

# 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); Titterton 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 Titterton (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

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

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  $X_1, \dots, X_n$ , iid from a mixture model with  $k_0$  components:

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

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

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

for all  $\gamma \in \Gamma$  and all values of the parameter in the form  $\theta = (p_1^0, \dots, p_j^0, p_{k_0}^0, p_{k+1}, \gamma_1^0, \dots, \gamma_{k_0}^0, \gamma_j^0)$  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  $\Theta_k$  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  $f_{\theta}$  is the main object of interest this is not of great importance, however in many situations recovering  $\theta$  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 reasonable 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 give 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  $\gamma$  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 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 as 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  $\theta$ . 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  $F_0(g) = \int f_0(x)g(x)dx$ . Let  $\nabla g_\gamma$  be the vector of first derivatives of  $g_\gamma$  with respect to  $\gamma$ , and  $D^2 g_\gamma$  be the matrix of second derivatives with respect to  $\gamma$ . 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 notation that is useful to characterise  $\Theta_k^0$ , following Liu and Shao (2003)'s presentation. Let  $\mathbf{t} = (t_i)_{i=0}^{k_0}$  with  $0 = t_0 < t_1 < \dots < t_{k_0} \leq k$  be a partition of  $\{1, \dots, k\}$ . 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,

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

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

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

and

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

Note that  $f_0$  corresponds to

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

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

Then we parameterize  $\theta$  as  $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ , so that  $f_{\theta} = f_{(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})}$ ,  $f'_{(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})}$  where  $f''_{(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})}$  denotes 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  $P^{\pi}[\cdot | X^n]$  the posterior distribution, where  $X^n = (X_1, \dots, X_n)$ .

### Assumptions

A1 *L1 consistency*: Letting  $\delta_n = \sqrt{\log n} / \sqrt{n}$ , there exists  $M > 0$  such that in probability with respect to  $f_0$ ,

$$P^{\pi} [\|f_0 - f_{\theta}\| \geq M\delta_n | X^n] \rightarrow 0$$

A2 *Regularity*: The model  $\gamma \in \Gamma \rightarrow g_{\gamma}$  is three times differentiable and regular in the sense that for all  $\gamma \in \Gamma$  the Fisher information matrix associated with the model  $g_{\gamma}$  is positive definite at  $\gamma$ . 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  $i = 1, \dots, k_0$   $\gamma_i^0 \in \text{int}(\Gamma)$  the interior of  $\Gamma$ .

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  $\{1, \dots, k\}$  as defined above, let  $\theta \in \Theta_k$  and write  $\theta$  as  $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$ ; then

$$(\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T f'_{\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}} + \frac{1}{2} (\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0)^T f''_{\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}}} (\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0) = 0 \Leftrightarrow (\phi_{\mathbf{t}} = \phi_{\mathbf{t}}^0). \quad (3)$$

Assuming also that if  $\gamma \notin \{\gamma_1, \dots, \gamma_p\}$  then for all functions  $h_\gamma$  which are linear combinations of derivatives of  $g_\gamma$  of order less than or equal to 2 with respect to  $\gamma$ , and all functions  $h_1$  which are also linear combinations of derivatives of the  $g_{\gamma_j}$ 's  $j = 1, \dots, p$  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  $\Gamma$  is not compact then we also assume that for all sequences  $\gamma_n$  converging to a point in  $\partial\Gamma$  the frontier of  $\Gamma$ , considered as a subset of  $(\mathbb{R} \cup \{-\infty, +\infty\})^d$ ,  $g_{\gamma_n}$  converges pointwise either to a degenerate function, i.e. satisfying either  $\int g(x) d\mu(x) = +\infty$  or  $g(x) = +\infty \times \mathbb{1}_{x \in A}$  for some set  $A$  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  $D^2 g_{\gamma_i}^0$ ,  $i = 1, \dots, k_0$ .

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

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

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. Assumptions [A2] is a usual regularity assumption and assumption [A4], is a much weaker version than the assumptions in Liu and Shao (2004) or in Dacunha-Castelle and Gassiat (1999), since likelihood ratio need be integrable on some chosen

subset of  $\Gamma$  and not everywhere. Assumption [A4] (first part) is the same assumption as in Liu and Shao (2004). It is related to the linear independence of the functions  $g_\gamma$ ,  $\nabla g_\gamma$  and  $D_{r,s}^2 g_\gamma$ ,  $r \leq s$  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  $\Gamma$  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  $\Gamma$  is not compact under mild conditions, under 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  $[1, +\infty)$ , where the densities  $g_\gamma$  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  $g_{\gamma_i}$ 's and their derivatives. For instance, in the case of a mixture of Poisson distributions with parameters  $\lambda$ , when  $\lambda$  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  $\lambda^x(a_1 + a_2x + a_3x^2)$  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  $\mathcal{D}(\alpha_1, \dots, \alpha_k)$

