Penalized model-based clustering with group-dependent shrinkage estimation

Alessandro Casa, Andrea Cappozzo and Michael Fop

Abstract Gaussian mixture models (GMM) are the most-widely employed approach to perform model-based clustering of continuous features. Grievously, with the increasing availability of high-dimensional datasets, their direct applicability is put at stake: GMMs suffer from the curse of dimensionality issue, as the number of parameters grows quadratically with the number of variables. To this extent, a methodological link between Gaussian mixtures and Gaussian graphical models has recently been established in order to provide a framework for performing penalized model-based clustering in presence of large precision matrices. Notwithstanding, current methodologies do not account for the fact that groups may be under or over-connected, thus implicitly assuming similar levels of sparsity across clusters. We overcome this limitation by defining data-driven and component specific penalty factors, automatically accounting for different degrees of connections within groups. A real data experiment on handwritten digits recognition showcases the validity of our proposal.

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 & Statistics, University College Dublin, e-mail: michael.fop@ucd.ie

1 Introduction and motivation

In model-based clustering finite mixture models are employed to delineate a one-to-one correspondence between mixture components and sought clusters, with the Gaussian distribution being the conventional choice to group multivariate continuous samples (Bouveyron et al., 2019). Unfortunately, in the big data era the applicability of this well-established procedure is jeopardized as Gaussian mixture models (GMM) tend to be over-parameterized in high-dimensional settings (Bouveyron and Brunet-Saumard, 2014). To mitigate this issue, several solutions have been proposed that include constrained modelling, variable selection and sparse estimation (Fop and Murphy, 2018). Particularly, within the latter family, Zhou et al. (2009) proposed a penalized approach in which the number of parameters to be estimated is drastically reduced by enforcing a graphical lasso penalty in the objective function (Friedman et al., 2008). The resulting penalized likelihood allows to detect different sparsity patterns in the estimated precision matrices, but it falls short when these matrices have a substantially different number of non-zero entries, as the method explicitly assumes a common shrinkage factor for each and every component of the mixture. Such a behavior may hinder the resulting clustering in applications where sparse intensity is cluster-wise different. To overcome this limitation, the present paper extends the methodology of Zhou et al. (2009) by devising group-wise penalty factors which automatically enforce under or over-connectivity in the precision matrices. The approach is entirely data-driven and does not require any additional hyper-parameter specification.

The remainder of the paper is structured as follows. In Section 2 we introduce our new proposal and we discuss two strategies to compute cluster-specific penalty factors. Section 3 presents a digits recognition application, in which dependence structures between pixels differ across digits. Section 4 summarizes the novel contributions and highlights future research directions.

2 Proposed solution

Consider a set of n observed data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$, with $\mathbf{x}_i \in \mathbb{R}^p$ for $i = 1, \ldots, n$. With the aim of partitioning $\mathbf{X}$ in K subpopulations or clusters, the present work proposes to carry out parameter estimation by maximizing the following penalized log-likelihood function:

$$\sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(\mathbf{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k) - \lambda \sum_{k=1}^{K} \|\mathbf{P}_k * \boldsymbol{\Omega}_k\|_1 \,. \tag{1}$$

The first term in (1) is the log-likelihood of a GMM, with K the number of mixture components, π_ks the mixing proportions ($\pi_k > 0$, $\sum_k \pi_k = 1$), and $\phi(\cdot; \boldsymbol{\mu}_k, \boldsymbol{\Omega}_k)$ the density of a multivariate Gaussian distribution with mean vector $\boldsymbol{\mu}_k = (\mu_{1k}, \ldots, \mu_{pk})$ and precision matrix $\boldsymbol{\Omega}_k$, $k = 1, \ldots, K$. The second term in (1) identifies a graphical lasso penalty with shrinkage factor λ that is applied to the K precision matrices. In details, $\|\cdot\|_1$ is the L_1 norm taken element-wise ($\|A\|_1 = \sum_{ij} |A_{ij}|$), with $*$ we denote the Hadamard product, and $\mathbf{P}_k$s are weighting matrices that scale the effect of the common penalty λ depending on the component-specific sparsity underlying cluster k, $k = 1, \ldots, K$. Such a penalty forces some entries in the precision matrices to be shrunk to 0, uncovering group-wise conditional independence among the variables.

The original proposal by Zhou et al. (2009) implicitly assumed $\mathbf{P}_k$ to be an all-one matrix $\forall k$, our specification of $\mathbf{P}_1, \ldots, \mathbf{P}_K$ instead allows to encode information about class specific sparsity patterns, accounting for under or over-connectivity scenarios. We rely on carefully initialized sample precision matrices $\hat{\boldsymbol{\Omega}}_1^{(0)}, \ldots, \hat{\boldsymbol{\Omega}}_K^{(0)}$ (based on model-based and/or ensemble initialization strategies) to define $\mathbf{P}_k = f(\hat{\boldsymbol{\Omega}}_k^{(0)})$, with $f : \mathbb{S}_+^p \to \mathbb{S}^p$ a function from the space of positive semi-definite matrices to the space of symmetric matrices of dimension p. Two viable options for defining $f(\cdot)$ are briefly described hereafter.

