

Regularized EM Algorithm

Pierre Houdouin, Esa Ollila, Frédéric Pascal

▶ **To cite this version:**

Pierre Houdouin, Esa Ollila, Frédéric Pascal. Regularized EM Algorithm. ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), Jun 2023, Rhodes Island, Greece. 10.1109/ICASSP49357.2023.10096439 . hal-04254175

HAL Id: hal-04254175
https://hal.science/hal-04254175v1

Submitted on 23 Oct 2023

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

arXiv:2303.14989v1 [stat.ML] 27 Mar 2023

REGULARIZED EM ALGORITHM

Pierre Houdouin[†]*, Esa Ollila*[⋆]*, Frédéric Pascal*[†]

[†] Université Paris-Saclay, CNRS, CentraleSupélec, Laboratoire des signaux et systèmes, 91190, Gif-sur-Yvette, France.
[⋆] Department of Signal Processing and Acoustics, Aalto University, Finland

ABSTRACT

Expectation-Maximization (EM) algorithm is a widely used iterative algorithm for computing (local) maximum likelihood estimate (MLE). It can be used in an extensive range of problems, including the clustering of data based on the Gaussian mixture model (GMM). Numerical instability and convergence problems may arise in situations where the sample size is not much larger than the data dimensionality. In such low sample support (LSS) settings, the covariance matrix update in the EM-GMM algorithm may become singular or poorly conditioned, causing the algorithm to crash. On the other hand, in many signal processing problems, a priori information can be available indicating certain structures for different cluster covariance matrices. In this paper, we present a regularized EM algorithm for GMM-s that can make efficient use of such prior knowledge as well as cope with LSS situations. The method aims to maximize a penalized GMM likelihood where regularized estimation may be used to ensure positive definiteness of covariance matrix updates and shrink the estimators towards some structured target covariance matrices. We show that the theoretical guarantees of convergence hold, leading to better performing EM algorithm for structured covariance matrix models or with low sample settings.

Index Terms— Clustering, EM algorithm, Gaussian mixture model, Structured covariance matrix, Regularization.

1. INTRODUCTION

Expectation-Maximization algorithm is a widely used iterative algorithm for finding the (local) maxima of the likelihood with incomplete data. It iteratively estimates the unknown parameters of the model by increasing the expected likelihood of the complete data conditioned over the observed data and current estimates of the parameters. The EM algorithm was proposed in [1] where the authors showed that at each iteration, the likelihood of the samples increases at least as much as the conditional expected likelihood. This likelihood maximization method is very suitable for dealing with mixture models. Work in [2] considers EM algorithm for mixtures of exponential families. The problem of the Gaussian mixture model (GMM) has been treated in [3] and later extended by [4] to t- distributions in order to cope with heavy-tailed data and outliers. More recently, generalization to mixtures of elliptical distributions has been developed for clustering applications in [5] as well as for classification applications (see [6]). Then, [7] proposed an adaptation of the EM algorithm to deal with skewed normal distributions.

In signal processing problems, dimension of the data, m, is often very high while the number of available samples, n, is low or of similar magnitude as m. In such low sample support (LSS) settings, convergence issues may arise when applying the EM algorithm for GMM. Namely, the update of the covariance matrix at each step is no longer necessarily invertible or poorly conditioned, leading the M-step to be ill-defined. To avoid this problem, [8] proposes a regularized version of the EM algorithm for GMM based on resampling. Although the developed method ensures the invertibility of covariance matrix update, it does not exploit the possible underlying structure of the covariance matrix. More recently, [9] proposed the CHIME algorithm to deal with high dimensional GMM-s. It relies on the estimation of discriminant vectors $\boldsymbol{\Sigma}^{-1}(\boldsymbol{\mu}_i - \boldsymbol{\mu}_j)$ then combined with Fisher Discriminant Analysis.

Regularized covariance matrix estimation [10, 11, 12, 13, 14] is a popular technique to cope with low sample support (LSS) settings. In this paper, we present a regularized EM algorithm where we maximize a penalized version of the GMM likelihood that leads to better conditioned regularized estimators of the cluster covariance matrices as well as better clustering accuracy in LSS settings. We show that theoretical convergence guarantees hold and perform simulation experiments to validate the results.

