Graphical Modelling in Genetics and Systems Biology

Marco Scutari

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

October 30th, 2012

Current Practices in Bayesian Networks Modelling

Bayesian Networks Modelling Framework

Bayesian network modelling has focused on two sets of parametric assumptions, because of the availability of closed form results and computational tractability:

- discrete Bayesian networks, which assume that both the global and the local distributions are multinomial. Common association measures are mutual information (log-likelihood ratio) and Pearson's X^2;
- Gaussian Bayesian networks, which assume that the global distribution is multivariate normal and the local distributions are univariate normals linked by linear dependence relationships. Association is measured by various estimators of Pearson's correlation.

Open Problems

In applications to data in genetics and systems biology, these two sets of assumptions (and Bayesian networks in general) present some important limitations.

- Given the small sizes of available data sets ($n \ll p$), how effective is the classic Bayesian take on learning and inference?
- Are the discrete and Gaussian assumptions really sensible for these kinds of data?
- Can Bayesian networks be used to perform an effective feature selection?

Data in Genetics and Systems Biology

Overview

In genetics and systems biology, graphical models are employed to describe and identify interdependencies among genes and gene products, with the eventual aim to better understand the molecular mechanisms that link them. Data commonly made available for this task by current technologies fall into three groups:

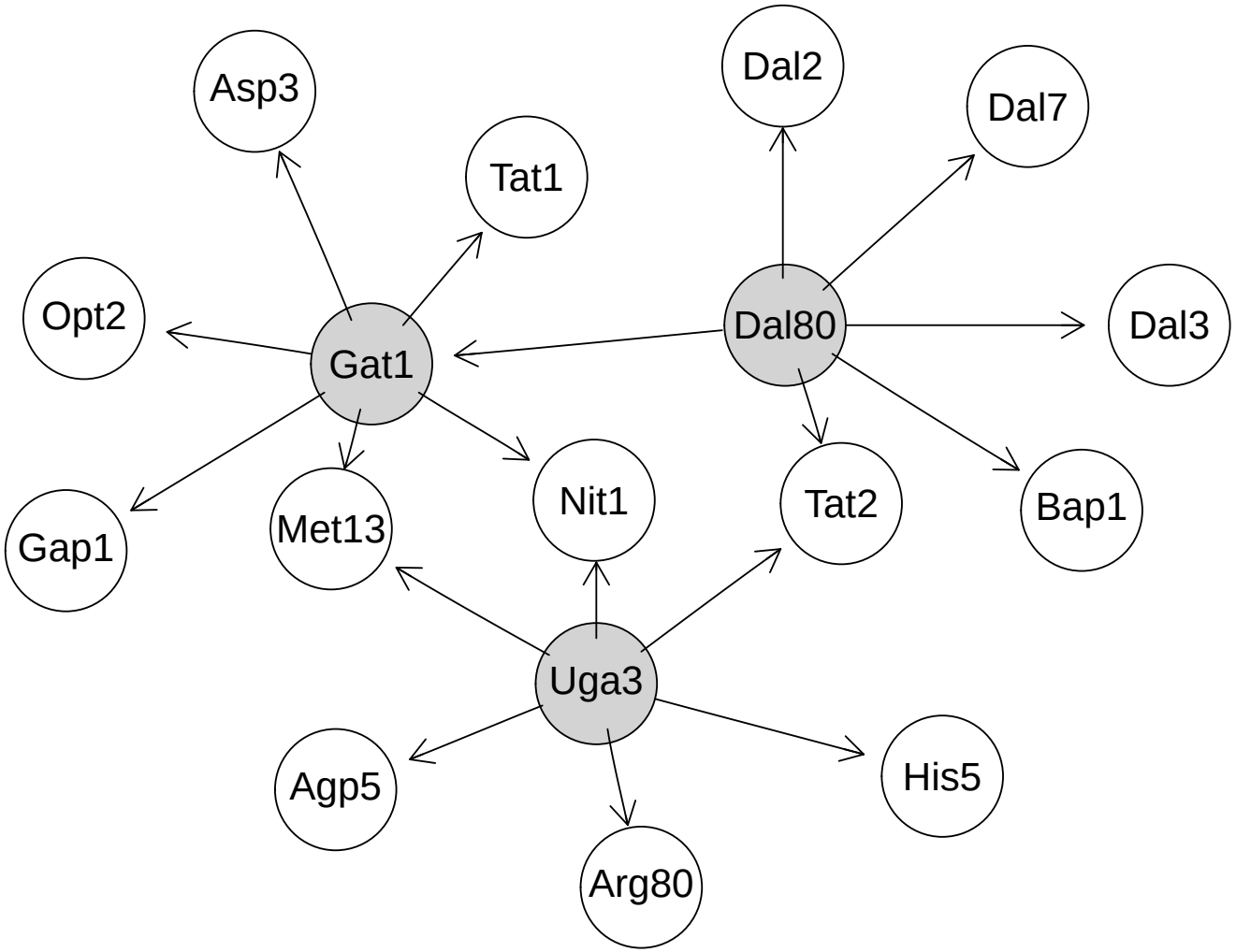
- gene expression data [6, 19], which measure the intensity of the activity of a particular gene through the presence of *messenger RNA* or other kinds of *non-coding RNA*;
- protein signalling data [17], which measure the proteins produced as a result of each gene's activity;
- sequence data [11], which provide the nucleotide sequence of each gene. For both biological and computational reasons, such data contain mostly *biallelic single-nucleotide polymorphisms* (SNPs).

