Appendix

Variance of the k-prop test statistics

Denote the number of cells as N, the number of genes as M, and let the observed gene expression matrix (UMI counts) be denoted as X. Each element indexed by X_{ij} represents the UMI count of the i-th gene in the j-th cell. We define k-proportion as the percentage of cells with gene expression less than a defined integer value k, $\hat{g}_i^{(k)} = P(X_{ij} \leq k)$. Our goal is to test whether the underlying true k-proportion is significantly higher than the expectation under a negative binomial model, i.e., $\Delta_i = \hat{g}_i^{(k)} - F_{NB}(k; \bar{X}_i, \Phi) = 0$, where F_{NB} is the cumulative distribution function under a negative binomial distribution, $\bar{X}_i$ is the mean and Φ is the dispersion level. We will use the delta method to derive the equation for variance of this test statistics.

We can write the test statistics with Tyler expansion approximation as follows:

$$
\begin{aligned}
\Delta_i &= \hat{g}_i^{(k)} - F_{NB}(k; \bar{X}_i, \Phi) \\
&= \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\big\{X_{ij} \leq k\big\} - F_{NB}(k; \bar{X}_i, \Phi) \\
&= \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\big\{X_{ij} \leq k\big\} - F_{NB}(k; \lambda_i, \Phi) + F_{NB}(k; \lambda_i, \Phi) - F_{NB}(k; \bar{X}_i, \Phi) \\
&\approx \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}\big\{X_{ij} \leq k\big\} - F_{NB}(k; \lambda_i, \Phi) + \frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} (\lambda_i - \bar{X}_i) \\
&= \frac{1}{N} \sum_{j=1}^{N} \Big(\mathbb{1}\big\{X_{ij} \leq k\big\} - \frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} X_{ij}\Big) - F_{NB}(k; \lambda_i, \Phi) + \frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} \lambda_i
\end{aligned}
$$

Only the two terms within the sum is changing with data and reflecting variability. We continue by calculating the variance of these two terms for each data point as follows.

$$
\begin{aligned}
&\mathbf{Var}\Big(\mathbb{1}\big\{X_{ij} \leq k\big\} - \frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} X_{ij}\Big) \\
&= \mathbf{Var}\Big(\mathbb{1}\big\{X_{ij} \leq k\big\}\Big) + \mathbf{Var}\Big(\frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} X_{ij}\Big) - 2\frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} \mathbf{Cov}\Big(\mathbb{1}\big\{X_{ij} \leq k\big\}, X_{ij}\Big) \\
&= g_i^{(k)}(1 - g_i^{(k)}) + \Big(\frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i}\Big)^2 \mathbf{Var}(X_{ij}) - 2\frac{\partial F_{NB}(k; \lambda_i, \Phi)}{\partial \lambda_i} \mathbf{Cov}\Big(\mathbb{1}\big\{X_{ij} \leq k\big\}, X_{ij}\Big)
\end{aligned}
\tag{1}
$$

The above equations need the formula of partial derivatives of $F_{NB}(k; \lambda_i, \Phi)$. Let $r = \Phi^{-1}$, we can write the PDF and CDF of negative binomial as follows:

$$
\begin{aligned}
F_{NB}(k;\lambda_i,r) &= \sum_k f_{NB}(k;\lambda_i,r) \\
&= \sum_k \binom{k+r-1}{k} (\frac{\lambda_i}{\lambda_i+r})^k (\frac{r}{\lambda_i+r})^r \\
&= \sum_k \binom{k+r-1}{k} r^r(\lambda_i+r)^{-k-r}\lambda_i^k \\
\frac{\partial f_{NB}(k;\lambda_i,\Phi)}{\partial \lambda_i} &= \binom{k+r-1}{k} r^{r+1}(\lambda_i+r)^{-k-r-1}\lambda_i^{k-1}(k+\lambda_i) \\
&= \frac{(k-\lambda_i)r}{(\lambda_i+r)\lambda_i} f_{NB}(k;\lambda_i,r) \\
\frac{\partial F_{NB}(k;\lambda_i,\Phi)}{\partial \lambda_i} &= \sum_k \Big(\frac{(k-\lambda_i)r}{(\lambda_i+r)\lambda_i} f_{NB}(k;\lambda_i,r) \Big)
\end{aligned}
$$

In addition, we know

$$
\begin{aligned}
\mathbf{Var}(X_{ij}) &= \lambda_i + \lambda_i^2/r \\
\mathbf{Cov}\Big(\mathbb{1}\big\{X_{ij} \le k\big\}, X_{ij}\Big) &= \mathbb{E}\big(\mathbb{1}\big\{X_{ij} \le k\big\} X_{ij}\big) - \mathbb{P}\big(X_{ij} \le k\big)\mathbb{E}X \\
&= \sum_k k f_{NB}(k;\lambda_i,r) - g_i^{(k)}\lambda_i
\end{aligned}
$$

This completes the calculation of variance using formula (1).