# Forward elements in graphical model search

Joe Whittaker (University of Lancaster), joe.whittaker@lancaster.ac.uk
Florian Martin and Yang Xiang (Philip Morris International Research and Development)

**Abstract**

We resurrect regression elements in context of graphical Gaussian model (GGM) search and define the forward elements of entropy. We give a detailed interpretation of the third order forward elements (triangular elements) including their use for data analysis and their relevance to GGM search. We propose a specific search procedure, FE3, and compare to the well known algorithms PC and Aracne and also to our coding of a penalised regression algorithm. Our study is work in progress, though our interim findings suggest FE3 is competitive with the best but more efficient.

## 1 Introduction

Graphs in which nodes represent random variables, and edges represent probabilistic dependencies between the variables, underly graphical models: Koller and Friedman (2009), Lauritzen (1996), Whittaker (1990). Our interest is the search for large scale Gaussian graphical models from experiments that typically generate data with a large number of variables $p$ but moderate or small number of repetitions $n$. The standard approach to global search for good or best models is to score the model using the maximised log-likelihood function; however this is computationally too difficult when $p$ is large. Local search for well fitting graphical models identify conditional independence (CI) statements, such as $X_1 \perp\!\!\!\perp X_2|X_3$ consistent with the data; these usually require analysis of low dimensional margins and so are computationally feasible whatever $p$. Progress has been made in recent years by researchers in at least three fields: statistics, machine learning and bioinformatics, though often working without reference to each other. Some key references are given in Section 3 below.

**Insight:** The marginal mutual information (MI) of two random variables and the conditional MI (CMI) of two variables given a third are, respectively

$$\begin{aligned} I_{12} &= \inf(X_1 \perp\!\!\!\perp X_2) = \mathsf{E}\log\frac{f_{12}}{f_1 f_2}, \\ I_{12|3} &= \inf(X_1 \perp\!\!\!\perp X_2|X_3) = \mathsf{E}\log\frac{f_{12|3}}{f_{1|3}f_{2|3}}, \end{aligned}$$

Cover and Thomas (2006). Here $f_A$ is the density function corresponding to the random variables indexed in the subset $A$. Consider the difference in MIs, $\Delta_{123}$ say, and using the definition of a conditional probabilty,

$$\begin{aligned} \Delta_{123} &= \inf(X_1 \perp\!\!\!\perp X_2) - \inf(X_1 \perp\!\!\!\perp X_2|X_3) \\ &= -\mathsf{E}\log\frac{f_{123}f_1 f_2 f_3}{f_{12}f_{13}f_{23}}. \end{aligned} \quad (1)$$

Surprisingly the right hand side is symmetric under permutation of the variables, and so

$$\Delta_{123} = I_{12} - I_{12|3} = I_{13} - I_{13|2} = I_{23} - I_{23|1}.$$

The corollary of interest to us is that the conditional MIs $\{I_{12|3}, I_{13|2}, I_{23|1}\}$ may be computed from four numbers: $\{I_{12}, I_{13}, I_{23}\}$ and $\Delta_{123}$; rather than from six. Our aim is to exploit this result.

**Regression and additive elements:** Following Newton and Spurrell (1967), Whittaker (1984) suggests that the interpretation of a fitted statistical model such as the classical linear or the generalized linear model is substantially

clarified by partitioning of the maximized log-likelihood ratio test statistic into additive elements. A consistent notation is developed in which the primary elements measure the unique contribution of each explanatory variable whereas the secondary and higher order elements measure the effective balance in the observed design. We extend this work to the analysis of conditional independence models for high dimensional multivariate data.

**Gaussian entropy, likelihood and mutual information:** Definitions of these terms in the context of conditional indepenendence modelling may be found in Whittaker (1990), Lauritzen (1996), Cover and Thomas (2006).

The Shannon entropy of a continuous random variable or vector $X$ is $-\mathsf{E}\log f_X(X)$ where $f_X$ is the density function of $X$. For the multivariate Normal distribution $\mathsf{N}(\mu, \Sigma)$ the Gaussian entropy is

$$-\mathsf{E}\log f_X(X) = const + \frac{1}{2}\log|\Sigma|,$$

where the constant is $\frac{1}{2}p(1+\log 2\pi)$. The MI and CMI at (1) above are differences in this entropy.

