A zero inflated log-normal model for inference of sparse microbial association networks

Vincent Prost [1,2,3], Stéphane Gazut [1] and Thomas Brüls [2,3*]

1 CEA, LIST, Laboratoire Sciences des Données et de la Décision, 91191 Gif-sur-Yvette, France
2 CEA, DRF, Institut François Jacob, Genoscope, 91057 Evry, France
3 CNRS-UMR8030, Université Paris-Saclay, UEVE, 91057 Evry, France.

*Corresponding author
E-mail: lastname@genoscope.cns.fr

Abstract

The advent of high-throughput metagenomic sequencing has prompted the development of efficient taxonomic profiling methods allowing to measure the presence, abundance and phylogeny of organisms in a wide range of environmental samples. Multivariate sequence-derived abundance data further has the potential to enable inference of ecological associations between microbial populations, but several technical issues need to be accounted for, like the compositional nature of the data, its extreme sparsity and overdispersion, as well as the frequent need to operate in under-determined regimes.

The ecological network reconstruction problem is frequently cast into the paradigm of Gaussian graphical models (GGMs) for which efficient structure inference algorithms are available, like the graphical lasso and neighborhood selection. Unfortunately, GGMs or variants thereof can not properly account for the extremely sparse patterns occurring in real-world metagenomic taxonomic profiles. In particular, structural zeros (as opposed to sampling zeros) corresponding to true absences of biological signals fail to be properly handled by most statistical methods.

We present here a zero-inflated log-normal graphical model specifically aimed at handling such "biological" zeros, and demonstrate significant performance gains over state-of-the-art statistical methods for the inference of microbial association networks, with most notable gains obtained when analyzing taxonomic profiles displaying sparsity levels on par with real-world metagenomic datasets.

1 Introduction

Metagenomics has increased our awareness that microbes most often live in communities structured by both environmental factors and ecological interactions between community members. Understanding the structure and dynamics of microbial communities thus requires to be able to detect such associations, and is of both fundamental [1] and practical importance as it may ultimately enable the design and engineering of consortia of interacting organisms in order to fulfill various needs (e.g. improving the efficiency of wastewater treatment plants [2, 3] and designing more robust fecal transplants [4]).

Competition, antagonism, symbiosis and commensalism are the most important types of ecological interactions, and because of the current inability to co-culture most of the microbes sampled from the environment [5, 6], computational methods play a key

role in predicting microbial associations. The advent of high-throughput sequencing and metagenomics resulted in the production of large amounts of multivariate microbial abundance data that could in theory be leveraged to reconstruct microbial association networks [7]. In practice however, only limited success has been met in terms of robust structure inference of real-world microbial interaction networks, and different methods frequently yield quite different results [8].

In principle, methods that consider the conditional dependency structure of microbial networks (like probabilistic graphical models) should perform better than methods using only univariate associations to predict pairwise interactions, because the former has the power to resolve direct from indirect associations (e.g. associations between two species mediated through a third species). For example, ref [9] showed that "univariate networks" (i.e. networks based on pairwise interactions predicted from univariate statistical associations) include high proportions of false-positive predictions when there is a substantial level of dependence between the samples, while the number of false positives was highly reduced when conditional networks were computed.

In the framework of Gaussian graphical models (GGMs), the structure of the inverse covariance matrix Σ^{-1} (also known as the precision matrix) encodes an undirected graph whose edges represent conditional dependencies between variables [10]. Powerful graph inference algorithms for the latter include the graphical lasso (glasso) [11] and neighborhood selection [12]. However and importantly, the gaussian assumption does not accomodate the excess of zeros typical of metagenomic datasets.

Indeed, a key difficulty for the statistical methods is the extreme sparsity and overdispersion of real-world sequence-based abundance data, including non random absences (structural zeros) that can lead to high levels of false positive results. Early population-level metagenomic studies highlighted extreme sparsity patterns in the human microbiome, with a large proportion of taxa being rare and absent in the majority of subjects, resulting in far more zero counts for each taxon than expected on the basis of Poisson, negative binomial, or Dirichlet-multinomial distributions [13].

Given their prevalence, zero values need to be properly handled by the modeling approaches, which requires accounting for the zero generating process in the first place. Zero values are often considered resulting from either the stochastic nature of sampling (sampling zeros), failed measurements arising from technical bias (technical zeros) or zero values resulting from a true absence of signal (structural or "biological" zeros) [14]. Previous work addressing technical zeros (absence of data) include ref [15], while ref [16] is an example of recent work dealing with sampling zeros.

Most of the existing work focuses on Gaussian graphical models, which also includes Poisson log-normal models where counts follow Poisson distributions with parameters sampled from a latent multivariate Gaussian variable (network inference proceeds in the latent space) as illustrated in refs [16–19]. Methods based on log-normal models in the GGM framework include gCoda [20], Banocc [21], and Gaussian copulas with zero-inflated marginals [22, 23]. Outside the realm of GGMs, previous work on Poisson graphical models include ref [15] and [24]. However, there is currently no consistent multivariate Poisson distribution to model dependencies between count variables, as Poisson graphical models fail to have proper joint distribution [24, 25] or to have both marginal and conditional Poisson distributions [26].

While there is an abundant literature about models developed for the reconstruction of ecological networks from microbial occurrence or abundance data, these typically make strong assumptions about the zero generating process (e.g. that all zeros are explained by a common probability model) and don't investigate the consequences of deviating from these assumptions. Yet, the classical distinction between sampling zeros and structural zeros is not refined enough to describe the assumptions underlying commonly used models, like zero-inflated ones, which can lead to substantial biases

under both simulated and real-world data settings [14].

On the other hand, because the number of sequences is not proportional to the total DNA content of samples, sequence data is compositional by nature, which prevents naive correlation-based analyses without appropriate correction [27], of which log-ratio based methods are the most commonly used. In the latter, the handling of zeros by adding a fixed pseudocount has been shown to introduce important biases when the datasets include rare taxa and/or heterogeneous sequence coverage. In that context, it is tempting to ignore samples in order to homogenize sequence depth, or to perform aggressive thresholding by removing rare taxa in order to alleviate this bias, though at the expense of discarding large amounts of valuable and biologically sensible data.

We present here a truncated log-normal graphical model for inferring microbial interaction networks that accounts for both the compositionality of microbial abundance data and specifically addresses zeros originating from biological absence (structural zeros), and discuss consistent methods for estimating sparse and high-dimensional inverse covariance matrices.

Regarding the generative aspect of the model, which is important for producing realistic abundance data (e.g. for benchmarking purposes), an attractive approach (implemented in the popular Spiec-Easi toolkit [22]) relies on Gaussian copulas coupled with count distribution marginals in order to generate multivariate count distributions. Dependencies between variables can be simulated with a latent multivariate Gaussian distribution, which is then composed with count distributions, typically zero-inflated negative binomials. This protocol yields multivariate zero-inflated negative binomial distributions, but it does not produce compositional data, nor does it generate realistic count distributions when the proportion of zeros and the variances are large. We propose a zero-inflated log normal model for generating sparse multivariate counts, and demonstrate that it produces more realistic counts (Figure 1).

