The Annals of Applied Statistics
2019, Vol. 13, No. 2, 931–957
https://doi.org/10.1214/18-AOAS1229

MODELING ASSOCIATION IN MICROBIAL COMMUNITIES WITH CLIQUE LOGLINEAR MODELS[1]

BY ADRIAN DOBRA[*,2], CAMILO VALDES[†,2], DRAGANA AJDIC[‡], BERTRAND CLARKE[§] AND JENNIFER CLARKE[§]

University of Washington, Florida International University*[†], *University of Miami*[‡] *and University of Nebraska-Lincoln*[§]

There is a growing awareness of the important roles that microbial communities play in complex biological processes. Modern investigation of these often uses next generation sequencing of metagenomic samples to determine community composition. We propose a statistical technique based on clique loglinear models and Bayes model averaging to identify microbial components in a metagenomic sample at various taxonomic levels that have significant associations. We describe the model class, a stochastic search technique for model selection, and the calculation of estimates of posterior probabilities of interest. We demonstrate our approach using data from the Human Microbiome Project and from a study of the skin microbiome in chronic wound healing. Our technique also identifies significant dependencies among microbial components as evidence of possible microbial syntrophy.

1. Introduction. Microbiomes—the communities of microorganisms peculiar to specific environments such as mammalian skin or managed agricultural soil—play key roles in a diverse set of biological phenomena, from plant growth to wine cultivation to human health and disease. Metagenomics is the study of genetic material recovered directly from a specific microbiome or environment without knowledge of the composition of the sample. Thus, metagenomic-based studies generate valuable information about the composition of microbiomes and differences in their composition that may be related to environmental differences. Traditionally, studying complex microbiome samples relied on intensive microbiological techniques involving the isolation and culturing of individual organisms followed by phenotypic or genotypic analysis. These techniques precluded microbial community profiling within a single sample. However, recent advances in high-throughput DNA sequencing technologies now permit whole-genome metagenomic sequencing (i.e., whole metagenome sequencing) without such isolation or

Received January 2018; revised November 2018.

[1]Supported in part by NSF Grants DMS/MPS-1737746 and DMS-1120255 to University of Washington, and Grants DMS-1410771 and DMS-1419754 to University of Nebraska-Lincoln.

[2]A. Dobra and C. Valdes contributed equally to this work and are joint first authors.

Key words and phrases. Contingency tables, graphical models, model selection, microbiome, next generation sequencing.

culturing. This means that characterization of complex microbial communities is now possible.

Whole metagenome sequencing (WMS) has served as the primary tool for several high profile, collaborative research endeavors such as the U.S. National Institute of Health Human Microbiome Project [Peterson et al. (2009)], the U.S. Department of Energy Joint Genome Institute's Integrated Microbial Genomes (IMG) system [Markowitz et al. (2014)] and the Canadian Institutes of Health Research Canadian Microbiome Initiative. Often, metagenome sequencing means that next generation sequencing (NGS) techniques are used. These techniques differ from classical Sanger sequencing in that instead of sequencing a whole DNA molecule nucleotide by nucleotide, the sequencing is done in parallel at many points of the DNA molecule resulting in short reads, or simply reads, typically ranging in length from 50 to 250 nucleotides. Usually, a key step in the analysis of NGS data is aligning the reads to a collection of consensus sequences or reference genomes for a collection of organisms. WMS is the general case for which our formal reasoning is designed; our examples use whole genome sequencing (WGS) and 16S sequencing. In WGS sequencing libraries are prepared from the extracted whole-DNA sequences of bacteria in the sample to be analyzed. The resulting sequencing short reads consequently represent the putative DNA sequences of the bacterial populations in the sample [Hasman et al. (2014), Thoendel et al. (2016)]. In contrast 16S rRNA sequencing libraries are prepared from the sequences of the highly conserved 16S ribosomal gene. The reads from 16S sequencing represent the sequences of the 16S genes in the bacterial populations in the sample. In downstream analyses WGS data are analyzed at the sequencing-read and genome levels, while 16S reads are assembled into clusters of reads called Operational Taxonomic Units (OTU), and analyzed as abstract representations of taxonomic groups [Charuvaka and Rangwala (2011), Nguyen et al. (2016)]. WGS captures a larger region of a bacterium's genome than 16S and can achieve better detection given the appropriate depth and breadth of sequencing coverage [Ranjan et al. (2016)].

It is well known that compositional studies of microbiomes alone provide no information about potential symbiosis or syntrophy—settings in which the metabolic waste products from one microbe provide nutrients for another—among species or strains [Levy and Borenstein (2013)]. Indeed, microbial communities in diverse settings have been shown to form syntrophic relationships. Such relationships have been posited to drive pathogenicity [de Kievit and Iglewski (2000), Koch et al. (2014)]. A simple approach to infer possible syntrophic relationships is to examine rates of co-occurrence of microorganisms in the same habitat across samples [Hoffmann et al. (2013)]. However, these methods cannot be used with a single sample as they rely on co-occurrence across many samples. In addition in most metagenomic studies based on sequencing, there is a portion of sequencing reads that cannot be associated with any known microorganisms in a particular environment, and these reads are often discarded inappropriately.

In this paper we address these limitations by proposing methods to identify bacteria and associations among them at various taxonomic levels (e.g., genus, species or strain). In WMS a collection of reads is sampled from a biological community within one sample. By aligning these reads to a database of microbial reference genomes, a categorical dataset showing the reference genomes to which each read aligns is obtained. In these data one row corresponds to one read, and one column corresponds to a genome—indicating the genomes to which each read maps. The rows are independent if the reads are from different organisms and often nearly independent even when they are from the same organism. Although initially counterintuitive this is seen empirically in a Bayesian context in Clarke et al. (2015). In fact assuming independence among a large number of reads is a reasonable first approximation because: (i) the number of nucleotides in the DNA molecules is very large so dependence will be rare, and (ii) even when reads are regarded as dependent this rests partially on their gene products. Here, we are not looking at gene products so dependence among them is irrelevant, making the dependence among reads smaller than one would initially expect.

We introduce a statistical approach based on a special class of loglinear models, which we call clique loglinear models as a tool for statistical inference about the presence of various strains, species and genera of bacteria and their associations within a given taxonomic level. Our goal is to assess associations among bacteria within a single sample (or across samples), and the likelihood of a specific bacterium, including a previously unknown bacterium, being in the sample. To represent these associations, we produce connectivity graphs showing which bacterial genera (or other taxonomic unit) are related by higher order interaction terms. This is possible because a clique loglinear model is a compound of disjoint collections of higher order terms, each collection permitting all possible interactions amongst the categories at the taxonomic level under study. Clique loglinear models are a sparse subset of all hierarchical loglinear models [Bishop, Fienberg and Holland (2007)], and this is operationally satisfactory since the associations among bacterial strains are often sparse as well. Stated in another way, the class of clique loglinear models is small enough to be tractable, yet large enough to be used for data summarization and model selection. Given the increasing speed of computing and accumulating knowledge about which bacteria are in which microbiome, this task is likely to be easier in the future than it is now.

In Section 2 we formally introduce clique loglinear models, describe their properties and develop model selection methods. In Section 3 we present a series of simulations to verify that our methodology qualitatively generates the results one would anticipate. In Sections 4 and 5 we analyze two datasets, and interpret our results in their scientific contexts. For the first of these our results are consistent with the findings from a more traditional approach to analysis of the same data. For the second we generate results that seem plausible given the experimental context; there is no previous analysis for comparison purposes. This shows that our modeling framework provides a viable alternative to expensive laboratory work. Finally,

in Section 6 we discuss how several features of our formalism relate to the real biological questions we have addressed.

2. Analyzing NGS data using clique loglinear models. In this section we motivate and outline our methodology for using metagenomic NGS data to detect associations among, say, bacterial strains or genera.

2.1. *Representing NGS data as a sparse contingency table.* Because of the way sequencing reads are generated, they can match none, one or several bacterial strains on a list $\{C_1, \ldots, C_B\}$ of known bacterial genomes. A sample of R reads can be represented as a $R \times B$ matrix $(c_{rb})_{RB}$ that we call a connectivity matrix, in which

$$c_{rb} = \begin{cases} 1, & \text{if read } r \text{ aligns to strain } b, \\ 0, & \text{if read } r \text{ does not align to strain } b. \end{cases}$$

Each row of $(c_{rb})_{RB}$ may be regarded as a vector valued outcome of the vector valued random variable $\mathbf{X}_{\mathcal{B}} = (X_1, \ldots, X_B)$ in which $X_b = X_b(r)$ is the indicator variable for a sampled read r to align (or match) to genome b. Each outcome of $\mathbf{X}_{\mathcal{B}}$ assumes one of 2^B patterns of zeroes and ones in $\mathcal{X}_{\mathcal{B}} = \{0, 1\}^B$. These vectors of length B generate a B-dimensional contingency table $\mathbf{n}_{\mathcal{B}}$ in which the count $\mathbf{n}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}})$ in cell $\mathbf{x}_{\mathcal{B}} \in \mathcal{X}_{\mathcal{B}}$ gives the number of reads that share the same pattern of alignments to the B genomes. Table 1 gives an example connectivity matrix.

We want to model the joint distribution of $\mathbf{X}_B$ to obtain estimates of interesting cell probabilities $\mathsf{p}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) = \mathsf{P}(\mathbf{X}_{\mathcal{B}} = \mathbf{x}_{\mathcal{B}})$ and relations amongst them. For instance

$$\mathsf{P}(X_1 = 0, \ldots, X_B = 0) \tag{2.1}$$

is the probability that a sampled read aligns to none of the B reference genomes. If the estimate of (2.1) is high, we might infer that we have found a bacterium or other microbial source not amongst the C_b's. By contrast

$$\mathsf{P}\big(X_{b^*} = 1, \{X_b = 0, \forall b \neq b^*\}\big) \tag{2.2}$$

TABLE 1

An example connectivity matrix. The first read only matches bacterial strain 1, *while reads* $2, \ldots, R$ *each match at least two strains*

Read	Genome 1	Genome 2	$\cdots$	Genome B
1	*Match*	No match	$\cdots$	No match
2	*Match*	*Match*	$\cdots$	No match
3	*Match*	No match	$\cdots$	*Match*
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\vdots$
R	*Match*	*Match*	$\cdots$	*Match*

is the probability that a sampled read comes from C_{b*} and does *not* come from any of the other $(B-1)$ genomes. To identify the bacteria that are most likely to be present, we choose the C_b's with the highest values of (2.2).

