arXiv:2304.01956v1 [stat.ME] 4 Apr 2023

相关环境中的微生物组相互作用随机图模型

Veronica Vinciotti，特伦托大学数学系；Ernst C. Wit 和 Francisco Richter，瑞士意大利语区大学计算研究所

摘要

微生物组由相互作用的微生物组成，构成复杂的微生物生态系统，并调节宿主的重要通路。微生物丰度测量是了解微生物之间复杂相互作用网络的关键。不同身体部位的微生物群落通常具有某些共同的总体结构，同时也会因局部环境的需求而呈现多样性。我们提出一种计算方法，利用宏基因组数据对多个身体部位的微生物群系系统进行联合推断。随机图模型（RGM）能够处理不同环境间的异质性，同时量化它们在结构层面的关联性。此外，该模型还允许在微生物层面和相互作用层面纳入外部协变量，从而进一步适应微生物组数据的丰富性和复杂性。结果表明：RGM 方法能够捕捉不同身体部位间不同程度的结构相似性，且这种相似性得到其分类学分类的支持；RGM 的贝叶斯实现能够全面量化参数不确定性；微生物组网络后验分布不仅呈现出稳定的核心结构，还揭示了不同身体部位之间有趣的个体差异，以及不同微生物纲之间可解释的关系。

1 引言

微生物组由相互作用的组成部分构成复杂的微生物生态，调节宿主内的重要通路。近年来，微生物系统得到了深入研究，并发现其与多种健康状况相关，例如肥胖（Le Chatelier 等，2013）、糖尿病（Pedersen 等，2016）以及对免疫疗法的反应（Lee 等，2022）。丰富的微生物组高通量数据来源，例如人类微生物组计划（HMP Consortium，2012年）和人类肠道宏基因组项目（MetaHIT）（Qin 等，2010年）生成的数据，是了解微生物群落间复杂相互作用网络的关键。

随着微生物组与局部环境相互作用，微生物组在不同身体部位的组成特征也各不相同（Sharon 等人，2022年）。例如，Segata 等人

（2012年）发现，消化道部位可分为四组，各组具有不同的细菌组成和代谢过程。尽管存在这种异质性，从结构角度来看，不同身体部位之间的相互作用特征预计大体相同。这构成了一个核心微生物组网络，描述了微生物组相互作用中跨时间、身体部位和人群保持稳定的组成部分。

大多数研究仅依赖粪便样本来代表肠道微生物组，并依赖唾液样本来描述口腔微生物组（Sharon 等人，2022年）。同样，所有现有的方法和实现，例如常用的 SparCC（Friedman 和 Alm，2012年）及 SPIEC-EASI（Kurtz 等人，2015），都是从给定数据集中推断一个微生物群系统。因此，这些方法适用于从针对某一环境的研究中学习环境特异性系统，或从不同身体部位的汇总数据中学习核心微生物组网络。与此不同，我们提出一种计算方法，用于联合推断多个身体部位的宏基因组数据所对应的微生物群系统，既能捕捉核心代谢网络，也能反映个体差异。

与本文提出的方法最为接近的是 Vinciotti 等人（2022年）开发的一种高斯耦合函数图模型，用于从基因组计数数据中推断微生物群系统。该方法能够恢复生成数据的核心微生物组网络；同时，其边缘分布所采用的参数形式既能捕捉不同身体部位之间微生物丰度的异质性，也能体现微生物数据的典型特征，例如零膨胀和组成性。高效的贝叶斯推断程序可用于量化估计参数的不确定性。恢复的网络也具有这一特性：微生物之间的相互作用会同时给出相应的后验边概率和偏相关分布。

本文旨在刻画微生物相互作用在结构层面上的异质性，同时量化不同环境中的微生物群系统之间可能存在的关联性。为此，我们在 Vinciotti 等人（2022年）的模型基础上，加入一个作用于条件独立图的随机图模型；这些条件独立图描述了各身体部位的微生物联合计数分布。我们将随机图模型定义为图模型与这一随机独立图模型的结合。

借鉴网络科学文献（Hoff 等，2002），我们将随机图模型形式化为一个潜在 probit 网络模型，其中，特定微生物群系系统中边的概率取决于一个潜在空间，该空间涵盖可能存在关联的环境；也就是说，如果某一身体部位在该潜在空间中与另一个存在该边的环境相近，那么这条边出现的概率就会增加。此外，边的概率还取决于各身体部位各自的网络稀疏度水平，以及网络层面上的外部协变量。对于后者，我们考察分类共享对微生物相互作用倾向的影响；但原则上，也可以在此阶段纳入任何其他协变量或外部知识。

