Bioinformatics
doi.10.1093/bioinformatics/xxxxxx
Advance Access Publication Date: Day Month Year
Manuscript Category

Genome analysis

Analyzing results of Bootstraps and Gibbs Samples for Transcription Abundance with Salmon

Xin Cao, Yu Hu

Department of Applied Math & Statistics , Stony Brook University, NY 11790

Abstract

We are provided with 6 pairs of pre-computed transcription abundance data vectors generated by Salmon, and compared the accuracy and efficiency of the two resampling methods (between Bootstrapping and Gibbs Sampling) between each pair. Using different metrics to compare the results from Bootstraps and Gibbs samples, we could know that the later yields expected results in most cases, while Bootstrap provides higher credible level given certain conditions of data sets. Source codes are available at https://github.com/YuHu1/CSE549_Project.

1 Introduction

In study of gene expression, Salmon is developed as a useful tool to get access to the relative abundance of sample transcripts based off of sequencing reads. The de facto way of implementing it is to sequence and quantify directly from RNA-reads (either from reference or de novo assembly). The software uses a Maximum Likelihood (ML) approach to obtain the abundance of a vector of transcripts under which the distribution of reads reaches its ML.

Salmon provides a quicker approach to this ML problem for large number of reads. However the raw result from Salmon only contains one vector of ML estimate, and we have no idea about whether the output is reliable and accurate, say, one small perturbation on the original reads can either switch the estimate slightly or significantly. In other words, we need to verify the robustness of the estimate since the Maximum Likelihood solution is not the global minimum of the function. Salmon has two build-in options of sampling from the posterior distribution- a Bootstrapping approach and a Gibbs sampling method.

Gibbs sampling is one specific case of Metropolis-Hastings algorithm for implementing MCMC (Monte Carlo Markov Chain) sampling. What we get from Gibbs sampling should be an entire probability distribution of abundances. The idea in Gibbs sampling is to generate posterior samples by sweeping through each variable (abundance of each transcript) iteratively through all equivalence classes, to sample from its conditional distribution with the remaining variables fixed to their current values. For instance, consider the set of random variables $\boldsymbol{X} = (X_1, X_2, \cdots, X_N)$. We start by setting these variables to their initial values $x_1^{(0)}, x_2^{(0)}, \cdots, x_N^{(0)}$. At iteration i, we sample

$$\begin{aligned} x_j^{(i)} &\sim p(X_j = x_j | X_1 = x_1^{(i-1)}, \cdots, \\ &\quad X_{j-1} = x_{j-1}^{(i-1)}, X_{j+1} = x_{j+1}^{(i-1)}, \cdots, X_N = x_N^{(i-1)}). \end{aligned}$$

This process continues until the system converges (normed difference between resampling and posterior distribution less than ϵ) or the maximum iterations are attained. Essentially, the Gibbs Sampler tries to estimate the true posterior distribution by randomly drawing samples from it: more samples drawn yields more accurate estimate to the true posterior distribution.

Bootstrapping works differently from Gibbs sampling. Instead of generating an entire distribution of each transcript abundance by iterating through all equivalence classes in one Gibbs sample, it returns a set of point estimates from drawing one abundance value (with replacement) from posterior distributions in one bootstrap. In other words, for original ML estimate $\boldsymbol{X}$, the i-th bootstrap returns a point estimate

$$\begin{aligned} \boldsymbol{X}^{(i)} &= (x(i_1), x(i_2), \cdots, x(i_N)), \\ \boldsymbol{i} &= \{i_1, \cdots, i_N\} \in \text{ permutation group } S_N \end{aligned}$$

with given set of statistical coefficients (for example $\hat{\boldsymbol{P}} = \{\boldsymbol{\Theta}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \cdots\}$) being reassigned per resample to keep robustness. Thus we resample the data many times, getting an ML estimate from every resampling. We can compare these ML estimates with the original output. The Bootstrap gives an idea of how confident in point estimates.

