# Asymptotic behaviour of the posterior distribution in overfitted mixture models 

Judith Rousseau, Kerrie Mengersen

## To cite this version:

Judith Rousseau, Kerrie Mengersen. Asymptotic behaviour of the posterior distribution in overfitted mixture models. Journal of the Royal Statistical Society: Series B, 2011, 73 (5), pp.689-710. 10.1111/j.1467-9868.2011.00781.x . hal-00641475

HAL Id: hal-00641475

## https://hal.science/hal-00641475

Submitted on 15 Nov 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

# Asymptotic behaviour of the posterior distribution in overfitted mixture models 

Judith Rousseau<br>ENSAE-CREST and CEREMADE, Université Paris Dauphine 3 avenue Pierre Larousse, 92245 Malakoff Cedex France<br>Kerrie Mengersen<br>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

$$
\begin{equation*}
f_{\theta}(x)=\sum_{j=1}^{k} p_{j} g_{\gamma_{j}}(x), \quad k \geq 1, \quad \gamma_{j} \in \Gamma, \quad \theta=\left(p_{1}, \ldots, p_{k}, \gamma_{1}, \ldots, \gamma_{k}\right) \in \Theta_{k} \quad \Gamma \subset \mathbb{R}^{d} \tag{1}
\end{equation*}
$$

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 $X_{1}, \ldots, X_{n}$, iid from a mixture model with $k_{0}$ components:

$$
\begin{equation*}
f_{0}(x)=\sum_{j=1}^{k_{0}} p_{j}^{0} g_{\gamma_{j}^{0}}(x), \quad k \geq 1, \quad \gamma_{j}^{0} \in \Gamma, \quad 1 \leq k_{0}<k \tag{2}
\end{equation*}
$$

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

$$
\theta=\left(p_{1}^{0}, \ldots, p_{k_{0}}^{0}, 0, \gamma_{1}^{0}, \ldots, \gamma_{k_{0}}^{0}, \gamma\right)
$$

for all $\gamma \in \Gamma$ and all values of the parameter in the form $\theta=\left(p_{1}^{0}, ., p_{j} ., p_{k_{0}}^{0}, p_{k+1}, \gamma_{1}^{0}, \ldots, \gamma_{k_{0}}^{0}, \gamma_{j}^{0}\right)$ 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 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 $\gamma$ 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 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 $\Theta_{k}^{0}=\left\{\theta \in \Theta_{k} ; f_{\theta}=f_{0}\right\}$ and let $l_{n}(\theta)$ be the log-likelihood calculated at $\theta$. Denote by $\|f-g\|=\int|f-g|(x) d x$ the $L_{1}$ distance and $\mathbb{P}_{n}(g)=\sum_{i=1}^{n} g\left(X_{i}\right) / n$ and $\mathbb{G}_{n}(g)=\sqrt{n}\left[\mathbb{P}_{n}(g)-F_{0}(g)\right]$ where $F_{0}(g)=\int f_{0}(x) g(x) d x$ and denote by $\operatorname{Leb}(A)$ the Lebesgue measure of a set $A$. We also use the symbol $a \wedge b$ to designate $\min (a, b)$.

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 $\delta \geq 0$

$$
\bar{g}_{\gamma}=\sup _{\left|\gamma^{\prime}-\gamma\right| \leq \delta} g_{\gamma^{\prime}}, \quad \underline{g} \gamma=\inf _{\left|\gamma^{\prime}-\gamma\right| \leq \delta} g_{\gamma^{\prime}}
$$

We now introduce some notations that are useful to characterise $\Theta_{k}^{0}$, following Liu and Shao
(2004)'s presentation. Let $\mathbf{t}=\left(t_{i}\right)_{i=0}^{k_{0}}$ with $0=t_{0}<t_{1}<\ldots<t_{k_{0}} \leq k$ be a partition of $\{1, \ldots, k\}$. 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,

$$
\forall i=1, \ldots, k_{0} \quad \gamma_{t_{i-1}+1}=\ldots=\gamma_{t_{i}}=\gamma_{i}^{0}, \quad p(i)=\sum_{j=t_{i-1}+1}^{t_{i}} p_{j}=p_{i}^{0}, \quad p_{t_{k_{0}}+1}=\ldots=p_{k}=0
$$

In other words $I_{i}=\left\{t_{i-1}+1, \ldots, t_{i}\right\}$ represents the cluster of components in $\{1, \ldots, k\}$ having the same parameter as $\gamma_{i}^{0}$. Then define the following parameterisation of $\theta \in \Theta_{k}$ (up to a permutation)

$$
\phi_{\mathbf{t}}=\left(\left(\gamma_{j}\right)_{j=1}^{t_{k_{0}}},\left(s_{i}\right)_{i=1}^{k_{0}-1},\left(p_{j}\right)_{j=t_{k_{0}+1}}^{k}\right) \in \mathbb{R}^{d t_{k_{0}}+k_{0}+k-t_{k_{0}}-1}, \quad s_{i}=p(i)-p_{i}^{0}, i=1, \ldots, k_{0}
$$

and

$$
\psi_{\mathbf{t}}=\left(\left(q_{j}\right)_{j=1}^{t_{k_{0}}}, \gamma_{t_{k_{0}}+1}, \ldots, \gamma_{k}\right), \quad q_{j}=\frac{p_{j}}{p(i)}, \quad \text { when } \quad j \in I_{i}=\left\{t_{i-1}+1, \ldots ., t_{i}\right\}
$$

Note that $f_{0}$ corresponds to

$$
\phi_{\mathbf{t}}^{0}=\left(\gamma_{1}^{0}, \ldots, \gamma_{1}^{0}, \gamma_{2}^{0}, \ldots, \gamma_{2}^{0}, \ldots, \gamma_{k_{0}}^{0}, \ldots, \gamma_{k_{0}}^{0}, 0 \ldots 0 \ldots 0\right)
$$

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 $\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)$, so that $f_{\theta}=f_{\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)}$ and we denote $f_{\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)}^{\prime}$ and $f_{\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)}^{\prime \prime}$ the first and second derivatives of $f_{\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)}$ with respect to $\phi_{\mathbf{t}}$ and computed at $\theta_{0}=\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)$.

We also denote by $P^{\pi}\left[. \mid X^{n}\right]$ the posterior distribution, where $X^{n}=\left(X_{1}, \ldots, X_{n}\right)$.

## Assumptions

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

$$
\lim _{M \rightarrow+\infty} \limsup _{n} E_{0}^{n}\left\{P^{\pi}\left[\left\|f_{0}-f_{\theta}\right\| \geq M \delta_{n} \mid X^{n}\right]\right\}=0
$$

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

$$
\begin{gathered}
F_{0}\left(\frac{\bar{g}_{\gamma_{i}^{0}}^{3}}{\underline{g}_{\gamma_{i}^{0}}^{3}}\right)<+\infty, \quad F_{0}\left(\frac{\sup _{\left|\gamma-\gamma_{i}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right|^{3}}{g_{\gamma_{i}^{0}}^{3}}\right)<+\infty, \quad F_{0}\left(\frac{\left|\nabla g_{\gamma_{i}^{0}}\right|^{4}}{f_{0}^{4}}\right)<+\infty \\
F_{0}\left(\frac{\sup _{\left|\gamma-\gamma_{i}^{0}\right| \leq \delta}\left|D^{2} g_{\gamma}\right|^{2}}{\underline{g}_{\gamma_{i}^{0}}^{2}}\right)<+\infty, \quad F_{0}\left(\frac{\sup _{\left|\gamma-\gamma_{i}^{0}\right| \leq \delta}\left|D^{3} g_{\gamma}\right|}{\underline{g}_{\gamma_{i}^{0}}}\right)<+\infty
\end{gathered}
$$

