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]. To motivate the development of the methods, 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, 4]). 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 (e.g., UMI count) of gene j.[1]

The statistic (1) is a measure of *absolute* change in expression level between two conditions, but sometimes it is preferrable to measure *relative* change, say, by normalizing each sample by the total expression. As we will see, a topic modeling perspective provides a natural way to analyze both absolute or relative changes in gene expression.

1.1. The binomial model.

1.2. The Poisson model.

2. 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}:

$$x_{i1}, \ldots, x_{im} \sim \text{Multinom}(x_{i1}, \ldots, x_{im}; s_i, \pi_i), \tag{2}$$

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},

$$\pi_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk}. \tag{3}$$

*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 cells in condition k, x_j as the total gene expression for gene j amon all cells, n_k as the number of cells in condition k expression profiles, and n as the total number of cells, the log-fold change can be computed as $\mathsf{lfc}(j,k) = \log_2 \left\{ \frac{x_{jk}}{x_j - x_{jk}} \times \frac{n - n_k}{n_k} \right\}$.

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

$$\log p(x) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \log(\textstyle\sum_{k=1}^{K} l_{ik} f_{jk}). \tag{4}$$

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

$$x_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{5}$$

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.

3. 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 ($z \neq k$):

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) \equiv \log_2 \frac{p(w = j \,|\, z = k)}{p(w = j \,|\, z \neq k)}. \tag{6}$$

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

$$\begin{aligned} \mathsf{lfc}^{\mathsf{topics}}(j,k) &= \log_2 \left\{ \frac{p(w = j, z = k)}{p(w = j, z \neq k)} \times \frac{p(z \neq k)}{p(z = k)} \right\} \\ &= \log_2 \left\{ \frac{\sum_{i=1}^{n} \sum_{t=1}^{s_i} \delta_j(w_{it})\, \phi_{ijkt}}{\sum_{i=1}^{n} \sum_{t=1}^{s_i} \delta_j(w_{it})(1 - \phi_{ijkt})} \right. \\ &\qquad \left. \times \frac{\sum_{i=1}^{n} \sum_{j'=1}^{m} \sum_{t=1}^{s_i} \delta_{j'}(w_{it})(1 - \phi_{ij'kt})}{\sum_{i=1}^{n} \sum_{j'=1}^{m} \sum_{t=1}^{s_i} \delta_{j'}(w_{it})\, \phi_{ij'kt}} \right\}, \end{aligned} \tag{7}$$

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

$$\begin{aligned} \phi_{ijkt} &\equiv p(z_{it} = k \,|\, w_{it} = j) \\ &= \frac{p(w_{it} = j \,|\, z_{it} = k)\, p(z_{it} = k)}{\sum_{k'=1}^{K} p(w_{it} = j \,|\, z_{it} = k')\, p(z_{it} = k')} \\ &= \frac{l_{ik} f_{jk}}{\sum_{k'=1}^{K} l_{ik'} f_{jk'}}. \end{aligned} \tag{8}$$

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:

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^{n} x_{ij}\, \phi_{ijk}}{\sum_{i=1}^{n} x_{ij}(1 - \phi_{ijk})} \times \frac{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij'}(1 - \phi_{ij'k})}{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij'} \phi_{ij'k}} \right\}. \tag{9}$$

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

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^n x_{ij}\,\phi_{ijk}}{\sum_{i=1}^n x_{ij}(1-\phi_{ijk})} \times \frac{\sum_{i=1}^n s_i(1-l_{ik})}{\sum_{i=1}^n s_i l_{ik}} \right\}. \tag{10}$$

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* (7, 10) 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 (10), we obtain the following simple expression for the log-fold change:

$$\mathsf{lfc}^{\mathsf{topics}}(j,k) = \log_2 \left\{ \frac{\hat{f}_{jk}\sum_{i=1}^n \hat{l}_{ik}}{\sum_{k'\neq k}\hat{f}_{jk'}\sum_{i=1}^n \hat{l}_{ik'}} \times \frac{\sum_{i=1}^n s_i(1-\hat{l}_{ik})}{\sum_{i=1}^n s_i \hat{l}_{ik}} \right\}. \tag{11}$$

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