Summer 2021

Developing variational Bayesian inference for applications to gene expression data

David Walker

Developing variational Bayesian inference for applications to gene expression data

by

David Cannon Walker

A Creative Component submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Statistics

Program of Study Committee:
Peng Liu, Major Professor
Vivek Roy
Yumou Qiu

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this dissertation/thesis. The Graduate College will ensure this dissertation/thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2021

ABSTRACT

Bayesian hierarchical generalized linear models are intuitively appealing for applications to gene sequencing data. However, they can be computationally costly to fit in high-dimensional settings using standard Markov-chain Monte Carlo methods. Here we explore the use of variational inference techniques to approximate the posterior of Bayesian hierarchical GLMMs for detecting differential expression in RNA-Seq data or differential translational efficiency in Ribo-Seq data. We find that in simulation studies the variational approach is comparable to two common methods for detecting differential expression, and that the variational posterior is close to the Markov-chain Monte Carlo posterior.

1. INTRODUCTION

Next-generation sequencing (NGS) has been a revolutionary tool in genetics and genomics over the past 15 years. By simultaneously increasing throughput and decreasing cost, NGS has allowed the investigation of new phenomena and driven scientific discoveries in genetics, genomics, and epigenomics (Auer and Doerge, 2010). NGS has driven rapid advances in the study of rare genetic diseases, cancer genomics, and noninvasive prenatal testing (Koboldt et al., 2013), and this is only the tip of the iceberg.

The statistical analysis of NGS data presents several interesting challenges. Statistical tools are used throughout the data processing pipeline, to map the raw output of the NGS platforms to a set of read counts (Mitra et al., 2014). Our focus here is on statistical tools to analyze the final product of the pipeline: read counts. At this stage, the high dimension of the data is a critical concern (Mitra et al., 2014). In studies and experiments that utilize NGS, there will be thousands of features that the researchers are interested in — features may be a genes within a single genome, a population of microbes in a shared environment, or something else, depending on the study. In all cases, there will be many more features than biological samples, and it will be necessary to carry out inference for each feature.

Below, we introduce two applications of NGS — RNA-Seq and Ribo-Seq — and the statistical methods under investigation.

RNA-Seq

One application of NGS technology is RNA sequencing (RNA-Seq). RNA-Seq uses NGS technology to analyze the cellular transcriptome by sequencing, mapping, and quantifying the population of transcripts present in a biological sample (Auer and Doerge, 2010). A common question of interest in the analysis of RNA-Seq data is to identify those genes that are

differentially expressed between experimental conditions, by comparing the RNA-Seq counts between conditions.

Ribo-Seq

Ribosome profiling (Ribo-Seq) measures which transcripts are actively translated, by mapping the position of translating ribosomes over the transcriptome (Calviello and Ohler, 2017). Only RNA sequences protected by ribosomes are sequenced. A Ribo-Seq experiment is usually paired with a complimentary RNA-Seq component, that measures the abundance of all transcripts in the sample (Perkins et al., 2019; Ingolia et al., 2009). This makes it possible to control for transcript abundance when analyzing differential translational efficiency between treatment conditions (Ingolia et al., 2009). Paired RNA-Seq and Ribo-Seq counts are produced from the same biological sample, which is separated into Ribosome footprinting (Ribo-Seq) and control (RNA-Seq) samples after several preparatory steps (Ingolia et al. (2009), supplement). A common goal in Ribo-Seq analysis is to identify genes that have differential translational efficiency between treatments.

Bayesian hierarchical models for RNA-Seq

Bayesian hierarchical models are an appealing choice for the analysis of RNA-Seq and Ribo-Seq experiment data, and have been proposed for a variety of applications in this area (Leng et al., 2013; Skelly et al., 2011; Vardhanabhuti et al., 2013). Hierarchical models borrow information across genes, which helps to make up for the relatively small sample sizes in most RNA-Seq or Ribo-Seq experiments (Ritchie et al., 2015).

Bayesian hierarchical generalized linear mixed models (GLMM) are particularly appealing because they can model a wide variety of study designs; Bayesian hierarchial GLMM are flexible enough to represent increasingly complex RNA-Seq studies, that can include random effects, multiple treatment conditions or covariates, and repeated measures (Vestal et al., 2020). For paired Ribo-Seq and RNA-Seq data, a Bayesian hierarchical GLMM can account for the correlation between Ribo-Seq and RNA-Seq counts from the same biological sample with a random block effect.

Variational inference

Fitting fully Bayesian hierarchical models for RNA-Seq data can be practically difficult. The posterior distribution is usually not available analytically, and inference based on sampling from the posterior (as in Markov chain Monte Carlo) is computationally costly because of the high dimension of the posterior in RNA-Seq applications, where the number of parameters of interest usually grows linearly with the number of genes and a typical experiment will include thousands of genes.

Variational inference (VI) is a method (or class of methods) for approximating difficult probability distributions, originally developed in the field of machine learning (Blei et al., 2017). Variational methods have been used to fit high-dimensional Bayesian models in several applications to NGS data, and RNA-Seq data in particular (Ferreira et al., 2018; Hensman et al., 2015; Thorne, 2018). Although these works apply VI methods to NGS data, the applications they consider are distinct from the differential expression and translational efficiency detection of interest to us, and the models they use are not applicable to this problem. Ferreira et al. (2018) explores a variational approach to fitting a latent variable model intended to isolate a technical effect ("dropouts") from the biological signal in single-cell RNA-Seq data; they do not consider an experimental design structure, or attempt to perform hypothesis testing. Hensman et al. (2015) uses VI to fit a model for correctly aligning sequence fragments to genes; they also do not consider experimental design structures. Thorne (2018) does consider a GLM structure, but uses a NB model and a regularized regression approach to attempt to identify a set of 'parent' gene predictors for 'child' gene expression in a time-series experimental setting. None of these methods uses the non-conjugate variational message passing approach we aim to utilize here, but rather employ other approaches to VI.

VI expresses the problem of finding the posterior of a fully Bayesian model as an optimization problem, and then solves the optimization problem numerically. Variational methods can be substantially faster than sampling-based methods for posterior inference (Blei et al., 2017). However, in order to formulate a tractable optimization problem, VI must impose restrictions on

the class of posterior distributions. For this reason, the optimal distribution selected by VI will not be the true posterior, but instead the best approximation to the true posterior that meets the restrictions.

Summary

In the rest of this paper, we first introduce some of the questions of interest in the statistical analysis of RNA sequencing data and ribosome profiling data, and define Bayesian hierarchical models for those two applications. We develop variational algorithms to fit the hierarchical Bayesian model to RNA-Seq and Ribo-Seq data. Finally, we conduct simulation studies to evaluate the quality of the variational fit in comparison to other methods, and discuss problems and topics of future interest.

2. METHODS

In this section, we will present our Bayesian hierarchical models for RNA-Seq and Ribo-Seq, and an outline of the VI algorithm used to fit those models. The VI algorithm is an extension of work in Tan and Nott (2013) and Wand (2014), to capture the hierarchical structure of the NGS data and to incorporate the point-mass mixture prior that is key to the representation of the scientific questions of interest.

