Selection of GLM mixtures: a new criterion for clustering purpose

Olivier Lopez, Milhaud Xavier

Selection of GLM mixtures: a new criterion for clustering purpose

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 is a challenging issue which has undergone many recent developments (Hennig and Liao (2013), Hannah, Blei, and Powell (2011), Aitkin (1999)). In practice, the model selection step is usually performed by using AIC or BIC penalized criteria. Though, simulations show that they tend to overestimate the actual dimension of the model. These evidence led us to consider a new criterion close to ICL, firstly introduced in Baudry (2009). Its definition requires to introduce a contrast embedding an entropic term: using concentration inequalities, we derive key properties about the convergence of the associated M-estimator. The consistency of the corresponding classification criterion then follows depending on some classical requirements on the penalty term. Finally a simulation study enables to corroborate our theoretical results, and shows the effectiveness of the method in a clustering perspective.

Keywords: Conditional classification likelihood, GLM, Model selection.

1. Introduction

The use of mixture modeling has boomed since the publication of Dempster, N.M., and D.B. (1977), who provided with how to estimate the pa-

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

rameters of a finite mixture model thanks to the EM algorithm. Applications involving finite mixtures are usually focused on dividing explicitly the population structure into subpopulations, given that we originally do not know to which subpopulation each individual belongs. This is what is commonly called an incomplete data problem, where the crucial point lies in determining the right number of subpopulations (or components) in order to perform a model-based clustering at the end. Meanwhile, generalized linear models (GLM) have undergone vigorous development since the early 1980's. Their great popularity may be due to their high flexibility and ability to consider both categorical and continuous risk factors (see McCullagh and Nelder (1989) and references therein). In many situations, finite mixtures of GLM can be very useful to deal with a large heterogeneity concerning the impact of those risk factors on some phenomenon in the population under study. This could be interpreted as complex interactions between a response and some covariates.

We focus in this paper on the topic of selecting the "best" GLM mixture; where "best" should be understood as the best trade-off between the fit and the clustering confidence (this notion will be more detailed further). In this view, we study a criterion which was originally proposed by Baudry (2009). We derive general exponential bounds for this type of criterion, and show how these results can be applied to the selection of the order (number of components) of GLM mixtures. Selecting one model within a collection is an important statistical problem that has undergone vigorous developments in the literature. In the mixture framework, a universal solution has failed to emerge to answer the question of selecting the right order. Many articles have been dedicated to the implementation of algorithmic mixture calibration techniques. Nevertheless, they often suffer from a lack of theoretical justification with respect to their convergence properties.

Based on the information theory, Oliviera-Brochado and Vitorino Martins (2005) point that there are basically two main approaches to infer the order of a mixture: hypothesis tests, and information and classification criteria. Garel (2007) highlights the difficulty of establishing multimodality by means of the generalized likelihood ratio test in the first approach, because the classical result according to which the test statistic is χ^2-distributed is generally not applicable where mixtures are concerned. Azais, Gassiat, and Mercadier (2006) and Azais, Gassiat, and Mercadier (2009), building on Gassiat (2002), offer a detailed solution to overcome this issue. When using information criteria, most authors (McLachlan and Peel (2000), Fraley and Raftery (1998))

agree that the BIC criterion gives better results than AIC since it seeks to minimize the Kullback-Leibler (KL) divergence to the true distribution (Raftery (1994), Ripley (1995)). For example, the convergence of BIC to estimate the order of gaussian mixtures has been proved in Keribin (1999). More generally, Gassiat and Van Handen (2013) unrolls the convergence properties (towards the theoretical model) of a model selected by a likelihood-penalized criterion (where the penalty linearly depends on the model dimension) in the context of mixtures. In practice, it is a well known fact that these two criteria tend to overestimate the theoretical number of components, especially when the model is misspecified (Baudry (2009)). This statement is not really surprising: AIC, BIC 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. Celeux and Soromenho (1996) and Biernacki (2000) were the firsts to introduce a *classification* criterion to avoid this overestimation, namely the ICL criterion. However, ICL consistency has not been proved in the context of maximum likelihood theory. Baudry (2009) recently demonstrated the consistency of a slightly modified version of ICL under the gaussian mixture framework; but no such property has been established in the context of GLM mixtures.

The aim of this paper is two-fold: obtaining new theoretical results concerning the classification criterion introduced by Baudry (2009), and developing its application in the context of GLM mixtures. In section 2, we consider a general mixture framework and define the conditional classification likelihood. By maximizing this contrast, we obtain an estimator of the mixture parameters which differs from the maximum likelihood estimator. Indeed, its purpose is to find a compromise between a small classification error and a good fit to data. We obtain a general bound for the estimation error based on concentration inequalities. In section 3, general penalized criteria are considered to select the order of the mixture, and we determine conditions for the consistency of such procedures. The application of these results to GLM mixtures follows in section 4. The practical behavior of this approach is then investigated through simulation studies in section 5.

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 mixands. If ν has finite support, f_ν is a finite mixture density. In the present paper, our interest lies in discrete mixtures: for any $y \in \mathcal{Y}$, the density f_ν is assumed to belong to a collection of densities M_g defined as

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

where $\Psi_g = \big(\Pi_{n_g} \times \Theta^{n_g}\big)$, with $\Pi_{n_g} \subset \big\{(\pi_1, ..., \pi_{n_g}) :\ \sum_{i=1}^{n_g} \pi_i = 1 \text{ and } \pi_i \geq 0\big\}$ and $\Theta^{n_g} = (\theta_1, ..., \theta_{n_g})$.
We will denote K_g the dimension of the parameter set Ψ_g. Based on i.i.d. observations $\mathbf{Y} = (Y_1, \ldots, Y_n)$, the corresponding likelihood is given by

$$\forall \psi_g \in \Psi_g, \quad L(\psi_g; Y_1, ..., Y_n) = L(\psi_g) = \prod_{j=1}^{n} \sum_{i=1}^{n_g} \pi_i\, f_i(Y_j;\theta_i). \quad (2)$$

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) $\hat{\psi}_g^{MLE}$ is defined as the maximizer of $L(\psi_g)$ over Ψ_g (in full generality, it may not be unique). Under some regularity conditions, $\hat{\psi}_g^{MLE}$ converges towards ψ_g^{MLE}, which is the true parameter when the model is correctly specified.

