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

June 19, 2024

Outline

Motivation

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

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

- Y_{ij}: count of transcript j in cell i
- **Non-continuous data** $\implies$ Linear Gaussian models do not apply
- **High percentage of zeros (**$\approx 80\%$**)** $\implies$ zero-inflation is needed

Model

- Dataset:
  - $\mathbf{Y}$: $n \times p$ count matrix ($n \approx p \approx 10^4$)
  - $\mathbf{X}$: $n \times d$ 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:

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

Modelling the zero inflation

The zero-inflation can take several forms:

- $\pi_{ij} = \pi \in [0, 1]$ (non-dependent)
- $\pi_{ij} = \sigma\left(\mathbf{X}\mathbf{B}^0\right)_{ij}, \ \mathbf{X} \in \mathbb{R}^{n \times d}, \ \mathbf{B}^0 \in \mathbb{R}^{d \times p}$ (column-wise dependence)
- $\pi_{ij} = \sigma\left(\bar{\mathbf{B}}^0\bar{\mathbf{X}}\right)_{ij}, \bar{\mathbf{B}}^0 \in \mathbb{R}^{n \times d}, \ \bar{\mathbf{X}} \in \mathbb{R}^{d \times p}$ (row-wise dependence)

Inference

- We aim at solving:
$$\boldsymbol{\theta^\star = \underset{\theta}{\mathbf{argmax}} \sum_{i=1}^{n} \log p_\theta(Y_i)}$$
- Exact inference and Expectation-Maximization (EM) are untractable
- Approximate solution via Variational EM (**VEM**), maximizing the tractable Evidence Lower BOund (**ELBO**):
$$J_Y(\theta, q) \triangleq \log p_\theta(\mathbf{Y}) - KL\left[q(\cdot) \| p_\theta(\cdot \mid \mathbf{Y})\right]$$
where q is a variational distribution approximating $p_\theta(\cdot \mid \mathbf{Y})$

Variational EM

- VE step: update the variational parameters ψ: choose the best approximation q_ψ of $p_\theta(\cdot|\mathbf{Y})$

$$\begin{aligned}\psi^{(h+1)} &= \arg\max_{\psi} J_Y(\theta^{(h)}, q_\psi) \\ &= \arg\min_{\psi} KL\left[q_\psi(Z, W) \| p_{\theta^{(h)}}(Z, W \mid Y)\right]\end{aligned}$$

- M step: update θ (usually via closed forms) :

$$\begin{aligned}\theta^{(h+1)} &= \arg\max_{\theta} J_Y(\theta, q_{\psi^{(h+1)}}) \\ &= \arg\max_{\theta} \mathbb{E}_{q_{\psi^{(h+1)}}}\left[\log p_\theta(Y, Z, W)\right]\end{aligned}$$

Choice of the variational family

- **Standard** variational approximation: $\mathbf{Z}|\mathbf{Y} \perp\!\!\!\perp \mathbf{W}|\mathbf{Y}$:

$$q^{(1)}_{\psi_i}(\mathbf{Z}_i, \mathbf{W}_i) \triangleq q_{\psi_i}(\mathbf{Z}_i) q_{\psi_i}(\mathbf{W}_i) = \otimes_{j=1}^{p} q_{\psi_i}(Z_{ij}) q_{\psi_i}(W_{ij})$$
$$= \boxed{\otimes_{j=1}^{p} \mathcal{N}\left(M_{ij}, S_{ij}^2\right) \mathcal{B}\left(P_{ij}\right)}.$$

- **Enhanced** variational approximation: Use dependance

$$Z_{ij}|W_{ij}, Y_{ij} = \left(Z_{ij}|Y_{ij}, W_{ij} = 1\right)^{W_{ij}} \left(Z_{ij}|Y_{ij}, W_{ij} = 0\right)^{1-W_{ij}}.$$

giving

$$q^{(2)}_{\psi_i}(\mathbf{Z}_i, \mathbf{W}_i) = \boxed{\otimes_{j=1}^{p} \mathcal{N}(\mu_j, \Sigma_{jj})^{W_{ij}} \mathcal{N}(M_{ij}, S_{ij}^2)^{1-W_{ij}} W_{ij}},$$

with $W_{ij} \sim^{\text{indep}} \mathcal{B}\left(P_{ij}\right)$. Variational parameters are $\psi_{ij} = (M_{ij}, S_{ij}, P_{ij})$.

- **Bi-concavity** holds for the standard variational approximation.

