中国科学 : 数学　2017 年　第 47 卷　第 12 期 : 1735 ~ 1760
SCIENTIA SINICA　Mathematica
《中国科学》杂志社
SCIENCE CHINA PRESS
综　述

微生物组学中的高维计数和成分数据分析

献给钱敏教授 90 华诞

吴昌晶[1], 何顺[1], 邓明华[1,2,3*]

1. 北京大学数学科学学院, 北京 100871;
2. 北京大学定量生物学中心, 北京 100871;
3. 北京大学统计科学中心, 北京 100871
E-mail: wcj@pku.edu.cn, heshun@pku.edu.cn, dengmh@pku.edu.cn

收稿日期: 2017-07-03; 接受日期: 2017-09-29; 网络出版日期: 2017-11-16; * 通信作者
国家重大科学研究计划 (批准号: 2015CB910303)、国家重点研发计划 (批准号: 2016YFA0502303) 和国家自然科学基金 (批准号: 31471246) 资助项目

摘要　人体微生物组对人体健康和疾病起着重要作用. 高通量测序技术的发展使得我们可以定量分析微生物组中所有菌种的成分. 本文回顾近来在微生物组学研究中的高维计数和成分数据分析方法, 其中包括 Dirichlet 多项分布模型及其拓展、从大维稀疏计数矩阵估计成分数据、高维成分回归和基于对数基底的成分数据统计推断方法.

关键词　高维计数数据　成分数据　Dirichlet 多项分布模型　稀疏性　可识别性　回归模型
MSC (2010) 主题分类　62-02, 62F15, 62H12, 62H15, 62J07, 62P10

1　引言

人体中生活着大量微生物, 其细胞总量据估计是我们人细胞的 10 倍[1]. 它们分布在人体的各个部位, 如皮肤、口腔和肠道等, 其中尤以肠道中的微生物数量最多、种类最丰富. 微生物往往以群落的形式存在, 它们之间通过协同作用来影响自身所在的环境, 反过来环境也影响着微生物群落的组成和结构. 近年来, 涌现出了大量探索微生物与人体健康关系的研究. 例如, 肠道微生物与肥胖[2]、糖尿病[3]和心血管疾病[4] 等有着密切联系, 而人们的饮食习惯和生活方式等也会影响肠道菌群的结构[5]. 如何综合分析微生物组学数据及这些相关信息, 从而为人类健康造福, 是当前生物统计领域一个重要的课题.

飞速发展的高通量测序技术使得人们可以定量分析人体内微生物群落的组成信息, 这给人们研究微生物群落之间、微生物与宿主间的关联性提供了极大便利. 当前广泛使用的测序技术有两种, 一种是 16S rRNA 测序, 另一种是全基因组鸟枪法测序 (whole genome shotgun sequencing). 16S rRNA 是原核细胞核糖体的一段结构性区域, 广泛存在于细菌等原核生物中. 它既包含了高度保守的区域, 有利于扩增时引物的结合, 又具有在不同菌种中高度分化的区域, 从而可以通过该段序列鉴别微生物的

英文引用格式: Wu C J, He S, Deng M H. High-dimensional count and compositional data analysis in microbiome studies (in Chinese). Sci Sin Math, 2017, 47: 1735–1760, doi: 10.1360/N012017-00147

种类. 在实验得到测序数据后, 人们往往先根据序列相似度 (如 97%) 把不同序列聚类成不同的操作分类单元 (operational taxonomic unit, OTU), 然后将得到的 OTU 与已有的 16S rRNA 数据库对比, 进而得到不同菌种所对应的序列个数 (读段数). 此外, 利用已知的序列信息可以构建一个系统发生树, 测序所得的 OTU 在不同层级累加后能够反映出不同分类等级的微生物含量, 如不同的门、纲、目、科、属、种的微生物的读段数.

与 16S rRNA 针对一段标记基因测序不同, 全基因组鸟枪法测序是把样品全部基因组一起测序, 从而得到宿主和微生物的全部基因信息, 然后与数据库中已有的微生物基因组序列信息比对后确定微生物的丰度信息. 这种测序方法亦称为宏基因组测序. 不过, 由于待测序的基因较多, 宏基因组测序的准确性对实验准备阶段要求较高[6], 数据分析也因读段数过多较为复杂. 关于鸟枪法测序的实验细节和数据处理, 可以参见文献 [7].

本文主要回顾微生物组学研究中的相关统计方法, 包括计数 (count data) 和成分数据 (compositional data) 分析. 微生物组学中的测序数据有如下特点. 第一, 所得数据反映了微生物在某个分类层级下的读段数, 从而原始数据为非负整数. 第二, 由于微生物群落中主导菌的存在, 少数 OTU 的读段数可以占到同一样本内总读段数的绝大部分, 而大部分 OTU 只有很少的读段数, 甚至很多 OTU 都只在个别样本中测得了数据, 在其他样本中计数为零. 第三, 由于测序深度和样本大小不同, 不同样本的总读段数相差很大, 一个样本的数据只反映了该样本内不同微生物种类的相对成分, 其绝对含量难以测得. 主导菌群的存在更使得这一成分数据属性不能在统计分析中简单忽略. 第四, 实验中得到的 OTU 总数往往成千上万, 即使只统计到 “属” 这一级别的微生物, 种类个数也往往上百, 而样本数通常只有几十到几百. 这意味着我们需要对高维计数或成分数据进行分析, 经典的变量个数固定、样本个数趋于无穷的统计方法可能失去功效甚至无法使用. 最后, 系统发生树的结构可以给微生物组数据分析提供更多的信息. 如图 1 所示, 每个结点代表不同等级的分类单元, 叶结点代表较低水平的分类单元 (如 OTU 或属、种), 而内部结点随着向根部汇合代表较高水平的分类单元 (如界、门). 因此, 对应到较高分类单元的读段数是其所有子结点的读段数之和.

在微生物组学研究中, 我们关心的科学问题主要包括以下几个. 首先是微生物相对丰度的估计, 以及对两总体或多总体间微生物丰度是否相同进行假设检验. 其次, 我们希望研究肠道微生物种群间以及它们与人体健康 (如衡量肥胖的 BMI 指数, body mass index) 和饮食摄入的影响, 这可以通过回归模型和图模型对数据进行建模, 目标是进行参数估计和相应的统计推断. 此外, 在数据分析过程中, 我们需要充分考虑微生物组数据的几个特点, 以保证统计推断的合理性. 关于微生物组、宏基因组学和高维成分数据分析, 可参见文献 [8]. Li 等[9] 和 Layeghifard 等[10] 也对微生物间相互作用的课题从计算生物学和网络的角度进行了回顾. 本文将进一步从高维计数数据和高维成分数据两个方面综述近年来微生物组学数据分析的一些统计方法.