2.2. A new contrast: the conditional classification likelihood

To figure out in what the optimization by conditional classification likelihood consists, we first have to define the conditional classification likelihood itself. This function is derived from the general principle of the EM algorithm and approximates the j$^{\text{th}}$ individual likelihood of the complete data (Y_j, δ_j),

where $\delta_j = (\delta_{ij})_{i\in[\![1,n_g]\!]}$ is the latent component indicator (more precisely δ_{ij} equals 1 if observation j belongs to component i, 0 otherwise).
Several authors have attempted to exploit the link between the likelihood of the observed data and the likelihood of the complete data (Celeux and Govaert (1992)), originally noted by Hathaway (1986). A specific term appears while writing the likelihood relatively to the complete data $(\mathbf{Y}, \delta)$, hereafter named the classification likelihood: $\forall \psi_g \in \Psi_g$,

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

$\tau_i(Y_j;\psi_g)$ is the *a posteriori* probability that observation j belongs to component i. The term that binds the two likelihoods is very close to what is commonly called the entropy:

$$\forall \psi_g \in \Psi_g,\ \forall y_j \in \mathbb{R}^d, \quad Ent(\psi_g; y_j) = -\sum_{i=1}^{n_g} \tau_i(y_j;\psi_g)\, \ln\Big(\tau_i(y_j;\psi_g)\Big).$$

This function results from the expectation (w.r.t.δ) taken in the first member of the right-hand term in (3), hence the "conditional classification likelihood" denoted further by L_{cc}:

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

where $Ent(\psi_g;\mathbf{Y}) = \sum_{j=1}^{n} Ent(\psi_g;Y_j)$.
The entropy is maximum in case of 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 highlighted by equation (4), this term can be seen as a penalization of the observed likelihood: the bigger the lack of confidence when making the a posteriori

classification (via the *Bayes rule*), the greater the penalization (and vice versa). In fact, the entropy has a zero limit when τ_i tends to 0 or 1. 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 will be a key point in the definition of the parameters space that is acceptable to ensure the convergence of the estimator based on the conditional classification likelihood. We must therefore avoid the *a posteriori* proportions of the mixture tending to zero. Also essential is to keep in mind that the mixture model should be identifiable, see McLachlan and Peel (2000) (p.26) for a further discussion on this issue. More generally, the L_{cc} expression enables to deduce additional constraints to be imposed on the parameters space (so that L_{cc} does not diverge) by studying its limits. Most of time, this suggests that critical situations correspond mainly to parameters that would not be bounded (Baudry (2009)). Define the maximum conditional classification likelihood estimator ($ML_{cc}E$)

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

It should converge towards $\psi_g^{ML_{cc}E} = \arg\max_{\psi_g \in \Psi_g} \mathbb{E}_{f^0}[\ln L_{cc}(\psi_g, Y)]$. Baudry (2009) provides us with the following example so as to catch in what $\psi_g^{ML_{cc}E}$ differs from ψ_g^{MLE}. Recall that the latter aims at minimizing the KL divergence between $f(.;\, \psi_g)$ and the theoretical distribution $f^0(.)$.

Example. *f^0 is the normal density $\mathcal{N}(0,1)$. Consider the model*

$$M = \left\{ \frac{1}{2} f_{\mathcal{N}}(.;\, -\mu, \sigma^2) + \frac{1}{2} f_{\mathcal{N}}(.;\, \mu, \sigma^2)\,; \quad \mu \in \mathbb{R},\ \sigma^2 \in \mathbb{R}^{+*} \right\},$$

where no further condition is imposed. There is no closed-form expression for $\psi_g^{ML_{cc}E}$ in this example (even when σ^2 is fixed!). However one can compute it numerically: $(\mu^{ML_{cc}E}, \sigma^{ML_{cc}E}) = (0.83, \sqrt{0.31})$. This means that there exists a unique maximizer of $\mathbb{E}_{f^0}[\ln L_{cc}(\mu, \sigma^2)]$ in Ψ_g (up to a label switch), which is obviously different from ψ_g^{MLE}. Indeed, $\psi_g^{MLE} = (\mu^{MLE}, \sigma^{MLE}) = (0, 1)$, which leads to nothing else than the theoretical distribution itself.

It shows that the $ML_{cc}E$ does not aim at recovering the theoretical distribution, even when contained in the model under consideration. Here, MLE has no rule to designate two suitable classes (components) for this model: it would therefore construct the same two exactly superimposed $f_{\mathcal{N}}(.;\, 0, 1)$.

The allocation of observations to one or any of these components would then be completely arbitrary (with probability 0.5, hence maximum entropy). In contrast, the compromise sought by the $ML_{cc}E$, which penalizes such excessive entropy, leads to find another estimator resulting in greater confidence in the assignment of observations to mixture components.

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

To shorten the notation, let $\psi_g^b = \psi_g^{ML_{cc}E}$ and $\hat{\psi}_g = \hat{\psi}_g^{ML_{cc}E}$. Define $\phi(\psi_g; y) = \ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)$, and $d(\psi_g, \psi_g^b) = -\mathbb{E}\left[\phi(\psi_g; Y)\right]$. If the model is correctly specified, that is if $f^0(\cdot) = f(\cdot; \psi_g^b)$, the function d is the Kullback-Leibler divergence between $f(\cdot; \psi_g)$ and $f(\cdot; \psi_g^b)$. In full generality, this quantity will be different from the Kullback-Leibler, but will express some pseudo-distance between the parameters ψ_g and ψ_g^b. Let us observe that, by definition of ψ_g^b, $d(\psi_g, \psi_g^b) \geq 0$ 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 L_{cc} contrast, centered by its expectation in the case where $\ln L_{cc}$ is bounded. 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 $d(\hat{\psi}_g, \psi_g^b)$. Let us note that, in view of applying our result to GLM inference, we require to have a result which is adapted to unbounded contrasts $\ln L_{cc}$ (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 of 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. $\nabla_\theta^2 f_i(y; \theta_i)$) the vector (resp. the matrix) of partial derivatives of f_i with respect to each component of θ_i. Assume that, for all $i = 1, ..., n_g$ and all $\theta \in \Theta$,*

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

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

In the case where the functions f_i are not bounded, we require some moment assumptions on both $\tilde{A}(y)$ defined in Assumption 1 and some functions related to the contrast evaluated at the true parameter.

Assumption 2. *Using the notations of Assumption 1, assume that there exists $m > 0$ such that*

$$\mathbb{E}[\tilde{A}(Y)^m] + \mathbb{E}[|\nabla_{\psi_g} \ln f(Y; \psi_g^b)|^m] + \mathbb{E}[\sup_{i=1,...,n_g} |g_{i,\psi_g^b}(Y)|^m] < \infty,$$

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

We now state the main result of this section.

Theorem 1. *Let*

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

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 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

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

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

Proof. Let $\phi_{\psi_g}(y) = \frac{\{\ln L_{cc}(\psi_g; y) - \ln L_{cc}(\psi_g^b; y)\}}{\|\psi_g - \psi_g^b\|}$.

