Zhou *et al. BioData Mining* (2020) 13:16
https://doi.org/10.1186/s13040-020-00226-7

BioData Mining

METHODOLOGY

Open Access

Identification of microbial interaction network: zero-inflated latent Ising model based approach

Jie Zhou[1], Weston D. Viles[3], Boran Lu[1], Zhigang Li[4], Juliette C. Madan[2], Margaret R. Karagas[2], Jiang Gui[1] and Anne G. Hoen[1,2*]

*Correspondence:
Anne.G.Hoen@dartmouth.edu
[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
[2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA
Full list of author information is available at the end of the article

Abstract

Background: Throughout their lifespans, humans continually interact with the microbial world, including those organisms which live in and on the human body. Research in this domain has revealed the extensive links between the human-associated microbiota and health. In particular, the microbiota of the human gut plays essential roles in digestion, nutrient metabolism, immune maturation and homeostasis, neurological signaling, and endocrine regulation. Microbial interaction networks are frequently estimated from data and are an indispensable tool for representing and understanding the conditional correlation between the microbes. In this high-dimensional setting, zero-inflation and unit-sum constraint for relative abundance data pose challenges to the reliable estimation of microbial interaction networks.

Methods and Results: To identify the microbial interaction network, the *zero-inflated latent Ising* (ZILI) model is proposed which assumes the distribution of relative abundance relies only on finite latent states and provides a novel way to solve issues induced by the unit-sum and zero-inflation constrains. A two-step algorithm is proposed for the model selection of ZILI. ZILI is evaluated through simulated data and subsequently applied to an infant gut microbiota dataset from New Hampshire Birth Cohort Study. The results are compared with results from Gaussian graphical model (GGM) and dichotomous Ising model (DIS). Providing ZILI is the true data-generating model, the simulation studies show that the two-step algorithm can identify the graphical structure effectively and is robust to a range of parameter settings. For the infant gut microbiota dataset, the final estimated networks from GGM and ZILI turn out to have significant overlap in which the ZILI tends to select the sparser network than those from GGM. From the shared subnetwork, a hub taxon Lachnospiraceae is

(Continued on next page)

(Continued from previous page)
identified whose involvement in human disease development has been discovered recently in literature.

Conclusions: Constrains induced by relative abundance of microbiota such as zero inflation and unit sum render the conditional correlation analysis unreliable for conventional methods such as GGM. The proposed optimal categoricalization based ZILI model provides an alternative yet elegant way to deal with these difficulties. The results from ZILI have reasonable biological interpretation. This model can also be used to study the microbial interaction in other body parts.

Keywords: Gut microbiota, Microbial interaction network, Latent Ising model, Dynamic programming, High-dimensional data, Sparse estimation

Introduction

The human microbiome, the collection of trillions of microbial organisms that live in our body spaces, belong to one of thousands of different species [1, 2]. The organisms that inhabit the human gut are an additional source of genetic diversity that can influence metabolism and modulate drug interactions [3]. Recent advances in genomic technologies enable production of thousands of 16S rRNA sequences per sample [4] and are powerful tools to explore the basic biology about human microbiome. Nevertheless, analyzing microbiome data and converting them into meaningful biological insights are still challenging tasks. First, the observed absolute abundance in sequencing experiment cannot inform the real absolute abundance of molecules in the sample which can be attributed to the sequence depth associated with the experiment. Multiple normalization methods have been proposed in literature to solve this problem among which total sum scaling (TSS) has been widely used in practice [5–9]. TSS scales each sample by the total read count and yields the relative abundance. However, the statistical analysis based on relative abundance can easily lead to spurious association due to the unit-sum constraint [10–14]. Further complicating the analysis of microbiome data is the zero-inflated distribution of read count [3]. As for the dataset in "Restults from the relative abundance of gut microbiota" section, among the 134 taxa, there are only 6 taxa for which the proportions of nonzero observations are greater than 80%. Zero inflation stems from the fact that the majority of the amplicon sequence variants (ASVs) either physically do not exist in the subject or are below the detection threshold for the given sample [2]. Another hurdle for analyzing the microbiome data is its high-dimensionality which usually involves hundreds of microbes; consequently, models equipped for this modeling task should be employed.

Microbial interaction network (MIN) is an indispensable tool for representing and understanding the relationships among the microbes [1, 15–17]. Traditionally, the interactions among the microbes are discovered through co-culture experiments which routinely involve only small number of species in an artificial community [18, 19]. Modern researches try to use the data from real environments such as human gut to infer the association among the microbes [20–23]. The corresponding statistical inferences of MIN based on these observational data have received much attention in recent years; however, the roadblocks mentioned above hinder the effective inference of MIN. As a compromise, most of the existing studies infer the MIN under the oversimplified assumptions [24, 25]. Especially, [24] ignores the unit-sum constraint and only considers the microbes for which the proportions of nonzero observation are higher than a given threshold; while in order to deal with the problem of zero inflation, [25] pools all the sparse taxa together and forms a composite taxon which is no longer sparse.

In light of the difficulties in MIN inference, in this paper we propose the zero-inflated latent Ising model (ZILI) for MIN aiming to address the roadblocks for analyzing the relative abundance of microbiome, i.e., (1) unit-sum constraint; (2) zero inflation; (3) high dimensionality. Latent models such as hidden Markov models [26], state space models [27] et al. have been widely used in economics, engineering and biology among many others. Despite their popularity across disciplines, latent models have not been investigated for microbiome data yet. Incidentally, [28] finds that the microbiota in human vagina could be characterized by finite states which provided a simple and intuitive understanding about the MIN in vagina. Inspired by the work in [28], in ZILI we assume that each of the p microbes in microbiota can be characterized by a latent discrete random variable $Z_j(1 \leq j \leq p)$. While for the random vector $\mathbf{Z} = \left(Z_1, \cdots, Z_p\right)$, the multiclass Ising model is employed to characterize the joint distribution of $\mathbf{Z}$. The relative abundances for each microbe are assumed to come from a zero-inflated mixture distribution which depends on the realization of $\mathbf{Z}$. Under this modeling framework, we propose a two-step algorithm for the model selection of ZILI. Specifically, in first step we estimate the states for each component of $\mathbf{Z}$ by transforming the relative abundances into categorical data. This step is implemented by an efficient dynamic programming algorithm. Based on the estimated state, in second step we use L_1-penalized group logistic regression to select the nonzero parameters involved in ZILI. In this way, the difficult issues, such as the unti-sum constraint and zero inflation, will not be the concerns for the data analysis. However, the cost for such simplication is the possible information loss brought by the categorization of relative abundance. Through simulated data, we investigate the performance of two-step algorithm and demonstrate its superiority over traditional Gaussian graphical model (GGM) and dichotomous Ising models (DIS) given that ZILI is the underlying data-generating model. We then apply both ZILI and GGM to an infant gut microbiome dataset from the New Hampshire Birth Cohort Study. It turns out the networks estimated by ZILI and GGM share a statistically significant part and ZILI shows the tendency to select sparser network than GGM. Within the shared subnetwork, Lachnospiraceae is identified as the hub taxon. On the other hand, recent researches have found that Lachnospiraceae widely exists in human gut [29] and is related to some severe diseases such as non-alcoholic fatty liver disease and inflammatory bowel diseases et al [30, 31]. Since this important taxon is identified by both models, this indicates that both ZILI and GGM can explain part of the information encoded in the relative abundance and the ZILI model can serve as a competitive tool for the MIN selection.

