

# Differential Markov random field analysis with an application to detecting differential microbial community networks

BY T. T. CAI

*Department of Statistics, University of Pennsylvania, 3720 Walnut Street, Philadelphia, Pennsylvania 19104, U.S.A.*

tcai@wharton.upenn.edu

H. LI, J. MA

*Department of Biostatistics, University of Pennsylvania, 423 Guardian Drive, Philadelphia, Pennsylvania 19104, U.S.A.*

hongzhe@upenn.edu jinma@upenn.edu

AND Y. XIA

*Department of Statistics, School of Management, Fudan University, Shanghai 200433, China*

xiayin@fudan.edu.cn

## SUMMARY

Micro-organisms such as bacteria form complex ecological community networks that can be greatly influenced by diet and other environmental factors. Differential analysis of microbial community structures aims to elucidate systematic changes during an adaptive response to changes in environment. In this paper, we propose a flexible Markov random field model for microbial network structure and introduce a hypothesis testing framework for detecting differences between networks, also known as differential network analysis. Our global test for differential networks is particularly powerful against sparse alternatives. In addition, we develop a multiple testing procedure with false discovery rate control to identify the structure of the differential network. The proposed method is applied to data from a gut microbiome study on U.K. twins to evaluate how age affects the microbial community network.

*Some key words:* Differential network; High-dimensional logistic regression; Microbiome; Multiple testing.

## 1. INTRODUCTION

### 1.1. Markov random field model for microbial networks

High-throughput sequencing technologies provide comprehensive surveys of the human microbiome using either 16S rRNA or shotgun metagenomics sequencing (Kuczynski et al., 2012). An important objective in microbiome studies is to infer the interactions among various microorganisms (Faust & Raes, 2012). There is growing interest in studying microbial community structures, because the underlying ecological structures are highly dynamic and undergo differential changes in response to changes in their environment. An example that motivates this paper is to understand how age-related physiological changes in the gut and modifications in lifestyle influence

gut microbial interaction structures (Biagi et al., 2010; Claesson et al., 2011). Motivated by this goal of uncovering changes in microbial interactions associated with age, we develop a method for detecting differential microbial community networks.

Microbiome data present at least two challenges. First, the observations in sequencing-based microbiome studies are often relative abundances of different bacteria, and the absolute microbial abundances are unavailable. The observation from each sample can be represented by a compositional vector  $R = (R_1, \dots, R_p)^T$  with  $p$  taxa and the unit-sum constraint  $\sum_{j=1}^p R_j = 1$  ( $R_j \geq 0$ ,  $j = 1, \dots, p$ ). Because typical microbial communities consist of both rare and common taxa, the second obstacle is the sparsity of the compositional vector  $R$ ; many taxa are absent from the sample or their abundances are below the detection level. There is a lack of flexible statistical models that can capture the complex dependency structures among the taxa in a given community. For instance, methods based on Gaussian graphical models cannot be directly applied to study the conditional dependence relationships, due to the unit-sum constraint, and naive application of such models can lead to spurious associations (Aitchison, 1982).

Gaussian graphical models have been applied to centred log-ratio transformed data to study conditional dependence relationships among microbes (Kurtz et al., 2015), but this transformation has difficulty in dealing with zeros and the resulting transformed data are not even close to being Gaussian or sub-Gaussian. Furthermore, graphical models based on the centred log-ratio transformation are difficult to interpret. Biswas et al. (2016) proposed studying conditional microbial interactions by modelling the residuals from a Poisson regression, but their approach does not account for the compositional nature of the data. Fang et al. (2017) modelled the conditional dependence among the latent absolute abundances that generate the compositional data using a logistic normal distribution, but this method does not address the sparsity issue. Finally, existing works based on graphical models do not assign uncertainties or statistical significance to the conditional associations, and are not developed with a focus on differential network analysis.

To address the above challenges, we propose to discretize the compositional vector  $R$  into a vector  $X$  of binary measurements  $\{-1, 1\}$  based on a prespecified abundance threshold such as the median relative abundance. In particular,  $-1$  represents absence or a relative abundance below the given threshold, whereas  $1$  represents presence or a relative abundance above the threshold. After discretizing the data, one can use a binary Markov random field to model conditional dependencies between the components of the discretized vector  $X$ . The pairwise relationships can be captured by an undirected graph  $G = (V, E)$  whose vertex set is  $V = \{1, \dots, p\}$  and whose edge set  $E$  corresponds to conditional dependencies. The binary Markov random field associated with graph  $G$  has the joint distribution

$$\text{pr}_{\Theta}(X) \propto \exp \left( \sum_{(r,t) \in E} \theta_{r,t} X_r X_t \right), \quad (1)$$

subject to a normalizing constant, where  $\theta_{r,t}$  quantifies the conditional dependence between taxa  $r$  and  $t$ . With binary  $X$ , this model has a clear biological interpretation: positive  $\theta_{r,t}$  implies co-existence and negative  $\theta_{r,t}$  implies co-exclusiveness.

### 1.2. Differential network analysis

Let  $\Theta_1 = (\theta_{r,t,1})$  and  $\Theta_2 = (\theta_{r,t,2})$  be the matrices that represent the microbial network among discretized taxa abundances for two age groups. In this paper, we are interested in the global test

$$H_0 : \Theta_1 = \Theta_2 \quad \text{versus} \quad H_1 : \Theta_1 \neq \Theta_2. \quad (2)$$

If the global null in (2) is rejected, it becomes of interest to test for entrywise changes in the differential network  $\Delta = \Theta_1 - \Theta_2 = (\delta_{r,t})$ ,

$$H_{0,r,t} : \delta_{r,t} = 0 \quad \text{versus} \quad H_{1,r,t} : \delta_{r,t} \neq 0 \quad (1 \leq r < t \leq p), \quad (3)$$

while controlling the false discovery rate at a prespecified level.

The gut microbial networks of young adults and the elderly often differ only in a small number of links (Goodrich et al., 2016). Our proposed global test uses the maximum of the standardized entrywise differences as the test statistic, and is thus particularly powerful against such sparse alternatives, compared to tests based on the Frobenius norm (Schott, 2007; Li & Chen, 2012). Our multiple testing procedure for differential network analysis accounts for the multiplicity in testing the  $p(p - 1)/2$  hypotheses, with both the false discovery proportion and the false discovery rate controlled asymptotically. The proposed global and multiple testing procedures are implemented in the R package `TestBMN` (R Development Core Team, 2019), which is available on GitHub. The merits of the proposed tests are further demonstrated through extensive simulations and a gut microbiome study of U.K. twins.

### 1.3. Our contribution

In an unpublished 2016 technical report from the University of Pittsburgh, Z. Ren developed a method for estimation of individual entries in  $\Theta$  for a single binary Markov random field model. In contrast, this paper studies global and multiple testing of two Markov random field models with multiple testing error control. Nodewise logistic regressions are used to develop the entrywise test statistics, whose dependence structure is much more complicated than in the Gaussian case (Xia et al., 2015; Cai & Liu, 2016; Xia et al., 2018). The debiased entrywise estimators for testing Gaussian graphical models are based on the residuals from nodewise linear regressions whose correlations are straightforward to characterize. However, the estimators in the current paper depend not only on the residual from each nodewise logistic regression but on carefully defined projection directions needed for bias correction. By overcoming these challenges, we establish theoretical properties of the proposed testing procedures.

