Variational Inference of Sparse Network from Count Data

Julien Chiquet [1] **Mahendra Mariadassou** [2] **Stéphane Robin** [1]

Abstract

The problem of network reconstruction from continuous data has been extensively studied and most state of the art methods rely on variants of Gaussian Graphical Models (GGM). GGM are unfortunately badly suited to sparse count data spanning several orders of magnitude. Most inference methods for count data (SparCC, REBACCA, SPIEC-EASI, gCoda, etc) first transform counts to pseudo-Gaussian observations before using GGM. We rely instead on a Poisson-LogNormal (PLN) model where counts follow Poisson distributions with parameters sampled from a latent multivariate Gaussian variable, and infer the network in the latent space using a variational inference procedure. This model allows us to (i) control for confounding covariates and differences in sampling efforts and (ii) integrate data sets from different origins. It is also competitive in terms of speed and accuracy with state of the art methods.

1. Introduction

Networks are the *de facto* mathematical object used to model and represent pairwise interactions between entities of interest. Examples include air traffic between airports, social interactions between participants of a conference, trophic relationships between species, gene regulations, ecological interactions between microbial species, etc. However, most networks are not observed directly but must be reconstructed first from indirect node-level observations using some kind of statistical procedure. In this perspective, graphical models are popular among statisticians to explore relationships between nodes in graphs since undirected graphical models (Lauritzen, 1996), also called Markov random fields (Harris, 2016), are a convenient class of models with sound theoretical groundings for capturing conditional dependence relationships between nodes: i and j are linked in $\mathbf{G}$ (noted $i \sim j$) if and only if features i and j are conditionally dependent given all the others. Powerful inference procedures exist for Gaussian Graphical Models (GGM) for continuous data and Ising or voter models for binary data and the research field is still active. An informative and non-exhaustive set of seminal papers in this field may include Yuan & Lin (2007); Banerjee et al. (2008); Ravikumar et al. (2010); Meinshausen & Bühlmann (2006). GGM have been successfully used to understand complex genetic regulations (Moignard et al., 2015; Fiers et al., 2018), to identify direct contacts between protein subunits (Drew et al., 2017) or to identify functional pathways associated to a disease (Yu et al., 2015). Unfortunately, we lack similar powerful procedures for count data, which are the focus of this work.

Count data arise naturally in fields such as ecology (species count at a given site), transcriptomics (number of transcripts in a tissue) and quite broadly, all subfields of biology based on molecular markers and high-througput sequencing. They also arise in political sciences (voting outcomes), tourism management (number of visitors to sightseeing spots), etc. By analogy to the Gaussian graphical setting, many efforts have been devoted throughout the years to develop multivariate Poisson distribution in order to model dependencies between count variables (see Inouye et al., 2017, for a review). Unfortunately, there is no satisfying Poisson counterpart to the multivariate Gaussian. Besag (1974) proved that so-called Poisson Graphical Models (PGM) are limited to negative dependencies to ensure proper joint distribution. Yang et al. (2012) proposed several variants of PGM but all of them fail to have both marginal and conditional Poisson distributions. Similarly, Allen & Liu (2012)'s local PGM and Gallopin et al. (2013)'s log-normal models both satisfy the local Markov property but have no proper joint distribution. Both methods estimate the neighborhood of a node by performing a generalized linear regression *à la* Meinshausen & Bühlmann (2006). Last but not least, the observed count data often display a variance larger than expected under the Poisson assumption, so that a model that induces over-dispersion is highly desirable.

A different and more recent line of work – used for microbial ecology in SPIEC-EASI (Kurtz et al., 2015), gCoda (Fang et al., 2017) or BAnoCC (Schwager et al., 2017) – addresses

[1]MIA 518, AgroParitech/INRA, Université Paris-Saclay, Paris, France [2]MaIAGE, INRA, Université Paris-Saclay, Jouy-en-Josas, France. Correspondence to: Julien Chiquet <julien.chiquet@agroparistech.fr>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

the problem differently, by i) replacing counts with (regularized) frequencies, and ii) taking their log-ratios before iii) moving back to the GGM framework. A positive side effect of this transformation is to remedy the issue referred to as the *compositionality problem*: counts can only be compared to each other within a sample but not across samples as they depend on a sample-specific size-factor, which may induce spurious negative correlations. The transformation is simple but prevents integration of heterogeneous data sources and thus discovery of interactions between nodes from different sources (*e.g.* bacteria and fungi), although important ones have been experimentally documented (Lima-Mendez et al., 2015). Finally, although their statistical framework can accomodate it, previous methods do not offer a way to correct for confounding factors and may thus confuse systematic effects (*e.g.* reliance on common nutrients) for interactions (Vacher et al., 2016).

In this paper, we use the framework of hierachical Poisson log-normal (PLN) model with a latent Gaussian layer and an observed Poisson layer. We use the GGM formulation to model direct interactions in the Gaussian layer and the GLM formulation to control for confounding factors in the Poisson layer, in line with Chib & Greenberg (1995); Park & Lord (2007); Ma et al. (2008). By using source-specific offsets (Agresti, 1996), we both correct for compositionality and integrate data from different sources. The model is similar to Biswas et al. (2016) with a major difference: we do not neglect the uncertainty of the latent Gaussian vector during the inference step but use instead a variational procedure that consistently accounts for it.

We introduce the model in Section 2, the variational inference procedure in Section 3, simulation results in Section 4 and a reanalysis of two datasets in Section 5

2. A Graphical Model for Count Data

2.1. Multivariate Poisson Log-Normal (PLN) Model

In the multivariate PLN model (Aitchison & Ho, 1989) an *i.i.d.* sample is drawn as follows: for each observed p-dimensional count vector Y_i ($1 \leq i \leq n$), a Gaussian latent (*i.e.* hidden) p-dimensional vector Z_i is drawn and the coordinates of Y_i are sampled independently from a Poisson distribution, conditionally on Z_i:

$$Z_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Sigma}), \quad Y_{ij} \mid Z_{ij} \sim \mathcal{P}(\exp\{\mu_j + Z_{ij}\}). \quad (1)$$

The parameters involved are $\boldsymbol{\mu} = (\mu_j)_{1\leq j,\leq p}$ and $\boldsymbol{\Sigma} = (\sigma_{jk})_{1\leq j,k\leq p}$. In the following, all count vectors Y_i are gathered into the $n \times p$ matrix $\mathbf{Y} \triangleq (Y_{ij})_{1\leq i\leq n, 1\leq j\leq p}$. The $n \times p$ matrix $\mathbf{Z}$ is defined as $\mathbf{Z} \triangleq (Z_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ in the same way. The PLN distribution displays several interesting properties such as over-dispersion with respect to the Poisson distribution. Indeed: $\mathbb{E}(Y_{ij}) = e^{\mu_j + \sigma_{jj}/2}$ and $\mathbb{V}(Y_{ij}) = \mathbb{E}(Y_{ij}) + (e^{\sigma_{jj}} - 1)\mathbb{E}(Y_{ij})^2 \geq \mathbb{E}(Y_{ij})$. Furthermore, the covariance between coordinates can take an arbitrary sign as $\mathbb{C}\text{ov}(Y_{ij}, Y_{ik}) = (e^{\sigma_{jk}} - 1)\mathbb{E}(Y_{ij})\mathbb{E}(Y_{ik})$, which means that $\mathbb{C}\text{ov}(Y_{ij}, Y_{ik})$ has the same sign as $\mathbb{C}\text{ov}(Z_{ij}, Z_{ik}) = \sigma_{jk}$.

