CytoOne

Yuqiu (Ian) Yang

Introduction

CyTOF (cytometry by time-of-flight) is a single-cell proteomic technique that uses heavy metal ions to detect the number of various proteins present on the surfaces of individual cells.

Goal

- Dimension reduction
- Cell clustering
- Batch effect correction
- Differential analysis on cell abundances
- Differential analysis on expressions

Model

Suppose we have $b = 1, \cdots, B$ batches, $c = 1, \cdots, C$ conditions (treatments), and $s = 1, \cdots, S$ subjects. In total, we have $n = 1, \cdots, N$ cell events, $m = 1, \cdots, M$ protein markers, and $k = 1, \cdots, K$ cell types. We further assume that we desire to reduce the dimension of the data to D.

Data
Denote by $y_n \in \mathbb{R}^{1\times M}$ the nth cell event. Therefore, the observed (stacked) data would be an $N \times M$ expression matrix.

Denote by $F^B \in \mathbb{R}^{N\times B}$ the fixed effect design matrix for batch effect. We will use one-hot encoding.

Similarly, let $F^C \in \mathbb{R}^{N\times C}$ be the fixed effect design matrix for condition (treatment) effect. We will use one-hot encoding.

Finally, let $R^S \in \mathbb{R}^{N\times S}$ be the random effect design matrix for subject effect. We will use one-hot encoding.
Global parameters
Denote by $\sigma^2_{\gamma^\Pi} \in \mathbb{R}$ the variance of each component of the patient random effect on cell type probabilities.

Denote by $\sigma^2_{\gamma^{z,\mu}} \in \mathbb{R}$ the variance of each component of the patient random effect on the mean of cell expressions.

Denote by $\sigma^2_{\gamma^{z,\Sigma}} \in \mathbb{R}$ the variance of each component of the patient random effect on the variance of cell expressions.

Denote by $\sigma^2_{\gamma^{w,\Sigma}} \in \mathbb{R}$ the variance of each component of the patient random effect on the mean of zero inflation.

Given the variances, we now define the coefficients for various random effects:
Denote by $\gamma^\Pi \in \mathbb{R}^{S\times K}$ the patient random effect on cell type probabilities.

Denote by $\gamma^{z,\mu} \in \mathbb{R}^{S\times M\times D}$ the patient random effect on means of cell expressions.

Denote by $\gamma^{z,\Sigma} \in \mathbb{R}^{S\times M\times D}$ the patient random effect on variances of cell expressions.

Denote by $\gamma^{w,\mu} \in \mathbb{R}^{S\times M\times D}$ the patient random effect on zero probabilities.

Now we define the coefficients for fixed effects:
Denote by $\beta^\Pi \in \mathbb{R}^{C\times K}$ the condition effect on cell type probabilities.

Denote by $\beta^{z,\mu} \in \mathbb{R}^{C\times M\times D}$ the condition effect on means of cell expressions.

Denote by $\beta^{z,\Sigma} \in \mathbb{R}^{C\times M\times D}$ the condition effect on variances of cell expressions.

Denote by $\beta^{w,\mu} \in \mathbb{R}^{C\times M\times D}$ the condition effect on zero probabilities.

Denote by $\alpha^{z,\mu} \in \mathbb{R}^{B\times M\times D}$ the batch effect on means of cell expressions.

Denote by $\alpha^{z,\Sigma} \in \mathbb{R}^{B\times M\times D}$ the batch effect on variances of cell expressions.

Denote by $\alpha^{w,\mu} \in \mathbb{R}^{B\times M\times D}$ the batch effect on zero probabilities.

Local parameters

Denote by $\Pi_n \in \mathbb{R}^{1\times K}$ the one hot encoding of cell types

Denote by $x_n \in \mathbb{R}^{1\times D}$ the low-dimensional embedding of the nth cell.

Denote by $w_n \in \mathbb{R}^{1\times M}$ the zero probability of a protein marker.

Denote by $z_n \in \mathbb{R}^{1\times M}$ the expression of a protein marker.

Hyperparameter

Let $\Theta \in \mathbb{R}^{K\times D}$ be the collection of cell prototypes. Therefore, each row of Θ, $\Theta_{k,\cdot}$ represents the cell prototype of the kth cell type in the D-dimensional space. The prototypes will be learned via gradient descent.

