HAL
open science

Efficient semiparametric estimation and model selection for multidimensional mixtures

Elisabeth Gassiat, Judith Rousseau, Elodie Vernet

To cite this version:

Elisabeth Gassiat, Judith Rousseau, Elodie Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. Electronic Journal of Statistics , 2018, 12 (1), pp.703-740. hal-01345919v2

HAL Id: hal-01345919

https://hal.science/hal-01345919v2

Submitted on 12 Dec 2017

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.

Efficient semiparametric estimation and model selection for multidimensional mixtures

Elisabeth Gassiat[1], Judith Rousseau[2], and Elodie Vernet[3]

[1]*Laboratoire de Mathématiques d'Orsay, Univ. Paris-Sud, CNRS, Université Paris-Saclay, 91405 Orsay, France*
[2]*CEREMADE, Université Paris Dauphine, Paris, Franc*
[3]*CMAP, École Polytechnique, route de Saclay, 91128 Palaiseau Cedex, France*

Abstract

In this paper, we consider nonparametric multidimensional finite mixture models and we are interested in the semiparametric estimation of the population weights. Here, the i.i.d. observations are assumed to have at least three components which are independent given the population. We approximate the semiparametric model by projecting the conditional distributions on step functions associated to some partition. Our first main result is that if we refine the partition slowly enough, the associated sequence of maximum likelihood estimators of the weights is asymptotically efficient, and the posterior distribution of the weights, when using a Bayesian procedure, satisfies a semiparametric Bernstein-von Mises theorem. We then propose a cross-validation like method to select the partition in a finite horizon. Our second main result is that the proposed procedure satisfies an oracle inequality. Numerical experiments on simulated data illustrate our theoretical results.

Keywords: Semiparametric statistics, mixture models, efficiency, Bernstein von Mises Theorem

Contents

1 Introduction **2**

2 Asymptotic efficiency **5**
2.1 Model and notations 5
2.2 Efficient influence functions and efficient Fisher informations 8
2.3 General result 11

3 Model selection **13**
3.1 Behaviour of the MLE as M increases 13
3.2 Criterion for model selection 14

4 Simulation study **17**
4.1 On the estimation of the risk and the selection of M 17
4.2 On the choice of M_0 22

5 Conclusion and discussion **23**
5.1 On the structural assumptions on the model 24
5.2 Extensions to Hidden Markov models 25

6 Proofs **25**
6.1 Proof of Proposition 1 25
6.2 Proof of Proposition 2 26
6.3 Proof of Lemma 1 27
6.4 Proof of Proposition 3 29
6.5 Proof of Corollary 1 33
6.6 Proof of Theorem 2 34
6.7 Proof of Proposition 4 34

1 Introduction

We consider in this paper multidimensional mixture models that describe the probability distribution of a random vector X with at least three coordinates. The model is a probability mixture of k populations such that the coordinates of X can be grouped into 3 blocks of random variables which are conditionally independent given the population. We call emission distributions the conditional distributions of the coordinates and θ the parameter that contains the probability weights of each population. It has been known for some time that such a model is identifiable under weak assumptions. When the coordinates of X take finitely many values, Kruskal [23] in 1977 provided an algebraic sufficient condition under which he proved identifiability. See also [28]. Kruskal's result was recently used by [1] to obtain identifiability under almost no assumption on the possible emission distributions: only the fact that, for each coordinate, the k emission distributions are linearly independent. Spectral methods were proposed by [2], which allowed [10] to derive estimators of the emission densities having the minimax rate of convergence when the smoothness of the emission densities is known. Moreover, [11] proposes an estimation procedure in the case of repeated measurements (where the emission distributions of each coordinate given a population are the same).

Our paper focuses on the semiparametric estimation of the population weights when nothing is known about the emission distributions. This is a semiparametric model, where the finite dimensional parameter of interest is θ and the infinite dimensional nuisance parameters are the emission distributions. In applications, the populations weights have direct interpretation. As an example, in [25] the problem is to estimate the proportion of cells of different types for diagnostic purposes, on the basis of flow cytometry data. Those data give the intensity of several markers responses and may be modelled as multidimensional mixtures.

We are in particular interested in constructing optimal procedures for the estimation of θ. Optimal may be understood as efficient, in Le Cam's theory point of view which is about asymptotic distribution and asymptotic (quadratic) loss. See [24], [8], [33], [34]. The first question is: is the parametric rate attainable in the semiparametric setting? We know here, for instance using spectral estimates, that the parametric rate is indeed attainable. Then, the loss due to the nuisance parameter may be seen in the efficient Fisher information and efficient estimators are asymptotically equivalent to the empirical process on efficient influence functions. The next question is thus: how can we construct asymptotically efficient estimators? In the parametric setting, maximum likelihood estimators (MLEs) do the job, but the semiparametric situation is more difficult, because one has to deal with the unknown nuisance parameter, see the theorems in chapter 24 of [33] where it is necessary to control various bias/approximation terms.

From a Bayesian perspective, the issue is the validity of the Bernstein-Von Mises property of the marginal posterior distribution of the parameter of interet θ. In other words: is the marginal posterior distribution of θ asymptotically Gaussian? Is it asymptotically centered around an efficient estimator? Is the asymptotic variance of the posterior distribution the inverse of the efficient Fisher information matrix? Semiparametric Bernstein-Von Mises theorems have been the subject of recent research, see [31], [12], [29], [15], [14], [9], [17] and [29].

The results of our paper are twofold: first we obtain asymptotically efficient semiparametric estimators using a likelihood strategy, then we propose a data driven method to perform the strategy in a finite horizon with an oracle inequality as theoretical guarantee.

Let us describe our ideas.

For the multidimensional mixture model we consider, we will take advantage of the fact that, we can construct a parametric mixture model based on an approximation of the emission densities by piecewise constant functions - i.e histograms - which acts as a correct model for a coarsened version of the observations (the observations are replaced by the number of points in each grid of the histograms). So that as far as the parameter of interest is concerned, namely the weights of the mixture, this approximate model is in fact well specified, in particular the Kullback-Leibler divergence between the true distribution and the approximate model is minimized at the true value of the parameter of interest, see Section 2.1 for more details. For each of these finite dimensional models, the parameter of interest, i.e. the weights of the mixture, may then be efficiently estimated within the finite dimensional model. Then, under weak assumptions, and using the fact that one can approximate any density on $[0, 1]$ by such histograms based on partitions with radius (i.e. the size of the largest bin) going to zero, it is possible to prove that asymptotically efficient semiparametric estimators may be built using the sequence of MLEs in a growing (with sample size) sequence of approximation models. In the same way, using Bayesian posteriors in the growing sequence of approximation models, one gets a Bernstein-Von Mises result. One of the important implications of the Bernstein-von Mises property is that credible regions, such as highest posterior density

regions or credible ellipsoids are also confidence regions. In the particular case of the semiparametric mixtures, this is of great interest, since the construction of a confidence region is not necessarily trivial. This is our first main result which is stated in Theorem 1: by considering partitions refined slowly enough when the number of observations increases, we can derive efficient estimation procedures for the parameter of interest θ and in the Bayesian approach for a marginal posterior distribution on θ which satisfies the renowned Bernstein-von Mises property.

We still need however in practice to choose a good partition, for a finite sample size. This can be viewed as a model selection problem. There is now a huge literature on model selection, both in the frequentist and in the Bayesian literature. Roughly speaking the methods can be split into two categories: penalized likelihood types of approaches, which include in particular AIC (Akaike's Information Criterion), BIC (Bayesian Information Criterion), MDL (Minimum Description Length) and marginal likelihood (Bayesian) criteria or approaches which consist in estimating the risk of the estimator in each model using for instance bootstrap or cross-validation methods. In all these cases theory and practice are nowadays well grounded, see for instance [22], [30], [6], [26], [7], [5], [16], [3]. Most of the existing results above cover parametric or nonparametric models. Penalized likelihoods in particular target models which are best in terms of Kullback-Leibler divergences typically and therefore aim at estimating the whole nonparametric parameter. Risk estimation via bootstrap or cross-validation methods are more naturally defined in semiparametric (or more generally set-ups with nuisance parameters) models, however the theory remains quite limited in cases where the estimation strategy is strongly non linear as encountered here.

The idea is to estimate the risk of the estimator in each approximation model, and then select the model with the smallest estimated risk. We propose to use a cross-validation method similar to the one proposed in [13]. To get theoretical results on such a strategy, the usual basic tool is to write the cross-validation criterion as a function of the empirical distribution which is not possible in our semiparametric setting. We thus divide the sample in non overlapping blocks of size a_n (n being the the sample size) to define the cross validation criterion. This enables us to prove our second main result: Theorem 2 which states an oracle inequality on the quadratic risk associated with a sample of size a_n observations, and which also leads to a criterion to select a_n. Simulations indicate moreover that the approach behaves well in practice.

In Section 2, we first describe the model, set the notations and our basic assumptions. We recall the semiparametric tools in Section 2.2, where we define the score functions and the efficient Fisher information matrices. Using the fact that spectral estimators are smooth functions of the empirical distribution of the observations, we obtain that, for large enough approximation model, the efficient Fisher information matrix is full rank, see Proposition 1. Intuition says that with better approximation spaces, more is known about all parameters of the distribution, in particular about θ. We prove in Proposition 2 that indeed, when the partition is refined, the Fisher information associated to this partition increases. This leads to our main general result presented in Theorem 1, Section 2.3: it is possible to let the approximation parametric models grow with the sample size

so that the sequence of maximum likelihood estimators are asymptotically efficient in the semiparametric model and so that a semiparametric Bernstein-von Mises Theorem holds. To prove this result we prove in particular, in Lemma 1, that semiparametric score functions and semiparametric efficient Fisher information matrix are the limits of the parametric ones obtained in the approximation parametric models, which has interest in itself, given the non explicit nature of semi-parametric efficient score functions and information matrices in such models. This implies in particular that the semiparametric efficient Fisher information matrix is full rank follows. In Section 3, we propose a model selection approach to select the number of bins. We first discuss in Section 3.1 the reasons to perform model selection and the fact that choosing a too large approximation space does not work, see Proposition 3 and Corollary 1. Then we propose in Section 3.2 our cross-validation criterion, for which we prove an oracle inequality in Theorem 2 and Proposition 4. Results of simulations are described in Section 4, where we investigate several choices of the number and length of blocks for performing cross validation, and investigate practically also V-fold strategies. In Section 5 we present possible extensions, open questions and further work. Finally Section 6 is dedicated to proofs of intermediate propositions and lemmas.

2 Asymptotic efficiency

2.1 Model and notations

Let $(X_n)_{n\geq 1}$ be a sequence of independent and identically distributed random variables taking values in the product of at least three compact subsets of Euclidean spaces which, for the sake of simplicity, we will set as $[0,1]^3$. We assume that the possible marginal distribution of an observation X_n, $n \geq 1$, is a population mixture of k distributions such that, given the population, the coordinates are independent and have some density with respect to the Lebesgue measure on $[0,1]$. The possible densities of X_n, $n \geq 1$, are, if $\mathbf{x} = (x_1, x_2, x_3) \in [0,1]^3$:

$$g_{\theta,\mathbf{f}}(\mathbf{x}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} f_{j,c}(x_c), \quad \sum_{j=1}^{k} \theta_j = 1, \quad \theta_j \geq 0, \forall j \tag{1}$$

Here, k is the number of populations, θ_j is the probability to belong to population j for $j \leq k$ and we set $\theta = (\theta_1, \ldots, \theta_{k-1})$. For each $j = 1, \ldots, k$, $f_{j,c}$, $c = 1, 2, 3$, is the density of the c-th coordinate of the observation, given that the observation comes from population j, and we set $\mathbf{f} = ((f_{j,c})_{1\leq c\leq 3})_{1\leq j\leq k}$. We denote by $\mathbb{P}^\star$ the true (unknown) distribution of the sequence $(X_n)_{n\geq 1}$, such that $\mathbb{P}^\star = P_{\theta^\star,\mathbf{f}^\star}^{\otimes\mathbb{N}}$, $dP_{\theta^\star,\mathbf{f}^\star}(\mathbf{x}) = g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x}$, for some $\theta^\star \in \Theta$ and $\mathbf{f}^\star \in \mathcal{F}^{3k}$, where Θ is the set of possible parameters θ and $\mathcal{F}$ the set of probability densities on $[0,1]$.

We approximate the densities by step functions on some partitions of $[0,1]$. We assume that we have a collection of partitions $\mathcal{I}_M$, $M \in \mathcal{M}$, $\mathcal{M} \subset \mathbb{N}$, so that for each $M \in \mathcal{M}$,

$\mathcal{I}_M = (I_m)_{1\le m\le M}$ is a partition of $[0,1]$ by Borel sets. Then I_m changes when M changes. For each $M \in \mathcal{M}$, we now consider the model of possible densities

$$g_{\theta,\boldsymbol{\omega};M}(\mathbf{x}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \left(\sum_{m=1}^{M} \frac{\omega_{j,c,m}}{|I_m|} 1\!\!1_{I_m}(x_c) \right). \tag{2}$$

Here, $|I|$ is the Lebesgue measure of the set I, $\boldsymbol{\omega} = (\omega_{j,c,m})_{1\le m\le M-1, 1\le c\le 3, 1\le j\le k}$, and for each $j = 1,\dots,k$, each $c = 1,2,3$, each $m = 1,\dots,M-1$, $\omega_{j,c,m} \ge 0$, $\sum_{m=1}^{M-1} \omega_{j,c,m} \le 1$, and we denote $\omega_{j,c,M} = 1 - \sum_{m=1}^{M-1} \omega_{j,c,m}$.
We denote Ω_M the set of possible parameters $\boldsymbol{\omega}$ when using model (2) with the partition $\mathcal{I}_M$.

For any partition $\mathcal{I}_M$, any $\omega = (\omega_m)_{1\le m\le M-1}$ such that $\omega_m \ge 0$, $m = 1,\dots,M$, with $\omega_m = 1 - \sum_{m=1}^{M-1} \omega_m$, denote f_ω the step function given by

$$f_\omega(x) = \sum_{m=1}^{M} \frac{\omega_m}{|I_m|} 1\!\!1_{I_m}(x), \quad \omega_m = \int_0^1 f_\omega(u) 1\!\!1_{I_m}(u) du. \tag{3}$$

When $\boldsymbol{\omega} \in \Omega_M$, let $\mathbf{f}_{\boldsymbol{\omega}} = ((f_{\omega_{j,c}})_{1\le c\le 3})_{1\le j\le k}$.

Possible extensions of our results to model (1) with non compact support, or with more than three coordinates, or with multivariate coordinates, and to model (2) with different sequences of partitions for each coordinate are discussed in Section 5.

