DIFFERENTIAL COUNT ANALYSIS WITH A TOPIC MODEL

PETER CARBONETTO*

1. Motivation, and overview of methods. The aim of this document is to derive, from first principles, a method for analysis of differential gene expression using a topic model (also known as "grade of membership model" [3]). This method may have other uses — say, to identify "key words" in a topic modeling analysis of text documents — but since our main motivation is analysis of gene expression data, we describe the methods with that application in mind.

Topic modeling brings a new twist to analysis of differential gene expression: In conventional differential expression analysis, each sample is assigned to a single condition; in topic modeling, the assignments to each topic are *proportional.* This needs to be accounted for in developing the new methods for differential expression analysis.

For motivation, we begin with the "log-fold change" statistic commonly used in microarray and RNA sequencing experiments to quantify expression differences between two conditions (e.g., [2, 6]). The log-fold change for gene j and condition k is a ratio of two conditional expectations,

$$\beta_{jk} \equiv \log_2 \frac{E[\,x_j \,|\, \text{condition} = k\,]}{E[\,x_j \,|\, \text{condition} \neq k\,]}, \tag{1}$$

where x_j is the measured expression level of gene j.[1] The way in which the expression level x_j is defined can lead to different log-fold change statistics. For example, several popular methods for analyzing differential expression compare changes in *relative* expression (say, relative to total expression in a cell) [1, 4, 5, 7]. As we will see, a topic modeling perspective provides a natural way to analyze differential gene expression when comparing either relative or absolute expression levels.

2. A binomial model. To develop the methods, we begin with a simple binomial model that predicts expression of a single gene given the proportional assignments to a topic:

$$x_i \sim \text{Binom}(s_i, \pi_i). \tag{2}$$

Here, x_i is the expression level of the target gene in sample i, s_i is the total expression in sample i, and $\text{Binom}(n, \theta)$ denotes the binomial distribution with n trials and success probability θ. In this simple model, the binomial probabilities are defined as

$$\pi_i = (1 - q_i)p_0 + q_i p_1, \tag{3}$$

where $q_i \in [0, 1]$ is the (known) proportion of sample i that is attributed to the topic, and $p_0, p_1 \in [0, 1]$ are two unknowns to be estimated.[2]

*Dept. of Human Genetics and the Research Computing Center, University of Chicago, Chicago, IL

[1] Defining x_{jk} as the total gene expression for gene j among all samples (expression profiles) in condition k, x_j as the total gene expression for gene j in all samples, n_k as the number of samples in condition k, and n as the total sample size, the log-fold change can be computed as $\beta_{jk} = \log_2 \left\{ \frac{x_{jk}}{x_j - x_{jk}} \times \frac{n - n_k}{n_k} \right\}$.

[2] This is a special case of the *linear probability model* in econometrics called [8]. The linear probability model would be typically written as $\pi_i = b_0 + q_i b$, where $b_0 = p_0$ and $n = p_1 - p_0 \in [-1, +1]$. This is a regression model for the binomial probability π_i, in which π_i increases linearly with topic proportion q_i.

Statistical inference with this simple binomial model implements analysis of differential gene expression. In particular, $\beta \equiv \log_2(p_1/p_0)$ is the *log-fold change in relative expression.* To show that this is so, consider the following statistical process for generating the counts $x_1, \ldots, x_n$:

- for $i = 1, \ldots, n$
  1. for $t = 1, \ldots, s_i$
     (a) Sample a topic, $z_{it} \sim \text{Binom}(1, q_i)$.
     (b) Sample a gene, $w_{it} \,|\, z_{it} \sim \begin{cases} \text{Binom}(1, p_1) & \text{if } z_{it} = 1 \\ \text{Binom}(1, p_0) & \text{otherwise.} \end{cases}$
  2. Generate the final gene count, $x_i \leftarrow w_{i1} + \cdots + w_{is_i}$.

