Stats4Astro 2017

MCMC Computer Lab

This lab will focus on writing MCMC samplers for high energy spectral analysis. This will largely involve writing code for the samplers described in the notes. We will use a number of simulated data sets rather than real data to avoid technical difficulties that would arise in fully accounting for the actual data generation mechanisms. The simulated data sets are among the files provided for the course. The following Table lists the simulated data sets and calibration products that you will need to complete this lab.

file name	columns	description
`spectral-1.txt`	Energy, Count	a basic spectral data set
`spectral-2.txt`	Energy, Count	a low-count spectral data set
`spectral-3.txt`	Energy, Count	a spectral data set with an emission line (Not provided, see below.)
`spectral-4.txt`	Energy, Count	a spectral data set that includes absorption, effective area, and background
`eff-area.txt`	Energy, Effective Area	an effective area curve

Each of the spectral data sets is a simulated high-energy spectrum and consist of photon counts from a number of energy bins. The first column reports the mean energy (keV) of each bin and the second the photon count. I refer to the mean energy in bin i as E_i and the count in bin i as Y_i for $i = 1, \ldots, n$. The energy range for the first three simulated spectra is [1.905, 10.730] and the energy range for the fourth is [1.905, 7.365].

We model the photon counts as independent Poisson random variables, $Y_i \overset{\text{indep}}{\sim} \text{Poisson}(\Lambda_i)$, where Λ_i is the expected photon count in bin i and is constrained according to a parameterized model. Three parametrized models will be considered. In Model 1 Λ_i follows a simple powerlaw, i.e.,

$$\text{Model 1} : \Lambda_i(\theta_1) = \alpha E_i^{-\beta},$$

where $\theta_1 = (\alpha, \beta)$ is a model parameter of direct scientific interest.

In Model 2, we add a spectral line to Model 1.

$$\text{Model 2} : \Lambda_i(\theta_2) = \alpha E_i^{-\beta} + \gamma I\{i \in \mathcal{L}(\delta)\},$$

where $I\{i \in \mathcal{L}(\delta)\}$ is an indicator function that is one if i is contained in the set $\mathcal{L}(\delta)$ and zero otherwise, $\mathcal{L}(\delta) = \{\delta - 1, \delta, \delta + 1\}$ for $\omega = 1, \ldots, n$, and $\theta_2 = (\alpha, \beta, \gamma, \delta)$ is a parameter of scientific interest. Thus, the spectral line adds a constant γ to the powerlaw in each of three adjacent bins and the location of the spectral line in indexed by δ.

In Model 3, we add absorption, effective area and background contamination to Model 1, but do not include the spectral line.:

$$\text{Model 3} : \Lambda_i(\theta_3) = (\alpha E_i^{-\beta} + \kappa) \cdot e^{-\omega/E_i} \cdot A(E_i)$$

where κ models the background intensity, $e^{-\omega/E_i}$ is the probability of a photon not being absorbed and $A(E_i)$ represents the effective area. Here $\theta_2 = (\alpha, \beta, \kappa, \omega)$ is a parameters of scientific interest.

When fitting these models you may use flat priors on all parameters.

1. Write code for a Random Walk Metropolis sampler to fit MODEL 1 to `spectral-1.txt`. What do you consider when you select your jumping rule? How efficiently does your sampler explore the posterior distribution?
2. Write code for an Independence Sampler to fit MODEL 1 to `spectral-1.txt`. Construct your jumping rule by using the R command `GLM` to derive an approximation to the posterior distribution. How efficiently does your sampler explore the posterior distribution?
   `GLM` *hint: You can fit the model with* `GLM` *using*
   `glm.fit <- glm( Count ~ I(-log(Energy)),family=poisson(link="log"))`
   You can save the fitted values of (α, β) *into* `fit` *with*
   `fit <- c(exp(glm.fit$coef[1]), glm.fit$coef[2])`
   and compute the asymptotic variance-covariance matrix via the delta method with
   `vmat <- diag(c(fit[1],1))%*%vcov(glm.fit)%*%diag(c(fit[1],1))`.
3. Run the Random Walk Metropolis sampler and the Independence Sampler that you derived in parts 1 and 2 to fit MODEL 1 to `spectral-2.txt`. How does the convergence behavior compare with the two data sets? You may want to reconsider the jumping rules that you settled on in parts 1 and 2 to design a more robust pair of samplers.
4. Write code for a new MCMC sampler that switches between your best Random Walk Metropolis sampler and your best Independence Sampler at each iteration. That is, write a sampler that uses the Random Walk Metropolis update in even numbered iterations and the Independence Sampler in odd numbered iterations. Run this mixed sampler to fit MODEL 1 to each of `spectral-1.txt` and `spectral-2.txt`. How does the mixed sampler compare with the samplers your derived in ran in parts 1, 2, and 3?
5. Based on your findings in parts 1–4. Write an efficient sampler to fit MODEL 3 to the dataset `spectral-4.txt`. You may want to use the method of Data Augmentation to handle the background contamination.
6. *Bonus:* Add five counts to `Counts[200]`, `Counts[201]`, and `Counts[202]` in `spectral-1.txt`. Call the new file `spectral-3.txt`. Now write an MCMC sampler to fit MODEL 2 to `spectral-3.txt`. Use the strategy outlined in the notes to fit the model using Data Augmentation and Metropolis within Gibbs. You might try repeating your run with a weaker or stronger line, or with the low-count spectral model in `spectral-2.txt`.