Latent Tree based Inference of Ecological Network using the Poisson Log-Normal Model

Acknowledgments

Our sincere thanks go to Corinne Vacher, who gave us access to experimental data on oak trees and the opportunity to test our model on a concrete issue.

I. Introduction

In the past decade, ecological networks have become a key tool to describe interactions between species and better understand the dynamics of a whole ecosystem or anticipate its response to a given change. Such interaction networks can be inferred based on the observation of the respective abundance of each species. Metagenomics relies on Next-Generation Sequencing (NGS) technologies to evaluate the (relative) abundance of microbial species in a given medium under varying experimental conditions or across replicates. A typical metabarcoding experiment results in a vector of read counts associated with each species under study.

From a statistical perspective, network inference is usually considered in the framework of probabilistic graphical models. A huge statistical literature exists about this problem in the Gaussian case, that is when the data consists in continuous observations. These methods need to be adapted to count data.

In this work, we propose a comprehensive statistical framework for the inference of ecological networks based on metagenomic counts. To this aim, we use the Poisson log-normal (PLN) model which provides a generic description for multivariate count data. The PLN model accounts for the specificities of metagenomic data such as over-dispersion or sequencing depth heterogeneity. More importantly, the PLN model allows to correct for the effect of covariates, which is critical to avoid the detection of spurious edges in the graph.

II. Model

PLN model. The negative binomial distribution has become the reference distribution for the analysis of NGS read counts. This distribution is also known as the Poisson-Gamma distribution as it consists in a Poisson distribution combined with a latent Gamma layer. Unfortunately, this model does not generalizes easily to multivariate count data as no generic version of the Gamma distribution exists. The Poisson log-normal distribution is similar to the Poisson-Gamma distribution, except that the latent Gamma layer is replaced with a latent log-normal layer. This PLN distribution displays the same over-dispersion feature as the Poisson-Gamma but generalizes easily to multivariate data via the multivariate normal distribution (Aitchison and Ho, 1989). The model can be described as follows: for each observation, a random Gaussian vector with as many dimensions as

species is first drawn; each observed count is then drawn conditionally on the corresponding coordinate of the latent unobserved Gaussian vector. The dependency between the counts is therefore encoded in the covariance matrix of the latent Gaussian vector. An important feature is that, as opposed to other multivariate count distributions (Inoue & al, 2018), the correlations between species abundances can be either positive or negative, preserving the sign of the terms of the Gaussian covariance matrix.

Graphical models. A graphical model is a graphical representation of the dependency structure between a set of variables. Briefly speaking, an edge is drawn between two variables if the dependence between them does not result from the effect of the other variables. In our example, the variables are the respective species abundances and two species are connected if they are in direct interaction. One major advantage of the PLN model is that it can take advantage of the methods that were developed for network inference in the framework of Gaussian Graphical Models (GGM). Our idea is to define the ecological network as the graphical model of the Gaussian latent layer of the PLN model.

Tree-based network inference. All network inference methods have to face the fact that the number of possible network grows super-exponentially with the number of species. This makes the exhaustive exploration of the set of all possible graphs combinatorially intractable. To circumvent this problem, we choose to model the network as a random sample in the set of spanning trees. This assumption is consistent with the expectation that ecological networks are sparse. It also allows us to take advantage of combinatorial results about the optimization or the summation over the whole set of spanning trees.

Proposed model. Put together, the statistical model we present is a hierarchical model composed of two layers of hidden parameters:

- the dependence tree of the Gaussian layer of the PLN model,
- the Gaussian layer of the PLN model itself.

III. Inference

Because of the presence of Gaussian latent layer, the PLN model is an incomplete data model for which the Expectation-Maximization (EM) algorithm could be considered. Unfortunately, the conditional distribution of the hidden layer given the observed data is intractable so the EM algorithm does not apply directly. However, a proxy of this distribution can be obtained using variational techniques (Wainwright & Jordan, 2008). This results in a Variational EM (VEM) that has been implemented in the ‘PLNmodels' R-package available on github (https://github.com/jchiquet/PLNmodels, Chiquet, Mariadassou & Robin, 2018).

The inference of the PLN models provides an estimate of the covariance matrix of the Gaussian layer. Hence we are brought to a network inference problem in the GGM context, where a usual method is the Graphical LASSO (Glasso). This penalized approach allows for a sparse inference.

As explained in Section II, we adopt a different approach, assuming the graphical model is drawn in the set of spanning trees. This model is similar to the mixture of tree-shaped graphical models considered by (Meila & Jaakola, 2006). The set of spanning trees displays several interesting combinatorial features, which makes maximization (Chow & Liu, 1968) or summation (Chaiken & al, 1978) achievable in polynomial time. Observe that the mixture assumption widens the range of graphs we are able to infer as it allows the presence of cycles and cliques.

Because this second layer of the model is a mixture, its inference can be carried out via an EM algorithm. Part of our contribution is to develop a new EM algorithm along which the conditional distribution of the tree given the data is computed. Unlike what is usually found in the literature, the conditional probability given the data for each edge to be part of the graphical model is updated, and not considered fixed. Once the conditional probability of each edge is computed, the inferred graph is defined by the most probable edges.