In this statistical process, $q_i = p(z_{it} = 1)$ is the topic probability, $p_1 = p(w_{it} = 1 \,|\, z_{it} = 1)$ is the conditional probability that the gene is expressed given membership to the topic, and $p_0 = p(w_{it} = 1 \,|\, z_{it} = 0)$ is the probability that the gene is expressed when not belonging to the topic. Therefore, we have

$$\beta \equiv \log_2 \frac{p_1}{p_0} = \log_2 \frac{p(w_{it} = 1 \,|\, z_{it} = 1)}{p(w_{it} = 1 \,|\, z_{it} = 0)}. \tag{4}$$

This is the log-fold change statistic for relative expression given proportional topic assignments $q_1, \ldots, q_n$. The binomial model (2) can in fact be derived from this statistical process (proof not shown). Therefore, estimating p_0, p_1 for the binomial model (2) provides an estimate of the log-fold change (4).

Before continuing, we point out that the s_i's, in practice, do not need to be the total expression for each sample i (*i.e.*, the total counts); for example, some researchers have suggested setting s_i to some pre-defined quantile of the sample's nonzero count distribution (e.g., [1]). This approach has been motivated in settings where a small number of genes are much more highly expressed than the others, and therefore these few genes have a large effect relative expression levels. In short, the binomial model is quite flexible, and can accommodate different relative differential expression analyses defines the s_i's. The only constraint on s_i is that it cannot be smaller than x_i.

3. A Poisson model. The Poisson likelihood with rate $\lambda = n\theta$ closely approximates the Binomial likelihood $\text{Binom}(n, \theta)$ when θ is small and n is large. (The Poisson arises as the limiting distribution of the binomial as $n \to \infty, \pi \to 0, n\theta \to \lambda$.) This is usually the case in gene expression studies; the total gene expression is large, whereas the contribution of each gene is small, even for genes with the highest levels of expression. This suggests a Poisson model of expression x_i given topic proportions q_i:

$$x_i \sim \text{Poisson}(t_i \lambda_i), \tag{5}$$

in which the Poisson rates are defined as

$$\lambda_i = (1 - q_i) f_0 + q_i f_1, \tag{6}$$

and the two unknowns to be estimated are $f_0, f_1 \geq 0$. Observe that, unlike the binomial model, the unknowns for the Poisson model are not constrained to be in $[0, 1]$, and so they do not necessarily represent probabilities. However, by setting $t_i = s_i$, we have that

$$f_0 \approx p_0$$
$$f_1 \approx p_1,$$

when p_0, p_1 are close to zero and all s_i are large. (See the Appendix for an alternative motivation of the Poisson model.)

In the next section, we derive the mathematical expressions needed to implement differential count analysis based on the Poisson model.

4. Poisson model derivations. To derive an algorithm for computing MLEs of f_0, f_1 in the Poisson model (5), we begin with the log-likelihood,

$$\ell(f_0, f_1) \equiv \log p(x \mid f_0, f_1) = \sum_{i=1}^{n} x_i \log(t_i \lambda_i) - t_i \lambda_i + \text{const}, \tag{7}$$

in which the "const" captures all terms that do not depend on f_0 or f_1. A useful identity is that the partial derivative of the log-likelihood with respect to λ_i is

$$\frac{\partial \ell}{\partial \lambda_i} = \frac{x_i}{\lambda_i} - t_i.$$

Making use of this result, the partial derivatives of the log-likelihood with respect to the model parameters f_0 and f_1 are

$$\frac{\partial \ell}{f_0} = \sum_{i=1}^{n} (x_i/\lambda_i - t_i) \times (1 - q_i) \tag{8}$$

$$\frac{\partial \ell}{f_1} = \sum_{i=1}^{n} (x_i/\lambda_i - t_i) \times q_i. \tag{9}$$

