EVRY-VAL D'ESSONNE
N° d'ordre : 99EVRY0006

THÈSE

présentée devant

L' UNIVERSITÉ D'EVRY-VAL D'ESSONNE

pour obtenir le titre de

Docteur en sciences de l'Université d'EVRY VAL D'ESSONNE

Spécialité : Mathématique

par

Christine Keribin

Sujet de la thèse

Tests de modèles par maximum de vraisemblance

Soutenue le 31 mai 1999 devant la commission d'examen composée de :

Mmes-Mrs.	Fabrice	GAMBOA	Rapporteur
	Elisabeth	GASSIAT	Directeur de thèse
	Dominique	HAUGHTON	
	Monique	JEANBLANC	
	Xavier	MILHAUD	Rapporteur
	Larry	WASSERMAN	Rapporteur

Résumé

L'ordre d'un modèle paramétrique est, grossièrement, le nombre de paramètres nécessaires pour définir le modèle. Lorsque cet ordre n'est pas connu, l'estimateur du maximum de vraisemblance va surestimer l'ordre, pour s'ajuster le mieux aux données. Mais un modèle sur-paramétré ne donnera pas de bons résultats de prédiction. Ainsi, il est intéressant d'étudier des estimateurs et des tests de l'ordre. Nous étudions des tests de rapport de vraisemblance dans le cadre de trois modèles: Le modèle à observations indépendantes et identiquement distribuées (i.i.d.) identifiable, même quand l'ordre surestimé, le modèle de mélange à observations i.i.d. (qui n'est pas identifiable quand le nombre de composantes est surestimé), et le modèle de chaîne de Markov cachée (CMC) à observations continues (qui présente les mêmes problèmes d'identifiabilité que le précédent, avec en plus la dépendance Markovienne). Dans le premier cas, nous utilisons des résultats connus de consistance de l'estimateur de l'ordre pour déterminer la vitesse asymptotique de la probabilité de se tromper d'ordre. Dans le second cas, nous démontrons d'abord la consistance de l'estimateur du maximum de vraisemblance pénalisée sous certaines hypothèses, puis nous donnons un majorant du niveau d'un test de contamination. Dans le dernier cas, nous testons un modèle à observations i.i.d. contre un modèle CMC à deux états cachés. Dans ce cas, sur un sous-ensemble des paramètres, nous montrons que le rapport de vraisemblance tend en loi vers la moitié du supremum du carré d'un processus gaussien tronqué à ses valeurs positives. Puis agrandissant le domaine, nous montrons que le rapport de vraisemblance tend vers l'infini en probabilité. Prenant un cas particulier de CMC, le modèle MA bruité, nous étudions, à ordre connu, un estimateur du filtre et des paramètres de la chaîne cachée, quand celle-ci est, elle aussi, à états continus. Des simulations permettent d'illustrer les résultats. Ces études sont accompagnées d'un travail bibliographique les positionnant dans le contexte actuel, les éclairant par des exemples, montrant leur apport, et proposant des perspectives de recherche.

Absract

The order of a parametric model is, roughly, the number of parameters needed to define the model. When this order is unknown, the maximum likelihood estimator over-estimates the order, to fit the data as best as possible. But an over-parameterized model will not give good prediction results. Hence, it is interesting to study estimators or tests of the model order. We study likelihood ratio tests for three models: the model with independent and identically distributed (i.i.d.) observations, identifiable even if the order is over-estimated, the mixture model with i.i.d. observations (which is not identifiable when the number of components is over-estimated), and the hidden Markov chain (HMC) model with continuous observations (which has the same identifiability problem, with, moreover, the Markovian dependence). In the first case, we use known consistence results of the order estimator to determine the asymptotic speed of the probability of choosing a wrong order. In the second case, we first prove, with appropriate assumptions, the consistence of the maximum penalized likelihood estimator, we give then an upper bound for the asymptotic level of a contamination test. In the last case, we test a model with i.i.d. observations against a HMC model with two hidden states. We prove in this case that, on a restricted set of parameters, the likelihood ratio converges in distribution to half the supremum of the square of a Gaussian process, restricted to its positive values. Then, increasing the parameter domain, we show that the likelihood ratio converges to infinity in probability. Taking a particular HMC case, the MA noisy model, we study, at a given known order, an estimator of the filter and the hidden chain parameters, when this chain has also continuous states. Simulations allow us to illustrate the results. A bibliographic work is added to these studies, in order to define their present context, to enlighten them with examples, to show their contribution, and to propose further works.

Remerciements

Je tiens à exprimer ma gratitude à Elisabeth Gassiat, qui a accepté la direction de ma thèse, alors que je ne pouvais y travailler qu'à mi-temps. Ses conseils, son suivi et ses encouragements m'ont été très précieux tout au long de ce travail et ont été la clé de son aboutissement.

Je suis très reconnaissante à Fabrice Gamboa, Xavier Milhaud et Larry Wasserman d'avoir accepté la tâche de rapporteur. Leurs remarques m'ont été précieuses pour la mise au point de ce mémoire. Thank you Larry for reviewing my work, in spite of some French in it!

Je remercie vivement Fabrice Gamboa, Elisabeth Gassiat, Dominique Haughton, Monique Jeanblanc et Xavier Milhaud pour leur participation au Jury.

Je tiens à remercier Dominique Haughton et Olivier Bermond. J'ai apprécié leur dynamisme, leur disponibilité et leur bonne humeur lors de nos collaborations. Merci aussi à Marc Lavielle et à Jean-Marc Bardet pour leurs conseils. Et merci à Valérie, secrétaire du labo d'Analyse et Probabilité.

Enfin, je porte une pensée spéciale à ma famille, qui, par ses encouragements et son soutien logistique, a tout mis en oeuvre pour me permettre de mener à bien ce travail.

A ma maman

Table des Matières

Chapitre Premier

Présentation générale

Le travail présenté ici porte sur l'utilisation de techniques de maximum de vraisemblance pour l'étude de modèles qui comportent des données cachées, que ces données soient noyées dans un bruit, ou dans un ensemble de paramètres trop grand. Ce dernier cas est apparenté à la partie des statistiques appelée choix de modèle ("model selection" dans la littérature anglo-saxone), dont l'intérêt ne se dément pas depuis les travaux pionniers de Akaike (1973) et Schwartz (1978) : il s'agit d'estimer ou de tester l'ordre d'un modèle, grossièrement, le nombre de paramètres qui décrivent le modèle, et donc d'en mesurer la complexité.

Nous commencerons par l'étude de l'ordre dans le cas où le modèle est identifiable, même quand il y a plus de paramètres que nécessaire. Utilisant le maximum de vraisemblance compensée (on dit aussi pénalisée), nous étudierons les probabilités de sous-estimer ou surestimer l'ordre réel du modèle. Ainsi, nous déterminerons à quelle vitesse le niveau de la statistique de test basée sur le maximum de vraisemblance compensée tend vers 0, et à quelle vitesse sa puissance tend vers 1.

Puis, nous nous intéresserons à la détermination de l'ordre des mélanges finis, c'est-à-dire à la détermination du nombre de leurs composantes. Le manque d'identifiabilité de ce type de modèle est connu, et la théorie asymptotique standard ne s'applique plus. Cependant, de récents résultats (Dacunha-Castelle et Gassiat (1999, 1997)) ont permis un pas décisif dans l'étude des modèles de mélanges. Utilisant une re-paramétrisation du modèle appelée "paramétrisation localement conique", ces auteurs ont démontré la convergence et calculé la distribution asymptotique de la statistique du maximum de vraisemblance . Ces résultats nous permettront de définir un estimateur de l'ordre d'un modèle de mélange par maximum de vraisemblance compensée , et nous montrerons sous quelles conditions il est consistant, et même fortement consistant. Nous étudierons alors les vitesses de l'estimateur du maximum de vraisemblance compensée , dans le cas d'un modèle de contamination.

Nous avons, jusqu'à présent, traité d'échantillons i.i.d. (indépendants et identiquement distribués). La suite de l'étude est consacrée à des modèles à dépendance markovienne, et plus particulièrement du type chaîne de Markov cachée . Ici, la source est une chaîne de Markov qui n'est pas directement observée, car noyée dans un bruit. Dans un premier temps, nous nous consacrerons à l'étude d'un modèle MA (Moving Average) bruité, et, à ordre connu, nous verrons comment il est possible d'identifier le système en utilisant une pseudo-vraisemblance et les avantages numériques de cette méthode. Dans un second temps, nous nous poserons le problème de test suivant : modèle i.i.d. contre chaîne de Markov cachée à deux états, qui est le pendant en régime markovien du test de contamination que nous avons étudié précédemment.

Ces études sont accompagnées d'exemples numériques, ce qui nous permettra de faire le point sur les algorithmes typiquement utilisés dans les situations que nous avons décrites.

Ce premier chapitre d'introduction est organisé de la façon suivante : nous commencerons par rappeler ce qu'est l'ordre d'un modèle, ainsi que les critères d'information associés. Puis une section sera réservée à chaque grand type de modèle : modèle identifiable à observations i.i.d., modèle de mélange à observations i.i.d., chaîne de Markov cachée . Dans chacune de ces sections, nous allons définir le modèle, l'éclairer par des exemples d'application, positionner notre étude dans le champ des recherches actuelles, et préciser notre contribution ainsi que les grands axes de démonstration. Quelques perspectives de recherche pourront être proposées. De plus, une section particulière présentera les techniques de simulation.

Tous les détails et preuves techniques sont consultables dans des chapitres dédiés à chaque étude, et dans lesquels nous avons rappelé les principaux théorèmes utilisés : Test de l'ordre d'un modèle de familles régulières identifiables, Estimation de l'ordre d'un modèle de mélange, Test de contamination, Estimation de paramètres dans un modèle MA bruité, Test de l'ordre d'un modèle de mélange à régime markovien.

1.1 Ordre d'un modèle et critères de choix

Définissons $\mathcal{T}_p = \{\theta^p = (t_1 \dots t_p) \in T_1 \times \dots \times T_p\}$, où $T_i, i = 1, \dots, p$ sont des ensembles compacts de $\mathbb{R}^{r_i}$. La dimension du paramètre θ^p est $R_p = \sum_{i=1}^{p} r_i$ et $\theta^p = (\theta_1, \dots, \theta_{R_p})$. Nous considérons maintenant la suite de modèles paramétriques à densité par rapport à une mesure ν, emboîtés de la façon suivante :

$$(\mathcal{G}_p = g(\theta^p, .), \theta^p \in \mathcal{T}_p)_{p \leq Q} \ , \ \mathcal{G}_{p_1} \cap \mathcal{G}_{p_2} = \mathcal{G}_{p_1 \wedge p_2} \quad (1)$$

où $p_1 \wedge p_2 = \min(p_1, p_2)$ Nous définissons alors l'ordre réel q d'un modèle de densité $g \in \mathcal{G}_Q$ comme le plus petit nombre de t_i nécessaires à décrire le modèle. Q est un majorant connu de l'ordre du modèle. Cette définition peut bien sûr s'étendre au cas de modèles non emboîtés, mais qui ont une intersection non vide, et pour lesquels la loi d'intérêt appartient à l'intersection. Remarquons que l'ordre d'un modèle n'est pas forcément égal au nombre de paramètres qui gouvernent sa densité, mais est une fonction de ce nombre. Typiquement, l'ordre d'un modèle de mélange est le nombre de ses composantes distinctes, qui est inférieure au nombre total des paramètres qui gouvernent la loi du mélange.

Nous observons un échantillon i.i.d. $(X_1, \ldots, X_n)$ de loi $g_0 \in \mathcal{G}_Q$ ayant comme ordre réel (inconnu) q et comme vrai paramètre (inconnu) $\theta^q = \theta_0^q$. Pour tout ordre $p \in \{1, .., Q\}$ et tout $g \in \mathcal{G}_p$, définissons la vraisemblance des observations :

$$l_n(g) = \sum_{i=1}^{n} \log g(X_i)$$

et la statistique du maximum de vraisemblance dans un modèle de dimension p

$$T_n(p) = \sup_{g \in \mathcal{G}_p} l_n(g)$$

Une question fondamentale en modélisation statistique est le choix d'un modèle bien ajusté, et comprenant le moins de paramètres possible. On pourrait songer à estimer q en choisissant l'entier maximisant les $T_n(p)$, mais cela conduirait presque toujours à $\hat{q} = Q$, et donc à un estimateur ne convergeant pas lorsque n tend vers l'infini. En maximisant sur l'ordre les maxima de vraisemblances, on obtient un modèle qui s'ajuste le mieux possible à l'observation, mais avec trop de paramètres. Si maintenant, on utilise le modèle estimé pour la prédiction, celle-ci sera de qualité très moyenne. Il faut donc un compromis entre la qualité de l'ajustement et celle de la prédiction, qui va privilégier les modèles ayant moins de paramètres, et aboutir à un estimateur consistant de l'ordre : c'est le principe de l'estimateur du maximum de vraisemblance compensée que nous présentons maintenant.

La vraisemblance compensée est définie en soustrayant au maximum de vraisemblance un terme de compensation $a_{n,p}$, définissant ainsi un critère d'information

$$W_n(p) = T_n(p) - a_{n,p}$$

Le choix de la compensation pose deux problèmes :

- Eviter la sous-estimation, c'est-à-dire l'oubli de certains paramètres. Quand l'ordre est sous-estimé, en général, la statistique du maximum de vraisemblance normalisée $T_n(p)/n$ converge vers le minimum de la distance de Kullback entre le modèle d'ordre p et la vraie loi. Le choix de $a_{n,p} = o(n)$ amène donc à un estimateur qui, asymptotiquement, ne sous-estime pas l'ordre.

- Eviter la surestimation, c'est-à-dire l'ajustement à un trop grand nombre de paramètres, qui nuiront ensuite à la qualité de la prédiction.

Le terme $a_{n,p}$ doit donc être croissant avec p, afin de compenser l'augmentation de vraisemblance d'un modèle sur-paramétré, et sera souvent mis sous la forme $a_{n,p} = p \cdot v(n)$. Il est maintenant connu que $v(n)$ doit tendre vers l'infini. Notons la première proposition $v(n) = 1$ de Akaike (1973), définissant ainsi le critère d'information AIC. Mais ce critère surestime l'ordre. Un peu plus tard, Schwarz (1978), en suivant des considérations bayésiennes, introduisit un critère BIC dont la compensation est $a_{n,p} = \frac{p}{2}\log(n)$. D'autres critères ont été proposés tels que le critère ΦIC défini en $\log\log(n)$ par Hannan et Quinn (1979) dans un cadre non i.i.d..

Nous définissons donc l'estimateur du maximum de vraisemblance compensée $\hat{q}_n$ de l'ordre q comme la valeur de p maximisant la vraisemblance compensée :

$$\hat{q}_n = \arg \sup_{p\in\{1,\dots,Q\}} (W_n(p))$$

Nous nous placerons dans les deux cas suivants :

- Le modèle est identifiable si l'ordre est bien ou sur-estimé, c'est-à-dire : supposons que la loi g_0 du modèle étudié est d'ordre q et de paramètre donné θ_0^q. Alors, parmi toutes les lois $g(\theta^p)$ d'ordre p supérieur ou égal à q :

$$\forall p \geq q \ , \ \{g(\theta^p, .) = g_0(.)\} \Leftrightarrow \{\theta^p = (\theta_0^q, 0, \dots, 0)\}$$

  Dans ce cas, les tests de rapport de vraisemblance suivent la théorie du χ^2 (Wilks (1967)).

- Le modèle n'est pas identifiable quand l'ordre est surestimé. Le cas typique est celui des modèles de mélange : en effet, soit $\pi \in [0,1]$, g_0 peut s'exprimer sous la forme $(1-\pi)g_0 + \pi g_0$, contredisant l'hypothèse d'identifiabilité, et nécessitant des développements théoriques plus importants que ceux de la théorie du χ^2.

Nous allons étudier ces cas dans les deux sections suivantes.

1.2 L'ordre dans un modèle identifiable

Nous nous plaçons ici dans le cas où le modèle est identifiable, quand l'ordre est bien ou surestimé et nous observons un échantillon i.i.d..

1.2.1 Les applications

Le choix du degré dans une régression polynomiale est un premier exemple d'application, si l'on considère que les variables de la régression sont supposées être des variables aléatoires et non des "variable de conception", ainsi que l'a montré Haughton (1988). Cet auteur a également étudié (Haughton (1997)) le modèle ANOVA à effet aléatoire, qui permet de tester une présence ou une absence d'effet de groupe. Les lois intervenant dans ces deux exemples appartiennent aux familles de lois exponentielles.

La famille des binomiales négatives, définie sur les entiers par :

$$g(\alpha, r, x) = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-r)^{\alpha} r^{x}, \ x = 0, 1, \cdots. \tag{2}$$

donne aussi un exemple de choix de modèle intéressant. Les paramètres de cette famille sont $\alpha > 0$ et $0 < r < 1$. Nous pouvons supposer que $\alpha \in [\alpha_1, \alpha_2]$ et $r \in [r_1, r_2]$, où α_1, α_2, r_1 et r_2 sont convenablement choisis de telle sorte que $1 \in (\alpha_1, \alpha_2)$. Il est en effet bien connu (Brown (1986)) que la famille des binomiales négatives n'est pas exponentielle quand α et r sont tous les deux inconnus. Le problème de sélection de modèle est motivé par la question : est-on dans le cas d'une famille de lois géométriques ($\alpha = 1$) ou non ($\alpha \neq 1$). En effet, si $\alpha = 1$, les observations proviennent d'une loi dite *sans mémoire*, c'est-à-dire telle que :

$$P(X > x | X > y) = P(X > x - y), \quad y > x.$$

1.2.2 L'état de l'art

Le critère AIC n'est pas consistant, car la probabilité de choisir un mauvais modèle ne tend pas forcément vers 0 quand la taille de l'échantillon grandit (voir par exemple Shibata (1978) pour le cas des processus ARMA, et Woodroofe (1982) sous des conditions générales de régularité). Il a en effet tendance à surestimer l'ordre. En revanche, le critère BIC, basé sur les termes principaux du développement asymptotique de la probabilité a posteriori (cadre bayésien) est consistant pour les familles exponentielles courbes (Haughton (1988)). Senoussi (1988) a proposé un travail très complet, donnant les conditions de pénalisation qui assurent la consistance du maximum de vraisemblance compensée dans le cadre des familles régulières. Enfin, dans le cas non i.i.d. des modèles ARMA, citons par exemple les travaux de Hannan (1980) et Azencott et Dacunha-Castelle (1984).

Connaissant les critères sur la pénalisation pour assurer la consistance, nous savons que la probabilité de commettre une erreur de choix de modèle en utilisant l'estimateur du maximum de vraisemblance compensée , tend vers 0 quand la taille de l'échantillon tend vers l'infini. La question intéressante est maintenant de savoir à quelle vitesse. Or, le comportement de la statistique du maximum de vraisemblance $T_{n,p}$ dépend de p :

- si $p > q$, $T_n(p) - T_n(q)$ converge vers une loi du χ^2 à $R_p - R_q$ degrés de liberté (Wilks (1967)).
- si $p < q$, $T_n(p) - T_n(q)$ converge vers $-\infty$, à une vitesse de l'ordre de n (avec des arguments similaires à ceux de Leroux (1992a) par exemple)

$$\lim_{n\to+\infty} \frac{1}{n}(T_n(p) - l_n(g_0)) = -K < 0$$

De même, les probabilités de sur-estimer $P(W_n(q+k) > W_n(q))$ ou de sous-estimer $P(W_n(q-k) > W_n(q))$ se comporteront de façons différentes. Ces probabilités asymptotiques de surestimer ou de sous-estimer l'ordre ont été étudiées dans le cas des familles exponentielles (Woodroofe (1982), Haughton (1989)) et dans le cas de la régression linéaire Nishii (1984) pour le critère BIC.

1.2.3 Contribution

Les résultats présentés dans cette section ont été réalisés en collaboration avec D. Haughton (Keribin et Haughton (1999)). Nous y étendons les résultats de Haughton (1989) au cas des observations i.i.d. provenant de familles régulières, qui ne sont pas nécessairement exponentielles. D'autre part, nous donnons ces probabilités pour une gamme plus large de pénalisation, incluant le cas du critère BIC. Ainsi, nous montrons que, sous des hypothèses de moments exponentiels des dérivées de $\log g(\theta, .)$ la probabilité de surestimer l'ordre du modèle est telle que :

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n > q)] = -1$$

et la probabilité $P(\hat{q}_n < q)$ de sous-estimer l'ordre est telle que :

$$\lim_{n\to+\infty} \frac{1}{n} \log[P(\hat{q}_n < q)] < 0$$

pour des compensations $a_{n,p} = p \cdot v(n)$, satisfaisant $\lim_{n\to+\infty} v(n) = +\infty$ et $v(n) = o(n)$. Ainsi, la vitesse de l'erreur de sous-estimation est un résultat de grandes déviations, tandis que celui de la vitesse de l'erreur de surestimation est un résultat de déviations modérées.

Dans le cas des familles exponentielles, et pour $v(n) = \frac{1}{2}\log n$, les résultats de Haughton (1989) donnent une borne supérieure de la taille de l'erreur de surestimation (probabilité d'estimer que l'ordre d'un modèle est $q+k$ alors que son ordre réel est q) de l'ordre de $O(n^{-\frac{k}{8}})$, et de la taille de l'erreur de sous-estimation de l'ordre de $O(e^{-n\cdot a})$, pour $a > 0$. Nos résultats retrouvent donc la vitesse de l'erreur de sous-estimation et améliorent même la majoration de l'erreur de surestimation que nous trouvons de l'ordre de $O(n^{-\frac{k-\varepsilon}{2}})$, pour $\varepsilon > 0$.

La probabilité d'erreur de sous-estimation converge beaucoup plus vite vers 0 que la probabilité d'erreur de surestimation. C'est pour cela que nous n'avons pas cherché un résultat optimal pour la sous-estimation, car c'est l'erreur totale qui est intéressante.

Considérons maintenant le test de $p = 1$ contre $p = 2$, avec $T_n(2) - T_n(1)$ comme statistique de test.

- si le seuil est fixé, alors le niveau est fixé (par les quantiles du χ^2), et la puissance tend vers 1.
- si le seuil est une fonction de $v(n)$, le niveau α_n est tel que $\lim_{n\to+\infty} \frac{1}{v(n)}\alpha_n \leq -1$, et tend donc vers 0, tandis que l'erreur de seconde espèce β_n est telle que $\lim_{n\to+\infty} \frac{1}{n}\beta_n \leq -a$, $a > 0$, la puissance tend donc vers 1.

Le maximum de vraisemblance compensée nous a donc permis de définir un test intéressant, car son niveau et son erreur de seconde espèce sont asymptotiquement nuls.

Enfin, les hypothèses de moments exponentiels introduites ici ont été vérifiées pour les familles de binomiales négatives, qui rappelons le, sont intéressantes car elles illustrent un cas de famille non exponentielle.

1.3 Les modèles de mélange fini à observations i.i.d.

Soit $\mathcal{F} = (f_\gamma)_{\gamma\in\Gamma}$ une famille de densités de probabilité par rapport à la mesure ν. Nous prendrons pour Γ un compact de $\mathbb{R}^k$ pour un entier k donné. $\mathcal{G}_p$ est maintenant l'ensemble de tous les p-mélanges finis de densités de $\mathcal{F}$:

$$\begin{aligned}\mathcal{G}_p \quad = \quad & \{g_{\pi,\alpha} = \sum_{i=1}^{p} \pi_i \cdot f_{\gamma^i} \ : \ \pi = (\pi_1, \ldots, \pi_p),\ \alpha = (\gamma^1, \ldots, \gamma^p), \\ & \forall i = 1, \ldots, p,\ \gamma^i \in \Gamma,\ 0 \leq \pi_i \leq 1,\ \sum_{i=1}^{p} \pi_i = 1\}\end{aligned} \tag{3}$$

Nous observons un échantillon i.i.d. $(X_1, \ldots, X_n)$ de loi g_0, pour $g_0 \in \mathcal{G}_p$. L'ordre réel q de g_0, est le nombre de composantes distinctes de g_0, c'est-à-dire le plus petit entier tel que tous les $\gamma^i, i = 1, \ldots, q$ soient différents, alors que leurs poids associés π_i sont non nuls.

Les mélanges sont source d'une nombreuse littérature, tant du point de vue théorique, que du point de vue des applications. On peut dès à présent citer quatre ouvrages de référence : les livres de Everitt et Hand (1981), Titterington, Smith et Makov (1985), McLachlan et Basford (1988) et Lindsay (1995).

1.3.1 Les applications

Les mélanges servent dans de multiples applications, en économie, médecine, génétique, botanique, géologie ou fiabilité par exemple. Titterington, Smith et Makov (1985) font un bilan très complet sur les applications et les techniques employées. Un exemple très connu, étudié par Cassie en 1954, est celui de l'étude de la longueur des poissons en fonction de leur âge. Cassie a dressé un histogramme de la longueur des poissons d'un échantillon de 256 individus. En observant les modes, il a détecté au moins deux groupes de poissons, ceux nés dans l'année, et ceux nés l'année précédente. Il a de plus suggéré qu'il y avait probablement deux groupes supplémentaires, correspondant à des individus plus âgés. Dans un autre domaine, Titterington, Smith et Makov (1985) citent aussi les travaux de Hasselblad en 1969, qui a étudié les avis de décès de femmes âgées de plus de 80 ans parus chaque jour dans le *Times of London* pendant la période 1910-1912. Ainsi, il a montré qu'une simple loi de Poisson ne s'ajustait que médiocrement aux données, tandis qu'un mélange de deux lois de Poisson améliorait l'ajustement. Cela pourrait indiquer deux schémas de mortalité dans les données, l'un pour l'hiver, l'autre pour l'été. Enfin, de façon récente, Chernoff et Lander (1995) ont travaillé sur le test que l'hypothèse d'un mélange de deux binomiales $\mathcal{B}(k,r)$ et $\mathcal{B}(k,\frac{1}{2})$ est le cas dégénéré d'une seule binomiale $\mathcal{B}(k,\frac{1}{2})$. Ce problème a une application directe en génétique, lorsque l'on veut déterminer s'il existe un marqueur pour un caractère génétique hétérogène, c'est-à-dire un caractère qui peut être causé par la mutation d'un gène parmi plusieurs possibles à différentes localisations. Si $p=\frac{1}{2}$, c'est qu'il n'y a pas de lien entre le marqueur proposé et un des gènes pour ce caractère.

Un sujet étroitement relié aux mélanges est celui très ancien (Gilmour (1951) en retrace déjà un historique commençant en 1851) de la classification ("clustering analysis"), à laquelle Gordon (1981) consacre une monographie, et qui est extensivement étudiée dans McLachland et Basford (1988) : comment construire une classification de n observations, éventuellement vectorielles, en sous-ensembles d'intérêt ("clusters") : Dans ce cas, les clusters sont autant de composantes du mélange. On peut ainsi vouloir estimer la densité du mélange, assigner les observations à un groupe, ou évaluer la pertinence de l'hypothèse du nombre de composantes.

1.3.2 Différentes approches des mélanges

Nous pouvons réécrire le modèle de mélanges (3) dans un cadre plus général :

$$g(x)=\int_\phi f(x|\phi)dG(\phi) \tag{4}$$

où $dG(\phi)$ est est la *mesure de mélange* associée au modèle de mélange.

Pour traiter des mélanges, on peut suivre une approche paramétrique, non paramétrique, ou semi-paramétrique.

- Quand aucune hypothèse n'est faite sur $dG(\phi)$ (discrète, continue, ou appartenant à une famille de loi donnée), le modèle est dit non-paramétrique.

- Si la mesure de mélange est modélisée sans hypothèse paramétrique, mais la famille mélangée f est paramétrique, on est dans le cas d'un modèle semi-paramétrique :

$$g(x; G, \theta) = \int_\phi f(x|\phi; \theta) dG(\phi)$$

  Lindsay et Lespérance (1995) ont fait une revue de ces modèles. Ce schéma peut-être utilisé pour des modèles mis sous la forme (Lindsay (1995)) :

$$g(x; \pi, \theta) = (1 - \pi) f(\theta, x) + \pi e(X)$$

  où $f(\theta, X)$ appartient à un modèle paramétrique, $e(X)$ est arbitraire (non paramétrique) et représente le manque d'ajustement du modèle (composantes manquantes, erreurs, etc).

- Enfin, si la mesure de mélange dégénère en une mesure discrète

$$dG(\phi) = \sum_{i=1}^{p} \pi_i \cdot \delta_{\gamma_i}(\phi)$$

  et que la famille f est paramétrique, on est dans le cadre d'un modèle paramétrique de mélange, ainsi que nous l'avons introduit en (3), et que nous nous proposons d'étudier ici.

Nous détaillons maintenant quelques techniques d'étude des paramètres et de l'ordre d'un modèle de mélange.

- *Les méthodes graphiques.* Basées sur les densités des lois, elles ne permettent souvent d'obtenir que des paramètres grossiers. De plus, elles sont délicates à employer pour déterminer l'ordre du mélange, spécialement lorsque le nombre de modes de la fonction est inférieur au nombre réel de composantes. On peut noter cependant les travaux de Roedner (1994), qui a utilisé une propriété de modalité du rapport de la densité d'un mélange de deux normales à une densité normale standardisée, pour définir un test graphique.

- *La méthode des moments.* Les paramètres sont exprimés en fonction des moments de la loi du mélange, eux-mêmes estimés par les moments empiriques. Les estimateurs des moments ne sont pas très efficaces, ce qui peut amener à une grande variance, mais ils peuvent donner de bonnes valeurs de départ dans les algorithmes. Voir par exemple Lindsay (1983) et de façon plus récente Lindsay et Roedner (1997). En ce qui concerne la détermination de l'ordre, Dacunha-Castelle et Gassiat (1997b), dans le cas des mélanges de lois à support discret, utilisent une propriété de dégénérescence des matrices de Hankel de dimension supérieure au nombre de composantes du modèle.

- *Les méthodes Bayésiennes.* Ayant mis le modèle sous forme (4), on peut considérer qu'il existe une variable aléatoire dite "latente" Φ, de loi a priori $G(\phi)$, et traiter le problème avec l'ensemble des données "complètes" (X, Φ). Ainsi, cette représentation en variables latentes lie les modèles de mélange au paradigme Bayésien. Dans sa forme la plus simple, $G(\phi)$ a une forme paramétrique simple, et on met des lois a priori sur les paramètres de cette loi.

- *Le maximum de vraisemblance :* les paramètres sont estimés par des estimateurs maximisant la vraisemblance du mélange. Leroux (1992a) a montré la consistance de l'estimateur du maximum de vraisemblance de la mesure de mélange quand le nombre de composants est inconnu. De plus, il a utilisé le fait que le ratio entre les vraisemblances du vrai modèle et des modèles d'ordre inférieur divisé par la taille de l'échantillon tend vers la distance de Kullback entre ces modèles pour montrer que l'estimateur du maximum de vraisemblance compensée ne sous-estime pas l'ordre du modèle. En revanche, il n'a pas donné de résultat de surestimation. En effet, pour tester l'existence d'un mélange de $p = q + k$ lois de la même famille paramétrique contre un mélange de q lois, l'utilisation du test général du rapport de vraisemblance n'est pas évidente, car le modèle n'est pas régulier.
  D'abord, l'hypothèse nulle est sur la frontière de l'espace des paramètres, plutôt qu'à l'intérieur, comme supposé dans la théorie classique.
  Plus important, le manque d'identifiabilité fait que la théorie asymptotique classique du maximum de vraisemblance ne s'applique plus, car le manque d'identifiabilité amène en général à la dégénérescence de l'information de Fisher. Ceci explique par exemple les résultats de Chen (1995), qui montre que l'estimation de la loi du mélange ne se fait qu'à la vitesse $n^{-1/4}$ quand le nombre de composants est inconnu, alors qu'elle est de $n^{-1/2}$ quand le nombre de composants est connu. Cependant, cela ne signifie pas que la statistique du maximum de vraisemblance ne converge pas. De fait, Dacunha-Castelle et Gassiat (1997) ont défini une nouvelle paramétrisation du modèle, appelée paramétrisation localement conique : un paramètre identifiable permet des développements de Taylor dans son voisinage, tandis que les autres contiennent toute la part non identifiable. Ainsi, la statistique du maximum de vraisemblance converge vers le supremum du carré d'un processus gaussien tronqué quand le modèle est dominé et que les paramètres sont bornés. S'il ne le sont pas, la statistique du maximum de vraisemblance peut en effet tendre vers vers l'infini (Hartigan 1985) à faible vitesse (Bickel et Chernoff,1993), de

l'ordre de $\frac{1}{2}\log\log(n)$.

Le cas mélange simple (test d'une composante donnée contre un mélange à deux composantes dont une est la composante donnée) a déjà donné de nombreux articles, comme celui de Chernoff et Lander (1995) et de Chen (1995) que nous avons précédemment cités, ou celui de Berdaï et Garel (1996), qui ont tabulé un test de rapport de vraisemblance dans un cas de mélange de gaussiennes.

1.3.3 Première contribution

Les résultats précédents (Dacunha-Castelle et Gassiat (1997, 1999)) nous ont servi pour démontrer la consistance de l'estimateur du maximum de vraisemblance compensée pour l'estimation de l'ordre d'un modèle de mélange lorsque la pénalisation vérifie $\lim_{n\to+\infty} v(n) = +\infty$ et $\frac{v(n)}{n} = o(n)$. Ainsi, nous complétons le résultat de Leroux (1992a), qui avait prouvé que l'estimateur du maximum de vraisemblance compensée ne sous-estime pas l'ordre, en ajoutant qu'il ne le surestime pas non plus. De plus, montrant que

$$\limsup_{n\to+\infty} \frac{T_n(p) - l_n(g_0)}{\log\log(n)}$$

est presque sûrement borné pour tout p tel que $1 \leq p \leq Q$ et imposant que la compensation vérifie de plus $\lim_{n\to+\infty} \frac{\log\log(n)}{v(n)} = 0$, nous avons obtenu la consistance presque sûre de l'estimateur du maximum de vraisemblance compensée (Keribin (1997, 1998)).

Enfin, nous avons étudié numériquement, dans le cas où le vrai modèle est un mélange de deux gaussiennes, l'influence du terme de compensation et comparé avec une méthode de moments (Dacunha-Castelle et Gassiat (1997b))

1.3.4 Deuxième contribution

En collaboration avec D. Haughton (Keribin et Haughton (1999)), nous avons ensuite cherché la vitesse asymptotique des probabilités de se tromper d'ordre, ainsi que nous l'avions fait dans le cas identifiable et nous présentons les résultats dans le cadre du mélange simple (ou modèle de contamination). Dans ce modèle, nous voulons tester l'hypothèse loi simple f_{γ^0} contre mélange à deux composantes dont l'une est f_{γ^0}. L'idée est que la plupart des observations appartiennent à la première composante de loi f_{γ^0}, mais qu'occasionnellement, l'observation est une valeur isolée provenant d'un deuxième composant.

$$\mathcal{G} \;=\; \{g_{\pi,\gamma} = (1-\pi)f_{\gamma^0} + \pi f_\gamma \;,\; f_{\gamma^0} \in \mathcal{F}, \;,\; f_\gamma \in \mathcal{F}, \;, 0 \leq \pi \leq 1 \;\}$$

Nous définissons donc le test

$$(H_0 :\; f_{\gamma^0}) \quad \text{contre} \quad (H1 :\; g \in \mathcal{G} \setminus \{f_{\gamma^0}\})$$

de domaine de rejet

$$T_n - l_n(g_0) \geq v(n)$$

et de niveau associé

$$\alpha_n := P_0(T_n - l_n(g_0) \geq v(n))$$

et montrons que le niveau décroît asymptotiquement à une vitesse exponentielle :

$$\lim \sup_{n\to+\infty} \frac{1}{v(n)} \log \alpha_n < -1$$

1.3.5 Perspectives de recherche

De façon générale, la choix de la compensation reste à investiguer. De façon théorique, elle peut varier dans une large plage de valeur, mais qu'en est-il pratiquement, et avec des échantillons qui sont finis? L'investigation numérique est donc une voie pour tabuler la compensation en fonction de la taille des échantillons par exemple.

1.4 Les modèles de chaîne de Markov cachée (CMC)

Jusqu'à présent, nous n'avions traité que d'observations i.i.d.. Nous allons maintenant nous pencher sur des modèles à dépendance markovienne du type chaîne de Markov cachée, qui sont un outil de modélisation puissant dans des applications de reconnaissance de la parole, de neurophysiologie, de biologie ou d'économie par exemple. Un modèle à chaîne de Markov cachée (CMC) est un processus stochastique à temps discret $(Y_k)_{k\in\mathbb{N}}$, tel que la loi du processus peut être bien expliquée par une chaîne de Markov non observable $(X_k)_{k\in\mathbb{N}}$, prenant des valeurs dans un espace fini ou non. Conditionnellement à $(X_k)_{k\in\mathbb{N}}$, les observations $(Y_k)_{k\in\mathbb{N}}$ sont une séquence de variables aléatoires indépendantes, et la loi de Y_n conditionnellement à $(X_k)_{k\in\mathbb{N}}$ est celle de Y_n conditionnellement à X_n :

$$P(Y_1,\ldots,Y_n/X_1,\ldots,X_n) = \prod_{k=1}^{n} P(Y_k/X_k) = \prod_{k=1}^{n} f_{X_k}(Y_k)$$

Ainsi, la loi marginale de dimension 1 du processus, c'est-à-dire la loi de Y_n est un mélange de populations. Dans le cas où la chaîne cachée est à nombre d'états finis m, la loi de Y_n est un mélange de m populations :

$$P(Y_n) = \sum_{l=1}^{m} P(Y_n/X_n = x_l)P(X_n = x_l)$$

Les observations peuvent donc être vues comme suivant un "mélange à régime markovien" et l'étude de ces modèles est une suite logique à l'étude des mélanges à observations i.i.d. que nous avons vue précédemment.

1.4.1 Les applications

Nous trouvons des applications dans des domaines très variés, comme par exemple :

- Reconnaissance de la parole (Rabiner (1989)). Chaque mot est une chaîne de Markov qui a ses propres états qu'on ne connaît pas. En revanche, toutes les chaînes de Markov ont le même nombre d'état (il semblerait que six états soit un optimum). Le mot est d'abord acquis par le processus suivant : 300 échantillons de 10 ms sont enregistrés, puis on en retire des grandeurs caractéristiques, par exemple des informations sur le spectre. On observe alors le mot à reconnaître, et on cherche la meilleure séquence d'états a posteriori.

- Analyse de séquences d'ADN (Churchill (1992)). L'ADN est vu comme une mosaïque de segments homogènes appartenant à un petit nombre de classes distinctes. La probabilité d'observer une base particulière (Cytosine, Thymine, Adénine, Guanime) à un site donné de l'ADN dépend du type de segment auquel elle appartient. L'organisation sous-jacente des segments de l'ADN est supposée suivre une chaîne de Markov. Ainsi, les états cachés sont les segments, tandis que les bases sont les observations. Ce modèle permet un cadre général où des hypothèses sur la structure du génome peuvent être formulées et testées. L'utilisation des CMC en biologie suscite toujours un vif intérêt. On peut trouver dans le récent ouvrage de Durbin *et al.* (1998), des applications à l'alignement multiple de séquences, à la construction d'arbres phylogéniques ou à l'analyse de la structure de l'ARN.

- Les modèles MA bruités sont un cas particulier de CMC. Ils sont notamment utilisés pour modéliser des investigations sismiques, comme dans la prospection pétrolière. Le sol est modélisé par des réflecteurs (indices de réflexion des différentes couches) en différents sites du sous-sol. Des ondes sont envoyées dans le sous-sol, et on étudie leur retour, c'est-à-dire comment les réflecteurs, variable aléatoires cachées, ont agi : c'est l'observation (Y_n). Cette observation est supposée être la convolution de l'onde avec les réflecteurs du sous-sol, altérée par un bruit indépendant. Mendel (1983) a proposé une modélisation 1D, que Lavielle (1991) a étendue au 2D.

- Les chaînes de Markov cachées non homogènes : récemment, ce type plus complexe a été utilisé comme modèle de précipitations (Hughes, Guttorp et Charles (1999)). Les états cachés sont des états météorologiques, les observations sont les précipitations à différents lieux à un certain nombre de dates ainsi que les pressions à ces mêmes dates. Les précipitations sont considérées sommes indépendantes étant donné l'état météorologique. En revanche, les états météorologiques ne suivent pas un modèle de chaîne de Markov indépendant du temps, mais ils dépendent de l'état météorologique et de l'état des pressions atmosphériques précédents.

Il est aussi question d'ordre dans les CMC : combien la chaîne cachée a-t-elle d'états? MacDonald et Zucchini (1997) donnent des applications qui cherchent à estimer ou tester ce nombre d'états cachés. Nous pouvons y trouver entre autres :

- Durée des éruptions successives du geyser Old Faithful. En 1990, Azzalini et Bowman (1990) ont d'abord proposé un modèle de chaîne de Markov du premier ordre, qui ne s'ajustait pas bien aux données, puis un modèle de chaîne de Markov du second ordre (la probabilité de transition dépend des deux états précédents), qui s'ajustait mieux, mais pour lequel ils n'arrivaient pas à trouver d'interprétations géophysiques. MacDonald et Zucchini ont utilisé les critères AIC et BIC pour comparer avec des CMC à deux, trois et quatre états. Deux modèles se sont détachés de l'ensemble : la chaîne de Markov du second ordre et le CMC à deux états, qui ont tous les deux 2 paramètres. Le modèle CMC à deux états semble plus simple à expliquer, mais il lui manque encore une explication géophysique convaincante.
- Dénombrement des mouvements d'un foetus de mouton. Leroux et Putterman (1992) ont eux aussi utilisé un estimateur du maximum de vraisemblance compensée pour comparer des modèles de mélange de lois de Poisson de 1 à 4 composantes, et des CMC dont la marginale d'ordre 1 sont des mélanges de lois de Poisson de 1 à 4 composantes. Leurs simulations ont montré que le critère BIC préférait le CMC à deux composantes, ce qui s'explique par deux états du foetus, la veille et le sommeil, et par le fait que les observations dans un court intervalle de temps sont reliées.

1.4.2 Etat de l'art

Gilbert (1959) s'est d'abord intéressé aux problèmes d'identifiabilité dans le cas particulier des fonctions déterministes de chaîne de Markov à états discrets. Puis Baum et Petrie (1969) ont démontré la consistance et la normalité asymptotique de l'estimateur du maximum de vraisemblance pour les fonctions stochastiques de chaîne de Markov à valeurs discrètes telles que Y_n/X_n soit i.i.d. de loi discrète; dans ce cadre, Vandekerkhove (1997) a démontré les propriétés LAN des fonctions stochastiques de chaînes de Markov dans le cas non-stationnaire. Baras et Finesso (1991) ont démontré la consistance de l'estimateur du maximum de vraisemblance compensée de l'ordre (nombre d'états cachés) dans le cas stationnaire, en imposant une restriction forte sur l'espace des paramètres, de manière à empêcher les problèmes de non-identifiabilité quand on surestime le nombre d'états de la chaîne cachée.

Leroux (1992b) a considéré les CMC à chaîne discrète et telles que (Y_n/X_n) soit de loi continue, et a montré, dans un espace quotient, la consistance de l'estimateur du maximum de vraisemblance quand l'ordre n'est pas connu, mais n'en a pas donné la vitesse.

A ordre fixé, Bickel et Ritov (1996) ont montré la propriété LAN du modèle CMC, Bickel, Ritov et Ryden (1998), la normalité asymptotique de l'estimateur du maximum de vraisemblance , et Mevel (1997), la normalité asymptotique du maximum de vraisemblance quand la chaîne n'est pas générée sous la loi stationnaire. Ryden (1994), de son côté, a obtenu la consistance et la normalité asymptotique d'un estimateur du maximum d'une pseudo-vraisemblance, découpant les observations en paquets de taille m et considérant que les paquets sont indépendants. Poursuivant dans cette direction, Ryden (1995) a montré que l'estimateur du maximum de pseudo-vraisemblance compensée ne sous-estimait pas l'ordre.

1.4.3 Première contribution

Avec O. Bermond (Bermond et Keribin (1999)), nous avons étudié le modèle MA(q) bruité où la source est à états continus et où l'ordre q est fixé :

$$Y_n = \sum_{l=0}^{q} b_0 X_{n-l} + \varepsilon_n$$

Le processus MA(q) bruité a une représentation markovienne. $(Y_t)_{t\in\mathbb{Z}}$ est une CMC où le processus markovien caché est $(X_n, X_{n-1}, \ldots, X_{n-q})$ et la loi de $(Y_n/(X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_{n-q} = x_{n-q}))$, l'observation conditionnellement à la chaîne cachée, est la loi du bruit (ε_n) décentrée de la valeur $\sum_{l=0}^{q} b_0 x_{n-l}$. Les nombreux résultats rappelés précédemment s'appliquent donc au cas MA bruité si le processus caché est discret, mais on en sait beaucoup moins quand la chaîne cachée est à états continus. Ainsi, reprenant l'idée de la pseudo-vraisemblance de Ryden (1994), nous avons démontré dans ce cas la consistance et la normalité asymptotique de l'estimateur de pseudo-vraisemblance. D'autre part, nous avons proposé une condition générale d'identifiabilité pour les processus MA bruités utilisant des propriétés des fonctions caractéristiques. Ainsi, nous montrons que si la loi de X n'est pas Gaussienne, et si elle possède un moment d'ordre 1 fini, alors le paramètre $\mathbf{b} = (b_0, b_1, \ldots, b_q)$ est identifiable à un facteur d'échelle près à partir d'une marginale m-dimensionnelle de Y, pour $m > q$. De plus, les lois de X et ε sont identifiables à un facteur de translation près. Ce résultat étend notablement celui de Giannakis et Swami (1990), supposant l'existence de cumulants d'ordre au moins 3. Enfin, nous proposons un algorithme et des simulations numériques dans le cas où la loi de X est un mélange de gaussiennes. Ceci nous permettra, par un algorithme EM, de réestimer conjointement les paramètres d'intérêt (le filtre $\mathbf{b}$ et la variance du bruit ε) ainsi que les paramètres de nuisance (ceux du mélange de gaussiennes). La pseudo-vraisemblance montre un autre de ses atouts lors de ces simulations : elle permet en effet de diminuer considérablement les calculs. En effet, supposons que le mélange ait K composantes et qu'on ait fait S paquets de m observations. Avec la pseudo-vraisemblance, les densités a posteriori nécessaires à calculer dans la phase d'Estimation demanderont le calcul de $S * K^{m+q}$ termes seulement, à comparer avec K^{S*m+q} termes dans le cas de la vraisemblance complète.

1.4.4 Deuxième contribution

Les résultats sur la détermination de l'ordre sont pour l'instant très incomplets. Ceci est du à la difficulté intrinsèque du modèle, que nous avons déjà rencontrée dans le cas des mélanges i.i.d. : les conditions usuelles de régularité ne sont pas remplies et la théorie du χ^2 ne s'applique pas. Dans cette partie, réalisée avec E. Gassiat (Gassiat et Keribin (1999)), nous avons réutilisé l'intuition et les méthodes de la paramétrisation localement conique de Dacunha-Castelle et Gassiat (1997); mais il est apparu que le calcul de la loi asymptotique dans le cadre général est beaucoup plus compliqué que sous un régime i.i.d.. Appliquant la méthode à un sous problème, nous avons obtenu des résultats intéressants. Ainsi, considérant le cas i.i.d. contre CMC à deux états, nous avons re-paramétré le modèle et sommes arrivés aux résultats suivants :

- Lorsque les paramètres sont restreints à un sous-ensemble des paramètres sur lequel la covariance du rapport de vraisemblance ne s'annule pas, la statistique du rapport de vraisemblance converge vers la moitié du supremum du carré de la partie positive d'un processus gaussien.
- Quand on fait croître l'espace des paramètre vers un certain domaine limite, la covariance d'un nombre infini de membres du processus gaussien tend vers 0, ce qui fait que la statistique du rapport de vraisemblance tend vers l'infini.

Nous avons déjà parlé de convergence vers l'infini du rapport de vraisemblance dans le cas des mélanges i.i.d., par exemple quand l'ensemble des paramètres n'est pas borné, ou quand la famille des composantes du mélange n'est pas régulière. Ici, l'espace des paramètres est compact et la famille est régulière. On peut alors comparer à un résultat similaire, qui découle du Théorème 4.5 de Dacunha-Castelle et Gassiat (1999) que la statistique du rapport de vraisemblance de l'ordre d'un modèle ARMA tend vers l'infini quand les pôles et racines ne sont pas restreints à un sous-ensemble compact du disque unité ouvert. Des simulations nous ont permis de voir un accroissement du rapport de vraisemblance, mais à une vitesse très lente, qui n'est pas sans rappeler la vitesse en $\frac{1}{2} \log \log(n)$ diagnostiquée par Hartigan (1985) dans le cas d'un mélange i.i.d. de deux Gaussiennes.

1.4.5 Perspectives de recherche

L'étude de la vitesse de convergence de la statistique du rapport de vraisemblance vers l'infini et sa loi asymptotique sont des points très intéressants; de même, l'étude sur le domaine complémentaire du domaine limite ci-dessus mentionné est à faire, ainsi que l'extension à des tests p contre q états cachés. D'autre part, on peut envisager d'utiliser la méthode employée par Dacunha-Castelle et Gassiat (1997b) pour les mélanges finis à loi discrète à l'étude de l'ordre de CMC.

1.5 Algorithmes et simulations

Dans la pratique, nous allons rencontrer deux types de questions :

- Comment arriver à calculer l'estimateur du maximum de vraisemblance d'un échantillon? Le point est d'autant plus délicat lorsque l'on surparamètre, car la vraisemblance a alors de nombreux maxima locaux et points selle.
- Comment essayer de tabuler les lois limites que nous avons rencontrées sous forme de supremum du carré d'un processus Gaussien tronqué?

C'est le sujet de cette section.

1.5.1 Calcul de l'estimateur du maximum de vraisemblance

Ce problème est un problème d'optimisation de la log-vraisemblance, ou de recherche de 0 de sa dérivée :

$$\nabla_\theta l_n(\theta) = 0;$$

La première approche, est d'utiliser une *méthode de Newton*, où l'étape $k+1$ est définie en fonction des résultats de l'étape k par :

$$\theta^{(k+1)} = \theta^{(k)} - [H(\theta^{(k)})]^{-1}\nabla_\theta l_n(\theta^{(k)})$$

avec H le Hessien de $l_n(\theta)$. La méthode de Newton converge très vite, mais

- Elle dépend des conditions initiales
- Elle nécessite le calcul du Hessien à chaque pas, et de son inversion, ce qui peut être très lourd en taille mémoire et temps de calcul si la dimension du paramètre est grande.

Afin de pallier le problème du calcul du Hessien, la *méthode de quasi-Newton* en remplace le calcul exact par une approximation B à chaque itération.

$$\theta^{(k+1)} = \theta^{(k)} - B^{-1}\nabla_\theta l_n(\theta^{(k)})$$

Plus rapide, cette méthode est cependant plus délicate à utiliser, car il ne faut pas que B soit trop loin du calcul exact pour conserver les propriétés de convergence. Une forme particulière pour B est de la choisir telle que : $B = -nI(\theta^{(k)})$ où $I(\theta^{(k)})$ est l'information de Fisher du modèle sous le paramètre $\theta^{(k)}$, simplement approximée par une loi des grands nombres. L'algorithme ainsi défini est la *méthode de score*. Ces méthodes, simples du point de vue théorique, ont cependant un inconvénient important : elles n'assurent pas forcément l'accroissement de la vraisemblance à chaque itération, comme l'ont noté Redner et Walker (1984).

Les *algorithmes récursifs* fournissent une autre voie d'approche. Ils font partie de la classe des algorithmes à dynamique Markovienne introduit par Benveniste, Métivier et Priouret (1987), de la forme :

$$\theta^{(k+1)} = \theta^{(k)} + \eta_k L(\theta^{(k)}, Z_k(\theta^{(k)}))$$

où $Z_k(\theta)$ est un processus de Markov dont la loi de transition d'un état à un autre dépend de θ, et éventuellement d'un autre paramètre. Voir également Duflo (1990) pour un ouvrage de référence sur les algorithmes stochastiques. Remarquons qu'on peut s'inspirer des méthodes du gradient et de Newton pour chercher un estimateur récursif du maximum de vraisemblance, ce qui donne, pour l'estimateur de Newton récursif :

$$\theta^{(k+1)} = \theta^{(k)} + \frac{1}{k}[H(\theta^{(k)})]^{-1}\nabla_\theta g(\hat{\theta}^{(k)}, X_{k+1})/g(\hat{\theta}^{(k)}, X_{k+1})$$

Ainsi, pour des CMC, Ryden (1997) a étudié un estimateur récursif de la pseudo-vraisemblance et Mevel (1997) des estimateurs récursifs du maximum de vraisemblance et des moindres carrés. Ces estimateurs ont eux aussi le problème du choix de la condition initiale, qui, si elle est mal choisie, peut entraîner l'estimateur vers un minimum local de la fonction à minimiser.

Reprenant l'estimation de la densité d'un mélange comme un problème d'estimation des données incomplètes, nous pouvons relier la recherche du maximum de vraisemblance à l'*algorithme EM* et à l'approche bayésienne. En effet, maximiser la log-vraisemblance des données observées (X_t) revient à maximiser l'espérance conditionnelle sous la loi a posteriori du logarithme de la vraisemblance $c(X_{1:n}, \Phi_{1:n}, \theta)$ des données complètes (X_t, Φ_t) :

$$Q(\theta|\theta^{(k)}) = E(\log c(X_{1:n}, \Phi_{1:n}, \theta)|X_{1:n}, \theta^{(k)})$$

L'algorithme EM procède donc en deux étapes itérativement :

- Etape E : Espérance : calcul de $Q(\theta|\theta^{(k)})$
- Etape M : Maximisation de $Q(\theta|\theta^{(k)})$: $\theta^{(k+1)} \in \arg\max_\theta Q(\theta|\theta^{(k)})$

La convergence de l'algorithme est assurée par l'augmentation de la vraisemblance à chaque pas. D'autre part, il est évidemment bien adapté quand il est difficile de maximiser la vraisemblance directement, mais facile de travailler avec les données complètes. Enfin, il faut noter les problèmes des conditions initiales. Voir Dempster, Laird et Rubin (1977), ou Redner et Walker (1984). Le même cadre est valide pour l'identification de modèle ARMA bruité ou le cas des CMC où les observations sont les Y_t et les données manquantes la chaîne de Markov cachée X_t. La principale difficulté de ce schéma est le calcul direct de l'espérance des données manquantes sous la loi a posteriori, qui est souvent impossible. Quand la source est discrète cependant (cas de nos CMC par exemple), elle peut être calculée efficacement en utilisant la procédure de *Forward-backward* (Rabiner (1989)).

Des variantes stochastiques ont été introduites, qui demandent le tirage des données manquantes sous la loi a posteriori. Ceci, (1) parce qu'on ne sait par forcément calculer analytiquement l'espérance des données manquantes sous la loi a priori d'une part, et (2) parce qu'on voulait mieux ajuster l'estimation du paramètre en prenant en compte une estimée de ces données manquantes.

- *MCEM: Monte Carlo EM.* Il répond à la première motivation et donne une procédure d'estimation de $Q(\theta|\theta^{(k)})$. La k-iéme étape E est remplacée par (1) le tirage de $M(k)$ réalisations indépendantes des données cachées sous la probabilité a posteriori des données manquantes conditionnellement aux observations et à l'estimation courante du paramètre, (2) une intégration de Monte Carlo (moyenne empirique de la log vraisemblance des données complètes).

  $$\hat{Q}(\theta|\theta^{(k)}) = \frac{1}{M(k)} \sum_{j=1}^{M(k)} \log c(X_{1:n}, \Phi_{1:n}^{(j)}, \theta^{(k)})$$

  Notons que le MCEM n'accroît pas de façon déterministe la vraisemblance à chaque itération. Les seuls résultats de convergence sont obtenus quand $M(k)$ tend vers l'infini (ce qui équivaut alors à un EM).

- *SEM: Stochastic EM.* L'idée est de remplacer, à chaque itération, le calcul et la maximisation de $Q(\theta|\theta^{(k)})$ par le calcul plus simple de la loi a posteriori $k(\Phi_{1:n}|X_{1:n}, \theta^{(k)})$ et la simulation sous cette loi des données manquantes $\hat{\Phi}_{1:n}$. $\theta^{(k+1)}$ est choisi comme maximisant la vraisemblance complète $c(X_{1:n}, \hat{\Phi}_{1:n}, \theta)$. De ce fait, il n'y a plus de mécanisme de stabilisation assurant la convergence de la suite des estimés vers une valeur déterministe. Celeux, Chauveau et Diebolt (1996) ont montré quelques résultats de convergence dans le cas de mélange de gaussiennes, mais qui ne s'appliquent pas aux modèles de convolution.

- *SAEM: "Simulated Annealing" EM.* Afin d'éviter le comportement erratique du SEM, ces auteurs essaient de le stabiliser en choisissant $\theta^{(k+1)}$ comme la combinaison convexe d'une version de l'EM et d'une version du SAEM, avec un pas η_k

décroissant vers 0 (l'analogie avec le recuit simulé s'arrêtant à la décroissance de la température). L'algorithme privilégie donc de plus en plus la version déterministe :

$$\theta^{(k+1)} = (1 - \eta_{k+1})\theta^{(k)}_{EM} + \eta_{k+1}\theta^{(k)}_{SEM}$$

Comme pour le SEM, il y a peu de résultats théoriques de convergence, et les problèmes de convergence vers un minimum local subsistent.