Once the data form a connectivity matrix, capturing associations in the resulting multiway contingency table can be done by determining models for the joint distributions of the observed categorical variables while recognizing that these random variables do not vary independently of each other. However, the association structure within a microbial community is likely to be sparse because most of the possible higher order interaction terms are likely to be discarded. This happens because, given the length of the list of reference genomes, most bacteria only occur jointly with a relatively small number of other bacteria. We argue that classes of hierarchical loglinear models [Bishop, Fienberg and Holland (2007)] are well suited to represent associations among bacterial taxa in a community.

2.2. *Clique loglinear models.* For a set $C \subseteq \mathcal{B}$, we denote $\mathcal{X}_C = \{0,1\}^{|C|}$, where $|C|$ stands for the number of elements of C. The subvector $\mathbf{X}_C$ of $\mathbf{X}_\mathcal{B}$ takes values $\mathbf{x}_C \in \mathcal{X}_C$. The C-marginal $\mathbf{n}_C$ of $\mathbf{n}_\mathcal{B}$ has cell counts $\mathbf{n}_C(\mathbf{x}_C) = \sum_{\mathbf{x}_{\mathcal{B}\setminus C}} \mathbf{n}_\mathcal{B}(\mathbf{x}_C, \mathbf{x}_{\mathcal{B}\setminus C})$. The corresponding marginal cell probabilities are $\mathsf{p}_C(\mathbf{x}_C) = \mathsf{P}(\mathbf{X}_C = \mathbf{x}_C)$.

Consider a hierarchical loglinear model M with k generators $\mathcal{C}(\mathsf{M}) = \{C_1, \ldots, C_k\}$, where $C_j \subseteq \mathcal{B}$, for $j = 1, \ldots, k$ and $k \geq 1$ [Bishop, Fienberg and Holland (2007), Edwards and Havránek (1985)]. Under this model the cell probabilities associated with $\mathbf{X}_\mathcal{B}$ are represented as [Whittaker (1990)]

$$\log \mathsf{p}_\mathcal{B}(\mathbf{x}_\mathcal{B}) = u_\varnothing + \sum_{\{C:\varnothing\neq C\subseteq C_j \text{ for some } j\in\{1,\ldots,k\}\}} u_C(\mathbf{x}_C). \tag{2.3}$$

Here, $u_\varnothing$ is an intercept, and $\{u_C(\mathbf{x}_C) : \mathbf{x}_C \in \mathcal{X}_C\}$ is the $|C|$-way interaction associated with the subvector $\mathbf{X}_C$ of $\mathbf{X}_B$. This model can be made identifiable either by imposing the sum to zero constraints $\sum_{\mathbf{x}_C\in\mathcal{X}_C} u_C(\mathbf{x}_C) = 0$ or by imposing the baseline equal with zero constraints that set $u_C(\mathbf{x}_C) = 0$ if one element of $\mathbf{x}_C$ is zero. For the latter, the loglinear expansion (2.3) becomes

$$\log \mathsf{p}_\mathcal{B}(\mathbf{x}_\mathcal{B}) = u_\varnothing + \sum_{\{C:\varnothing\neq C\subseteq C_j \text{ for some } j\in\{1,\ldots,k\}\}} u_C \prod_{i\in C} x_i, \tag{2.4}$$

where $u_C = u_C(1,\ldots,1)$.

A hierarchical loglinear model M is a clique loglinear model if its generators form a partition of $\mathcal{B}$: $\bigcup_{j=1}^k C_j = \mathcal{B}$, $C_{j_1} \cap C_{j_2} = \varnothing$ for $j_1 \neq j_2$. In this case the cell probabilities (2.4) are written as

$$\log \mathsf{p}_\mathcal{B}(\mathbf{x}_\mathcal{B}) = u_\varnothing + \sum_{j=1}^{k} \sum_{\{C:\varnothing\neq C\subseteq C_j\}} u_C \prod_{i\in C} x_i. \tag{2.5}$$

Thus, under a clique loglinear model the log cell probabilities are decomposed as a sum of groups of interaction terms in which each group represents a collection of categorical variables that may interact with each other in all possible ways but do not interact at all with categorical variables in other groups. Formally, the interpretation of clique loglinear models comes from this result:

PROPOSITION 2.1. *Let D_1 and D_2 be two subsets of $\mathcal{B}$ that are also subsets of two different generators of a clique loglinear model* M. *Then the random subvectors $\mathbf{X}_{D_1}$ and $\mathbf{X}_{D_2}$ are independent.*

PROOF. We collapse across the levels of $\mathbf{X}_{\mathcal{B}\setminus(D_1\cup D_2)}$ in the loglinear expansion (2.5). The marginal cell probabilities associated with $\mathbf{X}_{D_1\cup D_2}$ have the form

$$\log \mathsf{p}_{D_1\cup D_2}(\mathbf{x}_{D_1}, \mathbf{x}_{D_2}) = u_\varnothing + \sum_{\{C:\varnothing\neq C\subseteq D_1\}} u_C \prod_{i\in C} x_i + \sum_{\{C:\varnothing\neq C\subseteq D_2\}} u_C \prod_{i\in C} x_i.$$

Since the first term is a constant, the second term is a function of the levels of $\mathbf{X}_{D_1}$, and the third term is a function of the levels of $\mathbf{X}_{D_2}$, it follows that $\mathbf{X}_{D_1}$ and $\mathbf{X}_{D_2}$ are indeed independent. □

A consequence of Proposition 2.1 is that the cell probabilities of M decompose as a product of marginal cell probabilities associated with its generators:

$$\mathsf{p}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) = \prod_{j=1}^{k} \mathsf{p}_{C_j}(\mathbf{x}_{C_j}). \tag{2.6}$$

We denote by $\mathbf{u}_{\mathsf{M}}$ all the interaction terms that appear in (2.5). Under Multinomial sampling the log-likelihood function is written as a function of the interaction terms as follows:

$$l(\mathbf{u}_{\mathsf{M}}, \mathbf{n}_{\mathcal{B}}) = R u_\varnothing + \sum_{j=1}^{k} \sum_{\{C:\varnothing\neq C\subseteq C_j\}} u_C \mathbf{n}_C(\mathbf{x}_C) \prod_{i\in C} x_i. \tag{2.7}$$

By using Lagrange multipliers in (2.7) and (2.6) [Whittaker (1990)], it can be shown that the MLEs of the cell probabilities under M are

$$\widehat{\mathsf{p}}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) = R^{-k} \prod_{j=1}^{k} \mathbf{n}_{C_j}(\mathbf{x}_{C_j}). \tag{2.8}$$

Equation (2.8) shows that the MLEs of the cell probabilities of a clique loglinear model exist if and only if the counts in the marginal tables associated with its generators are strictly positive. This existence criterion is easily applicable in a computationally efficient manner. By contrast determining the existence of the MLEs for arbitrary hierarchical loglinear models is a difficult problem that has been solved theoretically [Fienberg and Rinaldo (2007)]. However, at the present

time there do not seem to exist any implementable algorithms for assessing the existence of MLEs of hierarchical loglinear models that are also computationally efficient when the number B of categorical variables involved is large.

We named this class of loglinear models based on the representation of the interaction structure defined by the u-terms $\mathbf{u}_{\mathsf{M}}$ as an independence graph [Whittaker (1990)]. This is an undirected graph G with vertices $\mathcal{B}$ and set of edges E. Each element $b \in \mathcal{B}$ is associated with the component X_b of $\mathbf{X}_{\mathcal{B}}$. An edge $e = (b_1, b_2)$ is included in E if there is a generator C_j of M such that $\{b_1, b_2\} \subseteq C_j$. Proposition 2.1 implies that the independence graph G of a clique loglinear model M has a special structure; the generators of M are the connected components of G and are also maximal complete subgraphs or cliques [Lauritzen (1996)]. As such, the independence graph of a clique loglinear model is obtained by putting together complete subgraphs without adding any edge between them. These cliques are the generators of the loglinear model, and uniquely identify it.

The class of clique loglinear models is a subset of decomposable loglinear models which, in turn, is a subset of graphical loglinear models that are themselves a subclass of hierarchical models; see the Supplementary Material [Dobra et al. (2019)], Section 3.1. The restriction to clique loglinear models offers key computational advantages. In addition to an easy way to calculate the MLEs and check their existence, these models are straightforward to interpret (Proposition 2.1) and allow the development of computationally efficient model determination algorithms that scale well when R or B become large.

The number of clique loglinear models for B categorical variables is the number of decompositions of B into positive integers [Abramowitz and Stegun (1972)]:

$$\mathcal{P}(B) = \frac{1}{\pi\sqrt{2}} \sum_{j=1}^{\infty} \sqrt{j} A_j(B) \frac{\mathrm{d}}{\mathrm{d}B} \frac{\sinh\left(\frac{\pi}{j}\sqrt{\frac{2}{3}\left(B - \frac{1}{24}\right)}\right)}{\sqrt{B - \frac{1}{24}}}, \tag{2.9}$$

where

$$A_j(B) = \sum_{0<h\le j,(h,j)=1} e^{\pi i\left(s(h,j) - \frac{2hB}{j}\right)}, \qquad s(h,j) = \sum_{l=1}^{j-1} \frac{l}{j}\left(\left(\frac{hl}{j}\right)\right),$$

with $((x)) = x - [x] - \frac{1}{2}$ if x is an integer and 0 otherwise, and (h, j) is the greatest common divisor of h and j. For example $\mathcal{P}(100) = 190{,}569{,}292$, $\mathcal{P}(200) \approx 3.973e + 12$ and $\mathcal{P}(1000) \approx 1.321e + 19$ [Hankin (2006)]. Therefore, although this is the smallest class of hierarchical loglinear models, it still contains a significantly large number of possible models that allow modeling various patterns of interactions among many categorical variables.

2.3. *Existing loglinear model selection methods.* Selection for loglinear models has been well studied in the literature [Edwards and Havránek (1985),

Whittaker (1990)]. Since large values of B arise naturally in metagenomics, answering questions about interactions within taxonomic levels with loglinear models must involve methods for model selection in high-dimensional contingency tables. The sparsity of these tables is extremely problematic as it leads to the invalidation of asymptotic approximations to the null distribution of the generalized likelihood ratio test statistic [Fienberg and Rinaldo (2007)]. Another key difficulty is the size of the space of possible loglinear models. For example, when $B = 5$ the number of possible hierarchical loglinear models is 7580; for $B = 8$ variables this number increases to 5.6×10^{22} [Dellaportas and Forster (1999)].