Introducing covariates. Interestingly, covariates can be easily introduced in the PLN model, by replacing the constant vector $\boldsymbol{\mu}$ with a regression term. Furthermore, in applications dealing with counts, an offset term is often required to account for some effects such as the sampling effort. Denote $x_i = (x_{i\ell})_{1\leq \ell\leq d}$ the vector of covariates for observation i and $\mathbf{B} = (\beta_{\ell j})_{1\leq \ell\leq d, 1\leq j\leq p}$ the corresponding matrix of regression coefficients. Also denote by o_{ij} the offset term for count Y_{ij}. Both can be accounted for by modifying (1) into

$$Y_{ij} \mid Z_{ij} \sim \mathcal{P}\left(\exp\{o_{ij} + x_i^\intercal \beta_j + Z_{ij}\}\right). \quad (2)$$

We define the offset matrix $\mathbf{O} = (o_{ij})_{1\leq i\leq n, 1\leq j\leq p}$ and the design matrix $\mathbf{X} = (x_{i\ell})_{1\leq i\leq n, 1\leq \ell\leq d}$ in the same way as $\mathbf{Y}$ and $\mathbf{Z}$.

2.2. The PLN-network Graphical Model

As mentioned in Section 1, no generic multivariate model is available for counts and existing models often impose undesired constraints on the dependency structure. To circumvent this issue, we use the PLN model to push the structure inference problem to the latent space and to infer the dependency structure relating the coordinates of the latent vector Z_i. We use the framework of graphical models (Lauritzen, 1996) to model this dependency structure. Intuitively, the graph encodes the conditional dependence structure between random variables. Formally, Z_i and Z_j are connected in the graph if and only Z_i and Z_j are independent conditionally on all other variables, that is: $Z_i \not\perp Z_j \mid Z_{\setminus\{i,j\}}$. Since the Z_i's are jointly Gaussian, so is $(Z_i, Z_j | Z_{-\{i,j\}})$. In particular, the partial correlation between Z_i and Z_j given the $(Z_k)_{k\neq i,j}$ is $\rho_{ij} = -\Omega_{ij}/\sqrt{\Omega_{ii}\Omega_{jj}}$ where $\boldsymbol{\Omega} \triangleq \boldsymbol{\Sigma}^{-1}$ is the precision matrix. Therefore Z_i and Z_j are conditionally independent if and only if $\Omega_{ij} = 0$ and the structure inference problem reduces to inferring the support of $\boldsymbol{\Omega}$, which is assumed to be sparse. In this perspective, it is critical to account for covariates that may have an effect on the observed counts to avoid spurious edges in the inferred graphical model (see e.g. Chandrasekaran et al., 2012, and discussions). As a consequence, we adopt the following parametrization for the distribution of Z_i given in (1):

$$Z_i \sim \mathcal{N}(\mathbf{0}_p, \boldsymbol{\Omega}^{-1}), \qquad \boldsymbol{\Omega} \text{ sparse} \quad (3)$$

which separates the structure parameter $\boldsymbol{\Omega}$ from the other effect parameters $\mathbf{O}$ and $\mathbf{B}$ appearing in the emission distribution (2).

3. Sparse Variational Inference

We now describe the adopted inference strategy, which aims to provide an estimate of the parameter $\boldsymbol{\theta} = (\mathbf{B}, \boldsymbol{\Omega})$.

3.1. Incomplete Data Model

In the incomplete data model (3), the evaluation of the log-likelihood of the observed data $\log p_{\boldsymbol{\theta}}(\mathbf{Y}) = \log \int p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) \mathrm{d}\mathbf{Z}$ is intractable, as well as its maximization with respect to $\boldsymbol{\theta}$. In this setting, the most popular strategy to perform maximum likelihood is to use the EM algorithm of (Dempster et al., 1977), which requires the evaluation of the conditional expectation of the complete log-likelihood $\mathbb{E}_{\boldsymbol{\theta}}\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z}) | \mathbf{Y}\right]$. Unfortunately, this amounts to computing (some moments of) the conditional distribution of each latent vector Z_i conditionally on the corresponding count vector Y_i, which has no close form in the PLN model. (Karlis, 2005) suggests to achieve this task via numerical or Monte-Carlo integration, but this approach is computationally too demanding when dealing with even a moderate number of variables.

Variational approximation. To circumvent this issue, we resort to a variational approximation (Wainwright & Jordan, 2008), which consists in finding a proxy for the conditional distribution $p_{\boldsymbol{\theta}}(Z_i|Y_i)$. This approach relies on a divergence measure between the true conditional distribution and the approximate distribution, chosen within a class $\mathcal{Q}$ of simple distributions, here the set of Gaussian distributions. Namely, each conditional distribution $p_{\boldsymbol{\theta}}(Z_i|Y_i)$ is approximated with a multivariate Gaussian distribution q_i with mean vector $\mathbf{m}_i$ and diagonal covariance matrix $\mathbf{S}_i = \mathrm{diag}(\mathbf{s}_i^2)$. The variational parameters are gathered into $\boldsymbol{\psi} = (\mathbf{M}, \mathbf{S})$, where $\mathbf{M} = [\mathbf{m}_1^\intercal \dots \mathbf{m}_n^\intercal]^\intercal$, $\mathbf{S} = [(\mathbf{s}_1^2)^\intercal \dots (\mathbf{s}_n^2)^\intercal]^\intercal$.

Choosing the Kullback-Leibler divergence to measure the quality of the approximation leads to the "variational" EM (VEM) algorithm, which aims to maximize the lower bound of the log-likelihood of the observed data, defined by

$$\begin{aligned} J(\mathbf{Y}; \boldsymbol{\psi}, \theta) &\triangleq \log p_{\boldsymbol{\theta}}(\mathbf{Y}) - KL\left[q_{\boldsymbol{\psi}}(\mathbf{Z}) || p_{\boldsymbol{\theta}}(\mathbf{Z}|\mathbf{Y})\right] \\ &= \mathbb{E}_q\left[\log p_{\boldsymbol{\theta}}(\mathbf{Y}, \mathbf{Z})\right] - \mathbb{E}_q\left[\log q_{\boldsymbol{\psi}}(\mathbf{Z})\right], \end{aligned} \tag{4}$$

where $\mathbb{E}_q$ is the expectation w.r.t. the distribution $q_{\boldsymbol{\psi}}$.