Gene Expression Data

Gene expression data are composed of a set of intensities from a microarray measuring the abundance of several RNA patterns, each meant to probe a particular gene.

- Microarrays measure abundances only in terms of relative probe intensities, so comparing different studies or including them in a meta-analysis is difficult in practice.
- Furthermore, even within a single study abundance measurements are systematically biased by batch effects introduced by the instruments and the chemical reactions used in collecting the data.
- Gene expression data are modelled as continuous random variables either assuming a Gaussian distribution or applying results from robust statistics.

Gene Expression Data



Network with regulator (grey) and target (white) genes from Friedman *et al.* [6].

Models for Gene Expression Data

Two classes of undirected graphical models are in common use:

- relevance networks [2], also known in statistics as correlation graphs, which are constructed using marginal dependencies.
- gene association networks, also known as concentration graphs or graphical Gaussian models [24], which consider conditional rather than marginal dependencies.

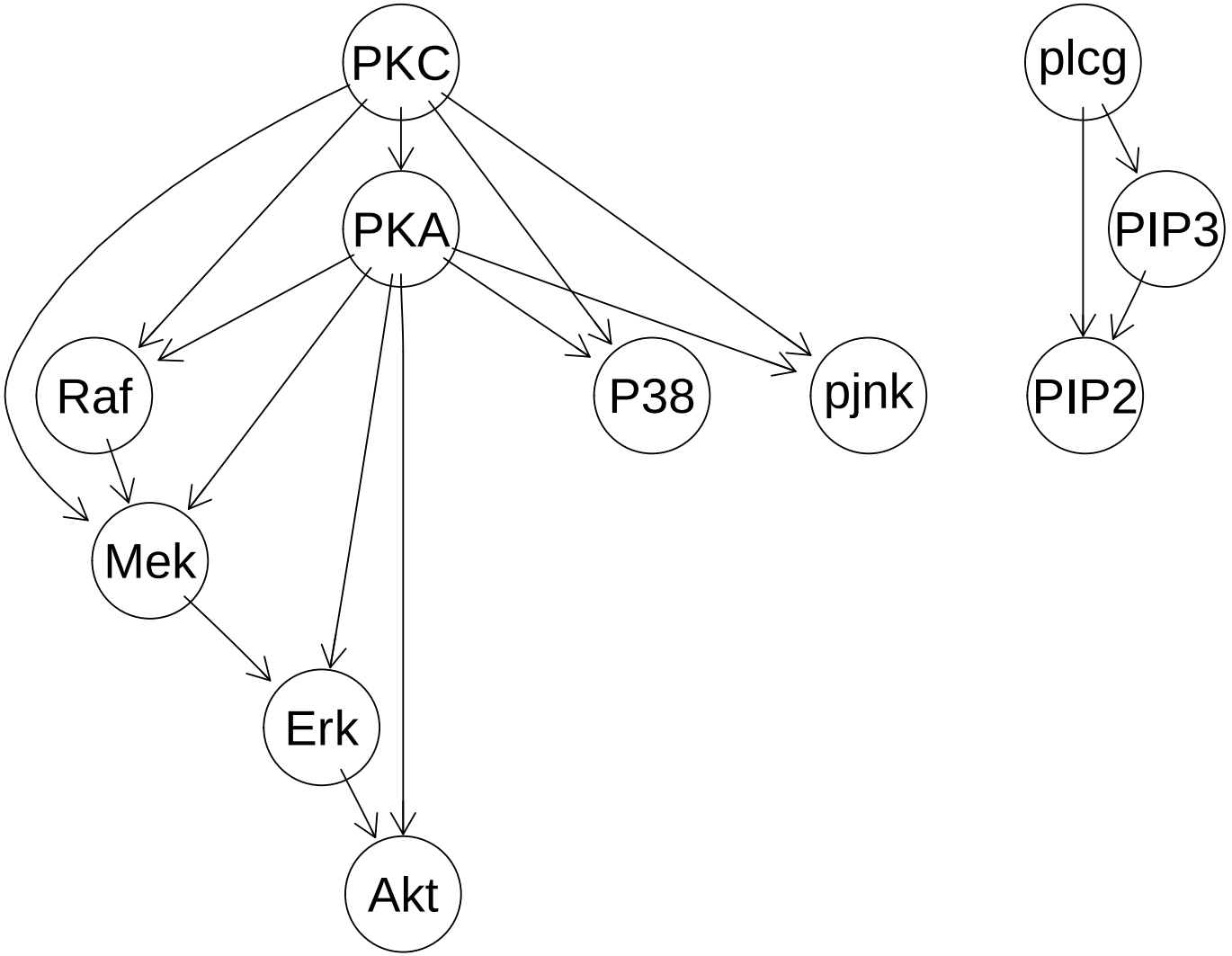
Bayesian network use by Friedman *et al.* [7], and has also been reviewed more recently in Friedman [4]. Inference procedures are usually unable to identify a single best BN, settling instead on a set of equally well behaved models. For this reason, it is important to incorporate prior biological knowledge into the network through the use of informative priors [12].

Protein Signalling Data

Protein signalling data are similar to gene expression data in many respects.

- In fact, they are often used to investigate indirectly the expression of a set of genes.