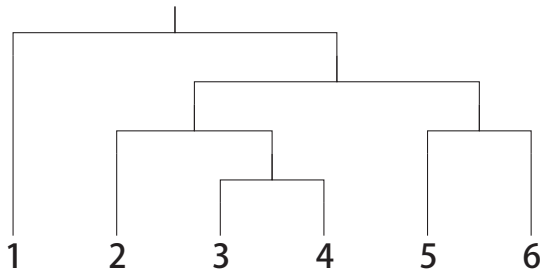

图 1　系统发生树示意图

在本文中, 我们用大写字母表示矩阵或随机变量/向量, 小写字母表示确定的向量或标量, 且矩阵和向量均用黑体表示. 对向量 $\boldsymbol{a}\in\mathbb{R}^p$, 记 $\|\boldsymbol{a}\|_q$ $(q=1,2,\infty)$ 为其 ℓ_q 范数. 对矩阵 $\boldsymbol{A}$, 令 $\|\boldsymbol{A}\|_1=\sum_{i,j}|a_{ij}|$ 为 $\boldsymbol{A}$ 的 ℓ_1 范数, $\|\boldsymbol{A}\|_F=(\sum_{i,j}a_{ij}^2)^{\frac{1}{2}}$ 为 $\boldsymbol{A}$ 的 Frobenius 范数, $\|\boldsymbol{A}\|_{\max}=\max_{i,j}|a_{ij}|$ 为元素的 ℓ_∞ 范数, $\|\boldsymbol{A}\|_{\ell_1}=\max_j\sum_i|a_{ij}|$ 为 $\boldsymbol{A}$ 的矩阵 ℓ_1 范数, $\|\boldsymbol{A}\|_2=\max_i\sigma_i(A)$ 为 $\boldsymbol{A}$ 的谱范数, $\|\boldsymbol{A}\|_*=\sum_i\sigma_i(A)$ 为 $\boldsymbol{A}$ 的核范数, 其中 $\sigma_i(\boldsymbol{A})$ 为 $\boldsymbol{A}$ 的第 i 大的奇异值. 当 $\boldsymbol{A}$ 是方阵时, 记 $\|\boldsymbol{A}\|_{1,\mathrm{off}}=\sum_{i\neq j}|a_{ij}|$ 为非对角元的 ℓ_1 范数.

2 高维计数数据分析

本节回顾微生物组学中的高维计数模型和统计方法. 使用计数模型可以更好地考虑到测序数据中采样的随机性, 因为测序所得的计数数据可以视为微生物种群丰度的一个带噪声的实现. 与之相对的, 由于在建模过程中需要考虑到抽样的随机性, 我们往往会使用分层模型 (hierarchical model), 这通常会使统计模型较为复杂, 难以推导高维情形下相应的统计性质. 此外, 高维数据意味着变量个数可能与样本量相当甚至远大于样本量, 我们需要对模型加一定结构性假设才有可能进行相应的统计分析. 首先介绍最常见的 Dirichlet 多项分布 (Dirichlet-multimomial, DM) 模型, 以及它的一些扩展.

2.1 DM 模型及相应的统计推断方法

假设我们观测到 n 个样本, 每个样本测量了 p 种微生物的读段数. 令 x_{ij} $(i=1,\dots,n;j=1,\dots,p)$ 表示样本 i 在菌种 j 上的读段数, 并记 $\boldsymbol{x}_i=(x_{i1},\dots,x_{ip})^{\mathrm{T}}$ 和 $N_i=\sum_{j=1}^p x_{ij}$ 为样本 i 的总读段数. 由于 x_{ij} 只能取非负整数, 并考虑到微生物组数据是成分数据的特点, 我们自然考虑多项分布模型, 其概率分布为

$$f_{\mathrm{M}}(\boldsymbol{x}_i;\boldsymbol{\pi}_i)=\binom{N_i}{\boldsymbol{x}_i}\prod_{j=1}^p\pi_{ij}^{x_{ij}},\tag{2.1}$$

其中 $\boldsymbol{\pi}_i=(\pi_{i1},\dots,\pi_{ip})^{\mathrm{T}}$ 是样本 i 对应的多项分布的参数, 满足 $\pi_{ij}>0$, $j=1,\dots,p$, $\sum_{j=1}^p\pi_{ij}=1$. 对于多项分布, 其均值和方差均可由该参数决定, 即设 X_{ij} 是相应位置的随机变量, 则有

$$\mathrm{E}(X_{ij})=N_i\pi_{ij},\quad \mathrm{Var}(X_{ij})=N_i\pi_{ij}(1-\pi_{ij}).\tag{2.2}$$

对于微生物组计数数据而言, 不同样本间的成分参数可能相差很大, 具有所谓的 “超散布性” (overdispersion), 即观测到的不同样本成分参数的方差显著大于多项分布模型下给出的方差. 这意味着每个样本应有各自的多项分布参数. 为了在建模中考虑进这种超散布性, 并减少模型需要的参数, 我们假设不同样本的成分参数来自同一随机变量的不同实现. 由于这一随机变量的取值空间在单纯形 (simplex) 上, 一个常用的分布是 Dirichlet 分布, 它也是 Bayes 理论中与多项分布共轭的分布, 其概率分布为

$$f_{\mathrm{D}}(\boldsymbol{\pi}_i;\boldsymbol{\alpha})=\frac{\Gamma(\alpha_+)}{\prod_{j=1}^p\Gamma(\alpha_j)}\prod_{j=1}^p\pi_{ij}^{\alpha_j-1},\tag{2.3}$$

其中 $\boldsymbol{\alpha}=(\alpha_1,\dots,\alpha_p)^{\mathrm{T}}$ 是 Dirichlet 分布参数, 满足 $\alpha_j>0$, $j=1,\dots,p$, 而 $\alpha_+=\sum_{j=1}^p\alpha_j$. 结合 (2.1) 和 (2.3), 我们可以写出 DM 模型[11] 的概率分布为

$$f_{\mathrm{DM}}(\boldsymbol{x}_i;\boldsymbol{\alpha})=\int f_{\mathrm{M}}(\boldsymbol{x}_i;\boldsymbol{\pi}_i)f_{\mathrm{D}}(\boldsymbol{\pi}_i;\boldsymbol{\alpha})\,d\boldsymbol{\pi}_i$$

$$= \binom{N_i}{\boldsymbol{x}_i} \frac{\Gamma(\alpha_+)}{\Gamma(N_i+\alpha_+)} \prod_{j=1}^{p} \frac{\Gamma(x_{ij}+\alpha_j)}{\Gamma(\alpha_j)}. \tag{2.4}$$

DM 模型的另一种等价的参数化方法是令 $\phi_j = \frac{\alpha_j}{\alpha_+}$, $\theta = \frac{1}{1+\alpha_+}$, 则概率分布可写为

$$f_{\mathrm{DM}}(\boldsymbol{x}_i;\boldsymbol{\phi},\theta) = \binom{N_i}{\boldsymbol{x}_i} \frac{\prod_{j=1}^{p}\prod_{k=1}^{x_{ij}}\{\phi_j(1-\theta)+(k-1)\theta\}}{\prod_{k=1}^{x_{ij}}\{1-\theta+(k-1)\theta\}}.$$

这种参数化方式的好处是参数的含义更易解释, 因为有

$$\mathrm{E}(X_{ij}) = N_i\phi_j, \quad \mathrm{Var}(X_{ij}) = N_i\phi_j(1-\phi_j)\{1+\theta(N_i-1)\}, \tag{2.5}$$

从而 ϕ_j 刻画了均值, 而 θ 称为散布参数 (dispersion parameter), 刻画了方差的散布性. 比较 (2.2) 和 (2.5) 可以看出, DM 模型的方差比多项分布的多乘了一个 $1+\theta(N_i-1)$ 因子.