**Loglikelihood, entropy and MI:** There is a close connection between the empirical Shannon entropy and the log-likelihood from a independent sample of $n$ identically distributed random variables $\ell = n \times \overline{\log f(x)}$. Log-likelihood ratio tests (deviances) are related to linear contrasts of empirical entropies. In particular it is not surprising to find that tests of independence or conditional independence within the Gaussian family are equivalent to making comparisons of the empirical MI or CMI with a threshold. In the Gaussian case computation of CMIs devolves to evaluation of determinants of minors of the empirical variance matrix. Consequently we describe the search procedures in terms of MI and cutoffs rather than in terms of hypothesis tests and their significance levels.

**Outline:** In Section 2 we generalise the computation of $\Delta$ to an arbitrary (finite) set of random variables, and define the forward elements from the Mobius inversion of the entropy function on a binary lattice. The application of forward elements in the local search for Gaussian graphical models is discussed in Section 3, and an algorithm FE3 based on third order elements is suggested. In Section 4 the results of comparing this with other competitors are given.

## 2 Forward elements from Mobius inversion of the entropy

We write $f_A$ for the density of the random vector $X_A$ indexed by the elements of $A$. The entropy is monotone function over the set of subsets of random variables in the sense that

$$-\mathsf{E}\log f_A \geq -\mathsf{E}\log f_B \text{ whenever } A \subseteq B,$$

Differences in entropy have interesting statistical interpretations, for instance $-\mathsf{E}\log f_B + \mathsf{E}\log f_A = -\mathsf{E}\log f_{B|A}$ is the conditional entropy. The set of subsets is a binary lattice.

**Example:** The lattice diagram of the power set of all subsets of $\{1,2,3\}$ is in Figure 1.

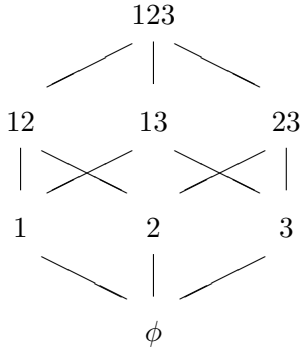



Figure 1: 3-dimensional binary lattice

Forward elements are constructed from a Mobius inversion of the entropies on this binary lattice. More generally, a lattice is the pair $(L, \leq)$ where $L$ is a (finite) set of points and $<$ is a partial order. Consider a function $h$ defined on a lattice and satisfying the additivity relation

$$h(a) = \sum_{b \leq a} g(b) \text{ for all } a \in L. \qquad (2)$$

We call the values of $g$ the forward elements of

$h$. The sum is taken over all points in the lattice $L$ that lie at or below $a$. This is a triangular system of equations which can be solved by back substitution. For example, on the lattice of the first $k$ integers the forward elements of a function $h$ are just its first differences. The forward elements are the solution of (2) for $g$, and can be written explicitly as

$$g(a) \quad = \quad \sum_{b \leq a} (-1)^{|a|-|b|} h(b) \text{ for all } a \in L. \quad (3)$$

This is an instance of Mobius inversion (Rota, 1964) and also can be obtained from the principle of inclusion-exclusion.

Using the notation of our application

$$\Delta_A \quad = \quad \sum_{B \subseteq A} (-1)^{|A|-|B|} (-\mathsf{E} \log f_B) \quad (4)$$

for all $A \in L$, with the entropy function taking the role of $h$. The first order forward element is the original entropy and the second order forward element is the negative of the mutual information for marginal independence:

$$\Delta_1 = -\mathsf{E} \log f_1 \text{ and } \Delta_{12} = -\mathsf{E} \log \frac{f_{12}}{f_1 f_2},$$

as $-\mathsf{E} \log f_\phi = 0$. The third order forward element is the symmetric function given above at (1).

The entropy function on the 3-dimensional lattice and the corresponding forward elements are displayed in Figures 2 and 3.

