GROUP-WISE SHRINKAGE FOR MULTICLASS GAUSSIAN GRAPHICAL MODELS

Alessandro Casa ∗1, Andrea Cappozzo∗2, and Michael Fop1

1School of Mathematics & Statistics, University College Dublin
2Department of Mathematics, Politecnico di Milano

Abstract

Gaussian Graphical Models are widely employed for modelling dependence among variables. Likewise, finite Gaussian mixtures are often the standard way to go for model-based clustering of continuous features. With the increasing availability of high-dimensional datasets, a methodological link between these two approaches has been established in order to provide a framework for performing penalized model-based clustering in the presence of large precision matrices. Notwithstanding, current methodologies do not account for the fact that groups may possess different degrees of association among the variables, thus implicitly assuming similar levels of sparsity across the classes. We overcome this limitation by deriving group-wise penalty factors, automatically enforcing under or over-connectivity in the estimated graphs. The approach is entirely data-driven and does not require any additional hyper-parameter specification. Simulated data experiments showcase the validity of our proposal.

Keywords: Model-based clustering, Penalized likelihood, Sparse precision matrices, Gaussian graphical models, Graphical lasso, EM algorithm

1 Introduction

In their recent work, Gelman and Vehtari (2020) include regularized estimation procedures among the most important contributions to the statistical literature in the last fifty years. In fact, technological advancements and the booming of data complexity, both from a magnitude and from a structural perspective, have fostered the development of complex models, often involving an increasingly large number of parameters. Different regularization strategies have been proposed to obtain good estimates and predictions in these otherwise troublesome settings, usually implemented via penalized estimation procedures.

In this framework, a considerable amount of effort has been put into the estimation of sparse structures (see Hastie et al., 2015, for a thorough review). Broadly speaking, the rationale underlying the sparsity concept assumes that only a small subset of parameters of a given statistical model is truly relevant. As a consequence, sparse procedures usually include regularization terms in the objective function to optimize, forcing the estimates of some parameters to be exactly equal to zero. These machineries generally lead to an improvement in terms of interpretability and stability of the results, as well as to advantages from a computational perspective, while simultaneously reducing the risk of falling into the overfitting trap.

∗These authors contributed equally to this work.

1Corresponding author: School of Mathematics and Statistics, University College Dublin, Belfield, Dublin 4, Ireland. Email: alessandro.casa@ucd.ie
Sparse modelling has been proven useful in a lot of different settings, spanning from regression to classification. Recently, a fruitful line of research has focused on the application of these strategies to estimate inverse covariance, or precision, matrices in high-dimensional scenarios (Pourahmadi, 2013). Estimation of such matrices may be difficult, if not impossible, when the number of variables is comparable or greater than the sample size. As a consequence, sparsity inducing procedures are particularly appealing as they allow to obtain estimates even when the number of features exceeds the number of observations. Moreover, they generally yield better results with respect to the non-penalized solutions when the dimension is large. Lastly, from an interpretative point of view, setting some elements to zero is particularly beneficial, since zeros in a precision matrix corresponds to pairs of variables being conditionally independent, given all the other variables. Note that these conditional dependencies can be conveniently represented by means of Gaussian Graphical Models (GGM, Whittaker, 1990), where the variables are seen as nodes of a graph, with edges connecting pairs of conditionally dependent ones.

Within this framework, the majority of the proposed procedures implicitly assumes that the observed sample has been drawn from a single multivariate Gaussian distribution. Nonetheless, in practical applications, observations might arise from distinct sub-populations, corresponding to observed classes or clusters, each of them characterized by a specific data generating mechanism with its own feature dependence patterns. As a consequence, the estimation of a single GGM turns out to be inappropriate and tools specifically conceived for the multiclass framework are needed. In this context, most of the existing proposals adopt a borrowing strength strategy, encouraging the estimates to be similar across classes. Nonetheless, purposely shrinking the differences among precision matrices can be inappropriate in supervised and unsupervised classification, since such approach might hinder groups discrimination and ultimately jeopardize the output of the analysis.

The study of discriminant analysis and/or clustering tools that automatically embed penalization schemes to sparsely represent the relationships among variables has been scarcely pursued in the literature. A work to mention is the one by Zhou et al. (2009) where a model-based approach is proposed. Here the intensity of the penalization imposed on the different precision matrices is common, thus implicitly assuming that the expected number of connections among features is similar across classes. As a direct consequence, the proposed method may fail to properly characterize the unknown data partition and underlying group-specific GGMs in those settings where sparse intensity are cluster-wise different. For example, the method can be inappropriate to classify subjects affected by autism spectrum disorder that might present under or over-connected fMRI networks with respect to control individuals (see Hull et al., 2017, for a review on the topic).

A first viable solution would consist in considering class-specific penalization intensities. While reasonable, this approach would imply a rapidly increasing computational burden and it becomes impractical with a moderate number of classes. In this work, taking our step from Bien and Tibshirani (2011) where the single class covariance estimation issue is considered, we introduce a generalization of the method by Zhou et al. (2009) which may be consequently seen as a particular case of our proposal. Even if considering a single penalization parameter, thus avoiding the troublesome selection of more shrinkage terms, our approach turns out to be more flexible and adaptive since it penalizes a class-specific transformation of the precision matrices rather than directly the matrices themselves. In such a way, we are able to encompass under or over-connectivity situations as well as scenarios where the GGMs share similar structures among the groups. As it will be clear in the next sections, our proposal can be effectively considered in the supervised scenario as well as in the semi-supervised one; nonetheless, in the following we will mainly focus on the clustering framework as it introduces more challenges.

