FGa1: Gibbs Sampling

303034908

November 2018

Comparison of RNA-seq data from two laboratories

Function `simulate.data`

The `simulate.data` function was run with the parameters α_1 = 100, β_1 =2, α_2 = 100, β_2=3, J =10 and K=10. The resulting data was plotted and is shown in figure 1.

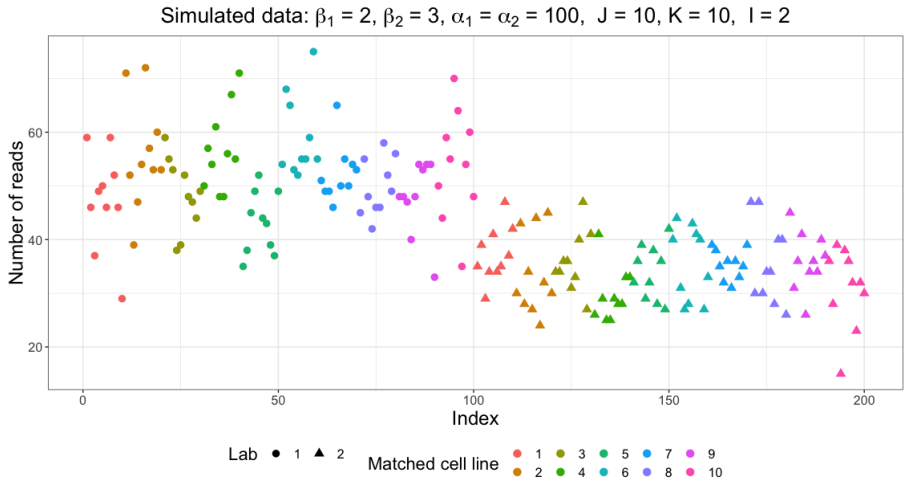

Figure 1: Plot of simulated data ordered by lab by cell line by experiment

One hundred Poisson

100 Poisson mass functions were generated using the equations below, with a=0.5, b=0.1, and α=100.

$$\beta \sim Gamma(a, b) \tag{1}$$

$$\lambda \sim Gamma(\alpha, \beta) \tag{2}$$

$$PoissonDistribution(\lambda) \tag{3}$$

The distributions were plotted with an limit on the x axis of 300 and on the y axis of 0.12 (figure 2) so as to include the highest number of plots given the large variance of lambdas created in equation 2.

Gibbs Sampler

Using a Gibbs sampler (code in appendix), β_1, β_2, λ_1 and λ_2 were evaluated. A plot of the sampling for β_1 and β_2 compared to the real values over the iteration is shown in figure 3.