2.1 RNA-Seq

RNA-Seq measures the expression of transcriptome features, e.g. gene expression, by enumerating the number of reads mapped to each transcript. The RNA-Seq data is often presented as a table of read counts, with each column corresponding to a biological sample and each row corresponding to a gene (Table 2.1). In experiments with samples from more than one treatment condition or population, a common goal is to detect genes with different mean expression levels between the treatment conditions; in other words, to detect genes that are differentially expressed between conditions (Lorenz et al., 2014; Robinson et al., 2010).

2.1.1 Model 1

For $g = 1, \ldots, G$, $i = 1, \ldots, N$, let y_{gi} represent the read count for gene g, sample i.

$$\begin{aligned} y_{gi} &\sim \text{Pois}(\lambda_{gi}) \\ \log(\lambda_{gi}) &= x_i^T \boldsymbol{\beta}_g \end{aligned} \tag{2.1}$$

For an experiment with two treatment conditions, $\boldsymbol{\beta}_g = (\beta_{g0}, \beta_{g1})^T$ and $x_i = (1, \text{I}(i \in \text{trt } 2))^T$

In order to determine differential abundance, we want to test

$$H_0 : \beta_{g1} = 0 \tag{2.2}$$

against $H_A : \beta_{g1} \neq 0$ for each gene. In most RNA-Seq experiments, there will be a substantial proportion of genes that are not differentially expressed. To reflect this in the model, and to make the model capable testing the sharp null in equation 2.2, we use a mixture of a Gaussian and a point-mass at zero as a prior for each β_{g1}.

$$\begin{aligned} \beta_{g1} &= (1 - D_g)W_g \\ D_g &\overset{iid}{\sim} \text{Bern}(\pi_0) \end{aligned} \tag{2.3}$$

The null hypothesis in equation 2.2 is then equivalent to 2.4 for each gene.

$$H_0 : D_g = 1 \tag{2.4}$$

We use Gaussian priors for the remaining regression parameters, a Gaussian prior for the hierarchical mean parameter, and an Inverse Gamma prior for the variance parameters.

$$\begin{aligned} (\beta_{g0}, W_g)^T &\overset{iid}{\sim} \text{MVN}\left(\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta\right) \\ \boldsymbol{\Sigma}_\beta &= \text{diag}(\sigma^2_{\beta 1}, \sigma^2_{\beta 2}) \\ \sigma^2_{\beta p} &\overset{ind}{\sim} \text{IG}(\alpha_{\beta p}, \gamma_{\beta p}) \\ \boldsymbol{\mu}_\beta &\sim \text{MVN}(0, \sigma^2_\mu \text{I}_2) \end{aligned} \tag{2.5}$$

Table 2.1: Example table of RNA-Seq read counts

	Treatment 1		Treatment 2	
Gene	Sample 1	Sample 2	Sample 1	Sample 2
1	12	18	1	7
2	44	171	45	29
⋮	⋮	⋮	⋮	⋮
5000	191	281	57	86
Total	950,106	1,311,245	558,009	767,510

2.2 Ribo-Seq

Ribo-Seq measures which features of the transcriptome are actively translated. Ribo-Seq is similar to RNA-Seq, but all the transcripts that are not protected by a translating ribosome are removed before read counts are produced (Calviello and Ohler, 2017). Ribo-Seq is not the only method for detecting translation and comes with some extra technical requirements, but it provides very detailed information about translation, including both ribosome abundance and ribosome location on transcripts (Stark et al., 2019).

Like RNA-Seq, Ribo-Seq data can be visualized as a table of counts, with rows for genes and columns for samples. Ribo-Seq read counts can be used in conjunction with RNA-Seq read counts from the same set of samples to identify genes that are translated at different rates (relative to the volume of transcripts) between treatment conditions. This is called differential translational efficiency (see table 2.2 for an example of paired Ribo-Seq and RNA-Seq read counts).

Because paired Ribo-Seq and RNA-Seq read counts are derived from the same biological sample, and undergo several technical preparation steps together before being processed in parallel, they will have some random biological and technical effects in common.

2.2.1 Model 2

For $g = 1, \ldots, G$, $i = 1, \ldots, N$, and $j = 1, 2$, let y_{gij} represent the read count for gene g, sample i, and preparation j, where $j = 1$ corresponds to Ribo-Seq.

$$\begin{aligned} y_{gij} &\sim \text{Pois}(\lambda_{gij}) \\ \log(\lambda_{gij}) &= x_{ij}^T \boldsymbol{\beta}_g + u_{gi} \\ u_{gi} &\overset{iid}{\sim} \text{N}(0, \sigma_u^2) \end{aligned} \tag{2.6}$$

The random effects common to each pair of Ribo-Seq and RNA-Seq counts are represented by the u_{gi}. For an experiment with two treatment conditions, $\boldsymbol{\beta}_g = (\beta_{g0}, \beta_{g1}, \beta_{g2}, \beta_{g3})^T$ and $x_{ij} = (1, \text{I}(i \in T_2),\ \text{I}(j = 2),\ \text{I}(j = 2)\text{I}(i \in T_2))^T$, so that β_{g3} is the parameter representing differential translational efficiency between treatments for gene g.

In order to determine differential translational efficiency, we want to test

$$H_0 : \beta_{g3} = 0 \tag{2.7}$$

against $H_A : \beta_{g3} \neq 0$ for each gene. Analogous to differential expression in Ribo-Seq, a substantial proportion of genes will not exhibit differential translational efficiency in most experiments, so we use a mixture of a Gaussian and a point-mass at zero as a prior for each β_{g3}.

$$\begin{aligned} \beta_{g3} &= (1 - D_g)W_g \\ D_g &\overset{iid}{\sim} \text{Bern}(\pi_0) \end{aligned} \tag{2.8}$$

The null hypothesis in equation 2.7 is then equivalent to 2.9 for each gene.

$$H_0 : D_g = 1 \tag{2.9}$$

We can use Gaussian distributions for the remaining regression parameters, if we are not concerned with testing other sharp nulls. This also simplifies the VI algorithm required to fit the model. We use conjugate Inverse Gamma (IG) prior distributions for the variance parameters of the Gaussian distributions. Let $\boldsymbol{\beta}_g^* = (\beta_{g0}, \beta_{g1}, \beta_{g2})^T$ and $\boldsymbol{u}_g = (u_{g1}, \ldots, u_{gN})^T$.

$$\begin{aligned} (\boldsymbol{\beta}_g^{*T}, W_g, \boldsymbol{u}_g^T)^T &\overset{iid}{\sim} \text{MVN}\left((\boldsymbol{\mu}_\beta, 0_N)^T, \begin{bmatrix} \boldsymbol{\Sigma}_{\boldsymbol{\beta}} & 0 \\ 0 & \sigma_u^2 \text{I}_N \end{bmatrix} \right) \\ \boldsymbol{\Sigma}_\beta &= \text{diag}(\{\sigma_{\beta p}^2\}]),\ p = 1, \ldots, 4 \\ \sigma_u^2 &\sim \text{IG}(\alpha_u, \gamma_u) \\ \sigma_{\beta p}^2 &\overset{ind}{\sim} \text{IG}(\alpha_{\beta p}, \gamma_{\beta p}),\ p = 1, \ldots, 4 \\ \boldsymbol{\mu}_\beta &\sim \text{MVN}(0, \sigma_\mu^2 \text{I}_4) \end{aligned} \tag{2.10}$$

Table 2.2: Example table of paired Ribo-Seq / RNA-Seq read counts

	Treatment 1				Treatment 2			
	Sample 1		Sample 2		Sample 1		Sample 2	
Gene	Ribo	RNA	Ribo	RNA	Ribo	RNA	Ribo	RNA
1	9	7	1	7	12	18	1	7
2	44	65	22	29	4	1100	45	786
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
5000	63	281	57	92	2287	482	37	86
Total	150,106	1,861,201	528,992	2,100,011	950,106	1,311,245	558,009	767,510

2.3 Variational inference algorithm

In this section, we describe a variational inference algorithm to approximate the posterior for model 2. The algorithm to approximate the posterior for model 1 is a simplification of this algorithm.

