On the Prior and Posterior Distributions Used in Graphical Modelling

Marco Scutari

m.scutari@ucl.ac.uk
Genetics Institute
University College London

August 5, 2014

The Problem

A large part of the 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). However:

- 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.
- 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 (e.g. Hamming distance [5] or SHD [10]), and are difficult to interpret.
- When learning causal graphical models often the focus is not on the parameters but in the presence of particular patterns of edges in the graph (e.g. [8]).

Looking for a Solution

Focusing on graph structures $\mathcal{G}$ sidesteps some of these problems and acknowledges causal graphical modelling literature [9].

0. We need to know more about the properties of priors $\mathrm{P}(\mathcal{G})$ and posteriors $\mathrm{P}(\mathcal{G} \mid \mathcal{D})$ distributions over the space of graphs, preferably as a function of arc and edge sets, say $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$ with $\mathcal{E} = \{(v_i, v_j), i \neq j\} \in \mathcal{O}(|\mathbf{V}|^2)$.

And then:

1. It would be good to have a measure(s) of spread for $\mathcal{G}$, to assess the noisiness of $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$ and the informativeness of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$.
2. Using such a measure(s), it would be interesting to study the convergence speed of structure learning algorithms and the influence of their tuning parameters.
3. It would also be interesting to investigate how to use higher order moments of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ to define new priors.

Edge Sets as Multivariate Bernoulli

Each edge e_{ij} in an undirected graph $\mathcal{G} = (\mathbf{V}, E)$ has only two possible states and therefore can be modelled as a Bernoulli random variable:

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

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

$$\mathbf{p} = \{p_I : I \subseteq \{1, \ldots, k\}, i \neq \varnothing\}, \qquad k = \frac{|\mathbf{V}|(|\mathbf{V}| - 1)}{2}$$

which represents the dependence structure [6] among the marginal distributions $B_i \sim Ber(p_i)$, $i = 1, \ldots, k$ of the edges.
The parameter set $\mathbf{p}$ can be estimated using a large number m of bootstrap samples as in Friedman *et al.* [2] or Imoto *et al.* [4], or MCMC samples as in Friedman & Koller [3].

Arc Sets as Multivariate Trinomial

Each arc a_{ij} in $\mathcal{G} = (\mathbf{V}, A)$ has three possible states, and therefore it can be modelled as a Trinomial random variable A_{ij}:

$$a_{ij} \sim A_{ij} = \begin{cases} -1 & \text{if } a_{ij} = \overleftarrow{a_{ij}} = \{v_i \leftarrow v_j\} \\ 0 & \text{if } a_{ij} \notin A, \text{ denoted with } \mathring{a_{ij}} \\ 1 & \text{if } a_{ij} = \overrightarrow{a_{ij}} = \{v_i \rightarrow v_j\} \end{cases}.$$

As before, the natural extension to model any set of arcs is to use a multivariate Trinomial random variable $\mathbf{T} \sim Tri_k(\mathbf{p})$.
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).

Second Order Properties of $Ber_k(\mathbf{p})$ and $Tri_k(\mathbf{p})$

All the elements of the covariance matrix Σ of an edge set $\mathcal{E}$ are bounded,

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

and similar bounds exist for the eigenvalues $\lambda_1, \ldots, \lambda_k$,

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

These bounds define a closed convex set in $\mathbb{R}^k$,

$$\mathcal{L} = \left\{ \Delta^{k-1}(c) : c \in \left[0, \frac{k}{4}\right] \right\}$$

where $\Delta^{k-1}(c)$ is the non-standard $k-1$ simplex

$$\Delta^{k-1}(c) = \left\{ (\lambda_1, \ldots, \lambda_k) \in \mathbb{R}^k : \sum_{i=1}^{k} \lambda_i = c, \lambda_i \geqslant 0 \right\}.$$

Similar results hold for arc sets, with $\sigma_{ii} \in [0,1]$ and $\lambda_i \in [0,k]$.

