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 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

Graphical Models

Marco Scutari University of Padova
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
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 $\{V \times V\}$) as a multivariate Bernoulli random variable $W \sim Ber_k(p)$. $W$ is uniquely identified by the parameter set

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

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 $p$ 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

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

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

\[ 0 \leq \lambda_i \leq \frac{k}{4} \quad \text{and} \quad 0 \leq \sum_{i=1}^{k} \lambda_i \leq \frac{k}{4}. \]

Furthermore, if \( W_1 \) and \( W_2 \) are two multivariate Bernoulli random variables, then they are independent if and only if

\[ W_1 \perp \! \! \! \perp W_2 \iff \text{COV}(W_1, W_2) = 0. \]
Measures of Structure Variability
Consider the graphical models $U_1, \ldots, U_m$ 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:
  
  $$p_i = \begin{cases} 
  1 & \text{if } e_i \in E \\
  0 & \text{otherwise} 
  \end{cases} \quad \text{and} \quad \Sigma = O;$$

- **intermediate entropy**: several models are observed with different frequencies $m_b$, $\sum m_b = m$, so
  
  $$\hat{p}_i = \frac{1}{m} \sum_{b: e_i \in E_b} m_b \quad \text{and} \quad \hat{p}_{ij} = \frac{1}{m} \sum_{b: e_i \in E_b, e_j \in E_b} m_b;$$

- **maximum entropy**: all possible models appear with the same frequency, which results in
  
  $$p_i = \frac{1}{2} \quad \text{and} \quad \Sigma = \frac{1}{4} I_k.$$
Entropic Variability

Entropy of the Bootstrapped Network Structures

\[ \Sigma = \frac{1}{4} I_2 \]

Minimum entropy \( \Sigma = O \)

Maximum entropy

\[ (0, \frac{1}{2}) \]

\[ (0, 0) \]

\[ \frac{1}{2}, 0 \]
Univariate Measures of Variability

- **The generalised variance**

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

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

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

- **The squared Frobenius matrix norm**

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

\[
\bar{\text{VAR}}_T(\Sigma) = \frac{4}{k} \text{VAR}_T(\Sigma), \quad \bar{\text{VAR}}_G(\Sigma) = 4^k \text{VAR}_G(\Sigma), \quad \bar{\text{VAR}}_N(\Sigma) = \frac{k^3 - 16 \text{VAR}_N(\Sigma)}{k(2k - 1)}.
\]

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.

\[
4m \text{tr}(\hat{\Sigma}) \sim \chi^2_{mk}
\]

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

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

\[
|||\hat{\Sigma} - \frac{1}{4}|||_F^2 \sim \frac{1}{8m} \chi^2_{\frac{3}{2}k(k+1)}
\]
Measures of Structure Variability

Structure Variability (Total Variance)

\[ \Sigma = \frac{1}{4} I_2 \]

minimum entropy
\[ \Sigma = 0 \]

maximum entropy

\[ (0, \frac{1}{2}) \]

\[ (0, 0) \]

\[ \frac{1}{4}, \frac{1}{4} \]

\[ \Sigma_1 \]

\[ \Sigma_2 \]

\[ \Sigma_3 \]

\[ (1/2, 0) \]
Measures of Structure Variability

Structure Variability (Squared Frobenius Matrix Norm)

(minimum entropy) \( \Sigma = O \)

(maximum entropy) \( \Sigma = \frac{1}{4} I_2 \)

\( (0, \frac{1}{2}) \)

\( (0, 0) \)

\( (\frac{1}{2}, 0) \)
Modelling Directed Acyclic Network Structures
Each arc $a_{ij}$ in a directed acyclic graph $G = (V, A)$ has three possible states,

$$ a_{ij} = \begin{cases} 
-1 & \text{if } a_{ij} = \overleftarrow{a_{ij}} = \{v_i \leftarrow v_j\} \\
0 & \text{if } a_{ij} \notin A, \text{ denoted with } \hat{a}_{ij} \\
1 & \text{if } a_{ij} = \overrightarrow{a_{ij}} = \{v_i \rightarrow v_j\} 
\end{cases} $$

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

$$ \text{VAR}(A_i) = \text{VAR}(E_i) + 4P(\overrightarrow{a_{ij}})P(\overleftarrow{a_{ij}}) \in [0, 1] $$
As before, the natural extension to model any set $W$ of arcs is to use a multivariate Trinomial random variable $W \sim Tri_k(p)$ 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 $Tri_k(p)$ 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]:

\[ P(\alpha_{ij}^+) = P(\hat{\alpha}_{ij}) \approx \frac{1}{4} + \frac{1}{4(n-1)} \quad \text{and} \quad P(\hat{\alpha}_{ij}) \approx \frac{1}{2} - \frac{1}{2(n-1)}. \]

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

\[ \text{VAR}(A_{ij}) \approx \frac{1}{2} + \frac{1}{2(n-1)} \rightarrow \frac{1}{2} \quad \text{as} \quad n \rightarrow \infty \]

and

\[ |\text{COV}(A_{ij}, A_{kl})| \lesssim 4 \left[ \frac{3}{4} - \frac{1}{4(n-1)} \right]^2 \left[ \frac{1}{4} + \frac{1}{4(n-1)} \right]^2 \rightarrow \frac{9}{64} \quad \text{as} \quad n \rightarrow \infty. \]
Measures of Structure Variability

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

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

We can also compute $\overline{\text{VAR}}_N(\Sigma)$ using a Monte Carlo estimate for $\text{COV}(A_{ij}, A_{kl})$ 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 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 Tuning Parameters

\[ \text{VAR}_T(\Sigma) \]

- ZF
- MI-SH
- MI
- COR

Marco Scutari
University of Padova
Determining Statistically Significant Functional Relationships

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

- PPARγ
- FoxC2
- CEBPα
- Myogenin
- Myo-D1
- Wnt5a
- LRP5
- DDIT3
- Myf-5

\[ n = 34 \]
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


J. S. Ide and F. G. Cozman.  
Random Generation of Bayesian Networks.  

Bootstrap Analysis of Gene Networks Based on Bayesian Networks and Nonparametric Regression.  

D. Jungnickel.  
Graphs, Networks and Algorithms.  

K. Korb and A. Nicholson.  
Bayesian Artificial Intelligence.  

F. Krummenauer.  
Limit Theorems for Multivariate Discrete Distributions.  
G. Melançon, I. Dutour, and M. Bousquet-Mélou.
Random Generation of Dags for Graph Drawing.

R. Nagarajan, S. Datta, and M. Scutari.
*Graphical Models in R.*
In preparation.

Functional Relationships Between Genes Associated with Differentiation Potential of Aged Myogenic Progenitors.

M. Scutari.
Structure Variability in Bayesian Networks.
M. Scutari.
Learning Bayesian Networks with the bnlearn R Package. 

M. Scutari.
Measures of Variability for Bayesian Network Graphical Structures. 
Submitted for publication.

M. Scutari.
*bnlearn: Bayesian network structure learning*, 2011. 

M. Scutari and A. Brogini. 
Constraint-based Bayesian Network Learning with Permutation Tests. 
References V

M. Scutari and K. Strimmer.
Introduction to Graphical Modelling.
In print.

Activation of an Adipogenic Program in Adult Myoblasts with Age.

I. Tsamardinos, L. E. Brown, and C. F. Aliferis.
The Max-Min Hill-Climbing Bayesian Network Structure Learning Algorithm.

Wnt10b Deficiency Promotes Coexpression of Myogenic and Adipogenic Programs in Myoblasts.