2 Motivation of the proposed model

The key motivation to use a multivariate zero-inflated model with a latent multivariate normal variable to account for the dependency structure of the model is partly inspired by the work of Lee et al., in the context of variable selection [13]. Our main motivations are:

- to have a better representation of real counts (e.g. improving over commonly used zero inflated negative binomial models) (Figure 1)
- to have a truly compositional data generator
- to take structural (biological) zeros into account at the network inference level
- to leverage the vast literature and experience available on log-normal models (e.g. [20], [16], [17], among others)

3 Materials and Methods

We describe next the zero inflated log-normal model and a corresponding network inference method. Observed counts are noted y_{ij} where $i = 1..n$ indexes the samples and $j = 1..p$ the variables (taxa). We denote the vector $\boldsymbol{y}_i$ of size p the observation of counts for sample i and more generally in bold letters vectors of size p. The count vectors are gathered into a $n \times p$ matrix Y.

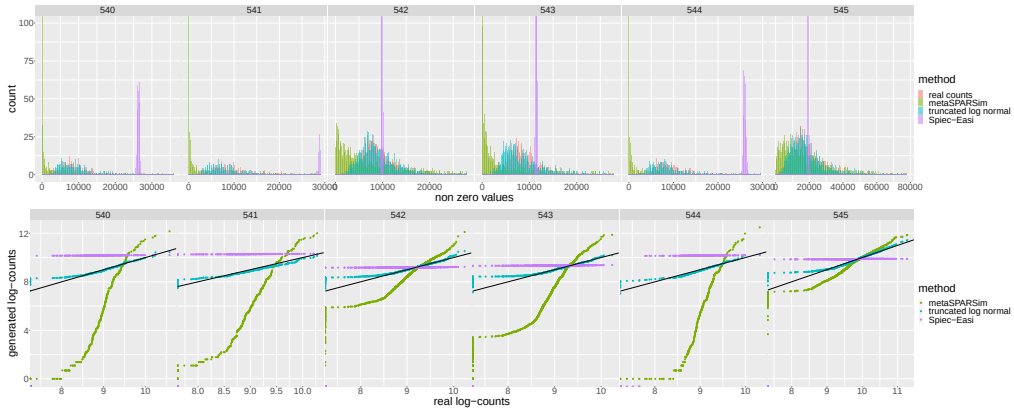

Fig 1. Top: histograms showing real and simulated count data. Real counts obtained from shotgun sequencing the microbiomes of the LifeLines-Deep cohort are shown in red; green, purple and blue represent synthetic counts generated respectively by metaSPARSim [28], spiec-easi [22] and our model (truncated log normal). The different panels show results for different randomly selected taxa. Down: quantile-quantile (QQ)-plots showing the logarithm of real counts vs the logarithm of synthetic counts.

3.1 The model

Our model is derived from the multivariate Poisson log-normal model [27] to which a zero-inflated component is added. We consider a latent multivariate normal variable $\boldsymbol{z}_i$ and a variable $\boldsymbol{a}_i$ representing the real (unknown) abundances. We define the multivariate distribution of $\boldsymbol{a}_i$ as following:

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\ a_{ij} &= \mathbb{1}_{z_{ij} > \delta_j} e^{z_{ij}} \end{aligned} \tag{1}$$

where $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ represents the multivariate normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ and $\mathbb{1}$ is the indicator function.

In the Gaussian Graphical Model, the structure of Σ^{-1} encodes an undirected graph with edges representing conditional dependencies between variables [10]. In this model, zeros are interpreted as biological zeros and are related to a latent gaussian variable.

In metagenomics, the observed counts $\boldsymbol{y}_i$ reflect the abundance proportions $\boldsymbol{\pi}_i$ of microbes present in the sample: $\pi_{ij} = \frac{a_{ij}}{\sum_j a_{ij}}$. The number of reads observed for each sample is generally variable, and we will note it N_i. The counts can then be modeled with a Poisson or multinomial distribution whose parameter is proportional to $\boldsymbol{\pi}_i$:

$$\boldsymbol{y}_i \sim \mathcal{M}(N_i, \boldsymbol{\pi}_i) \tag{2}$$

For example, ref [28] describes a method for simulating 16S rRNA gene count data based on gamma distributions, but it only moderately fits real data when sparsity and overdispersion are high, as shown in Figure 1. The same figure makes apparent that the zero-inflated negative binomial distribution is neither a good fit when parameters of the distribution are adjusted to real-world metagenomic data by using the data generation protocol of ref [22].

3.2 Network inference method

We next describe a network structure inference method suitable for metagenomic data consistent with the generative model presented above.

We will first present a centered log-ratio (clr)-like data transformation preserving scale invariance in the presence of zeros, and then propose an inference method for learning the sparse structure of Σ^{-1} that is inspired by the work of ref [17].

3.2.1 Data transformation

Because of variations in sequencing depths across samples, the total number of reads in a sample only reflects the relative proportion of microbes it contains [29]. This compositional nature of sequence data prevents naive correlation-based analyses without appropriate correction [27], of which log-ratio based methods are the most commonly used. In the latter, the handling of zeros by adding a fixed pseudocount (i.e., irrespective of the sequencing depth of the various samples) has been shown to introduce important biases when the datasets include rare taxa and/or are inhomogeneous in coverage [9]. Downsampling or ignoring samples to level off sequence depth, or aggressive thresholding to remove rare taxa, can somewhat alleviate this bias, although this has been criticised as statistically unsound [30] and is performed at the cost of discarding large amounts of possibly biologically relevant data.

One way to address this difficulty, as originally shown by Aitchison [27] and implemented in popular toolkits (e.g. ref [22]), is to apply a centered-log-ratio (clr) transformation to the counts:

$$y_{ij}^{\text{clr}} = \log(y_{ij}) - \frac{1}{p}\sum_{k=1}^{p}\log(y_{ik}) \tag{3}$$

An important property of this transformation is scale-invariance, which guarantees that two samples with similar read proportions will have similar transformed count data. Indeed, if counts represent absolute abundances with a scaling factor $\boldsymbol{a}_i \approx \lambda \boldsymbol{y}_i$, then we would have $\boldsymbol{a}_i^{\text{clr}} \approx \boldsymbol{y}_i^{\text{clr}}$.

As clr is not defined in zero, a typical workaround consists in adding a unit pseudocount to the original data in order to avoid numerical problems. However, doing so breaks the scale invariance property, as shown in Appendix B. We therefore propose a slightly different version of clr, preserving scale invariance in the presence of zeros and defined as following:

$$\tilde{y}_{ij} = \begin{cases} \log(y_{ij}) - \frac{\sum_{k,\ y_{ik}\neq 0}\log(y_{ik})}{|k,\ y_{ik}\neq 0|} & \text{if } y_{ij} \neq 0 \\ 0 & \text{if } y_{ij} = 0 \end{cases} \tag{4}$$

where $|k,\ y_{ik} \neq 0|$ denotes the number of non zero entries of $\boldsymbol{y_i}$. A proof that this transformation preserves scale invariance in the presence of zeros can be found in Appendix B.

