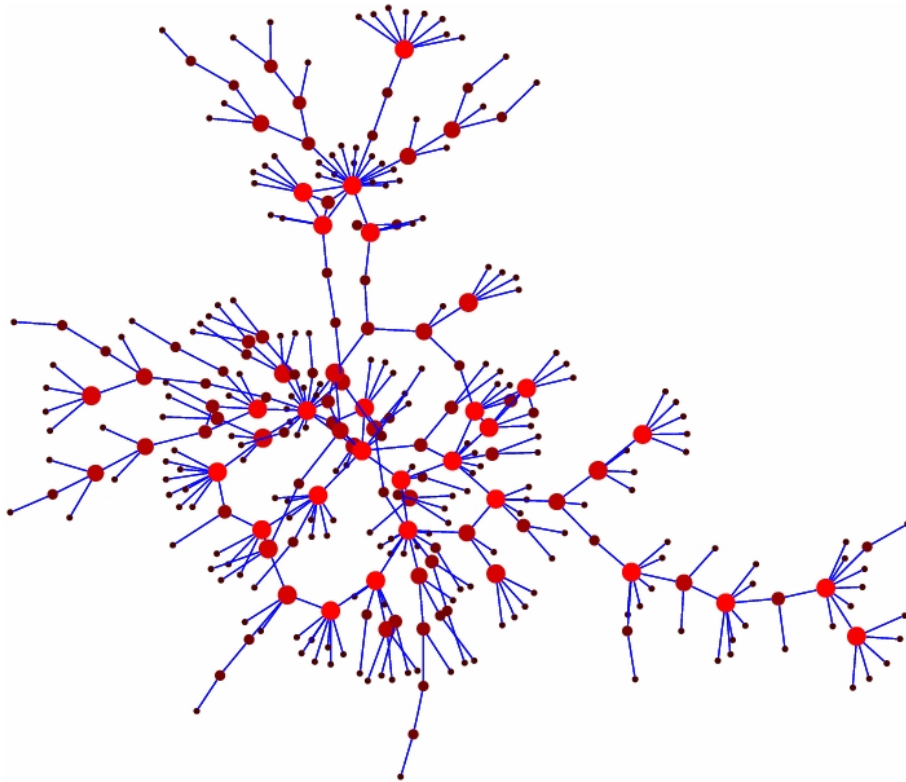


PH.D THESIS ON STATISTICS

Learning Probabilistic Networks in Large and Structured Domains

MASSIMILIANO MASCHERINI



DEPARTMENT OF STATISTICS

UNIVERSITY OF FLORENCE, ITALY

The image in the previous page:
M. E. J. Newman, The structure and function of complex networks, SIAM Review 45, 167-256 (2003).

UNIVERSITÀ DEGLI STUDI DI FIRENZE
DIPARTIMENTO DI STATISTICA “*Giuseppe Parenti*”

DOTTORATO IN STATISTICA APPLICATA
XVIII CICLO



Learning Probabilistic Networks in Large and Structured Domains

Massimiliano Mascherini

Supervisor:
Prof. Federico M. Stefanini

Director of Graduate Studies:
Prof. Fabrizia Mealli

Introduction

...πολλῶν χαλιῶν ἐργον οἰάων θ' ἀμα,
Σοφοκλέα

...*the bridle and the rudder too,*
Sophocles
(496-406 B.C.)

The main focus of this thesis is on learning Bayesian Networks in large and structured domain. This work is composed by four papers and one appendix. Each paper is self-contained with its bibliography and figures, tables and equation numbering. Parts and bits therefore appear in more than one paper. Original methods developed in such papers have been implemented in the R environment, [13], and the user's manual of the suite of functions created is included in the appendix.

The starting point of this thesis has been gene expression problem through microarray experiments. Without enter in technicalities, and addressing the reader to [34] for a wider review of microarray experiments, DNA microarrays have been usefully exploited to screen gene expression in comparative experiments, for example, different classes of patients affected by breast cancer, [32, 33]. The problem of the analysis of microarray data is that they produce data structures characterized by a large number of variables and very few replications as well as many different sources of bias affect the raw data, making the reproducibility and reliability of results low, [26].

Given all these problems, the question that a statistician has to face is: How to extract information from the huge amount of microarray data, to rebuild a genetic network, i.e. a causal model describing the gene regulation process?

