Structure Variability in Graphical Models

Marco Scutari

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

November 5, 2009

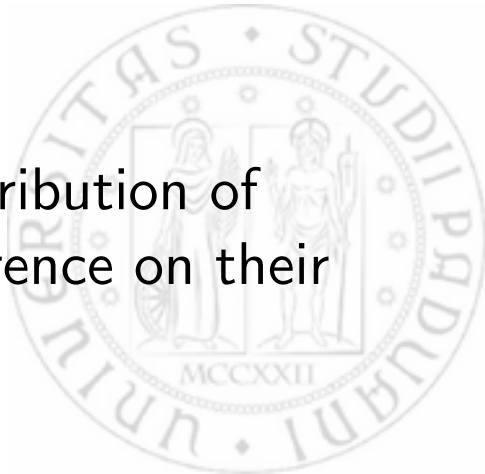
Introduction

Graphical models

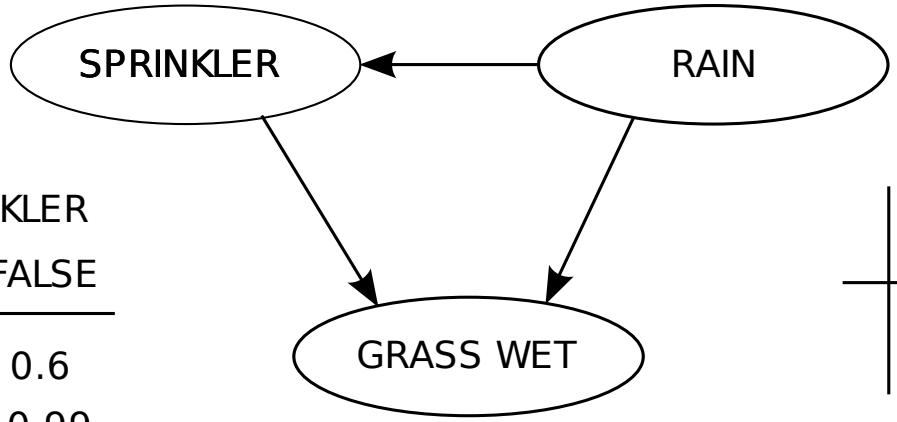
Graphical models are defined by the combination of:

- a network structure, either an undirected (Markov networks [2], gene association networks, correlation networks, etc.) or a directed graph (Bayesian networks [7]). Each node corresponds to a random variable.
- a global probability distribution which can be factorized into a set of local probability distributions (one for each node) according to the topology of the graph.

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



A simple Bayesian network: Watson's lawn



RAIN	SPRINKLER	
	TRUE	FALSE
FALSE	0.4	0.6
TRUE	0.01	0.99

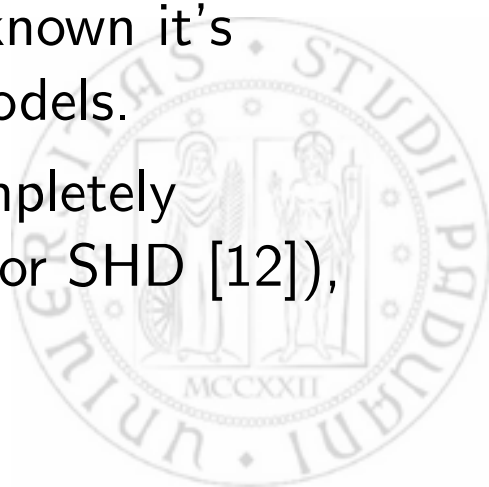
RAIN	
TRUE	FALSE
0.2	0.8

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 linear correlations).

- this makes comparing models learned with different algorithms difficult, because they maximize different scores, use different estimators for the parameters, work under different sets of hypotheses, etc. [11].
- unless the true global probability distribution is known it's difficult to assess the quality of the estimated models.
- the few measures of structural difference are completely descriptive in nature (i.e. Hamming distance [6] or SHD [12]), and have no easy interpretation.



Modeling undirected network structures

Edges and univariate Bernoulli random variables

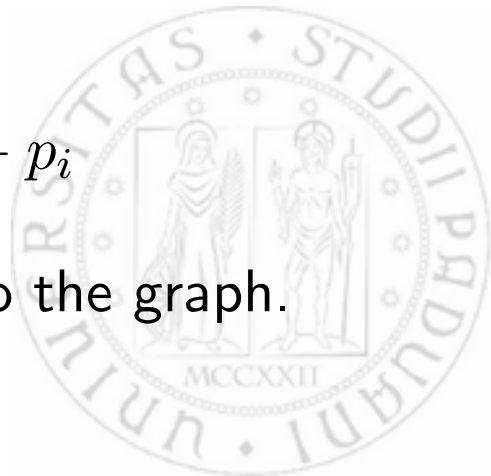
Each edge e_i in an undirected graph $\mathcal{U} = (\mathbf{V}, E)$ has only two possible states,

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

Therefore it can be modeled 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 belongs to the graph. Let's 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 $\mathbf{W} \sim Ber_k(\mathbf{p})$. It is uniquely identified by the parameter set

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

which represents the dependence structure [8] among the marginal distributions $W_i \sim Ber(p_i)$, $i = 1, \ldots, k$ of the edges.

