Bayesian Networks over clusters: application to gene interactions exploration

Badih Ghattas

Université d'Aix-Marseille
badih.ghattas@univ-amu.fr

ENBIS-SFDS, IHP, Paris, April 2014

Outline

- Bayesian Networks
- Clustering variables using Mutual information
- Networks over clusters
- Simulations
- Application to gene expresion interactions

Motivation

- Work with high dimensional data
- Model interaction between variables which would be difficult to confirm
- Reduce complexity
- Avoid redundancy
- Simplify presentation, give a readable model.

What is often done in high dimensional cases

- Reduce algorithmic complexity, restrict parents
- Eliminate non informative variables
- Statistically anlyse structure, (clusters in the network)
- Reduce presentation comlexity by aggregation

BN over clusters

We suggest a strategy based on "clustering" the set of variables in order to reduce the complexity of learning BNs using the following steps:

- First we apply a suitable clustering procedure to the variables.
- For each cluster we define a representative individual.
- We learn a BN using these individuals.

In the first step we use *mutual information* to construct clusters for the discrete case.
The representative individual for each cluster is chosen using majority vote in the discrete case, and the mean individual in the continuous case.

Clustering

Hierarchical clustering + similarity measure between each pair of variables + heuristic based on the pseudo F and t^2 statistics.

- n: sample size
- G: number of clusters at a given level of the hierarchy
- C_K: the k_{th} cluster, a subset of $\{1, 2, \cdots, n\}$
- N_K: the number of observations in C_K
- $\bar{x}$: the mean observation
- $\bar{x}_K$: the mean observation in cluster C_K
- $T = \sum_{i=1}^{n} \left\| x_i - \bar{x} \right\|^2$
- $W_K = \sum_{i \in C_K} \left\| x_i - \bar{x}_K \right\|^2$
- $P_G = \sum W_j$, where the sum is over the G clusters at level G of the hierarchy.
- $B_{KL} = W_M - W_K - W_L$ if $C_M = C_K \cup C_L$

$$pseudo.F = \frac{(T - P_G) \div (G - 1)}{P(G) \div (n - G)} \quad pseudo.t^2 = \frac{B_{KL}}{(W_K + W_L) \div (N_K + N_L - 2)}$$

The optimal number of clusters corresponds to the first local minimum of the t^2 statistic associated to a local maximum of the pseudo F statistic ([?]).

Distances and dissimilarities

- For the continuous case: Euclidean distance is used.
- For the discret case: Normalized mutual information is used to compute a dissimilarity measure between variables. For two variables X_i and X_j their dissimilarity is computed by:

$$\tilde{I}(X_i, Y_j) = \max_{(s,t)} (I(X_s, X_t)) - I(X_i, Y_j)$$

For the continuous case, mutual information may be used after a suitable discretization of the data.

Cluster representatives

Once the clusters are defined, we will choose a representative for each class. Our choice is different for discrete datasets and for continuous.

- In the continuous case, the representative is the individual which maximizes the correlation with the mean individual of the cluster.
- In the discrete case, the representative is computed as follows:
  - We first determine the "center" of the class by majority vote.
  - The representative is the variable within the class which maximizes the mutual information with the "center".

Experiments

In order to validate our procedure we use simulations to show separately that:

- We may retrieve the right clusters when they exist.
- The BN we construct using class representatives is stable whith respect to the number of classes, and the choice of the representatives.

The R Package "deal" is used for the experiments.

Discrete case simulations

Initial BN: three variables A, B and C, $n = 200$ observations.

