Clustering and Model Selection via Penalized Likelihood for Different-sized Categorical Data Vectors

Esther Derman, Erwan Le Pennec

▶ **To cite this version:**

Clustering and Model Selection via Penalized Likelihood for Different-sized Categorical Data Vectors

Esther Derman[*1] and Erwan Le Pennec[†1]

[1]CMAP, Ecole polytechnique, CNRS, Université Paris-Saclay, 91128, Palaiseau, France.

Abstract

In this study, we consider unsupervised clustering of categorical vectors that can be of different size using mixture. We use likelihood maximization to estimate the parameters of the underlying mixture model and a penalization technique to select the number of mixture components. Regardless of the true distribution that generated the data, we show that an explicit penalty, known up to a multiplicative constant, leads to a non-asymptotic oracle inequality with the Kullback-Leibler divergence on the two sides of the inequality. This theoretical result is illustrated by a document clustering application. To this aim a novel robust expectation-maximization algorithm is proposed to estimate the mixture parameters that best represent the different topics. Slope heuristics are used to calibrate the penalty and to select a number of clusters.

Keywords: Document clustering; Expectation-maximization algorithm; Multinomial mixture; Model selection; Penalized likelihood; Slope heuristics.

1 Introduction

1.1 Motivation

This study explores unsupervised clustering and model selection for multidimensional categorical data. Our basic block will be observations X_l of n_l independent and identically distributed instances of a categorical variable. We will assume we observe L independent such random vectors $(X_1, \ldots, X_L)$, that they share the same categories but not necessarily the same distribution and the same size. Such an observation is equivalent to a vector of L independent multinomial random variables, i.e. L barplots, with the same number of modalities but not necessarily the same distribution and the same size. Our objective is to cluster these observations, or the corresponding barplots, according to their similarities.

Clustering of categorical data is used in many fields such as social sciences, health, genetics, text analysis, etc. Several clustering techniques exist in the literature such as K-means or hierarchical clustering to name a few, we have chosen to use a mixture model to cluster the data. The first challenge encountered in our framework is that our observations X_ls are not of the same size. In other words, all the barplots do not necessarily represent the same number

---

[*]esther.boccara@polytechnique.edu

[†]erwan.le-pennec@polytechnique.edu

of instances. To our knowledge, this particularity regarding clustering has not yet been introduced in the literature. Gassiat et al. (2016) studied mixtures of multidimensional vectors in the case of independent and continuous marginal densities where they assume every observation has the same number of coordinates. In our framework, the X_ls may be of different dimensions besides being categorical, although a generalization to continuous variables should lead to similar properties. The most comparable work into this direction is the one of Matias et al. (2017) who model interaction events between individuals assuming they are clustered according to their pairwise interactions. They approximate interaction intensities by piecewise constant functions and penalize the resulting likelihood to do model selection.

A second challenge when dealing with our mixture model is the classical problem of choosing a proper number of clusters. An expectation-maximization algorithm is commonly used to estimate the best parameters in a given model (McLachlan and Peel, 2000), but there is no direct method to do model selection. Several techniques have been studied to select the number of clusters efficiently. Silvestre et al. (2014) propose a data-driven criterion for choosing the optimal number of clusters for categorical data. Based on a minimum message length criterion, they incorporate a model selection process inside the iterations of an expectation-maximization algorithm. Thanks to this method, they optimize time-consuming computation and avoid several runs of the same algorithm. On the same vein, Rigouste et al. (2007) estimate parameters of a multinomial mixture model for document clustering in a Bayesian framework and consider the model selection problem of choosing a dictionary that best discriminates the documents. Based on a text corpus, they alleviate the problem of high variability of the estimates in high dimension by deleting the most rare words from the dictionary. Using a categorical dataset of low back pain symptoms, Fop et al. (2017) address a Bayes factor criterion to select the most discriminative symptoms and cluster patients according to their symptoms' similarities.

We consider a penalized maximum likelihood approach where the penalty addresses the problem of simultaneously selecting the best multinomial mixture within a collection and the best group assignment of the vectors underlying the selected mixture. A classical trade-off between bias and variance naturally appears, which depends on this twofold complexity. Our model selection criterion takes these two steps into account. We show that it estimates mixture parameters efficiently and that vectors are robustly assigned to a proper cluster whatever the size of the data.

1.2 Results

The main result of this paper is a non-asymptotic oracle inequality that gives a sufficient condition on the penalty such that our estimator performs almost as well as the best one. Non-asymptotic oracle inequalities for categorical observations have already been introduced in Toussile and Gassiat (2009) and Bontemps and Toussile (2013) in the context of genomics, where the authors cluster a population into subpopulations according to their alleles' categories. The observations studied in their analysis are considered as independent and identically distributed vectors that are equally informative and of the same length as in our setting, they are conditionally independent and the information our data give depends on the length of the observed variables. Moreover, we aditionaly take into account cluster assignment of the observations when evaluating the risk of our estimation.

Our theoretical work is built on Massart's methodology (Massart, 2007) to compute the penalty function. Technical proofs are also inspired by the technical report of Cohen and Le Pennec (2011) (see also Cohen and Le Pennec (2012)) in which the authors use a penalized model selection technique with a maximum likelihood approach for conditional density estimation in a random design setting. Assuming we know which of the L density laws generated each obser-

vation, we place ourselves in a fixed design setting with deterministic covariates, although our result can easily be generalized to random covariates cases. Under mild assumptions on model structure, the non-asymptotic oracle inequality we address keeps the form of a Kullback-Leibler risk on the left side of the inequality without passing by a weaker one as it was the case in previous works (see Massart (2007), Maugis and Michel (2011), Cohen and Le Pennec (2012), Bontemps and Toussile (2013) and Montuelle and Le Pennec (2014)).

This theoretical study is applied to document clustering of NIPS conferences from 1987 to 2015. In order to estimate the parameters, we use a robust expectation-maximization algorithm that avoids classical numerical issues of local maxima and convergence to space boundaries. Moreover, since our theoretical penalty is known up to a multiplicative constant, we calibrate an optimal one by using slope heuristics and select the best model accordingly. The resulting visualization of different topics result to our intuitive a priori on their characteristics and effectively describes an evolution of machine learning overtime.

2 Model Selection Criterion

2.1 Framework and notations

In our model, we observe a family of L independent random vectors $(X_1, \ldots, X_L)$ where each X_l represents n_l independent and identically distributed instances X_l^i of a random variable having s_l as a true categorical density distribution with respect to the counting measure over B modalities, i.e.

$$pr\left((X_1, \ldots, X_L) = (x_1, \ldots, x_L)\right) = \prod_{l=1}^{L} pr\left((X_l^1, \ldots, X_l^{n_l}) = (x_l^1, \ldots, x_l^{n_l})\right)$$
$$= \prod_{l=1}^{L} \prod_{i=1}^{n_l} s_l(x_l^i).$$

This corresponds to the observation of $n = \sum_{l=1}^{L} n_l$ independent but only groupwise identical random variables. Assume, for a moment, that those L vectors can further be regrouped in K groups sharing the same density distribution. A natural model for such a situation is a mixture model in which the latent group comes from a K-dimensional multinomial variable Z that takes values in the set $\{1, \ldots, K\}$ of the different group labels and each group is characterized by its own density distribution f_k. Let $\pi_k = pr(Z = k)$. The distribution of Z is determined by the vector $\pi = (\pi_1, \ldots, \pi_K)$ which belongs to the $(K-1)$-dimensional simplex $\mathbb{S}_{K-1}$ and the density distribution becomes

$$pr\left((X_1, \ldots, X_L) = (x_1, \ldots, x_L)\right) = \prod_{l=1}^{L} pr(X_l = x_l) = \prod_{l=1}^{L} \left(\sum_{k=1}^{K} \pi_k pr(X_l = x_l \mid Z_l = k)\right)$$
$$= \prod_{l=1}^{L} \left\{\sum_{k=1}^{K} \pi_k \left(\prod_{i=1}^{n_l} f_k(x_l^i)\right)\right\}.$$

It is then usual to assign to each vector X_l its group tanks to a maximum a posteriori principle:

$$k_l = \underset{k \in \{1, \ldots, K\}}{\operatorname{argmax}} \, pr(Z = k | X_l = x_l) = \operatorname{argmax} \left\{\pi_k \left(\prod_{i=1}^{n_l} f_k(x_l^i)\right)\right\}$$

3

We will never assume that such a model holds but use the corresponding density functions as approximations of the true one and infer a group for each vector X_l with the same maximum a posteriori principle. For any K, such a mixture density distribution is entirely specified by $\theta = (\pi, f) = ((\pi_1, \ldots, \pi_K), (f_1, \ldots, f_K))$ in $\Theta_K = \mathbb{S}_{K-1} \times \mathcal{F}_K$ with $\mathcal{F}_K$ referring to a set of K categorical functions that will be specified later on. With a slight abuse of notation, we will denote by $P_{(K,\theta)}(x)$ the likelihood of an observation

$$P_{(K,\theta)}(x) = \prod_{l=1}^{L} P_{(K,\theta)}(x_l) = \prod_{l=1}^{L} \left\{ \sum_{k=1}^{K} \pi_k \left(\prod_{i=1}^{n_l} f_k(x_l^i) \right) \right\}.$$

For a fixed K, the distribution density is chosen in the model $\mathcal{S}_K = \{P_{(K,\theta)} : \theta \in \Theta_K\}$, while K has to be chosen among the set of positive integers smaller than L. Our task is thus to estimate a good couple (K, θ).

By construction, f_k is such that the class of multinomial mixtures is identifiable as long as $B \geq 2K - 1$, meaning that it is fully determined by its parameters up to label switching, as shown in Elmore and Wang (2003).

2.2 Maximum likelihood estimation

We use a classical maximum likelihood estimator to obtain the mixture parameters and use those parameters to assign groups.

More precisely, we define as $\gamma_n(\pi, f)$ the empirical contrast:

$$\gamma_n(\pi, f) = \sum_{l=1}^{L} -\log \left(P_{(K,\theta)}(x_l) \right).$$

The estimated parameter denoted by $(\widehat{\pi}, \widehat{f})$ maximizes the log-likelihood of the observations, which is equivalent to minimizing the empirical contrast:

$$(\widehat{\pi}, \widehat{f}) = \underset{(\pi, f) \in \Theta_K}{\operatorname{argmin}} \gamma_n(\pi, f).$$

To avoid existence issue, we will work with an almost minimizer and define an η -log-likelihood minimizer as any $(\widehat{\pi}, \widehat{f})$ that satisfies:

$$\gamma_n(\widehat{\pi}, \widehat{f}) \leq \inf_{(\pi, f) \in \Theta_K} \gamma_n(\pi, f) + \eta, \quad \eta > 0. \tag{1}$$

Since maximum likelihood estimation assigns a zero probability to unobserved categories, the likelihood can be infinite in some cases. This drawback is classical in discrete settings. To avoid this issue, we make the following assumption:

Assumption 2.1 (Model Structure). *Any candidate density f satisfies $f \geq \epsilon_n = e^{-\tau_n}$.*

The value ϵ_n can typically be taken as $1/n$. Thus, $-\log(f) \leq \tau_n$, where $\tau_n = \log(n)$. This assumption is legitimate in a context of document clustering because categories correspond to words that appear in at least one text.

We can now define a set $\mathcal{F}_K$ of K density functions on B categories so that any candidate satisfies Assumption 2.1. This set $\mathcal{F}_K$ is a simple product of multinomial densities satisfying the assumption:

$$\mathcal{F}_K = \left\{ (f(1), \ldots, f(B)); \sum_{b=1}^{B} f(b) = 1, e^{-\tau_n} \leq f(b) \leq 1 \right\}^K = \mathcal{F}^K.$$

Once parameters are estimated, an observation x_l is assigned to the cluster it most likely belongs to. We use the maximum a posteriori method:

$$\widehat{k_l} = \underset{k \in [\![1..K]\!]}{\operatorname{argmax}} \left\{ \widehat{\pi}_k \left(\prod_{i=1}^{n_l} \widehat{f}_k(x_l^i) \right) \right\}.$$

Finally, we note $\widehat{f}_{\widehat{k_l}}$ the estimated distribution of observation x_l.

2.3 Single model risk bound

In this section, we prove that for a fixed K such a scheme allows controlling the error between the true density function by a multiple of the best possible mixture. We will measure the error with the Kullback-Leibler loss. Recall that Kullback-Leibler divergence is a natural quality measure to be considered in maximum likelihood approach. It is defined by

$$\mathbf{KL}(s,t) = \int \log\left(\frac{s}{t}\right) s d\lambda,$$

where λ is a known positive measure such that $sd\lambda$ is absolutely continuous with respect to $td\lambda$.

Our first result is an oracle inequality that upper bound the expectation of Kullback-Leibler divergence between our estimate and the truth by the best possible one in the model $\mathcal{S}_K$ up to a multiplicative factor and a *variance* term that measures the model complexity. Cohen and Le Pennec (2012) show that under some general assumption on this bracketing entropy, model complexity is proportional to the model dimension. In our case, as we have an assigment step, the dimensions of the considered models correspond to the number of mixture parameters plus the cluster assignment cost. The dimension of a mixture model with K categorical functions of $\mathcal{F}_K$ is $KB-1$. For sake of simplicity, we will rather use a *dimension* $D_K = KB$.

