

# Influential Observations in a Graphical Model

Avner Bar-Hen

MAP5, Paris Descartes

Joint work with  
Jean-Michel Poggi  
(LMO, Paris Sud, Orsay & Paris Descartes)

# Graphical Model

Let  $X = (X_1, \dots, X_p) \sim \mathcal{N}_p(\mu, \Sigma)$  be a  $p$ -dimensional multivariate normal distributed random variable supposed to be such that  $\Sigma$  is invertible

- ▶ Graphical models encode random variables and their conditional dependencies
- ▶ Directed acyclic graph in which nodes  $\Gamma = \{1, \dots, p\}$  represent random variables and edges represent conditional probabilistic dependencies among them
- ▶ A pair  $(a, b)$  is in the set of edges if and only if  $X_a$  is dependent on  $X_b$  conditionally to the remaining variables  $\{X_k, k \in \Gamma \setminus \{a, b\}\}$
- ▶  $\text{cor}(X_a, X_b | \{X_k, k \in \Gamma \setminus \{a, b\}\}) = 0$  corresponds to a zero entry in  $\Theta = \Sigma^{-1}$

# Graphical Model

The  $L_1$ -penalized log-likelihood is

$$\ell_\lambda^S(\Theta) = \log \det \Theta - \text{tr}(\Theta S) - \lambda \|\Theta\|_1$$

$\lambda \geq 0$  being the tuning parameter.

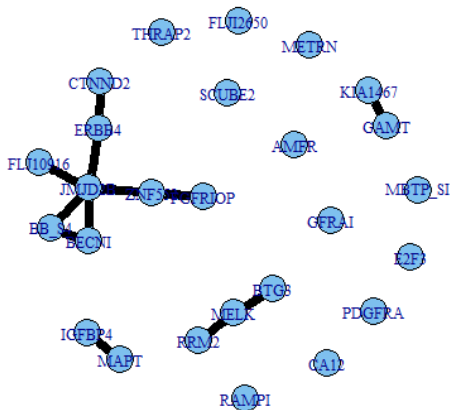
- ▶  $S = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})(X_i - \bar{X})'$  is the empirical covariance matrix
- ▶  $\hat{\Theta} = \arg \max \ell_\lambda^S(\Theta)$  is the ML estimate of the inverse of the concentration matrix  $\Sigma^{-1}$

The non-null entries of  $\hat{\Theta}$  define the edges of the estimated graphical model.

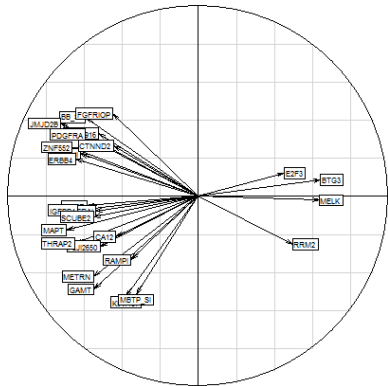
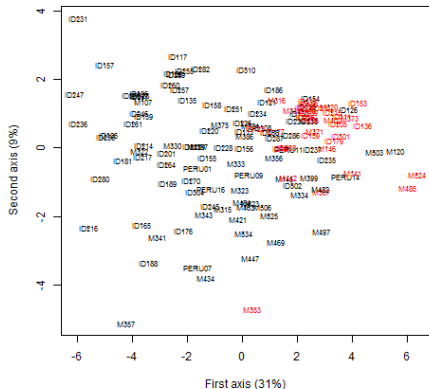
## Example

- ▶ 133 patients with stage I-III breast cancer (Hess et al., 2006) treated with chemotherapy prior to surgery
- ▶ Hess et al. (2006), Natowicz et al.(2008) developed and tested a multigene predictor for treatment response on this data set. They focused on a set of 26 genes having a high predictive value
- ▶ Patient response to the treatment is classified as either a pathologic complete response (pCR) 34 individuals or a residual disease (not-pCR) 99 individuals
- ▶ Data: 26 columns and 133 rows. The  $n$ th row gives the expression levels of the 26 identified genes for the  $n$ th patient. The  $p$  columns are named according to the genes
- ▶ Data already considered by Ambroise et al. (2009) and Giraud et al. (2012) : Gaussian Graphical Model to obtain genes interaction graph ( $L_1$  –penalized likelihood criterion).