## 2. GLOBAL AND MULTIPLE TESTING OF MARKOV NETWORKS

### 2.1. Notation and problem set-up

Let  $\{X^{(1)}, \dots, X^{(n_1)}\}$  be the  $n_1$  independent binary observations from the first population and  $\{Y^{(1)}, \dots, Y^{(n_2)}\}$  the  $n_2$  independent binary observations from the second population, often conveniently written as matrices  $X \in \mathbb{R}^{n_1 \times p}$  and  $Y \in \mathbb{R}^{n_2 \times p}$ . For a matrix  $X \in \mathbb{R}^{n_1 \times p}$ ,  $X_{-r}$  denotes the  $n_1 \times (p - 1)$  submatrix with the  $r$ th column removed. For a matrix  $\Theta_k = (\theta_{r,t,k})_{1 \leq r, t \leq p}$  and  $k = 1, 2$ , let  $\theta_{r,-r,k} = \{\theta_{r,t,k}, t \neq r\}$  denote the  $(p - 1)$ -dimensional subvector of parameters. For a symmetric matrix  $A$ ,  $\phi_{\max}(A)$  and  $\phi_{\min}(A)$  denote the largest and smallest eigenvalues of  $A$ . For a vector  $a \in \mathbb{R}^p$ , the usual vector  $\ell_1, \ell_2$  and  $\ell_\infty$  norms are denoted, respectively, by  $\|a\|_1, \|a\|_2$  and  $\|a\|_\infty$ . The indicator function is denoted by  $I(\cdot)$ .

To leverage the sparsity in the differential network  $\Delta$ , we propose a test based on the maximum of the standardized entrywise differences between the two matrices (Xia et al., 2015). Our motivation is that the global null hypothesis in (2) is equivalent to

$$H_0 : \max_{1 \leq r < t \leq p} |\theta_{r,t,1} - \theta_{r,t,2}| = 0.$$

Therefore one can construct the test statistic  $M_{n,p}$  by first obtaining nearly unbiased estimators of  $\theta_{r,t,k}$  ( $k = 1, 2$ ) and then deriving the standardized entrywise difference  $W_{r,t}$ , such that  $M_{n,p} = \max_{1 \leq r < t \leq p} W_{r,t}^2$ .

If the global null  $\Theta_1 = \Theta_2$  is rejected, simultaneous testing of entrywise differences is often of interest to identify where the two networks differ. To this end, we also consider multiple testing of all entries in  $\Delta$ , as formally defined in (3). A natural test statistic for each individual test in (3) is the standardized entrywise difference  $W_{r,t}$ . For any given threshold level  $\tau > 0$ , the null hypothesis  $H_{0,r,t}$  is rejected if  $|W_{r,t}| \geq \tau$ . Our goal is to choose an optimal threshold  $\tau$  such that the proposed multiple testing procedure rejects as many true positives as possible while controlling the false discovery rate at a prespecified level.

### 2.2. Derivation of the standardized entrywise statistics

To obtain  $W_{r,t}$ , we first derive nearly unbiased estimators of  $\theta_{r,t,k}$  ( $k = 1, 2$ ) and evaluate their variances. Without loss of generality, it suffices to focus on estimation of  $\theta_{r,t,1}$  from  $X$ .

Given  $X$ , one can recover the association parameters in a binary Markov random field using  $\ell_1$ -penalized nodewise logistic regression (Ravikumar et al., 2010), because the conditional distribution of  $X_r^{(i)}$  given  $X_{-r}^{(i)}$  is

$$\text{pr}(X_r^{(i)} | X_{-r}^{(i)}) = \frac{\exp(X_r^{(i)} \sum_{j \neq r} X_j^{(i)} \theta_{r,j,1})}{\exp(-X_r^{(i)} \sum_{j \neq r} X_j^{(i)} \theta_{r,j,1}) + \exp(X_r^{(i)} \sum_{j \neq r} X_j^{(i)} \theta_{r,j,1})}. \tag{4}$$

Thus the variable  $X_r$  can be viewed as the response in a logistic regression where all remaining variables  $X_{-r}$  act as covariates. However,  $\hat{\theta}_{r,t,1}$  obtained via  $\ell_1$ -penalized estimation (Ravikumar et al., 2010) is biased.

To correct for the bias in  $\hat{\theta}_{r,t,1}$ , it is instructive to write the  $r$ th variable as a nonlinear function of all remaining variables, i.e., for every node  $r = 1, \dots, p$ ,

$$X_r^{(i)} = E(X_r^{(i)} | X_{-r}^{(i)}) + \varepsilon_{r,1}^{(i)} = \dot{f}(u_{r,1}^{(i)}) + \varepsilon_{r,1}^{(i)} \quad (i = 1, \dots, n_1), \tag{5}$$

where  $u_{r,1}^{(i)} = X_{-r}^{(i)} \theta_{r,-r,1}$ ,  $\dot{f}(u) = \tanh(u)$  and  $\varepsilon_{r,1}^{(i)}$  are random variables satisfying  $E(\varepsilon_{r,1}^{(i)} | X_{-r}^{(i)}) = 0$ . We adopt the projection-based debiasing approach of Zhang & Zhang (2014). Specifically, for a suitably chosen score vector  $v_{r,t,1} \in \mathbb{R}^{n_1}$  and  $\hat{u}_{r,1}^{(i)} = X_{-r}^{(i)} \hat{\theta}_{r,-r,1}$ , one can show that the estimator

$$\check{\theta}_{r,t,1} = \hat{\theta}_{r,t,1} + \frac{\sum_{i=1}^{n_1} v_{r,t,1}^{(i)} \{X_r^{(i)} - \dot{f}(\hat{u}_{r,1}^{(i)})\}}{\sum_{i=1}^{n_1} v_{r,t,1}^{(i)} \ddot{f}(\hat{u}_{r,1}^{(i)}) X_r^{(i)}} \quad (t \neq r) \tag{6}$$

is nearly unbiased under mild conditions on the initial estimate  $\hat{\theta}_{r,-r,1}$ . The second term on the right-hand side of (6) projects the residual  $X_r^{(i)} - \dot{f}(\hat{u}_{r,1}^{(i)})$  onto the direction of  $v_{r,t,1}^{(i)}$ , thereby reducing the bias in  $\hat{\theta}_{r,t,1}$  to an acceptable level. As shown in the unpublished 2016 technical report by Z. Ren, the desired score vector  $v_{r,t,1}$  ( $t \neq r$ ) can be defined as

$$v_{r,t,1}^{(i)} = (X_t^{(i)} + 1)/2 - g(X_{-\{r,t\}}^{(i)}, \hat{\theta}_{r,-r,1}, \hat{\theta}_{t,-t,1}) \quad (i = 1, \dots, n_1), \tag{7}$$

where for  $\ddot{f}(u) = 4e^{2u}/(e^{2u} + 1)^2$ ,

$$g(X_{-\{r,t\}}, \theta_{r,-r,1}, \theta_{t,-t,1}) = \frac{E\{\dot{f}(X_{-r}\theta_{r,-r,1})(X_t + 1)/2 \mid X_{-\{r,t\}}\}}{E\{\dot{f}(X_{-r}\theta_{r,-r,1}) \mid X_{-\{r,t\}}\}}.$$

Intuitively, the score vector  $v_{r,t,1}$  resembles the residual for regressing  $X_t$  on  $X_{-\{r,t\}}$ . The choice in (7) ensures that  $v_{r,t,1}$  is uncorrelated with  $\varepsilon_{r,1}$  and that  $\check{\theta}_{r,t,1}$  achieves asymptotic efficiency. Given  $Y$  and the initial estimate  $\hat{\Theta}_2$ , the nearly unbiased  $\check{\theta}_{r,t,2}$  can be derived similarly.

*Remark 1.* Inference for high-dimensional models via the  $\ell_1$  penalty typically requires bias correction, because the  $\ell_1$  regularization achieves variable selection by shrinking every component of the regression coefficients towards zero (Zhang & Zhang, 2014; van de Geer et al., 2014). Existing work on bias correction focuses primarily on linear models, with the exception of van de Geer et al. (2014), which is based on inverting the Karush–Kuhn–Tucker conditions and is applicable to generalized linear models. In contrast, our debiasing approach is based on a local Taylor expansion of  $\dot{f}(u_{r,1}^{(i)})$  about  $\hat{u}_{r,1}^{(i)}$  in (5), which yields an approximately linear surrogate to the nonlinear conditional mean and facilitates the construction of entrywise statistics.