Sparse structure inference. To infer the structure – that is the underlying 'network' – we need to determine the support of $\boldsymbol{\Omega}$. To this end we add an ℓ_1 sparsity-inducing penalty to the lower bound of the likelihood, mimicking the Gaussian case like in the Graphical-Lasso. The corresponding objective function is thus

$$\begin{aligned} J_{\text{struct}}(\mathbf{Y}; \boldsymbol{\psi}, \boldsymbol{\theta}) &\triangleq J(\mathbf{Y}; \boldsymbol{\psi}, \boldsymbol{\theta}) - \lambda \left\|\boldsymbol{\Omega}\right\|_{\ell_1, \text{off}} \\ &\leq \log p_{\boldsymbol{\theta}}(\mathbf{Y}) - \lambda \left\|\boldsymbol{\Omega}\right\|_{\ell_1, \text{off}}, \end{aligned} \tag{5}$$

where $\|\boldsymbol{\Omega}\|_{\ell_1, \text{off}} = \sum_{j \neq k} |\Omega_{jk}|$ is the off-diagonal ℓ_1-norm of $\boldsymbol{\Omega}$ and $\lambda > 0$ is a tuning parameter controling the amount of sparsity. Note that, by construction, J_{struct} is a lower bound of the penalized log-likelihood.

3.2. Inference Algorithm

Objective function. The objective function J_{struct} inherits its properties from J since they only differ by the $\|\boldsymbol{\Omega}\|_{\ell_1, \text{off}}$ term. Thanks to Gaussian properties, J can be expressed in compact matrix notation. Let $\mathbf{S}_+ = \sum_{i=1}^n \mathbf{S}_i$ be the accumulated variance matrix; $\hat{\boldsymbol{\Sigma}} = n^{-1}\left(\mathbf{M}^\intercal \mathbf{M} + \mathbf{S}_+\right)$ be the estimated covariance matrix and $\mathbf{A} \triangleq (A_{ij})_{1 \leq i \leq n, 1 \leq j \leq p}$ the $n \times p$ matrix of expected counts, with entries:

$$A_{ij} \triangleq \mathbb{E}_q\left(Y_{ij}\right) = \exp(o_{ij} + x_i^\intercal \beta_j + m_{ij} + s_{ij}^2/2).$$

Then, the approximated log-likelihood writes:

$$\begin{aligned} J(\mathbf{Y}; \boldsymbol{\psi}, \boldsymbol{\theta}) = \mathbf{1}_n^\intercal \left(\mathbf{Y} \odot (\mathbf{O} + \mathbf{X}\mathbf{B} + \mathbf{M}) - \mathbf{A} + \frac{1}{2}\log \mathbf{S}\right) \mathbf{1}_p \\ + \frac{n}{2}\log\det \boldsymbol{\Omega} - \frac{n}{2}\,\mathrm{tr}\left(\hat{\boldsymbol{\Sigma}}\boldsymbol{\Omega}\right) + \frac{np}{2} - K(\mathbf{Y}), \end{aligned} \tag{6}$$

where $K(\mathbf{Y}) = \sum_{i,j} \log(Y_{ij}!)$ and $\odot$ is the Hadamard (term-to-term) product.

We now state the biconcavity of J and, consequently, of J_{struct}. This is the building block of the proposed alternating optimization algorithm. Proofs are given in Section S1.

Proposition 1 (Biconcavity of J)**.** *J is biconcave in $(\mathbf{B}, \mathbf{M}, \mathbf{S})$ and $\boldsymbol{\Omega}$. Furthermore, if $\mathbf{X}$ has full rank, J is strictly biconcave.*

Corollary 1 (Biconcavity of J_{struct})**.** *J_{struct} is biconcave in $(\mathbf{B}, \mathbf{M}, \mathbf{S})$ and $\boldsymbol{\Omega}$. Furthermore, if $\mathbf{X}$ has full rank, J_{struct} is strictly biconcave.*

The corollary follows from the concavity of $-\lambda\|\boldsymbol{\Omega}\|_{1,\text{off}}$ and from the fact that the sum of a strictly concave function with a concave function remains strictly concave.

Unfortunately, J (and consequently J_{struct}) is not jointly convex in $(\mathbf{B}, \mathbf{M}, \mathbf{S}, \boldsymbol{\Omega})$ in general and counter-examples can be found. In particular, this means that although gradient descent will converge to a stationary point of J (resp. J_{struct}), this stationary point is not guaranteed to be the global optimum of J (resp. J_{struct}) and may depend on the starting point of the iterative algorithm. Note that the same caveat applies to alternating optimization schemes such as the (V)EM algorithm.

Alternate optimization. To estimate both the variational and the model parameters we need to maximize J_{struct} with the additional box constraint that $\mathbf{S} \succ 0$, *i.e.*, that variance parameters in the variational distribution are positive. Thanks to the biconcavity of J_{struct}, we iterate two

updates until convergence. Iteration h consists of update (h_1) where $(\mathbf{B}^{(h)}, \mathbf{M}^{(h)}, \mathbf{S}^{(h)})$ are obtained by maximizing $J(\mathbf{Y}; (\mathbf{M}, \mathbf{S}), (\mathbf{B}, \mathbf{\Omega}^{(h-1)}))$, such that $\mathbf{S}$ is positive-definite, and update (h_2) where $\mathbf{\Omega}^{(h)}$ is obtained by maximizing $J_{\text{struct}}(\mathbf{Y}; (\mathbf{M}^{(h)}, \mathbf{S}^{(h)}), (\mathbf{B}^{(h)}, \mathbf{\Omega}))$.

Problem (h_1) can be solved by a gradient ascent with box-constraints for the variational variances $\mathbf{S}$ that must remain positive. We use the gradients which are given by $\nabla_{\mathbf{B}} J = \mathbf{X}^\intercal(\mathbf{Y} - \mathbf{A})$, $\nabla_{\mathbf{M}} J = \mathbf{Y} - \mathbf{A} - \mathbf{M}\mathbf{\Omega}$, $\nabla_{\mathbf{S}} J = \frac{1}{2}\left(\mathbf{S}^{\oslash} - \mathbf{A} - \mathbf{1}_n \operatorname{diag}(\mathbf{\Omega})^\intercal\right)$.

When $\lambda > 0$, Problem (h_2) is equivalent to minimizing $-n \log\det \mathbf{\Omega} + n \operatorname{tr}\left(\hat{\mathbf{\Sigma}}\mathbf{\Omega}\right) + 2\lambda \left\|\mathbf{\Omega}\right\|_{\ell_1,\text{off}}$ over the set of positive definite matrices. We recognize a sparse multivariate Gaussian maximum likelihood problem (Yuan & Lin, 2007; Banerjee et al., 2008), efficiently solved by the graphical-Lasso algorithm (Friedman et al., 2008).