The organization of this paper is as follows. In “Zero-inflated latent Ising model for MIN” section, the ZILI model is detailed. The related estimation procedures for ZILI are described in “MIN selection based on ZILI” section. Simulation studies are carried out in “Results from the simulated data” section. “Restults from the relative abundance of gut microbiota” section is devoted to compare ZILI and GGM through gut microbiome dataset. “Discussion” section concludes with a brief review about ZILI model.

Method

Zero-inflated latent Ising model for MIN

In this section, we introduce the zero-inflated latent Ising (ZILI) model for the microbial interaction network which provides an alternative way to handle the problem of unit-sum constraint and zero inflation. Suppose that there are p taxa in the microbiota of interest. For jth taxon ($j = 1, \cdots, p$), let Z_j denote its latent state variable which has the following multinomial distribution,

$$P(Z_j = k) = p_{jk} \tag{1}$$

for $k = 0, 1, \cdots, K_j - 1$ with $\sum_{k=0}^{K_j-1} p_{jk} = 1$, where K_j represents the number of the latent states for jth microbe ($1 \leq j \leq p$). For example, there may be three states for Z_j corresponding to three different states of relative abundance, (high, medium, low). This assumption can be partly justified by the existing findings in literature [28]. The studies in [28] found that the composition of vaginal bacterial communities can be characterized by five states. The microbiota for a given subject can be classified into one of these five states. The state may be affected by the exogenous factors such as sexual activity, menstruation et al. In order to study the general relationship among the microbiota, Eq. (1) generalizes the results in [28] and assumes there are finite states for each microbe. For ease of exposition, in the following we assume that all Z_j's are K-level variables. The arguments can be generalized to the more general situation straightforwardly for which K_j may differ for different microbes. We pool all the Z_j's together and form the vector $\mathbf{Z} = (Z_1, \cdots, Z_p)$ for which multiclass Ising model is employed to characterize its joint distribution,

$$P(\mathbf{z}) = c \exp\left\{ \sum_{s=1}^{p} \phi_s(z_s) + \sum_{s=1}^{p} \sum_{t=1}^{p} \phi_{st}(z_s, z_t) \right\}, \tag{2}$$

where ϕ_s and ϕ_{st} are the potential functions associated with sth and tth microbes respectively. It should be noted that conventionally the variables in Ising model are dichotomous. The general cases of multiple values are usually referred as Potts model in literature [32]. However, in order to keep the notation consistent with recent studies, e.g., [33], with a little abuse of notation, we use the multiclass Ising model to refer to the Potts model. Our aim is to estimate the conditional relationship among Z_j's these potential functions can be parameterized as follows. For each $1 \leq s \leq p$, and $l \in \{0, \cdots, K-1\}$, define $I\left[z_s = l\right] = 1$ if $z_s = l$ and 0 otherwise. Then we have

$$\phi_s(z_s) = \sum_{l \in \mathcal{A}} \theta_{s;l} I[z_s = l] \tag{3}$$

for $s \in \left\{1, 2, \cdots, p\right\}$ and $\mathcal{A} = \{1, \cdots, K-1\}$ while

$$\phi_{st}\left(z_s, z_t\right) = \sum_{(l,h) \in \mathcal{B}} \theta_{st;lh} I\left[z_s = l; z_t = h\right] \tag{4}$$

for $(s, t) \in \{1, \cdots, p\}^2$ and $\mathcal{B} = \mathcal{A} \times \mathcal{A}$. The unknown parameters in (3)-(4) include $\boldsymbol{\theta} = \left\{\theta_{j;l}, \theta_{jt;lh} j = 1, \cdots, p, t = 1, \cdots, p, j \neq t, l = 1, \cdots, K-1, h = 1, \cdots, K-1\right\}$.

Based on (2)-(4), for $1 \leq i \leq n$, $1 \leq j \leq p$, we have the following equation hold [33, 34],

$$p_{jl} \triangleq \operatorname{logit}\left(P\left[Z_{ij} = l | \mathbf{Z}_{i(-j)} = \mathbf{z}_{i(-j)}\right]\right) = \theta_{j;l} + \sum_{t \neq j} \sum_{h=1}^{K-1} \theta_{jt;lh} I[z_{it} = h], \tag{5}$$

where $\mathbf{Z}_{i(-j)} = \left(Z_{i1}, \cdots, Z_{i(j-1)}, Z_{i(j+1)}, \cdots, Z_{ip}\right)^T$ with Z_{ij} the ith observation of Z_j. From (5), it can be shown that $\theta_{j;l}$ is the log odds for event $Z_j = l$ given that the other Z_t's, $t \neq j$ are all zero. Similarly, $\theta_{jt;lh}$ is the log-odds ratio describing the association between events $Z_j = l$ and $Z_t = h$ given that all the other components of $\mathbf{Z}$ are fixed to zero. For more details about the interpretation of these quantities, see [33] and references there. Let $\boldsymbol{\theta}_{jt} = \left(\theta_{jt;11}, \cdots, \theta_{jt;1(K-1)}, \theta_{jt;(K-1)1}, \cdots, \theta_{jt;(K-1)(K-1)}\right)^T$. Vector $\boldsymbol{\theta}_{jt}$ reflects the relationship between Z_j and Z_t. If all the components of $\boldsymbol{\theta}_{jt}$ are zero, Z_j and Z_t turn out to be independent. If there exist nonzero components in $\boldsymbol{\theta}_{jt}$, then Z_j and Z_t are related. In other words, there is an edge connecting microbe j and t in the microbial interaction network.