## 2.2. Main result

**THEOREM 1.** *Under the assumptions [A1]-[A5] the posterior distribution satisfies:*

- *If  $\max(\alpha_j, j \leq k) < d/2$ , then for all  $\epsilon > 0$ ,*

$$P^\pi \left[ \exists I = \{j_1, \dots, j_{k-k_0}\}, \sum_{i \in I} p_i < n^{-1/2+\epsilon} \middle| X^n \right] \rightarrow 1 \quad P_0.$$

- *If  $\min(\alpha_j, j \leq k) > d/2$ , then for all  $\epsilon < 1/2$  for all permutation  $\sigma$  of  $\{1, \dots, k\}$*

$$P^\pi \left[ \exists I = \{j_1, \dots, j_{k-k_0}\}, \sum_{i \in I} p_i < \epsilon \wedge n^{-1/2+\epsilon} \middle| X^n \right] \rightarrow 0 \quad P_0.$$

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

$$\sum_{j=k_0+1}^k E^\pi [p_j | X^n] \rightarrow 0$$

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. 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 components 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-neglectable 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  $\alpha_j$ 's and on the difference between  $k$  and  $k_0$ . In particular, in the case where all  $\alpha_j$ '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  $\alpha_j$ '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  $\mathcal{D}(1/2, \dots, 1/2)$  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  $\gamma_j$  are independently and identically distributed according to some common distribution  $h_\eta$  indexed by a parameter  $\eta$  where  $\eta$  is itself given a prior  $\pi_0$  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.

The proof of Theorem 1 is given in the appendix. However we present some aspects of it that are of interest. Set  $A_n = \{\exists I = \{j_1, \dots, j_{k-k_0}\}, \sum_{i \in I} p_i > n^{-1/2+\epsilon}\}$ , i.e. 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  $\theta_0 \in \Theta_0$ . 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  $\sigma$  of  $\{1, \dots, k\}$ , by considering approximations of  $\Theta_0$  along paths of the form:

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

In contrast and by definition,  $A_n$  corresponds to paths approximating  $\Theta_0$  where at least two components merge, i.e. associated with partitions  $\mathbf{t}$  of  $\{1, \dots, k_0\}$  such that there exists  $i \leq k_0 - 1$  with  $t_{i+1} \geq t_i + 2$ . We prove in the appendix that each of these paths has a prior mass bounded by  $O((\log n)^q n^{-0.5(k_0 d + k_0 - 1 + d(k - k_0)/2)}) = o(n^{-(dk_0 + k_0 - 1 + \sum_{j \geq k_0 + 1} \alpha_{\sigma(j)})/2})$  when  $d/2 > \max\{\alpha_j, j = 1, \dots, k\}$ . Hence, when  $d/2 > \max\{\alpha_j, j = 1, \dots, k\}$   $dk_0 + k_0 - 1 + \sum_{j \geq k_0 + 1} \alpha_j$  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  $\sum_{j \geq k_0 + 1} p_j \approx 0$ , due to the influence of the prior via  $\alpha_j, j \geq k_0 + 1$ .

In contrast again, when  $d/2 \leq \min\{\alpha_j, j = 1, \dots, k\}$  a reverse phenomenon takes place, where we bound from below  $D_n$  by considering approximations of  $\Theta_0$  along paths of the following form: if  $I_1 = \{1, \dots, k - k_0 + 1\}$ ,  $I_i = \{k - k_0 + i\}$ ,  $i = 2, \dots, k_0$

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

i.e. by forcing all the parameters of the extra components to be close to  $\gamma_1^0$ . 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

$$\pi(A_n^c) \leq \delta_n^{dk_0 + k_0 - 1 + \sum_{j \geq k_0 + 1} \alpha_{\sigma(j)}} n^{\epsilon \sum_{j=k_0+1}^k (\alpha_{\sigma(j)} - d/2)}$$

which is of much smaller than  $D_n$  as soon as  $\epsilon > 1/2$  so that

$$P^\pi [A_n^c | X^n] = o_p(1).$$

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 test problems but also in point estimation problems.

In the following section we conduct a simulation study first to illustrate the above results but also to study the possible behaviours one could expect when  $\max \alpha_j \geq d/2$ .



### 3. Examples

We illustrate the results derived in the previous section with a very simple study of fitting a two-component Gaussian mixture model to a sample of data,  $Y = \{y_i, i = 1, \dots, n\}$ , generated from a single-component Gaussian distribution, say  $\mathcal{N}(0, 1)$ . Note that assumptions [A1]-[A5] are satisfied in the case of location mixtures of Gaussians or location 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 mixtures of Gaussians. We consider noninformative priors on  $p$  and  $\mu$  so that [A5] is also satisfied.

