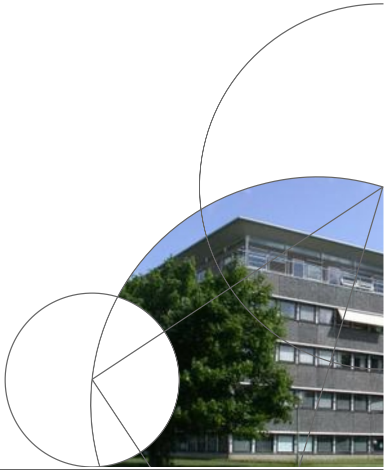
Faculty of Science

Markov Properties and the Multivariate Gaussian Distribution

Steffen Lauritzen
Department of Mathematical Sciences



Overview of lectures

Lecture 1 Markov Properties and the Multivariate Gaussian Distribution

Lecture 2 Likelihood Analysis of Gaussian Graphical Models

Lecture 3 Gaussian Graphical Models with Symmetry

For reference, if nothing else is mentioned, see Lauritzen (1996), Chapters 3 and 4.

We recall that two random variables X and Y are *independent* if

$$P(X \in A \mid Y = y) = P(X \in A)$$

or, equivalently, if

$$P\{(X \in A) \cap (Y \in B)\} = P(X \in A)P(Y \in B).$$

For continuous variables the requirement is a factorization of the joint density:

$$f_{XY}(x, y) = f_X(x) f_Y(y).$$

When X and Y are independent we write $X \perp\!\!\!\perp Y$.

Formal definition

Random variables X and Y are *conditionally independent* given the random variable Z if

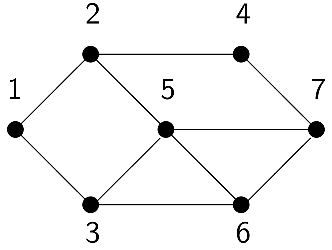
$$\mathcal{L}(X \mid Y, Z) = \mathcal{L}(X \mid Z).$$

We then write $X \perp\!\!\!\perp Y \mid Z$ (or $X \perp\!\!\!\perp_P Y \mid Z$)

Intuitively: Knowing Z renders Y *irrelevant* for predicting X.

Factorisation of densities:

$$\begin{aligned} X \perp\!\!\!\perp Y \mid Z \quad &\iff \quad f_{XYZ}(x, y, z) f_Z(z) = f_{XZ}(x, z) f_{YZ}(y, z) \\ &\iff \quad \exists a, b : f(x, y, z) = a(x, z) b(y, z). \end{aligned}$$



For several variables, complex systems of conditional independence can for example be described by undirected graphs.

Then *a set of variables A is conditionally independent of a set B, given the values of a set of variables C, if C separates A from B.*

For example in picture above

$$1 \perp\!\!\!\perp \{4,7\} \mid \{2,3\}, \qquad \{1,2\} \perp\!\!\!\perp 7 \mid \{4,5,6\}.$$

For random variables X, Y, Z, and W it holds

(C1) If $X \perp\!\!\!\perp Y \mid Z$ then $Y \perp\!\!\!\perp X \mid Z$;

(C2) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp U \mid Z$;

(C3) If $X \perp\!\!\!\perp Y \mid Z$ and $U = g(Y)$, then $X \perp\!\!\!\perp Y \mid (Z, U)$;

(C4) If $X \perp\!\!\!\perp Y \mid Z$ and $X \perp\!\!\!\perp W \mid (Y, Z)$, then $X \perp\!\!\!\perp (Y, W) \mid Z$;

If density w.r.t. product measure $f(x, y, z, w) > 0$ also

(C5) If $X \perp\!\!\!\perp Y \mid (Z, W)$ and $X \perp\!\!\!\perp Z \mid (Y, W)$ then $X \perp\!\!\!\perp (Y, Z) \mid W$.

Conditional independence can be seen as encoding abstract irrelevance: *Knowing C, A is irrelevant for learning B,* (C1)–(C4) translate into:

(I1) If, knowing C, learning A is irrelevant for learning B, then B is irrelevant for learning A;

(I2) If, knowing C, learning A is irrelevant for learning B, then A is irrelevant for learning any part D of B;

(I3) If, knowing C, learning A is irrelevant for learning B, it remains irrelevant having learnt any part D of B;

(I4) If, knowing C, learning A is irrelevant for learning B and, having also learnt A, D remains irrelevant for learning B, then both of A and D are irrelevant for learning B.

