THE PACKAGE *EMGLASSO*

HARRISON WATTS

1. INTRODUCTION

2. METHOD

Consider a d-dimensional mixture distribution with k mixture components associated with the parameter vector $\boldsymbol{\Theta} = (\boldsymbol{\tau}, \boldsymbol{\mu}, \Sigma)$, where

$$\boldsymbol{\tau} = \begin{bmatrix} \tau_1 \\ \tau_2 \\ \vdots \\ \tau_k \end{bmatrix}, \qquad \boldsymbol{\mu} = \begin{bmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \\ \vdots \\ \boldsymbol{\mu}_k \end{bmatrix}, \qquad \boldsymbol{\Sigma} = \begin{bmatrix} \boldsymbol{\Sigma}_1 \\ \boldsymbol{\Sigma}_2 \\ \vdots \\ \boldsymbol{\Sigma}_k \end{bmatrix}$$

represent respectively the mixture probability, the mean vector, and the covariance matrix for each component in the MMN distribution. For a random sample $\mathbf{X}' = (X_1', ..., X_n')$ of size n drawn independently from the population, the unobserved data are the random variables $\mathbf{Z}' = (Z_1, ..., Z_n)$ that determine with probability τ_j from which mixture the observation originates:

$$\mathbf{X}_i | (Z_i = j) \sim \mathcal{N}_d(\mu_j, \boldsymbol{\Sigma}_j), \qquad P(Z_i = j) = \tau_j, \qquad j = 1, ..., k,$$

$$\text{subject to} \qquad \sum_{j=1}^{k} \tau_j = 1.$$

These variables constitute a random sample on the multinomial random variable $\mathbf{Z}$. Therefore $Z_1, ..., Z_n$ are independently and identically distributed multinomial random variables with probabilities $\boldsymbol{\tau}$. Let $\mathbf{x}' = (\mathbf{x}_1', ..., \mathbf{x}_n')$ be the observed values of $\mathbf{X}$ given by the sample. And $\mathbf{z}' = (z_1, ..., z_n)$ are the latent values associated with the realization of the random sample. The observed likelihood function is $L(\boldsymbol{\Theta}|\mathbf{x}) = \prod_{i=1}^{n} \sum_{j=1}^{k} \tau_j f(\mathbf{x}_i; \boldsymbol{\mu}_j \boldsymbol{\Sigma}_j)$, where where f

Date: September 2018.

is the pdf of the jth multivariate normal distribution in the mixture and is given by

$$f(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{\exp(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}))}{\sqrt{(2\pi)^d|\boldsymbol{\Sigma}|}},$$

where $|\boldsymbol{\Sigma}|$ is the determinant of the covariance matrix $\boldsymbol{\Sigma}$. The complete likelihood function is the product

$$L^c(\boldsymbol{\Theta};\mathbf{x},\mathbf{z}) = \prod_{i=1}^{n}\prod_{j=1}^{k}\tau_j f(x_n;\mu_j,\boldsymbol{\Sigma}_j)^{\mathbf{1}_{(z_i=j)}},$$

where $\mathbf{1}$ is the indicator function. The product is used with an indicator function in the exponent because the latent values are given in the complete likelihood function, which is unknown in practice. Given the mostly exponential structure of the complete likelihood function, calculating the exapected value is simplified by considering $\ell^c(\boldsymbol{\Theta};\mathbf{x},\mathbf{z}) = \log L^c(\boldsymbol{\Theta};\mathbf{x},\mathbf{z})$ the logarithm of the complete likelihood:

$$\ell^c(\boldsymbol{\Theta};\mathbf{x},\mathbf{z}) = \sum_{i=1}^{n}\sum_{j=1}^{k}\log\tau_j - \frac{1}{2}\Big((\mathbf{x}_i-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j) + d\log(2\pi) + \log|\boldsymbol{\Sigma}_j|)\Big)$$

