Contents

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

June 28, 2017

Abstract

In this study, we consider unsupervised clustering of categorical vectors that can be of different dimensions. We use a likelihood approach to estimate the parameters of the underlying mixture model. A penalized model selection technique selects the best model within a collection in a non-asymptotic framework. Regardless of the true distribution that generated the data, we show that under weak assumptions on the penalty choice, an oracle inequality with the same Kullback-Leibler divergence on the two sides of the inequality is satisfied. A numerical application to document clustering is studied, where mixture parameters are estimated via a new robust EM. Slope heuristics are used to calibrate an optimal penalty and select the best model accordingly.

1 Introduction

This study explores unsupervised clustering and model selection for multidimensional categorical data. Assume we observe L random vectors $(X_l)_{1\leq l\leq L}$, where each X_l corresponds to n_l independent and identically distributed (iid) instances of a categorical variable. Since L may be big, we assume the observations to be structured into K subgroups and to be modeled by a multinomial mixture. Our goal is to simultaneously estimate the proper number of groups and the underlying probability laws that generated these vectors.

Gassiat, Rousseau and Vernet [GRV16] also studied mixtures of multidimensional vectors in the case of independent and continuous marginal densities, but 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.

Clustering of categorical data is used in many fields such as social sciences, health, genetics, text analysis, etc. An EM algorithm is commonly used to estimate the best parameters in a given model [MP00], but there is no direct method to do model selection. Several techniques have been studied to select the number of clusters efficiently. Silvestre, Cardoso and Figueiredo [SCF14] 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 EM algorithm. Through this method, they optimize time-consuming computation and avoid several runs of the same algorithm. On the same vein, Rigouste, Cappé and Yvon [RCY06] 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. Fop, Smart and Murphy [FSM17] use a Bayes factor criterion to select the most discriminative symptoms for low back pain diagnosis. Matias, Rebafka and Villers [MRV15] 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.

We consider a penalized maximum likelihood approach, where the penalty adresses the problem of selecting the best multinomial mixture within a collection and the best assignment group of the vectors underlying the selected mixture. We give a theoretical sense of the classical trade-off that exists between the bias term and the variance term which depends on model complexity. The main result of this paper is a non-asymptotic oracle inequality. It 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 [TG09] and [BT13] in the context of genomics, where they cluster a population into subpopulations according to their alleles' categories. The observations studied in their analysis are considered as iid 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.

Our work is built on Massart's methodology [Mas07] for the computation of the penalty. The process of our proof is based on the technical report of [CP11] (see also [CP12]) 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 observation, we place ourselves in a fixed design setting with deterministic covariates, although our result can easily be generalized to random covariates cases. Under a weak assumption on the structure of the models, 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 the previous works (see [Mas07], [MM08a], [CP12], [BT13], [MP14]).

The model selection we propose naturally depends on the dimension of the model which is twofold. It involves the dimension of the parameters' space on the one hand, and the cluster assignment complexity on the other hand. Our model selection criterion takes these two steps into account as it penalizes high dimensional models and their corresponding assignment complexity. We show that it estimates well the mixture parameters, and that clustering association of the vectors is robust whatever the size of the data.

Section 2 presents the multinomial mixture framework and the penalized maximum likelihood technique for model selection in our paradigm. A theoretical risk bound for maximum likelihood estimation is established for a single model in Section 2.3 and is extended to the multiple model case in Section 2.4. Section 3 is devoted to a practical application of our result in which the dataset corresponds to a corpus of research articles published in NIPS journal. Mixture parameters are estimated thanks to a new robust EM that annihilates underrepresented clusters, and an optimal penalty is calibrated using slope heuristics. All the technical results are proved at the Appendix.

2 Model Selection Criterion

2.1 Framework and notations

Consider a family of L independent random vectors $(X_l)_{1\leq l\leq L}$, where each X_l represents n_l iid instances of a random variable that has s_l as a true categorical density distribution with respect to a known positive measure. The studied sample is assumed to originate from a certain number K_m of sub-groups, where each cluster is characterized by its own density. The latent sub-group a vector comes from is modeled by a K_m-dimensional multinomial variable Z that takes values in the set $\{1, ..., k, ..., K_m\}$ of the different group labels. The distribution of Z is given by the vector $\pi^m := (\pi_k^m)_{1\leq k\leq K_m}$, where $\pi_k^m = \mathbb{P}(Z = k)$ belongs to the $(K_m - 1)$-dimensional simplex $\mathbb{S}_{K_m-1}$. In fact, each X_l follows a mixture distribution given by:

$$\sum_{k=1}^{K_m} \pi_k^m \mathbb{P}(X_l \mid Z_l = k).$$

Thus, the probability distribution of an observation $x_l = (x_l^i)_{1\leq i\leq n_l}$ within a group k is given by:

$$\begin{aligned} \mathbb{P}(x_l \mid Z_l = k) &= \prod_{i=1}^{n_l} \mathbb{P}(x_l^i \mid Z_l = k), \\ \mathbb{P}(x_l^i \mid Z_l = k) &=: f_k^m(x_l^i), \end{aligned} \tag{1}$$

where $f_k^m(b)$ is the probability of falling in category b given group k. Therefore, the likelihood of one observation can be expressed as follows:

$$P_{(K_m,B_m,\theta_m)}(x_l) = \sum_{k=1}^{K_m} \pi_k^m \left(\prod_{i=1}^{n_l} f_k^m(x_l^i) \right), \tag{2}$$

where $\theta_m := (\pi^m, f^m)$ corresponds to the parameters of one model with:

$$\begin{aligned} \pi^m &:= (\pi_k^m)_{1\leq k\leq K_m} \\ f^m &:= (f_k^m(b))_{\substack{1\leq k\leq K_m \\ 1\leq b\leq B_m}}. \end{aligned}$$

For a given couple (K_m, B_m) of K_m clusters and B_m categories, a parameter $\theta_m = (\pi^m, f^m)$ ranges in the set:

$$\Theta_m := \mathbb{S}_{K_m-1} \times \mathcal{F}_m,$$

where $\mathcal{F}_m$ refers to a set of categorical functions that will be defined later on. The collection of models that will be considered is denoted by

$$\mathcal{S}_m := \{P_{(K_m,B_m,\theta_m)} : \theta \in \Theta_m\},$$

with $(K_m, B_m) \in \mathcal{M}$, $\mathcal{M}$ being the collection of models

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

2.2 Maximum Likelihood Estimation

The parameter estimation is twofold. First, the global parameters of the mixture are obtained according to a maximum likelihood estimator (MLE). Then, each vector is affected to its closest cluster.

Denote by $n := \sum_l n_l$ the overall number of instances that are observed. We define as $\gamma_n(\pi^m, f^m)$ the empirical contrast of the observations also known as a MLE:

$$\gamma_n(\pi^m, f^m) := \sum_{l=1}^{L} -\log\left(P_{(K_m,B_m,\theta_m)}(x_l)\right). \tag{3}$$

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

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

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

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

with $\eta > 0$.

Since MLE 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 (Model Structure - $\mathbf{MS_m}$). *For all candidate density f taken in any model,*

$$\epsilon_n := e^{-\tau_n} \leq f.$$

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 the categories that are considered correspond to words that appeared in at least one text. More generally, we can use this lower-bound in the case of continuous densities with compact support. However, this assumption is no longer applicable to densities that can take extreme values, because these cannot be controlled in the estimation. Under Assumption ($\mathbf{MS_m}$), we can now define the set $\mathcal{F}_m$ of K_m density functions on B_m categories. It can be viewed as a product set that will be denoted by:

$$\mathcal{F}_m := \mathcal{F}_{m,1} \times ... \times \mathcal{F}_{m,K_m} \quad \text{with}$$
$$\mathcal{F}_{m,k} := \{(f_k(b))_{1\leq b\leq B_m} \mid \sum_{b=1}^{B_m} f_k(b) = 1, f_k(b) \geq e^{-\tau_n}\}, k \in \{1, ..., K_m\}.$$

Once parameters are estimated, for $l = 1, ..., L$, the observation x_l is assigned to the cluster it most likely belongs to. We use the Maximum a Posteriori (MAP) method:

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

We note $\widehat{f_{\widehat{k_l}}}$ the estimated distribution of observation x_l.