Notation:

1. $\boldsymbol{y}$ represents all the observed data
2. $\boldsymbol{\theta}$ represents all the parameters in the model
3. $\boldsymbol{\phi}$ represents the collection of hierarchical mean and variance parameters
4. $\boldsymbol{D}$ is the collection $\{D_g\}_{g \leq G}$
5. $p(\boldsymbol{\theta}|\boldsymbol{y})$ represents the true posterior
6. $p(\boldsymbol{y}, \boldsymbol{\theta})$ is the full joint distribution
7. $q(\boldsymbol{\theta})$ represents the variational joint distribution of all parameters in the model
8. $q(\boldsymbol{\phi})$ represents the joint distribution of the hierarchical parameters
9. q_j represents a single factor of the mean-field variational distribution, for sub-vector of parameters $\boldsymbol{\theta}_j$

2.3.1 Overview of Algorithm

The goal of the VI algorithm is to find the variational posterior distribution $q(\boldsymbol{\theta})$ that minimizes the Kullbeck-Liebler divergence to the true posterior: $\text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y}))$. Minimizing the KL divergence is equivalent to maximizing the evidence lower bound, $\mathcal{L}$.

$$\begin{aligned}\text{KL}(q(\boldsymbol{\theta})||p(\boldsymbol{\theta}|\boldsymbol{y})) &= \int q(\boldsymbol{\theta}) \ln \frac{q(\boldsymbol{\theta})}{p(\boldsymbol{y},\boldsymbol{\theta})} d\boldsymbol{\theta} + \ln p(\boldsymbol{y}) \\ &= -\int q(\boldsymbol{\theta}) \ln \frac{p(\boldsymbol{y},\boldsymbol{\theta})}{q(\boldsymbol{\theta})} d\boldsymbol{\theta} - \ln p(\boldsymbol{y}) \\ &= -\mathcal{L} - \ln p(\boldsymbol{y})\end{aligned} \tag{2.11}$$

The distribution $q(\boldsymbol{\theta})$ must satisfy the mean-field assumption in equation 2.12.

$$q(\boldsymbol{\theta}) = q(\boldsymbol{\phi}) q(\boldsymbol{D}) \prod_{g \leq G} q(\boldsymbol{\beta}_g^*, \boldsymbol{u}_g, W_g) \tag{2.12}$$

2.3.2 Conjugate updates

For a factorized distribution, the optimal form for a single factor while holding the other factors fixed is given by equations 2.13 and 2.14 (see either Ch. 10 of Bishop (2006) or Blei et al. (2017) for a derivation). All expectations are with respect to the current variational distributions of the other parameters.

$$\ln(q_j^*) = \underset{i \neq j}{\text{E}}[\ln(p(\boldsymbol{y},\boldsymbol{\theta}))] + \text{constant} \tag{2.13}$$

Taking the exponential of both sides of 2.13 and normalizing gives 2.14.

$$q_j^* = \frac{\exp\left(\underset{i \neq j}{\text{E}}[\ln(p(\boldsymbol{y},\boldsymbol{\theta})]\right)}{\int \exp\left(\underset{i \neq j}{\text{E}}[\ln(p(\boldsymbol{y},\boldsymbol{\theta})]\right) d\boldsymbol{\theta}_j} \tag{2.14}$$

For a parameter with exponential-family prior, that is conjugate to all neighboring distributions in the factor graph of the full joint, the optimal variational distribution will have the same form as the prior (Winn et al., 2005; Tan and Nott, 2013). When the expressions on the

right hand of equations 2.13 and 2.14 have a form that can be evaluated no matter the state of the variational distributions of the other parameters, these expressions provide the basis for an iterative update of $q_j(\boldsymbol{\theta}_j)$. Repeat:

1. Given current values for $q_i,\ \forall i \neq j$, set q_j to the right hand side of 2.14
2. Update $q_i,\ \forall i \neq j$

Below is the update for $q(D)$. Expressions for updating the variational distributions $q(\mu_\beta), q(\sigma^2_\beta)$, and $q(\sigma^2_u)$ are given in the Appendix.

Update for $q(\boldsymbol{D})$

$$\begin{aligned} \ln(q^*(D)) &= \underset{-q(D)}{\mathrm{E}}[\ln(p(\boldsymbol{y}, \boldsymbol{\theta}))] + \text{constant} \\ &= \sum_g [D_g \rho_{g1} + (1 - D_g)\rho_{g2}] + \text{constant} \\ \rho_{g1} &= (\mathrm{E}[\ln(p(\boldsymbol{y_g}|\boldsymbol{\beta}^*_g, \boldsymbol{u}_g, D_g = 1))] + \ln(\pi_0)) \\ \rho_{g2} &= \mathrm{E}[\ln(p(\boldsymbol{y_g}|\boldsymbol{\beta}^*_g, \boldsymbol{u}_g, W_g, D_g = 0))] + \ln(1 - \pi_0) \end{aligned} \tag{2.15}$$

So $q^*(D)$ factors into $\prod_g q^*(D_g)$, each of which is a Bernoulli distribution with mean parameter π^*_g in 2.16.

$$\begin{aligned} \pi^*_g &= \frac{r_{g1}}{r_{g2} + r_{g1}} \\ r_{g1} &= \exp(\rho_{g1}) \\ r_{g2} &= \exp(\rho_{g2}) \end{aligned} \tag{2.16}$$

Looking at this update more closely:

$$\begin{aligned} \pi^*_g &= (1 + \frac{r_{g2}}{r_{g1}})^{-1} \\ &= (1 + \exp(\rho_{g2} - \rho_{g1}))^{-1} \end{aligned} \tag{2.17}$$

This will be close to 0 if the sum of the expected value of the log likelihood and log prior for $D_g = 0$ (corresponding to the alternative hypothesis) is larger than that for $D_g = 1$, under the current variational distribution.

2.3.3 Non-conjugate updates

Since the Poisson distribution used for counts is not conjugate to the Normal priors for the parameters in the linear-predictor, the optimal density produced by evaluating the right hand side of equation 2.13 won't belong to a recognizable family.