Estimation of the parameters of $\mathbf{W}$

The parameter set $\mathbf{p}$ of $\mathbf{W}$ can be estimated via bootstrap [3] as in Friedman *et al.* [4] or Imoto *et al.* [5]:

1. For $b = 1, 2, \ldots, m$
   - 1.1 re-sample a new data set $\mathbf{D}^*_{\mathbf{b}}$ from the original data $\mathbf{D}$ using either parametric or nonparametric bootstrap.
   - 1.2 learn a graphical model $\mathcal{U}_b = (\mathbf{V}, E_b)$ from $\mathbf{D}^*_{\mathbf{b}}$.
2. Estimate the probability of each subset w of W as

$$\hat{p}_w = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{w \subseteq E_b\}}(\mathcal{U}_b).$$

Properties of the multivariate Bernoulli distribution

Moments

The first two moments of a multivariate Bernoulli variable $\mathbf{W} = [W_1, W_2, \ldots, W_k]$ are

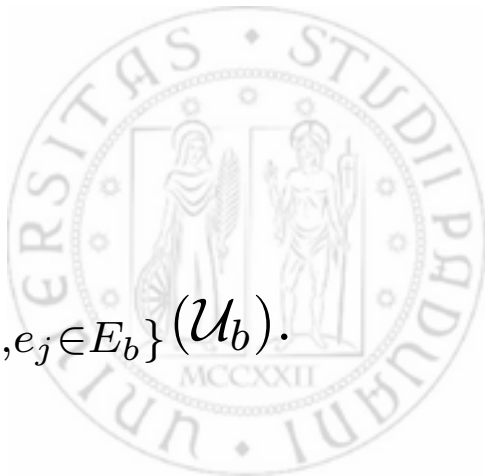
$$P = [\mathsf{E}(W_1), \ldots, \mathsf{E}(W_k)]^T \qquad \Sigma = [\sigma_{ij}] = [\mathsf{COV}(W_i, W_j)]$$

where

$$\begin{aligned} \mathsf{E}(W_i) &= p_i \\ \mathsf{COV}(W_i, W_j) &= \mathsf{E}(W_i W_j) - \mathsf{E}(W_i)\mathsf{E}(W_j) = p_{ij} - p_i p_j \\ \mathsf{VAR}(W_i) &= \mathsf{COV}(W_i, W_i) = p_i - p_i^2 \end{aligned}$$

and can be estimated using

$$\hat{p}_i = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{e_i \in E_b\}}(\mathcal{U}_b) \quad \text{and} \quad \hat{p}_{ij} = \frac{1}{m} \sum_{b=1}^{m} \mathbb{I}_{\{e_i \in E_b, e_j \in E_b\}}(\mathcal{U}_b).$$



Uncorrelation and independence

Theorem

Let B_i and B_j be two Bernoulli random variables. Then B_i and B_j are independent if and only if their covariance is zero:

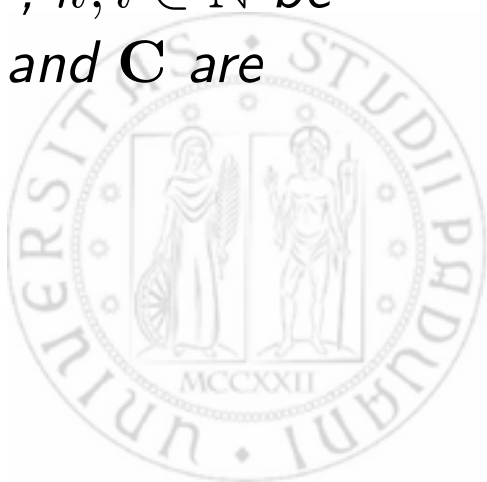
$$B_i \perp\!\!\!\perp B_j \iff \mathsf{COV}(B_i, B_j) = 0$$

Theorem

Let $\mathbf{B} = [B_1, B_2, \ldots, B_k]^T$ and $\mathbf{C} = [C_1, C_2, \ldots, C_l]^T$, $k, l \in \mathbb{N}$ be two multivariate Bernoulli random variables. Then $\mathbf{B}$ and $\mathbf{C}$ are independent if and only if

$$\mathbf{B} \perp\!\!\!\perp \mathbf{C} \iff \mathsf{COV}(\mathbf{B}, \mathbf{C}) = \mathbf{O}$$

where $\mathbf{O}$ is the zero matrix.



Constraints on the covariance matrix Σ

The marginal variances of the edges are bounded, because

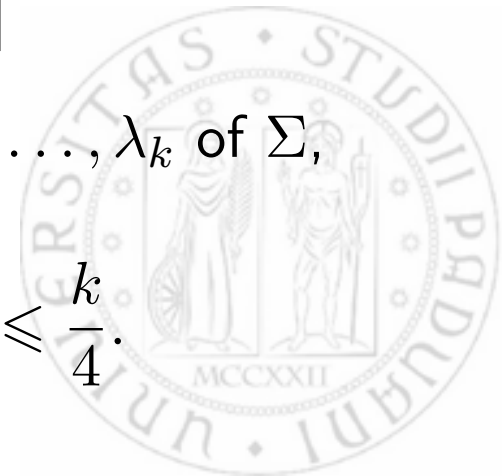
$$p_i \in [0,1] \Longrightarrow \sigma_{ii} = p_i - p_i^2 \in \left[0, \frac{1}{4}\right].$$

The maximum is attained for $p_i = \frac{1}{2}$, and the minimum for both $p_i = 0$ and $p_i = 1$. For the Cauchy-Schwartz theorem [1] then covariances are bounded too:

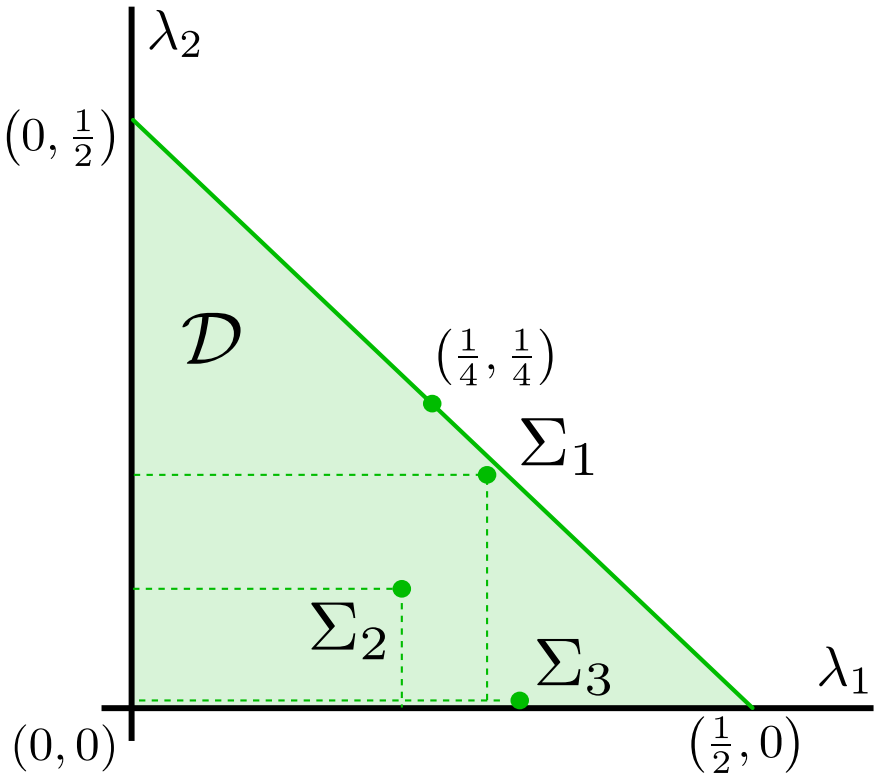
$$0 \leqslant \sigma_{ij}^2 \leqslant \sigma_{ii}\sigma_{jj} \leqslant \frac{1}{16} \Longrightarrow |\sigma_{ij}| \in \left[0, \frac{1}{4}\right].$$

These result in similar bounds on the eigenvalues $\lambda_1, \ldots, \lambda_k$ of Σ,

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



Constraints on Σ: a graphical representation



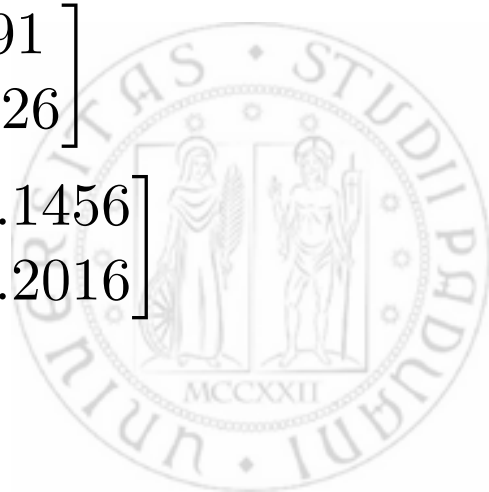
$$\Sigma_1 = \frac{1}{25}\begin{bmatrix} 6 & 1 \\ 1 & 6 \end{bmatrix} = \begin{bmatrix} 0.24 & 0.04 \\ 0.04 & 0.24 \end{bmatrix}$$

$$\Sigma_2 = \frac{1}{625}\begin{bmatrix} 66 & -21 \\ -21 & 126 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1056 & -0.0336 \\ -0.0336 & 0.2016 \end{bmatrix}$$

$$\Sigma_3 = \frac{1}{625}\begin{bmatrix} 66 & 91 \\ 91 & 126 \end{bmatrix}$$

$$= \begin{bmatrix} 0.1056 & 0.1456 \\ 0.1456 & 0.2016 \end{bmatrix}$$



Measures of Structure Variability

Entropy of the bootstrapped models

Let's consider the graphical models $\mathcal{U}_1, \ldots, \mathcal{U}_m$ learned from the bootstrap samples.

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

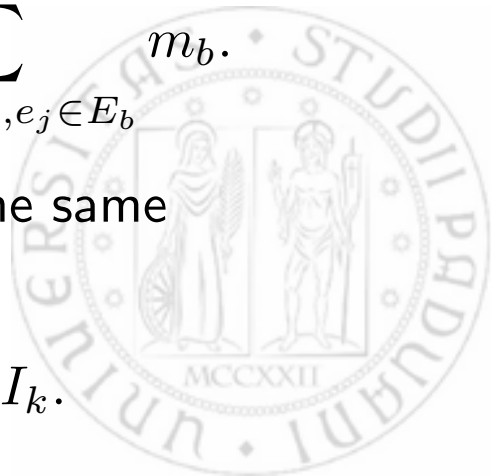
$$p_i = \begin{cases} 1 & \text{if } e_i \in E \\ 0 & \text{otherwise} \end{cases} \qquad \text{and} \qquad \Sigma = \mathbf{O}.$$