我们将所提出的随机图模型应用于一项微生物组研究，该研究是人类微生物组计划（HMP）的一部分。经过预处理后，我们使用了来自 13 个身体部位的 87 种微生物的 4500 个样本。分析表明，潜在空间能够捕捉这 13 个微生物系统之间的生物学相关性。事实上，推断出的潜在空间中各身体部位的位置，与 Segata 等人（2012年）所作的分类以及 Uberon 身体部位解剖分类（Mungall 等人，2012年）都高度吻合。我们还可以进一步查询环境特异性网络，尤其是其关联的估计边概率，以刻画各个单独

网络，并揭示这 13 种环境之间的共性与差异。除了能够利用所提出的模型从数据中发现的信息外，我们还发现，与针对每个身体部位分别进行单独分析相比，这种新方法能够更稳定地恢复微生物系统。

在第 2 节中，我们详细介绍随机图模型和贝叶斯推断过程；随后在第 3 节中展示真实数据和模拟数据的结果，并在第 4 节进行讨论和总结。

2 方法

2.1 随机图模型

本节定义用于从多个环境的异质微生物组数据中进行网络推断的随机图模型。令 $\mathbf{Y}^{(k)} = (Y_{(k)\,1}, \ldots, Y_{(k)p})$ 表示感兴趣的随机向量，即环境 k 中 p 操作分类单元（OTU）的丰度，其中 $k = 1, \ldots, B$。在我们的研究中，$B = 13$。在抽象层面上，我们假设 $\mathbf{Y}^{(k)}$ 服从某个图模型（GM），

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

相对于某个条件独立图 $G^{(k)}$ 及其相关参数$\mathbf{\Omega}^{(k)}$。这些图 $G = \left\{G^{(k)}\right\}_k$ 本身服从一个联合随机图模型，

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

对于某个参数向量 $\mathbf{\Theta}$。

图模型的类型和随机图模型的类型取决于所考虑的情形。对于图模型，我们采用高斯 copula 图模型，因为它的数学形式简单。它是适用于多元非高斯数据的灵活模型，例如这里所考虑的计数型微生物组数据。因此，与（Cougoul等人，2019；Vinciotti 等人，2022）类似，我们作如下假设：

$$P(Y_1^{(k)} \leq y_1, \ldots, Y_p^{(k)} \leq y_p \mid G^{(k)}, \mathbf{\Omega}^{(k)}) = \Phi_p\big(\Phi^{-1}(F_1(y_1)), \ldots, \Phi^{-1}(F_p(y_p); \mathbf{R}^{(k)})\big),$$

其中，Φ_p 是一个 p 维多元正态分布的累积分布函数，其均值向量为零，相关矩阵为 $\mathbf{R}^{(k)} = \mathbf{\Omega}^{(k)^{-1}}$；$\Phi$是标准单变量正态分布函数，而 F_j 是 OTU j的边缘分布。该模型在条件 k 下诱导的依赖结构由条件独立图 $G^{(k)}$表示。根据高斯图模型理论（Lauritzen，1996），该图由相关矩阵$\mathbf{\Omega}^{(k)}$逆矩阵中的零元素模式给出，该逆矩阵通常称为精度矩阵或浓度矩阵。

为适应微生物组数据的丰富性和异质性，OTU j 的边缘分布可以与外部协变量相关联，包括身体部位以及测序深度等标准化因子。与（Vinciotti 等人，2022年）相同，我们通过使用参数化边缘模型来对此进行形式化。具体而言，我们考虑离散 Weibull 分布

回归边缘分布（Peluso 等，2019），即 $j = 1, \ldots\ldots, p$,

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

具有节点协变量 $\boldsymbol{x} = (1, x_1, \ldots\ldots, x_m)^\top$，以及分别对应于定义离散 Weibull 分布的两个参数的回归系数 $\boldsymbol{\eta}_j$ 和 $\boldsymbol{\gamma}_j$。由于数据具有离散性，耦合函数的潜在高斯空间 $z_j = \Phi^{-1}(F_j(y_j))$ 映射并不唯一。事实上，每个观测值 $(y_j, \boldsymbol{x})$ 在潜在空间中对应一个区间，表示为

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

与联合随机图模型类似，我们尤其关注如何对不同环境之间的关联性进行建模，以及在微生物相互作用层面上如何与外部协变量或现有知识建立可能的连接。为此，我们将该模型形式化为以下潜在 probit 网络模型（Hoff 等，2002）

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

其中，$G_{j_1,\ j_2(k)} = 1$ 表示条件 k 下节点 Y_{j_1} 与节点 Y_{j_2} 之间的一条边，$\boldsymbol{w} \in \mathbb{R}^d$ 是边特异性协变量向量，$\mathbf{c}_1, \ldots\ldots, \mathbf{c}_B \in \mathbb{R}^2$ 是各条件对应的潜在空间变量，α_k 是模型的截距，与图 $G^{(k)}$ 的整体稀疏度有关。我们将与联合随机图模型相关的参数向量记为 $\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c})$。