The estimators  $\check{\theta}_{r,t,1}$  and  $\check{\theta}_{r,t,2}$  have unequal variances, so we must adjust for their variances before constructing the test statistic for the global null hypothesis. Denote by  $v_{r,t,1}^{(i),o}$  the oracle score vector calculated based on  $\Theta_1$ ,

$$v_{r,t,1}^{(i),o} = (X_t^{(i)} + 1)/2 - g(X_{-\{r,t\}}^{(i)}, \theta_{r,-r,1}, \theta_{t,-t,1}) \quad (i = 1, \dots, n_1),$$

and let

$$F_{r,t,1} = E[\{X_t - 2g(X_{-\{r,t\}}, \theta_{r,-r,1}, \theta_{t,-t,1})\}^2 \ddot{f}(u_{r,1})].$$

By definition,  $E\{v_{r,t,1}^{(i),o} \ddot{f}(u_{r,1}^{(i)}) \mid X_{-\{r,t\}}^{(i)}\} = 0$ , so  $E\{v_{r,t,1}^{(i),o} \ddot{f}(u_{r,1}^{(i)}) X_t^{(i)}\} = 2E\{(v_{r,t,1}^{(i),o})^2 \ddot{f}(u_{r,1}^{(i)})\} = F_{r,t,1}/2$ . Hence, for a reasonably good initial estimate  $\hat{\theta}_{r,-r,1}$ , one expects  $\check{\theta}_{r,t,1}$  to be close to

$$\tilde{\theta}_{r,t,1} = \theta_{r,t,1} + \frac{1}{n_1} \sum_{i=1}^{n_1} \frac{2v_{r,t,1}^{(i),o} \varepsilon_{r,1}^{(i)}}{F_{r,t,1}}.$$

Indeed, it can be shown that  $|\check{\theta}_{r,t,1} - \tilde{\theta}_{r,t,1}| = o_p\{(n_1 \log p)^{-1/2}\}$  for  $1 \leq r < t \leq p$ , under certain conditions. With  $s_{r,t,1} = \text{var}(2v_{r,t,1}^{(i),o} \varepsilon_{r,1}^{(i)}/F_{r,t,1}) = 1/F_{r,t,1}$ , one can use  $s_{r,t,1}/n_1$  as an approximation to  $\text{var}(\check{\theta}_{r,t,k})$ .

Finally, let the empirical variance estimates be defined as

$$\check{s}_{r,t,1} = \left\{ 4n_1^{-1} \sum_{i=1}^{n_1} (v_{r,t,1}^{(i)})^2 \ddot{f}(\hat{u}_{r,1}^{(i)}) \right\}^{-1}, \quad \check{s}_{r,t,2} = \left\{ 4n_2^{-1} \sum_{i=1}^{n_1} (v_{r,t,2}^{(i)})^2 \ddot{f}(\hat{u}_{r,2}^{(i)}) \right\}^{-1}.$$

The standardized entrywise difference is thus

$$W_{r,t} = \frac{\check{\theta}_{r,t,1} - \check{\theta}_{r,t,2}}{(\check{s}_{r,t,1}/n_1 + \check{s}_{r,t,2}/n_2)^{1/2}} \quad (1 \leq r < t \leq p).$$

### 2.3. Implementation of the testing procedure

Once we have the standardized entrywise statistic  $W_{r,t}$ , we can construct the test statistic for the global null in (2):

$$M_{n,p} = \max_{1 \leq r < t \leq p} W_{r,t}^2 = \max_{1 \leq r < t \leq p} \frac{(\check{\theta}_{r,t,1} - \check{\theta}_{r,t,2})^2}{\check{s}_{r,t,1}/n_1 + \check{s}_{r,t,2}/n_2}. \quad (8)$$

Section 3.2 shows that under certain mild conditions,  $M_{n,p} - 4 \log p + \log \log p$  converges to a Type I extreme value distribution under the null. Indeed, the  $W_{r,t}$  ( $1 \leq r < t \leq p$ ) are asymptotically standard normal under the null and are only weakly dependent. Hence the maximum of  $p(p-1)/2$  such variables squared, that is,  $M_{n,p}$ , should be close to  $2 \log\{p(p-1)/2\} \approx 4 \log p$ . Let

$$\Psi_\alpha = I(M_{n,p} \geq q_\alpha + 4 \log p - \log \log p), \quad (9)$$

where  $q_\alpha$  is the  $(1-\alpha)$  quantile of the Type I extreme value distribution with the cumulative distribution function  $\exp\{-(8\pi)^{-1/2} \exp(-z/2)\}$ . We reject the hypothesis  $H_0 : \Theta_1 = \Theta_2$  whenever  $\Psi_\alpha = 1$ .

In the multiple testing problem (3), for any given threshold level  $\tau > 0$  and  $1 \leq r < t \leq p$ , each individual hypothesis  $H_{0,r,t}$  is rejected if  $|W_{r,t}| \geq \tau$ . Let  $\mathcal{H}_0 = \{(r, t) : \theta_{r,t,1} = \theta_{r,t,2}, 1 \leq r < t \leq p\}$  denote the set of true nulls. Let  $R_0(\tau) = \sum_{(r,t) \in \mathcal{H}_0} I(|W_{r,t}| \geq \tau)$  be the total number of false positives and  $R(\tau) = \sum_{1 \leq r < t \leq p} I(|W_{r,t}| \geq \tau)$  the total number of rejections. The false discovery proportion and false discovery rate are defined, respectively, as

$$\text{FDP}(\tau) = \frac{R_0(\tau)}{\max\{R(\tau), 1\}}, \quad \text{FDR}(\tau) = E\{\text{FDP}(\tau)\}.$$

For a prespecified level  $\alpha$ , an ideal choice of  $\tau$  that is able to control the false discovery proportion and false discovery rate is

$$\tau_0 = \inf \left\{ 0 \leq \tau \leq 2(\log p)^{1/2} : \text{FDP}(\tau) \leq \alpha \right\}.$$

Here the choice of  $\tau$  is restricted to  $[0, 2(\log p)^{1/2}]$  because the asymptotic null distribution of  $W_{r,t}$  is standard normal so that  $\text{pr}\{\max_{(r,t) \in \mathcal{H}_0} |W_{r,t}| \geq 2(\log p)^{1/2}\} \rightarrow 0$  as  $n_1, n_2, p \rightarrow \infty$ .

However, the ideal  $\tau_0$  is unavailable because  $\mathcal{H}_0$  is unknown. To estimate  $\tau_0$ , it is helpful to understand the properties of  $W_{r,t}$ . Let  $\Phi(\tau)$  be the standard normal cumulative distribution function and let  $G(\tau) = 2 - 2\Phi(\tau)$ . Under the null hypothesis in (3) and some regularity conditions, one can show that as  $n_1, n_2 \rightarrow \infty$ ,

$$\sup_{0 \leq \tau \leq c(\log p)^{1/2}} \left| \frac{\text{pr}(|W_{r,t}| \geq \tau)}{G(\tau)} - 1 \right| \rightarrow 0 \quad (10)$$

uniformly for all  $1 \leq r < t \leq p$ , where  $p = n_k^\gamma$ , for any  $c > 0$  and any  $\gamma > 0$ . Therefore one can estimate  $R_0(\tau)$  by  $2\{1 - \Phi(\tau)\}|\mathcal{H}_0|$  as in Cai & Liu (2016), where  $|\mathcal{H}_0|$  can be estimated by  $q = (p^2 - p)/2$  due to the sparsity of  $\Delta$ . In fact, for weakly dependent  $W_{r,t}$ , it can be shown that

$$\sup_{0 \leq \tau \leq b_p} \left| \frac{R_0(\tau)}{G(\tau)|\mathcal{H}_0|} - 1 \right| \rightarrow 0 \quad (11)$$

in probability for  $b_p = \{4 \log p - 2 \log(\log p)\}^{1/2}$  as  $n_1, n_2, p \rightarrow \infty$ . So, we estimate  $FDP(\tau)$  by

$$\frac{G(\tau)(p^2 - p)/2}{\max\{R(\tau), 1\}}.$$

This leads to the following estimate of  $\tau_0$ :

$$\hat{\tau} = \inf \left[ 0 \leq \tau \leq b_p : \frac{G(\tau)(p^2 - p)/2}{\max\{R(\tau), 1\}} \leq \alpha \right]. \tag{12}$$

If the solution to (12) does not exist, we set  $\hat{\tau} = 2(\log p)^{1/2}$ . For  $1 \leq r < t \leq p$ , the null hypothesis  $H_{0,r,t} : \delta_{r,t} = 0$  is rejected if and only if  $|W_{r,t}| \geq \hat{\tau}$ .

