e. for all directions u in $U(\gamma^0)$, with right continuous derivative at $t = 0$.

There exist functions h, l, m, such that:

$$\forall \gamma = \gamma^0 + t.u \in \Gamma, u \in U(\gamma^0),\ |\log g_\gamma| \le h,\ |\frac{1}{g_\gamma}\frac{\partial g_{\gamma^0 + t.u}}{\partial t}| \le l,$$

$$|\frac{1}{g_\gamma}\frac{\partial^2 g_{\gamma^0+t.u}}{\partial t^2}| \leq m, \;\; E_{g_{0\nu}}[h] < +\infty, \;\; E_{g_{0\nu}}[l^2] < +\infty, \;\; E_{g_{0\nu}}[m] < +\infty.$$

Define the Fisher information at point γ^0 as the $p \times p$ matrix $I(\gamma^0)$ such that:

$$\forall u \in U(\gamma^0), \; Var\left(\frac{1}{g_{\gamma^0}}\frac{\partial g_{\gamma^0+t.u}}{\partial t}|_{t=0}\right) = u^T \cdot I(\gamma^0) \cdot u.$$

(RP3) $I(\gamma^0)$ is non degenerated.

COMMENTS

- Assumption (RP1) allows to define the Fisher information unambiguously since derivatives exist in at least p linearly independent directions.
- For γ^0 in the interior of Γ, the Fisher information so defined reduces to the usual Fisher information, and assumptions (RP2) and (RP3) state the regularity of the model in the usual way (see Dacunha-Castelle and Duflo (1986)).
- The geometric interpretation of (RP1) is that Γ possesses the following property: at a boundary point, there exists a small ball B centered at the boundary point such that $\Gamma \cap B$ is inside the tangent cone.

3.3.1. TESTING $\gamma = \gamma^0$ AGAINST $\gamma \neq \gamma^0$. The locally conic parameterization will be:

$$\gamma = \gamma^0 + \theta \cdot \beta,$$

$$\theta = \sqrt{(\gamma - \gamma^0)^T I(\gamma^0)(\gamma - \gamma^0)},$$

$$\beta = \frac{\gamma - \gamma^0}{\theta}.$$

It is then obvious that all assumptions (A1),(A2), (A3), (AN), (AC), (A4) hold, with:

$$\mathcal{B} = \overline{\mathcal{B}} = \{\frac{\gamma - \gamma^0}{\sqrt{(\gamma - \gamma^0)^T I(\gamma^0)(\gamma - \gamma^0)}}, \gamma \in \Gamma\}.$$

Now, let $\beta_1, \ldots, \beta_p$ be p independent directions in $\mathcal{B}$. Then:

$$\mathcal{D} = \overline{\mathcal{D}} = \{d = \frac{1}{g_0}\sum_{i=1}^{p} b_i d_i, \;\; \text{where} \;\; d_i = \frac{\partial g_{\gamma^0+\theta.\beta_i}}{\partial \theta}|_{\theta=0}, \; \text{and} \; \sum_{i=1}^{p} b_i \beta_i \in \mathcal{B}\}.$$

Then (AD) holds since $\overline{\mathcal{D}}$ is a compact subset of a finite dimensional linear space, and (AN) holds by construction. Let $l_n(\gamma)$ be the log-likelihood with n i.i.d. observations. We have:

THEOREM 3.8. *Under the assumptions (RP1) to (RP3),* $\sup_{\gamma \in \Gamma}(l_n(\gamma) - l_n(0))$ *converges in distribution to:*

$$\frac{1}{2}\sup_{u \in I(\gamma^0)^{1/2}.\mathcal{B}} (\langle u, V\rangle)^2 \, 1_{\langle u,V\rangle \geq 0}$$

where V is a p dimensional standard Gaussian random variable and $\langle .,. \rangle$ the usual scalar product on $\mathbb{R}^p$.

Proof. Theorem 3.1 gives that $\sup_{\gamma\in\Gamma} l_n(\gamma) - l_n(0)$ converges in distribution to:

$$\frac{1}{2} \sup_{\sum_{i=1}^p b_i\beta_i \in \mathcal{B}} (\langle b, d\rangle)^2 \, 1_{\langle b,d\rangle \geq 0}$$

with $d = (d_1, \ldots, d_p)$. The distribution of $\langle b, d\rangle$ is that of $\langle \beta, W\rangle$ where $\beta \in \mathcal{B}$ and W follows centered p-dimensional Gaussian distribution with variance $I(\gamma^0)$. The change of variables $u = I(\gamma^0)^{1/2} \cdot \beta$ gives the result. □

Theorem 3.8 allows one to know the asymptotic distribution of the maximum likelihood statistic depending on the geometric structure of Γ around γ_0:

- If γ_0 is in the interior of Γ, $\mathcal{B}$ contains all possible directions, and also $I(\gamma^0)^{1/2} \cdot \mathcal{B}$, so that (as already known) we obtain that the asymptotic distribution of $\sup_{\gamma\in\Gamma} l_n(\gamma) - l_n(0)$ is $\frac{1}{2} \cdot \chi^2(p)$, since the supremum in the Theorem is attained for $u = V/\|V\|$.
- If γ_0 is on the boundary of Γ, the asymptotic distribution does depend on γ^0 only through the set $\mathcal{B}$, that is only through the shape of Γ at the boundary point γ^0. More precisely:

COROLLARY 3.9. *Under the assumptions (RP1) to (RP3),* $\sup_{\gamma\in\Gamma} l_n(\gamma) - l_n(0)$ *converges in distribution to:*

$$\frac{1}{2} \left(\langle V, P_{I(\gamma^0)^{1/2}\cdot\mathcal{B}}(V)\rangle \right)^2 1_{\langle V, P_{I(\gamma^0)^{1/2}\cdot\mathcal{B}}(V)\rangle \geq 0}$$

where $P_{I(\gamma^0)^{1/2}\cdot\mathcal{B}}$ *the orthogonal projection onto the set* $I(\gamma^0)^{1/2} \cdot \mathcal{B}$ *and* V *is a* p*-dimensional standard Gaussian random variable.*

3.3.2. TESTING WITH A FINITE DIMENSIONAL NUISANCE PARAMETER. We suppose here that we are interested only in a part of the parameter. Namely, $\gamma = (\alpha, \delta)$, $\alpha \in \mathbb{R}^k$, $\delta \in \mathbb{R}^l$, $k + l = p$, and we want to test $\alpha = \alpha^0$ against $\alpha \neq \alpha^0$. Say that $\gamma^0 = (\alpha^0, \delta^0)$. We have:

THEOREM 3.10. *Under the assumptions (RP1) to (RP3) for the whole model and if* γ^0 *lies in the interior of* Γ*:* $\sup_{\gamma\in\Gamma} l_n(\gamma) - \sup_{\delta,(\alpha^0,\delta)\in\Gamma} l_n(\alpha^0, \delta)$ *converges in distribution to* $\frac{1}{2} \cdot \chi^2(k)$.

Proof. For the full model, we have that $\mathcal{B}$ is an ellipsoid in $\mathbb{R}^p$. For the submodel, the locally conic parameterization will be:

$$\gamma = \gamma^0 + \theta_1 \cdot (0, \beta_1),$$

$$\theta_1 = \sqrt{(\delta - \delta^0)^T I_1(\delta^0)(\delta - \delta^0)},$$

$$\beta_1 = \frac{\delta - \delta^0}{\theta_1},$$

where $I_1(\delta^0)$ is defined at point δ^0 as in (RP2) for the submodel. Then, with obvious notations, $\mathcal{B}_1 = (0)_k \times \mathcal{E}_l$ where $(0)_k$ is the null point in $\mathbb{R}^k$ and $\mathcal{E}_l$ is the l-dimensional ellipsoid defined with the matrix $I_1(\delta^0)$. Following the same change of variables than for the proof of Theorem 3.8, we have that: $\sup_{\gamma\in\Gamma} l_n(\gamma) - \sup_{\delta,(\alpha^0,\delta)\in\Gamma} l_n(\alpha^0, \delta)$ converges in distribution to

$$\frac{1}{2} \sup_{u\in\mathcal{B}} (\langle u, V\rangle)^2 \, 1_{\langle u,V\rangle \geq 0} - \frac{1}{2} \sup_{u_1\in\mathcal{E}_l} (\langle((0)_k, u_1), V\rangle)^2 \, 1_{\langle((0)_k,u_1),V\rangle \geq 0}$$

where V is a p dimensional standard Gaussian random variable. This in turns exactly equals $\frac{1}{2}(V_1^2 + V_2^2 + \ldots + V_k^2)$. □

Though the simple mixture (4.1) is also an application of the general result of section 3, we chose to present the problem of testing the number of populations in a mixture in a separate section.

4. Population mixtures

In this section, we show how the theory of locally conic models applies to population mixtures.
Let $\mathcal{F} = (f_\gamma)_{\gamma\in\Gamma}$ be a family of probability densities with respect to ν. Γ is a compact subset of $\mathbb{R}^k$ for some integer k. $\mathcal{G}_p$ is the set of all $p-$mixtures of densities of $\mathcal{F}$:

$$\mathcal{G}_p = \{g_{\pi,\alpha} = \sum_{i=1}^{p} \pi_i \cdot f_{\gamma^i} \ : \ \pi = (\pi_1, \ldots, \pi_p),\ \alpha = (\gamma^1, \ldots, \gamma^p),$$
$$\forall i = 1, \ldots, p,\ \gamma^i \in \Gamma,\ 0 \le \pi_i \le 1,\ \sum_{i=1}^{p} \pi_i = 1\}.$$

Obviously, the model is not identifiable for the parameters $\pi = (\pi_1, \ldots, \pi_p)$ and $\alpha = (\gamma^1, \ldots, \gamma^p)$. There exist mixtures g in $\mathcal{G}_p$ which have different representations $g_{\pi,\alpha}$ with different parameters π and α. For instance, we have for any permutation σ of the set $\{1, \ldots, p\}$:

$$\sum_{i=1}^{p} \pi_i \cdot f_{\gamma^i} = \sum_{i=1}^{p} \pi_{\sigma(i)} \cdot f_{\gamma^{\sigma(i)}}.$$

Another example which may not be avoided by taking some quotient with respect to permutations is:

$$f_{\gamma^0} = \sum_{i=1}^{p} \pi_i \cdot f_{\gamma^0}$$

for any (π_i) such that $\pi_i \ge 0$ and $\sum_{i=1}^{p} \pi_i = 1$.
However, we assume that $\mathcal{G}_p$ is identifiable in the weak following sense:

$$g_{\pi^0,\alpha^0} = g_{\pi^1,\alpha^1} \quad \nu\ a.e. \iff \left| \begin{array}{r} \sum_{i=1}^{p} \pi_i^0 \cdot \delta_{\gamma_i^0} = \sum_{i=1}^{p} \pi_i^1 \cdot \delta_{\gamma_i^1} \text{ as probability} \\ \text{distributions on } \Gamma. \end{array} \right.$$

In other words, $\mathcal{G}_p$ is identifiable if the parameter is the mixing discrete probability distribution on Γ. Teicher (see Teicher (1965)) or Yakowitz and Spragins (Yakowitz and Spragins (1968)) give sufficient conditions for such weak identifiability, which hold for instance for finite mixtures of Gaussian or gamma distributions.
We address the following problems:

- For a particular density $f_{\gamma_0} = f_0$, test f_0 against a simple mixture in the model (4.1) stated in the introduction.
- For a particular density $f_{\gamma_0} = f_0$, test f_0 against a general mixture, or test one population against a mixture.
- For an integer q less than p, test q populations against p populations.

