Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Sparse Poisson Log Normal Model for Zero-Inflated Multivariate Count Data

Togo Jean Yves KIOYE

kioyejeanyves@gmail.com

Université Clermont Auvergne

Paul-Marie GROLLEMUND

Université Clermont Auvergne

Jocelyn CHAUVET

Institut catholique de Vendée

Christophe CHASSARD

Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

Research Article

Keywords: Information criteria, Microbial communities, Multivariate count data, Variational EM algorithm, Variable selection, Zero inflation

Posted Date: October 29th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11381825/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Sparse Poisson Log Normal Model for Zero-Inflated Multivariate Count Data

Togo Jean Yves KIOYE[1*], Paul-Marie GROLLEMUND[1,2], Jocelyn CHAUVET[3,4], Christophe CHASSARD[1]

[1*]Université Clermont Auvergne, INRAE, Unité Mixte de Recherche sur le Fromage (UMRF), Aurillac, 15000, France.
[2]Université Clermont Auvergne, Laboratoire de Mathématiques Blaise Pascal (LMBP), Aubière, 63178, France.
[3]Institut catholique de Vendée (ICES), Centre de recherche de l'ICES, La Roche-sur-Yon, 85017, France.
[3]Université d'Angers, Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), Angers, 49000, France.

*Corresponding author(s). E-mail(s): togo_jean_yves.kioye@ext.uca.fr;
Contributing authors: paul_marie.grollemund@uca.fr; jchauvet@ices.fr;
christophe.chassard@inrae.fr;

Abstract

The Zero-Inflated Poisson Log-Normal model (ZIPLN) is a model for count data with excess zeros. It assumes that the observed zeros come from a mixture of a degenerate distribution at zero and a Poisson Log-Normal model (PLN), with both components influenced by explanatory covariates. For example, in abundance of microbial communities depending on the experimental conditions, some species may be modelled by a multivariate PLN distribution, while other species may be unobserved leading to an excess of zeros. When the observed abundances depend on a large number of covariates, it becomes necessary to implement variable selection procedures to improve interpretability, reduce noise, and simplify the model. This paper extends the Smooth Information Criterion (SIC) to the ZIPLN model. The proposed method consists of penalizing an Evidence Lower BOund (ELBO) of the marginal log-likelihood associated with the ZIPLN model using a smooth approximation of the $\boldsymbol{L_0}$-norm. The resulting algorithm combines a specific algorithm for the SIC, ε-telescoping, and a Variational Expectation Maximisation algorithm (VEM). Our study aims to induce sparsity on the regressors explaining the random occurrences of the ZIPLN model. To support our proposal, we conduct simulation studies and apply it to data from a microbial communities study involved in the milk production process.

Keywords: Information criteria, Microbial communities, Multivariate count data, Variational EM algorithm, Variable selection, Zero inflation

MSC Classification: 62H12 , 62J99

1 Introduction

There are numerous fields that exhibit multivariate count data, such as ecology, microbiology, industry, and medicine. In ecology and microbiology, they are commonly employed to study species distributions (Warton et al. 2015; Lee et al. 2020; Chiquet et al. 2021). Genetic marker amplification and sequencing produce high-dimensional count data that are often characterized by an excess of zeros relative to what is expected under conventional count distributions (Lambert 1992). This phenomenon reflects the true or apparent absence of species, and poses specific challenges for statistical modeling.

To deal with the excess of zeros, a widely adopted approach is to model the observed data as a mixture between a degenerate distribution at zero (Dirac mass at zero) and a counting distribution (Lambert 1992), with the transition from one to the other controlled by a probability w. Two families of models are often used: Zero-Inflated (ZI) models (Johnson and Kotz 1969; Cohen 1963) and zero-altered (ZA) models (Heilbron 1994), also known as hurdle models. ZA models consider that all zeros come from the degenerate distribution, while strictly positive values are derived from a zero-truncated counting model. Conversely, ZI models allow for a double origin for zeros, which can be generated either by the degenerate component or by the counting model.

We focus here on models developed using the ZI strategy, which accounts for two possible sources of zeros: a true zero, meaning that the species is absent from the ecosystem or under a given experimental condition, and a random zero due to sampling, which originates from the counting component of the mixture (e.g., the Poisson component). The first ZI models, introduced by Johnson and Kotz (1969); Cohen (1963), did not take covariates into account. Later, several extensions were proposed to incorporate covariates. Mullahy (1986) introduced a ZI Poisson (ZIP) regression model in which the Poisson intensity parameter, γ, depends on covariates. Lambert (1992) proposed a ZIP model where both γ and w are functions of covariates.

Since then, various ZI models have been developed to address the specificities of certain data. For example, in an industrial context, Li et al. (1999) have proposed a multivariate ZIP model for multiple types of defects. It is important to note that the term multivariate, here, refers to the number of defect types, rather than a joint multivariate response. Some extensions incorporate random effects into ZIP or ZI binomial regression models (Hall 2000), while others address censored data (Saffari and Adnan 2011). Certain recent extensions take into account joint responses and overdispersion. In microbiology, Lee et al. (2020) have proposed a ZI Bayesian regression

model, incorporating complex dependence modelling between microbial species, variable selection and MCMC inference. Batardière et al. (2025) have introduced a ZI version of the multivariate PLN model (ZIPLN).

In this paper, We extend the work initiated in Kioye et al. (2026) by considering the ZIPLN model. To achieve this, we adopt the extension proposed by Batardière et al. (2025), to which we consider an additional variable selection step during the inference procedure. As in the extension of Lambert (1992), the mixing probability in the model in Batardière et al. (2025) depends on covariates, which may be either the same as or different from those involved in the Poisson intensity parameter. In this context, variable selection can serve three distinct purposes. First, to identify the relevant covariates for each count variable (or each species) involved in the construction of the mixing probabilities; second, to select only the covariates that explain the observed abundances; and third, to select covariates that play a role in both the construction of the mixing probabilities and the explanation of abundances. In this paper, we focus on the second objective, namely the efficient selection of relevant covariates involved in the Poisson component of the ZIPLN model.

Variable selection is a key issue in multivariate analysis, and numerous approaches continue to be developed to address it. To promote sparsity in estimation, the ideal approach consists of adding the L_0 pseudo-norm to the loss function. However, optimization with the L_0 norm is computationally intractable due to the non-convex and discontinuous nature of the L_0 pseudo-norm (Natarajan 1995). A widely adopted strategy is to replace the L_0-norm with its convex surrogate, the L_1-norm (Chen and Donoho 1994; Tibshirani 1996). However, the L_1-norm introduces bias in the coefficient estimates (Fan and Li 2001; Zou 2006), making it suboptimal for identifying truly non-zero covariate effects in regression models (Candes et al. 2008; Marmin et al. 2019). To more closely approximate the L_0 norm while maintaining computational feasibility and overcoming the limitations of the L_1 penalty, several methods have been proposed. Among them, we can mention the reweighted L_1-norm (Candes et al. 2008), the L_q-quasinorm minimization (Foucart and Lai 2009), the log-sum penalty (Weston et al. 2003), and the Geman–McClure criterion (Geman 1987). These penalties depend on a parameter $\varepsilon > 0$ that controls the approximation of the L_0-norm and offers a suitable trade-off between reducing the bias induced by regularization and accurately approaching the L_0 norm as ε tends to zero. Significant advances have been achieved in signal processing over the years, aiming to develop sparse and unbiased estimators (Castella and Pesquet 2015; Chouzenoux et al. 2013; Florescu et al. 2014). More recently, the Geman–McClure criterion has been applied to variable selection in

statistics (O'Neill and Burke 2022; O'Neill and Burke 2023). O'Neill and Burke (2023) established a connection between the Geman–McClure criterion, defined in Section 3.1, and an information criterion called the Smooth Information Criterion (SIC) for fixed values of a parameter λ that controls the penalty level in the loss function. Although the SIC avoids the use of cross-validation for tuning λ, it still relies on an iterative procedure known as ε-telescoping to mitigate the instability in parameter estimation that arises when ε remains fixed.

The approach developed in this paper consists in building a sparse and less biased variable selection procedure for the multivariate ZIPLN model, based on variational inference (Jaakkola and Jordan 2000; Blei et al. 2017a) and using SIC, in order to obtain stable and parsimonious estimates of the model parameters. This facilitates interpretation by focusing on the relevant variables for each column of the sparse multivariate count data. This contribution is of practical importance, particularly in biological and ecological applications. Our work differs from ZIPLN-network (Tous et al. 2025), which aims to identify meaningful associations between the columns of the count matrix by applying an L_1-penalty to the latent precision matrix. Instead, we propose to induce sparsity in the coefficients that encode the effect of covariates on each column of the count matrix.

The remainder of the chapter is organized as follows. Section 2 provides a review of the ZIPLN model construction and describes the associated inference method. Section 3 presents the coupling strategy between the SIC algorithm and the Variational Expectation Maximisation algorithm (VEM) for variable selection. Sections 4 and 5 are devoted to empirical studies on simulated data and real data, respectively, the latter coming from a study on the abundance of microbial communities involved in the milk production process.

2 Background

2.1 Adapting PLN with zero inflation

Poisson Log-Normal (PLN) model. The PLN model is introduced by Aitchison and Ho (1989) and can be rewritten as:

$$\begin{aligned} y_{ij} \mid z_{ij} &\sim \mathcal{P}\left(e^{o_{ij}+z_{ij}}\right), && \text{(observed)} \\ \mathbf{z}_i &\sim \mathcal{N}_p\left(\boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\right), && \text{(latent)} \end{aligned} \tag{1}$$

where y_{ij} corresponds to the number of occurrences of species (or feature) j observed in the sample (or site) i. $\boldsymbol{x}_i \in \mathbb{R}^d$ is the vector of covariates associated with the sample i. $\boldsymbol{z}_i \in \mathbb{R}^p$ denotes the latent variable vector for the sample i, with

entries z_{ij}. o_{ij} represents the sampling effort associated with observation i and feature j. We denote by $\boldsymbol{Y} \in \mathbb{N}^{n\times p}$, $\boldsymbol{Z} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{O} \in \mathbb{R}^{n\times p}$ respectively the matrices associated to a collection of n samples of the abundances of p species, p latent variables and p sampling efforts . $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ is the matrix of d covariates associated to the n samples. The matrix $\boldsymbol{B} \in \mathbb{R}^{d\times p}$ encodes the effects of the d covariates on the p species of $\boldsymbol{Y}$. The matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{p\times p}$ captures the dependence structure among the species. $\mathcal{P}\big(e^{o_{ij}+z_{ij}}\big)$ represents the Poisson distribution with parameter $e^{o_{ij}+z_{ij}}$, and $\mathcal{N}_p\big(\boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big)$ denotes the multivariate normal distribution, where $\boldsymbol{x}_i^\top \boldsymbol{B}$ represents the mean vector of p components and Σ represents the variance-covariance matrix.

