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" [4]). 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.

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., [3, 7]). The log-fold change for gene j and condition k is a ratio of two conditional expectations,

$$\mathsf{lfc}(j,k) \equiv \log_2 \frac{E[\, x_j \,|\, \text{condition} = k\,]}{E[\, x_j \,|\, \text{condition} \neq k\,]}, \tag{1}$$

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) [2, 5, 6, 8]. 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.

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.

1.1. The binomial model. We begin with a simple binomial model that predicts expression of a single gene given the proportional assignments to a topic:

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

Here, x_i is the expression level of the target gene in sample i, s_i is the total expression in sample i, and $\mathrm{Binom}(n, \theta)$ 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 $\mathsf{lfc}(j,k) = \log_2 \left\{ \frac{x_{jk}}{x_j - x_{jk}} \times \frac{n - n_k}{n_k} \right\}$.

[2] This is a special case of a well-studied model in econometrics called the *linear probability model* [9]. The linear probability model would be typically written as $\pi_i = \beta_0 + q_i\beta$, where $\beta_0 = p_0$ and $\beta = p_1 - p_0 \in [-1, +1]$. 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, $\log_2(p_1/p_0)$ 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, $w_{it} \mid z_{it} \sim \begin{cases} \text{Binom}(1, p_1) & \text{if } z_{it} = 1 \\ \text{Binom}(1, p_0) & \text{otherwise.} \end{cases}$
  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, $p_1 = p(w_{it} = 1 \mid z_{it} = 1)$ is the conditional probability that the gene is expressed given membership to the topic, and $p_0 = p(w_{it} = 1 \mid z_{it} = 0)$ is the probability that the gene is expressed when not belonging to the topic. Therefore, we have

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

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).

1.2. The Poisson model. A Poisson model similar to the binomial model (2) leads to a method for estimating log-fold change in absolute expression levels. We proceed in a similar way. The Poisson model predicts expression x_i given the topic proportions q_i:

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

in which the Poisson rates are defined as

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

and the unknowns to be estimated are $f_0, f_1 \geq 0$. In the following, we show that $\log_2(f_1/f_0)$ is the log-fold change in absolute expression.

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

$$\log_2 \frac{f_1}{f_0} = \log_2 \frac{E[a_i]}{E[b_i]} \tag{7}$$

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 (7) 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 (7).

In summary, we will use the binomial model (2) to implement the log-fold change analysis for relative expression, and the Poisson model (5) to implement the log-fold change analysis for absolute expression. The next sections derive the mathematical expressions needed to implement these two analyses.

2. Binomial model derivations. *Add derivations here.*

3. Poisson model derivations. *Add derivations here.*

4. The multinomial topic model and Poisson non-negative matrix factorization. Here we briefly describe the multinomial topic model, and its connection to Poisson non-negative matrix factorization (Poisson NMF).

We begin with the "bag of words" description, which was used to describe the LDA model [1]. In this view, each document (or gene expression profile) i is represented as a vector of terms/genes, $w_i = (w_{i1}, \ldots, w_{is_i})$, where s_i is the size of document i. (The order of the words or genes appearing in this vector doesn't matter, hence the "bag of words.") Each $w_{it} \in \{1, \ldots, m\}$ is term/gene j with probability $p(w_{it} = j \,|\, z_{it} = k) = f_{jk}$, in which we have introduced z_{it}, a variable indicating which topic $k \in \{1, \ldots, K\}$ the word/gene is drawn from. The topic indicator variables for document i are in turn generated according to $p(z_{it} = k) = l_{ik}$.

This process also defines a *multinomial* model for an $n \times m$ matrix of counts x_{ij}:

$$x_{i1}, \ldots, x_{im} \sim \text{Multinom}(x_{i1}, \ldots, x_{im}; s_i, \pi_i), \tag{8}$$

where $x_{ij} = \sum_{t=1}^{s_i} \delta_j(w_{it})$ is the number of times term/gene j appears in document/cell i, and the probabilities π_{ij} are weighted sums of the "factors" f_{jk},

$$\pi_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk}. \tag{9}$$

The log-likelihood for the multinomial topic model, ignoring terms that do not depend on the model parameters, has a simple expression:

$$\log p(x) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \log(\textstyle\sum_{k=1}^{K} l_{ik} f_{jk}). \tag{10}$$

As we have shown elsewhere, the multinomial topic model is closely related to a Poisson non-negative matrix factorization of the count data,

$$x_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{11}$$

where $\lambda_{ij} = \sum_{k=1}^{K} \hat{l}_{ik} \hat{f}_{jk}$. Given a Poisson NMF fit, an equivalent multinomial topic model can be easily recovered, as we have shown elsewhere.

5. Gene expression differences in topics. Returning to the question of assessing differential gene expression, there are two new twists when done in the context of topic modeling:

1. The cluster (topic) assignments are probabilistic.
2. The cluster assignments are made at the level of genes, not cells.