- *SAEM : "Stochastic Approximation" EM* (Lavielle *et al* (1998)). A ne pas confondre avec le précédent, il réalise une approximation stochastique pour mettre à jour l'estimée courante de $Q(\theta|\theta^{(k)})$.

$$\hat{Q}(\theta|\theta^{(k)}) = \hat{Q}(\theta|\theta^{(k-1)}) + \eta_k(\frac{1}{M(k)}\sum_{j=1}^{M(k)} \log c(X_{1:n}, \Phi^{(j)}_{1:n}, \theta^{(k)}) - \hat{Q}(\theta|\theta^{(k-1)}))$$

où η_k est une suite de pas positifs. Avec $\eta_k = 1$, on retrouve un algorithme MCEM. L'intérêt de l'algorithme Stochastic Approximation EM est qu'il peut réutiliser le cadre général des algorithmes d'approximation stochastique pour déterminer les conditions sur la suite de pas, afin d'offrir la convergence vers un maximum local de la vraisemblance. Voir Cappé *et all* (1999) pour son utilisation dans le cas de l'identification d'un filtre linéaire bruité (dont le modèle MA bruité que nous avons déjà rencontré est un exemple).

Le calcul de l'espérance conditionnelle demande de pouvoir simuler la loi a posteriori des données manquantes conditionnellement aux observations. Si cette loi est très complexe, on peut choisir une méthode indirecte pour en obtenir des échantillons : on construit une chaîne de Markov irréductible et apériodique dont la probabilité stationnaire est la loi à simuler, qu'on fait fonctionner suffisamment longtemps. Ce sont les algorithmes de *Metropolis-Hastings* ou d'*échantillonnage de Gibbs*.

Tous les algorithmes présentés ont un inconvénient commun : ils ont besoin d'une bonne condition initiale pour parvenir au maximum global. Dans le cas des mélanges i.i.d., Chuang et Mendell (1997) et Leroux et Putterman(1992) ont constaté qu'un grand nombre de valeurs initiales aléatoirement choisies ne donnait pas de meilleurs résultats qu'un petit nombre de valeurs systématiques. Furman et Lindsay (1994) ont constaté que les estimateurs basés sur les moments donnent de bonnes valeurs initiales. Enfin, Dunmur et Titterington (1998) ont montré que dans le cas d'une chaîne de Markov cachée à deux états, le choix des conditions initiales ont une grande influence sur l'atteinte du maximum global.

Pour avoir des méthodes insensibles à l'initialisation, on peut avoir recours au *recuit simulé* (Guyon (1993) donne une bonne revue de ces méthodes). Le principe est de construire une chaîne de Markov non homogène dont la loi limite est la loi uniforme sur les valeurs des paramètres réalisant le maximum de vraisemblance. Ainsi, les sauts de la chaîne permettent de balayer tout l'espace des paramètres. Les différentes transitions de la chaîne sont liées à un schéma de refroidissement, en ne permettant que des sauts de plus en plus petits afin de converger vers le maximum global. Vandekerkhove (1997) a utilisé ce principe pour maximiser une pseudo-vraisemblance d'une chaîne de Markov cachée.

1.5.2 Loi limite du rapport de vraisemblance

La loi limite du rapport de vraisemblance est, dans de nombreux cas que nous avons rencontrés, proportionnelle au supremum du carré de la partie positive d'un processus gaussien. Azaïs et Wschebor (1997), puis Azaïs, Cierco-Ayrolles et Croquette (1998) ont donné des formules pour calculer la loi du supremum d'un processus gaussien stationnaire. Dietrich et Newsam (1997) ont proposé un algorithme de simulation d'un tel processus. Dans notre cas cependant, le processus gaussien n'est pas stationnaire.

Berdaï et Garel (1996) ont donné une tabulation du niveau et de la puissance du test de rapport de vraisemblance de l'absence de mélange contre un mélange à deux composantes pour des normales à moyennes inconnues γ_1 et γ_2 et variance commune inconnue σ. Leurs simulations ont mis en évidence la dépendance des résultats avec l'intervalle de variation de $d = \frac{|\gamma_1 - \gamma_2|}{\sigma}$: l'approximation admise que la loi est entre un $\chi^2(1)$ et un $\chi^2(2)$ n'est valable que pour un d pas trop grand, corroborant les résultats de Hartigan (1980). Atwood *et al* (1996) ont investigué la loi du même rapport de vraisemblance. Ils ont simulé 3200 trajectoires, chacune représentant un échantillon de taille 256000. Ils ont trouvé que la convergence était extrêmement lente. Ils ont de plus trouvé que le poids de la loi en 0 n'était pas nul, ceci pouvant venir d'une recherche non exhaustive du maximum, ou d'une manifestation de l'"effet Hartigan", mais que le $\chi^2(2)$ correspondait à une approximation à 15% de la loi limite. Enfin, Chuang et Mendell (1997) ont eux aussi constaté des convergences très lentes du rapport de vraisemblance dans le cas d'un mélange de deux normales bivariées. Notons enfin l'utilisation de techniques de bootstrap par Basford et McLachlan (1985)), McLachlan(1987) et plus récemment Feng et McCulloch (1996), le principe étant le suivant : à partir de l'échantillon à tester, on estime les paramètres maximisant la vraisemblance sous l'hypothèse nulle. Puis, un échantillon de bootstrap est généré à partir de la loi ayant les paramètres estimés précédents, et le rapport de vraisemblance est calculé après avoir ajusté le modèle sous chacune des hypothèses. Ce processus est alors répété K fois, et les différentes valeurs du rapport de vraisemblance formées à partir des échantillons de bootstrap successifs fournissent une évaluation du bootstrap, et donc de la vraie distribution du rapport de vraisemblance.

1.5.3 Contribution

Des simulations ont été menées pour l'identification du modèle MA bruité, en utilisant un EM, quand la source est un mélange de gaussiennes. Ce cas est intéressant, car on peut introduire de nouvelles variables cachées, celles des coefficients de mélanges. De plus, l'utilisation d'une pseudo-vraisemblance a permis de considérablement alléger le temps de calcul, mais qui reste cependant un problème quand le nombre de composantes de la source dépasse 4.

Nous avons voulu investigué, pour un mélange de deux gaussiennes, l'influence du terme de compensation de l'estimateur du maximum de vraisemblance compensée. Pour ce faire, nous avons calculé le maximum de vraisemblance dans un modèle d'ordre donné par un algorithme EM, en partant d'une grille de points initiaux, puis appliqué différentes compensations. Nous avons pu voir que plus la compensation est forte, plus l'estimateur met de temps à converger, en particulier quand les deux composantes sont peu discernables (un seul mode visible). Le cas du critère BIC donne d'excellents résultats. Cependant, quand le nombre de composantes augmente, le temps de calcul devient très important, ce qui demeure un handicap pour cette méthode. Toutefois, elle donne de meilleurs résultats que la méthode des moments (Dacunha-Castelle et Gassiat (1997b)) qui est cependant beaucoup plus rapide en temps d'éxécution.

Enfin, nous avons étudié numériquement le comportement asymptotique du rapport de vraisemblance dans le cas du mélange "à régime markovien". Pour calculer le maximum de vraisemblance, nous avons privilégié ici une méthode de Newton, étant donné le peu de paramètres et le fait, vérifié sur un grand nombre d'échantillons, que la log-vraisemblance a un unique maximum dans l'intervalle d'étude. D'ailleurs, d'autres simulations avec un algorithme EM et une grille de points de départ, n'ont pas donné de meilleurs résultats. Nous avons constaté que la convergence vers le processus limite est très lente, que l'on restreigne l'espace des paramètres ou non. Dans ce dernier cas, le rapport de vraisemblance tend théoriquement vers l'infini. Nos simulations montrent effectivement un très lent accroissement du rapport de vraisemblance, qui atteint 2 pour une taille d'échantillon de $10,000$. Ces résultats peuvent être rapprochés de ceux de Hartigan (1985), qui a donné, pour un mélange i.i.d. de deux Gaussiennes, une vitesse en $\frac{1}{2}\log\log n$.

Bibliographie

[1] Akaike H. Information theory and an extension of the maximum likelihood principle. *2nd International Symposium on Information Theory, eds B.N. Petrov and F. Czàki, Akademiai Kiadò, Budapest*, 267–281, 1973.

[2] Atwood L D, Wilson A F, Bailey-Wilson J E, Carruth J N, Elston R C.. On the distribution of the likelihood ratio test statistic for a mixture of two normal distributions. *Commun. Statist.-Simula.*, 25(3): 733–740, 1996.

[3] Azaïs J-M, Cierco-Ayrolles C, Croquette A.. Bounds for the distribution of the maximum of a smooth stationary gaussian process. *ESAIM: Proba. and Stat.*, soumis, 1998.

[4] Azaïs J-M, Wschebor M. Une formule pour calculer la distribution du maximum d'un processus stochastique. *C.R.A.S., série I, Paris.*, 324: 225–230, 1997.

[5] Azzalini A, Bowman A W. A look at some data on the Old Faithful geyser. *Applied Stat.*, 39: 357–365, 1990.

[6] Azencott R, Dacunha-Castelle D. Séries d'observations irrégulières, modélisation et prévision. *Masson, Paris*, 1984.

[7] Baum L E, Petrie T. Statistical inference for probabilistic functions of finite state markov chain. *Ann. MAth. Stat.*, 37: 1554–1563, 1966.

[8] Baras J S, Finesso L. Consistent estimation of the order of hidden markov chains. *Workshop 26-28 sept 1991, Lawrence, Kensas*, 1991.

[9] Basford K E, McLachlan G J. Likelihood Estimation with Normal Mixture Models. *Appl. Statist.*, 34(3): 282–289, 1985.

[10] Benveniste A, Metivier M, Priouret P. Algorithmes adaptatifs et approximations stochastiques. *Masson, Paris*, 1987.

[11] Berdaï A, Garel B. Detecting univariate normal mixture with two components. *Statistics et Decisions*, 14: 35–51, 1996.

[12] Bermond O, Keribin C. M-likelihood estimation for noisy processes. *J. Math. Syst. and Control*, soumis, 1999.

[13] Bickel P J, Chernoff H. Asymptotic Distribution of the likelihood Ratio Statistic in a Prototypical Non Regular Problem. *Statistics and Probabilities*, 83–96, 1993.

[14] Bickel P J, Ritov Y. Inference in hidden Markov models I: Local asymptotic normality in the stationnary case. *Bernouilli*, 2: 199–228, 1996.

[15] Bickel P J, Ritov Y, Ryden T. Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Ann. Statist.*, 26: 1614–1635, 1998.

[16] Brown L. Fundamentals of statistical exponential families. Lecture Notes-monograph series. *Institute of Mathematical Statistics, Hayward, California*, 1986.

[17] Cappé O, Doucet A, Lavielle M, Moulines E. Simulation-Based Methods for Blind Maximum-Likelihood Filter Identification. *Signal Processing*, 73: 3–24, 1999.

[18] Cassie R M. Some uses of probability paper in the analysis of size frequency distributions. *Austral. J. Mar. Fish. Freshw. Res.*, 5: 513–322, 1954.

[19] Celeux G, Chauveau D, Diebold J. Stochastic versions of the EM algorithm: an experimental study in the mixture case. *J. Statist. Comput. Simul.*, 55: 287–314, 1996.

[20] Chen J. Optimal rate of convergence for finite mixture models. *Ann. Statist.*, 23(1): 221–233, 1995.

[21] Chernoff H, Lander E. Asymptotic distribution of the likelihood ratio test that a mixture of two binomials is a single binomial. *Journal of statistical planning and inference*, 43: 19–40, 1995.

[22] Chuang R-J, Mendell N R. The approximate null distribution of the likelihood ratio test for a mixture of two bivariate normal distributions with equal covariance. *Commun. Statist.-Simula.*, 26(2): 631–648, 1997.

[23] Churchill G A. Hidden markov chains and the analysis of genome structure. *Computers Chem.*, 16(2): 107–115, 1992.

[24] Dacunha-Castelle D, Duflo M. Probabilités et statistiques: 2- Problèmes à temps mobile. *Masson, Paris*, 1993.

[25] Dacunha-Castelle D, Gassiat E. Testing in locally conic models, and application to mixture models. *ESAIM Prob. and stat., http: //www.emath.fr/ps/*,volume 1, 285–317, 1997.

[26] Dacunha-Castelle D, Gassiat E. The estimation of the order of a mixture model. *Bernoulli*, 3: 279–299, 1997b.

[27] Dacunha-Castelle D, Gassiat E. Testing the order of a model using locally conic parameterization: population mixtures and stationary ARMA processes. *Ann. Statist.*, à paraître, 1999.

[28] Delyon B, Lavielle M, Moulines E. Convergence of a stochastic approximation version of the EM algorithm. *Annals of Stat.*, 27(1),1999.

[29] Dembo A, Shao Q-M. Self-normalized moderate deviations and lils. *Stoch. Proc. and Appl.*, 75: 51–65, 1998.

[30] Dempster A P, Laird N M, Rubin D B. Large Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39: 1–38, 1977.

[31] Dietrich C R, Newsam G N. Fast and exact simulations of stationary gaussian processes through circulant embedding of the covariance matrix. *SIAM J. Sci. Comput.*, 18(4): 1088–1107, 1997.

[32] Duflo M. Méthodes récursives aléatoires. *Masson, Paris*, 1990.

[33] Dunmur A P, Titterington D M. The influence of initial conditions on maximum likelihood estimation of the parameters of binary hidden Markov model. *Stat. and Proba. Letters*, 40: 67–73, 1998.

[34] Durbin R, Eddy S R, Krogh A, Mitchison G J. Biological sequence analysis. Probabilistic models of proteins and nucleic acids. *Cambridge University Press, Cambridge UK*, 1998.

[35] Everit B S, Hand D J. Finite Mixture Distributions. *Monographs in Applied Probability and Statistics, Chapman and Hall, Londres*, 1981.

[36] Feng Z D, McCulloch C E. Using bootstrap Likelihood Ratio in Finite Mixture Models. *J. R. Statist. Soc. B*, 58(3): 609–617, 1996.

[37] Furman W D, Lindsay B G. Measuring the relative effectiveness of moment estimators as starting values in maximizing likelihoods. *Comput. Stat. and Data Analysis*, 17: 493–507, 1994.

[38] Gassiat E, Keribin C. The likelihood ratio test for the number of components in a mixture with Markov regime. , manuscrit, 1999.

[39] Giannakis G B, Swami A. On estimating Noncausal Nonminimum Phase ARMA Models of Non-Gaussian Processes. *IEEE Transactions on acoustics, speech and signal processing*, 38(3): 478–495, 1990.

[40] Gilbert E J. On the identifiability problem for functions of finite Markov chains. *Ann. Math. Statist.*, 30: 688–697, 1959.

[41] Gilbert J S L. The developpement of taxonomic theory since 1851. *Nature*, 168: 400–402, 1951.

[42] Gordon A D. Classification. *Chapman and Hall, Londres*, 1981.

[43] Hannan E J. The estimation of the order of an ARMA process. *Ann. Stat.*, 8(5),1071–1081, 1980.

[44] Hannan E J, Quinn B G. The determination of the order of an autoregression. *J.R. Statist. Soc. B*, 69: 190–195, 1979.

[45] Hartigan J A The estimation of the order of an ARMA process. *Ann. Stat.* , 8: 1071–1081, 1980.

[46] Hartigan, J.A. A failure of likelihood ratio asymptotics for normal mixtures. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L.M. Le Cam and R.A. Olshen eds, 807-810, 1985.

[47] Hasselblad V. Estimation of Finite Mixture of Distributions from the exponential Family. *Journal of the American Statistical Association*, 64: 1459–1471, 1969.

[48] Haughton D. On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16: 342–355, 1988.

[49] Haughton D. Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā A*, 51: 45–58, 1989.

[50] Haughton D, Dudley R. Information criteria for multiple data sets and restricted parameters. *Stat. Sin.*, 7(2): 265–284, 1997.

[51] Hughes J P, Guttorp P, Charles S P. A non-homogeneous hidden Markov model for precipitation occurrence. *Appl. Statist.*, 48(1): 15–30,1999.

[52] Keribin C. Consistent Estimation of the Order of Mixture Models. *Sankhyā A*, soumis, 1997.

[53] Keribin C. Estimation consistante de l'ordre de modèles de mélange. *C. R. Acad. Sci. Paris, Série I*, 326(2), 1998.

[54] Keribin C., Haughton D. Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families. *J. Statis. Plan. and Inference*, soumis, 1999.

[55] Keribin C, Haughton D. On the rate of convergence of a test of contamination using maximum likelihood techniques. , manuscrit, 1999.

[56] Lavielle M. 2-D Bayesian deconvolution. *Geophysics*, 56(12),2008–2018,1991.

[57] Ledoux M. Sur les déviations modérées des sommes de variables vectorielles indépendantes de même loi. *Ann. Inst. Henri Poincaré*, 28(2): 267–280, 1992.

[58] Leroux B G. Consistent estimation of a mixing distribution. *Ann. of Statist.*, 20(3): 1350–1360, 1992.

[59] Leroux B G. Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. and their Appl.*, 40: 127–143, 1992.

[60] Leroux B G, Putterman M L. Maximum-Penalized-Likelihood Estimation for Independent and Markov-Dependent Mixture Models. *Biometrics*, 48: 545–558, 1992.

[61] Lindsay B G. Moment Models: Theory, Geometry and Applications. *NSF-CBMS Regional Conference Series in Probability and Statistics, IMS, Hayward, CA*, Vol. 5 1995.

[62] Lindsay B G. Moment matrices: application in mixtures. *Ann. of Statist.*, 17(2): 722–1340, 1983.

[63] Lindsay B G, Roedner K. Moment-based oscillation properties of mixture models. *Ann. of Statist.*, 25(1): 378–386, 1997.

[64] Lindsay B G, Lesperance M L. A review of semiparametric mixture models. *J. Stat. Plann. Inference*, 47(1,2): 29–39(1995).

[65] MacDonald I L, Zucchini W. Hidden Markov and Other Models for Discrete-valued Time Series. *Chapman and Hall, Londres*, 1997.

[66] McLachlan G J On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *App. Statist.*, 36(3): 318–324, 1987.

[67] McLachlan G J, Basford K E. Mixture Models: Inference and Applications to Clustering. *Marcel Dekker Inc. , New York*, 1988.

[68] Mendel J F. Optimal sismeic deconvolution. *Academic Press Inc*, 1983.

[69] Mevel L. Statistique asymptotique pour modèles de Markov cachés. *Thèse, Université de Rennes 1*, 1997.

[70] Nishii R. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12: 758–765, 1984.

[71] Rabiner L R. Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition. *Proceedings of the IIIE* 77(2): 257–286, 1989.

[72] Redner R A, Walker H F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26(2): 195–223, 1984.

[73] Roedner K. A Graphical Technique for Determining the Number of Components in a Mixture of normals. *American Stat. Assoc.* 89(426): 487–495, 1994.

[74] Ryden T. Consistent and asymptotically normal parameter estimates for hidden Markov models. *Ann. Statist.* 22(4): 1884–1895, 1994.

[75] Ryden T. Estimating the order of hidden markov models. *Statistics* 26: 345-354, 1995.

[76] Ryden T. On recursive estimation for hidden Markov models. *Stoch. Proc. and Appl.* 66: 79–96, 1997.

[77] Schwarz G. Estimating the dimension of a model. *Annals of Statistics* 6: 461–646, 1978.

[78] Senoussi D. Statistique asymptotique presque sûre de modèles statistiques convexes. *Ann. Inst. Henri Poincaré*, 26: 342–355, 1988.

[79] Shibata R. Selection of the order of an autoregressive model by Aikaike's criterion. *Biometrika*, 63: 117–126, 1978.

[80] Titterington D M, Smith A F M, Makov U E. Statistical Analysis of Finite Mixture Distributions. *Wiley series in probability and mathematical statistics, John Wiley ans sons, Londres*, 1985.

[81] Vandekerkhove P. Contribution à l'étude statistique des chaînes de Markov cachées. *Thèse, Université de Montpellier II*, 1988.

[82] Wilks S S. Mathematical statistics. *Wiley*, New York, 1967.

[83] Woodroofe M. On model selection and the arc sines laws. *Annals of Statistics*, 10(4): 1182–1194, 1982.

Chapitre 2

Probabilités de sous-estimer ou surestimer l'ordre d'un modèle de familles régulières

Ce chapitre détaille notre contribution à l'étude des modèles identifiables, ainsi que nous l'avons présentée dans la section 1.2.3 de la présentation générale. Ce travail, réalisé en collaboration avec Dominique Haughton, a été soumis après révision au *Journal of Statistical Planning and Inference* en mars 2001.

Dans cet article, nous considérons des modèles emboîtés, identifiables si l'ordre est bien ou surestimé.

Définissons $\mathcal{T}_p = \{\theta^p = (t_1 \dots t_p) \in T_1 \times \dots \times T_p\}$, où $T_i, i = 1, \dots, p$ sont des ensembles compacts de $\mathbb{R}^{r_i}$. La dimension du paramètre θ^p est $R_p = \sum_{i=1}^p r_i$ et $\theta^p = (\theta_1, \dots, \theta_{R_p})$. Nous considérons maintenant la suite de modèles paramétriques à densités par rapport à une mesure ν, emboîtés de la façon suivante :

$$(\mathcal{G}_p = g(\theta^p, .), \theta^p \in \mathcal{T}_p)_{p \leq Q} \ , \mathcal{G}_{p_1} \cap \mathcal{G}_{p_2} = \mathcal{G}_{p_1 \wedge p_2}$$

où $p_1 \wedge p_2 = \min(p_1, p_2)$. Nous définissons alors l'ordre réel q d'un modèle de densité $g \in \mathcal{G}_Q$ comme le plus petit nombre de t_i nécessaires à décrire le modèle. Q est un majorant connu de l'ordre du modèle. Cette définition peut bien sûr s'étendre au cas de modèles non emboîtés, mais qui ont une intersection non vide, et pour lesquels la loi d'intérêt appartient à l'intersection. Remarquons que l'ordre d'un modèle n'est pas forcément égal au nombre de paramètres qui gouvernent sa densité, mais est une fonction de ce nombre. Typiquement, l'ordre d'un modèle de mélange est le nombre de ses composantes distinctes, qui est inférieure au nombre total des paramètres qui gouvernent la loi du mélange.

Le modèle est identifiable si l'ordre est bien ou surestimé, c'est-à-dire : supposons que la loi g_0 du modèle étudié est d'ordre q et de paramètre donné θ_0^q. Alors, parmi toutes les lois $g(\theta^p)$ d'ordre p supérieur ou égal à q :

$$\forall p \geq q \ , \ \{g(\theta^p,.) = g_0(.)\} \Leftrightarrow \{\theta^p = (\theta_0^q, 0, \ldots, 0)\}$$

Dans ce cas, les tests de rapport de vraisemblance suivent la théorie du χ^2 (Wilks (1967)).

Nous savons, par les travaux de Senoussi (1988), que l'estimateur du maximum de vraisemblance compensée est consistant pour l'estimation de l'ordre du modèle. Nous voulons étudier les vitesses asymptotiques des probabilités de se tromper d'ordre en utilisant un tel estimateur dans le cas de familles régulières générales, étendant ainsi les résultats d'Haughton (1989) sur le critère BIC dans les familles exponentielles.

Nous montrons que, sous des hypothèses de moments exponentiels des dérivées de $\log g(\theta,.)$, la probabilité asymptotique de surestimer l'ordre du modèle est telle que :

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n > q)] = -1$$

et la probabilité $P(\hat{q}_n < q)$ de sous-estimer l'ordre est telle que :

$$\limsup_{n\to+\infty} \frac{1}{n} \log[P(\hat{q}_n < q)] < 0$$

pour des compensations $a_{n,p} = p \cdot v(n)$, satisfaisant $\lim_{n\to+\infty} v(n) = +\infty$ et $v(n) = o(n)$. Ainsi, la vitesse de l'erreur de sous-estimation est un résultat de grandes déviations, tandis que celui de la vitesse de l'erreur de surestimation est un résultat de déviations modérées.

Ces résultats découlent de l'étude des probabilités asymptotiques des rapports de vraisemblance :

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(W_n(q+k) - W_n(q) > 0)] = -k, \tag{1}$$

$$\limsup_{n\to+\infty} \frac{1}{n} \log[P(W_n(q-k) - W_n(q) > 0)] < 0. \tag{2}$$

Avant de donner le principe de la démonstration, remarquons que nous noterons la vraisemblance $l_n(\theta)$, en lieu et place de $l_n(g(\theta))$. Dans le cas de la surestimation, pour $p \geq q$, un développement de Taylor de $T_n(p) - l_n(\theta_0^p)$ en θ_0^p, et de sa dérivée par rapport à θ^p permet d'écrire, avec $\partial l_n(\theta^p)$ le gradient de $l_n(\theta^p)$ dans $\mathcal{T}_p$, T l'opérateur de transposition, et $J(\theta_0^p)$ la matrice d'information de Fisher en θ_0^p :

$$T_n(p) - l_n(\theta_0) = \frac{1}{2}\Big(\frac{1}{n}(\partial l_n(\theta_0^p))^T \cdot [J(\theta_0^p)]^{-1} \cdot \partial l_n(\theta_0^p)\Big)(1 + o(1))$$

L' application de ce résultat pour $p = q + k$ et $p = q$ nous conduit à :

$$T_n(p) - T_n(q)(\theta_0) = \frac{1}{2}\Big((S_n^k)^T \cdot \Sigma_k^{-1} \cdot S_n^k\Big)(1 + o(1))$$

où S_n^k est un vecteur de dimension k qui tend vers la loi normale centrée de matrice de covariance Σ_k. Ainsi, le terme principal de (1) est :

$$P\Big((S_n^k)^T \cdot \Sigma_k^{-1} \cdot S_n^k \geq 2kv(n)\Big)$$

qui est un terme de déviations modérées pouvant se traiter par le théorème 3.7.1 dans Dembo et Zeitouni (1993) :

Théorème 2.0.1 *Soient $X_1, \ldots, X_n$ une séquence i.i.d. de vecteurs aléatoires de $\mathbb{R}^k$ telle que $\log E[e^{<\lambda, X_i>}] < \infty$ dans un voisinage de l'origine, $E(X_i) = 0$. On suppose que Σ, la matrice de covariance de X_1 est inversible. Fixons $a_n \to 0$ tel que $na_n \to \infty$ quand $n \to \infty$. Soit $Z_n \stackrel{\Delta}{=} \sqrt{a_n/n}\sum_{i=1}^n X_i$. Alors, pour tout ensemble mesurable Γ,*

$$\begin{aligned} -\frac{1}{2}\inf_{x\in\Gamma^o} < x, \Sigma^{-1}x > \quad &\leq \quad \liminf_{n\to\infty} a_n \log P(Z_n \in \Gamma) \\ &\leq \quad \limsup_{n\to\infty} a_n \log P(Z_n \in \Gamma) \\ &\leq \quad -\frac{1}{2}\inf_{x\in\bar{\Gamma}} < x, \Sigma^{-1}x > \end{aligned}$$

Quant aux restes, on montre que ce sont des termes négligeables en grandes déviations sur $\hat{\theta}$ et de grandes déviations banachiques sur des lois des grands nombres. Les théorèmes utilisés sont celui de Mogulskii (1988) pour les déviations sur l'estimateur du maximum de vraisemblance quand le paramètre appartient à un fermé borné de $\mathbb{R}^{R_Q}$.

Théorème 2.0.2 *Supposons que pour tout réel $N < +\infty$,*

$$\limsup_{r\to 0} E_{g_0\nu}(e^{N\omega(r)})$$

est fini, où $\omega(r)$ est le module de continuité de $\log g(\theta, x)$:

$$\omega(r) = \sup_{\theta_1,\theta_2\in\mathcal{T}_p, ||\theta_1-\theta_2||<r} |\log g(\theta_1, x) - \log g(\theta_2, x)| \,.$$

Alors, pour tout fermé $\Gamma \in \mathcal{T}_p$, de complémentaire Γ^c dans $\mathcal{T}_p$:

$$\limsup_{n\to\infty} \frac{1}{n} \log P\Big(\hat{\theta} \in \Gamma^c\Big) \leq -\inf_{\theta\in\Gamma^c} \mathcal{H}(\theta)$$

avec

$$\mathcal{H}(\theta) = -\log \inf_{u>0, m\in\mathcal{P}} E_{g_0\nu}\Big(e^{u\int_{\mathcal{T}_p}\log\frac{g(\theta,X)}{g(t,X)}m(dt)}\Big)$$

où $\mathcal{P}$ représente l'ensemble des mesures de probabilités m sur $\mathcal{T}_p$.

et le Théorème 4.5.3 de Dembo et Zeitouni (1993) adapté aux lois des grands nombres, que nous utiliserons dans le cas où $\mathcal{X}$ est l'ensemble des fonctions continues sur un compact et $h_n = \frac{1}{n}\partial^2 l_n(.)$:

Théorème 2.0.3 *Soit $\mathcal{X}$ un espace vectoriel topologique, soit h_n une suite de $\mathcal{X}$ telle que $\Lambda(\lambda) = \log E_{g_0\nu}(e^{<\lambda,h_n>})$ où λ appartient au dual $\mathcal{X}^*$ de $\mathcal{X}$ et $<,>$ indique le crochet de dualité. Alors, pour tout compact $\Gamma \in \mathcal{X}$,*

$$\limsup_{n\to\infty} \frac{1}{n} \log P(h_n \in \Gamma) \leq -\inf_{x\in\mathcal{X}} \Lambda^*(x)$$

où $\Lambda^*(.)$ est la transformée de Fenchel-Legendre de $\Lambda(.)$:

$$\Lambda^*(x) \stackrel{\Delta}{=} \sup_{\lambda\in\mathcal{X}^*}\{<\lambda, x> -\Lambda(\lambda)\}, \quad x \in \mathcal{X}.$$

Quand la familles de mesures associée à h_n est exponentiellement tendue, cette propriété s'étend alors à tous les ensembles fermés (Remarque c) suivant le théorème). La tension exponentielle est la propriété suivante :

Définition 2.0.1 *La suite h_n est exponentiellement tendue dans $\mathcal{X}$ si, pour tout réel $\varepsilon > 0$, il existe un compact $K \subset \mathcal{X}$ tel que*

$$\limsup_{n\to+\infty} \frac{1}{n} \log P(h_n \in K^c) < -\varepsilon.$$

Dans le cas de la sous-estimation, on suppose que $\hat{\theta}^{q-k}$ converge vers un unique $\breve{\theta}^{q-k}$ qui minimise l'information de Kullback entre $g(\theta_0^q, .)$ and $\mathcal{G}_{q-k}$:

$$K_{q-k} = \min_{\theta^{q-k}\in\mathcal{T}_{q-k}} E_{g_0\nu}\Big(\log \frac{g(\theta_0^q, X)}{g(\theta^{q-k}, X)}\Big).$$

On décompose alors $\{T_n(q-k) - l_n(\theta_0) \geq -kv(n)\}$ en deux évènements, pour un $A < K_{q-k}$:

- $\{l_n(\breve{\theta}^{q-k}) - l_n(\hat{\theta}_0^q) \geq n(-K_{q-k} + A) - kv(n)\}$, qui est contrôlable car on sait que $l_n(\breve{\theta}^{q-k}) - l_n(\hat{\theta}_0^q)$ tend en loi vers un χ^2.
- $\{l_n(\hat{\theta}^{q-k}) - l_n(\breve{\theta}^{q-k}) \geq n(K_{q-k} - A)\}$, qui est contrôlable car $l_n(\hat{\theta}^{q-k}) - l_n(\breve{\theta}^{q-k})$ peut être décomposé en termes de lois des grands nombres et de grandes déviations sur $\hat{\theta}^{q-k}$.

Ces deux parties se traitent avec des grandes déviations. Ceci montre que la probabilité de sous-estimer est négligeable par rapport à la probabilité de surestimer.

Les hypothèses de moments exponentiels considérées dans ces théorèmes sont vérifiées pour les familles de binomiales négatives, qui ne sont pas des familles exponentielles.

Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families

Christine Keribin
Université d'Evry-Val d'Essonne, France [1]

Dominique Haughton
Bentley College and M.I.T., USA [2]

Abstract: Large and moderate deviation principles are applied to find the limiting behavior of the probability of under-estimating or over-estimating the order of a model when using a maximum penalized likelihood estimator. The models are assumed to be identifiable even if the order is over-estimated. This paper extends known results concerning models in exponential families to more general regular models. Applications using negative binomial distributions are given.

2.1 Introduction

In this paper we investigate the over- and under-estimation of the order of a parametric model when the order is obtained by a maximum penalized likelihood method. The main tools used are moderate and large deviations principles.
Let $\mathcal{T}_p = \{\theta^p = (t_1 \dots t_p) \in T_1 \times \dots \times T_p\}$, where $T_i, i = 1, \dots, p$ are compact sets of $\mathbb{R}^{r_i}$. Note that the dimension of θ^p is $R_p = \sum_{i=1}^{p} r_i$ and $\theta^p = (\theta_1, \dots, \theta_{R_p})$. We consider the following sequence of parametric sets of probability densities with respect to a measure ν:

$$(\mathcal{G}_p = g(\theta^p, .), \theta^p \in \mathcal{T}_p)_{p \le Q} \ , \mathcal{G}_{p_1} \cap \mathcal{G}_{p_2} = \mathcal{G}_{p_1 \wedge p_2} \tag{3}$$

where $p_1 \wedge p_2 = \min(p_1, p_2)$. We define the true order q of $g \in \mathcal{G}_Q$ as the smallest integer such that all the $t_i, i = q+1, \dots, Q$ are null. Then for $1 \le k \le Q-q$, $\theta^{q+k} = (\theta^q, 0, \dots, 0)$.

[1](1) Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Boulevard des Coquibus, 91025 Evry

[2](2) Department of Mathematical Sciences, Bentley College, 175 Forest Street, Waltham, MA 02452-4705, USA

[3]*Key words and phrases.* penalized maximum likelihood, large deviations, moderate deviations, information criteria, model selection, negative binomial distributions.

We observe an i.i.d. n-sample $(X_1, \dots, X_n)$ of the distribution $g_0 \in \mathcal{G}_Q$ with true order q and true parameter $\theta^q = \theta_0^q := \theta_0$. We assume that

- **ID**: The model is identifiable for $\theta = \theta_0$, even if the order is well or over-estimated, that is:
$$\forall p \geq q \ , \ g(\theta^p, .) = g_0(.) \Leftrightarrow \theta^p = (\theta_0, 0, \dots, 0) = \theta_0^p.$$

Regression models are examples of such models, provided that the independent variables in the regression are assumed to be random variables as opposed to "design" variables (see Haughton (1988)).

For any order $p \in \{1, .., Q\}$ and any $g \in \mathcal{G}_p$, define the log-likelihood $l_n(\theta^p)$ to be:
$$l_n(\theta^p) = \sum_{i=1}^{n} \log g(\theta^p, X_i)$$
and the maximum likelihood statistic $T_n(p)$ over a model of dimension R_p to be
$$T_n(p) = \sup_{\theta^p \in \mathcal{T}_p} l_n(\theta^p).$$
Under appropriate assumptions, the maximum likelihood estimator $\hat{\theta}^p$ exists and we write: $T_n(p) = l_n(\hat{\theta}^p)$. The behavior of the maximum likelihood statistic depends on whether the order p is less or greater than the true order q:

- if $p > q$, $T_n(p) - T_n(q)$ converges to a chi-square law with $R_p - R_q$ degrees of freedom (Wilks (1967))
- if $p < q$, $T_n(p) - l_n(\theta_0)$ converges to $-\infty$, at a rate of order n (with similar arguments as Leroux (1992)):
$$\lim_{n \to +\infty} \frac{1}{n}(T_n(p) - l_n(\theta_0)) = -K < 0$$

Now, *the penalized likelihood* is defined by subtracting from the maximum likelihood some compensation term, yielding what are called information criteria:
$$W_n(p) = T_n(p) - a(p)v(n)$$
where $a(p)$ is a function that increases with the order (we assume in this paper that $a(p) = p$) and $v(n)$ is a function of the sample size n. The *maximum penalized likelihood estimator* $\hat{q}_n$ of the order q maximizes the penalized likelihood,
$$\hat{q}_n = arg \max_{p \in \{1, \dots, Q\}} (W_n(p)).$$

Information criteria are widely used and studied for model selection, and include (when $r_i = 1$, for $i = 1, \cdots p$) for example the AIC (Akaike (1984)) for which $v(n) = 1$, and the BIC (Schwarz (1978)) for which $v(n) = (1/2)\log n$. The use of penalized likelihood methods for determining the order of a model has been already studied (see Azencott and Dacunha-Castelle (1984)) and the consistency properties of these methods are known under appropriate assumptions (see section 2 for a summary of known results). Our goal here is to find the rate of convergence of this maximum penalized likelihood estimator to the true order; so we investigate the asymptotic probabilities of over- or under-estimating the true order by a value k. Such asymptotic probabilities have been investigated in the case of exponential families (see Woodroofe (1982), Haughton (1989)) and in the case of linear regression (see (Nishii (1984)).

This paper extends results of Haughton (1989) to the case of i.i.d. observations from regular families which are not necessarily exponential.

After introducing our notations and recalling known consistency results in section 2, we give our assumptions and statements of results in section 3. Section 4 gives applications of our results to model selection problems involving the negative binomial family of distributions. Detailed proofs of the results are given in section 5.

2.2 Assumptions and statements of known consistency results

We first define the notations of partial derivatives with respect to θ^p and of the different norms that will be used:

- Let $f(\theta^p)$ be a function of $\theta^p \in \mathcal{T}_p$. If $f(.)$ is sufficiently differentiable with respect to θ^p, its gradient at $\theta^p = \tilde{\theta}^p$ is denoted by $\partial f(\tilde{\theta}^p)$. For $i = 1, \ldots, R_p$, $\partial_i f(\tilde{\theta}^p)$ is the partial derivative of f at $\tilde{\theta}^p$ with respect to the $i-$th coordinate of θ.
- In the same way, the matrix $\partial^2 f(\tilde{\theta}^p) = [\partial^2_{i,j} f(\tilde{\theta}^p)]$ is the matrix of the second derivatives with respect to θ^p at $\theta^p = \tilde{\theta}^p$.
- $\mathbb{R}^{R_Q}$ will be given the usual euclidean norm $\| \, . \, \|$ and its associated product $<,>$.

- Let $\mathcal{M}_{R_Q}$ be the set of square $\mathbb{R}^{R_Q}$-matrices and let $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_Q})$ be the set of continuous functions from $\mathcal{Z}$ to $\mathcal{M}_{R_Q}$. The norm on $\mathcal{M}_{R_Q}$ is defined as

$$\| M \|_m = \sup_{\|x\|=1, x \in \mathbb{R}^{R_Q}} \| Mx \|$$

and the norm on $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_Q})$ is

$$\| M(.) \|_c = \sup_{\tilde{\theta} \in \mathcal{Z}} \| M(\tilde{\theta}) \|_m .$$

We now state the assumptions that lead to the consistency of the maximum penalized likelihood estimator:

- **H0-1:**
  - There exists a $g_0\nu$-integrable function f such that $\sup_{\theta \in \mathcal{T}_Q} | \log g(\theta, .) | \leq f(.)$.
  - For all x, $g(., x)$ is continuous on $\mathcal{T}_Q$.
- **H0-2**: The model is regular in θ_0^Q:
  - For all x, $\log g(., x)$ is twice continuously differentiable in a neighborhood $\mathcal{V}$ of θ_0^Q.
  - $\partial l_n(\theta_0^Q)$ has a zero mean and is square-integrable for $g_0\nu$.
  - For $1 \leq i, j \leq R_Q$:

$$E_{g_0\nu}(\partial l_n(\theta_0^Q) \partial l_n(\theta_0^Q)^T) = -E_{g_0\nu}(\partial^2 l_n(\theta_0^Q))$$

    is non-singular. Since the observations are iid,

$$E_{g_0\nu}(\partial l_n(\theta_0^Q) \partial l_n(\theta_0^Q)^T) = n E_{g_0\nu}(\frac{\partial g_0}{g_0} \frac{\partial g_0}{g_0}^T) = nJ(\theta_0^Q)$$

    where $J(\theta_0^Q)$ is the Fisher information matrix at $\theta = \theta_0^Q$.
- **H0-3:** There exists a θ_0^Q-neighborhood $\mathcal{W} \subset \mathcal{V}$ and a $g_0\nu$-integrable function h such that for all $i, j = 1, \ldots, R_Q$, $\sup_{\theta^Q \in \mathcal{W}} | \partial^2_{i,j} \log g(\theta^Q, .) | \leq h(.)$.
- **H0-4**: $q \leq Q$ and θ_0^Q belongs to the interior of $\mathcal{T}_Q$.
- **H0-5:** The compensation $v(n)$ satisfies $\lim_{n \to +\infty} v(n) = +\infty$ and $v(n) = o(n)$.

Under **Id** and **H0-1** the maximum likelihood estimator (MLE) of θ_0 in $\mathcal{T}_q$ converges to θ_0 (see (Dacunha-Castelle et Duflo (1993)). Assumptions **H0-2** and **H0-3** give the asymptotic normality of the MLE (see Dacunha-Castelle et Duflo (1993)). The convergence of the maximum likelihood estimator of θ_0^{q+k} in $\mathcal{T}_{q+k}$, $1 \leq k \leq Q-q$ holds if **H0-4** is assumed. Under these assumptions, if the compensation satisfies **H0-5**, the estimator of the order based on the maximum penalized likelihood method converges in probability to the true order (see Azencott and Dacunha-Castelle (1984)).

2.3 Assumptions and statements of theorems on over-estimation and under-estimation

We first investigate the asymptotic probabilities of over- and under-estimating the order by a fixed value k. The asymptotic probabilities of over- and under-estimating the order will be a direct consequence of these results.

2.3.1 Asymptotic behavior of the probability of over-estimating the order by a value k

Consider the following assumptions:

- **H1**: The following set is a neighborhood of 0:

$$D^1 = \{t \in \mathbb{R}^{R_Q}; E_{g_0\nu}(e^{<t, \partial \log g(\theta_0^Q, x)>}) < \infty\}.$$

- **H2-1**: There exists a neighborhood $\mathcal{Z}_1 \subset \mathcal{W}$ of θ_0^Q, where, for all x, $\log g(., x)$ is three times continuously differentiable in θ.

On this neighborhood, the Fisher information $J(\tilde{\theta}) = [J_{i,j}(\tilde{\theta})] = -E_{g_0\nu}(\frac{1}{n}\partial^2 l_n(\tilde{\theta}))$ belongs to $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_Q})$. Using now **H0-3**, there exists $\mathcal{Z} \subset \mathcal{Z}_1$ such that $0 < \| J(.) \|_c < \infty$.

- **H2-2**: For all $1 \leq i, j, l \leq R_Q$, the following sets are neighborhoods of 0:

$$\begin{aligned} D^2_{i,j} &= \{t \in \mathbb{R}; E_{g_0\nu}(e^{t \cdot \sup_{\tilde{\theta} \in \mathcal{Z}} |\partial^2_{i,j} \log g(\tilde{\theta}, X)|}) < \infty\} \\ D^3_{i,j,l} &= \{t \in \mathbb{R}; E_{g_0\nu}(e^{t \cdot \sup_{\tilde{\theta} \in \mathcal{Z}} |\partial^3_{i,j,l} \log g(\tilde{\theta}, X)|}) < \infty\}. \end{aligned}$$

- **H3**: For any $N < +\infty$,
$$\limsup_{r\to 0} E_{g_0\nu}(e^{N\omega(r)})$$
is finite, where $\omega(r)$ is the modulus of continuity of $\log g(\theta,x)$:
$$\omega(r) = \sup_{\theta_1,\theta_2\in\mathcal{T}_Q, \|\theta_1-\theta_2\|<r} \mid \log g(\theta_1,x) - \log g(\theta_2,x) \mid .$$

Theorem 2.3.1 *For $k \geq 1$, define the event B_k to be $B_k = \{W_n(q+k) - W_n(q) > 0\}$. Let assumptions* **ID**, **H0** *to* **H3** *hold. We have:*
$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(B_k)] = -k.$$

2.3.2 Asymptotic behavior of the probability of under-estimating the order by a value k

Here we need some additional assumptions in order to use large deviations properties. For $0 < k < q$, let
$$K_{q-k} = \min_{\theta^{q-k}\in\mathcal{T}_{q-k}} E_{g_0\nu}(\log \frac{g(\theta_0^q, X)}{g(\theta^{q-k}, X)})$$
be the minimum Kullback-Leibler distance between $g(\theta_0^q,.)$ and $\mathcal{G}_{q-k}$. Note that K_{q-k} decreases with k.

Consider the following assumptions:

- **H4**: For each $1 \leq k < q$, the minimum Kullback-Leibler distance K_{q-k} is attained at only one $\breve{\theta}^{q-k} \in \mathcal{T}_{q-k}$ and $E_{g_0\nu}(\partial^2 \log g(\breve{\theta}^{q-k}, X)) = J(\breve{\theta}^{q-k})$ is non-singular.

- **H5-1**: For each $1 \leq k < q$, there exists a neighborhood $\breve{\mathcal{Z}}^k \subset \mathcal{W}$ of $\breve{\theta}^{q-k}$ where, for all x, $\log g(.,x)$ is three times continuously differentiable in θ and $0 <\| J(\breve{\theta}^{q-k}) \|_c< \infty$.

- **H5-2**: For all $1 \leq i,j,l \leq R_Q$, and all $0 < k < q$, the following sets are neighborhoods of 0:
$$\begin{aligned} D^2_{i,j} &= \Big\{t \in \mathbb{R}; E_{g_0\nu}(e^{t\cdot\left(\sup_{\breve{\theta}\in\breve{\mathcal{Z}}^k} \sup|\partial^2_{i,j}\log g(\breve{\theta},X)|\right)}) < \infty\Big\} \\ D^3_{i,j,l} &= \Big\{t \in \mathbb{R}; E_{g_0\nu}(e^{t\cdot\left(\sup_{\breve{\theta}\in\breve{\mathcal{Z}}^k} \sup|\partial^3_{i,j,l}\log g(\breve{\theta},X)|\right)}) < \infty\Big\}. \end{aligned}$$
This assumption differs from assumption **H2** because the supremum is taken over $\breve{\mathcal{Z}}^k$ instead of $\mathcal{Z}$.

- **H6:** The following set is a neighborhood of 0:

$$\Big\{t; E_{g_0\nu}(e^{t\cdot\log\frac{g(\breve{\theta}^{q-k},X)}{g(\theta_0^q,X)}}) < \infty\Big\}.$$

Theorem 2.3.2 *For $k = 1, \cdots, q-1$, define the event B_{-k} to be*

$$B_{-k} = \{W_n(q-k) - W_n(q) > 0\}.$$

Let assumptions **ID**, **H0**, **H3** *to* **H6** *hold. We have:*

$$\limsup_{n\to+\infty} \frac{1}{n} \log[P(B_{-k})] < 0$$

It follows from Theorem 2.3.2 that asymptotically, these probabilities are negligible compared to those we studied in the previous section. Hence, their lower bound is not very important and is not computed here.

2.3.3 Rate of convergence of the maximum penalized likelihood estimator to the true order

We are now ready to state our main result, as a consequence of Theorems 2.3.1 and 2.3.2:

Theorem 2.3.3 *Under assumptions* **ID**, **H0** *to* **H6**, *we have:*

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n > q)] = -1$$

and

$$\limsup_{n\to+\infty} \frac{1}{n} \log[P(\hat{q}_n < q)] < 0.$$

Proof of Theorem 2.3.3
First write:

$$P(\hat{q}_n > q) \leq \sum_{k=1}^{Q-q} P(B_k)$$

where $P(B_1)$ is the leading term, and

$$P(\hat{q}_n < q) \leq \sum_{k=1}^{q-1} P(B_{-k})$$

where the probabilities $P(B_{-k})$ are negligible compared to $P(\hat{q}_n > q)$. So

$$\limsup_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n > q)] \leq -1,$$

and

$$\limsup_{n\to+\infty} \frac{1}{n} \log[P(\hat{q}_n < q)] < 0.$$

To complete the proof of the first statement of Theorem 2.3.3, we write:

$$P(B_1) - P(\hat{q}_n < q) \leq P(\hat{q}_n > q).$$

Since

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(B_1)] = -1$$

by Theorem 2.3.1, and since $P(\hat{q}_n < q)$ is asymptotically negligible compared to $P(B_1)$, it follows that

$$\liminf_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n > q)] \geq -1$$

which completes the proof.

A similar argument yields the following corollary:

Corollary: The probability $P(\hat{q}_n = q+1)$ can be asymptotically estimated as follows:

$$\lim_{n\to+\infty} \frac{1}{v(n)} \log[P(\hat{q}_n = q+1)] = -1.$$

2.4 Application to model selection problems involving the negative binomial family

2.4.1 Verification of assumptions

Consider the following family of laws on the integers, referred to as the negative binomial family:

$$g_{\alpha,p} = \frac{\Gamma(x+\alpha)}{\Gamma(x+1)\Gamma(\alpha)}(1-p)^{\alpha} p^{x}, \; x = 0, 1, \cdots. \tag{4}$$

The parameters of this family are $\alpha > 0$ and $0 < p < 1$. We will assume in our applications that $\alpha \in [\alpha_1, \alpha_2]$ and $p \in [p_1, p_2]$, where α_1, α_2, p_1 and p_2 are suitably chosen for a given application in such a way that $1 \in (\alpha_1, \alpha_2)$. It is known that the negative binomial family is not an exponential family when α and p are both unknown (see Brown, 1986, p. 27). Note that when $\alpha = 1$ we get a family of geometric distributions.

We also note that (as is well known) the negative binomial family arises as a compound mixture of Poisson distributions, with the Poisson rate λ distributed as a Gamma. For that reason the negative binomial family has been used to model counts where it is felt that the rate of occurence is not homogeneous (see for example Nelson (1980) and references herein).

We prove below that this family satisfies assumptions **H0** to **H6**. Here $\theta = (\alpha, p)$ and $g(\theta, x) = g_{\alpha,p}(x)$. We have $g(\theta, x) = \log\Gamma(x+\alpha) - \log\Gamma(x+1) - \log\Gamma(\alpha) + \alpha\log(1-p) + x\log p$. Using known facts (see e.g. Abramowitz and Stegun) about the Gamma function and the Digamma function (the latter defined as the derivative of the logarithm of the Gamma function), we have:

$$\sup_{\alpha} |\log\Gamma(x+\alpha) - \log\Gamma(x+1)| \leq \max\{|\log .886 - \log\Gamma(x+1)|, |\log\Gamma(x+\alpha_2) - \log\Gamma(x+1)|,$$

$$|\log\Gamma(x+\alpha_1) - \log\Gamma(x+1)|)\}.$$

By classical asymptotic expansions (given in Abramowitz and Stegun) for $\Gamma(x)$ as $x \to \infty$, we have

$$\Gamma(x+\alpha_i)/\Gamma(x+1) \sim x^{\alpha_i - 1},$$

for $i = 1, 2$. Since $\log x$ is absolutely summable with respect to $g_{\alpha,p}$ as is easily checked, the only remaining question is whether $\log\Gamma(x+1)$ is summable with respect to $g_{\alpha,p}$, which follows from the asymptotic expansion for $\Gamma(x)$ as $x \to \infty$ and the fact that $x\log x$ is summable.

In addition, $\log g_{\alpha,p}(x)$ is clearly continuous on $[\alpha_1, \alpha_2] \times [p_1, p_2]$. So **H0-1** is satisfied.

To check **H0-2**, note first that $\log g(\theta, x)$ is C^∞ on $(0, \infty) \times (0, 1)$ by C^∞ properties of the Gamma function on that range. Also, the derivatives $\partial_p l_n$ and $\partial_\alpha l_n$ both have zero means because the order of summation and differentiation can be exchanged.

The question of whether $\partial l_n(\theta_0)$ is square-integrable reduces to the question of whether $\psi(X_i + \alpha)$ is square-integrable, where ψ denotes the Digamma function, which in turns follows from the fact that $\psi(x) \sim \log x$ as $x \to \infty$ (see Abramowitz and Stegun p. 259).

As for the non-singularity of the Fisher information matrix, its determinant equals

$$[E_{g_0\nu}(\partial g_0/\partial\alpha)^2][E_{g_0\nu}(\partial g_0/\partial p)^2] - [E(\partial g_0/\partial\alpha)(\partial g_0/\partial p)]^2.$$

This determinant is strictly positive since $[E(XY)]^2 \leq (EX^2)(EY^2)$ with equality if and only if $Y = \lambda X$ almost surely for some constant λ.

To check **H0-3**, we have, by monotonicity of ψ' and ψ'':

$$\sup_{\alpha,p} |\partial^2 \log g/\partial\alpha^2| \leq |\psi'(x+\alpha_1)| + |\psi'(\alpha_1)|,$$

$$\sup_{\alpha,p} |\partial^2 \log g/\partial\alpha\partial p| = 1/(1-p_1),$$

and

$$\sup_{\alpha,p} |\partial^2 \log g/\partial p^2| \leq \alpha_2/(1-p_2)^2 + x/p_1^2.$$

These bounds are all clearly summable with respect to $g_{\alpha,p}$, using for the first bound the fact that $\psi'(x) \sim 1/x$ as $x \to \infty$.

Assumption **H0-4** will be satisfied in the applications we will consider, and we will use information criteria for which the compensation $v(n)$ satisfies **H0-5**.

Checking **H1** reduces to checking whether $\exp(t_1\psi(x+\alpha))$ and $\exp(xt_2/p)$ are summable with respect to $g_{\alpha,p}$ for t_1, t_2 small enough, which is straightforward. So **H1** holds.

The function $\log g(\cdot, x)$ is C^∞ on $(0,\infty)\times(0,1)$ so **H2-1** is satisfied.

As for **H2-2**, we have, as is easily checked:

$$\sup |\partial^2 \log g/\partial\alpha^2| = \psi'(\alpha_1) - \psi'(x+\alpha_1),$$

$$\sup |\partial^2 \log g/\partial\alpha\partial p| = 1/(1-p_2),$$

and

$$\sup |\partial^2 \log g/\partial p^2| \leq \alpha_2/(1-p_2)^2 + x/p_1^2.$$

As for suprema of absolute values of third derivatives, one is zero, the other one constant in x and the remaining ones equal:

$$\sup_{\alpha,p} |\partial^3 \log g/\partial\alpha^3| = \psi''(x+\alpha_1) - \psi''(\alpha_1),$$

$$\sup_{\alpha,p} |\partial^3 \log g/\partial p^3| \leq 2\alpha_2/(1-p_2)^3 + 2x/p_1^3.$$

Assumption **H2-2** now follows from the fact that $\exp(-t\psi'(x+\alpha_1))$, $\exp(tx/p_1^2)$, $\exp(t\psi''(x+\alpha_1))$, and $\exp(2tx/p_1^3)$ are summable with respect to $g_{\alpha,p}$, for t small enough. This in turns follows from the fact that $\psi'(x+\alpha_1) \sim 1/(x+\alpha_1)$ and $\psi''(x+\alpha_1) \sim -1/(x+\alpha_1)^2$, as $x \to \infty$.

To check **H3**, we note that, as can be checked easily, we have

$$\omega(r) \leq \sqrt{(x/p_1)^2 + [\psi(x+\alpha_1) + C]^2 + D},$$

where C and D do not depend on x ($D > 0$). It is clear that each function of x

$$x \to \exp[Nr\sqrt{(x/p_1)^2 + [\psi(x+\alpha_1) + C]^2 + D}]$$

is integrable with respect to the discrete probability measure given on the integers by $g_{\alpha,p}$. As r decreases to zero, these functions decrease to one. So by the monotone convergence theorem, the desired limit is finite, and **H3** is satisfied.

As for assumption **H4**, in the applications that follow, we will consider the question of whether $\alpha = 1$ or $\alpha \neq 1$. Let α_0, p_0 be the true parameters, with $\alpha_0 \neq 1$. Minimizing the Kullback-Leibler distance in K_{q-k} is equivalent to maximizing (over p) the function $\phi(p) = \log(1-p) + (\log p)\ \mu_{\alpha_0,p_0}$ where μ_{α_0,p_0} is the mean of the distribution g_{α_0,p_0} on the integers, so $\mu_{\alpha_0,p_0} = \alpha_0 p_0/(1-p_0) > 0$. It is clear by simple analyses that $\phi(p)$ has a unique maximum over $[p_1, p_2]$ at $\hat{p} = \mu_0/(\mu_0+1)$, provided that $\mu_0/(\mu_0+1) \in [p_1, p_2]$. In a given application, we will know of conceivable upper and lower bounds for α_0 and p_0 (thus for μ_0) and we will work with a compact set $[\alpha_1, \alpha_2] \times [p_1, p_2]$ such that $\mu_0/(\mu_0+1)$ will be inside (p_1, p_2) for any conceivable value of μ_0.

With this precaution taken, $\phi(p)$ has a unique maximum. The Fisher information matrix was proved to be non-singular at any (α, p) in $(0, \infty) \times (0, 1)$ so **H4** is satisfied.

Assumption **H5-1** is clearly satisfied because in fact **H0-3** holds over the whole compact set $[\alpha_1, \alpha_2] \times [p_1, p_2]$.

Assumption **H5-2** is satisfied because **H2** was proved with suprema over the whole compact set $[\alpha_1, \alpha_2] \times [p_1, p_2]$, and since we have assumed that the compact $[\alpha_1, \alpha_2] \times [p_1, p_2]$ contains $(1, \mu_0/(\mu_0+1))$ for any conceivable value of μ_0 for each of our applications.

The verification of **H6** is straightforward.

2.4.2 Application to counts of wars and burglaries

Given a set of observations from a family as defined in (4), we consider the question of whether $\alpha = 1$, in which case the negative binomial family is in fact a geometric family, or $\alpha \neq 1$. If in fact $\alpha = 1$, the observations are coming from a distribution which satisfies the *memoryless* or *lack of aging* property, which means that

$$P(X > x|X > y) = P(X > x - y).$$

Consider the following two situations:

Number of outbreaks of war per year during 1500 - 1931

Ortiz (1990) fits a negative binomial to the number of outbreaks of war per year for a period of 432 years (1500 - 1931). During that period, war broke out in the world a total of 299 times. Ortiz found maximum likelihood estimators (MLE's) $\hat{\alpha}$ and $\hat{p}$ for α and p equal to $\hat{\alpha} = 7.3772$ and $\hat{p} = .085773$. These estimators yield a maximum likelihood of -481.05 and a value of the BIC criterion equal to -487.12. Here the BIC criterion, to be maximized to select a model, is defined as:

$$BIC = \text{sup log likelihood} - (1/2)\text{dim(model)}\log n$$

(Schwarz, 1978).