Theorem 2.1. *Consider the observed vectors $(x_1, \ldots, x_L)$ as above and denote by n the overall number of observations. Consider also one model $\mathcal{S}_K$ as defined above and assume it satisfies Assumption 2.1. Let $(\widehat{\pi}, \widehat{f})$ be an η-likelihood minimizer in $\mathcal{S}_K$ as defined in equation 1. Denote the corresponding cluster assignment for each vector $x_l = (x_l^1, \ldots, x_l^{n_l})$:*

$$\widehat{k_l} = \underset{k \in [\![1..K]\!]}{\operatorname{argmax}} \left\{ \widehat{\pi}_k \left(\prod_{i=1}^{n_l} \widehat{f}_k(x_l^i) \right) \right\}.$$

Then, for any $C_1 > 1$, there exist two constants λ_0 and C_2 depending only on C_1 such that the estimate $(\widehat{\pi}, \widehat{f})$ satisfies

$$E\left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}})\right] \leq C_1 \left\{ \inf_{\substack{f \in \mathcal{F}_K \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) + \lambda_0 \frac{L \log K + \mu_n D_K}{n} \right\} + \frac{C_2}{n} + \frac{\eta}{n},$$

where $\mu_n = 2(\log(2\tau_n)^{1/2} + \pi^{1/2})^2 + 1 + \log(n)$.

In previous works (see Massart (2007), Cohen and Le Pennec (2012), Bontemps and Toussile (2013) and Maugis and Michel (2011)), a lower divergence appears on the left side of the inequality, such as Hellinger distance or Jensen-Kullback-Leibler divergence. The main contribution of

our result is that the empirical risk of the estimated density is measured according to the same averaged Kullback-Leibler divergence as the risk model appearing on the right side of the inequality. The upper bound addressed here is sharper and explicits more clearly the bias underlying density estimation because risk functions are the same on the left and right side of the inequality. This is made possible thanks to Assumption 2.1, which enables to bound the moments of log-ratios of density distributions. A concentration inequality on the empirical Kullback-Leibler divergence can then be deduced. Without Assumption 2.1, density distributions must be weighted by the true ones in order to bound a log-ratio which leads to a weaker divergence than **KL** (Massart (2007) and Cohen and Le Pennec (2012)).

Moreover, this oracle inequality takes the two steps of the estimation into account, namely mixture parameters and clustering assignment of the observations. This explains the term in $L\log(K)$ appearing on the left side of the inequality. It measures the cost for assigning each observation to one cluster, as D_K measures the complexity of mixture models. The overall *variance* $\lambda_0 \left(L\log K + \mu_n D_K \right)$ clearly balances the bias term appearing in the inequality as the minimum of the divergence over the parameter set.

Theorem 2.1 turns out to be crucial in our analysis because the model complexity that appears explicitly is the one that we will use to define a suitable penalized criterion.

2.4 Model selection theorem

We focus now on the choice of K and prove that this can be done with a simple penalization of the likelihood. Indeed, for any K, the oracle inequality of Theorem 2.1 holds as soon as Assumption 2.1 holds. One of the models minimizes the right hand side of the inequality but there is no way of knowing which one without also knowing the true densities $s_1, \ldots, s_L$.

We propose thus a data-driven strategy to select K that performs almost as well as we had known the best K. We use a penalized criterion

$$\mathbf{crit}(K) = \min_{(\pi, f) \in \Theta_K} \gamma_n(\pi, f) + \mathbf{pen}(K),$$

where $\mathbf{pen} : K \to \mathbb{R}^+$ denotes the penalty function. Our analysis suggests the use of a penalty of the form $\mathbf{pen}(K) = \lambda_0(D_K + L\log(K))$, where D_K plays the role of mixture complexity and $L\log(K)$ corresponds to cluster assignment cost, similar to the quantity . This relates to classical penalties such as AIC $= D_K/n$ and BIC $= D_K \log(n)/(2n)$ that are proportional to model dimension, but such penalties require large sample sizes in order to be consistent. The penalty we propose is consistent with non-asymptotic sample size and performs well with respect to a theoretical loss function **KL**.

The main result of this study is the following theorem that compares the risk of the selected model with the risk of the unknown best model and shows that for a suitable choice of the constant in the penalty function, the estimated model still performs almost as well as the best one.

Theorem 2.2. *Consider the observed vectors $(x_1, \ldots, x_L)$ and denote by n the overall number of observations. Consider also the collection $(\mathcal{S}_K)_{K \in [\![1..L]\!]}$ of models satifying Assumption 2.1 defined above.*

Let **pen** *be a non-negative penalty function and $\hat{K}$ any η-minimizer of*

$$\mathbf{crit}(K) = \min_{(\pi, f) \in \Theta_K} \gamma_n(\pi, f) + \mathbf{pen}(K).$$

Let $(\widehat{\pi}, \widehat{f})$ be the corresponding η-likelihood minimizers in $\mathcal{S}_{\widehat{K}}$ and define the resulting cluster

assignment for each vector $x_l = (x_l^1, \ldots, x_l^{n_l})$:

$$\widehat{k_l} = \operatorname*{argmax}_{k \in [\![1..\widehat{K}]\!]} \left\{ \widehat{\pi_k} \left(\prod_{i=1}^{n_l} \widehat{f_k}(x_l^i) \right) \right\}.$$

Then, for any constant $C_1 > 1$, there exist two constants λ_0' and C_2 depending only on C_1 such that if the penalty function is defined as:

$$\mathbf{pen} : K \longmapsto \lambda_0' \left(\mu_n D_K + L \log(K) + K \log(2) \right)$$

with $\mu_n = 2(\log(2\tau_n)^{1/2} + \pi^{1/2})^2 + 1 + \log(n)$, then, whatever the underlying true densities $s_1, \ldots, s_L$,

$$E \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f_{\widehat{k_l}}}) \right) \leq C_1 \inf_K \left\{ \inf_{\substack{f \in \mathcal{F}_K \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) + \frac{\mathbf{pen}(K)}{n} \right\}$$
$$+ \frac{C_2}{n} + \frac{\eta}{n}.$$

The inequality looks similar to the one of Theorem 2.1 and indeed the two differences are the infimum over K, which was not present in the first theorem and the penalty function, which is equal to the *variance* term of the first theorem up to an additional term $K \log(2)$ appearing in the penalty function. This small correction is not intrinsic but is required to apply a union bound over all K in the proof. The main gain is the infimum over K that ensures that our estimate is almost as good as the one leading the best single model oracle inequality. Remark that the infimum is not restricted to $K \leq L$ because any solution with $K > L$ can advantageously be replaced by one with L groups.

3   Model Selection on a text corpus

3.1   NIPS Conference Papers

In this section, we illustrate our model selection method by an application to document clustering. Note that we will note use the constant appearing in Theorem 2.2 but only the shape of the penalty to select the number of cluster. Note that the code and the datasets used are available at https://github.com/EstherBoc/SourceCodeDocumentClustering.

The dataset "NIPS Conference Papers 1987-2015 Data Set" contains distribution of words used in NIPS conference papers published from 1987 to 2015 (Perrone et al., 2016). Based on a dictionary of 11463 unique words that appear in 5811 conference papers, the dataset is represented as a matrix with 11463 rows and 5811 columns. Each row represents the number of occurrences of the corresponding word in each document. Problems of dimensionality and low performance are avoided by removing rare words (Rigouste et al., 2007).

To allow for better distinction between clusters, words that appear in more that 80% of the documents are removed. Only $B = 300$ most frequent words are finally considered in the remaining counting matrix. Documents that are empty with respect to this reduced dictionary are removed as well, which leads to $L = 5804$ text documents in the exploited dataset. In our analysis, the vocabulary is constructed a priori from the most frequent and discriminative words and stays fixed. In the supplementary material, model selection with varying dictionary size is addressed in a case where the set of words is assumed to be ordered. In practice, the whole

vocabulary is considered there so that different models can be compared to each other. However, the proportions of the first B words can be different, whereas all the others are uniform on the remaining probability.

The objective is to calibrate the penalty and select the best mixture model, that is, a number of clusters $\widehat{K}$ that has a good bias-variance tradeoff: as too many components may result to an over-fitting, a mixture with too few components may be too restrictive to approximate well the mixture underlying the data. Although time structure of the corpus is ignored in our model selection, it is analyzed in Section 3.4. One could also consider the corpus as a spatial mixture model with mixture proportions modeling time structure in the articles (Cohen and Le Pennec, 2014).

3.2   Clustering algorithm

Mixture parameters are estimated thanks to the expectation-maximization algorithm. It heavily depends on its initial parameters, so the log-likelihood often converges to a local maximum. We ran 500 short expectation-maximization algorithms from random initializing parameters to analyze the sensitivity of the log-likelihood with respect to the initialization. Figure 1 shows that in practice, despite high model dimension (9000 in this case), the log-likelihood keeps the same order of magnitude with high probability. Although some runs perform poorly, a high majority of them result to the same order of log-likelihood. Thus, a natural way of avoiding local maxima is to run several short expectation-maximization algorithms (with 15 iterations by default) from randomly chosen initializing parameters and run a long one from the most performing parameter in terms of likelihood (Toussile and Gassiat, 2009).

The expectation-maximization algorithm has another issue: it may converge to the boundary of the parameter space, in which case the output estimates are unstable and some mixture proportions become under-represented (Figueiredo and Jain, 2002). This especially occurs in high dimension. B. Zhang, C. Zhang and Xing Yi Zhang et al. (2004) proposed a *split and merge* method that divides or merges clusters according to their entropy information but despite random initializations, splitting may decrease stability of the estimates.

It has been developed in Figueiredo and Jain (2002) that component annihilation leads to more robust results. This observation introduced the expectation-maximization-minimum-message-length algorithm (EM-MML) that starts from a high number of clusters (Silvestre et al., 2014). At each iteration, mixture proportions are penalized, and if one of them goes under zero, the corresponding cluster is annihilated. The procedure continues until an optimal number of clusters is reached. By simultaneously dealing with the number of components and the estimates, this technique avoids time-consuming computation. However, besides keeping some initialization dependency, the EM-MML penalizes parameters according to a BIC-type function that leads to non robust results in a non asymptotic framework. Further work of Yan, Lai and Lin Yang et al. (2012) considers all the data set as an initializing parameter. The algorithm progressively annihilates clusters that give decreasing information with respect to Shannon entropy. Although less sensitive to initialization, this technique is time-consuming and based on a penalty that does not take data dimension into account.

We address a robust expectation-maximization algorithm (EM) that takes these issues into account. It proceeds as follows. For each model, clusters with low proportion estimates are annihilated while other parameters remain the same. After renormalization, additional EM are initialized from the remaining parameters and run until no low proportion mixture is met. This prevents the EM algorithm from approaching the boundary of the parameter space. It thus leads to more robust results in a sense that a 100 component mixture often performs as well as a well-balanced 50 component mixture. Algorithm 1 shows the detailed pseudocode of the algorithm.
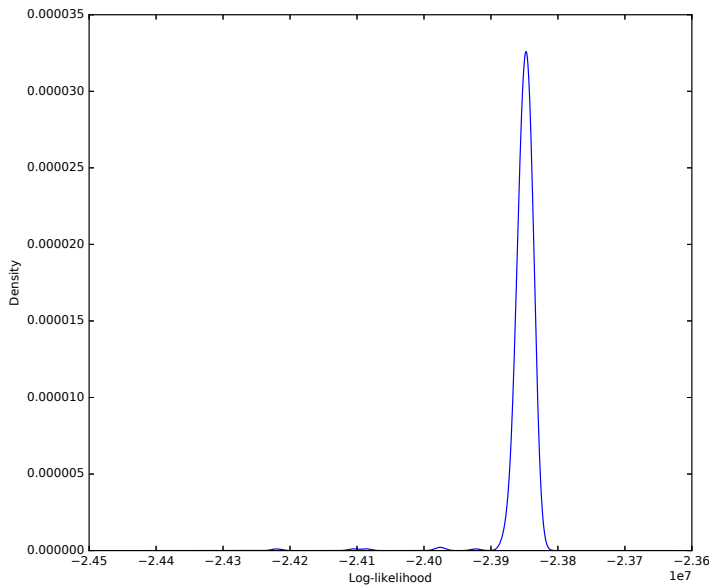
Figure 1: Kernel Density Estimation of log-likelihood distribution
based on 500 random initializing parameters after 15 iterations of
EM (K = 30, B = 300)

We remarked that despite annihilation, the number of remaining clusters still increases with respect to the number of components initially input. Several models with different dimensions can thus be explored for penalty calibration.

3.3 Slope heuristics

Besides parameter estimation, we want to calibrate our theoretical penalty for selecting a model according to the penalized criterion thus obtained.

Penalized log-likelihood enables to select a model that has the right bias-variance tradeoff. Unlike AIC and BIC functions, the penalty introduced in Theorem 2.2 is adapted to a non-asymptotic framework. However, it is defined up to an unknown multiplicative constant. In fact, any greater penalty satisfies the oracle inequality but may lead to a model with high bias. Through slope heuristics, Baudry et al. (2010) provide a practical technique to calibrate the constant that leads to an optimal penalty. It relies on the fact that the empirical contrast of the estimated parameter is linear with respect to the model dimension when the model is complex enough. Indeed, for most complex models, the bias term becomes stable so the risk behaves as the variance term and becomes linear with respect to the dimension. Denoting by $\lambda_{\min}$ the slope of the linear part of the empirical contrast, the optimal penalty function is

$$\mathbf{pen}_{\mathrm{opt}}(K) = 2\lambda_{\min}D_K,$$

with D_K the model dimension. The derivation of this formula is detailed in Maugis and Michel (2008) and Baudry et al. (2010).

