Afrika Statistika

Afrika Statistika
Vol. 11(2), 2016, pages 1041–1059.
DOI: http://dx.doi.org/10.16929/as/2016.1041.91

ISSN 2316-090X

Sparse Gaussian graphical mixture model

Anani Lotsi[1,*] **and Ernst Wit**[2]

[1]University of Ghana Legon, Department of Statistics, P.O.Box: 115 LG
[2] Department of Statistics and Probability, University of Groningen Nijenborgh 9, 9747 AG The Netherlands Email: e.c.wit@rug.nl

Received November 1, 2016; Accepted December 14, 2016

Abstract. This paper considers the problem of networks reconstruction from heterogeneous data using a Gaussian Graphical Mixture Model (GGMM). It is well known that parameter estimation in this context is challenging due to large numbers of variables coupled with the degenerate nature of the likelihood. We propose as a solution a penalized maximum likelihood technique by imposing an l_1 penalty on the precision matrix. Our approach shrinks the parameters thereby resulting in better identifiability and variable selection.

Résumé. Cet article considère le problème de la reconstruction de réseaux à partir de données hétérogènes en utilisant le modèle graphique gaussien mémangé (GGMM en Anglais). Il est connu que l'estimation paramétrique, dans ce contexte, n'est pas aisé à cause du grand nombre de variable et de la nature dégénérée de la vraisemblance. Nous proposons comme une solution une méthode de pénalisation du maximum de vraisemblance en imposant une pénalité de type L1 sur la précision de la matrice. Notre méthode réduit les paramètres et ainsi aboutit à une meilleure identification et à un meilleur choix des variables.
.

Key words: Gaussian graphical mixture model; Expectation maximization algorithm; Graphical LASSO.
AMS 2010 Mathematics Subject Classification : Primary 62-09, Secondary 62H12, 62J07.

1. Introduction

Biologists aim to describe the dependency structure among large number of genes. This is often done without taking into consideration the heterogeneity nature of the samples.

*Corresponding author Anali Losti: ALotsi@ug.edu.gh
Ernst Wit : e.c.wit@rug.nl

By heterogeneity, we mean networks may be different for different subgroups of samples. Our population of individuals may come from several distinct subpopulations each with their own underlying dependency structure. However, typically little information is known about an individual's group membership. In this setting, parameters may change for different subgroups of observations. We want to model such heterogeneity and recover the underlying networks from such data with some sparsity constraint. The problem becomes more complex if the number of components that made up the population is unknown. Statistical methods for analyzing such data are subject to active research currently (Agakov *et al.*, 2012). Gaussian graphical mixture models (GGMM) are ways to model such data.

A Gaussian graphical model (GGM) for a random vector $Y = (Y_1, ..., Y_p)$ is a pair $(\mathbb{G}, \mathbb{P})$ where $\mathbb{G}$ is an undirected graph and $\mathbb{P} = \left\{N(\mu, \Theta^{-1})\right\}$ is the model comprising all multivariate normal distributions whose inverse covariance matrix or precision matrix entries satisfies $(u, v) \in \mathbb{G} \iff \Theta_{uv} \neq 0$. The conditional independence relationship among nodes are captured in the precision matrix Θ. Consequently, the problem of selecting the graph is equivalent to estimating the off-diagonal zero-pattern of the concentration matrix. Further details on these models as well interpretation of the conditional independency on the graph can be found in (Lauritzen, 1996).

Mixture distributions are often used to model heterogeneous data or observations supposed to have come from one of K different networks or components. Under Gaussian mixtures, each component is suitably modelled by a family of Gaussian probability density. This paper deals with the problem of structural learning in reconstructing the underlying graphical networks (using a graphical Gaussian model) from a data supposed to have come from a mixture of Gaussian distributions.

A natural way for parameter estimation in GMMs is via a maximum likelihood estimation. However some performance degradation is encountered owing to the identifiability of the likelihood and the high dimensional setting. To overcome these problems, Banfield *et al.* (1993) proposed a parameter reduction technique by re-parameterizing the covariance matrix through eigenvalue decomposition. In doing so, some parameters are shared across clusters. As a result of a continuous increasing number of dimensions, this approach can not totally alleviate the $(n << p)$ phenomena. Recently proposals to overcome the high dimensionality problem involve estimating sparse precision matrix. Among these proposals is the penalized log likelihood technique of Friedman *et al.* (2008), an L_1 regularization approach which encourages many of the entries of the precision matrix to be 0. Our method is based on this idea. The L_1 penalty promotes sparsity. We provide sufficient conditions for consistency of the penalized MLE.

In this work we propose a penalized likelihood approach in the context of Gaussian graphical mixture model, which constraints the networks to be sparse. The parameters in the networks are estimated by incorporating an existing Graphical LASSO (GLASSO) method for covariance estimation into an EM algorithm. In effect, we view each network as an instance of a particular GGM. Therefore we aim at recovering the underlining various networks from which the data originate from. Additionally, we assess how well the resultant graphs obtained through GLASSO relate to the true graphs and we provide consistency

results of the estimates. Throughout this article we assume K, the number of components of mixture models is known.

The reminder of this article is organized as follows. We introduce the model, set up the Penalized Maximum Likelihood Estimate (PMLE) approach and summarize the main result in connection with the consistency of the estimates obtained from the mixture model in section 2. We then proceed with the inference procedure through a penalized version of the EM algorithm in section 3. In section 4 we present some simulations and an example of applications to illustrate our results. We conclude with a brief discussion and future works in section 5.

2. Penalized maximum likelihood estimation

In this section we introduce the Gaussian Graphical Mixture Model, then we derive the penalized likelihood upon which statistical inference via the EM algorithm is based and prove consistency of the Penalized Maximum Likelihood Estimates (PMLE).

2.1. The Mixture model

Mixture models are very popular for the analysis of complex data. A mixture model represents the given data as a mixture of K networks or components, each of which has different characteristics. We introduce our model in Figure (1), where we assume a genetic population. We suppose sample of expression level of these genes comes from two different networks after observing their metabolism structure. We then fit two Gaussian distribution $N(\mu_1, \Theta_1)$ and $N(\mu_2, \Theta_2)$ for these clusters. Figure (1) represents the above via a mixture model. The question now is how can we infer the underlying networks from which the data come from?

Suppose we are given a training data set $\{\mathbf{Y}_1, ..., \mathbf{Y}_n\}$, assumed to be a random sample from K mixture components. Our model consists of assuming that the variable Z_i, describing which network an individual originates, is a multinomial random variable with parameters, π_k, denoting the mixture proportions or the mixing coefficients with $(0 < \pi_k < 1)$, $\sum_{k=1}^{K} \pi_k = 1$, and K is known. In essence

$$P(Z_i = k) = \pi_k.$$

We wish to model the data by specifying a joint distribution

$$P(\mathbf{Y}_i, Z_i) = P(\mathbf{Y}_i|Z_i)p(Z_i).$$

We model each subpopulation separately by assuming a GGM where $(\mathbf{Y}_i|Z_i = k) \sim N(\mu_k, \Sigma_k)$. Our model posits that each $\mathbf{Y}_i$ was generated by randomly choosing Z_i from $\{1, ...K\}$, or $\mathbf{Y}_i$ was drawn from one of the k Gaussian depending on Z_i.

