

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,

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

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 $\text{lfc}(j, k)$ can be computed as

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

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 an 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, $w_i = (w_{i1}, \dots, w_{is_i})$, 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 $w_{it} \in \{1, \dots, m\}$ is term/gene j with probability $p(w_{it} = j \mid z_{it} = k) = f_{jk}$, in which we have introduced z_{it} , a variable indicating which topic $k \in \{1, \dots, K\}$ 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} :

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

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} ,

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

*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:

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

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

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

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 :

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

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

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

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

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

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:

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

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

$$(11) \quad \text{tclr}(j, k) = \log_2 \left\{ \frac{\sum_{i=1}^n x_{ij} \phi_{ijk}}{\sum_{i=1}^n x_{ij} (1 - \phi_{ijk})} \times \frac{\sum_{i=1}^n s_i l_{ik}}{\sum_{i=1}^n s_i (1 - l_{ik})} \right\}.$$

This is because, at the MLE, the loadings s_{ik} , $k = 1, \dots, K$, for a given document/cell i should be proportional to the sums $\sum_{j=1}^m x_{ij} \phi_{ijk}$.

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.