Zero-inflation in the MPLN Family

Bastien Batardière [1], Julien Chiquet [1], François Gindraud [2]& Mahendra Mariadassou [3]

[1] *MIA Paris-Saclay, Université Paris-Saclay, AgroParisTech, INRAE, France,* {*prenom.nom*}*@inrae.fr*
[2] *Université de Lyon, Université Lyon 1, CNRS, Laboratoire de Biométrie et Biologie Evolutive UMR 5558, francois.gindraud@inria.fr*
[3] *Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France, mahendra.mariadassou@inrae.fr*

Résumé. Les données de comptage en haute dimension sont difficiles à analyser telles quelles, et les approches basées sur des modèles statistiques restent efficaces et appropriées tout en préservant l'interprétabilité. Le modèle Poisson-Log-Normal (PLN) (multivarié) en est un exemple et suppose que les données de comptage sont influencées par une variable gaussienne latente structurée, exploitant les dépendances entre les comptages. Bien que les données de comptage du monde réel soient fréquemment caractérisées par des excès de zéros, un tel modèle ne prend pas en compte cette réalité. Nous proposons ici le modèle Zero-Inflated PLN (ZI-PLN), ajoutant une composante multivariée à excès de zéros au modèle, sous la forme d'une variable latente Bernoulli supplémentaire. L'inflation de zéros peut être fixe, spécifique au site, spécifique à la variable ou dépendre de covariables. Nous estimons les paramètres du modèle en utilisant une inférence variationnelle et comparons deux approximations : (i) distributions variationnelles gaussiennes et bernoulli indépendantes ou (ii) distribution gaussienne variationnelle conditionnée à la distribution bernoulli. La méthode est évaluée sur des données synthétiques. Tous les algorithmes sont disponibles dans un package Python `pyPLNmodels` et un package R `PLNmodels`.

Mots-clés. Données de comptages, modèle Poisson Log Normal, zéro-inflation, inférence variationnelle, optimisation alternée.

Abstract. High-dimensional count data are hard to analyze as is and statistical model-based approaches remain efficient and adequate, while preserving explainability. The (multivariate) Poisson-Log-Normal (PLN) model is one of them and assumes count data are driven by an underlying structured latent Gaussian variable, exploiting dependencies between counts. While real-world count data are frequently zero-inflated, such a model does not account for this reality. Here we propose the Zero-Inflated PLN (ZI-PLN) model, adding a multivariate zero-inflated component to the model, as an additional Bernoulli latent variable. The Zero-Inflation can be fixed, site-specific, feature-specific or depend on covariates. We estimate model parameters using variational inference and compare two approximations: (i) independent Gaussian and Bernoulli variational distributions or (ii) Gaussian variational conditioned on the Bernoulli one. The method is assessed on synthetic data. All the algorithms are available in a python package `pyPLNmodels` and an R package `PLNmodels`

Keywords. Count data, Poisson lognormal model, zero inflated model, variational Inference, alternate optimisation

1 Introduction

Count data appears in many different fields such as ecology, accidents analysis, single-cell RNA (scRNA) sequencing and metagenomics. For example, researchers may be interested in estimating the correlation between abundances of different species or expression of different genes in a cell. More specifically, the model introduced in this paper is motivated by the increasing importance of microbiome studies. Broadly speaking, a microbiome is a collection of microbes, together with their genomes, found in a given habitat (*e.g.* plant leaves, human gut, waste water, etc.). The most widespread way of studying microbiomes is to amplify and sequence a marker gene, which acts as a molecular barcode. The sequences are processed through bioinformatics pipelines [Escudié et al., 2017] to produce Operational Taxonomic Units (OTUs) / Amplicon Sequence Variants (ASVs), a proxy for microbial species in microbial ecology, and enumerated to create count tables, recording the abundance of each OTU/ASV in each sample. Those tables are characterized by a very high fraction (ranging from 80 to 95%) of zero counts and a high number of variables.

The (multivariate) Poisson-Log-Normal [in short PLN, see Aitchison and Ho, 1989] model offers a general framework to multivariate count data by offering flexibility to describe dependencies between counts by means of a latent Gaussian variable. By means of Poisson emission law with Log-Normal parameter, PLN models results in a overdispersed models. The underlying Gaussian structure inherent to the PLN model makes correlation between variables natural, unlike its NB counterpart. More generally, the PLN model falls in the family of latent variable models (LVMs), and more specifically of multivariate generalized linear mixed models (mGLMMs) sometimes also called generalized linear latent variable models (GLLVMs). In those models, the distribution of observed responses usually belongs either to the exponential family (Bernoulli, Binomial, Poisson, Negative-Binomial, with or without Zero-Inflation, etc.) or the exponential dispersion model (Tweedie, etc.). Model parameters are related to linear combinations of latent variables (and possibly covariates) through a simple link function. Parameter estimation for common GLLVMs is efficiently implemented in some packages [Niku et al., 2019, Seabold and Perktold, 2010], making a popular option for multivariate count data. However, while some models allows for dependency between variables in the latent space and other accounts for zero-inflation, no model accounts, to the best of our knowledge, for both at the same time.

