Group-wise penalized estimation schemes in model-based clustering

Strategie di stima penalizzata a livello di gruppo nel clustering basato su modello

Alessandro Casa, Andrea Cappozzo and Michael Fop

Abstract Gaussian mixture models provide a probabilistically sound clustering approach. However, their tendency to be over-parameterized endangers their utility in high dimensions. To induce sparsity, penalized model-based clustering strategies have been explored. Some of these approaches, exploiting the link between Gaussian graphical models and mixtures, allow to handle large precision matrices, encoding variables relationships. By assuming similar components sparsity levels, these methods fall short when the dependence structures are group-dependent. Our proposal, by penalizing group-specific transformations of the precision matrices, automatically handles situations where under or over-connectivity between variables is witnessed. The performances of the method are shown via a real data experiment.
Abstract *La sovra-parametrizzazione dei modelli di mistura Gaussiani, che rappresentano un approccio probabilistico al clustering, mette a rischio la loro utilità in dimensioni elevate. Per questo motivo sono state proposte strategie di stima penalizzate che permettono di gestire matrici di precisioni di grandi dimensioni, sfruttando il legame tra modelli grafici Gaussiani e modelli mistura. Questi metodi, assumendo sparsità simile tra tutte le componenti, falliscono quando la struttura di dipendenza varia di gruppo in gruppo. La nostra proposta, penalizzando una trasformazione delle matrici di precisione differente per ogni componente, gestisce situazioni in cui il numero di connessioni tra le variabili è diverso tra i gruppi. La validità del metodo è evidenziata grazie ad un'applicazione a dati reali.*

Alessandro Casa
Faculty of Economics and Management, Free University of Bozen-Bolzano
e-mail: alessandro.casa@unibz.it

Andrea Cappozzo
MOX - Laboratory for Modeling and Scientific Computing, Politecnico di Milano
e-mail: andrea.cappozzo@polimi.it

Michael Fop
School of Mathematics and Statistics, University College Dublin
e-mail: michael.fop@ucd.ie

Key words: Model-based clustering, Graphical lasso, EM algorithm, Gaussian graphical models

1 Introduction

Model-based clustering [2] represents a widely known and probabilistic-based strategy to cluster analysis. Here, the data generative mechanism is assumed to be adequately described by means of finite mixture models, with the Gaussian distribution being commonly considered as the component density when dealing with continuous data. Partitions are then practically obtained by drawing a one-to-one correspondence between mixture components and groups.

While being fruitfully adopted in a lot of different applications, one of the major shortcomings of this approach lies in its tendency to be over-parameterized in high-dimensional spaces. In fact, the number of parameters to estimate scales quadratically with the number of the observed variables, endangering the practical applicability of the method in some scenarios. To overcome this issue, several different approaches have been proposed in the literature (see [1] for a review on the topic).

Here, we focus specifically on a class of strategies that aims to induce parsimony by adopting penalized estimation schemes. More specifically, in [6] the number of association parameters to be estimated is drastically reduced by penalizing the component precision matrices via a graphical lasso penalty. Conveniently, zero entries in these matrices imply conditional independence between the corresponding variables, and the dependence structure can be visually represented by means of Gaussian graphical models. The adoption of a common shrinkage factor for all the component implies that the number of non-zero entries is similar across precision matrices for different components. This assumption can hinder group discrimination as it can be quite restrictive in those settings where the associations between the variables show cluster-dependent patterns.

To overcome this drawback, in this work we propose a generalization of the approach by [6], where we penalize component-specific transformations of the precision matrices rather than the matrices themselves. As a result, our method turns out to be more flexible and adaptive, without requiring the specification of additional hyper-parameters, as it automatically encompasses those situations where under or over-connectivity is witnessed in the class-specific graphical models. The rest of the paper is structured as follows. In Section 2 we outline the proposal, while in Section 3 we show the validity of the approach by applying it on a real data example. Lastly, in Section 4 we conclude with some remarks and highlighting possible future research directions.

2 Proposed methodology

Let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_i, \dots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathbb{R}^p$, be the set of the observed data. Coherently with the model-based formulation, to cluster the data into K different groups, we consider Gaussian mixture models. Given the considerations in the previous section, the parameters of the model are estimated by maximizing a penalized log-likelihood function which reads as:

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} ||\mathbf{P}_k * \Omega_k|| \tag{1}$$

where π_k's denote the mixing proportions, with $\pi_k > 0, \forall k$ and $\sum_k \pi_k = 1$, and $\phi(\cdot; \mu_k, \Omega_k)$ is the density of a multivariate Gaussian distribution with mean vector $\mu_k \in \mathbb{R}^p$ and $p \times p$ precision matrix Ω_k. Therefore, the first term in (1) represents the log-likelihood of a Gaussian mixture model, while the second one introduces the graphical lasso penalty, with shrinkage hyper-parameter λ. More specifically, with $||\cdot||$ we denote the element-wise L_1 norm, with $*$ the Hadamard product, and $\mathbf{P}_k$s are the matrices that drive the component-specific transformation of Ω_k. By penalizing the elements of $\mathbf{P}_k * \Omega_k$, instead of the ones in Ω_k as conversely done in [6], we scale the effect of λ and we uncover group-wise conditional dependence structures among the variables. As a consequence, our approach automatically encompasses those settings where the number of non-zero entries in Ω_ks is dissimilar.