The rest of the paper is structured as follows. In Section 2 we briefly introduce Gaussian Graphical Models and the sparse estimation procedures proposed in this framework both from a single class and from a multi-class perspective. Afterwards, we motivate and present our pro-
positional framing it in a classification framework both from a supervised and from an unsupervised standpoint. In Section 3, we assess the performances of the proposed method on some synthetic data. Lastly, the paper ends with a brief discussion in Section 4.

2 Methods

2.1 Preliminaries

Consider a sample of \( n \) observations \( x_i \in \mathbb{R}^p \), for \( i = 1, \ldots, n \), independent and identically distributed as \( \mathcal{N}(\mu, \Sigma) \), with \( \mu \in \mathbb{R}^p \) and \( \Sigma \) the \( p \times p \) covariance matrix. The precision matrix \( \Omega = \Sigma^{-1} \) embeds the conditional dependencies among the variables, whereby zero entries between pairs of variables imply that the two variables are conditionally independent given the other variables. In this framework, Gaussian graphical models are a convenient way to model and visualize the dependence structure among the features, defining a correspondence between a sparse \( \Omega \) and a graph, in which the variables are the nodes and with edges connecting only those being conditionally dependent (Whittaker, 1990).

In high-dimensional data settings when \( n \ll p \) (but not only), estimation of a sparse \( \Omega \) is often carried out resorting to a regularization framework based on the following objective function, which corresponds to a penalized log-likelihood:

\[
\ell(\Omega) = \log \det(\Omega) - \text{tr}(S\Omega) - \lambda \|\Omega\|_1,
\]

where \( S \) is the sample covariance matrix. The penalty term \( \lambda \|\Omega\|_1 \) is the standard lasso penalty (Tibshirani, 1996), with \( \|A\|_1 = \sum_{ij} |A_{ij}| \) the element-wise \( L_1 \) norm, and \( A_{ij} \) denoting the \((i,j)\)-th element of \( A \); \( \lambda \geq 0 \) is a regularization parameter. For a suitable value of \( \lambda \), when maximizing (1), some of the off-diagonal elements of the precision matrix are shrunk to zero, thus providing a sparse estimate of \( \Omega \). In this way, the regularization allows to obtain an estimate for \( \Omega \) even when \( p \gg n \) and it helps the interpretation of the relationships among features by means of the related graphical model (Hastie et al., 2015). This framework defines the graphical lasso problem, and different approaches to solve it have been proposed in the literature (see e.g. Banerjee et al., 2008; Friedman et al., 2008; Scheinberg et al., 2010; Witten et al., 2011).

The standard approach delineated above assumes that the observations come from a single population, adequately described by a single multivariate Gaussian distribution indexed by a unique precision matrix, thus a single GGM. However, this assumption does not hold in a lot of different applications where the observed data arise from \( K \) different sub-populations, which might be characterized by different association patterns and GGMs. Such scenarios require the estimation of \( K \) class-specific GMMs, defined in terms of a collection of sparse precision matrices \( \Omega_1, \ldots, \Omega_K \). In the case of multiple populations, the matrices \( \Omega_k \) are usually considered to have possible commonalities and shared macro sparsity patterns. Several works introduced modifications to the penalization in (1) to induce the estimated GGMs to be similar while allowing for structural differences, see e.g. Guo et al. (2011); Mohan et al. (2014); Danaher et al. (2014); Lyu et al. (2018).

The multi-class approaches mentioned above have often been considered with an explorative aim in mind, to obtain a parsimonious and interpretable characterization of the relationships among the variables within and between classes. Nonetheless, sparse estimation of the precision matrices can be fruitfully embedded within a probabilistic classification framework, for both supervised classification and clustering purposes. In fact, sparsity in the \( \Omega_k \)s might be helpful when dealing with high-dimensional data by inducing parsimonious representations of the data generative mechanisms. Furthermore, class-specific relationships between the observed features, adequately reflected in differences among the corresponding precision matrices, might be helpful in driving the classification. Here, to some extent, the approaches borrowing strength across classes, thus encouraging similarity among groups, may be inappropriate if not harmful as they
might hinder the classification task itself. This particularly holds in the case of clustering, where
the classes are not readily available and need to be inferred from the data.