In this work we assume that $\forall\ k$, $\mu_k = 0$. In practice, this means that the data is assumed to be normalized by subtracting the mean. Since $\mathbf{Y}_i$ is dependent on Z_i, we say that Z_i represents the class that produced $\mathbf{Y}_i$ and we know $\mathbf{Y}_i$ fully if we know which class Z_i

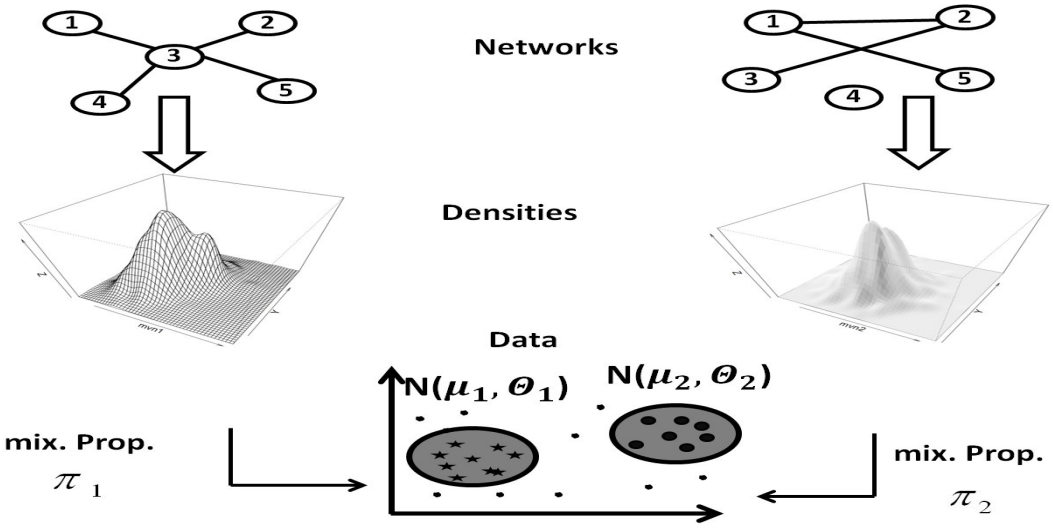

Fig. 1: Mixture models: we assume the data is composed of 2 separate mixtures of Gaussian (MOG), each with a corresponding graphical model or network.

falls. Also note that the Z_i's are latent random variables, meaning that they are hidden or unobserved. The density of each $\mathbf{Y}_i$ can be written as

$$\begin{aligned} f_\gamma(\mathbf{y}_i) &= \sum_{k=1}^{K} \pi_k \varphi_k(\mathbf{y}_i | Z_i = k) \\ f_\gamma(\mathbf{y}_i) &= \sum_{k=1}^{K} \pi_k \varphi_k(\mathbf{y}_i | \Theta_k) \end{aligned} \tag{1}$$

where $\varphi(\mathbf{y}_i|\Theta_k)$ denotes the density of Gaussian distribution with mean 0 and inverse covariance covariance matrix Θ_k; f_γ represents the "incomplete" mixture data density of the sample i.e $\mathbf{y} \sim f_\gamma$. We introduce the parameter set of mixture namely

$$\Omega = \left\{ \{\Theta_k\}_{k=1}^{K} | \Theta_k \succ 0, \quad k = 1, ..., K \right\},$$

$\Theta \succ 0$ indicates that Θ is positive-definite matrix, and

$$J = \left\{ \{\pi_k\}_{k=1}^{K} | \pi_k > 0, \quad k = 1, ..., K \right\}$$

and

$$\Gamma = \Omega \times J \tag{2}$$

denotes the parameter space with the true parameter defined as $\gamma_0 = (\Theta_0, \pi_0) \in \Gamma$.
In order to characterize the mixture model and estimate its parameters thereby recovering the underlying graphical structure from the data (seen as mixture of multivariate densities), several approaches may be considered. These approaches include graphical methods, methods of moments, minimum-distance methods, maximum likelihood (Ruan *et al.*, 2011; Zhou *et al.*, 2009) and Bayesian methods (Bernardo *et al.*, 2003; Biernacki *et al.*, 2000). In our case we adopt the penalized maximum likelihood method in a graphical model set up.

2.2. The penalized model-based likelihood

We can now write the likelihood of the incomplete data density as

$$L_{\mathbf{y}}(\gamma) = \prod_{i=1}^{n} \left[\sum_{k=1}^{K} \pi_k \varphi_k(\mathbf{y}_i | \Theta_k^{-1}) \right],$$

whose log-likelihood function is given by

$$l_{\mathbf{y}}(\gamma) = \sum_{i=1}^{n} \log f_\gamma(\mathbf{y}_i) \tag{3}$$

The goal is to maximize the log-likelihood in (3) with respect to γ. Unfortunately, a unique global maximum likelihood estimate does not exist because of the permutation symmetries of the mixture subpopulation; (Day, 1969; Lindsay *et al.*, 2006). Also the likelihood function of normal mixture models is not a bounded function on γ as was put forward by Kiefer *et al.* (1956). On the question of consistency of the MLE, Chanda (1954), Cramer (1946) focus on local ML estimation and mathematically investigate the existence of a consistent sequence of local maximizers. These results are mainly based on Wald's technique (Wald, 1949). Redner (1981) later extended these results to establish the consistency of the MLE for mixture distributions with restrained or compact parameter spaces. It was proved that the MLE exists and it is globally consistent in a compact subset $\hat{\Gamma}$ of Γ that contains γ_0; i.e

$$\text{given} \quad \hat{\gamma}_n | l_{\mathbf{y}}(\hat{\gamma}_n) = \max_{\gamma \in \Gamma} \quad l_{\mathbf{y}}(\gamma), \quad \hat{\gamma}_n \to \gamma_0 \text{ in probability, for n} \to \infty$$