- The relationships between proteins are indicative of their physical location within the cell and of the development over time of the molecular processes (pathways) they are involved in.
- Protein signalling data sometimes have sample sizes that are much larger than either gene expression or sequence data.

Protein Signalling Data



Network from the multi-parameter single-cell data from Sachs *et al.* [17].

Sequence Data

Sequence data analysis focuses on modelling the behaviour of one or more phenotypic traits (*e.g.* the presence of a disease in humans, yield in plants, milk production in cows) by capturing direct and indirect causal genetic effects:

- the identification of the genes that are strongly associated with a trait is called a genome-wide association study (GWAS);
- the prediction of a trait for the purpose of implementing a selection program (*i.e.* to decide which plants or animals to cross so that the offspring exhibit) is called genomic selection (GS).

Models for Sequence Data

From a graphical modelling perspective, modelling each SNP as a discrete variable is the most convenient option; multinomial models have received much more attention in literature than Gaussian or mixed ones. On the other hand, the standard approach in genetics is to recode the alleles as numeric variables,

$$X_i = \begin{cases} 1 \text{ if the SNP is "AA"} \\ 0 \text{ if the SNP is "Aa"} \\ -1 \text{ if the SNP is "aa"} \end{cases} \quad \text{or} \quad X_i = \begin{cases} 2 \text{ if the SNP is "AA"} \\ 1 \text{ if the SNP is "Aa"} \\ 0 \text{ if the SNP is "aa"} \end{cases},$$

and use additive Bayesian linear regression models [3, 10, 14] of the form

$$\mathbf{y} = \mu + \sum_{i=1}^{n} X_i g_i + \boldsymbol{\varepsilon}, \qquad g_i \sim \pi_{g_i},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \Sigma).$$

