

# 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.<sup>1</sup>

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  $x_{ij}$ :

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

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  $\pi_{ij}$  are weighted sums of the “factors”  $f_{jk}$ ,

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

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

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

---

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

<sup>1</sup>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  $\text{lfc}(j, k) = \log_2 \{n_{jk} / (n_j - n_{jk}) \times (n - n_k) / n_k\}$ , 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)  $k$ .

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

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

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.

**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  $k$  ( $z \neq k$ ):

$$(6) \quad \text{lfc}^{\text{topics}}(j, k) \equiv \log_2 \frac{p(w = j \mid z = k)}{p(w = j \mid z \neq k)}.$$

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

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

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

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

Since the topic assignments  $z_{it}$  do not depend on  $t$ —that is, we can drop the “ $t$ ” subscript from the  $\phi_{ijkt}$ ’s—the expression for the  $\text{lfc}$  simplifies:

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

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

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

This is because, at the MLE, the loadings  $l_{ik}$ ,  $k = 1, \dots, K$ , 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, we have that

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

Plugging this relationship into (10), we obtain the following simple expression for the log-fold change:

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

What is nice about this about this expression is that it can be computed without seeing the data. This same expression also works with the for the parameters of multinomial topic model  $l_{ik}, f_{jk}$  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] J. QUACKENBUSH, *Microarray data normalization and transformation*, Nature Genetics, 32 (2002), pp. 496–501.