For the model where $\alpha = 1$, the MLE for p is easily calculated as $\hat{p_G} = 1/(1 + (n/\sum X_i))$, where $n = 432$ is the sample size, and $\sum X_i = 299$ is the total number of times war broke out during the period considered. So $\hat{p_G} = .409$, yielding a maximum log likelihood

$$\text{max log lik}_G = n\log(1 - \hat{p_G}) + [\textstyle\sum X_i]\log \hat{p_G}$$

equal to -494.52 and a BIC equal to -497.56. So the BIC selects the model where α is unrestricted, and thus rejects the memoryless property. One possible interpretation of this result is that "war breeds war".

The frequency table for the number of outbreaks of wars per year during 1500 - 1931 was reconstructed from the original reference used by Ortiz (Wright, 1976) and is given below:

0	223
1	142
2	48
3	15
4	4
Total	432

Number of burglaries per year for a sample of households in five large cities in the United States

Nelson (1980) fits negative binomials to data on the number of burglaries during 1974 - 1975 for a sample of households from the National Crime Survey for five large cities in the United States (Chicago, Detroit, Los Angeles, New York, Philadelphia), and for samples of black households in Chicago from different age groups (16 - 24, 25 - 34, 35 - 49, 50 and above). Approximations to the MLE's $\hat{\alpha}$ and $\hat{p}$ for α and p were calculated by Nelson (1980), and are given below. The frequency tables for the number of burglaries experienced by households in the five cities are given in Table 2 of Nelson (1980). Using these data and the MLEs calculated by Nelson, we obtain the maximum log likelihood and BIC values given below for the model with α unrestricted, and for the model where $\alpha = 1$ (for which we use subscripts G, for geometric).

	Chicago	Detroit	Los Angeles	New York	Philadelphia
$\hat{\alpha}$	.498	.507	.599	.364	.376
$\hat{p}$	.195	.233	.199	.175	.195
sup log lik	-4090.42	-4328.92	-4644.07	-2737.76	-2757.29
BIC	-4099.69	-4338.09	-4653.33	-2746.96	-2766.51
sup log lik$_G$	-4104.44	-4347.77	-4653.84	-2752.99	-2762.96
BIC$_G$	-4109.08	-4352.36	-4658.47	-2757.59	-2767.57
$\hat{p_G}$	.108	.134	.130	.072	.071
n	10,675	9,586	10,505	9,906	10,005

In this context, the memoryless property implies that the probability of being further victimized (given that one will be victimized at all) is in fact equal to the probability of being victimized at all. The question of whether this property holds for victimization data was posed by Nelson at the end of his paper (see Nelson, 1980, p. 890).

For all five cities the model where $\alpha = 1$ (and thus the memoryless property) is rejected by the BIC; however, Philadelphia comes close to an $\alpha = 1$ selection.

For the number of burglaries for the four age groups of black households, MLEs for α and p were calculated by Nelson and are given below; the frequency tables for the number of burglaries are not available so we could not calculate BIC values in this case.

Note that of course BIC would select the geometric model for the age group 50 and above, since $\hat{\alpha} = 1.0$ for that group.

	16 - 24	25 - 34	35 - 49	50 and above
$\hat{\alpha}$	2.8	1.5	.76	1.0
$\hat{p}$	.08	.11	.16	.11

2.5 Proofs of Theorems

2.5.1 Proof of Theorem 2.3.1

The study of the asymptotic behavior of the likelihood ratio

$$P(B_k) = P(T_n(q+k) - T_n(q) > kv(n)),$$

leads to the main following steps:

- Performing a Taylor expansion of the likelihood ratio $T_n(q+k) - T_n(q)$
- Identifying its leading term to be $\frac{1}{2}(L_n(\theta_0))^T \cdot \Sigma^{-1} \cdot L_n(\theta_0)$, with $L_n(\theta_0)$ a vector of $\mathbb{R}^{R_{q+k}-R_q}$ that converges to a normal law $\mathcal{N}(0, \Sigma)$
- Applying a moderate deviations theorem to the leading term
- Proving that the remaining terms are negligible compared to the leading term, by applying large deviations theorems on Law of Large Numbers terms in Banach spaces, and on the maximum likelihood estimator.

We first study $T_n(q+k) - l_n(\theta_0) = l_n(\hat{\theta}^{q+k}) - l_n(\theta_0)$. For $k = 0, \cdots, Q-q$, applying a Taylor expansion to the gradient of the likelihood leads to the existence of a $\tilde{\theta}^{q+k}$ with each coordinate such that $|\ \tilde{\theta}_i^{q+k} - \theta_{i,0}^{q+k}\ | \leq |\ \hat{\theta}_i^{q+k} - \theta_{i,0}^{q+k}\ |$, $1 \leq i \leq R_{q+k}$, satisfying the following equality:

$$0 = \partial l_n(\hat{\theta}^{q+k}) = \partial l_n(\theta_0^{q+k}) + \partial^2 l_n(\tilde{\theta}^{q+k}) \cdot (\hat{\theta}^{q+k} - \theta_0^{q+k})$$

which yields, when $\partial^2 l_n(\tilde{\theta}^{q+k})$ is non-singular:

$$(\hat{\theta}^{q+k} - \theta_0^{q+k}) = -[\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})]^{-1} \cdot \frac{1}{n}\partial l_n(\theta_0^{q+k}) \tag{5}$$

Now, the expansion of the likelihood leads to the following equation, with $|\ \ddot{\theta}_i^{q+k} - \theta_{i,0}^{q+k}\ | \leq |\ \hat{\theta}_i^{q+k} - \theta_{i,0}^{q+k}\ |$, $1 \leq i \leq R_{q+k}$ and I_{q+k} the identity matrix of $\mathcal{M}_{R_{q+k}}$, set of square matrices of dimension R_{q+k}:

$$\begin{aligned}
l_n(\hat{\theta}^{q+k}) - l_n(\theta_0) &= (\partial l_n(\theta_0^{q+k}))^T \cdot (\hat{\theta}^{q+k} - \theta_0^{q+k}) + \frac{1}{2}(\hat{\theta}^{q+k} - \theta_0^{q+k})^T \cdot \partial^2 l_n(\ddot{\theta}^{q+k}) \cdot (\hat{\theta}^{q+k} - \theta_0^{q+k}) \\
&= -\frac{1}{2}\frac{1}{n}\partial l_n(\theta_0^{q+k})^T \cdot [\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})]^{-1} \cdot \\
&\quad \Big(I_{q+k} + (\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k}) - \frac{1}{n}\partial^2 l_n(\ddot{\theta}^{q+k})) \cdot [\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})]^{-1}\Big) \cdot \partial l_n(\theta_0^{q+k}) \\
&= \frac{1}{2}\frac{1}{n}\partial l_n(\theta_0^{q+k})^T \cdot [J(\theta_0^{q+k})]^{-1} \cdot (I_{q+k} + M_n^{q+k}) \cdot \partial l_n(\theta_0^{q+k}),
\end{aligned}$$

where $J(\theta_0^{q+k})$ is the Hessian and where M_n^{q+k} will be proved to be the negligible part of the expression:

$$\begin{aligned}
M_n^{q+k} &= (I_{q+k} + Q_1).(I_{q+k} + Q_2).(I_{q+k} + Q_3) - I_{q+k}, \text{ where} \\
Q_1 &= (-J(\theta_0^{q+k}) - \frac{1}{n}\partial^2 l_n(\theta_0^{q+k})) \cdot [\frac{1}{n}\partial^2 l_n(\theta_0^{q+k})]^{-1}, \\
Q_2 &= (\frac{1}{n}\partial^2 l_n(\theta_0^{q+k}) - \frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})) \cdot [\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})]^{-1}, \\
Q_3 &= (\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k}) - \frac{1}{n}\partial^2 l_n(\ddot{\theta}^{q+k})) \cdot [\frac{1}{n}\partial^2 l_n(\tilde{\theta}^{q+k})]^{-1}.
\end{aligned}$$

Using the same decomposition for $l_n(\theta_0) - T_n(q) = l_n(\theta_0) - l_n(\hat{\theta}^q)$, and adding both decompositions, we can write the leading term for $T_n(q+k) - T_n(q)$ as:

$$\mathcal{L}_0 = L_n(\theta_0)^T \cdot [E_{g_0\nu}(L_n(\theta_0) \cdot L_n(\theta_0)^T)]^{-1} \cdot L_n(\theta_0),$$

where $L_n(\theta_0)$ is the following vector of dimension $R_{q+k} - R_q$:

$$L_n(\theta_0) = \begin{pmatrix} \frac{1}{n}\frac{\partial l_n(\theta_0^{q+k})}{\partial \theta_{R_q+1}} \\ \cdots \\ \frac{1}{n}\frac{\partial l_n(\theta_0^{q+k})}{\partial \theta_{R_{q+k}}} \end{pmatrix} - J_R \cdot J(\theta_0^q)^T \cdot \begin{pmatrix} \frac{1}{n}\frac{\partial l_n(\theta_0^{q+k})}{\partial \theta_1} \\ \cdots \\ \frac{1}{n}\frac{\partial l_n(\theta_0^{q+k})}{\partial \theta_{R_q}} \end{pmatrix},$$

and J_R is the part of $J(\theta_0^{q+k})$ between the first and R_q-th columns, and the R_q+1-th and R_{q+k}-th rows:

$$J_R = \begin{pmatrix} E_{g_0\nu}(\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_q+1}}\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_1}) & \cdots & E_{g_0\nu}(\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_q+1}}\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_q}}) \\ \vdots & \cdots & \vdots \\ E_{g_0\nu}(\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_{q+k}}}\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_1}) & \cdots & E_{g_0\nu}(\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_{q+k}}}\frac{\partial \log(g(\theta_0^{q+k}))}{\partial \theta_{R_q}}) \end{pmatrix}.$$

Using these results, $\frac{2}{n}(T_n(q+k) - T_n(q))$ is expressed as the sum of the leading term $\mathcal{L}_0$ and of the negligible terms $\mathcal{L}_1$ and $\mathcal{L}_2$:

$$\frac{2}{n}(T_n(q+k) - T_n(q)) = \mathcal{L}_0 + \mathcal{L}_1 + \mathcal{L}_2,$$

where

$$\begin{aligned}\mathcal{L}_1 &= \frac{1}{n}\partial l_n(\theta_0^{q+k})^T \cdot [J(\theta_0^{q+k})]^{-1} \cdot M_n^{q+k} \cdot \frac{1}{n}\partial l_n(\theta_0^{q+k}),\\ \mathcal{L}_2 &= \frac{1}{n}\partial l_n(\theta_0^{q})^T \cdot [J(\theta_0^{q})]^{-1} \cdot M_n^{q} \cdot \frac{1}{n}\partial l_n(\theta_0^{q}).\end{aligned}$$

For every ε_1 such that $0 < \varepsilon_1 < 14^{-\frac{1}{3}}$, we can now bound $P(B_k) = P\Big(T_n(q+k) - T_n(q)) \geq kv(n)\Big)$ as follows:

$$P(B_k) \leq P\Big(\mathcal{L}_0 > \frac{2kv(n)(1-14\varepsilon_1^3)}{n}\Big) + P\Big(\parallel \mathcal{L}_1 \parallel_c > 7\varepsilon_1^3 \frac{2kv(n)}{n}\Big) + P\Big(\parallel \mathcal{L}_2 \parallel_c > 7\varepsilon_1^3 \frac{2kv(n)}{n}\Big),$$

where the $\parallel . \parallel_c$ norm involves the supremum over the compact set where $\tilde{\theta}^{q+k}$ and $\ddot{\theta}^{q+k}$ are defined.

The leading term $P\Big(\mathcal{L}_0 > \frac{2kv(n)(1-14\varepsilon_1^3)}{n}\Big)$ is treated below in Lemma 2.5.1. We study now the neglegtible parts involving $\mathcal{L}_1$ and $\mathcal{L}_2$:

$$\parallel \mathcal{L}_1 \parallel_c \leq \parallel M_n^{q+k} \parallel_c \cdot \parallel [J(\theta_0^{q+k})]^{-1} \parallel_m \cdot \parallel \frac{1}{n}\partial l_n(\theta_0^{q+k}) \parallel^2$$

Thus,

$$P\Big(\parallel \mathcal{L}_1 \parallel_c > 7\varepsilon_1^3 \frac{2kv(n)}{n}\Big) \leq P\Big(\parallel \frac{1}{n}\partial l_n(\theta_0^{q+k}) \parallel^2 \geq \frac{2kv(n)}{n} \cdot \frac{2}{\lambda_0^{q+k}}\Big) + P\Big(\parallel M_n^{q+k} \parallel_c \geq \frac{7\varepsilon_1^3}{2} . \lambda_0^{q+k}\Big) \tag{6}$$

with $\lambda_0^{q+k} = \inf_{x \in \mathcal{A}_1} < x, [J(\theta_0^{q+k})]^{-1} x >$, and $\mathcal{A}_1 = \{x \in \mathrm{I\!R}^{R_{q+k}}; \parallel x \parallel \geq 1\}$. Note that $\lambda_0^{q+k} > 0$ by **H0-2**. As M_n^{q+k} is the sum of seven matrices, each of them being the product of at most three of the matrices Q_1, Q_2, and Q_3, it is sufficient to determine the asympotic behavior of $\parallel Q_1 \parallel_c$, $\parallel Q_2 \parallel_c$ and $\parallel Q_3 \parallel_c$ to have the behavior of $\parallel M_n^{q+k} \parallel_c$. The first term in inequality (6) is treated in Lemma 2.5.2 below, and the second term is treated in Lemmas 2.5.3-2.5.5 below, leading to the upper bound for $P(B_k)$ needed for Theorem 2.3.1.

As for a lower bound for $P(B_k)$, one just has to note that, for $\varepsilon > 0$,

$$P\Big(\mathcal{L}_0 > \frac{2kv(n)(1+\varepsilon)}{n}\Big) - P\Big(\parallel \mathcal{L}_1 \parallel_c \geq \frac{\varepsilon kv(n)}{n}\Big) - P\Big(\parallel \mathcal{L}_2 \parallel_c \geq \frac{\varepsilon kv(n)}{n}\Big) \leq P(B_k),$$

and to use the same Lemmas 2.5.1 to 2.5.5, thus achieving the proof of Theorem 2.3.1.

We now give Lemmas 2.5.1-2.5.5, where, to simplify the notations, the $q+k$ superscripts have been omitted.

Lemma 2.5.1 *Under assumptions* **(H0-5)** *and* **(H1)**, *for any ε such that $-1 < \varepsilon < 1$,*

$$\lim_{n \to +\infty} \frac{1}{kv(n)} \log P\Big(\mathcal{L}_0 > \frac{2kv(n)(1-\varepsilon)}{n}\Big) = -(1-\varepsilon).$$

Lemma 2.5.2 *Under assumptions* **(H0-5)** *and* **(H1)***, and for* n *large enough,*

$$P\Big(\parallel \frac{1}{n}\partial l_n(\theta_0) \parallel^2 \geq \frac{2kv(n)}{n} \cdot \frac{2}{\lambda_0}\Big)$$

is $o(e^{-kv(n)})$.

Lemma 2.5.3 *Under assumption* **(H2-1)***, for any* $\varepsilon > 0$*, and for* n *large enough,*

$$P(\parallel Q_1 \parallel_c \geq \varepsilon) = P\Big(\parallel (-J(\theta_0) - \frac{1}{n}\partial^2 l_n(\theta_0)) \cdot [\frac{1}{n}\partial^2 l_n(\theta_0)]^{-1} \parallel_c \geq \varepsilon\Big)$$

is $o(e^{-kv(n)})$.

In the expression of M_n, in particular in Q_2 and Q_3, appear $\tilde{\theta}$ and $\ddot{\theta}$, which depend on $\hat{\theta}$. It is then also necessary to study the probabilities of deviations of $\tilde{\theta}$ and $\ddot{\theta}$. This is done in Lemma 2.5.4 below.

Lemma 2.5.4 *Under assumptions* **(H0-1)** *to* **(H2-3)**, **(H2-1)**, **(H3)***, and for* n *large enough,*

$$P\Big(\hat{\theta} \in \mathcal{T} \backslash \mathcal{Z}\Big)$$

is $o(e^{-kv(n)})$*, where* $\mathcal{Z}$ *is the neighborhood defined right after the statement of assumption* **H2-1**.

Lemma 2.5.5 *Under assumptions* **(H0-1)** *to* **(H2-3)** *and* **(H2)***, for any* $\varepsilon > 0$*, and for* n *large enough,*

$$P(\parallel Q_3 \parallel_c \geq \varepsilon) = P\Big(\parallel (\frac{1}{n}\partial^2 l_n(\tilde{\theta}) - \frac{1}{n}\partial^2 l_n(\ddot{\theta})) \cdot [\frac{1}{n}\partial^2 l_n(\tilde{\theta})]^{-1} \cdot 1_{[\hat{\theta} \in \mathcal{Z}]} \parallel_c \geq \varepsilon\Big)$$

is $o(e^{-kv(n)})$.

Note that a Lemma similar to Lemma 2.5.5 and proved in the same way shows that

$$P(\parallel Q_3 \parallel_c \geq \varepsilon) = o(e^{-kv(n)}).$$

Proof of Lemma 2.5.1 The matrix $[E_{g_0\nu}(L_n(\theta_0) \cdot L_n(\theta_0)^T)]$ is symmetric, so there exists a matrix Q and a diagonal matrix J_d such that $QQ^T = I$ and $[E_{g_0\nu}(L_n(\theta_0) \cdot L_n(\theta_0)^T)] = Q^T J_d Q$. Define now $J_d^{\frac{1}{2}}$ as the diagonal matrix satisfying $J_d^{\frac{1}{2}} J_d^{\frac{1}{2}} = (J_d^{\frac{1}{2}})^T J_d^{\frac{1}{2}} = J_d$.
We have, with $\mathcal{A}_1 = \{x \in \mathbb{R}^{R_{q+k}}; \parallel x \parallel \geq 1\}$ and $w(n) = (\frac{2kv(n)(1-\varepsilon)}{n})^{-\frac{1}{2}}$:

$$\begin{aligned} P\Big(\mathcal{L}_0 \geq \frac{2kv(n)(1-\varepsilon)}{n}\Big) &= P\Big(\frac{w(n)}{n} \parallel J_d^{\frac{1}{2}} Q L_n(\theta_0) \parallel \geq 1\Big) \\ &= P\Big(\frac{w(n)}{n} J_d^{\frac{1}{2}} Q L_n(\theta_0) \in \mathcal{A}_1\Big) \end{aligned}$$

We now check that the assumptions are sufficient to apply moderate deviations Theorem 3.7.1 of Dembo and Zeitouni (1993):

- $E_{g_0\nu}(J_d^{\frac{1}{2}} Q\partial L_n(\theta_0)(x)) = 0$, and using **H1**,

$$\{t \in \mathbb{R}^{R_{q+k}}; E_{g_0\nu}(e^{<t, J_d^{\frac{1}{2}} Q\partial L_n(\theta_0)(x)>}) < \infty\}$$

  is a neighborhood of 0.

- C, the covariance of $J_d^{\frac{1}{2}} Q\partial L_n(\theta_0)(x)$ is the identity matrix of $\mathcal{M}_{R_{q+k}}$ and is thus invertible.

$$\begin{aligned} C &= E_{g_0\nu}((J_d^{\frac{1}{2}} Q\partial L_n(\theta_0)(x))(J_d^{\frac{1}{2}} Q\partial L_n(\theta_0)(x))^T) \\ &= J_d^{\frac{1}{2}} Q E_{g_0\nu}(L_n(\theta_0)L_n(\theta_0)^T)Q^T(J_d^{\frac{1}{2}})^T \\ &= J_d^{\frac{1}{2}} QQ^T J_d^{-1} QQ^T (J_d^{\frac{1}{2}})^T \\ &= I \end{aligned}$$

- Theorem 3.7.1 of Dembo and Zeitouni (1993) is stated for $w(n) = n^\beta, \beta \in]0, \frac{1}{2}[$. However, a careful reading of the proof leads to the fact that the result holds for $w(n)$ such that $w(n) = o(\sqrt{n})$ and $\lim_{n\to+\infty} w(n) = +\infty$, which holds in our case by **H0-5**.

We now apply this theorem and write, with $\bar{\mathcal{A}}_1$ and $\mathcal{A}_1^o$ the closure and interior of $\mathcal{A}_1$:

$$\begin{aligned} -\frac{1}{2}\inf_{x\in\mathcal{A}_1^o} < x, C^{-1}x > &\leq \liminf_{n\to+\infty} \frac{w(n)^2}{n}\log P_1 \\ &\leq \limsup_{n\to+\infty} \frac{w(n)^2}{n}\log P_1 \\ &\leq -\frac{1}{2}\inf_{x\in\bar{\mathcal{A}}_1} < x, C^{-1}x > \end{aligned} \tag{7}$$

so that

$$\lim_{n\to+\infty} \frac{w(n)^2}{n}\log P_1 = -\frac{1}{2},$$

thus concluding the proof of Lemma 2.5.1

Proof of Lemma 2.5.2 We can use here the same moderate deviations theorem as in Lemma 2.5.1, with now $w(n) = (\frac{2kv(n)}{n}\cdot\frac{2}{\lambda_0})^{-\frac{1}{2}}$ and $C = J(\theta_0)$. This leads to:

$$\begin{aligned} \lim_{n\to+\infty}\frac{1}{kv(n)}\log P\Big(\parallel \frac{1}{n}\partial l_n(\theta_0)\parallel \geq \sqrt{\frac{2kv(n)}{n}\cdot\frac{2}{\lambda_0}}\Big) &= -\frac{2}{\lambda_0}\inf_{x\in\mathcal{A}_1} < x, [J(\theta_0)]^{-1}x > \\ &= -2. \end{aligned}$$

So that, for all $0 < \eta < 1$ and for n large enough,

$$P(\parallel \frac{1}{n}\partial l_n(\theta_0)\parallel \geq \sqrt{\frac{2kv(n)}{n}\cdot\frac{2}{\lambda_0}}) \leq e^{-2kv(n)(1-\eta)}$$

which is a $o(e^{-kv(n)})$.

Proof of Lemma 2.5.3 Q_1 is the product of two terms containing $\frac{1}{n}\partial^2 l_n(\theta_0)$, so that we have here a case of large deviations on $\frac{1}{n}\partial^2 l_n(\theta_0)$. Write, for any $\varepsilon_2 > 0$,

$$P\Big(\| (-J(\theta_0) - \frac{1}{n}\partial^2 l_n(\theta_0)) \cdot [\frac{1}{n}\partial^2 l_n(\theta_0)]^{-1} \|_c \geq \varepsilon \Big) \leq P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{A}_2\Big) + P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{A}_3\Big),$$

with

$$\mathcal{A}_2 = \Big\{ M \in \mathcal{M}_{R_{q+k}}, \| M + J(\theta_0) \|_m \geq \varepsilon \frac{1}{\| [J(\theta_0)]^{-1} \|_m + \varepsilon_2} \Big\}$$

and

$$\mathcal{A}_3 = \Big\{ M \in \mathcal{M}_{R_{q+k}}, \| M^{-1} \|_m \geq \| [J(\theta_0)]^{-1} \|_m + \varepsilon_2 \Big\}.$$

To use Theorem 4.5.3 of Dembo and Zeitouni (1993)on large deviations, we first need to prove that the sequence of matrices $\{\frac{1}{n}\partial^2 l_n(\theta_0)\}$ is exponentially tight.

Indeed, for any $K > 0$, let $\mathcal{K}_1(K)$ be the following compact set of $\mathcal{M}_{R_{q+k}}$:

$$\mathcal{K}_1(K) = \Big\{ M \in \mathcal{M}_{R_{q+k}} \ , \ \| M \|_m \leq \| J(\theta_0) \|_m + K \Big\}.$$

We write:

$$\begin{aligned} P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{K}_1(K)^C\Big) &= P\Big(\| \frac{1}{n}\partial^2 l_n(\theta_0) \|_m > \| J(\theta_0) \|_m + K \Big) \\ &\leq P\Big(\frac{1}{n} \sum_{i=1}^{n} \| \partial^2 \log g(\theta_0, X_i) \|_m \geq \| J(\theta_0) \|_m + K \Big) \\ &\leq 2e^{-n\mathcal{H}_1(\|J(\theta_0)\|_m + K)} \end{aligned}$$

by applying Chernoff's Theorem on $\mathbb{R}$ (Chernoff's Theorem applies by **H2-2**), where $\mathcal{H}_1$ is the Cramer function associated with $\| \partial^2 \log g(\theta_0, X) \|_m$, see chapter 2.2 of Dembo and Zeitouni (1993)for the properties of such function. Now, for all $\eta > 0$, there exists K such that $\mathcal{H}_1(\| J(\theta_0) \|_m + K) > \eta$. Then

$$P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{K}_1(K)^C\Big) \leq 2e^{-n\eta}$$

Hence

$$\limsup_{n\to+\infty} \frac{1}{n} \log P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{K}_1(K)^C\Big) \leq -\eta,$$

and $\{\frac{1}{n}\partial^2 l_n(\theta_0)\}$ is exponentially tight.

Now, Theorem 4.5.3 of Dembo and Zeitouni (1993) and the exponential tightness of $\{\frac{1}{n}\partial^2 l_n(\theta_0)\}$ lead to the following upper bounds:

$$\limsup_{n\to+\infty} \frac{1}{n} \log P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{A}_2\Big) \leq - \inf_{M \in \mathcal{A}_2} \Lambda^*(M) = -h_2$$

$$\limsup_{n\to+\infty} \frac{1}{n} \log P\Big(\frac{1}{n}\partial^2 l_n(\theta_0) \in \mathcal{A}_3\Big) \leq - \inf_{M \in \mathcal{A}_3} \Lambda^*(M) = -h_3$$

where Λ^* is the convex rate function defined as the Fenchel-Legendre transform of the generating moment function Λ associated with $\{\frac{1}{n}\partial^2 l_n(\theta_0)\}$:

$$\forall M \in \mathcal{M}_{R_{q+k}} \ , \ \Lambda^*(M) = \sup_{t\in\mathcal{M}^*_{R_{q+k}}} \{\sum_{i,j=1}^{R_{q+k}} t_{i,j}M_{i,j} - \Lambda(t)\}$$

$$\forall t \in \mathcal{M}^*_{R_{q+k}} \ , \ \Lambda(t) = \log E_{g_0\nu}(e^{\sum_{i,j=1}^{R_{q+k}} t_{i,j}\partial^2_{i,j}\log g(\theta_0,X)}).$$

It remains to see that $h_2 > 0$ and $h_3 > 0$. Note first that $\frac{1}{n}\partial^2 l_n(\theta_0)$ converges to $-J(\theta_0)$ in probability: so that, for $\nu > 0$,

$$\limsup_{n\to+\infty} \frac{1}{n}\log P(\| \frac{1}{n}\partial^2 l_n(\theta_0) + J(\theta_0) \|_m \leq \nu) = 0.$$

Write now

$$\limsup_{n\to+\infty} \frac{1}{n}\log P(\| \frac{1}{n}\partial^2 l_n(\theta_0) + J(\theta_0) \|_m \leq \nu) \leq -\inf_{\|M+J(\theta_0)\|_m\leq\nu} \Lambda^*(M).$$

Let ν tend to 0 to obtain $\Lambda^*(-J(\theta_0)) = 0$. As $\mathcal{A}_2$ and $\mathcal{A}_3$ are closed subsets that do not intersect a certain neighborhood of $-J(\theta_0)$, the convexity of the rate function Λ^* allows us to conclude that $h_2 > 0$ and $h_3 > 0$. To complete the proof of the Lemma, we note that $e^{-h_2 n}$ is a $o(e^{-kv(n)})$ by Assumption **H0-5**.

Proof of Lemma 2.5.4 Assumption **H3** allows us to use a large deviations theorem for the maximum likelihood estimator $\hat{\theta}$ Mogulskii (1988):

$$\begin{aligned}
\limsup_{n\to+\infty} \frac{1}{n}P\Big(\hat{\theta}\in\mathcal{T}\backslash\mathcal{Z}\Big) &= \limsup_{n\to+\infty}\frac{1}{n}P\Big(\sup_{\tilde{\theta}\in\mathcal{Z}}(l_n(\tilde{\theta}) - l_n(\theta_0)) \leq \sup_{\tilde{\theta}\in\mathcal{T}\backslash\mathcal{Z}}(l_n(\tilde{\theta}) - l_n(\theta_0))\Big) \\
&\leq -\inf_{\theta\in\mathcal{T}\backslash\mathcal{Z}} \mathcal{H}(\theta) \\
\text{with } \mathcal{H}(\theta) &= -\log \inf_{u>0,m\in\mathcal{P}} E_{g_0\nu}(e^{u\int_{\mathcal{T}}\log\frac{g(\theta,X)}{g(t,X)}m(dt)})
\end{aligned}$$

where $\mathcal{P}$ denotes the class of probability measures m on $\mathcal{T}$.
We now prove that $\inf_{\theta\in\mathcal{T}\backslash\mathcal{Z}}\mathcal{H}(\theta) > 0$. As $m(dt)$ is a probability measure, Jensen's inequality leads to

$$e^{u\int_{\mathcal{T}}\log\frac{g(\theta,X)}{g(t,X)}m(dt)} \leq \int_{\mathcal{T}} e^{u\log\frac{g(\theta,X)}{g(t,X)}}m(dt) = \int_{\mathcal{T}}\Big(\frac{g(\theta,X)}{g(t,X)}\Big)^u m(dt)$$

Choosing a particular measure $m(dt) = \delta_{\theta_0}$ and a real number $u = \frac{1}{2}$ yields:

$$\begin{aligned}
\inf_{u>0,m\in\mathcal{P}} E_{g_0\nu}(e^{u\int_{\mathcal{T}}\log\frac{g(\theta,X)}{g(t,X)}m(dt)}) &\leq E_{g_0\nu}\Big(\Big(\frac{g(\theta,X)}{g(\theta_0,X)}\Big)^{\frac{1}{2}}\Big) \\
&= 1 - \frac{1}{2}E_{g_0\nu}\Big(\Big(\sqrt{g(\theta,X)} - \sqrt{g(\theta_0,X)}\Big)^2\Big) \leq 1
\end{aligned}$$

Under **Id**, the expression on the right-hand-side is equal to 1 if and only if $\theta = \theta_0$. Hence, for $\theta \in \mathcal{T}\backslash\mathcal{Z}$

$$\inf_{u>0, m\in\mathcal{P}} E_{g_0\nu}(e^{u\int_{\mathcal{T}} \log\frac{g(\theta,X)}{g(t,X)}m(dt)}) < 1$$

so that, $\inf_{\theta\in\mathcal{T}\backslash\mathcal{Z}} \mathcal{H}(\theta) > 0$ and $P\Big(\hat\theta \in \mathcal{T}\backslash\mathcal{Z}\Big)$ is $o(e^{-kv(n)})$ for n large enough.

Proof of Lemma 2.5.5 To simplify notations, we do not repeat $1_{[\hat\theta\in\mathcal{Z}]}$ in the arguments below. We deal first with $\frac{1}{n}\partial^2 l_n(\tilde\theta) - \frac{1}{n}\partial^2 l_n(\ddot\theta)$. Since $\tilde\theta$ and $\ddot\theta$ lie in the same neighborhood, a Taylor expansion for each component of $\frac{1}{n}\partial^2 l_n(\tilde\theta) - \frac{1}{n}\partial^2 l_n(\ddot\theta)$ leads to the existence of $\alpha^{(ij)} \in \mathcal{Z}$ such that

$$\frac{1}{n}\partial^2_{i,j} l_n(\tilde\theta) - \frac{1}{n}\partial^2_{i,j} l_n(\ddot\theta) = \sum_{l=1}^{R^{q+k}} (\tilde\theta_l - \ddot\theta_l)\frac{1}{n}\partial^3_{i,j,l} l_n(\alpha^{(ij)}).$$

We define $H^{(l)}$ as the matrix whose i-th row and j-th column is $\frac{1}{n}\partial^3_{i,j,l} l_n(\alpha^{(ij)})$ and write:

$$\frac{1}{n}\partial^2 l_n(\tilde\theta) - \frac{1}{n}\partial^2 l_n(\ddot\theta) = \sum_{l=1}^{R_{q+k}} (\tilde\theta_l - \ddot\theta_l) H^{(l)}.$$

Now,

$$\begin{aligned}
\| \frac{1}{n}\partial^2 l_n(\tilde\theta) - \frac{1}{n}\partial^2 l_n(\ddot\theta) \|_m &\leq \sum_{l=1}^{R_{q+k}} \mid \tilde\theta_l - \ddot\theta_l \mid \cdot \| H^{(l)} \|_m \\
&\leq \| \tilde\theta - \ddot\theta \| \sum_{i,j,l=1}^{R_{q+k}} \mid \frac{1}{n}\partial^3_{i,j,l} l_n(\alpha^{(ij)}) \mid \\
&\leq 2 \| \theta_0 - \tilde\theta \| \sum_{i,j,l=1}^{R_{q+k}} \mid \frac{1}{n}\partial^3_{i,j,l} l_n(\alpha^{(ij)}) \mid \\
&\leq 2 \| [\frac{1}{n}\partial^2 l_n(\tilde\theta)]^{-1} \|_m \| \frac{1}{n}\partial l_n(\theta_0) \| \sum_{i,j,l=1}^{R_{q+k}} \mid \frac{1}{n}\partial^3_{i,j,l} l_n(\alpha^{(ij)}) \mid,
\end{aligned}$$

using (5). This leads to, for any $\varepsilon_3 > 0$ and $\varepsilon_4 > 0$:

$$\begin{aligned}
P(\| Q_3 \|_c \geq \varepsilon) &\leq P\Big(\sum_{i,j,l=1}^{R_{q+k}} \sup_{\tilde\theta\in\mathcal{Z}} \mid \frac{1}{n}\partial^3_{i,j,l} l_n(\tilde\theta) \mid \cdot \| [\frac{1}{n}\partial^2 l_n(.)]^{-1} \|_c^2 \cdot \| \frac{1}{n}\partial l_n(\theta_0) \| \geq \frac{\lambda_0}{2}\varepsilon\Big) \\
&\leq P_1 + \sum_{i,j,l=1}^{R_{q+k}} P_2^{i,j,l} + P_3
\end{aligned}$$

with

$$P_1 = P\left(\| \frac{1}{n}\partial l_n(\theta_0) \| \geq \frac{\lambda_0\varepsilon}{2} \frac{(\| [E_{g_0\nu}(\partial^2 \log g(.,X))]^{-1} \|_c + \varepsilon_3)^{-2}}{\sum_{i,j,l=1}^{R_{q+k}} E_{g_0\nu}(\sup_{\tilde\theta\in\mathcal{Z}} \mid \partial^3_{i,j,l} \log g(\tilde\theta, X) \mid) + \varepsilon_4}\right),$$

$$P_2^{i,j,l} = P\Big(\sup_{\tilde{\theta}\in\mathcal{Z}} \mid \frac{1}{n}\partial^3_{i,j,l} l_n(\tilde{\theta}) \mid \geq E_{g_0\nu}(\sup_{\tilde{\theta}\in\mathcal{Z}} \mid \partial^3_{i,j,l} \log g(\tilde{\theta}, X) \mid) + \varepsilon_4\Big),$$
$$P_3 = P\Big(\parallel [\frac{1}{n}\partial^2 l_n(.)]^{-1} \parallel_c \geq \parallel [E_{g_0\nu}(\partial^2 \log g(., X))]^{-1} \parallel_c + \varepsilon_3\Big).$$

By assumption **H1**, P_1 is $o(e^{-kv(n)})$ by using a Chernoff inequality. For all $1 < i, j, l < R_{q+k}$, we write:

$$P_2^{i,j,l} \leq P\Big(\frac{1}{n}\sum_{k=1}^{n} \sup_{\tilde{\theta}\in\mathcal{Z}} \mid \partial^3_{i,j,l} \log g(\tilde{\theta}, X_k) \mid \geq E_{g_0\nu}(\sup_{\tilde{\theta}\in\mathcal{Z}} \mid \partial^3_{i,j,l} \log g(\tilde{\theta}, X) \mid) + \varepsilon_4\Big),$$

and by assumption **H2-2**, we use a Chernoff inequality to conclude that P_2 is $o(e^{-kv(n)})$. P_3 can be written as

$$P_3 = P\Big(\frac{1}{n}\partial^2 l_n(.) \in \mathcal{A}_4\Big),$$

where $\mathcal{A}_4$ the following closed subset of $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$:

$$\mathcal{A}_4 = \{M(.) \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}}), \parallel M(.)^{-1} \parallel_c \geq \parallel [E_{g_0\nu}(\partial^2 \log g(., X))]^{-1} \parallel_c + \varepsilon_3\}.$$

An upper bound for P_3 can be found by applying again Theorem 4.5.3 of Dembo and Zeitouni (1993) on the Banach space $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$ of continuous functions of one variable from $\mathcal{Z}$ to $\mathcal{M}_{R_{q+k}}$, since $\frac{1}{n}\partial^2 l_n(.)$ belongs to this space. The theorem applies as follows:

- Recall first the norm on $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$:

$$\forall M(.) \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}}) \ , \ \parallel M(.) \parallel_c = \sup_{\tilde{\theta}\in\mathcal{Z}} \parallel M(.) \parallel_m$$

- The dual space of $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$ (space of all continuous linear functionals on $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$) is denoted throughout by $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})^*$. Let $t \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})^*$. For all $M(.) \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$,

$$t(M(.)) = \int_{\tilde{\theta}\in\mathcal{Z}} \sum_{i,j=1}^{R_{q+k}} [M(\tilde{\theta})]_{i,j} dt_{i,j}(\tilde{\theta})$$

- The logarithmic moment generating function $\Lambda : \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})^* \to [-\infty, \infty]$ exists as an extended real number.

$$\forall t \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})^* \ , \Lambda(t) = \log E_{g_0\nu}(e^{t(\frac{1}{n}\partial^2 l_n(.))})$$

- The Fenchel-Legendre transform of $\Lambda(.)$ denoted by $\Lambda^*(.)$ is defined as

$$\forall M(.) \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}}) \ , \Lambda^*(M(.)) = \sup_{t\in\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})^*} \{t(M(.)) - \Lambda(t)\}$$

- The family of probability measures associated with $\{\frac{1}{n}\partial^2 l_n(.)\}$ is exponentially tight, as proved in Lemma 2.5.6 below.

- Under these assumptions, Theorem 4.5.3 of Dembo and Zeitouni (1993) with its c) remark, leads to the convexity and non-negativity of Λ^* and gives an upper bound for

$$\limsup_{n\to+\infty} \frac{1}{n}\log P\Big(\frac{1}{n}\partial^2 l_n(.) \in \mathcal{A}_4\Big) \leq -\inf_{M(.)\in\mathcal{A}_4} \Lambda^*(M(.)) = -h_4.$$

It remains to see that $h_4 > 0$. As $\frac{\partial^2}{\partial\theta^2}\log g(\theta, X)$ is Lipschitz in θ by **H2-1**, the sequence $\{\frac{1}{n}\partial^2 l_n(.)\}$ is a Glivenko-Cantelli class in probability (see Van der Vaart and Wellner (1996)). Hence it converges to $-J(.)$ uniformly in probability so that $\Lambda^*(-J(.)) = 0$. Since $\mathcal{A}_4$ is a closed subset that does not contain a neighborhood of $-J(.)$, the convexity of the rate function allows us to conclude that $h_4 > 0$. Hence, for n large enough, $P_3 \leq e^{-n\cdot h_4}$.

This leads us to conclude that, for n large enough, $P(Q_3 \geq \varepsilon)$ is $o(e^{-kv(n)})$.

Lemma 2.5.6 *The family of probability measures associated with $\frac{1}{n}\partial^2 l_n(.)$ is exponentially tight: for all η, there exists a compact K_η of $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$ such that, for n large enough,*

$$P\Big(\frac{1}{n}\partial^2 l_n(.) \in K_\eta^c\Big) \leq e^{-n\eta} \tag{8}$$

Proof of Lemma 2.5.6: For the exponential tightness, it suffices to have a pre-compact set K_η for which (8) holds. Let $\mathcal{K}(K_2, K_3)$ be the set of functions $M(.) \in \mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$, such that

$$\| M(.) \|_c \leq E_{g_0\nu}(\| \partial^2 \log g(., X) \|_c) + K_2$$

and

$$\| \partial_l M(.) \|_c \leq E_{g_0\nu}(\| H^l(., X) \|_c) + K_3, 1 \leq l \leq R_{q+k}.$$

This set is a pre-compact subset of $\mathcal{C}(\mathcal{Z}, \mathcal{M}_{R_{q+k}})$, because the functions are bounded with bounded derivatives. We write:

$$P\Big(\frac{1}{n}\partial^2 l_n(.) \in \mathcal{K}(K_2, K_3)^C\Big) \leq P_{K2} + \sum_{l=1}^{R_{q+k}} P_{K3}^l,$$

where

$$P_{K2} = P\Big(\| \frac{1}{n}\partial^2 l_n(.) \|_c \geq E_{g_0\nu}(\| \partial^2 \log g(., X) \|_c) + K_2\Big)$$

and

$$P_{K3}^l = \sum_{l=1}^{R_{q+k}} P\Big(\| \partial_l(\frac{1}{n}\partial^2 l_n(.)) \|_c \geq E_{g_0\nu}(\| H^l(., X) \|_c) + K_3\Big).$$

By assumption **H2-2**, Chernoff's Theorem on $\mathbb{R}$ applies with $\mathcal{H}_2$, the Cramer function associated with $\| \partial^2 \log g(., X) \|_c$:

$$\begin{aligned} P_{K2} &\leq P\Big(\frac{1}{n}\sum_{i=1}^{n} \| \partial^2 \log g(., X_i) \|_c \geq E_{g_0\nu}(\| \partial^2 \log g(., X) \|_c) + K_2\Big) \\ &\leq 2e^{-n\mathcal{H}_2(E_{g_0\nu}(\|\partial^2 \log g(.,X)\|_c)+K_2)}. \end{aligned}$$

Similarly, for $l = 1, \dots, R_{q+k}$, and $\mathcal{H}_3^l$ the Cramer function associated with $\| \partial_l(\partial^2 \log g(., X)) \|_c$:

$$\begin{aligned} P_{K3}^l &\leq P\Big(\frac{1}{n}\sum_{i=1}^n \| \partial_l(\partial^2 \log g(., X_i) \|_c) \geq E_{g_0\nu}(\| H^l(., X) \|_c) + K_3\Big) \\ &\leq 2e^{-n\mathcal{H}_3^l(E_{g_0\nu}(\|\partial_l(\partial^2 \log g(.,X))\|_c)+K_3)} \end{aligned}$$

Now, for all $\eta > 0$, there exists K_2, K_3 such that $\mathcal{H}_2(E_{g_0\nu}(\| \partial^2 \log g(., X) \|_c) + K_2) > \eta$ and $\mathcal{H}_3^l(E_{g_0\nu}(\| \partial_l(\partial^2 \log g(., X)) \|_c) + K_3) > \eta$ for any $l = 1, \cdots, R_{q+k}$. Then,

$$P\Big(\frac{1}{n}\partial^2 l_n(.) \in \mathcal{K}(K_2, K_3)^C\Big) \leq 4e^{-n\eta}.$$

Hence,

$$\limsup_{n\to+\infty} \frac{1}{n} \log P\Big(\frac{1}{n}\partial^2 l_n(.) \in \mathcal{K}(K_2, K_3)^C\Big) \leq -\eta$$

and the sequence $\{\frac{1}{n}\partial^2 l_n(.)\}$ is exponentially tight.

2.5.2 Proof of Theorem 2.3.2

Note first that

$$P(\hat{q}_n = q - k) \leq P(l_n(\hat{\theta}^{q-k}) - l_n(\theta_0^q) \geq -kv(n))$$

Now, for $A < K_{q-k}$, we have

$$P\Big(l_n(\hat{\theta}^{q-k}) - l_n(\theta_0^q) \geq -kv(n)\Big) \leq P_a + P_b$$

where

$$\begin{aligned} P_a &= P\Big(l_n(\hat{\theta}^{q-k}) - l_n(\breve{\theta}^{q-k}) \geq n(K_{q-k} - A)\Big), \\ P_b &= P\Big(l_n(\breve{\theta}^{q-k}) - l_n(\theta_0^q) \geq n(-K_{q-k} + A) - kv(n)\Big). \end{aligned}$$

Applying a classical Cramer large deviation result (Dembo and Zeitouni (1993), chapter2.2) to P_b yields:

$$\limsup_{n\to+\infty} \frac{1}{n} \log P\Big(\frac{l_n(\breve{\theta}^{q-k}) - l_n(\theta_0^q)}{n(-K_{q-k} + A - \frac{kv(n)}{n})} \leq 1\Big) \leq -\inf_{x\in\mathcal{A}_5} \Lambda^*(x)$$

where

$$\begin{aligned} \mathcal{A}_5 &= \Big\{x \in \mathbb{R}, x \leq 1\Big\}, \\ \Lambda^*(x) &= \sup_{\lambda\in\mathbb{R}}(\lambda x - \Lambda(\lambda)), \\ \Lambda(\lambda) &= \limsup_{n\to+\infty} \frac{1}{n} \log E_{g_0\nu}\Big(e^{\lambda\frac{l_n(\breve{\theta}^{q-k}) - l_n(\theta_0^q)}{n(-K_{q-k}+A-\frac{kv(n)}{n})}}\Big) \\ &= \log E_{g_0\nu}\Big(e^{\lambda\frac{\log g(\breve{\theta}^{q-k}, X) - \log g(\theta_0^q, X)}{-K_{q-k}+A}}\Big). \end{aligned}$$

As the expected value $E_{g_0\nu}\Big(\frac{\log g(\breve{\theta}^{q-k},X)-\log g(\theta_0^q,X)}{-K_{q-k}+A}\Big)$, vanishing Λ^*, does not belong to $\mathcal{A}_5$, we conclude that $\inf_{x\in\mathcal{A}_5}\Lambda^*(x)>0$, by using the convexity of the rate function.

We now study P_a. Applying a Taylor expansion in $\mathcal{T}_{q-k}$ to the gradient of the likelihood leads to the existence of a $\tilde{\theta}$ such that $\mid \tilde{\theta}_i-\breve{\theta}_i^{q-k}\mid\leq\mid\hat{\theta}_i-\breve{\theta}_i^{q-k}\mid$, $1\leq i\leq R_{q-k}$ which satisfies:

$$0=\partial l_n(\hat{\theta})=\partial l_n(\breve{\theta}^{q-k})+\partial^2 l_n(\tilde{\theta})\cdot(\hat{\theta}-\breve{\theta}^{q-k}).$$

This yields, when $\partial^2 l_n(\tilde{\theta})$ is non-singular:

$$(\hat{\theta}-\breve{\theta}^{q-k})=-[\frac{1}{n}\partial^2 l_n(\tilde{\theta})]^{-1}\cdot\frac{1}{n}\partial l_n(\breve{\theta}^{q-k}). \tag{9}$$

Now, expanding the likelihood leads to the following equation, with
$\mid\ddot{\theta}_i-\breve{\theta}_i^{q-k}\mid\leq\mid\hat{\theta}_i-\breve{\theta}_i^{q-k}\mid$, $1\leq i\leq R_{q-k}$ and I the identity matrix of $\mathcal{M}_{R_{q-k}}$:

$$\begin{aligned} l_n(\hat{\theta}^{q-k})-l_n(\breve{\theta}^{q-k}) &= -\frac{1}{2}\frac{1}{n}\partial l_n(\breve{\theta}^{q-k})^T\cdot[\frac{1}{n}\partial^2 l_n(\tilde{\theta})]^{-1}\cdot \\ &\Big(I_{q+k}+(\frac{1}{n}\partial^2 l_n(\tilde{\theta})-\frac{1}{n}\partial^2 l_n(\ddot{\theta}))\cdot[\frac{1}{n}\partial^2 l_n(\tilde{\theta})]^{-1}\Big)\cdot\partial l_n(\breve{\theta}^{q-k}). \end{aligned}$$

Similarly as in the proof of Theorem 2.3.1, replacing θ_0 by $\breve{\theta}^{q-k}$, P_b can be upper bounded by the sum of three terms:

$$P\Big(\frac{1}{n}\partial l_n(\breve{\theta}^{q-k})^T\cdot[J(\breve{\theta})]^{-1}\frac{1}{n}\partial l_n(\breve{\theta}^{q-k})\geq 2(K_{q-k}-A)(1-7\varepsilon_1^3)\Big),$$

$$P(\parallel\frac{1}{n}\partial l_n(\breve{\theta}^{q-k})\parallel\geq 2(K_{q-k}-A)\cdot\frac{2}{\breve{\lambda}}),$$

$$P(\parallel M\parallel_m\geq 7\varepsilon_1^3\cdot\frac{\breve{\lambda}}{2}),$$

where $\breve{\lambda}=\inf_{x\in\bar{\mathcal{A}}_6}<x,[J(\breve{\theta})]^{-1}x>$, $\mathcal{A}_6=(x\in\mathbb{R}^{R_{q-k}};\parallel x\parallel\geq 1)$. Each term can then be treated by using the lemmas proved for the over-estimation. This completes the proof of Theorem 2.3.2.

Bibliographie

[1] Abramowitz M, Stegun I. Handbook of mathematical functions, ninth printing. *Dover, New-York*, 1970.

[2] Akaike H. Information theory and an extension of the maximum likelihood principle, Second International Symposium on Information Theory (editors B.N. Petrov and F. Czàki): 267–281. *Akademiai Kiadò, Budapest*, 1984.

[3] Azencott R, Dacunha-Castelle D. Séries d'observations irrégulières. *Masson, Paris*, 1984.

[4] Brown, L. Fundamentals of statistical exponential families. Lecture Notes-monograph series, *Institute of Mathematical Statistics, Hayward, California*, 1986.

[5] Dacunha-Castelle D, Duflo M. Probabilités et statistiques: 2- Problèmes à temps mobile. *Masson, Paris*, 1993.

[6] Dacunha-Castelle D, Gassiat E. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.*, 27, 4, 1178-1209, 1999.

[7] Dacunha-Castelle D, Gassiat E. The estimation of the order of a mixture model. *Bernoulli*, 3: 279–299, 1997.

[8] Dembo A, Zeitouni O. Large Deviations Techniques and Applications. *Jones and Bartlett Publishers, Boston*, 1993.

[9] Haughton, D. On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16: 342–355, 1988.

[10] Haughton, D. Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā A*, 51: 45–58,1989.

[11] Keribin, C. Consistent Estimation of the Order of Mixture Models. *Sankhyā A*, 62-1: 49–66,2000.

[12] Ledoux, M. Sur les déviations modérées des sommes de variables vectorielles indépendantes de même loi. *Ann. Inst. Henri Poincaré*, 28(2): 267–280, 1992.

[13] Leroux, B G. Consistent estimation of a mixing distribution. *The Annals of Statistics*, 20(3): 1350–1360, 1992.

[14] Mogulskii, A A. Large deviations for maximum likelihood estimators. *Lecture Notes in Mathematics*, 1299: 326–331, 1988.

[15] Nelson, J. Multiple victimizations in American cities: a statistical analysis of rare events. *American Journal of Sociology*, 85(4): 870–891, 1980.

[16] Nishii, R. Asymptotic properties of criteria for selection of variables in multiple regression. *Annals of Statistics*, 12: 758–765, 1984.

[17] Ortiz, E. A stochastic model of the distribution of wars in time: the negative binomial distribution. *ASA Proceedings of the section on social statistics*, 285–290, 1990.

[18] Schwarz, G. Estimating the dimension of a model. *Annals of Statistics* 6: 461–464, 1978.

[19] Van der Vaart, A W, Wellner, J A. Weak convergence and empirical processes. *Springer Series in Statistics*, Springer Verlag, 1996.

[20] Wilks, S S. Mathematical statistics. *Wiley*, New York, 1967.

[21] Woodroofe, M. On model selection and the arc sines laws. *Annals of Statistics*, 10(4): 1182–1194, 1982.

[22] Wright, Q. A study of war. *University of Chicago Press, Chicago*, 1970.

Chapitre 3

Estimation consistante de l'ordre d'un modèle de mélange

Le travail présenté dans ce chapitre a été soumis à la revue *Sankhyā A*. Afin de pouvoir utiliser le même test que celui que nous venons de voir au cas du test de l'ordre (nombre de composantes) d'un mélange, il nous faut d'abord la consistance de l'estimateur du maximum de vraisemblance compensée pour les mélanges. Leroux (1992) a montré que, sous certaines conditions, cette méthode conduisait à un estimateur qui, asymptotiquement ne sous-estimait pas le nombre de composants presque sûrement. Pour cela, il utilise le fait que le rapport de la vraisemblance du vrai modèle à la vraisemblance d'un modèle ayant moins de composantes est de l'ordre de n fois la distance de Kullback entre ces deux modèles. Notons que son résultat est encore valable dans le cas des mélanges généralisés (le nombre de composantes n'est plus fini), dans le sens où l'estimateur du maximum de vraisemblance compensée converge vers l'infini.

Pour traiter la surestimation, il manquait d'en savoir plus sur le comportement du rapport de vraisemblance dans ce cas. Les travaux récents de Dacunha-Castelle et Gassiat (1997) ont apporté la réponse : le logarithme du rapport de vraisemblance $T_n(p) - l_n(g_0)$ tend vers la moitié du supremum du carré d'un processus Gaussien restreint à ses valeurs positives. Ce résultat nous permet de montrer que, l'estimateur du maximum de vraisemblance compensée ne surestime pas asymptotiquement, en probabilité, l'ordre d'un mélange.

De plus, en renforçant les hypothèses sur la compensation et la famille de loi, la convergence de l'estimateur de l'ordre est presque sûre, ce qui nous permet d'avoir une convergence presque sûre des paramètres dans la topologie quotient (prenant en compte les différentes permutations de paramètres donnant la même loi). Pour ce faire, nous avons d'une part proposé une preuve différente de celle de Leroux de la non sous-estimation de l'ordre, plus simple, mais ne prenant pas en compte le cas des mélanges généralisés. Puis nous avons étendu les résultats de Dacunha-Castelle et Gassiat (1999) en proposant, sous certaines conditions, une "loi du logarithme itéré" pour le rapport de vraisemblance. Après avoir vérifié les conditions d'entropie, suffisantes pour répondre à nos hypothèses, dans le cas des mélanges de Gaussiennes ou de loi de Poisson, nous avons étudié numériquement l'influence du terme de compensation dans le cas d'un mélange de deux lois Gaussiennes. Enfin, nous avons comparé les résultats avec ceux donnés par une méthode de moments (Dacunha-Castelle et Gassiat (1997b)).

Consistent Estimation of the Order of Mixture Models

Christine Keribin
Université d'Evry-Val d'Essonne, France [1]

Abstract: In this paper, we consider the estimation of the number of components for mixture models using a maximum penalized likelihood method. Leroux (1992) proved, under some assumptions, that the method leads to an estimator which, asymptotically, does not underestimate the number of components a.s.. We prove here the almost sure consistency of the maximum penalized likelihood estimator for an appropriate penalization sequence. The proof uses the locally conic parameterization developed by Dacunha-Castelle and Gassiat (1997). A numerical study of the choice of the penalization term is proposed, as well as a comparison with a moment method.

3.1 Introduction

We study in this paper the estimation of the size of a mixture using a maximum penalized likelihood method. 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. $\mathcal{G}_Q$ is the set of all $Q-$mixtures of densities of $\mathcal{F}$:

$$\begin{aligned}\mathcal{G}_Q \quad &= \quad \{g_{(\pi,\alpha)} = \sum_{i=1}^{Q} \pi_i \cdot f_{\gamma^i} \ : \ \pi = (\pi_1, \ldots, \pi_Q), \ \alpha = (\gamma^1, \ldots, \gamma^Q), \qquad (1)\\ &\qquad \forall i = 1, \ldots, Q, \ \gamma^i \in \Gamma, \ 0 \leq \pi_i \leq 1, \ \sum_{i=1}^{Q} \pi_i = 1\}\end{aligned}$$

We observe an n-sample $(X_1, \ldots, X_n)$ of the distribution g_0 for some $g_0 \in \mathcal{G}_Q$. We define the true order q of g_0 as the smallest integer such that all the $\gamma^i, i = 1, \ldots, q$ are different and all the associated π_i are not null. We want to estimate q and the parameters (π_0,α_0) of g_0 that are unknown.

[1](1) Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Boulevard des Coquibus, 91025 Evry

[2]*Key words and phrases.* model selection, maximum penalized likelihood, non identifiable models, mixtures.

The estimation of the order for a mixture has been studied in different ways. See for instance Henna (1985), Izenman and Sommer (1988), Roedner (1994), for various non-parametric techniques or Lindsay (1983) and Dacunha-Castelle and Gassiat (1997b) who used moment techniques.

We focus here on the maximum penalized likelihood method. For any order $p \in \{1, .., Q\}$ and any $g \in \mathcal{G}_p$, define:

$$l_n(g) = \sum_{i=1}^{n} \log g(X_i)$$

and the maximum likelihood statistic in a model of dimension p

$$T_{n,p} = \sup_{g \in \mathcal{G}_p} l_n(g)$$

The penalized likelihood is defined by subtracting to the maximum likelihood some compensation term, yielding what is called information criteria:

$$W_n(p) = T_{n,p} - a_{n,p}$$

The estimator $\hat{q}_n$ of the order q is the maximum of the penalized likelihood,

$$\hat{q}_n = arg \sup_{p \in \{1,\ldots,Q\}} (W_n(p))$$

This idea of using such a penalization term was first introduced by Akaike (1970), who defined the AIC criterion, in which the compensation is the dimension p of the embedded models. Later, Schwarz (1978), following Bayesian considerations, gave a BIC criterion for exponential families with a penalty term $(p/2)\log(n)$. Haughton (1988, 1989) proved the consistency of the BIC criterion for iid observations of smoothly curved submodels of exponential families, and studied its equivalence with a Bayesian approach. Senoussi (1988) worked on the a.s. properties of convex statistical models.

