Prediction of the Nash through Penalized Mixture of Logistic Regression Models

Marie Morvan, Emilie Devijver, Madison Giacofci, Valérie Monbet

PREDICTION OF THE NASH THROUGH PENALIZED MIXTURE OF LOGISTIC REGRESSION MODELS

By Marie Morvan *, Emilie Devijver †, Madison Giacofci * and Valérie Monbet *

* *Univ Rennes, CNRS, IRMAR - UMR 6625, F-35000 Rennes, France*
† *Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP‡, LIG, 38000 Grenoble, France*

Abstract

In many medical problems, it is common to face heterogeneous data with unknown patients profiles leading to difficulties to build a good diagnosis model. In this paper, our aim is to build a suitable and interpretable diagnosis tool to predict the Non-Alcoholic Steatohepatitis (NASH), taking into account the structure and the dimension of the spectrometric data. Thus, we introduce a penalized mixture of logistic regression model that allows the prediction of a binary response. Parameters estimation is done using the EM algorithm. In the presence of a high number of covariates, estimation of the full covariance matrix and interpretation of the regression coefficients is not trivial. To highlight relevant covariates for the prediction and their links, we apply a penalization to the covariance matrix and the regression coefficients. The estimated model depends on regularization parameters that allow to adjust the strength of the penalization. Automatic selection tools are used to choose the best model, namely with respect to the AIC criterion. A simulation study is performed to evaluate the proposed method, and the application on the NASH data set is presented. This model leads to better prediction performance than the competitive methods and provides useful tools to better understand the data.

1. Introduction

Non-Alcoholic Fatty Liver Diseases (NAFLD) is nowadays one of the leading cause of liver disease in Western countries. NAFLD is characterized by a built-up of fat in the liver. Due to the worlwide increase of obesity and type 2 diabetes, its prevalence, currently estimated to 24%, is expected to further grow in the future (Younossi et al., 2018a). Combined with liver cell injuries and inflammation, subgroups of NAFLD patients may derive to Non-Alcoholic Steatohepatitis (NASH), a more serious form of NAFLD, which represents the second cause of liver transplants in the USA. While being crucial for patients and healthcare systems, diagnosis of NASH is still an issue as the disease is essentially asymptomatic with low specific syndromes. Properly detecting and staging the severity of NASH patients requires a liver biopsy with the disadvantages of being invasive, costly, source of potential surgical complications and affected with sampling and inter-observer variability. In addition, liver biopsy can not be envisaged at a large scale, nor be repeated in time. At the time being, although diagnosis of NASH based on noninvasive modalities is an active field of research, no prediction procedure achieved consensus in the medical community (see *e.g.* Younossi et al., 2018b).

‡. Institute of Engineering Univ. Grenoble Alpes

. *MSC 2010 subject classifications:* 62H30, 62J02, 62P10

Keywords and phrases: Mixture regression model, Prediction, Variable selection, Heterogeneous data

Mid-infrared spectroscopy provides a molecular fingerprint of biological sample, as for example blood sera or urine, and may represent a promising approach to predict and understand the physiological consequences of the disease. In this project we study a data set containing blood serum spectrum of 395 morbidly obese patients (Anty et al., 2010), including 66 patients diagnosed as NASH. Diagnosis has been established on the basis of liver biopsy conducted within the Hepatology unit of the Nice University Hospital. From a statistical perspective, our goal is to propose a statistical learning model to assign a score to a patient spectrum.
Basically, mid-infrared spectra represent the absorbance of a biological sample discretely measured on a given range of wavelength. Such data are complex to analyze as they rise several statistical issues.

1. **Dimension.** Each individual spectrum comprises hundreds of measurement points and more specifically, more variables than individuals in the sample. Hence the problem lies in a high-dimensional framework.
2. **Inner nature of the data.** Mid-infrared spectrum contains the whole molecular composition of a biological sample. However, not every molecular group is expected to be associated with the disease. It is then required to use dedicated method to select and identify wavelength associated with the disease progression.
3. **Inter-individual variability.** The studied population is heterogeneous and yields high individual fluctuations. This may be attributed to external or metabolic factors that are directly linked to the pathology development.

Taking account of subject-specific variability is a central point in our approach. It is now commonly accepted that individual metabolisms may strongly differ depending on lifestyle, feeding or global medical path. Compelling a cohort of patients to fit a rigid model appears then to be naive. A popular approach is to decompose the cohort onto few reference profiles summarizing as much as possible metabolic behaviors. Usually referred as *disease trajectories* in the literature (Ross and Dy, 2013), such reference profiles provide experts and practitioners interpretable knowledge on patient conditions and valuable information to predict the diagnosis.
In supervised learning, Generalized Linear Mixed-effects Models (GLMM) is a flexible setting to structure variability across individuals through a grouping structure (Breslow and Clayton, 1993). However, it requires the grouping structure to be known in advance which is a difficult task as the NASH disease remains poorly known. Mixture models become then the appropriate tool by modelling the unknown sub-population through a discrete latent variable. As a model-based clustering approach, posterior cluster membership probabilities are obtained for each individual based on its observations. Such approaches have been widely studied in the regression context through the mixture of regression models paradigm (Grün and Leisch, 2007; Khalili and Chen, 2007; Städler et al., 2010). In dedicated research works the central idea is to model the conditional distribution of the response variable given the predictors as a mixture distribution. The covariates are then treated as non-random variables and the heterogeneity is assume to be fully contained in the conditional distribution. Such a modelling is often unrealistic for observational data as predictors can display significantly different behaviours depending on cluster. Most importantly, those models are not appropriate in a prediction framework: as the response variable for a new observation is not

observed, posterior cluster membership can not be computed. As a matter of fact, Grün and Leisch (2007) examine model fitting strategies for finite mixture of generalized linear regression but do not discuss the prediction issue in this framework. Concurrently, the machine learning community developed the mixture of experts models. Such models are dedicated to prediction when dealing with heterogeneous regions in the input space. Based on the principle of divide-and-conqueer, mixture of experts models aim at estimating the distribution of the response variable conditionally on the covariates as for mixture of regressions but model the prior cluster size as functional mixing weigths depending on the covariates. Classical mixing weights includes exponential weights with the softmax gating network or more general ones (Yuksel et al., 2012). In this respect, mixture of experts models generalize the mixture of regression framework but can not be considered as a proper mixture model. Hence, although appealing from a prediction perspective, estimated model parameters have no clear biological interpretation.

In this paper, we consider an in-between approach where the joint distribution of the predictors and the response is defined as a mixture. Thus, our model has the advantage to exploit grouping structure information carried out in the conditional distribution as well as in the predictors. In addition, prediction of the response variable can be straightforwardly computed as the posterior cluster membership probabilities do not depend on the unobserved response for a new observation. For the NASH disease data, we are dealing with binary response. Hence the model considered is a mixture of logistic regression models, where the mixture is defined for the joint distribution, and the covariates are assumed to be Gaussian. Inference is performed through the Expectation-Maximation (EM) algorithm which is adapted to the latent variable setting.

As previously mentioned, spectrometry data carry out the whole molecular information contained in a biological sample whereas only few wavelengths are expected to be informative regarding to the prediction of the disease. The regression coefficient vector is thus expected to be sparse and variable selection should be considered to achieve accurate parameter estimation of our model. We choose to consider an L_1-penalized likelihood approach to simultaneously achieve variable selection and parameter estimation. Such approach can be straightforwardly slot into the EM mechanism and benefit from theoretical guarantees in the mixture of regression framework (Khalili and Chen, 2007; Städler et al., 2010). Similarly, GLasso estimator is considered for the precision matrix of the predictors in the clusters. Besides, this second penalization highlights dependence between covariates and reduces the dimension.

Our method leads to the estimation of patient profiles and the estimated parameters allow the interpretation of the molecular variables involved in the disease. Their interactions can be represented with graphical models using the estimated precision matrices and has brought valuable insights about the NASH disease for the experts we collaborate with.

The rest of the paper is organized as follows. First, the mixture of logistic regression model is presented in Section 2, and the prediction step is described thereby. Then, parameter estimation by maximum likelihood maximization is described in Section 3.1 and in Section 3.2, regularization is introduced to reduce the dimension and to allow interpretation. Model selection is discussed in Section 3.3 to select the number of clusters. Details for the EM algorithm are then provided in Section 4. Section 5 investigates the numerical performance in a simulation study. An R Markdown file available at http://rpubs.com/morvan_ma/PMLR

allows to replay some of the analysis on simulated data. Finally, Section 6 illustrates our result on the data set concerning the NASH disease.