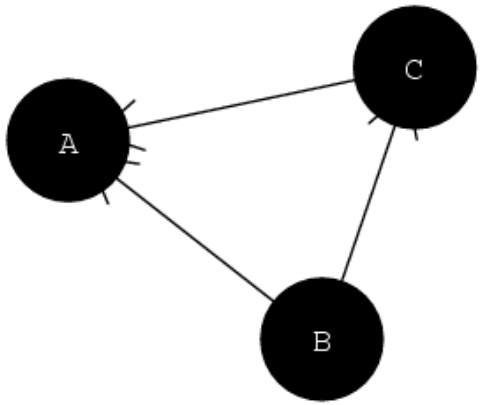


Figure: *Bayesian Network used for simulations in the discrete case.*

The conditional probabilities used in each node are fixed.

A, B et C are the centers of the clusters, 99 neighbors are drawn (perturbing randomly 10% observations). We thus obtain 3 clusters centered on A, B and C containing each 100 variables.

Simulated dataset- MI

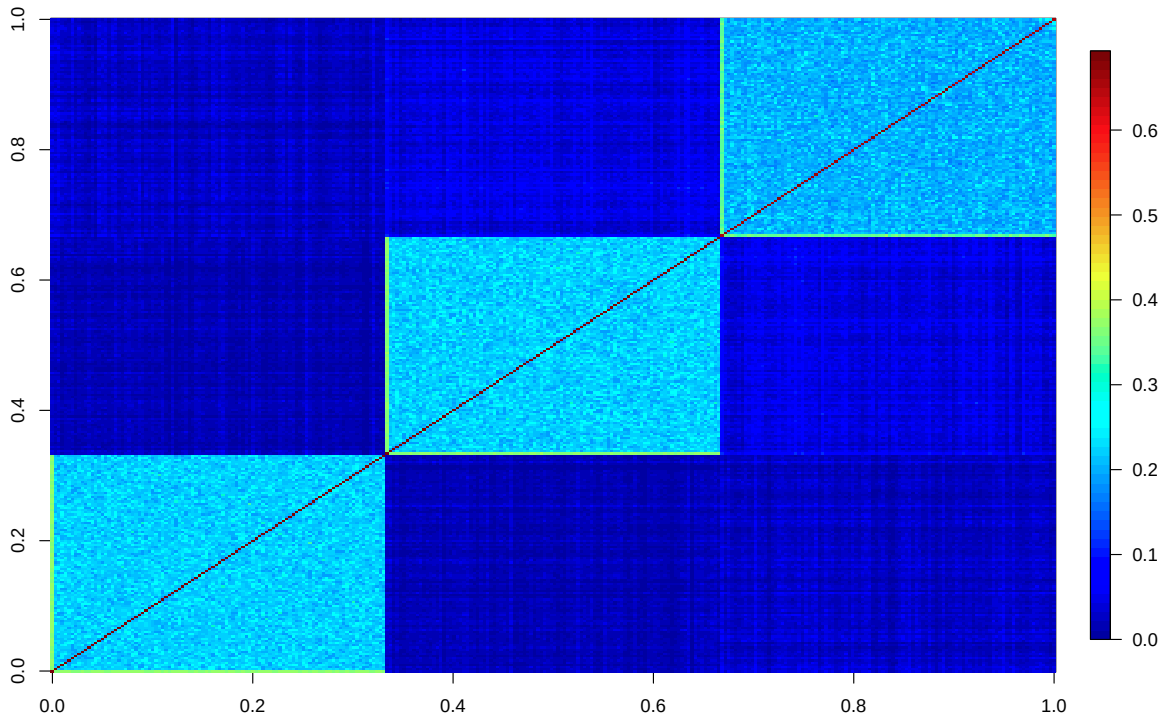


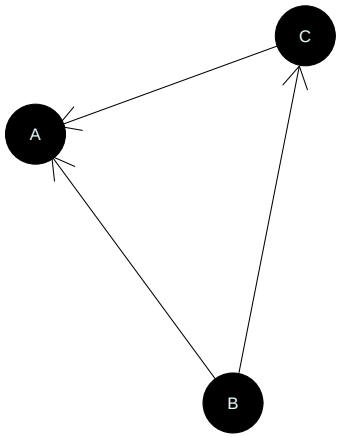
Figure: Mutual Information between variables

