Model selection of GLM mixtures with a clustering perspective

Olivier Lopez[a,b,c], Xavier Milhaud[a,b,*]

[a] *ENSAE ParisTech, 3 Avenue Pierre Larousse, 92245 Malakoff Cedex, France*
[b] *CREST (LFA lab), 15 Boulevard Gabriel Péri, 92245 Malakoff Cedex, France*
[c] *Sorbonne Universités, UPMC Université Paris VI, EA 3124, LSTA, 4 place Jussieu 75005 Paris, France*

Abstract

Model-based clustering from finite mixtures of generalized linear models (GLM) is a challenging issue which has undergone many recent developments. In practice model selection is often performed with AIC or BIC penalized criteria, although simulations show that they tend to overestimate the actual dimension of the model. As an alternative, we adapt another existing strategy to GLM mixtures: it consists in adding an entropic term to the log-likelihood in order to select the best mixture model, where "best" should be understood as the best trade-off between the fit to data and some clustering confidence. We derive key properties about the convergence of the associated M-estimator using concentration inequalities. The consistency of the corresponding selection criterion naturally follows under some classical requirements on the penalty. Finally, simulations enable to corroborate our theoretical results in the GLM mixtures framework and shows the effectiveness of the method.

Keywords: Clustering, Regression, Mixture, Model selection

1. Introduction

Selecting one model within a collection is an important statistical problem which has undergone vigorous developments in the literature. However,

*Corresponding author
Email addresses: olivier.lopez@ensae.fr (Olivier Lopez), xavier.milhaud@ensae.fr (Xavier Milhaud)

a universal solution to answer the question of selecting the right order in the mixture framework has failed to emerge. Many articles have been dedicated to the implementation of algorithmic mixture calibration techniques (e.g. [24], [35]). Nevertheless, they often suffer from a lack of theoretical justification with respect to their convergence properties. [31] point that there are basically two main approaches to infer the order of a mixture: hypothesis tests and information criteria. Concerning the former, [14] highlights the difficulty of establishing multimodality by means of the generalized likelihood ratio test. Indeed, the classical result according to which the test statistic is χ^2-distributed is generally not applicable where mixtures are concerned. [3], and [2], building on [15], suggest a potential solution to overcome this issue. When using information criteria, most authors agree that the BIC criterion gives better results than AIC ([26], [13], [33]). First, the consistency of BIC to estimate the order of gaussian mixtures has been proved in [22]. Recently, [16] have unrolled the convergence properties of mixture models selected by a likelihood-penalized criterion, where the penalty linearly depends on the model dimension. But their findings rely on some strong assumptions, in particular the boundedness of the log-likelihood under study. Despite these results, a commonly accepted fact is that both AIC and BIC tend to overestimate the theoretical number of components, especially when the model is misspecified. This statement is not really surprising: these two criteria and their variations were originally proposed for model selection problems in regular statistical models, and thus their usage is not well-supported or motivated for model selection in non-standard models such as mixtures. This is all the more important that applications involving finite mixtures are usually focused on describing a population structure, and that a key feature in mixture models for expliciting this structure is precisely their order.

To avoid this problem, [9] and [7] introduced a *classification* criterion, namely the ICL criterion. Unfortunately, there does not exist any consistency result about ICL in the context of maximum likelihood theory. Adopting another approach, [4] has recently demonstrated the consistency of a modified version of ICL, say ICL*, for gaussian mixtures. In this view he introduced a new contrast called the *conditional classification likelihood*, denoted by L_{cc}. The estimator maximizing the empirical L_{cc} differs from the maximum likelihood one in that it aims at finding the best compromise between a small *a posteriori* classification error and a good fit to data. Following his approach, we extend these results to make them fulfill the requirements needed in our context, the GLM mixtures framework. Our aim is to develop new theoretical

results about ICL* that apply to this more general context.

For the paper to be self contained, parts of sections 2 and 3 are devoted to the presentation of innovative concepts introduced and studied in [5], respectively the L_{cc} contrast and the ICL* criterion. Based on concentration inequalities, we provide a new result that states a general bound for the estimation error in section 2, even when considering unbounded log-contrasts. In section 3, we determine conditions for the consistency of general penalized criteria that are considered to select the order of the mixture. An application to GLM mixtures follows in section 4, and the practical behavior of this approach is investigated through simulations in section 5. Our results are expressed for a one-dimensional outcome Y but remain valid when Y is k-dimensional ($k > 1$).

2. The maximum conditional classification likelihood estimator

2.1. Context of mixtures

Let $(\mathcal{Y}, \mathcal{F})$ be a measurable space and let $(f_\theta)_{\theta\in\Theta}$ be a parametric family of densities on $\mathcal{Y}$. The parameter θ is assumed to range over a set $\Theta \in \mathbb{B}(\mathbb{R}^d)$; where $\mathbb{B}(.)$ denotes the Borel sets and $d \geq 1$. For any probability measure ν on $(\Theta, \mathbb{B}(\Theta))$, the mixture density f_ν is defined on $\mathcal{Y}$ by

$$f_\nu(y) = \int_\Theta f_\theta(y)\, \nu(d\theta) = \int_\Theta f(y;\theta)\, \nu(d\theta).$$

ν is the mixing distribution and (f_θ) is called the mixand. If ν has finite support, f_ν is a finite mixture density. In the present paper, our interest lies in finite discrete mixtures: f_ν is assumed to belong to a collection of densities M_g such that

$$M_g = \Big\{ f(y;\psi_g) = \sum_{i=1}^{n_g} \pi_i\, f_i(y;\theta_i) \mid \psi_g = (\pi_1, ..., \pi_{n_g}, \theta_1, ..., \theta_{n_g}) \in \Psi_g \Big\}, \quad (1)$$

where $\Psi_g = \big(\Pi_{n_g} \times \Theta^{n_g}\big)$, with $\Theta^{n_g} = (\theta_1, ..., \theta_{n_g})$ and the set Π_{n_g} is such that $\Pi_{n_g} \subset \big\{(\pi_1, ..., \pi_{n_g}) : \sum_{i=1}^{n_g} \pi_i = 1, \text{ and } \pi_i \geq \pi_{i+1} \geq 0 \text{ for } 1 \leq i \leq n_g - 1\big\}$. Denote K_g the dimension of the parameter set Ψ_g. Except for specific cases, mixture models belonging to M_g are identifiable ([26], p.26). Based on i.i.d. observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the likelihood reads

$$\forall \psi_g \in \Psi_g, \quad L(\psi_g; \mathbf{Y}) = L(\psi_g) = \prod_{j=1}^{n} \sum_{i=1}^{n_g} \pi_i\, f_i(Y_j;\theta_i). \quad (2)$$

Let us note that the true density function, $f^0(y)$, may not belong to any M_g in a misspecified case. The maximum likelihood estimator (MLE) $\hat{\psi}_g^{ML}$ is defined as the maximizer of $L(\psi_g)$ over Ψ_g (in full generality, it may not be unique). Under some regularity conditions, $\hat{\psi}_g^{ML}$ converges towards the unknown parameter $\psi_g^{ML} = \arg\max_{\psi_g} E_{f^0}[\ln L(\psi_g; Y)]$, which is the true parameter ψ^0 when the model is correctly specified ($f^0(y) = f(y; \psi^0)$).

2.2. The conditional classification likelihood

The conditional classification likelihood is derived from the general principle of the EM algorithm, and more precisely from the likelihood of the complete data $(Y_j, \delta_j)_{1\leq j\leq n}$, where $\delta_j = (\delta_{ij})_{1\leq i\leq n_g}$ is the latent component indicator (δ_{ij} equals 1 if observation j belongs to component i, 0 otherwise). Several authors have attempted to exploit the link between the observed likelihood and the likelihood of the complete data (or *classification likelihood*, see [8]), originally noted by [20]. A specific term appears while writing the likelihood relatively to the complete data $(\mathbf{Y}, \delta)$: $\forall \psi_g \in \Psi_g$,

$$\begin{aligned}
\ln L_c(\psi_g;\, \mathbf{Y}, \delta) &= \sum_{j=1}^{n}\sum_{i=1}^{n_g} \delta_{ij}\, \ln\left(\pi_i\, f_i(Y_j;\theta_i)\right) \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n_g} \delta_{ij}\, \ln\Big(\underbrace{\frac{\pi_i f_i(Y_j;\theta_i)}{\sum_{k=1}^{n_g}\pi_k f_k(Y_j;\theta_k)}}_{\tau_i(Y_j;\psi_g)}\Big) + \underbrace{\sum_{j=1}^{n}\overbrace{\sum_{i=1}^{n_g}\delta_{ij}}^{=1}\ln\left(\sum_{k=1}^{n_g}\pi_k f_k(Y_j;\theta_k)\right)}_{\ln L(\psi_g;\mathbf{Y})} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n_g} \delta_{ij}\, \ln\Big(\tau_i(Y_j;\psi_g)\Big) \;+\; \ln L(\psi_g;\mathbf{Y}) \qquad (3)
\end{aligned}$$

$\tau_i(Y_j;\psi_g)$ is the *a posteriori* probability that observation j belongs to the i$^{\text{th}}$ component. The term that binds the two likelihoods is close to what is often called the entropy: $\forall \psi_g \in \Psi_g,\ Ent(\psi_g; Y_j) = -\sum_{i=1}^{n_g}\tau_i(Y_j;\psi_g)\, \ln\Big(\tau_i(Y_j;\psi_g)\Big)$. This function results from the conditional expectation with respect to δ taken in (3), and leads to the "conditional classification likelihood":

$$\begin{aligned}
\ln L_{cc}(\psi_g;\mathbf{Y}) &= \ln L(\psi_g;\mathbf{Y}) \;+\; \sum_{j=1}^{n}\sum_{i=1}^{n_g}\mathbb{E}_\delta[\delta_{ij}|Y_j]\ln\big(\tau_i(Y_j;\psi_g)\big) \\
&= \ln L(\psi_g;\mathbf{Y}) \;-\; Ent(\psi_g;\mathbf{Y}), \qquad (4)
\end{aligned}$$

where $Ent(\psi_g;\mathbf{Y}) = \sum_{j=1}^{n} Ent(\psi_g; Y_j)$.

Note that $0 \leq Ent(\psi_g; Y_j) \leq \ln n_g$. The entropy is maximum at equiprobability ($\tau_1(Y_j; \psi_g) = ... = \tau_{n_g}(Y_j; \psi_g)$); and minimum when one of the *a posteriori* probabilities is worth 1. As can be seen in (4), this term somewhat represents a penalization of the observed likelihood: the lower the confidence when making the *a posteriori* classification of individuals via the Bayes rule, the greater the penalization. For instance, the entropy has a zero limit when τ_i tends to 0 or 1 in a two-component mixture. However, it is not differentiable at 0: consider the function $h(\tau_i) = \tau_i \ln \tau_i$, then we have $\lim_{\tau_i \to 0^+} h^{'}(\tau_i) = -\infty$. This is a key point in the definition of the parameter space Ψ_g that is acceptable to ensure the convergence of the estimator based on the L_{cc} contrast, and one should therefore avoid that the *a posteriori* proportions of the mixture tend to zero. Additional constraints to be imposed on the parameter space are typically deduced by studying the limits of (4) as well as its derivatives. The idea is to prevent them from diverging.

