2023-2024 Graphical Model Notes

Chapter 3 Exponential Families and Contingency Tables

- Denote $X_V \equiv (X_v : v \in V)$, indexed by $V = \{1, \cdots, p\}$. Each X_v takes values in the set $\mathscr{X}_v$. For a subset $A \subseteq V$, we write X_A to denote $(X_v : v \in A)$.

- Let $p(\cdot;\theta)$ be a collection of probability densities over $\mathscr{X}$ indexed by $\theta \in \Theta$. We say that p is an exponential family if it can be written as

$$\begin{aligned} p(x;\theta) &= \exp\left\{\sum_i \theta_i \phi_i(x) - A(\theta) - C(x)\right\} \\ &= \exp\{<\theta, \phi(x)> - A(\theta) - C(x)\} \end{aligned}$$

  - The family is said to be regular if Θ is a non-empty open set.
  - The functions ϕ_i are the sufficient statistics.
  - The components θ_i are the canonical/natural parameters.
  - The function $A(\theta)$ is the cumulant function such that the distribution normalises:

$$A(\theta) = \log \int \exp\{<\theta, \phi(x)> - C(x)\}\,\mathrm{d}x$$

  - The function $Z(\theta) \equiv e^{A(\theta)}$ is the partition function.

- We have

$$\nabla A(\theta) = \mathbb{E}_\theta \phi(X) \qquad \nabla\nabla^{\mathrm{T}} A(\theta) = \mathrm{Cov}_\theta \phi(X)$$

  $A(\theta)$ and $-\log p(x;\theta)$ are convex in θ, and the map $\mu(\theta) : \theta \mapsto \nabla A(\theta)$ is bijective, named the mean function. We care about convexity because it does not have multiple local minima, which in turn facilitates computation.

- Let $X_V = (X_1, \cdots, X_p)^{\mathrm{T}} \in \mathbb{R}^p$ be a random vector. Let $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p\times p}$ be a positive definite symmetric matrix. We say that X_V has a multivariate Gaussian distribution with μ and Σ if the joint density is

$$\begin{aligned} f(x_V) &= \frac{1}{(2\pi)^{\frac{p}{2}}|\Sigma|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(x_V - \mu)^{\mathrm{T}}\Sigma^{-1}(x_V - \mu)\right\} \\ &= \frac{1}{(2\pi)^{\frac{p}{2}}} \exp\left\{-\frac{1}{2}x_V^{\mathrm{T}} K x_V + \mu^{\mathrm{T}} K x_V - \frac{1}{2}\mu^{\mathrm{T}} K \mu + \frac{1}{2}\log|K|\right\} \qquad x_V \in \mathbb{R}^p \end{aligned}$$

  Here, $K \equiv \Sigma^{-1}$ is called the concentration matrix. Let

$$\phi(x_V) = \left(x_v, -\frac{1}{2}x_V x_V^{\mathrm{T}}\right) \qquad \theta = (K\mu, K)$$

  we could easily tell that the multivariate Gaussian distribution is an exponential family[1].

[1] For two matrices A and B, we have $<A, B> = \mathrm{tr}(A, B^{\mathrm{T}})$.

- Let X_V have a multivariate Gaussian distribution with concentration matrix $K = \Sigma^{-1}$, then $X_i \perp\!\!\!\perp X_j \mid X_{V\setminus\{i,j\}}$ iff $k_{ij} = 0$.

- Let $X_V^{(i)} = (X_1^{(1)}, \cdots, X_p^{(i)})$ be sampled over individuals $i = 1, \cdots, n$ and define

$$n(x_V) \equiv \sum_{i=1}^{n} \mathbb{1}\{X_1^{(i)} = x_1, \cdots, X_p^{(i)} = x_p\}$$

  the number of individuals who have the response pattern x_V. These counts are the sufficient statistics for the multinomial model with log-likelihood

$$\begin{aligned} l(p;n) &= \sum_{x_V} n(x_V) \log p(x_V) \\ &= \sum_{x_V \neq 0_V} n(x_V) \log \frac{p(x_V)}{p(0_V)} + n \log p(0_V) \end{aligned}$$

  where 0_V is the vector of zeros, $p(x_V) \geq 0$, and $\sum_{x_V} p(x_V) = 1$. The multinomial distribution is also an exponential family with

  - Sufficient statistics given by $n(x_V)$.
  - Canonical parameters given by $\log \dfrac{p(x_V)}{p(0_V)}$.
  - Convex cumulant function given by

$$-\log p(0_V) = \log\left(1 + \sum_{x_V \neq 0_V} e^{\theta(x_V)}\right)$$

  Each possibility x_V is called a cell of the table. Given $A \subseteq V$,

$$n(x_A) \equiv \sum_{x_B} n(x_A, x_B)$$

  where $B = V \setminus A$ is called the marginal table.