As noted before, the model is not identifiable for the parameters $\pi = (\pi_1, \ldots, \pi_p)$ and $\alpha = (\gamma^1, \ldots, \gamma^p)$. If reparameterized in an identifiable manner, lack of differentiability appears. When using the non identifiable parameterization with parameter (π, α), the lack of identifiability leads to a degeneracy of the Fisher information, so that, when using classical Taylor expansions for the log likelihood statistics, the remainder terms may not be bounded uniformly. Moreover, the asymptotic variance of the maximum likelihood estimator (which is the inverse of the Fisher information), when one of the parameters π or α is fixed is unbounded. This is why Ghosh and Sen had to separate strongly the γ parameters to develop the asymptotics of the maximum likelihood statistic (see Ghosh and Sen (1985)) when testing two populations against one population; that is, they assumed that the model for two populations verified $\|\gamma^1 - \gamma^2\| \geq \epsilon$ for a fixed positive ϵ and some norm $\|.\|$ on Γ. This assumption is rather unnatural.
For each mixture problem, we exhibit a locally conic parameterization that will solve the problem completely with no such separation on the parameters of the mixing family.
We make the following assumptions on the mixing family $\mathcal{F}$:

(M1) There exists a function h in $L_1(g_0\nu)$ such that: $\forall f \in \mathcal{F}$, $|\log f| \leq h$ ν-a.e.

4.1. Simple mixture

Here, the model is the subset of $\mathcal{G}_2$ given by:

$$g_{\pi,\gamma} = (1-\pi) f_{\gamma^0} + \pi f_\gamma \tag{4.1}$$

where $\pi \in [0,1]$, $\gamma \in \Gamma$ and the true density is $g_0 = f_{\gamma^0}$, γ^0 in the interior of Γ. Recall that H is the Hilbert space $L^2(g_0\nu)$. Define (θ, β) by:

$$\theta = \|\frac{g_{\pi,\gamma} - g_0}{g_0}\|_H = \pi \cdot \|\frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}}\|_H \ ; \ \ \beta = \gamma.$$

So that the new parameterization is given by:

$$g_{\pi,\gamma} = g_{(\theta,\beta)} = g_0 \left(1 + \theta \cdot \frac{f_\beta - f_{\gamma^0}}{f_{\gamma^0}} / \|\frac{f_\beta - f_{\gamma^0}}{f_{\gamma^0}}\|_H\right). \tag{4.2}$$

It is easy to see that here:

$$\mathcal{D} = \left\{ \frac{f_\beta - f_{\gamma^0}}{f_{\gamma^0}} \cdot \|\frac{f_\beta - f_{\gamma^0}}{f_{\gamma^0}}\|_H^{-1}, \beta \in \Gamma \right\}.$$

We make the following assumption:

(M2) f_γ is continuously differentiable ν almost everywhere with respect to $\gamma = (\gamma_1, \ldots, \gamma_k)$ in the interior of Γ. Moreover, there exists a function l such that:

$$\forall \gamma \in \Gamma, \ \ |\frac{1}{f_\gamma}\frac{\partial f_\gamma}{\partial \gamma_i}| \leq l, \ \ i = 1, \ldots, k \ \ E_{g_0\nu}[l^2] < +\infty.$$

Theorem 4.1. *Under assumptions (M1), (M2), and if*

(S2) $\mathcal{D}$ is a Donsker class and ξ_d has continuous sample paths

then the parameterization verifies all assumptions (A1), (A2), (A3), (AD), (AN), (AC), (A4), and Theorem 3.1 holds.

Proof. (M1) implies (AC), (M2) implies (A1), (A2), (A3) and (A4). In particular

$$g_{(\theta,\beta)} = g_0 \Longleftrightarrow \theta = 0$$

is obviously a consequence of the weak identifiability, since

$$\mathcal{T} \subset \{(\theta,\beta) \ : \ \theta \leq \| \frac{f_\beta - f_{\gamma^0}}{f_{\gamma^0}} \|_H \}.$$

Then, (AN) holds by construction. □

REMARK 4.2. A simple and useful example where the assumptions hold is the case of a mixture of Gaussian variables. If the mixture is only on the means, it is enough to assume that the means are in a compact set. If the mixture is also on the variances, easy computations show that the variances have to be restricted in a set of the form $[\epsilon, 2\sigma_0 - \epsilon]$, if σ_0 is the variance of g_0, for the assumption (AC) to hold. However, in this case, a close look at the expansions giving the form of the likelihood statistic shows that theorem 4.1 still holds even when the variance is allowed to take bigger values: the bigger values play a negligible role when taking the maximum. This will be fully developed in further work.

4.2. One population against two populations

In the case of the simple mixture, the locally conic parameterization is linear in the parameter θ. The Taylor expansion till order 2 is trivial, and the simple asymptotic result Theorem 3.1 holds. This will be the same for contamination models as explained in section 5. But this will no longer be the case for mixtures of unknown populations. Let us explain the situation in the most simple case of real parameters. Here, Γ will be a compact subset of $\mathbb{R}$. We suppose again that the underlying distribution is $g_0 = f_{\gamma^0}$, γ^0 in the interior of Γ. But the model is the whole $\mathcal{G}_2$. Define: $\beta = (\gamma, \delta)$ $\gamma \in \Gamma$, $\delta \in [0, M]$. A locally conic parameterization is given by:

$$g_{(\theta,\beta)} = \frac{\theta}{N(\beta)} f_\gamma + (1 - \frac{\theta}{N(\beta)}) f_{\gamma^0 + \frac{\theta}{N(\beta)}\delta}$$

where

$$N(\beta) = \| \delta \frac{1}{g_0} \frac{\partial f_\gamma}{\partial \gamma}|_{\gamma=\gamma^0} + \frac{f_\gamma - g_0}{g_0} \|_H .$$

If f_γ possesses sufficiently many derivatives with respect to γ, we have the following derivatives for $g_{\theta,\beta}$ ($g^{(k)}$ denotes the k-th derivative of $g_{(\theta,\beta)}$ with respect to θ and $f^{(k)}$ the k-th derivative of f_γ with respect to γ):

$$g'_{(\theta,\beta)} = \frac{\delta}{N(\beta)}(1 - \frac{\theta}{N(\beta)})(f'_{\gamma^0 + \frac{\theta}{N(\beta)}\delta}) + \frac{1}{N(\beta)}(f_\gamma - f_{\gamma^0 + \frac{\theta}{N(\beta)}\delta}),$$

$$g^{(k)}_{(\theta,\beta)} = -\frac{k\delta^{k-1}}{N(\beta)^k} f^{(k-1)}_{\gamma^0 + \frac{\theta}{N(\beta)}\delta} + \frac{\delta^k}{N(\beta)^k}(1 - \frac{\theta}{N(\beta)}) f^{(k)}_{\gamma^0 + \frac{\theta}{N(\beta)}\delta}.$$

Observe now that $N(\beta)$ goes to 0 as soon as γ goes to γ^0 and δ goes to 0. Then it can be seen that $\frac{\delta}{N(\beta)^2}$ can not be uniformly bounded, so that $g"_{(\theta,\beta)}$ divided by $g_{(\theta,\beta)}$ may not be uniformly dominated by an integrable function as required in (A4). To find the result, the locally conic parameterization is

still a key point, but the Taylor expansion has to be made till an order bigger than 2 in a region where $N(\beta)$ goes to 0. We shall need the assumptions:

(M3) $N(\beta) = 0$ if and only if $\gamma = \gamma^0$ and $\delta = 0$.

(M4) f_γ possesses derivatives till order 5. For all $k \leq 5$, $\frac{f^{(k)}_{\gamma^0}}{g_0} \in L^2(g_0\nu)$. Moreover, there exist functions m_2, m_5 and a positive ϵ such that:

$$\sup_{\gamma-\gamma^0\leq\epsilon} |\frac{f"_\gamma}{g_0}| \leq m_2 \quad E_{g_0\nu}[m_2^2] < +\infty,$$

$$\sup_{\gamma-\gamma^0\leq\epsilon} |\frac{f^{(5)}_\gamma}{g_0}| \leq m_5 \quad E_{g_0\nu}[m_5^2] < +\infty.$$

Define $\mathcal{D}$ as the set of functions d

$$d = \frac{1}{N(\beta)} \left(\frac{f_\gamma - f_{\gamma^0} + \delta f'_{\gamma^0}}{g_0} \right).$$

Define also

$$d_1 = \frac{f'_{\gamma^0}/g_0}{\|f'_{\gamma^0}/g_0\|_H},\ d_2 = \frac{f"_{\gamma^0}/g_0}{\|f"_{\gamma^0}/g_0\|_H},\ u = \langle d_1, d_2 \rangle.$$

Notice that d_1, d_2 are in $\overline{\mathcal{D}}$ as well as $(\lambda d_1 + \mu d_2)/\sqrt{\lambda^2+\mu^2+2u\lambda\mu}$. We may now state the Theorem:

THEOREM 4.3. *Assume that (M1), (M3), (M4) hold, that $\mathcal{D}$ is a Donsker class and that ξ_d has continuous sample paths. Then $T_n - l_n(0)$ converges in distribution to the following variable:*

$$\sup \left\{ \frac{1}{2} \cdot \sup_{d\in\mathcal{D}} (\xi_d)^2 \cdot 1_{\xi_d \geq 0}; \frac{1}{2}\xi^2_{d_1} + \frac{1}{2}\xi^2_{(d_2-ud_1)/\sqrt{1-u^2}} 1_{\xi_{(d_2-ud_1)/\sqrt{1-u^2}}>0} \right\}.$$

The asymptotic distribution is the supremum of two terms. The first one is the sup term which was expected, and which is obtained for parameters that do not approach too fast the non identifiable point. The second term comes from the boundary of the set $\mathcal{D}$, that is from approaching the non identifiable point. This second term has an unexpected form, since it seems to be twice than an ordinary term (it adds two terms), and appears as a boundary term coming from second order.