The study of the asymptotic properties of the estimators (and so the choice of the penalization term) splits into two subproblems which are of different nature:

- Avoiding underestimation: when the order is underestimated, in general the normalized maximum likelihood statistic $\frac{T_{n,p}}{n}$ converges to the minimum of the Kullback distance between the model of order p and the true distribution. The choice of $a_{n,p} = o(n)$ leads to an estimator which, asymptotically, does not underestimate the order.

- Avoiding overestimation: in case the model is identifiable even for overestimated order, and in case the parameters may be consistently estimated, it is often possible to have (using a Taylor expansion of the likelihood):

$$T_{n,p} - l_n(g_0) = \frac{1}{n}(\sum_{i=1}^{n} H_p(X_i))^2(1 + o(1))$$

  where the $H_p(X_i)$ are centered iid variables (in general, H_p is the score function). If moreover the $H_p(X_i)$ follow an iterated logarithm law, and if the compensation $a_{n,p}$ satisfies $\lim_{n\to+\infty} \log\log(n)/a_{n,p} = 0$, then $\hat{q}_n$ tends to q a.s..

For mixture models, Leroux (1992a) showed that the maximum penalized likelihood estimator does not underestimate a.s. the true order and applied the method to foetal movement data (1992b). In this paper, we obtain the same result with lighter assumptions. Moreover, we prove that, asymptotically, for an appropriate penalty term, the maximum penalized likelihood estimator of the number of components for a mixture model does not overestimate the order either. The problem encountered when the order is overestimated is that the model is not identifiable. Take the simple mixture model: $g_{(\pi,\alpha)}$ in $\mathcal{G}_2$, with $\pi = (1-u, u)$, $\alpha = (\gamma_0, \gamma)$, γ_0 fixed, the true model being $g_0 = f_{\gamma_0}$. In this case, g_0 has different representations $g_{\pi,\alpha}$ with different parameters:

$$g_0 = (1-u)\cdot f_{\gamma_0} + u \cdot f_{\gamma_0} = 1 \cdot f_{\gamma_0} + 0 \cdot f_\gamma$$

Define the directional estimators $\hat{u}_\alpha^n$ as the maximum likelihood estimator of u when α is fixed, and $\hat{\alpha}_u^n$ as the maximum likelihood estimator of α when u is fixed. In a standard way, $\sqrt{n}\hat{u}_\alpha^n$ and $\sqrt{n}\hat{\alpha}_u^n$ converge to Gaussian random variables, and define Gaussian processes. But the covariance of these processes is unbounded as γ tends to 0 or u tends to 0, since the directional Fisher informations tend to 0. So, the global maximum likelihood estimators do not converge at all. To deal with this problem, Dacunha-Castelle and Gassiat (1997, 1999) developed the locally conic parameterization, which driving idea is to reparameterize the model. One parameter is identifiable at the previous non-identifiable point and the other parameters contain all the non-identifiability. It is then possible to do a Taylor expansion of the likelihood in the neighborhood of the identifiable parameter to find the asymptotic behavior of the likelihood ratio statistic. Using these results, we prove the consistency of the maximum penalized likelihood estimator. Moreover, with strengthened assumptions, we also show the almost sure consistency. The results are applied to Gaussian and Poisson mixtures. In addition, we give simulation results that show that the method is relevant and not very sensitive to the penalty term.

The paper is organized as follows: in section 2, we give the definition of a locally conic parameterization, define the model, the assumptions and state the convergence in probability. Section 3 deals with the a.s. consistency, and section 4 studies the assumptions for mixtures of Gaussian families or Poisson distributions. The last section is devoted to a numerical study of the penalty term and compares the method with the moment method of Dacunha-Castelle and Gassiat (1997b).

3.2 Convergence in Probability

3.2.1 Notations, Assumptions and Locally Conic Parameterization

Let $X^{(n)}$ be an n-sample of a mixture of q populations, that is

$$P_0^{(n)} = (g_0\nu)^{\otimes n}$$

where g_0 is a mixture of q populations in the parametric family $(f_\gamma)_{\gamma\in\Gamma}$, $\Gamma \subset \mathbb{R}^k$, Γ compact:

$$g_0 = \sum_{l=1}^{q} \pi_l^0 f_{\gamma^{l,0}}$$

The general model is that of Q-mixtures $\mathcal{G}_Q$ given by (1), since the only knowledge is that q is upperbounded by Q.
We assume that $\mathcal{G}_Q$ is identifiable in the weak following sense:

- **(Id)** $\sum_{l=1}^{Q} \pi_l^0 f_{\gamma^{l,0}} = \sum_{l=1}^{Q} \pi_l^1 f_{\gamma^{l,1}} \quad \nu\ a.e. \Longleftrightarrow \sum_{l=1}^{Q} \pi_l^0 \cdot \delta_{\gamma^{l,0}} = \sum_{l=1}^{Q} \pi_l^1 \cdot \delta_{\gamma^{l,1}}$

as probability distributions on Γ. In other words, $\mathcal{G}_Q$ is identifiable if the parameter is the mixing discrete probability distribution on Γ. Teicher (1965) or Yakowitz and Spragins (1968) give sufficient conditions for this kind of weak identifiability, which holds for instance for finite mixtures of Gaussian or Gamma distributions.

We first recall the principle of the local parameterization proposed by Dacunha-Castelle and Gassiat (1999). The main idea is to define a new parameterization for the model: a first parameter θ is identifiable and a second β contains all the non identifiability. Henceforth, we denote the probability density of the mixture as $g_{(\theta,\beta)}$ when we use it with the locally conic parameterization. We have:

$$g_{(\theta,\beta)} = g_0 \Leftrightarrow \theta = 0 \tag{2}$$

The θ parameter can be viewed as a distance to the real model, and β represents the directions for which the submodels approach $\theta = 0$. For all $p \in \{q+1, \dots, Q\}$, define $\mathcal{B}_p$ 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

$$\lambda_i \geq 0,\ \gamma^i \in \Gamma,\ i = 1, \dots, p-q,\ \delta^l \in \mathbb{R}^k,\ \rho_l \in \mathbb{R},\ l = 1, \dots, q \tag{3}$$

$$\sum_{i=1}^{p-q} \lambda_i + \sum_{l=1}^{q} \rho_l = 0 \ \text{ and } \ \sum_{i=1}^{p-q} \lambda_i^2 + \sum_{l=1}^{q} \rho_l^2 + \sum_{l=1}^{q} \|\delta^l\|^2 = 1 \tag{4}$$

Since we shall use partial derivatives of f_γ with respect to γ, we introduce some notations: $D^h_{i_1 \dots i_h}$ will be the h-th partial derivative operator with respect to $\gamma_{i_1} \dots \gamma_{i_h}$. So that $D^h_{i_1 \dots i_h} f_\gamma$ will be the value of this partial derivative of f at point γ. Let now H be the Hilbert space $L^2(g_0 \nu)$. Let then

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

For any β in $\mathcal{B}_p$ 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:

$$g_{(\theta,\beta)} = \sum_{i=1}^{p-q} \lambda_i \frac{\theta}{N(\beta)} f_{\gamma^i} + \sum_{l=1}^{q} (\pi_l^0 + \rho_l \frac{\theta}{N(\beta)}) f_{\gamma^{l,0} + \frac{\theta}{N(\beta)} \delta^l} \tag{5}$$

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 their weight tending to 0.
It leads to the fact that there exists a set $\mathcal{T} = [0, M] \times \tilde{\mathcal{B}}$ such that $\mathcal{G}_p = \{g_{(\theta,\beta)}, (\theta, \beta) \in \mathcal{T}\}$ and (2) holds (see Dacunha-Castelle and Gassiat (1999) for the detailed proof).

We shall use the following assumptions:

- **(P1-a)** There exists a function h in $L_1(g_0 \nu)$ such that: $\forall g \in \mathcal{G}_Q$, $|\log g| \leq h$ ν-a.e.
  (P1-b) f_γ possesses partial derivatives till order 5. For all $h \leq 5$, and all $i_1 \dots i_h$,

$$\frac{D^h_{i_1 \dots i_h} f_{\gamma^0}}{g_0} \in L^3(g_0 \nu)$$

Moreover, there exist random variables m_2, m_3, m_5, independent of γ, and a positive ϵ such that:

$$\sup_{\|\gamma-\gamma^0\|\leq\epsilon} |\frac{D^2_{i_1 i_2} f_\gamma}{g_0}| \leq m_2 \quad E_{g_0\nu}[m_2^3] < +\infty$$

$$\sup_{\|\gamma-\gamma^0\|\leq\epsilon} |\frac{D^3_{i_1 i_2 i_3} f_\gamma}{g_0}| \leq m_3 \quad E_{g_0\nu}[m_3^3] < +\infty$$

$$\sup_{\|\gamma-\gamma^0\|\leq\epsilon} |\frac{D^5_{i_1 \ldots i_5} f_\gamma}{g_0}| \leq m_5 \quad E_{g_0\nu}[m_5^3] < +\infty$$

- **(P2)** For any integer p_1, p_2, such that $p_1 + p_2 \leq Q - q$, for any set of distinct points γ^l, $l = 1, \ldots, p_1$, distinct from any $\gamma^{l,0}$, the set of functions

$$((\frac{f_{\gamma^l}}{g_0})_{l=1,\ldots,p_1}, (\frac{f_{\gamma^{l,0}}}{g_0})_{l=1,\ldots,q}, (\frac{D^1_i f_{\gamma^{l,0}}}{g_0})_{l=1,\ldots,q,i=1,\ldots,k}, (\frac{D^2_{ij} f_{\gamma^{l,0}}}{g_0})_{l=1,\ldots,p_2,i,j=1,\ldots,k})$$

  is linearly independent in H.

- **(P3)** Define $\mathcal{D}$, the subset of the unit sphere of H of functions of form

$$\frac{1}{N(\beta)}\left(\sum_{l=1}^{q} \pi_l^0 \sum_{i=1}^{k} \delta_i^l \frac{D^1_i f_{\gamma^{l,0}}}{g_0} + \sum_{i=1}^{Q-q} \lambda_i \frac{f_{\gamma^i}}{g_0} + \sum_{l=1}^{q} \rho_l \frac{f_{\gamma^{l,0}}}{g_0}\right)$$

  with β in $\tilde{\mathcal{B}}$. Define also ξ_d the Gaussian process indexed by $\mathcal{D}$ with covariance the usual hilbertian product in H. We suppose that $\mathcal{D}$ is a Donsker class (see Van der Vaart and Wellner (1996)), and ξ_d has continuous sample paths. Notice that this assumption allows us to have results of uniform convergence in distribution on $\mathcal{D}$.

- **(C1)** For all $n \geq 1$, the compensation satisfies $a_{n,p+1} \geq a_{n,p} > 0$, $\lim_{n\to+\infty} a_{n,p} = +\infty$, $a_{n,p} = o(n)$ and $\lim_{n\to+\infty} a_{n,p}/a_{n,q} > 1$ for $q < p \leq Q$.

These assumptions **(P1)-(P3)** are identical to those used by Dacunha-Castelle and Gassiat (1999) to prove the convergence in distribution of the maximum likelihood and to identify the limiting distribution. To check **(P3)**, one may use metric entropy or bracketing entropy Theorems (see for instance Ossiander (1987) or Van der Vaart and Wellner (1996)). Those entropy results may also be used to prove the continuity of the process (ξ_d). In fact, computation of entropy behavior is in general easy, since $\overline{\mathcal{D}}$ is a parametric set described with a finite number of parameters varying in a compact euclidian space. Note that the compensation of the BIC criterion satisfies **(C1)**.

3.2.2 Convergence in Probability

We state now our first asymptotic result:

Theorem 3.2.1 *Assume* **(P1)**, **(P2)**, **(P3)**, **(C1)** *and* **(Id)** *hold. If an upper bound* Q *of* q *is known, then* $\hat{q}_n$ *tends to* q *in probability.*

Proof of Theorem 3.2.1

Consider first the case $p < q$. In this case, the proof will be directly done for the almost sure case, and without the need of the assumption on a known upper bound Q. For all $g_p \in \mathcal{G}_p$, $E_{g_0\nu}(\log g_p)$ is defined using **(P1-a)** assumption. Define

$$K(g_0, \mathcal{G}_p) = \inf_{g_p \in \mathcal{G}_p} K(g_0, g_p).$$

with $K(g_0, g_p) = E_{g_0\nu}(\log(g_0/g_p))$ the Kullback distance between g_0 and g_p. Since the set of the π and α parameters is compact and the Kullback distance is continuous with respect to these parameters, the function $K(g_0, g_p)$ attains its infimum for some $\tilde{g}_p \in \mathcal{G}_p$, and then $K(g_0, \mathcal{G}_p) > 0$ using **(Id)**. We first prove that, a.s.,

$$\lim_{n\to+\infty} \frac{1}{n}(T_{n,p} - l_n(g_0)) \to -K(g_0, \mathcal{G}_p).$$

By definition of the maximum likelihood, we have:

$$\frac{1}{n}(T_{n,p} - l_n(g_0)) \geq \frac{1}{n}(l_n(\tilde{g}_p) - l_n(g_0)),$$

Applying the strong law of large numbers, we get almost surely

$$\liminf_{n\to+\infty} \frac{1}{n}(T_{n,p} - l_n(g_0)) \geq -K(g_0, \tilde{g}_p).$$

Define $m_\eta(x, \mathcal{G}_p) = \sup_{d((\pi_1,\alpha_1),(\pi_2,\alpha_2))\leq\eta} |\log g_{(\pi_1,\alpha_1)_p}(x) - \log g_{(\pi_2,\alpha_2)_p}(x)|$. As the set of parameters is compact, there exists a finite number N_η of balls with diameter $\eta/2$ covering it and with center $c_i = (\pi_i, \alpha_i)_p$, $i = 1 \ldots N_\eta$. Now, obviously

$$\frac{1}{n}(T_{n,p} - l_n(g_0)) \leq \sup_{i=1\ldots N_\eta} \frac{1}{n}(l_n(g_{c_i}) - l_n(g_0)) + \frac{1}{n}\sum_{j=1}^{n} m_\eta(X_j)$$

so that a.s.

$$\begin{aligned}
\limsup_{n\to+\infty} \frac{1}{n}(T_{n,p} - l_n(g_0)) &\leq \sup_{i=1\ldots N_\eta} (-K(g_0, g_{c_i})) + E_{g_0\nu}(m_\eta(X)) \\
&\leq -\inf_{i=1\ldots N_\eta} (-K(g_0, g_{c_i})) + E_{g_0\nu}(m_\eta(X)) \\
&\leq -K(g_0, \tilde{g}_p) + E_{g_0\nu}(m_\eta(X))
\end{aligned}$$

Now, we have $\lim_{\eta\to 0} m_\eta(x) = 0$ a.s.. Moreover, **(P1-a)** implies $\lim_{\eta\to 0} E_{g_0}(m_\eta(X)) = 0$, so that $\limsup_{n\to+\infty} \frac{1}{n}(T_{n,p} - l_n(g_0)) \leq -K(g_0, \tilde{g}_p)$, and we may conclude that $\lim_{n\to+\infty} \frac{1}{n}(T_{n,p} - l_n(g_0))$ converges a.s. to $-K(g_0, \mathcal{G}_p)$. The rest of this part of the proof is restricted to this event of probability 1. According to the condition on the rate of growth of $a_{n,p}$ and the fact that $K(g_0, \mathcal{G}_p) > 0$, we can write for large n and for all $p < q$

$$T_{n,q} - T_{n,p} > a_{n,q} - a_{n,p}$$

We deduce that $\liminf_{n\to+\infty} \hat{q}_n \geq q$.

Consider now the case $p \geq q$.
Dacunha-Castelle and Gassiat (1999) used the locally conic parameterization of a mixture to prove that under the assumptions **(P1)** to **(P3)**, **(Id)**, $\hat{\theta}_p$ is consistent in probability and $T_{n,p} - l_n(g_0)$ converges in distribution to the following variable:

$$\sup\left\{\sup_{d\in\mathcal{D}} \frac{1}{2}\xi_d^2 1_{\xi_d\geq 0};\ \sup_{d_1\in\mathcal{D}_1, d_2\in\mathcal{D}_2} \frac{1}{2}(\xi_{d_1}^2 + \xi_{d_2}^2 1_{\xi_{d_2}\geq 0})\right\}$$

where $\mathcal{D}_1$ is the subset of the unit sphere of H of functions of form

$$\sum_{l=1}^{q}\sum_{i=1}^{k} \lambda_{l,k} \frac{D_i^1 f_{\gamma^{l,0}}}{g_0}$$

and $\mathcal{D}_2$ is the subset of the unit sphere of H of functions of form

$$\sum_{l=1}^{p_1} \mu_l \frac{f_{\gamma^l}}{g_0} + \sum_{l=1}^{q} \tilde{\rho}_l \frac{f_{\gamma^{l,0}}}{g_0} + \sum_{l=1}^{q}\sum_{i=1}^{k} \lambda_{l,i} \frac{D_i^1 f_{\gamma^{l,0}}}{g_0} + \sum_{l=1}^{p_2}\sum_{i,j=1}^{k} \tau_l a_i a_j \frac{D_{ij}^2 f_{\gamma^{l,0}}}{g_0}$$

with $p_1 \leq p-q-1$, $p_1 + p_2 \leq p - q$, $\mu_l \geq 0$, $\sum_{l=1}^{p_1} \mu_l + \sum_{l=1}^{q} \tilde{\rho}_l = 0$, $\tau_l \geq 0$, which are orthogonal to $\mathcal{D}_1$.
Notice that $\mathcal{D}_1$ and $\mathcal{D}_2$ are subsets of $\overline{\mathcal{D}}$, the (compact) closure of $\mathcal{D}$ in H. More precisely, $\overline{\mathcal{D}}$ is the subset of the unit sphere of H of functions of form

$$\sum_{l=1}^{p_1} \lambda_l \frac{f_{\gamma^l}}{g_0} + \sum_{l=1}^{q} \tilde{\rho}_l \frac{f_{\gamma^{l,0}}}{g_0} + \sum_{l=1}^{q}\sum_{i=1}^{k} \mu_{l,i} \frac{D_i^1 f_{\gamma^{l,0}}}{g_0} + \sum_{l=1}^{p_2}\sum_{i,j=1}^{k} \tau_l a_i a_j \frac{D_{ij}^2 f_{\gamma^{l,0}}}{g_0}$$

with $p_1 + p_2 \leq p - q$, $\lambda_l \geq 0$, $\sum_{l=1}^{p_1} \lambda_l + \sum_{l=1}^{q} \tilde{\rho}_l = 0$, $\tau_l \geq 0$.

Consider now the probability of $\hat{q}_n$ being greater than q:

$$P(\{\hat{q}_n > q\}) \leq \sum_{p=q+1}^{Q} P(\{W_n(p) > W_n(q)\})$$

We have

$$P(\{W_n(p) > W_n(q)\}) = P(\{\frac{T_{n,p} - T_{n,q}}{a_{n,q}} > \frac{a_{n,p}}{a_{n,q}} - 1\})$$

Now, for $\epsilon > \sup_{p \in q+1,\ldots,Q}(\frac{a_{n,p}}{a_{n,q}} - 1)$, $\lim_{n \to +\infty} P(\{\frac{T_{n,p}-T_{n,q}}{a_{n,q}} \geq \epsilon\}) = 0$ because $T_{n,p}$ converges in distribution for $p \geq q$ and $\frac{1}{a_{n,q}}$ tends to 0.
Therefore, $\frac{T_{n,p}-T_{n,q}}{a_{n,q}} - \frac{a_{n,p}}{a_{n,q}} + 1$ tends to $1 - \frac{a_{n,p}}{a_{n,q}}$ in probability, which is a strictly negative term.
So, $\lim_{n \to +\infty} P(\{W_n(p) > W_n(q)\}) = 0$ and $\hat{q}_n$ tends to q in probability.

Define now $\hat{\pi}_{\hat{q}_n}$ and $\hat{\alpha}_{\hat{q}_n}$ as the maximum likelihood estimator of the parameters of the mixture when the estimated order is $\hat{q}_n$.

Corollary 3.2.2 *Suppose that the assumptions of Theorem 3.2.1 hold. Then,* $(\hat{\pi}_{\hat{q}_n}, \hat{\alpha}_{\hat{q}_n})$ *tends to* (π_0, α_0) *in probability.*

Fix $\gamma^f = (\gamma^{1,f}, \ldots, \gamma^{Q,f})$. For a mixture density of order r, and of parameters
$(\pi, \alpha)_r = (\pi_1, \ldots, \pi_r, \gamma^1, \ldots, \gamma^r)$ define, for $p > r$:
$(\pi, \alpha)_p = (\pi_1, \ldots, \pi_r, 0, \ldots, 0, \gamma^1, \ldots, \gamma^r, \gamma^{r+1,f} \ldots, \gamma^{p,f})$. We can write:

$$P(\{\|(\hat{\pi}_{\hat{q}_n}, \hat{\alpha}_{\hat{q}_n})_Q - (\pi_0, \alpha_0)_Q\| > \epsilon\}) \leq P(\{\hat{q}_n \neq q\}) + P(\{\|(\hat{\pi}_q, \hat{\alpha}_q) - (\pi_0, \alpha_0)\| > \epsilon\})$$

Now, $P(\{\hat{q}_n \neq q\})$ tends to 0 using Theorem 3.2.1. Moreover, when the order is known, the maximum likelihood estimator of the mixture parameters is consistent in the quotient topology defined by the permutations of the parameters (Leroux (1992)). So that the second term also tends also to 0 in the quotient topology and this achieves the proof of corollary 3.2.2.

3.3 Almost Sure Consistency of the Estimated Number of Components

3.3.1 The Order Is not Underestimated

As Leroux (1992a) proved it under stronger assumptions, but for non finite mixtures, $\hat{q}_n$ does not asymptotically underestimate the number of components of the true mixing distribution.

Theorem 3.3.1 *Under the assumptions* **(P1-a), (C1)** *and* **(Id)**, $\liminf_{n\to+\infty} \hat{q}_n \geq q$ *with probability 1.*

For the proof of this theorem, see section 3.2.1.

3.3.2 Almost Sure Consistency

In this section, we strengthen the Donsker property on $\mathcal{D}$ in a uniform law of the iterated logarithm:

- **(P4)** For all $d \in \mathcal{D}$ define $Z_n(d) = \sum_{i=1}^{n} d(X_n)$. Then
$$\limsup_{n\to+\infty} \frac{\|Z_n(d)\|_s}{(2n\log\log(n))^{1/2}} \leq K \quad a.s.$$
with $\|.\|_s$ the supremum norm on $\mathcal{D}$, $K > 0$.

We also strengthen the $a_{n,p}$ behavior in the following sense:

- **(C2)** The compensation satisfies $\lim_{n\to+\infty} \frac{loglogn}{a_{n,p}} = 0$

A sufficient condition to have **(P4)**, as proved in Dudley and Philipp (1983), is that $\mathcal{D}$ satisfies **(P3)** and $E(\frac{\|Z_1(d)\|_s^2}{L_2\|Z_1(d)\|_s}) < \infty$, with $L_2 = L \circ L$ and $Lx = \log(x \vee e)$. These assumptions are studied for mixtures of Gaussian families or Poisson distributions in section 4. We have the following asymptotic result:

Theorem 3.3.2 *Assume* **(P1)** *to* **(P4)**, **(C1)**, **(C2)** *and* **(Id)** *hold. If a superior value* Q *of* q *is known, then*

- **(a)** $\hat{q}_n$ *tends to* q *a.s..*
- **(b)** $(\hat{\pi}_{\hat{q}}, \hat{\alpha}_{\hat{q}})$ *tends to* (π_0, α_0) *a.s..*

It remains to be proved that $\hat{q}_n$ does not asymptotically overestimate the order a.s.. Let $\hat{\theta}_p$ be the estimator of $\theta_0 = 0$ in $\mathcal{G}_p$, for $q \leq p \leq Q$. Theorem 2.1 of Dacunha-Castelle and Gassiat (1997) states the convergence of $\hat{\theta}_p$ in probability. A careful read to their proof allows us to see that the convergence holds a.s.. Moreover, $\sup_\beta \hat{\theta}_{p,\beta}$, where $\hat{\theta}_{p,\beta}$ is the directional estimator when β is fixed, also converge a.s.. Now, the following lemma will give us the key of the proof:

Lemma 3.3.3 *Assume that the assumptions of Theorem 3.3.2 hold. Then,*

$$\limsup_{n\to+\infty} \frac{T_{n,p} - l_n(g_0)}{\log\log(n)}$$

is a.s bounded for $Q \geq p \geq q$.

Consider now $\Omega^* = \{\omega; \hat{q}_n(\omega) > q\} \subset \bigcup_{p=q+1}^{Q} \Omega^*_{n,p}$, with $\Omega^*_{n,p} = \{\omega; \hat{q}_n(\omega) = p\}$.
For $\omega \in \Omega^*_{n,p}$, $W_n(p) > W_n(q)$ and we have

$$W_n(p) > W_n(q) \Leftrightarrow \frac{T_{n,p} - T_{n,q}}{a_{n,q}} - \frac{a_{n,p}}{a_{n,q}} + 1 > 0$$

Now, $\lim_{n\to+\infty} \frac{T_{n,p}-T_{n,q}}{a_{n,q}}$ tends a.s. to 0 because of lemma 3.3.3, **(P4)** and **(C2)**. Therefore, $\frac{T_{n,p}-T_{n,q}}{a_{n,q}} - \frac{a_{n,p}}{a_{n,q}} + 1$ tends a.s. to a strictly negative term.
So $P(\limsup_{n\to+\infty}\{W_n(p) > W_n(q)\}) = 0$, which gives $P(\limsup_n \Omega^*_{n,p}) = 0$ for any $p = q+1, \ldots, Q$. We deduce that $\hat{q}_n$ tends to q a.s..

Proof of lemma 3.3.3

Let $V_{n,p}(\beta)$ be the log likelihood ratio statistic when β is fixed: $V_{n,p}(\beta) = l_n(g_{(\hat{\theta}_{p,\beta},\beta)})$. We have:

$$T_{n,p} = \sup_{\beta\in\tilde{\mathcal{B}}} V_{n,p}(\beta)$$

The proof of lemma 3.3.3 follows the same steps as those used by Dacunha-Castelle and Gassiat (1999) for their Theorem 3.1. Henceforth, we have to carefully check that all the steps of their proof still holds in the a.s. case, with the new assumption **(P4)**. For the sake of simplicity, we will now omit the p subscript. The method consists in separating the domain in two regions:

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

$$B_n = \{\beta \ : \ \exists l = 1, \ldots, q, \ \frac{\|\delta^l\|}{N(\beta)^2} \geq \frac{1}{\eta_n^\alpha}\}$$

for $\alpha < 3/4$ and $\eta_n = \sup_{\beta\in\mathcal{B}} \hat{\theta}_\beta$. The reason of these two domains is that on B_n, $N(\beta)$ and all δ^l tends uniformly to 0, so that there exists l such that $\frac{||\delta^l||}{N(\beta)^2}$ is not uniformly bounded. As this expression is found in the Taylor expansion of the likelihood, this leads to do this expansion until order 5 on B_n, while it is sufficient to do it until order 2 on

A_n, to prove the result.
Then we have a.s.

$$T_{n,p} - l_n(g_0) = \sup\left\{\sup_{\beta\in A_n} l_n(g_{(\hat\theta_\beta,\beta)}), \sup_{\beta\in B_n} l_n(g_{(\hat\theta_\beta,\beta)})\right\} - l_n(g_0)$$

Let us study what happens on each region A_n and B_n.

On A_n, they prove

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

with

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

for β in A_n and where the $o(\cdot)$ is uniform in probability over β in A_n. The delicate point is to check the almost sure uniformity, which will flow from the use of **(P4)** during the careful read of their proof. We can then deduce that:

$$\limsup_{n\to+\infty}\sup_{\beta\in A_n}\frac{l_n(g_{(\hat\theta_\beta,\beta)}) - l_n(g_0)}{\log\log(n)}$$

is a.s bounded.

On the second domain B_n, they show that it is necessary to perform the Taylor expansion until order 5 to obtain:

$$V_n(\beta) =$$
$$\left\{\frac{1}{2n}\left(\frac{(D_n(\beta) - \frac{F_n(\beta)U(\beta)}{S(\beta)})^2}{1-\frac{(U(\beta))^2}{S(\beta)}}\right)1_{D_n(\beta)-\frac{F_n(\beta)U(\beta)}{S(\beta)}\geq 0} + \frac{1}{2n}\left(\frac{(F_n(\beta))^2}{S(\beta)}\right)\right\}(1+o(1))$$

uniformly in probability, with

$$F_n(\beta) = \sum_{l=1}^{q}\sum_{i=1}^{k}\rho_l\delta_i^l\sum_{j=1}^{n}\frac{D_i^1 f_{\gamma^{l,0}}}{g_0}(X_j)$$

$$S(\beta) = \sum_{l,l'=1}^{q}\sum_{i,i'=1}^{k}\rho_l\rho_{l'}\delta_i^l\delta_{i'}^{l'}\langle\frac{D_i^1 f_{\gamma^{l,0}}}{g_0}, \frac{D_{i'}^1 f_{\gamma^{l',0}}}{g_0}\rangle_H$$

$$U(\beta) = \sum_{l=1}^{q}\sum_{i=1}^{k}\rho_l\delta_i^l\langle\frac{D_i^1 f_{\gamma^{l,0}}}{g_0}, \frac{g'_{(0,\beta)}}{g_0}\rangle_H$$

Again, **(P4)** enables the achievement of the same result a.s. uniformly, so that

$$\limsup_{n\to+\infty} \sup_{\beta\in B_n} \frac{l_n(g_{(\hat{\theta}_\beta,\beta)}) - l_n(g_0)}{\log\log(n)}$$

is a.s bounded.

Then, $\limsup_{n\to+\infty} \frac{T_{n,p}-l_n(g_0)}{\log\log(n)}$ is a.s. bounded.

3.4 Examples

We now check that Theorems 3.3.2 and 3.2.1 hold for Gaussian and Poisson mixtures.

3.4.1 The Gaussian Case

Theorem 3.4.1 *Theorems 3.3.2 and 3.2.1 hold for $\mathcal{F} = (f_M)_{M\in\Gamma}$, $M = (m_1, \ldots, m_k)$, Γ compact of $\mathbb{R}^k$, being a family of densities of the form*

$$f_M(x) = \frac{1}{(\sigma\sqrt{2\pi})^k} \exp -(\frac{\sum_{i=1}^k (x_i - m_i)^2}{2\sigma^2})$$

where M is unkown.

Proof of 3.4.1

In this case, the identifiability assumption **(Id)** is checked by Yakowitz and Spragins (1968) and **(P2)** is immediate.
The expressions of the partial first and second derivatives are:

$$D_i^1 f_M(x) = -\frac{x_i - m_i}{\sigma^2} f_M(x)$$

$$D_{i,i'}^2 f_M(x) = \frac{(x_{i'} - m_{i'})(x_i - m_i)}{\sigma^4} f_M(x)$$

$$D_{i,i}^2 f_M(x) = \frac{(x_i - m_i)^2 - \sigma^2}{\sigma^4} f_M(x)$$

Define now $\bar{m} = \max_{i=1,\ldots,k}(\sup_{M\in\Gamma} m_i)$, $\underline{m} = \min_{i=1,\ldots,k}(\inf_{M\in\Gamma} m_i)$, $\bar{M} = (\bar{m}, \ldots, \bar{m})$, $\underline{M} = (\underline{m}, \ldots, \underline{m})$ and $\underline{\pi} = \min_{l=1,\ldots,q} \pi_l^0$ (which is starkly positive). Observe we have the following majorations for $f_M \in \mathcal{F}$: $\underline{f} \leq f_M \leq \bar{f}$ with

$$\underline{f}(x) = f_{\bar{M}}(x) \quad , \quad \bar{f}(x) = f_{\underline{M}}(x) \quad \text{for} \quad x_i \leq \underline{m}$$

$$\underline{f}(x) = f_{\underline{M}}(\frac{\bar{m} + \underline{m}}{2})(x) \quad , \quad \bar{f}(x) = \frac{1}{(\sigma\sqrt{2\pi})^k} \quad \text{for} \quad \underline{m} \leq x_i \leq \bar{m}$$

$$\underline{f}(x) = f_{\underline{M}}(x) \quad , \quad \bar{f}(x) = f_{\bar{M}}(x) \quad \text{for} \quad x_i \geq \bar{m}$$

$$q\underline{\pi}\underline{f} = \underline{g_0} \leq g_0 \leq \bar{g_0} = q\bar{f}$$

$$\sup_{M \in \Gamma} \frac{f_M}{g_0} \leq \frac{\bar{f}}{\underline{g_0}}$$

All these functions are square integrable with respect to $g_0\nu$. In the same way, the supremum norm of products of functions of $(\frac{f_M}{g_0}, \frac{D_i^1 f_M}{g_0}, \frac{D_{ij}^2 f_M}{g_0})$ is square integrable for $g_0\nu$. This proves **(P1)**.

We are going to prove now that $E_{g_0\nu}(\|d(\beta)(X)\|_s^2) < \infty$. In the region where $N(\beta)$ does not tend to 0, the result is immediate using the previous majorations. If $N(\beta)$ tends to 0, there exists at least one M^l which converges to any of the $M^{l,0}$. Writing a Taylor expansion, the possible accumulation points of $d(\beta)$ are the $\tilde{d}(\mu, \tilde{\rho}, \lambda, \tau, a)(\cdot)$ in the unit sphere of H:

$$\sum_{l=1}^{p_1} \mu_l \frac{f_{M^l}}{g_0} + \sum_{l=1}^{q} \tilde{\rho}_l \frac{f_{M^{l,0}}}{g_0} + \sum_{l=1}^{q}\sum_{i=1}^{k} \lambda_{l,i} \frac{D_i^1 f_{M^{l,0}}}{g_0} + \sum_{l=1}^{p_2}\sum_{i,j=1}^{k} \tau_l a_i a_j \frac{D_{ij}^2 f_{M^{l,0}}}{g_0}$$

with $p_1 \leq p - q - 1$, $p_1 + p_2 \leq p - q$, $\mu_l \geq 0$, $\sum_{l=1}^{p_1} \mu_l + \sum_{l=1}^{q} \tilde{\rho}_l = 0$, $\tau_l \geq 0$, where all the coefficients are bounded. On the subsequence,

$$d(\beta) = \tilde{d}(\mu, \tilde{\rho}, \lambda, \tau, a)(1 + o(1))$$

We can then find a $\bar{d}$ which is $g_0\nu$ -square integrable such that $|d(\beta)(x)| \leq \bar{d}(x)$. $E(\|d(\beta)(X)\|_s^2) < \infty$ follows, which gives us **(P4)**.

We check now the Donsker property **(P3)**: $\int_o^{+\infty} \sqrt{\log N(\epsilon)} d\epsilon < \infty$ where $\mathcal{N}(\epsilon)$ is the bracketing number. Recall that given two functions l and u, the $[l, u]$ bracket is the set of all functions f with $l \leq f \leq u$. An ϵ-bracket is a bracket with $\|u - l\|_H < \epsilon$. The bracketing number $\mathcal{N}(\epsilon)$ is the minimum number of ϵ-brackets needed to cover $\mathcal{D}$. Note that l and u need not belong to $\mathcal{D}$ themselves, but are assumed to have finite norms. $N(\epsilon)$ being equal to 1 for a sufficiently large ϵ, we only have to check the order of $\mathcal{N}(\epsilon)$ when ϵ tends to 0.

If we divide each set of parameter of $\tilde{\mathcal{B}}$ in intervals of length ϵ, we obtain a partition of $\mathcal{D}$ of cardinal of order $O(\frac{1}{\epsilon^K})$, with $K = (k+1)Q$. Let β^* be the parameter of the center of a pavement $P = \otimes_{i=1}^{K}[\beta_i^* - \frac{\epsilon}{2}, \beta_i^* + \frac{\epsilon}{2}]$. We check that the bracket $[\inf_{\beta \in P} d(\beta), \sup_{\beta \in P} d(\beta)]$

is of order ϵ by writing a Taylor expansion of $d(\beta)$ around $d(\beta^*)$ on the pavements where $N(\beta)$ does not tend to 0, and around $\tilde{d}(\beta^*)$ on the pavements where $N(\beta)$ tends to 0. We obtain that $E_{g_0\nu}(\inf_{\beta\in P} d(\beta) - \sup_{\beta\in P} d(\beta))^2$ is an $O(\epsilon)$ thanks to the fact that the products of functions of $(\frac{f_M}{g_0}, \frac{D_i^1 f_M}{g_0}, \frac{D_{ij}^2 f_M}{g_0})$ are square integrable for $g_0\nu$. $\mathcal{N}(\epsilon)$ is then of the order of $\frac{1}{\epsilon^K}$ and the property holds.

Proposition 3.4.2 *Theorems 3.2.1 and 3.3.2 hold for Gaussian families with an unknown variance lower bounded by a strictly positive number.*

The proof follows the same lines as 3.4.1, using the fact that the variance is lower bounded by a strictly positive number.

3.4.2 Poisson Distributions

Theorem 3.4.3 *Theorems 3.2.1 and 3.3.2 hold for* $\mathcal{F} = (f_\gamma)_{\gamma\in\Gamma}$*,* $\Gamma = [\underline{\gamma}, \bar{\gamma}]$*,* $\underline{\gamma} > 0$ *being a family of densities of the form*

$$f_\gamma(x) = e^{-x}\frac{\gamma^x}{x!}$$

with respect to the measure $\nu = \sum_{x=0}^{+\infty} \delta_x$*.*

Proof of 3.4.3

The useful majorations are as follows:

$$e^{-\bar{\gamma}+\underline{\gamma}} f_{\underline{\gamma}} \le f_\gamma \le e^{-\underline{\gamma}+\bar{\gamma}} f_{\bar{\gamma}}$$

$$\underline{\pi} q f_{\underline{\gamma}} \le \sum_{l=1}^{q} \pi_l^0 f^{l,0} \le q \bar{f}_\gamma$$

The derivatives are:

$$D^1 f_\gamma(x) = (\frac{x}{\gamma}) f_\gamma(x)$$

$$D^2 f_\gamma(x) = ((\frac{x}{\gamma})^2 - \frac{x}{\gamma^2}) f_\gamma(x)$$

and the products of functions of $(\frac{f_\gamma}{g_0}, \frac{D^1 f_\gamma}{g_0}, \frac{D^2 f_\gamma}{g_0})$ are square integrable for $g_0\nu$.

3.5 Numerical Experiments

We numerically study the influence of the penalty term and compare our method with the moment method of Dacunha-Castelle and Gassiat (1997b) for two different true distributions of two Gaussian components of same variance $\sigma = 1/2$:

$$g_0(x) = \pi_1 \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\gamma^1)^2}{2\sigma^2} + \pi_2 \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(x-\gamma^2)^2}{2\sigma^2}.$$

- **D1**: $\gamma^1 = -1$, $\gamma^2 = 1$, $\pi_1 = \pi_2 = \frac{1}{2}$
- **D2**: $\gamma^1 = -\frac{1}{2}$, $\gamma^2 = \frac{1}{2}$, $\pi_1 = \frac{9}{10}$, $\pi_2 = \frac{1}{10}$

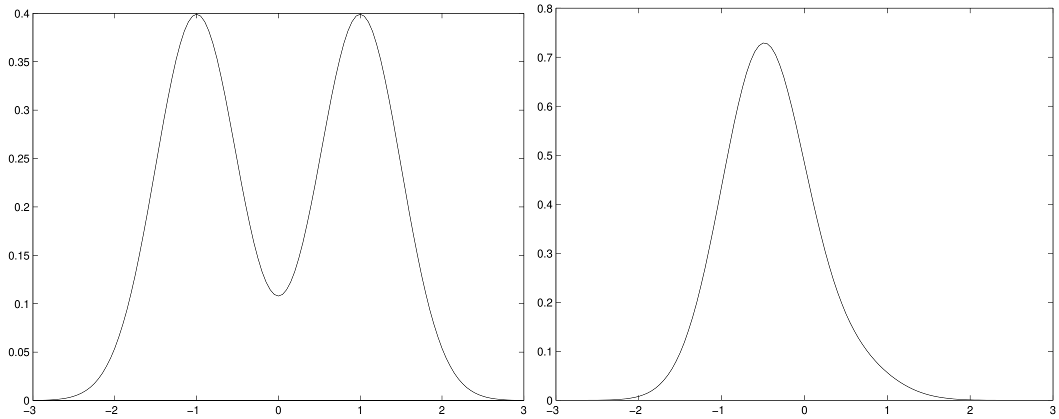

Figure 3.1 : *The distribution D1 (on left) and D2 (on right)*

On the first distribution clearly appears the two components, while it is not the case for the second one, as shown on Fig. 3.1. For each model, we draw 20 samples of lengths $n = 200, 500, 1000, 1500, 2000$ and apply the maximum penalized likelihood method and the moment method with the upper bound $Q = 3$.

For the maximum penalized likelihood method, we use, for $p > 1$, the EM algorithm (Dempster *and al.* (1977), Redner and Walker (1984)) to find the maximum value of the likelihood in $\mathcal{G}_1$, $\mathcal{G}_2$, and $\mathcal{G}_3$. This algorithm is particularly adapted to models with incomplete data. In our case, write the random variables X_i as

$$X_i = \Phi_i + Y_i, \tag{6}$$

where Φ_i and Y_i are mutually independent, Y_i is the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ and Φ_i is the mixing distribution $\phi = \sum_{l=1}^{q} \pi_l^0 \delta_{\gamma^{l,0}}$. The complete data set is (X_i, Φ_i). The EM algorithm is an iterative algorithm such that, at each step s, two consecutive operations are done:

- The computation of the Expectation, conditionally to the a posteriori distribution of the observations, of the log-likelihood of the complete data. For this, we have to compute the following quantity:

$$\hat{u}_l^{(s)}(i) = \frac{\pi_l^{(s)} f_{\gamma^{l,(s)}}(X_i)}{\sum_{l=1}^{p} \pi_l^{(s)} f_{\gamma^{l,(s)}}(X_i)}, \quad l = 1, \ldots, p$$

  which is the current estimated value of the conditional probability that observation i was generated by the l-th component, and s is the iteration number.

- The Maximization of this expectation to find a new set of parameters. We have here the explicit formula:

$$\pi_l^{(s+1)} = \frac{\sum_{i=1}^{n} \hat{u}_l^{(s)}(i)}{\sum_{l=1}^{p} \sum_{i=1}^{n} \hat{u}_l^{(s)}(i)}, \quad \lambda^{l,(s+1)} = \frac{\sum_{i=1}^{n} \hat{u}_l^{(s)}(i) X_i}{\sum_{i=1}^{n} \hat{u}_l^{(s)}(i)}, \qquad l = 1, \ldots, p.$$

This algorithm is known to increase the likelihood at each step, but to only reach a local minimum of the likelihood function, so that we start from a grid of initial values. For each starting value, we stop the algorithm as soon as the difference between the current likelihood value and the next estimated is less that 10^{-10} or when the number of steps is larger than 2500, for time execution reasons. Then, we penalized with the following functions:

$$a_{n,p} = \frac{p}{2}(\log n)^{\alpha} = \frac{p}{2} v_n, \quad \alpha = 1/2, 1, 2, 3, 4.$$

Notice that $\alpha = 1$ is the compensation term of the BIC criterion.

In case of the moment method, we follow the method and algorithm proposed by Dacunha-Castelle and Gassiat (1997b). The base principle is to characterize the number of support points of the mixing distribution by means of the Hankel matrices of the first algebraic moments. Namely, for one-dimensional distributions, the cardinality q of the mixing distribution is proved to be the least integer such that the Hankel matrix degenerates:

$$q = \inf\{p, \det H(p) = 0\} \tag{7}$$

where $H(p)$ is the Hankel matrix of order p of the moments of the mixing distribution ϕ:

$$H(p)_{i,j} = E_\phi(\Phi^{i+j-2}), \quad 1 \leq i, j \leq p+1.$$

Assume now we have a consistent estimator $\hat{c}_n^p$ of $E_\phi(\Phi^p)$, based on independent observations $X_1, \ldots, X_n$, and denotes through $\hat{H}(p)$ the Hankel matrix based on the estimated values $\hat{c}_n^p$. Equation (7) could lead to choose an estimator of the order q as the minimizer of $|\det \hat{H}(p)|$, and would lead a.s. to choose p as large as possible. To overcome this problem, a penalty term $A(p)b_n$ is introduced, such that $A(p)$ is a strictly increasing function and $\lim_{n\to+\infty} b_n = 0$. The estimator $\hat{q}_n$ is finally defined as the minimizer of

$$|\det \hat{H}(p)| + A(p)b_n$$

over all $\mathbb{N}$. In case of translation mixture, when $\gamma \subset \mathbb{R}$ and $\lim_{n\to+\infty} \sqrt{n} b_n = +\infty$, as soon as the components have finite moments of all orders, $\hat{q}_n$ tends to q in probability.

The algorithm is simple. Using (6), we have:

$$E_{g_0\nu}(X_i^p) = \sum_{l=0, l \text{ even}}^{p} \frac{p!}{l!(p-l)!} \frac{\sigma^{2l}(2l)!}{l! 2^l} E_\phi(\Phi_i^{p-l}),$$

leading to a triangular linear system for the computations of $\hat{c}_n^p$, that can be solved by replacing $E_{g_0\nu}(X_i^p)$ by the moment estimator $\frac{1}{n}\sum_{i=1}^n X_i^p$. Then, its remains to compute the determinant and apply the penalization term. The rate b_n of the penalization in this case is different from the rate v_n in the maximum penalized likelihood case. However, b_n is of the same order as $\sqrt{v_n/n}$, so that we choose

$$A(p)b_n = \frac{p}{2}\sqrt{\frac{(\log n)^\alpha}{n}}, \quad \alpha = 1/2, 1, 2, 3, 4.$$

The results are summarized in the following tables. For each sample size, we write the estimated orders for the two methods and for the 20 samples. When the two components are clearly visible, case **D1**, the maximum penalized likelihood method gives excellent results, and is not very sensitive to the choice of the penalty term. The convergence is fast: the heaviest penalty term leads to under-estimate the value of the order for lowest size samples, but as soon as $n = 1000$, all the penalty terms lead to the true value of the order. The results are less evident for the moment method. For $n = 200$ and $n = 500$, it remains sensitive to the value of the penalization. For higher values of n, the results are better, but not as best as these of the maximum penalized likelihood method.

		Likelihood			Moment		
		$\hat{q}_n = 1$	$\hat{q}_n = 2$	$\hat{q}_n = 3$	$\hat{q}_n = 1$	$\hat{q}_n = 2$	$\hat{q}_n = 3$
n=200	α=1/2	0	19	1	0	6	14
	α=1	0	20	0	0	11	9
	α=2	0	20	0	0	19	1
	α=3	0	20	0	0	20	0
	α=4	20	0	0	18	2	0
n=500	α=1/2	0	18	2	0	9	11
	α=1	0	20	0	0	9	11
	α=2	0	20	0	0	20	0
	α=3	0	20	0	0	20	0
	α=4	20	0	0	5	15	0
n=1000	α=1/2	0	18	2	0	8	12
	α=1	0	20	0	0	11	9
	α=2	0	20	0	0	20	0
	α=3	0	20	0	0	20	0
	α=4	0	20	0	0	20	0
n=1500	α=1/2	0	18	2	0	5	15
	α=1	0	20	0	0	9	11
	α=2	0	20	0	0	19	1
	α=3	0	20	0	0	20	0
	α=4	0	20	0	0	20	0
n=2000	α=1/2	0	18	2	0	9	11
	α=1	0	20	0	0	13	7
	α=2	0	20	0	0	19	1
	α=3	0	20	0	0	20	0
	α=4	0	20	0	0	20	0

Comparison of the two methods for the case **D1**

In the second case **D2**, the rate of the maximum penalized likelihood method is slower than in the first case, and the results are under-estimated for the heaviest penalizations, even for $n = 2000$: it is more difficult to separate the two components, that are seen in a same mode. However, for $\alpha = 1, 2$ the results gives excellent results for every n, and this is an argument to use the BIC criterion ($\alpha = 1$). The moment method remains more dependent on the penalization, and we clearly see that it converges slower than the maximum penalized method: as an example, for $\alpha = 2$, it has not converged yet, whereas it has already converged for $n = 1000$ in case of the likelihood method.

		Likelihood			Moment		
		$\hat{q}=1$	$\hat{q}=2$	$\hat{q}=3$	$\hat{q}=1$	$\hat{q}=2$	$\hat{q}=3$
n=200	α=1/2	0	19	1	4	16	0
	α=1	4	16	0	10	10	0
	α=2	20	0	0	20	0	0
	α=3	20	0	0	20	0	0
	α=4	20	0	0	20	0	0
n=500	α=1/2	0	19	1	0	20	0
	α=1	0	20	0	1	19	0
	α=2	11	9	0	20	0	0
	α=3	20	0	0	20	0	0
	α=4	20	0	0	20	0	0
n=1000	α=1/2	0	18	2	0	20	0
	α=1	0	19	1	0	20	0
	α=2	2	18	0	19	1	0
	α=3	20	0	0	20	0	0
	α=4	20	0	0	20	0	0
n=1500	α=1/2	0	18	2	0	20	0
	α=1	0	20	0	0	20	0
	α=2	0	20	0	11	9	0
	α=3	20	0	0	20	0	0
	α=4	20	0	0	20	0	0
n=2000	α=1/2	0	19	1	0	20	0
	α=1	0	20	0	0	20	0
	α=2	0	20	0	5	15	0
	α=3	20	0	0	20	0	0
	α=4	20	0	0	20	0	0

Comparison of the two methods for the case **D2**

On these examples, the moment method, even if it is simple and fast to implement, is not better than the maximum likelihood estimator, which gives stable results, with faster rate, and few dependent on the compensation. It remains to enhance the way to find the maximum value of the likelihood, because the response time becomes prohibitive when the number of components increases. Other algorithm to find the maximum likelihood value could be investigated, such as simulated annealing method, which are not dependent on the initial point. These simulations are a first attempt to study the effect of the penalization on the result. A complete study still needs to be driven in a further work. In any cases, the BIC criterion is seen to be very reliable.

Bibliographie

[1] Akaike H. Statistical predictor identification. *Ann. Inst. Statis. Math.*, 22: 202–217, 1970.

[2] Dacunha-Castelle D, Gassiat E. Testing in locally conic models. *ESAIM Prob. and stat.,http: //www.emath.fr/ps/*, volume 1, 285–317, 1997.

[3] Dacunha-Castelle D, Gassiat E. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.*, (to appear) 1999.

[4] Dacunha-Castelle D, Gassiat E. The estimation of the order of a mixture model. *Bernoulli*, 3: 279–299, 1997b.

[5] Dempster A.P., Laird N.M., Rubin D.B. Large Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39: 1–38, 1977.

[6] Dudley, R.M., Philipp, W. Variance Principles for Sums of Banach Space Valued Random Elements of Empirical Processes. *Zeitschrift fur Warscheinlischkeitstheory verw. Gebiete*, 62: 509–552, 1983

[7] Henna, J. On estimating the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.*, 37: 235–240, 1985.

[8] Haughton, D. On the choice of a model to fit data from an exponential family. *Ann. Statist.*, 16: 342–355, 1988.

[9] Haughton, D. Size of the error in the choice of a model to fit data from an exponential family. *Sankhyā A*, 51: 45–58, 1989.

[10] Izenman, A.J., Sommer, C. Philatelic mixtures and multivariate densities. *Journal of the American Math. Soc.*, 83–94, 1988.

[11] Leroux, B.G. Consistent estimation of a mixing distribution. *Ann. Statist.*, 20: 1350–1360, 1992.

[12] Leroux, B.G. Maximum Penalized Likelihood Estimation for Independent and Markov-Dependent Mixture Models. *Biometrics*, 48: 545–558, 1992.

[13] Lindsay, B.G. Moment matrices: application in mixtures. *Ann. Statist.*,17(2): 722–740, 1983.

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

[15] Redner R A, Walker H F. Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review* 26(2): 195–223, 1984.

[16] Roeder, K. A graphical technique for determining the number of components in a mixture of normals. *Journal of the American Statis. Assoc.*, 89(426): 487–495, 1994.

[17] Schwarz, G. Estimating the dimension of a model. *Ann. Statist.*, 6: 461-464, 1978.

[18] Senoussi, D. Statistique asymptotique presque sûre de modèles statistiques convexes. *Ann. Inst. Henri Poincaré*, 26: 342-355, 1988.

[19] Teicher, H. Identifiability of finite mixtures. *Ann. Math. Statist.*, 36: 423–439, 1965.

[20] Van der Vaart, A. W., Wellner, J.A. Empirical Processes. *Springer Verlag*, 1996.

[21] Yakowitz, S.J., Spragins, J.D. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39: 209–214, 1968.

Chapitre 4

Test de contamination

Dans ce travail, réalisé en collaboration avec D. Haughton, nous allons étudier la vitesse de convergence de la probabilité de se tromper d'ordre, mais l'étude générale du cas p contre q s'est révélée plus difficile que prévue. Prenons le cas i.i.d. de loi $g_0 = f_{\gamma_0}$ contre un mélange de deux populations quelconques pour comprendre ce qui se passe. Dans ce cas, la re-paramétrisation du modèle donne :

$$g_{\theta,\delta,\gamma} = (1 - \frac{\theta}{N})f_{\gamma_0 + \frac{\delta\theta}{N}} + \frac{\theta}{N}f_\gamma$$

avec

$$N^2 = E_{g_0\nu}\Big((\delta \frac{\partial_\gamma f_{\gamma_0+\frac{\delta\theta}{N}}}{g_0} + \frac{f_\gamma - f_{\gamma_0}}{f_{\gamma_0}})^2\Big)$$

où θ est le paramètre identifiable. La consistance du paramètre localement conique nous permet un développement de Taylor de la vraisemblance dans le voisinage de $\theta = 0$. Ainsi, notant maintenant la vraisemblance $l_n(\theta, \delta, \gamma)$ avec les nouveaux paramètres, et $l_n(0, \delta, \gamma) = l_n(0)$:

$$l_n(\hat{\theta}, \delta, \gamma) - l_n(0) = -\frac{1}{2}\frac{(\nabla_{\theta=0} l_n(\theta, \delta, \gamma))^2}{\nabla^2_{\theta=0} l_n(\theta, \delta, \gamma)} 1_{\{\nabla_{\theta=0} l_n(\theta,\delta,\gamma) \geq 0\}}(1 + R_n)$$

Or, la maximisation de $l_n(\hat{\theta}, \delta, \gamma) - l_n(0)$ sur δ et γ peut amener à des valeurs nulles du dénominateur. De plus, le terme R_n de reste doit être un $o(1)$ uniformément en probabilité. Dacunha-Castelle et Gassiat (1997) résolvent le point en partageant l'espace en deux parties. Sur l'une d'entre elle, le développement de la vraisemblance doit être poursuivi jusqu'à l'ordre 5. Dans notre cas, le problème du dénominateur subsiste, car nous traitons le numérateur et le dénominateur de façon séparée. Afin de supprimer le problème du dénominateur, nous nous plaçons sous l'hypothèse mélange simple ou

modèle de contamination.
Soit $\mathcal{G} = \{g_{\pi,\gamma} = (1-\pi)f_{\gamma^0} + \pi f_\gamma \quad , \quad f_{\gamma^0} \in \mathcal{F}, \quad , \quad f_\gamma \in \mathcal{F}, \quad , 0 \leq \pi \leq 1 \quad \}$, nous définissons le test :

$$(H0 : \; f_{\gamma^0}) \quad \text{contre} \quad (H1 : \; g \in \mathcal{G} \setminus \{f_{\gamma^0}\})$$

de domaine de rejet

$$T_n - l_n(g_0) \geq v(n)$$

et de niveau associé

$$\alpha_n := P_0\left(T_n - l_n(g_0) \geq v(n)\right)$$

où $T_n = \sup_{g \in \mathcal{G}} l_n(g)$. Dans ce cadre, notant $h_\gamma = \frac{1}{N}\frac{f_\gamma - f_{\gamma_0}}{f_{\gamma_0}}$ et $N^2 = E_{g_0\nu}(h_\gamma^2)$, $\nabla^2_{\theta=0} l_n(\theta, \gamma) = -\sum_{i=1}^n h_\gamma^2(X_i)$ qui ne peut s'annuler grâce à la condition de normalisation. En suivant le même cheminement que pour les familles identifiables, nous parvenons donc a une vitesse asymptotique de l'ordre de $O(e^{-v(n)(1-\varepsilon)})$ pour $0 < \varepsilon < 1$. Remarquons que le terme principal fait apparaître maintenant, non pas une déviation sur $\mathbb{R}^n$, mais dans l'espace de Banach des fonctions continues, à cause du supremum sur γ. Le théorème clé utilisé est le Théorème de Ledoux (1992) suivant :

Théorème 4.0.1 *Soit X une variable aléatoire à valeurs dans un espace de Banach réel séparable B. Supposons que $\mathbb{E}(f(X)) = 0$ et $\mathbb{E}(f^2(X)) < \infty$ pour tout f de B^*, le dual de B. Notons $(H, |\cdot|)$ l'espace de Hilbert autoreproduisant associé à la covariance de X. Désignons en outre par I la fonctionnelle sur B définie par $I(x) = |x|^2/2$ ou $+\infty$ selon que x appartient ou non à H. Notons encore $||.||^*$ la norme sur B^*.*

Supposons que la famille $\{f^2(X) : f \in B^*, ||f||^* \leq 1\}$ est uniformément intégrable. Soit $(b_n)_{n\geq 1}$ une suite croissante de réels positifs vérifiant $b_n/\sqrt{n} \to \infty$ et $b_n/n \to 0$ et telle qu'il existe $0 < \delta < 1$ et $A \geq 1$ tels que pour tous les entiers $n, k \geq 1$,

$$b_{nk} \leq A k^{1-\delta} b_n \tag{1}$$

Alors, pour que

$$\limsup_{n\to\infty} \frac{n}{b_n^2} \log \mathbb{P}\{\frac{1}{b_n}\sum_{i=1}^n X_i \in F\} \leq -\inf_{x \in F} I(x)$$

pour tout fermé F de B, il faut et il suffit que les deux conditions suivantes soient réalisées :

- *(i) $\frac{1}{b_n}\sum_{i=1}^n X_i \to 0$ en probabilité;*

- *(ii) il existe $M > 0$ tel que, quel que soit $u > 0$,*

$$\limsup_{n\to\infty} \frac{1}{b_n^2} \log(n\mathbb{P}\{||X|| > ub_n\}) \leq -\frac{u^2}{M}$$

Notons que :

- Si X vérifie le théorème central limite, (i) est automatiquement satisfaite.
- En présence de moments exponentiels de tous ordres, la propriété (ii) est aussi satisfaite, car $b_n/n \to 0$.
- La condition (1) empêche la suite de se rapprocher trop vite de $b_n = n$, mais autorise cependant les suites usuelles.
- Si F est le fermé de B, $F = \{x \in B; ||x|| \geq u\}$, $u > 0$, on a :

$$\inf_{x\in F} I(x) = \frac{u^2}{2\sigma^2}$$

  où $\sigma = \sup_{||f||^* \leq 1} \left(\mathbb{E}(f^2(X))^{1/2}\right) < \infty$

Les hypothèses prises dans cet article sont des moments exponentiels pour $(f_\gamma - f_{\gamma_0})/f_{\gamma_0}$, et ne sont donc pas vérifiées dans le cas de mélanges de Gaussiennes. Elles restreignent donc les applications à des lois concentrées, par exemple des lois à support compact.

