SC7 Bayes Methods

Second problem sheet (Sections 4.3-6 of lecture notes).

Section A questions

1. The Savage axioms (as formulated by DeGroot) characterise coherent prior preference for events stated in terms of inequalities, so that $A \preceq B$ says we think $\pi(A) \leq \pi(B)$.

   (a) Write down the first three axioms (see Lecture notes).

   > **Solution:** Let S be a sample space and let $\mathcal{S}$ be a σ-field of sets in S. In the following events A, B etc are subsets of S and members of $\mathcal{S}$.
   >
   > Axiom 1. For any two events A and B exactly one of the following relations must hold: $A \succ B$, $A \prec B$, $A \sim B$.
   >
   > Axiom 2. If $A_1 \cap A_2 = B_1 \cap B_2 = \emptyset$ and $A_i \succeq B_i, i = 1, 2$ then $A_1 \cup A_2 \succeq B_1 \cup B_2$. If in addition either $A_1 \succ B_1$ or $A_2 \succ B_2$ then $A_1 \cup A_2 \succ B_1 \cup B_2$.
   >
   > Axiom 3. If $A \in \mathcal{S}$ then $\emptyset \preceq A$. Furthermore $\emptyset \prec S$.

   (b) Suppose a probability space $(S, \mathcal{S}, \pi)$ expressing prior preferences exists. For $A, B \in \mathcal{S}$ let A^c, B^c give the complements of A and B. Show $A \preceq B \;\Rightarrow\; A^c \succeq B^c$ from the Axioms of Probability.

   > **Solution:** $\pi(A \cup A^c) = 1$ (as $\pi(S) = 1$)and $\pi(A \cup A^c) = \pi(A) + \pi(A^c)$ (countable additivity) so $A \preceq B \Rightarrow \pi(A) \leq \pi(B)$ (π expresses preferences) $\Rightarrow 1 - \pi(A^c) \leq 1 - \pi(B^c)$ (substituting) $\Rightarrow \pi(A^c) \succeq \pi(B^c) \Rightarrow A^c \succeq B^c$ (expresses preferences).

   (c) No longer assuming a probability space expressing prior preferences exists, suppose preferences over sets in $\mathcal{S}$ satisfy the first three Savage Axioms. Show (from the Savage Axioms alone) that if $A \preceq B$ then $A^c \succeq B^c$.

   > **Solution:** We are given $A \preceq B$. Suppose in addition $A^c \prec B^c$. Since A^c and A are exclusive, and similarly B^c and B we have from Axiom 2 that $A \cup A^c \prec B \cup B^c$, ie $S \prec S$, which is a contradiction, so $A^c \prec B^c$ is not possible and by Axiom 1 we must have either $A^c \sim B^c$ or $A^c \succ B^c$, ie $A^c \succeq B^c$.

2. Let X be an $n \times p$ design matrix with rows $x_i, i = 1, 2, ..., n$ and $\theta = (\theta_1, \theta_2, ..., \theta_p)^T$ a p-component vector of parameters. Let $z = (z_1, ..., z_n)$ be jointly independent normal

random variables, $z \sim N(X\theta, I_n)$ with I_n the $n \times n$ identity. In the probit observation model for $y = (y_1, ..., y_n)$, we observe $y_i = 1$ if $z_i > 0$ and $y_i = 0$ if $z_i \leq 0$.

Denote by $\pi(\theta, z) = \pi(\theta)\pi(z|\theta)$ the joint density of θ and z with $\pi(\theta) = N(\theta; 0, \Sigma)$ a normal prior for θ and Σ a $p \times p$ covariance matrix.

(a) Show that $y_i \sim$ Bernoulli$(\Phi(x_i\theta))$.

Solution: $\Pr(Y_i = 1) = \Pr(z_i > 0) = \Pr(x_i\theta + w_i > 0)$ if $w_i \sim N(0, 1)$, so $\Pr(Y_i = 1) = \Pr(w_i > -x_i\theta) = \Phi(x_i\theta)$.

(b) Write the posterior $\pi(\theta, z|y)$ in terms of the model elements.

Solution:

$$\pi(\theta, z|y) \propto p(y|z, \theta)\pi(\theta)\pi(z|\theta)$$

with $p(y|z, \theta) = \prod_i \mathbb{I}_{y_i = \mathbb{I}_{z_i > 0}}$, $\pi(\theta) = N(\theta; 0, \Sigma)$ and $\pi(z|\theta) = N(z; X\theta, I_n)$.

