10 Statistics

10.9 Markov Chain Monte Carlo (6 units)

Bayesian inference is covered in the IB Statistics course, and developed further in the II(D) Principles of Statistics course. Knowledge of the IB Markov Chains course, whilst useful, is not necessary, and there is no requirement to quote results from it.

Introduction

In Bayesian statistics, it is essential to be able to sample from the posterior distribution of unknown parameters given some data. In arbitrary, high-dimensional problems, this is not possible analytically, but in recent years Markov Chain Monte Carlo methods (MCMC) have become a popular alternative.

Markov Chain

The key idea is quite simple. We want to sample from a distribution $\pi(\mathbf{x})$, $\mathbf{x} \in \mathbb{R}^m$, but cannot do so directly. Instead we create a discrete-time Markov chain $\mathbf{X}(n)$ (taking values in $\mathbb{R}^m$) such that $\mathbf{X}(n)$ has equilibrium distribution π. Then

$$\mathbf{X}(n) \to \mathbf{X} \sim \pi \quad \text{in distribution, as } n \to \infty \tag{1}$$

and

$$\frac{1}{N}\sum_{n=1}^{N} f\big(\mathbf{X}(n)\big) \to \mathbb{E}_\pi\big(f(\mathbf{X})\big) \quad \text{as } N \to \infty$$

where f is any real-valued function on $\mathbb{R}^m$ for which the right-hand side above is well-defined. The second of these limits can be used to calculate means and variances of components of $\mathbf{X}$, as well as approximations to the distribution functions. For example, $f(\mathbf{x}) = x_i$ gives the mean of the ith component X_i, and $f(\mathbf{x}) = I(x_i \leqslant b)$ gives the distribution function of X_i at point b.

Gibbs sampler

Suppose we do not have a tractable closed-form expression for the equilibrium density $\pi(\mathbf{x}) = \pi(x_1, \ldots, x_m)$, but we do know the induced full conditional densities $\pi(x_i|\mathbf{x}_{-i})$, where $\mathbf{x}_{-i}$ is the vector $\mathbf{x}$ omitting the ith component, $\mathbf{x}_{-i} = (x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_m)$.

A systematic form of the Gibbs sampler algorithm proceeds as follows. First, pick an arbitrary starting value $\mathbf{x}^0 = (x_1^0, \ldots, x_m^0)$. Then successively make random drawings from the full conditional distributions $\pi(x_i|\mathbf{x}_{-i})$, $i = 1, \ldots, m$, as follows:

$$\begin{aligned}
&x_1^1 \text{ from } \pi(x_1|\mathbf{x}_{-1}^0)\\
&x_2^1 \text{ from } \pi(x_2|x_1^1, x_3^0, \ldots, x_m^0)\\
&x_3^1 \text{ from } \pi(x_3|x_1^1, x_2^1, x_4^0, \ldots, x_m^0)\\
&\vdots\\
&x_m^1 \text{ from } \pi(x_m|\mathbf{x}_{-m}^1).
\end{aligned}$$

This cycle completes a transition from $\mathbf{x}^0 = (x_1^0, \ldots, x_m^0)$ to $\mathbf{x}^1 = (x_1^1, \ldots, x_m^1)$. Repeating the cycle produces a sequence $\mathbf{x}^0, \mathbf{x}^1, \mathbf{x}^2, \ldots$ which is a realization of a Markov chain, which is

known as the Gibbs sampler. We call $\pi(\mathbf{x}, \mathbf{y})$ the transition probability density of this Markov chain.

Question 1 Assume that the Markov chain $\mathbf{X}(n)$ takes values in a finite subset $S \subset \mathbb{R}^m$. Verify that π is an equilibrium distribution for this chain. That is, check that for all $\mathbf{y} \in S$,

$$\sum_{\mathbf{x}} \pi(\mathbf{x})\pi(\mathbf{x}, \mathbf{y}) = \pi(\mathbf{y}).$$

It can be shown that this implies that π is the equilibrium distribution of the Gibbs sampler, in the sense of (1), but do not attempt to prove it. Thus our estimate of $\mathbb{E}_\pi\big(f(\mathbf{X})\big)$, taken over N iterations, is

$$\frac{1}{N}\sum_{n=1}^{N} f(\mathbf{x}^n).$$

Football data

Data from the performance of K football teams, over T years has been scored on a scale of 0 (no wins) to 114 (win in all 38 games), with a win scoring three and a draw scoring one point. Let us model Y_{kt}, the score of the kth team in year t, as

$$Y_{kt}|\text{parameters} \sim \mathrm{N}(\mu_k, \sigma_k^2), \quad \text{for } k = 1, \ldots, K \text{ and } t = 1, \ldots, T$$

with the hierarchical prior structure that the team mean μ_k and variance σ_k^2 are independently distributed, given θ, as

$$\begin{aligned} \mu_k|\theta &\sim \mathrm{N}(\theta, \sigma_0^2) \\ \sigma_k^{-2} &\sim \Gamma(\alpha_0, \beta_0), \end{aligned}$$

where σ_0^2, α_0 and β_0 are known parameters, and θ is a second-stage prior with distribution

$$\theta \sim \mathrm{N}(\mu_0, \tau_0^2),$$

where μ_0 and τ_0^2 are known parameters.