2. Model

2.1 Mixture of logistic regression model

Let $(\mathbf{X}, Y)$ be two random variables , Y being a binary response in $\{0, 1\}$ and $\mathbf{X} \in \mathbb{R}^p$ a set of p covariates. In a latent class framework, we assume that individuals are spread among K unknown clusters of prior size $(\pi_k)_{k=1,...,K}$ with $0 < \pi_k$ and $\sum_{k=1}^{K} \pi_k = 1$. We denote by $\mathbf{Z} = (Z_1, ..., Z_K)$ the latent random variable, where Z_k equals to 1 if the individual is in cluster k and 0 otherwise. Considering that both covariates and the regression model depend on the underlying unknown cluster structure, the mixture of regression model appears to be an appropriate setting. The logistic model for the response Y in the mixture setting is then defined as

$$Y|\{\mathbf{X} = \mathbf{x}, Z_k = 1\} \sim \mathcal{B}\left(p^{(k)}(\mathbf{x})\right),$$

with $\mathbf{x}$ a realization of $\mathbf{X}$ and $p^{(k)}(\mathbf{x}) = \mathbb{P}(Y = 1|\{\mathbf{X} = \mathbf{x}, Z_k = 1\})$. The predictors are related to the response variable Y using the logistic link function

$$\text{logit}\big(p^{(k)}(\mathbf{x})\big) = \mathbf{x}\boldsymbol{\beta}_k,$$

where logit : $x \mapsto \log(\frac{x}{1-x})$ and $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})$ is the vector of regression coefficients in the k-th cluster. Moreover, given that $\{Z_k = 1\}$, the covariate $\mathbf{X}$ is modelled as a multivariate Gaussian distribution such as

$$\mathbf{X}|\{Z_k = 1\} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k),$$

with $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and Σ_k respectively the mean and the covariance matrix of the predictors in the k-th cluster.
Finally, the probability distribution of Y given $\mathbf{X} = \mathbf{x}$ can be written as a finite mixture of logistic regression model. As the response is binary (0 or 1), the probability distribution is given by

$$\mathbb{P}(Y = 1|\mathbf{X} = \mathbf{x}; \Phi_K) = \sum_{k=1}^{K} \pi_k \mathbb{P}(Y = 1|\{\mathbf{X} = \mathbf{x}, Z_k = 1; \boldsymbol{\phi}_k\}) = \sum_{k=1}^{K} \pi_k p^{(k)}(\mathbf{x}),$$

where $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\beta}_k)$ is the vector of parameters of the kth cluster and the whole set of parameters to be estimated for the mixture model with K clusters is denoted by $\Phi_K = (\pi_1, ..., \pi_K, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$.

2.2 Prediction

Finite mixture of regression models, where the mixture relies on the conditional distribution, are used to model a cluster structure within regression models, emphasize on parameters and estimation. However, one of our goal is to predict the response variable Y (in the real data set, disease or not), and one would adapt the model to do prediction. On the other hand, mixture of experts models, introduced in Jacobs et al. (1991) and reviewed in Yuksel

et al. (2012), deal with prediction, and our model can be compared to this class of models for the prediction step. When considering mixture of experts models, one has to choose a particular metric to weight the prediction. In our case, the mixture model introduced in section 2.1 is determined by the multinomial distribution for the cluster variable Z, inducing posterior probabilities as weights, as some particular cases of mixture of experts using Gaussian parametric forms in the gate (Yuksel et al., 2012).
Formally, the prediction rule is the following,

$$\begin{aligned}\mathbb{E}(Y_0|\mathbf{X}_0=\mathbf{x}_0) &= \sum_{k=1}^{K}\mathbb{P}(Y_0=1,Z_0=k|\mathbf{X}_0=\mathbf{x}_0)\\ &= \sum_{k=1}^{K}\mathbb{P}(Y_0=1|Z_{0k}=1,\mathbf{X}_0=\mathbf{x}_0)\mathbb{P}(Z_{0k}=1|\mathbf{X}_0=\mathbf{x}_0) \qquad (1)\end{aligned}$$

As a matter of fact, the theoretical prediction rule does not depend on the unobserved response variable y_0, meaning that for prediction purpose, only the cluster membership information contained in $\mathbf{X}$ is required. However, for the estimation step, the observed information $(\mathbf{X}_i, Y_i)$ have been used to take into account the link between the predictors and the response variable structuring the data into clusters. Hence, the estimated model parameter depends on posterior probability $\tau_{ik} = P(Z_{ik}=1|\mathbf{X}_i=\mathbf{x}_i, Y_i=y_i)$ that implicitly accounts for the grouping structure carried out by the conditional distribution. The estimation of parameters detailed in Section 3 leads to an estimate for (1), to predict y_0. Let $\tau'_{0k} = \mathbb{P}(Z_{0k}=1|\mathbf{X}_0=\mathbf{x}_0)$, then

$$\hat{\tau}'_{0,k} = \frac{\hat{\pi}_k f_{\mathbf{X}_0}(\mathbf{x}_0;\hat{\boldsymbol{\mu}}_k,\hat{\Sigma}_k)}{\sum_{l=1}^{K}\hat{\pi}_l f_{\mathbf{X}_0}(\mathbf{x}_0;\hat{\boldsymbol{\mu}}_l,\hat{\Sigma}_l)},$$

and,

$$\hat{y}_0 = \sum_{k=1}^{K}\hat{\tau}'_{0,k}\frac{\exp(\mathbf{x}_0^t\hat{\boldsymbol{\beta}}_k)}{1+\exp(\mathbf{x}_0^t\hat{\boldsymbol{\beta}}_k)}.$$

3. Estimation method

3.1 Parameters estimation by penalized maximum likelihood

Let $\{(y_i,\mathbf{x}_i)_{i=1,\dots,n}\}$ be a sample of size n of realizations of the random variables $(\mathbf{X},Y)$ introduced previously. Parameters estimation is done via maximum likelihood estimation. As many variables are considered in the real data set, we suppose that some of them are not relevant. Therefore, feature selection is considered in order to highlight relevant variables and get an interpretable model. Two types of parameters have to be estimated for each cluster, the regression coefficients and the covariance matrices. A consistent estimation of a full covariance matrix remains challenging even in a moderate dimension. Moreover, for our application, some covariates are conditionally independent to the others, but there is a strong structure of correlation between covariates, which is encoded in the Gaussian model with covariance matrix Σ by the precision matrix Θ, corresponding to the inverse of the covariance matrix: $\Theta = \Sigma^{-1}$. Thus, a Graphical Lasso penalization (as proposed by Friedman et al.

(2008)) is used to constrain some values of the cluster specific precision matrices associated to the covariates to be equal to zero. In a similar way, in our situation, we suppose that only a subset of the covariates is relevant for the prediction of the response variable, and the regression coefficients are expected to be sparse. Hence, a Lasso penalty (Tibshirani, 1994) is used to constrain some elements of the vector $\boldsymbol{\beta}_k$ to be exactly equal to zero. Therefore, an ℓ_1-regularized estimation strategy is adopted to obtain sparse estimates of both the precision matrices of the predictors $\Theta_1, \ldots, \Theta_K$ and the regression coefficients $\boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_K$. According to the sample $(y_i, \mathbf{x}_i)_{i=1,...,n}$, the penalized likelihood problem we aim to solve is thus given by, for $\lambda_k \geq 0, \rho_k \geq 0$, for all $k = 1, \ldots, K$,

$$(2) \quad \hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = \arg\max_{\Phi_K} \left\{ \ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \Phi_K) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k\|_1 \right\},$$

where $\|\boldsymbol{\beta}_k\|_1 = \sum_{j=1}^{p} |\beta_{k,j}|$, $\Theta_k = \Sigma_k^{-1}$ is the precision matrix in the kth cluster and $\|\Theta_k\|_1$ denotes the sum of the absolute values of Θ_k. The quantities λ_k and ρ_k correspond to regularization parameters driving the amount of shrinkage on the parameters $\boldsymbol{\beta}_k$ and Θ_k for every cluster k, $k = 1, \ldots, K$. A method to select them is described in Section 3.2.