We introduce here the Zero-Inflated Poisson Log-Normal (ZI-PLN) model, based on the Poisson Log-Normal (PLN) model. ZI-PLN benefits from the Gaussian structure of the PLN model, with an extra zero-inflated component. This extra layer adds flexibility to the model as its parameters can be chosen to be shared across the individuals, across the genes or even to depend on its own set of covariates. As exact inference of (ZI)PLN is intractable and conditional laws are only partially known, we cannot rely on the Expectation-Maximization (EM) algorithm [Dempster et al., 1977], as done for optimizing classical latent variable models. We rely on variational inference [Jaakkola and Jordan, 2000, Wainwright and Jordan, 2008, Hui et al., 2017, Blei et al., 2017]. Other approaches based on Monte Carlo techniques have been proposed [JAC, 2007, Cap] to infer the maximum likelihood estimator, but it does not scale with the dimension of the observations. Numerical integration can be performed [Aitchison and Ho, 1989] as an alternative to the variational approximation used

here but becomes prohibitive when the number of dimensions exceeds 5. Here, we develop a Variational-EM algorithm where we propose two different variational approximations. The first assumes conditional independence between both components, resulting in a fast M step. By contrast, the second is slightly slower but leverages the dependence between components to use a more complex variational approximation.

Related work ZINBWaVE, proposed by Risso et al. [2018] is the closest work to ours, modelling zero-inflation (resp. counts) as a logistic (resp. log-linear) regression involving sample-level, gene-level and (unobserved) sample-level covariates, where the unobserved covariates are presumed to be unwanted variations and captured through latent factors. This model however suffers from a lack of identifiability and is mostly interested in estimating the probability that a null count arises from zero-inflation. We distinguish ourselves from ZINBWaVE via identifiability of parameters and most importantly via the inherent and explicit dependency structure between variables.

2 Model

Background: Multivariate Poisson lognormal-model The multivariate Poisson lognormal model relates a p-dimensional observation count vector $\mathbf{Y}_i = (Y_{i1}, \ldots, Y_{ip}) \in \mathbb{N}^p$ to a p-dimensional vector of Gaussian latent variables $\mathbf{Z}_i \in \mathbb{N}^p$ with precision matrix $\mathbf{\Omega}$ (that is, covariance matrix $\mathbf{\Sigma} \triangleq \mathbf{\Omega}^{-1}$). We adopt a formulation of PLN close to a multivariate generalized linear model, where the main effect is due to a linear combination of d covariates $\mathbf{x}_i \in \mathbb{R}^d$ (including a vector of intercepts). We also let the possibility to add some offsets for the p variables in in each sample, that is $\boldsymbol{o}_i \in \mathbb{R}^p$:

$$\begin{aligned} \text{latent space} \quad & \mathbf{Z}_i \sim \mathcal{N}(\mathbf{o}_i + \mathbf{x}_i^\intercal \boldsymbol{B}, \mathbf{\Omega}^{-1}), \\ \text{observation space} \quad & Y_{ij} \,|\, Z_{ij} \quad \text{indep.} \qquad \boldsymbol{Y}_i | \boldsymbol{Z}_i \sim \mathcal{P}\left(\exp\{\boldsymbol{Z}_i\}\right). \end{aligned} \tag{1}$$

The $d \times p$ matrix $\boldsymbol{B}$ is the latent matrix of regression parameters. The latent covariance matrix $\mathbf{\Sigma}$ describes the underlying residual structure of dependence between the p variables, once the covariates are accounted for. We denote by $\boldsymbol{Y}, \boldsymbol{O}, \boldsymbol{X}$ the observed matrices with respective sizes $n \times p, n \times p$ and $n \times d$ stacking row-wise the vectors of counts, offsets and covariates (respectively $\mathbf{Y}_i, \mathbf{x}_i$ and $\mathbf{o}_i$). We also denote by $\boldsymbol{Z}$ the $n \times p$ matrix of unobserved latent Gaussian vectors $\mathbf{Z}_i$.

Zero-inflated PLN regression model We now aim to model an excess of zeros in the data by adding zero-inflation to the standard PLN model eq. (1), so that the zeros in $\boldsymbol{Y}_i$ arise from two different sources: either from a component where zero is the only possible value, or from a standard PLN component like eq. (1). This two-component mixture is described thanks to an additional latent vector $\boldsymbol{W}_i = (W_{i1,\ldots,ip}) \in \mathbb{R}^p$ of Bernoulli random variables, parametrized by probabilities $\pi_i = (\pi_{i1}, \ldots, \pi_{ip})$ describing the probability that variable j in

sample i belongs to the pure zero component:

$$\begin{aligned}
\text{PLN latent space} \quad & \boldsymbol{Z}_i = (Z_{ij})_{j=1\ldots p} && \sim \mathcal{N}(\boldsymbol{o}_i + \mathbf{x}_i^\intercal \boldsymbol{B}, \boldsymbol{\Omega}^{-1}), \\
\text{excess of zero} \quad & \boldsymbol{W}_i = (W_{ij})_{j=1\ldots p} && \sim \otimes_{j=1}^{p} \mathcal{B}(\pi_{ij}), \\
\text{observation space} \quad & Y_{ij} \,|\, W_{ij}, Z_{ij} && \sim^{\text{indep}} W_{ij}\delta_0 + (1 - W_{ij})\mathcal{P}\left(\exp\{o_{ij} + Z_{ij}\}\right),
\end{aligned} \tag{2}$$

