

# Bayesian Structural Learning with Parametric Marginals for Count Data: An Application to Microbiota Systems

**Veronica Vinciotti**<sup>†</sup>

*Department of Mathematics  
University of Trento  
Italy*

VERONICA.VINCIOTTI@UNITN.IT

**Pariya Behrouzi**<sup>†</sup>

*Applied Mathematics and Statistics Group  
Wageningen University and Research  
Netherlands*

PARIYA.BEHROUZI@WUR.NL

**Reza Mohammadi**

*Department of Business Analytics  
Amsterdam Business School  
University of Amsterdam, Netherlands*

A.MOHAMMADI@UVA.NL

**Editor:** Samuel Kaski

## Abstract

High dimensional and heterogeneous count data are collected in various applied fields. In this paper, we look closely at high-resolution sequencing data on the microbiome, which have enabled researchers to study the genomes of entire microbial communities. Revealing the underlying interactions between these communities is of vital importance to learn how microbes influence human health. To perform structural learning from multivariate count data such as these, we develop a novel Gaussian copula graphical model with two key elements. Firstly, we employ parametric regression to characterize the marginal distributions. This step is crucial for accommodating the impact of external covariates. Neglecting this adjustment could potentially introduce distortions in the inference of the underlying network of dependences. Secondly, we advance a Bayesian structure learning framework, based on a computationally efficient search algorithm that is suited to high dimensionality. The approach returns simultaneous inference of the marginal effects and of the dependence structure, including graph uncertainty estimates. A simulation study and a real data analysis of microbiome data highlight the applicability of the proposed approach at inferring networks from multivariate count data in general, and its relevance to microbiome analyses in particular. The proposed method is implemented in the R package **BDgraph**.

**Keywords:** Copula graphical models, Discrete Weibull, Link prediction, Structure learning, Microbiome

## 1. Introduction

Graphical modelling approaches allow to learn statistical dependences from multivariate data. Among these, Gaussian graphical models are by far the most popular, thanks also to their efficient implementations for high dimensional problems (Friedman et al., 2008;

---

0. Authors <sup>†</sup> contributed equally to this work.

Mohammadi et al., 2023). In many applied fields, however, data are far from Gaussian. In this paper, we consider the case of count data, such as the high-resolution sequencing data collected routinely in genomic studies. It is not uncommon for these data to feature marginal distributions that are skewed and with a large mass at zero. For this reason, transformations, such as the logarithm or the centered log ratio, are typically applied to genomic data, followed by Gaussian graphical modelling approaches on the transformed data. This is for example the case of the two most used methods for microbiome data, SparCC (Friedman and Alm, 2012) and SPIEC-EASI (Kurtz et al., 2015). These transformations require a pseudo-count adjustment to be able to handle zeros and may therefore impact also the network inference conducted downstream.

In the literature, extensions of Gaussian graphical models to non-Gaussian data can take different forms but there is generally little research for the case of unbounded count data, such as the genomic data that we discuss above. Roy and Dunson (2020) have recently proposed a pairwise Markov random field model with flexible node potentials, while Cougoul et al. (2019) have proposed a Gaussian copula graphical model to couple the contribution from the marginal distributions with that of the underlying dependence structure. Considering microbiota systems as the specific application, they propose zero-inflated negative Binomial marginals. Our work is linked to this second paper. On the one hand the use of a Gaussian copula facilitates the integration of novel approaches with existing ones that rely on Gaussianity, without the need for ad-hoc transformations. On the other hand, the use of parametric marginal distributions, rather than the non-parametric empirical distributions as in the popular non-paranormal approach (Liu et al., 2009), facilitates the inclusion in the model of additional covariates, which are, for example, typically available in genomic studies but often ignored due to methodological restrictions. These marginal effects, if left unaccounted for, could distort the inference of the underlying network of dependences.

While there are no specific constraints in the choice of the parametric marginal model, in this paper, we advocate the use of discrete Weibull regression for linking the marginal distributions to external covariates (Klakattawi et al., 2018; Haselimashhadi et al., 2018; Peluso et al., 2019). The simplicity of this distribution (a two-parameter distribution), combined with the fact that the two parameters can jointly capture broad levels of dispersion (from under to over), makes it quite an appealing candidate for multivariate count data with a high number of random variables and/or external covariates, such as the microbiota data that we consider in the real data example. This is because, firstly, for a large number of count variables, one wants to avoid tuning the type of distribution for each variable, and, secondly, for a large number of external covariates, a global requirement of over dispersion at all levels of the covariates could prove too restrictive. Finally, an important feature of the discrete Weibull distribution in the context of Gaussian copula graphical models, is the fact that it is generated as a discretized form of a continuous Weibull distribution (see Figure 1). This creates a latent non-Gaussian space in the vicinity of the data, with a one-to-one mapping with the latent Gaussian data, where the conditional independence graph resides.

A fundamental problem of copula graphical models for discrete data, bounded or unbounded, is the fact that the marginal distributions are not strictly monotonic. In this setting, while the existence of a copula can still be guaranteed by Sklar’s theorem (Sklar, 1959), its uniqueness can not. In fact, the class of copulas compatible with a given discrete dataset can be quite large, leading to potential biases in the inferential procedure

(Genest and Nešlehová, 2007). On the one hand, this problem is alleviated by the presence of covariate dependent marginals, particularly when the covariates are continuous and the underlying network does not depend on the covariates (Yang et al., 2020). This can be seen as a second advantage of incorporating covariates in the marginal models, when the objective is to perform structural learning for count data. On the other hand, more advanced inferential procedures are required, that account for the fact that each observed count is associated with an interval in the latent Gaussian space. This relies on the ideas of extended rank likelihood (Hoff, 2007) and has been used also in the context of Gaussian copula graphical models, both in a frequentist setting (Behrouzi and Wit, 2019) and in a Bayesian setting (Dobra and Lenkoski, 2011; Dobra and Mohammadi, 2018; Mohammadi et al., 2017; Murray et al., 2013). While extended rank likelihood has been developed for ordinal (bounded) data, in this paper we develop these approaches for Gaussian copula graphical models on unbounded count data with parametric marginals. The use of these approaches avoids the need for ad-hoc data transformation procedures that condense each interval into one point, with choices such as the right-most point of the interval (essentially using the non-paranormal approach of Liu et al. (2009, 2012) on count data) or the point corresponding to the median of the distribution function at the two extremes of the interval (Cougoul et al., 2019). These choices, while efficient, may not work well with skewed distributions, or generally distribution functions that are highly stepwise.

Finally, we conduct inference in a Bayesian framework, leading to a novel Bayesian structure learning procedure in the context of Gaussian copula graphical models with parametric marginals, extending the efficient computational approaches that have been recently proposed for this class of models (Mohammadi and Wit, 2015; Mohammadi et al., 2023), and providing an alternative to frequentist approaches (Cougoul et al., 2019). Appropriate choices of a prior distribution on the graphs can be made to encourage sparsity. Importantly, uncertainty on the graph learning is fully quantified by the procedure and can be summarized in various ways, such as by calculating posterior probabilities for each edge via Bayesian averaging. This plays a crucial role, particularly in high dimensional settings, where model selection methods for regularized approaches do not work well and where there is typically a large uncertainty around the optimal graph.

In conclusion, this paper presents a novel methodology for structural learning from high dimensional heterogeneous count data. Section 2 will describe the details of the methodology proposed, whose implementation has been included in the R package `BDgraph` (Mohammadi and Wit, 2019). A simulation study in Section 3 and a real data analysis of microbiome data from the Human Microbiome Project (HMP Consortium, 2012) in Section 4 will show the usefulness of the proposed approach at inferring networks from high dimensional count data in general, and in the context of microbiota data analyses in particular. Finally, Section 5 will draw some conclusions and point to future research directions.

## 2. Methods

In this section, we present the technical details of the proposed method, starting with the definition of a Gaussian copula graphical model and of the discrete Weibull (DW) regression used for the marginal components, followed by the Bayesian inferential procedure.

## 2.1 Gaussian copula graphical model with DW marginals

Let  $\mathbf{Y} = (Y_1, \dots, Y_p)$  be a vector of count variables. In the case of microbiota systems that we consider in Section 4, these are abundances of the individual microbes or, more commonly, of the Operating Taxonomic Units (OTUs) into which they are clustered, e.g., bacterial species. Let  $F_j(\cdot)$ ,  $j = 1, \dots, p$ , be the cumulative distribution functions associated to the  $p$  variables, respectively. In a copula graphical model, the joint distribution of the variables is described via a copula function  $C(\cdot)$  that couples the marginal distributions  $F_j(\cdot)$  into their joint dependence. Formally,

$$P(Y_1 \leq y_1, \dots, Y_p \leq y_p) = C(F_1(y_1), \dots, F_p(y_p) \mid \Theta),$$

where  $\Theta$  are the parameters describing the copula function  $C(\cdot)$ . In the case of a Gaussian copula (Hoff, 2007; Mohammadi et al., 2017)

$$P(Y_1 \leq y_1, \dots, Y_p \leq y_p) = \Phi_p(\Phi^{-1}(F_1(y_1)), \dots, \Phi^{-1}(F_p(y_p)) \mid \mathbf{R}),$$

where  $\Phi_p(\cdot)$  is the cumulative distribution function of a  $p$ -dimensional multivariate normal with a zero mean vector and correlation matrix  $\mathbf{R}$ , while  $\Phi(\cdot)$  is the standard univariate normal distribution function.

The dependence structure is captured by the inverse of the correlation matrix  $\mathbf{K} = \mathbf{R}^{-1}$ , typically called the precision or concentration matrix. In particular, the zero patterns in this matrix define the conditional independence graph in the latent Gaussian space, following from the theory of Gaussian graphical models (Lauritzen, 1996). In general, the Markov properties of the discrete variables  $\mathbf{Y}$  cannot be recovered fully from the Markov properties of the latent Gaussian variables. Indeed, discrete data are characterized by higher-order dependences than the second-order dependences that characterize Gaussian variables. Abegaz and Wit (2015) show formally how the closer the discrete marginal distributions are to continuous distributions, the better the recovery of the conditional independence structure of  $\mathbf{Y}$  from the Gaussian copula. In the presence of highly discrete data, additional dependences among the discrete data that are not modelled by the Gaussian copula can be considered as being less relevant, since they emerge from the marginals (Dobra and Lenkoski, 2011).