[4] defined the maximum conditional classification likelihood estimator:

$$ML_{cc}E \quad := \quad \hat{\psi}_g^{ML_{cc}} = \underset{\psi_g \in \Psi_g}{\arg\max} \frac{1}{n} \sum_{j=1}^{n} \ln L_{cc}(\psi_g;\, Y_j). \tag{5}$$

The $ML_{cc}E$ has to be found as the empirical counterpart of the parameter $\psi_g^{ML_{cc}} = \arg\max_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g, Y)]$. [5] (p.10) provides us with an example to catch in what $\psi_g^{ML_{cc}}$ differs from ψ_g^{ML}. In particular, he shows that the $ML_{cc}E$ does not aim at recovering the theoretical distribution, even when contained in the model under consideration. The compromise sought by the $ML_{cc}E$, which penalizes an excessive entropy, leads to find another estimator resulting in greater confidence in the *a posteriori* assignment of observations to mixture components. This is crucial in our clustering objective.

2.3. Exponential bound for the $ML_{cc}E$

To shorten the notation, let $\psi_g^b = \psi_g^{ML_{cc}}$ and $\hat{\psi}_g = \hat{\psi}_g^{ML_{cc}}$. Define $\mathcal{L}(\psi_g; y) = \ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)$, and $d(\psi_g, \psi_g^b) = -\mathbb{E}_{f^0}[\mathcal{L}(\psi_g; Y)]$. The function d is defined in the same idea as the Kullback-Leibler (KL) divergence between $f(\cdot; \psi_g)$ and $f(\cdot; \psi_g^b)$. In full generality, this quantity is different from the KL, but expresses some pseudo-distance between the parameters ψ_g and ψ_g^b. Observe that, by definition, $d(\psi_g, \psi_g^b) \geq 0$ for all $\psi_g \in \Psi_g$.

The main result of this section is to provide an exponential bound for the deviation probability of the $\ln L_{cc}$ contrast. If the contrast is unbounded, up to some additional moment condition, the exponential bound is perturbed by

a polynomial term multiplied by a constant which is a decreasing function of the sample size. As a corollary, we deduce bounds for $d(\hat{\psi}_g, \psi_g^b)$. Notice the necessity for our result to be also adapted to unbounded contrasts in view of applying it to GLM inference (for many GLM distributions, the logarithm of the response density is unbounded). To obtain the exponential bound, we first require an assumption which ensures a domination of the components f_i as well as their derivatives.

Assumption 1. *Assume that Θ is a compact subset of $\mathbb{R}^d$. Denote by $\nabla_\theta f_i(y;\theta_i)$ (resp. $\nabla^2_\theta f_i(y;\theta_i)$) the vector (resp. matrix) of partial derivatives of f_i w.r.t. each component of θ_i. Assume that $\forall \theta \in \Theta$, and all $i = 1, ..., n_g$,*

$$\begin{aligned} f_i(y;\theta) &\geq \tilde{\Lambda}_-(y) > 0, \\ f_i(y;\theta) &\leq \tilde{\Lambda}_0(y) < \infty, \\ |\nabla_\theta f_i(y;\theta)| &\leq \tilde{\Lambda}_1(y), \\ |\nabla^2_\theta f_i(y;\theta)| &\leq \tilde{\Lambda}_2(y), \end{aligned}$$

with $\sup_{l=0,1,2} \tilde{\Lambda}_l(y)\tilde{\Lambda}_-(y)^{-1} \leq \tilde{A}(y)$.

In the case where the functions f_i would not be bounded, we require some moment assumptions on $\tilde{A}(y)$. Let us note that unlike the classical assumptions in the literature ([16], [5], [22]), we do not require the boundedness of the log-contrast and/or of its derivatives with respect to θ.

Assumption 2. *Using the notations of Assumption 1, assume that there exists $m > 0$ such that $\mathbb{E}_{f_0}[\tilde{A}(Y)^{2m}] < \infty$.*

Theorem 1. *Let*

$$P(x;g) = \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \frac{\left\{ \ln L_{cc}(\psi_g; Y_j) - \ln L_{cc}(\psi_g^b; Y_j) + d(\psi_g, \psi_g^b) \right\}}{\|\psi_g - \psi_g^b\|} \right| > x \right),$$

where Ψ_g is a set of parameters such that, for all $\psi_g = (\pi_1, ..., \pi_{n_g}, \theta_1, ..., \theta_{n_g}) \in \Psi_g$, for all $1 \leq i \leq n_g$, $\pi_i \geq \pi_- > 0$. Assume that ψ_g^b is an interior point of Ψ_g. Under Assumptions 1 and 2 with $m - \varepsilon \geq 2$ for some $\varepsilon \geq 0$, there exists four positive constants A_3, A_4, A_5 and A_6 (depending on the parameter space Θ and on the functions f_i only) such that

$$P(x;g) \leq 4\left\{ \exp\left(-\frac{A_3\, x^2}{n} \right) + \exp\left(-\frac{A_4\, x}{n^{1/2-\varepsilon}} \right) \right\} + \frac{A_5}{x^{(m-\varepsilon)/2}},$$

for $x > A_6\, n^{1/2}\, [\ln n]^{1/2}$.

Sketch of the proof.
Define for a single observation, $\phi_{\psi_g}(y) = \phi_{1\psi_g}(y) - \phi_{2\psi_g}(y)$ where

$$\phi_{1\psi_g}(y) = \frac{\left\{\ln f(y;\psi_g) - \ln f(y;\psi_g^b)\right\}}{\|\psi_g - \psi_g^b\|}, \quad \phi_{2\psi_g}(y) = \frac{Ent(\psi_g;y) - Ent(\psi_g^b;y)}{\|\psi_g - \psi_g^b\|}.$$

The proof consists of applying the concentration inequality of Proposition A1 along with Proposition A2 (available in Appendix A) to the class of functions $\mathcal{F}_l = \{\phi_{l\psi_g} : \psi_g \in \Psi_g\}$, for $l = 1, 2$. To apply Proposition A2, we have to check that polynomial bounds on the covering numbers of these two classes hold (condition (i) in Proposition A2). This is done in the first step of the proof. Nevertheless, Proposition A1 and A2 require the boundedness of the class of functions that one considers. Therefore it can only be obtained for a truncated version of these two classes, that is $\mathcal{F}_l \mathbf{1}_{F_l(y)\leq M}$ for some M going to infinity at some reasonable rate, and some appropriate function F_l. The application of the concentration inequality to the truncated version is performed in the second step of the proof. In a third step, the difference between the truncated version and the remainder term is considered. Finally, in a fourth step, all the results are gathered.
These different steps are detailed in Appendix A.

Remark 1: in the bound of $P(x;g)$, two terms decrease exponentially, while a third one decreases in a polynomial way. This additional term is the price to pay for considering potentially unbounded variables Y (see [15] and [16] for related bounds in the bounded case). If we increase the assumptions on Y, by assuming the existence of an exponential moment for $\tilde{A}(y)$ instead of a finite m^{th} moment for m large enough (Assumption 2), a better bound can be obtained (see Appendix C). The existence of such an exponential moment typically holds in the case where one considers bounded variables Y, which lead to a bounded function $\tilde{A}(y)$.

Remark 2: it is easy to see, from the proof of Theorem 1, that a similar bound holds if $\ln L_{cc}$ is replaced by the log-likelihood, and ψ_g^b is the limit of the MLE. Indeed, the proof is divided into proving bounds for the classical log-likelihood, and for the entropy term. In this last situation, note that the restriction of the probabilities π_i to values larger than π_- is not required. This restriction in Theorem 1 was imposed by the behavior of the derivative of the entropy near 0, which could explode otherwise.

2.4. Convergence rates for the $ML_{cc}E$

Corollary 1. *Assume that* $\ln L_{cc}$ *is twice differentiable with respect to* ψ_g, *and denote by* H_{ψ_g} *the Hessian matrix of* $\mathbb{E}_{f^0}[\ln L_{cc}(\psi_g; Y)]$ *evaluated at* ψ_g. *Assume that, for some* $\mathfrak{c} > 0$, $\psi_g^T H_{\psi_g} \psi_g > \mathfrak{c}\|\psi_g\|_2^2$ *for all* $\psi_g \in \Psi_g$, *where* $\|\cdot\|_2$ *denotes the* L^2-*norm. Then, under the assumptions of Theorem 1, for* $m \geq 2$ *in Assumption 2 and for the norm* $\|\cdot\|_2$, *we have*

$$\|\hat{\psi}_g - \psi_g^b\|_2 = O_P\left(\frac{1}{n^{1/2}}\right).$$

If $m > 4$,

$$\|\hat{\psi}_g - \psi_g^b\|_2 = O_{a.s.}\left(\frac{[\ln n]^{1/2}}{n^{1/2}}\right).$$

Proof. Observe that, from a second order Taylor expansion, $d(\hat{\psi}_g, \psi_g^b) \geq \mathfrak{c}\|\hat{\psi}_g - \psi_g^b\|_2^2$. By definition of $\hat{\psi}_g$, we have

$$\sum_{j=1}^{n} \frac{\ln L_{cc}(\hat{\psi}_g; Y_j) - \ln L_{cc}(\psi_g^b; Y_j)}{\|\hat{\psi}_g - \psi_g^b\|_2} \geq 0.$$

Therefore,

$$\sum_{j=1}^{n} \frac{\{\ln L_{cc}(\hat{\psi}_g; Y_j) - \ln L_{cc}(\psi_g^b; Y_j)\}}{\|\hat{\psi}_g - \psi_g^b\|_2} + \frac{nd(\hat{\psi}_g, \psi_g^b)}{\|\hat{\psi}_g - \psi_g^b\|_2} \geq \frac{nd(\hat{\psi}_g, \psi_g^b)}{\|\hat{\psi}_g - \psi_g^b\|_2} \geq \mathfrak{c}n\|\hat{\psi}_g - \psi_g^b\|_2.$$

Applying Theorem 1, we get, for $x > A_6 n^{1/2}[\ln n]^{1/2}$,

$$\mathbb{P}\left(\mathfrak{c}n\|\hat{\psi}_g - \psi_g^b\|_2 > x\right) \leq P(x; g) \leq 4\left\{\exp\left(-\frac{A_3 x^2}{n}\right) + \exp\left(-\frac{A_4 x}{n^{1/2-\varepsilon}}\right)\right\} + \frac{A_5}{x^{(m-\varepsilon)/2}}.$$

Define $E_n(u) = \left\{(n^{1/2}\|\hat{\psi}_g - \psi_g^b\|_2 > u[\ln n]^{1/2}\right\}$. We have $\mathbb{P}(E_n(u)) \leq P(x; g)$ with $x = u\mathfrak{c}n^{1/2}[\ln n]^{1/2}$, if $u > A_6$. Proving the almost sure rate of Corollary 1 is done by applying the Borel-Cantelli Lemma to the sets $\{E_n(u)\}_{n\in\mathbb{N}}$, for some u large enough. We need to show that for some u large enough, $\sum_{n\geq 1} \mathbb{P}(E_n(u)) < \infty$. We have, for $u > A_6$,

$$\begin{aligned}\sum_{n=1}^{\infty} \mathbb{P}(E_n(u)) \quad &\leq \quad \sum_{n=1}^{\infty} \frac{4}{n^{A_3\mathfrak{c}^2u^2}} + \sum_{n=1}^{\infty} 4\exp\left(-A_4 n^{\varepsilon}[\ln n]^{1/2} u\mathfrak{c}\right) \\ &\qquad + \sum_{n=1}^{\infty} \frac{A_5}{u^{\frac{m-\epsilon}{2}} \mathfrak{c}^{\frac{m-\epsilon}{2}} n^{\frac{m-\epsilon}{4}} [\ln n]^{\frac{m-\epsilon}{4}}}.\end{aligned}$$

We see that the first sum in the right-hand side is finite provided that $u > \mathfrak{c}^{-1}A_3^{-1/2}$. The second sum is finite if $\varepsilon > 0$. The third is finite if $m > 4$ and ε taken sufficiently small.

