Statistical analysis of RNA sequencing data
Functional Genomics Assignment 1

Ernest Turro
University of Cambridge

26 Oct 2018

1 Instructions

- The report must be one PDF and an accompanying R source code text file (no other formats are accepted)
- The file names must follow the `fga1_XXX.pdf` and `fga1_XXX.R` convention, where `XXX` is your CRSid (Ernest would save his report as `fga1_et341.pdf` and `fga1_et341.R`)
- The PDF should contain plots and answers to the questions below
- The R code should run without throwing any errors
- Elegance of writing, plotting and coding will be taken into account
- The report must be a maximum of 4 pages long
- This course work comprises 30% of the overall mark for this module

2 Comparison of RNA-seq data from two laboratories

In the second lecture on RNA-seq analysis, you heard about an algorithm called Gibbs sampling. A Gibbs sampler generates random samples from the joint posterior distribution of the parameters in a Bayesian statistical model (introductory documents and examples can be found online, e.g. `http://www.mit.edu/~ilkery/papers/GibbsSampling.pdf`). At each iteration of the algorithm, a sample is drawn from the conditional posterior distribution of a particular parameter, given the values of all other parameters. The objective of the particular Gibbs sampler described during the lecture was to infer parameters representing the expression levels of transcripts. That algorithm, which is an example of a Gibbs sampler, is described in detail in the paper by Turro et al. [1]. For this assignment, you should implement a Gibbs algorithm that aims to fit a different model.

Suppose two laboratories wish to pool their resources to conduct a large experiment. First, they would like to know whether the are any systematic differences between how they generate sequencing data.

To address this question, two different sets of J randomly chosen cell lines are selected without replacement. One set is distributed to one lab and the other set is distributed to the other lab. Within each lab, each cell line is sequenced K times through identical repetitions of the library preparation and sequencing protocol.

Let y_{ijk} denote the read counts for a particular gene for the kth repetition of the sequencing experiment for the jth cell line within lab i. Assume the following hierarchical model:

$$y_{ijk} \sim \text{Pois}(\lambda_{ij}), \tag{1}$$
$$\lambda_{ij} \sim \text{Gamma}(\alpha_i, \beta_i), \tag{2}$$

where $i \in \{1, 2\}$, $j \in \{1, \ldots, J\}$ and $k \in \{1, \ldots, K\}$. Here and throughout, the gamma distribution is parameterised firstly by a shape parameter and secondly by a rate parameter.

- Implement an R function called `simulate.data` that, given α_1, β_1, α_2, β_2, J and K, generates a random vector of read counts from the model above. The elements of the vector should be ordered by experiment ordered within cell line ordered within lab.
- Using the function above, simulate a vector of counts with $\alpha_1 = 100$, $\beta_1 = 2$, $\alpha_2 = 100$, $\beta_2 = 3$, $J = 10$ and $K = 10$, and plot it against the indices of the elements in the vector, using a different point shape for each laboratory and a different point colour for each of the $2 \times J$ experiments.

In order to fit a Bayesian model, prior distributions for parameters need to be specified. For convenience (to ensure conjugacy), let us fix $\alpha_i = 100$ for $i \in \{1, 2\}$ and specify the following prior distribution for β_i:

$$\beta_i \sim \text{Gamma}(a, b), \tag{3}$$

where we set the hyperparameters $a = 0.5$ and $b = 0.1$.

- Plot a set of 100 Poisson mass functions in a 10×10 grid for which the rate parameters have been randomly generated according to the gamma distribution shown in Equation 2, whose parameters have, in turn, been generated by the distribution in Equation 3. Use the same limits for all the x-axes.

The joint posterior mass function of the parameters is given by:

$$P(\lambda, \beta | y) \propto \prod_i P(\beta_i) \prod_j P(\lambda_{ij} | \beta_i) \prod_k P(y_{ijk} | \lambda_{ij}). \tag{4}$$

A Gibbs sampler will allow us to sample from this joint posterior distribution in a simple manner and this will allow us to assess the difference between β_1 and β_2. Because the conjugate prior of the rate parameter of a Poisson distribution is a gamma distribution and the conjugate prior of the rate parameter of a gamma distribution with known shape is a gamma distribution, we can write out the following conditional posterior distributions for the parameters:

$$P(\lambda_{ij}|\lambda_{-ij}, \beta, y) \propto \text{Gamma}\left(\alpha_i + \sum_l y_{ijl}, \beta_i + L\right), \tag{5}$$

$$P(\beta_i|\lambda, \beta_{-i}, y) \propto \text{Gamma}\left(a + \alpha_i J, b + \sum_j \lambda_{ij}\right), \tag{6}$$

where λ_{-ij} denotes all elements of λ except for λ_{ij} and β_{-i} represents β_j such that $i \neq j$.

- Implement an R function called `gibbs.sampler` that runs a Gibbs sampler using the posterior conditionals in Equations 5 and 6 for a pre-specified number of iterations (e.g. 10,000). Set the values for the λs at iteration t conditional on the values of the βs at iteration $t-1$, then set the values of the βs at iteration t using the recently sampled values of the λs for iteration t. You should record the values throughout all iterations. For example, the values for λ need to be recorded in a $2J \times T$ matrix, where T is the number of iterations. You will need to initialise a vector (the first column of the corresponding matrix) of λs and βs with sensible starting values. After the T iterations have been completed, the function should return the two matrices in a list with two components named `lambda` and `beta`.

- Simulate count data with the function `simulate.data` using $\alpha_1 = \alpha_2 = 100$, $\beta_1 = 0.12$, $\beta_2 = 0.1$, $J = 10$ and $K = 10$ and fit the model using `gibbs.sampler`. Make a plot showing the traces of β_1 and β_2 and superimpose horizontal lines indicating the true values.

The mean of a gamma distribution with shape α and rate β is $\frac{\alpha}{\beta}$. We can assess the evidence that there is a systematic difference between the sequencing approaches used by the two labs by comparing the posterior means of the two beta distributions.

- Plot a histogram of the Gibbs trace of $\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}$, after discarding the values for the first 100 iterations.

- What is the posterior probability that lab 2 generates more reads, on average, than lab 1?

- For each pair $(J, K) \in \{10, 100\} \times \{10, 100\}$, plot a pair of histograms depicting the posterior distributions of β_1 and β_2. Comment on the effects of J and K on the posterior variances of these parameters.

3 Adequacy of modelling and experimental design

- Comment on inadequacies, if any, of the experimental design.
- Comment on inadequacies or underlying simplifying assumptions, if any, of the modelling approach.
- How would you combine inferences obtained using the method above across many genes?

References

[1] Turro, E., Su, S.-Y., Goncalves, A., Coin, L. J. M., Richardson, S., and Lewin, A. 2011. Haplotype and isoform specific expression estimation using multi-mapping RNA-seq reads. *Genome Biol*, 12(2):R13.