Evaluation Project Documentation

Edited by Bo Li

September 7, 2011

1 Introduction

We use the graphical model showed in Figure 1. We denote the observed data as D. We denote $|D|$ as N, number of reads in the data. We assume there are M transcripts in the reference and they are numbered from 1 to M. In addition, our model has an extra "noise" transcript to account for reads coming from background noise, numbered as 0. θ is the probability distribution of a read is sequenced from a particular transcripts. We have $\theta_i = \tau_i l_i, i = 1...M$ and $|\theta| = M + 1$. For details about RSEM model, please see reference[1][2].

We denote an assembly (the reference set used in RSEM's model) as A.

This project's goal is to evaluate which assemble method performs better, given a fixed data set D. That is to say, we want to find a function f, such that given any two assemblers, for their assemblies A_1 and A_2 made from D, we have :

$$f(A_1) > f(A_2) \Leftrightarrow A_1 \text{ is better than } A_2$$

Currently, we have four candidates for f. They are likelihood score, BIC, model evidence by Monte Carlo approximation and model evidence by convex approximation. We want to show that the latter three performs better than the first one. Ideally, we also want to find that the latter two are better than BIC.

In the following four sections, I'll describe the four measures. In addition, I'll omit notation A int all following formulae. We just need to know for all formulae, "given A" is omitted.

2 Loglikelihood

First, pick up θ_{MLE} (MLE means maximum likelihood estimator) :

$$\theta_{MLE} = \underset{\theta}{\operatorname{argmax}} \log P(D|\theta)$$

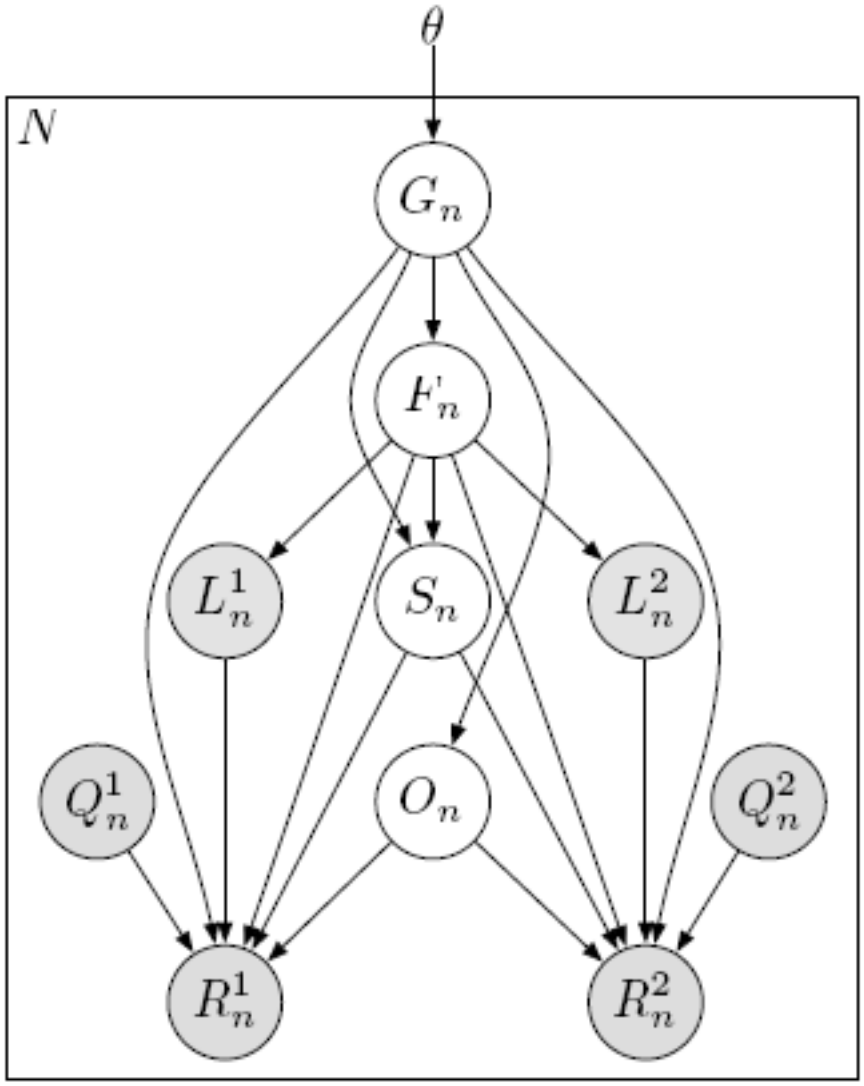

Figure 1: RSEM's graphical model

Then Loglikelihood score is defined as

$$\log P(D|\theta_{MLE})$$

3 Bayesian information criterion

We are interested in $P(D)$, model evidence.

$$P(D) = \int P(D|\theta)p(\theta)d\theta$$

Under certain condition [3], by the Laplace approximation, we have

$$\log P(D) \simeq \log P(D|\theta_{MAP}) + \log P(\theta_{MAP}) + \frac{M+1}{2}\log(2\pi) - \frac{1}{2}\log|H|$$

θ_{MAP} is the Maximum a posteriori estimator. It is the model of the posterior distribution $P(\theta|D)$. H is the Hessian matrix of second derivatives of the negative log posterior at θ_{MAP}.