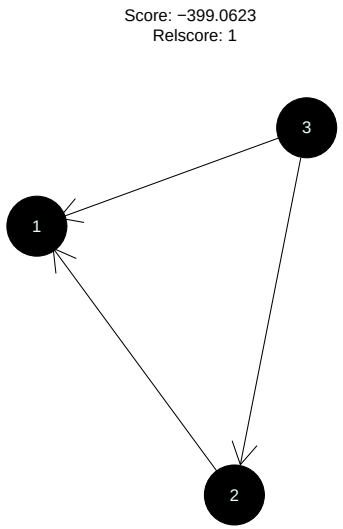
Similarities to Centers

signif	A	B	C	Rep	Nearest	Farest
C1	0.5260	0.0258	0.0591	0.526	0.526	0.269
C2	0.0431	0.5150	0.0753	0.539	0.515	0.262
C3	0.0460	0.0985	0.5470	0.547	0.443	0.278

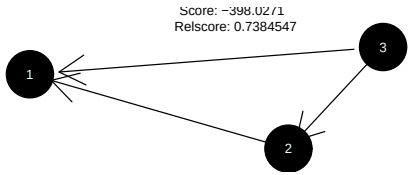
BN over representatives

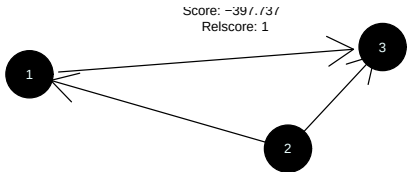
We apply the clustering approach based on mutual information for the 150 variables. The three clusters are retrieved without difficulties.
We will successively replace each variable of the estimated network once by its nearest and farest neighbor within its class.

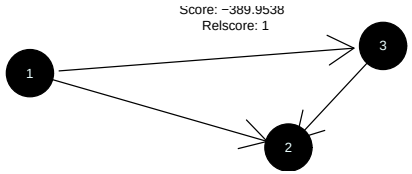


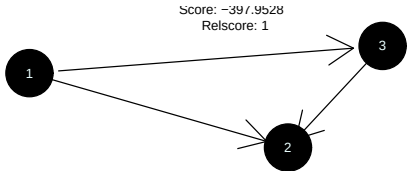


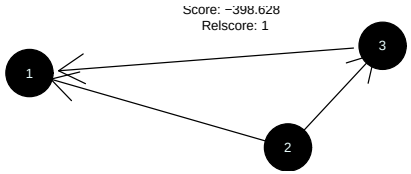
BN over modified representatives











Comparing the Networks

	Change in node 1		Change in node 2		Change in node 3	
	nearest	farest	nearest	farest	nearest	farest
Estimated Network before perturbation						
number of entering vertices	0	0	2	2	1	1
number of exiting vertices	2	2	0	0	1	1
Comparing the estimated network and the perturbed ones						
number of similar vertices	2	3	3	3	3	0
number of inversed vertices	1	0	0	0	0	3
number of omitted vertices	0	0	0	0	0	0
number of new vertices	0	0	0	0	0	0
Distances between representative and its suppliant						
Mutual information	0.64703	0.61369	0.64941	0.65025	0.6199	0.58942

Table: *Perturbing the network on representatives.*

The perturbed networks are quite similar to the estimated ones. This shows that classes are enough homogeneous and the choice of the representatives is good enough.

Continuous Case

Consider the same BN as in the discrete case, for a continuous node A having two continuous parents B and C, its probability distribution is Gaussian $N(m_0 + m_B \times B + m_C \times C, \sigma^2)$ (m_0, m_B and m_C are the regression coefficients of A over B and C).

The simulation scheme is the same as in the discrete case.
The representatives are identified once the neighbors are generated and clustering is done.

The BN obtained has the same score as the initial one. Variables perturbation (replacing a class representative by another individual of the same class) degrades the results more than in the discrete case.