Accounting for zero inflation. In count data, the number of observed zeros can exceed the expected number predicted by a standard count model. This can arise from two main mechanisms: true absences, which may be linked to experimental conditions, covariates, or random factors, and zeros that are the result of the counting process. This is referred to as ZI. For example, when many entries of $\boldsymbol{Y}$ are zeros, the classical PLN model underestimates the proportion of zero observations. To address this limitation, a common approach is to model the observed counts as a mixture of two components: a Dirac mass at zero (a degenerate distribution at zero), denoted by δ_0, and a count model, describing random occurrences. A direct extension of ZI to the PLN model involves introducing a probability w_{ij} that the observed count y_{ij} takes the value 0, along with a degenerate distribution at 0 in the observation space of (1). The resulting model can be expressed as:

$$
\begin{aligned}
y_{ij} \mid z_{ij}, w_{ij} &\sim w_{ij}\delta_0 + (1-w_{ij})\mathcal{P}\big(e^{(o_{ij}+z_{ij})}\big), \\
\boldsymbol{z}_i &\sim \mathcal{N}_p\big(\boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big).
\end{aligned}
$$

The observation space is equivalent to:

$$
y_{ij} \mid z_{ij}, w_{ij} \sim \begin{cases} 0 & \text{w.p. } w_{ij}, \\ \mathcal{P}(e^{o_{ij}+z_{ij}}) & \text{w.p. } 1-w_{ij}, \end{cases}
$$

where w.p. stands for with probability.

The density of y_{ij} can be written as follows:

$$
p(y_{ij}) = \begin{cases} w_{ij} + (1-w_{ij})e^{-e^{o_{ij}+z_{ij}}} & \text{if } y_{ij} = 0, \\ (1-w_{ij})\mathcal{P}(e^{o_{ij}+z_{ij}}) & \text{if } y_{ij} \in \mathbb{N}^*. \end{cases}
$$

Several strategies exist to define the probability w_{ij}. Both w_{ij} and $e^{o_{ij}+z_{ij}}$ can be related through covariates. Sometimes w_{ij} and $e^{o_{ij}+z_{ij}}$ are unrelated; other times w_{ij} is a function of $e^{o_{ij}+z_{ij}}$, such as $w_{ij} = 1/(1+(e^{o_{ij}+z_{ij}})^\tau)$, where τ is an unknown constant (Lambert 1992). Here, we consider that it may depend on species j at site i through explanatory variables $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^{d_0}$. For example, in

ecological or microbiological contexts, the environmental characteristics specific to a site may influence the over-representation of zeros. Similarly, certain species exhibit specific behaviors or distributions that alter the probability of observing zeros in the data. These explanatory variables can be the same as those used for the Poisson intensity parameter. The probability w_{ij} is generally modeled by a logistic regression defined as follows:

$$w_{ij} = \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j}}{1 + e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j}}, \text{ and } \text{logit}(w_{ij}) = \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j,$$

where $\widetilde{\boldsymbol{x}}_i \in \mathbb{R}^{d_0}$ represents the covariate vector that influences ZI at site i and $\widetilde{\boldsymbol{b}}_j$ denotes the corresponding coefficients for species j.

Zero-Inflated PLN (ZIPLN) regression model. Batardière et al. (2025) modeled the excess of zeros by incorporating ZI into the standard PLN model (2). To achieve this, they modeled w_{ij} using an additional latent layer following a Bernoulli distribution, parameterized by a probability π_{ij} which encodes the probability that a species j at site i belongs to the degenerate zero distribution. The resulting model is as follows:

$$\begin{aligned} y_{ij} \mid z_{ij}, w_{ij} &\sim w_{ij}\delta_0 + (1 - w_{ij})\mathcal{P}\big(o_{ij} + e^{z_{ij}}\big), \\ w_{ij} &\sim \mathcal{B}\big(\pi_{ij}\big), \\ \mathbf{z}_i &\sim \mathcal{N}_p\big(\boldsymbol{x}_i^\top \boldsymbol{B}, \boldsymbol{\Sigma}\big), \end{aligned} \tag{2}$$

where π_{ij} is defined by a logistic function

$$\pi_{ij} = \frac{e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j}}{1 + e^{\widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j}}, \text{ and } \text{logit}(\pi_{ij}) = \widetilde{\boldsymbol{x}}_i^\top \widetilde{\boldsymbol{b}}_j.$$

We define the matrices $\boldsymbol{\Pi}$ and $\boldsymbol{W}$ of dimensions $n \times p$, where the entries correspond to π_{ij} and w_{ij} respectively. Additionally, $\widetilde{\boldsymbol{B}} \in \mathbb{R}^{d_0 \times p}$ is the coefficient matrix for the p species associated with $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times d_0}$. We also define $\boldsymbol{\theta} = \left(\boldsymbol{B}, \boldsymbol{\Sigma}, \widetilde{\boldsymbol{B}}\right)$ as the set of parameters for the ZIPLN model.

2.2 Inference

Inference problem. The inference problem in (2) is to maximize the marginal log-likelihood in the presence of latent variables. However, this maximization requires evaluating high-dimensional integrals over the latent spaces, making the inference computationally intractable. The marginal log-likelihood is written as follows:

$$\ell(\boldsymbol{\theta}, \boldsymbol{Y}) = \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(\boldsymbol{y}_i), \tag{3}$$

where

$$p_{\boldsymbol{\theta}}(\boldsymbol{y}_i) = \int_{\boldsymbol{z}_i, \boldsymbol{w}_i} p_{\boldsymbol{\theta}}(\boldsymbol{y}_i, \boldsymbol{z}_i, \boldsymbol{w}_i) d\boldsymbol{z}_i d\boldsymbol{w}_i.$$

Inference by Variational Expectation Maximisation (VEM). The estimation of the parameters of the ZIPLN model (2) requires the use of the EM algorithm (Dempster et al. 1977).

However, the E-step involves computing the conditional expectation under the posterior distribution $p_{\boldsymbol{\theta}}(\boldsymbol{z}_i, \boldsymbol{w}_i \mid \boldsymbol{y}_i)$, which does not have a closed form. To circumvent this issue, one can use either approximation methods or sampling techniques to deal with the intractable component during the E-step. To this end, we can rely on the variational approximation (Blei et al. 2017a; Jaakkola and Jordan 2000). This approach approximates the true posterior $p_{\boldsymbol{\theta}}(\boldsymbol{z}_i, \boldsymbol{w}_i \mid \boldsymbol{y}_i)$ with a variational distribution $q_{\boldsymbol{\Psi}_i}(\boldsymbol{z}_i, \boldsymbol{w}_i)$, obtained by minimizing the Kullback–Leibler (KL) divergence. $\boldsymbol{\Psi}_i$ denotes the set of variational parameters associated with observation i. Recently, Batardière et al. (2025) extended this approach to the ZIPLN model. In the following, we recall its application in the context of the ZIPLN model (see Batardière et al. (2025) for more details).

In the case of the ZIPLN model, *mean-field* variational inference assumption is employed. This approach assumes that the variational distribution $q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W})$ factorizes as follows:

$$\begin{aligned} q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W}) &= \prod_{i=1}^{n} q_{\boldsymbol{\psi}_i^1}(\boldsymbol{z}_i) q_{\boldsymbol{\psi}_i^2}(\boldsymbol{w}_i) \\ &= \prod_{i=1}^{n} \bigotimes_{j=1}^{p} q_{\boldsymbol{\psi}_i^1}(\boldsymbol{z}_{ij}) q_{\boldsymbol{\psi}_i^2}(\boldsymbol{w}_{ij}). \end{aligned} \tag{4}$$

This factorization assumes that the latent variables $\boldsymbol{z}_i$ and $\boldsymbol{w}_i$ are independent under the variational distribution. The choice of the variational distributions ($q_{\boldsymbol{\psi}_i^1}(\boldsymbol{z}_i)$ and $q_{\boldsymbol{\psi}_i^2}(\boldsymbol{w}_i)$) directly impacts the complexity of the optimization problem (Blei et al. 2017b). Batardière et al. (2025) assumed that $q_{\boldsymbol{\psi}_i^1}(\boldsymbol{z}_i)$ follows a multivariate Gaussian distribution, $\mathcal{N}_p\big(\boldsymbol{m}_i, \operatorname{diag}(\boldsymbol{s}_i^2)\big)$ with $\boldsymbol{\psi}_i^1 = (\boldsymbol{m}_i, \boldsymbol{s}_i)$, while, $q_{\boldsymbol{\psi}_i^2}(\boldsymbol{w}_i)$ follows a Bernoulli distribution, $\mathcal{B}(p_{ij})$ with $\boldsymbol{\psi}_i^2 = \boldsymbol{p}_i$ where $0 \leq p_{ij} \leq 1$. The set of variational parameters is collected into $\boldsymbol{\Psi} = (\boldsymbol{\psi}^1, \boldsymbol{\psi}^2)$, where $\boldsymbol{\psi}^1 = (\boldsymbol{M}, \boldsymbol{S})$ and $\boldsymbol{\psi}^2 = \boldsymbol{P}$. The matrices $\boldsymbol{M}$, $\boldsymbol{S}$, and $\boldsymbol{P}$ are $n \times p$ matrices whose rows correspond respectively to the p-dimensional vectors $\boldsymbol{m}_i$, $\boldsymbol{s}_i$, and $\boldsymbol{p}_i$.