*Remark 2.* It is important to restrict  $\tau \in [0, b_p]$  in (12) for false discovery proportion control, because the convergence in (11) may not hold for  $\tau > b_p$ . In such cases, direct thresholding on  $W_{r,t}$  with  $\hat{\tau} = 2(\log p)^{1/2}$  is used to control the false discovery rate. Without the constraint  $\tau \leq b_p$ , our multiple testing procedure reduces to the Benjamini–Hochberg procedure, which may not control the false discovery proportion with some positive probability if the number of true alternatives  $|\mathcal{H}_0^c|$  is fixed as  $p \rightarrow \infty$ . An alternative approach to approximating  $R_0(\tau)$  is to bootstrap, as in Cai & Liu (2016).

### 3. THEORETICAL PROPERTIES

#### 3.1. Assumptions

We make assumptions to establish the theoretical properties of the proposed testing procedures. For  $r = 1, \dots, p$ , let

$$Q_{r,1} = E_{\Theta_1} \{ \ddot{f}(X_{-r} \theta_{r,-r,1}) X_{-r} X_{-r}^T \}, \quad Q_{r,2} = E_{\Theta_2} \{ \ddot{f}(Y_{-r} \theta_{r,-r,2}) Y_{-r} Y_{-r}^T \}$$

denote the Hessians of the likelihood functions associated with the  $r$ th logistic regression in the first and second populations, respectively. For  $k = 1, 2$ , the matrix  $Q_{r,k}$  is the Fisher information matrix associated with the local conditional probability distribution, and is analogous to the  $(p - 1) \times (p - 1)$  submatrix of the precision matrix in Gaussian graphical models.

*Assumption 1.* Assume that  $\log p = o(n_k^{1/3})$  and  $n_1 \asymp n_2$ . For each  $r = 1, \dots, p$ , there exist constants  $C_{\min}, C_{\max} > 0$  such that

$$0 < C_{\min} \leq \phi_{\min}(Q_{r,k}) \leq \phi_{\max}(Q_{r,k}) \leq C_{\max} < \infty.$$

Further, assume that the initial estimators  $\hat{\Theta}_k$  ( $k = 1, 2$ ) satisfy

$$\max_{1 \leq r \leq p} \|\hat{\theta}_{r,-r,k} - \theta_{r,-r,k}\|_1 = o_p\{(\log p)^{-1}\}, \tag{13}$$

$$\max_{1 \leq r \leq p} \|\hat{\theta}_{r,-r,k} - \theta_{r,-r,k}\|_2 = o_p\{(n_k \log p)^{-1/4}\}. \tag{14}$$



The bounded eigenvalue assumption on  $Q_{r,k}$  is standard in inference for high-dimensional Ising models (Ravikumar et al., 2010), and ensures that the  $(p - 1)$  variables do not become overly dependent. It is also crucial for establishing the asymptotic normality of  $\check{\theta}_{r,t,k}$ . Initial estimators satisfying (13) and (14) can be obtained via  $\ell_1$ -penalized nodewise logistic regression if the maximum node degree is  $d_k = o\{n_k^{1/2}/(\log p)^{3/2}\}$ . The conditions in (13) and (14) are slightly stronger than those required for entrywise normality in the unpublished 2016 technical report by Z. Ren, because smaller biases are required for testing procedures in order to provide significance quantification for each pair of edges.

*Assumption 2.* For each  $r = 1, \dots, p$ , there exists a constant  $C_w > 0$  such that the maximum neighbourhood weight satisfies

$$\max_{1 \leq r \leq p} \sum_{t:(r,t) \in E} |\theta_{r,t,k}| \leq C_w < \infty.$$

Assumption 2 implies that  $X_{-r}\theta_{r,-r,1}$  and  $Y_{-r}\theta_{r,-r,2}$  are sub-exponential random variables, and therefore  $\min_{1 \leq r < t \leq p} F_{r,t,k} > c^* > 0$  for  $k = 1, 2$  and some  $c^* > 0$ . In view of the conditional distribution (4), violation of Assumption 2 may lead to degenerate marginal distributions. This assumption is crucial for controlling the dependence in Lemma 1, and is also used in Santhanam & Wainwright (2012).

To ensure good performance of the proposed testing procedures, it is important to characterize the dependence between the standardized entrywise statistics  $W_{r,t}$  and  $W_{r',t'}$  for  $(r, t) \neq (r', t')$ , and to understand when such dependences are weak. The reason for  $W_{r,t}$  and  $W_{r',t'}$  being dependent is the correlation among  $v_{r,t,k}^o \varepsilon_{r,k}$  for  $1 \leq r < t \leq p$  and  $k = 1, 2$ . In contrast to the Gaussian case (Xia et al., 2015), it is difficult to evaluate these correlations analytically in the current setting due to the unknown normalizing constant in (1). However, the following lemma says that these correlations are bounded under mild conditions.

LEMMA 1. Under Assumption 2, for  $r \neq r', t \neq t'$  and  $k = 1, 2$ ,

$$\begin{aligned} |\text{cor}(v_{r,t,k}^o \varepsilon_{r,k}, v_{r',t',k}^o \varepsilon_{r',k})| &\leq \frac{4C_0}{c^*} \{ |\sinh(2\theta_{r,r',k})| + |\sinh(2\theta_{r,t,k}) \sinh(2\theta_{r',t,k})| \}, \\ |\text{cor}(v_{r,t,k}^o \varepsilon_{r,k}, v_{r,t',k}^o \varepsilon_{r,k})| &\leq \frac{4C_0}{c^*} \{ |\sinh(2\theta_{t,t',k})| + |\sinh(2\theta_{r,t,k}) \sinh(2\theta_{r,t',k})| \}, \\ |\text{cor}(v_{r,t,k}^o \varepsilon_{r,k}, v_{r',t',k}^o \varepsilon_{r',k})| &\leq \frac{C_1}{c^*} \left\{ \sum_{a=r',t'} |\sinh(2\theta_{r,a,k})| \right\} \left\{ \sum_{b=r',t'} |\sinh(2\theta_{t,b,k})| \right\}, \end{aligned}$$

where  $C_0$  and  $C_1$  are absolute constants that depend only on  $C_w$ .

Lemma 1 implies that the correlation among the nearly unbiased estimators  $\tilde{\theta}_{r,t,k}$ , or equivalently the correlation among the  $W_{r,t}$ , is bounded above by a function of  $\sinh(2\theta_{r,t,k})$  for  $1 \leq r < t \leq p$  and  $k = 1, 2$ . However, establishing such results for high-dimensional binary Markov random fields is challenging. The proof of Lemma 1 requires careful analysis based on the conditional distributions  $\text{pr}(X_r, X_{r'} \mid X_{-\{r,r'\}})$ ,  $\text{pr}(X_r, X_t, X_{r'} \mid X_{-\{r,t,r'\}})$  and  $\text{pr}(X_r, X_t, X_{r'}, X_{t'} \mid X_{-\{r,t,r',t'\}})$ .