3.2.2 Effect of clr transformation on the estimation of Σ^{-1}

In "canonical" (non zero-inflated) log-normal models, when $\boldsymbol{a}_i \sim \mathcal{LN}(\boldsymbol{\mu}, \Sigma)$, we have $\boldsymbol{a}_i^{\text{clr}} \sim \mathcal{N}(F\boldsymbol{\mu}, F\Sigma F)$, where $F = I - \frac{1}{p}J$ and J is a $p \times p$ matrix filled with ones. Even though there is no guarantee that $F\Sigma F$ is close to Σ (in the sense of the Frobenius norm) when $p >> 0$, as shown in Appendix A, the approximation appears reasonably good in practice (see e.g. [22]).

In the zero inflated log-normal model, we will assume similarly that if $\boldsymbol{a}_i$ follows a distribution as defined in Equation 1, then $\tilde{\boldsymbol{a}}_i$ follows a zero inflated normal distribution defined as:

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{N}(\tilde{\boldsymbol{\mu}}, \tilde{\Sigma}) \\ \tilde{a}_{ij} &= \mathbb{1}_{z_{ij} > \tilde{\delta}_j} z_{ij} \end{aligned} \tag{5}$$

and assume that $\tilde{\Sigma}$ is a good approximation of Σ.

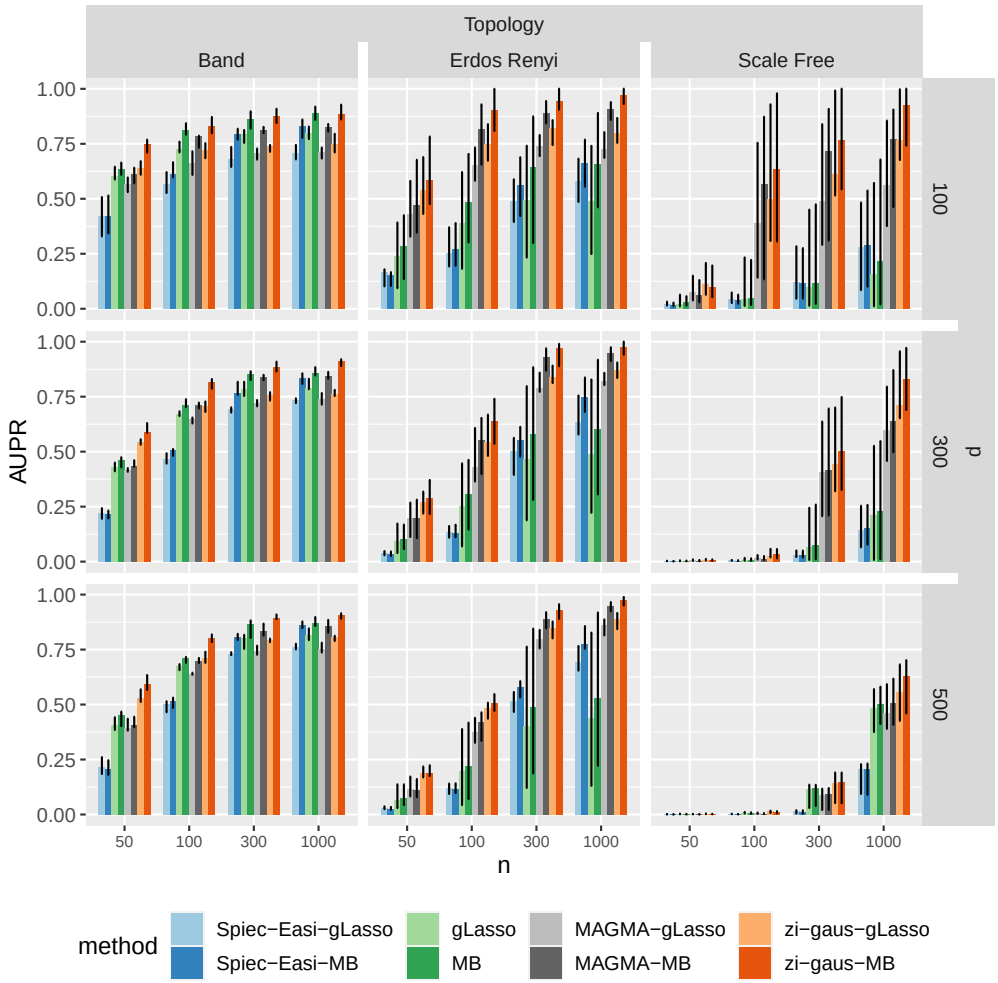

Fig 2. Effect of the number of samples (n), number of taxa (p) and network topology on the performance of the various methods, measured with the area under precision-recall curve (AUPR), on synthetic datasets. Bars represent the median over 10 runs, and error bars +25% and −25% quantiles. Each method (Spiec-Easi (blue) [22], MAGMA (gray) [23] and ours (zi-gaus, orange), as well as no transformation at all (green)), was tested with two structure inference algorithms (glasso and neighborhood selection).

3.2.3 Sparse network inference

We will assume that $\tilde{\boldsymbol{y}}_i \approx \tilde{\boldsymbol{a}}_i$ follows the zero inflated normal distribution defined in equation 5. For network structure inference, we follow a procedure similar in spirit to [17] and [23], with distinct data transformation steps to estimate the latent layer, and whose key steps can be summarized as follows:

- obtain initial estimates of parameters $\hat{\mu}_j$, $\hat{\Sigma} = (\hat{\Sigma}_{11}, \hat{\Sigma}_{22}, .., \hat{\Sigma}_{pp})$ and $\hat{\delta}_j$
- transform data to posterior mean of $Z|\tilde{Y}$ to obtain $\tilde{Z}$
- infer the structure of Σ^{-1} with either the graphical lasso (glasso) [11] using the empirical correlation matrix of $\tilde{Z}$, or neighborhood search (Meinshausen and Bühlmann, MB) [12] on $\tilde{Z}$.

The glasso [11] solves a penalized likelihood maximization problem for the multivariate normal distribution, and Ambroise and Chiquet have shown (personal communication and Appendix D) that the MB algorithm solves a penalized pseudo-likelihood maximization problem. On the other hand, ref [17] showed that this last step is therefore the maximization step of an EM algorithm, the two first steps yielding the expectation step.

3.2.4 Initial parameter estimates

We describe here how to obtain initial estimates for the model parameters $\hat{\boldsymbol{\delta}}, \hat{\boldsymbol{\mu}}, \hat{\Sigma}$. As shown in [17], a diagonal estimation of Σ can be made in order to reduce the computational burden of the method (note ref [31] provides an algorithm for maximizing the penalized likelihood without the diagonality assumption). Under the constraint that $\hat{\Sigma}$ is diagonal, all variables are independent and marginal parameters $\hat{\mu}_j$ and $\hat{\Sigma}_{jj}$ can be estimated separately for each variable. We have the following likelihood from the marginal distributions, which have both a continuous and a discrete part:

$$\mathcal{L}(\mu_j, \Sigma_{jj}, \delta_j | \tilde{\boldsymbol{y}}_i) = \prod_i \left(\mathbb{1}_{\tilde{y}_{ij}=0} \Phi_{\mu,\Sigma_{jj}}(\delta_j) + \mathbb{1}_{\tilde{y}_{ij}\neq 0} f_{\mu,\Sigma_{jj}}(\tilde{y}_{ij}) \right) \tag{6}$$

where $f_{\mu_j,\Sigma_{jj}}$ and $\Phi_{\mu_j,\Sigma_{jj}}$ are respectively the normal distribution and the cumulative normal distribution of mean μ_j and variance Σ_{jj}.