3.2 Regularization parameters selection

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ and $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)$ determine the amount of regularization, and their choice is important in the penalized likelihood approach. Large values of tuning parameters tend to select a simple model whose parameters estimates have smaller variance, whereas small values of the tuning parameters lead to complex models, with smaller bias. The choice of the optimal tuning parameters is based on a trade-off between the bias and variance. There are $2K$ parameters to be tuned. Generalized Cross-validation is a popular method for the tuning parameters selection in penalized likelihood approaches (see for example Fan and Li, 2001; Khalili and Chen, 2007). The tuning parameters are chosen one at a time by minimizing the Cross-Validation criterion over a set of possible values. However, Cross-Validation methods are computationally heavy. Moreover, a study by Wang et al. (2007) shows that Generalized Cross-validation may lead to the selection of some irrelevant variables. Different studies (for example Wang et al., 2007; Jiang et al., 2018; Lloyd-Jones et al., 2018; Khalili and Lin, 2013) suggest the use of the Bayesian Information Criterion (BIC) for tuning parameters selection. The BIC (Schwarz, 1978) makes a trade-off between the number of free parameters and the fit to the data (through the likelihood function), and selects a sparse model fitting well the data. For a given number of clusters K, for a parameter estimate $\hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$, obtained with tuning parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$, the BIC is defined as

$$BIC^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = -2 \ln \mathcal{L}(\hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + \nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})} \ln(n),$$

with $\hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the maximizer of the penalized log-likelihood function and $\nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the number of free parameters of the model, corresponding to the number of non-zero coefficients of the model. In our setting, two different tuning vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ have to be chosen. An automatic procedure is proposed to build a grid of possible tuning parameters for each cluster: first a classification by MAP rule is derived after few iterations of the EM algorithm, and the grids for $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are constructed for each cluster. Parameters are then estimated for each

combination of the possible values of the two dimensions tuning parameters grid and finally, the parameters minimizing the BIC are retained.

3.3 Selection of the number of clusters

The number of clusters K is a sensible parameter because it is linked with the heterogeneity of the population. However, it is latent and has to be selected. The previous selection tools are also used to select the number of clusters K. In the framework of mixture models the BIC is commonly used to select the number of clusters K (Keribin, 2000). It is defined by

$$BIC_K = -2\ln \mathcal{L}(\hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + \nu_K \ln(n),$$

with $\hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the maximum likelihood estimator restricted to the relevant variables and ν_K the number of free parameters of the model estimated with K clusters. However, the BIC was designed for non-structured data, and Biernacki et al. (2000) shows that in some mispecified situations, it can lead to a wrong choice of K. For that reason, the ICL criterion was developed (see Biernacki et al., 2000; McLachlan and Peel, 2000), and adds to the BIC formulation an entropy term taking into account the concentration shape of the clusters. For a model estimated with K clusters, it is defined as

$$ICL_K = BIC_K - 2\sum_{i=1}^{n}\sum_{k=1}^{K} \tau_{ik}\ln\tau_{ik}$$

with τ_{ik} the posterior probabilities estimated for $i = 1, \ldots, n$ and $k = 1, \ldots, K$. This criterion is more adapted to the model-based clustering framework. In the case of well-separated clusters, the entropy term is close to zero, and the ICL criterion value is close to the BIC value. In case of non separated clusters, the entropy term is highly negative and the value of ICL_K increases. Thus, the ICL criterion favors models with well separated clusters.
Finally, for predictive purpose, the AIC is known to be more suitable (Shmueli, 2010). With the previous notations, the AIC is defined as

$$AIC_K = -2\ln \mathcal{L}(\hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + 2\nu_K.$$

The model is estimated for different possible values of K, and the model selected is the one minimizing the chosen criterion value. In the sequel, the three criteria are compared.

4. EM algorithm

4.1 Formulae

An Expectation-Maximization (EM) algorithm is used to optimize (2). The EM algorithm is an iterative algorithm used in incomplete-data problems to approximate maximum likelihood estimates, introduced in Dempster et al. (1977). In our model, the observed data $(y_i, \mathbf{x}_i)_{i=1,\ldots,n}$ can be viewed as being incomplete, considering that the latent information of cluster is missing, for all $i = 1, \ldots, n$. We denote by $(y_i, \mathbf{x}_i, \mathbf{z}_i)$ the augmented or complete data, with $\mathbf{z}_i = (z_{i1}, ..., z_{iK})$ referring to the unobserved data, where $z_{ik} = 1$ if observation i belongs to cluster k, 0 otherwise. In the presence of clusters (non-observed), it is of common

use to consider the likelihood of the complete data, which thanks to the Bayes' Rule can be decomposed into

$$\begin{aligned}\ln \mathcal{L}(y_1,\dots,y_n,\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\Phi_K) = &\ln \mathcal{L}(y_1,\dots,y_n|\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\beta})\\ &+\ln \mathcal{L}(\mathbf{x}_1,\dots,\mathbf{x}_n|\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\mu},\boldsymbol{\Theta})\\ &+\ln \mathcal{L}(\mathbf{z}_1,\dots,\mathbf{z}_n;\boldsymbol{\pi}).\end{aligned}$$

The ordinary EM algorithm consists in maximizing the conditional expectation of the complete log-likelihood (conditionally on the observed data) given a current set of parameters Φ_K^* instead of maximizing the likelihood alone. In our situation, we consider the conditional expectation of the penalized complete log-likelihood such as

$$\begin{aligned}\arg\max_{\Phi_K} \quad \Bigg\{ &\mathbb{E}_{\Phi_K^*}\left[\ln \mathcal{L}(y_1,\dots,y_n,\mathbf{x}_1,\dots,\mathbf{x}_n,\mathbf{z}_1,\dots,\mathbf{z}_n;\Phi_K)|\mathbf{x}_1,\dots,\mathbf{x}_n,y_1,\dots,y_n\right]\\ &-\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k\|_1-\sum_{k=1}^{K}\rho_k\|\Theta_k\|_1\Bigg\}.\end{aligned}$$

Thus, the EM algorithm alternates between two steps, called E-step and M-step, until convergence. We first describe those two steps in our context.
The E-step of the algorithm comes up to predict the non-observed latent class by their conditional expectation for all individuals $i=1,\dots,n$ using posterior probabilities $\tau_{ik}=\mathbb{P}(Z_{ik}=1|\mathbf{X}_i=\mathbf{x}_i,Y_i=y_i;\boldsymbol{\phi}_k^{[h]})$ for all $k=1,\dots,K$, given the current parameters $\pi_k^{[h]},\boldsymbol{\mu}_k^{[h]},\Sigma_k^{[h]},\boldsymbol{\beta}_k^{[h]}$, at iteration $[h]$. For all $i=1,\dots,n$ and for all clusters $k=1,\dots,K$, there are given by

$$\tau_{ik}^{[h+1]}=\frac{\pi_k^{[h]}f_{\mathbf{X},Y}\left(\mathbf{x}_i,y_i;\boldsymbol{\mu}_k^{[h]},\Sigma_k^{[h]},\boldsymbol{\beta}_k^{[h]}\right)}{\sum_{\ell=1}^{K}\pi_\ell^{[h]}f_{\mathbf{X},Y}\left(\mathbf{x}_i,y_i;\boldsymbol{\mu}_\ell^{[h]},\Sigma_\ell^{[h]},\boldsymbol{\beta}_k^{[h]}\right)},$$

where $f_{\mathbf{X},Y}(.)$ is the joint density function of $(\mathbf{X},Y)$. The joint density is computed thanks to the relation $f_{\mathbf{X},Y}(.)=f_{Y|\mathbf{X}}(.)f_{\mathbf{X}}(.)$ for which the distributions are known to be a binomial distribution with parameter $p^{(k)}(\mathbf{x})$ for $Y|\mathbf{X}=\mathbf{x}$ and a normal distribution with parameters $\boldsymbol{\mu}_k$ and Σ_k for $\mathbf{X}$.
Thus, it is possible to compute the expected value of the penalized complete log-likelihood $Q(\Phi_K,\Phi_K^{[h]})$ given the observed data $(y_i,\mathbf{x}_i)_{i=1,\dots,n}$ and $\Phi_K^{[h]}$,

$$\begin{aligned}Q(\Phi_K,\Phi_K^{[h]})=\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{ik}^{[h+1]}\Bigg(&\ln\pi_k^{[h]}-\frac{1}{2}\left(y_i\mathbf{x}_i^t\boldsymbol{\beta}_k^{[h]}-\mathbf{x}_i^t\boldsymbol{\beta}_k^{[h]}-1\right)+\frac{1}{2}\ln|\Theta_k^{[h]}|\\ &-\frac{1}{2}\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{[h]}\right)^T\Theta_k^{[h]}\left(\mathbf{x}_i-\boldsymbol{\mu}_k^{[h]}\right)\Bigg)-\sum_{k=1}^{K}\lambda_k\|\boldsymbol{\beta}_k^{[h]}\|_1-\sum_{k=1}^{K}\rho_k\|\Theta_k^{[h]}\|_1.\end{aligned}$$