In addition to the degenerate nature of the likelihood (Kiefer *et al.*, 1956) on the set Γ, the "high dimensional, low sample size setting"- where the number of observations n is smaller that the number of nodes or features p- is another complication. Estimating the parameters in the GGMM by maximizing criterion (3) is a complex one. The penalized likelihood-based method (Friedman *et al.*, 2008; Yuan *et al.*, 2007) is a promising approach to counter the degeneracy of $l_{\mathbf{y}}(\gamma)$ while keeping the parameter space Γ unaltered. However, to make the PMLE work, one has to solve the problem of what kind of penalty functions are eligible. We opt for a penalty function that prevents the likelihood from degenerating under the multivariate mixture model. We assume that the penalty function $P : \Gamma \to \mathbb{R}_0^+$ given by

$$P(\Theta) = \exp(-\lambda ||\Theta||_1),$$

satisfies:

$$\lim_{|\Theta_k|\to\infty} P(\Theta_k)|\Theta_k|^n = 0 \quad \forall k \in \{1, 2, ..., K\} \quad \forall n \tag{4}$$

where $\lambda > 0$ is a user-defined tuning parameter that regulates the sparsity level, $|\Theta|$ denotes determinant of Θ, and $||.||_1$ is the L_1 norm or the sum of absolute values of the entries of a matrix or a vector i.e $||\mathbf{X}||_1 = \sum_{i=1}^{n} |X_i|$.
This results in placing an L_1 penalty on the entries of the concentration matrices so that the resulting estimates are sparse and zeroes in these matrices correspond to conditional independency between the nodes similar to (Meinshausen *et al.*, 2006). Numerous advantages result from this approach. First of all, the corresponding penalized likelihood is bounded and the penalized likelihood function does not degenerate in any point of the closure of parameter space Γ and therefore the existence of the penalized maximum likelihood estimator is guaranteed. Next, in the context of GGM, penalizing the precision matrix results in better estimates and sparse models are more interpretable and often preferred in application.

We define the L_1 penalized log-likelihood as:

$$l_{\mathbf{y}}^{p}(\gamma) = l_{\mathbf{y}}(\gamma) - \lambda_n \sum_{k=1}^{K} ||\Theta_k||_1 \tag{5}$$

where $\lambda_n \propto \frac{\lambda}{\sqrt{n}}$, $||\Theta||_1 = \sum_{i,j} |\theta_{ij}|$, K is the number of mixing components assumed fixed. The hyperparameters K and λ determine the complexity of the model. The corresponding PMLE are defined as

$$\hat{\gamma}_{\lambda_n} = \arg\max_{\gamma} \quad l_y^p(\gamma) \tag{6}$$

Our method penalizes all the entries of the precision matrix including the diagonal elements. We do this in order to avoid the likelihood to degenerate. To see this, consider a special case of a model consisting of two univariate normal mixtures $\pi_1\varphi(\mathbf{y}|\sigma_1) + \pi_2\varphi(\mathbf{y}|\sigma_2)$. By letting $\sigma_1 \to 0$ with other parameters remaining constant, the log-likelihood tends to infinity for values of $y = 0$, i.e the log-likelihood degenerates due to mixture formulation whereby a single observation mixture component with a decreasing variance on top of the observation explodes the likelihood. For that matter an L_1 penalty which does not penalize the diagonal elements tend to result in a degenerate ML estimator especially when $n \to \infty$.

2.3. Consistency

At this stage we want to characterize the solution obtained in Equation (6). The general theorem concerning the consistency of the MLE (Redner, 1980; Wald, 1949) can be extended to cover our type of penalized MLE. This is because if a likelihood function which yields a strong consistent estimate over a compact set is given, then our L_1 penalty would not alter the consistency properties. Consistency of the PMLE is given in theorem 1. The latter uses results in (Wald, 1949) under the classical MLE over a compact set.

Before we present our result relating to the consistency of our PMLE, we summarize the corresponding MLE version in the following lemmas. First the following assumptions will be needed.

A1: There is a neighborhood ρ of γ_0 such that for all $\gamma \in \rho$; for almost all $\mathbf{y} \in R^n$; and for for l, j and $s = 1, ..., \upsilon$; $\frac{\partial f}{\partial \gamma_l}$, $\frac{\partial^2 f}{\partial \gamma_l \partial \gamma_j}$, $\frac{\partial^3 f}{\partial \gamma_l \partial \gamma_j \partial \gamma_s}$ exist and satisfy

$$\left|\frac{\partial f}{\partial \gamma_l}\right| < g_l(\mathbf{y}); \left|\frac{\partial^2 f}{\partial \gamma_l \partial \gamma_j}\right| < g_{lj}(\mathbf{y}); \left|\frac{\partial^3 f}{\partial \gamma_l \partial \gamma_j \partial \gamma_s}\right| < g_{ljs}(\mathbf{y}),$$

where g_l, and g_{lj} are integrable and $g_{ljs}(\mathbf{y})$ satisfies

$$\int_{R^n} g_{ljs}(\mathbf{y}) f_{\gamma_0}(\mathbf{y}) d\mathbf{y} < \infty.$$

A2: The matrix $\delta(\gamma) = \left(\int_{R^n} \frac{\partial \ln f}{\partial \gamma_l} \frac{\partial \ln f}{\partial \gamma_j} f d\mathbf{y}\right)$ is positive definite at γ_0.

Lemma 1. *If conditions A1 and A2 are satisfied, then, given any sufficiently small neighborhood ρ_0 of γ_0 with probability equals 1 as the sample size n approaches infinity, there is a unique solution to the likelihood equations in ρ_0 and this solution is an MLE.*

Lemma 1 indicates that, by restricting attention to a fixed neighborhood of γ_0, we have a unique and consistent solution to the likelihood equations.

The next lemma considers a situation where the likelihood is an unbounded function. For that one must assume a compact (closed and bounded) parameter space. It will be assumed that there is a σ- finite measure μ such that for each $\gamma \in \Gamma$ the probability measure μ_γ is absolutely continuous w.r.t. μ. We let $f_\gamma(\mathbf{y})$ denote any representative of the density of μ_γ w.r.t μ. The following assumptions are made in addition:

A3: The parameter space Γ is a closed and bounded subset of R^l for some positive number l. In particular, $T = \{(\Theta_1, ..., \Theta_K)\} \mid \quad s.t. \quad ||\Theta_k||_1 \leq M^*$ and $||\Theta_k||_2 \geq \epsilon^*, k = 1, ..., K$, for some positive number M^* and ϵ^*.

A4: Let $B_r(\gamma)$ be the closed ball of radius r about γ. Then for any positive real number r, let:

$$f_\gamma(\mathbf{y}, r) = \sup_{\eta \in B_r(\gamma)} f_\gamma(\mathbf{y}, \eta); \quad f^*_\gamma(\mathbf{y}, r) = max\left[1, f_\gamma(\mathbf{y}, r)\right].$$

Then for each γ and for sufficiently small r

$$\int \ln f^*_\gamma(\mathbf{y}, r) d\mu_{\gamma_0} < \infty.$$

A5:

$$\int \left|\ln f_{\gamma_0}(\mathbf{y})\right| d\mu_{\gamma_0} < \infty.$$

A6: if $\gamma_l \to \gamma$, then $f_{\gamma_l}(\mathbf{y}) \to f_\gamma(\mathbf{y})$.