In the context of copula graphical models, the marginal distributions  $F_j(\cdot)$  are typically considered as nuisance parameters and estimated by their empirical counterpart. However, in real-world applications, such as in genomic studies, external covariates are often available and there is an interest in estimating their effect on the outcome while accounting for the multivariate nature of the data. In this paper, we argue how, accounting for external covariates at the marginal level is important also for structural learning. Indeed, when the dependence structure does not vary with the covariates, adjusting for marginal effects has the two benefits of widening the range of the marginal distributions at each discrete point and of correcting for the bias in the estimation of multivariate dependences induced by the marginal effects, respectively.

In this paper, we propose to model the marginal components, and their link with covariates, via a discrete Weibull regression (Peluso et al., 2019), although other count distributions can be used at this stage. Formally, let  $\mathbf{X} = (1, X_1, \dots, X_d)^t$  be a vector of covariates.

Then, the conditional distribution of  $Y_j$  given  $\mathbf{X}$  is modelled by:

$$F_j(y_j|\mathbf{X} = \mathbf{x}) = 1 - q_j(\mathbf{x})^{(y_j+1)^{\beta_j(\mathbf{x})}}, \quad y_j = 0, 1, \dots, \quad (1)$$

where the function  $q_j(\cdot)$ , corresponding to the parameter  $q$  of the distribution, takes values between 0 and 1, while  $\beta_j(\cdot)$  is associated to the parameter  $\beta$  and takes values in the positive real line. We link the parameters to the external covariates using the logit and the log links, respectively, that is

$$q_j(\mathbf{x}) = \frac{\exp(\mathbf{x}^t \boldsymbol{\theta}_j)}{1 + \exp(\mathbf{x}^t \boldsymbol{\theta}_j)}, \quad \beta_j(\mathbf{x}) = \exp(\mathbf{x}^t \boldsymbol{\gamma}_j), \quad (2)$$

with  $\boldsymbol{\theta}_j$  and  $\boldsymbol{\gamma}_j$  denoting the regression coefficients associated to the  $Y_j$  marginal component of the model. For other choices of link functions, see Haselimashhadi et al. (2018). The simplest case of only the intercept in each model corresponds to the case of no external covariates, i.e., simply discrete Weibull marginal distributions. In the real data analysis, we also consider a model with an additional zero inflation component  $\pi_j(\mathbf{x})$ . This is common in the microbiome literature due to the sparsity of the data (Cougoul et al., 2019), although we find that this zero-inflated model is rarely selected against the simpler model.

A few properties of a discrete Weibull distribution make it an ideal candidate for modelling high dimensional count data. In particular:

1.  $F_j(0|\mathbf{X} = \mathbf{x}) = P(Y_j = 0|\mathbf{X} = \mathbf{x}) = 1 - q_j(\mathbf{x})$ , thus the parameter  $q$  models directly the proportion of zeros in the data and the effect of covariates on this. This may be useful for datasets with a large percentage of zeros. In the case of a zero-inflated model,  $F_j(0|\mathbf{X} = \mathbf{x}) = \pi_j(\mathbf{x}) + (1 - \pi_j(\mathbf{x}))(1 - q_j(\mathbf{x}))$ , with  $\pi_j(\mathbf{x})$  accounting for the excess of zeros not contemplated by the discrete Weibull distribution.
2. The two parameters of the distribution are sufficient to capture both under and over dispersion levels, while still being a parsimonious choice (e.g., same number of parameters as the commonly used negative Binomial distribution). This has been shown to be useful on real data analyses of count data, particularly in the case of under dispersion (Peluso et al., 2019). Although microbiome data are typically highly over dispersed, the presence of a large number of external covariates could make a global requirement of over dispersion at all levels of the covariates  $\mathbf{x}$  too restrictive. Moreover, capturing both over and under dispersion is appealing when modelling any generic high dimensional multivariate count data, as it avoids the fine tuning of the most appropriate marginal distribution for each variable.
3. The quantiles of the distribution have a closed-form expression, with the  $\tau$  quantile, for  $\tau \geq 1 - q$ , given by Peluso et al. (2019),

$$\mu(\tau) = \left\lceil \left( \frac{\log(1 - \tau)}{\log(q)} \right)^{1/\beta} - 1 \right\rceil,$$

where  $\lceil \cdot \rceil$  denotes the ceiling function. This means that a re-parametrization based on the median is also possible, e.g., when quantification of the covariate effects is of primary interest (Burger et al., 2020).

- The distribution is developed as a discretized form of the continuous Weibull distribution (Chakraborty, 2015). Namely, by defining the cumulative distribution function (cdf) of a continuous Weibull distribution by

$$F_{CW}(y; q, \beta) = 1 - \exp \left[ - \left( \frac{y}{(-\log q)^{-\frac{1}{\beta}}} \right)^\beta \right], \quad y \geq 0,$$

one can easily show that the probability mass function of the discrete Weibull distribution, associated to the cdf in Equation 1, is given by

$$f(y; q, \beta) = q^{y^\beta} - q^{(y+1)^\beta} = F_{CW}(y+1) - F_{CW}(y) = \int_y^{y+1} f_{CW}(t) dt \quad y = 0, 1, 2, \dots$$

This creates a one-to-one connection between the latent continuous Weibull space, with the same parameters as the discrete Weibull distribution, and the Gaussian space, as depicted schematically in Figure 1.

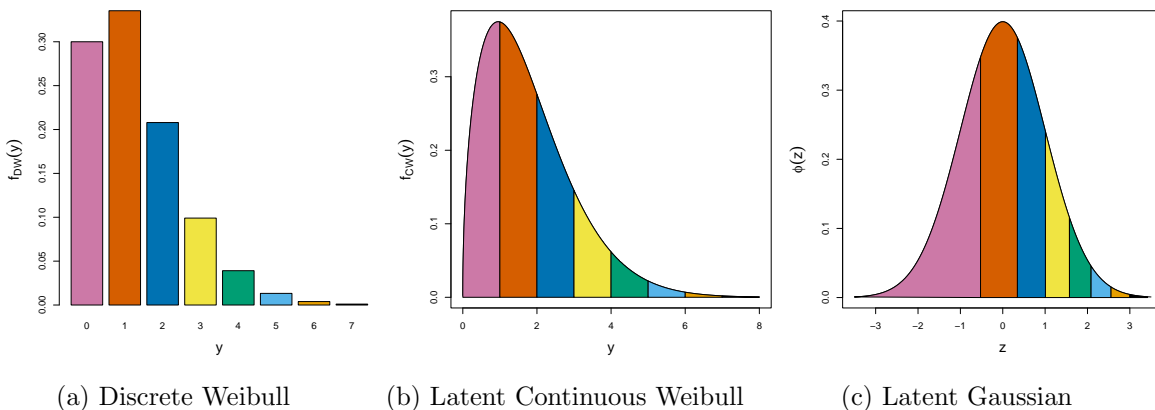


Figure 1: Schematic connection between (a) the discrete Weibull probability mass function  $f(y; q = 0.7, \beta = 1.5)$ , (b) the underlying continuous Weibull density  $f_{CW}(y; q, \beta)$  and (c) the latent Gaussian  $z = \Phi^{-1}(F_{CW}(y; q, \beta))$ . Each colour relates to the probability associated to the corresponding value (here microbial abundance) of the discrete random variable.

As it is clear also from the figure, each discrete observation is linked to an interval in the continuous space. This is the case for copula models on discrete data in general and will require special attention when it comes to inference, as we will discuss more in details in the next section.

## 2.2 Bayesian inference for a DW graphical model

Inference for copula graphical models involves estimation of the marginals and of the network component. A copula formulation enables us to learn the marginals separately from the dependence structure of the  $p$  random variables.

We first concentrate on the marginal components, that is the estimation of the regression coefficients  $\theta_j$  and  $\gamma_j$ ,  $j = 1, \dots, p$ . Given  $n$  observations on component  $Y_j$ , denoted with the vector  $\mathbf{y}_j$ , and on the  $d$ -dimensional vector of covariates, stored in the  $n \times d$  matrix  $\mathbf{x}$  with  $\mathbf{x}_i$  the vector corresponding to the  $i^{\text{th}}$  row, the likelihood for component  $j$  is given by

$$L_j(\mathbf{y}_j, \mathbf{x} \mid \theta_j, \gamma_j) = \prod_{i=1}^n \left[ \left( \frac{e^{\mathbf{x}_i^t \theta_j}}{1 + e^{\mathbf{x}_i^t \theta_j}} \right)^{y_{ij}^{(e^{\mathbf{x}_i^t \gamma_j})}} - \left( \frac{e^{\mathbf{x}_i^t \theta_j}}{1 + e^{\mathbf{x}_i^t \theta_j}} \right)^{(y_{ij}+1)^{(e^{\mathbf{x}_i^t \gamma_j})}} \right],$$

where we consider the logit and log links on the  $q$  and  $\beta$  parameters, respectively. Based on this likelihood, we perform inference on the marginal components using an adaptive Metropolis-Hastings scheme, as in Haselimashhadi et al. (2018). For the simulations and real data analysis in this paper, we set standard Gaussian priors on the regression coefficients  $\theta_j$  and  $\gamma_j$ . This step of marginal fitting is conducted offline, prior to the next steps for the fitting of the copula. The two-step procedure is an efficient approximation of the joint estimation problem and is often used in copula graphical modelling (Joe, 2005).

Once the marginals are estimated, inference of the network component requires an inverse mapping from the observed to the latent Gaussian space. As depicted visually in Figure 1, each observed discrete value corresponds to an interval in the latent Gaussian space with the same associated probability. Formally, given the  $n \times p$  observed data  $\mathbf{y}$  and the fitted marginals, the Gaussian latent variables  $\mathbf{z}$  are constrained in the intervals

$$\mathcal{D}_F(\mathbf{y}) = \{ \mathbf{z} \in R^{n \times p} : \Phi^{-1}(F_{ij}(y_{ij} - 1)) < z_{ij} \leq \Phi^{-1}(F_{ij}(y_{ij})) \}, \quad (3)$$

where we indicate with  $F_{ij}$  the cdf of  $Y_j$  when  $\mathbf{X} = \mathbf{x}_i$ . Rather than condensing these intervals into a single point, as in Cougoul et al. (2019), we retain this information within the MCMC sampling scheme, similar to the approach of Dobra and Lenkoski (2011) and Mohammadi et al. (2017) for ordinal data.