If we further assume the Gaussian prior, then in the asymptotic case, we have

$$\log P(D) \simeq \log P(D|\theta_{MAP}) - \frac{1}{2}(M+1)\log N$$

The above formula is what we used in this project for BIC. Because we assume θ follows $Dir(1)$, the MAP estimator is the same as MLE estimator.

However, because the "certain condition" is not satisfied here, we do not prefer this measure.

For details, please read P213-P217 of Pattern Recognition and Machine Learning(PRML).

4 Model evidence by Monte Carlo approximation

Our goal is to compute the *model evidence*, $P(D)$. Using Bayes rule, we can express the model evidence as

$$P(D) = \frac{P(D|\theta')P(\theta')}{P(\theta'|D)} \tag{1}$$

Here, θ' can be any particular value of the parameters. For example, we might choose $\theta' = \theta_{PME}$ for numerical issues. PME means posterior mean estimator.The numerator of this fraction is easily computed, as it is simply the product of the likelihood and the prior. The challenge is to compute the denominator, $P(\theta'|D)$. One way to compute this value is via sampling of the latent variables,

Z, from their posterior distribution:

$$\begin{aligned} P(\theta'|D) &= \sum_z P(\theta', z|D) && (2)\\ &= \sum_z P(\theta'|z, D)P(z|D) && (3)\\ &= \sum_z P(\theta'|z)P(z|D) && (4)\\ &\approx \frac{1}{N_s}\sum_{i=1}^{N_s} P(\theta'|z^{(i)}) && (5) \end{aligned}$$

where $z^{(1)}, \ldots, z^{(N_s)}$ are samples from $P(z|D)$, possibly via Gibbs sampling.

After we get $P(D)$, the f is defined as $\log P(D)$.

5 Model evidence by convex approximation

This is another way to approximate $P(D)$ and our goal is to calculate $\log P(D)$ here, too.

Refresh: There are N reads and M transcripts. So $|\theta| = M + 1$ (including the noise transcript).

The data likelihood $\log P(D|\theta)$ can be decomposed as follows:

$$\begin{aligned} \log P(D|\theta) &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{P(Z|D, \theta)} \frac{q(Z)}{q(Z)}\\ &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{q(Z)} + \sum_Z q(Z) \log \frac{q(Z)}{P(Z|D, \theta)}\\ &= F(q, \theta) + KL(q(Z)||P(Z|D, \theta))\\ F(q, \theta) &= \sum_Z q(Z) \log \frac{P(D, Z|\theta)}{q(Z)} \end{aligned}$$

For any given θ^*, let $q(Z) = P(Z|D, \theta^*)$, we have

$$\log P(D|\theta) \geq F(P(Z|D, \theta^*), \theta)$$

In addition, when $\theta = \theta^*$, $logP(D|\theta) = F(P(Z|D, \theta^*), \theta)$ for that $KL(q(Z)||P(Z|D, \theta)) = 0$.

Therefore, assume a dirichlet prior of $\alpha_i = 1$, we have $P(D) \geq \int_\theta p(\theta) e^{F(P(Z|D,\theta^*),\theta)} d\theta$ for any θ^*. We use $\theta^* = \theta_{MLE}$.

Because

$$F(P(Z|D, \theta^*), \theta) = \sum_{i=0}^{M} c_i^* \log \theta_i + \sum_Z P(Z|D, \theta^*) \log \frac{P(D|Z)}{P(Z|D, \theta^*)}$$

We have

$$\begin{aligned} P(D) &\geq \int_\theta p(\theta) e^{F(P(Z|D,\theta^*),\theta)} d\theta \\ &= e^{\sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}} \int_\theta p(\theta) \prod_{i=0}^{M} \theta_i^{c_i^*} d\theta \\ &= e^{\sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)}} \frac{\Gamma(M+1)\prod_{i=0}^{M}\Gamma(c_i^*+1)}{\Gamma(M+1+N)} \end{aligned}$$

So

$$\begin{aligned} \log P(D) &\geq \log\Gamma(M+1) + \sum_{i=0}^{M} \log\Gamma(c_i^*+1) - \log\Gamma(M+1+N) + \sum_Z P(Z|D,\theta^*) \log \frac{P(D|Z)}{P(Z|D,\theta^*)} \\ &= \log\Gamma(M+1) + \sum_{i=0}^{M} \log\Gamma(c_i^*+1) - \log\Gamma(M+1+N) + \sum_{n=1}^{N} \sum_{z_{ni} \in \pi_n^x} P(z_{ni}|r_n,\theta^*) \log \frac{P(r_n|z_{ni})}{P(z_{ni}|r_n,\theta^*)} \end{aligned}$$

To have a better understand of this part, I'd suggest to read P450-P455 of PRML.

6 References

[1] Li, B., Ruotti, V., Stewart, R. M., Thomson, J. A., Dewey, C. N. (2010). **RNA-Seq gene expression estimation with read mapping uncertainty**. *Bioinformatics*, 26(4), 493-500.

[2] Li, B. and Dewey, C. N. **RSEM: accurate quantification from RNA-Seq data with or without a reference genome.** Submitted.

[3] Schwarz, G. (1978). **Estimating the dimension of a model.** *Annals of Statistics* **6**, 461-464.

[4] Bishop, C. M. (2006). **Pattern recognition and machine learning**. *Springer*