Asymptotic behaviour of the posterior distribution in overfitted mixture models

Judith Rousseau, Kerrie Mengersen

Asymptotic behaviour of the posterior distribution in over-fitted mixture models

Judith Rousseau

ENSAE-CREST and CEREMADE, Université Paris Dauphine 3 avenue Pierre Larousse, 92 245 Malakoff Cedex France

Kerrie Mengersen

Queensland University of Technology, Brisbane, Queensland 4001 Australia.

Abstract.

In this paper we study the asymptotic behaviour of the posterior distribution in a mixture model when the number of components in the mixture is larger than the true number of components, a situation commonly referred to as overfitted mixture. We prove in particular that quite generally the posterior distribution has a stable and interesting behaviour, since it tends to empty the extra components. This stability is achieved under some restriction on the prior, which can be used as a guideline for choosing the prior. Some simulations are presented to illustrate this behaviour.

Keywords: Asymptotic; Bayesian; mixture models; overfitting; posterior concentration

1. Introduction

Finite mixture models provide a very flexible and often biologically or physically interpretable model for describing complex distributions (Marin and Robert (2007); Frühwirth-Schnatter (2006); MacLachlan and Peel (2000); Titterington et al. (1985)) An important concomitant problem of choosing the appropriate number of components in a mixture distribution has entertained and concerned a large number of researchers and attracted a correspondingly large literature (Akaike (1973); Dempster et al. (1977); Lee et al. (2008); McGrory and Titterington (2007); Richardson and Green (1997); Robert and Wraith (2009); Schwarz (1978)). When the number of components is unknown, the analyst can intentionally or unintentionally propose an over-fitting model, that is, one with more components than can be supported by the data. The problem of non-identifiability in estimation of over-fitted mixture models is well known; in her review of this problem, for example, Frühwirth-Schnatter (2006) observes that identifiability will be violated as either one of the component weights is zero or two of the component parameters are equal. Examples of this behaviour are provided and possible solutions are presented, including choosing priors that bound the posterior away from the unidentifiability sets or that induce shrinkage for elements of the component parameters, although the opportunity to reduce the mixture model to the true model is forfeited by this practice.

In this paper, we contribute to this growing understanding of how over-fitted mixtures behave in Bayesian analysis, particularly as the dimension of the component parameters grows. Consider a mixture model of the form

$$f_\theta(x) = \sum_{j=1}^{k} p_j g_{\gamma_j}(x), \quad k \geq 1, \quad \gamma_j \in \Gamma, \quad \theta = (p_1, ..., p_k, \gamma_1, ..., \gamma_k) \in \Theta_k \quad \Gamma \subset \mathbb{R}^d. \tag{1}$$

The number of components k can be known or unknown. Estimating k can be difficult in practice and it is often the case that one prefers to choose a large k, with the risk that the *true distribution* has less components. However the non-identifiability of the parameter in cases where the true distribution has a smaller number of components leads to the following question: how can we interpret the posterior distribution in such cases? To answer such a question we investigate the asymptotic behaviour of the posterior distribution.

More precisely, assume that we have observations $X_1, ..., X_n$, iid from a mixture model with k_0 components:

$$f_0(x) = \sum_{j=1}^{k_0} p_j^0 g_{\gamma_j^0}(x), \quad k \geq 1, \quad \gamma_j^0 \in \Gamma, \quad 1 \leq k_0 < k. \tag{2}$$

In such cases the model is non-identifiable since all values of the parameter in the form

$$\theta = (p_1^0, ..., p_{k_0}^0, 0, \gamma_1^0, ..., \gamma_{k_0}^0, \gamma),$$

for all $\gamma \in \Gamma$ and all values of the parameter in the form $\theta = (p_1^0, ., p_j .., p_{k_0}^0, p_{k+1}, \gamma_1^0, ..., \gamma_{k_0}^0, \gamma_j^0)$ with $p_j + p_{k+1} = p_j^0$ satisfy $f_0 = f_\theta$. This non-identifiability is much stronger than the non identifiability corresponding to permutations of the labels in the mixture representation. In such cases, it is well known that the asymptotic behaviour of the likelihood is not regular, although under mild conditions the maximum likelihood converges to the set of values in Θ_k satisfying $f_\theta = f_0$, see Feng and McCulloch (1996). In such cases where the true parameter lies on the boundary of the parameter set, the multiplicity of the limiting set implies that the maximum likelihood estimator does not have a stable asymptotic behaviour. When f_θ is the main object of interest this is not of great importance, however in many situations recovering θ is of major interest. A particular example in which such estimates are particularly useful is time evolving mixture models, where the estimation of the number of components at each time period would be too time consuming to make. In such cases, using a quite large number of components, which can be regarded as a reasonnable upper bound on the number of components over the different time periods is computationally easier. It thus becomes crucial to know that the posterior distribution under overfitted mixtures gives interpretable results.

In this paper we study the asymptotic behaviour of the posterior distribution, inducing some results on the asymptotic behaviour of Bayesian estimates such as the posterior mean. It turns out,

that the posterior distribution has a much more stable behaviour than the maximum likelihood estimator if the prior on the weights is reasonable. In particular we prove that if the dimension d of γ is larger than some value depending on the prior, then asymptotically the extra components in the k-mixture are emptied under the posterior distribution. This result is of interest in particular because it validates the use of Bayesian estimation in mixture models with too many components. It is also of interest since it is one of the few example where the prior can actually have an impact asymptotically, even to first order (consistency) and where choosing a *less informative prior* leads to better results. It is to be noted, that the usual *less informative prior* are designed to favour weights close to 0, so that in the present framework they actually bring the correct information, as opposed to more informative priors which would prevent the weights to become small. It also shows that the penalization effect of integrating out the parameter, as considered in the Bayesian framework is not only useful in model choice or testing contexts but also in estimating contexts.

In Section 2 we state our main result, where we link conditions on the prior to the asymptotic behaviour of the posterior distribution. A simulation study is presented in Section 3 where we illustrate our theoretical results and also consider a case for which no theoretical asymptotic results have been obtained.

2. Consistency issues : main results

In this section we state the main results of the paper, namely that the posterior distribution concentrates on the subset of parameters for which $f_\theta = f_0$ so that $k - k_0$ components have weight 0. The reason for this stable behaviour as opposed to the unstable behaviour of the maximum likelihood estimator is that integrating out the parameter acts as a penalization: the posterior is essentially putting mass on the sparsest way to approximate the true density.

We first give some notation and state the assumptions needed to describe the asymptotic behaviour of the posterior distribution.

2.1. Assumptions and notation

We denote $\Theta_k^0 = \{\theta \in \Theta_k; f_\theta = f_0\}$ and let $l_n(\theta)$ be the log-likelihood calculated at θ. Denote by $\|f-g\| = \int |f-g|(x)dx$ the L_1 distance and $\mathbb{P}_n(g) = \sum_{i=1}^n g(X_i)/n$ and $\mathbb{G}_n(g) = \sqrt{n}[\mathbb{P}_n(g) - F_0(g)]$ where $F_0(g) = \int f_0(x)g(x)dx$ and denote by $\mathrm{Leb}(A)$ the Lebesgue measure of a set A. We also use the symbol $a \wedge b$ to designate $\min(a, b)$.

Let ∇g_γ be the vector of first derivatives of g_γ with respect to γ, and $D^2 g_\gamma$ be the matrix of second derivatives with respect to γ. Define for $\delta \geq 0$

$$\bar{g}_\gamma = \sup_{|\gamma' - \gamma| \leq \delta} g_{\gamma'}, \quad \underline{g}_\gamma = \inf_{|\gamma' - \gamma| \leq \delta} g_{\gamma'}$$

We now introduce some notations that are useful to characterise Θ_k^0, following Liu and Shao

(2004)'s presentation. Let $\mathbf{t} = (t_i)_{i=0}^{k_0}$ with $0 = t_0 < t_1 < ... < t_{k_0} \leq k$ be a partition of $\{1, ..., k\}$. For all $\theta \in \Theta_k$ such that $f_\theta = f_0$ there exists $\mathbf{t}$ as defined above such that, up to a permutation of the labels,

$$\forall i = 1, ..., k_0 \quad \gamma_{t_{i-1}+1} = ... = \gamma_{t_i} = \gamma_i^0, \quad p(i) = \sum_{j=t_{i-1}+1}^{t_i} p_j = p_i^0, \quad p_{t_{k_0}+1} = ... = p_k = 0.$$

In other words $I_i = \{t_{i-1} + 1, ..., t_i\}$ represents the cluster of components in $\{1, ..., k\}$ having the same parameter as γ_i^0. Then define the following parameterisation of $\theta \in \Theta_k$ (up to a permutation)

$$\phi_{\mathbf{t}} = \left((\gamma_j)_{j=1}^{t_{k_0}}, (s_i)_{i=1}^{k_0-1}, (p_j)_{j=t_{k_0+1}}^{k}\right) \in \mathbb{R}^{dt_{k_0}+k_0+k-t_{k_0}-1}, \quad s_i = p(i) - p_i^0, i = 1, ..., k_0$$

and

$$\psi_{\mathbf{t}} = \left((q_j)_{j=1}^{t_{k_0}}, \gamma_{t_{k_0}+1}, ..., \gamma_k\right), \quad q_j = \frac{p_j}{p(i)}, \quad \text{when} \quad j \in I_i = \{t_{i-1} + 1, ..., t_i\}.$$

Note that f_0 corresponds to

$$\phi_{\mathbf{t}}^0 = \left(\gamma_1^0, ..., \gamma_1^0, \gamma_2^0, ..., \gamma_2^0, ..., \gamma_{k_0}^0, ..., \gamma_{k_0}^0, 0...0...0\right)$$

where γ_i^0 is repeated $t_i - t_{i-1}$ times in the above vector, for any $\psi_{\mathbf{t}}$.

Then we parameterize θ as $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$, so that $f_\theta = f_{(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}$ and we denote $f'_{(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})}$ and $f''_{(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})}$ the first and second derivatives of $f_{(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}$ with respect to $\phi_{\mathbf{t}}$ and computed at $\theta_0 = (\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})$.

We also denote by $P^\pi[.|X^n]$ the posterior distribution, where $X^n = (X_1, ..., X_n)$.

Assumptions

A1 *L1 consistency*: There exists $\delta_n \leq (\log n)^q/\sqrt{n}$, for some $q \geq 0$ such that,

$$\lim_{M \to +\infty} \text{limsup}_n E_0^n \left\{P^\pi \left[\|f_0 - f_\theta\| \geq M\delta_n | X^n\right]\right\} = 0$$

A2 *Regularity*: The model $\gamma \in \Gamma \to g_\gamma$ is three times differentiable and regular in the sense that for all $\gamma \in \Gamma$ the Fisher information matrix associated with the model g_γ is positive definite at γ. Denote by $D^{(3)}g_\gamma$ the array whose components are

$$\frac{\partial^3 g_\gamma}{\partial\gamma_{i_1}\partial\gamma_{i_2}\partial\gamma_{i_3}}.$$

For all $i \leq k_0$, there exists $\delta > 0$ such that

$$F_0\left(\frac{\bar{g}_{\gamma_i^0}^3}{\underline{g}_{\gamma_i^0}^3}\right) < +\infty, \quad F_0\left(\frac{\sup_{|\gamma-\gamma_i^0|\leq\delta}|\nabla g_\gamma|^3}{\underline{g}_{\gamma_i^0}^3}\right) < +\infty, \quad F_0\left(\frac{|\nabla g_{\gamma_i^0}|^4}{f_0^4}\right) < +\infty$$

$$F_0\left(\frac{\sup_{|\gamma-\gamma_i^0|\leq\delta}|D^2 g_\gamma|^2}{\underline{g}_{\gamma_i^0}^2}\right) < +\infty, \quad F_0\left(\frac{\sup_{|\gamma-\gamma_i^0|\leq\delta}|D^3 g_\gamma|}{\underline{g}_{\gamma_i^0}}\right) < +\infty$$

Assume also that for all $i = 1, ..., k_0$ $\gamma_i^0 \in \text{int}(\Gamma)$ the interior of Γ.