An interesting feature of step function approximation is that the Kullback-Leibler divergence $KL((\theta^\star, \mathbf{f}^\star), (\theta, \mathbf{f}_{\boldsymbol{\omega}}))$ between the distribution with density $g_{(\theta^\star,\mathbf{f}^\star)}$ and that with density $g_{(\theta,\boldsymbol{\omega};M)}$, when $(\theta,\boldsymbol{\omega}) \in \Theta \times \Omega_M$, is minimised at

$$\theta = \theta^\star,\ \boldsymbol{\omega} = \boldsymbol{\omega}_M^\star := \left(\int_0^1 f_{j,c}^\star(u) 1\!\!1_{I_m}(u) du \right)_{1\le m\le M-1, 1\le c\le 3, 1\le j\le k}.$$

Indeed

$$\begin{aligned}
KL((\theta^\star, \mathbf{f}^\star), (\theta, \mathbf{f}_{\boldsymbol{\omega}})) &:= \mathbb{E}^\star \left[\log \left(\frac{g_{\theta^\star,\mathbf{f}^\star}(X)}{g_{\theta,\boldsymbol{\omega};M}(X)} \right) \right] \\
&= \mathbb{E}^\star \left[\log \left(\frac{g_{\theta^\star,\mathbf{f}^\star}(X)}{g_{\theta^\star,\boldsymbol{\omega}^\star;M}(X)} \right) \right] + \mathbb{E}^\star \left[\log \left(\frac{g_{\theta^\star,\boldsymbol{\omega}^\star;M}(X)}{g_{\theta,\boldsymbol{\omega};M}(X)} \right) \right] \\
&= KL((\theta^\star, \mathbf{f}^\star), (\theta^\star, \mathbf{f}_{\omega^\star})) + \mathbb{E}^\star \Bigg(\sum_{\substack{1\le m_1\le M \\ 1\le m_2\le M \\ 1\le m_3\le M}} \prod_{c=1}^{3} 1\!\!1_{X_c\in I_{m_c}} \log \left(\frac{g_{\theta^\star,\boldsymbol{\omega}^\star;M}(X)}{g_{\theta,\boldsymbol{\omega};M}(X)} \right) \Bigg) \\
&= KL((\theta^\star, \mathbf{f}^\star), (\theta^\star, \mathbf{f}_{\boldsymbol{\omega}^\star})) + KL((\theta^\star, \mathbf{f}_{\boldsymbol{\omega}^\star}), (\theta, \mathbf{f}_{\omega})).
\end{aligned}$$

This particularity can also be obtained considering $Y_i := ((1\!\!1_{I_m}(X_{i,c}))_{1\le m\le M})_{1\le c\le 3}$ when the X_i's have probability density (1). Indeed, the density of the observations Y_i is exactly the probability density (2) with

$$\theta = \theta^\star,\ \boldsymbol{\omega} = \boldsymbol{\omega}_M^\star$$

This cornerstone property is specific to the chosen approximation, i.e. the step function approximation.
Let $\ell_n(\theta, \boldsymbol{\omega}; M)$ be the log-likelihood using model (2), that is

$$\ell_n(\theta, \boldsymbol{\omega}; M) = \sum_{i=1}^{n} \log g_{\theta,\boldsymbol{\omega};M}(X_i).$$

We denote, for each $M \in \mathcal{M}$, $(\widehat{\theta}_M, \widehat{\boldsymbol{\omega}}_M)$ the maximum likelihood estimator (shortened as MLE), that is a maximizer of $\ell_n(\theta, \boldsymbol{\omega}; M)$ over $\Theta \times \Omega_M$.

Let Π_M denote a prior distribution on the parameter space $\Theta \times \Omega_M$. The posterior distribution $\Pi_M(\cdot|X_1, \ldots, X_n)$ is defined as follows. For any Borel subset A of $\Theta \times \Omega_M$,

$$\Pi_M(A|X_1, \ldots, X_n) = \frac{\int_A \prod_{i=1}^n g_{\theta,\boldsymbol{\omega};M}(X_i) d\Pi_M(\theta, \boldsymbol{\omega})}{\int_{\Theta\times\Omega_M} \prod_{i=1}^n g_{\theta,\boldsymbol{\omega};M}(X_i) d\Pi_M(\theta, \boldsymbol{\omega})}.$$

The first requirement to get consistency of estimators or posterior distributions is the identifiability of the model. We use the following assumption.

Assumption (A1)**.**
• *For all* $j = 1, \ldots, k$, $\theta_j^\star > 0$.
• *For all* $c = 1, 2, 3$, *for all* $j = 1, \ldots, k$, $f_{j,c}^\star$ *is continuous and almost surely positive and*

$$\forall j_1, j_2 \leq k, \quad 0 < \inf_x \frac{f_{j_1,c}^\star(x)}{f_{j_2,c}^\star(x)} \leq \sup_x \frac{f_{j_1,c}^\star(x)}{f_{j_2,c}^\star(x)} < +\infty \tag{4}$$

• *For all* $c = 1, 2, 3$, *the measures* $f_{1,c}^\star dx, \ldots, f_{k,c}^\star dx$ *are linearly independent.*

Note that the two first points in Assumption (A1) imply that for all M, $(\theta^\star, \boldsymbol{\omega}_M^\star)$ lies in the interior of $\Theta \times \Omega_M$.

It is proved in Theorem 8 of [1] that under (A1) identifiability holds up to label switching, that is, if $\mathcal{T}_k$ is the set of permutations of $\{1, \ldots, k\}$,

$$\forall \theta \in \Theta,\ \forall \mathbf{f} \in \mathcal{F}^{3k},\ g_{\theta,\mathbf{f}} = g_{\theta^\star,\mathbf{f}^\star} \Longrightarrow \exists \sigma \in \mathcal{T}_k \text{ such that } {}^\sigma\theta = \theta^\star,\ {}^\sigma\mathbf{f} = \mathbf{f}^\star,$$

where ${}^\sigma\theta \in \Theta$ (resp. ${}^\sigma\mathbf{f} \in \mathcal{F}^{3k}$, ${}^\sigma\theta_j = \theta_{\sigma(j)}$, ${}^\sigma f_{j,c} = f_{\sigma(j),c}$, $c \in \{1, 2, 3\}$, $j \in \{1, , \ldots, k\}$) denotes the image of θ after permuting the labels using σ. This also implies the identifiability of model (2) if the partition is refined enough. We also need the following assumption to ensure that all functions $f_{j,c;M}^\star$ tend to $f_{j,c}^\star$ Lebesgue almost everywhere, where $f_{j,c;M}^\star$ is the function defined in (3) with $\omega = (\omega_{j,c,m}^\star)_m$.

Assumption (A2)**.**
• *For all* M, *the sets* I_m *in* $\mathcal{I}_M$ *are intervals with non empty interior.*
• *As* M *tends to infinity,* $\max_{1\leq m\leq M} |I_m|$ *tends to* 0.

Then, if (A1) and (A2) hold, for M large enough, we have that for all $c = 1, 2, 3$, the measures $f^\star_{1,c;M}dx, \ldots, f^\star_{k,c;M}dx$ are linearly independent. We give a formal proof of this fact in Section 6.1. Thus, using again the identifiability result in [1], under (A1) and (A2), for M large enough,

$$\forall \theta \in \Theta,\ \forall \boldsymbol{\omega} \in \Omega_M,\ g_{\theta,\boldsymbol{\omega};M} = g_{\theta^\star,\boldsymbol{\omega}^\star_M;M} \implies \exists \sigma \in \mathcal{T}_k \text{ such that } {}^\sigma\theta = \theta^\star,\ {}^\sigma\boldsymbol{\omega} = \boldsymbol{\omega}^\star_M,$$

where ${}^\sigma\boldsymbol{\omega} \in \Omega_M$ and ${}^\sigma\omega_{j,c,m} = \omega_{\sigma(j),c,m}$, for all $m \in \{1, \ldots, M\}$, $c \in \{1, 2, 3\}$, $j \in \{1, \ldots, k\}$.

2.2 Efficient influence functions and efficient Fisher informations

We now study the estimation of θ in model (1) and in model (2) from the semiparametric point of view, following Le Cam's theory. We start with model (2) which is easier to analyze since it is a parametric model. For any M, $g_{\theta,\boldsymbol{\omega};M}(\mathbf{x})$ is a polynomial function of the parameter $(\theta, \boldsymbol{\omega})$ and the model is differentiable in quadratic mean in the interior of $\Theta \times \Omega_M$. Denote by $S^\star_M = (S^\star_{\theta,M}, S^\star_{\boldsymbol{\omega},M})$ the score function for parameter $(\theta, \boldsymbol{\omega})$ at point $(\theta^\star, \boldsymbol{\omega}^\star_M)$ in model (2). We have for $j = 1, \ldots, k-1$

$$\left(S^\star_{\theta,M}\right)_j = \frac{\prod_{c=1}^3 f_{\omega^\star_{j,c;M}} - \prod_{c=1}^3 f_{\omega^\star_{k,c;M}}}{g_{\theta^\star,\boldsymbol{\omega}^\star_M;M}}, \tag{5}$$

and for $j = 1, \ldots, k$, $c = 1, 2, 3$, $m = 1, \ldots, M-1$

$$\left(S^\star_{\boldsymbol{\omega},M}\right)_{j,c,m} = \frac{\theta^\star_j \left(\frac{1\!\!1_{I_m}(x_c)}{|I_m|} - \frac{1\!\!1_{I_M}(x_c)}{|I_M|}\right) \prod_{c' \neq c} f_{\omega^\star_{j,c';M}}}{g_{\theta^\star,\boldsymbol{\omega}^\star_M;M}}. \tag{6}$$

Denote by J_M the Fisher information, that is the variance of $S^\star_M(X)$:

$$J_M = \mathbb{E}^\star\left[S^\star_M(X) S^\star_M(X)^T\right]$$

Here, $\mathbb{E}^\star$ denotes expectation under $\mathbb{P}^\star$, and $S^\star_M(X)^T$ is the transpose vector of $S^\star_M(X)$.

When considering the question of efficient estimation of θ in the presence of a nuisance parameter, the relevant mathematical objects are the efficient influence function and the efficient Fisher information. Recall that the efficient score function is the projection of the score function with respect to parameter θ on the orthogonal subspace of the closure of the linear subspace spanned by the tangent set with respect to the nuisance parameter (that is the set of scores in parametric models regarding the nuisance parameter), see [33] or [34] for details. The efficient Fisher information is the variance matrix of the efficient score function. For parametric models, a direct computation gives the result. If we partition the Fisher information J_M according to the parameters θ and $\boldsymbol{\omega}$, that is

$$[J_M]_{\theta,\theta} = \mathbb{E}^\star\left[S^\star_{\theta,M}(X) S^\star_{\theta,M}(X)^T\right],\ [J_M]_{\boldsymbol{\omega},\boldsymbol{\omega}} = \mathbb{E}^\star\left[S^\star_{\boldsymbol{\omega},M}(X) S^\star_{\boldsymbol{\omega},M}(X)^T\right],$$

$$[J_M]_{\theta,\boldsymbol{\omega}} = \mathbb{E}^\star\left[S^\star_{\theta,M}(X)S^\star_{\boldsymbol{\omega},M}(X)^T\right],\ [J_M]_{\boldsymbol{\omega},\theta} = [J_M]^T_{\theta,\boldsymbol{\omega}},$$

we get that, in model (2), if we denote $\tilde{\psi}_M$ the efficient score function for the estimation of θ,

$$\tilde{\psi}_M = S^\star_{\theta,M} - [J_M]_{\theta,\boldsymbol{\omega}}([J_M]_{\boldsymbol{\omega},\boldsymbol{\omega}})^{-1}S^\star_{\boldsymbol{\omega},M},$$

and the efficient Fisher information $\tilde{J}_M$ is the $(k-1)\times(k-1)$-matrix given by

$$\tilde{J}_M = [J_M]_{\theta,\theta} - [J_M]_{\theta,\boldsymbol{\omega}}([J_M]_{\boldsymbol{\omega},\boldsymbol{\omega}})^{-1}[J_M]^T_{\theta,\boldsymbol{\omega}}.$$

To discuss efficiency of estimators, invertibility of the efficient Fisher information is needed. Spectral methods have been proposed recently to get estimators in model (2), see [2]. It is possible to obtain upper bounds of their local maximum quadratic risk with rate $n^{-1/2}$, which as a consequence excludes the possibility that the efficient Fisher information be singular. This is stated in Proposition 1 below and proved in Section 6.1.

Proposition 1. *Assume (A1) and (A2). Then, for large enough M, $\tilde{J}_M$ is non singular.*

In the context of mixture models, all asymptotic results are given up to label switching. We define here formally what we mean by 'up to label switching' for frequentist efficiency results with Equation (8) and Bayesian efficiency results with Equation (10).

By Proposition 1, if (A1) and (A2) hold, for large enough M, $\tilde{J}_M$ is non singular, and an estimator $\widehat{\theta}$ is asymptotically a regular efficient estimator of $\theta^\star$ if and only if

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}_M^{-1}}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}_M\left(X_i\right) + o_{\mathbb{P}^\star}(1), \quad \text{up to label switching,} \tag{7}$$

which formally means that there exists a sequence $(\sigma_n)_n$ belonging to $\mathcal{T}_k$ such that

$$\sqrt{n}\left({}^{\sigma_n}\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}_M^{-1}}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}_M\left(X_i\right) + o_{\mathbb{P}^\star}(1). \tag{8}$$

To get an asymptotically regular efficient estimator, one may for instance use the MLE $\widehat{\theta}_M$ (see the beginning of the proof of Theorem 1). One may also apply a one step improvement (see Section 5.7 in [33]) of a preliminary spectral estimator, such as the one described in [2].

In the Bayesian context, Bernstein-von Mises Theorem holds for large enough M if the prior has a positive density in the neighborhood of $(\theta^\star, \boldsymbol{\omega}^\star_M)$, see Theorem 10.1 in [33]. That is, if $\|\cdot\|_{TV}$ denotes the total variation distance, with $\Pi_{M,\theta}$ the marginal distribution on the parameter θ,

$$\left\|\Pi_{M,\theta}\left(\cdot|X_1,\ldots,X_n\right) - \mathcal{N}\left(\widehat{\theta};\frac{\tilde{J}_M^{-1}}{n}\right)\right\|_{TV} = o_{\mathbb{P}^\star}(1), \quad \text{up to label switching,} \tag{9}$$

where $\widehat{\theta}$ verifies Equation (7), which formally means that

$$\sup_{A\subset\Theta}\left|\Pi_{M,\theta}\big(\exists\sigma\in\mathcal{T}_k:\ {}^{\sigma}\theta\in A\big|X_1,\ldots,X_n\big) - \mathcal{N}\left({}^{\sigma_n}\widehat{\theta};\frac{\tilde{J}_M^{-1}}{n}\right)(A)\right| = o_{\mathbb{P}^\star}(1), \tag{10}$$

where (σ_n) and $\widehat{\theta}$ satisfy Equation (8).
A naive heuristic idea is that, when using the Y_i's as summaries of the X_i's, one has less information, but more and more if the partition $\mathcal{I}_M$ is refined. Thus, the efficient Fisher information should grow when partitions $\mathcal{I}_M$ are refined. The following proposition is proved in Section 6.2.

Proposition 2. *Let $\mathcal{I}_{M_1}$ be a coarser partition than $\mathcal{I}_{M_2}$, that is such that for any $I \in \mathcal{I}_{M_1}$, there exists $A \subset \mathcal{I}_{M_2}$ such that $I = \cup_{I' \in A} I'$. Then*

$$\tilde{J}_{M_2} \geq \tilde{J}_{M_1}$$

in which" $\geq$" denotes the partial order between symmetric matrices.

Thus, it is of interest to let the partitions grow so that one reaches the largest efficient Fisher information.

Let us now come back to model (1). Let, for $j = 1, \ldots, k$, $c = 1, 2, 3$, $\mathcal{H}_{j,c}$ be the subset of functions h in $L^2(f^\star_{j,c}dx)$ such that $\int h f^\star_{j,c} dx = 0$. Then the tangent set for $\mathbf{f}$ at point $(\theta^\star, \mathbf{f}^\star)$ is the subspace $\dot{\mathcal{P}}$ of $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x})$ spanned by the functions

$$\mathbf{x} \mapsto \frac{h(x_c)\prod_{c'=1}^3 f^\star_{j,c'}(x_{c'})}{g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})},\ h \in \mathcal{H}_{j,c},\ j = 1, \ldots, k,\ c = 1, 2, 3.$$

Notice that for each $j = 1, \ldots, k$ and $c = 1, 2, 3$, $\mathcal{H}_{j,c}$ is a closed linear subset of $L^2(f^\star_{j,c}dx)$ and that $\dot{\mathcal{P}}$ is a closed linear subset of $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x})$. The efficient score function $\tilde{\psi}$ for the estimation of θ in the semiparametric model (1) is given, for $j = 1, \ldots, k-1$, by

$$\tilde{\psi}_j = \left(S^\star_\theta\right)_j - \mathbb{A}\left(S^\star_\theta\right)_j,\ \left(S^\star_\theta\right)_j = \frac{\prod_{c=1}^3 f^\star_{j,c} - \prod_{c=1}^3 f^\star_{k,c}}{g_{\theta^\star,\mathbf{f}^\star}}, \tag{11}$$

with $\mathbb{A}$ the orthogonal projection onto $\dot{\mathcal{P}}$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x})$. Then, the efficient Fisher information $\tilde{J}$ is the variance matrix of $\tilde{\psi}$.
If $\tilde{J}$ is non singular, an estimator $\widehat{\theta}$ is asymptotically a regular efficient estimator of $\theta^\star$ if and only if

$$\sqrt{n}\left(\widehat{\theta} - \theta^\star\right) = \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^n \tilde{\psi}\left(X_i\right) + o_{\mathbb{P}^\star}(1), \quad \text{up to label switching} \tag{12}$$

and a Bayesian method using a nonparametric prior Π satisfies a semiparametric Bernstein-von Mises theorem if, with Π_θ the marginal distribution on the parameter θ,

$$\left\|\Pi_\theta\left(\cdot|X_1, \ldots, X_n\right) - \mathcal{N}\left(\widehat{\theta}; \frac{\tilde{J}^{-1}}{n}\right)\right\|_{TV} = o_{\mathbb{P}^\star}(1), \quad \text{up to label switching} \tag{13}$$

for a $\widehat{\theta}$ satisfying (12).

