

Algorithms for Graphical Models (AGM)

Conditional independence in factored distributions

\$Date: 2008/10/16 09:46:34 \$

AGM-06

In this lecture

- Factored distributions
- Hypergraphs
- Hierarchical models
- Interaction graphs
- The Hammersley-Clifford theorem

Factor multiplication, generally

- We have multiplied factors representing independent random variables.
- But, *since all the factors can be (implicitly) defined on the same table* the ‘broadcasting followed by pointwise multiplication’ algorithm can be applied to multiply *any* two factors.
- Cue `prod_gen` from `gPy.Examples`

Factored distributions

- Given a set of factors $\{f_1, f_2, \dots, f_n\}$ with non-negative values, let $Z = \sum f_1 \times f_2 \times \dots \times f_n$.
- The sum is over all variables, it's the scalar you get by 'marginalising away' all the variables.
- As long as $Z > 0$ the factors define a distribution:
$$P = Z^{-1} f_1 \times f_2 \times \dots \times f_n$$

Representing factored distributions

- Have $P = Z^{-1} f_1 \times f_2 \times \cdots \times f_n$
- So $P = f_1 \times f_2 \times \cdots \times (f_i/Z) \times \cdots \times f_n$ for any i
- Cue norming from gPy.Examples
- Computing Z (the *partition function*) is a pain.
- If we only need probabilities up to common factor, better to represent P (implicitly) by $\{f_1, f_2, \dots, f_n\}$

Independence properties of factored distributions

- If $P = f_1 \times f_2 \times \cdots \times f_n$ what can we deduce about the independence properties of P ?
- If each f_i is a univariate factor for a distinct variable, then as we have seen, each variable is independent.
- But what about the general case?

Conditional distributions

- Let X and Z be subsets of variables of joint distribution P .
- Let $P(X, Z)$ denote the distribution produced by projecting P down onto $X \cup Z$ by marginalisation. Similarly $P(Z)$ is the marginal distribution just on Z .
- The distribution on X *conditional on* Z is $P(X|Z) = \frac{P(X,Z)}{P(Z)}$.
- $P(X|Z)$ contains conditional distributions $P(X|Z = z)$, one for each instantiation z of Z .
- $P(X|Z = z)$ is a distribution for X , *given that* $Z = z$.

Computing conditional distributions

- If $P(X, Z)$ and $P(Z)$ are both represented by factors, then $P(X|Z)$ can be computed by pointwise *division*.
- $P(X|Z)$ is undefined if there's division by zero.
- Cue cond from gPy.Examples

Conditional independence

Let X and Z be subsets of variables of joint distribution P .

X is conditionally independent of Y given Z under P if:

$$P(X, Y|Z) = P(X|Z)P(Y|Z)$$

Write this as: $X \perp Y|Z[P]$ or just: $X \perp Y|Z$ when the P is obvious.

“ X and Y are independent once we know Z .”

It turns out that factored distributions have *conditional* independence properties which depend on their *structure*.

The structure of a factored distribution

- What matters is not so much the numbers in the factors but the relationship between the *variables* in the various factors.
- Consider $\mathcal{H} = \{\text{set of variables used by } f_i\}_i$
- So in the example from `norming`
 $\mathcal{H} = \{\{A, B\}, \{B, C\}, \{C, D\}, \{A, D\}\}$.
- \mathcal{H} is a *hypergraph*.

Hypergraphs

- A hypergraph \mathcal{H} is a set of subsets of a finite set H , the *base set*.
- The elements $h \in \mathcal{H}$ are called the *hyperedges*.
- We will only consider hypergraphs where $H = \cup_{h \in \mathcal{H}} h$
- $\text{red}(\mathcal{H})$ is the *reduced hypergraph* produced by removing all hyperedges in \mathcal{H} that are contained in some other hyperedge.
- The hyperedges that are left after reduction are called a *generating class*.