A3 *Integrability*: There exists $\Gamma_0 \subset \Gamma$ satisfying $\text{Leb}(\Gamma_0) > 0$, and for all $i \leq k_0$

$$d(\gamma_i^0, \Gamma_0) = \inf_{\gamma \in \Gamma_0} |\gamma - \gamma_i^0| > 0$$

and such that for all $\gamma \in \Gamma_0$,

$$F_0\left(\frac{g_\gamma^4}{f_0^4}\right) < +\infty, \quad F_0\left(\frac{g_\gamma^3}{g_{\gamma_i^0}^3}\right) < +\infty, \quad \forall i \leq k_0$$

A4 *Stronger identifiability* : For all $\mathbf{t}$ partitions of $\{1, ..., k\}$ as defined above, let $\theta \in \Theta_k$ and write θ as $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$; then

$$\begin{aligned} &(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T f'_{\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}} + \frac{1}{2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T f''_{\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) = 0 \Leftrightarrow \\ &\forall i \leq k_0, s_i = 0 \quad \& \quad \forall j \in I_i \quad q_j(\gamma_j - \gamma_i^0) = 0, \quad \forall i \geq t_{k_0} + 1 \quad p_i = 0 \end{aligned} \tag{3}$$

Assuming also that if $\gamma \notin \{\gamma_1, ..., \gamma_k\}$ then for all functions h_γ which are linear combinations of derivatives of g_γ of order less than or equal to 2 with respect to γ, and all functions h_1 which are also linear combinations of derivatives of the g_{γ_j}'s $j = 1, ..., k$ and its derivatives of order less than or equal to 2, then $\alpha h_\gamma + \beta h_1 = 0$ if and only if $\alpha h_\gamma = \beta h_1 = 0$.

Extension to non compact cases : If Γ is not compact then we also assume that for all sequences γ_n converging to a point in $\partial\Gamma$ the frontier of Γ, considered as a subset of $(\mathbb{R} \cup \{-\infty, +\infty\})^d$, g_{γ_n} converges pointwise either to a degenerate function, i.e. satisfying $\int g(x)d\mu(x) \in \{+\infty, 0\}$ or to a proper density g such that g is linearly independent of any non null combinations of $g_{\gamma_i}^0$, $\nabla g_{\gamma_i}^0$ and $D^2 g_{\gamma_i}^0$, $i = 1, ..., k_0$.

A5 *Prior* : The prior density, with respect to Lebesgue measure on Θ, is continuous and positive and the prior $\pi(p)$ on $(p_1, ..., p_k)$ satisfies

$$\pi(p) = C(p) p_1^{\alpha_1 - 1} .... p_k^{\alpha_k - 1}$$

where $C(p)$ is a continuous function on the Simplex bounded from above and from below by positive constants.

These assumptions are weaker versions of the kind of assumptions that can be found in the literature on asymptotic properties of mixture models. Assumption [A1] is quite mild and there are quite a few results in the literature proving such a consistency of the posterior for various classes of priors; see for instance Ghosal and der Vaart (2001) and Scricciolo (2001) for Gaussian mixtures or Rousseau (2007) for Beta mixtures. Since here the model corresponds to a finite k, the L_1 posterior concentration rate δ_n will typically be of order $n^{-1/2}$, i.e. $q = 0$. This will imply

sharp results on the behaviour of the posterior distribution on the weights of the *extra components*, see Theorem 1 and its comments.

Assumption [A2] is a usual regularity assumption and assumption [A3], is a weaker version than the assumptions in Liu and Shao (2004) or in Dacunha-Castelle and Gassiat (1999), since the likelihood ratio needs only be integrable on some chosen subset of Γ and not everywhere. Assumption [A4] (first part) is also weaker than in Liu and Shao (2004). It is related to the linear independence of the functions g_γ, ∇g_γ and $D^2_{r,s} g_\gamma$, $r \leq s$ and is weaker than requiring that these functions are linearly independent. Note that in the case of an overfitted mixture the compactness assumption is important, and in particular the likelihood ratio statistic is not a consistent test statistic in cases where the parameter space Γ is not compact; see Azais et al. (2006). Here, however we prove that it is not a necessary assumption and that the result remains valid when Γ is not compact under mild conditions, i.e. the second part of assumption [A4]. These conditions are in particular satisfied for most regular exponential families, including Gaussian, exponential and student mixtures if the degrees of freedom varies in a compact subset of $[1, +\infty)$, in these three cases the densities g_γ converge to degenerate functions near the boundary of the set. In the case of discrete distributions, such as Poisson mixtures, it is to be expected that the limit is still a distribution at least for some of the points of the boundary. However, the limit will often be linearly independent of the g_{γ_i}'s and their derivatives. For instance, in the case of a mixture of Poisson distributions with parameters λ, when λ goes to 0 the density converges to 0 except at $x = 0$ where it converges to 1, so that the limit is a proper distribution. However this limit is linearly independent of any function (of x) in the form $\lambda^x(a_1 + a_2 x + a_3 x^2)$ unless $a_1 = a_2 = a_3 = 0$ and [A4] is satisfied. The assumption [A5] on the prior on p is valid for instance in the case of Dirichlet priors on the weights $\mathcal{D}(\alpha_1, ..., \alpha_k)$.

2.2. Main result : asymptotic behaviour of the posterior distribution on the weights

We now state the main result.

THEOREM 1. *Under the assumptions [A1]-[A5] the posterior distribution satisfies: Let $\mathcal{S}_k$ be the set of permutations of $\{1, ..., k\}$, $\bar{\alpha} = \max(\alpha_j, j \leq k)$ and $\underline{\alpha} = \min(\alpha_j, j \leq k)$*

- *If $\bar{\alpha} < d/2$, set $\rho = (dk_0 + k_0 - 1 + \bar{\alpha}(k - k_0))/(d/2 - \bar{\alpha})$ then*

$$\lim_{M \to +\infty} limsup_n E_0^n \left\{ P^\pi \left[\min_{\sigma \in \mathcal{S}_k} \sum_{i=k_0+1}^{k} p_{\sigma(i)} > M n^{-1/2} (\log n)^{q(1+\rho)} \middle| X^n \right] \right\} = 0$$

- *If $\underline{\alpha} > d/2$ and $\rho' = (dk_0 + k_0 - 1 + d(d - k_0)/2)/((\underline{\alpha} - d/2)(k - k_0))$*

$$\lim_{\epsilon \to 0} limsup_n E_0^n \left\{ P^\pi \left[\min_{\sigma \in \mathcal{S}_k} \sum_{i=k_0+1}^{k} p_{\sigma(i)} < \epsilon (\log n)^{-q(1+\rho')} \middle| X^n \right] \right\} = 0.$$

Recall that $(\alpha_1, ..., \alpha_k)$ are the hyperparameter appearing in the prior distribution on the weights, and controlling its behaviour when some of the weights are close to 0 and that q is given in assumption [A1]. As a consequence of Theorem 1, if $\max(\alpha_j, j \leq k) < d/2$, the posterior estimates verify

$$\sum_{j=k_0+1}^{k} E^{\pi}[p_j|X^n] = O_p(n^{-1/2}(\log n)^{q(\rho+1)})$$

as n goes to infinity, under the convention that the classes are labelled such that the posterior means of the weights p_j are in decreasing order. An important special case corresponds to $q = 0$, since in that case the posterior expectation of the weights of the *extra components* is of order $O_p(n^{-1/2})$.

Hence if none of the components are small, it implies that k is probably not larger than k_0. Also in the case of longitudinal data, it is possible to choose the largest possible k for all time periods and to estimate the parameters with this value of k; the Bayesian answer would make sense and be interpretable, since at each time a component is allocated with a small weight only if it corresponds to an empty component.

In contrast, if $\min(\alpha_j, j \leq k) > d/2$ and if the number of components is larger than it should be, then 2 or more components will tend to merge with non-negligible weights each. This will lead to less stable behaviour since the weights of each of these 2 components can vary, and the selection of the components that will merge can also vary. In the intermediate case, if $\min(\alpha_j, j \leq k) \leq d/2 \leq \max(\alpha_j, j \leq k)$, then the situation varies depending on the α_j's and on the difference between k and k_0. In particular, in the case where all α_j's are equal to $d/2$ then although we have no definite result we conjecture that the posterior distribution does not have a stable limit.

One of the consequences of the above result is in the choice of the prior on the weights in mixture models. Since it is more interesting to have the posterior distribution concentrated on the configuration where the extra components receive no weights as opposed to a merging of some of the components, it is better to choose small values of the α_j's. In particular in the case of location - scale mixtures then choosing $\alpha_j < 1$ is preferable in this regard. Note that the special case of a Dirichlet $\mathcal{D}(1/2, ..., 1/2)$ which is the marginal Jeffreys prior (associated with the Multinomial model) is among such priors.

The usual case of a hierarchical mixture where the component's parameters γ_j are independently and identically distributed according to some common distribution h_η indexed by a parameter η where η is itself given a prior π_0 falls into the setup of condition [A5] since the prior mass of sets in the form $\{\gamma; |\gamma_0 - \gamma| \leq \epsilon\}$ is still equivalent to the Lebesgue measure of this set.

A possible practical use of this theorem is therefore to compute the posterior distribution in a mixture model with a rather large number of components and a Dirichlet - type prior on the weights with small hyperparameters parameters (α_j) and check for small weights in the posterior distribution. The threshold for deciding which are the small weights is case dependent. This is

illustrated in the real data analysis performed in Section 3.2.

