**Graphical models**
Summer school 2016

*V. Monbet*

## 1. Undirected graphical models, continuous variables

### Multivariate Gaussian distribution

*needed packages: ade4, FactoMineR, qgraph, fields, glmnet*

### Data download

```
library(ade4)
data(deug)
str(deug)
```

This dataset gives the exam results of 104 students in the second year of a French University onto 9 subjects. The best mark is 20 and the worse 0. Options 1 and 2 are not precisely described. Sport gives additive points if the student get a mark over 10. For instance, one student with 12 obtains 2 points.
We suppose that the data are independant observations of a multivariate Gaussian distribution. We aim at finding conditional independencies. Sport can be considered as a supplementary variable.

### Questions

1. Data exploration
   (a) Look at scatter plots to find the strongest correlations between the variables

   ```
   pairs(deug$tab,pch=20)
   ```

   (b) Use principal component analysis to give an overview of the data.

   ```
   library(FactoMineR)
   PCA(deug$tab,quanti.sup=9)
   ```

   Does it confirm what you observed in the previous question?
   (c) Plot the empirical correlation precision matrices of the data. You can used the following function for the plot.

```
plot.mat <- function(A,zr=NULL,col = tim.colors(64),main=NULL){
par(mar=c(4, 5, 5, 5) + 0.1)
if (is.null(zr)) {zr = range(A)}
n1 = dim(A)[1]
ticks = seq(0,1,length.out=dim(A)[1])
image(t(A)[,rev(1:n1)],zlim = zr,axes=FALSE,col=col,main=main)
box(col="black")
axis(1,at=ticks,labels=FALSE)
text(ticks, par("usr")[3] - .12, labels = colnames(C), srt = 45,
pos = 2, xpd = TRUE)
image.plot( legend.only=TRUE, zlim= zr,col=col)
}
```

Some coefficients of the precision matrix are small. How can you interpret it?

2. Undirected graphical model.

   (a) Plot a graph of the data from the correlation matrix $C$

   ```
   linrary(qgraph)
   C = cor(data$deug[,-9])
   qgraph(C,labels=colnames(C))
   ```

   Can you explain why all the variables are connected? Is it a good representation of the true model?
   The graph can also be plot from the precision matrix as follows.

   ```
   C = cor(data$deug[,-9])
   C.inv = solve(C)
   qgraph(wi2net(C.inv),labels=colnames(C))
   ```

   (b) Hard thresholding for edge selection.
   For the sequel, we can work with standardized data and exclude the sport variable. It will leads to simpler programs without changing the results.

   ```
   Xc = scale(deug$tab[,-9])
   ```

   We will try to simplify the graph by finding the null partial correlations. We first use hard thresholding. It means taht we will set to zeros all the cefficients of the transition matrix which are below a fixed threshold. The "best" threshold can be chosen by comparing BIC criteria.
   Write a program which, for a sequence of thresholds $s$,
   - set to zero the coefficients of the precision matrix which are least than $s$
   - compute the BIC criteria corresponding to the new precision matrix
   - select the best threshold $s$
   - plot the corresponding graph.
   Now interpret the obtained graph.

   (c) Link with the linear model and stepwise variable selection.
   Choose a variable and fit a linear model to predict this variable given the others (use R function `lm`).

Compare the estimated coefficients to the graph.
Use the step function to select a parcimonious model. Compare it to the graph. The step function works according to AIC cirteria. Cross-validation is another interesting way to compare models.
Compare the complete linear model and the ones you selected using 10 folds cross-validation.

(d) Soft thresholding for edge selection.
Soft thresholding can also be used for edge selection using Lasso penalty. For Lasso penalization, there is no threshold to select but a reguralisation constant.

i. Variable per variable analysis.
Choose a variable as exemple and use *glmnet* to find a good set of predictors of this variable.

ii. All at one time
Use *glasso* to build a sparse graph.

iii. Compare both results with each other and with the results obtained in the previous questions.

3. More classical tests for zero partial correlation.
Zero partial correaltion can be tested using conditional independence test. For testing the independance of $X$ andd $Y$ given $\mathbf{Z}$ in a Gaussian model, the test statistic is defined as follows

$$z(\rho_{XY|\mathbf{Z}} = \frac{1}{2}\log\left(\frac{1+\hat{\rho}_{XY|\mathbf{Z}}}{1-\hat{\rho}_{XY|\mathbf{Z}}}\right)$$

$$S = \sqrt{n - dim(\mathbf{Z}) - 3}|z(\rho_{XY|\mathbf{Z}}|.$$

$S$ tends in distribution to a standard Gaussian variable when $n$ tends to infinity.

(a) Use this to test conditional independencies you found with the graphical model.

(b) Propose a bootstrap test to do the same.

4. Give a conclusion to resume the results you found in the previous questions.

5. Latent factors.
According to the previous result, one can assume that the dataset could be described by a small number of latent factors.

(a) Use function `factanal()`

(b) Propose a graph structure and write an EM algorithm to estimate the parameters of the model and infer the latent variables given the observations.

**2. Undirected graphical models, binary variables**

*needed packages: IsingSampler, IsingFit, glmnet, qgraph*

**Data download**

We will work with a survey on Portoguese student habits concerning alcohool consumption. Here, we will mainly focus on understanding the life environment of the students. The data, their description and a file to load them are available on my teaching webpage under the SummerSchool link.
Some variables are binary and others are nominal. At a first step we transform the variables to have only binary variables in order to be able to use the binary Ising model.

**Questions**

1. Data exploration

   (a) Use multivariate correspondance analysis (MCA) to give an overview of the data. For this analysis, data have to be defined as factor.

   ```
   library(FactoMineR)
   mca = MCA(d3)
   ```

   It is also interesting to look at the results when the `G1` to `G3` variables are considered as supplementary variables.

   ```
   p = ncol(d3)
   mca = MCA(d3,quali.sup=(p-2):p)
   ```

   (b) Can you already see significant dependence between the variables?

2. Ising model

   (a) Use the `IsingFit` function to fit and plot an Ising model. This function use a *elasso* algorithm to fit the model
   (more details on `http://www.nature.com/articles/srep05918`)[1].
   Can you explain why all the variables are not connected?

   (b) Compare the graph to the results of CMA. What appends if the `G1` to `G3` variables are kept out from the graph estimation?

   (c) Compare the graph to the weighted adjacency matrix obtained as an output of the IsingFit function.

   (d) Give an interpretation of the fitted model. Are there interesting subgraphs that one could focus on?

3. Link with the logistic regression and stepwise variables selection.

   (a) Choose a variable and fit a logistic regression to predict this variable given the others (use R function `glm`). Compare the estimated coefficients to the graph.

[1] Caution: definition of the penalty is such that $\gamma$ can be negative?

(b) Use the `step` function to select a parcimonious model. Compare it to the graph.

(c) Use the `glmnet` function ot fit a regularized logistic regresion with lasso penalty. Compare it to the graph.

4. Now, we wish to verify the robustness of the estimator. You can choose to work on a smaller set of variables corresponding to a subgraph.

   (a) Fit a model on the observations of the subset of variables.

   (b) Use the `IsingSampler` function to sample several realizations of this model. The sample size can be fixed to the number of observations.

   (c) Fit again the model on the simulations and compare it to the reference model (the one fitted on the observations).

   (d) Propose a bootstrap test to test conditional dependencies.

5. Latent factors
   We remarked earlier that the nodes of the graph can be grouped in clusters. Latent discrete varables can be infered to resume the data.
   Propose an algorithm to estimate such latent variables. You can just give the ideas without implement them.