The next assumption requires that the entrywise conditional associations not be too large and ensures that  $W_{r,t}$  ( $1 \leq r < t \leq p$ ) only weakly dependent.



Assumption 3. For a constant  $\xi > 0$ , let

$$\mathcal{A}_t(\xi) = \{r : |\sinh(2\theta_{r,t,1})| \geq (\log p)^{-2-\xi} \text{ or } |\sinh(2\theta_{r,t,2})| \geq (\log p)^{-2-\xi}\}$$

be small enough that  $\max_{1 \leq t \leq p} |\mathcal{A}_t(\xi)| = o(p^\gamma)$  for  $0 < \gamma < 1/3$ .

Variants of Assumption 3 are commonly used (Cai et al., 2013; Cai & Liu, 2016; Xia et al., 2015, 2018). This assumption is also mild for microbial networks because evolutionary stability and robustness often result in only a small number of strong microbial interactions (Leclerc, 2008).

### 3.2. Limiting null distribution and optimality of the global test

For  $0 < \eta < 1$ , let

$$\Lambda(\eta) = \{1 \leq r \leq p : |\sinh(2\theta_{r,t,1})| > \eta \text{ or } |\sinh(2\theta_{r,t,2})| > \eta \text{ for some } t \neq r\}$$

denote the set of indices  $r$  such that either  $X_r$  is highly correlated with some  $X_t$  or  $Y_r$  is highly correlated with some  $Y_t$ , given all remaining variables. Let  $\eta^* = \min\{0.5(c^*/C_1)^{1/2}, 0.125c^*/C_0\}$ .

**THEOREM 1.** *Suppose that Assumptions 1–3 hold. If there exist  $0 < \eta < \eta^*$  and a sequence of numbers  $\Lambda_{p,\eta} > 0$  such that  $|\Lambda(\eta)| \leq \Lambda_{p,\eta} = o(p)$ , then under the null hypothesis in (2), for any  $z \in \mathbb{R}$ ,*

$$\text{pr}(M_{n,p} - 4 \log p + \log \log p \leq z) \rightarrow \exp\{-(8\pi)^{-1/2} \exp(-z/2)\} \tag{15}$$

as  $n_1, n_2, p \rightarrow \infty$ , where  $M_{n,p}$  is defined in (8). Under the null hypothesis, the convergence in (15) is uniform for all  $X$  and  $Y$  satisfying Assumptions 1–3.

*Remark 3.* The bounded cardinality condition on  $\Lambda(\eta)$  is similar to that in Cai et al. (2013), and is mild because the magnitude of  $\sinh(2\theta_{r,t,k})$  reflects the conditional association between each pair of microbial taxa. Most of the entries in  $\Theta_k$  should be bounded from above, or the rare taxa might go extinct, which is undesirable for maintaining a robust ecological system. This condition together with Assumptions 2 and 3 guarantee weak dependence among the  $W_{r,t}$ , and allows us to apply strategies for extreme values in the Gaussian case.

In addition to the limiting null distribution, our global testing procedure also maintains rate-optimal power. Consider the matrices

$$\mathcal{U}(c) = \left\{ (\Theta_1, \Theta_2) : \max_{1 \leq r < t \leq p} \frac{|\theta_{r,t,1} - \theta_{r,t,2}|}{(s_{r,t,1}/n_1 + s_{r,t,2}/n_2)^{1/2}} \geq c(\log p)^{1/2} \right\}. \tag{16}$$

Let  $\mathcal{T}_\alpha$  be the set of all  $\alpha$ -level tests, that is,  $\text{pr}(T_\alpha = 1) \leq \alpha$  for any  $T_\alpha \in \mathcal{T}_\alpha$ .

**THEOREM 2.** (i) *Suppose that Assumptions 1–3 hold. The test  $\Psi_\alpha$  defined in (9) satisfies*

$$\inf_{(\Theta_1, \Theta_2) \in \mathcal{U}(4)} \text{pr}(\Psi_\alpha = 1) \rightarrow 1$$

as  $n_1, n_2, p \rightarrow \infty$ .

(ii) Let  $\log p = o(n_k)$  for  $k = 1, 2$  and  $\alpha, \beta > 0$  with  $\alpha + \beta < 1$ . Then there exists a constant  $c_0 > 0$  such that for all sufficiently large  $n_1, n_2$  and  $p$ ,

$$\inf_{(\Theta_1, \Theta_2) \in \mathcal{U}(c_0)} \sup_{T_\alpha \in \mathcal{T}_\alpha} \text{pr}(T_\alpha = 1) \leq 1 - \beta.$$

*Remark 4.* Statement (i) in Theorem 2 says that  $\Psi_\alpha$  rejects the null hypothesis in (2) with high probability when  $(\Theta_1, \Theta_2) \in \mathcal{U}(4)$ , and statement (ii) says that the lower bound of order  $(\log p)^{1/2}$  in (16) cannot be further improved.

### 3.3. False discovery rate control in multiple testing

Let  $q_0 = |\mathcal{H}_0|$ , let  $q = (p^2 - p)/2$ , and let

$$\mathcal{S}_\rho = \left\{ (r, t) : 1 \leq r < t \leq p, \frac{|\theta_{r,t,1} - \theta_{r,t,2}|}{(s_{r,t,1}/n_1 + s_{r,t,2}/n_2)^{1/2}} \geq (\log p)^{1/2+\rho} \right\}.$$

The following theorem shows that our multiple testing procedure ensures asymptotic control of the false discovery proportion and false discovery rate at the prespecified level  $\alpha$ .

**THEOREM 3.** Suppose  $|\mathcal{S}_\rho| \geq \{(8\pi)^{-1/2}\alpha^{-1} + \zeta\}(\log \log p)^{1/2}$  for some  $\rho > 0$  and  $\zeta > 0$ . Assume further that  $q_0 \geq c_1 p^2$  for some  $c_1 > 0$  and that  $p \leq c_2 n_k^\gamma$  for some  $c_2 > 0$  and  $\gamma > 0$ . Then under the assumptions of Theorem 1,

$$\lim_{(n_1, n_2, p) \rightarrow \infty} \frac{\text{FDR}(\hat{\tau})}{\alpha q_0/q} = 1, \quad \frac{\text{FDP}(\hat{\tau})}{\alpha q_0/q} \rightarrow 1$$

in probability as  $n_1, n_2, p \rightarrow \infty$ .

*Remark 5.* The condition on the size of  $\mathcal{S}_\rho$  is mild in that it only requires a few standardized entrywise differences to be at least of order  $(\log p)^{1/2+\rho}$ , which is necessary to ensure that (11) holds. The proof of Theorem 3 again relies heavily on knowledge of the dependence among the  $W_{r,t}$  as characterized in Lemma 1.

## 4. SIMULATIONS

### 4.1. Data and model selection

We consider (a) the Erdős–Rényi random graph  $G_{\text{ER}}(p, d/p)$  (Erdős & Rényi, 1960) with average degree  $d = 4$ , (b) the Watts–Strogatz model  $G_{\text{WS}}$  (Watts & Strogatz, 1998), which forms networks with small-world properties, and (c) the Barabasi–Albert scale-free network model  $G_{\text{BA}}$  (Barabási & Albert, 1999), as illustrated in Fig. 1. In all comparisons, the graph sizes and the sample sizes are taken as  $p \in \{50, 100\}$  and  $n_1 = n_2 = 100$ . For a given graph, the nonzero entries  $\theta_{r,t,k}$  were drawn uniformly from  $[-0.5, -0.1] \cup [0.1, 0.5]$ . Given the two matrices  $\Theta_1$  and  $\Theta_2$ , binary data  $\{X^{(i)}\}_{i=1}^{n_1} \sim \text{pr}_{\Theta_1}(X)$  and  $\{Y^{(i)}\}_{i=1}^{n_2} \sim \text{pr}_{\Theta_2}(Y)$  were generated by Gibbs sampling.

