Random Graphical Model of Microbiome Interactions in Related Environments

Veronica VINCIOTTI, Ernst C. WIT, and Francisco RICHTER

The microbiome constitutes a complex microbial ecology of interacting components that regulates important pathways in the host. Most microbial communities at various body sites tend to share common substructures of interactions, while also showing diversity related to the needs of the local environment. The aim of this paper is to develop a method for inferring both the common core and the differences in such microbiota systems. The approach combines two elements: (i) a random graph model generating networks across environments, and capturing potential relatedness at the structural level, with (ii) a Gaussian copula graphical model for the inference of environment-specific networks from multivariate microbial data. We propose a Bayesian approach for the joint inference of microbiota systems from metagenomic data for a number of body sites. The analysis of human microbiome data shows how the proposed random graphical model is able to capture varying levels of structural similarity across the different body sites and how this is supported by their taxonomical classification. Beyond a stable core, the inferred microbiome systems show interesting differences between the body sites, as well as interpretable relationships between various classes of microbes.

Key Words: Microbiome; Graphical models; Random graph model; Bayesian inference.

1. INTRODUCTION

The microbiome constitutes a complex microbial ecology of interacting components that regulates important pathways in the host. Microbiotic systems have been intensively studied in recent years, and they have been found to shape the health of plants and animals (Kost et al. 2023). In humans, associations have been found with a number of health conditions, such as obesity (Le Chatelier et al. 2013), diabetes (Pedersen et al. 2016) and the response to immunotherapy (Lee et al. 2022). Rich sources of high-throughput data of the microbiome, such as those generated by the Human Microbiome Project (Consortium 2012) and the Metagenomics of the Human Intestinal Tract (MetaHIT) project (Qin et al. 2010), or the CRAMdb database for animal microbiomes (Lei et al. 2022), are key to learning the intricate network of interactions among microbial communities.

V. Vinciotti (✉) Department of Mathematics, University of Trento, Trento, Italy
(E-mail: *veronica.vinciotti@unitn.it*).
E. C. Wit · F. Richter, Institute of Computing, Università della Svizzera Italiana, Lugano, Switzerland.

Journal of Agricultural, Biological, and Environmental Statistics, Volume 31, Number 1, Pages 46–59
https://doi.org/10.1007/s13253-024-00638-6

As the microbiome interacts with the local environment, the microbiome varies in constitution profile at different sites in the host (Sharon et al. 2022). For example, Segata et al. (2012) find four groups of digestive tract sites in the human body, characterised by distinct bacterial compositions and metabolic processes. Despite this heterogeneity, it is expected that the interaction profile is largely shared between different body sites from a structural perspective. This constitutes a core microbiome network, describing stable components of the microbiome interactions across time, body sites and populations.

Most studies on animals and humans rely solely on faecal samples to represent the gut microbiome and on saliva samples to describe the oral microbiome (Kim et al. 2023; Sharon et al. 2022). Equally, available methods and implementations, such as the commonly used SparCC (Friedman and Alm 2012) and SPIEC-EASI (Kurtz et al. 2015), infer a single microbiota system from abundance data obtained from a single body site. As such, they are suited to learn either environment-specific systems from microbiome data on that environment, or some consensus microbiome network from pooled data across different body sites. Instead, we propose a Bayesian approach for the joint inference of microbiota systems from metagenomic data for a number of body sites that captures both the core metabolic network as well as individual differences.

Vinciotti et al (2022) developed a Gaussian copula graphical model to infer microbiota systems from count genomic data. While the parametric form used for the marginals is able to capture both the heterogeneity of microbial abundances across different body sites and the typical features of microbial data, such as zero inflation and compositionality, the approach recovers only a consensus microbiome network. In this paper, we extend the method to infer structured body site-specific microbiome networks.

In Sect. 2, we describe the random graphical model and the Bayesian inference procedure in detail, while in Sect. 3, we validate the method on simulated data, before presenting the results on the Human Microbiome Project study on 87 microbes across 13 body sites. Our analysis shows that the latent space is able to capture the biological relatedness between the 13 microbiotic systems. Indeed, the locations of the body sites in the inferred latent space match closely both with the classification made by Segata et al. (2012) and with the Uberon anatomy classification of body sites (Mungall et al. 2012). The environment-specific networks, and in particular their associated estimated edge probabilities, can be queried further, in order to characterise the individual networks as well as to highlight commonalities and differences between the 13 environments. Beyond the information that can be discovered from the data using the proposed model, we find that the new approach leads to a more stable recovery of the microbiotic systems, compared to individual analyses conducted for each body site separately. In Sect. 4, we discuss the wider implications of the method and present a conclusion.

2. METHODS

We propose a model for capturing heterogeneity at the structural level of microbial interactions, while quantifying the possible relatedness among microbiota systems from different environments. To this end, we augment the model of Vinciotti et al (2022) with a

random graph model on the conditional independence graphs that describe the joint microbial count distributions at each body site. We define a novel *random graphical model* as the combination of a graphical model with an associated random independence graph model.

The random conditional independence graph model can depend on external covariates (Ni et al. 2022) or be defined endogenously. Borrowing from the network science literature (Hoff et al. 2002), we formalise the random graph model as a latent probit network model, where the probability of an edge in a particular microbiota system depends on a latent space of potentially related environments, i.e. it will increase if the body site is close in this latent space to another environment where that particular edge is present. In addition, the edge probability depends on individual network sparsity levels for each body site and on external covariates at the network level. For the latter, we consider the effect of taxonomy sharing on the propensity of microbes to interact, but, in principle, any other covariate or external knowledge can be included at this stage.