9

Algorithm 1 Robust EM algorithm

Input: $k_{\max}$
Output: $k_{\mathrm{opt}}, \widehat{\pi}^{(t)}_{i_{\max}}, \widehat{f}^{(t)}_{i_{\max}}, \widehat{\mathcal{L}}^{(t)}_{i_{\max}}$

1: **procedure** ROBUSTEM
2:      $t \leftarrow 0$
3:      $k_{\mathrm{current}} \leftarrow k_{\max}$
4:      $\pi_{\mathrm{threshold}} \leftarrow \frac{1}{100 \cdot k_{\mathrm{current}}}$
5:      $\mathcal{L}^{(0)} \leftarrow -\infty$
6:      **for** i = 1 to 15 **do**
7:          $\pi^{(0)}_i, f^{(0)}_i \leftarrow$ INITIALIZE(π, f)
8:          $\widehat{\pi}^{(10)}_i, \widehat{f}^{(10)}_i, \widehat{\mathcal{L}}^{(10)}_i \leftarrow$ SHORTEM$(k_{\mathrm{current}}, \pi^{(0)}_i, f^{(0)}_i)$
9:      $t \leftarrow 10$
10:     $i_{\max} \leftarrow \mathrm{argmax}_i\, \widehat{\mathcal{L}}^{(10)}_i$
11:     **while** $\widehat{\pi}^{(t)}_{i_{\max}} < \pi_{\mathrm{threshold}}$ **do**
12:        $k_{\mathrm{current}} \leftarrow \sharp\{k : \widehat{\pi}^{(t)}_{i_{\max}}(k) \geq \pi_{\mathrm{threshold}}\}$
13:        $\pi_{\mathrm{threshold}} \leftarrow \frac{1}{100 \cdot k_{\mathrm{current}}}$
14:        $\widehat{\pi}^{(t+u)}_{i_{\max}}, \widehat{f}^{(t+u)}_{i_{\max}}, \widehat{\mathcal{L}}^{(t+u)}_{i_{\max}} \leftarrow$ EM$(k_{\mathrm{current}}, \pi^{(t)}_{i_{\max}}, f^{(t)}_{i_{\max}})$
15:        $t \leftarrow t + u$
16:     $k_{\mathrm{opt}} \leftarrow k_{\mathrm{current}}$

We use slope heuristics to calibrate penalty for NIPS data. Practical methodologies and visualizing graphics were implemented thanks to the R package **capushe**. The maximum number of clusters in the collection is set to $K_{\max} = 100$. Linear regression is operated with different number of points from which we obtain different slope coefficients. The technique used to deduce the minimal constant is described in Baudry et al. (2010). Figure 2 shows linear regression of the log-likelihood with respect to the selected number of points. It gives a slope of $\widehat{\lambda}_{\min} \approx 15$. Finally, the model that minimizes the resulting criterion corresponds to $\widehat{K} = 31$ clusters.

3.4 Back-mapping of NIPS topics

We used the selected model of $\widehat{K} = 31$ clusters and its corresponding parameter estimates to analyze the evolution of some representative NIPS topics overtime. Topics are characterized by their distribution of words. Thus, each cluster can be labeled by looking at the most likely words.

At each year, the average posterior probability of the corresponding articles over every cluster gives a year-scaled evolution of topics from 1987 to 2015. Figures 3 and 4 show time evolution of some representative topics and their most likely words.

The word-cloud in Fig. 3b represents the categorical distribution of words. The bigger is a word, the higher its probability is. Clusters become easily interpretable thanks to this representation. Based on its biggest words, we can deduce that Cluster 30 corresponds to reinforcement learning. This topic appeared in the late 80s and knew some ups and downs from then on (Fig. 3). In Fig. 4, only the most probable words are represented. Figure 4a indicates that the corresponding topic is Bayesian inference and that it constantly increased overtime. So did the popularity of optimization problems (Fig. 4d) and sparsity (Fig. 4f). Classification and support vector machines in particular got most popular in the late 90s before going through a decline (Fig. 4c). Clustering methods knew the same trend, except that their popularity decreased later
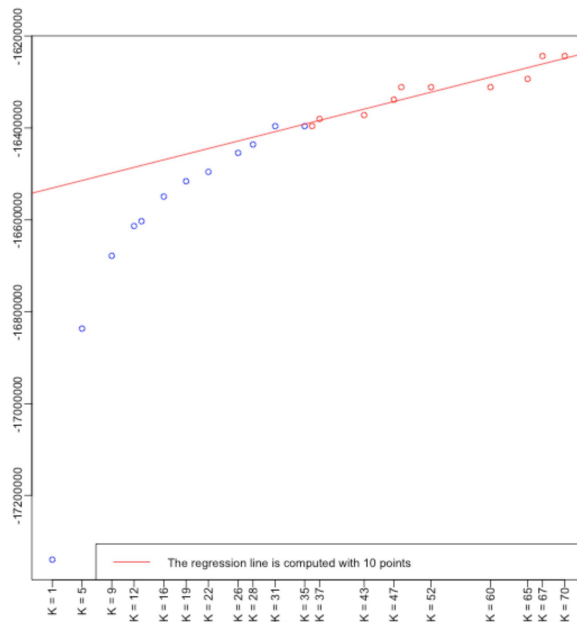
Figure 2: Slope heuristics on NIPS data

on. A constant decline of popularity can also be noticed regarding neural networks architecture (Fig. 4e). This could be explained by the fact that the approach and vocabulary on neural networks clearly evolved since 2011, leading to deep learning and its new vocabulary. These trends give a clear overview of how NIPS conferences evolved since 1987 and are similar to those obtained by Perrone et al. (2016) who use Poisson random fields to model the dynamic of the documents' features.

4 Discussion

In this paper, we have considered the problem of estimating several distributions of different-sized categorical variables. We have been able to prove that the number of components can be estimated almost optimally by a penalized maximum likelihood principle. The theoretical analysis has been completed by a numerical illustration on text clustering in which we have proposed a better initialization scheme for the expectation-maximization algorithm used to estimate the parameters as well as a calibration of the penalty.

The results obtained with categorical variables can directly be extended to bounded continuous ones. Consider a family of L independent random vectors $(X_1, \ldots, X_L)$ where each X_l represents n_l independent and identically distributed instances of a random variable that has s_l as a true continuous density distribution. Under the same theoretical assumptions as those addressed in our work, the same kind of results can be obtained if we approximate each density by lower bounded piecewise constant functions. The maximum likelihood estimator on a model of densities that are piecewise constant on a given partition would then correspond to a histogram estimator and the objective would be to estimate K clusters and the best corresponding histograms. In Massart (2007), the problem of estimating the best partition when observing independent and identically distributed variables is addressed. The theoretical results that are

obtained there could be extended to conditionally independent variables that are clustered according to their densities' similarities. However, the lower-bound on the model densities as stated in Assumption 2.1 can be no longer applicable if densities can take extreme values (Gaussian density for example), because these cannot be controlled in the estimation.

(a) Time evolution of reinforcement learning

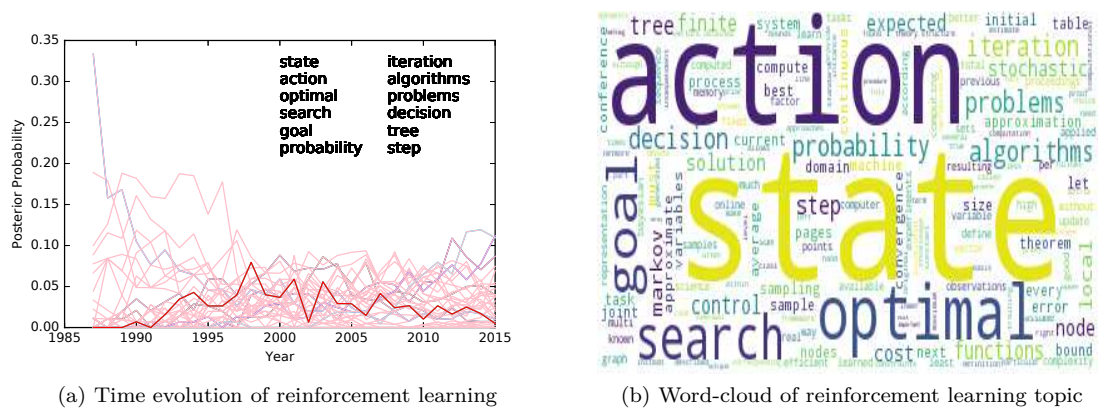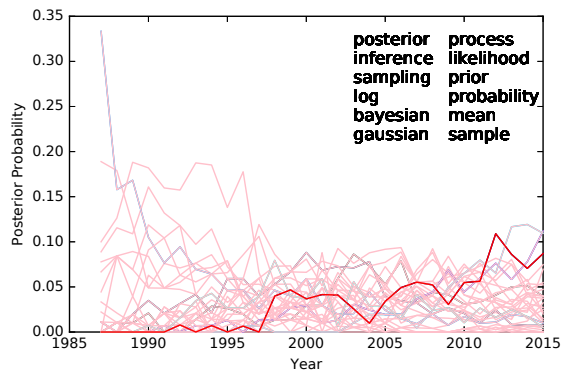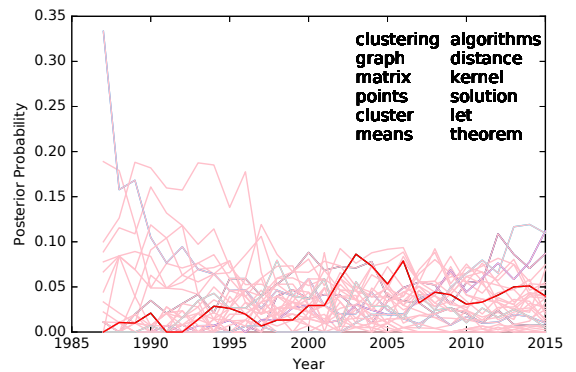

(b) Word-cloud of reinforcement learning topic

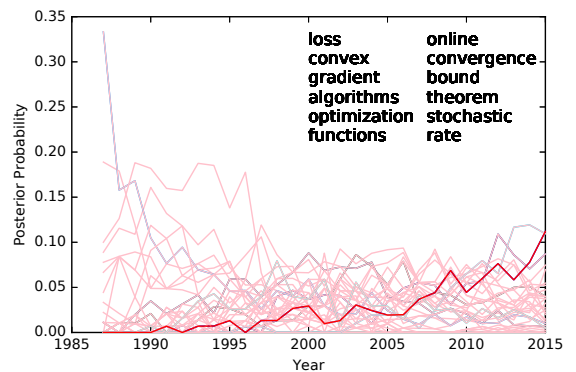Figure 3: Time evolution and word-cloud of reinforcement learning
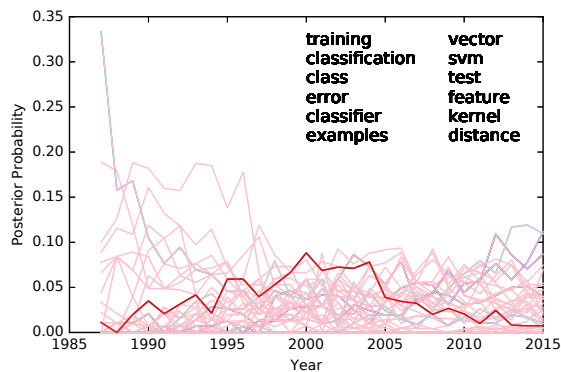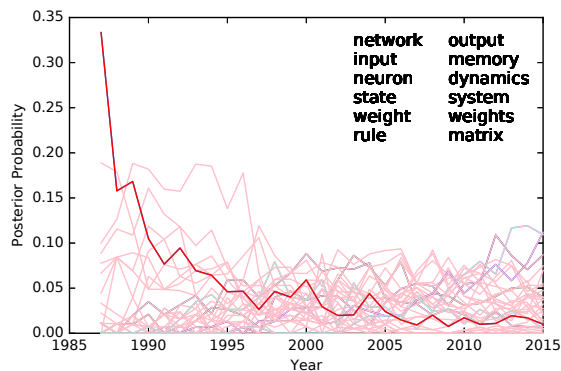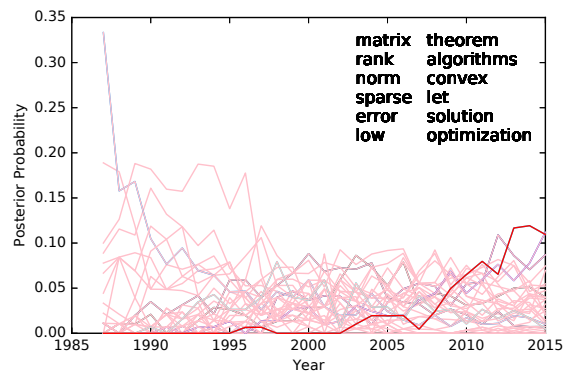
Figure 4: Comparative evolution of some representative NIPS topics from 1987 to 2015 with the 12 most likely words. The red line corresponds to the labeled topic and the transparent one to the other topics.

13

Bibliography

Baudry, J.-P., Maugis, C., and Michel, B. (2010). Slope heuristics: Overview and implementation. *Statist. Comput.*, 22:455–470.

Bontemps, D. and Toussile, W. (2013). Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics*, 7:2344–2371.

Cohen, S. and Le Pennec, E. (2012). Partition-based conditional density estimation. *ESAIM: P & S*, 17:672–697.

Cohen, S. X. and Le Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *INRIA Technical Report.*

Cohen, S. X. and Le Pennec, E. (2014). Unsupervised segmentation of spectral images with a spatialized gaussian mixture modeland model selection. *Oil and Gas Science and Technology.*

Elmore, R. and Wang, S. (2003). Identifiability and estimation in finite mixture models with multinomial components. *Technical Report.*

Figueiredo, M. A. and Jain, A. K. (2002). Unsupervised learning of finite mixture models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(3).

Fop, M., Smartzand, K. M., and Murphy, T. B. (2017). Variable selection for latent class analysis with application to low back diagnosis. In press.

Gassiat, E., Rousseau, J., and Vernet, E. (2016). Efficient semiparametric estimation and model selection for multidimensional mixtures. (Available at ArXiv:1607.05430).

Genovese, C. R. and Wasserman, L. (2000). Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 28:1105–1127.

Massart, P. (2007). *Concentration Inequalities and Model Selection*, volume 1896. Springer.

Matias, C., Rebafka, T., and Villers, F. (2017). Estimation and clustering in a semiparametric poisson process stochastic block model for longitudinal networks. (Available at ArXiv:1512.07075).

Maugis, C. and Michel, B. (2008). Slope heuristics for variable selection and clustering via gaussian mixtures. *RR-INRIA*, 7223.

Maugis, C. and Michel, B. (2011). A non asymptotic penalized criterion for gaussian mixture model selection. *ESAIM: P&S (2011)*, 15.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Probability and Statistics.