2.3 General result

When the size of the bins in the partition decreases, we expect that the efficient score functions in (2) are good approximations of the efficient score functions in (1) so that asymptotically efficient estimators in model (2) become efficient estimators in model (1). This is what Theorem 1 below states, under the following additional assumption :

Assumption (A3). *For all M large enough, $\mathcal{I}_M$ is a coarser partition than $\mathcal{I}_{M+1}$*

We first obtain:

Lemma 1. *Under Assumptions (A1), (A2) and (A3), the sequence of score functions $(\tilde{\psi}_M)_M$ converges in $L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{x})$ to the score function $\tilde{\psi}$, and the sequence of efficient Fisher informations $(\tilde{J}_M)_M$ converges to the efficient Fisher information matrix $\tilde{J}$, which is non singular.*

The invertibility of $\tilde{J}$ is a consequence of Proposition 1, Proposition 2, and the convergence of $\tilde{J}_M$ to $\tilde{J}$.

We are now ready to state Theorem 1.

Theorem 1. *Under Assumptions (A1), (A2), and (A3), there exists a sequence M_n tending to infinity sufficiently slowly such that the MLE $\widehat{\theta}_{M_n}$ is asymptotically a regular efficient estimator of $\theta^\star$ and satisfies*

$$\sqrt{n}\left(\widehat{\theta}_{M_n}-\theta^\star\right)=\frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}\left(X_i\right)+o_{\mathbb{P}^\star}(1),\quad \textit{up to label switching}. \tag{14}$$

Moreover, under the same assumptions and if for all M, the prior Π_M has a positive and continuous density at $(\theta^\star,\boldsymbol{\omega}_M^\star)$, then there exists a sequence L_n tending to infinity sufficiently slowly such that

$$\left\|\Pi_{L_n,\theta}\left(\cdot|X_1,\ldots,X_n\right)-\mathcal{N}\left(\theta^\star+\frac{\tilde{J}^{-1}}{n}\sum_{i=1}^{n}\tilde{\psi}\left(X_i\right);\frac{\tilde{J}^{-1}}{n}\right)\right\|_{TV}=o_{\mathbb{P}^\star}(1),$$
$$\textit{up to label switching}. \tag{15}$$

Note that, in Theorem 1, any sequence $M_n'\leq M_n$ going to infinity also satisfies (14) and similarly (15) holds for any sequence $L_n'\leq L_n$ going to infinity.

Proof. As explained in Section 2.1, model (2) is the correct model associated with the observations made of the counts per bins and under Assumption (A1) it is a regular model in the neighborhood of the true parameter. Also, using the identifiability of the model and the trick given in [33] p. 63, we get consistency of the MLE. Thus, it is possible to apply Theorem 5.39 in [33] to get that for each M, the MLE $\widehat{\theta}_M$ is regular and asymptotically efficient, that is

$$\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M-\theta^\star\right)=\frac{\tilde{J}_M^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_M\left(X_i\right)+R_n(M),$$

where for each M, $\sigma_{n,M}$ is a sequence of permutations in $\mathcal{T}_k$, and $(R_n(M))_{n\geq 1}$ is a sequence of random vectors converging to 0 in $\mathbb{P}^\star$-probability as n tends to infinity. Therefore, there exists a sequence M_n tending to infinity sufficiently slowly so that, as n tends to infinity, $R_n(M_n)$ tends to 0 in $\mathbb{P}^\star$-probability. Now,

$$\begin{aligned}
\frac{\tilde{J}_{M_n}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}_{M_n}(X_i) &= \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i)+\frac{\tilde{J}_{M_n}^{-1}-\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i)\\
&\quad+\frac{\tilde{J}_{M_n}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{\psi}_{M_n}-\tilde{\psi})(X_i)\\
&= \frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i)+o_{\mathbb{P}^\star}(1)
\end{aligned}$$

since, by Lemma 1, $\mathbb{E}^\star\|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}(\tilde{\psi}_{M_n}-\tilde{\psi})(X_i)\|^2=\|\tilde{\psi}_{M_n}-\tilde{\psi}\|^2_{L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x})}$ tends to 0 as n tends to infinity and $(\tilde{J}_{M_n})^{-1}$ converges to $(\tilde{J})^{-1}$ as n tends to infinity, so that the first part of the theorem is proved.
On the Bayesian side, for all M, there exists a sequence $V_n(M)$ of random vectors converging to 0 in $\mathbb{P}^\star$-probability as n tends to infinity such that

$$\sup_{A\subset\Theta}\left|\Pi_{M,\theta}\big(\exists\sigma\in\mathcal{T}_k:\ {}^{\sigma}\theta\in A\big|X_1,\ldots,X_n\big)-\mathcal{N}\left({}^{\sigma_{n,M}}\widehat{\theta}_M;\frac{\tilde{J}_M^{-1}}{n}\right)(A)\right|=V_n(M).$$

Arguing as previously, there exists a sequence L_n tending to infinity sufficiently slowly so that, as n tends to infinity, both $V_n(L_n)$ and $R_n(L_n)$ tend to 0 in $\mathbb{P}^\star$-probability. Using the fact that the total variation distance is invariant through one-to-one transformations we get

$$\begin{aligned}
&\left\|\mathcal{N}\left({}^{\sigma_{n,M}}\widehat{\theta}_M;\frac{\tilde{J}_M^{-1}}{n}\right)-\mathcal{N}\left(\theta^\star+\frac{\tilde{J}^{-1}}{n}\sum_{i=1}^{n}\tilde{\psi}(X_i);\frac{\tilde{J}^{-1}}{n}\right)\right\|_{TV}\\
&\quad=\left\|\mathcal{N}\left(\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M-\theta^\star\right)-\frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i);\tilde{J}_M^{-1}\right)-\mathcal{N}\left(0;\tilde{J}^{-1}\right)\right\|_{TV}\\
&\quad=\left\|\mathcal{N}\left(\tilde{J}_M^{1/2}[\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M-\theta^\star\right)-\frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i)];Id\right)-\mathcal{N}\left(0;\tilde{J}_M\tilde{J}^{-1}\right)\right\|_{TV}\\
&\quad\leq\left\|\mathcal{N}\left(\tilde{J}_M^{1/2}[\sqrt{n}\left({}^{\sigma_{n,M}}\widehat{\theta}_M-\theta^\star\right)-\frac{\tilde{J}^{-1}}{\sqrt{n}}\sum_{i=1}^{n}\tilde{\psi}(X_i)];Id\right)-\mathcal{N}(0;Id)\right\|_{TV}\\
&\qquad\qquad+\left\|\mathcal{N}(0,Id)-\mathcal{N}\left(0;\tilde{J}_M\tilde{J}^{-1}\right)\right\|_{TV}.
\end{aligned}$$

But for vectors in $m\in\mathbb{R}^{k-1}$ and symmetric positive $(k-1)\times(k-1)$ matrices Σ we have

$$\|\mathcal{N}(m,Id)-\mathcal{N}(0;Id)\|_{TV}\leq\|m\|$$

and

$$\|\mathcal{N}(0, Id) - \mathcal{N}(0; \Sigma)\|_{TV} \leq \mathbb{P}\left(\|\Sigma^{1/2}U\|^2 - \|U\|^2 \geq \log[det(\Sigma)]\right) - \mathbb{P}\left(\|U\|^2 - \|\Sigma^{-1/2}U\|^2 \geq \log[det(\Sigma)]\right)$$

where $U \sim \mathcal{N}(0, Id)$. Thus the last part of the theorem follows from the triangular inequality and the fact that using Lemma 1, as n tends to infinity, $\tilde{J}_{L_n}\tilde{J}^{-1}$ tends to Id, the identity matrix, and $V_n(L_n)$ and $R_n(L_n)$ tend to 0 in $\mathbb{P}^\star$-probability. □

3 Model selection

In Theorem 1, we prove the existence of some increasing partition leading to efficiency. In this section, we propose a practical method to choose a partition when the number of observations n is fixed. In Section 3.1 we prove that one has to take care to choose not too large M_n's since sequences $(M_n)_n$ tending too quickly to infinity lead to inconsistent estimators. In Section 3.2, we propose a cross-validation method to estimate the oracle value $M_n^\star$ minimizing the unknown risk as a function of M: $M \mapsto \mathbb{E}^\star\left[\|\widehat{\theta}_M - \theta^\star\|^2\right]$ (up to label switching).

3.1 Behaviour of the MLE as M increases

We first explain why the choice of the model is important. We have seen in Proposition 2 that for a sequence of increasing partitions, the efficient matrix is non decreasing. The question is then: can we take any sequence tending to infinity wih n? Or, for a fixed n, can we take any M arbitrarily large? As is illustrated in Figure 2, we see that if M is too large (or equivalently if M_n goes to infinity too fast) the MLE (or the Bayesian procedure) is biased.

In Proposition 3, we give the limit of the MLE when the number n of observations is fixed but M tends to infinity.

Proposition 3. *Under Assumptions (A1) and (A2). For almost all observations $X_1, \ldots, X_n$, $\widehat{\theta}_M(X_1, \ldots, X_n)$ tends to*

$$\underline{\theta}_n = (\underbrace{\lfloor n/k \rfloor/n, \ldots, \lfloor n/k \rfloor/n}_{r:=n-k\lfloor n/k \rfloor}, \underbrace{\lceil n/k \rceil/n, \ldots, \lceil n/k \rceil/n}_{k-r})$$

up to label switching, when M tends to infinity.

Proposition 3 is proved in Section 6.4.

Using Proposition 3, we can deduce a constraint on sequences M_n leading to consistent estimation of $\theta^\star$, depending on the considered sequence of partitions $(\mathcal{I}_M)_{M\in\mathcal{M}}$, which may give an upper bound on sequences M_n leading to efficiency. We believe that this constraint is very conservative and leads to very conservative bounds. Corollary 1 below is proved in Section 6.5.

Corollary 1. *Assume that (A1), (A2) and (A3) hold. If $\widehat{\theta}_{M_n}$ tends to $\theta^\star$ in probability and if $\theta^\star$ is different from $(1/k, \dots, 1/k)$, then there exists $N > 0$ and a constant $C > 0$ such that for all $n \geq N$,*

$$n^2 \left(\max_{m \leq M_n} |I_m| \right)^2 M_n \geq C.$$

In particular, if there exists $0 < C_1 \leq C_2$ such that for all $n \in \mathbb{N}$ and $1 \leq m \leq M_n$,

$$\frac{C_1}{M_n} \leq |I_m| \leq \frac{C_2}{M_n} \tag{16}$$

then there exists a constant $C > 0$ such that,

$$M_n \leq C n^2.$$

Note that Assumption (16) holds as soon as the partition is regular, and in particular for the dyadic regular partitions which forms an embedded sequence of partitions where $\mathcal{M} = \{2^p, p \in \mathbb{N}^\star\}$ and for all $M \in \mathcal{M}$ $I_m = [(m-1)/M, m/M)$ for all $m < M$, $I_M = [(M-1)/M, 1]$.

3.2 Criterion for model selection

In this section, we propose a criterion to choose the partition when n is fixed. This criterion can be used to choose the size M of a family of partitions but also to choose between two families of partition. For each dataset, we can compute the MLE or the posterior mean or other Bayesian estimators under model (2) with partition $\mathcal{I}$. We thus shall index all our estimators by $\mathcal{I}$. Note that the results of this section are valid for any family of estimators $(\tilde{\theta}_{\mathcal{I}})$ and not only for the MLE $(\widehat{\theta}_{\mathcal{I}})$. But we illustrate our results using the MLE.

Proposition 3 and Corollary 1 show the necessity to choose an appropriate partition among a collection of partitions $\mathcal{I}_M$, $M \in \mathcal{M}$. To choose the partition we need a criterion. Since the aim is to get efficient estimators, we choose the quadratic risk as the criterion to minimize. We thus want to minimize over all possible partitions

$$R_n(\mathcal{I}) = \mathbb{E}^\star \left[\|\tilde{\theta}_{\mathcal{I}}(X_{1:n}) - \theta^\star\|^2_{\mathcal{T}_k} \right], \tag{17}$$

where $X_{1:n} = (X_i)_{i \leq n}$ and for all $\theta, \tilde{\theta} \in \Theta$,

$$\|\theta - \tilde{\theta}\|_{\mathcal{T}_k} = \min_{\sigma \in \mathcal{T}_k} \|^\sigma \theta - \tilde{\theta}\|_2. \tag{18}$$

As usual, this criterion cannot be computed in practice (since we do not know $\theta^\star$) and we need for each partition $\mathcal{I}$ some estimator $C(\mathcal{I})$ of $R_n(\mathcal{I})$.

We want to emphasize here that the choice of the criterion for this problem is not easy. Indeed, the quadratic risk $R_n(\mathcal{I})$ cannot be written as the expectation of an excess loss expressed thanks to a contrast function, i.e. in the form $\mathbb{E}^\star \left[\mathbb{E}^\star \left[\gamma(\tilde{\theta}(X_{1:n}), X) - \gamma(\theta^\star, X) | X_{1:n} \right] \right]$,

where $\gamma : \Theta \times \mathcal{X} \to [0, +\infty)$. Yet, the latter is the framework of most theoretical results in model selection, see [5] or [26] for instance. Moreover decomposing the quadratic risk as an approximation error plus an estimation error as explained in [5]:

$$R_n(\mathcal{I}) = \underbrace{\inf_{\theta \in \Theta_{\mathcal{I}}} \|\theta - \theta^\star\|^2_{\mathcal{T}_k}}_{\text{approximation error}} + \underbrace{R_n(\mathcal{I}) - \inf_{\theta \in \Theta_{\mathcal{I}}} \|\theta - \theta^\star\|^2_{\mathcal{T}_k}}_{\text{estimation error}}, \quad \text{where } \Theta_{\mathcal{I}} = \Theta,$$

we see that the approximation error is always zero in our model (and not decreasing as often when the complexity of the models increases). Hence

$$R_n(\mathcal{I}_M) = \underbrace{Var^\star\left[\tilde{\theta}_{\mathcal{I}_M}(X_{1:n})\right]}_{\text{variance}} + \underbrace{\left\|\mathbb{E}^\star\left[\tilde{\theta}_{\mathcal{I}_M}(X_{1:n})\right] - \theta^\star\right\|^2_{\mathcal{T}_k}}_{\text{bias}} \tag{19}$$

where $Var^\star(.)$ is to be understood as the trace of the variance matrix. Here the bias is only an estimation bias and not a model mispecification bias.

In the case of the MLE, using Theorem 1, for all fixed M (large enough), the regularity of the mixture of these multivariate distributions implies that the bias is $O(1/n)$ and the variance converges to the inverse Fisher information matrix so that

$$R_n(\mathcal{I}_M) = Var^\star\left[\widehat{\theta}_{\mathcal{I}_M}(X_{1:n})\right] + O(1/n) = tr\left(\tilde{J}_M^{-1}\right) + O(1/n)$$

and if $M \geq M_\epsilon$ so that $\|\tilde{J}_{M_\epsilon}^{-1} - \tilde{J}^{-1}\| \leq \epsilon$ we obtain that for n large enough

$$\text{tr}\left(\tilde{J}^{-1}\right) - 2\epsilon \leq R_n(\mathcal{I}_M) \leq \text{tr}\left(\tilde{J}^{-1}\right) + 2\epsilon.$$

Minimizing $R_n(\mathcal{I}_M)$ therefore corresponds to choosing M such that $\tilde{J}_M$ is close enough to $\tilde{J}$ (i.e. M large enough) while not deteriorating too much the approximation of $R_n(\mathcal{I}_M)$ by $\text{tr}(\tilde{J}_M^{-1})$ (i.e. M not too large).