The M-step of the algorithm consists in maximizing the conditional expectation computed in the E-step, with respect to the parameter Φ, to obtain $\Phi^{[h+1]}$, according to

$$\Phi_K^{[h+1]}=\arg\max_{\Phi_K}\quad\left\{Q(\Phi_K,\Phi_K^{[h]})\right\}.$$

Then, the update of every parameter at iteration $[h+1]$ is given by, for all $k = 1, \ldots, K$,

$$\widehat{\pi}_k^{[h+1]} = \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{[h+1]},$$

$$\widehat{\boldsymbol{\mu}}_k^{[h+1]} = \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \mathbf{x}_i,$$

$$\hat{\Theta}_k^{[h+1]} = \arg\max_{\Theta_k} \left\{ \sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(\log\det \Theta_k - \frac{1}{2} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h+1]} \right)^T \Theta_k \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h+1]} \right) \right) - \rho_k \|\Theta_k\|_1 \right\},$$

$$\hat{\boldsymbol{\beta}}_k^{[h+1]} = \arg\max_{\boldsymbol{\beta}_k} \left[\sum_{i=1}^{n} \tau_{ik}^{[h+1]} \left(y_i \mathbf{x}_i^T \boldsymbol{\beta}_k - \ln\left(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)\right) \right) \right] - \lambda_k \|\boldsymbol{\beta}_k\|_1.$$

4.2 Tuning the EM algorithm

The convergence of the EM algorithm to the optimal solution can highly depend on the initial parameters. Thus, the algorithm has to start from sensible initial parameters in order to avoid convergence to a local maximum of the likelihood function. Here, we adopt a Search/Run/Select (S/R/S) strategy as developed in Biernacki et al. (2003):

- Find t initial positions of parameters: obtain a partition of the set of observations of the explanatory variables $(\mathbf{x}_i)_{i=1,\ldots,n}$ into K clusters with a k-means algorithm (Macqueen (1967)). According to this clustering, compute the logistic regression estimators in each cluster, for each of the t trials.
- Run a small fixed number of iterations of the EM algorithm at the t initial positions previously found.
- Among these t possible starting values, select the one which maximizes the log-likelihood to start the EM algorithm.

The E and M steps are repeated until the log-likelihood does not improve more than a particular threshold or a maximum number of iterations is reached.

5. Experiments on simulated data

We perform simulations to evaluate the performance of the prediction through a penalized mixture of regressions. Our objectives are (i) to evaluate the quality of the estimation of parameters, (ii) to evaluate the prediction performance of our model and (iii) to evaluate the interest of using our model in situations where non homogeneous data are collected, when clusters modulate the prediction rule.

5.1 Competing methods and evaluation criteria

Strategies are compared over 30 replicated data sets for each case. The method is assessed from two points of view: estimation and prediction.

To evaluate the estimators, we compare the performance in estimation of the proposed method, denoted PMLR (Penalized Mixture of Logistic Regression) with the penalized lo-

gistic regression, denoted PLR, and with a finite mixture of logistic regression models introduced in Grün and Leisch (2007) and denoted MLR. Remark that the main difference between this method and ours is the penalization, but also the clustering, which relies on $Y|\mathbf{X}$ whereas ours relies on $(\mathbf{X}, Y)$, allowing the prediction.
The number of clusters varies in the estimation step, and performance of selection procedure is studied for the AIC, the BIC and the ICL criteria.
Bias and variance of every estimator $\hat{\boldsymbol{\mu}}, \hat{\boldsymbol{\Sigma}}, \hat{\boldsymbol{\beta}}$ are computed.
The support of $\boldsymbol{\beta}$ and $\boldsymbol{\Theta}$ are also compared with the true one, and summarized through the relevant variable detection (RVD) and irrelevant variable elimination (IVE) [1] of the regression coefficients. The RVD of an estimate $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,p})$ is defined as $RVD_k = TP_k/(TP_k + FN_k)$ with TP_k the number of coefficients $(\beta_{k,j})_{j=1,...,p}$ correctly predicted as non zero and FN_k the number of coefficients $(\beta_{k,j})_{j=1,...,p}$ predicted zero while being non zero. The IVE of an estimate $\hat{\boldsymbol{\beta}}_k = (\hat{\beta}_{k,1}, ..., \hat{\beta}_{k,p})$ is defined as $IVE_k = TN_k/(TN_k + FP_k)$ with TN_k the number of coefficients $(\beta_{k,j})_{j=1,...,p}$ correctly predicted as zero and FP_k the number of coefficients $(\beta_{k,j})_{j=1,...,p}$ predicted non zero while being equal to zero. The RDV is equal to 1 if all the relevant variables are retained in the model. The IVE is equal to 1 if all the irrelevant variables are eliminated from the model.
Performance in clustering is studied through the Adjusted Rand Index (ARI) which measures the similarity between two partitions (Hubert and Arabie, 1985). A value of ARI of 1 means that the predicted partition is equal to the theoretical partition.

To evaluate the quality of prediction of the binary response, we compare our results with PLR and logistic regression (denoted LR), for which we are able to do prediction.
We consider as a criterion the Area Under the Receiver Operating Characteristic curve (AUROC). It measures the diagnostic ability of a binary classifier system at all discrimination thresholds and is frequently used in medical diagnosis problems. To give more precise results, we also add for few cases the full Receiver Operating Characteristic curve (ROC curve).
Remark that prediction performance is assessed for a fixed number of clusters K.

5.2 Simulation settings

Simulated data comes from a mixture with $K = 3$ clusters of proportions $\boldsymbol{\pi} = (0.3, 0.3, 0.4)$, $n = 250$ or 500 observations, and $p = 40$ variables, with only 8 relevant variables (non-zero coefficients). Four different cases are studied, described in details in Annex 9.1. In the first two cases, non-zero coefficients concern the same variables in every cluster (same support), but the concentration of the clusters and the balance between the two classes in each cluster are different (easy case (1) and difficult case (2)). In the third and the fourth cases, non-zero coefficients concern different variables according to the cluster (different supports). In Case 4 (difficult case), the clustering relies on $Y|\mathbf{X}$ whereas $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_3$ are the same, where in Case 3 (easy case), the clustering relies on $Y|\mathbf{X}$ and $\mathbf{X}$.

5.3 Results and interpretation

Model Selection

1. These criteria correspond to the sensibility and specificity criteria in binary classifier evaluation, renamed here to avoid confusion.

Criterion	AIC				BIC				ICL			
Method	PMLR		MLR		PMLR		MLR		PMLR		MLR	
Sample size n	250	500	250	500	250	500	250	500	250	500	250	500
Same support												
Easy case (1)	**3**	**3**	1	1	2	2	1	1	2	2	1	1
Difficult case (2)	**3**	**3**	1	1	**3**	**3**	1	1	**3**	**3**	1	1
Different supports												
Easy case (3)	**3**	**3**	1	2	**3**	2	1	1	**3**	2	1	1
Difficult case (4)	2	2	1	2	1	2	1	1	1	1	1	1

TABLE 1

Selected number of clusters. We compare our method PMLR with the mixture of logistic regression MLR. For each method, we provide the number of clusters leading to the smallest value of the following criterion: AIC, BIC and ICL, as a majority vote over the 30 repetitions for each of the 4 simulation cases. The simulations are done for two sample sizes (n=250 and n=500). The true number of clusters, 3, is in bold.

The majority vote of the model selection according to the lowest criterion values is computed over the repetition of each case for AIC, BIC and ICL, for the models estimated with 1 to 4 clusters. For each criterion and each case the number of clusters mostly selected is collected in Table 1. For our method, the AIC leads to the selection of the right number of cluster in almost all the cases, except for the case 4 where two clusters are very similar, and it selects two clusters. The most important information for the clustering is carried by $\mathbf{X}$ and in this case the two cluster-model is selected. ICL criterion and BIC lead to the selection of wrong number of clusters for five out of the eight cases. However, these criteria lead to a better model selection for our method than the classical mixture of regression method, where the criteria lead to a one-cluster model selection in almost all cases. To conclude, model selection is better done with our method with respect to those criteria, and we will focus on AIC which has the best performance.