Denote by σ_x^2 the variance of the low-dimensional cell embedding. This parameter is fixed apriori.

Denote by σ_y^2 the normal variance of the mollified uniform. This parameter is fixed apriori.

Denote by δ the log variance (in the case of Gaussian noise) or the length of support (in the case of uniform noise). This parameter can be learned or can be fixed apriori.

Priors

$$
\begin{aligned}
\log(\sigma^2_{\gamma^\Pi}) &\sim N(\mu_0, \sigma_0^2)\\
\log(\sigma^2_{\gamma^{z,\mu}}) &\sim N(\mu_1, \sigma_1^2)\\
\log(\sigma^2_{\gamma^{z,\Sigma}}) &\sim N(\mu_2, \sigma_2^2)\\
\log(\sigma^2_{\gamma^{w,\mu}}) &\sim N(\mu_3, \sigma_3^2)\\
\beta^{\Pi}_{c,\cdot} &\sim N(\mathbf{0}, \sigma^2_{\beta^\Pi}\mathbf{I})\\
\beta^{z,\mu}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\beta^{z,\mu}}\mathbf{I})\\
\beta^{z,\Sigma}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\beta^{z,\Sigma}}\mathbf{I})\\
\beta^{w,\mu}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\beta^{w,\mu}}\mathbf{I})\\
\alpha^{z,\mu}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\alpha^\Pi}\mathbf{I})\\
\alpha^{z,\Sigma}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\alpha^{z,\Sigma}}\mathbf{I})\\
\alpha^{w,\mu}_{c,\cdot,d} &\sim N(\mathbf{0}, \sigma^2_{\alpha^{w,\mu}}\mathbf{I})
\end{aligned}
$$

Full likelihood & priors

Let

$$v_n = \{z_n, w_n, x_n, \Pi_n\}$$

$$\Delta = \{\sigma^2_{\gamma^\Pi}, \sigma^2_{\gamma^{z,\mu}}, \sigma^2_{\gamma^{z,\Sigma}}, \sigma^2_{\gamma^{w,\mu}}, \beta^\Pi, \beta^{z,\mu}, \beta^{z,\Sigma}, \beta^{w,\mu}, \alpha^{z,\mu}, \alpha^{z,\Sigma}, \alpha^{w,\mu}, \gamma^{z,\mu}, \gamma^{z,\Sigma}, \gamma^{w,\mu}, \gamma^\Pi\}$$

$$
\begin{aligned}
p(\{y_n, v_n\}_{n=1}^N, \Delta|F^B, F^C, R^S) &= p(\{y_n\}_{n=1}^N|\{v_n\}_{n=1}^N, \Delta, F^B, F^C, R^S)p(\{v_n\}_{n=1}^N|\Delta, F^B, F^C, R^S)p(\Delta|F^B, F^C, R^S)\\
&= [\prod_{n=1}^N p(y_n|v_n, \Delta, F^B, F^C, R^S)][\prod_{n=1}^N p(v_n|\Delta, F^B, F^C, R^S)]p(\Delta|F^B, F^C, R^S)
\end{aligned}
$$

We will look at $p(\Delta|F^B, F^C, R^S)$ first.