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

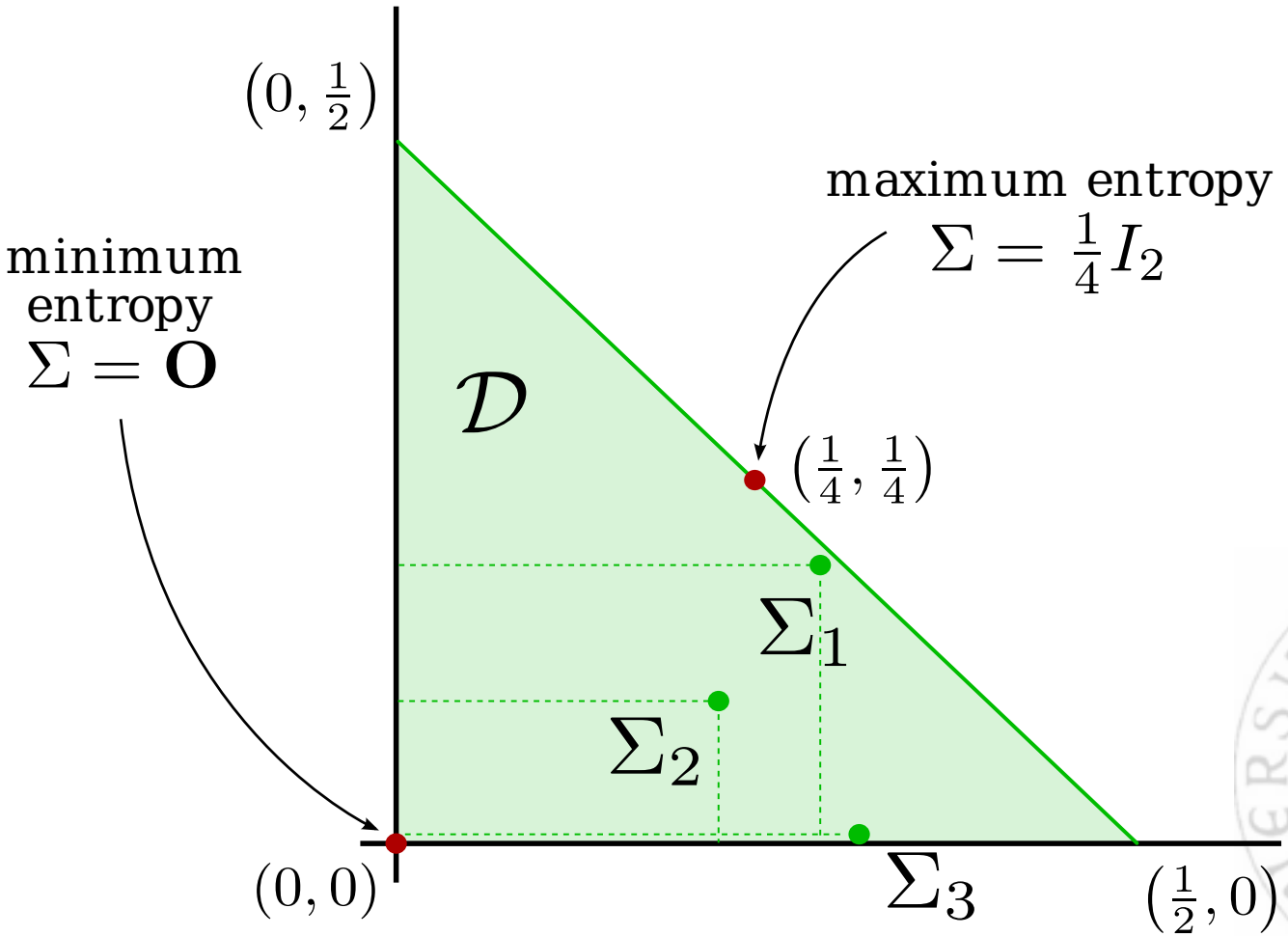
$$\hat{p}_i = \frac{1}{m} \sum_{b \,:\, e_i \in E_b} m_b \qquad \text{and} \qquad \hat{p}_{ij} = \frac{1}{m} \sum_{b \,:\, e_i \in E_b, e_j \in E_b} m_b.$$

- 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 models



Univariate measures of variability

- the *generalized variance*

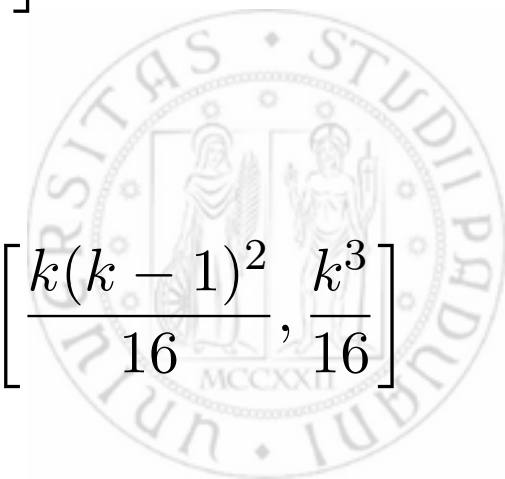
$$\text{VAR}_G(\Sigma) = \det(\Sigma) = \prod_{i=1}^{k} \lambda_i \in \left[0, \frac{1}{4^k}\right]$$