We have assumed that the relationship among microbes can be characterized by the multiclass Ising model (2)-(4). The state variables Z_j's in Ising model, however, are latent and can not be observed directly. Instead, the observable quantities are the relative abundances of the microbes which ae denoted by X_j's here. For each X_j, we assume its distribution can be characterized by a mixture distribution which relies on the realization of $\mathbf{Z}$. Specifically, we have the following conditional distribution for X_j given $z_j = l$ for $1 \leq l \leq K-1$,

$$f\left(x_j | z_j = l\right) = f_{jl}(x_j), \tag{6}$$

where f_{jl} $(1 \leq l \leq K-1)$ can be any continuous distribution defined on $[0, 1]$. When z_j is in state zero, i.e., $l = 0$, we assume

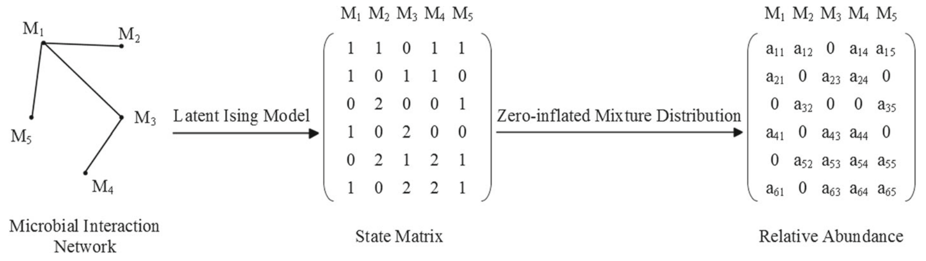
$$f_{j0}(x) = \begin{cases} \pi_j & \text{for } x = 0 \\ g_{j0}(x) & \text{otherwise} \end{cases}$$

for some $0 < \pi_j < 1$. Here g_{j0} can be any continuous distribution defined on $[0, 1]$. In other words, $f_{j0}(x)$ is a zero-inflated distribution. Let $\mu_{jl} = E\left(X_j | Z_j = l\right)$. For $l = 0$, μ_{jl} is understood as the expectation with respect to the density function g_{j0}. Note if there are states, $l \neq h$ such that $\mu_{jl} = \mu_{jh}$, then it is impossible to identify state l from h based on absolute or relative abundance data. In order to ensure the model identifiability, without loss of generalization, we assume,

$$\mu_{j0} < \mu_{j1} < \cdots < \mu_{j(K-1)} \tag{7}$$

for $1 \leq j \leq p$. Given $\mathbf{X} = \left(X_1, \cdots, X_p\right)$ and its n i.i.d observations, $\mathbf{X}_1, \cdots, \mathbf{X}_n$, we aim to estimate the MIN through (1)$\sim$(7) which we call zero-inflated latent Ising model (ZILI). The data-generating process of ZILI is depicted in Fig. 1.

Remarks 1 *(1) We have adopted a zero-inflated form for density function f_{j0} while continuous form for f_{jl} $(1 \leq l \leq K-1)$. In other words, the zero observations can only arise from*


M1
M2
M3
M4
M5
Latent Ising Model
Zero-inflated Mixture Distribution
Microbial Interaction Network
State Matrix
Relative Abundance

$$\begin{array}{ccccc} M_1 & M_2 & M_3 & M_4 & M_5 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 0 & 1 & 1 & 0 \\ 0 & 2 & 0 & 0 & 1 \\ 1 & 0 & 2 & 0 & 0 \\ 0 & 2 & 1 & 2 & 1 \\ 1 & 0 & 2 & 2 & 1 \end{array}$$

$$\begin{array}{ccccc} M_1 & M_2 & M_3 & M_4 & M_5 \\ a_{11} & a_{12} & 0 & a_{14} & a_{15} \\ a_{21} & 0 & a_{23} & a_{24} & 0 \\ 0 & a_{32} & 0 & 0 & a_{35} \\ a_{41} & 0 & a_{43} & a_{44} & 0 \\ 0 & a_{52} & a_{53} & a_{54} & a_{55} \\ a_{61} & 0 & a_{63} & a_{64} & a_{65} \end{array}$$

Fig. 1 Diagram of data-generating process in ZILI model

f_{j0} which has the smallest mean relative abundance among $f_{j0}, \cdots, f_{j(K-1)}$. This assumption serves to ensure the identifiability of ZILI model. In literature, the zero observations in microbiome data are usually classified into two categories by their nature [2, 6]. In first category, the zero means the corresponding microbe physically does not exist in the subject, or true zero. In second category, the microbe does exist in the subject; nevertheless, for this sample, this microbe happens not to exist or be below the threshold of the testing procedure, i.e., false zero. So our assumption about f_{jl} for $0 \leq l \leq K-1$ means that both true and false zero's can only come from f_{j0}. Though there is possibility that there are zeros' that do come from $f_{jl}(l \neq 0)$, we assume such probability is negligible compared with the former case, which we believe is a reasonable simplication to the real situations. (2) Conventional latent models, e.g state space model, typically assume the observed variables can be represented by a small number of latent variables and in this way the model dimensionality can be reduced. In ZILI, however, the latent variables are for the observations instead of the observed variables. Similar ideas have been used in factor analysis with the name R-type or Q-type factor analysis respectively [35].

MIN selection based on ZILI

From Eq. (5), it can be seen that the selection of MIN is equivalent to the selection of the nonzero components of $\boldsymbol{\theta}$ involved in ZILI model. In this section, we propose a two-step algorithm to select such nonzero components of $\boldsymbol{\theta}$ based on $\mathbf{X}_1, \cdots, \mathbf{X}_n$, the observations of relative abundance.

Step 1: state estimation

In this step, for each microbe, we aim to estimate the state Z_j $(1 \leq j \leq p)$ for each observation. For any given microbe, the proposed algorithm only involves its own observations. So for ease of exposition, we suppress the subscript j and use the generic notation (Z, X) to introduce the algorithm. The corresponding number of classes is denoted by $K_j = K$.

With the observations of relative abundance, $X_1, X_2, \cdots, X_n$, in hand, the estimation of Z is carried out through the following optimal classification of $X_1, X_2, \cdots, X_n$. Without loss of generality, we assume that the observations have been ordered, i.e., $X_1 \leq X_2 \leq \cdots \leq X_n$. For a given integer $k \geq 2$, let $b(n, K)$ denote a classification scheme which classifies $(X_1, \cdots, X_n)$ into k classes. Such classification can be depicted by the following notations,

$$\begin{aligned}
G_1 &= \left\{X_1, X_2, \cdots, X_{i_1}\right\}, \\
G_2 &= \left\{X_{i_1+1}, X_{i_1+2}, \cdots, X_{i_2}\right\}, \\
\cdots & \cdots \cdots\cdots\cdots\cdots, \\
G_k &= \left\{X_{i_{K-1}+1}, \cdots X_n\right\}.
\end{aligned} \tag{8}$$