We consider three cases corresponding to dimensions  $d = 1, 2, > 2$ , respectively:

- Case 1 :  $\alpha_1 = \alpha_2 = 1 > d/2$  and  $\alpha_j - d/2 = 1/2$ .

$$y_i \sim N(0, 1) ; \quad G = p\mathcal{N}(\mu_1, 1) + (1 - p)\mathcal{N}(\mu_2, 1),$$

where  $\mathcal{N}(\mu, \tau)$  denotes the univariate normale distribution with mean  $\mu$  and variance  $\tau$ . In this case Theorem 1 implies that for any  $\epsilon > 0$ ,  $P^\pi [p_2 < n^{-\epsilon} | X^n] = o_p(1)$ , which might still contain the possibility that the extra component eventually becomes empty as the sample size becomes very large, but at a very slow rate.

- Case 2 :  $\alpha_1 = \alpha_2 = d/2$ .

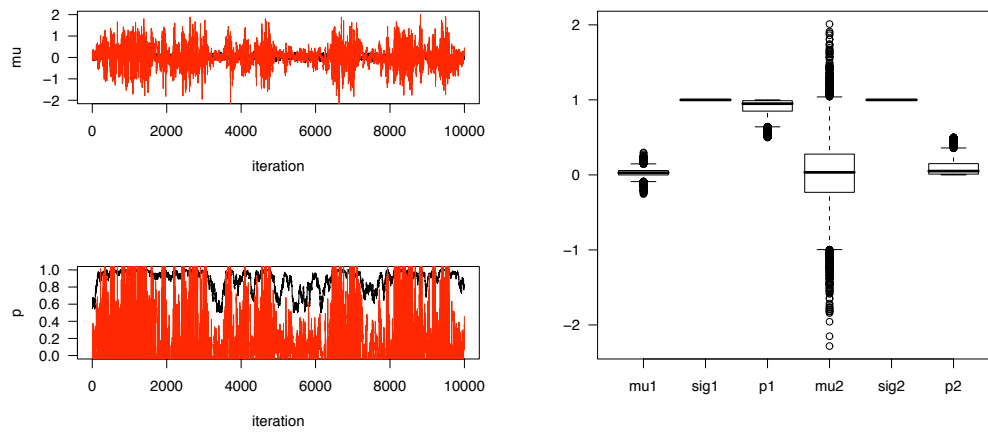
$$y_i \sim N(0, 1) ; \quad G = p\mathcal{N}(\mu_1, \sigma_{12}) + (1 - p)\mathcal{N}(\mu_2, \sigma_{22})$$

The theorem does not cover this case. It is our belief, however, that the posterior might still concentrate on the configuration where the extra component becomes eventually empty as the sample size increases.

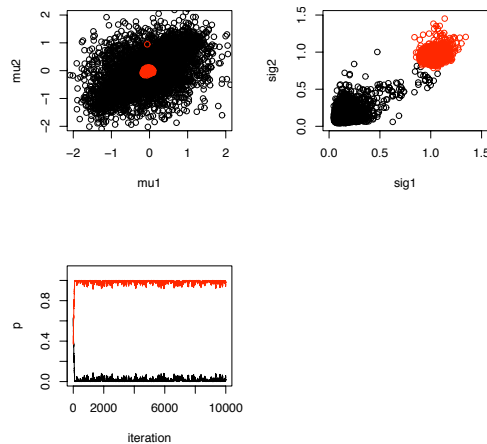
- Case 3 :  $\alpha_1 = \alpha_2 < d/2$ .

$$y_i \sim \mathcal{N}_2(0, 1) ; \quad G = p\mathcal{N}_2(\underline{\mu}_1, \Sigma_1) + (1 - p)\mathcal{N}_2(\underline{\mu}_2, \Sigma_2)$$

where, in the last case,  $\mathcal{N}_2(\underline{\mu}, \Sigma)$  represents a bivariate normal distribution with mean  $\underline{\mu} = (\mu_1, \mu_2)'$  and covariance matrix  $\Sigma$ . In the present case we only consider covariance matrices  $\Sigma_j$  to be diagonal with off-diagonals (covariances) equal to zero and diagonals (variances) given by  $\sigma_{j1}^2, \sigma_{j2}^2, j = 1, 2$ . In the above cases the generating distributions are  $\mathcal{N}(0, 1)$  and  $G$  denotes the model to be estimated. For each case, replicate samples of size 100, 500, 1000 and 5000 were generated and estimates of the mixture parameters were obtained based on 10,000 iterations of a standard Gibbs algorithm using conjugate, noninformative priors. Figures 1 and 2 depict the behaviour of the estimated parameters for one representative replicate of sample size 1000, for  $d = 1$  and  $d = 3$ , respectively. Other replicates revealed similar behaviour although, as expected, the results based