The proof of Theorem 1 is given in the appendix, however we present some aspects of it that are of interest. In the case where $\bar{\alpha} < d/2$, writing $M_n = M(\log n)^{q(1+\rho)}$ and $A_n = \{\min_{\sigma\in\mathcal{S}_k} \sum_{i\geq k_0+1} p_{\sigma(i)} > n^{-1/2}M_n\} \cap \{\|f_0 - f_\theta\| \leq \delta_n\}$, i.e. the complementary of the event where the extra components are emptied at a rate of order slightly larger than $n^{-1/2}$. Then posterior probability of A_n can be written as

$$P^\pi [A_n|X^n] = \frac{\int_{A_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi(\theta)}{\int e^{l_n(\theta)-l_n(\theta_0)} d\pi(\theta)} := \frac{N_n}{D_n}$$

where $l_n(\theta)$ is the log-likelihood and $\theta_0 \in \Theta_0$, and we need to prove that $P^\pi [A_n|X^n] = o_p(1)$; which corresponds to the first case of Theorem 1. To do so we bound from above N_n and from below D_n. First we prove that with probability going to 1,

$$D_n \geq C n^{-(dk_0+k_0-1+\sum_{j\geq k_0+1}\alpha_{\sigma(j)})/2},$$

for any permutation σ of $\{1, ..., k\}$, by considering approximations of Θ_0 along paths of the form:

$$|\gamma_{\sigma(i)} - \gamma_i^0| \leq n^{-1/2}, \quad |p_{\sigma(i)} - p_i^0| \leq n^{-1/2} \quad \sum_{j\geq k_0+1} p_{\sigma(j)} \leq n^{-1/2}.$$

These paths correspond to the configuration where the extra components are emptied at a rate of order $n^{-1/2}$. In contrast and by definition, A_n corresponds to paths approximating Θ_0 where at least two components merge, in other words they are associated with partitions $\mathbf{t}$ of $\{1, ..., k_0\}$ such that there exists $i \leq k_0 - 1$ with $t_{i+1} \geq t_i + 2$ and at least two components merging with γ_i^0 have weights much larger than $n^{-1/2}$. We prove in the appendix that each of these paths has a prior mass bounded by $o_p(D_n)$ when $d/2 > \max\{\alpha_j, j = 1, ..., k\}$. In this case, $dk_0 + k_0 - 1 + \min_{\sigma\in\mathcal{S}_k} \sum_{j\geq k_0+1} \alpha_j$ appears as an effective dimension of the model, which is different from the number of parameters, $dk + k - 1$, or even from some "effective number of parameters" that would be given by the number of parameters used to parameterize the path $\min_{\sigma\in\mathcal{S}_k} \sum_{j\geq k_0+1} p_{\sigma(j)} = O(n^{-1/2})$, due to the influence of the prior via the α_j's. This can be understood as a measure of complexity of the model induced by the prior.

In the case where $d/2 < \min\{\alpha_j, j = 1, ..., k\}$, we need to prove that $P^\pi [B_n|X^n] = o_p(1)$, with $B_n = \{\min_{\sigma\in\mathcal{S}_k} \sum_{i\geq k_0+1} p_{\sigma(i)} < \epsilon_n\} \cap \{\|f_0 - f_\theta\| \leq \delta_n\}$ with $\epsilon_n = \epsilon(\log n)^{-q(1+\rho')}$. This is done also by bounding from below D_n and from above a numerator similar to N_n, with B_n instead on A_n in its definition. A reverse phenomenon takes place: we bound from below D_n by considering approximations of Θ_0 along paths of the following form: if $I_1 = \{1, ..., k - k_0 + 1\}$, $I_i = \{k - k_0 + i\}$,

$i = 2, ..., k_0$

$$\left| \sum_{j \in I_i} \frac{p_j}{\sum_{j \in I_i} p_j} \gamma_j - \gamma_i^0 \right| \leq n^{-1/2}, \quad |\sum_{j \in I_i} p_j - p_i^0| \leq n^{-1/2} \quad \forall j \in I_i, i = 1, ..., k_0 | \gamma_j - \gamma_i^0 | \leq n^{-1/4},$$

i.e. by forcing all the parameters of the extra components to be close to γ_1^0. This leads to

$$D_n \geq C n^{-0.5(k_0 d + k_0 - 1 + d(k-k_0)/2)},$$

with large probability whereas $\pi(B_n) = o_p(D_n)$ so that

$$P^\pi \left[B_n | X^n \right] = o_p(1).$$

An interesting feature of this argument is that it shows that the asymptotic behaviour of the posterior distribution is driven by prior mass of approximating paths to the true density f_0. This acts as a penalization factor in a way which is more subtle than the mere dimension of the parameter. This phenomenon is also observed in Rousseau (2007) in the framework of consistency of Bayes factors. It is of interest to note that the natural penalization induced by Bayesian approaches is not only crucial in testing problems but also in point estimation problems.

In the following section we conduct a simulation study to illustrate the above results but also to study the possible behaviours one could expect when $\max \alpha_j \geq d/2$. We also consider a real data analysis to present some of the ways to use Theorem 1 in practice, a good reference for such uses is also Frühwirth-Schnatter (2006).

3. Examples

We first consider a simulation study illustrating the theoretical results . Two cases are considered: $d/2 < \min(\alpha_j)$ and $d/2 > \max(\alpha_j)$.

3.1. Simulated example

We consider a very simple study of fitting a two-component Gaussian mixture model to a sample of data, $Y = \{y_i, i = 1, .., n\}$, generated from a single-component Gaussian distribution, say $\mathcal{N}_q(0, 1)$, which the q dimensional vector whose components are independent and identically distributed standard Gaussian random variables; to simplify notations $\mathcal{N}(0, 1) = \mathcal{N}_1(0, 1)$. Note that assumptions [A1]-[A5] are satisfied in the case of location mixtures of Gaussians and location-scale mixtures of Gaussians. In particular, condition [A1] has been proved by Ghosal and van der Vaart (2006), [A2]-[A4] are weaker versions of the hypothesis required in Chambaz and Rousseau (2008) and are therefore satisfied for both types of mixtures of Gaussians. We consider for p a Uniform prior on $[0, 1]$ and the component's parameters are assumed independent and identically distributed with

a $\mathcal{N}(0, 10^4)$ prior on the means and a Uniform $\mathcal{U}(0, 100)$ on the variances, so that [A5] is also satisfied. In all cases, we computed the posterior distribution for M replications of samples of sizes $n = 50, 100, 500, 1000, 5000, 10000$ of standard Gaussian random variables $\mathcal{N}(0, 1)$, where $M = 50$ for the sample sizes $n = 50, 100, 500$ and $M = 20$ for $n = 1000, 5000, 10000$. The figures show boxplots of the posterior means of the p, where p is the largest among the two possible weights (i.e. $p > 1 - p$). In the following description G denotes the model to be estimated. We consider three cases corresponding to the dimensions $d = 1, 4, 2$ respectively:

- Case 1 : $d = 1$ and $\alpha_1 = \alpha_2 = 1 > d/2$: location mixture of univariate Gaussian distributions with fixed variance.

$$\begin{aligned} \text{Generating distribution :} \qquad & y_i \sim \mathcal{N}(0, 1) \in \mathbb{R} \\ \text{Model} \qquad & G = p\mathcal{N}(\mu_1, 1) + (1 - p)\mathcal{N}(\mu_2, 1), \end{aligned}$$

  where $\mathcal{N}(\mu, \tau)$ denotes the univariate normal distribution with mean μ and variance τ. In this case Theorem 1 implies that $P^\pi [p > 1 - \epsilon | X^n]$ $(p > 1 - p)$ is small as n gets large for ϵ small enough but fixed. Thus although we expect to see a component with smaller weight, this weight should not go to 0, we also expect to see the two means μ_1 and μ_2 to become closer and closer to 1.

- Case 2 : $d = 4$ and $\alpha_1 = \alpha_2 = 1 < d/2$: location - scale mixture of bivariate Gaussian distributions.

$$\begin{aligned} \text{Generating distribution :} \qquad & y_i \sim \mathcal{N}_2(0, 1) \in \mathbb{R}^2 \\ \text{Model} \qquad & G = p\mathcal{N}_2(\mu_1, \Sigma_1^2) + (1 - p)\mathcal{N}_2(\mu_2, \Sigma_2^2) \end{aligned}$$

  where $\mu_j = (\mu_{j1}, \mu_{j2})'$ and Σ_j is the diagonal matrix with diagonal elements $(\sigma_{j1}^2, \sigma_{j2}^2)$. In this case we expect the posterior mean of p $(p > 1 - p)$ to increase to 1 as n increases. The convergence being quite noticable we have run simulations up to $n = 5000$ only.

- Case 3 : $d = 2$ and $\alpha_1 = \alpha_2 = 1 = d/2$: location-scale mixture of univariate Gaussian distributions.

$$\begin{aligned} \text{Generating distribution :} \qquad & y_i \sim \mathcal{N}(0, 1) \in \mathbb{R} \\ \text{Model} \qquad & G = p\mathcal{N}(\mu_1, \sigma_1^2) + (1 - p)\mathcal{N}(\mu_2, \sigma_2^2), \end{aligned}$$

  This is the case where there is no theoretical answer.

The empirical findings support the theoretical asymptotic behaviour described in the previous section: for $d = 1$ (case 1) the posterior distribution of the weights is unstable, even with increasing

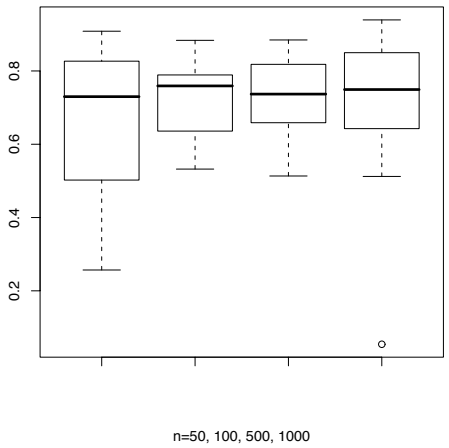

Figure 1. boxplot of posterior means of p in the case $d = 1 < 2\alpha$

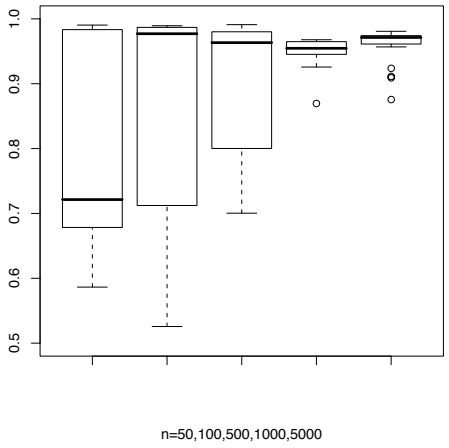

Figure 2. boxplot of posterior means of p in the case $d = 4 > 2\alpha$

sample size, see Figure 1. The means of the components become closer and closer to 0 as the sample size increases. On the contrary, when $d = 4$ (case 2) one component becomes effectively empty as $n \geq 1000$, see Figure 2. In the case where $d = 2 = 2\alpha$ (case 3) the posterior expectation of p does not seem to clearly converge to 1, or if it did it would be very slowly since at $n = 10000$ we still observe a large proportion of posterior means of p to be less than 0.95. Still there is a large difference between $n = 50$ and $n \geq 100$ and for larger values of n, the posterior means of p seem to be reasonably close to 1, see Figure 3.