I propose a log-fold change statistic to address these two points. It compares the probability of gene j occurring ($w = j$) given topic k ($z = k$) versus the probability given assignment a topic other than k ($z \neq k$):

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) \equiv \log_2 \frac{p(w = j \,|\, z = k)}{p(w = j \,|\, z \neq k)}. \tag{12}$$

For a given gene j and topic k, $\mathsf{lfc}(j,k)$ can be calculated as

$$\begin{aligned}
\mathsf{lfc}^{\mathsf{topics}}(j,k) &= \log_2 \left\{ \frac{p(w = j, z = k)}{p(w = j, z \neq k)} \times \frac{p(z \neq k)}{p(z = k)} \right\} \\
&= \log_2 \left\{ \frac{\sum_{i=1}^n \sum_{t=1}^{s_i} \delta_j(w_{it})\, \phi_{ijkt}}{\sum_{i=1}^n \sum_{t=1}^{s_i} \delta_j(w_{it})(1 - \phi_{ijkt})} \right. \\
&\qquad \left. \times \frac{\sum_{i=1}^n \sum_{j'=1}^m \sum_{t=1}^{s_i} \delta_{j'}(w_{it})(1 - \phi_{ij'kt})}{\sum_{i=1}^n \sum_{j'=1}^m \sum_{t=1}^{s_i} \delta_{j'}(w_{it})\, \phi_{ij'kt}} \right\},
\end{aligned} \tag{13}$$

where ϕ_{ijkt} denotes the posterior probability of $z_{it} = k$ given $w_{it} = j$,

$$\begin{aligned}
\phi_{ijkt} &\equiv p(z_{it} = k \,|\, w_{it} = j) \\
&= \frac{p(w_{it} = j \,|\, z_{it} = k)\, p(z_{it} = k)}{\sum_{k'=1}^K p(w_{it} = j \,|\, z_{it} = k')\, p(z_{it} = k')} \\
&= \frac{l_{ik} f_{jk}}{\sum_{k'=1}^K l_{ik'} f_{jk'}}.
\end{aligned} \tag{14}$$

Since the topic assignments z_{it} do not depend on t—that is, we can drop the "t" subscript from the ϕ_{ijkt}'s—the expression for the *lfc* simplifies:

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^n x_{ij}\, \phi_{ijk}}{\sum_{i=1}^n x_{ij}(1 - \phi_{ijk})} \times \frac{\sum_{i=1}^n \sum_{j'=1}^m x_{ij'}(1 - \phi_{ij'k})}{\sum_{i=1}^n \sum_{j'=1}^m x_{ij'} \phi_{ij'k}} \right\}. \tag{15}$$

At the maximum-likelihood solution (MLE) of the l_{ik}'s and f_{jk}'s, the *lfc* statistic simplifies further:

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^n x_{ij}\, \phi_{ijk}}{\sum_{i=1}^n x_{ij}(1 - \phi_{ijk})} \times \frac{\sum_{i=1}^n s_i(1 - l_{ik})}{\sum_{i=1}^n s_i l_{ik}} \right\}. \tag{16}$$

This is because, at the MLE, the loadings l_{ik}, $k = 1, \ldots, K$, for a given document/cell i should be equal to the average of the weighted counts $\frac{1}{s_i} \sum_{j=1}^m x_{ij} \phi_{ijk}$.

Finally, it is convenient that the *lfc* (13, 16) will be the same if we replace the multinomial topic model parameters l_{ik} and f_{jk} with the corresponding parameters of the Poisson NMF, $\hat{l}_{ik}$ and $\hat{f}_{jk}$ (proof not given). From the derivation of the EM algorithm for Poisson NMF, this identity holds at the MLE:

$$\hat{f}_{jk} = \frac{\sum_{i=1}^n \phi_{ijk}}{\sum_{i=1}^n \hat{l}_{ik}}.$$

Plugging this relationship into (16), we obtain the following simple expression for the log-fold change:

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\hat{f}_{jk} \sum_{i=1}^n \hat{l}_{ik}}{\sum_{k' \neq k} \hat{f}_{jk'} \sum_{i=1}^n \hat{l}_{ik'}} \times \frac{\sum_{i=1}^n s_i(1 - \hat{l}_{ik})}{\sum_{i=1}^n s_i \hat{l}_{ik}} \right\}. \tag{17}$$

What is nice about this about this expression is that it can be computed without seeing the data. It is also plain to see from this expression that to arrive at a log-fold change, one must weight the factors f_{jk} by the sample-wide topic probabilities $\sum_i l_{ik}$ across This same expression also works with the for the parameters of multinomial topic model l_{ik}, f_{jk}, again, so long as they are MLEs (proof not shown).

REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan, *Latent Dirichlet allocation*, Journal of Machine Learning Research, 3 (2003), pp. 993–1022.

[2] 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.

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

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

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