By now, the use of penalized estimation and shrinkage methods in a clustering context has
been quite scantily explored. The works by Hao et al. (2018), Gao et al. (2016), Huang et al.
(2017), and Zhou et al. (2009) connect sparse GGMs and clustering based on Gaussian mixture
models, falling under the framework of model-based clustering (see Bouveyron et al., 2019, for
a recent review on this topic). In their general formulation, these approaches place sparsity-
inducing penalties on mean and inverse covariance parameters of the Gaussian components. The
penalty on the cluster-specific means is employed to perform variable selection, while the penalty
on the component precision matrices is for regularization and sparse estimation of association
parameters in high-dimensional settings. Since the main focus of this work is on characterizing
the penalization effect on the estimation of sparse dependence structures in a clustering scenario,
for ease of exposition we consider no penalization on the mean parameters. In this case, following
the seminal work by Zhou et al. (2009), parameter estimation and the subsequent clustering step
are typically carried out by maximizing the following penalized log-likelihood:

$$
\ell_P(\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} \|\Omega_k\|_1,
$$

(2)

where the first term corresponds to the log-likelihood of a Gaussian mixture, while the second
one is the graphical lasso penalty function applied to all component-specific precision matrices.
Here $K$ represents the number of clusters, $\phi(\cdot)$ is the multivariate Gaussian density function,
and $\Theta = \{\pi_k, \mu_k, \Omega_k\}_{k=1}^{K}$ is the whole set of mixture parameters. Maximization of (2)
is difficult, since some strategies to estimate sparse precision matrices do not guarantee positive-definiteness,
that is required to compute a valid log-likelihood. In order to overcome this issue, the authors
note that, even without knowing the class membership of the observations, the estimation of the
component-wise precision matrices can be carried out similarly to what is done in the standard
GGM framework. More specifically, an EM algorithm (Dempster et al., 1977) is employed to
obtain the maximum penalized likelihood estimates for the model parameters. In the algorithm,
it turns out that the objective function to be maximized with respect to $\Omega_k$ (for $k = 1, \ldots, K$)
in the M step is related to the penalized log-likelihood in (1), where quantities are weighted
by the cluster posterior probability estimates at a given iteration. Therefore, a graphical lasso
strategy can be adopted with the penalization parameter in (1) being replaced by the scaled
coefficient $2\lambda/n_k^{(t)}$, where $n_k^{(t)}$ denote the estimated expected $k$-th cluster sample size at iteration
$t$ of the algorithm. While further details about the estimation process can be found in Zhou
et al. (2009), we focus here on how to obtain cluster-specific sparse precision matrices to account
for cluster-wise distinct degrees of sparsity.

2.2 Proposal

All the multi-class GGM estimation strategies introduced in Section 2.1 assume that different
classes are characterized by similar structure in the precision matrices, either explicitly or
implicitly. This assumption is explicit for those approaches where similarities among the precision
matrices are encouraged by the considered penalty term. On the other hand, in the approach
proposed by Zhou et al. (2009) the assumption is implicitly entailed by the use of a single pe-
nalization parameter $\lambda$. The adopted penalization scheme, even if somehow weighted by the
clusters sample sizes as shown at the end of Section 2.1, can be profitably considered only in
those situations where the number of non-zero entries in the precision matrices is similar across
classes. Therefore these approaches do not contemplate under or over-connectivity scenarios,
where the groups are characterized by strongly different amounts of sparsity. This constitutes a
serious limitation in those applications where different degrees of connectivity could ultimately
characterize the resulting data partition.
Figure 1: On the top: true data generating component-specific precision matrices. Black squares denote the presence of an edge between the two variables in the corresponding GGM. On the bottom: $F_1$ score behavior, as a function of the parameter $\lambda$.

To better explain this issue, here we provide an illustrative example where we simulate $n = 200$ $p$-dimensional observations with $p = 20$ from a mixture with $K = 2$ Gaussian components and mixing proportions $\pi = (0.5, 0.5)$. The considered component precision matrices are associated with the two sparse at random structures depicted in the top panel of Figure 1. The first component is characterized by an almost diagonal precision matrix while the second presents a dense structure, thus mimicking a scenario where the connectivity is drastically different among the two classes. We estimate these matrices employing the penalization scheme in (2), with $\lambda$ ranging over a suitable interval. The ability in recovering the association patterns inherent to the two clusters is evaluated by means of the $F_1$ score:

$$F_1 = \frac{tp}{tp + 0.5(fp + fn)},$$

where $tp$ denotes the correctly identified edges (i.e., the number of non-zero entries in the precision matrix correctly estimated as such), while $fp$ and $fn$ represent respectively the number of incorrectly identified edges and the number of missed edges (i.e. the number of non-zero entries wrongly shrunk to 0). Line plots displaying the $F_1$ patterns for the two components are reported in the bottom panel of Figure 1. A trade-off is clearly visible, indicating how a common penalty term prevents the possibility to obtain a proper estimation of both the precision matrices. In fact, if for the second component a mild penalization might be adequate to estimate the dense dependence structure, for the first component the high degree of sparsity is recovered only when a stronger penalty is considered.

This illustrative example shows clearly how the simple penalization strategy proposed by Zhou et al. (2009) does not represent a suitable solution when dealing with unbalanced class-specific degree of connections among the variables. A possible alternative would consist in un-
coupling the precision matrices estimation by considering the following penalized log-likelihood:

$$\ell_P(\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Omega_k) - \sum_{k=1}^{K} \lambda_k \|\Omega_k\|_1. \quad (4)$$

The increased flexibility of this approach is then obtained by considering component specific penalization coefficients. While in principle reasonable, the introduction of $K$ different penalties may be problematic, since in the graphical lasso framework little guidance is provided on how to tune these parameters. Time consuming grid searches are generally considered, with the optimal parameter selected according to the values of information criteria computed on the corresponding model or resorting to computational intensive cross-validation strategies. The simultaneous presence of $K$ penalty parameters would make this approach more complex, also from a computational perspective.

