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 this paper, a suitable and interpretable diagnosis statistical model is proposed to predict the Non-Alcoholic Steatohepatitis (NASH) from near infrared spectrometry data. In this disease, unknown patients profiles are expected to lead to different diagnosis. The model has then to take into account the heterogeneity of the data and the dimension of the spectrometric data. To this end, we propose to fit a mixture on the joint distribution of the diagnosis binary variable and the covariates selected in the spectra. Because of the high dimension of the data, a penalized maximum likelihood estimator is considered. In practice, a twofold penalty on both regression coefficients and covariance parameters is imposed. Automatic selection criteria such as the AIC and BIC are used to select the amount of shrinkage and the number of clusters. Performance of the overall procedure is evaluated through a simulation study and its application on the NASH data set is analysed. The model leads to higher prediction performance than competitive methods and provides highly interpretable results.

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 build-up of fat in the liver. Due to the worlwide increase of obesity and type 2 diabetes, its prevalence, currently estimated at 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 subject to 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). Mid-infrared spectroscopy provides a molecular fingerprint of biological fluid, such as 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

‡. Institute of Engineering Univ. Grenoble Alpes

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

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

serum spectra of 395 morbidly obese patients (Anty et al., 2010), including 66 patients diagnosed as NASH. Diagnosis has been based on 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 raise 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 all the 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 appropriate methods to select the range of wavelengths that are 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 to 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) are 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 causes remain poorly known. Mixture models become then the appropriate tool by modelling the unknown sub-population through a discrete latent variable. 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 (Hoshikawa, 2013): posterior cluster membership probabilities indeed depends on the unobserved response variable and cannot be computed. As a matter of fact, Grün and Leisch (2007) examine model fitting strategies for finite mixture of generalized linear regressions but do not discuss the prediction issue. Several works examine alternative strategies to allow computation of posterior probabilities as in Misiti et al. (2015), Ahonen et al. (2019) or Bougeard et al. (2018) but require to use a separate learning algorithm to do so. Concur-

rently, the machine learning community developed mixture of experts (MoE) models. Such models are dedicated to prediction when dealing with heterogeneous regions in the input space. Based on the principle of divide-and-conqueer, MOE 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 weights depending on the covariates. Classical mixing weights include exponential weights with the softmax gating network or more general ones (Yuksel et al., 2012). As pointed by Ahonen et al. (2019) MOE by strictly partitioning the covariates space can not be considered as proper mixture models. 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-Maximization (EM) algorithm which is adapted to the latent variable setting.

As previously mentioned, spectrometry data contain all the 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.

First, the mixture of logistic regression model is presented in Section 2.1, and the prediction step is described thereby. Then, parameter estimation by maximization of the likelihood is described in Section 2.2 and in Section 2.3, regularization is introduced to reduce the dimension and to allow interpretation. Model selection is discussed in Section 2.4 to select the number of clusters. Section 3 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 4 illustrates our result on the data set concerning the NASH disease.

2. Penalized mixture of logistic regressions model with random design

2.1 Random design model specification and prediction

Let $(\mathbf{X}, Y) \in \mathbb{X} \times \{0,1\}$ be a pair of random variables with $\mathbb{X} \subset \mathbb{R}^p$, for $p \in \mathbb{N}$. The variable Y stands for a binary dependent variable whereas $\mathbf{X}$ is a random vector of p quantitative covariates. In a learning context, one seek to describe the conditional relationship $f(y|\mathbf{X} = \mathbf{x})$ between the response Y and the predictors $\mathbf{X}$. We assume this relationship depends on a latent (unobserved) class variable $\mathbf{Z} = (Z_1, \ldots, Z_k)$, $K \in \mathbb{N}$, which follows a multinomial distribution such that

$$\mathbf{Z} \sim \mathcal{M}(1, \pi_1, \cdots, \pi_K),$$

where $\pi_k = P(Z_k = 1)$ for $k = 1, \ldots, K$ with $\pi_k > 0$ and $\sum_{k=1}^{K} \pi_k = 1$. Conditionally on $\mathbf{Z}$, we define a logistic regression model with random design hierarchically as

$$\begin{aligned} &\mathbf{X}|\{Z_k = 1\} \sim \mathcal{N}_p(\boldsymbol{\mu}_k, \Sigma_k), \\ &Y|\{\mathbf{X} = \mathbf{x}, Z_k = 1\} \sim \mathcal{B}\left(p^{(k)}(\mathbf{x})\right). \end{aligned} \tag{1}$$

The notations $\mathcal{N}_p$, $\mathcal{B}$ denote respectively the Gaussian and binomial distributions. Parameters $\boldsymbol{\mu}_k \in \mathbb{R}^p$ and $\Sigma_k \in \mathbb{R}^{p \times p}$ are respectively the mean and the covariance matrix of the random covariates $\mathbf{X}$ conditionally on $\{Z_k = 1\}$. Moreover, given $\{Z_k = 1\}$, the covariates $\mathbf{X}$ are related to the response variable Y through the logistic link function such as

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

where $\text{logit} : x \mapsto \log(x/(1-x))$ and $\boldsymbol{\beta}_k = (\beta_{k,1}, \ldots, \beta_{k,p})$ are the regression coefficients of the generalized linear model in cluster k.
The marginal joint distribution of $(\mathbf{X}, Y)$ (unconditional on $\mathbf{Z}$) is then defined as a mixture of logistic regression with a random design. In particular, the joint density can be written as

$$f_{Y,\mathbf{X}}(y, \mathbf{x}) = \sum_{k=1}^{K} \pi_k f_{Y|\{\mathbf{X},\mathbf{Z}\}}(y; \boldsymbol{\beta}_k) f_{\mathbf{X}|\mathbf{Z}}(\mathbf{x}; \boldsymbol{\mu}_k, \Sigma_k). \tag{2}$$

