SC7 Bayes Methods

First problem sheet (Sections 1-4.2 of lecture notes).

Section A questions (optional)

1. Let $x_1, x_2, x_3, ...$ be an infinite exchangeable sequence of binary random variables with generative model $q \sim F$ and $x_i \sim p(\cdot|q)$ jointly independent for $i \geq 1$ given q. Show that $\text{cov}(x_i, x_j) \geq 0$ for all $i, j \in \{1, 2, 3, ...\}$.

Solution: By exchangeability $E(x_i) = E(x_1)$ for all $i = 1, 2, 3, ...$, and x_i, x_j are independent given q so

$$\begin{aligned}
\text{cov}(x_i, x_j) &= E(x_i x_j) - E(x_i)E(x_j) \\
&= E(E(x_i x_j|q)) - E(x_1)^2 \\
&= E(E(x_i|q)E(x_j|q)) - E(x_1)^2 \\
&= E(E(x_1|q)^2) - E(E(x_1|q))^2 \\
&= \text{var}(E(x_1|q)) \\
&\geq 0
\end{aligned}$$

2. Let Σ be an $n \times n$ covariance matrix and for $x \in \mathbb{R}^n$ let $p_{1:n}(x) = N(x; 0_n, \Sigma)$. Let $\mathcal{O}_n$ be the set of all non-empty subsets of $[n]$ and let $o \in \mathcal{O}_n$ be a set with m elements. For $x_o = (x_{o_1}, \ldots, x_{o_m})$ with $x_o \in \mathbb{R}^m$ let $p_o(x_o) = N(x_o; 0_m, \Sigma_{o,o})$ where 0_m is a vector of m zeros and $\Sigma_{o,o}$ is the sub-matrix of Σ obtained by taking the rows and columns in $(o_1, \ldots, o_m)$. Define what it means for the family of distributions $\{p_o; o \in \mathcal{O}_n\}$ to be marginally consistent and show that this holds. *Hint: it might help to take a look at Section 3.1.3 of our notes.*

Solution: MC holds if

$$p_o(x_o) = \int_{\mathbb{R}^{n-m}} p_{1:n}(x) dx_{[1:n]\setminus o}$$

for each $o \in \mathcal{O}_n$. Since marginals of multivariate normals $N(\mu, \Sigma)$ are simply obtained by dropping the entries in μ and Σ corresponding to the integrated variables, we have

$$\int_{\mathbb{R}^{n-m}} p_{1:n}(x) dx_{[1:n]\setminus o} = N(x_o; 0_m, \Sigma_{o,o})$$

so the family is MC. To verify this briefly, let A be a $m \times n$ matrix of zeros except that $A_{i,o_i} = 1$ for $i = 1, \ldots, m$ and let $z \sim N(0_n, \Sigma)$. Let $z_o = Az$ so we just extract the variables we want. Linear combinations of normals are normal so we just need the mean and covariance of z_o. We have $E(z_o) = 0_m$ and $\text{var}(Az) = A\Sigma A^T = \Sigma_{o,o}$.

Section B questions (hand in solutions to these questions)

3. In the radiocarbon dating example (with the same key question about span), suppose the dated materials are found in layers (strata) piled up on one another, with $y_{i,j}$ the radiocarbon date for $\theta_{i,j}$, the j'th date in the i'th layer. Let $L < \psi_1 < \psi_2 < ... < \psi_M < U$ be the age parameters for the layer boundaries. If we have n_i dates from the ith layer we know that for $i = 1, 2, ..., M-1$, and $j = 1, 2, ..., n_i$, $\psi_i < \theta_{i,j} < \psi_{i+1}$ (so specimen dates in higher layers are not as old as dates in lower layers). Let $\psi = (\psi_1, ..., \psi_M)$ and $\theta = (\theta_1, ...\theta_{M-1})$ with $\theta_i = (\theta_{i,1}, ..., \theta_{i,n_i})$.

   Derive a prior density $\pi(\theta, \psi)$ for the parameters θ, ψ with reference to the prior elicitation checklist given in lectures. Hint: how are the layer boundary dates $\psi_2, ..., \psi_{M-1}$ generated?

4. (From Cox and Hinkley *Theoretical Statistics*) For $i = 1, ..., n$, let $\theta_i \in \{0, 1\}$ be the indicator for the event that student i enjoys the course this year and let $\theta = (\theta_1, ..., \theta_n)$. Suppose our prior probability for $\theta_i = 1,\ i = 1, ..., n$ is that they are iid with $P(\theta_i = 1) = p$ with p our prior probability that an individual student enjoys the course and we take a fixed value of p expressing our prior expectation for the proportion enjoying the course (based perhaps on past years).

