INFERRING GENE NETWORKS FROM MICROARRAY WITH GRAPHICAL MODELS

Antonino Abbruzzo[1] and Angelo M. Mineo[1]

SEAS - Dipartimento di Scienze Economiche Finanziarie e Statistiche
Università degli Studi di Palermo
Viale delle Scienze Ed., 13, 90128 Palermo, Italy
(e-mail: antonino.abbruzzo@unipa.it)
(e-mail: angelo.mineo@unipa.it)

ABSTRACT. Microarray technology allows to collect a large amount of genetic data, such as gene expression data. The activity of the genes are coordinate by a complex network that regulates their expressions controlling common functions, such as the formation of a transcriptional complex or the availability of a signalling pathway. Understanding this organization is crucial to explain normal cell physiology as well as to analyse complex pathological phenotypes. Graphical models are a class of statistical models that can be used to infer gene regulatory networks. In this paper, we examine a class of graphical models: the strongly decomposable graphical models for mixed variables. Among others properties, explicit expressions of maximum likelihood estimators are available for decomposable graphical models. This property makes the use of decomposable model suitable for high-dimensional data. We apply decomposable graphical models to a real dataset example.

1 INTRODUCTION

Microarray technology has become more prevalent in biology over the last decade. A microarray is a collection of microscopic DNA spots attached to a solid surface. Microarrays are used to measure the expression levels of large numbers of genes, simultaneously. The activity of the genes can be described by a complex network that regulates their expressions controlling common functions, such as the formation of a transcriptional complex or the availability of a signalling pathway. Understanding how our genes work together as a network could i) hold the potential for new treatments and preventive measures in disease, ii) add a new level of complexity to scientists' knowledge of how DNA works to integrate and regulate cell functionality. So, the need of statistical tools to analyse and extract information from such data has become crucial.

Graphical models are useful to infer conditional independence relationships between random variables. The conditional independence relationships can be visualized as a network with a graph. Graphs are object with two components: nodes and links. Nodes are in one-to-one correspondence with random variables and links represent relations between genes. If a link between two genes is absent this means that these two genes are conditional independent given the rest. Pairwise, local and global Markovian properties are the connections between graph theory and statistical modelling. Unfortunately, classic graphical models cannot be applied to high-dimensional data due to computational reasons. Recently, different techniques have been proposed to overcome this computational limitation.

A branch of research works on penalized graphical models. The idea is to penalize the maximum likelihood function, for example with the ℓ_1-norm, to produce sparse solutions. The main assumption of these models is that the networks are sparse, which means many of the variables are conditionally independent from the others. The most known algorithm to estimate sparse graphs is probably the graphical lasso (glasso) proposed by Friedman *et al.* (2008). This models cannot deal with dataset which include mixed variables. A dataset contains mixed variables if both qualitative and quantitative variables have been collected which is a common situation in microarray studies.

One approach to deal with mixed data, where the measurements can be binary, ordinal and continuous, is the semiparametric Bayesian copula graphical models (Hoff (2007)). The semiparametric Bayesian copula graphical model uses the assumption of copula Gaussianity on the multivariate latent variables which are in one-to-one corresponds with the observed variables. So, the methodology can also be seen as a latent variable method for non-Gaussian multivariate data. Furthermore, the model can deal with missing data at random. Conditional dependence and regression coefficients as well as credible intervals can be obtained from the analysis. Moreover, copula Gaussian graphical models allow to impute missing data. Imputation of missing data is essential in microarray analysis since the number of statistical units is usually very small. This means that every piece of information becomes precious. For example, in the analysis of breast cancer data 65 measurements were collected over 62 units. If we remove the missing data, we would reduce the units to 25. However, for higher-dimensional problems the Bayesian copula approach becomes problematic due to its computational complexity and convergences of the proposal distribution.

In this paper, we examine strongly decomposable graphical models for analysing mixed data. A strongly decomposable graphical model is a graphical model whose graph neither contains cycles of length more than three nor forbidden path. A path exits between nodes A and B if one can reach A from B in a finite number of steps. A forbidden path is a path between two not adjacent discrete nodes which pass through continuous nodes. The distributional assumption is that random variables are conditional Gaussian distributed. Even tough, this assumption could be too restrictive, there exits some techniques to transform the data, for example Box and Cox transformation. A model selection procedure is necessary in order to obtain a parsimonious graph which fits the data. A modification of the Chow and Liu's (1968) algorithm can be used to do model selection for this class of models. This procedure is based on an initial estimation of a forest. A forest is a disjoint union of trees. A tree is a connected acyclic undirected graph.