In particular, the extended rank likelihood function for a given graph  $G$  and associated precision matrix  $\mathbf{K} = \mathbf{R}^{-1}$  is defined as

$$L_E(\mathbf{z} \in \mathcal{D}_F(\mathbf{y}); \mathbf{K}, G) = \int_{\mathcal{D}_F(\mathbf{y})} P(\mathbf{z} \mid \mathbf{K}, G) d\mathbf{z}$$

where  $P(\mathbf{z} \mid \mathbf{K}, G)$  is the profile likelihood in the Gaussian latent space:

$$P(\mathbf{z} \mid \mathbf{K}, G) \propto |\mathbf{K}|^{n/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\mathbf{K}\mathbf{U}) \right\}$$

with  $\mathbf{U} = \mathbf{z}^t \mathbf{z}$  the sample moment. The likelihood is combined to priors to lead to the posterior

$$P(\mathbf{K}, G \mid \mathbf{z} \in \mathcal{D}_F(\mathbf{y})) \propto L_E(\mathbf{z} \in \mathcal{D}_F(\mathbf{y}); \mathbf{K}, G) P(\mathbf{K} \mid G) P(G) \quad (4)$$

where  $P(\mathbf{K} \mid G)$  denotes the prior distribution on the precision matrix  $\mathbf{K}$  for a given graph structure  $G$  and  $P(G)$  denotes a prior distribution for the graph  $G$ . Similar to Mohammadi

et al. (2017), one can show that the posterior distribution of each marginal  $Z_j$  conditional on the other  $\mathbf{Z}$  and on the precision matrix  $\mathbf{K}$  is given by a Gaussian distribution

$$Z_j | \mathbf{K}, \mathbf{Z}_{V \setminus \{j\}} = \mathbf{z} \sim N \left( - \sum_k \frac{K_{jk} z_k}{K_{jj}}, \frac{1}{K_{jj}} \right),$$

truncated on the interval

$$\left( \Phi^{-1}(F_j(Y_j - 1)) < Z_j \leq \Phi^{-1}(F_j(Y_j)) \right],$$

with  $F_j(\cdot)$  the discrete Weibull cdf linking  $Y_j$  to  $\mathbf{x}$ .

As regards to the prior specification on the graph  $G$ , we consider an Erdős-Rényi random graph with a link inclusion probability set to 0.2, representing the case of a sparse graph, unless stated otherwise. For other options, see Dobra and Lenkoski (2011) and Mohammadi and Wit (2015). As for the precision matrix  $\mathbf{K}$ , conditional on a given graph  $G$ , we consider a G-Wishart distribution, defined by

$$P(\mathbf{K} | G) = \frac{1}{I_G(b, \mathbf{D})} |\mathbf{K}|^{(b-2)/2} \exp \left\{ -\frac{1}{2} \text{Tr}(\mathbf{D}\mathbf{K}) \right\},$$

where  $b > 2$  is the degree of freedom,  $\mathbf{D}$  is a symmetric positive definite matrix, and  $I_G(b, \mathbf{D})$  is a normalizing constant (Roverato, 2002). For the simulations and real data analysis in this paper, we set  $b = 3$  and  $\mathbf{D} = \mathbb{I}_p$ , following Mohammadi et al. (2023).

As the space of possible graphs is very large, computationally efficient search algorithms are needed to sample from the posterior distribution (4). To efficiently explore the graph space, we consider the birth-death Markov chain Monte Carlo (BDMCMC) search algorithm developed by Mohammadi and Wit (2015). In particular, the algorithm explores the graph space by either adding (birth) or deleting (death) an edge to a graph  $G = (V, E)$ , independently of the rest and via a Poisson process with birth/death rates given by

$$R_e(G, \mathbf{K}) = \min \left\{ \frac{P(G^*, \mathbf{K}^* | \mathbf{z})}{P(G, \mathbf{K} | \mathbf{z})}, 1 \right\}, \text{ for each } e \in \{E \cup \bar{E}\}, \quad (5)$$

where  $G^* = (V, E \cup \{e\})$  for the birth of an edge  $e \in \bar{E}$ , while  $G^* = (V, E \setminus \{e\})$  for the death of an edge  $e \in E$ , and  $\mathbf{K}^*$  is the corresponding precision matrix. Since the birth/death events are independent Poisson processes, the time between two successive events has a mean waiting time given by

$$W(G, \mathbf{K}) = \frac{1}{\sum_{e \in \{E \cup \bar{E}\}} R_e(G, \mathbf{K})}. \quad (6)$$

Based on the above birth/death rates and waiting times, the birth and death probabilities that govern the move to a new graph are given by

$$P(\text{birth/death of edge } e \in \{E \cup \bar{E}\}) = R_e(G, \mathbf{K}) \times W(G, \mathbf{K}). \quad (7)$$

The pseudo-code for the BDMCMC search algorithm for sampling from the target posterior distribution (4) is reported in Algorithm 1. The first step of Algorithm 1 is to update



---

**Algorithm 1:** BDMCMC search algorithm for GCGM with DW marginals
 

---

**Input:** A graph  $G = (V, E)$  with a precision matrix  $\mathbf{K}$  and data  $\mathbf{y}$  and  $\mathbf{x}$ .

**for**  $N$  iteration **do**

**Step 1:** Sample the latent data for each marginal  $j$ , updating the latent  $n$  values  $z_j$  from their full conditional distribution:

$$Z_j | \mathbf{K}, \mathbf{Z}_{V \setminus \{j\}} = z \sim N \left( - \sum_k \frac{K_{jk} z_k}{K_{jj}}, \frac{1}{K_{jj}} \right),$$

each truncated on its corresponding interval in  $\mathcal{D}_{F_j}(\mathbf{y})$  from Equation (3);

**Step 2:** **for all the possible jumps in parallel do**

    Compute the birth and death rates by Equation 5;

    Compute the waiting time by Equation 6;

    Sample the graph based on the birth/death probabilities in Equation 7;

**Step 3:** Sample the precision matrix, according to the updated graph;

**Output:** Samples from the target posterior distribution (4).

---

the latent variables given the observed data. Then, in step 2, on the basis of the sampled latent data, the algorithm computes the birth/death rates. This is done in parallel since the rates associated to each edge can be calculated independently of each other. For details on how to calculate the birth/death rates see (Mohammadi et al., 2023, Section 2), while Figure 2 provides a visualization of the algorithm. Finally, step 3 of the algorithm can be

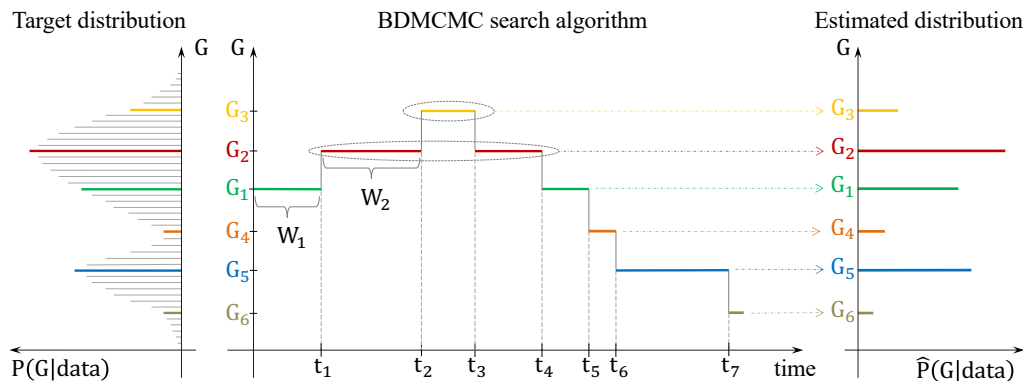


Figure 2: Graphical representation of the BDMCMC search algorithm over the graph space for the Step 2 of Algorithm 1. The left panel shows the target posterior distribution of the graphs, while the right panel represents its estimation based on the total waiting times of the graphs visited by the algorithm. The middle panel visualizes how the algorithm explores the graph space, where  $\{W_1, W_2, \dots\}$  are the waiting times and  $\{t_1, t_2, \dots\}$  are the jumping times of the algorithm. This figure is adapted from Mohammadi et al. (2023).

done by exact sampling from a G-Wishart distribution, as in Lenkoski (2013).

Following from the Bayesian inference of the Gaussian copula graphical model with discrete Weibull marginals, one can extract any information of interest for the analysis. In particular, from the marginal components, one obtains the posterior distribution of the regression coefficients and can investigate any effect of interest, while from the graph posterior, one can calculate the posterior edge inclusion probabilities:

$$P(\text{edge } e \in E \mid \text{data}) = \frac{\sum_{t=1}^N 1(e \in G^{(t)})W(G^{(t)}, \mathbf{K}^{(t)})}{\sum_{t=1}^N W(G^{(t)}, \mathbf{K}^{(t)})}, \quad (8)$$

where  $N$  denotes the MCMC iterations (after burn-in) and  $1(e \in G^{(t)}) = 1$  if  $e \in G^{(t)}$  and zero otherwise. These probabilities capture the full uncertainty on the graph learning, which is particularly useful in high dimensional settings such as the microbiome data.

### 3. Simulation study

The main objective of the proposed method is that of learning the underlying structure of dependence from complex and heterogeneous count data, which are routinely generated in genomic studies. We therefore conduct a simulation study to measure the performance of the method in this setting.

For the simulations, we consider networks with  $p$  nodes and a sparse random graph structure. Given a graph  $G$  and marginals  $F_j(\cdot)$ ,  $j = 1, \dots, p$ , we use the following procedure to simulate count data. We first generate a precision matrix from a G-Wishart distribution with  $b = 3$  and  $\mathbf{D} = \mathbb{I}_p$ , and standardize it to the inverse of a correlation matrix. We then draw  $n$  multivariate normal samples from  $N_p(\mathbf{0}, \mathbf{K}^{-1})$ . This generates a matrix  $\mathbf{z}$  of dimension  $n \times p$ . Finally, we obtain the discrete data using  $y_{ij} = F_j^{-1}(\Phi(z_{ij}))$ , for  $i = 1, \dots, n$  and  $j = 1, \dots, p$ , with  $\Phi(\cdot)$  the standard normal distribution and  $F_j(\cdot)$  a distribution function of a specified shape as detailed in each study below.

