Measures of Variability for Graphical Models

Marco Scutari

marco.scutari@stat.unipd.it
Department of Statistical Sciences
University of Padova

March 14, 2011

Graphical Models

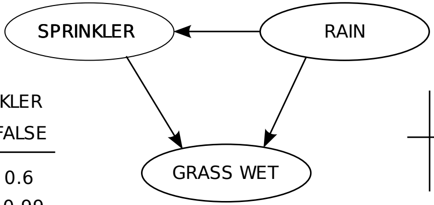
Graphical Models

Graphical models are defined by:

- a network structure, either an undirected graph (Markov networks [3], gene association networks, correlation networks, etc.) or a directed graph (Bayesian networks [9]). Each node corresponds to a random variable;
- a global probability distribution, which can be factorised into a small set of local probability distributions according to the topology of the graph.

This combination allows a compact representation of the joint distribution of large numbers of random variables and simplifies inference on its parameters.

A Simple Bayesian Network: Watson's Lawn



RAIN	SPRINKLER	
	TRUE	FALSE
FALSE	0.4	0.6
TRUE	0.01	0.99

RAIN	
TRUE	FALSE
0.2	0.8

SPRINKLER	RAIN	GRASS WET	
		TRUE	FALSE
FALSE	FALSE	0.0	1.0
FALSE	TRUE	0.8	0.2
TRUE	FALSE	0.9	0.1
TRUE	TRUE	0.99	0.01

The Problem

Most literature on the analysis of graphical models focuses on the study of the parameters of local probability distributions (such as conditional probabilities or partial correlations).

- Comparing models learned with different algorithms is difficult, because they maximise different scores, use different estimators for the parameters, work under different sets of hypotheses, etc. [15].
- Unless the true global probability distribution is known it is difficult to assess the quality of the estimated models.
- The few available measures of structural difference are completely descriptive in nature (i.e. Hamming distance [8] or SHD [21]), and are difficult to interpret.

Focusing on network structures sidesteps most of these issues.

Modelling Undirected Network Structures

Edges and Univariate Bernoulli Random Variables

Each edge e_i in an undirected graph $\mathcal{U} = (\mathbf{V}, E)$ has only two possible states,

$$e_i = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{otherwise} \end{cases}.$$

Therefore it can be modelled as a Bernoulli random variable E_i,

$$e_i \sim E_i = \begin{cases} 1 & e_i \in E \text{ with probability } p_i \\ 0 & e_i \notin E \text{ with probability } 1 - p_i \end{cases},$$

where p_i is the probability that the edge e_i appears in the graph. We will denote it as $E_i \sim Ber(p_i)$.

Edge Sets as Multivariate Bernoulli

The natural extension of this approach is to model any set W of edges (such as E or $\{\mathbf{V} \times \mathbf{V}\}$) as a multivariate Bernoulli random variable $\mathbf{W} \sim Ber_k(\mathbf{p})$. $\mathbf{W}$ is uniquely identified by the parameter set

$$\mathbf{p} = \{p_w : w \subseteq W, w \neq \varnothing\},$$

which represents the dependence structure [10] among the marginal distributions $W_i \sim Ber(p_i)$, $i = 1, \ldots, k$ of the edges.

The parameter set $\mathbf{p}$ can be estimated using m bootstrap samples [4] as suggested in Friedman *et al.* [5] or Imoto *et al.* [7].

Second Order Properties

The marginal variances of the edges are bounded, because

$$p_i \in [0,1] \Longrightarrow \sigma_{ii} = p_i - p_i^2 \in \left[0, \frac{1}{4}\right].$$

Covariances are bounded in the same interval (in modulus). Similar bounds exist for the eigenvalues $\lambda_1, \ldots, \lambda_k$ of the covariance matrix Σ,

$$0 \leqslant \lambda_i \leqslant \frac{k}{4} \qquad \text{and} \qquad 0 \leqslant \sum_{i=1}^{k} \lambda_i \leqslant \frac{k}{4}.$$