where δ_0 is the Dirac distribution and we note $\pi = (\pi_{ij})_{i=1..n, j=1..p}$. We focus on the probability π_{ij} of belonging to the pure zero component. Our model is flexible enough to accomodate different parametrizations for π_{ij} based on the availability of covariates and/or modeling choices made by the user:

$$\pi_{ij} = \pi \in [0, 1] \qquad \text{(non-dependent – ND)} \tag{3a}$$

$$\pi_{ij} = \text{logit}^{-1}(\boldsymbol{X}^0 \boldsymbol{B}^0)_{ij}, \;\; \boldsymbol{X}^0 \in \mathbb{R}^{n \times d_0}, \;\; \boldsymbol{B}^0 \in \mathbb{R}^{d_0 \times p} \qquad \text{(column-wise dependence – CD)} \tag{3b}$$

$$\pi_{ij} = \text{logit}^{-1}(\bar{\boldsymbol{B}}^0 \overline{\boldsymbol{X}}^0)_{ij}, \;\; \bar{\boldsymbol{B}}^0 \in \mathbb{R}^{n \times d_0}, \;\; \overline{\boldsymbol{X}}^0 \in \mathbb{R}^{d_0 \times p} \qquad \text{(row-wise dependence – RD)} \tag{3c}$$

where $\text{logit}^{-1}(\cdot)$ is the logistic (or inverse logit) function, $d_0 \geqslant 1$, $\boldsymbol{B}^0$ (resp. $\bar{\boldsymbol{B}}^0$) are regression coefficients associated with row-wise matrix of covariates $\boldsymbol{X}^0$ (resp. column-wise covariates $\bar{\boldsymbol{X}}^0$), obtained by stacking the vectors $((\mathbf{x}_1^0)^\intercal, \ldots, (\mathbf{x}_n^0)^\intercal)$, which may or may not be the same as in $\boldsymbol{X}$, the matrix of covariates in the PLN component.

Using standard results on Poisson and Gaussian distribution, we easily derive the expectation and variance of the ZI-PLN regression model. Letting $A_{ij} \triangleq \exp\left(o_{ij} + \mu_{ij} + \sigma_{jj}/2\right)$ with $mu_{ij} = \mathbf{x}_i^\intercal \boldsymbol{B}_j$, then

$$\begin{aligned}
\mathbb{E}\left(Y_{ij}\right) &= (1 - \pi_{ij}) A_{ij} > 0, \\
\mathbb{V}\left(Y_{ij}\right) &= (1 - \pi_{ij}) A_{ij} + (1 - \pi_{ij}) A_{ij}^2 \left(e^{\sigma_{jj}} - (1 - \pi_{ij})\right).
\end{aligned}$$

In the following, we are interested in inferring the vector of parameters $\theta = (\boldsymbol{\Omega}, \boldsymbol{B}, \pi) \in \mathbb{S}_p^{++} \times \mathcal{M}_{p,d}(\mathbb{R}) \times \mathcal{M}_{n,p}([0, 1])$ where $\mathbb{S}_p^{++}$ is the set of $p \times p$ positive-definite matrices and $\mathcal{M}_{p,d}(\mathbb{R})$ the set of $p \times d$ real-valued matrices. We first show that model 2 is identifiable.

Identifiability of ZI-PLN models Identifiability results are available for the ZI Poisson model [Li, 2012] and can be generalized to the ZI-PLN regression model. To this end, we first consider the simple ZI-PLN model, (*i.e.* a ZI-PLN model without covariate and a single parameter π), with a single sample, in order to drop the index i:

$$\begin{aligned}
\mathbf{W} = (W_j)_{j=1\ldots p} &\sim \mathcal{B}^{\otimes}(\pi) = (\pi_1) \otimes \cdots \otimes (\pi_p) \\
\mathbf{Z} = (Z_j)_{j=1\ldots p} &\sim \mathcal{N}_p(\mu, \boldsymbol{\Omega}^{-1}) \\
Y_j | W_j, Z_j &\sim W_j \delta_0 + (1 - W_j)\mathcal{P}(e^{Z_j}), \quad Y_j \perp Y_k | \mathbf{W}, \mathbf{Z}
\end{aligned} \tag{4}$$

Proposition 1. *The simple ZI-PLN model defined in 4 with parameter $\theta = (\pi, \mu, \boldsymbol{\Omega})$ and parameter space $(0, 1)^p \times \mathbb{R}^p \times \mathbb{S}_p^{++}$ is identifiable.*

Proposition 2. *The ZI-PLN regression model 2 with zero-inflation defined as in eq. (3b) and parameter $\theta = (\boldsymbol{\Omega}, \boldsymbol{B}, \boldsymbol{B}^0)$ and parameter space $\mathbb{S}_p^{++} \times \mathcal{M}_{p,d}(\mathbb{R}) \times \mathcal{M}_{p,d}(\mathbb{R})$ is identifiable if and only if the $n \times d$ matrix of covariates $\mathbf{X}^0$ is full rank.*