We can decompose $\phi_{\psi_g}(y) = \phi_{1\psi_g}(y) - \phi_{2\psi_g}(y)$, where

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

The proof consists of applying the concentration inequality of Proposition A1, along with Proposition A2 to the classes 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 of the proof, all the results are gathered.

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

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

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 Pakes and Pollard (1989) 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 Nolan and Pollard (1987) 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. 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 Einmahl and Mason (2000)). Let

$$P_1^{(l)}(x;g) = \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \sum_{j=1}^{n} \{\phi_{l\psi_g}(Y_j) - \mathbb{E}[\phi_{l\psi_g}(Y)]\} \mathbf{1}_{F_l(Y_j)\leq M_n} > x \right).$$

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

$$\mathbb{E}\left[\sup_{\psi_g\in\Psi_g}\left|\sum_{j=1}^{n}\varepsilon_j\phi_{l\psi_g}(Y_j)\mathbf{1}_{F_l(Y_j)\leq M_n}\right|\right] \leq C_g\, n^{1/2}[\log(M_n)]^{1/2}, \tag{6}$$

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 $x > 2A_1C_g n^{1/2}[\log M_n]^{1/2}$, that the probability $P_1^{(l)}(x;g)$ is bounded by

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

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

$$P_1^{(l)}(x;g) \leq 2\left\{\exp\left(-\frac{C_2x^2}{n}\right) + \exp\left(-\frac{C_3x}{M_n}\right)\right\},$$

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

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

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

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

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

$$\mathbb{P}(S_{l,M_n} > x) \;\leq\; \frac{n^k}{x^k\, M_n^{k'/2}}\,\mathbb{E}[F_l(Y)^{k'}]^{1/2}\mathbb{E}[F_l(Y)^{2k}]^{1/2}. \tag{7}$$

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,

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

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

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

Hence, for $x > C_5$,

$$\mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left|\sum_{j=1}^n \mathbb{E}[\phi_{l\psi_g}^{M_n^c}(Y)]\right| > x\right) = 0. \tag{8}$$

Let

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

It follows from (7) and (8) that

$$P_2^{(l)}(x;g) \leq \mathbb{P}(S_{l,M_n} > x/2) + \mathbb{P}\left(\sup_{\psi_g \in \Psi_g} \left|\sum_{j=1}^n \mathbb{E}[\phi_{l\psi_g}^{M_n^c}(Y)]\right| > x/2\right) \leq \frac{C_6}{x^{(m-\varepsilon)/2}},$$

for $x > C_5$.

Step 4: summary.
We have

$$P(x;g) \leq \sum_{l=1}^2 P_1^{(l)}(x/4;g) + P_2^{(l)}(x/4;g).$$

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. □

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 Gassiat (2002) and Gassiat and Van Handen (2013) 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 in Assumption 2, a better bound can be obtained. This will especially be the case when one considers bounded variables Y, which lead to a bounded function $\tilde{A}(y)$. In appendix Appendix C, we show how this bound can be obtained under this more restrictive assumption.

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. This problem does not appear when one only considers the log-likelihood.

2.4. Almost sure 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 $\mathbb{E}[\ln L_{cc}(\psi_g;Y)]$ 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 $E_n(u) = \mathbb{P}(n^{1/2}\|\hat{\psi}_g - \psi_g^b\|_2 > u[\ln n]^{1/2})$. 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 $u > A_6$,

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

We see that the first sum in the right-hand side is finite provided that $u > A_3^{-1/2}$. 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 $m \geq 2$,

$$p_n(u) \leq 4\exp(-A_3 u^2) + 4\exp(-A_4[\ln n]^{1/2} u) + 2^{4m} A_5 \mathfrak{c}^{-m/2} u^{-m/2},$$

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

Remark 3: 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

$$\frac{1}{n}\sum_{j=1}^{n} \{\ln L_{cc}(\hat{\psi}_g; Y_j) - \ln L_{cc}(\psi_g^b; Y_j)\} + d(\hat{\psi}_g, \psi_g^b) = O_{a.s.}\left([\ln n] n^{-1}\right). \quad (9)$$

3. A new penalized selection criterion: ICL^*

As mentioned before, a crucial issue in clustering and mixture analysis is to determine the appropriate order of the mixture to correctly describe the data. Biernacki (2000) 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 likelihood, and eliminate the problem of overestimating the order in the mixture. 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, called ICL, is henceforth well suited to issues of population clustering. But a particular attention should be paid to the definition of the penalty: early works used to consider entropy as part of the penalty. Unfortunately no theoretical result could be demonstrated from this viewpoint, despite promising results in practical applications (Biernacki et al. (2006)). Baudry (2009) then proposed to redefine the ICL criterion by combining it to the L_{cc} contrast. The penalty thus becomes identical to that of BIC, and the estimator ($ML_{cc}E$) used to express the "new" ICL criterion differs from the maximum likelihood estimator. In this regard, we have in the previous section shown the strong convergence of this estimator 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 $M_g,\ g = \{1, ..., G\}$.

3.1. Previous works on ICL criteria

The ICL criterion was defined on the same basis as the BIC criterion: Biernacki (2000) 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 $\hat{\psi}_g^{MLE}$ when n gets large, which is clearly questionable since the contrast is different from the classical likelihood. Besides, the label vector δ is not observed, so that the Bayes rule is used on a posteriori probabilities to assign observations to each mixture component: the predicted label is denoted $\hat{\delta}^B$ and also depends on

the MLE. This leads to consider the following procedure:

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

McLachlan and Peel (2000) suggest to use the *a posteriori* probabilities $\tau_i(y; \hat{\psi}_g^{MLE})$ instead of $\hat{\delta}^B$:

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

In fact, ICL_a and ICL_b are really different in practice only if $\forall i,\ \tau_i(Y_j; \hat{\psi}_g^{MLE}) \simeq 1/n_g$. 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. Biernacki (2000) and McLachlan and Peel (2000) have shown, through various simulated and real-life examples, that the ICL criterion is more robust than the BIC criterion 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. However, there is no clear relationship between the maximum likelihood theory and the entropy. In addition, 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 (Nishii (1988)). 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 $Ent(\psi_g; \mathbf{Y}) = O(n)$. Until very recently, there was therefore clearly a gap between the practical interest aroused by ICL and its theoretical justification. This was partly plugged by Baudry (2009) who introduced a

new version of ICL integrating a "BIC-type penalty":

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

In the context of gaussian mixtures, Baudry (2009) has rigorously shown that the number of components selected using this criterion converges weakly towards the theoretical one, but only in the bounded case.

3.2. Consistency of selection criteria

Still in the mixture modelling framework, let M_{g^*} denote the model with smallest dimension K_{g^*} such that $\mathbb{E}[\ln L_{cc}(\psi_{g^*}^b)] = \max_{g=1,\ldots,G} \mathbb{E}[\ln L_{cc}(\psi_g^b)]$. The following theorem provides consistency properties of a class of penalized estimators. Related results can be found in Baudry (2009).