At each iteration t of the EM algorithm, the estimates for $\boldsymbol{\Theta}$ are

$$\boldsymbol{\Theta}^{(t)} = \begin{bmatrix}\tau^{(t)}\\ \mu^{(t)}\\ \boldsymbol{\Sigma}^{(t)}\end{bmatrix}.$$

The goal is to maximize the objective function

$$\begin{aligned}Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)}) &= E_{\mathbf{Z}|\mathbf{x},\boldsymbol{\Theta}^{(t)}}[\ell^c(\boldsymbol{\Theta};\mathbf{x},\mathbf{Z})]\\ &= \sum_{i=1}^{n}\sum_{j=1}^{k}T_{i,j}^{(t)}\left[\log\tau_j^- \frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_j)'\boldsymbol{\Sigma}_j^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_j) - \frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Sigma}_j|)\right]\end{aligned}$$

where

$$T_{i,j}^{(t)} = P(\mathbf{Z}_i = j|\mathbf{X}_i = \mathbf{x}_i,\boldsymbol{\Theta}^{(\mathbf{t})}) = \frac{\tau_j^{(t)}}{\sum_{j=1}^{k}\tau_j^{(t)}}f(\mathbf{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)$$

by Bayes' theorem.

Obtain the $(t+1)$th estimates for $\boldsymbol{\Theta}$ as the arguments for Q that produce a maximum, i.e. $\boldsymbol{\Theta}^{(t+1)} = \text{argmax}Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^{(t)})$. Because $\boldsymbol{\tau}$ is in a separate term of a linear combination of the parameters that constitute $\boldsymbol{\Theta}$, it can be optimized separately from $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. The new

estimate for $\boldsymbol{\tau}$ is then obtained by maximizing Q in the direction of $\boldsymbol{\tau}$. Define the function the terms of the sum that involve only $\boldsymbol{\tau}^{(t)}$

$$\begin{aligned} &\text{maximize} && Q_{\boldsymbol{\tau}} = \sum_{i=1}^{n}\sum_{j=1}^{k} T_{i,j}^{(t)} \log \tau_j \\ &\text{subject to} && R_{\boldsymbol{\tau}} = \sum_{j=1}^{k} \tau_j - 1 = 0. \end{aligned}$$

This constrained optimization problem can be solved with a Lagrange multiplier λ. Solve the system of equations $(\nabla Q_{\boldsymbol{\tau}}, R_{\boldsymbol{\tau}}) = (\nabla \lambda R_{\boldsymbol{\tau}}, 0)$ where ∇ denotes the gradient with respect to $\boldsymbol{\tau}$. Expressions for each component in $\boldsymbol{\tau}$ can be found in terms of λ. Then the constraint equation can be used to find a constant value for λ. This procedure leads to the following system:

$$\begin{aligned} \sum_{j=1}^{k} \frac{T_{i,1}^{(t)}}{-\lambda} &= \tau_1 \\ \sum_{j=1}^{k} \frac{T_{i,2}^{(t)}}{-\lambda} &= \tau_2 \\ &\vdots \\ \sum_{j=1}^{k} \frac{T_{i,k}^{(t)}}{-\lambda} &= \tau_k \\ \sum_{j=1}^{k} \tau_j &= 1. \end{aligned}$$

It is clear that the system is consistent only when

$$\sum_{j=1}^{k}\sum_{i=1}^{n} \frac{T_{i,j}^{(t)}}{-\lambda} = 1.$$

Because $\mathbf{Z}_i|\mathbf{x}_i, \boldsymbol{\Theta}^{(\mathbf{t})}$ is a discrete random variable with support $\{1, 2, ..., k\}$, it must be that the sum $\sum_{j=1}^{k} T_{i,j}^{(t)} = \sum_{j=1}^{k} P(\mathbf{Z}_i = j|\mathbf{X}_i = \mathbf{x}_i, \boldsymbol{\Theta}^{(\mathbf{t})}) = 1$ by the definition of probability mass

function. Therefore $\sum_{i=1}^{n} \frac{1}{-\lambda} = 1$ so that $\lambda = -n$. Use this value to obtain

$$\boldsymbol{\tau}^{(t+1)} = \begin{bmatrix} \frac{1}{n}\sum_{j=1}^{k} T_{i,1}^{(t)} \\ \frac{1}{n}\sum_{j=1}^{k} T_{i,2}^{(t)} \\ \vdots \\ \frac{1}{n}\sum_{j=1}^{k} T_{i,k}^{(t)} \end{bmatrix}.$$