Lemma 2. *Given assumptions (A3-A6), and let $C = \{\gamma \in \Gamma | f_\gamma(\mathbf{y}) = f_{\gamma_0}(\mathbf{y}) \quad almost\ everywhere\}$. If S is any open neighborhood containing C, then with probability equals 1, the MLE is eventually in S.*

The two lemmas show that the MLE converges to the set C. Since C is the set of all parameters for which the density is the true density, it may be said that the MLE converges strongly to the true set of parameters.

We then define two further conditions upon which our theorem 1 holds.

A7: Let $\bar{\Gamma}$ denotes the quotient topological space obtained from Γ and suppose that $\bar{\Gamma}$ is any compact subset containing γ_0.

A8:

$$\int |\ln f_l(\mathbf{y}, \gamma_l)| \, d\mu_{\gamma_j} < \infty \quad for \quad \gamma_l \in \Gamma_l \quad \text{and} \quad \gamma_j \in \Gamma_j.$$

Theorem 1. *Suppose that the mixing distributions satisfy conditions (A3-A8). Define* $|\gamma_0| = ||\pi_0||_2 + ||\Theta_0||_F$. *Suppose that* π_k *is bounded away from zero, and the penalty is set as* $\lambda_n \propto (1/\sqrt{n})$. *It follows that for a fixed p, the penalized likelihood solution* $\hat{\gamma}_{\lambda_n}$ *is consistent in the quotient topological space* $\bar{\Gamma}$, *i.e* $\forall \epsilon > 0$

$$\lim_{n\to\infty} P\left(|\hat{\gamma}_{\lambda_n} - \gamma_0| > \epsilon\right) = 0.$$

Proof.

$$\begin{aligned} P(|\hat{\gamma}_{\lambda_n} - \gamma_0| > \epsilon) &= P(|\hat{\gamma}_{\lambda_n} - \hat{\gamma} + \hat{\gamma} - \gamma_0| > \epsilon) \\ &\leq P(|\hat{\gamma}_{\lambda_n} - \hat{\gamma}| > \epsilon/2) + P(|\hat{\gamma} - \gamma_0| > \epsilon/2) \end{aligned}$$

Want to show:

$$\lim_{n\to\infty} P(|\hat{\gamma}_{\lambda_n} - \hat{\gamma}| > \epsilon/2) = 0 \tag{7}$$

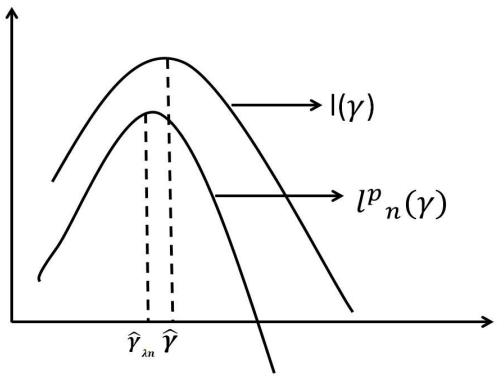

Fig. 2: Sketch of the likelihood and the penalized likelihood function

We have the followings:

$$l_n^p(\gamma) = l(\gamma) - \lambda_n ||\Theta||_1$$

with

$$\hat{\gamma}_{\lambda_n} = \arg\max_{\gamma} \quad l_n^p(\gamma)$$

and

$$\hat{\gamma} = \arg\max_{\gamma} \quad l(\gamma)$$

$\forall n > n_1$, we have

$$\frac{\partial^2 l}{\partial \gamma^2}(\hat{\gamma}) = O_p(-nM) \tag{8}$$

and suppose that l_n^p uniformlly converges to l.

We know that given $\epsilon/2$, we can find δ s.t $|\gamma - \hat{\gamma}| > \epsilon/2$

$$l(\gamma) < l(\hat{\gamma}) - \delta \tag{9}$$

We also know that $\exists n_2$ s.t $\forall \quad n > n_2$

$$l_n^p(\gamma) > l(\gamma) - \delta \tag{10}$$

Now let assume that

$$|\hat{\gamma}_{\lambda_n} - \hat{\gamma}| > \epsilon/2 \tag{11}$$

then from (9), we have

$$\begin{aligned} l^p(\hat{\gamma}_{\lambda_n}) &< l(\hat{\gamma}_{\lambda_n}) \\ &< l(\hat{\gamma}) - \delta \end{aligned} \tag{12}$$

But from (10),

$$l^p(\hat{\gamma}) > l(\hat{\gamma}) - \delta.$$

So $\hat{\gamma}_{\lambda_n}$ did not maximize l^p because at $\hat{\gamma}$, it is higher. Therefore assumption (11) is flase. This completes the proof

3. Penalized EM algorithm

In order to maximize the penalized likelihood function (5) we consider a penalized version of the EM algorithm of Dempster *et al.* (1997). To do that we first augment our data $\mathbf{Y}_i$ with $\mathbf{Z}_i$ so that the complete data associated with our model now becomes $\mathbf{C}_i = (\mathbf{Y}_i, \mathbf{Z}_i)$ and an EM algorithm iteratively maximizes, instead of the penalized observed log-likelihood $l_{\mathbf{y}}^p$ in (5), the conditional expectation of the penalized log-likelihood of the augmented data.

Suppose $\mathbf{c}_i \sim h_{\mathbf{c}_i}(\gamma)$, i.e $h_{\mathbf{c}_i}(\gamma)$ is the density of the augmented data $\mathbf{c}_i$. Now the penalized log-likelihood of the augmented data can be written as

$$
\begin{aligned}
l^p_{\mathbf{c}}(\gamma) &= \ln\left[h_{\mathbf{c}_i}(\gamma)\right] - \lambda \sum_{k=1}^{K} ||\Theta_k||_{l_1} \\
l^p_{\mathbf{c}}(\gamma) &= \sum_{i=1}^{n} \left(\ln \pi_k + \ln \varphi_k(\mathbf{y}_i|\Theta_k^{-1})\right) - \lambda \sum_{k=1}^{K} ||\Theta_k||_{l_1} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} 1_{\{Z_i=k\}} \left[\ln \pi_k + \ln \varphi_k(\mathbf{y}_i|\Theta_k^{-1})\right] - \lambda \sum_{k=1}^{K} ||\Theta_k||_{l_1} \qquad (13)
\end{aligned}
$$

The indicator function $1_{\{Z_i=k\}}$ simply says that if you knew which component the observation i came from, we would simply use its corresponding Θ_k for the likelihood. For illustration purpose, and suppose we have 3 observations and we are certain that the first two were generated by the Gaussian density $N(0, \Theta_2)$, and the last came from $N(0, \Theta_1)$. Then we write the full log-likelihood as follows:

$$
l_{\mathbf{c}}(\Theta) = l_{\mathbf{y}_1}(\Theta_2) + l_{\mathbf{y}_2}(\Theta_2) + l_{\mathbf{y}_3}(\Theta_1) \qquad (14)
$$