The rest of this paper is organized as follows. In Section 2, we briefly recall the methodology used to infer decomposable graphical models for mixed data. In Section 3, we show an application of the methodology to a real dataset which contains mixed variables that are the expression level of genes collected in a microarray experiment and some clinical information of the patients. Finally, we conclude with a discussion.

2 METHODOLOGY

A graph is a couple $G = (V,E)$ where V is a finite set of nodes and $E \subset V \times V$ is a subset of ordered couples of V. Nodes are in one-to-one correspondence with random variables. Sup-

pose we have p discrete and q continuous nodes and write the sets of nodes as Δ and Γ, where $V = \{\Delta \cup \Gamma\}$. Let the corresponding random variables be (I,Y), and a typical observation be (i,y). Here, i is a p-tuple containing the values of the discrete variables, and y is a real vector of length q. Let $\Im$ indicates the set of possible i. Let the probability that $I = i$ be p_i, and assume the distribution of Y given $I = i$ is multivariate normal $N(\mu_i, \Sigma_i)$ so that both the conditional mean and covariance may depend on i. This is called the conditional Gaussian (CG) distribution. The density can be written as:

$$f(i,y) = p(i)\,(2\pi|\Sigma(i)|)^{-\frac{1}{2}} \exp\left(-\frac{1}{2}\left(y-\mu(i)\right)^T \Sigma_i^{-1}(y-\mu(i))\right) = \exp\left(g(i) + h(i)^T y - \frac{1}{2} y^T K(i) y\right),$$

where:

$$K(i) = \Sigma(i)^{-1}, \qquad h(i) = K(i)\mu(i), \qquad g(i) = \log p(i) + \frac{1}{2}\log|K(i)| - \frac{1}{2}h(i)^T K(i) h(i) - \frac{q}{2}\log(2\pi),$$

are called canonical parameters. This models are also called mixed interaction models and are defined by constrained canonical parameters of the CG-distribution. For example, let $\Delta = \{A,B\}$ and $\Gamma = \{X,Z\}$, and let the levels of the factors A and B be denoted j and k. So in this case $i = (j,k)$ and $y = (x,z)$. The joint density can be written:

$$f(i,y) = \exp\{g(i) + h(i)^x x + h(i)^z z - \frac{1}{2}(k_{xx}x^2 + 2k_{xz}xz + k_{zz}z^2)\},$$

and we can write the unrestricted (or saturated) model as

$$g(i) = u + u_j^a + u_k^b + u_{jk}^{ab}, \qquad h(i)^x = v + v_j^a + v_k^b + v_{jk}^{ab}, \qquad h(i)^z = w + w_j^a + w_k^b + w_{jk}^{ab},$$

$$K = \begin{pmatrix} k_{xx} & k_{xz} \\ k_{zx} & k_{zz} \end{pmatrix}.$$

In our example the covariance matrix $\Sigma(i)$ is constant over i and we refereed to such models as homogeneous. More details on the conditional Gaussian distribution can be found in Lauritzen and Wermuth (1989). Explore the space of all possible models is infeasible even for low-dimensional problems. The number of possible graphs is $2^{\binom{n}{2}}$, so for example we would have to explore 1024 with 5 nodes.

Model selection for high-dimensional mixed data can be carried out through the methodology proposed by Edwards *et al.* (2010). This technique uses a modification of the Chow and Liu's (1968) algorithm. This algorithm requires a square matrix of weights of dimension $(p+q)\times(p+q)$ which indicates the magnitude of the relation between the nodes. Edwards *et al.* (2010) proposed measures based on the minimization of a measure such as BIC-type or AIC penalized mutual information criteria.

Next, we give some details on this procedure. Firstly, the maximum weights spanning tree is found. Chow and Liu showed that finding the maximum likelihood tree can be formulated as finding a maximum weight spanning tree, a task for which highly efficient algorithms exist. This approach requires that all edges-weights are calculated. Then, Kruskal's algorithm (Kruskal (1956)) is applied to find a maximum weight spanning tree. This algorithm starts with the null graph and successively selects edges $\{e_1,\ldots,e_r\}$, where $e_k = (v_i, v_j) \in V$. If

edges $e_1, \ldots, e_r$ are selected, the algorithm selects an edge e such that: 1) $e \notin \{e_1, \ldots, e_r\}$ and $\{e_1, \ldots, e_r, e\}$ is a forest, and 2)e has maximum weight among all edges satisfying the first point.