We evaluate the performance of the method in terms of parameter estimation and graph recovery. For the first, we compare the true precision matrix  $\mathbf{K}$  with the posterior mean estimate  $\widehat{\mathbf{K}}$  using the Kullback-Leibler divergence between the corresponding mean-zero multivariate Gaussian distributions. This is available in closed form and is given by (Pardo, 2018):

$$KL(\widehat{\mathbf{K}}) = 0.5(\text{Tr}(\mathbf{K}\widehat{\mathbf{K}}^{-1}) + \text{Tr}(\widehat{\mathbf{K}}\mathbf{K}^{-1}) - p).$$

Instead, for graph recovery, we use the function `auc` in the `pROC` R package to calculate the area under the Receiver Operating Characteristic (ROC) curve. The latter is obtained by setting cutoffs on the posterior edge inclusion probabilities in Equation (8). At each cutoff, the  $x$  and  $y$  coordinates of the point on the curve are given by the false positive rate and the true positive rate, respectively, of the estimated graph for that cutoff and with respect to the true graph structure.

#### 3.1 Effect of increase in $p$ , graph density and zero inflation

In a first simulation study, we evaluate the performance of the proposed Gaussian copula graphical model with discrete Weibull marginals, which we abbreviate to DWGM. In particular, we test how the performance of the method is affected by the dimensionality  $p$ , the sparsity of the graph  $G$  and the level of zero inflation  $\pi$ .

We generate data as described before, with marginals  $F_j(\cdot)$  given by discrete Weibull distributions linked to one external covariate. In particular, we consider a binary covariate  $X$  drawn from a Bernoulli(0.5), e.g., observations split into two groups (like the two environments, stool and saliva, in the real application in Section 4). For the regression parameters in Equation (2), we set a constant  $\log(\beta(x)) = \gamma_0 = \log(0.7)$  for all  $p$  variables, while  $q$  values that differ across the two conditions and the  $p$  variables are obtained by setting  $\log(q(x)/(1-q(x))) = \theta_0 + \theta_1 x$ , with  $\theta_0$  drawn from a  $N(0,0.1)$  and  $\theta_1$  from a  $N(2,0.01)$ . This choice of parameters leads to generally over-dispersed data (Peluso et al., 2019).

We set  $n = 100$ ,  $p = 50$ , a link inclusion probability in  $G$  equal to 0.05 and no zero inflation ( $\pi = 0$ ). We then measure how the performance of the method varies when increasing  $p$  to 100, the graph density to 20% or  $\pi$  to 0.2 for all predictors. Figure 3 shows the AUC and KL values across 50 simulated datasets. For each dataset, we run the Bayesian inferential procedures for 1000 iterations for each marginal and 10k iterations for the structure learning (Algorithm 1), and set the priors as specified in the description of the method. As expected, and in line with similar studies in the literature, we find that the performance deteriorates as  $p$  increases, both in terms of graph recovery (Figure 3, top left) and parameter estimation (Figure 3, top right). Similarly, we find a deterioration in both performance measures as the link inclusion probability increases from 0.05 to 0.2, i.e., the denser the graph becomes (Figure 3, middle panel), and as the level of zero inflation increases from 0 to 0.2 (Figure 3, bottom panel). Indeed, an increase in zero inflation results in an increase in the percentage of zeros, which goes on average from 32% when  $\pi = 0$  to 45% when  $\pi = 0.2$  across the 50 predictors. This induces a weaker signal on the dependence structure.

Zero inflation is expected to have an effect also on the estimation of the marginal effects. Figure 4 shows how estimation of the marginal effect  $\theta_1$  is more accurate the lower the level of zero inflation, with a bias that decreases the larger the sample size  $n$  is. The bias has an effect also on the identification of the covariate  $X$  as an important predictor. When  $\pi = 0.8$  and  $n = 100$ , 2.8% of the times the predictor is not included in the marginal model according to the 90% High Posterior Density (HPD) interval on  $\theta_1$  across the 50 predictors. This percentage reduces to 0.01% when there is no zero inflation ( $\pi = 0$ ), and goes down to 0% when  $n = 1000$  for all cases.

The computational time of this simulation study is mostly affected by  $p$ . In particular, an analysis on one of the datasets with  $p = 50$ ,  $n = 100$ , link inclusion probability equal to 0.2 and zero inflation  $\pi = 0$  required approximately 80 seconds, compared to 333 seconds when  $p = 100$ .

### 3.2 Effect of covariate adjustment on structural learning

In a second simulation study, we evaluate the effect of covariate adjustment on structural learning. To this end, we consider the same generative process as in the previous section and concentrate on the case  $p = 50$ ,  $n = 100$  and 0.2 link inclusion probability. As a benchmark, we also consider the case of no covariates in the marginal models, i.e.,  $\theta_1 = 0$ . We then compare the proposed DWGM with the Gaussian copula graphical model (GCGM) for ordinal data of Mohammadi et al. (2017), implemented in the R package `BDgraph`. We use 10k iterations also in this case and the same prior specifications. In the absence of covariates,

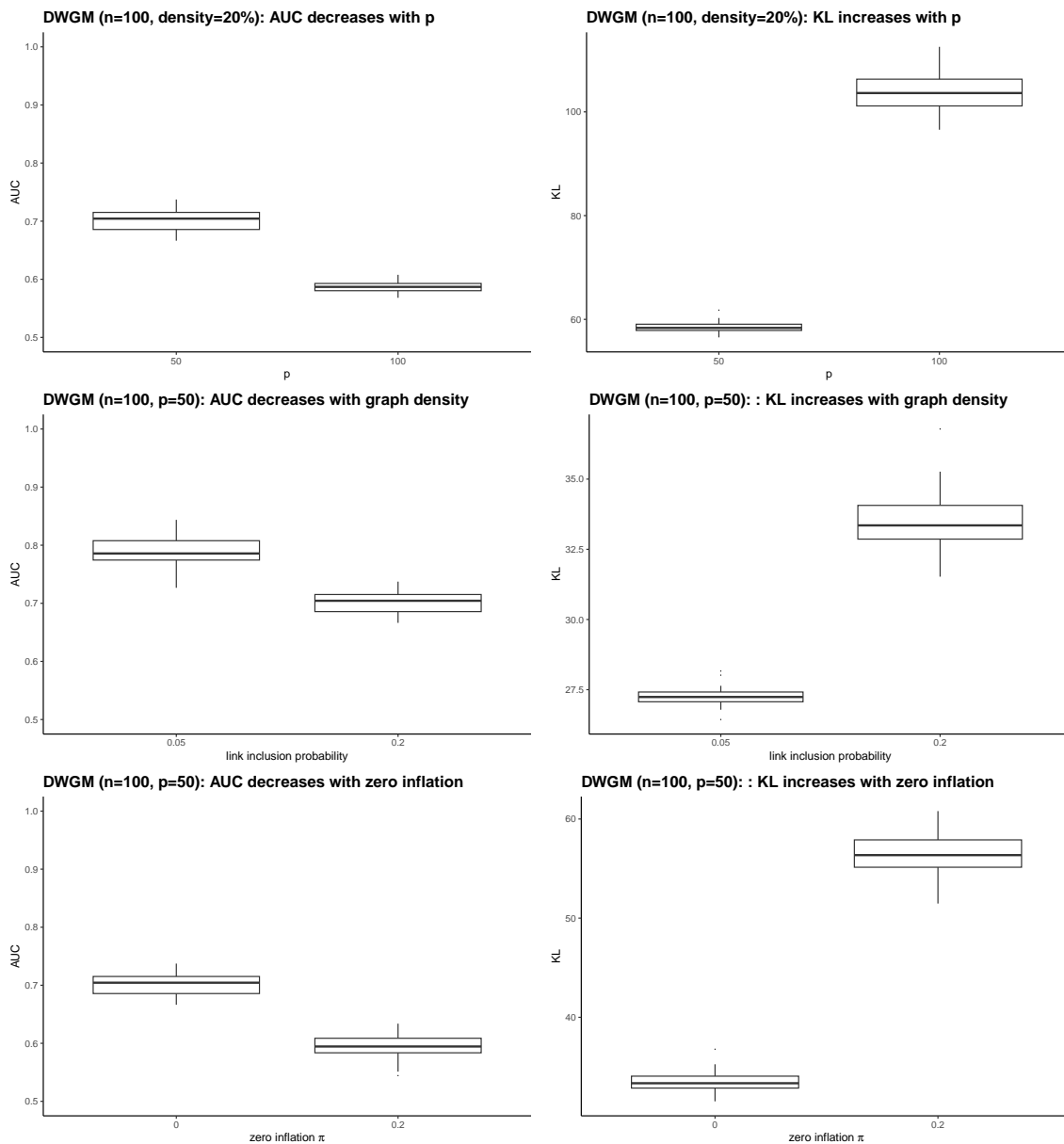


Figure 3: Performance of DWGM, in terms of accuracy of network recovery (left) and parameter estimation (right), decreases with increasing dimensions (top), graph density (middle) and zero inflation (bottom). Data are simulated from a Gaussian copula graphical model with DW marginals. Boxplots are the result of 50 simulations.

the GCGM method is similar to our approach, the only difference being the (non-parametric) empirical marginal distributions used in GCGM versus the parametric distributions with an unbounded support for the marginals in DWGM. Clearly, being non-parametric, GCGM does not lend itself easily to the inclusion of external covariates in the marginals.

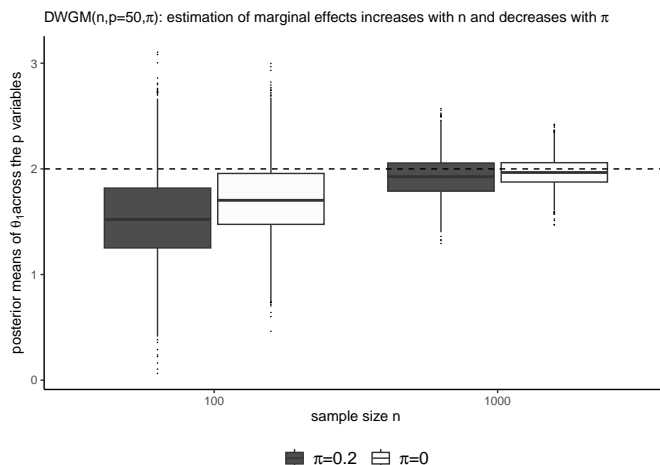


Figure 4: Performance of DWGM, in terms of estimation of the marginal effects, increases with sample size  $n$  and decreases with zero inflation  $\pi$ . Data are simulated from a Gaussian copula graphical model with DW marginals with zero inflation ( $\pi = 0.2$ ) and without ( $\pi = 0$ ), under two sample sizes ( $n \in \{100, 1000\}$ ). Boxplots are for the mean posterior estimates of  $\theta_1$  across the  $p$  predictors and 50 simulations. The dotted horizontal line corresponds to the average true value of  $\theta_1$  across the 50 predictors.