The solution for $\hat{\delta}_j$ is straightforward and independent of other parameters:

$$\hat{\delta}_j = \min_{i,\, y_{ij}\neq 0} \tilde{y}_{ij} \tag{7}$$

The other parameters can be obtained by maximizing the log-likelihood:

$$\begin{aligned} \hat{\mu}_j, \hat{\Sigma}_{jj} = \quad & \text{argmax}_{\mu,\sigma^2} \textstyle\sum_{i,\tilde{y}_{ij}=0} \log\left(\Phi_{\mu,\sigma^2}\left(\hat{\delta}_j\right)\right) \\ & + \textstyle\sum_{i,\tilde{y}_{ij}\neq 0} \log\left(f_{\mu,\sigma^2}\left(\log\left(y_{ij}\right)\right)\right) \end{aligned} \tag{8}$$

We will note $\hat{\Sigma} = \text{diag}(\hat{\Sigma}_{11}, \hat{\Sigma}_{22}, .., \hat{\Sigma}_{pp})$.

3.2.5 Posterior mean transformation

Following ref [17], we use the algorithm on the inferred latent Gaussian variable $\tilde{\boldsymbol{z}}_i$. $\tilde{Z}$ is obtained by posterior mean transformation:

$$\tilde{Z} := \mathbb{E}_{\hat{\boldsymbol{\delta}},\hat{\boldsymbol{\mu}},\hat{\Sigma}} \left[Z | \tilde{Y} \right] \tag{9}$$

As $\hat{\Sigma}$ is diagonal, each value of $\tilde{Z}$ can be computed separately (see Appendix C):

$$\tilde{z}_{ij} = \mathbb{E}_{\hat{\delta}_j, \hat{\mu}_j, \hat{\Sigma}_{jj}} \left[z_{ij} | \tilde{y}_{ij} \right] \tag{10}$$

For which the computation is easy:

$$\tilde{z}_{ij} = \begin{cases} \frac{\int_{-\infty}^{\hat{\delta}_j} y . f_{\hat{\mu}_j, \hat{\sigma}_j}(y) dy}{\Phi_{\hat{\mu}_j, \hat{\sigma}_j}(\hat{\delta}_j)} & \text{if } \tilde{y}_{ij} = 0 \\ \tilde{y}_{ij} & \text{if } \tilde{y}_{ij} \neq 0 \end{cases} \tag{11}$$

Interestingly, we noticed that fixing a small inaccuracy in ref [17] during data transformation at the level of the diagonal terms of $\tilde{S}$ in Equation (19) (see Appendix C) did not increase the support of inferred networks. Our correction actually leads to increased variance of individual observations with less covariance information being shifted across variables, therefore possibly leading to poorer support for the network structure.

3.3 Synthetic datasets

In order to compare the performance of various methods without knowing the true underlying interactions among taxa in real datasets, we designed synthetic datasets with a controlled ground truth.

For our simulations, we tested two generative models. The first is the model described in section 3.1 with equations (1) and (2) (zi-gaus). The second is the model implemented in [22] using Gaussian copulas and zero-inflated negative binomial marginals (combined in the NorTA protocol). The covariance matrix Σ associated to a given graph topology is generated using the Spiec-Easi framework [22], with condition number $\kappa = 100$ and the number of edges e equal to the number of variables p.

3.3.1 Synthetic networks for Simulation 1

In the first simulation, we used the zi-gaus generative model in conjunction with different network topologies (band, Erdos-Renyi and scale-free), and varying sample numbers $n = 50, 100, 300, 1000, 2000$ and taxa numbers $p = 100, 200, 500$. The library sizes N_i were drawn from negative binomial distributions $N_i \sim \mathcal{NB}(\text{mean} = 1.5 \times 10^6, \text{size} = 5)$. The sparsity levels of taxa s_j (i.e. the expected proportion of zeros) was drawn from a uniform distribution $s_j \sim \text{unif}(0, 0.9)$, bringing the total number of zeros to 45% of count values. The correspondence to the parameter $\boldsymbol{\delta}$ is given by: $\delta_j = \Phi^{-1}_{\mu_j, \sigma_j^2}(s_j)$, and the mean of the gaussian distribution was also chosen randomly from a uniform distribution: $\mu_j \sim \text{unif}(0, 3)$.

3.3.2 Synthetic networks for Simulation 2

In the second simulation, the sparsity level of the variables is varied while p and n are fixed to $p = 300$ and $n = 100$. We used a deterministic formulation for $\boldsymbol{s}$: $s_j = \frac{j}{d}^t$ where t is chosen so that the mean of s_j is equal to the target global proportion of zeros which can take values from $0, 10, 50, 70$ and 90%. The parameters $\boldsymbol{\mu}$ and N_i are chosen as in Simulation 1.

For this second simulation, we also used the NorTA framework from ref [22] for comparative purposes, with a latent normal distribution $\boldsymbol{z}_i \sim \mathcal{N}(0, \Sigma)$ and count values generated by composition with marginal inverse cumulative distribution functions (cdf) $\boldsymbol{y}_i = \left(F_1^{-1}(\Phi_{0,\Sigma_{11}}(z_{i1})), F_2^{-1}(\Phi_{0,\Sigma_{22}}(z_{i2})), .., F_p^{-1}(\Phi_{0,\Sigma_{pp}}(z_{ip})) \right)$, where F_j are cdf of zero-inflated negative binomial distributions. As parameters for those distributions, we took means as $\text{mean}_j = e^{\mu_j}$, a size ν fixed to $\nu = 10$, and a probability of zero equal to s_j.

4 Results

4.1 Accounting for sparsity and overdispersion

In order to assess the ability of our model to deal with the high levels of sparsity and overdispersion that are inherent to natural microbial abundance datasets, we compared the extent to which synthetic count data generated under our model and two other model-based approaches are able to reproduce abundances of taxonomic profiles generated from human gut microbiomes of the population-based LifeLines-Deep cohort [32] (using the metapalette profiling tool [33]).

The performance of our model was compared to two other methods generating synthetic count data by integrating information from experimental count distributions in their model-based generative process: spiec-easi [22] and metasparsim [28]. Spiec-easi implements a Normal to Anything (NorTA) approach [34] to generate correlated count data by coupling Gaussian copulas with count distribution marginals (typically zero-inflated negative binomials) fitted on real-world count data [22], while metasparsim is a tool designed to generate 16S rRNA gene counts based on a two-step gamma multivariate hypergeometric model [28].

The figure 1 provides two different views of the performance of the three methods, with the top panel showing count histograms for different randomly selected taxa, while the bottom panel presents the same results in the form of quantile-quantile plots (QQ-plots). Both representations make apparent that the zero inflated log-normal model provides a better fit to the very sparse and overdispersed data extracted from the real-world microbiomes.