In this work we propose to carry out parameter estimation by maximizing a penalized log-likelihood function defined as follows:

$$\ell_P(\Theta) = \sum_{i=1}^{n} \log \sum_{k=1}^{K} \pi_k \phi(x_i; \mu_k, \Omega_k) - \lambda \sum_{k=1}^{K} \|P_k \ast \Omega_k\|_1 \quad (5)$$

where $\ast$ denotes the element-wise product between two matrices and the $P_k$ are weighting matrices employed to scale the effect of the penalty. The focus is then shifted towards the specification of $P_1, \ldots, P_K$, which, when properly encoding information about class specific sparsity patterns, it introduces a degree of flexibility that accounts for under or over-connectivity scenarios. Hereafter, we describe two data-driven procedures for inferring $P_k$s by means of carefully initialized sample precision matrices $\hat{\Omega}_k^{(0)}, \ldots, \hat{\Omega}_K^{(0)}$. Our proposals rely on the definition of a functional $f: \mathbb{S}^p_+ \to \mathbb{S}^p$, where $\mathbb{S}^p_+$ and $\mathbb{S}^p$ respectively denote the space of positive semi-definite and symmetric matrices of dimension $p$, such that $P_k = f(\hat{\Omega}_k^{(0)})$. Recommendations on how to compute $\hat{\Omega}_k^{(0)}$, and how to subsequently define $P_k$, $k = 1, \ldots, K$ will be the object of the next subsections.

### 2.2.1 Initializing the sample precision matrices $\hat{\Omega}_k^{(0)}$

The definition of a proper strategy to initialize the matrices $\hat{\Omega}_k^{(0)}$, for $k = 1, \ldots, K$, strongly depends on the framework and on the specific task of the analysis. In fact, in supervised and semi-supervised scenarios where the class labels are known for at least a subset of observations, the initialization step turns out to be straightforward. More formally, let us denote with $n$ the number of observations in the training set, with $m$ the number of observations with known label where $m = \sum_{k=1}^{K} m_k$ with $m_1, \ldots, m_K$ denoting the class specific sample sizes and $m_k > 0$, for $k = 1, \ldots, K$. In the supervised setting, where $m = n$, and in the semi-supervised one, where $0 < m < n$, we simply define $\hat{\Omega}_k^{(0)}$ to be the $k$-th class sample precision matrix estimated on the pertaining $m_k$ observations. Note that, if $p > m_k$, the initialized sample precision matrix might be obtained by means of $K$ graphical lasso algorithms.

On the other hand, in a clustering framework where $m = 0$, the specification of $\hat{\Omega}_k^{(0)}$ is more tricky. The absence of clear indications about the group membership of the observations makes the approach introduced for the supervised and semi-supervised context impractical. Nonetheless, it is possible to find suitable workarounds in order to employ our proposal even in an unsupervised scenario. From a practical point of view we consider a general multi-step procedure as follows:

1. Run any clustering algorithm on the observed data $X = \{x_1, \ldots, x_n\}$, to obtain an initial partition $Z = \{z_1, \ldots, z_n\}$ with $z_i = k$ if the $i$-th observation belongs to $k$-th cluster having cardinality equal to $n_k$;
2. Given the obtained partition \( Z \), estimate the cluster specific precision matrices \( \hat{\Omega}_1^{(0)}, \ldots, \hat{\Omega}_K^{(0)} \) using only those observations assigned to that specific group; again \( \hat{\Omega}_k^{(0)} \) might be obtained as the sample estimate when \( p < n_k \) or as the graphical lasso solution if \( p \geq n_k \).

When resorting to this procedure, the choice of which clustering strategy to adopt to obtain \( Z \) is subjective and needs to be discussed. In fact, inadequate choices might provide incoherent indications of the true clustering structure and hinder the possibility to subsequently recover it and to obtain an accurate reconstruction of the dependencies among the variables when maximizing (5). In principle, different clustering strategies may be adopted, and providing specific suggestions about the more adequate ones is beyond the scope of this work. Nonetheless, model-based techniques (see e.g. Fraley and Raftery, 2002) might constitute a clever choice, being them coherent with the framework considered here. Also ensemble strategies can be adequate, as they aim to combine the strengths of different algorithms and lessen the impact of some otherwise cumbersome choices (see Russell et al., 2015; Wei and McNicholas, 2015; Casa et al., 2020, for some proposals from a model-based clustering standpoint). Lastly, powerful initialization strategies for partitioning the data into \( K \) groups can as well appropriately serve the purpose (Scrucca and Raftery, 2015).