The marginal log-likelihood (3) can be rewritten as:

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{Y}) =& J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) + \\ & \mathrm{KL}\Big[q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W}) \,||\, P_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{W} \mid \boldsymbol{Y})\Big]. \end{aligned} \tag{5}$$

The first term $J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi})$ is the ELBO of the marginal log-likelihood, and the second term is called the KL divergence between the variational distribution $q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W})$ and the posterior $p_{\boldsymbol{\theta}}(\boldsymbol{Z}, \boldsymbol{W} \mid \boldsymbol{Y})$, this last term is always positive. So (5) can be written as:

$$\begin{aligned} \ell(\boldsymbol{\theta}, \boldsymbol{Y}) \geq J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y}, \boldsymbol{Z}, \boldsymbol{W})] \\ & - \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W})}[\log q_{\boldsymbol{\Psi}}(\boldsymbol{Z}, \boldsymbol{W})]. \end{aligned} \tag{6}$$

Proposition 2.1. *The matrix formulation of the ELBO (6) for the ZIPLN model can be written as:*

$$
\begin{aligned}
J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) = & \mathbf{1}_n^\top [\boldsymbol{Q} \odot (\boldsymbol{Y} \odot (\boldsymbol{O} + \boldsymbol{M}) - \boldsymbol{A} \\
& - \log(\boldsymbol{Y}!)) + \boldsymbol{P} \odot \boldsymbol{\delta}_0(\boldsymbol{Y})] \mathbf{1}_p \\
& \frac{-1}{2} \big[- n \log(|\boldsymbol{\Omega}|) + \\
& \mathbf{1}_n^\top \left((\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B}) \boldsymbol{\Omega} (\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B})^\top \right) \mathbf{1}_n \\
& + \mathit{trace}(\boldsymbol{\Omega} \tilde{\boldsymbol{S}}^2) \big] + \\
& \mathbf{1}_n^\top (\boldsymbol{P} \odot \widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{B}} - \log(\mathbf{1}_{n,p} + e^{\widetilde{\boldsymbol{X}} \widetilde{\boldsymbol{B}}})) \mathbf{1}_p \\
& + \frac{np}{2} + \frac{1}{2} \mathbf{1}_n^\top \log(\boldsymbol{S}^2) \mathbf{1}_p + \\
& \mathbf{1}_n^\top (-\boldsymbol{P} \odot \log(\boldsymbol{P}) - \boldsymbol{Q} \odot \log(\boldsymbol{Q})) \mathbf{1}_p,
\end{aligned} \tag{7}
$$

where $\odot$ denotes the Hadamard product, $\boldsymbol{S}^2$ is given by $\boldsymbol{S}^2 = \boldsymbol{S} \odot \boldsymbol{S}$. We define $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$, $|\boldsymbol{\Omega}|$ represents the determinant of $\boldsymbol{\Omega}$. $\tilde{\boldsymbol{S}}^2 = \sum_{i=1}^n \mathit{diag}(\boldsymbol{s}_i^2)$ represents the accumulated matrix formed by summing the n matrices $\mathit{diag}(\boldsymbol{s}_i^2)$. The matrix $\boldsymbol{A}$ is constructed element-wise as $\boldsymbol{A} = e^{\boldsymbol{O} + \boldsymbol{M} + \boldsymbol{S}^2/2}$, and finally, we set $\boldsymbol{Q} = \mathbf{1}_{n,p} - \boldsymbol{P}$.

The proof of Proposition 2.1 is postponed to the appendix A.

3 Multivariate ZIPLN regression and variable selection

Variable selection plays an important role in multivariate regression models when the number of explanatory variables is high. It can improve prediction accuracy, enhance model interpretability, and help identify the variables that have a significant influence on the response variables. Variable selection methods are well established for many univariate regression models and some multivariate regression models. However, only a limited number of studies have investigated their application to the multivariate ZIPLN model.

Variable selection in ZIPLN. To the best of our knowledge, the recent work addressing variable selection in the ZIPLN model is that of Tous et al. (2025), which aims to identify meaningful associations among the columns of the count matrix by imposing an L_1-penalty on the latent precision matrix $\boldsymbol{\Omega}$. In contrast, our objective is to induce sparsity in the regression coefficient matrix $\boldsymbol{B}$ in order to identify the active variables in the count component of the ZIPLN model. Instead using an L_1-penalty, we employ a smooth approximation of the L_0-penalty to promote sparsity in the estimation of the regression coefficients (see section 3.1 for more details). The use of a smooth approximation of the L_0-penalty is motivated by the bias introduced by the uniform shrinkage of all parameters, regardless of their true relevance, induced by the L_1-penalty (Candes et al. 2008; Fan and Li 2001; Zou 2006), rendering the L_1-penalty suboptimal for accurately recovering the true support of the nonzero components.

Limits of the traditional methods. The dependent nature of the columns in multivariate count data makes the application of traditional two-step variable selection methods (namely, a selection phase followed by an estimation phase) difficult. This limitation arises from the fact that a variable may play a determining role in modeling the observed abundances within a specific column while being uninformative for other columns. Moreover, these variables can shape the inter-column dependency relationships. In this context, managing variables independently across columns through selective removal or addition can be unrealistic, as it ignores the underlying interactions.

Unified one-step procedure. To overcome these limitations, we propose a unified one-step procedure that combines variable selection and less biased parameter estimation. This approach pursues a dual objective: sparsity and less biased estimators compared to other regularization methods. To the best of our knowledge, the most recent work addressing these dual objectives for multivariate count data is proposed in Kioye et al. (2026). In the following section, we propose to extend these strategies to multivariate sparse count data.

3.1 Smooth Information Criterion and ZIPLN (SICZIPLN)

Smooth information criterion (SIC). The SIC is introduced by O'Neill and Burke (2023) which is formulated as follows:

$$\text{SIC} = \ell(\boldsymbol{\theta}) - \frac{\lambda}{2}\Big[\|\widetilde{\boldsymbol{\theta}}\|_{0,\varepsilon} + k\Big], \tag{8}$$

where $\ell(\boldsymbol{\theta})$ denotes the log-likelihood, $\boldsymbol{\theta}$ represents the model parameters, $\tilde{\boldsymbol{\theta}}$ corresponds to the parameters subject to regularization, and k is the number of unregularized parameters. The penalty coefficient λ is set to $\lambda = \log(n)$ for the Bayesian Information Criterion (BIC) and $\lambda = 2$ for the Akaike Information Criterion (AIC), respectively. $\|.\|_{0,\varepsilon}$ is a penalty function determined through a function $\phi_\varepsilon(x)$ defined as follows:

$$\phi_\varepsilon(x) = \frac{x^2}{x^2 + \varepsilon^2}.$$

Note that the function ϕ_ε is differentiable for $\varepsilon > 0$, and we have: $\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$. The parameter ε can be interpreted as a tuning parameter controlling the approximation of the L_0 norm. This penalty provides a good trade-off between reducing the bias induced by regularization and obtaining sparse estimates as $\varepsilon \longrightarrow 0$, compared to the L_1 penalty (Chen and Donoho 1994; Tibshirani 1996) and the SCAD penalty (Fan and Li 2001). However, relying on a penalty function

based on ϕ_ε requires the calibration of the parameter ε, which is not straightforward. Small values of ε allow a closer approximation of the L_0-norm but lead to unstable parameter estimates. In contrast, large values of ε yield more stable parameter estimates but fail to approximate the L_0-norm. To address stability issues related to parameter variations as a function of ε (see Kioye et al. (2026) for the geometrical interpretation), O'Neill and Burke (2023) proposed an iterative procedure called ε-telescoping, which updates the parameters successively over a grid of ε ranging from $\varepsilon = 100$ to $\varepsilon = 10^{-5}$. Although this approach avoids the use of cross-validation for calibrating λ, it still requires an iterative procedure to address the stability issue. The ε-telescoping strategy consists of defining an exponentially decreasing sequence of ε values approaching zero. For each value in the sequence, The SIC problem (8) is optimized, using the solution from the previous step as the initialization for the next, thereby improving convergence and stability.

Penalized ELBO of ZIPLN. Depending on the objectives, we may penalize all the parameters of the ZIPLN model or focus on a specific parameter of interest. For example, Tous et al. (2025) introduced an L_1 penalty on the latent precision matrix $\mathbf{\Sigma}^{-1}$ to identify significant associations between the columns of the count matrix. Here, our goal is to induce sparsity in the regressor matrix $\boldsymbol{B}$ in order to identify the active variables for each column of the count matrix and to obtain less biased estimates. To achieve this, we adapt the SIC criterion to the ZIPLN model, in which the log-likelihood is replaced by its corresponding ELBO. The resulting penalized ELBO is formulated as follows:

$$J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) = J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) - \frac{\lambda}{2}\|\boldsymbol{B}\|_{0,\varepsilon},$$

where $\|\boldsymbol{B}\|_{0,\varepsilon} = \sum_{k=1}^{d}\sum_{j=1}^{p} \frac{b_{kj}^2}{b_{kj}^2+\varepsilon^2}$.

The detailed expression of the penalized ELBO for the ZIPLN is given in Proposition 3.1.

Proposition 3.1. *The penalized ELBO for the ZIPLN model can be expressed in matrix form as follows:*

$$\begin{aligned}
J_{pen}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \mathbf{1}_n^\top [\boldsymbol{Q} \odot (\boldsymbol{Y} \odot (\boldsymbol{O} + \boldsymbol{M}) - \boldsymbol{A} \\
& - \log(\boldsymbol{Y}!)) + \boldsymbol{P} \odot \boldsymbol{\delta}_0(\boldsymbol{Y})]\mathbf{1}_p \\
& \frac{-1}{2}\big[- n\log(|\boldsymbol{\Omega}|) + \\
& \mathbf{1}_n^\top \left((\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Omega}(\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B})^\top\right)\mathbf{1}_n \\
& + trace(\boldsymbol{\Omega}\tilde{\boldsymbol{S}}^2)\big] + \\
& \mathbf{1}_n^\top (\boldsymbol{P} \odot \widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}} - \log(\mathbf{1}_{n,p} + e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}}}))\mathbf{1}_p \\
& + \frac{np}{2} + \frac{1}{2}\mathbf{1}_n^\top \log(\boldsymbol{S}^2)\mathbf{1}_p + \\
& \mathbf{1}_n^\top (-\boldsymbol{P} \odot \log(\boldsymbol{P}) - \boldsymbol{Q} \odot \log(\boldsymbol{Q}))\mathbf{1}_p \\
& - \frac{\lambda}{2}\|\boldsymbol{B}\|_{0,\varepsilon}. \qquad (9)
\end{aligned}$$

3.2 Optimization strategies

Variational Expectation-Maximization (VEM). To optimize (9), we employ the VEM algorithm, which can be interpreted as a double maximization procedure. In the E-step, we optimize the variational parameters $\boldsymbol{\Psi}$ while keeping the model parameters $\boldsymbol{\theta}$ fixed. In the M-step, we optimize the model parameters $\boldsymbol{\theta}$ given the current estimates of the variational parameters $\boldsymbol{\Psi}$. Both steps rely on a gradient-based optimization approach, implemented via the R package `nloptr` (Ypma 2014) for each value of ε during the ε-telescoping. This method requires the precise computation of the gradient of the penalized ELBO with respect to each parameter. The detailed gradient expressions for each parameter are the same as those provided in Batardière et al. (2025), except for the gradient with respect to the penalized parameter $\boldsymbol{B}$. The gradient with respect to each parameter can be written as follows:

$$
\begin{aligned}
\nabla_{\boldsymbol{B}} J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{B}\boldsymbol{\Omega} - \boldsymbol{X}^\top \boldsymbol{M}\boldsymbol{\Omega} \\
& - \frac{\lambda}{2}\phi'_\varepsilon(\boldsymbol{B}), \\
\nabla_{\widetilde{\boldsymbol{B}}} J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \widetilde{\boldsymbol{X}}^\top \boldsymbol{P} - \widetilde{\boldsymbol{X}}^\top \left(\frac{e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}}}}{1 + e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}}}} \right), \\
\nabla_{\boldsymbol{\Omega}} J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \frac{n}{2}\boldsymbol{\Omega}^{-1} + \tilde{\boldsymbol{S}}^2 \\
& - \frac{1}{2}\big((\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B})^\top (\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B}), \\
\nabla_{\boldsymbol{M}} J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \boldsymbol{Q} \odot [\boldsymbol{Y} - \boldsymbol{A}] - (\boldsymbol{M} - \boldsymbol{X}\boldsymbol{B})\boldsymbol{\Omega}, \\
\nabla_{\boldsymbol{S}} J_{\text{pen}}(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \frac{1}{\boldsymbol{S}} - (\boldsymbol{Q}) \odot \boldsymbol{S} \odot \boldsymbol{A} - \boldsymbol{S}\text{diag}(\boldsymbol{\Omega}),
\end{aligned}
$$

where $\phi'_\varepsilon(x) = x\varepsilon^2/(x^2+\varepsilon^2)^2$ denotes the element-wise gradient of the $\phi_\varepsilon(x)$ penalty applied to the coefficient matrix $\boldsymbol{B}$. The gradient corresponding to the intercept term is set to zero, as the intercept is excluded from penalization.