2.1. Random Graphical Model

In this section, we define the *random graphical model* for network inference from heterogeneous microbiome data from a number of environments. For environment $k = 1, \ldots, B$, let $\mathbf{Y}^{(k)} = (Y_1^{(k)}, \ldots, Y_p^{(k)})$ be the random p-dimensional vector of interest, consisting of the abundances of p Operational Taxonomic Units (OTUs). In our study, the number of environments corresponds with the $B = 13$ different body sites, in which we measure $p = 87$ microbes. We assume $\mathbf{Y}^{(k)}$ constitutes a *graphical model* (GM),

$$\mathbf{Y}^{(k)} | G^{(k)} \sim \mathcal{L}_{G^{(k)}}(\mathbf{\Omega}^{(k)}),$$

relative to some conditional independence graph $G^{(k)}$ with some associated parameters $\mathbf{\Omega}^{(k)}$. Furthermore, we assume that the graphs $G = \left\{G^{(k)}\right\}_k$ are themselves distributed according to a joint **random graph model**,

$$G^{(k)} \sim P(\mathbf{\Theta}), \quad k = 1, \ldots, B$$

for some vector of parameters $\mathbf{\Theta}$.

The type of graphical model and the type of random graph model can depend on the situation under consideration. As for the graphical model, we consider the Gaussian copula graphical model, due to its easy mathematical formulation and its flexibility in modelling multivariate non-Gaussian data, such as the count microbiome data under consideration. Thus, similarly to Cougoul et al. (2019) and Vinciotti et al (2022), we assume:

$$P(Y_1^{(k)} \leq y_1, \ldots, Y_p^{(k)} \leq y_p \mid G^{(k)}, \mathbf{\Omega}^{(k)}) = \Phi_{\mathbf{\Omega}^{(k)}}\left(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_p(y_p))\right),$$

where $\Phi_{\mathbf{\Omega}^{(k)}}$ is the cumulative distribution function of a p-dimensional multivariate normal with a zero mean vector and precision matrix $\mathbf{\Omega}^{(k)}$, Φ is the standard univariate normal distribution function, and F_j is the marginal distribution of OTU j. The dependency structure induced by this model in condition k is represented by the conditional independence graph

$G^{(k)}$. Following from the theory of Gaussian graphical models (Lauritzen 1996), this is given by the zero-patterns of the precision matrix $\boldsymbol{\Omega}^{(k)}$.

In order to adapt to the richness and heterogeneity of microbiome data, the marginal distribution F_j of OTU j should be linked to external covariates, such as body site and sequencing depth. We formalise this with the use of a parametric marginal model. In particular, as in Vinciotti et al (2022), we consider discrete Weibull regression marginals, i.e. for each $j = 1, \ldots, p$,

$$\begin{aligned} F_j(y_j|\boldsymbol{x}) &= 1 - q_j(\boldsymbol{x})^{(y_j+1)^{b_j(\boldsymbol{x})}} \\ \log\left(\frac{q_j(\boldsymbol{x})}{1-q_j(\boldsymbol{x})}\right) &= \mathbf{x}^t\boldsymbol{\eta}_j, \quad \log\big(b_j(\boldsymbol{x})\big) = \boldsymbol{x}^t\boldsymbol{\gamma}_j \end{aligned} \tag{1}$$

with node covariates $\boldsymbol{x} = (1, x_1, \ldots, x_m)^\top$ and regression coefficients $\boldsymbol{\eta}_j$ and $\boldsymbol{\gamma}_j$ associated with the two parameters defining the discrete Weibull distribution, respectively. These two parameters allow to capture both the case of over and under dispersion relative to Poisson (Peluso et al. 2019). As such, this distribution is particularly suited to our case. On the one hand, the presence of external covariates may generate broad dispersion levels across the covariate levels and the different OTUs. On the other hand, fine tuning of each marginal model across a selection of candidate distributions is time-consuming for a large number of OTUs and/or covariates.

Due to the discreteness of the data, the mapping from the discrete to the latent Gaussian space $z_j = \Phi^{-1}(F_j(y_j))$ of the copula is not unique. Indeed, each observation $(y_j, \boldsymbol{x})$ is associated with an interval in the latent space, given by

$$\mathcal{I}_{F_j}(y_j|\boldsymbol{x}) = \big(\Phi^{-1}\big(F_j(y_j-1|\boldsymbol{x})\big), \Phi^{-1}\big(F_j(y_j|\boldsymbol{x})\big)\big]. \tag{2}$$

As for the joint random graph model, we are particularly interested in modelling the relatedness of the different environments as well as a possible link with external covariates/existing knowledge at the microbial interaction level. To this end, we formalise the model with the following latent probit network model (Hoff et al. 2002)

$$P(G_{j_1,j_2}{}^{(k)} = 1 \mid G^{(-k)}_{j_1,j_2}, \boldsymbol{\Theta}, \boldsymbol{w}) = \Phi\Big(\alpha_k + \boldsymbol{w}_{j_1,j_2}{}^t\boldsymbol{\beta} + \mathbf{c}_k^t \sum_{k' \neq k} \mathbf{c}_{k'} 1_{\{G_{j_1,j_2}{}^{(k')}=1\}}\Big), \tag{3}$$

where $G_{j_1,j_2}{}^{(k)} = 1$, with $j_1, j_2 \in \{1, \ldots, p\}$, $j_1 \neq j_2$, defines an edge between the random variables Y_{j_1} and Y_{j_2} in condition k, $\boldsymbol{w} \in \mathbb{R}^d$ is the vector of edge-specific covariates, $\mathbf{c}_1, \ldots, \mathbf{c}_B \in \mathbb{R}^2$ are the latent space variables for each condition, α_k is the intercept of the model and relates to the overall sparsity level of graph $G^{(k)}$. We denote with $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c})$ the vector of parameters associated to the joint random graph model.