Because the space of possible models is extremely large, various stochastic search schemes have been used to identify models with high posterior probability. Dellaportas and Forster (1999) is a key reference, although there are other papers that develop stochastic search schemes for discrete data [Dellaportas and Tarantola (2005), Dobra and Massam (2010), Madigan and Raftery (1994), Madigan and York (1995, 1997), Tarantola (2004)]. One feature of these and other stochastic searches on spaces of hierarchical, graphical and decomposable loglinear models (see, e.g., Massam, Liu and Dobra (2009)) is that these models involve repeated transitions from one model to another model. This necessitates ensuring the next model is still in the target space of models. For instance, for decomposable graphs transitioning from a current decomposable graph to another decomposable graph involves checks that the decomposability property is preserved. While such checks can be done relatively quickly for graphs with few vertices, for graphs that involve hundreds of vertices the running time of stochastic searches increases rapidly. Since considerable computational effort is required to visit loglinear models sequentially by adding and removing higher order terms, restricting the model space to, say, clique loglinear models provides a necessary reduction in the running time of model determination algorithms. This makes the required computations intensive yet feasible.

Copula Gaussian graphical models [Dobra and Lenkoski (2011)] have successfully been used to analyze a 16-dimensional table. Several related methods based on efficiently determining Gaussian graphical models with many variables in a latent space have also been subsequently proposed [Mohammadi et al. (2017), Mukherjee and Rodriguez (2016)]. However, conditional independence relationships in a latent space do not necessarily translate into similar relationships in the observed discrete variables space. For this reason inferring multivariate interactions from latent Gaussian graphical models has limited practical relevance.

Dirichlet process mixture models have recently emerged in the analysis of categorical data. Canale and Dunson (2011) developed Bayesian nonparametric kernel mixtures for multivariate count data. Ultra-sparse high-dimensional contingency tables have been analyzed using probabilistic low rank tensor factorizations induced through a Dirichlet process mixture model of product multinomial distributions [Bhattacharya and Dunson (2012), Dunson and Xing (2009), Kunihama and Dunson (2013), Zhou et al. (2015)]. These papers present simulation studies

and real-world data examples that involve up to 172 categorical variables [Zhou et al. (2016)]. While promising in terms of their computational efficiency, Bayesian sparse tensor factorization methods do not easily translate into interpretable multivariate associations. In fact, although Johndrow, Bhattacharya and Dunson (2017) provided bounds on the tensor ranks for sparse weakly hierarchical loglinear models, it is still unclear whether low rank tensor factorizations exist for certain classes of sparse loglinear models. Interpreting low rank tensor factorization models for categorical data does not seem to be straightforward. On the other hand, clique loglinear models are easily interpretable. This represents the key advantage of this class of models over sparse tensor factorization methods.

2.4. *A stochastic search method for clique loglinear models.* Let M vary over the collection $\mathcal{M}$ of B-dimensional clique loglinear models for which the MLEs exist. We want to find M's that fit the data well and are parsimonious. For this purpose we develop a stochastic search procedure based on the Bayesian information criterion (BIC). For large sample sizes it is well known that the BIC is an approximation to the mode of a posterior distribution over a model space. The BIC is also optimal in a Bayes testing sense [Schwarz (1978)]. The calculation of BIC for clique loglinear models proceeds as follows. Denote by $\mathcal{C}(\mathsf{M}) = \{C_1, \ldots, C_k\}$ the generators of M. The MLEs of the mean cell values under M are calculated based on (2.8),

$$\log \widehat{m}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) = \log\bigl(R\widehat{p}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}})\bigr) = \sum_{j=1}^{k} \log \mathbf{n}_{C_j}(\mathbf{x}_{C_j}) - (k-1)\log R, \tag{2.10}$$

for all $\mathbf{x}_{\mathcal{B}} \in \mathcal{X}_{\mathcal{B}}$. From (2.5) we see that the number of free interaction terms that appear in M is equal with the sum of the number of nonempty subsets of the generators of M. Therefore, the BIC of M is given by

$$\begin{aligned}\mathrm{BIC}(\mathsf{M}) = {} & -2 \sum_{\{\mathbf{x}_{\mathcal{B}} \in \mathcal{X}_{\mathcal{B}} : \mathbf{n}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) > 0\}} \mathbf{n}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) \log \widehat{m}_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) \\ & + \left(\sum_{j=1}^{k} 2^{|C_j|} - k + 1\right) \log R. \end{aligned} \tag{2.11}$$

Equations (2.10) and (2.11) show that the BIC of a clique loglinear model can be efficiently calculated even for large contingency tables since no iterative numerical optimization methods are involved as it would have been the case for arbitrary graphical and hierarchical loglinear models. The calculation of the log mean cell values can also be performed using a formula for decomposable loglinear models [Lauritzen (1996)], but the calculation of the number of free interaction terms of these models would have been complicated by their overlapping sets of generators. For this reason the calculation of BIC for clique loglinear models is easier as compared to any other loglinear model that does not belong to this class.

Consider the following distribution over $\mathcal{M}$:

$$\pi(\mathsf{M}) \propto \exp\bigl(-\mathrm{BIC}(\mathsf{M})\bigr). \tag{2.12}$$

Finding clique loglinear models with smaller values of BIC is equivalent to finding models at or close to the modes of the distribution (2.12). We can think of $\pi(\mathsf{M})$ as a posterior distribution over $\mathcal{M}$ obtained by assuming a flat prior over $\mathcal{M}$. Thus, the $\pi(\mathsf{M})$'s can be considered to be the Bayes model weights, and, in the sequel, these weights will be used to perform model averaging using Occam's window. This methodology originates in Madigan and Raftery (1994) and has been developed in numerous other contexts, for example, dynamic linear models [Onorante and Raftery (2016)], "large p" regression [Dobra (2009), Hans, Dobra and West (2007)] and graphical models [Dobra and Massam (2010), Lenkoski and Dobra (2011), Madigan, Gavrin and Raftery (1995)].

Our goal is to find clique loglinear models that have large posterior weights (2.12). The largest would achieve

$$\widehat{\mathsf{M}}_{\mathsf{opt}} = \arg\max_{\mathcal{M}} \pi(\mathsf{M}).$$

However, models that have posterior weights comparable to that of the optimal model $\widehat{\mathsf{M}}_{\mathsf{opt}}$ are also relevant. The stochastic search algorithm we propose below is devised to seek the set of models

$$\mathcal{S}(c) = \bigl\{\mathsf{M} \in \mathcal{M} : \pi(\mathsf{M}) \geq c\pi(\widehat{\mathsf{M}}_{\mathsf{opt}})\bigr\}, \tag{2.13}$$

where $c \in (0, 1)$ is a constant that needs to be specified before the start of the algorithm. The clique loglinear models that do not belong to $\mathcal{S}(c)$ are discarded. The idea of eliminating models with low posterior probability compared to the highest posterior probability model is based on the Occam's window principle of Madigan and Raftery (1994).

For ease of exposition we begin by stating our procedure informally. Our stochastic search procedure moves toward models with larger values of $\pi(\mathsf{M})$. The models that are visited in a run are collected as if in a bag. Each run of the stochastic search algorithm collects models until it appears to reach a local optimum. At that point the stochastic search algorithm will likely visit only models that are already in the bag. We use many different runs, and combine all the bags of models collected in each run into a larger bag $\mathcal{S}$. Out of this bag we only retain those models that have comparable posterior weights with the best model identified across all runs:

$$\widehat{\mathcal{S}}(c) = \Bigl\{\mathsf{M} \in \mathcal{S} : \pi(\mathsf{M}) \geq c \max_{\mathsf{M}\in\mathcal{S}} \pi(\mathsf{M})\Bigr\}. \tag{2.14}$$

Across multiple runs that were sufficiently long, we would hope that $\widehat{\mathcal{S}}(c)$ from (2.14) will approximate well $\mathcal{S}(c)$ in (2.13). This is very likely to happen if $\widehat{\mathsf{M}}_{\mathsf{opt}}$ has been visited and included in $\mathcal{S}$. An empirical test for figuring out whether $\widehat{\mathsf{M}}_{\mathsf{opt}}$ was indeed identified is to determine the proportion of runs that reached

$\arg\max_{\mathcal{S}} \pi(\mathsf{M})$. A high proportion of runs that ended up visiting the best model in $\mathcal{S}$ represents a good indication that $\widehat{\mathsf{M}}_{\mathsf{opt}}$ might indeed be in $\mathcal{S}$. In the sequel we perform Bayes model averaging using the models in $\widehat{\mathcal{S}}(c)$ with weights in (2.12), and this lets us estimate the quantities of interest. Models not in $\widehat{\mathcal{S}}(c)$ are discarded; this is justified if $\widehat{\mathcal{S}}(c)$ comprises most models that have large posterior probabilities.

Our stochastic algorithm for identifying $\mathcal{S}(c)$ from (2.13) proceeds as follows. We start with a randomly generated clique loglinear model. If any of the marginals associated with the generators of this model contain counts of zero, the MLEs of this model do not exist and another random model is generated. We repeat these steps until a valid clique loglinear model is generated; we denote this model with M_0. Starting with M_0, we generate a chain of models $\langle \mathsf{M}_t \rangle$ for $t = 1, 2, \ldots$. At step t, with equal probability, we select one of the following four ways of producing a valid (i.e., for which the MLEs (2.8) exist) candidate clique loglinear model M':

(i) Split a random clique of M_t into two cliques.
(ii) Join two random cliques of M_t into a clique.
(iii) Switch two random elements that belong to two random cliques of M_t.
(iv) Move a random element of a random clique of M_t to another random clique of M_t.

After sampling a move of type (i), (ii), (iii) or (iv), we produce a clique loglinear model M' by applying a move of that type to model M_t. For moves of type (i), we uniformly select a clique of M_t and divide the elements in that clique $\mathcal{C}'$ into two disjoint sets $\mathcal{C}'_1$ and $\mathcal{C}'_2$ that become two new cliques of M'. The other cliques of M_t are also cliques for M'. For moves of type (ii), we uniformly sample two cliques $\mathcal{C}'_1$ and $\mathcal{C}'_2$ of M_t and form a new clique $\mathcal{C}' = \mathcal{C}'_1 \cup \mathcal{C}'_2$. The other cliques of M_t together with $\mathcal{C}'$ are the cliques of M'. For moves of type (iii), we uniformly sample two cliques $\mathcal{C}'_1$ and $\mathcal{C}'_2$ of M_t and also uniformly sample a element $v_1 \in \mathcal{C}'_1$ and a element $v_2 \in \mathcal{C}'_2$. We form two new cliques $\mathcal{C}''_1 = \mathcal{C}'_1 \setminus \{v_1\} \cup \{v_2\}$ and $\mathcal{C}''_2 = \mathcal{C}'_2 \setminus \{v_2\} \cup \{v_1\}$. The cliques $\mathcal{C}''_1, \mathcal{C}''_2$ together with the other cliques of M_t give the candidate model M'. For moves of type (iv), we uniformly sample two cliques $\mathcal{C}'_1$ and $\mathcal{C}'_2$ of M_t and also uniformly sample a element $v_1 \in \mathcal{C}'_1$. We form two new cliques $\mathcal{C}''_1 = \mathcal{C}'_1 \setminus \{v_1\}$ and $\mathcal{C}''_2 = \mathcal{C}'_2 \cup \{v_1\}$. The cliques $\mathcal{C}''_1, \mathcal{C}''_2$ together with the other cliques of M_t give the candidate model M'.