Proposed algorithm. Algorithm 1 presents our optimization procedure. It is initialized using the estimator for the unpenalized ZIPLN model, provided in the R package `PLNmodels`.

4 Simulations

In this section, we investigate the performance of parameter estimation of the proposed SICZIPLN method within a controlled experimental study. Data are generated from the ZIPLN model (2) by varying three factors: the sample size, the sparsity level of $\boldsymbol{B}$, and the number of columns (or species) in the count matrix (see the supplementary materials for more details). The sparsity level is controlled by varying the proportion of zero entries in $\boldsymbol{B}$, corresponding to an increasing number of irrelevant variables for some species. The goal is to assess the impact of the penalization of $\boldsymbol{B}$ on parameter estimation accuracy under increasing sample sizes, increasing sparsity levels, and increasing numbers of species.

Data generation. Each entry of $\boldsymbol{X} \in \mathbb{R}^{n \times d}$ and $\widetilde{\boldsymbol{X}} \in \mathbb{R}^{n \times d_0}$ is generated independently from the uniform distribution ($\mathcal{U}(0,\, 1.5)$). We fixed the values of d and d_0 at 5. The entries of $\boldsymbol{B}$ take values

Algorithm 1 SICZIPLN algorithm

Initialization:
- $\pi^0 = (\boldsymbol{B}^0, \boldsymbol{\Sigma}^0, \boldsymbol{M}^0, \boldsymbol{S}^0, \widetilde{\boldsymbol{B}}^0)$,
- $\boldsymbol{E} = (\varepsilon_1, \cdots, \varepsilon_T)$ a vector of decreasing values, with $\varepsilon_t = \varepsilon_1 r^{t-1}$ where $r \in]0, 1[$.

for ε in $\boldsymbol{E}$ **do** ▷ ε-telescoping

repeat ▷ VEM

VE Step: Updating $\boldsymbol{M}$, $\boldsymbol{S}$ and $\boldsymbol{P}$

$$\boldsymbol{M}^{(t+1)} = \arg\max_{\boldsymbol{M}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}),$$
$$\boldsymbol{S}^{(t+1)} = \arg\max_{\boldsymbol{S}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}),$$
$$\boldsymbol{P}^{(t+1)} = \arg\max_{\boldsymbol{P}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}).$$

VM Step: Updating $\boldsymbol{B}$, $\boldsymbol{\Sigma}$ and $\widetilde{\boldsymbol{B}}$

$$\boldsymbol{B}^{(t+1)} = \arg\max_{\boldsymbol{B}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}),$$
$$\boldsymbol{\Omega}^{(t+1)} = \arg\max_{\boldsymbol{\Omega}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}),$$
$$\widetilde{\boldsymbol{B}}^{(t+1)} = \arg\max_{\widetilde{\boldsymbol{B}}} J_{\text{pen}}(\boldsymbol{Y}; \boldsymbol{\theta}, \boldsymbol{\Psi}).$$

until convergence; ▷ End VEM

Update $\pi^{(\varepsilon)} = (\boldsymbol{B}^{(t+1)}, \boldsymbol{\Omega}^{(t+1)}, \widetilde{\boldsymbol{B}}^{(t+1)}, \boldsymbol{M}^{(t+1)}, \boldsymbol{S}^{(t+1)}, \boldsymbol{P}^{(t+1)})$.

end for ▷ End ε-telescoping

Post-treatment: Set at 0 all entries of $\boldsymbol{B}$ below w ($w = 10^{-5}$), and update $\pi^{(\varepsilon_T)}$.

Return $\pi^{(\varepsilon_T)}$.

in $\{1, 0.5, 0\}$. All the entries of $\widetilde{\boldsymbol{B}}$ are set to 1. The covariance matrix $\boldsymbol{\Sigma}$ is defined as Toeplitz matrix with entries $\boldsymbol{\Sigma}_{hj} = 0.5^{|j-h|}$ where $(h, j) \in \{1, \ldots, p\}^2$.

Performance evaluation. Our evaluation framework relies on four standard metrics.

- **Root Mean Squared Error (RMSE)**: This measures the accuracy of the estimates by comparing the estimated values $\widehat{\boldsymbol{B}}$, $\widehat{\boldsymbol{Y}}$ and $\widehat{\boldsymbol{\Omega}}$ with the true values of $\boldsymbol{B}$, $\boldsymbol{Y}$ and $\boldsymbol{\Omega}$.
- **True Positive Rate (TPR) and True Negative Rate (TNR)**: The TPR measures the proportion of nonzero coefficients in $\boldsymbol{B}$ that are correctly identified as nonzero by $\widehat{\boldsymbol{B}}$, whereas the TNR measures the proportion of zero coefficients in $\boldsymbol{B}$ that are correctly identified as zero by $\widehat{\boldsymbol{B}}$. They are defined as follows:

$$\text{TPR}(\widehat{\boldsymbol{B}}, \boldsymbol{B}) = \frac{\text{Card}\left\{(k,j) : \hat{b}_{kj} \neq 0 \text{ and } b_{kj} \neq 0\right\}}{\text{Card}\left\{(k,j) : b_{kj} \neq 0\right\}},$$
$$\text{TNR}(\widehat{\boldsymbol{B}}, \boldsymbol{B}) = \frac{\text{Card}\left\{(k,j) : \hat{b}_{kj} = 0 \text{ and } b_{kj} = 0\right\}}{\text{Card}\left\{(k,j) : b_{kj} = 0\right\}},$$

where $\text{Card}\{.\}$ denotes the cardinality of the set.

- **Information criterion.** For each simulation, we evaluated the BIC and the Integrated Completed Likelihood (ICL). We recall below the BIC and ICL formulas for PLN and ZIPLN:

$$\text{BIC} = \text{ELBO} - \frac{K \log(n)}{2},$$
$$\text{ICL} = \text{BIC} - H(\Psi),$$

where K denotes the number of parameters, and $H(.)$ the entropy term of PLN (Chiquet et al. 2021) or ZIPLN (Batardière et al. 2025). For PLN, $K = p(p+1)/2 + pd$, and for ZIPLN, $K = p(p+1)/2 + pd + pd_0$.

Compared methods. To contextualize the performance of our proposed method, SICZIPLN, relative to existing approaches, we evaluate the following models in our simulation study.

- **PLN:** The standard PLN model as proposed by Chiquet et al. (2021).
- **SICPLN:** A penalized version of the PLN model (SICPLN) proposed by Kioye et al. (2026).
- **ZIPLN:** The standard multivariate Zero-Inflated Poisson Log-Normal Model proposed by Batardière et al. (2025).
- **SICZIPLN:** Our proposed method.

4.1 Experiment 1: Performance evaluation across sample sizes

4.1.1 Experiment design

In this study, we consider a fixed coefficient matrix $\boldsymbol{B}$ in which 25% of the entries are randomly set to zero. The number of columns in the count matrix is fixed at $p = 5$, while the sample size varies from 100 to 1000. For each sample size, we estimate the parameters and assess the RMSE between the true parameter values and their corresponding estimates. Additionally, we compute the BIC and the ICL to evaluate model performance. This experiment is repeated across 100 independent datasets.

4.1.2 Experiment Results

RMSE of $\boldsymbol{B}$. The left panel of Figure 1 presents the RMSE values obtained by comparing the coefficients of the matrix $\boldsymbol{B}$ with their corresponding estimates. We observe that the models accounting for zero inflation (ZIPLN and SICZIPLN) achieve lower RMSE values than their non zero inflated counterparts (PLN and SICPLN), regardless of the sample size. The penalized versions of PLN and ZIPLN models, denoted by SICPLN and SICZIPLN, respectively, promote sparsity while maintaining RMSE values close to those of their unpenalized counterparts. This behavior may be attributed to the relatively low bias induced by the penalty function used in the SIC framework, which

allows the penalized models to achieve sparse estimates without compromising estimation accuracy.

RMSE of $\boldsymbol{Y}$. The right panel of Figure 1 shows the RMSE of prediction errors of the different methods. The RMSE associated with SICPLN is higher than the RMSE of the other methods. The RMSE of PLN is lower than those of ZIPLN and SICZIPLN. The gap between the true value and the predicted value can be large for some observations with PLN and SICPLN, thereby pulling the violin plots toward extreme points. The RMSE of ZIPLN and SICZIPLN remain close to each other and avoid large prediction errors for some observations. This highlights that ZIPLN and SICZIPLN can be a good trade-off when $\boldsymbol{Y}$ is sparse.

For the SICPLN and SICZIPLN models, the regularization parameter λ was fixed at $\log(n)$ for minimizing the BIC criterion. It is worth noting that the value of λ minimizing the BIC may differ from the value that minimizes the estimation and the prediction errors.

TNR and TPR. Figure 2 presents the performance of the different methods in terms of the true positive rate (TPR) and true negative rate (TNR) over 100 datasets. The left plot presents the TPR results. The TPR of PLN and ZIPLN remains equal to 1, indicating that these two cannot promote sparsity in the estimations. Each estimated coefficient takes a non-zero value. In contrast, SICPLN and SICZIPLN promote sparsity and can estimate a large proportion of coefficients that take non-zero values as non-zero. This ability increases when the sample size increases, particularly for SICZIPLN, which has a TPR value greater than 0.75. The right plot presents the results for the TNR. These results highlight SICZIPLN's ability to identify zero coefficients by estimating them exactly at zero. In contrast, SICPLN appears less effective in detecting zero coefficients in the matrix $\boldsymbol{B}$, likely due to the absence of zero inflation. Finally, the PLN and ZIPLN models exhibit a zero TNR, as they do not allow for the estimation of certain coefficients exactly at zero.

BIC and ICL. Figure 3 presents the results of the absolute values of BIC and ICL for each method obtained over 100 datasets. The left panel shows the BIC values, and the right panel presents the ICL values. In both cases, the lowest value is obtained by SICZIPLN, followed by the ZIPLN model. These results highlight the benefits of accounting for zero inflation in improving model performance. The performance of the ZIPLN model further improves when variable selection is incorporated. This improvement is consistent with the reduction in irrelevant information resulting from variable selection.