3.1. The E-step

From Equation (13) we compute the quantity $Q(\gamma|\gamma^{(t)})$ as follows

$$
\begin{aligned}
Q(\gamma|\gamma^{(t)}) &= E_{\mathbf{Z}_i}\left[l_{\mathbf{c}}(\gamma) - \lambda||\Theta||_1|\mathbf{y}; \gamma^{(t)}\right] \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \left[\ln \varphi_k(\mathbf{y}_i|\Theta_k^{-1}) + \ln \pi_k\right] E_{\mathbf{Z}_i}\left[1_{\{Z_i=k\}}|\mathbf{y}_i; \gamma^{(t)}\right] - \lambda||\Theta_k||_1 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \left[\ln \varphi_k(\mathbf{y}_i|\Theta_k^{-1}) + \ln \pi_k\right] P\left(Z_i = k|\mathbf{y}_i; \gamma^{(t)}\right) - \lambda||\Theta_k||_1 \\
&= \sum_{i=1}^{n}\sum_{k=1}^{K} \left[\ln \varphi_k(\mathbf{y}_i|\Theta_k^{-1}) + \ln \pi_k\right] \omega_{ik}^{(t)} - \lambda||\Theta_k||_1 \qquad (15)
\end{aligned}
$$

The E-step actually consists of calculating ω_{ik}, the probabilities (conditional on the data and $\gamma^{(t)}$) that $\mathbf{Y}_i$'s originate from component k. It can also be seen as the responsibility that component k takes for explaining the observation $\mathbf{Y}_i$ and it tells us for which group an individual actually belongs. Using Bayes theorem, we have:

$$
\begin{aligned}
\omega_{ik}^{(t)} &= P\left(Z_i = k|\mathbf{y}_i, \gamma^{(t)}\right) \\
&= \frac{P(\mathbf{y}_i|Z_i = k; \gamma^{(t)})P(Z_i = k, \gamma^{(t)})}{\sum_{l=1}^{K} P(\mathbf{y}_i|Z_i = l; \gamma^{(t)})P(Z_i = l, \gamma^{(t)})} \\
&= \frac{\varphi_k^{(t)}(\mathbf{y}|\Theta_k^{-1})\pi_k^{(t)}}{\sum_{l=1}^{K} \varphi_l^{(t)}(\mathbf{y}_i|\Theta_k^{-1})\pi_l^{(t)}}
\end{aligned} \qquad (16)
$$

3.2. The M-step

The M-step for our mixture model can be split in to two parts, the maximization related to π_k and the maximization related to Θ_k.

1. M-step for π_k:
   For the maximization over π_k we make use of the constraint that $\sum_{k=1}^{K} \pi_k = 1$ i.e $\pi_K = 1 - \sum_{k=1}^{K-1} \pi_k$ and $\pi_k > 0$. It turns out that there is an explicit form for π_k. Let $k_0 \in \{1, ..., K-1\}$. Then

$$\frac{\partial Q}{\partial \pi_{k_0}} = \sum_{i=1}^{n} \left[\frac{\omega_{ik_0}^{(t)}}{\pi_{k_0}} - \frac{\omega_{iK}^{(t)}}{1 - \sum_{k=1}^{K-1} \pi_k} \right] \tag{17}$$

   Setting $\frac{\partial Q}{\partial \pi_{k_0}} = 0$, yields the following:

$$\omega_{.k_0}^{(t)} \sum_{k=1}^{K-1} \pi_k + \pi_{k_0} \omega_{.K}^{(t)} = \omega_{.k_0}^{(t)} \tag{18}$$

   It can be shown that a unique solution to Equation (18) is

$$\begin{aligned} \pi_{k_0}^{(t+1)} &= \omega_{.k_0}^{(t)}/n \\ &= \sum_{i=1}^{n} \omega_{ik_0}^{(t)}/n \end{aligned} \tag{19}$$

2. M-step for Θ_k:
   Next, to maximize (15) over Θ_k, we only need the term that depends on Θ_k. The first thing we do here is to try to formulate the maximization problem for a mixture component to be similar to that for Gaussian graphical modeling with the aim of applying graphical LASSO method. The latter applies LASSO penalty to the inverse covariance matrix Θ with the aim of estimating sparse graphs.
   Now from Equation (15), for a specific cluster k_0, the term that depends on the cluster specific covariance matrix Θ_{k_0} is given by

$$\begin{aligned} Q(\Theta_{k_0}) &= \sum_{i=1}^{n} \omega_{ik_0}^{(t)} \ln \varphi_{k_0}(\mathbf{y}_i | \Theta_{k_0}^{-1}) - \lambda ||\Theta_{k_0}||_1 \\ &= \sum_{i=1}^{n} \omega_{ik_0}^{(t)} \left[\frac{1}{2} \ln |\Theta_{k_0}| - \frac{1}{2} \mathbf{y}_i' \Theta_{k_0} \mathbf{y}_i \right] - \lambda ||\Theta_{k_0}||_1 \\ &= \sum_{i=1}^{n} \frac{\omega_{ik_0}^{(t)}}{2} \ln |\Theta_{k_0}| - \frac{1}{2} tr \left(\sum_{i=1}^{n} \omega_{ik_0}^{(t)} (\mathbf{y}_i \mathbf{y}_i') \Theta_{k_0} \right) - \lambda ||\Theta_{k_0}||_1 \\ &= \frac{\omega_{.k_0}^{(t)}}{2} \left[\ln |\Theta_{k_0}| - tr \left(\tilde{S}_{k_0} \Theta_{k_0} \right) - \frac{2\lambda}{\omega_{.k_0}^{(t)}} ||\Theta_{k_0}||_1 \right] \\ &= \frac{\omega_{.k_0}^{(t)}}{2} \left[\ln |\Theta_{k_0}| - tr \left(\tilde{S}_{k_0} \Theta_{k_0} \right) - \lambda_n ||\Theta_{k_0}||_1 \right] \end{aligned} \tag{20}$$

where

$$\omega_{.k_0}^{(t)} = \sum_{i=1}^{n} \omega_{ik_0}^{(t)}$$

$$\tilde{S}_{k_0} = \frac{\sum_{i=1}^{n} \omega_{ik_0}^{(t)} (\mathbf{y}_i \mathbf{y}_i^{'})}{\omega_{.k_0}^{(t)}} \tag{21}$$

is the weighted empirical covariance matrix, and

$$\hat{\Theta}_{k_0} = \arg\max_{\Theta} \left\{ \ln |\Theta_{k_0}| - tr(\tilde{S}_{k_0} \Theta_{k_0}) - \lambda_n ||\Theta_{k_0}||_1 \right\} \tag{22}$$

subject to the constraint that Θ_{k_0} is positive definite with $\lambda_n = \frac{2\lambda}{\omega_{.k_0}^{(t)}}$.