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

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

Then, if $m > 2$ in Assumption 2 and if $nu_n \to \infty$, we get $\forall g \neq g^*$

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

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

$$\varepsilon_g = \mathbb{E}\left[\ln L_{cc}(\psi_{g^*}^b; Y)\right] - \mathbb{E}\left[\ln L_{cc}(\psi_g^b; Y)\right].$$

Decompose

$$\begin{aligned}
\frac{1}{n} \sum_{j=1}^{n} \ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\hat{\psi}_g; Y_j) =& \varepsilon_g + \frac{1}{n} \sum_{j=1}^{n} \Big\{ \ln L_{cc}(\psi_{g^*}^b; Y_j) - \mathbb{E}[\ln L_{cc}(\psi_{g^*}^b; Y)] \Big\} \\
&- \frac{1}{n} \sum_{j=1}^{n} \Big\{ \ln L_{cc}(\psi_g^b; Y_j) - \mathbb{E}[\ln L_{cc}(\psi_g^b; Y)] \Big\} \\
&+ \frac{1}{n} \sum_{j=1}^{n} \Big\{ \ln L_{cc}(\hat{\psi}_{g^*}; Y_j) - \ln L_{cc}(\psi_{g^*}^b; Y_j) + d(\hat{\psi}_{g^*}, \psi_{g^*}^b) \Big\} \\
&- \frac{1}{n} \sum_{j=1}^{n} \Big\{ \ln L_{cc}(\hat{\psi}_g; Y_j) - \ln L_{cc}(\psi_g^b; Y_j) + d(\hat{\psi}_g, \psi_g^b) \Big\}.
\end{aligned} \tag{10}$$

It follows from the remark following Corollary 1 that the last two terms in (10) are $O_{a.s.}([\ln n]n^{-1})$ (or $O_P(n^{-1})$). We now distinguish two cases: $\epsilon_g > 0$ and $\epsilon_g = 0$.

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

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

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

$$\frac{1}{n}\sum_{j=1}^{n}\ln L_{cc}(\hat{\psi}_{g^*};Y_j) - \ln L_{cc}(\hat{\psi}_{g};Y_j) - pen(g^*) + pen(g) < 0. \tag{11}$$

However, due to the previous remarks, if we take $u_n = o(1)$, the left-hand side in (11) 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, $\mathbb{P}(\hat{g} = g) = o(1)$).

Step 2: $\epsilon_g = 0$**.**
Since $\psi^b_g = \psi^b_{g^*}$,

$$\frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g^*};Y_j) - \mathbb{E}[\ln L_{cc}(\psi^b_{g^*};Y)]\right\} = \frac{1}{n}\sum_{j=1}^{n}\left\{\ln L_{cc}(\psi^b_{g};Y_j) - \mathbb{E}[\ln L_{cc}(\psi^b_{g};Y)]\right\},$$

and $\varepsilon_g = 0$, which shows that the first three terms in (10) are zero. This leads to

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

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. □

4. Application to the selection of GLM mixture models

4.1. Description of the GLM framework

GLM are a common way to integrate specific risk factors; and notably 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, the topic of GLM has undergone vigorous development in the 1980's and these models are nowadays used in many fields among which marketing, economics, medicine, astronomy. For example, it has become a standard tool in insurance pricing (Ohlson and Johansson (2010)). 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 the exponential family. Among many others, Young and Hunter (2010), Gruen and Leisch (2007) and Leisch (2008) are currently interested in GLM mixtures. However, no theoretical development exists about selection criteria that satisfy classifying objectives in the context of GLM mixtures: this section thus aims at giving the suitable convergence conditions using the ICL^* criterion.

To be in line with the previous notations, consider i.i.d. replications $(Y_j)_{1\leq j\leq n}$ with $Y_j = (Z_j, X_j)$. Z_j is the **random response** for the j$^{\text{th}}$ individual, and $X_j^T = (1, X_{j1}, ..., X_{jp})$ its vector of **covariates** (we use superscript T to denote the matrix transpose). Introduce an invertible **link function** l such that $l(\mathbb{E}[Z_j]) = X_j^T\beta$, with $\beta^T = (\beta_0, ..., \beta_p)$.
Moreover we assume that the conditional distribution of Z_j given X_j belongs to an exponential family, that is

$$f_{Z|X}(z_j; \alpha, \phi) = \exp\left(\frac{z_j\alpha - b(\alpha)}{a(\phi)} + c(z_j, \phi)\right), \tag{12}$$

where $a(.)$, $b(.)$ and $c(.)$ are specific functions depending on the model under study, and α and ϕ are the parameters to be estimated. This is similar to the vector θ in section 2 if $\theta = (\alpha_1 a(\phi)^{-1}, ..., \alpha_d a(\phi)^{-1}, \phi) = (\tilde{\theta}^T, \phi)$. Hence, it is easy to check that Assumptions 1 and 2 hold with $\tilde{A}(y) = z^2 \exp(z \sup_{\theta\in\Theta} |\tilde{\theta}^T x|)$. As a matter of fact, the choice of the response distribution determines how we explicit the relation between α, ϕ and the parameters of the outcome distribution itself. The interested reader can learn more about GLM in the seminal book by McCullagh and Nelder (1989).
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 McLachlan and Peel (2000) (p.146) and Wang (1994), and special cases about the Poisson regression model as well as the binomial regression model are available in Wang et al. (1996) and Follmann and Lambert (1991) respectively.

4.2. L_{cc} likelihoods for two members of the GLM family

We focus here on discrete support mixtures with components all belonging to the same GLM family, which is actually the type of model resorted to in practice. One thus considers the set M_g of density functions given by $\left\{ f(.;\psi_g) = \sum_{i=1}^{n_g} \pi_i \, f_i(.;\alpha_i,\phi_i) \mid \psi_g = (\pi_1,...,\pi_{n_g},\alpha_1,...,\alpha_{n_g},\phi_1,...,\phi_{n_g}) \in \Psi_g \right\}$, where $f_i(.;\alpha_i,\phi_i)$ follows (12).
Our goal is to study deeply the characteristics of the L_{cc} contrast for most famous distributions of the GLM family. We would like to formulate the constraints to impose on the parameters space for each density function, as well as on auxiliary functions ($a(), b()$ and $c()$). To simplify, our results are expressed for a one-dimensional outcome Y but remain valid when Y is k-dimensional ($k > 1$). For the sake of conciseness, we focus more intensively on the two families that will be considered in the simulation study.