Furthermore, if $\mathbf{W_1}$ and $\mathbf{W_2}$ are two multivariate Bernoulli random variables, then they are independent if and only if

$$\mathbf{W_1} \perp\!\!\!\perp \mathbf{W_2} \Longleftrightarrow \mathrm{COV}(\mathbf{W_1}, \mathbf{W_2}) = \mathbf{O}.$$

Measures of Structure Variability

Entropy of the Bootstrapped Network Structures

Consider the graphical models $\mathcal{U}_1, \ldots, \mathcal{U}_m$ learned from the bootstrap samples. Three scenarios are possible:

- minimum entropy: all the models learned from the bootstrap samples have the same structure. In this case:

$$p_i = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Sigma = \mathbf{O};$$

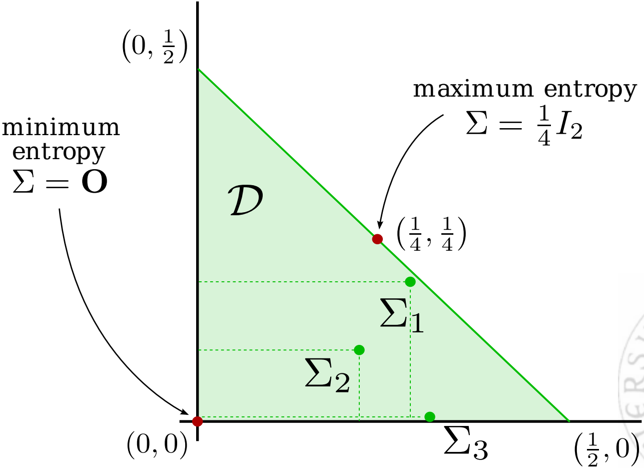
- intermediate entropy: several models are observed with different frequencies m_b, $\sum m_b = m$, so

$$\hat{p}_i = \frac{1}{m} \sum_{b\,:\,e_i \in E_b} m_b \qquad \text{and} \qquad \hat{p}_{ij} = \frac{1}{m} \sum_{b\,:\,e_i \in E_b, e_j \in E_b} m_b;$$

- maximum entropy: all possible models appear with the same frequency, which results in

$$p_i = \frac{1}{2} \qquad \text{and} \qquad \Sigma = \frac{1}{4} I_k.$$

Entropy of the Bootstrapped Network Structures



Univariate Measures of Variability

- The *generalised variance*

$$\mathsf{VAR}_G(\Sigma) = \det(\Sigma) = \prod_{i=1}^{k} \lambda_i \in \left[0, \frac{1}{4^k}\right].$$

- The *total variance* (or *total variability*)

$$\mathsf{VAR}_T(\Sigma) = \operatorname{tr}(\Sigma) = \sum_{i=1}^{k} \lambda_i \in \left[0, \frac{k}{4}\right].$$

- The squared *Frobenius matrix norm*

$$\mathsf{VAR}_N(\Sigma) = |||\Sigma - \frac{k}{4} I_k|||_F^2 = \sum_{i=1}^{k} \left(\lambda_i - \frac{k}{4}\right)^2 \in \left[\frac{k(k-1)^2}{16}, \frac{k^3}{16}\right].$$

Measures of Structure Variability

All of these measures can be rescaled to vary in the $[0, 1]$ interval and to associate high values to networks whose structure display a high entropy in the bootstrap samples:

$$\overline{\mathsf{VAR}}_T(\Sigma) = \frac{4}{k}\mathsf{VAR}_T(\Sigma), \quad \overline{\mathsf{VAR}}_G(\Sigma) = 4^k\mathsf{VAR}_G(\Sigma), \quad \overline{\mathsf{VAR}}_N(\Sigma) = \frac{k^3 - 16\mathsf{VAR}_N(\Sigma)}{k(2k-1)}.$$

Furthermore, these measures can be easily translated into asymptotic or Monte Carlo tests (via parametric bootstrap) having the maximum entropy covariance matrix as the null hypothesis.