2.3 Single Model Risk Bound

In this section, a theoretical bound for a Kullback-Leibler type loss is addressed, which introduces a penalized model selection technique for choosing a best model within a collection. In penalized maximum likelihood approach, the selected model $(\widehat{K_m}, \widehat{B_m})$ is a minimizer of the following penalized criterion:

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

where $\mathbf{pen} : \mathcal{M} \to \mathbb{R}^+$ denotes the penalty function. It depends on model complexity which involves a technical notion of bracketing entropy. Under some general assumption on this bracketing entropy, it has been shown in [CP12] that model complexity is proportional to the dimension of the model. In our case, the dimensions of the considered models correspond to the number of mixture parameters plus the cluster assignment cost. By denoting $D_{(K_m,B_m)}$ the number of parameters that determine a mixture model with K_m categorical functions of $\mathcal{F}_m$, we have

$$D_m = K_m(B_m - 1) + (K_m - 1) = K_m B_m - 1.$$

The main result of our study suggests a penalty of the form $\mathbf{pen}(K_m, B_m) = \lambda_0(D_m + L\log(K_m))$, where D_m plays the role of the mixture model complexity, and $L\log(K_m)$ corresponds to the cluster assignment cost. This relates to classical penalties such as AIC and BIC functions that are proportional to the 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.

The following theorem states a theoretical risk bound for a single model, which penalizes too complex models according to the number of observations. It turns out to be crucial in model selection, as model complexity appears explicitly and will be used to construct the penalty.

Theorem 2.1. *Consider the observed vectors $((x_l^i)_{1\le i\le n_l})_{1\le l\le L}$ as above, and denote by $n := \sum_l n_l$ the overall number of observations. Consider also one model $\mathcal{S}_m$ as defined above and assume it satisfies* **(MS$_\mathbf{m}$)**. *Let $(\widehat{\pi^m}, \widehat{f^m})$ be a η-MLE in $\mathcal{S}_m$:*

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

and define the corresponding cluster assignment for each vector $x_l := (x_l^i)_{1\le i\le n_l}$:

$$\widehat{k_l} = \operatorname*{argmax}_{1\le k\le K_m} \left\{ \widehat{\pi_k^m} \left(\prod_{i=1}^{n_l} \widehat{f_k^m}(x_l^i) \right) \right\}.$$

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

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^m_{\widehat{k_l}}})] \le C_1 \left(\inf_{(k_l)_l\in\{1,\dots,K_m\}^L} \inf_{f^m\in\mathcal{F}_m} \left\{ \frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l}) \right\} + \lambda_0 \left(\frac{L\log K_m}{n} + \frac{D_m}{n} \right) \right) + \frac{C_2}{n} + \frac{\eta}{n}.$$

In previous works (see [Mas07], [CP12], [BT13], [MM08a]), a lower divergence appeared 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 **KL** 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, as risk functions are the same on the left and on the right side of the inequality. This is made possible thanks to Assumption **(MS$_\mathbf{m}$)**, 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 **(MS$_\mathbf{m}$)**, density distributions must be weighted by the true ones in order to bound a log-ratio which leads to a weaker divergence than **KL** ([Mas07], [CP12]).

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 $L\log(K_m)$ appearing on the left side of the inequality. It penalizes the cost for assigning each observation to one cluster, as D_m penalizes complex mixture models.

2.4 Model Selection Theorem

The main result of this study answers the natural question of the choice of the model. Instead of one model $\mathcal{S}_m$, we have a collection of models $(\mathcal{S}_m)_{m\in\mathcal{M}}$ that all satisfy Assumption $(\mathbf{MS_m})$. Then, by applying the previous theorem, we have the following inequality:

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^m_{\widehat{k}_l}})] \leq C_1 \left(\inf_{(k_l)_l \in \{1,\dots,K_m\}^L} \inf_{f^m \in \mathcal{F}_m} \left\{ \frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l}) \right\} + \lambda_0 \left(\frac{L\log K_m}{n} + \frac{D_m}{n} \right) \right) + \frac{C_2}{n} + \frac{\eta}{n}.$$

One of the models minimizes the right hand side of the inequality, but there is no way of detecting which one without knowing the oracle densities $(s_l)_{1\leq l\leq L}$. A data-driven strategy is therefore proposed to select an estimate among the collection according to a selection rule that performs almost as well as if we had known the true family of densities.

The following theorem compares the selected estimator with the risk of this estimator associated with the best model and shows that under some condition on the penalty function, the estimated model still performs almost as well as the best one. The penalty function is naturally increasing with respect to the model complexity, as simpler models are preferable to the true one which may overfit the data.

Theorem 2.2. *Consider the observed vectors $((x_l^i)_{1\leq i\leq n_l})_{1\leq l\leq L}$ as above, and denote by $n := \sum_l n_l$ the overall number of observations. Consider also the collection $(\mathcal{S}_m)_{m\in\mathcal{M}}$ of models as defined above. Let $(\widehat{\pi^m}, \widehat{f^m})_{m\in\mathcal{M}}$ be the corresponding collection of η-MLEs:*

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

Define the corresponding cluster assignment for each vector $x_l := (x_l^i)_{1\leq i\leq n_l}$:

$$\widehat{k_l^m} = \operatorname*{argmax}_{1\leq k\leq K_m} \left\{ \widehat{\pi_k^m} \left(\prod_{i=1}^{n_l} \widehat{f_k^m}(x_l^i) \right) \right\}.$$

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:

$$\begin{aligned} \mathbf{pen}: \quad \mathcal{M} &\to \mathbb{R}^+ \\ (K_m, B_m) &\mapsto \lambda_0' \left(D_m + L\log(K_m) + K_m B_m \log(2) + 2(\sqrt{\log(2)} + \sqrt{\pi})^2 + 1 + \log(n) \right) \end{aligned}$$

for all $(K_m, B_m) \in \mathcal{M}$, then $\widehat{m} := (\widehat{K_m}, \widehat{B_m})$ exists, where $(\widehat{K_m}, \widehat{B_m})$ minimizes the penalized log-likelihood criterion:

$$\mathbf{crit}(K_m, B_m) := \gamma_n(\widehat{\pi^m}, \widehat{f^m}) + \mathbf{pen}(K_m, B_m).$$

Moreover, whatever the underlying true densities $(s_l)_{1\leq l\leq L}$,

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

The additional terms appearing in the penalty function play the role of a union bound on the collection of models and enables the oracle inequality to simultaneously apply for all models. It results from a Kraft type assumption that is not intrinsic to the models but can be chosen in a way such that it is small compared to the model complexity $D_m + L\log(K_m)$. Furthermore, the penalty term added to the likelihood estimate compensates for both the variance term and the bias between the empirical **KL**-divergence and the true infimum of **KL** within the model. For a given penalty, the selected model is the one that minimizes the penalized criterion as defined over the collection.

3 Model Selection on a text corpus

3.1 Clustering algorithm

The dataset "NIPS Conference Papers 1987-2015 Data Set" contains distribution of words used in NIPS conference papers published from 1987 to 2015 [PJST16]. Based on a dictionary of 11463 unique words that appear in 5811 conference papers, the dataset is represented as a 11463×5811 matrix. 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 [RCY06]. Moreover, 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. Our 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. The selected model must keep a good balance between bias and variance.

Although time structure of the corpus is ignored in our model selection, it is analyzed as a prediction in Section 3.3. One could also consider the corpus as a spatialized mixture model (see [CP14]) with mixture proportions modeling time structure in the articles. In [PJST16], a probabilistic model for time-dependent data is used and applied to NIPS corpus.

Mixture parameters are estimated thanks to the EM algorithm. The EM heavily depends on its initial parameters, so the log-likelihood often converges to a local maxima. We ran 500 short EMs 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 dimension (8999 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 EMs (with 15 iterations by default) from randomly chosen initializing parameters and run a long EM from the most performing parameter in terms of likelihood [TG09].

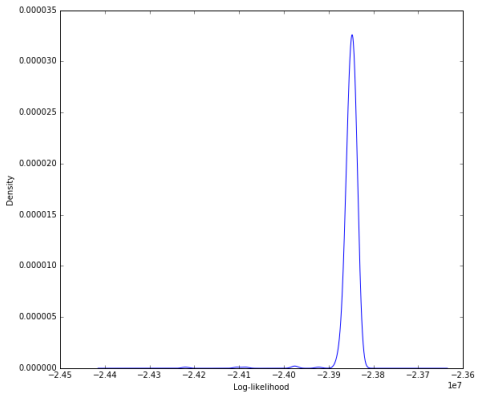


Figure 1: Kernel Density Estimation of log-likelihood distribution based on 500 random initializing parameters after 15 iterations of EM (K = 30, B = 300)

The EM 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 [FJ02]. This especially occurs in high dimension. B. Zhang, C. Zhang and Xing Yi [ZZY04] propose 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 [FJ02] that component annihilation leads to more robust results. This observation introduced the EM-MML-algorithm (see [SCF14]) that starts from a high number of clusters. 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 [YLL12] 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.

Besides parameter estimation, we want to calibrate our theoretical penalty before selecting a model according to the penalized criterion thus obtained. As will be further explained, a range of models are explored to calibrate

the penalty. For each model, clusters with low proportion estimate 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, due to its high variance, a 100 component mixture often performs as well as a well-balanced 50 component mixture. Algorithm 1 shows the detailed pseudocode of the algorithm. 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.