Figure 5 evaluates the two approaches in terms of graph recovery (left) and parameter estimation (right) across 50 simulations. As expected, in the absence of covariates ( $\theta_1 = 0$ ),

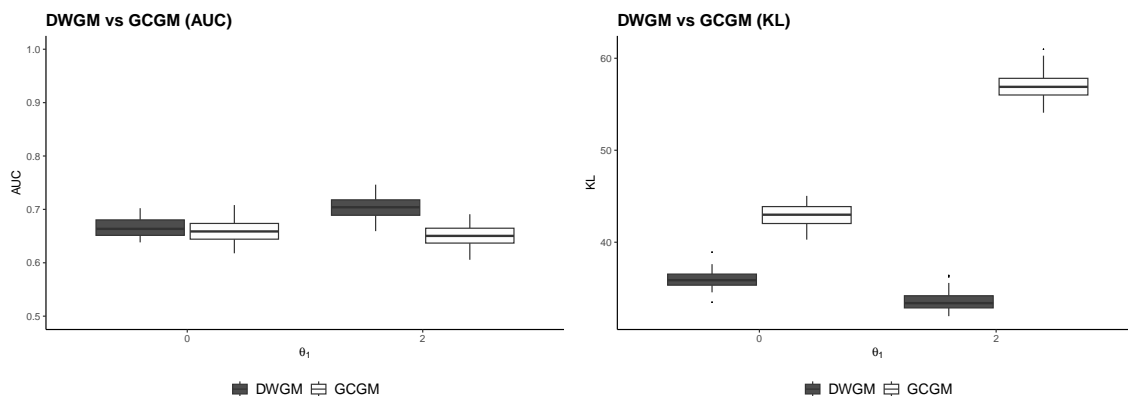


Figure 5: Performance of DWGM, in terms of accuracy of network recovery (left) and parameter estimation (right), is compared with a Gaussian copula graphical model with non-parametric marginals (GCGM). Data are simulated from a Gaussian copula graphical model with  $n = 100$ ,  $p = 50$ , a link inclusion probability of 0.2 and DW marginals without ( $\theta_1 = 0$ ) and with ( $\theta_1 = 2$ , on average across the  $p$  variables) the presence of an external covariate  $X$ . Boxplots are the result of 50 simulations.

the two approaches perform similarly, although the correctly specified parametric marginals lead to a better estimation of the precision matrix. In contrast to this, the improvement in structural learning after adjusting for covariate effects is clear when data are simulated with a  $\theta_1$  marginal effect equal to 2 on average across the  $p$  variables, both in terms of graph recovery and parameter estimation. Looking closely at graph recovery, and estimating a graph structure by setting a 0.5 cutoff on the posterior edge probabilities, we find that the **GCGM** estimated graph is in general denser (14% density on average versus 9% for **DWGM**), with a worse false positive rate (9% on average versus 3% for **DWGM**) and only a marginally better false negative rate (66% versus 67% for **DWGM**). This suggests that not correcting for the covariates at the marginal level may result in the detection of spurious dependences between the variables, as noted also in other studies (Vinciotti et al., 2016, 2023).

The computational time for one of the datasets was about 79 seconds for **DWGM** and 36 seconds for **GCGM**. The difference is mostly attributed to the fitting of the marginal distributions. This has a negligible computational time in **GCGM**, while MCMC sampling for the marginal fitting in **DWGM** required about 45 seconds. Future versions of the **BDgraph** package will consider a parallel implementation for the marginal fitting across the  $p$  variables.

### 3.3 Effect of miss-specified marginal distributions

In a third simulation study, we evaluate the robustness of **DWGM** to count data simulated with negative Binomial marginals. We consider a similar setting to the first simulation, i.e., one binary covariate, a constant dispersion parameter  $\phi = 0.5$  and a mean  $\mu$  dependent on  $X$ , with  $\log(\mu(x)) = \theta_0 + \theta_1 x$ , with  $\theta_0$  drawn from a  $N(0, 0.1)$  and  $\theta_1$  from a  $N(2, 0.01)$  across the  $p$  variables. As with the second simulation, we fix  $n = 100$ ,  $p = 50$  and a link inclusion probability equal to 0.2.

We compare our proposed method with a Gaussian copula graphical model with negative Binomial regression marginals, implemented in the R package **rMAGMA** (Cougoul et al., 2019). As well as using a different distribution for the marginals, inference in **rMAGMA** is conducted using a frequentist paradigm and, in addition, it does not make use of the extended rank likelihood approach. Indeed, the fitted marginals are used to transform the data into the latent variables by taking the mean of the interval, i.e.,  $z_{ij} = \Phi^{-1}\left(\frac{F_j(y_{ij} - 1) + F_j(y_{ij})}{2}\right)$ , and then graphical lasso is used on the transformed data.

Figure 6 reports the accuracy of the methods in terms of graph recovery (left) and parameter estimation (right). For **rMAGMA**, which is based on penalised inference, the ROC curve is constructed across the path of solutions generated by the tuning penalty parameter, while the Kullback-Leibler is calculated based on the optimal model selected using the stability approach for regulation selection (stars) criterion (Liu et al., 2010).

The results show, firstly, similar results for **DWGM** compared to the previous simulations (Figure 3 and Figure 5 for  $p = 50$ ,  $n = 100$  and 20% graph density), suggesting a robustness of **DWGM** against the marginal model specification. Secondly, **DWGM** has a better performance than **rMAGMA**, particularly when it comes to the accuracy of the estimated precision matrix  $\mathbf{K}$  of the optimal model. Since **rMAGMA** is in fact correctly specified in this simulation, we speculate that the posterior edge probabilities of a Bayesian structural learning procedure lead to a better separation between the presence/absence of links than the edge weights calculated across the penalised path of the frequentist **rMAGMA** approach.

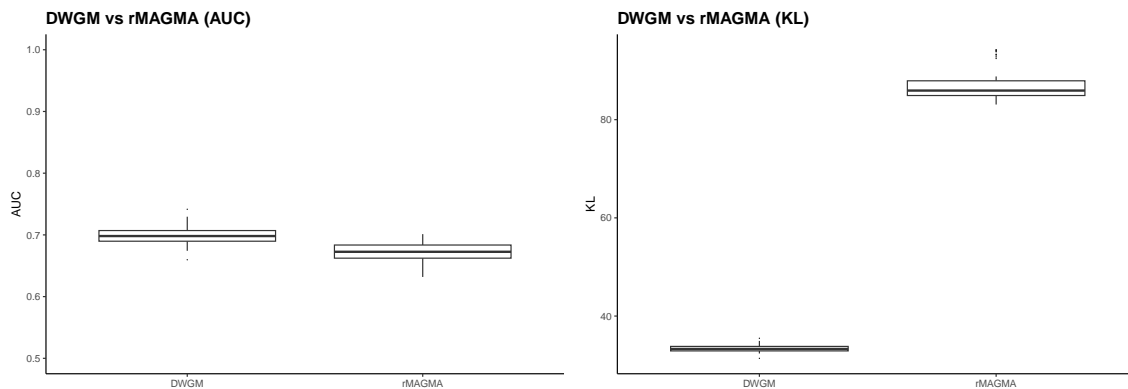


Figure 6: Performance of DWGM, in terms of accuracy of network recovery (left) and parameter estimation (right), is compared with a Gaussian copula graphical model with parametric NB marginals (rMAGMA). Data are simulated from a Gaussian copula graphical model with  $n = 100$ ,  $p = 50$ , a link inclusion probability of 0.2 and NB marginals. Boxplots are the result of 50 simulations.

#### 4. Inferring the network of the microbiota

In this section, we use Gaussian copula graphical models with discrete Weibull marginals to recover the network of interactions between microbial species. Interactions between microbes are fundamental in shaping the structure and functioning of the human microbiota, and their malfunctioning has been linked to a number of medical conditions. A lack of understanding of how these interactions shape and evolve makes it difficult to predict their relevance in biomedical fields. For these reasons, microbiota systems have been intensively studied in recent years. Large consortia have developed technologies for the collection of high-throughput data of the microbiome, e.g., the Human Microbiome Project (HMP Consortium, 2012) and the Metagenomics of the Human Intestinal Tract (MetaHIT) project (Qin et al., 2010). These have paved the way for further studies investigating the association of the microbiota functioning with a number of medical conditions, such as obesity (Le Chatelier et al., 2013) and diabetes (Pedersen et al., 2016), as well as with the response to certain treatments, such as immunotherapy (Lee et al., 2022).

In this illustration, we focus in particular on the gut and oral microbiome. The microbial communities in the mouth and colon are connected anatomically via the saliva. However, the extent to which oral microbes reach and colonize the gut is yet under debate (Rashidi et al., 2021). To resolve this long-standing controversy, many studies have been devoted to study jointly the human stool and saliva microbiome profiles. To this end, we apply our methodology to recover a core network of interactions between microbes across the two different environments. Crucially, the method that we have developed takes into account both the fact that the OTU abundances may differ marginally between the body sites (stool and saliva) and that the data may be affected by potential experimental effects.

As in Cougoul et al. (2019), we retrieve the 16S variable region V3-5 data from the Human Microbiome Project (HMP Consortium, 2012) and perform the analysis at the level

of Operating Taxonomic Units (OTUs). After filtering samples with less than 500 reads, we consider microbiomes from 663 healthy individuals, with microbial concentration measured from either stool or saliva. We then restrict our attention to the 155 OTUs which are present in at least 25% of the samples and with more than two distinct observed values in both the saliva and stool samples. Finally, in order to account for varying sequencing depths change significantly between samples, we estimate the library size of each sample by the geometric mean of pairwise ratios of OTU abundances of that sample with all other samples (Cougoul et al., 2019).

In the next sections, we use the proposed approach on the microbiome data with  $p = 155$  OTUs (the nodes of the network) and  $n = 663$  samples, accounting for the marginal effects of the location in the body (stool or saliva) and the library size of each biological sample.

#### 4.1 Accounting for covariates via DW regression marginals

We fit discrete Weibull marginal models, linking both parameters  $q$  and  $\beta$  to body site and sequencing length. We take sequencing length in the log scale, which is more in line with its use in the literature as an offset of a negative Binomial model (Cougoul et al., 2019). Including also an interaction between the two covariates leads to 8 regression parameters per marginal component. As the data are sparse (with on average a percentage of zeros per OTU equal to 62%, and ranging from 40.7% to 75%), we fit also a zero-inflated discrete Weibull distribution, with a zero inflation parameter  $\pi_j(\mathbf{x})$  for component  $j$ , which we let vary between the two body sites. On each of the two additional parameters, we place a Beta(1,1) prior distribution.