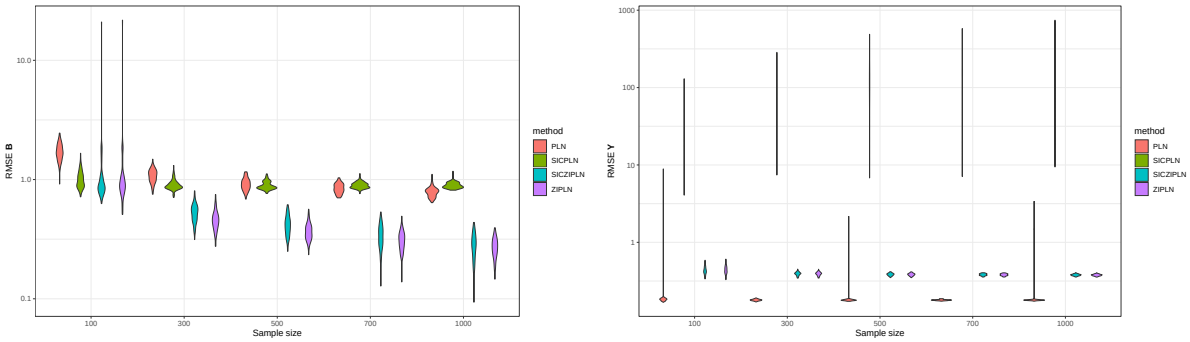


Fig. 1: Violin plots of 100 replication values depicting the RMSE between true values (denoted by $\boldsymbol{B}$, and $\boldsymbol{Y}$) and their corresponding estimates with variations in sample size $n \in \{100, 300, 500, 700, 1000\}$. Left plot shows the violin plots for coefficient matrix. Right plot presents the violin plots for prediction error

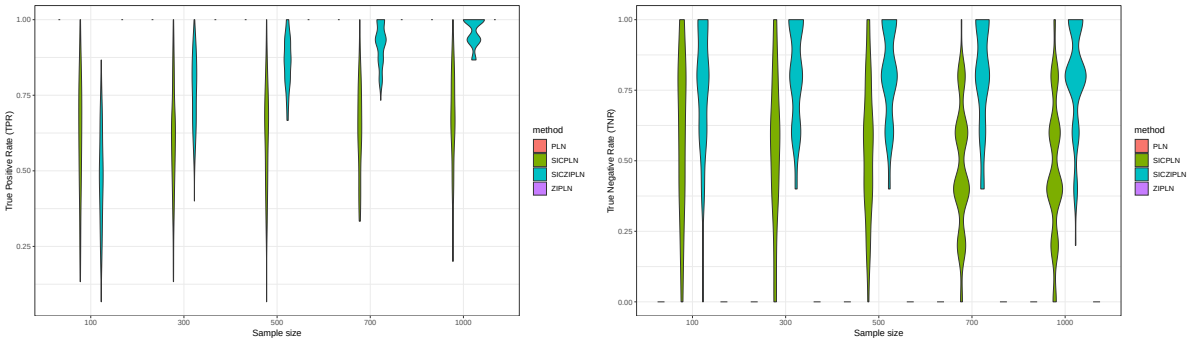


Fig. 2: True Positive Rate (TPR) and True Negative Rate (TNR) when sample sizes varying between 100 and 1000 for PLN, SICPLN, SICZIPLN and ZIPLN. Left plot shows the violin plots for the TPR. Right plot presents the the violin plots for the TNR

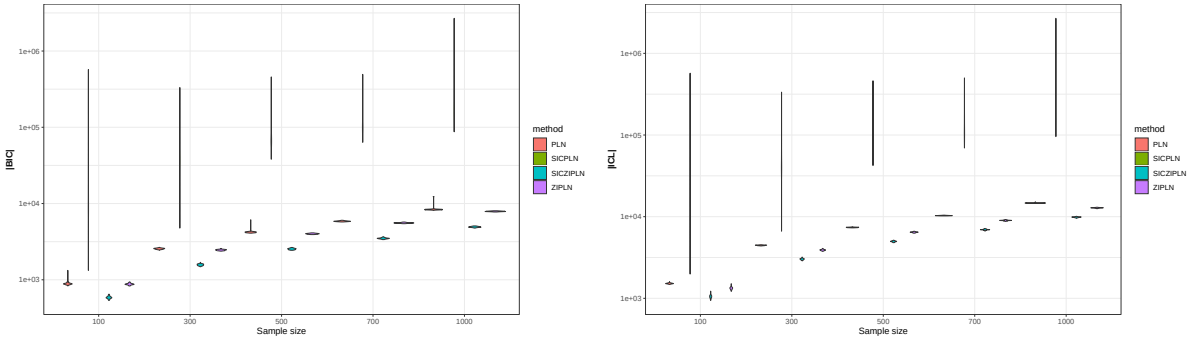


Fig. 3: Violin plots of 100 datasets values depicting BIC and the ICL with variations in sample size $n \in \{100, 300, 500, 7000, 1000\}$. Left plot shows the violin plots of the BIC. Right plot presents the violin plots for the ICL

4.2 Experiment 2: Effect of the proportion of zero coefficients in $\boldsymbol{B}$

4.2.1 Experiment design

In this simulation framework, we fixed the sample size at $n = 1000$ and the number of explanatory variables for both PLN and ZI components at 5. The proportion of null coefficients in the matrix $\boldsymbol{B}$ was varied from 25% to 90%. For each configuration, we generated $\boldsymbol{Y}$ According to ZIPLN models, and then fitted the PLN, SICPLN, ZIPLN, and SICZIPLN models.

The estimation accuracy was assessed by computing the RMSE between the true parameters $\boldsymbol{B}$ and $\boldsymbol{Y}$ and their corresponding estimates. Additionally, model fit was evaluated using the BIC and the ICL. Finally, we assessed each method's ability to correctly identify null and non-null coefficients by computing the TNR and the TPR.

4.2.2 Experiment results

RMSE of $\boldsymbol{B}$. In the left panel of Figure 4, we observe that methods accounting for zero inflation (ZIPLN and SICZIPLN) outperform those that do not consider this feature (PLN and SICPLN) in terms of RMSE. The superior performance these methods come from the additional latent layer that accounts for zero inflation in the ZIPLN model. Moreover, as the proportion of zeros increases, the estimates obtained using the SICZIPLN method converge towards the true values by shrinking some coefficients exactly to zero. SICPLN and SICZIPLN are based on the same penalty function, which presents similar properties to SCAD (Fan and Li 2001) and other reweighted methods (Zou 2006; Candes et al. 2008), taking zero inflation into account can help reduce estimation errors.

RMSE of $\boldsymbol{Y}$. In the right panel of Figure 4, the prediction errors indicate that PLN produces lower prediction errors compared to SICPLN, ZIPLN and SICZIPLN. ZIPLN and SICZIPN present similar prediction errors, which are lower than those of SICPLN.This result shows that taking zero inflation into account may not be necessary to reduce prediction errors, but it could help avoid predictions toward extreme values, as observed when the proportion of zeros in $\boldsymbol{B}$ is 0.5. In contrast, when zero inflation is taken into account, variation in the proportion of zeros in $\boldsymbol{B}$ does not have a significant impact on the prediction error.

TNR and TPR. Figure 5 presents the performance of different methods in terms of TPR and TNR over 100 datasets, as a function of the proportion of zeros in the matrix $\boldsymbol{B}$. In the left panel, the TPR of both PLN and ZIPLN are equal to 1. This was expected because these two methods estimate each coefficient as different from zero. Although SICPLN and SICZIPLN shrink the estimated coefficients toward zero, these methods

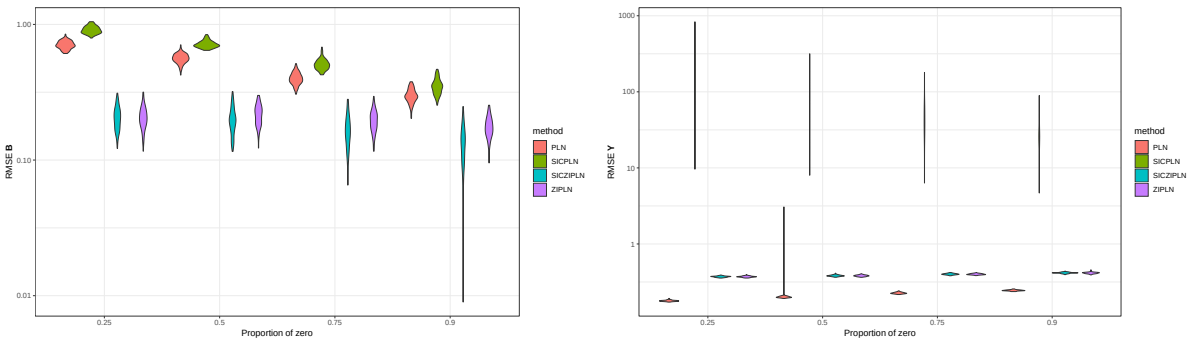


Fig. 4: Violin plots of 100 replication values depicting the RMSE between true values (denoted by $\boldsymbol{B}$, and $\boldsymbol{Y}$) and their corresponding estimates with variations in proportion $p \in \{0.25, 0.5, 0.75, 0.9\}$ of zeros in $\boldsymbol{B}$. Left plot shows the violin plots for the RMSE of $\boldsymbol{B}$. Right plot presents the violin plots for $\boldsymbol{Y}$

estimate a large proportion of non-null coefficients as non-null. In fact, the SIC penalty induces non-uniform shrinkage, which avoids shrinking the large coefficients. However, SICZIPLN presents a higher TPR than SICPLN, and its TPR is closer to those of PLN and ZIPLN for some proportions of zeros.

Unlike the left panel, the TNR of both PLN and ZIPLN are equal to 0 because these methods do not induce sparsity in $\boldsymbol{B}$. As expected, SICPLN and SICZIPLN promote sparsity, with TNR different from zero when the proportion of zeros in $\boldsymbol{B}$ is large. SICZIPLN obtains a good TNR (greater than 0.75), regardless of the proportion of zeros in $\boldsymbol{B}$.

BIC and ICL. Figure 6 presents the results of the absolute values of BIC and ICL as a function of the proportion of zeros. In both the left and right panels, the best information criterion is obtained by SICZIPLN, whereas those of PLN and ZIPLN are sometimes closer to each other. This shows that variable selection improves the performance of ZIPLN model, and this improvement is relevant when $\boldsymbol{B}$ is sparse. The absolute values of the BIC and ICL for SICPLN highlight that performing variable selection when $\boldsymbol{Y}$ is sparse, without accounting for zero inflation in the model, does not improve the information criteria.