Algorithm 1 Robust EM algorithm

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

procedure ROBUSTEM
  $t \leftarrow 0$
  $k_{\text{current}} \leftarrow k_{\max}$
  $\pi_{\text{threshold}} \leftarrow \frac{1}{100 \cdot k_{\text{current}}}$
  $\mathcal{L}^{(0)} \leftarrow -\infty$
  for i = 1 to 15 **do**
    $\pi_i^{(0)}, f_i^{(0)} \leftarrow \text{INITIALIZE}(\pi, f)$
    $\widehat{\pi}_i^{(10)}, \widehat{f}_i^{(10)}, \widehat{\mathcal{L}}_i^{(10)} \leftarrow \text{SHORTEM}(k_{\text{current}}, \pi_i^{(0)}, f_i^{(0)})$
  $t \leftarrow 10$
  $i_{\max} \leftarrow \text{argmax}_i \, \widehat{\mathcal{L}}_i^{(10)}$
  while $\widehat{\pi}_{i_{\max}}^{(t)} < \pi_{\text{threshold}}$ **do**
    $k_{\text{current}} \leftarrow \sharp\{k : \widehat{\pi}_{i_{\max}}^{(t)}(k) \geq \pi_{\text{threshold}}\}$
    $\pi_{\text{threshold}} \leftarrow \frac{1}{100 \cdot k_{\text{current}}}$
    $\widehat{\pi}_{i_{\max}}^{(t+u)}, \widehat{f}_{i_{\max}}^{(t+u)}, \widehat{\mathcal{L}}_{i_{\max}}^{(t+u)} \leftarrow \text{EM}(k_{\text{current}}, \pi_{i_{\max}}^{(t)}, f_{i_{\max}}^{(t)})$
    $t \leftarrow t + u$
  $k_{\text{opt}} \leftarrow k_{\text{current}}$

3.2 Slope heuristics

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, Maugis and Michel [BMM10] 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 given by:

$$\mathbf{pen}_{\text{opt}}(m) = 2\lambda_{\min} D_m, \tag{4}$$

with D_m the model dimension. The derivation of this formula is detailed in [MM08b] and [BMM10].

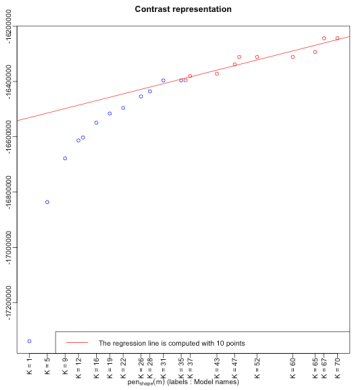
We use slope heuristics to calibrate penalty for NIPS data. The practical methodology and the visualizing graphics were implemented thanks to the R package "**capushe**".

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 [BMM10]. Figure 2 shows linear regression of the log-likelihood with respect to the selected number of points. It corresponds to a slope of $\widehat{\lambda}_{\min} \approx 15$. The model that minimizes the resulting criterion is given by $\widehat{K} = 31$ clusters.

3.3 Back-mapping of NIPS topics

Based on the selected model of $\widehat{K} = 31$ clusters, we analyze the evolution of representative NIPS topics overtime. For all articles published in a given year, average posterior probabilities are computed over the

Figure 2: Slope heuristics on NIPS data



clusters. This gives a year-scaled evolution of topics from 1985 to 2015. Figure **??** shows how clusters evolved over the years. Some representative clusters were highlighted in Figure **??**. The word cloud of their corresponding categorical distribution figuring on the left side gives a hint on the subject they deal with.

4 Conclusion

In this paper, the problem of estimating several distributions of different-sized categorical variables has been introduced. An approximation with a categorical mixture model has been proposed and a theoretical oracle inequality has been established, that measures the efficiency of both parameter estimation and cluster assignment of the observations. Under a weak assumption, the model risk and the expected empirical risk have been compared under the same divergence, which lead to a sharper inequality. This result has been extended to study the risk for a collection of models and enabled to explicit a non-asymptotic penalty function for model selection. Under the same theoretical assumptions, the same kind of result can be obtained in the case of continuous variables if we approximate them by piecewise constant functions.

Model selection through penalized contrast has been applied to a text corpus of NIPS conference papers. We have introduced an adjusted EM algorithm that lead to more robust and reliable estimations. These estimates have been used to calibrate the constant appearing in the theoretical penalty. An optimal penalty function has been deduced thanks to slope heuristics. Finally, we visualized the obtained clusters and noticed an actual evolution of NIPS topics overtime since 1985.

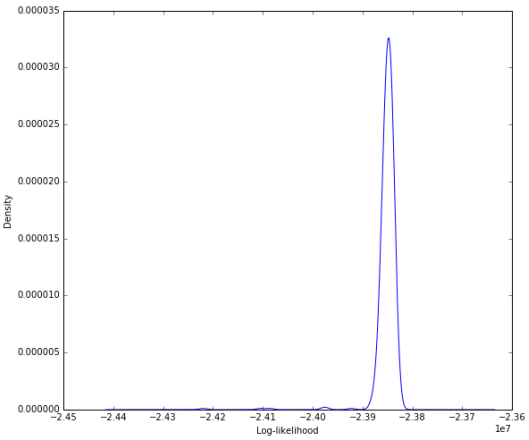


Figure 3: toto

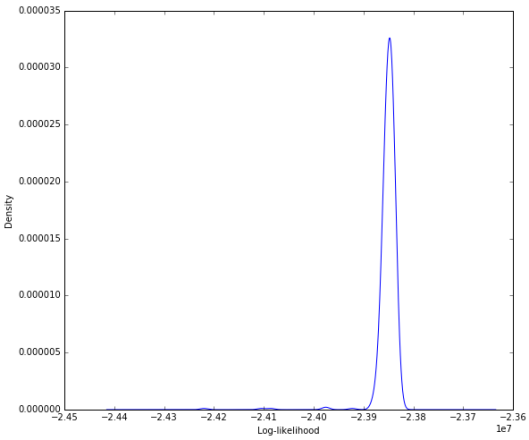


Figure 4: titi

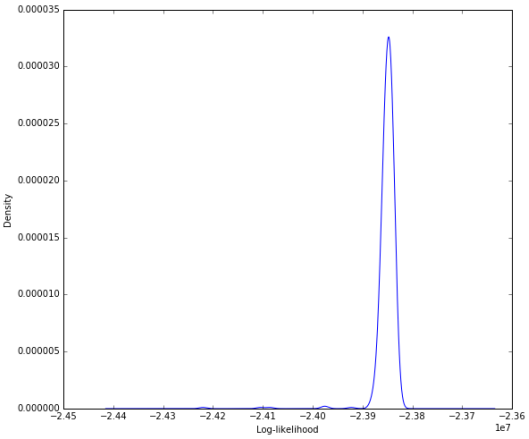


Figure 5: tata

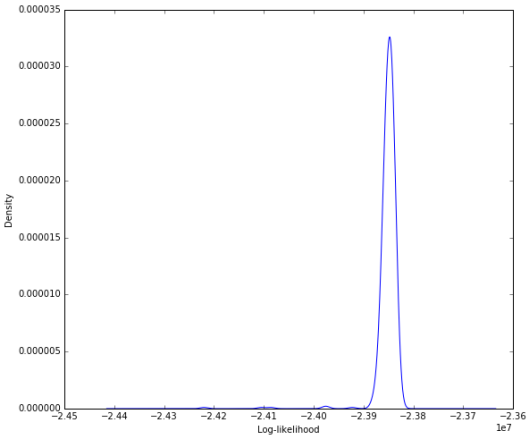


Figure 6: tutu

5 Proofs for Section 2

5.1 Preliminary

In this section, Theorems 2.1 and 2.2 are proved in a more general setting, where some assumptions are relaxed. Denoting by s_l the true density of observation l, we assume the structure of the models to be such that:

Assumption (Generalized Model Structure - $\mathbf{GMS_m}$). *For all $k, k' \in \{1, ..., K\}$ and $l \in \{1, ..., L\}$, any density function $f_k^m \in \mathcal{F}_{m,k}, f_{k'}^m \in \mathcal{F}_{m,k'}$ taken in the model satisfies*

(i) $e^{-\tau_n} \leq \frac{f_k^m}{f_{k'}^m} \leq e^{\tau_n}$

(ii) $e^{-\tau_n} \leq \frac{f_k^m}{s_l}$, *$s_l$ being the true density of observation l.*

Remark that this assumption implies the Assumption ($\mathbf{MS_m}$) described above. Moreover, the second part of the assumption is made on the model density rather than the true one.