Because the approximation error is always zero we cannot apply the usual methods and we use instead a variant of the cross-validation technique.

Consider a partition of $\{1, \cdots, n\}$ in the form $(B_b, B_{-b}, b \leq b_n)$, in other words the partition is made of $2 \times b_n$ subsets of $\{1, \cdots, n\}$. By definition $B_{b_1} \cap B_{-b_2} = \emptyset$ for all $b_1, b_2 \leq b_n$. Because an arbitrary estimator, e.g. the MLE, based on any finite sample size is not unbiased, the following naive estimator of the risk is not appropriate:

$$C_{CV1}(\mathcal{I}) = \frac{1}{2b_n} \sum_{b=1}^{b_n} \|\tilde{\theta}_{\mathcal{I}}(X_{B_b}) - \tilde{\theta}_{\mathcal{I}}(X_{B_{-b}})\|^2_{\mathcal{T}_k}.$$

This can be seen by decomposing the risk $R_n(\mathcal{I})$ as in Equation (19) and by computing the expectation of $C_{CV1}(\mathcal{I})$ in the case where the sizes of B_b, B_{-b}, $b \leq b_n$, are all equal,

$$\mathbb{E}^\star\left[C_{CV1}(\mathcal{I})\right] = Var^\star\left[\tilde{\theta}_{\mathcal{I}}(X_{B_b})\right].$$

Then, the criterion $C_{CV1}(\mathcal{I})$ do not capture the bias of the estimator $\tilde{\theta}_{\mathcal{I}}$.

In the case of the MLE, using Proposition 3, $C_{CV1}(\mathcal{I})$ is tending to 0 when $\max_m |I_m|$ tends to 0. So that minimizing this criterion leads to choosing a partition $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}} C_{CV1}(\mathcal{I})$ which has a large number of sets and so $\widehat{\theta}_{\widehat{\mathcal{I}}_n}(X_{1:n})$ may be close to $(1/k, \ldots, 1/k)$ and then may not even be consistent. As an illustration, see Figure 2 where $R_n(\mathcal{I})$, $Var^\star\left[\widehat{\theta}_{\mathcal{I}}(X_{1:n})\right]$ and $\left\|\mathbb{E}^\star\left[\widehat{\theta}_{\mathcal{I}}(X_{1:n})\right] - \theta^\star\right\|^2_{\mathcal{T}_k}$ are plotted as a function of M, for three simulation sets and various values of the sample size n, see Section 4 for more details. It is quite clear from these plots that the variances remain either almost constant with M or tend to decrease, while the bias increases with M and becomes dominant as M becomes larger. As a result $R_n(\mathcal{I})$ tends to first decrease and then increase as M increases.

To address the bad behaviour of $C_{CV1}(\mathcal{I})$, we use an idea of [13]. Choose a fixed base partition $\mathcal{I}_0$ with a small number of bins (although large enough to allow for identifiability). Then compute

$$C_{CV}(\mathcal{I}) = \frac{1}{b_n}\sum_{b=1}^{b_n} \|\tilde{\theta}_{\mathcal{I}}(X_{B_b}) - \tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})\|^2_{\mathcal{T}_k}.$$

Ideally we would like to use a perfectly unbiased estimator $\tilde{\theta}$ in the place of $\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})$, see Assumption (A5.2) used in Theorem 2 and Proposition 4. We discuss the choice of $\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})$ at the end of the section.

Figure 3 gives an idea of the behaviour of $C_{CV}(\cdot)$ and $C_{CV1}(\cdot)$ using the MLE. It shows in particular that in our simulation study $C_{CV}(\cdot)$ follows the same behaviour as $R_n(\cdot)$, contrarywise to $C_{CV1}(\cdot)$. More details are given in Section 4.

We now provide some theoretical results on the behaviour of the minimizer of $C_{CV}(\cdot)$ over a finite family of candidate partitions $\mathcal{M}_n$ compared to the minimizer of $R_{a_n}(\cdot)$ over the same family. Let $m_n = \#\mathcal{M}_n$ be the number of candidate partitions. We consider the following set of assumptions:

Assumption (A5)**.**

(A5.1) B_b, B_{-b}, $b \leq b_n$ are disjoint sets of equal size

$$\#B_b = \#B_{-b} = a_n, \text{ for all } b \leq b_n$$

(A5.2) $\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})$ is not biased i.e. $\mathbb{E}^\star[\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})] = \theta^\star$.

We obtain the following oracle inequality.

Theorem 2. *Suppose Assumption (A5). For any sequences $0 < \epsilon_n, \delta_n < 1$, with probability greater than*

$$1 - 2m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right),$$

we have

$$R_{a_n}(\widehat{\mathcal{I}}_n) \leq \frac{1+\epsilon_n}{1-\epsilon_n}\inf_{\mathcal{I}\in\mathcal{M}_n} R_{a_n}(\mathcal{I}) + \frac{2\delta_n}{1-\epsilon_n}, \tag{20}$$

where $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}\in\mathcal{M}_n} C_{CV}(\mathcal{I})$.

As a consequence of Theorem 2, the following Proposition holds. Recall that $n = 2b_n a_n$.

Proposition 4. *Assume (A5). If* $b_n \gtrsim n^{2/3}\log^2(n)$, $a_n \lesssim n^{1/3}/(\log^2(n))$, *and* $m_n \leq C_\alpha n^\alpha$, *for some* $C_\alpha > 0$ *and* $\alpha \geq 0$, *then*

$$\mathbb{E}^\star\left[a_n R_{a_n}(\widehat{\mathcal{I}}_n)\right] \leq \inf_{\mathcal{I}\in\mathcal{M}_n} a_n R_{a_n}(\mathcal{I}) + o(1),$$

where $\widehat{\mathcal{I}}_n \in \arg\min_{\mathcal{I}\in\mathcal{M}_n} C_{CV}(\mathcal{I})$.

Note that for each $\mathcal{I}$, $R_{a_n}(\mathcal{I})$ is of order of magnitude $1/a_n$ so that the main term in the upper bound of Proposition 4 is $\inf_{\mathcal{I}\in\mathcal{M}_n} a_n R_{a_n}(\mathcal{I})$. Note also that this is an exact oracle inequality (with constant 1).

In Theorem 2 and Proposition 4, $\widehat{\mathcal{I}}_n$ is built on n observations while the risk is associated with $a_n < n$ observations. This leads to a conservative choice of $\widehat{\mathcal{I}}_n$, i.e. we may choose a sequence $\widehat{\mathcal{I}}_n$ (optimal with a_n observations) increasing more slowly than the optimal one (with n observation). We think however that this conservative choice should not change the good behaviour of $\widehat{\theta}_{\widehat{\mathcal{I}}_n}$, since Theorem 1 implies that any sequence of partitions which grows slowly enough to infinity leads to an efficient estimator. Hence, once the sequence M_n growing to infinity is chosen, then any other sequence growing to infinity more slowly also leads to an efficient estimator.

In Proposition 4 and Theorem 2, the reference point estimate $\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})$ is assumed to be unbiased. This is a strong assumption, which is not exactly satisfied in our simulation study. To consider a reasonable approximation of it, $\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})$ is chosen as the MLE associated with a partition with a small number of bins. Recall that the maximum likelihood estimator is asymptotically unbiased and for a fixed M, the bias of the MLE for the whole parameter θ, ω is of order $1/n$. The heuristic is that a small number of bins implies a smaller number of parameters to estimate, so that the asymptotic regime is attained faster. Our simulations confirm this heuristic, see Section 4.

To take a small number of bins but large enough to get identifiability, we observe in Section 4.2 a great heterogeneity among different estimators and also that some estimators have null components or cannot be computed, when the number of bins is too small.

4 Simulation study

4.1 On the estimation of the risk and the selection of M

In this section, we illustrate the results obtained in Sections 3.1 and 3.2 with simulations. We compare six criteria for the model selection based on C_{CV} with different choices of size of training and testing sets. We choose the regular embedded dyadic partitions, i.e. when $\mathcal{M} = \{2^p, p \in \mathbb{N}^*\}$ and for all $M \in \mathcal{M}$, $I_m = [(m-1)/M, m/M)$ for all

$m < M$, $I_M = [(M-1)/M, 1]$. Following Corollary 1, when n is fixed, we only consider $M = 2^P \leq M_n = n^3$ (i.e. $P \leq P_n := \lfloor 3/2 \log(n) \rfloor$). In this part, we only consider MLE estimators with ordered components and approximated thanks to the EM algorithm.

For n fixed, the choice of the model, through P, is done using the criterion C_{CV} based on two types of choice for $(B_b), (B_{-b})$. First, we use the framework under which we were able to prove something, i.e. Assumption (A5.1) where all the training and testing sets are disjoints. In this context we use different sizes a_n and b_n:

- $b_n = \lceil n^{2/3} log(n)/(20) \rceil$ and $a_n = \lfloor n/(2b_n) \rfloor$ (Assumption of Proposition 4, up to $\log(n)$), leading to the criterion $C_{CV}^{D,1}$ and the choice of P noted $\widehat{P}_n^{D,1} \in \arg\min_{P \leq P_n} C_{CV}^{D,1}(\mathcal{I}_{2^P})$,
- $b_n = \lceil n^{1/3} \rceil$, $a_n = \lfloor n/(2b_n) \rfloor$, leading to the criterion $C_{CV}^{D,2}$ and the choice of P noted $\widehat{P}_n^{D,2} \in \arg\min_{P \leq P_n} C_{CV}^{D,2}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n/10 \rfloor$, $b_n = \lfloor n/(2a_n) \rfloor$, leading to the criterion $C_{CV}^{D,3}$ and the choice of P noted $\widehat{P}_n^{D,3} \in \arg\min_{P \leq P_n} C_{CV}^{D,3}(\mathcal{I}_{2^P})$

We also consider the famous V-fold, where the dataset is cut into b_n disjoint sets $\tilde{B}_b$ of size a_n, leading to training sets $B_b = \tilde{B}_b$ and testing sets $B_{-b} = \{1, \ldots n\} \setminus \tilde{B}_b$. We also use different sizes a_n and b_n:

- $a_n = \lfloor n^{1/3} \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,1}$ and the choice of P noted $\widehat{P}_n^{V,1} \in \arg\min_{P \leq P_n} C_{CV}^{V,1}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n^{2/3}/2 \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,2}$ and the choice of P noted $\widehat{P}_n^{V,2} \in \arg\min_{P \leq P_n} C_{CV}^{V,2}(\mathcal{I}_{2^P})$,
- $a_n = \lfloor n/10 \rfloor$, $b_n = \lfloor n/a_n \rfloor$, leading to the criterion $C_{CV}^{V,3}$ and the choice of P noted $\widehat{P}_n^{V,3} \in \arg\min_{P \leq P_n} C_{CV}^{V,3}(\mathcal{I}_{2^P})$.

Note that for criteria

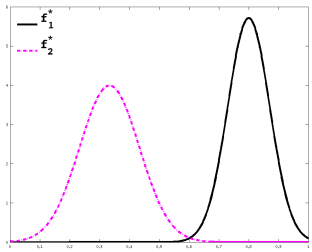
- $C_{CV}^{j,1}$, $j \in \{D, V\}$, a_n is proportional to $n^{1/3}$ up to a logarithm term,
- $C_{CV}^{j,2}$, $j \in \{D, V\}$, a_n is proportional to $n^{2/3}$,
- $C_{CV}^{j,3}$, $j \in \{D, V\}$, a_n is proportional to n.

We now explain how we choose $\mathcal{I}_0$. As explained earlier M_0 has to be taken small, but not too small since otherwise the model would not be identifiable. We propose to choose the smallest $M_0 = 2^{P_0}$ such that $M_0 \geq k+2$ (equivalently $P_0 \geq \log(k+2)/\log(2)$). This lower bound ensures that generically on $\mathcal{I}_0$ the model (2) is identifiable.

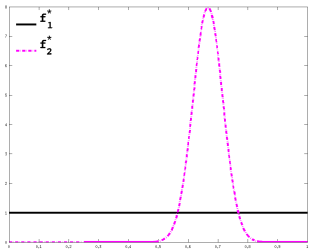
We consider three different simulation settings. In each one of them we consider the conditionally repeated sampling model, i.e. $f_{j,1} = f_{j,2} = f_{j,3}$, both for the true distribution and for the model. In the three cases, $k = 2$ and the other parameters are given in Table 1. So that, we work with $P_0 = 2$ and $M_0 = 2^2 = 4$.
The different emission distributions are represented in Figure 1.

Simu.	k	$\theta^\star$	$f^\star_{1,1}dx = f^\star_{1,2}dx = f^\star_{1,3}dx$	$f^\star_{2,1}dx = f^\star_{2,2}dx = f^\star_{2,3}dx$
1	2	$(0.3, 0.7)$	$\mathcal{N}(4/5, 0.07^2)$ truncated to $[0,1]$	$\mathcal{N}(1/3, 0.1^2)$ truncated to $[0,1]$
2	2	$(0.2, 0.8)$	$\mathcal{U}((0,1))$	$\mathcal{N}(2/3, 0.05^2)$ truncated to $[0,1]$
3	2	$(0.3, 0.7)$	$\beta(1,2)$	$\beta(5,3)$

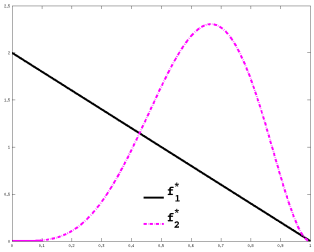
Table 1: Values of the true parameters for simulation 1 to 3



(a) Simulation 1



(b) Simulation 2

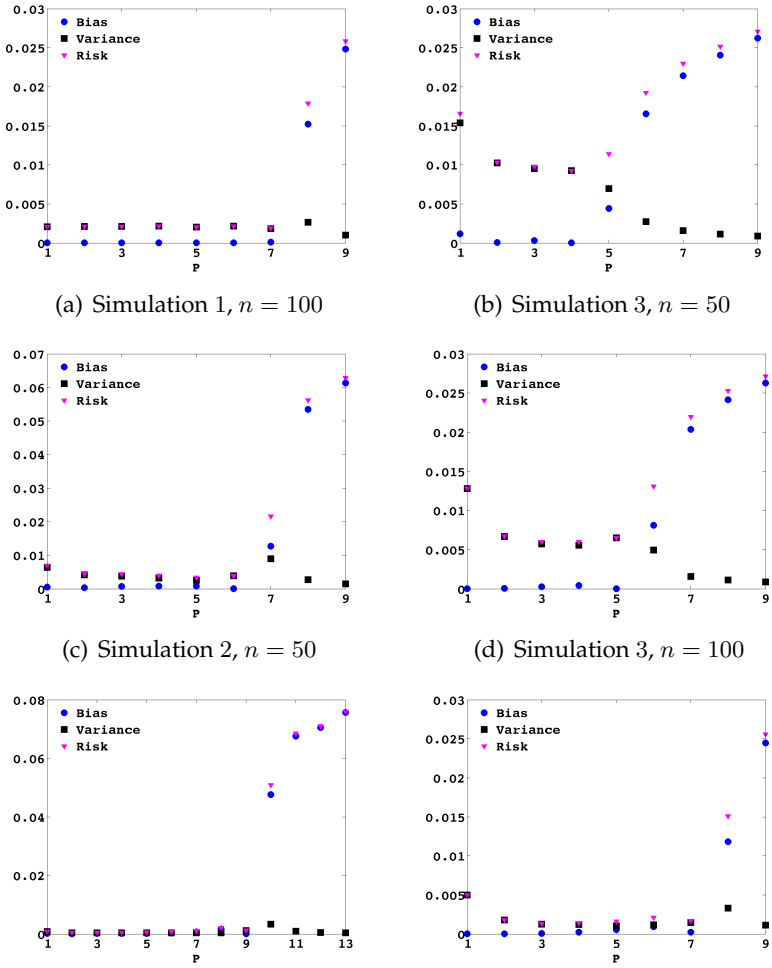


(c) Simulation 3

Figure 1: Representation of the true emission distributions for simulations 1, 2 and 3.

