MATH 829: Introduction to Data Mining and Analysis
Graphical Models I

Dominique Guillot

Departments of Mathematical Sciences
University of Delaware

May 2, 2016

Independence and conditional independence: motivation

We begin with a classical example (Whittaker, 1990):

- We study the examination marks of 88 students in five subjects: mechanics, vectors, algebra, analysis, statistics (Mardia, Kent, and Bibby, 1979).
- Mechanics and vectors were closed books.
- All the remaining exams were open books.

We can examine the results using a stem and leaf plot.

	algebra	statistics
0-		9
10-		45778889
20-	1	012455699
30-	1266677889	0011122333344556677799
40-	01133333455666677778889999999	000000011234445555556679
50-	0000001111223333344455666778899	0113346
60-	000111123455578	11233447888
70-	12	033
80-	0	1111
90-		

Note: Data appears to be normally distributed.

Example (cont.)

We compute the correlation between the grades of the students:

mech	1.0				
vect	0.55	1.0			
alg	0.55	0.61	1.0		
anal	0.41	0.49	0.71	1.0	
stat	0.39	0.44	0.66	0.61	1.0
	mech	vect	alg	anal	stat

- We observe that the grades are positively correlated between subjects (performance of a student across subjects (good or bad) is consistent).

We now examine the inverse correlation matrix:

mech	1.60				
vect	-0.56	1.80			
alg	-0.51	-0.66	3.04		
anal	**0.00**	**-0.15**	-1.11	2.18	
stat	**-0.04**	**-0.04**	-0.86	-0.52	1.92
	mech	vect	alg	anal	stat

Example (cont.)

Interpreting the inverse correlation matrix:

- Diagonal entries $= 1/(1-R^2)$ are related to the proportion of variance explained by regressing the variable on the other variables.
- Off-diagonal entries: proportional to the correlation of pairs of variables, **given the rest of the variables**.

For example, $R^2_{\text{mech}} = (1.60-1)/1.60 = 37.5\%$.
For the off-diagonal entries, we first scale the inverse correlation matrix $\Omega = (\omega_{ij})$ by computing $\frac{\omega_{ij}}{\sqrt{\omega_{ii}\omega_{jj}}}$:

mech	1.0				
vect	-0.33	1.0			
alg	-0.23	-0.28	1.0		
anal	**0.00**	**-0.08**	-0.43	1.0	
stat	**-0.02**	**-0.02**	-0.36	-0.25	1.0
	mech	vect	alg	anal	stat

The off-diagonal entries of the scaled inverse correlation matrix are the **negative of the conditional correlation coefficients** (i.e., the correlation coefficients after conditioning on the rest of the variables).

Example (cont.)

Notation:

- We denote the fact that two random variables X and Y are independent by $X \perp\!\!\!\perp Y$.
- We write $X \perp\!\!\!\perp Y | \{Z_1, \dots, Z_n\}$ when X and Y are independent given $Z_1, \dots, Z_n$.
- When the context is clear (i.e. when working with a fixed collection of random variables $\{X_1, \dots, X_n\}$, we write

$$X_i \perp\!\!\!\perp X_j \mid \text{rest}$$

instead of $X_i \perp\!\!\!\perp X_j | \{X_k : 1 \leq k \leq n, k \neq i, j\}$.

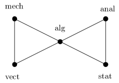
Important: In general, uncorrelated variables are not independent. This is true however for the multivariate Gaussian distribution.

Example (cont.)

We noted before that our data appears to be Gaussian. Therefore it appears that:

1. anal $\perp\!\!\!\perp$ mech | rest.
2. anal $\perp\!\!\!\perp$ vect | rest.
3. stat $\perp\!\!\!\perp$ mech | rest.
4. stat $\perp\!\!\!\perp$ vect | rest.

We represent these relations using a **graph**:



We put **no edge** between two variables iff they are conditionally independent (given the rest of the variables).