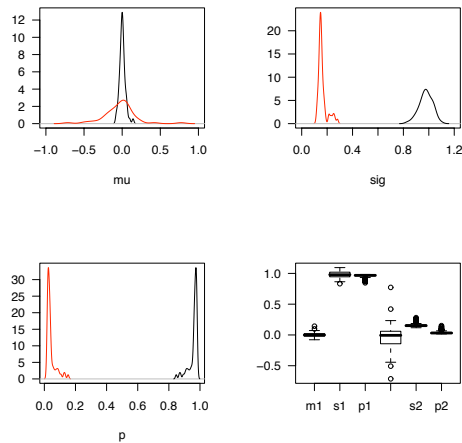
**Figure 1.** Left : trace of MCMC outputs, Right : boxplot of posterior means (60 replicates) in the case  $d=1$  and  $n=1000$



**Figure 2.** (representative) trace of MCMC outputs in the case  $d=3$  and  $n=1000$

on smaller samples ( $n = 100$ ) were less compelling than those based on larger samples. The empirical findings support the theoretical asymptotic behaviour described in the previous section: for  $d = 1$  the posterior distribution of the weights is unstable, even with increasing sample size, but the parameters of the components become closer and closer to mean equal to 0 and variances equal to 1, as the sample size increases. On the contrary, when  $d = 3$  one component is effectively empty.

Figures 3 and 4 summarise the estimated parameters for the case  $d = 2$ . For this case, no theory was established; it was proposed that the weights could behave quite differently across replicates, although they could stabilise to an empty component in a longer MCMC run or with larger sample size. This is indeed reflected in the Figures.



**Figure 3.** density and boxplots of posterior means (based on 60 replicates) in the case  $d=2$  and  $n = 1000$

#### 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  $\gamma$  ( $d \leq 2$ ) it becomes necessary to *favour* small weights with a prior in the form  $p_1^{-1/2} \dots p_k^{-1/2}$ , which interestingly corresponds to the noninformative prior in a multinomial model. It thus appears that in this subtil 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. *aos* 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. Titterton (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 Eighth 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.
- Titterton, D., A. Smith, and U. Makov (1985). *Statistical Analysis of Finite Mixture Distributions*. New York: John Wiley.

## 5. Appendix: Proof of Theorem 1

Set  $A_n = \{\exists I = \{j_1, \dots, j_{k-k_0}\}, \sum_{i \in I} p_i > n^{-1/2+\epsilon}\}$  for some  $\epsilon > 0$ . The posterior probability of interest is bounded by

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

where  $A'_n = A_n \cap \{\|f - f_0\| \leq M\delta_n\}$ . We denote by

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

To prove the first part of Theorem 1 we first prove that for all  $\epsilon > 0$ , there exists  $C_\epsilon$  such that for all permutations  $\sigma : \{1, \dots, k\} \rightarrow \{1, \dots, k\}$ ,

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

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

$$P^\pi [A_n | X^n] = o_p(1)$$

which terminates the proof of the first part of Theorem 1. Similarly if  $d/2 < \min\{\alpha_j, j = 1, \dots, k\}$ , we obtain

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

which leads to

$$P^\pi [A_n^c | X^n] = o_p(1) .$$

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

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

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  $\Gamma_0$ ,  $\min_{l \leq k_0} |\gamma_j^* - \gamma_l^0| > C\epsilon_1$  and by definition of  $S_n$ ,  $\sum_{j \geq k_0 + 1} p_j \leq k_0 \delta_n$ . Such a path to approach  $\Theta_0$  corresponds to the partition  $\mathbf{t} = (0, 1, 2, \dots, k_0)$ . Let  $(\phi_{\mathbf{t}}, \psi_{\mathbf{t}})$  be the parameterisation of  $\theta$  associated to the partition  $\mathbf{t}$ . We consider a Taylor expansion of  $l_n(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}) - l_n(\phi_{\mathbf{t}}^0, \psi_{\mathbf{t}})$ . By convention and without loss of generality we write  $p_1 = 1 - \sum_{j \geq 2} p_j$  and write  $p_j^0 = 0$  and  $\gamma_j^0 = \gamma_j$  for  $j = k_0 + 1, \dots, k$ . Then