In Figure 2 we display the evolution of the risk $R_n(\mathcal{I}_{2^P})$, the variance $Var^\star\left[\widehat{\theta}_{2^P}(X_{1:n})\right]$ and the squared bias $\left\|\mathbb{E}^\star\left[\widehat{\theta}_{2^P}(X_{1:n})\right] - \theta^\star\right\|^2_{\mathcal{T}_k}$ defined in Equation (19) as the number of bins 2^P increases, for different values of n and for each of the three true distributions. The risks, bias and variances are estimated by Monte Carlo, based on 1000 repeated samples and for each of them we compute the MLE using the EM algorithm. We notice that typically the bias first is either constant or slightly decreasing as P increases and then increases rapidly for larger values of P until it stabilizes to the value $\|\underline{\theta}_n - \theta^\star\|_{\mathcal{T}_k}$, which is what was proved in Proposition 3. On the other hand the variance is monotone non increasing as P increases until P becomes quite large and then it decreases to zero (which also is a consequence of Proposition 3) when P gets large. As a result the risk, which is the sum of the squared bias and the variance, is typically constant or decreasing for small increasing values of P and then increasing to $\|\underline{\theta}_n - \theta^\star\|^2_{\mathcal{T}_k}$ when P gets large.

In real situations $R_n(\mathcal{I}_{2^P})$ is unknown, we now illustrate the behaviour of the different



(a) Simulation 1, $n = 100$ (b) Simulation 3, $n = 50$

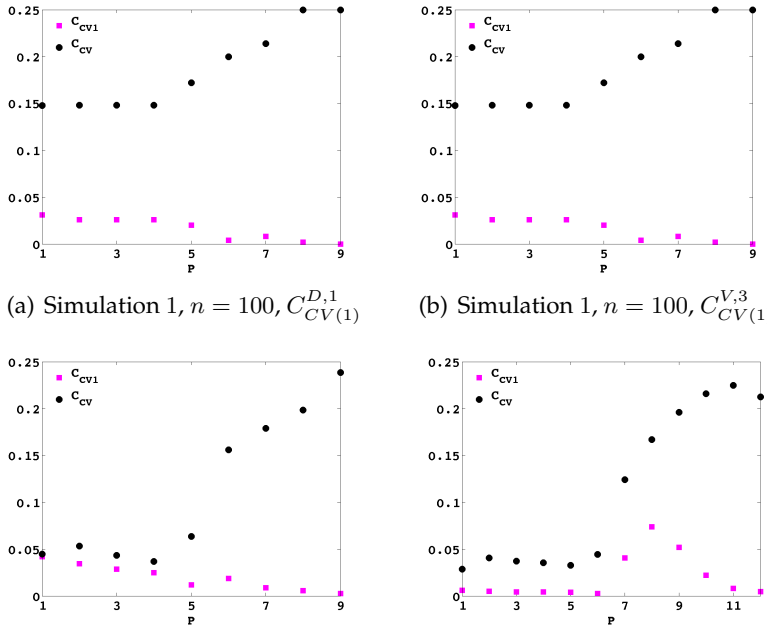
(c) Simulation 2, $n = 50$ (d) Simulation 3, $n = 100$

(e) Simulation 2, $n = 500$ (f) Simulation 3, $n = 500$

Figure 2: Patterns of the risk (with black squares), the squared bias (with blue dots) and variance (with magenta triangles) with respect to $P = \log(M)/\log(2)$ for simulations 1, 2 and 3 and different values of n.

criteria C_{CV} and C_{CV1} and see how close to $R_n(\mathcal{I}_{2^P})$ they are. For the sake of conciseness we only display results for simulated data 1 and 2 and for $n = 100, 500$ since they are very typical of all other simulation studies we have conducted. The results are presented in Figure 3, where the criteria C_{CV}, C_{CV1} are computed based on a single data $X_{1:n}$. We see in figure 3 that contrarywise to C_{CV}, the basic cross-validated criterion C_{CV1} does not recover the behaviour of $R_n(\mathcal{I}_{2^P})$ correctly as it fails to estimate the bias. Note that we do not compare the values but the behaviour. Indeed, the criteria are used to choose the best P by taking the minimum of the criterion so that the values are not important by themselves. Besides, we know that the criterion C_{CV} is biased by a constant depending on $\mathcal{I}_0$. As theoretically explained in Section 3 and as a consequence of Proposition 3,

we can see that the criteria C_{CV1} are tending to 0 when P increases while it is not the case for the criteria C_{CV}. It is interesting to note that from Figure 3, the minimizer in P of C_{CV} corresponds to values of the risk that are close to the minimum, we precise this impression with table 2.



(a) Simulation 1, $n = 100$, $C^{D,1}_{CV(1)}$ (b) Simulation 1, $n = 100$, $C^{V,3}_{CV(1)}$

(c) Simulation 2, $n = 500$, $C^{D,1}_{CV(1)}$ (d) Simulation 2, $n = 500$, $C^{V,3}_{CV(1)}$

Figure 3: behaviour of C_{CV1} vs C_{CV} as a function of P

Finally we compare the six criteria $C^{j,c}_{CV}$, $j \in \{D, V\}$, $c \in \{1, 2, 3\}$, by estimating the squared risk of the associated estimator $\widehat{\theta}_{2^{\hat{P}^{j,c}_n}}$, presented in Table 2 across different sample sizes n and the three simulation set-ups (simulated data 1, 2 and 3) described above. We can compare the six squared risk to $\sqrt{\min_{P \leq P_n} R_n(2^P)}$ and $\sqrt{R_n(2^{P_0})}$. The different risks are estimated by Monte Carlo by repeating 100 times the estimation. The differences of performance between the different criteria are not obvious. Besides, the performances of all the criteria are satisfactory, compared to $\sqrt{\min_{P \leq P_n} R_n(2^P)}$. Yet, we suggest not to use criterion $C^{V,1}_{CV}$ because it is computationally more intensive than the others, particularly when n is large (because of large b_n). In our simulation study $C^{D,1}_{CV}$ and $C^{V,2}_{CV}$ seem to behave slightly better than the others.

These results confirm that by using M_0 small, the criterion behaves correctly. Moreover, the fact that the choice of $\widehat{I}_n$ corresponds to a risk associated with $a_n < n$ observations does not seem to be a conservative choice even in a finite horizon (i.e. when n is fixed). We were expecting this behaviour asymptotically but not necessarily in a finite horizon.

Simulation	1	1	1	1	1	2	2	2	3	3	3
n	50	100	500	1000	2000	50	100	500	50	100	500
$\sqrt{\min_{P\leq P_n} R_n(2^P)}$	0.062	0.043	0.020	0.014	0.010	0.058	0.046	0.020	0.096	0.078	0.036
$\sqrt{R_n(2^{\hat{P}_0})}$	0.063	0.046	0.021	0.015	0.010	0.067	0.046	0.022	0.10	0.082	0.042
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{D,1}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.069	0.047	0.019	0.014	0.011	0.075	0.056	0.019	0.12	0.087	0.037
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{D,2}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.073	0.046	0.022	0.015	0.010	0.065	0.056	0.025	0.10	0.087	0.046
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{D,3}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.086	0.047	0.021	0.014	0.010	0.087	0.056	0.026	0.11	0.087	0.041
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{V,1}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.091	0.046	0.021	0.013	0.009	0.104	0.055	0.022	0.11	0.087	0.053
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{V,2}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.069	0.046	0.019	0.013	0.010	0.070	0.049	0.022	0.12	0.084	0.036
$\sqrt{\mathbb{E}^\star\left[\lVert\widehat{\theta}_{2^{\hat{P}_n^{V,3}}}(X_{1:n})-\theta^\star\rVert^2\right]}$	0.103	0.046	0.019	0.014	0.009	0.10	0.049	0.022	0.14	0.083	0.035

Table 2: Comparison of the squared risk of estimators associated to different criteria

4.2 On the choice of M_0

To compute the different criteria $C_{CV}^{j,i}$, $j \in \{D, V\}$, $i \in \{1, 2, 3\}$, we proposed to choose M_0 as small as possible but for which the model is identifiable up to label switching. Given $\min_{1\leq j\leq k} \theta_j^\star > 0$, the model associated to the parameter space is identifiable as soon as the k vectors $(\omega_{1,c,\cdot}^\star)_{M_0}, \ldots, (\omega_{k,c,\cdot}^\star)_{M_0}$ in Δ_{M_0} are linearly independent for all $c \in \{1, 2, 3\}$. Considering the dimension of linear spaces, we may choose $M_0 = k+1$ or $k+2$. Then we should generically avoid issues with identifiability. We chose such a M_0 in the previous simulations. We now study the impact of a choice of M_0 that would be too small .

To do so we have simulated data from $f_{2,1}^\star(y) = f_{2,2}^\star(y) = f_{2,3}^\star(y) = 1$ and $f_{1,1}^\star(y) = f_{1,2}^\star(y) = f_{1,3}^\star(y) = 1 + \cos(2\pi(y + \epsilon))$, for $\epsilon = 0.25$, 0.3, 0.4 and 0.5 and $n = 500$, 1000 and 2000. In this case the smallest possible value for M_0 based on the regular grid on $[0, 1]$ is $M_0 = 2$ for $\epsilon \neq 0.5$ whereas if $\epsilon = 0.5$ the model is non identifiable at $M_0 = 2$ but becomes identifiable when $M_0 = 4$. For each of these simulation data we have computed various estimators: MLEs based on the EM algorithm initiated at different values, the posterior mean and the MAP estimator computed from a Gibbs sample algorithm with a Dirichlet prior distribution on θ and independently a Dirichlet prior distribution on each $\omega_{j,1}$, for $1 \leq j \leq k$. We noticed that the EM algorithm with different initializations were very heterogeneous. Moreover, the MAP estimator, posterior mean and spectral estimators often had one of the $\hat{\theta}_j$ null or close to 0. Sometimes, the spectral estimator could not be computed.

The explanation for such behaviour is that when the model is not identifiable, one $\theta_j^\star$ may be null or the vectors $(\omega_{1,c,\cdot}^\star)_{M_0}, \ldots, (\omega_{k,c,\cdot}^\star)_{M_0}$ may be linearly dependent for some $c \in \{1, 2, 3\}$. In this case, the likelihood will have multiple modes (apart from those arising because of label switching). Hence a way to check that M_0 is not too small is to compute multiple initialisation of the EM algorithm if the MLE is estimated or to look for very small values of $\hat{\theta}_j$ in the case of Bayesian estimators (possibly also running multiple MCMC chains with different initial values). In practice we suggest that this analysis be conducted for a few number of values M_0 and then to select the value that leads to the

most stable results.

To illustrate this, we present a simulation study where the number of estimators is $S = 14$. The 10 first estimators were obtained using the EM algorithm with different random initializations, we also considered the spectral estimator proposed in [2] and an estimator obtained with the EM algorithm with the spectral estimator as initialization. The last two estimators were the MAP estimator and the posterior mean. We considered regular partitions with $M_0 = 2$, 4 and 6 bins. To present the results in a concise way we have summarized them in Figure 4, using the indicator

$$ID(M_0) = \frac{S\,\mathbb{1}_{\forall s \leq S,\ j \leq 2,\ \hat{\theta}^s_{M_0,j} > 1/\sqrt{n}}}{\sum_{s=1}^{S} \left(\min_{j=1,2} \hat{\theta}^s_{M_0,j} - \frac{1}{S} \sum_{t=1}^{S} \min_{j=1,2} \hat{\theta}^t_{M_0,j} \right)^2}$$

with $k = 2$ and $\theta^\star = (0.7, 0.3)$. Thus when $ID(M_0) = 0$, there is a suspicion that the associated model is not identifiable and another partition should be chosen. It appears that for all n, when $\epsilon = 0.5$ $M_0 = 2$ always appears as having a pathological behaviour and for the other values of ϵ this value is accepted for large values of n.

	$\epsilon = 0.25$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
$M_0 = 2$	0	0	0	0
$M_0 = 4$	**24**	0	**58**	0
$M_0 = 6$	24	**10**	0	**28**

(a) $n = 500$

	$\epsilon = 0.25$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
$M_0 = 2$	0	0	0	0
$M_0 = 4$	**26**	**28**	**25**	**63**
$M_0 = 6$	0	26	16	10

(b) $n = 1000$

	$\epsilon = 0.25$	$\epsilon = 0.3$	$\epsilon = 0.4$	$\epsilon = 0.5$
$M_0 = 2$	**113**	**116**	0	0
$M_0 = 4$	31	38	**38**	**47**
$M_0 = 6$	18	40	14	19

(c) $n = 2000$

Figure 4: The identifiability criterion $ID(M_0)$ for different M_0, n and ϵ.

5 Conclusion and discussion

To sum up our results, we propose semiparametric estimators of the mixing weights in a mixture with a finite number of components and unspecified emission distributions. These estimators are constructed using an approximate model for the mixture where the emission densities are modelled as piecewise constant functions on fixed partitions of the sampling space. This approximate model is thus parametric and regular and more importantly well specified as far as the weight parameters θ are concerned. From Theorem

1 we have that for all $M \geq M_0$, $\sqrt{n}\left(\widehat{\theta}_{\tilde{M},n} - \theta^\star\right) \rightsquigarrow \mathcal{N}\left(\theta^\star, \tilde{J}_{\tilde{M}}^{-1}\right)$ as n goes to infinity and that $\tilde{J}_{\tilde{M}}^{-1} \to \tilde{J}^{-1}$ as M goes to infinity (and similarly from a Bayesian point of view). Moreover we have proved in Section 3.1 that for all n, as M goes to infinity, $\widehat{\theta}_{M,n} \to \bar{\theta}_n$ and that as $n \to +\infty$, $\bar{\theta}_n \to (1/k, \cdots, 1/k)$ whatever the true value θ^* of the parameter. These two results show that we can find a sequence M_n going to infinity such that $\sqrt{n}\left(\widehat{\theta}_{\tilde{M}_n,n} - \theta^\star\right) \rightsquigarrow \mathcal{N}\left(\theta^\star, \tilde{J}^{-1}\right)$ but also that we cannot choose M_n going to infinity arbitrarily fast. It is thus important to determine a procedure to select M, for finite n.

To choose M_n in practice, for finite n, we propose in Section 3.2 an approach which consists in minimizing an estimate of the quadratic risk $R_n(\mathcal{I})$ in the partition $\mathcal{I}$, as a way to ensure that the asymptotic variance of $\sqrt{n}(\widehat{\theta} - \theta)$ is close to $\tilde{J}^{-1}$ and that the quantity $\sqrt{n}(\widehat{\theta} - \theta)$ is asymptotically stable. The construction of an estimator of $R_n(\mathcal{I})$ is not trivial due to the strong non linearity of the maximum likelihood estimator in mixture models and we use a reference model with a small number of bins M_0 as a proxy for an unbiased estimator θ, together with a cross validation approach to approximate $R_{a_n}(\mathcal{I}$ for all partition $\mathcal{I}$ with $a_n = o(n)$. This leads at best to a minimization of the risk $R_{a_n}(\mathcal{I}_M)$ instead of $R_n(\mathcal{I}_M)$, however this is it not per se problematic since a major concern is to ensure that M_n is not too large.

In the construction of our estimation procedure (either by MLE or based on the posterior distribution) we have considered the same partitionning of $[0, 1]$ for each coordinate $c \in \{1, 2, 3\}$. This can be relaxed easily by using different partitions accross coordinates, if one wishes to do so to adapt to different smoothness of emission densities for instance. However, this would require choices of M for each coordinate. We believe that our theoretical results would stay true. We did more simulations in this setting and we observed that, when the emission distributions are distinct in each direction, choosing different M for each coordinates is time consuming and does not really improve the estimations of θ, at least in our examples.

We have also presented our results under some seemingly restrictive assumptions, which we now discuss.

5.1 On the structural assumptions on the model

In model (1), it is assumed that each individual has three conditionally independent observations in $[0, 1]$ each. Obviously this assumptions can be relaxed to any number p of conditionally independent observations with $p \geq 3$ without modifying the conclusions of our results.