Assume also that for all $i=1, \ldots, k_{0} \gamma_{i}^{0} \in \operatorname{int}(\Gamma)$ the interior of $\Gamma$.
A3 Integrability: There exists $\Gamma_{0} \subset \Gamma$ satisfying $\operatorname{Leb}\left(\Gamma_{0}\right)>0$, and for all $i \leq k_{0}$

$$
d\left(\gamma_{i}^{0}, \Gamma_{0}\right)=\inf _{\gamma \in \Gamma_{0}}\left|\gamma-\gamma_{i}^{0}\right|>0
$$

and such that for all $\gamma \in \Gamma_{0}$,

$$
F_{0}\left(\frac{g_{\gamma}^{4}}{f_{0}^{4}}\right)<+\infty, \quad F_{0}\left(\frac{g_{\gamma}^{3}}{\underline{g}_{\gamma_{i}^{0}}^{3}}\right)<+\infty, \quad \forall i \leq k_{0}
$$

A4 Stronger identifiability : For all $\mathbf{t}$ partitions of $\{1, \ldots, k\}$ as defined above, let $\theta \in \Theta_{k}$ and write $\theta$ as $\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)$; then

$$
\begin{align*}
& \left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} f_{\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}}^{\prime}+\frac{1}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} f_{\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}}^{\prime \prime}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)=0 \Leftrightarrow \\
& \forall i \leq k_{0}, s_{i}=0 \quad \& \quad \forall j \in I_{i} \quad q_{j}\left(\gamma_{j}-\gamma_{i}^{0}\right)=0, \quad \forall i \geq t_{k_{0}}+1 \quad p_{i}=0 \tag{3}
\end{align*}
$$

Assuming also that if $\gamma \notin\left\{\gamma_{1}, \ldots, \gamma_{k}\right\}$ 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 $j=1, \ldots, k$ 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 $\Gamma$ 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 $\int g(x) d \mu(x) \in\{+\infty, 0\}$ 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 $D^{2} g_{\gamma_{i}}^{0}, i=1, \ldots, k_{0}$.

A5 Prior : The prior density, with respect to Lebesgue measure on $\Theta$, is continuous and positive and the prior $\pi(p)$ on $\left(p_{1}, \ldots, p_{k}\right)$ satisfies

$$
\pi(p)=C(p) p_{1}^{\alpha_{1}-1} \ldots . p_{k}^{\alpha_{k}-1}
$$

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 $\delta_{n}$ 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 $\Gamma$ 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 $g_{\gamma}, \nabla g_{\gamma}$ 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, 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 $g_{\gamma}$ 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 $\lambda^{x}\left(a_{1}+a_{2} x+a_{3} x^{2}\right)$ 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 $\mathcal{D}\left(\alpha_{1}, \ldots, \alpha_{k}\right)$.

### 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 $\{1, \ldots, k\}, \bar{\alpha}=\max \left(\alpha_{j}, j \leq k\right)$ and $\underline{\alpha}=\min \left(\alpha_{j}, j \leq k\right)$

- If $\bar{\alpha}<d / 2$, set $\rho=\left(d k_{0}+k_{0}-1+\bar{\alpha}\left(k-k_{0}\right)\right) /(d / 2-\bar{\alpha})$ then

$$
\lim _{M \rightarrow+\infty} \limsup _{n} E_{0}^{n}\left\{P^{\pi}\left[\min _{\sigma \in \mathcal{S}_{k}} \sum_{i=k_{0}+1}^{k} p_{\sigma(i)}>M n^{-1 / 2}(\log n)^{q(1+\rho)} \mid X^{n}\right]\right\}=0
$$

- If $\underline{\alpha}>d / 2$ and $\rho^{\prime}=\left(d k_{0}+k_{0}-1+d\left(d-k_{0}\right) / 2\right) /\left((\underline{\alpha}-d / 2)\left(k-k_{0}\right)\right)$

$$
\lim _{\epsilon \rightarrow 0} \limsup _{n} E_{0}^{n}\left\{P^{\pi}\left[\min _{\sigma \in \mathcal{S}_{k}} \sum_{i=k_{0}+1}^{k} p_{\sigma(i)}<\epsilon(\log n)^{-q\left(1+\rho^{\prime}\right)} \mid X^{n}\right]\right\}=0
$$

Recall that $\left(\alpha_{1}, \ldots, \alpha_{k}\right)$ 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 $\max \left(\alpha_{j}, j \leq k\right)<d / 2$, the posterior estimates verify

$$
\sum_{j=k_{0}+1}^{k} E^{\pi}\left[p_{j} \mid X^{n}\right]=O_{p}\left(n^{-1 / 2}(\log n)^{q(\rho+1)}\right)
$$

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 $O_{p}\left(n^{-1 / 2}\right)$.

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 $\min \left(\alpha_{j}, j \leq k\right)>d / 2$ 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 $\min \left(\alpha_{j}, j \leq\right.$ $k) \leq d / 2 \leq \max \left(\alpha_{j}, j \leq k\right)$, 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 $\mathcal{D}(1 / 2, \ldots, 1 / 2)$ 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 $\left\{\gamma ;\left|\gamma_{0}-\gamma\right| \leq \epsilon\right\}$ 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 $\left(\alpha_{j}\right)$ 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 $A_{n}=$ $\left\{\min _{\sigma \in \mathcal{S}_{k}} \sum_{i \geq k_{0}+1} p_{\sigma(i)}>n^{-1 / 2} M_{n}\right\} \cap\left\{\left\|f_{0}-f_{\theta}\right\| \leq \delta_{n}\right\}$, 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

$$
P^{\pi}\left[A_{n} \mid X^{n}\right]=\frac{\int_{A_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi(\theta)}{\int e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi(\theta)}:=\frac{N_{n}}{D_{n}}
$$

where $l_{n}(\theta)$ is the log-likelihood and $\theta_{0} \in \Theta_{0}$, and we need to prove that $P^{\pi}\left[A_{n} \mid X^{n}\right]=o_{p}(1)$; 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 ,

$$
D_{n} \geq C n^{-\left(d k_{0}+k_{0}-1+\sum_{j \geq k_{0}+1} \alpha_{\sigma(j)}\right) / 2}
$$

for any permutation $\sigma$ of $\{1, \ldots, k\}$, by considering approximations of $\Theta_{0}$ along paths of the form:

$$
\left|\gamma_{\sigma(i)}-\gamma_{i}^{0}\right| \leq n^{-1 / 2}, \quad\left|p_{\sigma(i)}-p_{i}^{0}\right| \leq n^{-1 / 2} \quad \sum_{j \geq k_{0}+1} p_{\sigma(j)} \leq n^{-1 / 2}
$$

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 $\Theta_{0}$ where at least two components merge, in other words they are associated with partitions $\mathbf{t}$ of $\left\{1, \ldots, k_{0}\right\}$ such that there exists $i \leq k_{0}-1$ with $t_{i+1} \geq t_{i}+2$ and at least two components merging with $\gamma_{i}^{0}$ have weights much larger than $n^{-1 / 2}$. We prove in the appendix that each of these paths has a prior mass bounded by $o_{p}\left(D_{n}\right)$ when $d / 2>\max \left\{\alpha_{j}, j=1, \ldots, k\right\}$. In this case, $d k_{0}+k_{0}-1+$ $\min _{\sigma \in \mathcal{S}_{k}} \sum_{j \geq k_{0}+1} \alpha_{j}$ appears as an effective dimension of the model, which is different from the number of parameters, $d k+k-1$, or even from some "effective number of parameters" that would be given by the number of parameters used to parameterize the path $\min _{\sigma \in \mathcal{S}_{k}} \sum_{j \geq k_{0}+1} p_{\sigma(j)}=$ $O\left(n^{-1 / 2}\right)$, due to the influence of the prior via the $\alpha_{j}$ 's. This can be understood as a measure of complexity of the model induced by the prior.

