Composite likelihood inference for the Poisson log-normal model

Julien Stoehr[1, 2] and Stéphane Robin[3]

[1]CEREMADE, Université Paris-Dauphine, Université PSL, CNRS, 75016 Paris, France.
[2]Université Paris-Saclay, INRAE, AgroParisTech, UMR MIA Paris-Saclay, 91120 Palaiseau, France.
[3]Sorbonne Université, CNRS, Laboratoire de Probabilités, Statistique et Modélisation, 75005 Paris, France.

Abstract

The Poisson log-normal model is a latent variable model that provides a generic framework for the analysis of multivariate count data. Inferring its parameters can be a daunting task since the conditional distribution of the latent variables given the observed ones is intractable. For this model, variational approaches are the golden standard solution as they prove to be computationally efficient but lack theoretical guarantees on the estimates. Sampling based solutions are quite the opposite. Starting from already available variational approximations, we define a first Monte Carlo EM algorithm to obtain maximum likelihood estimators. We then extend this algorithm to the case of a composite likelihood in order to be able to handle higher dimensional count data.

Keywords. Composite likelihood, importance sampling, Monte Carlo EM algorithm, multivariate count data

1 Introduction

Latent variable models have become a prominent tool for statistical modeling in almost all application domains (life sciences, economy, industry, medicine, environmental sciences, to name a few). From a general point of view, a latent variable model aims at describing the variation of response variables $\mathbf{Y}$ conditionnaly to unobserved (*i.e.*, *latent*) variables $\mathbf{Z}$ (plus, possibly, observed covariates $\mathbf{x}$). We denote by $\boldsymbol{\theta}$ the set of unknown parameters ruling the distributions of both the latent variables $\mathbf{Z}$ and the observed variables $\mathbf{Y}$.

1.1 The Poisson log-normal model

All along this paper we focus on the multivariate Poisson log-normal model [PLN: Aitchison and Ho, 1989] that is a generic model for the joint distribution of count data, accounting for covariates. As a typical example, one may consider n sites (indexed by $1 \leq i \leq n$) and p animal species (indexed by $1 \leq j \leq p$) and denote by Y_{ij} the number of individuals from species j observed in site i. We will further assume that a d-dimensional vector of covariates (descriptors) $\mathbf{x}_i$ describing the environmental conditions is recorded for each site i. The PLN

model assumes that, for each site i, we can relate the p-dimensional observations to a latent p-dimensional Gaussian random vector $\mathbf{Z}_i = (Z_{ij})_{1\leq j\leq p}$ with covariance matrix $\mathbf{\Sigma}$. Namely the observed counts Y_{ij} conditionally on Z_i are independent and distributed according to a Poisson distribution with rate $\exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + Z_{ij})$:

$$\begin{aligned} \{\mathbf{Z}_i\}_{1\leq i\leq n} \ i.i.d. : & \qquad \mathbf{Z}_i \sim \mathcal{N}(0, \mathbf{\Sigma}); \\ \{Y_{ij}\}_{1\leq i\leq n, 1\leq j\leq p} \text{ independent} \mid \{\mathbf{Z}_i\}_{1\leq i\leq n} : & \quad Y_{ij} \mid Z_{ij} \sim \mathcal{P}\left(\exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + Z_{ij})\right), \end{aligned} \tag{1}$$

where $\boldsymbol{\beta}_j = [\beta_{1j} \dots \beta_{dj}]^\top \in \mathbb{R}^d$ is a species-specific vector of regression parameters and where the covariance matrix $\mathbf{\Sigma} = [\Sigma_{jk}]_{1\leq j,k\leq p}$ encodes the dispersion and dependency structure between the counts from a same site.

In this model, the regression coefficient $\beta_{\ell j}$ $(1 \leq \ell \leq d,\ 1 \leq j \leq p)$ is interpreted as the effect of covariate ℓ of the mean abundance of species j, whereas Σ_{jk} $(1 \leq j, k \leq p)$ is the latent covariance coefficient between the abundances of species j and k. Gathering the vectors of regression coefficients into a $d \times p$ matrix $\boldsymbol{B} = [\boldsymbol{\beta}_1 \cdots \boldsymbol{\beta}_p]$, the parameter of the PLN model becomes $\boldsymbol{\theta} = (\boldsymbol{B}, \mathbf{\Sigma})$.

1.2 Maximum likelihood and the EM algorithm

Maximum likelihood inference for latent variable models is usually not straightforward because the likelihood of the data $p_{\boldsymbol{\theta}}(\mathbf{Y})$ results from an high-dimensional integration over the unobserved $\mathbf{Z}$: $p_{\boldsymbol{\theta}}(\mathbf{Y}) = \int p_{\boldsymbol{\theta}}(\mathbf{Y} \mid \mathbf{Z}) p_{\boldsymbol{\theta}}(\mathbf{Z}) \mathrm{d}\mathbf{Z}$. This problem can be circumvented using the celebrated expectation-maximization (EM) algorithm [Dempster et al., 1977], which relies on the evaluation of certain moments of the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ during the E step of the algorithm. Importantly, the resulting estimate inherits the general properties of maximum likelihood estimates; in particular, its asymptotic variance can be evaluated using side products of the EM algorithm [Louis, 1982].

Unfortunately, the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ can become intractable even for moderately complex models. Many strategies have been proposed to deal with this complex conditional distribution. Thesse can be cast into two main approaches: either approximate it or sample from it. The former approach can typically rely on variational approximations, whereas the latter gave rise to a series of stochastic version of EM.

Variational approximation of EM. Variational approximations have received a lot of attention in the last decade because of their computational efficiency and their (reasonable) ease of implementation. Broadly speaking, they rely on the determination of an approximate distribution $q(\mathbf{Z}) = q_{\boldsymbol{\theta},\mathbf{Y}}(\mathbf{Z}) \simeq p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ chosen within a certain class of distributions and optimal according to a given divergence measure, [see e.g. Jaakkola and Jordan, 2000, Wainwright and Jordan, 2008, Blei et al., 2017]. Still, although variational EM (VEM) algorithms usually have good empirical performances, the resulting estimate is not guaranteed, in general, to have any desirable statistical property, such as consistency or asymptotic normality, and no measure of uncertainty or significance test can be provided to practitioners in the end.

A VEM algorithm has been developped by Chiquet et al. [2018, 2019] for the inference of the PLN model, which is implemented in the R package `PLNmodels` available from the CRAN (`cran.r-project.org`).

1.3 Monte-Carlo approximations of EM

An alternative way to address the intractability of the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ is to resort to Monte Carlo EM (MCEM) algorithm [Wei and Tanner, 1990]. The latter is an

extension of the standard EM algorithm specifically designed to deal with E-step which cannot be performed analytically or is computationally cumbersome. In such situations, the E-step can be replaced with a Monte Carlo sampling method, such as important sampling [Booth and Hobert, 1999, Levine and Casella, 2001] or Markov Chain Monte Carlo [McCulloch, 1997, Fort and Moulines, 2003], to provide estimates of the needed conditional moments. However, all these methods require to sample from the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$, which is time consuming and can be poorly efficient, especially when the dimension of $\mathbf{Z}$ is large.

Composite likelihood. Composite likelihoods refer to a broad variety of approaches where the likelihood of the data is replaced with a modified version of it. Such approaches trace back to Besag [1974] and Lindsay [1988] proposed an (almost) general definition, which consists in splitting the data into a series of (possibly overlapping) blocks and to replace the log-likelihood with a weighted sum of log-likelihoods corresponding to each block of data. Splitting the dataset into blocks allows dealing with large dimension data and/or complex dependency structure, while keeping desirable statistical properties [see, e.g. Varin et al., 2011]. We will see that, in presence of latent variables, the classical EM algorithm can be extended to deal with the composite likelihood, as well as its variational or Monte-Carlo counterparts.

1.4 Contribution

Our contribution is three-fold. First, we introduce a Monte Carlo EM for the inference of the Poisson log-normal model. Importantly, the proposed algorithm gives access to the asymptotic variance of the parameters and, consequently, to confidence intervals and tests of the parameters, as opposed to the available VEM algorithm. Our second contribution is the use of composite likelihood inference, using blocks of species. This allows to consider latent variables living in a space with smaller dimension (actually, the number of species in each block), so to make importance sampling efficient. To the best of our knowledge, this results in the first EM algorithm combining importance sampling and composite likelihood to address inference on such a model. Third, we show how to define a mixture proposal distribution that ensures finite variance on bounded test functions and controls the efficiency of the importance sampling steps.

Outline. In Section 2, after reviewing a set of useful notions and properties about the EM algorithm, we briefly recall the principle of the variational approximation of EM. Then we describe how the methods and properties associated with the EM algorithm can be extended to composite likelihood inference. In Section 3, we introduce a Monte Carlo EM algorithm, based on importance sampling (IS), which applies to both likelihood and composite likelihood inference. The construction of the blocks for the latter is discussed at the beginning of Section 4. Then we illustrate the use of our algorithm, first on synthetic data, then to analyse fish abundances in the Barents see. Throughout the paper, the Poisson log-normal serves as a common thread for the proposed methodology.

2 A reminder on inference for latent variable models

2.1 Maximum likelihood inference using EM

EM algorithm. The inference of any incomplete data model can be tackled using the Expectation-Maximization (EM) algorithm [Dempster et al., 1977], which relies on the de-

composition of the log-likelihood of the observed data $\mathbf{Y}$:

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\theta}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}] + \mathcal{H}_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}) \tag{2}$$

where $\mathcal{H}_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ stands for the conditional entropy of the unobserved variables $\mathbf{Z}$ given the observed ones:

$$\mathcal{H}_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}) = -\int_{\mathbb{R}^p} p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{Y}) \log p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{Y}) \mathrm{d}\mathbf{z}$$

and $\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})$ is the so-called complete log-likelihood.

The EM algorithm requires the evaluation of the conditional expectation of the complete log-likelihood (first term of (2)) using the current estimates $\boldsymbol{\theta}^{(h)}$ of the model parameters, defining for $\boldsymbol{\theta} \in \Theta$:

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) = \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}]. \tag{3}$$

The formal EM algorithm is given by Algorithm 1. By iterating these two steps, the algorithm generates a sequence that converges under regularity conditions to the maximum likelihood (ML) estimator [Wu, 1983, Boyles, 1983].

Algorithm 1: EM

repeat

E step: compute the moments of the conditional distribution $p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Z} \mid \mathbf{Y})$ required to evaluate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)})$;

M step: update the parameter estimate as:

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) ;$$

until *convergence*;

Asymptotic variance of the maximum likelihood estimates. Because EM is a maximum likelihood method, under fairly general regularity conditions, the resulting estimators are asymptically unbiased, Gaussian, with known asymptotic variance. Evaluating asymptotic precisions of parameter estimates then requires the computation of the Fisher information matrix defined for all $\boldsymbol{\theta} \in \Theta$ as

$$I(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}) \{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})\}^{\top}]$$

where the expectation is taken with respect to $\mathbf{Y}$. It is somewhat usual that the latter is intractable and therefore need to be estimated. Straightforward solution is to resort to a simple sample mean estimate based on a n-sample of observations $(\mathbf{Y}_1, \ldots, \mathbf{Y}_n)$, namely one uses the observed Fisher information matrix

$$\widehat{I}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i) \{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)\}^{\top}. \tag{4}$$

Since integrating over the latent space might be cumbersome, Louis [1982] provides a reformulation of the score function for incomplete data models, which makes the exact computation of $\widehat{I}(\boldsymbol{\theta})$ possible on some models using side information resulting from EM inference. More specifically, he shows that

$$\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}]. \tag{5}$$

Regular models admits an alternative definition of the Fisher information matrix, namely

$$I(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})\big].$$

Similarly to the score version, one can derive a sample mean estimate where the computation of the Hessian is made possible, for some models, using the second Louis' formula

$$\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \,\big|\, \mathbf{Y}\big] + \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}]. \tag{6}$$

Combining (5) and (6), the resulting estimator writes as

$$\widehat{I}(\boldsymbol{\theta}) - \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i) + \nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i)\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i)\}^\top \,\big|\, \mathbf{Y}_i\big]$$

Interestingly, this estimator forms a special instance of control variates estimator. Indeed, the term in the conditional expectation integrates to 0 with respect to the joint distribution $p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})$ since it can be written as $\nabla^2_{\boldsymbol{\theta}} p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z})/p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z})$. The practical benefits of such an estimator may nonetheless be limited. To ensure a variance reduction compared to the observed Fisher information matrix, it requires introducing and tuning a control coefficient to weight the above sum, a task that might not be done without introducing biased or a computational burden.

2.2 Variational EM

One main limitation of the EM algorithm lies in the determination of the conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ or at least, of the evaluation of the moments of this distribution that are needed to evaluate $Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)})$ or the observed Fisher information matrix. In many situations, this distribution is intractable and so are the needed moments. It is also quite frequent that sampling from $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ in an efficient manner is not trivial, so that no simple Monte-Carlo alternative exists. The Poisson log-normal model is one of theses inconvenient models, as, even in the one dimensional case, the conditional distribution of the latent Z given the observed count Y has no close form.

Variational inference circumvent this problem, by replacing $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$ with an approximate distribution $q_{\boldsymbol{\psi}}(\mathbf{Z})$, chosen within a certain class of distributions, parametrized with $\boldsymbol{\psi}$. The VEM algorithm then amounts at maximizing a lower bound of the log-likelihood, defined as

$$\log p_{\boldsymbol{\theta}}(\mathbf{Y}) - \mathrm{KL}\left[q_{\boldsymbol{\psi}}(\mathbf{Z}) \| p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})\right] = \mathbb{E}_{q_{\boldsymbol{\psi}}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})] + \mathcal{H}_{q_{\boldsymbol{\psi}}}(\mathbf{Z})$$

where KL stands for the Kullback-Leibler divergence between $q_{\boldsymbol{\psi}}$ and $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y})$:

$$\mathrm{KL}\left[q_{\boldsymbol{\psi}} \| p_{\boldsymbol{\theta}}(\cdot \mid \mathbf{Y})\right] = \int_{\mathbb{R}^p} q_{\boldsymbol{\psi}}(\mathbf{z}) \log \frac{q_{\boldsymbol{\psi}}(\mathbf{z})}{p_{\boldsymbol{\theta}}(\mathbf{z} \mid \mathbf{Y})} \mathrm{d}\mathbf{z}.$$

The VEM alternates the update of the *variational* parameter $\boldsymbol{\psi}$, with this of the model parameter $\boldsymbol{\theta}$. The reader may refer to Jaakkola [2001] or Wainwright and Jordan [2008] for a general introduction.

PLN model. The complete likelihood associated to the PLN model writes (up to constants with respect to the parameters)

$$\begin{aligned}
\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) &= \log p_{\boldsymbol{\Sigma}}(\mathbf{Z}) + \log p_{\boldsymbol{B}}(\mathbf{Y} \mid \mathbf{Z}) \\
&= -\frac{n}{2} \log |\boldsymbol{\Sigma}| - \frac{1}{2} \sum_{i=1}^{n} \mathbf{Z}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z}_i \\
&\quad + \sum_{i=1}^{n} \sum_{j=1}^{p} \left\{ Y_{ij}(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + Z_{ij}) - \exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + Z_{ij}) \right\}.
\end{aligned}$$

Therefore, the E-step of Algorithm 1 requires to evaluate , for all $1 \leq i \leq n$ and $1 \leq j \leq p$, the conditional moments :

$$\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[Z_{ij} \mid Y_{ij}], \qquad \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\exp(Z_{ij}) \mid Y_{ij}] \qquad \text{and} \qquad \mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[\mathbf{Z}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z}_i \,\middle|\, \mathbf{Y}_i\right]. \tag{7}$$

Unfortunately, these conditional moment can not be evaluated in close form and their numerical evaluation computationally to heavy, even for a small number of species, which hampers the use of EM. Chiquet et al. [2018, 2019] developped a variationnal EM algorithm for the PLN model, using a Gaussian approximate distribution q. Because the sites $1 \leq i \leq n$ are supposed to be independent, each conditional distribution $p_{\boldsymbol{\theta}}(\mathbf{Z}_i \mid \mathbf{Y}_i)$ is approximated with a specific normal distribution $\mathcal{N}(\mathbf{m}_i, \mathbf{S}_i)$ with diagonal covariance matrix $\mathbf{S}_i$. The variational parameter is hence $\boldsymbol{\psi} = \{(\mathbf{m}_i, \mathbf{S}_i)\}_{1 \leq i \leq n}$.