3 Estimation by Variational Inference

Our goal is to maximize the marginal likelihood. In the framework of latent models, a standard approach (e.g. with *Expectation-Maximization* algorithms) uses the following decomposition by integrating over the latent variables $\mathbf{W}, \mathbf{Z}$

$$\log p_\theta(\mathbf{Y}) = \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} = \int_{\mathbf{W},\mathbf{Z}} \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y}) d\mathbf{W} d\mathbf{Z}. \tag{5}$$

However, for the ZI-PLN model, it is untractable since the conditional distribution $p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})$ has no closed-form. To overcome this issue, we rely on a variational approximation of this distribution which will yield a lower bound of $\log p_\theta(\cdot)$ to be optimized: for observation i, we denote by $\widetilde{p}_\psi(\mathbf{Z}_i, \mathbf{W}_i)$ the approximation of $p_\theta(\mathbf{Z}_i, \mathbf{W}_i|\mathbf{Y}_i)$ where ψ is a set of variational parameters to be optimized. Subtracting to the untractable expression eq. (5) of the log-likelihood the positive (and also untractable) quantity (known as the Kullback-Lebler divergence)

$$KL(\widetilde{p}_\psi(.) \| p_\theta(.|\mathbf{Y})) = \int_{\mathbf{W},\mathbf{Z}} \log \frac{\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})}{p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})} \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W}) d\mathbf{W} d\mathbf{Z}$$

results after some rearrangements in the following Evidence Lower BOund (ELBO):

$$\begin{aligned} J(\theta, \psi) &= \log p_\theta(\mathbf{Y}) - KL(\widetilde{p}_\psi(.) \| p_\theta(.|\mathbf{Y})) \\ &= \int_{\mathbf{W},\mathbf{Z}} \log \frac{p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})}{\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})} \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W}) d\mathbf{W} d\mathbf{Z} \\ &= \widetilde{\mathbb{E}}[\log p_\theta(\mathbf{Z}, \mathbf{W}, \mathbf{Y})] - \widetilde{\mathbb{E}}[\log \widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})], \end{aligned} \tag{6}$$

which also looks like a plugin of integral eq. (5) with $p_\theta(\mathbf{Z}, \mathbf{W}|\mathbf{Y})$ replaced with $\widetilde{p}_\psi(\mathbf{Z}, \mathbf{W})$. An appropriate choice of variational approximation will make the integral calculation tractable, while leading to an acceptable approximation of the log-likelihood [Blei et al., 2017].

3.1 Choice of the variational family

Standard variational approximation A straightforward, yet efficient, approach is to consider the mean field approximation, which breaks all dependencies between the vectors $\mathbf{Z}_i$ and $\mathbf{W}_i$ and their respective coordinates and approximates the conditional distribution as the product of its coordinate-wise marginals:

$$\widetilde{p}^{(1)}_{\psi_i}(\mathbf{Z}_i, \mathbf{W}_i) \triangleq \widetilde{p}_{\psi_i}(\mathbf{Z}_i)\widetilde{p}_{\psi_i}(\mathbf{W}_i) = \otimes_{j=1}^p \widetilde{p}_{\psi_i}(\mathbf{Z}_{ij})\widetilde{p}_{\psi_i}(\mathbf{W}_{ij}).$$

On top of that, we assume Gaussian and Bernoulli distribution for $\widetilde{p}_{\psi_i}(\mathbf{Z}_{ij})$ and $\widetilde{p}_{\psi_i}(\mathbf{W}_{ij})$ respectively, giving rise to the following variational approximation

$$\widetilde{p}^{(1)}_{\psi_i}(\boldsymbol{Z}_i, \boldsymbol{W}_i) = \otimes_{j=1}^p \mathcal{N}\left(M_{ij}, S_{ij}^2\right) \mathcal{B}\left(P_{ij}\right) \tag{7}$$

with $0 \leqslant P_{ij} \leqslant 1$ and $\psi_i = (M_{ij}, S_{ij}, P_{ij})_{1 \leqslant j \leqslant p}$. We denote $\boldsymbol{M}, \boldsymbol{S}$ and $\mathbf{P}$ the $n \times p$ matrices with respective entries M_{ij}, S_{ij} and P_{ij} $(1 \leqslant i \leqslant n, 1 \leqslant j \leqslant p)$. This approximation therefore requires the estimation of $3np$ additional variational parameters on top of θ.

Enhanced variational approximation. As W_{ij} can take only two values, the dependence between $\mathbf{Z}_{ij}$ and $\mathbf{W}_{ij}$ can easily be highlighted by noting that

$$Z_{ij}|W_{ij}, Y_{ij} = (Z_{ij}|Y_{ij}, W_{ij} = 1)^{W_{ij}} \, (Z_{ij}|Y_{ij}, W_{ij} = 0)^{1-W_{ij}} \, . \tag{8}$$