3.2. Case Study: School performance in maths and science

The performance of school students in maths and science is a key indicator for educators as well as governments across the world. It is often of interest to identify whether subgroups of students

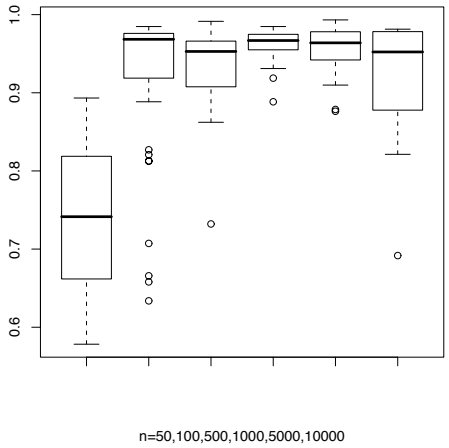

Figure 3. boxplot of posterior means of p in the case $d = 2 = 2\alpha$

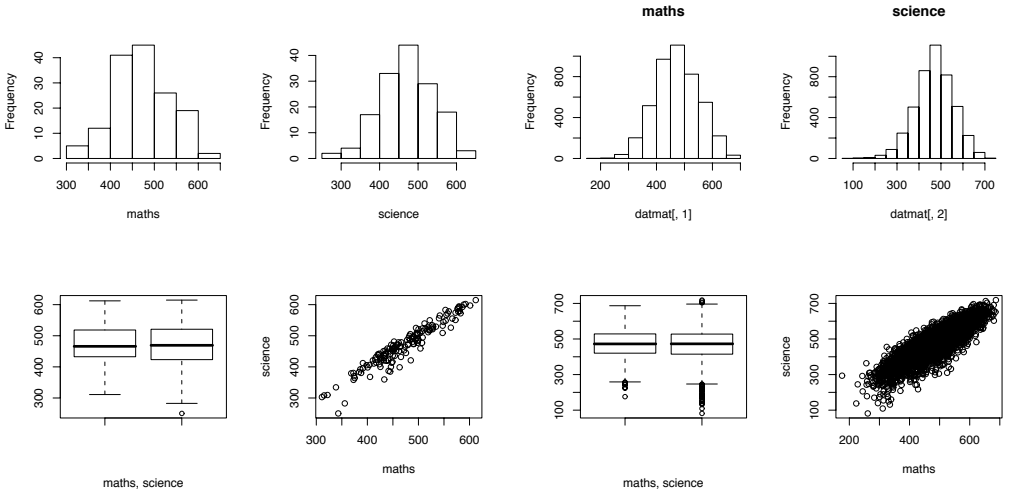

Figure 4. scores in math and science. LEFT (columns $1\&2$): at the school level ; RIGHT (columns $3\&4$) : at the student level

or schools can be identified on the basis of common tests in maths and science, and the features of such subgroups if they exist. In the study considered here, scores on common maths and science tests were obtained for 4500 students in 150 schools across a single country. A visual representation of the data is provided in Figure 4.

This dataset closely reflects the simulation study, since both univariate and multivariate (bivariate) analyses can be undertaken, at both the school level (averaging over students), for which the sample size is modest ($n = 150$) and at the student level (ignoring schools), for which the sample size better reflects the asymptotic situation. Of course, for a formal analysis of these data, a mixed model would be more appropriate, with students nested within schools. Although this hierarchical structure is easily accommodated in the Bayesian framework, it is also arguable that such a model might reduces the ability to classify students across schools into different subgroups

based on their test results. The univariate analyses focused on the maths scores. If y denotes the math score for either a school or a student, the model for y is $y \sim \sum_{j=1}^{k} p_j \mathcal{N}(\mu_j, \sigma_j^2)$. At the school level, we let y_i denote the average maths score for the ith school, $i = 1, .., n = 150$ with sample mean 473.0 and sample standard deviation 63.5 and at the student level y_i is the score of a student, $i = 1, ..., n = 4500$. In all instances the prior on the μ's were diffuse normal priors : $\mathcal{N}(0, 10000)$ and the priors on the σ's were $\mathcal{U}(0, 100)$.

First, a single normal model was fitted using a diffuse normal prior. The posterior estimates of μ and σ were then 471 (s.d. 5.213) and 64.09 (s.d. 3.76), respectively. Second, a two-component mixture model was fitted, assuming that the component means, variances and weights are unknown ($d = 2$). A Dirichlet ($\alpha_1 = \alpha_2 = 1 = d/2$) prior was placed on (p_1, p_2). We obtained the following posterior estimates and 95% credible intervals for the parameters $p_1 = 0.07$, $(4.610^{-4}, 0.62)$, $p_2 = 0.93$ $(0.38, 1.0)$, $\mu_1 = 315.3$ $(128.7, 451.4)$, $\mu_2 = 476.4$ $(463.5, 499.0)$, $\sigma_1 = 49.7$ $(4.0, 96.8)$, $\sigma_2 = 62.5$ $(52.4, 72.2)$. An analysis with a Dirichlet $(1/2, 1/2)$ prior on the wieghts was conducted leading to similar results. An analogous three-component normal mixture with Dirichlet $(1, 1, 1)$ prior on the weights produced estimates (with 95% credible intervals) $p_1 = 0.07$ $(4.410^{-4}, 0.56)$, $p_2 = 0.36$ $(4.210^{-3}, 0.94)$, $p_3 = 0.58$ $(0.03, 0.99)$, $\mu_1 = 301.5$ $(114.6, 454.9)$, $\mu_2 = 436.9$ $(467.2, 581.8)$, $\mu_3 = 504.2$ $(467.2, 581.8)$, $\sigma_1 = 49.1$ $(3.99, 96.68)$, $\sigma_2 = 49.7$ $(8.3, 91.0)$, $\sigma_3 = 53.2$ $(9.7, 76.8)$. Fitting the model with a Dirichlet $(1/2, 1/2, 1/2)$ lead to posterior means and 95% credible intervals : $p_1 = 0.007$ $(4.710^{-6}, 0.04)$, $p_2 = 0.20$, $(4.710^{-5}, 0.98)$ and $p_3 = 0.80$, $(8.510^{-3}, 0.98)$. Given the large credible intervals of the posterior distribution, it is difficult, in this case to make any statement about a possible over-fitted model, even for $k = 3$ under a Dirichlet $(1, 1, 1)$ however the results under the Dirichlet $(1/2, 1/2, 1/2)$ suggest that a two component mixture model would be sufficient to represent the data. We also tried a four component mixture model with a prior Dirichlet(1,1,1,1), but again the credible intervals were considerably overlapping.

Continuing the univariate analysis, at the student level, the sample mean and standard deviation of the $n = 4500$ students' maths scores were 474.4 and 78.26, respectively. Assuming first that the standard deviation was known, the posterior estimates of p_1 and p_2 were 0.53 and 0.46. An analogous three-component normal mixture gave posterior estimates for p_1, p_2 and p_3 of 0.23, 0.51 and 0.26. Fitting a two-component normal mixture (location-scale) gave estimates (with 95% credible intervals) for p_1 and p_2 of 0.90 (0.85, 0.93) and 0.1 (0.060, 0.15). A three-component normal mixture produced estimates for p_1, p_2 and p_3 of 0.029 $(1.310^{-3}, 9.910^{-2})$, 0.72 (0.35, 0.92) and 0.25 (0.073, 0.64). Interestingly, although we are in the case $d/2 = \alpha_j$, for which we have no asymptotical result, the posterior distribution puts most of its mass on the configuration with one empty and two non empty components. The credible intervals, compared to those obtained at the school level, are much narrower, indicating that an adequate model would be two-component mixture model, this is in agreement with the bivariate analyses.

The bivariate analyses included both the maths and science scores. Thus at the school level,

(y_{i1}, y_{i2}) denotes the average maths and science scores for the ith school, $i = 1, .., 150$, and $(y_1, y_2)' \sim \mathcal{N}_2((\mu_1, \mu_2)', \Sigma)$. Fitting a two-component normal mixture with component means, weights and diagonal variance-covariance matrix unknown gave p_1 and p_2 as 0.93 (0.85,0.98) and 0.073 $(210^{-3}, 0.25)$. The analogous three-component normal mixture produced p_1, p_2 and p_3 of 0.81(0.68,0.90), 0.19 (0.09,0.31) and 0.0067 $(4.410^{-7}, 4.310^{-2})$. In the bivariate anlayses of the student-level data, fitting a two-component mixture gave estimates for p_1 and p_2 of 0.89 (0.76-0.98) and 0.11, (0.008-0.24) and a three-component mixture gave estimates of p_1, p_2 and p_3 of 0.54 (0.45, 0.61) , 0.40 (0.31, 0.49) and 0.06 (0.03, 0.12). These analyses broadly confirm the results proposed in this paper. Visual assesment of Figure 4 suggests that a single normal distribution adequately described the school-based data.

The student-based analyses assuming a known variance represents the first situation described in the Theorem, in which asymptotically the superfluous components will be non-empty and consequently exhibit poor convergence. On the contrary the bivariate analysis at the student level strongly indicates that there are at most 2 components in the mixture since the weight of the third component is very small with large probability. The existence of a second component seems confirmed at least for the student-level bivariate data, but it is quite possible that both components are close.

The school univariate data seem to follow the theory, at least as far as point estimates are concerned since with known variances ($d = 1 < 2\alpha$) we have balanced weights when $k = 2$ and 3 non negligible weights when k=3 whereas with unkown mean and variance the analyses represented the middle situation, for which we do not have results, and we observe that the extra- components seem empty asymptotically although this process may be slow. However the credible regions are large, which makes it difficult to have a definite answer based only on these data.

The bivariate plots suggested that a two-component mixture might be required to describe the slight irregularity in the tail of the distribution; this was particularly noticeable for the full student-level dataset. These bivariate analyses represented the third situation, in which excess components were expected to empty out. This was indeed observed for the three-component mixtures fitted to both the school-level and the student-level data.

4. Discussion

This paper has contributed to an increased understanding of an important problem in mixture modelling, namely the concern about the impact of over-fitting the number of components in the mixture. This practice is ubiquitous and its impact is felt both in situations in which the mixture components and associated parameters are literally interpreted, and in situations in which the mixture is used as a convenient model-fitting framework.

The results presented in this paper contribute to the partial solutions provided in previous literature by describing the asymptotic behaviour of the posterior distribution when the typical

additive mixture distribution is over-fitted. The main consistency result indicates that the posterior distribution concentrates on a sparse representation of the true density; this is exhibited by a subset of components that adequately describe the density remaining well described and any superfluous components becoming empty. Estimators based on the posterior distribution thus exhibit quite stable behaviour in the presence of over-fitting, as opposed to alternatives such as the maximum likelihood estimator which can be quite unstable in this situation.

Importantly, the asymptotic behaviour appears to depend on the dimension of the mixture parameters in relation with the form of the prior distribution on the weights, in particular in cases of low dimensional parameters γ ($d \leq 2$) it becomes necessary to *favour* small weights with a prior in the form $p_1^{-1/2}...p_k^{-1/2}$, which interestingly corresponds to the noninformative prior in a multinomial model. It thus appears that in this subtle framework, the prior has an impact to first order since the asymptotic behaviour of the posterior distribution depends heavily on the form of the prior.

These results thus provide practical guidelines for the cases that they address. Overfitted mixtures can thus be used as an alternative to estimating the number of components and it also provides some guidelines as to the choice of the prior distribution.

The paper has also identified cases for which further research is required, such as the intermediate case where $\min(\alpha_j) \leq d/2 \leq \max(\alpha_j)$, for which no description of the asymptotic behaviour of the posterior distribution is obtained.

References

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. pp. 267–281. Second International Symposium on Information Theory.

Azais, J.-M., E. Gassiat, and C. Mercadier (2006). Asymptotic distribution and power of the likelihood ratio test for mixtures: bounded and unbounded case. *Bernoulli 12*, 775–799.

Chambaz, A. and J. Rousseau (2008). Bounds for Bayesian order identification with application to mixtures. *Ann. Statist. 36*, 938–962.