Our goal is to determine which of these methods is a better approach at estimating the true posterior distribution around Maximum Likelihood solution. The term 'better' should be evaluated through multiple criterion, accuracy, mapping rates, as well as efficiency.

2 Implementation

Salmon requires a `fasta` file containing reference transcripts and a (set of) `fasta/fastq` file(s) containing reads to run quantification, which are precomputed by the instructor due to their sizes- exceeding 4GB each. We are provided with the output `sf` file with transcript indexes, (effective) lengths, Transcripts Per Million (TPM), and abundance fc corresponding equivalent classes listed in columns. Besides, through python program of converting bootstrap outputs to `tsv` files, we also obtai the resampled abundance distribution for transcripts through Bootstraps c Gibbs sampler.

2.1 Processing abundance data

Note that the source file of read paragraphs `SRR493366-7` come from human gene paragraphs and could be found o `ftp://ftp.sra.ebi.ac.uk/vol1/fastq/SRR493`. It contain 6 pairs of resampled data (each pair contains BS and GS results), each ha 100 bootstraps or Gibbs samples, with 213,622 equivalence classes. W need to compare the accuracy (or confidence level) of these 100 resample results within pairs, analyzing those quantification. We import the origina `sf` and `tsv` data to Matlab, then do the transpose and export those dat to `csv` files, with identical index with equivalence classes.

2.2 Metrics for accuracy

We tested the accuracy for resamples using mostly measures in section 5.2 of paper (1), with additional one measuring confidence levels.

For the first evaluation we use the width 95% credible interval i ratio to the mean of each Bootstrap/Gibbs sample, in order to measur the confidence level of the two resampling methods. For transcript i an resample $\boldsymbol{X}_i$, that ratio equals

$$T_i = \frac{\theta_U - \theta_L}{\mu(\boldsymbol{X}_i)}$$

for interval $[\theta_L, \theta_U]$, where $\Pr(\theta_L \leq \boldsymbol{X}_i \leq \theta_U) \approx 0.95$. We implemen this measure by extracting the approximate values

$$\theta_L = \frac{1}{2}\left(x_i^{(2)} + x_i^{(3)}\right),\ \theta_U = \frac{1}{2}\left(x_i^{(97)} + x_i^{(98)}\right)$$

from vector of resample abundances sorted in ascending order.

The second measure is proportionality correlation which determine the (linear) proportionality between two vectors, recommended by Lovel et al. in (2). It is defined as

$$\rho_p = \frac{2\mathrm{Cov}\{\log \boldsymbol{x}, \log \boldsymbol{y}\}}{\mathrm{Var}\{\log \boldsymbol{x}\} + \mathrm{Var}\{\log \boldsymbol{y}\}}.$$

For perfect correlation $\rho_p = 1$, and theoretically ρ_p swings between -1 and 1 for all $\boldsymbol{x}$ and $\boldsymbol{y}$. We add a correction constant 10^{-2} to all values when computing ρ_p for each Bootstrap/Gibbs sample, preventing the unexpected values resulted by all 0 abundance for certain transcripts.

Another criterion is mean absolute relative difference (MARD) introduced in (1), computed for each transcript i per one resample (Bootstrap/Gibbs) using average absolute difference:

$$\mathrm{ARD}_i = \begin{cases} 0 & \text{if } x_i = y_i = 0 \\ \frac{2|x_i - y_i|}{|x_i + y_i|} & \text{otherwise} \end{cases}$$

in which x_i and y_i represent the posterior value (true abundance) and resampled value. ARD has an upper bound of 2 and a lower bound of 0, when the resample perfectly match original distribution. We compute the mean value of all transcripts per resample by $\mathrm{MARD} = (1/M) \cdot \sum_{i=1}^{M} \mathrm{ARD}_i$.