On the rate of convergence of a test of contamination using a maximum likelihood technique

Christine Keribin
Université d'Evry-Val d'Essonne, France [1]

Dominique Haughton
Bentley College and M.I.T., USA [2]

Abstract: Large and moderate deviations are applied to find the asymptotical behavior of the level of a contamination test, using maximum likelihood techniques.

4.1 Introduction

We consider the following simple mixture model:

$$\mathcal{G} = \{g_{\pi,\gamma} = (1-\pi)f_{\gamma^0} + \pi f_\gamma \ ,\ f_{\gamma^0} \in \mathcal{F},\ \ f_\gamma \in \mathcal{F},\ 0 \le \pi \le 1\} \tag{2}$$

where $\mathcal{F} = (f_\gamma)_{\gamma\in\Gamma}$ is a family of probability densities with respect to a measure ν, Γ is a compact subset of $\mathbb{R}$. The choice of $\mathbb{R}$ is driven by the fact, seen later, that $\frac{f_\gamma - g_0}{g_0}$ must be continuous on a finite number of intervals, to apply the Theorem of moderate deviations of Ledoux (1992). $\mathcal{G}$ is the set of the mixtures of a given density $f_{\gamma^0} = g_0$ with any other density of $\mathcal{F}$. We suppose that we observe a n-sample $(X_i)_{i=1,n}$ of true density g_0, where γ^0 lies in the interior of Γ. We want to test the contamination of the n-sample, that is

$$H_0 : g_0 \text{ against } H1 : g \in \mathcal{G} \setminus \{g_0\}$$

with rejecting domain:

$$T_n - l_n(g_0) \ge v(n)$$

[1](1) Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Boulevard des Coquibus, 91025 Evry

[2](2) Department of Mathematical Sciences, Bentley College, 175 Forest Street, Waltham, MA 02452-4705, USA

[3]*Key words and phrases.* penalized maximum likelihood, large deviations, moderate deviations, information criteria, model selection

and associated level:

$$\alpha_n := P_0\left(T_n - l_n(g_0) \geq v(n)\right)$$

where, for any $g \in \mathcal{G}$, $l_n(g) = \sum_{i=1}^{n} \log g(X_i)$ is the log-likelihood of the observations $(X_i)_{i=1,n}$, $T_n = \sup_{g\in\mathcal{G}} l_n(g)$ is the maximum log-likelihood statistic in the model $\mathcal{G}$ and $v(n) = o(n)$ with $\lim_{n\to+\infty} v(n) = +\infty$.

The idea of this kind of rejecting domain comes from the study of the maximum penalized likelihood estimator of the order (roughly, the number of parameters the model depends on). In case of identifiable models, the asymptotic behavior of the level of a test of the order has already been studied with such kind of rejecting domain (Keribin and Haughton (1999)). In this case, the basic principles that have been used are:

- a Taylor expansion of the likelihood with respect to the parameters, using the consistency of the maximum likelihood estimator of the parameters in a model of a given order,
- the application of theorems of moderate deviations in $\mathbb{R}$ and of large deviations in Banach spaces.

The case of a mixture is a little more complicated. Since g_0 has different representations $g_{\pi,\gamma}$ in $\mathcal{G}$ with different parameters

$$\{g_{\pi,\gamma} = g_0\} \Leftrightarrow \{\pi = 0 \text{ or } \gamma = \gamma^0\}$$

the maximum likelihood estimator of the global parameters $\hat{\pi}^n$ and $\hat{\gamma}^n$ does not converge at all and the Taylor expansion cannot be done with respect to these parameters. However, this does not imply that the maximum likelihood statistic does not converge. In fact, D. Dacunha-Castelle and E. Gassiat (1997) defined a new parameterization (θ, γ), called locally conic, for which the maximum likelihood estimator of θ was consistent, proved the consistency of the maximum likelihood statistic and found the limit distribution.

Hence, the study of the limiting behavior of the level can be driven in the same way as in the identifiable case. But now, the Taylor expansion is made with respect to θ in a neighbourhood of 0. Moreover, we have to take a supremum of the likelihood over all the parameters γ. Hence, we apply a recent development on moderate deviations in Banach spaces (Ledoux (1992)), which gives us an upper bound of the rate, and take strong assumptions of the existence of exponential moments for $(f_\gamma - f_{\gamma^0})/f_{\gamma^0}$. These assumptions on the sets of functions are not fullfilled in the Gaussian mixture case and limit the applications to very concentrated densities such as densities with a compact support.

A first section details the notations and assumptions. The second section is devoted to the study of the asymptotical behavior . The proofs are detailed in the last section.

4.2 Assumptions

In this section, we first recall some assumptions and known results on the locally conic parameterization. Then, we state the assumptions that are necessary for the use of the deviations theorems.

4.2.1 Locally conic parameterization

The main principle of the locally conic parameterization is to define a new parameterization of the model: a first positive and real parameter θ contains some "distance" to the true model, and a second parameter β is some direction to approach the true model. Hence, in the case of a simple mixture, the locally conic parameters are defined as follows:

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

so that, the new parameterization is given by

$$g_{\pi,\theta} = g_0 \cdot \Big(1 + \theta \cdot \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \cdot \| \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \|_H^{-1} \Big) = g_0 \cdot (1 + \theta h_\gamma)$$

where $\| \cdot \|_H$ represents the norm in the Hilbert space $H = L^2(g_0\nu)$.
Notice that $\| \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \|_H$ is the χ^2 distance beween f_γ and f_{γ^0}. As $\theta = \pi \| \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \|_H$, the parameter θ is bounded and belongs to $\mathcal{T} = [0, M]$.

The following assumptions **(M1)** to **(M3)** lead to the consistency of the maximum likelihood estimator of the parameter θ (Dacunha-Castelle and Gassiat (1997)):

- **(M0)** $\mathcal{G}$ is identifiable in the weak sense (see Teicher (1965) or Yakowitz and Spragings (1968))

- **(M1)** $\gamma \to f_\gamma$ is a continuous fonction on Γ. Moreover, there exists a function h in $L_1(g_0\nu)$, such that $\forall f \in \mathcal{F}, \mid \log f \mid \leq h$ ν-a.e.

- **(M2)** There exists two random variables, independent of γ, l_1 and l_2 such that, $\forall \gamma \in \Gamma$:

$$\begin{aligned} \mid \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \mid \leq l_1, \quad E_{g_0\nu}(l_1^2) < +\infty \\ \frac{f_{\gamma^0}}{f_\gamma} \leq l_2, \quad E_{g_0\nu}(l_2^2) < +\infty \end{aligned}$$

- **(M3)** $\mathcal{D} = \left\{ h_\gamma = \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \cdot \parallel \frac{f_\gamma - f_{\gamma^0}}{f_{\gamma^0}} \parallel_H^{-1}, \gamma \in \Gamma \right\}$ is a Donsker class and ξ_d, the gaussian process on $\mathcal{D}$ with covariance the usual scalar product in $L_2(g_0\nu)$, has continuous path.

4.2.2 Assumptions for the use of deviations theorems

- **(M4)** Let $l_1(x,\gamma) = h_\gamma^2(x)$, $l_2(x,\gamma) = \frac{\partial}{\partial\gamma} h_\gamma^2(x)$. For $j = 1, 2$, the following sets are neigbourhoods of 0:

$$S_j = \left\{ t; \int \exp\left(t \cdot \sup_{\gamma \in \Gamma} |l_j(x,\gamma)| \right) g_0(x)\nu(x) \right\} < +\infty$$

- **(M5)** Let $l_3(x,\gamma,\theta) = \left(\frac{h_\gamma(x)}{1+\theta h_\gamma(x)} \right)^2$, $l_4(x,\gamma,\theta) = \frac{h_\gamma^3(x)}{(1+\theta h_\gamma(x))^2}$, $l_5(x,\gamma,\theta) = \frac{h_\gamma^4(x)}{(1+\theta h_\gamma(x))^2}$. There exists $\mathcal{Z}$, a neighbourhood of 0, such that, for all $j = 3, 4, 5$, the following sets are neigbourhoods of 0:

$$\left\{ t; \int \exp\left(t \cdot \sup_{\gamma \in \Gamma, \theta \in \mathcal{Z}} |l_j(x,\gamma,\theta)| \right) g_0(x)\nu(x) < +\infty \right\}$$

- **(M6)** $\int \exp\left(t \cdot h_\gamma(x)\right) g_0(x)\nu(x) < +\infty$ for all $t \in \mathbb{R}$

- **(M7)** For any $N > +\infty$, $\limsup_{r \to 0} E_{g_0\nu}(\exp\left(N\omega(r)\right))$ is finite, where

$$\omega(r) = \sup_{0 \leq \theta_1, \theta_2 \leq M, \gamma_1 \in \Gamma, \gamma_2 \in \Gamma, (\theta_1 - \theta_2)^2 + (\gamma_1 - \gamma_2)^2 \leq r^2} |\log g_{\theta_1,\gamma_1} - \log g_{\theta_2,\gamma_2}|$$

is the modulus of continuity of $\log g_{\theta,\gamma}$.

4.3 Asymptotic behavior of the level.

Theorem 4.3.1 *Let $v(n)$ be a sequence such that $\lim_{n\to+\infty} v(n) = +\infty$, $v(n) = o(n)$ and $b_n = \sqrt{nv(n)}$ satisfies the following property: there exists $0 < \delta < 1$ and $A \geq 1$ such that for all integers $n, k \geq 1$,*

$$b_{nk} \leq Ak^{1-\delta} b_n.$$

Consider the test of

$$(H_0 : g_0) \quad \textit{against} \quad (H1 : g \in \mathcal{G} \setminus \{g_0\}),$$

with rejecting domain

$$\{T_n - l_n(g_0) \geq v(n)\}$$

and associated level:

$$\alpha_n := P_0\left(T_n - l_n(g_0) \geq v(n)\right)$$

Suppose assumptions **(M1)** *to* **(M7)** *hold. The level decreases asymptotically at the following exponential rate:*

$$\limsup_{n\to+\infty} \frac{1}{v(n)} \log \alpha_n \leq -1 + \varepsilon \text{ , \textit{for all} } 0 < \varepsilon < 1$$

Notice that the assumption on the behavior of b_n avoids the series to be too near from $b_n = n$, but allows us to use the usual series.

Proof of theorem 4.3.1

The log-likelihood can be expressed in terms of θ and γ with the locally conic parameterization: $l_n(g) = l_n(\theta, \gamma)$. We will note $\partial l_n(\theta, \gamma)$ the first derivative with respect to θ, $\partial^2 l_n(\theta, \gamma)$ the second derivative with respect to θ and $\widehat{\theta}_\gamma$ the maximum likelihood estimator when γ is fixed. We know that $\widehat{\theta}_\gamma$ and $\sup_\gamma \widehat{\theta}_\gamma$ are consistent.The level can be written as:

$$\alpha_n = P\left(\sup_{\theta,\gamma} l_n(\theta,\gamma) - l_n(g_0) \geq v(n)\right)$$

where P denotes the probability under g_0.With γ fixed, we apply a Taylor expansion with respect to θ to the gradient of the log-likelihood. This leads to the existence of $\tilde{\theta}_\gamma$ such that $0 \leq \tilde{\theta}_\gamma \leq \widehat{\theta}_\gamma$ and:

$$0 = \partial l_n(\widehat{\theta}_\gamma, \gamma) = \partial l_n(0, \gamma) + \widehat{\theta}_\gamma \cdot \partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)$$

that gives, when $\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)$ is not null:

$$\widehat{\theta}_\gamma = -\frac{\partial l_n(0,\gamma)}{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)} \tag{3}$$

In the same way, there exists $\ddot{\theta}_\gamma$, $0 \leq \ddot{\theta}_\gamma \leq \widehat{\theta}_\gamma$, such that:

$$l_n(\widehat{\theta}_\gamma, \gamma) - l_n(g_0) = \widehat{\theta}_\gamma \cdot \partial l_n(0,\gamma) + \frac{\widehat{\theta}_\gamma^2}{2} \cdot \partial^2 l_n(\ddot{\theta}_\gamma, \gamma) \tag{4}$$

Combining these two equations leads to, with $E_{g_o\nu}\left(\partial^2 l_n(0,\gamma)\right) = E_{g_o\nu}(h_\gamma^2) = 1$:

$$\begin{aligned}
l_n(\widehat{\theta}_\gamma, \gamma) - l_n(g_0) &= -\frac{\left(\partial l_n(0,\gamma)\right)^2}{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)} + \frac{1}{2}\left(\frac{\partial l_n(0,\gamma)}{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}\right)^2 \cdot \partial^2 l_n(\ddot{\theta}_\gamma, \gamma) \\
&= -\frac{1}{2}\frac{\left(\partial l_n(0,\gamma)\right)^2}{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}\left(1 + \frac{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma) - \partial^2 l_n(\ddot{\theta}_\gamma, \gamma)}{\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}\right) \\
&= \frac{1}{2}\left(\frac{\frac{1}{n}\left(\partial l_n(0,\gamma)\right)^2}{-\frac{1}{n}\partial^2 l_n(0,\gamma)}\right) \cdot (1 + M_n) \\
1 + M_n &= \left(1 + \frac{\frac{1}{n}\partial^2 l_n(0,\gamma) - \frac{1}{n}\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}{\frac{1}{n}\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}\right) \cdot \\
&\quad \left(1 + \frac{\frac{1}{n}\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma) - \frac{1}{n}\partial^2 l_n(\ddot{\theta}_\gamma, \gamma)}{\frac{1}{n}\partial^2 l_n(\widetilde{\theta}_\gamma, \gamma)}\right)
\end{aligned}$$

We replace now the derivatives by their explicit formula: $\partial l_n(\theta,\gamma) = \sum_{i=1}^n \frac{h_\gamma(X_i)}{1+\theta h_\gamma(X_i)}$, $\partial^2 l_n(\theta,\gamma) = -\sum_{i=1}^n \left(\frac{h_\gamma(X_i)}{1+\theta h_\gamma(X_i)}\right)^2$. For all $0 < \varepsilon_1 < 1$, $0 < \varepsilon_2$, $0 < \varepsilon_3$, the level can be bounded as follows:

$$\begin{aligned}
\alpha_n \leq\ & P\left(\sup_{\gamma\in\Gamma} \frac{1}{n}\left(\sum_{i=1}^n h_\gamma(X_i)\right)^2 \geq 2v(n)\frac{1-\varepsilon_1}{(1+\varepsilon_2)(1+\varepsilon_3)}\right) \\
+\ & P\left(\sup_{\gamma\in\Gamma} \frac{1}{\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i)} \geq \frac{1}{1-\varepsilon_1}\right) \\
+\ & P\left(\sup_{\gamma\in\Gamma}\left|\frac{\frac{1}{n}\sum_{i=1}^n \left(h_\gamma^2(X_i) - \frac{h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)}{\frac{1}{n}\sum_{i=1}^n \left(\frac{h_\gamma(X_i)}{1+\tilde{\theta}_\gamma h_\gamma(X_i)}\right)^2}\right| \geq \varepsilon_2\right)
\end{aligned}$$

$$+ \quad P\left(\sup_{\gamma\in\Gamma}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_\gamma^2(X_i)}{\left(1+\ddot{\theta}_\gamma h_\gamma(X_i)\right)^2}-\frac{h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_\gamma(X_i)}{1+\tilde{\theta}_\gamma h_\gamma(X_i)}\right)^2}\right|\geq\varepsilon_3\right)$$

The proof is an immediate consequence of the following lemmas:

Lemma 4.3.2 *If **(M4)** and **(M6)** hold, for $v(n)$ such that $\lim_{n\to+\infty} v(n) = +\infty$ and $v(n) = o(n)$, we have:*

$$\lim\sup_{n\to+\infty}\frac{1}{v(n)}P\left(\sup_{\gamma\in\Gamma}\frac{1}{n}\left(\sum_{i=1}^{n}h_\gamma(X_i)\right)^2\geq 2v(n)\frac{1-\varepsilon_1}{(1+\varepsilon_2)(1+\varepsilon_3)}\right)\leq-\frac{1-\varepsilon_1}{(1+\varepsilon_2)(1+\varepsilon_3)}$$

Lemma 4.3.3 *If **(M4)** holds, there exists $\lambda_1 > 0$ such that:*

$$\lim\sup_{n\to+\infty}\frac{1}{n}P\left(\sup_{\gamma\in\Gamma}\frac{1}{\frac{1}{n}\sum_{i=1}^{n}h_\gamma^2(X_i)}\geq\frac{1}{1-\varepsilon_1}\right)\leq-\lambda_1$$

Lemma 4.3.4 *If **(M4)** and **(M5)** hold, there exists $\lambda_2 > 0$ such that:*

$$\lim\sup_{n\to+\infty}\frac{1}{n}P\left(\sup_{\gamma\in\Gamma}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\left(h_\gamma^2(X_i)-\frac{h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_\gamma(X_i)}{1+\tilde{\theta}_\gamma h_\gamma(X_i)}\right)^2}\right|\geq\varepsilon_3\right)\leq-\lambda_2$$

Lemma 4.3.5 *If **(M4)** and **(M5)** hold, there exists $\lambda_3 > 0$ such that:*

$$\lim\sup_{n\to+\infty}\frac{1}{n}P\left(\sup_{\gamma\in\Gamma}\left|\frac{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_\gamma^2(X_i)}{\left(1+\ddot{\theta}_\gamma h_\gamma(X_i)\right)^2}-\frac{h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)}{\frac{1}{n}\sum_{i=1}^{n}\left(\frac{h_\gamma(X_i)}{1+\tilde{\theta}_\gamma h_\gamma(X_i)}\right)^2}\right|\geq\varepsilon_2\right)\leq-\lambda_3$$

Lemma 4.3.5 is an immediate consequence of lemma 4.3.4

Lemma 4.3.6 *If **(M7)** holds, there exists $\lambda_4 > 0$ such that:*

$$\lim\sup_{n\to+\infty}\frac{1}{n}P\left(\sup_{\gamma\in\Gamma}\hat{\theta}_\gamma\in(\mathcal{T}\backslash\mathcal{Z})\times\Gamma\right)\leq-\lambda_4$$

4.4 Proofs of the lemma

Proof of Lemma 4.3.2

According to the rate of $v(n)$, this result is a direct application of the recent development of Ledoux (1992) on moderate deviations on a Banach space. However, this theorem deals with continuous functions, and h_γ is not continuous in $\gamma = \gamma_0$. So, we divide Γ in two domains (denoted Γ_j, $j = 1, 2$): $\gamma \leq \gamma_0$, $\gamma \geq \gamma_0$ where h_γ is continuous. Let $b_n = \sqrt{2nv(n)\frac{1-\varepsilon_1}{(1+\varepsilon_2)(1+\varepsilon_3)}}$:

$$P\left(\sup_{\gamma\in\Gamma}\left|\sum_{i=1}^n h_\gamma(X_i)\right| \geq b_n\right) \leq \sum_{j=1}^2 P\left(\sup_{\gamma\in\Gamma_j}\left|\sum_{i=1}^n h_\gamma(X_i)\right| \geq b_n\right)$$

Now, we consider the Banach space $\mathcal{C}_{\Gamma_j}$, set of the continuous functions of γ from Γ_j to $\mathbb{R}$ and check the required assumptions:

- $E_{g_0\nu}\left(h_\gamma(X)\right) = 0$ is clear
- $E_{g_0\nu}\left(\sup_{\gamma\in\Gamma_j} h_\gamma^2(X)\right) < +\infty$ is true under **(M4)**
- The existence of exponential moments of all orders for $h_\gamma(X)$ is true under **(M6)**
- We have $\frac{b_n}{\sqrt{n}} \to +\infty$ and $\frac{b_n}{n} \to 0$ according to the behavior of $v(n)$.

Now, $P\left(\sup_{\gamma\in\Gamma_j}\left|\sum_{i=1}^n h_\gamma(X_i)\right| \geq b_n\right) = P\left(\frac{1}{b_n}\sum_{i=1}^n h_\gamma(X_i) \in A_1\right)$,
with $A_1 = \left\{l(.) \in \mathcal{C}_{\Gamma_j}, \sup_{\gamma\in C_j}|l(\gamma)| \geq 1\right\}$. For the closed subset A_1 of $\mathcal{C}_{\Gamma_j}$, using Ledoux,

$$\limsup_{n\to+\infty}\frac{n}{b_n^2}\log P\left(\frac{\sum_{i=1}^n h_\gamma(X_i)}{b_n} \in A_1\right) \leq -\frac{1}{2\sigma^2}$$

where $\sigma^2 = \sup_{m\in\mathcal{M}_{\Gamma_j}\int dm(\gamma)\leq 1} E\left(\int h_{\gamma_1}dm(\gamma_1)\int h_{\gamma_2}dm(\gamma_2)\right)$, $\mathcal{M}_{\Gamma_j}$ being the dual space of $\mathcal{C}_{\Gamma_j}$. Now, using a Hölder inequality and the fact that $E(h_\gamma^2) = 1$, we have $E\left(\int h_{\gamma_1}dm(\gamma_1)\int h_{\gamma_2}dm(\gamma_2)\right) \leq \int\int dm(\gamma_1)dm(\gamma_2)$, so that $\sigma^2 = 1$. This achieves the proof of the lemma.

Proof of lemma 4.3.3

The upper bound will be given by applying a theorem of large deviations in the Banach space $\mathcal{C}_\Gamma$ of the continuous functions from Γ to $\mathbb{R}$. The norm on $\mathcal{C}_\Gamma$ is denoted

$\|\cdot\|_\infty$ such that, for $s_. \in \mathcal{C}_\Gamma$, $\|s_.\|_\infty = \sup_{\gamma\in\Gamma}|s_\gamma|$. Let $\mathcal{A}_2$ be the closed subset of $\mathcal{C}_\Gamma$ defined as $\mathcal{A}_2 = \left\{s_. \in \mathcal{C}_\Gamma,\ \left\|\frac{1}{s_.}\right\|_\infty \geq \frac{1}{1-\varepsilon_1}\right\}$. We can write:

$$P\left(\sup_{\gamma\in\Gamma}\frac{1}{\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i)} \geq \frac{1}{1-\varepsilon_1}\right) = P\left(\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i) \in \mathcal{A}_2\right)$$

We denote $\mathcal{C}_\Gamma^*$ the dual space of $\mathcal{C}_\Gamma$, and $<,>$ the duality bracket: for $s_. \in \mathcal{C}_\Gamma$, for $m \in \mathcal{C}_\Gamma^*$, $< m, s_. >= \int_\Gamma s_\gamma \cdot m(d\gamma)$. The logarithmic moment generating function is:

$$\Lambda_1 : \mathcal{C}_\Gamma^* \to (-\infty, +\infty]\ ,\ \Lambda_1(m) = \log E_{g_0 v}\exp(< m, h_\gamma^2 >)$$

and the Fenchel-Legendre transform of $\Lambda_1^*(.)$ is defined as:

$$\Lambda_1^*(s_.) = \sup_{m\in\mathcal{C}_\Gamma^*}\{< m, s_. > -\Lambda_1(m)\}\ ,\ s_. \in \mathcal{C}_\Gamma.$$

Now, if $\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i)$ is exponentially tight, the remark c) of Theorem 4.5.3 of Dembo and Zeitouni (1993) allows us to have the following upper bound:

$$\limsup_{n\to+\infty}\frac{1}{n}\log P\left(\sup_{\gamma\in\Gamma}\frac{1}{\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i)} \geq \frac{1}{1-\varepsilon_1}\right) \leq -\inf_{s_.\in\mathcal{A}_2}\Lambda_1^*(s_.) = -\lambda_1$$

(M3) imply that $(\overline{\mathcal{D}})^2$ is a Glivenko-Cantelli class in probability and that $\overline{\mathcal{D}}$ is a compact subset of the unit sphere . Hence, $\frac{1}{n}\left\|\sum_{i=1}^n h_\gamma^2(X_i)\right\|_\infty$ converges in probability to 1, for $\tau > 0$, so:

$$\limsup_{n\to+\infty}\frac{1}{n}\log P\left(\left\|\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i) - 1\right\|_\infty \leq \tau\right) = 0$$

Write now

$$\limsup_{n\to+\infty}\frac{1}{n}\log P\left(\left\|\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i) - 1\right\|_\infty \leq \tau\right) \leq -\inf_{\left\{s_.\in\mathcal{C}_\Gamma, \|s_.-1\|_\infty\leq\tau\right\}}\Lambda_1^*(s_.)$$

so $\Lambda_1^*(1) = 0$. Now, $\mathcal{A}_2$ is a closed subset that does not contain a neighbourhood of 1 and Λ_1^* is convex; so $\lambda_1 > 0$ and the lemma is proved. It remains to be seen that $\frac{1}{n}\sum_{i=1}^n h_\gamma^2(X_i)$ is exponentially tight.

Let $\mathcal{K}_1(K_1, K_1^{'})$ be the following precompact of $\mathcal{C}_\Gamma$:

$$\left\{s_. \in \mathcal{C}_\Gamma\ ,\ \|s_\gamma\|_\infty \leq E_{g_0 v}\left(\left\|h_\gamma^2(X)\right\|_\infty\right) + K_1\ ,\ \left\|\frac{\partial}{\partial\gamma}s_\gamma\right\|_\infty \leq E_{g_0 v}\left(\left\|\frac{\partial}{\partial\gamma}h_\gamma^2(X)\right\|_\infty\right) + K_1^{'}\right\}$$

Write:

$$
\begin{aligned}
& P\left(\frac{1}{n}\sum_{i=1}^{n} h_\gamma^2(X_i) \in \mathcal{K}_1^c(K_1, K_1')\right) \\
\leq\ & P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} h_\gamma^2(X_i)\right\|_\infty \geq E_{g_0 v}\left(\left\|h_\gamma^2(X)\right\|_\infty\right) + K_1\right) \\
+\ & P\left(\left\|\frac{1}{n}\sum_{i=1}^{n} \frac{\partial}{\partial\gamma} h_\gamma^2(X_i)\right\|_\infty \geq E_{g_0 v}\left(\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty\right) + K_1'\right) \\
\leq\ & P\left(\frac{1}{n}\sum_{i=1}^{n} \left\|h_\gamma^2(X_i)\right\|_\infty \geq E_{g_0 v}\left(\left\|h_\gamma^2(X)\right\|_\infty\right) + K_1\right) \\
+\ & P\left(\frac{1}{n}\sum_{i=1}^{n} \left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X_i)\right\|_\infty \geq E_{g_0 v}\left(\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty\right) + K_1'\right) \\
\leq\ & e^{-n\mathcal{H}_1 E_{g_0 v}\left(\left\|h_\gamma^2(X)\right\|_\infty + K_1\right)} + e^{-n\mathcal{H}_2\left(E_{g_0 v}\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty + K_1'\right)}
\end{aligned}
$$

where, using **(M4)**, Chernof's Theorem on $\mathbb{R}$ has been applied with $\mathcal{H}_1$ (resp. $\mathcal{H}_2$) the Cramer function associated with $\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty$ (resp. $\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty$). Now, for all $\eta > 0$, there exists K_1 and K_1' such that $\mathcal{H}_1\left(E_{g_0 v}\left(\left\|h_\gamma^2(X)\right\|_\infty\right) + K_1\right) > \eta$ and $\mathcal{H}_2\left(E_{g_0 v}\left(\left\|\frac{\partial}{\partial\gamma} h_\gamma^2(X)\right\|_\infty\right) + K_1'\right) > \eta$. Then,

$$
P\left(\frac{1}{n}\sum_{i=1}^{n} h_\gamma^2(X_i) \in \mathcal{K}_1^c(K_1, K_1')\right) \leq 2e^{-n\eta}
$$

Hence,

$$
\limsup_{n\to+\infty} \frac{1}{n} \log P\left(\frac{1}{n}\sum_{i=1}^{n} h_\gamma^2(X_i) \in \mathcal{K}_1^c(K_1, K_1')\right) \leq -\eta
$$

and $\frac{1}{n}\sum_{i=1}^{n} h_\gamma^2(X_i)$ is exponentially tight.

Proof of lemma 4.3.4

Remember that M is the maximal value for θ, and that $\tilde{\theta}_\gamma \leq \widehat{\theta}_\gamma \leq M$. We write:

$$
\begin{aligned}
& \left|\sum_{i=1}^{n}\left(h_\gamma^2(X_i) - \frac{h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)\right| \\
=\ & \left|\sum_{i=1}^{n} h_\gamma^2(X_i)\left(\frac{2\tilde{\theta}_\gamma h_\gamma(X_i) + \tilde{\theta}_\gamma^2 h_\gamma^2(X_i)}{\left(1+\tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}\right)\right|
\end{aligned}
$$

$$\leq \quad \widehat{\theta}_\gamma \cdot \left(\sum_{i=1}^n \frac{2 \cdot \left| h_\gamma^3(X_i) \right|}{\left(1 + \tilde{\theta}_\gamma h_\gamma(X_i)\right)^2} + \sum_{i=1}^n \frac{M \cdot h_\gamma^4(X_i)}{\left(1 + \tilde{\theta}_\gamma h_\gamma(X_i)\right)^2} \right)$$

Define, for $k \in \{2, 3, 4\}$:

$$V_{k-1} = E_{g_0\nu} \left(\sup_{\gamma \in \Gamma, \widetilde{\theta} \in \mathcal{Z}} \frac{|h_\gamma^k(X)|}{\left(1 + \widetilde{\theta} h_\gamma(X)\right)^2} \right)$$

For $0 < \varepsilon_4 < V_1$, $0 < \varepsilon_5$, $0 < \varepsilon_6$, write, for a small enough ε_2 such that

$$\varepsilon_2 \cdot \frac{V_1 - \varepsilon_4}{2 \cdot (V_2 + \varepsilon_5) + M \cdot (V_3 + \varepsilon_6)} \in \mathcal{Z} :$$

$$\begin{aligned}
& P \left(\sup_{\gamma \in \Gamma} \left(\widehat{\theta}_\gamma \cdot \frac{\frac{1}{n} \sum_{i=1}^n \frac{2 \cdot \left| h_\gamma^3(X_i) \right| + M \cdot h_\gamma^4(X_i)}{\left(1 + \tilde{\theta}_\gamma h_\gamma(X_i)\right)^2}}{\frac{1}{n} \sum_{i=1}^n \left(\frac{h_\gamma(X_i)}{1 + \tilde{\theta}_\gamma h_\gamma(X_i)} \right)^2} \right) \geq \varepsilon_2 \right) \\
\leq \quad & P \left(\sup_{\gamma \in \Gamma} \widehat{\theta}_\gamma \geq \varepsilon_2 \frac{V_1 - \varepsilon_4}{2 \cdot (V_2 + \varepsilon_5) + M \cdot (V_3 + \varepsilon_6)} \right) \\
& + P \left(\sup_{\gamma \in \Gamma, \widetilde{\theta} \in \mathcal{Z}} \frac{1}{\frac{1}{n} \sum_{i=1}^n \left(\frac{h_\gamma(X_i)}{1 + \widetilde{\theta} h_\gamma(X_i)} \right)^2} \geq \frac{1}{V_1 - \varepsilon_4} \right) \\
& + P \left(\sum_{i=1}^n \sup_{\gamma \in \Gamma, \widetilde{\theta} \in \mathcal{Z}} \frac{\left| h_\gamma^3(X_i) \right|}{\left(1 + \widetilde{\theta} h_\gamma(X_i)\right)^2} \geq V_2 + \varepsilon_5 \right) \\
& + P \left(\sum_{i=1}^n \sup_{\gamma \in \Gamma, \widetilde{\theta} \in \mathcal{Z}} \frac{h_\gamma^4(X_i)}{\left(1 + \widetilde{\theta} h_\gamma(X_i)\right)^2} \geq V_3 + \varepsilon_6 \right)
\end{aligned}$$

Taking **(M4)** into account, a Chernof's inequality can be used to see that, for a large enough n, the last three terms are of the order $e^{-\lambda n}$, with $\lambda > 0$. The rate of the first term is also a large deviation, according to Lemma 4.3.6, so that Lemma 4.3.4 is proved.

Proof of lemma 4.3.6. We first write:

$$\left\{ \sup_{\gamma \in \Gamma} \hat{\theta}_\gamma \in \mathcal{T} \backslash \mathcal{Z} \right\} \subset \left\{ \sup_{\gamma \in \Gamma} \sup_{\theta \in \mathcal{T} \backslash \mathcal{Z}} l_n(\theta, \gamma) \geq \sup_{\gamma \in \Gamma} \sup_{\theta \in \mathcal{Z}} l_n(\theta, \gamma) \right\}.$$

We now divide Γ in the two domains Γ_j, $j = 1,2$ where h_γ is continuous. On each domain, assumption **(M7)** together with the continuity of the likelihood, allow us to use Mogulskii's Theorem (1988) of large deviations for the maximum likelihood estimators:

$$\begin{aligned}
& \lim\sup_{n\to+\infty} \frac{1}{n} P\left(\sup_{\gamma\in\Gamma_j} \hat{\theta}_\gamma \in \mathcal{T}\backslash\mathcal{Z}\right) \\
\leq\ & P\left(\sup_{(\theta,\gamma)\in\mathcal{T}\backslash\mathcal{Z}\times\Gamma_j} (l_n(\theta,\gamma) - l_n(0,\gamma)) \geq \sup_{(\theta,\gamma)\in\mathcal{Z}\times\Gamma_j} (l_n(\theta,\gamma) - l_n(0,\gamma))\right) \\
\leq\ & -\inf_{(\theta,\gamma)\in\mathcal{T}\backslash\mathcal{Z}\times\Gamma_j} \mathcal{H}(\theta,\gamma) \\
\text{with}\mathcal{H}(\theta,\gamma) =\ & -\log \inf_{u>0, m\in\mathcal{P}} E_{g_0\nu}\left(\exp\left(u\int_{\mathcal{Z}\times\Gamma_j} \log\frac{g_{\theta,\gamma}}{g_{v,w}}(X)\cdot m(dv,dw)\right)\right).
\end{aligned}$$

$\mathcal{P}$ denotes here the class of probability measures m on $\mathcal{Z}\times\Gamma_j$.

We now prove that $-\inf_{(\theta,\gamma)\in\mathcal{T}\backslash\mathcal{Z}\times\Gamma_j} \mathcal{H}(\theta,\gamma) > 0$. As $m(dv,dw)$ is a probability measure, Jensen's inequality leads to:

$$\exp\left(u\int_{\mathcal{Z}\times\Gamma_j} \log\frac{g_{\theta,\gamma}}{g_{v,w}}(X)\cdot m(dv,dw)\right) \leq \int_{\mathcal{Z}\times\Gamma_j} \left(\frac{g_{\theta,\gamma}}{g_{v,w}}(X)\right)^u m(dv,dw).$$

The choice of the particular measure $m(dv,dw) = \delta_0\otimes\delta_\gamma$ and the value $u=\frac{1}{2}$ gives:

$$\inf_{u>0,m\in\mathcal{P}} E_{g_0\nu}\left(\exp\left(u\int_{\mathcal{Z}\times\Gamma_j} \log\frac{g_{\theta,\gamma}}{g_{v,w}}(X)\cdot m(dv,dw)\right)\right) \leq E_{g_0\nu}\left(\left(\frac{g_{\theta,\gamma}}{g_{0,\gamma}}(X)\right)^{\frac{1}{2}}\right).$$

Now, $E_{g_0\nu}\left(\left(\frac{g_{\theta,\gamma}}{g_{0,\gamma}}(X)\right)^{\frac{1}{2}}\right) = 1 - E_{g_0\nu}\left((\sqrt{g_{\theta,\gamma}(X)} - \sqrt{g_0(X)})^2\right) \leq 1$. According to the locally conic parameterization, the right term is 1 if and only if $\theta = 0$.

Hence, for $(\theta,\gamma)\in(\mathcal{T}\backslash\mathcal{Z})\times\Gamma_j$,

$$\inf_{u>0,m\in\mathcal{P}} E_{g_0\nu}\left(\exp\left(u\int_{\mathcal{Z}\times\Gamma_j} \log\frac{g_{\theta,\gamma}}{g_{v,w}}(X)\cdot m(dv,dw)\right)\right) < 1,$$

so that, $\inf_{(\mathcal{T}\backslash\mathcal{Z})\times\Gamma_j} \mathcal{H}(\theta,\gamma) > 0$, ending the proof.

Bibliographie

[1] Dacunha-Castelle D., Gassiat E. Testing in locally conic models. *ESAIM Prob. and stat.,http: //www.emath.fr/ps/*, volume 1, 285-317, 1997.

[2] Dembo A., Zeitouni O. Large Deviations Techniques and Applications. *Jones and Bartlett Publishers, Boston*, 1993.

[3] Keribin, C. Consistent Estimation of the Order of Mixture Models. *Sankhyā A*, submitted, 1997.

[4] Keribin C., Haughton D. Asymptotic probabilities of over-estimating and under-estimating the order of a model in general regular families. *J. Statis. Plan. and Inference*, submitted, 1999.

[5] Ledoux, M. Sur les déviations modérées des sommes de variables vectorielles indépendantes de même loi. *Ann. Inst. Henri Poincaré*, 28(2): 267-280, 1992.

[6] Mogulskii, A. A. Large deviations for maximum likelihood estimators. *Lecture Notes in Mathematics*, 1299: 326-331, 1988.

[7] Teicher, H. Identifiability of finite mixtures. *Ann. Math. Statist.*, 36: 423-439, 1965.

[8] Van der Vaart, A. W., Wellner, J. A. Empirical Processes. *Springer Verlag.*, 1996.

[9] Yakowitz, S. J., Spragins, J. D. On the identifiability of finite mixtures. *Ann. Math. Statist.*, 39: 209-214, 1968.

Chapitre 5

Estimation des paramètres d'un modèle MA bruité

Réalisé en collaboration avec O. Bermond, et soumis à la revue *Statistical Inference for Stochastic Processes*, cet article étudie le processus MA(q) bruité $(y_t)_{t\in\mathbb{N}}$ à ordre q fixé suivant,

$$y_t = \sum_{l=0}^{q} b_0 x_{t-l} + \varepsilon_t = \mathbf{b}^T X_t + \varepsilon_t$$

est une CMC où la chaîne est $(X_t)_{t\in\mathbb{N}}$, avec $X_t = (x_t, \ldots, x_{t-q})$. $(\varepsilon_t)_{t\in\mathbb{N}}$ est un bruit i.i.d. indépendant de $(x_t)_{t\in\mathbb{N}}$, de densité f_ε par rapport à la mesure μ. Quand x_t est à valeurs discrètes, nous sommes dans le cadre d'un modèle classique d'une chaîne finie observée dans du bruit, pour laquelle on sait filtrer, lisser et estimer les paramètres. En effet les coefficients de la matrice de transition sont :

$$a_{i,j} = P(X_t = X^j / X_{t-1} = X^i) = 1_{x_1^j = x_0^i} 1_{x_2^j = x_1^i} \ldots 1_{x_q^j = x_{q-1}^i} P(x_t = x_0^j)$$

et la loi des observations conditionnelle à la chaîne est

$$P(y_t / X_t = X^i) = f_\varepsilon(y_t - \mathbf{b}^T X^i) := f_i(y_t).$$

Ce modèle peut donc être vu sous l'angle des modèles à données incomplètes (la chaîne de Markov), où $(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T})$ représente l'ensemble des T observations complétées des données manquantes. Ainsi, on peut utiliser l'algorithme d'EM pour l'estimation. Si le bruit ε_t est Gaussien par exemple, la log vraisemblance des données complètes est proportionnelle à

$$\log p_n(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T}) \sim -\frac{1}{n}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{T}(y_t - \mathbf{b}^T X_t)^2$$

La phase d'Estimation nécessite le calcul de l'espérance conditionnelle à la loi a posteriori de la log-vraisemblance complète $E(\log p_n/\mathbf{y}_{1:T})$, pour lequel il suffit d'évaluer $E(X_t/\mathbf{y}_{1:T})$ et $E(X_t X_t^T/\mathbf{y}_{1:T})$. On évalue ainsi $P(X_t = i/\mathbf{y}_{1:T})$ en utilisant un algorithme de *Forward - Backward.* Puis on Maximise l'espérance a posteriori pour trouver une estimation des paramètres.

Il n'y a plus de forward backward quand x_t est continue. Si cependant x_t est gaussien, on retrouve le principe du filtre de Kalman. Le problème est bien sûr le calcul de l'espérance conditionnelle sous la loi a posteriori. Une approche est celle des variantes stochastiques de l'EM, où l'on cherche à simuler les données manquantes sous la loi a posteriori, mais c'est lourd numériquement, et la convergence n'est pas toujours assurée. L'approche proposée ici est l'utilisation d'une pseudo vraisemblance, définie de la façon suivante : pour $m > q$, $r \in \{1, \ldots, m\}$ et S la partie entière de $(T-m)/r$:

$$l_T(\theta; r) = \frac{1}{S+1} \sum_{s=0}^{S} \log p_m(y_{sr+1}, \ldots, y_{sr+m}; \theta)$$

qui représente le processus de contraste. La fonction de contraste associée est

$$K(\theta_0, \theta) = E_{\theta_0}\left(\log p_m(\mathbf{y}_{1:m}; \theta_0)\right) - E_{\theta_0}\left(\log p_m(\mathbf{y}_{1:m}; \theta)\right)$$

qui est positive, comme information de Kullback-Leibler.

L'idée d'une telle pseudo-vraisemblance a d'abord été introduite par Ryden (1994) pour une CMC à états finis. Cet auteur ne considère que les cas $m = r$. Il apparaît cependant que le fait de couper les données en paquets qui ne se recouvrent pas n'est pas nécessaire pour obtenir la consistance et la normalité asymptotique, que nous montrons sous couvert d'une hypothèse d'identifiabilité de la marginale d'ordre m, et d'hypothèses classiques d'existence de moments et de régularité de cette marginale d'ordre m. Nous en étudions d'ailleurs des conditions suffisantes dans le cas d'un MA(q) à entrée continue paramétrique et à bruit dont la densité n'est connue qu'à une échelle près. Comme la dérivée normalisée du processus est une somme faisant intervenir un processus au plus $m+q$-dépendant, nous utiliserons le Théorème de Hoeffding et Robbins, ainsi énoncé dans Brockwell et Davis (1990) :

Théorème 5.0.1 *Si (X_t) est un processus stationnaire m-dépendant de variables aléatoires de moyenne nulle et fonction d'autocovariance $\gamma(.)$, et si*
$\Gamma_m = \gamma(0) + 2\sum_{j=1}^{m} \gamma(j) \neq 0$, *alors*

- *(i)* $\lim_{T\to\infty} \frac{1}{T} \cdot var(\sum_{t=1}^{T} X_t) = \Gamma_m$ *et*
- *(ii)* $\frac{1}{\sqrt{T}} \sum_{t=1}^{T} X_t$ *converge vers la loi gaussienne* $\mathcal{N}(0, \Gamma_m)$.

Nous proposons de plus un critère simple de moyenne finie, pour l'identifiabilité de la marginale d'ordre $m > q$, utilisant des propriétés des fonctions caractéristiques qui sont rappelées au moment de la preuve. Ceci permet d'alléger notablement les hypothèses de cumulants d'ordre supérieur à 3 de Giannakis (1989). Ces résultats sont utilisés dans le cas où l'entrée est un mélange de K gaussiennes :

$$g_\gamma(x) = \sum_{i=1}^{K} \pi_i \phi(x, \mu_i, \sigma_i^2)$$

L'intérêt de cet exemple est qu'on peut introduire de nouveaux paramètres cachés (z_t) (appelés paramètres de "nuisance" par rapport aux paramètres dits d'"intérêt" du filtre et de la variance du bruit) qui sont les variables aléatoires discrètes prenant leur valeurs dans $\{1, \ldots, K\}$ telles que $P(z_t = i) = \pi_i$ soit le poids de la composante $\phi(x, \mu_i, \sigma_i^2)$ et les paramètres des composantes. Les données complètes sont donc $(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T}, \mathbf{z}_{1-q:T})$. Utilisant un algorithme de EM, on cherche à maximiser l'espérance a posteriori de la pseudo-vraisemblance complète $Q(\theta|\theta') = E[L_T(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T}, \mathbf{z}_{1-q:T}, \theta, r)|\mathbf{y}_{1:T}; \theta']$, qui, de part la forme de L_T est une somme de $S+1$ termes dont l'un est :

$$\begin{aligned} & E(\log p(\mathbf{y}(s), \mathbf{x}(s), \mathbf{z}(s), \theta)|\mathbf{y}(s); \theta'] \\ = \; & E(\log p(\mathbf{y}_{sr+m:sr+1}, \mathbf{x}_{sr+m:sr+1-q}, \mathbf{z}_{sr+m:sr+1-q}, \theta)|\mathbf{y}(s); \theta'] \\ = \; & Q_i(\mathbf{b}, \sigma^2|\theta') + Q_n(\gamma|\theta') \end{aligned}$$

avec

$$\begin{aligned} Q_i(\eta|\theta') &= E[\log p(\mathbf{y}(s)|\mathbf{x}(s); \eta)|\mathbf{y}(s); \theta'] \\ Q_n(\gamma|\theta') &= E[\log p(\mathbf{x}(s), \mathbf{z}(s); \gamma)|\mathbf{y}(s); \theta'] \end{aligned}$$

Le séparation des variables d'intérêt et des paramètres de nuisance vient de la structure particulière de notre modèle pour lequel nous avons noté $\gamma = (\pi_i, \mu_i, \sigma_i^2)_{1\leq i\leq K}$. En maximisant $Q_i(\eta|\theta')$ on obtient $\hat{\sigma}^2$ et $\hat{\mathbf{b}}$ en fonction des espérances conditionnelles suivantes : $E[\mathbf{x}_{sr+i:sr+i-q}|\mathbf{y}(s); \theta^{(k)}]$ et $E[\mathbf{x}_{sr+i:sr+i-q}\mathbf{x}_{sr+i:sr+i-q}^T|\mathbf{y}(\mathbf{s}); \theta^{(k)}]$. Or, de part la structure du bruit,

$$P(\mathbf{x}_{sr+i:sr+i-q}|\mathbf{y}(s); \theta^{(k)}) = \sum_{\mathbf{z}(s)} P([\mathbf{x}_{sr+i:sr+i-q}|\mathbf{y}(s), \mathbf{z}(s) = \mathbf{z}; \theta^{(k)})P(\mathbf{z}(s) = \mathbf{z}|\mathbf{y}(\mathbf{s}), \theta^k).$$

Les espérances conditionnelles sont donc la moyenne et variance d'un mélange de Gaussiennes, que l'on peut calculer par le Théorème (Kay (1993)) suivant :

Théorème 5.0.2 *Si les données observées* $\mathbf{y}$ *peuvent être modélisées par :*

$$\mathbf{y} = \mathbf{H}\mathbf{x} + \mathbf{w}$$

où $\mathbf{y}$ *est un vecteur* $N \times 1$, $\mathbf{H}$ *est une matrice connue de dimension* $N \times p$, $\mathbf{x}$ *est un vecteur aléatoire de dimension* $p \times 1$ *avec comme loi a priori une Gaussienne* $\mathcal{N}(\mu_x, C_x)$, $\mathbf{w}$ *est un vecteur* $N \times 1$ *de loi normale* $\mathcal{N}(0, C_w)$ *et indépendant de* $\mathbf{x}$, *alors la loi a posteriori* $p(\mathbf{x}|\mathbf{y})$ *est Gaussienne de moyenne*

$$E(\mathbf{x}|\mathbf{y}) = \mu_x + C_x H^T (H C_x H^T + C_w)^{-1} (\mathbf{y} - H\mu_x)$$

et covariance :

$$C_{x|y} = C_x - C_x H^T (H C_x H^T + C_w)^{-1} H C_x$$

Ainsi, un terme Q_i nécessite le calcul de K^{m+q} poids $P(\mathbf{z}(s) = \mathbf{z}|\mathbf{y}(\mathbf{s}), \theta^k)$. Utilisant ces résultats, la maximisation de Q_n permet d'obtenir une réestimation des paramètres de nuisance. Le calcul de $Q(\theta|\theta')$ demande donc le calcul d'un nombre de termes de l'ordre de SK^{m+q}, à comparer avec K^{T+q} termes dans la cas où la vraisemblance n'aurait pas été coupée en paquets. Les simulations faites montrent que la convergence est assez rapide, et que les performances asymptotiques sont peu altérées dès que le ratio signal sur bruit dépasse 20dB. Notons toutefois que, même si on en a diminué la complexité, le nombre de termes à calculer reste exponentiel, ce qui limite, pour des simulations dans des temps raisonnables à un filtre de dimension inférieure à 4 et une source ayant 4 composantes.

M-likelihood estimation for noisy MA processes

Olivier Bermond
ENST, France[1]

Christine Keribin
Université d'Evry-Val d'Essonne, France [2]

Abstract: In this paper, the so-called 'blind' identification of a finite-order MA model observed in noise is addressed. We propose an estimator based on the maximization of an approximate likelihood function. The idea is to split the data in pseudo-independent pack. It is shown that the resulting estimator is consistent and asymptotically Gaussian under some regularity assumptions. Sufficient conditions for these assumption are presented in the particular case of continuous input, and an algorithm is presented. A limited Monte-Carlo experiment is then presented to illustrate our findings.

Introduction

In this contribution, the estimation of a finite-order (possibly non-minimum phase) MA model observed in additive white noise is addressed. This kind of model has been considered by many authors in recent years (see [6, 20, 15] for early contributions, and more recently the works of Nikias and Mendel [17], Haykin [9] and Cadzow [4]). This interest followed the pioneering contribution of Lii and Rosenblatt [14], who proved (in a constructive manner) the identifiability of such models using higher-order spectra. Most of the techniques known to date rely either implicitly or explicitly on higher-order statistics in both time and frequency domain (see [7] and [23] among others). These methods are most often straightforward to implement but are far from being optimal from a statistical standpoint, in particular because higher-order methods (most often based on third- or fourth-order cumulants or spectra) do not exploit the underlying statistical

[1](1) Département Signal, 46, rue Barrault 75634 Paris Cedex 13
[2](2) Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Boulevard des Coquibus, 91025 Evry
[3]*Key words and phrases.* penalized maximum likelihood, large deviations, moderate deviations, information criteria, model selection

distribution of the driving noise (input). Thousands data samples are often required to obtain reasonably accurate estimates. In many applications, an *a priori* information is available on the distribution of the input. Examples of this situation may be found in geophysics or in digital communication problems: in these cases, the distribution of the driving noise is either known or at least, may be modeled accurately (using, *e.g.* some parametrical model). For such situations, important improvements in the performance of the estimates can be expected, by taking into account this information.

To alleviate this drawback, maximum likelihood methods have been considered by several authors, especially for the particular case of discrete input (see [1, 11, 22] and the references therein). The problem is that, in general, the likelihood function cannot be evaluated explicitly. Hence, the maximum likelihood estimator can only be constructed by resorting to some sort of Monte-Carlo integration methods, which are known to be computationally very demanding. The method proposed in this contribution is an alternative to maximum likelihood. The basic idea consists in constructing a minimum of contrast estimator based on a family of contrast function that satisfy the two design criteria: **(i)** it can be evaluated with a significantly lower numerical cost than ML estimation, **(ii)** the contrast function is an approximation of the true likelihood function, so that the minimum of contrast estimator is consistent and asymptotically normal.

More specifically, the observed data, say, y_t can be written as

$$y_t = b_0 x_t + b_1 x_{t-1} + \cdots + b_q x_{t-q} + \epsilon_t \quad t \in \mathbb{Z} \tag{1}$$

Here, the $(x_t)_{t\in\mathbb{Z}}$ are independent identically distributed (i.i.d.) real random variables with common distribution G_x, and the $(\epsilon_t)_{t\in\mathbb{Z}}$ are i.i.d. and independent of $(x_t)_{t\in\mathbb{Z}}$ with common distribution $F_\varepsilon = f_\varepsilon d\mu$ where μ is a dominating measure on $\mathbb{R}$, and f_ε is the density w.r.t to μ. The process $(y_t)_{t\in\mathbb{Z}}$ is then (strict sense) stationary and the finite-dimensional distributions are all dominated by $\mu^{\otimes\mathbb{N}}$. The problem addressed in this contribution is to estimate some parameter θ, including the filter coefficients $\mathbf{b} = (b_0, b_1, \ldots, b_q)$, and possibly some parameters of the distributions F_ε and G_x, on the basis of T observations $y_1, \ldots, y_T$.

Denote m_t the noise-free hidden process

$$m_t = b_0 x_t + b_1 x_{t-1} + \cdots + b_q x_{t-q}$$

m_t is a moving average of order q (MA(q)), and has thus a Markovian representation. y_t is then a hidden Markov model (HMM) where the hidden Markov process is $(x_t, x_{t-1}, \ldots, x_{t-q})$.

Many algorithms were proposed to estimate the parameters of a HMM, when the underlying Markov chain has a finite number of states. This type of models have been studied in a series of papers by Baum and its colleagues (see [18]). Recent results have been obtained, showing that the maximum likelihood estimator (under some regularity conditions) is consistent [13] and asymptotically normal [2]. Not much is known about the estimation of HMM when the underlying process is not finite-dimensional (which is the case in our model for continuous input signal).

In this contribution, we propose a generalized maximum split data likelihood estimate for θ; it is proved to be consistent and asymptotically normal under mild assumptions on the driving and measurement noise distribution. The paper is organized as follows: in section 1, conditions upon which the model is identifiable are stated. In section 2, the maximum split data likelihood estimate is presented; the consistency and the asymptotic normality is established, under some regularity assumptions. In section 3, it is shown that, under practically verifiable conditions on the marginal distribution of the driving noise and of the measurement noise, the maximum split data likelihood estimate is consistent and asymptotically normal.

When compared to classical fourth-order cumulants techniques, it is established that the estimator has a smaller asymptotic variance (at the expense of the knowledge of a parametric model for the distribution of the driving noise). In a last section, a practical implementation of the algorithm is presented, in the particular case where the driving noise distribution can be modeled as a mixture of Gaussian. A Monte-Carlo experiment is performed to illustrate our findings.

5.1 Model identifiability

The identification of linear non-minimum phase (in the noise-free case) systems has been first investigated in the seminal paper by Lii and Rosenblatt [14]; these authors prove the identifiability of the parameters model in a constructive manner, using polyspectra. This property was shown to hold under the assumption that the driving noise has a non-vanishing cumulant of order strictly greater than 2, of course excluding Gaussian distributions. This proof of identifiability may be extended to the noisy case provided that the distribution of the measurement noise (or, more specifically, the cumulants) is known up to scale factor. Unfortunately, such polyspectral methods require the knowledge of all the cumulants of given order (typically 3 or 4). Significant improvements have later been obtained by several authors (see Giannakis, Mendel [7] and the references therein) , showing that the model remains identifiable using only a finite number of cumulants. This property formed the basis of many different estimation methods.

In this section, we prove a general identifiability condition for the "blind" estimation of the noisy MA(q) model. Roughly speaking, the identifiability result states that, provided a maximizer of the order of the model, say q, is known , and that the distribution of the input is not Gaussian, then the filter coefficients $b_0, \ldots, b_q$ can be identified uniquely (up to an inherent scale indeterminacy, see below), and the distribution function of the input and the measurement noise can be recovered as well, up to an irrelevant translation factor. It should be stressed that the result holds even when the distribution of the driving and the measurement noise are not assumed to belong to some known parametric family of distribution function, opening the road to adaptive estimation of the filter coefficients (in situations where these distributions are to be considered as nuisance parameters). Finally, it is worthwhile to note that the identifiability result holds when solely the m-dimensional "marginal" distribution of $(y_1, \cdots, y_m)$ is known. This result thus extends significantly the results by Giannakis and Swami [8].

Notations: throughout this article, we will denote real numbers by lower case letters, vectors by lower case bold-faced letter (*e.g.* u_i is the i-th coordinate of the vector $\mathbf{u}$), and matrices by upper case letters, if A is a matrix, $\mathbf{a}_i$ will denote its i-th column and a_{kl} its entry on the k-th row and l-th column. We shall also use the notation for any couple of integers (n_1, n_2) (so that the vector below may be indexed either in increasing or decreasing order):

$$\mathbf{y}_{n_1:n_2} \triangleq (y_{n_1}, \ldots, y_{n_2})^T$$

where $.^T$ denotes transposition.

Scaling convention: the model (1) considered carries an inherent scale indeterminacy; obviously the filter may be multiplied by a non zero scalar factor, and multiplying the input process by the inverse scalar yields the same observations. In order to fix this indeterminacy, we choose to enforce the first filter coefficient to be 1:

$$b_0 = 1$$

(we could have fixed the variance of x_t as well). Moreover, we will assume that there exist at least one other coefficient than b_0 that is non zero.

