Sparse inference in Poisson log-normal model by approximating the L_0-norm

Togo Jean Yves KIOYE[1], Paul-Marie GROLLEMUND[12], Jocelyn CHAUVET[34], Christophe CHASSARD[1]

[1] Unité Mixte de Recherche sur le Fromage (UMRF), Aurillac, France
[2] Laboratoire de Mathématiques Blaise Pascal (LMBP), Clermont Ferrand, France
[3] Centre de recherche de l'ICES, Roche-sur-Yon, France
[4] Laboratoire Angevin de Recherche en Ingénierie des Systèmes (LARIS), Angers, France

E-mail for correspondence: `togo_jean_yves.kioye@uca.fr`

Abstract: Variable selection methods are essential in statistical modelling to improve interpretability by identifying the most relevant predictors. This article focuses on the Poisson Log Normal (PLN) model, widely used for analysing multivariate count data in fields like ecology and agronomy. Recent advancements, such as those by Chiquet et al. (2021), highlight sparse network inference using the evidence lower bound of the likelihood combined with an L_1-penalty on the precision matrix. This paper introduces an alternative approach based on the Smooth Information Criterion (SIC, O'Neill and Burke (2023)), which smoothly approximates the L_0-penalty, removing the need for cross-validation to tune regularisation parameters. The study targets the coefficient matrix of the PLN model, proposing an inference procedure for effective variable selection. The method integrates the SIC penalisation algorithm with the PLN model fitting algorithm, a variational EM algorithm. To support our proposal, we provide theoretical results and insights about the penalisation method, we perform simulation studies to assess the method, which is also applied on real datasets from a study of microbial communities in milk production.

Keywords: Variable selection; Multivariate count data; Variational EM algorithm; Information criteria; Bayes estimate; Microbial communities.

1 Introduction

Multivariate count data analysis plays a central role in statistical modeling, providing valuable insights across various fields. Applications include the

This paper was published as a part of the proceedings of the 39th International Workshop on Statistical Modelling (IWSM), Limerick, Ireland, 13–18 July 2025.

study of the simultaneous abundance of different species in ecological communities, modelling the occurrence of multiple diseases within a population, and the analysis of the relative abundances of microorganisms in specific ecosystems (Chiquet et al. (2021)). However, multivariate count data analysis presents particular challenges, especially in managing dependencies between variables and addressing overdispersion. In recent years, several advanced statistical approaches have been developed to tackle these challenges, including multivariate Poisson regression (Chiquet et al. (2021)), multivariate negative binomial regression (Shi and Valdez (2014)), and copula-based models (Nikoloulopoulos and Karlis (2009)).

In this work, we focus on the PLN model, which offers great flexibility to consider additional extensions (Aitchison and Ho (1989)). This model relies on a latent layer, which complicates its inference. The variational approximation is an efficient method to overcome this difficulty, and the model is thus fitted using the Variational Expectation-Maximization (VEM) algorithm. However, it is also possible to perform inference for this model using Monte Carlo-based approaches or the Laplace method. The PLN model allows for the exploration of network structures through the estimated coefficients of the precision matrix in practical settings. It also enables examination of the impact of dependent variables on count data by estimating the regression coefficients. In this context, it is important to develop an estimation procedure alongside a variable selection method to ensure that the fitted model retains only those coefficients that are significantly different from zero.

The use of L_1-type penalties, such as the Lasso method (Tibshirani (1996)), provides a classical approach to address this trade-off. However, these methods require the tuning of a regularization parameter, typically accomplished through computationally intensive procedures such as cross-validation. Also the L_1-penalty is often suboptimal to accurately estimate the support of nonzero components, thus introducing bias in parameter estimates.

Recently, O'Neill and Burke (2023) introduced a new penalization approach that aims to smoothly approximate the norm L_0, enabling efficient variable selection without requiring calibration of the regularization parameter through procedures such as cross-validation. This approach, termed the Smooth Information Criterion (SIC), has been applied to distributional regression models.