In the case where $d / 2<\min \left\{\alpha_{j}, j=1, \ldots, k\right\}$, we need to prove that $P^{\pi}\left[B_{n} \mid X^{n}\right]=o_{p}(1)$, with $B_{n}=\left\{\min _{\sigma \in \mathcal{S}_{k}} \sum_{i \geq k_{0}+1} p_{\sigma(i)}<\epsilon_{n}\right\} \cap\left\{\left\|f_{0}-f_{\theta}\right\| \leq \delta_{n}\right\}$ with $\epsilon_{n}=\epsilon(\log n)^{-q\left(1+\rho^{\prime}\right)}$. 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 $\Theta_{0}$ along paths of the following form: if $I_{1}=\left\{1, \ldots, k-k_{0}+1\right\}, I_{i}=\left\{k-k_{0}+i\right\}$,

$$
\begin{aligned}
& i=2, \ldots, k_{0} \\
& \left|\sum_{j \in I_{i}} \frac{p_{j}}{\sum_{j \in I_{i}} p_{j}} \gamma_{j}-\gamma_{i}^{0}\right| \leq n^{-1 / 2}, \quad\left|\sum_{j \in I_{i}} p_{j}-p_{i}^{0}\right| \leq n^{-1 / 2} \quad \forall j \in I_{i}, i=1, \ldots, k_{0}\left|\gamma_{j}-\gamma_{i}^{0}\right| \leq n^{-1 / 4},
\end{aligned}
$$

i.e. by forcing all the parameters of the extra components to be close to $\gamma_{1}^{0}$. This leads to

$$
D_{n} \geq C n^{-0.5\left(k_{0} d+k_{0}-1+d\left(k-k_{0}\right) / 2\right)}
$$

with large probability whereas $\pi\left(B_{n}\right)=o_{p}\left(D_{n}\right)$ so that

$$
P^{\pi}\left[B_{n} \mid X^{n}\right]=o_{p}(1)
$$

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: $d / 2<\min \left(\alpha_{j}\right)$ and $d / 2>\max \left(\alpha_{j}\right)$.

### 3.1. Simulated example

We consider a very simple study of fitting a two-component Gaussian mixture model to a sample of data, $Y=\left\{y_{i}, i=1, . ., n\right\}$, 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 $\mathcal{N}(0,1)=\mathcal{N}(0,1)$. 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 $\mathcal{N}\left(0,10^{4}\right)$ 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.

$$
\begin{aligned}
\text { Generating distribution : } & y_{i} \sim \mathcal{N}(0,1) \in \mathbb{R} \\
\text { Model } & G=p \mathcal{N}\left(\mu_{1}, 1\right)+(1-p) \mathcal{N}\left(\mu_{2}, 1\right),
\end{aligned}
$$

where $\mathcal{N}(\mu, \tau)$ denotes the univariate normal distribution with mean $\mu$ and variance $\tau$. In this case Theorem 1 implies that $P^{\pi}\left[p>1-\epsilon \mid X^{n}\right](p>1-p)$ is small as $n$ gets large for $\epsilon$ 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 $\mu_{1}$ and $\mu_{2}$ 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.

$$
\begin{aligned}
\text { Generating distribution : } & y_{i} \sim \mathcal{N}_{2}(0,1) \in \mathbb{R}^{2} \\
\text { Model } & G=p \mathcal{N}_{2}\left(\mu_{1}, \Sigma_{1}^{2}\right)+(1-p) \mathcal{N}_{2}\left(\mu_{2}, \Sigma_{2}^{2}\right)
\end{aligned}
$$

where $\mu_{j}=\left(\mu_{j 1}, \mu_{j 2}\right)^{\prime}$ and $\Sigma_{j}$ is the diagonal matrix with diagonal elements $\left(\sigma_{j 1}^{2}, \sigma_{j 2}^{2}\right)$. 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.

$$
\begin{aligned}
\text { Generating distribution : } & y_{i} \sim \mathcal{N}(0,1) \in \mathbb{R} \\
\text { Model } & G=p \mathcal{N}\left(\mu_{1}, \sigma_{1}^{2}\right)+(1-p) \mathcal{N}\left(\mu_{2}, \sigma_{2}^{2}\right),
\end{aligned}
$$

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

$n=50,100,500,1000$

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

$n=50,100,500,1000,5000,10000$

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 $y \sim \sum_{j=1}^{k} p_{j} \mathcal{N}\left(\mu_{j}, \sigma_{j}^{2}\right)$. At the school level, we let $y_{i}$ denote the average maths score for the $i$ th 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, $i=1, \ldots, n=4500$. In all instances the prior on the $\mu$ 's were diffuse normal priors : $\mathcal{N}(0,10000)$ and the priors on the $\sigma$ 's were $\mathcal{U}(0,100)$.

First, a single normal model was fitted using a diffuse normal prior. The posterior estimates of $\mu$ and $\sigma$ 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 $\left(\alpha_{1}=\alpha_{2}=1=d / 2\right)$ prior was placed on $\left(p_{1}, p_{2}\right)$. We obtained the following posterior estimates and $95 \%$ credible intervals for the parameters $p_{1}=0.07,\left(4.610^{-4}, 0.62\right), p_{2}=$ $0.93(0.38,1.0), \mu_{1}=315.3(128.7,451.4), \mu_{2}=476.4$ (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) $p_{1}=0.07\left(4.410^{-4}, 0.56\right), p_{2}=0.36$ $\left(4.210^{-3}, 0.94\right), p_{3}=0.58(0.03,0.99), \mu_{1}=301.5(114.6,454.9), \mu_{2}=436.9(467.2,581.8), \mu_{3}=$ 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 : $p_{1}=0.007\left(4.710^{-6}, 0.04\right), p_{2}=0.20,\left(4.710^{-5}, 0.98\right)$ and $p_{3}=0.80,\left(8.510^{-3}, 0.98\right)$. 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 $0.029\left(1.310^{-3}, 9.910^{-2}\right), 0.72(0.35,0.92)$ 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,
$\left(y_{i 1}, y_{i 2}\right)$ denotes the average maths and science scores for the $i$ th school, $i=1, \ldots, 150$, and $\left(y_{1}, y_{2}\right)^{\prime} \sim \mathcal{N}_{2}\left(\left(\mu_{1}, \mu_{2}\right)^{\prime}, \Sigma\right)$. 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 $0.073\left(210^{-3}, 0.25\right)$. 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 $0.0067\left(4.410^{-7}, 4.310^{-2}\right)$. 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 (0.760.98 ) 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 $\mathrm{k}=3$ 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 studentlevel 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 $\gamma(d \leq 2)$ it becomes necessary to favour small weights with a prior in the form $p_{1}^{-1 / 2} \ldots p_{k}^{-1 / 2}$, 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 $\min \left(\alpha_{j}\right) \leq d / 2 \leq \max \left(\alpha_{j}\right)$, 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: SpringerVerlag, 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 $A_{n}^{\prime}=\left\{\exists I=\left\{j_{1}, \ldots, j_{k-k_{0}}\right\}, \sum_{i \in I} p_{i}>M_{n} \delta_{n}\right\}$ for some large $M_{n}$ (depending on $\delta_{n} M_{n}$ will be chosen as a power of $\log n$ or a a large constant. The posterior probability of interest is bounded by

$$
\begin{aligned}
P^{\pi}\left[A_{n}^{\prime} \mid X^{n}\right] & =P^{\pi}\left[A_{n}^{\prime} \cap\left\{\left\|f-f_{0}\right\| \leq M \delta_{n}\right\} \mid X^{n}\right]+o_{P}(1) \\
& =\frac{\int_{A_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta)}{\int_{\left\|f_{0}-f_{\theta}\right\| \leq M \delta_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta)}+o_{P}(1)
\end{aligned}
$$

where $A_{n}=A_{n}^{\prime} \cap\left\{\left\|f-f_{0}\right\| \leq M \delta_{n}\right\}$ and $M$ is a fixed positive constant. We denote by

$$
\begin{equation*}
N_{n}=\int_{A_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta) \quad \text { and } \quad D_{n}=\int_{\left\|f_{0}-f_{\theta}\right\| \leq M \delta_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta) \tag{4}
\end{equation*}
$$

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 $C_{\epsilon}$ and there exists a permutation $\sigma:\{1, \ldots, k\} \rightarrow\{1, \ldots, k\}$, s.t.