Performance in estimation

Method	PMLR		MLR	
Sample size n	250	500	250	500
Same support - easy case (1)	0.89	0.91	0	0.01
Same support - difficult case (2)	1	1	0.01	0.01
Different supports - easy case (3)	0.96	0.97	0.01	0.01
Different supports - difficult case (4)	0.42	0.42	0	0.01

TABLE 2

Performance in clustering via ARI. We compare PMLR with the mixture of logistic regressions MLR. For each method, for a number of clusters fixed to $K = 3$, we compute the Adjusted Rand Index: closest to 1, better the clustering. It is done over the 30 repetitions for each of the 4 simulation cases. The simulation are also done for two sample sizes, n=250 and n=500.

The quality of classification is checked with the Adjusted Rand Index (ARI), detailed in Table 2. ARI values are higher for clustering obtained with our method than with a mixture of logistic regressions done on the conditional variable, even for models with the wrong number of clusters (K=2). Performance is similar for $n = 250$ and $n = 500$, meaning that asymptotic is already achieved. The dataset has been generated according to PMLR in which the joint distribution $(\mathbf{X}, Y)$ in the mixture is considered. The information carried by

$\mathbf{X}$ is used explicitly for the clustering. In MLR, only the conditionnal distribution is used in the estimation task, so the clustering is driven by the relation between $\mathbf{X}$ and Y. Those results show that, if the covariates are also structured into clusters, MLR will not have good performance.

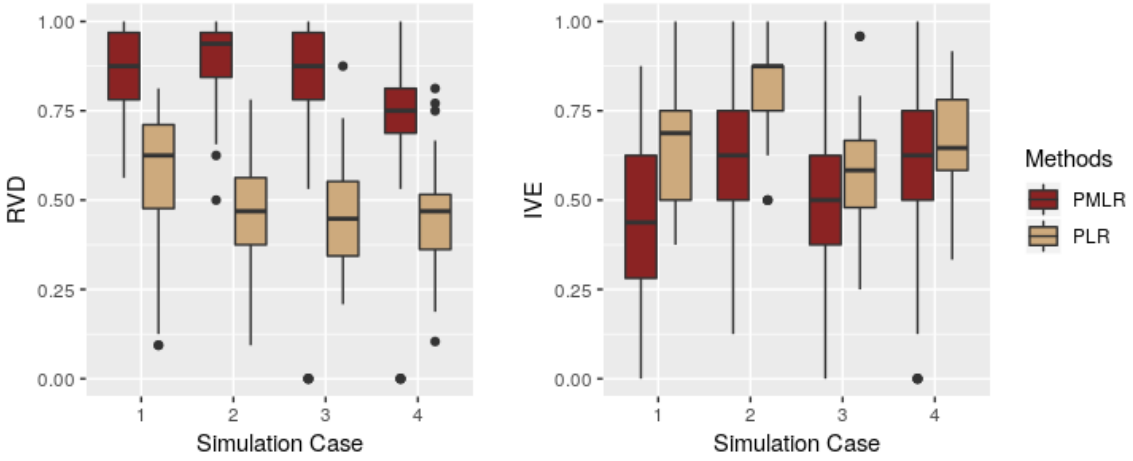

Figure 1: Performance in variable selection for $\hat{\boldsymbol{\beta}}$. We provide boxplots for the Relevant Variable Detection (RVD) on the left and boxplots for the Irrelevant Variable Elimination (IVE) on the right. We compare our method (PMLR) with the Penalized Logistic Regression (PLR), which selects relevant variables in the regression matrix with an ℓ_1-penalty. Regularization parameters are selected with the BIC. Every case introduced in the simulation setting is described in abscissa.

In Figure 1, we illustrate the ability to find the relevant variables for all simulation cases. We remark that it is higher with our method than with the penalized logistic regression. The ability to eliminate the irrelevant variables is slightly lower with our method in the cases 1 and 2, and similar than the penalized logistic regression for cases 3 and 4. These results show that our method keeps too many variables, but succeeds in selecting the relevant ones, unlike penalized logistic regression that eliminates to many variables including important ones. In medicine, it is particularly interesting to not delete relevant variables, as interpretation is of great interest.
We also pay attention to the bias and variance of each estimator. Performance was good for our method, better for higher sample size (as the maximum likelihood has good performance asymptotically).

Performance in prediction
The prediction performance is shown in Figure 2. Our method leads clearly to the best performance for cases 2 and 3. Case 1 is an easy case, every method is performing well. For the case 4, the best performance is obtained with our method but the wrong number of clusters is chosen due to the strong similarities between two clusters. We could also conclude that our method has a small variability in AUROC among the 30 repetitions.

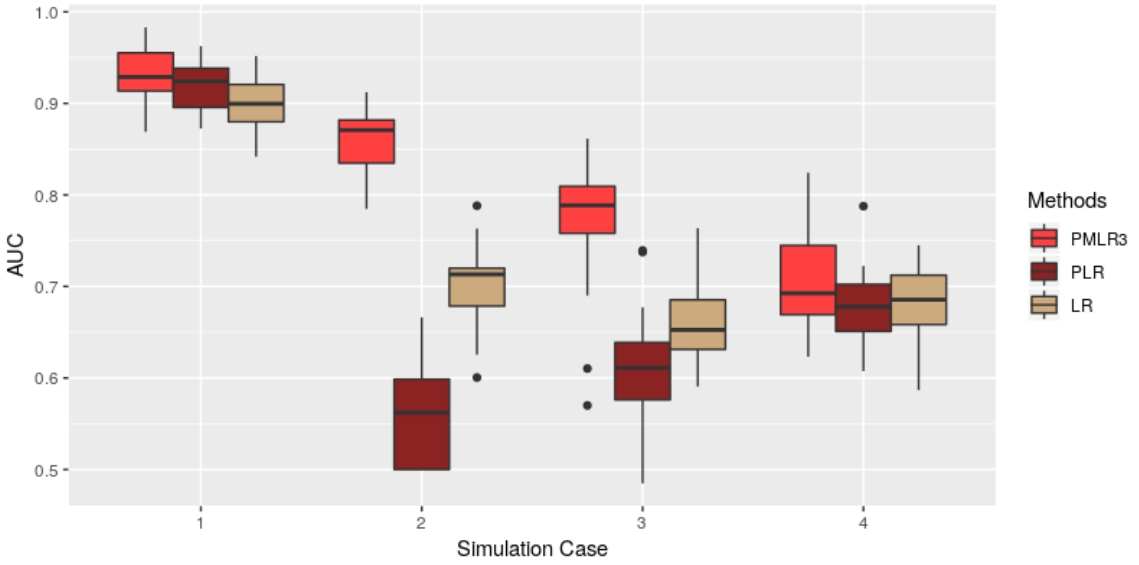

Figure 2: Prediction performance. We provide boxplots for Area Under the Receiver Operating Characteristic (AUROC) obtained for the 4 simulation cases for each method: Logistic Regression (LR), Penalized Logistic Regression (PLR) and Penalized Mixture of Logistic Regression with 3 clusters (PMLR-3). Simulations are repeated 30 times.

To illustrate those results, the ROC curves obtained for the 30 repetitions of the simulation cases 2 and 3 for the competitive prediction methods PMLR (with 3 clusters), PLR and LR are shown in Figure 3. For these cases, the ROC curves obtained with our prediction method are above the ROC curves obtained with the two competitive methods, showing better prediction performance in these cases with our method. Remark that PLR and LR have very close curves, whereas their AUROC are different.

6. Application to the NASH data set

On biological applications, we usually face to individual effects changing the prediction rule. We assume here that there exist homogeneous clusters of observations, relying on biological or genetic similarities. Those similarities might be independent of the severity of the disease that we want to diagnose. A better prediction of the disease is achieved considering this cluster structure and more importantly, a better understanding of the disease evolution process is given by our modeling. First, we describe in more details the data set. Then, we describe the results by our procedure PMLR, from estimation, interpretation and biological points of view.