The Gibbs sampler is well suited to the analysis of hierarchical models, since the full one-dimensional conditional distributions often have extremely simple forms. For example, in the above model,

$$\begin{aligned} \mu_k|\boldsymbol{\mu}_{-k}, \theta, \boldsymbol{\sigma}^2, \mathbf{y} &\sim \mathrm{N}\left(\frac{\sigma_k^{-2}\sum_{t=1}^{T} y_{kt} + \theta\sigma_0^{-2}}{T\sigma_k^{-2} + \sigma_0^{-2}}, \frac{1}{T\sigma_k^{-2} + \sigma_0^{-2}}\right) \\ \theta|\boldsymbol{\sigma}^2, \boldsymbol{\mu}, \mathbf{y} &\sim \mathrm{N}\left(\frac{\sigma_0^{-2}\sum_{k=1}^{K} \mu_k + \mu_0\tau_0^{-2}}{K\sigma_0^{-2} + \tau_0^{-2}}, \frac{1}{K\sigma_0^{-2} + \tau_0^{-2}}\right) \\ \sigma_k^{-2}|\boldsymbol{\sigma}^2_{-k}, \boldsymbol{\mu}, \theta, \mathbf{y} &\sim \Gamma\left(\alpha_0 + \frac{T}{2}, \beta_0 + \frac{1}{2}\sum_{t=1}^{T}(y_{kt} - \mu_k)^2\right). \end{aligned}$$

Question 2 Verify the one-dimensional conditional distributions given above. What is the marginal prior distribution of μ_k?

Question 3 Implement the Gibbs sampler to sample from the posterior distribution of $(\boldsymbol{\mu}, \boldsymbol{\sigma}^2, \theta)$ given $\mathbf{y}$. You can find the data for $\mathbf{y}$ in the file `II-10-9-2019football.csv` on the CATAM website. Take as known $\sigma_0 = 10$, $\alpha_0 = 10^{-5}$, $\beta_0 = 10^{-3}$, $\mu_0 = 60$, $\tau_0 = 20$. Briefly discuss what these prior parameter values have assumed about the football data. With reference to these priors, how did you choose the initial state of the Markov chain?

If you wish, you can use a package to simulate distributions, but you should implement the Gibbs sampler yourself without using library routines.

Question 4 Use your Gibbs sampler to estimate the posterior mean of each parameter θ, μ_k, σ_k^2. Plot a histogram of the posterior distribution of θ and comment on your histogram. Explain how you obtained it.

Question 5 Now choose a team k. Estimate the posterior probability that your chosen team is above average, $\mathbb{P}(\mu_k > \theta|\mathbf{y})$.

Question 6 Build up an idea of how accurate your estimates for μ_k and $\mathbb{P}(\mu_k > \theta|\mathbf{y})$ are, for your chosen team k, by performing independent runs of the Gibbs sampler, computing estimates for each of the parameters on each run and then computing the sample variances of these estimates. Comment on how fast your estimates converge by considering sample variances at different values of N.

Question 7 Now try letting the algorithm run for an initial period of M cycles before calculating estimates based on a further N iterations. This might allow the distribution to settle down to equilibrium before being measured. Calculate sample variances (as the previous question) for a few suitable values of M to see if this makes any noticeable difference. Explain why you do or do not see a difference.

Question 8 In any MCMC procedure we must ensure that we are exploring the full sample space. One way to check this is to run a number of chains that start from different points. Using a few widely dispersed starting points confirm, or otherwise, that your results are independent of the starting point. What is the effect of running an initial M cycles in this situation?

Hint. Recall that the Gamma distribution $\Gamma(\gamma, \lambda)$ has density

$$f(x) = \frac{\lambda^\gamma x^{\gamma-1} e^{-\lambda x}}{\Gamma(\gamma)},$$

with mean γ/λ and variance γ/λ^2. Also recall that a Gamma $\Gamma(n/2, \lambda)$ has the same distribution as the scaled chi-squared $(2\lambda)^{-1}\chi_n^2$.

You can assume that a $\Gamma(2.50001, \lambda)$ is approximately distributed as a χ_5^2 (suitably scaled), which in turn is exactly equal to the sum of two independent exponentials plus an independent normal squared.