The property analogous to (C5) is slightly more subtle and not generally obvious.

An *independence model* (Studený, 2005) $\perp_\sigma$ is a ternary relation over subsets of a finite set V. It is *graphoid* if for all subsets A, B, C, D:

(S1) if $A \perp_\sigma B \mid C$ then $B \perp_\sigma A \mid C$ (symmetry);

(S2) if $A \perp_\sigma (B \cup D) \mid C$ then $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid C$ (decomposition);

(S3) if $A \perp_\sigma (B \cup D) \mid C$ then $A \perp_\sigma B \mid (C \cup D)$ (weak union);

(S4) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid (B \cup C)$, then $A \perp_\sigma (B \cup D) \mid C$ (contraction);

(S5) if $A \perp_\sigma B \mid (C \cup D)$ and $A \perp_\sigma C \mid (B \cup D)$ then $A \perp_\sigma (B \cup C) \mid D$ (intersection).

Semigraphoid if only (S1)–(S4). It is *compositional* if

(S6) if $A \perp_\sigma B \mid C$ and $A \perp_\sigma D \mid C$ then $A \perp_\sigma (B \cup D) \mid C$ (composition).

Separation in undirected graphs

Let $\mathcal{G} = (V, E)$ be finite and simple undirected graph (no self-loops, no multiple edges).

For subsets A, B, S of V, let $A \perp_{\mathcal{G}} B \mid S$ denote that S *separates* A *from* B *in* $\mathcal{G}$, i.e. that all paths from A to B intersect S.

Fact: *The relation* $\perp_{\mathcal{G}}$ *on subsets of* V *is a compositional graphoid.*

This fact is the reason for choosing the name 'graphoid' for such independence model.

Probabilistic Independence Model

For a system V of *labeled random variables* $X_v, v \in V$, we use the shorthand

$$A \perp\!\!\!\perp B \mid C \iff X_A \perp\!\!\!\perp X_B \mid X_C,$$

where $X_A = (X_v, v \in A)$ denotes the variables with labels in A.

The properties (C1)–(C4) imply that $\perp\!\!\!\perp$ *satisfies the semi-graphoid axioms* for such a system, and the *graphoid axioms if the joint density of the variables is strictly positive.*

A regular multivariate Gaussian distribution defines a compositional graphoid independence model, as we shall see later.

Geometric orthogonality

Let L, M, and N be linear subspaces of a Hilbert space H and

$$L \perp M \mid N \iff (L \ominus N) \perp (M \ominus N),$$

where $L \ominus N = L \cap N^{\perp}$.$L$ and M are said to *meet orthogonally in* N.

- (O1) If $L \perp M \mid N$ then $M \perp L \mid N$;
- (O2) If $L \perp M \mid N$ and U is a linear subspace of L, then $U \perp M \mid N$;
- (O3) If $L \perp M \mid N$ and U is a linear subspace of M, then $L \perp M \mid (N + U)$;
- (O4) If $L \perp M \mid N$ and $L \perp R \mid (M + N)$, then $L \perp (M + R) \mid N$.

Intersection does not hold in general whereas *composition* (S6) does.

$\mathcal{G} = (V, E)$ simple undirected graph; An independence model $\perp_\sigma$ satisfies

(P) *the pairwise Markov property* if

$$\alpha \not\sim \beta \Rightarrow \alpha \perp_\sigma \beta \mid V \setminus \{\alpha, \beta\};$$

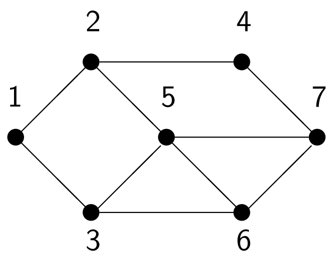
(L) *the local Markov property* if

$$\forall \alpha \in V : \alpha \perp_\sigma V \setminus \mathrm{cl}(\alpha) \mid \mathrm{bd}(\alpha);$$

(G) *the global Markov property* if

$$A \perp_{\mathcal{G}} B \mid S \Rightarrow A \perp_\sigma B \mid S.$$

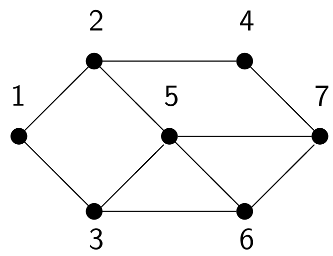
Pairwise Markov property



Any non-adjacent pair of random variables are conditionally independent given the remaning.