Variance of the variational estimates. An important feature of variational inference is that the resulting estimator of $\boldsymbol{\theta}$ does not enjoy the same properties as the regular maximum-likelihood estimator; in particular, no general results ensures its asymptotic normality, nor even its consistency. In summary, the computational efficiency of variational approximation comes at the price of the loss of general statistical guarantees. In particular, the (asymptotic) variance of the variational estimator is not known. Still, this variance can be estimated using jackknife [Efron and Stein, 1981], which requires to run the VEM algorithm n times. This option is implemented in the `PNmodels` R package.

2.3 Composite likelihood inference

Composite likelihood is another admissible contrast for grounded statistical inference that may prove useful especially when dealing with a relatively high-dimensional latent variable $\mathbf{Z}$. The observed variables $\mathbf{Y}$ are distributed into C (possibly overlapping) blocks $\{\mathcal{C}_b\}_{1 \leq b \leq C}$ of data and the composite log-likelihood is defined as

$$c\ell_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{b=1}^{C} \lambda_b \, \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)}), \tag{8}$$

where λ_b is the weight associated with block $\mathcal{C}_b$ and $\mathbf{Y}^{(b)}$ is the dataset reduced to block $\mathcal{C}_b$. Most of the time, all blocks are chosen to have the same size k, pairwise composite likelihood referring to $k = 2$.

The maximum composite likelihood (MCL) estimate is then defined as $\widehat{\boldsymbol{\theta}}_{c\ell} = \arg\max_{\boldsymbol{\theta}} c\ell_{\boldsymbol{\theta}}(\mathbf{Y})$. Importantly, $\widehat{\boldsymbol{\theta}}_{c\ell}$ enjoys properties similar to these of maximum likelihood estimator, namely consistency and asymptotic normality [see Varin et al., 2011, for a general review], although, possibly, with higher asymptotic variance [Zhao and Joe, 2005, Xu et al., 2016]. These properties actually derive from the more general properties of M-estimators [van der Vaart, 1998, Chap. 5].

Composite likelihood inference using EM. Composite likelihood can be used for the inference of incomplete data model, using an adaptation of the EM Algorithm 1. Indeed, one may observe that the decomposition given in Equation (2) holds for each block of data $\mathbf{Y}^{(b)}$, so Equation (8) can be rephrased as

$$c\ell_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{b=1}^{C} \lambda_b \mathbb{E}_{\boldsymbol{\theta}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)}, \mathbf{Z}) \,\middle|\, \mathbf{Y}^{(b)}\right] + \sum_{b=1}^{C} \lambda_b \mathcal{H}_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}^{(b)}). \tag{9}$$

However, this formulation still requires to deal with the conditional distribution of the whole latent variables $p_{\boldsymbol{\theta}}(\mathbf{Z} \mid \mathbf{Y}^{(b)})$. A computational advantage is only obtained when also distributing

the latent variables $\mathbf{Z}$ into C blocks of small size, so that Equation (8) becomes

$$c\ell_{\boldsymbol{\theta}}(\mathbf{Y}) = \sum_{b=1}^{C} \lambda_b \mathbb{E}_{\boldsymbol{\theta}}\Big[\log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \,\Big|\, \mathbf{Y}^{(b)}\Big] + \sum_{b=1}^{C} \lambda_b \mathcal{H}_{\boldsymbol{\theta}}(\mathbf{Z}^{(b)} \mid \mathbf{Y}^{(b)}), \tag{10}$$

where we only need to deal with the conditional distribution of a subset of random variables: $p_{\boldsymbol{\theta}}(\mathbf{Z}^{(b)} \mid \mathbf{Y}^{(b)})$. One may then design the CL-EM Algorithm 2.

Algorithm 2: CL-EM

repeat

CL-E step: compute the moments of each of the C conditional distribution $p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Z}^{(b)} \mid \mathbf{Y}^{(b)})$ required to evaluate

$$Q_{c\ell}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) = \sum_{b=1}^{C} \lambda_b \mathbb{E}_{\boldsymbol{\theta}^{(h)}}\Big[\log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \,\Big|\, \mathbf{Y}^{(b)}\Big];$$

CL-M step: update the parameter estimate as:

$$\boldsymbol{\theta}^{(h+1)} = \arg\max_{\boldsymbol{\theta}} Q_{c\ell}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)});$$

until *convergence*;

Algorithm 2 enjoys the same general property as the EM Algorithm 1. The proof follows the same line as this for EM and is given in Appendix A.1.1.

Proposition 1. *Using Algorithm 2 yields a sequence* $(\boldsymbol{\theta}^{(h)})_{h\in\mathbb{N}}$ *such that* $c\ell_{\boldsymbol{\theta}^{(h+1)}}(\mathbf{Y}) \geq c\ell_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y})$.

PLN model. In the context of the PLN model, the blocks will consists of subsets of k species, that is $\mathcal{C}_b \subset \{1, \ldots, p\}$ and $\mathbf{Y}^{(b)} = \{Y_{ij}\}_{1\leq i\leq n, j\in\mathcal{C}_b}$. We shall consider the same blocks of species for $\mathbf{Z}$ as for $\mathbf{Y}$, that is $\mathbf{Z}^{(b)} = \{Z_{ij}\}_{1\leq i\leq n, j\in\mathcal{C}_b}$. Obviously, the conditional moments to be evaluated at the CL-E step are the same than these given in (7), replacing $\mathbf{Y}$ (resp. $\mathbf{Z}$) with $\mathbf{Y}^{(b)}$ (resp. $\mathbf{Z}^{(b)}$) for each block $\mathcal{C}_b$, but with a further simplification on the parameter. The complete likelihood $p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})$ of block $\mathcal{C}_b$ does not depend on all the elements of $\boldsymbol{\theta}$. More specifically, it only involves $\boldsymbol{\theta}^{(b)} = (\boldsymbol{B}^{(b)}, \boldsymbol{\Sigma}^{(b)})$ where

$$\boldsymbol{B}^{(b)} = [\boldsymbol{\beta}_j]_{j\in\mathcal{C}_b}, \quad \text{and} \quad \boldsymbol{\Sigma}^{(b)} = [\boldsymbol{\Sigma}_{jk}]_{(j,k)\in\mathcal{C}_b}.$$

We thus need to estimate on each block $1 \leq b \leq C$, each site $1 \leq i \leq n$ and each specie $j \in \mathcal{C}_b$

$$\mathbb{E}_{\boldsymbol{\theta}^{(b,h)}}\Big[Z_{ij}^{(b)} \,\Big|\, Y_{ij}^{(b)}\Big], \quad \mathbb{E}_{\boldsymbol{\theta}^{(b,h)}}\Big[\exp(Z_{ij}^{(b)}) \,\Big|\, Y_{ij}^{(b)}\Big] \quad \text{and} \quad \mathbb{E}_{\boldsymbol{\theta}^{(b,h)}}\Big[\mathbf{Z}_i^{(b)\top} \boldsymbol{\Sigma}^{(b)^{-1}} \mathbf{Z}_i^{(b)} \,\Big|\, \mathbf{Y}_i^{(b)}\Big]. \tag{11}$$

In the sequel, while $\mathbf{Z}^{(b)}$ denotes the $n \times k$ matrix of latent variables for the block $\mathcal{C}_b$, $Z_{ij}^{(b)}$ does not refer to an element of the j-th column of $\mathbf{Z}^{(b)}$ but to an element of the column of $\mathbf{Z}^{(b)}$ associated to specie j.

Conversely, the data block $\mathbf{Y}^{(b)}$ contributes only to the estimation of the elements $\boldsymbol{\theta}^{(b)}$. As a consequence, to be able to estimate each element of the covariance matrix $\boldsymbol{\Sigma}$, we need to resort to overlapping blocks, so that each pair of species (j, ℓ) appears at least once in the same block. The construction of the blocks will be discussed in Section 4.2.

Asymptotic variance of the maximum composite likelihood estimates. As aforementioned, the estimator $\widehat{\boldsymbol{\theta}}_{c\ell}$ resulting from the maximization of the composite log-likelihood $c\ell(\boldsymbol{\theta})$ is consistent and asymptotically normal. Its asymptotic variance is given by the inverse of the so-called Godambe information matrix [Varin et al., 2011] defined for all $\boldsymbol{\theta} \in \Theta$ as

$$G(\boldsymbol{\theta}) = H(\boldsymbol{\theta})J(\boldsymbol{\theta})^{-1}H(\boldsymbol{\theta}) \quad \text{where} \quad \begin{cases} J(\boldsymbol{\theta}) & = \mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} c\ell(\boldsymbol{\theta})], \\ H(\boldsymbol{\theta}) & = -\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} c\ell(\boldsymbol{\theta})\big]. \end{cases} \tag{12}$$

where the expectation and the variance are taken with respect to $\mathbf{Y}$. Observe that, for regular models, because of the identity

$$\mathbb{V}\mathrm{ar}_{\boldsymbol{\theta}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})] = -\mathbb{E}_{\boldsymbol{\theta}}\big[\nabla^2_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y})\big], \tag{13}$$

when $c\ell(\boldsymbol{\theta})$ is the regular log-likelihood, we have $H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$. Consequently, $G(\boldsymbol{\theta}) = H(\boldsymbol{\theta}) = J(\boldsymbol{\theta})$ and fits the Fisher information matrix. In the general framework, the computation of $G(\boldsymbol{\theta})$ can be achieved using solely the gradient of the log-marginal on each block.

Proposition 2. *For all $\boldsymbol{\theta} \in \Theta$, the matrix H writes as a convex sum of Fisher information matrices of each block*

$$H(\boldsymbol{\theta}) = \sum_{b=1}^{C} \lambda_b \mathbb{E}_{\boldsymbol{\theta}}\big[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)})\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)})\}^\top\big]^{\blacksquare}.$$

where $A^{\blacksquare}$ stands for a $k \times k$ matrix A that has been embedded into a $p \times p$ matrix M with $M_{j\ell}$ taking value 0 if the pair (j, ℓ) does not belong to the possible pair of species of block $\mathcal{C}_b$ and being equal to the coefficient of A related to the pair (j, ℓ) otherwise.

This expression stems from that identity (13) holds for each block of data $\mathbf{Y}^{(b)}$ and that the score of the likelihood on a given block has a zero expectation.

Just as in the case of the regular likelihood framework, one can derive a sample variance estimator for $J(\boldsymbol{\theta})$ and a sample mean estimator for $H(\boldsymbol{\theta})$. For both estimators $\{\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)})\}_{1 \leq b \leq C}$ terms can be computed applying Louis' formula (5). Note that Louis' formula (6) could also be used in the context of composite likelihood to derive control variates estimators.

Model selection. Model selection procedure also exists in the context of composite likelihood inference and analogues of the AIC [Varin and Vidoni, 2005] and BIC [Gao and Song, 2010] criteria have been derived in this context. The latter reference proves the consistency of the composite likelihood BIC criterion, defined as

$$BIC = c\ell_{\boldsymbol{\theta}}(\mathbf{Y}) - \frac{\log(n)}{2}\dim(\boldsymbol{\theta}), \quad \text{where} \quad \dim(\boldsymbol{\theta}) = \mathrm{tr}[H(\boldsymbol{\theta})G(\boldsymbol{\theta})^{-1}]. \tag{14}$$

Observe that, in the regular likelihood context, both $H(\boldsymbol{\theta})G(\boldsymbol{\theta})^{-1} = J(\boldsymbol{\theta})H(\boldsymbol{\theta})^{-1}$ reduces to the identity matrix and $\dim(\boldsymbol{\theta})$ is the number of independent parameters, which yields in the regular BIC formulagit status. In Section 3.2, we will derive estimates of the matrices $H(\boldsymbol{\theta})$, $J(\boldsymbol{\theta})$ and $G(\boldsymbol{\theta})$, which will enable us to evaluate the composite likelihood BIC criterion.

3 Importance Sampling within EM algorithm

Monte-Carlo approximations of EM (MCEM) are among the most popular alternative to deterministic (*e.g.*, variational) approximations to deal with intractable E steps. The aim is to compute an estimate of the objective function (3) that is used to carry out the updates of the model parameters. While the MCEM method offers a wide range of solutions due to the extensive choices of sampling routines, it is not always straightforward to have an appropriate strategy for each problem. In what follows we focus on importance sampling as we will see it can benefit from existing variational approaches reminded in Section 2.2.

3.1 Maximum likelihood inference using importance sampling

In the framework of the PLN model (1), because the sites $1 \leq i \leq n$ are supposed to be independent, the function Q decomposes into

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i) \mid \mathbf{Y}_i]$$

The subsequent method remains nonetheless valid in a general framework by writing things in terms of the joint distribution of $\mathbf{Y}$ and $\mathbf{Z}$, though the sampling method might be more intricate in presence of dependence between, say, the sites. For each site $1 \leq i \leq n$, the core of the problem is to estimate quantities of the form

$$\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[f(\mathbf{Z}_i) \mid \mathbf{Y}_i] = \int_{\mathbb{R}^p} f(\mathbf{z}) p_{\boldsymbol{\theta}^{(h)}}(\mathbf{z} \mid \mathbf{Y}_i) \mathrm{d}\mathbf{z},$$

for some measurable function f. Importance sampling aims at estimating such expectation with respect to $p_{\boldsymbol{\theta}^{(h)}}(\cdot \mid \mathbf{Y}_i)$ by approximating this conditional distribution with a random probability measure based on weighted samples from a probability density function $q_i^{(h)}$, referred to as proposal distribution, such that $p_{\boldsymbol{\theta}^{(h)}}(\cdot \mid \mathbf{Y}_i)$ is absolutely continuous with respect to $q_i^{(h)}$. The latter assumption ensures that the Radon-Nikodym derivative of the distribution measure of $\mathbf{Z}_i \mid \mathbf{Y}_i$ with respect to the distribution mesure with density $q_i^{(h)}$ exists, namely it can be written as

$$\frac{p_{\boldsymbol{\theta}^{(h)}}(\cdot \mid \mathbf{Y}_i)}{q_i^{(h)}} \triangleq \frac{\rho_i^{(h)}}{\displaystyle\int_{\mathbb{R}^d} \rho_i^{(h)}(\mathbf{z}) q_i^{(h)}(\mathbf{z}) \mathrm{d}\mathbf{z}} \quad \text{where} \quad \rho_i^{(h)} = \frac{p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y}_i, \cdot)}{q_i^{(h)}}.$$

We thus have

$$\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[f(\mathbf{Z}_i) \mid \mathbf{Y}_i] = \frac{\displaystyle\int_{\mathbb{R}^p} f(\mathbf{z}) \rho_i^{(h)}(\mathbf{z}) q_i^{(h)}(\mathbf{z}) \mathrm{d}\mathbf{z}}{\displaystyle\int_{\mathbb{R}^d} \rho_i^{(h)}(\mathbf{z}) q_i^{(h)}(\mathbf{z}) \mathrm{d}\mathbf{z}},$$

and the related so-called self-normalized importance sampling estimator using $N \in \mathbb{N}^*$ independent samples $(\mathbf{V}_{i1}, \ldots, \mathbf{V}_{iN})$ from $q_i^{(h)}$ is

$$\widehat{\mathbb{E}}_{q_i^{(h)}}[f(\mathbf{Z}_i)] \triangleq \sum_{r=1}^{N} w_{ir}^{(h)} f(\mathbf{V}_{ir}), \quad \text{with} \quad w_{ir}^{(h)} = \frac{\rho_i^{(h)}(\mathbf{V}_{ir})}{\sum_{s=1}^{N} \rho_i^{(h)}(\mathbf{V}_{is})}. \tag{15}$$

The latter provides a consistent estimator that converges at a rate $\sqrt{N}$ and is asymptotically unbiased. Self-normalized estimates are however biased for fixed N (it is a special instance of estimator of a ratio of expected values). While it may be neglected at first since it decreases faster than the variance with N, it is possible to account for it using the method developed by Middleton et al. [2019] in the more comprehensive yet very useful particle filters framework.