If the MLEs of M' do not exist, we set $\mathsf{M}_{t+1} = \mathsf{M}_t$. If the MLEs of M' exist, we set $\mathsf{M}_{t+1} = \mathsf{M}'$ with probability $\min\{1, \pi(\mathsf{M}')/\pi(\mathsf{M}_t)\}$. Otherwise we set $\mathsf{M}_{t+1} = \mathsf{M}_t$. This stochastic search algorithm typically moves to models with larger $\pi(\mathsf{M})$. If the sampled candidate model M' happens to have a smaller $\pi(\mathsf{M})$ than the current model M_t, the algorithm could still visit it with positive probability. This is useful because sometimes models with smaller $\pi(\mathsf{M})$ must be visited before finding models with larger $\pi(\mathsf{M})$. This is the case when M is a local maximum but not a global maximum. The geometry of the space of models affects this: getting stuck in a local maximum is not a problem if it is a global maximum; on the other hand,

the model space is discrete, so it is possible that the models with the largest $\pi(\mathsf{M})$ are not very similar to each other.

Only moves of type (i) and (ii) are needed to connect a clique loglinear model for which the MLE exists with any other clique loglinear model in $\mathcal{M}$. However, we empirically found that an algorithm that included moves of types (iii) and (iv) was less likely to get stuck in local maxima of $\pi(\cdot)$. We note that this is a stochastic search procedure, not a Markov Chain Monte Carlo procedure so that the acceptance probabilities are not relevant; see the Supplementary Material [Dobra et al. (2019)], Section 3.2 for a discussion of this point. Furthermore, sparse contingency tables such as we are studying here frequently have unbalanced counts. This is rarely a problem for the BIC, as discussed in the Supplementary Material [Dobra et al. (2019)], Section 3.3.

2.5. *Bayesian model selection and inference for clique loglinear models.* As we will see in the simulated and real world data analysis examples, selecting clique loglinear models based on the BIC leads to results that are easily interpretable and are obtained with a low to moderate computational effort. However, there are several shortcomings related to the use of the BIC. First, the BIC limits the candidate clique loglinear models to those models for which the MLEs exist. Second, although asymptotically approximating posterior distributions using the BIC is a well-established technique [Berger, Ghosh and Mukhopadhyay (2003)], for sparse contingency tables the assumed limiting behavior might not occur. Third, the BIC is equivalent to using a uniform prior over the model space. In some applications, employing other priors over the model space could be desirable. For these reasons, in this section we describe a full Bayesian framework for model selection and inference.

In the sequel we follow Dawid and Lauritzen (1993). We start with a fictive prior table of positive numbers $\mathbf{n}^0_{\mathcal{B}}$. A default specification for this prior table involves setting its grand total (the sum of its cell counts or sample size) to a value $\alpha > 0$ much smaller than the grand total R of the observed table $\mathbf{n}_{\mathcal{B}}$ (e.g., $\alpha = 1$), then setting all its cells to an equal value: $\mathbf{n}^0_{\mathcal{B}}(\mathbf{x}_{\mathcal{B}}) = \alpha/|\mathcal{B}|$, for all $\mathbf{x}_{\mathcal{B}} \in \mathcal{X}_{\mathcal{B}}$. The effect of the choice of the values of α on the loglinear models selected based on Bayes factors has been studied empirically in Massam, Liu and Dobra (2009) and theoretically from a geometrical perspective in Letac and Massam (2012). These papers found that, for larger values of α, more interaction terms appear in the hierarchical loglinear models with the largest posterior probabilities. When α becomes smaller with values close to 0, the hierarchical loglinear models selected contain fewer interaction terms that involve a smaller number of variables.

Consider a clique loglinear model M with generators $\mathcal{C}(\mathsf{M}) = \{C_1, \ldots, C_k\}$. For each generator $C \in \mathcal{C}(\mathsf{M})$ of M, we let $\mathsf{Dir}(\mathbf{n}^0_C)$ denote the Dirichlet distribution for the marginal cell probabilities p_C:

$$\mathsf{P}(\mathsf{p}_C \mid \mathbf{n}^0_C) \propto \prod_{\mathbf{x}_C \in \mathcal{X}_C} \mathsf{p}_C(\mathbf{x}_C)^{\mathbf{n}^0_C(\mathbf{x}_C)-1}. \tag{2.15}$$

Since the generators of M do not overlap, the Dirichlet priors (2.15) are pairwise hyperconsistent and define a unique prior for the cell probabilities $\mathsf{p}_{\mathcal{B}}$ under model M:

$$\mathsf{P}(\mathsf{p}_{\mathcal{B}} \mid \mathbf{n}^0_{\mathcal{B}}) = \prod_{j=1}^{k} \mathsf{P}(\mathsf{p}_{C_j} \mid \mathbf{n}^0_{C_j}). \tag{2.16}$$

This prior is called the hyper Dirichlet prior for $\mathsf{p}_{\mathcal{B}}$, and it is denoted by $\mathsf{HyperDir}_{\mathsf{M}}(\mathbf{n}^0_{\mathcal{B}})$. The hyper Dirichlet prior is conjugate for the multinomial likelihood and yields a posterior distribution for the cell probabilities

$$\mathsf{P}(\mathsf{p}_{\mathcal{B}} \mid \mathbf{n}_{\mathcal{B}}) \propto \mathsf{P}(\mathbf{n}_{\mathcal{B}} \mid \mathsf{p}_{\mathcal{B}})\mathsf{P}(\mathsf{p}_{\mathcal{B}} \mid \mathbf{n}^0_{\mathcal{B}}),$$

that is, hyper Dirichlet $\mathsf{HyperDir}_{\mathsf{M}}(\mathbf{n}^*_{\mathcal{B}})$, where $\mathbf{n}^*_{\mathcal{B}} = \mathbf{n}_{\mathcal{B}} + \mathbf{n}^0_{\mathcal{B}}$. The marginal likelihood under model M is

$$\mathsf{P}(\mathbf{n}_{\mathcal{B}} \mid \mathsf{M}, \mathbf{n}^0_{\mathcal{B}}) = \prod_{j=1}^{k} \mathsf{P}(\mathbf{n}_{C_j} \mid \mathbf{n}^0_{C_j}), \tag{2.17}$$

where

$$\mathsf{P}(\mathbf{n}_C \mid \mathbf{n}^0_C) = \frac{\Gamma(\alpha)}{\Gamma(\alpha + R)} \prod_{\mathbf{x}_C \in \mathcal{X}_C} \frac{\Gamma(\mathbf{n}^*_C(\mathbf{x}_C))}{\Gamma(\mathbf{n}^0_C(\mathbf{x}_C))},$$

for each $C \in \mathcal{C}(\mathsf{M})$. The posterior distribution of M is

$$\mathsf{P}(\mathsf{M} \mid \mathbf{n}_{\mathcal{B}}) \propto \mathsf{P}(\mathbf{n}_{\mathcal{B}} \mid \mathsf{M}, \mathbf{n}^0_{\mathcal{B}})\mathsf{P}(\mathsf{M}), \tag{2.18}$$

where $\mathsf{P}(M)$ is a prior distribution on the set of all clique loglinear models. A prior on this space that penalizes for model complexity is [Jones et al. (2005)]

$$\mathsf{P}(\mathsf{M}) \propto \prod_{j=1}^{k} \left(\frac{\beta}{1-\beta}\right)^{\binom{|C_j|}{2}}, \tag{2.19}$$

where $|C|$ denotes the number of elements in the set C, and $\beta \in (0, 1)$ represents the probability of including an additional edge in the corresponding independence graph. A related prior with desirable multiplicity correction properties is obtained by marginalizing out β in (2.19) with respect to a $\mathsf{Beta}(a, b)$ distribution [Carvalho and Scott (2009)]. A more general type of priors on the models space is defined through product distributions for random partitions [Barry and Hartigan (1992)]:

$$\mathsf{P}(\mathsf{M}) \propto \prod_{j=1}^{k} q(C_j), \tag{2.20}$$

where $q(C)$ is a cohesion function for subsets $C \subseteq \mathcal{B}$. Priors (2.20) exploit the special structure of clique loglinear models that partition variables into cliques. The stochastic search method from Section 2.4 can be used to identify clique loglinear models with high posterior probabilities. For this purpose the distribution (2.12) needs to be replaced by the posterior distribution (2.18). The rest of the model search procedure remains unchanged.

3. Simulation results. To benchmark the performance of the method, we created a synthetic experiment with a known community dependency structure. We obtained 2273 bacterial genomes from the National Center for Biotechnology Information (NCBI) GenBank database. These genomes were collected from GenBank's complete genome set, or those genomes that are considered to have a final DNA sequence per genomic structure (chromosomes and/or plasmids). From these 2273 genomes we randomly chose 200 genomes and created a population connectivity matrix representing 1000 synthetic genomic reads that indicates the connectivity among the genomes. Each simulated read has a corresponding row in the connectivity matrix with a match for at least one genome; this is indicated by 1s. The matrix is based on a file supplied to the simulation program that indicates which genomes are present and what cliques they form. If two genomes are in the same clique, they are given 1s for 80% of their joint reads (as assigned by i.i.d. Bernoulli(0.8) random variables). The remaining cells in the 1000×200 connectivity matrix are randomly filled with 0s and 1s sampled from a Bernoulli(0.2) distribution. This procedure gives a connectivity matrix consistent with a chosen clique structure on genomes. Note that not all 200 genomes are shown because only some were in nontrivial cliques. Further details are given in the Supplementary Material [Dobra et al. (2019)], Section 4.

We use the connectivity matrix to generate a connectivity graph. A connectivity graph has vertices that represent distinct organisms and edges that represent higher order interactions between their reads. This definition will be made more precise in Section 4 when we deal with real data. The connectivity graph for the synthetic reads is in Figure 1. We verified that two genomes are connected in Figure 1 if and only if they are connected by reads in the connectivity matrix.