### 2.2.2 Obtaining \( P_k \) via inversely weighted sample precision matrices

The first viable proposal for defining \( P_k \) reads as follows:

\[
P_{k,ij} = \frac{1}{|\hat{\Omega}_{k,ij}^{(0)}|},
\]

where \( P_{k,ij}, \hat{\Omega}_{k,ij}^{(0)} \) are respectively the \((i,j)\)-th elements of the matrices \( P_k \) and \( \hat{\Omega}_k^{(0)} \). Notice that it is sufficient to set \( P_{k,ii} = 0, \forall i = 1, \ldots, p \), whenever the diagonal entries of \( \Omega_k \) shall not be penalized. Intuitively, with (6) we are inflating/deflating the penalty enforced on the \((i,j)\)-th element of the matrix \(\Omega_k\) according to the magnitude of \(\hat{\Omega}_{k,ij}^{(0)}\). Clearly, values of \(|\hat{\Omega}_{k,ij}^{(0)}|\) close to 0 induce a higher penalty on \(\Omega_{k,ij}\). This strategy shares connections with the proposal by Bien and Tibshirani (2011) developed in a covariance estimation context, and it might be seen as a multiclass extension of the approach proposed in Fan et al. (2009), where the adaptive lasso (Zou, 2006) is generalized to the estimation of sparse precision matrices.

An hard-thresholding version of (6) may also be considered:

\[
P_{k,ij} = \begin{cases} 
\frac{1}{|\hat{\Omega}_{k,ij}^{(0)}|} & \text{if } |\hat{\Gamma}_{k,ij}^{(0)}| < \gamma, \\
0 & \text{otherwise},
\end{cases}
\]

where \(\hat{\Gamma}_{k,ij}^{(0)}\) is the initialized partial correlation matrix for the \(k\)-th class, and \(0 \leq \gamma \leq 1\) is a positive threshold to be specified. This approach is related to the fixed-zero problem of Chaudhuri et al. (2007): when \(\lambda\) is sufficiently large, it leads to an estimate equivalent to that one obtained from a given association graph where the zero entries correspond to partial correlations of magnitude lower than the specified threshold \(\gamma\). The idea is somewhat connected to the thresholding operator for sparse covariance matrix estimation (Bickel and Levina, 2008; Pourahmadi, 2013) and has been explored in the \texttt{covglasso} R package (Fop, 2020). The hard-thresholded approach is not further considered in the following as it requires a sensitive choice of \(\gamma\), whereas the other proposals do not rely on any hyper-parameter specification.

### 2.2.3 Obtaining \( P_k \) via distance measures in the \( \mathbb{S}_+^p \) space

An alternative approach consists in setting the elements of \( P_k \) proportional to the distance between \(\hat{\Omega}_k^{(0)}\) and \(\text{diag} \left( \hat{\Omega}_k^{(0)} \right)\), where \(\text{diag} \left( \hat{\Omega}_k^{(0)} \right)\) indicates a diagonal matrix whose diagonal
elements are equal to the ones in $\hat{\Omega}_k^{(0)}$. We propose to compute $P_{k,ij}$ as follows:

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

with $D(\cdot,\cdot)$ being a suitably chosen measure of distance between positive semi-definite matrices. Since the $S_p^+$ space is non-Euclidean, several distances may be employed: the reader is referred to Dryden et al. (2009) for a thorough dissertation on the topic. Alternatively, to account for the diagonal elements in defining $P_k$, one may employ the following quantity:

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

where $I_p$ denotes the identity matrix of dimension $p$. The definitions of (8) and (9) simply stems from the conjecture that the “closer” $\hat{\Omega}_k^{(0)}$ is to a diagonal matrix, the higher the group-wise penalty shall be, thus forcing some of the entries in $\Omega_k$ to be shrunk to 0.

### 2.2.4 Obtaining $P_k$: methods comparison

The above-mentioned strategies share the same underlying rationale as they aim to impose stronger penalization on those entries of the precision matrices corresponding to weaker sample conditional dependencies among variables. Moreover, being the specification class specific, they fruitfully encompass situations where one or more groups present different sparsity patterns and magnitudes. While in principle the solution in Section 2.2.2 induces an entry-wise different penalty, it heavily depends on the estimation of $\hat{\Omega}_k^{(0)}$. Therefore, in those situations when the reliability of the sample estimates for the group precision matrices is difficult to assess, it might be convenient to let $P_k$ depend on a group specific constant, as reported in Section 2.2.3.

The proposed approaches generalize the one by Zhou et al. (2009), as the strategies coincide when $P_k$ is set to be equal to a matrix of ones, for all $k = 1,\ldots,K$. Once the definition of $P_k$ has been established, coherently to Zhou et al. (2009), the model is estimated employing an EM algorithm where the M step adopts a graphical lasso strategy to compute $\Omega_1,\ldots,\Omega_K$ with the penalty term $2\lambda P_k/n_k^{(t)}$, where $n_k^{(t)}$ as defined at the end of Section 2.1. Routines for doing so have been implemented in R (R Core Team, 2020), and the source code is freely available at https://github.com/AndreaCappozzo/sparsemix. Promising results are obtained when performing penalized model-based clustering with $P_k$ defined as in (8) and (9), as reported in the next Section.
3 Numerical illustration

3.1 Experimental setup

We illustrate via numeric experiments how the proposed procedures are effective in recovering the true group-wise conditional structure in a multi-class population. We generate $n = 1500$ observations from a Gaussian mixture distribution with $K = 3$ components, with the precision matrices $\Omega_k$ having various sparsity patterns and embedding different association structures. Three different scenarios are considered:

- **Equal proportion of edges in $\Omega_k$:** for each repetition of the simulated experiment, the precision matrices $\Omega_k$ are generated according to a sparse at random Erdős-Rényi graph structure (Erdős and Rényi, 1960) characterized by the same probabilities of connection, equal to 0.5. The number of variables is $p = 20$.