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

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

$$U_n(i-1) = \mathbb{G}_n \left( \frac{f_{\gamma_i^0} - f_{\gamma_1^0}}{f_0} \right), i = 2, \dots, k, \quad U_n(i) = \mathbb{G}_n \left( \frac{\nabla_l g_{\gamma_j^0}}{f_0} \right), \quad i = k-1+l+(j-1)*d, \quad j \leq k_0$$

then  $U_n = o_p(1)$ . Denote by  $\Omega_n(c_0, C) = \{(\theta, X^n); J(\bar{\phi}) \leq c_0 n I(\theta); |U_n| \leq C\}$  where  $I(\theta)$  is a Fisher information matrix defined as  $E_0^n [J(\phi^0)]$ . Assumptions [A2] and [A3], together with basic algebra imply that  $I$  is bounded from above by a constant times the identity matrix uniformly on  $S_n$  so that 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

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

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  $c_1 p_{k_0+1}^{\alpha_{k_0+1}-1} \dots p_k^{\alpha_k-1}$ . 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  $\sigma^*$  defined in Theorem 1. Set  $A(k) = (dk_0 + k_0 - 1 + \sum_{j>k_0} \alpha_j)/2$ , and  $a > 0$  as small as need be; then

$$\begin{aligned} P_0^n [D_n < cn^{-A(k)}] &\leq P_0^n \left[ \int_{S_n \cap \Omega_n} e^{-\frac{nC_0}{2} \|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1} U_n / \sqrt{n}\|^2} d\pi_k(\theta) < 2an^{-A(k)} \right] \\ &\leq P_0^n \left[ \int_{S_n \cap \Omega_n^c} e^{-\frac{nC_0}{2} \|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1} U_n / \sqrt{n}\|^2} d\pi_k(\theta) > \int_{S_n} e^{-\frac{nC_0}{2} \|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1} U_n / \sqrt{n}\|^2} d\pi_k(\theta)/2 \right] \\ &\leq \frac{E_0^n \left[ \int_{S_n \cap \Omega_n^c} e^{-\frac{nC_0}{2} \|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0 - C_0^{-1} U_n / \sqrt{n}\|^2} d\pi_k(\theta) \right]}{Cn^{-A(k)}} + o(1) \\ &\leq C'n^{A(k)} \int_{S_n} P_0^n [\Omega_n^c] d\pi_k(\theta), \end{aligned}$$

The lower bound in (5) is then proved by determining an upper bound on  $P_0^n [\Omega_n^c(c_0, C)]$ . 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  $P_0^n [J(i, i') - nI(i, i') < cn]$ . We have if  $i, i' \leq k - 1$ ,

$$J(i, i') - nI(i, i') = \sqrt{n} \mathbb{G}_n \left( \frac{(g_{\gamma_i^0} - g_{\gamma_1^0})(g_{\gamma_{i'}^0} - g_{\gamma_1^0})}{f_0^2} \right) + n\mathbb{P}_n [\Delta_{\bar{\theta}}(i, i')]$$

with

$$\Delta_{\bar{\theta}}(i, i') = \left( \frac{(g_{\bar{\gamma}_i} - g_{\bar{\gamma}_1})(g_{\bar{\gamma}_{i'}} - g_{\bar{\gamma}_1})}{f_{\bar{\theta}}^2} \right) - \left( \frac{(g_{\gamma_i^0} - g_{\gamma_1^0})(g_{\gamma_{i'}^0} - g_{\gamma_1^0})}{f_0^2} \right)$$

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

$$Cn^{-1} F_0 \left[ \left( \frac{(g_{\gamma_i^0} - g_{\gamma_1^0})(g_{\gamma_{i'}^0} - g_{\gamma_1^0})}{f_0^2} \right)^2 \right] \leq \frac{Ck_0}{n \min_{j \leq k_0} (p_j^0)^4} + n^{-1} \max_{\gamma \in \Gamma_0} F_0 \left( \frac{g_{\gamma}^4}{f_0^4} \right)$$

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  $i, i' \leq k - 1$ , Then if  $i, i' \leq k - 1$

$$\begin{aligned} \left| \frac{\partial \Delta(i, i')}{\partial p_j} \right| &= \left| \frac{(g_{\gamma_i} - g_{\gamma_1})(g_{\gamma_{i'}} - g_{\gamma_1})(g_{\gamma_j} - g_{\gamma_1})}{f_\theta^3} \right| \\ &\leq \frac{[\bar{g}_{\gamma_i^0} + \bar{g}_{\gamma_{i'}^0}][\bar{g}_{\gamma_{i'}^0} + \bar{g}_{\gamma_1^0}][\bar{g}_{\gamma_j^0} + \bar{g}_{\gamma_1^0}]}{(1 - \delta_n)^3 \left( \sum_{j=1}^k p_j^0 \bar{g}_{\gamma_j^0} \right)^3} \end{aligned}$$