(c) Show that

$$p(\theta|z) = N(\theta; \mu, V)$$

with $\mu = VX^Tz$ and $V = (\Sigma^{-1} + X^TX)^{-1}$.

Solution: From the prior models $p(\theta|z) = \pi(\theta)\pi(z|\theta)$ so

$$p(\theta|z) \propto \exp(-\tfrac{1}{2}[|z - X\theta|^2 + \theta^T\Sigma^{-1}\theta]).$$

On the other hand the Q claims

$$p(\theta|z) \propto \exp(-\tfrac{1}{2}(\theta - \mu)^TV^{-1}(\theta - \mu))),$$

so it is sufficient to substitute μ and V in this expression and check we recover the first expression.

$$(\theta - \mu)^TV^{-1}(\theta - \mu) = (\theta - VX^Tz)^TV^{-1}(\theta - VX^Tz)$$

expand, and $z^TX\theta = \theta^TX^Tz$ as these are scalars,

$$\begin{aligned} &= \theta^TV^{-1}\theta - 2z^TX\theta + z^TXVX^Tz \\ &= \theta^T\Sigma^{-1}\theta + \theta^TX^TX\theta - 2z^TX\theta + \text{const wrt } \theta \\ &= |z - X\theta|^2 + \theta^T\Sigma^{-1}\theta + \text{const wrt } \theta, \end{aligned}$$

so we are done as this is equal to the first exponent.

(d) Show that

$$\pi(z_i|y_i,\theta) \propto \begin{cases} N(z_i; x_i\theta, 1)\mathbb{I}_{z_i \leq 0} & \text{if } y_i = 0 \\ N(z_i; x_i\theta, 1)\mathbb{I}_{z_i > 0} & \text{if } y_i = 1 \end{cases}$$

Solution: By Bayes rule $\pi(z_i|y_i,\theta) \propto p(y_i|z_i,\theta)\pi(z_i|\theta)$. Now

$$p(y_i|z_i,\theta)\pi(z_i|\theta) = \begin{cases} N(z_i; x_i\theta, 1)\mathbb{I}_{z_i \leq 0} & \text{if } y_i = 0 \\ N(z_i; x_i\theta, 1)\mathbb{I}_{z_i > 0} & \text{if } y_i = 1 \end{cases}$$

(e) Give a Gibbs sampler sampling $\pi(\theta|y)$ (Hint: $\pi(\theta, z|y)$ would be easier).

Solution: In a Gibbs sampler we iterate between updating $\theta|z$ and then for $i = 1, ..., n$ we update $z_i|\theta, z_{-i}$, where z_{-i} is the vector of z's with z_i removed. Suppose $X_t = (\theta, z)$. By the results above we simulate

$$\theta' \sim N(\mu, V), \text{with } \mu = VX^Tz \text{ and } V = (\Sigma^{-1} + X^TX)^{-1},$$

and then for $i = 1, ..., n$,

$$z_i' \sim N(x_i\theta', 1)\mathbb{I}_{z_i' \leq 0} \quad \text{if } y_i = 0$$

and

$$z_i' \sim N(x_i\theta', 1)\mathbb{I}_{z_i' > 0} \quad \text{if } y_i = 1.$$

Then set $X_{t+1} = (\theta', z')$. This targets $\pi(\theta, z|y)$ with marginal $\pi(\theta|y)$.

3. Let $\mathcal{M} = \{1, 2\}$ and consider two generative models $\pi_m(\theta)p_m(y|\theta),\ m \in \mathcal{M}$ and corresponding marginal likelihoods $p_m(y),\ m \in \mathcal{M}$ for continuous parameters $\theta \in \Omega$ and data $y \in \mathcal{Y}$. Let $q(\theta) = c\tilde{q}(\theta)$ be an arbitrary density over Ω satisfying $q(\theta) > 0$ for all $\theta \in \Omega$.

   Show that the Bayes factor $B_{1,2} = p_1(y)/p_2(y)$ is given by[1]

$$B_{1,2} = \frac{E_{\theta\sim q}(\pi_1(\theta)p_1(y|\theta)/\tilde{q}(\theta))}{E_{\theta\sim q}(\pi_2(\theta)p_2(y|\theta)/\tilde{q}(\theta))}$$

   and state how this might be estimated using Monte Carlo samples.