With notation $i_0 = 1, i_k = n$, we define the following loss function for $b(n, K)$,

$$L[\,b(n, K)] = \sum_{h=0}^{K-1} D\left(i_h, i_{h+1}\right),$$

where

$$D\left(i_h, i_{h+1}\right) = \sum_{i=i_h}^{i_{h+1}} \left(X_i - m_h\right)^2, \tag{9}$$

$$m_h = \frac{1}{i_{h+1} - i_h + 1} \sum_{i=i_h}^{i_{h+1}} X_i.$$

We aim to find a classification scheme $b(n, K)$ which can minimize loss function $L[\,b(n, K)]$. Such optimal classification scheme is denoted by $p(n, K)$. It should be noted that other more complex forms of loss function $D(\cdot)$ are also possible, e.g., the absolute deviation based loss function, which is more robust for the data with outliers. However, given the popularity of squared loss function, we will focus on (9) and leave other possible forms for the future studies. We employ the following top-down dynamic programming algorithm to find $p(n, K)$ [36]. Specifically, the algorithm involves the following recursive procedures,

$$L[\,p(n, 2)] \; = \min_{2 \le i \le n} \{D(1, i-1) + D(i, n)\}, \tag{10}$$

$$L[\,p(n, K)] \; = \min_{K \le i \le n} \left\{L\left[p(i-1, K-1)\right] + D(i, n)\right\}. \tag{11}$$

Based on (10)-(11), for given K, the algorithm can be implemented as follows. First, find i_{K-1} such that

$$L[\,p(n, K)] = L\left[p(i_{K-1} - 1, K-1)\right] + D(i_{K-1}, n). \tag{12}$$

Based on i_{K-1}, denote the Kth class by $G_K = \{i_{K-1}, i_{K-1} + 1, \cdots, n\}$. In second step, find i_{K-2} such that

$$L\left[p\left(i_{K-1} - 1, K-1\right)\right] = L\left[p\left(i_{K-2} - 1, K-2\right)\right] + D\left(i_{K-2}, i_K - 1\right),$$

then we get the $(K-1)$th class $G_{K-1} = \{i_{K-2}, i_{K-2} + 1, \cdots, i_{K-1} - 1\}$. By the same fashion, all the classes $G_1, G_2, \cdots, G_K$ can be derived, which is the optimal solution $p(n, K)$. Based on $p(n, K)$, the estimate of Z for observations in class G_k is defined as $\hat{Z} = k - 1$ for $k = 1, \cdots, K$.

The algorithm above assumes that K, the number of the classes, is known as a priori. In practice, K usually is unknown and has to be determined based on the data. Though several methods have been proposed in literature, such as likelihood ratio test in R package *mixtools* [37], or BIC method in package *sBIC* [38], these methods have poor performances when the data are zero-inflated. Instead, we propose the following criterion to select K. For a given upper bound, say, $\bar{K}$, and each K with $2 \le K \le \bar{K}$, the minimum loss $L(p(n, K))$ is calculated. Define $d_K = L(p(n, K+1)) - L(p(n, K))$ for $K = 2, \cdots, \bar{K} - 1$ and let $\bar{d}$ be the mean of d_K's. Then the first K with $d_K \le \bar{d}$ will be selected as the class number. This criterion turns out to have a better performance than the methods mentioned above in the simulation studies in "Results from the simulated data" section. However, it should be noted, as suggested by one of the reviewers, that the choice of K may potentially have big impact on the final selected model. Consequently, in practice the robust way for the determination of K is to compare multiple methods, from which domain knowledge may be employed to choose the optimal one.

Step 2: network selection

Equation (5) shows that, after the logit transformation, the conditional probability p_{jl} defined in (5) is a linear function of $\boldsymbol{\theta}$. Here the covariates are the indicator functions of events $\{Z_{it} = h\}$ $\left(t \neq j, h = 1, \cdots, K-1\right)$. Based on this observation, the neighborhood method is proposed in [33] to select the nonzero components in $\boldsymbol{\theta}$ for dichotomous Ising model. In this paper the same method will be employed for the MIN selection. Specifically, for jth microbe, let $\boldsymbol{\theta}_j = \left(\boldsymbol{\theta}_{j1}^T, \cdots, \boldsymbol{\theta}_{j(j-1)}^T, \boldsymbol{\theta}_{j(j+1)}^T, \cdots, \boldsymbol{\theta}_{jp}^T\right)^T$ where $\boldsymbol{\theta}_{jt}$ is defined in "Method" section. Based on the Eq. (5), we consider the following penalized group logistic regression problem,

$$\hat{\boldsymbol{\theta}}_j = \arg\min_{\boldsymbol{\theta}_j} \left\{ -l\left(\boldsymbol{\theta}_j | \hat{\mathbf{Z}}_{(-j)}\right) + \lambda \sum_{t \neq j} \sqrt{m_{jt}} \|\boldsymbol{\theta}_{jt}\|_2 \right\}, \tag{13}$$

where $l(\boldsymbol{\theta}_j | \hat{\mathbf{Z}}_{(-j)}) = \sum_{i=1}^{n} \left\{ I(Z_{ij} = l) \log p_{jl} + (1 - I(Z_{ij} = l)) \log(1 - p_{jl}) \right\}$ with p_{jl} being defined in eq. (5), λ is the tuning parameter, m_{jt} is the length of vector $\boldsymbol{\theta}_{jt}$ and $\| \cdot \|_2$ is the Euclidean norm. The form of $\sqrt{m_{jt}}$ aims to account for the varying group size of $\boldsymbol{\theta}_{jt}$ [39]. Such form of penalty in (13) tends to shrink the components in same group $\boldsymbol{\theta}_{jt}$ to zero simultaneously. For given λ, the coordinate decent algorithm [40, 41] is employed to solve (13). As for the selection of λ, extended BIC proposed in [42] is adopted which favors sparser model compared with the standard BIC. The minimization problem in (13) is solved for each Z_j $(1 \leq j \leq p)$. With the final estimate $\hat{\boldsymbol{\theta}}$ in hand, we define an edge between Z_j and Z_t if there exists at least one nonzero component in either $\hat{\boldsymbol{\theta}}_{jt}$ or $\hat{\boldsymbol{\theta}}_{tj}$. An alternative way to define an edge requires there exists at least one nonzero component in both $\hat{\boldsymbol{\theta}}_{jt}$ and $\hat{\boldsymbol{\theta}}_{tj}$. It turns out these two strategies are asymptotically equivalent [33, 43] and so we just employ the former one to select the MIN in the numerical studies. The magnitude of the components of $\hat{\boldsymbol{\theta}}_{jt}$ plays no role in the determination of the edges [33, 44].

In the above, the proposed algorithm estimates the interaction network by separately solving p conditional penalized maximum likelihood estimation problems. Alternatively, we can form a joint conditional likelihood function for $\boldsymbol{\theta}$ and estimate the network through the penalized version of the joint conditional likelihood function; however, this approach is not computationally as stable as (13) [44]. We therefore put the focus on the individual regression method (13). Figure 2 illustrates the workflow of the two-step algorithm through a toy MIN.