下一节中，我们将讨论如何根据微生物组数据推断完整的模型参数集，即边际层面的 $\boldsymbol{\eta}_1, \ldots, \boldsymbol{\eta}_p, \boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_p$，以及结构层面的 $G^{(1)}, \ldots, G^{(B)}, \boldsymbol{\Omega}^{(1)}, \ldots, \boldsymbol{\Omega}^{(B)}, \boldsymbol{\Theta}$。

2.2 贝叶斯推断

图 1 展示了所提出的随机图模型，以及它如何从不同且可能相互关联的环境中生成微生物组数据。具体而言，给定定义潜在网络模型的参数$\boldsymbol{\Theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}, \mathbf{c})$后，在给定其余图的条件下，$G^{(k)}$ 中的边相互独立，因此由伯努利抽样生成。给定每种条件 $k = 1, \ldots, B$ 对应的图 $G^{(k)}$ 后，数据随后通过高斯 copula 图模型（Vinciotti 等人，2022）生成；也就是说，通过 G-Wishart 分布抽取正定精度矩阵，再将所得的多元正态抽样与参数化边缘分布结合，以生成计数型微生物组数据。

为了量化参数估计中的全部不确定性，我们采用贝叶斯推断程序。为此，我们考虑以下先验分布：对 $\boldsymbol{\Theta}$ 中的每个参数采用 $(0, 10)$ 先验；在给定图 $G^{(k)}$ 的条件下，对精度矩阵 $\boldsymbol{\Omega}^{(k)} \sim W_G$ 采用 $(3, \mathbb{I}_p)$ G-Wishart 先验；对 $(\boldsymbol{\eta}_j, \boldsymbol{\gamma}_j)$，$j = 1, \ldots, p$ 中的每个回归系数采用 $N(0, 1)$ 先验。

我们设计了一种马尔可夫链蒙特卡罗（MCMC）方案，用于从参数的后验分布中生成样本。具体而言，该方案由以下步骤组成：

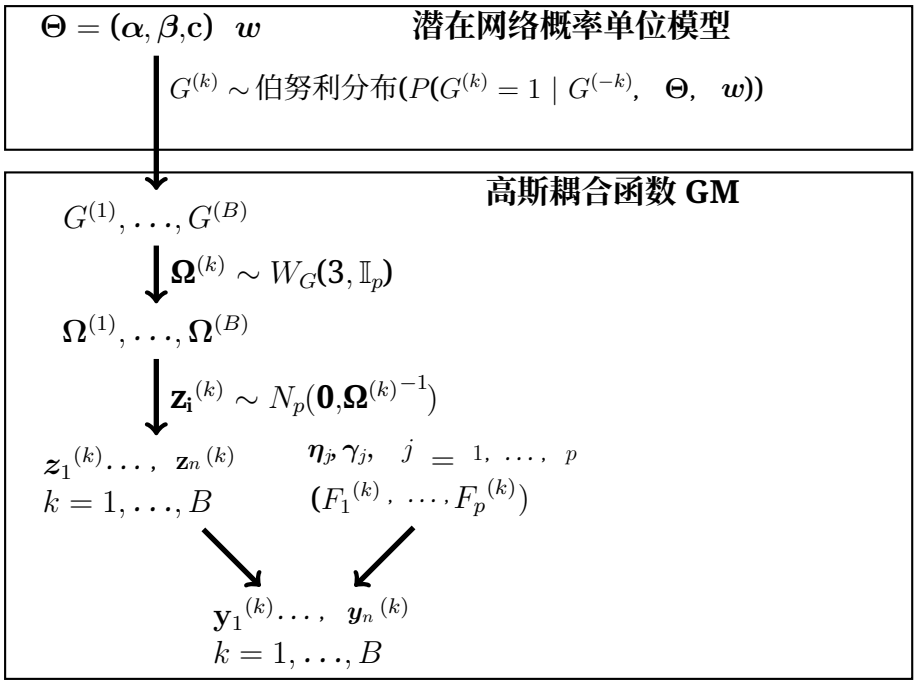

图 1：用于生成不同相关环境中微生物组数据的随机图模型示意图。该模型结合了潜在网络概率单位模型和高斯 copula 图模型。

1. 对边际回归参数进行 Metropolis-Hastings 抽样 η_j, γ_j (Haseli- mashhadi 等, 2018)

2. 通过一系列概率单位回归（带偏移量）从后验分布中对参数进行吉布斯抽样:

- $\alpha|\beta, \mathbf{c}$ $\{G^{(k)}\}_k, w$
- $\beta|\alpha, \mathbf{c}$ $\{G^{(k)}\}_k, w$
- $\mathbf{c}_k|\alpha, \beta, \mathbf{c}_{-k}, \{G^{(k)}\}_k, w$

3. 对 $z_{ij}{}^{(k)}|\Omega^{(k)}, z_{i,-j}{}^{(k)}, y_i{}^{(k)}$ 在 $\mathcal{I}_{\cdot F_j}(y_{ij}|x_i)$上采用截断正态分布进行吉布斯抽样（Vin- ciotti 等，2022年）

4. 采用 G-威沙特后验分布对$\Omega^{(k)}|G^{(k)}, z^{(k)}$ 进行吉布斯抽样（Mohammadi 和 Wit, 2015）

5. 对图进行连续时间出生-死亡MCMC采样

$$(G^{(k)})^{\pm e}|\Omega^{(k)}, z^{(k)}, G^{(k)}, \Theta, w$$

从当前图生成一张新图 $G^{(k)}$，其中一条边 e 被添加、移除或保留（Mohammadi 和 Wit，2015）。

收敛后，将返回所有参数的后验分布。我们尤其关注参数 $\boldsymbol{\Theta}$，它们提供了有关生成这些图的潜在过程，以及不同环境在结构层面上的关联程度的信息；还关注图 $G^{(k)}$，它们与后验边纳入概率相关。

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

其中，N为 MCMC 迭代次数，$W(\boldsymbol{\Omega}_t{}^{(k)}, \boldsymbol{\Theta})$为具有精度矩阵$\boldsymbol{\Omega}_t{}^{(k)}$的图$G_t{}^{(k)}$的等待时间（Vinciotti 等人，2022）。精度矩阵$\boldsymbol{\Omega}^{(k)}$ 的后验分布也可获得，并可通过以下方式转换为每条边的偏相关:

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

这些值还能提供有关各环境中依赖关系符号的信息。类似地，也可以根据返回的图的 MCMC 链，推导出任何感兴趣的网络统计量的后验分布。

3 结果

3.1 模拟研究

为了阐明所提出的随机图模型背后的数据生成过程（图 1），并评估其从数据中推断参数的性能，我们针对 $B = 13$ 个环境，模拟生成$n = 346$ 个观测值和 $p = 87$ 个变量，样本量和维度与真实数据相匹配。在模拟中，我们构建一个潜在空间 $\boldsymbol{\Theta}$，其中包含以下组成部分：从 $N(-2,\ 1)$分布中抽取的 $\boldsymbol{\alpha}$ 参数，即网络稀疏性较高；一个从 $U(-0.5,\ 0.5)$分布中抽取的边协变量 W，并为其设定一个关联参数 $\beta = 2.5$；潜在向量$\mathbf{c} \in \mathbb{R}^2$，其每个分量均从 $N(0,\ 0.3)$分布中抽取。给定 $\boldsymbol{\Theta}$后，我们首先针对每条边，在其他边的条件下通过伯努利抽样生成 13 个图 $G^{(k)}$。我们重复抽样 100 次，以便从图的实际联合分布中抽样。给定真实图后，我们从 $W_G(3,\ \mathbb{I}_p)$分布中抽取相应的精度矩阵，最后通过多元高斯抽样获得各环境的数据。此处我们省略离散边缘分布的情形，重点关注潜在空间的推断和网络的恢复。上述方法构造的数据可通过运行本文配套 `rgm` 软件包中的函数 `sim.rgm` 获取，输入参数采用默认值。

图2展示了运行函数 `rgm` 后经过10000次MCMC迭代得到的结果，所用先验分布如第2.2节所述。每次迭代的计算时间约为14秒。我们保留最后25%的迭代结果，用于根据公式（4）计算边的后验分布，以及随机图模型参数$\boldsymbol{\Theta}$ 的后验分布。第一幅图通过比较公式（3）中的真实 Probit概率与使用 $\boldsymbol{\alpha}$、β 和$\mathbf{c}$ 参数的后验估计均值所得的概率，展示了潜在网络空间 $\boldsymbol{\Theta}$得到了较好的恢复。第二幅图通过比较重构的图与