The algorithm is initialized using the estimator of the graphical-Lasso obtained by shrinking the covariance matrix computed on the Pearson residuals of a linear model predicting $\log(1 + \mathbf{Y})$ from $\mathbf{X}$ and $\mathbf{O}$.

Model Selection. Model selection is a notoriously hard in network inference. Several procedures have been proposed to select an optimal value of λ in GGM and we rely on both (i) the Stability Approach to Regularization Selection (StARS) introduced in Liu et al. (2010) and (ii) variants of BIC taylored for the high-dimensional setting, such as EBIC (Chen & Chen, 2008).

Briefly, StARS relies on resampling a large number B of subsamples of size m and inferring a network $\mathbf{\Omega}^{(b,\lambda)}$ on each subsample b for each value of λ in a grid Λ. The frequency of inclusion of edge $e = i \sim j$ is computed as $p_e^\lambda = \#\{b : \Omega_{ij}^{(b,\lambda)} \neq 0\}/B$ and its variance as $v_e^\lambda = p_e^\lambda(1 - p_e^\lambda)$. The stability $\text{stab}(\lambda)$ of the network is then simply $\text{stab}(\lambda) = 1 - 2\bar{v}^\lambda$ where $\bar{v}^\lambda$ is the average of the v_e^λ. Note that $\text{stab}(\lambda)$ decreases from 1 for $\lambda = \infty$ (empty network) to a nonnegative value for small λ. StARS selects the smallest λ (densest network) for which $\text{stab}(\lambda) \geq 1 - 2\beta$. We use $B = 50$ subsamples of size $m = \lfloor 10\sqrt{n} \rfloor$ and $2\beta = 0.05$ as default, as suggested in Liu et al. based on theoretical results.

By contrast, EBIC is a non-resampling based alternative with no computational overhead, that penalizes both the number of unknown parameters and the complexity of the model space (Chen & Chen, 2008). In our framework, $\text{EBIC}_\gamma(\hat{\mathbf{B}}, \hat{\mathbf{\Omega}}_\lambda) = -2\,\text{loglik}(\mathbf{Y}; \hat{\mathbf{B}}, \hat{\mathbf{\Omega}}_\lambda) + \log(n)(|\mathcal{E}_\lambda| + pd) + \gamma \log\binom{p(p+1)/2}{|\mathcal{E}_\lambda|}$, where $\mathcal{E}_\lambda$ is the edge set of a candidate graph and $\binom{m}{n}$ is the binomial coefficient. The first penalty term is the usual BIC penalization: our model has pd unknown regression parameters in $\mathbf{B}$ plus $|\mathcal{E}_\lambda|$ inferred terms in $\hat{\mathbf{\Omega}}_\lambda$. The second penalty term, tuned by $\gamma \in [0, 1]$, is used to adjust the tendency of the usual BIC – recovered for $\gamma = 0$ – to choose overly dense graphs in the high-dimensional setting. Here, we replace loglik in EBIC by its variational surrogate $J(\mathbf{Y}; \hat{\mathbf{\Omega}})$ and use $\gamma = 0.5$ as recommended by Foygel & Drton for GGM.

Implementation. We implemented our algorithm in a `R/C++` package **PLNmodels**, available on github `https://github.com/jchiquet/PLNmodels`. using Gradient ascent with box constraints, as implemented in the **nlopt** library (Johnson, 2011), for the first step and the implementation of the graphical-Lasso found in the **glassofast** `R` package (Sustik & Calderhead, 2012) for the second step.

4. Simulation Study

4.1. Simulation Protocol

Network generation. We generate ground truth graphs according to an Erdös-Rényi model (no particular structure), a preferential attachment model (scale-free property) or an affiliation model (community structure). These models are used to generate a binary adjacency matrix $\mathbf{G}$ and in turn a positive-definite precision matrix $\mathbf{\Omega}$ with the sparsity pattern of $\mathbf{G}$. $\mathbf{\Omega}$ is defined in two steps as follows: $\tilde{\mathbf{\Omega}} = \mathbf{G} \times v$, $\mathbf{\Omega} = \tilde{\mathbf{\Omega}} + \operatorname{diag}(|\min(\mathbf{eig}(\tilde{\mathbf{\Omega}}))| + u)$, with $u, v > 0$. The two scalars u, v control both the difficulty of the network inference problem – they are related to the strength of the partial correlations, and in turn of the interactions in the network – and the conditioning of $\mathbf{\Omega}$. Higher v leads to stronger correlations and higher u to better conditioning. We always set $v = 0.3, u = 0.1$ in our simulations. This is similar to the protocol used in the `R` package **huge**.

Compositional data generation. To ensure fair comparisons, we did not simulate count data from a PLN distribution. Instead, we introduce a compositional model frequently used in community ecology and genomics. The simulation process is sketched in Section S2 and consists in 3 steps:

i) Draw the 'real' (unreachable) *abundances* $\mathbf{a}_i$ of the p species in sample i from a lognormal: $\mathbf{a}_i \sim \mathcal{LN}(\mathbf{XB}, \mathbf{\Omega}^{-1})$ where $\mathbf{X}$ accounts for some covariates and $\mathbf{\Omega}$ is the latent network between species generated as explained above.

ii) Transform abundances $\mathbf{a}_i$ to *proportions* $\boldsymbol{\pi}_i$ with a logistic-transform, i.e. $\pi_{ij} = e^{a_{ij}} / \sum_j e^{a_{ij}}$.

iii) For random value of N_i – the sampling effort in sample i, typically the sequencing depth – draw observed *counts* Y_i via a multinomial distribution $\mathcal{M}(N_i, \boldsymbol{\pi}_i)$.