Bayesian Statistics

Bayesian Basics: Priors and Posteriors

Following Bayes' theorem, the posterior distribution of the parameters in the model (say $\boldsymbol{\theta}$) given the data is

$$p(\boldsymbol{\theta} \mid \mathbf{X}) \propto p(\mathbf{X} \mid \boldsymbol{\theta}) \cdot p(\boldsymbol{\theta}) = L(\boldsymbol{\theta}; \mathbf{X}) \cdot p(\boldsymbol{\theta})$$

or, equivalently,

$$\log p(\boldsymbol{\theta} \mid \mathbf{X}) = c + \log L(\boldsymbol{\theta}; \mathbf{X}) + \log p(\boldsymbol{\theta}).$$

It is important to note two fundamental properties:

- $\log L(\boldsymbol{\theta}; \mathbf{X})$ is a function of the data and scales with the sample size, as $n \to \infty$;
- $\log p(\boldsymbol{\theta})$ does not scale as $n \to \infty$.

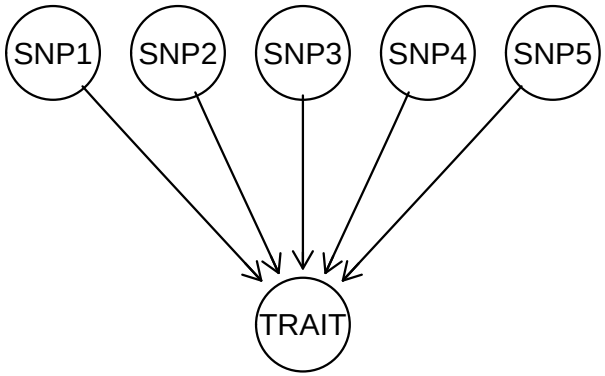
Posteriors in "Small n, Large p" Settings

Therefore, as the sample size increases, the information present in the data dominates the information provided in the prior and determines the overall behaviour of the model. For small sample sizes:

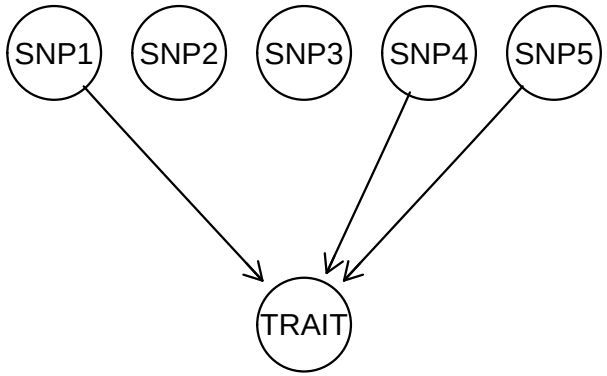
- the prior distribution plays a much larger role because there is not enough data available to disprove the assumptions the prior encodes;
- information is introduced by prior is defined not only through is hyperparameters, but from the probabilistic structure of the prior itself;
- even non-informative priors are never completely non-informative, only "least informative" [20, 21].