4.3 Conclusion of the simulation studies

The SICZIPLN is suitable for achieving variable selection when $\boldsymbol{Y}$ is sparse, compared to its version that does not incorporate zero inflation. In fact, it allows to: 1) reduce the RMSE of the penalized parameters when the sample size increases and when many components of $\boldsymbol{B}$ need to be estimated exactly at 0; 2) provide a good trade-off between achieving accurate estimation and inducing sparsity; and 3) achieve good BIC and ICL values compared to PLN, ZIPLN, and SICPLN. We provide an additional simulation study in section 1

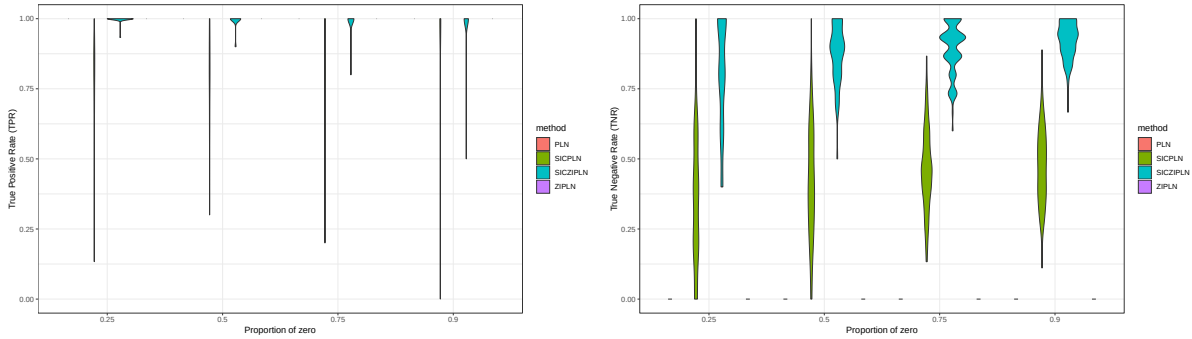


Fig. 5: True Positive Rate (TPR) and True Negative Rate (TNR) when the proportion of zero varying between 0.25 to 0.90 for PLN, SICPLN, SICZIPLN and ZIPLN. Left plot shows the violin plots for the TPR. Right plot presents the the violin plots for the TNR

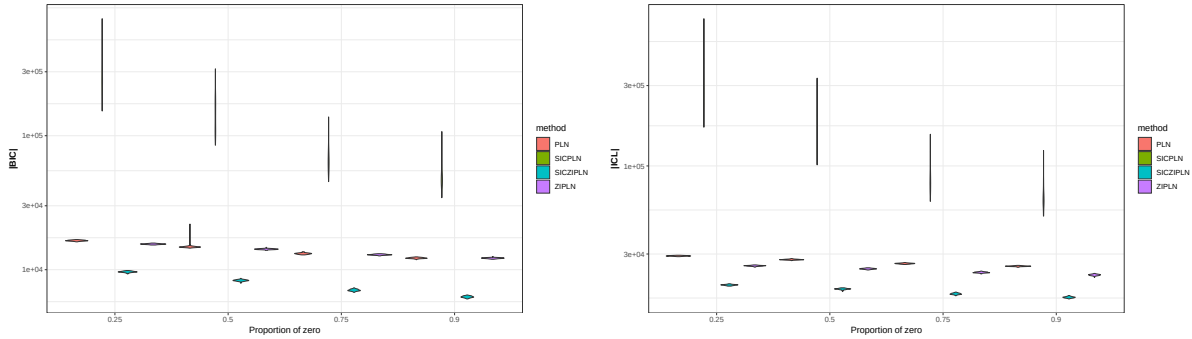


Fig. 6: Violin plots of 100 replication values depicting BIC and the ICL with variations in proportion $p \in \{0.25, 0.5, 0.75, 0.9\}$ of zeros in $\boldsymbol{B}$. Left plot shows the violin plots of the BIC. Right plot presents the violin plots for the ICL

of the supplementary materials to evaluate the performance of SICZIPLN when the number of columns in $\boldsymbol{Y}$ (or the number of species) increases. We also show the influence of the penalization in terms of estimation accuracy for the precision matrix $\boldsymbol{\Omega}$ in the Figure 4 of the supplementary materials.

5 Applications

5.1 Data description

We consider a study aimed at identifying relevant variables related to the microbiology of milk, diet, and milk biochemistry that could explain the abundance of certain microbial communities involved in the milk production process. These data come from the Amont Saint-Nectaire project.

The data were collected from 7 different ecosystems: air, water, teats, milk filter, bedding,

feces, and the milk. The samples were taken from 14 farms, with collections carried out in both summer and winter. Three samples were taken for each ecosystem, season, and farm. The total sample size is $n = 539$ when we concatenate the data from the 7 ecosystems by row. Samples of milk, teat, and bedding were prepared and stored as described in (Verdier-Metz et al. 2023) and feces samples as described by Musati et al. (2025). Air and bedding samples were prepared and stored as milk filter and teat samples, respectively. After applying bioinformatics preprocessing (Theil and Rifa 2019, 2021) the 16S amplicon sequence variants (ASV) were assigned using an environment specific database (Meola et al. 2019; Callahan et al. 2016) and a general database (SILVA r 138).

We consider two scenarios to evaluate how different methods perform depending on the proportion of zeros in the response matrix. In the first scenario, the 16S ASV with a prevalence of less than 15% were excluded (scenario 1). In the second scenario, the 16S ASV with a prevalence of less than 5% were excluded (scenario 2). As the abundances of ASV are strongly associated with ecosystems and seasons, we have chosen to focus here on data from the teats with the genus as the taxonomic level.

Regarding the teats data in winter, for scenario 1, we obtained a sample size of $n = 42$ with $p = 49$ genera (response variables), with 37% of the entries in the response matrix being zero. For scenario 2, the sample size remained $n = 42$, but p increased to 84 genera, with 52% of the entries in the response matrix being zero.

We consider the following covariates: Diet_Milk_FromPasture (proportion of milk produced on pasture), LAB (enumeration of lactic acid bacteria) and MBC_CCS (enumeration of somatic cell counts in milk). We then fitted the PLN, ZIPLN, SICPLN, and SICZIPLN models in order to identify the relevant variables that influence each response variable.

5.2 Results and interpretations

Information criteria. The data were fitted using four models: PLN, SICPLN, ZIPLN, and SICZIPLN. Model performance was then compared using BIC. The results are presented in Table 1 for scenarios 1 and 2. For both scenarios, the best fit in terms of BIC is obtained with SICZIPLN, with values of -9630.629 and -15874.45, respectively. The superior performance of SICZIPLN compared to the other methods can be explained by the combined effect of accounting for zero-inflation in the PLN model and implementing variable selection, which enhances the overall performance of the models. These results further indicate that, in a non-penalized setting, the ZIPLN model does not necessarily outperform a standard PLN model. Indeed, the BIC of PLN is higher than that of ZIPLN. SICPLN demonstrates poorer performance comparatively to others methods. The

observed BIC results in these two real-world scenarios are consistent with those observed in the simulation studies.

The heatmaps of the estimated coefficients for scenario 1 of PLN are presented in Figure 7a. Those of SICPLN are shown in Figure 8a , those of ZIPLN are illustrated in Figure 7b, and finally those of SICZIPLN are presented in Figure 8b. The heatmaps of the estimated coefficients for scenario 2 are provided in the supplementary materials (5a, 6a, 5b, and 6b for respectively PLN, SICPLN, ZIPLN and SICZIPLN). These visualizations highlight the differences in coefficient estimates across methods and emphasize the improvements brought by SICZIPLN as a good trade-off between inducing sparsity and achieving good information criteria.

Model	BIC scenario 1	BIC scenario 2
PLN	-10194.78	-16501.87
SICPLN	-240345.2	-298421.6
ZIPLN	-10216	-16841.38
SICZIPLN	**-9630.629**	**-15874.45**

Table 1: BIC values for the four methods (PLN, SICPLN, ZIPLN, and SICZIPLN). The best value for each scenario is shown in bold (higher the better)

Sparsity. Figure 8 and Figure 6 in the supplementary materials show that a larger number of coefficients are estimated exactly at zero when zero-inflation is not modeled. Consequently, the SICPLN estimates appear more parsimonious than those obtained with the SICZIPLN model with more than 65% and 73% of coefficients estimated at 0 respectively for the scenarios 1 and 2. The corresponding proportions of coefficients estimated at zero for each method are reported in Table 2. The lower BIC observed for SICPLN relative to SICZIPLN can be attributed to the fact that SICPLN tends to shrink to zero some coefficients that should not be 0 in practice, leading to an overly sparse representation. SICZIPLN avoids an overly sparse representation of the support of the estimated coefficients with more than 50% and 22% of coefficients estimated at 0 respectively for the scenarios 1 and 2. This is observed through the TPR and TNR in the simulation studies in which the true parameters are known. Finally, it should be noted that in these analyses, the regularization parameter was fixed at $\log(n)$ for all methods.

Bias-parsimony trade-off. Variable selection relies on a trade-off between model parsimony and the bias introduced in the estimates. From this perspective, SICZIPLN, which accounts for zero inflation, helps reduce bias while maintaining a reasonable level of parsimony. This observation is consistent with the results presented in Tables 1 and 2, as well as in Section 4 which underscore the superiority of the SICZIPLN model over the others methods in terms of the bias-parsimony trade-off.

Assessing the impact of bias on data reconstruction. Given that selection procedures inherently introduce bias in parameter estimates, it is

also relevant to assess the impact of this bias on the models ability to reconstruct the count matrices in the two scenarios. Table 3 reports the overall prediction errors, expressed in terms of RMSE, for each method. As in the simulation studies, the PLN model demonstrates a better ability to reconstruct the data, with lower RMSE values than SICPLN, ZIPLN, and SICZIPLN in the both scenarios.

Table 4 presents the RMSE of the prediction errors associated with the zero entries in the count matrices. We can remark that ZIPLN and SICZIPLN have lower RMSE than PLN and SICPLN. This indicates that the additional latent layers following a Bernoulli distribution help to predict the entries corresponding to zeros in the data.

The performance of SICPLN and SICZIPLN in Tables 3 and 4 could be improved by finding another value of λ that minimizes the prediction errors. This new value of λ could be achieved by accepting a degradation in model performance in terms of BIC.