For example, $1 \perp_\sigma 5 \mid \{2,3,4,6,7\}$ and $4 \perp_\sigma 6 \mid \{1,2,3,5,7\}$.

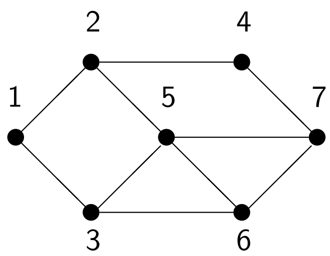
Local Markov property



Every variable is conditionally independent of the remaining, given its neighbours.

For example, $5 \perp_\sigma \{1,4\} \mid \{2,3,6,7\}$ and $7 \perp_\sigma \{1,2,3\} \mid \{4,5,6\}$.

Global Markov property



To find conditional independence relations, one should look for separating sets, such as $\{2,3\}$, $\{4,5,6\}$, or $\{2,5,6\}$

For example, it follows that $1 \perp_\sigma 7 \mid \{2,5,6\}$ and $2 \perp_\sigma 6 \mid \{3,4,5\}$.

For any semigraphoid it holds that

$$(\mathrm{G}) \Rightarrow (\mathrm{L}) \Rightarrow (\mathrm{P})$$

If $\perp_\sigma$ satisfies graphoid axioms it further holds that

$$(\mathrm{P}) \Rightarrow (\mathrm{G})$$

so that *in the graphoid case*

$$(\mathrm{G}) \iff (\mathrm{L}) \iff (\mathrm{P}).$$

The latter holds in particular for $\perp\!\!\!\perp$, when $f(x) > 0$.

A d-dimensional random vector $X = (X_1, \ldots, X_d)$ has a *multivariate Gaussian distribution* or *normal* distribution on $\mathcal{R}^d$ if there is a vector $\xi \in \mathcal{R}^d$ and a $d \times d$ matrix Σ such that

$$\lambda^\top X \sim \mathcal{N}(\lambda^\top \xi, \lambda^\top \Sigma \lambda) \quad \text{for all } \lambda \in R^d. \tag{1}$$

We then write $X \sim \mathcal{N}_d(\xi, \Sigma)$.

Taking $\lambda = e_i$ or $\lambda = e_i + e_j$ where e_i is the unit vector with i-th coordinate 1 and the remaining equal to zero yields:

$$X_i \sim \mathcal{N}(\xi_i, \sigma_{ii}), \quad \text{Cov}(X_i, X_j) = \sigma_{ij}.$$

Hence ξ is the *mean vector* and Σ the *covariance matrix* of the distribution.

The definition (1) makes sense if and only if $\lambda^\top \Sigma \lambda \geq 0$, i.e. if Σ is *positive semidefinite*. Note that we have allowed distributions with variance zero.

The multivariate moment generating function of X can be calculated using the relation (1) as

$$m_d(\lambda) = E\{e^{\lambda^\top X}\} = e^{\lambda^\top \xi + \lambda^\top \Sigma \lambda / 2}$$

where we have used that the univariate moment generating function for $\mathcal{N}(\mu, \sigma^2)$ is

$$m_1(t) = e^{t\mu + \sigma^2 t^2 / 2}$$

and let $t = 1$, $\mu = \lambda^\top \xi$, and $\sigma^2 = \lambda^\top \Sigma \lambda$.

In particular this means that *a multivariate Gaussian distribution is determined by its mean vector and covariance matrix.*

Assume $X^\top = (X_1, X_2, X_3)$ with X_i independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$. Then

$$\lambda^\top X = \lambda_1 X_1 + \lambda_2 X_2 + \lambda_3 X_3 \sim \mathcal{N}(\mu, \tau^2)$$

with

$$\mu = \lambda^\top \xi = \lambda_1 \xi_1 + \lambda_2 \xi_2 + \lambda_3 \xi_3, \quad \tau^2 = \lambda_1^2 \sigma_1^2 + \lambda_2^2 \sigma_2^2 + \lambda_3^2 \sigma_3^2.$$