$$
\begin{equation*}
P_{0}^{n}\left[D_{n} \geq C_{\epsilon} n^{-\frac{d k_{0}+k_{0}-1+\sum_{j=k_{0}+1}^{k} \alpha_{\sigma(j)}}{2}}\right]>1-\epsilon, \quad \pi\left(A_{n}\right) \leq C \frac{\delta_{n}^{d k_{0}+k_{0}-1+\sum_{j=k_{0}+1}^{k} \alpha_{\sigma(j)}}}{M_{n}^{\left(d / 2-\max _{j} \alpha_{j}\right)}} \tag{5}
\end{equation*}
$$

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

$$
E_{0}^{n}\left[P^{\pi}\left[A_{n} \mid X^{n}\right]\right]=o(1)
$$

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

$$
\begin{equation*}
P_{0}^{n}\left[D_{n} \geq C_{\epsilon} n^{-\left(d k_{0}+k_{0}-1+d\left(k-k_{0}\right) / 2\right) / 2}\right]>1-\epsilon, \quad \pi\left(A_{n}^{c}\right) \leq C \max _{\sigma} \delta_{n}^{d k_{0}+k_{0}-1+\sum_{j=k_{0}+1}^{k} \alpha_{\sigma(j)}} \tag{6}
\end{equation*}
$$

which leads to

$$
E_{0}^{n}\left[P^{\pi}\left[A_{n}^{c} \mid X^{n}\right]\right]=o(1)
$$

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

$$
D_{n} \geq \int_{S_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta)
$$

where
$S_{n}=\left\{\left(p_{1}, \ldots, p_{k}, \gamma_{1}, \ldots, \gamma_{k}\right) ;\left|p_{j}-p_{j}^{0}\right| \leq n^{-1 / 2} ;\left|\gamma_{j}-\gamma_{j}^{0}\right| \leq n^{-1 / 2}, j=1, \ldots, k_{0} ;\left|\gamma_{j}-\gamma_{j}^{*}\right| \leq \epsilon_{1}, j \geq k_{0}+1\right\}$
where $\gamma_{j}^{*} \in \Gamma_{0}, j \geq k_{0}+1$ and satisfy $\min _{k_{0}<l \neq j}\left|\gamma_{j}^{*}-\gamma_{l}^{*}\right|>C \epsilon_{1}$, with $C, \epsilon_{1}>0$ fixed. By definition of $\Gamma_{0}, \min _{l \leq k_{0}}\left|\gamma_{j}^{*}-\gamma_{l}^{0}\right|>C \epsilon_{1}$ 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 $\Theta_{0}$ corresponds to the partition $\mathbf{t}=\left(0,1,2, \ldots, k_{0}\right)$. Let $\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)$ be the parameterisation (of $\theta$ ) associated to the partition $\mathbf{t}$. We consider a Taylor expansion of $l_{n}\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)-l_{n}\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)$, corresponding to $\theta=\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right) \in S_{n}$ and $\theta_{0}=\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)$. By convention and without loss of generality we write $p_{j}^{0}=0$ and $\gamma_{j}^{0}=\gamma_{j}^{*}$ for $j=k_{0}+1, \ldots, k$ and following the definition of $\phi_{\mathbf{t}}$, $p_{k_{0}}=1-\sum_{i \neq k_{0}} p_{i}$. Then

$$
\begin{equation*}
l_{n}\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)-l_{n}\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)=\sqrt{n}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} U_{n}-\frac{n}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} J(\bar{\theta})\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right) \tag{7}
\end{equation*}
$$

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

$$
\begin{aligned}
U_{n}(i) & =\mathbb{G}_{n}\left(\frac{\nabla_{l} g_{\gamma_{j}^{0}}}{f_{0}}\right), \quad i=l+(j-1) * d, \quad j \leq k_{0} \quad l=1, \ldots, d \\
U_{n}(i) & =\mathbb{G}_{n}\left(\frac{f_{\gamma_{j}^{0}}-f_{\gamma_{1}^{0}}}{f_{0}}\right), \quad i=k_{0} d+j-\mathbb{1}_{j \geq k_{0}+1}, \quad 1 \leq j \neq k_{0} \leq k
\end{aligned}
$$

then $U_{n}=0_{p}(1)$. Denote $i d$ the identity matrix and $\Omega_{n}\left(c_{0}, C\right)=\left\{X^{n} ; \sup _{\theta \in S_{n}}\|J(\theta)\| \leq c_{0} n ;\left|U_{n}\right| \leq\right.$ $C\}$. Therefore the log likelihood ratio is bounded from below by $\sqrt{n}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} U_{n}-\frac{C_{0} n}{2}\left\|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right\|^{2}$ for some positive constant $C_{0}$ on $\Omega_{n}\left(c_{0}, C\right)$. This leads to, on $\Omega_{n}\left(c_{0}, C\right)$ :

$$
\begin{aligned}
\int_{S_{n}} e^{l_{n}(\theta)-l_{n}\left(\theta_{0}\right)} d \pi_{k}(\theta) & \geq e^{\frac{1}{2 C_{0}}\left\|U_{n}\right\|^{2}} \int_{S_{n}} e^{-\frac{n C_{0}}{2}\left\|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}-C_{0}^{-1} U_{n} / \sqrt{n}\right\|^{2}} d \pi_{k}(\theta) \\
& \geq \int_{S_{n}} e^{-\frac{n C_{0}}{2}\left\|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}-C_{0}^{-1} U_{n} / \sqrt{n}\right\|^{2}} d \pi_{k}(\theta)
\end{aligned}
$$

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 $c_{1} p_{k_{0}+1}^{\alpha_{k_{0}+1}-1} \ldots p_{k}^{\alpha_{k}-1}$. Thus on $\Omega_{n}\left(c_{0}, C\right)$, we have

$$
\begin{aligned}
\int_{S_{n}} e^{-\frac{C_{0} n}{2}\left\|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}-C_{0}^{-1} U_{n} / \sqrt{n}\right\|^{2}} \pi_{k}(\theta) d \theta & \geq n^{-\left(d k_{0}+k_{0}-1\right) / 2} \prod_{j=k_{0}+1}^{k} \int_{0}^{\delta_{n} / k_{0}} e^{-\frac{n C_{0}}{2}\left(p_{j}-U_{n}(j)\right)^{2}} p_{j}^{\alpha_{j}-1} d p_{j} \\
& \geq n^{-\left(d k_{0}+k_{0}-1+\sum_{j>k_{0}}\left(\alpha_{j}-1\right)\right) / 2} \prod_{j=k_{0}+1}^{k} \int_{c / \sqrt{n}}^{\delta_{n} / k_{0}} e^{-\frac{n C_{0}}{2}\left(p_{j}-U_{n}(j)\right)^{2}} d p_{j} \\
& \geq n^{-\left(d k_{0}+k_{0}-1+\sum_{j>k_{0}} \alpha_{j}\right) / 2} \prod_{j=k_{0}+1}^{k} \Phi\left(C_{1}\right) \\
& \geq n^{-\left(d k_{0}+k_{0}-1+\sum_{j>k_{0}} \alpha_{j}\right) / 2}
\end{aligned}
$$

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 $\sigma^{*}$ 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

$$
P_{0}^{n}\left[D_{n}<a n^{-\left(d k_{0}+k_{0}-1+\sum_{j>k_{0}} \alpha_{j}\right) / 2}\right] \leq P_{0}^{n}\left[\Omega_{n}^{c}\left(c_{0}, C\right)\right] .
$$