This observation also hints that state of the art methods (refs [22] and [28]) used to simulate multivariate count data from rRNA amplicons fail to properly account for counts derived from shotgun metagenomic data, probably because of higher overdispersion in the latter.

4.2 Accuracy of microbial network inference

We compared our method against two state of the art methods for inferring microbial association networks. Magma [23] is a method for detecting microbial associations specifically aiming to deal with an excess of zero counts, while also taking compositionality and overdispersion into account; it is based on Gaussian copula models with marginals modeled with zero-inflated negative binomial generalized linear models, and the core algorithm for network structure inference is the graphical lasso, with its inference procedure relying on the estimation of the latent data by a specific medium imputation procedure [23]. Spiec-easi [22] is a popular and widely used toolkit for inferring microbial ecological networks; it is designed to deal with both compositionality and sparsity, and relies on algorithms for sparse neighborhood and inverse covariance selection (i.e. the graphical lasso) for network structure inference. We performed a comparison of these three model-based methods in terms of their ability to accurately recover the structure of controlled networks of different dimensions, topologies and sampling regimes (described in the subsection 3.3.1 of the Material and Methods). The performance of the methods was quantified by the area under precision-recall curves (AUPR, which should be preferred over ROC-AUC on imbalanced datasets with a lot of true negatives like microbiome taxonomic data), and is displayed in Figure 2 for synthetic datasets encompassing different topologies (band, Erdos-Renyi, scale-free), dimensions ($p = 100, 300, 500$) and sample numbers ($n = 50, 100, 300, 1000, 2000$).

Besides the expected dependence of the results on network topology, dimension and sample number, this figure makes apparent significant performance gains of the proposed method, in particular for the recovery of networks endowed with scale-free

topologies that were shown to be the most difficult to reconstruct accurately in previous studies [22]. It should be noted however that all methods show only limited accuracy for this network topology when operating in the under-determined regime, where the number of variables (taxa, p) exceeds the number of samples (n). It should also be noted that the MB algorithm appears to perform better than the graphical lasso. A similar figure, presenting network reconstruction accuracy metrics for a distinct dataset generated using the NORTA protocol with zero inflated negative binomials (see section 3.3 of the Methods), is shown in the Supplemental Information.

On the other hand, the experiments designed to probe the effects of increasing levels of sparsity (using the datasets described in the subsection 3.3.2 of the Material and Methods) make clear that all three methods are strongly affected by increasing proportions of zeros, as shown in Figure 3, with no method performing accurately under very high levels of sparsity (i.e., 90% of zeros).

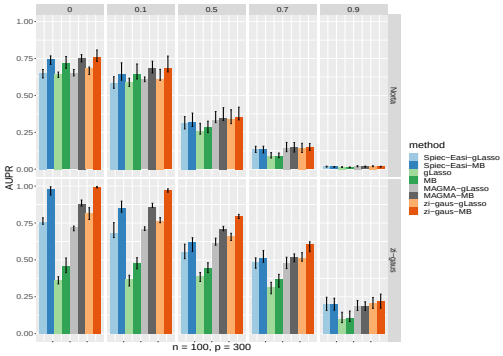

Fig 3. Effect of the amount of zeros in the data (vertical panels) on method performance. Top panel: network reconstruction accuracy for various methods using synthetic data generated with the NORTA protocol using zero inflated negative binomial marginals (see Methods). Bottom panel: network reconstruction accuracy using data generated under the zero inflated log-normal (zi-gaus) model. Notations are as in Figure 2, and results are shown for the Erdos-Renyi topology.

4.3 Inference of real-world microbial association networks

We finally applied our method to infer microbial association networks from real-world taxonomic profiles generated from healthy gut microbiomes of the LifeLines-Deep [32] population cohort (n=1135) (using the metapalette software [33]). Even though confirmed ecological interactions within natural microbial communities are only scarcely known, several global network properties have been reproducibly documented, including the salient observation of preferential associations among phylogenetically close organisms [35, 36]. This property, known as assortativity, is manifest at the level of highly connected components in our analysis, and is displayed for the species taxonomic level in Figure 4.

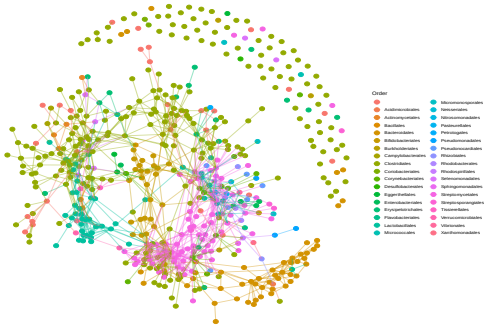

Fig 4. Species-level association network inferred from the LifeLines-Deep microbiomes. Nodes represent species and are colored according to their taxonomic order.

5 Conclusion

Accurate inference of microbial interaction networks is necessary in order to gain a deeper understanding of the functioning of microbial consortia and to take advantage of the increasingly large body of data that is generated worldwide through environmental genomics initiatives. The formalism of Gaussian graphical models naturally lends itself to the problem of network reconstruction, but despite attractive mathematical attributes and the availability of powerful inference algorithms, application of this framework to the identification of robust microbial associations has only been met with moderate success. It appears unlikely though that the problem lies in the GGM formalism itself or in inherently flawed inference algorithms; instead, it is most likely rooted in difficulties to properly account for the idiosyncrasies of biological data. Beyond compositionality and overdispersion, high levels of sparsity are common attributes of real-world biological datasets, which are challenging to account for in a statistically sound manner. In particular, an important distinction between sampling zeros and structural zeros is often neglected, and nearly all of the actual analyses of multidimensional metagenomic count data perform some form of aggressive thresholding to discard rare but possibly biologically relevant data.

We presented here a simple statistical model specifically aimed at accounting for large amounts of structural (biological) zeros in the network inference process, and provide evidence of its usefulness by showing performance gains with respect to several state of the art methods for microbial network inference. Overall, this work should be viewed as a small contribution towards the goal of more fully integrating key attributes of biological data into established and powerful computational frameworks, like Gaussian graphical models.

6 Acknowledgments

We would like to thank Christophe Ambroise and Julien Chiquet for helpful discussions and advices, and Sawsan Kanj for generating the taxonomic profiles of the LifeLines-Deep cohort.

References

1. Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. science. 2008;320(5879):1034–1039.

2. Wagner M, Loy A. Bacterial community composition and function in sewage treatment systems. Current opinion in biotechnology. 2002;13(3):218–227.

3. Schmidt I, Sliekers O, Schmid M, Bock E, Fuerst J, Kuenen JG, et al. New concepts of microbial treatment processes for the nitrogen removal in wastewater. FEMS microbiology reviews. 2003;27(4):481–492.

4. Ianiro G, Rossi E, Thomas AM, Schinzari G, Masucci L, Quaranta G, et al. Faecal microbiota transplantation for the treatment of diarrhoea induced by tyrosine-kinase inhibitors in patients with metastatic renal cell carcinoma. Nature communications. 2020;11(1):1–6.

5. Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Current opinion in microbiology. 2016;31:217–226.