Meynet, C. (2012). *Variable selection in model-based clustering for high-dimensional data.* PhD thesis, Université Paris-Sud XI.

Montuelle, L. and Le Pennec, E. (2014). Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics.*

Perrone, V., Jenkins, P. A., Spanó, D., and Teh, Y. W. (2016). Poisson random fields for dynamic feature models. *arXiv preprint arXiv:1611.07460.*

Rigouste, L., Cappé, O., and Yvon, F. (2007). Inference and evaluation of the multinomial mixture model for text clustering. *Information Processing and Management*, 43(5):1260–1280.

Silvestre, C., Cardoso, M. G. M. S., and Figueiredo, M. A. T. (2014). Identifying the number of clusters in discrete mixture models. (Available at ArXiv:1409.7419).

Toussile, W. and Gassiat, E. (2009). Variable selection in model-based clustering using multilocus genotype data. *Adv. Data Anal. Classif*, 3:109–134.

Yang, M.-S., Lai, C.-Y., and Lin, C.-Y. (2012). A robust em clustering algorithm for gaussian mixture models. *Pattern Recognition*, 45(11):3950–3961.

Zhang, B., Zhang, C., and Yi, X. (2004). Competitive em algorithm for finite mixture models. *Pattern Recognition*, 37(1):131–144.

A   Supplementary material

A.1   Preliminary

In this section, Theorems 2.1 and 2.2 are proved in a more general setting where some assumptions are relaxed. Notation is generalized so that each model is indexed by m rather than K. Thus, a parameter set of a model is denoted by Θ_m, the corresponding model by $\mathcal{S}_m$ and $\mathcal{F}_m$ is the associated K-uplet of mixture distributions. An element $f \in \mathcal{F}_m$ can also be described as a family $(f_1, \ldots, f_K)$ of K categorical probability distributions on B categories. Denoting by s_l the true density of observation l, we assume the structure of the models to be such that:

Assumption A.1 (**Model structure**). *Any density functions $f, f' \in \mathcal{F}_m$ taken in a model satisfy $e^{-\tau_n} \le f/f' \le e^{\tau_n}$ and $e^{-\tau_n} \le f/s_l$, $l \in \{1, \ldots, L\}$, s_l being the true density of observation l.*

This assumption implies Assumption 2.1 as formulated inside the article. Moreover, the second part of the assumption is made on the model density rather than on the true one which is unknown. Our main result adresses an oracle inequality that links some averaged Kullback-Leibler divergence **KL** of the selected estimator to the averaged **KL** between the true densities and every model within the collection. This inequality leads to some penalty function which heavily relies on a notion of bracketing entropy. A bracket $[f_k^-, f_k^+]$ is a pair of real-valued functions such that for all $x \in \mathcal{X}$, $f_k^-(x) \le f_k^+(x)$. A density function f_k is said to belong to the bracket $[f_k^-, f_k^+]$ if $f_k^-(x) \le f_k(x) \le f_k^+(x)$ for all $x \in \mathcal{X}$. Take $f^-, f^+ \in \mathcal{F}_m$ such that for all $k = 1, \ldots, K$, $[f_k^-, f_k^+]$ forms a bracket. Fix $(k_1, \ldots, k_L)$ a cluster assignment of the observations. We define the width of such a family of brackets as follows:

$$\mathbf{a}(f^-, f^+) = \frac{1}{n} \sum_{l=1}^{L} n_l E_{s_l} \left[| \log \left(\frac{f_{k_l}^-}{f_{k_l}^+} \right) |^2 \right].$$

The bracketing entropy $H_{[.], \mathbf{a}}(\delta, \mathcal{F}_m)$ of a set of functions $\mathcal{F}_m$ is defined as the logarithm of the minimum number of brackets of width smaller than δ such that every function of $\mathcal{F}_m$ belongs to one of these brackets. The model complexity that will be considered rather depends on the localized models, which in our framework can be written as:

$$\mathcal{F}_m(\tilde{f}, \sigma) = \left\{ f \in \mathcal{F}_m \mid \mathbf{a}(\tilde{f}, f) \le \sigma^2 \right\}.$$

We also impose a structural assumption on the localized models:

Assumption A.2. *There exists a real-valued function ϕ_m on $[0, +\infty)$ such that ϕ_m is non-decreasing, the mapping $\delta \mapsto \frac{1}{\delta} \phi_m(\delta)$ is non-increasing on $(0, +\infty)$ and for every $\sigma \ge 0$ and every $f \in \mathcal{F}_m$,*

$$\int_0^\sigma H_{[.], \mathbf{a}}(\delta, \mathcal{F}_m(f, \sigma))^{1/2} d\delta \le \phi_m(\sigma).$$

The use of divergence $\mathbf{a}$ instead of **KL** has some benefit, as it allows to take advantage of the metric entropy of the models and deduce the bracketing entropy. It is smaller, up to a certain constant, than the **KL** divergence:

Proposition A.1. *Assume that Assumption A.1 is satisfied. Then*

$$\frac{1}{n} \sum_{l=1}^{L} n_l E_{s_l} \left[| \log \left(\frac{f_{k_l}}{s_l} \right) |^2 \right] \le \frac{\tau_n^2}{e^{-\tau_n} + \tau_n - 1} \frac{1}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(f_{k_l}, s_l).$$

Proof. The proof can be deduced from Meynet's result (Meynet, 2012) that states the following.

Lemma A.1. *Let P and Q be two probability measures with $P \ll Q$. Assume there exists $\tau > 0$ such that $\log(\|\frac{dP}{dQ}\|_\infty) \leq \tau$. Then*

$$\int \left(\log \frac{dP}{dQ}\right)^2 dP \leq \frac{\tau^2}{e^{-\tau} + \tau - 1} \mathbf{KL}(P, Q).$$

By taking $\tau = \tau_n$, $dP = f_{k_l} d\mu$ and $dQ = s_l d\mu$, one can deduce the result. $\square$

In order to avoid measurability issues, we also impose a separability condition on the models:

Assumption A.3 (Separability). *There exists a countable subset $\mathcal{F}'_m$ of $\mathcal{F}_m$ and a set $\mathcal{X}'_m$ of measure $\lambda(\mathcal{X} \setminus \mathcal{X}'_m) = 0$ such that for every $f \in \mathcal{F}_m$, there exists a sequence $(f_j)_{j \geq 1}$ of elements of $\mathcal{F}'_m$ such that for every $x \in \mathcal{X}'_m$, $\log(f_j(x))$ goes to $\log(f(x))$ as j goes to infinity.*

A.2 Single Model Risk Bound

Theorem A.1. *Let $(x_1, \ldots, x_L)$ be L observations of independent random vectors $(X_1, \ldots, X_L)$ where each X_l consists of n_l independent and identically distributed instances of a multinomial vector that has s_l as a true density with respect to some known positive measure. Assume $\mathcal{S}_m$ is a model for which Assumptions A.1, A.2 and A.3 hold. Let $(\widehat{\pi}, \widehat{f}) \in \Theta_m$ be an η -log-likelihood minimizer in $\mathcal{S}_m$:*

$$\gamma_n(\widehat{\pi}, \widehat{f}) \leq \inf_{(\pi, f) \in \Theta_m} \gamma_n(\pi, f) + \eta,$$

and define its corresponding cluster assignment for each observation x_l:

$$\widehat{k}_l = \underset{k \in [\![1..K]\!]}{\operatorname{argmax}} \left\{ \widehat{\pi}_k \left(\prod_{i=1}^{n_l} \widehat{f}_k(x_l^i) \right) \right\}.$$

Then, for any $C_1 > 1$, there exists a constant C_2 depending only on C_1 such that, for $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = n^{1/2}\sigma$, the estimate $(\widehat{\pi}, \widehat{f})$ satisfies

$$E\left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f_{\widehat{k}_l}})\right] \leq C_1 \left\{ \inf_{\substack{f \in \mathcal{F}_m \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) + \frac{(2 + \kappa_0)L \log K + \kappa_0 \mathfrak{D}_m}{n} + \frac{\eta}{n} \right\}$$
$$+ \frac{C_2}{n}.$$

Proof. By definition of $\widehat{k}_l$ and using the fact that all mixture probabilities π_k are less than 1, we can write:

$$\eta + \inf_{(\pi, f) \in \Theta_m} \gamma_n(\pi, f) \geq \sum_{l=1}^{L} -\log \left(K\widehat{\pi}_{\widehat{k}_l} \left[\prod_{i=1}^{n_l} \left(\frac{\widehat{f_{\widehat{k}_l}}}{s_l} \right)(x_l^i) \right] \right)$$
$$\geq -L\log(K) - \sum_{l=1}^{L} \sum_{i=1}^{n_l} \log \left(\left(\frac{\widehat{f_{\widehat{k}_l}}}{s_l} \right)(x_l^i) \right). \tag{2}$$

On the other hand, for any cluster assignment $(k_1, \ldots, k_L)$ of the observations we have

$$\gamma_n(\pi, f) \leq \sum_{l=1}^{L} -\log(\pi_{k_l}) + \sum_{l=1}^{L} -\log\left(\prod_{i=1}^{n_l}\left(\frac{f_{k_l}}{s_l}\right)(x_l^i)\right)$$

because $-\log$ is non-increasing. Therefore, by inequality (2),

$$\inf_{(\pi,f)\in\Theta_m} \gamma_n(\pi, f) \leq \inf_{(k_l)_l\in[\![1..K]\!]^L} \left[\inf_{(\pi,f)\in\Theta_m} \left\{\sum_{l=1}^{L} -\log(\pi_{k_l}) + \sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{f_{k_l}}{s_l}\right)(x_l^i)\right)\right\}\right],$$

which leads to

$$\sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{\widehat{f_{\widehat{k_l}}}}{s_l}\right)(x_l^i)\right) \leq L\log(K) + \eta$$

$$+ \inf_{(k_l)_l\in[\![1..K]\!]^L}\left[\inf_{(\pi,f)\in\Theta_m}\left\{\sum_{l=1}^{L} -\log(\pi_{k_l})\right.\right. \quad (3)$$

$$\left.\left.+ \sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{f_{k_l}}{s_l}\right)(x_l^i)\right)\right\}\right].$$

We define by $\bar{f} = (\bar{f}_1, \ldots, \bar{f}_K)$ a family of densities that satisfy for all $\delta > 0$ and all cluster assignment $(k_1, \ldots, k_L)$:

$$\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \bar{f}_{k_l}) \leq \inf_{f\in\mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) + \delta. \quad (4)$$

For all $l \in \{1, \ldots, L\}$ and any density f_{k_l} at l, we denote the empirical Kullback-Leibler divergence $\mathbf{kl}(f_{k_l})$ by:

$$\mathbf{kl}(f_{k_l}) = \sum_{i=1}^{n_l} -\log\left(\frac{f_{k_l}}{s_l}\right).$$

We have $E\left[\sum_{l=1}^{L}\mathbf{kl}(f_{k_l})(X_l)\right] = \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l})$. Eventually, define:

$$\nu_l(f_{k_l}) = \mathbf{kl}(f_{k_l})(x_l) - E_{s_l}\left[\mathbf{kl}(f_{k_l})(X_l)\right] \quad (5)$$

the centered version of the empirical $\mathbf{KL}$-divergence. Thanks to equation (3) and using the

definition of $\bar{f}$ in (4), we can write:

$$
\sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k_l}}) + \sum_{l=1}^{L} E_{s_l} \left[\mathbf{kl}(\widehat{f}_{\widehat{k_l}})(X_l) \right] \leq L \log K + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \inf_{\pi \in \mathbb{S}_{k-1}} \left(\sum_{l=1}^{L} - \log(\pi_{k_l}) \right) \right.
$$

$$
+ \sum_{l=1}^{L} \left(\sum_{i=1}^{n_l} - \log \left(\left(\frac{\bar{f}_{k_l}}{s_l} \right) (x_l^i) \right) - E_{s_l} \left[\sum_{i=1}^{n_l} \log \left(\left(\frac{\bar{f}_{k_l}}{s_l} \right) (X_l^i) \right) \right] \right)
$$

$$
+ \sum_{l=1}^{L} E_{s_l} \left[\sum_{i=1}^{n_l} \log \left(\left(\frac{\bar{f}_{k_l}}{s_l} \right) (X_l^i) \right) \right] \right\} + \eta
$$

$$
\leq L \log K + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \inf_{\pi \in \mathbb{S}_{k-1}} \left(\sum_{l=1}^{L} - \log(\pi_{k_l}) \right) + \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) \right.
$$

$$
\left. + \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \bar{f}_{k_l}) \right\} + \eta
$$

which can be rewritten as:

$$
\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) \leq L \log K + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \sum_{l=1}^{L} - \log(\frac{1}{K}) + \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) \right.
$$

$$
+ \inf_{f \in \mathcal{F}_m} \left(\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) \right) \right\} - \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k_l}}) + \delta + \eta \tag{6}
$$

$$
\leq 2L \log K + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m} \left(\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) \right) \right\}
$$

$$
- \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k_l}}) + \delta + \eta,
$$

where the infimum over the simplex is upper-bounded by a uniform distribution. It remains to get an upper bound of the deviation $-\sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k_l}})$, which is based on the following lemmas. First define for all random variable Z and any event A of positive probability, $E^A[Z] = \frac{E[Z \mathbb{1}_A]}{pr(A)}$. We have the following results.