The paper is structured as follows. Section 2 recalls the GMM and provides the main derivations of this work for penalized GMM. Section 3 contains experiments on synthetic data. Conclusions, remarks, and perspectives are drawn in Section 4.

2. REGULARIZED EM ALGORITHM

Let us assume that each observation $\mathbf{x}_i \in \mathbb{R}^m$ is drawn from a GMM. Each cluster has its own mean vector $\boldsymbol{\mu}_k \in \mathbb{R}^m$ and symmetric positive definite (SPD) $m \times m$ covariance matrix

$\mathbf{\Sigma}_k$. We then have the following probability density function for $\mathbf{x}_i$ with priors (mixing proportions) $\pi_1, ..., \pi_K \in [0, 1]$ that verify $\sum_k \pi_k = 1$, mean vectors $\boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K$ and covariance matrices $\mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_K$:

$$f(\mathbf{x}_i|\boldsymbol{\theta}) = (2\pi)^{-\frac{m}{2}} \sum_{k=1}^{K} \pi_k \, |\mathbf{\Sigma}_k|^{-\frac{1}{2}} \, e^{-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu}_k)^\top \mathbf{\Sigma}_k^{-1}(\mathbf{x}_i - \boldsymbol{\mu}_k)}$$

where $\boldsymbol{\theta} = (\pi_1, ..., \pi_K, \boldsymbol{\mu}_1, ..., \boldsymbol{\mu}_K, \mathbf{\Sigma}_1, ..., \mathbf{\Sigma}_K)$ collects all unknown parameters.

A priori knowledge of the structure of the cluster covariance matrices can be brought into the GMM estimation problem in forms of fixed SPD target $m \times m$ covariance matrices $\mathbf{T}_k$, $k = 1, \ldots, K$. Namely, the covariance matrix of kth cluster is assumed to be close (but not identical) to a prespecified target matrix $\mathbf{T}_k$. To ensure that such structure is exploited, we penalize the likelihood with the Kullback-Leibler divergence between each $\mathbf{\Sigma}_k$ and $\mathbf{T}_k$, defined as [10]

$$\Pi_{\text{KL}}(\mathbf{\Sigma}_k, \mathbf{T}_k) = \frac{1}{2}\big(\text{tr}(\mathbf{\Sigma}_k^{-1}\mathbf{T}_k) - \log|\mathbf{\Sigma}_k^{-1}\mathbf{T}_k| - m\big).$$

Let $\mathbf{X} = (\mathbf{x}_1 \quad \cdots \mathbf{x}_n)$ denote the data matrix of n i.i.d. samples from the GMM model. Then the log-likelihood of $\boldsymbol{\theta}$ based on the data is

$$\ell(\boldsymbol{\theta}|\mathbf{X}) = \sum_{i=1}^{n} \log f(\mathbf{x}_i|\boldsymbol{\theta})$$

and the aim is then to find a (local) maximizer of the penalized GMM likelihood function:

$$\ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\mathbf{X}) = \ell(\mathbf{X}|\boldsymbol{\theta}) - \sum_{k=1}^{K} \eta_k \Pi_{\text{KL}}(\mathbf{\Sigma}_k, \mathbf{T}_k)$$

where $\eta_1, ..., \eta_K \geq 0$ denote the penalization parameters for each cluster specified by the user. Let $\mathbf{z} = (z_1, \ldots, z_n)^\top$ represent the vector of latent unobserved labels. Thus $z_i \in \{1, \ldots, K\}$ specifies to which population ith observation belongs to. However, this information is missing or is unobservable. The data $(\mathbf{X}, \mathbf{z})$ is considered to as complete data, while the observed data $\mathbf{X}$ are referred to as incomplete data. Let $\log f(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})$ denote the log-likelihood function of the complete data.

Proposition 2.1. *Starting with an initial value $\boldsymbol{\theta}^{(0)}$, the iterative ($t = 1, 2, \ldots$) maximization of penalized conditional expected likelihood of the complete data*

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log f(\mathbf{z}, \mathbf{X}|\boldsymbol{\theta})\right] - \sum_{k=1}^{K} \eta_k \Pi_{\text{KL}}(\mathbf{\Sigma}_k, \mathbf{T}_k)$$

leads to sequence that ascents the penalized likelihood, i.e., $\ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(t+1)}|\mathbf{X}) \geq \ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(t)}|\mathbf{X})$.