PLN model. Self-normalized importance sampling estimators can therefore be applied to estimating the conditional moments (7), or more practically, their gradients counterpart involved in the M-step. Appendix A.2 shows how to perform a stochastic gradient scheme to achieve the update for the regression coefficients $\boldsymbol{B}$ and provide an update for the covariance matrix $\boldsymbol{\Sigma}$ from the same weighted sample of a proposal distribution $q_i^{(h)}$. These updates solely requires to compute for each site i

$$\left\{\widehat{\mathbb{E}}_{q_i^{(h)}}[\exp(Z_{ij})]\right\}_{1 \leq j \leq p} \quad \text{and} \quad \widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i \mathbf{Z}_i^\intercal]. \tag{16}$$

The whole inference procedure is described in Algorithm 3.

As for the variance of the estimators, the PLN model forms an example where we cannot either exactly compute the conditional expectation (5). Nonetheless, stemming from this identity, we can derive a plug-in estimator for the score function $\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i)$ by recycling the particles and the weights from the importance sampling scheme of the last iteration of ISEM Algorithm 3. Namely, the estimator (4) becomes

$$\widehat{I}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} \widehat{\mathbb{E}}_{q_i^{(h)}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i)] \widehat{\mathbb{E}}_{q_i^{(h)}}[\nabla_{\boldsymbol{\theta}} \log p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{Z}_i)]^\top \tag{17}$$

We shall remark that, to compute the plug-in estimator, we do not have to store the particles and the weights associated to each site. Indeed, for each site i, the plug-in estimator is based on the same self-normalized importance sampling estimators (16) used in the M Step (see Appendix A.2).

3.2 Maximum composite likelihood inference using importance sampling

Importance sampling performances may quickly degrade when the dimension of the sampling space increases. Indeed, Agapiou et al. [2017] shows that, over the class of bounded test function, one can control the L^2 error of the estimator if the sample size N increases at least exponentially with respect to the 2-Rényi divergence. Chatterjee and Diaconis [2018] achieve a similar result in high probability and for the KL divergence on another set of problems. The limitation of importance sampling arises from the difficulty to control such discrepancy measure and finding a proposal close to the, possibly, intricate target distribution. As a consequence, Algorithm 3 may become inefficient when dealing with a large number p of latent variables $\mathbf{Z}$ and reducing the dimension of the space in which importance sampling is performed becomes paramount.

The benefit of Algorithm 2 is that, when the conditional moments in the CL-E step are intractable, they can be estimated using importance sampling but with a sampling space of smaller dimension, *i.e.*, of the size of the block. Specifically, we can consider a proposal distribution $q_i^{(b,h)}$ specific to each iteration h, block $\mathcal{C}_b$ and site i and the self-normalized weights for a N-sample $(\mathbf{V}_{i1}, \ldots, \mathbf{V}_{iN})$ from $q_i^{(b,h)}$ writes as

$$w_{ir}^{(b,h)} = \frac{\rho_i^{(b,h)}(\mathbf{V}_{ir})}{\sum_{s=1}^{N} \rho_i^{(b,h)}(\mathbf{V}_{is})}, \quad \text{where} \quad \rho_i^{(b,h)} = \frac{p_{\boldsymbol{\theta}^{(b,h)}}(\mathbf{Y}_i^{(b)}, \cdot)}{q_i^{(b,h)}}. \tag{18}$$

PLN model. As in the case of the regular likelihood, importance sampling is used to estimating gradients counterpart of the moments (11). Appendix A.3 states that unlike the regular framework, we do not have access to a direct importance sampling estimator for $\boldsymbol{\Sigma}$ but we can still perform a stochastic gradient scheme to achieve the updates for both parameters that relies, for each site i, each block b and each specie j, on

$$\sum_{i=1}^{n} \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)} \mathbf{Z}_i^{(b)\top}\right] \quad \text{and} \quad \sum_{b \in \mathcal{C}(j)} \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\exp\left(Z_{ij}^{(b)}\right)\right].$$

The whole inference procedure is described in Algorithm 4.

The estimation of the Godambe information matrix (12) relies on following Monte Carlo estimators

$$\widehat{J}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \widehat{S}_i \widehat{S}_i^\top \quad \text{and} \quad \widehat{H}_n(\theta) = \frac{1}{n} \sum_{i=1}^{n} \sum_{b=1}^{C} \lambda_b \widehat{S}_i^{(b)} \widehat{S}_i^{(b)\top}, \tag{19}$$

where

$$\widehat{S}_i^{(b)} = \widehat{\mathbb{E}}_{q_i^{(h)}}\left[\nabla \log p_\theta\left(Y_i^{(b)}, Z_i^{(b)}\right)\right]^{\blacksquare} \quad \text{and} \quad \widehat{S}_i = \sum_{b=1}^{C} \lambda_b \widehat{S}_i^{(b)},$$

where $S^{\blacksquare}$ stands for a vector S of length $d{\times}k{+}k(k{+}1)/2$ that has been embedded into a vector T of length $d{\times}p{+}p(p{+}1)/2$ with T_j taking value 0 if the parameter $\boldsymbol{\theta}_j$ is not involved in the block $\mathcal{C}_b$ and being equal to the coefficient of S related to the parameter $\boldsymbol{\theta}_j$ otherwise. From a numerical perspective, the computation of the former estimators is somewhat less straightforward than those used for the Fisher information matrix. These matrices do not depend on precisely the same statistics as those involved in the M Step, owing to differences in the cross-products utilized. We need to store the statistics specific to each block and each site.

3.3 Importance proposal distribution

As aforementioned, the choice of the proposal distribution $q_i^{(h)}$ or $q_i^{(b,h)}$ is paramount to get reliable estimates with a reasonable simulation cost. Finding an appropriate proposal poses a challenge that can then be treated by adaptive sequential methods [*e.g.,* Cornuet et al., 2012, Daudel et al., 2021, Korba and Portier, 2022, Daudel et al., 2023] but the latter remain computationally cumbersome to be plugged in an MCEM scheme as the target changes at each iteration.

We choose here to build an initial proposal distribution upon the variational approximation from Chiquet et al. [2018, 2019] and subsequently adapt it at each iteration. Our approach relies on the following result that gives conditions, when the proposal distribution $q_i^{(h)}$ is Gaussian, to achieve non-normalized weights $\rho_i^{(h)}(\mathbf{V}_i)$ that (i) have finite variance and (ii) are bounded. Proposition 3 give conditions under which both properties hold; its proof is given in Appendix A.1.2.

Proposition 3. *Under the PLN model (1), given $1 \leq i \leq n$ and a parameter value $\boldsymbol{\theta} = (\boldsymbol{B}, \boldsymbol{\Sigma})$, letting $m \in \mathbb{R}^p$ and S be a symmetric positive definite matrix and denoting $\varphi(\cdot\,; m, S)$ the density function on $\mathbb{R}^p$ of the multivariate normal distribution $\mathcal{N}(m, S)$, it holds that:*

(i) If $2\boldsymbol{\Sigma}^{-1} - S^{-1}$ is positive definite, then

$$\int_{\mathbb{R}^p} \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v})^2}{\varphi(\mathbf{v}\,; m, S)} \mathrm{d}\mathbf{v} < \infty;$$

(ii) If $\boldsymbol{\Sigma}^{-1} - S^{-1}$ is positive definite, then we also have

$$\sup_{\mathbf{v} \in \mathbb{R}^p} \frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v})}{\varphi(\mathbf{v}\,; m, S)} < \infty.$$

Vanilla Gaussian proposal. In the first instance, we might consider the normal distribution that best fits the conditional distribution against which we integrate. From that perspective, the proposal $q_i^{(h)}$ is set, first (when $h = 0$) to the Gaussian approximation resulting from the variational inference, then to a Gaussian distribution whose first two moments matches the ones of $p_{\boldsymbol{\theta}^{(h-1)}}(\cdot \mid \mathbf{Y}_i)$.

This choice is surprisingly marred by a major drawback. The computation of the non-normalized weight $\rho_i^{(h)}(\mathbf{V}_i)$ still involves in the numerator a dominant quadratic term related not to the conditional variance, but to the marginal variance $\boldsymbol{\Sigma}^{(h)}$ (see Equation (23) in Appendix A.1.2). Using a normal proposal based on the conditional variance yields a distribution that is more concentrated than the joint distribution involved in the non-normalized weight.

The latter are thus never bounded in the tails. Moreover the condition given by Proposition 3 to get at least finite variance does not hold in all generality. Such a proposal distribution thus provides a fair idea of where to sample, but may nonetheless lead to a zero effective sample size.

Conversely, if we consider a Gaussian proposal that scales to the variance of the marginal distribution of $\mathbf{Z}_i$. Even though the associated non-normalized weights are solely bounded under specific assumptions about m, they always have a finite variance. However, samples from such a proposal are significantly more spread out compared to samples from a distribution that scales to the variance of the targeted conditional distribution of $\mathbf{Z}_i \mid \mathbf{Y}_i$. This results in a suboptimal exploration of the target and a low effective sample size.

ISEM algorithm for the PLN model. In what follows, we thus set $q_i^{(h)}$ to a mixture distribution that strikes a balance between the benefits offered by both aforementioned ones, namely, for $\alpha \in [0, 1)$

$$q_i^{(h)} = \alpha\varphi\left(\cdot; \mathbf{m}_i^{(h)}, \mathbf{S}_i^{(h)}\right) + (1-\alpha)\varphi\left(\cdot; \mathbf{m}_i^{(h)}, \boldsymbol{\Sigma}^{(h)}\right), \tag{20}$$

where $(\mathbf{m}_i^{(0)}, \mathbf{S}_i^{(0)})$ and $\boldsymbol{\Sigma}^{(0)}$ are respectively the variational parameter $(\mathbf{m}_i, \mathbf{S}_i)$ and the variational estimation of $\boldsymbol{\Sigma}$, and for $h \geq 1$, $\mathbf{m}_i^{(h)}$ and $\mathbf{S}_i^{(h)}$ are Monte Carlo estimates of the mean and variance of the conditional distribution $p_{\boldsymbol{\theta}^{(h-1)}}(\cdot \mid \mathbf{Y}_i)$. The following proposition states that the importance weights associated with the mixture inherit the same properties as the non-normalized weight related to $\varphi(\cdot; m, \boldsymbol{\Sigma}^{(h)})$. The proof follows directly from Proposition 3 and is given in Appendix A.1.2.

Proposition 4. *Under the same condition as Proposition 3, with* $\boldsymbol{\theta}^{(h)} = (\boldsymbol{B}^{(h)}, \boldsymbol{\Sigma}^{(h)})$*, for all* $\alpha \in [0, 1)$*, the non-normalized weights associated to the proposal distribution* (20) *satisfy*

$$\mathbb{E}_{q_i^{(h)}}\left[\rho_i^{(h)}(\mathbf{V})^2\right] = \int_{\mathbb{R}^p} \frac{p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y}_i, \mathbf{v})^2}{q_i^{(h)}(\mathbf{v})} \mathrm{d}\mathbf{v} < \infty.$$

The mixture proportion α controls how close we aim to be from the optimal Gaussian proxy of the conditional, while the second component is solely dedicated to ensuring finite variance of the non-normalized weights in Proposition 4. Algorithm 3 summarizes our MCEM algorithm for the inference of the PLN model (1) based on the mixture proposal 20.

Composite ISEM algorithm for the PLN model. The proposal distribution (20) can easily be extended to the composite likelihood framework, where the aim is to have a mixture on each block. Namely, the proposal distribution for block $\mathcal{C}_b$ at iteration h is

$$q_i^{(b,h)} = \alpha\varphi\left(\cdot; \mathbf{m}_i^{(b,h)}, \mathbf{S}_i^{(b,h)}\right) + (1-\alpha)\varphi\left(\cdot; \mathbf{m}_i^{(b,h)}, \boldsymbol{\Sigma}^{(b,h)}\right), \quad \alpha \in (0, 1),$$

where the parameter corresponds to those from (20) reduced to block $\mathcal{C}_b$, *i.e.*,

$$\{(\mathbf{m}_i^{(b,0)}, \mathbf{S}_i^{(b,0)})\}_{1\leq i\leq n} = \{([\mathbf{m}_{i,j}], [\mathbf{S}_{i,j}])\}_{1\leq i\leq n, j\in\mathcal{C}_b}, \qquad \boldsymbol{\Sigma}^{(b,0)} = [\boldsymbol{\Sigma}_{jk}^{(0)}]_{(j,k)\in\mathcal{C}_b},$$

the updates $\mathbf{m}_i^{(b,h)}$ and $\mathbf{S}_i^{(b,h)}$ are the estimated mean and variance of the conditional distribution $p_{\boldsymbol{\theta}^{(b,h-1)}}(\cdot \mid \mathbf{Y}_i^{(b)})$, while $\boldsymbol{\Sigma}^{(b,h)}$ is the current estimate of $\boldsymbol{\Sigma}^{(b)}$, that is $\boldsymbol{\Sigma}^{(b,h)} = [\boldsymbol{\Sigma}_{jk}^{(h)}]_{(j,k)\in\mathcal{C}_b}$. The whole inference procedure for the PLN model and this peculiar choice of proposal is summarized in Algorithm 4.