La Rosa 等[12] 考虑了 DM 模型下关于微生物成分组成的检验问题. DM 模型参数的极大似然估计没有解析解, 但从 (2.5) 易推知参数 ϕ_j 的矩估计为 $\hat{\phi}_j = \frac{\sum_{i=1}^{n} x_{ij}}{N}$, 其中 $N = \sum_{i=1}^{n} N_i$. 而散布参数 θ 的矩估计为

$$\hat{\theta} = \frac{\sum_{j=1}^{p}(S_j - G_j)}{\sum_{j=1}^{p}\{S_j + (N_c-1)G_j\}},$$

其中

$$N_c = \frac{1}{n-1}\left(N - \frac{\sum_{i=1}^{n} N_i^2}{N}\right), \quad S_j = \frac{1}{n-1}\sum_{i=1}^{n} N_i\left(\frac{x_{ij}}{N_i} - \hat{\phi}_j\right)^2, \quad G_j = \frac{1}{N-n}\sum_{i=1}^{n} N_i \frac{x_{ij}}{N_i}\left(1-\frac{x_{ij}}{N_i}\right).$$

在经典的 p 固定而 N 趋于无穷的渐近理论框架下, 矩估计具有优良的统计性质, 如渐近无偏和渐近正态性[13], 可以据此给出关于微生物成分组成的假设检验的统计量及其渐近分布. 例如, 如果想检验某总体的微生物成分 $\boldsymbol{\phi}$ 是否与一预先给定的成分向量 $\boldsymbol{\phi}_0$ 相等,

$$H_0 : \boldsymbol{\phi} = \boldsymbol{\phi}_0 \quad \text{vs.} \quad H_1 : \boldsymbol{\phi} \neq \boldsymbol{\phi}_0,$$

则可考虑如下 Wald 类型的检验统计量:

$$T_1 = (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0)^{\mathrm{T}} (V(\boldsymbol{\phi}_0, \hat{\theta}, N))^{-} (\hat{\boldsymbol{\phi}} - \boldsymbol{\phi}_0), \tag{2.6}$$

其中 $(\cdot)^{-}$ 是矩阵的 Moore-Penrose 广义逆, 而

$$V(\boldsymbol{\phi}_0, \hat{\theta}, N) = \frac{1}{N^2}\left\{\hat{\theta}\left(\sum_{i=1}^{n} N_i^2 - N\right) + N\right\}(\boldsymbol{D}(\boldsymbol{\phi}_0) - \boldsymbol{\phi}_0\boldsymbol{\phi}_0^{\mathrm{T}}),$$

其中 $\boldsymbol{D}(\boldsymbol{\phi}_0)$ 是对角元为 $\boldsymbol{\phi}_0$ 的对角矩阵. 在零假设下, T_1 的渐近分布为自由度是矩阵 $\boldsymbol{D}(\boldsymbol{\phi}_0) - \boldsymbol{\phi}_0\boldsymbol{\phi}_0^{\mathrm{T}}$ 的秩的卡方统计量, 故可据此计算检验统计量的 p 值. 类似地, 也可以考虑两总体或多总体下成分向量的假设检验问题. 例如, 对于两总体检验

$$H_0 : \boldsymbol{\phi}_1 = \boldsymbol{\phi}_2 \quad \text{vs.} \quad H_1 : \boldsymbol{\phi}_1 \neq \boldsymbol{\phi}_2,$$

其相应的检验统计量为

$$T_2 = (\hat{\boldsymbol{\phi}}_1 - \hat{\boldsymbol{\phi}}_2)^{\mathrm{T}} \hat{\boldsymbol{S}}^{-1} (\hat{\boldsymbol{\phi}}_1 - \hat{\boldsymbol{\phi}}_1), \tag{2.7}$$

其中 $\hat{\boldsymbol{S}}$ 的定义可参见文献 [12, (8)]. 不过, 这些检验统计量的渐近分布是在经典的 p 固定 N 趋于无穷的框架下成立的, 当微生物种类个数较多且存在大量低计数甚至零计数的菌种时, 收敛到渐近分布的速度很慢, p 值的计算需要借助自举法[14] (bootstrap) 或置换[15] (permutation) 的方法, 对计算资源要求较高.

除了对微生物成分组成进行检验, Chen 和 Li [16] 在 DM 模型下研究了环境协变量 (如饮食摄入) 对微生物成分组成的影响. 在他们使用的实际数据中, 通过调查问卷测得了 98 个志愿者在 214 种微量营养物的摄入数据[5]. 假设除了微生物组计数矩阵, 还观测到了一个协变量矩阵 $\boldsymbol{Z}=(z_{ij})_{n\times q}$ (不妨设 $\boldsymbol{Z}$ 的第一列为 1), 其中 q 可能是高维的. 为将微生物组成成分与协变量联系起来, Chen 和 Li 考虑了 DM 模型第一种参数化的方法, 并假设了如下对数线性模型:

$$\alpha_j(\boldsymbol{z}_i)=\exp\left(\sum_{k=1}^{q}\beta_{jk}z_{ik}\right), \tag{2.8}$$

其中 $\boldsymbol{z}_i=(z_{i1},\dots,z_{iq})^{\mathrm{T}}$, β_{jk} 代表了协变量 k 对菌种 j 的作用. 令 $\boldsymbol{B}=(\beta_{jk})_{p\times q}$, $\boldsymbol{\beta}_k=(\beta_{1k},\dots,\beta_{pk})^{\mathrm{T}}$, 并记 $l(\boldsymbol{B};\boldsymbol{X},\boldsymbol{Z})$ 为将上式代入 (2.4) 后的对数似然函数. 当 p 和 q 较大时, 直接优化似然函数会使估计结果不稳定, Chen 和 Li 进一步考虑了稀疏加分组的惩罚项. 特别地, 令 $\boldsymbol{B}$ 的估计值为

$$\hat{\boldsymbol{B}}=\arg\min_{\boldsymbol{B}}\left\{-l(\boldsymbol{B};\boldsymbol{X},\boldsymbol{Z})+\lambda_1\sum_{k=2}^{q}\|\boldsymbol{\beta}_k\|_2+\lambda_2\sum_{k=2}^{q}\|\boldsymbol{\beta}_k\|_1\right\}, \tag{2.9}$$

其中 λ_1 和 λ_2 分别为分组和稀疏的惩罚参数, Chen 和 Li 建议使用交叉验证 (cross validation) 或者 BIC 准则 (Bayesian information criterion) 来选择惩罚参数. 第一部分关于 $\|\boldsymbol{\beta}_k\|_2$ 的惩罚可以导致分组水平的稀疏性[17], 即对某些 k 而言, $\boldsymbol{\beta}_k$ 估计值的分量全为 0, 代表协变量 k 对微生物组分没有影响. 第二部分针对 $\|\boldsymbol{\beta}_k\|_1$ 的惩罚则可以导致元素水平的稀疏性[18], 即只有少部分 β_{jk} 不为 0, 意味着协变量对菌种成分的作用是稀疏的. Chen 和 Li 提出了分块坐标下降法 (block coordinate descent) 来求解优化问题 (2.9), 但由于模型较为复杂, 没有推导相应的统计性质, 如估计量的相合性以及对参数进行假设检验的方法.

类似地, Wardsworth 等[19] 在 Bayes 框架下考虑了相同的问题, 并在参数 β_{jk} 上使用了针板先验 (spike-and-slab prior) 来达到变量选择的效果. 特别地, Wardsworth 等令 β_{jk} 的先验为

$$\beta_{jk}\sim\xi_{jk}\mathcal{N}(0,\sigma_k^2)+(1-\xi_{jk})\delta_0(\beta_{jk}),$$

其中 ξ_{jk} 是取 0 或 1 的二值变量, δ_0 是 Dirac δ 函数, σ_k^2 刻画了协变量 k 对微生物组分作用的方差. 而对于二值变量 ξ_{jk} 的概率参数 p_{jk}, Wardsworth 等进一步假设其先验是较为平坦的 Beta 分布. 随后, Wardsworth 等提出了使用 MCMC 算法 (Markov chain Monte Carlo) 对后验分布进行统计推断. Wardsworth 等考虑的是完全 Bayes 模型, 在先验的选择上充分利用了共轭先验. 但由于针板先验的存在, 对后验分布的采样难以扩展到更高维度的情况, Bayes 模型往往难以得到解析解, 抽样比较困难也在一定程度上限制了这个方法的应用.