REMARK 4.4. If moreover f'_γ/g_0 is uniformly bounded in H, it is easily seen that

$$\| \frac{\frac{f_\gamma - f_{\gamma^0} + \delta f'_{\gamma^0}}{g_0}}{\|\frac{f_\gamma - f_{\gamma^0} + \delta f'_{\gamma^0}}{g_0}\|} - \frac{\frac{f_{\gamma'} - f_{\gamma^0} + \delta' f'_{\gamma^0}}{g_0}}{\|\frac{f_{\gamma'} - f_{\gamma^0} + \delta' f'_{\gamma^0}}{g_0}\|} \|$$

is upper bounded by

$$2 \frac{\|\frac{f_\gamma - f_{\gamma'} + (\delta-\delta') f'_{\gamma^0}}{g_0}\|}{\|\frac{f_\gamma - f_{\gamma^0} + \delta f'_{\gamma^0}}{g_0}\|}.$$

The number of covering balls in H is then easily seen to be of order $1/\epsilon^2$ when $N(\beta)$ does not approach zero, and of order $1/\epsilon^4$ when $N(\beta)$ approaches zero. The Donsker condition then holds.

4.3. One population against a mixture

Here, we suppose again that the underlying distribution is $g_0 = f_{\gamma^0}$, γ^0 in the interior of Γ. But the model is the whole $\mathcal{G}_p$ for some known integer p. Define:

$$\beta = (\lambda_1, \dots, \lambda_{p-1}, \gamma^1, \dots, \gamma^{p-1}, \delta), \quad \lambda_i \geq 0, \sum_{i=1}^{p-1} \lambda_i = 1,$$

$$\gamma^i \in \Gamma,\ i = 1, \dots, p-1 \text{ and } \delta \in \mathbb{R}.$$

The locally conic parameterization is given by:

$$g_{(\theta,\beta)} = \sum_{i=1}^{p-1} \lambda_i \frac{\theta}{N(\beta)} f_{\gamma^i} + (1 - \frac{\theta}{N(\beta)}) f_{\gamma^0 + \frac{\theta}{N(\beta)}\delta}.$$

$\mathcal{D}$ is the subset of the unit sphere of H of functions of the form:

$$\left(\delta \frac{1}{g_0} \frac{\partial f_\gamma}{\partial \gamma}|_{\gamma=\gamma^0} + \sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - g_0}{g_0} \right) \frac{1}{N(\beta)}$$

where $\lambda_i \geq 0$, $\gamma^i \in \Gamma$, $i = 1, \dots, p-1$, $\delta \in \mathbb{R}$, and

$$\sum_{i=1}^{p-1} \lambda_i = 1,$$

and

$$N(\beta) = \| \delta \frac{1}{g_0} \frac{\partial f_\gamma}{\partial \gamma}|_{\gamma=\gamma^0} + \sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - g_0}{g_0} \|_H.$$

The following assumption will replace (M3):

(M3*) $N(\beta) = 0$ if and only if $\sum_{i=1}^{p-1} \lambda_i (\gamma - \gamma^0)^2 = 0$ and $\delta = 0$.

For non negative $\lambda_1, \dots, \lambda_{p-1}$, any λ and $\epsilon = 0$ or 1, if $\Lambda = (\lambda_1, \dots, \lambda_{p-1}, \lambda, \epsilon)$, define

$$d(\Lambda) = \frac{\sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - f_{\gamma^0}}{g_0} + \lambda d_1 + \epsilon d_2}{\| \sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - f_{\gamma^0}}{g_0} + \lambda d_1 + \epsilon d_2 \|_H}$$

and define $\mathcal{D}_1$ to be the subset of $\overline{\mathcal{D}}$ of functions $d(\Lambda)$ which are orthogonal to d_1. Then

Theorem 4.5. *Under (M1), (M3*) and (M4), if $\mathcal{D}$ is a functional Donsker class and ξ_d has continuous sample paths, then $T_n - l_n(0)$ converges in distribution to the following variable:*

$$\sup \left\{ \frac{1}{2} \cdot \sup_{d \in \mathcal{D}} (\xi_d)^2 \cdot 1_{\xi_d \geq 0}; \frac{1}{2} \xi_{d_1}^2 + \sup_{d \in \mathcal{D}_1} \frac{1}{2} \xi_d^2 \cdot 1_{\xi_d \geq 0} \right\}.$$

Theorem 4.5 is obviously an extension of Theorem 4.3 since in case $p = 2$, $\mathcal{D}_1$ contains only one direction.

The addressed problem of testing one population (known or unknown) against a p-mixture can now clearly be solved using Theorem 3.1 when the population is known, and using Theorem 3.6 together with Theorem 3.8 when the population is unknown.

4.3.1. q POPULATIONS AGAINST p POPULATIONS. In the first version of the paper, we claimed that we believed that the asymptotic distribution of the maximum likelihood statistic in the general model could be derived using the following locally conic parameterization. This claim has further be proved in subsequent work, see Dacunha-Castelle and Gassiat (1996).
Define $\mathcal{B}_0$ the set of parameters $\beta = (\lambda_1, \dots, \lambda_{p-q}, \gamma^1, \dots, \gamma^{p-q}, \delta^1, \dots, \delta^q, \rho_1, \dots, \rho_q)$ such that $\lambda_i \geq 0$, $\gamma^i \in \Gamma$, $i = 1, \dots, p-q$, $\delta^l \in \mathbb{R}^k$, $\rho_l \in \mathbb{R}$, $l = 1, \dots, q$, and $\sum_{i=1}^{p-q} \lambda_i + \sum_{l=1}^{q} \rho_l = 0$. Let then

$$N(\beta) = \| \sum_{l=1}^{q} \sum_{i=1}^{k} \pi_l^0 \delta_i^l \frac{1}{g_0} \frac{\partial f_\gamma}{\partial \gamma_i}|_{\gamma=\gamma^{l,0}} + \sum_{i=1}^{p-q} \lambda_i \frac{f_{\gamma^i}}{g_0} + \sum_{l=1}^{q} \rho_l \frac{f_{\gamma^{l,0}}}{g_0} \|_H .$$

For any β in $\mathcal{B}_0$ and any non negative θ such that for any integer $l = 1, \dots, q$, $\pi_l^0 + \rho_l \frac{\theta}{N(\beta)} \geq 0$, define the mixture:

$$g_{(\theta,\beta)} = \sum_{i=1}^{p-q} \lambda_i \frac{\theta}{N(\beta)} f_{\gamma^i} + \sum_{l=1}^{q} (\pi_l^0 + \rho_l \frac{\theta}{N(\beta)}) f_{\gamma^{l,0} + \frac{\theta}{N(\beta)} \delta^l} . \tag{4.3}$$

Such parameterization may be viewed as a perturbation of g_0 in the following way: perturb the q mixture g_0 through a perturbation of the parameters $\gamma^{l,0}$ and the weights π_l^0, and add a perturbation as a $p-q$-mixture with weight tending to 0. Such equation does not completely set a locally conic parameterization. Indeed, the equation (4.3) does not define unambiguously (θ, β) for a given mixture. For instance, different sets of parameters may give g_0. It is then important to define the set $\mathcal{B}$ such that $g_{(\theta,\beta)} = g_0 \Longleftrightarrow \theta = 0$, which is not an immediate consequence of the definition of $g_{(\theta,\beta)}$. We shall then precisely describe the set $\mathcal{B}$. The asymptotic distribution of the likelihood ratio will take a similar form than when testing 1 against p populations.
THE LOCALLY CONIC PARAMETERIZATION. Let g be any p-mixture:

$$g = \sum_{i=1}^{p} \pi_i \cdot f_{\gamma^i} .$$

To describe it through equation (4.3), one has to associate the parameters of g to those of g_0, that is to give a special order to the parameters. In other words: for any permutation σ of the set $\{1, \dots, p\}$, we define the parameters θ_σ such that $g_{(\theta_\sigma, \beta_\sigma)} = g$. This leads to:

$$\beta_\sigma = (\lambda_{1,\sigma}, \dots, \lambda_{p-q,\sigma}, \gamma^{1,\sigma}, \dots, \gamma^{p-q,\sigma}, \delta^{1,\sigma}, \dots, \delta^{q,\sigma}, \rho_{1,\sigma}, \dots, \rho_{q,\sigma})$$

with:

$$\forall i = 1, \dots, p-q, \quad \lambda_{i,\sigma} \cdot \theta_\sigma = \pi_{\sigma(i)} \cdot N(\beta_\sigma),$$

$$\forall i = 1, \dots, p-q, \quad \gamma^{i,\sigma} = \gamma^{\sigma(i)},$$

$$\forall i = 1, \dots, q, \quad \delta^{i,\sigma} \cdot \theta_\sigma = (\gamma^{\sigma(p-q+i)} - \gamma^{i,0}) \cdot N(\beta_\sigma),$$

$$\forall i = 1, \dots, q, \quad \rho_{i,\sigma} \cdot \theta_\sigma = (\pi_{\sigma(p-q+i)} - \pi_i^0) \cdot N(\beta_\sigma).$$

It is easily seen that

$$\theta_\sigma = \| \sum_{l=1}^{q} \sum_{i=1}^{k} (\gamma_i^{\sigma(p-q+l)} - \gamma_i^{l,0}) \frac{1}{g_0} \frac{\partial f_\gamma}{\partial \gamma_i}|_{\gamma=\gamma^{l,0}} + \sum_{i=1}^{p-q} \pi_{\sigma(i)} \frac{f_{\gamma^{\sigma(i)}}}{g_0} + \sum_{l=1}^{q} (\pi_{\sigma(p-q+l)} - \pi_l^0) \frac{f_{\gamma^{l,0}}}{g_0} \|_H.$$

The system is then ambiguous on the scale of β_σ since a multiplication by a scalar of β leads to the same result to $N(\beta)$.
The problem is then to choose between the permutations. The following choice is a good one. The idea is to associate step by step the nearest points γ^i involved in g to the set of points $\gamma^{l,0}$ involved in g_0. Look for:

$$\min_{l=1,\dots,q, i=1,\dots,p} \|\gamma^{l,0} - \gamma^i\|.$$

It is attained for l_1 and i_1. Define then $\sigma(p-q+l_1) = i_1$. Look then for

$$\min_{l=1,\dots,q, l\neq l_1, i=1,\dots,p, i\neq i_1} \|\gamma^{l,0} - \gamma^i\|.$$

It is attained for l_2 and i_2. Set then $\sigma(p-q+l_2) = i_2$. By induction, define in the same way $\sigma(p-q+l_j) = i_j$ for $j = 1, \dots, q$. In this algorithm, consider only points truly involved in g (eventually less than p points). Then complete the permutation σ in some ordered way. You then have defined a permutation $\sigma(g)$. The locally parameterization is then given by equation (4.3) with:

$$\mathcal{T} = \{(\theta, \beta_{\sigma(g)}) \ : \ \theta \leq \theta_{\sigma(g)}, \ \ g \in \mathcal{G}\}.$$

This induces the set $\mathcal{B}$ as the intersection of all directions approaching 0 in $\mathcal{T}$. Such parameterization is locally conic.