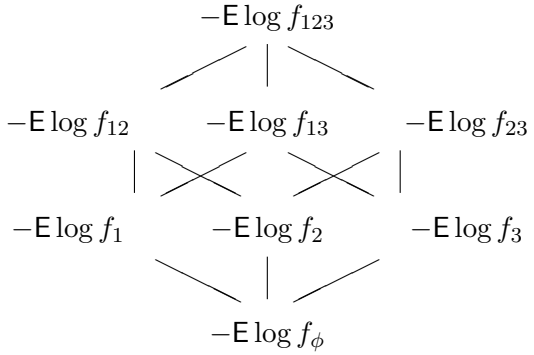



Figure 2: 3-dimensional entropies

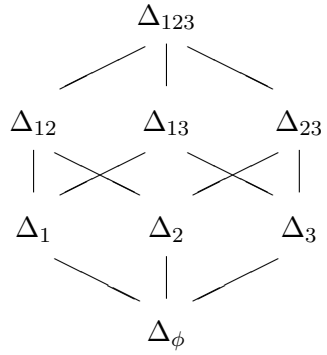



Figure 3: 3-dimensional forward elements

**Gaussian forward elements:** When evaluated for the Gaussian entropy the third order (triangular) forward element, $\Delta_{123}$, is

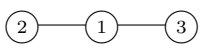
$$\frac{1}{2} \log \frac{|\operatorname{var}(X_{123})||\operatorname{var}(X_1)||\operatorname{var}(X_2)||\operatorname{var}(X_3|)}{|\operatorname{var}(X_{12})||\operatorname{var}(X_{13})||\operatorname{var}(X_{23})|},$$

and is clearly symmetric in the variables. Its empirical counterpart replaces $\Sigma = \operatorname{var}(X)$ by the sample variance matrix.

**Some interpretation of $\Delta$:** There are three features we wish to draw attention to: the interpretation of large values in their own right; the efficient computation of $\inf(X_i \perp\!\!\!\perp X_j | X_k)$ for all values of $i, j, k$; and the application of forward elements to local search procedures for graphical models.

Firstly note that a large value of $\Delta_{123}$ indicates that variable 3 affects the relationship between variables 1 and 2, as is evident by (1). This is compatible with a graph in which 3 is connected to both 1 and 2, and hence by symmetry when $1, 2, 3$ are all connected.

However this is not a necessary condition. Consider the chain CI graph




defined by $I_{23|1} = 0$. In this graph it appears that $I_{12|3}$ and $I_{12}$ might be equal, as 3 should not affect the relation between 1 and 2, and thus $\Delta_{123}$ should be 0. However this is not necessar-

ily true.

$$\begin{aligned} I_{12|3} &= \mathsf{E}\log\frac{f_{12|3}}{f_{1|3}f_{2|3}} = \mathsf{E}\log\frac{f_{123}f_3}{f_{13}f_{23}} \\ &= \mathsf{E}\log\frac{f_{12}f_{13}f_3}{f_1 f_{13} f_{23}}, \text{ using } (X_2 \perp\!\!\!\perp X_3|X_1, \\ &= \mathsf{E}\log\frac{f_{12}}{f_1 f_{2|3}}. \end{aligned}$$

The last term is not $I_{12}$ so that $\Delta_{123} \neq 0$.

**Efficient computation of first order CMIs:** To obtain $\inf(X_i \perp\!\!\!\perp X_j|X_k)$ for all triples of variables, without appeal to symmetry, requires $p\binom{p-1}{2}$ CMI calculations, To obtain these using forward elements requires $\binom{p}{2}$ evaluations of marginal MIs and $\binom{p}{3}$ evaluations of $\Delta_{ijk}$. There is an improvement of efficiency by a factor of 1/3, supposing that each calculation is roughly of the same order of magnitude.

## 3 Local search using forward elements

Local search algorithms identify conditional independence statements from data and use these to eliminate edges in the graph. The first suggested appears to be the PC algorithm Spirtes *et al.* (2000) which proved effective and has theoretical properties that guarantee the correct choice of model. The starting point is to compute all pairs of MIs and threshold these to find the initial set of adjacencies; in effect this is the relevance network proposed by Butte *et al.* (2000) in the bioinformatics literature. Then, for each edge $(i, j)$ in the graph, PC searches for a subset $A$ from the neighbours of that edge for which $X_i \perp\!\!\!\perp X_j|X_A$; if such a subset is found the edge is dropped from the graph. The first pass considers conditional independence of order $|A| = 1$, The procedure increases the size of $A$ sequentially until either the edge is dropped or a limit on the size of $A$ is breached.

