Graphical Models: Worksheet 2 MT 2023

Questions will not be marked, but solutions will be provided.

A: Warm Up

A1. Exponential Families. For each of the following, show that the family of distributions is an exponential family, and find the: (i) canonical and mean parameters; (ii) sufficient statistics; (iii) maximum likelihood estimate; (iv) cumulant function.

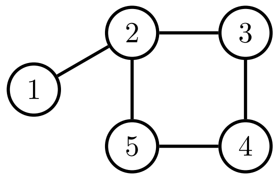
(a) The set of Binomial(n, p) distributions, with n fixed.

(b) The set of Gamma distributions with parameters (a, b).

For (a), show directly that the derivatives of the cumulant function give the first two centred moments of the sufficient statistics (see Lemma 3.1).

A2. Graphical Separation. Consider the graph below. List all the independences implied by the pairwise Markov property.

Give one conditional independence that follows from the global Markov property but is not already in your list.



B: Core Questions

B1. Markov Properties.

Let $\mathcal{G}$ be a graph, and define the *boundary* of a vertex v by

$$\mathrm{bd}_{\mathcal{G}}(v) \equiv \{w \in V \setminus \{v\} \mid w \sim v\}.$$

A distribution obeys the *local Markov property* with respect to $\mathcal{G}$ if

$$X_v \perp\!\!\!\perp X_{V \setminus (\mathrm{bd}_{\mathcal{G}}(v) \cup \{v\})} \mid X_{\mathrm{bd}_{\mathcal{G}}(v)}, \qquad \forall v \in V.$$

(a) Show that if p obeys the local Markov property then this implies that p obeys the pairwise Markov property.

(b) Show that the global Markov property implies the local Markov property.

(c) Show that, if p is strictly positive and obeys the pairwise Markov property with respect to $\mathcal{G}$, then p also obeys the global Markov property with respect to $\mathcal{G}$.
[Hint: Property 5 of the graphoid axioms and the proof of Theorem 4.10 may be helpful.]

(d) Give an example of a graph in which property 5 is required for the pairwise Markov property to imply the local Markov property. Hence or otherwise find a distribution in which the pairwise property holds with respect to this graph, but the local property does not.

B2. Decomposability

Complete the proof of Theorem 4.20 from lectures; that is, show that if $\mathcal{G}$ is an undirected graph, (iii) the fact that every minimal a, b-separator is complete implies that (iv) the cliques satisfy the running intersection property starting with a given C; and that (iv) implies (i): the graph is decomposable.

Does decomposability imply that every minimal A, B-separator is complete, for sets A and B?

B3. Whittaker Data

Using R, load the Whittaker data from lectures with the commands:

```
> library(ggm)
> data(marks)
> head(marks, 8)  # inspect the first few
> solve(cov(marks))  # empirical concentration matrix
```

(if not already installed, you may have to call `install.packages("ggm")`, to use the `ggm` package).

(a) Manually find the MLE for the covariance matrix Σ, under the model from lectures in which 'analysis' and 'statistics' are independent of 'mechanics' and 'vectors' conditional on 'algebra'.

[Hint: R commands you might need are `solve()`, which inverts matrices, and the use of square brackets `[]` for subsetting. See the MSc R Programming lecture notes for details.]

(b) Suppose we have i.i.d. observations $x_V^{(1)}, \dots, x_V^{(n)}$ from a multivariate Gaussian with known mean μ and unknown covariance Σ. Show that the log-likelihood for Σ can be written as

$$l(\Sigma; X) = -\frac{n}{2} \log |\Sigma| - \frac{n}{2} \operatorname{tr}(S\Sigma^{-1}).$$

where $S = \frac{1}{n} \sum_{i=1}^{n} (x_V^{(i)} - \mu)(x_V^{(i)} - \mu)^T$ and tr() is the trace operator.

[Hint: $\operatorname{tr}(AB) = \operatorname{tr}(BA)$]

(c) Hence carry out a likelihood ratio test to see whether this model is a good fit to the data.

B4. Iterative Proportional Fitting

(a) Show that the iterative proportional fitting algorithm does not decrease the likelihood at any iteration. Argue also that, if the likelihood does not strictly increase after a full cycle of updates then the algorithm has converged to a solution.