GWAS/GS Models vs Bayesian Networks

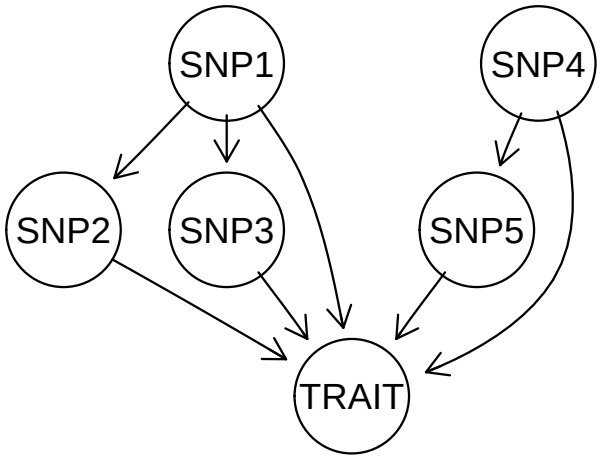
GWAS/GS Model



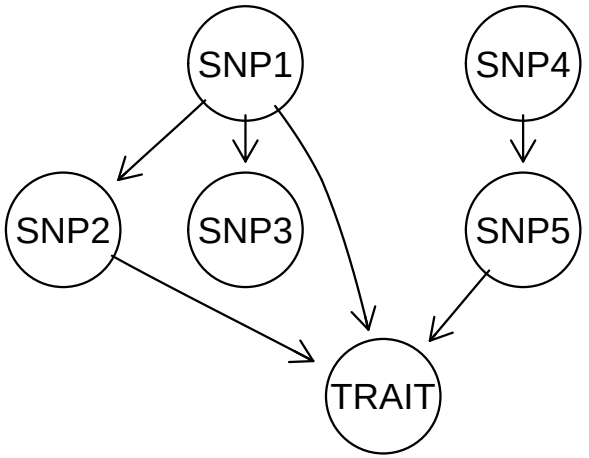
GWAS/GS Model with Feature Selection



Restricted Bayesian Network



General Bayesian Network



Parametric Assumptions

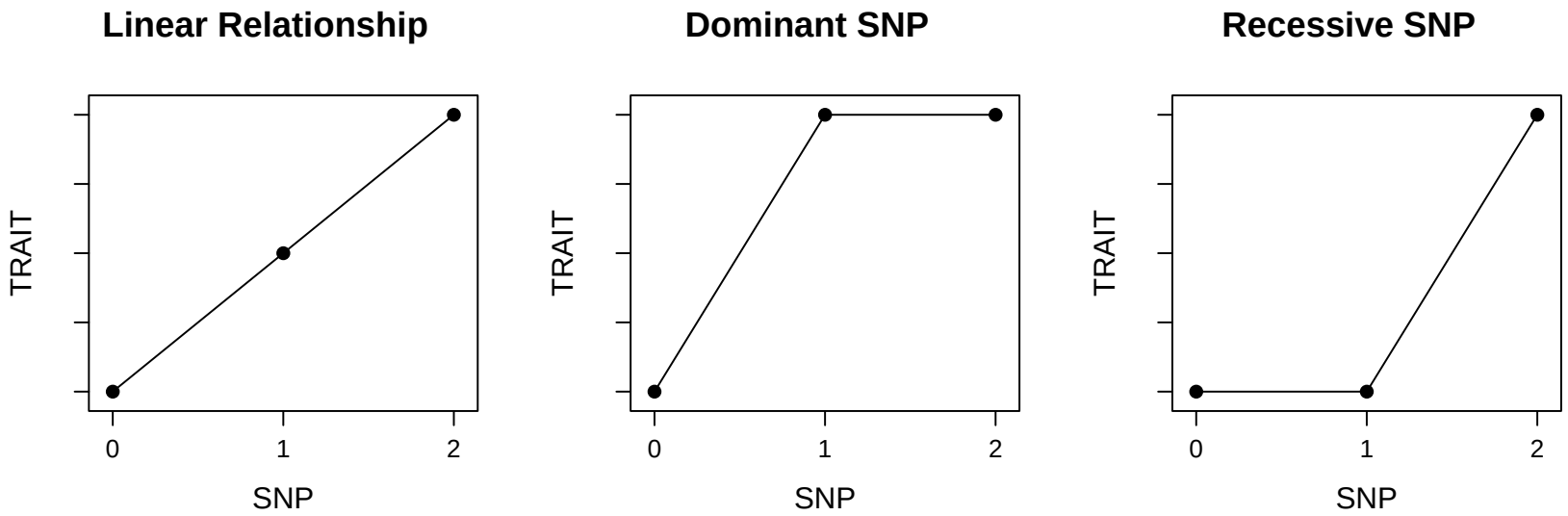
Limits of Bayesian Networks' Parametric Assumptions

Distributional assumptions underlying BNs present important limitations:

- Gaussian BNs assume that the global distribution is multivariate normal, which is unreasonable for sequence data (discrete), gene expression and protein signalling data (significantly skewed);
- Gaussian BNs are only able to capture linear dependencies;
- discrete BNs assume a multinomial distribution and disregard the ordering of the intervals (for discretised data) or of the alleles (in sequence data) is ignored.