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

Dempster, A., N. Laird, and D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Statist. Society Series B 39*, 1–38.

Feng, Z. and C. McCulloch (1996). Using bootstrap likelihood ratio in finite mixture models. *J. Royal Statist. Society Series B 58*, 609–617.

Frühwirth-Schnatter, S. (2006). *Finite Mixture and Markov Switching Models*. New York: Springer-Verlag, New York.

Ghosal, S. and A. V. der Vaart (2001). Entropies and rates of convergence for maximum likelihood and bayes estimation for mixtures of normal densities. *Ann. Statist. 29*, 1233–1263.

Ghosal, S. and A. van der Vaart (2006). Convergence rates of posterior distributions for non iid observations. *Ann. Statist. 35*(1), 192–223.

Lee, K., J.-M. Marin, K. Mengersen, and C. Robert (2008). Bayesian inference on mixtures of distributions. In N. N. Sastry (Ed.), *Platinum Jubilee of the Indian Statistical Institute*. Indian Statistical Institute, Bangalore.

Liu, X. and Y. Shao (2004). Asymptotics for likelihood ratio tests under loss of identifiability. *Ann. Statist. 31*, 807–832.

MacLachlan, G. and D. Peel (2000). *Finite Mixture Models*. New York: John Wiley.

Marin, J.-M. and C. Robert (2007). *Bayesian Core*. Springer-Verlag, New York.

McGrory, C. and D. Titterington (2007). Variational approximations in bayesian model selection for finite mixture distributions. *Computational Statistics and Data Analysis 51*, 5352–5367.

Richardson, S. and P. Green (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Royal Statist. Society Series B 59*, 731–792.

Robert, C. and D. Wraith (2009). Computational methods for Bayesian model choice. In A. I. of Physics (Ed.), *MaxEnt 2009 proceedings*. (To appear.).

Rousseau, J. (2007). Approximating interval hypotheses: p-values and Bayes factors. In J. M. Bernardo, J. O. Berger, A. P. Dawid, and A. F. M. Smith (Eds.), *Bayesian Statistics 8: Proceedings of the Eigth International Meeting*. Oxford University Press.

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Statist. 6*, 461–464.

Scricciolo, C. (2001). Convergence rates of posterior distributions for dirichlet mixtures of normal densities. Technical report, University of Padova.

Titterington, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley.

5. Appendix: Proof of Theorem 1

Set $A_n' = \{\exists I = \{j_1, ..., j_{k-k_0}\}, \sum_{i\in I} p_i > M_n\delta_n\}$ for some large M_n (depending on δ_n M_n will be chosen as a power of $\log n$ or a a large constant. The posterior probability of interest is bounded by

$$\begin{aligned} P^\pi\left[A_n^{'}|X^n\right] &= P^\pi\left[A_n^{'} \cap \{\|f - f_0\| \leq M\delta_n\}|X^n\right] + o_P(1) \\ &= \frac{\int_{A_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta)}{\int_{\|f_0-f_\theta\|\leq M\delta_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta)} + o_P(1) \end{aligned}$$

where $A_n = A_n^{'} \cap \{\|f - f_0\| \leq M\delta_n\}$ and M is a fixed positive constant. We denote by

$$N_n = \int_{A_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta) \quad \text{and} \quad D_n = \int_{\|f_0-f_\theta\|\leq M\delta_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta). \tag{4}$$

To prove the first part of Theorem 1 we first prove if $\max_j \alpha_j < d/2$ that for all $\epsilon > 0$, there exists C_ϵ and there exists a permutation $\sigma : \{1, ..., k\} \to \{1, ..., k\}$, s.t.

$$P_0^n\left[D_n \geq C_\epsilon n^{-\frac{dk_0+k_0-1+\sum_{j=k_0+1}^k \alpha_{\sigma(j)}}{2}}\right] > 1 - \epsilon, \quad \pi(A_n) \leq C\frac{\delta_n^{dk_0+k_0-1+\sum_{j=k_0+1}^k \alpha_{\sigma(j)}}}{M_n^{(d/2-\max_j \alpha_j)}} \tag{5}$$

The combination of these two inequalties implies that for all $\epsilon > 0$, with probability larger than $1 - \epsilon$,

$$E_0^n\left[P^\pi\left[A_n|X^n\right]\right] = o(1)$$

which terminates the proof of the first part of Theorem 1. Similarly if $d/2 < \min\{\alpha_j, j = 1, ..., k\}$, we obtain with $M_n = \epsilon_n\delta_n^{-1}$, where ϵ_n is either a small constant or a sequence converging to 0 at the rate a power of $(\log n)^{-1}$,

$$P_0^n\left[D_n \geq C_\epsilon n^{-(dk_0+k_0-1+d(k-k_0)/2)/2}\right] > 1 - \epsilon, \quad \pi(A_n^c) \leq C\max_\sigma \delta_n^{dk_0+k_0-1+\sum_{j=k_0+1}^k \alpha_{\sigma(j)}} \tag{6}$$

which leads to

$$E_0^n\left[P^\pi\left[A_n^c|X^n\right]\right] = o(1)\ .$$

We now establish (5) and (6). We start with (5). Throughout the proof we write all constants whose value are of no consequence to be equal to 1. First

$$D_n \geq \int_{S_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta)$$

where

$$S_n = \{(p_1, ..., p_k, \gamma_1, ..., \gamma_k); |p_j - p_j^0| \leq n^{-1/2}; |\gamma_j - \gamma_j^0| \leq n^{-1/2}, j = 1, ..., k_0; |\gamma_j - \gamma_j^*| \leq \epsilon_1, j \geq k_0+1\}$$

where $\gamma_j^* \in \Gamma_0$, $j \geq k_0+1$ and satisfy $\min_{k_0<l\neq j} |\gamma_j^* - \gamma_l^*| > C\epsilon_1$, with $C, \epsilon_1 > 0$ fixed. By definition of Γ_0, $\min_{l\leq k_0} |\gamma_j^* - \gamma_l^0| > C\epsilon_1$ and by definition of S_n, $\sum_{j\geq k_0+1} p_j \leq k_0 n^{-1/2}$. Such a path to approach Θ_0 corresponds to the partition $\mathbf{t} = (0, 1, 2, ..., k_0)$. Let $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ be the parameterisation (of θ) associated to the partition $\mathbf{t}$. We consider a Taylor expansion of $l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) - l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})$, corresponding to $\theta = (\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) \in S_n$ and $\theta_0 = (\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})$. By convention and without loss of generality we write $p_j^0 = 0$ and $\gamma_j^0 = \gamma_j^*$ for $j = k_0 + 1, ..., k$ and following the definition of $\phi_{\mathbf{t}}$, $p_{k_0} = 1 - \sum_{i\neq k_0} p_i$. Then

$$l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) - l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}) = \sqrt{n}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T U_n - \frac{n}{2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T J(\bar{\theta})(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) \quad (7)$$

where $J(\bar{\theta}) = -\partial^2 l_n(\bar{\phi}_{\mathbf{t}}, \psi_{\mathbf{t}})/\partial\phi_{\mathbf{t}}\partial\phi_{\mathbf{t}}^T$, $\bar{\phi}_{\mathbf{t}} \in (\phi_{\mathbf{t}}, \phi_{\mathbf{t}}^0)$. and

$$\begin{aligned} U_n(i) &= \mathbb{G}_n\left(\frac{\nabla_l g_{\gamma_j^0}}{f_0}\right), \quad i = l + (j-1)*d, \quad j \leq k_0 \quad l = 1, ..., d \\ U_n(i) &= \mathbb{G}_n\left(\frac{f_{\gamma_j^0} - f_{\gamma_1^0}}{f_0}\right), \quad i = k_0 d + j - \mathbb{1}_{j\geq k_0+1}, \quad 1 \leq j \neq k_0 \leq k \end{aligned}$$

then $U_n = 0_p(1)$. Denote id the identity matrix and $\Omega_n(c_0, C) = \{X^n; \sup_{\theta\in S_n} \|J(\theta)\| \leq c_0 n; |U_n| \leq C\}$. Therefore the log likelihood ratio is bounded from below by $\sqrt{n}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T U_n - \frac{C_0 n}{2}\|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0\|^2$ for some positive constant C_0 on $\Omega_n(c_0, C)$. This leads to, on $\Omega_n(c_0, C)$:

$$\begin{aligned} \int_{S_n} e^{l_n(\theta)-l_n(\theta_0)} d\pi_k(\theta) &\geq e^{\frac{1}{2C_0}\|U_n\|^2} \int_{S_n} e^{-\frac{nC_0}{2}\|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1}U_n/\sqrt{n}\|^2} d\pi_k(\theta) \\ &\geq \int_{S_n} e^{-\frac{nC_0}{2}\|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1}U_n/\sqrt{n}\|^2} d\pi_k(\theta) \ . \end{aligned}$$

Recall that on S_n, $p_j \geq 0$ for $j \geq k_0 + 1$. Using assumption [A5] we can bound from below $\pi_k(\theta)$ by $c_1 p_{k_0+1}^{\alpha_{k_0+1}-1} ... p_k^{\alpha_k - 1}$. Thus on $\Omega_n(c_0, C)$, we have

$$\begin{aligned} \int_{S_n} e^{-\frac{C_0 n}{2}\|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1}U_n/\sqrt{n}\|^2} \pi_k(\theta) d\theta &\geq n^{-(dk_0+k_0-1)/2} \prod_{j=k_0+1}^{k} \int_0^{\delta_n/k_0} e^{-\frac{nC_0}{2}(p_j - U_n(j))^2} p_j^{\alpha_j - 1} dp_j \\ &\geq n^{-(dk_0+k_0-1+\sum_{j>k_0}(\alpha_j-1))/2} \prod_{j=k_0+1}^{k} \int_{c/\sqrt{n}}^{\delta_n/k_0} e^{-\frac{nC_0}{2}(p_j - U_n(j))^2} dp_j \\ &\geq n^{-(dk_0+k_0-1+\sum_{j>k_0}\alpha_j)/2} \prod_{j=k_0+1}^{k} \Phi(C_1) \\ &\geq n^{-(dk_0+k_0-1+\sum_{j>k_0}\alpha_j)/2} \end{aligned}$$

where $c > 0$ is chosen small enough and where C_1 depends on C, c, C_0. To obtain the best lower bound we can choose the permutation σ^* which minimizes $\sum_{j>k_0} \alpha_{\sigma(j)}$ which we set equal to the

identity to simplify notations and if $a > 0$ is small enough

$$P_0^n \left[D_n < a n^{-(dk_0+k_0-1+\sum_{j>k_0}\alpha_j)/2} \right] \quad \leq \quad P_0^n \left[\Omega_n^c(c_0, C) \right].$$