Mixture of linear regression models. The gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ belongs to the exponential family, and is thus a potential choice to model the error in a GLM mixture. The density of this random variable can be written in the exponential form (12) by operating the following transformations: $\alpha = \mu$ (hence $\alpha \in \mathbb{R}$), $b(\alpha) = \mu^2/2$ (so that $b(\alpha) \in \mathbb{R}^+$), $\phi = a(\phi) = \sigma^2$ (hence $\phi \in \mathbb{R}^{+*}$), and $c(y;\phi) = -1/2\,(y^2/\sigma^2 + \ln 2\pi\sigma^2)$ (so that $c(y;\phi) \in \mathbb{R}$).

Considering an *identity* link and a *gaussian* error in the GLM mixture model, we fall back on gaussian mixtures to which shall be added some dependence in function of observed covariates. Following (4), the conditional classification likelihood for a single observation y_j reads

$$\ln L_{cc}(\psi_g; y_j) = \ln \left(\sum_{i=1}^{n_g} \frac{\pi_i}{\sqrt{2\pi\sigma_i^2}} \, e^{-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}} \right) + \sum_{i=1}^{n_g} \tau_i(y_j;\psi_g) \, \ln \tau_i(y_j;\psi_g),$$

$$\text{where } \tau_i(y_j;\psi_g) = \frac{\pi_i}{\sqrt{2\pi\sigma_i^2}} \, e^{-\frac{1}{2}\frac{(y_j - X_j\beta_i)^2}{\sigma_i^2}} \left(\sum_{k=1}^{n_g} \frac{\pi_k}{\sqrt{2\pi\sigma_k^2}} \, e^{-\frac{1}{2}\frac{(y_j - X_j\beta_k)^2}{\sigma_k^2}} \right)^{-1}.$$

Clearly, the same constraints as those on gaussian mixtures should be imposed: constraints on μ_i and σ_i^2 are to be passed onto β_i and σ_i^2, and therefore also onto α_i and ϕ_i. One should work in a properly selected compact space to ensure the bounded nature of the L_{cc} log-likelihood as well as its derivative. To summarize, the study of these limits shows that:

i) σ_i^2 has to be upper-bounded, so that ϕ_i should be bounded;
ii) σ_i^2 should not tend to 0, hence the same constraint on ϕ_i;
iii) regression coefficients must remain bounded ($\forall i \in [\![1, n_g]\!], |\beta_i| \neq \infty$). Given that $\alpha_i = \mu_i = X\beta_i$, we deduce that α_i should also be bounded.

Mixture of Poisson regression models. When dealing with counting data, another option for modeling the error might be a Poisson law $\mathcal{P}(\mu)$. Table 1 provides the correspondence between μ and the parameters of the exponential family. The individual likelihood in such a mixture follows (after some computations): $\forall \psi_g \in \Psi_g$,

$$L(\psi_g; y_j) = \sum_{i=1}^{n_g} \pi_i \; e^{-e^{X_j\beta_i}} \; \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!}.$$

From (4), the conditional classification likelihood is thus given by

$$\ln L_{cc}(\psi_g; y_j) = \ln \left(\sum_{i=1}^{n_g} \pi_i e^{-e^{X_j\beta_i}} \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!} \right) + \sum_{i=1}^{n_g} \tau_i(y_j; \psi_g) \; \ln \tau_i(y_j; \psi_g),$$

$$\text{where } \tau_i(y_j; \psi_g) = \pi_i \, e^{-e^{X_j\beta_i}} \frac{\left[e^{X_j\beta_i}\right]^{y_j}}{y_j!} \left(\sum_{k=1}^{n_g} \pi_k \, e^{-e^{X_j\beta_k}} \frac{\left[e^{X_j\beta_k}\right]^{y_j}}{y_j!} \right)^{-1}.$$

Let us make the parameter μ_i tend towards the bounds of its domain and study the limits of the L_{cc} likelihood. We obtain, after some tedious computations, the following constraint: β_i coefficients must remain bounded ($\forall i \in [\![1, n_g]\!], |\beta_i| \neq \infty$), which means the same constraint for parameters α_i.

Summary of constraints on the parameter space. Table 1 compiles the overall results for other classical distribution of the GLM family. It permits to recap the constrained support of both α and ϕ parameters, which guarantees the convergence results of the $ML_{cc}E$ estimator and the ICL^* criterion. Indeed, it has become obvious that the GLM family members behave in the same way regarding the constraints to be imposed upon the parameters of the

Law:	Normal $\mathcal{N}(\mu, \sigma^2)$	Binomial $B(n, \mu)$	Poisson $\mathcal{P}(\mu)$	Gamma $\mathcal{G}(\mu, \nu)$	Inverse Gaussian $\mathcal{IN}(\mu, \sigma^2)$
Support	$y \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma^2 \in \mathbb{R}^{+*}$	$y \in [\![0, n]\!]$ $n \in \mathbb{N}^*$ $\mu \in [0, 1]$	$y \in \mathbb{N}$ $\mu \in \mathbb{R}^+$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\nu \in \mathbb{R}^{+*}$	$y \in \mathbb{R}^+$ $\mu \in \mathbb{R}^{+*}$ $\sigma^2 \in \mathbb{R}^{+*}$
$\alpha(\mu)$ Support	μ $\alpha \in \mathbb{R}$	$\ln[\mu/(1-\mu)]$ $\alpha \in \mathbb{R}$	$\ln \mu$ $\alpha \in \mathbb{R}$	$-\mu^{-1}$ $\alpha \in \mathbb{R}^{-*}$	$-(2\mu^2)^{-1}$ $\alpha \in \mathbb{R}^{-*}$
ϕ Support	σ^2 $\phi \in \mathbb{R}^{+*}$	1	1	ν^{-1} $\phi \in \mathbb{R}^{+*}$	σ^2 $\phi \in \mathbb{R}^{+*}$
$b(\alpha)$	$\alpha^2/2$	$\ln(1+e^\alpha)$	e^α	$-\ln(-\alpha)$	$-(-2\alpha)^{1/2}$
$c(y, \phi)$	$-\frac{1}{2}\left(\frac{y^2}{\phi} + \ln(2\pi\phi)\right)$	$\ln(C_n^{ny})$	$-\ln(y!)$		$-\frac{1}{2}\left(\ln(2\pi\phi y^3) + \frac{1}{\phi y}\right)$
$\mu(\alpha) = \mathbb{E}[Y]$	α	$e^\alpha/(1+e^\alpha)$	e^α	$-1/\alpha$	$(-2\alpha)^{-1/2}$
Constraints	$\lvert\alpha\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$	$\lvert\alpha\rvert < +\infty$	$\lvert\alpha\rvert < +\infty$	$\lvert\alpha\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$	$\lvert\alpha\rvert < +\infty$ $\phi < +\infty$ $\phi \nrightarrow 0$

Table 1: Constraints to be applied on the parameters of the exponential family (12).

exponential family. Provided the dispersion does not tend to 0 and once the trend parameter and / or the dispersion are bounded, the conditional classification log-likelihood as well as its derivative (including, particularly, the entropy derivative) remains finite. These results confirm the necessity to choose compact sets for the parameters space.