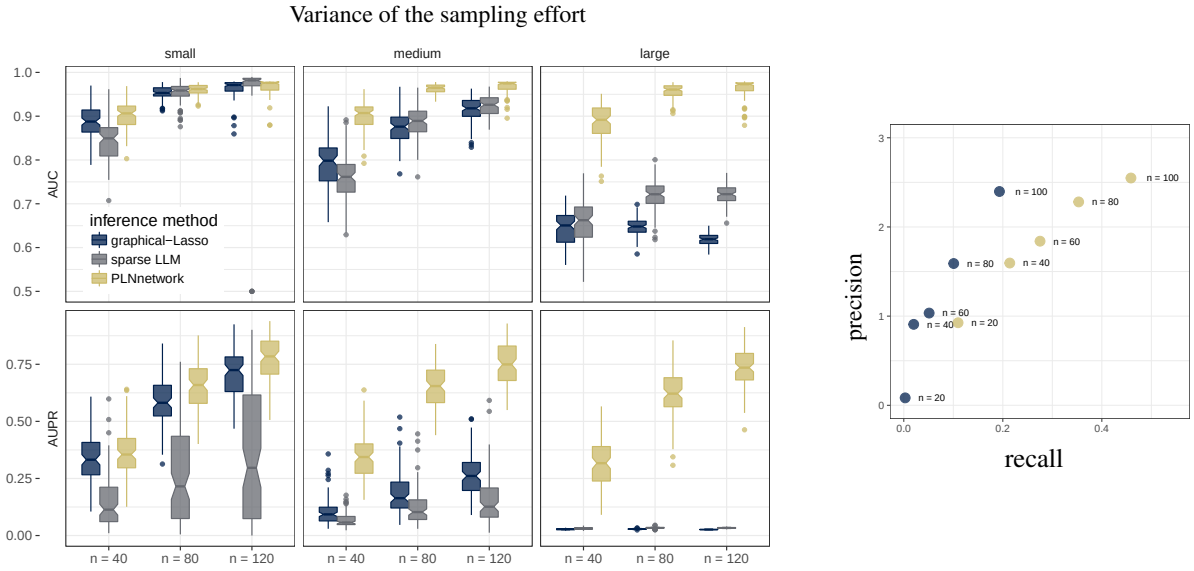

Figure 1. Left: Effect of the variability of the sampling effort, a.k.a the compositional problem, on the quality of the reconstruction of 50-node random networks measured with AUC (top) or AUPR (bottom) on 100 simulations. Right: Performance of the model selection procedures (BIC or StARS) in `PLNnetwork` for reconstructing 50-node randoms networks, averaged over 100 simulations.

Experimental setup. We fix the number of variables to $p = 50$ in all our experiments. This represents both the largest network size that can be reasonably recovered considering typical sample sizes ($n \sim 50 - 100$) and the (approximate) number of nodes in the datasets of Section 5. The sampling efforts N_i are drawn from a negative binomial distribution with mean 1000 and variance $1000 + 1000^2/\nu$: $N_i \sim^{\text{i.i.d}} \mathcal{NB}(\mu = 1000, \nu)$. The covariates are chosen so that $\mathbf{X}$ is the design matrix of a one-way ANOVA with 3 balanced groups and the regression coefficients are sampled in a uniform distribution: $B_{jk} \sim^{\text{i.i.d.}} \mathcal{U}(-b, b)$. These parameters were chosen to mimic (marginal) count distributions – in terms of location and dispersion – commonly observed in microbial ecology applications. To control the difficulty of the problem, we vary (i) the sample size n, (ii) the overdispersion parameter ν, larger ν corresponding to more similar sampling efforts, (iii) the effect of covariates b, larger b corresponding to larger Signal to Noise Ratio (SNR) in the underlying linear model and therefore smaller fraction of variance explained by $\mathbf{\Omega}$.

Competitors. All methods are referred to using `teletype` family font. For instance, we refer to our method as `PLNnetwork`. Among the many competitors, we pick one representative from each family:

i) Sparse GGM methods (Friedman et al., 2008; Meinshausen & Bühlmann, 2006) applied to log-transformed counts, or `graphical-Lasso`, as implemented in the `R`-package **glasso** (Friedman et al., 2008).

ii) Sparse log-linear graphical models (Yang et al., 2012; Allen & Liu, 2012), or `sparse LLM`, as implemented in the `R`-package **RNAseqNet** (Imbert et al., 2017).

iii) Methods dedicated to compositional count data, such as `SPiEC-Easi` (Kurtz et al., 2015) for the precision matrix and `sparCC` (Friedman & Alm, 2012) for the covariance one. Both methods correct for compositionality by using pseudo-counts plus log-transformation. The former then applies graphical-Lasso and nonparanormal transformation (Liu et al., 2009) while the latter uses resampling and thresholded correlations. Both are implemented in the `R`-package **spieceasi**.

Performance assessment. Each competitor produces a sequence of networks, varying from an empty to a full graph and ordered by a tuning parameter λ that controls the number of edges in the graph. Since the problem of choosing λ is particularly troublesome in network inference, it was left aside by comparing the methods in terms of the precision-recall (PR) or Receiver operating characteristic (ROC) curves. We recall that ROC curves are obtained by plotting the true positive rate (or recall) as a function of the false positive rate (or fall-out), while PR curve represents the positive predictive value (or precision) as a function of the recall. Although ROC curves are more present in the literature, PR curves are more informative in unbalanced cases with a small proportion of positives as they give less weight to regions with large false positive rates, which are of limited interest in practice (Davis & Goadrich, 2006). The curves for one simulation were summarized using the

area under the ROC curve (AUC) and area under the PR curve (AUPR): the larger the AUC/AUPR, the better the reconstruction.

4.2. Results

Our first batch of experiments illustrate the effect of sampling effort and external covariates on network quality.

Non-compositional methods fail. We study the heterogeneity of sampling efforts by varying ν in $\{100, 10, 2\}$ (resp. small, medium and large heterogeneity) and comparing `graphical-Lasso`, `sparse LLM` and `PLNnetwork`. The latter is the only one that accounts for compositionality by introducing a sample-specific offset, computed as the sum of counts in each sample. Results averaged over 100 replicates are displayed in Figure **??**. They show that, as expected, `PLNnetwork` is not sensitive to differences in sampling effort, contrary to `graphical-Lasso` and `sparse LLM`. Both methods completely fail in terms of AUC, and even more so in terms of AUPR, when sampling efforts vary wildly. For large sample sizes ($n = 120$) and medium variability, although the AUC is close to 1, the low AUPR proves that the first, and supposedly most reliable, edges inferred by `graphical-Lasso` and `sparse LLM` are in fact mostly false positives.

Accounting for covariates effect does matter. We now focus on the effect of an external covariate in the data, and how it affects the performance of the methods. Regarding the sampling effort, we fix $\nu = 2$ in this experiment, and we only compare the compositional methods together since the other approaches fail in this setting. The strength of the covariate effect is controlled by the parameter b in our compositional model. The larger b, the larger the effect of the covariate and the harder the problem of network reconstruction when not accounting for the covariate. We vary b in $\{1, 2, 3\}$, (resp. small, medium and large effects). On top of that, we vary the sample size and consider the three network topologies (scale-free, random and community networks), always with $p = 50$ nodes. We evaluate the performance of `SPiEC-Easi`, `sparCC` and `PLNnetwork` in terms of AUC and AUPR on 100 simulation of each kind and report the average values in Section S3.

Model Selection issue. We now focus on `PLNnetwork` to address the question of choosing λ and compare StARS and EBIC, as presented in Section 3. Figure **??** reports the results of our experiments in terms of precision/recall averaged over 100 simulation. It shows that StARS systematically outperforms BIC in terms of recall and precision. However, this increase in performance comes at the cost of a huge computational burden.

5. Illustrations

We illustrate our methodology with two examples from different fields. The French election example shows that our method can handle large datasets. We illustrate how to interpret the results and propose some validation checks. We then consider a metagenomic example (Oak mildew), on which we show how to decompose the effects of the different covariates on the inferred interactions and propose some biological interpretations.