Option 1: $f(\cdot)$ via inversely weighted sample precision matrices

The first proposal for defining $\mathbf{P}_k$ is as follows:

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

where $P_{k,ij}$, $\hat{\Omega}_{k,ij}^{(0)}$ are respectively the (i,j)-th elements of the matrices $\mathbf{P}_k$ and $\hat{\boldsymbol{\Omega}}_k^{(0)}$. Intuitively, an high $|\hat{\Omega}_{k,ij}^{(0)}|$ value induces a deflation on the penalty enforced on the (i,j)-th element of $\boldsymbol{\Omega}_k$, whereas when $|\hat{\Omega}_{k,ij}^{(0)}|$ is close to 0 we are imposing an extra shrinkage on $\Omega_{k,ij}$. This strategy can be seen as a multiclass extension of the approach proposed in Fan et al. (2009).

Option 2: $f(\cdot)$ via distance measures in the $\mathbb{S}_+^p$ space

A second data-driven alternative involves setting $\mathbf{P}_k$ entries proportional to the distance between $\hat{\boldsymbol{\Omega}}_k^{(0)}$ and $\text{diag}\left(\hat{\boldsymbol{\Omega}}_k^{(0)}\right)$, where $\text{diag}\left(\hat{\boldsymbol{\Omega}}_k^{(0)}\right)$ is a diagonal matrix whose diagonal elements are equal to the ones in $\hat{\boldsymbol{\Omega}}_k^{(0)}$. Such a strategy mathematically reads as follows:

$$P_{k,ij} = \frac{1}{\mathcal{D}\left(\hat{\boldsymbol{\Omega}}_k^{(0)}, \text{diag}\left(\hat{\boldsymbol{\Omega}}_k^{(0)}\right)\right)}, \quad \forall i, j = 1, \ldots, p \quad \text{and} \quad i \neq j, \tag{3}$$

where with $\mathcal{D}(\cdot, \cdot)$ we identify a distance measure in the space of positive semi-definite matrices. Given the non-Euclidean nature of the $\mathbb{S}_+^p$ space several $\mathcal{D}(\cdot, \cdot)$ may be considered when defining (3): we subsequently employ Frobenius and Riemannian distances, but other options are at our disposal (see, e.g., Dryden et al., 2009).

The two above-described strategies for defining $f(\cdot)$ force entries corresponding to weaker sample conditional dependencies to be more strongly penalized. Once the definition of P_k has been established, coherently to Zhou et al. (2009), the model is estimated employing an EM algorithm where, in the M step, a graphical lasso strategy is adopted to compute $\boldsymbol{\Omega}_1, \ldots, \boldsymbol{\Omega}_K$ with $\lambda_{\text{gl}} = 2\lambda P_k / n_k^{(t)}$ with $n_k^{(t)}$ denoting the estimated sample size of the k-th cluster at the t-th iteration of the algorithm.

3 Application to handwritten digits recognition

The methodology presented in the previous section is employed to perform automatic handwritten digits recognition. The considered dataset is publicly available in the University of California Irvine Machine Learning data repository (`http://archive.ics.uci.edu/ml/datasets/optical+recognition+of+handwritten+digits`) and it contains $n = 5620$ handwritten samples of $K = 10$ digits. After having performed a preprocessing step to eliminate the near-zero variance pixels, we are left with $p = 47$ features onto which perform model based clustering. This translates to a challenging modeling task due to the narrow separation between classes and the high dimensionality of the parameter space. Indeed, a standard GMM with full precision matrices would require the estimation of $(K-1)+Kp+Kp(p-1)/2 = 11759$ parameters. We fit the penalized GMM methodology in (1) to the handwritten digits recognition dataset with different specification of P_ks: results are reported in Table 1. In details, $\mathbf{P}_k = \boldsymbol{J}$ identifies the original procedure of Zhou et al. (2009), with $\boldsymbol{J}$ the all-one matrix, while the remaining models describe the novel proposals of Section 2.

The penalized methods are able to shrink the estimates in a group-wise manner, recovering fairly well the underlying data partition. This is especially true in our proposals for which, even though the resulting Adjusted Rand Index Rand (1971) is not dramatically affected, the number of covariance parameters shrunk to 0 is digits-wise different thanks to the $\mathbf{P}_k$ specification. In Figure 1 we report the averaged images for digits 0, 5, and 9 and the estimated graphs in the precision matrices for the $\mathbf{P}_k$ via Riemannian distance approach. This method showcases the highest ARI and we can appreciate

Table 1 BIC, Adjusted Rand Index (ARI), number of estimated parameters in the precision matrices for different penalized model-based clustering methods and for digits 0, 5 and 9. Handwritten digits dataset.