Limits of Bayesian Networks' Parametric Assumptions

However, most biological phenomena are not linear nor unordered:



and both learning and subsequent inference are not aware that dependencies are likely to take the form of (non-linear) stochastic trends, especially in the case of sequence data.

A Test for Trend

An constraint-based approach that has the potential to outperform both discrete and Gaussian BNs has been recently proposed by Musella [13] using the Jonckheere-Terpstra test for trend among ordered alternatives [8, 22].
The null hypothesis is that of homogeneity; if we denote with $F_{i,k}(x_3)$ the distribution function of $X_3 \mid X_1 = i, X_2 = k$,

$$H_0 : F_{1,k}(x_3) = F_{2,k}(x_3) = \ldots = F_{T,k}(x_3) \quad \text{for } \forall x_3 \text{ and } \forall k.$$

The alternative hypothesis $H_1 = H_{1,1} \cup H_{1,2}$ is that of stochastic ordering, either increasing

$$H_{1,1} : F_{i,k}(x_3) \geqslant F_{j,k}(x_3) \qquad \text{with } i < j \text{ for } \forall x_3 \text{ and } \forall k$$

or decreasing

$$H_{1,2} : F_{i,k}(x_3) \leqslant F_{j,k}(x_3) \qquad \text{with } i < j \text{ for } \forall x_3 \text{ and } \forall k.$$

The Jonckheere-Terpstra Test Statistic

Consider a conditional independence test for $X_1 \perp\!\!\!\perp X_3 \mid X_2$, where X_1, X_2 and X_3 have T, L and C levels respectively. The test statistic is defined as

$$JT = \sum_{k=1}^{L} \sum_{i=2}^{T} \sum_{j=1}^{i-1} \left[\sum_{s=1}^{C} w_{ijsk} n_{isk} - \frac{n_{i+k}(n_{i+k}+1)}{2} \right]$$

where the w_{ijsk} are Wilcoxon scores, defined as

$$w_{ijsk} = \sum_{t=1}^{s-1} \left[n_{itk} + n_{jtk} + \frac{n_{isk} + n_{jsk} + 1}{2} \right],$$

and has an asymptotic normal distribution with mean and variance defined in Lehmann [9] and Pirie [16].

Feature Selection

Feature Selection in Genetics and Systems Biology

It is not possible, nor expected, for all genes in modern, genome-wide data sets to be relevant for the trait or the molecular process under study:

- for sequence data, we aim to find the subset of genes $\mathbf{S} \subset \mathbf{X}$ for a trait $\mathbf{y}$ such that

$$\mathrm{P}(\mathbf{y} \mid \mathbf{X}) = \mathrm{P}(\mathbf{y} \mid \mathbf{S}, \mathbf{X} \setminus \mathbf{S}) \approx \mathrm{P}(\mathbf{y} \mid \mathbf{S}),$$

  which is none other than the Markov blanket of the trait.
- for gene expression and protein signalling data, we need to know at least part of the pathways under investigation to initialise the feature selection. Otherwise, we can only enforce sparsity using shrinkage tests [18] or non-uniform structural priors [5].

Markov Blankets for GWAS/GS Models

After using a (reasonably fast) Markov blanket learning algorithm identify such a subset $\mathbf{S}$, we can either fit one of the Bayesian linear regression models in common use or learn a BN from $\mathbf{y}$ and $\mathbf{S}$.

PROS: in both cases, the smaller number of variables makes models more regular.

CONS: the conditional independence tests used by Markov blanket learning algorithms assume that observations are independent. Such an assumption is likely to be violated in animal and plant genetics, which make heavy use of inbred populations.

