Many experiments in medicine and ecology can be conveniently modelled by finite Gaussian mixtures but face the problem of dealing with small data sets. We propose a robust version of the estimator based on self-regression and sparsity promoting penalization in order to estimate the components of Gaussian mixtures in such contexts. A space alternating version of the penalized EM algorithm is obtained and we prove that its cluster points satisfy the Karush-Kuhn-Tucker conditions. Monte Carlo experiments are presented in order to compare the results obtained by our method and by standard maximum likelihood estimation. In particular, our estimator is seen to perform well better than the maximum likelihood estimator.

Key words: finite Gaussian mixtures, maximum likelihood estimation, Kullback Proximal Point algorithms, EM algorithm, l1 penalization, LASSO, sparsity, regression mixtures, model based clustering

Finite Gaussian mixture models are widely used in a great number of application fields as a means to perform model based classification. From pattern recognition to biology, from quality control to finance, many examples have shown the pertinence of the Gaussian mixture model approach. The book [13] is the most comprehensive reference for finite non necessarily Gaussian mixture models with many application examples. In Gaussian mixture models, the data $Y_1, \ldots, Y_n$ are assumed i.i.d. and to be drawn from the density

$$\sum_{k=1}^{K} p_K^* \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k^*)}} \exp\Big(-\frac{1}{2}(x-\mu_k^*)\Sigma_k^{*-1}(x-\mu_k^*)\Big) \qquad (1)$$

where the vector $\theta^* = (p_1^*, \ldots, p_K^*, \mu_1^*, \ldots, \mu_K^*, \Sigma_1^*, \ldots, \Sigma_k^*)$ is an unknown multidimensional parameter. To this model, we traditionally associate an extended model using the notion of complete data. In mixture models, the complete data are independent and identically distributed couples of the form (Y_i, Z_i)

where Z_i is a multinomial random variable taking values in $\{1,\dots,K\}$ with $P(Z_i = k) = \pi_k^*$ and which represents the index of the mixture component from which observation i was drawn. We assume that conditionally on the event $Z_i = k$, Y_i has density $\frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k^*)}} \exp\Big(-\frac{1}{2}(x-\mu_k^*)\Sigma_k^{*-1}(x-\mu_k^*)\Big)$. The variables $Z_1,\dots,Z_n$ being unobserved, they are usually called latent variables.

The standard approach for estimating θ^* is the maximum likelihood methodology which consists of finding $\hat{\theta}$ which maximizes the log-likelihood function

$$l(\theta) = \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} p_K \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k)}} \exp\Big(-\frac{1}{2}(Y_i-\mu_k)\Sigma_k{}^{-1}(Y_i-\mu_k)\Big)\Big) \quad (2)$$

over the set $\Theta = \{(\pi_1,\dots,\pi_K,\mu_1,\dots,\mu_K,\Sigma_1,\dots,\Sigma_K) \mid \pi_k \in \mathbb{R}_+,\ \mu_k \in \mathbb{R}^d,\ \Sigma_k \in \mathbb{S}_d^+,\ \text{and } \sum_{k=1}^K \pi_k = 1\}$ where $\mathbb{S}_d^+$ denotes the set of all symmetric positive semidefinite matrices.

Interestingly enough, the supremum of the log-likelihood function over Θ is equal to $+\infty$ and is obtained for singular covariance matrices. A study of the one dimensional case was made in [2]. Thus, exact likelihood maximization is not the good approach for this problem. However, many researchers and practitioners have noticed that some local maximizer of the log-likelihood function is in fact consistent in practice. From the numerical viewpoint, local maximizers of the log-likelihood function are usually obtained using the Expectation-Maximization (EM) algorithm of Dempster Laird and Rudin [8]. This algorithm is a nice procedure with closed form expression of each iteration in the Gaussian mixture case. The EM algorithm for mixture models is available in the MIXMOD package [1] within Matlab or Scilab for instance.

Beside the question of finding the right local optimizer of the likelihood function, one of the main problems for estimating θ^* relies in having a sufficiently great sample size. Usually, large sample sizes may sometimes be available in a number of applications such as pattern recognition or financial time series analysis but in ecology for instance the sample size may be very small in situations where finite mixture models are suspected to be very pertinent due to the biological context. The goal of this paper is to remedy this problem by proposing a new methodology for Gaussian mixture model estimation in the case where the sample size is extremely small. Our approach needs to provide a certain amount of robustness. In the same spirit as for the median in the one dimensional case, the main idea is to express the estimators of the μ_k's as a combination of a small number of data in the middle of each cluster. This is simply done by restricting the search to the data's span, i.e. to obtain the μ_k's as a regression with covariates the data themselves and to impose an additional sparsity constraint on the regression vectors. In order to simplify the analysis, we will assume the covariance matrices to be of the form $\sigma_k^2 I$. The σ_k's and the π_k's can also be estimated using for instance a maximum

likelihood approach conditioned on the estimated value of the μ_k's.

The wole procedure is formally equivalent to joint variable selection and estimation in a mixture of regression model. Variable selection and estimation are performed using l_1-penalized EM steps which reduce the complexity of the regression model just as for the LASSO [14]. Encouraging simulations results show that the proposed approach correctly estimates the class of 8 over 10 points on average for a mixture of 3 Gaussians in dimension two. Monte carlo experiments are performed for samples sizes of 10 points and dimension growing up to to 50 showing a good behavior of the method which outperforms the standard maximum likelihood estimator.