- **Different proportion of edges in $\Omega_k$:** for each repetition of the simulated experiment, the precision matrices $\Omega_k$ are generated according to a sparse at random Erdős-Rényi graph structure, with probabilities of connection equal to 0.1, 0.8 and 0.4 for $\Omega_1$, $\Omega_2$ and $\Omega_3$ respectively. The number of variables is $p = 20$.

- **High-dimensional and different proportion of edges in $\Omega_k$:** for each repetition of the simulated experiment, the precision matrices $\Omega_k$ are generated according to a sparse at random Erdős-Rényi graph structure with different probabilities of connection as per the previous scenario. The number of variables is $p = 100$.

In all scenarios, we take equal mixing proportions $\pi_k = 1/3$, $k = 1, 2, 3$ and mean vectors equal to:

$$
\mu_1 = -1.5e_{20}, \quad \mu_2 = 0e_{20}, \quad \mu_3 = 1.5e_{20},
$$

for the first two scenarios, while

$$
\mu_1 = 5e_{100}, \quad \mu_2 = 0e_{100}, \quad \mu_3 = 5e_{100},
$$

for the high dimensional case, with $e_{20}$ and $e_{100}$ identifying the all-ones vector in $\mathbb{R}^{20}$ and $\mathbb{R}^{100}$, respectively. Such parameters induce a moderate degree of overlapping between components in the lower-dimensional case, whilst producing well-separated clusters in the high-dimensional scenario. We point out again that given that the primary objective of the study is assessing the recovering of the true underlying sparse precision matrices, we do not impose any regularization on the mean parameters. An example of resulting graph structures from the first two scenarios are reported in Figure 2 and 3 respectively. We repeat the experiment $B = 100$ times, and for each repetition we fit model in Equation (5) computing $P_k$ as follows:

![Figure 3: Example of simulation setting Different proportion of edges in $\Omega_k$. Black squares denote the presence of an edge between the two variables.](image-url)
Zhou et al. (2009): $P_k$ set equal to the all-one matrix for $k = 1, 2, 3$

- $P_k$ via Frobenius distance in $S^+_p$: $P_k$ computed as in (8), with $D(\cdot, \cdot)$ the Frobenius distance in the $S^+_p$ space,

- $P_k$ via Riemannian distance in $S^+_p$: $P_k$ computed as in (8), with $D(\cdot, \cdot)$ the Riemannian distance in the $S^+_p$ space,

- $P_k$ via inversely weighted $|\hat{\Omega}_k^{(0)}|$: $P_k$ computed as in (6).

A grid of equispaced 100 elements for the penalty term $\lambda$ is considered, with lower and upper extremes set to:

$$\left[0, \max_k \left\{ \max \left\{ |\hat{S}_k^{(0)} - I_p|, \frac{\hat{n}_{k}^{(0)}}{2} \right\} \right\} \right],$$

where the inner max operation is taken element-wise, and with $\hat{S}_k^{(0)}$ and $\hat{n}_{k}^{(0)}$ the starting estimates of the sample covariance matrices and their associated sample sizes, initialized via the patterned Gaussian Mixture Model provided by the mclust software (Scrucca et al., 2016). Other initialization strategies are clearly possible, as described in Section 2.2.1. Notice that the upper term in (10) forces the final estimates $\hat{\Omega}_k$ to be approximately diagonal when $P_k$ is equal to the all-ones matrix (Zhao et al., 2012). Models performance is evaluated via component-wise $F_1$ scores (see Equation 3), where the problem of matching the estimated clustering to the actual classification is dealt with by means of the matchClasses routine of the e1071 R package (Meyer et al., 2020). Simulation results are reported in the next subsection.
3.2 Simulation results

Results for the simulated experiments are summarized in Figure 4, depicting smoothed lines plots of the $F_1$ scores for the estimated $\Omega_k$, $k = 1, 2, 3$, under the three considered scenarios, varying fitting method and shrinkage factor $\lambda$. By visually exploring Figure 4, several interesting patterns emerge.

First off, it is immediately noticed that the models performance in the first scenario does not vary across components, with $F_1$ score trajectories for $\Omega_1$, $\Omega_2$, and $\Omega_3$ being pretty much the same. This is expected, as each precision matrix is generated with a probability of connection equal to 0.5. This is the “gold-standard” scenario for the method described in Zhou et al. (2009), since in principle a common $\lambda$ should be sufficient for achieving group-wise sparsity. Indeed, the highest $F_1$ values (around 0.7) are achieved by all proposals when small penalty values are considered. Notwithstanding, multiplying the common shrinkage term by a group-specific factor $P_k$ moderates the rapid decline in performance when $\lambda$ increases. Particularly, computing the $P_k$s as a function of the Frobenius distance between $\hat{\Omega}_k^{(0)}$ and $\text{diag}(\hat{\Omega}_k^{(0)})$ greatly down-weights the impact of the common penalty term, making the procedure less sensitive to the selection of $\lambda$.