[1] Ming-Hui Chen, Qi-Man Shao, *On Monte Carlo methods for estimating ratios of normalizing constants*, Ann. Statist. 25(4), 1563-1594, (1997a)

Solution: For $m \in \mathcal{M}$,

$$\begin{aligned} E_{\theta\sim q}(\pi_m(\theta)p_m(y|\theta)/\tilde{q}(\theta)) &= \int_\Omega q(\theta)\pi_m(\theta)p_m(y|\theta)/\tilde{q}(\theta)d\theta \\ &= c\int_\Omega \pi_m(\theta)p_m(y|\theta)d\theta \\ &= cp_m(y), \end{aligned}$$

so c cancels in the ratio and we get the Bayes factor. To estimate take $\theta^{(t)} \sim q,\ t = 1, ..., T$ and form the estimate

$$\hat{B}_{1,2} = \frac{\sum_{t=1}^T \pi_1(\theta^{(t)})p_1(y|\theta^{(t)})/\tilde{q}(\theta^{(t)})}{\sum_{t=1}^T \pi_2(\theta^{(t)})p_2(y|\theta^{(t)})/\tilde{q}(\theta^{(t)})}.$$

Remark: this is the ratio of two importance sampling estimators, but something nice happens - we dont need to estimate the normalising constant c as it cancels. It is important the same samples are used top and bottom as this has a stabilising effect on the estimator. For the choice of q you want a distribution central between the two posteriors. The physicists sometimes called this "umbrella sampling" and used $q(\theta) \propto \sqrt{\pi_1(\theta|y)\pi_2(\theta|y)}$ which is tractable (up to a constant). Chen and Shao call the estimator "Ratio Importance Sampling" and show the optimal choice is $q(\theta) \propto |\pi_1(\theta|y) - \pi_2(\theta|y)|$ which is not tractable.

Section B questions

4. Let $\succeq$ be a system of preferences over sets in $\mathcal{S}$ which satisfy the first three Savage Axioms. Show that if $A \cap D = B \cap D = \emptyset$ then $A \cup D \succ B \cup D$ if and only if $A \succ B$. Meaning: we can add the same set to both sides of an inequality, if it doesn't intersect the sets appearing in the inequality, and we can remove the same set from both sides of an inequality. This will be useful for Question 8.

5. Let $\Gamma(x; \alpha, \beta)$ be the Gamma density. Consider Poisson observations $Y = (Y_1, Y_2, ..., Y_n)$ with means $\lambda = (\lambda_1, \lambda_2, ..., \lambda_n)$ given by a mixture of Gamma densities: for shape parameters α_1, α_2 and rate parameters β_1, β_2, a known mixture proportion $0 < p < 1$ and $i = 1, 2, ..., n$, we observe

$$Y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$$

(all iid) with

$$\lambda_i \sim p\Gamma(\lambda_i; \alpha_1, \beta_1) + (1-p)\Gamma(\lambda_i; \alpha_2, \beta_2).$$

(a) Denote by $\pi(\alpha_1, \beta_1, \alpha_2, \beta_2)$ a prior for the unknown shape and rate parameters. Write down the joint posterior for $\alpha_1, \beta_1, \alpha_2, \beta_2$ and λ given $Y_1, Y_2, ..., Y_n$. Give an MCMC algorithm sampling $\alpha_1, \beta_1, \alpha_2, \beta_2, \lambda | Y_1, ..., Y_n$.

(b) Integrate λ out of the joint posterior to obtain a marginal posterior denisty for $\alpha_1, \beta_1, \alpha_2, \beta_2 | Y_1, ..., Y_n$. Comment briefly on how you would alter your MCMC algorithm for the new target. What considerations would guide your choice of simulation method (ie, whether to simulate the joint or the marginal posterior density)?

6. Let $\pi(\theta), \theta \in R$ be a prior density for a scalar parameter, let $p(y|\theta), y \in R^n$ be the observation model density and let $\pi(\theta|y) \propto \pi(\theta)p(y|\theta)$ be the posterior density. Consider a Markov chain simulated in the following way. Suppose $\theta^{(0)} \sim \pi(\cdot)$ is a draw from the prior and for $t = 0, 1, 2, ...$ we generate a Markov chain by simulating data $y^{(t)} \sim p(\cdot|\theta^{(t)})$ and then $\theta^{(t+1)} \sim \pi(\cdot|y^{(t)})$.