In the sequel, the set of parameters for the mixture density (2) with K clusters will be denoted by $\Phi_K = (\pi_1, ..., \pi_K, \boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_K)$, where $\boldsymbol{\phi}_k = (\boldsymbol{\mu}_k, \Sigma_k, \boldsymbol{\beta}_k)$, $k = 1, \ldots, K$, is the vector of parameters of the k-th cluster.
The originality of our model compared to traditional finite mixture of regressions model usually explored in the literature (see *e.g* Grün and Leisch, 2007; Khalili and Chen, 2007) is the random design framework. It has two important consequences. First, as stated in Model (1), the random covariates are distributed themselves as a mixture and hence are also impacted by the unknown grouping structure. As pointed out by Hoshikawa (2013), it is then possible to infer cluster membership probabilities conditionally on the values observed for the covariates. Besides, the random design framework allows to consider prediction of the response for out-of-sample observations. Regression based on a fixed design indeed suffer from an optimism excess as emphasized in Rosset and Tibshirani (2019) and hence should be used for prediction only with in-sample observations.

Formally, for a given out-of-sample observation $\mathbf{x}_0$ drawn from Model (1), the prediction rule is given by,

$$\mathbb{E}(Y_0|\mathbf{X}_0=\mathbf{x}_0)=\sum_{k=1}^{K}\mathbb{P}(Y_0=1|Z_{0k}=1,\mathbf{X}_0=\mathbf{x}_0)\mathbb{P}(Z_{0k}=1|\mathbf{X}_0=\mathbf{x}_0), \tag{3}$$

with $\mathbf{Z}_0=(Z_{01},...,Z_{0K})$ being the latent random variable associated to the new observation $\mathbf{X}_0$.
Replacing both quantities with their expressions in the joint mixture of logistic regressions framework, one obtains a predicted value

$$\widehat{Y_0}=\mathbb{E}(Y_0|\mathbf{X}_0=\mathbf{x}_0)=\sum_{k=1}^{K}\tau'_{0,k}\frac{\exp(\mathbf{x}_0^t\boldsymbol{\beta}_k)}{1+\exp(\mathbf{x}_0^t\boldsymbol{\beta}_k)},$$

where

$$\tau'_{0,k}=\mathbb{P}(Z_{0k}=1|\mathbf{X}_0=\mathbf{x}_0)=\frac{\pi_k f_{\mathcal{N}}(\mathbf{x}_0;\boldsymbol{\mu}_k,\Sigma_k)}{\sum_{\ell=1}^{K}\pi_\ell f_{\mathcal{N}}(\mathbf{x}_0;\boldsymbol{\mu}_\ell,\Sigma_\ell)}, \tag{4}$$

with $f_{\mathcal{N}}(\,\cdot\,;\boldsymbol{\mu},\Sigma)$ being the multivariate Gaussian density with mean $\boldsymbol{\mu}$ and covariance $\boldsymbol{\Sigma}$.
Model (1) can be compared to mixture of experts models (MoE), initially introduced in Jacobs et al. (1991) and reviewed in Yuksel et al. (2012). Such class of models are also dedicated for prediction in the mixture context. The mixing proportions depend on the covariates through a link function which is chosen according to the application. The outcome variable distribution depends then on both the covariates $\mathbf{X}$ and on the latent cluster membership variable $\mathbf{Z}$. In our setting, the cluster variable $\mathbf{Z}$ is assumed to follow a multinomial distribution as prior distribution, inducing posterior probabilities as weights depending on the covariates, and in that sense, can be viewed as some particular case of mixture of experts using Gaussian parametric forms in the gate (Yuksel et al., 2012).

2.2 Estimation using penalized likelihood

Given a sample $\{(\mathbf{x}_i,y_i)_{i=1,...,n}\}$ of n independent realizations of the random variables $(\mathbf{X},Y)$, the unknown parameters $\Phi_K=(\pi_1,...,\pi_K,\boldsymbol{\phi}_1,...,\boldsymbol{\phi}_K)$ are estimated by maximum likelihood. Our work takes place in a moderate dimension context for the covariates $\mathbf{X}$. However, even under this assumption, some variables may not be relevant for predicting the response variable Y and finding a consistent estimation of a full unstructured covariance matrix remains challenging. A penalized likelihood approach is therefore adopted with a double penalty. The first penalty is a Lasso penalty (Tibshirani, 1994) dedicated to perform clusterwise selection of features relevant for the logistic regressions in the model. The second part of the penalization controls the estimation of the covariates covariance matrices involved in the clustering part of the model. The maximum likelihood estimator of covariance matrices is indeed known to poorly behave even in a moderate dimension. A Graphical Lasso penalty (Friedman et al., 2008) is thus used to estimate covariance matrices, under the assumption that their inverses are sparse. Besides, this assumption seems reasonable for the real data studied in the sequel.