Proof. First note that

$$\begin{aligned}
&\ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\mathbf{X}) \\
&= \log f(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta}) - \log f(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}) - \sum_{k=1}^{K} \eta_k \Pi_{\text{KL}}(\mathbf{\Sigma}_k, \mathbf{T}_k) \\
&= \mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log f(\mathbf{X}, \mathbf{z}|\boldsymbol{\theta})\right] - \sum_{k=1}^{K} \eta_k \Pi_{\text{KL}}(\mathbf{\Sigma}_k, \mathbf{T}_k) \\
&\qquad - \mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log f(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta})\right] \\
&= Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - \mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}\left[\log f(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta})\right]
\end{aligned}$$

Gibb's inequality ensures that

$$\mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log f(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta})] \leq \mathbb{E}_{\mathbf{z}|\mathbf{X},\boldsymbol{\theta}^{(t)}}[\log f(\mathbf{z}|\mathbf{X}, \boldsymbol{\theta}^{(t)})].$$

The above then implies that

$$\ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\mathbf{X}) - \ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}^{(t)}|\mathbf{X}) \geq Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)}|\boldsymbol{\theta}^{(t)})$$

so maximizing $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ causes $\ell_{\boldsymbol{\eta}}(\boldsymbol{\theta}|\mathbf{X})$ to increase at least as much. This means that we only need

$$Q(\boldsymbol{\theta}^{(t+1)} \mid \boldsymbol{\theta}^{(t)}) - Q(\boldsymbol{\theta}^{(t)} \mid \boldsymbol{\theta}^{(t)}) \geq 0$$

for the ascent property to hold. This holds since $\boldsymbol{\theta}^{(t+1)}$ is chosen to maximize $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$. □

Proposition 2.2. *The E-step of the regularized EM algorithm remains unchanged, we have $\forall i \in [\![1, n]\!], \forall k \in [\![1, K]\!]$:*

$$p_{ik}^{(t)} = \frac{\hat{\pi}_k^{(t)} |\hat{\mathbf{\Sigma}}_k^{(t)}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k^{(t)})^\top \hat{\mathbf{\Sigma}}_k^{t-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k^{(t)})}}{\sum_{j=1}^{K} \hat{\pi}_j^{(t)} |\hat{\mathbf{\Sigma}}_j^{t}|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(t)})^\top \hat{\mathbf{\Sigma}}_j^{t-1}(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_j^{(t)})}} \quad (1)$$

Proof. This is obvious since the penalization term has no impact on the expected conditional likelihood. Thus evaluating $Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)})$ one can notice that, similar to non-penalized case, only

$$p_{ik}^{(t)} = \Pr(z_j = k|\mathbf{x}_i, \boldsymbol{\theta}^{(t)})$$

depends on $\boldsymbol{\theta}^{(t)}$. One can then apply the Bayes theorem and use the fact that conditional class distributions are Gaussian, i.e., $\mathbf{x}_i|z_j = k \sim \mathcal{N}_m(\boldsymbol{\mu}_k, \mathbf{\Sigma}_k)$ and $\pi_k = \Pr(z_j = k)$. □

Proposition 2.3. *The M-step consists in :*

$$\pi_k^{(t+1)} = \frac{1}{n}\sum_{i=1}^{n} p_{ik}^{(t)}, \qquad \hat{\boldsymbol{\mu}}_k^{(t+1)} = \sum_{i=1}^{n} w_{ik}^{(t)} \mathbf{x}_i$$

$$\hat{\mathbf{\Sigma}}_k^{(t+1)} = \beta_k^{(t+1)} \sum_{i=1}^{n} w_{ik}^{(t)} (\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k^{(t)})(\mathbf{x}_i - \hat{\boldsymbol{\mu}}_k^{(t)})^\top + (1 - \beta_k^{(t+1)})\mathbf{T}_k,$$

where $\beta_k^{(t+1)} = \dfrac{n\pi_k^{(t+1)}}{\eta_k + n\pi_k^{(t+1)}}$ *and* $w_{ik}^{(t)} = \dfrac{p_{ik}^{(t)}}{\sum_{i=1}^{n} p_{ik}^{(t)}}$

Algorithm 1 L-fold cross-validation of η_k for kth cluster

Input: Initial set of indices $\mathcal{D}^0$ and scale $\hat{\theta}_k^0$ for the kth cluster. A set $\{\eta_j\}_{j=1}^J$ of candidate penalty parameter values.

