Homework 7: The Intensity of 19th Century Literature

36-350, Fall 2013

Due at 11:59 pm on Thursday, 31 October 2013

We continue our investigation into the history of British literature, using the data from lab. Refer to the lab solutions.

More subtle tests than the one used in lab indicate that there is some evidence of bursts of genre formation. To further investigate this, we look at more complex models. In an **inhomogeneous Poisson process**, the number of new events during year t follows a Poisson distribution with mean λ_t — λ_t is called the **intensity** during that year. The likelihood is then

$$L(\lambda_1, \lambda_2, \ldots \lambda_n) = \prod_{t=1}^{n} \frac{\lambda_t^{x_t} e^{-\lambda_t}}{x_t!}$$

To keep things simple, we will assume that the intensity is constant over each decade of our data, but can switch from decade to decade. (That is, all years from 1740 to 1749 have one intensity, 1750 to 1759 another, etc.) There are thus 16 decades covered by the data[1].

1. (25) We first find the maximum likelihood estimate of the 16 decades' intensities.
   (a) (10) Suppose that there is one intensity for all time periods. Use calculus to maximize the log-likelihood, and find an algebraic expression for the maximum likelihood estimate in terms of the data. (Do not just copy a book or website.)
   (b) (10) If there are multiple periods, each with their own intensity, show that the log-likelihood adds across periods. Show that the MLE for each period can be found by using your formula from problem 1a on that period's data.
   (c) (5) Report the MLE of λ for each decade, and the code you used to find it.
2. (45) Even if intensities vary over decades, it seems implausible that they would all be very wildly different from each other. We use a **prior distribution** $p(\lambda)$ to express our sense of what we think the intensities ought

[1] For these purposes we'll ignore 1900.

to be like, and then use Bayes's rule to update this. To sample from the resulting **posterior distribution**, we will use the Metropolis algorithm, from lecture 16.

(a) (5) Our prior distribution for λ is a gamma distribution with a shape of 2 and scale 0.1. Plot the density, $p(\lambda)$. (*Hint:* `curve` and `dgamma`.) Is the most-likely uniform rate you found in lab near the peak of the prior?

(b) (5) Write a function, `rinitial`, which takes no arguments, and returns a single draw from the prior distribution. How do you know that this is working?

(c) (10) To use the code from the notes, you will need to write an `rproposal` function. This should take as its argument a value for λ, and return a small random perturbation to it. Write this function, using one of the built-in random variable generators. The intensity must be non-negative, so make sure that the proposal returned is ≥ 0. How do you know this is working?

(d) (10) Write a function, `dposterior`, to calculate the product of the prior density and the likelihood (not the log-likelihood). It should take two arguments, the value of λ and the vector of data, and return a single number. How do you know that it works?

(e) (10) Using the code from lecture[2], your `rinitial` function, your `rproposal` function, and your `dposterior` function, generate a sample of 100,000 draws from the posterior distribution for the decade 1850–1859. Discard the first 10,000 as "burn in", and report the mean and standard deviation of the last 90,000. How different are the mean and standard deviation if you don't discard the first 10,000?

(f) (5) Plot a histogram of the retained draws, and add the *prior* density curve. Do they match? Should they?

3. (30) We now extend Bayesian estimation from one decade to sixteen.

(a) (15) Write a single function (or expression) to return sample from the posterior distribution of intensity *vectors*, i.e., it should return an array of `n` rows and 16 columns, each column being the intensity for a different vector. (You may need to modify the code from lecture, your `rproposal` function or your `dposterior` function.) Generate 100,000 draws, discard the first 10,000 as burn-in, and report the mean and standard deviation for each decade's intensity.

(b) (10) Use `boxplot()` to make a summary display of the posterior distributions for each decade. Compare them to the maximum likelihood intensity estimates. How well do they match? *Hint:* you will find the graph easier to read with `outline=FALSE`.

[2]Which you might need to modify, depending on how you've written your other functions.

(c) (5) Explain what you would have to change in your code to estimate a separate intensity for each year, not just each decade. (You do not have to make the changes.)

4. (20) Extra Credit If the intensity does vary over time, it is plausible that varies smoothly over time, rather than making abrupt jumps. One way to do this is to use a prior distribution in which the different years's intensities are dependent. For instance, we can put a prior distribution on $\log \lambda_t$, where each of them is Gaussian, with mean -1.3 and standard deviation 0.6, and the correlation between $\log \lambda_i$ and $\log \lambda_j$ is $0.7^{|i-j|}$.

   (a) (8) Modify your `dposterior` function so that it takes a vector of *log* intensities, as well as a data vector, and returns the appropriate product of prior density and likelihood. *Hint 1:* Remember to undo the logarithm on the intensities when calculating likelihoods. *Hint 2:* There are several packages with multivariate Gaussian density functions, such as `mvtnorm` and `mixtools`. *Hint 3:* You will need the 160×160 covariance matrix whose i, j^{th} entry is $(0.6)^2 0.7^{|i-j|}$.

   (b) (2) Modify your `rinitial` function so that it returns a vector λ of the appropriate length. *Hint:* There are multivariate normal generators.

   (c) (3) Modify your `rposterior` function so that it takes in a vector and returns a vector. Do you still need to ensure all components are positive? If it's not strictly needed, is it a good idea?

   (d) (7) Combine these steps, modifying `rmetropolis` if need be, to generate 10,000 draws from the posterior distribution. Discard the first 9,000 and plot ten randomly-selected draws on the same plot, connecting points belonging to the same draw visually. Undo the logs so that the results can be compared to your earlier work. Describe what you see, and how it compares to the figure.

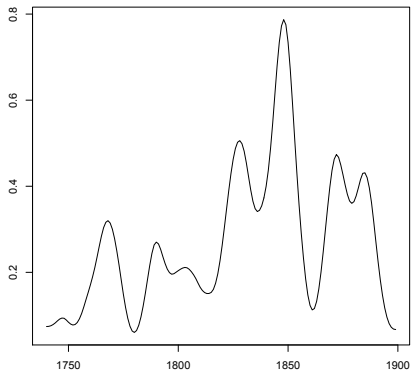

Figure 1: Mystery figure, for those attempting the extra credit.