1 Presentation of the method

1.1 Recalls on regression mixtures and model selection with nondifferentiable penalties

In order to introduce our method, we will need some preliminary notions on mixtures of regressions and the LASSO algorithm for sparse variable selection in regression models.

1.1.1 Regression mixtures

The Gaussian regression mixture assumes that the observations are couples of the form (Y, X) where X takes its values in $\mathbb{R}^p$ and conditionally on X, the real valued random variable Y follows the mixture density

$$f_{Y|X}(y) = \sum_{k=1}^{K} p_k \, \mathcal{N}(X\beta_k, \sigma_k^2). \tag{3}$$

Such mixture models are frequent in econometrics and chemometrics as described in the introduction of [11]. Estimation in these models can be performed using likelihood maximization as in [13] using the EM algorithm or a Bayesian methodology as studied in [11] using Monte Carlo Markov Chain techniques.

1.1.2 Variable and model selection with nondifferentiable penalties

An important problem in the study of regression mixtures is the one of variable selection. This problem has been recently addressed in [12] and several other works using various penalties based on nondifferentiable norm-like functions

of the regression vectors like the l_1 norm as in the LASSO [14] or SCAD introduced by [9].

The extension of the LASSO technique to the present situation consists of penalizing the log-likelihood like function l given by

$$\tilde{l}(\theta) = \sum_{i=1}^{n} \log\Big(\sum_{k=1}^{K} p_k \frac{1}{(2\pi)^d \det(\Sigma_k)} \exp\Big(-\frac{1}{2}(Y_i - X\beta_k)\Sigma_k{}^{-1}(Y_i - X\beta_k)\Big)\Big), \quad (4)$$

by the sum of the l_1-norms of the regression vectors in order to obtain the following estimator

$$\hat{\theta} \in \text{argmax}_{\theta\in\Theta} \tilde{l}(\theta) - \lambda \sum_{k=1}^{K} \|\beta_k\|_1 \quad (5)$$

where λ is a parameter to be adapted to the sample size. In the case of one regression with known variance σ^2 instead of a mixture, the recent theory of the LASSO prescribes to take λ proportional to $\sqrt{2\log(p)\sigma^2}$ even in the case where $n \ll p$; see for instance [3]. On the other hand, the paper [12] studies the case of proper mixtures with unknown variances but restricts the analysis to $n \gg p$ and p is constant as a function of n. A lot of work still remains to be done in the case where $n \ll p$ and the variance is unknown.

1.2 Our proposal: The mixture of self-regression with sparsity constraint

1.2.1 The estimator

In all what follows, we will assume that the data have been centered. Our proposal relies on the following simple idea: if the cluster probabilities π_k^*'s, the class indices Z_i's and the variances σ_k^*'s where known ahead of time, the estimators of the μ_k's could be chosen, in the small sample setting, as linear combinations of the data vectors themselves, in the same spirit as the median is preferable to the mean based on the theory of robust statistics. However, chosing the right sample vectors to estimate the mean seems a priori to be a very hard task. The main ingredient in our proposal is to force our estimator to be very sparsely represented as a linear combination of the data. Thus, the estimation of the mean vectors can be seen as a estimating a mixture of sparse regressions. The problem of estimating the right covariates being NP-hard in general, the estimation can be performed using with a penalization enforcing sparsity just as the l_1 penalty in the LASSO.

In the more general case where the indices Z_i are unobserved, and the cluster probabilities π_k^*'s and the variances σ_k^*'s are unknown, one can consider

maximizing the l_1-penalized log-likelihood like function given by

$$\sum_{i=1}^{n} \log \Big(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi\sigma_k^2)^d}} \exp\Big(- \frac{(Y_i - \mu_k)^t (Y_i - \mu_k)}{2\sigma_k} \Big) \Big) - \lambda_n \sum_{k=1}^{K} \|\beta_k\|_1 \quad (6)$$

under the data-driven constraints $\mu_k = Y\beta_k$ for $k = 1, \ldots, n$ where the matrix Y is given by $Y = [Y_1, \ldots, Y_n]$. In other words, we would like to maximize the l_1-penalized likelihood like function

$$\tilde{l}_{pen}(\theta) = \tilde{l}(\theta) - \lambda_n \sum_{k=1}^{K} \|\beta_k\|_1, \quad (7)$$

where

$$\tilde{l}(\theta) = \sum_{i=1}^{n} \log \Big(\sum_{k=1}^{K} \pi_k \frac{1}{\sqrt{(2\pi\sigma_k^2)^d}} \exp\Big(- \frac{(Y_i - Y\beta_k)^t (Y_i - Y\beta_k)}{2\sigma_k^2} \Big) \Big). \quad (8)$$

1.2.2 The Space-Alternating l1-EM algorithm

Optimizing the l_1-penalized function (6) can be performed using an EM-type algorithm. The Expectation Step consists of computing the conditional expectation of the complete l_1-penalized likelihood like function given the observations $Y_1, \ldots, Y_n$ where the distribution of the latent variables is taken to be their marginal density parametrized by the approximation $\bar{\theta}$ of the true parameter θ^*. The resulting quantity is traditionally denoted by $Q(\theta, \bar{\theta})$ and we will use the same notation in our l_1-penalized context. We will use the general form of covariance matrices instead of $\sigma_k^2 I$ in order to present a more general form of the method.