Hence $X \sim \mathcal{N}_3(\xi, \Sigma)$ with $\xi^\top = (\xi_1, \xi_2, \xi_3)$ and

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & 0 \\ 0 & \sigma_2^2 & 0 \\ 0 & 0 & \sigma_3^2 \end{pmatrix}.$$

If Σ is *positive definite*, i.e. if $\lambda^\top \Sigma \lambda > 0$ for $\lambda \neq 0$, the distribution has density on $\mathcal{R}^d$

$$f(x \mid \xi, \Sigma) = (2\pi)^{-d/2} (\det K)^{1/2} e^{-(x-\xi)^\top K (x-\xi)/2}, \tag{2}$$

where $K = \Sigma^{-1}$ is the *concentration matrix* of the distribution. Since a positive semidefinite matrix is positive definite if and only if it is invertible, we then also say that Σ is *regular*.

If $X_1, \ldots, X_d$ are independent and $X_i \sim \mathcal{N}(\xi_i, \sigma_i^2)$ their joint density has the form (2) with $\Sigma = \mathrm{diag}(\sigma_i^2)$ and $K = \Sigma^{-1} = \mathrm{diag}(1/\sigma_i^2)$.

Hence *vectors of independent Gaussians are multivariate Gaussian.*

In the bivariate case it is traditional to write

$$\Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_1\sigma_2\rho \\ \sigma_1\sigma_2\rho & \sigma_2^2 \end{pmatrix},$$

with ρ being the *correlation* between X_1 and X_2. Then

$$\det(\Sigma) = \sigma_1^2\sigma_2^2(1-\rho^2) = \det(K)^{-1}$$

and

$$K = \frac{1}{\sigma_1^2\sigma_2^2(1-\rho^2)} \begin{pmatrix} \sigma_2^2 & -\sigma_1\sigma_2\rho \\ -\sigma_1\sigma_2\rho & \sigma_1^2 \end{pmatrix}.$$

Thus the density becomes

$$f(x \mid \xi, \Sigma) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{(1-\rho^2)}} \times e^{-\frac{1}{2(1-\rho^2)}\left\{\frac{(x_1-\xi_1)^2}{\sigma_1^2} - 2\rho\frac{(x_1-\xi_1)(x_2-\xi_2)}{\sigma_1\sigma_2} + \frac{(x_2-\xi_2)^2}{\sigma_2^2}\right\}}.$$

The contours of this density are ellipses and the corresponding density is bell-shaped with maximum in (ξ_1, ξ_2).

The marginal distributions of a vector X can all be Gaussian without the joint being multivariate Gaussian:

For example, let $X_1 \sim \mathcal{N}(0,1)$, and define X_2 as

$$X_2 = \begin{cases} X_1 & \text{if } |X_1| > c \\ -X_1 & \text{otherwise.} \end{cases}$$

Then, using the symmetry of the univariate Gausssian distribution, X_2 is also distributed as $\mathcal{N}(0,1)$.

However, the joint distribution is not Gaussian unless $c = 0$ since, for example, $Y = X_1 + X_2$ satisfies

$$P(Y=0) = P(X_2 = -X_1) = P(|X_1| \le c) = \Phi(c) - \Phi(-c).$$

Note that for $c = 0$, the correlation ρ between X_1 and X_2 is 1 whereas for $c = \infty$, $\rho = -1$.

It follows that *there is a value of c so that X_1 and X_2 are uncorrelated,* and still not jointly Gaussian.

Adding two independent Gaussians yields a Gaussian:
If $X \sim \mathcal{N}_d(\xi_1, \Sigma_1)$ and $X_2 \sim \mathcal{N}_d(\xi_2, \Sigma_2)$ and $X_1 \perp\!\!\!\perp X_2$

$$X_1 + X_2 \sim \mathcal{N}_d(\xi_1 + \xi_2, \Sigma_1 + \Sigma_2).$$

To see this, just note that

$$\lambda^\top (X_1 + X_2) = \lambda^\top X_1 + \lambda^\top X_2$$

and use the univariate addition property.

Linear transformations preserve multivariate normality:
If L is an $r \times d$ matrix, $b \in \mathcal{R}^r$ and $X \sim \mathcal{N}_d(\xi, \Sigma)$, then

$$Y = LX + b \sim \mathcal{N}_r(L\xi + b, L\Sigma L^\top).$$

Again, just write

$$\gamma^\top Y = \gamma^\top (LX + b) = (L^\top \gamma)^\top X + \gamma^\top b$$

and use the corresponding univariate result.