Consider a 7-dimensional table, and suppose that we have an undirected model based on the cliques $\{1, 2, 3\}$, $\{1, 2, 4\}$, $\{2, 3, 5\}$, $\{1, 3, 6\}$ and $\{5, 7\}$.

(b) Show that this model is decomposable and that if the IPF algorithm is run in the order given above, it will return the MLE after a single iteration of updating each clique in turn.

(c) Is the same true if we choose the order $\{1, 2, 4\}$, $\{2, 3, 5\}$, $\{1, 3, 6\}$, $\{5, 7\}$ and $\{1, 2, 3\}$?

C: Optional

C1. Marginal Models

Let $\mathcal{G}$ be a graph containing a path $\pi : i - k_1 - \cdots - k_m - j$, for $m \geq 0$.

(a) Construct a distribution p that factorizes according to $\mathcal{G}$, and such that for any set $C \subseteq V \setminus \{i, j, k_1, \dots, k_m\}$ we have $X_i \not\perp\!\!\!\perp X_j \mid X_C \, [p]$.
[Hint: remember that undirected graphs generalize Markov chain models.]

(b) Deduce that for any graph $\mathcal{G}$ and sets A, B, S such that $A \not\perp_s B \mid S$ in $\mathcal{G}$, there exists a distribution p which factorizes according to $\mathcal{G}$ and for which $X_A \not\perp\!\!\!\perp X_B \mid X_S$ in p. (In this sense the global Markov property is *complete*; separation represents **all** the independences guaranteed by factorization.)

Given an undirected graph $\mathcal{G}$ and subset of vertices $W \subseteq V$, define $\mathcal{G}|^W$ as the undirected graph with vertex set W, and an edge $i - j$ if and only if there is a path from i to j in $\mathcal{G}$ with all intermediate vertices in $V \setminus W$. [Note that this is quite different to the induced subgraph $\mathcal{G}_W$.]

(c) Let $p(x_V)$ be globally Markov with respect to $\mathcal{G}$. Show that $p(x_W) = \sum_{x_{V \setminus W}} p(x_V)$ is globally Markov with respect to $\mathcal{G}|^W$.

(d) Show further that, in general, $p(x_W)$ is **not** globally Markov with respect to any edge subgraph of $\mathcal{G}|^W$.

C2. Hierarchical Models.

Let $\mathcal{C}$ be a collection of non-empty subsets of a set V, such that:

- $\bigcup_{C \in \mathcal{C}} C = V$;
- for any distinct $C, D \in \mathcal{C}$ we have $C \not\subseteq D$.

In other words, this is a set of inclusion maximal subsets. We call $\mathcal{C}$ a *generating class*.

(a) Show that the cliques of a graph are a generating class.

(b) List, up to symmetry, all the generating classes on the set $V = \{1, 2, 3\}$. Do all generating classes correspond to the cliques of a graph?

Given a generating class $\mathcal{C}$, we can define a corresponding log-linear model by requiring that $\lambda_A = 0$ whenever A is not a subset of any element of $\mathcal{C}$. Such models are called *hierarchical*.

(c) Show that the counts $n(x_C)$ for $C \in \mathcal{C}$ are sufficient statistics for this model.

The data below consist of answers from high schoolers to a Dayton, Ohio survey on substance use.

Alcohol	Tobacco	Marijuana Yes	Marijuana No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

They are available in a text file, `substance.txt`, on the class website. After downloading the file, you can read these data into R using

```
> dat <- read.table("substance.txt", header=TRUE)
```

(d) Using `glm()` with `family=poisson`, fit a hierarchical model to these data with the generating class $\mathcal{C} = \{\{A, M\}, \{T, M\}, \{A, T\}\}$. Is it a good fit? Is any smaller hierarchical model a good fit?

(e) Verify that the fitted distribution has the same sufficient statistics as the data.

(f) Try adding the vector $(+1, -1, -1, +1, -1, +1, +1, -1)$ to your counts. Verify that the parameter estimates are unchanged with this 'new data'. Can you explain why?