Also, the method of estimation relies heavily on the fact that the $X_{i,c}$'s belong to $[0, 1]$. This is not such a restrictive assumption since one can transform any random variable on $\mathbb{R}$ into $[0, 1]$, writing $X_{i,c} = G_c(\tilde{X}_{i,c})$, where G_c is a given cumulative distribution on $\mathbb{R}$ and $\tilde{X}_{i,c}$ is the original observation. Then the conditional densities are obtained as

$$f^\star_{j,c}(x_c) = f_{\tilde{X};j,c}(G_c^{-1}(x_c))/g_c(G_c^{-1}(x_c)), \quad j \leq k, c \in \{1, \cdots, 3\}$$

and Assumption (4) becomes that for all $c \in \{1, \cdots, 3\}$,

$$0 < \text{liminf}_{x_c} \frac{f_{\tilde{X};j_1,c}(G_c^{-1}(x_c))}{f_{\tilde{X};j_2,c}(G_c^{-1}(x_c))} \leq \text{limsup}_{x_c} \frac{f_{\tilde{X};j_1,c}(G_c^{-1}(x_c))}{f_{\tilde{X};j_2,c}(G_c^{-1}(x_c))} < +\infty, \tag{21}$$

which means that the densities of the observations within each group have all the same tail behaviour. Note that a common assumption found in the literature for estimation of densities on $[0, 1]$ is that the densities are bounded from above and below, which in the above framework of transformations G_c amounts to saying that $f_{\tilde{X};j_1,c}$'s have all the same tail behaviours as $g(\cdot)$. This is a much stronger assumption because it would mean that the tail behaviour of the densities $f_{\tilde{X};j_1,c}$ is known a priori, whereas (21) only means that the tails are the same between the components of the mixtures but they not need to be the same as those of g.

Finally we have considered univariate conditional observations $X_{i,c} \in \mathbb{R}$, again this can be relaxed easily by considering partitions of $[0, 1]^d$ with $d \geq 1$ if $X_{i,c} \in [0, 1]^d$. In this case the first part of Assumption (A2) needs to be replaced by :

• There exists $a > 0$ such that for all M, for all I_m in $\mathcal{I}_M$, there exists an open ball I such that $I_m \subset I$ and $|I_m| \geq a|I|$.

5.2 Extensions to Hidden Markov models

Finite mixture models all have the property that, when the approximation space for the emission distributions is that of step functions (histograms), then the model stays true for the observation process, but associated to the summary of the observations made of the counts in each bins. This leads to a proper and well specified likelihood for the parameter θ, w and there is no problem of model misspecification as fgar as θ is concerned even when the number of bins is fixed and small. We expect the results obtained in this paper to remain valid for nonparametric hidden Markov models with translated emission distributions studied in [21] or for general nonparametric finite state space hidden Markov models studied in [18], [35] and [19]. In the latter, the parameter describing the probability distribution of the latent variable is the transition matrix of the hidden Markov chain. However, semiparametric asymptotic theory for dependent observations is much more involved, see [27] for the ground principles. It seems difficult to identify the score functions and the efficient Fisher information matrices for hidden Markov models even in the parametric approximation model, so that to get results such as Theorem 1 could be quite challenging, nevertheless we think that the results obtained here pave the way to obtaining semi-parametric efficient estimation of the transition matrix in nonparametric hidden Markov models.

6 Proofs

6.1 Proof of Proposition 1

Let us first prove that for large enough M, the measures $f^\star_{1,c;M}dx, \ldots, f^\star_{k,c;M}dx$ are linearly independent. Indeed, if it is not the case, there exists a subsequence M_p tending to infinity

as p tends to infinity and a sequence $(\alpha^{(p)})_{p\geq 1}$ in the unit ball of $\mathbb{R}^k$ such that for all $p \geq 1$,

$$\sum_{j=1}^{k} \alpha_j^{(p)} f^{\star}_{j,c;M_p}(x) = 0$$

Lebesgue a.e. Let $\alpha = (\alpha_1, \ldots, \alpha_k)$ be a limit point of $(\alpha^{(p)})_{p\geq 1}$ in the unit ball of $\mathbb{R}^k$. Using Assumption (A.2) and Corollary 1.7 in Chapter 3 of [32], we have that as p tends to infinity, $f^{\star}_{j,c;M_p}(x)$ converges to $f^{\star}_{j,c}(x)$ Lebesgue a.e. so that we obtain $\sum_{j=1}^{k} \alpha_j f^{\star}_{j,c}(x) = 0$ Lebesgue a.e., contradicting Assumption (A1).
Fix now M large enough so that the measures $f^{\star}_{1,c;M}dx, \ldots, f^{\star}_{k,c;M}dx$ are linearly independent. Then, one may use the spectral method described in [2] to get estimators $\widehat{\theta}_{sp}$ and $\widehat{\boldsymbol{\omega}}_{M;sp}$ of the parameters θ and $\boldsymbol{\omega}_M$ from a sample of the multinomial distribution associated to density $g_{\theta,\boldsymbol{\omega};M}$. The estimator uses eigenvalues and eigenvectors computed from the empirical estimator of the multinomial distribution. But in a neighborhood of $\theta^{\star}$ and $\boldsymbol{\omega}^{\star}$, this is a continuously derivative procedure, and since on this neighborhood, classical deviation probabilities on empirical means hold uniformly, we get easily that for any vector $V \in \mathbb{R}^k$, there exists $K > 0$ such that for all $c > 0$, for large enough n (the size of the sample):

$$\sup_{\|\theta-\theta^{\star}\|\leq \frac{c}{\sqrt{n}}} \mathbb{E}_{\theta}\left[\left(\sqrt{n}\langle \widehat{\theta}_{sp} - \theta, V\rangle\right)^2\right] \leq K.$$

Now, the multinomial model is differentiable in quadratic mean, and following the proof of Theorem 4 in [20] one gets that, if $V^T \tilde{J}_M V = 0$, then

$$\lim_{c\to+\infty} \lim_{n\to+\infty} \sup_{\|\theta-\theta^{\star}\|\leq \frac{c}{\sqrt{n}}} \mathbb{E}_{\theta}\left[\left(\sqrt{n}\langle \widehat{\theta}_{sp} - \theta, V\rangle\right)^2\right] = +\infty.$$

Thus for all $V \in \mathbb{R}^k$, $V^T \tilde{J}_M V \neq 0$, so that $\tilde{J}_M$ is not singular.

6.2 Proof of Proposition 2

We prove the proposition when $M_1 = M$, $M_2 = M+1$, $\mathcal{I}_M = \{I_1, \ldots, I_M\}$ and $\mathcal{I}_{M+1} = \{I_1, \ldots, I_{M,0}, I_{M,1}\}$ with $I_M = I_{M,0} \cup I_{M,1}$, which is sufficient by induction. We denote $(\omega^{(M)}_{j,c,m})_{j,c,m}$ the parameter $\boldsymbol{\omega}$ in the model with partition $\mathcal{I}_M$ and $(\omega^{(M+1)}_{j,c,m})_{j,c,m}$ the parameter $\boldsymbol{\omega}$ in the model with partition $\mathcal{I}_{M+1}$. Define $b \in (0,1)$, $\alpha_{j,c} \in (0,1)$, $j = 1, \ldots, k$, $c = 1, 2, 3$ so that

$$|I_{M,0}| = (1-b)|I_M|,\ |I_{M,1}| = b|I_M|,\ \omega^{(M+1)}_{j,c,M} = (1-\alpha_{j,c})\omega^{(M)}_{j,c,M},$$
$$\omega^{(M+1)}_{j,c,M+1} = \alpha_{j,c}\omega^{(M)}_{j,c,M}.$$

Then, we may write

$$g_{\theta,\boldsymbol{\omega};M}(\mathbf{x}) = \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M} \left(\frac{\omega^{(M)}_{j,c,m}}{|I_m|}\right)^{\mathbb{1}_{I_m}(x_c)}$$

and

$$
\begin{aligned}
&g_{\theta,\boldsymbol{\omega};M+1}(\mathbf{x}) \\
&= \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M-1} \left(\frac{\omega_{j,c,m}^{(M+1)}}{|I_m|} \right)^{\mathbb{1}_{I_m}(x_c)} \left[\left(\frac{\omega_{j,c,M}^{(M+1)}}{|I_{M,0}|} \right)^{\mathbb{1}_{I_{M,0}}(x_c)} \left(\frac{\omega_{j,c,M+1}^{(M+1)}}{|I_{M,1}|} \right)^{\mathbb{1}_{I_{M,1}}(x_c)} \right] \\
&= \sum_{j=1}^{k} \theta_j \prod_{c=1}^{3} \prod_{m=1}^{M} \left(\frac{\omega_{j,c,m}^{(M)}}{|I_m|} \right)^{\mathbb{1}_{I_m}(x_c)} \left[\left(\frac{\alpha_{j,c}}{b} \right)^{\mathbb{1}_{I_{M,1}}(x_c)} \left(\frac{1-\alpha_{j,c}}{1-b} \right)^{\mathbb{1}_{I_{M,0}}(x_c)} \right].
\end{aligned}
$$

Thus, when $x_c \notin I_M$ for $c = 1, 2, 3$, $g_{\theta,\boldsymbol{\omega};M+1}(\mathbf{x}) = g_{\theta,\boldsymbol{\omega};M}(\mathbf{x})$ and computations have to take care of $\mathbf{x}$'s such that for some c, $x_c \in I_M$. If we parametrize the model with partition $\mathcal{I}_{M+1}$ using the parameter $\left(\theta, (\omega_{j,c,m}^{(M)}), (\alpha_{j,c})\right)$ we get the same efficient Fisher information for θ as when parametrizing with $\left(\theta, (\omega_{j,c,m}^{(M+1)})\right)$. Define the function D as the difference between the gradient of $\log g_{\theta,\boldsymbol{\omega};M+1}$ and that of $\log g_{\theta,\boldsymbol{\omega};M}(\mathbf{x})$ with respect to the parameter $\left(\theta, (\omega_{j,c,m}^{(M)}), (\alpha_{j,c})\right)$:

$$
D(\mathbf{x}) := \nabla \log g_{\theta,\boldsymbol{\omega};M+1}(\mathbf{x}) - \nabla \log g_{\theta,\boldsymbol{\omega};M}(\mathbf{x}),
$$

in particular the last coordinates of $\nabla \log g_{\theta,\boldsymbol{\omega};M}(\mathbf{x})$ corresponding to the derivatives with respect to $(\alpha_{j,c})$ are zero. Let us denote $K^{(M+1)}$ the Fisher information obtained for this new parametrization, that is

$$
K^{(M+1)} = \mathbb{E}^\star[(\nabla \log g_{\theta,\boldsymbol{\omega};M+1}(X))(\nabla \log g_{\theta,\boldsymbol{\omega};M+1}(X))^T].
$$

Easy but tedious computations give

$$
\mathbb{E}^\star[(\nabla \log g_{\theta,\boldsymbol{\omega};M}(X))(D(X))^T] = \begin{pmatrix} 0 & \cdots & 0 \\ \vdots & \vdots & \vdots \\ 0 & \cdots & 0 \end{pmatrix},
$$

so that

$$
K^{(M+1)} = \begin{pmatrix} J_M & 0 \\ 0 & 0 \end{pmatrix} + \Delta
$$

where $\Delta = \mathbb{E}^\star[D(X)D(X)^T]$ is positive semi-definite. As said before, $\tilde{J}_{M+1}$ is obtained from $K^{(M+1)}$ using the similar formula as from J_{M+1}. Then usual algebra gives that $\tilde{J}_{M+1} \geq \tilde{J}_M$ since Δ is positive semi-definite.

6.3 Proof of Lemma 1

Under (A1), the functions $f_{j,c}^\star$ are upper bounded. Let, for any M, $\mathbb{A}_M$ be the orthogonal projection in $L^2(g_{\theta^\star,\mathbf{f}^\star} d\mathbf{x})$ onto $\dot{\mathcal{P}}_M$, the set of step functions spanned by the functions $\left(S_{\boldsymbol{\omega},M}^\star\right)_{j,c,m}$, $j = 1, \ldots, k$, $c = 1, 2, 3$, $m = 1, \ldots, M-1$. Then for all $j = 1, \ldots, k-1$,

$$
(\tilde{\psi}_M)_j = \left(S_{\theta,M}^\star\right)_j - \mathbb{A}_M \left(S_{\theta,M}^\star\right)_j,
$$

so that

$$(\tilde{\psi})_j - (\tilde{\psi}_M)_j = (S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j - \mathbb{A}_M \left[(S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j\right] + (\mathbb{A}_M - \mathbb{A})\,(S^\star_\theta)_j\,. \tag{22}$$

Using Assumption (A2) and Corollary 1.7 in Chapter 3 of [32], we have that as M tends to infinity, $\left(S^\star_{\theta,M}\right)_j$ converges to $(S^\star_\theta)_j$ Lebesgue a.e. Both functions are uniformly upper bounded by the finite constant $1/\theta^\star_j$ using Assumption (A.1), so that $\left(S^\star_{\theta,M}\right)_j$ converges to $(S^\star_\theta)_j$ in $L^2(g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})d\mathbf{x})$ as M tends to $+\infty$ and $\left\|(S^\star_\theta)_j - \left(S^\star_{\theta,M}\right)_j\right\|_{L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{x})}$ converges to 0 as M tends to $+\infty$. Thus to prove that $(\tilde{\psi})_j - (\tilde{\psi}_M)_j$ converges to 0 in $L^2(g_{\theta^\star,\mathbf{f}^\star}d\mathbf{x})$ when M tends to $+\infty$, we need only to prove that $\|\,(\mathbb{A}_M - \mathbb{A})\,(S^\star_\theta)_j\,\|_{L^2(g_{\theta^\star,\mathbf{f}^\star})}$ converges to 0. So we now prove that, for all $S \in L^2(g_{\theta^\star,\mathbf{f}^\star})$, $\|\mathbb{A}_M S - \mathbb{A}S\|_{L^2(g_{\theta^\star,\mathbf{f}^\star})}$ converges to 0 when M tends to $+\infty$.
First we prove that $\mathbb{A}_M S$ converges in $L^2(g_{\theta^\star,\mathbf{f}^\star})$. Let $L \geq M$ and set $\psi_S(M,L) = \|\mathbb{A}_M S - \mathbb{A}_L S\|_{L^2(g_{\theta^\star,\mathbf{f}^\star})}$. For large enough M, we have $\psi_S(M,L) = \|\mathbb{A}_M(\mathbb{A}_L S) - \mathbb{A}_L S\|_{L^2(g_{\theta^\star,\mathbf{f}^\star})}$, since using (A3), $\dot{\mathcal{P}}_M \subset \dot{\mathcal{P}}_L$. It is easy to see that, for all M, $\psi_S(M,L)$ is a monotone sequence, non decreasing in L, and bounded, so that it converges to some $\psi^\star_S(M) \geq 0$. Moreover, since for all L, $\psi_S(M+1,L) \leq \psi_S(M,L)$, at the limit $\psi^\star_S(M)$ is monotone non increasing in M and non-negative so that it converges. Let $\psi^\star_S$ be its limit. Because

$$\begin{aligned}
&\mathbb{A}_L S = \mathbb{A}_M(\mathbb{A}_L S) + (I - \mathbb{A}_M)(\mathbb{A}_L S) = \mathbb{A}_M S + (I - \mathbb{A}_M)(\mathbb{A}_L S),\\
&\mathbb{A}_M S \perp (I - \mathbb{A}_M)(\mathbb{A}_L S),
\end{aligned}$$

we get that

$$\|\mathbb{A}_L S\|^2 = \|\mathbb{A}_M S\|^2 + \|(I - \mathbb{A}_M)(\mathbb{A}_L S)\|^2 = \|\mathbb{A}_M S\|^2 + \psi_S(M,L)^2.$$