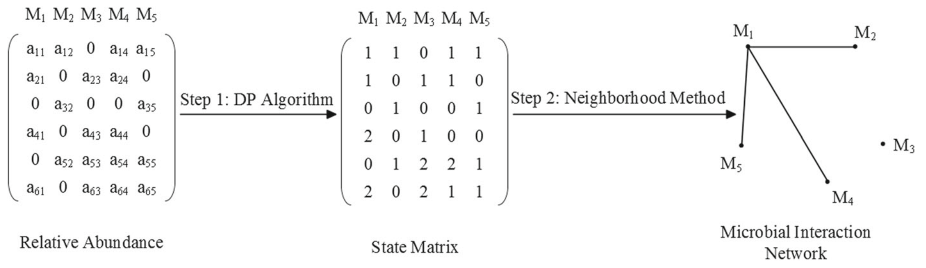


Fig. 2 Procedures of two-step algorithm for ZILI model

Remarks 2 *For the two-step algorithm proposed above, it is expected that the selection of MIN will be improved if we can improve the state estimates $\hat{Z}_{ij}$'s; however, the misclassification is inevitable in two-step algorithm which will adversely impact the final network selection. In "Results from the simulated data" section, we investigate how the misclassification impacts the MIN selection through simulation studies.*

Results

Results from the simulated data

In this section, we investigate the performance of the two-step algorithm when ZILI is the underlying data-generating model. As a comparison, the popular Gaussian graphical model (GGM) and dichotomous Ising model (DIS) will also be fitted using the same dataset. Here DIS is constructed by transforming the relative abundance into 0 or 1 according to whether it is less than the median. The same algorithm in "Step 2: network selection" section will be employed to estimate the structure of this dichotomous Ising model.

Specifically, assume that there are p microbes with state variables $\mathbf{Z} = \left(Z_1, \cdots, Z_p\right)$. Each realization of Z_j $(j = 1, \cdots, p)$ takes value from the set $\{0, 1, 2\}$. The conditional distribution of Z_j $(j \neq 1)$ given all the other components of $\mathbf{Z}$ only depends on microbe Z_{j-1}. As for microbe 1, the distribution of Z_1 depends on microbe Z_p. For such a model, the nonzero parameters involved in Eq. (5) include $\left(\theta_{j;1}, \theta_{j;2}, \theta_{j(j-1);11}, \theta_{j(j-1);12}, \theta_{j(j-1);21}, \theta_{j(j-1);22}\right)$ which are assumed to be same for all j's. For each replication, these parameters are sampled from the multivariate normal distribution $N_6(\mu, \Sigma)$ with $\mu = (-1, 3, -0.8, 2, -3, -4)^T$ and $\Sigma = \text{diag}\left(0.1^2, 0.3^2, 0.08^2, 0.2^2, 0.3^2, 0.4^2\right)$.

Given the Ising model above, the Gibbs sampler is employed to generate the samples of $\mathbf{Z}$. Specifically, first a p-dimensional vector is generated where the states for each Z_j are independently sampled from the set $\{0, 1, 2\}$ with equal probability 1/3. Then given all Z_t, $(t \neq j)$, the state of Z_j is updated based on Eq. (5). By the same fashion, the states of all the other Z_j can be updated recursively. We run this process 200 times and the final state of $\mathbf{Z}$ will be deemed a qualified representative of the underlying Ising model. Based on the samples of $\mathbf{Z}$, the samples of absolute abundance $\mathbf{X} = \left(X_1, \cdots, X_p\right)$ are generated according to $X_j|Z_j = z \sim \text{N}\left(\mu_z, \sigma^2\right)$ with $\mu_0 = 10, \mu_1 = 15, \mu_2 = 20$ and a given σ^2. Pooling all the samples of $\mathbf{X}$ together leaves us a $n \times p$ matrix which represents n absolute abundance observations for p microbes. For each column, the absolute abundances which are less than a given percentile with rank u are replaced by zero. Here u is sampled from uniform distribution $U[0, \tau]$ for a given $0 < \tau < 1$. For each row in this zero-inflated matrix, we then transform the absolute abundances to relative abundances by dividing each entry by the corresponding row sum. Figure 2 shows the diagram for the data-generating process.

To compare the performances of different models, two criteria, true positive rate (TPR) and false positive rate (FPR) will be used which are defined respectively as,

$$\text{TPR} = \frac{\#\{\text{identified true edges}\}}{\#\{\text{all true edges}\}}, \tag{14}$$

$$\text{FPR} = \frac{\#\{\text{falsely identified edges}\}}{\#\{\text{all none edges}\}}. \tag{15}$$

An ideal algorithm should have a relatively high TPR and low FPR. There are multiple factors that can influence the performance of the algorithm, which include the variance

σ^2, the sample size n, and the zero proportion z. For three choices σ, two choices of n and three choices of τ, Table 1 lists the results of TPR and FPR for ZILI, DIS and GGM respectively. Here the number of the microbes is set to be $p = 60$ and the number of replication is 100. Note for GGM, there are different estimation methods available such as graphical lasso [45], or neighborhood method [43] et al. Here in order to facilitate the comparison with ZILI and DIS, we adopt the neighborhood method of [43]. The same model selection criterion extended BIC is used in all cases. It can be seen from Table 1 that for all the cases considered, the proposed two-step algorithm does can select the network structure effectively while both GGM and DIS have low TPR and can not properly select the true edges. On the other hand, all the three factors considered, i.e., variance, sample size and zero proportion have significant impact on the performances of two-step algorithm. Two-step algorithm has the best performance with the small σ^2, τ and large n which is in accordance with our expectation. In particular, a large σ^2 will lead to a high misclassification rate for the state estimation which in turn results in a poor network selection, i.e., low TPR and high FPR.

Restults from the relative abundance of gut microbiota

In this section, ZILI is employed to investigate the conditional association among the microbes in the infant stool samples from New Hampshire Birth Cohort Study (NHBCS), a cohort of mother-infant pairs in New Hampshire. For this dataset, stool samples were collected from infants at six weeks and twelve months of age, who were followed in the NHBCS. The stool samples were characterized by 16S rRNA sequencing. The R software package *DADA21* was used to infer the abundance of amplicon sequence variants in each sequenced sample [46]. Taxonomy at the family level was obtained by classifying the sequences against the reference training dataset from the GreenGenes Database