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 $x \in \mathcal{X}$, $f_k^-(x) \leq f_k^+(x)$. A density function f_k is said to belong to the bracket $[f_k^-, f_k^+]$ if $f_k^-(x) \leq f_k(x) \leq f_k^+(x)$ for all $x \in \mathcal{X}$. Take $f^-, f^+ \in \mathcal{F}_m$ such that for all $k \in \{1, ..., K\}$, $[f_k^-, f_k^+]$ forms a bracket. Fix $(k_l)_{1 \leq l \leq L}$ a cluster assignment of the observations. We define the width of such a family of brackets as follows:

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

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:

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

Remark that the localized model includes the set product of the localized sets

$$\mathcal{F}_{m,k}(\tilde{f}_k, \sigma) := \{f_k \in \mathcal{F}_{m,k} \mid \mathbf{a}(\tilde{f}_k, f_k) \leq \sigma^2\}, k = 1, ..., K.$$

We also impose a structural assumption on the localized models:

Assumption ($\mathbf{H_m}$). *There exists a real-valued function ϕ_m on $[0, +\infty)$ such that*

(i) *ϕ_m is non-decreasing*

(ii) *$\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0, +\infty)$*

(iii) *For every $\sigma \geq 0$ and every $f \in \mathcal{F}_m$*

$$\int_0^\sigma \sqrt{H_{[.],\mathbf{a}}(\delta, \mathcal{F}_m(f, \sigma))} d\delta \leq \phi_m(\sigma).$$

The use of the divergence **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, as it can be deduced from Meynet's result [Mey12]:

Proposition 5.1. *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*

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

Indeed, by taking $\tau := \tau_n$, $dP := f_{k_l} d\mu$ and $dQ := s_l d\mu$, one can deduce that

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

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

Assumption ($\mathbf{Sep_m}$). *There exists a countable subset $\mathcal{F}'_m$ of $\mathcal{F}_m$ and a set $\mathcal{X}'_m$ with $\mu(\mathcal{X} \setminus \mathcal{X}'_m) = 0$ such that for every $f \in \mathcal{F}_m$, for all $k = 1, ..., K$, there exists a sequence $(f_{k,j})_{j\geq 1}$ of elements of $\mathcal{F}'_{m,k}$ such that for every $x \in \mathcal{X}'_m$, $\log(f_{k,j}(x))$ goes to $\log(f_k(x))$ as j goes to infinity.*

5.2 Single Model Risk Bound

Theorem 5.1. *Let $(X_l)_{1\leq l\leq L}$ be independent random vectors where each X_l consists of n_l independent and identically distributed (iid) instances of a multinomial vector that has s_l as a true density with respect to some positive measure μ. Assume $\mathcal{S}_m$ is a model for which Assumptions ($\mathbf{GMS_m}$), ($\mathbf{H_m}$) and ($\mathbf{Sep_m}$) hold. Let $(\widehat{\pi}, \widehat{f}) \in \Theta_m$ be a η-log-likelihood minimizer in $\mathcal{S}_m$:*

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

and define its corresponding assignment for each vector X_l:

$$\widehat{k_l} = \operatorname*{argmax}_{k\in\{1,...,K\}} \left\{ \widehat{\pi_k} \left(\prod_{i=1}^{n_l} \widehat{f}_k(X_l^i) \right) \right\}.$$

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 $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$, the estimate $(\widehat{\pi}, \widehat{f})$ satisfies

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

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

$$\begin{aligned} \eta + \inf_{(\pi,f)\in\Theta_m} \gamma_n(\pi, f) &\geq \sum_{l=1}^{L} -\log\left(K\widehat{\pi}_{\widehat{k_l}} \left[\prod_{i=1}^{n_l} \left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l} \right)(X_l^i) \right] \right) \\ &\geq -L\log(K) - \sum_{l=1}^{L}\sum_{i=1}^{n_l} \log\left(\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l} \right)(X_l^i) \right). \end{aligned} \tag{5}$$

On the other hand, for any family $(k_l)_{1\leq l\leq L}$ of cluster assignment,

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

because $-\log$ is non-increasing. Therefore, by inequality (5),

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

which leads to

$$\sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{\widehat{f}_{\widehat{k_l}}}{s_l} \right)(X_l^i) \right) \leq L\log(K) + \inf_{(k_l)_l\in\{1,...,K\}^L} \left\{ \inf_{(\pi,f)\in\Theta_m} \left\{ \sum_{l=1}^{L} -\log(\pi_{k_l}) + \sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{f_k}{s_l} \right)(X_l^i) \right) \right\} \right\} + \eta. \tag{6}$$

We define by $\bar{f} := (\bar{f}_1, ..., \bar{f}_K)$ a family of densities that satisfy for all $\delta > 0$ and all cluster assignment $(k_l)_{1\leq l\leq L}$:

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

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

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

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

$$\nu_l(f_{k_l}) := \mathbf{kl}(f_{k_l})(X_l) - \mathbb{E}_{s_l}[\mathbf{kl}(f_{k_l})(X_l)] \tag{8}$$

the centered version of the empirical **KL**-divergence. According to equation (6) and using the definition of $\bar{f}$ in (7), we can write:

$$\begin{aligned}
\sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l}) + \sum_{l=1}^{L} \mathbb{E}_{s_l}[\mathbf{kl}(\widehat{f}_{\widehat{k}_l})(X_l)] &\leq L \log K + \inf_{(k_l)_l \in \{1,\dots,K\}^L} \Big\{ \inf_{\pi \in \mathbb{S}_{k-1}} \left\{ \sum_{l=1}^{L} -\log(\pi_{k_l}) \right\} + \sum_{l=1}^{L} -\log\left(\left(\frac{\bar{f}_{k_l}}{s_l} \right)^{\otimes n_l} (X_l) \right) \\
&\quad + \sum_{l=1}^{L} \left(\mathbb{E}_{s_l}\left[\sum_{i=1}^{n_l} -\log\left(\left(\frac{\bar{f}_{k_l}}{s_l} \right)(X_l^i) \right) \right] - \mathbb{E}_{s_l}\left[\sum_{i=1}^{n_l} -\log\left(\left(\frac{\bar{f}_{k_l}}{s_l} \right)(X_l^i) \right) \right] \right) \Big\} + \eta \\
&\leq \inf_{(k_l)_l \in \{1,\dots,K\}^L} \left\{ \inf_{\pi \in \mathbb{S}_{k-1}} \left\{ \sum_{l=1}^{L} -\log(\pi_{k_l}) \right\} + \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \bar{f}_{k_l}) \right\} \\
&\quad + L \log K + \eta,
\end{aligned}$$

which can be rewritten as:

$$\begin{aligned}
\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f}_{\widehat{k}_l}) &\leq L \log K + \inf_{(k_l)_l \in \{1,\dots,K\}^L} \left\{ \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{\pi \in \mathbb{S}_{k-1}} \left\{ \sum_{l=1}^{L} -\log(\pi_{k_l}) \right\} + \inf_{f \in \mathcal{F}_m} \left\{ \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) \right\} \right\} \\
&\quad - \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l}) + \delta + \eta \\
&\leq L \log K + \inf_{(k_l)_l \in \{1,\dots,K\}^L} \left\{ \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \sum_{l=1}^{L} -\log(\frac{1}{K}) + \inf_{f \in \mathcal{F}_m} \left\{ \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) \right\} \right\} \\
&\quad - \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l}) + \delta + \eta \\
&\leq 2L \log K + \inf_{(k_l)_l \in \{1,\dots,K\}^L} \left\{ \sum_{l=1}^{L} \nu_l(\bar{f}_{k_l}) + \inf_{f \in \mathcal{F}_m} \left\{ \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f_{k_l}) \right\} \right\} - \sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l}) + \delta + \eta.
\end{aligned}$$

It remains to get an upper bound of the deviation $-\sum_{l=1}^{L} \nu_l(\widehat{f}_{\widehat{k}_l})$, which is stated in the following lemmas.