Split $\mathcal{D}^0$ into L distinct folds $\mathcal{D}_1, \ldots, \mathcal{D}_L$ s.t. $\mathcal{D}^0 = \cup_{l=1}^L \mathcal{D}_l$ and set $\mathbf{T}_k^0 = \hat{\theta}_k^0 \cdot \mathbf{I}_m$ as the target matrix
Set $\mathrm{Err}_j \equiv \mathrm{Err}(\eta_j) = 0$ for $j = 1, \ldots, J$.
for $l \in [\![1, L]\!]$ **do**
    Set $\mathcal{D}_{\text{val}} = \mathcal{D}_l$ and $\mathcal{D}_{\text{tr}} = \mathcal{D} / \mathcal{D}_{\text{val}}$
    $\mathbf{S}_{\text{val}} = \frac{1}{|\mathcal{D}_{\text{val}}|} \sum_{i \in \mathcal{D}_{\text{val}}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\text{val}})(\mathbf{x}_i - \bar{\mathbf{x}}_{\text{val}})^\top$
    $\hat{\mathbf{\Sigma}} = \frac{1}{|\mathcal{D}_{\text{tr}}|} \sum_{i \in \mathcal{D}_{\text{tr}}} (\mathbf{x}_i - \bar{\mathbf{x}}_{\text{tr}})(\mathbf{x}_i - \bar{\mathbf{x}}_{\text{tr}})^\top$
    for $\eta \in \{\eta_1, \ldots, \eta_J\}$ **do**
        $\hat{\mathbf{\Sigma}}_\eta = \frac{|\mathcal{D}_{\text{tr}}|}{\eta + |\mathcal{D}_{\text{tr}}|} \hat{\mathbf{\Sigma}} + \frac{\eta}{\eta + |\mathcal{D}_{\text{tr}}|} \mathbf{T}_k^0$
        $\mathrm{Err}_l = \mathrm{Err}_l + \mathrm{tr}(\hat{\mathbf{\Sigma}}_\eta^{-1} \mathbf{S}_{\text{val}}) + \log |\mathbf{\Sigma}_\eta|$
Choose η_j that minimizes $\{\mathrm{Err}(\eta_j)\}_{j=1}^J$

Proof. We wish to maximize

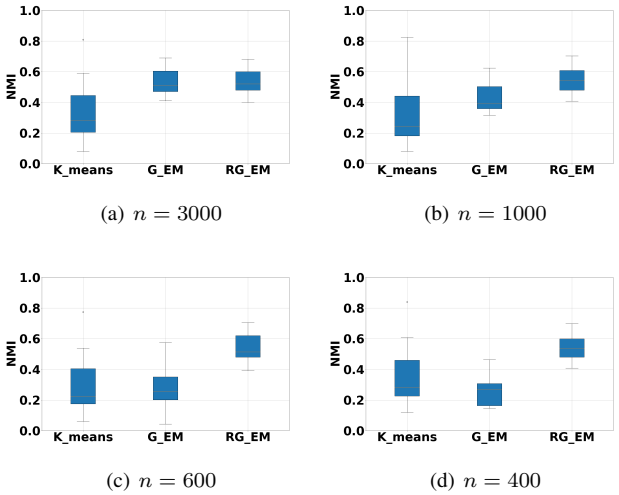
$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \sum_{i=1}^{n} \sum_{k=1}^{K} p_{ik}^{(t)} (\log f(\mathbf{x}_i \in \mathcal{C}_k|\boldsymbol{\theta}) + \log \pi_k) - \frac{1}{2} \sum_{k=1}^{K} \eta_k \left(\mathrm{tr}(\mathbf{\Sigma}_k^{-1} \mathbf{T}_k) - \log |\mathbf{\Sigma}_k^{-1} \mathbf{T}_k| - m \right)$$

