Zero-inflation in the Multivariate Poisson Lognormal Family

Bastien Batardière, François Gindraud, Julien Chiquet and Mahendra Mariadassou

Université Paris-Saclay, AgroParisTech, INRAE, UMR MIA Paris-Saclay, MaIAGE

May 22, 2024

- In Single cell analysis, it is usual to deal with high-dimensional count data:

$$\boldsymbol{Y} = \begin{pmatrix} 12 & 0 & \cdots & 0 & 9 \\ 2 & 0 & \cdots & 0 & 0 \\ \vdots & & & & \vdots \\ 341 & 5 & \cdots & 1 & 0 \end{pmatrix}$$

- Y_{ij}: count of transcript j in cell i
- **Non-continuous data** $\implies$ Linear Gaussian models do not apply
- **High percentage of zeros ($\approx 90\%$)** $\implies$ zero-inflation is needed

- Dataset:
  - $\boldsymbol{Y}$: $n \times p$ count matrix ($n \approx p \approx 10^4$)
  - $\boldsymbol{X}$: $n \times d$ or $d \times p$ covariates ($d \approx 10$)
- Parameter $\theta = (\mathbf{B}, \boldsymbol{\Sigma}, \boldsymbol{\pi})$
  - $\mathbf{B} \in \mathbb{R}^{d \times p}$ regression coefficient.
  - $\boldsymbol{\Sigma} \in \mathbb{R}^{p \times P}$ covariance matrix.
  - $\boldsymbol{\pi} \in \mathbb{R}^{n \times p}$ zero-inflation coefficient.
- Model:

$$
\begin{aligned}
\mathbf{W_i} &\sim \mathcal{B}(\pi_i) \\
\boldsymbol{Z_i} &\sim \mathcal{N}(\mathbf{XB}, \boldsymbol{\Sigma}) \\
(Y_{ij} \mid Z_{ij}, W_{ij}) &\sim (1 - W_{ij})\, \mathcal{P}(\exp(Z_{ij}))
\end{aligned}
$$

The zero-inflation can take several forms:

$$\pi_{ij} = \pi \in [0,1] \qquad \text{(non-dependent)}$$

$$\pi_{ij} = \sigma\left(\mathbf{X}\mathbf{B}^0\right)_{ij}, \ \mathbf{X} \in \mathbb{R}^{n\times d}, \ \mathbf{B}^0 \in \mathbb{R}^{d\times p} \qquad \text{(column-wise dependence)}$$

$$\pi_{ij} = \sigma\left(\mathbf{B}^0\mathbf{X}\right)_{ij}, \mathbf{B}^0 \in \mathbb{R}^{n\times d}, \ \mathbf{X} \in \mathbb{R}^{d\times p} \qquad \text{(row-wise dependence)}$$