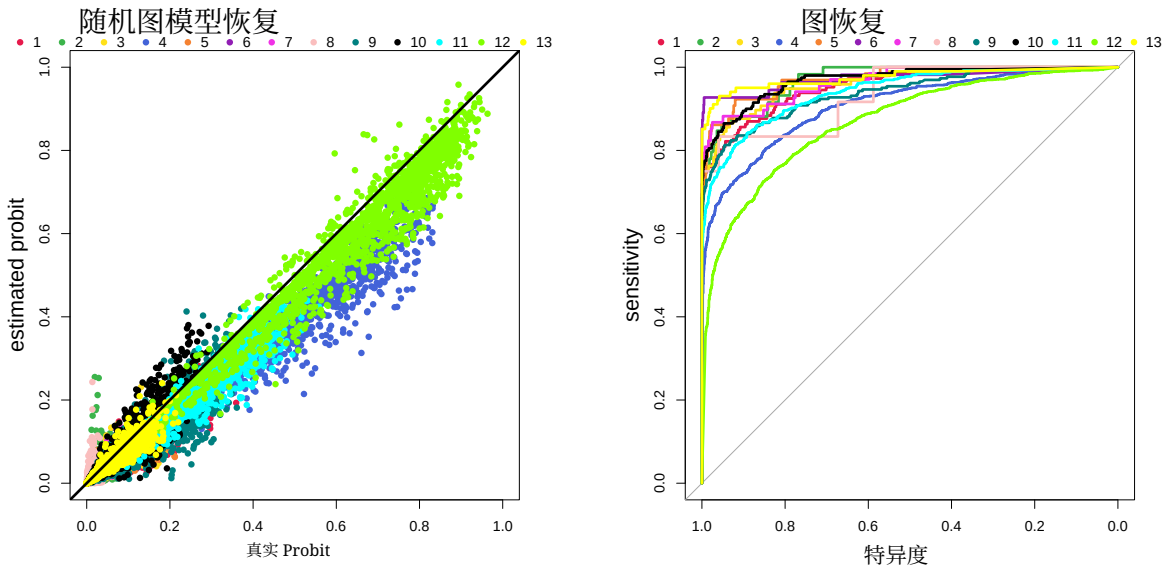

图2：模拟研究的结果，在最后2500次MCMC迭代上进行评估。左：公式（3）中的真实Probit 概率与使用 α、 β 和 c 参数的后验估计均值计算得到的概率。右：对于每个环境，在后验边概率序列的阈值范围内，将恢复的图与真实图进行比较得到的受试者工作特征曲线，后验边概率来自公式（4）。13种颜色分别区分13个环境。

真实图，展示了网络{v8}得到了准确重构；比较涵盖每个环境，并采用一系列后验边概率阈值。13个环境中，受试者工作特征曲线下面积的平均值为0.95。

3.2 J跨身体部位的微生物群系系统联合推断

我们使用 `rMAGMA` package中的微生物组数据（Cougoul等人，2019），这些数据基于人类微生物组计划（HMP联盟，2012年）中健康个体的16S可变区V3-5数据，在操作分类单元（OTU）水平收集微生物丰度。剔除测序读数少于500的样本后，我们聚焦于样本量最大的13个身体部位，分别为“前鼻孔”（下文称为 `nose`）、“附着角化龈”（`ker-gingiva`）、“颊黏膜”（`cheek`）、“硬腭”（`palate`）、“左耳后沟”（`L-ear`）、“腭扁桃体”（`tonsils`）、“右耳后沟”（`R-ear`）、“唾液”（`saliva`）、“粪便”（`stool`）、“龈下菌斑”（`sub-gingiva`）、“舌背”（`tongue`）、“咽喉”（`throat`）和“龈上菌斑”（`sup-gingiva`）。平均而言，每个身体部位有346个样本。最后，我们将关注范围限定在87个操作分类单元上，这些操作分类单元在上述每种环境中都具有两个以上不同的观测值。微生物群落是相互作用的单位，因此构成网络的节点。

我们在每个操作分类单元的边际层面以及网络层面都考虑了多个协变量。作为每个操作分类单元离散 Weibull 分布边际分布（公式 1）的协变量 x，我们纳入每个样本的文库大小，该值通过几何平均数估算，