6.1 The NASH data set

Non Alcoholic Steatohepatitis (NASH) is a disease affecting the liver, and characterized by a fat deposit in the liver cells. In the long term, this disease results in important perturbations of the patient's metabolism. Currently, the diagnosis is obtained after a liver biopsy and an

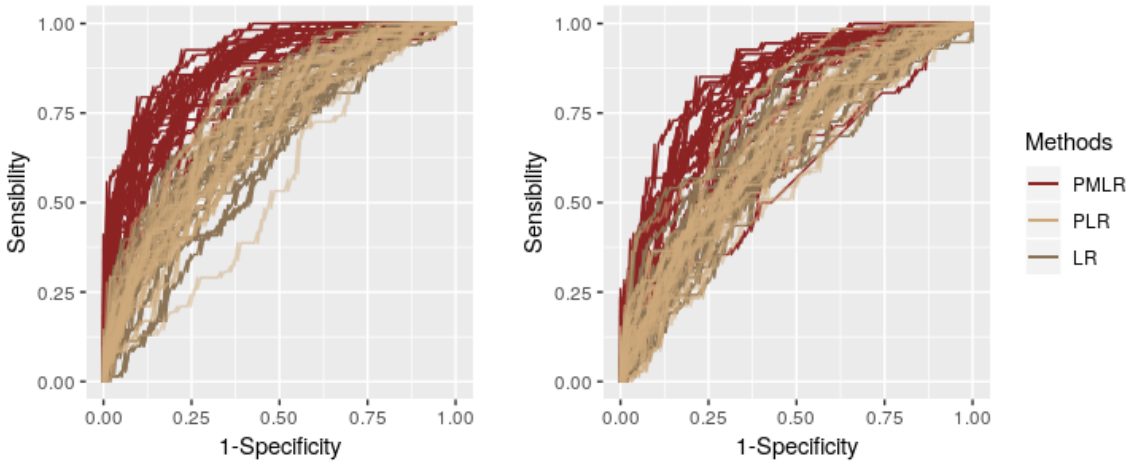

Figure 3: Prediction performance. We provide Receiver Operating Characteristic curves (ROC) for each method: Penalized Mixture of Logistic Regression (PMLR) with 3 clusters, Penalized Logistic Regression (PLR), Logistic Regression (LR). Plot on the left corresponds to the case 2, whereas plot on the right corresponds to case 3. Simulations are replicated over 30 data sets, and we plot a curve per data set to highlight the variance.

histological study of the sample, which is an invasive method. The method we propose is based on spectrum measured on blood serum, then is non invasive and leads to a prediction of the disease. Moreover, as the method is model-based, interpretation of the prediction results is possible, leading to a better understanding from experts of the disease and its evolution.

Experts suggest that different unknown patient typologies exist, and for each typology the molecular signature to establish the diagnosis is different. The proposed model, based on the existence of a discrete latent class (cluster), takes into account this feature of our data. As we deal with high-dimensional data (large number of wavelengths in each spectrum), experts suspect that among the information available, some variables are irrelevant for the NASH diagnosis. One of the aim is to select the relevant variables to improve the diagnosis and to allow a better interpretation of the data.

The data set we consider is the following. We observe 395 patients, including 66 NASH patients ($\sim 17\%$), coming from Nice hospital, in France. Clinical variables and spectrometric measures on sera samples are available. The spectrometric curves represent a molecular fingerprint of the sample and reflect the metabolic profile of each patient, affected by the liver condition. Portions of the spectrometric curves are selected by experts for there ability to describe metabolism variations that could be linked to the liver condition of the patients. Spectral variables are used to construct the prediction model, whereas biological and clinical variables only help for the interpretation.

6.2 Analyses and results

Model selection

The data set is randomly split in a calibration set containing 4/5 of the individuals (316 individuals including 53 NASH patients) and a validation set containing the individuals left (79 individuals including 13 NASH patients). These sets are randomly chosen but contain the same proportion of NASH patients and no significant differences are observed between clinical variables within the two sets. The model is estimated on the calibration set, for 1 to 3 clusters. The model selection criteria are evaluated for each model and represented in Table 3. The lowest AIC and BIC values are obtained for the model estimated with two clusters. The lowest ICL value is obtained for the model estimated with one cluster, but with a slight difference with the ICL value corresponding to the model with two clusters. Following the conclusions detailed in Section 5.3 we select the model with 2 clusters according to the AIC value.

	K = 1	K = 2	K = 3
AIC	-50180	**-50459**	-30957
BIC	-49936	**-49970**	-30627
ICL	**-49936**	-49901	-30365

TABLE 3

Model selection. Comparison of the model selection criteria AIC, BIC and ICL for models estimated by PMLR on the calibration set of the NASH data set for 1 to 3 clusters. The bold values indicate the best values of the criteria obtained.

Estimators and models

Graphical models obtained from the sparsely estimated precision matrices for each cluster are represented in Figure 4. We observe different relationships between variables according to the cluster. Considering only the relationship between the variables, we can see that for the first cluster, there is a group of variables from X2 to X11 with a lot of links. For the second cluster, we observe two groups of strongly linked variables: first with the variables X2, X3, X4, X6, X7, X8, and second with the variables X1, X5, X12, X14, X17, X19. The links between variables are completely different according to the cluster. The node color represents the coefficient value of the variable for the model applying to the considered cluster. For the first cluster, we observe that a lot of regression coefficients are close or equal to zero. For the second cluster, coefficients have more extreme values. The completely different variables links and effects on the prediction for each cluster suggest different metabolic mechanisms of the patients, depending on the cluster.

Statistical interpretation of the constructed model

The proportions of each cluster are $0.66, 0.34$. The proportion of disease cases changes according to the cluster: 19 % in cluster 1 and 12 % in cluster 2.

The performance obtained with PMLR1 and PLR are similar. Indeed, when considering only 1 cluster, PMLR1 consists in a penalized logistic regression.

In Table 4, we observe the highest AUROC values and good classification rate values for the model estimated with two clusters, that shows the lowest AIC and BIC values. The model selection with the AIC and BIC is consistent with the cross validation performance obtained. Compared to competing method, the chosen model has the best performance with the highest AUROC (0.75) and good classification rate values (0.76), and a high negative predictive value indicating a good screening test. The repartition of the predicted scores

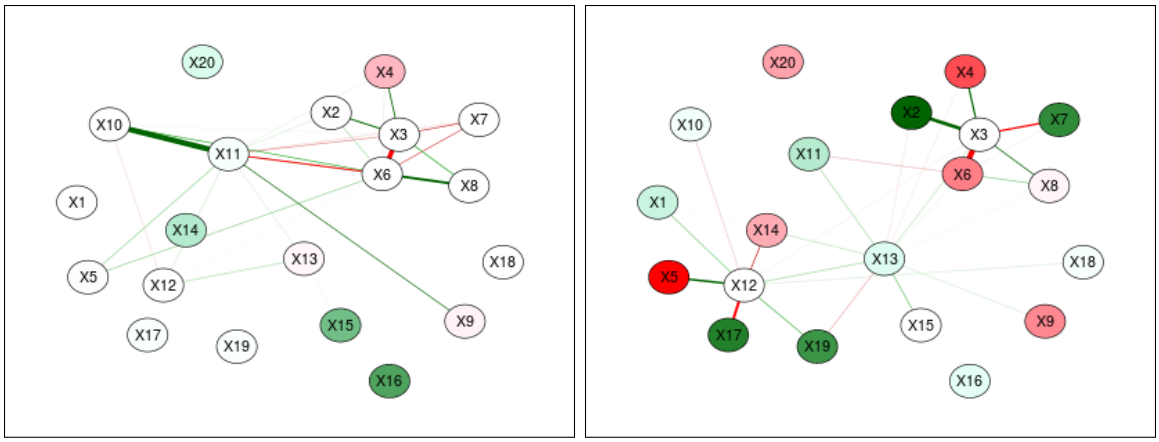

Figure 4: Graphical models. The selected model has 2 clusters, the network related to the first cluster is on the left and the network related to the second cluster is on the right. Precision matrices are sparsely estimated, so the networks are sparse. Arrow colors correspond to the sign of the partial correlation (green for positive correlation, red for negative correlation) and the intensity of the edges correspond to the value of the correlation (stronger is the color, stronger is the correlation). Node colors correspond to the value of the regression coefficient for each variable for the model applying to the considered cluster. No color indicates a value of zero for the regression coefficient.

	PMLR-1	PMLR-2	PMLR-3	PLR	LR
AUROC	0.64	**0.75**	0.68	0.64	0.67
Se	0.62	**0.77**	0.85	0.62	0.69
Sp	0.62	**0.76**	0.5	0.62	0.7
NPV	0.89	**0.94**	0.94	0.89	0.92
PPV	0.24	**0.38**	0.25	0.24	0.31
CR	0.62	**0.76**	0.56	0.62	0.7

TABLE 4

Comparison of the prediction performance obtained with different methods: our method with 1 to 3 clusters (PMLR-1, PMLR-2, PMLR-3), penalized logistic regression (PLR) and logistic regression (LR). The chosen model is represented in bold script. The quantities we use for the comparison are the following: Area Under the Receiver Operating Characteristic (AUROC), sensibility (Se), specificity (Sp), negative predictive value (NPV), positive predictive value (PPV), classification rate (CR).

according to the real class of the individuals from the validation set is represented by Figure 5.