Algorithm 3: ISEM for PLN

Input: number of iterations n_{iter}, number of draws N, mixture proportion α, variational parameter $\{(\mathbf{m}_i^{(0)}, \mathbf{S}_i^{(0)})\}_{1\leq i\leq n}$, variational estimation $\boldsymbol{\Sigma}^{(0)}$ of $\boldsymbol{\Sigma}$.

for $h = 0$ **to** $n_{\text{iter}} - 1$ **do**

  for $i = 1$ **to** n **do**

    sample $(\mathbf{V}_{i1}, \ldots, \mathbf{V}_{iN})$ from $q_i^{(h)} \sim \alpha\mathcal{N}\left(\mathbf{m}_i^{(h)}, \mathbf{S}_i^{(h)}\right) + (1-\alpha)\mathcal{N}\left(\mathbf{m}_i^{(h)}, \boldsymbol{\Sigma}^{(h)}\right)$;

    compute $\left(w_{i1}^{(h)}, \ldots, w_{iN}^{(h)}\right)$ according to (15);

    compute $\widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i]$, $\widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i\mathbf{Z}_i^\top]$ and $\left\{\widehat{\mathbb{E}}_{q_i^{(h)}}[\exp(Z_{ij})]\right\}_{1\leq j\leq p}$;

    set

$$\mathbf{m}_i^{(h+1)} = \widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i] \quad \text{and} \quad \mathbf{S}_i^{(h+1)} = \widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i\mathbf{Z}_i^\top] - \mathbf{m}_i^{(h+1)}\mathbf{m}_i^{(h+1)^\top};$$

  end

  /* See Appendix A.2 for update formulas */

  set $\boldsymbol{\Sigma}^{(h+1)} = \dfrac{1}{n}\displaystyle\sum_{i=1}^{n} \widehat{\mathbb{E}}_{q_i^{(h)}}[\mathbf{Z}_i\mathbf{Z}_i^\top]$;

  for $j = 1$ **to** p **do**

    set $\boldsymbol{\beta}_j^{(h+1)}$ with a gradient scheme based on $\left\{\widehat{\mathbb{E}}_{q_i^{(h)}}[\exp(Z_{ij})]\right\}_{1\leq i\leq n}$;

  end

end

Relation to adaptive importance sampling. Adaptive version of Gaussian mixture proposal distributions can be achieved with the M-PMC algorithm of Cappé et al. [2008]. Specifically for the PLN model, in order to update the proposal distribution at each iteration h for each site i, and potentially each block b, we may consider their defensive option, incorporating a Gaussian distribution with a covariance matrix that fulfils condition (i) of Proposition 3 as defensive component, *e.g.*, for $0 \leq \alpha < 1$ and $K \in \mathbb{N}^*$,

$$\widetilde{q}_i^{(h)} = (1-\alpha)\varphi\left(\cdot; \mathbf{m}_i^{(h)}, \boldsymbol{\Sigma}^{(h)}\right) + \alpha\sum_{k=1}^{K} \eta_k \varphi\left(\cdot; \widetilde{\mathbf{m}}_{ik}^{(h)}, \widetilde{\mathbf{S}}_{ik}^{(h)}\right). \tag{21}$$

If we set the proposal distribution to lie within the family of two-component Gaussian mixtures ($K = 1$), the method resume to recursively update the mean and the covariance matrix of the free component to minimizing the Kullback-Leibler divergence between the conditional distribution of interest and the mixture distribution. The latter optimum differs from our solution that solely aims at minimizing the Kullback-Leibler divergence between the conditional distribution of interest and the free component. A numerical comparison of the performances will be presented in Section 4.1.

4 Illustrations

In this section, we first provide a comparison of our importance sampling setting with an adaptive importance sampling scheme in the context of Gaussian mixtures. Secondly, we discuss the construction of the blocks in view of composite likelihood inference. Then, we compare and assess the performances of the different inference algorithms previously proposed on synthetic datasets. We also present a comparison with the original VEM algorithm. Eventually, we use them to analyse populations of fishes from the Barents sea.

Algorithm 4: Composite ISEM for PLN

Input: number of iterations n_{iter}, number of draws N, mixture proportion α, variational parameter on each block $\{(\mathbf{m}_i^{(b,0)}, \mathbf{S}_i^{(b,0)})\}_{1\leq i\leq n}$, variational estimation $\boldsymbol{\Sigma}^{(0)}$ of $\boldsymbol{\Sigma}$.

for $h = 0$ **to** $n_{\text{iter}} - 1$ **do**
  for $b = 1$ **to** C **do**
    set $\boldsymbol{\Sigma}^{(b,h)} = [\boldsymbol{\Sigma}_{jk}^{(h)}]_{(j,k)\in\mathcal{C}_b}$;
    for $i = 1$ **to** n **do**
      sample $(\mathbf{V}_{i1}^{(b)}, \ldots, \mathbf{V}_{iN}^{(b)})$ from $q_i^{(b,h)} \sim \alpha\mathcal{N}\left(\mathbf{m}_i^{(b,h)}, \mathbf{S}_i^{(b,h)}\right) + (1-\alpha)\mathcal{N}\left(\mathbf{m}_i^{(b,h)}, \boldsymbol{\Sigma}^{(b,h)}\right)$;
      compute $\left(w_{i1}^{(b,h)}, \ldots, w_{iN}^{(b,h)}\right)$ according to (18);
      compute $\widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)}\right]$, $\widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)}\mathbf{Z}_i^{(b)\top}\right]$ and $\left\{\widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\exp(Z_{ij}^{(b)})\right]\right\}_{j\in\mathcal{C}_b}$;
      set

$$\mathbf{m}_i^{(b,h+1)} = \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)}\right] \quad \text{and} \quad \mathbf{S}_i^{(b,h+1)} = \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)}\mathbf{Z}_i^{(b)\top}\right] - \mathbf{m}_i^{(b,h+1)}\mathbf{m}_i^{(b,h+1)\top};$$

    end
  end
  `/* See Appendix A.3 for update formulas */`
  set $\boldsymbol{\Sigma}^{(h+1)}$ with a gradient scheme based on $\left\{\sum_{i=1}^n \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\mathbf{Z}_i^{(b)}\mathbf{Z}_i^{(b)\top}\right]\right\}_{1\leq b\leq C}$;
  for $j = 1$ **to** p **do**
    set $\boldsymbol{\beta}_j^{(h+1)}$ with a gradient scheme based on $\left\{\sum_{b\in\mathcal{C}(j)} \widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\exp\left(Z_{ij}^{(b)}\right)\right]\right\}_{1\leq i\leq n}$;
  end
end

4.1 Comparison with adaptative importance sampling in a Gaussian mixture family

Denote $\widetilde{w}_{ir}^{(h)}$, $1 \leq r \leq N$, the importance weights associated with a N-sample from the mixture proposal (21) and $\mathcal{R}$ the relative efficiency in terms of the normalized perplexity of the importance weights, that is

$$\mathcal{R} = \frac{\exp\left(-\sum_{r=1}^N w_{ir}^{(h)} \log w_{ir}^{(h)}\right)}{\exp\left(-\sum_{r=1}^N \widetilde{w}_{ir}^{(h)} \log \widetilde{w}_{ir}^{(h)}\right)}.$$

Over a data set of 1000 count matrices randomly draws from the PLN model (1) with $n = 100$, $p = 2, 3, 5$ and 7, and $d = 3$, Table 1 shows that the adaptive scheme converges towards a mixture for which the Shannon entropy of the associated normalized weight is equivalent to our proposal.

For the same data set, whether the number of free mixture component in the adaptive scheme is $K = 1$ or $K = 2$, we can also observe that the 5% and 95% quantiles of Monte Carlo estimates of the Kullback-Leibler divergence $\mathrm{KL}[\widetilde{q}_i^{(h)} \| q_i^{(h)}]$ is close to zero (Table 2). In the context of the PLN model, the adaptive scheme applied to Gaussian mixture models proves to be of little use. It implies an additional computational burden and does not lead to significant improvement as the proposal $q_i^{(h)}$ is already close to the optimal adaptive version.

Table 1: Normalized perplexity relative efficiency $\mathcal{R}$ for $N = 5000$ draws. Quantiles of the relative efficiency observed over a data set of a 1000 count matrices corresponding to 100 independent draws from the PLN model (1) ($n = 100$ sites, $d = 3$ covariates) for 10 different random parameter configurations.

Number of species p	2	3	5	7
5%-quantile	0.997	0.994	0.990	0.981
95%-quantile	1.123	1.119	1.067	1.036

Table 2: Kullback-Leibler divergence $\mathrm{KL}[\widetilde{q}_i^{(h)} \| q_i^{(h)}]$ for $N = 5000$ draws and mixture proposal $\widetilde{q}_i^{(h)}$ with $K = 1$ or $K = 2$ free components. Quantiles of the Monte Carlo estimates for a data set of a 1000 count matrices corresponding to 100 independent draws from the PLN model (1) ($n = 100$ sites, $d = 3$ covariates) for 10 different random parameter configurations.

	$K = 1$				$K = 2$			
Number of species p	2	3	5	7	2	3	5	7
5%-quantile ($\times 10^{-3}$)	4.28	1.77	2.90	5.87	6.27	4.77	7.98	13.7
95%-quantile ($\times 10^{-1}$)	1.31	1.36	0.80	0.59	1.35	1.38	0.83	0.71

4.2 Determining the blocks for composite likelihood inference

We first discuss the way species can be spread into blocks for composite-likelihood inference. In this paper, we only considered blocks with constant size $k \leq p$. Obviously, for a given block size k, using a small number of blocks C alleviates the computational burden.

To get an estimate of each entry of the covariance matrix Σ, it is sufficient that each pair of distinct species (j, j'), $1 \leq j < j' \leq p$, appears at least once in a same block $\mathcal{C}_b$, $1 \leq b \leq C$. Hence, it is unnecessary to explore every possible combinations of blocks with size k, so $C \leq \binom{p}{k}$. On the other hand, because there are $p(p-1)/2$ pairs of species and because each block contains $k(k-1)/2$ pairs (giving a total of $Ck(k-1)/2$ pairs), we need that $C \geq p(p-1)/[k(k-1)]$. Remark that $k = 2$ is a trivial case as each blocks contributes to estimate one single covariance parameter so $C = p(p-1)/2$.

Spreading the p species into k blocks is equivalent to build an incomplete block design in terms of design of experiments. In the general case, finding an incomplete block design with a minimal number of block is a challenging combinatorial task. We conceived a greedy stochastic algorithm to build such a design. Figure 1 gives the number of blocks returned by this algorithm for the various configurations of the simulation study.

We observe that the upper bound $\binom{p}{k}$ is much too pessimistic, as compared to the obtained number of blocks C. Our algorithm find a number of blocks close to the lower bound $p(p-1)/[k(k-1)]$. Furthermore, we note a strong dependence of C on k. For instance, taking $k = 5$ yields in a smaller the number of blocks than taking $k = 2$. Consequently, we expect better computational efficiency for the CL5 algorithm than for the CL2. Due to C increasing as p^2 for $k = 2$, we did not run the CL2 algorithm for $p > 30$.

4.3 Simulation study

4.3.1 Simulation design

To mimic typical datasets encountered in community ecology or biogeography, we fixed the number of sites to $n = 100$ and the number of covariates to $d = 3$ (that is, one intercept and two covariates) and made the number of species vary from $p = 5$ to $p = 50$. The offset term was set

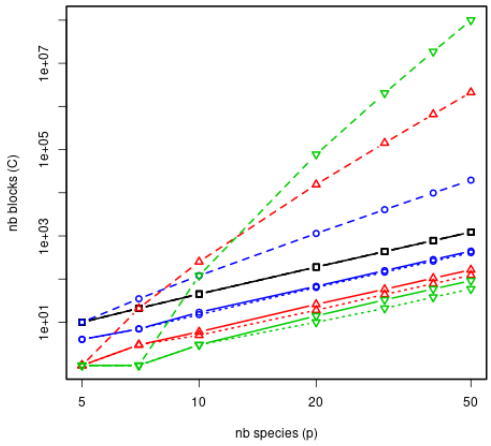

Figure 1: Number of blocks C as a function of the number of species p (in log-log-scale) for blocks of size $k = 2$ (black squares ■), $k = 3$ (blue circles ○), $k = 5$ (red triangles up △) and $k = 7$ (green triangles down ▽). Solid line: number of blocks actually used, dashed line: upper bound $\binom{p}{k}$, dotted line: lower bound $p(p-1)/[k(k-1)]$.

to zero. For each dimension p, we fixed the $n \times d$ matrix of covariates $\mathbf{X}$, the $d \times p$ matrix of regression coefficient $\boldsymbol{B}$ and the $p \times p$ covariance matrix $\boldsymbol{\Sigma}$ and sampled $M = 100$ count matrices $\mathbf{Y}^m$ $(1 \leq m \leq M)$ with dimension $n \times p$ according to the PLN model (1).

Estimation algorithms. For small number of species ($p \leq 10$), we carried out maximum likelihood inference using our ISEM algorithm 3, and referred to as "full likelihood" (FL). For each simulated dataset $\mathbf{Y}^m$, we obtained the parameter estimates $\widehat{\boldsymbol{B}}^m$ and $\widehat{\boldsymbol{\Sigma}}^m$, and their respective estimated asymptotic variance related to the estimated Fisher information matrix (17).

For all number of species and each simulated dataset, we also performed composite likelihood inference using our composite ISEM algorithm 4 with blocks of size $k = 2, 3, 5$ and 7. It is further referred to as "composite likelihood" or "CLk". For each simulation setting, we obtained the parameter estimates $\widehat{\boldsymbol{B}}^m$ and $\widehat{\boldsymbol{\Sigma}}^m$, and their respective estimated asymptotic variance using the Godambe matrix based on estimates (19).

For all ISEM algorithms, we used a linearly increasing number of particles along the iterations: at iteration h, we used $N^{(h)} = hN^{(0)}$ particles. We considered $N^{(0)} = 50, 100$ and 200 initial particles. We did not observe any significant difference, except in the situations (not shown) that are actually doomed to fail with all considered numbers of particles due to the aforementioned limitation of IS when the dimension p or k grows. We report hereafter the results obtained with $N^{(0)} = 200$ initial particles. We let the algorithms run at most 1000 iterations and used a lag of 50 steps for the stopping criterion. Regarding the mixture proposal distribution defined in Equation (20), we used the mixture proportion $\alpha = 0.9$.

Normality of the estimators. We are interested in the validity of the model parameters inference, especially in the ability to provide valid tests and/or confidence intervals, which are not provided by variational inference. To this aim, for each algorithm under consideration and each, say, regression parameter $\beta_{\ell j}$, we examined the standardized estimates

$$\widetilde{\beta}_{\ell j} = \left(\widehat{\beta}_{\ell j} - \beta_{\ell j}\right) \Big/ \sqrt{\widehat{\mathbb{V}\mathrm{ar}}\left(\widehat{\beta}_{\ell j}\right)} \tag{22}$$

where $\beta_{\ell j}$ stands for the true value, $\widehat{\beta}_{\ell j}$ for its estimate provided by the algorithm and $\widehat{\mathbb{V}\mathrm{ar}}(\widehat{\beta}_{\ell j})$ for the estimated variance of $\widehat{\beta}_{\ell j}$. According to the M-estimator theory, for a given set of parameter $\boldsymbol{\theta}$ and a given ISEM algorithm, the distribution of the $\widetilde{\beta}_{\ell j}^{(m)}$ across simulations $m = 1, \ldots, M$ should be close to a standard normal. We used the Kolmogorov-Smirnov (KS) test to assess this normality and reported the resulting p-value.

4.3.2 Simulation results

Full likelihood vs composite likelihood. Figure 2 gives the distributions of the KS test for both the FL and the CLk algorithms, for $k = 2, 3, 5$ and 7. The boxplots consist of KS p-values, associated with the $p \times d$ regression parameters $\beta_{j\ell}$. This figure shows that normality is not rejected for the CL algorithms, even for large number of species ($p = 30$ or 50). A departure from normality is however observed with full likelihood inference, for a moderate number of species ($p \simeq 10$). This illustrates the difficulty encountered by importance sampling to perform well, even for a moderate values of p, while keeping a reasonable computational budget.

