Graphical Models: Worksheet 0 MT 2023

This sheet is designed to revise some bits from previous courses that will be particularly useful for Graphical Models.

1. **Sufficient Statistics.**

   Let $X_i \sim \text{Binom}(n_i, \theta)$, $i = 1, \dots, k$ be independent binomial random variables with known sizes n_i, and unknown $\theta \in [0, 1]$.

   (a) Write down the log-likelihood for θ, and find a sufficient statistic.

   The log-likelihood is

   $$\begin{aligned} l(\theta) &= \left(\sum_i X_i\right) \log\theta + \left(\sum_i \{n_i - X_i\}\right) \log(1-\theta) \\ &= r\log\theta + (n-r)\log(1-\theta) \end{aligned}$$

   where $n = \sum_i n_i$ *and* $r = \sum_i X_i$. *Hence* r *is a (minimal) sufficient statistic.*

   (b) Find the MLE for θ and its asymptotic distribution.

   Differentiating we find that the MLE is $\hat\theta = r/n$, *and taking the second derivative we get*

   $$\begin{aligned} l''(\theta) &= -\frac{r}{\theta^2} - \frac{n-r}{(1-\theta)^2} \\ I_n(\theta) = -\mathbb{E}l''(\theta) &= \frac{n}{\theta} + \frac{n}{(1-\theta)} = \frac{n}{\theta(1-\theta)} \end{aligned}$$

   since $\mathbb{E}r = n\theta$. *Hence by standard asymptotic results,*

   $$\sqrt{n}(\hat\theta - \theta) \approx N(0, \theta(1-\theta)).$$

   (c) What would constitute a conjugate prior for θ?

   We want something that keeps the form $\theta^x(1-\theta)^{n-x}$, *so something like* $\pi(\theta) \propto \theta^{a-1}(1-\theta)^{b-1}$ *would work. You might recognise this as a Beta distribution.*

   (d) Suppose you have $n_1 = n_2 = 100$, and data $X_1 = 48$, $X_2 = 52$. Build a confidence interval for θ using your answer to (b).

   The MLE is (48+52)/200 = 0.5, and confidence interval will be $0.5 \pm 1.96\sqrt{0.5(1-0.5)/200}$.

   (e) Now suppose $X_1 = 10$ and $X_2 = 90$. How does your answer differ?

   The answer is the same, since the sufficient statistic r *(and* n*) is the same. However, the model is clearly inappropriate, since it's extremely unlikely we would observe 10 or 90 from the same* $\text{Binom}(100, \theta)$ *distribution, regardless of the value of* θ.

2. **Conditional Distributions.**

   Suppose that X, W are independent Exponential(λ) random variables. Define $Y = X + W$. Find the joint density of X and Y. Are X and Y independent?

   The joint density is

   $$f_{XY}(x, y) = \begin{cases} \lambda^2 e^{-\lambda y} & \text{if } y > x > 0, \\ 0 & \text{otherwise} \end{cases}.$$

Note that the expression within the valid range for x, y factorizes, so when performing the usual change of variables one may mistakenly conclude that X and Y are independent. They are clearly dependent, since in particular $Y > X$ with probability 1.

Find the conditional density of X given Y.

This is just proportional to the joint density, which doesn't depend upon x. Hence X must be uniform on its valid range $[0, Y]$. So $X|Y \sim \text{Unif}[0, Y]$.

3. **Conditional Events**

   Let X, Y, and Z be discrete random variables taking values in the sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$.

   (a) Write down and briefly justify the law of total probability for discrete random variables X and Y.

   This is $\sum_{y'} P(X = x \mid Y = y')P(Y = y') = P(X = x)$. To prove, note that the sum is just $\sum_{y'} P(X = x, Y = y')$ by definition, and since the sum is over all states of Y, it is clearly $P(X = x)$.

   (b) Prove Bayes' Formula:

   $$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y) \cdot P(Y = y)}{\sum_{y' \in \mathcal{Y}} P(X = x \mid Y = y') \cdot P(Y = y')}.$$

   Using the definition of conditional probability twice, we get

   $$P(Y = y \mid X = x) = \frac{P(X = x \mid Y = y) \cdot P(Y = y)}{P(X = x)}.$$

   Applying the law of total probability to $P(X = x)$ gives the result.