3 Results

We compute the measurements in section 2 and collect the running time of Bootstraps & Gibbs samples from `salmon_quant.log` in each directory. For confidence level $\boldsymbol{T}$, we have $N = 213,622 = (\#\text{ of transcripts})$ values per vector; that number reduces to $M = 100$ when comes to MARD and correlation ρ_p. We generate a simulated probability distribution (PDF) function for each vector above, and plot the PDF's of 6 resembling Bootstraps/Gibbs in clusters.

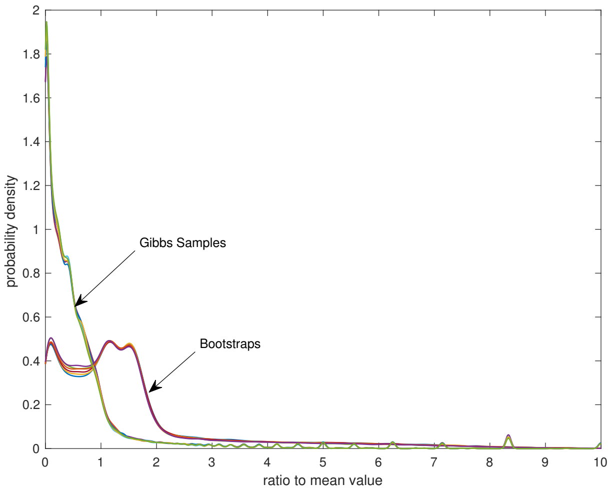

Fig. 1. Probability density for credible level T of all pairs of Bootstraps and Gibbs Samples

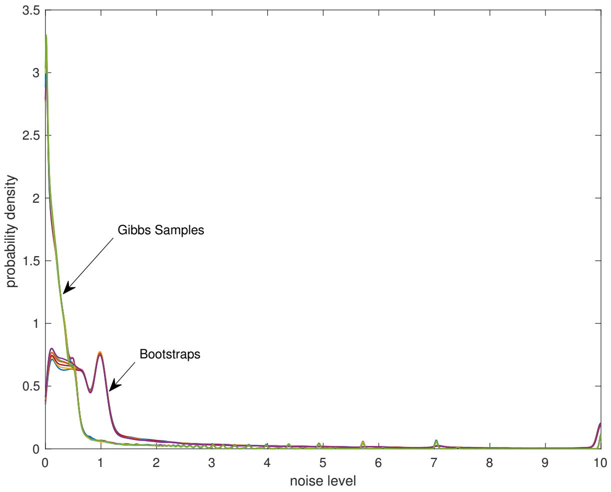

Fig. 2. Probability density for noise of all pairs of Bootstraps and Gibbs Samples

3.1 Accuracy

Figure 1 shows the ratio of 95% confidence interval width against mean abundance, depicted as PDF in two clusters: Bootstraps and Gibbs samples. We observe that the most of Gibbs samples have credible width $\frac{1}{2}(\theta_U - \theta_L)$ less than mean value, while for Bootstraps the ratio distribution extends to approximately 2. We also plot the noise $\mathcal{Z} = \sigma/\mu$ of N transcripts from Bootstraps and Gibbs samples in Figure 2, also in form of PDF. Generally, resamples from Gibbs samples are less noisy.

The result suggests that Gibbs sampling appears to be more accurate than Bootstrapping in confidence level, which means resamples from Gibbs samples are significantly more robust, but the tail probability of Gibbs sampling is larger.