5.1. French Presidential Elections, 2017

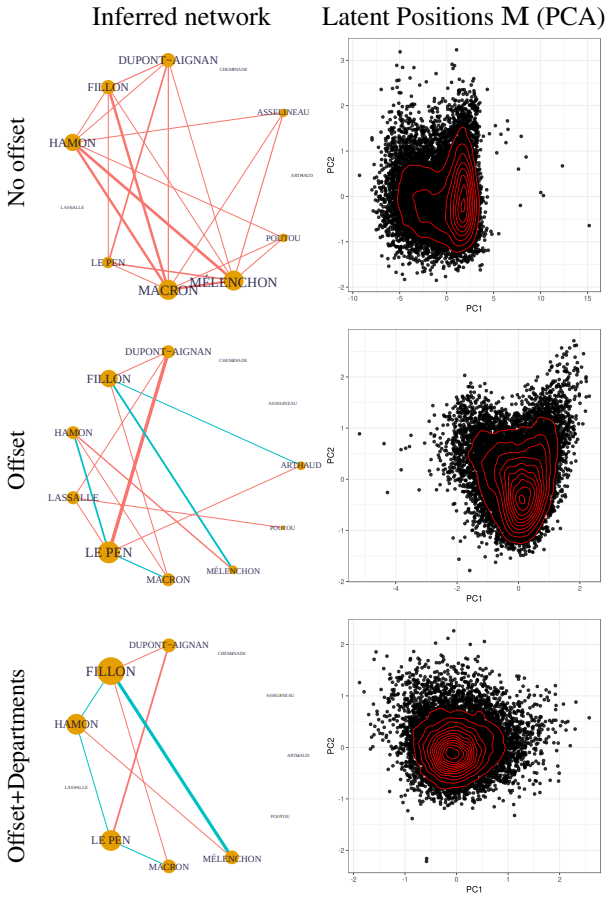

Figure 2. Network between candidates of last French presidential elections. Left column: Networks inferred under different models. Edges represent partial correlations ρ_{ij}: their thickness is proportional to $|\rho_{ij}|$ and they are colored red if $\rho_{ij} > 0$ and blue if $\rho_{ij} < 0$. A node's size and label size are proportional to its degree. Right column: Positions of the polling stations $\mathbf{M}$ in the Gaussian latent space. Since the latent space has dimension 11, we performed a PCA of $\mathbf{M}$ and show only the principal plane. Red lines represent contour lines of the density estimated with a 2D Gaussian kernel. Roughly elliptic curves reveal that there is no obvious remaining structure in the latent space.

We first consider the first round of the French presidential

election of 2017. The dataset consists in the votes cast for each of the 11 candidates in the more than 63 000 polling stations. Our goal here is to find *competing* candidates, who appeal to different voters, and *compatible* candidates, who appeal to the same voters, after accounting for the fact that elections are a zero-sum game.

Data were downloaded from the French open data platform `data.gouv.fr` and filtered to remove stations with no votes. To reduce inference times, we consider a random subset of 13,704 stations that accounted for 20% of the registered population. The voting population in those booths varied wildly, ranging from 10 to 105,891 (6th district of French citizens living abroad) registered voters, with a median at 736 and 99.5% of the stations with less than 1,700 voters. We consider the log-registered population of voters as an offset to account for different station sizes. This means in particular that votes are not affected by the *compositionality effect* as much as in other settings, as they do not sum up to the offset. Voting patterns are well-known to depend on geography and we therefore consider department (a French administrative division) as a proxy for geography.

We consider three models in total: without offset, with offset but no covariate, with offset and covariate and use the same grid of λ – decreasing geometrically from 1 to $1e^{-3}$ in 31 steps – for all. The optimal value $\lambda^\star$ was selected using StARS with 100 subsamples of size 1170 ($\simeq 10\sqrt{n}$).

The offset matters. Figure 2 shows that the inclusion of an offset drastically reduces the density of the reconstructed network and alters the sign and strength of partial correlations. Failing to accoung for varying station sizes leads to a spurious positive partial correlations between most candidates: the shift of all stations towards the positive orthant in the latent space are mistaken for positive correlations between all coordinates. The offset counteracts this by translating back all stations towards the origin along the direction $\mathbb{R}\mathbf{1}$.

Correcting for geography is important. Figure 2 shows that correcting for geography also changes the graph but to a lesser extent. When including only the offset, the inferred latent positions ($\mathbf{M}$) do not display the expected elliptic distribution of a multivariate Gaussian (Fig. 2, right panel). Taking the department of origin into account helps recover ellipticity and confirms that geography is indeed a strong structuring factor in the latent space.

Political interactions. If we consider the network reconstructed with the offset and geographic covariate as the most reliable, results show that candidate with similar political leaning appeal to the same voters (M. Le Pen and N. Dupont-Aignan (both far right), B. Hamon (left) and J.-L. Mélenchon (far left), E. Macron (center) and J.-F. Fillon (right)) whereas candidate with different leanings appeal to different voters (M. Le Pen versus B. Hamon and E. Macron, J.-F. Fillon versus J.-L. Mélenchon). More precisely, a negative partial correlation between candidates A and B means that, all other things being equal, a high vote for one in a station is correlated to a low vote for the other.

This may explain the absence of negative correlation between far left and far right: although their electorates are different, they vote in the same stations. Similarly, the fact that the positive partial correlation between E. Macron and B. Hamon disappears when controlling for geography means that they have high voter shares in the same departments but not necessarily in the same polling stations. This is confirmed by the high correlation (0.76) of their respective regression coefficients across departments.

5.2. Oak Mildew

The metagenomic dataset introduced in (Jakuschkin et al., 2016) consists of microbial communities sampled on the surface of oak leaves (the samples). The leaves were collected on trees with different resistance levels to the fungal pathogenic species *E. alphitoides*, responsible for the oak powdery mildew. The available information about the operating taxononomic units (OTU – a proxy for species) are given in Section S4. Unfortunately, not all OTU can been identified at the species level and some OTU are not related to any known species. In the following, we consider two groups of samples labeled by Jakuschkin et al.: $n_r = 39$ resistant samples (where *E. alphitoides* was essentially absent) and $n_s = 39$ susceptible samples (where a significant activity of *E. alphitoides* was detected). In addition to the sampling tree, several covariates were measured for each leaf: orientation, distance to trunk, distance to ground, distance to base. The total number of OTU considered is $p = 114$ in this data set (66 bacterial ones and 48 fungal ones, including *E. alphitoides*).