In this paper, we propose to adapt this approach to the PLN model to obtain a more parsimonious estimate of the regression coefficients and reduce the computational cost associated with tuning the regularization parameter.

The following section presents the PLN model, details the method for estimating its parameters, and describes the penalization function used to obtain sparse estimates of the regression coefficients.

2 Inference

Poisson Log Normal (PLN) model. The multivariate PLN model was introduced by Aitchison and Ho (1989) to model joint count data using environmental factors, while accounting for the dependency structure between counts. Each observed count vector $\mathbf{y}_i = (y_{i1}, \ldots, y_{ip}) \in \mathbb{N}^p$ is associated with a latent vector $\mathbf{z}_i = (z_{i1}, \ldots, z_{ip}) \in \mathbb{R}^p$, which follows a multivariate Gaussian distribution with covariance matrix $\boldsymbol{\Sigma}$. The observed counts y_{ij} are assumed to follow a Poisson distribution with parameter $\exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + z_{ij})$, where o_{ij} represents the sampling effort for species j at site i, and $\mathbf{x}_i \in \mathbb{R}^d$ is the covariate vector for site i. The model is formulated as follows:

$$\begin{aligned} y_{ij} \mid z_{ij} &\sim \mathcal{P}\big(\exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + z_{ij})\big) && \text{(observed space)} \\ \mathbf{z}_i &\sim N_p\big(\mathbf{0}, \boldsymbol{\Sigma}\big) && \text{(latent space)}, \end{aligned} \tag{1}$$

In this model, $\boldsymbol{\beta}_j = (\beta_{1j}, \ldots, \beta_{dj})^\top$ encodes the effect of the d environmental covariates on the abundance of species j. Let $\mathbf{Y} \in \mathbb{N}^{n \times p}$ denote the count matrix, $\mathbf{X} \in \mathbb{R}^{n \times d}$ the matrix of environmental variables, $\mathbf{O} \in \mathbb{R}^{n \times p}$ the offset matrix (or sampling effort), and $\mathbf{B} \in \mathbb{R}^{d \times p}$ the regression coefficient matrix. We also define $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\Sigma})$ as the set of parameters for the PLN model, and $\mathbf{Z} \in \mathbb{R}^{n \times p}$ as the matrix of latent vectors $\mathbf{z}_i$.

Estimation methods. Estimation of the parameters of the PLN model requires the use of the EM algorithm. However, the E-step involves computing the conditional expectation under the distribution $p_{\boldsymbol{\theta}}(\mathbf{z}_i \mid \mathbf{y}_i)$, which is intractable. To circumvent this issue, Chiquet et al. (2021) propose using a variational inference strategy to approximate $p_{\boldsymbol{\theta}}(\mathbf{z}_i \mid \mathbf{y}_i)$ by a distribution $q_{\boldsymbol{\psi}}(\mathbf{z}_i)$ that minimizes the Kullback-Leibler (KL) divergence between $q_{\boldsymbol{\psi}}(\mathbf{z}_i)$ and $p_{\boldsymbol{\theta}}(\mathbf{z}_i \mid \mathbf{y}_i)$. This strategy leads to the maximization of a lower bound of the log-likelihood known as the ELBO (*Evidence Lower BOund*):

$$J(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\psi}) = \log(p_{\boldsymbol{\theta}}(\mathbf{Y})) - \mathrm{KL}\Big[q_{\boldsymbol{\psi}}(\mathbf{Z}|\mathbf{Y}) \mid\mid P_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})\Big]. \tag{2}$$

Chiquet et al. (2021) considered the mean-field approach in which:

$$q_{\boldsymbol{\psi}}(\mathbf{Z}|\mathbf{Y}) = \prod_i^n q_{\boldsymbol{\psi}}(\mathbf{z}_i) = \prod_i^n \mathcal{N}(\mathbf{z}_i; \mathbf{m}_i, \mathrm{diag}(\mathbf{s}_i^2))$$