The penalized log-likelihood is finally defined, for $\lambda_k \geq 0, \rho_k \geq 0$, for all $k = 1, \ldots, K$, as

$$\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \Phi_K) - \sum_{k=1}^{K} \lambda_k \|\boldsymbol{\beta}_k\|_1 - \sum_{k=1}^{K} \rho_k \|\Theta_k\|_1, \tag{5}$$

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 the elements of Θ_k. The regularization constants λ_k and ρ_k drive the amount of shrinkage on the parameters $\boldsymbol{\beta}_k$ and Θ_k for every cluster k, $k = 1, \ldots, K$. Their selection is described in Section 2.3.

Penalized maximum likelihood estimation consists in maximizing the convex function (5) with respect to parameters Φ_K. In a latent variable framework, this objective is usually achieved through an Expectation-Maximisation (EM) algorithm (Dempster et al., 1977), which alternates between two steps, E-step and M-step. Convergence properties of the EM algorithm for penalized likelihood are studied in Green (1990). At iteration $[h]$, the E-step of the algorithm resumes to compute, for all individuals $i = 1, \ldots, n$, the cluster membership posterior probabilities $\tau_{ik}^{[h]}$ given the current parameters value $\Phi_K^{[h-1]}$. The posterior probabilities are computed thanks to

$$\tau_{ik}^{[h]} = \mathbb{P}(Z_{ik} = 1 | \mathbf{X}_i = \mathbf{x}_i, Y_i = y_i) = \frac{\pi_k f_{\mathcal{B}}(y_i; p_{\boldsymbol{\beta}_k}(\mathbf{x}_i)) f_{\mathcal{N}}(\mathbf{x}_i; \boldsymbol{\mu}_k, \Sigma_k)}{\sum_{\ell=1}^{K} \pi_\ell f_{\mathcal{B}}(y_i; p_{\boldsymbol{\beta}_\ell}(\mathbf{x}_i)) f_{\mathcal{N}}(\mathbf{x}_0; \boldsymbol{\mu}_\ell, \Sigma_\ell)}, \tag{6}$$

where $f_{\mathcal{B}}(\cdot; p_{\boldsymbol{\beta}_k}(\mathbf{x}_i))$ denotes the density of the Bernoulli distribution with parameters $p_{\boldsymbol{\beta}_k}(\mathbf{x}_i) = \exp(\mathbf{x}_i^t \boldsymbol{\beta}_k)/(1 + \exp(\mathbf{x}_i^t \boldsymbol{\beta}_k))$.
Then, in the M-step, the conditional expectation of the penalized likelihood given the current parameters value is maximized. It leads to the following updates of every parameter at iteration $[h]$, for all $k = 1, \ldots, K$:

$$\begin{aligned}
\widehat{\pi}_k^{[h]} &= \frac{1}{n} \sum_{i=1}^{n} \tau_{ik}^{[h]}, \\
\widehat{\boldsymbol{\mu}}_k^{[h]} &= \left[\sum_{i=1}^{n} \tau_{ik}^{[h]} \right]^{-1} \sum_{i=1}^{n} \tau_{ik}^{[h]} \mathbf{x}_i, \\
\hat{\Theta}_k^{[h]} &= \arg\max_{\Theta_k} \left\{ \sum_{i=1}^{n} \tau_{ik}^{[h]} \left(\log\det \Theta_k - \frac{1}{2} \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h]} \right)^T \Theta_k \left(\mathbf{x}_i - \widehat{\boldsymbol{\mu}}_k^{[h]} \right) \right) - \rho_k \|\Theta_k\|_1 \right\}, \\
\hat{\boldsymbol{\beta}}_k^{[h]} &= \arg\max_{\boldsymbol{\beta}_k} \left[\sum_{i=1}^{n} \tau_{ik}^{[h]} \left(y_i \mathbf{x}_i^T \boldsymbol{\beta}_k - \ln\left(1 + \exp(\mathbf{x}_i^T \boldsymbol{\beta}_k)\right) \right) \right] - \lambda_k \|\boldsymbol{\beta}_k\|_1.
\end{aligned}$$

Few iterations of a Newton-Raphson algorithm are run to compute $\hat{\Theta}_k^{[h]}$ and $\hat{\boldsymbol{\beta}}_k^{[h]}$. Note that it is not necessary to let the optimization run until convergence since an improvement of the penalized likelihood at each iteration is enough to guarantee the convergence of the EM algorithm to a local maxima (Meng and Rubin, 1993).

The convergence of the EM algorithm to the global maxima is known to highly depend on the initial parameters. Here, we adopt a Search/Run/Select (S/R/S) strategy as developed in Biernacki et al. (2003) and we perform the initialization in 3 steps:

1. Find t initial positions of parameters: obtain a partition of the set of observations of the explanatory variables $(\mathbf{x}_i)_{i=1,...,n}$ 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.

2. Run a small fixed number of iterations of the EM algorithm at the t initial positions previously found.

3. Among these t possible starting values, select the one which maximizes the log-likelihood to start the EM algorithm.

2.3 Tuning the regularization parameters

The tuning parameters $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_K)$ and $\boldsymbol{\rho} = (\rho_1, ..., \rho_K)$ are related to the amount of regularization, and their selection is a critical issue in a penalized likelihood approach. It is usually based on a trade-off between bias and variance: 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. 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). However, this method is computationally heavy and Wang et al. (2007) show that Generalized Cross-Validation may lead to the selection of some irrelevant variables. Besides, different studies (for example Wang et al., 2007; Khalili and Lin, 2013; Jiang et al., 2018; Lloyd-Jones et al., 2018) suggest the use of the Bayesian Information Criterion (BIC) to that purpose. The BIC (Schwarz, 1978) is constructed to realise a trade-off between the number of free parameters and the quality of the model measured by the likelihood. For a given number of clusters K, the BIC is defined as