and if  $i, i' \leq k - 1, j \leq k_0$ ,

$$\begin{aligned} \left| \frac{\partial \Delta_\gamma(i, i')}{\partial \gamma_j} \right| &= \left| \frac{(g_{\gamma_i} - g_{\gamma_1})(g_{\gamma_{i'}} - g_{\gamma_1}) \nabla g_{\gamma_j}}{f_\theta^3} + \mathbb{1}_{j=i} \frac{\nabla g_{\gamma_i}(g_{\gamma_{i'}} - g_{\gamma_1})}{f_\theta^2} + \mathbb{1}_{j=i'} \frac{\nabla g_{\gamma_{i'}}(g_{\gamma_i} - g_{\gamma_1})}{f_\theta^2} \right. \\ &\quad \left. - \mathbb{1}_{j=1} \frac{\nabla g_{\gamma_1}(g_{\gamma_{i'}} + g_{\gamma_i} - 2g_{\gamma_1})}{f_\theta^2} \right| \\ &\leq \frac{[\bar{h}_{\gamma_i^0} + g_{\gamma_k}][\bar{h}_{\gamma_{i'}^0} + g_{\gamma_k}][\bar{h}_{\gamma_j^0} + g_{\gamma_k}]}{(1 - \delta_n)^3 \left( \sum_{j=1}^k p_j^0 \bar{h}_{\gamma_j^0} \right)^3} + \mathbb{1}_{j=i} \frac{\sup_{|\gamma - \gamma_i^0| \leq \delta} |\nabla g_\gamma| (\bar{h}_{\gamma_{i'}^0} + g_{\gamma_k})}{(1 - \delta_n)^3 \left( \sum_{j=1}^k p_j^0 \bar{h}_{\gamma_j^0} \right)^2} \\ &\quad + \mathbb{1}_{j=i'} \frac{\sup_{|\gamma - \gamma_{i'}^0| \leq \delta} |\nabla g_\gamma| (\bar{h}_{\gamma_i^0} + g_{\gamma_k})}{(1 - \delta_n)^3 \left( \sum_{j=1}^k p_j^0 \bar{h}_{\gamma_j^0} \right)^2}. \end{aligned}$$

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

$$F_0 \left( \sup_{\theta \in S_n} \left| \frac{\partial \Delta_\gamma(i, i')}{\partial \gamma_j} \right| \right) \leq M < +\infty \quad \forall j, \forall i, i' \leq k - 1,$$

so that for all  $c > 0$ , there exist  $\delta_0$  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  $i \leq k - 1$  and  $i' \geq k$ ,

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

with

$$\Delta_{\bar{\theta}}(i, i') = \frac{(g_{\bar{\gamma}_i} - g_{\gamma_k}) \nabla g_{\bar{\gamma}_{i'}}}{f_{\bar{\theta}}^2} - \frac{(g_{\gamma_i^0} - g_{\gamma_k}) \nabla g_{\gamma_{i'}^0}}{f_0^2}$$

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  $\Delta_\gamma(i, i')$  we obtain an upper bound with terms of the form

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



and

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

so that

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

The same calculations can be made for the terms  $J(i, i')$  when  $i, i' \geq k$ , so that finally there exists  $c_0, C > 0$  such that for all  $\theta \in S_n$   $P_0^n [\Omega_n^c(c_0, C)] \leq 2\epsilon$  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  $\theta \in A_n$ , consider a partition  $\mathbf{t}^*$  and a permutation  $\sigma^*$  which minimizes  $2|\gamma| + |\lambda|^2$ , where

$$\gamma = \left( [(s_i + p_i^0) \left( \sum_{j \in I_i} q_j \gamma_j - \gamma_i^0 \right)]_{i=1}^{k_0}, (s_i)_{i=2}^{k_0}, (p_j)_{j=t_{k_0}+1}^k \right) \quad \lambda = ([\sqrt{q_j}(\gamma_j - \gamma_i^0)]_{j \in I_i}, i = 1, \dots, k_0).$$

Since  $\theta \in A_n$ , for such a  $(\mathbf{t}^*, \sigma^*)$ ,  $(2|\gamma| + |\lambda|^2)(\mathbf{t}^*, \sigma^*)$  goes to zero as  $n$  goes to infinity, and we can consider a Taylor expansion of  $f_\theta$  in terms of  $\phi_{\mathbf{t}}$  around  $\phi_{\mathbf{t}}^0$ :