To prove the O_P-rate of Corollary 1, we need to show that $p_n(u) = \mathbb{P}(E_n(u/[\ln n]^{1/2}))$ tends to zero when u tends to infinity. Using the same arguments as before, for $m \geq 2$,

$$p_n(u) \leq 4\exp(-A_3u^2) + 4\exp(-A_4[\ln n]^{1/2}u) + 2^{4m}A_5\mathfrak{c}^{-m/2}u^{-m/2},$$

where the right-hand side tends to zero when u tends to infinity. □

Remark 4: it also follows the proof of Corollary 1 the stronger result

$$\frac{1}{n}\sum_{j=1}^{n}\frac{\{\ln L_{cc}(\hat{\psi}_g;Y_j) - \ln L_{cc}(\psi_g^b;Y_j)\}}{\|\hat{\psi}_g - \psi_g^b\|} + \frac{d(\hat{\psi}_g,\psi_g^b)}{\|\hat{\psi}_g - \psi_g^b\|} = O_{a.s.}\left([\ln n]^{1/2}n^{-1/2}\right).$$

This implies

$$\frac{1}{n}\sum_{j=1}^{n}\{\ln L_{cc}(\hat{\psi}_g;Y_j) - \ln L_{cc}(\psi_g^b;Y_j)\} + d(\hat{\psi}_g,\psi_g^b) = O_{a.s.}\left([\ln n]n^{-1}\right). \quad (6)$$

3. A recent selection criterion designed for clustering: ICL*

As mentioned earlier, a crucial issue in clustering and mixture analysis is to determine the appropriate order of the mixture to correctly describe the population structure. [7] tried to circumvent the challenge faced by BIC as for selecting the right number of classes, especially in the case of a misspecified mixture model. He wanted to emulate the BIC approach by replacing the observed likelihood by the classification one, which should partially eliminate the problem of overestimating this order. This way, he expected to find a criterion that allows achieving a better compromise between the classification quality and the fit to data. This criterion, the so-called ICL, is henceforth well suited to issues of population clustering. However, a particular attention should be paid to the definition of the penalty term: early works used to consider the entropy as part of the penalty, but no theoretical results have been demonstrated from this viewpoint despite promising results in practical applications ([6]). [4] then proposed a redefinition of ICL, leading to his new ICL* criterion. The idea behind the construction of ICL* by [4] is to combine BIC with the L_{cc} contrast. In this regard, we have in the previous section

shown the strong convergence of the $ML_{cc}E$ towards the theoretical parameter of the underlying distribution under particular regularity conditions. We now focus on the selection process from a finite collection of nested models M_g, with $g = \{1, ..., G\}$.

3.1. Previous works on ICL criteria

ICL was defined on the same basis as the BIC criterion: [7] suggests to select in the collection the model satisfying

$$M^{ICL} = \underset{M_g \in \{M_1,...,M_G\}}{\arg\min} \Big(- \max_{\psi_g \in \Psi_g} \ln L_c(\psi_g; \mathbf{Y}, \delta) + \frac{K_g}{2} \ln n \Big).$$

In practice, one approximates $\arg\max_{\psi_g} L_c(\psi_g; \mathbf{Y}, \delta)$ by $\hat{\psi}_g^{ML}$ when n gets large, which leads to

$$\begin{aligned} M^{ICL_a} &= \underset{M_g \in \{M_1,...,M_G\}}{\arg\min} \Big(- \ln L_c(\hat{\psi}_g^{ML}; \mathbf{Y}, \hat{\delta}^B) + \frac{K_g}{2} \ln n \Big) \\ &= \underset{M_g \in \{M_1,...,M_G\}}{\arg\min} \Big(- \ln L(\hat{\psi}_g^{ML}; \mathbf{Y}) - \sum_{j=1}^{n} \sum_{i=1}^{n_g} \hat{\delta}_{ij}^B \ln \tau_i(Y_j; \hat{\psi}_g^{ML}) + \frac{K_g}{2} \ln n \Big). \end{aligned}$$

This approximation is questionable since the associated contrast is different from the classical likelihood. Besides, the label vector δ is inferred from using the Bayes rule on *a posteriori* probabilities: this implies that the predicted labels, denoted further $\hat{\delta}^B$, also depends on the MLE. Almost simultaneously, [26] suggest to use the *a posteriori* probabilities $\tau_i(y; \hat{\psi}_g^{ML})$ instead of $\hat{\delta}^B$:

$$\begin{aligned} M^{ICL_b} &= \underset{M_g \in \{M_1,...,M_G\}}{\arg\min} \Big(- \ln L_c(\hat{\psi}_g^{ML}; \mathbf{Y}, \tau(\hat{\psi}_g^{ML})) + \frac{K_g}{2} \ln n \Big) \\ &= \underset{M_g \in \{M_1,...,M_G\}}{\arg\min} \Big(- \ln L(\hat{\psi}_g^{ML}; \mathbf{Y}) + \underbrace{Ent(\hat{\psi}_g^{ML}) + \frac{K_g}{2} \ln n}_{pen^{ICL_b}(K_g)} \Big). \end{aligned}$$

In fact, ICL_a and ICL_b are really different only if $\forall i,\ \tau_i(Y_j; \hat{\psi}_g^{ML}) \simeq 1/n_g$. Some basic algebra shows that $ICL_a \geq ICL_b$: this means that ICL_a penalizes to a greater extent a model whose observations allocation is uncertain than does ICL_b. [7] and [26] have shown, through various simulated and real-life examples, that ICL is more robust than BIC when the model is

misspecified (which is often the case in reality). Granted, BIC and ICL have similar behaviors when the mixture components are distinctly separated; but ICL severely penalizes the likelihood in the reverse case, still taking into account its complexity. Nevertheless there is no clear relationship between the maximum likelihood theory and the entropy, so that the criterion defined as such is not fully satisfactory from a theoretical viewpoint. Indeed, its properties have not been proved yet: for instance it is not consistent in the sense that BIC is, because its penalty does not satisfy Nishii's conditions ([28]). In particular, it is not negligible in front of n:

$$\frac{1}{n} Ent(\psi_g;\, \mathbf{Y}) \xrightarrow[n\to\infty]{\mathbb{P}} \mathbb{E}_{f^0}\left[Ent(\psi_g;\, Y)\right] \;>\; 0$$

It follows that $Ent(\psi_g;\, \mathbf{Y}) = O(n)$. This gap between the practical interest aroused by ICL and its theoretical justification was plugged by [4], who defined ICL* as a new version of ICL with a BIC-type penalty:

$$M^{ICL^*} = \underset{M_g \in \{M_1,\ldots,M_G\}}{\arg\min} \Big(-\ln L_{cc}(\hat{\psi}_g^{ML_{cc}}) + \frac{K_g}{2} \ln n \Big).$$

In the context of gaussian mixtures for bounded variables, [4] has shown that the number of components selected using this criterion converges weakly towards the theoretical one.

3.2. Consistency of selection criteria

Consider the mixture framework, and let M_{g^*} denote the model with smallest dimension K_{g^*} such that $\mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_{g^*})] = \max_{g=1,\ldots,G} \mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_g)]$. The following theorem provides consistency properties for a large class of penalized estimators. Related results can be found in [5].

Theorem 2. *Consider a collection of models $(M_1, ..., M_G)$ satisfying the assumptions of Theorem 1. Consider a penalty function $pen(M_g) = K_g u_n$,*

$$\hat{g} = \underset{g=1,\ldots,G}{\arg\max} \left(\frac{1}{n} \sum_{j=1}^{n} \ln L_{cc}(\hat{\psi}_g; Y_j) - pen(M_g) \right),$$

and assume that $\psi^b_g = \psi^b_{g^*}$ for M_g such that $K_g \geq K_{g^*}$.
Then, if $m > 2$ in Assumption 2 and if $nu_n \to \infty$, we get $\forall g \neq g^*$

$$\mathbb{P}(\hat{g} = g) = o(1).$$

Also, if $m > 4$ in Assumption 2, there exists some constant C such that, if $nu_n > C \ln n$, almost surely, $\hat{g} \neq g$ for n large enough and for all $g \neq g^*$.

Proof. Let $\varepsilon_g = \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi^b_{g^*}; Y)\right] - \mathbb{E}_{f^0}\left[\ln L_{cc}(\psi^b_g; Y)\right]$.

Decompose

$$\begin{aligned}\frac{1}{n}\sum_{j=1}^{n} \ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\hat{\psi}_g; Y_j) \leq \varepsilon_g &+ \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g^*}; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_{g^*}; Y)]\right\} \\ &- \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g}; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_{g}; Y)]\right\} \\ &+ \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\psi^b_{g^*}; Y_j)\right\} \\ &- \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\hat{\psi}_{g}; Y_j) - \ln L_{cc}(\psi^b_{g}; Y_j)\right\}. \end{aligned} \tag{7}$$

It follows from the remark following Corollary 1 that the last two terms in (7) are $O_{a.s.}([\ln n]n^{-1})$ (or $O_P(n^{-1})$). We now distinguish two cases: $\epsilon_g > 0$ and $\epsilon_g = 0$.

Case 1: $\epsilon_g > 0$**.**
It follows from the law of iterated logarithm that

$$\begin{aligned}&\left|\frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g^*}; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_{g^*}; Y)]\right\}\right| \\ &+\left|\frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g}; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi^b_{g}; Y)]\right\}\right| \quad = \quad O_{a.s.}([\ln\ln n]^{1/2}n^{-1/2}).\end{aligned}$$

Note that these two terms are $O_P(n^{-1/2})$ if we only focus on O_P-rates.

If $\hat{g} = g$, we have

$$\frac{1}{n}\sum_{j=1}^{n}\left(\ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\hat{\psi}_g; Y_j)\right) - pen(M_{g^*}) + pen(M_g) < 0. \tag{8}$$

However, due to the previous remarks, if we take $u_n = o(1)$, the left-hand side in (8) converges almost surely towards ε_g (in probability rates, is equal to $\varepsilon_g + o_P(1)$). This ensures that M_g is almost surely not selected for n large enough (in probability rates, $\mathbb{P}(\hat{g} = g) = o(1)$).

Case 2: $\epsilon_g = 0$**.**
Since $\psi_g^b = \psi_{g^*}^b$,

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi_{g^*}^b; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi_{g^*}^b; Y)]\right\} = \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi_g^b; Y_j) - \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g^b; Y)]\right\},$$

and $\varepsilon_g = 0$, which shows that the first three terms in (7) are zero. This leads to

$$\frac{1}{n}\sum_{j=1}^{n}\left(\ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\hat{\psi}_g; Y_j)\right) - pen(M_{g^*}) + pen(M_g) \geq \begin{cases} u_n + O_{a.s.}\left(\frac{\ln n}{n}\right) \\ u_n + O_P\left(\frac{1}{n}\right), \end{cases}$$

since $K_g - K_{g^*} > 1$. This shows that there exists a constant $C > 0$ such that, if $nu_n > C \ln n$, M_g is almost surely not selected when n tends to infinity. To obtain that $\mathbb{P}(\hat{g} = g) = o(1)$, it is sufficient to have nu_n tending to infinity in this context. □

4. Application to the selection of GLM mixture models

GLM have been introduced as an extension of classical linear models where the response variable is assumed to be the realization of a random variable belonging to an exponential family. They are a common way to integrate specific risk factors; and include analysis-of-variance models, logit and probit models for quantal responses, log-linear models and multinomial response models for counts, but also classical models for survival data. Due to this flexibility, GLM are nowadays used in many fields among which marketing, economics, medicine, astronomy. Meanwhile the popularity of finite mixtures of GLM has steadily grown, notably because they are a natural way to deal with heterogeneous impacts of some risk factors on the phenomenon under study. For instance, it has become a standard tool in insurance pricing ([27], [30]). The pioneering work by **(author?)** [38] on the likelihood optimization in GLM mixtures was then followed by some seminal papers (**(author?)** [1], **(author?)** [23], **(author?)** [26], **(author?)** [17], **(author?)** [18] and [19]). However, the question of selecting an appropriate GLM mixture to perform clustering of a population has not really been tackled in this context. This question of optimizing the clustering properties of the mixture is essential in applications such as insurance pricing, in which classes of insurance customers need to be identified. In this section, we show how the results on ICL* model selection apply to GLM, by checking the assumptions required for consistency.