$$BIC^{(\boldsymbol{\lambda},\boldsymbol{\rho})} = -2 \ln \mathcal{L}\left(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}\right) + \nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})} \ln(n),$$

with $\hat{\Phi}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ the arguments of the maxima of the penalized log-likelihood function with tuning parameters $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$. The quantity $\nu^{(\boldsymbol{\lambda},\boldsymbol{\rho})}$ counts the number of free parameters of the model, corresponding to the number of non-zero coefficients of the model. As tuning parameters λ_k and ρ_k have to be chosen for each cluster $k = 1, \ldots, K$, an automatic procedure is proposed. First a classification by MAP rule is derived after few iterations of the EM algorithm, and grids for $\boldsymbol{\lambda}$ and $\boldsymbol{\rho}$ are deduced for each cluster k. Parameters are then estimated for each combination of the possible values of the two dimensional grid and finally, the values of parameters minimizing the BIC are retained.

2.4 Selection of the number of clusters

The number of clusters K is a sensible parameter because it describes the heterogeneity of the population. In an unsupervised setting, K is unknown and thus has to be estimated as well. Besides variable selection, the BIC is also commonly used to determine the number of clusters K in a mixture models framework (Keribin, 2000). For a given number of clusters

K, the BIC is defined as

$$BIC_K = -2\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + \nu_K \ln(n),$$

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) show that in some misspecified situations, it can lead to a wrong choice of K. For that reason, the Integrated Classification Likelihood (ICL) criterion, better adapted to the model-based clustering framework, was developed (see Biernacki et al., 2000; McLachlan and Peel, 2000). An entropy term taking into account the concentration shape of the clusters is added to the usual BIC. For a model estimated with K clusters, the ICL is defined as

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

with τ_{ik} being the posterior probabilities. 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. In other words, the ICL criterion favors models with well separated clusters.
Finally, for predictive purpose, the Akaike Information Criterion (AIC) is known to be more suitable (Shmueli, 2010). With the previous notations, the AIC is defined as

$$AIC_K = -2\ln \mathcal{L}(y_1, \ldots, y_n, \mathbf{x}_1, \ldots, \mathbf{x}_n; \hat{\hat{\Phi}}_K^{(\boldsymbol{\lambda},\boldsymbol{\rho})}) + 2\nu_K.$$

Those three criteria will be compared in the simulation study.

3. Simulation study

Performance of our procedure are assessed through a simulation study. In particular, the study focuses on the model selection properties and the ability of the model to correctly predict a diagnostic, *i.e* the response variable.

3.1 Synthetic data generation

Simulated data are randomly generated according to a mixture of $K = 3$ clusters with proportions $\boldsymbol{\pi} = (0.3, 0.3, 0.4)$. The dimension of the predictors is set to $p = 40$ variables with only 8 variables being relevant in each cluster. Four different cases described in details in Appendix 7 are investigated. As a brief summary, the first two cases (Case 1 and 2) consider the same support for the regression coefficients in every cluster, *i.e.* the non-zero coefficients are associated with the same variables, but the concentration in each cluster differs and the parameters of the regression models lead to unequal class balances with respect to the cluster. The clustering task in Case 1 can be considered to be easier than the one in Case 2 because the ratio of means difference over the within variance (also called signal to noise ratio) is larger. The two last cases (Case 3 and 4) consider different supports for the regression coefficients, *i.e.* the non-zero coefficients are associated with different variables

depending on the cluster. In Case 4, an extra difficulty is further added: the cluster means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_3$ are assumed to be equal so that information about clusters is carried by the conditional distributions $Y|\mathbf{X}, Z$, in other words, by the values of the regression parameters. For each model, the impact of the sample size is measured, n being respectively equal to 250 and 500. These values are chosen because they are plausible for the applications in mind. Each experiment is repeated 30 times and majority votes or confidence intervals are computed.

3.2 Model selection

The proposed penalized mixture of logistic regression model, referred to as PMLR, is compared to a finite mixture of logistic regression model introduced in Grün and Leisch (2007) and denoted MLR. Note that in MLR, the clustering structure is only driven by the conditional distribution of $Y|\mathbf{X}$ so that the posterior distribution of the latent class variable can not be defined and the model can not be used for any prediction task.

Number of clusters Let us first consider the selection of the number of clusters. As stated earlier, several information criteria are compared. Table 1 reports the majority vote of each of them for the PMLR and the MLR models. For PMLR, the AIC leads, whatever of the size of the dataset, to the selection of 3 clusters which is the right number. For Case 4, where two of the three clusters only differ through the coefficients of the regression model, the criteria fails and selects models with only 2 clusters. The penalty terms of BIC and ICL criteria are stronger than the one of the AIC and they tend to under estimate the number of clusters. These trend is more severe for the MLR than for PLMR. Finally, we decide to use AIC for the selection of K.

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 quality of the clustering can be further evaluated using the Ajusted Rank Index (see Vinh et al. (2010)) which is abbreviated ARI and which measures the similarity between the simulated and the estimated clusterings. This index is bounded by 1 and the greater the better. The mean of ARI values computed for the models selected by the AIC are reported in Table 2. They are close to 1 for the PLRM models except for the difficult Case 4, where two clusters (among the three) are close. For the MLR model all the values are close to 0 and it is mainly due to the bad selection of the number of clusters. But it is also explained by the lack of information brought by the model about the cluster. Indeed, in MLR model,

the clusters are only defined by the coefficients of the regression models while in PMLR the covariates are also involved in the clustering structure.