6. Goers L, Freemont P, Polizzi KM. Co-culture systems and technologies: taking synthetic biology to the next level. Journal of The Royal Society Interface. 2014;11(96):20140065.

7. Faust K, Raes J. Microbial interactions: from networks to models. Nature Reviews Microbiology. 2012;10(8):538–550.

8. Weiss S, Van Treuren W, Lozupone C, Faust K, Friedman J, Deng Y, et al. Correlation detection strategies in microbial data sets vary widely in sensitivity and precision. The ISME journal. 2016;10(7):1669–1681.

9. Tackmann J, Rodrigues JFM, von Mering C. Rapid inference of direct interactions in large-scale ecological networks from heterogeneous microbial sequencing data. Cell systems. 2019;9(3):286–296.

10. Lauritzen SL. Graphical Models. Oxford University Press; 1996.

11. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical LASSO. Biostatistics (Oxford, England). 2008;9:432–41. doi:10.1093/biostatistics/kxm045.

12. Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Statist. 2006;34(3):1436–1462. doi:10.1214/009053606000000281.

13. Lee KH, Coull BA, Moscicki AB, Paster BJ, Starr JR. Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. Biostatistics. 2018;doi:10.1093/biostatistics/kxy067.

14. Silverman JD, Roche K, Mukherjee S, David LA. Naught all zeros in sequence count data are the same. bioRxiv. 2018;doi:10.1101/477794.

15. Choi H, Gim J, Won S, Kim Y, Kwon S, Park C. Network analysis for count data with excess zeros. BMC Genetics. 2017;18. doi:10.1186/s12863-017-0561-z.

16. Chiquet J, Robin S, Mariadassou M. Variational Inference for sparse network reconstruction from count data. In: Chaudhuri K, Salakhutdinov R, editors. Proceedings of the 36th International Conference on Machine Learning. vol. 97 of Proceedings of Machine Learning Research. Long Beach, California, USA: PMLR; 2019. p. 1162–1171. Available from: `http://proceedings.mlr.press/v97/chiquet19a.html`.

17. Sinclair D, Hooker G. Sparse inverse covariance estimation for high-throughput microRNA sequencing data in the Poisson log-normal graphical model. Journal of Statistical Computation and Simulation. 2019;89(16):3105–3117. doi:10.1080/00949655.2019.1657116.

18. Biswas S, Mcdonald M, Lundberg D, Dangl J, Jojic V. Learning Microbial Interaction Networks from Metagenomic Count Data. Journal of Computational Biology. 2016;23:526–535. doi:10.1089/cmb.2016.0061.

19. Wu H, Deng X, Ramakrishnan N. Sparse Estimation of Multivariate Poisson Log-Normal Models from Count Data; 2016.

20. Fang H, Huang C, Zhao H, Deng M. gCoda: Conditional Dependence Network Inference for Compositional Data. Journal of Computational Biology. 2017;24. doi:10.1089/cmb.2017.0054.

21. Schwager E, Mallick H, Ventz S, Huttenhower C. A Bayesian method for detecting pairwise associations in compositional data. PLOS Computational Biology. 2017;13:e1005852. doi:10.1371/journal.pcbi.1005852.

22. Kurtz ZD, Müller CL, Miraldi ER, Littman DR, Blaser MJ, Bonneau RA. Sparse and Compositionally Robust Inference of Microbial Ecological Networks. PLOS Computational Biology. 2015;11(5):1–25. doi:10.1371/journal.pcbi.1004226.

23. Cougoul A, Bailly X, Wit EC. MAGMA: inference of sparse microbial association networks. bioRxiv. 2019;doi:10.1101/538579.

24. Gallopin M, Rau A, Jaffrézic F. A Hierarchical Poisson Log-Normal Model for Network Inference from RNA Sequencing Data. PloS one. 2013;8:e77503. doi:10.1371/journal.pone.0077503.

25. Allen GI, Liu Z. A log-linear graphical model for inferring genetic networks from high-throughput sequencing data. In: 2012 IEEE International Conference on Bioinformatics and Biomedicine. IEEE; 2012. p. 1–6.

26. Yang E, Allen G, Liu Z, Ravikumar PK. Graphical models via generalized linear models. In: Advances in Neural Information Processing Systems; 2012. p. 1358–1366.

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

28. Patuzzi I, Baruzzo G, Losasso C, Ricci A, Di Camillo B. metaSPARSim: a 16S rRNA gene sequencing count data simulator. BMC Bioinformatics. 2019;20. doi:10.1186/s12859-019-2882-6.

29. Weiss S, Xu Z, Peddada S, Amir A, Bittinger K, González A, et al. Normalization and microbial differential abundance strategies depend upon data characteristics. Microbiome. 2017;5. doi:10.1186/s40168-017-0237-y.

30. McMurdie PJ, Holmes S. Waste not, want not: why rarefying microbiome data is inadmissible. PLoS Comput Biol. 2014;10(4):e1003531.

31. Gégout-Petit A, Muller-Gueudin A, Karmann C. Graph estimation for Gaussian data zero-inflated by double truncation; 2019. Available from: `https://hal.archives-ouvertes.fr/hal-02367344`.

32. Zhernakova A, Kurilshikov A, Bonder MJ, Tigchelaar EF, Schirmer M, Vatanen T, et al. Population-based metagenomics analysis reveals markers for gut microbiome composition and diversity. Science. 2016;352(6285):565–569.

33. Koslicki D, Falush D. MetaPalette: A K-mer painting approach for metagenomic taxonomic profiling and quantification of novel strain variation. MSystems. 2016;1(3):e00020–16.

34. Nelsen RB. An introduction to copulas. Springer Science & Business Media; 2007.

35. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS computational biology. 2012;8(7).

36. Hall CV, Lord A, Betzel R, Zakrzewski M, Simms LA, Zalesky A, et al. Co-existence of network architectures supporting the human gut microbiome. iScience. 2019;22:380–391.