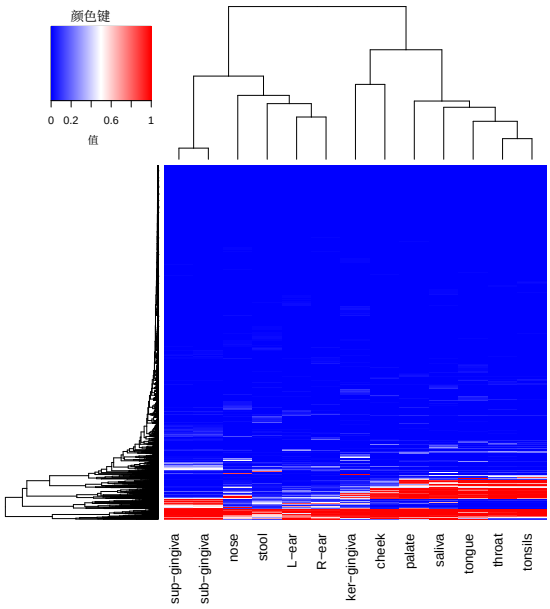

图 3：根据微生物组数据推断的随机图模型：每个环境中各条边的后验边概率，经行列聚类重新排列。

计算方法是将该样本的 OTU丰度与所有其他样本的 OTU丰度进行两两比值计算（ rMAGMA中的函数GMPR ）、每个身体部位的虚拟变量，以及身体部位与文库大小之间的相互作用。这样，每个操作分类单元有 26 个参数。对于公式 3 中的随机图模型，每个环境由一个稀疏度参数 α_k 和一个潜在位置$\boldsymbol{c}_k \in \mathbb{R}^2$定义；此外，还有一个回归系数向量 $\boldsymbol{\beta}$ ，与六个二元变量（$\boldsymbol{w}$）相关联，这些变量用于编码一对操作分类单元是否属于相同分类层级。具体而言，我们考虑六个分类层级，分别为细菌的phylum、 class、 order、 family、 genus 和 species。

我们通过 50000 次 MCMC迭代为每个 OTU 拟合离散威布尔参数边缘分布（ BDgraph 包中的函数 bdw.reg （Mohammadi and Wit, 2019））。与高斯 copula 图模型的推断方法通常采用的做法一样，这一步首先执行，随后使用参数的后验均值估计（在迭代的最后 25% 部分进行评估）计算公式 2 中的区间。然后通过迭代第 2.2 节所述过程的步骤 2–5，利用这些区间学习结构依赖关系（随本文提供的 rgm 包中的函数 rgm ）。鉴于图的空间极其庞大，我们让贝叶斯结构学习过程运行 300 万次 MCMC迭代。所有后续结果均根据最后 7,500 次迭代进行评估。

结果的解读。 分析最直接的产出是针对每个环境推断出的 13 个网络。图 3 根据公式（4）计算出的后验边概率，对这些网络进行了汇总。显然，这些网络往往较为稀疏，而且在不同条件之间存在一定差异。造成这种环境网络变异的原因可从公式（3）所述的随机图生成过程中找到。图 3 显示，所有网络整体上都高度稀疏（在 13 个环境中，后验边概率的平均值为 6.8%）。这一点体现在拟合的随机图模型取值较低

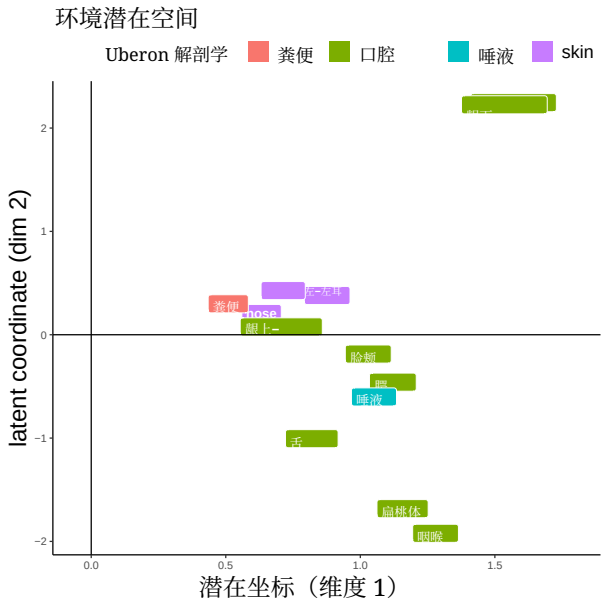

图 4：根据微生物组数据推断出的随机图模型：身体部位（$\boldsymbol{c}$）在二维潜在空间中的后验平均位置。颜色表示身体部位的 Uberon 解剖分类。

（在 13 个环境中，后验估计的平均值为 -2.7）。

图 3 显示，某些环境之间在结构上高度相似；相似环境之间，高概率边和低概率边都有较高的共享程度。随机图模型的潜在位置 $\boldsymbol{c}$ 捕捉到了这种相似性，其后验均值绘于图 4。例如， `sup-gingiva` 和 `sub-gingiva` 是高度相关的环境， `throat` 和 `tonsils` 也同样如此。事实上，在这两种情况下，相关的两个 $\boldsymbol{c}$ 向量的内积都很大，因为它们在空间中彼此接近，且远离零点。公式（3）中的指示函数进一步促进这些网络之间共享边。事实上，后验边概率大于 0.5 的边中， `sup-gingiva` 和 `sub-gingiva` 网络共有 93%， `throat` 和 `tonsils` 网络共有 95%。查看根据精度矩阵通过公式（5）计算得到的偏相关后验均值，我们还发现依赖关系的符号也一致： `sup-gingiva` 和 `sub-gingiva` 的各边偏相关值相关系数为 0.90， `throat` 和 `tonsils` 的相关系数为 0.93。最后，由于这两对网络在潜在空间中几乎相互正交，我们预计 `sup-gingiva`/`sub-gingiva` 网络与 `throat`/`tonsils` 网络之间的结构共享较少。事实上，在所有网络对中，它们的高概率边一致性最低，平均共享比例为 21.3%。