Microarray dataset

Drosophila Melanogaster is one of the most studied organisms in biology ans serves as a model to analyze the genetic and developmental processes which are common for pluri-cellular eucaryots, humans in particular. The genome codes for 14000 genes.
The hypothesis usually admitted is that genes having the same profile, should share the same transcriptional regulation mecanism. In this situation such genes should also participate to common functional processes.

Consider n genes $g_1, \ldots, g_n$, n being too big (~ 14000). The random variable X_i corresponds to the expression level of gene g_i, and the set $\{X_1, .., X_n\}$ constitutes the nodes of the graph. Vertices of the graph represent interactions between genes (an oriented arc between X_i X_j corresponds to an interaction between genes g_i and g_j.

Microarray dataset

The available data come from Affymetrix microarrays analyzed by T. Lecuit group from the LGPD. A collection of embryons is studied at 5 instances of the celularization and development stages of the Drosophila:

- T_0 is the earlier stage of the development. After fecondation, only one cell is available with several kernels. That is the syncitial state. During clivage stage (which takes two hours thirty), kernels are subdivied but not the cells. They migrate then to the cell periphery.
- During cellularization $(T_1, T_2) = 1h$, an epithelium is formed ....
- The gastrulation stage (T_3, T_4) is characterized by morphological movements of the embryon. Embryons genome activation begins at the end of clivage. Zygotic genes are expressed mainly from the cellularization stage and on.

For each time point (T_0 T_4), three independent experiences are carried. We have thus 13986 variables (the genes) and 15 observations (time*repetition).

Applying our procedure

We apply the different stages of our procedure as follows:

- First each variable is discretized to produce binary variables using the median.
- We determine the number of clusters in the data: five clusters are detected.
- Data are then clustered using the mutual information dissimilarity and Ward criterion for clusters aggregation.
- Representatives of each class are computed by majority vote over the discrete version of the data.
- Finally, a Bayesian network is inferred over the representatives.

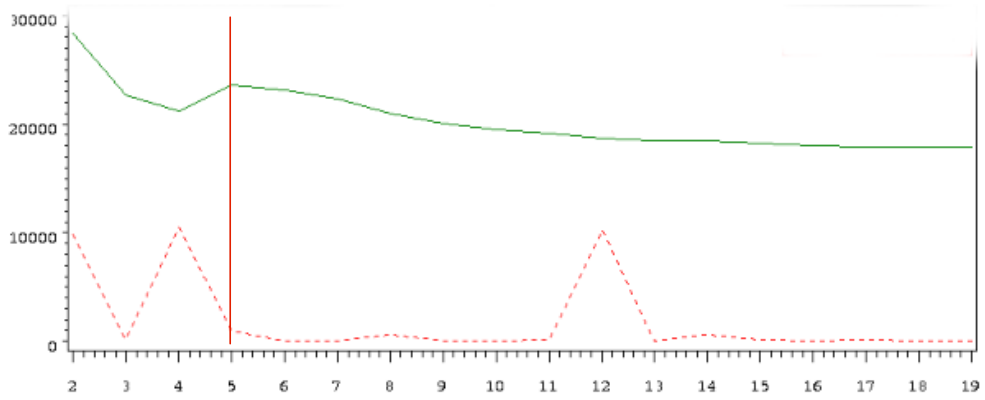


Figure: *Number of clusters(x-axis) estimation. The dashed line correspond to the pseudo t^2 statistic and the full one to the pseudo F.*

Analyzing annotations within the clusters

First we do a functional analysis of the genes within each cluster using the GOToolBox.
Annotations distributions within the clusters are analysed using GOToolBox and Gene Ontology resources (GO), in terms of the biological processes involving the considered genes.
Within each cluster we detect the over represented functionalities.
The most frequent annotations as well as the overrepresented ones are different from one cluster to the other. The distributions within a cluster are conserved when a cluster is split to two new clusters.
As an example lets consider one of the nodes where we have 2184 genes. Only 940 genes are annotated in Gene Ontology.