Method	PMLR		MLR	
Sample size n	250	500	250	500
Same support				
Easy case (1)	0.89	0.91	0	0.01
Difficult case (2)	1	1	0.01	0.01
Different supports				
Easy case (3)	0.96	0.97	0.01	0.01
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.

Variable selection In order to evaluate the ability of the penalization method to select the relevant predictors, relevant variable detection (RVD) and irrelevant variable elimination (IVE) criteria[1] are used. The RVD associated to 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 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, whereas the IVE is equal to 1 if all the irrelevant variables are eliminated from the model. Then, a selection is good is both criteria are close to 1. However, in practice, a trade-off between both is often observed. The distributions of RDV and IVE estimated on the 30 repetitions of the simulation-estimation tasks are represented by boxplots in Figure 1. The regularization procedure leads to high values of the RVD and low values of the IVE for the PMLR and the opposite occurs for the MLR. It means that, in the PMLR, the effect of the penalization is conservative in the sense that it tends to select models keeping to much predictors. It is usually a good strategy for the prediction, where models slightly too rich generally perform better. At the opposite, in the MLR, the regularization eliminates too much variables including some of the important ones. Distributions of Figure 1 also illustrate that RVD significantly decreases when the clusters are difficult to identify (as in Case 4).

3.3 Prediction performances

To evaluate the quality of the prediction of the binary response, we consider the PMLR model for a fixed number of clusters $K = 3$. For a sake of comparison, results are also reported for the one cluster model referred to as PLR for Penalized Logistic Regression and the logistic regression (LR) model which corresponds to the model with one cluster and no

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

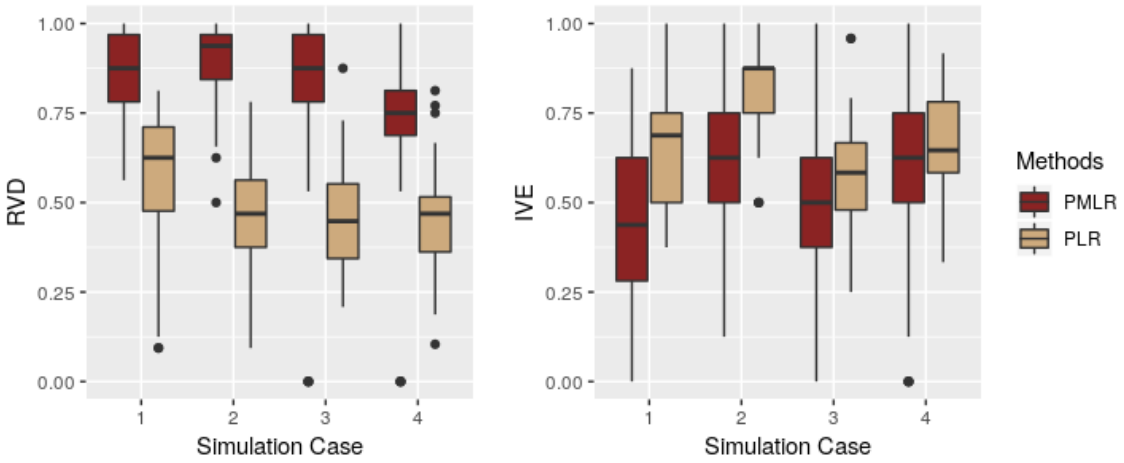

Figure 1: Performance in variable selection for $\hat{\boldsymbol{\beta}}$. We provide boxplots for the Relevant Variable Detection (RVD) on the left and boxplots for the Irrelevant Variable Elimination (IVE) on the right. We compare our method (PMLR) with the Penalized Logistic Regression (PLR), which 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.

penalization. For each of them, the Area Under the Receiver Operating Characteristic curve (AUROC) is computed from out of sample observations. AUROC is a common summary statistic for the goodness of a predictor in a binary classification task. It is equal to the probability that a predictor will rank a randomly chosen positive instance higher than a randomly chosen negative one.

The distributions of the AUROC are plotted in Figure 2. In Case 1, the classes of the binary variable are easy to discriminate so that all the models behave well although the PMLR model performs slightly better. In simulation cases 2 and 3 the population is highly heterogeneous and the regression models differ depending on the clusters. In theses cases, the PMLR model, which takes into account the heterogeneity, is significantly better than the two other models without any mixture. In order to go into more details, Received Operating Characteristic curves are plotted for Cases 2 and 3. The curves corresponding to the 30 estimations-predictions of the PMLR model are almost all above the ones of the PLR and LR models highlighting better predictions. It is interesting to note that the curves of the PLR and LR models are not very different indicating that over-learning in the LR model does not deteriorate significantly the prediction. The identification of the heterogeneity is clearly more important than the selection of the relevant predictor for the prediction task. Finally in simulation Case 4, where two clusters have the same mean, all the models have similar poor performances. In general, the prediction becomes better according to the AUROC criteria if the model is easier to calibrate. This is illustrated by the decreasing of the median AUROC from simulation Case 1 to simulation Case 4.