- The log-linear parameters for $p(x_V) > 0$ are defined by the relation

$$\begin{aligned} \log p(x_V) &= \sum_{A \subseteq V} \lambda_A(x_A) \\ &= \lambda_\varnothing + \lambda_1(x_1) + \cdots + \lambda_V(x_V) \end{aligned}$$

  and the identifiability constraint $\lambda_A(x_A) = 0$ whenever $x_a = 1$ for some $a \in A$.

- Consider a 2×2 contingency table with probabilities π_{ij}. The log-linear parametrisation has

$$\begin{aligned} \log \pi_{11} &= \lambda_\varnothing & \log \pi_{21} &= \lambda_\varnothing + \lambda_X \\ \log \pi_{12} &= \lambda_\varnothing + \lambda_Y & \log \pi_{22} &= \lambda_\varnothing + \lambda_X + \lambda_Y + \lambda_{XY} \end{aligned}$$

We can deduce that

$$\lambda_{XY} = \log \frac{\pi_{11}\pi_{22}}{\pi_{21}\pi_{12}}$$

and $e^{\lambda_{XY}}$ is called the odds ratio between X and Y.

- Let $X_i \sim \mathrm{P}(\mu_i)$ independently, and let $N = \sum_{i=1}^{k} X_i$. Then,

$$N \sim \mathrm{P}(\sum_i \mu_i)$$

$$(X_1, \cdots, X_k)^{\mathrm{T}} \mid N = n \sim \mathrm{Multinom}(n, (\pi_1, \cdots, \pi_k)^{\mathrm{T}})$$

where $\pi_i = \dfrac{\mu_i}{\sum_j \mu_j}$.

- Let $p > 0$ be a discrete distribution on X_V with associated log-linear parameters $\lambda_C, C \subseteq V$. The conditional independence $X_a \perp\!\!\!\perp X_b \mid X_{V \setminus \{a,b\}}$ holds if and only if $\lambda_C = 0$ for all $\{a,b\} \subseteq C \subseteq V$.

Chapter 4 Undirected Graphical Models

- Let V be a finite set. An undirected graph $\mathscr{G}$ is a pair (V, E) where
  - V are the vertices/nodes.
  - $E \subseteq \{\{i, j\} : i, j \in V, i \neq j\}$ is a set of unordered distinct pairs of V called edges.

  We represent graphs by drawing the vertices and then joining pairs of vertices by a line if there is an edge between them.

  - We write $i \sim j$ if $\{i, j\} \in E$, and say they are adjacent in the graph. The vertices adjacent to i are called the neighbours of i, and the set of neighbours is often called the boundary of i and denoted by $\mathrm{bd}_{\mathscr{G}}(i)$.
  - A path in a graph is a sequence of adjacent vertices without repetition. The length of a path is the number of edges in it.
  - Given a subset of vertices $W \subseteq V$, we define the induced subgraph $\mathscr{G}_W$ of $\mathscr{G}$ to be the graph with vertices W, and all edges from $\mathscr{G}$ whose endpoints are contained in W.
  - We say $C \subseteq V$ is complete if $i \sim j$ for every $i, j \in C$. A maximal complete set is called a clique. The set of cliques in a graph is denoted by $\mathscr{C}(\mathscr{G})$.

- Let $A, B, S \subseteq V$. We say that A and B are separated by S in $\mathscr{G}$ ($A \perp_s B \mid S[\mathscr{G}]$) if every path from any $a \in A$ to any $b \in B$ contains at least one vertex in S. A and B are separated by S (where $S \cap A = S \cap B = \varnothing$) iff A and B are separated by $\varnothing$ in $\mathscr{G}_{V \setminus S}$.

- Let $\mathscr{G}$ be a graph with vertices V, and let p be a probability distribution over the random variables X_V. We say that p satisfies the pairwise Markov property for $\mathscr{G}$ if

$$i \not\sim j \text{in} \mathscr{G} \implies X_i \perp\!\!\!\perp X_j \mid X_{V \setminus \{i,j\}} [p]$$

We say that p satisfies the global Markov property for $\mathscr{G}$ if for any disjoint sets A, B, S

$$A \perp_s B \mid S \text{in} \mathscr{G} \implies X_A \perp\!\!\!\perp X_B \mid X_S [p]$$