4.1. Definitions relative to the GLM framework

Following the notations of section 2, consider i.i.d. replications $(Y_j)_{1\leq j\leq n}$ where $Y_j = (Z_j, X_j)$. Here, Z_j stands for the j$^{\text{th}}$ **response** associated to the j$^{\text{th}}$ column vector of **covariates** $X_j = (1, X_{j1}, ..., X_{jp})$, where X_j belongs to a compact set $\mathcal{X} \subset \mathbb{R}^{p+1}$. Introduce the column vector of regression coefficients $\beta = (\beta_0, ..., \beta_p)$, where $\beta \in B$ for some compact $B \subset \mathbb{R}^{p+1}$. In a Generalized Linear Model (see the seminal book of [25]), one assumes that the two following assumptions hold:

i) the conditional distribution of $Z|X$ belongs to the exponential family,
ii) there exists an invertible **link** w, such that $w(\mathbb{E}[Z|X]) = X^T\beta$.

This yields that the conditional distribution of $Z|X = x$ admits the following probability density with respect to an appropriate dominating measure,

$$f_{Z|X=x}(z, \phi) = \exp\left(\frac{z\alpha(x^T\beta) - b(\alpha(x^T\beta))}{a(\phi)} + c(z, \phi)\right), \tag{9}$$

where a is a positive function, b is a twice differentiable strictly convex function, and where $\alpha(u) = b'^{-1}(w^{-1}(u))$. A particular case consists of choosing the canonical link function, that is $w = b'$, which ensures that α is the identity function. The parameter ϕ corresponds to a dispersion parameter. Table 1 summarizes the canonical links $w(k)$ as well as the relations between the parameters of the exponential family and those of some well-known distributions. Only β is involved in the expression of $E[Z|X]$, which makes it

Law	$\mathcal{N}(\mu, \sigma^2)$	$\mathcal{B}(n, p)$	$\mathcal{P}(\mu)$	$\mathcal{G}(\nu, \nu/\eta)$	$\mathcal{IN}(\mu, \sigma^2)$
	$z \in \mathbb{R}$	$z \in [\![0, n]\!]$	$z \in \mathbb{N}$	$z \in \mathbb{R}^+$	$z \in \mathbb{R}^+$
	$\mu \in \mathbb{R}$	$p \in [0, 1]$	$\mu \in \mathbb{R}^+$	$\eta \in \mathbb{R}^{+*}$	$\mu \in \mathbb{R}^{+*}$
	$\sigma^2 \in \mathbb{R}^{+*}$			$\nu \in \mathbb{R}^{+*}$	$\sigma^2 \in \mathbb{R}^{+*}$
$\mathbb{E}[Z\|X]$	μ	np	μ	η	μ
$w(k)$	$Id(k)$	$\ln(k/(1-k))$	$\ln(k)$	$1/k$	$1/k^2$
Model	$\mu = X\beta$	$p = (1 + e^{-X\beta})^{-1}$	$\mu = e^{X\beta}$	$\eta = (X\beta)^{-1}$	$\mu = (X\beta)^{-1/2}$
α	μ	$\ln(p/(1-p))$	$\ln(\mu)$	$-\eta^{-1}$	$-1/(2\mu^2)$
ϕ	σ^2	1	1	ν^{-1}	$1/\sigma^2$
$b(\alpha)$	$\alpha^2/2$	$n\ln(1 + e^{\alpha})$	e^{α}	$-\ln(-\alpha)$	$-(-2\alpha)^{1/2}$
$a(\phi)$	ϕ	ϕ	ϕ	ϕ	ϕ
$c(z, \phi)$	$-\frac{1}{2}\frac{z^2}{\phi}$	$\ln(C_n^z)$	$-\ln(z!)$	$\phi^{-1}\ln(\phi^{-1}z)$	$-\frac{1}{2}\ln(2\pi\phi z^3)$
	$-\frac{1}{2}\ln(2\pi\phi)$			$-\ln(z) - \ln(\Gamma(\phi^{-1}))$	$-\frac{1}{2}(\phi z)^{-1}$

Table 1: Classical members of the exponential family and related specifications.

the most important quantity to estimate, whereas ϕ is often considered as a nuisance parameter. Apart from the model selection issue, potential difficulties in GLM mixtures concern the identifiability due to the existence of the covariates. Further details can be found in [36] and [26] (p.146); and special cases about the Poisson regression model as well as the binomial regression model are available in [37] and [12] respectively.

4.2. Conditions for applying the ICL methodology to GLM mixtures*

Back to the notations of the previous sections, mixtures of GLM consist of assuming that the distribution of $Y = (Z, X)$ is $f(y; \psi_g) = \sum_{i=1}^{n_g} \pi_i f_i(y; \theta_i)$, with

$$f_i(y; \theta_i) = \exp\left(\frac{z\alpha_i(x^T\beta_i) - b_i(\alpha_i(x^T\beta_i))}{a_i(\phi_i)} + c_i(z, \phi_i)\right) f_X(x),$$

where f_X denotes the density of X (since f_X does not depend on the parameters, this quantity does not play any role in the definition of the log-likelihood and can be removed), and $\theta_i = (\beta_{i0}, ..., \beta_{ip}, \phi_i)$. Note that, in full generality, we can mix different members of the exponential family, which is the reason for considering functions $(\alpha_i, b_i, a_i, c_i)$ which may be different for each i. Nevertheless, because this is the model class resorted to in practice, we will consider the same exponential family for each component of the mixture. We recall that $\alpha_i(u) = b_i^{'-1}(w_i^{-1}(u))$, and we assume that for each component of the mixture, w_i is twice continuously differentiable, with $|w_i'| > 0$.

To ensure consistency of the ICL* model selection procedure, we need to check that Assumptions 1 and 2 hold. We first determine appropriate functions Λ_- and Λ_j for $j = 0, 1, 2$, following the notations of Assumption 1. We first assume that

$$c_-(z) \leq c(z, \phi_i) \leq c_+(z), \tag{10}$$

for all z, i, and ϕ_i. If there is no dispersion parameter (as in the Poisson case, for instance), $c(z, \phi) = c_-(z) = c_+(z)$. Next, define $\alpha_{min} = \inf_{i,\beta_i,x} \alpha_i(x^T\beta)$, $\alpha_{max} = \sup_{i,\beta_i,x} \alpha_i(x^T\beta)$, $a_{min} = \inf_{i,\phi_i} a(\phi_i)$, and $a_{\max} = \sup_{i,\phi_i} a(\phi_i)$. This quantities are finite, since we assumed that the parameters and covariates belong to a compact set. Without loss of generality, we will assume that b is positive, which can always be the case up to some modification of function c. The function Λ_- corresponds to a lower bound for the densities f_i. Using the previous notations, it is straightforward to check that we can take

$$\Lambda_-(y) = \exp\left(\frac{z\alpha_{min}}{a_{max}} + c_-(z)\right).$$

Similarly, $\Lambda_0(y)$ is an upper bound for the densities, and can be taken as

$$\Lambda_0(y) = \exp\left(\frac{z\alpha_{max}}{a_{min}} + c_+(z)\right).$$

Functions Λ_1 and Λ_2 can be obtained by looking at the derivatives of f_i with respect to the components of β_i and ϕ_i. Observe that, since b_i' is strictly convex and w_i is invertible with $|w_i'| > 0$, the first and second order derivatives of $\alpha_i(u)$ are bounded by a constant $C > 0$ if we restrict the values of u to the compact of all possible values for $x^T\beta_i$ when $x \in \mathcal{X}$ and $\beta_i \in B_i$.

We easily get

$$\begin{aligned}
\Lambda_1(y) &= \left(\frac{z(C\sup_{x\in\mathcal{X}}\|x\|_\infty + \alpha_{max}\sup_{i,\phi_i}|a_i'(\phi_i)a_i(\phi_i)^{-1}|)}{a_{min}} + \sup_{i,\phi_i}|c_i'(z,\phi_i)|\right)\Lambda_0(z),\\
\Lambda_2(y) &= \Big(\frac{(z+z^2)(\sup_{x\in\mathcal{X}}\|x\|_\infty^2(C+\alpha_{max}^2))\sup_{i,\phi_i}|a_i'(\phi_i)a_i(\phi_i)^{-1}|)(1+\sup_{i,\phi_i}|c'(z,\phi_i)|)}{a_{min}^2}\\
&\quad + \frac{z\alpha_{max}}{a_{min}}\sup_{i,\phi_i}\left|\frac{a_i''(\phi_i)a_i(\phi_i) - 2a_i'(\phi_i)}{a_i(\phi_i)}\right| + \sup_{i,\phi_i}|c''(z,\phi_i)|\Big)\,\Lambda_0(z).
\end{aligned}$$

where $c_i'(z,\phi_i)$ (resp. $c_i''(z,\phi_i)$) denotes the first (resp. second) order derivative of c_i with respect to ϕ_i. If we impose:

- $\max(\sup_{i,\phi_i}|a_i'(\phi_i)a_i(\phi_i)^{-1}|), \sup_{i,\phi_i}\left|\frac{a_i''(\phi_i)a_i(\phi_i)-2a_i'(\phi_i)}{a_i(\phi_i)}\right|) < \infty$,
- $|c'(z,\phi_i)| + |c''(z,\phi_i)| \le C_0 z^2$,

for some constant C_0, then Assumption 1 holds with

$$\tilde{A}(y) = C_1 z^2 \exp(C_2 z)\exp(c_+(z) - c_-(z)),$$

for some constants C_1 and C_2. For most of classical families, Assumption 2 holds, eventually up to some restriction of the parameter space. Let us also note that, if we were performing classical maximum likelihood inference (instead of $ML_{cc}E$), we could separate the problem of estimating ϕ_i of the problem of estimating β_i. In this case, the term $\exp(c_+(z)-c_-(z))$ disappears, and the condition of Assumption 2 can be simplified.

5. Simulation study

In this section, we perform a simulation study to check the previous theoretical results and validate our convergence properties. By sampling observations coming from finite mixture models (first, mixtures of normal regressions, then mixtures of Poisson regressions), we show that i) the $ML_{cc}E$ tends towards the parameter maximizing the expected log-contrast, ii) ICL* looks consistent while being adapted for clustering purposes. The well-known tendency to overestimate the order of the mixture (when using AIC or BIC) is less obvious, which is a very good news since ICL* was designed to this end. For practical considerations, the latter result is also interesting: the model dimension is lowered, which should bring more robustness to the parameters estimation and probably lead to more relevant predictions. Moreover we mecanically decrease the probability to make a mistake when assigning observations to mixands in less complex mixture models, a nice and desired feature in a clustering perspective.