(a) i. Calculate the joint density, $p(\theta^{(0)}, \theta^{(1)})$ say, for $\theta^{(0)}, \theta^{(1)}$ and show that $p(\theta^{(0)}, \theta^{(1)}) = p(\theta^{(1)}, \theta^{(0)})$ (ie they are exchangeable).

ii. Show that marginally, $\theta^{(t)} \sim \pi(\cdot)$ for all $t = 0, 1, 2, ...$

iii. Give the transition probability density $K(\theta, \theta')$ for the chain and show the chain is reversible with respect to the prior $\pi(\theta)$.

(b) Suppose we are given an MCMC algorithm $\theta^{(T)} = \mathcal{M}(\theta^{(0)}, T, y)$, initialised at $\theta^{(0)}$, and targeting the posterior $\pi(\theta|y) \propto \pi(\theta)p(y|\theta)$, so $\theta^{(T)} \xrightarrow{D} \pi(\cdot|y)$ as $T \to \infty$. Here $\mathcal{M}$ is a function that moves us T steps forward in the MCMC run and this Markov chain is just some MCMC algorithm for simulating $\pi(\theta|y)$ and so not related to the Markov chain in the previous part.

Suppose we think we have chosen T sufficiently large that the chain has converged, and so we believe $\theta^{(T)} \sim \pi(\cdot|y)$ is a good approximation.

Consider the following procedure simulating pairs (ϕ_i, θ_i), $i = 1, 2, ..., K$: (Step 1) parameter $\phi_i \sim \pi(\cdot)$ is an independent draw from the prior; (Step 2) synthetic data $y'_i \sim p(\cdot|\phi_i)$ is an independent draw from the observation model; (Step 3) the MCMC algorithm $\mathcal{M}$ is initialised with a draw $\theta_i^{(0)} \sim \pi^{(0)}$ from an arbitrary fixed initial distribution $\pi^{(0)}$ and (Step 4) we set $\theta_i = \mathcal{M}(\theta_i^{(0)}, T, y'_i)$.

Let $\phi = (\phi_1, ..., \phi_K)$ and $\theta = (\theta_1, ..., \theta_K)$ be samples generated in this way.

i. Suppose the chain has indeed converged by T steps for all starting states $\theta^{(0)}$. Let $p(\phi, \theta)$ be the joint distribution of the random vectors ϕ and θ. Show that $p(\phi, \theta) = p(\theta, \phi)$.

ii. Give a non-parametric test for MCMC convergence which makes use of the result in Question 6(b)i. Hint: the null is $\theta^{(T)} \sim \pi(\cdot|y)$.

7. (a) Consider two models with parameter spaces respectively $\theta \in \Re^p$ and $\phi = (\theta, \psi)$ with $\psi \in \Re^q$, so that $\phi \in \Re^{p+q}$. We want to compare model 1 with prior $\pi_1(\theta)$, observation model $p_1(y|\theta)$ and marginal likelihood $p_1(y)$ with model 2 where we have $\pi_2(\phi)$, $p_2(y|\phi)$, and $p_2(y)$ correspondingly.

   Let $Q(\psi)$ be a probability density on $\Re^q$. Show that

$$\frac{p_1(y)}{p_2(y)} = \frac{E_{(\theta,\psi)|y,m=2}(Q(\psi)\pi_1(\theta)p_1(y|\theta)h(\theta,\psi))}{E_\psi(E_{\theta|y,m=1}(\pi_2(\theta,\psi)p_2(y|\theta,\psi)h(\theta,\psi)))}$$

   where $\psi \sim Q$ in the expectation in the denominator and $h : \Re^{p+q} \to \Re$ is a function chosen so that the expectations exist. Comment briefly on how this last identity may be used for model comparison for models defined on spaces of unequal dimension.[2]

   (b) Briefly outline any assumptions we are making about the densities above.