One resolution is to impose an additional exponential-family assumption on $q_g(\boldsymbol{\beta}_g^*, W_g, \boldsymbol{u}_g)$, as in equation 2.18. In this case, the exponential-family in 2.18 will be the MVN distribution; $\boldsymbol{\lambda}_g$ and $t(\boldsymbol{\theta}_g)$ will be the natural parameter and sufficient statistic of the MVN.

$$\begin{aligned} \boldsymbol{\theta}_g &:= (\boldsymbol{\beta}_g^*, W_g, \boldsymbol{u}_g)^T \\ q(\boldsymbol{\theta}_g) &= \exp(\boldsymbol{\lambda}_g^T t(\boldsymbol{\theta}_g) - h(\boldsymbol{\lambda}_g)) \end{aligned} \tag{2.18}$$

Knowles and Minka (2011) show the following (in more general terms). For the algorithm that consists of cycling over the steps in algorithm 1, a fixed point occurs when the gradient of $\mathcal{L}$ with respect to each of the $\boldsymbol{\lambda}_g$ is 0. Non-conjugate message passing (Knowles and Minka, 2011) reduces to conjugate message passing (Winn et al., 2005) for every factor where the distributions are conjugate, which motivates the use of conjugate updates wherever possible in the algorithm.

while $\mathcal{L}$ *has not converged* **do**
  for $g = 1, \ldots, G$ **do**
    $\boldsymbol{\lambda}_g \leftarrow \mathcal{V}(\boldsymbol{\lambda}_g)^{-1} \frac{\partial \mathrm{E}[\ln(p(\boldsymbol{y}, \boldsymbol{\theta})]}{\partial \boldsymbol{\lambda}_g}$;
    $\pi_g^* \leftarrow \frac{r_{g1}}{r_{g2}+r_{g1}}$
  end
  for $\boldsymbol{\phi}_j \in \boldsymbol{\phi}$ **do**
    Update $q(\boldsymbol{\phi}_j)$ according to eq. 2.14
  end
end

Algorithm 1: ncvi

Updates for $q(\boldsymbol{\theta}_g)$

For the case where $q(\boldsymbol{\theta}_g)$ is the MVN distribution, Wand (2014) offers a simplified update in terms of the mean and variance parameters. This is reproduced in equation 2.19, where 'vec' denotes the operation defined in Magnus and Neudecker (2019) that maps a matrix to a vector,

and $\mathrm{D}_x s$ denotes the vector of derivatives of s w.r.t. x. The expansion of equation 2.19 for the special case of model 2 is in the Appendix.

$$\begin{aligned} \boldsymbol{\Sigma}_g &\leftarrow (-2 \text{ vec}^{-1}\mathrm{D}_{\text{vec}(\Sigma_g)}\mathrm{E}[\ln p(\boldsymbol{y}, \boldsymbol{\theta})])^{-1} \\ \boldsymbol{\mu}_g &\leftarrow \boldsymbol{\mu}_g + \boldsymbol{\Sigma}_g^{-1}\mathrm{D}_{\mu_g}\mathrm{E}[\ln p(\boldsymbol{y}, \boldsymbol{\theta})] \end{aligned} \tag{2.19}$$

2.4 Multiple testing

Detecting either differential expression (for RNA-Seq) or differential translational efficiency (for paired Ribo-Seq / RNA-Seq) amounts to testing for each gene whether the mean read count is the same across different sets of samples, although the details of how this is tested will be different for each method. Since there will be thousands of genes to test in most experiments, multiple test correction is an important consideration in the statistical analysis of Ribo-Seq and RNA-Seq.

In the literature on RNA-Seq applications, false discovery rate (Benjamini and Hochberg, 1995) control is the preferred method for multiple test correction (Li et al., 2012; Bi and Liu, 2016). In these applications (similar to applications in ecology) it is more important to limit the proportion of type I errors relative to the number of actual discoveries than it is to ensure that the absolute number of type I errors is controlled (Verhoeven et al., 2005).

For models 1 and 2, we can calculate the Bayesian false discovery rate using posterior probabilities that each gene is not differentially expressed. If we have a collection of G posterior probabilities $p_1, \ldots, p_G$, representing the posterior probability that each gene is not differentially expressed, then we control Bayesian FDR at level α by rejecting H_0^g if $1 - p_g > c^*$, where

$$c^* = sup\{c : \widehat{FDR}(c) < \alpha\} \tag{2.20}$$

and

$$\widehat{FDR}(c) = \frac{\sum_g p_g \mathrm{I}(1 - p_g > c)}{\sum_g \mathrm{I}(1 - p_g > c)} \tag{2.21}$$

Then Bayesian FDR controlled at level α is calculated by

$$\widehat{BFDR}(\alpha) = \frac{\sum_g p_g \mathrm{I}(1 - p_g > c^*)}{\sum_g \mathrm{I}(1 - p_g > c^*)} \tag{2.22}$$

3. SIMULATIONS

We present simulation studies to compare the performance of the VI for model 1 to the performance of 'edgeR' (Robinson et al., 2010) and 'baySeq' (Hardcastle and Kelly, 2010), as well as to model 1 fit by MCMC. Additionally, we present simulation studies to compare the VI posterior for model 2 to the multi-group capabilities of 'edgeR' and to the MCMC fit for model 2. Finally, we validate the *accuracy* (Wand, 2014) of the VI posterior against the MCMC posterior for model 2.

3.1 Detecting differential expression

The R-package 'edgeR' implements the statistical methods in Robinson et al. (2010); Robinson and Smyth (2008); McCarthy et al. (2012). These papers assume a negative binomial model for read counts, use empirical Bayes to estimate gene-specific dispersion parameters, and develop an exact test for differential expression under the negative binomial model. McCarthy et al. (2012) extends the methods to function GLMs. 'edgeR' has been one of the most popular methods for differential expression analysis of RNA-Seq data over the last decade, and is a common point of comparison in the literature. 'edgeR' uses frequentist methods for FDR control, like Benjamini and Hochberg (1995).

The R-package 'baySeq' implements an empirical Bayes approach to differential expression analysis described in Hardcastle and Kelly (2010). The method assumes a negative binomial model for read counts and uses empirical Bayes to estimate hierarchical distributions for the collection of gene specific parameters. 'baySeq' uses posterior probabilities for the null and alternative models to test differential expression for each gene. These posterior probabilities can be used to control the Bayesian FDR as described in section 2.

We simulated data sets according to a Poisson and a negative binomial model for read counts. Parameter choices were based on values from Kvam et al. (2012) Simulation 2. To initialize VI ('ncvi'), we estimated the overall mean and gene-specific fixed effects with the 'glm()' function, initialized gene-specific random effects at 0, and initialized variance parameters at their prior mean.

Simulation 1

For our first setting, we simulated from a Poisson model for counts, with each data set containing 5000 genes observed over 8 samples, with 4 samples from each of 2 treatments. Counts for each gene were drawn according to equation 2.1. We set the overall baseline mean to 60, and gene-specific regression parameters were drawn from a multivariate Gaussian distribution with mean $\boldsymbol{\mu}_\beta = (0,0)^T$ and variance parameters $\sigma^2_{\beta 1} = 8, \sigma^2_{\beta 2} = 0.5$. The overall proportion of non-differentially expressed genes was set to $\pi_0 = 0.8$.

For each method, we fit the model to each data set and tested for differential expression. For 'edgeR' and 'baySeq', we ran the method according to the instructions given in the package vignettes for the appropriate experimental setting. Mean FDR and ROC curves are given in figures 3.1a and 3.1b

Adjusting the prior for variance components

Within setting 1, we modified the values for the hyper-priors of the variance parameters, given in table 3.1. FDR and ROC curves are given in figure 3.3.