Supporting information

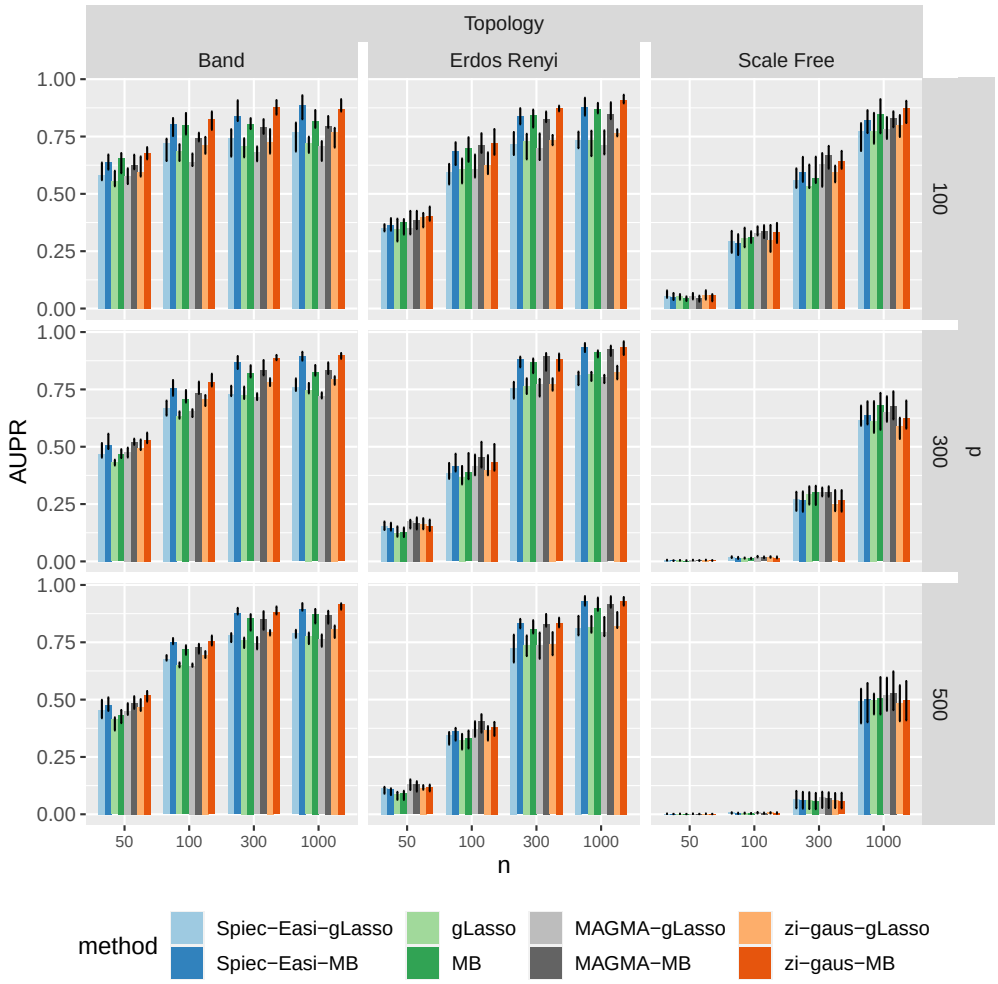

Fig 5. Effect of the number of samples (n), number of taxa (p) and network topology on the performance of the various methods, measured with the area under precision-recall curve (AUPR), on synthetic datasets generated with the NORTA approach using zero inflated negative binomial marginals (see Methods). Bars represent the median over 10 runs, and error bars +25% and −25% quantiles. Each method (Spiec-Easi (blue) [22], MAGMA (gray) [23] and ours (zi-gaus, orange), as well as no transformation at all (green)), was tested with two structure inference algorithms (glasso and neighborhood selection).

A Problem with compositionality under the gaussian assumption

If we assume that $\boldsymbol{x}$ follows a gaussian distribution $\boldsymbol{x} \sim \mathcal{N}(\mu, \Sigma)$, then $\boldsymbol{x}^{\text{clr}}$ also follows a gaussian distribution $\mathcal{N}(F\boldsymbol{\mu}, F\Sigma F)$. However, $F\Sigma F$ does not necessarily converge to Σ when p increases, as shown by taking the Frobenius norm of the difference between Σ and $F\Sigma F$:

$$\|\Sigma - F\Sigma F\|_{\mathcal{F}}^2 = \left\| \frac{2}{p} J\Sigma - \frac{1}{p} J\Sigma J \right\|^2 \tag{12}$$

$$= \sum_{ij} \left(\frac{2}{p} \sum_k \sigma_{ik} - \frac{1}{p^2} \sum_{kl} \sigma_{kl} \right)^2 \tag{13}$$

$$= p \sum_i l_i^2 \tag{14}$$

with

$$l_i = \frac{2}{p} \sum_k \sigma_{ik} - \frac{1}{p^2} \sum_{kl} \sigma_{kl}$$

B Problem with the clr transformation when there is an excess of zeros

A key property of the clr transformation is the equivalence between log-ratios of compositional and log-ratios of absolute data. For example, if we assume that $\boldsymbol{x}_i$ is scaled by N,

$$\boldsymbol{y}_i = N\boldsymbol{x}_i$$

then we have

$$y_{ij}^{\text{clr}} = x_{ij}^{\text{clr}}$$

If we combine now a clr transformation with pseudo counts, we obtain $y_{ij}^{\text{clr, s-e}}$ as

$$y_{ij}^{\text{clr, s-e}} = \log(y_{ij} + 1) - \frac{1}{p} \sum_{k=1}^{p} \log(y_{ik} + 1)$$

If $\boldsymbol{y}_i$ contains some zero values, we will have

$$\begin{aligned} y_{ij}^{\text{clr, s-e}} &= \log(y_{ij} + 1) - \tfrac{1}{p} \textstyle\sum_{k=1}^{p} \log(y_{ik} + 1) \\ &\approx \log(y_{ij}) - \tfrac{1}{p} \textstyle\sum_{k, y_{ik} \neq 0} \log(y_{ik}) \\ &= \log(N x_{ij}) - \tfrac{1}{p} \textstyle\sum_{k, y_{ik} \neq 0} \log(N x_{ik}) \\ &= \log(x_{ij}) - \tfrac{1}{p} \textstyle\sum_{k, y_{ik} \neq 0} \log(x_{ik}) + \tfrac{1}{p} \left|k, y_{ik} = 0\right| \log(N) \\ &= x_{ij}^{\text{clr}} + \tfrac{1}{p} \left|k, y_{ik} = 0\right| \log(N) \end{aligned} \tag{15}$$

with $|k, y_{ik} = 0|$ the number of zero entries of $\boldsymbol{y}_i$. It is apparent that log transformed ratios are not scale-invariant in the presence of zero values, and that the difference increases with the number of zeros in the vector.

Thus, if the data is very sparse, and if the sequencing depth further has high variance, then the relevance of the clr transformation diminishes.

C One step EM

Assuming the following probabilistic model:

$$\begin{aligned} \boldsymbol{z}_i &\sim \mathcal{N}(\boldsymbol{\mu}, \Sigma) \\ \tilde{y}_{ij} &= I(z_{ij} > \delta_j) z_{ij}, \end{aligned} \tag{16}$$

an EM argument can show that the penalized likelihood can be increased by maximizing:

$$L^1(\Sigma^{-1}) = \mathbb{E}_{Z|\tilde{Y}}\left(\log(\mathbb{P}(Z|\boldsymbol{\mu}, \Sigma^{-1}))\right) - \lambda\|\Sigma^{-1}\|_1 \tag{17}$$

$$= \frac{n}{2}\log|\Sigma^{-1}| - \frac{n}{2}\mathbb{E}_{Z|\tilde{Y}}\left(tr(S\Sigma^{-1})\right) - \lambda\|\Sigma^{-1}\|_1 + \text{const} \tag{18}$$

where $S := \frac{1}{n}\sum_i(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}})(\boldsymbol{z}_i - \hat{\boldsymbol{\mu}})^T$.