Model	Scenario 1	Scenario 2
PLN	0	0
SICPLN	**0.6598639**	**0.7380952**
ZIPLN	0	0
SICZIPLN	0.5170068	0.2222222

Table 2: Proportion of zero coefficients estimated by the four methods (PLN, SICPLN, ZIPLN, and SICZIPLN). The high value for each scenario is shown in bold

Model	Scenario 1	Scenario 2
PLN	**3.49279**	**2.720131**
SICPLN	23.86808	20.87189
ZIPLN	4.147834	3.421849
SICZIPLN	741.4447	576.655

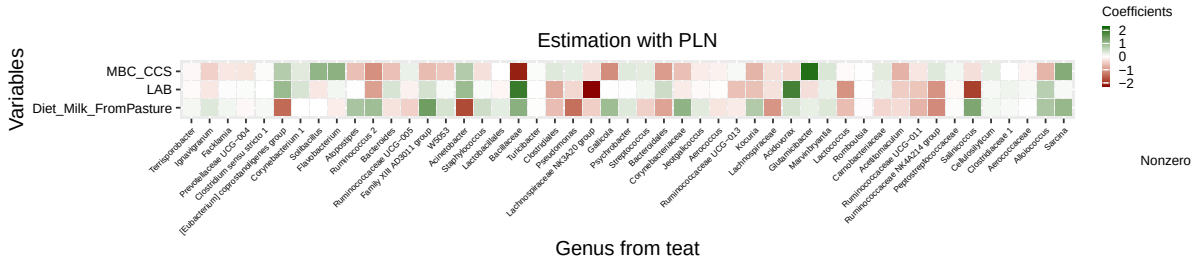
Table 3: RMSE of prediction errors across all entries of the count matrix for the four methods (PLN, SICPLN, ZIPLN, and SICZIPLN). The best value for each scenario is shown in bold

Model	Scenario 1	Scenario 2
PLN	0.6164954	0.3659356
SICPLN	0.3882336	0.303811
ZIPLN	0.3265603	0.2835277
SICZIPLN	**3.91489e-05**	**2.742393e-05**

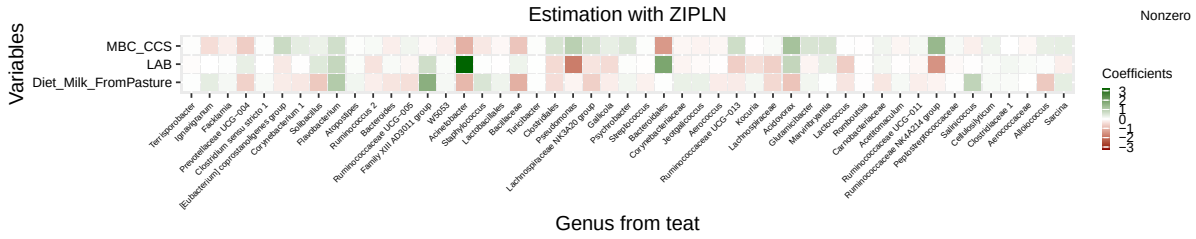
Table 4: RMSE of prediction errors for zero entries in the count matrix for the four methods (PLN, SICPLN, ZIPLN, and SICZIPLN). The best value for each scenario is shown in bold

6 Conclusion and discussion

In this paper, we proposed an extension of variable selection based on coupling the SIC and ZIPLN model. This approach led us to incorporate a recent variable selection method into the ZIPLN model, which is defined as a mixture of a degenerate distribution at zero and a standard PLN model. The inference was then carried out by combining the algorithm specific to the SIC, known as ε-telescoping, which progressively approximates the L_0 norm, with the VEM algorithm for the penalized ZIPLN. This combination allowed us to gradually shrink the coefficients associated with irrelevant variables to exactly zero. The resulting



(a) Estimation with PLN



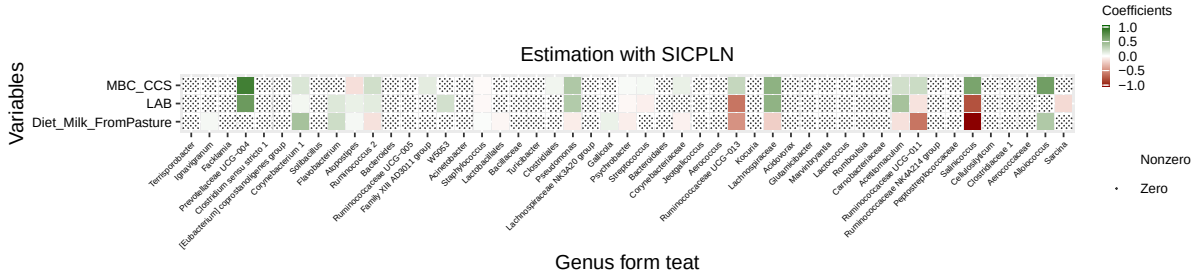
(b) Estimation with ZIPLN

Fig. 7: Scenario 1: Coefficient estimates from unpenalized PLN and ZIPLN models on winter teat microbiota with 37% zeros in the count matrix

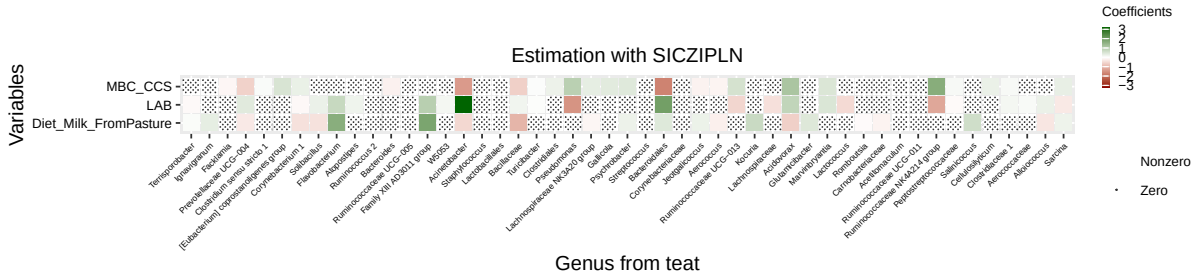
method was first applied to simulated data under various scenarios to assess its relevance, and then to real data from a study aimed at identifying variables that influence the abundances of microbial communities involved in the milk production process. The results demonstrate improved performance of the proposed extension, particularly in its ability to estimate coefficients while promoting sparsity.

Empirically, the prediction errors obtained with the proposed method is higher to those produced by a the other competive methods. Achieving a lower prediction could consist of finding a different regularization parameter from $\log(n)$ as suggested by O'Neill and Burke (2023). This could come at the cost of degrading the BIC of SICZI-PLN in the worst-case scenario. Even though the simulation study shows comparable results between the penalized and unpenalized models in terms of parameters estimation, we do not evaluate how the penalization impacts the variational parameters because these parameters are involved in the inference step, not in the data generation process.

Given the potential offered by approximating the L_0 norm and modeling multivariate



(a) Estimation with SICPLN



(b) Estimation with SICZIPLN

Fig. 8: Scenario 1: Coefficient estimates from penalized SICPLN and SICZIPLN models on winter teat microbiota with 37% zeros in the count matrix

count data, it seems relevant to further explore these extensions. Currently, telescoping requires a thresholding step at the end of the algorithm, which can be interpreted as a direct consequence of the penalty function. We believe that this thresholding could be avoided by adopting an alternative penalty function, such as those proposed in Candes et al. (2008); Castella et al. (2018). Being non-differentiable, this function have led some authors, notably Castella et al. (2018), to use semidefinite programming relaxation techniques to induce sparsity in penalized least squares problems. Furthermore, another possible improvement would be to also induce sparsity on the precision matrix as well as on the coefficients associated with zero inflation. However, this approach would introduce additional bias, which could compromise the performance of the variable selection.

Supplementary information. Supplementary material containing additional simulation and application results is provided.

Acknowledgements. We would like to thank Celine Delbès, Isabelle Verdier, and Sebastien Theil for providing the data and for fruitful discussions.

Funding. This work was supported by the DATA program and IRC-SAE related to the ISITE/CAP2025 funding program of Clermont Auvergne University.

Code availability. The proposed algorithm has been implemented in an R (R Core Team et al. 2022) package called SICZIPLN, which is available on GitHub: https://github.com/kioye-jean-yves/SICZIPLN

Conflict of interest/Competing interests. The authors declare no competing interests.

CRediT authorship contribution statement. Togo Jean Yves KIOYE: Conceptualization, Methodology, Software, Formal analysis, Writing Original Draft. **Paul-Marie GROLLEMUND:** Conceptualization, Methodology, Review of the Original Draft, Supervision, Project administration, Funding acquisition. **Jocelyn CHAUVET:** Supervision, Conceptualization, Methodology, Review of the Original Draft. **Christophe CHASSARD:** Supervision, Project administration, Funding acquisition.

Appendix A Proofs and mathematical details

Proof of Proposition 2.1 The final expression of the ELBO, as introduced in Equation (6) can be written as:

$$\begin{aligned} J(\boldsymbol{Y}, \boldsymbol{\theta}, \boldsymbol{\Psi}) =& \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W})] + \\ & \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Z})] + \\ & \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{W})] - \\ & \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log q_{\psi_1}(\boldsymbol{Z})] - \\ & \mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log q_{\psi_2}(\boldsymbol{W})], \end{aligned}$$

each of these expectation terms can be derived as follows:

- **Derivation of** $\mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Y} \mid \boldsymbol{Z}, \boldsymbol{W})]$

$$\begin{aligned} =& \sum_{i=1}^{n}\sum_{j=1}^{p}(1-p_{ij})(y_{ij}(o_{ij}+m_{ij}) \\ & - e^{o_{ij}+m_{ij}+s_{ij}^2/2} - \log(y_{ij}!)) + p_{ij}\delta(y_{ij}) \\ =& \mathbf{1}_n^\top[\boldsymbol{Q} \odot (\boldsymbol{Y} \odot (\boldsymbol{O}+\boldsymbol{M}) - \boldsymbol{A} - \log(\boldsymbol{Y}!)) + \\ & \boldsymbol{P} \odot \boldsymbol{\delta}_0(\boldsymbol{Y})]\mathbf{1}_p, \end{aligned}$$

  where $\boldsymbol{Q} = \mathbf{1}_{n,p} - \boldsymbol{P}$ and $\boldsymbol{\delta}_0(\boldsymbol{Y})$ is the Dirac mass at zero.