Theorem 5.1.1 *Let $m > q$ and denote $P_m(\mathbf{b}, G_x, F_\varepsilon)$ the m-dimensional distribution of $\mathbf{y}_{m:1}$ under the model (1). Assume that G_x is not Gaussian, and in addition that its first moment exist and is finite. Then the identity*

$$P_m(\mathbf{b}, G_x, f_\varepsilon) = P_m(\tilde{\mathbf{b}}, \tilde{G}_x, \tilde{f}_\varepsilon) \tag{2}$$

implies

$$\mathbf{b} = \tilde{\mathbf{b}} \tag{3}$$

and there exists a real number λ such that:

$$G_x(\cdot) = \tilde{G}_x(\cdot - \lambda), \quad f_\varepsilon(\cdot) = \tilde{f}_\varepsilon(\cdot + \lambda(1 + b_1 + \ldots + b_q)) \tag{4}$$

Proof of Theorem 5.1.1. The proof relies upon the results and the techniques used in Linnik, Kagan and Rao [10]. It is based on properties of characteristic functions. We preface the proof by some useful results: we shall need the following Theorem (Theorem 10.3.1 page 311 of [10]) to prove the first part of the Theorem (identifiability of the filter coefficients):

Theorem 5.1.2 *Let $\mathbf{u} = A\mathbf{v}$ and $\mathbf{u} = B\mathbf{w}$ be two representations of an n dimensional random vector where A and B are constant matrices of order $n \times d_a$ and $n \times d_b$ respectively, and $\mathbf{v}$ and $\mathbf{w}$ are random vectors with* **independent** *components. Then, the following two assertions hold:*

(1) If the i-th column of A, $\mathbf{a}_i$, is not proportional to any column of B then v_i is normal (Gaussian).

(2) If $\mathbf{a}_i$ is proportional to $\mathbf{b}_j$, then the logarithms of the characteristic functions of v_i and w_j differ by a polynomial in a neighborhood of the origin.

We will also use the lemmas below to show the last part of Theorem 5.1.1

Lemma 5.1.3 *Consider a nonnegative real δ_0, and assume the validity of the following equation, for all u, v, $|u| < \delta_0$, $|v| < \delta_0$:*

$$\psi_1(u + c_1 v) + \cdots + \psi_d(u + c_d v) = \lambda(u) + \mu(v) + p_k(u, v) \tag{5}$$

where p_k is a polynomial of degree k, $\psi_1, \ldots, \psi_d, \lambda, \mu$ are complex valued functions of real variables. It is moreover assumed that:

(1) the numbers c_j are all distinct,

(2) the functions $\psi_1, \ldots, \psi_d, \lambda, \mu$ are continuous.

Then, there exists a neighborhood of the origin where the $\psi_1, \ldots, \psi_d, \lambda$ and μ are polynomials of degree at most $max(d, k)$.

This first result is lemma 1.5.1 page 29 of [10] and we shall rather use its extension to the multidimensional case; that is lemma 1.5.2 page 32 of [10].

Lemma 5.1.4 *Let A be a $n \times d$ matrix, and denote $\mathbf{a}_i$ its i-th column. Consider the extension of equation (5) for $\mathbf{t} \in \mathbb{R}^m$ such that $\forall i \quad |t_i| < \delta_0$, for continuous functions $\psi_1, \ldots, \psi_d$ and $\xi_1, \ldots, \xi_p$:*

$$\psi_1(\mathbf{a}_1^T \mathbf{t}) + \ldots + \psi_d(\mathbf{a}_d^T \mathbf{t}) = \xi_1(t_1) + \ldots + \xi_p(t_p) \tag{6}$$

Suppose that $\mathbf{a}_i$ is not proportional to any other column of A nor to any column of the identity matrix of order n, I_n. Then any function ψ_i satisfying (6) is a polynomial of degree d at most, in some neighborhood of the origin.

Let us now prove the Theorem 5.1.1. Consider $\mathcal{T}(\mathbf{b})$ (resp. $\mathcal{T}(\tilde{\mathbf{b}})$) the $m \times (m+q)$ toeplitz matrix built from the filter $\mathbf{b}$ (resp. $\tilde{\mathbf{b}}$) [3]:

$$\mathcal{T}(\mathbf{b}) = \begin{pmatrix} b_0 & b_1 & \ldots & b_q & 0 & \ldots & \ldots & 0 \\ 0 & b_0 & b_1 & \ldots & b_q & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & & \ddots & & \vdots \\ \vdots & & \ddots & \ddots & \ddots & & & 0 \\ 0 & \ldots & \ldots & 0 & b_0 & b_1 & \ldots & b_q \end{pmatrix}$$

and let M be the $m \times (2m+q)$ matrix: $M = [\mathcal{T}(\mathbf{b}), I_m]$. Moreover, let us denote $\mathbf{u} = (\mathbf{x}_{m:1-q}^T, \varepsilon_{m:1}^T)^T$ where the distribution of x_i is G_x and the density of ε_i is f_ε. Let $\tilde{M}$ and $\tilde{\mathbf{u}}$ denote similar quantities for $\tilde{\mathbf{b}}, \tilde{G}_x$ and $\tilde{f}_\varepsilon$.

If $P_m(\mathbf{b}, G_x, f_\varepsilon) = P_m(\tilde{\mathbf{b}}, \tilde{G}_x, \tilde{f}_\varepsilon)$, we have two representations for the random vector $\mathbf{y}_{m:1}$:

$$\begin{cases} \mathbf{y}_{m:1} &= M\mathbf{u}, \\ \mathbf{y}_{m:1} &= \tilde{M}\tilde{\mathbf{u}}. \end{cases} \tag{7}$$

[3] the matrix represented here corresponds to the particular value $m = q + 2$

Now if $\mathbf{b} \neq \tilde{\mathbf{b}}$, and as soon as one of the coefficient of $\mathbf{b}$ (else than $b_0 = 1$) is not zero, then there exists a column, say the k-th, of M that is not proportional to any column of $\tilde{M}$ (more precisely it is one of the columns of $\mathcal{T}(\mathbf{b})$). So, according to Theorem 5.1.2, the corresponding variate x_{m-k} is Gaussian, and this contradicts the hypothesis on G_x. This proves the first assertion (3):

$$\mathbf{b} = \tilde{\mathbf{b}}$$

Let us prove the second assertion (4). Denote ϕ_x and ϕ_ε the characteristic functions of G_x and F_ε respectively, and analogous notations for $\tilde{G}_x$ and $\tilde{F}_\varepsilon$. The equality (7) of the two representations of $\mathbf{y}_{m:1}$ implies the equality of their characteristic functions as well, so $\forall \mathbf{t} \in \mathbb{R}^m$ (E denotes the expectation operator):

$$\begin{aligned} E\left[e^{i\mathbf{t}^T M\mathbf{u}}\right] &= E\left[e^{i\mathbf{t}^T \tilde{M}\tilde{\mathbf{u}}}\right] \\ &= E\left[e^{i\mathbf{t}^T \mathcal{T}\mathbf{x}_{m:1-q}}\right] E\left[e^{i\mathbf{t}^T \varepsilon_{m:1}}\right] \\ &= \prod_{k=1}^{m+q} E\left[e^{i\mathbf{t}^T \boldsymbol{\tau}_k x_{m+1-k}}\right] \prod_{l=1}^{m} E\left[e^{it_l \varepsilon_{m+1-l}}\right] \end{aligned}$$

so that:

$$\prod_{k=1}^{m+q} \tilde{\phi}_x(\boldsymbol{\tau}_k^T \mathbf{t}) \prod_{l=1}^{m} \tilde{\phi}_\varepsilon(t_l) = \prod_{k=1}^{m+q} \phi_x(\boldsymbol{\tau}_k^T \mathbf{t}) \prod_{l=1}^{m} \phi_\varepsilon(t_l) \tag{8}$$

where $\boldsymbol{\tau}_k$ is the k-th column of $\mathcal{T}(\mathbf{b}) = \mathcal{T}(\tilde{\mathbf{b}})$ (by (3)). We have used the temporal and the mutual independence of the process x_t and ε_t. Now, if we introduce the functions ψ and ξ such that:

$$\begin{aligned} \psi(u) &= \log \phi_x(u) - \log \tilde{\phi}_x(u), \\ \xi(u) &= \log \phi_\varepsilon(u) - \log \tilde{\phi}_\varepsilon(u). \end{aligned}$$

then taking the logarithm of (8), we obtain:

$$\psi(\boldsymbol{\tau}_1^T \mathbf{t}) + \ldots + \psi(\boldsymbol{\tau}_{m+q}^T \mathbf{t}) = \xi(t_1) + \ldots + \xi(t_m) \tag{9}$$

Application of lemma 5.1.4, is made possible by the particular structure of matrix $\mathcal{T}(\mathbf{b})$ once again; indeed at least one of its column is not proportional to any other column of $\mathcal{T}(\mathbf{b})$, and to any column of the identity matrix. Moreover the functions ψ and ξ are continuous, and then, in a neighborhood of the origin, ψ is a polynomial of degree $m + q$ at most. It is easy to see that ξ is then a polynomial of same degree in the same neighborhood. Moreover, equality (9) implies that the terms in $\psi(\boldsymbol{\tau}_k^T \mathbf{t})$ (for any integer k) must be monomials. Write for u, real in a neighborhood of the origin

$$\psi(u) = \sum_{k=1}^{m+q} p_k u^k$$

so that identifying the terms of degree k that are not monomials in equation (9) implies that if $k \geq 2$ and for any $(l_1, \ldots, l_k)$ such that at least two of them are distinct, we have:

$$\forall \mathbf{t} \text{ such that } |t_i| < \delta\,; \qquad p_k \sum_{j=1}^{m+q} \tau_{l_1 j} \times \ldots \times \tau_{l_k j} t_{l_1} \cdots t_{l_k} = 0.$$

For the particular k-uplet $(l_1, \ldots, l_k) = (1, \ldots, 1, q_0 + 1)$ where q_0 is such that $b_{q_0} \neq 0$ one obtains:

$$\forall |t_1| < \delta, \forall |t_{q_0+1}| < \delta\,; \qquad p_k \times b_{q_0} t_1^{k-1} t_{q_0+1} = 0$$

and thus $p_k = 0$, and this is true $\forall 2 \leq k \leq m + q$. This proves that ψ is linear in some neighborhood of zero, and as $\psi(0) = 0$, there exist a real λ such that $\psi(u) = i\lambda u$, (as x_t and $\tilde{x}_t$ are real random variables, the linear function ψ must be imaginary).

Moreover, the characteristic functions considered are continuously differentiable with respect to the variable $\mathbf{t}$ (we have supposed that the variables have finite first moments) so that ψ must be continuously differentiable as well, and then, the identity $\psi(u) = i\lambda u$ is valid on the whole real line. Then, it is straightforward to see that we have:

$$G_x(\cdot) = \tilde{G}_x(\cdot - \lambda)$$

The same equality may be shown to hold for f_ε and $\tilde{f}_\varepsilon$, but with the appropriate translation factor μ given by the equation (9), so that $\mu = \lambda(b_0 + \ldots + b_q)$

5.2 Split likelihood

Let m be an integer, and p_m the density of $(y_1, \ldots, y_m)$ with respect to $\mu_m = \mu^{\otimes m}$. Denote $\theta \in \Theta$ the whole parameter to estimate, that is the filter $\mathbf{b}$, and some parameters corresponding to distributions of the driving and the measurement noise. Finally let θ_0 be the true parameter, which is assumed to be an interior point of Θ. For the estimation of θ we shall minimize the following contrast function, where E_{θ_0} denotes expectation under the distribution $p_m(\mathbf{y}_{1:m}; \theta_0)$:

$$K(\theta_0, \theta) = E_{\theta_0} \left(\log p_m(\mathbf{y}_{1:m}; \theta_0) \right) - E_{\theta_0} \left(\log p_m(\mathbf{y}_{1:m}; \theta) \right) \tag{10}$$

which is the Kullback-Leibler information of a group of m observations. It is known that $K(\theta_0, \theta) \geq 0; \forall \theta \in \Theta$ and that $K(\theta_0, \theta) = 0$ if and only if $p_m(\mathbf{y}_{1:m}; \theta_0) = p_m(\mathbf{y}_{1:m}; \theta)$, μ_m almost surely. Hence, provided the identifiability assumption

(A1) $p_m(\mathbf{y}_{1:m}; \theta_0) = p_m(\mathbf{y}_{1:m}; \theta_1)$ μ_m *a.s. if and only if* $\theta_0 = \theta_1$.

holds, then $\theta \longrightarrow K(\theta_0, \theta)$ has a unique minimum at $\theta = \theta_0$. Assumption **(A1)** can be studied using the identifiability result presented in the previous section. Now, we shall consider a particular 'split' contrast process. Let r be a fixed integer in $\{1, \ldots, m\}$, and $S = [(T - m)/r]$ ($[u]$ is the integer part of u). The contrast process is:

$$l_T(\theta; r) = \frac{1}{S+1} \sum_{s=0}^{S} \log p_m(y_{sr+1}, \ldots, y_{sr+m}; \theta) \tag{11}$$

We then define $\hat{\theta}_r(T)$ as a maximizer of $l_T(\theta, r)$ over the set Θ. Because y_t is a m-dependent process, it follows from the Ergodic Theorem that under appropriate moments conditions, $l_T(\theta; r)$ converges a.s. (as T tends to infinity) to $K(\theta, \theta_0)$. To prove that $\hat{\theta}_r(T)$ is consistent and asymptotically normal, a somewhat stronger form of central limit Theorem (namely, a function CLT) is required. This stronger form of the CLT is proved by adding to the conventional moment assumptions, some regularity conditions on the 'm-likelihood' w.r.t to the parameter θ. The study of contrast of the form (11) is by now a very standard issue, and many types of different conditions can be imposed. We use here the classical Wald's type assumption (see [24]), introduced by this author to study the maximum likelihood estimate. They are threefold;

(A2) *The function* $\theta \rightarrow p_m(\mathbf{y}_{1:m}; \theta)$ *is continuous for* μ_m*- a.e.* $\mathbf{y}_{1:m}$*, and*

$$\lim_{|\theta| \rightarrow \infty} p_m(\mathbf{y}_{1:m}; \theta) = 0$$

.

(A3) $E_{\theta_0} |\log p_m(\mathbf{y}_{1:m}; \theta_0)| < \infty$*, and for each* $\theta \in \Theta$ *there exists* $\delta > 0$ *such that*

$$E_{\theta_0}(\sup_{|\theta - \theta'| \leq \delta} (\log p_m(\mathbf{y}_{1:m}; \theta')^+) < \infty$$

and there exists $M > 0$ *such that*

$$E_{\theta_0}(\sup_{|\theta'| \geq M} (\log p_m(\mathbf{y}_{1:m}; \theta')^+) < \infty.$$

and the final assumption (required for asymptotic normality):

(A4) *$p_m(\mathbf{y}_{1:m};\theta)$ is twice differentiable w.r.t. θ, μ_m a.s. in a neighborhood of θ_0, and:*

$$\forall i, \quad E_{\theta_0}\left(\frac{\partial}{\partial\theta_i}\log p_m(\mathbf{y}_{1:m};\theta_0)\right)^2 < \infty$$
$$\forall i,j \quad E_{\theta_0}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_m(\mathbf{y}_{1:m};\theta_0)\right| < \infty$$

in addition $E_{\theta_0}(\log p_m(\mathbf{y}_{1:m};\theta))$ may be differentiated twice under the integral sign with respect to θ at θ_0. Moreover for all i,j there exists a function H, $\delta > 0$ and $\alpha > 0$ such that:

$$\sup_{|\theta-\theta_0|\leq\delta}\left|\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_m(\mathbf{y}_{1:m};\theta) - \frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_m(\mathbf{y}_{1:m};\theta_0)\right| \leq |\theta-\theta_0|^\alpha H(\mathbf{y}_{1:m})$$

with $E_{\theta_0}H(\mathbf{y}_{1:m}) < \infty$.

Finally, we shall introduce some notations for the asymptotic bound obtained in Theorem 5.2.1. Define the matrix $J_r(\theta_0)$ by:

$$J_r(\theta_0) = I_0(\theta_0) + \sum_{s=1}^{[(m+q-1)/r]}\left(I_s(\theta_0) + I_s(\theta_0)^T\right)$$

where

$$(I_s)_{ij}(\theta_0) = E_{\theta_0}\left[\frac{\partial}{\partial\theta_i}\log p_m(\mathbf{y}_{1:m};\theta_0)\cdot\frac{\partial}{\partial\theta_j}\log p_m(\mathbf{y}_{sr+1:sr+m};\theta_0)\right]$$

and finally the matrix $D(\theta_0)$, that equals $-I_0(\theta_0)$

$$d_{ij}(\theta_0) = E_{\theta_0}\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_m(\mathbf{y}_{1:m};\theta_0)\right]$$

We may then state:

Theorem 5.2.1 *Assume* **(A1), (A2), (A3)** *hold.*
Then $\hat\theta_r(T)$ converges a.s. to θ_0 as T tends to infinity.

If moreover **(A4)** *holds and the matrix D is invertible, then*

$$\sqrt{T}(\hat\theta_r(T) - \theta_0) \longrightarrow \mathcal{N}(0, \Sigma_r(\theta_0))$$

where

$$\begin{aligned}\Sigma_r(\theta_0) &= r \cdot D(\theta_0)^{-1} J_r(\theta_0) D(\theta_0)^{-1} \\ &= r \cdot \left(I_0(\theta_0)^{-1} + \sum_{s=1}^{[(m+q-1)/r]} I_0(\theta_0)^{-1} \left[I_s(\theta_0) + I_s(\theta_0)^T \right] I_0(\theta_0)^{-1} \right). \qquad (12)\end{aligned}$$

Proof of the Theorem: The proof is a straightforward adaptation of the proof of consistency and asymptotic normality in [21]): the processes $(x_t)_{t\in\mathbb{Z}}$ and $(\varepsilon_t)_{t\in\mathbb{Z}}$ are ergodic and stationary so that we may apply strong laws of large numbers to prove the convergence of the empirical contrast towards the Kullback-Leibler information. The consistency of $\hat{\theta}_r(T)$ is then a straight consequence.

To prove the asymptotic normality, a Taylor expansion together with the CLT for the derivative (w.r.t θ) process yields the result. As the normalized derivative process is a sum involving an at most $m+q$-dependent process, we may apply the extension of the central limit Theorem of Hoeffding and Robbins (see [3]) to it.

Remark : The idea of such 'split-likelihood' originates in the work of Ryden ([21]) for HMM with finite state hidden Markov chain. This author only considers the case $r = m$. However, it appears that there is no need to split the data in non overlapping groups to obtain consistency and asymptotic normality. Only the asymptotic variance will be affected by taking different r's, and it may happen that a smaller r leads to smaller asymptotic variance.

For a given model, to apply Theorem 5.2.1, one has to verify assumptions **(A1)** to **(A4)**. We deal with the three latter assumptions in some details in the next section for the continuous input signal model.

5.3 Continuous input signal model

In this section, the particular case of continuous input in model (1) is addressed. We derive sufficient conditions, easily checkable, for assumptions **(A2)** to **(A4)** to hold. We shall first precise our particular model:

Continuous model (CM): Assume model (1), with the following conditions on the driving and measurement noise distributions:

- f_ε is known up to scale, that is $f_\varepsilon(\cdot) = f(\cdot) = \frac{1}{\sigma} f_1(\frac{\cdot}{\sigma})$ where f_1 is some known density and σ an unknown positive parameter.

- G_x is absolutely continuous with respect to some measure ν on $\mathbb{R}$, $dG_x = g_\gamma d\nu$, where γ is bounded and represents a possible parameterization for the input density.

The parameter to estimate for such a model is $\theta = (b_1, \ldots, b_q, \gamma, \sigma)$. Consider the following assumption:
(H1) *f is continuous, positive μ-a.s., bounded ($\|f\|_\infty < +\infty$), $f(u)$ vanishes as $|u|$ tends to infinity. Define for any real ρ*

$$S(\rho) = \int f(u-\rho) |\log f(u)| d\mu(u)$$

then $S(\rho)$ is finite for all ρ. Moreover, the parameter σ is lower and upper bounded $\sigma \in [\sigma_{min}, \sigma_{max}]$ with $\sigma_{min} > 0$.

Consider the integral form of p_m:

$$p_m(\mathbf{y}_{1:m}; \theta) = \int \prod_{i=1}^{m} f(y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}) \prod_{k=1-q}^{m} g_\gamma(x_k) \tag{13}$$

Assumption **(A2)** is then an obvious consequence of **(H1)** by application of the dominated convergence Theorem of Lebesgue. To see if **(A3)** holds, notice first that p_m is uniformly bounded:

$$\|p_m(\mathbf{y}_{1:m}, \theta)\|_\infty \leq \frac{1}{\sigma_{min}^m} \|f_1\|_\infty^m$$

so that almost surely and for all θ':

$$\log p_m(\mathbf{y}_{1:m}; \theta') \leq m(\log \|f_1\|_\infty - \log \sigma_{min})$$

and

$$(\log p_m(\mathbf{y}_{1:m}; \theta'))^+ \leq m(|\log \|f_1\|_\infty| + |\log \sigma_{min}|)$$

and the second part of **(A3)** follows. Moreover, Jensen's inequality (by concavity of the logarithm) applied to (13) implies that

$$\log p_m(\mathbf{y}_{1:m}; \theta_0) \geq \sum_{i=1}^{m} \int \log \left(\frac{1}{\sigma} f_1 (\frac{y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}}{\sigma}) \right) \prod_{k=1-q}^{m} g_\gamma(x_k) d\nu_{m+q}(\mathbf{x}_{1-q:m})$$

so that

$$\begin{aligned}|\log p_m(\mathbf{y}_{1:m};\theta_0)| &\leq m(|\log \|f_1\|_\infty| + |\log \sigma_{min}|) + \\ &\sum_{i=1}^{m} \int \left|\log f_\sigma(y_i - \mathbf{b}^T \mathbf{x}_{i:i-q})\right| \prod_{k=1-q}^{m} g_\gamma(x_k) d\nu_{m+q}(\mathbf{x}_{1-q:m})\end{aligned}$$

and

$$\begin{aligned}&E_{\theta_0}|\log p_m(\mathbf{y}_{1:m};\theta_0)| \leq m(|\log \|f_1\|_\infty| + 2|\log \sigma_{min}|) + \\ &m \int f_1\left(z - \frac{\mathbf{b}^T(\mathbf{x}_{0:q} - \mathbf{u}_{0:q})}{\sigma}\right) |\log f_1(z)| d\mu(z) \prod_{k=0}^{q} g_\gamma(x_k) g_\gamma(u_k) d\nu_{q+1}(\mathbf{x}_{0:q}) d\nu_{q+1}(\mathbf{u}_{0:q})\end{aligned} \tag{14}$$

Finally the first part of **(A3)** will stem from the following appropriate assumption:

(H2) *For any positive* λ *and any* ρ, $\int S(\rho + \lambda x) g_\gamma(x) d\nu(x) < +\infty$. These two assumptions **(H1)** and **(H2)** will ensure the consistency of the split-likelihood estimate. Now to study assumption **(A4)** used for asymptotic normality, let us recall some already known results for dominated models, commonly used for EM algorithms [19]. Let $p(\mathbf{y}_{1:m}, \mathbf{x}_{1-q:m};\theta)$ be the joint density of $\mathbf{y}_{1:m}$ and $\mathbf{x}_{1-q:m}$ under parameter θ with respect to $\mu_m \otimes \nu_{m+q}$. Define the EM functional for this model:

$$Q(\theta, \theta') = E\left[\log p(\mathbf{y}_{1:m}, \mathbf{x}_{1-q:m};\theta) | \mathbf{y}_{1:m};\theta'\right].$$

When it is possible to differentiate under the expectation it is well known that:

$$\begin{aligned}\frac{\partial}{\partial\theta_i} Q(\theta,\theta) &= \frac{\partial}{\partial\theta_i} \log p_m(\mathbf{y}_{1:m};\theta) \\ &= E\left[\frac{\partial}{\partial\theta_i} \log p(\mathbf{y}_{1:m}, \mathbf{x}_{1-q:m};\theta)\middle| \mathbf{y}_{1:m};\theta\right]\end{aligned} \tag{15}$$

furthermore, if we can differentiate twice,

$$\begin{aligned}\frac{\partial^2}{\partial\theta_i\partial\theta_j} \log p_m(\mathbf{y}_{1:m};\theta) &= \frac{\partial^2}{\partial\theta_i\partial\theta_j} Q(\theta,\theta) \\ &+ E\left[\frac{\partial}{\partial\theta_i} \log p(\mathbf{x}_{1-q:m}|\mathbf{y}_{1:m};\theta) \frac{\partial}{\partial\theta_j} \log p(\mathbf{x}_{1-q:m}|\mathbf{y}_{1:m};\theta)\middle| \mathbf{y}_{1:m};\theta\right]\end{aligned}$$

and we may use the identity

$$\log p(\mathbf{x}_{1-q:m}|\mathbf{y}_{1:m};\theta) = \log p(\mathbf{x}_{1-q:m}, \mathbf{y}_{1:m};\theta) - \log p_m(\mathbf{y}_{1:m};\theta)$$

to derive:

$$
\begin{aligned}
&\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p_m(\mathbf{y}_{1:m};\theta) &&(16)\\
&= E\left[\frac{\partial^2}{\partial\theta_i\partial\theta_j}\log p(\mathbf{x}_{1-q:m},\mathbf{y}_{1:m};\theta)\middle|\,\mathbf{y}_{1:m},\theta\right]\\
&+E\left[\frac{\partial}{\partial\theta_i}\log p(\mathbf{x}_{1-q:m},\mathbf{y}_{1:m};\theta)\frac{\partial}{\partial\theta_j}\log p(\mathbf{x}_{1-q:m},\mathbf{y}_{1:m};\theta)\middle|\,\mathbf{y}_{1:m};\theta\right] &&(17)\\
&-E\left[\frac{\partial}{\partial\theta_i}\log p(\mathbf{x}_{1-q:m},\mathbf{y}_{1:m};\theta)\middle|\,\mathbf{y}_{1:m};\theta\right]E\left[\frac{\partial}{\partial\theta_j}\log p(\mathbf{x}_{1-q:m},\mathbf{y}_{1:m};\theta)\middle|\,\mathbf{y}_{1:m};\theta\right]
\end{aligned}
$$

Let us introduce the following last assumption:

(H3) *f is twice differentiable, $\log p_m$ and its expectation may be differentiated twice under the integral sign. Assume that $uf(u)$, $f'(u)$ and $u^2f'(u)$ vanish as $|u|$ goes to infinity, and*

$$
\begin{aligned}
&\int \frac{f'^2}{f}(u)d\mu(u) < +\infty\\
&\int u^2\frac{f'^2}{f}(u)d\mu(u) < +\infty\\
&\int |f''(u)|d\mu(u) < +\infty
\end{aligned}
$$

Moreover, suppose that g_γ has a finite moment of order 2 and

$$
\begin{aligned}
E_{\theta_0}\left[\left(\frac{\partial}{\partial\gamma}\log g_\gamma(x)\right)^2\right] &< +\infty\\
E_{\theta_0}\left[\frac{\partial^2}{\partial\gamma_i\partial\gamma_j}\log g_\gamma(x)\right] &< +\infty
\end{aligned}
$$

Let us state:

Theorem 5.3.1 *For the model* **(CM)** *with continuous input signals,* **(A2)** *and* **(A3)** *hold under* **(H1)** *and* **(H2)**. **(A4)** *holds under* **(H1)** *and* **(H3)**.

Proof of Theorem 5.3.1.

The first part of this Theorem has previously been proved, we shall now prove that assumption **(A4)** holds, assuming **(H1)** and **(H3)**. Looking at the factorized expression of $p(\mathbf{y}_{1:m}, \mathbf{x}_{1-q:m}; \theta)$:

$$p_m(\mathbf{y}_{1:m}, \mathbf{x}_{1-q:m}; \theta) = \prod_{i=1}^{m} f_\sigma(y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}) \prod_{i=1-q}^{m} g_\gamma(x_i) \tag{18}$$

we may write the derivative of $p_m(\mathbf{y}_{1:m}; \theta)$ w.r.t. θ. Indeed, inserting (18) in equation (15) yields:

$$\frac{\partial}{\partial \mathbf{b}} \log p_m(\mathbf{y}_{1:m}; \theta) = -E\left[\sum_{i=1}^{m} \mathbf{x}_{i:i-q} \frac{f'_\sigma}{f_\sigma}(y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}) \middle| \mathbf{y}_{1:m}; \theta\right] \tag{19}$$

$$\frac{\partial}{\partial \sigma} \log p_m(\mathbf{y}_{1:m}; \theta) = -E\left[\sum_{i=1}^{m} \frac{y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}}{\sigma^2} \frac{f'_\sigma}{f_\sigma}(y_i - \mathbf{b}^T \mathbf{x}_{i:i-q}) \middle| \mathbf{y}_{1:m}; \theta\right] - \frac{m}{\sigma} \tag{20}$$

$$\frac{\partial}{\partial \gamma} \log p_m(\mathbf{y}_{1:m}; \theta) = E\left[\sum_{i=1-q}^{m} \frac{\partial}{\partial \gamma} \log g_\gamma(x_i) \middle| \mathbf{y}_{1:m}; \theta\right] \tag{21}$$

now, we shall use the following conditional Jensen's inequality:

$$\begin{aligned} E_\theta\left[\left(E\left[\alpha(\mathbf{x}_{1-q}, \mathbf{y}_{1:m}; \theta) | \mathbf{y}_{1:m}; \theta\right]\right)^2\right] &\leq E_\theta\left[E\left[\left(\alpha(\mathbf{x}_{1-q}, \mathbf{y}_{1:m}; \theta)\right)^2 | \mathbf{y}_{1:m}; \theta\right]\right] \\ &\leq E_\theta\left[\left(\alpha(\mathbf{x}_{1-q}, \mathbf{y}_{1:m}; \theta)\right)^2\right] \end{aligned} \tag{22}$$

where α is any measurable function. Applying inequality (22) to the (co-)variance of (19) proves that it is finite as soon as we have:

$$\int \frac{f'^2}{f}(u) d\mu(u) < +\infty \tag{23}$$

Same computations prove that the (co-)variance of (20) is finite as soon as $uf(u)$ vanishes as $|u|$ tends to infinity, and:

$$\int u^2 \frac{f'^2}{f}(u) d\mu(u) < +\infty \tag{24}$$

and finally, the variance of (21) is obviously finite as soon as the variance of $\frac{\partial}{\partial \gamma} \log g_\gamma(x)$ is.

Now, to see that

$$E_{\theta_0} \left| \frac{\partial^2}{\partial \theta_k \partial \theta_l} \log p_m(\mathbf{y}_{1:m}; \theta_0) \right|$$

is finite use equation (16) and similar computations as before to find that it is bounded as soon as the last parts of **(H3)** are verified.

Remark: If the input signal is a finite Gaussian mixture, and the noise is Gaussian it is easily seen that all the assumptions **(H1)**, **(H2)** and **(H3)** are verified as soon as the variances in the mixture are upper and lower bounded. This is the case on which we will focus in section 5.4.

5.4 Algorithm and numerical experiments

In this section, we present an algorithm that computes an approximation of the split-likelihood estimate, for the estimation of the parameters of the model **(CM)**:

$$y_t = x_t + b_1 x_{t-1} + \cdots + b_q x_{t-q} + \varepsilon_t \quad t \in \mathbb{Z}$$

We will moreover assume Gaussian noise in the following; $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$. Recall that we denote $\theta = (\mathbf{b}, \sigma^2, \gamma)$ where γ is some parameter of the input distribution. As the input process $\{x_t\}$ is not observed, it is useful to resort to the EM paradigm ([5]). The EM algorithm is an iterative method used for minimum contrast estimation. It consists of two steps; given a current parameter, one computes an approximation of the contrast by introducing some latent (unobserved) data (E-step) and then one updates the parameter by maximizing this approximation (M-step). The natural latent data for the model is clearly the input signal, but we will introduce auxiliary unobserved variables in the particular case where the input distribution is modeled by a mixture of Gaussian distributions. This will allow us to *jointly* reestimate the parameters of interest (the filter and the noise variance) and the nuisance parameters (those of the Gaussian mixture). We shall in a first time recall known results on Gaussian mixtures and on the EM algorithm, afterwards, we will concentrate on our particular model, and present a practically implementable algorithm (a version of this algorithm has already been presented in [16]).

Gaussian mixtures: We model the distribution of x_t as a mixture of K Gaussian densities:

$$g_\gamma(x) = \sum_{i=1}^{K} \pi_i \phi(x, \mu_i, \sigma_i^2) \tag{25}$$

where $\phi(\cdot, m, v)$ denotes the Gaussian kernel with mean m and variance v. The parameter $\gamma = (\pi_i, \mu_i, \sigma_i^2)_{1 \le i \le K}$ contains respectively the weights, means and variances associated to the mixture. The interest of such a model is twofold; as we shall see, the EM steps involve discrete sums of moments of Gaussian variables, thus allowing exact computation through an enumeration. Moreover, most distributions may be approximated with arbitrary precision by such a Gaussian mixture. When dealing with mixtures of densities, it is very convenient to introduce a hidden 'labeling' discrete variable z_t taking its values in $\{1, \ldots, K\}$ such that:

$$\begin{aligned} &P(z_t = i) = \pi_i \\ &p(x_t | z_t = i; \gamma) = \phi(x_t, \mu_i, \sigma_i^2) \end{aligned}$$

The conditional distribution of x_t with respect to the label z_t is then Gaussian.

EM algorithm: Let us recall the EM algorithm for our particular contrast: given an initial parameter, we repeat these two steps until convergence:

- **E-step:** compute the conditional expectation:

$$Q(\theta|\theta') = E[L_T(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T}, \mathbf{z}_{1-q:T}, \theta, r)|Y; \theta'] \tag{26}$$

- **M-step:** update the current parameter $\theta^{(k)}$ by:

$$\theta^{(k+1)} = \arg\max_{\theta} Q(\theta|\theta^{(k)}),$$

where we have defined the 'complete' split contrast depending on the overlap parameter r (S denoting $[(T-m)/r]$):

$$L_T(\mathbf{y}_{1:T}, \mathbf{x}_{1-q:T}, \mathbf{z}_{1-q:T}, \theta, r) = \frac{1}{S+1} \sum_{s=0}^{S} \log p(\mathbf{y}(s), \mathbf{x}(s), \mathbf{z}(s); \theta) \tag{27}$$

For convenience, we have denoted $\mathbf{y}(s) = \mathbf{y}_{sr+m:sr+1}$, $\mathbf{x}(s) = \mathbf{x}_{sr+m:sr+1-q}$, and $\mathbf{z}(s) = \mathbf{z}_{sr+m:sr+1-q}$. The key of the convergence of the EM is that at each step, the split likelihood increases. Now we shall use the particular nested structure of our model to simplify the evaluation of the EM functional (26). Indeed, we may write:

$$\begin{aligned} p(\mathbf{y}(s), \mathbf{x}(s), \mathbf{z}(s), \theta) &= p(\mathbf{y}(s)|\mathbf{x}(s), \mathbf{z}(s), \theta) \times p(\mathbf{x}(s)|\mathbf{z}(s), \theta) \times p(\mathbf{z}(s), \theta) \\ &= p(\mathbf{y}(s)|\mathbf{x}(s), \eta) \times p(\mathbf{x}(s)|\mathbf{z}(s); \gamma) \times p(\mathbf{z}(s); \gamma) \end{aligned} \tag{28}$$

this factorization allows us to write Q as a sum of two terms, the first one depending only on the parameter of interest $(\mathbf{b}, \sigma^2)$ and the second one on γ:

$$Q(\theta|\theta') = Q_i(\mathbf{b}, \sigma^2|\theta') + Q_n(\gamma|\theta').$$

where

$$\begin{aligned} Q_i(\eta|\theta') &= \frac{1}{S+1}\sum_{s=0}^{S} E[\log p(\mathbf{y}(s)|\mathbf{x}(s);\eta)|\mathbf{y}(s);\theta'] \\ Q_n(\gamma|\theta') &= \frac{1}{S+1}\sum_{s=0}^{S} E[\log p(\mathbf{x}(s),\mathbf{z}(s);\gamma)|\mathbf{y}(s);\theta'] \end{aligned}$$

This decomposition allows the M-step to be performed in two stages; the maximization of Q_i, which is straightforward for our particular prior, it yields the next reestimate of the filter and the noise variance. The second stage is the re-estimation of the parameter of the Gaussian mixture, which amounts to fitting such a prior distribution on noisy (yet not convolved) data. We shall precise these two steps, and see that the modeling chosen for the prior density is a promising approach as it merely simplifies the computations involved.

Parameters of interest η: Because the additive measurement noise is assumed to be Gaussian, the function to maximize at each step of the EM for the parameter $\mathbf{b}$ is quadratic. That is, up to an additive constant:

$$\begin{aligned} Q_i(\eta|\theta^{(k)}) &= -\frac{1}{2\sigma^2(S+1)}\sum_{s=0}^{S}\sum_{i=1}^{m} E[(y_{sr+i} - \mathbf{b}^T\mathbf{x}_{sr+i:sr+i-q})^2|\mathbf{y}(s);\theta^{(k)}] - \frac{m}{2}\log(2\pi\sigma^2) \\ &= -\frac{1}{2\sigma^2}\left(\hat{R}_{yy} - 2\hat{R}_{yx}\mathbf{b} + \mathbf{b}^T\hat{R}_{xx}\mathbf{b}\right) - \frac{m}{2}\log(2\pi\sigma^2) \end{aligned}$$

Where we have defined the empirical 'split' posterior covariances and correlation of y and x (for the current parameter $\theta^{(k)}$:

$$\begin{aligned} \hat{R}_{yy} &= \frac{1}{S+1}\sum_{s=0}^{S}\sum_{i=1}^{m} \mathbf{y}_{sr+i}^2 \\ \hat{R}_{yx} &= \frac{1}{S+1}\sum_{s=0}^{S}\sum_{i=1}^{m} \mathbf{y}_{sr+i} E[\mathbf{x}_{sr+i:sr+i-q}|\mathbf{y}(s);\theta^{(k)}]^T \\ \hat{R}_{xx} &= \frac{1}{S+1}\sum_{s=0}^{S}\sum_{i=1}^{m} E[\mathbf{x}_{sr+i:sr+i-q}\mathbf{x}_{sr+i:sr+i-q}^T|\mathbf{y}(s);\theta^{(k)}] \end{aligned} \tag{29}$$

To derive the re-estimate of $\mathbf{b}$ with the constraint that $\mathbf{b} = (1, \tilde{\mathbf{b}})$, we have to use extracted elements of these matrices; i.e. $\hat{R}_{yx} = (R_1, R_2)$ with the scalar $R_1 = \left(\hat{R}_{yx}\right)_1$ and in the same manner, consider the block decomposition:

$$\hat{R}_{xx} = \begin{pmatrix} R_{11} & R_{12} \\ R_{21} & R_{22} \end{pmatrix}$$

where R_{11} is scalar and R_{22} is a $q \times q$ matrix. With these notations we have:

$$\tilde{\mathbf{b}}^{(k+1)} = R_{22}^{-1}(R_2 - R_{12})^T \tag{30}$$

Similar computations yield the re-estimate of σ^2:

$$\begin{aligned} \hat{\sigma}^{2(k+1)} &= \frac{1}{m}\left(\hat{R}_{yy} - \hat{R}_{yx}(\hat{R}_{xx})^{-1}\hat{R}_{yx}^T\right) \\ &= \frac{1}{m}\left(\hat{R}_{yy} - 2R_1 + R_{11} - (R_2 - R_{12})R_{22}^{-1}(R_2 - R_{12})^T\right) \end{aligned}$$

So, for the re-estimation of the parameters of interest, it is only needed to compute the empirical posterior mean and covariance of $\mathbf{x}_{sr+i:sr+i-q}$ conditionally on the observation of $\mathbf{y}(s)$, (for all i and and all s) under the current parameter. This is in general intractable, due to the complexity of the posterior distribution, but we will show that the modeling of the input distribution as a mixture of Gaussian allows exact computation of these moments.

With $\varepsilon(s) = (\varepsilon_{sr+m}, \ldots, \varepsilon_{sr+1})^T$, the model may be seen as a pseudo source separation (instantaneous mixture) model, except for the temporal independence of the 'sources' (in fact, the use of a split contrast is equivalent to suppose this block independence):

$$\mathbf{y}(s) = \mathcal{T}\mathbf{x}(s) + \varepsilon(s) \quad 0 \leq s \leq S. \tag{31}$$

Where $\mathcal{T} = \mathcal{T}(\mathbf{b})$ is the $m \times m + q$ Toeplitz matrix associated with the filter:

$$\mathcal{T} = \begin{pmatrix} b_0 & b_1 & \ldots & b_q & 0 & \ldots & 0 \\ 0 & b_0 & b_1 & \ldots & b_q & \ddots & \vdots \\ \vdots & \ddots & \ddots & & \ddots & 0 & \\ 0 & \ldots & 0 & b_0 & b_1 & \ldots & b_q \end{pmatrix}$$

So conditionally on the label $\mathbf{z}(s)$, the model (31) is a linear Gaussian model, for which the posterior moments of the input are easily computed. Indeed, for each s the conditional expectations sought may be expressed as:

$$E[h(\mathbf{x}(s))|\mathbf{y}(s), \eta] = \sum_{\mathbf{z}} E[h(\mathbf{x}(s))|\mathbf{y}(s), \mathbf{z}(s) = \mathbf{z}, \eta]p(\mathbf{z}|\mathbf{y}(s), \eta) \tag{32}$$

where the sum over $\mathbf{z}$ should be read as the sum over all indices $z_{1-q}, \dots, z_m$ (thus involving K^{m+q} terms). Let us introduce the notations for the mean vector and diagonal covariance of $\mathbf{x}(s)$ conditionally on the label $\mathbf{z}$:

$$\begin{aligned}
\mu(\mathbf{z}) &= (\mu_{z_{sr+m}}, \dots, \mu_{z_{sr+1-q}})^T \\
R_{\mathbf{z}} &= \begin{pmatrix} \sigma^2_{z_{sr+m}} & & O \\ & \ddots & \\ O & & \sigma^2_{z_{sr+1-q}} \end{pmatrix}
\end{aligned}$$

and finally let

$$\Gamma_{\mathbf{z}} = \mathcal{T} R_{\mathbf{z}} \mathcal{T}^T + \sigma^2 I$$

so that $p(\mathbf{y}(s)|\mathbf{z}(s) = \mathbf{z}, \theta)$ is a Gaussian density $\mathcal{N}(\mathcal{T}\mu(\mathbf{z}), \Gamma_{\mathbf{z}})$ and Bayes' formula yields the posterior probability of $\mathbf{z}(s)$ conditionally on $\mathbf{y}(s)$ under θ:

$$p(\mathbf{z}|\mathbf{y}(s);\theta) = \frac{p(\mathbf{y}(s)|\mathbf{z}(s) = \mathbf{z};\theta) p(\mathbf{z};\theta)}{p(\mathbf{y}(s);\theta)}, \tag{33}$$

(where $p(\mathbf{z};\theta)$ is a product of π_i), so that:

$$\begin{aligned}
\omega_s(\mathbf{z}_{1-q:m}, \theta) &\triangleq p(\mathbf{z}(s) = (z_m, \dots, z_{1-q})|\mathbf{y}(s), \theta) && (34) \\
&\propto \frac{\pi_{\mathbf{z}_m} \cdots \pi_{\mathbf{z}_{1-q}}}{|\det(2\pi\Gamma_{\mathbf{z}})|^{\frac{1}{2}}} \exp\left[(\mathbf{y}(s) - \mathcal{T}\mu(\mathbf{z}))^T \Gamma_{\mathbf{z}}^{-1} (\mathbf{y}(s) - \mathcal{T}\mu(\mathbf{z}))\right] && (35)
\end{aligned}$$

Finally, for each $\mathbf{z}$, the previous conditional expectations (32) are given by explicit formulas (see for instance [12]).

$$\begin{aligned}
E[\mathbf{x}(s)|\mathbf{y}(s), \mathbf{z}(s) = \mathbf{z}] &= \mu(\mathbf{z}) + R_{\mathbf{z}} \mathcal{T}^T \Gamma_{\mathbf{z}}^{-1} (\mathbf{y}(s) - \mathcal{T}\mu(\mathbf{z})) = M_t(\mathbf{z}) \\
E[\mathbf{x}(s)\mathbf{x}(s)^T|\mathbf{y}(s), \mathbf{z}(s) = \mathbf{z}] &= R_{\mathbf{z}} - R_{\mathbf{z}} \mathcal{T}^T \Gamma_{\mathbf{z}}^{-1} \mathcal{T} R_{\mathbf{z}} + M_t(\mathbf{z}) M_t(\mathbf{z})^T
\end{aligned}$$

Nuisance parameter γ: The re-estimation of the parameters of the Gaussian mixture is almost as straightforward as the case of a noisy observation of this mixture. Only the posterior weights on the labels are different. We may further decompose the functional Q_n by:

$$\begin{aligned}
Q_n(\gamma, \theta') &= \frac{1}{S+1} \sum_{s=0}^{S} E[\log p(\mathbf{x}(s), \mathbf{z}(s); \gamma)|\mathbf{y}(s); \theta'] \\
&= \frac{1}{S+1} \sum_{s=0}^{S} \sum_{\mathbf{z}_{1-q}, \dots, \mathbf{z}_m} E\left[\log \pi_{\mathbf{z}} \phi(\mathbf{x}(s), \mu(\mathbf{z}), R_{\mathbf{z}})|\mathbf{y}(s), \mathbf{z}; \theta'\right] \omega_s(\mathbf{z}_{1-q:m}; \theta') && (36) \\
&= Q_1(\mu_{1:K}, \sigma^2_{1:K}; \theta') + Q_2(\pi_{1:K}; \theta'), && (37)
\end{aligned}$$

where $\pi_{\mathbf{z}}$ is the product of the corresponding weights $\pi_{z_m} \times \cdots \times \pi_{z_{1-q}}$. It is interesting to notice that the posterior weights ω_s do not depend on the parameter γ, but only on the current version θ' of the whole parameter. The last expression of Q_n shows that the re-estimation of the weights may be performed independently from this of the means and variances.

A constrained maximization of Q_2, with the constraint that the weights add to one, $\sum_{i=1}^{K} \pi_i = 1$, yields:

$$\begin{aligned}\hat{\pi}_i^{(k+1)} &= \frac{1}{(S+1)(m+q)} \sum_{s=0}^{S} \sum_{\mathbf{z}_{1-q:m}} [\delta(z_{1-q} = i) + \cdots + \delta(z_m = i)] \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)}), \\ &= \frac{1}{(S+1)(m+q)} \sum_{s=0}^{S} \sum_{\mathbf{z}_{1-q:m}} \left(\mathbf{d}(\mathbf{z}, i)^T \mathbf{1}\right) \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)}), \end{aligned} \tag{38}$$

where $\mathbf{d}(\mathbf{z}, i)$ is the vector whose k-th entry is $\delta(z_{m+1-k} = i)$ (that is the Kronecker symbol; one when $z_{m+1-k} = i$ and zero otherwise), and $\mathbf{1}$ is the vector of $\mathbb{R}^{m+q}$ composed of one's: $(1, \ldots, 1)^T$.

For the means and variances, the re-estimation is straightforward, thanks to the simplicity of the function Q_1 (up to a constant):

$$\begin{aligned}& Q_1(\mu_{1:K}, \sigma^2_{1:K}; \theta') \\ =\ & \frac{-1}{S+1} \sum_s \sum_{\mathbf{z}_{1-q:m}} \sum_{k=1-q}^{m} \left[\frac{E[(x_{sr+k} - \mu_{z_k})^2 | \mathbf{y}(s), \mathbf{z}; \theta']}{\sigma^2_{z_k}} + \frac{1}{2} \log \sigma^2_{z_k} \right] \omega_s(\mathbf{z}_{1-q:m}, \theta')\end{aligned}$$

Cancelling the derivatives of Q_1 turns into $\forall 1 \leq i \leq K$:

$$\hat{\mu}_i^{(k+1)} = \frac{\sum_s \sum_{\mathbf{z}_{1-q:m}} \mathbf{d}(\mathbf{z}, i)^T E[\mathbf{x}_{sr+m:sr+1-q} | \mathbf{y}(s), \mathbf{z}; \theta^{(k)}] \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)})}{\sum_s \sum_{z_{1-q:m}} (\mathbf{d}(\mathbf{z}, i)^T \mathbf{1}) \, \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)})} \tag{39}$$

that is $\forall 1 \leq i \leq K$:

And finally for the variances $\forall 1 \leq i \leq K$:

$$\begin{aligned}\hat{\sigma}_i^{2(k+1)} &= \frac{\sum_s \sum_{\mathbf{z}_{1-q:m}} \mathbf{d}(\mathbf{z}, i)^T E[\mathbf{x}^2_{sr+m:sr+1-q} | \mathbf{y}(s), \mathbf{z}; \theta^{(k)}] \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)})}{\sum_s \sum_{z_{1-q:m}} (\mathbf{d}(\mathbf{z}, i)^T \mathbf{1}) \, \omega_s(\mathbf{z}_{1-q:m}, \theta^{(k)})} \\ &\quad -(\hat{\mu}_i^{(k+1)})^2.\end{aligned}$$

where $\mathbf{x}^2_{sr+m:sr+1-q}$ denotes the vector $(x^2_{sr+m}, \ldots, x^2_{sr+1-q})^T$.

As we have seen, the modeling of the input by Gaussian mixtures allows explicit computation of the posterior density of the input, and thus that of the EM functional, and the parameters of such a mixture may be actualized at each step of the EM, for a low cost. Indeed, the main quantities; the posterior weights ω_s and the posterior moments of the input x for each of the labels have already been computed for the re-estimation of the filter and the noise.

Moreover, it is also possible to compute an approximate asymptotic covariance of the estimator, the Cramer Rao Bound (CRB) Σ_r as defined in Theorem 5.2.1, equation (12). Indeed, equation (15) shows that it is only necessary to compute the derivative of the functional Q (for $\theta = \theta' = \theta_0$) in order to have an expression for the CRB. This derivative may be computed by the very same conditioning on the labels. The expectation is then empirically achieved throughout a Monte-Carlo integration for instance.

5.5 Simulations:

We present here a brief illustration of the previous algorithm in a simple case. We consider $T = 1000$ samples of a noisy MA(2) model with random filter coefficients. We present the convergence of the algorithm for a Signal to Noise Ratio (SNR) of 10 dB.

Remark: we have defined the SNR as $10\log_{10}(\frac{\sigma_s^2}{\sigma^2}\sum_{k=0}^{q} b_k^2)$; it is the ratio of the variance of the noiseless observation over the variance of the noise (expressed in dB). We consider two different mixtures of Gaussian, and present for each some empirical results on the asymptotic Cramer Rao Bound.

- the first input distribution is composed of two major modes , i.e. a mixture of two Gaussian distributions, weighted by $\pi_1 = \pi_2 = 0.5$, with means and variances respectively $\mu_1 = -1, \mu_2 = 1$ and $\sigma_1 = \sigma_2 = 0.2$. It is a good way for modeling binary (or multi-modal) signals (e.g. in numerical communications). It is zero-mean, with $\sigma_s^2 = 1.2$, and negative kurtosis of order -1.4.

- The second one is heavy-tailed, corresponding to sources often zeros with some outliers; this kind of distribution may capture the general features of so-called 'sparse' signals. Here, the first Gaussian with mean 0 and variance 5 is weighted by 0.2, while the second with same zero mean and variance 0.2 (much more peaky) is highly weighted by 0.8. It is zero-mean, with $\sigma_s^2 = 1.16$, and kurtosis of 8.2.

In figure 5.1 we show the convergence of both the parameters of interest (filter and noise variance) and those of nuisance (input distribution), in the sub-figures "variances", the square root of the variances have been plotted. Figure 5.3 shows the evolution of the input distribution, showing the capability of the algorithm to capture the shape of the distributions.

The initialization was random for the filter, and the initial input distribution was composed of two opposite random means with same weights (=0.5) and random variances. Finally, the initial noise variance was set to have an initial SNR of 10dB.

The convergence is reasonably fast, requiring really about 200 iterations. The precision of the estimation was of order $1/\sqrt{(T)}$.

Next, we focus on the asymptotic performance; figure 5.2 shows the evolution of the CRB (for the parameter $\mathbf{b}$ only) when the SNR increases, for the two previous distributions. In fact, we have represented the trace of the CRB, so it corresponds to the mean square error on the filter. For convenience, we have represented $10\log_{10}(\mathrm{trace}(I_\theta))$. It is computed for the previous model throughout a Monte-Carlo integration (with 20000 samples). We have represented three different bounds; one referred as "Input/Output" is the expression of the complete CRB, when the input signal X is also observed; it is linear as a function of the SNR (in dB). It is computed the same way as Σ_r, except that instead of using the derivative of the log-density (score function) of the packs of $y_{sr+1},\ldots,y_{sr+m}$, we use the score function associated with the complete data packs $\mathbf{y}_{sr+1:sr+m}, \mathbf{x}_{sr+1-q:sr+m}$. It is in fact the very same computations by replacing the posterior moments of the input by their empirical (observed) ones.

The second bound is the CRB for noise free observations; computed for the same contrast, with no additional noise those two bounds are references for the bound of the model (blindness and noise). Indeed, we know that the bound for noisy observation must be higher than the noiseless one, but we observe that they are in fact very close as soon as the SNR is greater than 20 dB (in fact, there is a continuity of the bound as a function of σ^2 in the neighborhood of zero). For low SNR, the bound tends to increase linearly, but about twice faster than the Input/Output bound. For SNR between -20 to 20 dB, we can see that the bound evolves parallel to the I/O, but is still higher.

We have generally observed that for different r's (here $r = 1, 2$ or 3), the behavior of the asymptotic bound is the same, but for the middle overlap parameter ($r = 2$), the bound is a bit lowered (this phenomenon has happened for every simulation). So there might be a possibility to increase the performances by choosing the good overlap (but we have no theoretical result in this way yet).

