Graphical Modeling for Discrete Random Variables with Application to Tissue Microarray (TMA) Experiments

Corinne Dahinden
Life Science Zurich PhD Program on Systems Biology of Complex Diseases
Seminar für Statistik, ETH Zürich Switzerland

Keywords: Graphical models, Interactions, Lasso, Log-linear models
AMS: 68R01, 62P10

Abstract

Tissue microarrays (TMA) are composed of hundreds of tissue sections from different patients arrayed on a single glass slide. With the use of immunohistochemical staining, they provide a high-throughput method of analyzing potential biomarkers on large patient samples. The assessment of the expression level of a biomarker is usually performed by the pathologist on a categorical scale.

The analysis of the interaction of these biomarkers and in particular the estimation of the graphical model associated with the underlying discrete random variables, are of biological importance. Questions such as how the interaction pattern changes with survival time are of direct biological interest. However, the estimation of the interaction structure requires sophisticated techniques. Our approach is to fit an ℓ_1-regularized log-linear model assuming a multinomial sampling scheme in order to obtain the graphical model. The regularization becomes necessary as after cross-tabulation of the samples in contingency tables, many cell entries remain zero, leading to so-called sparse contingency tables, where standard procedures fail to work.

We compare our graphical model fiting procedure with other algorithms in a simulation study. Then the proposed method is applied to a dataset consisting of tissue microarray measurements from renal cell carcinoma patients to graphically model the distribution of the underlying discrete random variables for patients with long survival time compared to patients where early death occurs. To further argue in favor of the reasonability of the proposed models, we had a closer look at the differing edges in the models and used these to validate the graphs by examining the corresponding survival curves.

1. Introduction

The central motivation that led to this work was the analysis of Tissue Microarray experiments for the use of validating known pathways and in addition to detect changes to those in groups with e.g. high survival time compared to short survival. Tissue microarray technology allows rapid visualization of

molecular targets in thousands of tissues at a time, either at DNA, RNA or protein level. They are powerful for validation and extension of findings obtained from genomic surveys such as cDNA microarrays. cDNA microarrays are useful to analyse a huge number of genes like a couple of thousand in one specimen at a time. In contrast, TMAs are applicable to the analysis of one target, denoted as biomarker in the following, at a time, but in up to 1000 tissues on each slide.

Our primary intererst is now in the association among these biomarkers. The goal as far as this question is concerned is to model the association in a so-called graphical model.

2. Graphical Models

Graphical models are a marriage of graph and probability theory. They provide a framework to graphically represent dependence and independence relationsof a given distribution. Generally speaking, graphical models consist of a graph in which the nodes represent random variables in a probability distribution and the edges of the graph represent the conditional dependence structure of the probability distribution. More formally graphs and graphical models are defined as follows:

Definition (Graph): A graph $\mathbb{G} = (V, E)$ consists of a set of vertices or nodes V and a set of edges $E \subseteq V \times V$. These can either be directed or undirected. An edge is called *directed* if $(X, Y) \in E$ but $(Y, X) \notin E$.

Terminology: For a directed graph $\mathbb{G} = (V, E)$, the *parents* of X is the set of nodes $Pa_X = \{Y|(Y, X) \in E\}$. The *children* or *descendant* of X is the set of nodes $\{Y|(X, Y) \in E\}$.

Definition (Graphical Model): The *directed* graph $\mathbb{G} = (V, E)$ together with the probability distribution P is called a *Graphical Model* $(\mathbb{G}, P)$, if $(\mathbb{G}, P)$ satisfies the *Markov Condition:*

$$Y \perp ND_Y | PA_Y \ \ \forall \ \ Y \in V,$$

where ND_Y stands for the *non-descendants* of Y. An *undirected* graph together with a probability distribution P is called *Graphical Model* if P satisfies the *undirected Markov Condition*, which means that for all distinct vertex sets $A, B, S \in V$ it holds that:

$$A \perp B|S, \text{ whenever S separates A from B.}$$

S separates A from B in V if there are no paths from A to B using only vertices in $V \backslash S$. The graph which represents the conditional independence relations

$$X \perp Y | V \backslash \{X, Y\}, \forall \, X, Y \in V$$

is called *Conditional Independence Graph.*

It is not true that one class, either directed or undirected, is more powerful.