- **Derivation of** $\mathbb{E}_{q_{\boldsymbol{\Psi}}}[\log p_{\boldsymbol{\theta}}(\boldsymbol{Z})]$

$$\begin{aligned} =& -\frac{n}{2}\log(|\boldsymbol{\Sigma}|) - \frac{np}{2}\log(2\pi) \\ & - \frac{1}{2}\sum_{i=1}^{n}(\boldsymbol{m}_i - \boldsymbol{x}_i^\top\boldsymbol{B})^\top\boldsymbol{\Omega}(\boldsymbol{m}_i - \boldsymbol{x}_i^\top\boldsymbol{B}) \end{aligned}$$

$$
\begin{aligned}
&- \frac{1}{2}\text{trace}(\boldsymbol{\Omega}\text{diag}(\boldsymbol{s}_i^2)) \\
=&-\frac{1}{2}\big[n\log(|\boldsymbol{\Sigma}|) + np\log(2\pi) + \\
&\mathbf{1}_n^\top\left((\boldsymbol{M}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Omega}(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{B})^\top\right)\mathbf{1}_n + \\
&\text{trace}(\boldsymbol{\Omega}\tilde{\boldsymbol{S}}^2)\big].
\end{aligned}
$$

- **Derivation of** $\mathbb{E}_{q_{\boldsymbol{\Psi}}(\boldsymbol{Z},\boldsymbol{W})}[\log p_{\boldsymbol{\theta}}(\boldsymbol{W})]$

$$
\begin{aligned}
=&\sum_{i=1}^{n}\sum_{j=1}^{p}\mathbb{E}_{q_{\boldsymbol{\Psi}}}[w_{ij}]\log(\pi_{ij}) + \\
&(1-\mathbb{E}_{q_{\boldsymbol{\Psi}}}[w_{ij}])\log(1-\pi_{ij}), \\
=&\sum_{i=1}^{n}\sum_{j=1}^{p} p_{ij}\ \log(\pi_{ij}) + (1-p_{ij})\log(1-\pi_{ij}), \\
=&\mathbf{1}_n^\top(\boldsymbol{P}\odot\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}} - \log(\mathbf{1}_{n,p}+e^{(\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}})})\mathbf{1}_p.
\end{aligned}
$$

- **Derivation of** $\mathbb{E}_{q_{\boldsymbol{\Psi}}}[\log q_{\boldsymbol{\psi}_1}(\boldsymbol{Z})]$

$$
\begin{aligned}
=&-\frac{np}{2}\log(2\pi) - \sum_{i=1}^{n}\frac{n}{2}\log(|\text{diag}(s_i^2)|) + \frac{np}{2}, \\
=&-\frac{np}{2}\log(2\pi) + \frac{np}{2} + \mathbf{1}_n^\top\log(\boldsymbol{S}^2)\mathbf{1}_p.
\end{aligned}
$$

- **Derivation of** $\mathbb{E}_{q_{\boldsymbol{\Psi}}}[\log q_{\boldsymbol{\psi}_2}(\boldsymbol{W})]$

$$
\begin{aligned}
=&\sum_{i=1}^{n}\sum_{j=1}^{p} p_{ij}\log p_{ij} + (1-p_{ij})\log(1-p_{ij}) \\
=&\mathbf{1}_n^\top(\boldsymbol{P}\odot\log(\boldsymbol{P}) + (\mathbf{1}_{n,p}-\boldsymbol{P})\odot\log(\mathbf{1}_{n,p}-\boldsymbol{P}))\mathbf{1}_p.
\end{aligned}
$$

Finally, by regrouping all the terms and expressing them in matrix form, we obtain the full ELBO expression:

$$
\begin{aligned}
J(\boldsymbol{Y},\boldsymbol{\theta},\boldsymbol{\Psi}) =&\mathbf{1}_n^\top[\boldsymbol{Q}\odot(\boldsymbol{Y}\odot(\boldsymbol{O}+\boldsymbol{M}) - \boldsymbol{A} \\
&- \log(\boldsymbol{Y}!)) + \boldsymbol{P}\odot\boldsymbol{\delta}_0(\boldsymbol{Y})]\mathbf{1}_p \\
&\frac{-1}{2}\big[-n\log(|\boldsymbol{\Omega}|) + \\
&\mathbf{1}_n^\top\left((\boldsymbol{M}-\boldsymbol{X}\boldsymbol{B})\boldsymbol{\Omega}(\boldsymbol{M}-\boldsymbol{X}\boldsymbol{B})^\top\right)\mathbf{1}_n \\
&+\text{trace}(\boldsymbol{\Omega}\tilde{\boldsymbol{S}}^2)\big] + \\
&\mathbf{1}_n^\top(\boldsymbol{P}\odot\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}} - \log(\mathbf{1}_{n,p}+e^{\widetilde{\boldsymbol{X}}\widetilde{\boldsymbol{B}}}))\mathbf{1}_p \\
&+\frac{np}{2} + \frac{1}{2}\mathbf{1}_n^\top\log(\boldsymbol{S}^2)\mathbf{1}_p + \\
&\mathbf{1}_n^\top(-\boldsymbol{P}\odot\log(\boldsymbol{P}) - \boldsymbol{Q}\odot\log(\boldsymbol{Q}))\mathbf{1}_p.
\end{aligned}
$$

□

References

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

Batardière, B., Chiquet, J., Gindraud, F., Mariadassou, M.: Zero-inflation in the multivariate poisson lognormal family: B. batardière et al. Statistics and Computing **35**(6), 196 (2025)

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

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

Chen, S., Donoho, D.: Basis pursuit. In: Proceedings of 1994 28th Asilomar Conference on Signals, Systems and Computers, vol. 1, pp.

41–44 (1994). IEEE

Chouzenoux, E., Jezierska, A., Pesquet, J.-C., Talbot, H.: A majorize-minimize subspace approach for \ell_2-\ell_0 image regularization. SIAM Journal on Imaging Sciences **6**(1), 563–591 (2013)

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J.A., Holmes, S.P.: Dada2: High-resolution sample inference from illumina amplicon data. Nature methods **13**(7), 581–583 (2016)

Chiquet, J., Mariadassou, M., Robin, S.: The poisson-lognormal model as a versatile framework for the joint analysis of species abundances. Frontiers in Ecology and Evolution **9**, 588292 (2021)

Cohen, A.C.: Estimation in Mixtures of Discrete Distributions. Statistical Pub. Society Calcutta, ??? (1963)

Castella, M., Pesquet, J.-C.: Optimization of a geman-mcclure like criterion for sparse signal deconvolution. In: 2015 IEEE 6th International Workshop on Computational Advances in Multi-Sensor Adaptive Processing (CAMSAP), pp. 309–312 (2015). IEEE

Castella, M., Pesquet, J.-C., Marmin, A.: Rational optimization for nonlinear reconstruction with approximate l0 penalization. IEEE Transactions on Signal Processing **67**(6), 1407–1417 (2018)

Candes, E.J., Wakin, M.B., Boyd, S.P.: Enhancing sparsity by reweighted l 1 minimization. Journal of Fourier analysis and applications **14**, 877–905 (2008)

Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the em algorithm. Journal of the royal statistical society: series B (methodological) **39**(1), 1–22 (1977)

Florescu, A., Chouzenoux, E., Pesquet, J.-C., Ciuciu, P., Ciochina, S.: A majorize-minimize memory gradient method for complex-valued inverse problems. Signal Processing **103**, 285–295 (2014)

Fan, J., Li, R.: Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American statistical Association **96**(456), 1348–1360 (2001)

Foucart, S., Lai, M.-J.: Sparsest solutions of underdetermined linear systems via lq-minimization for $0 \leq q \leq 1$. Applied and Computational Harmonic Analysis **26**(3), 395–407 (2009)

Geman, S.: Statistical methods for tomographic image restoration. Bull. Internat. Statist. Inst.

52, 5–21 (1987)

Hall, D.B.: Zero-inflated poisson and binomial regression with random effects: a case study. Biometrics **56**(4), 1030–1039 (2000)

Heilbron, D.C.: Zero-altered and other regression models for count data with added zeros. Biometrical Journal **36**(5), 531–547 (1994)

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

Johnson, N., Kotz, S.: Distributions in Statistics: Di. Continuous (1969)

Kioye, T.J.Y., Grollemund, P.-M., Chauvet, J., Druilhet, P., Saint-Loubert-Bie, E., Chassard, C.: Sparse inference in poisson log-normal model by approximating the l0-norm. Journal of Multivariate Analysis, 105686 (2026)

Lambert, D.: Zero-inflated poisson regression, with an application to defects in manufacturing. Technometrics **34**(1), 1–14 (1992)

Lee, K.H., Coull, B.A., Moscicki, A.-B., Paster, B.J., Starr, J.R.: Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. Biostatistics **21**(3), 499–517 (2020)

Li, C.-S., Lu, J.-C., Park, J., Kim, K., Brinkley, P.A., Peterson, J.P.: Multivariate zero-inflated poisson models and their applications. Technometrics **41**(1), 29–38 (1999)

Musati, M., Coppa, M., Delbès, C., Verdier-Metz, I., Popova, M., Niderkorn, V., Bouchon, M., Farizon, Y., Enjalbert, F., Renna, M., Lussiana, C., Mangione, G., Martin, B., Ferlay, A.: The ruminal and faecal microbiota, digestion processes, and milk composition of dairy cows are modified by the botanical biodiversity of pastures. animal **19**(6), 101537 (2025) https://doi.org/10.1016/j.animal.2025.101537

Marmin, A., Castella, M., Pesquet, J.-C.: How to globally solve non-convex optimization problems involving an approximate l0 penalization. In: ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), pp. 5601–5605 (2019). IEEE

Meola, M., Rifa, E., Shani, N., Delbès, C., Berthoud, H., Chassard, C.: Dairydb: a manually curated reference database for improved taxonomy annotation of 16s rrna gene sequences from dairy products. BMC genomics **20**(1), 560 (2019)

Mullahy, J.: Specification and testing of some modified count data models. Journal of econometrics **33**(3), 341–365 (1986)

Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM journal on computing

24(2), 227–234 (1995)

O'Neill, M., Burke, K.: Robust distributional regression with automatic variable selection. arXiv preprint arXiv:2212.07317 (2022)

O'Neill, M., Burke, K.: Variable selection using a smooth information criterion for distributional regression models. Statistics and Computing **33**(3), 71 (2023)

R Core Team, R., Team, R.C., et al.: R: A Language and Environment for Statistical Computing; R Foundation for Statistical Computing: Vienna, Austria, 2021 (2022)

Saffari, S.E., Adnan, R.: Zero-inflated poisson regression models with right censored count data. Matematika, 21–29 (2011)

Tous, J., Chiquet, J., Deacon, A.E., Fontrodona-Eslava, A., Fraser, D.F., Magurran, A.E.: A jsdm with zero-inflation to increase the ecological relevance of analyses of species distribution data. bioRxiv, 2025–07 (2025)

Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society Series B: Statistical Methodology **58**(1), 267–288 (1996)

Theil, S., Rifa, E.: Utopia: An automatically updated, complete and consistent its reference database. In: 5. Colloque de Génomique Environnementale, p. (2019)

Theil, S., Rifa, E.: ranomaly: Amplicon workflow for microbial community analysis. F1000Research **10**, 7 (2021)

Verdier-Metz, I., Delbès, C., Bouchon, M., Rifa, E., Theil, S., Chaucheyras-Durand, F., Chevaux, E., Dunière, L., Chassard, C.: Dietary live yeast supplementation influence on cow's milk, teat and bedding microbiota in a grass-diet dairy system. Microorganisms **11**(3), 673 (2023)

Warton, D.I., Blanchet, F.G., O'Hara, R.B., Ovaskainen, O., Taskinen, S., Walker, S.C., Hui, F.K.: So many variables: joint modeling in community ecology. Trends in ecology & evolution **30**(12), 766–779 (2015)

Weston, J., Elisseeff, A., Schölkopf, B., Tipping, M.: Use of the zero-norm with linear models and kernel methods. Journal of machine learning research **3**(Mar), 1439–1461 (2003)

Ypma, J.: Introduction to nloptr: an r interface to nlopt. R Package **2**, 1–15 (2014)

Zou, H.: The adaptive lasso and its oracle properties. Journal of the American statistical association **101**(476), 1418–1429 (2006)

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- Supplementary.pdf