   Our prior on the function $q(\theta) = n^{-1} \sum_i \theta_i$ has mean p (that's good) and variance $p(1-p)/n$. If n is large this prior expresses near certainty in the proportion enjoying the course. This doesn't represent prior knowledge.

   Modify the prior model so that the variance doesn't go to zero as $n \to \infty$ whilst retaining $E(q(\theta)) = p$. *Hint: your solution will be a hierarchical model of some sort, but beyond that there are many correct answers.*

5. Let X_1, X_2 be binary random variables. Table entries below give probabilities, $p(x_1, x_2) = \Pr(X_1 = x_1, X_2 = x_2)$, for outcomes $(X_1, X_2) = (x_1, x_2)$ indicated by row and column.[1]

	$X_1 = 0$	$X_1 = 1$
$X_2 = 0$	0	1/2
$X_2 = 1$	1/2	0

   (a) Show that X_1 and X_2 are exchangeable.

[1] From P. Diaconis and Freedman D. (1980). *Finite Exchangeable Sequences.* Ann. Probab. v8 p745–764.

(b) Show that there does not exist a distribution F such that

$$p(x_1, x_2) = \int_0^1 \prod_{i=1,2} p^{x_i}(1-p)^{1-x_i} dF(p),$$

ie, de Finetti's theorem need not hold if the exchangeable sequence is finite.

6. Consider two urns. In the first urn there are 50 black balls and 50 red balls. In the second urn there are 100 balls, the number of each color unknown. Suppose the proportion of black balls in the second urn is equal ϕ.

Jane's ϕ-prior, $\pi(\phi)$, satisfies $E(\phi) = 1/2$. Jane is offered a choice of urn and color and *two* balls are drawn (with replacement) from the chosen urn. Jane receives a £1 reward for each ball matching her chosen color. Her utility function is $U(0) = 0, U(1) = v, U(2) = 1$ with $1/2 < v < 1$.

Jane is offered red from the first urn or black from the second.

(a) Show that the expected utility of choosing the second urn given ϕ is

$$E(U|\phi) = 2\phi(1-\phi)v + \phi^2.$$

(b) Jane chooses the first urn. Show that this choice maximises the expected utility.

(c) Jane is now offered black from the first urn or red from the second. Show that Jane should again choose the first urn.

Section C questions (optional)

7. An 18×24 square-foot area of grass in a field was split into 3×3 square-foot cells (so $n = 48$). An indicator for army-worm damage was recorded for the centre square in each cell. When the grass was planted the plough and harrows ran down the columns.

	Easting							
Northing	1	2	3	4	5	6	7	8
1	1	1	1	0	0	0	0	1
2	1	1	1	0	0	1	0	1
3	0	1	1	1	1	0	1	1
4	1	1	1	0	0	1	1	1
5	1	1	0	0	0	0	0	1
6	0	0	1	1	0	1	1	1

For $i = 1, ..., n$, let $Y_i \in \{0, 1\}$ indicate army-worm damage observed in cell i. The data are displayed above. Suppose $Y_i \sim$ Bernoulli(μ_i) independently in each cell, with

$$\mu_i = \frac{\exp(z_i)}{1 + \exp(z_i)}.$$

where $z_i \in \Re$ is an unobserved real latent variable associated with cell i. Let $z = (z_1, ..., z_n)$. The z-values are modelled by an auto-regression with an $n \times n$ weight matrix W

$$z = (I - W)^{-1}\epsilon, \quad \epsilon \sim N(0, \sigma^2 I_n)$$

It is of interest to estimate a map of the log-odds for damage, and to test for an effect due to ploughing. Assume a prior for σ is available and consider eliciting a prior for z by specifying W.

(a) Suppose that for $i = 1, ..., n$, $W_{ii} = 0$. Let $\pi(z_i|z_{-i}, \sigma)$ denote the conditional density of z_i given the other z-values and σ. Show that

$$p(z_i|z_{-i}, \sigma) = N(z_i; \textstyle\sum_{j \neq i} W_{i,j} z_j, \sigma^2).$$

Solution: Since $z - Wz = \epsilon$ and $W_{ii} = 0$, we have for each $i = 1, ..., n$ that $z_i = \sum_{j \neq i} W_{i,j} z_j + \epsilon_i$. The conditional density of z_i is then normal mean $\sum_{j \neq i} W_{i,j} z_j$ and variance σ^2.