The lower bound in (5) is then proved by determining an upper bound on $P_{0}^{n}\left[\Omega_{n}^{c}\left(c_{0}, C\right)\right]$. Note first that for all $\epsilon>0$ there exists $C>0$ such that with probability greater than $1-\epsilon,\left|U_{n}\right| \leq C$. Then we bound for each $i, i^{\prime} \leq k-1+k_{0} d$, and some $c>0$ small enough $P_{0}^{n}\left[J\left(i, i^{\prime}\right)-n I\left(i, i^{\prime}\right)<c n\right]$, where $I$ is a Fisher information matrix defined as $E_{0}^{n}\left[J\left(\theta_{0}\right)\right]$ with the constraint here that $\theta_{0}=$ $\left(p_{1}^{0}, \ldots, p_{k_{0}}^{0}, 0, \ldots, 0, \gamma_{1}^{0}, \ldots \gamma_{k}^{0}\right)$ writing $\gamma_{j}^{0}=\gamma_{j}^{*}$ when $j=k_{0}+1, \ldots, k$. We have if $i, i^{\prime} \geq d k_{0}+1$,

$$
J\left(i, i^{\prime}\right)-n I\left(i, i^{\prime}\right)=\sqrt{n} \mathbb{G}_{n}\left(\frac{\left(g_{\gamma_{j}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)\left(g_{\gamma_{j^{\prime}}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)}{f_{0}^{2}}\right)+n \mathbb{P}_{n}\left[\Delta_{\bar{\theta}}\left(i, i^{\prime}\right)\right]
$$

with $j, j^{\prime}$ the indices corresponding to $i, i^{\prime}$ as in the definition of $U_{n}$,

$$
\Delta_{\bar{\theta}}\left(i, i^{\prime}\right)=\left(\frac{\left(g_{\bar{\gamma}_{j}}-g_{\bar{\gamma}_{k_{0}}}\right)\left(g_{\bar{\gamma}_{j^{\prime}}}-g_{\bar{\gamma}_{k_{0}}}\right)}{f_{\bar{\theta}}^{2}}\right)-\left(\frac{\left(g_{\gamma_{j}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)\left(g_{\gamma_{j^{\prime}}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)}{f_{0}^{2}}\right)
$$

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

$$
C n^{-1} F_{0}\left[\left(\frac{\left(g_{\gamma_{j}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)\left(g_{\gamma_{j^{\prime}}^{0}}-g_{\gamma_{k_{0}}^{0}}\right)}{f_{0}^{2}}\right)^{2}\right] \leq \frac{C k_{0}}{n \min _{l \leq k_{0}}\left(p_{l}^{0}\right)^{4}}+n^{-1} \max _{\gamma \in \Gamma_{0}} F_{0}\left(\frac{g_{\gamma}^{4}}{f_{0}^{4}}\right)
$$

Assumption [A3] implies that the second term on the right hand side of the above inequality is of order $O\left(n^{-1}\right)$, so that the above probability is $O\left(n^{-1}\right)$. To study the behaviour of $\Delta_{\theta}\left(i, i^{\prime}\right)$ we consider its derivatives : if $i, i^{\prime}>d k_{0}$,

$$
\begin{aligned}
\left|\frac{\partial \Delta_{\theta}\left(i, i^{\prime}\right)}{\partial p_{l}}\right| & =\left|\frac{\left(g_{\gamma_{j}}-g_{\gamma_{k_{0}}}\right)\left(g_{\gamma_{j^{\prime}}}-g_{\gamma_{k_{0}}}\right)\left(g_{\gamma_{l}}-g_{\gamma_{k_{0}}}\right)}{f_{\theta}^{3}}\right| \\
& \leq \frac{\left[\bar{g}_{\gamma_{j}^{0}}+\bar{g}_{\gamma_{k_{0}}}\right]\left[\bar{g}_{\gamma_{j^{\prime}}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right]\left[\bar{g}_{\gamma_{l}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right]}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0} \underline{g}_{\gamma_{j}^{0}}\right)^{3}}
\end{aligned}
$$

and if $i, i^{\prime}>d k_{0}, l \leq k$,

$$
\begin{aligned}
& \left|\frac{\partial \Delta_{\theta}\left(i, i^{\prime}\right)}{\partial \gamma_{l}}\right|=\left\lvert\, \frac{\left(g_{\gamma_{j}}-g_{\gamma_{k_{0}}}\right)\left(g_{\gamma_{j^{\prime}}}-g_{\gamma_{k_{0}}}\right) \nabla g_{\gamma_{l}}}{f_{\theta}^{3}}+\mathbb{1}_{j=l} \frac{\nabla g_{\gamma_{j}}\left(g_{\gamma_{i^{\prime}}}-g_{\gamma_{k_{0}}}\right)}{f_{\theta}^{2}}+\mathbb{1}_{l=j^{\prime}} \frac{\nabla g_{\gamma_{j^{\prime}}}\left(g_{\gamma_{j}}-g_{\gamma_{k_{0}}}\right)}{f_{\theta}^{2}}\right. \\
& \left.-\mathbb{1}_{l=k_{0}} \frac{\nabla g_{\gamma_{k_{0}}}\left(g_{\gamma_{j^{\prime}}}+g_{\gamma_{j}}-2 g_{\gamma_{k_{0}}}\right)}{f_{\theta}^{2}} \right\rvert\, \\
& \leq \frac{\left[\bar{\gamma}_{\gamma_{j}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}^{0}\left[\bar{g}_{\gamma_{j^{\prime}}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right] \sup _{\mid \gamma-\gamma_{l}^{0}}^{0} \leq \delta\right.}{}\left|\nabla g_{\gamma}\right| \left\lvert\, 1_{j=l} \frac{\sup _{\left|\gamma-\gamma_{j}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right|\left(\bar{h}_{\gamma_{j^{\prime}}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right)}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0} \underline{\gamma}_{\gamma_{j}^{0}}\right)^{3}}\right.
\end{aligned}
$$

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

$$
F_{0}\left(\sup _{\theta \in S_{n}}\left|\frac{\partial \Delta_{\gamma}\left(i, i^{\prime}\right)}{\partial \gamma_{l}}\right|\right) \leq M<+\infty \quad \forall l \leq k, \forall i, i^{\prime}>d k_{0},
$$

so that for all $c>0$, there exist $\delta_{0}$ such that for all $\delta<\delta_{0}$,

$$
P_{0}\left[\mathbb{P}_{n}\left|\sup _{\theta \in S_{n}}\right| \Delta\left(i, i^{\prime}\right)| |>c\right] \leq \frac{\delta M}{c},
$$

which can be made as small as necessary. Similarly if $i>d k_{0}$ and $i^{\prime} \leq d k_{0}$,

$$
J\left(i, i^{\prime}\right)-n I\left(i, i^{\prime}\right)=\sqrt{n} \mathbb{G}_{n}\left(\frac{\left(g_{\gamma_{j}^{0}}-g_{\gamma_{\gamma_{0}^{0}}^{0}}\right) \nabla g_{\gamma_{j^{\prime}}^{0}}}{f_{0}^{2}}\right)+n \mathbb{P}_{n}\left[\Delta_{\bar{\theta}}\left(i, i^{\prime}\right)\right]
$$

with

$$
\Delta_{\bar{\theta}}\left(i, i^{\prime}\right)=\frac{\left(g_{\bar{\gamma}_{j}}-g_{\gamma_{k_{k}}}\right) \nabla g_{\bar{\gamma}_{j^{\prime}}}}{f_{\bar{\theta}}}-\frac{\left(g_{\gamma_{j}^{0}}-g_{\gamma_{k_{0}^{0}}^{0}}\right) \nabla g_{\gamma_{j^{\prime}}}}{f_{0}^{2}}
$$

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

$$
\frac{\sup _{\left|\gamma-\gamma_{j}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right| \sup _{\left|\gamma-\gamma_{j^{\prime}}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right|}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0} \underline{q}_{\gamma_{j}^{0}}\right)^{2}}, \frac{\sup _{\left|\gamma-\gamma_{j}^{0}\right| \leq \delta}\left|D^{2} g_{\gamma}\right|\left(\bar{g}\left(\gamma_{j^{\prime}}^{0}\right)+\bar{g}_{\gamma_{k_{0}^{0}}}\right)}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0}{\underline{q_{\gamma}^{0}}}_{j}\right)^{2}}
$$

and

$$
\frac{\sup _{\left|\gamma-\gamma_{j^{\prime}}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right|\left[\bar{g}_{\gamma_{i}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right]\left[\bar{g}_{\gamma_{j}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right]}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0}{\underline{\gamma_{j}^{0}}}^{\gamma_{j}^{0}}\right)^{3}}, \frac{\sup _{\left|\gamma-\gamma_{j^{\prime}}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right| \sup _{\left|\gamma-\gamma_{j}^{0}\right| \leq \delta}\left|\nabla g_{\gamma}\right|\left[\bar{g}_{\gamma_{i}^{0}}+\bar{g}_{\gamma_{k_{0}}^{0}}\right]}{\left(1-\delta_{n}\right)^{3}\left(\sum_{j=1}^{k} p_{j}^{0}{\underline{\gamma_{\gamma}^{\prime}}}\right)^{3}}
$$

so that

$$
P_{0}^{n}\left[\left|\mathbb{P}_{n}\left[\sup _{\theta \in S_{n}}\left|\Delta\left(i, i^{\prime}\right)\right|\right]\right|<c\right] \leq C \delta / c
$$

The same calculations can be made for the terms $J\left(i, i^{\prime}\right)$ when $i, i^{\prime} \leq d k_{0}$, so that finally there exists $c_{0}, C>0$ such that for all $\theta \in S_{n} P_{0}^{n}\left[\Omega_{n}^{c}\left(c_{0}, C\right)\right] \leq 2 \epsilon$ and the lower bound of $D_{n}$ in (5) is established.