Analytic law of W_{ij}

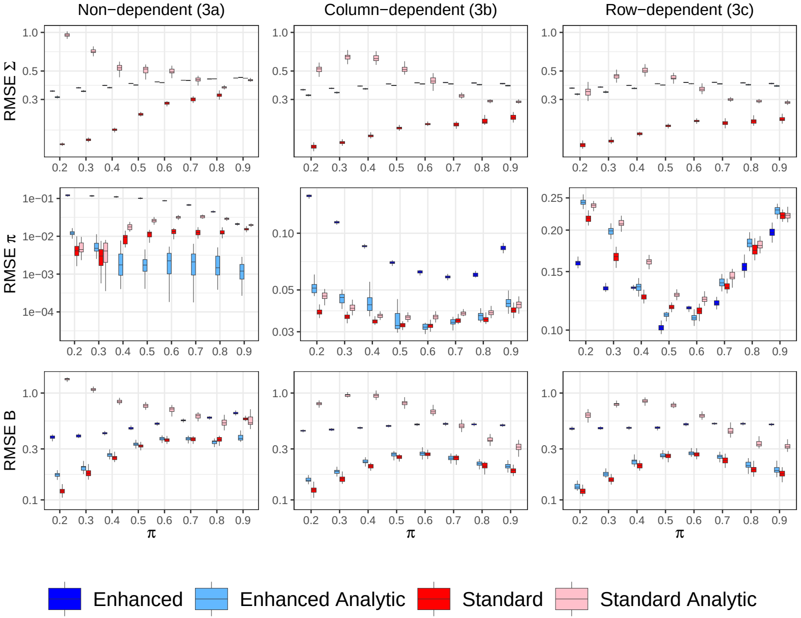
The conditional law $W_{ij} | Y_{ij} = 0$ is known and given by:

$$W_{ij} | Y_{ij} \sim \mathcal{B}\left(\frac{\pi_{ij}}{\varphi\left(\mathbf{X}_i^\top \beta_j, \Sigma_{jj}\right)(1-\pi_{ij}) + \pi_{ij}}\right) \mathbf{1}_{Y_{ij}=0},$$

with φ given by the Lambert function

- P_{ij} can be removed from optimization
- bi-concavity lost due to non-concavity of φ.

Simulations



Applications on real data: ZIPLN vs PLN

- Data: Microbiota of 45 lactating cows $\implies$ $n = 899$ samples .
- After removing Amplicon Sequence Variants (ASV) with more than 5 % prevalence $\implies$ $p = 259$ ASV.
- Still 90 % of zeros.

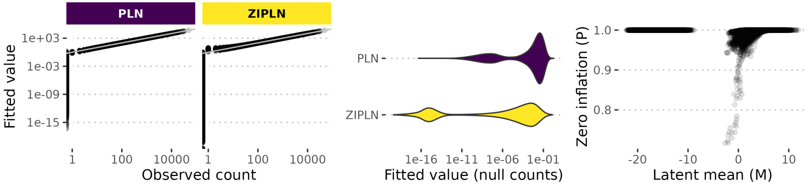
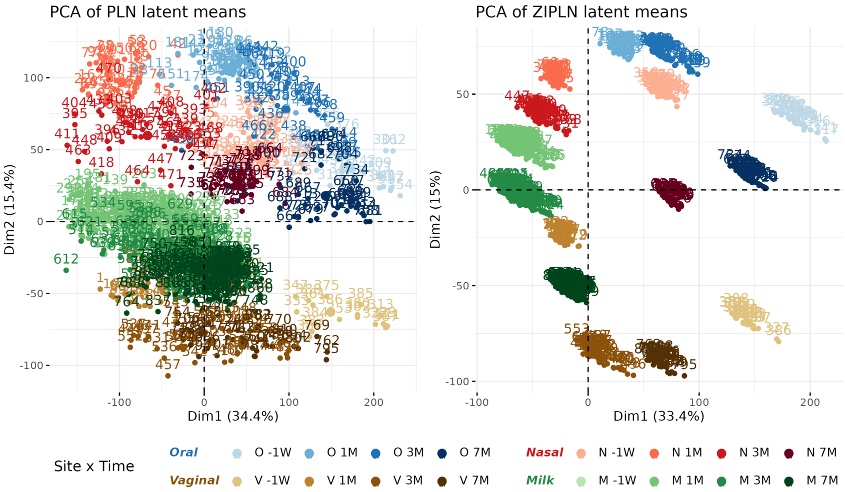


Figure: Model fits of PLN and ZIPLN in terms of fitted versus observed counts (left panel), fitted values for null counts (middle panel) and comparaison of P_{ij} and M_{ij} estimated for null counts by ZIPLN(right panel).

Applications on real data: PLN vs ZIPLN



Thanks for your attention