The threshold from which a patient is label as a NASH patient is automatically chosen as the threshold maximizing the sum of the sensibility and specificity. This threshold is represented in Figure 5 and allows a good separation between patients with NASH and patients without NASH.

Biological interpretation of the constructed model

We characterize the clusters obtained with the selected model with the clinical variables available in Table 5. Values correspond to the mean over individuals for each cluster. We

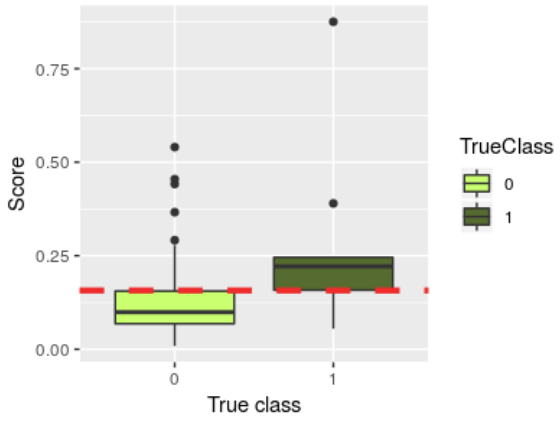

Figure 5: Performance in prediction. We provide boxplots of the score in function of the true class, for the Penalized Mixture of Logistic Regression (PMLR). The red dashed line corresponds to the threshold automatically learned.

	Cluster 1	Cluster 2	p-value	Signif
Age	40	39	0.6	
Sex	0.84	0.88	0.4	
Weight	119	120	0.4	
BMI	44	45	0.6	
Height	164	164	0.7	
AST	28	26	0.2	
ALT	38	29	0.001	**
AST.ALT	0.88	1	6.10^{-4}	**
GGT	47	34	10^{-3}	**
Gluc	6.2	5.7	0.06	
Insuline	24	21	0.2	
HBA1C	6.1	5.7	0.01	*
chol	5.5	5.2	4.10^{-3}	**
HDL	1.4	1.4	0.5	
LDL	3.2	3.1	0.4	
TG	2	1.4	4.10^{-7}	**

TABLE 5

Characterization of the clusters obtained with the clinical variables. The model selected has two clusters. Values correspond to the mean over individuals for each cluster. We compare the mean in each cluster with a t-test, reporting the p-value and the significance. For the variable Sex, the women's rate is precised, and a Fisher test is used for the comparison.

compare the mean in each cluster with a t-test, reporting the p-value and the significance. We observe that there is a significant difference between the two clusters for the variables ALT (corresponding to the alanine transaminase), AST.ALT (ratio aspartate aminotransferase-alanine transaminase), GGT (Gamma-glutamyltransferase), HBA1C (glycated hemoglobin), chol (cholesterol) and TG (triglyceride). In the first cluster, the variables linked with diabete (Gluc and HBA1C) have higher values as well as the variables indicating liver problems (ALT, GGT). More generally, patients from the first cluster seems to have more severe liver complication than patients from the second cluster according to the seric indicators. Moreover, there is no significant difference between the two clusters for the morphological

variables (weight, height, BMI), so that model allows to recognize the severity of the liver injury even when patients are not different for morphological variables.
We also represent the distribution of the predicted scores according to the different stage of histological variables. We can see in Figure 6 that the predicted score is a good indicator of the histological caracteristics of the patient. Indeed, the score increase with the Steatosis, ballooning and inflammation stage, indicators used to establish the diagnosis. Fibrosis doesn't enter in the NASH definition and thus is not predicted by our model.

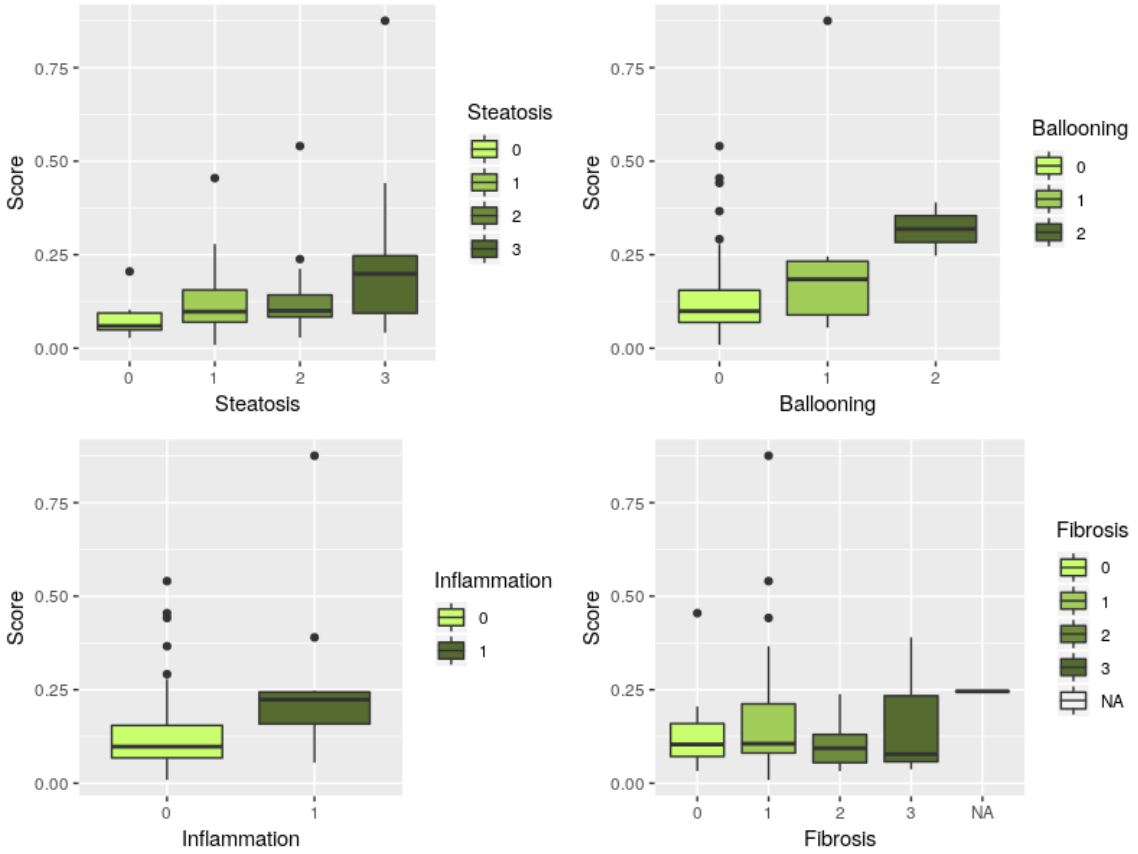

Figure 6: Boxplots of the repartition of the score predicted with the selected model according to the stage of the histological variable considered. Top left: steatosis, top rigth: ballooning, bottom left: inflammation, bottom right: fibrosis.

7. Conclusion

In this paper we have presented a predictive method that allows to build a model on data structured in clusters, including non-relevant variables. This method provides interpretable tools to help for a better understanding of the data, with similar or higher prediction performance than competitive predictive models. This work was conducted on a real data problem concerning the use of the spectrometric technology to develop a non-invasive diagnosis tool to predict the NASH disease. We obtained encouraging results both in terms of prediction

performance and in terms of interpretation, with clusters characterized by clinical variables and a prediction score linked to histological variables. Moreover, it is common in medical problems to face structured data with unknown patients profiles. Thus, our method could be broader used to handle this kind of situations.
In this reported work, the analysis focused on a selection of wavenumbers performed by experts, but an interesting direction for further research would be to consider the whole spectra. Thus, we would like to adapt this method to handle functional data and perform the selection of specific areas of the spectra. This would highlight the type of molecules involved in the disease and offer the possibility to link the different areas of the spectra. Moreover, the discriminant information allowing the best prediction could be held at the same time by the intensity values of the spectra at particular wavenumbers and by the shape of the spectra at specific areas of the spectra. A different projection scale could be considered according to the area of the spectrum selected.

8. Acknowledgements

Part of this work was supported by the CNRS and AMIES institutions through the exploratory projects PEPS I3A AppSpec and PEPS I3A STATOPO. The authors are grateful to Rodolphe Anty (Centre Hospitalier Universitaire de Nice) and the Hepatology unit for the provision of the data set and to Diafir and more precisely Hugues Tariel and Maëna Le Corvec for the spectrometric measures. The authors also thank Olivier Loréal (INSERM, Univ. Rennes) for his precious medical expertise and helpful suggestions.