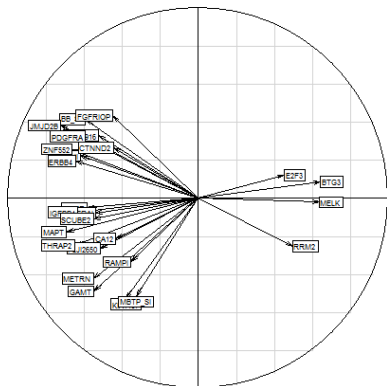
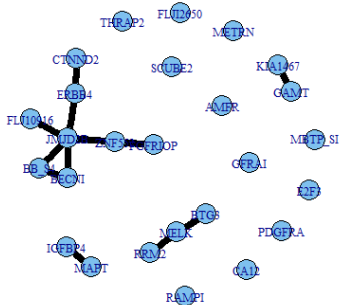
## Example: network (R package huge)



## Example: PCA



## Example: PCA



## Influence of the observations

### What about graph stability?

- ▶ Classically robustness deals with model stability (and considered globally)
- ▶ Focus on **individual observations diagnosis** issues rather than model properties or variable selection problems
- ▶ We use here Graphical Models to perform diagnosis on observations
- ▶ We use influence function, a classical diagnostic method to measure the perturbation induced by a single observation: **stability issue through jackknife**



## Influence function

- ▶  $X_1, \dots, X_n$  r.v. of common distribution function (df)  $F$  on  $\mathbb{R}^p$  ( $p \geq 1$ )
- ▶ The influence of an infinitesimal perturbation along  $\delta_x$  on statistic  $T(F)$

$$IC_{T,F}(x) = \lim_{\epsilon \rightarrow 0} \frac{T((1 - \epsilon)F + \epsilon \delta_x) - T(F)}{\epsilon}$$

- ▶ Statistic  $T(F)$  naturally estimated by  $T(F_n)$   
where  $F_n = \frac{1}{n} \sum_{i=1}^n \delta_{x_i}$  is the empirical df
- ▶  $IC_{T,F_n}(x_i)$  is used to evaluate the **importance** of an **observation**  $x_i \in \mathbb{R}^p$
- ▶ **Connection between influence function and jackknife** (Miller, 1974):  
let  $F_{n-1}^{(i)} = \frac{1}{n-1} \sum_{j \neq i} \delta_{x_j}$ , then  $F_n = \frac{n-1}{n} F_{n-1}^{(i)} + \frac{1}{n} \delta_{x_i}$ . If  $\epsilon = -\frac{1}{n-1}$ , we have:

$$\begin{aligned} IC_{T,F_n}(x_i) &\approx \frac{T((1 - \epsilon)F_n + \epsilon \delta_{x_i}) - T(F_n)}{\epsilon} \\ &\approx (n-1)(T(F_n) - T(F_{n-1}^{(i)})) \end{aligned}$$

## A first remark about jackknifed covariance matrix

- ▶  $\mathbf{S} = \frac{1}{n} \sum_i^n (\mathbf{X}_i - \bar{\mathbf{X}})(\mathbf{X}_i - \bar{\mathbf{X}})'$  covariance matrix
- ▶  $\mathbf{S}_{-j} = \frac{1}{n-1} \sum_{i \neq j} (\mathbf{X}_i - \bar{\mathbf{X}}_{-j})(\mathbf{X}_i - \bar{\mathbf{X}}_{-j})'$  jackknifed covariance matrix

It can be shown that

$$\mathbf{S}_{-j} = \frac{n}{n-1} \mathbf{S} - \frac{2}{n} (\mathbf{X}_j - \bar{\mathbf{X}}_{-j})(\mathbf{X}_j - \bar{\mathbf{X}}_{-j})'$$

which quantifies the size of the perturbation

# A first influence index

$$\ell_\lambda^S(\Theta) = \log \det \Theta - \text{tr}(\Theta S) - \lambda \|\Theta\|_1 ; \lambda \geq 0$$

- ▶  $\hat{\Theta} = \arg \max \ell_\lambda^S(\Theta)$ : MLE of  $\Sigma^{-1}$ , the inverse of the concentration matrix, based on  $X_i, i = 1, \dots, n$
- ▶  $\widehat{\Theta}_{-j} = \arg \max \ell_\lambda^{S-j}(\Theta)$ : MLE of  $\Sigma^{-1}$  based on  $X_i, i \neq j$