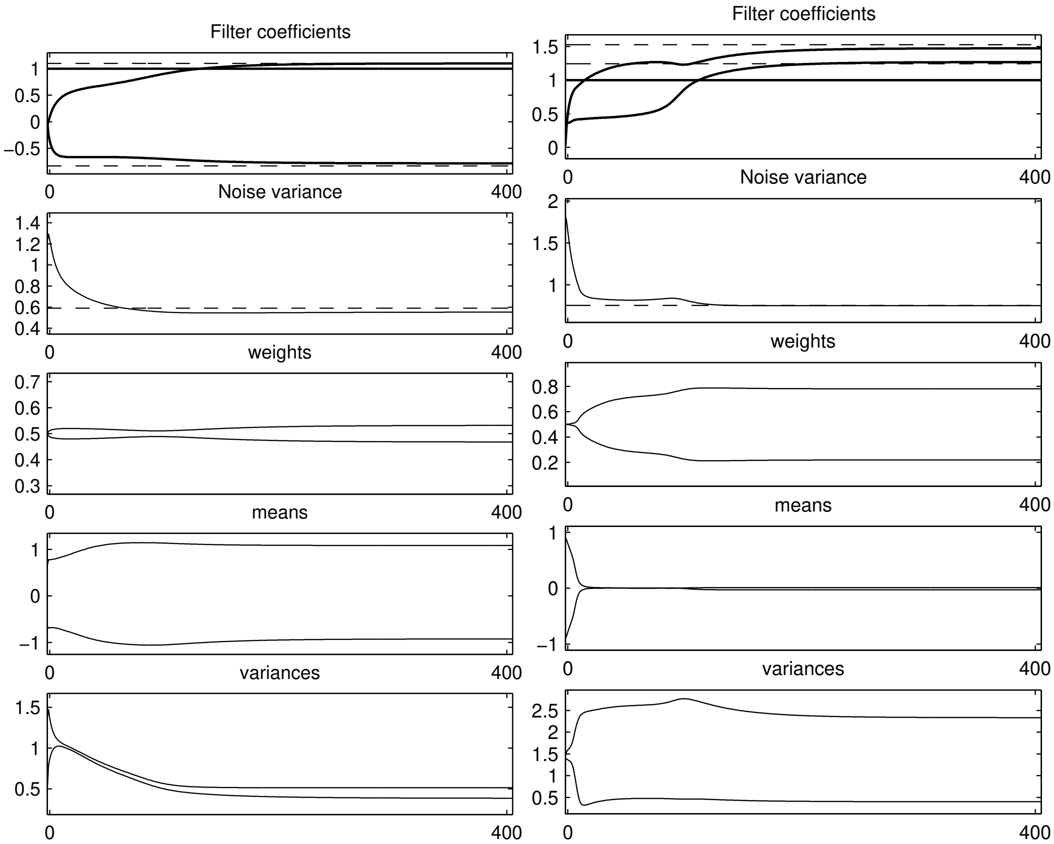

Figure 5.1 : *Convergence of the EM re-estimates (SNR = 10dB). Left: bimodal distribution, right: heavy-tailed distribution.*

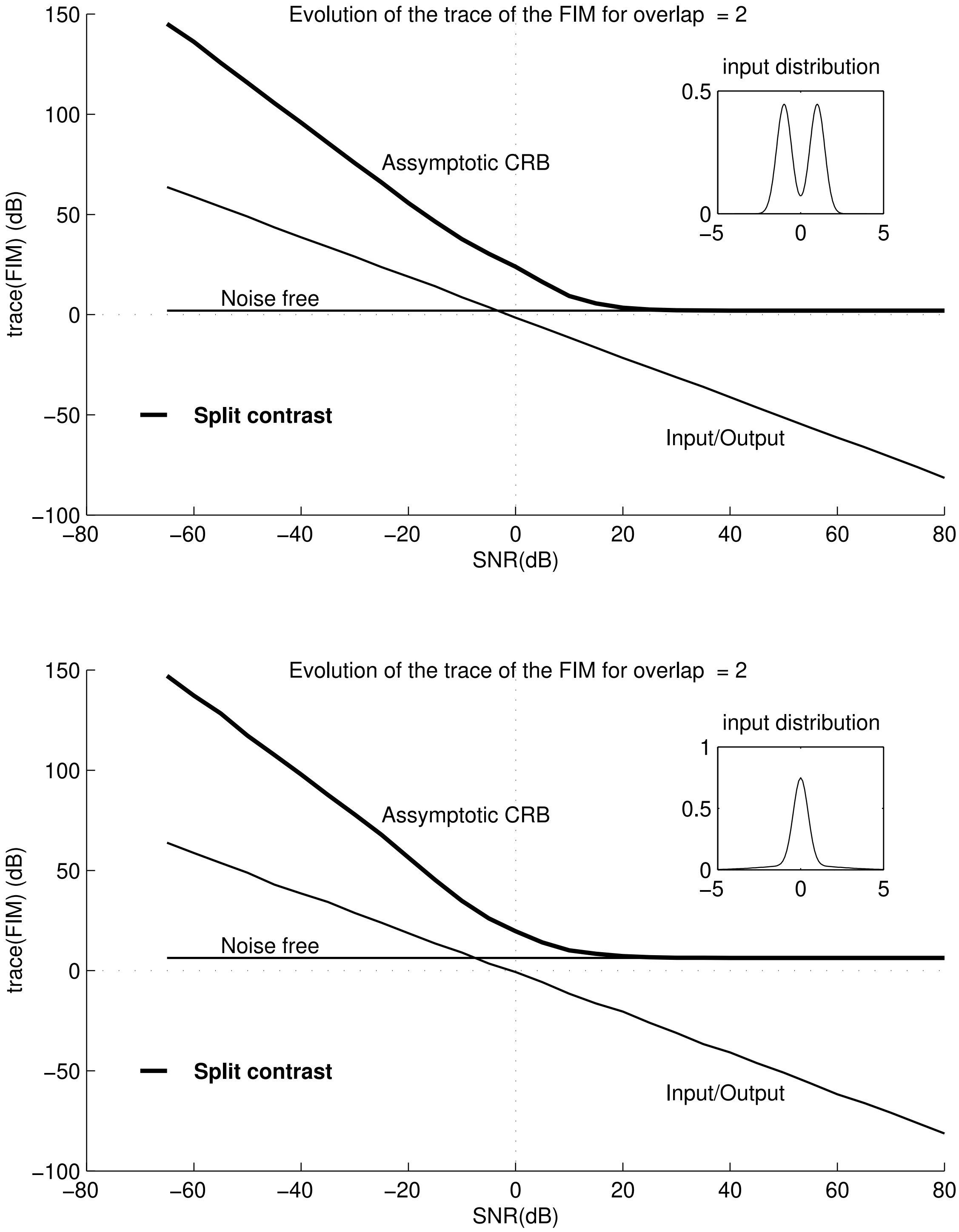

Figure 5.2 : *CRB vs SNR.*

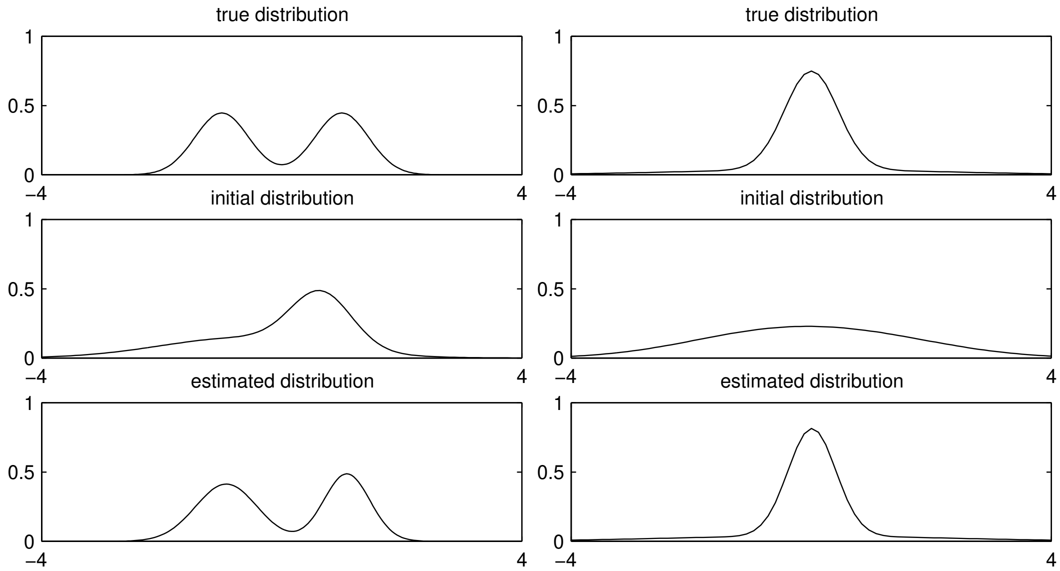

Figure 5.3 : *Evolution of sources distributions. Left: bimodal, right: heavy-tailed.*

Bibliographie

[1] C. Antón-Haro, J.A.R. Fonollosa, and J.R. Fonollosa. Blind channel estimation and data detection using hidden markov models. In *IEEE Trans. Signal Processing*, volume 45 (1), January 1997.

[2] P.J. Bickel, Y. Ritov and T. Ryden. Asymptotic normality of the maximum likelihood estimator for general hidden Markov models. *Annals of Stat.*, 26: 1614-1635, 1998.

[3] P. J. Brockwell and R. A. Davis. *Times series: theory and methods.* Springer series in statistics, second edition, 1991.

[4] J.A. Cadzow. Blind deconvolution via cumulant extrema. *IEEE Signal Proc. Magazine*, 1996.

[5] A. P. Dempster, N. M. Laird, and D.B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Stat. Soc. B*, 39: 1–38, 1977.

[6] D.L. Donoho. On minimum entropy deconvolution. *Applied Time-Series Analysis. Academic Press*, 1981.

[7] G. Giannakis and J.M. Mendel. Identification of non-minimum phase systems using higher-order statistics. *IEEE Trans. Acoust., Speech, Signal Processing*, 37: 360–377, 1989.

[8] G.B. Giannakis and A. Swami. On estimating noncausal nonminimum phase ARMA models of non-gaussian processes. *IEEE Trans. on Acoustics, Speech, and Signal Processing*, 38: 478–495, 1990.

[9] S. Haykin. Blind Deconvolution. *Prentice Hall, Englewood Cliffs, NJ*, 1994.

[10] A.M. Kagan, Y.V. Linnik, and C.R. Rao. *Characterization Problems in Mathematical Statistics.* Wiley, 1973.

[11] G.K. Kaleh and R. Vallet. Joint parameter estimation and symbol detection for linear or non-linear unknown channels. In *IEEE Trans. Communications*, volume 42 (7), 1994.

[12] S.M. Kay. Fundamentals of statistical signal processing. *Signal Proc. series. Prentice Hall*, 1993.

[13] B. Leroux. Maximum likelihood estimation for hidden markov models. *Stoch. Processes and their Applications*, 40: 127–143, 1992.

[14] K.S. Lii and M. Rosenblatt. Deconvolution and estimation of transfer function phase and coefficients for non-gaussian linear processes. *Annals of Statistics*, 10: 1195–1208, 1982.

[15] J. Mendel. Optimal seismic deconvolution. *Academic Press*, 1983.

[16] E. Moulines, J.-F. Cardoso, and E. Gassiat. Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models. In *IEEE Int. Conf. Acoust., Speech, Signal Processing*, pages 3617–3620, 1997.

[17] C.L. Nikias and J.M. Mendel. Signal processing with higher-order spectra. *IEEE Signal Processing Magazine*, pages 10–37, July 1993.

[18] L.R. Rabiner. A tutorial on hidden markov models and selected applications in speech recognition. In *IEEE*, volume 77 (2), pages 257–285, February 1989.

[19] A. Redner and H.F. Walker. Mixture densities, maximum likelihood and the em algorithm. *SIAM Review*, 26-2: 195–239, 1984.

[20] M. Rosenblatt. Stationary sequences and random fields. *Birkhäuser, Boston*, 1985.

[21] T. Ryden. Consistent and asymptotically normal parameter estimates for hidden markov models. *Annals of Stat.*, 22: 1884–1895, 1994.

[22] N. Seshadri. Joint data and channel estimation using blind trellis search techniques. In *IEEE Trans. Communications*, volume 42 (2), pages 1606–1616, 1994.

[23] J.K. Tugnait. Blind equalization and estimation of digital communication FIRq channels using cumulant matching. *IEEE Trans. on Communications*, 43: 1240–1260, 1995.

[24] A. Wald. Note on the consistency of the maximum likelihood estimate. *Annals of Math. Statist.*, 20: 595–601, 1949.

Chapitre 6

Test de rapport de vraisemblance du nombre de composantes d'un mélange à régime Markovien

Cet article a été réalisé en collaboration avec E. Gassiat. Nous y étudions l'ordre d'une CMC, l'ordre étant son nombre d'états cachés. La question est d'estimer ou de tester cet ordre. Nous avons déjà parlé de la difficulté intrinsèque liée au modèle : la loi d'une observation conditionnellement à la chaîne cachée est une loi de mélange, et de ce fait hérite les problèmes de la non-identifiabilité, augmenté de la non dépendance des observations. Quand l'ordre est inconnu et fini, Leroux (1992b) a démontré la consistance de l'estimateur du maximum de vraisemblance et Ryden (1995) a montré que l'estimateur de pseudo-vraisemblance compensée ne sous-estimait pas l'ordre. A ordre connu, Bickel, Ritov et Ryden (1998) ont montré la normalité asymptotique de l'estimateur du maximum de vraisemblance . Considérant le test d'un modèle i.i.d. de loi donnée contre CMC à deux états, nous avons reparamétré le modèle en réutilisant les idées de Dacunha-Castelle et Gassiat (1997). Nous allons maintenant présenter le modèle étudié, et attirons l'attention sur le fait que les notations sont modifiées par rapport à ce que nous avions rencontré jusqu'à présent : l'ordre du modèle sera noté m, et q devient le paramètre (non identifiable sous l'hypothèse nulle) de la matrice de transition de la chaîne cachée.

(X_k) prend ses valeurs dans $\{\gamma_0, \gamma_1\}$, avec comme matrice de transition :

$$\Pi_{u,q} = \begin{pmatrix} 1 - \frac{qu}{1-u} & \frac{qu}{1-u} \\ q & 1-q \end{pmatrix}$$

et comme loi initiale la loi stationnaire :

$$P(X_0 = \gamma_0) = 1 - u, \quad P(X_0 = \gamma_1) = u$$

Les lois des observations conditionnelles à la chaîne sont deux densités f_{γ_0} and f_{γ_1} dans $\mathcal{F}$. Ici, $0 \leq q \leq 1$, $0 \leq u \leq 1$, et q et u sont aussi contraints par

$$0 \leq \frac{qu}{1-u} \leq 1.$$

Nous considérons donc le test de H_0 : "$u = 0$" contre H_1 : "$u \neq 0$". Pour tout η positif, soit :

$$\mathcal{Q}_\eta = \{q \in [0,1] \ : \ (1-q)^2(N^2+1) \leq 1 - \eta\} = [q_\eta, 1]$$

avec

$$q_\eta = 1 - \sqrt{\frac{1-\eta}{N^2+1}} > 0, \quad N^2 = \|\frac{f_{\gamma_1} - f_{\gamma_0}}{f_{\gamma_0}}\|_{L^2(f_{\gamma_0}\cdot\nu)},$$

et $\mathcal{P}_\eta$ l'ensemble de toutes les valeurs possibles pour les paramètres (u, q).

Soit $\ell_n(\Pi, \gamma_1, \ldots, \gamma_m; m)$, la log-vraisemblance des observations sous le paramètre $(\Pi, \gamma_1, \ldots, \gamma_m)$ où, rappelons le, m est le nombre d'états cachés. Définissons aussi :

$$T_n(m) = \sup\{\ell_n(\Pi, \gamma_1, \ldots, \gamma_m; m) \ : \ \Pi \text{ matrice stochastique } m \times m, (\gamma_1, \ldots, \gamma_m) \in \Gamma^m\}, \tag{1}$$

le maximum de la log-vraisemblance dans un modèle d'ordre m, et notons $(\hat{u}, \hat{q})$, l'estimateur du maximum de vraisemblance pour (u, q) et $\hat{u}_q$ l'estimateur du maximum de vraisemblance de u quand q est fixé.

- Lorsque les paramètres sont restreints à l'ensemble $\mathcal{P}_\eta$, la statistique du logarithme du rapport de vraisemblance sous H0

  $$W_n(\eta) = \sup\{\ell_n(u, q) - \ell_n(0, q) \ : \ (u, q) \in \mathcal{P}_\eta\}$$

  converge sous H_0 vers

  $$\frac{1}{2} \sup_{q \in \mathcal{Q}_\eta} \xi_q^2 \cdot 1_{\xi_q \geq 0},$$

  la moitié du supremum du carré de la partie positive du processus gaussien ξ_q de covariance

  $$R(q_1, q_2) = \frac{C(q_1, q_2)}{C(q_1, q_1)^{1/2} C(q_2, q_2)^{1/2}},$$

où
$$C(q_1, q_2) = \frac{1-(1-q_1)(1-q_2)}{1-(1-q_1)(1-q_2)(N^2+1)}.$$

- Quand η tend vers 0, la covariance d'un nombre infini de membres du processus gaussien tend vers 0, de telle sorte que $T_n(2) - T_n(1)$ tend vers l'infini.

Les idées importantes nous permettant d'arriver à ce résultat sont :

- Utiliser le filtre de prédiction $(p_k(u,q))_{k\geq 1}$:
$$p_k(u,q) = P(X_k = 1/Y_1, \ldots, Y_{k-1}).$$
et son équation récursive
$$p_{k+1}(u,q) = \frac{\frac{qu}{(1-u)}(1-p_k)f_{\gamma_0}(Y_k) + (1-q)p_k f_{\gamma_1}(Y_k)}{(1-p_k)f_{\gamma_0}(Y_k) + p_k f_{\gamma_1}(Y_k)}$$
- Exprimer la log-vraisemblance en fonction du filtre de prédiction :
$$nS_n(u,q) = \ell_n(u,q) - \ell_n(0,q) = \sum_{k=1}^{n} \log[1 + p_k(u,q)\frac{f_{\gamma_1} - f_{\gamma_0}}{f_{\gamma_0}}(Y_k)]$$
- Utiliser des propriété d'ergodicité des chaînes de Markov $(Y_k, p_k(u,q))_{k\in\mathbb{N}}$, $(Y_k,$ dérivées de $p_k(u,q))_{k\in\mathbb{N}}$. Ce point de vue a d'ailleurs déjà été utilisé par Mevel (1997) dans sa thèse pour démontrer les propriétés asymptotiques de l'estimateur du maximum de vraisemblance quand le nombre de populations m est connu.

La convergence en loi du logarithme du rapport de vraisemblance $W_n(\eta)$ sous H0 se passe en deux étapes :

- Montrer que $\hat{u}$ tend en probabilité vers 0. Pour ce faire, on remarque que la chaîne $(Y_k, p_k(u,q))_{k\in\mathbb{N}}$ est uniformément ergodique de probabilité invariante $\pi = f_{\gamma_0}\nu \otimes \mu_{u,q}$. Le processus de contraste $S_n(u,q)$ converge donc presque sûrement vers la fonction de contraste
$$S(u,q) = -\int K(f_{\gamma_0}, (1-p)f_{\gamma_0} + pf_{\gamma_1})d\mu_{u,q}(p),$$
dans laquelle $K(\cdot,\cdot)$ est l'information de Kullbak. De plus,
$$S(u,q) = 0 \Leftrightarrow u = 0.$$

Enfin un critère d'uniformité

$$\limsup_{n\to\infty} P\left(\sup_{|u-u'|\leq h, |q-q'|\leq h, (u,q)\in\mathcal{P}_\eta^\delta} |S_n(u,q) - S_n(u',q')| \geq \epsilon(\delta,\eta)\right) = 0$$

permet de conclure.

- Faire un développement de Taylor de la vraisemblance en 0, et en étudier tous les termes :
$$S_n(\hat{u}_q, q) = -\frac{1}{2}\frac{(\frac{\partial}{\partial u}S_n(0,q))^2}{\frac{\partial^2}{\partial u^2}S_n(0,q)} \cdot (1 + R_n(q))$$

Cette partie a trois étapes principales :

- La variable aléatoire
$$\frac{\partial^2}{\partial u^2}S_n(0,q)$$
converge en probabilité vers $N^2C(q,q)$ uniformément sur $\mathcal{Q}_\eta$. Pour cela, nous utiliserons une fonction de Lyapounov pour démontrer l'existence d'une probabilité invariante de la chaîne $\Phi_k = (W_k, \frac{\partial p_k}{\partial u}(0,q), \frac{\partial^2 p_k}{\partial u^2}(0,q))$. Puis, nous obtiendrons que cette chaîne possède la propriété de l'oubli de la condition initiale. Enfin, montrant que cette chaîne a un état atteignable, nous aurons une loi faible des grands nombres. Enfin, nous montrerons que la convergence est uniforme.

- Pour toute variable aléatoire $\tilde{u}$ convergeant vers 0 uniformément en q,
$$\frac{\partial^2}{\partial u^2}S_n(\tilde{u},q) - \frac{\partial^2}{\partial u^2}S_n(0,q)$$
converge en probabilité vers 0 uniformément sur $\mathcal{Q}_\eta$. Ainsi, $R_n(q) = o_P(1)$. La difficulté est de trouver les bonnes fonctions pour lesquelles on calcule facilement les récurrences sur les moments. Nous nous servirons alors extensivement du moyen suivant : si la suite réelle déterministe $(z_k)_{k\in\mathbb{N}}$ vérifie, pour un nombre positif δ, pour tout entier k
$$z_{k+1} \leq M + (1-\delta)z_k$$
et aussi $z_1 \leq M$, alors pour tout entier k,
$$z_k \leq Cte \cdot M,$$
Ceci sera particulièrement utilisé pour M fonction d'une variable $\overline{u}$, et tendant vers 0 avec $\overline{u}$.

- Enfin, le processus
$$\sqrt{n}\left(\frac{\partial}{\partial u}S_n(0,q)\right)_{q\in\mathcal{Q}_\eta}$$
converge uniformément vers le processus Gaussien de covariance $N^2C(.,.)$. Nous commencerons par voir que, pour tout q dans $\mathcal{Q}_\eta$, le processus
$$\left(\frac{\partial p_k}{\partial u}(0,q);W_k\right)_{k\in\mathbb{N}}$$
est une chaîne de Markov Harris positive, V-uniformément ergodique pour $V(x) = 1+\parallel x \parallel_2^{2+\epsilon_*}$. Puis nous obtiendrons la convergence des marginales finies dimensionnelles en vérifiant la condition de Lindeberg (voir Dacunha-Castelle et Duflo (1993)) que nous allons rappeler ci-dessous et la convergence de
$$\frac{1}{n}\sum_{k=1}^{n}\frac{\partial p_k}{\partial u}(0,q_i)\frac{\partial p_k}{\partial u}(0,q_j).$$
Enfin, nous montrerons que le processus
$$\left(\sqrt{n}\frac{\partial}{\partial u}S_n(0,q)\right)_{q\in\mathcal{Q}_\eta}$$
est uniformément tendu, en utilisant le critère de Kolmogorov pour les processus continus, qui sera démontré à l'aide de l'inégalité de Rosenthal pour les martingales.

Les principales étapes étant décrites, nous rappelons maintenant quelques théorèmes utilisés dans les preuves. Nous renvoyons en revanche le lecteur au livre de Meyn and Tweedie (1993) pour toutes les définitions, théorèmes et propriétés sur les chaînes de Markov.

Théorème 6.0.1 de limite centrale pour les martingales, (Dacunha-Castelle et Duflo (1993), Théorème 2.8.41).
Soit M une martingale vectorielle de carré intégrable de dimensions d et de crochet $< M >$. Soit Γ une matrice de covariance déterministe. On suppose que, pour une suite réelle déterministe (a_n) positive et croissant vers l'infini :

- *(i) $a_n^{-1} < M >_n$ tend vers Γ en probabilité.*

- *(ii) La "condition de Lindeberg" est satisfaite, c'est-à-dire, pour tout* $\varepsilon > 0$

$$\frac{1}{a_n}\sum_{k=1}^{n} E[\| M_k - M_{k-1} \|^2 \, 1_{[\|M_k - M_{k-1}\| \geq \varepsilon\sqrt{a_n}]}/\mathcal{F}_{k-1}]$$

tend en probabilité vers 0. Alors, $a_n^{-1} < M >_n$ *tend presque sûrement vers 0 et* $a_n^{-1/2} < M >_n$ *tend vers la loi normale* $\mathcal{N}(0,\Gamma)$.

Théorème 6.0.2 Inégalité de Rosenthal pour les martingales (Hall et Heyde (1980)).
Si $(M_k, \mathcal{F}_k, 1 \leq k \leq n)$ *est une martingale et* $2 \leq d < \infty$, *alors il existe une constante* K *dépendant seulement de* d *telle que :*

$$E \mid M_n \mid^d \leq K\Big(\cdot E\left[\left(\sum_{k=1}^{n} E((Z_k)^2/\mathcal{F}_{k-1})\right)^{d/2}\right] + \sum_{k=1}^{n} E \mid Z_k \mid^d \Big),$$

où $Z_1 = M_1$ *et* $Z_k = M_k - M_{k-1}$, $2 \leq k \leq n$.

Théorème 6.0.3 Oubli de la condition initiale (Duflo (1996), Théorème 1.IV.22)
Soit un modèle itératif féllérien sur $\mathbb{R}^d$. *On suppose que, pour tout couple* (x,y) *de points de* $\mathbb{R}^d$, $(X_n^x - X_n^y)$ *converge en loi vers* 0. *Alors, il existe au plus une probabilité invariante. S'il existe une probabilité invariante* μ, *quelque soit l'état initial,* (X_n) *converge en loi vers* μ.

Théorème 6.0.4 Critère de Lyapounov de stabilité (Duflo (1996), Théorème 1.IV.24)
Soit un modèle itératif markovien sur $\mathbb{R}^d$, $X_n = F(X_n, \eta_{n+1})$. *On suppose qu'il existe un* $a \geq 1$, *une constante* α, $0 \leq \alpha \leq 1$ *et une norme* $|.|$*sur* $\mathbb{R}^d$ *telles que :*

$$E(|F(0,\eta_1)|^a) < \infty,$$

$$E(|F(x,\eta_1) - |F(y,\eta_1)|^a)^{1/a} \leq \alpha|x-y|.$$

Alors, le modèle est stable de loi stationnaire μ. *Pour toute fonction* ϕ *de* $\mathbb{R}^d$ *dans* $\mathbb{R}$ *continue* μ*-p.s. et telle que*

$$|\phi(x)| \leq \; Cte(|x|^a + 1),$$

et toute loi initiale, on a la loi des grands nombres :

$$\frac{1}{n}\sum_{k=0}^{n} \phi(X_k) \overset{p.s.}{\to} \mu(\phi).$$

The likelihood ratio test for the number of components in a mixture with Markov regime

Elisabeth Gassiat
Université d'Orsay, France[1]

Christine Keribin
Université d'Evry-Val d'Essonne, France [2]

Abstract: We study the LRT statistic for testing a single population i.i.d. model against a mixture of two populations with Markov regime. We prove that the LRT statistic converges to infinity in probability as the number of observations tends to infinity. This is a consequence of a convergence result of the LRT statistic for a subproblem where the parameters are restricted to a subset of the whole parameter set.

6.1 Introduction

Hidden Markov Models have been used for a long time as a useful modeling tool. Applications include speech processing (Rabiner, 1989), neurophysiology (Fredkin, Rice, 1992), biology (Churchill, 1989), see also Mac Donald and Zucchini (1997). A hidden Markov model (HMM) is a discrete time stochastic process $(Y_k)_{k\in\mathbb{N}}$ such that the distribution of the process may be well explained through a non observable Markov chain $(X_k)_{k\in\mathbb{N}}$ taking values in a finite state space. Given (X_k), (Y_k) is a sequence of independent random variables, and the distribution of Y_n given (X_k) is that of Y_n given X_n. Thus, the one dimensional marginal of the process, namely the distribution of Y_n, appears to be a mixture of populations where the number of populations m is the cardinality of the state space of the hidden Markov chain. The observations may then be said to follow a "mixture model with Markov regime".

[1](1) Laboratoire Modélisation Stochastique et Statistique, Université d'Orsay, Bat. 425, 91 405 Orsay, France

[2](2) Laboratoire Analyse et Probabilité, Université d'Evry-Val d'Essonne, Boulevard des Coquibus, 91025 Evry

[3]*AMS 1991 subject classifications.* 62 F 05, 62 H 30, 62 A 10 .

[4]*Key words and phrases.* Likelihood Ratio Test; Hidden Markov Model; Order of a Mixture.

Asymptotic properties of maximum likelihood estimators (m.l.e.) for the estimation of the parameters in a HMM have received early interest when the (Y_k) take values in a finite set (Baum, Petrie, 1966), and recent interest in the general case (Leroux, 1992; Ryden, 1995; Bickel and Ritov, 1996; Bickel, Ritov and Ryden, 1998; Mevel, 1997; Vandekerkhove, 1997). Various algorithms have been proposed to achieve the m.l.e., see for instance Mevel (1997) or Vandekerkhove (1997).

The problem of estimating m, the number of populations, or of testing between different values of m, is an important one either for numerical features (a larger value than the true one leads to flat likelihoods) or for its interpretation as a meaningful parameter. Earlier works on the subject propose to use likelihood methods (Leroux, 1992; Ryden, 1995; Finesso, 1990), but fail to prove the consistency of the estimator (or the asymptotic level of the test) because of an intrinsic difficulty of the model: the usual regularity conditions are not fulfilled under the null hypothesis, so that the χ-square theory may not apply. Corresponding problem for mixtures with i.i.d. regime has a long history (see Lindsay, 1995) but the asymptotic distribution of the LRT (likelihood ratio test) has been recently solved by Dacunha-Castelle and Gassiat (1999, 1997b) and the consistency of the penalized likelihood estimator has been proved by Keribin (1998).

The aim of this paper is the study of the LRT for testing the number of populations in a mixture with Markov regime. Intuition and methods of Dacunha-Castelle and Gassiat (1999, 1997b) may be applied, and it appears that the computation of the asymptotic distribution in the general setting is much more complicated than in the i.i.d. regime situation. However, by applying the method to a subproblem of the general problem, we obtain meaningful results. Let us now be more precise.

Let

$$\mathcal{F} = \{f_\gamma, \gamma \in \Gamma\}$$

be a set of probability densities with respect to a given measure ν on a metric space, where Γ is a compact subset of some Euclidean space. Let $(X_k)_{k\in\mathbb{N}}$ be a stationary Markov chain taking values in a set of m distinct values of Γ, $\{\gamma_1, \dots, \gamma_m\}$, with transition matrix Π. The observations are $(Y_1, \dots, Y_n)$. Conditionally to $(X_1, \dots, X_n)$, $(Y_1, \dots, Y_n)$ are independent variables, each Y_i having conditional distribution $f_{X_i} \cdot \nu$.

Define

$$T_n(m) = \sup\{\ell_n(\Pi, \gamma_1, \ldots, \gamma_m; m) \ : \ \Pi \ m \times m \text{ stochastic matrix}, (\gamma_1, \ldots, \gamma_m) \in \Gamma^m\}, \tag{2}$$

in which $\ell_n(\Pi, \gamma_1, \ldots, \gamma_m; m)$ is the log-likelihood of the observations under the parameter $(\Pi, \gamma_1, \ldots, \gamma_m)$ when the number of hidden states is m. One of our main results is that the LRT statistic $T_n(2) - T_n(1)$ for testing i.i.d. observations of one population against 2 populations with Markov regime converges in probability to infinity under the null hypothesis (see Theorem 6.2.5). This result is a consequence of another convergence result in the sub-problem of testing a fixed contamination with Markov regime. Let f_{γ_0} and f_{γ_1} be fixed densities in $\mathcal{F}$, $\gamma_0 \neq \gamma_1$. Let $(X_k)_{k \in \mathbb{N}}$ take values in $\{\gamma_0, \ \gamma_1\}$, with initial distribution

$$P(X_0 = \gamma_0) = 1 - u, \ \ P(X_0 = \gamma_1) = u$$

and transition matrix

$$\Pi_{u,q} = \begin{pmatrix} 1 - \frac{qu}{1-u} & \frac{qu}{1-u} \\ q & 1 - q \end{pmatrix} \tag{3}$$

Here, $0 \leq q \leq 1$, $0 \leq u \leq 1$, and q and u are also constrained by

$$0 \leq \frac{qu}{1-u} \leq 1.$$

We want to test that the Y_i's are i.i.d. with common density f_{γ_0}, against that $(Y_n)_{n \in \mathbb{N}}$ follows a HMM for some parameters u, q, and with fixed γ_0 and γ_1. Obviously, the HMM reduces to i.i.d. variables with common density f_{γ_0} if and only if $u = 0$, so that the test we consider is that of $H_0 :$ "$u = 0$" against $H_1 :$ "$u \neq 0$". Let H be the Hilbert space $L^2(f_{\gamma_0} \cdot \nu)$, and

$$N^2 = \| \frac{f_{\gamma_1} - f_{\gamma_0}}{f_{\gamma_0}} \|_H^2.$$

For any $\eta \in]0, 1[$, let

$$\mathcal{Q}_\eta = \{q \in [0, 1] \ : \ (1 - q)^2 (N^2 + 1) \leq 1 - \eta\} = [q_\eta, 1] \tag{4}$$

with

$$q_\eta = 1 - \sqrt{\frac{1 - \eta}{N^2 + 1}} > 0.$$

When the q's are restricted to belong to $\mathcal{Q}_\eta$, let $\mathcal{P}_\eta$ be the parameter set of possible (u, q). The log-likelihood $\ell_n(\Pi_{u,q}, \gamma_0, \gamma_1; 2)$ will now be denoted by $\ell_n(u, q)$. Notice that under the null hypothesis, the parameter q is undefined, the log-likelihood does not depend on q, and is $\ell_n(0,q)$ for any q. The LRT statistic is now, when the q's are restricted to belong to $\mathcal{Q}_\eta$:

$$V_n(\eta) = \sup\{\ell_n(u,q) - \ell_n(0,q) \ : \ (u,q) \in \mathcal{P}_\eta\}. \qquad (5)$$

The first main convergence result is that $V_n(\eta)$ converges to half the supremum of the square of the positive part of some gaussian process (see Theorem 6.2.1). This result is similar to that obtained for testing the number of populations in a mixture, or for testing the order of a stationary ARMA process, see Dacunha-Castelle and Gassiat (1999, 1997b). As η goes to 0, the covariance of infinitely many members of the process goes to 0, and the convergence to infinity of $T_n(2) - T_n(1)$ is thus an easy consequence.
Convergence to infinity of LRT's has been observed earlier in two papers about mixtures with i.i.d. regime: when the parameter set is unbounded, see Hartigan (1985), and when the family $\mathcal{F}$ is not regular, two situations in which the set of scores may be proved to be non compact, see Ciuperca (1999). Here, the parameter set is compact, and the family $\mathcal{F}$ is regular. Our result may be compared to a similar result, not proved in any paper but which is also an easy consequence of Theorem 4.5 in Dacunha-Castelle and Gassiat (1999), that the LRT for testing the order of an ARMA model tends to infinity when the poles and roots are not restricted to a compact subset of the open unit complex disk.
Our result leaves open the problem of the choice of the penalization rate when estimating the number of populations using penalized likelihood estimators, for mixtures with Markov regime. However, for translation or scale mixtures, the method proposed by Dacunha-Castelle and Gassiat (1997a) may be applied to mixtures with Markov regime to obtain a consistent estimate of the number of populations.

The paper is organized as follows. In section 2, we state the assumptions and our main results. Section 3 is devoted to numerical experiments, and section 4 to the proofs of intermediate results.

6.2 Main results

We study the asymptotic properties of the LRT under the null hypothesis, in which the Y_i's are assumed to be i.i.d. with distribution $f_{\gamma_0}\nu$. Let us introduce the assumptions:

- (A1) For $\gamma \in \{\gamma_0, \gamma_1\}$, under H_0, $E(|\log f_\gamma(Y_1)|) < +\infty$.

- (A2) Under H_0, the support of the random variable $\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y_1)$ is $\mathbb{R}^+$.
- (A3) Under H_0, $E(\frac{f_{\gamma_0}}{f_{\gamma_1}}(Y_1)) < +\infty$.
- (A4) Under H_0, $E\left(\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y_1)\right)^3 < +\infty$.

Let now $(\xi_q)_{q\in\mathcal{Q}_\eta}$ be the continuous gaussian process with covariance

$$R(q_1,q_2) = \frac{C(q_1,q_2)}{C(q_1,q_1)^{1/2}C(q_2,q_2)^{1/2}} \tag{6}$$

with

$$C(q_1,q_2) = \frac{1-(1-q_1)(1-q_2)}{1-(1-q_1)(1-q_2)(N^2+1)}, \tag{7}$$

Then

Theorem 6.2.1 *Assume that (A1), (A2), (A3) and (A4) hold. Under H_0, $V_n(\eta)$ converges in distribution to*

$$\frac{1}{2}\sup_{q\in\mathcal{Q}_\eta} \xi_q^2 \cdot 1_{\xi_q\geq 0}.$$

Proof of Theorem 6.2.1.

Theorem 6.2.1 will be proved by investigating the asymptotic behavior of the maximum of the likelihood for fixed q as a process with $\mathcal{Q}_\eta$ the time set, and then taking the maximum for q over $\mathcal{Q}_\eta$. For an HMM with parameters $\Pi_{u,q}$, γ_0 and γ_1, let $(p_k(u,q))_{k\geq 1}$ be the prediction filter, that is, for any positive k:

$$p_k(u,q) = P(X_k = \gamma_1/Y_1,\dots,Y_{k-1}). \tag{8}$$

It is well known that the $p_k(u,q)$'s satisfy the recursive equation:

$$p_{k+1}(u,q) = \frac{\frac{qu}{(1-u)}(1-p_k)f_{\gamma_0}(Y_k) + (1-q)p_k f_{\gamma_1}(Y_k)}{(1-p_k)f_{\gamma_0}(Y_k) + p_k f_{\gamma_1}(Y_k)} \tag{9}$$

with initialization $p_1(u,q) = u$. Parameter (u,q) is omitted for p_k on the right hand of the recurrence equation for readability. This will be the case also in further recurrence equations concerning the derivatives of the prediction filter. Now the log-likelihood may be written:

$$\ell_n(u,q) - \ell_n(0,q) = \sum_{k=1}^n \log[1 + p_k(u,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k)] \tag{10}$$

Define

$$S_n(u,q) = \frac{1}{n}[\ell_n(u,q) - \ell_n(0,q)].$$

Asymptotic properties of $S_n(u,q)$ and its derivatives will be investigated using the properties of the Markov chains $(Y_k, p_k(u,q))_{k\in\mathbb{N}}$, $(Y_k,$ derivatives of $p_k(u,q))_{k\in\mathbb{N}}$. This point of view is the one adopted by Mevel (1997), Mevel and LeGland (1999) to obtain the asymptotic properties of the maximum likelihood estimator (m.l.e.) of the parameters when the number of populations m is known.
Let now $(\widehat{u},\widehat{q})$ be the m.l.e. for (u,q), and $\widehat{u}_q$ the m.l.e. for u when q is fixed. We have:

Proposition 6.2.2 *Assume (A1), (A2) (A3) and (A4). Then* $\sup_{q\in\mathcal{Q}_\eta}\widehat{u}_q$ *and* $\widehat{u}$ *converge to* 0 *in probability under* H_0.

Remark: This result could be derived from Theorem 3 of Leroux (1992). However, we give here an alternative specific proof which shows the effect of overestimating m.
The limiting contrast function of $S_n(u,q)$ will be proved to be

$$S(u,q) = -\int K(f_{\gamma_0}, (1-p)f_{\gamma_0} + pf_{\gamma_1})d\mu_{u,q}(p), \tag{11}$$

in which $K(\cdot,\cdot)$ is the Kullbak information (see for instance Dacunha-Castelle and Duflo, 1993, for a definition) and $\mu_{u,q}$ is a probability distribution on $[0,1]$ such that, for $u=0$, $\mu_{0,q}$ is the Dirac at 0. Notice that $S(u,q)$ is always finite by (A1), and, since the Kullback information is always non negative, for any (u,q), $S(u,q)\leq 0$, and $S(0,q)=0$. Proposition 6.2.2 will follow from the fact that $S(u,q)=0$ only for $u=0$, and that the convergence of $S_n(u,q)$ to $S(u,q)$ is uniform enough, see section 4. We have:

$$V_n(\eta) = \sup_{q\in\mathcal{Q}_\eta} nS_n(\hat{u}_q, q).$$

But using Proposition 6.2.2, for any q in $\mathcal{Q}_\eta$, Taylor expansion leads to

$$S_n(\widehat{u}_q, q) = -\frac{1}{2}\frac{(\frac{\partial}{\partial u}S_n(0,q))^2}{\frac{\partial^2}{\partial u^2}S_n(0,q)}\cdot(1+R_n(q)) \tag{12}$$

in which $R_n(q)$ converges in probability to 0 uniformly in q as soon as

$$\frac{\frac{\partial^2}{\partial u^2}S_n(\tilde{u},q)}{\frac{\partial^2}{\partial u^2}S_n(0,q)}$$

converges to 1 uniformly in q for any $\tilde{u}$ converging in probability to 0 uniformly in q. Now,

$$\frac{\partial^2}{\partial u^2}S_n(u,q) = -\frac{1}{n}\sum_{i=1}^{n}\frac{1}{(1+p_k(u,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k))^2}(\frac{\partial p_k}{\partial u}(u,q))^2(\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k))^2$$
$$+\frac{1}{n}\sum_{i=1}^{n}\frac{1}{1+p_k(u,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k)}\frac{\partial^2 p_k}{\partial u^2}(u,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k). \quad (13)$$

In particular, for $u = 0$, one obtains

$$\frac{\partial^2}{\partial u^2}S_n(0,q) = -\frac{1}{n}\sum_{i=1}^{n}(\frac{\partial p_k}{\partial u}(0,q))^2(\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k))^2+\frac{1}{n}\sum_{i=1}^{n}\frac{\partial^2 p_k}{\partial u^2}(0,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k). \quad (14)$$

In this sum, the second term is centered, so that as soon as laws of large numbers apply, the limiting value is $-N^2$ times the expectation of $(\frac{\partial p_k}{\partial u}(0,q))^2$ under the stationary distribution. This convergence result holds more generally and we have:

Proposition 6.2.3 *Assume that (A2), (A3) and (A4) hold. Then, for any random $\tilde{u}$ converging in probability to 0 uniformly in q,*

$$\frac{\partial^2}{\partial u^2}S_n(\tilde{u},q)$$

converges in probability to $-N^2C(q,q)$ uniformly over $\mathcal{Q}_\eta$.

Proposition 6.2.3 leads to

$$V_n(\eta) = \frac{1}{2}\left[\sup_{q\in\mathcal{Q}_\eta}\frac{(\sqrt{n}\frac{\partial}{\partial u}S_n(0,q))^2}{N^2C(q,q)}\right](1+o_P(1)) \quad (15)$$

and Theorem 6.2.1 will follow from

Proposition 6.2.4 *Assume (A1), (A2) and (A3). Then, the process*

$$\sqrt{n}\left(\frac{\partial}{\partial u}S_n(0,q)\right)_{q\in\mathcal{Q}_\eta}$$

converges uniformly to the continuous gaussian process with covariance $N^2C(\cdot,\cdot)$.

Comments.
The score process at $u = 0$ and for fixed q is:

$$\frac{\partial}{\partial u}S_n(0,q) = \frac{1}{n}\sum_{k=1}^{n}\frac{\partial p_k}{\partial u}(0,q)\frac{f_{\gamma_1}-f_{\gamma_0}}{f_{\gamma_0}}(Y_k)$$

in which $(\frac{\partial p_k}{\partial u}(0,q))_{k\in\mathbb{N}}$ is a Markov chain satisfying the recursion:

$$\frac{\partial p_{k+1}}{\partial u}(0,q) = q + (1-q)\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y_k)\frac{\partial p_k}{\partial u}(0,q) \tag{16}$$

with initialization $\frac{\partial p_1}{\partial u}(0,q) = 1$. But in case an invariant distribution exists and has a second moment σ^2, (16) leads to:

$$\sigma^2 = q^2 + 2q(1-q) + (1-q)^2(N^2+1)\sigma^2.$$

Since for any q, $q^2 + 2q(1-q) \geq 0$, this equation leads to a non negative σ^2 if and only if $1-(1-q)^2(N^2+1) \geq 0$. This last condition is thus a necessary one to obtain a convergence in distribution of $\sqrt{n}\frac{\partial}{\partial u}S_n(0,q)$ and a law of large numbers for $\frac{\partial^2}{\partial u^2}S_n(0,q)$. This is why the parameter set is cut in two different regions, the union of all $\mathcal{P}_\eta$, and its complementary set. Our conjecture is that, on this complementary set, $nS_n(\hat{u}_q,q)$ converges to half the square of a standard gaussian white noise.

To state the unboundedness of the LRT statistic for testing between one or two populations in a mixture with Markov regime, we introduce the following assumption:

- **(I)** $T_n(1) - \ell_n(0,q)$ is uniformly tight under H_0.

Notice that a sufficient condition for (I) to hold is that the family $\mathcal{F}$ is regular enough. For instance, differentiability in quadratic mean with respect to γ and existence and inversibility of the Fisher information suffice since in this case $T_n(1) - \ell_n(0,q)$ converges to a $\chi^2(1)$ (see for instance Van der Vaart, 1999). We now obtain the unconstrained result:

Theorem 6.2.5 *Assume (I), and that (A1), (A2), (A3) and (A4) hold for any γ_0 and γ_1 in Γ. Then, under H_0, $T_n(2) - T_n(1)$ converges to infinity in probability.*

Remark.
The same result holds if the parameter space is restricted to all possible (u,q) such that

$$(1-q)^2(N^2+1) \leq 1,$$

as may be seen from the proof. This means that the unboundedness of the LRT statistic is not due to the fact that the model contains non ergodic hidden Markov chains (by taking $q = 0$), but to the fact that the alternative may be too far from the null hypothesis, also with respect to the distance between the populations involved in the mixture.

Proof of Theorem 6.2.5.

Let $f_{\gamma_0}\nu$ be the true (unknown) distribution of the i.i.d. variables (Y_k). Let γ_1 be a point in Γ distinct of γ_0.

Since the LRT statistic for i.i.d. observations with the model $\mathcal{F}$ is tight, $T_n(2)-T_n(1)$ converges to infinity in probability as soon as $V_n(0)$ does.

Let $\rho < 1$, and define for all integer k:

$$r_k = \rho^{k^2}.$$

Define also $q_k \in [0, 1]$ by

$$(1 - q_k)^2(N^2 + 1) = 1 - r_k,$$

Obviously,

$$V_n(0) \geq \sup\{\ell_n(\widehat{u_{q_k}}, q_k) - \ell_n(0, q_k) \ : \ k \in \mathbb{N}\}. \tag{17}$$

Now, for the particular sequence $r_k = \rho^{k^2}$, observe that

$$\lim_{k \neq l \to +\infty} R(q_k, q_l) = 0,$$

so that

$$\sup_{k \leq L} \xi_{q_k}^2 \cdot 1_{\xi_{q_k} \geq 0}$$

converges in probability to $+\infty$ as L tends to $+\infty$. For any positive M, for any positive ϵ, there exists L_0 such that

$$P(\sup_{k \leq L_0} \xi_{q_k}^2 \cdot 1_{\xi_{q_k} \geq 0} > M) \geq 1 - \epsilon.$$

But using portmanteau's Lemma, the fact that $\sup\{\ell_n(\widehat{u_{q_k}}, q_k) - \ell_n(0, q_k) \ : \ k \leq L_0\}$ converges in distribution to $\sup_{k \leq L_0} \xi_{q_k}^2 \cdot 1_{\xi_{q_k} \geq 0}$, and inequality (17), we have that

$$\liminf_{n \to +\infty} P(V_n(0) > M) \geq 1 - \epsilon.$$

Since this holds for any ϵ, $V_n(0)$ tends to $+\infty$ in probability.

6.3 Numerical experiments

In this section, we numerically investigate the behavior of the LRT statistic. The true model is an i.i.d. process with gaussian distribution $\mathcal{N}(0,1)$ of mean 0 and variance 1. We test i.i.d. observations of one population with distribution $\mathcal{N}(0,1)$ against two populations with Markov regime, the distributions of the observations conditionally to the Markov chain being $\mathcal{N}(0,1)$ and $\mathcal{N}(1,1)$. In other words $f_{\gamma_0}(x) = (2\pi)^{-1/2} \exp -\frac{x^2}{2}$ and $f_{\gamma_1}(x) = (2\pi)^{-1/2} \exp -\frac{(x-1)^2}{2}$.

We first draw the likelihood map, then we study the convergence in distribution of $V_n(\eta)$ (Theorem (6.2.1)) and we finally experiment the convergence to infinity of the LRT statistic (Theorem (6.2.5)).

6.3.1 Likelihood map

We compute the log-likelihood ratio (LLR)

$$\ell_n(u,q) - \ell_n(0,q) = \sum_{k=1}^{n} \log[1 + p_k(u,q)\frac{f_{\gamma_1} - f_{\gamma_0}}{f_{\gamma_0}}(Y_k)],$$

using the recursive equation of the prediction filter (9) for $(q,u) \in [0,1] \times [0,1]$, such that $qu/(1-u) \leq 1$, by step of 0.01. The typical aspect of all our LLR maps is the following: we observe the LLR to fast decrease with u, and refine the map for u near 0. Two cases happen:

- Case A: the maximum is 0, reached for $\hat{u} = 0$, and the LLR is an always decreasing function of u (Fig. 6.1),
- Case B: the maximum is stricly positive and is reached for $\hat{u}$ close to 0 (Fig. 6.2).

When the number of observations increases, $\hat{u}$ gets closer to 0. For 10000 observations, $\hat{u}$ is of the order of 0.01, for 50000 observations, $\hat{u}$ is of the order of 0.005, and for 100000 observations, $\hat{u}$ is of the order of 0.001.

As the LLR can take large negative values, we cut off the values larger than -3 and replaced them by -3 to allow the visualization of the maximum. In the same way, we represent the points of the domain where $qu/(1-u) \geq 1$ with the value -4.

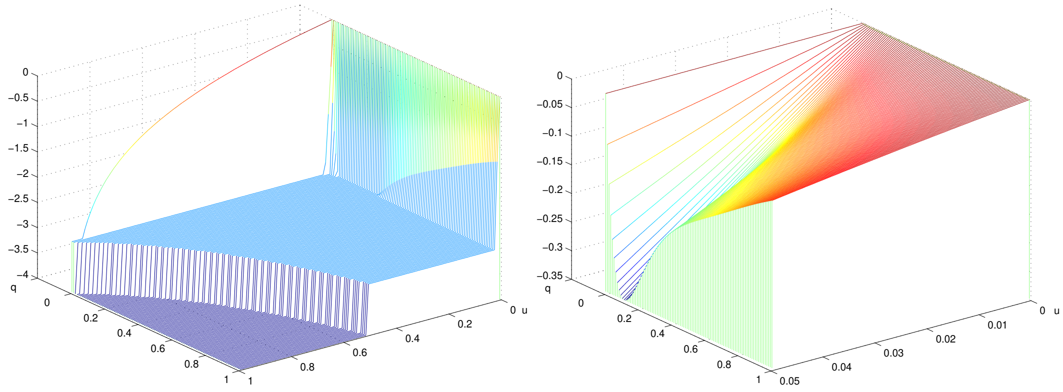

Figure 6.1 : *Case A: On left: on the whole domain* $(q, u) \in [0, 1] \times [0, 1]$ *by steps of* 0.01. *On right: refinement* $(q, u) \in [0, 1] \times [0, 0.05]$ *by steps of* 5.10^{-4} *for* u *and* 0.01 *for* q

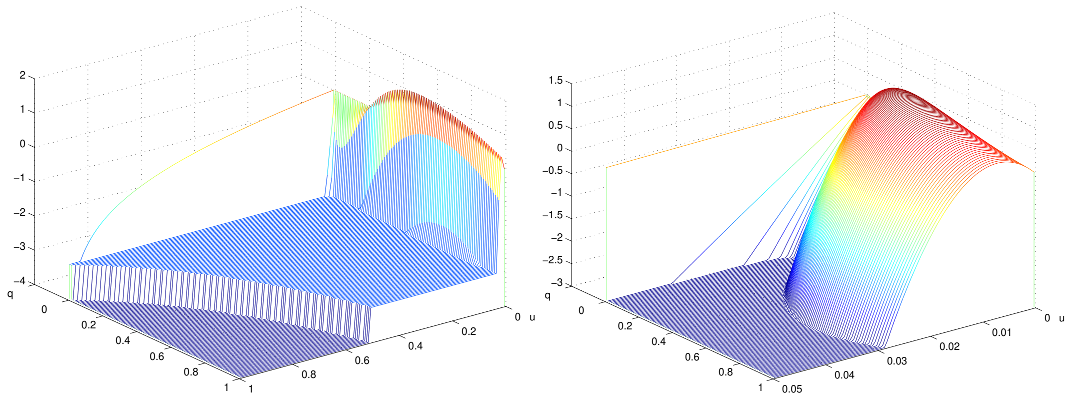

Figure 6.2 : *Case B. On left: on the whole domain* $(q, u) \in [0, 1] \times [0, 1]$ *by steps of* 0.01. *On right: refinement* $(q, u) \in [0, 1] \times [0, 0.05]$ *by steps of* 5.10^{-4} *for* u *and* 0.01 *for* q

6.3.2 Distribution of $V_n(\eta)$

The aim of this section is to study the rate in the distributional convergence of the LRT statistic on $\mathcal{P}_\eta$ to the asymptotic distribution given in Theorem (6.2.1):

$$\frac{1}{2} \sup_{q \in \mathcal{Q}_\eta} \xi_q^2 \cdot 1_{\xi_q \geq 0},$$

to be able to say for how many observations the asymptotic approximation holds. We encounter difficulties to have a significant result on the whole $\mathcal{P}_\eta$ domain, so we do the simulations on a simplified model, where the supremum is taken over two values of q.

$$\frac{1}{2} \sup_{q \in \{1, 0.4\}} \xi_q^2 \cdot 1_{\xi_q \geq 0}.$$

In this case, we can easily simulate the theoretical law, using the following property: if X_1 and X_2 are independent normal deviates $\mathcal{N}(0,1)$ and if

$$Y_1 = \sigma_1 X_1, \quad Y_2 = \sigma_2(\rho X_1 + \sqrt{1-\rho^2} X_2),$$

then Y_1 and Y_2 are dependent random variables, normally distributed with standard deviations σ_1 and σ_2, and with correlation coefficient ρ.

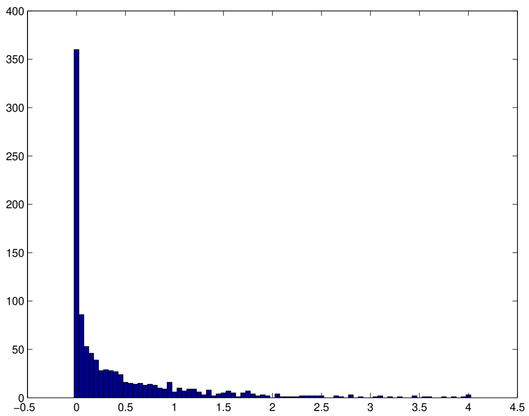

Figure 6.3 : *Histogram of the simulated law* $\frac{1}{2} \sup_{q \in \{1, 0.4\}} \xi_q^2 \cdot 1_{\xi_q \geq 0}$. *with* 1000 *points*

Fig. 6.3 shows the histogram of the simulated law with 1000 points, with $\rho = 0.18$ corresponding to the correlation when $q = 1$ and $q = 4$. The outline does not significantly change when we increase the number of points. Notice that the weight for $[0, 0.05]$ is around 370. For two independent normal deviates, this weight is around 340, whereas for $\rho = 1$, this weight is around 620.

To approximate the distribution of the LRT statistic $V_n(0.4, 1)$ in this case, we draw m samples of length n, and compute for each sample the maximum likelihood ratio value for the two values of q. For this computation, we use a Newton algorithm, using the fact, checked on a great number of samples, that the log-likelihood has an unique maximum near $u = 0$. Other simulations with an EM algorithm (Dempster *and al.* (1977)) and various initial points did not give better results. We first take $m = 500$ and $n = 2,000$, $n = 5,000$, $n = 10,000$ (Fig. 6.4 left). The weight of $[0, 0.05]$ is around 250. Compared with the simulated law $370/2 = 185$, it is much more higher. If we now increase the number of samples to $m = 5,000$ of size $n = 10,000$ (Fig. 6.4 right), this weight $2450/10 = 245$ is not significantly modified.

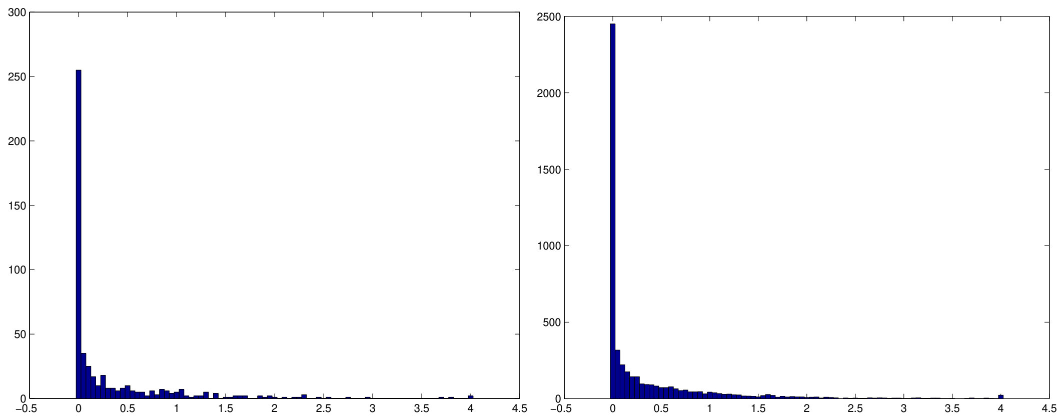

Figure 6.4 : *Histogram of the simulated LRT statistic* $V_n(0.4, 1)$*,* $m = 500$ *(on left) and* $m = 5000$ *(on right) samples of size* $n = 10,000$

If we increase the size of the samples to $n = 100,000$ for 500 samples (Fig. 6.5 left), this weight begins to decrease (around 230), and this is more significant for $m = 50$ samples of size $1,000,000$ (Fig. 6.5 right), where the weight is of the order of $20 \times 10 = 200$.

This allows us to conclude that the convergence is extremely slow, and we can relate this fact to the slow convergence of the LRT statistic for a mixture of two normal distributions as pointed out by Atwood *and al.* (1996) or Chuang and Mendell (1997). Bootstrap techniques might also be another way to investigate the distribution, see McLachlan (1987) or Feng and McCulloch (1996) for the case of i.i.d mixtures.

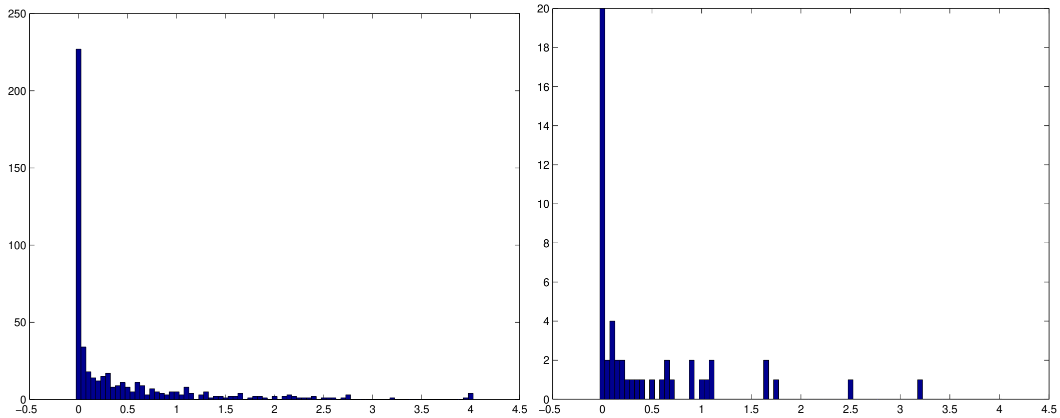

Figure 6.5 : *Histogram of the simulated LRT statistic* $V_n(0.4, 1)$*,* $m = 500$ *samples of size* $n = 100,000$ *(on left) and* $m = 50$ *samples of size* $n = 1,000,000$ *(on right)*

6.3.3 Convergence of the LRT statistic to infinity

At which rate does the LRT on $\mathcal{P}_0$ converge to infinity? We numerically investigate the evolution of the LRT with the length n of the sample.