在 DM 模型的基础上, 学者们还考虑了如何对微生物菌种间的相互关系进行建模, 尤其是刻画菌种间的直接相关关系. 这里先简单介绍对于变量相关性的度量方法. 对于随机向量 $\boldsymbol{X}=(X_1,\dots,X_p)^{\mathrm{T}}$, 分量间的相关性由协方差矩阵 $\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$ 给出, 其中 $\sigma_{ij}=\mathrm{Cov}(X_i,X_j)$. 但是, 协方差仅刻画了两个分量间相关性大小, 不能衡量二者的直接关联性. 直接相关性在统计学中往往用偏相关系数或条件独立性描述, 它们分别表示在消除其他变量的线性影响和全部影响后, 对两个变量相互依赖关系的度

量. 当 $\boldsymbol{X}$ 服从 p 维正态分布时, 分量 i 和 j 的偏相关系数为 0 等价于条件独立, 亦等价于精度矩阵 (precision matrix) $\boldsymbol{\Omega}=(\boldsymbol{\Sigma})^{-1}=(\omega_{ij})_{p\times p}$ 中的 $\omega_{ij}=0$.

Yang 等[20] 构造了一个分层模型 mLDM (metagenomic Lognormal-Dirichlet-Multinomial) 以同时估计微生物菌种间的直接相关以及微生物与环境协变量的作用. 在 Chen 和 Li 考虑的对数线性模型 (2.8) 的基础上, Yang 等进一步假设

$$\alpha_j(\boldsymbol{z}_i)=\exp\left(\sum_{k=1}^{q}\beta_{jk}z_{ik}+w_{ij}\right), \tag{2.10}$$

其中

$$\boldsymbol{w}_i=(w_{i1},\ldots,w_{ip})^{\mathrm{T}}\sim\mathcal{N}_p(\boldsymbol{\mu},(\boldsymbol{\Omega})^{-1}),$$

这里 $\boldsymbol{\mu}$ 为未知参数. 此时我们感兴趣的参数是 $\boldsymbol{B}$ 和 $\boldsymbol{\Omega}$, 它们分别描述了协变量对微生物组分的作用及菌种间的直接相关关系. 考虑到菌种和协变量个数均为高维, Yang 等使用了带惩罚的似然方法, 对 $\|\boldsymbol{B}\|_1$ 和 $\|\boldsymbol{\Omega}\|_1$ 都进行了惩罚, 使得估计出的两种关联性都是稀疏的. 不过, 由于全似然函数过于复杂, Yang 等把中间层本应需要的隐变量 $\boldsymbol{w}_i$ 也作为参数同 $\boldsymbol{B}$、$\boldsymbol{\mu}$ 和 $\boldsymbol{\Omega}$ 写进似然, 然后使用分块坐标下降法进行参数估计.

与此同时, Biswas 等[21] 也考虑了相同的问题, 他们使用了 Poisson 对数正态模型, 即假设读段数服从 Poisson 分布, 同一样本不同菌种的 Poisson 分布参数取对数后形成的向量具有形如 (2.10) 的表示形式. 其方法的稀疏性假设和算法求解与 mLDM 方法大同小异, 这里不再赘述. 需要指出, Biswas 等的模型 MInt (microbial interaction) 没有直接考虑微生物组数据的成分特征, 而是在协变量中加入了测序深度作为弥补, 另外该模型也没有考虑协变量是高维的情形. 以上这两种方法均是从计算生物学的角度出发, 用计数数据对菌种直接相关性进行建模. 由于分层模型涉及层数和参数都较多, 在计算和优化上都有一定的难度, 也不适用于维度较高的情形.

2.2 DM 模型的扩展

DM 模型虽然部分解决了微生物组数据超散布性的问题, 但其相较于多项分布而言也仅增加了一个散布参数, 能刻画的数据结构仍然较为有限, 尤其是对菌种间相互作用的描述具有很大的局限性. 事实上, 一个 p 维的 Dirichlet 分布可以视为由 p 个相互独立的 Gamma 分布经归一化生成, 这蕴含了 DM 模型中不同变量间的相关性结构: 一方面, 多项分布的总计数固定导致不同分量间有负相关的趋势; 另一方面, Dirichlet 分布的生成机制意味着在不考虑总量限制时各分量是相互 “独立” 的. 但实际上, 微生物菌种间既有负相关, 也有正相关[22].

一个自然的改进是把单一的 Dirichlet 分布改为有限个 Dirichlet 分布的混合 (Dirichlet multinomial mixture, DMM), 该模型也利于对样本进行聚类分析. Holmes 等[23] 考虑了 DMM 模型在微生物组学中的应用. 设模型由 K 个 Dirichlet 分布混合而成, 参数分别是 $\boldsymbol{\alpha}_1,\ldots,\boldsymbol{\alpha}_K$, 混合的权重为 $w_1,\ldots,w_K$, 满足 $\sum_{k=1}^{K}w_k=1$. 参考 (2.3), DMM 模型下关于多项分布参数 $\boldsymbol{\pi}_i$ 的概率分布可表为

$$f_{\mathrm{D}}^{\mathrm{mix}}(\boldsymbol{\pi}_i;\boldsymbol{\alpha}_k,w_k,k=1,\ldots,K)=\sum_{k=1}^{K}w_kf_{\mathrm{D}}(\boldsymbol{\pi}_i;\boldsymbol{\alpha}_k).$$

Holmes 等进一步在 α_{kj} 上引入了 Gamma 分布作为先验, 对 DMM 模型进行了 Bayes 统计推断. 不过, 混合模型中一个常见的问题是 K 的选取, Holmes 等提出在完全 Bayes 框架下通过比较不同 K 下

模型的不确定性进行模型选择的方法. DMM 模型可以自然地对样本进行分类或聚类, 但处于同一类中的样本仍然面临着 DM 模型中协方差结构不够丰富的问题.

为了能更灵活地描述计数数据的生成机制, Xia 等[24] 提出了针对微生物组学数据分析的加性逻辑正态多项分布模型[25,26] (additive logistic normal multinomial, ALNM). ALNM 模型假设多项分布的参数在经过加性对数比 (additive log-ratio, alr) 变换后服从正态分布. 特别地, 在选定一个指标 p 作为基准 (baseline) 变量后, alr 变换定义为

$$\boldsymbol{\psi}_i = \mathrm{alr}(\boldsymbol{\pi}_i) = \left\{ \log\left(\frac{\pi_{i1}}{\pi_{ip}}\right), \ldots, \log\left(\frac{\pi_{i,p-1}}{\pi_{ip}}\right) \right\}^{\mathrm{T}}. \tag{2.11}$$

Xia 等研究了 ALNM 模型下微生物组分和环境协变量的关系. 设 $\boldsymbol{\psi}_i \sim \mathcal{N}_{p-1}(\boldsymbol{\mu}_i, \boldsymbol{\Sigma})$, 仍记 $\boldsymbol{z}_i \in \mathbb{R}^q$ 为样本 i 的协变量, Xia 等假设 $\boldsymbol{\mu}_i = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{z}_i$, 其中 $\boldsymbol{B} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_q)^{\mathrm{T}} \in \mathbb{R}^{q\times(p-1)}$ 为系数矩阵. 由于 Xia 等考虑的是 p 较小但协变量维数 q 较大的情形, 他们提出了带分组稀疏惩罚的似然方法估计系数矩阵, 即

$$(\hat{\boldsymbol{B}}, \hat{\boldsymbol{\Sigma}}) = \mathop{\arg\min}_{(\boldsymbol{B},\boldsymbol{\Sigma})} \left\{ -\sum_{i=1}^{n} l(\boldsymbol{B}, \boldsymbol{\Sigma}; \boldsymbol{x}_i, \boldsymbol{z}_i) + \lambda \sum_{j=1}^{q} \|\boldsymbol{\beta}_j\|_2 \right\},$$