(b) Scientist A believes a small number of army worm midges blew in after planting, so that the first arrivals were scatted at random over the area and these then expanded through the ground to form local clusters. Using the auto-regression, elicit a prior $\pi_A(z|\sigma, u)$ for z with a single scalar parameter u.

Solution: Let $i \sim j$ if i is a neighbor of j on the lattice. The cells are all equally spaced and there is nothing to distinguish easting and northing. Also, we would expect some sort of local Markov structure, since the army-worm spread through the ground. We set $p(z_i|z_{-i}, \sigma) = p(z_i|z_{j \sim i}, \sigma, u)$ with $u > 0$ a coupling parameter appearing in $W^{(A)}$,

$$W^{(A)}_{i,j} = \begin{cases} u & \text{if } i \sim j \text{ and} \\ 0 & \text{otherwise.} \end{cases}$$

Note that $W^{(A)}$ is symmetric since $i \sim j$ is reflexive. This gives

$$\pi_A(z|\sigma, u) = N(z; 0, (I - W^{(A)})^{-1}(I - W^{(A)})^{-1}\sigma^2).$$

Remark: (not required) This is likely to leave a boundary effect in which the prior variance of z_i for i a perimeter cell is larger than that for interior cells. Also there is likely to be some overall anisotropy due to the lattice orientation.

(c) Scientist B believes the army worms blew in before planting and formed small clusters a foot or so across which were subsequently smeared out along the columns by ploughing and seeding. Using the auto-regression, elicit a prior $\pi_B(z|\sigma, v)$ for z with a single scalar parameter v.

Solution: (2,S) By the same reasoning we set $p(z_i|z_{-i}, \sigma) = p(z_i|z_{j\sim i}, \sigma, v)$ with $v > 0$ a parameter in $W^{(B)}$,

$$W^{(B)}_{i,j} = \begin{cases} v & \text{if } i \sim j \text{ and } j \text{ is directly north or south of } i \\ 0 & \text{otherwise.} \end{cases}$$

This will make the columns independent in the prior. Note that $W^{(B)}$ is again symmetric. This gives

$$\pi_B(z|\sigma, v) = N(z; 0, (I - W^{(B)})^{-1}(I - W^{(B)})^{-1}\sigma^2).$$

(d) Suppose priors for σ, u and v have been elicited. Say how you would check that these priors and the priors for z reflect available knowledge. Indicate any further information you would need from Scientists A and B.

Solution: (2, S) We could simulate synthetic data by simulating $z' \sim \pi_A(z|\sigma, u)$ (or similarly B) and then for $i = 1, ..., n$ simulate $Y'_i \sim$ Bernoulli$(\mu_i(z'_i))$. We could check with the scientists that Y' shows plausible variation by estimating for example the correlation of Y_i and Y_j for i and j neighboring cells and i and j distant pairs. We would reference this to their previous experience with army-worm, not to the data in hand. We could also consider model elaboration, incorporating different north and south weights, or additional diagonal weights, and seeing if it influenced our analysis (ie checking conclusions dont change).

8. Consider a process generating $x_1, x_2, x_3, ...$ in which $x_1 = 1$ with fixed and known probability p and for $n = 1, 2, ...,$

$$p(x_{n+1} = 1|x_n, ..., x_1) = \frac{p + k_n}{1 + n},$$

where $k_n = \sum_{i=1}^{n} x_i$ and $p(x_{n+1} = 0|x_n, ..., x_1) = 1 - p(x_{n+1} = 1|x_n, ..., x_1)$.

(a) Is the process Markov?

Solution: No. For eg $p(1|1,1) = (p+2)/3$ and $p(1|1,0) = (p+1)/3$ which are not equal for any p so the process depends on its history.

(b) Show that this process generates an infinite exchangeable sequence.