References

R. Anty, A. Iannelli, S. Patouraux, S. Bonnafous, V. Lavallard, M. Senni-Buratti, I. Ben Amor, A. Staccini-Myx, MC. Saint-Paul, F. Berthier, P. Huet, Y. Le Marchand-Brustel, J. Gugenheim, P. Gual, and A. Tran. A new composite model including metabolic syndrome, alanine aminotransferase and cytokeratin-18 for the diagnosis of non-alcoholic steatohepatitis in morbidly obese patients. *Alimentary pharmacology & therapeutics*, 32 11-12:1315–22, 2010.

C. Biernacki, G. Celeux, and G. Govaert. Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(7):719–725, 2000.

C. Biernacki, G. Celeux, and G. Govaert. Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models. *Computational Statistics & Data Analysis*, 41(3):561–575, 2003.

N. E. Breslow and D. G. Clayton. Approximate inference in generalized linear mixed models. *Journal of the American Statistical Association*, 88(421):9–25, 1993.

A.P. Dempster, N.M. Laird, and D.B. Rubin. Maximum likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society*, 39(1):1–38, 1977.

J. Fan and J. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456):1348–1360, 2001.

J. Friedman, T. Hastie, and R. Tibshirani. Sparse Inverse Covariance Estimation with the Graphical Lasso. *Biostatistics*, 9(3):432–441, 2008.

B. Grün and F. Leisch. Fitting finite mixtures of generalized linear regressions in R. *Computational Statistics & Data Analysis*, 51(11):5247–5252, July 2007. .

L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985.

R.A. Jacobs, M.I. Jordan, S.J. Nowlan, and G.E. Hinton. Adaptive mixtures of local experts. *Neural Computation*, 3(1):79–87, 1991.

Y. Jiang, Y. Conglian, and J. Qinghua. Model selection for the localized mixture of experts models. *Journal of Applied Statistics*, 45(11):1994–2006, 2018.

C. Keribin. Consistent estimation of the order of mixture models. *Sankhya: The Indian Journal of Statistics, Series A*, 62(1):49–66, 2000.

A. Khalili and J. Chen. Variable selection in finite mixture of regression models. *Journal of the American Statistical Association*, 102(479):1025–1038, 2007.

A. Khalili and S. Lin. Regularization in finite mixture of regression models with diverging number of parameters. *Biometrics*, 69(2):436–446, 2013.

L.R. Lloyd-Jones, H.D. Nguyen, and G. J. McLachlan. A globally convergent algorithm for lasso-penalized mixture of linear regression models. *Computational Statistics & Data Analysis*, 119:19–38, 2018.

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

J. Ross and J. Dy. Nonparametric mixture of gaussian processes with constraints. *Proc. 30th Int. Conf. Mach. Learn.*, 28:1346–1354, 2013.

G. Schwarz. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

N. Städler, P. Bühlmann, and S. van de Geer. ℓ1-penalization for mixture regression models. *TEST*, 19(2): 209–256, Aug 2010.

R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58:267–288, 1994.

H. Wang, R. Li, and C.-L. Tsai. Tuning parameter selectors for the smoothly clipped absolute deviation method. *Biometrika*, 94(3):553–568, 2007.

Z. Younossi, Q.M. Anstee, M. Marietti, T. Hardy, L. Henry, M. Eslam, J. George, and E. Bugianesi. Global burden of nafld and nash: trends, predictions, risk factors and prevention. *Nature Reviews Gastroenterology & Hepatology*, 15:11–20, 2018a.

Z. Younossi, R. Loomba, Q. Anstee, M. Rinella, E. Bugianesi, G. Marchesini, B. Neuschwander-Tetri, L. Serfaty, F. Negro, S. Caldwell, V. Ratziu, K. Corey, S. Friedman, M. Abdelmalek, S. Harrison, A. Sanyal, J. Lavine, P. Mathurin, M. Charlton, Z. Goodman, N. Chalasani, K. Kowdley, J. George, and K. Lindor. Diagnostic modalities for nonalcoholic fatty liver disease, nonalcoholic steatohepatitis, and associated fibrosis. *Hepatology*, 68(1):349–360, 2018b.

S. E. Yuksel, J. N. Wilson, and P. D. Gader. Twenty years of mixture of experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193, 2012. .

9. Appendix

9.1 Parameters definition

We summarize the parameters chosen to simulate data in Table 6.

Address of the First, Third and Fourth author
Institut de recherche mathématique de Rennes
263 avenue du Général Leclerc
35000 Rennes, France
E-mail: marie.morvan@univ-rennes1.fr
E-mail: joyce.giacofci@univ-rennes2.fr
E-mail: valerie.monbet@univ-rennes1.fr

Address of the Second author
Laboratoire d'Informatique de Grenoble - LIG
Université Grenoble Alpes
700 Avenue centrale 38 400 Saint Martin d'Hères, France
E-mail: emilie.devijver@univ-grenoble-alpes.fr

	k	Same support - Easy case (1)	Same support - Difficult case (2)
	1	$(-1, 1, 0.5, 1, \mathbf{0}_{32}, -1, -1, 0.5, 0.2)$	$(-2, 1, 0.5, 1, \mathbf{0}_{32}, -1, -0.2, 0.5, 0.2)$
$\boldsymbol{\beta}$	2	$(1, 0.5, 0.5, -1, \mathbf{0}_{32}, 1, -0.2, 1, -0.5)$	$(1, -0.5, -0.5, -1, \mathbf{0}_{32}, 1, -0.2, 1, -0.5)$
	3	$(-1, 0.5, 1, -1, \mathbf{0}_{32}, 1, 1, 2, -1)$	$(-2, 0.5, -2, -1, \mathbf{0}_{32}, 1.5, 1, 2, -1)$
	1	$(-\mathbf{1}_{20}, \mathbf{1}_{20})$	$(-\mathbf{2}_{20}, \mathbf{1}_{20})$
$\boldsymbol{\mu}$	2	$\mathbf{1}_{40}$	$\mathbf{1}_{40}$
	3	$(\mathbf{1}_{20}, \mathbf{2}_{20})$	$\mathbf{3}_{40}$
	1	diag(0.5)	diag(2/3)
$\boldsymbol{\Sigma}$	2	band$(0.5, 0.2, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.5, 0.1, 0.1, \mathbf{0}_{36})$
	3	band$(0.8, 0.4, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.4, 0.1, 0.1, \mathbf{0}_{36})$
	k	Different supports - Easy case (3)	Different supports - Difficult case (4)
	1	$(-2, 1, 0.5, 1, \mathbf{0}_{32}, -1, -0.2, 0.5, 0.2)$	same as case 3
$\boldsymbol{\beta}$	2	$(\mathbf{0}_4, 1, -0.5, -0.5, -1, \mathbf{0}_{16}, 1, -0.2, 1, -0.5, \mathbf{0}_4)$	same as case 3
	3	$(\mathbf{0}_8, -2, 0.5, -2, -1, 1.5, 1, 2, -1, \mathbf{0}_{16})$	same as case 3
	1	$(-\mathbf{1}_{20}, \mathbf{1}_{20})$	$(\mathbf{1}_{20}, \mathbf{2}_{20})$
$\boldsymbol{\mu}$	2	$\mathbf{1}_{40}$	$\mathbf{1}_{40}$
	3	$(\mathbf{1}_{20}, \mathbf{2}_{20})$	$(\mathbf{1}_{20}, \mathbf{2}_{20})$
	1	diag(0.5)	diag(2/3)
$\boldsymbol{\Sigma}$	2	band$(0.8, 0.3, 0.1, 0.1, \mathbf{0}_{36})$	diag(2/3)
	3	band$(0.6, 0.3, 0.1, 0.1, \mathbf{0}_{36})$	band$(1, 0.4, 0.1, 0.1, \mathbf{0}_{36})$

TABLE 6

Summary of the parameters for the 4 cases. To define a model, $\boldsymbol{\beta}, \boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ have to be defined. As the mixture model has 3 clusters, there are 3 different parameters in each case (the cluster is represented with k). In Case 1 and Case 2, the relevant variables are the same within clusters. The differences are on the concentration of the clusters and the balance between the two classes in each cluster. In Case 3 and Case 4, the relevant variables are different within clusters. In Case 4, the clustering relies on $Y|\mathbf{X}$ whereas $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_3$ are the same, where in Case 3, the clustering relies on $Y|\mathbf{X}$ and $\mathbf{X}$. For the covariance matrices $\boldsymbol{\Sigma}$, we have used diagonal matrices and banded matrices.