Section C questions

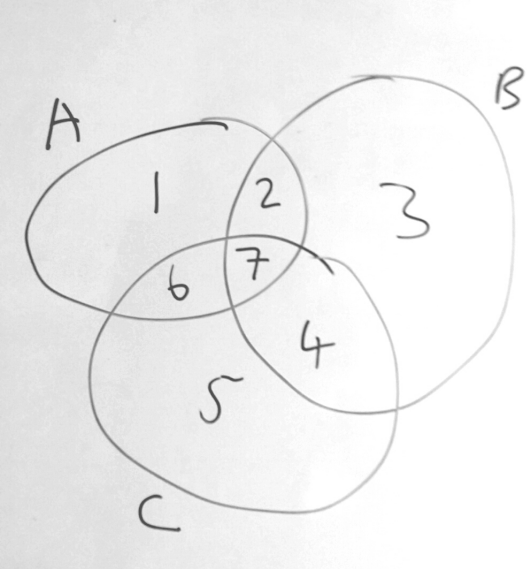
8. Show that prior preferences respecting the first three Savage Axioms are transitive, that is, if $A \preceq B$ and $B \preceq C$ then $A \preceq C$.

Solution: In a shorthand from the figure below, where $1 = A \cap B^c \cap C^c$, $2 = A \cap B \cap C^c$ etc, we let sequences of numbers denote unions of the corresponding sets so $12 = (A \cap B^c \cap C^c) \cup (A \cap B \cap C^c)$ etc.

By Question 4, $A \preceq B$ implies $16 \preceq 34$ (the set D we chop off A and B is 27) and since $B \preceq C$ we have $23 \preceq 56$ (removing the common element 47). Now $16 \cap 23 = \emptyset$ and $34 \cap 56 = \emptyset$ so taking the unions of both sides using axiom 2, we have $1623 \preceq 3456$. We can remove the common overlap 36 using the lemma. So $12 \preceq 45$. But if we now

[2] Chen, M.H. and Shao, Q.M. (1997b). *Estimating ratios of normalizing constants for densities with different dimensions.* Statistica Sinica v7, p607–630.

add 67 to both sides using the lemma we have $1267 \preceq 4567$ which is $A \preceq C$.



9. (MSc 2020 exam - students had a related practical in 2020) A book club with n members wants to decide what book to read next. They have a shortlist of B books with labels $\mathcal{B} = \{1, ..., B\}$. Let $\mathcal{P}_\mathcal{B}$ be the set of all permutations of the labels in $\mathcal{B}$. For $i = 1, ..., n$ the i'th reader gives a ranked list of the books $y_i = (y_{i,1}, ..., y_{i,B}),\ y_i \in \mathcal{P}_\mathcal{B}$, ranking them from most to least interesting. The data are $y = (y_1, ..., y_n)$.

   In a Plackett-Luce model each book $b = 1, ..., B$ has interest measure $\theta_b > 0$. Let $\theta = (\theta_1, ..., \theta_B),\ \theta \in R^B$. Let $Y_i \in \mathcal{P}_\mathcal{B}$ denote the random ranking from the i'th reader. In the Plackett-Luce model, given $Y_{i,1} = y_{i,1}, ..., Y_{i,a-1} = y_{i,a-1}$, the a'th entry (ie, the next entry) is decided by choosing book b with probability proportional to θ_b from the books $\mathcal{B} \setminus \{y_{i,1}, ..., y_{i,a-1}\}$ remaining. The $Y_1, ..., Y_n$ are jointly independent given θ.

   (a) i. Show that the likelihood $L(\theta; y)$ is

$$L(\theta; y) = \prod_{i=1}^{n} \prod_{a=1}^{B} \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{B} \theta_{y_{i,b}}}.$$

Solution: Since

$$\begin{aligned} \Pr\{Y_i = y_i|\theta\} &= \Pr\{Y_{i,1} = y_{i,1}|\theta\} \times \Pr\{Y_{i,2} = y_{i,2}|Y_{i,1} = y_{i,1}, \theta\} \times ... \\ &\quad \times \Pr\{Y_{i,B} = y_{i,B}|Y_{i,1:(B-1)} = y_{i,1:(B-1)}, \theta\} \end{aligned}$$

and

$$\Pr\{Y_{i,a} = y_{i,a} | Y_{i,1:(a-1)} = y_{i,1:(a-1)}, \theta\} = \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{B} \theta_{y_{i,b}}}$$

from the definition of the process, we must have

$$\Pr\{Y_i = y_i|\theta\} = \prod_{a=1}^{B} \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{B} \theta_{y_{i,b}}}, \qquad i = 1, ..., n$$

and therefore $L(\theta; y) = \prod_{i=1}^{n} \Pr\{Y_i = y_i|\theta\}$ by independence.

ii. The prior is $\pi_{\mathcal{B}}(\theta) = \prod_{b=1}^{B} \pi(\theta_i)$ with $\pi(\theta_i) = \Gamma(\theta_i; \alpha', 1)$ with $\alpha' > 0$ given. Write down the posterior density $\pi(\theta|y)$ and give an MCMC algorithm targeting $\pi(\theta|y)$.