其中 $l(\boldsymbol{B}, \boldsymbol{\Sigma}; \boldsymbol{x}_i, \boldsymbol{z}_i)$ 是观测样本 i 在 ALNM 模型下的对数似然, λ 是惩罚参数. 这样的惩罚可以导致系数矩阵 $\hat{\boldsymbol{B}}$ 的分组稀疏性, 即某些协变量对微生物组分没有影响. Xia 等随之提出了 MCEM (Monte Carlo expectation maximization) 算法求解上述优化问题, 使用 5 折交叉验证来选择惩罚参数. ALNM 模型更为灵活的代价是参数个数的增加, 且失去了共轭分布计算的简便性, 这使得上述 MCEM 算法不能应用于菌种个数较多的情形中. 此外, ALNM 模型的另一个问题是, 不同的基准变量选择可能导致后续的模型选择不具有不变性, 亦即, 当选定另一变量 (如第一个) 作为基准变量时, alr 变换后的变量仍服从正态分布, 但均值和方差均需进行线性变换, 这可能影响后续建模中对参数结构的假设 (如稀疏性).

此外, DM 模型还没有考虑系统发生树的结构, 而该信息往往有助于微生物组数据的分析. 为此, Wang 和 Zhao[27] 提出了 Dirichlet 树多项分布 (Dirichlet-tree multinomial, DTM) 模型[28]. 与 DM 模型在所有菌种组成的成分向量上假设 Dirichlet 分布不同, DTM 模型对树中的每个子树都引入了一个 Dirichlet 分布. 以图 1 为例, 树中共有 6 个叶结点, 5 个内部结点. 记叶结点组成的集合为 $\mathcal{L}$, 内部结点组成的集合为 $\mathcal{V}$. 对每个内部结点 $v \in \mathcal{V}$, 记 $\mathcal{C}_v$ 为其子结点组成的集合, 则每个内部结点与其子结点组成的子树都可刻画一个多项分布, 因父结点的读段数是其子结点读段数之和. 设由 $v \in \mathcal{V}$ 和 $\mathcal{C}_v$ 诱导的多项分布参数为 $\boldsymbol{b}_v = \{b_{vc} : c \in \mathcal{C}_v\}$, 则有 $b_{vc} > 0$, $\sum_{c\in\mathcal{C}_v} b_{vc} = 1$. 再定义 $\delta_{vc}(l)$ 为由 v 经 c 是否能最终到达叶结点 l 的示性函数, 即当 l 在以 c 为根结点的子树内时 $\delta_{vc}(l) = 1$, 否则 $\delta_{vc}(l) = 0$. 此时 DM 模型中多项分布的参数 π_j 可以由一系列 b_{vc} 表示,

$$\pi_j = \prod_{v\in\mathcal{V}} \prod_{c\in\mathcal{C}_v} b_{vc}^{\delta_{vc}(j)}.$$

这里暂时忽略了关于样本 i 的脚标, 并记

$$x_{vc} = \sum_{j\in\mathcal{L}} \delta_{vc}(j) x_j, \quad x_{v+} = \sum_{c\in\mathcal{C}_v} x_{vc},$$

则多项分布的概率分布可表示为

$$f_{\mathrm{M}}(\{\boldsymbol{b}_v, v \in \mathcal{V}\}; \boldsymbol{x}) = \prod_{v\in\mathcal{V}} \frac{\Gamma(x_{v+}+1)}{\prod_{c\in\mathcal{C}_v}\Gamma(x_{vc}+1)} \prod_{c\in\mathcal{C}_v} b_{vc}^{x_{vc}}.$$

此时, 我们在每个子树诱导的多项分布参数 $\boldsymbol{b}_v$ 上都引入一个 Dirichlet 分布, 记其参数为 $\boldsymbol{\alpha}_v$, 则 DTM 模型的概率分布可表为

$$\begin{aligned} f_{\mathrm{DTM}}(\{\boldsymbol{\alpha}_v, v\in\mathcal{V}\};\boldsymbol{x}) &= \prod_{v\in\mathcal{V}}\int f_{\mathrm{M}}(\{\boldsymbol{b}_v, v\in\mathcal{V}\};\boldsymbol{x}) f_{\mathrm{D}}(\boldsymbol{b}_v;\boldsymbol{\alpha}_v)\,d\boldsymbol{b}_v \\ &= \prod_{v\in\mathcal{V}}\frac{\Gamma(x_{v+}+1)\Gamma(\alpha_{v+})}{\Gamma(x_{v+}+\alpha_{v+})}\prod_{c\in\mathcal{C}_v}\frac{\Gamma(x_{vc}+\alpha_{vc})}{\Gamma(x_{vc}+1)\Gamma(\alpha_{vc})}, \end{aligned}$$

其中 $\alpha_{v+}=\sum_{c\in\mathcal{C}_v}\alpha_{vc}$. Wang 和 Zhao 考虑了协变量对微生物组分的回归问题, 即对每个 α_{vc} 假设如下对数线性模型:

$$\log(\alpha_{vc}) = \boldsymbol{z}^{\mathrm{T}}\boldsymbol{\beta}_{vc},$$

其中 $\boldsymbol{z}$ 是协变量, $\boldsymbol{\beta}_{vc}$ 刻画了所有协变量对树中 v 到 c 这条枝的作用. 将代表样本的脚标 i 代回 b、α、x 和 z 中后, 我们希望估计参数 $\boldsymbol{\beta}=\{\boldsymbol{\beta}_{vc}, v\in\mathcal{V}, c\in\mathcal{C}_v\}$. 记 DTM 模型的对数似然为 $l_{\mathrm{DTM}}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{Z})$, Wang 和 Zhao 考虑了如下稀疏加分组惩罚的极大似然方法估计参数,

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}}\left\{-l_{\mathrm{DTM}}(\boldsymbol{\beta};\boldsymbol{X},\boldsymbol{Z}) + \lambda_1\sum_{v\in\mathcal{V}}\sum_{c\in\mathcal{C}_v}\|\boldsymbol{\beta}_{vc}\|_1 + \lambda_2\sum_{v\in\mathcal{V}}\sum_{c\in\mathcal{C}_v}\|\boldsymbol{\beta}_{vc}\|_2\right\},$$

其中 λ_1 和 λ_2 是惩罚参数, 惩罚思路与 Chen 和 Li 的方法 (2.9) 类似. 不过需要指出, Chen 和 Li 的方法中分组是一个协变量对所有的菌种成分的影响为一组, 组内系数为 0 意味着该协变量完全不影响微生物组分; 而 Wang 和 Zhao 的 DTM 模型中分组是所有协变量对树中的某条枝为一组, 组内系数为 0 意味着这条枝不被任何协变量影响. 最后, Wang 和 Zhao 提出了一个加速近似点梯度法 (accelerated proximal gradient) 求解上述优化问题.