To get an initial estimator of  $\Theta_k$ , we run nodewise  $\ell_1$ -regularized logistic regression with penalty parameter  $\lambda_{r,k}$ , using the R package `glmnet` (Friedman et al., 2010). Symmetric estimates were obtained by averaging the nodewise estimates. The optimal tuning parameters  $\lambda_{r,k}$  ( $r = 1, \dots, p; k = 1, 2$ ) were chosen to maximize the performance of the global and the

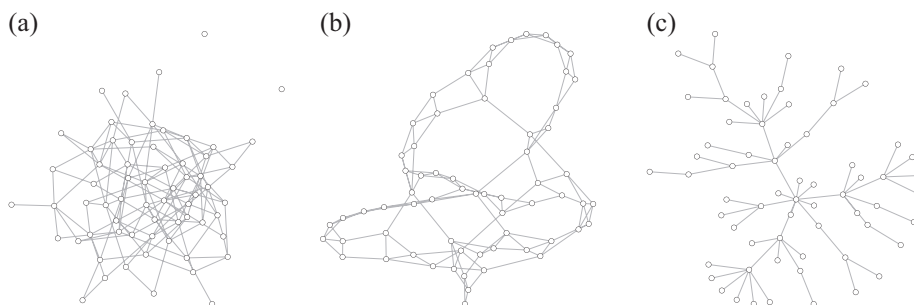


Fig. 1. Illustrations of different graphs used in our simulations: (a) random graph  $G_{ER}(p, d/p)$ , (b) small-world network  $G_{WS}$ , and (c) scale-free network  $G_{BA}$ .

Table 1. Empirical Type I errors and powers (%) of the global test with  $\alpha = 5\%$  and  $n_1 = n_2 = 100$ . The Type I errors were evaluated over 500 replications, and the powers were calculated over 200 replications

$p$	Method	$G_{ER}$	$G_{WS}$	$G_{BA}$	$G_{ER}$	$G_{WS}$	$G_{BA}$
		Type I error			Power		
50	Proposed method	3.0	1.6	3.6	94.5	88.0	91.5
	Xia et al. (2015)	8.4	9.4	7.2	69.5	70.0	91.0
	PermBMN	5.8	5.8	5.0	97.0	93.5	96.5
	PermMLE	4.6	5.2	3.8	98.5	97.0	99.5
100	Proposed method	2.0	3.0	2.4	95.5	94.0	97.5
	Xia et al. (2015)	6.2	7.8	7.4	80.0	87.0	94.5
	PermBMN	5.4	5.0	4.2	98.0	97.0	99.0
	PermMLE	4.6	4.8	4.6	97.0	99.0	100.0

PermBMN, permutation-based test using  $\hat{\Theta}_k$ ; PermMLE, permutation-based test using the modified maximum likelihood estimator of  $\Theta_k$  assuming that the true graph structures are known.

multiple testing procedures. For the global test, the tuning parameter  $\lambda_{r,k}$  in each logistic regression was selected based on the extended Bayesian information criterion (Barber & Drton, 2015). As the multiple testing procedure relies on the approximation of  $2\{1 - \Phi(\tau)\}|\mathcal{H}_0|$  to  $R_0(\tau)$ , the tuning parameters needed in multiple testing were chosen by ensuring  $2\{1 - \Phi(\tau)\}|\mathcal{H}_0|$  was as close to  $\sum_{(r,t) \in \mathcal{H}_0} I(|W_{r,t}| \geq \tau)$  as possible.

#### 4.2. Results

We compare our approach with two-sample testing based on Gaussian graphical models (Xia et al., 2015) and two other permutation-based methods for global testing. The first permutation method uses the same  $\hat{\theta}_{r,t,k}$  and  $\check{s}_{r,t,k}$  as in our approach and takes the minimum  $p$ -value of the entrywise test  $H_{0,r,t}$  in (3) as the test statistic. The significance of the global test is then calibrated using a permutation approach. The second permutation method differs from the first only in the way the parameters  $\theta_{r,t,k}$  and  $s_{r,t,k}$  are estimated. Specifically, it estimates  $\theta_{r,t,k}$  and its variance using maximum likelihood with specified support of  $\Theta_k$  and finite-sample bias correction (Firth, 1993). In our comparisons, the per replication  $p$ -values for both permutation methods were evaluated over 1000 permutations.

Table 1 presents the empirical Type I errors of the global test, and confirms that our method and the two permutation tests control the Type I errors well in all settings. In comparison, the Type I

Table 2. Empirical false discovery rates and true positive rates (%) for multiple testing with false discovery rate  $\alpha = 10\%$  and  $n_1 = n_2 = 100$  over 200 replications

$p$	Method	$G_{ER}$	$G_{WS}$	$G_{BA}$	$G_{ER}$	$G_{WS}$	$G_{BA}$
		False discovery rate			True positive rate		
50	Proposed method	10.0	9.1	10.6	52.9	59.9	61.0
	Xia et al. (2015)	13.8	13.7	13.5	51.7	54.9	59.9
100	Proposed method	8.3	9.4	9.2	56.1	57.7	62.9
	Xia et al. (2015)	12.2	12.3	11.7	50.5	49.6	57.2

errors of the method of Xia et al. (2015) are slightly inflated, which is not surprising because the data are not multivariate normal.

To evaluate the power of the global test, we constructed the differential network  $\Delta$  such that five entries in the upper triangular part of  $\Delta$  were uniformly drawn from  $\{-1.5(\log p/n_k)^{1/2}, 1.5(\log p/n_k)^{1/2}\}$ . Let  $\Theta_0$  be generated from one of the three models  $G_{ER}$ ,  $G_{WS}$  or  $G_{BA}$ , and let  $\Theta_1 = \Theta_0 - \Delta$  and  $\Theta_2 = \Theta_0 + \Delta$ . The empirical powers, which are proportions of null hypothesis rejections, are shown in Table 1. Our method yields very high powers for all settings and uniformly outperforms the method of Xia et al. (2015). The two permutation-based methods, especially the second one based on the modified maximum likelihood estimator of  $\Theta_k$ , demonstrate superior performance in terms of power. The downside is that both permutation methods require heavy computation.

Finally, using the same design  $\Theta_1 = \Theta_0 - \Delta$  and  $\Theta_2 = \Theta_0 + \Delta$ , we examined multiple testing of individual entries in the differential network  $\Delta$  while controlling the false discovery rate at  $\alpha = 10\%$ . The true differential network  $\Delta$  was constructed to be sparse such that the number of edges is approximately  $0.01p(p-1)$ , with nonzero entries drawn uniformly from  $[-0.5, -0.1] \cup [0.1, 0.5]$ . The empirical false discovery and true positive rates of our method and that of Xia et al. (2015) were estimated by

$$\text{Average} \left[ \frac{\sum_{(r,t) \in \mathcal{H}_0} I(|W_{r,t}| \geq \hat{t})}{\max\{R(\hat{t}), 1\}} \right], \quad \text{Average} \left\{ \frac{\sum_{(i,j) \in \mathcal{H}_1} I(|W_{r,t}| \geq \hat{t})}{\sum_{1 \leq r < t \leq p} I(\theta_{r,t,1} \neq \theta_{r,t,2})} \right\},$$

where  $\mathcal{H}_1$  denotes the set of nonzero locations. Permutation-based methods are inapplicable to multiple testing. Table 2 shows that our multiple testing procedure controls the false discovery rate well in all scenarios, and returns reasonably high true positive rates. In contrast, the multiple testing method of Xia et al. (2015) yields slightly higher false positive and lower true positive rates. The difference in true positive rates between the methods increases with  $p$ .