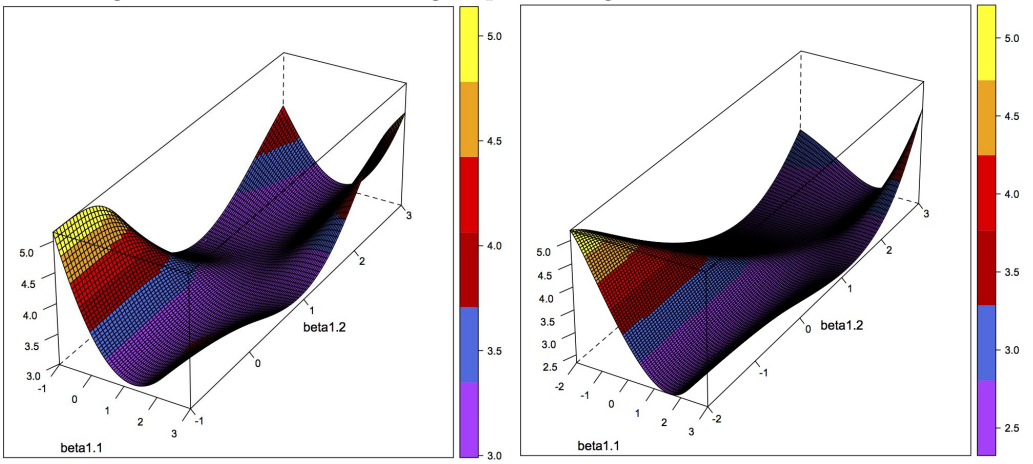
For the sake of simplicity, we consider two-component GLM mixtures with no intercept and a unique covariate: the random design is generated from a uniform distribution on some interval $[a, b]$. Thanks to the maximum likelihood estimation properties, the theoretical maximizer of the classical log-likelihood is obviously the theoretical parameter itself. On the contrary and not surprisingly, it could be quite difficult to find the maximizer of the L_{cc} contrast because of the entropic term. However, this is the first mandatory step so as to check the convergence properties of the $ML_{cc}E$.

5.1. Empirical convergence of the $ML_{cc}E$

In the sequel, 10 000 uniformly-distributed observations X_j are sampled to compute ψ^0. In both applications mixture weights are set constant in the optimization process to gain some computation time, as well as variances in the normal regression case. We thus have $\pi_1 = \pi_2 = 0.5$, with standard deviations $\sigma_1 = \sqrt{10}$, $\sigma_2 = 2$ in the normal regression mixture setting. Table 2 gives the theoretical parameters to be reached by the M-estimator, and Figure 1 illustrates how the Kullback-Leibler divergence between the L_{cc} contrast and the true distribution behaves at the MLE neighborhood. Notice that the theoretical maximizer of the L_{cc} contrast is not very close to the theoretical maximizer of the log-likelihood, while still being comparable.

Now we simulate random samples of normal and Poisson regression mixtures (respectively with the same true densities f^0 as previously, see Table 2),

Figure 1: KL divergence between the true distribution and the L_{cc} contrast. On the left: normal regression mixture. On the right: poisson regression mixture.

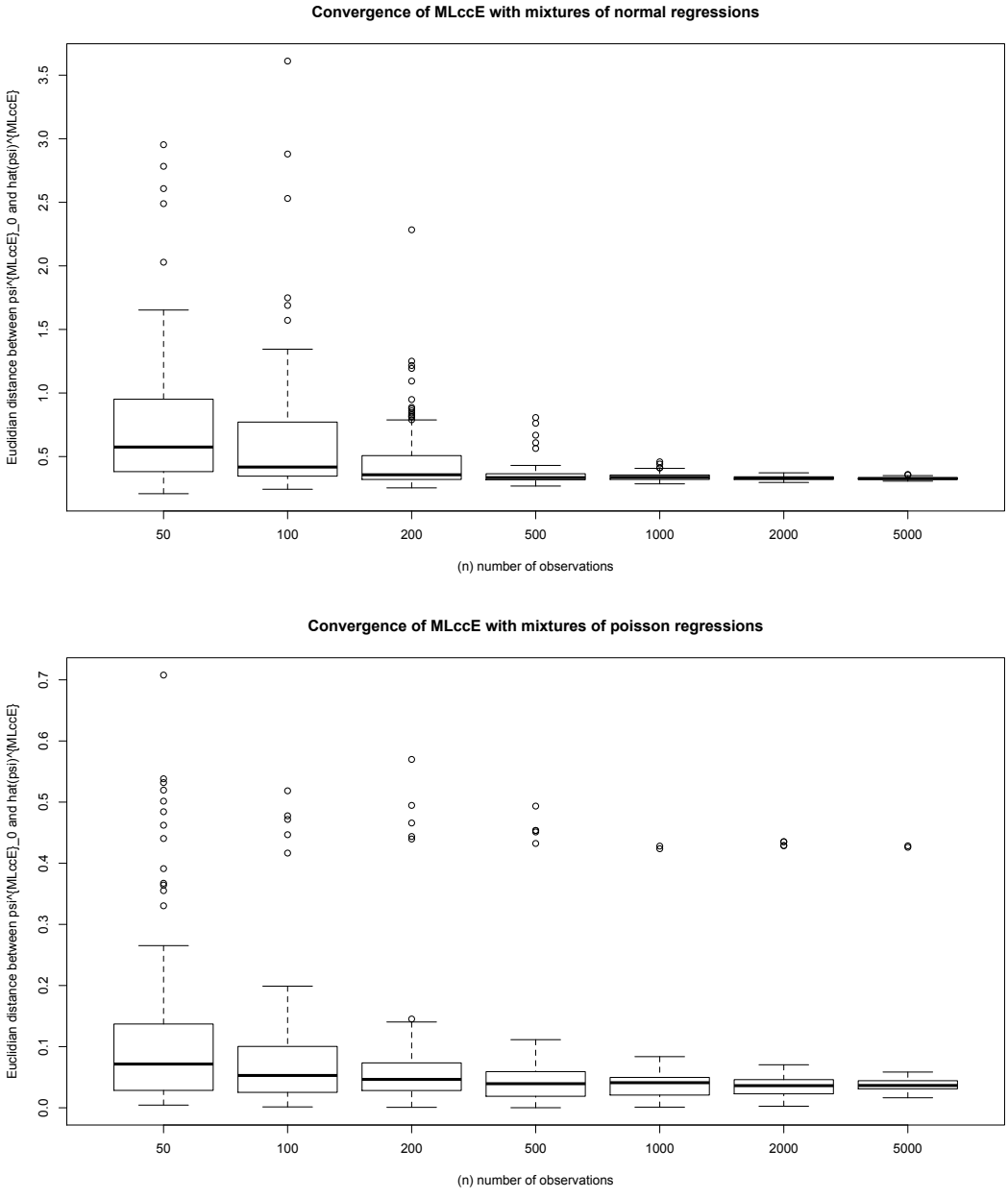


and see whether the $ML_{cc}E$ tends towards the L_{cc} maximizer. The idea is to repeat this procedure 100 times, and then study the mean and the standard deviation of the estimator values. This way, the $ML_{cc}E$ empirical behaviour can be investigated adequately. We expect that the mean of the euclidian distance between the $ML_{cc}E$ and the L_{cc} maximizer tends to 0, with a dispersion that narrows down when the number of observations increases. Results are summarized in Figure 2, and confirm this convergence whatever the type of the random variable considered. Indeed, mixtures of linear regressions stand for the continuous case whereas mixtures of Poisson regressions represent the discrete case. Despite the high number of observations, notice that the maximization can still lead to some erroneous estimations (especially in the Poisson case): this could be explained by the contrast complexity and some difficulties experienced in the optimization algorithms.

Model class:	X	True β_1	True β_2	L_{cc} maximizer β_1^0	L_{cc} maximizer β_2^0
Linear regression	$\sim \mathcal{U}(0,1)$	1	1.3	0.5	1.66
Poisson regression	$\sim \mathcal{U}(0,1)$	1.1	1.6	-0.16	1.75

Table 2: Minimizers of the KL divergence between f^0 and the L_{cc} contrast.

Figure 2: Boxplot (100 experiments) $ML_{cc}E$ convergence towards the maximizer of the expected L_{cc} contrast. From top to bottom: normal and poisson regression mixtures.



5.2. Illustration of ICL consistency*

As Figure 2 suggests, we should consider at least 2000 observations to ensure reasonable convergence properties of the $ML_{cc}E$. There are two interesting situations in which the consistency of ICL* should be tested: the first one relates to the selection of a mixture density where components are strongly overlapping, whereas the other one corresponds to well-separated component densities. Theoretically speaking, ICL may not be too different from BIC in the latter case because the entropic term must be negligible. In other words, these two criteria should lead to similar results as they use the same estimator (MLE) apart from that. On the contrary, although the penalty term is exactly alike for ICL* and BIC, the ICL* selection process is based on the $ML_{cc}E$ and this is clearly censed to affect the model selection in a different manner in the overlapping situation. The entropic term is obviously no longer negligible in such a case, and the selection criteria have no reason to behave similarly. In particular, highly overlapping component densities is closely linked to low confidence when assigning observations to mixture components once the model fitted. The use of ICL* should therefore result in selecting a simpler model and tend to avoid the problem of overestimating g. To check it, let us consider 30 experiments for which the following steps are undertaken:

1. $(X_j)_{1\leq j\leq 2000}$ is sampled from the uniform distribution;
2. draw a 3-component mixture with user-defined parameters;
3. fit 4 different mixture models (from 2 to 5 components): for each one,
   (a) find the MLE and $ML_{cc}E$ corresponding to the empirical density,
   (b) compute the model selection criteria (AIC, BIC and ICL from the MLE; and ICL* from the $ML_{cc}E$ with the same penalty as BIC);
4. for each model selection criterion, the selected model corresponds to the minimum over the 4 available criterion values.

This algorithm is performed for both normal regression mixtures and poisson regression mixtures respectively. Concerning mixture parameters, they are stored in Table 3 (except for the standard deviations in the normal regression case which all equal to $\sqrt{3}$). These parameters were randomly chosen, and we checked that this choice had no influence on our final results to guarantee their robustness (by changing these values to other coherent ones).
Tables 4 and 5 offers an overview of ICL* performance by summarizing the statistics over these 30 experiments for these two model classes: the goal is to

Mixture parameters:	π_1	π_2	π_3	β_1	β_2	β_3	X
Normal regression							
Well-separated case	1/3	1/3	1/3	0.5	20	40	$\sim \mathcal{U}(1,2)$
Overlapping case	1/3	1/3	1/3	0.5	6	12	$\sim \mathcal{U}(1,2)$
Poisson regression							
Well-separated case	0.3	0.4	0.3	-0.5	2	4	$\sim \mathcal{U}(1,1.5)$
Overlapping case	0.3	0.4	0.3	-1	0.2	0.5	$\sim \mathcal{U}(1,4)$

Table 3: True parameters for the simulation of mixture models.

Model complexity (# components):	2	3	4	5	% overestimation	% right g
Distinct components						
AIC	4	8	7	11	60%	27%
BIC	4	8	7	11	60%	27%
ICL	4	9	6	11	57%	30%
ICL*	0	21	3	6	**30%**	**70%**
Overlapping components						
AIC	4	13	5	8	43%	43%
BIC	4	13	5	8	43%	43%
ICL	6	13	5	6	37%	43%
ICL*	7	23	0	0	**0%**	**77%**

Table 4: Consistency of ICL* in the case of mixtures of normal regressions.

Model complexity (# components):	2	3	4	5	% overestimation	% right g
Distinct components						
AIC	0	10	12	8	67%	33%
BIC	0	11	12	7	63%	37%
ICL	0	14	10	6	53%	47%
ICL*	4	17	3	6	**30%**	**57%**
Overlapping components						
AIC	2	10	6	12	60%	**33%**
BIC	2	10	5	13	60%	**33%**
ICL	20	5	4	1	**17%**	17%
ICL*	11	8	9	0	30%	27%

Table 5: Consistency of ICL* in the case of mixtures of poisson regressions.

see whether using the ICL* criterion leads to select an appropriate mixture model, knowing that the true model has only 3 components. In most of cases, its performance looks satisfactory: it generally avoids the problem of selecting too much complex mixtures (% overestimation) and looks better than AIC, BIC and ICL when trying to recover the right number of com-

ponents. However the case of overlapping components in poisson regression mixtures is somehow problematic, in the sense that the percentage of right predictions for the number of components is not really satisfying (even if the probability to overestimate it is once again diminished). This is certainly linked to what was observed on Figure 2, and shows that there are still come cases where the $ML_{cc}E$ is far from the best possible estimator. In this case the selection process simply looses it efficiency because it is based on a poor estimator, and its consistency deteriorates.