Reducing hypergraphs

- $\mathcal{H} = \{\{A, B\}, \{B, C\}, \{C, D\}, \{A, D\}\}$ is already reduced.
- Let $\mathcal{H}' = \{\{\text{Smoking}\}, \{\text{Cancer, Smoking}\}, \{\text{Bronchitis, Smoking}\}\}$,
- $\text{red}(\mathcal{H}') = \{\{\text{Cancer, Smoking}\}, \{\text{Bronchitis, Smoking}\}\}$
- Redundant hyperedges correspond to redundant factors.
- Cue `redund` from `gPy.Examples`

Hierarchical models

- A generating class (i.e. a reduced hypergraph) $\mathcal{H} = \{h_1, h_2, \dots, h_n\}$ defines a set of factored probability distributions.
- It is simply the set of probability distributions of the form: $f_1 \times f_2 \times \dots \times f_n$ where $h_i =$ set of variables used by f_i
- Such a set of probability distributions is called a *hierarchical model*.
- So each generating class defines a hierarchical model.

(Undirected) Graphs

- A graph is a pair (V, E) where V is a set of vertices and E a set of edges. $E \subset V \times V$.
- If (α, β) is an edge and so is (β, α) then it is an *undirected edge*.
- Write $\alpha \sim \beta$ if there is an undirected edge between α and β .
- If all edges in a graph are undirected then call the graph undirected.
- The great thing about graphs is that you can see them!

To each hypergraph an undirected graph

Each hypergraph \mathcal{H} has an associated undirected graph $\mathcal{H}_{[2]} = (V, E)$ where:

- $V = H$
- $(\alpha, \beta) \in E \Leftrightarrow \{\alpha, \beta\} \subset h$, for some $h \in \mathcal{H}$

$\mathcal{H}_{[2]}$ is called the *2-section* of \mathcal{H} . If \mathcal{H} is the hypergraph for a factored distribution then $\mathcal{H}_{[2]}$ is the *interaction graph* for the distribution. Cue igs from gPy.Examples

To each undirected graph a hypergraph

- A set of vertices $A \subseteq V$ is *complete* if for all distinct $\alpha, \beta \in A$:
 $\alpha \sim \beta$.
- A set of vertices C is a *clique* if it is *maximally* complete.
- That is, it is complete and adding any new vertex to C would produce a set which is not complete.
- For any graph \mathcal{G} , the set of all its cliques $\mathcal{C}(\mathcal{G})$ is the *clique hypergraph* of the graph.

Factorising according to a graph

- Let \mathcal{G} be a graph whose vertices are the variables of a joint distribution P .
- \mathcal{G} has a clique hypergraph $\mathcal{C}(\mathcal{G})$. Each clique $c \in \mathcal{C}(\mathcal{G})$ is a subset of the variables in the joint distribution P .

The distribution P is said to *factorise according to the graph \mathcal{G}* if there are factors f_c such that:

$$P = \prod_{c \in \mathcal{C}(\mathcal{G})} f_c$$

where each factor f_c only depends on the variables in c .

Factorising according to interaction graphs

- Recall that each factored distribution has an associated hypergraph and interaction graph.
- Because the interaction graph is generated from the distribution's hypergraph, it is easy to see that *a factored distribution always factorises according to its own interaction graph.*
- But why should we care about factorising according to a graph?

The global Markov property

A joint distribution P over variables V obeys the *global Markov property* relative to a graph $\mathcal{G} = (V, E)$, if for any triple (A, B, S) of disjoint subsets of V such that S separates A from B in \mathcal{G} , we have:

$$A \perp B | S [P]$$

So if the global Markov property obtains we can use the graph to ‘read off’ conditional independence relations.

Factorisation implies global Markov property

- You guessed it: if P factorises according to \mathcal{G} , then P obeys the global Markov property relative to \mathcal{G} .
- This means a distribution's interaction graph can be used to read off its conditional independence relations.
- Cue igs again.

The Hammersley-Clifford theorem

The Hammersley-Clifford theorem: If a distribution P is everywhere positive then it factorises according to a graph \mathcal{G} if and only if it obeys the global Markov property relative to \mathcal{G} .

- If there are zeroes in P then we can have the global Markov property without factorisation holding.
- This is a simplified presentation: there are other Markov properties apart from the global one—they are equivalent if P is everywhere positive.