Figure 3 suggests that the proportionality correlations ρ_p of Gibb samples are also appreciably larger than those from Bootstraps. From th derivation below

$$\begin{aligned}\mathrm{Var}(\log(\boldsymbol{x}/\boldsymbol{y})) &= \mathrm{Var}(\log \boldsymbol{x} - \log \boldsymbol{y}) \\ &= \mathrm{Var}(\log \boldsymbol{x}) + \mathrm{Var}(\log \boldsymbol{y}) - 2\mathrm{Cov}(\log \boldsymbol{x}, \log \boldsymbol{y}\end{aligned}$$

we know that the closer ρ_p to 1, the closer $\mathrm{Var}(\log(\boldsymbol{x}/\boldsymbol{y}))$ to 0, henc the stronger ratio correlation exist between $\boldsymbol{x}$ and $\boldsymbol{y}$. The mean level c correlation is $\boldsymbol{\mu}(\rho_p) = 0.9218$ for Bootstraps, and $\boldsymbol{\mu}(\rho_p) = 0.9868$ fo Gibbs samples, thus Gibbs samples are also more accurate than Bootstrap in term of proportionality correlation.

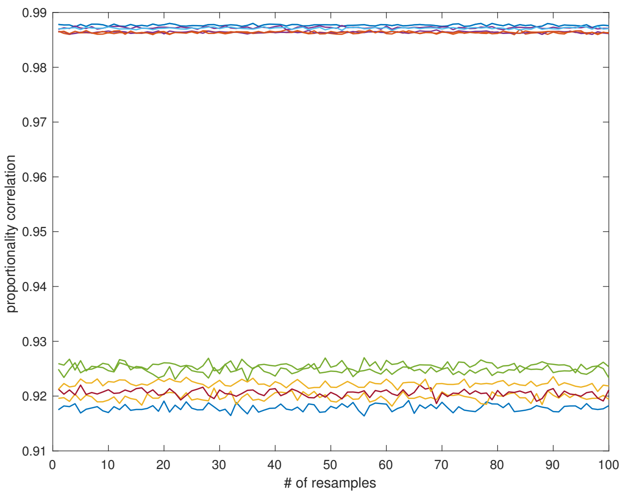

Fig. 3. Proportionality correlation

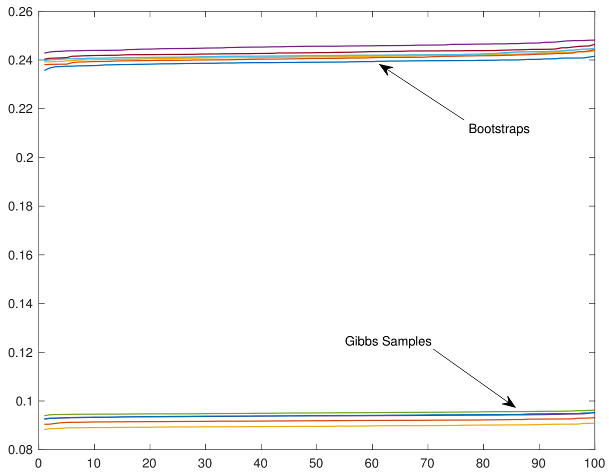

Fig. 4. MARD difference

For MARD difference, one could observe from Figure 4 that results from Bootstraps and Gibbs Samples also fall in clusters (which indicates same resample method for different random reads are highly similar). Gibbs samples have MARD around 0.09, whereas MARD for Bootstraps congregate at approximately 0.24.