There are independence relations, which can only be expressed using directed graphical models, whereas there are distributions which cannot be represented in a directed graph. For both models it holds that we can directly read off the conditional independence relations by the concept of *d-separation* in the case of directed graphical models (for details see [2]) and *separation* in undirected models. If the distribution allows a graphical representation with a directed acyclic graph, where all conditional independence relations can be read off by the concept of *d-separation*, $\mathbb{G}$ and P are called to be *faithful* to each other. Although DAGs and CIGs can represent different families of distributions, there is a correspondence between the two which is embodied by the moralization rule ([2]). Because of this result we can obtain the CIG from a faithful DAG representation by transforming the arrows into lines and linking unlinked parents with moral edges.

Our goal is to estimate Conditional Independence Graphs in a first step. Future work will include a second step, where whenever a consistent DAG extension is possible, the corresponding CIG is further processed into a DAG. In the following, we describe different algorithms to fit CIG. For the directed model fitting procedures, the moralized graph of the DAG will be considered, which is the underlying conditional independence graph.

3. Algorithms

In this section we describe the algorithms which we will consider in the following. Our methodological contribution is described in detail, whereas for the methods to compare, only a brief overview is given with the corresponding reference.

3.1. Regularization Approach

We fit a conditional independence graph using log-linear models, where the logarithm of the contingency table cell probability is a linear function of some specific design matrix and $\boldsymbol{\beta}$.

$$\log \mathbf{p} = \mathbf{U}\boldsymbol{\beta}$$

$$span(U) = span(u_1) \bigoplus span(u_2) \bigoplus \ldots \bigoplus span(u_{12}) \bigoplus \ldots$$

The design matrix U consists of the popular $u-$term parametrization first introduced by [1]. Vectors belonging to vector spaces spanned by different $u - terms$ are orthogonal. And we choose orthogonal subspaces of $u - terms$. We now estimate our $\boldsymbol{\beta}$-vector by minimizing the *group-ℓ_1-penalized* negative log-likelihood l.

$\boldsymbol{\beta} = (\beta_1, \ldots, \beta_{12}, \ldots, \beta_{1\ldots k}) = (\beta_k, k \in \mathcal{P}(V))$ or $\boldsymbol{\beta}$ restricted to $(\beta_k, k \in \mathcal{P}_t(V))$

$$\widehat{\boldsymbol{\beta}}^{\lambda} = \arg\min_{\boldsymbol{\beta}} \left[-l(\boldsymbol{\beta}) + \lambda \sum_{j \in \mathcal{P}(V)} ||\beta_j||_{\ell_2} \right], \text{ where}$$

$$l(\boldsymbol{\beta}) = \log P_{\boldsymbol{\beta}}(\mathbf{n}) \propto \sum_i n_i (\mathbf{U}\boldsymbol{\beta})_i / \sum_i n_i$$

The advantage of choosing a group-penalty, is that model selection on interaction level is performed instead of including or excluding single parameters. This is done in a level-wise approach. To decide upon the model complexity, the interaction level is chosen as the one minimizing the cross-validated negative log-likelihood. The $\boldsymbol{\beta}$ vector is then translated into a graphical model. The CIG builds up by connecting the nodes corresponding to non-zero β_k components.

3.2. Graphical Model Fitting Algorithms

PC Algorithm: Is a two-stage algorithm, where first the skeleton of the graphical model is identified by performing a number of conditional independence tests. Then the edges are orientated using information from the first stage (see [4]).

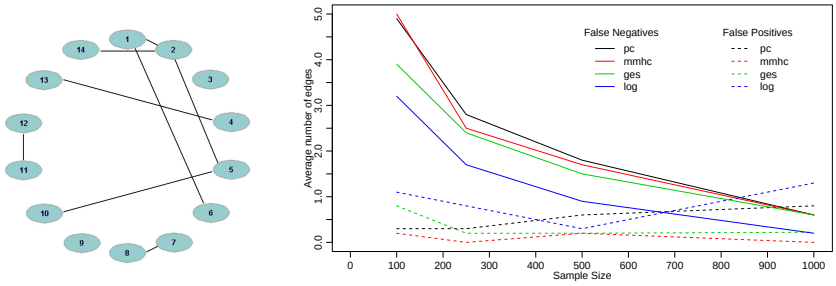
Greedy Equivalent Search (GES): A greedy search algorithm which under some assumptions for the prior distribution identifies the most probable a posteriori graphical model. The detailed algorithm can be found in [3].

Max-Min Hill-Climbing (MMHC): Max-Min Hill-Climbing, identifies first the parents and children set of each variable and then performs a greedy hill-climbing search beginning with an empty graph only adding edges in the parents or children set. The algorithm is explained in [5].

