Zero-inflated model for single-cell RNA-seq data

Davide Risso

2015-10-20

Model Specification

The model is as follows. We have single-cell RNA-seq measurements for $j = 1, ..., J$ genes and $i = 1, ..., n$ cells. We denote the **true unknown expression** of gene j in cell i as μ_{ij}. Y_{ij} denotes the **observed** number of reads for gene j in cell i; Z_{ij} is a binary **unobserved** random variable that models the excess of zeros: $Pr(Z_{ij} = 1) = \pi_{ij}$ is the probability that gene j is **not measured** in cell i, i.e., that it is not present in the cDNA library and not sequenced. $Y_{ij}|Z_{ij} = 0 \sim NB(\mu_{ij}, \phi_j)$. (Note that the choice of Z=1 means not expressed is for consistency with standard zero-inflation parametrizations, e.g., the one in the pscl R package).

We assume that the probability of measuring a gene depends on some gene-level technical covariates, by considering the following regression across genes (i.e., one per cell):

$$\pi_{ij} = \text{logit}^{-1}(\alpha_i W_j)$$

In practice, we consider the "technical covariates" to be the population-level average expression, the length and the GC-content, i.e.,

$$\pi_{ij} = \text{logit}^{-1}(\alpha_{0i} + \alpha_{1i}\tilde{\mu}_j + \alpha_{2i} l_j + \alpha_{3i} g_j),$$

where l_j is the length of the gene j, g_j its GC-content and $\tilde{\mu}_j$ is the population-level average expression of gene j. Note that unlike l_j and g_j, $\tilde{\mu}_j$ is not known a priori: we can estimate it from bulk samples (if available) or we can substitute $\tilde{\mu}_j$ with μ_{ij}.

Similarly, we can consider a across-cell regression to relate the true expression μ_{ij} to biological and technical covariates of interest, i.e.,

$$\log \mu_{ij} = \beta_j X_i,$$

where the regressors X_i can be biological, e.g., sub-population labels, and/or technical, e.g., quality features. Note that in a typical application, (part of) X can be unkown (clustering).

As a first simple model we consider a constant $log\mu_{ij} = \beta_{0j}$, but it should be straightforward to add **known** covariates. The addition of a clustering step to estimate the labels could be doable by, e.g., iterating between a clustering step and a zero-inflation estimation step.

Note that

$$Pr(Y_{ij} = y_{ij}|X, W) = \begin{cases} (1 - \pi_{ij})\, p_{NB}(y_{ij}; \mu_{ij}, \phi_j) & y_{ij} > 0 \\ \pi_{ij} + (1 - \pi_{ij})\, p_{NB}(0; \mu_{ij}, \phi_j) & y_{ij} = 0. \end{cases}$$

The quantity of interest seems to be the probability of being a dropout given that we do not observe any expression, i.e.,

$$Pr(Z_{ij} = 1|Y_{ij} = 0) = \frac{\pi_{ij}}{\pi_{ij} + (1 - \pi_{ij})(1 + \phi_j \mu_{ij})^{-1/\phi_j}}.$$

These probabilities can be used as weights in a weighted PCA, or can be used to impute the drop-outs, e.g., by

$$y^*_{ij} = y_{ij}\, Pr(Z_{ij} = 0|Y_{ij} = y_{ij}) + \mu_{ij}\, Pr(Z_{ij} = 1|Y_{ij} = y_{ij}).$$

Hence,

$$y_{ij}^{*} = \begin{cases} y_{ij} & y_{ij} > 0, \\ \mu_{ij}\, Pr(Z_{ij} = 1 | Y_{ij} = 0) & y_{ij} = 0. \end{cases}$$

Parameter Estimation

To estimate the parameters in a reasonable time, we need to numerically optimize the likelihood. We can in principle use a classical implementation of a regular zero-inflated model (e.g., the implementation in the pscl package), but this requires a $(n\,J) \times p$ design matrix, where $p = n\,(k_Z + 1) + J\,(k_Y + 1) + J$, k_Z is the number of covariates of the logistic regression, k_Y is the number of covariates in the log-linear regression and the last J parameters are the gene-specific dispersions. Note that we may consider a simpler model where there is only one global dispersion parameter, reducing the number of parameters by $J - 1$.

This design matrix will be very sparse, hence a more efficient way to maximize the likelihood is to compute the partial likelihood per each cell (or gene) and then optimize the sum of the partial likelihoods.

We can write the log-likelihood in the following way.

$$l(\alpha, \beta, \theta) = \sum_{i,j:y_{ij}>0} l_1(\alpha_i, \beta_j, \theta_j) + \sum_{i,j:y_{ij}=0} l_0(\alpha_i, \beta_i, \theta_j),$$

where

$$l_1(\alpha, \beta, \theta) = \log(1-\pi) + \log\Gamma(y+\theta) - \log\Gamma(\theta) + \theta\,[\log\theta - \log(\mu+\theta)] + y\,[\log\mu - \log(\mu+\theta)], \quad (1)$$

$$l_0(\alpha, \beta, \theta) = \log[\pi + (1-\pi)\,p_{Y|Z}(0)]; \quad (2)$$

where

$$\mu = e^{X\beta}, \quad (3)$$

$$\pi = \frac{1}{1+e^{-W\alpha}}, \quad (4)$$

$$\theta = 1/\phi, \quad (5)$$

$$p_{Y|Z}(0) = Pr_{Y|Z=0}(y=0;\mu,\theta) = \left(\frac{\theta}{\mu+\theta}\right)^{\theta}. \quad (6)$$

Using the chain rule, we can derive the following gradient functions.

$$\frac{\partial l_1(\alpha,\beta,\theta)}{\partial\alpha} = -\frac{1}{1-\pi}\,\frac{W\,e^{-W\alpha}}{(1+e^{-W\alpha})^2}, \quad (7)$$

$$\frac{\partial l_0(\alpha,\beta,\theta)}{\partial\alpha} = \frac{1-p_{Y|Z}(0)}{p_Y(0)}\,\frac{W\,e^{-W\alpha}}{(1+e^{-W\alpha})^2}, \quad (8)$$

$$\frac{\partial l_1(\alpha,\beta,\theta)}{\partial\beta} = \left[y - \left(\frac{y+\theta}{\mu+\theta}\right)\mu\right] X, \quad (9)$$

$$\frac{\partial l_0(\alpha,\beta,\theta)}{\partial\beta} = -\frac{1-\pi}{p_Y(0)}\,p_{Y|Z}(0)\,\frac{\theta}{\mu+\theta}\,\mu\,X, \quad (10)$$

$$\frac{\partial l_1(\alpha,\beta,\theta)}{\partial\theta} = \Psi(y+\theta) - \Psi(\theta) + \log\theta - \log(\mu+\theta) - \frac{y+\theta}{\mu+\theta} + 1, \quad (11)$$

$$\frac{\partial l_0(\alpha,\beta,\theta)}{\partial\theta} = \frac{1-\pi}{p_Y(0)}\,p_{Y|Z}(0)\,\left[\log\left(\frac{\theta}{\mu+\theta}\right) + \frac{\mu}{\mu+\theta}\right], \quad (12)$$

where $\Psi(x)$ denotes the digamma function and

$$p_Y(0) = Pr_Y(y = 0; \pi, \mu, \theta) = \pi + (1 - \pi) \left(\frac{\theta}{\mu + \theta} \right)^{\theta}.$$