$$
\begin{aligned}
p(\Delta|F^B, F^C, R^S) &= p(\sigma^2_{\gamma^\Pi})p(\sigma^2_{\gamma^{z,\mu}})p(\sigma^2_{\gamma^{z,\Sigma}})p(\sigma^2_{\gamma^{w,\mu}})* \\
&\quad p(\gamma^\Pi|\sigma^2_{\gamma^\Pi})p(\gamma^{z,\mu}|\sigma^2_{\gamma^{z,\mu}})p(\gamma^{z,\Sigma}|\sigma^2_{\gamma^{z,\Sigma}})p(\gamma^{w,\mu}|\sigma^2_{\gamma^{w,\mu}})* \\
&\quad p(\beta^\Pi)p(\beta^{z,\mu})p(\beta^{z,\Sigma})p(\beta^{w,\mu})* \\
&\quad p(\alpha^{z,\mu})p(\alpha^{z,\Sigma})p(\alpha^{w,\mu}) \\
&= logN(\sigma^2_{\gamma^\Pi}|\mu_0, \sigma_0^2)logN(\sigma^2_{\gamma^{z,\mu}}|\mu_1, \sigma_1^2)logN(\sigma^2_{\gamma^{z,\Sigma}}|\mu_2, \sigma_2^2)logN(\sigma^2_{\gamma^{w,\mu}}|\mu_3, \sigma_3^2)* \\
&\quad [\prod_{s=1}^{S} N(\gamma^\Pi_{s,\cdot}|\mathbf{0}, \sigma^2_{\gamma^\Pi}\mathbf{I})][\prod_{d=1}^{D}\prod_{s=1}^{S} N(\gamma^{z,\mu}_{s,\cdot,d}|\mathbf{0}, \sigma^2_{\gamma^{z,\mu}}\mathbf{I})][\prod_{d=1}^{D}\prod_{s=1}^{S} N(\gamma^{z,\Sigma}_{s,\cdot,d}|\mathbf{0}, \sigma^2_{\gamma^{z,\Sigma}}\mathbf{I})][\prod_{d=1}^{D}\prod_{s=1}^{S} N(\gamma^{w,\mu}_{s,\cdot,d}|\mathbf{0}, \sigma^2_{\gamma^{w,\mu}}\mathbf{I})]* \\
&\quad [\prod_{c=1}^{C} N(\beta^\Pi_{c,\cdot}|\mathbf{0}, \sigma^2_{\beta^\Pi}\mathbf{I})][\prod_{d=1}^{D}\prod_{c=1}^{C} N(\beta^{z,\mu}_{c,\cdot,d}|\mathbf{0}, \sigma^2_{\beta^{z,\mu}}\mathbf{I})][\prod_{d=1}^{D}\prod_{c=1}^{C} N(\beta^{z,\Sigma}_{c,\cdot,d}|\mathbf{0}, \sigma^2_{\beta^{z,\Sigma}}\mathbf{I})][\prod_{d=1}^{D}\prod_{c=1}^{C} N(\beta^{w,\mu}_{c,\cdot,d}|\mathbf{0}, \sigma^2_{\beta^{w,\mu}}\mathbf{I})]* \\
&\quad [\prod_{d=1}^{D}\prod_{b=1}^{B} N(\alpha^{z,\mu}_{b,\cdot,d}|\mathbf{0}, \sigma^2_{\alpha^{z,\mu}}\mathbf{I})][\prod_{d=1}^{D}\prod_{b=1}^{B} N(\alpha^{z,\Sigma}_{b,\cdot,d}|\mathbf{0}, \sigma^2_{\alpha^{z,\Sigma}}\mathbf{I})][\prod_{d=1}^{D}\prod_{b=1}^{B} N(\alpha^{w,\mu}_{b,\cdot,d}|\mathbf{0}, \sigma^2_{\alpha^{w,\mu}}\mathbf{I})]
\end{aligned}
$$

We then look at $p(v_n|\Delta, F^B, F^C, R^S)$

$$
\begin{aligned}
p(v_n|\Delta, F^B, F^C, R^S) &= p(\Pi_n|\beta^\Pi, \gamma^\Pi, F^B, F^C, R^S)p(x_n|\Pi_n)* \\
&\quad p(w_n|x_n, \beta^{w,\mu}, \alpha^{w,\mu}, F^B, F^C, R^S)* \\
&\quad p(z_n|x_n, \beta^{z,\mu}, \beta^{z,\Sigma}, \alpha^{z,\mu}, \alpha^{z,\Sigma}, F^B, F^C, R^S) \\
&= \text{Categorical}([\frac{\exp(F_n^B\beta^\Pi_{\cdot,1} + R_n^S\gamma^\Pi_{\cdot,1})}{\sum_{k=1}^K \exp(F_n^B\beta^\Pi_{\cdot,k} + R_n^S\gamma^\Pi_{\cdot,k})}, \cdots, \frac{\exp(F_n^B\beta^\Pi_{\cdot,K} + R_n^S\gamma^\Pi_{\cdot,K})}{\sum_{k=1}^K \exp(F_n^B\beta^\Pi_{\cdot,k} + R_n^S\gamma^\Pi_{\cdot,k})}])* \\
&\quad N(x_n|\Pi_n\Theta, \sigma_x^2\mathbf{I})* \\
&\quad p(w_n|x_n, \beta^{w,\mu}, \alpha^{w,\mu}, F^B, F^C, R^S)* \\
&\quad p(z_n|x_n, \beta^{z,\mu}, \beta^{z,\Sigma}, \alpha^{z,\mu}, \alpha^{z,\Sigma}, F^B, F^C, R^S)
\end{aligned}
$$