We use a quasi-Newton method implemented in the R function `optim` to minimize the negative log-likelihood (7). Note that in the special case in which all the topic proportions are either 0 or 1, the MLEs have a simple closed-form solution:

$$f_0 = \frac{\sum_{i=1}^{n} (1 - q_i) x_i}{\sum_{i=1}^{n} (1 - q_i) t_i} \tag{10}$$

$$f_1 = \frac{\sum_{i=1}^{n} q_i x_i}{\sum_{i=1}^{n} q_i t_i} \tag{11}$$

4.1. EM for Poisson model. We also implement a simple EM algorithm for fitting the Poison model (5). The key is to introduce latent variables $a_i \sim \text{Poisson}(t_i(1 - q_i)f_0)$ and $b_i \sim \text{Poisson}(t_i q_i f_1)$, then work with the expected complete log-likelihood $E[\log p(x, a, b \mid f_0, f_1)]$. The "M-step" updates work out to

$$f_0 = \frac{\sum_{i=1}^{n} \phi_i}{\sum_{i=1}^{n} t_i (1 - q_i)} \tag{12}$$

$$f_1 = \frac{\sum_{i=1}^{n} \gamma_i}{\sum_{i=1}^{n} t_i q_i}, \tag{13}$$

where we have introduced notation for the posterior expectations of the latent variables, $\phi_i \equiv E[a_i]$ and $\gamma_i \equiv E[b_i]$. It can be shown that the posterior distribution of (a_i, b_i) is multinomial with number of trials x_i and event probabilities proportional to $(1 - q_i)f_0$ and $q_i f_1$. So the posterior expectations computed in the "E-step" are

$$\phi_i = x_i(1 - q_i)f_0/\lambda_i \tag{14}$$

$$\gamma_i = x_i q_i f_1/\lambda_i. \tag{15}$$

This completes the description of the EM algorithm for the Poisson model.

4.2. glm identity parameterization. Once we have obtained MLEs of f_0 and f_1, we would also like to characterize uncertainty in these estimates—that is, compute the standard error (s.e.). In particular, we are interested in the s.e. of the log-fold change statistic, β. As an intermediate step, we consider the "glm identity" parameterization $\lambda_i = b_0 + q_i b$, where $b = f_1 - f_0$ and $b_0 = f_0$. This parameterization can be implemented using `glm` in R with `family = poisson(link = "identity")`, and therefore can be used to verify our calculations.

Under the Laplace approximation to the likelihood at the MLE $(\hat{b}_0, \hat{b})$, the covariance matrix is $-H^{-1}$, where H is the 2×2 matrix of second-order partial derivatives,

$$\frac{\partial^2 \ell}{\partial b_0^2} = -\sum_{i=1}^{n} \frac{x_i}{\lambda_i^2}, \qquad \frac{\partial^2 \ell}{\partial b^2} = -\sum_{i=1}^{n} \frac{x_i q_i^2}{\lambda_i^2}, \qquad \frac{\partial^2 \ell}{\partial b_0 \partial b} = -\sum_{i=1}^{n} \frac{x_i q_i}{\lambda_i^2}.$$

This result can be used to obtain the standard errors and z-scores for $\hat{b}_0$ and $\hat{b}$.

4.3. Log-fold change parameterization. Here we derive the s.e. for the MLE of $\beta \equiv \log(f_1/f_0)$. (For convenience, we use the natural logarithm here rather than the base-2 logarithm; to obtain the base-2 log-fold change statistic and its s.e., divide by $\log 2$.) With this new parameterization, the Poisson rates are

$$\lambda_i = f_0 \times \{1 - q_i(1 - e^{\beta})\} \tag{16}$$

The second-order partial derivatives needed to compute the 2×2 Hessian are

$$\frac{\partial \ell}{f_0} = \frac{1}{f_0} \sum_{i=1}^{n} x_i - t_i \lambda_i \tag{17}$$

$$\frac{\partial \ell}{\beta} = f_1 \sum_{i=1}^{n} (x_i/\lambda_i - t_i) \times q_i \tag{18}$$

$$\frac{\partial^2 \ell}{f_0^2} = -\frac{1}{f_0^2} \sum_{i=1}^{n} x_i \tag{19}$$

$$\frac{\partial^2 \ell}{\beta^2} = -f_1 \sum_{i=1}^{n} t_i q_i - x_i f_0 q_i (1 - q_i)/\lambda_i^2 \tag{20}$$

$$\frac{\partial^2 \ell}{\partial f_0 \partial \beta} = -\frac{f_1}{f_0} \sum_{i=1}^{n} t_i q_i. \tag{21}$$