Same steps are used as in the usual EM algorithm to derive $\hat{\pi}_k^{(t+1)}$ and $\hat{\boldsymbol{\mu}}_k^{(t+1)}$. Solving for $\hat{\mathbf{\Sigma}}_k^{(t+1)}$, we can compute the solution of $\nabla_{\mathbf{\Sigma}_k^{-1}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}^{(t)}) = \mathbf{0}$ while keeping the other parameters of $\boldsymbol{\theta}$ fixed with their solutions for iteration t. By setting $\tilde{\mathbf{x}}_i = \mathbf{x}_i - \hat{\boldsymbol{\mu}}_k^{(t)}$, this leads to solving the following estimating equation:

$$\sum_{i=1}^{n} p_{ik}^{(t)} \left(\mathbf{\Sigma}_k - \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top \right) - \eta_k \left(\mathbf{\Sigma}_k - \mathbf{T}_k \right) = \mathbf{0}$$
$$\Leftrightarrow \sum_{i=1}^{n} p_{ik}^{(t)} \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_i^\top + \eta_k \mathbf{T}_k = \left(\eta_k + \sum_{i=1}^{n} p_{ik}^{(t)} \right) \mathbf{\Sigma}_k$$

whose solution is $\hat{\mathbf{\Sigma}}_k^{(t+1)}$ given in the proposition. □

Let us now turn the discussion to user-defined regularization parameters, $\mathbf{T}_k$-s and η_k-s. The fixed SPD target matrices $\mathbf{T}_k$ can bring prior knowledge to the estimation problem or simply provide a well-conditioned estimator of cluster covariance matrices by shrinking the covariance matrix updates towards a scaled identity matrix. In the latter case, the obvious target matrices to be used are $\mathbf{T}_k = \hat{\theta}_k^0 \cdot \mathbf{I}_m$ where $\hat{\theta}_k^0$ is initial estimate of the scale statistic $\theta_k = \mathrm{tr}(\mathbf{\Sigma}_k)/m$. Here we use $\hat{\theta}_k^0 = \mathrm{tr}(\hat{\mathbf{\Sigma}}_k^0)/m$, where $\hat{\mathbf{\Sigma}}_k^0$ is an initial estimate of the kth cluster covariance matrix obtained *e.g.*, with a first clustering obtained by K-means or after several iterations of the EM.



(a) $n = 3000$ (b) $n = 1000$ (c) $n = 600$ (d) $n = 400$

Fig. 1. Performances in terms of NMI for different sample lengths n. The dimension is $m = 50$.

The choice of regularization parameter is also important. We choose the regularization parameter using a cross-validation procedure minimizing the negative Gaussian log-likelihood as in [14]. Each η_k is estimated independently among a set of candidates values $\{\eta_1, \ldots, \eta_J\}$. The procedure is described in Algorithm 1.

3. EXPERIMENTS ON SYNTHETIC DATA

3.1. Simulation set-up

The performance of the proposed method is compared with K-means and vanilla Gaussian-EM (G-EM) algorithm. In order to make G-EM competitive in regimes with high dimensions and low sample size, we add a vanilla regularization to G-EM for the estimation of the covariance matrix. K-means is implemented using Scikit-learn library [15] while the proposed regularized Gaussian EM (RG-EM) algorithm is implemented from scratch. For both the G-EM and the proposed RG-EM, we grant a maximum of 200 iterations for parameter estimation. For our method, we recompute the optimal η_k every 20 iterations.

The number of clusters is $K = 3$, and data is generated from GMM with equal prior probabilities ($\pi_k = 1/3$, $\forall k$). The mean vectors of each cluster are drawn on a centered sphere of radius 2. Cluster covariance matrices possess autoregressive AR(1) structure $(\mathbf{\Sigma}_k)_{ij} = \theta_k \varrho_k^{|i-j|}$, where each cluster having its own AR correlation coefficient ϱ_k. The total number of points and the dimension vary across the simulations.

3.2. Performance in various settings

We test the performances of each method on various scenarios differing by the number of points n and dimension m. For each scenario, we run 25 simulations with different covariance matrices and mean vectors.