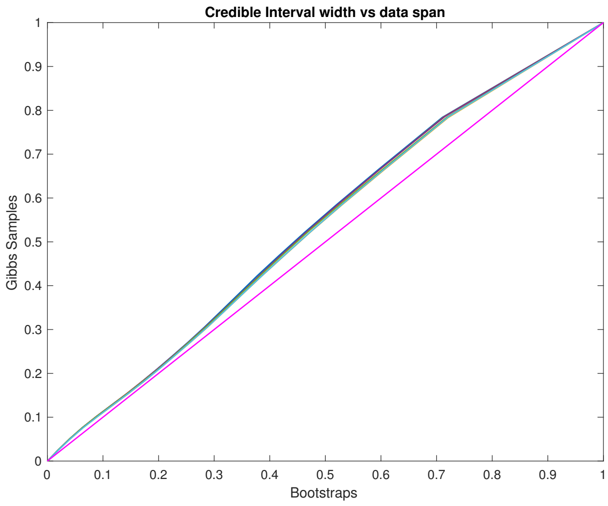

Fig. 5. Average credible levels of BS and GS from 5% to 100%

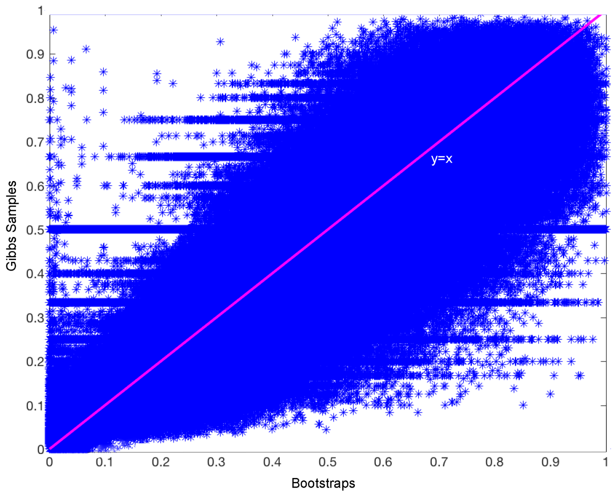

Fig. 6. Credible levels of BS and GS from 5% to 100% for all transcripts

3.2 Running efficiency

We listed the running time (seconds counted from beginning of resamples) of 100 Bootstraps and Gibbs samples for reads `SRR493366-71`, labeled with T_1 to T_6 in Table 1. It is no ambiguity that the efficiency of Gibbs samples are significantly higher than Bootstraps, which has a high probability of being resulted from the reassignment of parameters $\boldsymbol{\mathcal{P}}$ in each resample.

Reads	T_1	T_2	T_3	T_4	T_5	T_6
Bootstraps	240.544	255.576	287.328	265.048	288.952	288.816
Gibbs	7.460	7.811	8.124	7.847	7.952	8.433

Table 1. Time result comparison

3.3 Systematic failures

Systematic errors for either Gibbs sampling or Bootstrapping are unavoidable, and such errors has slight but significant impact on our observed results. We then test the 'failure' case, aka errors that cannot be diminished through improving the accuracy or efficiency of statistical methods, through the credible interval levels of resampled results.

For each reads of Gibbs sample/Bootstrap result $\boldsymbol{X}_i$, we draw credible intervals from 5% to 95% divided by its span, and take the average over all transcripts, denoted as

$$I_\alpha = \frac{1}{N}\sum_i \left(\frac{CI_\alpha}{\max(\boldsymbol{X}_i) - \min(\boldsymbol{X}_i)}\right),$$

where CI_α is the α-level credible interval of $\boldsymbol{X}_i$. Therefore we get two data sets - I_B for Bootstraps and I_G for Gibbs Samples, where $I_B = \{I_{0.05}, I_{0.1}, \cdots, 1\}$. We plot one of them against another below, along with the expected perfect condition that $I_B = I_G$.

From Figure 5 we observe that in term of 'span' of data sets, average credible intervals of posterior samples from Gibbs sampler are slightly wider than that from Bootstraps. That provides us from another aspect that BS has better credible levels than GS given an acceptable range of data. However, the actual width of credible intervals may be higher from BS, but that is probably resulted from the overestimating of data complexity by Bootstraps. We also provide the scattered plot of CI_α from all corresponding transcripts in `SRR493366` in cloud-shaped Figure 6.

References

[1]Patro, R., Duggal, G. and Kingsford, C.. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. bioRxiv 021592; doi: `http://dx.doi.org/10.1101/021592`

[2]Lovell, D., Pawlowsky-Glahn, V., Egozcue, J.J., Marguerat, S. & Bahler, J. Proportionality: a valid alternative to correlation for relative data. BioRxiv, 10.1101/008417 (2014)