Let  $\underline{\Theta} = (1_{\theta_{ij} \neq 0})_{1 \leq i, j \leq n}$  a matrix of 0's and 1's: adjacency matrix

Let  $h_1(j)$  be the number of edges affected by the removing observation  $j$

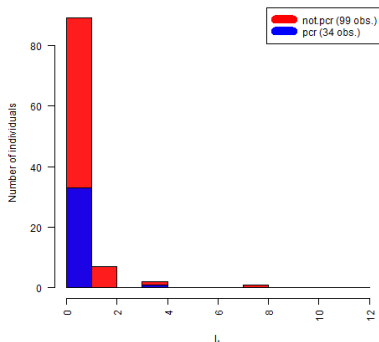
$$h_1(j) = \frac{1}{2} \|\underline{\hat{\Theta}} - \underline{\widehat{\Theta}_{-j}}\|_0$$

## A first influence index: example

Let  $\underline{\Theta} = (1_{\theta_{ij} \neq 0})_{1 \leq i, j \leq n}$  a matrix of 0's and 1's

Let  $h_1(j)$  be the number of edges affected by the removing observation  $j$  ( $j = 1, \dots, 133$ )

$$h_1(j) = \frac{1}{2} \|\hat{\underline{\Theta}} - \widehat{\underline{\Theta}}_{-j}\|_0$$

Histogram of  $h_1$  for cancer dataset

## Link between influence and likelihood

- ▶ Strong links between jackknife and likelihood (influence function as derivative of the statistic)
- ▶ the  $L_1$  penalized log-likelihood of  $S_{-j}$  can be expressed in terms of  $S$ :

$$\ell_{\lambda}^{S_{-j}}(\Theta) = \log \det \Theta - \frac{n}{n-1} \text{tr}(\Theta S) - \frac{1}{n} (x_j - \bar{x}_{-j})' \Theta (x_j - \bar{x}_{-j}) + \lambda \|\Theta\|_1$$

$$\ell_{\lambda}^{S_{-j}}(\Theta) = \ell_{\lambda}^S(\Theta) - \frac{1}{n} (x_j - \bar{x}_{-j})' \Theta (x_j - \bar{x}_{-j})$$

- ▶ The effect is to add a  $L_2$  term that taking into account the contribution of  $x_j$  to the penalized likelihood
- ▶ A natural definition of influence could be given by  $(x_j - \bar{x}_{-j})' \hat{\Theta} (x_j - \bar{x}_{-j})$

## A second influence index

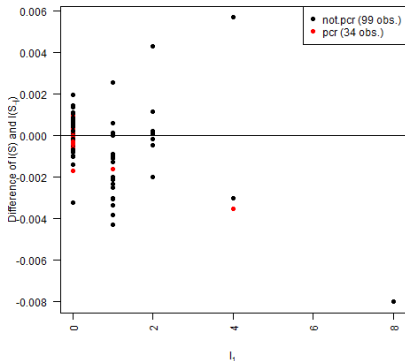
Let  $l_2(\cdot)$  be the difference of the likelihoods induced by the removing of one observation

$$\begin{aligned}l_2(j) &= \ell_{\lambda}^S(\hat{\Theta}) - \ell_{\lambda}^{S-j}(\widehat{\Theta}_{-j}) \\ &= \frac{1}{n}(x_j - \bar{x}_{-j})' \hat{\Theta} (x_j - \bar{x}_{-j})\end{aligned}$$

## Link between the two influence indices on the example

$$\blacktriangleright I_1(j) = \frac{1}{2} \|\hat{\Theta} - \hat{\Theta}_{-j}\|_0 \text{ versus } I_2(j) = \ell_{\lambda}^S(\hat{\Theta}) - \ell_{\lambda}^{S-j}(\hat{\Theta}_{-j})$$

for the 133 observations of cancer dataset



Fluctuations of maximum likelihood of concentration matrix ( $I_2(j)$ ) is not enough to infer stability of adjacency matrix ( $I_1(j)$ )