5. Possible extensions

We briefly show how the theory of locally conic models could be used in two other situations, leaving complete exposition and details for further investigation.

5.1. Contamination or perturbation models

Let g_0 be fixed. Suppose we want to test: $H_0 : \{g_0\}$ against the perturbation $H_1 : \{g_{(\theta,\beta)} = g_0 + \theta\beta\}$ (or against the contamination model $\{g_{(\theta,\beta)} = (1-\epsilon)g_0 + \epsilon g_1\}$ which is similar to the perturbation model).
Assume $\theta \in [0,1]$, $\beta \in \mathcal{B}$ where $\mathcal{B}$ is a subset of the unit ball of $L_2(1/g_0.\nu)$ such that all β in $\mathcal{B}$ verify $\int \beta d\nu = 0$. We then have here:

$$\mathcal{D} = \frac{1}{g_0}\mathcal{B}.$$

Assume there exist Banach spaces C_1 and C_2 with canonical injections $C_1 \to C_2 \to L^2(1/g_0.\nu)$ and real numbers K_1 and K_2 such that B is a subset of C_1 with:

$$\forall \beta \in \mathcal{B}, \ \|\beta\|_{C_1} \leq K_1, \ \|\beta\|_{C_2} \leq K_2.$$

$\mathcal{B}$ is equipped with the topology induced by C_2. Then, we may apply the theory of locally conic models as soon as:

- The image in C_2 of the unit ball of C_1 is compact in C_2,
- The continuity of the linear forms $\beta \to \beta(x)$ follows from the condition $|\beta_1(x) - \beta_2(x)| \leq \|\beta_1 - \beta_2\|_{C_2}$,
- $\frac{1}{g_0}\mathcal{B}$ is a Donsker class.

A simple example is the following: $C_1 = H^4$, $C_2 = H^2$ where H^p is the Sobolev space of functions with p derivatives, equipped with the norm $\sum_{j=0}^{p} \|f^{(j)}\|_2$. Then if we choose as perturbation set

$$\mathcal{B} = \{\beta \in H^4 \cup H^2 \cup L^2(1/g_0.\nu),\ \|\beta\|_{L^2(1/g_0.\nu)} = 1,$$
$$\|\beta\|_{H^4} \leq K_1,\ \|\beta\|_{H^2} \leq K_2,\ \beta(0) = 0\}.$$

We then have:

$$\begin{aligned} |\beta_1(x) - \beta_2(x)| &= |\int_0^x (\int_0^u [\beta"_1(v) - \beta"_2(v)]dv)| \\ &\leq \|\beta_1 - \beta_2\|_{H^2}. \end{aligned}$$

In such situations, Theorem 3.3 holds, so that:

$$\|\hat{g} - g_0\| \text{ converges to 0 at speed } 1/\sqrt{n}$$

where $\hat{g}$ is the maximum likelihood of g in the perturbation model. In other words, the norm of the density may be estimated at rate $1/\sqrt{n}$ in such non parametric model. The estimation of non linear functionals of a density in non parametric models is a widely studied problem with known results and still open questions. It is already known that some non linear functionals of a density may be estimated at rate $1/\sqrt{n}$ in non parametric settings, see for instance Donoho (Donoho (1988)). It is however not in the scope of this paper to discuss this subject. Let us only notice that it is also known that maximum likelihood estimators and functional Donsker class theory do not lead to the optimal results for some critical non parametric situations, compare for instance with the results of Laurent (Laurent (1993)).

5.2. ARMA MODELS

Let $(\epsilon_n)_{n\in\mathbb{N}}$ be a sequence of independent centered Gaussian random variables with common variance σ^2. An ARMA(p,q) process $(X_n)_{n\in\mathbb{N}}$ is given by the following equation (see for instance Azencott and Dacunha-Castelle (1984)):

$$X_n + a_1 X_{n-1} + \ldots + a_p X_{n-p} = \epsilon_n + b_1 \epsilon_{n-1} + \ldots + b_q \epsilon_{n-q}$$

where $a_1, \ldots, a_p, b_1, \ldots, b_q$ are real parameters.
Let $X = (X_n)_{n\in\mathbb{N}}$ be a given process, and suppose we have to test that X is an ARMA(p_0, q_0) process against X is an ARMA(p, q) process. As for the mixture model, the ARMA(p,q) model is non identifiable when using parameters $a_1, \ldots, a_p, b_1, \ldots, b_q$. For example an i.i.d. sequence has all 0 parameters, and also any equal parameters $a_1 = b_1, \ldots, a_k = b_k$, $k \leq p$ and $k \leq q$, the other parameters being set to 0. We shall prove in a forthcoming paper (Dacunha-Castelle and Gassiat (1996)) that it is possible to define a locally conic parameterization to deduce the asymptotic behavior of the maximum pseudo-likelihood statistic for the case of Gaussian processes, or of the minimum contrast statistic for general second order processes.

This leads to a simpler presentation than in Hannan (1982). This new presentation also makes clearer the reason why the asymptotic limit distribution is the supremum of a function of a Gaussian process over some space.

6. Proofs

Proof of Theorem 2.1. $K(g_0, g_{(\theta,\beta)})$, the Kullback information, is continuous with respect to the parameter (θ, β), thanks to (A1), (A2) and (AC). Define:

$$k(\theta) = \inf_{\beta \in \mathcal{B} \,:\, (\theta,\beta) \in \mathcal{T}} K(g_0, g_{(\theta,\beta)}).$$

Since $\overline{\mathcal{T}}$ is a compact set and using assumption (A1) we have that:

$$\forall \theta > 0,\ k(\theta) > 0.$$

Moreover, k is continuous. Define now:

$$U_n(\theta) = \sup_{\beta \in \mathcal{B} \,:\, (\theta,\beta) \in \overline{\mathcal{T}}} \frac{1}{n} \left(l_n(\theta, \beta) - l_n(0, \beta) \right).$$

First of all, we obviously have:

$$\liminf_{n \to +\infty} U_n(\theta) \geq -k(\theta) \text{ a.s.}$$

Define $m_\eta(x, \theta) = \sup_{d(\beta_1,\beta_2) \leq \eta} |\log g_{(\theta,\beta_1)}(x) - \log g_{(\theta,\beta_2)}(x)|$. Since $\mathcal{B}$ is precompact, for any positive η there exists a finite number N_η of balls with diameter $\eta/2$ covering $\mathcal{B}$ and with centers β_i, $i = 1, \ldots, N_\eta$. Now, obviously

$$U_n(\theta) \leq \sup_{i=1,\ldots,N_\eta} \frac{1}{n} \left(l_n(\theta, \beta_i) - l_n(0, \beta) \right) + \frac{1}{n} \sum_{i=1}^{n} m_\eta(X_i, \theta)$$

so that a.s.

$$\limsup_{n \to +\infty} U_n(\theta) \leq \sup_{i=1,\ldots,N_\eta} \left(-K(g_0, g_{(\theta,\beta_i)}) \right) + E_{g_0.\nu}(m_\eta(X, \theta))$$

so that

$$\limsup_{n \to +\infty} U_n(\theta) \leq -k(\theta) + E_{g_0.\nu}(m_\eta(X, \theta)).$$

Now, we have $\lim_{\eta \to 0} m_\eta(x, \theta) = 0$ a.s., and (AC) implies

$$\lim_{\eta \to 0} E_{g_0.\nu}(m_\eta(X, \theta)) = 0$$

so that

$$\limsup_{n \to +\infty} U_n(\theta) \leq -k(\theta)$$

and we may conclude that $U_n(\theta)$ converges a.s. to $-k(\theta)$ for all θ.
Now, $\hat{\theta}$ is a maximizer of $U_n(\theta)$. Let δ be a positive real number. We have, since k is continuous:

$$\exists \epsilon > 0,\ \forall \theta > \delta,\ k(\theta) \geq 2\epsilon.$$

Let η be a positive real number, and let $(\theta_i)_{i=1,\ldots,N}$ be N real numbers such that $\theta_i = \theta_{i-1} + \eta$, $\theta_1 = \delta + \eta$, $\theta_N \geq M$. We have:

$$U_n(\theta) = U_n(\theta_i) + U_n(\theta) - U_n(\theta_i).$$

We have:

$$\begin{aligned} P(\hat{\theta} \geq \delta) &\leq P(\sup_{\theta \geq \delta} U_n(\theta) > 0) \\ &\leq P(w_n(\eta) > \epsilon) + P([\inf_{i=1,\dots,N} U_n(\theta_i)] \geq -\epsilon) \end{aligned}$$

where

$$w_n(\eta) = \sup_{|\theta - \theta'| \leq \eta} |U_n(\theta) - U_n(\theta')|.$$

Now, $\inf_{i=1,\dots,N} U_n(\theta_i)$ converges a.s. to $\inf_{i=1,\dots,N} -k(\theta_i)$ which is less than -2ϵ so that $P([\inf_{i=1,\dots,N} U_n(\theta_i)] \geq -\epsilon)$ tends to 0 as n tends to infinity. It only remains to show that $P(w_n(\eta) > \epsilon)$ tends also to 0 for a good choice of η. To do this, notice that: if $r_\eta(x) = \sup_{|\theta-\theta'|\leq\eta} \sup_{\beta \in \mathcal{B}} |\log g_{(\theta,\beta)}(x) - \log g_{(\theta',\beta)}(x)|$ we have:

$$\lim_{\eta \to 0} E_{g_0.\nu}(r_\eta(X)) = 0$$

thanks to assumptions (A1) and (AC). Now:

$$w_n(\eta) \leq \frac{1}{n} \sum_{i=1}^{n} r_\eta(X_i).$$

So that almost surely:

$$\limsup_{n \to +\infty} w_n(\eta) \leq E_{g_0.\nu}(r_\eta(X))$$

which is smaller than ϵ for small enough η.