The conditional distribution of $Z_{ij}|Y_{ij}, W_{ij} = 1$ simplifies to $Z_{ij}|W_{ij} = 1$ and is thus known as Z_{ij} and W_{ij} are independent: it follows a gaussian distribution with mean $\boldsymbol{X}_i^\top B_j$ and variance Σ_{jj}. By contrast, $Z_{ij}|Y_{ij}, W_{ij} = 0$ is untractable and approximated by a gaussian distribution, giving rise to an alternative and slightly more involved variational approximation:

$$\widetilde{p}^{(2)}_{\psi_i}(\boldsymbol{Z}_i, \boldsymbol{W}_i) = \otimes_{j=1}^p \mathcal{N}(\boldsymbol{X}_i^\top B_j, \Sigma_{jj})^{W_{ij}} \mathcal{N}(M_{ij}, S_{ij}^2)^{1-W_{ij}} W_{ij}, \quad W_{ij} \sim^{\text{indep}} \mathcal{B}\left(P_{ij}\right). \tag{9}$$

Expected lower bounds We set $\psi = (\psi_i)_{1 \leqslant i \leqslant n}$ and variational distribution $\widetilde{p}^{(1)}_\psi = \prod_{i=1}^n \widetilde{p}^{(1)}_{\psi_i}$ (resp. $\widetilde{p}^{(2)}_\psi = \prod_{i=1}^n \widetilde{p}^{(2)}_{\psi_i}$) defined in eq. (7) (resp. eq. (9)), its expectation $\widetilde{\mathbb{E}}^{(1)}$ (resp. $\widetilde{\mathbb{E}}^{(2)}$) and its ELBO $J^{(1)}(\psi, \theta)$ (resp. $J^{(2)}(\psi, \theta)$) detailed in the next proposition.

Proposition 3. *The ELBO defined in eq. (6) with variational approximation $\widetilde{p}^{(1)}_\psi$ can be written in matrix form as*

$$\begin{aligned} J^{(1)}(\psi, \theta) =& \widetilde{\mathbb{E}}^{(1)}\left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] + \widetilde{\mathbb{E}}^{(1)}\left[\log p_\theta(\mathbf{W})\right] + H(\mathbf{P}) + \frac{np}{2} + \frac{1}{2} \operatorname{Tr}\left(\mathbf{1}_{n,p}^\top \log(\mathbf{S^2})\right) \\ &+ \frac{n}{2} \log|\boldsymbol{\Omega}| - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}\left(\operatorname{Diag}(\bar{\mathbf{S}}^2) + g\left(\mathbf{M} - \mathbf{XB}\right)\right)\right) \end{aligned}$$

and with variational approximation $\widetilde{p}^{(2)}_\psi$ we get

$$\begin{aligned} J^{(2)}(\psi, \theta) =& \widetilde{\mathbb{E}}^{(2)}\left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] + \widetilde{\mathbb{E}}^{(2)}\left[\log p_\theta(\mathbf{W})\right] + H(\mathbf{P}) + \frac{np}{2} + \frac{1}{2} \operatorname{Tr}\left(\mathbf{Q}^\top \log(\mathbf{S^2})\right) \\ &+ \frac{n}{2} \log|\boldsymbol{\Omega}| - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}\left(\operatorname{Diag}\left(\mathbf{1}_n^\top\left(\mathbf{Q} \odot \mathbf{S}^2\right)\right) + g\left(\mathbf{Q} \odot (\mathbf{M} - \mathbf{XB})\right)\right)\right) \\ &- \frac{1}{2} \operatorname{Tr}\left(\operatorname{diag}(\boldsymbol{\Omega}) 1_n^\top \left((1_n \operatorname{diag}(\boldsymbol{\Sigma})^\top) \odot \mathbf{P} + \mathbf{P} \odot \mathbf{Q} \odot (\mathbf{M} - \mathbf{XB})^2\right)\right) \\ &- \frac{1}{2} 1_n^\top \mathbf{P} \log(\operatorname{diag}(\boldsymbol{\Sigma})), \end{aligned}$$

where $\odot$ denotes the Hadamard product, diag *returns a vector constituted of the diagonal of the input squared matrix, $\mathbf{1}_n$ is a column-vector of size n filled with 1s, $\mathbf{1}_{n,p}$ is a matrix of size $n \times p$ filled with 1s,* Diag *takes a vector x and returns a diagonal matrix with diagonal x, logarithm and squared functions are applied component-wise, $\mathbf{Q} = \mathbf{1}_{n,p} - \mathbf{P}$ and $g(\mathbf{D}) = \mathbf{D}^\top \mathbf{D}$ for $\mathbf{D} \in \mathbb{R}^{n \times p}$. We denoted $\bar{\mathbf{S}}^2 = \mathbf{1}_n^\top \mathbf{S}^2$ and $\delta_{0,\infty}(x) = \begin{cases} 0 & \text{if } x = 0 \\ -\infty & \text{else} \end{cases}$ with the convention that $0 \times \delta_{0,\infty}(x) = 0$ for all x and $0 \times \log(0) = 0$. Note that both ELBOs share the following terms ($\widetilde{\mathbb{E}}^{(1)}$ and $\widetilde{\mathbb{E}}^{(2)}$ coincides for the following terms so that we drop the index):*