The expression for $\frac{\partial^2 \ell}{\partial f_0 \partial \beta}$ is the more complicated one, but fortuately it simplifies at the MLE, $\beta = \hat{\beta}$ (at the MLE, the gradient of the log-likelihood with respect to β vanishes):

$$\frac{\partial^2 \ell}{\beta^2} = -f_1^2 \sum_{i=1}^{n} x_i (q_i/\lambda_i)^2. \tag{22}$$

Therefore, at the MLE $\beta = \hat{\beta}$, the standard error is

$$\mathrm{se}(\hat{\beta}) = \frac{1}{f_1} \times \sqrt{\frac{\bar{a}}{\bar{a} \times \bar{c} - \bar{b}^2}}, \tag{23}$$

where I've defined

$$\bar{a} = \sum_{i=1}^{n} x_i, \qquad \bar{b} = \sum_{i=1}^{n} t_i q_i, \qquad \bar{c} = \sum_{i=1}^{n} x_i (q_i/\lambda_i)^2.$$

From this, the z-score is recovered as $z = \hat{\beta}/se(\hat{\beta})$.

Appendix A. More on Poisson model. Consider the following process for generating the counts $x_1, \ldots, x_n$:

- for $i = 1, \ldots, n$
  1. $a_i \sim \text{Poisson}(f_1)$ — Sample the within-topic gene count.
  2. $b_i \sim \text{Poisson}(f_0)$ — Sample the outside-topic gene count.
  3. $a'_i \sim \text{Binom}(a_i, q_i)$ — Subsample the within-topic genes.
  4. $b'_i \sim \text{Binom}(b_i, 1 - q_i)$ — Subsample the outside-topic genes.
  5. $x_i \leftarrow a'_i + b'_i$ — Generate the final gene count.

In this generative process, $f_1 = E[a_i]$ represents the within-topic gene rate, and $f_0 = E[b_i]$ is the outside-topic gene rate, and therefore

$$\beta^{\mathsf{abs}} \equiv \log_2 \frac{f_1}{f_0} = \log_2 \frac{E[a_i]}{E[b_i]} \tag{24}$$

is the *log-fold change statistic in (absolute) expression given proportional topic assignments* $q_1, \ldots, q_n$. For intuition, when all the topic proportions q_i are 0 or 1, this statistical process simplify to $x_i \sim \text{Poisson}(f_1)$ if $q_i = 1$, and $x_i \sim \text{Poisson}(f_0)$ if $q_i = 0$, and so (24) would reduce to the ratio of the mean expression levels inside and outside the topic. The Poisson model (5) can be derived from this statistical process, and so estimating f_0, f_1 for the Poisson model (5) provides an estimate of the log-fold change (24).

REFERENCES

[1] J. H. Bullard, E. Purdom, K. D. Hansen, and S. Dudoit, *Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments*, BMC Bioinformatics, 11 (2010), p. 94.

[2] X. Cui and G. A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments*, Genome Biology, 4 (2003).

[3] K. K. Dey, C. J. Hsiao, and M. Stephens, *Visualizing the structure of RNA-seq expression data using grade of membership models*, PLoS Genetics, 13 (2017), p. e1006599.

[4] C. W. Law, Y. Chen, W. Shi, and G. K. Smyth, *voom: precision weights unlock linear model analysis tools for RNA-seq read counts*, Genome Biology, 15 (2014), p. R29.

[5] M. I. Love, W. Huber, and S. Anders, *Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2*, Genome Biology, 15 (2014), p. 550.

[6] J. Quackenbush, *Microarray data normalization and transformation*, Nature Genetics, 32 (2002), pp. 496–501.

[7] M. D. Robinson, D. J. McCarthy, and G. K. Smyth, *edgeR: a Bioconductor package for differential expression analysis of digital gene expression data*, Bioinformatics, 26 (2010), pp. 139–140.

[8] J. H. Stock and M. W. Watson, *Introduction to Econometrics*, Pearson, Boston, MA, 3rd ed., 2015.