In order to answer to this question we developed methods to learn probabilistic networks in large and structured domains characterized by regularities supposed to be also present in genetic networks, as described in the following.

Paper1: A review on learning the structure of Bayesian Networks representing influence relations among genes

The intent of this paper is twofold: first, to provide a review on structural learning of Bayesian Networks; second, to describe some key features of microarray experiment which are performed to assess differential gene expression.

Bayesian Networks (BNs), [8, 14, 24], are among the leading technologies to describe and derive conditional independence relations existing among random variables. Addressing to [8, 14] for terminology and theoretical aspects, a Bayesian network is defined as a directed acyclic graph that encodes the joint probability distribution for a set of random variables. The nodes in the graph represent the random variables and missing arrows between the nodes, specify properties of conditional independence between the variables. Therefore BNs are an effective way to characterize probabilistic and causal relations among variables.

The appropriate graphical model that best depicts the problem domain is often provided by an expert, and the BN is ready to be used for inferential purpose. There are also many situations in which the appropriate domain expertise is not available, so from the beginning the user is obliged to look to data to suggest what modelling assumptions might be appropriate. The prevalence and the importance of such problem is reflected in the rapidly expanding interest in the general area of the structural learning of probabilistic network from data, an area that naturally merges with a number of more general field of research, including statistical modelling, where the extraction of information from data is a classical challenge, and machine learning.

In this paper we provide a wide review of different approaches of learning a Bayesian Networks from data. We focus our attention on the constraint-based approach, where we mainly describe the PC and the NPC algorithm, [29, 30] and on the score & search approach, where we depict Penalized Likelihood metrics, [1, 28], Bayesian metrics, [7, 10] and information theory based metrics, [16] as well as heuristic strategies as HillClimbing, [6, 27], Simulated Annealing, [15], and Genetic Algorithm, [12, 18].

Paper2: Encoding structural prior information to learn Bayesian Networks

By the review provided on paper 1, we decided to move in the score & search approach, focusing on bayesian metrics, in which the score value assigned to a candidate structure is its posterior probability, given the data and a prior distribu-

tion over structures. We noticed that if a lot of efforts have been done to encode the likelihood, [7, 10], not much interest has been dedicated to the prior distribution on structure. In large and structured domains as the microarray area, the expert knowledge domain is quit rich and the elicitation of this information can improve the overall performance of the structural learning process. Most of the approaches developed in the literature to elicit the a-priori distribution on Directed Acyclic Graphs (DAGs) require a full specification of graphs, [5, 10, 11]. Nevertheless, expert's prior knowledge about conditional independence relations may be weak, making the elicitation task troublesome. Moreover, the detailed specification of prior distributions for structural learning is NP-Hard, [6], namely in large networks the elicitation is not practical. This is the case, for example, of gene expression analysis, in which a small degree of graph connectivity is a priori plausible and where substantial information may regard dozens against thousands of nodes.

In order to solve this problem, we propose an elicitation procedure for DAGs which exploits prior knowledge on network topology, and that is suited to large Bayesian Networks. Then, we develop a new quasi-Bayesian score function, the *P*-metric, to perform structural learning following a score-and-search approach. We implement the new metric in the R environment, [13], using the package DEAL, [4].

Finally the proposed metric is tested with two different benchmark dataset, the ASIA network, [19], and the HGH network, [20], with successful results.

Parts of paper 2 have been already published in [21] as Working Paper 2005/13 - Statistics Department - University of Florence.

Paper3: M-GA, a genetic algorithm to search for the best Conditional Gaussian Bayesian Networks

Having defined a new metric, we then focus on the problem of the search of optimal Bayesian Network from a database of observations. Being this problem NP-hard, several heuristic search strategies have been found to be effective and fully justified, [6].

Moving on the framework of the Evolutionary Computation, [3, 17, 25, 35] we present here a new population-based algorithm, the M-GA, to learn the structure of Bayesian Networks without assuming any ordering of nodes and allowing for the presence of both discrete and continuous random variables.

The M-GA algorithm is a variation of the classic GA, proposed by [18] for Bayesian Networks. It differs by considering Conditional Gaussian BNs that are evaluated using the BDe metric, [10] as well as the offspring production process is here deeply changed and innovated in order to improve the genetic variability among generations.