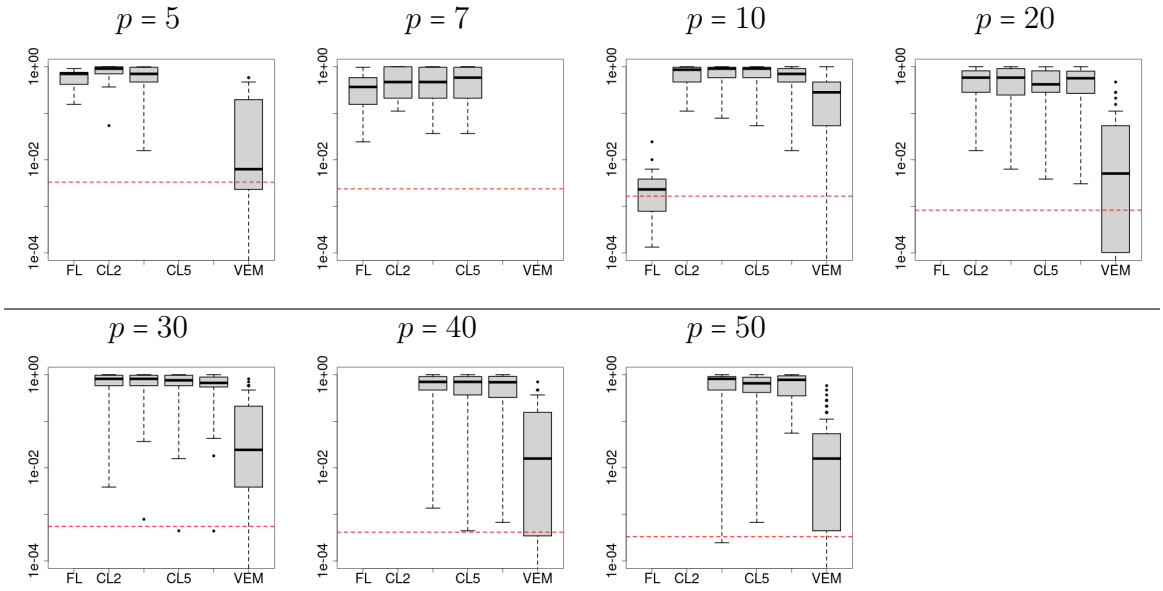

Figure 2: Distribution of the p-values of the Kolmogorov-Smirnov test for the distribution of the standardized estimates $\widetilde{\beta}_{\ell j}$ over the $M = 100$ simulations as a function of the inference algorithm (FL, CLk and VEM). Each boxplot is built across the $d \times p = 3p$ normalised coefficients $\widetilde{\beta}_{\ell j}$. Dotted red lines: $\alpha = 5\%$ significance threshold after Bonferroni correction (*i.e.*, $\alpha/(dp)$).

Figure 3 points out the good fit of the tests statistics defined in Equation (22) to the standard normal for all composite likelihood algorithms. These results strikingly contrast with the results corresponding to the specific case of full likelihood inference presented in Appendix A.4.1. Figure 10 shows a systematic increase of the departure from normality when the number of species grows from $p = 5$ to $p = 10$, p-values dropping faster once $p \geq 7$. The qq-plots from Figure 11 outlines that, while departing from the normal distribution, the FL algorithm tends to over-estimate $\mathbb{V}\mathrm{ar}(\widehat{\beta}_{\ell j})$ as p increases. The discrepancy between composite likelihood and full likelihood approaches can be further observed on Figure 12. On that instance for $p = 10$, the distribution of the $\widehat{\beta}_{\ell j}$ resulting from the different CLk algorithms do fit a standard Gaussian, whereas these resulting from FL does not.

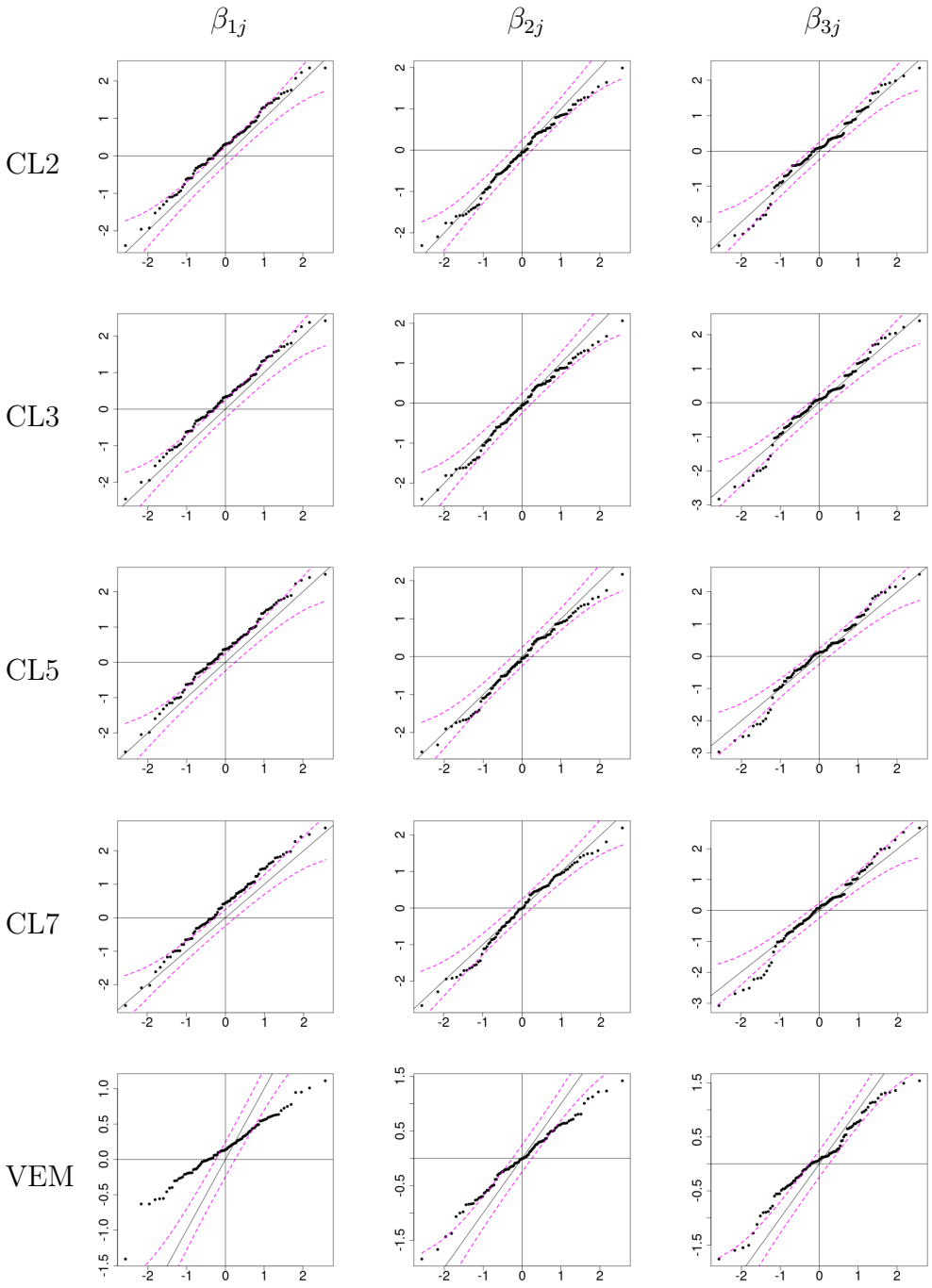

Figure 3: qq-plots of the normalized regression coefficient $\widetilde{\beta}_{\ell j}$ defined in Equation (22) for each of the $d = 3$ covariates and for the second simulated species (among $p = 30$) resulting from the composite likelihood algorithms CLk for $k = 2, 3, 5, 7$ and the VEM algoritmh (with jackknife variance estimates). x-axis: standard normal quantiles, y-axis: quantiles of $\widetilde{\beta}_{\ell j}$ (black dots [•]), magenta dashed lines [- -]: 95% bounds for the standard normal qq-plot.

Variance of the estimates. We then studied the effect of the block size k on the estimated asymptotic variance of the regression coefficient estimates provided by composite likelihood inference. Figure 4 displays the distribution (across coefficients) of the variance ratios, taking the CL5 algorithm as an arbitrary reference to account for the intrinsic variability between coefficients. This figure shows that the variance slightly decreases as k increases, but always remains close to one.

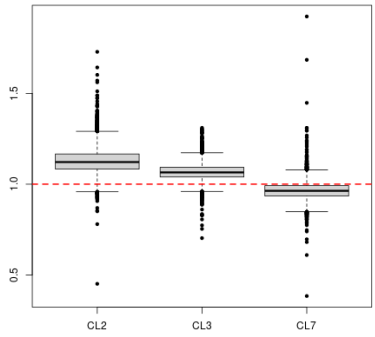

Figure 4: Boxplot of the relative variance of the estimates $\widehat{\beta}_{\ell j}$ of the regression coefficients obtained with the CL2, CL3 and CL7 algorithms, as compared to the CL5 algorithm for $p = 30$ species. Each boxplot is built across the $d \times p = 90$ normalised coefficients $\widetilde{\beta}_{\ell j}$.

Computational time. Table 3 gives the mean computational times required by the CLk algorithms for several number of species p, together with the number of iterations and the number of blocks. We first observe that the mean number of iterations is much smaller than the maximal number allowed (1000). As expected, the computational time mainly depends on the number of species p, which is both explained by the increasing number of iterations and the number of blocks required.

Table 3: Computational performances averaged over the $M = 100$ simulations for $p = 10, 30$ and 50 species for each of the algorithm CLk, with $k = 2, 3, 5$ and 7 and the VEM algorithm: mean computational time in seconds (left), mean number of iterations (center) number of blocks (right).

	Computational time (s)					Number of iterations				Number of blocks			
Algorithm	CL2	CL3	CL5	CL7	VEM	CL2	CL3	CL5	CL7	CL2	CL3	CL5	CL7
$p = 10$	44	42	216	149	10	124	115	107	101	45	17	6	3
$p = 30$	4061	8936	592	3196	18	299	256	243	216	435	159	60	34
$p = 50$	–	10769	8744	4249	34	–	522	443	123	–	448	166	93

Comparison with VEM estimates. Our work is mainly motivated by the estimation of the (asymptotic) variance of the estimate, so to provide confidence intervals or significance tests for the model's parameters. In this perspective, we considered the jackknife estimate of the variance of the VEM estimators described in Section 2.2. Table 3 shows that, even if the jackknife procedure requires to run $n = 100$ times the VEM algorithm, the over computational burden of VEM remains (much) lighter than this of the FL and CLk algorithms we propose. Resorting to the jackknife procedure is hence appealing to carry out proper inference in an efficient manner.

Unfortunately, Figure 2 shows that the discrepancy between the distribution of the standardized statistics $\widetilde{\beta}_{k\ell}$ and a standard normal is poor (normality is rejected in a large proportion of simulations). Figure 3 (last row) shows that this poor fit is mostly due to an over-estimation of the variance, resulting in too wide confidence intervals ans poor statistical power for hypothesis testing.

4.4 Fish abundances in the Barents sea

We now illustrate the use of the proposed methdology on fish abundances in the Barents sea. The dataset consists of the abundances of $p = 30$ fish species in $n = 89$ stations (sites) from the Barents sea that have been collected between April and May 1997 and described by Fossheim et al. [2006]. For each sample, the latitude and longitude of the station as well as the temperature of the water and the depth were recorded. The data are available from the `PLNmodels` R package [Chiquet et al., 2021]. Abundances were all obtained with the same experimental protocol, so no offset term is required in the model.

For these illustrations, we started with $N^{(0)} = 200$ particles, we set the maximum number of iterations to 10000 and used a lag of 50 steps for the stopping rule.

4.4.1 Reduced datas et: $p = 7$ species

We first consider only the $p = 7$ more abundant species of the dataset, in order to compare full likelihood (FL) inference with composite likelihood (CL) inference described in Sections 3.1 and 3.2. For composite likelihood, we used the same block sizes as in Section 4.3, that is $k = 2, 3, 5$ and 7. The number of iterations for each algorithm were FL = 2885, CL2 = 4172, CL3 = 4471 and CL5 = 3484, while the variational EM algorithm from the `PLNmodels` packages took 89 before convergence.

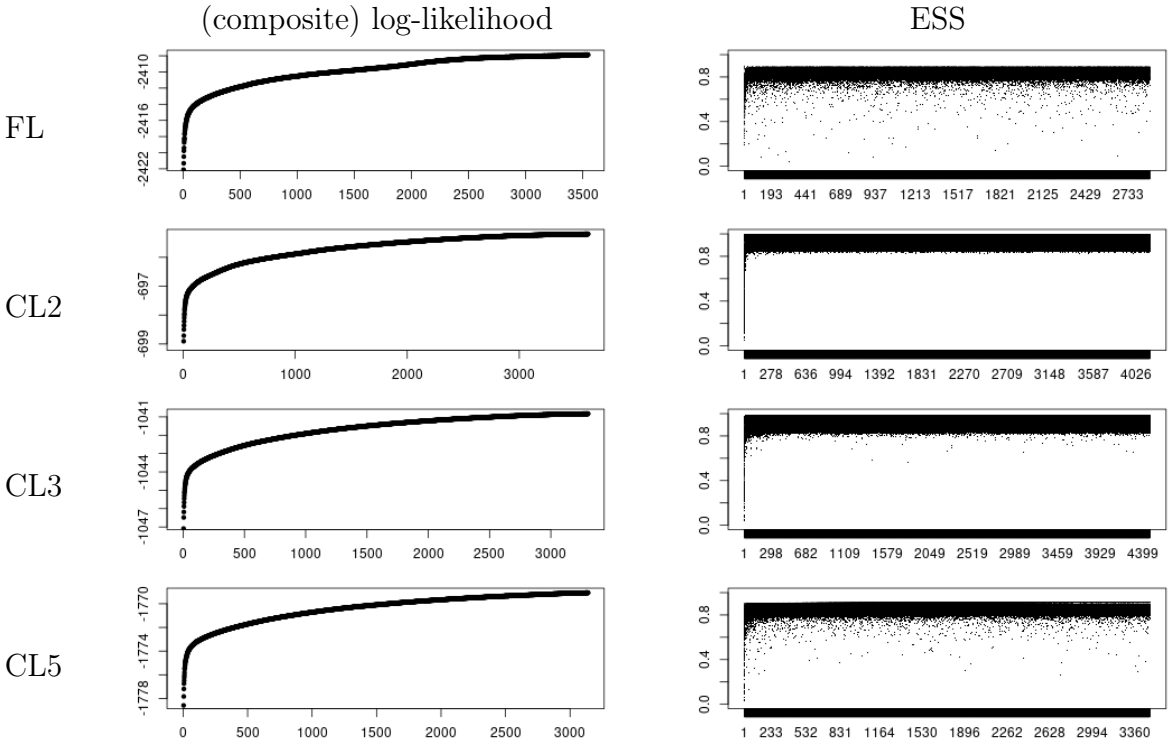

Figure 5: (Composite) log-likelihood (left) and effective sample size (right) along iterations for the different algorithms and Barents sea reduced data set ($p = 7$ species). From top to bottom: FL, CL2, CL3, CL5. Left: one dot for each iteration, right: one boxplot for each iteration (one value for each block and each site).

Figure 5 displays the evolution of the estimated (composite) log-likelihood along the iterations, as well as the distribution of the effective sampling sizes (ESS) across each site and each block. We observe the typical behavior of EM-algorithms, with a dramatic increase of the log-likelihood during the first steps, then a slower convergence before reaching a plateau. In

terms of sampling efficiency, we observe that the ESS goes rapidly (less than 50 iterations) to 80%.