- the *total variance*

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

- 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

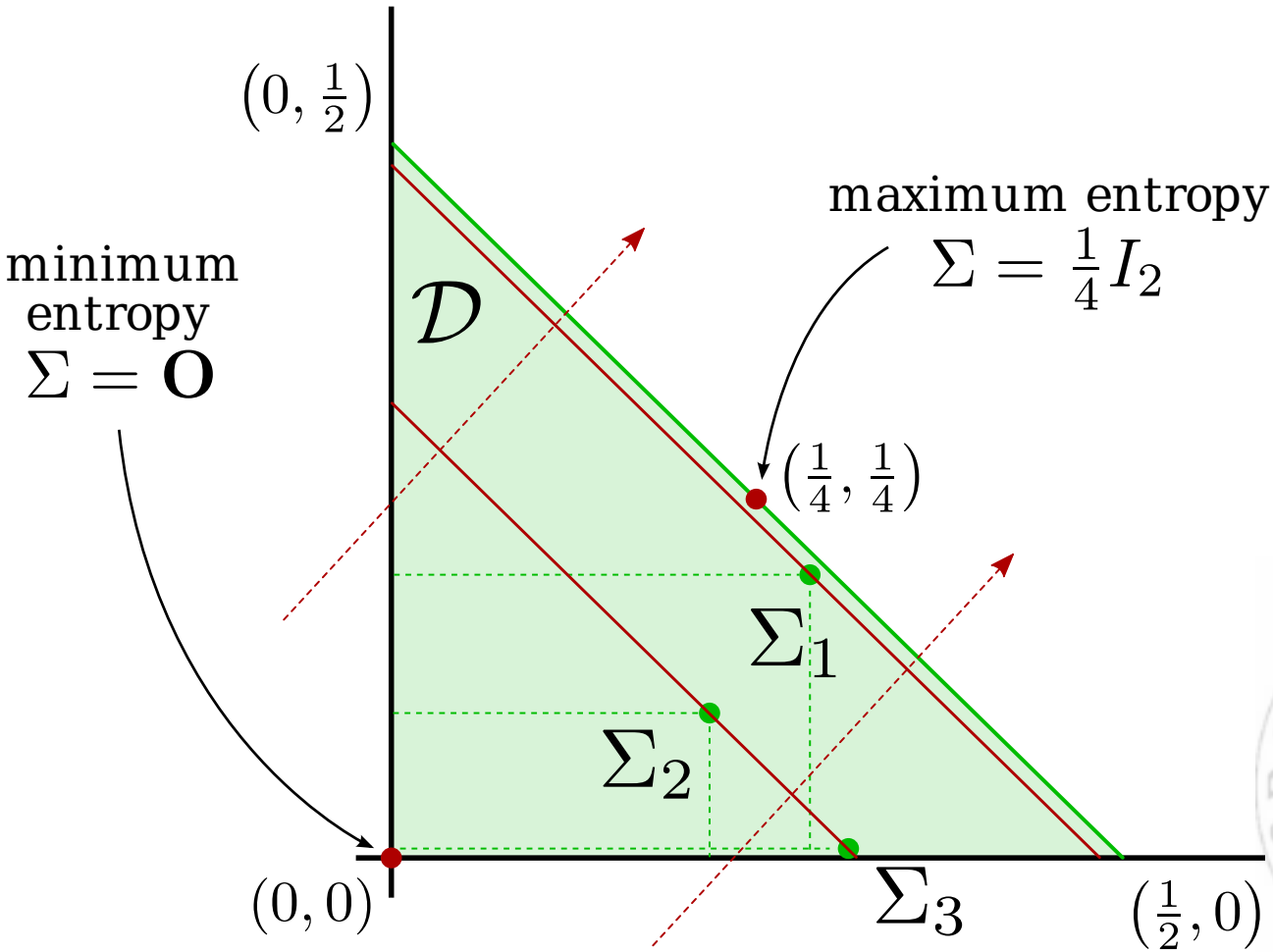
$$\overline{\mathsf{VAR}}_T(\Sigma) = \frac{\mathsf{VAR}_T(\Sigma)}{\max_\Sigma \mathsf{VAR}_T(\Sigma)} = \frac{4}{k}\mathsf{VAR}_T(\Sigma)$$

$$\overline{\mathsf{VAR}}_G(\Sigma) = \frac{\mathsf{VAR}_G(\Sigma)}{\max_\Sigma \mathsf{VAR}_G(\Sigma)} = 4^k\mathsf{VAR}_G(\Sigma)$$