In the next section, we discuss inference of the full set of model parameters from microbiome data, namely $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$ at the marginal level and $G^{(1)}, \ldots, G^{(B)}$, $\boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(B)}$, $\boldsymbol{\Theta}$ at the structural level.

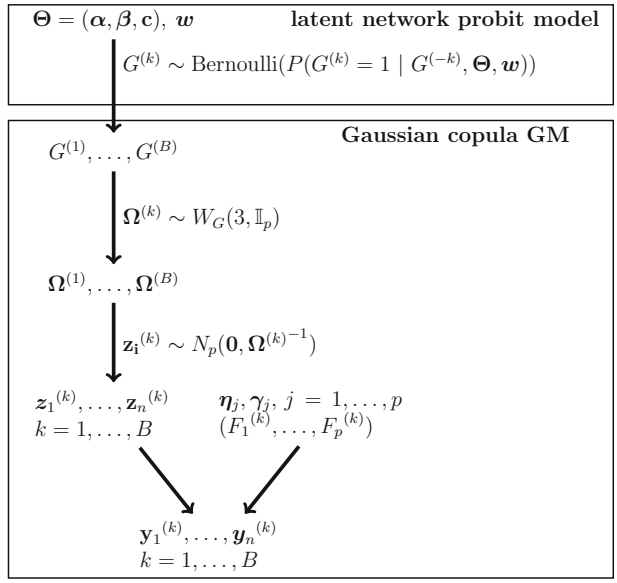

Figure 1. Hierarchical representation of the random graphical model responsible for generating the microbiome data across different related environments. The model combines a latent network probit model with a Gaussian copula graphical model.

2.2. Bayesian Inference

Figure 1 describes the proposed random graphical model and how it generates microbiome data from different, possibly related, environments. In particular, given the parameters $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c})$ that define the latent network model, edges in $G^{(k)}$ are independent, conditional on the remaining graphs, and are, therefore, the result of Bernoulli draws. Given graphs $G^{(k)}$ for each condition $k = 1, \ldots, B$, the data are then generated via a Gaussian copula graphical model (Vinciotti et al 2022), i.e. positive-definite precision matrices are drawn via G-Wishart distributions, and the resulting multivariate normal draws are combined with the parametric marginals to generate count microbiome data.

In order to quantify the full uncertainty in the estimation of the parameters, we opt for a Bayesian inferential procedure. To this end, we consider the following prior distributions: non-informative $N(0, 10)$ priors on each parameter in $\boldsymbol{\Theta}$, weakly informative $N(0, 1)$ priors on each regression coefficient in $(\boldsymbol{\eta}_j, \boldsymbol{\gamma}_j)$, $j = 1, \ldots, p$, non-informative G-Wishart priors for the precision matrix $\boldsymbol{\Omega}^{(k)} \sim W_G(3, \mathbb{I}_p)$ conditional on the graph $G^{(k)}$ (Mohammadi and Wit 2015). We use Markov Chain Monte Carlo (MCMC) sampling scheme for generating samples from the posterior distribution of the parameters, described in Table 1.

Upon convergence, posterior distributions of all parameters are returned. We focus particularly on the parameters $\boldsymbol{\Theta}$, which provide information on the latent process generating the graphs and how related the different environments are at the structural level, and on the graphs $G^{(k)}$, which are associated with posterior edge inclusion probabilities

$$P(G_{j_1 j_2}{}^{(k)} = 1 | \boldsymbol{y}, \boldsymbol{x}, \boldsymbol{w}) = \frac{\sum_{t=1}^{N} 1((j_1, j_2) \in G_t{}^{(k)}) W(\boldsymbol{\Omega}_t{}^{(k)}, \boldsymbol{\Theta})}{\sum_{t=1}^{N} W(\boldsymbol{\Omega}_t{}^{(k)}, \boldsymbol{\Theta})}, \tag{4}$$

Table 1. MCMC scheme of random graphical model inference

1. Metropolis-hastings sampling of the marginal regression parameters $\boldsymbol{\eta}_j, \boldsymbol{\gamma}_j$ (Haselimashhadi et al. 2018)
2. Gibbs sampling of the parameters $\boldsymbol{\Theta}$ from their posterior distribution via a sequence of probit regressions (with offset):
• $\boldsymbol{\alpha}\|\boldsymbol{\beta}, \mathbf{c}, \{G^{(k)}\}_k, \boldsymbol{w}$
• $\boldsymbol{\beta}\|\boldsymbol{\alpha}, \mathbf{c}, \{G^{(k)}\}_k, \boldsymbol{w}$
• $\mathbf{c}_k\|\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c}_{-k}, \{G^{(k)}\}_k, \boldsymbol{w}$
3. Gibbs sampling of $z_{ij}{}^{(k)}\|\boldsymbol{\Omega}^{(k)}, \boldsymbol{z}_{i,-j}{}^{(k)}, \boldsymbol{y}_i{}^{(k)}$ via a truncated normal on $\mathcal{I}_{\hat{F}_j}(y_{ij}\|\boldsymbol{x}_i)$ (Vinciotti et al 2022)
4. Gibbs sampling of $\boldsymbol{\Omega}^{(k)}\|G^{(k)}, \boldsymbol{z}^{(k)}$ via a G-Wishart posterior distribution (Mohammadi and Wit 2015)
5. Continuous time birth-death MCMC sampling of the graph $(G^{(k)})^{\pm e}\|\boldsymbol{\Omega}^{(k)}, \boldsymbol{z}^{(k)}, G^{(k)}, \boldsymbol{\Theta}, \boldsymbol{w}$ generating a new graph from the current graph $G^{(k)}$ with an edge e added, removed or kept (Mohammadi and Wit 2015). The move to a larger dimension (birth) or a smaller one (death) is regulated by an exponential waiting time between two successive events and it is always accepted, making these approaches particularly efficient for high dimensional search spaces, as in this case