Secondly, the forest is used as a starting point in order to find the best decomposable graphical model according to a measure. The estimation strategy then consists in restricting the search space to models with explicit estimates, i.e. decomposable models. A key result is that: if $M_0 \subset M_1$ are decomposable models differing by one edge $e = (v_i, v_j)$ only, then e is contained in one clique C of M_1 only, and the likelihood ratio test for M_0 versus M_1 can be performed as a test of $v_i \perp v_j | C_{\setminus \{v_i, v_j\}}$. These computations only involve the variables in C. It follows that for likelihood-based scores such as AIC or BIC, score differences can be calculated locally which is far more efficient then fitting both M_0 and M_1 and then stored, indexed by v_i, v_j and C, so that they can be reused again if needed in the course of the search. This can lead to considerable efficiency gains.

3 ANALYSIS OF BREAST CANCER DATA

In this section we analyse breast cancer data. The data comes from a study performed on 62 biopsies of breast cancer patients over 59 genes. These genes were identified using comparative genomic hybridization. Continuous measures of expression levels of those 59 genes were collected. In order to link gene amplification/deletion information to the aggressiveness of the tumours in this experiment, clinical information is available about each of the patients: age at diagnosis (AGE), follow-up time (Surv.Time), whether or not the patient died of breast cancer (C.Death), the grade of the tumour (C.Grade), the size of the tumour (Size.Lesion), and the Nottingham Prognostic Index (NPI). C.Death is a dichotomous variable, C.Grade is ordinal with three categories and NPI is a continuous index used to determine prognosis following surgery for breast cancer. NPI values are calculated using three pathological criteria: the size of the lesion; the number of involved lymph nodes; and the grade of the tumour. The complete dataset results in 62 units and 65 variables with a 3.6% of missing data which are imputed by using the semiparametric Bayesian copula graphical model proposed by Hoff (2007). The advantage of using this methodology to impute missing data is that it takes into account the multivariate structure of the data and it can deal with binary, ordinal and continuous measurements.

Our aim is to find a network which could describe the relationships between the 65 variables which are the nodes of the graph and highlight the relationships between gene expression levels and clinical variables. We use the package `gRapHD` (Abreu *et al.* (2010)) to analysis the breast cancer data. Firstly, the forest that minimize the BIC is found by applying the function `minForest`. This result in a quite simple graph with at last $p+q-1$ links. A more complex model can be found by applying the function `stepw`. This function performs a forward search strategy through strongly decomposable models starting from a given decomposable graphical model. At each step, the edge giving the greatest reduction in BIC is added. The process ends when no further improvement is possible. A convenient choice of the starting model is the minimal BIC forest, but other arbitrary decomposable models may be used.

We start the analysis by selecting the minimal BIC forest. Then, we use the stepwise function. The estimated graph appears to be incoherent with our expectation. NPI is a linear combination of Size.Lesion, C.Grade and N.lymph. So, we expect a graph which contains links between NPI and Size.Lesion and NPI and C.Grade. Instead, the links between Size.Lesion and NPI were missing in the estimated graph. These missing links were due to a forbidden path. The forbidden path C.Death - Size.Lesion - NPI - C.Death would had appeared if we connect NPI with Size.Lesion. So, we use a different graph as a starting point. We add a link between C.Grade and C.Death to the minimal BIC forest. Note that the probability of dying due to cancer given that the cancer grade is at level three is 11 times greater with respect to the probability of dying due to cancer given that the cancer grade is at level one.

Figure 1 shows the heterogeneous strongly decomposable graphical model with starting point the minimum BIC forest with the addition of the link between C.Grade and C.Death.