Proof of Theorem 3.1. An obvious consequence of assumption (AC) together with (A1) and (A2) is that, if a submodel is fixed by the parameter β, the estimator of maximum likelihood $\hat{\theta}_\beta$ converges to 0 as n tends to infinity. Moreover:

Lemma 6.1. *Under assumptions (A1), (A2), (AC), $\hat{\theta}_\beta$ converges to 0 in probability uniformly in the parameter β.*

Proof. We have:

$$\{\sup_\beta \hat{\theta}_\beta > \delta\} \subset \{\sup_\beta \sup_{\theta \geq \delta} (l_n(\theta, \beta) - l_n(0, \beta)) > 0\}$$

so that:

$$P(\sup_\beta \hat{\theta}_\beta > \delta) \leq P(\sup_{\theta \leq \delta} U_n(\theta) > 0)$$

and the end of the proof is the same as that of Theorem 2.1. □

Let $V_n(\beta)$ be the log likelihood ratio statistic in the submodel: $V_n(\beta) = l_n(\hat{\theta}_\beta, \beta)$. We have:

$$T_n = \sup_{\beta \in \mathcal{B}} V_n(\beta).$$

Now: Assumption (A2) implies that, if $\hat{\theta}_\beta > 0$, the derivative of $l_n(\theta, \beta)$ with respect to θ is zero at the point $\hat{\theta}_\beta$: On $\{\hat{\theta}_\beta > 0\}$,

$$\sum_{i=1}^{n} \frac{g'_{(\hat{\theta}_\beta,\beta)}}{g_{(\hat{\theta}_\beta,\beta)}}(X_i) = 0.$$

Expanding this equation leads to:

$$0 = \sum_{i=1}^{n} \frac{g'_{(0,\beta)}}{g_{(0,\beta)}}(X_i) - \hat{\theta}_\beta \cdot \sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_{(0,\beta)}}\right)^2 (X_i)(1+R_n)$$

where

$$R_n = \int_0^1 \frac{Z_n(t\hat{\theta}_\beta, \beta)}{A_n(\beta)} dt,$$

with

$$Z_n(u,\beta) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{g'_{(u,\beta)}}{g_{(u,\beta)}}\right)^2 (X_i) - A_n(\beta) - \frac{1}{n}\sum_{i=1}^{n} \left(\frac{g"_{(u,\beta)}}{g_{(u,\beta)}}\right)(X_i)$$

and

$$A_n(\beta) = \frac{1}{n}\sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_{(0,\beta)}}\right)^2 (X_i).$$

Now, define:

$$Z(u,\beta) = \int \left((\frac{g'_{(u,\beta)}}{g_{(u,\beta)}})^2 g_0 - (\frac{g'_{(0,\beta)}}{g_0})^2 g_0 - \frac{g"_{(u,\beta)}}{g_{(u,\beta)}} g_0 \right) d\nu.$$

Using the same tricks as for Theorem 2.1, we have:

$$\lim_{\delta\to 0} \sup_{\beta} \sup_{|u|\le\delta} |Z(u,\beta)| = 0$$

and then

$$\lim_{\delta\to 0} \limsup_{n\to+\infty} \sup_{\beta} \sup_{|u|\le\delta} |Z_n(u,\beta)| = 0 \text{ in probability.}$$

An immediate consequence of this result together with Theorem 2.1 and (A5) is that $R_n = o_P(1)$ in Probability uniformly over β. We may then state:

LEMMA 6.2. *The following equation holds:*

$$\hat{\theta}_\beta = \frac{\sum_{i=1}^{n} \frac{g'_{(0,\beta)}}{g_0}(X_i)}{\sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_0}\right)^2 (X_i)} \cdot (1+o_P(1)) \cdot 1_{\sum_{i=1}^{n} \frac{g'_{(0,\beta)}}{g_0}(X_i) > 0}$$

where the $o_P(.)$ holds in probability uniformly over β.

Expansion of the logarithm in $V_n(\beta)$ and similar arguments lead to:

$$\begin{aligned} V_n(\beta) - l_n(0) &= \sum_{i=1}^{n} \frac{g_{(\hat{\theta}_\beta,\beta)} - g_0}{g_0}(X_i) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n} \left(\frac{g_{(\hat{\theta}_\beta,\beta)} - g_0}{g_0}\right)^2 (X_i)(1+o_P(1)) \\ &= \hat{\theta}_\beta \sum_{i=1}^{n} \frac{g'_{(0,\beta)}}{g_0}(X_i)(1+o_P(1)) \\ &\quad -\frac{(\hat{\theta}_\beta)^2}{2}\sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_0}\right)^2 (X_i)(1+o_P(1)). \end{aligned}$$

We may then state:

LEMMA 6.3.

$$V_n(\beta) - l_n(0) = \frac{1}{2} \frac{\left(\sum_{i=1}^n \frac{g'_{(0,\beta)}}{g_0}(X_i) \right)^2}{\sum_{i=1}^n \left(\frac{g'_{(0,\beta)}}{g_0} \right)^2 (X_i)} \cdot (1 + o_P(1)) \cdot 1_{\sum_{i=1}^n \frac{g'_{(0,\beta)}}{g_0}(X_i) > 0}$$

where the $o_P(.)$ holds uniformly over β.

Theorem 3.1 is then an immediate consequence of the previous lemma and assumptions (AD) and (A5).

PROOF OF THEOREM 3.3. Previous results lead to:

$$T_n - l_n(0) = \frac{1}{2} (\hat{\theta})^2 \cdot \left(\sum_{i=1}^n (\frac{g'_{0,\hat{\beta}}}{g_{0,\hat{\beta}}})^2 (X_i) \right) (1 + o_P(1))$$

where the $o_P(1)$ holds uniformly over β, so that:

$$\sqrt{n}\hat{\theta} = \sqrt{2(T_n - l_n(0))} \cdot \sqrt{\frac{1}{n} \sum_{i=1}^n (\frac{g'_{0,\hat{\beta}}}{g_{0,\hat{\beta}}})^2 (X_i)} \cdot (1 + o_P(1)). \tag{6.1}$$

Now:

LEMMA 6.4.

$$\frac{1}{n} \sum_{i=1}^n (\frac{g'_{0,\hat{\beta}}}{g_{0,\hat{\beta}}})^2 (X_i)$$

converges to 1 in $g_0.\nu$ probability as n tends to infinity.

Proof. Since a Donsker class is Glivenko-Cantelli in probability, we have:

$$\lim_{n \to +\infty} \sup_{d \in \mathcal{D}} | \frac{1}{n} \sum_{i=1}^n d^2(X_i) - \|d\|_H^2 | = 0 \ \ in\ g_0.\nu\ probability.$$

Now, using assumption (A5) we have

$$\lim_{n \to +\infty} \sup_{d \in \mathcal{D}} | \frac{1}{n} \sum_{i=1}^n d^2(X_i) - 1 | = 0 \ \ in\ g_0.\nu\ probability.$$

Moreover, denoting $\frac{g'_{0,\hat{\beta}}}{g_{0,\hat{\beta}}}$ by $\hat{d}$ we have:

$$| \frac{1}{n} \sum_{i=1}^n (\hat{d})^2 (X_i) - 1 | \leq \sup_{d \in \mathcal{D}} | \frac{1}{n} \sum_{i=1}^n d^2(X_i) - 1 |$$

and the lemma follows. □

Now, equation (6.1) and lemma 6.4 prove Theorem 3.3.

PROOF OF THEOREM 3.4. First of all, $(g_{(\theta_n,\beta_0)} \cdot \nu)^{\otimes n}$ and $(g_0 \cdot \nu)^{\otimes n}$ are contiguous. Indeed the log-likelihood ratio is:

$$\begin{aligned}\Lambda_n &= \sum_{i=1}^{n} \log \frac{g_{(\theta_n,\beta_0)}}{g_0}(X_i) \\ &= \frac{c}{n}\sum_{i=1}^{n} \frac{g'_{(0,\beta_0)}}{g_0}(X_i) + \frac{c^2}{2n}\sum_{i=1}^{n}\left(\frac{g"_{(0,\beta_0)}}{g_0} - \left(\frac{g'_{(0.\beta_0)}}{g_0}\right)^2\right)(X_i)(1+o_P(1))\end{aligned}$$

which converges in distribution under $g_0 \cdot \nu$ to the Gaussian distribution $\mathcal{N}(-c^2/2, c^2)$. This proves the contiguity, see Roussas (1970) Proposition 3.1 p. 11. This implies, see Roussas (1970) p.7:

- $\hat{\theta}$ converges to 0 in $g_{(\theta_n,\beta_0)} \cdot \nu$ probability,
- For all β in B^c, $\hat{\theta}_\beta$ converges to 0 in $g_{(\theta_n,\beta_0)} \cdot \nu$ probability,

so that lemma 6.3 stays true under $g_{(\theta_n,\beta_0)} \cdot \nu$. Now, definition 2.1 p.7 of Roussas (1970) again implies that:

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g'_{(0,\beta)}}{g_0}\right)^2(X_i)$$

converges to 1 in probability under $g_{(\theta_n,\beta_0)} \cdot \nu$. Moreover, applying Theorem 7.1 p. 33 of Roussas (1970) we see that:

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g'_{(0,\beta)}}{g_0}(X_i)$$

converges in distribution under $g_{(\theta_n,\beta_0)} \cdot \nu$ to the Gaussian distribution with mean $c\langle \frac{g'_{(0,\beta)}}{g_0}, \frac{g'_{(0,\beta_0)}}{g_0}\rangle$ and variance 1, and Theorem 3.4 follows.

PROOF OF THEOREM 3.5. The proof of the theorem follows the same lines as that of Theorem 3.4, except that we use Le Cam's third Lemma for metric spaces (see Van der Vaart and Wellner (1996) p. 404), which gives that

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{g'_{(0,\beta)}}{g_0}\right)^2(X_i)$$

converges to 1 uniformly (over $\mathcal{B}$) in probability under $g_{(\theta_n,\beta_0)} \cdot \nu$, and that the process

$$\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\frac{g'_{(0,\beta)}}{g_0}(X_i)\right)_{\beta\in\mathcal{B}}$$

converges in distribution under $g_{(\theta_n,\beta_0)} \cdot \nu$ to the Gaussian process with mean function $c\langle \frac{g'_{(0,\beta)}}{g_0}, \frac{g'_{(0,\beta_0)}}{g_0}\rangle$ and covariance the scalar product in H. Theorem 3.5 follows.

PROOF OF THEOREM 4.3. Define $\eta_n = \sup_\beta \hat{\theta}_\beta$, which is known to tend to 0 in probability. The proof relies on separating the domain into two regions:

$$A_n = \{\beta \ : \ \frac{\delta}{N(\beta)^2} \leq \frac{1}{\eta_n^\alpha}\}$$

$$B_n = \{\beta \ : \ \frac{\delta}{N(\beta)^2} \geq \frac{1}{\eta_n^\alpha}\}$$

for a suitable choice of α. Then we have:

$$T_n - l_n(0) = \sup \left\{ \sup_{\beta \in A_n} l_n(\hat{\theta}_\beta, \beta) - l_n(0), \sup_{\beta \in B_n} l_n(\hat{\theta}_\beta, \beta) - l_n(0) \right\}.$$

Then we prove:

LEMMA 6.5. *Under the assumptions of Theorem 4.3,* $\sup_{\beta \in A_n} l_n(\hat{\theta}_\beta, \beta) - l_n(0)$ *converges in distribution to*

$$\frac{1}{2} \sup_{d \in \mathcal{D}} (\xi_d)^2 \cdot 1_{\xi_d} \geq 0.$$

and

LEMMA 6.6. *Under the assumptions of Theorem 4.3,* $\sup_{\beta \in B_n} l_n(\hat{\theta}_\beta, \beta)$ *converges in distribution to*

$$\frac{1}{2}\xi_{d_1}^2 + \frac{1}{2}\xi^2_{(d_2 - u d_1)/\sqrt{1+u^2}} 1_{\xi_{(d_2 - u d_1)/\sqrt{1+u^2}} > 0\}}.$$

The following lemma will be a basic tool.