Partition X into into X_A and X_B, where $X_A \in \mathcal{R}^A$ and $X_B \in \mathcal{R}^B$ with $A \cup B = V$.
Partition mean vector, concentration and covariance matrix accordingly as

$$\xi = \begin{pmatrix} \xi_A \\ \xi_B \end{pmatrix}, \quad K = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix}, \quad \Sigma = \begin{pmatrix} \Sigma_{AA} & \Sigma_{AB} \\ \Sigma_{BA} & \Sigma_{BB} \end{pmatrix}.$$

Then, if $X \sim \mathcal{N}(\xi, \Sigma)$ it holds that

$$X_B \sim \mathcal{N}_s(\xi_B, \Sigma_{BB}).$$

This follows simply from the previous fact using the matrix

$$L = (0_{AB}\ I_B)\,.$$

where 0_{AB} is a matrix of zeros and I_B is the $B \times B$ identity matrix.

If Σ_{BB} is regular, it further holds that

$$X_A \,|\, X_B = x_B \sim \mathcal{N}_A(\xi_{A|B}, \Sigma_{A|B}),$$

where

$$\xi_{A|B} = \xi_A + \Sigma_{AB}\Sigma_{BB}^{-1}(x_B - \xi_B) \quad \text{and} \quad \Sigma_{A|B} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA}.$$

In particular, *$\Sigma_{AB} = 0$ if and only if X_A and X_B are independent.*

To see this, we simply calculate the conditional density.

$$f(x_A \mid x_B) \propto f_{\xi,\Sigma}(x_A, x_B)$$
$$\propto \exp\left\{-(x_A-\xi_A)^\top K_{AA}(x_A-\xi_A)/2 - (x_A-\xi_A)^\top K_{AB}(x_B-\xi_B)\right\}.$$

The linear term involving x_A has coefficient equal to

$$K_{AA}\xi_A - K_{AB}(x_A-\xi_B) = K_{AA}\left\{\xi_A - K_{AA}^{-1}K_{AB}(x_B-\xi_B)\right\}.$$

Using the matrix identities

$$K_{AA}^{-1} = \Sigma_{AA} - \Sigma_{AB}\Sigma_{BB}^{-1}\Sigma_{BA} \quad (3)$$

and

$$K_{AA}^{-1}K_{AB} = -\Sigma_{AB}\Sigma_{BB}^{-1}, \quad (4)$$

we find

$$f(x_A \mid x_B) \propto \exp\left\{-(x_A-\xi_{A|B})^\top K_{AA}(x_A-\xi_{A|B})/2\right\}$$

and the result follows.

From the identities (3) and (4) it follows in particular that then the conditional expectation and concentrations also can be calculated as

$$\xi_{A|B} = \xi_A - K_{AA}^{-1}K_{AB}(x_B-\xi_B) \quad \text{and} \quad K_{A|B} = K_{AA}.$$

Note that the *marginal covariance is simply expressed in terms of Σ* whereas the *conditional concentration is simply expressed in terms of K.*

Further, since

$$\xi_{A|B} = \xi_A - K_{AA}^{-1}K_{AB}(x_B-\xi_B) \quad \text{and} \quad K_{A|B} = K_{AA},$$

X_A and X_B are independent if and only if $K_{AB}=0$, giving *$K_{AB}=0$ if and only if $\Sigma_{AB}=0$.*

More generally, if we partition X into X_A, X_B, X_S, the conditional concentration matrix of $X_{A\cup B}$ given $X_C = x_C$ is simply

$$K_{A\cup B|C} = \begin{pmatrix} K_{AA} & K_{AB} \\ K_{BA} & K_{BB} \end{pmatrix},$$

so

$$X_A \perp\!\!\!\perp X_B \mid X_C \iff K_{AB} = 0.$$

It follows that a Gaussian independence model is a compositional graphoid.