Therefore the maximization of Θ_k consists of running the graphical LASSO procedure (Friedman *et al.*, 2008) for each cluster where each observation $\mathbf{Y}_i$ for Θ_k gets a weight and the sampling covariance matrix S_k is transformed to a weighted sampling covariance. This is a major innovation in our work where we formulate the Gaussian mixture modelling problem in a Gaussian graphical modelling framework. We summarize the algorithm below:

Initialize $\pi_1, ..., \pi_{Kmax}$, $\Theta_1, ..., \Theta_{Kmax}$
repeat
for $\lambda \in (\lambda_1, ..., \lambda_K)$
Compute:

1. E-step: $\omega_{ik} = \frac{\varphi_k(\mathbf{y}|\Theta_k^{-1})\pi_k}{\sum_{l=1}^{K} \varphi_l(\mathbf{y}_i|\Theta_k^{-1})\pi_l}$
2. M-step:
   - $\hat{\pi}_k = \sum_{i=1}^{n} \omega_{ik}/n$
   - $\hat{\Theta}_k = \arg\max_{\Theta} \left\{ \ln |\Theta_k| - tr(\tilde{S}_k \Theta_k) - \lambda_n ||\Theta_k||_1 \right\}$, where

$$\tilde{S}_k = \frac{\sum_{i=1}^{n} \omega_{ik} (\mathbf{y}_i \mathbf{y}_i^{'})}{\omega_{.k}}$$

and

$$\lambda_n = \frac{2\lambda}{\omega_{.k}}.$$

---

4. Simulation and Real-data Example

We generate data from two component mixtures and consider two different schemes based on λ. We study the consistency properties of the PMLE by allowing the sample size to grow. We subsequently applied our method to two real data "Mathematics scores" and "CellSignal" data.

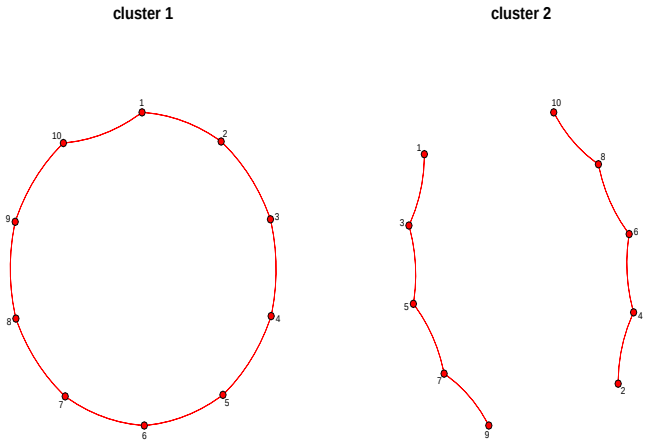

Fig. 3: True graphical model of the 2 clusters

4.1. Simulation

We investigate the consistency properties of the PMLE using our penalized EM algorithm described in section 2. We simulate data $\mathbf{Y}_1, ..., \mathbf{Y}_n$ from two-component multivariate normal mixture models each with probability (true mixture proportion) equals 0.5 and inverse covariance matrix Θ_k built according to the following schemes.

$$\Theta_1(i,j) = \begin{cases} 1 \quad if \quad i=j \\ -0.4, \quad if|i-j|=1 \\ 0, \text{elsewhere} \end{cases} \tag{23}$$

$$\Theta_2(i,j) = \begin{cases} 1 \quad if \quad i=j \\ -0.4, \quad if|i-j|=2 \\ 0, \text{elsewhere} \end{cases} \tag{24}$$

The corresponding graphical model structures are depicted in Figure (3). For a fixed p, we consider two schemes one with $\lambda \propto \sqrt{n \log p}$ where $\lambda_n \propto \frac{1}{\sqrt{n}}$ and the other with $\lambda \propto \sqrt{\log p}$, where $\lambda_n \propto \frac{1}{n}$ each with increasing sample sizes, $n = (100, 300, 800, 2000, 5000)$ to examine the consistency of the PMLEs. In all cases, parameter estimation is achieved by maximizing the likelihood function via our penalized EM-algorithm. The results of our penalized EM-algorithm approach are compared based on the two different schemes corresponding to different values of λ.

Due to the effect of label switching, we are not able to assign correctly each parameter estimate to the right class. As a result, the estimates
$\{(\pi_1, \Theta_1), (\pi_2, \Theta_2)\}$ will be interchangeably represented. We compute the Absolute Deviation (AD) of the mixture proportions, and compare the Frobenuis norm of the difference between the true and estimated precision matrices for each cluster. In addition we compute

Model	Bias(AD)/Frobenuis	F_1 score	TP	FP	Precison
n=100					
π	AD=0.1125				
Θ_1	F=1.7280	0.555	5	5	0.5
Θ_2	F=1.6221	0.529	9	15	0.375
n=300					
π	AD=0.067				
Θ_1	F= 0.9702	0.5333	8	14	0.3636
Θ_2	F= 0.8432	0.5882	10	14	0.4167
n=800					
π	AD=0.0625				
Θ_1	F=0.9279	0.5882	10	14	0.4166
Θ_2	F=0.4804	0.4705	8	18	0.3076
n=2000					
π	AD=0.0263				
Θ_1	F=0.4170	0.5925	8	11	0.4210
Θ_2	F=0.4465	0.625	10	12	0.4545
n=5000					
π	AD=0.002				
Θ_1	F=0.3529	0.6153	8	10	0.444
Θ_2	F=0.2883	0.6060	10	13	0.4347

Table 1: The Absolute Deviation (AD), Frobenuis norm (F), the F_1 score, the True Positive (TP), the False Positive (FP) and the Precision of the PMLE for two-component mixture with $\lambda \propto \sqrt{n \log p}$.

the F_1 score, True positive (TP), False positive (FP), Precision and Recall for the PMLE.

Example 1. We considered the simulated two-component multivariate normal mixture models above and choose sequence of values of λ such that $c_1\sqrt{n \log p} \leq \lambda \leq c_2\sqrt{n \log p}$. On experimental basis we set $(c_1, c_2) = (0.1, 0.25)$. The performances of the penalized EM-algorithm corresponding to different sample sizes are presented in Table 1.

The results show that as the sample size increases, the AD (for the mixture proportions) and the Frobenuis norms (for the precision matrices) decrease indicating the consistency of the PMLEs. At $n = 5000$, the AD for the mixture proportion is almost 0, indicating that our method has recovered precisely the true mixture distribution. We reported also the F_1 score, the True Positive (TP), the False Positive (FP), the Precision and the Recall of the PMLE. We recorded an overall improvement in the F_1 score as n increases.