Table 3.1: Variance parameter settings

Option	$\boldsymbol{\alpha}_\beta$	$\boldsymbol{\gamma}_\beta$	Prior E$(\sigma^2_{\beta 1}, \sigma^2_{\beta 2})$
1	$(2,5)^T$	$(8,2)^T$	$(8, 0.5)$
2	$(5,9)^T$	$(32,4)^T$	$(8, 0.5)$
3	$(17,33)^T$	$(128,16)^T$	$(8, 0.5)$

Mis-matched values for π_0

Since the π_0 parameter is given to the VI algorithm rather than estimated, we tested the fit when the algorithm received an incorrect value for π_0. Parameter settings were the same as

setting 1, with the exception of π_0 which was actually 0.5. The variational algorithm and MCMC fit the model using the incorrect value $\pi_0 = 0.8$, or the correct value. FDR and ROC curves comparing the two options are given in 3.2.

Simulation 2

We simulated data sets according to a negative binomial model for counts, again with 5000 genes and 8 samples. The mean structure for this settings was the same as for setting 1. All genes shared a common dispersion parameter, set to either 0.05 or 0.17. The overall proportion of non-differentially expressed genes was set to $\pi_0 = 0.8$. We fit 'baySeq' and 'edgeR' in the same way as for Simulation 1. Mean FDR and ROC curves are given in figures 3.4.

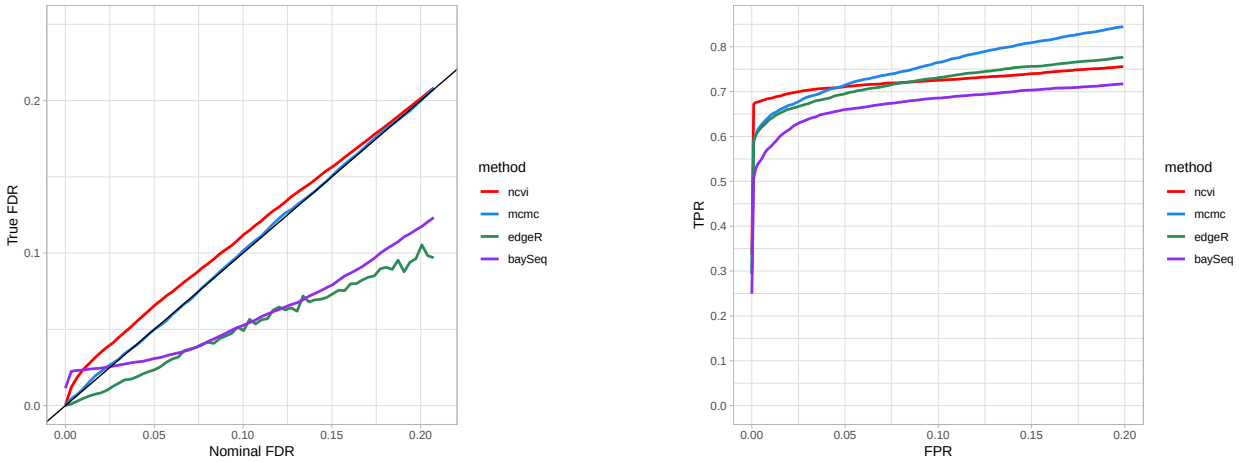
3.2 Detecting differential translational efficiency

Simulation 3

We simulated from a Poisson model for counts, with 5000 genes and 16 samples, this time with a fixed and random effects structure matching equation 2.6 from model 2. The overall baseline mean was $\exp(5) = 148$, and gene-specific regression parameters were drawn from a multivariate Gaussian distribution with $\boldsymbol{\mu}_\beta = (0, 0, 0, 0)^T$, $\sigma^2_{\beta 1} = 8$, $\sigma^2_{\beta 2} = \sigma^2_{\beta 3} = \sigma^2_{\beta 4} = 1/2$. Random effects were drawn from a Gaussian distribution with mean 0 and variance $\sigma^2_u = 2$. The overall proportion of non-differentially expressed genes was set to $\pi_0 = 0.8$. We fit model 2 to each data set using both VI and MCMC; we also fit 'edgeR', which can accommodate the fixed effects structure, although it doesn't handle random effects. Mean FDR and ROC curves are given in figures 3.5a and 3.5b. The nominal FDR for 'edgeR' was always above 0.2, and usually equal to 1 for any number of nulls rejected, so it is not included in the plot. The variational fit is somewhat sensitive to the initial values for the random effects $\{u_{gi}\}$; the best fit achieved is shown below.

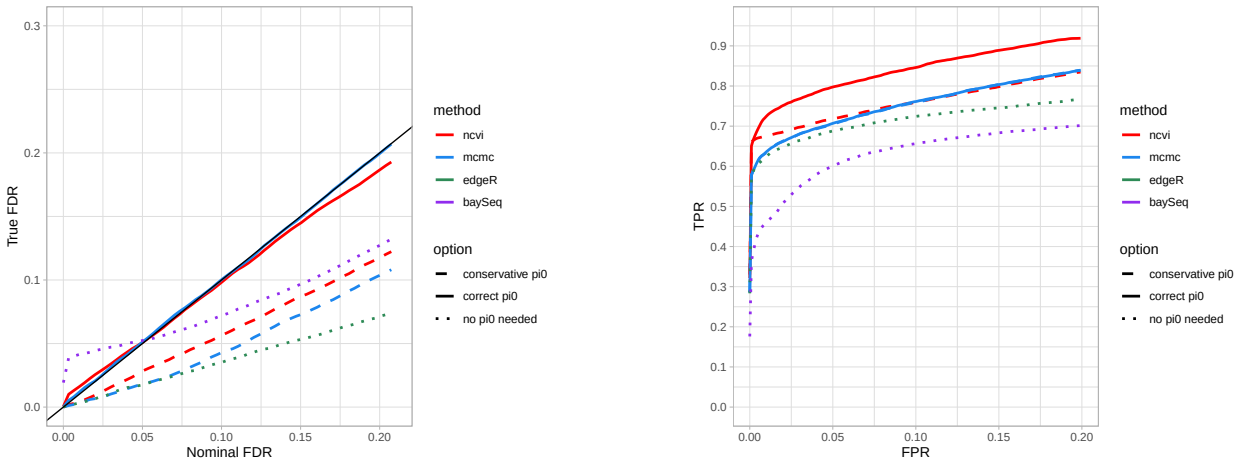
3.3 Accuracy of VI approximation

Since the comparisons of FDR and ROC curves for MCMC and VI above reflect differences in posterior estimates of the D_g, but don't allow for a comparison of the posterior distributions of