Table 1 Comparison of ZILI, DIS and GGM based on simulated data

			ZILI		**DIS**		**GGM**	
τ	σ	n	TPR	FPR	TPR	FPR	TPR	FPR
10	0.5	60	0.8322	0.0110	0.0115	0.0008	0.1940	0.0347
		120	0.9650	0.0024	0.0173	0.0006	0.0721	0.0058
	1	60	0.7945	0.0123	0.0106	0.0007	0.0145	0.0059
		120	0.9615	0.0043	0.0163	0.0005	0.1683	0.0051
	2	60	0.2688	0.0256	0.0115	0.0010	0.0123	0.0061
		120	0.5041	0.0187	0.0096	0.0003	0.0368	0.0046
40	0.5	60	0.7775	0.0134	0.0106	0.0009	0.1961	0.0274
		120	0.9445	0.0059	0.0180	0.0005	0.1923	0.0056
	1	60	0.7260	0.0139	0.0163	0.0007	0.1895	0.0276
		120	0.9353	0.0067	0.0120	0.0003	0.1821	0.0057
	2	60	0.2085	0.0247	0.016	0.0013	0.1328	0.030
		120	0.4043	0.0194	0.0115	0.0004	0.0991	0.0055
80	0.5	60	0.4443	0.0217	0.0720	0.0086	0.2346	0.0465
		120	0.6090	0.0148	0.1115	0.0071	0.2248	0.0144
	1	60	0.4088	0.0214	0.0681	0.0085	0.2321	0.0477
		120	0.6020	0.0149	0.1138	0.0071	0.2196	0.0146
	2	60	0.1538	0.0247	0.0493	0.0084	0.1851	0.0514
		120	0.2820	0.0207	0.0938	0.0065	0.1620	0.0142

Consortium (Version 13.8). There were 398 six-week and 316 twelve-month samples with varying abundances across 134 taxonomic families.

For each taxon, if the proportion of nonzero observations is less than 1%, then the number of classes is set to be $K_j = 2$ and the observations are classified according to whether it is zero or not. Otherwise, the upper bound of K_j is set to be $\bar{K} = 6$. Then we follow the two-step algorithm to select the network. In order to gain insights from the difference between ZILI and GGM, the networks based on GGM have also been selected using the neighborhood method. In light of the severe zero inflation in the dataset, it is inappropriate to assume the GGM for the whole dataset. To alleviate the problem of zero inflation, we choose to use the subsets of this dataset to construct the GGM networks. Specifically, for each $s = 10\%, 20\%, \cdots, 80\%$, we extract the corresponding subset from the original dataset which only includes the microbes whose proportions of nonzero observations are greater than s. For each of these subsets, GGM is fitted using the neighborhood method. The ZILI network involves 134 microbe taxa while the eight GGM networks only involves eight subsets of these 134 taxa. So in order to compare the ZILI network with the eight GGM networks, we extract the subnetworks from ZILI networks for each s. For each of the extracted network, we then compare it with the corresponding GGM network in terms of their connectivity and the results are listed in Table 2.

In Table 2, each row corresponds to a pair of ZILI and GGM networks. For two microbes, (0,0) represents there is no edge connecting them in both ZILI and GGM network; (0,1) represents there is an edge in ZILI network while no edge in GGM network; (1,0) represents there is an edge in GGM network while no edge in Ising network; (1,1) represents there is an edge connecting them in both ZILI and GGM network. The columns 3-6 in Table 2 list the numbers of the edges falling into these four categories respectively. The relationship of ZILI and GGM is our primary interest. To this end, the χ^2 test for the independence of ZILI and GGM is carried out and the corresponding adjusted p-value's are listed in the last column of Table 2. Note the p-value here is based on the estimated networks rather than the relative abundance. So we call them conditional p-value. These p-value's suggest that the networks of ZILI and GGM are closely related, even though ZILI and GGM are based on entirely different assumptions about how the data are generated. A more detailed inspection reveals that most of the edges selected by ZILI are also selected by GGM and GGM selects far more edges than ZILI. In other words, ZILI is more conservative than GGM in terms of edge selection.

Table 2 Comparison of microbial interaction networks selected by GGM and ZILI

		ZILI				
		(0,0)	**(1,0)**	**(0,1)**	**(1,1)**	***p*-value**
GGM	10%	589	58	6	13	0.0000
	20%	357	37	3	9	0.0000
	30%	182	38	2	9	0.0000
	40%	137	25	2	7	0.0000
	50%	89	23	2	6	0.0022
	60%	55	17	0	6	0.0005
	70%	46	15	0	5	0.0252
	80%	19	6	0	3	0.0445

The data are the relative abundances of microbiota in infant gut from NHBCS

The ZILI network and all the GGM networks corresponding to the threshold $s = 10\%, 20\%, \cdots, 80\%$ are available in the supplementary materials. Figure 3 presents the subnetwork that is shared by the ZILI networks and GGM network corresponding to $s = 10\%$. From Fig. 3, it can be seen that Lachnospiraceae is selected as hub taxon by both ZILI and GGM. It has been discovered in literature that R. gnavus, one of the members in Lachnospiraceae family, has high frequency in infant gut [29]. Lachnospiraceae has close connections with severe human diseases, such as inflammatory bowel diseases (IBD) [30], non-alcoholic fatty liver disease [31]. The R. gnavus ATCC 29149 strain possesses the complete Nan cluster involved in sialic acid metabolism for the production of an intramolecular trans-sialidase [47]. It has also been demonstrated recently that R. gnavus produces iso-bile acids. The iso-bile acids detoxification pathway influences the growth of one of the predominant genera in the human gut, i.e., the Bacteroides [48]. In summary, Lachnospiraceae plays an active role in human metabolism which in turn impacts the growth of the other taxa in the gut microbiota. In this respect, it is not surprising to find its wide connections with other members of the microbiota.

Discussion

The prosperous microbiome datasets have led us to a new level of biological researches. Nevertheless, how to gain scientific insight from these complex datasets through novel statistical methods remains a big challenge for researchers. In light of the difficulties in MIN selection, we propose a novel zero-inflated latent Ising model (ZILI) to this problem. In ZILI, the relative abundances of microbiota are assumed to follow a mixture distribution which relies on the realization of a latent Ising model. Through simulation studies, it is shown that under given scenarios, the proposed two-step algorithm for the inference