If there is no noise, and therefore, the observations are truly zero-inflated,

$$
\begin{aligned}
p(w_n|x_n, \beta^{w,\mu}, \alpha^{w,\mu}, F^B, F^C, R^S) &= \text{Delta}(1) \\
p(z_n|x_n, \beta^{z,\mu}, \beta^{z,\Sigma}, \alpha^{z,\mu}, \alpha^{z,\Sigma}, \beta^{w,\mu}, \alpha^{w,\mu}, F^B, F^C, R^S) &= ZILN(z_n|\mu_z(x_n) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{z,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{z,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{z,\mu}, x_n), \\
&\quad \exp(\log(\Sigma_z(x_n)) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{z,\Sigma}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{z,\Sigma}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{z,\Sigma}, x_n)), \\
&\quad \text{logit}^{-1}(\mu_w(x_n) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{w,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{w,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{w,\mu}, x_n)))
\end{aligned}
$$

If the noise is Gaussian or uniform

$$
\begin{aligned}
p(w_n|x_n, \beta^{w,\mu}, \alpha^{w,\mu}, F^B, F^C, R^S) &= N(w_n|(\mu_w(x_n) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{w,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{w,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{w,\mu}, x_n)), \\
&\quad \exp(\log(\Sigma_w(x_n)))) \\
p(z_n|x_n, \beta^{z,\mu}, \beta^{z,\Sigma}, \alpha^{z,\mu}, \alpha^{z,\Sigma}, F^B, F^C, R^S) &= N(z_n|\mu_z(x_n) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{z,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{z,\mu}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{z,\mu}, x_n), \\
&\quad \exp(\log(\Sigma_z(x_n)) + \text{einsum}(\text{“}nb, bmd, nd \to nm\text{”}, F_n^B, \alpha^{z,\Sigma}, x_n)+ \\
&\quad + \text{einsum}(\text{“}nc, cmd, nd \to nm\text{”}, F_n^C, \beta^{z,\Sigma}, x_n)+ \\
&\quad + \text{einsum}(\text{“}ns, smd, nd \to nm\text{”}, R_n^S, \gamma^{z,\Sigma}, x_n)))
\end{aligned}
$$

Finally, we look at $p(y_n|v_n, \Delta, F^B, F^C, R^S)$
If there is no noise, and therefore, the observations are truly zero-inflated

$$\begin{aligned}p(y_n|v_n, \Delta, F^B, F^C, R^S) &= p(y_n|w_n, z_n, F^B, F^C, R^S)\\ &= \mathrm{Delta}([\frac{1}{1+\exp(-w_n)}]\exp(z_n))\end{aligned}$$

If the noise is Gaussian,

$$\begin{aligned}p(y_n|v_n, \Delta, F^B, F^C, R^S) &= p(y_n|w_n, z_n, F^B, F^C, R^S)\\ &= N([\frac{1}{1+\exp(-w_n)}]\exp(z_n), \exp(\delta)\mathbf{I})\end{aligned}$$

If the noise is uniform,

$$\begin{aligned}p(y_n|v_n, \Delta, F^B, F^C, R^S) &= p(y_n|w_n, z_n, F^B, F^C, R^S)\\ &= MU([\frac{1}{1+\exp(-w_n)}]\exp(z_n) - \exp(\delta)\mathbf{I}, [\frac{1}{1+\exp(-w_n)}]\exp(z_n), \sigma_y^2\mathbf{I})\end{aligned}$$

Variational distributions

$$q(\{v_n\}_{n=1}^N, \Delta|\{y_n\}_{n=1}^N, F^B, F^C, R^S) = q(\Delta|\{y_n, v_n\}_{n=1}^N, F^B, F^C, R^S)q(\{v_n\}_{n=1}^N|\{y_n\}_{n=1}^N, F^B, F^C, R^S)$$

We first look at $q(\Delta|\{y_n, v_n\}_{n=1}^N, F^B, F^C, R^S)$
Our approximation is mean-field.