The lower bound in (5) is then proved by determining an upper bound on $P_0^n \left[\Omega_n^c(c_0, C) \right]$. Note first that for all $\epsilon > 0$ there exists $C > 0$ such that with probability greater than $1 - \epsilon$, $|U_n| \leq C$. Then we bound for each $i, i' \leq k - 1 + k_0 d$, and some $c > 0$ small enough $P_0^n \left[J(i, i') - nI(i, i') < cn \right]$, where I is a Fisher information matrix defined as $E_0^n \left[J(\theta_0) \right]$ with the constraint here that $\theta_0 = (p_1^0, ..., p_{k_0}^0, 0, ..., 0, \gamma_1^0, ...\gamma_k^0)$ writing $\gamma_j^0 = \gamma_j^*$ when $j = k_0 + 1, ..., k$. We have if $i, i' \geq dk_0 + 1$,

$$J(i, i') - nI(i, i') \quad = \quad \sqrt{n}\mathbb{G}_n \left(\frac{(g_{\gamma_j^0} - g_{\gamma_{k_0}^0})(g_{\gamma_{j'}^0} - g_{\gamma_{k_0}^0})}{f_0^2} \right) + n\mathbb{P}_n \left[\Delta_{\bar{\theta}}(i, i') \right]$$

with j, j' the indices corresponding to i, i' as in the definition of U_n,

$$\Delta_{\bar{\theta}}(i, i') = \left(\frac{(g_{\bar{\gamma}_j} - g_{\bar{\gamma}_{k_0}})(g_{\bar{\gamma}_{j'}} - g_{\bar{\gamma}_{k_0}})}{f_{\bar{\theta}}^2} \right) - \left(\frac{(g_{\gamma_j^0} - g_{\gamma_{k_0}^0})(g_{\gamma_{j'}^0} - g_{\gamma_{k_0}^0})}{f_0^2} \right)$$

Using a Tchebychev inequality the first term is less than $nc/2$ with probability

$$Cn^{-1} F_0 \left[\left(\frac{(g_{\gamma_j^0} - g_{\gamma_{k_0}^0})(g_{\gamma_{j'}^0} - g_{\gamma_{k_0}^0})}{f_0^2} \right)^2 \right] \leq \frac{Ck_0}{n \min_{l \leq k_0} (p_l^0)^4} + n^{-1} \max_{\gamma \in \Gamma_0} F_0 \left(\frac{g_\gamma^4}{f_0^4} \right)$$

Assumption [A3] implies that the second term on the right hand side of the above inequality is of order $O(n^{-1})$, so that the above probability is $O(n^{-1})$. To study the behaviour of $\Delta_\theta(i, i')$ we consider its derivatives : if $i, i' > dk_0$,

$$\begin{aligned} \left| \frac{\partial \Delta_\theta(i, i')}{\partial p_l} \right| \quad &= \quad \left| \frac{(g_{\gamma_j} - g_{\gamma_{k_0}})(g_{\gamma_{j'}} - g_{\gamma_{k_0}})(g_{\gamma_l} - g_{\gamma_{k_0}})}{f_\theta^3} \right| \\ &\leq \quad \frac{[\bar{g}_{\gamma_j^0} + \bar{g}_{\gamma_{k_0}^0}][\bar{g}_{\gamma_{j'}^0} + \bar{g}_{\gamma_{k_0}^0}][\bar{g}_{\gamma_l^0} + \bar{g}_{\gamma_{k_0}^0}]}{(1 - \delta_n)^3 \left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0} \right)^3} \end{aligned}$$

and if $i, i' > dk_0$, $l \leq k$,

$$\begin{aligned}\left|\frac{\partial \Delta_\theta(i,i')}{\partial \gamma_l}\right| &= \left|\frac{(g_{\gamma_j}-g_{\gamma_{k_0}})(g_{\gamma_{j'}}-g_{\gamma_{k_0}})\nabla g_{\gamma_l}}{f_\theta^3} + \mathbb{1}_{j=l}\frac{\nabla g_{\gamma_j}(g_{\gamma_{i'}}-g_{\gamma_{k_0}})}{f_\theta^2} + \mathbb{1}_{l=j'}\frac{\nabla g_{\gamma_{j'}}(g_{\gamma_j}-g_{\gamma_{k_0}})}{f_\theta^2}\right. \\ &\quad \left. - \mathbb{1}_{l=k_0}\frac{\nabla g_{\gamma_{k_0}}(g_{\gamma_{j'}}+g_{\gamma_j}-2g_{\gamma_{k_0}})}{f_\theta^2}\right| \\ &\leq \frac{[\bar{g}_{\gamma_j^0}+\bar{g}_{\gamma_{k_0}^0}][\bar{g}_{\gamma_{j'}^0}+\bar{g}_{\gamma_{k_0}^0}]\sup_{|\gamma-\gamma_l^0|\leq\delta}|\nabla g_\gamma|}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^3} + \mathbb{1}_{j=l}\frac{\sup_{|\gamma-\gamma_j^0|\leq\delta}|\nabla g_\gamma|(\bar{h}_{\gamma_{j'}^0}+\bar{g}_{\gamma_{k_0}^0})}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^2} \\ &\quad + \mathbb{1}_{l=j'}\frac{\sup_{|\gamma-\gamma_{j'}^0|\leq\delta}|\nabla g_\gamma|(\bar{h}_{\gamma_i^0}+\bar{g}_{\gamma_{k_0}^0})}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^2} + \mathbb{1}_{l=k_0}\frac{\sup_{|\gamma-\gamma_{k_0}^0|\leq\delta}|\nabla g_{\gamma_{k_0}}|(\bar{g}_{\gamma_j^0}+\bar{g}_{\gamma_{j'}^0}+2\bar{g}_{\gamma_{k_0}^0})}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^2}\end{aligned}$$

Assumptions [A2] and [A3] imply that there exists $\delta, M > 0$ such that

$$F_0\left(\sup_{\theta\in S_n}\left|\frac{\partial\Delta_\gamma(i,i')}{\partial\gamma_l}\right|\right) \leq M < +\infty \quad \forall l \leq k, \forall i, i' > dk_0,$$

so that for all $c > 0$, there exist δ_0 such that for all $\delta < \delta_0$,

$$P_0\left[\mathbb{P}_n\left|\sup_{\theta\in S_n}|\Delta(i,i')|\right| > c\right] \leq \frac{\delta M}{c},$$

which can be made as small as necessary. Similarly if $i > dk_0$ and $i' \leq dk_0$,

$$J(i,i') - nI(i,i') = \sqrt{n}\mathbb{G}_n\left(\frac{(g_{\gamma_j^0}-g_{\gamma_{k_0}^0})\nabla g_{\gamma_{j'}^0}}{f_0^2}\right) + n\mathbb{P}_n\left[\Delta_{\bar{\theta}}(i,i')\right]$$

with

$$\Delta_{\bar{\theta}}(i,i') = \frac{(g_{\bar{\gamma}_j}-g_{\gamma_{k_0}})\nabla g_{\bar{\gamma}_{j'}}}{f_{\bar{\theta}}^2} - \frac{(g_{\gamma_j^0}-g_{\gamma_{k_0}^0})\nabla g_{\gamma_{j'}^0}}{f_0^2}$$

Assumptions [A2] and [A3] imply that using a Tchebychev inequality $|J(i,i') - nI(i,i')| < cn$ for all $c > 0$ with probability of order $o(1)$. Also looking at the derivative of $\Delta_\theta(i,i')$ we obtain an upper bound with terms of the form

$$\frac{\sup_{|\gamma-\gamma_j^0|\leq\delta}|\nabla g_\gamma|\sup_{|\gamma-\gamma_{j'}^0|\leq\delta}|\nabla g_\gamma|}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^2}, \quad \frac{\sup_{|\gamma-\gamma_j^0|\leq\delta}|D^2 g_\gamma|(\bar{g}(\gamma_{j'}^0)+\bar{g}_{\gamma_{k_0}^0})}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^2}$$

and

$$\frac{\sup_{|\gamma-\gamma_{j'}^0|\leq\delta}|\nabla g_\gamma|[\bar{g}_{\gamma_i^0}+\bar{g}_{\gamma_{k_0}^0}][\bar{g}_{\gamma_j^0}+\bar{g}_{\gamma_{k_0}^0}]}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^3}, \quad \frac{\sup_{|\gamma-\gamma_{j'}^0|\leq\delta}|\nabla g_\gamma|\sup_{|\gamma-\gamma_j^0|\leq\delta}|\nabla g_\gamma|[\bar{g}_{\gamma_i^0}+\bar{g}_{\gamma_{k_0}^0}]}{(1-\delta_n)^3\left(\sum_{j=1}^k p_j^0 \underline{g}_{\gamma_j^0}\right)^3}$$

so that

$$P_0^n \left[|\mathbb{P}_n \left[\sup_{\theta \in S_n} |\Delta(i,i')| \right] | < c \right] \leq C\delta/c.$$

The same calculations can be made for the terms $J(i,i')$ when $i, i' \leq dk_0$, so that finally there exists $c_0, C > 0$ such that for all $\theta \in S_n$ $P_0^n \left[\Omega_n^c(c_0, C)\right] \leq 2\epsilon$ and the lower bound of D_n in (5) is established.

To bound $\pi(A_n)$ in (5), we need to characterize $\theta \in A_n$. For each $\theta \in A_n$, as n increases θ converges towards $\Theta_0 = \cup_{\mathbf{t}} \Theta_{0\mathbf{t}}$, i.e. $\min_{\sigma,\mathbf{t}} |\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}| \to 0$, where the minimum is taken over the set of possible permutations of $\{1, ..., k\}$ and partitions $\mathbf{t}$. By convention we assume that for all $\theta_0 \in \Theta_0$, $\theta_0 \in \Theta_{0\mathbf{t}}$ if for all $j \geq t_{k_0} + 1$, $\gamma_j \notin \{\gamma^0_1, ..., \gamma^0_{k_0}\}$. Consider a partition $\mathbf{t}$ and a permutation σ which minimizes $|\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}|$, for a given θ. A Taylor expansion of f_θ in terms of $\phi_{\mathbf{t}}$ around $\phi^0_{\mathbf{t}}$ leads to:

$$f_\theta = f_0 + (\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}})^T f^{'}_{(\phi^0_{\mathbf{t}}, \psi_{\mathbf{t}})} + \frac{1}{2}(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}})^T f^{"}_{(\phi^0_{\mathbf{t}}, \psi_{\mathbf{t}})}(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}) + \frac{1}{6}(\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}})^{(3)} f^{(3)}_{(\bar{\phi}_{\mathbf{t}}, \psi_{\mathbf{t}})}$$

where $\bar{\phi}_{\mathbf{t}} \in (\phi_{\mathbf{t}}, \phi^0_{\mathbf{t}})$. The last term of right hand side of the above equation is bounded by $C|\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}|^3$ in L_1 for some positive constant $C > 0$, it is thus $o(|\phi_{\mathbf{t}} - \phi^0_{\mathbf{t}}|^2)$. Define then

$$\begin{aligned} \gamma &= \gamma(\mathbf{t}, \sigma) = \left([(s_i + p^0_i)(\sum_{j \in I_i} q_j \gamma_j - \gamma^0_i)]_{i=1}^{k_0}, (s_i)_{i=2}^{k_0}, (p_j)_{j=t_{k_0}+1}^{k} \right) \\ \lambda &= \lambda(\mathbf{t}, \sigma) = \left([\sqrt{q_j}(\gamma_j - \gamma^0_i)]_{j \in I_i}, i = 1, ..., k_0 \right). \end{aligned}$$

then $(2|\gamma| + |\lambda|^2)(\mathbf{t}, \sigma)$ goes to zero as n goes to infinity, and droping the dependence on $\sigma, \mathbf{t}$ in the notation, we can write with $\eta = |\lambda|^2/(2|\gamma| + |\lambda|^2)$,

$$f_\theta - f_0 = \frac{1}{2}(2|\gamma| + |\lambda|^2) \left((1-\eta) w(\gamma)^T L' + \eta w(\lambda)^T L" w(\lambda) + o(1) \right) \tag{8}$$

where $w(x) = x/|x|$ if $x \neq 0$ and

$$\begin{aligned} L' &= \left((\nabla g_{\gamma^0_1})^T, ..., (\nabla g_{\gamma^0_{k_0}})^T, g_{\gamma^0_1} - g_{\gamma^0_{k_0}}, ..., g_{\gamma^0_{k_0-1}} - g_{\gamma^0_{k_0}}, g_{\gamma^0_{t_{k_0}+1}} - g_{\gamma^0_{k_0}}, ..., g_{\gamma^0_k} - g_{\gamma^0_{k_0}} \right), \\ L" &= \text{diag} \left(p^0_1 D^2 g_{\gamma^0_1}, ..., p^0_{k_0} D^2 g_{\gamma^0_{k_0}} \right). \end{aligned}$$