5. Simulation study

In this section, we perform a simulation study to check the previous theoretical results. One would like to validate the convergence properties of both the $ML_{cc}E$ and the ICL^* criterion. By sampling observations coming from finite mixtures (firstly mixtures of normal regressions, then mixtures of Poisson regressions), we show that i) the new estimator seems to tend towards the true parameter (maximizing the expected log-contrast), ii) the selection criterion looks consistent while being adapted for clustering purposes. Indeed, the well-known tendency to overestimate the order of the mixture when using AIC or BIC tends to disappear, which is a very good news since ICL^* was initially designed to this end. For practical considerations, the latter result is also interesting because it enables to lower the model dimension: this should bring more robustness to the parameters estimation, probably leading to more relevant predictions. Moreover we mecanically lower the probability to make an error when assigning observations to mixands in less complex

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

Model class:	X	True β_1	True β_2	L_{cc} minimizer β_1^0	L_{cc} minimizer 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

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 our 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 $ML_{cc}E$ is not very close to the theoretical MLE (true parameters), 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), and see whether the $ML_{cc}E$ tends towards the L_{cc} minimizer. 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} minimizer tends to 0, with a dispersion that narrows down when the number of observation increases. Results are summarized in Figure 2, which confirm this convergence whatever the random variables type. Indeed, mixtures of linear regressions stand for the continuous case whereas

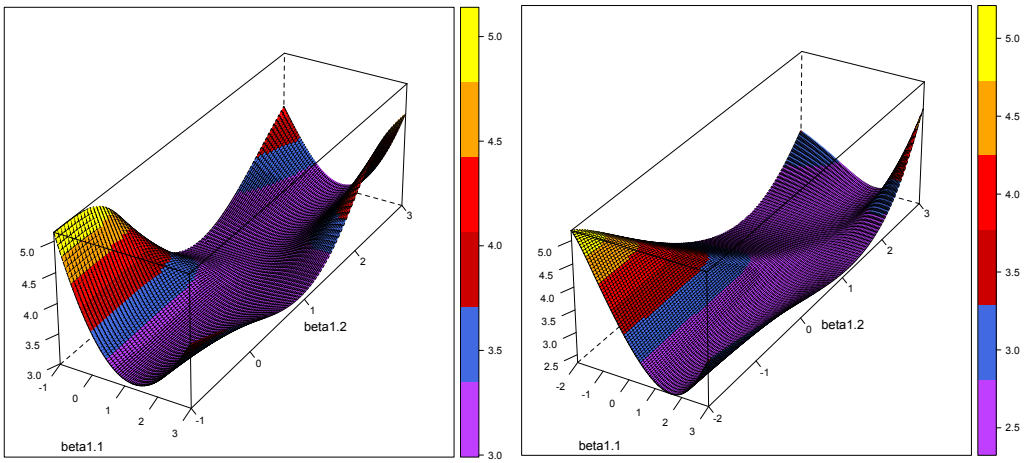

Figure 1: KL divergence between the true distribution and the L_{cc} contrast. On the left: normal regression mixture. On the right: poisson regression mixture.

mixtures of Poisson regressions represent the discrete case. Despite the high number of observations, notice that the optimization can still lead to some erroneous estimations (especially in the Poisson case): this could partly be explained by the contrast complexity and some difficulties experienced in the maximization algorithms.

5.2. Illustration of ICL^* consistency

As Figure 2 suggests, we 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 stands for the selection of a mixture density where components are strongly overlapping, whereas the other one corresponds to well-separated component densities. Theoretically speaking, the ICL selection criterion may not be too different from the BIC one 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 $ML_{cc}E$. This is clearly censed to affect the model selection in a different manner, to be identified in this case study. Of course, it is much more exciting to look at what is happening with strongly overlapping components. The entropic term is obviously not negligible in such a case and the selection criteria have no