The M-GA algorithm is implemented in the R environment using the package DEAL, [4]. Numerical performances of our Mixed-Genetic Algorithm, (M-GA), are investigated on a case study taken from the literature, [2] and compared with the greedy search with successful results.

Many of the results of the paper 3 have been already presented at the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation, Vienna, 28-30 November 2005, and published in [22] by the IEEE Computational Intelligence Society.

Paper4: The normalization of DNA microarrays using spike controls: an additive model

When working on learning probabilistic networks, meanwhile we cooperated with a team of molecular biologists of the Department of Internal Medicine of the University of Florence, in the analysis of the DNA microarrays data. The paper 4 was born from this cooperation and it is focused on a model to normalize DNA microarrays using spike controls.

The study of a metabolic pathway is focused on a limited number of genes in comparison to whole-genome studies. Therefore, a large number of (spike) controls may be printed within slide together with several ESTs replicates to reduce bias and variance of estimates, [9, 31].

In this paper, a linear additive mixed effect model is developed to remove dye and spatial biases using spike controls. The iterated weighted least squares algorithm is adapted to obtain a fast algorithm to search for the optimal model and to perform point estimates of model parameters. Actual data from a very noisy calibration slide have been successfully normalized following our model.

Many of the results of this paper have been submitted in [23], for publication.

Appendix: MASTINO: a suite of R functions to learn Bayesian Networks.

All the methods developed in paper 2 and paper 3 have been implemented in MASTINO, a suite of R functions, written in R, [13], and coded on the top of the package DEAL, [4].

In particular, MASTINO extends the package DEAL, and it provides the implementation of the P -metric to evaluate Bayesian Networks, of the M-GA to search for the best Conditional Gaussian Bayesian Networks as well as a lot of other utility functions suited to manipulating Bayesian Networks.

The package MASTINO can be downloaded from the web page <http://www.ds.unifi.it/mascherini/MASTINO> and may be used under the terms of the GNU, General Public License Version 2.

References

- [1] H. Akaike. Information theory and an extension of the maximum likelihood principle. *Int Symp Information Theory 2nd*, pages 267–281.
- [2] J. H. Badsberg. *An environment for Graphical Models*. Aalborg University, Aalborg, Denmark, 1995.
- [3] R. Blanco, I. Inza, and P. Larrañaga. Learning bayesian networks in the space of structures by estimation of distribution algorithms. *International Journal of Intelligent Systems*, 18:205–220, 2003.
- [4] S. G. Bøttcher and C. Dethlefsen. DEAL: A package for learning bayesian networks. *Journal of Statistical Software*, 8(20):1–40, 2003.
- [5] W. L. Buntine. Theory of refinement on bayesian networks. *Proceedings of 7th Conference on Uncertainty in Artificial Intelligence*, pages 52–60, 1991.
- [6] D. M. Chickering, D. Geiger, and D. Heckerman. Learning bayesian networks: Search methods and experimental results. *Preliminary papers of the 5th Intl. Workshop on Artificial Intelligence and Statistics*, pages 112–128, 1995.
- [7] G. F. Cooper and E. Herskovitz. A bayesian method for the induction of probabilistic networks from data. *Machine Learning*, 9:309–357, 1992.
- [8] R. G. Cowell, P. A. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic Networks and Expert Systems*. Springer-Verlag, New York, 1999.
- [9] B. Eickhoff, B. Korn, M. Schick, A. Poustka, and J. van der Bosch. Normalization of array hybridization experiments in differential gene expression analysis. *Nucleic Acids Research*, e33(27), 1999.
- [10] D. Heckerman, D. Geiger, and D. M. Chickering. Learning bayesian network: A combination of knowledge and statistical data. *Proceedings of 10th Conf. Uncertainty in Artificial Intelligence*, pages 293–301, 1994.
- [11] D. Heckerman, C. Meek, and G. Cooper. A bayesian approach to causal discovery. *Technical Report MSR-TR-97-05*, 1997.
- [12] J. H. Holland. *Adaptation in Natural and Artificial System*. University of Michigan Press, Ann Arbor, MI, 1975.
- [13] R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5:299–314, 1996.