Bioinformatic papers that suggest networks based on low order conditional independence are Magwene and Kim (2004) and Wille and Bühlmann (2006). The Aracne algorithm Margolin *et al.* (2006) also starts with the relevance network and then considers triples of variables with mutually adjacent edges in the graph, i.e. triangles. It applies the information inequality (Cover and Thomas, 2006) to eliminate the weakest of the three edges, and the one most likely to correspond to a first order conditional independence. This makes Aracne computationally efficient. In both the PC and the Aracne algorithms the graph is sequentially updated as each edge test is considered.

We propose an algorithm based on computing some 3rd order forward elements of the entropy. In common with the others, the relevance network provides the initial adjacency matrix. In this graph we find a triple of 3 nodes, $(i, j, k)$ say, whose adjacencies form a triangle or form a chain [O—O—O]; and then evaluate the corresponding forward element $\Delta_{ijk}$. The triple determines three possible edges which are now tested for first order conditional independence using the 3 marginal MIs and $\Delta_{ijk}$. Note that if $\Delta_{ijk} < 0$ an edge may be added, unlike other local search procedures. If $\Delta_{ijk} > 0$ and the initial adjacencies form a triangle, then if one edge is eliminated it is the weakest. This leads to a graph in which some edges are eliminated by thresholding $\inf(X_i \perp\!\!\!\perp X_j)$ and others by thresholding $\inf(X_i \perp\!\!\!\perp X_j|X_k)$.

There is one difficulty. When two triples overlap, for instance $(1, 2, 3)$ and $(1, 2, 4)$, the order in which the procedure is applied may lead to different outcomes, as $I_{12|3}$ and $I_{12|4}$ may have different implications for edge $(1, 2)$. This makes a sequential procedure difficult to implement; it is resolved below within a batch procedure that considers all triangular elements together.

**FE3: a forward elements batch procedure:** The steps of the FE3 are

1. Compute $I_{ij}$ for all pairs of nodes. Eliminate the edge $(i, j)$ from the complete graph whenever

   $$I_{ij} < m,$$

   to build $A^1$.

2. Find all triples of nodes with one that has two neigbours in $A^1$.

3. Find $\Delta_{ijk}$ for all such triples and store.

4. Order the triples in ascending $\Delta$.

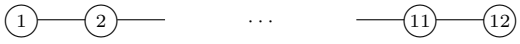
5. Use this order of triples to threshold the three conditional MIs $(I_{ij|k}, I_{ik|j}, I_{jk|i}) = (I_{ij} - \Delta_{ijk}, I_{ik} - \Delta_{ijk}, I_{jk} - \Delta_{ijk})$ against $m$, or equivalently

$$(I_{ij}, I_{ik}, I_{jk}) < m + \Delta_{ijk}, \tag{5}$$

and update $A^1$ to get $A^2$.

**Remarks:** Several points are worth making: This procedure visits nodes rather than edges, unlike some other algorithms. Edges may be added, unlike other algorithms where they are only subtracted. The effect of ordering is to remove the arbitrariness in updating the edges sequentially. As the procedure evaluates all relevant third order elements before updating the adjacencies it is a batch-like procedure. Updating the adjacency matrix in ascending order of $\Delta$ implies edges that might be added are considered before edges that might be subtracted. Consequently a putative CI in the update is not overwritten by another in the same update. The update at (5) might be interpreted as the application of a local threshold to a relevance network. The procedure could be repeated on $A^2$ until convergence.