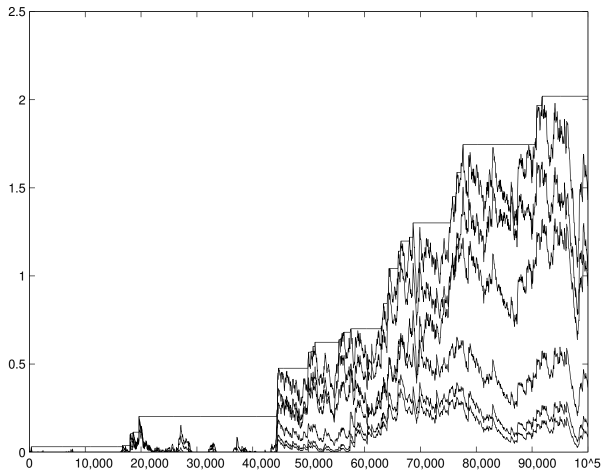

Figure 6.6 : $V_n^{(1)}(q_k)$ *for the six values of* q_k *defined by (18) with* $\rho = 0.85$*,* $k = 0, \ldots, 5$*. The sample size increases from size* $n = 50$ *to* $n = 100,000$ *by steps of* 50

As pointed out in the proof of Theorem (6.2.5), it is sufficient to take a sequence q_k such that

$$(1 - q_k)^2 (N^2 + 1) = 1 - \rho^{k^2}, \tag{18}$$

and observe that

$$\sup\{V_n^{(1)}(q_k)\ :\ k \in \mathbb{N}\}$$

tends to infinity with n, where

$$V_n^{(1)}(q_k) = \ell_n(\hat{u}_{q_k}, q_k) - \ell_n(0, q_k).$$

We choose $\rho = 0.85$, $k = 0, \ldots, 5$, and compute $V_n^{(1)}(q_k)$ from $n = 50$ to $n = 100,000$ by steps of 50 (Fig. 6.6). The result shows that the convergence is very slow, but can be seen. For the i.i.d. mixture of two normals, Hartigan (1985) gave a $\frac{1}{2}\log\log(n)$ rate. Such rate cannot be experimentally derived, it has to be theoretically investigated for mixtures with Markov regime in a further work.

6.4 Proofs.

6.4.1 Proof of Proposition 6.2.2.

Proposition 6.2.2 is a consequence of the following Lemmas.

Lemma 6.4.1 *Assume (A2). Let (u, q) be a fixed parameter. Then under H_0, $(Y_k, p_k(u, q))_{k\in\mathbb{N}}$ is a uniformly ergodic chain with invariant probability $\pi = f_{\gamma_0}\nu \otimes \mu_{u,q}$. For $u = 0$, $\mu_{0,q}$ is the dirac at 0, and for $u \neq 0$,*

$$\int p d\mu_{u,q} \geq \frac{qu}{2(1-u)}. \tag{19}$$

Proof of Lemma 6.4.1. Under H_0, $(Y_k, p_k(u, q))_{k\in\mathbb{N}}$ is a Markov chain since the (Y_k) are i.i.d. and by equation (9). Moreover, for any k, Y_k and $p_k(u, q)$ are independent variables so that the Lemma follows as soon as $(p_k(u, q))_{k\in\mathbb{N}}$ is a uniformly ergodic chain with invariant probability $\mu_{u,q}$. For $u = 0$, $p_k(0, q) = 0$ for all k. For $u \neq 0$, $(p_k(u, q))_{k\in\mathbb{N}}$ is a Feller chain, ψ-irreductible with ψ the Lebesgue measure on $(\frac{qu}{1-u}, 1-q)$ thanks to (A2), thus a T-chain by Theorem 6.2.9, and then a uniformly ergodic chain by Theorem 16.2.5 of Meyn and Tweedie (1993). Now using equation (9) one obtains, using the fact that $p_k(u, q)$ is $\sigma(Y_1, \ldots, Y_{k-1})$-measurable for all k, independent of Y_k, and using Jensen's inequality applied to the function $1/x$:

$$E(p_{k+1}(u, q)/Y_1, \ldots, Y_{k-1}) \geq \frac{qu}{(1-u)}(1 - p_k(u, q))$$

so that

$$E(p_{k+1}(u,q)) \geq \frac{qu}{(1-u)}(1 - E(p_k(u,q)))$$

and equation (19) follows by taking the limit as k goes to infinity.

Lemma 6.4.2 *Assume (A1) and (A2). Then under H_0, $S_n(u,q)$ converges a.s. to $S(u,q)$. Moreover, for any positive δ,*

$$\sup_{u\geq\delta,(u,q)\in\mathcal{P}_\eta} S(u,q) = -2\epsilon(\delta,\eta) < 0$$

Proof of Lemma 6.4.2. The a.s. convergence of $S_n(u,q)$ to $S(u,q)$ follows from the first point of Lemma 6.4.1. Now,

$$S(u,q) = \int f_{\gamma_0}(x)d\nu(x)[\int \log(1 - p + p\frac{f_{\gamma_1}}{f_{\gamma_0}}(x))d\mu_{u,q}(p)]$$

and by Jensen's inequality

$$\int \log(1-p+p\frac{f_{\gamma_1}}{f_{\gamma_0}}(x))d\mu_{u,q}(p) \leq \log(1 - \int pd\mu_{u,q}(p) + \int pd\mu_{u,q}(p)\frac{f_{\gamma_1}}{f_{\gamma_0}}(x))$$

so that

$$S(u,q) \leq -K(f_{\gamma_0}, (1 - \int pd\mu_{u,q}(p))f_{\gamma_0} + \int pd\mu_{u,q}(p)f_{\gamma_1}).$$

Using now the second point of Lemma 6.4.1 one obtains

$$\sup_{u\geq\delta,(u,q)\in\mathcal{P}_\eta} S(u,q) \leq -\inf\{K(f_{\gamma_0},(1-p)f_{\gamma_0}+pf_{\gamma_1}) \ : \ u\geq\delta,(u,q)\in\mathcal{P}_\eta, p \geq \frac{qu}{2(1-u)}\}.$$

Now using (A1) $K(f_{\gamma_0},(1-p)f_{\gamma_0}+pf_{\gamma_1})$ is a continuous function of p and the result follows.

Lemma 6.4.3 *There exist a positive number M_1 such that, for any integer k,*

$$E\left(\sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k}{\partial u}(u,q)|\right) < M_1, \quad E\left(\sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k}{\partial q}(u,q)|\right) < M_1$$

Proof of Lemma 6.4.3.

Let us give the iterative model for the first derivative with respect to u and with respect to q of $p_k(u,q)$ for any value of (u,q). From now on, we shall denote by W_k the random variable $\frac{f_{\gamma_1}(Y_k)}{f_{\gamma_0}(Y_k)}$. W_k is non negative and its expectation under H_0 equals 1. The initial value is

$$\frac{\partial p_1}{\partial u}(u,q) = 1, \frac{\partial p_1}{\partial q}(u,q) = 0,$$

and for any positive integer k:

$$\begin{aligned} \frac{\partial p_{k+1}}{\partial u}(u,q) &= \frac{q}{(1-u)^2}\frac{(1-p_k)}{(1-p_k)+p_kW_k} \\ &\quad + \left(1-\frac{q}{(1-u)}\right)\frac{W_k}{[(1-p_k)+p_kW_k]^2}\frac{\partial p_k}{\partial u} \end{aligned} \tag{20}$$

$$\begin{aligned} \frac{\partial p_{k+1}}{\partial q}(u,q) &= \frac{\frac{u}{(1-u)}(1-p_k)-p_kW_k}{(1-p_k)+p_kW_k} \\ &\quad + \left(1-\frac{q}{(1-u)}\right)\frac{W_k}{[(1-p_k)+p_kW_k]^2}\frac{\partial p_k}{\partial q}. \end{aligned} \tag{21}$$

Direct computation leads to

$$|\frac{\partial p_{k+1}/\partial u}{(1-p_{k+1})^2}| \leq \frac{q}{(1-u)^2[(1-\frac{qu}{1-u})\bigwedge q]^2} + \frac{1-\frac{q}{(1-u)}}{(1-\frac{qu}{(1-u)})^2}W_k|\frac{\partial p_k/\partial u}{(1-p_k)^2}|, \tag{22}$$

in which $a\bigwedge b$ means $\inf\{a,b\}$. We have

$$\sup\{\frac{q}{(1-u)^2[(1-\frac{qu}{1-u})\bigwedge q]^2} \;:\; (u,q)\in\mathcal{P}_\eta\} = M_2 < +\infty$$

and also

$$\sup\{\frac{1-\frac{q}{(1-u)}}{(1-\frac{qu}{(1-u)})^2} \;:\; (u,q)\in\mathcal{P}_\eta\} = 1-h_1 < 1$$

so that

$$\sup_{(u,q)\in\mathcal{P}_\eta}|\frac{\partial p_{k+1}/\partial u}{(1-p_{k+1})^2}| \leq M_2 + (1-h_1)W_k\sup_{(u,q)\in\mathcal{P}_\eta}|\frac{\partial p_k/\partial u}{(1-p_k)^2}|. \tag{23}$$

This leads to

$$E\left[\sup_{(u,q)\in\mathcal{P}_\eta}|\frac{\partial p_{k+1}/\partial u}{(1-p_{k+1})^2}|\right] \leq M_2 + (1-h_1)E\left[\sup_{(u,q)\in\mathcal{P}_\eta}|\frac{\partial p_k/\partial u}{(1-p_k)^2}|\right].$$

By induction one obtains that for any integer k

$$E\left[\sup_{(u,q)\in\mathcal{P}_\eta}|\frac{\partial p_k/\partial u}{(1-p_k)^2}|\right] \leq \frac{M_2}{h_1}.$$

Since obviously

$$\sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k}{\partial u}| \leq \sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k/\partial u}{(1-p_k)^2}|$$

one obtains the first inequality of Lemma 6.4.3 with $M_1 = \frac{M_2}{h_1}$. The second inequality may be obtained similarly.

Let us come back to the proof of Proposition 6.2.2. For any positive δ, we have

$$P\left(\sup_{q\in\mathcal{Q}_\eta} \widehat{u}_q \geq \delta\right) \geq P\left(\widehat{u} \geq \delta\right)$$

so that it is enough to prove the Proposition for $\sup_{q\in\mathcal{Q}_\eta} \widehat{u}_q$. Now, let (u_i, q_i) be the centers of $N(h)$ squares of edge of lentgh h covering $\mathcal{P}_\eta^\delta = \{(u,q) \in \mathcal{P}_\eta : u \geq \delta\}$. We have

$$\begin{aligned} P\left(\sup_{q\in\mathcal{Q}_\eta} \widehat{u}_q \geq \delta\right) &\leq P\left(\sup_{(u,q)\in\mathcal{P}_\eta^\delta} S_n(u,q) \geq 0\right) \\ &\leq \sum_{i=1}^{N(h)} P\left(S_n(u_i,q_i) \geq -\epsilon(\delta,\eta)\right) \\ &+ P\left(\sup_{|u-u'|\leq h, |q-q'|\leq h, (u,q)\in\mathcal{P}_\eta^\delta} |S_n(u,q) - S_n(u',q')| \geq \epsilon(\delta,\eta)\right). \end{aligned}$$

Now, as soon as $|u-u'| \leq h, |q-q'| \leq h, (u,q) \in \mathcal{P}_\eta^\delta$,

$$|S_n(u,q) - S_n(u',q')| \leq \frac{h}{n}\sum_{k=1}^{n}(\sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k}{\partial u}(u,q)| + \sup_{(u,q)\in\mathcal{P}_\eta} |\frac{\partial p_k}{\partial q}(u,q)|)m(Y_k),$$

where m is the function

$$m(x) = \max\left(\frac{f_{\gamma_0}}{f_{\gamma_1}}(x); \frac{f_{\gamma_1}}{f_{\gamma_0}}(x)\right) - 1$$

Use Lemma 6.4.1, Lemma 6.4.2, Markov inequality, the fact that Y_k is independent of the p_k and its derivatives, and Lemma 6.4.3, to conclude that

$$\limsup_{n\to+\infty} P\left(\sup_{q\in\mathcal{Q}_\eta} \widehat{u}_q \geq \delta\right) = 0$$

and Proposition 6.2.2 follows.

6.4.2 Proof of Proposition 6.2.3.

We shall first prove:

Lemma 6.4.4 *Assume (A2), (A3) and (A4). The random variable*

$$\frac{\partial^2}{\partial u^2} S_n(0,q)$$

converges in probability to $N^2 C(q,q)$ uniformly over $\mathcal{Q}_\eta$.

Proof of Lemma 6.4.4.
We have from (14)

$$\frac{\partial^2}{\partial u^2} S_n(0,q) = -\frac{1}{n}\sum_{k=1}^{n}(\frac{\partial p_k}{\partial u}(0,q))^2(W_k-1)^2 + \frac{1}{n}\sum_{k=1}^{n}\frac{\partial^2 p_k}{\partial u^2}(0,q)(W_k-1).$$

We shall use the following trick **(T1)** for proving weak laws of large numbers for Markov chains. Let $(\Phi_k)_{k\in\mathbb{N}}$ be a Markov chain in $\mathbb{R}^d$ for some dimension d. Assume that g is a norm-like function, that is a non negative function such that $\lim_{|x|\to+\infty} g(x) = +\infty$.

- 1. If there exists an initial value x such that
$$\liminf_{k\to+\infty} E[g(\Phi_k)/\Phi_0 = x] < +\infty$$
then by Proposition 12-1-3 of Meyn and Tweedie (1993), the Markov chains admits an invariant probability μ.

- 2. If moreover for any initial points v_1 and v_2, $\Phi_k^{v_1} - \Phi_k^{v_2}$ tends in distribution to 0, then by using 1.IV.22 of Duflo (1996), the invariant probability is unique, and Φ_k converges in distribution to μ.

- 3. Then, by using Proposition 6-4-2 of Meyn and Tweedie (1993), $(\Phi_k)_{k\in\mathbb{N}}$ is an e-chain.

- 4. If moreover there exists a reachable state, then for any real function f such that $f \leq V$,
$$\frac{1}{n}\sum_{k=1}^{n} f(\Phi_k)$$
converges in probability to $\int f d\mu$ by a similar proof as for Theorem 18-5-1 of Meyn and Tweedie (1993).

We shall apply **(T1)** to $\Phi_k = (W_k, \frac{\partial p_k}{\partial u}(0,q), \frac{\partial^2 p_k}{\partial u^2}(0,q))$. Let us recall the recursive equations

$$\frac{\partial p_{k+1}}{\partial u}(0,q) = q + (1-q)W_k \frac{\partial p_k}{\partial u}(0,q) \tag{24}$$

$$\begin{aligned}\frac{\partial^2 p_{k+1}}{\partial u^2}(0,q) &= 2q - 2qW_k \frac{\partial p_k}{\partial u}(0,q) - 2(1-q)(W_k-1)W_k \left(\frac{\partial p_k}{\partial u}(0,q)\right)^2 \\ &\quad + (1-q)W_k \frac{\partial^2 p_k}{\partial u^2}(0,q)\end{aligned} \tag{25}$$

with initialization $\frac{\partial p_1}{\partial u}(0,q) = 1$ and $\frac{\partial^2 p_1}{\partial u^2}(0,q) = 0$. Notice that $\frac{\partial p_k}{\partial u}(u,q)$ is non negative for all integer k. (24) leads to

$$E\left(\frac{\partial p_{k+1}}{\partial u}(0,q)\right) = 1$$

for any integer k, and

$$E\left(\frac{\partial p_{k+1}}{\partial u}(0,q)\right)^2 = q^2 + 2q(1-q) + (1-q)^2(N^2+1)E\left(\frac{\partial p_k}{\partial u}(0,q)\right)^2$$

so that by induction, for any integer k,

$$E\left(\frac{\partial p_k}{\partial u}(0,q)\right)^2 \leq \frac{2}{\eta}.$$

Thus

$$E\left|\frac{\partial^2 p_{k+1}}{\partial u^2}(0,q)\right| \leq \frac{8}{\eta N^2} + (1-q)E\left|\frac{\partial^2 p_k}{\partial u^2}(0,q)\right|$$

and, by induction, for any integer k,

$$E\left|\frac{\partial^2 p_{k+1}}{\partial u^2}(0,q)\right| \leq \frac{8}{\eta N^2 q}.$$

Point 1. of **(T1)** holds with $g(x,y,z) = x^2 + y^2 + |z|$. To prove Point 2., notice first that the W_k are i.i.d. so that it is enough to prove it for the second and third coordinates of Φ_k. Let $v_1 = (y_1, z_1)$ and $v_2 = (y_2, z_2)$ be two different initial points for $(\frac{\partial p_k}{\partial u}(0,q), \frac{\partial^2 p_k}{\partial u^2}(0,q))$. Let $\Delta_k(y_1,y_2)$ be the difference of $\frac{\partial p_k}{\partial u}(0,q)$ with initial values y_1 and y_2. Let now $S_k(y_1,y_2)$ be the sum of $\frac{\partial p_k}{\partial u}(0,q)$ with initial values y_1 and y_2, and $\Delta_k^2(v_1,v_2)$ be the difference of $\frac{\partial^2 p_k}{\partial u^2}(0,q)$ when the initial values are v_1 and v_2. Equations (24) and (25) give

$$\Delta_{k+1}(y_1,y_2) = (1-q)W_k \Delta_k(y_1,y_2)$$

$$S_{k+1}(y_1,y_2) = 2q + (1-q)W_kS_k(y_1,y_2)$$
$$\Delta_{k+1}(y_1,y_2)S_{k+1}(y_1,y_2) = 2q(1-q)W_k\Delta_k(y_1,y_2) + (1-q)^2W_k^2\Delta_k(y_1,y_2)S_k(y_1,y_2)$$
$$\Delta_{k+1}^2(v_1,v_2) = -2qW_k\Delta_k(y_1,y_2) - 2(1-q)(W_k-1)W_k\Delta_k(y_1,y_2)S_k(y_1,y_2)$$
$$+(1-q)W_k\Delta_k^2(v_1,v_2)$$

By iterating these equations, one obtains that, for any integer k,

$$E|\Delta_k(y_1,y_2)| = (1-q)^{k-1}|y_1-y_2|,$$
$$E|\Delta_k(y_1,y_2)S_k(y_1,y_2)| \leq \frac{2q}{\eta_*}(1-q)^{k-1}|y_1-y_2|,$$

as soon as η_* is such that $\eta_* \leq \eta$ and $|x_1^2 - y_1^2| \leq 2q/\eta_*$, and then

$$E|\Delta_k^2(v_1,v_2)| \leq 2(1-q)^{k-1}|y_1-y_2| + \frac{4}{\eta_*}(1-q)^kN^2|y_1-y_2|$$

as soon as also $|z_1 - z_2| \leq 2|y_1-y_2| + \frac{4}{\eta_*}(1-q)N^2|y_1-y_2|$, which is always possible by choosing η_* small enough. Consequently, $\Delta_k(y_1,y_2)$ and $\Delta_k^2(v_1,v_2)$ tend in distribution to 0, and point 2 of **(T1)** holds. Since by using assumption (A2) the point $(0,q,2q)$ is obviously reachable, point 4. also holds, and for any q in $\mathcal{Q}_\eta$, the random variable

$$\frac{\partial^2}{\partial u^2}S_n(0,q)$$

converges in probability to $N^2C(q,q)$.

Let us now prove that this convergence is uniform. We have, using (24), for any q and q':

$$\begin{aligned}\frac{\partial p_{k+1}}{\partial u}(0,q) - \frac{\partial p_{k+1}}{\partial u}(0,q') &= (q-q')[1 - W_k\frac{\partial p_k}{\partial u}(0,q)] \\ &\quad +(1-q')W_k\left(\frac{\partial p_k}{\partial u}(0,q) - \frac{\partial p_k}{\partial u}(0,q')\right).\end{aligned}$$

Using Lemma 6.4.3 one obtains by induction

$$E\sup_{|q-q'|\leq\delta,q,q'\in\mathcal{Q}_\eta}|\frac{\partial p_k}{\partial u}(0,q) - \frac{\partial p_k}{\partial u}(0,q')| \leq \delta\cdot\frac{1+M_1}{q_\eta} \tag{26}$$

Also,

$$\begin{aligned}\left(\frac{\partial p_{k+1}}{\partial u}(0,q)\right)^2 &= q^2 + 2q(1-q)W_k\frac{\partial p_k}{\partial u}(0,q) \\ &\quad +(1-q)^2W_k^2\left(\frac{\partial p_k}{\partial u}(0,q)\right)^2,\end{aligned}$$

so that,

$$E \sup_{q \in \mathcal{Q}_\eta} \left(\frac{\partial p_k}{\partial u}(0,q) \right)^2 \leq \frac{M_1 + 2}{\eta}. \tag{27}$$

Further, for any q and q',

$$\begin{aligned} \left(\frac{\partial p_{k+1}}{\partial u}(0,q) \right)^2 - \left(\frac{\partial p_{k+1}}{\partial u}(0,q') \right)^2 &= q^2 - q'^2 + (2q(1-q) - 2q'(1-q'))W_k \frac{\partial p_k}{\partial u}(0,q) \\ & 2q'(1-q')W_k \left(\frac{\partial p_k}{\partial u}(0,q) - \frac{\partial p_k}{\partial u}(0,q') \right) \\ & +((1-q)^2 - (1-q')^2)W_k^2 \left(\frac{\partial p_k}{\partial u}(0,q) \right)^2 \\ & +(1-q')^2 W_k^2 \left[\left(\frac{\partial p_k}{\partial u}(0,q) \right)^2 - \left(\frac{\partial p_k}{\partial u}(0,q') \right)^2 \right] \end{aligned}$$

and by induction,

$$E \sup_{|q-q'| \leq \delta, q, q' \in \mathcal{Q}_\eta} \left| \left(\frac{\partial p_k}{\partial u}(0,q) \right)^2 - \left(\frac{\partial p_k}{\partial u}(0,q') \right)^2 \right| \leq \delta \cdot M_4 \tag{28}$$

with

$$M_4 = \frac{1}{\eta} \cdot [2 + 2M_1 + \frac{1 + M_1}{q_\eta} + \frac{2(N^2+1)M_1}{\eta}].$$

Following the same lines, using (25) one obtains

$$E \sup_{|q-q'| \leq \delta, q, q' \in \mathcal{Q}_\eta} \left| \frac{\partial^2 p_k}{\partial u^2}(0,q) - \frac{\partial^2 p_k}{\partial u^2}(0,q') \right| \leq \frac{\delta}{q_\eta} \cdot [2 + 2M_1 + 2\frac{1+M_1}{q_\eta} + 2N^2(\frac{M_1}{\eta} + M_4)]. \tag{29}$$

Using (27) and (29), it follows that for some constant M_5 one has

$$E \sup_{|q-q'| \leq \delta, q, q' \in \mathcal{Q}_\eta} \left| \frac{\partial^2}{\partial u^2} S_n(0,q) - \frac{\partial^2}{\partial u^2} S_n(0,q') \right| \leq \delta \cdot M_5$$

and Lemma 6.4.4 follows.
To finish the proof of Proposition 6.2.3 we shall prove:

Lemma 6.4.5 *Assume that (A2), (A3) and (A4) hold. Then, for any random $\tilde{u}$ converging to* 0 *uniformly in q,*

$$\frac{\partial^2}{\partial u^2} S_n(\tilde{u}, q) - \frac{\partial^2}{\partial u^2} S_n(0,q)$$

converges in probability to 0 *uniformly over $\mathcal{Q}_\eta$.*

Proof of Lemma 6.4.5.

For any non negative u, we have

$$\begin{aligned}\frac{\partial^2}{\partial u^2}S_n(u,q)-\frac{\partial^2}{\partial u^2}S_n(0,q) &= \frac{1}{n}\sum_{k=1}^n (W_k-1)^2[\frac{\partial p_k}{\partial u}(0,q)^2-\frac{\partial p_k}{\partial u}(u,q)^2 \\ &+\frac{\partial p_k}{\partial u}(u,q)^2(1-\frac{1}{(1-p_k(u,q)+p_k(u,q)W_k)^2})] \\ &+\frac{1}{n}\sum_{k=1}^n (W_k-1)[-\frac{\partial^2 p_k}{\partial u^2}(0,q)+\frac{\partial^2 p_k}{\partial u^2}(u,q) \\ &+\frac{\partial^2 p_k}{\partial u^2}(u,q)(\frac{1}{(1-p_k(u,q)+p_k(u,q)W_k)}-1)]\end{aligned}$$

We shall use the following fact that holds as soon as $\overline{u}$ is small enough. For any integer k, any q and any $u \leq \overline{u}$:

$$\frac{qu}{1-u} \leq p_k(u,q) \leq 1-q. \tag{30}$$

Also,

$$1-\frac{1}{1-p_k(u,q)+p_k(u,q)W_k}=\frac{p_k(u,q)(W_k-1)}{1-p_k(u,q)+p_k(u,q)W_k}$$

and

$$\left|\frac{1}{1-p_k(u,q)+p_k(u,q)W_k}\right| \leq \frac{1}{q_\eta}$$

so that

$$\begin{aligned}E\left(\sup_{q,u\leq\overline{u}}\left|\frac{\partial^2}{\partial u^2}S_n(u,q)-\frac{\partial^2}{\partial u^2}S_n(0,q)\right|\right) &\leq \frac{1}{n}\sum_{k=1}^n [N^2h_1(\overline{u})+\frac{1+q_\eta}{q_\eta^2}E|W_k-1|^3h_2(\overline{u}) \\ &+E|W_k-1|h_3(\overline{u})+\frac{1}{q_\eta}E|W_k-1|^2h_4(\overline{u})]\end{aligned}$$

with

$$h_1(\overline{u}) \geq E\left(\sup_{q,u\leq\overline{u}}\left|\frac{\partial p_k}{\partial u}(0,q)^2-\frac{\partial p_k}{\partial u}(u,q)^2\right|\right),$$

$$h_2(\overline{u}) \geq E\left(\sup_{q,u\leq\overline{u}}\left|p_k(u,q)\frac{\partial p_k}{\partial u}(u,q)^2\right|\right),$$

$$h_3(\overline{u}) \geq E\left(\sup_{q,u\leq\overline{u}}\left|\frac{\partial^2 p_k}{\partial u^2}(0,q)-\frac{\partial^2 p_k}{\partial u^2}(u,q)\right|\right)$$

and

$$h_4(\overline{u}) \geq E\left(\sup_{q,u\leq\overline{u}}\left|p_k(u,q)\frac{\partial^2 p_k}{\partial u^2}(u,q)\right|\right).$$

Lemma 6.4.5 will thus follow from the fact that it is possible to choose h_1, h_2, h_3 and h_4 such that for any $i = 1, \dots, 4$, $\lim_{u \to 0} h_i(u) = 0$. From now on, Cte will designate a universal constant number. We shall use the fact **(T2)** that if the deterministic real valued sequence $(z_k)_{k \in \mathbb{N}}$ satisfies, for a positive number δ, for any integer k

$$z_{k+1} \leq M + (1-\delta) z_k$$

and also $z_1 \leq M$, then for any integer k

$$z_k \leq Cte \cdot M.$$

(T2) will often be used with M involving $\overline{u}$, and tending to 0 when $\overline{u}$ does.

Lemma 6.4.6 *As soon as $\overline{u}$ is small enough,*

$$E\left(\sup_{q,u \leq \overline{u}} \frac{\partial p_k}{\partial u}(u,q)^2\right) \leq Cte < +\infty$$

Proof of Lemma 6.4.6.

Direct computation leads to, for any integer k, any q and any $u \leq \overline{u}$:

$$\begin{aligned} \frac{1}{(1-p_{k+1}(u,q))^4}\left(\frac{\partial p_{k+1}}{\partial u}(u,q)\right)^2 \quad \leq \quad & \frac{1}{q^2(1-\overline{u})^4} + \frac{2}{q^4(1-\overline{u})^2}(1-q) W_k \frac{\partial p_k}{\partial u}(u,q) \\ & \frac{(1-\frac{q}{1-u})^2}{(1-\frac{qu}{1-u})^4} W_k^2 \frac{1}{(1-p_k(u,q))^4}\left(\frac{\partial p_k}{\partial u}(u,q)\right)^2. \end{aligned}$$

Now there exists $\overline{u}$ small enough such that:

$$\sup_{q,u \leq \overline{u}} \frac{(1-\frac{q}{1-u})^2}{(1-\frac{qu}{1-u})^4} E(W_k^2) \leq 1 - \frac{\eta}{2}.$$

Taking the supremum over q and $u \leq \overline{u}$, then taking expectation, and then using **(T2)** one obtains that for any integer k:

$$E\left[\sup_{q,u \leq \overline{u}} \frac{1}{(1-p_k(u,q))^4}\left(\frac{\partial p_k}{\partial u}(u,q)\right)^2\right] \leq Cte < +\infty$$

and lemma 6.4.6 follows.

Lemma 6.4.7 *As soon as $\overline{u}$ is small enough,*

$$E\left(\sup_{q,u \leq \overline{u}} \frac{\partial p_k}{\partial u}(u,q)^{2+\epsilon_*}\right) \leq Cte < +\infty,$$

in which ϵ_* is defined below.

Proof of Lemma 6.4.7.
Using (A4),

$$3^{\epsilon}\sup_{q\in\mathcal{Q}_\eta}(1-q)^{2+\epsilon}E(W_k)^{2+\epsilon}$$

is a finite, continuous function of (q,ϵ) for $\epsilon<\epsilon_0$, which is upper bounded by $1-\eta$ for $\epsilon=0$, so that there exists a positive ϵ_* such that

$$3^{\epsilon_*}\sup_{q\in\mathcal{Q}_\eta}(1-q)^{2+\epsilon_*}E(W_k)^{2+\epsilon_*}\leq 1-\eta/2. \tag{31}$$

We now have for any integer k, any q and any $u\leq\overline{u}$:

$$\begin{aligned}
&\left(\frac{1}{(1-p_{k+1}(u,q))^2}\frac{\partial p_{k+1}}{\partial u}(u,q)\right)^{2+\epsilon_*}\\
=\;&\left(\frac{1}{(1-p_{k+1}(u,q))^4}\frac{\partial p_{k+1}}{\partial u}(u,q)^2\right)^{1+\epsilon_*/2}\\
\leq\;&\Big(Cte+CteW_k\frac{\partial p_k}{\partial u}(u,q)\\
+\;&\frac{(1-\frac{q}{1-u})^2}{(1-\frac{qu}{1-u})^4}W_k^2\frac{1}{(1-p_k(u,q))^4}\frac{\partial p_k}{\partial u}(u,q)^2\Big)^{1+\epsilon_*/2}\\
\leq\;&3^{\epsilon_*/2}\Big(Cte+CteW_k^{1+\epsilon_*/2}\frac{\partial p_k}{\partial u}(u,q)^{1+\epsilon_*/2}\\
&+\frac{(1-\frac{q}{1-u})^{2+\epsilon_*}}{(1-\frac{qu}{1-u})^{4+2\epsilon_*}}W_k^{2+\epsilon_*}\left(\frac{1}{(1-p_k(u,q))^4}\frac{\partial p_k}{\partial u}(u,q)^2\right)^{1+\epsilon_*/2}\Big).
\end{aligned}$$

Now there exists $\overline{u}$ small enough such that:

$$\sup_{q,u\leq\overline{u}}3^{\epsilon_*}\frac{(1-\frac{q}{1-u})^{2+\epsilon_*}}{(1-\frac{qu}{1-u})^{4+2\epsilon_*}}E(W_k^{2+\epsilon_*})\leq 1-\frac{\eta}{4}.$$

Taking the supremum over q and $u\leq\overline{u}$, then taking expectation, and then using **(T2)** one obtains that for any integer k:

$$E\left[\sup_{q,u\leq\overline{u}}\frac{1}{(1-p_k(u,q))^{4+2\epsilon_*}}\left(\frac{\partial p_k}{\partial u}(u,q)\right)^{2+\epsilon_*}\right]\leq Cte<+\infty$$

and lemma 6.4.7 follows.

Lemma 6.4.8 *As soon as $\overline{u}$ is small enough,*

$$E\left(\sup_{q,u\leq\overline{u}}p_k(u,q)\right)\leq Cte\cdot\overline{u}.$$

Proof of Lemma 6.4.8.
Direct computation leads to, for any integer k, any q and any $u \leq \overline{u}$:

$$\frac{p_{k+1}(u,q)}{1-p_{k+1}(u,q)} \leq Cte \cdot \overline{u} + \frac{1-q_\eta}{1-\frac{u}{1-u}} W_k \frac{p_k(u,q)}{1-p_k(u,q)}.$$

Notice that $p_1(u,q) = u$. Taking the supremum over q and $u \leq \overline{u}$, then taking expectation, and then using **(T2)** one obtains Lemma 6.4.8.

Lemma 6.4.9 *As soon as $\overline{u}$ is small enough, it is possible to choose h_2 such that*

$$\lim_{u\to 0} h_2(u) = 0.$$

Proof of Lemma 6.4.9.
Using Minkowsky's inequality, one may choose

$$h_2(\overline{u}) \leq E\left((\sup_{q,u\leq\overline{u}} p_k(u,q))^{1+2/\epsilon_*} \right)^{1/(1+2/\epsilon_*)} E\left(\sup_{q,u\leq\overline{u}} \frac{\partial p_k}{\partial u}(u,q)^{2+\epsilon_*} \right)^{1/(1+\epsilon_*/2)}$$

Since the p_k are ≤ 1, we have for any integer k, any q and any u

$$p_k(u,q))^{1+2/\epsilon_*} \leq p_k(u,q)$$

and using Lemma 6.4.7 and Lemma 6.4.8 one may choose

$$h_2(\overline{u}) = Cte \cdot (\overline{u})^{1/(1+2/\epsilon_*)}$$

Lemma 6.4.10 *As soon as $\overline{u}$ is small enough, it is possible to choose h_1 such that*

$$\lim_{u\to 0} h_1(u) = 0.$$

Proof of Lemma 6.4.10.
Direct computation leads to, for any integer k, any q and any $u \leq \overline{u}$:

$$\begin{aligned} \left|\frac{\partial p_{k+1}}{\partial u}(u,q) - \frac{\partial p_{k+1}}{\partial u}(0,q)\right| \quad \leq \quad & Cte \cdot \overline{u} + Cte \cdot W_k p_k(u,q) + Cte \cdot \overline{u} W_k \frac{\partial p_k}{\partial u}(u,q) \\ & + Cte \cdot W_k^2 p_k(u,q) \frac{\partial p_k}{\partial u}(u,q) \\ & + (1-q) W_k \left|\frac{\partial p_k}{\partial u}(u,q) - \frac{\partial p_k}{\partial u}(0,q)\right|. \end{aligned}$$

Taking the supremum over q and $u \leq \overline{u}$, then taking expectation, and then using **(T2)**, Lemma 6.4.3 and Lemma 6.4.8, one obtains that for any integer k,

$$E\left(\sup_{q,u\leq\overline{u}} |\frac{\partial p_k}{\partial u}(u,q) - \frac{\partial p_k}{\partial u}(0,q)|\right) \leq h_5(\overline{u}) \tag{32}$$

with

$$h_5(u) = \frac{1}{q_\eta} \cdot Cte \cdot u. \tag{33}$$

Now, since

$$|\frac{\partial p_k}{\partial u}(u,q)^2 - \frac{\partial p_k}{\partial u}(0,q)^2| = |\frac{\partial p_k}{\partial u}(u,q) - \frac{\partial p_k}{\partial u}(0,q)| \cdot \left(\frac{\partial p_k}{\partial u}(u,q) + \frac{\partial p_k}{\partial u}(0,q)\right)$$

direct computation leads to, for any integer k, any q and any $u \leq \overline{u}$:

$$\begin{aligned}
|\frac{\partial p_{k+1}}{\partial u}(u,q)^2 - \frac{\partial p_{k+1}}{\partial u}(0,q)^2| \quad \leq \quad & Cte \cdot \overline{u} + Cte \cdot \overline{u} \sup_{q,u\leq\overline{u}} \frac{\partial p_k}{\partial u}(u,q) \\
& + Cte \cdot W_k \sup_{q,u\leq\overline{u}} p_k(u,q) \sup_{q,u\leq\overline{u}} \frac{\partial p_k}{\partial u}(u,q) \\
& + Cte \cdot \overline{u} W_k \sup_{q,u\leq\overline{u}} \frac{\partial p_k}{\partial u}(u,q)^2 \\
& + Cte \cdot W_k^2 \sup_{q,u\leq\overline{u}} p_k(u,q) \frac{\partial p_k}{\partial u}(u,q)^2 \\
& + Cte \cdot W_k |\frac{\partial p_k}{\partial u}(u,q) - \frac{\partial p_k}{\partial u}(0,q)| \\
& + Cte \cdot \overline{u} W_k^2 \frac{\partial p_k}{\partial u}(u,q)^2 \\
& + Cte \cdot W_k^3 \sup_{q,u\leq\overline{u}} p_k(u,q) \sup_{q,u\leq\overline{u}} \frac{\partial p_k}{\partial u}(u,q)^2 \\
& + (1-q)^2 W_k^2 |\frac{\partial p_k}{\partial u}(u,q)^2 - \frac{\partial p_k}{\partial u}(0,q)^2|.
\end{aligned}$$

Taking the supremum over q and $u \leq \overline{u}$, then taking expectation, and then using **(T2)**, Lemma 6.4.6, Lemma 6.4.8, equation (32) and equation (33) one obtains Lemma 6.4.10.

Lemma 6.4.11 *As soon as $\overline{u}$ is small enough,*

$$E\left(\sup_{q,u\leq\overline{u}} |\frac{\partial^2 p_k}{\partial u^2}(u,q)|\right) \leq Cte < +\infty$$

Proof of Lemma 6.4.11.

First of all, let us give the recursive equations for the second derivative of the prediction filter:

$$\begin{aligned}\frac{\partial^2 p_{k+1}}{\partial u^2}(u,q) &= \frac{2q}{(1-u)^3}\frac{(1-p_k)}{(1-p_k)+p_kW_k} \\ & -\frac{2q}{(1-u)^2}\frac{W_k}{[(1-p_k)+p_kW_k]^2}\frac{\partial p_k}{\partial u}(u,q) \\ -2\left(1-\frac{q}{(1-u)}\right) &\cdot (W_k-1)\frac{W_k}{[(1-p_k)+p_kW_k]^3}\left(\frac{\partial p_k}{\partial u}(u,q)\right)^2 \\ & +\left(1-\frac{q}{(1-u)}\right)\frac{W_k}{[(1-p_k)+p_kW_k]^2}\frac{\partial^2 p_k}{\partial u^2}(u,q)\end{aligned} \tag{34}$$

with initial value $\frac{\partial^2 p_1}{\partial u^2}(u,q)=0$. Direct computation leads to, for any integer k, any q and any $u\leq\overline{u}$:

$$\begin{aligned}\frac{1}{(1-p_{k+1}(u,q))^2}\left|\frac{\partial^2 p_{k+1}}{\partial u^2}(u,q)\right| &\leq Cte+Cte\cdot W_k\frac{\partial p_k}{\partial u}(u,q)+Cte\cdot W_k^2\frac{\partial p_k}{\partial u}(u,q)^2 \\ & +\frac{(1-\frac{q}{1-u})}{(1-\frac{qu}{1-u})^2}W_k\frac{1}{(1-p_k(u,q))^2}\left|\frac{\partial^2 p_k}{\partial u^2}(u,q)\right|\end{aligned}$$

Now there exists $\overline{u}$ small enough such that:

$$\sup_{q,u\leq\overline{u}}\frac{(1-\frac{q}{1-u})}{(1-\frac{qu}{1-u})^2}\leq 1-\frac{\eta}{2}.$$

Taking the supremum over q and $u\leq\overline{u}$, then taking expectation, and then using **(T2)** one obtains that for any integer k:

$$E\left[\sup_{q,u\leq\overline{u}}\frac{1}{(1-p_k(u,q))^2}\left|\frac{\partial^2 p_k}{\partial u^2}(u,q)\right|\right]\leq Cte<+\infty$$

and lemma 6.4.11 follows.

Lemma 6.4.12 *As soon as $\overline{u}$ is small enough, it is possible to choose h_4 such that*

$$\lim_{u\to 0}h_4(u)=0.$$

Proof of Lemma 6.4.12.

Follows from the recursive inequality:

$$\frac{p_{k+1}(u,q)}{(1-p_{k+1}(u,q))^3}\left|\frac{\partial^2 p_{k+1}}{\partial u^2}(u,q)\right| \leq Cte\cdot p_{k+1}(u,q)+Cte\cdot\overline{u}\cdot W_k\frac{\partial p_k}{\partial u}(u,q)$$

$$
\begin{aligned}
&+Cte\cdot\overline{u}W_k^2\frac{\partial p_k}{\partial u}(u,q)^2\\
&+Cte\cdot W_k^2 p_k(u,q)\frac{\partial p_k}{\partial u}(u,q)\\
&+Cte\cdot W_k^3 p_k(u,q)\frac{\partial p_k}{\partial u}(u,q)^2\\
&+Cte\cdot\overline{u}W_k|\frac{\partial^2 p_k}{\partial u^2}(u,q)|\\
&+\frac{(1-\frac{q}{1-u})^2}{(1-\frac{qu}{1-u})^3}W_k^2\frac{p_k(u,q)}{(1-p_k(u,q))^3}|\frac{\partial^2 p_k}{\partial u^2}(u,q)|.
\end{aligned}
$$

Lemma 6.4.13 *As soon as $\overline{u}$ is small enough, it is possible to choose h_3 such that*

$$\lim_{u\to 0} h_3(u) = 0.$$

Proof of Lemma 6.4.13.
Follows from the recursive inequality:

$$
\begin{aligned}
|\frac{\partial^2 p_{k+1}}{\partial u^2}(u,q)-\frac{\partial^2 p_{k+1}}{\partial u^2}(0,q)| \;\le\;& Cte\cdot\overline{u}+Cte\cdot W_k p_k(u,q)\\
&+Cte\cdot W_k|\frac{\partial p_k}{\partial u}(u,q)-\frac{\partial p_k}{\partial u}(0,q)|\\
&+Cte\cdot\overline{u}W_k\frac{\partial p_k}{\partial u}(u,q)+Cte\cdot W_k^2 p_k(u,q)\frac{\partial p_k}{\partial u}(u,q)\\
&+Cte\cdot W_k^2|\frac{\partial p_k}{\partial u}(u,q)^2-\frac{\partial p_k}{\partial u}(0,q)^2|\\
&+Cte\cdot\overline{u}W_k^2\frac{\partial p_k}{\partial u}(u,q)^2\\
&+Cte\cdot W_k^3 p_k(u,q)\frac{\partial p_k}{\partial u}(u,q)^2\\
&+Cte\cdot\overline{u}W_k|\frac{\partial^2 p_k}{\partial u^2}(u,q)|\\
&+Cte\cdot W_k^2 p_k(u,q)|\frac{\partial^2 p_k}{\partial u^2}(u,q)|\\
&+(1-q)W_k|\frac{\partial^2 p_k}{\partial u^2}(u,q)-\frac{\partial^2 p_k}{\partial u^2}(0,q)|.
\end{aligned}
$$

6.4.3 Proof of Proposition 6.2.4.

To prove Proposition 6.2.4, we shall prove the finite dimensional marginal convergence of the process and its tightness. We shall use:

Lemma 6.4.14 *Under the assumptions of Proposition 6.2.4, for any q in $\mathcal{Q}_\eta$,*

$$\left(\frac{\partial p_k}{\partial u}(0,q); W_k\right)_{k\in\mathbb{N}}$$

is a positive Harris Markov chain, g-uniformly ergodic for $g(x) = 1+ \parallel x \parallel_2^{2+\epsilon_*}$.

Proof of Lemma 6.4.14.

It is enough to prove the results of the Lemma for the Markov chain $\Phi = (\frac{\partial p_k}{\partial u}(0,q))_{k\in\mathbb{N}}$. Using (A2), and applying Proposition 4.2.1 of Meyn and Tweedie (1993), Φ is ψ-irreducible with ψ the Lebesgue measure on $[q,+\infty[$. Then, since Φ is a Feller chain by (16), applying Propositions 6.2.8 and 6.2.9 of Meyn and Tweedie (1993), all compact subsets of $[q,+\infty[$ are petite and Φ is a T-chain. Also, Φ is strongly aperiodic. Taking now $g(x) = 1+ \parallel x \parallel_2^{2+\epsilon_*}$, we have

$$\begin{aligned} E(V(\frac{\partial p_{k+1}}{\partial u}(0,q))/\frac{\partial p_k}{\partial u}(0,q)=p) &= E\left((q+(1-q)\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y)p)^2\right)^{1+\epsilon_*/2}+1 \\ &\leq 1+3^{\epsilon_*}\Big(q^{2+\epsilon_*} \\ &\quad +[2q(1-q)]^{1+\epsilon_*/2}E(\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y))^{1+\epsilon_*/2}p^{1+\epsilon_*/2} \\ &\quad +(1-q)^{2+\epsilon_*}E(\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y))^{2+\epsilon_*}(g(p)-1)\Big). \end{aligned}$$

The right side of the last inequality divided by $V(p)$ tends to

$$3^{\epsilon_*}(1-q)^{2+\epsilon_*}E(\frac{f_{\gamma_1}}{f_{\gamma_0}}(Y))^{2+\epsilon_*}$$

as p tends to ∞, which is less or equal to $1-\eta/2$, so that there exists a M such that, for $p \geq M$,

$$E(V(\frac{\partial p_{k+1}}{\partial u}(0,q))/\frac{\partial p_k}{\partial u}(0,q)=p) \leq (1-\eta/3)g(p),$$

and we have for some positive b,

$$E(g(\frac{\partial p_{k+1}}{\partial u}(0,q))/\frac{\partial p_k}{\partial u}(0,q)=p) - g(p) \leq -\eta/3\cdot g(p) + b1_{p\leq M}. \tag{35}$$

Consequently Φ is Harris positive by Theorem 11.3.4 and V-uniformly ergodic by Theorem 16.1.2. of Meyn and Tweedie (1993).

Lemma 6.4.15 *For any integer m and any $q_1, \ldots, q_m$ in $\mathcal{B}_\eta$, the vector*

$$\sqrt{n}(\frac{\partial}{\partial u} S_n(0, q_1), \ldots, \frac{\partial}{\partial u} S_n(0, q_m))$$

converges in distribution to the centered gaussian distribution with covariance $N^2 C(q_i, q_j)$.

Proof of Lemma 6.4.15.
Define

$$M_n = n(\frac{\partial}{\partial u} S_n(0, q_1), \ldots, \frac{\partial}{\partial u} S_n(0, q_m)).$$

M_n is a martingale with respect to the σ-field $\mathcal{F}_n = \sigma(Y_1, \ldots, Y_n)$, with hook

$$\langle M_n \rangle_{i,j} = N^2 \sum_{k=1}^{n} \frac{\partial p_k}{\partial u}(0, q_i) \frac{\partial p_k}{\partial u}(0, q_j).$$

We shall apply Theorem 2.8.41 of Dacunha-Castelle and Duflo (1993). The Lindeberg condition holds since, for any positive δ,

$$\frac{1}{n} \sum_{k=1}^{n} E[\| M_k - M_{k-1} \|^2 1_{\|M_k - M_{k-1}\| \geq \delta\sqrt{n}} / \mathcal{F}_{k-1}] \leq \frac{Cte}{(\delta\sqrt{n})^{\epsilon_*}} \max_{i=1,\ldots,m} \frac{1}{n} \sum_{k=1}^{n} (\frac{\partial p_k}{\partial u}(0, q_i))^{2+3\epsilon_*}$$

which converges to 0 as n tends to infinity thanks to Lemma 6.4.14. Theorem 1.IV.24 of Duflo (1996) may be applied (with $a = 2$ and $\alpha = 1 - \eta$) to the iterative markovian model leading to $(\frac{\partial p_k}{\partial u}(0, q_1), \ldots, \frac{\partial p_k}{\partial u}(0, q_m))$, so that

$$\frac{1}{n} \sum_{k=1}^{n} \frac{\partial p_k}{\partial u}(0, q_i) \frac{\partial p_k}{\partial u}(0, q_j)$$

converges, as n tends to infinity, to the covariance of $\frac{\partial p_k}{\partial u}(0, q_i)$ and $\frac{\partial p_k}{\partial u}(0, q_j)$ with respect to the invariant distribution of the Markov chain. This covariance with respect to the invariance distribution may be computed using equation (16). The obtained limiting value is

$$\frac{1 - (1 - q_1)(1 - q_2)}{1 - (1 - q_1)(1 - q_2)(N^2 + 1)}$$

so that

$$\lim_{n \to +\infty} \frac{1}{n} \langle M_n \rangle_{i,j} = N^2 C(q_i, q_j)$$

and Lemma 6.4.15 follows.

Lemma 6.4.16 *The process*

$$\left(\sqrt{n}\frac{\partial}{\partial u}S_n(0,q)\right)_{q\in\mathcal{Q}_\eta}$$

is uniformly tight.

Proof of Lemma 6.4.16.

For any q_1, q_2 and any positive integer k let

$$Z_k = \left(\frac{\partial p_k}{\partial u}(0,q_1) - \frac{\partial p_k}{\partial u}(0,q_2)\right)(W_k - 1).$$

Using the Rosenthal inequality for martingales (Hall and Heyde, 1980) one has for any $p \geq 2$:

$$\begin{aligned} E \mid \sqrt{n}(\frac{\partial}{\partial u}S_n(0,q_1) - \frac{\partial}{\partial u}S_n(0,q_2)) \mid^p \quad &\leq \quad Cte \cdot E\left[\left(\frac{1}{n}\sum_{k=1}^{n} E(Z_k^2/Y_1,\ldots,Y_{k-1})\right)^{p/2}\right] \\ &\quad + \frac{1}{n^{p/2}}\sum_{k=1}^{n} E \mid Z_k \mid^p . \end{aligned} \tag{36}$$

We have now

$$E(Z_k^2/Y_1,\ldots,Y_{k-1}) = \left(\frac{\partial p_k}{\partial u}(0,q) - \frac{\partial p_k}{\partial u}(0,q_2)\right)^2 N^2.$$

By iterating equation (16), one obtains for any q:

$$\frac{\partial p_k}{\partial u}(0,q) = q + \sum_{l=1}^{k-2} q(1-q)^l W_{k-1}\cdots W_{k-l} + (1-q)^{k-1}W_{k-1}\cdots W_1$$

so that

$$\mid \frac{\partial p_k}{\partial u}(0,q_1) - \frac{\partial p_k}{\partial u}(0,q_2) \mid \leq \mid q_1 - q_2 \mid [1 + \sum_{l=1}^{k-1}(1-\tilde{q})^{l-1} \mid 1-(l+1)\tilde{q} \mid W_{k-1}\cdots W_{k-l}]$$

where $\tilde{q}$ is some number between q_1 and q_2.

Define

$$\alpha = 1 - \frac{2}{2+\epsilon_*}\frac{\log 1-\eta/4}{\log 1-q_\eta}.$$

We have $0 < \alpha < 1$, and for big enough l one has

$$(l+1)(1-q_\eta)^{l-1} \leq (1-q_\eta)^{\alpha l},$$

so that:

$$| \frac{\partial p_k}{\partial u}(0,q_1) - \frac{\partial p_k}{\partial u}(0,q_2) |\leq Cte \mid q_1 - q_2 \mid [1 + \sum_{l=1}^{k-1}(1-q_\eta)^{\alpha l} W_{k-1}\cdots W_{k-l}]. \tag{37}$$

Define the Markov chain $(D_k)_{k\in\mathbb{N}}$ by $D_1 = 0$ and

$$D_{k+1} = 1 + (1-q_\eta)^\alpha W_k D_k.$$

Following the arguments of the proof of Lemma 6.4.14, it can be proved that this Markov chain is Harris positive and $1+x^{2+\epsilon_*}$-uniformly ergodic. Now, equation (37) implies that

$$| \frac{\partial p_k}{\partial u}(0,q_1) - \frac{\partial p_k}{\partial u}(0,q_2) |\leq Cte \mid q_1 - q_2 \mid D_k. \tag{38}$$

We thus obtain

$$\begin{aligned} E\left[\left(\frac{1}{n}\sum_{k=1}^n E(Z_k^2/Y_1,\ldots,Y_{k-1})\right)^{p/2}\right] &\leq Cte \mid q_1 - q_2 \mid^p E\left[\left(\frac{1}{n}\sum_{k=1}^n D_k^2\right)^{p/2}\right] \\ &\leq Cte \mid q_1 - q_2 \mid^p E\left[\frac{1}{n}\sum_{k=1}^n D_k^p\right] \end{aligned} \tag{39}$$

where the last inequality follows by Jensen's inequality since $p/2 \geq 1$. But, as soon as $p \leq 2+\epsilon_*$, $\frac{1}{n}\sum_{k=1}^n ED_k^p$ converges and we get:

$$E\left[\left(\frac{1}{n}\sum_{k=1}^n E(Z_k^2/Y_1,\ldots,Y_{k-1})\right)^{p/2}\right] \leq Cte \mid q_1 - q_2 \mid^p \tag{40}$$

for big enough n. Also, using (38), as soon as $p \leq 2+\epsilon_*$:

$$\frac{1}{n^{p/2}}\sum_{k=1}^n E \mid Z_k \mid^p \leq Cte \mid q_1 - q_2 \mid^p \frac{1}{n^{p/2}}\sum_{k=1}^n ED_k^p \leq Cte \mid q_1 - q_2 \mid^p \tag{41}$$

for big enough n.

It follows from equations (36), (40) and (41) that for big enough n, for any q_1 and q_2,

$$E \mid \sqrt{n}(\frac{\partial}{\partial u}S_n(0,q_1) - \frac{\partial}{\partial u}S_n(0,q_2)) \mid^{2+\epsilon_*} \leq Cte \cdot (q_1 - q_2)^{2+\epsilon_*}, \tag{42}$$

which is the Kolmogorov criterion for the uniform tightness of a continuous process, and Lemma 6.4.16 follows.

Bibliographie

[1] Atwood L.D., Wilson A.F., Bailey-Wilson J.E., Carruth J.N., Elston R.C. On the distribution of the likelihood ratio test statistic for a mixture of two normal distributions. *Commun. Statist.-Simula.*, 25(3): 733–740, 1996.

[2] Baum, L.E. and Petrie, T. Statistical inference for probabilistic functions of finite state Markov chains. *Ann. Math. Stat.*, 37, 1554-1563, 1966.

[3] Bickel, P.J. and Ritov, Y. Inference in hidden Markov models I: local asymptotic normality in the stationary case. *Bernoulli*, 2, 199-228, 1996.

[4] Bickel, P.J., Ritov, Y. and Ryden, T. Asymptotic normality of the maximum-likelihood estimator for general hidden Markov models. *Annals of Stat.*, 26, 1614-1635, 1998.

[5] Chuang R-J., Mendell N.R. The approximate null distribution of the likelihood ratio test for a mixture of two bivariate normal distributions with equal covariance. *Commun. Statist.-Simula.*, 26(2): 631–648, 1997.

[6] Churchill, G.A. Stochastic models for heterogeneous DNA sequences. *Bulletin of Math. Biology*, 51, 79-94, 1989.

[7] Ciuperca, G. Sur le test de maximum de vraisemblance pour le mélange de populations *Note aux C.R.A.S.*, 328, série I, 4, 351-358, 1999.

[8] Dacunha-Castelle, D. and Duflo, M. *Probabilités et statistiques, tome 2.* Masson, 1993.

[9] Dacunha-Castelle, D. and Gassiat, E. Estimation of the number of components in a mixture *Bernoulli*, 3, 279-299, 1997a.

[10] Dacunha-Castelle, D. and Gassiat, E. Testing in locally conic models. *ESAIM Prob. and stat.*, 1, 1997b.

[11] Dacunha-Castelle, D. and Gassiat, E. Testing the order of a model using locally conic parametrization: population mixtures and stationary ARMA processes. *Ann. Statist.*, (to appear) 1999.

[12] Dempster A.P., Laird N.M., Rubin D.B. Large Maximum-likelihood from incomplete data via the EM algorithm. *J. Royal Statist. Soc. Ser. B*, 39: 1–38, 1977.

[13] Duflo, M. *Algorithmes stochastiques.* Springer, 1996.

[14] Feng Z.D., McCulloch C.E. Using bootstrap Likelihood Ratio in Finite Mixture Models. *J. R. Statist. Soc. B*, 58(3): 609–617, 1996.

[15] Finesso, L. *Consistent Estimation of the Order for Markov and Hidden Markov Chains* Ph. D. Thesis, University of Maryland, 1990.

[16] Fredkin, D.R. and Rice, J.A. Maximum likelihood estimation and identification directly from single-channel recordings. *Proc. Royal Soc. London Ser. B*, 249, 125-132, 1992.

[17] Hall, P. and Heyde, C.C. *Martingale Limit Theory and Its Application.* Academic Press, 1980.

[18] Hartigan, J.A. A failure of likelihood ratio asymptotics for normal mixtures. *Proc. Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer*, L.M. Le Cam and R.A. Olshen eds, 807-810, 1985.

[19] Henna, J. On estimating the number of constituents of a finite mixture of continuous distributions. *Ann. Inst. Statist. Math.*, 37, 235-240, 1985.

[20] Keribin, C. Consistent estimation of the Order of Mixture Models. *submitted*, 1997.

[21] Leroux, B.G. Maximum-likelihood estimation for hidden Markov models. *Stoch. Proc. and their Appl.*, 40, 127-143, 1992.

[22] Leroux, B.G. and Puterman, M.L. Maximum Penalized Likelihood Estimation for Independent and Markov-Dependent Mixture Models. *Biometrics*, 48, 545-558, 1992.

[23] Lindsay, B.G. *Mixture models: Theory, Geometry and Applications.* 1995.

[24] Mac Donald, I.L. and Zucchini, W. *Hidden Markov and Other Models for Discrete-valued Time Series.* Chapman and Hall, 1997.

[25] McLachlan G.J. On Bootstrapping the Likelihood Ratio Test Statistic for the Number of Components in a Normal Mixture. *App. Statist.*, 36(3): 318–324, 1987.

[26] Mevel, L. *Statistique asymptotique pour les modèles de Markov cachés* Thèse de l'Université de Rennes I, 1997.

[27] Mevel, L. and LeGland F Exponential forgetting and Geometric Ergodicity in Hidden Markov models. *Mathematics of Control, Signals, and Systems*, to appear.

[28] Meyn, S.P. and Tweedie, R.L. *Markov chains and stochastic stability* Springer-Verlag, 1993.

[29] Rabiner, L.R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, 77, 257-284, 1989.

[30] Ryden, T. Estimating the order of hidden Markov models. *Statistics*, 26, 345-354, 1995.

[31] Vandekerkhove, P. *Identification de l'ordre des processus ARMA stables. Contribution à l'étude statistique des chaines de Markov cachées.* Thèse de l'Université de Montpellier II, 1997.

[32] Van der Vaart, A. *Asymptotic Statistics.* Cambridge ed., 1999.