To check that our method is able to recover the connectivity graph from Figure 1, we applied the stochastic search method described in Section 2.4 using 200 chains each of length 200,000. We set $c = 10^{-4}$ in (2.14). We calculated BMA estimates of the posterior inclusion probabilities of edges in the corresponding independence graph based on the models in the set $\widehat{\mathcal{S}}(c)$. This generates a connectivity graph that is shown in Figure 1. We comment that there is nothing unique about the value 10^{-4}. It was chosen for convenience, was not discredited by the individual posterior probabilities we found and a sensitivity analysis showed that it was a reasonable choice within a range of possible cutoff values. In practice the choice of cutoff value would be data driven to ensure appropriate robustness of the inferences.

In Figure 1 two genomes are connected if and only if the sum of the posterior weights of the best models containing higher order terms between the two genomes is above 0.1. Loosely, this is intuitively equivalent to saying that the posterior probability that the two genomes are associated (in the sense of higher order interaction terms in loglinear models) is at least 0.1. We see that true connectivity structure has been fully recovered by the clique loglinear models, but the clique loglinear

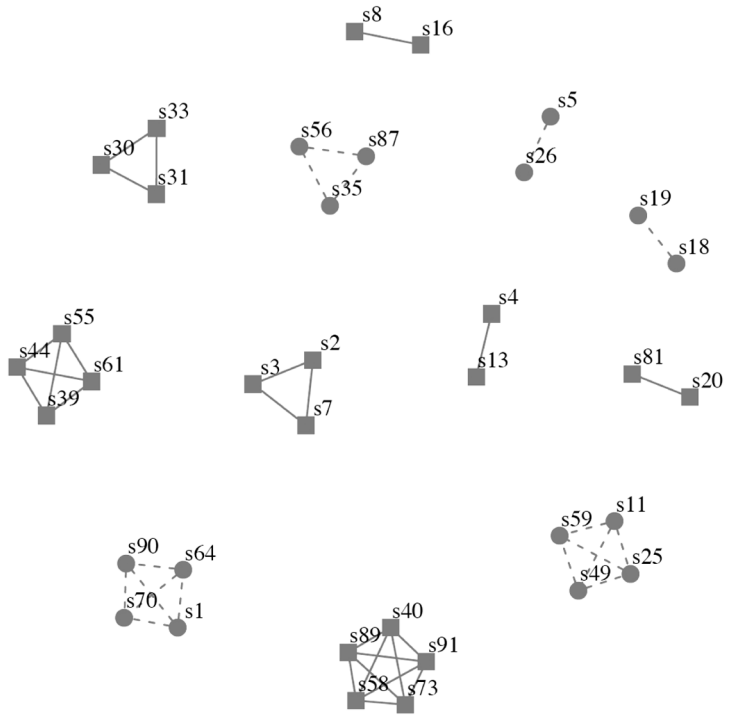

FIG. 1. *Actual and recovered connectivities in the simulated data. The connectivities that were recovered by the clique loglinear models but were absent in the true connectivity graph are marked with dotted edges and round vertices. The connectivities marked with solid edges and square vertices give the true connectivity graph.*

models identified additional interactions that were absent from the true connectivity graph. So, in this simple case our method returns the full set of cliques built into the data. Nevertheless, even though clique loglinear models are a restricted class of loglinear models, they can over detect interactions. The reason is that, in the simulated connectivity matrix, the cliques that are present are strongly built into the matrix; they will be found as long as enough reads are included. However, cliques that are not present may also be found by our method from the random 1s in the connectivity matrix that do not correspond to genomes in any predefined cliques.

The average computation time for calculating the BIC of a clique loglinear model for 200 genomes and 1000 reads was 0.26 seconds. Our implementation in plain R is slower due to the limits imposed by this software package and could be made significantly faster if implemented in C. For a discussion of comparison with other methods, please see Section 3.5 in the Supplementary Material [Dobra et al. (2019)].

4. Example: Characterizing associations in a microbial community. The Human Microbiome Project (HMP) is an ongoing collaborative study funded by the U.S. National Institutes of Health (NIH) to provide data and tools for studying the role of human microbiomes in human health and disease. Started in 2007 it has generated ground-breaking publications [Fierer et al. (2010), Minot et al. (2013), Zhao et al. (2012)] and a plethora of metagenomic data on human microbiomes. Our method from Section 2.4 can represent the associations from an HMP sample with an independence graph so we can infer the bacterial taxa present and their associations.

Human metagenome sample SRS015072, obtained from the vaginal microbiome of a female participant of the HMP Core Microbiome Sampling Protocol A (HMP-A) dbGaP study, was downloaded via FTP from the HMP Data Analysis and Coordination Center (DACC). The sample consisted of 495,256 paired-end, 100 base pair reads (with an average mate-distance of 81bp) sequenced and provided in Illumina FASTQ format. These reads were aligned to the collection of 4940 bacterial genomes, from the Integrated Microbial Genomes and Metagenomes (IMG, version 4.0) database [Markowitz et al. (2014)] using the Bowtie2 aligner [Langmead and Salzberg (2012)]. Of the sample reads 369,633 aligned to one or more of the reference genomes. The number of reads that aligned to each bacterial strain, species and genera was calculated and connectivity tables were generated for analysis at the genera level.

The first step in each analysis is to identify those genera that cannot be involved in higher order interactions (i.e., cannot be part of a clique with two or more vertices). Note any two genera that define a marginal two-by-two table (disregarding all other genera) whose counts are not strictly positive cannot be part of the same clique because the MLEs of any clique loglinear model that involves that two-way interaction do not exist. We refine the definition of a connectivity graph as follows: it is a graph whose vertices correspond to categorical variables, that is, the presence of a genus. Given two vertices, there is an edge joining them if the two-way marginal contingency table associated with the two categorical variables contains only strictly positive counts. Within each analysis we ran the stochastic search from Section 2.4 for 100,000 iterations from 100 random starting clique graphs.

A total of 95 genera had component species or strains to which reads aligned. Two genera were said to be connected if and only if each had at least one strain that shared a read. The genera with shared reads are shown in Figure 2. It is seen that 15 genera did not share reads across other genera (though each did within its own genus). This shows two facts: (i) 15 genera can be dropped from subsequent analysis at the genus level; and (ii) the hairball showing the 80 genera that share at least one read is complex enough that further analysis is worth doing, that is, it is worthwhile to use clique loglinear models to seek higher order interaction terms.

Table 1 from Supplementary Material [Dobra et al. (2019)] gives the degree (the number of neighbors) of each element (genus) in the raw connectivity graph. Several of these genera, including Lactobacillus, Prevotella and Staphylococcus,

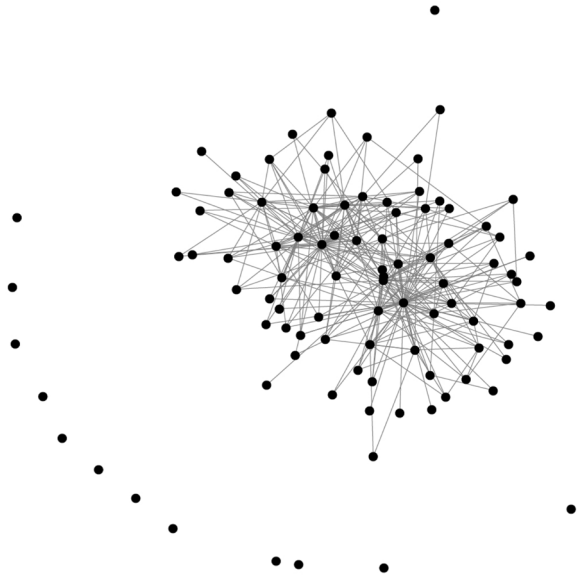

FIG. 2. *Genera connectivity graph for the example in Section 4.*

have been identified as common members of vaginal microbiome communities [Huang et al. (2014)]. Note that the 15 isolated genera in Figure 2 cannot form cliques with any of the other genera because in the reduced table, that is, grouping all strains into genera, the two-way marginals they form contain two counts of zero. As such, they cannot be accommodated by clique loglinear models—except as cliques of size one—and dropping them amounts to a significant reduction in computational running time. This is important because the number of possible clique loglinear models increases rapidly with the number of vertices.

We have reduced the data to an 80-dimensional contingency table. It has 377 cells with strictly positive counts. The largest positive count is 332,117 while the second largest is 11,614. We perform 100 runs of 100,000 iterations of the stochastic search procedure from Section 2.4. Of all the clique loglinear models visited, we found 1133 whose BICs were within $c = 10^{-4}$ of the BIC of the best clique loglinear model identified across all 100 chains. As in Section 3 we used an Occam's window form of BMA limited to the best models visited while renormalizing (2.12) to reflect this. The clique structure of the best model is shown in Figure 3. Forty of the 100 runs identified the same best model, while the rest of the chains were trapped in local modes and did not reach this model.

We have also generated the independence graph resulting from the Occam's window BMA in Figure 4. The strength of the connectivities between genera are indicated by different types of lines and reflect ranges of posterior probabilities calculated from the Occam's window BMA probability using the models amongst the 1133 best models for which a given collection of higher order terms is present.