More precisely, the complete l_1-penalized log-likelihood like function $\tilde{l}^c_{pen}(\theta)$, i.e. the penalized log-likelihood like function of the complete data (Y_1, Z_1), $\ldots$, (Y_n, Z_n) is given by

$$\tilde{l}^c_{pen}(\theta) = \sum_{i=1}^{n} \log \Big(\pi_{Z_i} \frac{1}{\sqrt{(2\pi\sigma_k^2)^d}} \exp\Big(- \frac{(Y_i - Y\beta_k)^t (Y_i - Y\beta_k)}{2\sigma_k} \Big) \Big). \quad (9)$$

Thus, we obtain

$$Q(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \log \Big(\pi_k \frac{1}{\sqrt{(2\pi\sigma_k^2)^d}} \exp\Big(- \frac{(Y_i - Y\beta_k)^t (Y_i - Y\beta_k)}{2\sigma_k} \Big) \Big) \tau_{i,k} \quad (10)$$

where we used the standard notation $\tau_{i,k} = P_{\bar{\theta}}(Z_i = k \mid Y_1, \ldots, Y_n)$.

The Maximization Step consists of maximizing

$$Q(\theta,\bar{\theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K}\log\left(\pi_k\frac{1}{\sqrt{(2\pi\sigma_k^2)^d}}\exp\Big(-\frac{(Y_i-Y\beta_k)^t(Y_i-Y\beta_k)}{2\sigma_k^2}\Big)\right)\tau_{i,k} - \lambda\sum_{k=1}^{K}\|\beta_k\|_1. \tag{11}$$

In order to make the computational part easy, the π_k's, β_k's and σ_k's can be optimized alternatively in the manner of the Gauss-Seidel approach. In fact, the separability of the problem into two subproblems, the first being optimization over the π_k's and the second being optimization over the β_k's and σ_k's is already well known and the solution of the first subproblem is of the form

$$\pi_k = \frac{\sum_{i=1}^{n}\tau_{i,k}}{\sum_{i=1}^{n}\sum_{k=1}^{K}\tau_{i,k}}. \tag{12}$$

On the other hand, joint optimization in β_k's and the σ_k's is not separable and space alternating option is necessary to keep the computational complexity of each step at a low level. In order to address this problem, we need a generalization of the EM algorithm allowing for componentwise optimization at each step. Such penalized EM algorithms have been recently studied in the broader framework of Space Alternating Kullback Proximal Point Algorithms in [7]. Optimizing successively over the β_k's at one iteration and over the σ_k's at the next iteration should be sufficiently efficient in practice. Here, we will also optimize one cluster at a time in order to obtain the injectivity conditions which are needed in the theoretical analysis of the algorithm. A simple way to accelerate this extreme version of the Gauss-Seidel methodology would be to average the new iterates $\beta^{(l)}$ and $\sigma^{(l)}$ with the previous respective iterates in order to smooth the algorithm's trajectory.

2 Convergence analysis

When using a maximum likelihood approach, incorporation of a nondifferentiable penalty in the EM algorithm may cause some technical difficulties. A rigourous analysis has been proposed in [7] in the case of general nondifferentiable penalties and space alternating optimization versions of the EM algorithm. The convergence analysis is made easier after interpreting the EM algorithm as a Proximal Point Algorithm which was first done in [5] (see also [6] for more precise results).

In our special case, we only need to show that our Space-Alternating l1-EM is a Space-Alternating Kullback Proximal Point Algorithm of the form studied in [7] [1].

[1] for a definition of the Clarke subdifferential, see the Appendix of [7]

Algorithm 1 Space-Alternating l_1-EM algorithm

Input $L \in \mathbb{N}_*$

Choose intial iterate $\theta^{(0)} = (\pi_1^{(0)}, \ldots, \pi_K^{(0)}, \beta_1^{(0)}, \ldots, \beta_K^{(0)}, \sigma_1^{(0)}, \ldots, \sigma_K^{(0)})$

$l = 1$

while $l \leq L$ **do**

(E-Step) Compute the conditional probabilities $P_{\theta^{(l-1)}}(Z_i = k \mid Y)$ given the observations $Y_1, \ldots, Y_n$ for $i = 1, \ldots, n$ and $k = 1, \ldots, K$ using the following formula

$$\tau_{i,k}^{(l)} = \frac{\pi_k^{(l-1)} \frac{1}{\sqrt{(2\pi {\sigma_k^{(l-1)}}^2)^d}} \exp\Big(- \frac{(Y_i - Y\beta_k^{(l-1)})^t (Y_i - Y\beta_k^{(l-1)})}{2{\sigma_k^{(l-1)}}^2} \Big)}{\sum_{k=1^K} \pi_k^{(l-1)} \frac{1}{\sqrt{(2\pi {\sigma_k^{(l-1)}}^2)^d}} \exp\Big(- \frac{(Y_i - Y\beta_k^{(l-1)})^t (Y_i - Y\beta_k^{(l-1)})}{2{\sigma_k^{(l-1)}}^2} \Big)} \tag{13}$$

compute

–**either** the $\pi_k^{(l)}$'s by the formula

$$\pi_k^{(l)} = \frac{\sum_{i=1}^n \tau_{i,k}^{(l)}}{\sum_{i=1}^n \sum_{k=1}^K \tau_{i,k}^{(l)}} \tag{14}$$

–**or** $\beta_k^{(l)}$ as the solution of the LASSO-like optimization problem

$$\beta_k^{(l)} \in \operatorname{argmin}_{\beta \in \mathbb{R}^n} \|Y\tau_k^{(l)} - Y\beta\|_2^2 - \lambda \|\beta\|_1. \tag{15}$$

for the index k updated in cyclic order along iterations.