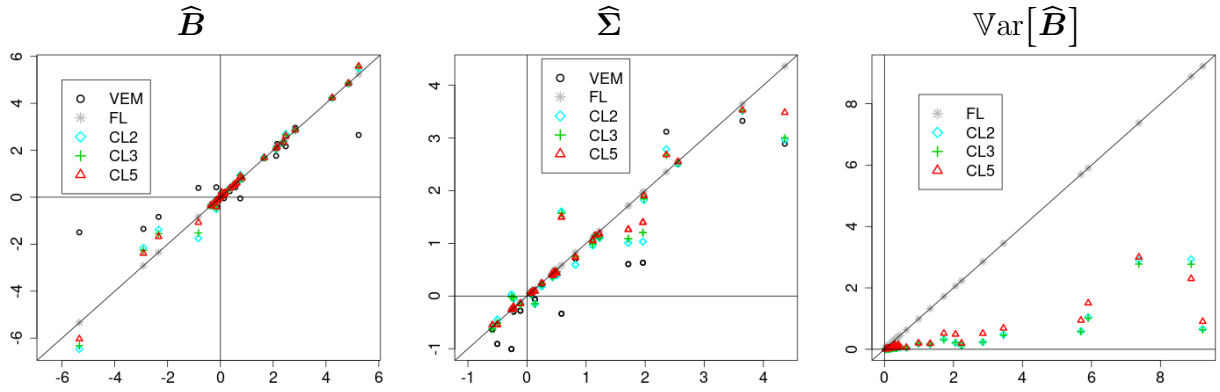

Figure 6: Comparison of the estimates with the different algorithms for the Barents sea reduced data set ($p = 7$ species). Left: regression coefficients $\widehat{\beta}_{hj}$, center: covariance parameters $\widehat{\Sigma}_{jk}$, right: estimated variances of the regression coefficients $\widehat{\beta}_{hj}$. x-axis = full likelihood inference (FL), y-axis: likelihood inference (FL = gray asterisk [$*$] = reference), composite likelihood inference (CL2 = cyan diamond [$\diamond$], CL3 = green plus sign [$+$], CL5 = red triangle [$\triangle$) and variational EM (VEM = black circles [$\circ$]).

Figure 6 compares the estimates obtained with the different algorithms. We observe that the estimates of both the regression coefficients β_{hj} are fairly consistent for FL and CL algorithms, but depart from the variational estimates (VEM), which was used to initialize each of them. The same observation holds for covariance parameters Σ_{jk}, with a higher diversity among the FL and CL algorithms. Importantly, the variance of the estimators of the regression coefficients $\widehat{\beta}_{hj}$ vary importantly between full-likelihood and composite-likelihood inference, the latter providing less variant estimates. This suggests that, in the case of the Poisson log-normal model, the composite likelihood inference yields a higher power to detect potential effects of the covariates.

4.4.2 Full data set: $p = 30$ species

We then considered the full dataset of Fossheim et al. [2006] and ran the composite likelihood algorithm described in Section 2.3 with $k = 3, 5$ and 7 blocks (CL3, CL5, CL7), as well as the VEM algorithm using the `PLNmodels` package.

Comparison of the different algorithms. The left part of Figure 7 shows that the estimates of the regression coefficients $\beta_{\ell j}$ obtained with the three CL algorithms were all very close. Moreover, it shows that the difference with the variational estimates is also quite small. The same observation can be made about the covariance parameters Σ_{jk} (same Figure, center). These results show that the estimates are robust to the choice of the block size k. It also confirms the general observation that VEM provides accurate estimates of the parameters (without matching them with any uncertainty measure).

The right part of Figure 7 assess the sampling efficiency of the CL5 algorithm, which stopped after 2571 iterations: the ESS for each site and block is higher than 80% after less than 100 iterations. The CL3 and CL7 algorithms displayed a similar behavior (not shown).

To further inquire difference between the VEM and CL alogithms, we checked the proportion of VEM estimates covered by the CL confidence intervals. We observed that only 4 variational estimates of the regression coefficients $\beta_{\ell j}$ (out of $q \times p = 150$) felt outside the CL5 composite likelihood confidence interval and that all variational estimates of the variance parameters σ_{kj} where inside the CL5 confidence intervals. To this respect, only a very small fraction of the

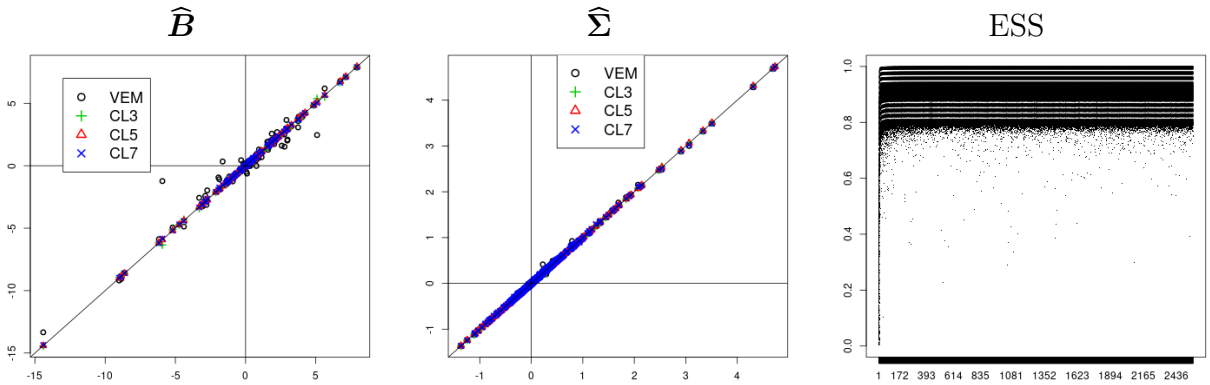

Figure 7: Barents sea full data set (p = 30 species). Estimates of the regression coefficients $\beta_{\ell j}$ (left) and the covariance parameters σ_{jk} (center). x-axis: estimates obtained with the CL5 algorithm (red triangle [△] = reference), y-axis: estimates obtained with the VEM (black circles [○]), CL3 (green plus sign [+]) and CL7 (blue 'times' sign [×]). Right: boxplot of the efficient sample size for all sites and blocks as a function of the iterations for the CL5 algorithm.

VEM estimates turned out to significantly different from the corresponding CL5 estimates. Still, Table 4 shows that, when such a discrepancy occurs, it may lead to substantially different estimations of the regression parameters.

Table 4: Comparison of the VEM and CL5 estimates when the VEM estimate does not meet the CL5 95% confidence interval for the Barents sea full data set (p = 30 species).

Regression parameter	$\beta_{1,1}$	$\beta_{2,2}$	$\beta_{2,2}$	$\beta_{1,11}$
VEM	-1.23	2.46	2.09	-0.94
CL5	-6.33	5.39	3.10	-1.93

Model selection. As reminded at the end of Section 2.3, a BIC criterion can be associated with composite likelihood inference. Combining the four available covariate, we fitted the 2^4 possible models (from Model 1: intercept only, Model 2: intercept plus Latitude, Model 3: intercept plus Longitude, up to Model 16: intercept plus the four covariates) to the full Barents data set. For each model, we estimated the dimension $\dim(\boldsymbol{\theta})$ defined in (14) as $\widehat{\dim}(\boldsymbol{\theta}) = \mathrm{tr}[\widehat{H}_n(\boldsymbol{\theta})\widehat{G}_n(\boldsymbol{\theta})^{-1}]$ where $\widehat{H}_n(\boldsymbol{\theta})$ and $\widehat{J}_n(\boldsymbol{\theta})$ are given in (12) and $\widehat{G}_n(\boldsymbol{\theta}) = \widehat{H}_n(\boldsymbol{\theta})\widehat{J}_n(\boldsymbol{\theta})^{-1}\widehat{H}_n(\boldsymbol{\theta})$.

The left panel of Figure 8 gives the values of the composite log-likelihoods obtained with the CL5 and CL7 algorithms. We subtracted the composite log-likelihood for Model 1 to emphasize that $c\ell_{\widehat{\boldsymbol{\theta}}}(bY)$ does not increase at the same speed, depending on the block size k = 5 or 7. This is consistent with the notion of adaptive dimension $\dim(\boldsymbol{\theta})$ appearing in the BIC derived by Gao and Song [2010] (see (14)).

The center panel of Figure 8 shows that, although the estimate $\widehat{\dim}(\boldsymbol{\theta})$ of this adaptive dimension is a combination of the two Monte-Carlo estimates $\widehat{H}_n(\boldsymbol{\theta})$ and $\widehat{J}_n(\boldsymbol{\theta})$, it remains almost constant among models involving the same number of variables.

The right panel of Figure 8 gives the final BIC criteria for the 16 models and the two algorithms CL5 and CL7 (still subtracting the composite log-likelihood of Model 1). When considering models with same number of covariates, the BIC criterion does select the same model with one covariate (Latitude), two covariates (Latitude + Depth) and four covariates for the CL5 and CL7 algorithm. However the model Latitude + Depth + Temperature is selected

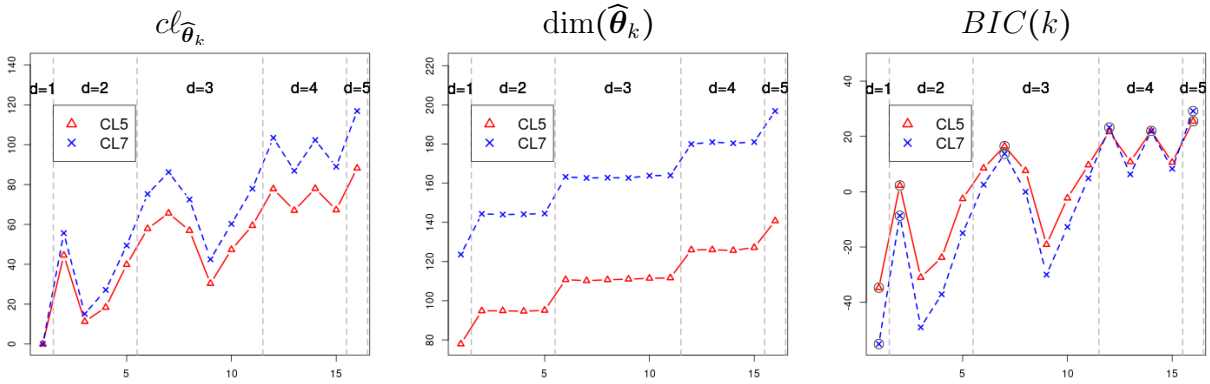

Figure 8: Model selection among the 16 possible models (x-axis) for the Barents sea full data set (p = 30 species) with the CL5 and CL7 algorithms. Dashed vertical lines: limits between models involving $d = 1, \dots 5$ covariates (including the intercept). Left: composite likelihood (subtracting the composite likelihood for Model 1); Center: dimension of the parameter, as defined in Equation (14); Right: BIC criterion (subtracting the composite likelihood for Model 1). CL5 algorithm: red triangle [△], CL7: blue 'times' sign [×] (the best model within each dimension is circled for each algorithm).

with CL5, whereas CL7 yields the model Latitude + Depth + Longitude. This differences are likely to result from the strong correlation between the Temperature and both the Latitude and the Longitude that is observed in the data set.

Overall, for both algorithms, the BIC criterion opts for the full model (Model 16), suggesting that all covariates should be kept in the model. This is consistent with the fact that, when considering a separate univariate PLN model for each of the p = 30 species, each of the four covariates turns out to be significant for at least one of the species, even after Bonferroni correction (not shown). We will therefore focus on the interpretation of the full model, in the next section.

Interpretation. A main interest of the algorithm which we propose here is to provide an accurate estimate of the variance of the parameter estimates of the PLN model. This enables the practitioner to determine (*i*) which environmental covariates have a significant effect on each species under study and (*ii*) which pairs of species display a correlation, that does not result from environmental variations.

Figure 9 displays an example of such results. The left panel shows a contrasted pattern as for the effect of local (that is, within the Barents sea) variations of the environmental conditions on the different fish species: some display large effect sizes (displayed on top), indicating a strong sensitivity to changes, whereas others seem to be poorly affected (displayed at bottom), suggesting a wider ecological niche.

The center and right panels of Figure 9 presents the estimate of the latent covariance matrix $\boldsymbol{\Sigma}$ (and of the associated correlation matrix $\text{cor}(\boldsymbol{\Sigma})$), where only significant terms are displayed. This results in a fairly sparse pattern, indicating that only a small fraction of species pairs (41 among 435) have correlated abundance variations, once accounted for the environmental effects.

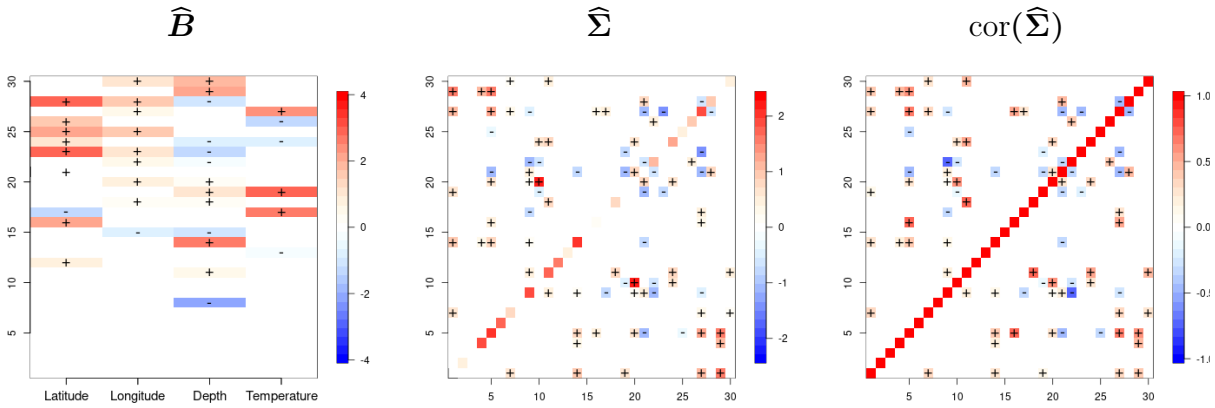

Figure 9: Colormap of the parameter estimates the Barents sea full data set ($p = 30$ species). Left: regression coefficients $\widehat{\beta}_{\ell j}$, center: covariance parameters $\widehat{\Sigma}_{jk}$, right: corresponding correlations $\widehat{\rho}_{jk} = \widehat{\Sigma}_{jk}/\sqrt{\widehat{\Sigma}_{jj}\widehat{\Sigma}_{kk}}$. Blank cells correspond to non-significant estimates, red cells (marked with '+') to positive estimates and blue cells (marked with '−') to negative estimates.

Code availability

The proposed algorithms have all been implemented in R [R Core Team, 2015] and C++, and will be included in the `PLNmodels` package [Chiquet et al., 2021] soon. For the time being, the codes are available from the authors upon request.

Acknowledgements

The authors thank Julien Chiquet (INRAE MIA-Paris-Saclay, France) and Mahendra Mariadassou (INRAE MaIAGE, France) for helpful discussions and advice. The authors are grateful to the INRAE MIGALE bioinformatics facility (MIGALE, INRAE, 2020. Migale bioinformatics Facility, doi: 10.15454/1.5572390655343293E12) for providing computing resources. The first author has been partly funded by the European Union (ERC-2022-SYG-OCEAN-101071601). Views and opinions expressed are however those of the author only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart. Importance Sampling: Intrinsic Dimension and Computational Cost. *Statistical Science*, 32(3):405–431, 2017.

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

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society: Series B (Methodological)*, 36(2):192–225, 1974.

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

J. G. Booth and J. P. Hobert. Maximizing generalized linear mixed model likelihoods with an automated Monte Carlo EM algorithm. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 61(1):265–285, 1999.

R. A. Boyles. On the Convergence of the EM Algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 45(1):47–50, 1983.

O. Cappé, R. Douc, A. Guillin, J.-M. Marin, and C. P. Robert. Adaptive importance sampling in general mixture classes. *Statistics and Computing*, 18(4):447–459, 2008.

S. Chatterjee and P. Diaconis. The sample size required in importance sampling. *The Annals of Applied Probability*, 28(2):1099–1135, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for probabilistic Poisson PCA. *The Annals of Applied Statistics*, 12(4):2674–2698, 2018.

J. Chiquet, M. Mariadassou, and S. Robin. Variational inference for sparse network reconstruction from count data. In *International Conference on Machine Learning*, 2019.