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

where  $\bar{\phi}_{\mathbf{t}} \in (\phi_{\mathbf{t}}, \phi_{\mathbf{t}}^0)$ . The last term of right hand side of the above equation is bounded by  $C|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0|^3$  in  $L_1$  for some positive constant  $C > 0$ , it is thus  $o(|\phi_{\mathbf{t}} - \phi_{\mathbf{t}}^0|^2)$ . Therefore, dropping the dependence on  $\sigma^*, \mathbf{t}^*$  set  $\eta = |\lambda|^2 / (2|\gamma| + |\lambda|^2)$ , and

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

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

$$\begin{aligned} L' &= \left( (\nabla g_{\gamma_1^0})^T, \dots, (\nabla g_{\gamma_{k_0}^0})^T, g_{\gamma_1^0} - g_{\gamma_{k_0}^0}, \dots, g_{\gamma_{k_0-1}^0} - g_{\gamma_{k_0}^0}, g_{\gamma_{t_{k_0}+1}^0} - g_{\gamma_{k_0}^0}, \dots, g_{\gamma_k^0} - g_{\gamma_{k_0}^0} \right), \\ L'' &= \text{diag} \left( p_1^0 D^2 g_{\gamma_1^0}, \dots, p_{k_0}^0 D^2 g_{\gamma_{k_0}^0} \right). \end{aligned}$$

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  $\theta_{r_n}$  satisfying

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

Thus to prove that  $|\lambda|^2 + 2|\eta| \leq \delta_n/c$  for some  $c$ , it is enough to find a subsequence of  $\theta_n$  which contradicts (9). Thus to simplify notations we write without loss of generality all subsequences  $\theta_n$ . Since the set of possible partitions  $\mathbf{t}^*$  and  $\sigma^*$  is finite, there is a subsequence of  $\theta_n$  along which  $\mathbf{t}^*$  and  $\sigma^*$  are constants. From now on we work with this  $\mathbf{t}^*$  and  $\sigma^*$  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  $w(\gamma), w(\lambda), \eta$  on the unit spheres of dimensions  $k_0$  and  $k - k_0 - 1$  and on  $[0, 1]$  respectively, and which we still denote  $w(\gamma_n), w(\lambda_n), \eta_n$ .

We first consider the case where  $\Gamma$  is compact. Then  $\theta_n$  belongs to a compact set and there exists a subsequence such that  $L'_n$  is consistent 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  $(\phi, \psi)$  based on  $w(\gamma), w(\lambda)$  and  $\eta$  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  $i = 1, \dots, k_0$ ,  $\sum_{j \in I_i} q_j \gamma_j = \gamma_i^0$ ,  $s_i = 0$ ,  $p_j = 0$ ,  $i = 1, \dots, k_0 - 1$  and  $j = t_{k_0} + 1, \dots, k$   $\sqrt{q_j}(\gamma_j - \gamma_i^0) = uw_{t_{i-1}+j}$  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  $\gamma_i^0$  is a singleton, since then, having  $\tilde{\eta}_n \rightarrow 1$  means that  $|\tilde{\gamma}_n| = o(|\tilde{\lambda}_n|^2)$  and

$$|\tilde{\gamma} - \gamma_i^0| = o\left(\sum_i \sum_{j \in I_i} q_j (\tilde{\gamma}_j - \gamma_i^0)^2\right). \quad (10)$$

Therefore if  $w_{t_i}(\tilde{\lambda}_n) \rightarrow w_{t_i} \neq 0$  then there exists  $c_1 > 0$  such that

$$(\gamma_{t_i} - \gamma_i^0)^2 \geq c_1 \left( \sum_i \sum_{j \in I_i} q_j (\gamma_j - \gamma_i^0)^2 \right)$$

which contradicts (10). 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,  $j = t_{k_0} + 1, \dots, k$ , imply that for all  $i = 1, \dots, k_0 - 1$ ,  $w_i(\gamma)^t \nabla g_{\gamma_i^0} + w_{k_0+i} g_{\gamma_i^0}^0 = 0$  and