–**or** $\sigma_k^{(l)}$ using the formula

$$\sigma_k^{(l)} = \frac{1}{\sum_{i=1}^k \tau_{i,k}^{(l)}} \sum_{i=1}^n \|Y_i - Y\beta_k^{(l-1)}\|^2 \tau_{i,k}^{(l)}. \tag{16}$$

for one index k updated in cyclic order along iterations.

cyclically

end while

Output $\pi_k^{(L)}$, $\beta_k^{(L)}$ and $\sigma_k^{(L)}$ for $k = 1, \ldots, K$.

Proposition 1 *The space alternating l1-EM algorithm is a Space Alternating Kullback Proximal Point Algorithm as defined in Definition 2.1.1. of [7], i.e. the iterations can be written* [2]

$$\theta^{k+1} = \operatorname{argmax}_{\theta \in \Theta_{k-1(\text{mod } R)+1} \cap D_{ildel} \cap D_{I,\theta^k}} \tilde{l}(\theta) - \sum_{r=1}^R \lambda_r p_n(\psi_r(\theta)) - I_y(\theta, \theta^k). \tag{17}$$

where ψ*:* $\mathbb{R}^p \mapsto \mathcal{S}_1 \times \cdots \times \mathcal{S}_R$ *be a continuously differentiable mapping,* λ

[2] The relaxation sequence of Definition 2.1.1. of [7] is taken to be constant and equal to one

be a positive real vector in $\mathbb{R}^R$, p_n *be a possibly nonsmooth penalty function with bounded Clarke subdifferential on compact sets and the parameter space is decomposed into subspaces* $\Theta_r = \Theta \cap \mathcal{S}_r$, $r = 1, \ldots, R$ *where* $\mathcal{S}_1, \ldots, \mathcal{S}_R$ *are subspaces of* $\mathbb{R}^p$ *and* $\mathbb{R}^p = \oplus_{r=1}^R \mathcal{S}_r$.

Proof. First, we adopt the decomposition of the parameter space into the cartesian product of the π_k's space, the β_k's space and the σ_k's space. More precisely Θ_1 is the simplex in $\mathbb{R}^K$ and $\mathcal{S}_1 = \mathbb{R}^K$, $\Theta_{2,k} = \mathbb{R}^n = \mathcal{S}_{2,k}$, and $\Theta_{3,k} = \mathbb{R}_+$ and $\mathcal{S}_{3,k} = \mathbb{R}$ for $k = 1, \ldots, K$. Thus r takes its values in the list $\{1, (2,1), \ldots, (2,K), (3,1), \ldots, (3,K)\}$.

Then the mappings Ψ_r are just the orthogonal projections onto $\mathcal{S}_r$ for $r \in \{1, (2,1), \ldots, (2,K), (3,1), \ldots, (3,K)\}$. Moreover $\lambda_1 = 0$ and $\lambda_{(3,k)} = 0$ for $k = 1, \ldots, K$ because the class probabilities and the variances are not penalized. Moreover $\lambda_{(2,k)} = \lambda$ for $k = 1, \ldots, K$.

Next, the Q-function can be written [3]

$$Q(\theta, \bar{\theta}) = \tilde{l}(\theta) - I_y(\theta, \bar{\theta}) \tag{18}$$

with

$$I_y(\theta, \bar{\theta}) = \sum_{i=1}^{n} \sum_{k=1}^{K} t_{ik}(\bar{\theta}) \log\left(\frac{t_{ik}(\bar{\theta})}{t_{ik}(\theta)}\right). \tag{19}$$

where

$$t_{ik}(\theta) = \frac{\pi_k \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(Y_i - X\beta_k)^2}{2\sigma_k^2}\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi\sigma_l^2}} \exp\left(-\frac{(Y_i - X\beta_l)^2}{2\sigma_l^2}\right)}. \tag{20}$$

Thus, the space alternating LASSO-EM algorithm is a special case of the Space Alternating Kullback Proximal Point Algorithm for which the sequence $(\mu_k)_{k \in \mathbb{N}}$ is constant and the terms are all equal to one. □

In the following, we will denote the list of values for the parameter r by $\mathcal{R} = \{1, (2,1), \ldots, (2,K), (3,1), \ldots, (3,K)\}$. We then have the following theorem.

Theorem 2 *Let* θ^* *be a cluster point of the Space Alternating Penalized Kullback Proximal sequence. If* θ^* *lies in the interior of* D_{ildel}, *then* θ^* *satisfies the following property: there exists a set of subsets* $I_r^{**} \subset I^*$ *where* I^* *denotes the index of the active constraints at* θ^*, *i.e.* $I^* = \{(i,j)$ *s.t.* $t_{i,j}(\theta^*) = 0$, *and there is a family of real numbers* λ_{ij}, $(i,j) \in \mathcal{I}_r^{**}$, $r \in \mathcal{R}$ *such that the following Karush-Kuhn-Tucker condition for optimality holds at cluster point* θ^**:*

$$0 \in \nabla \tilde{l}(\theta^*) - \sum_{r \in \mathcal{R}} \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r \in \mathcal{R}} \sum_{(i,j) \in \mathcal{I}_r^{**}} \lambda_{ij}^* \nabla t_{ij}(\theta^*).$$

[3] see Section 4.1 of [7] for more details.

