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., [1, 2]). To motivate the ideas below, I write the log-fold change for gene j and condition k as a ratio of two conditional expectations,

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

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

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

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.

2. Poisson non-negative matrix factorization and the multinomial topic model. *Add text here.*

REFERENCES

[1] X. Cui and G. A. Churchill, *Statistical tests for differential expression in cDNA microarray experiments*, Genome Biology, 4 (2003).

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

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