所提出的方法检测到的环境之间的相似性，在一定程度上得到了 Uberon 解剖分类中身体部位分类的支持，尤其是三个与皮肤相关的身体部位。这些部位在图 4 的潜在空间中彼此接近，平均有 63% 的高概率边是共同的（图 3）。另一方面，与口腔相关的身体部位似乎进一步分成了两组。这与 Segata 等人（2012年）的分析一致，他们发现有四组

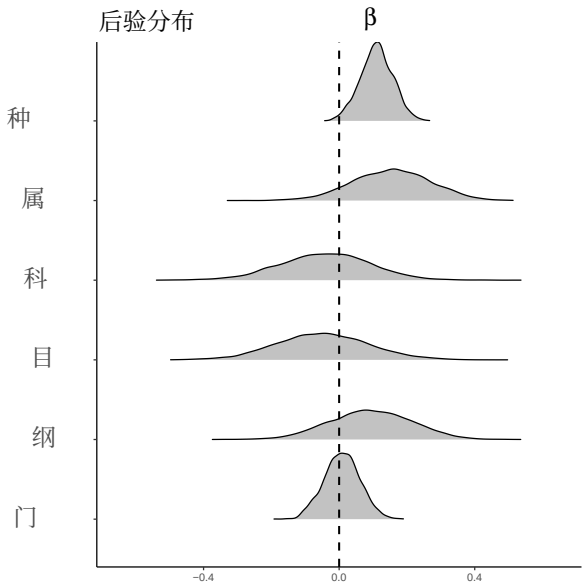

图5：根据微生物组数据推断出的随机图模型：估计与六个虚拟变量相关的 β 参数，这些变量用于指示构成一条边的一对节点是否属于同一分类层级

基于群落组成相似性的身体部位，分别为： cheek、 ker-gingiva、palate； saliva、 tongue、 tonsils、 throat； sub-gingiva和 sup-gingiva；以及 stool。这些组在图4中也清晰可见。

最后，微生物的分类学亲缘关系会促进它们之间形成连接。图5显示，在任何环境中，两个操作分类单元相连的概率与它们是否属于所考虑的某些分类层级呈正相关，尤其是 species、 genus 和 class 分类。

与其他方法的比较。 图6显示，与估计单独网络相比，随机图模型还能够更稳定地恢复单独网络。事实上，该图显示，与对每个环境分别拟合单独的高斯 copula 图模型相比，所提出的 rgm 方法得到的后验边概率方差更小。对于后者，我们采用 Vinciotti 等人（2022年）的方法（ BDgraph 软件包中的函数 bdgraph.dw ），该方法使用与本文相同的参数化边缘分布，但对每个环境采用更为传统的 Erd¨os-R´enyi 随机图先验。为便于比较，将 Erd¨os-R´enyi 先验的先验边概率设定为与通过 rgm 分析恢复的网络稀疏度相匹配。图6显示，与替代方法相比，通过rgm 检测到的后验概率更加集中于0或1，从而使本文提出的方法恢复的结构依赖关系中的噪声水平更低。

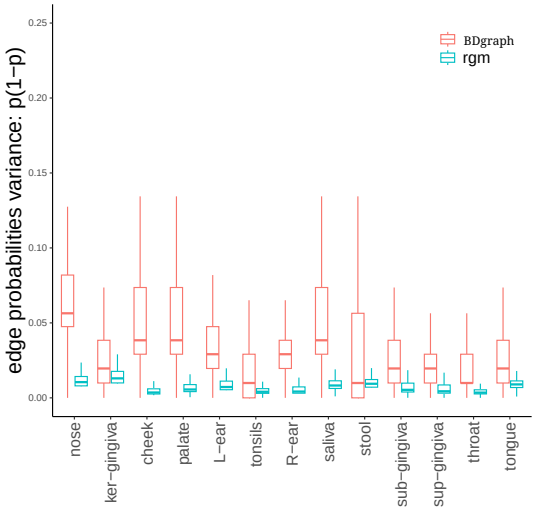

图6：比较每种环境下的联合 `rgm` 高斯耦合函数图模型与各自的高斯耦合函数图模型（Vinciotti 等人，2022）。对于每种方法和每种环境，图中展示了后验边概率方差的箱线图。