Independence and factorizations

Graphical models (a.k.a Markov random fields) are multivariate probability models whose independence structure is characterized by a graph.

Recall: Independence of random vectors is characterized by a factorization of their joint density:

- **Independent variables:** For two random vectors X, Y:

$$X \perp\!\!\!\perp Y \quad \Leftrightarrow \quad f_{X,Y}(x,y) = g(x)h(y) \qquad \forall x, y$$

for some functions g, h.

- **Conditionally independent variables:** Similarly, for three random vectors X, Y, Z:

$$X \perp\!\!\!\perp Y | Z \quad \Leftrightarrow \quad f_{X,Y,Z}(x,y,z) = g(x,z)h(y,z)$$

for all x, y and all z for which $f_Z(z) > 0$.

Independence graphs

Let $X = (X_1, \dots, X_p)$ be a random vector.

- The **conditional independence graph** of X is the graph $G = (V, E)$ where $V = \{1, \dots, p\}$ and

$$(i,j) \notin E \quad \Leftrightarrow \quad X_i \perp\!\!\!\perp X_j \mid \text{rest}.$$

- A subset $S \subset V$ is said to separate $A \subset V$ from $B \subset V$ if every path from A to B contains a vertex in S.

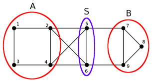
Notation: If $X = (X_1, \dots, X_p)$ and $A \subset \{1, \dots, p\}$, then $X_A := (X_i : i \in A)$.

Theorem: (the separation theorem) Suppose the density of X is positive and continuous. Let $V = A \cup S \cup B$ be a partition of V such that S separates A from B. Then

$$X_A \perp\!\!\!\perp X_B \mid X_S.$$

Independence graphs (cont.)

Example: $X = (X_1, X_2, X_3, X_4, X_5, X_6, X_7, X_8, X_9)$:



Then

$$(X_1, X_2, X_3, X_4) \perp\!\!\!\perp (X_7, X_8, X_9) | (X_5, X_6).$$

Markov properties

Let $X = (X_1, \ldots, X_p)$ be a random vector and let G be a graph on $\{1, \ldots, p\}$. The vector is said to satisfy:

1. The **pairwise Markov property** if $X_i \perp\!\!\!\perp X_j \mid \text{rest}$ whenever $(i,j) \notin E$.
2. The **local Markov property** if for every vertex $i \in V$,
$$X_i \perp\!\!\!\perp X_{V \setminus \text{cl}(i)} | X_{\text{ne}(i)},$$
where $\text{ne}(i) = \{j \in V : (i,j) \in E, j \neq i\}$ and $\text{cl(i)} = \{i\} \cup \text{ne}(i)$.
3. The **global Markov property** if for every disjoint subsets $A, S, B \subset V$ such that S separates A from B in G, we have
$$X_A \perp\!\!\!\perp X_B \mid X_S.$$

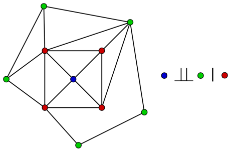
- Clearly, global $\Rightarrow$ local $\Rightarrow$ pairwise.
- When X has a positive and continuous density, by the separation theorem,

pairwise $\Rightarrow$ global

and so all three properties are equivalent.

Example: the local Markov property

Illustration of the local Markov property:



The Hammersley–Clifford theorem

- An **undirected graphical model** (a.k.a. Markov random field) is a set of random variables satisfying a Markov property.
- Independence and conditional independence correspond to a factorization of the joint density.
- It is natural to try to characterize Markov properties via a factorization of the joint density.
- The Hammersley–Clifford theorem provides a necessary and sufficient condition for a random vector to have a Markov random field structure.

Theorem:(Hammersley–Clifford) Let X be a random vector with a positive and continuous density f. Then X satisfies the *pairwise Markov property* with respect to a graph G if and only if

$$f(x) = \prod_{C \in \mathcal{C}} \psi_C(x_C),$$

where $\mathcal{C}$ is the set of (maximal) cliques (complete subgraphs) of G.