Since the $\mathbf{P}_k$s encode information about class-specific sparsity patterns, they play a pivotal role in our proposal, therefore the focus is shifted towards their specification. We adopt a data-driven procedure, relying on estimated sample precision matrices $\hat{\Omega}_1^{(0)}, \dots, \hat{\Omega}_K^{(0)}$, obtained conditionally on carefully initialized partitions. The weight matrices are then defined as $\mathbf{P}_k = f(\hat{\Omega}_k^{(0)})$, with $f : \mathbb{S}_+^p \to \mathbb{S}^p$ a function from the space of positive semi-definite matrices to the space of p-dimensional symmetric matrices.

Hereafter we describe two viable options to define $f(\cdot)$. Nonetheless, we are aware that different routes can be taken when specifying $f(\cdot)$, with subjectivity and prior information potentially playing a relevant role in the process.

- According to the first proposal, which can be seen as a multiclass generalization of the approach by [4], $\mathbf{P}_k$ is defined as

$$P_{k,ij} = 1/|\hat{\Omega}_{k,ij}^{(0)}| \tag{2}$$

  with $P_{k,ij}$ and $\hat{\Omega}_{k,ij}^{(0)}$ denoting the (i,j)-th elements of the matrices $\mathbf{P}_k$ and $\hat{\Omega}_k^{(0)}$ respectively. This specification allows to inflate/deflate the penalty terms on the elements of Ω_k according to the element-wise magnitude of $\hat{\Omega}_k^{(0)}$. In fact, when $|\hat{\Omega}_{k,ij}^{(0)}|$ is close to 0, $P_{k,ij}$ would impose an extra shrinkage on $\Omega_{k,ij}$.
- The second proposal sets the elements of $\mathbf{P}_k$ proportional to the distance between $\hat{\Omega}_k^{(0)}$ and $\mathrm{diag}(\hat{\Omega}_k^{(0)})$, where $\mathrm{diag}(\hat{\Omega}_k^{(0)})$ is a diagonal matrix whose elements are

Table 1 ARI, number of estimated parameters d_Ω and Median Frobenius Distance, for different penalized model-based clustering methods.

	ARI	d_Ω	MFD
Zhou et al.(2009)	0.6724	320	830
$\mathbf{P}_k$ as in (2)	0.7199	242	421
$\mathbf{P}_k$ as in (3), Frobenius	0.6875	312	701
$\mathbf{P}_k$ as in (3), Riemannian	0.6812	314	798

equal to the ones in $\hat{\Omega}_k^{(0)}$. The entries of $\mathbf{P}_k$ are practically computed as

$$P_{k,ij} = \frac{1}{\mathscr{D}\left(\hat{\Omega}_k^{(0)}, \mathrm{diag}\left(\hat{\Omega}_{\mathrm{k}}^{(0)}\right)\right)}, \tag{3}$$

for $i, j = 1, \dots, p$. With $\mathscr{D}(\cdot,\cdot)$ we denote a suitable measure of distance between positive semi-definite matrices. Since $\mathbb{S}_+^p$ is a non-Euclidean space, we employ Frobenius and Riemannian distances (see [3] for a detailed discussion).

These two strategies share the same rationale, as they aim to penalize more strongly those entries corresponding to weaker sample conditional dependencies. Nonetheless, while for the second approach $\mathbf{P}_k$s depend on a group specific constant, in the first one the induced penalty is entry-wise different, thus possibly more accurate when the sample estimates $\hat{\Omega}_k^{(0)}$s are regarded as reliable. Lastly note that in [6] $\mathbf{P}_k$ is assumed to be a matrix of ones for all $k = 1, \dots, K$.

Once $\mathbf{P}_1, \dots, \mathbf{P}_K$ are specified, the model is estimated employing an EM-algorithm with the graphical lasso embedded in the M-step, when estimating sparse precision matrices.

3 Application

Our proposal is here employed on the Olive Oil dataset, which is publicly available in the `R` package `pgmm` [5]. The data report the percentage composition of $p = 8$ fatty acids for $n = 572$ samples of olive oil, coming from $K = 9$ different regions in Italy. The aim of the analyses consists in recovering the group structure, given by the geographical partition, of the oils by using their lipidic characteristics.