## 5. APPLICATION TO GUT MICROBIOME DATA IN U.K. TWINS

We applied the proposed testing procedures to data from a gut microbiome study of U.K. twins. The original study (Goodrich et al., 2016) investigated whether host genotype shapes the gut microbiome composition, using 16S rRNA sequencing data collected from faecal samples of 2731 individuals. The data are available at <http://www.ebi.ac.uk/ena/data/view/PRJEB13747>. We are interested in whether the microbial community structures summarized as conditional dependence relationships among the microbial taxa are associated with the age of the host.

To ensure robust detection of microbial interactions, samples with total read counts less than 20 were first removed. For each of the remaining 2714 samples, the relative abundance of each

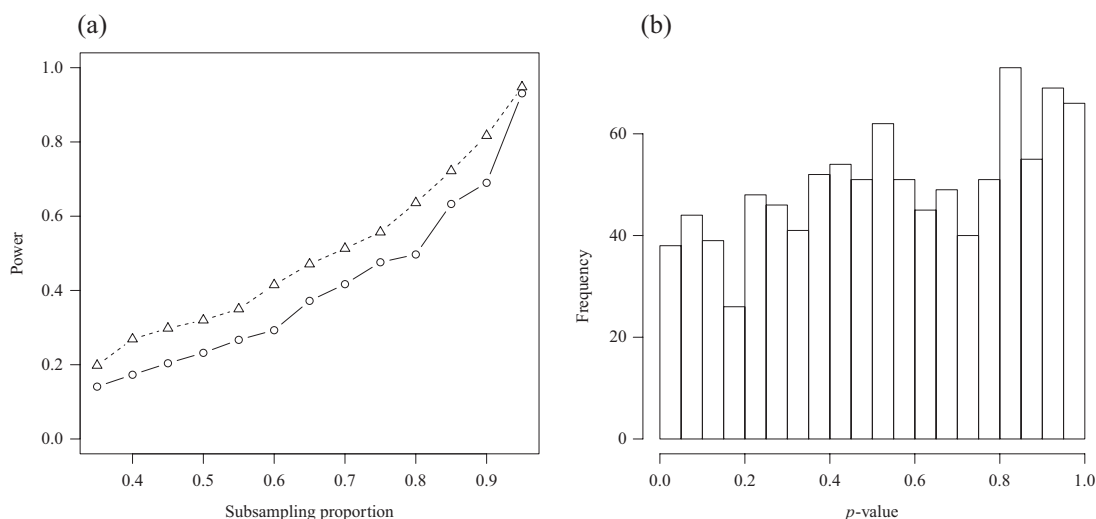


Fig. 2. (a) Empirical power curves of the proposed method (circles) and the permutation method using  $\check{\Theta}_k$  (triangles) by subsampling the two datasets at different proportions. (b) A histogram of the  $p$ -values from our method in back-testing by randomly generating two datasets 1000 times.

genus was calculated by dividing the read count for each genus by the total number of reads in the sample. Of the 294 genera, only those with at least 0.001% of relative abundance in at least 75% of the 2714 samples were used, which reduces the total number of genera to  $p = 59$ . To examine the association of microbial interactions with host age, we selected 286 young adults who were at most 43 years and 284 elderly subjects aged at least 74 years. Since these samples included twins, we randomly chose one individual from each pair of twins, which gave  $n_1 = 171$  independent young adults and  $n_2 = 180$  elderly subjects in our analysis.

Because of possible errors in sequencing reads, genera with relative abundance lower than 0.001% are expected to be due to noise or sequencing errors. We therefore first discretized the relative abundances using a cut-off of 0.001%, where for a given genus,  $-1$  represents absence or extremely low abundance of the genus and  $1$  represents presence. However, some more abundant genera are present in over 75% of the samples. For these genera, we used the 25th percentile as the cut-off to discretize the data, so  $-1$  represents low abundance of the genus and  $1$  represents high abundance. One should keep the definitions in mind when interpreting the  $\Theta$  parameters.

We applied the global testing procedure to the two groups of subjects and obtained a  $p$ -value of 0.009, indicating that the microbial networks for the two age groups are significantly different. To assess the stability of our method and perform power comparisons, we generated 1000 subsamples within each age group, with sampling proportions ranging from 0.35 to 0.95. The empirical powers of our method and the permutation method using  $\check{\Theta}_k$  are presented in Fig. 2(a). Our method performs slightly worse than the permutation method, but we believe that the gap reduces with larger samples. Another advantage of our method is that it runs much faster than the permutation method. We also performed back-testing for the global null hypothesis by randomly splitting the pooled data into two groups with  $n_1$  and  $n_2$  subjects over 1000 replications. Figure 2(b) shows the resulting  $p$ -values, which are slightly skewed towards the right, indicating that the proposed test is conservative.

To recover the differential microbial network, we applied the proposed multiple testing procedure with  $FDR = 15\%$  and show the results in Fig. 3. The same differential links were selected by FDR between 14% to 16%. For FDR between 12% to 14%, all edges in Fig. 3 except that

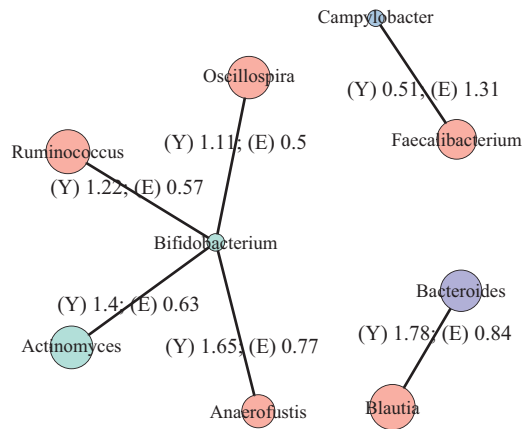


Fig. 3. Estimated differential microbial network between young and elderly individuals using our multiple testing procedure with  $FDR = 15\%$ , where isolated genera are not shown. Each node represents a genus coloured by phylum, and node size is proportional to the prevalence of 1s among all samples for the corresponding genus. Edge width is proportional to the magnitude of the entrywise test statistic. Turquoise circles, Actinobacteria; purple circles, Bacteroidetes; red circles, Firmicutes; blue circles, Proteobacteria.

between *Bacteroides* and *Blautia* were selected. Here the two values associated with each differential edge represent the estimated odds ratios between the two genera for the two age groups. For example, the odds ratio between *Faecalibacterium* and *Campylobacter* among young adults is 0.51, indicating that high abundance in *Faecalibacterium* is associated with lower odds of high abundance in *Campylobacter*. In other words, *Faecalibacterium* and *Campylobacter* have a competitive relationship within the microbial community, which could be due to their competition for space, nutrients, and so on. In contrast, the relationship between *Faecalibacterium* and *Campylobacter* is collaborative for the elderly group, since the odds ratio is 1.31. A collaborative relationship between two microbes can be due to cross-feeding, co-colonization or other reasons (Faust & Raes, 2012). Each pair of values associated with each edge shows a significant difference, suggesting changes in the microbial community structure as people age. Indeed, the abundance of *Faecalibacterium* has been found to be negatively associated with age (Franceschi et al., 2017). Further, *Ruminococcus* was significantly enriched in immune-mediated inflammatory diseases (Forbes et al., 2016), and *Oscillospira* was enriched in inflammatory diseases (Konikoff & Gophna, 2016). Thus *Ruminococcus* and *Oscillospira* may also play an important role in the ageing process, because age is characterized by chronic low-grade inflammation. Figure 3 provides additional evidence on how these genera are implicated in the ageing process. More importantly, genera involved in the differential network may be potential microbial targets that can be manipulated through dietary or medical interventions for healthy ageing. As the true differential network is unknown, further experimental validation is needed to confirm these results.