Conclusion

We developed new results in clustering population from mixtures of generalized linear models. In this context this is a key matter since most of famous selection criteria have a well-known tendency to overestimate the order of the mixture. This means that the actual impact of covariates over the response variable is not adequately captured. Motivated by this, our technique is based on some theoretical extensions to the works by [5]: it embraces both the convergence rate of the $ML_{cc}E$ and the consistency of ICL* in the case of unbounded contrasts, which is crucial when dealing with GLM mixtures. The bounds that we obtained for the estimation error, through concentration inequalities, hold even in a non-asymptotic framework. Concerning the model selection step, the simulation study confirms that using ICL* makes the tendency to overestimate the number of components disappear (whatever the type of outcome, continuous or categorical). The position of ICL* for segmentation purpose when observing large heterogeneity within the population under study is thus strengthened. For future research, it would be tempting to adapt this concept to other practical matters ([21]) and seek the theoretical properties of such estimators: integrating specific quantities within the contrast should lead to promising developments.

Acknowledgements

We gratefully thank Jean-Patrick Baudry, Bernard Garel and Victor-Emmanuel Brunel for their insightful and constructive comments on this project.

Appendix A. Steps of the proof of Theorem 1

Proof. **Step 1: covering numbers requirements.**
Let $\mathcal{A}_i = \{\pi_i\, f_i(y;\theta) : \theta \in \Theta, \pi_i \in [\pi_-, 1]\}$. Due to Assumption 1, a first order Taylor expansion shows that

$$\left| \frac{\ln f(y;\psi_g) - \ln f(y;\psi_g^b)}{\|\psi_g - \psi_g^b\|} \right| \leq n_g\, d\, \frac{\tilde{\Lambda}_1(y)}{\tilde{\Lambda}_-(y)},$$

where we recall that d is the dimension of Θ. So the class $\mathcal{F}_1$ admits the envelope $F_1(y) = \nabla_{\psi_g} \ln f(y;\psi_g^b) + n_g\, d\tilde{A}(y)\mathrm{diam}(\Psi_g)$, where $\mathrm{diam}(\Psi_g)$ denotes the diameter of Ψ_g with respect to $\|\cdot\|$. Observe that, from Assumption 1, it follows from a second order Taylor expansion and Lemma 2.13 in [32] that $N_{F_1}(\varepsilon, \mathcal{A}_i) \leq C_1 \varepsilon^{-V_1}$, for some constants $C_1 > 0$ and $V_1 > 0$. Since $\mathcal{F}_1 = \sum_{i=1}^{n_g} \mathcal{A}_i$, Lemma 16 in [29] applies, so that $N_{F_1}(\varepsilon, \mathcal{F}_1) \leq C_1 n_g^{n_g V_1} \varepsilon^{-n_g V_1}$.

For the class $\mathcal{F}_2$, the bound on the covering number is a consequence of Lemma B2 in Appendix B. The assumptions of Lemma B1, required to obtain Lemma B2, clearly hold from Assumption 1, with $\Lambda_-(y) = \tilde{\Lambda}(y)$, $\Lambda_0(y) = \tilde{\Lambda}_0(y)$, $\Lambda_1(y) = d\tilde{\Lambda}_1(y) + \tilde{\Lambda}_0(y)$, $\Lambda_2(y) = 2^{-1}d^2\tilde{\Lambda}_2(y)$. The envelope of $\mathcal{F}_2$ is $F_2(y) = n_g[\Lambda_3(y)\mathrm{diam}(\Psi_g) + \sup_{i=1,\ldots,n_g} |g_{i,\psi_0}(y)|]$, with $\Lambda_3(y) = CA(y)^3$.

Step 2: concentration inequality for truncated classes.
Introduce a constant $M_n > 0$, and consider the classes $\mathcal{F}_l^{M_n} = \mathcal{F}_l \mathbf{1}_{F_l(y) \leq M_n}$ for $l = 1, 2$, where the functions F_l are the envelope functions defined in Step 1. Observe that the covering number of $\mathcal{F}_l^{M_n}$ can be bounded using the same bound as in Step 1, since truncation does not alter this bound (this can be seen as a consequence of Lemma A.1 in [10]). Let

$$P_{1l}(x;g) = \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \sum_{j=1}^{n} \{\phi_{l\psi_g}(Y_j) - \mathbb{E}_{f_0}[\phi_{l\psi_g}(Y)]\} \mathbf{1}_{F_l(Y_j) \leq M_n} > x \right).$$

To bound this probability, we combine Proposition A1 and Proposition A2. The requirements (ii) and (iii) of Proposition A2 hold with $M = M_n$, $\sigma^2 = \mathbb{E}[F_l(Y)^2]$, while the requirement (i) follows from Step 1. Observe that M_n can be taken large enough so that $M_n \geq \sigma$ (in Step 4 of the proof, we will make M_n tend to infinity). Following the notations of Proposition A2, introducing a sequence of Rademacher variables $(\varepsilon_j)_{1 \leq j \leq n}$ independent from $(Y_j)_{1 \leq j \leq n}$, we get

$$\mathbb{E}_{f_0}\left[\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \varepsilon_j \phi_{l\psi_g}(Y_j) \mathbf{1}_{F_l(Y_j) \leq M_n} \right| \right] \leq C_g\, n^{1/2} [\log(M_n)]^{1/2}, \tag{A.1}$$

where C_g is a constant depending on K_g, the dimension of the model.

Taking $u = x(2A_1)^{-1}$ in Proposition A1, we get, for $x > 2A_1 C_g n^{1/2}[\log M_n]^{1/2}$, that the probability $P_1^{(l)}(x;g)$ is bounded by

$$\mathbb{P}\left(\sup_{\psi_g\in\Psi_g}\left|\sum_{j=1}^{n}\{\phi_{l\psi_g}^{M_n}(Y_j) - \mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n}(Y)]\}\right| > A_1\left(\mathbb{E}_{f_0}\left[\sup_{\psi\in\Psi_g}\left|\sum_{j=1}^{n}\varepsilon_j\phi_{l\psi_g}(Y_j)\mathbf{1}_{F_l(Y_j)\le M_n}\right|\right] + u\right)\right),$$

where $\phi_{l\psi_g}^{M_n}(y) = \phi_{l\psi_g}(y)\mathbf{1}_{F_l(y)\le M_n}$. Hence, from Proposition A1 with $\sigma^2_{\mathcal{F}_l^M} = \sigma^2$, we get

$$P_{1l}(x;g) \le 2\left\{\exp\left(-\frac{C_2x^2}{n}\right) + \exp\left(-\frac{C_3x}{M_n}\right)\right\},$$

with $C_2 = A_2[4A_1^2\sigma^2]^{-1}$, and $C_3 = A_2[2A_1]^{-1}$.

Step 3: remainder term.
Define $\phi_{l\psi_g}^{M_n^c}(y) = \phi_{l\psi_g}(y)\mathbf{1}_{F_l(y)>M_n}$. We have

$$\left|\sum_{j=1}^{n}\phi_{l\psi_g}^{M_n^c}(Y_j)\right| \;\le\; \sum_{j=1}^{n}F_l(Y_j)\mathbf{1}_{F_l(Y_j)>M_n} =: S_{l,M_n}.$$

Hence, from Markov's inequality, $\mathbb{P}(S_{l,M_n} > x) \;\le\; \frac{n^k}{x^k}\mathbb{E}_{f_0}[F_l(Y)^k\mathbf{1}_{F_l(Y)>M_n}]$. Next, from Cauchy-Schwarz inequality,

$$\mathbb{E}_{f_0}[F_l(Y)^k\mathbf{1}_{F_l(Y)>M_n}] \;\le\; \mathbb{E}_{f_0}[F_l(Y)^{2k}]^{1/2}\mathbb{P}(F_l(Y) > M_n)^{1/2}.$$

Again, from Markov's inequality, $\quad\mathbb{P}(F_l(Y) > M_n) \;\le\; \frac{\mathbb{E}_{f_0}[F_l(Y)^{k'}]}{M_n^{k'}}$. This finally leads to

$$\mathbb{P}(S_{l,M_n} > x) \;\le\; \frac{n^k}{x^k\, M_n^{k'/2}}\,\mathbb{E}_{f_0}[F_l(Y)^{k'}]^{1/2}\mathbb{E}_{f_0}[F_l(Y)^{2k}]^{1/2}. \tag{A.2}$$

Take $M_n = n^{1/2-\varepsilon}$. Then $n^k M_n^{k'/2}$ is equal to 1 provided that $k' = 2k+\varepsilon$. We take $k' = m$, which corresponds to $k = m/2 - \varepsilon/2$. Next,

$$\begin{aligned}\left|\mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n^c}(Y)]\right| &\le \mathbb{E}_{f_0}\left[F_l(Y)^2\right]^{1/2}\mathbb{P}(F_l(Y) > M_n)^{1/2}\\ &\le \frac{\mathbb{E}_{f_0}[F_l(Y)^m]^{1/2}}{M_n^{m/2}}\,\mathbb{E}_{f_0}[F_l(Y)^2]^{1/2}.\end{aligned}$$

Therefore, since $(m-\varepsilon) \geq 2$,

$$\left| \sum_{j=1}^{n} \mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n^c}(Y)] \right| \leq \mathbb{E}_{f_0}[F_l(Y)^{4k}]^{1/2} \, \mathbb{E}_{f_0}[F_l(Y)^2]^{1/2} =: C_5.$$

Hence, for $x > C_5$,

$$\mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n^c}(Y)] \right| > x \right) = 0. \tag{A.3}$$

Let

$$P_{2l}(x;g) = \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \{\phi_{l\psi_g}^{M_n^c}(Y_j) - \mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n^c}(Y)]\} \right| > x \right).$$

It follows from (A.2) and (A.3) that

$$P_{2l}(x;g) \;\leq\; \mathbb{P}(S_{l,M_n} > x/2) + \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \mathbb{E}_{f_0}[\phi_{l\psi_g}^{M_n^c}(Y)] \right| > x/2 \right) \leq \frac{C_6}{x^{(m-\varepsilon)/2}},$$

for $x > C_5$.

Step 4: summary.
We have

$$P(x;g) \leq \sum_{l=1}^{2} P_{1l}(x/4;g) + P_{2l}(x/4;g).$$

From Step 2 and 3, we deduce that

$$P(x;g) \;\leq\; 4\left\{ \exp\left(-\frac{C_2 x^2}{16n} \right) + \exp\left(-\frac{C_3 x}{4M_n} \right) \right\} + \frac{C_7}{x^{(m-\varepsilon)/2}},$$

for $x > \max(C_5, 2A_1 C_g n^{1/2} \log M_n)$. The result follows from the fact that we can impose C_g large enough so that $C_5 \leq 2A_1 C_g n^{1/2} \log M_n$, and from the fact that we imposed $M_n = n^{1/2-\varepsilon/2}$ at Step 3. □

In the sequel, we present the concentration inequality that we use to derive our exponential bounds. This inequality is due to [34]. We use a formulation of this inequality similar to the one used in [11].