同样在 DTM 模型下, Tang 等[29] 考虑了微生物组分的检验问题, 并加入了树结构的信息. 对于 DTM 模型中每一颗子树的 DM 模型, 都可以用类似于 (2.6) 的统计量进行关于组分的假设检验. 局部的检验可能比全局的 DM 模型检验功效更高, 这是因为在子树上的检验既可能增强其父结点检验的功效, 也可能被父结点的其他子结点对应的检验抵消, 所以, DTM 模型下的检验是否能增强功效取决于树结构是否本身蕴含了足够的信息. 此外, 不妨考虑每个内部结点 $v\in\mathcal{V}$ 对应的检验为

$$H_{0,v}:\boldsymbol{\alpha}_v=\boldsymbol{\alpha}_v^0 \quad \text{vs.} \quad H_{1,v}:\boldsymbol{\alpha}_v\neq\boldsymbol{\alpha}_v^0,$$

其中 $\boldsymbol{\alpha}_v^0$ 是事先给定的向量, 则在全局零假设 $H_0=\bigcap_{v\in\mathcal{V}}H_{0,v}$ 下, 每个子树的检验得到的 p 值是渐近独立的, 因为每个检验只需在给定自己所在子树根结点的总读段数条件下, 即可得到统计量的渐近分布 (参见文献 [29, 定理 1] 或 [30]). 对于这种由树结构诱导的一系列假设检验, Soriano 和 Ma [31] 指出, 局部的检验可能会和与它临近或相互嵌套的检验结果相似, 故联合考虑可能会增强检验的功效. 这种假设在微生物组学中系统发生树很可能成立, 因为在两个分类单元近似或相互包含时, 其组分也往往类似. 因此, 为了对全局零假设 H_0 进行检验, Tang 等[29] 进一步发展了 PhyloScan (phylogenetic scan) 检验方法, 使用树中的所有三元组构造检验统计量. 特别地, 对于每个非根结点且满足 $\mathcal{C}_v\cap\mathcal{V}\neq\emptyset$ 的中间结点 v, 考察其父结点和任一个子结点组成的三元组, 则这三个结点均对应了一个假设检验. 由于在全局零假设下三个检验统计量渐近独立, 容易推导一个把该三元组作为一个整体时对应的统计量及其渐近分布. 此时为了对全局零假设做检验, 需要考虑所有可能的三元组, 显然, 它们对应的统计量 (因三元组之间有交集) 是相互关联的. 结合树的结构, Tang 等给出了这些检验的 p 值取最小值后的近似矫正方法.

除了以上这些 DM 模型的扩展, 还有学者们考虑了更自由的分布假设. 当微生物组测序数据包含了同一个体不同身体部位或不同时间点的样本时, Shi 和 Li [30] 考虑了成对多项分布 (paired multinomial, PairMN) 模型, 在多项分布上层没有对参数进行分布的假设. 假设样本 i 有两组观测值 $\boldsymbol{X}_i = (\boldsymbol{x}_i^1, \boldsymbol{x}_i^2) \in \mathbb{R}^{p\times 2}$, Shi 和 Li 设每组观测仍满足多项分布, 参数分别是 $\boldsymbol{\pi}_i^1$ 和 $\boldsymbol{\pi}_i^2$, 但仅对参数的一阶和二阶矩做了如下假设:

$$\mathrm{E}(\boldsymbol{\pi}_i^t) = \boldsymbol{\mu}^t, \quad \mathrm{Var}(\boldsymbol{\pi}_i^t) = \boldsymbol{\Sigma}^t, \quad t = 1, 2, \quad \mathrm{Cov}(\boldsymbol{\pi}_i^1, \boldsymbol{\pi}_i^2) = \boldsymbol{\Sigma}^{12}.$$

尽管 PairMN 模型没有假设参数的具体分布, 但是仍可使用矩估计的办法估计参数 (参见文献 [30, 引理 1]). 因此, 对于两总体成分的检验

$$H_0 : \boldsymbol{\mu}^1 = \boldsymbol{\mu}^2 \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}^1 \neq \boldsymbol{\mu}^2,$$

仍可考虑类似于 (2.7) 形式的统计量, 并可推导其在 p 固定、N 趋于无穷时的渐近分布. 当潜在模型仍然是 DM 模型时, PairMN 模型估计的协方差不如在正确的模型假定下使用的极大似然估计量更有效; 反过来, PairMN 模型的矩估计量可能在其他分布下比 DM 模型的检验统计量功效更高. PairMN 模型也可以扩展到树结构上, 只要类似于 DTM 模型在每棵子树上引入 PairMN 分布即可. 而且, 不同子树对应的检验统计量也是渐近独立的, 可以进行上文提到的针对全局零假设的检验. 但是需要指出, 当检验目标是全局零假设时, 尽管这两种方法的局部检验统计量是渐近独立的, 但由于系统发生树的内点较多, 如何更有效地将这些 p 值结合取决于备择假设的形式. 不同的组合形式在零假设下的渐近水平是一致的, 但在不同备择假设下的渐近功效不同.

既然多项分布的参数不必来自 Dirichlet 分布, 同样可以不假设多项分布, 只对计数数据成分的期望做假设. Tang 等[32] 提出了一个更一般的 QCAT (quasi-conditional association test) 方法对微生物组分和环境协变量的关系进行估计和假设检验. 他们提出的模型与 Xia 等[24] 的 ALNM 模型在一阶矩的意义下是等价的, 但没有具体假设多项分布和加性逻辑正态分布. 特别地, 如 (2.11), 设

$$\mathrm{E}(\boldsymbol{x}_i) = N_i\boldsymbol{\pi_i}, \quad \boldsymbol{\psi}_i = \mathrm{alr}(\boldsymbol{\pi}_i), \quad \boldsymbol{\psi}_i = \boldsymbol{B}^{\mathrm{T}}\boldsymbol{z}_i,$$

则可推知对 $j = 1, \ldots, p-1$, 有

$$\pi_{ij} = \frac{\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_j)}{\sum_{j=k}^{p-1}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_k) + 1}, \tag{2.12}$$

而 $\pi_{ip} = 1/\{\sum_{j=1}^{p-1}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_j) + 1\}$. Tang 等提出通过求解如下估计方程估计参数 $\boldsymbol{B}$:

$$\sum_{i=1}^{n}\boldsymbol{S}_i(\boldsymbol{B}) = \sum_{i=1}^{n}(\boldsymbol{x}_i - N_i\boldsymbol{\pi_{ij}})\otimes\boldsymbol{x}_i = \boldsymbol{0}_{p\times q},$$

其中 $\boldsymbol{S}_i(\boldsymbol{B}) = (\boldsymbol{x}_i - N_i\boldsymbol{\pi_{ij}})\otimes\boldsymbol{x}_i$, $\otimes$ 代表 Kronecker 积. 在经典的渐近意义下, 只要矩条件假设成立, 该方法的估计量在计数数据不是多项分布时也是渐近相合的. 后续的假设检验, 例如, 检验 $H_0 : \boldsymbol{\beta}_1 = \boldsymbol{0}$, 可以使用传统的记分统计量 (score statistics), 这里不再赘述. 此外, Tang 等[32] 还考虑了计数数据中 0 太多时的解决办法, 即设 $\delta_{ij} = I\ (x_{ij} \neq 0)$, 其中 $I(\cdot)$ 为示性函数, 则 (2.12) 可改写为

$$\pi_{ij} = \frac{\delta_{ij}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_j)}{\sum_{k=1}^{p-1}\delta_{ik}\exp(\boldsymbol{z}_i^{\mathrm{T}}\boldsymbol{\beta}_k) + \delta_{ip}},$$

从而可以针对数据的正部使用类似的估计方程进行参数估计和假设检验. 对计数数据为 0 的部分, 可以使用广义估计方程 (generalized estimation equations) 的方法进行建模分析. 除了对 0 计数的考虑, Tang 等提出的 QCAT 方法也可以扩展到树上, 其思想与前述方法是类似的. Tang 等[32] 的估计方程中没有惩罚项或约束条件限制, 因此, 当菌种个数较多或协变量个数较多时, 估计方法和检验统计量较为不稳定.