Minimum and Maximum Entropy

These results provide the foundation for characterising three cases corresponding to different configurations of the probability mass in $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$:

- minimum entropy: the probability mass is concentrated on a single graph structure. This is the best possible configuration for $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$, because only one edge set E (or one arc set A) has a non-zero posterior probability.
- intermediate entropy: several graph structures have non-zero probabilities. This is the case for informative priors $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and for the posteriors $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$ resulting from real-world data sets.
- maximum entropy: all graph structures have the same probability. This is the worst possible configuration for $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$, because it corresponds to a non-informative prior. In other words, the data $\mathcal{D}$ do not provide any information useful in identifying a high-posterior graph $\mathcal{G}$.

Properties of the Multivariate Bernoulli

In the minimum entropy case, only one configuration of edges E has non-zero probability, which means that

$$p_{ij} = \begin{cases} 1 & \text{if } e_{ij} \in E \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Sigma = \mathbf{O}$$

where $\mathbf{O}$ is the zero matrix.
The uniform distribution over $\mathbf{G}$ arising from the maximum entropy case has been studied extensively in random graph theory [1]; its two most relevant properties are that all edges e_{ij} are independent and have $p_{ij} = \frac{1}{2}$. As a result, $\Sigma = \frac{1}{4} I_k$; all edges display their maximum possible variability, which along with the fact that they are independent makes this distribution non-informative for $\mathcal{E}$ as well as $\mathcal{G}(\mathcal{E})$.

Properties of the Multivariate Trinomial

In the maximum entropy case we have that [7]

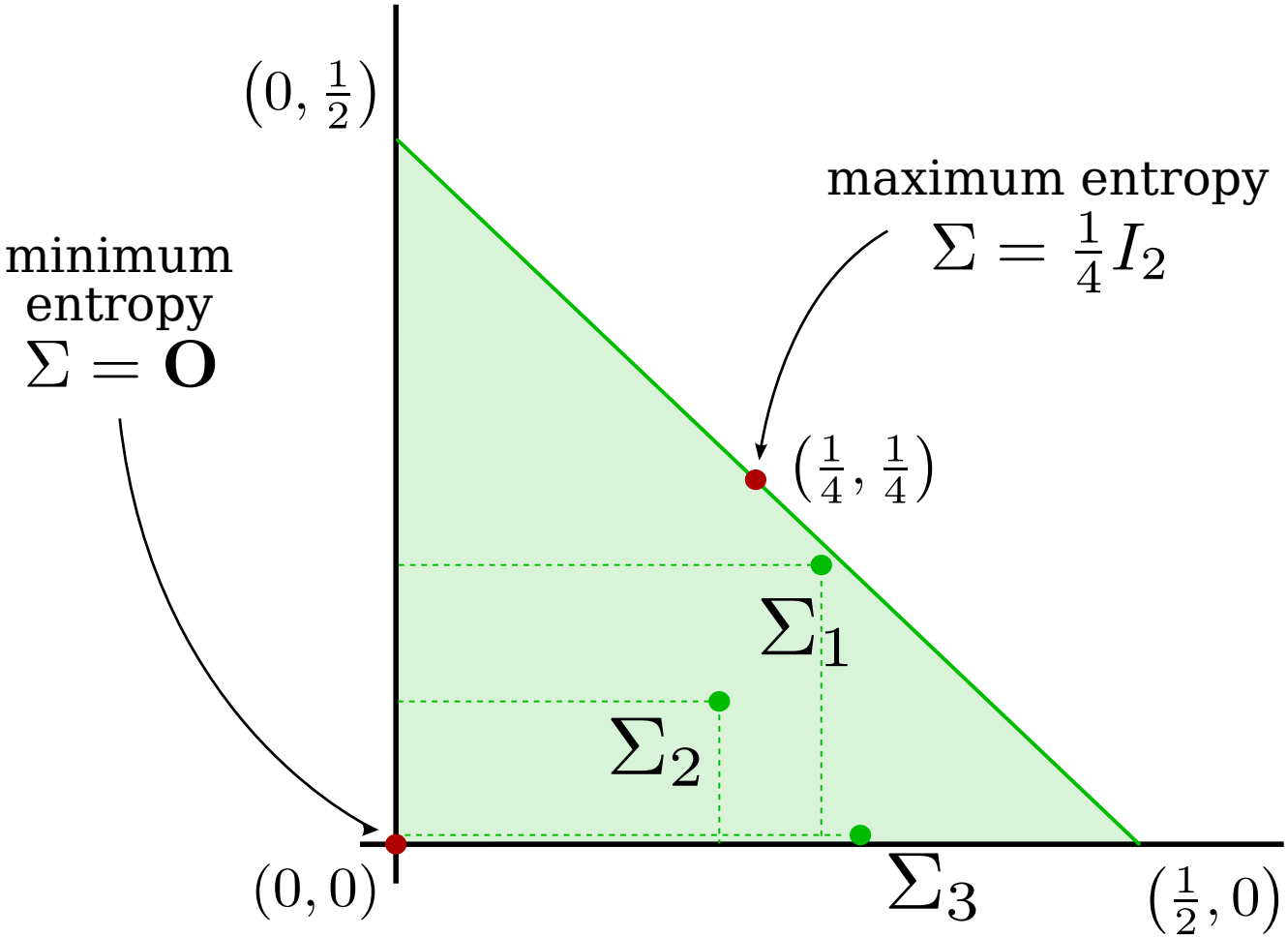
$$\mathrm{P}(\overrightarrow{a_{ij}}) = \mathrm{P}(\overleftarrow{a_{ij}}) \simeq \frac{1}{4} + \frac{1}{4(n-1)} \to \frac{1}{4}$$