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

Therefore for all  $i = 1, \dots, k_0 - 1$

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

Since  $E_{\gamma_i^0} \nabla \log g_{\gamma_i^0}(X) = 0$ , the above equality implies that for all  $i = k_0 + 1, \dots, 2k_0 - 1$ ,  $w_{k_0+i}(\gamma) = 0$ . The regularity assumption [A2] (positivity of the Fisher information matrix) of each model  $g_\gamma$  implies that  $w^t \nabla \log g_\gamma = 0 \Leftrightarrow w = 0$ . We finally obtain that  $w(\gamma) = 0$  which contradicts the fact that  $w(\gamma)$  belongs to the sphere with radius 1. If  $\Gamma$  is not compact, for any converging sub-sequence of  $\theta_n$  to a point  $\Theta_0$  for which all components parameters  $\gamma_j$  belong to  $\Gamma$  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  $\theta_n$  converges to a point in  $\bar{\Theta}_0$ , where at least one of the components' parameters belongs to  $\partial\Gamma$ . Let  $J = \{j \leq k; \gamma_{j,n} \rightarrow \partial\Gamma\} \neq \emptyset$ . By definition of  $\mathbf{t}^*$ ,  $J \subset \{t_{k_0} + 1, \dots, k\}$  and choosing  $\sigma^*$  accordingly we can write  $J = \{k_1, \dots, k\}$  with  $k_1 \geq t_{k_0} + 1$ . Hence for all  $j < k_1$ , there exists  $\gamma_j \in \Gamma$  such that  $\gamma_{j,n} \rightarrow \gamma_j$ . We split  $L'_n$  into  $L'_{n,(1)}$  and  $L'_{n,(2)}$  where  $L'_{n,(2)} = (g_{\gamma_{j,n}} - g_{\gamma_{k_0}}^0, j = k_1, \dots, k)$  and by definition of  $k_1$ ,  $L_{n,(1)}$  converges to  $L'_{\infty,(1)}$  so that (9) becomes in the limit,

$$\left| (1 - \eta)w_{(1)}^t(\gamma)L'_{(1)} + (1 - \eta)w_{(2)}^t(\gamma)L'_{n,(2)} + \eta w(\lambda)^t L'' w(\lambda) \right|_1 \rightarrow 0 \quad (11)$$

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

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

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

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

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 assumption [A4 *non compact*]. 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 the compact case since  $L'_{n,(2)}$  has no influence.

Therefore on  $A_n$

$$|\lambda|^2 + 2|\eta| \leq \delta_n, \quad \sum_{j \geq k_0+1} p_j > n^{-1/2+\epsilon}$$

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

$$\exists i \leq k_0, \quad \text{card}(I_i) \geq 2, \quad \exists j_1, j_2 \in I_i, \quad q_{j_1} > \epsilon/k_0, \quad q_{j_2} > n^{-1/2+\epsilon}$$

without loss of generality we set  $i = 1$  and  $j_1 = 1, j_2 = 2$ . Then we obtain

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

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

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

Also on  $I_1$

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

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

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

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 there exists  $q > 0$  such that

$$V_1 \leq (\log n)^q \delta_n^{d+d/2} \left(1 + \delta_n^{(1-2\epsilon)(\alpha_2-d/2)}\right) \delta_n^{\sum_{j=3}^{t_1} \alpha_j \wedge d/2}$$

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

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

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

$$V_{\mathbf{t}} \leq \delta_n^{k_0-1+dk_0+d/2+\sum_{j=3}^k \alpha_j - \sum_{i=1}^{k_0-1} \alpha_{t_i+1}}$$

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

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

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

$$P^\pi [A_n^c | X^n] = o_p(1). \quad (12)$$

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  $\mathbf{t} = (0, k - k_0 + 1, k - k_0 + 2, \dots, k)$  so that  $t_{k_0} = k$  and define

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

where  $\bar{\gamma}_i = \sum_{j \in I_i} q_j \gamma_j$ ,  $\phi_{\mathbf{t}} = (\gamma_j, j \leq k; s_i, i = 2, \dots, k_0)$  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)

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

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

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

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

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

and

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

We have

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

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

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

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

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

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

$$\pi[B_n] \quad \text{where} \quad B_n := \{|f_0 - f_\theta| \leq \delta_n\} \cap \left\{ \sum_{i=k_0+1}^k p_i \leq n^{-1/2+\epsilon} \right\}$$

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  $\mathbf{t}$

$$|s_i| \leq \delta_n; \quad \left\| \sum_{j \in I_i} q_j \gamma_j - \gamma_i^0 \right\| \leq \delta_n \quad q_{t_i+j} \leq n^{-1/2+\epsilon}, j = 2, \dots, t_{i+1} - 1 \quad \forall i = 1, \dots, k_0$$

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

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

$$\delta_n^{dk_0+k_0-1+\sum_{j \geq k_0+1} \alpha_{\sigma(j)}} n^{\epsilon \sum_{j=k_0+1}^k (\alpha_{\sigma(j)} - d/2)}$$

so that  $P^\pi [B_n | X^n] = o_p(1)$  if  $d(k - k_0)/2 < \sum_{j \geq k_0+1} \alpha_{\sigma(j)} - 2\epsilon \sum_{j=k_0+1}^k (\alpha_{\sigma(j)} - d/2)$  which is satisfied as soon as  $\epsilon < 1/2$ .