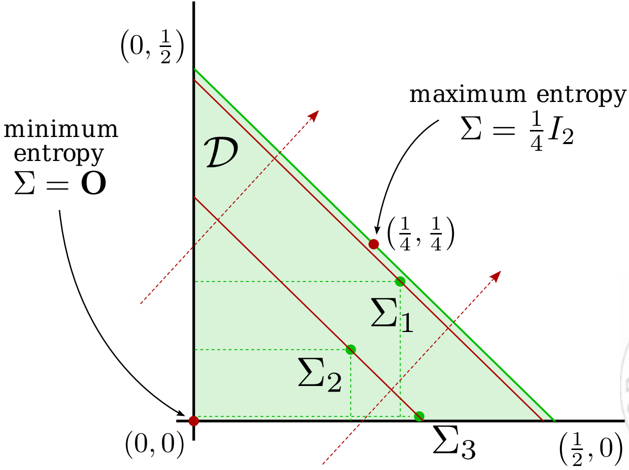
$$4m\,\mathrm{tr}(\hat{\Sigma}) \mathrel{\dot\sim} \chi^2_{mk}$$

$$\sqrt{n}\left[4^k\det(\hat{\Sigma}) - 1\right] \mathrel{\dot\sim} N(0, 2k)$$

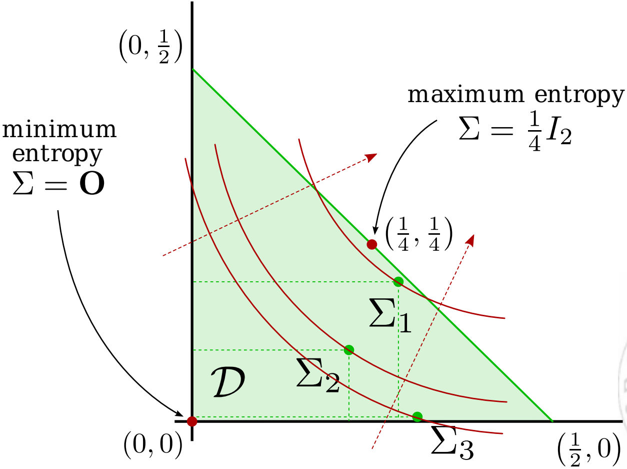
$$\frac{mk}{2}\sqrt[k]{4^k\det(\hat{\Sigma})} \mathrel{\dot\sim} Ga\left(\frac{k(m+1-k)}{2}, 1\right)$$

$$|||\hat{\Sigma} - \frac{1}{4}|||_F^2 \mathrel{\dot\sim} \frac{1}{8m}\chi^2_{\frac{1}{2}k(k+1)}$$

Structure Variability (Total Variance)



Structure Variability (Squared Frobenius Matrix Norm)



Modelling Directed Acyclic Network Structures

Edges and Univariate Trinomial Random Variables

Each arc a_{ij} in a directed acyclic graph $\mathcal{G} = (\mathbf{V}, A)$ has three possible states,

$$a_{ij} = \begin{cases} -1 & \text{if } a_{ij} = \overleftarrow{a_{ij}} = \{v_i \leftarrow v_j\} \\ 0 & \text{if } a_{ij} \not\in A\text{, denoted with } \mathring{a}_{ij} \\ 1 & \text{if } a_{ij} = \overrightarrow{a_{ij}} = \{v_i \rightarrow v_j\} \end{cases},$$

and therefore it can be modelled as a Trinomial random variable A_i, which is essentially a multinomial random variable with three states. Variability measures (and their normalised variants) can be extended from the undirected case as

$$\mathsf{VAR}(A_i) = \mathsf{VAR}(E_i) + 4\mathsf{P}(\overrightarrow{a_{ij}})\mathsf{P}(\overleftarrow{a_{ij}}) \in [0, 1]$$

Edge Sets as Multivariate Trinomials

As before, the natural extension to model any set W of arcs is to use a multivariate Trinomial random variable $\mathbf{W} \sim Tri_k(\mathbf{p})$ and to estimate its parameters via nonparametric bootstrap.