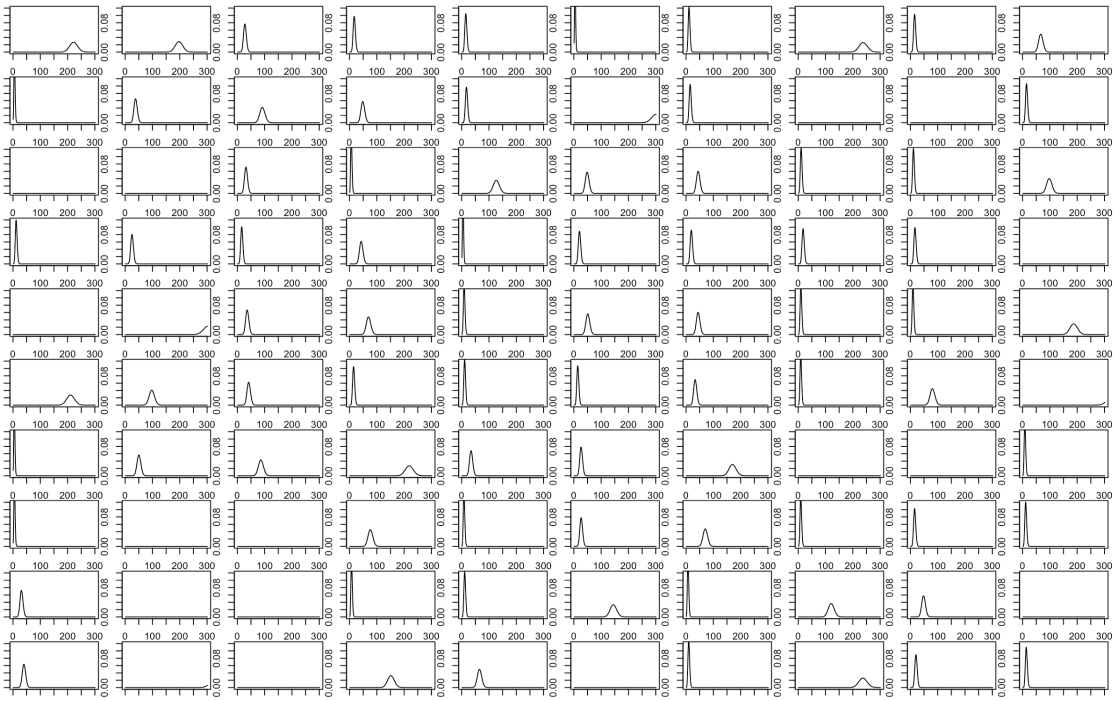

Figure 2: 100 by 100 Poisson distributions

Comparing posterior means of beta distributions

A histogram of $\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}$ is shown in figure 4. Testing normality with Shapiro test came out positive, showing that the difference of the mean posterior follows a normal distribution with mean = -164 and standard deviation = 42.

Two methods were used to evaluate the posterior probability that lab 2 generates more reads than lab 1. First, a Mann Whitney test was performed between $\frac{\alpha_1}{\beta_1}$ and $\frac{\alpha_2}{\beta_2}$ with both α_1 and α_2 = 100 and $beta_1$ and β_2 being the values evaluated from the Gibbs sampler without the burn in of 100 iterations. A positive result to the alternative hypothesis shows that mean posterior distribution of reads for lab 2 is indeed statistically different and greater than lab 1, with a p-value of $2.2 * e^{-16}$.

Secondly, a t-test was performed between $\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}$ and a normal distribution with mean of 0 and same standard deviation of 42. This confirmed the previous results, as the two distributions were statistically different with a p-value of $2.2e^{-16}$.

In both cases, this showed that the posterior probability that lab 2 generates more reads, on average, than lab 1 is $1 - 2.2e^{-16} \sim 1$.

Effects of J vs K

For each pair (J,K) $\in$ {10,100} x{10,100}, a pair of histograms (figure 5) were plotted depicting the posterior distributions of β_1 and β_2. As shown in table 1 increasing J decreased the variance substantially, whilst increasing K slightly increased it.

betas	J=10, K=10	Jx10	Kx10	Jx10 and Kx10
$/beta_1$	$1.29e^{-05}$	- 89.2 %	+5.8%	-89.0 %
$/beta_1$	$1.02e^{-05}$	-90.1%	+5.3%	-89.9%

Table 1: Percentage change in variance with different combiniations of J and K multiplication

Adequacy of modelling and experimental design

Experimental design

The aim of this experiment is to infer if there is a systematic difference how they generate sequencing data. However, to achieve this answer from a number of reads for a specific gene, there are many assumptions that