Markov Blankets for Gene Expression Data

CONS:

- we must learn the Markov blanket of each gene, which is an embarrassingly parallel task but a computationally intensive one;
- if we use backtracking and other optimisations to share information between different runs, significant speed-ups are possible at the cost of an increased error rate;
- in both cases, merging the Markov blankets requires the use of symmetry corrections [1, 23] that violate the proofs of correctness of the learning algorithms.

A better approach is the feature selection algorithm by Peña *et al.* [15].

PROS:

- it identifies in a single run all the nodes required to compute the conditional probability distribution for a given set of variables;
- it uses only pairwise measures of dependence, so it is computationally and sample efficient.

Thanks!

References

References I


C. F. Aliferis, A. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos.
Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation.
Journal of Machine Learning Research, 11:171–234, 2010.


A. J. Butte, P. Tamayo, D. Slonim, T. R. Golub, and I. S. Kohane.
Discovering Functional Relationships Between RNA Expression and Chemotherapeutic Susceptibility Using Relevance Networks.
PNAS, 97:12182–12186, 2000.


R. L Fernando D. Habier, K. Kizilkaya, and D. J. Garrick.
Extension of the Bayesian Alphabet for Genomic Selection.
BMC Bioinformatics, 12(186):1–12, 2011.


N. Friedman.
Inferring Cellular Networks Using Probabilistic Graphical Models.
Science, 303:799–805, 2004.


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.

References II


N. Friedman, M. Linial, and I. Nachman.
Using Bayesian Networks to Analyze Expression Data.
Journal of Computational Biology, 7:601–620, 2000.


N. Friedman, M. Linial, I. Nachman, and D. Pe'er.
Using Bayesian Networks to Analyze Gene Expression Data.
Journal of Computational Biology, 7:601–620, 2000.


A. Jonckheere.
A Distribution-Free k-Sample Test Against Ordered Alternatives.
Biometrika, 41:133–145, 1954.


E. L. Lehmann.
Nonparametrics: Statistical Methods Based on Ranks.
Springer, 2006.


T. H. E. Meuwissen, B. J. Hayes, and M. E. Goddard.
Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps.
Genetics, 157:1819–1829, 2001.

References III

G. Morota, B. D. Valente, G. J. M. Rosa, K. A. Weigel, and D. Gianola.
An Assessment of Linkage Disequilibrium in Holstein Cattle Using a Bayesian Network.
Journal of Animal Breeding and Genetics, 2012.
In print.

S. Mukherjee and T. P. Speed.
Network Inference using Informative Priors.
PNAS, 105:14313–14318, 2008.

F. Musella.
Learning a Bayesian Network from Ordinal Data.
Working Paper 139, Dipartimento di Economia, Università degli Studi "Roma Tre", 2011.

T. Park and G. Casella.
The Bayesian Lasso.
Journal of the American Statistical Association, 103(482), 2008.

References IV

J. Peña, R. Nilsson, J. Björkegren, and J. Tegnér.
Identifying the Relevant Nodes Without Learning the Model.
In *Proceedings of the 22nd Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-06)*, pages 367–374, 2006.

W. Pirie.
Jonckheere Tests for Ordered Alternatives.
In *Encyclopaedia of Statistical Sciences*, pages 315–318. Wiley, 1983.

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 and A. Brogini.
Bayesian Network Structure Learning with Permutation Tests.
Communications in Statistics – Theory and Methods, 41(16–17):3233–3243, 2012.

P. Spirtes, C. Glymour, and R. Scheines.
Causation, Prediction, and Search.
MIT Press, 2000.

References V


H. Steck.
Learning the Bayesian Network Structure: Dirichlet Prior versus Data.
In *Proceedings of the 24th Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-08)*, pages 511–518, 2008.


H. Steck and T. Jaakkola.
On the Dirichlet Prior and Bayesian Regularization.
In *Advances in Neural Information Processing Systems (NIPS)*, pages 697–704, 2002.


T. J. Terpstra.
The Asymptotic Normality and Consistency of Kendall's Test Against Trend When the Ties Are Present in One Ranking.
Indagationes Mathematicae, 14:327–333, 1952.


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.


J. Whittaker.
Graphical Models in Applied Multivariate Statistics.
Wiley, 1990.