The set of variational parameters is collected in the vector $\boldsymbol{\psi} = (\mathbf{M}, \mathbf{S})$, where $\mathbf{M} = (\mathbf{m}_1^\top, \ldots, \mathbf{m}_n^\top)^\top$ and $\mathbf{S} = (\mathbf{s}_1^{2^\top}, \ldots, \mathbf{s}_n^{2^\top})^\top$. We can express the

ELBO (2) in matrix form as follows:

$$\begin{aligned} J(\mathbf{Y}, \boldsymbol{\theta}, \boldsymbol{\psi}) \;=\; & \mathbf{1}_n^\top \left(\mathbf{Y} \odot (\mathbf{O} + \mathbf{X}\mathbf{B} + \mathbf{M}) - \mathbf{A}\right) \mathbf{1}_p - \sum_{i=1}^{n}\sum_{j=1}^{p} \log(y_{ij}!) \\ & + \frac{n}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}\mathrm{trace}\left(\mathbf{M}\boldsymbol{\Omega}\mathbf{M}^\top\right) - \frac{1}{2}\mathrm{trace}\left(\hat{\mathbf{S}}\boldsymbol{\Omega}\right) \\ & + \frac{1}{2}\mathbf{1}_n \log(\mathbf{S})\mathbf{1}_p + \frac{np}{2}, \end{aligned} \tag{3}$$

where $\odot$ is the Hadamard product, $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ is the precision matrix, $|\cdot|$ its determinant, $\hat{\mathbf{S}} = \sum_{i=1}^{n} \mathbf{S}_i$, and $\mathbf{A}$ is the $n \times p$ matrix of expected count with entries $a_{ij} = \exp(o_{ij} + \mu_{ij} + m_{ij} + s_{ij}^2/2)$.

Variables selection. We consider the ELBO of the PLN model (2), with a focus on the parameter of interest, namely the regression coefficient matrix $\mathbf{B}$, for which we seek to obtain a sparse estimate. To this end, we add a penalty ϕ_ε to the ELBO to constrain certain regression coefficients to be exactly equal to 0. The penalization function ϕ_ε was introduced by O'Neill and Burke (2023) with the goal of approximating the L_0 norm.

The corresponding penalized objective function is then

$$J_{\mathrm{pen}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\psi}) \;= J(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\psi}) - \frac{\lambda}{2}\Big[\phi_\varepsilon(\mathbf{B}) + k\Big], \tag{4}$$

where $\lambda = \log(n)$, k is the number of parameters that are not to be penalized, and

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

Note that the function ϕ_ε is differentiable for $\varepsilon > 0$, and we have: $\lim_{\varepsilon \to 0} \phi_\varepsilon(x) = \|x\|_0$.

Optimization algorithm (SICPLN). To sparsely estimate the support of the entries in $\mathbf{B}$, we incorporate a Variational Expectation-Maximization (VEM) algorithm into the ε-telescoping framework proposed by O'Neill and Burke (2023). This approach eliminates the calibration a fixed value for ε, which controls the approximation of the L_0 norm, and enables stable estimation of $\mathbf{B}$.

The ε-telescoping strategy consists of defining an exponentially decreasing sequence of ε values approaching zero. For each value in the sequence, $J_{\mathrm{pen}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\psi})$ is optimized, using the solution from the previous step as the initialization for the next, thereby improving convergence and stability.

In the VEM algorithm, the classical E-step of the EM algorithm is replaced with a variational step, in which we estimate the variational parameters $\boldsymbol{\psi}$ by maximizing the penalized ELBO. In the M-step, we update the model parameters $\boldsymbol{\theta}$ by also maximizing the same penalized ELBO. The optimiza-

tion problems for both steps are formulated as follows:

$$\boldsymbol{\psi}^{(t+1)} = \arg\max_{\boldsymbol{\psi}} J_{\text{pen}}(\mathbf{Y}; \boldsymbol{\theta}^{(t)}, \boldsymbol{\psi}) \qquad \textbf{(VE step)} \tag{5}$$

$$\boldsymbol{\theta}^{(t+1)} = \arg\max_{\boldsymbol{\theta}} J_{\text{pen}}(\mathbf{Y}; \boldsymbol{\theta}, \boldsymbol{\psi}^{(t+1)}) \qquad \textbf{(VM step)} \tag{6}$$

3 Simulation study

Compared methods. To provide a context for the performance of the proposed method compared to existing methods, we apply the following methods in this simulation study.