总体而言, 直接使用计数数据分析微生物组成的相互关系或者与协变量的关系时, 由于要对抽样过程进行建模, 关联性分析一般都是在成分参数上考虑回归模型或更复杂的结构. 即使是简单的成分向量假设检验问题, 也要考虑到数据的超散布性, 引入上层分布. 大部分分层模型的优化算法都不能很好地扩展到高维情形, 即菌种个数 p 和协变量个数 q 都很大的情况, 且相应的统计理论也难以建立. 如何直接对计数数据建模, 把微生物组数据的特点考虑进来, 并发展出统计上可靠、计算上可扩展 (scalable) 的方法, 是目前微生物组学研究中的一个重要方向. 最后, 我们将各个方法的目标、特点以及是否可处理高维数据总结在表 1 中以供参考. 这里高维的标准是估计或检验方法是否适用于样本数、菌种数和变量数同时增大的情形. 如果方法建立在经典极限意义下, 则不算高维; 如果方法考虑了稀疏性信息, 则算作高维.

2.3 从计数数据到成分数据

除了对计数数据建模, 还可以把计数数据转化为成分数据后, 再使用针对成分数据的统计模型和方法. 与绝对量数据不同, 成分数据描述了一个总体中各个部分所占的比例, 因此只承载了相对量的信息. 特别地, 设 $\boldsymbol{\pi}$ 为 p 维成分数据, 则 $\boldsymbol{\pi}$ 取值在 $p-1$ 维单形 $\mathcal{S}^{p-1}$ 上, 定义为

$$\mathcal{S}^{p-1}=\left\{\boldsymbol{\pi}=(\pi_1,\ldots,\pi_p)^{\mathrm{T}}:\pi_i>0,\sum_{i=1}^{p}\pi_i=1\right\}.$$

注意, 这里考虑的是各分量严格大于零的情形, 这给后续的基于对数比变换的成分数据分析方法提供了极大便利. 对应到微生物组学研究中, 相当于假设各菌种的绝对丰度不是零, 而测序中的零计数是由于样本太小及测序深度不足造成的. 当考虑的分类单元不是十分接近底层 (如 OTU) 时, 这个假设可以认为是合理的.

对于成分数据分析, 目前最常用的方法都是建立在 Aitchison 提出的对数比变换[25,33] 的基础上, 例如, (2.11) 定义的加性对数比变换, 亦或由下式定义的中心对数比 (centred log-ratio, clr) 变换:

$$\mathrm{clr}(\boldsymbol{\pi})=\left(\log\frac{\pi_1}{g(\boldsymbol{\pi})},\ldots,\log\frac{\pi_p}{g(\boldsymbol{\pi})}\right),\tag{2.13}$$

其中 $g(\boldsymbol{\pi})=(\prod_{i=1}^{p}\pi_i)^{1/p}$ 是 $\boldsymbol{\pi}$ 的几何均值. 对于成分数据的统计分析方法, 我们留待下节中详细讨论. 这里先介绍一些从计数数据估计成分的方法.

最简单的从计数数据到成分数据的办法是把同一样本的读段数进行归一化, 即

$$\hat{\boldsymbol{\pi}}_i^{\mathrm{mle}}=\left(\frac{x_{i1}}{\sum_{j=1}^{p}x_{ij}},\ldots,\frac{x_{ip}}{\sum_{j=1}^{p}x_{ij}}\right)^{\mathrm{T}},$$

这也是多项分布模型 (2.1) 下的极大似然估计. 然而, 由于微生物计数数据中存在着大量的 0, 简单地归一化使得成分数据中有过多的 0, 从而无法继续进行基于对数比变换的成分数据分析方法. 一个简单的填补零计数的方法是给每个数据点增加一定量的伪计数 (pseudocount), 而伪计数大小的选择具

表 1 DM 模型及其拓展模型的比较

方法	目标	特点	微生物数是否高维	协变量数是否高维
DM [12]	估计并检验微生物的相对丰度	利用 Dirichlet 分布来刻画数据的超散布性	否	无协变量
Chen 和 Li [16]	估计微生物的相对丰度, 并研究协变量对微生物成分组成的影响	在 DM 模型的基础上利用对数线性模型考虑协变量的作用, 并同时考虑对稀疏和分组的惩罚	是	是
Wardsworth 等 [19]	估计微生物的相对丰度, 并研究协变量对微生物成分组成的影响	在 Bayes 的框架下研究协变量的作用, 并使用针板先验实现变量选择	是	是
mLDM [20]	同时估计微生物间的直接作用以及微生物和协变量间的作用	同时将微生物间的协方差矩阵以及微生物和协变量间的作用系数矩阵写进似然函数	是	是
MInt [21]	同时估计微生物间的直接作用以及微生物和协变量间的作用	使用 Poisson- 对数正态分层模型建模	是	否
DMM [23]	估计并检验微生物的相对丰度	利用混合 Dirichlet 分布作为先验来刻画超散布性	否	无协变量
ALNM [24]	估计微生物的相对丰度, 并研究协变量对微生物成分组成的影响	利用加性逻辑正态多项分布模型来描述计数数据的生成机制	否	是
DTM [27]	利用系统发生树的结构信息研究协变量对微生物成分组成的影响	利用 Dirichlet 树多项分布考察了协变量对处于不同层级微生物的影响	是	是
PhyloScan [29]	在 DTM 模型下, 考虑微生物组分的检验问题	在检验统计量中加入了树结构的信息	否	无协变量
PairMN [30]	对成对数据估计并检验微生物的相对丰度	以成对多项分布为先验分布, 在多项分布上层没有对参数进行分布的假设	否	无协变量
QCAT [32]	估计微生物的相对丰度, 并研究协变量对微生物成分组成的影响	没有假设具体分布, 只对计数数据的期望进行假设	否	否

有一定的主观性, 其基本原则是不超过数据生成过程中的最小探测精度 (计数数据为 1). 例如, 设伪计数为 c, 则估计量为

$$\hat{\pi}_{ij}^{\text{naive}} = \frac{x_{ij}+c}{\sum_{j=1}^{p}(x_{ij}+c)}.$$

这种非参数的数据填补方法在效果上同 Bayes DM 模型取后验期望是等价的, 即设 $\boldsymbol{\pi}_i$ 的先验是参数为 $\boldsymbol{\alpha}$ 的 Dirichlet 分布, 则给定观测 $\boldsymbol{x}_i$ 下 $\boldsymbol{\pi}_i$ 的后验分布是参数为 $\boldsymbol{x}_i+\boldsymbol{\alpha}$ 的 Dirichlet 分布, 其期望为

$$\mathrm{E}(\pi_{ij} \mid \boldsymbol{x}_i) = \frac{x_{ij}+\alpha_j}{N_i+\alpha_+}.$$

当没有先验知识时, 最常见的做法是令所有 α_j 相等, 即 $\boldsymbol{\alpha} = s\boldsymbol{\phi} = s \cdot (1/p, \ldots, 1/p)^{\mathrm{T}}$, 其中 $\boldsymbol{\phi} = (1/p, \ldots, 1/p)^{\mathrm{T}}$, 满足 $\sum_j \phi_j = 1$. 表 2 给出了不同 s 的选取时对 0 计数的填补值.

不同的先验选择会导致截然不同的对 0 的填补值, 合适的选择往往需要考虑到数据采集的方式, 并具有一定主观性. 此外, 因为填充值十分接近 0, 后续基于对数比变换的分析方法可能对先验参数的选择较为敏感.