LEMMA 6.7. *Let* $\phi = \frac{\delta}{N(\beta)^2}$ *Then*

$$\delta \leq \frac{A}{\phi^{1/3}}\ ,\ N(\beta) \leq \frac{B}{\phi^{2/3}}$$

where A and B are some fixed constants.

The lemma says that, when $N(\beta)$ goes to 0, its speed and that of δ may be controlled via ϕ.

Proof. It is enough to prove that

$$\frac{\delta^2}{N(\beta)}$$

is uniformly bounded. Indeed, if not, let β_n be a sequence such that

$$\lim_n \frac{\delta_n^2}{N(\beta_n)} = +\infty.$$

Then, using (M3), δ_n tends to 0 and γ_n tends to γ^0. Letting

$$\frac{f"_{\gamma^0}}{g_0} = a \frac{f'_{\gamma^0}}{g_0} + t, \quad t \neq 0,$$

be an orthogonal decomposition in H, we have:

$$\begin{aligned} N(\beta_n)^2 \quad = \quad & \Big((\gamma_n - \gamma^0 + \delta_n + a/2.(\gamma_n - \gamma^0)^2)^2 \cdot \| \frac{f'_{\gamma^0}}{g_0} \|_H^2 \\ & + a^2/4.((\gamma_n - \gamma^0)^4 \|t\|_H^2 \Big) (1 + o(1)). \end{aligned}$$

Let $\gamma_n - \gamma^0 + \delta_n = a_n(\gamma_n - \gamma^0)^2$. Then:

$$\frac{\delta_n^2}{N(\beta_n)} = \frac{(a_n(\gamma_n - \gamma^0) - 1)^2}{\sqrt{(a_n + a/2)^2 \| \frac{f'_{\gamma^0}}{g_0} \|_H^2 + a^2/4. \|t\|_H^2}}$$

which is always bounded. □

REMARK 6.8. Notice that the only constraint on the parameters is given by

$$\frac{\theta}{N(\beta)} \leq 1.$$

In particular the speed of ϕ is unconstrained, this will be useful when optimizing the approximating polynomial for proving Lemma 6.6.

PROOF OF LEMMA 6.5. First, the following expansion holds for θ tending to 0:

$$\begin{aligned} l_n(\theta,\beta) - l_n(0) &= \sum_{i=1}^{n} \frac{g_{(\theta,\beta)} - g_0}{g_0}(X_i) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n} \left(\frac{g_{(\theta,\beta)} - g_0}{g_0}\right)^2 (X_i) \left(1 + O\left(\frac{g_{(\theta,\beta)} - g_0}{g_0}(X_i)\right)\right). \end{aligned} \tag{6.2}$$

Let us now write an expansion of $g_{(\theta,\beta)}$ till order 2:

$$g_{(\theta,\beta)}(x) = g_0(x) + \theta \cdot g'_{(0,\beta)}(x) + \frac{\theta^2}{2} \cdot g"_{(\theta*,\beta)}(x)$$

for a $\theta* \leq \theta$ and depending on x. Now as θ tends to 0:

$$g"_{(\theta*,\beta)}(x) = -2\frac{\delta}{N(\beta)^2} f'_{\gamma^0}(x) + 0(\frac{\delta^2}{N(\beta)^3}\theta m_2(x) g_0(x))$$

since δ is bounded and using (M4). Write:

$$D_n(\beta) = \sum_{i=1}^{n} \frac{g'_{(0,\beta)}}{g_0}(X_i),$$

$$F_n = \sum_{i=1}^{n} \frac{f'_{\gamma^0}}{g_0}(X_i).$$

Define also

$$a^2 = \|\frac{f'_{\gamma^0}}{g_0}\|_H^2$$

and

$$u(\beta) = \langle d_1, \frac{g'_{(0,\beta)}}{g_0}\rangle_H.$$

Notice that

$$\sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_0}\right)^2 (X_i) = n \cdot (1 + o_P(1)),$$

$$\sum_{i=1}^{n} \left(\frac{g'_{(0,\beta)}}{g_0}\right) d_1(X_i) = nu(\beta) \cdot (1 + o_P(1)),$$

where the $o(1)$ are uniform in probability, thanks to (AD) and (A5). Let us now see what happens on A_n and for $\theta \leq \eta_n$. Applying lemma 6.7 we obtain

$$\frac{\delta^2}{N(\beta)^3}\theta \leq \eta_n^{1-4\alpha/3}$$

which goes to 0 as soon as $\alpha < 3/4$. It is now not too hard to prove:

$$\begin{aligned} l_n(\theta,\beta) - l_n(0) &= \theta D_n(\beta) - \frac{\delta}{N(\beta)^2} F_n \theta^2 + o_P(n\theta^2) \\ &\quad - \frac{\theta^2}{2} n(1+o_P(1)) + \frac{\delta}{N(\beta)^2}\theta^3 nau(\beta)(1+o_P(1)) \\ &\quad - \frac{\theta^4}{2}\frac{\delta^2}{N(\beta)^4} na^2(1+o_P(1)) + o_P(n\theta^2) \end{aligned}$$

where all the $o_P(\cdot)$ are uniform in probability over β in A_n. Now,

$$\frac{\delta}{N(\beta)^2}\theta \leq \eta_n^{1-\alpha}$$

and since $(D_n(\beta), F_n)/\sqrt{n}$ converges uniformly in distribution using (AD) we have easily

$$\frac{\delta}{N(\beta)^2} F_n \theta^2 = o_P(\theta D_n(\beta))$$

where the $o_P(\cdot)$ is uniform in probability over β in A_n. We finally get for β in A_n and for $\theta \leq \eta_n$:

$$l_n(\theta,\beta) - l_n(0) = \left(\theta D_n(\beta) - \frac{\theta^2}{2} n\right)(1+o_P(1))$$

where again the $o(\cdot)$ is uniform in probability over β in A_n. Since $\hat{\theta}_\beta \leq \eta_n$ this obviously leads, by maximizing $\theta D_n(\beta) - \frac{\theta^2}{2} n$ to:

$$V_n(\beta) = \frac{1}{2}\frac{D_n(\beta)^2}{n} 1_{D_n(\beta)\geq 0}(1+o_P(1))$$

for β in A_n and where the $o(\cdot)$ is uniform in probability over β in A_n. The conclusion of Lemma 6.5 follows using (M4) and the fact that $\cup_n A_n = \mathcal{D}$.

PROOF OF LEMMA 6.6. We shall use again expansion (6.2), but the expansion for $g_{(\theta,\beta)}$ has now to be done till order 5:

$$g_{(\theta,\beta)}(x) = g_0(x) + \theta \cdot g'_{(0,\beta)}(x) + \sum_{i=2}^{4} \frac{\theta^i}{i!} \cdot g^{(i)}_{(0,\beta)}(x) + \frac{\theta^5}{5!} \cdot g^{(5)}_{(\theta*,\beta)}(x)$$

for a $\theta* \leq \theta$ and depending on x. The aim is now to prove that for $\theta \leq \eta_n$ and for $\beta \in B_n$ we have:

$$l_n(\theta,\beta) - l_n(0) = P_n(\theta,\beta)(1+o_P(1)) \tag{6.3}$$

where all the $o(\cdot)$ are uniform in probability over β in B_n and with

$$\begin{aligned} P_n(\theta,\beta) &= \theta D_n(\beta) - \frac{\delta}{N(\beta)^2} F_n \theta^2 - \frac{\theta^2}{2} n \\ &\quad + \frac{\delta}{N(\beta)^2}\theta^3 nau(\beta) - \frac{\theta^4}{2}\frac{\delta^2}{N(\beta)^4} na^2. \end{aligned}$$

First of all, notice that on B_n, δ and $N(\beta)$ are bounded by $\eta_n^{\alpha/3}$ and tend uniformly to 0. So that we may write:

$$g_{(\theta,\beta)}(x) = g_0(x) + \theta \cdot g'_{(0,\beta)}(x) - \sum_{i=2}^{4} \frac{\theta^i}{i!} \cdot \frac{i\delta^{i-1}}{N(\beta)^i} f_{\gamma^0}^{(i-1)}(x)(1+o_P(1)) - \frac{\theta^5}{4!}\frac{\delta^4}{N(\beta)^5} f_{\gamma^0}^{(5)}(x) + O(\frac{\delta^5}{N(\beta)^5}\theta^5 m_5(x)).$$

From now on, all the $o(\cdot)$ will be in probability uniformly for β in B_n. Now, using expansion (6.2) together with the previous result leads to

$$\begin{aligned} l_n(\theta,\beta) - l_n(0) = & \\ & \sum_{i=1}^{n}\Bigg(\theta \cdot \frac{g'_{(0,\beta)}}{g_0}(X_i) - \sum_{k=2}^{5} \frac{\theta^k}{(k-1)!}\frac{\delta^{k-1}}{N(\beta)^k}\frac{f_{\gamma^0}^{(k-1)}}{g_0}(X_i)(1+o_P(1)) \\ & + O(\frac{\delta^5}{N(\beta)^5}\theta^5\frac{m_5}{g_0}(X_i))\Bigg) - \frac{1}{2}\sum_{i=1}^{n}\Bigg(\theta \cdot \frac{g'_{(0,\beta)}}{g_0}(X_i) \\ & - \sum_{k=2}^{5} \frac{\theta^k}{(k-1)!}\frac{\delta^{k-1}}{N(\beta)^k}\frac{f_{\gamma^0}^{(k-1)}}{g_0}(X_i)(1+o_P(1)) + O(\frac{\delta^5}{N(\beta)^5}\theta^5\frac{m_5}{g_0}(X_i))\Bigg)^2 \\ & + O\Bigg(\sum_{i=1}^{n}(\theta \cdot \frac{g'_{(0,\beta)}}{g_0}(X_i) - \sum_{k=2}^{5} \frac{\theta^k}{(k-1)!}\frac{\delta^{k-1}}{N(\beta)^k}\frac{f_{\gamma^0}^{(k-1)}}{g_0}(X_i)(1+o_P(1)) \\ & + O(\frac{\delta^5}{N(\beta)^5}\theta^5\frac{m_5}{g_0}(X_i)))^3\Bigg) \end{aligned}$$

which, when keeping only the two first terms in the first sum and when taking the squares in the second sum, leads to the fact that $l_n(\theta,\beta) - l_n(0)$ equals $P_n(\theta,\beta)(1+o_P(1))$ plus terms which may be bounded with one of the following forms:

$$\frac{\theta^k\delta^{k-1}}{N(\beta)^k}\sum_{i=1}^{n}\frac{f_{\gamma^0}^{(k-1)}}{g_0}(X_i), \quad k \geq 3,$$

$$\frac{\theta^5\delta^5}{N(\beta)^5}n \quad \frac{\theta^{k+1}\delta^{k-1}}{N(\beta)^k}n \quad \frac{\theta^{k+l}\delta^{k+l-2}}{N(\beta)^{k+l}}n, \quad k,l \geq 3,$$

$$\theta^3 n \quad \frac{\theta^{k+2}\delta^{k-1}}{N(\beta)^k}n \quad \frac{\theta^{k+l+1}\delta^{k+l-2}}{N(\beta)^{k+l}}n \quad \frac{\theta^{k+l+m}\delta^{k+l+m-3}}{N(\beta)^{k+l+m}}n, \quad k,l,m \geq 2.$$