Let M be fixed. Then we get that

$$\lim_{L\to+\infty} \|\mathbb{A}_L S\|^2 = \|\mathbb{A}_M S\|^2 + \psi^\star_S(M)^2$$

and if we write ℓ the limit on the lefthandside of the equation, by letting now M tend to infinity we get that $\ell = \ell + \psi^\star_S$. This in turns implies that $\psi^\star_S = 0$. Now let L_p, M_p converge to infinity as p goes to infinity in such a way that for all p $L_p \geq M_p$ (which we can always assume by symmetry). Then

$$\psi_S(M_p, L_p) \leq \psi^\star_S(M_p) \overset{p\to+\infty}{\to} 0$$

so that the sequence $\mathbb{A}_M S$ is Cauchy in $L^2(g_{\theta^\star,\mathbf{f}^\star})$ and converges. Denote $\overline{\mathbb{A}}S$ its limit. Let us prove that $\overline{\mathbb{A}}S \in \dot{\mathcal{P}}$. Any function in $\dot{\mathcal{P}}_M$ is a finite linear combination of functions $S_{j,c,M}$ of form

$$S_{j,c,M}(\mathbf{x}) = \frac{h_{j,c,M}(x_c)\prod_{c'} f_{\omega^\star_{j,c';M}}(x_{c'})}{g_{\theta^\star,\boldsymbol{\omega}^\star_M;M}(\mathbf{x})} \tag{23}$$

with $h_{j,c,M} \in \mathcal{H}_{j,c}$ is such that $h_{j,c,M} f_{\omega^\star_{j,c;M}}$ is a linear combination of indicator functions. It is thus enough to prove that the limit of any converging sequence of such functions is in $\dot{\mathcal{P}}$. Consider a sequence $S_{j,c,M}$ converging to some variables $S^\star$ in $L^2(g_{\theta^\star,\mathbf{f}^\star})$ as M tends to infinity. Note that

$$h_{j,c,M}(x_c) = \sum_{\ell=1}^{k} \theta^\star_\ell S_{j,c,M}(\mathbf{x}) \times \prod_{c'} \frac{f_{\omega^\star_{\ell,c';M}}(x_{c'})}{f_{\omega^\star_{j,c';M}}(x_{c'})}$$

so that condition (4) together with the almost sure convergence of $f_{\omega^\star_{\ell,c';M}}(x_{c'})/f_{\omega^\star_{j,c';M}}(x_{c'})$ towards $f^\star_{\ell,c'}(x_{c'})/f^\star_{j,c'}(x_{c'})$ implies that

$$S_{j,c,M}(\mathbf{x}) \times \prod_{c'} \frac{f_{\omega^\star_{\ell,c';M}}(x_{c'})}{f_{\omega^\star_{j,c';M}}(x_{c'})} \overset{L^2(g_{\theta^\star,\mathbf{f}^\star})}{\to} S^\star \prod_{c'} \frac{f^\star_{\ell,c'}(x_{c'})}{f^\star_{j,c'}(x_{c'})} := h,$$

which in turns implies that $h_{j,c,M}$ converges in $L^2(f^\star_{j,c})$ to h. Since $h_{j,c,M} \in \mathcal{H}_{j,c}$ for all M, $h \in \mathcal{H}_{j,c}$ and $S^\star \in \dot{\mathcal{P}}$.
We now prove that all function in $\dot{\mathcal{P}}$ is a limit in $L^2(g_{\theta^\star,\mathbf{f}^\star})$ of functions in $\dot{\mathcal{P}}_M$. As before, it is enough to prove it for functions $S_{j,c}$ of form

$$S_{j,c}(\mathbf{x}) = \frac{h(x_c) \prod_{c'} f^\star_{j,c'}(x_{c'})}{g_{\theta^\star,\mathbf{f}^\star}(\mathbf{x})}$$

with $h \in \mathcal{H}_{j,c}$. We are thus looking for a sequence of functions $h_M \in \mathcal{H}_{j,c}$ such that $h_M f_{\omega^\star_{j,c;M}}$ is a linear combination of indicator functions and such that $S_{j,c,M}$ as defined by (23) converges to $S_{j,c}$ in $L^2(g_{\theta^\star,\mathbf{f}^\star})$. Using Lemma 1.2 in [32], h may be approximated by a continuous function, which in turns may be approximated by a (centred) linear combination of indicator functions and the result follows using again that $\prod_{c'} f_{\omega^\star_{j,c';M}}/g_{\theta^\star,\boldsymbol{\omega}^\star_M;M}$ and $\prod_{c'} f^\star_{j,c'}/g_{\theta^\star,\mathbf{f}^\star}$ are bounded (by $1/\min_j \theta^\star_j$). Thus, we easily get that for all S in $L^2(g_{\theta^\star,\mathbf{f}^\star})$, $\mathbb{A}_M \mathbb{A} S - \mathbb{A} S$ converges to 0 in $L^2(g_{\theta^\star,\mathbf{f}^\star})$, so that $\overline{\mathbb{A}}\mathbb{A} = \mathbb{A}$.
Now, one easily deduces that $\overline{\mathbb{A}} = \mathbb{A}$. Indeed: if S is in $\dot{\mathcal{P}}$, one has $\mathbb{A}S = S$, and then $\overline{\mathbb{A}}S = S$. If now S is in the orthogonal of $\dot{\mathcal{P}}$, then for any $\tilde{S} \in \dot{\mathcal{P}}$, one has $\langle \mathbb{A}_M S, \tilde{S}\rangle = \langle S, \mathbb{A}_M \tilde{S}\rangle$ which leads to $\langle \overline{\mathbb{A}} S, \tilde{S}\rangle = \langle S, \overline{\mathbb{A}}\tilde{S}\rangle = 0$, so that $\overline{\mathbb{A}}S$ is in the orthogonal of $\dot{\mathcal{P}}$ and in $\dot{\mathcal{P}}$ so that $\overline{\mathbb{A}}S$=0.

6.4 Proof of Proposition 3

Proposition 3 is easily implied by Lemma 2 which formalizes the following. When the sequence of observations $X_1, \ldots, X_n$ and n are fixed, then almost surely there exists a sufficiently fine partition $\mathcal{I}_M$ such that there exists at most one component of an observation in each set I_m, $m \leq M$. Then we can reorder the sets I_m so that $X_{i,c} \in I_{i+n(c-1)}$, for all $c \in \{1,2,3\}$ and $i \leq n$. In this case, the likelihood $\ell_n(\cdot,\cdot;M)$ is maximised at each parameter $(\theta, \boldsymbol{\omega})$ belonging to the set $\mathcal{S}_M \subset \Delta_k \times (\Delta_M)^{3k}$ that we explain now (and formalise in Lemma 2). Each element of $\mathcal{S}_M$ corresponds to one clustering of the observations in k sets

(represented by the $(A_j^\star)_{j\leq k}$ in Lemma 2) of size as equal as possible. For each clustering, for all $j \leq k$,
• $\theta_j = \#A_j^\star/n$ is the proportion of observations associated to $A_j^\star$ (then the θ_j are almost equal to $1/k$),
• for all $c \in \{1,2,3\}$ and for all $l \leq M$,

$$\omega_{j,c,l} = \begin{cases} 1/\#A_j^\star & \text{if } l-n(c-1) \in A_j^\star \text{ (i.e. } X_{l-n(c-1)} \in I_l \text{ is associated to} \\ & \text{the hidden state } j\text{),} \\ 0 & \text{if } l-n(c-1) \in \{1,\dots n\} \setminus A_j^\star \text{ (i.e. } X_{l-n(c-1)} \in I_l \text{ is not} \\ & \text{associated to } j\text{),} \\ 0 & \text{otherwise (i.e. there is no observation in } I_l\text{).} \end{cases}$$

Lemma 2. *Let $X_1,\dots,X_n$ be fixed observations, as soon as for all $i \leq n$ and $c \in \{1,2,3\}$, $X_{i,c} \in I_{i+n(c-1)}$ then the likelihood $\ell_n(\cdot,\cdot;M)$ is maximised at $(\widehat{\theta}_M, \widehat{\boldsymbol{\omega}}_M)$ if and only if $(\widehat{\theta}_M, \widehat{\boldsymbol{\omega}}_M) \in \mathcal{S}_M$ where*

$$\begin{aligned} \mathcal{S}_M = \big\{(\theta,\boldsymbol{\omega}) : \theta_j = \#A_j^\star/n,\ \omega_{j,c,l} = \mathbb{1}_{l-n(c-1)\in A_j^\star}/\#A_j^\star, \\ (J_1,J_2) \textit{ partition of } \{1,\dots,k\},\ \#J_2 = n-k\lfloor n/k\rfloor =: r \\ (A_j^\star)_{j\leq k} \textit{ partition of } \{1,\dots,n\}, \#A_{j_1}^\star = \lfloor n/k\rfloor =: q, \textit{ for } j_1 \in J_1, \\ \#A_{j_2}^\star = \lfloor n/k\rfloor + 1 =: q+1, \textit{ for } j_2 \in J_2\big\}, \end{aligned}$$

and $n = kq + r,\ 0 \leq r \leq k-1$.

Proof. Since the set of parameters is compact and the likelihood is a continuous function of the parameters then the maximum is attained.

If $(\theta,\boldsymbol{\omega})$ maximises the likelihood $\ell_n(\cdot,\cdot;M)$,

(P1) then, for all $1 \leq i \leq n$, there exists $1 \leq j \leq k$ such that $\omega_{j,c,i+n(c-1)} > 0$ for all $c \in \{1,2,3\}$.
Indeed, if there exists $1 \leq i \leq n$ such that for all $1 \leq j \leq k$, $\omega_{j,c,i+n(c-1)} = 0$ for some $c \in \{1,2,3\}$, then

$$\ell_n(\theta,\omega;M) = \sum_{i=1}^n \log\left(\sum_{j=1}^k \theta_j \prod_{c=1}^3 \omega_{j,c,i+n(c-1)}\right) + \underbrace{\sum_{i=1}^n \log\left(1/(|I_i||I_{i+n}||I_{i+2n}|)\right)}_{\text{constant}} = -\infty.$$

(P2) and if there exists j,c,i such that $\omega_{j,c,i+n(c-1)} = 0$ and $\theta_j > 0$ then $\omega_{j,d,i+n(d-1)} = 0$ for all d.
Indeed otherwise you can give the weight $\omega_{j,d,i+n(d-1)}$, to one of the other $\omega_{j,d,s+n(d-1)}$ for which $\omega_{j,e,s+n(e-1)} > 0$, for all $e \neq d$ (which exist otherwise take $\theta_j = 0$ which would increase the likelihood) and this increases the likelihood.

(P3) and if $\theta_j > 0$, then $\omega_{j,c,l} = 0$ if $l - n(c-1) \notin \{1, \dots, n\}$.
Indeed, in this case, there is no observation in I_l so that $\omega_{j,c,l}$ does not appear in the likelihood and we conclude similarly as the previous point.

Combining all the previous remarks, we know that the maximum can only be attained (and is at least once) in one of the following sets, indexed by $J \subset \{1, \dots, k\}$ which determines the zeros of θ and $A_j \subset \{1, \dots, n\}$, $j \leq k$, which determine the zeros of ω:

$$\begin{aligned}\mathcal{S}_{J,A_1,\dots,A_k} =& \{\theta \in \Delta_k : \ \theta_j > 0, \ j \in J, \theta_j = 0, \ j \in J^c\} \\ &\times \prod_{j \leq k} \Big\{ (\omega_{j,1,\cdot}, \omega_{j,2,\cdot}, \omega_{j,3,\cdot}) \in (\Delta_M)^3 : \\ &\qquad \text{if } j \in J, \quad \underbrace{\omega_{j,c,i+n(c-1)}}_{\text{using (P2)}} > \underbrace{0}_{\text{using (P3)}}, \text{ if } i \in A_j, \ c \in \{1,2,3\} \\ &\qquad \text{and } \omega_{j,c,l} = 0, \text{ if } l \in \{1, \dots, M\} \setminus \{i + n(c-1), \ i \in A_j\} \Big\}.\end{aligned}$$

Note that we do not assume that $(A_j)_{j \in J}$ is a partition of $\{1, \dots, n\}$.

We fix $J \subset \{1, \dots, k\}$ and $A_j \subset \{1, \dots, n\}$, $j \in J$. Now we search for parameters $(\bar{\theta}, \bar{\boldsymbol{\omega}})$ in $\mathcal{S}_{J,A_1,\dots,A_k}$ which maximize the likelihood. They are zeros of the derivative of

$$(\theta, \boldsymbol{\omega}, \lambda, \mu) \mapsto \ell_n(\theta, \boldsymbol{\omega}; M) + \lambda \left(\sum_{j=1}^{k} \theta_j - 1 \right) + \sum_{c=1}^{3} \mu_{j,c} \left(\sum_i \omega_{j,c,i} - 1 \right), \tag{24}$$

with respect to non zero components (θ_j, $\omega_{j,c,i+n(c-1)}$, λ and $\mu_{j,c}$, for $j \in J$, $i \in A_j$, $1 \leq c \leq 3$). Annulling the partial derivatives give

$$\sum_{i \in A_j} \frac{\bar{\omega}_{j,1,i} \bar{\omega}_{j,2,i+n} \bar{\omega}_{j,3,i+2n}}{\sum_{s \in J(i)} \bar{\theta}_s \bar{\omega}_{s,1,i} \bar{\omega}_{s,2,i+n} \bar{\omega}_{s,3,i+2n}} = -\lambda, \qquad \forall j \in J \tag{25}$$

$$\frac{\bar{\theta}_j \prod_{d \neq c} \bar{\omega}_{j,d,i+n(d-1)}}{\sum_{s \in J(i)} \bar{\theta}_s \bar{\omega}_{s,1,i} \bar{\omega}_{s,2,i+n} \bar{\omega}_{s,3,i+2n}} = -\mu_{j,c}, \qquad \forall j \in J, \ i \in A_j, \ c \in \{1,2,3\} \tag{26}$$

$$\sum_{j \in J} \bar{\theta}_j = 1, \tag{27}$$

$$\sum_{i \in A_j} \bar{\omega}_{j,c,i+n(c-1)} = 1, \qquad \forall j \in J, \ c \in \{1,2,3\}, \tag{28}$$

where $J(i) = \{s \in J : \ i \in A_s\}$.

Multiplying Equation (26) by $\bar{\omega}_{j,c,i+n(c-1)}$ and then summing the result over $i \in A_j$ and using Equation (28), we obtain that $\mu_{j,c}$ does not depend on c. Then using Equations (26) for $c = 1$, $c = 2$ and $c = 3$, we obtain

$$\bar{\theta}_j \bar{\omega}_{j,1,i} \bar{\omega}_{j,2,i+n} = \bar{\theta}_j \bar{\omega}_{j,1,i} \bar{\omega}_{j,3,i+2n} = \bar{\theta}_j \bar{\omega}_{j,2,i+n} \bar{\omega}_{j,3,i+2n},$$

so that

$$\bar{\omega}_{j,1,i} = \bar{\omega}_{j,2,i+n} = \bar{\omega}_{j,3,i+2n}. \tag{29}$$

Furthermore, multiplying Equation (25) by $\bar{\theta}_j$ and summing the result over $j \in J$ and using Equation (27), we obtain $\lambda = -n$. Moreover by multiplying Equation (26) by $\bar{\omega}_{j,c,i+n(c-1)}$, and then summing the result over $i \in A_j$ and finally subtracting (25) multiplied by $\bar{\theta}_j$ to the result (ie making $\sum_{i\in A_j}(-\bar{\theta}_j)(25) + \bar{\omega}_{j,c,i+n(c-1)}(26)$), we get

$$0 = -\mu_{j,c} - n\bar{\theta}_j. \tag{30}$$

Then using again Equations (26), (29) and (30), we get

$$\bar{\omega}^2_{j,c,i+n(c-1)} = n \sum_{s\in J(i)} \bar{\theta}_s \bar{\omega}^3_{s,1,i}, \quad \forall j \in J(i),\ \forall c \in \{1,2,3\},$$

so that $\bar{\omega}_{j,c,i+n(c-1)}$ does not depend on $j \in J(i)$ and

$$\bar{\omega}_{j,c,i+n(c-1)} = \mathbb{1}_{i\in A_j} / \left(n \sum_{s\in J(i)} \bar{\theta}_s \right), \quad \forall j \in J(i). \tag{31}$$

For each $\mathcal{S}_{J,A_1,\ldots,A_k} =: \mathcal{S}$, we have obtained the zeros of the derivative of the log-likelihood, that we now denote $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\boldsymbol{\omega}})$, to emphasize the dependence with the considered set $\mathcal{S}$. We now want to know which of these zeros $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\boldsymbol{\omega}})$ are local maxima thanks to the second partial derivatives.