The beneficial effect of a group-wise $P_k$ becomes apparent in the second scenario, where the true precision matrices possess dissimilar a priori probability of connection. $F_1$ trajectories are component-wise different: the almost diagonal $\Omega_1$ would require a higher shrinkage for recovering the underlying sparse graph structure, whereas the highly connected $\Omega_2$ is well-estimated when low values of $\lambda$ are considered. This congenital trade-off is mitigated by the $P_k$ factor, which adjusts for under or over-connectivity within the estimation process. In particular, every suggested approach succeeds in improving the strategy of Zhou et al. (2009), with $F_1$ patterns for our proposals almost always dominating the common shrinkage method. This behavior is intensified even further in the high-dimensional setting, where as soon as $\lambda$ increases, the proportion of incorrectly missed edges produces a huge drop in the $F_1$ score for the second component. As previously highlighted, the $P_k$ via Frobenius distance in $S^p_+$ is the solution enforcing the greatest discount on the common $\lambda$, greatly improving the results for $\Omega_2$ at the expense of overestimating the true number of edges for $\Omega_1$.

With Figure 4, we have demonstrated the overall superiority of our proposals with respect to a common penalty framework in group-wise recovering of sparse precision matrices. Nevertheless, when it comes to performing the analysis a single value of $\lambda$ must be chosen. Providing a solution on how to select the common penalty term goes beyond the scope of the present manuscript, and we therefore rely on previous results (Pan and Shen, 2007; Zou et al., 2007), which propose...
to select $\lambda$ by maximizing a modified version of the Bayesian Information Criterion (BIC):

$$BIC = 2 \log L(\hat{\Theta}) - d_0 \log(n),$$

(11)

where $\log L(\hat{\Theta})$ is the log-likelihood evaluated at $\hat{\Theta}$, obtained maximizing (5), and $d_0$ is the number of non-zero parameters. The resulting empirical distribution of the mean $F_1$ score, averaged over $\hat{\Omega}_1$, $\hat{\Omega}_2$ and $\hat{\Omega}_3$, for the $B = 100$ simulations in the three scenarios is reported in Figure 5. For each simulation and method, $\lambda$ providing the highest value of (11) has been selected. As expected, the overall results do not dramatically change in the equal proportion of edges in $\Omega_k$ case. On the other hand, for the other two scenarios, our proposals perform substantially better with respect to Zhou et al. (2009), especially in the high-dimensional case. None of the introduced methods for computing $P_k$ seems to outperform the others; nonetheless, it is clear that, whenever the degree of conditional dependence varies greatly across components, weighting the common penalty $\lambda$ by a group-specific factor improves the recovering of the group-wise different sparse structures.

4 Discussion

The present paper has highlighted the limitations of imposing a single penalty when performing regularized precision matrices estimation in a multiclass setting. We have argued that methods enforcing similarities in the graphical models across groups may not be adequate for classification since they have detrimental effect when it comes to groups discrimination, in particular in the case of clustering. Thus, we have focused our attention on the penalized model-based method with sparse precision matrices framework of Zhou et al. (2009), where class-specific differences are preserved. Nonetheless, this methodology does not account for situations in which a component displays under or over-connectivity with respect to the remaining ones. To this extent, we have proposed some procedures to incorporate group-specific differences in the estimation, enforcing a carefully initialized solution to drive the algorithm in under or over penalizing specific components. Numerical illustrations have confirmed the validity of our proposals. By means of our solutions we have achieved both group-wise flexibility in the precision matrices reconstruction and we have mitigated the impact the common shrinkage factor has in the overall sparse estimation.

Note that the primary aim of this work has been to bring to light a shortcoming of the available multiclass precision matrix estimation approach and to propose some viable solutions to overcome it. Nonetheless, the proposals we have discussed are not meant to be a panacea useful for all the possible settings as other plausible strategies might be fruitfully adopted. In this context, a work that is worth mentioning is the one by Fop et al. (2019), where the authors consider a purpose-dependent and group specific penalization terms in order to parsimoniously characterize the marginal association among the variables in a Gaussian mixture modelling setting.

Lastly, note that we believe that the specification of group-specific penalties should never leave aside prior information, subject-matter knowledge and the characteristics of the data to be analyzed, as their incorporation in the methodology can be strongly beneficial for the analysis.

References

Banerjee, O., Ghaoui, L. E., and d’Aspremont, A. (2008). Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. Journal of Machine Learning Research, 9:485–516.

Bickel, P. J. and Levina, E. (2008). Covariance regularization by thresholding. The Annals of Statistics, 36(6):2577 – 2604.
Bien, J. and Tibshirani, R. J. (2011). Sparse estimation of a covariance matrix. *Biometrika*, 98(4):807–820.

Bouveyron, C., Celeux, G., Murphy, T. B., and Raftery, A. E. (2019). *Model-based clustering and classification for data science: with applications in R*. Cambridge University Press.