For each marginal and for each of the two models (discrete Weibull and zero-inflated discrete Weibull), we use 10k MCMC iterations, retaining the last 25% as samples from the posterior distribution. Trace plots of the regression parameters showed that this number of iterations was sufficient to reach convergence. A comparison of the zero-inflated versus the standard DW regression model using the Bayesian Information Criterion (BIC) showed that 24 out of the 155 OTUs necessitated the zero-inflated component of the model. As a matter of comparison, we also fitted negative Binomial, using its most common formulation with a mean dependent on the covariates and a constant dispersion parameter, as in Cougoul et al. (2019). Here we found that 15 out of the 155 OTUs were better fitted with a zero-inflated NB model. These results show how using a zero-inflated model upfront because the data are very sparse, as done in most of the literature on microbiome analyses, may not necessarily be the best option and that conducting model selection between the two models, as done in this paper, is a better choice.

Figure 7 reports the results, whereby for each OTU we consider the best model between the zero-inflated and the non-zero inflated version. The BIC comparison shows similar performance between discrete Weibull and negative Binomial models, with some cases showing a significantly better fit for discrete Weibull. The top right plot shows the dispersion ratio (i.e., the variance divided by the mean) from the fitted DW marginals for each observation and each OTU. The plot shows how the data are highly over-dispersed in both conditions, a setting where NB is typically the default choice. Finally, the bottom plots show the OTU with the largest BIC difference (left plot), i.e., the OTU best fitted by DW when compared to NB, and the OTU with the lowest BIC difference (right plot). In both cases, we plot the



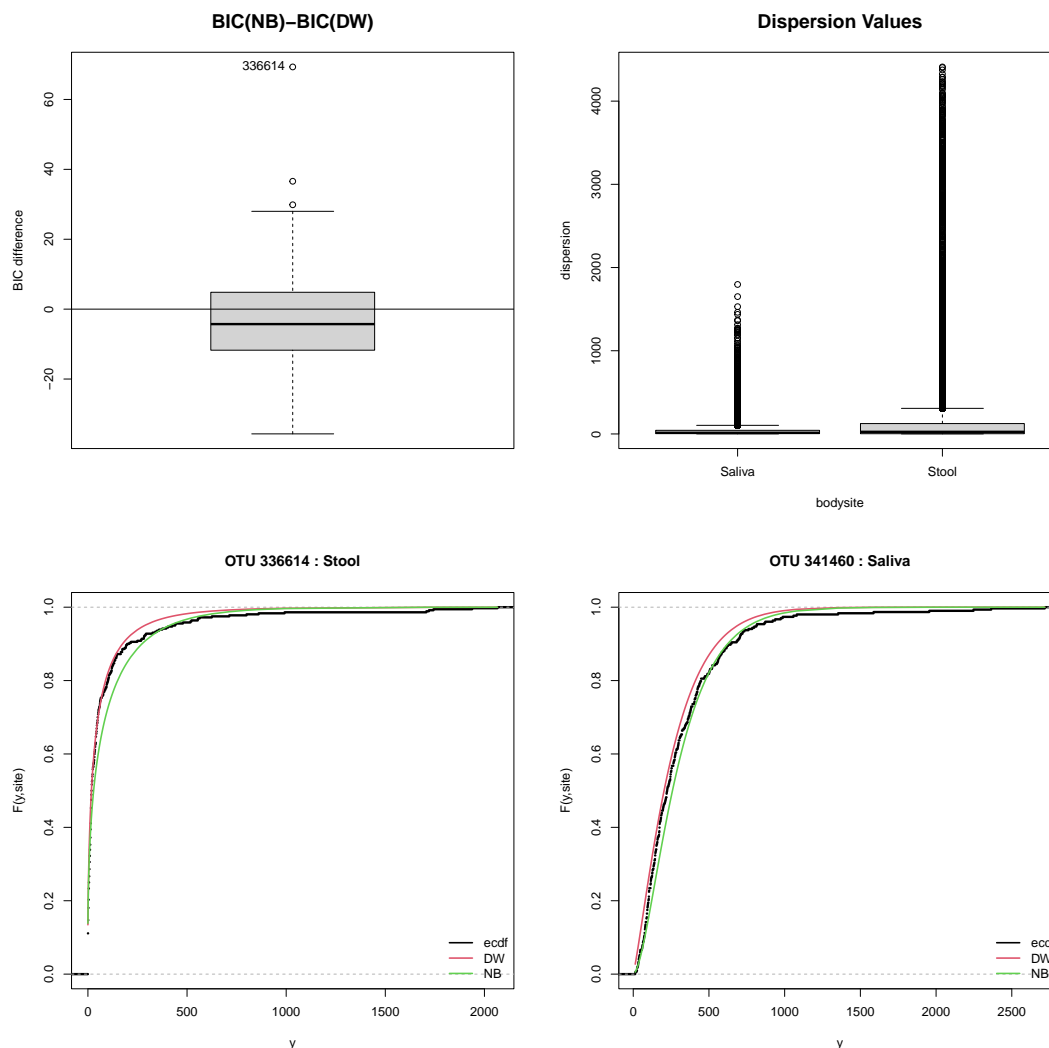


Figure 7: Top: Boxplot of BIC differences between the NB model and the DW model across all 155 OTUs (left) and dispersion levels from the fitted DW model for each OTU and each observation, split by body site (right). Bottom: Cumulative distribution functions (empirical and fitted) corresponding to a specific body site for the OTU best fitted by DW (BIC difference = 68.26235, left) and by NB (BIC difference = -31.07967, right) .

cumulative distribution functions of DW and NB associated to the body site which shows the biggest difference, while taking an average of the parameters across the normalizing factor. Superimposing these fitted distributions on the empirical cumulative distribution functions associated to the two groups shows the extent of the discrepancy between the two models.

Including covariates in the inference of microbiota systems has the advantage that analyses that are typically conducted on a microbe by microbe basis, such as Lee et al. (2020), are

now naturally embedded in the overall joint model. Indeed, one can inspect the estimation and inference of any marginal effect of interest. In this particular analysis, there is interest in detecting the OTUs that are differentially expressed between the two different body sites. Figure 8 shows how all 155 OTUs differ significantly between the two body sites. Further-

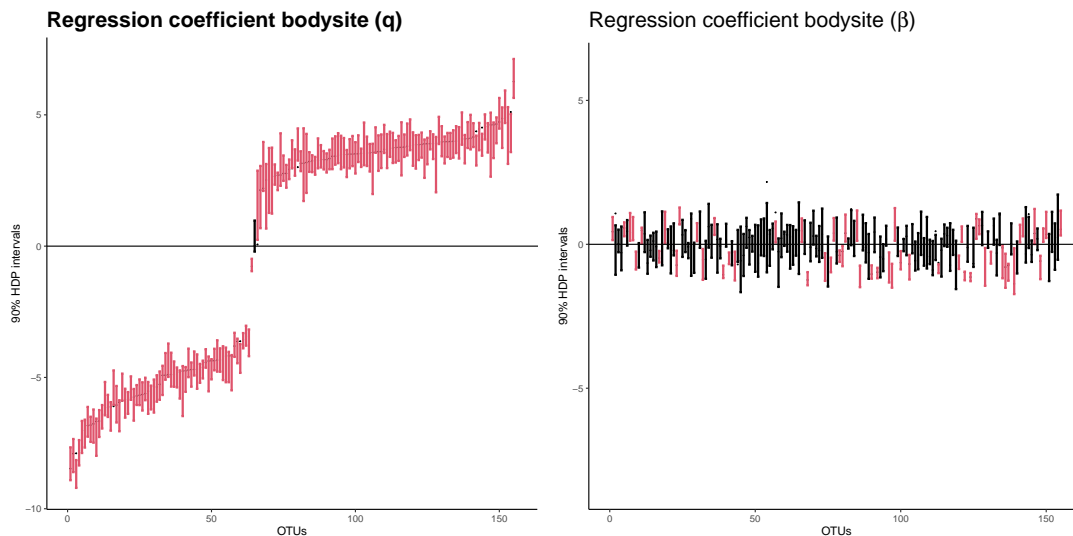


Figure 8: 90% High Posterior Density (HPD) intervals of the  $\theta$  (left) and  $\gamma$  (right) regression parameter corresponding to the body site covariate, sorted according to the median of the  $\theta$  regression coefficient across the posterior samples. Intervals that do not contain the zero are coloured in red.

more, the plots show how the regression coefficient of the  $q$  parameter is highly significant, suggesting large differences between the proportion of zeros in the two environments for most OTUs. In contrast to this, the regression coefficient of the  $\beta$  parameter is less significant. This may in fact indicate that a simpler DW regression model, with a constant  $\beta$  parameter, may be sufficient for some of the OTUs. As shown in the simulation (Figure 5), these marginal effects, if left unaccounted for, could significantly distort the inference of the microbiota system, which we discuss in the next section.

## 4.2 Bayesian structure learning of the microbiota system

We now turn to the main task of recovering the underlying network of dependencies between the OTUs. The space of possible graphs among 155 nodes is huge, creating a statistical and computational challenge at a level that has not been considered before in the context of Bayesian structure learning. Thus a few checks and considerations were made. Firstly, we start the MCMC chain by setting the initial graph to the empty graph, as we expect a sparse graph. Secondly, we perform the structure learning for a long number of iterations, namely 10 million MCMC iterations. Thirdly, we check the trace of the posterior edge probabilities and graph sizes for convergence. We also check the sensitivity to the graph prior, by setting the link inclusion probability once to 0.2, the default value in `BDgraph`, and a second time to

0.04, which is the sparsity level of the graph detected by **rMAGMA** using zero-inflated negative Binomial as marginals and the stability selection criterion **stars** for model selection (Liu et al., 2010). Overall, we observe a high correlation among the edge posterior probabilities from the two chains (0.97). For the rest of the analysis, we consider the chain with link inclusion probability 0.04, which resulted in the highest log-likelihood when evaluated at the posterior estimates of the marginal distributions and precision matrix.

We firstly investigate the impact that the inclusion of covariates has on the inference of the underlying graph. To this end, we compare our proposed approach with a Gaussian copula graphical model that uses the empirical distribution for the marginals (**GCGM**), i.e., a model that does not make use of covariate information. We run also **GCGM** using 10 million iterations for the structure learning part and using the same prior on the graph (link inclusion probability 0.04). Figure 9 shows how the posterior distribution on graph sizes from the **DWGM** model (left) is concentrated on a sparser graph compared with the posterior distribution from the **GCGM** model. This was found also in the simulation study (Figure 5) and suggests that spurious links may have been detected by this second analysis where covariates have been omitted.