- [14] F. V. Jensen. *An introduction to Bayesian Networks*. Springer Verlag, New York, N.Y., 1996.
- [15] S. Kirkpatrick, D. J. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.
- [16] W. Lam and F. Bacchus. Learning bayesian belief networks: An approach based on the mdl principle. *Computational Intelligence*, 10(4):269–293, 1994.
- [17] P. Larrañaga and D. Lozano. *Estimation of Distribution Algorithms. A New Tool for Evolutionary Computation*. Kluwer Academic Publisher, Boston, MI, 2002.
- [18] P. Larrañaga and M. Poza. Structure learning of bayesian networks by genetic algorithms: A performance analysis of control parameters. *IEEE Journal on Pattern Analysis and Machine Intelligence*, 18(9):912–926, 1996.
- [19] S. Lauritzen and D. Spiegelhalter. Local computation with probabilities on graphical structures and their application to expert system. *Journal of the Royal Statistical Society - B Series*, 50(2):157–192, 1988.
- [20] P. Le, A. Bahl, and L. Ungar. Using prior knowledge to improve genetic network reconstruction from microarray data. *InSilico Biology*, 27(4), 2004.
- [21] M. Mascherini and F. M. Stefanini. Encode prior information to learn bayesian networks. *WP of the Department of Statistics - University of Florence*, 13, 2005.
- [22] M. Mascherini and F. M. Stefanini. M-GA: A genetic algorithm to learn conditional gaussian bayesian networks. *Proceedings of the IEEE International Conference on Computational Intelligence for Modelling, Control and Automation*, 2005.
- [23] C. Mavilia, M. Mascherini, V. Martineti, A. Tanini, M. L. Brandi, and F. M. Stefanini. The normalization of dna microarrays using spike controls: an additive model. *Submitted*.
- [24] J. Pearl. *Probabilistic Reasoning in Intelligent Systems*. Morgan Kaufmann, San Mateo, CA, 1988.
- [25] M. Pelikan, D. E. Goldberg, and E. Cantu-Paz. Boa: The bayesian optimization algorithm. *Proceedings of the Genetic and Evolutionary Computation Conference*, pages 525–532, 1999.

- [26] E. Petricoin III, J. Hackett, L. Lesko, R. Puri, S. Gutman, K. Chumakov, J. Woodcock, D. Feigal Jr., K. Zoon, and F. Sistare. Medical applications of microarray technologies: a regulatory science perspective. *Nature Genetics*, 32:474–479, 2002.
- [27] E. Rich and K. Knight. *Artificial Intelligence, 2nd Edition*. McGraw Hill, 1991.
- [28] G. Schwartz. Estimating the dimension of the model. *Annals of Statistics*, 7(2):461–464, 1978.
- [29] P. Spirtes, C. Glymour, and R. Scheines. *Causation, Prediction, and Search, 2nd ed.* MIT Press, New York, N.Y., 2000.
- [30] H. Steck. *Constraint-Based structural learning in bayesian networks using finite data sets*. PhD thesis - University of Munich, Munich, Germany, 2001.
- [31] O. Thellin, W. Zorzi, B. Lakaye, B. De Borman, B. Coumans, G. Hennen, T. Grisar, A. Igout, and E. Heinen. Housekeeping genes as internal standards: use and limits. *Journal of Biotechnology*, 75:291–295, 1999.
- [32] M. van de Vijver, Y. He, L. van't Veer, H. Dai, A. Hart, D. Voskuil, G. Schreiber, J. Peterse, C. Roberts, M. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, T. van der Velde, H. Bartelink, S. Rodenhuis, E. Rutgers, S. Friend, and R. Bernards. A gene-expression signature as a predictor of survival in breast cancer. *New England Journal of Medicine*, 347(25):1999–2009, 2002.
- [33] L. van't Veer, H. Dai, M. van de Vijver, Y. He, A. Hart, M. Mao, J. Peterse, K. van der Kooy, M. Marton, A. Witteveen, G. Schreiber, R. Kerkhoven, C. Roberts, P. Linsley, R. Bernards, and S. Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415:530–536, 2002.
- [34] E. Witt and J. McClure. *Statistics for Microarray: design, analysis, and inference*. John Wiley & Sons, Chichester, 2004.
- [35] M. L. Wong, W. Lam, and K. S. Leung. Using evolutionary computation and minimum description length for data mining of probabilistic knowledge. *IEEE Trans. Pattern Anal. Mac. Intell.*, 21(2):174–178, 1999.