Lemma 5.1. *Let $(k_l)_{1 \leq l \leq L} \in \{1, \dots, K\}^L$ be a group assignment for each vector X_l. Then, there exist three absolute constants $\kappa_0' > 4$, κ_1' and κ_2' such that, under Assumption ($\mathbf{H_m}$), for all $m \in \mathcal{M}$, for all $y_m > \sigma_m$ and every measurable event A such that $\mathbb{P}(A) > 0$,*

$$\mathbb{E}^A\Big[\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{l=1}^{L} \frac{-\nu_l(f_{k_l})}{y_m^2 + \kappa_0' \frac{1}{n} \sum_{l=1}^{L} n_l \mathbb{E}\left[|\log\left(\frac{f_{k_l}}{s_l} \right)|^2 \right]} \Big] \leq \kappa_1' \frac{\sigma_m}{y_m} + \frac{\kappa_2'}{\sqrt{n y_m^2}} \sqrt{\log\left(\frac{1}{\mathbb{P}(A)} \right)} + \frac{9\tau_n}{n y_m^2} \log\left(\frac{1}{\mathbb{P}(A)} \right).$$

Lemma 5.2. *Let Z be a random variable and Ψ a non-decreasing function such that for all measurable event A satisfying $\mathbb{P}(A) > 0$, $\mathbb{E}^A[Z] \leq \Psi\left(\log\left(\frac{1}{\mathbb{P}(A)} \right) \right)$. Then, for all $x \geq 0$, $\mathbb{P}(Z > \Psi(x)) \leq e^{-x}$.*

Using Lemma 5.1, for all $\lambda > 0$ we derive that:

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

Therefore, thanks to Lemma 5.2, for all $x>0$, except on a set of probability less than e^{-x},

$$\sum_{(k_l)_l\in\{1,\dots,K\}^L}\exp\left(\sup_{f\in\mathcal{F}_m}\frac{\lambda}{n}\sum_{l=1}^{L}\frac{-\nu_l(f_{k_l})}{y_m^2+\kappa_0'\frac{1}{n}\sum_{l=1}^{L}n_l\mathbb{E}\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}\right)\leq\sum_{(k_l)_l\in\{1,\dots,K\}^L}\exp\left(\frac{\lambda\kappa_1'\sigma_m}{y_m}+\frac{\lambda\kappa_2'}{\sqrt{ny_m^2}}\sqrt{x}+\lambda\left(\frac{9\tau_n}{ny_m^2}\right)\right)$$
$$\leq K^L\exp\left(\frac{\lambda\kappa_1'\sigma_m}{y_m}+\frac{\lambda\kappa_2'}{\sqrt{ny_m^2}}\sqrt{x}+\lambda\left(\frac{9\tau_n}{ny_m^2}\right)x\right).$$

A fortiori, we have

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

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

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

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

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

and obtain an upper bound for $\frac{1}{n}\sum_{l=1}^{L}-\nu_l(\widehat{f}_{\widehat{k}_l})$. By using equation (5.2),

$$\sum_{l=1}^{L}\frac{n_l}{n}\mathbf{KL}(s_l,\widehat{f}_{\widehat{k}_l})\leq\frac{2L\log K}{n}+\inf_{(k_l)_l\in\{1,\dots,K\}^L}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}_{k_l})+\inf_{f\in\mathcal{F}_m}\left\{\sum_{l=1}^{L}\frac{n_l}{n}\mathbf{KL}(s_l,f_{k_l})\right\}\right\}+\frac{\delta+\eta}{n}$$
$$+\left(y_m^2+\kappa_0'\frac{1}{n}\sum_{l=1}^{L}n_l\mathbb{E}\left[|\log\left(\frac{\widehat{f}_{\widehat{k}_l}}{s_l}\right)|^2\right]\right)\left(\frac{9\tau_n+1}{\theta^2}+\frac{\kappa_1'+\kappa_2'}{\theta}\right).$$

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 5.1, we obtain

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

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

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

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

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

Since $\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} \nu_l(\bar{f}_{k_l})] = 0$ for any family of clustering assignment $(k_l)_l \in \{1,\dots,K\}^L$, we derive:

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

and

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

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

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

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

5.2.1 Proof of Lemma 5.1

Proof. Consider a class of real-valued and measurable functions defined below. For a fixed family of K functions denoted by $\tilde{f} := (\tilde{f}_k)_{1 \leq k \leq K}$:

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

We are thus 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

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

and

$$g_{k_l}^+ - g_{k_l}^- = \log\left(\frac{f_{k_l}^+}{f_{k_l}^-} \right).$$

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

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

Recall Theorem 6.8 from [Mas07]:

Theorem 5.2. *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 $j \geq 2$,*

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

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 $k \geq 2$,

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

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 $\mathbb{P}(A) > 0$,

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

where $E = \frac{\kappa}{\epsilon}\frac{1}{\sqrt{n}}\int_0^{\epsilon\sqrt{V}}\sqrt{H(\gamma)\wedge n}d\gamma + \frac{2(b+\sqrt{V})}{n}H(\sqrt{V})$. 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 $b = \tau_n$. However, the set $\mathcal{G}(\tilde{f},\sigma)$ is not necessarily countable so the supremum $W(\tilde{f},\sigma)$ is not measurable. We rather define the countable subset:

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

with $\mathcal{F}'_m$ satisfying Assumption (**Sep$_\mathbf{m}$**). 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 5.2 to our context, we can conclude that

$$\mathbb{E}^A\left[\sup_{f\in\mathcal{G}(\tilde{f},\sigma)}\frac{1}{n}\sum_{l=1}^{L}(-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l}))\right] \leq E + \frac{(1+6\epsilon)\sqrt{2V}}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} + \frac{2b}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right),$$

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

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

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

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

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,

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

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, $\frac{\phi_m(\sigma_m)}{\sqrt{n}\sigma_m^2} = 1$. Thus, when $\sigma \geq \sigma_m$,

$$E \leq (\kappa + 2(b+\sigma))\frac{\phi_m(\sigma)}{\sqrt{n}} \leq (27 + 2(b+\sigma))\frac{\phi_m(\sigma)}{\sqrt{n}}.$$

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

$$\frac{1}{n}\sum_{l=1}^{L}\sum_{i=1}^{n_l}\mathbb{E}_{s_l}\left[|\log\left(\frac{f_{k_l}}{g_{k_l}}\right)|^2\right]\leq\frac{1}{n}\sum_{l=1}^{L}n_l\tau_n^2=\tau_n^2$$

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

$$\begin{aligned}\mathbb{E}^A[\sup_{f\in\mathcal{G}(\tilde{f},\sigma)}\frac{1}{n}\sum_{l=1}^{L}(-\nu_l(f_{k_l})+\nu_l(\tilde{f}_{k_l}))]&\leq(27+2(\tau_n+\sigma))\frac{\phi(\sigma)}{\sqrt{n}}+\frac{7\sqrt{2}}{\sqrt{n}}\sigma\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)}+\frac{2\tau_n}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)\\&\leq(27+4\tau_n)\frac{\phi(\sigma)}{\sqrt{n}}+\frac{7\sqrt{2}}{\sqrt{n}}\sigma\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)}+\frac{2\tau_n}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right)\end{aligned}$$

We now use the pealing lemma in order to bound the supremum on the overall model.

Lemma 5.3 (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*

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

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

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

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

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

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

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

We have chosen $(\tilde{f}_{k_l})$ such that for any (f_{k_l}) family of K-uplet and any $\epsilon_d>0$,

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

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

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

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

Lemma 5.4 (Bernstein Inequality)**.** *Assume there exist $V',b'\geq 0$ such that for all integer $j\geq 2$*

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

Then, for all measurable event A such that $\mathbb{P}(A)>0$,

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

This yields for all $l = 1, ..., L$:

$$\frac{1}{n}\sum_{l=1}^{L}\mathbb{E}^A[-\nu_l(\tilde{f}_{k_l})] \leq \frac{\sqrt{2V'}}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} + \frac{b'}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right),$$

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

$$\begin{aligned}\mathbb{E}^A[\frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + \kappa'^2\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2\right]}] &\leq \frac{1}{y_m^2 + \kappa'^2\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2\right]}\frac{\sqrt{2V'}}{\sqrt{n}}\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} \\ &+ \frac{1}{y_m^2 + \kappa'^2\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2\right]}\frac{b'}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right) \\ &\leq \frac{1}{\kappa'}\frac{\sqrt{2}}{\sqrt{ny_m^2}}\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} + \frac{1}{y_m^2}\frac{b'}{n}\log\left(\frac{1}{\mathbb{P}(A)}\right).\end{aligned}$$

Therefore,

$$\begin{aligned}\mathbb{E}^A[\sup_{f\in\mathcal{F}_m}\frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(f_{k_l}) + \nu_l(\tilde{f}_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]} + \frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(\tilde{f}_{k_l})}{y_m^2 + \kappa'^2\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{s_l}{\tilde{f}_{k_l}}\right)|^2\right]}] \\ \leq 4(27+4\tau_n)\frac{\sigma_m}{y_m} + \left(\frac{28\sqrt{2}}{\sqrt{ny_m^2}} + \frac{1}{\kappa'}\frac{\sqrt{2}}{\sqrt{ny_m^2}}\right)\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} + \left(\frac{9\tau_n}{ny_m^2}\right)\log\left(\frac{1}{\mathbb{P}(A)}\right).\end{aligned}$$

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

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

we have

$$\begin{aligned}&\mathbb{E}^A[\sup_{f\in\mathcal{F}_m}\frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(f_{k_l})}{y_m^2 + 2(2+\epsilon_d)\frac{1}{n}\sum_{l=1}^{L} n_l\mathbb{E}\left[|\log\left(\frac{f_{k_l}}{s_l}\right)|^2\right]}] \\ \leq 4(27+4\tau_n)\frac{\sigma_m}{y_m} + &\left(\frac{28\sqrt{2}}{\sqrt{ny_m^2}} + \frac{1}{\kappa_d'}\frac{\sqrt{2}}{\sqrt{ny_m^2}}\right)\sqrt{\log\left(\frac{1}{\mathbb{P}(A)}\right)} + \left(\frac{9\tau_n}{ny_m^2}\right)\log\left(\frac{1}{\mathbb{P}(A)}\right).\end{aligned}$$

