Inferring gene regulation networks

Christophe Giraud[1,2]

1. Université de Nice - Sophia Antipolis,
Parc Valrose,
06108 Nice, France

2. INRA, Laboratoire MIA,
Domaine de Vilvert,
78352 Jouy-en-Josas, France

e-mail: christophe.giraud@unice.fr

Abstract

A current challenge in system biology is to infer the regulation network of a family of p genes from a n-sample of microarrays, with n (much) smaller than p. Gaussian graphical models are simple models to describe these regulation networks. We propose a procedure that performs Gaussian graph estimation by model selection. We introduce a collection of candidate graphs and then select one of them by minimizing a penalized empirical risk. We pay a special attention to the maximum degree of the graphs that we can handle and asses the performance of the procedure in a non-asymptotic setting. The good theoretical properties of the procedure are confirmed on numerical examples.

Keywords. Gene Regulation Networks, Gaussian Graphs, Model Selection, Sparsity.

Biological systems involve complex networks of interactions between entities such as genes or proteins. These networks can be conveniently represented by a graph. Each vertices of the graph corresponds to a protein or a gene, and an edge between two vertices represents a direct interaction. For example, Figure 1 records 1948 (known) interactions between 1458 proteins of the yeast. Recent biotechnological tools enable to produce a huge amount of proteomic or transcriptomic data. One of the challenge of the post-genomic is to infer the functional interactions between the genes or the proteins from these data. The task is challenging for the statistician due to the very high-dimensional nature of the data. For example, microarrays measure the expression level of a few thousand genes (typically 4000) whereas the sample size n is no more than a few tens. Since the number of possibles interactions between p genes is $p(p-1)/2$ (nearly ten millions if $p = 4000$), it seems hopeless to try to infer these interactions from $n \approx 20$ microarrays. This task is actually possible (up to some extend) thanks to the sparsity of the interaction network.

Valuable tools for analyzing the network of interactions are the Gaussian Graphical Models. The vector of the expression levels of the p genes is modeled by a Gaussian variable in $\mathbf{R}^p$. The Gaussian graph then represents the conditional dependences between the coordinates. More precisely, if $X = (X_1, \ldots, X_p)$ represent the expression levels of the p genes, the graph has an edge between the genes i and j if and only if X_i *is not independent of* X_j *conditionally on the other variables*. The goal of the statistician is to infer these edges from a n-sample of the variables X. The edges correspond to the non-zero entries of the partial correlation matrix, so when the sample size n is larger than p, a possible algorithm to infer the edges is to threshold the inverse of the empirical covariance matrix. This strategy is no more possible when n is (much) smaller than p and several new algorithms have been proposed. Unfortunately, the real performance of these algorithms are mostly unknown: the few theoretical results are only valid under restrictive conditions on the covariance matrix and they assume that the sample size n tends to infinity.

We propose a new statistical procedure to estimate the graph of conditional dependences of X. We first introduce a collection of candidate graphs and then select one of them by minimizing a penalized empirical risk. The performance of the procedure is assessed in a non-asymptotic setting without any hypotheses on the covariance matrix. These good theoretical properties of

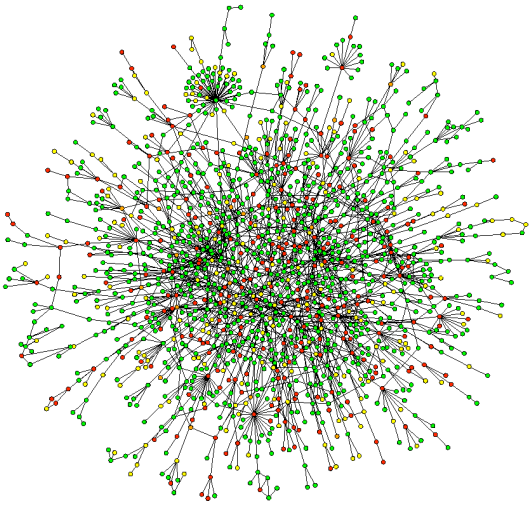

Figure 1: Protein-Protein interaction network of the yeast.

the procedure are confirmed by numerical results. Since we are interested on the maximal "size" of the graph that we can infer, we pay a special attention to the maximal degree D of the graphs that we can handle. This maximal degree turns to be roughly $n/(2\log(p/D))$, which means that p should stay small compared to $de^{n/(2d)-1}$, where d is the degree of the graph of conditional dependences.