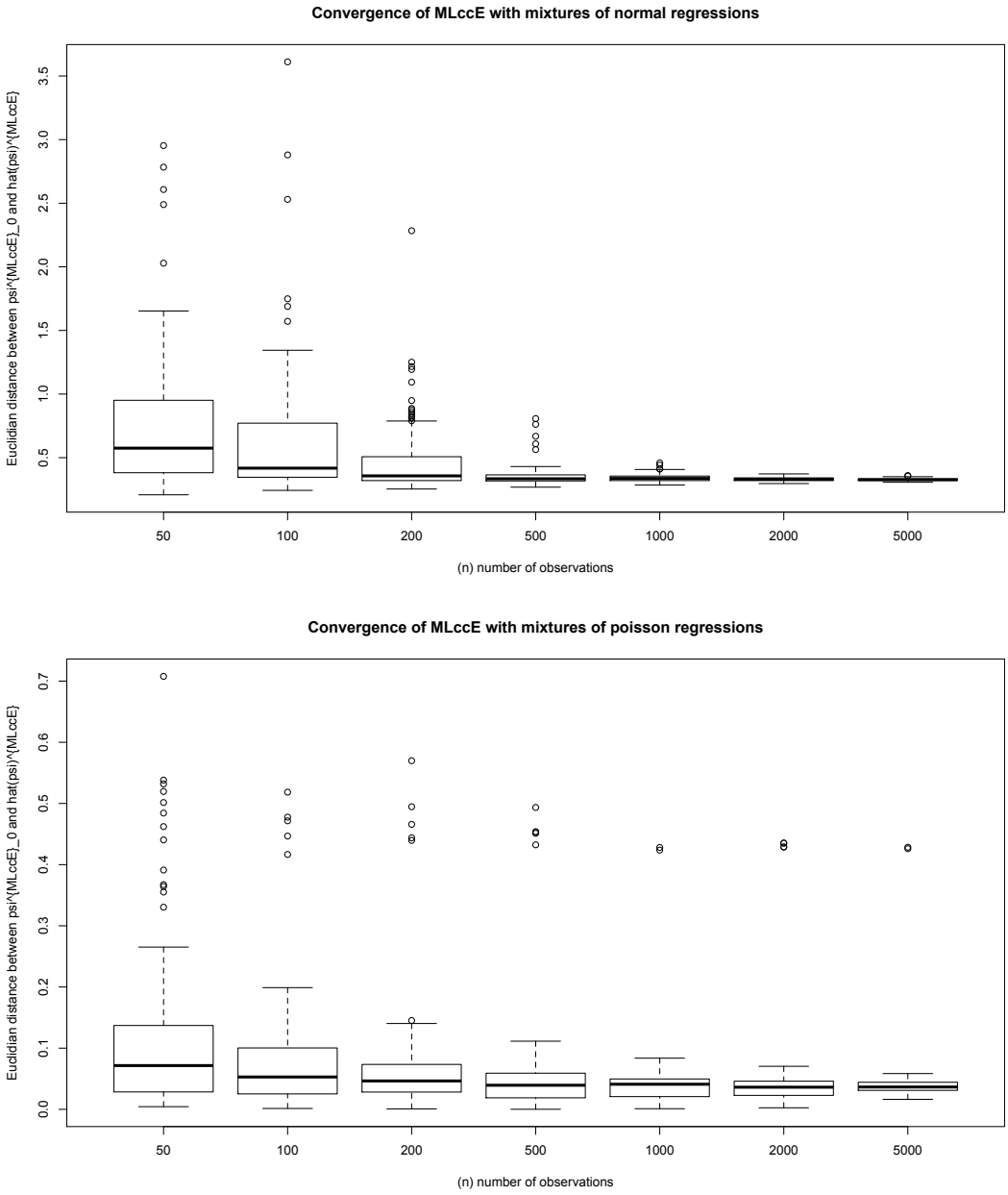

Figure 2: Boxplot (100 experiments) $ML_{cc}E$ convergence towards the maximizer of the expected log-contrast. From top to bottom: normal and poisson regression mixtures.

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

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	$\sim \mathcal{U}(1,4)$

reason to behave analogously. In particular, do we still observe the famous issue of overestimating g? To overcome the problem of little confidence when assigning observations to mixture components, the ICL^* may strongly penalize a mixture density embedding highly overlapping components: naturally, this should result in a simpler model (which sometimes could even become too simplistic). To check this, 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 values of the model selection criteria (AIC, BIC and ICL from the MLE; and ICL^* from the $ML_{cc}E$);
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 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

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	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 5: Consistency of ICL^* in the case of mixtures of poisson 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%

problem of selecting too much complex mixtures (% overestimation) and looks better than AIC, BIC and ICL when trying to recover the right number of components. 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 g is not really satisfying (even if the probability to overestimate g is once again diminished). This is certainly linked with what was observed on Figure 2: indeed 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

In this paper, we developed a new approach in clustering population from mixtures of generalized linear models. In this context this is a key matter since most of model selection criteria such as AIC or BIC 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 Baudry (2009): it embraces both the convergence of a specific M-estimator (adapted to the clustering purpose) and the consistency of the ICL^* criterion (a derivative of ICL). The bounds that we obtained through concentration inequalities hold even in a non-asymptotic framework. Moreover, they are valid even when the considered density is unbounded, a crucial feature when dealing with GLM mixtures. Concerning the ICL^* criterion, empirical studies on simulated gaussian mixtures in Baudry (2009) reveal that the overestimation of the number of components tends to disappear: this is also confirmed in our simulation study involving different GLM members for mixture components. 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 and seek the theoretical properties of such estimators: integrating specific quantities within the contrast instead of considering them as part of the penalty is innovative, and should lead to promising developments.

Appendix A. Concentration inequality

In this section, we present the concentration inequality that we use to derive our exponential bounds. This inequality is due to Talagrand (1994). We use a formulation of this inequality similar to the one used in Einmahl and Mason (2005).

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. Einmahl and Mason (2005) 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 Einmahl and Mason (2005) 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) $\sup_{f\in\mathcal{F}} E\left[f(Y, X)^2\right] \leq \sigma^2$,
(iii) $M \leq \frac{1}{4\nu}\sqrt{n\sigma^2/\log(C_1M/\sigma)}$, *with* $C_1 = \max(e, C^{1/\nu})$.

Then,

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

Appendix B. Covering numbers

Lemma B1. *Let* $\psi_g^b = (\pi_{10}, ..., \pi_{n_g 0}^b, \theta_{10}^b, ..., \theta_{n_g 0}^b)$, $\psi_1 = (\pi_{11}, ..., \pi_{n_g 1}, \theta_{11}, ..., \theta_{n_g 1})$, $\psi_2 = (\pi_{11}, ..., \pi_{n_g 1}, \theta_{11}, ..., \theta_{n_g 1})$, *with* $\sum_{i=1}^{n_g} \pi_{il} = 1$ *for* $l = 0, 1, 2$. *Assume that, for all* $i \in \{1, ..., n_g\}$,

$$|\pi_{i1} f_i(y; \theta_{i1}) - \pi_{i2} f_i(y; \theta_{i2})| \leq \Lambda_1(y) \|\psi_1 - \psi_2\|, \tag{B.1}$$

$$\left| \frac{\pi_{i1} f_i(y; \theta_{i1}) - \pi_{i0} f_i(\theta_{i0}; y)}{\|\psi_1 - \psi_g^b\|} - \frac{\pi_{i2} f_i(y; \theta_{i2}) - \pi_{i0} f_i(\theta_{i0}; y)}{\|\psi_2 - \psi_g^b\|} \right| \leq \Lambda_2(y) \|\psi_1 - \psi_2\|. \tag{B.2}$$

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* $A(y) < \infty$,

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

Consider the classes of functions

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

with $g_{i,\psi_0}(y) = \lim_{\psi \to \psi_g^b} g_{i,\psi}(y)$. *Then, assuming that, for all* $\psi \in \Psi_g$, $\pi_l \geq \pi_- > 0$ *for all* $l = 1, ..., n_g$,

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

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

Proof. Define, for $l = 0, 1, 2$,

$$\begin{aligned} g_l(y) &= \pi_{il} f_i(y; \theta_{il}) + \sum_{j=1}^{n_g} \mathbf{1}_{j \neq i} \pi_{jl} f_j(y; \theta_{jl}), \\ h_l(y) &= \frac{\pi_{il} f_i(y; \theta_{il})}{g_l(y)}. \end{aligned}$$

Write, for $l = 0, 2$,

$$\begin{aligned} h_1(y) - h_l(y) &= \frac{\pi_{i1} f_i(y; \theta_{i1}) - \pi_{il} f_i(y; \theta_{il})}{g_1(y)} \\ &\quad + \left\{ \frac{g_l(y) - g_1(y)}{g_1(y) g_l(y)} \right\} \pi_l f_i(y; \theta_{il}). \end{aligned} \tag{B.5}$$

Observe that $g_1(y) \geq \Lambda_-(y)$, so using equation (B.1), we get

$$|h_1(y) - h_2(y)| \leq \frac{\Lambda_1(y)\|\psi_1 - \psi_2\|}{\Lambda_-(y)} + \frac{\Lambda_1(y)\Lambda_0(y)\|\psi_1 - \psi_2\|}{\Lambda_-(y)^2}.$$