Proposition A1. *Let $\mathcal{F}$ be a pointwise measurable class of functions bounded by M. Let $(\varepsilon_j)_{1\leq j\leq n}$ denote an i.i.d. sequence of Rademacher variables independent from $(Y_j)_{1\leq j\leq n}$, that is $\mathbb{P}(\varepsilon_j = 1) = \mathbb{P}(\varepsilon_j = -1) = 1/2$. Then, we have for all u,*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}}\|\sum_{j=1}^{n} f(Y_j) - E[f(Y)]\| > A_1\left\{E\left[\sup_{f\in\mathcal{F}}\left\|\sum_{j=1}^{n} f(Y_j)\varepsilon_j\right\|\right] + u\right\}\right) \leq 2\left\{\exp\left(-\frac{A_2u^2}{n\sigma_{\mathcal{F}}^2}\right) + \exp\left(-\frac{A_2u}{M}\right)\right\},$$

with $\sigma_{\mathcal{F}}^2 = \sup_{f\in\mathcal{F}} Var(f(Y))$, and where A_1 and A_2 are universal constants.

Proposition A1 introduces the expectation of the supremum of a symmetrized sum that can make this inequality difficult to handle in full generality. [11] proposed a simple result to bound this expectation under generic conditions on the class of functions $\mathcal{F}$. Before stating their result, let us introduce the concept of covering numbers. For a probability measure $\mathbb{Q}$, define $\|\cdot\|_{2,\mathbb{Q}}$ as the L^2-norm associated to measure $\mathbb{Q}$. For a class $\mathcal{F}$ with envelope F (that is such that, for all $f \in \mathcal{F}$, $\|f(y)\| \leq F(y)$), define $\mathfrak{N}(\varepsilon, \|\cdot\|_{2,\mathbb{Q}})$ as the minimal number of balls (with respect to the $\|\cdot\|_{2,\mathbb{Q}}-$metric) of radius ε required to cover $\mathcal{F}$, and define

$$N_F(\varepsilon, \mathcal{F}) = \sup_{\mathbb{Q}:\mathbb{Q}(F^2)<\infty} \mathfrak{N}(\varepsilon\mathbb{Q}(F^2), \|\cdot\|_{2,\mathbb{Q}}).$$

The proposition below, due to [11] is valid up to some control on $N_F(\varepsilon, \mathcal{F})$ (which should not increase too fast when ε tends to zero) and some condition on the second order moments in the class $\mathcal{F}$.

Proposition A2. *Let $\mathcal{F}$ be a pointwise measurable class of functions bounded by M such that, for some constants $C, \nu \geq 1$, and $0 \leq \sigma \leq M$, we have*

(i) $N_M(\varepsilon, \mathcal{F}) \leq C\varepsilon^{-\nu}$, *for* $0 < \varepsilon < 1$,
(ii) $\sup_{f\in\mathcal{F}} E\left[f(Y)^2\right] \leq \sigma^2$,
(iii) $M \leq \frac{1}{4\nu}\sqrt{n\sigma^2/\log(C_1M/\sigma)}$, *with* $C_1 = \max(e, C^{1/\nu})$.

Then,

$$E\left[\sup_{f\in\mathcal{F}}\left\|\sum_{j=1}^{n} f(Y_j)\varepsilon_j\right\|\right] \leq A\sqrt{\nu n\sigma^2\log(C_1M/\sigma)}.$$

Appendix B. Covering numbers

Lemma B1. *Let* $\psi_g^b = (\pi_1^{(b)}, ..., \pi_{n_g}^{(b)}, \theta_1^{(b)}, ..., \theta_{n_g}^{(b)})$, $\psi_g^{(1)} = (\pi_1^{(1)}, ..., \pi_{n_g}^{(1)}, \theta_1^{(1)}, ..., \theta_{n_g}^{(1)})$, $\psi_g^{(2)} = (\pi_1^{(2)}, ..., \pi_{n_g}^{(2)}, \theta_1^{(2)}, ..., \theta_{n_g}^{(2)})$, *with* $\sum_{i=1}^{n_g} \pi_i^{(l)} = 1$ *for* $l = \{1, 2, b\}$. *Assume that, for all* $i \in \{1, ..., n_g\}$,

$$|\pi_i^{(1)} f_i(y; \theta_i^{(1)}) - \pi_i^{(2)} f_i(y; \theta_i^{(2)})| \quad \leq \quad \Lambda_1(y) \|\psi_g^{(1)} - \psi_g^{(2)}\|, \tag{B.1}$$

$$\left| \frac{\pi_i^{(1)} f_i(y; \theta_i^{(1)}) - \pi_i^{(b)} f_i(y; \theta_i^{(b)})}{\|\psi_g^{(1)} - \psi_g^b\|} - \frac{\pi_i^{(2)} f_i(y; \theta_i^{(2)}) - \pi_i^{(b)} f_i(y; \theta_i^{(b)})}{\|\psi_g^{(2)} - \psi_g^b\|} \right| \leq \Lambda_2(y) \|\psi_g^{(1)} - \psi_g^{(2)}\| \tag{B.2}$$

Moreover, assume that for all $\theta_i \in \Theta$, $0 < \Lambda_-(y) \leq f_i(y; \theta_i) \leq \Lambda_0(y) < \infty$, *with, for some function* $A(y) < \infty$,

$$\sup_{l=0,1,2} \left(\frac{\Lambda_j(y)}{\Lambda_-(y)} \right) \leq A(y). \tag{B.3}$$

Let $H(x) = x \ln(x)$. *Consider the classes of functions*

$$\mathcal{G}_i \quad = \quad \left\{ y \to g_{i,\psi_g}(y) = \frac{H(\tau_i(y)) - H(\tau_i^{(b)}(y))}{\|\psi_g - \psi_g^b\|} : \psi_g \in \Psi_g \right\},$$

Then, assuming that, for all $\psi_g \in \Psi_g$, $\pi_l \geq \pi_- > 0$ *for all* $l = 1, ..., n_g$,

$$\forall (\psi_g, \psi_g') \in \Psi_g, \ |g_{i,\psi_g}(y) - g_{i,\psi_g'}(y)| \leq \Lambda_3(y) \|\psi_g - \psi_g'\|, \tag{B.4}$$

for some function $\Lambda_3(y) \leq C A(y)^3$ *for some constant* $C > 0$.

Proof. Define, for $l = \{1, 2, b\}$

$$\begin{aligned} g^{(l)}(y) \quad &= \quad \pi_i^{(l)} f_i(y; \theta_i^{(l)}) + \sum_{k=1}^{n_g} \pi_k^{(l)} f_k(y; \theta_k^{(l)}) \, \mathbf{1}_{k \neq i}, \\ \tau_i^{(l)}(y) \quad &= \quad \frac{\pi_i^{(l)} f_i(y; \theta_i^{(l)})}{g^{(l)}(y)}. \end{aligned}$$

Write, for $l = \{2, b\}$,

$$\begin{aligned} \tau_i^{(1)}(y) - \tau_i^{(l)}(y) \quad &= \quad \frac{\pi_i^{(1)} f_i(y; \theta_i^{(1)}) - \pi_i^{(l)} f_i(y; \theta_i^{(l)})}{g^{(1)}(y)} \\ &\quad + \left\{ \frac{g^{(l)}(y) - g^{(1)}(y)}{g^{(1)}(y) g^{(l)}(y)} \right\} \pi_i^{(l)} f_i(y; \theta_i^{(l)}). \end{aligned} \tag{B.5}$$

Observe that $g^{(l)}(y) \geq \Lambda_-(y)$, so using equation (B.1), we get

$$|\tau_i^{(1)}(y) - \tau_i^{(2)}(y)| \leq \frac{\Lambda_1(y)\|\psi_g^{(1)} - \psi_g^{(2)}\|}{\Lambda_-(y)} + \frac{\Lambda_1(y)\Lambda_0(y)\|\psi_g^{(1)} - \psi_g^{(2)}\|}{\Lambda_-(y)^2}.$$

Due to assumption (B.3),

$$|\tau_i^{(1)}(y) - \tau_i^{(2)}(y)| \leq (A(y) + A(y)^2)\|\psi_g^{(1)} - \psi_g^{(2)}\|, \tag{B.6}$$

Next, observe that, again from (B.1),

$$\frac{|\tau_i^{(2)}(y) - \tau_i^{(b)}(y)|}{\|\psi_g^{(2)} - \psi_g^b\|} \leq \frac{\Lambda_1(y)}{\Lambda_-(y)} \leq A(y), \tag{B.7}$$

where we used again (B.3) and the fact that $\min(g^{(2)}(y), g^{(b)}(y)) \geq \Lambda_-(y)$. Using again (B.5), but this time for $l = \{b\}$, we get, according to (B.2),

$$\begin{aligned}\left|\frac{\tau_i^{(1)}(y) - \tau_i^{(b)}(y)}{\|\psi_g^{(1)} - \psi_g^b\|} - \frac{\tau_i^{(2)}(y) - \tau_i^{(b)}(y)}{\|\psi_g^{(2)} - \psi_g^b\|}\right| &\leq \frac{\Lambda_2(y)\|\psi_g^{(1)} - \psi_g^{(2)}\|}{\Lambda_-(y)} + \frac{\Lambda_2(y)\Lambda_0(y)\|\psi_g^{(1)} - \psi_g^{(2)}\|}{\Lambda_-(y)^2} \\ &\leq (A(y) + A(y)^2)\|\psi_g^{(1)} - \psi_g^{(2)}\|.\end{aligned}$$

Moreover, note that

$$|\ln(\tau_i^{(b)}(y))| \leq \frac{1}{\tau_i^{(b)}(y)}, \tag{B.8}$$

and that

$$\frac{\tau_i^{(1)}(y)}{\min(\tau_i^{(b)}(y), \tau_i^{(1)}(y))} \leq A(y), \tag{B.9}$$

from (B.3). Finally, again due to (B.3), note that, for $l = \{1, 2, b\}$,

$$\frac{1}{\tau_i^{(l)}(y)} \leq \frac{A(y)}{\pi_-}. \tag{B.10}$$

Let $H(x) = x\ln(x)$. Observe that $\sum_{i=1}^{n_g} H(\tau_i^{(l)}(y)) = Ent(\psi_g^{(l)}; y)$. Then, using that $|\ln(x/x')| \leq |x - x'|/\min(x, x')$, decompose

$$\begin{aligned}\left|\frac{[H(\tau_i^{(1)}(y)) - H(\tau_i^{(b)}(y))]}{\|\psi_g^{(1)} - \psi_g^b\|} - \frac{[H(\tau_i^{(2)}(y)) - H(\tau_i^{(b)}(y))]}{\|\psi_g^{(2)} - \psi_g^b\|}\right| &\leq \left|\frac{\tau_i^{(1)}(y) - \tau_i^{(b)}(y)}{\|\psi_g^{(1)} - \psi_g^b\|} - \frac{\tau_i^{(2)}(y) - \tau_i^{(b)}(y)}{\|\psi_g^{(2)} - \psi_g^b\|}\right| \\ &\quad \times \left(|\ln(\tau_i^{(b)}(y))| + \frac{\tau_i^{(1)}(y)}{\min(\tau_i^{(b)}(y), \tau_i^{(1)}(y))}\right) \\ &\quad + \frac{|\tau_i^{(2)}(y) - \tau_i^{(b)}(y)|\,|\tau_i^{(1)}(y) - \tau_i^{(2)}(y)|}{\min(\tau_i^{(b)}(y), \tau_i^{(1)}(y), \tau_i^{(2)}(y))\|\psi_g^{(2)} - \psi_g^b\|},\end{aligned}$$

Combining this with (B.6), (B.7), (B.8), (B.9) and (B.10) shows that

$$\left| \frac{[H(\tau_i^{(1)}(y)) - H(\tau_i^{(b)}(y))]}{\|\psi_g^{(1)} - \psi_g^b\|} - \frac{[H(\tau_i^{(2)}(y)) - H(\tau_i^{(b)}(y))]}{\|\psi_2 - \psi_g^b\|} \right| \leq \Lambda_3(y) \|\psi_g^{(1)} - \psi_g^{(2)}\|,$$

where

$$\Lambda_3(y) = (A(y)^2 + A(y)^3) \left(\frac{1}{\pi_-} + 1 \right) + \frac{A(y)^3}{\pi_-}.$$

□

Observe that, due to a Taylor expansion and under Assumption 1, the class $\mathcal{G}_i$ is bounded by the envelope

$$\mathfrak{g}(y) = \frac{\tilde{\Lambda}_0(y) + \tilde{\Lambda}_1(y)}{\tilde{\Lambda}_-(y)} + \frac{\tilde{\Lambda}_0(y)^2 + \tilde{\Lambda}_0(y)\,\tilde{\Lambda}_1(y)}{\tilde{\Lambda}_-(y)^2} \tag{B.11}$$

Hence, it is also bounded by $2(\tilde{A}(y) + \tilde{A}(y)^2)$.

