SC7 Bayes Methods

Second problem sheet (Sections 4.3-5 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).

   (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.

   (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 $A^c \succeq B^c$.

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 $y_i \sim \text{Bernoulli}(\Phi(x_i\theta))$.

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

   (c) Show that

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

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

   (d) Show that

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

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

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.

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).

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

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.

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

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 $A \preceq C$.

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

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

   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 $\pi(\theta|y)$.

   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.

   (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

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

   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.

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 $\pi(\theta|x)$.

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.

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

[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)