Due to assumption (B.3),

$$|h_1(y) - h_2(y)| \leq (A(y) + A(y)^2)\|\psi_1 - \psi_2\|, \tag{B.6}$$

for some constant $A > 0$. Next, observe that, again from (B.1),

$$\frac{|h_2(y) - h_0(y)|}{\|\psi_2 - \psi_g^b\|} \leq \frac{\Lambda_1(y)}{\Lambda_-(y)} \leq A(y), \tag{B.7}$$

where we used again (B.3) and the fact that $\min(g_2(y), g_0(y)) \geq \Lambda_-(y)$. Using again (B.5), but this time for $l = 0$, we get, according to (B.2),

$$\begin{aligned}\left|\frac{h_1(y) - h_0(y)}{\|\psi_1 - \psi_g^b\|} - \frac{h_2(y) - h_0(y)}{\|\psi_2 - \psi_g^b\|}\right| &\leq \frac{\Lambda_2(y)\|\psi_1 - \psi_2\|}{\Lambda_-(y)} + \frac{\Lambda_2(y)\Lambda_0(y)\|\psi_1 - \psi_2\|}{\Lambda_-(y)^2} \\ &\leq (A(y) + A(y)^2)\|\psi_1 - \psi_2\|.\end{aligned}$$

Moreover, note that

$$|\log(h_0(y))| \leq \frac{1}{h_0(y)}, \tag{B.8}$$

and that

$$\frac{h_1(y)}{\min(h_0(y), h_1(y))} \leq A(y), \tag{B.9}$$

from (B.3). Finally, again due to (B.3), note that, for $l = 0, 1, 2$,

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

Let $H(x) = x\ln(x)$. Observe that $H(h_l(y)) = Ent(\psi_l; y)$. Then decompose

$$\begin{aligned}\left|\frac{[H(h_1(y)) - H(h_0(y))]}{\|\psi_1 - \psi_g^b\|} - \frac{[H(h_2(y)) - H(h_0(y))]}{\|\psi_2 - \psi_g^b\|}\right| &\leq \left|\frac{h_1(y) - h_0(y)}{\|\psi_1 - \psi_g^b\|} - \frac{h_2(y) - h_0(y)}{\|\psi_2 - \psi_g^b\|}\right| \\ &\quad \times \left(|\log(h_0(y))| + \frac{h_1(y)}{\min(h_0(y), h_1(y))}\right) \\ &\quad + \frac{|h_2(y) - h_0(y)|\,|h_1(y) - h_2(y)|}{\min(h_0(y), h_1(y), h_2(y))\|\psi_2 - \psi_g^b\|},\end{aligned}$$

where we used that $|\log(x/x')| \leq |x - x'|/\min(x, x')$. Combining this with (B.6), (B.7), (B.8), (B.9) and (B.10) shows that

$$\left| \frac{[H(h_1(y)) - H(h_0(y))]}{\|\psi_1 - \psi_g^b\|} - \frac{[H(h_2(y)) - H(h_0(y))]}{\|\psi_2 - \psi_g^b\|} \right| \leq \Lambda_3(y)\|\psi_1 - \psi_2\|,$$

where

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

□

Lemma B2. *Using the notations of Lemma B1, let $\mathcal{G}^g = \sum_{i=1}^{n_g} \mathcal{G}_i$. Then $\mathcal{G}^g$ is a class of functions bounded by $G(y) = n_g[\Lambda_3(y)\,\mathrm{diam}(\Psi_g) + g_{\psi_0}(y)]$.*

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

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

Proof. Due to (B.4) in Lemma B1, for all $\mathfrak{g} \in \mathcal{G}_i$, $|\mathfrak{g}(y)| \leq \tilde{G}(y) = \Lambda_3(y)\,\mathrm{diam}(\Psi_g) + g_{\psi_0}(y)$, where $\mathrm{diam}(\Psi_g)$ denotes the diameter of Ψ_g for the norm $\|\cdot\|$. Then, $\mathcal{G}^g$ is bounded by $G(y) = n_g[\Lambda_3(y)\,\mathrm{diam}(\Psi_g) + g_{\psi_0}(y)]$. From Lemma 2.13 in Pakes and Pollard (1989), we get $N_{\tilde{G}}(\varepsilon, \mathcal{G}_i) \leq C\varepsilon^{-V}$, for some constants $C > 0$ and $V > 0$. The result then follows from Lemma 16 in Nolan and Pollard (1987). □

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*

$$E[\exp(2\rho[\tilde{A}(y) + |\nabla_{\psi_g} \ln f(\psi_g^b; y)| + \sup_{i=1,\ldots,g} |g_{i,\psi_0}(Y)|])] < \infty,$$

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*

$$P(x; g) \leq 4\left\{\exp\left(-\frac{A_3 x^2}{n}\right) + \exp\left(-\frac{A_4 x}{\ln n}\right)\right\} + A_7 \exp(-\rho x/2),$$

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

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

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.

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.

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

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.

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

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.

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.

A.P. Dempster, Laird N.M., and Rubin D.B. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society*, 39:1–38, 1977.

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. ISSN 0894-9840. doi: 10.1023/A:1007769924157. URL http://dx.doi.org/10.1023/A:1007769924157.

Uwe Einmahl and David M. Mason. Uniform in bandwidth consistency of kernel-type function estimators. *Ann. Statist.*, 33(3):1380–1403, 2005. ISSN 0090-5364. doi: 10.1214/009053605000000129. URL http://dx.doi.org/10.1214/009053605000000129.

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

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.

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

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

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

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

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.

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

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.

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

Friderich Leisch. Modelling background noise in finite mixtures of generalized linear regression models. Technical Report 37, Department of Statistics, University of Munich, 2008.

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

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

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

Deborah Nolan and David Pollard. U-processes: rates of convergence. *Ann. Statist.*, 15(2):780–799, 1987. ISSN 0090-5364. doi: 10.1214/aos/1176350374. URL http://dx.doi.org/10.1214/aos/1176350374.

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

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

Ariél Pakes and David Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057, 1989. ISSN 0012-9682. doi: 10.2307/1913622. URL http://dx.doi.org/10.2307/1913622.

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

B.D. Ripley. *Pattern Recognition and Neural Networks*. Cambridge University Press. Cambridge, 1995.

M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Ann. Probab.*, 22(1):28–76, 1994. ISSN 0091-1798. URL http://links.jstor.org/sici?sici=0091-1798(199401)22:1<28:SBFGAE>2.0.CO;2-W&origin=MSN.

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

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

D.S. Young and D.R. Hunter. Mixtures of regressions with predictor-dependent mixing proportions. *Computational Statistics & Data Analysis*, 54(10):2253–2266, 2010.