Fully Bayesian Analysis of RNA-seq Counts for the Detection of Gene Expression Heterosis

roko

July 2022

1 Introduction

fully Bayesian: any parameter used in the vanilla Bayes formula is itself parameterized (aka there is a hyperprior for every prior parameter)

2 eBayes

empirical Bayes: using the data itself to estimate prior distribution parameters

"Empirical Bayes is an approximation to more exact Bayesian methods- and with the amount of data we have, it's a very good approximation." - David Robinson Ph. D.

eBayes is an approximation b/c prior parameters become point estimates from data instead of optimized parameters in a fully Bayesian approach.

eBayes has drawbacks though of the data is not representative enough of a sample to speak for an entire population prior belief.

SPOILER ALERT:

The paper conclude that eBayes is perfectly fine when hyperparameter estimates are available and accurate, suggesting Method of Moments as one possible way to get those estimates. However, this will only work for distributions who can be defined by their moments.

THIS IS WHY FULLY BAYESIAN MODEL MIGHT BE BENEFICIAL!! NO NEED TO RELY ON A POTENTIALLY INACCURATE PRIOR.

ESS: the minimum size of a set of posterior samples that have the same efficency for posterior density estimation as the samples obtained from an MCMC chain

$ESS = \frac{N}{\sum_{t=-\infty}^{\infty} \rho_t} = \frac{N}{1+2\sum_{t=1}^{\infty} \rho_t}$

Ex: A chain with 2000 samples and an ESS of 100 may not be as good as another chain with 1000 samples but 200 ESS. More samples need not mean more efficient chain.

$\underline{\hat{R}}$: $\frac{\text{total chain sample sd}}{\sqrt{\frac{1}{m}\sum_{i=1}^{m}(\text{chain sample variance})_i}}$

The R-hat is a metric that is evaluated to see if it's close to 1. If it is, then that means the total variance and the average variances for each chain are roughly the same and the chains have all converged on the same posterior.

This metric is a debatable diagnostic as pointed out by Aki Vehtari in 2021 in his paper "Rank-normalization, folding, and localization: An improved Rb for assessing convergence of MCMC" COWRITTEN BY ANDREW GELMAN WHO DESIGNED THE R-HAT DIAGNOSTIC IN THE FIRST PLACE.

I only bring this up because they make the claim that they know convergence happened because most genes had $\hat{R} < 1.1$.

3 Scientific Motivation

phenotype: observable characteristics or traits of an organism

heterosis: the phenomenon where hybrid child plant surpasses each of its inbred parents w.r.t some characteristic

Applications: agriculture yield, extinction prevention, nutrition, lifespan, resistance to chemicals

It is not clear if gene expression heterosis (genetic disparity) is the underlying cause for phenotypic heterosis.

Ex: There may be a genetic component to addiction or sexual orientation, but we can't point to the specific causal gene yet. Same idea here.

Examining gene expression disparities is a good place to start.

Genetic expression heterosis $\overset{?}{\to}$ phenotype heterosis?

Motivation: Provide a measure of (un)certainty for each kind of heterosis considered. This is done utilizing posterior probabilities.

4 Data

In section 5, the authors apply the methodology to a real maize dataset (Iowans amirite).

The simulation - while notationally confusing - actually demonstrates how the authors want to measure posterior probabilities of heterosis.

(GO TO DEMONSTRATION IN END OF PAPER IN TABLET)

5 Model Choice

$$
\begin{aligned}
y_{gn} &\overset{\text{ind}}{\sim} \text{Poisson}(\exp(h_n + \varepsilon_{gn} + X_n \beta_g)) \\
\varepsilon_{gn} &\overset{\text{ind}}{\sim} \text{Normal}(0, \gamma_g^2) \\
\frac{1}{\gamma_g^2} &\overset{\text{ind}}{\sim} \text{Gamma}\left(\frac{\nu}{2}, \frac{\nu\tau}{2}\right) \\
\nu &\sim \text{Uniform}(0, d) \\
\tau &\sim \text{Gamma}(a, b) \\
\beta_{g\ell} &\overset{\text{ind}}{\sim} \text{Normal}(\theta_\ell, \sigma_\ell^2) \\
\theta_\ell &\sim \text{Normal}(0, c_\ell^2) \\
\sigma_\ell &\sim \text{Uniform}(0, s_\ell)
\end{aligned}
$$

Figure 1: Model Utilized

Notice this allows for information to be shared across genes since ALL β_{gl} priors are identical. This has a shrinkage towards the mean effect pulling posterior β_{gl} estimates closer to the mean prior mean *which is shared by all β_{gl} posterior estimates.*

This model borrows information across genes in β and borrows information across replicates with ϵ. Competing methods do not do as much borrowing across replicates.

6 Model Evaluation

The end goal is to obtain posterior probabilities of a particular heterosis occurring.

This is calculated using the posterior estimates of the regression coefficients as follows:

$$\frac{1}{M}\sum_{m=1}^{M} I\left(2\beta_{g2}^{(m)} + \beta_{g4}^{(m)} > 0 \text{ and } 2\beta_{g3}^{(m)} + \rho_{g4}^{(m)} > 0\right)$$

Intuition: Count the number of posterior beta values that yield heterosis in count data.

7 Design

Posterior estimates were collected for all hyperparameters and a subset of $\beta_{g.}$ values. They claim to not have had enough memory to load all posterior estimates back on the the CPU and instead opt to summarize the posterior estimates as a normal distribution using posterior sample mean and variance.

NOTE: I don't actually know if a poisson-(log-normal) mixture likelihood and normal prior result in a normal posterior. Figure 4 certainly seems to suggest so, but I did not look into the algebra.

8 Results

Q: What's the ground truth? How do we know that simulation results are accurate?

A: Simulation studies include a simple mode where betas have known distributions and so heterosis is easy to determine. For real data, posterior estimates don't have a ground truth. Only apply an appropriate shrinkage in the hopes of better generalization.

One such comparison is to an eBayes method called edgeR. They tried to recover the eBayes estimated parameter values from their fully Bayesian model.

Figure S4 show calibration curves (aka) true proportions vs estimated propotion of posterior probability. The curved is smoothed, so the effect of shrinkage is less exaggerated.

Figure S5 shows coverages, showcasing all methods have roughly the same coverage.

Figures S6 and S7 shows the ROC curve when a decision rule is made at each posterior probability threshold. It seems the less control the model has over the ground truth of β, the worse the ROC curve gets, though even the less controlled βs yield good AUCs.

Not really sure how Figures S9 and S10 were generated since I doubt they ran the model for every possible true proportion of heterosis.

9 Takeaways

- The field of heterosis and statistical analysis is incredibly fresh with the authors frequently citing themselves.

- A fully Bayesian model avoids inaccuracies that might come with eBayes and provides a baseline for approximate inference approaches.

- Results seem to line up with eBayes approaches.

- They mention a lot of computation limitations that I think are either not true or will cease to be true in the near future.