SC7 Bayes Methods

Second problem sheet (Sections 3-4 of lecture notes).

Section A questions

1. 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.

2. Suppose $X = x$ is a draw from an Ising model distribution $X \sim \pi(\cdot|\theta)$ on an $m \times m$ lattice, so that for $i = 1, 2, ..., n$ with $n = m^2$, $x_i \in \{0, 1\}$ and $x = (x_1, ..., x_n)$ with $n = m^2$. Let N_i be the set of neighbors of pixel i on the square lattice. Let $\#x$ denote the number of disagreeing neighbors, that is

   $$\#x = \frac{1}{2}\sum_{i=1}^{m^2}\sum_{j\in N_i}\mathbb{I}_{x_i \neq x_j}.$$

   Under the Ising model, $\pi(x|\theta) = \exp(-\theta\#x)/Z(\theta)$ where $Z(\theta)$ is a normalising constant.

   Suppose we don't observe X itself but instead observe $Y = y_{obs}$ with $y_{obs} = (y_{obs\,1}, ..., y_{obs\,n})$ and $Y_i|x_i \sim N(x_i, \sigma^2)$ iid for $i = 1, 2, ..., n$. Here $\sigma > 0$ is known and a prior $\theta \sim \text{Exp}(2)$ is elicited for θ.

   (a) Write down the posterior $\pi(\theta, x|y_{obs})$ in terms of the model elements and explain why it is doubly intractable when $m \gg 1$.

   (b) Consider the statistic for $y \in R^n$

   $$S(y) = \frac{1}{2}\sum_{i=1}^{m^2}\sum_{j\in N_i}(y_i - y_j)^2$$

   and distance measure $d(s - s') = |s - s'|$. Briefly motivate this choice of of the ABC statistic $S(y)$ [Hint: what happens when $\sigma \ll 1$.].

   (c) An MCMC algorithm for $X \sim \pi(x|\theta)$ is available. Give an ABC algorithm targeting $\pi(\theta|d(S(Y), S(y_{obs})) < \delta)$ using the MCMC algorithm to simulate $X \sim \pi(x|\theta)$.

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

Section B questions

3. (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. Explain briefly why

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

for h a function we are free to choose.

   (b) 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]

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

4. (ABC) We considered a version of ABC related to the rejection algorithm. Consider the following MCMC-ABC algorithm[3], targeting $\pi(\theta|y)$ (approximately) using the statistics $S(y)$, distance $d(S, S')$ and threshold δ. The observation model is $p(y|\theta)$ and the prior is $\pi(\theta)$. Suppose $X_t = \theta$.

   Step 1. Simulate $\theta' \sim q(\theta'|\theta)$ and $y' \sim p(y'|\theta')$.

   Step 2. If $d(S(y), S(y')) < \delta$ then accept θ' (set $X_{t+1} = \theta'$) with probability

$$\alpha(\theta'|\theta) = \min\left\{1,\ \frac{\pi(\theta')q(\theta|\theta')}{\pi(\theta)q(\theta'|\theta)}\right\}$$

   and otherwise reject θ' (set $X_{t+1} = \theta$).

   (a) Show this algorithm targets $\pi(\theta|d(S(Y), S(y)) < \delta)$ (y is fixed here and $Y \sim p(\cdot|\theta)$).

   (b) Suppose we have $y_i \sim \text{Poisson}(\Lambda)$, $i = 1, 2, ..., n$ with $n = 5$. Prior $\lambda \sim \Gamma(\alpha = 1, \beta = 1)$. Give the ABC-MCMC algorithm targeting $\pi(\lambda|y)$ (approximately). Take $S(y) = \bar{y}$, $d(\bar{y}', \bar{y}) = |\bar{y}' - \bar{y}|$ and $\delta = 0.5$.

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

[3] Marjoram et al, "Markov chain Monte Carlo without likelihoods", PNAS (2003).

5. 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 back 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.

6. If $\pi(\theta)$ is a prior for θ then an inference scheme is a rule $\psi(\theta; \pi, y)$ for updating belief for θ given data y. For example in Bayesian inference $\psi_{Bayes}(\theta; \pi, y) = \pi(\theta|y)$ but in ABC at fixed δ, $\psi_{\Delta,\delta}(\theta; \pi, y) = \pi(\theta|Y \in \Delta_y(\delta))$.

   For $1 \leq j < n$ let $y_{1:j} = (y_1, ..., y_j)$ and $y_{j+1:n} = (y_{j+1}, ..., y_n)$ so we split the data into two sets. Suppose the data are conditionally independent, so

   $$p(y|\theta) = \prod_{i=1}^{n} p(y_i|\theta).$$

   A belief update is *order-coherent* for conditionally independent data if

   $$\psi(\theta; \pi, y) = \psi(\theta; \psi(\theta; \pi, y_{1:j}), y_{j+1:n})$$

   for all $j \in \{1, 2, ..., n - 1\}$ (the posterior from the first data set is the prior for the next).

   (a) Show that Bayesian inference is order-coherent.

   (b) Show that ABC with fixed δ is not in general order coherent. *Hint: take summary statistic $S(y) = y$ and Euclidean distance measure $d(y, y') = ||y - y'||$ and give a counter-example.*

   (c) Let $C_y(\delta)$ be the rectangular prism $C_y(\delta) = \{y' \in R^n : |y_i - y'_i| < \delta \; \forall \; i = 1, ..., n\}$. Show that inference with $\psi_{C,\delta}(\theta; \pi, y) = \pi(\theta|Y \in C_y(\delta))$ is order-coherent.

Section C questions

7. 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

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

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[4]

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

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`

[4] 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)