## Remark: Influence measuring stability of the links through jackknife

Reference graph is generated from the whole dataset and influence of a perturbation induced by the deletion of an observation can be measured by any distance between  $\Theta$  and  $\Theta_{-i}$

Let  $J_1(a, b)$  be the number of times that status of edge  $(a, b)$  is changed by the removing of one observation

$$J_1(a, b) = \sum_{i=1}^n \mathbb{1}_{|\widehat{\Theta}(a,b) - \widehat{\Theta}_{-i}(a,b)| \neq 0}$$

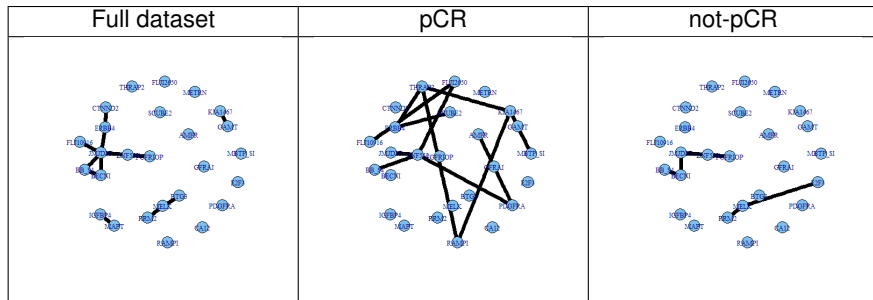
$25 \cdot 26 / 2 = 325$  possible edges and for each edge the theoretical range of  $J_1$  is between 0 and 133.

0	1	2	3	4	5	7	9	11
295	7	3	2	1	1	1	1	1



# The two groups of cancer data set

Patient response to the treatment is classified as either a pathologic complete response (pCR) 34 individuals or a residual disease (not-pCR) 99 individuals



## Above influence functions

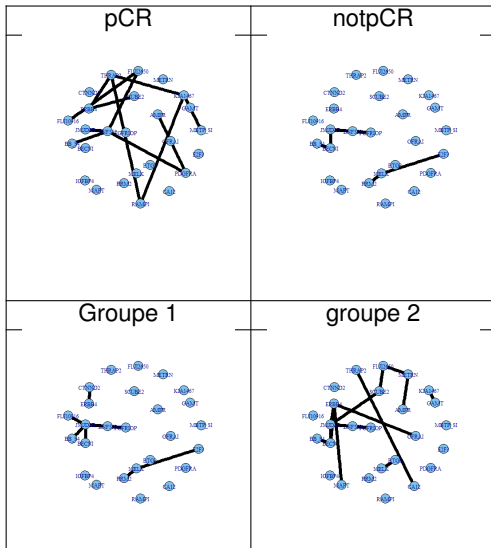
Which class is the less affected by removing or adding observation  $i$ ?

- ▶ Two classes: pCR/not-pCR and two adjacency matrices  $\underline{\Theta}^{(1)}$  and  $\underline{\Theta}^{(2)}$
- ▶ Let  $\underline{\Theta}^{(k \vee i)} = \underline{\Theta}^{(k)}$  if the observation  $i$  is from class  $k$  and  $\underline{\Theta}^{(k \vee i)}$  is the adjacency matrix computed from (individuals of class  $k$  + individual  $i$ )
- ▶  $I_1^k(i)$  be the number of edges of  $\underline{\Theta}^{(k \vee i)}$  affected by removing of observation  $i$  ( $k = 1, 2$ ).

$$I_1^k(i) = \frac{1}{2} \|\widehat{\underline{\Theta}^{(k \vee i)}} - \widehat{\underline{\Theta}_{-i}^{(k \vee i)}}\|_0$$

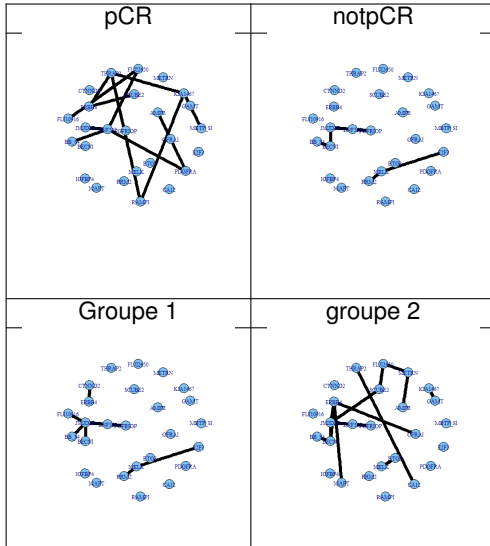
- ▶ For each  $i$  we can compute  $\arg \min_k I_1^k(i)$ .

