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 differences 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,

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

where x_j is the measured expression level (e.g., UMI count) of gene j. 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 of 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).

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}:

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

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},

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

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

$$x_{ij} \sim \text{Poisson}(\lambda_{ij}), \tag{6}$$

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; by "equivalent", we mean the likelihoods of the two models are the same.

3. The "term-count log-ratio" (*tclr*). 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 statistic, the "term-count log-ratio," to address these two points. It is 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:

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

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

$$\begin{aligned} \mathsf{tclr}(j,k) &= \log_2 \left\{ \frac{E[\, x_j, \mathsf{topic}(j) = k \,]}{E[\, x_j, \mathsf{topic}(j) \neq k \,]} \times \frac{p(\mathsf{topic}(j) \neq k)}{p(\mathsf{topic}(j) = k)} \right\} \\ &= \log_2 \left\{ \frac{\sum_{i=1}^{n} \sum_{t=1}^{s_i} \delta_j(w_{it})\, \phi_{ijkt}}{\sum_{i=1}^{n} \sum_{t=1}^{s_i} \delta_j(w_{it})\, (1 - \phi_{ijkt})} \times \frac{\sum_{i=1}^{n} \sum_{t=1}^{s_i} \phi_{ijkt}}{\sum_{i=1}^{n} \sum_{t=1}^{s_i} 1 - \phi_{ijkt}} \right\}, \end{aligned} \tag{8}$$

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

$$\begin{aligned} \phi_{ijkt} &\equiv p(z_{it} = k \,|\, w_{it} = j) \\ &= \frac{p(w_{it} = j \,|\, z_{it} = k)\, p(z_{it} = k)}{\sum_{k'=1}^{K} p(w_{it} = j \,|\, z_{it} = k')\, p(z_{it} = k')} \\ &= \frac{l_{ik} f_{jk}}{\sum_{k'=1}^{K} l_{ik'} f_{jk'}}. \end{aligned} \tag{9}$$

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

$$\mathsf{tclr}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^{n} x_{ij}\, \phi_{ijk}}{\sum_{i=1}^{n} x_{ij}\, (1 - \phi_{ijk})} \times \frac{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij} \phi_{ij'k}}{\sum_{i=1}^{n} \sum_{j'=1}^{m} x_{ij} (1 - \phi_{ij'k})} \right\} \tag{10}$$

At the maximum-likelihood solution (MLE) of the l_{ik}'s and f_{kl}'s, the *tclr* statistic simplifies slightly:

$$\mathsf{tclr}(j,k) = \log_2 \left\{ \frac{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} = k)}{\sum_{i=1}^{n} x_{ij}\, p(z_{ij} \neq k)} \times \frac{\sum_{i=1}^{n} m_i l_{ik}}{\sum_{i=1}^{n} m_i (1 - l_{ik})} \right\}. \tag{11}$$

This is because, at the MLE, the loadings l_{ik}, $k = 1, \ldots, K$, for a given document/cell i should be proportional to the sums $\sum_{j=1}^{m} x_{ij} p(z_{ij} = k)$.

Finally, it is convenient that the *tclr* (8) 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).

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.