Our aim here is to unravel the association between the different microbial and fungal species by reconstructing the ecological network. Obviously, we are especially interested in the interactions between *E. alphitoides* and the other species. We emphasize that unlike `SPiEC-Easi` or `sparCC`, that are limited to interactions between bacteria or between fungi due their normalisation step, we can actually investigate interactions between bacterial and fungi *E. alphitoides* although each type has its own sequencing depth. A similar target was already at the core of Jakuschkin et al.'s work. However, our approach differs from theirs from a methodological view-point as we jointly estimate the effect of the covariates $\mathbf{B}$ and the dependency structure $\mathbf{\Omega}$ while they only corrected the observed counts for the effect of the covariates using a regression model before feeding the residuals from that regression to a network inference

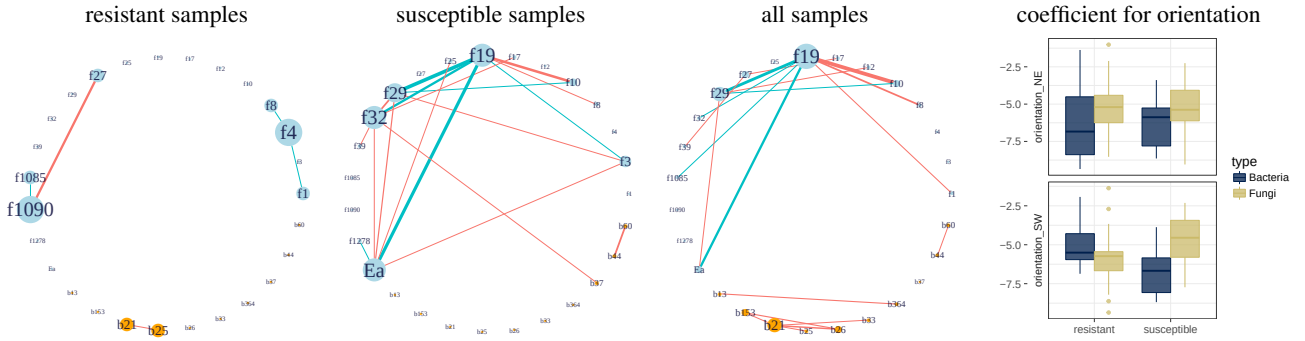

Figure 3. Oak mildew network analysis: networks inferred by `PLNnetwork` and selected by StARS for a stability of 0.995. Each network correspond to networks inferred using samples respectively from the resistant tree, the susceptible tree and both trees. Blue vertices represent Fungi; orange vertices represent bacteria. Edges represent partial correlations ρ_{ij}: edge thickness is proportional to $|\rho_{ij}|$ and are colored red if $\rho_{ij} > 0$ and blue if $\rho_{ij} < 0$. A node's size and label size are proportional to its degree. Only nodes having at least one edge among the three networks are included in the plots. Inset: Boxplot of regression coefficient of abundances against orientation.

method. This two-steps procedure fails to account for the fact that $\mathbf{B}$ is estimated and to propagate uncertainty from the first step to the second one. Moreover, Jakuschkin et al. focused their study on the set of susceptible samples, while we propose here to infer three networks: one for susceptible samples, one for resistant samples and one for all samples. By these means, we hope to obtain a more thorough map of interactions between the pathogen and its ecosystem.

The three PLN models including respectively the susceptible, the resistant and all samples were defined as follows: for the susceptible and the resistant models, we applied `PLNnetwork` by including simple effects of the orientation and of the distance to the trunk (the other distances were highly correlated with the former). For the model with all the samples, we added the tree status (resistant or susceptible) as a covariate, in addition to its interactions with the two other covariates (orientation and distance to trunk). These two approaches – separating or merging the samples – address different yet complementary goals: by separating the samples, we assume that the two underlying networks (and thus covariances) are different and need a specific analysis; the counterpart that merges all samples aims to render a synergistic network that encompasses important interactions from both situations after correction of the mean effects due to the tree status (resistant or susceptible).

Before getting into the interpretation of the results in terms of species interactions, we remind that the PLN models also enables to measure the effect of the covariates on each species. The right panel of Figure 3 displays the distribution of the regression parameters of the two orientation indicators (NE = north-east and SW = south-west), in each tree, across each species type. We do not discuss extensively these results but one may observe a strong interaction between SW orientation and tree type on both fungi and bacteria: bacteria are notably depleted in leaves facing SW in susceptible trees.

We now focus on the results of our analysis in terms of networks in Figure 3. All networks inferred with `PLNnetwork` where selected with StARS on a 50-size grid of penalties, using a high stability level of $1 - 2\beta = 0.995$ to drastically limit the number of false positive edges. The resistant and susceptible networks show very different patterns, while the consensus network seems to catch features from both of them. In the susceptible network, *E. alphitoides* is identified as (i) antagonist to fungi f1278, from the *Mycosphaerella punctiformis* species, which colonizes living oak leaves asymptomatically and may prevent infection by *E. alphitoides* and (ii) mutualist to fungi f29, from the *Xenosonderhenia syzygii* species, usually found in leaf spots, common on weakened and senescent leaves. The other mutualists of *E. alphitoides* unfortunately belong to unknown species and no similar observations can be made. Interestingly, in the susceptible network, the pathogen has less interactions than fungi f19, but is connected to it, whereas both have few connections in the resistant network. As *E. alphitoides* is known to be responsible for the mildew disease, the comparison of these networks suggest that its pathogenic effect is partially mediated by f19. In addition to the direct effect of the pathogen on a small set of species, its (negative) effect on fungi f19, which seems to play a central role in the phyllosphere, leverages its impact on the whole system. Finally, the consensus network encompassing both sources of samples resembles the susceptible network, with some notable discrepancies: a cluster composed by bacterial species b21, b25, b26, b153 and to a lesser extent b33 is found in the consensus network, which was only incipient in the resistant network. This is probably due to the gain in statistical power induced by a larger sample-size.

References

Agresti, A. *An introduction to categorical data analysis*, volume 135. Wiley New York, 1996.

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

Allen, G. I. and Liu, Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In *Bioinformatics and Biomedicine (BIBM), 2012 IEEE International Conference on*, pp. 1–6. IEEE, 2012.

Banerjee, O., Ghaoui, L. E., and d'Aspremont, A. Model selection through sparse maximum likelihood estimation for multivariate gaussian or binary data. *Journal of Machine learning research*, 9(Mar):485–516, 2008.

Besag, J. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 192–236, 1974.

Biswas, S., Mcdonald, M., Lundberg, D. S., Dangl, J. L., and Jojic, V. Learning microbial interaction networks from metagenomic count data. *Journal of Computational Biology*, 23(6):526–535, 2016.

Chandrasekaran, V., Parrilo, P. A., and Willsky, A. S. Latent variable graphical model selection via convex optimization. *The Annals of Statistics*, 40(4):1935–1967, 2012.

Chen, J. and Chen, Z. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

Chib, S. and Greenberg, E. Understanding the metropolis-hastings algorithm. *The American Statistician*, 49(4): 327–335, 1995.

Davis, J. and Goadrich, M. The relationship between precision-recall and roc curves. In *Proceedings of the 23rd international conference on Machine learning*, pp. 233–240. ACM, 2006.