where N is the number of MCMC iterations and $W(\boldsymbol{\Omega}_t{}^{(k)}, \boldsymbol{\Theta})$ is the waiting time for graph $G_t{}^{(k)}$ with precision matrix $\boldsymbol{\Omega}_t{}^{(k)}$, that is, the average time that the MCMC sampling has spent visiting the graph $G_t{}^{(k)}$ before jumping to other configurations (Mohammadi and Wit 2015). Posterior distributions on the precision matrices $\boldsymbol{\Omega}^{(k)}$ are also available and can be converted to partial correlations for each edge, via

$$\pi_{j_1 j_2} = -\frac{\omega_{j_1 j_2}}{\sqrt{\omega_{j_1 j_1}\omega_{j_2 j_2}}}, \tag{5}$$

with $\omega_{j_1 j_2}$ denoting the (j_1, j_2) entry of a precision matrix $\boldsymbol{\Omega}$. These values give information also about the sign of the dependencies in each environment. In a similar vein, posterior distributions of any network statistic of interest can be derived from the MCMC chain of graphs that is returned.

3. RESULTS

In this section, we present a simulation study to show the performance of the random graphical model, as well as an implementation on data from the Human Microbiome Project.

3.1. SIMULATION STUDY

In order to clarify the data generating process behind the proposed random graphical model described in Fig. 1, and to assess its performance in inferring parameters from data, we simulate $n = 346$ observations on $p = 87$ variables for $B = 13$ environments, with the sample size and dimensions matching those of the real data. For the simulation, we construct a latent space $\boldsymbol{\Theta}$ with the following components: $\boldsymbol{\alpha}$ parameters drawn from a $N(-2, 1)$ distribution, i.e. a low edge probability in Eq. (3), leading to a high level of network sparsity; one edge covariate W from a $U(-0.5, 0.5)$ distribution with an associated

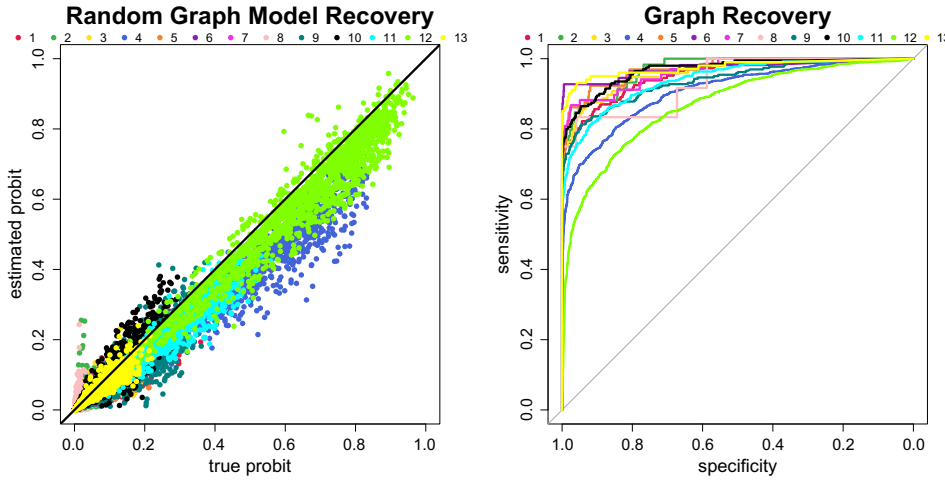

Figure 2. Results from the simulation study, evaluated on the last 2500 MCMC iterations. Left: True probit probabilities from Eq. (3) versus those calculated using the mean posterior estimates of the $\boldsymbol{\alpha}$, β and $\mathbf{c}$ parameters. Right: Receiver operating characteristic curves of the recovered graphs against the true ones for each environment, across a sequence of thresholds on the posterior edge probabilities from Eq. (4). The 13 colours distinguish the 13 environments, respectively.

parameter $\beta = 2.5$; latent vectors $\mathbf{c} \in \mathbb{R}^2$ with each component drawn from a $N(0, 0.3)$ distribution. Given $\boldsymbol{\Theta}$, we first generate $k = 1, \ldots, 13$ graphs via Bernoulli draws for each edge conditional on the others. We iterate the sampling in order to obtain the joint distribution of the graphs and check that the sampling successfully converged to this joint distribution by monitoring the density of the graphs being sampled. Given the sampled graphs $\{G^{(k)}\}$ at the final iteration, we sample their associated precision matrices $\{\Omega^{(k)}\}$ from a $W_G(3, \mathbb{I}_p)$ distribution. We finally obtain the observed data for each environment k from a multivariate Gaussian draw $N(0, \Omega^{(k)})$. We omit here the case of discrete marginals and concentrate on the inference of the latent space and recovery of the networks. The data constructed as described above can be retrieved by running the function `sim.rgm` of the `rgm` package accompanying this paper, using the default values for the inputs.