- GLMNET: a univariate Poisson regression model using Lasso regularization (implemented in the R `glmnet` package). As this modeling is univariate, we duplicate it separately for each column of the count matrix $\mathbf{Y}$. While this approach disregards the dependence between columns of $\mathbf{Y}$, at least it includes a variable selection procedure.
- PLN: the standard Poisson Log Normal model implemented in the R package `PLNmodels`. Unlike the GLMNET approach, this is a multivariate approach that takes the dependence between columns of $\mathbf{Y}$ into account, but it does not perform variable selection.

Numerical results for one configuration. To straightforwardly present numerical results of competing methods, we propose highlighting the estimated coefficients of each method in the configuration where: $n = 10000$, $p = 4$, and the structure of $\boldsymbol{\Sigma}$ is the Full case. Table 1 provides the results of coefficients estimation. We observe that SICPLN outperforms GLMNET for accurate coefficient estimation, probably because GLMNET uses a LASSO penalty, which can introduce bias into the estimates. In addition, unlike GLMNET, SICPLN exploits dependency structures between multidimensional responses to improve estimates. Finally, with SICPLN, coefficients associated with irrelevant variables are estimated as exactly zero, which is not always the case with GLMNET (see species 4).

4 Extensions and other results

During the presentation, we will introduce the formulation of the SIC as a prior distribution in the Bayesian framework, as well as additional numerical results based on simulated and real data (see Kioye et al. (2024) for details).

Acknowledgments: Special Thanks to Program DATA and IRC-SAE related to the ISITE/CAP2025 funding program of Clermont Auvergne University.

TABLE 1. Estimated coefficients with PLN, GLMNET and SICPLN.

Species	Estimation method	$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	$\mathbf{x_5}$	$\mathbf{x_6}$
Species 1	True coefficient	0	1	1	1	1	0
	PLN	0.08	1.04	1.09	1.07	1.10	0.09
	GLMNET	0	0.91	0.99	0.93	0.96	0
	SICPLN	0	0.95	1	0.98	0.92	0
Species 2	True coefficient	0.5	0	0	1	1	0
	PLN	0.55	0.07	0.15	1.04	0.99	0.13
	GLMNET	0.43	0	0	0.98	0.87	0
	SICPLN	0.47	0	0	0.98	0.92	0
Species 3	True coefficient	1	0.5	0.5	1	1	0
	PLN	1.10	0.58	0.54	1.05	1.06	0.10
	GLMNET	0.97	0.39	0.43	1.05	0.84	0
	SICPLN	1	0.48	0.44	0.96	0.97	0
Species 4	True coefficient	1	1	0	0	0.5	0
	PLN	0.91	0.95	0.05	0.10	0.52	0.02
	GLMNET	0.64	1.14	0.10	-0.14	0.14	-0.21
	SICPLN	0.94	0.98	0	0	0.54	0

References

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

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

Kioye, T. J. Y., Grollemund, P. M., Chauvet, J. et al (2024). Sparse inference in Poisson Log-Normal model by approximating the L0-norm. *arXiv preprint arXiv:2403.17087*.

Nikoloulopoulos, A. K. and Karlis, D. (2009). Modeling multivariate count data using copulas. *ommunications in Statistics-Simulation and Computation*, **39(1)**, 172 – 187.

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

Shi, P. and Valdez, E. A. (2014). Multivariate negative binomial models for insurance claim counts. *Insurance: Mathematics and Economics*, **55**, 18 – 29.

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