Casa, A., Scrucca, L., and Menardi, G. (2020). Better than the best? Answers via model ensemble in density-based clustering. *Advances in Data Analysis and Classification*, pages 1–25.

Chaudhuri, S., Drton, M., and Richardson, T. S. (2007). Estimation of a covariance matrix with zeros. *Biometrika*, 94(1):199–216.

Danaher, P., Wang, P., and Witten, D. M. (2014). The joint graphical lasso for inverse covariance estimation across multiple classes. *Journal of the Royal Statistical Society: Series B (Methodological)*, 76(2):373.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22.

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

Erdős, P. and Rényi, A. (1960). On the evolution of random graphs. *Publications of the Mathematical Institute of the Hungarian Academy of Sciences*, 5(1):17–60.

Fan, J., Feng, Y., and Wu, Y. (2009). Network exploration via the adaptive lasso and scad penalties. *The Annals of Applied Statistics*, 3(2):521.

Fop, M. (2020). *covglasso: Sparse Covariance Matrix Estimation*. R package version 1.0.2.

Fop, M., Murphy, T. B., and Scrucca, L. (2019). Model-based clustering with sparse covariance matrices. *Statistics and Computing*, 29(4):791–819.

Fraley, C. and Raftery, A. E. (2002). Model-based clustering, discriminant analysis, and density estimation. *Journal of the American Statistical Association*, 97(458):611–631.

Friedman, J., Hastie, T., and Tibshirani, R. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.

Gelman, A. and Vehtari, A. (2020). What are the most important statistical ideas of the past 50 years? *arXiv preprint arXiv:2012.00174*.

Guo, J., Levina, E., Michailidis, G., and Zhu, J. (2011). Joint estimation of multiple graphical models. *Biometrika*, 98(1):1–15.

Hao, B., Wei Sun, W., Liu, Y., and Cheng, G. (2018). Simultaneous clustering and estimation of heterogeneous graphical models. *Journal of Machine Learning Research*, 18(217):1–58.

Hastie, T., Tibshirani, R., and Wainwright, M. (2015). *Statistical learning with sparsity: the lasso and generalizations*. CRC press.
Huang, F., Chen, S., and Huang, S.-J. (2017). Joint estimation of multiple conditional Gaussian graphical models. *IEEE transactions on neural networks and learning systems*, 29(7):3034–3046.

Hull, J. V., Dokovna, L. B., Jacokes, Z. J., Torgerson, C. M., Irimia, A., and Van Horn, J. D. (2017). Resting-state functional connectivity in autism spectrum disorders: A review. *Frontiers in psychiatry*, 7:205.

Lyu, Y., Xue, L., Zhang, F., Koch, H., Saba, L., Kechris, K., and Li, Q. (2018). Condition-adaptive fused graphical lasso (CFGL): An adaptive procedure for inferring condition-specific gene co-expression network. *PLoS computational biology*, 14(9):e1006436.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2020). *e1071: Misc Functions of the Department of Statistics, Probability Theory Group (Formerly: E1071)*, TU Wien. R package version 1.7-4.

Mohan, K., London, P., Fazel, M., Witten, D., and Lee, S. (2014). Node-based learning of multiple gaussian graphical models. *The Journal of Machine Learning Research*, 15(1):445–488.

Pan, W. and Shen, X. (2007). Penalized model-based clustering with application to variable selection. *Journal of Machine Learning Research*, 8:1145–1164.

Pourahmadi, M. (2013). *High-Dimensional Covariance Estimation*. Wiley Series in Probability and Statistics. John Wiley & Sons, Inc.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Russell, N., Murphy, T. B., and Raftery, A. E. (2015). Bayesian model averaging in model-based clustering and density estimation. *arXiv preprint arXiv:1506.09035*.

Scheinberg, K., Ma, S., and Goldfarb, D. (2010). Sparse inverse covariance selection via alternating linearization methods. In Lafferty, J., Williams, C., Shawe-Taylor, J., Zemel, R., and Culotta, A., editors, *Advances in Neural Information Processing Systems*, volume 23.

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5: Clustering, Classification and Density Estimation Using Gaussian Finite Mixture Models. *The R Journal*, 8(1):289–317.

Scrucca, L. and Raftery, A. E. (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Advances in Data Analysis and Classification*, 9(4):447–460.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1):267–288.

Wei, Y. and McNicholas, P. D. (2015). Mixture model averaging for clustering. *Advances in Data Analysis and Classification*, 9(2):197–217.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley.

Witten, D. M., Friedman, J. H., and Simon, N. (2011). New insights and faster computations for the graphical lasso. *Journal of Computational and Graphical Statistics*, 20(4):892–900.

Zhao, T., Liu, H., Roeder, K., Lafferty, J., and Wasserman, L. (2012). The huge package for high-dimensional undirected graph estimation in R. *Journal of Machine Learning Research*, 13:1059–1062.
Zhou, H., Pan, W., and Shen, X. (2009). Penalized model-based clustering with unconstrained covariance matrices. *Electronic Journal of Statistics*, 3:1473.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476):1418–1429.

Zou, H., Hastie, T., and Tibshirani, R. (2007). On the “degrees of freedom” of the lasso. *The Annals of Statistics*, 35(5):2173–2192.