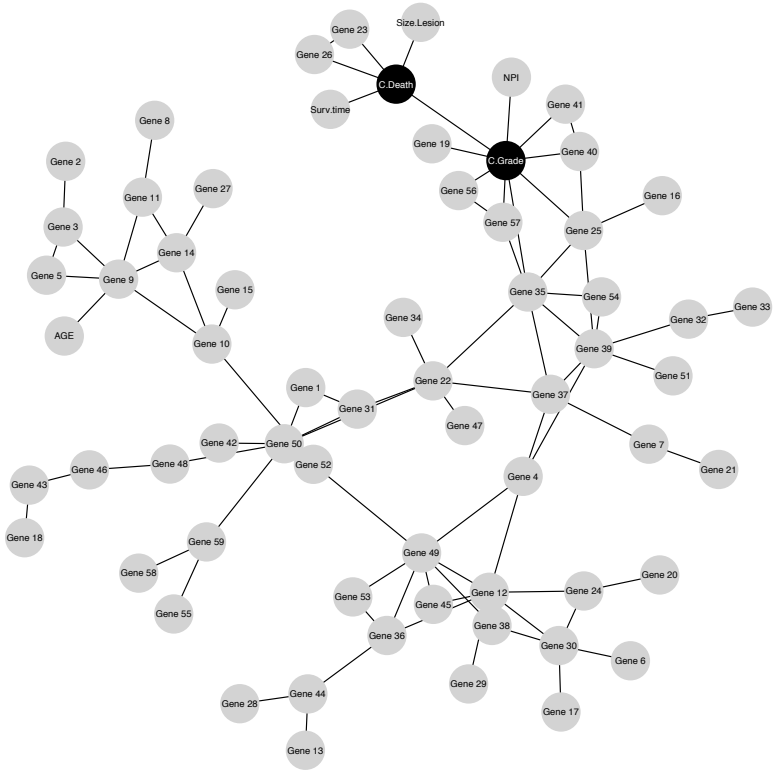

Figure 1. Graph obtained by applying the heterogeneous strongly decomposable graphical model to breast cancer data with starting point a minimum BIC forest with a link between C.Grade and C.Death. Black nodes indicate discrete variables while grey nodes represent continuous variables.

Even though the addition of this link in the starting graph creates a path between NPI and Surv.time, the expected link between Size.Lesion and NPI is still missing. This is due

to another model limitation which is no cycle of length more than three can be estimated. The most connected node is C.Grade which is connected with 9 other nodes namely: NPI, Gene 41, Gene 40, Gene 19, Gene 15, Gene 57, Gene 35 and Gene 25. Other high connected nodes are Gene 9 and Gene 12 which are connected with 8 other nodes and Gene 49 which is connected with other 7 nodes. Unfortunately, we do not have the names of the genes but only the labels, so we can only highlight some of the characteristics of the graph. None of the gene resulted isolated from the other. Gene 23 and 26 are conditional independent given C.Death of the rest. Although from graphs inspection, we could be tempted to say that a relation emerges between the 59 genes and of the clinical conditions of the patients, robust statistical procedures should be used to investigate the credibility of such relation.

4 DISCUSSION

In this paper, we have explored a class of graphical models, the strongly decomposable graphical models, which can be used to infer networks for high-dimensional mixed data. The algorithm performances is feasible and the theory of decomposable graphs have been investigated. There are some limitations due to the assumption of decomposable models. This means that neither cycle of length more than 3 nor forbidden path can be estimated. We have shown in a real example on breast cancer data that this can bring at wrong interpretation of the network. So, careful attention should be paid during the analysis. In this case, the identification of a missing important link was possible due to the way we construct the NPI index but in general case this identification could be a very difficult task. The algorithm cannot deal with missing data so we have used a Bayesian approach in order to impute such data. In future works we will try to extend the algorithm to manage dataset with missing data. In our opinion, model selection and robustness of the networks are other two points that deserve further investigations.

REFERENCES

ABREU, G.C., EDWARDS, D., LABOURIAU, R. (2010): High-Dimensional Graphical Model Search with the gRapHD R Package, *Journal of Statistical Software*, *37*, 1–18.

CHOW, C., LIU, C. (1968): Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, *14*, 462–467.

EDWARDS, D. (2000): *Introduction to graphical modelling*. Second edition. Springer Text in Statistics. Springer Verlag, New York.

EDWARDS, D., DE ABREU, C.G., LABOURIAU, R. (2010): Selecting high-dimensional mixed graphical models using minimal AIC or BIC forests. *BMC bioinformatics*, *11*, 11–18.

FRIEDMAN, J., HASTIE, T., TIBSHIRANI, R. (2008): Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, *9*, 432–441.

HOFF, P. (2007): Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, *1*, 265–283.

KRUSKAL, J.B (1956): On the Shortest Spanning Subtree of a Graph and the Traveling Salesman Problem. *Proceedings of the American Mathematical Society*, *7*, 48–50.

LAURITZEN, S., WERMUTH, N. (1989): Graphical models for association between variables, some of which are qualitative and some quantitative. *The Annals of Applied Statistics*, *17*, 31–57.