Lemma A.2. *Let Z be a random variable and Ψ a non-decreasing function such that for all measurable event A satisfying $pr(A) > 0$, $E^A[Z] \leq \Psi \left(\log \left(\frac{1}{pr(A)} \right) \right)$. Then, for all $x \geq 0$, $pr(Z > \Psi(x)) \leq e^{-x}$.*

Lemma A.3. *Let $(k_1, \ldots, k_L) \in [\![1..K]\!]^L$ be a group assignment for each observation x_l. Then, there exist three absolute constants $\kappa_0' > 4$, κ_1' and κ_2' such that, under Assumption A.2, for all $m \in \mathcal{M}$, for all $y_m > \sigma_m$ and every measurable event A such that $pr(A) > 0$,*

$$
E^A \left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E \left[| \log \left(\frac{f_{k_l}}{s_l} \right) |^2 \right]} \right] \leq \kappa_1' \frac{\sigma_m}{y_m} + \frac{\kappa_2'}{(ny_m^2)^{1/2}} \log \left(\frac{1}{pr(A)} \right)^{1/2}
$$

$$
+ \frac{9\tau_n}{ny_m^2} \log \left(\frac{1}{pr(A)} \right).
$$

19

Then, for all $\lambda > 0$ we can derive that:

$$E^A\left[\sup_{(k_l)_l \in [\![1..K]\!]^L} \sup_{f \in \mathcal{F}_m} \exp\left(\frac{\lambda}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right)\right]$$

$$= E^A\left[\sup_{(k_l)_l \in [\![1..K]\!]^L} \exp\left(\sup_{f \in \mathcal{F}_m} \frac{\lambda}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right)\right]$$

$$\leq E^A\left[\sum_{(k_l)_l \in [\![1..K]\!]^L} \exp\left(\sup_{f \in \mathcal{F}_m} \frac{\lambda}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right)\right]$$

Therefore, by Lemmas A.3 and A.2, for all $x > 0$, except on a set of probability less than e^{-x},

$$\sum_{(k_l)_l \in [\![1..K]\!]^L} \exp\left(\sup_{f \in \mathcal{F}_m} \frac{\lambda}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right)$$

$$\leq \sum_{(k_l)_l \in [\![1..K]\!]^L} \exp\left(\frac{\lambda \kappa_1' \sigma_m}{y_m} + \frac{\lambda \kappa_2'}{(n y_m^2)^{1/2}} x^{1/2} + \lambda\left(\frac{9\tau_n}{n y_m^2}\right) x\right)$$

$$\leq K^L \exp\left(\frac{\lambda \kappa_1' \sigma_m}{y_m} + \frac{\lambda \kappa_2'}{(n y_m^2)^{1/2}} x^{1/2} + \lambda\left(\frac{9\tau_n}{n y_m^2}\right) x\right).$$

A fortiori, we have

$$\exp\left(\frac{\lambda}{n} \sum_{l=1}^{L} \frac{-\nu_l(\widehat{f}_{\widehat{k_l}})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l}\right)|^2\right]}\right) \leq K^L \exp\left(\frac{\lambda \kappa_1' \sigma_m}{y_m} + \frac{\lambda \kappa_2'}{(n y_m^2)^{1/2}} x^{1/2} + \lambda\left(\frac{9\tau_n}{n y_m^2}\right) x\right),$$

and except on a set of probability less than e^{-x},

$$\frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\widehat{f}_{\widehat{k_l}})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l}\right)|^2\right]} \leq \frac{L \log(K)}{\lambda} + \frac{\kappa_1' \sigma_m}{y_m} + \frac{\kappa_2'}{(n y_m^2)^{1/2}} x^{1/2} + \left(\frac{9\tau_n}{n y_m^2}\right) x.$$

It remains to choose λ as $\lambda_m = n y_m^2 > 0$ and $y_m = \theta\left(\frac{x}{n} + \sigma_m^2 + \frac{L \log(K)}{n}\right)^{1/2}$, with $\theta > 1$ to be explicited later on. We can already write

$$\frac{1}{n} \sum_{l=1}^{L} -\nu_l(\widehat{f}_{\widehat{k_l}}) \leq \left(y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l}\right)|^2\right]\right)\left(\frac{9\tau_n + 1}{\theta^2} + \frac{\kappa_1' + \kappa_2'}{\theta}\right)$$

and obtain an upper bound for $\frac{1}{n} \sum_{l=1}^{L} -\nu_l(\widehat{f}_{\widehat{k_l}})$. By using equation (6),

$$\frac{1}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) \leq \frac{2L \log K}{n} + \inf_{(k_l)_l \in [\![1..K]\!]^L}\left(\frac{1}{n} \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m}\left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l})\right)\right) + \frac{\delta + \eta}{n}$$

$$+ \left(y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l}\right)|^2\right]\right)\left(\frac{9\tau_n + 1}{\theta^2} + \frac{\kappa_1' + \kappa_2'}{\theta}\right).$$

20

Now we define $C_{\tau_n} = \frac{e^{-\tau_n} + \tau_n - 1}{\tau_n^2}$ and choose $\epsilon_{pen} > 0$ such that $\left(\frac{9\tau_n + 1}{\theta_{pen}^2} + \frac{\kappa_1' + \kappa_2'}{\theta_{pen}} \right) \kappa_0' = C_{\tau_n} \epsilon_{pen}$. Using Proposition A.1, we obtain

$$\frac{1}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) \leq \frac{2L \log K}{n} + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\frac{1}{n} \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) \right)$$
$$+ \frac{\delta + \eta}{n} + \frac{\epsilon_{pen}}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) + \frac{y_m^2 C_{\tau_n} \epsilon_{pen}}{\kappa_0'}.$$

Therefore, with probability less than e^{-x},

$$\frac{1 - \epsilon_{pen}}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) - \frac{2L \log K}{n} - \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\frac{1}{n} \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) \right)$$
$$- \frac{\delta + \eta}{n} - \frac{\theta_{pen}^2 (\sigma_m^2 + \frac{L \log(K)}{n}) C_{\tau_n} \epsilon_{pen}}{\kappa_0'} > \frac{x}{n} \theta_{pen}^2 \frac{C_{\tau_n} \epsilon_{pen}}{\kappa_0'}.$$

For all $\alpha > 0$ and any non negative random variable, we have $E[Z] = \alpha \int_{x \geq 0} \mathrm{pr}(Z > \alpha x) dx$. Furthermore, let $\kappa_0 = \frac{C_{\tau_n} \epsilon_{pen} \theta_{pen}^2}{\kappa_0'}$, we get:

$$E \left[\frac{1 - \epsilon_{pen}}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) - \frac{2L \log K}{n} - \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\frac{1}{n} \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) \right) \right.$$
$$\left. - \frac{\delta + \eta}{n} - \kappa_0 (\sigma_m^2 + \frac{L \log(K)}{n}) \right] \leq \frac{\kappa_0}{n}.$$

Since $E[\frac{1}{n} \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l})(X_l)] = 0$ for any cluster assignment $(k_l)_l \in [\![1..K]\!]^L$ of the observations, we derive:

$$E \left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) \right] \leq \frac{1}{1 - \epsilon_{pen}} \left\{ \inf_{\substack{f \in \mathcal{F}_m \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) \right.$$
$$\left. + (2 + \kappa_0) \frac{L \log K}{n} + \kappa_0 \sigma_m^2 + \frac{\kappa_0}{n} + \frac{\delta + \eta}{n} \right\}.$$

Recalling that δ can be chosen arbitrary small, this leads to:

$$E \left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f}_{\widehat{k_l}}) \right] \leq \frac{1}{1 - \epsilon_{pen}} \left(\inf_{\substack{f \in \mathcal{F}_m \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) \right) \right.$$
$$\left. + (2 + \kappa_0) \frac{L \log K}{n} + \kappa_0 \sigma_m^2 + \frac{\kappa_0}{n} + \frac{\eta}{n} \right),$$

which concludes the proof by taking $C_1 = \frac{1}{1 - \epsilon_{pen}}$ and $C_2 = \frac{\kappa_0}{1 - \epsilon_{pen}}$. $\qquad \square$

A.2.1 Proof of Lemma A.2

Let $x \geq 0$. Take $A = \{Z < \Psi(x)\}$. Either $\mathrm{pr}(A) = 0 \leq e^{-x}$ or $\mathrm{pr}(A) > 0$ in which case, by assumption,

$$\Psi(x) < \frac{E[Z \mathbb{1}_{\{Z < \Psi(x)\}}]}{\mathrm{pr}(Z < \Psi(x))} \leq \Psi \left(\log \left(\frac{1}{\mathrm{pr}(Z < \Psi(x))} \right) \right).$$

Since Ψ is non-decreasing, $x < \log\left(\frac{1}{\mathrm{pr}(Z < \Psi(x))}\right)$, which leads to the conclusion.

A.2.2 Proof of Lemma A.3

Consider a class of real-valued and measurable functions defined below. For a fixed family of K functions denoted by $\tilde{f} = (\tilde{f}_1, \ldots, \tilde{f}_K)$:

$$
\begin{aligned}
\mathcal{G}(\tilde{f}, \sigma) &= \left\{ \log\left(\prod_{i=1}^{n_l} \frac{f_{k_l}}{\tilde{f}_{k_l}}\right)_{l \in \{1,\ldots,L\}} \mid f \in \mathcal{F}_m, \mathbf{a}(f, \tilde{f}) \leq \sigma^2 \right\} \\
&= \left\{ \left(-\mathbf{kl}(f_{k_l}) + \mathbf{kl}(\tilde{f}_{k_l})\right)_{l \in \{1,\ldots,L\}} \mid f \in \mathcal{F}_m, \mathbf{a}(f, \tilde{f}) \leq \sigma^2 \right\}
\end{aligned}
$$

We are focusing on $W(\tilde{f}, \sigma) = \sup_{f \in \mathcal{G}(\tilde{f}, \sigma)} \sum_{l=1}^{L} (-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l}))$. If $[f_{k_l}^-, f_{k_l}^+]$ is a bracket of size γ containing f_{k_l}, then

$$
g_{k_l}^- = \log\left(\frac{f_{k_l}^-}{\tilde{f}_{k_l}}\right) \leq \log\left(\frac{f_{k_l}}{\tilde{f}_{k_l}}\right) \leq \log\left(\frac{f_{k_l}^+}{\tilde{f}_{k_l}}\right) = g_{k_l}^+ \text{ and } g_{k_l}^+ - g_{k_l}^- = \log\left(\frac{f_{k_l}^+}{f_{k_l}^-}\right).
$$

According to Assumption A.1, $\left| \log\left(\frac{f_{k_l}^+}{f_{k_l}^-}\right) \right| \leq \tau_n$. Thus, for any integer $j \geq 2$:

$$
\frac{1}{n} \sum_{l=1}^{L} \sum_{i=1}^{n_l} E_{s_l}\left[\mid g_{k_l}^+ - g_{k_l}^- \mid^j\right] \leq \frac{j!}{2} \tau_n^{j-2} \gamma^2.
$$

Recall Theorem 6.8 from Massart (2007):

Theorem A.2. *Let $\mathcal{G}$ be a countable class of real-valued and measurable functions. Assume that there exist some positive numbers V and b such that for all $f \in \mathcal{G}$ and all integers $j \geq 2$,*

$$
E[\mid f \mid^j] \leq \frac{j!}{2} V b^{j-2}.
$$

Assume furthermore that for any positive number γ, there exist a finite set $\mathcal{B}(\gamma)$ of brackets covering $\mathcal{G}$ such that for any bracket $[g^-, g^+] \in \mathcal{B}(\gamma)$ and all integers $k \geq 2$,

$$
E[\mid g^+ - g^- \mid^j] \leq \frac{j!}{2} \gamma^2 b^{j-2}.
$$

Let $e^{H(\gamma)}$ denote the minimal cardinality of such a covering. Then, there exists an absolute constant κ such that for any $\epsilon \in (0, 1]$ and any measurable set A with $pr(A) > 0$,

$$
E^A\left[\frac{1}{n} \sup_{f \in \mathcal{G}} \sum_{l=1}^{L} \nu_l(f)\right] \leq E + \frac{(1 + 6\epsilon)(2V)^{1/2}}{n^{1/2}} \log\left(\frac{1}{pr(A)}\right)^{1/2} + \frac{2b}{n} \log\left(\frac{1}{pr(A)}\right),
$$

where $E = \frac{\kappa}{\epsilon} \frac{1}{\sqrt{n}} \int_0^{\epsilon V^{1/2}} (H(\gamma) \wedge n)^{1/2} d\gamma + \frac{2(b + \sqrt{V})}{n} H(V^{1/2})$. Furthermore, $\kappa \leq 27$.