Solution: First, $\pi(\theta|y) \propto \pi_{\mathcal{B}}(\theta)L(\theta; y)$ with $\pi_{\mathcal{B}}$ and L given.
Secondly, here is "an MCMC algorithm" targeting $\pi(\theta|y)$. Let $X_t = \theta$.

A. For $b = 1, ..., B$

S1 simulate $\theta'_b \sim \pi(\cdot)$ and set $\theta'_a = \theta_a,\ a = 1, ..., B,\ a \neq b$;

S2 with probability

$$\alpha(\theta'|\theta) = \min\left\{1,\ \frac{L(\theta'; y)}{L(\theta; y)}\right\}$$

set $\theta \leftarrow \theta'$.

B. Set $X_{t+1} = \theta$.

Some further simplification is possible but unnecessary.

iii. Explain why the scale β' in the prior $\Gamma(\alpha', \beta')$ for $\theta_i,\ i \in \mathcal{B}$ may be set equal one. Suppose odds of 1000 : 1 for ranking one book above another represent extreme preference and are a priori unlikely for books on the shortlist. Explain how a fixed numerical value of α' might be chosen, noting any assumptions.

Solution:
First, we can choose scale $\beta' = 1$. If we choose otherwise we can simply rescale $\theta' = \theta\beta'$. The likelihood is invariant so this factor is not identifiable anyway. Put another way, we can recover the posterior for $\beta' = c$ by by multiplying $\theta \sim \pi(\theta|y, \beta' = 1)$ by c.

Secondly, suppose there are just $B = 2$ books in the list and one reader. I assume (very reasonably!) that prior knowledge about θ is not changed by these choices. Then

$$O_{1,2} = \Pr\{Y_1 = (1,2)|\theta\} / \Pr\{Y_1 = (2,1)|\theta\} = \theta_1/\theta_2.$$

If a value of $O_{1,2}$ as large as 1000 represents extreme preference across readers then it should be unlikely. I assume "unlikely" to mean having probability less than 1%. Choose α' so that the 99% quantile of $= \theta_1/\theta_2$ is at 1000 (I assume this is possible - it is). This could be done using simulation if necessary. Many different choices and justifications may be given but concrete assumptions are required.

(b) Suppose B is large so each reader $i = 1, ..., n$ only reports the first N entries $x_i = (x_{i,1}, ..., x_{i,N})$ in their ranking, with $N \ll B$. Here $x_{i,j} = y_{i,j}$ for $i = 1, ..., n$ and $j = 1, ..., N$. The data are $x = (x_1, ..., x_n)$.

i. Show that the likelihood $L(\theta; x)$ for the new data is

$$L(\theta; x) = \prod_{i=1}^{n} \prod_{a=1}^{N} \frac{\theta_{x_{i,a}}}{\sum_{b=a}^{N} \theta_{x_{i,b}} + \sum_{d \in \mathcal{B} \setminus x_i} \theta_d}.$$

Solution:
From the process description we just take the first N outcomes but choose from the full list, so

$$\Pr\{Y'_{i,a} = x_{i,a} | Y'_{i,1:(a-1)} = x_{i,1:(a-1)}, \theta\} = \frac{\theta_{x_{i,a}}}{\sum_{b=a}^{N} \theta_{y_{i,b}} + \sum_{d \in \mathcal{B} \setminus y_i} \theta_d},$$

proceeding in the same way as before. Note that, for reader i, we always add $\sum_{d \in \mathcal{B} \setminus y_i} \theta_d$ in the denominator since these books didn't get picked by reader i so they never appear in the numerator. Taking the product,

$$\Pr\{Y'_i = x_i | \theta\} = \prod_{a=1}^{N} \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{N} \theta_{y_{i,b}} + \sum_{d \in \mathcal{B} \setminus y_i} \theta_d}, \qquad i = 1, ..., n$$

with $L(\theta; x) = \prod_{i=1}^{n} \Pr\{Y'_i = x_i | \theta\}$.

ii. Let $\mathcal{C} = \bigcup_{i=1}^{n} x_i$ give the books appearing in at least one ranking and $\mathcal{D} = \mathcal{B} \setminus \mathcal{C}$ be the books appearing in none. Let $\theta_{\mathcal{C}} = (\theta_b)_{b \in \mathcal{C}}$ and $V = \sum_{d \in \mathcal{D}} \theta_d$.