Example 2. In this example, we again choose the same two-component multivariate Gaussian mixture models. In contrast to the model used in example 1, we have fixed the tuning parameter λ such that $c_1\sqrt{\log p} \leq \lambda \leq c_2\sqrt{\log p}$ and (c_1, c_2) remain unchanged. The performances of the penalized EM-algorithm corresponding to different sample sizes are presented in Table 2. We again observe a decrease in both the Frobenuis norm and the AD as n increases even though we suffer from a deficiency in the AD of π for the case $n = 800$. However

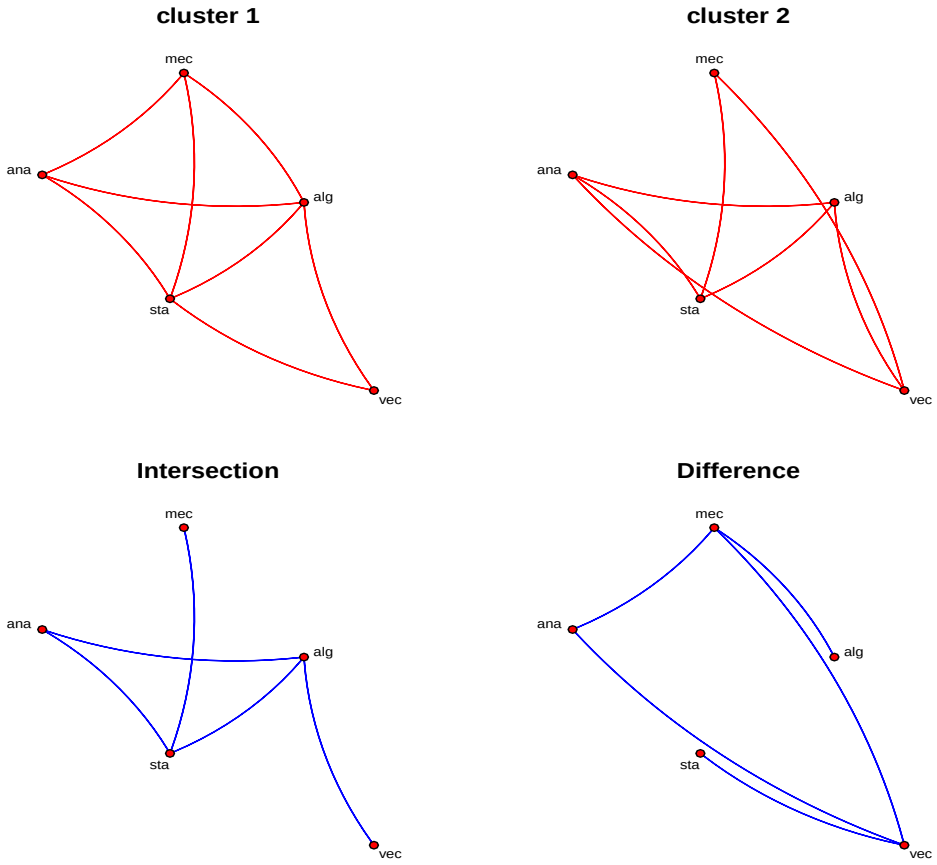

Fig. 4: Graphical model of the 2 group of students

the AD is almost 0 at $n = 5000$. We note that this penalty decreases to 0 faster and as result tends to produce full graph as can be seen in the higher value recorded for false positive. Comparing the two examples, we observe that the choice of λ plays a strong role in parameter estimation and graph selection consistency of the resultant networks. The consistency properties of the PMLEs was achieved in both cases but our results indicate that the overall performance of the asymptotic behavior of $\lambda \propto \sqrt{n \log p}$ is more satisfactory. Even though both penalty decrease to 0 as n increases, $\lambda \propto \sqrt{n \log p}$ decreases slower resulting in a relatively sparser networks as compared to $\lambda \propto \sqrt{\log p}$.

4.2. Real-data Examples

4.2.1. Mathematics Scores Data

As a simple example of a data set to which mixture models may be applied, we consider the data set on marks in five mathematics exams score. This data set consists of 88 students

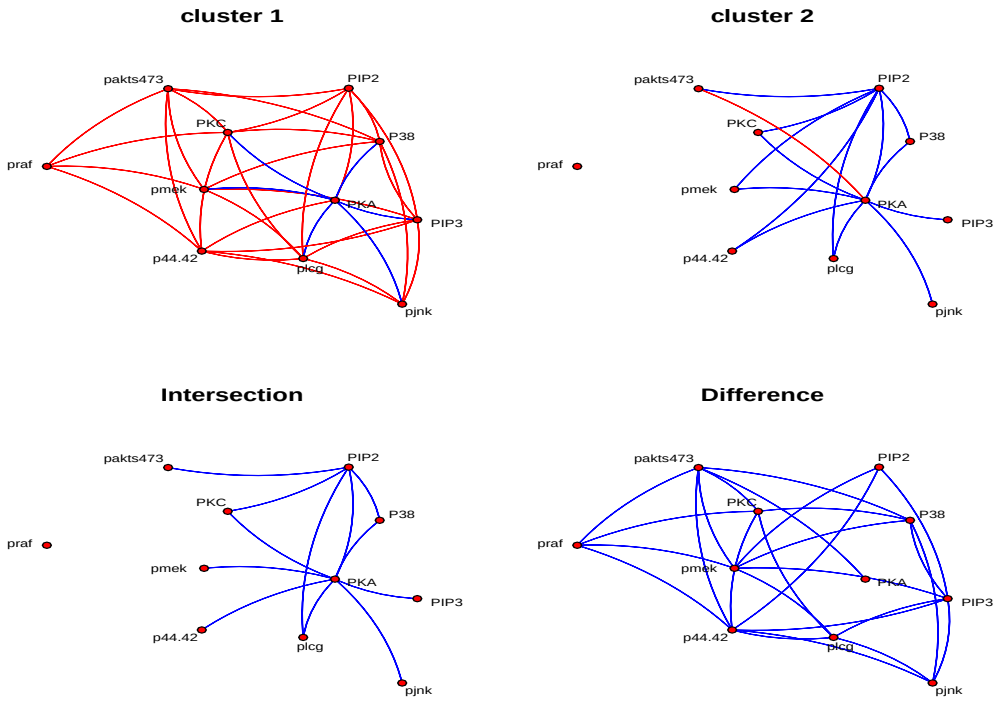

Fig. 5: Graphical models of the CellSignal data with two mixtures of Gaussian distributions

who took examinations in 5 subjects namely mechanics, vectors, algebra, analysis, statistics. Some were with open book and others with closed book. Mechanics and vectors were with closed book.