$$\begin{aligned} \tilde{\mathbb{E}}\left[\log p_\theta(\mathbf{Y}|\mathbf{Z}, \mathbf{W})\right] &= \operatorname{Tr}\left(\mathbf{Q}^\top\left(\mathbf{Y} \odot (\mathbf{O} + \mathbf{M}) - \mathbf{A} - \log(\mathbf{Y}!)\right) + \mathbf{P}^\top \delta_{0,\infty}\left(\mathbf{Y}\right)\right), \\ \tilde{\mathbb{E}}\left[\log p_\theta(\mathbf{W})\right] &= \operatorname{Tr}\left(\mathbf{P}^\top \mu_0 - \mathbf{1}_{n,p}^\top \log\left(\mathbf{1}_{n,p} + e^{\mu_0}\right)\right), \\ H(\mathbf{P}) &= -\operatorname{Tr}\left(\mathbf{P}^\top \log(\mathbf{P}) + \mathbf{Q}^\top \log\left(\mathbf{Q}\right)\right), \end{aligned}$$

where factorial and exponential are applied component-wise and the matrix $\mathbf{A}$ *denotes* $\exp(\mathbf{O} + \mathbf{M} + \mathbf{S}^2/2)$ *where* exp *is applied component-wise and* $\mu_0 = \mathbf{1}_{n,p} \times \text{logit}(\pi)$ *in the ND case,* $\mu_0 = \boldsymbol{X}^0\boldsymbol{B}^0$ *in the CD case and* $\mu_0 = \bar{\boldsymbol{B}}^0\bar{\boldsymbol{X}}^0$ *in the RD case.*

4 Optimization

Estimating θ is equivalent to solving the optimization problem

$$\arg\max_{\psi,\theta} J(\psi, \theta). \tag{10}$$

where J can be either $J^{(1)}$ (standard approximation) or $J^{(2)}$ (enhanced approximation).

4.1 Optimization of $J^{(1)}$

Past experience for standard PLN models [Chiquet et al., 2017, 2019, 2021] (and analytical properties of $J^{(1)}$ derived in this section) suggests solving the above problem using alternated gradient descent.

Proposition 4. *[Updates of* $\mathbf{B}, \boldsymbol{\Omega}, \mathbf{P}$ *and* $\mathbf{B}^0$*] For fixed* ψ*, the values of* $\boldsymbol{\Omega}, \mathbf{B}$ *maximizing* $J^{(1)}$ *are*

$$\widehat{\boldsymbol{\Omega}} = n\left[g(\boldsymbol{M} - \mathbf{XB}) + \bar{\boldsymbol{S}}^2\right]^{-1}, \qquad \widehat{\mathbf{B}} = [\mathbf{X}^\intercal\mathbf{X}]^{-1}\mathbf{X}^\intercal\mathbf{M}.$$

where $g(\mathbf{D}) = \mathbf{D}^\intercal\mathbf{D}$ *as in Proposition 3. Furthermore, if* $\mathbf{X}^0 = \mathbf{1}_n$ *,* $J^{(1)}$ *is maximized at* $\widehat{\mathbf{B}}^0 = \frac{1}{n}\mathbf{1}_n^\top\mathbf{P}$*. Likewise, when* θ *is fixed,* $J^{(1)}$ *is concave with respect to* $\mathbf{P}$ *and maximized at* $\widehat{\mathbf{P}} = \text{logit}^{-1}\left(\mathbf{A} + \mathbf{X}^0\mathbf{B}^0\right) \times \delta_0(\mathbf{Y})$.

Convergence of Algorithm 1 to a stationnary point of $J^{(1)}$ is a direct consequence of the following lemma.

Lemma 1 (Convergence properties)**.** $J^{(1)}$ *is (separately) concave in* θ *and* ψ*.*

4.2 Optimization of $J^{(2)}$

While optimization of $J^{(1)}$ is easily manageable using closed forms and benefits from a bi-concavity property, optimization of $J^{(2)}$ is more challenging. Indeed, the concavity in $\boldsymbol{\Omega}$ is lost and no closed form can be used for any parameter update.

We do not maximize the ELBO with respect to each parameter in an alternate coordinate-wise fashion but instead compute the gradient with respect to (ψ, θ) as if it were a single parameter. Formally, given $\psi^{(0)}, \theta^{(0)}$ and a learning rate $\eta > 0$, we perform the update step until an arbitrary criterion is reached or a number of iterations is done.

$$(\psi^{(s+1)}, \theta^{(s+1)}) = (\psi^{(s)}, \theta^{(s)}) + \eta\nabla_{\psi,\theta}J^{(2)}(\psi^{(s)}, \theta^{(s)}) \tag{11}$$

Algorithm 1: VEM

Input : $\theta^{(0)}, \psi^{(0)}$ initial point, $T \geqslant 1$ number of iterations.

for $s = 0, \dots T-1$ **do**

M-step

$$\boldsymbol{\Omega}^{(s+1)} = n \left[g\left(\boldsymbol{M}^{(s)} - \mathbf{X}\mathbf{B}^{(s)} \right) + \bar{\boldsymbol{S}}^{2(s)} \right]^{-1}$$

$$\mathbf{B}^{(s+1)} = [\mathbf{X}^\intercal \mathbf{X}]^{-1} \mathbf{X}^\intercal \mathbf{M}^{(s)}$$

$$\mathbf{B}^{0,(s+1)} = \arg\max_{\mathbf{B}^0} \operatorname{Tr}\left[\left(\mathbf{P}^{(s)} \right)^\intercal \mathbf{X}^0 \mathbf{B}^0 \right] - \operatorname{Tr}\left[\mathbf{1}_{n,p}^\intercal \log\left(1 + e^{\mathbf{X}^0 \mathbf{B}^0} \right) \right]$$