To bound $\pi\left(A_{n}\right)$ in (5), we need to characterize $\theta \in A_{n}$. For each $\theta \in A_{n}$, as $n$ increases $\theta$ converges towards $\Theta_{0}=\cup_{\mathbf{t}} \Theta_{0 \mathbf{t}}$, i.e. $\min _{\sigma, \mathbf{t}}\left|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right| \rightarrow 0$, where the minimum is taken over the set of possible permutations of $\{1, \ldots, k\}$ 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 $j \geq t_{k_{0}}+1, \gamma_{j} \notin\left\{\gamma_{1}^{0}, \ldots, \gamma_{k_{0}}^{0}\right\}$. Consider a partition $\mathbf{t}$ and a permutation $\sigma$ which minimizes $\left|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right|$, for a given $\theta$. A Taylor expansion of $f_{\theta}$ in terms of $\phi_{\mathbf{t}}$ around $\phi_{\mathbf{t}}^{0}$ leads to:

$$
f_{\theta}=f_{0}+\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} f_{\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)}^{\prime}+\frac{1}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} f_{\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)}^{\prime \prime}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)+\frac{1}{6}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{(3)} f_{\left(\bar{\phi}_{\mathbf{t}}, \psi_{\mathbf{t}}\right)}^{(3)}
$$

where $\bar{\phi}_{\mathbf{t}} \in\left(\phi_{\mathbf{t}}, \phi_{\mathbf{t}}^{0}\right)$. The last term of right hand side of the above equation is bounded by $C\left|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right|^{3}$ in $L_{1}$ for some positive constant $C>0$, it is thus $o\left(\left|\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right|^{2}\right)$. Define then

$$
\begin{aligned}
\gamma & =\gamma(\mathbf{t}, \sigma)=\left(\left[\left(s_{i}+p_{i}^{0}\right)\left(\sum_{j \in I_{i}} q_{j} \gamma_{j}-\gamma_{i}^{0}\right)\right]_{i=1}^{k_{0}},\left(s_{i}\right)_{i=2}^{k_{0}},\left(p_{j}\right)_{j=t_{k_{0}}+1}^{k}\right) \\
\lambda & =\lambda(\mathbf{t}, \sigma)=\left(\left[\sqrt{q_{j}}\left(\gamma_{j}-\gamma_{i}^{0}\right)\right]_{j \in I_{i}}, i=1, \ldots, k_{0}\right) .
\end{aligned}
$$

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

$$
\begin{equation*}
f_{\theta}-f_{0}=\frac{1}{2}\left(2|\gamma|+|\lambda|^{2}\right)\left((1-\eta) w(\gamma)^{T} L^{\prime}+\eta w(\lambda)^{T} L^{"} w(\lambda)+o(1)\right) \tag{8}
\end{equation*}
$$

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

$$
\begin{aligned}
L^{\prime} & =\left(\left(\nabla g_{\gamma_{1}^{0}}\right)^{T}, \ldots,\left(\nabla g_{\gamma_{k_{0}}^{0}}\right)^{T}, g_{\gamma_{1}^{0}}-g_{\gamma_{k_{0}}^{0}}, \ldots, g_{\gamma_{k_{0}-1}^{0}}-g_{\gamma_{k_{0}}^{0}}, g_{\gamma_{k_{k_{0}}+1}^{0}}-g_{\gamma_{k_{0}}^{0}}, \ldots, g_{\gamma_{k}^{0}}-g_{\gamma_{k_{0}}^{0}}\right) \\
L^{\prime \prime} & =\operatorname{diag}\left(p_{1}^{0} D^{2} g_{\gamma_{1}^{0}}, \ldots, p_{k_{0}}^{0} D^{2} g_{\gamma_{k_{0}}^{0}}\right)
\end{aligned}
$$

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

$$
\begin{equation*}
\left|\left(1-\eta_{r_{n}}\right) w\left(\gamma_{r_{n}}\right)^{T} L_{r_{n}}^{\prime}+\eta_{r_{n}} w\left(\lambda_{r_{n}}\right)^{T} L " w\left(\lambda_{r_{n}}\right)\right| \leq c_{n} \tag{9}
\end{equation*}
$$

Thus to prove that $|\lambda|^{2}+2|\eta| \leq \delta_{n} / c$ for some $c$, it is enough to find a subsequence of $\theta_{r_{n}}$ 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 $w\left(\gamma_{n}\right), w\left(\lambda_{n}\right), \eta_{n}$ 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 $w\left(\gamma_{n}\right), w\left(\lambda_{n}\right), \eta_{n}$. Despite the notation $w(\gamma)$ and $w(\lambda)$ the above statement does not imply that $\gamma_{n}$ or $\lambda_{n}$ converges towards $\gamma$ and $\lambda$.

We first consider the case where $\Gamma$ is compact. Then $\theta_{n}$ belongs to a compact set and there exists a subsequence such that $L_{n}^{\prime}$ converges to some vector $L_{\infty}^{\prime}$ corresponding to some $\theta \in \Theta_{0}$. At the limit, inequality (9) becomes

$$
(1-\eta) w(\gamma)^{t} L_{\infty}^{\prime}+\eta w(\lambda)^{t} L^{\prime \prime} w(\lambda)=0
$$

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

$$
f_{\phi^{0}, \psi}^{\prime}\left(\phi-\phi^{0}\right)+0.5\left(\phi-\phi^{0}\right)^{t} f_{\phi^{0}, \psi}^{\prime}\left(\phi-\phi^{0}\right)=u(1-\eta) w(\gamma)^{t} L_{\infty}^{\prime}+u \eta w(\lambda)^{t} L^{\prime \prime} w(\lambda)=0
$$

which contradicts assumption [A4]. If $\eta=1$ such a construction still exists and satisfies for all $i=1, \ldots, k_{0}, \sum_{j \in I_{i}} q_{j} \gamma_{j}=\gamma_{i}^{0}$, for all $i=1, \ldots, k_{0}-1 s_{i}=0$, for all $i=t_{k_{0}}+1, \ldots, k, p_{i}=0$ and for all $i=1, \ldots, k_{0}, j \in I_{i}, \sqrt{q_{j}}\left(\gamma_{j}-\gamma_{i}^{0}\right)=u w_{t_{i-1}+j}$ 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 $\gamma_{i}^{0}$ is a singleton, since $\eta_{n} \rightarrow \eta=1$ implies that $\left|\gamma_{n}\right|=o\left(\left.|\lambda|_{n}\right|^{2}\right)$ and

$$
\begin{equation*}
\left|\gamma_{t_{i}, n}-\gamma_{i}^{0}\right|=o\left(\sum_{i} \sum_{j \in I_{i}} q_{j}\left(\gamma_{j, n}-\gamma_{i}^{0}\right)^{2}\right) \tag{10}
\end{equation*}
$$

Therefore if $w_{t_{i}}\left(\tilde{\lambda}_{n}\right) \rightarrow 0$ and we can choose $\gamma_{t_{i}}=\gamma_{i}^{0}$. If $\eta=0$, then (9) leads to $w(\gamma)^{t} L_{\infty}^{\prime}=$ 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, $j=t_{k_{0}}+1, \ldots, k$, imply that for all $i=1, \ldots, k_{0}-1$, $w_{i}(\gamma)^{t} \nabla g_{\gamma_{i}^{0}}+w_{d k_{0}+i}(\gamma) g_{\gamma_{i}}^{0}=0$ and