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

5.3 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 on our collection of models:

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

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

The model selection theorem is then:

Theorem 5.3. *Let $(X_l)_{1\le l\le L}$ be independent random vectors where each X_l consists of n_l independent and identically distributed (iid) instances of a multinomial vector that has s_l as a true histogram density with respect to some positive measure μ. Assume $\mathcal{S} := (\mathcal{S}_m)_{m\in\mathcal{M}}$ is an at most countable collection of models for which Assumption (**K**) holds. For every model $S_m \in \mathcal{S}$, we also assume that Assumptions ($\mathbf{H_m}$) and ($\mathbf{Sep_m}$) hold. Let $(\widehat{\pi^m}, \widehat{f^m})$ be a η-MLE in $\mathcal{S}_m$:*

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

and define its corresponding assignment for each vector X_l:

$$\widehat{k_l^m} = \operatorname*{argmax}_{1\le k\le K} \left\{ \widehat{\pi_k^m} \left(\prod_{i=1}^{n_l} \widehat{f_k^m}(X_l^i) \right) \right\}.$$

Define $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$. Let also $\mathbf{pen} : \mathcal{M} \to \mathbb{R}^+$ and consider the penalized log-likelihood criterion:

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

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

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

the penalized log-likelihood estimate verifying for all $\eta' > 0$:

$$\mathbf{crit}(\widehat{m}) \le \inf_{m\in\mathcal{M}} \{\mathbf{crit}(m)\} + \eta'$$

satisfies

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{\widehat{m}}_{\widehat{k_l}}})] \le C_1 \inf_{m\in\mathcal{M}} \left\{ \inf_{k_l} \left\{ \inf_{f} \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l}) \right\} + \frac{\mathbf{pen}(m)}{n} \right\} + C_2 \frac{\Sigma}{n} + \frac{\eta+\eta'}{n}.$$

Proof. For any cluster assignment $(k_l)_{1\le l\le L}$ within the model, define $\overline{f^m} \in \mathcal{F}_m$ such that:

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

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

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

Then, for all $m' \in \mathcal{M}'$:

$$\begin{aligned}
&\sum_{l=1}^{L} -\log\left(\sum_{k=1}^{K_{m'}} \widehat{\pi_k^{m'}} \left(\prod_{i=1}^{n_l} \left(\frac{\widehat{f_k^{m'}}}{s_l} \right)(X_l^i) \right) \right) + \mathbf{pen}(m') \\
&\le \sum_{l=1}^{L} -\log\left(\sum_{k=1}^{K_m} \widehat{\pi_k^m} \left(\prod_{i=1}^{n_l} \left(\frac{\widehat{f_k^m}}{s_l} \right)(X_l^i) \right) \right) + \kappa_0(\mathfrak{D}_m + L\log(K_m) + x_m) + \eta'
\end{aligned}$$

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

$$\begin{aligned}
&\sum_{l=1}^{L} -\log\left(\sum_{k=1}^{K_m} \widehat{\pi_k^m} \left(\prod_{i=1}^{n_l} \left(\frac{\widehat{f_k^m}}{s_l} \right)(X_l^i) \right) \right) + \kappa_0(\mathfrak{D}_m + L\log(K_m) + x_m) + \eta' \\
&\le \inf_{(\pi^m, f^m)\in\Theta_m} \left\{ \sum_{l=1}^{L} -\log\left(\sum_{k=1}^{K_m} \pi_k^m \left(\prod_{i=1}^{n_l} \left(\frac{f_k^m}{s_l} \right)(X_l^i) \right) \right) \right\} + \eta + \kappa_0(\mathfrak{D}_m + L\log(K_m) + x_m) + \eta' \\
&\le \inf_{k_l} \left\{ \inf_{(\pi^m, f^m)\in\Theta_m} \left\{ \sum_{l=1}^{L} -\log(\pi^m_{k_l}) + \sum_{l=1}^{L}\sum_{i=1}^{n_l} -\log\left(\left(\frac{f^m_{k_l}}{s_l} \right)(X_l^i) \right) \right\} \right\} + \eta + \kappa_0(\mathfrak{D}_m + L\log(K_m) + x_m) + \eta'.
\end{aligned}$$

Using the definition of ν_l and by assumption on **pen**,

$$\inf_{k_l}\left\{\inf_{(\pi^m,f^m)\in\Theta_m}\left\{\sum_{l=1}^{L}-\log(\pi^m_{k_l})+\sum_{l=1}^{L}\sum_{i=1}^{n_l}-\log\left(\left(\frac{f^m_{k_l}}{s_l}\right)(X^i_l)\right)\right\}\right\}+\eta+\kappa_0(\mathfrak{D}_m+L\log(K_m)+x_m)+\eta'$$
$$\leq \inf_{k_l}\left\{\sum_{l=1}^{L}-\log(\frac{1}{K_m})+\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})+\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\bar{f}^m_{k_l})\right\}+\eta+\kappa_0(\mathfrak{D}_m+L\log(K_m)+x_m)+\eta'$$
$$\leq L\log(K_m)+\inf_{k_l}\left\{\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})+\inf_{f^m\in\mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f^m_{k_l})\right\}+\delta+\eta+\eta'+\kappa_0(\mathfrak{D}_m+L\log(K_m)+x_m)$$
$$\leq \inf_{k_l}\left\{\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})+\inf_{f^m\in\mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f^m_{k_l})\right\}+\delta+\eta+\eta'+\mathbf{pen}(m).$$

It follows that

$$\sum_{l=1}^{L}-\log\left(\sum_{k=1}^{K_{m'}}\widehat{\pi^{m'}_k}\left(\prod_{i=1}^{n_l}\left(\frac{\widehat{f^{m'}_k}}{s_l}\right)(X^i_l)\right)\right)+\mathbf{pen}(m')\leq\inf_{k_l}\left\{\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})+\inf_{f^m\in\mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f^m_{k_l})\right\}+\delta+\eta+\eta'+\mathbf{pen}(m)$$

and on the other hand,

$$\begin{aligned}\sum_{l=1}^{L}-\log\left(\sum_{k=1}^{K_{m'}}\widehat{\pi^{m'}_k}\left(\prod_{i=1}^{n_l}\left(\frac{\widehat{f^{m'}_{k_l}}}{s_l}\right)(X^i_l)\right)\right)+\mathbf{pen}(m')&\geq\sum_{l=1}^{L}-\log\left(K_{m'}\widehat{\pi^{m'}_{\widehat{k}_l}}\left(\prod_{i=1}^{n_l}\left(\frac{\widehat{f^{m'}_{k_l}}}{s_l}\right)(X^i_l)\right)\right)+\mathbf{pen}(m')\\&\geq\sum_{l=1}^{L}-\log\left(K_{m'}\left(\prod_{i=1}^{n_l}\left(\frac{\widehat{f^{m'}_{k_l}}}{s_l}\right)(X^i_l)\right)\right)+\mathbf{pen}(m')\\&\geq-L\log(K_{m'})+\sum_{l=1}^{L}\nu_l(\widehat{f^{m'}_{\widehat{k}_l}})+\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\widehat{f^{m'}_{\widehat{k}_l}})+\mathbf{pen}(m')\end{aligned}$$

Thus,

$$\begin{aligned}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,\widehat{f^{m'}_{\widehat{k}_l}})\leq L\log(K_{m'})-\sum_{l=1}^{L}\nu_l(\widehat{f^{m'}_{\widehat{k}_l}})-\mathbf{pen}(m')&\\+\inf_{k_l}\left\{\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})+\inf_{f^m\in\mathcal{F}_m}\sum_{l=1}^{L}n_l\mathbf{KL}(s_l,f^m_{k_l})\right\}+\delta+\eta+\eta'+\mathbf{pen}(m)&\end{aligned}$$

It remains to get an upper bound of the deviation $-\sum_{l=1}^{L}\nu_l(\widehat{f^{m'}_{\widehat{k}_l}})$. Using Lemmas 5.1 and 5.2, except on a set of probability less than $e^{-x'_m-x}$, for any $y_{m'}>\sigma_{m'}$,

$$\frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(\widehat{f}^{m'}_{\widehat{k}_l})}{y^2_{m'}+\kappa'_0\frac{1}{n}\sum_{l=1}^{L}n_l\mathbb{E}\left[|\log\left(\frac{\widehat{f^{m'}_{k_l}}}{s_l}\right)|^2\right]}\leq\frac{L\log(K_{m'})}{\lambda}+\frac{\kappa'_1\sigma_{m'}}{y_{m'}}+\frac{\kappa'_2}{\sqrt{ny^2_{m'}}}\sqrt{x+x_{m'}}+\left(\frac{9\tau_n}{ny^2_{m'}}\right)(x+x_{m'}).$$