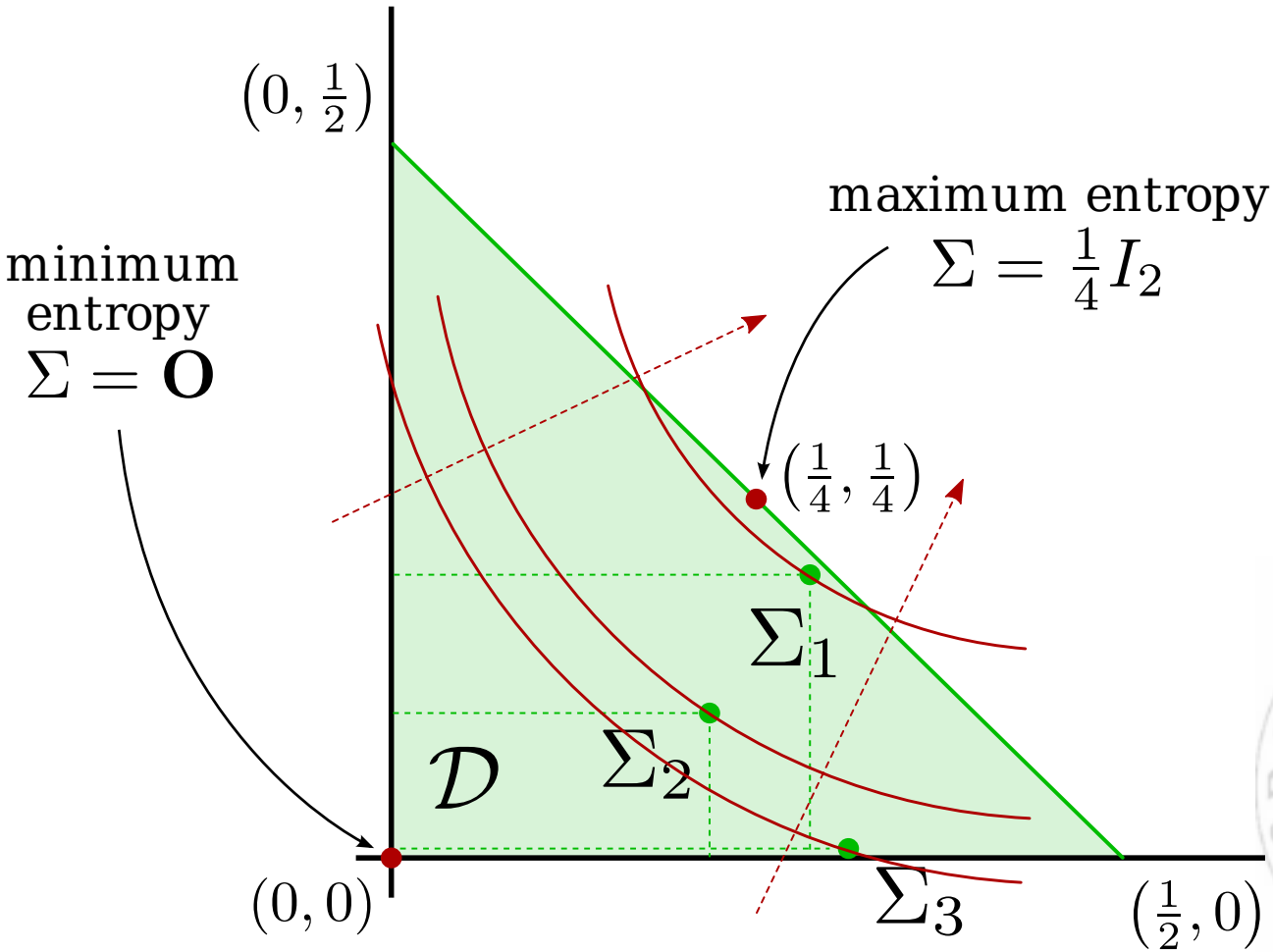
$$\begin{aligned}\overline{\mathsf{VAR}}_N(\Sigma) &= \frac{\max_\Sigma \mathsf{VAR}_N(\Sigma) - \mathsf{VAR}_N(\Sigma)}{\max_\Sigma \mathsf{VAR}_N(\Sigma) - \min_\Sigma \mathsf{VAR}_N(\Sigma)} \\ &= \frac{k^3 - 16\mathsf{VAR}_N(\Sigma)}{k(2k-1)}\end{aligned}$$

All of them vary in the $[0, 1]$ interval and associate high values to networks whose structure display a high entropy in the bootstrap samples.

Structure variability (total variance)



Structure variability (Frobenius norm)