Proof. We start by verifying that Assumptions 2.2.1, 2.2.3 and Assumptions 2.2.4 of [7] hold in our case. The differentiability requirement in Assumptions 2.2.1.(i). is obvious. However, if one β_k belongs to the kernel of X, it may be of any arbitrary large norm without leading the log-likelihood towards $-\infty$. However, note that, as is well known is Gaussian mixture models, $\tilde{l}$ tends to $+\infty$ only at finite number of degenerate points. Thus, since, the penalization term p_n tends to $+\infty$ as the norm of any β_k tends to $+\infty$, the difference $\tilde{l} - \lambda_2 p_n(\Psi_2(\cdot))$ tends to $-\infty$ if the norm of any β_k goes to $+\infty$. Since l also tends to $-\infty$ as any variance tends to $+\infty$, the term $\tilde{l} - \lambda_2 p_n$ tends to $-\infty$ when the norm of θ tends to $+\infty$.

The domain $D_{\tilde{l}}$ is defined by the fact that the term inside the log in (2) must be positive. On the other hand, for any $\bar{\theta}$ in $\Theta = \Theta_1 \times \Theta_{2,1} \times \cdots \times \Theta_{2,K} \times \Theta_{3,1} \times \cdots \times \Theta_{3,K}$, the domain $D_{I_y,\bar{\theta}}$ is the set of the θ's for which the $t_{ik}(\theta)$ are positive, and therefore, does not depend on $\bar{\theta}$. Moreover, the set of θ's for which the $t_{ik}(\theta)$ are positive is $D_{\tilde{l}}$. Thus, the projection of D_I onto the first coordinate is $D_{\tilde{l}}$ and Assumptions 2.2.1.(ii). are satisfied.

Assumptions 2.2.1.(iii). is immediate since here the relaxation sequence (denoted here by $(\mu_k)_{k\in\mathbb{N}}$) is constant. Assumptions 2.2.1.(iv). is also straightforward since the mappings Ψ_r are orthogonal projections onto $\mathcal{S}_r$, $r \in \mathcal{R}$.

In our context, based on (20), we have $\delta = t\log(t) - 1$ and Assumptions 2.2.3.(i)-(iii). are easily verified. Injectivity of the mapping t when restricted to $\cup_{j=1}^{3}\Theta_{j,k}$ is proved in [4] and thus, injectivity holds on each $\Theta_{1,k},\ldots,\Theta_{3,k}$ and Assumption 2.2.3.(iv) holds.

Moreover, since $t_{ik}(\theta) = 0$ implies that $\pi_k = 0$ and $\pi_k = 0$ implies

$$\frac{\partial t_{ik}}{\partial \beta_{jl}}(\theta) = 0 \tag{21}$$

for all $j = 1,\ldots,p$ and $l = 1,\ldots,K$ and

$$\frac{\partial t_{ik}}{\partial \sigma^2}(\theta) = 0, \tag{22}$$

it follows that $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = \nabla t_{ik}(\theta^*)$ if $\mathcal{S}_r$ is the vector space generated by the probability vectors π and $P_{\mathcal{S}_r}(\nabla t_{ik}(\theta^*)) = 0$ otherwise.

Let θ^* be a cluster point in the interior of $D_{\tilde{l}}$. Since the t_{ik} are clearly continuously differentiable around such a θ^*, Corollary 1 in [7] gives that θ^* satisfies the following property: there exists a set of subsets $I_r^* \subset I^*$ and a family of real numbers λ_{ij}, $(i,j) \in \mathcal{I}_r^*$, $r \in \mathcal{R}$ such that the following Karush-Kuhn-Tucker condition for optimality holds at cluster point θ^*:

$$0 \in \nabla\tilde{l}(\theta^*) - \sum_{r\in\mathcal{R}} \lambda_r \partial p_n(\psi_r(\theta^*)) + \sum_{r\in\mathcal{R}} \sum_{(i,j)\in\mathcal{I}_r^{**}} \lambda_{ij}^* \nabla t_{ij}(\theta^*),$$

which is the desired result. □

The meaning of this theorem is simply that a Karush-Kuhn-Tucker condition is satisfied at any cluster point in the domain of definition of the log-likelihood.

3 Simulation results

In this section, we address the question of testing the algorithm on simulated and datasets. The Space Alternating l_1-EM was first tested on simulated data sets. The experiments were built as follows: 10 samples in $\mathbb{R}^2$ were generated from three different gaussian distributions with the objective to recover the index of the distribution they were drawn from up to some index permutation. The class probabilities were taken as $\pi_1 = .3$, $\pi_2 = .2$ and $\pi_3 = .5$ and the variances as $\sigma_1^2 = 5$, $\sigma_2^2 = 7$ and $\sigma_3^2 = 10$ without change through all the simulation experiments. Various experiments were performed using different values for the expectation vectors μ_1, μ_2 and μ_3 since it could be easily suspected that the distance between them would play a major role in the class index recovery problem. The results presented below were obtained using the following Monte Carlo scheme: the expectations were isotropic dilations of three points in $\mathbb{R}^2$ drawn uniformly at random in the cube $[-\frac{1}{2}, \frac{1}{2}]^3$. We ran the code for dilation factors d going from 10 to 100 by steps of 10.

3.1 Two dimensional data

An example of the type of result we obtained is given in Figure 1 below where the 10 points were correctly classified.

Here is another example when the expectation vectors are chosen closer to each other and 8 points over 10 were correctly classified.

Our Monte Carlo experiments are given in Figure 3. For each dilation parameter d, 1000 Monte Carlo experiments were performed and the number of points correctly classified was computed by finding the best permutation of the set $\{1, 2, 3\}$ matching the class indices obtained by the output of the space alternating l_1 algorithm.