VE-step

$$\mathbf{P}^{(s+1)} = \operatorname{logit}^{-1}\left(\mathbf{A}^{(s)} + \mathbf{X}^0 \mathbf{B}^{0,(s+1)} \right) \times \delta_0(\mathbf{Y}), \quad \mathbf{Q}^{(s+1)} = \mathbf{1}_{n,p} - \mathbf{P}^{(s+1)}$$

$$\mathbf{M}^{(s+1)} = \arg\max_{\mathbf{M}} \left(\operatorname{Tr}\left((\mathbf{Y} \odot \mathbf{M} - \mathbf{A})^\top \mathbf{Q}^{(s+1)} \right) - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}^{(s+1)} g\left(\mathbf{M} - \mathbf{X}\mathbf{B}^{(s+1)} \right) \right) \right)$$

$$\mathbf{S}^{(s+1)} = \arg\max_{\mathbf{S}} \left(-\operatorname{Tr}\left(\mathbf{A}^\top \mathbf{Q}^{(s+1)} \right) - \frac{1}{2} \operatorname{Tr}\left(\mathbf{1}_{n,p}^\intercal \log\left(\mathbf{S}^2 \right) \right) - \frac{1}{2} \operatorname{Tr}\left(\boldsymbol{\Omega}^{(s+1)} \bar{\mathbf{S}}^2 \right) \right)$$

end

Output : $\theta^{(T)}, \psi^{(T)}$

4.3 Optimization using analytic law of $W_{ij}|Y_{ij}$

The exact conditional law $W_{ij}|Y_{ij}$ can be derived and is detailed in the next proposition.

Proposition 5. *Let $1 \leqslant j \leqslant p$. The conditional law of $W_{ij}|Y_{ij}$ is given by*

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

with $\varphi(\mu, \sigma^2) = \mathbb{E}\left[\exp(-X)\right]$, $X \sim \mathcal{LN}\left(\mu, \sigma^2\right)$.

In Section 3 we made the variational approximation $\widetilde{p}(W_{ij}) \sim \mathcal{B}(P_{ij})$, considered P_{ij} as free and optimized the ELBO with respect to P_{ij}. The above proposition suggests that P_{ij} can instead be derived directly from θ and not considered as a free variational parameter. We consider $\widetilde{J}^{(1)}$ (resp. $\widetilde{J}^{(2)}$) the ELBO $J^{(1)}$ (resp. $J^{(2)}$) with $P_{ij} = \Psi(\theta)_{ij}$ with

$$\Psi(\theta) \triangleq \frac{\pi}{\varphi\left(\mathbf{X}^\top \mathbf{B}, \mathbf{1}_n \operatorname{diag}(\boldsymbol{\Sigma})^\top \right) \odot (1 - \pi) + \pi} \odot \mathbf{1}_{\boldsymbol{Y}=\mathbf{0}},$$

where φ and the division are applied component-wise and $\mathbf{1}_{\boldsymbol{Y}=\mathbf{0}}$ is a $n \times p$ matrix such that $\left(\mathbf{1}_{\boldsymbol{Y}} = 0\right)_{ij} = 0$ if and only if $Y_{ij} = 0$. Formally, we have

$$\widetilde{J}^{(1)}(\mathbf{M}, \mathbf{S}, \boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0) = J^{(1)}\left(\mathbf{M}, \mathbf{S}, \Psi(\boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0), \boldsymbol{\Omega}, \mathbf{B}, \mathbf{B}^0 \right),$$

and the same formula applies to $\widetilde{J}^{(2)}$. Note that both $\widetilde{J}^{(1)}$ and $\widetilde{J}^{(2)}$ have np fewer variational parameters compared to $J^{(1)}$ and $J^{(2)}$ ($2np$ compared to $3np$) since $\mathbf{P}$ is now completely determined by θ. The function φ is intractable but a sharp (derivable) approximation $\tilde{\varphi}$ is available and detailed in the next section. A major drawback of this approach compared to optimizing $J^{(1)}$ is the lack of any closed form update as stationary points of $\tilde{\varphi}$ are intractable. For the optimization, we consider the gradient scheme defined in Equation (11) where ψ is replaced with $\psi_1 = (\mathbf{M}, \mathbf{S})$.

5 Simulation Study

5.1 Experimental details

We evaluate model 3a on simulated data. We set $p = 300$ and $d = 1$. Given $\theta^\star = (\mathbf{\Omega}^\star, \mathbf{B}^\star, \pi^\star)$, we simulate $n = 1000$ independent observations $\mathbf{Y}_i$, optimize each criterion, repeat this procedure 10 times and evaluate the Root Mean Squared Error (RMSE) between $\mathbf{\Omega}^\star$ and $\widehat{\mathbf{\Omega}}$, between $\mathbf{B}^\star$ and $\widehat{\mathbf{B}}$ and between $\pi^\star$ and $\widehat{\pi}$. The resuls are displayed in Figure 1.