Now, since $\theta/N(\beta) \leq 1$, the first term in this list may be bounded by:

$$\delta \cdot \frac{\theta^2\delta}{N(\beta)^2}\sum_{i=1}^{n}\frac{f_{\gamma^0}^{(k-1)}}{g_0}(X_i)$$

which is uniformly in probability

$$o_P(\frac{\theta^2\delta}{N(\beta)^2}F_n).$$

Some of the other terms will be proven to be $o_P(n\theta^2)$ using Lemma 6.7 and the fact that β is in B_n:

$$\frac{\theta^5\delta^5}{N(\beta)^5}n = O(n\theta^2\frac{\delta^5}{N(\beta)^2}) = O(n\theta^2\eta^{\alpha/3}),$$

$$\frac{\theta^{k+1}\delta^{k-1}}{N(\beta)^k}n = O(n\theta^2\frac{\delta^{k-1}}{N(\beta)}) = O(n\theta^2\eta^{\alpha(k-3)/3}),\ k \geq 4,$$

$$\frac{\theta^{k+l}\delta^{k+l-2}}{N(\beta)^{k+l}}n = O(n\theta^2\frac{\delta^{k+l-2}}{N(\beta)^2}) = O(n\theta^2\eta^{\alpha(k+l-6)/3}),\ \ k,l \geq 3,$$

$$\theta^3 n = o(n\theta^2)\quad \frac{\theta^{k+2}\delta^{k-1}}{N(\beta)^k} = O(n\theta^2\delta^{k-1}),$$

$$\frac{\theta^{k+l+1}\delta^{k+l-2}}{N(\beta)^{k+l}}n = O(n\theta^2\frac{\delta^{k+l-2}}{N(\beta)}) = O(n\theta^2\eta^{\alpha(k+l-4/3))}),\ \ k,l \geq 2,$$

$$\frac{\theta^{k+l+m}\delta^{k+l+m-3}}{N(\beta)^{k+l+m}}n = O(n\theta^2\frac{\delta^{k+l+m-3}}{N(\beta)^2}) = O(n\theta^2\eta^{\alpha(k+l+m-7)/3}),\ k+l+m \geq 8.$$

The remaining terms may be proven to be $o(n\frac{\theta^4\delta^2}{N(\beta)^4})$. They are:

$$\frac{n\theta^4\delta^2}{N(\beta)^3} = O(N(\beta)\cdot n\frac{\theta^4\delta^2}{N(\beta)^4}),$$

$$\frac{n\theta^5\delta^2}{N(\beta)^4} = O(\theta\cdot n\frac{\theta^4\delta^2}{N(\beta)^4}),$$

$$\frac{n\theta^6\delta^3}{N(\beta)^6} = O(\delta\cdot n\frac{\theta^4\delta^2}{N(\beta)^4}),$$

$$\frac{n\theta^7\delta^4}{N(\beta)^7} = O(\delta^2\cdot n\frac{\theta^4\delta^2}{N(\beta)^4}).$$

It is not possible now to conclude that (6.3) holds since the remaining terms are shown to be negligible with respect to one of the terms of P_n. However, they are uniformly negligible with respect to the involved term. Moreover, it will be seen that, at the optimizing value (θ,β), all terms in P_n have the same order. Our aim is to conclude that (6.3) holds and that to optimize $l_n(\theta,\beta) - l_n(0)$ we just have to maximize P_n and verify that all terms in P_n have the same order at the maximum point. To be able to conclude, we shall then only need to prove that, it is not possible that $P_n(\theta,\beta)$ becomes small together with the fact that some of its terms become of order bigger than that of the maximum value. Now, at the maximum value, all terms of P_n have the same order, which is $0(1)$, and not $o(1)$. Let us prove that the supremum of l_n is not reached when one of the terms of P_n tends to infinity, together with the fact that P_n is close to 0. Define $\phi = \frac{\delta}{N(\beta)^2}$;

- If $\phi F_n\theta^2$ tends to infinity, then it is small with respect to $-\theta^4\phi^2 na^2$ which is negative. If this last term is compared to $n\phi\theta^3$, it is much smaller only in case $\theta\phi$ tends to 0, in which case $n\phi\theta^3$ is small with respect to $-n\theta^2$ which is negative. We may conclude that in this case, P_n is not small.
- If $\theta D_n(\beta)$ tends to infinity, it is much smaller than $-n\theta^2$ which is negative. In case $n\phi\theta^3$ is much bigger than $-n\theta^2$, $\theta\phi$ tends to infinity and the only leading term is then $-n\phi^2\theta^4$ which is negative.

- If $\phi n\theta^3$ tends to infinity, then it has been seen that in case $\theta\phi$ tends to infinity, the only leading term is $-n\phi^2\theta^4$, and in case $\theta\phi$ tends to 0, the only leading term is $-n\theta^2$. Now, in case $\theta|\phi|$ is lower and upper bounded, let α be an accumulation value of $\phi\theta$. On the subsequence, $\alpha nau\theta^2 - \frac{1}{2}\alpha^2\theta^2 na^2 - \frac{1}{2}n\theta^2$ is negative.

We may conclude that the supremum value of l_n is attained in the region where all terms of P_n are $O(1)$, where (6.3) holds.
Now, we then have to optimize $P_n(\theta,\beta)$ for $\theta \le \eta_n$ and β in B_n. First notice that on B_n

$$D_n(\beta) = \frac{1}{N(\beta)}\left((\gamma-\gamma^0+\delta)\sum_{i=1}^n \frac{f'_{\gamma^0}}{g_0} + \frac{(\gamma-\gamma^0)^2}{2}\sum_{i=1}^n \frac{f"_{\gamma^0}}{g_0}\right)(X_i)(1+o_P(1))$$

where $N(\beta)$ has the same expansion. Depending on the leading terms in the expansion, the only possible approximations of $D_n(\beta)$ are the following:

$$D_n(\beta) = \left(\sum_{i=1}^n d_1(X_i)\right)(1+o_P(1)) \tag{6.4}$$

or

$$D_n(\beta) = \left(\sum_{i=1}^n \frac{\lambda d_1(X_i) + d_2(X_i)}{\sqrt{1+\lambda^2+2u\lambda}}\right)(1+o_P(1)) = D_n(\lambda)(1+o_P(1)) \tag{6.5}$$

for some real number λ. Moreover, the $o(1)$ terms may be uniformly bounded using a function of η_n. It follows that,

$$B_n = B_n(\infty) \cup (\cup_{\lambda\in\mathbb{R}^+} B_n(\lambda))$$

where $B_n(\infty)$ is the set of β such that $\delta/N(\beta)^2 \ge 1/\eta_n^\alpha$ and (6.4) holds, and $B_n(\lambda)$ is the set of β such that $\delta/N(\beta)^2 \ge 1/\eta_n^\alpha$ and (6.5) holds.
Maximization over $B_n(\infty)$. On this set, we have, up to a multiplying factor $1+o(1)$:

$$P_n(\theta,\beta) = \theta\frac{F_n}{a} - \frac{\theta^2}{2}n - \phi F_n\theta^2 + \phi na\theta^3 - \frac{\phi^2}{2}na^2\theta^4$$

where $\phi = \delta/N(\beta)^2$. We shall maximize it over ϕ and then over θ, and then verify that the optimizing values verify $\beta \in B_n(\infty)$ and $\theta/N(\beta) \le 1$. Maximizing in ϕ leads to

$$\phi = \frac{1}{a\theta} - \frac{F_n}{na^2\theta^2}$$

and the value of $P_n(\theta,\beta)$ for this value of ϕ is then

$$\frac{F_n^2}{2na^2}$$

which does not depend on θ, and converges to $1/2\cdot\xi_{d_1}^2$.
Let us now verify that the optimizing value may correspond to some $\beta \in B_n(\infty)$ and $\theta/N(\beta) \le 1$. Indeed, we may choose β such that $N(\beta) \sim c\delta$ for a constant c, so that $\phi \sim 1/c^2\delta$, and $\theta/N(\beta) \sim c\theta\phi$. Now for the optimizing value of ϕ we have

$$\theta\cdot\phi = \frac{1}{a} - \frac{F_n}{na^2\theta}$$

and since any θ now is an optimizing value we may choose $\theta = \frac{|F_n|}{n}$ where the constraints hold.

Maximization over $B_n(\lambda)$. On this set, we have, up to a multiplying factor $1 + o(1)$:

$$P_n(\theta, \beta) = \theta D_n(\lambda) - \frac{\theta^2}{2} n - \phi F_n \theta^2 + \phi n a u(\lambda) \theta^3 - \frac{\phi^2}{2} n a^2 \theta^4$$

where $\phi = \delta / N(\beta)^2$ and $u(\lambda) = (\lambda + u)/\sqrt{1 + \lambda^2 + 2u\lambda}$. We shall again maximize it over ϕ and then over θ, and then verify that the optimizing values verify $\beta \in B_n(\lambda)$ and $\theta / N(\beta) \leq 1$. Maximizing in ϕ leads to

$$\phi = \frac{u(\lambda)}{a\theta} - \frac{F_n}{n a^2 \theta^2},$$

and the value of $P_n(\theta, \beta)$ for this value of ϕ is then

$$\frac{F_n^2}{2na^2} + \theta (D_n(\lambda) - u(\lambda) \frac{F_n}{a}) - n \frac{\theta^2}{2} (1 - u^2(\lambda).$$

The maximization over θ leads to

$$\theta = \frac{1}{n} \left(\frac{D_n(\lambda) - F_n u(\lambda)/a}{1 - u^2(\lambda)} \right) 1_{D_n(\lambda) - F_n u(\lambda)/a > 0}.$$

On the event

$$1_{D_n(\lambda) - F_n u(\lambda)/a > 0} = 0,$$

we have

$$\sup_{(\theta, \beta), \beta \in B_n} P_n(\theta, \beta) \leq \frac{F_n^2}{2na^2}$$

and then, letting θ tend to 0,

$$\sup_{(\theta, \beta), \beta \in B_n} P_n(\theta, \beta) = \frac{F_n^2}{2na^2}.$$

On the event

$$1_{D_n(\lambda) - F_n u(\lambda)/a > 0} = 1,$$

easy computation gives

$$\theta = \frac{1}{n} \left(\frac{D_n(0) - u F_n/a}{1 - u^2} \sqrt{1 + \lambda^2 + 2\lambda u} \right) 1_{D_n(0) - u F_n/a > 0}.$$

The maximizing value of $P_n(\theta, \beta)$ is then

$$\frac{F_n^2}{2na^2} + \frac{1}{2n} \frac{(D_n(0) - u F_n/a)^2}{1 - u^2} 1_{D_n(0) - u F_n/a > 0}.$$

In all cases, the optimizing value of P_n converges to

$$\frac{1}{2} \xi_{d_1}^2 + \frac{1}{2} \xi_{(d_2 - u d_1)/\sqrt{1-u^2}}^2 1_{\xi_{(d_2 - u d_1)/\sqrt{1-u^2}} > 0}.$$

Let us now verify that the optimizing value may correspond to some $\beta \in B_n(\lambda)$ and $\theta / N(\beta) \leq 1$. Indeed, we may choose β such that $N(\beta) \sim c(\lambda) \delta^2$ for a constant $c(\lambda)$ (depending on λ), so that $\theta / N(\beta) \sim \tilde{c}(\lambda) \theta \phi^{2/3}$. Now for the optimizing value of ϕ we have

$$\theta^{3/2} \cdot \phi = \frac{u}{a} \sqrt{\theta} - \frac{F_n}{n a^2 \theta} \sqrt{\theta}.$$

Now, $\frac{F_n}{na^2\theta}$ converges in distribution for the optimizing value of θ, θ converges to 0 in probability, so that the constraints hold. Lemma 6.6 is thus proved.