In our context, the assumptions of the theorem are satisfied on $\mathcal{G}(\tilde{f}, \sigma)$ with $V = \sigma^2$ and $b = \tau_n$ except that the set $\mathcal{G}(\tilde{f}, \sigma)$ is not necessarily countable so the supremum $W(\tilde{f}, \sigma)$ can be non measurable. We rather define the countable subset:

$$
\mathcal{G}'(\tilde{f}, \sigma) = \left\{ \left(-\mathbf{kl}(f_{k_l}) + \mathbf{kl}(\tilde{f}_{k_l})\right)_{l \in \{1,\ldots,L\}} \mid f \in \mathcal{F}'_m, \mathbf{a}(f, \tilde{f}) \leq \sigma^2 \right\},
$$

with $\mathcal{F}'_m$ satisfying Assumption A.3. Thus, $W(\tilde{f}, \sigma) = \sup_{f \in \mathcal{G}'(\tilde{f}, \sigma)} \sum_{l=1}^{L} (-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l}))$ almost surely, and by applying Theorem A.2, we can conclude that

$$E^A \left[\sup_{f \in \mathcal{G}(\tilde{f}, \sigma)} \frac{1}{n} \sum_{l=1}^{L} (-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})) \right] \leq E + \frac{(1 + 6\epsilon)\sigma 2^{1/2}}{n^{1/2}} \log \left(\frac{1}{\mathrm{pr}(A)} \right)^{1/2} + \frac{2\tau_n}{n} \log \left(\frac{1}{\mathrm{pr}(A)} \right).$$

Let find an upper bound for E. Take $\epsilon = 1$. We have:

$$E = \frac{\kappa}{n^{1/2}} \int_0^\sigma (H(\gamma) \wedge n)^{1/2} d\gamma + \frac{2(\tau_n + \sigma)}{n} H(\sigma).$$

The mapping $\gamma \mapsto H(\gamma, \mathcal{F}_m(\tilde{f}, \sigma))$ is non-increasing. By Assumption A.2, if $\tilde{f} \in \mathcal{F}_m$,

$$\int_0^\sigma (H(\gamma, \mathcal{F}_m(\tilde{f}, \sigma)) \wedge n)^{1/2} d\gamma \leq \phi_m(\sigma).$$

Also,

$$H(\sigma, \mathcal{F}_m(\tilde{f}, \sigma)) = \frac{1}{\sigma} \int_0^\sigma H(\sigma, \mathcal{F}_m(\tilde{f}, \sigma)) d\gamma \leq \left(\frac{1}{\sigma} \int_0^\sigma H(\gamma, \mathcal{F}_m(\tilde{f}, \sigma)) d\gamma \right)^2 \leq \frac{\phi_m^2(\sigma)}{\sigma^2}.$$

By inserting these bounds,

$$E \leq \kappa \frac{1}{n^{1/2}} \phi_m(\sigma) + \frac{2(\tau_n + \sigma)}{n} \frac{\phi_m^2(\sigma)}{\sigma^2} \leq \left(\kappa + 2(\tau_n + \sigma) \frac{\phi_m(\sigma)}{n^{1/2}\sigma^2} \right) \frac{\phi_m(\sigma)}{n^{1/2}}.$$

Since $\delta \mapsto \delta^{-1}\phi_m(\delta)$ is non-increasing, so is $\delta \mapsto \delta^{-2}\phi_m(\delta)$. Also, by definition of σ_m, $\frac{\phi_m(\sigma_m)}{n^{1/2}\sigma_m^2} = 1$. Thus, when $\sigma \geq \sigma_m$,

$$E \leq (\kappa + 2(\tau_n + \sigma)) \frac{\phi_m(\sigma)}{n^{1/2}} \leq (27 + 2(b + \sigma)) \frac{\phi_m(\sigma)}{n^{1/2}}.$$

Notice also that for any family $f, g \in \mathcal{F}_m$:

$$\mathbf{a}(f, g) = \frac{1}{n} \sum_{l=1}^{L} n_l E_{s_l} \left[| \log \left(\frac{f_{k_l}}{g_{k_l}} \right) |^2 \right] \leq \frac{1}{n} \sum_{l=1}^{L} n_l \tau_n^2 = \tau_n^2.$$

Therefore, $\sigma \leq \tau_n$ and for all $\sigma \geq \sigma_m$:

$$E^A \left[\sup_{f \in \mathcal{G}(\tilde{f}, \sigma)} \frac{1}{n} \sum_{l=1}^{L} (-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})) \right] \leq (27 + 4\tau_n) \frac{\phi_m(\sigma)}{n^{1/2}} + \frac{7 \times 2^{1/2}}{n^{1/2}} \sigma \log \left(\frac{1}{pr(A)} \right)^{1/2} + \frac{2\tau_n}{n} \log \left(\frac{1}{pr(A)} \right).$$

We now use the pealing lemma as stated in Lemma 4.23 from Massart (2007) in order to bound the supremum on the overall model.

Lemma A.4 (Pealing lemma). *Let S be a countable set, $\tilde{f} \in S$ and $a : S \to \mathbb{R}^+$ such that $a(\tilde{f}) = \inf_{f \in S} a(f)$. Let Z be a random process indexed by S and $B(\sigma) = \{f \in S \mid a(f) \leq \sigma\}$. Assume that for any positive σ the non-negative random-variable $\sup_{f \in B(\sigma)}(Z(f) - Z(\tilde{f}))$ has finite expectation. Then, for any function ψ on $\mathbb{R}^+$ such that $\frac{\psi(x)}{x}$ is non-increasing on $\mathbb{R}^+$ and*

$$E \left[\sup_{f \in B(\sigma)} Z(f) - Z(\tilde{f}) \right] \leq \psi(\sigma), \sigma \geq \sigma_\star \geq 0,$$

one has for any positive $x \geq \sigma_\star$:

$$E\left[\sup_{f\in S} \frac{Z(f) - Z(\tilde{f})}{x^2 + a^2(f)}\right] \leq 4\frac{\psi(x)}{x^2}.$$

With $S = \mathcal{F}_m$, $\sigma_\star = \sigma_m$, $\tilde{f}$ to be specified later, $a(f) = \frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{f_{k_l}}{\tilde{f}_{k_l}}\right)|^2\right]$, $Z(f) = \frac{1}{n}\sum_{l=1}^L -\nu_l(f_{k_l})$ and $Z(\tilde{f}) = \frac{1}{n}\sum_{l=1}^L -\nu_l(\tilde{f}_{k_l})$, provided $y_m \geq \sigma_m$, we have:

$$E^A\left[\sup_{f\in\mathcal{F}_m} \frac{1}{n}\sum_{l=1}^L \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + \frac{1}{n}\sum_{l=1}^L n_l E_{s_l}\left[|\log\left(\frac{f_{k_l}}{\tilde{f}_{k_l}}\right)|^2\right]}\right] \leq 4(27 + 4\tau_n)\frac{\phi(y_m)}{n^{1/2}y_m^2} + \frac{28\times 2^{1/2}}{n^{1/2}y_m}\log\left(\frac{1}{pr(A)}\right)^{1/2}$$
$$+ \frac{8\tau_n}{ny_m^2}\log\left(\frac{1}{pr(A)}\right).$$

Using the monotonicity of $\delta \mapsto \phi(\delta)/\delta$ and the definition of σ_m,

$$E^A\left[\sup_{f\in\mathcal{F}_m} \frac{1}{n}\sum_{l=1}^L \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + \frac{1}{n}\sum_{l=1}^L n_l E_{s_l}\left[|\log\left(\frac{f_{k_l}}{\tilde{f}_{k_l}}\right)|^2\right]}\right] \leq 4(27 + 4\tau_n)\frac{\sigma_m}{y_m} + \frac{28\sqrt{2}}{n^{1/2}y_m}\log\left(\frac{1}{pr(A)}\right)^{1/2}$$
$$+ \frac{8\tau_n}{ny_m^2}\log\left(\frac{1}{pr(A)}\right).$$

We have chosen $\tilde{f}$ such that for any f family of K functions and any $\epsilon_d > 0$,

$$\frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2\right] \leq (1 + \epsilon_d)\frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{s_l}{f_{k_l}}\right)|^2\right].$$

Therefore, $\frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{f_{k_l}}{\tilde{f}_{k_l}}\right)|^2\right] \leq 2(2 + \epsilon_d)\frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]$ and

$$E^A\left[\sup_{f\in\mathcal{F}_m} \frac{1}{n}\sum_{l=1}^L \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2 + \epsilon_d)\frac{1}{n}\sum_{l=1}^L n_l E\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right] \leq 4(27 + 4\tau_n)\frac{\sigma_m}{y_m}$$
$$+ \frac{28\times 2^{1/2}}{n^{1/2}y_m}\log\left(\frac{1}{pr(A)}\right)^{1/2}$$
$$+ \frac{8\tau_n}{ny_m^2}\log\left(\frac{1}{pr(A)}\right).$$

We now use a Bernstein-type control, which is a rewriting of Bernstein's theorem:

Lemma A.5 (**Bernstein Inequality**). *Assume there exist $V', b' \geq 0$ such that for all integer $j \geq 2$*

$$\frac{1}{n}\sum_{l=1}^L n_l E\left[\log\left(\frac{f_{k_l}}{s_l}\right)_+^j\right] \leq \frac{j!}{2}V'b'^{j-2}.$$

Then, for all measurable event A such that $pr(A) > 0$,

$$E^A\left[\frac{1}{n}\sum_{l=1}^L -\nu_l(f_{k_l})\right] \leq \left(\frac{2V'}{n}\log(\frac{1}{pr(A)})\right)^{1/2} + \frac{b'}{n}\log\left(\frac{1}{pr(A)}\right).$$

24

By taking $V' = \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{\tilde{f}_{k_l}}{s_l}\right)|^2 \right]$ and $b' = \tau_n$ and applying this result to $\tilde{f}$, this yields for all $y_m, \kappa' > 0$:

$$E^A \left[\frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + \kappa'^2 \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2 \right]} \right] \leq \frac{1}{y_m^2 + \kappa'^2 \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2 \right]} \left(\frac{2V'}{n} \right)^{1/2} \log\left(\frac{1}{pr(A)}\right)^{1/2}$$

$$+ \frac{1}{y_m^2 + \kappa'^2 \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2 \right]} \frac{b'}{n} \log\left(\frac{1}{pr(A)}\right)$$

$$\leq \frac{1}{\kappa'} \left(\frac{2}{ny_m^2} \right)^{1/2} \log\left(\frac{1}{pr(A)}\right)^{1/2} + \frac{1}{y_m^2} \frac{b'}{n} \log\left(\frac{1}{pr(A)}\right).$$

Therefore,

$$E^A \left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2 \right]} + \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + \kappa'^2 \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2 \right]} \right]$$

$$\leq 4(27 + 4\tau_n)\frac{\sigma_m}{y_m} + \left(\frac{28 \times 2^{1/2}}{(ny_m^2)^{1/2}} + \frac{1}{\kappa'} \left(\frac{2}{ny_m^2} \right)^{1/2} \right) \log\left(\frac{1}{pr(A)}\right)^{1/2} + \left(\frac{9\tau_n}{ny_m^2} \right) \log\left(\frac{1}{pr(A)}\right).$$

Choosing κ' as $\kappa_d'^2 = \frac{2(2+\epsilon_d)}{1+\epsilon_d}$, we can conclude that since:

$$E^A \left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2 \right]} + \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + \kappa'^2 \frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2 \right]} \right]$$

$$\geq E^A \left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2 \right]} + \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2 \right]} \right],$$

we have

$$E^A \left[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n} \sum_{l=1}^{L} n_l E \left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2 \right]} \right]$$

$$\leq 4(27 + 4\tau_n)\frac{\sigma_m}{y_m} + \left(\frac{28 \times 2^{1/2}}{(ny_m^2)^{1/2}} + \frac{1}{\kappa_d'} \left(\frac{2}{ny_m^2} \right)^{1/2} \right) \log\left(\frac{1}{pr(A)}\right)^{1/2} + \frac{9\tau_n}{ny_m^2} \log\left(\frac{1}{pr(A)}\right).$$

Defining $\kappa_1' = 4(27 + 4\tau_n)$, $\kappa_2' = 2^{1/2}(28 + \frac{1}{\kappa_d'})$ and $\kappa_0' = 2(2 + \epsilon_d)$ leads to the conclusion.

A.3 Model selection theorem

Just as model complexity appeared in the single model inequality, the multi-model case involves a term that takes the global collection into account. Therefore, we assume the existence of the following Kraft inequality which bounds in a sense the complexity of our collection of models:

Assumption A.4 (**Kraft inequality**). *There exists a family $(x_m)_{m \in \mathcal{M}}$ of non-negative numbers such that*

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \Sigma < +\infty.$$

Theorem A.3. *Let $(x_1, \ldots, x_L)$ be L observations of independent random vectors $(X_1, \ldots, X_L)$ where each X_l consists of n_l independent and identically distributed instances of a multinomial vector that has s_l as a true categorical density with respect to some known positive measure. Assume $(\mathcal{S}_m)_{m \in \mathcal{M}}$ is an at most countable collection of models for which Assumption A.4 holds. For every model $S_m \in \mathcal{S}$, we also assume that Assumptions A.1, A.2 and A.3 hold. Let* **pen** *be a non-negative penalty function and $\widehat{m}$ any η'-minimizer of*

$$\mathbf{crit}(m) = \min_{(\pi^m, f^m) \in \Theta_m} \gamma_n(\pi^m, f^m) + \mathbf{pen}(m).$$

Let $(\widehat{\pi^{\widehat{m}}}, \widehat{f^{\widehat{m}}})$ be the corresponding η-likelihood minimizers in $\mathcal{S}_{\widehat{m}}$ and define the resulting cluster assignment for each vector $x_l = (x_l^1, \ldots, x_l^{n_l})$:

$$\widehat{k}_l = \operatorname*{argmax}_{k \in [\![1..\widehat{K}]\!]} \left\{ \widehat{\pi}_k^{\widehat{m}} \left(\prod_{i=1}^{n_l} \widehat{f}_k^{\widehat{m}}(x_l^i) \right) \right\}.$$

Define $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$. Then, for any $C_1 > 0$, there exist some constants κ_0 and C_2 that depend only on C_1, and such that whenever

$$\mathbf{pen}(m) \geq \kappa(\mathfrak{D}_m + L\log(K) + x_m) \ \text{with} \ \kappa > 1 + \kappa_0,$$

for all model $m \in \mathcal{M}$, the penalized log-likelihood estimate satisfies

$$E\left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f}_{\widehat{k}_l}^{\widehat{m}}) \right] \leq C_1 \inf_{m \in \mathcal{M}} \left\{ \inf_{\substack{f \in \mathcal{F}_m \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}^m) + \frac{\mathbf{pen}(m)}{n} \right) \right\}$$
$$+ C_2 \frac{\Sigma}{n} + \frac{\eta + \eta'}{n}.$$

Proof. For any cluster assignment $(k_1, \ldots, k_L)$ of the observations within the model m, define $\overline{f}^m \in \mathcal{F}_m$ such that:

$$\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \overline{f}_{k_l}^m) \leq \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) + \delta. \tag{7}$$

Fix also $m \in \mathcal{M}$ such that $\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \overline{f}_{k_l}^m) < +\infty$ and define

$$\mathcal{M}' = \left\{ m' \in \mathcal{M} \mid \gamma_n(\widehat{\pi}^{m'}, \widehat{f}^{m'}) + \mathbf{pen}(m') \leq \gamma_n(\widehat{\pi}^m, \widehat{f}^m) + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta' \right\}.$$