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

Drew, K., Mller, C. L., Bonneau, R., and Marcotte, E. M. Identifying direct contacts between protein complex subunits from their conditional dependence in proteomics datasets. *PLOS Computational Biology*, 13(10):1–23, 10 2017.

Fang, H., Huang, C., Zhao, H., and Deng, M. gCoda: conditional dependence network inference for compositional data. *Journal of Computational Biology*, 24(7):699–708, 2017.

Fiers, M. W. E. J., Minnoye, L., Aibar, S., Bravo Gonzlez-Blas, C., Kalender Atak, Z., and Aerts, S. Mapping gene regulatory networks from single-cell omics data. *Briefings in Functional Genomics*, pp. elx046, 2018.

Foygel, R. and Drton, M. Extended Bayesian information criteria for gaussian graphical models. In *Advances in neural information processing systems*, pp. 604–612, 2010.

Friedman, J. and Alm, E. J. Inferring correlation networks from genomic survey data. *PLOS Computational Biology*, 8(9):1–11, 09 2012.

Friedman, J., Hastie, T., and Tibshirani, R. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.

Gallopin, M., Rau, A., and Jaffrzic, F. A hierarchical Poisson Log-Normal model for network inference from RNA sequencing data. *PLOS ONE*, 8(10):1–9, 10 2013.

Harris, D. J. Inferring species interactions from co-occurrence data with Markov networks. *Ecology*, 97 (12):3308–3314, 2016. ISSN 1939-9170.

Imbert, A., Valsesia, A., Le Gall, C., Armenise, C., Lefebvre, G., Gourraud, P.-A., Viguerie, N., and Villa-Vialaneix, N. Multiple hot-deck imputation for network inference from rna sequencing data. *Bioinformatics*, pp. btx819, 2017.

Inouye, D. I., Yang, E., Allen, G. I., and Ravikumar, P. A review of multivariate distributions for count data derived from the Poisson distribution. *Wiley Interdisciplinary Reviews: Computational Statistics*, 9(3), 2017.

Jakuschkin, B., Fievet, V., Schwaller, L., Fort, T., Robin, C., and Vacher, C. Deciphering the pathobiome: Intra- and interkingdom interactions involving the pathogen Erysiphe alphitoides. *Microbial ecology*, pp. 1–11, 2016.

Johnson, S. G. *The NLopt nonlinear-optimization package*, 2011.

Karlis, D. EM algorithm for mixed Poisson and other discrete distributions. *Astin bulletin*, 35(01):3–24, 2005.

Kurtz, Z. D., Müller, C. L., Miraldi, E. R., Littman, D. R., Blaser, M. J., and Bonneau, R. A. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comput Biol*, 11(5):e1004226, May 2015.

Lauritzen, S. L. *Graphical Models*. Oxford Statistical Science Series. Clarendon Press, 1996.

Lima-Mendez, G., Faust, K., Henry, N., Decelle, J., Colin, S., Carcillo, F., Chaffron, S., Ignacio-Espinosa, J. C., Roux, S., Vincent, F., Bittner, L., Darzi, Y., Wang, J., Audic, S., Berline, L., Bontempi, G., Cabello, A. M.,

Coppola, L., Cornejo-Castillo, F. M., d'Ovidio, F., De Meester, L., Ferrera, I., Garet-Delmas, M.-J., Guidi, L., Lara, E., Pesant, S., Royo-Llonch, M., Salazar, G., Sánchez, P., Sebastian, M., Souffreau, C., Dimier, C., Picheral, M., Searson, S., Kandels-Lewis, S., Gorsky, G., Not, F., Ogata, H., Speich, S., Stemmann, L., Weissenbach, J., Wincker, P., Acinas, S. G., Sunagawa, S., Bork, P., Sullivan, M. B., Karsenti, E., Bowler, C., de Vargas, C., and Raes, J. Determinants of community structure in the global plankton interactome. *Science*, 348(6237), 2015. ISSN 0036-8075.

Liu, H., Lafferty, J., and Wasserman, L. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(Oct):2295–2328, 2009.

Liu, H., Roeder, K., and Wasserman, L. Stability approach to regularization selection (stars) for high dimensional graphical models. In *Proceedings of the 23rd International Conference on Neural Information Processing Systems - Volume 2*, NIPS'10, pp. 1432–1440, USA, 2010. Curran Associates Inc.

Ma, J., Kockelman, K. M., and Damien, P. A multivariate Poisson-lognormal regression model for prediction of crash counts by severity, using Bayesian methods. *Accident Analysis & Prevention*, 40(3):964–975, 2008.

Meinshausen, N. and Bühlmann, P. High-dimensional graphs and variable selection with the lasso. *Ann. Statist.*, 34(3):1436–1462, 06 2006.

Moignard, V., Woodhouse, S., Haghverdi, L., Lilly, A. J., Tanaka, Y., Wilkinson, A. C., Buettner, F., Macaulay, I. C., Jawaid, W., Diamanti, E., et al. Decoding the regulatory network of early blood development from single-cell gene expression measurements. *Nature biotechnology*, 33(3): 269, 2015.

Park, E. and Lord, D. Multivariate poisson-lognormal models for jointly modeling crash frequency by severity. *Transportation Research Record: Journal of the Transportation Research Board*, (2019):1–6, 2007.

Ravikumar, P., Wainwright, M. J., Lafferty, J. D., et al. High-dimensional ising model selection using l1-regularized logistic regression. *The Annals of Statistics*, 38(3):1287–1319, 2010.

Schwager, E., Mallick, H., Ventz, S., and Huttenhower, C. A Bayesian method for detecting pairwise associations in compositional data. *PLOS Computational Biology*, 13 (11):1–21, 11 2017.

Sustik, M. A. and Calderhead, B. Glassofast: An efficient glasso implementation. *UTCS Technical Report TR-12-29 2012*, 2012.

Vacher, C., Tamaddoni-Nezhad, A., Kamenova, S., Peyrard, N., Moalic, Y., Sabbadin, R., Schwaller, L., Chiquet, J., Smith, M. A., Vallance, J., Fievet, V., Jakuschkin, B., and Bohan, D. A. Learning ecological networks from next-generation sequencing data. In Woodward, G. and Bohan, D. A. (eds.), *Ecosystem Services: From Biodiversity to Society, Part 2*, volume 54 of *Advances in Ecological Research*, pp. 1 – 39. Academic Press, 2016.

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

Yang, E., Allen, G., Liu, Z., and Ravikumar, P. K. Graphical models via generalized linear models. In *Advances in Neural Information Processing Systems*, pp. 1358–1366, 2012.

Yu, X., Zeng, T., Wang, X., Li, G., and Chen, L. Unravelling personalized dysfunctional gene network of complex diseases based on differential network model. *Journal of translational medicine*, 13(1):189, 2015.

Yuan, M. and Lin, Y. Model selection and estimation in the Gaussian graphical model. *Biometrika*, 94(1):19–35, 2007.