4. Simulation Study

We simulate datasets of sizes 100, 250, 500, 1000 and 2500 according to the graph in Figure 1, which has 14 nodes and 9 edges. We then compare the conditional independence graphs identified by the different procedures. We

Figure 1: Left: The graphical model defining the simulation study. Right: Summary of the results of the simulation study.



see that the loglinear regularization approach is very competitive overall and clearly superior to the other algorithms as far as false negatives are concerned. The simulation study is averaged over 10 simulations of the according size. We see that for 1000 samples, the loglinear approach almost never includes a false negative. We can therefore assume that the conditional independence relations which can be derived from the resulting graph are very reliable. On the other hand by considering the false positives, we can conclude that there

might be more independence relations, which can't be read off from the graph. This would call for a hybrid version of the loglinear model approach, where the resulting CIG estimation would be plugged in as a very good starting model for another graphical model fitting procedure, such as the PC algorithm.

5. Application to Tissue Microarray Dataset

5.1. Dataset

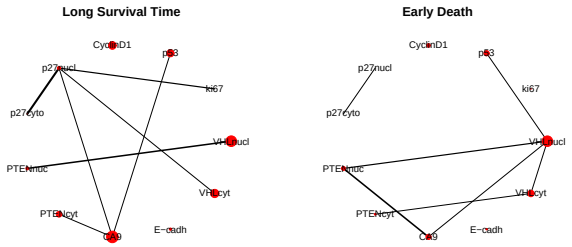
We do have data from 1116 renal cell carcinoma patients. As this cancer type is quite frequent, some interesting genes have been identified and there exist hypothesis about some pathways. TMA are now the perfect tool to further evaluate those genes and analyze what exactly happens in these pathways in a large scale study.

Up to now 11 biomarkers are observed in a total of 1116 patients. This number is constantly growing and new biomarkers are being analyzed. The measurements are categorical, some of them binary. We also have information such as tumor grade or survival time.

5.2. Results

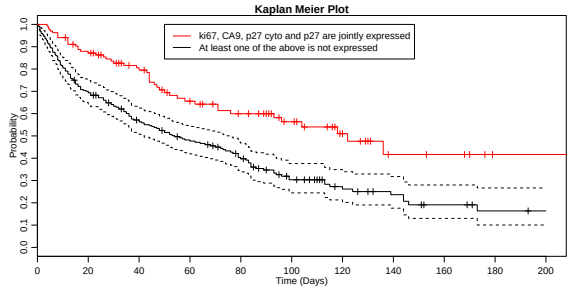
By dividing the samples in groups of patients with high survival time and such of low survival time, we hope to be able to see some differences in the corresponding distribution of the biomarkers i.e. we hope to see some difference in the graphical model representation between these two groups. The models are displayed in Figure 2.

Figure 2: Graphical Models for groups of patients with high survival time and low survival time



A cross-validation study revealed that it is sufficient to consider no higher order interaction than first order. In Figure 2, the thickness of the edges are proportional to the absolute value of the first order interaction, whereas the diameter of the nodes are proportional to the absolute value of the corresponding main effect. We can clearly see the difference as far as the biomarkers p27nucl, p27cyto, CA9 and ki67 are concerned. If we have a closer look at exactly these biomarkers, we observe, that if these are jointly expressed, the patient seems to have a much better prognosis in terms of survival, which can

Figure 3: Kaplan Meier plot of the survival time for patients with the indicated biomarkers expressed compared to patients which at least do not express one of them.



be seen in Figure 3. This supports the assumption that the corresponding graphical models are very reasonable and in addition is also an interesting biological result with direct consequences for the patient. In addition, which was highly surprising in these models, was that VHL is in boths graphs associated with PTEN, even though these two biomarkers are supposed to act in two different pathways where no interaction was expected. This will be subject of further exploration.

6. Summary

We have proposed an algorithm to fit graphical models for discrete random variables. This algorithm has been tested in a simulation study and then applied to a TMA dataset. The application provided some interesting insight into the underlying biological process and is currently subject of further exploration.

7. Bibliography

[1] N.W Birch. Maximum likelihood in three-way contingency tables. *Journal of the Royal Statistical Society. Series B*, 25:220–233, 1963.

[2] Steffen L. Lauritzen. *Graphical Models*. Oxford Statistical Science Series, 17. Oxford Clarendon Press, 1996.

[3] Richard E. Neapolitan. *Learning Bayesian Networks*. Prentice Hall, April 2003.

[4] Peter Spirtes, Clark Glymour, and Richard Scheines. *Causation, Prediction, and Search*, volume 1. The MIT Press, Cambridge, Massachusetts, London, England, 2nd edition, 2000.

[5] Tsamardinos, Ioannis, Brown, Laura, Aliferis, and Constantin. The max-

min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, October 2006.