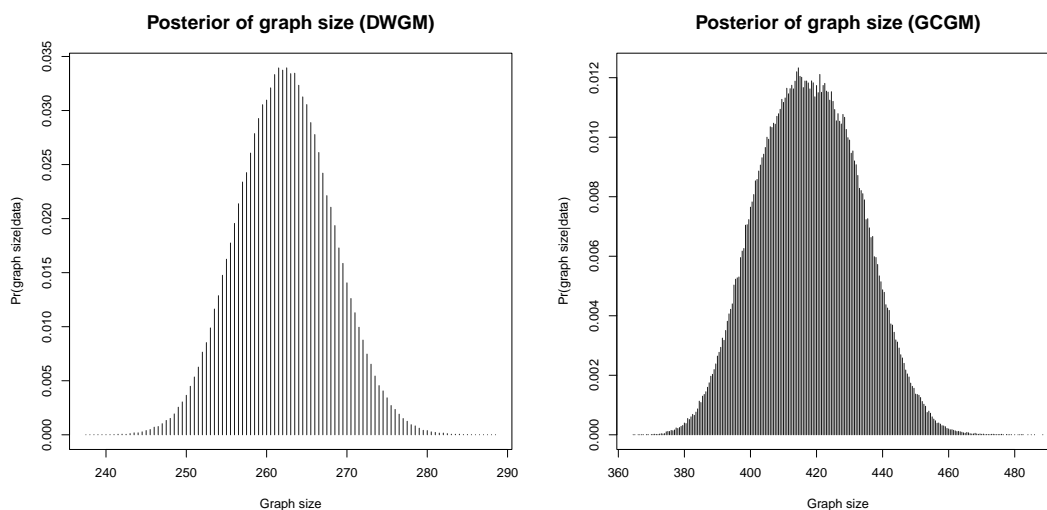


Figure 9: Posterior distribution of graph sizes for the model **DWGM** that accounts for covariates (left) versus the Gaussian copula graphical model that does not make use of covariates (**GCGM**, right).

Setting a cutoff of 0.5 on the edge posterior probabilities, the network contains 359 edges. Figure 10 shows the overlap between these edges and the optimal graph detected by **rMAGMA** (Cougoul et al., 2019). There is a moderate (178 edges) overlap with the total of 414 edges detected by **rMAGMA**. One of the major advantages of **DWGM**, which has not been considered before in the context of microbiome analyses, is that the uncertainty around the optimal graph is also measured. This is particularly important for structure learning in high dimensions, as noticed also in the simulation (Figure 6). Indeed, the right plot of Figure

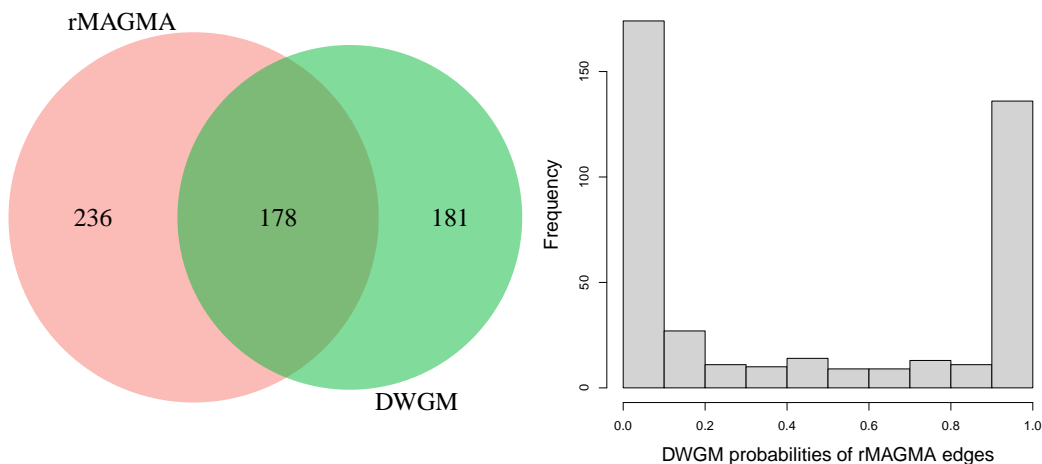


Figure 10: Left: Venn diagram comparing the optimal graph estimated by **rMAGMA** using the **stars** criterion and the optimal graph estimated by **DWGM** by setting a cutoff of 0.5 on the posterior edge probabilities. Right: Histogram of **DWGM** posterior edge probabilities associated to the 414 edges detected by **rMAGMA**.

10 shows how many of the edges detected by **rMAGMA** have a low posterior edge probability calculated by **DWGM**.

Finally, Figure 11 plots the network inferred by **DWGM**, with nodes coloured according to their phyla association (firmicutes, proteobacteria, bacteroidetes, actinobacteria, fusobacteria) and edges coloured according to their partial correlations, computed from the Bayesian averaging estimate of the precision matrix. The information on phyla association is useful in distinguishing between the stool and saliva microbiota. Indeed, firmicutes and bacteroidetes represent more than 90% of the total human gut microbiota (Qin et al., 2010) and have been found associated with several pathological conditions affecting the gastrointestinal tract, obesity and type 2 diabetes (Magne et al., 2020; Indiani et al., 2018). So we take this group of OTUs as representative of the gut microbiota. The remaining OTUs, with a mix of phylum levels, are instead associated to the saliva microbiota (Choi et al., 2020). In the optimal network (Figure 11), more connections are found within each group (average posterior edge probability equal to 4.7% in gut and 10% in saliva) than between the two groups (average posterior edge probability equal to 1.6% between gut and saliva). While the connections are particularly strong within each group (average of posterior edge probabilities larger than 0.5 equal to 89% in gut and 87% in saliva, with associated average absolute partial correlation equal to 0.23 and 0.21, respectively), strong connections are detected also between the two groups (average of posterior edge probabilities larger than 0.5 between gut and saliva equal to 86%, with associated average absolute partial correlations equal to 0.17), suggesting the presence of interactions between the two systems and supporting existing knowledge that oral microbes have the capacity to spread throughout the gastrointestinal system.

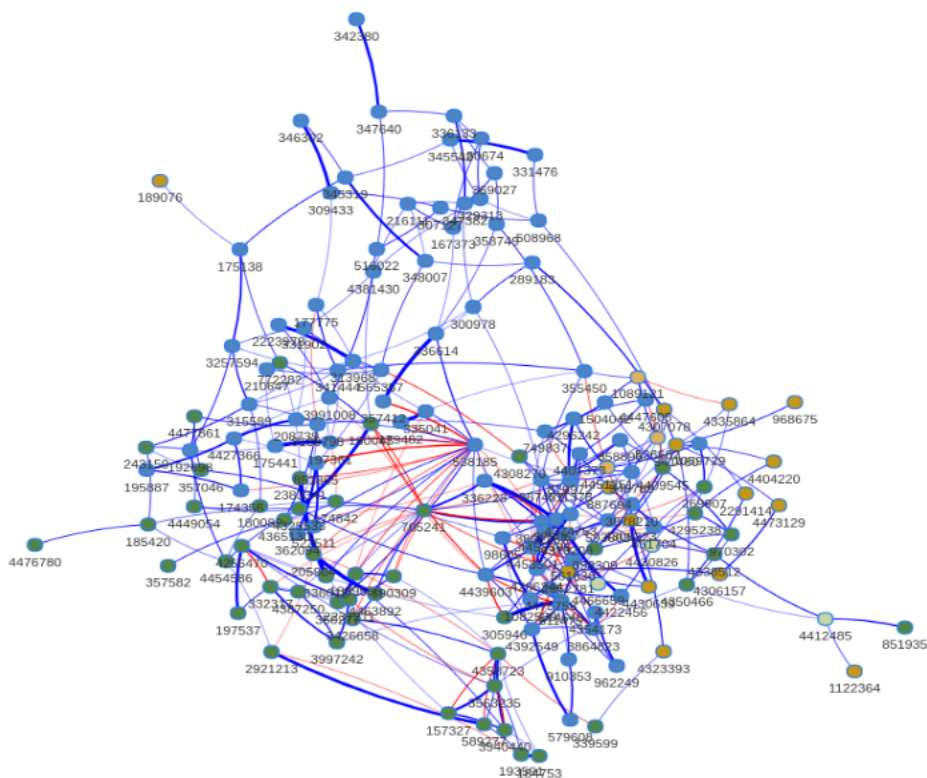


Figure 11: The inferred microbiota system with posterior edge probabilities greater than 0.5. Node colours are associated to OTU phyla, with ● Firmicutes, ● Bacterobacteria, ● Proteobacteria, ● Actinobacteria and ● Fusobacteria. Edge colors are associated to positive — and negative — partial correlations, and width of the edges to their absolute values.

## 5. Conclusion

In this paper, we have presented a copula graphical modelling approach that is able to recover the core dependence structure from high dimensional and heterogeneous count data. We have shown the usefulness of this approach in learning interactions between microbes from count data provided by the latest microbiome experiments, featuring high dimensionality, sparsity, heterogeneity and compositionality. The approach has three key features.

Firstly, it allows to adjust for the effect of covariates in the marginal components of the model. This is useful, both in quantifying the effect of covariates of interest on the count variables and in aiding network recovery. The latter is down to two reasons: on one hand, the inclusion of covariates removes spurious dependencies that may be induced by the effect of the covariates on the variables of interest; on the other hand, the inclusion of (particularly continuous) covariates at the marginal level expands the region of support for consistent estimation of the copula in the case of discrete variables.

Secondly, discrete Weibull regression is used for modelling the marginal distributions conditional on the covariates and is shown to be a simple (two parameters) yet flexible (broad

dispersion levels) choice compared to more commonly used distributions for count data. Moreover, its definition as a discretized continuous Weibull distribution provides a latent continuous space in the vicinity of the data with a one-to-one mapping with the inferred conditional independence graph. This may be useful in deriving theoretical properties of the proposed approach. The connection between discrete and continuous Weibull distributions may also be useful in defining novel multivariate rounded kernels (Canale and Dunson, 2011) and in using these to develop a mixture model for multivariate count data with parametric marginals.