$$\mathrm{P}(\mathring{a}_{ij}) \simeq \frac{1}{2} - \frac{1}{2(n-1)} \to \frac{1}{2}$$

as $n \to \infty$, where n is the number of nodes of the graph. As a result, we have that

$$\begin{aligned}
\mathrm{E}(A_{ij}) &= \mathrm{P}(\overrightarrow{a_{ij}}) - \mathrm{P}(\overleftarrow{a_{ij}}) = 0, \\
\mathsf{VAR}(A_{ij}) &= 2\,\mathrm{P}(\overrightarrow{a_{ij}}) \simeq \frac{1}{2} + \frac{1}{2(n-1)} \to \frac{1}{2}, \\
|\mathsf{COV}(A_{ij}, A_{kl})| &= 2\,[\mathrm{P}(\overrightarrow{a_{ij}}, \overrightarrow{a_{kl}}) - \mathrm{P}(\overrightarrow{a_{ij}}, \overleftarrow{a_{kl}})] \\
&\lessapprox 4 \left[\frac{3}{4} - \frac{1}{4(n-1)}\right]^2 \left[\frac{1}{4} + \frac{1}{4(n-1)}\right]^2 \to \frac{9}{64}.
\end{aligned}$$

A Geometric Representation of Entropy in $\mathcal{L}$



The space of the eigenvalues $\mathcal{L}$ for two edges in an undirected graph.

Univariate Measures of Variability

- The generalised variance, $\text{VAR}_G(\Sigma) = \det(\Sigma) = \prod_{i=1}^{k} \lambda_i \in \left[0, \frac{1}{4^k}\right]$.
- The total variance (or total variability)