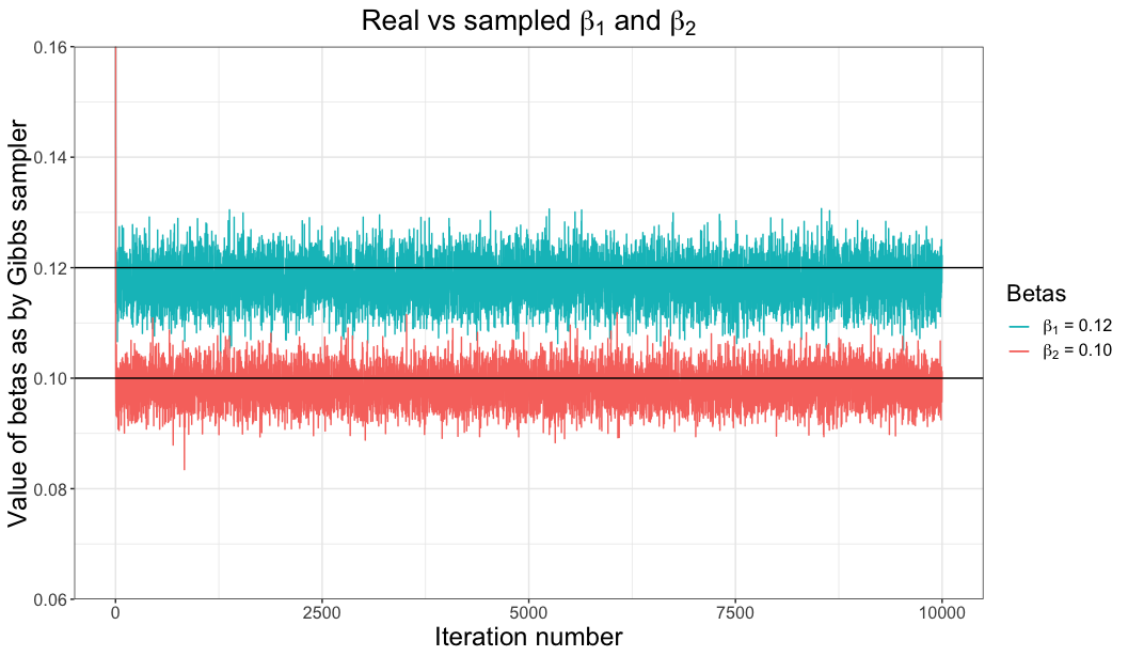

Figure 3: Plot of β_1 and β_2 as by Gibbs sampler vs real expected values

need to be made, and therefore many possible inadequacies to the experimental design.

Firstly, the cell lines sent to each lab were not the same, but rather, "two sets of randomly chosen cell lines", picked without replacement, meaning that full matches of cell lines between labs were not present. Each cell line could have a different gene expression, which is not taken into account in the model above. Moreover, not every cell line may have the same quality annotations for that cell line's reference genome, again affecting the number of reads per gene. Depending on the choice of organism for the final experiment, the results from each cell line could also be weighted so as to give a more selective insight.

Two other major experimental protocols that each lab should have been closely copying are controls and read normalisation. The number of reads per gene will increase with gene length and total number of mapped reads. Thus a Reads Per Kilobase per Million mapped reads (RPKM) normalisation would be needed. However, RPKM assumes that that the total RNA output (unknown) is the same for all libraries. As this may not be true, TMM normalisation or median log deviation normalisation with DESeq should be used ([1]).

Modelling approach

How the priors as constant are decided in the model affects hugely the results. These constants would include the number of iterations for the Gibbs sampler; the hyper parameters **a** and **b**, as well as the choice of alpha; and the 'burn in' period at the beginning of the sampling. If the constants (priors) are the same for all experiment, K, it assumes that gene expression for each cell line will not change over time, which may not be the case if gene expression is dynamic. This may be more or less of an issue depending on how temporally near each repeat was to each other. If the constants are the same for each cell line, J, it assumes that each cell line will have the same expression of that gene. This may also be very unlikely, as different cell lines are different in their nature by differential expression of genes. This may be solved by choosing a house keeping gene which should be expressed similarly across all cell lines.

All of these parameters could be taken into account in the model, even though it would become very high dimensional and computationally heavy. A question to ask then, would be if convergence during the Gibbs sampling could be an issue, and if that convergence is truly a global maximum/minimum or simply a local maximum/minimum.