Results are displayed in Fig. 1, and performance is evaluated using NMI index [16]. Computed using both the mutual information and the entropy of each cluster, it is a score between 0 and 1 that reflects how similar the set found are compared to the original clusters. In the first scenario, a lot of samples are available for each cluster. K-means performs poorly compared to EM-based methods, probably because clusters sometimes overlap. G-EM and proposed RG-EM perform equally well since there are enough data to accurately estimate the covariance matrices (regularization is not necessary). Then we gradually reduce n in the train set. As can be noted, the performance of K-means is not much affected, while the performance of G-EM drops heavily since there is not enough data for obtaining well-conditioned covariance matrix updates, although the vanilla regularization ensures invertibility. On the contrary, RG-EM handles very well the reduction of the train set size and keeps almost equal performances, using the target matrix to compensate for the lack of data.

Fig.2 displays the Frobenius error for estimation of the covariance matrix of cluster 1 across the iterations for $n = 600$ and $m = 50$. As expected, the Robust Gaussian EM method has a much smaller error compared to the vanilla EM method. It converges much faster with a smaller variance toward its final estimation. On the contrary, vanilla EM estimation converges toward a worse estimation of the covariance matrix, as the error is higher than the first guess with K-means.

We conclude our experiments by studying the evolution of NMI of each method when the dimension m increases for fixed values of $n = 600$ and $n = 1000$. As can be noted from Fig. 3, while RG-EM can cope well with the high-dimensionality of the data, with performances almost as good as in low dimensions, there is a breakpoint dimension for G-EM where its NMI drops. Such a threshold value depends on the number of data points: the larger the train set is, the higher the breakpoint is.

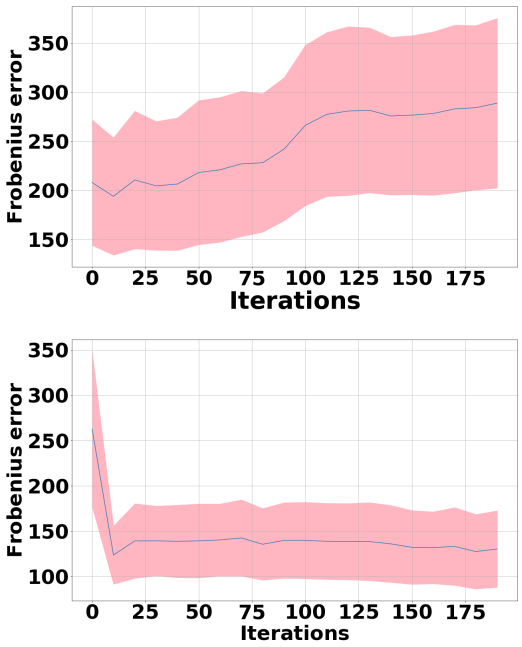


Fig. 2. Frobenius error for G-EM (*top panel*) and proposed RG-EM (*bottom panel*) for cluster 1; $n = 600$, $m = 50$.

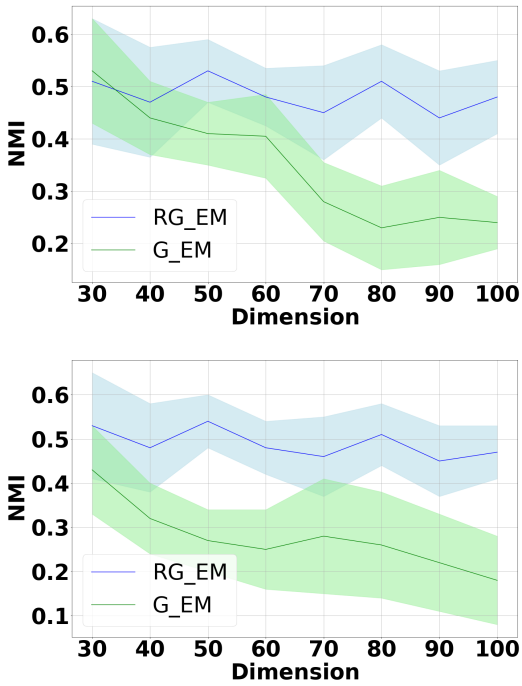


Fig. 3. Performance evaluation via NMI index for different dimensions m when $n = 1000$ (*top panel*) and $n = 600$ (*bottom panel*).