# Which class is the less affected by removing or adding observation $i$ ?



	0	1	Sum
1	95	28	123
2	4	6	10
Sum	99	34	133

# Which class is the less affected by removing or adding observation $i$ ?



	0	1	Sum
1	95	28	123
2	4	6	10
Sum	99	34	133

Full dataset

# Clustering?

	0	1	Sum
1	95	28	123
2	4	6	10
Sum	99	34	133

- ▶ What about iterate **But ...** one group becomes empty (small group have large variability)
- ▶ Second idea: define a class centroid (open question)
- ▶ What about stability with respect to starting point (related to centroid definition)

## References

- ▶ Ambroise, C., Chiquet, J., and Matias, C. (2009). *Inferring sparse Gaussian graphical models with latent structure*. *Electron. J. Stat.*, 3:205–238.
- ▶ Friedman, J., Hastie, T. and Tibshirani R. (2008). *Sparse inverse covariance estimation with the graphical Lasso*. *Biostatistics*, 9:432–441.
- ▶ Giraud, C. , Huet, S. and Verzelen, N. (2012). *Graph selection with GGMselect*. *SAGMB*, Vol. 11 (3) 1544–6115.
- ▶ Hess, K.R., Anderson, K., Symmans, W.F., Valero, V., Ibrahim, N., Mejia, J.A., Booser, D., Theriault, R.L., Buzdar, U., Dempsey, P.J., Rouzier, R., Sneige, N., Ross, J.S., Vidaurre, T., Gomez, H.L., Hortobagyi, G.N., and Pustzai, L. (2006). *Pharmacogenomic predictor of sensitivity to preoperative chemotherapy with paclitaxel and fluorouracil, doxorubicin, and cyclophosphamide in breast cancer*. *Journal of Clinical Oncology*, 24(26):4236-4244.
- ▶ Meinshausen N., Bühlmann P. (2006). *High-dimensional graphs and variable selection with the Lasso*. *Annals of Statistics*, 34:1436–1462.
- ▶ Natowicz, R., Incitti, R., Horta, E.G., Charles, B., Guinot, P., Yan, K., Coutant, C., Andr F., Pustzai, R., and Rouzier, L. (2008). *Prediction of the outcome of a preoperative chemotherapy in breast cancer using dna probes that provide information on both complete and incomplete response*. *BMC Bioinformatics*, 9(149).
- ▶ Scutari, M. (2010). *Learning Bayesian networks with the bnlearn R package*. *Journal of Statistical Software*, 35(3).

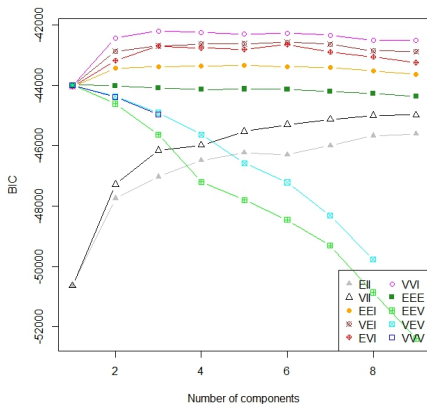
**Thank you for your attention  
(and your questions)**

# Model-based clustering

Let's try Gaussian mixture model :

$$f(x) = pf_1(x) + (1 - p)f_2(x)$$

where  $f_1 \sim \mathcal{N}_p(\mu_1, \Sigma_1)$  and  $f_2 \sim \mathcal{N}_p(\mu_2, \Sigma_2)$ .





## Model-based clustering

*Mclust*: best model: diagonal, varying volume and shape (VVI) with 3 components

	pCR	not-pCR	Sum
1	6	46	52
2	27	16	43
3	1	37	38
Sum	34	99	133

## Model-based clustering

*Mclust*: best model: diagonal, varying volume and shape (VVI) with 2 components

	0	1	Sum
1	29	32	61
2	70	2	72
Sum	99	34	133