**A Markov chain example:** An autoregression is generated on $p = 12$ variables from $X_{j+1} = 0.2X_j + \epsilon_j$, where the errors are iid, and 100 repetitions of the process are observed. The theoretical independence graph is just the chain




After the first pass the adjacency matrix of the relevance network, $A^1$, is

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [2,] | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [3,] | 1 | 1 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [4,] | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 | 0 |
| [5,] | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 1 | 0 | 0 | 0 |
| [6,] | 0 | 0 | 0 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| [7,] | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 | 0 |
| [8,] | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 1 | 0 |
| [9,] | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 1 | 0 | 1 | 1 | 0 |
| [10,] | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 | 1 |
| [11,] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 1 |
| [12,] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 |

The matrix is roughly banded, but there are clearly two bands, and maybe three bands beneath the diagonal, rather than the one band of the generating process. There are over 80 distinct triples in this graph. Applying the second pass of FE3 leads to an adjacency matrix $A^2$ that is correct apart from 1 wrong entry. The differences $A^1 - A^2$ are

| | [,1] | [,2] | [,3] | [,4] | [,5] | [,6] | [,7] | [,8] | [,9] | [,10] | [,11] | [,12] |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| [1,] | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [2,] | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [3,] | 1 | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| [4,] | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | -1 | 0 | 0 |
| [5,] | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 |
| [6,] | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| [7,] | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 | 0 |
| [8,] | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 1 | 0 |
| [9,] | 0 | 0 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | 0 | 1 | 0 |
| [10,] | 0 | 0 | 0 | -1 | 0 | 0 | 1 | 1 | 0 | 0 | 0 | 1 |
| [11,] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 1 | 0 | 0 | 0 |
| [12,] | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 | 0 |

One entry is negative corresponding to adding an edge.

## 4 Results

Important aspects of the fitted model include the quality of its predictions, and the accuracy of the revealed structure. Having the right edges is crucial to scientific understanding and having the right variance matrix determines predictive power. A global criterion such as the likelihood function or a penalised version is the natural candidate that combines these two aspects. However in our context of large $p$ this is infeasible, and we restrict attention to comparing the found graph with the true graph that generated the data in terms of the number of wrong edges.

We report here some of the interim results of a small scale simulation study, to test the efficacy of the FE3 algorithm.

**Evaluation:** We use a standard setup for simulation where a data set is sampled from the graphical Gaussian model analysed, and the result is compared to ground truth. First, the experimental parameters: sample size $n$, number of variables $p$, and the expected number of edges, are set. A random choice of $G_{true}$ is made by selecting edges with equal probability, A random choice of the precision (inverse variance) matrix $D_{true}$ respecting the edges of $G_{true}$ is generated by choosing a $\mathsf{Unif}(-1, 1)$ value of the precision for the edges selected at the first

step. The matrix is made diagonally dominant, checked to be positive definite and its inverse is returned as the variance $\Sigma$. A sample of $X$ is drawn from the corresponding multivariate Normal distribution. The data sample is searched using each algorithm in the study, and the process is repeated for enough replications of the given settings. Note that just one realisation of data is analysed for one choice of $G_{true}$.

**A comparison of some chosen local methods:** We evaluate several local methods for comparison with the forward element method of order 3 (FE3). The methods under consideration are

- RN the relevance network with a chi-sq threshold; this is included as a very simple benchmark, and also because it is the first step in several other algorithms including FE3.

- PC the path consistency algorithm, as this is well known for its good performance; the code is taken from the R-package pcalg.

- CN is our version of a sparse regression algorithm that examines each node with the R-package lars to find the neighbours.

- AR the Aracne algorithm, well thought of among bioinformaticians, that extends the relevance network to inspect triples of nodes. The code is contained in the Bioconductor package minet, Meyer *et al.* (2008).

**Threshold and tuning parameter:** The cutoff for the mutual information $MI_{cut}$ is determined by the 95% chi-squared quantile. This threshold is perturbed by a tuning parameter $\gamma \in (0, 1)$ to a cutoff of $\lambda = MI_{cut} \exp(3(\gamma - 0.5))$. In our experiments 10 equi-spaced values of the tuning parameter are considered. The simulation parameter settings are held fixed for $nreps = 8$ repetitions and the evaluation reports the median wrong edge count within the repetitions.

**Wrong edge evaluation:** The number of wrong edges, not distinguishing between false positives and false negatives, are plotted against the tuning parameter in Figure 4. The value $\gamma = 0.5$ corresponds to the default threshold cutoff.