We now prove that there exists $c > 0$ and $N \in \mathbb{N}$ such that for all $n \geq N$ and all $\theta \in A_n$, $|\lambda|^2 + 2|\eta| \leq \delta_n/c$. Indeed, were it not the case, we could construct a sequence c_n decreasing to 0 such that there would exist a subsequence θ_{r_n} satisfying

$$\left| (1 - \eta_{r_n}) w(\gamma_{r_n})^T L^{'}_{r_n} + \eta_{r_n} w(\lambda_{r_n})^T L" w(\lambda_{r_n}) \right| \leq c_n. \tag{9}$$

Thus to prove that $|\lambda|^2 + 2|\eta| \leq \delta_n/c$ for some c, it is enough to find a subsequence of θ_{r_n} which contradicts (9). Thus to simplify notations we write without loss of generality all subsequences θ_n. Since the set of possible partitions $\mathbf{t}$ and σ is finite, there is a subsequence of θ_n along which $\mathbf{t}$ and σ are constants. From now on we work with this $\mathbf{t}$ and σ which we drop from our notations hereafter. Since $w(\gamma_n), w(\lambda_n), \eta_n$ vary in a compact set there exists a subsequence which converges to some values $w(\gamma), w(\lambda), \eta$ on the unit spheres of dimensions k_0 and $k - k_0 - 1$ and on $[0, 1]$ respectively, and which we still denote $w(\gamma_n), w(\lambda_n), \eta_n$. Despite the notation $w(\gamma)$ and $w(\lambda)$ the above statement does not imply that γ_n or λ_n converges towards γ and λ.

We first consider the case where Γ is compact. Then θ_n belongs to a compact set and there exists a subsequence such that $L_n^{'}$ converges to some vector $L_\infty^{'}$ corresponding to some $\theta \in \Theta_0$. At the limit, inequality (9) becomes

$$(1 - \eta)w(\gamma)^t L_\infty^{'} + \eta w(\lambda)^t L" w(\lambda) = 0$$

and if $0 < \eta < 1$ we can construct (ϕ, ψ) based on $w(\gamma)$, $w(\lambda)$ and η such that there exists $u > 0$ for which

$$f'_{\phi^0,\psi}(\phi - \phi^0) + 0.5(\phi - \phi^0)^t f'_{\phi^0,\psi}(\phi - \phi^0) = u(1 - \eta)w(\gamma)^t L_\infty^{'} + u\eta w(\lambda)^t L" w(\lambda) = 0$$

which contradicts assumption [A4]. If $\eta = 1$ such a construction still exists and satisfies for all $i = 1, ..., k_0$, $\sum_{j \in I_i} q_j \gamma_j = \gamma_i^0$, for all $i = 1, ..., k_0 - 1$ $s_i = 0$, for all $i = t_{k_0} + 1, ..., k$, $p_i = 0$ and for all $i = 1, ..., k_0$, $j \in I_i$, $\sqrt{q_j}(\gamma_j - \gamma_i^0) = u w_{t_{i-1}+j}$ with $u > 0$ small. This is possible even if there exists $i \leq k_0$ such that $t_i = t_{i-1} + 1$, i.e. the class of components close to γ_i^0 is a singleton, since $\eta_n \to \eta = 1$ implies that $|\gamma_n| = o(|\lambda|_n|^2)$ and

$$|\gamma_{t_i,n} - \gamma_i^0| = o\left(\sum_i \sum_{j \in I_i} q_j (\gamma_{j,n} - \gamma_i^0)^2\right). \tag{10}$$

Therefore if $w_{t_i}(\tilde{\lambda}_n) \to 0$ and we can choose $\gamma_{t_i} = \gamma_i^0$. If $\eta = 0$, then (9) leads to $w(\gamma)^t L_\infty^{'} = 0$. Note that the constraints on $w(\gamma)$ are the following: for the components corresponding to p_j, $j \geq t_{k_0} + 1$, the terms w_l are greater or equal to 0. Assumption [A4] together with the positivity of the weights associated to the p_j's, $j = t_{k_0} + 1, ..., k$, imply that for all $i = 1, ..., k_0 - 1$, $w_i(\gamma)^t \nabla g_{\gamma_i^0} + w_{dk_0+i}(\gamma) g_{\gamma_i}^0 = 0$ and

$$\forall i \geq dk_0 + k_0, \quad w_i(\gamma) = 0, \quad \text{and} \quad g_{\gamma_{k_0}^0} \sum_{i=dk_0+1}^{2k_0-1} w_i(\gamma) - w_{k_0}(\gamma) \nabla g_{\gamma_{k_0}^0} = 0$$

Therefore for all $i = 1, ..., k_0 - 1$

$$w_{k_0+i}(\gamma) = -\frac{w_i(\gamma)^t \nabla g_{\gamma_i^0}}{g^0_{\gamma_i}} = -w_i(\gamma)^t \nabla \log g_{\gamma_i^0}$$

Since $E_{\gamma_i^0}\left(\nabla \log g_{\gamma_i^0}(X)\right) = 0$, the above equality implies that for all $i = dk_0 + 1, ..., (d+1)k_0$, $w_i(\gamma) = 0$. The regularity asumption [A2] (positivity of the Fisher information matrix) of each model g_γ implies that $w^t \nabla \log g_\gamma = 0 \quad \Leftrightarrow w = 0$. We finally obtain that $w(\gamma) = 0$ which contradicts the fact that $w(\gamma)$ belongs to the sphere with radius 1. If Γ is not compact, for any converging sub-sequence of θ_n to a point Θ_0 for which all components parameters γ_j belong to Γ or for which all components γ_j correspond to a probability density (see assumption [A4]) we can apply the arguments of the compact case, leading to a contradiction of (9). We thus only need consider sub-sequences which do not converge to such a point. In other words and without loss of generality we can assume that θ_n converges to a point in $\bar{\Theta}_0$, where at least one of the components' parameters belongs to $\partial\tilde{\Gamma}$, where $\partial\tilde{\Gamma} = \{\gamma \in \partial\Gamma; \int g_\gamma(x)d\mu(x) \in \{0, +\infty\}\}$. Let $J = \{j \leq k; \gamma_{j,n} \to \partial\tilde{\Gamma}\} \neq \emptyset$. By definition of $\mathbf{t}$, $J \subset \{t_{k_0} + 1, ..., k\}$ and choosing σ accordingly we can write $J = \{k_1, ..., k\}$ with $k_1 \geq t_{k_0} + 1$. Hence for all $j < k_1$, there exists $\gamma_j \in \Gamma$ such that $\gamma_{j,n} \to \gamma_j$. We split L'_n into $L^{'}_{n,(1)}$ and $L^{'}_{n,(2)}$ where $L^{'}_{n,(2)} = (g_{\gamma_{j,n}} - g^0_{\gamma_{k_0}}, j = k_1, ..., k)$ and by definition of k_1, $L'_{n,(1)}$ converges to $L^{'}_{\infty,(1)}$ so that (9) becomes in the limit,

$$\left|(1-\eta)w^t_{(1)}(\gamma)L^{'}_{(1)} + (1-\eta)w^t_{(2)}(\gamma)L^{'}_{n,(2)} + \eta w(\lambda)^t L^{"} w(\lambda)\right|_1 \to 0 \tag{11}$$

as n goes to infinity, where the only term depending on n is $L^{'}_{n,(2)}$. If $\eta < 1$ then (11) can be written as: there exists h integrable such that

$$\lim_{n\to\infty} \left| \sum_{j=1}^{k-k_1+1} w_{(2),j}(\gamma) g_{\gamma_{j+k_1-1,n}} - h \right|_1 = 0$$

if $w_{(2)}(\gamma) \neq 0$ then set $\bar{w}_2 = \sum_l w_{(2),l}$ and since $w_{(2),l} \geq 0$ for all l, then (11) can be expressed as

$$\left| \sum_{j=1}^{k-k_1+1} p_j g_{\gamma_{j+k_1-1,n}} - h/(1-\bar{w}_2) \right|_1 \to 0, \quad p_j = w_{(2),j}/\bar{w}_2$$

Thus $h/(1-\bar{w}_2)$ is a probability density and $\sum_{j=1}^{k-k_1+1} p_j g_{\gamma_{j+k_1-1,n}}$ converges towards a proper probability density which contradicts de definition of J. Hence $w_{(2)} = 0$ and we can apply the same arguments as in the compact case to conclude. If $\eta = 1$, then we can use the same argument as in in the compact case since $L^{'}_{n,(2)}$ has no influence.