The conclusions of the simulation study are that the proposed model selection procedures are efficient if the clusters are distinct enough. Furthermore, the introduction of the latent class

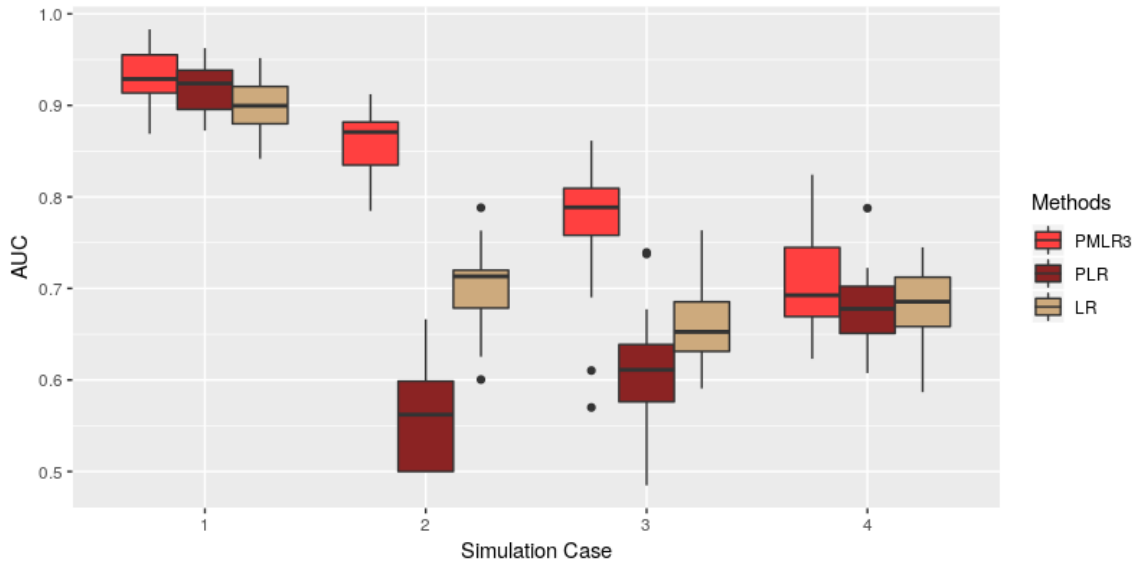

Figure 2: Prediction performance. We provide boxplots for Area Under the Receiver Operating Characteristic (AUROC) obtained for the 4 simulation cases for each method: Logistic Regression (LR), Penalized Logistic Regression (PLR) and Penalized Mixture of Logistic Regression with 3 clusters (PMLR-3).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

variable is critical for the prediction of the response variable. Finally a sample size of 250 seems to be large enough to get reasonable estimations. Bias and variance of estimators of $\hat{\boldsymbol{\mu}}$, $\hat{\boldsymbol{\Sigma}}$ and $\hat{\boldsymbol{\beta}}$ have been computed (not shown here). The bias is close to zero and the variances decreased with the sample size, as expected for likelihood estimates, and are already small enough for $n = 250$.

4. 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 get by our procedure PMLR, from estimation, interpretation and biological points of view.

4.1 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

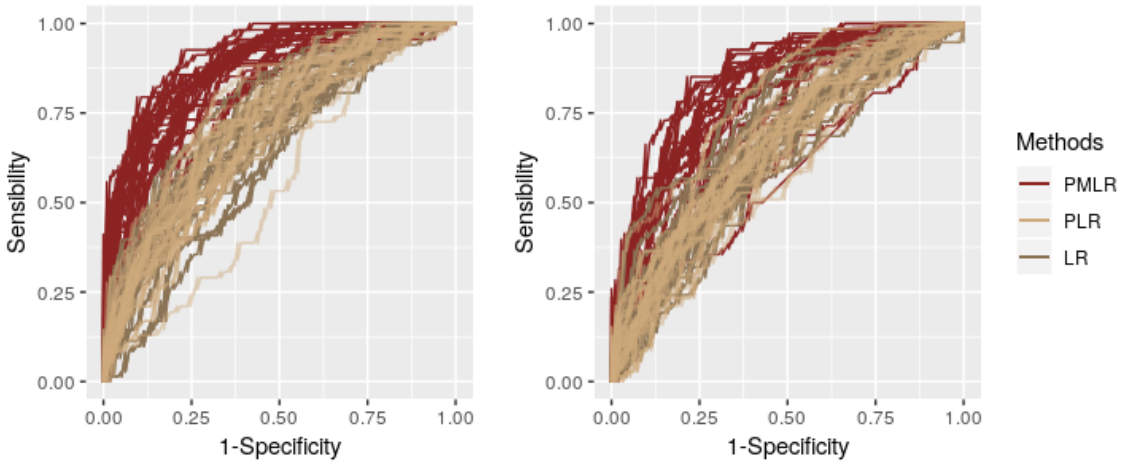

Figure 3: Prediction performance. We provide Receiver Operating Characteristic curves (ROC) for each method: Penalized Mixture of Logistic Regression (PMLR) with 3 clusters, Penalized Logistic Regression (PLR), Logistic Regression (LR). Plot on the left corresponds to the case 2, whereas plot on the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

of the patient's metabolism. Currently, the diagnosis is obtained after a liver biopsy and an histological study of the sample, which is an invasive method. The method we propose is based on spectrum measured on blood serum, thus 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.

4.2 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 K =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 3.2 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.

Statistical interpretation of the constructed model The proportions of each cluster are 0.66 and 0.34. The proportion of diseased patients changes according to the cluster: 19 % in cluster 1 and 12 % in cluster 2.