We consider sets $\mathcal{S}_{J,A_1,\ldots,A_k}$ for which there exists $i \le n$ such that there exist j and l are in $J(i)$ and $j \neq i$. We consider a second partial derivative of

$$\tilde{\ell}_n(\theta, \tilde{\boldsymbol{\omega}}; M) = \sum_{i=1}^{n} \log \left(\sum_{j=1}^{k} \theta_j (\tilde{\omega}_{j,1,i})^3 \right)$$

that is the log-likelihood (up to an additive constant) associated to the model where for all $1 \le m \le k$, $1 \le s \le n$, $\omega_{m,1,s} = \omega_{m,2,s+n} = \omega_{m,3,s+2n}$. Assume without loss of generality that $\theta_l \ge \theta_j$, then (using that $\theta_k = 1 - \sum_{m<k} \theta_m$ and $\omega_{j,1,n} = 1 - \sum_{s<n} \omega_{j,1,s}$),

$$\frac{\partial^2 \tilde{\ell}_n}{\partial \tilde{\omega}^2_{j,1,i}}({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\boldsymbol{\omega}}; M) = C \left(6\, {}^{\mathcal{S}}\bar{\theta}_j \sum_{m\in J(i)\setminus\{j\}} {}^{\mathcal{S}}\bar{\theta}_m - 3\, {}^{\mathcal{S}}\bar{\theta}^2_j \right) \ge C \left(6\, {}^{\mathcal{S}}\bar{\theta}_j\, {}^{\mathcal{S}}\bar{\theta}_l - 3\, {}^{\mathcal{S}}\bar{\theta}^2_j \right) > 0,$$

where $C > 0$. This implies that for all sets $\mathcal{S}_{J,A_1,\ldots,A_k} := \mathcal{S}$ where there exists $i \le n$ such that $\#J(i) > 1$, every zeros $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\boldsymbol{\omega}})$ is not a local maximum. So that the only possible local maxima of $\ell_n(\theta, \boldsymbol{\omega}; M)$ are the zeros $({}^{\mathcal{S}_{J,A_1,\ldots,A_k}}\bar{\theta}, {}^{\mathcal{S}_{J,A_1,\ldots,A_k}}\bar{\omega})$ where $\#J(i) = 1$ for all $i \le n$, i.e. when $(A_j)_{j\in J}$ forms a partition of $\{1,\ldots,n\}$.

So we now only consider sets A_j, $j \in J$ which form a partition of $\{1,\ldots,n\}$ and $\bar{\omega}_{j,c,i+n(c-1)} = \mathbb{1}_{i\in A_j}/(n\bar{\theta}_j)$ for $i \in A_j$, using Equation (31). As $\sum_{i\in A_j} \bar{\omega}_{j,1,i} = 1$, we then obtain that $\bar{\theta}_j = \#A_j/n = 1/(n\bar{\omega}_{j,1,i})$, for all $i \in A_j$. So that we now only have to choose the best partition $(A_j)_{j\in J}$ of $\{1,\ldots,n\}$ and J. Let $N_j = \#A_j$, we know that $\sum_j N_j = n$ and the log-likelihood at the local maximum $({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega})$ associated to $\mathcal{S}_{J,A_1,\ldots,A_k} =: \mathcal{S}$ is

$$\ell_n({}^{\mathcal{S}}\bar{\theta}, {}^{\mathcal{S}}\bar{\omega}; M) = \sum_{s\in J} N_s \log(N_s^{-2}) + constant.$$

So that we want to minimize

$$\sum_{s\in J} N_s \log(N_s) \text{ under the constraint } \sum_{s\in J} N_s = n \tag{32}$$

over $J \subset \{1, \dots k\}$ and $N_j \in \mathbb{N}$, $j \in J$. This minimization is equivalent to the minimization of

$$\sum_{s\leq k} N_s \log(N_s) \text{ under the constraint } \sum_{s\leq k} N_s = n \tag{33}$$

over $N_j \in \mathbb{N}$, $j \leq k$ (since then the problem (33) is less constrained than for the minimization of (32) when J is fixed).

And, when k divides n, the minimum of (33) is attained at $N_s = n/k$. Otherwise, when k does not divide n, consider only two indices s_1, s_2 in $\{1, \dots, k\}$ and assume that N_s, $s \notin \{s_1, s_2\}$ are fixed such that $N_{s_1} + N_{s_2} = S_N$ is also fixed. Then we want to minimise $-N_{s_1}\log(N_{s_1}) - (S_N - N_{s_1})\log(S_N - N_{s_1})$. Studying the function $x \in (0, S_N) \mapsto -x\log(S_N) - (S_N - x)\log(S_N - x)$, we obtain that the minimum is attained when N_{s_1} and $N_{s_2} = S_N - N_{s_1}$ are the closest of $N_S/2$. Then in both cases, the MLE is attained at every $(\theta, \boldsymbol{\omega}) \in \mathcal{S}_M$.

□

6.5 Proof of Corollary 1

Suppose that for all $N > 0$ and all $C > 0$, there exists $n \geq N$ such that

$$n^2 \left(\max_{m\leq M_n} |I_m|\right)^2 M_n \leq C.$$

So that there exists a subsequence $(\phi(n))_{n\in\mathbb{N}}$ of $(n)_{n\in\mathbb{N}}$ such that

$$(\phi(n))^2 \left(\max_{m\leq M_{\phi(n)}} |I_m|\right)^2 M_{\phi(n)} \underset{n\to\infty}{\longrightarrow} 0. \tag{34}$$

Set $\epsilon > 0$, by Proposition 3, there exists $N_1 > 0$ such that for all $n \geq N_1$,

$$\begin{aligned}
&P\left(\left|\widehat{\theta}_{M_n}(X_{1:\phi(n)}) - (1/k, \dots, 1/k)\right| \leq \epsilon\right) \\
&\quad \geq P\left(\left\{\exists\, 1 \leq i_1, i_2 \leq \phi(n),\ 1 \leq c, d \leq 3,\ m \leq M_{\phi(n)} :\ X_{i_1,c} \in I_m, X_{i_2,d} \in I_m\right\}^c\right) \\
&\quad \geq 1 - \sum_{i_1=1}^{\phi(n)} \sum_{i_2=1}^{\phi(n)} \sum_{m=1}^{M_{\phi(n)}} P\left(X_{i_1,c} \in I_m, X_{i_2,d} \in I_m\right) \\
&\quad \geq 1 - (\phi(n))^2 M_{\phi(n)} \max\left(\sup g, (\sup g)^2\right) \left(\max_{m\leq M_{\phi(n)}} |I_m|\right)^2.
\end{aligned} \tag{35}$$

Using Equations (34) and (35) and Assumption (A3), then $\widehat{\theta}_{M_n}(X_{1:\phi(n)})$ tends in probability to $(1/k, \dots, 1/k)$ which contradicts the convergence in law of $\widehat{\theta}_{M_n}$ to $\theta^\star$. This concludes the proof.

6.6 Proof of Theorem 2

We first recall Lemma 2.1 in [4]:

Lemma 3 (Sylvain Arlot). *Let $A, B, C, R : \mathcal{M} \to \mathbb{R}$. If for all $m, m' \in \mathcal{M}$,*

$$(C(m) - R(m)) - (C(m') - R(m')) \leq A(m) + B(m'),$$

then for all $\widehat{m} \in \mathcal{M}$ such that $C(\widehat{m}) \leq \inf_{m \in \mathcal{M}} C(m) + \rho$, $\rho > 0$,

$$R(\widehat{m}) - B(\widehat{m}) \leq \inf_{m \in \mathcal{M}} \{R(m) + A(m)\} + \rho.$$

We are going to use this lemma with $R(\mathcal{I}) = R_{a_n}(\mathcal{I})$, $C(\mathcal{I}) = C_{CV}(\mathcal{I})$ and

$$A(\mathcal{I}) = B(\mathcal{I}) = \epsilon_n R(\mathcal{I}) + \delta_n.$$

Using Hoeffding's inequality,

$$P\left(\{-B(\mathcal{I}) \leq C_{CV}(\mathcal{I}) - R_{a_n}(\mathcal{I}) - \mathbb{E}^\star\left[\|\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}}) - \theta^\star\|^2_{\mathcal{T}_k}\right] \leq A(\mathcal{I})\}^c\right) \leq 2\exp\left(-2b_n A(\mathcal{I})^2\right),$$

since $\|\tilde{\theta}_{\mathcal{I}}(X_{B_b}) - \tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}})\|^2 \leq 1$, for all b. We introduce the sets

$$\mathcal{S}_{\mathcal{I}} = \left\{-B(\mathcal{I}) \leq C_{CV}(\mathcal{I}) - R_{a_n}(\mathcal{I}) - \mathbb{E}^\star\left[\|\tilde{\theta}_{\mathcal{I}_0}(X_{B_{-b}}) - \theta^\star\|^2_{\mathcal{T}_k}\right] \leq A(\mathcal{I})\right\} \tag{36}$$

for all $\mathcal{I} \in \mathcal{M}_n$. Using Lemma 3, on the set $\cap_{\mathcal{I} \in \mathcal{M}_n} \mathcal{S}_{\mathcal{I}}$, Equation (20) holds and using Equation (36), we obtain

$$P(\cap_{\mathcal{I} \in \mathcal{M}_n} \mathcal{S}_{\mathcal{I}}) \geq 1 - 2m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I} \in \mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right).$$

6.7 Proof of Proposition 4

Using Theorem 2,

$$\begin{aligned}
&\mathbb{E}^\star\left[a_n R_{a_n}(\widehat{\mathcal{I}}_n)\right] \\
&\leq a_n\left(\frac{1+\epsilon_n}{1-\epsilon_n} \inf_{\mathcal{I} \in \mathcal{M}_n} R_{a_n}(\mathcal{I}) + \frac{2\delta_n}{1-\epsilon_n}\right) \\
&\quad + 2a_n m_n \exp\left(-2b_n\left(\epsilon_n \inf_{\mathcal{I} \in \mathcal{M}_n} R_{a_n}(\mathcal{I}) + \delta_n\right)^2\right)
\end{aligned}$$

we can conclude by taking $\epsilon_n = \delta_n = 1/(\log(n) a_n)$.

Acknowledgements

This work was partly supported by the grants ANR Banhdits and Calibration. We want to thank the reviewers and the associate editor for their helpful comments.

References

[1] E. S. Allman, C. Matias, and J. A. Rhodes. Identifiability of parameters in latent structure models with many observed variables. *Ann. Statist.*, 37(6A):3099–3132, 12 2009.

[2] A. Anandkumar, R. Ge, D. Hsu, S. M. Kakade, and M. Telgarsky. Tensor decompositions for learning latent variable models. *JMLR*, 15:2773–2832, 2014.

[3] T. Ando. *Bayesian model selection and statistical modeling*. Statistics: Textbooks and Monographs. CRC Press, Boca Raton, FL, 2010.

[4] S. Arlot. *Contributions to statistical learning theory: estimator selection and change-point detection*. Habilitation à diriger des recherches, University Paris Diderot, December 2014. Habilitation à diriger des recherches.

[5] S. Arlot and A. Celisse. A survey of cross-validation procedures for model selection. *Stat. Surv.*, 4:40–79, 2010.

[6] P. Barbe and P. Bertail. *The Weighted Bootstrap*, volume 98 of *Lecture Notes in Statistics*. Springer, 1995.

[7] J.-P. Baudry, C. Maugis, and B. Michel. Slope heuristics: overview and implementation. *Stat. Comput.*, (22):455–470, 2012.

[8] P. J. Bickel, C. A. J. Klaassen, Y. Ritov, and J. A. Wellner. *Efficient and adaptive estimation for semiparametric models*. Johns Hopkins Series in the Mathematical Sciences. Johns Hopkins University Press, Baltimore, MD, 1993.

[9] P. J. Bickel and B. J. K. Kleijn. The semiparametric Bernstein-von Mises theorem. *Ann. Statist.*, 40(1):206–237, 2012.

[10] S. Bonhomme, K. Jochmans, and J.-M. Robin. Estimating multivariate latent-structure models. *Ann. Statist.*, 44(2):540–563, 2016.

[11] S. Bonhomme, K. Jochmans, and J.-M. Robin. Non-parametric estimation of finite mixtures from repeated measurements. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(1):211–229, 2016.

[12] S. Boucheron and E. Gassiat. A Bernstein-von Mises theorem for discrete probability distributions. *Electron. J. Stat.*, 3:114–148, 2009.

[13] M. A. Brookhart and M. J. van der Laan. A semiparametric model selection criterion with applications to the marginal structural model. *Comput. Statist. Data Anal.*, 50(2):475–498, 2006.

[14] I. Castillo. Semiparametric Bernstein–von Mises theorem and bias, illustrated with Gaussian process priors. *Sankhya A*, 74(2):194–221, 2012.

[15] I. Castillo. A semiparametric Bernstein–von Mises theorem for Gaussian process priors. *Probab. Theory Related Fields*, 152(1-2):53–99, 2012.

[16] G. Claeskens and N. L. Hjort. *Model selection and model averaging*, volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 2008.

[17] P. De Blasi and N. L. Hjort. The Bernstein–von Mises theorem in semiparametric competing risks models. *J. Statist. Plann. Inference*, 139(7):2316–2328, 2009.

[18] Y. De Castro, E. Gassiat, and C. Lacour. Minimax adaptive estimation of nonparametric hidden Markov models. *JMLR*, 17(111), 2016.

[19] Y. De Castro, E. Gassiat, and S. Le Corff. Consistent estimation of the filtering and marginal smoothing distributions in nonparametric hidden Markov models. *I.E.E.E. Trans. Info. Th.*, 63(8):4758–4777, 2017.

[20] E. Gassiat, D. Pollard, and G. Stoltz. Revisiting the van Trees inequality in the spirit of Hajek and Le Cam. *unpublished manuscript*, 2013.

[21] E. Gassiat and J. Rousseau. Non parametric finite translation hidden Markov models and extensions. *Bernoulli*, 22(1):193–212, 2016.

[22] M. H. Hansen and B. Yu. Model selection and the principle of minimum description length. 96:746–774, 2001.

[23] J. B. Kruskal. Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics. *Linear Algebra and Appl.*, 18(2):95–138, 1977.

[24] L. Le Cam and G. Yang. *Asymptotics in Statistics. Some Basic Concepts, Second Edition*. Springer-Verlag, New-York, 2000.

[25] Gyemin Lee and Clayton Scott. EM algorithms for multivariate Gaussian mixture models with truncated and censored data. *Comput. Statist. Data Anal.*, 56(9):2816–2829, 2012.

[26] P. Massart. *Concentration inequalities and model selection*, volume 1896 of *Lecture Notes in Mathematics*. Springer, Berlin, 2007. Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003, With a foreword by Jean Picard.

[27] B. McNeney and J. A. Wellner. Application of convolution theorems in semiparametric models with non-i.i.d. data. *J. Statist. Plann. Inference*, 91(2):441–480, 2000. Prague Workshop on Perspectives in Modern Statistical Inference: Parametrics, Semi-parametrics, Non-parametrics (1998).

[28] J. A. Rhodes. A concise proof of kruskal's theorem on tensor decomposition. *Linear Algebra and Appl.*, 432(7):1818–1824, 2010.

[29] V. Rivoirard and J. Rousseau. Bernstein-von Mises theorem for linear functionals of the density. *Ann. Statist.*, 40(3):1489–1523, 2012.

[30] C.P. Robert. *The Bayesian Choice*. Springer-Verlag, New York, second edition, 2001.

[31] X. Shen. Asymptotic normality of semiparametric and nonparametric posterior distributions. *J. Amer. Statist. Assoc.*, 97(457):222–235, 2002.

[32] E. M. Stein and R. Shakarchi. *Real analysis*. Princeton Lectures in Analysis, III. Princeton University Press, Princeton, NJ, 2005. Measure theory, integration, and Hilbert spaces.

[33] A. W. van der Vaart. *Asymptotic statistics*, volume 3 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge University Press, Cambridge, 1998.

[34] A. W. van der Vaart. Semiparametric statistics. In *Lectures on probability theory and statistics (Saint-Flour, 1999)*, volume 1781 of *Lecture Notes in Math.*, pages 331–457. Springer, Berlin, 2002.

[35] E. Vernet. Posterior consistency for nonparametric Hidden Markov Models with finite state space. *Electronic Journal of Statistics*, 9:717–752, 2015.