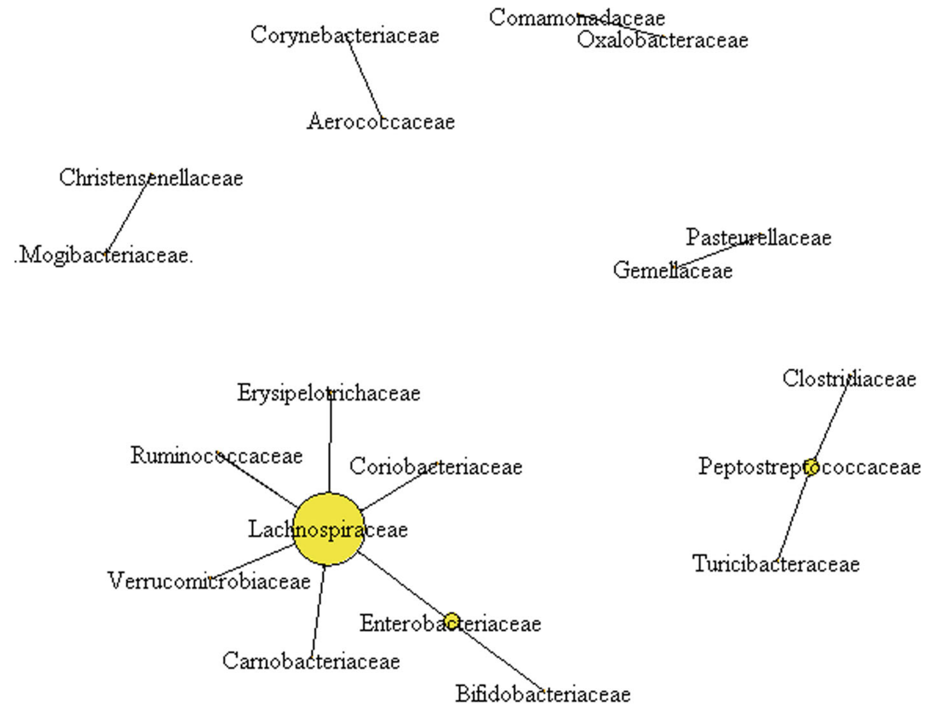


Fig. 3 The overlappted subnetwork in GGM network and ZILI nework

of ZILI can select the true network structure effectively while Gaussian graphical model and dichotomous Ising model have little power to recover the network structure. For a microbiome dataset from New Hampshire Birth Cohort Study, it is shown that ZILI is more conservative compared with Gaussian graphical model. Among the edges shared by these networks, a hub taxon is selected which has close connections with human metabolism. These findings indicate that ZILI can serve as an competitive model to estimate the microbial interaction network. On the other hand, we only consider the problem of model selection in this paper. In order to gain more insights about the conditional correlation beween microbes, quantitative characteristics like parameter estimates should be taken into consideration which will be studied in our future studies.

Supplementary information

Supplementary information accompanies this paper at https://doi.org/10.1186/s13040-020-00226-7.

Additional file 1: The file *Networks.pdf* includes the gut microbial interaction networks selected by ZILI model and Gaussian graphical model with thresholds, 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% respectively. These networks are used in Table 2 to investigate the relationship between ZILI and GGM.

Acknowledgements

We are grateful to the participants and staff of the New Hampshire Birth Cohort Study for their efforts in the collection and processing of microbiome data. This work was supported in part by US National Institutes of Health grants (R01LM012012; R01LM012723;P20GM104416; P01ES022832; P42ES007373; UG3/UH3OD023275), US Environmental Protection Agency grant (RD83544201).

Authors' contributions

Conceptualization: JG, AGH, JZ; Data curation: JCM, MRK; Formal analysis: JG, WDV, AGH, JZ; Funding acquisition: MRK, AGH, ZL; Investigation: JZ, JG, AGH, WDV; Methodology: JZ, JG; Project administration: AGH, ZL; Resources: MRK; Software: JZ; Supervision: JG, AGH; Validation: AGH, JG; Visualization: JZ; Writing: JZ, BL. The author(s) read and approved the final manuscript.

Availability of data and materials

The relative abundance datasets used for this work are available upon request from the corresponding author. The programs used to implement the algorithms in this paper is written with R langurage and can be freely downloaded at github website, https://github.com/hoenlab/Ising.

Competing interests

The authors declare that they have no competing interests.

Author details

[1]Department of Biomedical Data Science, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. [2]Department of Epidemiology, Geisel School of Medicine, Dartmouth College, Hanover, NH, USA. [3]Department of Mathematics and Statistics, University of Southern Maine, Portland, ME, USA. [4]Department of Biostatistics, University of Florida, Gainesville, FL, USA.

Received: 2 April 2020 Accepted: 22 September 2020
Published online: 07 October 2020

References

1. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012;10(8):538–50. https://doi.org/10.1038/nrmicro2832.
2. Li HZ. Microbiome, metagenomics, and high-dimensional compositional data analysis. Ann Rev Stat Appl. 2015;2: 73–94. https://doi.org/10.1146/annurev-statistics-010814-020351.
3. Ursell LK, Metcalf JL, Parfrey LW, Knight R. Defining the human microbiome. Nutr Rev. 2012;70(Suppl 1):38–44. https://doi.org/10.1111/j.1753-4887.2012.00493.x.
4. Ward D, Weller R, Bateson M. 16S rRNA sequences reveal numerous uncultured microorganisms in a natural community. Nature. 1990;345:63–5. https://doi.org/10.1038/345063a0.
5. Anders S, Huber W. Differential expression analysis for sequence count data. Genome Biol. 2010;11(10):R106. https://doi.org/10.1186/gb-2010-11-10-r106.
6. Chen L, Reeve J, Zhang L, Huang S, Wang X, Chen J. GMPR: A robust normalization method for zero-inflated count data with application to microbiome sequencing data. PeerJ. 2018;6:e4600. https://doi.org/10.7717/peerj.4600.
7. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are compositional: and this is not optional. Front Microbiol. 2017;8:2224. https://doi.org/10.3389/fmicb.2017.02224.
8. Lovén J, Orlando DA, Sigova AA, Lin CY, Rahl PB, Burge CB, Levens DL, Lee TI, Young RA. Revisiting global gene expression analysis. Cell. 2012;151(3):476–82.