By definition of the estimator and since $-\log$ is decreasing,

$$\gamma_n(\widehat{\pi}^m, \widehat{f}^m) + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta' \leq \inf_{(\pi^m, f^m) \in \Theta_m} \gamma_n(\pi^m, f^m) + \eta$$
$$+ \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta'$$
$$\leq \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \inf_{(\pi^m, f^m) \in \Theta_m} \left(\sum_{l=1}^{L} -\log(\pi_{k_l}^m) \right. \right.$$
$$+ \left. \left. \sum_{l=1}^{L} \sum_{i=1}^{n_l} -\log\left(\left(\frac{f_{k_l}^m}{s_l} \right)(X_l^i) \right) \right) \right\}$$
$$+ \eta + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta'.$$

Using the previous definition of ν_l in (5) besides equation (7) and by assumption on **pen**,

$$\inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\inf_{(\pi^m, f^m) \in \Theta_m} \left(\sum_{l=1}^{L} -\log(\pi_{k_l}^m) + \sum_{l=1}^{L} \sum_{i=1}^{n_l} -\log\left(\left(\frac{f_{k_l}^m}{s_l} \right)(x_l^i) \right) \right) \right)$$
$$+ \eta + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta'$$
$$\leq \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\sum_{l=1}^{L} -\log(\frac{1}{K}) + \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^m) + \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \bar{f}_{k_l}^m) \right)$$
$$+ \eta + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m) + \eta'$$
$$\leq L\log(K) + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) + \delta$$
$$+ \eta + \eta' + \kappa_0(\mathfrak{D}_m + L\log(K) + x_m)$$
$$\leq \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) + \delta + \eta + \eta' + \mathbf{pen}(m).$$

It follows that

$$\gamma_n(\widehat{\pi}^{m'}, \widehat{f}^{m'}) + \mathbf{pen}(m') \leq \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) + \delta$$
$$+ \eta + \eta' + \mathbf{pen}(m)$$

and on the other hand,

$$\gamma_n(\widehat{\pi}^{m'}, \widehat{f}^{m'}) + \mathbf{pen}(m') \geq \sum_{l=1}^{L} -\log\left(K'\widehat{\pi}_{k_l}^{m'} \left(\prod_{i=1}^{n_l} \left(\frac{\widehat{f}_{k_l}^{m'}}{s_l} \right)(x_l^i) \right) \right) + \mathbf{pen}(m')$$
$$\geq \sum_{l=1}^{L} -\log\left(K' \left(\prod_{i=1}^{n_l} \left(\frac{\widehat{f}_{k_l}^{m'}}{s_l} \right)(x_l^i) \right) \right) + \mathbf{pen}(m')$$
$$\geq -L\log(K') + \sum_{l=1}^{L} \nu_l(\widehat{f}_{\underset{\sim}{k_l}}^{m'}) + \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\underset{\sim}{k_l}}^{m'}) + \mathbf{pen}(m'),$$

where K' denotes the number of clusters given in model m'. Thus,

$$\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k}_l}^{m'}) \leq L\log(K') - \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l}^{m'}) - \mathbf{pen}(m')$$

$$+ \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^{m}) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^{m}) \right) + \delta$$

$$+ \eta + \eta' + \mathbf{pen}(m)$$

It remains to get an upper bound of the deviation $-\sum_{l=1}^{L} \nu_l(\widehat{f_{\widehat{k}_l}^{m'}})$. Using Lemmas A.2 and A.3, except on a set of probability less than $e^{-x'_m - x}$, for any $y_{m'} > \sigma_{m'}$,

$$\frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\widehat{f}_{\widehat{k}_l}^{m'})}{y_{m'}^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k}_l}^{m'}}{s_l}\right)|^2 \right]} \leq \frac{L\log(K')}{\lambda} + \frac{\kappa_1' \sigma_{m'}}{y_{m'}} + \frac{\kappa_2'}{(ny_{m'}^2)^{1/2}}(x + x_{m'})^{1/2}$$

$$+ \left(\frac{9\tau_n}{ny_{m'}^2}\right)(x + x_{m'}).$$

This time we choose λ as $\lambda_m = ny_{m'}^2 > 0$ and $y_{m'} = \theta(\frac{x + x_{m'}}{n} + \sigma_{m'}^2 + \frac{L\log(K')}{n})^{1/2}$, with $\theta > 1$ to be explicited later on. We deduce that except on a set of probability less than $e^{-x'_m - x}$, for any $y_{m'} > \sigma_{m'}$,

$$\frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(\widehat{f}_{\widehat{k}_l}^{m'})}{y_{m'}^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k}_l}^{m'}}{s_l}\right)|^2 \right]} \leq \left(\frac{9\tau_n + 1}{\theta^2} + \frac{\kappa_1' + \kappa_2'}{\theta}\right).$$

Now, we use Kraft condition A.4, and conclude that if we make a proper choice of $y_{m'}$ for all models $m' \in \mathcal{M}'$, this property holds simultaneously on $\mathcal{M}'$ except on a set of probability less than $e^{-x}\Sigma$. Therefore, except on that set, for all $m' \in \mathcal{M}'$,

$$\frac{1}{n} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k}_l}^{m'}) \leq \frac{L\log(K')}{n} + \left(y_{m'}^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k}_l}^{m'}}{s_l}\right)|^2 \right]\right)\left(\frac{9\tau_n + 1}{\theta^2} + \frac{\kappa_1' + \kappa_2'}{\theta}\right)$$

$$- \frac{\mathbf{pen}(m')}{n} + \inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\frac{1}{n}\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^{m}) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^{m}) \right)$$

$$+ \frac{\delta}{n} + \frac{\eta + \eta'}{n} + \frac{\mathbf{pen}(m)}{n}.$$

Define $C_{\tau_n} = \frac{e^{-\tau_n} + \tau_n - 1}{\tau_n^2}$ and choose $\epsilon_{pen} > 0$ such that $\left(\frac{9\tau_n + 1}{\theta_{pen}^2} + \frac{\kappa_1' + \kappa_2'}{\theta_{pen}}\right)\kappa_0' = C_{\tau_n}\epsilon_{pen}$. We

obtain

$$\frac{1}{n}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\widehat{f}_{\widehat{k}_l}^{m'}) \leq \frac{L\log(K')}{n} + \frac{y_{m'}^2 C_{\tau_n}\epsilon_{pen}}{\kappa_0'} + \frac{C_{\tau_n}\epsilon_{pen}}{n}\sum_{l=1}^{L}n_l E\left[|\log\left(\frac{\widehat{f}_{\widehat{k}_l}^{m'}}{s_l}\right)|^2\right]$$

$$-\frac{\mathbf{pen}(m')}{n} + \inf_{(k_l)_l \in [\![1..K]\!]^L}\left(\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f_{k_l}^m)\right)$$

$$+\frac{\delta}{n} + \frac{\eta+\eta'}{n} + \frac{\mathbf{pen}(m)}{n}.$$

Then, by using Proposition A.1, simultaneously for any $m' \in \mathcal{M}'$ and except on a set of probability less than $e^{-x}\Sigma$,

$$(1-\epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\widehat{f}_{\widehat{k}_l}^{m'}) \leq \frac{L\log(K')}{n} + \frac{y_{m'}^2 C_{\tau_n}\epsilon_{pen}}{\kappa_0'} - \frac{\mathbf{pen}(m')}{n}$$

$$+ \inf_{(k_l)_l \in [\![1..K]\!]^L}\left(\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f_{k_l}^m)\right) \quad (8)$$

$$+\frac{\delta}{n} + \frac{\eta+\eta'}{n} + \frac{\mathbf{pen}(m)}{n}.$$

Let now study the term $\frac{L\log(K')}{n} + \frac{y_{m'}^2 C_{\tau_n}\epsilon_{pen}}{\kappa_0'} - \frac{\mathbf{pen}(m')}{n}$. Define $\kappa_0 = \frac{\theta_{pen}^2 C_{\tau_n}\epsilon_{pen}}{\kappa_0'}$. We have

$$\frac{L\log(K')}{n} + \frac{y_{m'}^2 C_{\tau_n}\epsilon_{pen}}{\kappa_0'} - \frac{\mathbf{pen}(m')}{n} = \frac{L\log(K')}{n} + \kappa_0\left(\frac{x+x_{m'}}{n} + \sigma_{m'}^2 + \frac{L\log(K')}{n}\right) - \frac{\mathbf{pen}(m')}{n}$$

$$\leq \kappa_0\frac{x}{n} + (1+\kappa_0)\left(\frac{x_{m'}}{n} + \sigma_{m'}^2 + \frac{L\log(K')}{n}\right) - \frac{\mathbf{pen}(m')}{n}$$

$$\leq \kappa_0\frac{x}{n} - (1 - \frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(m')}{n}.$$

Thus, based on equation (8), except on a set of probability less than $e^{-x}\Sigma$, simultaneously for any $m' \in \mathcal{M}'$,

$$(1-\epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\widehat{f}_{\widehat{k}_l}^{m'}) + (1 - \frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(m')}{n} - \inf_{(k_l)_l \in [\![1..K]\!]^L}\left(\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f_{k_l}^m)\right)$$

$$\leq \frac{\delta}{n} + \frac{\eta+\eta'}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 x}{n}.$$

Since $\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}_{k_l}^m)$ is integrable (with null expectation), we deduce that $M = \sup_{m' \in \mathcal{M}'}\frac{\mathbf{pen}(m')}{n}$ is almost surely finite. By definition of the mapping $\mathbf{pen}$, $\kappa\frac{x_{m'}}{n} \leq M$ for all $m' \in \mathcal{M}'$. Therefore,

$$\Sigma \geq \sum_{m' \in \mathcal{M}'}e^{-x_{m'}} \geq |\mathcal{M}'|\, e^{-\frac{Mn}{\kappa}},$$

and $\mathcal{M}'$ is almost surely finite. Thus, for all fixed $m \in \mathcal{M}$ fixed, a minimizer $\widehat{m}$ over $\mathcal{M}'$ of

$$\mathbf{crit}(m') = \gamma_n(\widehat{\pi}^{m'},\widehat{f}^{m'}) + \mathbf{pen}(m')$$

29

exists. For this minimizer, with probability greater than $1 - e^{-x}\Sigma$, one has

$$(1 - \epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{\widehat{f_{k_l}^m}}) + (1 - \frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(\widehat{m})}{n} - \inf_{(k_l)_l \in [\![1..K]\!]^L} \left\{ \frac{1}{n}\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}^m) + \inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right\}$$

$$\leq \frac{\delta}{n} + \frac{\eta + \eta'}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 x}{n}.$$

Using the same integration technique than in the previous theorem,

$$E\left[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{\widehat{f_{k_l}^m}}) + \frac{(1 - \frac{1+\kappa_0}{\kappa})}{1 - \epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n} \right] \leq \frac{1}{1 - \epsilon_{pen}} \left(\inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) \right.$$

$$\left. + \frac{\delta}{n} + \frac{\eta + \eta'}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 \Sigma}{n} \right),$$

and since δ can be artbitrary small,

$$E\left[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{\widehat{f_{k_l}^m}}) + \frac{(1 - \frac{1+\kappa_0}{\kappa})}{1 - \epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n} \right] \leq \frac{1}{1 - \epsilon_{pen}} \left(\inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) \right.$$

$$\left. + \frac{\eta + \eta'}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 \Sigma}{n} \right).$$

This inequality is true for all $m \in \mathcal{M}$. Therefore,

$$E\left[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{\widehat{f_{k_l}^m}}) + \frac{(1 - \frac{1+\kappa_0}{\kappa})}{1 - \epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n} \right]$$

$$\leq \frac{1}{1 - \epsilon_{pen}} \inf_{m \in \mathcal{M}} \left(\inf_{(k_l)_l \in [\![1..K]\!]^L} \left(\inf_{f^m \in \mathcal{F}_m} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}^m) \right) + \frac{\mathbf{pen}(m)}{n} \right) + \frac{\kappa_0}{1 - \epsilon_{pen}}\frac{\Sigma}{n} + \frac{\eta + \eta'}{n},$$

and by taking $C_1 = \frac{1}{1 - \epsilon_{pen}}$ and $C_2 = \frac{\kappa_0}{1 - \epsilon_{pen}}$, we deduce an inequality stronger than the result stated in the theorem because the penalty appears with a positive coefficient on the left-side. $\square$

A.4 Proofs of Theorems 1 and 2

A.4.1 Bracketing Entropy

A number of lemmas concerning entropy with bracketing is provided in this paragraph. These results are intended to compute the bracketing entropy of the K-product set $\mathcal{F}_K$ with respect to $\mathbf{a}$, based on simpler sets of functions. Let first introduce two other metrics. For $f_k \in \mathcal{F}$, we denote by $\|.\|_\infty$ the L_∞-norm:

$$\|f_k\|_\infty = \max_{1 \leq b \leq B} | f_k(b) |$$

For $f, \tilde{f} \in \mathcal{F}_K$ and $(k_1, \ldots, k_L) \in [\![1..K]\!]^L$, let also $\mathbf{d}_\infty^2$ be the divergence defined by:

$$\mathbf{d}_\infty^2(\tilde{f}, f) = \frac{1}{n}\sum_{l=1}^{L} n_l \|\log\left(\frac{\tilde{f}_{k_l}}{f_{k_l}}\right)\|_\infty^2.$$

Lemma A.6. *Let* $(\delta_1, \ldots, \delta_K)$ *be a family of positive numbers. Then,*

$$H_{[.],\mathbf{d}_\infty^2}(\max_{k=1,\ldots,K} \delta_k^2, \mathcal{F}_K) \leq \sum_{k=1}^{K} H_{[.],\mathbf{d}_\infty^2}(\delta_k^2, \mathcal{F})$$

Proof. For all $k \in [\![1..K]\!]$, let $\delta_k > 0$ and $[f_k^-, f_k^+]$ a bracket of $\mathbf{d}_\infty^2$-diameter less than δ_k^2 in $\mathcal{F}$. Then,

$$\mathbf{d}_\infty^2((f_k^-)_{k=1,\ldots,K}, (f_k^+)_{k=1,\ldots,K}) = \frac{1}{n} \sum_{l=1}^{L} n_l \|\log\left(\frac{f_{k_l}^-}{f_{k_l}^+}\right)\|_\infty^2 \leq \max_{k \in [\![1..K]\!]} \delta_k^2.$$

Therefore, by covering all $\mathcal{F}$ with a number of brackets N_k of width less than δ_k^2, we can cover $\mathcal{F}_K$ with $\prod_k N_k$ brackets, which leads to the result. $\qquad\square$

Lemma A.7. *Let $\epsilon > 0$. Then*

$$H_{[.],\mathbf{d}_\infty^2}(\epsilon^2, \mathcal{F}) \leq H_{[.],\|.\|_\infty}(\epsilon, [-\tau_n, 0]^B).$$

Proof. By definition, $\mathcal{F} = \mathbb{S}_{B-1} \cap [e^{-\tau_n}, 1]^B$. Let $\mathcal{B}_k(\epsilon)$ be a set of brackets of $\|.\|_\infty$-width less than ϵ covering $[-\tau_n, 0]^B$. Let $f_k \in \mathcal{F}$. Then, $\log(f_k) \in [-\tau_n, 0]^B$ and there exists a bracket $[u_k^-, u_k^+] \in \mathcal{B}_k(\epsilon)$ such that $u_k^- \leq \log(f_k) \leq u_k^+$ with $\|u_k^- - u_k^+\|_\infty \leq \epsilon$. We can rewrite u_k^- and u_k^+ as $u_k^- = \log(v_k^-)$, $u_k^+ = \log(v_k^+)$ respectively. This leads to a bracket $[v_k^-, v_k^+]$ of width less than ϵ^2 with respect to $\mathbf{d}_\infty^2$. Therefore, an $\epsilon - \|.\|_\infty$-covering of $[-\tau_n, 0]^B$ induces an $\epsilon^2 - \mathbf{d}_\infty^2$-covering of $\mathcal{F}$ so we can conclude. $\qquad\square$

The following result is inspired by Lemma 2 from Genovese and Wasserman (2000).