	BIC	ARI	$d_{\boldsymbol{\Omega}}$	d_0	d_5	d_9
$\mathbf{P}_k = \boldsymbol{J}$	-388862	0.6837	4914	701	989	1271
$\mathbf{P}_k$ as in (2)	-368604	0.6820	3436	535	651	721
$\mathbf{P}_k$ as in (3), Frobenius distance	-391359	0.6827	6066	771	1003	1041
$\mathbf{P}_k$ as in (3), Riemannian distance	-388902	0.6841	5206	723	1059	1295

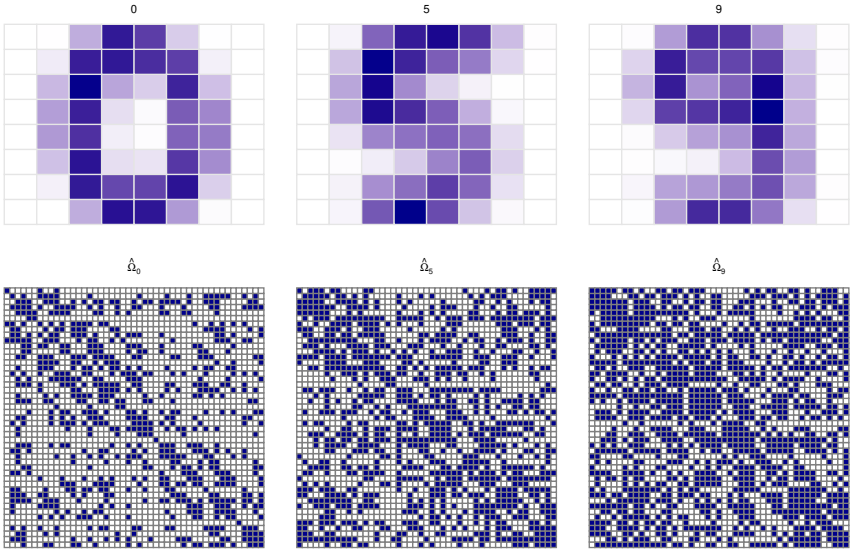

Fig. 1 Averaged images for digits 0, 5, and 9 and estimated graphs in the precision matrices for the $\mathbf{P}_k$ via Riemannian distance approach. Dark blue squares denote the presence of an edge between the two variables. Handwritten digits dataset.

how the number of estimated non-zero entries appreciably differ between the selected digits.

4 Conclusion and discussion

In this work we have proposed an extension to the approach outlined in Zhou et al. (2009). Two different procedures have been suggested to account for under or over-connected sparsity patterns in the precision matrices within a model-based clustering framework. The first solution provides an entry-wise inflation/deflation on the common penalty factor, while the second relies

on distance metrics in the space of positive semi-definite matrices to determine group-wise adjustments to the overall shrinkage term. An experiment on handwritten digits recognition has demonstrated the promising applicability of the devised procedure.

A direction for future research involves the development of a flexible mix-and-match methodology in which the penalization could interchangeably be applied to sparse precision and/or covariance matrices (Bien and Tibshirani, 2011). Such a framework, coupled with a penalty in the component means, can ultimately be employed to discard variables irrelevant for the clustering: ideas are being explored and they will be the object of future work.

References

Bien J, Tibshirani RJ (2011) Sparse estimation of a covariance matrix. Biometrika 98(4):807–820

Bouveyron C, Brunet-Saumard C (2014) Model-based clustering of high-dimensional data: A review. Computational Statistics and Data Analysis 71:52–78

Bouveyron C, Celeux G, Murphy TB, Raftery AE (2019) Model-Based Clustering and Classification for Data Science, vol 50. Cambridge University Press

Dryden IL, Koloydenko A, Zhou D (2009) Non-Euclidean statistics for covariance matrices, with applications to diffusion tensor imaging. The Annals of Applied Statistics 3(3):1102–1123

Fan J, Feng Y, Wu Y (2009) Network exploration via the adaptive LASSO and SCAD penalties. The Annals of Applied Statistics 3(2):521–541

Fop M, Murphy TB (2018) Variable selection methods for model-based clustering. Statistics Surveys 12:18–65

Friedman J, Hastie T, Tibshirani R (2008) Sparse inverse covariance estimation with the graphical lasso. Biostatistics 9(3):432–441

Rand WM (1971) Objective criteria for the evaluation of clustering methods. Journal of the American Statistical Association 66(336):846

Zhou H, Pan W, Shen X (2009) Penalized model-based clustering with unconstrained covariance matrices. Electronic Journal of Statistics 3:1473–1496