UROP Notes

Michael Diao

1 RCTD

Our ultimate goal is to determine the fractional contributions of each cell type to a particular sample. We do this by maximum likelihood, using the following hierarchical model:

$$\begin{aligned} \mathsf{Y}_{i,j} \mid \lambda_{i,j} &\sim \text{Poisson}(N_i \lambda_{i,j}) \\ \log \lambda_{i,j} &= \log(\vec{\beta}_i \cdot \vec{\mu}_j) + \alpha_i + \gamma_j + \varepsilon_{i,j}, \end{aligned}$$

where

- $\mathsf{Y}_{i,j}$ is the random variable corresponding to the observed expression of gene j at pixel i,
- N_i is the number of transcripts for pixel i,
- $\vec{\beta}_i$ is the K-dimensional row vector of contributions from each cell type (where K is the number of cell types in question) at pixel i,
- $\vec{\mu}_j$ is the K-dimensional column vector of mean expressions of gene j for each cell type,
- α_i is a fixed pixel-specific effect.
- γ_j and $\varepsilon_{i,j}$ are random effects that introduce noise. γ_j in particular is intended to account for platform effects that may over- or underrepresent certain genes. We let these be normally distributed with mean 0 and variance σ_γ, σ_ε respectively.

Therefore, determining the fractional contributions of each cell type reduces to finding the maximum likelihood parameter $\vec{\beta}_i$ for each i.

> **Question 1.1.** Now we have a ton of parameters, potentially thousands. β alone introduces $K \times J$ of them. How do we do any useful estimation here?

We proceed in the following steps:

1. **Supervised estimation of cell type profiles**.

   Using a reference dataset, we estimate the parameters μ_j of expression levels for gene j, giving $\hat{\vec{\mu}}_j$ which will be used in the next steps.

   We can do this by obtaining a (e.g. scRNA-seq) reference annotated with cell types, after which $\vec{\mu}_j$ can be estimated as the empirical average normalized expression of gene j within each cell type.

2. **Gene filtering**.

   Using the estimated expression profiles $\hat{\vec{\mu}}_j$, we filter out genes that are not highly variable across cell types.

   We can do this by taking the expression profiles $\hat{\vec{\mu}}_j$ and selecting genes with a minimum average expression and sufficiently high variance.

   does this discard genes that are, say, only expressed in one cell type? (average might be low, but gene could be good marker)

3. **Platform Effect Normalization**.

   With an estimate for $\vec{\mu}_j$, it turns out we now have a way to estimate the platform effects γ_j as well. The idea is that we can consider the average observed expression across pixels

$$\mathsf{M}_j = \frac{1}{I}\sum_{i=1}^{I} \mathsf{Y}_{i,j},$$

   whence

$$\begin{aligned}
\log \mathbb{E}_{\mathsf{M}_j|\vec{\lambda}_j}\left[M_j \,\middle|\, \lambda_{1,j}, \ldots, \lambda_{I,j}\right] &= \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i \lambda_{i,j}\right) \\
&= \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i \exp\left(\log(\vec{\beta}_i \cdot \vec{\mu}_j) + \alpha_i + \gamma_j + \varepsilon_{i,j}\right)\right) \\
&= \gamma_j + \log\left(\frac{1}{I}\sum_{i=1}^{I} N_i (\vec{\beta}_i \cdot \vec{\mu}_j) \exp\left(\alpha_i + \varepsilon_{i,j}\right)\right) \\
&= \gamma_j + \log\left(\frac{1}{I}\sum_{i=1}^{I} \left(\sum_{k=1}^{K} \beta_{i,k} \cdot \mu_{k,j}\right) N_i \exp\left(\alpha_i + \varepsilon_{i,j}\right)\right)
\end{aligned}$$

   ¡++¿

¡++¿