Therefore on A_n

$$|\lambda|^2 + 2|\gamma| \leq \delta_n, \quad \sum_{j \geq k_0+1} p_j > M_n \delta_n$$

so that for all $\theta \in A_n$, $(\mathbf{t}, \sigma)$ must satisfy :

$$\exists i \leq k_0, \quad \text{card}(I_i) \geq 2, \quad \exists j_1, j_2 \in I_i, \quad q_{j_1} > \epsilon/k_0, \quad q_{j_2} > (M_n - k)\delta_n > M_n\delta_n/2$$

if M_n is large enough; without loss of generality we set $i = 1$ and $j_1 = 1, j_2 = 2$. Then we obtain

$$|s_i| \leq \delta_n, \quad \forall i \leq k_0 - 1 \quad p_j \leq \delta_n, \quad j = t_{k_0} + 1, ..., k, \quad \left| \sum_{j \in I_i} q_j\gamma_j - \gamma_i^0 \right| \leq \delta_n, \quad q_j|\gamma_j - \gamma_i^0|^2 \leq \delta_n$$

We now bound the prior probability of such a set : The constraints on the s_i's and on the p_j's imply that

$$\pi(\{|s_i| \leq \delta_n, \forall i \leq k_0\}) \leq C\delta_n^{k_0 - 1}, \quad \pi\left(\{p_j \leq \delta_n, j = t_{k_0} + 1, ..., k\}\right) \leq \delta_n^{\sum_{j=t_{k_0}+1}^{k} \alpha_j}.$$

Also on I_1

$$q_1(\gamma_1 - \gamma_1^0) = -\sum_{j \in I_1, j>1} q_j(\gamma_j - \gamma_1^0) + 0(\delta_n), \quad q_2 > M_n\delta_n/2, \quad |\gamma_j - \gamma_1^0| \leq \sqrt{\delta_n/q_j}, \quad j \in I_1$$

the prior probability of the set of $(q_1, \gamma_1, q_2, \gamma_2, q_j\gamma_j, j > 2, j \in I_1)$ satisfying the above constraints is bounded by

$$V_1 \leq \delta_n^d \int_{q_2 \geq M_n\delta_n/2} (\delta_n/q_2)^{d/2} q_2^{\alpha_2 - 1} dq_2 \prod_{j>2, j \in I_1} \int_{q_j, \gamma_j} \mathbb{1}_{|\gamma_j - \gamma_1^0| \leq \sqrt{\delta_n/q_j}} q_j^{\alpha_j - 1} dq_j d\gamma_j$$

Note that

$$\begin{aligned} \int_{q_j, \gamma_j} \mathbb{1}_{|\gamma_j - \gamma_1^0| \leq \sqrt{\delta_n/q_j}} q_j^{\alpha_j - 1} dq_j d\gamma_j &\leq \delta_n^{\alpha_j} + \delta_n^{d/2} \int_{\delta_n}^{1} q^{\alpha_j - 1 - d/2} dq \\ &\leq \delta_n^{\alpha_j \wedge d/2} (\log n)^{\mathbb{1}_{\alpha_j = d/2}} \end{aligned}$$

and we finally obtain that if $q = \sum_{j=1}^{k_0} \mathbb{1}_{\alpha_j = d/2}$ such that

$$V_1 \leq (\log n)^q \delta_n^d \delta_n^{\sum_{j=2}^{t_1} \alpha_j \wedge d/2} M_n^{\max_j \alpha_j - d/2}$$

Similarly the prior probability of the set of parameters associated with I_i is bounded by

$$V_i \leq \delta_n^{d + \sum_{j=t_{i-1}+2}^{t_i} \alpha_j \wedge d/2}$$

Finally the volume of the set of $\theta \in A_n$ associated with the partition $\mathbf{t}$ is bounded

$$V_{\mathbf{t}} \leq \delta_n^{k_0 - 1 + \sum_{j=t_{k_0}+1}^{k} \alpha_j + dk_0 + \sum_{j=3}^{t_{k_0}} \alpha_j \wedge d/2 - \sum_{i=1}^{k_0 - 1} \alpha_{t_i + 1} \wedge d/2} (\log n)^q M_n^{\max_j \alpha_j - d/2}$$

If $\max_j \alpha_j < d/2$ then

$$V_{\mathbf{t}} \leq \delta_n^{k_0-1+dk_0+\sum_{j=2}^{k}\alpha_j-\sum_{i=1}^{k_0-1}\alpha_{t_i+1}} M_n^{\max_j \alpha_j - d/2}$$

So that with probability going to 1, $V_{\mathbf{t}} D_n \leq M_n^{\max_j \alpha_j - d/2}$ i.e. $\pi(A_n)D_n \leq M_n^{\max_j \alpha_j - d/2}$ and

$$P^{\pi}\left[A_n'|X^n\right] = o_p(1) \quad \text{if} \quad \max\{\alpha_j, j = 1, ..., k\} < d/2.$$

We now prove the second part of Theorem 1, where $\min\{\alpha_j j = 1, ..., k\} > d/2$ and we prove that

$$P^{\pi}\left[B_n|X^n\right] = o_p(1) \quad \text{where} \quad B_n := \{|f_0 - f_\theta| \leq \delta_n\} \cap \{\sum_{i=k_0+1}^{k} p_i \leq \epsilon_n\}, \tag{12}$$

with ϵ_n small (either converging to 0 or a fixed small constant). To prove (12) we need a different lower bound of D_n, based on a different approximative set $\tilde{S}_n$ of f_0, since the approximative path based on $\sum_{j=k_0+1}^{k} p_j \approx 0$ is not the most parcimonious in terms of prior mass. Consider $\mathbf{t} = (0, k-k_0+1, k-k_0+2, ..., k)$ so that $t_{k_0} = k$ and define

$$\tilde{S}_n = \{(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}); |\bar{\gamma}_i - \gamma_i^0| \leq n^{-1/2}; |s_i| \leq n^{-1/2}; q_j|\gamma_j - \gamma_i^0|^2 \leq n^{-1/2}, \quad \forall j \in I_i, i = 1, ..., k_0\}$$

where $\bar{\gamma}_i = \sum_{j\in I_i} q_j\gamma_j$, $\phi_{\mathbf{t}} = (\gamma_j, j \leq k; s_i, i = 2, ..., k_0)$ and $\psi_{\mathbf{t}} = (q_j, j \in I_i, i \leq k_0)$. Similar computations to those made on the terms $V_{\mathbf{t}}$ lead to (up to fixed multiplicative constants)

$$\pi\left(\tilde{S}_n\right) \leq n^{-(k_0-1+dk_0+d/2(k-k_0))/2}, \quad \pi\left(\tilde{S}_n\right) \geq n^{-(k_0-1+dk_0+d/2(k-k_0))/2}.$$

To lower bound D_n we consider a Taylor expansion of $l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ around $\phi_{\mathbf{t}} = \phi_{\mathbf{t}}^0$ to the order 3

$$l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) - l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}) \quad = \quad \sqrt{n}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T W_n - \frac{n}{2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T H(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) + R_n \tag{13}$$

where $H = -\frac{\partial^2 l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})}{\partial\phi_{\mathbf{t}}\partial\phi_{\mathbf{t}}^T}$, and noting

$$W_n(t) = \mathbb{G}_n\left(\frac{p_i^0 q_j \nabla_l g_{\gamma_i^0}}{f_0}\right), t = l+(j-1)d, j \in I_i, \quad W_n(kd+t) = \mathbb{G}_n\left(\frac{f_{\gamma_{t+1}}^0 - f_{\gamma_1}^0}{f_0}\right), t = 1, ..., k_0-1$$

and

$$R_n = \frac{1}{6}\sum_{r_1,r_2,r_3}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)_{r_1}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)_{r_2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)_{r_3}\frac{\partial l_n^3}{\partial\phi_{\mathbf{t},r_1}\phi_{\mathbf{t},r_2}\phi_{\mathbf{t},r_3}}(\bar{\phi}_{\mathbf{t}}, \psi_{\mathbf{t}}) \quad \bar{\phi}_{\mathbf{t}} \in (\phi_{\mathbf{t}}, \phi_{\mathbf{t}}^0).$$

We have

$$\begin{aligned}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T W_n &= \sum_{i=1}^{k_0} p_i^0 (\bar{\gamma}_i - \gamma_i^0)^T \mathbb{G}_n \left(\frac{\nabla g_{\gamma_i^0}}{f_0} \right) + \sum_{i=2}^{k_0} s_i \mathbb{G}_n \left(\frac{g_{\gamma_i^0} - g_{\gamma_1^0}}{f_0} \right) \\ &= O_p \left[\sum_{i=1}^{k_0} \|\bar{\gamma}_i - \gamma_i^0\| + \sum_{i=2}^{k_0} |s_i| \right] \\ &= O_p(n^{-1/2}). \end{aligned} \tag{14}$$

The difficulty in proving that the second term in (13) is of order $O_p(1)$ comes from the fact that $|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0|$ is not of order $n^{-1/2}$ since for each $j \in I_1$ $\|\gamma_j - \gamma_1^0\| = O(n^{-1/4})$. However simple computations lead to

$$\frac{n}{2}(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T H (\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) = O_p \left(n \left(\sum_i s_i^2 + \|\bar{\gamma}_i - \gamma_i^0\|^2 \right) \right) = O_p(1)$$

We now study R_n. Each term including at least one s_i or one $(\gamma_{k-k_0+i-1} - \gamma_i^0)$, $i \geq 2$ are of order $O_p(n^{-1})$, therefore we need only consider derivatives of the loglikelihood in the form:

$$\frac{\partial^3 l_n}{\partial \gamma_{j_1 l_1} \partial \gamma_{j_2 l_2} \partial \gamma_{j_3, l_3}}, \quad j_1, j_2, j_3 \in I_1$$

Straightforward computations computations imply that for all $l_1, l_2, l_3 \leq d$,

$$\begin{aligned} &\sum_{j, j_2, j_3 \in I_1} (\gamma_{j_1 l_1} - \gamma^0_{i_2 l_1})(\gamma_{j_2 l_2} - \gamma^0_{i_2 l_2})(\gamma_{j_3 l_3} - \gamma^0_{i_3 l_3}) \frac{\partial^3 l_n}{\partial \gamma_{j_1 l_1} \partial \gamma_{j_2 l_2} \partial \gamma_{j_3, l_3}} \\ = \; & O_p \left(n \left(\|\bar{\gamma}_1 - \gamma_1^0\| \sum_{j \in I_1} \|\gamma_j - \gamma_1^0\|^2 + n^{-1/2} \sum_{j \in I_1} \|\gamma_j - \gamma_1^0\|^3 + \sum_{j \in I_1} \|\gamma_j - \gamma_1^0\|^4 \right) \right) \end{aligned}$$

under the assumption that for all $i = 1, ..., k_0$,

$$F_0 \left[\sup_{|\gamma - \gamma_i^0| \leq \delta} \frac{|D^4 g_\gamma|}{g_\gamma} \right] < +\infty$$

Finally uniformly over $\tilde{S}_n$, $l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) - l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}) = O_p(1)$ and using similar computation as in the case $d/2 > \max_j \alpha_j$, for all $\epsilon > 0$ there exists $C_\epsilon > 0$ such that

$$P_0^n \left[D_n < n^{-(dk_0 + k_0 - 1 + d(k-k_0)/2)/2} C_\epsilon \right] \leq \epsilon.$$

We then bound $\pi[B_n]$. The argument used in the control of $\pi(A_n)$ imply that $\pi[B_n]$ is bounded

by the prior on the set constraint by : for all $\mathbf{t}$

$$|s_i| \leq \delta_n; \quad \| \sum_{j \in I_i} q_j \gamma_j - \gamma_i^0 \| \leq \delta_n \quad q_{t_i+j} \leq \epsilon_n, j = 2, ..., t_{i+1} - 1 \quad \forall i = 1, ..., k_0$$
$$q_j \|\gamma_j - \gamma_i^0\|^2 \leq \delta_n \quad \forall j \in I_i, i = 1, ..., k_0 \quad \text{and} \quad \sum_{j \geq t_{k_0}+1} p_j \leq \delta_n.$$

The prior probability of such a set is bounded by a term of order

$$\delta_n^{dk_0+k_0-1+d(t_{k_0}-k_0)/2+\sum_{j \geq t_{k_0}+1} \alpha_{\sigma(j)}} \epsilon_n^{\sum_i \sum_{j=2}^{t_{i+1}} (\alpha_{\sigma(j)} - d/2)}$$

so that $P^\pi [B_n | X^n] \leq \epsilon$ if $\epsilon_n \leq \epsilon(\delta_n/\sqrt{n})^a$ for an appropriate value of a. Hence if $\delta_n = n^{-1/2}$ it is enough to choose $\epsilon_n = \epsilon$ small , else ϵ_n must be a power of $(\log n)^{-1}$.