Lemma A.8. *Let $\epsilon > 0$. Then*

$$H_{[.],\|.\|_\infty}(\epsilon, [-\tau_n, 0]^B) \leq B \log(2) + B \left(\log\left(\frac{\tau_n}{\epsilon}\right)\right)_+$$

Proof. Divide the cube $[-\tau_n, 0]^B$ of $\mathbb{R}^B$ into a number of N disjoint cubes with sides parallels to the axes and of length ϵ. For one cube, let x_1 the closest vertex from 0, and y_1 the furthest vertex from 0. We have $\max_{b \in [\![1..B]\!]} | x_1(b) - y_1(b) | \leq \epsilon$. Thus, the family of vertices $\{(x_1, y_1), \ldots, (x_N, y_N)\}$ forms an $\epsilon - \|.\|_\infty$ bracketing of $[-\tau_n, 0]^B$. Clearly, we have

$$N \leq \left(1 + \frac{\tau_n}{\epsilon}\right)^B \leq \max\left(2^B, \left(\frac{2\tau_n}{\epsilon}\right)^B\right) \leq 2^B \max\left(1, \frac{\tau_n}{\epsilon}\right)^B.$$

$\qquad\square$

Proposition A.2 (Bracketing entropy of $\mathcal{F}_{\mathbf{K}}$). *For all $K \in \mathcal{M}$ and all $\delta \in (0, 1]$, we have*

$$H_{[.],\mathbf{a}}(\delta, \mathcal{F}_K) \leq KB \left(\log(2\tau_n) + \log\left(\frac{1}{\delta}\right)\right).$$

Proof. Let $\delta \in (0, 1]$. By definition, $\mathbf{a} \leq \mathbf{d}_\infty^2$. Therefore, $H_{[.],\mathbf{a}}(\delta^2, \mathcal{F}_K) \leq H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}_K)$. Recalling that $\mathcal{F} = \mathbb{S}_{B-1} \cap [e^{-\tau_n}, 1]^B$ and given Lemmas A.6, A.7 and A.8, we have

$$\begin{aligned} H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}_K) &\leq K H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathbb{S}_{B-1} \cap [e^{-\tau_n}, 1]^B) \\ &\leq K H_{[.],\|.\|_\infty}(\delta, [-\tau_n, 0]^B) \\ &\leq K \left(B \log(2) + B \left(\log\left(\frac{\tau_n}{\delta}\right)\right)_+\right) \end{aligned}$$

and $H_{[.],\mathbf{a}}(\delta, \mathcal{F}_K) \leq KB \left(\log(2) + \left(\log\left(\frac{\tau_n}{\delta}\right)\right)_+\right) \leq KB \left(\log(2\tau_n) + \log\left(\frac{1}{\delta}\right)\right)$ because $\delta \in (0, 1]$. $\qquad\square$

A.4.2 Bracketing and model dimension

Theorems 2.1 and 2.2 are obtained from Theorems A.1 and A.3 which address a penalty function related to geometrical properties of the models, namely bracketing entropy with respect to some distance $\mathbf{a}$. Recall that the function $\sigma \mapsto \int_0^\sigma H_{[.],\mathbf{a}}(\delta, \mathcal{S}_m)^{1/2} d\delta$ always satisfies Assumption A.2. The quantity $\mathfrak{D}_m$ is defined as $n\sigma_m^2$ where σ_m^2 is the unique root of $\phi_m(\sigma)/\sigma = n^{1/2}\sigma$. A good choice of ϕ_m is one which leads to a small upper bound of $\mathfrak{D}_m$. Although σ_m is not very explicit, it can be related to an entropic dimension of the model.

Define the bracketing dimension D_m of a compact set as the smallest real number D such that there exists a constant C such that

$$\forall \delta > 0, H_{[.],\mathbf{a}}(\delta, \mathcal{F}_m) \leq D(C + \log\left(\frac{1}{\delta}\right)).$$

In a parametric setting, the bracketing dimension is equivalent to the number of parameters to be estimated within a model. The following result from Cohen and Le Pennec (2011) states that under some assumption on the bracketing entropy, $\mathfrak{D}_m$ is proportional to the entropic dimension D_m.

Proposition A.3. *Assume for any $\delta \in (0,1]$, there exist $D_m > 0$ and $C_m \geq 0$ such that*

$$H_{[.],\mathbf{a}}(\delta, \mathcal{F}_m) \leq D_m\left(C_m + \log\left(\frac{1}{\delta}\right)\right).$$

Then, the function

$$\phi_m(\sigma) = \sigma D_m^{1/2}\left(C_m^{1/2} + \pi^{1/2} + \log\left(\frac{1}{\sigma \wedge e^{-1/2}}\right)^{1/2}\right)$$

satisfies the properties required in Assumption A.2 and $\mathfrak{D}_m$ satisfies

$$\mathfrak{D}_m \leq \left(2\left(C_m^{1/2} + \pi^{1/2}\right)^2 + 1 + \log\left(\frac{n}{e(C_m^{1/2} + \pi^{1/2})^2 D_m}\right)_+\right) D_m.$$

A.4.3 Proof of Theorem 1

Proposition A.2 directly indicates that if we choose the constants $D_m = KB$ and $C_m = \log(2\tau_n)$ and apply Proposition A.3, then the given ϕ_m satisfies Assumption A.2 in our setting. Therefore, we can apply Theorem A.1 and get the following oracle inequality:

$$E\left[\sum_{l=1}^L \frac{n_l}{n}\mathbf{KL}(s_l, \widehat{f_{\widehat{k_l}}})\right] \leq C_1\left(\inf_{\substack{f \in \mathcal{F}_K \\ (k_l)_l \in [\![1..K]\!]^L}}\left(\sum_{l=1}^L \frac{n_l}{n}\mathbf{KL}(s_l, f_{k_l})\right) + (2 + \kappa_0)\frac{L\log K}{n} + \kappa_0\frac{\mathfrak{D}_m}{n}\right)$$
$$+ \frac{C_2}{n} + \frac{\eta}{n}.$$

By definition, $\mu_n = 2\left(\log(2\tau_n)^{1/2} + \pi^{1/2}\right)^2 + 1 + \log(n)$. According to Proposition A.3, we also have

$$E\left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f_{\widehat{k_l}}})\right] \leq C_1 \left(\inf_{\substack{f \in \mathcal{F}_K \\ (k_l)_l \in [\![1..K]\!]^L}} \left(\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l})\right) + (2+\kappa_0)\frac{L \log K}{n} + \frac{\kappa_0 \mu_n D_K}{n}\right)$$
$$+ \frac{C_2}{n} + \frac{\eta}{n}.$$

It remains to choose $\lambda_0 = 2 + \kappa_0$ to conclude the proof.

A.4.4   Proof of Theorem 2

We need to find the weights x_m satisfying Assumption A.4 in order to apply Theorem A.3 in our framework. It is easy to show that $x_m \geq K \log(2)$ is a sufficient condition on the weights. Indeed, define $\delta = 1/2$. Then, $e^{-x_m} \leq \delta^K$. With a collection $\mathcal{M} = \mathbb{N} \setminus \{0\}$, we have

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \sum_{K \geq 1} \delta^K = \frac{\delta}{1-\delta} = 1.$$

We thus take $x_m = K \log(2)$. Then we have $\sum_{m \in \mathcal{M}} e^{-x_m} \leq 1$. We can apply Theorem A.3 and use Proposition A.3 to state that for any chosen $\lambda_0' > \kappa_0 + 1$, if

$$\mathbf{pen}(K) = \lambda_0'(\mu_n D_K + L \log(K) + x_m),$$

then the following inequality is satisfied:

$$E\left[\sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, \widehat{f_{\widehat{k_l}}})\right] \leq C_1 \inf_{K \in \mathcal{M}} \left(\inf_{\substack{f \in \mathcal{F}_K \\ (k_l)_l \in [\![1..K]\!]^L}} \sum_{l=1}^{L} \frac{n_l}{n} \mathbf{KL}(s_l, f_{k_l}) + \frac{\mathbf{pen}(K)}{n}\right)$$
$$+ \frac{C_2}{n} + \frac{\eta + \eta'}{n}.$$

This concludes the proof.

A.4.5   Varying number of categories

In this paragraph, the number B of different multinomial parameters varies besides the number of clusters K. We consider an ordered set of $B_{\max} \geq B$ categories where the first B categories can have different proportions that sum up to less than 1, whereas the others are uniformly distributed over the remaining probability. We still assume the lower bound of Assumption A.1 on density distributions. Thus, in this case, the set of density functions to be considered is defined by:

$$\mathcal{F}_{(K,B)} = \mathcal{F}^K \quad \text{with}$$
$$\mathcal{F} = \{(f(b))_{1 \leq b \leq B_{\max}} \mid \sum_{b=1}^{B_{\max}} f(b) = 1, e^{-\tau_n} \leq f(b) \leq 1, f(B+1) = \dots = f(B_{\max})\}.$$

In this framework, the following bound on the bracketing entropy can be stated:

Proposition A.4 (**Bracketing entropy of $\mathcal{F}_{(\mathbf{K},\mathbf{B})}$**)**.** *For all $\delta \in (0,1]$, we have*

$$H_{[.],\mathbf{a}}(\delta, \mathcal{F}_{(K,B)}) \leq K(B+1)\left(\log(2\tau_n) + \log\left(\frac{1}{\delta}\right)\right).$$

Proof. Since $\mathbf{a} \leq \mathbf{d}_\infty^2$, we still have $H_{[.],\mathbf{a}}(\delta^2, \mathcal{F}_{(K,B)}) \leq H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}_{(K,B)})$. Following Lemma A.6, we also have

$$H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}_{(K,B)}) \leq KH_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F})$$

Remarking that $B+1$ parameters determine elements of $\mathcal{F}$ (B first multinomial parameters plus one that is uniformly distributed over the remaining categories), Lemmas A.7 and A.8 lead to

$$\begin{aligned}
H_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}_{(K,B)}) &\leq KH_{[.],\mathbf{d}_\infty^2}(\delta^2, \mathcal{F}) \\
&\leq KH_{[.],\|.\|_\infty}(\delta, [-\tau_n, 0]^{B+1}) \\
&\leq K\left((B+1)\log(2) + (B+1)\left(\log\left(\frac{\tau_n}{\delta}\right)\right)_+\right),
\end{aligned}$$

and $H_{[.],\mathbf{a}}(\delta, \mathcal{F}_K) \leq K(B+1)\left(\log(2) + \left(\log\left(\frac{\tau_n}{\delta}\right)\right)_+\right) \leq K(B+1)\left(\log(2\tau_n) + \log\left(\frac{1}{\delta}\right)\right).$ $\qquad \square$

Theorem A.3 can thus be applied to this case, provided that the Kraft type assumption A.4 is satisfied. A model is defined as $m = (K, B)$ with $(K, B) \in \mathcal{M}$, $\mathcal{M}$ being the following collection of models:

$$\mathcal{M} = \{(1,1)\} \cup \mathbb{N} \setminus \{0\} \times \mathbb{N} \setminus \{0,1\}.$$

As a matter of fact,

Proposition A.5. *If $x_m \geq KB\log(2)$, then Kraft Assumption A.4 is satisfied and $\sum_{m \in \mathcal{M}} e^{-x_m} \leq 1$.*

Proof. Define $\delta = 1/2$. Then, $e^{-x_m} \leq \delta^{KB}$ and

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq \delta + \sum_{K \geq 1, B \geq 2} \delta^{KB} = \delta + \sum_{B \geq 2} \frac{\delta^B}{1 - \delta^B} \leq \delta + \sum_{B \geq 2} \frac{\delta^B}{1 - \delta} = \delta + \frac{\delta^2}{1 - \delta} = 1.$$

$\qquad \square$

The shape of the penalty thus obtained is given by $\mathbf{pen}(K, B) = \lambda_0'(\mu_n K(B+1) + L\log(K) + KB\log(2))$ with $\lambda_0' > \kappa_0 + 1$.