   (c) Express $P(Z = z)$ in terms of probabilities of the form $P(X = x), P(Y = y \mid X = x), P(Z = z \mid X = x, Y = y)$. In terms of the sizes of the sets $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$, how many calculations (additions, subtractions, multiplications, divisions) are required to evaluate it for all $z \in \mathcal{Z}$?

   $$\begin{aligned} P(Z = z) &= \sum_{x,y} P(X = x)P(Y = y \mid X = x)P(Z = z \mid X = x, Y = y) \\ &= \sum_{x} P(X = x) \sum_{y} P(Y = y \mid X = x)P(Z = z \mid X = x, Y = y). \end{aligned}$$

   The inner sum requires $|\mathcal{X}| \cdot |\mathcal{Y}| \cdot |\mathcal{Z}|$ multiplications and then $|\mathcal{X}||\mathcal{Z}|(|\mathcal{Y}| - 1)$ summations; the outer $|\mathcal{X}| \cdot |\mathcal{Z}|$ multiplications and $(|\mathcal{X}| - 1)|\mathcal{Z}|$ summations. This ends up being $O(|\mathcal{X}||\mathcal{Y}||\mathcal{Z}|)$ operations.

   (d) What difference does it make if $P(Z = z \mid X = x, Y = y) = P(Z = z \mid Y = y)$?

   In this case we can write

   $$\begin{aligned} &\sum_{x,y} P(X = x)P(Y = y \mid X = x)P(Z = z \mid Y = y) \\ &= \sum_{y} P(Z = z \mid Y = y) \sum_{x} P(Y = y \mid X = x)P(X = x) \end{aligned}$$

   the inner sum can be done in $O(|\mathcal{Y}||\mathcal{X}|)$ operations and the outer in $O(|\mathcal{Y}||\mathcal{Z}|)$. Hence only $O(|\mathcal{Y}|(|\mathcal{X}| + |\mathcal{Z}|))$ operations.

4. **Contingency Tables.**

Let (X_i, Y_i, Z_i), $i = 1, \dots, n$ be i.i.d. vectors of categorical variables such that $P(X = x, Y = y, Z = z) = \pi_{xyz}$. Define

$$n_{xyz} = \sum_{i=1}^{n} \mathbb{1}\{X_i = x, Y_i = y, Z_i = z\}.$$

The array $(n_{xyz})_{x,y,z}$ is called a *contingency table* (see Part A Stats).

(a) Write down the likelihood for $\boldsymbol{\pi} = (\pi_{xyz})_{x,y,z}$.

Just as with any multinomial form, we get

$$l(\boldsymbol{\pi}) = \sum_{x,y,z} n_{xyz} \log \pi_{xyz}, \qquad \pi_{xyz} \geq 0, \sum_{x,y,z} \pi_{xyz} = 1.$$

(b) We say X is *conditionally independent* of Y given Z if we can write

$$P(X = x, Y = y, Z = z) = P(Z = z) \cdot P(X = x \mid Z = z) \cdot P(Y = y \mid Z = z)$$

for all x, y, z. Show that, the MLE of $\boldsymbol{\pi}$ under this restriction is

$$\hat{\pi}_{xyz} = \frac{n_{x+z} \cdot n_{+yz}}{n_{++z} \cdot n},$$

where, for example, $n_{x+z} = \sum_y n_{xyz}$. *[Hint: this is similar to the two-dimensional independence case from Part A stats.]*

Writing $\pi_{xyz} = r_z s_{x|z} t_{y|z}$ *we can write the log-likelihood as*

$$\begin{aligned} l(\boldsymbol{\pi}) &= \sum_{x,y,z} n_{xyz} \log r_z s_{x|z} t_{y|z} \\ &= \sum_z n_{++z} \log r_z + \sum_{x,z} n_{x+z} \log s_{x|z} + \sum_{y,z} n_{+yz} \log t_{y|z}. \end{aligned}$$

Maximizing each term separately (subject to its own summation restrictions) gives $\hat{r}_z = n_{++z}/n$, $\hat{s}_{x|z} = n_{x+z}/n_{++z}$, $\hat{t}_{y|z} = n_{+yz}/n_{++z}$, *and so the result.*

5. **Multivariate Normal Distributions.**

[This is harder, but do-able.]