Thirdly, a Bayesian inferential procedure based on the extended rank likelihood and on an efficient continuous-time birth-death process allows to account for the full uncertainty both in the marginals, and thus in the covariate effects, and in the graph component. The latter is important, particularly in high dimensional settings where model selection methods for regularized approaches do not work well and where there is typically a large uncertainty around the optimal graph. The method proposed captures this uncertainty at the level of the graph structure (via posterior probabilities of each link) and intensity of the interactions (via posterior estimates of partial correlations), but any other graph statistics of interest can be estimated via Bayesian averaging.

The simulation study and the real data analysis of microbiome data show the usefulness of the proposed approach at inferring networks from high-dimensional count data in general, and its relevance in the context of microbiota data analyses in particular. Indeed, the inferred interactions between firmicutes and bacteroidetes in the gut microbiota can create an opportunity for microbiome research to develop new microbial targets for the nutritional or therapeutic prevention and management of pathological conditions affecting the gastrointestinal tract, such as inflammatory bowel diseases, obesity and type 2 diabetes. At the same time, the analysis proposed has shown the potential to detect crucial interactions between the gut and oral microbiota, which has been suggested only recently in the literature.

## Software

The method proposed in this paper is implemented in the R package `BDgraph` which is freely available from the Comprehensive R Archive Network (CRAN) at <http://cran.r-project.org/packages=BDgraph>.

## Acknowledgments

This project was partially supported by the European Cooperation in Science and Technology (COST) [COST Action CA15109 European Cooperation for Statistics of Network Data Science (COSTNET)].

## References

Fentaw Abegaz and Ernst C. Wit. Copula Gaussian graphical models with penalized ascent Monte Carlo EM algorithm. *Statistica Neerlandica*, 69:419–441, 2015.

- Pariya Behrouzi and Ernst Wit. Detecting epistatic selection with partially observed genotype data by using copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 68(1):141–160, 2019.
- Divan Burger, Robert Schall, Johannes Ferreira, and Ding-Geng Chen. A robust Bayesian mixed effects approach for zero inflated and highly skewed longitudinal count data emanating from the zero inflated discrete Weibull distribution. *Statistics in Medicine*, 39(9):1275–1291, 2020.
- Antonio Canale and David B. Dunson. Bayesian kernel mixtures for counts. *Journal of the American Statistical Association*, 106(496):1528–1539, 2011.
- Subrata Chakraborty. Generating discrete analogues of continuous probability distributions - A survey of methods and constructions. *Journal of Statistical Distributions and Applications*, 2(1):1–30, 2015.
- Da Hyeon Choi, Jiwon Park, Ju Kwang Choi, Kyeong Eun Lee, Won Hee Lee, Jinho Yang, Ju Yeon Lee, Yoon Jeong Park, Chan Oh, Ho-Ryun Won, Bon Seok Koo, Jae Won Chang, and Yoon Shin Park. Association between the microbiomes of tonsil and saliva samples isolated from pediatric patients subjected to tonsillectomy for the treatment of tonsillar hyperplasia. *Experimental & Molecular Medicine*, 52(9):1564–1573, 2020.
- Arnaud Cougoul, Xavier Bailly, and Ernst Wit. MAGMA: inference of sparse microbial association networks, 2019. bioRxiv.
- Adrian Dobra and Alex Lenkoski. Copula Gaussian graphical models and their application to modeling functional disability data. *The Annals of Applied Statistics*, 5(2A):969–993, 2011.
- Adrian Dobra and Reza Mohammadi. Loglinear model selection and human mobility. *The Annals of Applied Statistics*, 12(2):815–845, 2018.
- Jerome Friedman, Trevor Hastie, and Robert Tibshirani. Sparse inverse covariance estimation with the graphical lasso. *Biostatistics*, 9(3):432–441, 2008.
- Jonathan Friedman and Eric Alm. Inferring correlation networks from genomic survey data. *PLoS Computational Biology*, 8(9):e1002687, 2012.
- Christian Genest and Johanna Nešlehová. A primer on copulas for count data. *ASTIN Bulletin: The Journal of the IAA*, 37(2):475–515, 2007.
- Hamed Haselimashhadi, Veronica Vinciotti, and Keming Yu. A novel Bayesian regression model for counts with an application to health data. *Journal of Applied Statistics*, 45(6):1085–1105, 2018.
- HMP Consortium. A framework for human microbiome research. *Nature*, 486(7402):215–221, 2012.
- Peter Hoff. Extending the rank likelihood for semiparametric copula estimation. *The Annals of Applied Statistics*, 1(1):265–283, 2007.

- Cláudia Maria dos Santos Pereira Indiani, Karina Ferreira Rizzardi, Paula Midori Castelo, Lúcio Fábio Caldas Ferraz, Michelle Darrieux, and Thaís Manzano Parisotto. Childhood obesity and firmicutes/bacteroidetes ratio in the gut microbiota: a systematic review. *Childhood obesity*, 14(8):501–509, 2018.
- Harry Joe. Asymptotic efficiency of the two-stage estimation method for copula-based models. *Journal of Multivariate Analysis*, 94(2):401–419, 2005.
- Hadeel Klakattawi, Veronica Vinciotti, and Keming Yu. A simple and adaptive dispersion regression model for count data. *Entropy*, 20(2):142, 2018.
- Zachary Kurtz, Christian Müller, Emily Miraldi, Dan Littman, Martin Blaser, and Richard Bonneau. Sparse and compositionally robust inference of microbial ecological networks. *PLoS Computational Biology*, 11(5):e1004226, 2015.
- Steffen Lauritzen. *Graphical Models*. Clarendon Press, 1996.
- Emmanuelle Le Chatelier, Trine Nielsen, Junjie Qin, Edi Prifti, Falk Hildebrand, et al. Richness of human gut microbiome correlates with metabolic markers. *Nature*, 500(7464):541–546, 2013.
- Karla Lee, Andrew Thomas, Laura Bolte, Johannes Björk, Laura Kist de Ruijter, et al. Cross-cohort gut microbiome associations with immune checkpoint inhibitor response in advanced melanoma. *Nature Medicine*, 28:535–544, 2022.
- Kyu Lee, Brent Coull, Anna-Barbara Moscicki, Bruce Paster, and Jacqueline Starr. Bayesian variable selection for multivariate zero-inflated models: Application to microbiome count data. *Biostatistics*, 21(3):499–517, 2020.
- Alex Lenkoski. A direct sampler for g-wishart variates. *Stat*, 2(1):119–128, 2013.
- Han Liu, John Lafferty, and Larry Wasserman. The nonparanormal: Semiparametric estimation of high dimensional undirected graphs. *Journal of Machine Learning Research*, 10(80):2295–2328, 2009.
- Han Liu, Kathryn Roeder, and Larry Wasserman. Stability approach to regularization selection (StARS) for high dimensional graphical models. *Neural Information Processing Systems*, 24:1432–1440, 2010.
- Han Liu, Fang Han, Ming Yuan, John Lafferty, and Larry Wasserman. High-dimensional semiparametric Gaussian copula graphical models. *The Annals of Statistics*, 40(4):2293–2326, 2012.
- Fabien Magne, Martin Gotteland, Lea Gauthier, Alejandra Zazueta, Susana Pesoa, Paola Navarrete, and Ramadass Balamurugan. The firmicutes/bacteroidetes ratio: a relevant marker of gut dysbiosis in obese patients? *Nutrients*, 12(5):1474, 2020.
- Reza Mohammadi and Ernst Wit. Bayesian structure learning in sparse Gaussian graphical models. *Bayesian Analysis*, 10(1):109–138, 2015.



- Reza Mohammadi and Ernst Wit. BDgraph: An R package for Bayesian structure learning in graphical models. *Journal of Statistical Software*, 89(3):1–30, 2019.
- Reza Mohammadi, Fentaw Abegaz, Edwin van den Heuvel, and Ernst Wit. Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 66(3):629–645, 2017.
- Reza Mohammadi, H el ene Massam, and G erard Letac. Accelerating Bayesian structure learning in sparse Gaussian graphical models. *Journal of the American Statistical Association*, 118(542):1345–1358, 2023.
- Jared Murray, David Dunson, Lawrence Carin, and Joseph Lucas. Bayesian Gaussian copula factor models for mixed data. *Journal of the American Statistical Association*, 108(502):656–665, 2013.
- L Pardo. *Statistical Inference Based on Divergence Measures*. Chapman and Hall/CRC, 2018.
- Helle Krogh Pedersen, Valborg Gudmundsdottir, Henrik Bj orn Nielsen, Tuulia Hyotylainen, Trine Nielsen, et al. Human gut microbes impact host serum metabolome and insulin sensitivity. *Nature*, 535(7612):376–381, 2016.
- Alina Peluso, Veronica Vinciotti, and Keming Yu. Discrete Weibull generalized additive model: an application to count fertility data. *Journal of the Royal Statistical Society: Series C*, 68(3):565–583, 2019.
- Junjie Qin, Ruiqiang Li, Jeroen Raes, Manimozhiyan Arumugam, Kristoffer Solvsten Burgdorf, Chaysavanh Manichanh, Trine Nielsen, Nicolas Pons, Florence Levenez, Takuji Yamada, et al. A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, 464(7285):59–65, 2010.
- Armin Rashidi, Maryam Ebadi, Daniel Weisdorf, Massimo Costalonga, and Christopher Staley. No evidence for colonization of oral bacteria in the distal gut in healthy adults. *Proceedings of the National Academy of Sciences*, 118(42):e2114152118, 2021.
- Alberto Roverato. Hyper inverse Wishart distribution for non-decomposable graphs and its application to Bayesian inference for Gaussian graphical models. *Scandinavian Journal of Statistics*, 29(3):391–411, 2002.
- Arkaprava Roy and David Dunson. Nonparametric graphical model for counts. *Journal of Machine Learning Research*, 21(229):1–21, 2020.
- Abe Sklar. Fonctions de r epartition  a n dimensions et leurs marges. *Publications de l’Institut de Statistique de l’Universit e de Paris*, 8:229–231, 1959.
- Veronica Vinciotti, Ernst Wit, Rick Jansen, Eco Geus, B. Penninx, Dorret Boomsma, and Peter Hoen. Consistency of biological networks inferred from microarray and sequencing data. *BMC Bioinformatics*, 17, 2016. doi: 10.1186/s12859-016-1136-0.

Veronica Vinciotti, Luca De Benedictis, and Ernst Wit. Cultural data integration via random graphical modelling. *arXiv preprint arXiv:2311.14367*, 2023.

Lu Yang, Edward Frees, and Zhengjun Zhang. Nonparametric estimation of copula regression models with discrete outcomes. *Journal of the American Statistical Association*, 115(530):707–720, 2020.