J. Chiquet, M. Mariadassou, and S. Robin. The Poisson-Lognormal Model as a Versatile Framework for the Joint Analysis of Species Abundances. *Frontiers in Ecology and Evolution*, 9: 188, 2021. doi: 10.3389/fevo.2021.588292. URL https://www.frontiersin.org/article/10.3389/fevo.2021.588292.

J.-M. Cornuet, J.-M. Marin, A. Mira, and C. P. Robert. Adaptive multiple importance sampling. *Scandinavian Journal of Statistics*, 39(4):798–812, 2012.

K. Daudel, R. Douc, and F. Portier. Infinite-dimensional gradient-based descent for alpha-divergence minimisation. *The Annals of Statistics*, 49(4):2250–2270, 2021.

K. Daudel, R. Douc, and F. Roueff. Monotonic Alpha-divergence Minimisation for Variational Inference. *Journal of Machine Learning Research*, 24(62):1–76, 2023.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum Likelihood from Incomplete Data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39:1–38, 1977.

B. Efron and C. Stein. The jackknife estimate of variance. *The Annals of Statistics*, pages 586–596, 1981.

G. Fort and E. Moulines. Convergence of the Monte Carlo expectation maximization for curved exponential families. *The Annals of Statistics*, 31(4):1220 – 1259, 2003.

M. Fossheim, E. M. Nilssen, and M. Aschan. Fish assemblages in the Barents Sea. *Marine Biology Research*, 2(4):260–269, 2006.

X. Gao and P. X.-K. Song. Composite likelihood bayesian information criteria for model selection in high-dimensional data. *Journal of the American Statistical Association*, 105(492): 1531–1540, 2010.

T. Jaakkola. *Advanced mean field methods: theory and practice*, chapter Tutorial on variational approximation methods, pages 129–160. MIT Press, 2001.

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

A. Korba and F. Portier. Adaptive Importance Sampling meets Mirror Descent : a Bias-variance Tradeoff. In *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pages 11503–11527. PMLR, 2022.

R. A. Levine and G. Casella. Implementations of the Monte Carlo EM Algorithm. *Journal of Computational and Graphical Statistics*, 10(3):422–439, 2001.

B. Lindsay. Composite likelihood methods. *Contemporary mathematics*, 80(1):221–239, 1988.

T. A. Louis. Finding the observed information matrix when using the EM algorithm. *Journal of the Royal Statistical Society. Series B*, pages 226–233, 1982.

C. E. McCulloch. Maximum Likelihood Algorithms for Generalized Linear Mixed Models. *Journal of the American Statistical Association*, 92(437):162–170, 1997.

L. Middleton, G. Deligiannidis, A. Doucet, and P. E. Jacob. Unbiased Smoothing using Particle Independent Metropolis-Hastings. In K. Chaudhuri and M. Sugiyama, editors, *Proceedings of the Twenty-Second International Conference on Artificial Intelligence and Statistics*, volume 89 of *Proceedings of Machine Learning Research*, pages 2378–2387. PMLR, 2019.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015. URL https://www.R-project.org/.

A. van der Vaart. *Asymptotic statistics*, volume 27 of *Cambridge Series in Statistical and Probabilistic Mathematics*. Cambridge Univ. Press, New York, 1998.

C. Varin and P. Vidoni. A note on composite likelihood inference and model selection. *Biometrika*, 92(3):519–528, 2005.

C. Varin, N. Reid, and D. Firth. An Overview of Composite Likelihood Methods. *Statistica Sinica*, 21:5–42, 2011.

M. J. Wainwright and M. I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Found. Trends Mach. Learn.*, 1(1–2):1–305, 2008.

G. C. G. Wei and M. A. Tanner. A Monte Carlo Implementation of the EM Algorithm and the Poor Man's Data Augmentation Algorithms. *Journal of the American Statistical Association*, 85(411):699–704, 1990.

C. Wu. On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11(1): 95–103, 1983.

X. Xu, N. Reid, and L. Xu. Note on information bias and efficiency of composite likelihood. Technical Report 1612.06967, arXiv, 2016.

Y. Zhao and H. Joe. Composite likelihood estimation in multivariate data analysis. *Canadian Journal of Statistics*, 33(3):335–356, 2005.

A Appendix

A.1 Proofs

A.1.1 Proof of Proposition 1

As $\theta^{(h+1)}$ is the maximizer of $\sum_{b=1}^{B} \lambda_b \mathbb{E}_{\theta^{(h)}}\left[\log p_\theta(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \mid \mathbf{Y}^{(b)}\right]$, we have that

$$
\begin{aligned}
0 &\leq \sum_{b=1}^{B} \lambda_b \mathbb{E}_{\theta^{(h)}}\left[\log p_{\theta^{(h+1)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \,\middle|\, \mathbf{Y}^{(b)}\right] - \sum_{b=1}^{B} \lambda_b \mathbb{E}_{\theta^{(h)}}\left[\log p_{\theta^{(h)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \,\middle|\, \mathbf{Y}^{(b)}\right] \\
&= \sum_{b=1}^{B} \lambda_b \mathbb{E}_{\theta^{(h)}}\left[\log \frac{p_{\theta^{(h+1)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})}{p_{\theta^{(h)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})} \,\middle|\, \mathbf{Y}^{(b)}\right] \leq \sum_{b=1}^{B} \lambda_b \log \mathbb{E}_{\theta^{(h)}}\left[\frac{p_{\theta^{(h+1)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})}{p_{\theta^{(h)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})} \,\middle|\, \mathbf{Y}^{(b)}\right]
\end{aligned}
$$

using Jensen's inequality, so

$$
\begin{aligned}
0 &\leq \sum_{b=1}^{B} \lambda_b \log \int_{\mathbb{R}^k} p_{\theta^{(h)}}(\mathbf{Z}^{(b)} \mid \mathbf{Y}^{(b)}) \frac{p_{\theta^{(h+1)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})}{p_{\theta^{(h)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)})} \mathrm{d}\mathbf{Z}^{(b)} \\
&= \sum_{b=1}^{B} \lambda_b \log \left\{ \frac{1}{p_{\theta^{(h)}}(\mathbf{Y}^{(b)})} \int_{\mathbb{R}^k} p_{\theta^{(h+1)}}(\mathbf{Y}^{(b)}, \mathbf{Z}^{(b)}) \mathrm{d}\mathbf{Z}^{(b)} \right\} = c\ell(\theta^{(h+1)}) - c\ell(\theta^{(h)}). \qquad \blacksquare
\end{aligned}
$$

This proof relies on the decomposition given in (10). One may observe that it also holds for the decomposition (9), replacing $\mathbf{Z}^{(b)}$ with $\mathbf{Z}$. As a consequence, replacing $\mathbf{Z}^{(b)}$ with $\mathbf{Z}$ in Algorithm 2 would yield the same property, although this alternative algorithm would not bring any computational advantage.

A.1.2 Proof of Proposition 3

(i) For all $\mathbf{v} \in \mathbb{R}^p$,

$$
\begin{aligned}
\frac{p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y}_i, \mathbf{v})^2}{\varphi(\mathbf{v}; m, S)} &= \frac{|S|^{1/2}}{(2\pi)^{p/2}|\Sigma|} \exp\left\{-\frac{1}{2}\mathbf{v}^\top \left(2\Sigma^{-1} - S^{-1}\right)\mathbf{v}\right\} \\
&\quad \times \exp\left[2\sum_{j=1}^{p}\left\{Y_{ij}(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(h)} + v_j) - \exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(h)} + v_j)\right\}\right].
\end{aligned}
$$

Thus the latter is finite if the quadratic terms satisfy for all $\mathbf{v} \in \mathbb{R}^p$,

$$
\mathbf{v}^\top \left(2\Sigma^{-1} - S^{-1}\right)\mathbf{v} \geq 0.
$$

The result follows.
(ii) For all $\mathbf{v} \in \mathbb{R}^p$,

$$
\begin{aligned}
\frac{p_{\boldsymbol{\theta}}(\mathbf{Y}_i, \mathbf{v})}{\varphi(\mathbf{v}\,; m, S)} &= \frac{|S|^{1/2}}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\mathbf{v}^\top \left(\Sigma^{-1} - S^{-1}\right)\mathbf{v}\right\} \\
&\quad \times \exp\left[\sum_{j=1}^{p}\left\{Y_{ij}(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(h)} + v_j) - \exp(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j^{(h)} + v_j)\right\}\right]
\end{aligned} \tag{23}
$$

.

We thus have a finite upper bound if the quadratic terms satisfy for all $\mathbf{v} \in \mathbb{R}^p$,

$$
\mathbf{v}^\top \left(\Sigma^{-1} - S^{-1}\right)\mathbf{v} \geq 0.
$$

A.1.3 Proof of Proposition 4

For all $\mathbf{v} \in \mathbb{R}^p$,

$$q_i^{(h)}(\mathbf{v}) \geq (1-\alpha)\varphi(\mathbf{v}; \mathbf{m}_i^{(h)}, \boldsymbol{\Sigma}^{(h)})$$

and then,

$$\frac{p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y}_i, \mathbf{v})^2}{q_i^{(h)}(\mathbf{v})} \leq \frac{1}{(1-\alpha)} \frac{p_{\boldsymbol{\theta}^{(h)}}(\mathbf{Y}_i, \mathbf{v})^2}{\varphi(\mathbf{v}; \mathbf{m}_i^{(h)}, \boldsymbol{\Sigma}^{(h)})}.$$

The result follows from statement *(i)* of Proposition 3, taking $S = \boldsymbol{\Sigma} = \boldsymbol{\Sigma}^{(h)}$.

A.2 Update formulas (M step)

Objective function.

$$Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) = \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mid \mathbf{Y}] = \sum_{i=1}^{n} \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\log p_{\boldsymbol{\Sigma}}(\mathbf{Z}_i) + \log p_{\boldsymbol{B}}(\mathbf{Y}_i \mid \mathbf{Z}_i) \mid \mathbf{Y}_i].$$

The above decomposition allows to perform optimization in $\boldsymbol{B}$ and $\boldsymbol{\Sigma}$ separately.

Update for $\boldsymbol{B}$. For all $(i,j) \in \{1,\dots,n\} \times \{1,\dots,p\}$

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{\beta}_j} \log p_{\boldsymbol{B}}(\mathbf{Y}_i \mid \mathbf{Z}_i) &= \frac{\partial}{\partial \boldsymbol{\beta}_j} \left\{ (\mathbf{o}_i + \boldsymbol{B}^\top \mathbf{x}_i + \mathbf{Z}_i)^\top \mathbf{Y}_i - \sum_{k=1}^{p} \exp\left(o_{ik} + \mathbf{x}_i^\top \boldsymbol{\beta}_k + Z_{ik}\right) \right\} \\ &= \left\{ \mathbf{Y}_{ij} - \exp\left(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j + Z_{ij}\right) \right\} \mathbf{x}_i.\end{aligned}$$

The gradient in $\boldsymbol{\beta}_j$ thus writes as

$$\nabla^{(h)}(\boldsymbol{\beta}_j) \triangleq \frac{\partial}{\partial \boldsymbol{\beta}_j} Q(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)}) = \sum_{i=1}^{n} \left\{ \mathbf{Y}_{ij} - \exp\left(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j\right) \mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\exp\left(Z_{ij}\right) \mid \mathbf{Y}_{ij}] \right\} \mathbf{x}_i$$

and its self-normalised importance sampling estimate for a N-sample $(\mathbf{V}_{i1}, \dots, \mathbf{V}_{iN})$ from $q_i^{(h)}$

$$\begin{aligned}\widehat{\nabla}^{(h)}(\boldsymbol{\beta}_j) &= \sum_{i=1}^{n} \left\{ \mathbf{Y}_{ij} - \exp\left(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j\right) \widehat{\mathbb{E}}_{q_i^{(h)}}[\exp\left(Z_{ij}\right)] \right\} \mathbf{x}_i \\ &= \sum_{i=1}^{n} \left\{ \mathbf{Y}_{ij} - \exp\left(o_{ij} + \mathbf{x}_i^\top \boldsymbol{\beta}_j\right) \sum_{r=1}^{N} w_{ir}^{(h)} \exp\left(V_{irj}\right) \right\} \mathbf{x}_i\end{aligned}$$

Update for $\boldsymbol{\Sigma}$. For all $(i,j,k) \in \{1,\dots,n\} \times \{1,\dots,p\} \times \{1,\dots,p\}$,

$$\frac{\partial}{\partial \Sigma_{jk}} \log p_{\boldsymbol{\Sigma}}(\mathbf{Z}_i) = -\frac{1}{2} \frac{\partial}{\partial \Sigma_{jk}} \left(\log |\boldsymbol{\Sigma}| + \mathbf{Z}_i^\top \boldsymbol{\Sigma} \mathbf{Z}_i \right).$$

Denote E_{jk} the matrix unit with value 1 at coefficient (j,k) and δ_{jk} the Kronecker delta. Since $\boldsymbol{\Sigma}$ is symmetric,

$$\frac{\partial}{\partial \Sigma_{jk}} \log |\boldsymbol{\Sigma}| = \operatorname{tr}\left(\boldsymbol{\Sigma}^{-1} \frac{\partial}{\partial \Sigma_{jk}} \boldsymbol{\Sigma} \right) = \operatorname{tr}(\boldsymbol{\Sigma}^{-1} E_{jk}) + (1-\delta_{jk}) \operatorname{tr}(\boldsymbol{\Sigma}^{-1} E_{kj}) = (2-\delta_{jk}) \Sigma^{-1}_{jk}.$$

Furthermore, again using that $\boldsymbol{\Sigma}$ is symmetric

$$\begin{aligned}\frac{\partial}{\partial \Sigma_{jk}} \mathbf{Z}_i^\top \boldsymbol{\Sigma}^{-1} \mathbf{Z}_i &= \sum_{u=1}^{p} Z_{iu} \sum_{v=1}^{p} Z_{iv} \frac{\partial}{\partial \Sigma_{jk}} \Sigma^{-1}_{uv} \\ &= -\frac{2-\delta_{jk}}{2} \sum_{u=1}^{p} Z_{iu} \sum_{v=1}^{p} Z_{iv} (\Sigma^{-1}_{uj} \Sigma^{-1}_{kv} + \Sigma^{-1}_{uk} \Sigma^{-1}_{jv}) \\ &= (\delta_{jk} - 2) \sum_{u=1}^{p} Z_{iu} \Sigma^{-1}_{uj} \sum_{v=1}^{p} Z_{iv} \Sigma^{-1}_{kv} \\ &= (\delta_{jk} - 2) \left[\Sigma^{-1} Z_i Z_i^\top \Sigma^{-1} \right]_{jk}.\end{aligned}$$

The update for $\boldsymbol{\Sigma}$ is then solution of

$$\frac{\delta_{jk}-2}{2}\left[n\boldsymbol{\Sigma}^{-1}-\boldsymbol{\Sigma}^{-1}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\mathbf{Z}_i\mathbf{Z}_i^\top\mid\mathbf{Y}_i]\boldsymbol{\Sigma}^{-1}\right]_{jk}=0.$$

This leads to, for the same N-sample $(\mathbf{V}_{i1},\ldots,\mathbf{V}_{iN})$ from $q_i^{(h)}$

$$\boldsymbol{\Sigma}^{(h+1)}=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(h)}}[\mathbf{Z}_i\mathbf{Z}_i^\top\mid\mathbf{Y}_i]\quad\text{and}\quad\widehat{\boldsymbol{\Sigma}}^{(h+1)}=\frac{1}{n}\sum_{i=1}^{n}\sum_{r=1}^{N}w_{ir}^{(h)}\mathbf{V}_{ir}\mathbf{V}_{ir}^\top.$$

A.3 Update formulas (CL-M step)

Objective function.