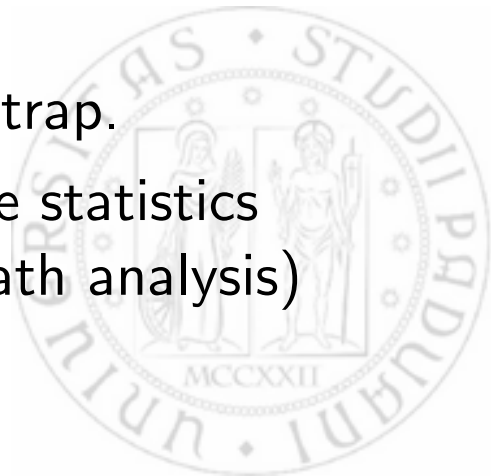


Applications

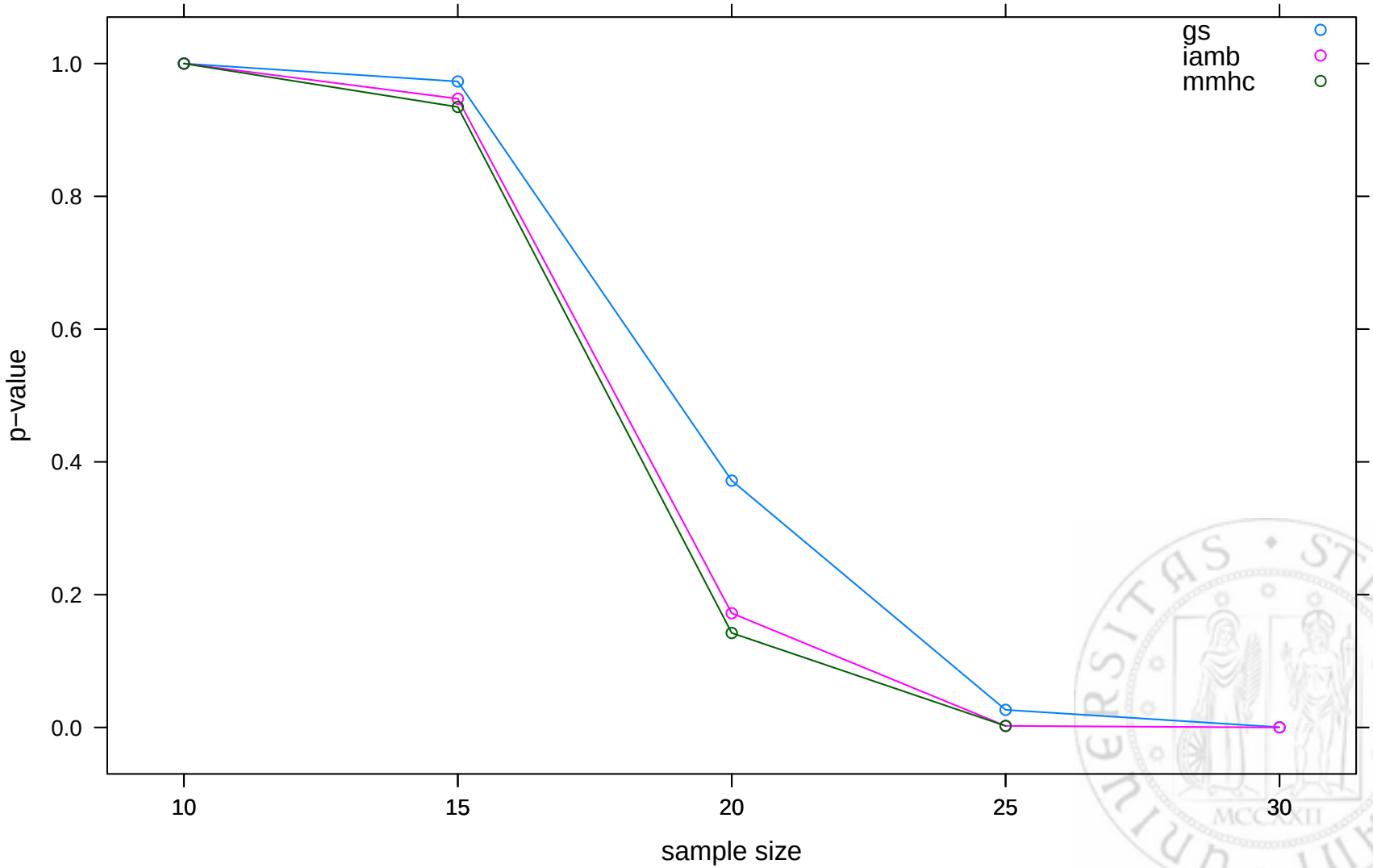
- compare the performance of different combinations of learning algorithms and network scores/independence tests on the same data.
- study the performance of an algorithm at different sample sizes by changing the size bootstrap samples. The simplest way is to test the hypothesis

$$H_0 : \Sigma = \frac{1}{4} I_k \qquad\qquad H_1 : \Sigma \neq \frac{1}{4} I_k$$

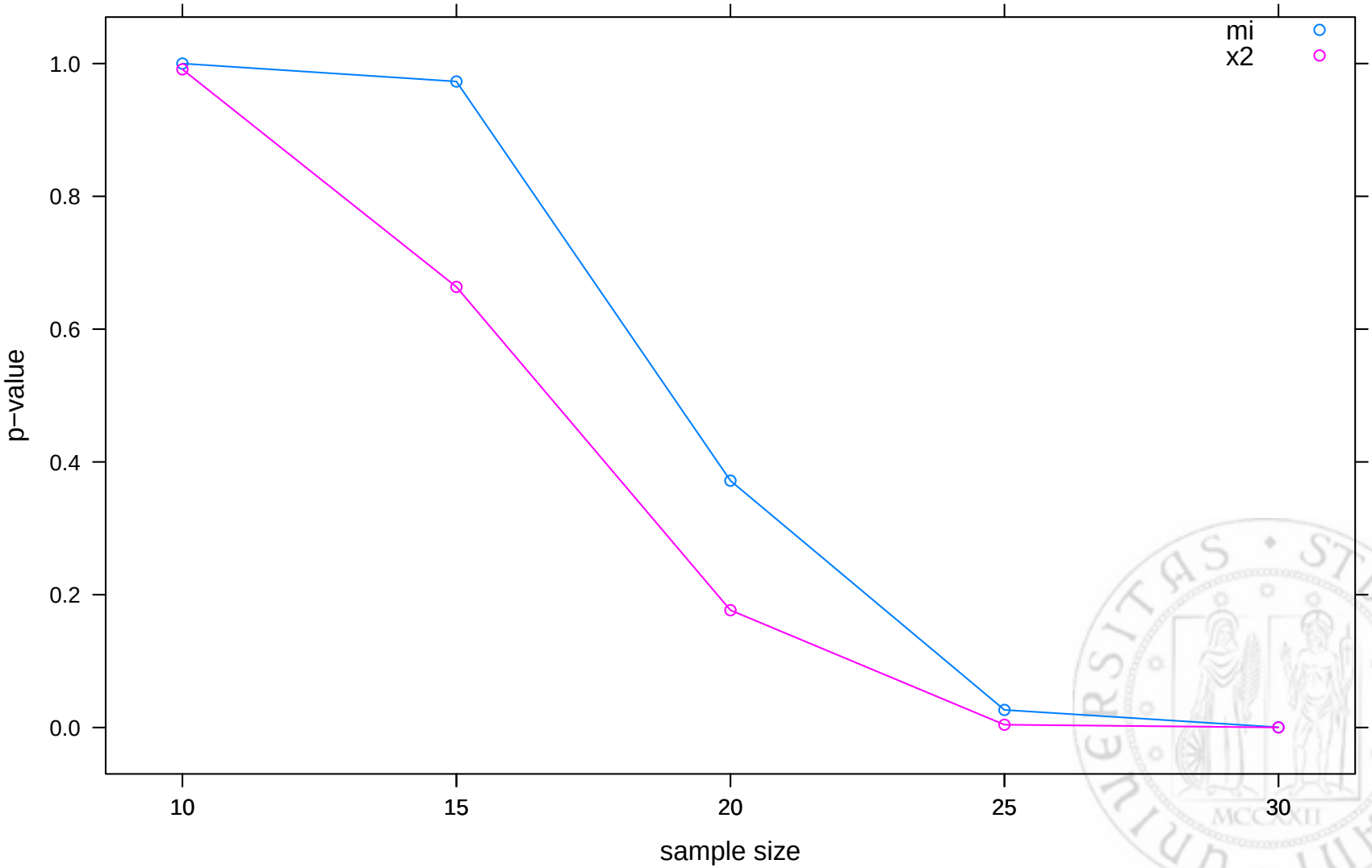
  using either parametric tests or parametric bootstrap.
- apply many techniques from classical multivariate statistics (such as principal components), graph theory (path analysis) and linear algebra (matrix decompositions).