We fit a two-mixture components to the data with a strong indication that there are two groups of students each with similar subjects interest. We applied our PMLE algorithm to the data with λ based on scheme 1. The pattern of interactions among the two groups were depicted in Figure (4). The network differences as well as similarities are also shown. The results indicate that 61% of students have similar subjects interest while 39% falls in other group of interest. In one group, we observe no interactions between mechanics and analysis nor statistics and vectors while in the other group such interactions do exist.

4.2.2. Analysis of cell signalling data

We consider the application of our method on the flow cytometry dataset (cell signalling data) of Sachs *et al.* (2005). The data set contains flow cytometry of $p = 11$ proteins measured on $n = 7466$ cells. The CellSignal data were collected after a series of stimulatory cues and inhibitory interventions with cell reactions stopped at 15 minutes after stimulation by fixation, to profile the effects of each condition on the intracellular signaling networks.

Model	Bias(AD)/Frobenuis	F_1 score	TP	FP	Precison
n=100					
π	AD=0.0307				
Θ_1	F= 3.4081	0.3446	10	32	0.2380
Θ_2	F= 3.4018	0.3181	7	29	0.1944
n=300					
π	AD=0.0356				
Θ_1	F=1.0539	0.3703	10	34	0.2272
Θ_2	F=0.8657	0.3137	8	35	0.1860
n=800					
π	AD=0.0669				
Θ_1	F=0.6419	0.3703	10	34	0.2272
Θ_2	F=0.7605	0.3018	8	37	0.1777
n=2000					
π	AD=0.0312				
Θ_1	F=0.5081	0.3168	8	34	0.1882
Θ_2	F=0.4150	0.3636	10	35	0.2222
n=5000					
π	AD=0.0065				
Θ_1	F=0.2771	0.3703	10	34	0.2272
Θ_2	F=0.2857	0.2692	7	37	0.1590

Table 2: The Bias(AD), Frobenuis norm (F), F_1 score, True Positive (TP), False Positive (FP) and the Precision of the PMLE for two-component mixture with $\lambda \propto \sqrt{\log p}$.

Each independent sample in the data set is made up of quantitative amounts of each of the 11 phosphorylated molecules, simultaneously measured from single cells.

We again fit a two-mixture component to the data. The result of applying our PMLE algorithm to the data set using the first scheme is shown Figure (5). The result indicates that 90% of the observation falls in one component whiles 10% falls in the other cluster. We also display the differences and similarities in the two components. The following proteins interaction were seen to be present in each of the two components: $(pakts473, PIP2)$, $(PKC, PIP2)$, $(PKA, pjnk)$, $(pmek, PKA)$ to mention but few. Differences in the interaction occur among the following proteins: $(pakts473, praf)$, $(PIP2, p44.42)$, $(PKC, plog)$; see Figure (5) for details.

5. Conclusion

We have developed a penalized likelihood estimator for Gaussian graphical mixture models. We have imposed an L_1 penalty on the precision matrix with extra condition preventing the likelihood not to degenerate. The estimates were efficiently computed through a penalized version of the EM-algorithm. By taking advantage of the recent development in Gaussian graphical models, we have implemented our method with the use of the graphical lasso algorithm. We have provided consistency properties for the penalized maximum likelihood estimator in Gaussian graphical mixture model. Our results indicate a better performance

in parameter consistency as well as in graph selection consistency for $\lambda = O(\sqrt{n \log p})$ or $\lambda_n \propto \frac{1}{\sqrt{n}}$. Our method is suitable for large networks recovering from non homogeneous data. Another interesting situation is when K, the number of mixture components in the model is unknown. This is a more practical problem than the one we have discussed and probably involves simultaneous model selection.

References

Agakov, F, V., Orchard, P, S., and Amos, J., 2012. Discriminative Mixtures of Sparse Latent Fields for Risk Management. Journal of Machine Learning Research - Proceedings Track, 2012, 22,10-18

Banfield, J, D., and Raftery, A. E., 1993. Model-Based Gaussian and Non-Gaussian Clustering. Biometrics, 1993, 49,3, 803-821

Bernardo, JM., 2003. Bayesian clustering with variable and transformation selections. Bayesian Statistics 7: Proceedings of the Seventh Valencia International Meeting., 2003,249.

Biernacki, C., Celeux, G., and Govaert, Gérard., 2000. Assessing a Mixture Model for Clustering with the Integrated Completed Likelihood. IEEE Trans. Pattern Anal. Mach. Intell., 2000, 22, 7, 719–725.

Chanda, K., C., 1954. A Note on the Consistency and Maxima of the Roots of Likelihood Equations. Biometrika., 1954, 41, 1/2, 56-61.

Cramer, H., 1946. Mathematical methods of statistics. Princeton University Press.

Day, N., E., 1969. Estimating the Components of a Mixture of Normal Distributions. Biometrika., Day1969, 56, 3, 463-474.

Dempster, A. P., Laird, N. M., and Rubin, D. B., 1997. Maximum Likelihood from Incomplete Data via the EM Algorithm. Journal of the Royal Statistical Society., 30, 1, 1-38.

Friedman, J., Hastie, T., and Tibshirani, R., 2008. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 2008, 3,9, 432-441

Kiefer, J., and Wolfowitz, J., 1956. Consistency of the Maximum Likelihood Estimator in the Presence of Infinitely Many Incidental Parameters. The Annals of Mathematical Statistics., 1956, 27, 4, 887-906.

Lauritzen, S, L., 1996. Graphical models. The Clarendon Press Oxford University Press, New York, 1996.

Lindsay, B. G., and Ray. S., 2006. The topography of multivariate normal mixtures. Annals of Statistics., 2006, 33, 5, 2042-2065.

Meinshausen, N., and Bühlmann, P., 2006. High dimensional graphs and variable selection with the Lasso. Annals of Statistics., 34, 1436–1462.

Redner, R. A., 1980. Maximum Likelihood Estimation for Mixture Models. JSC (Series)., Lyndon B. Johnson .Space Center, NASA.

Redner, R. A., 1981. Note on the Consistency of the Maximum Likelihood Estimate for Nonidentifiable Distributions. The Annals of Statistics., 9, 1, 225-228.

Ruan, L., Yuan, M., and Zou, H., 2001. Regularized parameter estimation in high-dimensional gaussian mixture models. Neural Comput., 2001, 23,6, 1605–1622

Sachs, K., Perez, O., Pe'er, D., Lauffenburger, D. A., and Nolan, G. P., 2005. Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data. Science., 30, 1, 523-529.

Wald, A., 1949. Note on the Consistency of the Maximum Likelihood Estimate. The Annals of Mathematical Statistics., 20, 4, 595-601.

Yuan, M., and Lin, Y., 2007. Model selection and estimation in the Gaussian graphical model. Biometrika., 94, 1, 19-35.

Zhou, H., Pan, W., and Shen, X., 2009. Regularized Penalized model-based clustering with unconstrained covariance matrices. Electron J Sta., 2009, 3, 1473-1496.