Finally, even in the simple case as shown above with the underlying assumptions, the Gibbs sampler is still an inaccurate algorithm. The results are themselves distributed with a certain error, and the distribution may be centered around a value quite different from reality. Please see code for further investigation of this. Part of this inaccuracy may be due to the choice of distributions and their intrinsic difference to reality. For example the Poisson is chosen to model read counts per gene. This assumes independence of 'rate' of reads, which may not be correct, and an implication that number of reads corresponds to gene expression. However, the reads for mapped gene are actually coming forming multiple transcripts of that original gene, namely isotypes, haplotypes and multiple splice variants.

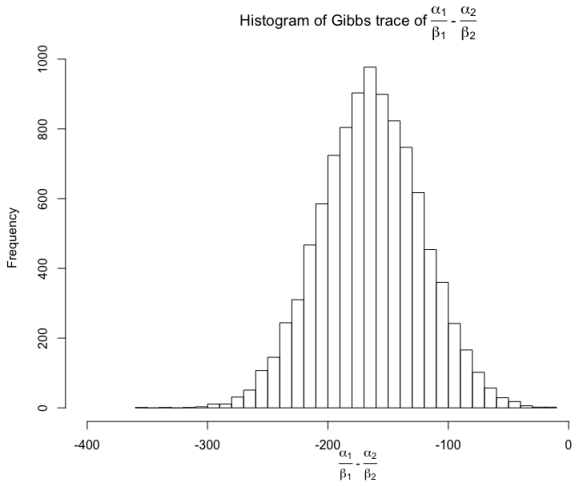

Figure 4: Histogram of $\frac{\alpha_1}{\beta_1} - \frac{\alpha_2}{\beta_2}$
with α_1 and $\alpha_2 = 100$ and β_1 and β_2 taken from the Gibbs sampler

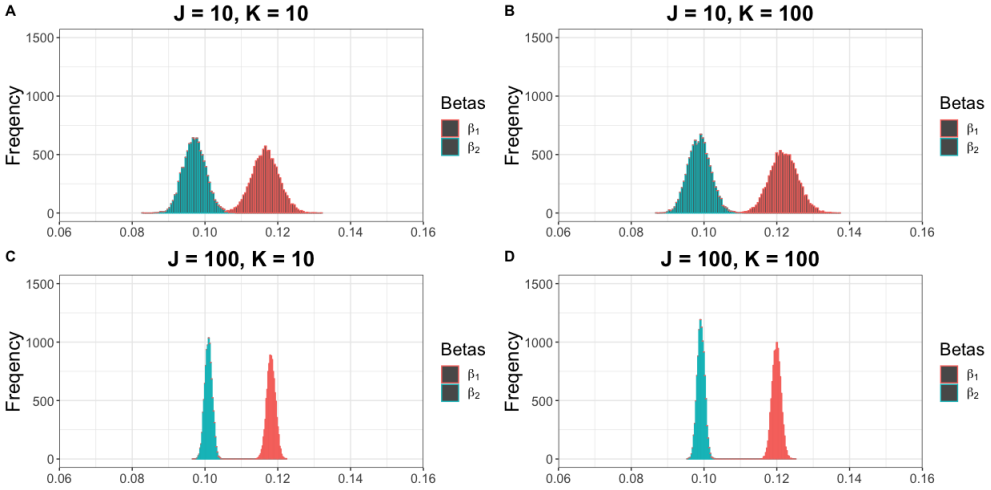

Figure 5: Histograms of β_1 and β_2 from Gibbs sampler with differnet values of J and K

Multiple genes

Experimentally, when comparing number of reads per gene across many genes one would need to correct for the fact that in RNA-seq the number of reads for different genes are not independent. Therefore, when comparing different genes across different cell lines in different labs, difference in reads per gene may not be due to systematic sequencing errors. Moreover, multiple comparisons would require statistical corrections such as Bonferroni. Finally, assuming all correct statistical and experimental corrections are made, evidence for a lab having systematic sequencing errors could be tested by comparing read counts for clusters of genes rather than individual comparisons.

References

[1] Robinson Oshlack 2010: "A scaling normalization method for differential expression analysis of RNA-seq data"