Our findings are that:

– The plots are fairly reproducible with $p = 32$, $n = 50 \times 32$, few errors and 8 repetitions. There are approximately 150 edges in any one graph and 350 missing edges.

– FE3 is competitive with PC and both are superior to AR.

– The number of wrong edges for FE3, PC, AR are flat when plotted against the tuning parameter.

– The fluctuations in CN are real but, relative to the other algorithms, it is repeatedly unstable.

– RN is improved by increasing $\gamma$ and surprisingly gets down to levels of PC and FE3 when it approaches 1.

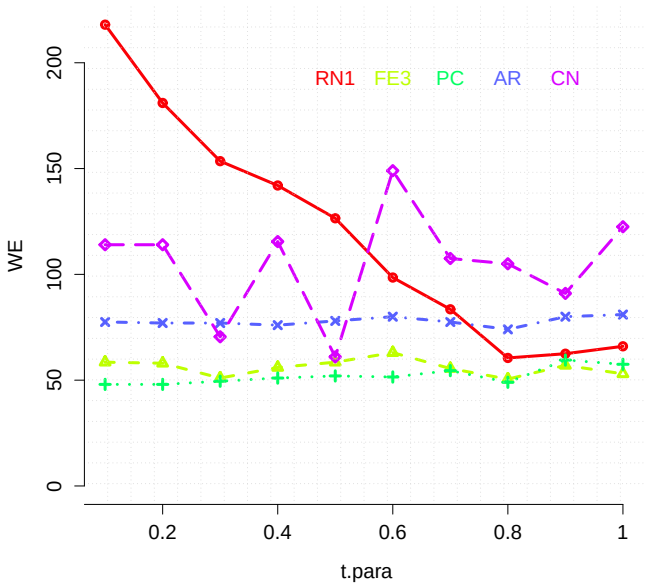



Figure 4: Wrong edges for several local search methods.

**Next steps:** These preliminary findings are encouraging though some issues have to be resolved before finalisation of this interim study.

**Acknowledgement:** This work is in part supported by Philip Morris International.

## References


Butte, A., Tamayo, P., Slonim, D., Golub, T., and Kohane, I. (2000). Discovering functional relationships between rna expression and chemotherapeutic susceptibility using relevance networks. *Proceedings of the National Academy of Sciences*, **97**(22), 12182–12186.

Cover, T. M. and Thomas, J. A. (2006). *Elements of Information Theory*. Wiley-interscience, 2nd edition.

Koller, D. and Friedman, N. (2009). *Probabilistic Graphical Models: Principles and Techniques*. Mit Press, Cambridge.

Lauritzen, S. L. (1996). *Graphical Models*. Oxford University Press.

Magwene, P. M. and Kim, J. (2004). Estimating genomic coexpression networks using first-order conditional independence. *Genome Biology*, **5:R100**.

Margolin, A., Nemenman, I., Basso, K., Wiggins, C., Stolovitzky, G., Dalla Favera, R., and Califano, A. (2006). Aracne: an algorithm for the reconstruction of gene regulatory networks in a mammalian cellular context. *BMC Bioinformatics*, **20;7**, Suppl 1:S7.

Meyer, P., Lafitte, F., and Bontempi, G. (2008). minet: A r/bioconductor package for inferring large transcriptional networks using mutual information. *BMC bioinformatics*, **9**(1), 461.

Newton, R. and Spurrell, D. (1967). A development of multiple regression for the analysis of routine data. *Applied Statistics*, pages 51–64.

Rota, G.-C. (1964). On the foundations of combinatorial theory I. Theory of Möbius functions. *Probability Theory and Related Fields*, **2**(4), 340–368.

Spirtes, P., Glymour, C., and Scheines, R. (2000). *Causation, Prediction and Search*. MIT Press, New York, 2nd edition.

Whittaker, J. (1984). Model interpretation from the additive elements of the likelihood function. *Applied Statistics*, pages 52–64.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. Wiley, Chichester.

Wille, A. and Bühlmann, P. (2006). Low-order conditional independence graphs for inferring genetic networks. *Statistical applications in genetics and molecular biology*, **5**(1), 1–32.