Lemma B2. *Using the notations of Lemma B1, let $\mathcal{G}^g = \sum_{i=1}^{n_g} \mathcal{G}_i$. Then, under Assumption 1, $\mathcal{G}^g$ is a class of functions bounded by $G(y) = n_g\, \mathfrak{g}(y)$.*

$$N_{\mathfrak{g}}(\varepsilon, \mathcal{G}^g) \leq C n_g^{n_g V} \varepsilon^{-n_g V},$$

for some constants $C > 0$ and $V > 0$.

Proof. From Lemma 2.13 in [32], we get $N_{\mathfrak{g}}(\varepsilon, \mathcal{G}_i) \leq C\varepsilon^{-V}$, for some constants $C > 0$ and $V > 0$. The result then follows from Lemma 16 in [29]. □

Appendix C. Improvement of the bound of Theorem 1 under an exponential moment assumption

Assumption 3. *Using the notations of Assumption 1, assume that there exists $\rho > 0$ such that*

$$E[\exp(2\rho[\tilde{A}(y) + |\nabla_{\psi_g} \ln f(\psi_g^b; y)| + \mathfrak{g}(Y)]) < \infty,$$

where the $g_{i,\psi}(y)$ corresponds to the notations of Lemma B1.

Theorem C1. *Using the notations and assumptions of Theorem 1, but with Assumption 2 replaced by Assumption 3, we have*

$$P(x;g) \;\leq\; 4\left\{\exp\left(-\frac{A_3x^2}{n}\right)+\exp\left(-\frac{A_4x}{\ln n}\right)\right\}+A_7\exp(-\rho x/2),$$

for $x > A_6 n^{1/2}[\ln\ln n]^{1/2}$, *and some constant* $A_7 > 0$.

Proof. The proof is similar as the one of Theorem 1, but with Step 3 replaced by:

Step 3': remainder term using the exponential moments assumption.

Using the same notations as in Step 3 of Theorem 1, from Chernoff's inequality

$$\mathbb{P}(S_{l,M_n} > x) \leq \exp(-\rho x)(1+E[\exp(\rho F_l(Y))\mathbf{1}_{F_l(Y)>M_n}])^n.$$

Next, from Cauchy-Schwarz inequality,

$$E[\exp(\rho F_l(Y))\mathbf{1}_{F_l(Y)>M_n}] \leq E[\exp(2\rho F_l(Y))]^{1/2}\mathbb{P}(F_l(Y)>M_n)^{1/2}.$$

Again, from Chernoff's inequality,

$$\mathbb{P}(F_l(Y) > M_n) \leq E[\exp(2\rho F_l(Y))]\exp(-2\rho M_n).$$

This finally leads to

$$\begin{aligned}\mathbb{P}(S_{l,M_n} > x) &\leq e^{-\rho x}\left(1+E[\exp(2\rho F_l(Y))]\exp(-\rho M_n)\right)^n \\ &\leq \exp(-\rho x)\exp\left(ne^{-\rho M_n}E[\exp(2\rho F_l(Y))]\right).\end{aligned}$$

Taking $M_n = \rho^{-1}\ln n$ leads to

$$\mathbb{P}(S_{l,M_n} > x) \leq C_\rho\exp(-\rho x). \tag{C.1}$$

Next,

$$\begin{aligned}\left|E[\phi_{l\psi_g}^{M_n^c}(Y)]\right| &\leq E\left[F_l(Y)^2\right]^{1/2}\mathbb{P}(F_l(Y)>M_n)^{1/2} \\ &\leq E\left[F_l(Y)^2\right]^{1/2}E[\exp(2\rho F_l(Y))]^{1/2}\exp(-\rho M_n).\end{aligned}$$

Again, since $M_n = \rho^{-1} \ln n$, we get $n \exp(-\rho M_n) = 1$. Therefore,

$$\left| \sum_{j=1}^{n} E[\phi_{l\psi_g}^{M_n^c}(Y)] \right| \leq E\left[F_l(Y)^2\right]^{1/2} E[\exp(2\rho F_l(Y))]^{1/2} =: C_8.$$

Hence, for $x > C_8$,

$$\mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} E[\phi_{l\psi_g}^{M_n^c}(Y)] \right| > x \right) = 0. \tag{C.2}$$

Let

$$P_2^{(l)}(x; g) = \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} \{\phi_{l\psi_g}^{M_n^c}(Y_j) - E[\phi_{l\psi_g}^{M_n^c}(Y)]\} \right| > x \right).$$

It follows from (C.1) and (C.2) that

$$\begin{aligned} P_2^{(l)}(x; g) &\leq \mathbb{P}(S_{l,M_n} > x/2) + \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left| \sum_{j=1}^{n} E[\phi_{l\psi_g}^{M_n^c}(Y)] \right| > x/2 \right) \\ &\leq C_\rho \exp(-\rho x/2), \end{aligned}$$

for $x > C_8$.

Combining the different steps similarly to Step 4 in the proof of Theorem 1 leads to the result. □

References

[1] M. Aitkin. A general maximum likelihood analysis of variance components in generalized linear models. *Biometrics*, 55:117–128, 1999.

[2] J.-M. Azais, E. Gassiat, and C. Mercadier. Asymptotic distribution and power of the likelihood ratio test for mixtures: bounded and unbounded case. *Bernoulli*, 12(5):775–799, 2006.

[3] J.-M. Azais, E. Gassiat, and C. Mercadier. The likelihood ratio test for general mixture models with possibly structural parameters. *ESAIM P&S*, 13:301–327, 2009.

[4] J.P. Baudry. *Sélection de modèle pour la classification non supervisée. Choix du nombre de classes.* PhD thesis, Univ. Paris Sud XI, 2009.

[5] J.P. Baudry. Estimation and selection for model-based clustering with the conditional classification likelihood. *Preprint*, pages 1–35, 2012.

[6] C. Biernacki, G. Celeux, G. Govaert, and F. Langrognet. Model-based cluster and discriminant analysis with the mixmod software. *Computational Statistics and Data Analysis*, 51(2):587–600, 2006.

[7] Christophe Biernacki. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on PAMI*, 22:719–725, 2000.

[8] G. Celeux and G. Govaert. A classification em algorithm for clustering and two stochastic versions. *Computational Statistics and Data Analysis*, 14(3):315–332, 1992.

[9] G. Celeux and G. Soromenho. An entropy criterion for assessing the number of clusters in a mixture model. *Classification Journal*, 13(2):195–212, 1996.

[10] Uwe Einmahl and David M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Theoret. Probab.*, 13(1):1–37, 2000.

[11] Uwe Einmahl and David M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403, 2005.

[12] D.A. Follmann and D. Lambert. Identifiability for non parametric mixtures of logistic regressions. *Journal of Statistical Planning and Inference*, 27:375–381, 1991.

[13] C. Fraley and A.E. Raftery. How many clusters? which clustering method? answer via model-based cluster analysis. *The Computer Journal*, 41(8):578–588, 1998.

[14] Bernard Garel. Recent asymptotic results in testing for mixtures. *Computational Statistics and Data Analysis*, 51:5295–5304, 2007.

[15] E. Gassiat. Likelihood ratio inequalities with applications to various mixtures. *Ann. Inst. Henri Poincaré*, 38:897–906, 2002.

[16] E. Gassiat and R. Van Handen. Consistent order estimation and minimal penalties. *IEEE Trans. Info. th*, 59(2):1115–1128, 2013.

[17] Bettina Gruen and Friderich Leisch. Fitting finite mixtures of generalized linear regressions in r. *Computational Statistics and Data Analysis*, 51(11):5247–5252, 2007.

[18] Bettina Gruen and Friderich Leisch. *Finite mixture of generalized linear regression models*, pages 205–230. Recent advances in linear models and related areas. Springer, 2008.

[19] L.A. Hannah, D.M. Blei, and W.B. Powell. Dirichlet process mixtures of generalized linear models. *Journal of Machine Learning Research*, 1:1–33, 2011.

[20] R. Hathaway. A constrained em algorithm for univariate normal mixtures. *Journal of Statistical Computation and Simulation*, 23(3):211–230, 1986.

[21] C. Hennig and T.F. Liao. How to find an appropriate clustering for mixed-type variables with application to socio-economic stratification. *Journal of the Royal Statistical Society: Series C*, 62(3):309–369, 2013.

[22] C. Keribin. *Tests de modèles par maximum de vraisemblance*. PhD thesis, Université d'Evry, 1999.

[23] P.J. Lenk and W.S. DeSarbo. Bayesian inference for finite mixtures of generalized linear models with random effects. *Psychometrika*, 65(1):93–119, 2000.

[24] W. Luo. Penalized minimum matching distance-guided em algorithm. *International Journal of Electronics and Communications*, 60:235–239, 2006.

[25] P. McCullagh and J. A. Nelder. *Generalized linear models, 2nd ed.* Monographs on Statistics and Applied Probability. Chapman and Hall, London, 1989.

[26] G. McLachlan and D. Peel. *Finite Mixture Models*. Wiley Series In Probability and Statistics. Wiley, New York, 2000.

[27] X. Milhaud. Exogenous and endogenous risk factors management to predict surrender behaviours. *ASTIN Bulletin*, 43(3):373–398, 2013.

[28] R Nishii. Maximum likelihood principle and model selection when the true model is unspecified. *Journal of Multivariate Analysis*, 27(2):392–403, 1988.

[29] Deborah Nolan and David Pollard. U-processes: rates of convergence. *Ann. Statist.*, 15(2):780–799, 1987.

[30] E. Ohlson and B. Johansson. *Non-Life Insurance Pricing with Generalized Linear Models*. Springer, 2010.

[31] A. Oliviera-Brochado and F. Vitorino Martins. Assessing the number of components in mixture models: a review. Working Paper, November 2005.

[32] Ariél Pakes and David Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057, 1989.

[33] A.E. Raftery. Bayesian model selection in social research (with discussion). Technical Report 94-12, Demography Center Working, University of Washington, 1994.

[34] M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994.

[35] H.X. Wang, B. Luo, Q.B. Zhang, and S. Wei. Estimation for the number of components in a mixture model using stepwise split-and-merge em algorithm. *Pattern Recognition Letters*, 25:1799–1809, 2004.

[36] P. Wang. *Mixed Regression Models for Discrete Data*. PhD thesis, University of British Columbia, Vancouver, 1994.

[37] P. Wang, M.L. Puterman, I. Cockburn, and N.D. Le. Mixed poisson regression models with covariate dependent rates. *Biometrics*, 52:381–400, 1996.

[38] M. Wedel and W.S. DeSarbo. A mixture likelihood approach for generalized linear models. *Journal of Classification*, 12:21–55, 1995.