4. CONCLUSION

In this paper, we presented a regularized version of the EM algorithm for GMM that outperforms several state-of-the-art methods in low sample size regimes on simulated datasets. In this new approach, the estimation of the covariance matrix is regularized with an additive penalty that shrinks the covariance matrix towards a preset target matrix. If the target matrix is close enough to the actual covariance matrix, we achieve very good performances even with few data in high dimensions. The optimal coefficients η_k controlling the regularization are chosen following a cross-validation procedure and updated regularly across the iterations. Such cross-validation procedure avoids regularization if enough data are

available, retrieving the performance of the classical EM algorithm. Our method can thus be seen as an enhancement of regular EM when the data aspect ratio $\frac{n}{m}$ is low.

5. REFERENCES

[1] 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. Series B (Methodological)*, vol. 39, no. 1, pp. 1–22, 1977.

[2] R. A. Redner and H. F. Walker, "Mixture densities, maximum likelihood and the EM algorithm," *SIAM Review*, vol. 26, no. 2, pp. 195–239, 1984.

[3] G. Xuan, W. Zhang, and P. Chai, "EM algorithms of Gaussian mixture model and hidden Markov model," in *Proceedings 2001 International Conference on Image Processing (Cat. No.01CH37205)*, vol. 1, 2001, pp. 145–148 vol.1.

[4] S. Ingrassia, S. C. Minotti, and G. Incarbone, "An EM algorithm for the student-t cluster-weighted modeling," in *Challenges at the Interface of Data Analysis, Computer Science, and Optimization*. Springer, 2012, pp. 13–21.

[5] V. Roizman, M. Jonckheere, and F. Pascal, "A flexible EM-like clustering algorithm for noisy data," *arXiv preprint arXiv:1907.01660*, 2019.

[6] P. Houdouin, A. Wang, M. Jonckheere, and F. Pascal, "Robust classification with flexible discriminant analysis in heterogeneous data," in *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2022, pp. 5717–5721.

[7] M. Teimouri, "EM algorithm for mixture of skew-normal distributions fitted to grouped data," *Journal of Applied Statistics*, vol. 48, no. 7, pp. 1154–1179, 2021.

[8] X. Yi and C. Caramanis, "Regularized em algorithms: A unified framework and statistical guarantees," 2015. [Online]. Available: https://arxiv.org/abs/1511.08551

[9] T. Cai, J. Ma, and L. Zhang, "CHIME: Clustering of high-dimensional gaussian mixtures with EM algorithm and its optimality," *Annals of Statistics*, vol. 47, no. 3, pp. 1234–1267, 2019.

[10] Y. Sun, P. Babu, and D. P. Palomar, "Regularized Tyler's scatter estimator: Existence, uniqueness, and algorithms," *IEEE Transactions on Signal Processing*, vol. 62, no. 19, pp. 5143–5156, 2014.

[11] F. Pascal, Y. Chitour, and Y. Quek, "Generalized robust shrinkage estimator and its application to STAP detection problem," *IEEE Transactions on Signal Processing*, vol. 62, no. 21, pp. 5640–5651, 2014.

[12] E. Ollila and E. Raninen, "Optimal shrinkage covariance matrix estimation under random sampling from elliptical distributions," *IEEE Transactions on Signal Processing*, vol. 67, no. 10, pp. 2707–2719, 2019.

[13] E. Ollila, D. P. Palomar, and F. Pascal, "Shrinking the eigenvalues of M-estimators of covariance matrix," *IEEE Transactions on Signal Processing*, vol. 69, pp. 256–269, 2021.

[14] M. Yi and D. E. Tyler, "Shrinking the covariance matrix using convex penalties on the matrix-log transformation," *Journal of Computational and Graphical Statistics*, vol. 30, no. 2, pp. 442–451, 2020.

[15] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[16] A. F. McDaid, D. Greene, and N. Hurley, "Normalized mutual information to evaluate overlapping community finding algorithms," 2011. [Online]. Available: https://arxiv.org/abs/1110.2515