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

TABLE 4

Comparison of the prediction performance obtained with different methods: our method with 2 to 3 clusters (PMLR-2, PMLR-3), penalized logistic regression (PLR) and logistic regression (LR). The chosen model is PMLR-2. The bold values indicate the best values. 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).

Remark that the model PMLR with one cluster is equivalent to PLR, so performance is summarized by PLR in Table 4.

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 methods, 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. Distribution of the predicted scores according to the real class of the individuals from the validation set is represented by Figure 4.

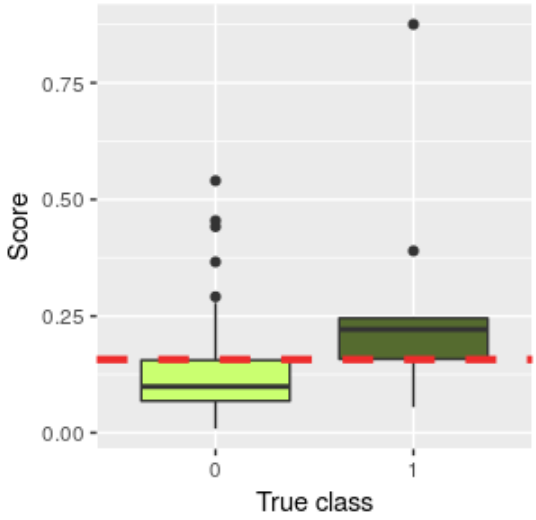

Figure 4: Performance in prediction. We provide boxplots of the scores with respect to the true class, for the Penalized Mixture of Logistic Regression with 3 clusters (PMLR). The red dashed line corresponds to the threshold automatically learned.

The threshold from which a patient is labelled as a NASH patient is automatically chosen as the value in $[0, 1]$ maximizing the sum of the sensibility and specificity. This threshold is represented in Figure 4 in red dashed line and allows a good separation between NASH patients and control.

Estimators and models Graphical models obtained from the sparsely estimated precision matrices for each cluster are represented in Figure 5. 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, a group of variables from X2 to X11 is densely connected. For the second cluster, we observe two groups of strongly linked variables: the first pool is concerned with the variables X2, X3, X4, X6, X7, X8, and the second with the variables X1, X5, X12, X14, X17, X19. The links between variables strongly differ according to the cluster. The node color represents the coefficient value of the variable for the model applied to the considered cluster. In the first cluster, the estimated model is very sparse, a large proportion of regression coefficients are close or equal to zero. In the second cluster, regression coefficients take more extreme values. The difference observed in the graphical structures and regression models for each cluster suggest different metabolic mechanisms of the patients, depending on the cluster.

Biological interpretation of the constructed model We characterize the clusters obtained with the selected model through the clinical variables available in Table 5. Val-

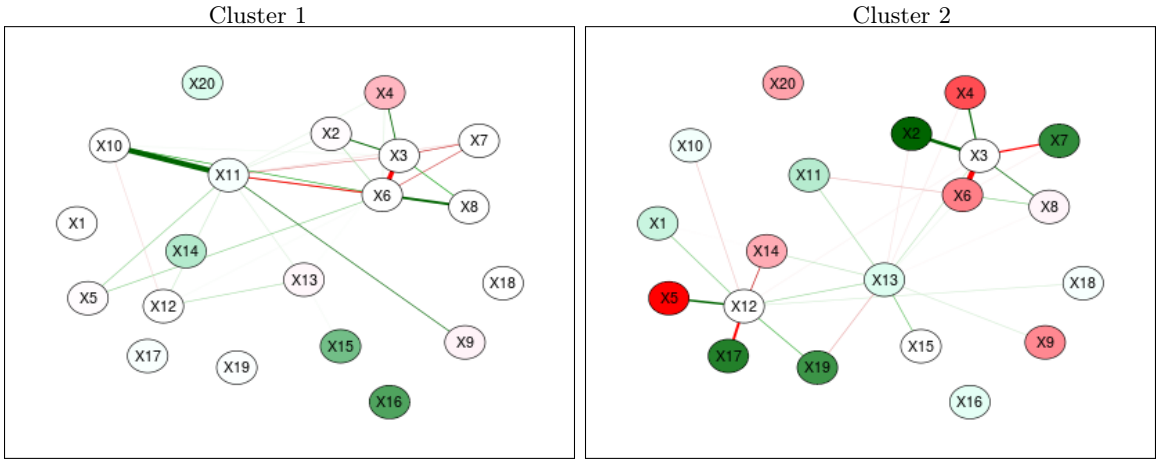

Figure 5: Graphical models. The selected model has 2 clusters, the network related to the first cluster is on the left and the network related to the second cluster is on the right. Precision matrices are sparsely estimated, so the resulting networks are sparse. Arrow colors correspond to the sign of the partial correlation (green for positive correlation, red for negative correlation) and the edge widths correspond to the value of the correlation (stronger is the color, stronger is the absolute correlation). Node color is associated with the value of the regression coefficient for each variable and cluster: strongest the color, larger is the value (green for positive values and red for negatives ones). No color indicates a value of zero for the regression coefficient (irrelevant variable).

	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 nominal variable Sex, the women's rate is indicated, and a Fisher test is used for the comparison.

ues 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. We observe 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 associated 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 seem to have more severe liver complications 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 increases with the steatosis, ballooning and inflammation stage, indicators used to establish the diagnosis. Fibrosis does not enter in the NASH definition and thus is not predicted by our model.