We choose this time λ as $\lambda_m:=ny^2_{m'}>0$ and $y_{m'}:=\theta\sqrt{\frac{x+x_{m'}}{n}+\sigma^2_{m'}+\frac{L\log(K_{m'})}{n}}$, 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 $y_{m'}>\sigma_{m'}$,

$$\frac{1}{n}\sum_{l=1}^{L}\frac{-\nu_l(\widehat{f}^{m'}_{\widehat{k}_l})}{y^2_{m'}+\kappa'_0\frac{1}{n}\sum_{l=1}^{L}n_l\mathbb{E}\left[|\log\left(\frac{\widehat{f^{m'}_{k_l}}}{s_l}\right)|^2\right]}\leq\left(\frac{9\tau_n+1}{\theta^2}+\frac{\kappa'_1+\kappa'_2}{\theta}\right).$$

Now, we use the Kraft condition (**K**), 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 $m'\in\mathcal{M}'$,

$$\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{m'}_{\widehat{k}_l}}) \leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} - \frac{\mathbf{pen}(m')}{n} + \frac{L\log(K_{m'})}{n} + \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\}$$
$$+ \left(y^2_{m'} + \kappa'_0 \frac{1}{n}\sum_{l=1}^{L} n_l \mathbb{E}\left[\left|\log\left(\frac{\widehat{f^{m'}_{\widehat{k}_l}}}{s_l}\right)\right|^2\right]\right)\left(\frac{9\tau_n+1}{\theta^2} + \frac{\kappa'_1+\kappa'_2}{\theta}\right).$$

Now we define $C_{\tau_n} := \frac{e^{-\tau_n}+\tau_n-1}{\tau_n^2}$ and choose $\epsilon_{pen} > 0$ such that $\left(\frac{18\tau_n+1}{\theta^2_{pen}} + \frac{\kappa'_1+\kappa'_2}{\theta_{pen}}\right)\kappa'_0 = C_{\tau_n}\epsilon_{pen}$. We obtain

$$\begin{aligned}
\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{m'}_{\widehat{k}_l}}) &\leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} - \frac{\mathbf{pen}(m')}{n} + \frac{L\log(K_{m'})}{n} + \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} \\
&\quad + \left(\frac{1}{n}\sum_{l=1}^{L} n_l \mathbb{E}\left[\left|\log\left(\frac{\widehat{f^{m'}_{\widehat{k}_l}}}{s_l}\right)\right|^2\right]\right) C_{\tau_n}\epsilon_{pen} + \frac{y^2_{m'} C_{\tau_n}\epsilon_{pen}}{\kappa'_0} \\
&\leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} - \frac{\mathbf{pen}(m')}{n} + \frac{L\log(K_{m'})}{n} + \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} \\
&\quad + \frac{\epsilon_{pen}}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{m'}_{\widehat{k}_l}}) + \frac{y^2_{m'} C_{\tau_n}\epsilon_{pen}}{\kappa'_0}.
\end{aligned}$$

Then, simultaneously for any $m' \in \mathcal{M}'$, except on a set of probability less than $e^{-x}\Sigma$,

$$\begin{aligned}
(1-\epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{m'}_{\widehat{k}_l}}) &\leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} - \frac{\mathbf{pen}(m')}{n} + \frac{L\log(K_{m'})}{n} \\
&\quad + \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} + \frac{y^2_{m'} C_{\tau_n}\epsilon_{pen}}{\kappa'_0}
\end{aligned} \tag{9}$$

Let now study the term $\frac{L\log(K_{m'})}{n} + \frac{y^2_{m'} C_{\tau_n}\epsilon_{pen}}{\kappa'_0} - \frac{\mathbf{pen}(m')}{n}$. Define $\kappa_0 := \frac{\theta^2_{pen} C_{\tau_n}\epsilon_{pen}}{\kappa'_0}$. We have

$$\begin{aligned}
\frac{L\log(K_{m'})}{n} + \frac{y^2_{m'} C_{\tau_n}\epsilon_{pen}}{\kappa'_0} - \frac{\mathbf{pen}(m')}{n} &= \frac{L\log(K_{m'})}{n} + \kappa_0\left(\frac{x+x_{m'}}{n} + \sigma^2_{m'} + \frac{L\log(K_{m'})}{n}\right) - \frac{\mathbf{pen}(m')}{n} \\
&\leq \kappa_0\frac{x}{n} + (1+\kappa_0)\left(\frac{x_{m'}}{n} + \sigma^2_{m'} + \frac{L\log(K_{m'})}{n}\right) - \frac{\mathbf{pen}(m')}{n} \\
&\leq \kappa_0\frac{x}{n} - (1 - \frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(m')}{n}.
\end{aligned}$$

Thus, based on equation (9), except on a set of probability less than $e^{-x}\Sigma$, simultaneously for any $m' \in \mathcal{M}'$,

$$\begin{aligned}
(1-\epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{m'}_{\widehat{k}_l}}) + (1 - \frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(m')}{n} - \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} \\
\leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 x}{n}.
\end{aligned}$$

Since $\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l})$ 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

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

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

$$(1-\epsilon_{pen})\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{\widehat{m}}_{\widehat{k}_l}}) + (1-\frac{1+\kappa_0}{\kappa})\frac{\mathbf{pen}(\widehat{m})}{n} - \inf_{k_l}\left\{\frac{1}{n}\sum_{l=1}^{L}\nu_l(\bar{f}^m_{k_l}) + \inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\}$$
$$\leq \frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n} + \frac{\kappa_0 x}{n}.$$

Using the same integration technique than in the previous theorem, one has

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{\widehat{m}}_{\widehat{k}_l}})] + \frac{(1-\frac{1+\kappa_0}{\kappa})}{1-\epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n} \leq \frac{1}{1-\epsilon_{pen}}(\frac{\eta+\eta'}{n} + \frac{\delta}{n} + \frac{\mathbf{pen}(m)}{n}$$
$$+ \inf_{k_l}\left\{\inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} + \frac{\kappa_0 \Sigma}{n}),$$

and since δ can be artbitrary small,

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{\widehat{m}}_{\widehat{k}_l}})] + \frac{(1-\frac{1+\kappa_0}{\kappa})}{1-\epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n}] \leq \frac{1}{1-\epsilon_{pen}}(\frac{\eta+\eta'}{n} + \frac{\mathbf{pen}(m)}{n} + \inf_{k_l}\left\{\inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} + \frac{\kappa_0 \Sigma}{n}).$$

Therefore,

$$\mathbb{E}[\frac{1}{n}\sum_{l=1}^{L} n_l \mathbf{KL}(s_l, \widehat{f^{\widehat{m}}_{\widehat{k}_l}})] + \frac{(1-\frac{1+\kappa_0}{\kappa})}{1-\epsilon_{pen}}\frac{\mathbf{pen}(\widehat{m})}{n}] \leq \frac{1}{1-\epsilon_{pen}} \inf_{m\in\mathcal{M}}\left\{\inf_{k_l}\left\{\inf_f \sum_{l=1}^{L} n_l \mathbf{KL}(s_l, f^m_{k_l})\right\} + \frac{\mathbf{pen}(m)}{n}\right\}$$
$$+ \frac{\kappa_0}{1-\epsilon_{pen}}\frac{\Sigma}{n} + \frac{\eta+\eta'}{n},$$

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. □

5.4 Proofs of Theorems 2.1 and 2.2

5.4.1 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 $\mathcal{F}_m$ with respect to $\mathbf{a}$, based on simpler sets of functions. Let first introduce two other metrics. For $f_k \in \mathcal{F}_{m,k}$, we denote by $\|.\|_\infty$ the L_∞-norm:

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

For $f, \tilde{f} \in \mathcal{F}_m$ and $(k_l)_{1\leq l\leq L} \in \{1,...,K\}^L$, let also $\mathbf{d}^2_\infty$ be the divergence defined by:

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

Lemma 5.5. *Let $(\delta_k)_{1\leq k\leq K}$ be a family of positive numbers. Then,*

$$H_{[.],\mathbf{d}^2_\infty}(\max_{k=1,...,K}\delta^2_k, \mathcal{F}_m) \leq \sum_{k=1}^{K} H_{[.],\mathbf{d}^2_\infty}(\delta^2_k, \mathcal{F}_{m,k})$$

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

$$\mathbf{d}^2_\infty((f^-_k)_{1\leq k\leq K}, (f^+_k)_{1\leq k\leq K}) = \frac{1}{n}\sum_{l=1}^{L} n_l \|\log\left(\frac{f^-_{k_l}}{f^+_{k_l}}\right)\|^2_\infty \leq \max_{k=1,...,K}\delta^2_k.$$

Therefore, by covering all $\mathcal{F}_{m,k}$ with a number of brackets N_k of width less than δ^2_k, we can cover $\mathcal{F}_m$ with $\prod_k N_k$ brackets, which leads to the result. □

Lemma 5.6. *Let $\epsilon > 0$. Then for all $k = 1, ..., K$,*

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

Proof. By definition, $\mathcal{F}_{m,k} = \mathbb{S}_{B-1} \cap [e^{-\tau_n}, 1]^B$. Let $\mathcal{B}_k(\epsilon)$ be a set of brackets of $\|.\|_\infty$-width less than ϵ covering $[-\tau_n, 0]^B$. Let $f_k \in \mathcal{F}_{m,k}$. 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 $\mathbf{d}^2_\infty$. Therefore, an $\epsilon - \|.\|_\infty$-covering of $[-\tau_n, 0]^B$ induces an $\epsilon^2 - \mathbf{d}^2_\infty$-covering of $\mathcal{F}_{m,k}$ so we can conclude. □

The following result is inspired by Lemma 2 from [GW00].