Figure 2 reports the results after 10,000 MCMC iterations, obtained by running the function `rgm` with prior distributions as described in Sect. 2.2. We retain the last 25% of the iterations for the calculation of posterior edge distributions from Eq. (4) and posterior distributions of the parameters $\boldsymbol{\Theta}$ of the random graph model. The first plot shows a good recovery of the latent network space $\boldsymbol{\Theta}$, by comparing the true probit probabilities from Eq. (3) with those obtained using the mean posterior estimates of the $\boldsymbol{\alpha}$, β and $\mathbf{c}$ parameters. The second plot shows an accurate reconstruction of the networks $G^{(k)}$, by comparing the recovered graphs with the true graphs, for each environment and across a sequence of thresholds on the posterior edge probabilities. The average area under the receiver operating characteristic curves is 0.95, across the 13 environments.

Beyond the specific example used in the simulation, and following related studies in the literature (Mohammadi et al. 2017; Vinciotti et al 2022), we expect the performance of the method to improve, in terms of parameter estimation and graph recovery, the lower p is

and the sparser and more structured the graphs are. Moreover, as we will show in the real-data application, we expect the joint model across conditions to lead to better performance compared to individual analyses per condition in the presence of similarities between graphs, as these induce a sharing of information across environments that is exploited only by the joint analysis.

3.2. JOINT INFERENCE OF MICROBIOTA SYSTEMS ACROSS BODY SITES

Microbiome data We use the microbiome data from a study conducted as part of the Human Microbiome Project (Consortium 2012), collecting microbial abundances at the level of Operational Taxonomic Units (OTUs) from 16S variable region V3-5 data of healthy individuals. The data are available in the `rMAGMA` package in R (Cougoul et al. 2019). After filtering out samples with less than 500 reads, we focus on the 13 body sites with the largest sample size, namely "Anterior_nares" (later referred to as `nose`), "Attached_Keratinized_gingiva" (`ker-gingiva`), "Buccal_mucosa" (`cheek`), "Hard_palate" (`palate`), "L_Retroauricular_crease" (`L-ear`), "Palatine_Tonsils" (`tonsils`), "R_Retroauricular_crease" (`R-ear`), "Saliva" (`saliva`), "Stool" (`stool`), "Subgingival_plaque" (`sub-gingiva`), "Tongue_dorsum" (`tongue`), "Throat" (`throat`), and "Supragingival_plaque" (`sup-gingiva`). On average, there are 346 samples for each body site. We finally restrict our attention to the 87 OTUs which have more than two distinct observed values in each of these environments. The microbal communities are the interacting units, and, therefore, constitute the nodes of the network.
Marginal models A number of covariates are considered at the marginal level of each OTU. It is well-known that the library size affects the reads of a particular OTU. The larger the library size, the larger the number of reads. The library size is estimated by the geometric mean of pairwise ratios of OTU abundances of that sample with respect to all other samples (function `GMPR` in `rMAGMA`). Furthermore, the abundance of each OTU varies at the different body sites. Therefore, we include the library size, dummy variables for each body site, and interactions between body sites and library size for each sample as covariates for the discrete Weibull marginal distribution for each OTU [Eq. (1)]. This results in 26 parameters per OTU. We also consider a more complex model with the inclusion of an additional zero-inflated parameter for each OTU and each environment, on which we place a Beta(1,1) prior distribution.

We fit discrete Weibull parametric marginals for each OTU via 50,000 MCMC iterations (function `bdw.reg` in the `BDgraph` package (Mohammadi and Wit 2019)). We select between a discrete Weibull and a zero-inflated discrete Weibull model for each marginal via a BIC criterion. As in Vinciotti et al (2022), we find that only a small percentage of OTUs (12.5%) necessitates the more complex zero-inflated model. In principle, further tuning of the marginal models could be conducted by considering also other distributions, such as the negative Binomial or hurdle distributions. To this end, Fig. 3 shows how the performance is similar between a discrete Weibull and a negative Binomial distribution, with a small number of OTUs being significantly better fitted by discrete Weibull. Although selecting the best distribution for each OTU is possible, and other models for count data are also available, this is time-consuming for a large number of variables and may not have a big

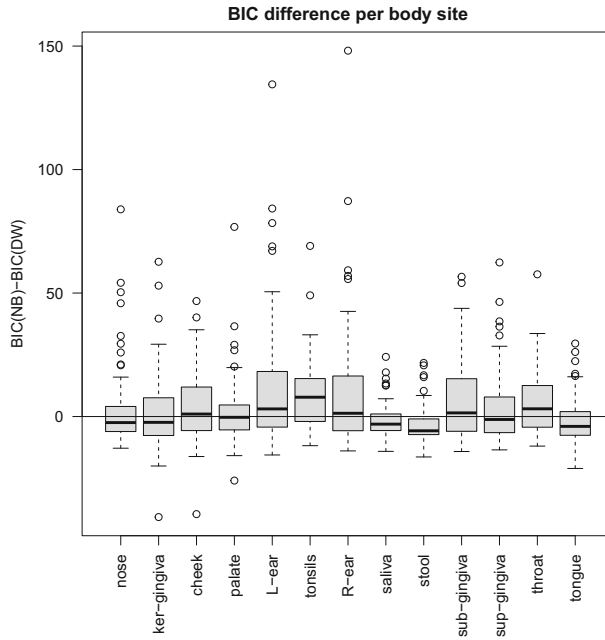

Figure 3. Difference of Bayesian Information Criterion (BIC) between the discrete Weibull and the negative Binomial model for each OTU and condition. In each case, the zero-inflated model is considered if it leads to a lower BIC.

impact on the structural learning procedure. Indeed, related studies have shown that the recovery of the network dependences is rather robust to miss-specifications of the marginal distributions (Cougoul et al. 2019; Vinciotti et al 2022).