Finally the parameter estimate $\boldsymbol{\Theta}^{(maxIt)}$ is used as the argument for the function *glasso* from its R package of the same name. The so-called *lstep* in the package *emglasso* initializes a vector $\mathcal{R} = (0.01, 0.02, ..., 2)$ of candidate penalties and evaluates *glasso* for each $\rho = 1, 2, ...200$ to obtain the estimate $\mathbf{W}_\rho$ for $\boldsymbol{\Sigma}$. For each $n = 1, ..., N$ and $k = 1, ..., K$ the Bayesian information criterion (BIC) is calculated by the formula $BIC_\rho = -2(loglik_\rho) + (D2_\rho)\ln(N)$, where $loglik_\rho$ is the natural logarithm of the likelihood $L(\mathbf{W}_\rho; \mathbf{x}, \mathbf{z})$ and is a value of the *glasso* function. The symbol $D2_\rho$ represents the number of non-zero parameters that is estimated by *glasso* for each $\rho = 1, 2, ..., 200$ and $k = 1, ..., K$. So $D2$ is the number of parameters in the upper triangle of the mixture's kth covariance matrix estimate in $\mathbf{W}_\rho$. A higher value of $\mathcal{R}$ can only yield higher sparsity in the estimate of the covariance matrix $\boldsymbol{\Sigma}$, so an increase in $\mathcal{R}$ may correspond to a decrease in $D2_\rho$ and thus a decrease in BIC_ρ. The penalty $\mathcal{R}_\rho^*$ that minimizes BIC_ρ is selected and used as a the penalty for *glasso* to obtain $\mathbf{W}_\rho^*$, the final estimate for $\boldsymbol{\Sigma}$.

3. Simulation

The package *emglasso* is equipped with functions for generating positive definite matrices of arbitrary sparsity and random samples from *MMN* distributions. The function $rspdmatrix(D, \lambda, epsilon = 1e-4)$ creates a block matrix of specified dimension D by drawing random values from a uniform distribution over (0,1). This is not the typical independence structure for practical applications. The nonzero parameters are random samples from the binomial distribution. Then entry indices are drawn from a binomial distribution with parameters (D^2, λ), where λ is the specified sparsity of the matrix. These entries are removed symmetrically about the diagonal. Finally the diagonal is modified to ensure the matrix is positive definite. Sparse matrices that arise from mathematical modeling of real

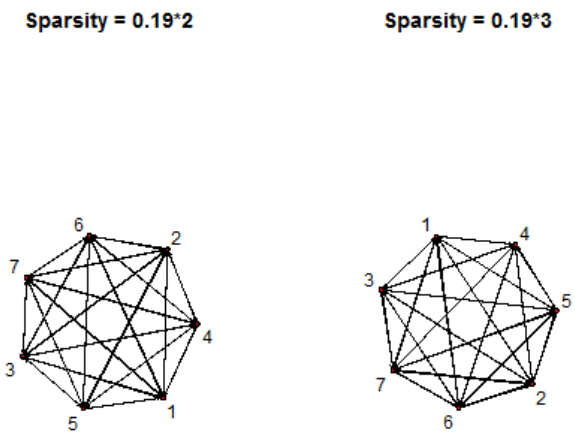

FIGURE 1. Low sparsity precision matrices yield highly dependent graphs. These graphs correspond to the covariance matrices Sigma0[[2]] and Sigma0[[3]].

data, however, tend to have a heavier distribution of nonzero parameters near the diagonal rather than randomly preserving only the positive definite property. This is the type of sparsity that *glasso* was developed to estimate. Below is the R code for the function *rspdmatrix* followed by a demonstration of some matrices obtained of various sparsity represented as the graph of its corresponding precision-matrix inverse.

