CHRISTOPHE GIRAUD, Université de Nice, laboratoire J.A. Dieudonné, Parc Valrose, 06108 Nice, France; et INRA, laboratoire MIA, Parc Vilvert, 78352 Jouy-en-Josas, France

Estimation of Gaussian Graphs by Model Selection

Biological systems involve complex networks of interactions between entities such as genes or proteins. One of the challenge of the post-genomic is to infer these networks from high-throughoutput data produced by recent biotechnological tools. 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.

Valuable tools for analyzing these network of interactions are the Gaussian Graphical Models. The vector $X = (X_1, \ldots, X_p)$ of the expression levels of the p genes is modeled by a Gaussian variable in $\mathbf{R}^p$. Then, the Gaussian 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.

We propose a statistical procedure to estimate the graph of conditional dependences from 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 the procedure are confirmed by numerical results. We pay a special attention to the maximal degree D of the graphs that we can handle, which turns to be roughly $n/\big(2\log(p/D)\big)$.