When taking as initial parameters the diagonal estimates $\Sigma_0 = \text{diag}(\Sigma_{11}, \Sigma_{22}, .., \Sigma_{pp})$, the equation can be simplified with the following equality (see [17]):

$$\mathbb{E}_{Z|\tilde{Y}}\left(\text{tr}(S\Sigma^{-1})\right) = \text{tr}(\tilde{S}\Sigma^{-1})) \tag{19}$$

where $\tilde{S}$ is the empirical covariance matrix of the posterior mean of Z, noted $\tilde{Z}$ and defined as following:

$$\tilde{z}_{ij} = \mathbb{E}_{z_{ij}|\tilde{y}_{ij}}(z_{ij})$$

We noticed here a small inaccuracy in [17], affecting equation (19) and based on the following observation (where $\boldsymbol{\mu} = 0$ for more simplicity; we also remind the reader of the notation $\Omega = \Sigma^{-1}$).

$$\mathbb{E}_{Z|\tilde{Y}}\left(\text{tr}(S\Omega)\right) = \mathbb{E}_{Z|\tilde{Y}}\left(\sum_{jk} S_{jk}\Omega_{jk}\right) \tag{20}$$

$$= \mathbb{E}_{Z|\tilde{Y}}\left(\sum_{jk}\left(\frac{1}{n}\sum_i z_{ij}z_{ik}\right)\Omega_{jk}\right) \tag{21}$$

$$= \sum_{jk}\Omega_{jk}\frac{1}{n}\sum_i \mathbb{E}_{\boldsymbol{z}_i|\tilde{\boldsymbol{y}}_i}\left(z_{ij}z_{ik}\right) \tag{22}$$

$$= \sum_{jk}\Omega_{jk}\frac{1}{n}\sum_i \mathbb{E}_{z_{ij}|\tilde{y}_{ij}}(z_{ij})\mathbb{E}_{z_{ij}|\tilde{y}_{ij}}(z_{ik}) \tag{23}$$

$$= \text{tr}(\tilde{S}\Sigma^{-1})) \tag{24}$$

The passage from (22) to (23) relies on the equality:

$$\mathbb{E}_{\boldsymbol{z}_i|\tilde{\boldsymbol{y}}_i}\left(z_{ij}z_{ik}\right) = \mathbb{E}_{z_{ij}|\tilde{y}_{ij}}(z_{ij})\mathbb{E}_{z_{ij}|\tilde{y}_{ij}}(z_{ik})$$

justified by the fact that Σ_0 is diagonal. This is true for all j, k when $j \neq k$ but doesn't hold when $j = k$.

D Neighbourhood selection as a penalized maximization problem

The argument follows unpublished work from Ambroise and Chiquet: "solving MB is equivalent to maximize a penalized pseudo likelihood".

The pseudo likelihood p-$\mathcal{L}$ of a gaussian vector $\boldsymbol{x}$ is an approximation of its likelihood and is defined as following:

$$\text{p-}\mathcal{L}(\boldsymbol{\mu}, \Sigma|\boldsymbol{x}) = \prod_i^p \mathbb{P}(x_i|\boldsymbol{x}_{\setminus i})$$

Let's assume that $\tilde{\boldsymbol{y}}_i$ follows a normal distribution: $\tilde{\boldsymbol{y}}_i \sim \mathcal{N}(\boldsymbol{\mu}, \Sigma)$, so we can compute the distribution of variable i conditioned on other variables:

$$\tilde{y}_{ij}|\tilde{\boldsymbol{y}}_{i\backslash j} \sim \mathcal{N}(\mu_{j|\backslash j}, \Sigma_{j|\backslash j})$$

with

$$\mu_{j|\backslash j} = \mu_j + \Sigma_{j\backslash j}\Sigma_{\backslash j\backslash j}^{-1}(\tilde{\boldsymbol{y}}_{\backslash j} - \boldsymbol{\mu}_{\backslash j})$$

and

$$\Sigma_{j|\backslash j} = \Sigma_{jj} - \Sigma_{j\backslash j}\Sigma_{\backslash j\backslash j}^{-1}\Sigma_{j\backslash j}^T$$

By taking the logarithm, we can develop

$$\log P(\tilde{y}_{ij}|\tilde{\boldsymbol{y}_{i\backslash j}}) = \frac{1}{2}\log(\frac{1}{\Sigma_{j|\backslash j}}) - \frac{1}{2\Sigma_{j|\backslash j}}(y_{ij} - \mu_{j|\backslash j})^2 + \text{const}$$

Now assume $\boldsymbol{\mu} = 0$ for simplicity, and note $\Omega = \Sigma^{-1}$. Developments for block matrix inversion give us:

$$\Omega_{j\backslash j} = -\left(\Sigma_{jj} - \Sigma_{j\backslash j}\Sigma_{\backslash j\backslash j}^{-1}\Sigma_{j\backslash j}^T\right)^{-1}\Sigma_{j\backslash j}\Sigma_{\backslash j\backslash j}^{-1}$$

and

$$\Omega_{jj} = \left(\Sigma_{jj} - \Sigma_{j\backslash j}\Sigma_{\backslash j\backslash j}^{-1}\Sigma_{j\backslash j}^T\right)^{-1}$$

We notice that:

$$\Sigma_{j\backslash j} = \frac{1}{\Omega_{jj}}$$

and

$$\mu_{j|\backslash j} = -\frac{\Omega_{j\backslash j}}{\Omega_{jj}}\tilde{\boldsymbol{y}}_{\backslash j}$$

Those terms can then be injected in the log pseudo likelihood expression:

$$\log \mathbb{P}\left(\tilde{y}_{ij}|\tilde{\boldsymbol{y}}_{i\backslash j}\right) = \frac{1}{2}\log\left(\Omega_{jj}\right) - \frac{1}{2}\Omega_{jj}\left(\tilde{y}_{ij} + \frac{\Omega_{j\backslash j}}{\Omega_{jj}}\tilde{\boldsymbol{y}}_{i\backslash j}\right)^2 + \text{const}$$

Fixing Ω_{jj} to a constant:

$$\log \mathbb{P}\left(\tilde{y}_{ij}|\tilde{\boldsymbol{y}}_{i\backslash j}\right) = -\frac{\Omega_{jj}}{2}\left(\tilde{y}_{ij} - \frac{\Omega_{j\backslash j}}{\Omega_{jj}}\tilde{y}_{i\backslash j}\right)^2 + \text{const}$$

Finally, the whole penalized log pseudo likelihood can be expressed as:

$$\begin{aligned}
\text{p-}\mathcal{L}\left(\Omega_{\text{off}}|\tilde{Y}\right) &= \sum_j\sum_i -\frac{\Omega_{jj}}{2}\left(\tilde{y}_{ij} - \frac{\Omega_{j\backslash j}}{\Omega_{jj}}\tilde{\boldsymbol{y}}_{i\backslash j}\right)^2 - \lambda\|\Omega_{\text{off}}\|_1 + \text{const} \\
&= \sum_j -\frac{\Omega_{jj}}{2}\|\tilde{Y}_{.j} - \frac{\Omega_{j\backslash j}}{\Omega_{jj}}\tilde{Y}_{.\backslash j}^T\|_2^2 - \lambda\|\Omega_{\text{off}}\|_1 + \text{const}
\end{aligned}$$

Where the parameter Ω_{off} is the off-diagonal values of Ω.

Solving the lasso regressions of the MB algorithm is therefore equivalent to maximizing the penalized pseudo likelihood with the off-diagonal values of the inverse covariance matrix as parameter.