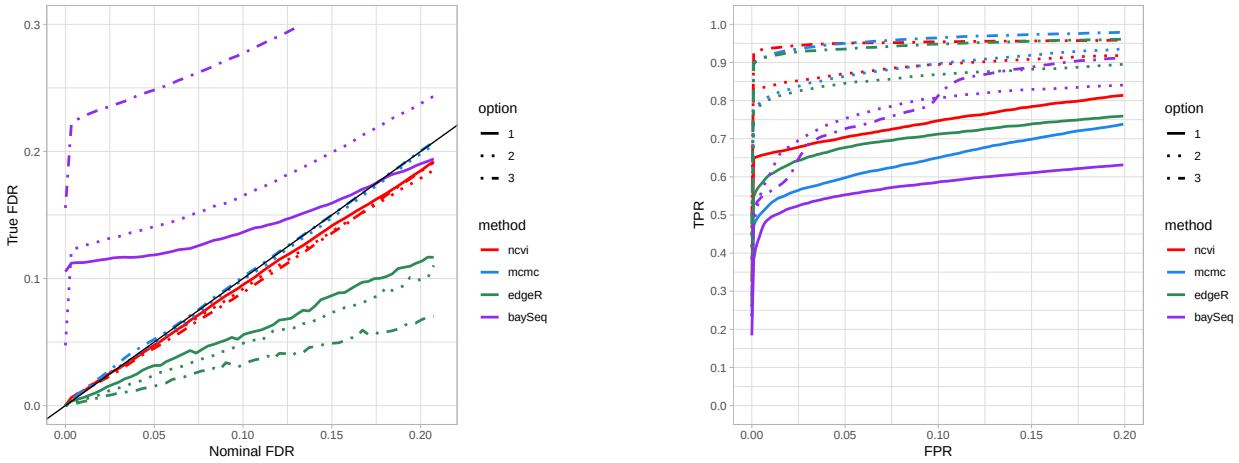
(a) (b)

Figure 3.1: FDR and ROC curves for setting 1



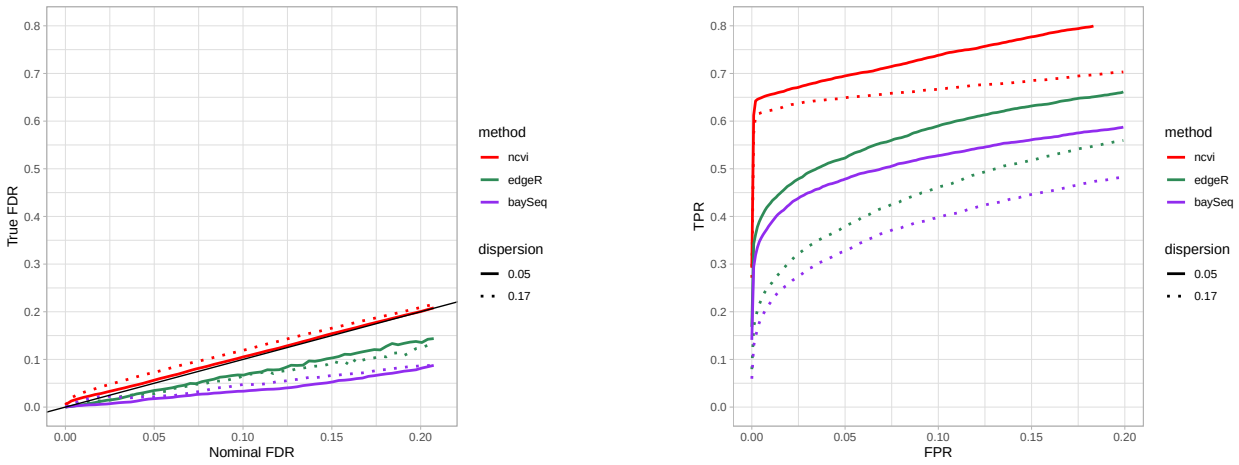
(a) (b)

Figure 3.2: FDR and ROC curves for mis-matched π_0



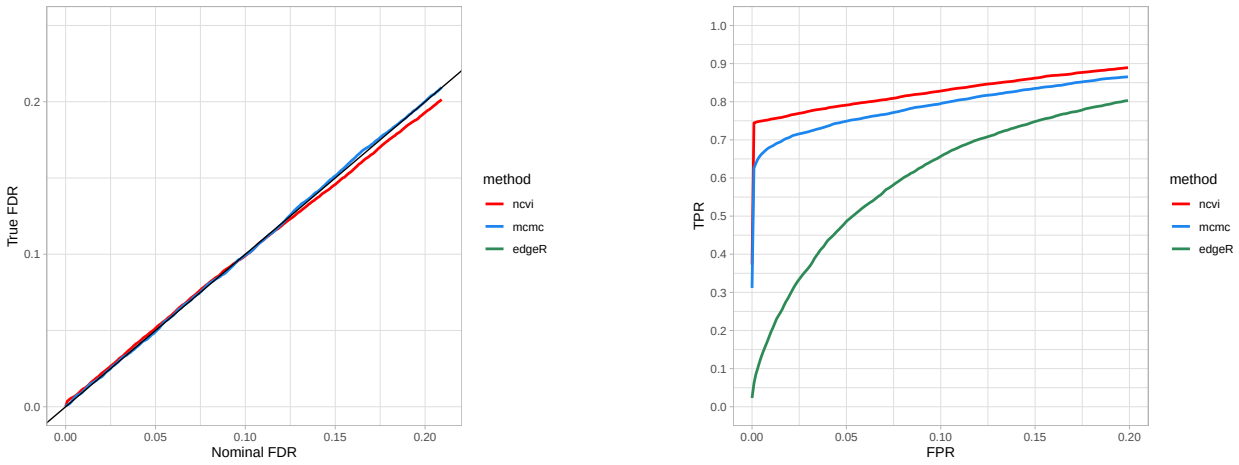
(a) (b)

Figure 3.3: FDR and ROC curves for setting 1 with range of variance hyper-parameters



(a) (b)

Figure 3.4: FDR and ROC curves for setting 2, with disperion 0.05 or 0.17



(a) (b)

Figure 3.5: FDR and ROC curves for setting 3

the continuous parameters, we conducted a simple accuracy analysis as in Wand (2014) to learn how the variational posterior for those parameters compares to the MCMC posterior. After getting both the variational approximation and MCMC posterior, we calculated the accuracy score for each parameter as in equation 3.1. Following Wand (2014), accuracy scores were calculated using a kernel density estimate based on the MCMC posterior to approximate $p(\theta|y)$. Accuracy scores for β_{g0}, β_{g1} were relatively high, with most falling between 60 and 80. Accuracy scores for β_{g2} and W_g were more variable, particularly for the latter, see figure 3.6. The majority of poor accuracy scores were in cases where the posterior probability that $D_g = 1$ was close to 1. There was no clear relationship between poor accuracy scores for the continuous parameters and mis-identification of D_g, see figure 3.7. These scores are somewhat lower than those achieved by Wand (2014), although the model he analyzes is substantially simpler than ours.

$$\text{accuracy}(q^*) = 100\left(1 - \frac{1}{2}\int |q^*(\theta) - p(\theta|y)|d\theta\right)\% \tag{3.1}$$