Write down the prior distribution of V and the likelihood $L(\theta_{\mathcal{C}}, V; x)$, and give the posterior $\pi(\theta_{\mathcal{C}}, V|x)$ as a function of $\theta_{\mathcal{C}}$ and V.

Solution:
First, $V \sim \Gamma(m\alpha', 1)$ with $m = |\mathcal{D}|$ in the prior.
Secondly, since $\mathcal{B} \setminus y_i = (\mathcal{C} \setminus y_i) \cup \mathcal{D}$ and the sets in the union are disjoint, the likelihood can be written

$$L(\theta_{\mathcal{C}}, V; x) = \prod_{i=1}^{n} \prod_{a=1}^{N} \frac{\theta_{y_{i,a}}}{\sum_{b=a}^{N} \theta_{y_{i,b}} + \sum_{d \in \mathcal{C} \setminus y_i} \theta_d + V}.$$

Only $\theta_b, b \in \mathcal{C}$ appear in this expression outside V so the parameterisation on the left is correct.
Thirdly, the posterior

$$\pi(\theta_C, V|x) \propto L(\theta_C, V; x)\Gamma(V; m\alpha', 1) \prod_{i \in C} \pi(\theta_i),$$

has the same marginal distribution for θ_C as $\pi(\theta|x)$.
Alternatively for this last bit, a more explicit calculation integrates $\pi(\theta|x)$ over dof in $\theta_{\mathcal{D}}|V$. These do not appear in the likelihood. We can write

$$\pi_{\mathcal{B}}(\theta) = \pi_{\mathcal{C}}(\theta_{\mathcal{C}})\pi_{\mathcal{D}}(\theta_{\mathcal{D}})$$

and

$$\pi_{\mathcal{D}}(\theta_{\mathcal{D}}) = p(V)p(\theta_{\mathcal{D}}|V)$$

and use the fact that we know the marginal prior $p(V) = \Gamma(V; m\alpha', 1)$.

iii. Give an MCMC algorithm targeting $\pi(\theta_{\mathcal{C}}, V|x)$. State briefly why it may be more efficient, for estimation of $\theta_{\mathcal{C}}$ in the case $|\mathcal{C}| \ll B$, than MCMC targeting $\pi(\theta|x)$.

Solution: First the MCMC. Let $X_t = (\theta_{\mathcal{C}}, V)$.

A. For $b \in \mathcal{C}$

S1 simulate $\theta'_b \sim \pi(\cdot)$ and set $\theta'_a = \theta_a,\ a \in \mathcal{C},\ a \neq b$; this determines $\theta'_{\mathcal{C}}$

S2 with probability

$$\alpha(\theta'_{\mathcal{C}}|\theta_{\mathcal{C}}) = \min\left\{1,\ \frac{L(\theta'_{\mathcal{C}}, V; x)}{L(\theta_{\mathcal{C}}, V; x)}\right\}$$

set $\theta_{\mathcal{C}} \leftarrow \theta'_{\mathcal{C}}$.

B. **T1** simulate $V' \sim \Gamma(m\alpha', 1)$

T2 with probability

$$\alpha(V'|V) = \min\left\{1,\ \frac{L(\theta_{\mathcal{C}}, V'; x)}{L(\theta_{\mathcal{C}}, V; x)}\right\}$$

set $V' \leftarrow V$.

C. Set $X_{t+1} = (\theta_{\mathcal{C}}, V)$.

Secondly, the efficiency remark. This target has a lower parameter dimension ($|\mathcal{C}| + 1 \ll B$) than $\pi(\theta|x)$ and similar computational cost per update so we expect more rapid MCMC convergence in a given computing time. Essentially we are targeting a marginal with uninteresting parameters integrated out.