PROOF OF THEOREM 4.5. The proof follows the same line as that of Theorem 4.3. We first prove that Lemma 6.7 still holds. Again, assume that β_n is a sequence such that $\frac{\delta_n^2}{N(\beta_n)}$ tends to infinity. Then, using (M3*), δ_n tends to 0 and also $\lambda_{i,n}(\gamma_n^i - \gamma^0)^2$ for each $i = 1, \ldots, p-1$. By eventually extracting convergent subsequences, let now I be the set of i such that γ_n^i converges to some $\gamma^{i,*}$ different from γ^0, and let J be the complementary set of indices. Then:

$$\begin{aligned} N(\beta_n) &= \|(\sum_{i\in I}\lambda_{i,n}\frac{f_{\gamma^{i,*}} - f_{\gamma^0}}{g_0})(1+o_P(1)) + (\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)+\delta_n)\frac{f'_{\gamma^0}}{g_0}) \\ &\quad + \frac{1}{2}(\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)^2\frac{f"_{\gamma^0}}{g_0})(1+o_P(1))\|. \end{aligned}$$

There are only three possibilities:

1). If

$$N(\beta_n) = \|(\sum_{i\in I}\lambda_{i,*}\frac{f_{\gamma^{i,*}} - f_{\gamma^0}}{g_0}) + \mu\frac{f'_{\gamma^0}}{g_0} + \frac{f"_{\gamma^0}}{g_0}\|\frac{1}{2}(\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)^2)(1+o_P(1))$$

where the $\lambda_{i,*}$ are non negative real numbers and μ is a real number. Then

$$\frac{\delta_n^2}{N(\beta_n)} = O\left(\frac{(\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0))^2}{\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)^2}\right) = O(1).$$

2). If

$$N(\beta_n) = \|(\sum_{i\in I}\lambda_{i,*}\frac{f_{\gamma^{i,*}} - f_{\gamma^0}}{g_0}) + \mu\frac{f'_{\gamma^0}}{g_0}\| \cdot |\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)+\delta_n|(1+o_P(1))$$

where now μ is 1 or -1. Then

$$\frac{\delta_n^2}{N(\beta_n)} = O\left(\frac{\delta_n^2}{(\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)+\delta_n)^2}\right) = O(1).$$

3). If

$$N(\beta_n) = \|(\sum_{i\in I}\lambda_{i,*}\frac{f_{\gamma^{i,*}} - f_{\gamma^0}}{g_0})\| \cdot C\max_{i\in I}\lambda_{i,n}(1+o_P(1))$$

where C is some constant. Then

$$\sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)+\delta_n = o(\max_{i\in I}\lambda_{i,n}) \quad \text{and} \quad \sum_{i\in J}\lambda_{i,n}(\gamma_n^i-\gamma^0)^2 = o(\max_{i\in I}\lambda_{i,n}).$$

This implies that

$$\delta_n = o(\sqrt{\max_{i\in I}\lambda_{i,n}})$$

so that

$$\frac{\delta_n^2}{N(\beta_n)} = O\left(\frac{\delta_n^2}{\max_{i\in I}\lambda_{i,n}}\right) = o(1)$$

and Lemma 6.7 is proved.

The formula for $g^{(k)}$ still holds for $k \geq 2$, and that for $k = 1$ is obviously

changed. Now, in all expansions, only Lemma 6.7 is used, and not the particular form of $D_n(\beta)$, till the end of the proof of Lemma 6.6. So that, following the same lines we see that Lemma 6.5 still holds, and that on $\delta/N(\beta) \geq 1/\eta_n^\alpha$ we have the uniform approximation of $l_n(\theta,\beta) - l_n(0)$ by $P_n(\theta,\beta)$ with the same formula. Difference in the proof comes when approximating $D_n(\beta)$. For non negative $\lambda_1,\ldots,\lambda_{p-1}$, any λ and $\epsilon = 0$ or 1, if $\Lambda = (\lambda_1,\ldots,\lambda_{p-1},\lambda,\epsilon)$, define

$$d(\Lambda) = \frac{\sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - f_{\gamma^0}}{g_0} + \lambda d_1 + \epsilon d_2}{\| \sum_{i=1}^{p-1} \lambda_i \frac{f_{\gamma^i} - f_{\gamma^0}}{g_0} + \lambda d_1 + \epsilon d_2 \|_H}$$

then the only possible approximations of $D_n(\beta)$ take the form

$$D_n(\beta) = \left(\sum_{i=1}^{n} d(\Lambda)(X_i) \right) (1 + o_P(1)) = D_n(\Lambda)(1 + o_P(1)). \tag{6.6}$$

Define $u(\Lambda) = \langle d(\Lambda), d_1 \rangle$. Following the same lines as for Lemma 6.6, we only need to maximize $P_n(\theta,\beta)$ replacing $D_n(\beta)$ by some $D_n(\Lambda)$ and $u(\beta)$ by $u(\Lambda)$ (The fact that we only need to maximize $P_n(\theta,\beta)$ follows the same arguments *a posteriori* than in the proof of Lemma 6.6). We perform the maximization similarly in ϕ then in θ, which leads to the optimizing values:

$$\phi = \frac{u(\Lambda)}{a\theta} - \frac{F_n}{na^2\theta^2}$$

$$\theta = \frac{D_n(\Lambda) - u(\Lambda)F_n/a}{n(1-u^2(\Lambda))} 1_{D_n(\Lambda) - u(\Lambda)F_n/a > 0}.$$

On the event

$$1_{D_n(\Lambda) - u(\Lambda)F_n/a > 0} = 0,$$

we again have $\sup P_n = \frac{F_n^2}{2na^2}$. On the event

$$1_{D_n(\Lambda) - u(\Lambda)F_n/a > 0} = 1,$$

computation leads to the maximum value for P_n:

$$\frac{F_n^2}{2na^2} + \frac{1}{2n} \frac{(D_n(\Lambda) - u(\Lambda)F_n/a)^2}{1-u^2(\Lambda)} 1_{D_n(\Lambda) - u(\Lambda)F_n/a > 0}.$$

Verification that the optimizing values lie in the right set are straightforward. Then, notice that $d(\Lambda) - u(\lambda)d_1$ is orthogonal to d_1, and use assumption (AD) to end the proof.

Acknowledgments

The authors want to thank an anonymous referee for pointing out the reference to prove theorem 3.6.

References

Adler, R. J. (1990). *An introduction to continuity, extrema, and related topics for general Gaussian processes.* IMS Lecture Notes-Monograph Series.

Azais, J. M. and Wschebor, M. (1995). A formula to compute the distribution of the maximum of a random process. Université P. Sabatier, Toulouse.

Azencott, R. and Dacunha-Castelle, D. (1984). *Séries d'observations irrégulières.* Masson.

Beran, R. and Millar, P. W. (1987). Stochastic estimation and testing. *Annals of Stat.*, **15** 1131–1154.

Berdai, A. and Garel, B. (1994). Performances d'un test d'homogénéité contre une hypothèse de mélange gaussien. *Rev. Stat. Appl.* **42** 63–79.

Bickel, P. and Chernoff, H. (1993). Asymptotic distribution of the likelihood ratio statistic in a prototypical non regular problem. In *Statistics and Probability: A Raghu Raj Bahadur Festschrift.*

Dacunha-Castelle, D. and Duflo, M. (1986). *Probability and Statistics.* Springer Verlag New-York.

Dacunha-Castelle, D. and Gassiat, E. (1996). Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Submitted.*

Donoho, D. L. (1988). One-sided inference about functionals of a density. *Annals of Stat.* **16** 1390–1420.

Dudley, R. M. (1967). The size of compact subsets of hilbert space and continuity of Gaussian processes. *Journal of Funct. Analysis* **1** 290–330.

Ghosh, J. and Sen, P. (1985). On the asymptotic performance of the log-likelihood ratio statistic for the mixture model and related results. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer.* Le Cam, L.M. and Olshen, R.A. eds.

Hannan, E. J. (1982). Testing for autocorrelation and akaike's criterion. In *Essays in Statistical Science*, p. 403-412. Gani, J.M., Hannan, E.J. eds.

Hartigan, J. A. (1985). A failure of likelihood ratio asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer.* Le Cam, L.M. and Olshen, R.A. eds.

Laurent, B. (1993). *Estimation de fonctionnelles intégrales non linéaires de la densité et de ses dérivées.* PhD thesis, Université de Paris XI, France.

Ossiander, M. (1987). A central limit theorem under metric entropy with l^2 bracketing. *Annals of Prob.* **15** 897–919.

Redner, R. (1981). Note on the consistency of the maximum likelihood estimate for nonidentifiable distributions. *Annals of Stat.* **9** 225–228.

Roussas, G. G. (1970). *Contiguity of probability measures: some applications in statistics.* Princeton University press.

Self, S. and Liang, K. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Jour. Amer. Stat. Assoc.* **823** 605–610.

Teicher, H. (1965). Identifiability of finite mixtures. *Annals of Math. Statist.* **36** 423–439.

Van der Vaart, A. W. and Wellner, J. A. (1996). *Empirical Processes.* Springer Verlag.

Yakowitz, S. J. and Spragins, J. D. (1968). On the identifiability of finite mixtures. *Annals of Math. Stat.* **39** 209–214.

Laboratoire Modélisation Stochastique et Statistique, Université d'Orsay, Bat. 425, 91405 Orsay, France.

Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Bd. des Coquibus, 91025 Evry, France. E-mail: gassiat@lami.univ-evry.fr