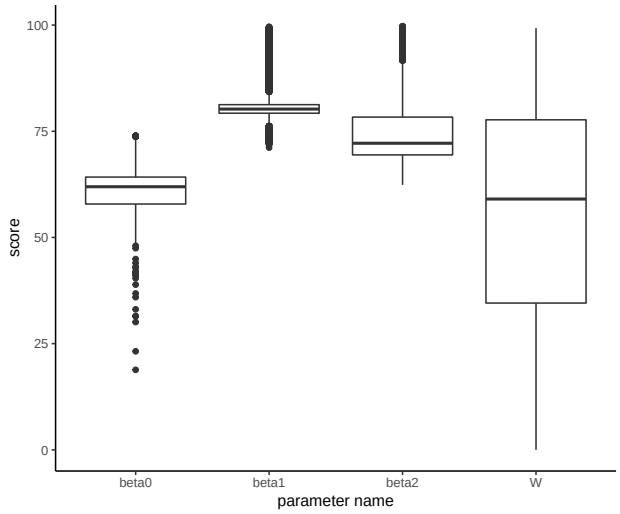
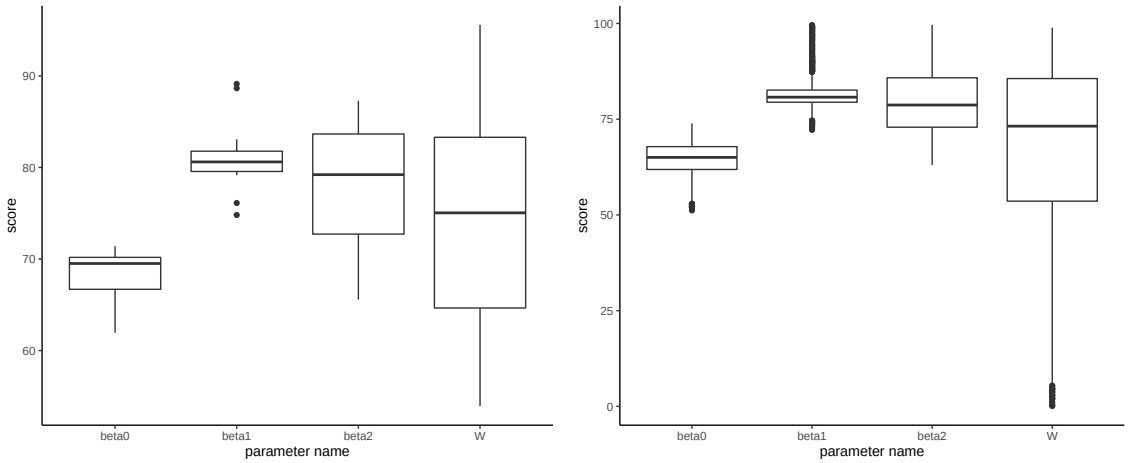


Figure 3.6: Accuracy scores



(a) $D_g = 1$ (b) $D_g = 0$

Figure 3.7: Accuracy scores when D_g was mis-identified

4. DISCUSSION

In simulation studies, the VI algorithm for fitting a Bayesian hierarchical GLM is comparable to commonly used methods for differential expression detection, 'edgeR' and 'baySeq', in terms of false discovery rate control and receiver operating characteristic. This remains true when counts are simulated from a negative binomial distribution, rather than a Poisson distribution as our model assumes. The performance of the VI fit is comparable to, and in some cases slightly better than, the MCMC fit for the same model. The VI algorithm is largely robust to differences in variance hyper-priors, and remains in the same relative position to the other methods in the settings tested; the algorithm does appear sensitive to the value of π_0 that it's passed, becoming noticeably more conservative or liberal when given large or small values respectively. Introducing the ability to estimate π_0 into the algorithm might improve upon this.

For the detection of differential translational efficiency, VI fit of the Bayesian hierarchical GLMM is comparable to the MCMC fit, when given good initial values. VI offers substantial speed improvements over MCMC in this setting, producing an approximate posterior in 2 minutes on average against 45 minutes on average for MCMC, run on a 2020 iMac with 3.3 GHz 6-core Intel Core i5 processor. The accuracy score of the VI posterior relative to the MCMC posterior is sometimes low for the continuous variables, although it is unclear whether or not that is a concern in this application, and may merit future consideration.

Taken together, the results of the simulation studies indicate the potential utility of a VI approach to fitting Bayesian hierarchical GLMM for applications to NGS data. However, VI also has some limitations, and is not the only method that might enhance the practicality of Bayesian hierarchical GLMM in these applications. There are fewer theoretical guarantees established for VI than for MCMC (Blei et al., 2017). However, this is an active area of research, with some encouraging recent developments for Bayesian GLMM (Wang and Blei, 2019). Additionally,

developing a VI algorithm often requires model specific derivation and implementation, which can be time-consuming for the researcher. Some alternatives exist; Kucukelbir et al. (2017) developed an automatic VI algorithm, that can fit the class of differentiable probability models without the help of analytic derivations. Finally, the quality of the VI fit can be sensitive to initial values, particularly in the Ribo-Seq case; the development of a high-quality systematic approach to initialization is likely to be important for the effective use of VI with real data.

Bibliography

Auer, P. L. and Doerge, R. (2010). Statistical design and analysis of rna sequencing data. *Genetics*, 185(2):405–416.

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, 57(1):289–300.

Bi, R. and Liu, P. (2016). Sample size calculation while controlling false discovery rate for differential expression analysis with rna-sequencing experiments. *BMC bioinformatics*, 17(1):1–13.

Bishop, C. M. (2006). Pattern recognition. *Machine learning*, 128(9).

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Calviello, L. and Ohler, U. (2017). Beyond read-counts: Ribo-seq data analysis to understand the functions of the transcriptome. *Trends in Genetics*, 33(10):728–744.

Ferreira, P. F., Carvalho, A. M., and Vinga, S. (2018). Variational inference in probabilistic single-cell rna-seq models. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pages 11–18. Springer.

Hardcastle, T. J. and Kelly, K. A. (2010). bayseq: empirical bayesian methods for identifying differential expression in sequence count data. *BMC bioinformatics*, 11(1):1–14.

Hensman, J., Papastamoulis, P., Glaus, P., Honkela, A., and Rattray, M. (2015). Fast and accurate approximate inference of transcript expression from rna-seq data. *Bioinformatics*, 31(24):3881–3889.

Ingolia, N. T., Ghaemmaghami, S., Newman, J. R., and Weissman, J. S. (2009). Genome-wide analysis in vivo of translation with nucleotide resolution using ribosome profiling. *science*, 324(5924):218–223.

Knowles, D. and Minka, T. (2011). Non-conjugate variational message passing for multinomial and binary regression. *Advances in Neural Information Processing Systems*, 24:1701–1709.

Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K., and Mardis, E. R. (2013). The next-generation sequencing revolution and its impact on genomics. *Cell*, 155(1):27–38.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18(1):430–474.

Kvam, V. M., Liu, P., and Si, Y. (2012). A comparison of statistical methods for detecting differentially expressed genes from rna-seq data. *American journal of botany*, 99(2):248–256.

Leng, N., Dawson, J. A., Thomson, J. A., Ruotti, V., Rissman, A. I., Smits, B. M., Haag, J. D., Gould, M. N., Stewart, R. M., and Kendziorski, C. (2013). Ebseq: an empirical bayes hierarchical model for inference in rna-seq experiments. *Bioinformatics*, 29(8):1035–1043.

Li, J., Witten, D. M., Johnstone, I. M., and Tibshirani, R. (2012). Normalization, testing, and false discovery rate estimation for rna-sequencing data. *Biostatistics*, 13(3):523–538.

Lorenz, D. J., Gill, R. S., Mitra, R., and Datta, S. (2014). Using rna-seq data to detect differentially expressed genes. In *Statistical analysis of next generation sequencing data*, pages 25–49. Springer.

Magnus, J. R. and Neudecker, H. (2019). *Matrix differential calculus with applications in statistics and econometrics*. John Wiley & Sons.

McCarthy, D. J., Chen, Y., and Smyth, G. K. (2012). Differential expression analysis of multifactor rna-seq experiments with respect to biological variation. *Nucleic acids research*, 40(10):4288–4297.