Solution: If $q \sim \text{Beta}(\alpha, \beta)$ and $x_i = 1$ wp q and otherwise $x_i = 0$ iid for $i = 1, ..., n$ then

$$p(x_1, ..., x_n|q) = q^k(1-q)^{n-k}$$

where $k = \sum_i x_i$. It follows that

$$\begin{aligned} p(x_1, ..., x_n) &= \int q^{k_n}(1-q)^{n-k_n} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} q^{\alpha-1}(1-q)^{\beta-1} dq \\ &= \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} \frac{\Gamma(\alpha+k)\Gamma(\alpha+n-k)}{\Gamma(\alpha+\beta+n)} \end{aligned}$$

and so

$$\frac{p(x_1, ..., x_n, 1)}{p(x_1, ..., x_n)} = \frac{\alpha + k_n}{\alpha + \beta + n}.$$

Now $\Pr(x_1 = 1) = p$ so $p = E(E(x_1|q)) = \alpha/(\alpha+\beta)$ so if we take $\alpha = p$ and $\beta = 1 - p$ then we match this and

$$p(x_{n+1} = 1|x_1, ..., x_n) = \frac{p + k_n}{1 + n}.$$

It follows by de Finetti that $x_1, x_2, ...$ is an infinite exchangeable sequence since

$$p(x_1, ..., x_n) = \int \prod_{i=1}^{n} q^{x_i}(1-q)^{1-x_i} \text{Beta}(q; p, 1-p) dq.$$

9. Let $G = (V, E)$ be a directed acyclic graph with nodes $V = \{1, ., ..., n\}$ and edges $E \subseteq V \times V$. A Directed Acyclic Graph (DAG) is a directed graph with no directed loops.

   For $n = 1, 2, 3, ...$ let $\mathcal{G}_n$ be the set of all distinct DAG's on n labeled nodes and let $|\mathcal{G}_n|$ give the number of such DAG's. There are three DAG's on $n = 2$ nodes with $V = \{1, 2\}$ and $G_i = (V, E_i),\ i = 1, 2, 3$. Possible edge sets are $E_1 = \emptyset$, $E_2 = \{(1, 2)\}$ and $E_3 = \{(2, 1)\}$, so $|\mathcal{G}_2| = 3$. There are $|\mathcal{G}_3| = 25$ DAG's on $n = 3$ nodes. See `https://oeis.org/A003024`.

   (a) Let $\pi_{U,n}(g),\ g \in \mathcal{G}_n$ give the uniform distribution on DAG's with n nodes, so that

$$\pi_{U,n}(g) = \frac{\mathbb{I}_{g \in \mathcal{G}_n}}{|\mathcal{G}_n|}.$$

For $G \sim \pi_{U,3}$ and $G = (V, E)$ let $G_{-3} \in \mathcal{G}_2$ be the DAG obtained from G by removing node 3 and any edge $(i, 3) \in E$ or $(3, i) \in E$, $i = 1, 2$ connected to node 3. Show that G_{-3} is not distributed according to $\pi_{U,2}$.

Solution: (3 marks, standard) 3 into 25 wont go. Let $g \in \mathcal{G}_2$ be the DAG $g = (V = \{1, 2\}, E = \emptyset)$ so $\pi_{U,2}(g) = 1/3$. It is sufficient to show that

$$\pi_{U,2}(g) \neq \sum_{g' \in \mathcal{G}_3} \pi_{U,3}(g') \mathbb{I}_{g'_{-3} = g}.$$

The LHS is equal $1/3$. The set $\{g' \in \mathcal{G}_3 : g'_{-3} = g\}$ has 9 elements (one with 0 edges, 4 with one edge and 4 with two edges) so since $1/3 \neq 9/25$ we have $\Pr(G_{-3} = g) \neq \pi_{U,2}(g)$. We dont actually have to count DAGs in $\{g' \in \mathcal{G}_3 : g'_{-3} = g\}$ as there is no integer m such that $m/25 = 1/3$.

(b) (research question) Construct a family of probability distributions $p_{1:n}(g),\ g \in \mathcal{G}_n$ over DAGs that is marginally consistent, in the sense that if $G \sim p_{1:n}$ then $G_{-n} \sim p_{1:n-1}$. Of course we can "cheat" and write down any $p_{1:n}$ we like for some fixed n and then *define* the probability $p_s,\ s \subset [n]$ for DAGs on the vertex set s to be the marginal! Not very practical at large n.