These preliminary results show that the number of correctly recovered class indices increases with the dilation factor of the unit cube into which the expectation vector were drawn uniformly at random. Moreover, the larger the dilation facteur, i.e. the better separated the Gaussians are, the closer to 80% of correctly identified class indices the space alternating l_1 EM with data-driven constraint provides.

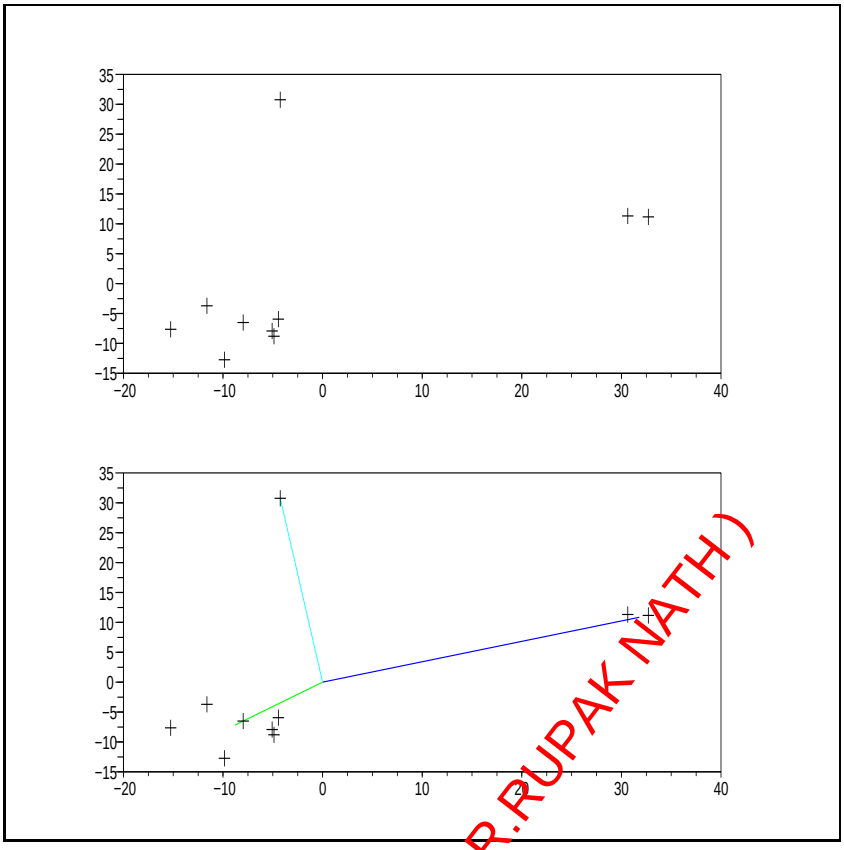

Fig. 1. The result obtained with the LASSO (or space alternating l_1)-EM for centers drawn uniformly in the cube $[-30, 30]^3$.

The number of correctly recovered class indices for 1000 Monte Carlo experiments in the case of 10 points as a function of the box into which the expectation vectors have been uniformly drawn are given in Table 1 below.

Initial cube	$[-5, 5]^2$	$[-10, 10]^2$	$[-15, 15]^2$	$[-20, 20]^2$	$[-25, 25]^2$
ANCRCI	6.113	6.982	7.494	7.891	8.233
Initial cube	$[-30, 30]^2$	$[-35, 35]^2$	$[-40, 40]^2$	$[-45, 45]^2$	$[-50, 50]^2$
ANCRCI	8.41	8.376	8.607	8.712	8.738

3.2 Large dimensional data

Recall that the estimation of the expectation vectors consists, for each cluster, of the selection of a very small number of data vectors, one linear combination of whose is suspected to lie near the center of the cluster. The main hope when we started to conceive this approach was that the dimension would have only a very small influence on the result and that estimation in large dimensions could be performed with almost as few data as in two dimensions.