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

Therefore for all $i=1, \ldots, k_{0}-1$

$$
w_{k_{0}+i}(\gamma)=-\frac{w_{i}(\gamma)^{t} \nabla g_{\gamma_{i}^{0}}}{g_{\gamma_{i}}^{0}}=-w_{i}(\gamma)^{t} \nabla \log g_{\gamma_{i}^{0}}
$$

Since $E_{\gamma_{i}^{0}}\left(\nabla \log g_{\gamma_{i}^{0}}(X)\right)=0$, the above equality implies that for all $i=d k_{0}+1, \ldots,(d+1) k_{0}$, $w_{i}(\gamma)=0$. The regularity asumption [A2] (positivity of the Fisher information matrix) of each model $g_{\gamma}$ 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 $\Gamma$ 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 $\Gamma$ or for which all components $\gamma_{j}$ 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 subsequences 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 $\partial \tilde{\Gamma}$, where $\partial \tilde{\Gamma}=\left\{\gamma \in \partial \Gamma ; \int g_{\gamma}(x) d \mu(x) \in\{0,+\infty\}\right\}$. Let $J=\left\{j \leq k ; \gamma_{j, n} \rightarrow \partial \tilde{\Gamma}\right\} \neq \emptyset$. By definition of $\mathbf{t}, J \subset\left\{t_{k_{0}}+1, \ldots, k\right\}$ and choosing $\sigma$ accordingly we can write $J=\left\{k_{1}, \ldots, k\right\}$ 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 $L_{n}^{\prime}$ into $L_{n,(1)}^{\prime}$ and $L_{n,(2)}^{\prime}$ where $L_{n,(2)}^{\prime}=\left(g_{\gamma_{j, n}}-g_{\gamma_{k_{0}}}^{0}, j=k_{1}, \ldots, k\right)$ and by definition of $k_{1}, L_{n,(1)}^{\prime}$ converges to $L_{\infty,(1)}^{\prime}$ so that (9) becomes in the limit,

$$
\begin{equation*}
\left|(1-\eta) w_{(1)}^{t}(\gamma) L_{(1)}^{\prime}+(1-\eta) w_{(2)}^{t}(\gamma) L_{n,(2)}^{\prime}+\eta w(\lambda)^{t} L^{"} w(\lambda)\right|_{1} \rightarrow 0 \tag{11}
\end{equation*}
$$

as $n$ goes to infinity, where the only term depending on $n$ is $L_{n,(2)}^{\prime}$. 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

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

Thus $h /\left(1-\bar{w}_{2}\right)$ 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 $L_{n,(2)}^{\prime}$ 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 :

$$
\exists i \leq k_{0}, \quad \operatorname{card}\left(I_{i}\right) \geq 2, \quad \exists j_{1}, j_{2} \in I_{i}, \quad q_{j_{1}}>\epsilon / k_{0}, \quad q_{j_{2}}>\left(M_{n}-k\right) \delta_{n}>M_{n} \delta_{n} / 2
$$

if $M_{n}$ is large enough; without loss of generality we set $i=1$ and $j_{1}=1, j_{2}=2$. Then we obtain

$$
\left|s_{i}\right| \leq \delta_{n}, \quad \forall i \leq k_{0}-1 \quad p_{j} \leq \delta_{n}, \quad j=t_{k_{0}}+1, \ldots, k, \quad\left|\sum_{j \in I_{i}} q_{j} \gamma_{j}-\gamma_{i}^{0}\right| \leq \delta_{n}, \quad q_{j}\left|\gamma_{j}-\gamma_{i}^{0}\right|^{2} \leq \delta_{n}
$$

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

$$
\pi\left(\left\{\left|s_{i}\right| \leq \delta_{n}, \forall i \leq k_{0}\right\}\right) \leq C \delta_{n}^{k_{0}-1}, \quad \pi\left(\left\{p_{j} \leq \delta_{n}, j=t_{k_{0}}+1, \ldots, k\right\}\right) \leq \delta_{n}^{\sum_{j=t_{k_{0}}+1}^{k} \alpha_{j}}
$$

Also on $I_{1}$

$$
q_{1}\left(\gamma_{1}-\gamma_{1}^{0}\right)=-\sum_{j \in I_{1}, j>1} q_{j}\left(\gamma_{j}-\gamma_{1}^{0}\right)+0\left(\delta_{n}\right), \quad q_{2}>M_{n} \delta_{n} / 2, \quad\left|\gamma_{j}-\gamma_{1}^{0}\right| \leq \sqrt{\delta_{n} / q_{j}}, \quad j \in I_{1}
$$

the prior probability of the set of $\left(q_{1}, \gamma_{1}, q_{2}, \gamma_{2}, q_{j} \gamma_{j}, j>2, j \in I_{1}\right)$ satisfying the above constraints is bounded by

$$
V_{1} \leq \delta_{n}^{d} \int_{q_{2} \geq M_{n} \delta_{n} / 2}\left(\delta_{n} / q_{2}\right)^{d / 2} q_{2}^{\alpha_{2}-1} d q_{2} \prod_{j>2, j \in I_{1}} \int_{q_{j}, \gamma_{j}} \mathbb{1}_{\left|\gamma_{j}-\gamma_{1}^{0}\right| \leq \sqrt{\delta_{n} / q_{j}}} q_{j}^{\alpha_{j}-1} d q_{j} d \gamma_{j}
$$

Note that

$$
\begin{aligned}
\int_{q_{j}, \gamma_{j}} \mathbb{1}_{\left|\gamma_{j}-\gamma_{1}^{0}\right| \leq \sqrt{\delta_{n} / q_{j}}} q_{j}^{\alpha_{j}-1} d q_{j} d \gamma_{j} & \leq \delta_{n}^{\alpha_{j}}+\delta_{n}^{d / 2} \int_{\delta_{n}}^{1} q^{\alpha_{j}-1-d / 2} d q \\
& \leq \delta_{n}^{\alpha_{j} \wedge d / 2}(\log n)^{1_{\alpha_{j}=d / 2}}
\end{aligned}
$$

and we finally obtain that if $q=\sum_{j=1}^{k_{0}} 1_{\alpha_{j}=d / 2}$ 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

$$
V_{\mathbf{t}} \leq \delta_{n}^{k_{0}-1+\sum_{j=t_{k_{0}}+1}^{k} \alpha_{j}+d k_{0}+\sum_{j=3}^{t_{k_{0}}} \alpha_{j} \wedge d / 2-\sum_{i=1}^{k_{0}-1} \alpha_{t_{i}+1} \wedge d / 2}(\log n)^{q} M_{n}^{\max _{j} \alpha_{j}-d / 2}
$$

If $\max _{j} \alpha_{j}<d / 2$ then

$$
V_{\mathbf{t}} \leq \delta_{n}^{k_{0}-1+d k_{0}+\sum_{j=2}^{k} \alpha_{j}-\sum_{i=1}^{k_{0}-1} \alpha_{t_{i}+1}} M_{n}^{\max _{j} \alpha_{j}-d / 2}
$$

So that with probability going to $1, V_{\mathbf{t}} D_{n} \leq M_{n}^{\max _{j} \alpha_{j}-d / 2}$ i.e. $\pi\left(A_{n}\right) D_{n} \leq M_{n}^{\max _{j} \alpha_{j}-d / 2}$ and

$$
P^{\pi}\left[A_{n}^{\prime} \mid X^{n}\right]=o_{p}(1) \quad \text { if } \quad \max \left\{\alpha_{j}, j=1, \ldots, k\right\}<d / 2
$$