$$\begin{aligned}Q_{c\ell}(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(h)})&=\sum_{b=1}^{C}\lambda_b\mathbb{E}_{\boldsymbol{\theta}^{(b,h)}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}^{(b)},\mathbf{Z}^{(b)})\;\middle|\;\mathbf{Y}^{(b)}\right]\\&=\sum_{b=1}^{C}\lambda_b\sum_{i=1}^{n}\mathbb{E}_{\boldsymbol{\theta}^{(b,h)}}\left[\log p_{\boldsymbol{\Sigma}^{(b)}}(\mathbf{Z}_i^{(b)})+\log p_{\boldsymbol{B}^{(b)}}(\mathbf{Y}_i^{(b)}\mid\mathbf{Z}_i^{(b)})\;\middle|\;\mathbf{Y}_i^{(b)}\right].\end{aligned}$$

Update for $\boldsymbol{B}$. For all $(i,j,b)\in\{1,\ldots,n\}\times\{1,\ldots,p\}\times\{1,\ldots,C\}$,

$$\frac{\partial}{\partial\boldsymbol{\beta}_j}\log p_{\boldsymbol{B}^{(b)}}(\mathbf{Y}_i^{(b)}\mid\mathbf{Z}_i^{(b)})=\begin{cases}\left\{\mathbf{Y}_{ij}-\exp\left(o_{ij}+\mathbf{x}_i^\top\boldsymbol{\beta}_j+Z_{ij}^{(b)}\right)\right\}\mathbf{x}_i, & \text{if } j\in\mathcal{C}_b,\\ \mathbf{0}_d, & \text{otherwise.}\end{cases}$$

If we denote $\mathcal{C}(j)=\{b\in\{1,\ldots,C\}\mid j\in\mathcal{C}_b\}$ the set of indices of blocks that contain j, the gradient in $\boldsymbol{\beta}_j$ writes as

$$\begin{aligned}\nabla^{(h)}(\boldsymbol{\beta}_j)&\triangleq\frac{\partial}{\partial\boldsymbol{\beta}_j}Q_{c\ell}(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(h)})\\&=\sum_{i=1}^{n}\left\{\mathbf{Y}_{ij}-\exp\left(o_{ij}+\mathbf{x}_i^\top\boldsymbol{\beta}_j\right)\left(\sum_{b\in\mathcal{C}(j)}\lambda_b\mathbb{E}_{\boldsymbol{\theta}^{(h)}}\left[\exp\left(Z_{ij}^{(b)}\right)\;\middle|\;\mathbf{Y}_{ij}\right]\right)\right\}\mathbf{x}_i\end{aligned}$$

and its self-normalised importance sampling estimate for N-samples $(\mathbf{V}_{i1}^{(b)},\ldots,\mathbf{V}_{iN}^{(b)})$ from $q_i^{(b,h)}$

$$\begin{aligned}\widehat{\nabla}^{(h)}(\boldsymbol{\beta}_j)&=\sum_{i=1}^{n}\left\{\mathbf{Y}_{ij}-\exp\left(o_{ij}+\mathbf{x}_i^\top\boldsymbol{\beta}_j\right)\left(\sum_{b\in\mathcal{C}(j)}\lambda_b\widehat{\mathbb{E}}_{q_i^{(b,h)}}\left[\exp\left(Z_{ij}^{(b)}\right)\right]\right)\right\}\mathbf{x}_i\\&=\sum_{i=1}^{n}\left[\mathbf{Y}_{ij}-\exp\left(o_{ij}+\mathbf{x}_i^\top\boldsymbol{\beta}_j\right)\left\{\sum_{b\in\mathcal{C}(j)}\lambda_b\sum_{r=1}^{N}w_{ir}^{(b,h)}\exp\left(V_{irj}^{(b)}\right)\right\}\right]\mathbf{x}_i\end{aligned}$$

Update for $\boldsymbol{\Sigma}$. Unlike the full data framework we do not have access to an estimate of Σ and should resort as well to a gradient based method. Denote $\Omega^{(b)}$ the inverse matrix of $\Sigma^{(b)}$ and for $(j,k)\in\{1,\ldots,p\}^2$, $\mathcal{C}(j,k)=\{b\in\{1,\ldots,C\}\mid j,k\in\mathcal{C}_b\}$

$$\frac{\partial}{\partial\Sigma_{jk}}Q_{c\ell}(\boldsymbol{\theta}\mid\boldsymbol{\theta}^{(h)})=\sum_{b\in\mathcal{C}(j,k)}\frac{(2-\delta_{jk})\lambda_b}{2}\left[\Omega^{(b)}\sum_{i=1}^{n}\mathbb{E}\left[\mathbf{Z}_i^{(b)}\mathbf{Z}_i^{(b)\top}\right]\Omega^{(b)}-n\Omega^{(b)}\right]_{jk}^{\blacksquare}$$

where $[A]^{\blacksquare}$ stands for a $k\times k$ matrix A that has been embedded into a $p\times p$ matrix M with M_{jk} taking value 0 if the pair (j,k) does not belong to the possible pair of species of block $\mathcal{C}_b$

and being equal to the coefficient of A related to the pair (j,k) otherwise. Each term of the gradient in $\boldsymbol{\Sigma}$ can thus be estimated using self-normalised importance estimates

$$\begin{aligned}\frac{\partial}{\partial \Sigma_{jk}} \widehat{Q_{c\ell}(\boldsymbol{\theta} \mid \boldsymbol{\theta}^{(h)})} &= \sum_{b \in \mathcal{C}(j,k)} \frac{(2-\delta_{jk})\lambda_b}{2} \left[\Omega^{(b)} \sum_{i=1}^{n} \widehat{\mathbb{E}}_{q_i^{(b,h)}} \left[\mathbf{Z}_i^{(b)} \mathbf{Z}_i^{(b)\top} \right] \Omega^{(b)} - n\Omega^{(b)} \right]_{jk}^{\blacksquare} \\ &= \sum_{b \in \mathcal{C}(j,k)} \frac{(2-\delta_{jk})\lambda_b}{2} \left[\Omega^{(b)} \sum_{i=1}^{n} \left(\sum_{r=1}^{N} w_i^{(b,r)} \mathbf{V}_{ir}^{(b)} \mathbf{V}_{ir}^{(b)\top} \right) \Omega^{(b)} - n\Omega^{(b)} \right]_{jk}^{\blacksquare}\end{aligned}$$

A.4 Additional simulation results

A.4.1 Full likelihood inference

We present here additional results regarding the distribution of the full likelihood estimates of the regression coefficients $\beta_{\ell j}$.

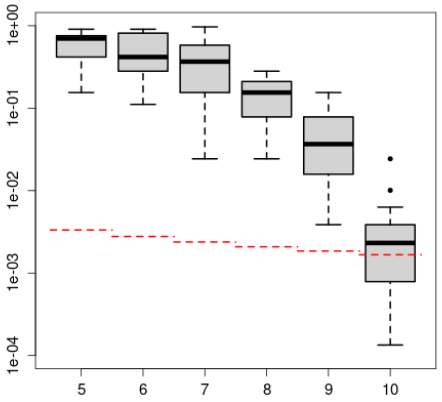

Figure 10: Distribution of the p-values of the Kolmogorov-Smirnov test for the distribution of the standardized estimates $\widetilde{\beta}_{\ell j}$ over the $M = 100$ simulations as a function of the number of species ($p = 5, \ldots, 10$) for full-likelihood inference (FL). Dotted red lines: $\alpha = 5\%$ significance threshold after Bonferronni correction (*i.e.*, $\alpha/(dp)$).

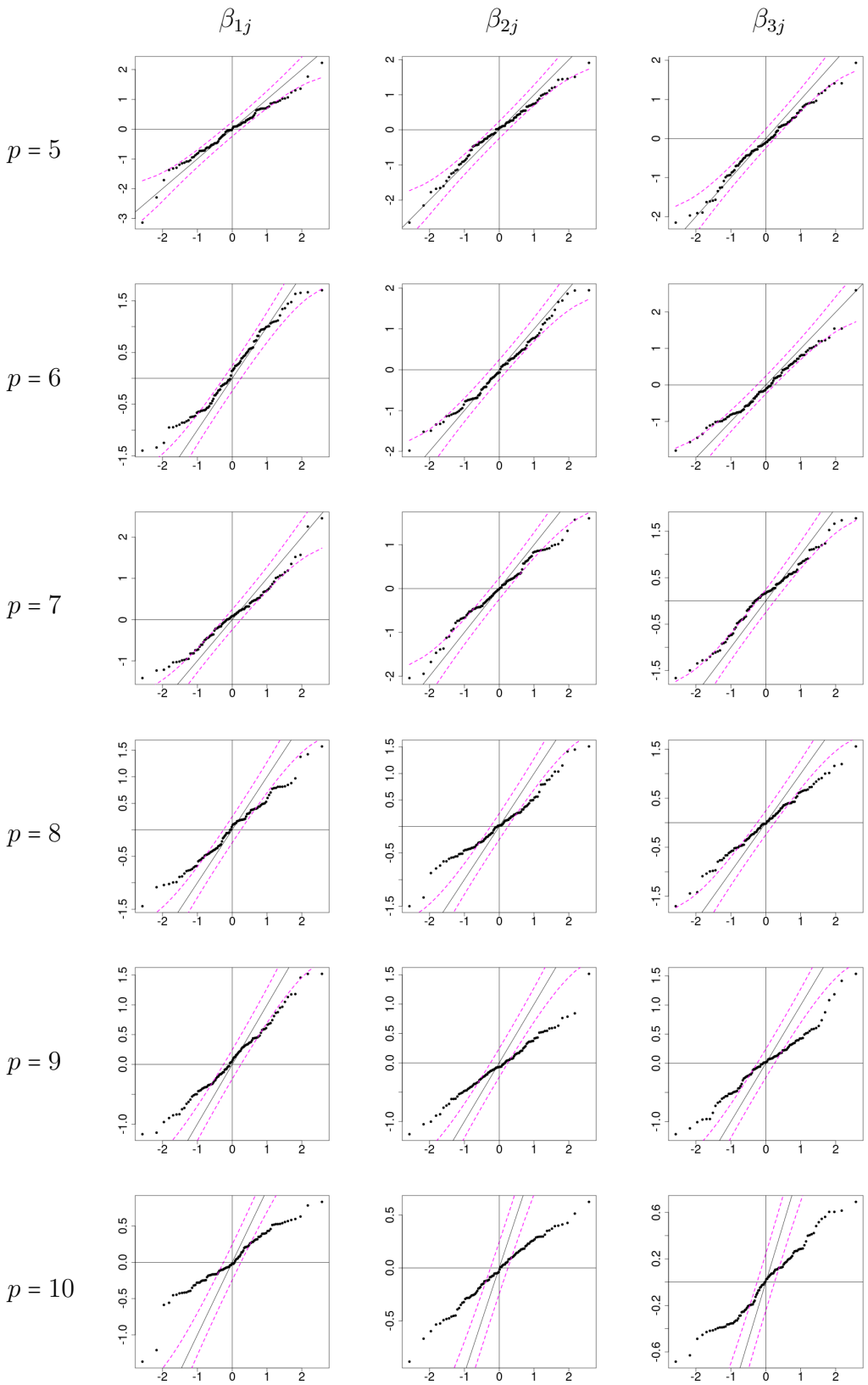

Figure 11: qq-plots of the normalized regression coefficient $\widetilde{\beta}_{\ell j}$ defined in Equation (22) for the second simulated species for each of the $d = 3$ covariates with full likelihood (FL) with $p = 5, \ldots, 10$ species. Same legend as Figure 3.

A.4.2 Composite likelihood inference

This section provides additional results regarding the distribution of the composite likelihood estimates of the regression coefficients $\beta_{\ell j}$.

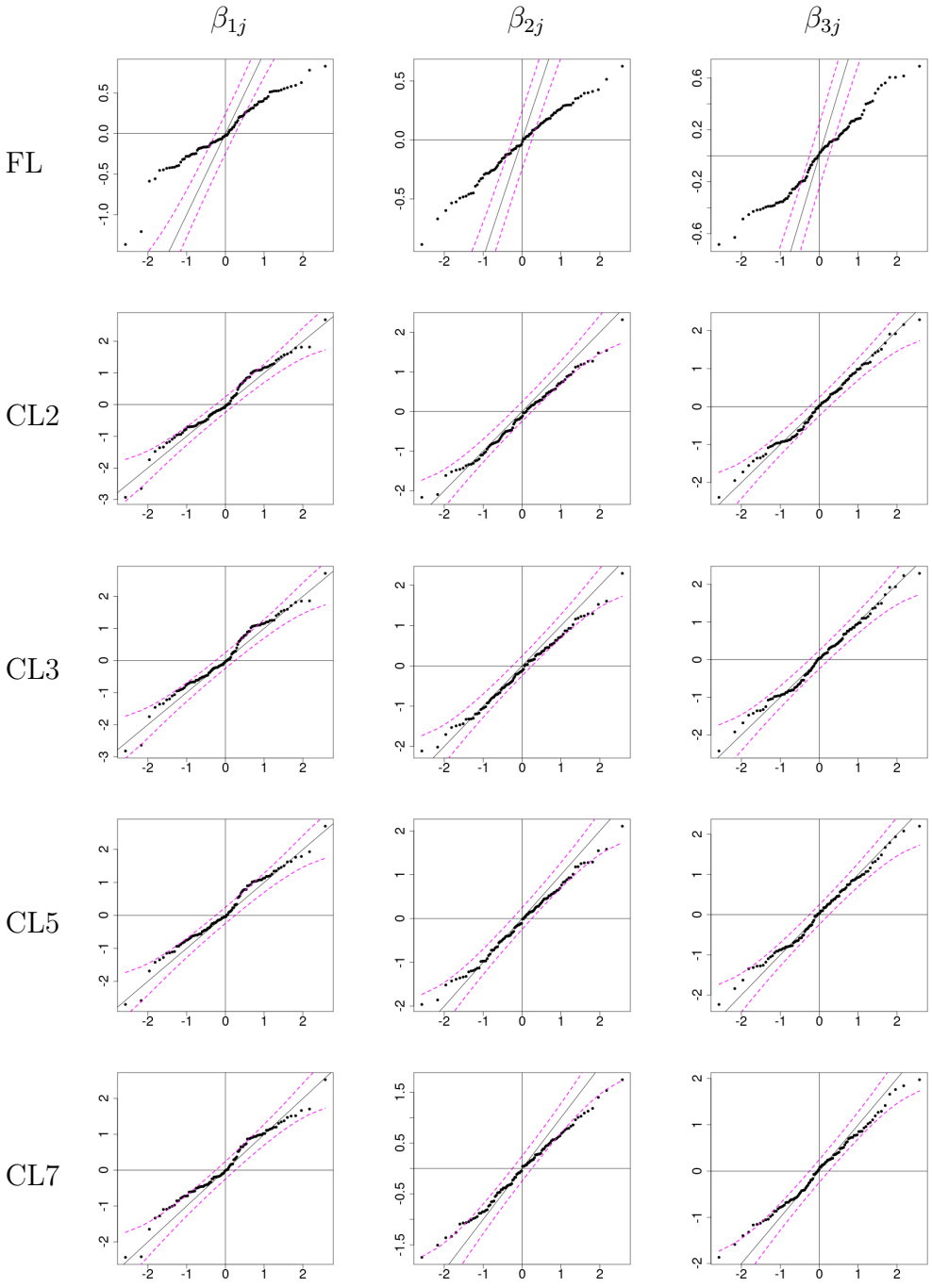

Figure 12: qq-plots of the normalized regression coefficient $\widetilde{\beta}_{\ell j}$ defined in Equation (22) for the second simulated species for each of the $d = 3$ covariates for $p = 10$ species with full likelihood (FL) or composite likelihood (CL) with blocks of size $k = 2, 3, 5, 7$. Same legend as Figure 3.