Mitra, R., Gill, R., Datta, S., and Datta, S. (2014). Statistical analyses of next generation sequencing data: an overview. *Statistical Analysis of Next Generation Sequencing Data*, pages 1–24.

Perkins, P., Mazzoni-Putman, S., Stepanova, A., Alonso, J., and Heber, S. (2019). Ribostreamr: a web application for quality control, analysis, and visualization of ribo-seq data. *BMC genomics*, 20(5):1–9.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for rna-sequencing and microarray studies. *Nucleic acids research*, 43(7):e47–e47.

Robinson, M. D., McCarthy, D. J., and Smyth, G. K. (2010). edger: a bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics*, 26(1):139–140.

Robinson, M. D. and Smyth, G. K. (2008). Small-sample estimation of negative binomial dispersion, with applications to sage data. *Biostatistics*, 9(2):321–332.

Skelly, D. A., Johansson, M., Madeoy, J., Wakefield, J., and Akey, J. M. (2011). A powerful and flexible statistical framework for testing hypotheses of allele-specific gene expression from rna-seq data. *Genome research*, 21(10):1728–1737.

Stark, R., Grzelak, M., and Hadfield, J. (2019). Rna sequencing: the teenage years. *Nature Reviews Genetics*, 20(11):631–656.

Tan, L. S. and Nott, D. J. (2013). Variational inference for generalized linear mixed models using partially noncentered parametrizations. *Statistical Science*, 28(2):168–188.

Thorne, T. (2018). Approximate inference of gene regulatory network models from rna-seq time series data. *BMC bioinformatics*, 19(1):1–12.

Vardhanabhuti, S., Li, M., and Li, H. (2013). A hierarchical bayesian model for estimating and inferring differential isoform expression for multi-sample rna-seq data. *Statistics in biosciences*, 5(1):119–137.

Verhoeven, K. J., Simonsen, K. L., and McIntyre, L. M. (2005). Implementing false discovery rate control: increasing your power. *Oikos*, 108(3):643–647.

Vestal, B. E., Moore, C. M., Wynn, E., Saba, L., Fingerlin, T., and Kechris, K. (2020). Mcmseq: Bayesian hierarchical modeling of clustered and repeated measures rna sequencing experiments. *BMC bioinformatics*, 21(1):1–20.

Wand, M. P. (2014). Fully simplified multivariate normal updates in non-conjugate variational message passing. *Journal of Machine Learning Research*.

Wang, Y. and Blei, D. M. (2019). Frequentist consistency of variational bayes. *Journal of the American Statistical Association*, 114(527):1147–1161.

Winn, J., Bishop, C. M., and Jaakkola, T. (2005). Variational message passing. *Journal of Machine Learning Research*, 6(4).

APPENDIX. Additional results

Variational updates

Update $q(\boldsymbol{\mu}_\beta)$

$$\begin{aligned}
\ln(q^*(\boldsymbol{\mu}_\beta)) &= \underset{-q(\mu_\beta)}{\mathrm{E}}[\ln(p(\boldsymbol{y},\boldsymbol{\theta}))] + \text{constant} \\
&= \mathrm{E}\left[\sum_g \ln(p(\boldsymbol{\beta}_g^*, W_g|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)) + \ln(p(\boldsymbol{\mu}_\beta))\right] + \text{constant} \\
&= \sum_g \left[\boldsymbol{\mu}_g^T (E\boldsymbol{\Sigma}_\beta^{-1})\boldsymbol{\mu}_\beta - \frac{1}{2}\boldsymbol{\mu}_\beta^T (E\boldsymbol{\Sigma}_\beta^{-1})\boldsymbol{\mu}_\beta\right] - \frac{P}{2}(E\sigma_0^{-2})\boldsymbol{\mu}_\beta^T\boldsymbol{\mu}_\beta + \text{constant} \\
&= -\frac{1}{2}(\boldsymbol{\mu}_\beta - M)^T R^{-1}(\boldsymbol{\mu}_\beta - M) + \text{constant} \\
R &= G(E\boldsymbol{\Sigma}_\beta^{-1}) + (E\sigma_0^{-2})I_P \\
M &= R[\sum_g (E\boldsymbol{\Sigma}_\beta^{-1})\boldsymbol{\mu}_g]
\end{aligned} \tag{.1}$$

Update $q(\sigma_{\beta p}^{-2})$ (update for $q(\sigma_u^{-2}$ is essentially the same form)

Let $\tau_{\beta p} = \sigma_{\beta p}^{-2}$

$$\begin{aligned}
\ln(q^*(\tau_{\beta p})) &= \underset{-q(\tau_{\beta p})}{\mathrm{E}}[\ln(p(\boldsymbol{y},\boldsymbol{\theta}))] + \text{constant} \\
&= \mathrm{E}\left[\sum_g \ln(p(\boldsymbol{\beta}_g^*, W_g|\boldsymbol{\mu}_\beta, \boldsymbol{\Sigma}_\beta)) + \ln(p(\tau_\beta))\right] + \text{constant} \\
&= -\frac{1}{2}\left[\gamma_p + \sum_g E(\beta_{gp} - \mu_{\beta p})^2\right]\tau_{\beta p} + (G + \alpha_p - 1)\ln\tau_{\beta p} + \text{constant}
\end{aligned} \tag{.2}$$

Update $q_g(\boldsymbol{\beta}_g^*, W_g, \boldsymbol{u}_g)$

$$\begin{aligned}
\boldsymbol{\Sigma}_g &\leftarrow (-2\ \mathrm{vec}^{-1}\mathrm{D}_{\mathrm{vec}(\Sigma_g)}\mathrm{E}[\ln\ p(\boldsymbol{y},\boldsymbol{\theta})])^{-1} \\
\boldsymbol{\mu}_g &\leftarrow \boldsymbol{\mu}_g + \boldsymbol{\Sigma}_g^{-1}\mathrm{D}_{\mu_g}\mathrm{E}[\ln\ p(\boldsymbol{y},\boldsymbol{\theta})]
\end{aligned} \tag{.3}$$

If $s = \mathrm{E}[\ln\, p(\boldsymbol{y}, \boldsymbol{\theta})]$, then:

$$
\begin{aligned}
\mathrm{vec}^{-1}(\mathrm{D}_{\mathrm{vec}(\Sigma_g)} s) &= -\frac{1}{2}[(E\boldsymbol{\Sigma}_{\beta}^{-1}) + (1-\pi_g) C^T \mathrm{diag}(\exp[A_1]) C + \pi_g C_0^T \mathrm{diag}(\exp[A_0]) C_0] \\
\mathrm{D}_{\mu_g} s &= (E\boldsymbol{\Sigma}_{\beta}^{-1})(M - \boldsymbol{\mu}_g) + (1-\pi_g) * C^T (y_g - \exp[A_1]) + \pi_g C_0^T (y_g - \exp[A_0]) \\
A_1 &= C\boldsymbol{\mu}_g + \frac{1}{2}\mathrm{diag}(C\boldsymbol{\Sigma}_g C^T) \\
A_0 &= C_0\boldsymbol{\mu}_g + \frac{1}{2}\mathrm{diag}(C_0\boldsymbol{\Sigma}_g C_0^T)
\end{aligned}
\tag{.4}
$$

Where $C = [X, Z]$ and C_0 is C with the column corresponding to β_{g3} set to 0.