Graph clustering with the Stochastic block model (using Daudin's ICL)

Martin Metodiev

Daudin's Paper

The following is a brief project with the goal to implement Daudin's ICL:

$$\begin{aligned} ICL(m_Q) &= \max_{\theta} \log \mathcal{L}(\mathcal{X}, \tilde{\mathcal{Z}}|\theta, m_Q) \\ &- \frac{1}{2} \times \frac{Q(Q+1)}{2} \log \frac{n(n-1)}{2} - \frac{Q-1}{2} \log(n) \end{aligned} \tag{1}$$

Simulation scenario

The following simulation scenario was inspired from a vignette from the greed package. However, none of this package's code is used. All new functions used were written by myself.

We begin by simulating from a hierarchically structured SBM model, with 2 large clusters, each composed of 3 smaller clusters with higher connection probabilities, making a total of 6 clusters.

```
N  <- 400           # Number of node
K  <- 6             # Number of cluster
pi <- rep(1/K,K)    # Clusters proportions
lambda   <- 0.1     # Building the connectivity matrix template
lambda_o <- 0.01
Ks <- 3
mu <- bdiag(lapply(1:(K/Ks), function(k){
  matrix(lambda_o,Ks,Ks)+diag(rep(lambda,Ks))}))+0.001
sbm <- sim_sbm(N,pi,mu) # Simulation
```

The method finds a very good solution. Model selection is done via Daudin's ICL as described above. Note that convergence thresholds are set to be quite small (1e-3), so large errors can (and do) occur, especially when the number of classes Q is large. In fact, setting Q<=6 is recommended.

```
sol = var_bayes_model_selection(sbm$x, Q=3)
#> [1] "ICL:  -7540.04840521306 for  3 clusters"
#> [1] "ICL:  -7433.89124634079 for  4 clusters"
#> [1] "ICL:  -7344.60699162115 for  5 clusters"
#> [1] "ICL:  -7308.1511059694 for  6 clusters"
#> [1] "ICL:  -7337.35087244629 for  7 clusters"
table(sol@cl, sbm$cl)
#>
#>      1   2   3   4   5   6
```

```
#>    1  0  0  0  0  0 68
#>    2  0  0  0  0 60  0
#>    3  0  0  0 79  0  0
#>    4  1  2 64  0  0  0
#>    5  0 62  0  0  0  0
#>    6 62  2  0  0  0  0
```