10. For $\theta \in \Omega$ and $i = 1, 2$ let $p_i(\theta) = q_i(\theta)/c_i$ and $\theta_i^{(t)} \sim p_i,\ t = 1, ..., T$ so c_i normalises q_i. Let h be defined so that $\int_\Omega q_1(\theta)q_2(\theta)h(\theta)d\theta$ exists. Let $r = c_1/c_2$ and

$$\hat{r}_h = \frac{\sum_{t=1}^T q_1(\theta_2^{(t)})h(\theta_2^{(t)})}{\sum_{j=1}^T q_2(\theta_1^{(t)})h(\theta_1^{(t)})}.$$

Let the relative mean square error be defined

$$RE(\hat{r}_h) = \frac{E[(\hat{r}_h - r)^2]}{r^2},$$

where the expectation is taken over the random samples $\theta_i^{(t)},\ t = 1, ..., T$ for $i = 1, 2$ which are assumed jointly independent. It may be shown (using the delta-rule) that

$$RE(\hat{r}_h) = \frac{1}{T}\int_\Omega \frac{p_1(\theta)p_2(\theta)(p_1(\theta) + p_2(\theta))h(\theta)^2 d\theta}{\left(\int_\Omega p_1(\theta)p_2(\theta)h(\theta)d\theta\right)^2} - \frac{2}{T} + O(T^{-2}).$$

Show that this expression is minimised over functions h by the choice[3]

$$h(\theta) \propto \frac{1}{p_1(\theta) + p_2(\theta)}.$$

Hint: Cauchy Schwarz or functional differentiation WRT h both lead to the result.

[3]following the proof in Meng, XL and Wong, WH, *Simulating ratios of normalizing constants via a simple identity: a theoretical exploration*, Statistica Sinica 6:831-860 (1996)

Solution: Dropping the explicit θ dependence,

$$\begin{aligned}\left(\int_\Omega p_1 p_2 h d\theta\right)^2 &= \left(\int_\Omega \sqrt{\frac{p_1 p_2}{p_1+p_2}}\sqrt{p_1 p_2(p_1+p_2)}h d\theta\right)^2 \\ &\leq \left(\int_\Omega \left|\sqrt{\frac{p_1 p_2}{p_1+p_2}}\sqrt{p_1 p_2(p_1+p_2)}h\right| d\theta\right)^2 \\ &\leq \int_\Omega \frac{p_1 p_2}{p_1+p_2} d\theta \int_\Omega p_1 p_2(p_1+p_2)h^2 d\theta\end{aligned}$$

using the Cauchy-Schwarz inequality. So

$$\int_\Omega \frac{p_1(\theta)p_2(\theta)(p_1(\theta)+p_2(\theta))h(\theta)^2 d\theta}{\left(\int_\Omega p_1(\theta)p_2(\theta)h(\theta)d\theta\right)^2} \geq \left(\int_\Omega \frac{p_1 p_2}{p_1+p_2} d\theta\right)^{-1}$$

and this bound is achieved by $h = h_O$ where $h_O \propto (p_1+p_2)^{-1}$.

Alternatively take $h(\theta) = h_O(\theta) + \epsilon f(\theta)$ (with h_O the unknown optimum), differentiate wrt to ϵ, and set $\epsilon = 0$. Let

$$\begin{aligned}A_h &= \int_\Omega p_1(\theta)p_2(\theta)(p_1(\theta)+p_2(\theta))h(\theta)^2 d\theta \\ B_h &= \left(\int_\Omega p_1(\theta)p_2(\theta)h(\theta)d\theta\right)^2\end{aligned}$$

Then

$$\left.\frac{\partial A_{h_O+\epsilon f}/B_{h_O+\epsilon f}}{\partial \epsilon}\right|_{\epsilon=0} = \frac{2\int_\Omega p_1 p_2(p_1+p_2)f h_O d\theta \times B_{h_O} - 2A_{h_O}\sqrt{B_{h_O}} \times \int_\Omega p_1 p_2 f d\theta}{B_{h_O}^2} = 0$$

and this must hold for every function f such that the integrals exist and this requires the integral argument to be zero,

$$\sqrt{B_{h_O}} p_1 p_2(p_1+p_2)h_O - A_{h_O} p_1 p_2 = 0$$

which is $(p_1+p_2)h_O = A_{h_O}/\sqrt{B_{h_O}}$. Since the RHS is not a function of θ we must have $h_O = c(p_1+p_2)^{-1}$ (and indeed the RHS is then $A_{h_O}/\sqrt{B_{h_O}} = c$).

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