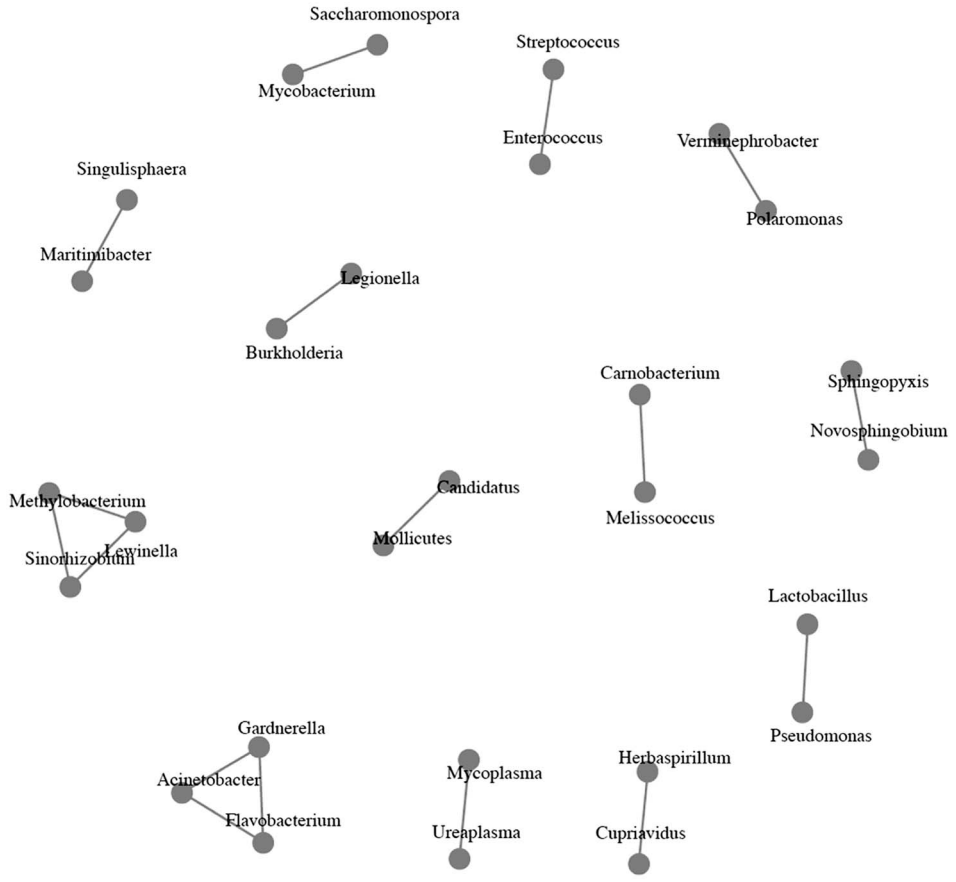

FIG. 3. *The best clique loglinear model identified for the genera data.*

Observe that Figure 4 does not have a clique structure because BMAs of clique loglinear models do not in general form another clique loglinear model. More specifically, in both Figures 3 and 4 the edges correspond to pairs of variables/genera that have strictly posterior probabilities of belonging to a collection of higher order terms in the 1133 clique loglinear models as evaluated by the posterior probabilities given by the Occam's window BMA.

Several of the links in Figure 4 are supported by biological findings regarding the vaginal microbiome. For instance: (i) Ureaplasma and Mycoplasma are bacterial genera from the same bacterial family and are commonly found in the reproductive tract of both men and women; (ii) Polaromonas and Verminephrobacter, and Yersinia and Caldicellulosiruptor, are from the same bacterial family and have been validated by experimentation; and (iii) Melissococcus and Carnobacterium are both main genera producers of bacteriocins, ribosomally synthesized antibacte-

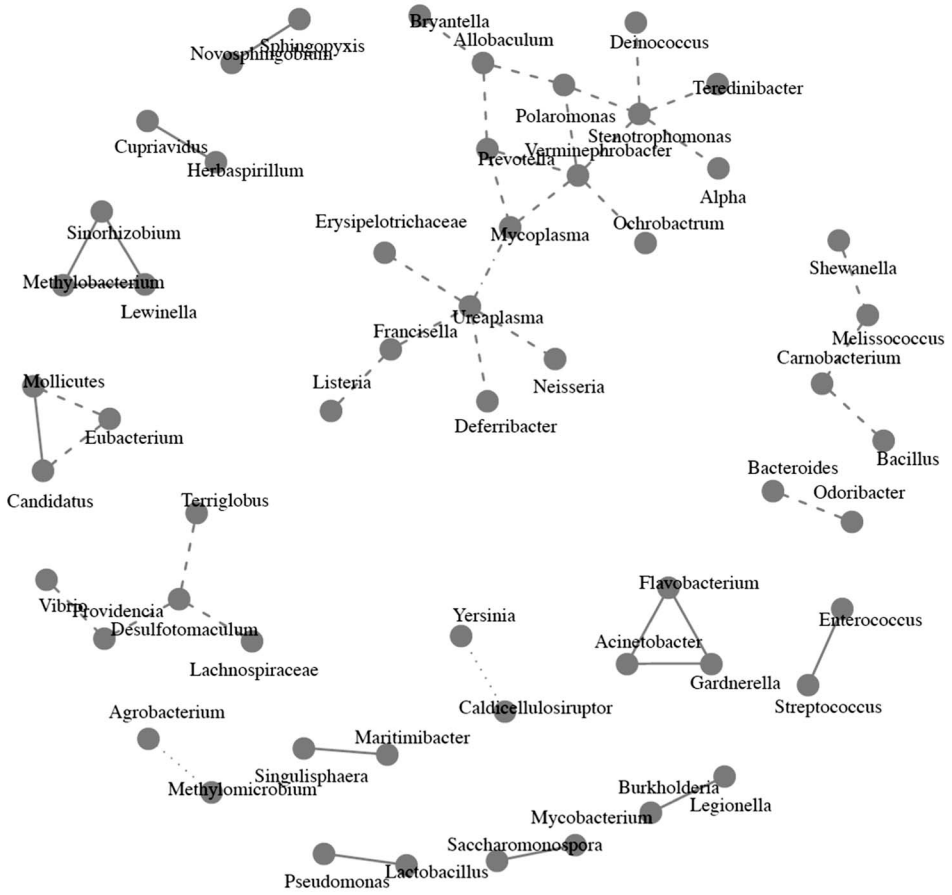

FIG. 4. *Pairs of variables that have strictly positive posterior probabilities of belonging to an higher order term of a clique loglinear model in the genera data under the Occam's window BMA. Solid, dot dashed, dotted and dashed lines denote posterior probabilities that belong to the intervals* (0.9, 1], (0.5, 0.9], (0.1, 0.5] *and* (0, 0.1] *respectively.*

rial peptides/proteins that either kill or inhibit the growth of closely related bacteria and are considered antimicrobial microbes.

We also produced estimates of the probabilities in (2.1) and (2.2) which, from here on, will be referred to as individual existence probabilities. The BMA estimates of the top five individual existence probabilities are as follows: Lactobacillus 0.86; "Unknown" 0.08; Pseudomonas 0.05; Acinobacter 0.01; and Gardnerella 0.002 (probabilities do not add to one because of rounding). While we do not have standard errors for these estimates, it is obvious that Gardnerella is present in only trace amounts and the presence of an unidentified genus is not zero. We return to the interpretation of unidentified genera in Section 6 but note that our findings are consistent with those from analyses of vaginal microbiome samples based on

16S rDNA sequence data that also identified the presence of previously unknown bacterial taxa [Fettweis et al. (2012)].

5. Example: The diabetic foot wound microbiome. One of the complications of diabetes, particularly in elderly patients, is the development and impaired healing of foot ulcers. Diabetes is the primary cause of nontraumatic lower extremity amputations in the United States; approximately 14–24% of patients with diabetes who develop a foot ulcer eventually require an amputation. A diabetic foot ulcer is an open sore or wound that occurs in about 15% of patients with diabetes, usually on the bottom of the foot.

It has been hypothesized that an altered skin microbiome may play a role in the compromised healing of diabetic foot ulcers [Smith et al. (2016)]. The purpose of this study was to investigate the bacteria involved in chronic wound healing. Samples were taken from three locations—the wound bed, the wound edge and the peripheral healthy skin of the foot—of 10 patients at two time points: the time of initial visit and one week after the initial visit. Half of the patients were considered healers, and the remaining patients were considered nonhealers based on clinical assessment of their wounds. Samples were prepared and submitted for V4 16S rRNA gene sequencing with the Illumina MiSeq platform with services provided by Second Genome, Inc. Out of the total of 60 samples (20 from wound base, 20 from wound edge and 20 from healthy skin), we could not not PCR-amplify bacterial DNA from 10 health skin samples. Sequences from the remaining 50 samples passed quality filtering and were mapped to a set of representative consensus sequences to generate an abundance table of Operational Taxonomic Units (OTUs); an OTU is simply a cluster of closely related reads. This table was analyzed using an overdispersed Poisson model [Robinson, McCarthy and Smyth (2010)] to identify OTUs that were significantly differentially expressed between healers and nonhealers at each of the three sample locations (FDR corrected p-value < 0.05). The results consisted of three lists of significant OTUs, one for each location. Further details can be found in the Supplementary Material [Dobra et al. (2019)], Section 5, that also extend our method to the comparison of two populations.

Our clique loglinear analysis is based on the counts of the number of sequencing reads assigned to significant OTUs for each sample with a separate table for the significant OTUs from each location. This analysis is different from our previous example in that: (i) our interest is on the association among samples that may be reflected in components of the microbiome; and (ii) any associations will be based on sharing of OTUs across samples as opposed to sharing of reads across genomes.

An initial exploratory data analysis using hierarchical clustering and principal components analysis revealed that samples from the same subject cluster together, and that subjects cluster into two groups with patients 4 and 5 (healers) and patient 7 (nonhealer) forming a cluster distinct from the remaining subjects; see the Supplementary Material [Dobra et al. (2019)], Section 5. For each of our three analyses (one for each significant OTU list), all samples could form cliques with any of the

other samples because all two-way marginals contain only nonzero counts. We ran the stochastic search from Section 2.4 for 100,000 iterations from 100 random starts. Due to the smaller size of the table relative to the previous example (50 vs. 80 binary variables), we noted convergence to a best graph in less than 50,000 iterations.

The strongest factor in clique formation is subject/patient origin followed by sample location; the distinction between healers and nonhealers is not evident despite the focus on significant OTUs.