4 讨论与结论

本文提出了一种新颖的方法，用于从不同但相关的环境中获得的微生物丰度多变量测量结果推断微生物系统。我们展示了，将每个环境的图模型与描述各环境间图分布的联合随机图模型相结合，能够了解单独的微生物群系系统及其结构相似性。为进一步适应微生物组数据的丰富性和复杂性，所提出的方法允许纳入外部协变量，这些协变量可能与边缘微生物丰度或其相互作用存在关联。

除了本文所呈现的分析外，该方法还可更广泛地用于在不同条件下测得的微生物组数据，以了解各环境内部的结构依赖关系及其在不同环境间的相似性。在这一更一般的层面上，本文提出的方法与针对多种条件下数据的图建模方法有一定相似之处，例如 Ni 等人（2022）所述的方法。根据具体情境，可能需要采用更适用的随机图模型，例如，对于随时间测量且依赖关系会随时间变化的微生物组数据。

软件

R 软件包可在 https://github.com/franciscorichter/rgm 获取

资助

EW 和 FRM 感谢瑞士国家科学基金会（SNF 188334）提供资助

参考文献

Cougoul, A., X. Bailly 和 E. Wit（2019）。MAGMA：稀疏微生物关联网络的推断。bioRxiv：538579。Friedman, J. 和 E. Alm（2012）。基于基因组调查数据推断相关网络。*PLoS Computational Biology 8*（9），e1002687。Haselimashhadi, H., V. Vinciotti 和 K. Yu（2018）。一种用于计数数据的新型贝叶斯回归模型及其在健康数据中的应用。应用统计学杂志 *45*（6），1085–1105。HMP联盟（2012年）。人类微生物组研究框架。《自然》*486*（7402），215–221。Hoff, P., A. Raftery 和 M. Handcock（2002）。社会网络分析的潜在空间方法。美国统计学会会刊 *97*（460），1090–1098。Kurtz, Z., C. M¨uller, E. Miraldi, D. Littman, M. Blaser 和 R. Bonneau（2015）。稀疏且对组成稳健的微生物生态网络推断。*PLoS ComputationalBiology 11*（5），e1004226。Lauritzen, S.（1996）。图模型。克拉伦登出版社。Le Chatelier, E., T. Nielsen, J. Qin, E. Prifti, F. Hildebrand 等（2013）。人类肠道微生物组的丰富度与代谢标志物相关。《自然》*500*（7464），541–546。Lee, K., A. Thomas, A. Bolte, J. Bj¨ork, L. Kist de Ruijter 等（2022年）。晚期黑色素瘤中跨队列肠道微生物组与免疫检查点抑制剂疗效的关联。《自然·医学》*28*，535–544。Mohammadi, R. 和 E. Wit（2015）。稀疏高斯图模型中的贝叶斯结构学习。《贝叶斯分析》*10*（1），109–138。Mohammadi, R. 和 E. Wit（2019）。BDgraph：用于图模型贝叶斯结构学习的 R 软件包。《统计软件杂志》*89*（3），1–30。Mungall, C., C. Torniai, G. Gkoutos, S. Lewis 和 M. Haendel（2012）。Uberon：一种整合性多物种解剖学本体。《基因组生物学》*13*（1），R5。Ni, Y., V. Baladandayuthapani, M. Vannucci 和 F. Stingo（2022年）。面向现代生物学应用的贝叶斯图模型。《统计方法与应用》*31*（2），197–225。Pedersen, H., V. Gudmundsdottir, H. Nielsen, T. Hyotylainen, T. Nielsen 等（2016）。人类肠道微生物影响宿主血清代谢组和胰岛素敏感性。《自然》*535*（7612），376–381。Peluso, A., V. Vinciotti 和 K. Yu（2019）。离散威布尔广义加性模型：在生育计数数据中的应用。《皇家统计学会杂志：*C*辑》*68*（3），565–583。

Qin, J.、R. Li、J. Raes、M. Arumugam、K. S. Burgdorf、C. Manichanh、T. Nielsen、N. Pons、F. Levenez、T. Yamada 等（2010年）。通过宏基因组测序建立的人类肠道微生物基因目录。《自然》*464*（7285），59–65。Segata, N.、S. Haake、P. Mannon、K. Lemon、L. Waldron 等（2012年）。基于口腔表面、扁桃体、咽喉和粪便样本的人类成人消化道细菌微生物组组成。《基因组生物学》*13*，R42。Sharon, I.、N. Quijada、E. Pasolli、M. Fabbrini、F. Vitali 等（2022年）。人类核心微生物组：它是否存在，我们如何找到它？对这一概念的批判性综述。《营养素》*14*（14），2872。Vinciotti, V.、P. Behrouzi 和 R. Mohammadi（2022年）。基于计数型宏基因组数据的微生物群系系统贝叶斯推断。arXiv:2203.10118。