However:

- the acyclicity constraint of Bayesian networks makes deriving exact results very difficult because it cannot be written in closed form;
- the score equivalence of most structure learning strategies makes inference on $Tri_k(\mathbf{p})$ tricky unless particular care is taken (i.e. both possible orientations of many arcs result in equivalent probability distributions, so the algorithms cannot choose between them).

Properties of the Multivariate Trinomial

In the maximum entropy case we have the following approximate results [11]:

$$\mathsf{P}(\overrightarrow{a_{ij}}) = \mathsf{P}(\overleftarrow{a_{ij}}) \simeq \frac{1}{4} + \frac{1}{4(n-1)} \quad \text{and} \quad \mathsf{P}(\mathring{a}_{ij}) \simeq \frac{1}{2} - \frac{1}{2(n-1)}.$$

where n is the number of nodes of the graph. Furthermore, we have that

$$\mathsf{VAR}(A_{ij}) \simeq \frac{1}{2} + \frac{1}{2(n-1)} \to \frac{1}{2} \text{ as } n \to \infty$$

and

$$|\mathsf{COV}(A_{ij}, A_{kl})| \lessapprox 4\left[\frac{3}{4} - \frac{1}{4(n-1)}\right]^2 \left[\frac{1}{4} + \frac{1}{4(n-1)}\right]^2 \to \frac{9}{64} \text{ as } n \to \infty.$$

Measures of Structure Variability

Since variances are bounded in $[0, 1]$ we can define again

$$\overline{\mathsf{VAR}}_T(\Sigma) = \frac{1}{k}\mathsf{VAR}_T(\Sigma) \quad \text{and} \quad \overline{\mathsf{VAR}}_G(\Sigma) = \mathsf{VAR}_G(\Sigma).$$

We can also compute $\overline{\mathsf{VAR}}_N(\Sigma)$ using a Monte Carlo estimate for $\mathsf{COV}(A_{ij}, A_{kl})$ based on Ide and Cozman's algorithm [6]. The same holds for hypothesis tests.

Determining Statistically Significant Functional Relationships

The Problem

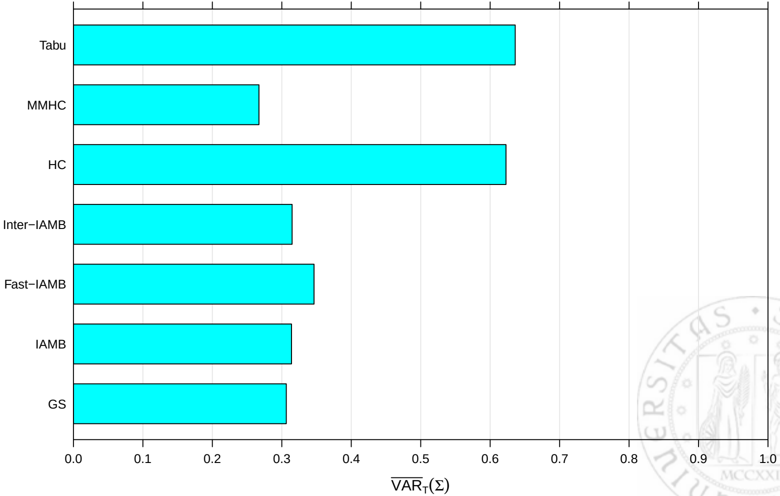
- transcriptions of regulatory (gene) networks controlling both myogenic and adipogenic differentiation are still under active investigation.
- myogenic and adipogenic differentiation pathways are typically considered non-overlapping, but Taylor-Jones et al. [20] has shown that myogenic progenitors from aged mice co-express some aspects of both myogenic and adipogenic gene programs.
- their balance is apparently regulated by Wnt signalling according to Vertino et al. [22], but there have been few efforts to understand the interactions between these two networks.