5. Conclusion

In this paper we have presented a predictive method that allows to build a model on data structured in unknown 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 suggested by a real data problem concerning the use of the spectrometric technology to develop a non-invasive diagnosis tool for the prediction of the NASH disease. We obtained very 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 much broader used.
In this reported work, the analysis focused on a preselection 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 automatic 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.

6. 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

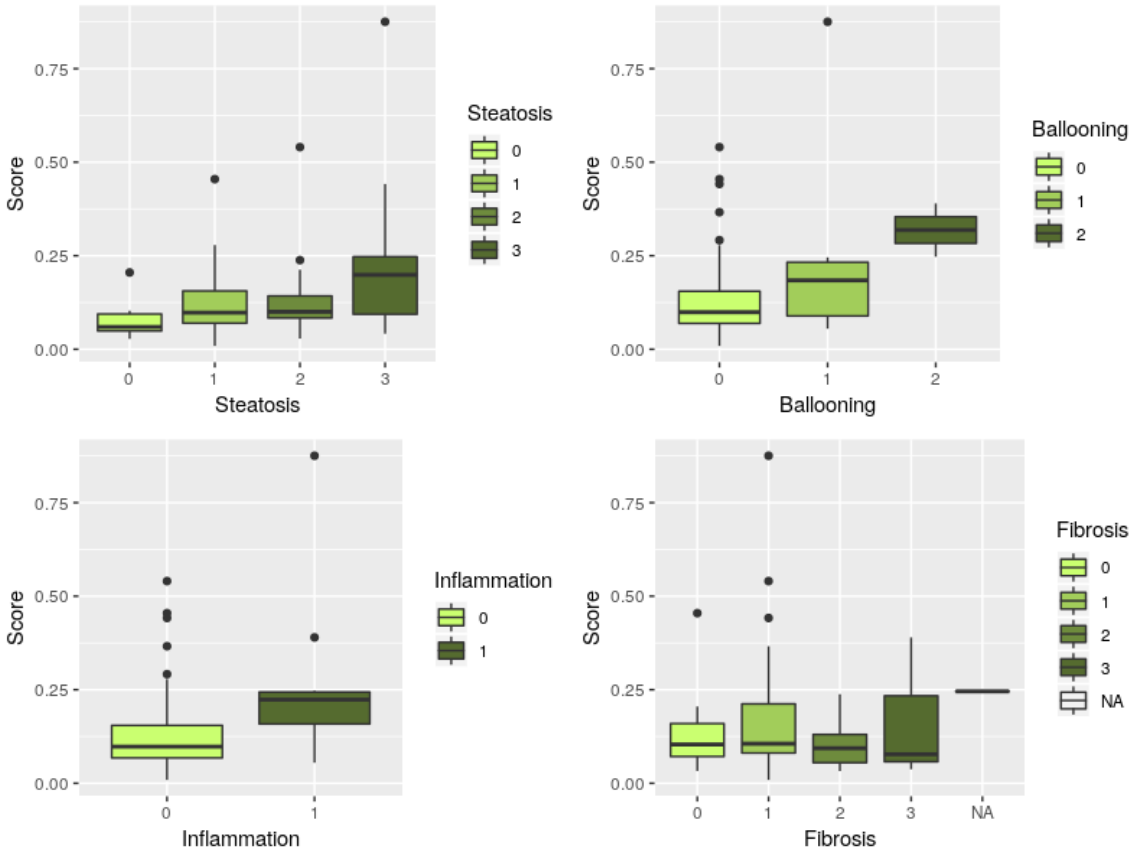

Figure 6: Boxplots of the repartition of the score predicted with the selected model according to the stage of the histological variable considered. Top left: steatosis, top rigth: ballooning, bottom left: inflammation, bottom right: fibrosis.

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

I. Ahonen, J. Nevalainen, and D. Larocque. Prediction with a flexible finite mixture-of-regressions. *Computational Statistics & Data Analysis*, 132:212 – 224, 2019. Special Issue on Biostatistics.

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.

S. Bougeard, H. Abdi, G. Saporta, and N. Niang. Clusterwise analysis for multiblock component methods. *Advances in Data Analysis and Classification*, 12(2):285–313, Jun 2018.

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.

P. J. Green. On use of the em algorithm for penalized likelihood estimation. *Journal of the Royal Statistical Society: Series B (Methodological)*, 52(3):443–452, 1990.

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

T. Hoshikawa. Mixture regression for observational data, with application to functional regression models. Available: http://arxiv.org/abs/1307.0170, 2013.

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.

J. Macqueen. Some methods for classification and analysis of multivariate observations. In *In 5-th Berkeley Symposium on Mathematical Statistics and Probability*, pages 281–297, 1967.

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

X-L. Meng and D.B. Rubin. Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, 80(2):267–278, 06 1993.

M. Misiti, Y. Misiti, J.-M. Poggi, and B. Portier. Mixture of linear regression models for short term pm10 forecasting in haute normandie (france). *CS-BIGS*, 6(1):47–60, 2015.

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

S. Rosset and R.J. Tibshirani. From fixed-x to random-x regression: Bias-variance decompositions, covariance penalties, and prediction error estimation. *Journal of the American Statistical Association*, 0(0):1–14, 2019.

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.

N. X. Vinh, J. Epps, and J. Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11(Oct): 2837–2854, 2010.

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. .

7. Appendix: parameters definition

We summarize the parameters chosen to generate synthetic datasets in Table 6.

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

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.

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