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

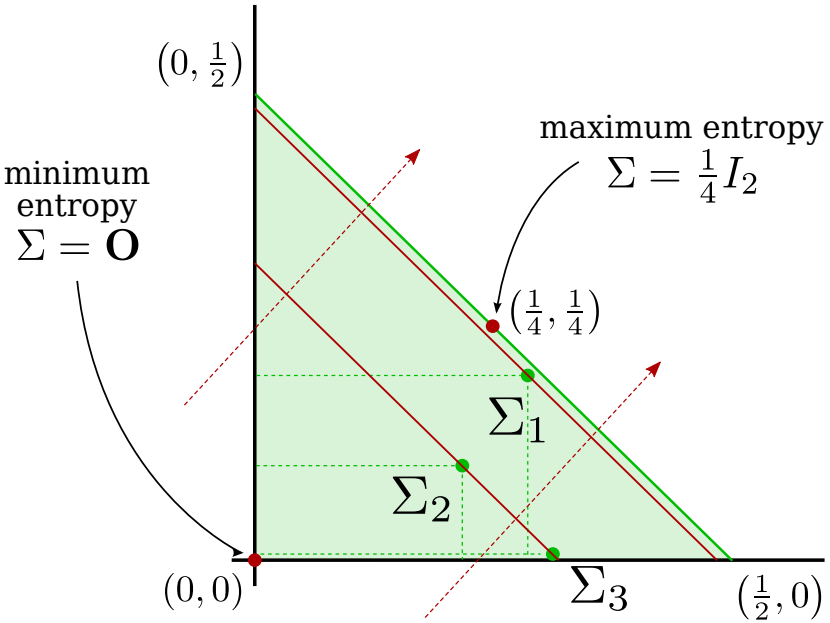
- The squared Frobenius matrix norm

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

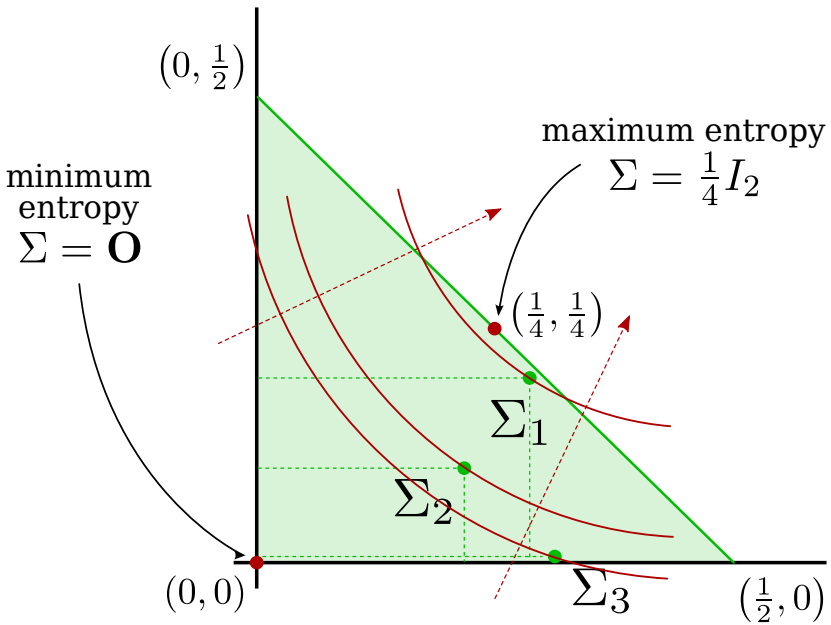
All of these measures can be rescaled to vary in the $[0, 1]$ interval and to associate high values to networks whose structure displays a high entropy.

The equivalent measures of variability for directed acyclic graphs can be derived in the same way, and they can be similarly normalised.

Structure Variability: Level Curves



Level curves in $\mathcal{L}$ for $\mathrm{VAR}_T(\Sigma)$.



Level curves in $\mathcal{L}$ for $\mathrm{VAR}_F(\Sigma)$.

Conclusions and Open Problems

- First and second order properties of $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ and $\mathrm{P}(\mathcal{G}(\mathcal{E}) \mid \mathcal{D})$ can be often derived in closed form, and have a geometric interpretation.
- First and second order properties of the uniform $\mathrm{P}(\mathcal{G}(\mathcal{E}))$ on directed acyclic graphs can be a basis for simulations and the definition of new priors; could they translate to the uniform prior over decomposable undirected graphs?
- Is there a way of identifying paths using covariance matrix decompositions?
- Shrinking the covariance matrix affects $\mathrm{P}(e_{ij})$ and $\mathrm{P}(a_{ij})$ as well, and it is possible to use it for regularisation purposes. Applications to Bayesian model averaging and significant edges/arcs identification?

References I

B. Bollobás.

Random Graphs.
Cambridge University Press, 2nd edition, 2001.

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.

N. Friedman and D. Koller.

Being Bayesian about Bayesian Network Structure: A Bayesian Approach to Structure Discovery in Bayesian Networks.
Machine Learning, 50(1–2):95–126, 2003.

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.

F. Krummenauer.

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

References II

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.

K. Sachs, O. Perez, D. Pe'er, D. A. Lauffenburger, and G. P. Nolan.

Causal Protein-Signaling Networks Derived from Multiparameter Single-Cell Data.

Science, 308(5721):523–529, 2005.

M. Scutari.

On the Prior and Posterior Distributions Used in Graphical Modelling (with discussion).

Bayesian Analysis, 8(3):505–532, 2013.

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.