IV. Simulation

We tested our method with several dependence structures and several densities of edges. In addition to spanning trees, we considered Erdös structures, which are random graphs, scale-free structures which are rather sparse and clusters. The latter are very different from the other structures and should be challenging for our method. The number of vertices in the graph has been set between 10 and 30, edge probability varies between 0.025 and 0.25 and the number of observations between 20 and 100.

Considering the original graph as ground truth, our approach allows the inference of a family of nested graphs derived from the thresholding of the estimated conditional edges probabilities. The Area Under the Receiver Operating Curve (AUC) is used as a summary measure of the graph reconstruction quality. The AUC of our method was compared to that of the glasso for all settings and dependence structures. In a specific experiment, inferences are done on 40 different graphs.

Our method performs well on trees, and as expected is less efficient on the other cases but it is still comparable to or better than the glasso. In all tested settings, in terms of median of AUC our method is about 5% above glasso with trees, about 3% in the scale-free structure and only by 1% in the cluster. On the Erdös structure the two methods perform identically. The medians of AUC are around 80% when only the number of vertices varies, however they increase significantly with the number of observations : 62% for 20 observation, and 85% for 100 with our method.

V. Illustration

The fungal Erysiphe alphitoides (EA) is the causal agent of oak powdery mildew. Jakushkin et al (2016) study its pathobiome via microbial network inference and emphasize the importance of covariates. The sampling of oak leaves microbiome was done on three different oaks with different infection status. The corresponding data table is composed of 116 samples of 94 species of fungi and bacteria of oak leaves, including the EA agent.

Several covariates are available, among which the tree identifier, the distance from the leaf to the tree base, and a measure of infection.
Relative species abundances were evaluated by metabarcoding, for which it is necessary to correct for depth of coverage. Treating the later as an offset, we fitted four PLN regression models (all including offsets) on these data, including covariates one by one. They are nested and take an additional variable among those previously mentioned.
For each of the four models, we computed the conditional probabilities of edges to be part of the network. To define the threshold above which an edge is included in the network, we evaluated the overall proportion of absent edges using a multiple testing technique proposed by Storey (2002).
As expected, the more covariates are included in the model, the less edges are inferred in the corresponding network, underlying the benefits of taking covariates into account. Edges removed at each step can be interpreted as spurious edges from the preceding step that were actually reflecting the effect of the included covariate. The model only adjusting for the offset contains 2630 edges, whereas the one with four covariates has 2300 edges. Between these two models all nodes lose 7 connexions on average. Regarding the pathogen EA across all models, its major role in the organization of the pathobiome is proven by its degree remaining stable at about 60 (59, 64, 63 and 60 respectively).

VI. Discussion

We provide a comprehensive statistical framework for the inference of ecological networks based on NGS read counts, which includes a formal probabilistic model and the associated estimation algorithm. Our model infers interaction networks and easily adapts to different experimental conditions by enabling the user to account for offsets and covariates.

Our final algorithm uses successively a VEM algorithm for the PLN and an EM algorithm for the inference of the tree structure. The latent layer of the PLN is first inferred using the PLNmodels package, then the EM algorithm infers the network. A technical perspective consists in building an algorithm which encompasses our EM algorithm in the M step of the VEM algorithm, within the PLNmodels package.

Identifying an estimator for the number of edges in the final graph is crucial to the network inference. The multiple testing heuristic we developed to estimate the overall density of the network seems to work well in practice, yet its reliability needs to be further investigated.

Finally, a challenging issue for network inference is the possibility that some species ~~of or~~ covariate having a strong impact on the ecosystem was not measured, resulting in spurious edges (see the illustration section). The automatic detection and estimation of such missing variable can be considered in the context of tree-shaped graphs.

Bibliography

AITCHISON, John et HO, C. H. The multivariate Poisson-log normal distribution. *Biometrika*, 1989, vol. 76, no 4, p. 643-653.

CHAIKEN, Seth et KLEITMAN, Daniel J. Matrix tree theorems. *Journal of combinatorial theory, Series A*, 1978, vol. 24, no 3, p. 377-381.

CHIQUET, Julien, MARIADASSOU, Mahendra, et ROBIN, Stéphane. Variational inference for probabilistic Poisson PCA. *arXiv preprint arXiv:1703.06633*, 2017.

CHOW, C. et LIU, Cong. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 1968, vol. 14, no 3, p. 462-467.

INOUYE, David I., YANG, Eunho, ALLEN, Genevera I., *et al.* A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2017, vol. 9, no 3.

JAKUSCHKIN, Boris, FIEVET, Virgil, SCHWALLER, Loïc, et al. Microbial ecology, 2016, vol. 72, no 4, p. 870-880.

MEILĂ, Marina et JAAKKOLA, Tommi. Tractable Bayesian learning of tree belief networks. *Statistics and Computing*, 2006, vol. 16, no 1, p. 77-92.

STOREY, John D. A direct approach to false discovery rates. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 2002, vol. 64, no 3, p. 479-498.

WAINWRIGHT, Martin J., JORDAN, Michael I., *et al.* Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 2008, vol. 1, no 1–2, p. 1-305.