Random graph model Covariates are considered also at the random graph level. In particular, the random graph model in Eq. (3) is defined by a sparsity parameter α_k and a latent location $\boldsymbol{c}_k \in \mathbb{R}^2$, for each body site k, as well as a vector $\boldsymbol{\beta}$ of regression coefficients associated with six binary variables ($\boldsymbol{w}$) that encode whether a pair of OTUs belong to the same taxonomy level. In particular, we consider the six taxonomy levels given by the bacterial `phylum`, `class`, `order`, `family`, `genus` and `species`.

Structural learning As typical of inferential approaches for Gaussian copula graphical models, the fitting of marginals is performed first, followed by a calculation of the intervals from Eq. (2) using posterior mean estimates of the parameters (evaluated on the last 25% of the iterations). These intervals are then used for the subsequent learning of the structural dependencies, by iterating steps 2–5 of the procedure described in Sect. 2.2 (function `rgm` in the `rgm` package that accompanies this paper). Given the huge space of graphs, we let the Bayesian structural learning procedure run for 3 million MCMC iterations. All subsequent results are evaluated on the last 7500 iterations.

Interpretation of results The most immediate output of the analysis are the 13 networks that are inferred for each environment. Figure 4 summarises these networks by the posterior edge probabilities, calculated from Eq. (4). It is clear that the networks tend to be sparse, and vary, to some extent, between the conditions. The reasons for this environmental network

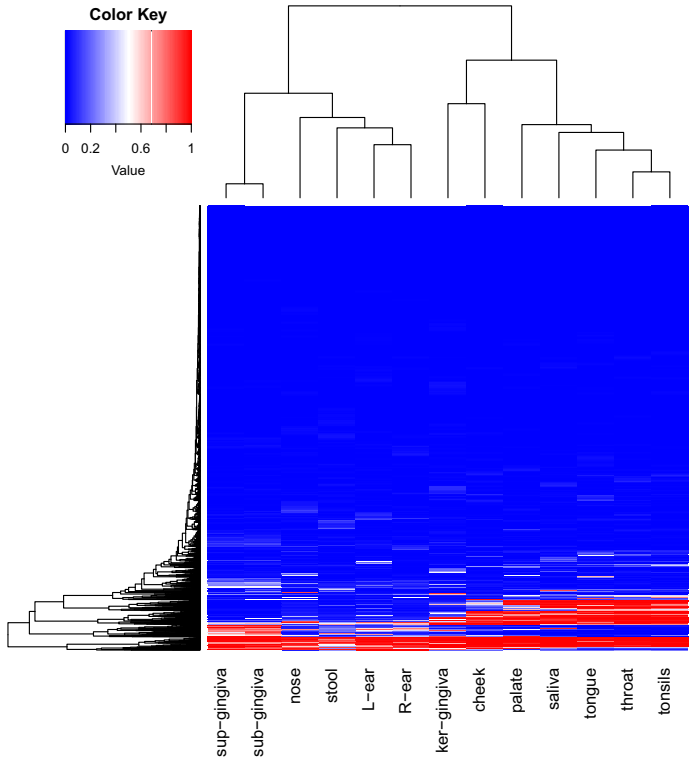

Figure 4. Random graphical model inferred from microbiome data: Posterior edge probabilities for each edge and each environment, rearranged via row and column clustering.

variation can be found from the random graph generative process, described by Eq. (3). Figure 4 shows a high level of sparsity across all networks (mean posterior edge probabilities equal to 6.8%, on average across the 13 environments). This is captured by low intercept values α of the fitted random graph model (mean posterior estimate -2.7, on average across the 13 environments).

Figure 4 shows high structural similarity between some environments, with significant sharing of edges and non-edges with high probability among similar environments. This is explained by the latent locations of the body sites in the random graph model, shown in Fig. 5. For example, `sup-gingiva` and `sub-gingiva` are highly related environments, and similarly `throat` and `tonsils`. Indeed, in both cases, the two associated latent location vectors $\boldsymbol{c}$ have a large inner product, as they are close to each other in the space and far from zero. The indicator function in Eq. (3) further encourages sharing of edges between these networks. Indeed, 93% of the edges with posterior edge probability greater than 0.5 are in common between the `sup-gingiva` and `sub-gingiva` networks, and 95% between the `throat` and `tonsils` networks. Looking at the posterior mean of partial correlations, calculated from the precision matrices via Eq. (5), we find an agreement also on the sign of the dependency, with a correlation of 0.90 between `sup-gingiva` and `sub-gingiva` partial correlation values for each edge, and 0.93 between `throat` and `tonsils`. Finally, as the two pairs of networks are almost orthogonal to each other in the latent space, we expect

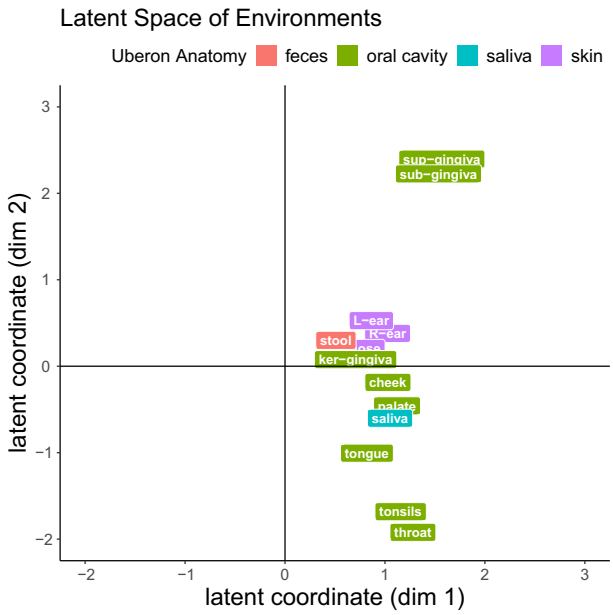

Figure 5. Random graphical model inferred from microbiome data: mean posterior locations of the body sites ($\boldsymbol{c}$) in a 2D latent space. Colours refer to the Uberon anatomy classification of the body sites.