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

$$
\begin{equation*}
P^{\pi}\left[B_{n} \mid X^{n}\right]=o_{p}(1) \quad \text { where } \quad B_{n}:=\left\{\left|f_{0}-f_{\theta}\right| \leq \delta_{n}\right\} \cap\left\{\sum_{i=k_{0}+1}^{k} p_{i} \leq \epsilon_{n}\right\} \tag{12}
\end{equation*}
$$

with $\epsilon_{n}$ 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 $\mathbf{t}=\left(0, k-k_{0}+1, k-k_{0}+2, \ldots, k\right)$ so that $t_{k_{0}}=k$ and define

$$
\tilde{S}_{n}=\left\{\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right) ;\left|\bar{\gamma}_{i}-\gamma_{i}^{0}\right| \leq n^{-1 / 2} ;\left|s_{i}\right| \leq n^{-1 / 2} ; q_{j}\left|\gamma_{j}-\gamma_{i}^{0}\right|^{2} \leq n^{-1 / 2}, \quad \forall j \in I_{i}, i=1, \ldots, k_{0}\right\}
$$

where $\bar{\gamma}_{i}=\sum_{j \in I_{i}} q_{j} \gamma_{j}, \phi_{\mathbf{t}}=\left(\gamma_{j}, j \leq k ; s_{i}, i=2, \ldots, k_{0}\right)$ and $\psi_{\mathbf{t}}=\left(q_{j}, j \in I_{i}, i \leq k_{0}\right)$. Similar computations to those made on the terms $V_{\mathbf{t}}$ lead to (up to fixed multiplicative constants)

$$
\pi\left(\tilde{S}_{n}\right) \leq n^{-\left(k_{0}-1+d k_{0}+d / 2\left(k-k_{0}\right)\right) / 2}, \quad \pi\left(\tilde{S}_{n}\right) \geq n^{-\left(k_{0}-1+d k_{0}+d / 2\left(k-k_{0}\right)\right) / 2}
$$

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

$$
\begin{equation*}
l_{n}\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)-l_{n}\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)=\sqrt{n}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} W_{n}-\frac{n}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} H\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)+R_{n} \tag{13}
\end{equation*}
$$

where $H=-\frac{\partial^{2} l_{n}\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)}{\partial \phi_{\mathbf{t}} \partial \phi_{\mathbf{t}}^{T}}$, and noting
$W_{n}(t)=\mathbb{G}_{n}\left(\frac{p_{i}^{0} q_{j} \nabla_{l} g_{\gamma_{i}^{0}}}{f_{0}}\right), t=l+(j-1) d, j \in I_{i}, \quad W_{n}(k d+t)=\mathbb{G}_{n}\left(\frac{f_{\gamma_{t+1}}^{0}-f_{\gamma_{1}}^{0}}{f_{0}}\right), t=1, \ldots, k_{0}-1$
and

$$
R_{n}=\frac{1}{6} \sum_{r_{1}, r_{2}, r_{3}}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)_{r_{1}}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)_{r_{2}}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)_{r_{3}} \frac{\partial l_{n}^{3}}{\partial \phi_{\mathbf{t}, r_{1}} \phi_{\mathbf{t}, r_{2}} \phi_{\mathbf{t}, r_{3}}}\left(\bar{\phi}_{\mathbf{t}}, \psi_{\mathbf{t}}\right) \quad \bar{\phi}_{\mathbf{t}} \in\left(\phi_{\mathbf{t}}, \phi_{\mathbf{t}}^{0}\right)
$$

We have

$$
\begin{align*}
\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} W_{n} & =\sum_{i=1}^{k_{0}} p_{i}^{0}\left(\bar{\gamma}_{i}-\gamma_{i}^{0}\right)^{T} \mathbb{G}_{n}\left(\frac{\nabla g_{\gamma_{i}^{0}}}{f_{0}}\right)+\sum_{i=2}^{k_{0}} s_{i} \mathbb{G}_{n}\left(\frac{g_{\gamma_{i}^{0}}-g_{\gamma_{1}^{0}}}{f_{0}}\right) \\
& =O_{p}\left[\sum_{i=1}^{k_{0}}\left\|\bar{\gamma}_{i}-\gamma_{i}^{0}\right\|+\sum_{i=2}^{k_{0}}\left|s_{i}\right|\right] \\
& =O_{p}\left(n^{-1 / 2}\right) . \tag{14}
\end{align*}
$$

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

$$
\frac{n}{2}\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)^{T} H\left(\phi_{\mathbf{t}}-\phi_{\mathbf{t}}^{0}\right)=O_{p}\left(n\left(\sum_{i} s_{i}^{2}+\left\|\bar{\gamma}_{i}-\gamma_{i}^{0}\right\|^{2}\right)\right)=O_{p}(1)
$$

We now study $R_{n}$. Each term including at least one $s_{i}$ or one $\left(\gamma_{k-k_{0}+i-1}-\gamma_{i}^{0}\right), i \geq 2$ are of order $O_{p}\left(n^{-1}\right)$, 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 $l_{1}, l_{2}, l_{3} \leq d$,

$$
\begin{aligned}
& \sum_{j, j_{2}, j_{3} \in I_{1}}\left(\gamma_{j_{1} l_{1}}-\gamma_{i_{2} l_{1}}^{0}\right)\left(\gamma_{j_{2} l_{2}}-\gamma_{i_{2} l_{2}}^{0}\right)\left(\gamma_{j_{3} l_{3}}-\gamma_{i_{3} l_{3}}^{0}\right) \frac{\partial^{3} l_{n}}{\partial \gamma_{j_{1} l_{1}} \partial \gamma_{j_{2} l_{2}} \partial \gamma_{j_{3}, l_{3}}} \\
= & O_{p}\left(n\left(\left\|\bar{\gamma}_{1}-\gamma_{1}^{0}\right\| \sum_{j \in I_{1}}\left\|\gamma_{j}-\gamma_{1}^{0}\right\|^{2}+n^{-1 / 2} \sum_{j \in I_{1}}\left\|\gamma_{j}-\gamma_{1}^{0}\right\|^{3}+\sum_{j \in I_{1}}\left\|\gamma_{j}-\gamma_{1}^{0}\right\|^{4}\right)\right)
\end{aligned}
$$

under the assumption that for all $i=1, \ldots, k_{0}$,

$$
F_{0}\left[\sup _{\left|\gamma-\gamma_{i}^{0}\right| \leq \delta} \frac{\left|D^{4} g_{\gamma}\right|}{g_{\gamma}}\right]<+\infty
$$

Finally uniformly over $\tilde{S}_{n}, l_{n}\left(\phi_{\mathbf{t}}, \psi_{\mathbf{t}}\right)-l_{n}\left(\phi_{\mathbf{t}}^{0}, \psi_{\mathbf{t}}\right)=O_{p}(1)$ 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

$$
P_{0}^{n}\left[D_{n}<n^{-\left(d k_{0}+k_{0}-1+d\left(k-k_{0}\right) / 2\right) / 2} C_{\epsilon}\right] \leq \epsilon .
$$

We then bound $\pi\left[B_{n}\right]$. The argument used in the control of $\pi\left(A_{n}\right)$ imply that $\pi\left[B_{n}\right]$ is bounded
by the prior on the set constraint by : for all $\mathbf{t}$

$$
\begin{aligned}
& \left|s_{i}\right| \leq \delta_{n} ; \quad\left\|\sum_{j \in I_{i}} q_{j} \gamma_{j}-\gamma_{i}^{0}\right\| \leq \delta_{n} \quad q_{t_{i}+j} \leq \epsilon_{n}, j=2, \ldots, t_{i+1}-1 \quad \forall i=1, \ldots, k_{0} \\
& q_{j}\left\|\gamma_{j}-\gamma_{i}^{0}\right\|^{2} \leq \delta_{n} \quad \forall j \in I_{i}, i=1, \ldots, k_{0} \quad \text { and } \quad \sum_{j \geq t_{k_{0}}+1} p_{j} \leq \delta_{n} .
\end{aligned}
$$

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

$$
\delta_{n}^{d k_{0}+k_{0}-1+d\left(t_{k_{0}}-k_{0}\right) / 2+\sum_{j \geq t_{k_{0}}+1} \alpha_{\sigma(j)} \sum_{n} \sum_{i} \sum_{j=2}^{t_{i+1}}\left(\alpha_{\sigma(j)}-d / 2\right)}
$$

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