Consider $\mathcal{N}_3(0,\Sigma)$ with covariance matrix

$$\Sigma = \begin{pmatrix} 1 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

The concentration matrix is

$$K = \Sigma^{-1} = \begin{pmatrix} 3 & -1 & -1 \\ -1 & 1 & 0 \\ -1 & 0 & 1 \end{pmatrix}.$$

The marginal distribution of (X_2, X_3) has covariance and concentration matrix

$$\Sigma_{23} = \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix}, \quad (\Sigma_{23})^{-1} = \frac{1}{3}\begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

The conditional distribution of (X_1, X_2) given X_3 has concentration and covariance matrix

$$K_{12} = \begin{pmatrix} 3 & -1 \\ -1 & 1 \end{pmatrix}, \quad \Sigma_{12|3} = (K_{12})^{-1} = \frac{1}{2}\begin{pmatrix} 1 & 1 \\ 1 & 3 \end{pmatrix}.$$

Similarly, $\mathbf{V}(X_1 \mid X_2, X_3) = 1/k_{11} = 1/3$, etc.

Consider $X = (X_v, v \in V) \sim \mathcal{N}_V(0, \Sigma)$ with Σ regular and $K = \Sigma^{-1}$.

The concentration matrix of the conditional distribution of (X_α, X_β) given $X_{V\setminus\{\alpha,\beta\}}$ is

$$K_{\{\alpha,\beta\}} = \begin{pmatrix} k_{\alpha\alpha} & k_{\alpha\beta} \\ k_{\beta\alpha} & k_{\beta\beta} \end{pmatrix},$$

Hence

$$\alpha \perp\!\!\!\perp \beta \mid V \setminus \{\alpha, \beta\} \iff k_{\alpha\beta} = 0.$$

Thus *a regular Gaussian distribution is pairwise, local, and global Markov w.r.t. the graph $\mathcal{G}(K)$* given by

$$\alpha \not\sim \beta \iff k_{\alpha\beta} = 0.$$

$\mathcal{S}(\mathcal{G})$ denotes the symmetric matrices A with $a_{\alpha\beta} = 0$ unless $\alpha \sim \beta$ and $\mathcal{S}^+(\mathcal{G})$ their positive definite elements.

A *Gaussian graphical model* for X specifies X as multivariate normal with $K \in \mathcal{S}^+(\mathcal{G})$ and otherwise unknown.

Note that the density then factorizes as

$$\log f(x) = \text{constant} - \frac{1}{2}\sum_{\alpha \in V} k_{\alpha\alpha} x_\alpha^2 - \sum_{\{\alpha,\beta\}\in E} k_{\alpha\beta} x_\alpha x_\beta,$$

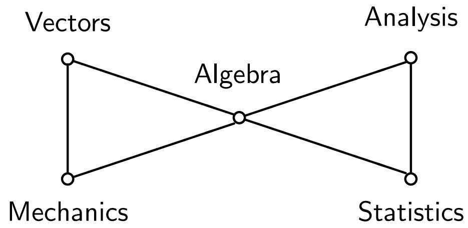
hence *no interaction terms involve more than pairs.*.

Mathematics marks

Examination marks of 88 students in 5 different mathematical subjects. The empirical concentrations (on or above diagonal) and partial correlations (below diagonal) are

	Mechanics	Vectors	Algebra	Analysis	Statistics
Mechanics	5.24	−2.44	−2.74	0.01	−0.14
Vectors	0.33	10.43	−4.71	−0.79	−0.17
Algebra	0.23	0.28	26.95	−7.05	−4.70
Analysis	−0.00	0.08	0.43	9.88	−2.02
Statistics	0.02	0.02	0.36	0.25	6.45

Graphical model for mathmarks



This analysis is from Whittaker (1990).

We have $\mathrm{An}, \mathrm{Stats} \perp\!\!\!\perp \mathrm{Mech}, \mathrm{Vec} \mid \mathrm{Alg}$.

UNIVERSITY OF COPENHAGEN DEPARTMENT OF MATHEMATICAL SCIENCES

Lauritzen, S. L. (1996). *Graphical Models*. Clarendon Press, Oxford, United Kingdom.

Studený, M. (2005). *Probabilistic Conditional Independence Structures*. Information Science and Statistics. Springer-Verlag, London.

Whittaker, J. (1990). *Graphical Models in Applied Multivariate Statistics*. John Wiley and Sons, Chichester, United Kingdom.

Steffen Lauritzen — Markov Properties and the Multivariate Gaussian Distribution — Swiss Winterschool 2015 — Lecture 1
Slide 37/37