10. Let g be a simple directed graph so for $(i, j) \in I_n$ with $I_n = \{(i, j) \in [n] \times [n] : i < j\}$ the possible values are $g_{i,j} \in \Omega_{i,j}$ with $\Omega_{i,j} = \{i \to j, i \leftarrow j, i \sim j\}$ where $i \sim j$ means no edge between i and j. Let $\mathcal{G}_n$ be the space of possible simple directed graphs on n vertices. Let $G \in \mathcal{G}_n$ be the unknown true graph and suppose we have data y informing G and a prior $\pi(g) = P(G = g)$ and posterior $\pi(g|y),\ g \in \mathcal{G}_n$. Let $\hat{g} \in \mathcal{G}_n$ be an estimate for G and let

$$l(\hat{g}, G) = n(n-1) - \sum_{i<j} \mathbb{I}_{\hat{g}_{i,j} = G_{i,j}}$$

be the loss if we estimate $\hat{g}$ when the truth is G.

(a) For $e \in \Omega_{i,j}$ let $p(e) = P(G_{i,j} = e|y)$ be the marginal posterior probability that $G_{i,j} = e$. Show that the Bayes estimator $\hat{g} = \arg\min_{g \in \mathcal{G}_n} E_{G|y}(l(g, G))$ is given by $\hat{g}_{i,j} = \arg\max_{e \in \Omega_{i,j}} p(e)$ for $(i, j) \in I_n$.

Solution: For $g \in \mathcal{G}_n$ let $\vec{S}(g) = \{(i, j) \in I_n : g_{i,j} = i \to j\}$ and let $\overleftarrow{S}$ and $\widetilde{S}$ be

defined similarly. The Expected Posterior Loss (EPL) is

$$\begin{aligned} E_{G|y}(l(g,G)) &= n(n-1) - \sum_{i<j} E_{G|y}(\mathbb{I}_{g_{i,j}=G_{i,j}}) \\ &= n(n-1) - \sum_{i<j} P(G_{i,j} = g_{i,j}|y) \\ &= n(n-1) - \sum_{(i,j)\in\vec{S}(g)} p(i\to j) - \sum_{(i,j)\in\overleftarrow{S}(g)} p(i\leftarrow j) - \sum_{(i,j)\in\widetilde{S}(g)} p(i\sim j) \end{aligned}$$

so we choose g to make the sum as large as possible. Since $(\vec{S}(g), \overleftarrow{S}(g), \widetilde{S}(g))$ is a partition of I_n, each (i,j) appears in one of the sums. We should choose each $g_{i,j}$ so that $p(g_{i,j})$ is as large as possible, hence $\hat{g}_{i,j} = \arg\max_{e\in\Omega_{i,j}} p(e)$.

(b) Rewrite the loss in terms of reward and utility and give the posterior reward distribution in terms of $\pi(\cdot|y)$. Relate the expected utility to the Expected Posterior Loss and show that the Bayes estimator maximises the expected utility.

Solution: Fix $g \in \mathcal{G}_n$. If $G \sim \pi(\cdot|y)$ then the reward $R(g,G) = \sum_{i<j} \mathbb{I}_{g_{i,j}=G_{i,j}}$ takes values in $\mathcal{R}_n = \{0, 1, \ldots, n(n-1)\}$. If we take the utility to be $U(R) = R$ (we have more degrees of freedom than we need) then $l(g,G) = n(n-1) - U(R(g,G))$. Partition $\mathcal{G}_n = \cup_{r\in\mathcal{R}_n} \mathcal{G}_{n,r}$ in sets of graphs $\mathcal{G}_{n,r} = \{h \in \mathcal{G}_n : R(g,h) = r\}$ with equal reward. The reward distribution $P_{g,y}(r) = P(R(g,G) = r|y)$ is

$$P_{g,y}(r) = \sum_{h\in\mathcal{G}_{n,r}} \pi(h|y), \quad \text{for } r \in \mathcal{R}_n.$$

and

$$\begin{aligned} E_{G|y}(l(g,G)) &= n(n-1) - \sum_{h\in\mathcal{G}_n} \pi(h|y) U(R(g,h)) \\ &= n(n-1) - \sum_{r\in\mathcal{R}_n} \left[\sum_{h\in\mathcal{G}_{n,r}} \pi(h|y) \right] U(r) \\ &= n(n-1) - \sum_{r\in\mathcal{R}_n} P_{g,y}(r) U(r) \\ &= n(n-1) - E_{P_{g,y}}(U(R)) \end{aligned}$$

so minimising the EPL $E_{G|y}(l(g,G))$ over $g \in \mathcal{G}_n$ is the same as maximising the expected utility $E_{P_{g,y}}(U(R))$ on the right.

Statistics Department, University of Oxford
Geoff Nicholls: nicholls@stats.ox.ac.uk