Although some cliques appear in all three best graphs, for example, samples from the wound edge of patient 2, most cliques shift subtly with the changes in the significant OTUs; see Figure 5. For example, all samples from patient 6 form a single clique in the best graph based on OTUs that are significantly different in the wound bed between healers and nonhealers. However, in the best graph based on significant OTUs in the wound edge, one of the wound bed samples from patient 6 forms a clique with the wound bed samples from patient 11, while the remaining samples from patient 6 maintain a clique. Not surprisingly, the cliques involving samples from patient 6 change again in the best graph based on significant OTUs in healthy samples (note that patient 6 has no healthy samples). As expected from

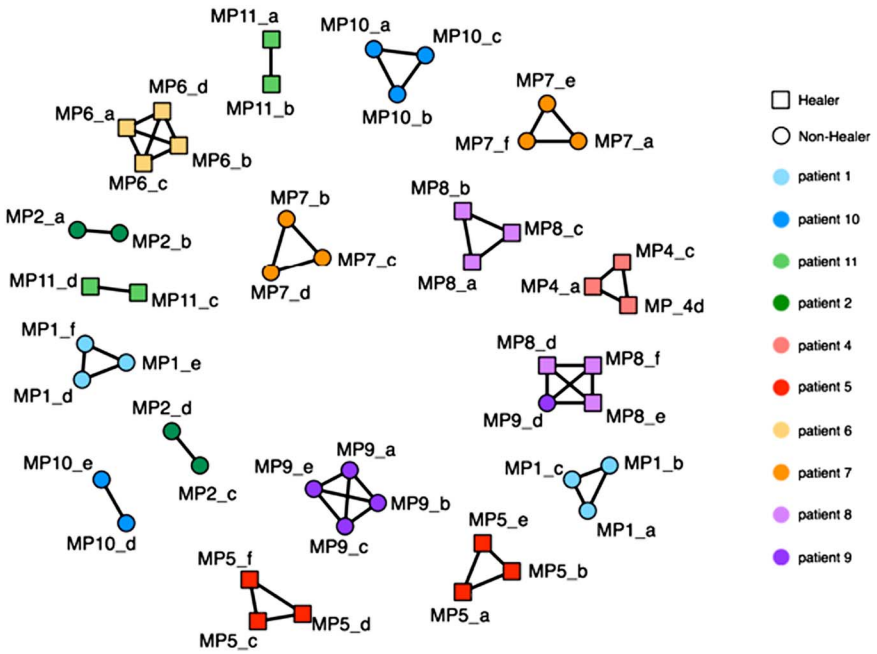

FIG. 5. *The best clique loglinear model of samples from the wound microbiome based on significant OTUs in the wound bed between healers and nonhealers. MP = microbiome patient. The sample location is wound bed* (*a and b*), *wound edge* (*c and d*) *or healthy skin* (*e and f*).

our exploratory analysis, samples from patients 4, 5 and 7 formed cliques only among themselves and not with samples from other patients.

Thus, we have analyzed the data to identify the OTUs and to search for associations among samples. To the best of our knowledge, the data in this example have not been analyzed in any way analogous to our methodology before, either biologically or statistically. Thus, we are unable to corroborate our findings from other sources, although a priori our findings do not appear unreasonable. This example demonstrates that our methodology has the potential to obviate a lot of expensive microbiological work.

6. Discussion. We have developed a statistical methodology for contingency table analysis based on clique loglinear models. This methodology can be broadly used in the context of high dimensional tables, and it accounts for model uncertainty by Bayesian model averaging. Our methodology can infer the presence/absence of specific taxa as well as associations among members of taxa. We have demonstrated the use of our approach in both simulated and real data contexts relevant to applications to metagenomics. We described an efficient method for selecting clique loglinear models based on the BIC. We also provided a Bayesian framework for model determination of clique loglinear models. We note that the advantages of using the Bayesian framework versus the BIC are that all clique loglinear models will be candidate models (as opposed to those models for which the MLEs exist) and that more flexible priors on the model space (2.19) and (2.20) can be employed. On the other hand, hyper Dirichlet priors for cell probabilities together with priors on the model space must be specified, hence the sensitivity of the results with respect to these priors should be investigated. In this sense the BIC could be seen as a conceptually simpler alternative to the full Bayesian framework especially from the point of view of applied researchers.

An issue that repeatedly comes up in this sort of analysis is how to account for the dependence among reads. As noted in Section 1, this dependence is frequently small and, in our experience, the higher the quality of the data the smaller the dependence among reads is, though it is never zero. As a first approximation, therefore, assuming independence is reasonable and parallels the bag-of-words approach that has been applied with much success in natural language processing. In recent years this has been improved to random "N-grams" and an analogous improvement may be possible with NGS reads. Of more immediate relevance the-bag-of-words model has been applied to several branches of bioinformatics with success [Lovato (2015)]. It is important to distinguish between genuine associations among taxa and simply having reads in common for some other reason—syntrophy or evolution, for example—something our methodology does not address. However, this level of study remains in its infancy.

An important question is how much information is really contained in the reads. In this context one can ask if there is adequate read converge to infer reliably which genomes are present in the sample and hence in the population. Obviously, this is

a function of the number of reads, diversity within the sample and complexity of the population. This is not a question that can be addressed statistically after the reads have been generated, although in principle it could be partially addressed at the design stage of the read generation. Read coverage will typically be incomplete and typically will be a limitation on analytic methods. This may increase the uncertainty of downstream inferences, but the task is to reflect uncertainty accurately not to under-represent it. Methodologies that compensate for uncertainty or evaluate uncertainty by, say, robustness criteria, remain to be developed.

More specifically, when reads are shared by two taxa they only mean that the two taxa are similar in the regions that were sampled. Strictly speaking this does not tell us anything about the coexistence of the two taxa. However, first, if shared reads from two taxa are found we have ruled out the case that neither taxon is present. Moreover, if we have reads present that are unique to one taxon, then we have established its presence. If we have reads that are unique to the other taxon, then we have established its presence. Finally, if we have reads that are unique to each of the taxa, we have unambiguously identified both taxa are present. We dropped reads that are unique to one taxon they were the singleton vertices in the connectivity graphs, so we could focus on higher order terms that represented two reads.

We comment that in our wound microbiome example, unlike the HMP example, our analysis may have failed to capture all of the information in the available data as the OTU table was converted from counts (i.e., each cell gives the *number* of reads that align to a given OTU in a given sample) to binary (i.e., each cell indicates if *any* reads align to a given OTU in a given sample) prior to analysis. An extension of our method to raw data consisting of counts, for example, with each subject S_j, $1 \leq j \leq b$, we associate a categorical random variable X_j that takes value k if k sampled reads align to OTU i, $1 \leq i \leq k$ and takes value 0; otherwise, it is a topic for future research.

On the other hand, our method extends to comparing two populations on the basis of their connectivity. The difference in dependencies can be regarded as indicators of which associations are present or absent in the normal case (say) versus the diseased case. This amounts to looking at the different structure of the graphs and interpreting what the cliques mean in terms of reads. Our treatment in Section 5 was subject-by-subject. In Section 5 of the Supplementary Material [Dobra et al. (2019)], we compare two collections of metagenomic samples from two populations, healers and nonhealers.

Our inference of the significant presence/absence of bacterial taxa, possibly unknown, is based on posterior estimates of probabilities (2.1) and (2.2). We refer to these as existence probabilities; however, this terminology belies the subtleties regarding their interpretation. These probabilities are estimated from the model averaged joint posterior distribution, and hence are conditional on clique loglinear models with high posterior probabilities, that is, an Occam's window approach. If a bacterial taxon (say, genus) to which reads uniquely align does not appear in any

of these models, these reads will impact our probability estimates. For example, the estimate of the probability of an unidentified taxon will be inflated, while the estimate of the probability of presence of a taxon appearing in the model average will be deflated. The extent of this impact may be small, as any taxon to which many reads align should appear in the model average, but this cannot be guaranteed and warrants further study.

Stated in another way, the category of genomes we have called unidentified may be only an artifact of the modeling. Indeed, if the genome list contains all the genomes in the sample, the probability of an unidentified genome is simply a residual reflecting the short reads that do not align in sufficiently large numbers to any genome in the models in the model average. On the other hand, if the genome list is incomplete, the probability of an unidentified genome is the sum of two parts—the probability of a genome we know but that was not included amongst the B reference genomes plus the probability of something that we have not encountered before.

SUPPLEMENTARY MATERIAL

Additional proofs, maps, figures and tables (DOI: 10.1214/18-AOAS1229SUPP; .pdf). In this online supplementary material, we describe the data that were used. We also present the computational experiments performed, the details of the simulations, and further details on the software that was developed in this article.

REFERENCES

ABRAMOWITZ, M. and STEGUN, I. A. (1972). *Handbook of mathematical functions with formulas, graphs, and mathematical tables*. U.S. Dept. of Commerce: US GPO, Washington, DC.

BARRY, D. and HARTIGAN, J. A. (1992). Product partition models for change point problems. *Ann. Statist.* **20** 260–279. MR1150343

BERGER, J. O., GHOSH, J. K. and MUKHOPADHYAY, N. (2003). Approximations and consistency of Bayes factors as model dimension grows. *J. Statist. Plann. Inference* **112** 241–258. MR1961733

BHATTACHARYA, A. and DUNSON, D. B. (2012). Simplex factor models for multivariate unordered categorical data. *J. Amer. Statist. Assoc.* **107** 362–377. MR2949366

BISHOP, Y. M. M., FIENBERG, S. E. and HOLLAND, P. W. (2007). *Discrete Multivariate Analysis: Theory and Practice*. Springer, New York. MR2344876

CANALE, A. and DUNSON, D. B. (2011). Bayesian kernel mixtures for counts. *J. Amer. Statist. Assoc.* **106** 1528–1539. MR2896854

CARVALHO, C. M. and SCOTT, J. G. (2009). Objective Bayesian model selection in Gaussian graphical models. *Biometrika* **96** 497–512. MR2538753

CHARUVAKA, A. and RANGWALA, H. (2011). Evaluation of short read metagenomic assembly. *BMC Genomics* **12** S8.

CLARKE, B., VALDES, C., DOBRA, A. and CLARKE, J. (2015). A Bayes testing approach to metagenomic profiling in bacteria. *Stat. Interface* **8** 173–185. MR3322164

DAWID, A. P. and LAURITZEN, S. L. (1993). Hyper-Markov laws in the statistical analysis of decomposable graphical models. *Ann. Statist.* **21** 1272–1317. MR1241267

DE KIEVIT, T. and IGLEWSKI, B. (2000). Bacterial quorum sensing in pathogenic relationships. *Infect. Immun.* **68** 4839–4849.

DELLAPORTAS, P. and FORSTER, J. J. (1999). Markov chain Monte Carlo model determination for hierarchical and graphical log-linear models. *Biometrika* **86** 615–633. MR1723782

DELLAPORTAS, P. and TARANTOLA, C. (2005). Model determination for categorical data with factor level merging. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **67** 269–283. MR2137325

DOBRA, A. (2009). Variable selection and dependency networks for genomewide data. *Biostatistics* **10** 621–639.

DOBRA, A. and LENKOSKI, A. (2011). Copula Gaussian graphical models and their application to modeling functional disability data. *Ann. Appl. Stat.* **5** 969–993. MR2840183

DOBRA, A. and MASSAM, H. (2010). The mode oriented stochastic search (MOSS) algorithm for log-linear models with conjugate priors. *Stat. Methodol.* **7** 240–253. MR2643600

DOBRA, A., VALDES, C., AJDIC, D., CLARKE, B. and CLARKE, J. (2019). Supplement to "Modeling association in microbial communities with clique loglinear models." DOI:10.1214/18-AOAS1229SUPP.

DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *J. Amer. Statist. Assoc.* **104** 1042–1051. MR2562004

EDWARDS, D. and HAVRÁNEK, T. (1985). A fast procedure for model search in multidimensional contingency tables. *Biometrika* **72** 339–351. MR0801773

FETTWEIS, J., SERRANO, M., GIRERD, P., JEFFERSON, K. and BUCK, G. (2012). A new era of the vaginal microbiome: Advances using next generation sequencing. *Chem. Biodivers.* **9** 965–976.

FIENBERG, S. E. and RINALDO, A. (2007). Three centuries of categorical data analysis: Log-linear models and maximum likelihood estimation. *J. Statist. Plann. Inference* **137** 3430–3445. MR2363267

FIERER, N., LAUBER, C., ZHOU, N., MCDONALD, D., COSTELLO, E. and KNIGHT, R. (2010). Forensic identification using skin bacterial communities. *Proc. Natl. Acad. Sci. USA* **107** 6477–6481.

NIH HMP WORKING GROUP, PETERSON, J., GARGES, S., GIOVANNI, M., MCINNES, P., WANG, L., SCHLOSS, J. A., BONAZZI, V., MCEWEN, J. E. et al. (2009). The NIH human microbiome project. *Genome Res.* **19** 2317–2323.

HANKIN, R. K. S. (2006). Additive integer partitions in R. *J. Stat. Softw.* **16**. Code Snippet 1.

HANS, C., DOBRA, A. and WEST, M. (2007). Shotgun stochastic search for "large p" regression. *J. Amer. Statist. Assoc.* **102** 507–516. MR2370849

HASMAN, H., SAPUTRA, D., SICHERITZ-PONTEN, T., LUND, O., SVENDSEN, C. A., FRIMODT-MØLLER, N. and AARESTRUP, F. M. (2014). Rapid whole-genome sequencing for detection and characterization of microorganisms directly from clinical samples. *Eur. J. Clin. Microbiol. Infect. Dis.* **52** 139–146.

HOFFMANN, C., DOLLIVE, S., GRUNBERG, S., CHEN, J., LI, H., WU, G., LEWIS, J. and BUSHMAN, F. (2013). Archaea and fungi of the human gut microbiome: Correlations with diet and bacterial residents. *PLoS ONE* **8** e66019.

HUANG, B., FETTWEIS, J., BROOKS, J. P., JEFFERSON, K. and BUCK, G. (2014). The changing landscape of the vaginal microbiome. *Clin. Lab. Med.* **34** 747–761.

JOHNDROW, J. E., BHATTACHARYA, A. and DUNSON, D. B. (2017). Tensor decompositions and sparse log-linear models. *Ann. Statist.* **45** 1–38. MR3611485

JONES, B., CARVALHO, C., DOBRA, A., HANS, C., CARTER, C. and WEST, M. (2005). Experiments in stochastic computation for high-dimensional graphical models. *Statist. Sci.* **20** 388–400. MR2210226

KOCH, G., NADAL-JIMENEZ, P., REIS, C., MUNTENDAM, R., BOKHOVE, M., MELILLO, E., DIJKSTRA, B., COOL, R. and QUAX, W. (2014). Reducing virulence of the human pathogen Burkholderia by altering the substrate specificity of the quorum-quenching acylase PvdQ. *Proc. Natl. Acad. Sci. USA* **111** 1568–1573.

KUNIHAMA, T. and DUNSON, D. B. (2013). Bayesian modeling of temporal dependence in large sparse contingency tables. *J. Amer. Statist. Assoc.* **108** 1324–1338. MR3174711

LANGMEAD, B. and SALZBERG, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9** 357–359.

LAURITZEN, S. L. (1996). *Graphical Models*. *Oxford Statistical Science Series* **17**. Clarendon Press, Oxford. MR1419991

LENKOSKI, A. and DOBRA, A. (2011). Computational aspects related to inference in Gaussian graphical models with the G-Wishart prior. *J. Comput. Graph. Statist.* **20** 140–157. MR2816542

LETAC, G. and MASSAM, H. (2012). Bayes factors and the geometry of discrete hierarchical log-linear models. *Ann. Statist.* **40** 861–890. MR2985936

LEVY, R. and BORENSTEIN, E. (2013). Metabolic modeling of species interaction in the human microbiome elucidates community-level assembly rules. *Proc. Natl. Acad. Sci. USA* **110** 12804–12809.

LOVATO, P. (2015). Bag of words approaches for Bioinformatics Ph. D. thesis, Dept. Informatics, Univ. Verona.

MADIGAN, D., GAVRIN, J. and RAFTERY, A. E. (1995). Eliciting prior information to enhance the predictive performance of Bayesian graphical models. *Comm. Statist. Theory Methods* **24** 2271–2292. MR1350662

MADIGAN, D. and RAFTERY, A. (1994). Model selection and accounting for model uncertainty in graphical models using Occam's window. *J. Amer. Statist. Assoc.* **89** 1535–1546.

MADIGAN, D. and YORK, J. (1995). Bayesian graphical models for discrete data. *Int. Stat. Rev.* **63** 215–232.

MADIGAN, D. and YORK, J. C. (1997). Bayesian methods for estimation of the size of a closed population. *Biometrika* **84** 19–31. MR1450189

MARKOWITZ, V. M., CHEN, I. M., PALANIAPPAN, K., CHU, K., SZETO, E., PILLAY, M., RATNER, A., HUANG, J., WOYKE, T. et al. (2014). IMG 4 version of the integrated microbial genomes comparative analysis system. *Nucleic Acids Res.* **42** D560–D567.

MASSAM, H., LIU, J. and DOBRA, A. (2009). A conjugate prior for discrete hierarchical log-linear models. *Ann. Statist.* **37** 3431–3467. MR2549565

MINOT, S., BRYSON, A., CHEHOUD, C., WU, G., LEWIS, J. and BUSHMAN, F. (2013). Rapid evolution of the human gut virome. *Proc. Natl. Acad. Sci. USA* **110** 12450–12455.

MOHAMMADI, A., ABEGAZ, F., VAN DEN HEUVEL, E. and WIT, E. C. (2017). Bayesian modelling of Dupuytren disease by using Gaussian copula graphical models. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 629–645. MR3632345

MUKHERJEE, C. and RODRIGUEZ, A. (2016). GPU-powered shotgun stochastic search for Dirichlet process mixtures of Gaussian graphical models. *J. Comput. Graph. Statist.* **25** 762–788. MR3533637

NGUYEN, N.-P., WARNOW, T., POP, M. and WHITE, B. (2016). A perspective on 16S rRNA operational taxonomic unit clustering using sequence similarity. *NPJ Biofilms and Microbiomes* **2** 16004.

ONORANTE, L. and RAFTERY, A. E. (2016). Dynamic model averaging in large model spaces using dynamic Occam's window. *Eur. Econ. Rev.* **81** 2–14.

RANJAN, R., RANI, A., METWALLY, A., MCGEE, H. S. and PERKINS, D. L. (2016). Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469** 967–977.

ROBINSON, M. D., MCCARTHY, D. J. and SMYTH, D. K. (2010). edgeR: A Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26** 139–140.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464. MR0468014

SMITH, K., COLLIER, A., TOWNSEND, E. M., O'DONNELL, L. E., BAL, A. M., BUTCHER, J., MACKAY, W. G., RAMAGE, G. and WILLIAMS, C. (2016). One step closer to understanding the

role of bacteria in diabetic foot ulcers: Characterising the microbiome of ulcers. *BMC Microbiol.* **16** 54.

TARANTOLA, C. (2004). MCMC model determination for discrete graphical models. *Stat. Model.* **4** 39–61. MR2037813

THOENDEL, M., JERALDO, P. R., GREENWOOD-QUAINTANCE, K. E., YAO, J. Z., CHIA, N., HANSSEN, A. D., ABDEL, M. P. and PATEL, R. (2016). Comparison of microbial DNA enrichment tools for metagenomic whole genome sequencing. *J. Microbiol. Methods* **127** 141–145.

WHITTAKER, J. (1990). *Graphical Models in Applied Multivariate Statistics. Wiley Series in Probability and Mathematical Statistics*: *Probability and Mathematical Statistics*. Wiley, Chichester. MR1112133

ZHAO, J., SCHLOSS, P., KALIKIN, L., CARMODY, L., FOSTER, B., PETROSINO, J., CAVALCOLI, J., VANDEVANTER, D., MURRAY, S. et al. (2012). Decade-long bacterial community dynamics in cystic fibrosis airway. *Proc. Natl. Acad. Sci. USA* **109** 5809–5814.

ZHOU, J., BHATTACHARYA, A., HERRING, A. H. and DUNSON, D. B. (2015). Bayesian factorizations of big sparse tensors. *J. Amer. Statist. Assoc.* **110** 1562–1576. MR3449055

ZHOU, J., HERRING, A. H., BHATTACHARYA, A., OLSHAN, A. F., DUNSON, D. B. and THE NATIONAL BIRTH DEFECTS PREVENTION STUDY (2016). Nonparametric Bayes modeling for case control studies with many predictors. *Biometrics* **72** 184–192. MR3500587

A. DOBRA
DEPARTMENT OF STATISTICS
UNIVERSITY OF WASHINGTON
BOX 354322
SEATTLE, WASHINGTON, 98195
USA
E-MAIL: adobra@uw.edu

C. VALDES
SCHOOL OF COMPUTING
AND INFORMATION SCIENCES
FLORIDA INTERNATIONAL UNIVERSITY
11200 SW 8TH STREET
ECS 354
MIAMI, FLORIDA 33199
USA
E-MAIL: hobbes182@gmail.com

D. AJDIC
DEPARTMENT OF DERMATOLOGY
AND CUTANEOUS SURGERY
AND DEPARTMENT OF MICROBIOLOGY
AND IMMUNOLOGY
MILLER SCHOOL OF MEDICINE
UNIVERSITY OF MIAMI
RMSB ROOM 2089
1600 NW 10TH AVENUE
MIAMI, FLORIDA 33136
E-MAIL: d.ajdic@med.miami.edu

B. CLARKE
DEPARTMENT OF STATISTICS
UNIVERSITY OF NEBRASKA-LINCOLN
340 HARDIN HALL NORTH WING
LINCOLN, NEBRASKA 68583
USA
E-MAIL: bclarke3@unl.edu

J. CLARKE
DEPARTMENT OF STATISTICS
AND DEPARTMENT OF FOOD SCIENCE AND TECHNOLOGY
UNIVERSITY OF NEBRASKA-LINCOLN
340 HARDIN HALL NORTH WING
LINCOLN, NEBRASKA 68583
USA
E-MAIL: jclarke3@unl.edu