Parameters simulation We employ a simulation approach to generate the matrix $\mathbf{\Sigma}^\star = \mathbf{\Omega}^\star$ in a block-wise structure, where the identity matrix is added on top of it to ensure invertibility. The matrices $\mathbf{X}$ and $\mathbf{B}^\star$ are simulated so that each term of $\mathbf{X}\mathbf{B}^\star$ follows a gaussian distribution with unit variance. The mean is chosen along $\{0, 0.5, \ldots, 4\}$ and displayed on the x-axis. The parameter $\pi^\star$ is set to 0.3.

References

Frédéric Escudié, Lucas Auer, Maria Bernard, Mahendra Mariadassou, Laurent Cauquil, Katia Vidal, Sarah Maman, Guillermina Hernandez-Raquet, Sylvie Combes, and Géraldine Pascal. FROGS: Find, Rapidly, OTUs with Galaxy Solution. *Bioinformatics*, 34(8):1287–1294, 12 2017. ISSN 1367-4803.

J. Aitchison and C. H. Ho. The multivariate poisson-log normal distribution. *Biometrika*, 76 (4):643–653, 1989.

Jenni Niku, Francis KC Hui, Sara Taskinen, and David I Warton. gllvm: Fast analysis of multivariate abundance data with generalized linear latent variable models in r. *Methods in Ecology and Evolution*, 10(12):2173–2182, 2019.

Skipper Seabold and Josef Perktold. statsmodels: Econometric and statistical modeling with python. In *9th Python in Science Conference*, 2010.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. 39:1–38, 1977.

T. S. Jaakkola and M. I. Jordan. Bayesian parameter estimation via variational methods. *Statistics and Computing*, 10(1):25–37, 2000.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

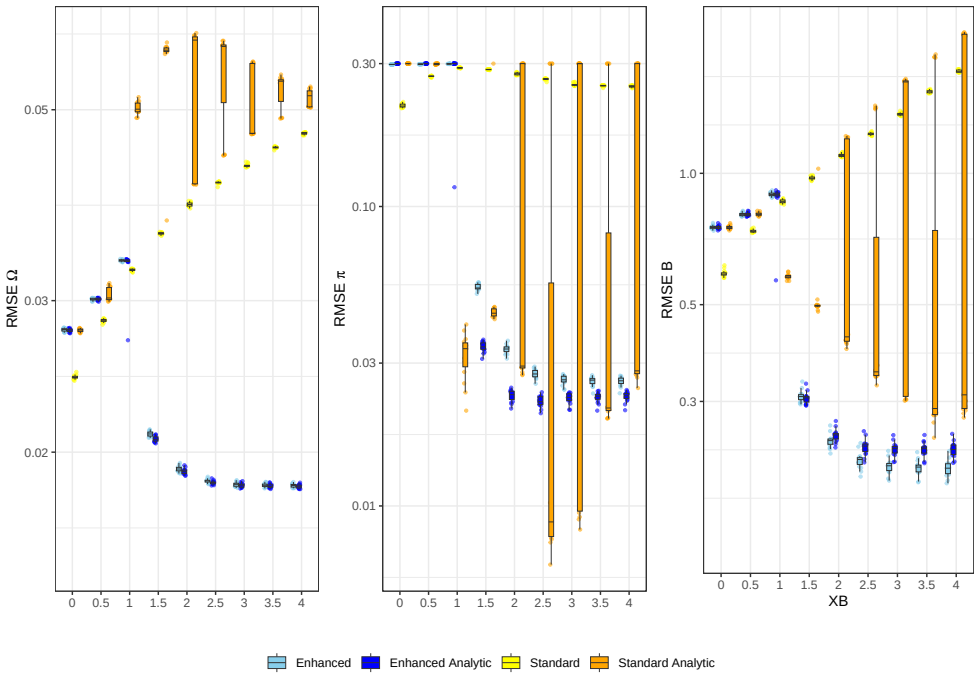

Figure 1: Simulations with $n = 1000, p = 300$ and $d = 1$. Enhanced corresponds to $J^{(2)}$, Enhanced Analytic to $\widetilde{J}^{(2)}$, Standard to $J^{(1)}$ and Standard Analytic to $\widetilde{J}^{(1)}$.

Francis KC Hui, David I Warton, John T Ormerod, Viivi Haapaniemi, and Sara Taskinen. Variational approximations for generalized linear latent variable models. *Journal of Computational and Graphical Statistics*, 26(1):35–43, 2017.

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877, 2017.

Mcmc maximum likelihood for latent state models. *Journal of Econometrics*, 137(2):615–640, 2007. ISSN 0304-4076.

ISSN 03036898, 14679469.

Davide Risso, Fanny Perraudeau, Svetlana Gribkova, Sandrine Dudoit, and Jean-Philippe Vert. A general and flexible method for signal extraction from single-cell rna-seq data. *Nature communications*, 9(1):1–17, 2018.

Chin-Shang Li. Identifiability of zero-inflated Poisson models. *Brazilian Journal of Probability and Statistics*, 26(3):306 – 312, 2012.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. Variational inference for probabilistic poisson pca. *Ann. Appl. Stat.*, March 2017.

Julien Chiquet, Stephane Robin, and Mahendra Mariadassou. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, pages 1162–1171. PMLR, 2019.

Julien Chiquet, Mahendra Mariadassou, and Stéphane Robin. The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. *Frontiers in Ecology and Evolution*, 9:188, 2021.