Comparing learning algorithms' performance



Comparing statistical tests' performance



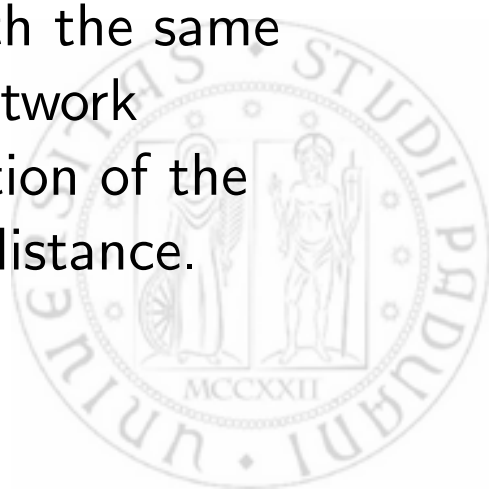
Further Applications

Distances in the space of graphs

The availability of the first two moments of the random vector $\mathbf{E}$ allows the computation of the Mahalanobis distance

$$D_{\mathcal{U}^*} = (E^* - \mathsf{E}(E))^T \Sigma^{-1} (E^* - \mathsf{E}(E))$$

of any possible graphical structure $\mathcal{U}^* = (\mathbf{W}, E^*)$ with the same vertex set. This method works even when the true network structure is not known, and gives a better representation of the geometry of the space of the graphs than Hamming distance.



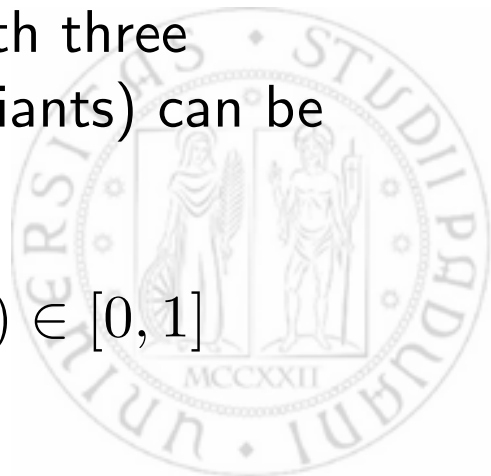
Extensions to directed graphs

Each arc $a_i = (v_j, v_k)$ in a directed graph $\mathcal{G} = (\mathbf{V}, A)$ has three possible states

$$a_i = \begin{cases} -1 & \text{if } a_i = \{v_j \leftarrow v_k\} \text{ (backward)} \\ 0 & \text{if } a_i \notin A \\ 1 & \text{if } a_i = \{v_j \rightarrow v_k\} \text{ (forward)} \end{cases}$$

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

$$\text{VAR}(A_i) = \text{VAR}(E_i) + 4\text{P}(\text{forward})\text{P}(\text{backward}) \in [0, 1]$$



Thank you.

References

References I

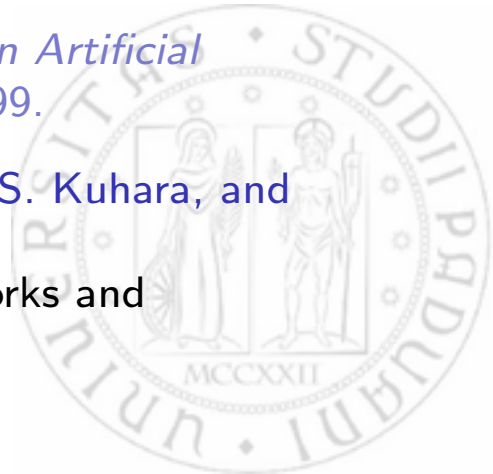

R. B. Ash.
Probability and Measure Theory.
Academic Press, 2nd edition, 2000.


D. I. Edwards.
Introduction to Graphical Modelling.
Springer, 2000.


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


Nir Friedman, Moises Goldszmidt, and Abraham Wyner.
Data Analysis with Bayesian Networks: A Bootstrap Approach.
In *Proceedings of the 15th Annual Conference on Uncertainty in Artificial Intelligence (UAI-99)*, pages 206 – 215. Morgan Kaufmann, 1999.


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.



References II


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


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


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


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.


Marco Scutari.
bnlearn: Bayesian network structure learning, 2009.
R package version 1.7, http://www.bnlearn.com/.

References III

Marco Scutari.
Learning Bayesian Networks with the bnlearn Package.
Journal of Statistical Software, 2009.
Submitted for publication.

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.