Lemma 5.7. *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 $\max_{1\leq b\leq B} \mid x_1(b) - y_1(b) \mid \leq \epsilon$. Thus, the family of vertices $\{(x_1, y_1), ..., (x_N, y_N)\}$ forms an $\epsilon - \|.\|_\infty$ bracketing of $[-\tau_n, 0]^B$. Clearly, we have

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

□

Proposition 5.2 (Bracketing entropy of $\mathcal{F}_m$)**.** *For all $m \in \mathcal{M}$ and all $\delta \in (0, 1]$, we have*

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

Proof. Let $\delta \in (0, 1]$. By definition, $\mathbf{a} \leq \mathbf{d}^2_\infty$. Therefore, $H_{[.],\mathbf{a}}(\delta^2, \mathcal{F}_m) \leq H_{[.],\mathbf{d}^2_\infty}(\delta^2, \mathcal{F}_m)$. Recalling that for all k, $\mathcal{F}_{m,k} = \mathbb{S}_{B-1} \cap [e^{-\tau_n}, 1]^B$, given Lemmas 5.5, 5.6 and 5.7, we have

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

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

□

5.4.2 Bracketing and dimension of models

Theorems 2.1 and 2.2 are obtained from Theorems 5.1 and 5.3 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 $\sigma \mapsto \int_0^\sigma \sqrt{H_{[.],\mathbf{a}}(\delta, \mathcal{S}_m)}d\delta$ always satisfies Assumption $(\mathbf{H_m})$. The quantity $\mathfrak{D}_m$ is defined as $n\sigma_m^2$ where σ_m^2 is the unique root of $\phi_m(\sigma)/\sigma = \sqrt{n}\sigma$. 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 [CP11] shows that under some assumption on the bracketing entropy, $\mathfrak{D}_m$ is proportional to the entropic dimension D_m.

Proposition 5.3. *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

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

satisfies the properties required in Assumption $(\mathbf{H_m})$ *, and $\mathfrak{D}_m$ satisfies*

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

5.4.3 Proof of Theorem 2.1

Proposition 5.2 directly shows that if we choose the constants $D_m := KB$ and $C_m := \log(2)$ and apply Proposition 5.3, there is a ϕ_m that satisfies Assumption $(\mathbf{H_m})$ in our setting. Therefore, we can apply Theorem 5.1 and get the following oracle inequality:

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

Define $\mu_n := 2\left(\sqrt{\log(2)} + \sqrt{\pi}\right)^2 + 1 + \log(n)$. According to Proposition 5.3, we also have

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

and

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

with $C_2' := C_2 + \kappa_0\mu_n$. Since $\mu_n \geq 1$, it remains to choose $\lambda_0 := 2 + \kappa_0\mu_n$ to conclude the proof.

5.4.4 Proof of Theorem 2.2

We need to find the weights x_m satisfying Assumption $(\mathbf{K})$ in order to apply Theorem 5.3 in our framework. The following Lemma states that $x_m \geq KB\log(2)$ is a sufficient condition on the weights. Its demonstration is inspired by Lemma 2's proof in [BT13]:

Proposition 5.4. *For any $m \in \mathcal{M}$ with $\mathcal{M}$ the collection of models defined above, let $x_m \geq KB\log(2)$. Then*

$$\sum_{m \in \mathcal{M}} e^{-x_m} \leq 1.$$

Proof. Define $\delta := 1/2$. Then, $e^{-x_m} \leq \delta^{KB}$. Recalling the collection to be $\mathcal{M} := \{(1,1) \cup (\mathbb{N}\setminus\{0\} \times \mathbb{N}\setminus\{0,1\})\}$, we have

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

□

Now take x_m such that $x_m = K_m B_m \log(2) + \mu_n$. It satisfies Kraft Assumption $(\mathbf{K})$, and $\sum_{m\in\mathcal{M}} e^{-x_m} \leq 1$. We can apply Theorem 5.3 and use Proposition 5.3 to state that if

$$\mathbf{pen}(m) \geq \kappa(D_{(K,B)}\mu_n + L\log(K_m) + x_m), \kappa > \kappa_0 + 1$$

the following inequality is satisfied:

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

It remains to define λ_0' as any $\lambda_0' > (\kappa_0+1)\mu_n$ in order to obtain the result stated.

References

[BMM10] Jean-Patrick Baudry, Cathy Maugis, and Bertrand Michel. Slope heuristics: Overview and implementation. 2010.

[BT13] Dominique Bontemps and Wilson Toussile. Clustering and variable selection for categorical multivariate data. *Electronic Journal of Statistics*, 7:2344–2371, 2013.

[CP11] S. X. Cohen and E. Le Pennec. Conditional density estimation by penalized likelihood model selection and applications. *INRIA*, Technical Report, 2011.

[CP12] S.X. Cohen and E. Le Pennec. Partition-based conditional density estimation. *ESAIM: Probability and Statistics*, 2012.

[CP14] S. X. Cohen and E. Le Pennec. Unsupervised segmentation of spectral images with a spatialized gaussian mixture modeland model selection. *Oil and Gas Science and Technology*, 2014.

[FJ02] Mario A.T. Figueiredo and Anil K. Jain. Unsupervised learning of finite mixture models. 2002.

[FSM17] Michael Fop, Keith M. Smartzand, and Thomas Brendan Murphy. Variable selection for latent class analysis with application to low back diagnosis. 2017.

[GRV16] Elisabeth Gassiat, Judith Rousseau, and Elodie Vernet. Efficient semiparametric estimation and model selection for multidimensional mixtures. 2016.

[GW00] Christopher R. Genovese and Larry Wasserman. Rates of convergence for the gaussian mixture sieve. *The Annals of Statistics*, 2000.

[Mas07] Pascal Massart. *Concentration Inequalities and Model Selection*, volume 1896. Springer, 2007.

[Mey12] Caroline Meynet. *Variable selection in model-based clustering for high-dimensional data*. PhD thesis, Université Paris-Sud XI, 2012.

[MM08a] Cathy Maugis and Bertrand Michel. A non asymptotic penalized criterion for gaussian mixture model selection. 2008.

[MM08b] Cathy Maugis and Bertrand Michel. Slope heuristics for variable selection and clustering via gaussian mixtures. 2008.

[MP00] Geoffrey McLachlan and David Peel. *Finite Mixture Models*. Wiley Series in Probability and Statistics, 2000.

[MP14] L. Montuelle and E. Le Pennec. Mixture of gaussian regressions model with logistic weights, a penalized maximum likelihood approach. *Electronic Journal of Statistics*, 2014.

[MRV15] Catherine Matias, Tabea Rebafka, and Fanny Villers. Estimation and clustering in a semiparametric poisson process stochastic block model for longitudinal networks. *HAL*, 2015.

[PJST16] Valerio Perrone, Paul A. Jenkins, Dario Spanó, and Yee Whye Teh. Poisson random fields for dynamic feature models. 2016.

[RCY06] Loïs Rigouste, Olivier Cappé, and François Yvon. Inference and evaluation of the multinomial mixture model for text clustering. 2006.

[SCF14] Clàudia Silvestre, Margarida G. M. S. Cardoso, and Màrio A. T. Figueiredo. Identifying the number of clusters in discrete mixture models. 2014.

[TG09] Wilson Toussile and Elisabeth Gassiat. Variable selection in model-based clustering using multilocus genotype data. 2009.

[YLL12] Miin-Shen Yang, Chien-Yo Lai, and Chih-Ying Lin. A robust em clustering algorithm for gaussian mixture models. 2012.

[ZZY04] Baibo Zhang, Changshui Zhang, and Xing Yi. Competitive em algorithm for nite mixture models. 2004.