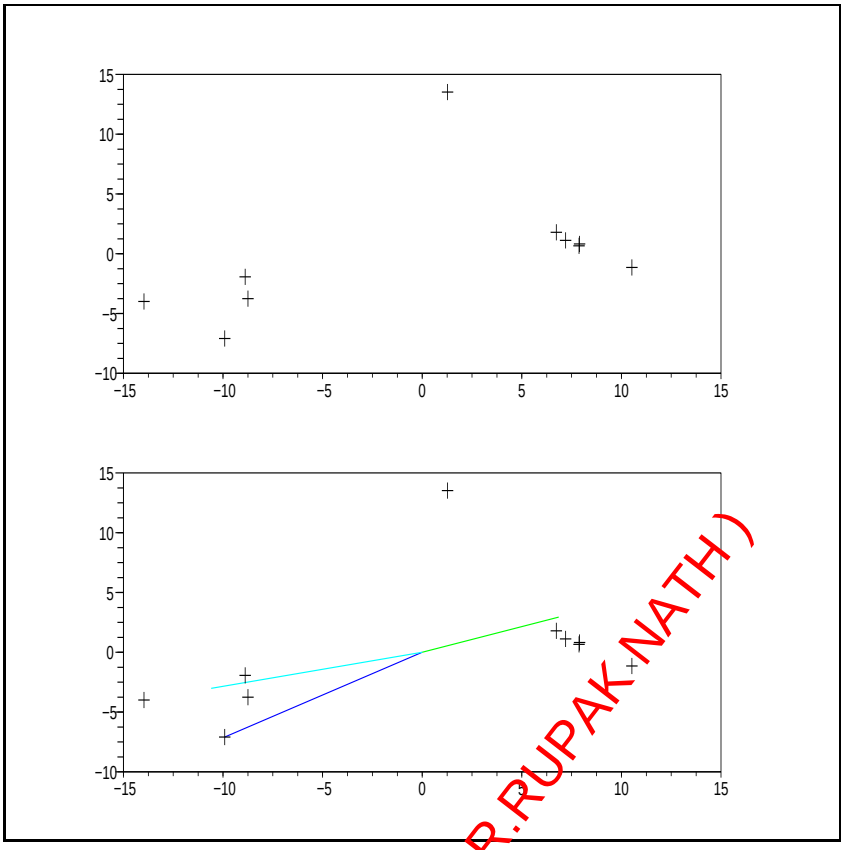

Fig. 2. The result obtained with the space alternating l_1-EM for centers drawn uniformly in the cube $[-15, 15]^3$.

In order to confirm this expectation, we performed Monte Carlo experiments in dimensions 5, 10, 15, 20, 25, 30, 35, 40, 45 and 50. The results are presented in Table 1. In order to compare with the standard likelihood approach for finite Gaussian mixtures, we gathered the results obtained for the same experiments in Table 2 [4].

As Table 1 shows, the class index recovery rate is still quite good in dimension 50 for well separed mixtures and confirm our primary intuition. A look at Table 2 shows that our method compares quite well with the standard likelihood approach for Gaussian mixtures estimation, especially in the higher dimensions where the average number of well classified data is better by often more than one unit. Giving a rigourous argument justifying these observations is currently under investigation but more experiments should be performed in order to explore in finer details the behavior of the method in more realistic context.

[4] We used the EM algorithm for Gaussian mixtures with covariance matrices equal to multiples of the identity matrix

	$[-10,10]^d$	$[-15,15]^d$	$[-20,20]^d$	$[-25,25]^d$	$[-30,30]^d$	$[-35,35]^d$	$[-40,40]^d$	$[-45,45]^d$	$[-50,50]^d$
$d=5$	7.044	7.672	8.182	8.373	8.543	8.591	8.682	8.733	8.63
$d=10$	6.499	7.498	7.834	8.174	8.419	8.425	8.599	8.471	8.6
$d=15$	6.241	7.074	7.581	8.074	8.154	8.249	8.347	8.338	8.422
$d=20$	5.878	6.544	7.391	7.736	7.905	8.01	8.037	8.121	8.126
$d=25$	5.665	6.213	6.796	7.283	7.427	7.649	7.613	7.738	7.87
$d=30$	5.661	6.189	6.663	7.063	7.303	7.412	7.386	7.576	7.441
$d=35$	5.684	6.309	6.538	6.922	7.132	7.185	7.278	7.234	7.396
$d=40$	5.769	6.296	6.788	6.977	6.974	7.218	7.292	7.369	7.371
$d=45$	5.763	6.392	6.842	6.945	7.19	7.161	7.334	7.318	7.306
$d=50$	5.798	6.545	6.816	7.007	7.247	7.223	7.376	7.369	7.422

Table 1. The average number of correctly recovered class indices (ANCRCI) over the 1000 Monte Carlo experiments shown in Figuremc is given for increasing sizes of the initial cubes where the expectation vectors are chosen uniformy at random and for increasing dimension of the sample space.

	$[-10, 10]^d$	$[-15, 15]^d$	$[-20, 20]^d$	$[-25, 25]^d$	$[-30, 30]^d$	$[-35, 35]^d$	$[-40, 40]^d$	$[-45, 45]^d$	$[-50, 50]^d$
$d = 5$	5.781	6.386	6.747	6.864	6.899	7.061	7.11	7.04	7.073
$d = 10$	5.79	6.206	6.642	6.822	6.878	6.813	6.896	6.981	7.011
$d = 15$	5.695	6.31	6.556	6.646	6.929	6.848	6.791	6.897	7.005
$d = 20$	5.648	6.019	6.374	6.609	6.655	6.567	6.86	6.837	6.802
$d = 25$	5.493	5.932	6.184	6.407	6.56	6.606	6.767	6.705	6.703
$d = 30$	5.557	5.961	6.169	6.355	6.48	6.508	6.492	6.485	6.546
$d = 35$	5.492	5.898	6.076	6.173	6.354	6.274	6.436	6.24	6.371
$d = 40$	5.514	5.819	6.151	6.134	6.195	6.248	6.212	6.291	6.437
$d = 45$	5.523	5.854	6.033	6.062	6.13	6.141	6.14	6.189	6.203
$d = 50$	5.467	5.737	6.039	6.023	6.134	6.047	6.091	6.194	6.118

Table 2. The average number of correctly recovered class indices (ANCRCI) over the 1000 Monte Carlo experiments shown in Figuremc is given for increasing sizes of the initial cubes where the expectation vectors are chosen uniformy at random and for increasing dimension of the sample space.