$$\begin{aligned}q(\Delta|\{y_n, v_n\}_{n=1}^N, F^B, F^C, R^S) &= q(\Delta)\\ &= q(\sigma^2_{\gamma^\Pi}|\gamma^\Pi)q(\sigma^2_{\gamma^{z,\mu}}|\gamma^{z,\mu})q(\sigma^2_{\gamma^{z,\Sigma}}|\gamma^{z,\Sigma})q(\sigma^2_{\gamma^{w,\mu}}|\gamma^{w,\mu})*\\ &\quad q(\gamma^\Pi)q(\gamma^{z,\mu})q(\gamma^{z,\Sigma})q(\gamma^{w,\mu})*\\ &\quad q(\beta^\Pi)q(\beta^{z,\mu})q(\beta^{z,\Sigma})q(\beta^{w,\mu})*\\ &\quad q(\alpha^{z,\mu})q(\alpha^{z,\Sigma})q(\alpha^{w,\mu})\\ &= logN(\sigma^2_{\gamma^\Pi}|\log(var(\gamma^\Pi)) - \frac{\sigma_0^2}{2}, \sigma_0^2)logN(\sigma^2_{\gamma^{z,\mu}}|\log(var(\gamma^{z,\mu})) - \frac{\sigma_1^2}{2}, \sigma_1^2)*\\ &\quad logN(\sigma^2_{\gamma^{z,\Sigma}}|\log(var(\gamma^{z,\Sigma})) - \frac{\sigma_2^2}{2}, \sigma_2^2)logN(\sigma^2_{\gamma^{w,\mu}}|\log(var(\gamma^{w,\mu})) - \frac{\sigma_3^2}{2}, \sigma_3^2)*\\ &\quad N(\gamma^\Pi|U_{\gamma^\Pi}, S^2_{\gamma^\Pi})N(\gamma^{z,\mu}|U_{\gamma^{z,\mu}}, S^2_{\gamma^{z,\mu}})N(\gamma^{z,\Sigma}|U_{\gamma^{z,\Sigma}}, S^2_{\gamma^{z,\Sigma}})N(\gamma^{w,\mu}|U_{\gamma^{w,\mu}}, S^2_{\gamma^{w,\mu}})*\\ &\quad N(\beta^\Pi|U_{\beta^\Pi}, S^2_{\beta^\Pi})N(\beta^{z,\mu}|U_{\beta^{z,\mu}}, S^2_{\beta^{z,\mu}})N(\beta^{z,\Sigma}|U_{\beta^{z,\Sigma}}, S^2_{\beta^{z,\Sigma}})N(\beta^{w,\mu}|U_{\beta^{w,\mu}}, S^2_{\beta^{w,\mu}})*\\ &\quad N(\alpha^{z,\mu}|U_{\alpha^{z,\mu}}, S^2_{\alpha^{z,\mu}})N(\alpha^{z,\Sigma}|U_{\alpha^{z,\Sigma}}, S^2_{\alpha^{z,\Sigma}})N(\alpha^{w,\mu}|U_{\alpha^{w,\mu}}, S^2_{\alpha^{w,\mu}})\end{aligned}$$

We then look at $q(\{v_n\}_{n=1}^N|\{y_n\}_{n=1}^N, F^B, F^C, R^S)$
Again, our approximation is mean-field. And we let

$$q(\{v_n\}_{n=1}^N|\{y_n\}_{n=1}^N, F^B, F^C, R^S) = \prod_{n=1}^N q(v_n|y_n, F^B, F^C, R^S)$$

Therefore,

$$\begin{aligned}q(v_n|y_n, F^B, F^C, R^S) &= q(\Pi_n|x_n, F^B, F^C, R^S)q(x_n|z_n, w_n, F^B, F^C, R^S)*\\ &\quad q(w_n|y_n, F^B, F^C, R^S)q(z_n|y_n, F^B, F^C, R^S)\\ &= \mathrm{Delta}(\text{nearest prototype})N(x_n|U_x([\frac{1}{1+\exp(-w_n)}]\exp(z_n), F_n^B, F_n^C, R_n^S), s_x^2\mathbf{I})*\\ &\quad N(w_n|U_w(y_n), S_w^2(y_n))N(z_n|U_z(y_n), S_z^2(y_n))\end{aligned}$$