The Experimental Setting

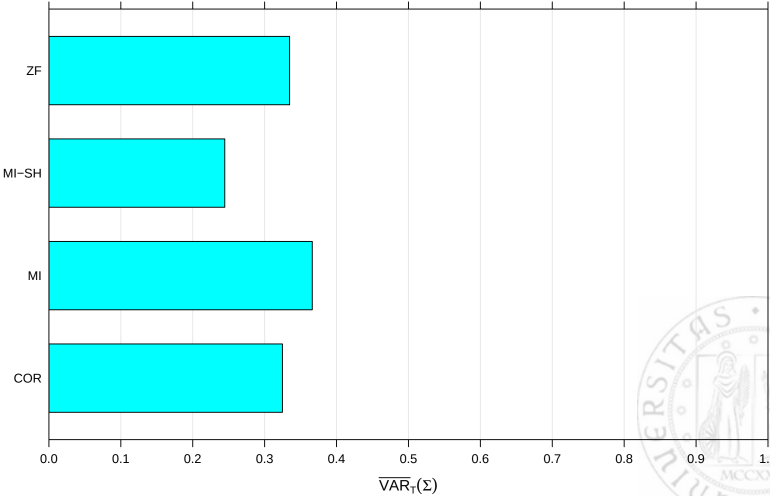
The clonal gene expression data was generated from RNA isolated from 34 clones of myogenic progenitors obtained from 24-months old mice, cultured to confluence and allowed to differentiate for 24 hours. RT–PCR was used to quantify the expression of 12 genes:

- myogenic regulatory factors: Myo-D1, Myogenin and Myf-5.
- adipogenesis-related genes: FoxC2, DDIT3, C/EPB and PPARγ.
- Wnt-related genes: Wnt5a and Lrp5.
- control genes: GAPDH, 18S and B2M.

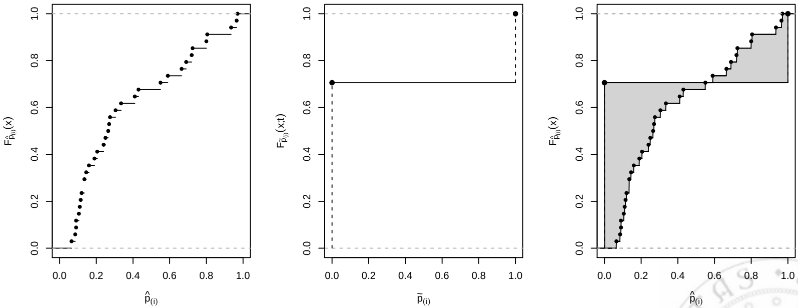
Choosing the Right Structure Learning Algorithm



Choosing the Right Tuning Parameters



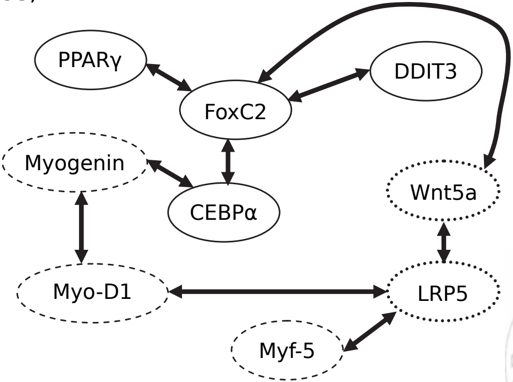
Determining Significant Functional Relationships



Significant functional relationships can be selected by filtering out the noise in the data or by finding the closest minimum-entropy configuration.

Statistically Significant FRs

control genes:
GAPDH, 18S, B2M



$n = 34$

Conclusions

Conclusions

- In literature inference on the structure of graphical models is usually overlooked in favour of the inference on the parameters of the global and local distributions.
- Rigorous inference on network structures is possible with the appropriate multivariate distributions: multivariate Bernoulli and multivariate Trinomial.
- In this setting we can define descriptive statistics and hypothesis tests which are easy to interpret and apply to any set of edges/arcs.

Thank you.

References

References I

R. B. Ash.
Probability and Measure Theory.
Academic Press, 2nd edition, 2000.

S. S. Chavan, M. A. Bauer, M. Scutari, and R. Nagarajan.
NATbox: a Network Analysis Toolbox in R.
BMC Bioinformatics, 10(Suppl 11):S14, 2009.
Supplement contains the Proceedings of the 6th Annual MCBIOS Conference (Transformational Bioinformatics: Delivering Value from Genomes).

