THE "TERM-COUNT LOG-RATIO" STATISTIC FOR TOPIC MODELING ANALYSIS OF DIFFERENTIAL GENE EXPRESSION

PETER CARBONETTO*

1. Differential gene expression. The "log-fold change" statistic is commonly used in microarray and RNA sequencing experiments to quantify expression changes between two conditions (e.g., [2, 3]). To motivate the ideas below, I write the log-fold change for gene j and condition k as a ratio of two conditional expectations,

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

where x_j is the measured expression level (e.g., UMI count) of gene j. In experiments where the conditions are inferred—for example, by running a machine learning algorithm to cluster the expression profiles—this quantity could represent the difference in gene expression between cells inside and outside a cluster.

Supposing n_k out of a total of n gene expression profiles (cells) are from condition k, then $\mathsf{lfc}(j,k)$ can be computed as

$$\mathsf{lfc}(j,k) = \log_2 \left\{ \frac{n_{jk}}{n_j - n_{jk}} \times \frac{n - n_k}{n_k} \right\}, \tag{2}$$

where n_j is the total expression of gene j among all expression profiles, and n_{jk} is the total expression of j among all cells in condition (or cluster) k.

The aim in the next sections is to define a analogue to the log-fold change statistic for topic modeling.

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).

The topic model describes a process for generating an $n \times m$ matrix of counts, X. We begin with the "bag of words" description, which is what is used to describe LDA [1]. In this view, each row i is a document (or gene expression profile), and let m_i be the size of this document; that is, $m_i = \sum_{j=1}^{m} x_{ij}$. The vector w_i is a vector of terms (or genes) of length m_i (the order of the words or genes appearing in this vector doesn't matter, hence the "bag of words"). For each $t = 1, \ldots, m_i$, the word/gene w_{it} is equal to j with probability $p(w_{it} \,|\, z_{it} = k) = f_{jk}$, where z_{it} is a variable indicating which topic, $k \in \{1, \ldots, K\}$ the word/gene belongs to. The topic indicator variable is in turn generated according to $p(z_{it} = k) = l_{ik}$, where $l_{i1}, \ldots, l_{iK}$ is a document-specific probability table.

This process defines a *multinomial* model for the counts $x_{i1}, \ldots, x_{im}$ in document/sample i, hence the "multinomial topic model":

$$x_{i1}, \ldots, x_{im} \sim \text{Multinom}(x_{i1}, \ldots, x_{im}; m_i, \pi_i), \tag{3}$$

where π_i is a vector of probabilities π_{ij} given by a weighted sum of the word/gene probabilities, or "factors", f_{jk},

$$\pi_{ij} = \sum_{k=1}^{K} l_{ik} f_{jk}. \tag{4}$$

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

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

$$\log p(x) = \sum_{i=1}^{n} \sum_{j=1}^{m} x_{ij} \log(\textstyle\sum_{k=1}^{K} l_{ik} f_{jk}) \tag{5}$$

As we have investigated in other documents, the multinomial topic model is closely related to a Poisson non-negative matrix factorization of the count data,

$$p() \tag{6}$$

3. The "term-count log-ratio" (*tclr*). Returning to the question of assessing differential gene expression, there are two twists relative to the standard analysis: one, the group (topic) assignments are probabilistic; two, the group assignments are made at the level of genes, not cells. With these two points in mind, I propose the "term-count log-ratio", the (logarithm of the) expected expression level of gene j conditioned on assignment to topic k over the expected expression level of gene j conditioned on not being assigned to topic k:

$$\mathsf{tclr}(j,k) = \log_2 \frac{E[\, x_j \,|\, z_j = k \,]}{E[\, x_j \,|\, z_j \neq k \,]}. \tag{7}$$

For a given gene j and topic k, $\mathsf{tclr}(j,k)$ is calculated as

$$\begin{aligned}
\mathsf{tclr}(j,k) &= \log_2 \left\{ \frac{E[\, x_j, z_j = k \,]}{E[\, x_j, z_j \neq k \,]} \times \frac{p(z_j \neq k)}{p(z_j = k)} \right\} \\
&= \log_2 \left\{ \frac{\sum_{i=1}^{n} E[\, x_{ij}, z_{ij} = k \,]}{\sum_{i=1}^{n} E[\, x_{ij}, z_{ij} \neq k \,]} \times \frac{\sum_{i=1}^{n} p(z_{ij} \neq k)}{\sum_{i=1}^{n} p(z_{ij} = k)} \right\} \\
&= \log_2 \left\{ \frac{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} = k)}{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} \neq k)} \times \frac{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij}\, p(z_{ij'} \neq k)}{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij}\, p(z_{ij'} = k)} \right\},
\end{aligned} \tag{8}$$

The probabilities $p(z_{ij} = k)$ in the above expressions are *posterior probabilities*, which, in the multinomial topic model, work out to simply

$$p(z_{ij} = k) = \frac{l_{ik} f_{jk}}{\sum_{k'=1}^{K} l_{ik'} f_{jk'}}. \tag{9}$$

Here I've made use of the property that the topic assignments z_{it} are the same for all $w_{it} = j$, so in a small abuse of notation I've written the topic assignments as z_{ij}.

At the maximum-likelihood solution of the l_{ik}'s and f_{kl}'s, the *tclr* statistic simplifies to

$$\mathsf{tclr}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} = k)}{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} \neq k)} \times \frac{\sum_{i=1}^{n} m_i l_{ik}}{\sum_{i=1}^{n} m_i (1 - l_{ik})} \right\}. \tag{10}$$

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