The function *rmmnorm*$(N, D, \tau, \mu, \mathbf{\Sigma})$ generates a random sample of size N from a collection of multivariate normal distributions with parameters μ and $\mathbf{\Sigma}$ mixed according to τ.

A code in R will run *emglasso* on a randomly generated MMN sample generated with *rmmnorm* and the *rspdmatrix* code from above.

The true parameters are recorded below, followed by the parameter estimates obtained from *emglasso*.

In this execution of *emglasso*, the covariance sparsity of each mixture component was gradually increased by 19%.

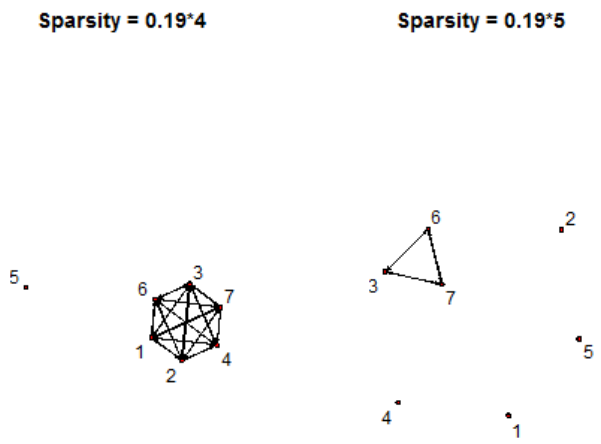

FIGURE 2. Higher sparsity precision matrices yield less dependent graphs. These graphs correspond to the covariance matrices Sigma0[[4]] and Sigma0[[5]].

The plots for the BIC for the last four components of the covariance matrix are in figures 3 and 4.

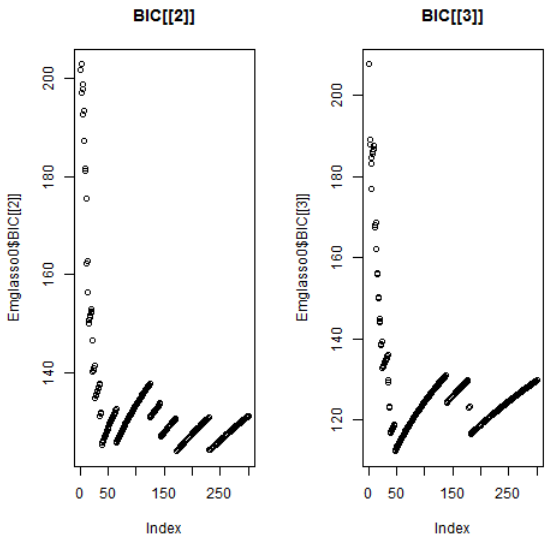

FIGURE 3. These plots correspond to the BIC for the covariance matrices Sigma0[[2]] and Sigma0[[3]].

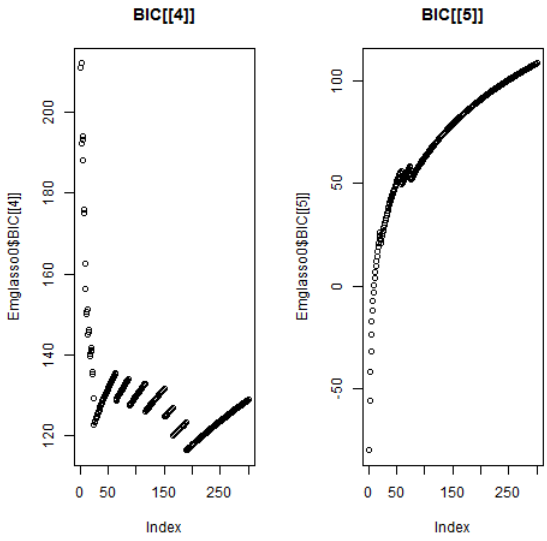

FIGURE 4. These plots correspond to the BIC for the covariance matrices Sigma0[[4]] and Sigma0[[5]].