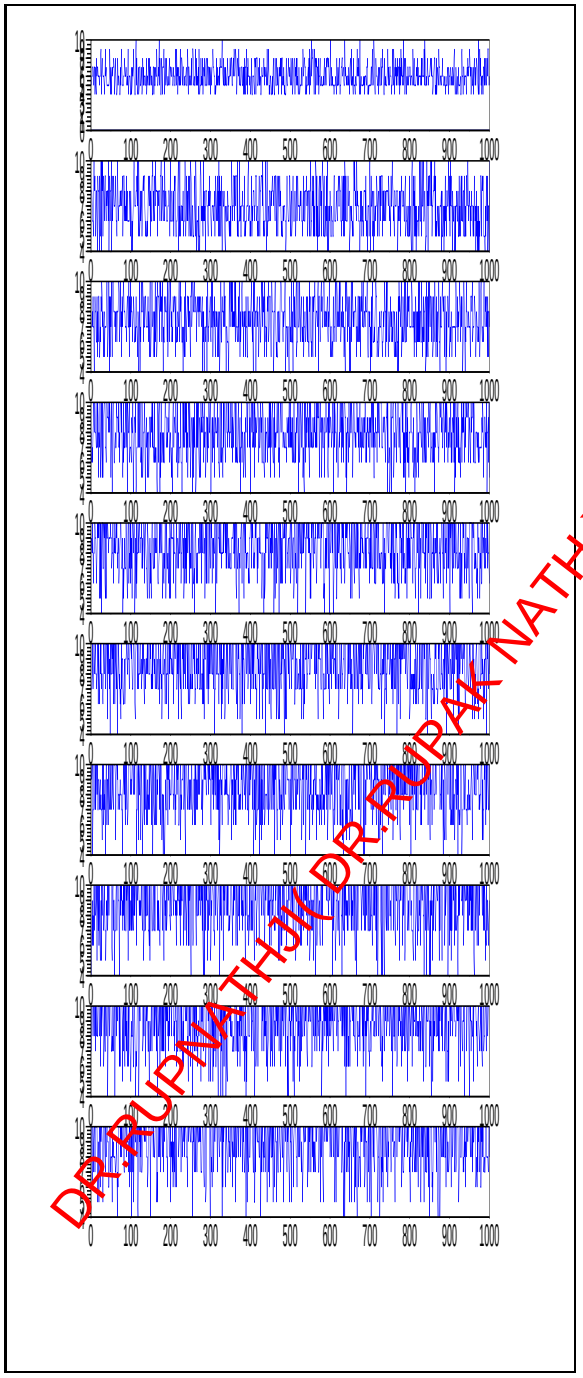

Fig. 3. For values of the dilation parameter going from 10 to 100 by steps of 10, the associated Monte Carlo simulation number is on the x-axis and the number of correctly identified class index among the 10 points is on the y-axis.

4 Conclusion

The goal of this paper was to propose a robust version of the maximum likelihood strategy for the estimation of finite Gaussian mixtures. Our approach is

based on self-regression and sparse variable selection. Sparsity was promoted by using an l_1 penalty as in the LASSO. We developed a space alternating version of the penalized EM algrithm and proved that the interesting cluster points satisfy the Karush-Kuhn-Tucker optimality conditions. Our method was then tested on simulated datasets. In particular, the Monte Carlo experiments showed that cluster identification was more robust with our approach than by using the standard maximum likelihood estimator. Theoretical justifications of these observations ought to be investigated in a near future in order to increase our understanding of the strengths and weaknesses of this approach.

Acknowledgement. The author would like to thank Amelie Vaniscotte for very helpful discussions on the results of this paper.

References

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

[2] C. Biernacki and S. Chrétien, (2003) "Degeneracy in the maximum likelihood estimation of univariate Gaussian mixtures with EM", Statist. Probab. Lett. 61, no. 4, 373–382.

[3] E. Candès and Y. Plan, (2009) "Near ideal model selection by l_1 penalization", The Annals of Statistics, to appear.

[4] G. Celeux, S. Chrétien, F. Forbes and A. Mkhadri (2001) "A Component-Wise EM Algorithm for Mixtures" Journal of Computational and Graphical Statistics, vol. 10, no. 4, 697-712.

[5] S. Chrétien and A. Hero, (2000) "Kullback proximal algorithms for maximum-likelihood estimation". Information-theoretic imaging. IEEE Trans. Inform. Theory 46, no. 5, 1800–1810

[6] S. Chrétien and A. Hero, (2008) " On EM algorithms and their proximal generalizations". ESAIM P&S 12, 308–326

[7] S. Chrétien, A. Hero and H. Perdry (2008) "Space Alternating Penalized Kullback Proximal Point Algorithms for Maximing Likelihood with Nondifferentiable Penalty", submitted. Available at http: //arxiv.org/abs/0901.0017

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, (1977) "Maximum likelihood from incomplete data via the EM algorithm," *J. Royal Statistical Society, Ser. B*, vol. 39, no. 1, pp. 1–38.

[9] J. Fan and R. Li (2001) “Variable selection via non-concave penalized likelihood and its oracle properties”, Journal of the American Statistical Association, 96, 1348–1360.

[10] J. A. Fessler, and A. O. Hero, (1994)“Space-alternating generalized expectation-maximization algorithm”, IEEE Trans. Signal Processing, vol. 42, no. 10, pp. 2664–2677.

[11] M. Hurn M, A. Justel and C.P. Robert, (2003) "Estimating Mixtures of Regressions", Journal of Computational and Graphical Statistics, 12, 55–79.

[12] A. Khalili and J. Chen, (2007) “Variable Selection in Finite Mixture of Regression Models”, Journal of the American Statistical Association, Volume 102, Number 479, pp. 1025-1038.

[13] G.J. McLachlan and D. Peel. (2000) *Finite Mixture Models*. Wiley

[14] R. Tibshirani, (1996) “ Regression shrinkage and selection via the LASSO”, Journal of the Royal Statistical Society, Series B, vol. 58, no. 1, pp. 267–288.