Let $M = (m_{ij})$ be a $p \times p$-matrix and $C \subseteq \{1, \dots, p\}$; let $D = \{1, \dots, p\} \setminus C$. We say that

$$M_{DD \cdot C} \equiv M_{DD} - M_{DC}(M_{CC})^{-1} M_{CD}$$

is the *Schur complement* of M with respect to C, and its entries are

$$m_{ij \cdot C} \equiv m_{ij} - M_{iC}(M_{CC})^{-1} M_{Cj} \qquad \text{for } i, j \in D.$$

Now let $X_V \sim N_p(\mu, \Sigma)$ have a multivariate normal distribution, meaning that it has Lebesgue density

$$f(x_V; \mu, \Sigma) = \frac{1}{(2\pi)^{p/2} (\det \Sigma)^{1/2}} \exp\left\{ -\frac{1}{2} (x_V - \mu)^T \Sigma^{-1} (x_V - \mu) \right\}, \quad x_V \in \mathbb{R}^p,$$

for $\mu \in \mathbb{R}^p$ and a symmetric positive definite matrix Σ.

(a) Let Σ be partitioned as

$$\Sigma = \begin{pmatrix} \Sigma_{CC} & \Sigma_{CD} \\ \Sigma_{DC} & \Sigma_{DD} \end{pmatrix}$$

with $|C| = p_1$, $|D| = p_2$. Show that

$$\Sigma^{-1} = \begin{pmatrix} \Sigma^{-1}_{CC\cdot D} & -\Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD} \\ -\Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D} & \Sigma^{-1}_{DD} + \Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD} \end{pmatrix}$$

[Note that Σ^{-1}_{DD} means $(\Sigma_{DD})^{-1}$.]

Multiplying the expression given by Σ (partitioned as above) and simplifying gives the result. For example, the first entry is

$$\begin{aligned} \Sigma^{-1}_{CC\cdot D}\Sigma_{CC} - \Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD}\Sigma_{DC} &= \Sigma^{-1}_{CC\cdot D}(\Sigma_{CC} - \Sigma_{CD}\Sigma^{-1}_{DD}\Sigma_{DC}) \\ &= \Sigma^{-1}_{CC\cdot D}\Sigma_{CC\cdot D} = I_{p_1}. \end{aligned}$$

and the final one is

$$\begin{aligned} &-\Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD} + \left(\Sigma^{-1}_{DD} + \Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD}\right)\Sigma_{DD} \\ &= I_{p_2} - \Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD} + \Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD} \\ &= I_{p_2}. \end{aligned}$$

(b) By considering the terms in the density which depend upon x_C, show that

$$X_C \mid X_D = x_D \;\sim N_{p_1}\big(\mu_C + \Sigma_{CD}\Sigma^{-1}_{DD}(x_D - \mu_D),\, \Sigma_{CC\cdot D}\big)\,.$$

where $\Sigma_{CC\cdot D} = \Sigma_{CC} - \Sigma_{CD}\Sigma^{-1}_{DD}\Sigma_{DC}$.

Applying the previous part to the log-pdf of X_V we obtain:

$$\begin{aligned} &\log f(x_V) \\ &= -\frac{1}{2}(x_V - \mu)^T\Sigma^{-1}(x_V - \mu) + \textit{const.} \\ &= \frac{1}{2}(x_C - \mu_C)^T\Sigma^{-1}_{CC\cdot D}(x_C - \mu_C) + (x_C - \mu_C)^T\Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD}(x_D - \mu_D) + \textit{const.} \end{aligned}$$

so completing the square and ignoring terms not depending on x_C we get

$$= \frac{1}{2}(x_C - \mu_{C\cdot D})^T\Sigma^{-1}_{CC\cdot D}(x_C - \mu_{C\cdot D}) + \textit{const.}$$

where $\mu_{C\cdot D} \equiv \mu_C + \Sigma_{CD}\Sigma^{-1}_{DD}(x_D - \mu_D)$. Consequently $X_C \mid X_D$ has the distribution given.

(c) Hence show that the marginal distribution $X_D \sim N_{p_2}(\mu_D, \Sigma_{DD})$.

Recall that

$$f_V(x_V) = f_{C|D}(x_C \mid x_D) \cdot f_D(x_D),$$

so the marginal distribution is whatever is left after dividing by the conditional distribution (subtracting on the log-scale). Close inspection of the term added in to complete the square shows that it is

$$\frac{1}{2}(x_D - \mu_D)^T\Sigma^{-1}_{DD}\Sigma_{DC}\Sigma^{-1}_{CC\cdot D}\Sigma_{CD}\Sigma^{-1}_{DD}(x_D - \mu_D),$$

and that this cancels with one of the two terms resulting from the DD component of Σ^{-1} in the likelihood. The other is $\frac{1}{2}(x_D - \mu_D)^T\Sigma^{-1}_{DD}(x_D - \mu_D)$, so the marginal distribution is as suggested.