9. Robinson MD, McCarthy DJ, Smyth GK. edgeR: a bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics. 2010;26(1):139–40. https://doi.org/10.1093/bioinformatics/btp616.
10. Aitchison J. The statistical analysis of compositional data. J R Stat Soc Ser B Methodol. 1982;44(2):139–60.
11. Lovell D, Pawlowsky-Glahn V, Egozcue JJ, Marguerat S, Bähler J. Proportionality: a valid alternative to correlation for relative data. PLoS Comput Biol. 2015;11(3):e1004075. https://doi.org/10.1371/journal.pcbi.1004075.
12. Mandal S, Van Treuren W, White RA, Eggesbø M, Knight R, Peddada SD. Analysis of composition of microbiomes: a novel method for studying microbial composition. Microb Ecol Health Dis. 2015;26(1):27663. https://doi.org/10.3402/mehd.v26.27663.
13. Morton JT, Sanders J, Quinn RA, McDonald D, Gonzalez A, Vazquez-Baeza Y, Navas-Molina JA, Song SJ, Metcalf JL, Hyde ER, Lladser M, Dorrestein PC, Knight R. Balance trees reveal microbial niche differentiation. MSystems. 2017;2(1):e0016216. https://doi.org/10.1128/msystems.00162-16.
14. Tsilimigras MC, Fodor AA. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. Ann Epidemiol. 2016;26(5):330–5. https://doi.org/10.1016/j.annepidem.2016.03.002.
15. Claesson MJ, Jeffery IB, Conde S, Power SE, O'connor EM, Cusack S, Harris HMB, Coakley M, Lakshminarayanan B, O'Sullivan O, et al. Gut microbiota composition correlates with diet and health in the elderly. Nature. 2012;488:178–84.
16. Claussen JC, Skiecevičienė J, Wang J, Rausch P, Karlsen TH, Lieb W, Baines JF, Franke A, Hütt MT. Boolean analysis reveals systematic interactions among low-abundance species in the human gut microbiome. PLoS Comput Biol. 2017;13:e1005361.
17. Friedman J, Alm E. Inferring correlation networks from genomic survey data. PLoS Comput Biol. 8:e1002687.
18. Gause GF. The Struggle for Existence. Baltimore: Williams & Wilkins; 1934.
19. Hsu RH, Clark RL, Tan JW, Ahn JC, Gupta S, Romero PA, Venturelli OS. Microbial interaction network inference in microfluidic droplets. Cell Syst. 2019;9(3):229–42. https://doi.org/10.1016/j.cels.2019.06.008.
20. Barberan A, Bates ST, Casamayor EO, Fierer N. Using network analysis to explore co-occurrence patterns in soil microbial communities. ISME J. 2012;6:343–51.
21. Berry D, Widder S. Deciphering microbial interactions and detecting keystone species with co-occurrence networks. Front Microbiol. 2014;5:219. https://doi.org/10.3389/fmicb.2014.00219.
22. Biswas S, McDonald M, Lundberg DS, Dangl JL, Jojic V. Learning microbial interaction networks from metagenomic count data. In: International Conference on Research in Computational Molecular Biology; 2015. p. 32–43.
23. Mitra K, Carvunis AR, Ramesh SK, Ideker T. Integrative approaches for finding modular structure in biological networks. Nat Rev Genet. 2013;14(10):719–32. https://doi.org/10.1038/nrg3552.
24. Chen I, Kelkar YD, Gu Y, Zhou J, Qiu X, Wu H. High-dimensional linear state space models for dynamic microbial interaction networks. PloS ONE. 2017;12(11):e0187822.
25. Marino S, Baxter NT, Huffnagle GB, Petrosino JF, Schloss PD. Mathematical modeling of primary succession of murine intestinal microbiota. Proc Natl Acad Sci. 2014;111(1):439–44.
26. Yoon BJ. Hidden Markov models and their applications in biological sequence analysis. Curr Genomics. 2009;10(6):402–15. https://doi.org/10.2174/138920209789177575.
27. Durbin J, Koopman SJ. Time Series Analysis by State Space Methods: Second Edition, 2nd Revised ed.: Oxford Statistical Science Series; 2009.
28. Gajer P, Brotman RM, Bai G, Sakamoto J, Schutte UM, Zhong X, Koenig SSK, Fu L, Ma ZS, Zhou X, et al. Temporal dynamics of the human vaginal microbiota. Sci Transl Med. 2012;4(132):132–52. https://doi.org/10.1126/scitranslmed.3003605PMID:22553250.
29. Sagheddu V, Patrone V, Miragoli F, Puglisi E, Morelli L. Infant early gut colonization by Lachnospiraceae: high frequency of Ruminococcus gnavus. Front Pediatr. 2016;4:57. https://doi.org/10.3389/fped.2016.00057.
30. Png CW, Lindén SK, Gilshenan KS, Zoetendal EG, McSweeney CS, Sly LI, McGuckin MA, Florin THJ. Mucolytic bacteria with increased prevalence in IBD mucosa augment in vitro utilization of mucin by other bacteria. Am J Gastroenterol. 2010;105:2420–8. https://doi.org/10.1038/ajg.2010.281.
31. Shen F, Zheng RD, Sun XQ, Ding WJ, Wang XY, Fan JG. Gut microbiota dysbiosis in patients with non-alcoholic fatty liver disease. Hepatobiliary Pancreat Dis Int. 2017;16(4):375–81. https://doi.org/10.1016/S1499-3872(17)60019-5. PMID: 28823367.
32. Potts RB. Some generalized order-disorder transformations. In: Mathematical Proceedings of the Cambridge Philosophical Society; 1952. p. 106–9, Cambridge University Press.
33. Ravikumar P, Wainwright MJ, Lafferty JD. High-dimensional Ising model selection using L_1 regularized logistic regression. Ann Stat. 2010;38:1287–319.
34. Wainwright MJ, Jordan MI. 2008. Graphical Models, Exponential Families, and Variational Inference, Foundations and Trends® in Machine Learning, Vol. 1. http://dx.doi.org/10.1561/2200000001.
35. Bennett S. An introduction to multivariate techniques for social and behavioural sciences. New York: Wiley; 1976.
36. Sniedovich M. Dynamic programming: Foundations and principles. New York: Taylor & Francis; 2010. ISBN 978-0-8247-4099-3.
37. Tatiana B, Didier C, David RH, Derek Y. mixtools: An R Package for analyzing finite mixture models. J Stat Softw. 2009;32(6):1–29.
38. Weihs L, Plummer M. Computing the singular BIC for multiple models. 2016. https://cran.rproject.org/web/packages/sBIC. R package, version 0.2.0.
39. Meier L, Geer S, Buhlmann P. The group lasso for logistic regression. J R Stat Soc Ser B Stat Methodol. 2008;70:53–71.
40. Friedman J, Hastie T, Tibshirani R. Regularization paths for generalized linear models via coordinate descent. J Stat Softw. 2010;33(1):1.
41. Fu WJ. Penalized regressions: the bridge versus the lasso. J Comput Graph Stat. 1998;7:397–416.
42. Chen J, Chen Z. Extended BIC for small-n-large-P sparse GLM. Stat Sin. 2012;22:555–74.
43. Meinshansen N, Buhlmann P. High dimensional graphs and variable selection with lasso. Ann Stat. 2006;34(3):1436?-62.
44. Cheng J, Levina E, Wang P, Zhu J. A sparse Ising model with covariates. Biometrics. 2014;70:943–53.

45. Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistcs. 2008;9: 432–41.
46. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Methods. 2016;13:581–3. https://doi.org/10.1038/nmeth.3869.
47. Tailford LE, Owen CD, Walshaw J, Crost EH, Hardy-Goddard J, Le Gall G, et al. Discovery of intramolecular trans-sialidases in human gut microbiota suggests novel mechanisms of mucosal adaptation. Nat Commun. 2015;6: 7624. https://doi.org/10.1038/ncomms8624.
48. Devlin AS, Fischbach MA. A biosynthetic pathway for a prominent class of microbiota-derived bile acids. Nat Chem Biol. 2015;11:685–90. https://doi.org/10.1038/nchembio.1864.

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.