With respect to annotations

Consider for example the *"Cytosolic ribosome"* annotation: 16 genes are annotated for this term within a cluster where 940 genes are annotated. One question arises: "for this term, if we randomly select 940 genes in the Drosophila genome, what is the probability that 16 within them are annotated knowing that there are 104 annotated genes in the whole genome ?". The p-value of the test gives the answer to that question.

This annotation appears at the first position, it is thus the most overrepresented.

These results show that the obtained clusters are different with respect to the co-expression and co-regulation, and that each cluster is homogeneous in this sense.

The networks

we have inferred the Bayesian network for several clustering results, 3, 4 and 5 clusters. The obtained networks are represented on figure 4.

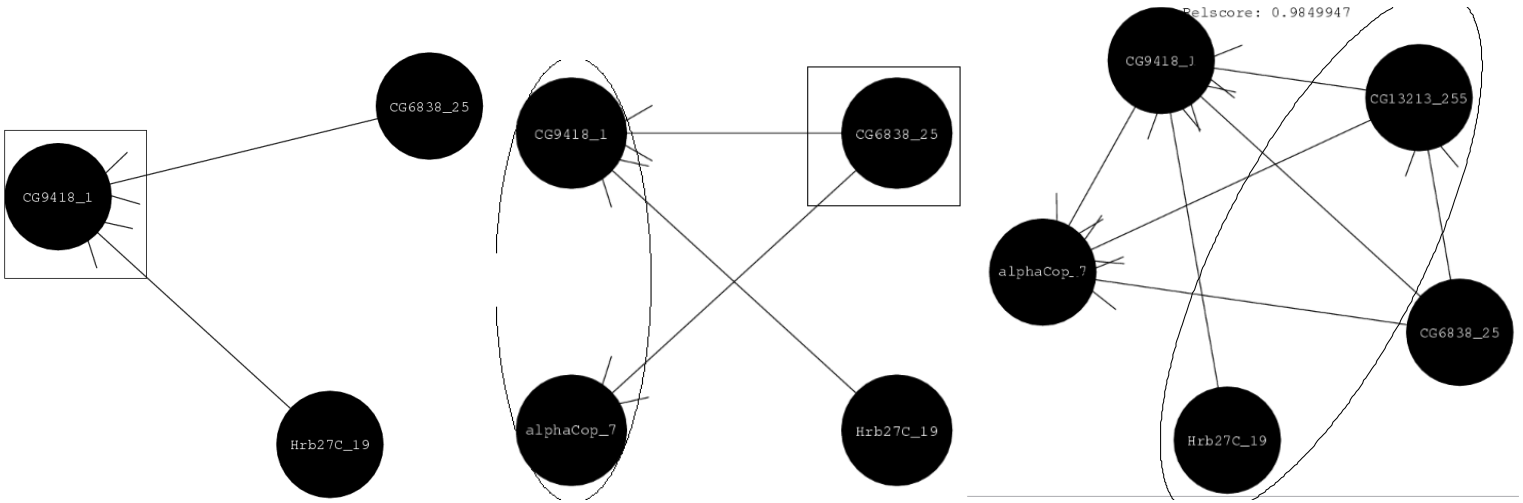


Figure: *Bayesian networks inferred over representatives using 3, 4 and 5 clusters*

When moving from 3 to 4 clusters, two clusters keep the same representatives. The third cluster from the 3 nodes network is split in two subclasses. The representative of this third cluster is still representative of one of the new sub clusters. The oriented vertices between the initial clusters are also kept. The same kind of observations appear when passing from 4 to 5 clusters.

Future Work

- Big Data
- Features selection in unsupervised learning
- Working with 3 dimensional data: MRI - PET scans Volumes. Networks over Regions.
- ...