Since the method of Xia et al. (2015) does not perform well for discrete data, we applied their method to the centred log-ratio transformed relative abundance data and obtained a  $p$ -value of 0.007 for the global test, which is consistent with the conclusion from the proposed global test. At 15% false discovery rate, application of the multiple testing procedure of Xia et al. (2015) only gave one differential edge between *Holdemania* from the phylum Firmicutes and *Butyrivibrio* from Bacteroidetes. The difference is largely due to the fact that the centred log-ratio transformed data are far from being normally distributed.

## 6. DISCUSSION

To deal with excessive zeros and the unit sum constraint, in this paper we proposed using a binary Markov random field to study microbial interactions characterized by their conditional dependence relationships. Such models are robust and largely free of distributional assumptions on the data and have a natural interpretation in terms of co-existence or co-exclusiveness of the microbial communities. For very rare taxa, we suggest using a very small cut-off to discretize the data. For more common taxa, sample median or quartiles can be used. Using different cut-offs may lead to different results and thus to models with different interpretations. One may explore different ways of dichotomizing the data to see whether consistent results can be obtained.

The proposed method for testing binary Markov random fields can be applied to other data with binary observations, such as cancer somatic mutation data in cancer genomics or characterization of neural firing patterns and the reconstruction of neural connections in neuroscience. Our methods can be extended to study more general Markov random fields with each node taking more than two discrete values.

## ACKNOWLEDGEMENT

This research was supported by the U.S. National Institutes of Health and National Science Foundation and by the National Natural Science Foundation of China.

## SUPPLEMENTARY MATERIAL

Supplementary material available at *Biometrika* online includes technical lemmas and proofs.

## REFERENCES

- AITCHISON, J. (1982). The statistical analysis of compositional data (with discussion). *J. R. Statist. Soc. B* **44**, 139–77.
- BARABÁSI, A.-L. & ALBERT, R. (1999). Emergence of scaling in random networks. *Science* **286**, 509–12.
- BARBER, R. F. & DRTON, M. (2015). High-dimensional Ising model selection with Bayesian information criteria. *Electron. J. Statist.* **9**, 567–607.
- BIAGI, E., NYLUND, L., CANDELA, M., OSTAN, R., BUCCI, L., PINI, E., NIKKĪLA, J., MONTI, D., SATOKARI, R., FRANCESCHI, C. et al. (2010). Through ageing, and beyond: Gut microbiota and inflammatory status in seniors and centenarians. *PLoS One* **5**, 1–14.
- BISWAS, S., McDONALD, M., LUNDBERG, D. S., DANGL, J. L. & JOJIC, V. (2016). Learning microbial interaction networks from metagenomic count data. *J. Comp. Biol.* **23**, 526–35.
- CAI, T. T. & LIU, W. (2016). Large-scale multiple testing of correlations. *J. Am. Statist. Assoc.* **111**, 229–40.
- CAI, T. T., LIU, W. & XIA, Y. (2013). Two-sample covariance matrix testing and support recovery in high-dimensional and sparse settings. *J. Am. Statist. Assoc.* **108**, 265–77.
- CLAESSON, M. J., CUSACK, S., O’SULLIVAN, O., GREENE-DINIZ, R., DE WEERD, H., FLANNERY, E., MARCHESI, J. R., FALUSH, D., DINAN, T., FITZGERALD, G. et al. (2011). Composition, variability, and temporal stability of the intestinal microbiota of the elderly. *Proc. Nat. Acad. Sci.* **108**, 4586–91.
- ERDŐS, P. & RÉNYI, A. (1960). On the evolution of random graphs. *Pub. Math. Inst. Hungar. Acad. Sci.* **5**, 17–61.
- FANG, H., HUANG, C., ZHAO, H. & DENG, M. (2017). gCoda: Conditional dependence network inference for compositional data. *J. Comp. Biol.* **24**, 699–708.
- FAUST, K. & RAES, J. (2012). Microbial interactions: From networks to models. *Nat. Rev. Microbiol.* **10**, 538–50.
- FIRTH, D. (1993). Bias reduction of maximum likelihood estimates. *Biometrika* **80**, 27–38.
- FORBES, J. D., VAN DOMSELAAR, G. & BERNSTEIN, C. N. (2016). The gut microbiota in immune-mediated inflammatory diseases. *Front. Microbiol.* **7**, 1081.
- FRANCESCHI, C., GARAGNANI, P., VITALE, G., CAPRI, M. & SALVIOLI, S. (2017). Inflammaging and ‘garb-aging’. *Trends Endocrinol. Metab.* **28**, 199–212.
- FRIEDMAN, J. H., HASTIE, T. J. & TIBSHIRANI, R. J. (2010). Regularization paths for generalized linear models via coordinate descent. *J. Statist. Software* **33**, 1–22.



- GOODRICH, J. K., DAVENPORT, E. R., BEAUMONT, M., JACKSON, M. A., KNIGHT, R., OBER, C., SPECTOR, T. D., BELL, J. T., CLARK, A. G. & LEY, R. E. (2016). Genetic determinants of the gut microbiome in UK twins. *Cell Host & Microbe* **19**, 731–43.
- KONIKOFF, T. & GOPHNA, U. (2016). Oscillospira: A central, enigmatic component of the human gut microbiota. *Trends Microbiol.* **24**, 523–4.
- KUCZYNSKI, J., LAUBER, C. L., WALTERS, W. A., PARFREY, L. W., CLEMENTE, J. C., GEVERS, D. & KNIGHT, R. (2012). Experimental and analytical tools for studying the human microbiome. *Nat. Rev. Genet.* **13**, 47–58.
- KURTZ, Z. D., MÜLLER, C. L., MIRALDI, E. R., LITTMAN, D. R., BLASER, M. J. & BONNEAU, R. A. (2015). Sparse and compositionally robust inference of microbial ecological networks. *PLoS Comp. Biol.* **11**, 1–25.
- LECLERC, R. D. (2008). Survival of the sparsest: Robust gene networks are parsimonious. *Molec. Syst. Biol.* **4**, 213.
- LI, J. & CHEN, S. X. (2012). Two sample tests for high-dimensional covariance matrices. *Ann. Statist.* **40**, 908–40.
- R DEVELOPMENT CORE TEAM (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, <http://www.R-project.org>.
- RAVIKUMAR, P., WAINWRIGHT, M. J. & LAFFERTY, J. D. (2010). High-dimensional Ising model selection using  $\ell_1$ -regularized logistic regression. *Ann. Statist.* **38**, 1287–319.
- SANTHANAM, N. P. & WAINWRIGHT, M. J. (2012). Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Trans. Info. Theory* **58**, 4117–34.
- SCHOTT, J. R. (2007). A test for the equality of covariance matrices when the dimension is large relative to the sample sizes. *Comp. Statist. Data Anal.* **51**, 6535–42.
- VAN DE GEER, S., BÜHLMANN, P., RITOV, Y. & DEZEURE, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Ann. Statist.* **42**, 1166–202.
- WATTS, D. J. & STROGATZ, S. H. (1998). Collective dynamics of small-world networks. *Nature* **393**, 440–2.
- XIA, Y., CAI, T. & CAI, T. T. (2015). Testing differential networks with applications to the detection of gene-gene interactions. *Biometrika* **102**, 247–66.
- XIA, Y., CAI, T. & CAI, T. T. (2018). Multiple testing of submatrices of a precision matrix with applications to identification of between pathway interactions. *J. Am. Statist. Assoc.* **113**, 328–39.
- ZHANG, C.-H. & ZHANG, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *J. R. Statist. Soc. B* **76**, 217–42.

[Received on 11 November 2017. Editorial decision on 3 November 2018]