D. I. Edwards.
Introduction to Graphical Modelling.
Springer, 2nd edition, 2000.

B. Efron and R. Tibshirani.
An Introduction to the Bootstrap.
Chapman & Hall, 1993.

N. Friedman, M. Goldszmidt, and A. Wyner.
Data Analysis with Bayesian Networks: A Bootstrap Approach.
In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence*, pages 206–215. Morgan Kaufmann, 1999.

References II

J. S. Ide and F. G. Cozman.
Random Generation of Bayesian Networks.
In *Proceedings of the 16th Brazilian Symposium on Artificial Intelligence*, pages 366–375. Springer-Verlag, 2002.

S. Imoto, S. Y. Kim, H. Shimodaira, S. Aburatani, K. Tashiro, S. Kuhara, and S. Miyano.
Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression.
Genome Informatics, 13:369–370, 2002.

D. Jungnickel.
Graphs, Networks and Algorithms.
Springer, 3rd edition, 2008.

K. Korb and A. Nicholson.
Bayesian Artificial Intelligence.
Chapman & Hall, 2004.

F. Krummenauer.
Limit Theorems for Multivariate Discrete Distributions.
Metrika, 47(1):47–69, 1998.

References III

G. Melançon, I. Dutour, and M. Bousquet-Mélou.
Random Generation of Dags for Graph Drawing.
Technical Report INS-R0005, Centre for Mathematics and Computer Sciences, Amsterdam, 2000.

R. Nagarajan, S. Datta, and M. Scutari.
Graphical Models in R.
Use R! series. Springer, 2011.
In preparation.

R. Nagarajan, S. Datta, M. Scutari, M. L. Beggs, G. T. Nolen, and C. A. Peterson.
Functional Relationships Between Genes Associated with Differentiation Potential of Aged Myogenic Progenitors.
Frontiers in Physiology, 1(21):1–8, 2010.

M. Scutari.
Structure Variability in Bayesian Networks.
Working Paper 13-2009, Department of Statistical Sciences, University of Padova, 2009.
Deposited on arXiv in the Statistics - Methodology archive, available from http://arxiv.org/abs/0909.1685.

References IV

M. Scutari.
Learning Bayesian Networks with the bnlearn R Package.
Journal of Statistical Software, 35(3):1–22, 2010.

M. Scutari.
Measures of Variability for Bayesian Network Graphical Structures.
Journal of Multivariate Analysis, 2010.
Submitted for publication.

M. Scutari.
bnlearn: Bayesian network structure learning, 2011.
R package version 2.4, http://www.bnlearn.com/.

M. Scutari and A. Brogini.
Constraint-based Bayesian Network Learning with Permutation Tests.
Communications in Statistics – Theory and Methods, 2011.
Special Issue containing the Proceedings of the Conference "Statistics for Complex Problems: the Multivariate Permutation Approach and Related Topics", Padova, June 14 – 15. In print.

References V

M. Scutari and K. Strimmer.
Introduction to Graphical Modelling.
In D. J. Balding, M. Stumpf, and M. Girolami, editors, *Handbook of Statistical Systems Biology*. Wiley, 2011.
In print.

J. M. Taylor-Jones, R. E. McGehee, T. A. Rando, B. Lecka-Czernik, D. A. Lipschitz, and C. A. Peterson.
Activation of an Adipogenic Program in Adult Myoblasts with Age.
Mechanisms of Ageing and Development, 123(6):649–661, 2002.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis.
The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.
Machine Learning, 65(1):31–78, 2006.

A. M. Vertino, J. M. Taylor-Jones, K. A. Longo, E. D. Bearden, T. F. Lane, R. E. McGehee, O. A. MacDougald, and C. A. Peterson.
Wnt10b Deficiency Promotes Coexpression of Myogenic and Adipogenic Programs in Myoblasts.
Molecular Biology of the Cell, 16(4):2039–2048, 2005.