In the analyses we compare our method, considering different specifications of $\mathbf{P}_k$ as outlined in the previous section, with the strategy proposed by [6]. Different competitors are compared in terms of clustering performances via the Adjusted Rand Index (ARI). Moreover, we evaluate also the number of non-zero parameters d_Ω, as a proxy of model complexity, and the Median Frobenius Distance (MFD) computed as:

$$\mathrm{median}_{k=1,\dots,K}\left(||\hat{\Omega}_k - \bar{\Omega}_k||_F\right)$$

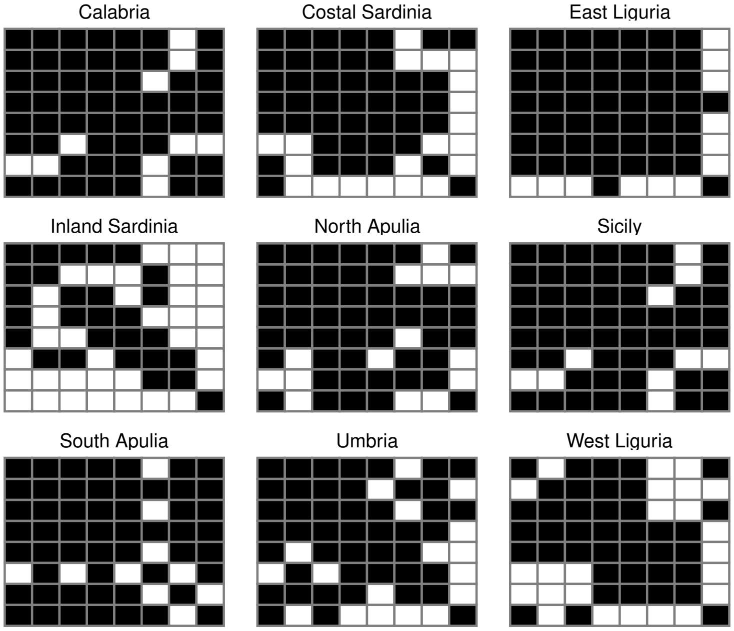

Fig. 1 Estimated precision matrices, with $\mathbf{P}_k$ defined as in (2). Black squares denote the presence of an edge between the two variables.

where $||\cdot||_F$ denotes the Frobenius norm, while $\bar{\Omega}_k$ is the k-th component empirical precision matrix, computed using the true labels, which allows to evaluate how the model identifies the conditional association structure among the variables. The results are reported in Table 1.

We immediately note that, including a data-driven specification for $\mathbf{P}_k$ slightly improves the clustering performance with respect to the all-one matrix as in [6]. Furthermore, our proposals are able to obtain a reduction in the total number of non-zero parameters d_Ω, especially when defining the weight matrices as in (2). This latter approach appears to be the best one also when considering the Median Frobenius Distance, thus when evaluating how good the method is in recovering the true conditional dependencies. Figure 1 displays the component precision matrices estimated using this method; from here we see that the association structure varies appreciably across regions, with our proposal exploiting this behaviour in the estimation step.

4 Conclusion and discussion

In this work we showed how, in the penalized clustering framework, partitions retrieval can be jeopardized when imposing a single penalty on the component precision matrices. In fact, automatically enforcing similarities in the estimated graphical models across groups, this can be harmful when it comes to groups discrimination.

More specifically, we have proposed a generalization of the approach outlined in [6]. Here the authors, by considering a single penalization parameter, implicitly assume that all the groups present a similar degree of sparsity. Therefore, this method does not account for those situations where one or more components shows under or over-connectivity with respect to the others. For this reason, we have devised a procedure which penalizes a group-specific transformation of the component precision matrices. The proposal automatically encompasses situations where the groups are characterized by a different amount of non-zero entries in the corresponding precision matrices. In our work, we proposed several different ways to define the transformed precision matrices to be penalized. Numerical explorations on real data have confirmed the validity of the method.

Lastly note that, while outlined for Gaussian mixtures parameterized in terms of precision matrices, this penalized approach can be fruitfully generalized to component covariance matrices. Moreover, if paired with a carefully chosen penalization term on the component means, this methodology can be used to perform variable selection in the model-based clustering context.

References

1. Bouveyron, C. & Brunet-Saumard, C. Model-based clustering of high-dimensional data: A review. Comput Stat Data An, **71**, 52-78 (2014)
2. Bouveyron, C., Celeux, G., Murphy, T.B. & Raftery, A.E. Model-based clustering and classification for data science: with applications in R. Cambridge University Press (2019)
3. Dryden, I.L., Koloydenko, A. & Zhou, D. Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. Ann Appl Stat, **3(3)**, 1102-1123 (2009)
4. Fan, J., Feng, Y. & Wu, Y. Network exploration via the adaptive LASSO and SCAD penalties. Ann Appl Stat, **3(2)**, 521-541 (2009)
5. McNicholas, P.D., ElSherbiny, A., McDaid, A.F. & Murphy, T.B. pgmm: Parsimonious Gaussian Mixture Models. R package version 1.2.4 (2019)
6. Zhou, H., Pan, W., & Shen, X. Penalized model-based clustering with unconstrained covariance matrices. Electron J Stat, **3**, 1473-1496 (2009)