little structural sharing between the `sup-gingiva`/`sub-gingiva` networks and the `throat`/`tonsils` networks. Indeed, they have the lowest agreement of high-probability edges across all pairs, with an average sharing of only 21.3%.

The similarities between the environments detected by the proposed method are partly supported by the Uberon anatomy classification of body sites, particularly when it comes to the three skin-related body sites. These are located close to each other in the latent space of Fig. 5 and have on average 63% of high-probability edges in common (Fig. 4). On the other hand, the oral cavity-related body sites appear to be further split into two groups. This is in line with the analysis of Segata et al. (2012) that found four groups of body sites based on similar community compositions, namely: `cheek`, `ker-gingiva`, `palate`; `saliva`, `tongue`, `tonsils`, `throat`; `sub-gingiva` and `sup-gingiva`; and `stool`. These groups are also clearly evident in Fig. 5.

Finally, the results show how the taxonomical relatedness of the microbes encourages the presence of a link between them. Indeed, Fig. 6 shows how the probability of two OTUs connecting, in any environment, is positively associated with their belonging to some of the taxonomy levels considered, in particular to the `species`, `genus` and `class` taxonomies.

Comparison with other methods Figure 7 shows that the random graphical model leads also to a more stable recovery of the individual networks, compared to estimating individual networks. Indeed, the figure shows that the variances of the posterior edge probabilities are smaller for the proposed `rgm` approach than when fitting individual Gaussian copula graphical models for each environment separately. For the latter, we used the approach of Vinciotti et al (2022). Implemented in the function `bdgraph.dw` in the `BDgraph` R package, we considered the same parametric marginals as those considered in this paper but a more traditional Erdös-Rényi random graph prior for each environment. To facilitate comparison,

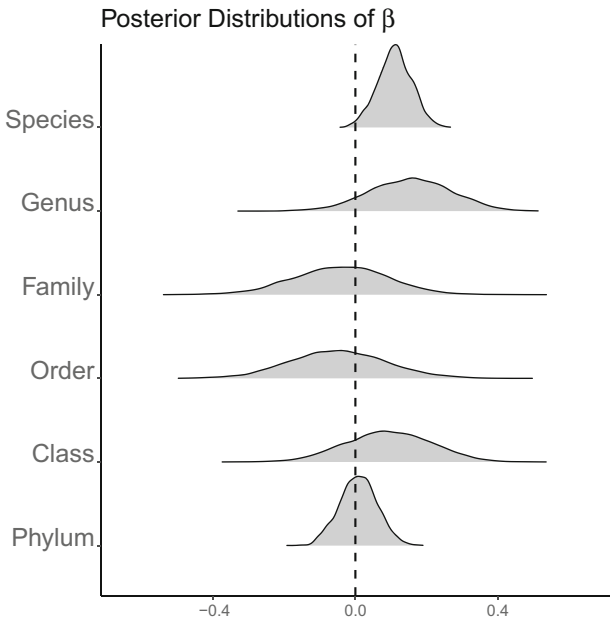

Figure 6. Random graphical model inferred from microbiome data: estimation of $\boldsymbol{\beta}$ parameters associated with six dummy variables indicating if a pair of nodes forming an edge belongs to the same taxonomy level .

the prior edge probability of the Erdös-Rényi prior is set to match the sparsity level of the networks recovered by the `rgm` analysis. Figure 7 shows how the posterior probabilities detected by `rgm` are more concentrated on either 0 or 1 than with the alternative approach. This means that the joint analysis proposed in this study leads to a more confident detection of structural dependencies, as it induces a sharing of information across environments, compared to separate analyses for each environment.

4. DISCUSSION AND CONCLUSION

In this paper, we have proposed a novel approach for the inference of microbiotic systems from multivariate measurements of microbial abundances across different, but related, environments. We have shown how the combination of graphical models for each environment with a joint random graph model describing the distribution of graphs across environments allows to learn about the individual microbiota systems as well as their structural similarities. In order to further adapt to the richness and complexity of microbiome data, the proposed approach allows for the inclusion of external covariates that may have an association with marginal microbial abundances or their interactions.

We have applied the methodology to the study of the human microbiome and have shown how the method is able to recover the microbiotic system between 87 microbes that are specific to each of the 13 body sites considered, as well as to capture the biological relatedness between the 13 microbiotic systems. Although, for this application, the number of samples for each body site was larger than the number of microbes ($p = 87$ and $n = 346$ on average across body sites), the Bayesian structural learning approach that is considered (Mohammadi and Wit 2015) can be used also when the number of observations per condition

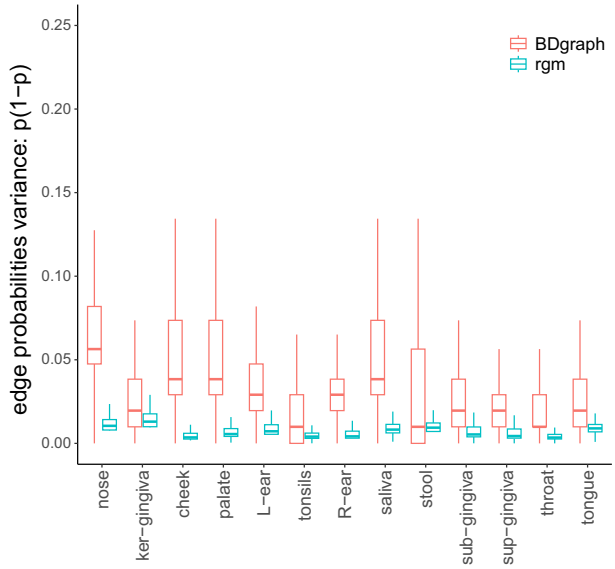

Figure 7. Comparison between the joint `rgm` and individual Gaussian copula graphical models for each environment (Vinciotti et al 2022). For each method and each environment, the plot shows the boxplot of the variance of the posterior edge probabilities .

is smaller than p, which is common for genomic data. In fact, as we show also in this paper, the joint modelling of the graphs across environments induces a sharing of information across environments which may be particularly beneficial in these cases.

Beyond the analysis presented in this paper, the method can be used more broadly on microbiome data measured across different conditions, where there is interest in learning structural dependencies within each environment and their similarities between environments. At this more general level, the proposed methodology share some similarities with graphical modelling approaches from data across multiple conditions, such as those described by Ni et al. (2022). More dedicated random graph models may be needed depending on the context, e.g. for the case of microbiome data measured over time with dependencies that change over time.

Software availability R package available at https://github.com/franciscorichter/rgm.

Funding Open access funding provided by Università degi Studi di Trento within the CRUI-CARE Agreement.

Declarations

Conflict of interest All authors declare that they have no conflict of interest.

[Received September 2023. Revised December 2023. Accepted June 2024. Published Online June 2024.]

REFERENCES

Consortium HMP (2012) A framework for human microbiome research. Nature 486(7402):215–221

Cougoul A, Bailly X, Wit E (2019) MAGMA: inference of sparse microbial association networks. bioRxiv: 538579

Friedman J, Alm E (2012) Inferring correlation networks from genomic survey data. PLoS Comput Biol 8(9):e1002687

Haselimashhadi H, Vinciotti V, Yu K (2018) A novel Bayesian regression model for counts with an application to health data. J Appl Stat 45(6):1085–1105

Hoff P, Raftery A, Handcock M (2002) Latent space approaches to social network analysis. J Am Stat Assoc 97(460):1090–1098

Kim J, Cho Y, Lim S, Seo M, Sohn J, Kim B, Rho M, Pai H (2023) Comparative analyses of the faecal resistome against β-lactam and quinolone antibiotics in humans and livestock using metagenomic sequencing. Sci Rep 13:20993

Kost C, Patil K, Friedman J, Garcia S, Ralser M (2023) Metabolic exchanges are ubiquitous in natural microbial communities. Nat Microbiol 8:2244–2252

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

Lauritzen S (1996) Graphical models. Clarendon Press

Le Chatelier E, Nielsen T, Qin J, Prifti E, Hildebrand F et al (2013) Richness of human gut microbiome correlates with metabolic markers. Nature 500(7464):541–546

Lee K, Thomas A, Bolte A, Björk J, Kist de Ruijter L et al (2022) Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. Nat Med 28:535–544

Lei B, Xu Y, Lei Y, Li C, Zhou P, Wang L, Yang Q, Li X, Li F, Liu C, Cui C, Chen T, Ni W, Hu S (2022) CRAMdb: a comprehensive database for composition and roles of microbiome in animals. Nucl Acids Res 51(D1):D700–D707

Mohammadi R, Wit E (2015) Bayesian structure learning in sparse Gaussian graphical models. Bayesian Anal 10(1):109–138

Mohammadi R, Wit E (2019) BDgraph: an R package for Bayesian structure learning in graphical models. J Stat Softw 89(3):1–30

Mohammadi R, Abegaz F, van den Heuvel E, Wit E (2017) Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. J R Stat Soc: Ser C (Appl Stat) 66(3):629–645

Mungall C, Torniai C, Gkoutos G, Lewis S, Haendel M (2012) Uberon, an integrative multi-species anatomy ontology. Genome Biol 13(1):R5

Ni Y, Baladandayuthapani V, Vannucci M, Stingo F (2022) Bayesian graphical models for modern biological applications. Stat Methods Appl 31(2):197–225

Pedersen H, Gudmundsdottir V, Nielsen H, Hyotylainen T, Nielsen T et al (2016) Human gut microbes impact host serum metabolome and insulin sensitivity. Nature 535(7612):376–381

Peluso A, Vinciotti V, Yu K (2019) Discrete Weibull generalized additive model: an application to count fertility data. J R Stat Soc Ser C 68(3):565–583

Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T et al (2010) A human gut microbial gene catalogue established by metagenomic sequencing. Nature 464(7285):59–65

Segata N, Haake S, Mannon P, Lemon K, Waldron L et al (2012) Composition of the adult digestive tract bacterial microbiome based on seven mouth surfaces, tonsils, throat and stool samples. Genome Biol 13:R42

Sharon I, Quijada N, Pasolli E, Fabbrini M, Vitali F et al (2022) The core human microbiome: Does it exist and how can we find it? A critical review of the concept. Nutrients 14(14):2872

Vinciotti V, Behrouzi P, Mohammadi R (2022) Bayesian structural learning with parametric marginals for count data: An application to microbiota systems. arXiv:2203.10118