由于样本间可能存在着可相互借鉴的信息, 考虑对某一样本的先验参数时可以借助于其他样本的信息. Martín-Fernádez 等[34] 提出了一个几何 Bayes 乘性 (geometric Bayesian-multiplicative, GBM) 的非完全 Bayes 方法进行零的填充. 其基本思想包含三个部分. 一是对 $\boldsymbol{\phi}$ 的选择, Martín-Fernádez 等建议对于样本 i, $\boldsymbol{\phi}_{ij}$ 应使用所有其他样本落在第 j 类的比例, 即

$$\phi_{ij} = \frac{\sum_{\substack{k=1 \\ k \neq i}}^{n} x_{kj}}{N - N_i}.$$

第二, 对于 s_i 的选取, Martín-Fernádez 等结合了 $\boldsymbol{\phi}_i$ 的信息, 令 $s_i = 1/g(\boldsymbol{\phi}_i)$, 其中 $g(\boldsymbol{\phi}_i)$ 是 $\boldsymbol{\phi}_i$ 的几何均值. 最后, Martín-Fernádez 等在使用此先验的基础上进行了乘性的调整, 以保持非零计数部分间的比例不因对零的填补而改变. 特别地, GBM 方法给出的成分的估计量为

$$\hat{\pi}_{ij}^{\mathrm{GBM}} = \begin{cases} \phi_{ij} \dfrac{s_i}{N_i + s_i}, & \text{当 } x_{ij} = 0, \\ \dfrac{x_{ij}}{N_i} \left(1 - \displaystyle\sum_{k, x_{ik}=0} \hat{\pi}_{ik}^{\mathrm{GBM}}\right), & \text{当 } x_{ij} \neq 0. \end{cases}$$

GBM 方法在估计菌种成分时使用了留一法 (leave-one-out) 选取先验, 适当使用了所有样本信息, 在一般的问题中是较为简洁的填充零数据的方法. 但在微生物组学研究中, 我们还可根据数据的特点做进一步假设. 由于不同微生物间相互协作, 菌种间存在着某些共生和排斥关系[35], 因此可以假设不同样本的微生物成分并不是自由地散布在 $p-1$ 维单形空间, 其主要部分是受少数几个公共的因子所影响. 由于成分数据经中心对数比变换后落在实空间, 且中心对数比变换后的数据往往是后续分析的基础, 故在本文作者的一篇尚未发表的文章中[1)], 我们假设每个样本的中心对数比变换后的参数 $\boldsymbol{y}_i = \mathrm{clr}(\boldsymbol{\pi}_i)$ 在堆叠后是一个近似的低秩矩阵, 并提出了核范数惩罚的极大似然估计量 LR^2. 特别地, 多项分布关于 clr 参数 $\boldsymbol{Y} = (\boldsymbol{y}_1, \ldots, \boldsymbol{y}_n)$ 的对数似然函数可写为

$$\log(f_{\mathrm{M}}(\boldsymbol{X}; \boldsymbol{Y})) = \sum_{i,j} x_{ij} \log \pi^j(\boldsymbol{y}_i) + C,$$

表 2　DM 模型几种常见的先验参数的选择 [34]

先验	s	$\alpha_j = s/p$	$x_{ij} = 0$ 时的填补值
Haldane	0	0	0
Perks	1	$\frac{1}{p}$	$\frac{1}{p(N_i+1)}$
Jeffreys	$\frac{p}{2}$	$\frac{1}{2}$	$\frac{1}{2N_i+p}$
Bayes-Laplace	p	1	$\frac{1}{N_i+p}$
平方根	$\sqrt{N_i}$	$\frac{\sqrt{N_i}}{p}$	$\frac{1}{p(\sqrt{n}+1)}$

1) Wu C, Gao Z, Deng M, et al. LR^2: Low-rank matrix recovery for large sparse count data. Technical Report, 2017.

其中 C 是与参数无关的常数, 而

$$\boldsymbol{\pi}(\boldsymbol{y}_i)=(\pi^1(\boldsymbol{y}_i),\ldots,\pi^p(\boldsymbol{y}_i))=\left(\frac{\exp(y_{i1})}{\sum_j\exp(y_{ij})},\ldots,\frac{\exp(y_{ip})}{\sum_k\exp(y_{ij})}\right) \tag{2.14}$$

是 clr 逆变换. 注意, 由于上式对应的变换不是一一映射, 为了保证参数的可识别性, 假设 $\sum_j y_{ij}=0$. 从而, LR^2 估计量中 $\hat{\boldsymbol{Y}}^{\mathrm{LR}^2}$ 定义为如下优化问题的解:

$$\min_{\|\boldsymbol{Y}\|_{\max}\leqslant\alpha,\boldsymbol{Y}\mathbf{1}_p=\mathbf{0}}\left\{-\frac{1}{N}\sum_{1\leqslant i\leqslant n,1\leqslant j\leqslant p}x_{ij}\log\pi^j(\boldsymbol{y}_i)+\lambda\|\boldsymbol{Y}\|_*\right\}, \tag{2.15}$$

其中 α 是参数矩阵 $\boldsymbol{Y}$ 元素无穷范的上界, $\mathbf{1}_p$ 是分量全为 1 的 p 维向量, λ 是控制核范数惩罚强度的参数. 矩阵的核范数是矩阵的秩的凸松弛[36], 求解上式可以使得所得参数矩阵是低秩的. 解出参数矩阵 $\hat{\boldsymbol{Y}}^{\mathrm{LR}^2}$ 后, 成分矩阵的估计可由 (2.14) 求出. 我们提出了 ADMM (alternating Direction method of multipliers) 算法求解优化问题 (2.15), 同时证明了当真实参数矩阵是低秩矩阵时 (设其秩为 r), 在合适的惩罚参数选取下, 有

$$\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}^{\mathrm{LR}^2}\|_F^2=O_p\left(\frac{rnp(n\vee p)\log(n+p)}{N}\right)\quad\text{和}\quad\mathrm{KL}(\boldsymbol{\Pi},\hat{\boldsymbol{\Pi}}^{\mathrm{LR}^2})=O_p\left(\frac{rn(n\vee p)\log(n+p)}{N}\right), \tag{2.16}$$

其中 $\boldsymbol{\Pi}=(\boldsymbol{\pi}_1,\ldots,\boldsymbol{\pi}_n)^{\mathrm{T}}$, $n\vee p=\max\{n,p\}$, $\mathrm{KL}(\boldsymbol{\Pi},\hat{\boldsymbol{\Pi}})$ 为两个成分矩阵每行的 KL 散度 (Kullback-Leibler divergence) 之和, 定义为

$$\mathrm{KL}(\boldsymbol{\Pi},\hat{\boldsymbol{\Pi}})=\sum_{i,j}\pi_{ij}\log\frac{\pi_{ij}}{\hat{\pi}_{ij}}.$$

此外, 我们还用极小极大 (minimax) 理论给出了估计的下界, 即当真实参数矩阵是低秩矩阵, 即 $\boldsymbol{Y}\in\mathcal{Y}=\{\boldsymbol{Y}\in\mathbb{R}^{n\times p}:\boldsymbol{Y}\mathbf{1}_p=\mathbf{0},\|\boldsymbol{Y}\|_{\max}\leqslant\alpha,\mathrm{rank}(\boldsymbol{Y})\leqslant r\}$ 时, 存在常数 $\delta>0$ 使得

$$\inf_{\hat{\boldsymbol{Y}}}\sup_{\boldsymbol{Y}\in\mathcal{Y}}\mathrm{P}_{\boldsymbol{Y}}\left(\mathrm{KL}(\boldsymbol{\Pi}(\boldsymbol{Y}),\boldsymbol{\Pi}(\hat{\boldsymbol{Y}}))\geqslant C_\alpha\frac{rn(n\vee p)}{N}\right)\geqslant\delta$$

和

$$\inf_{\hat{\boldsymbol{Y}}}\sup_{\boldsymbol{Y}\in\mathcal{Y}}\mathrm{P}_{\boldsymbol{Y}}\left(\|\boldsymbol{Y}-\hat{\boldsymbol{Y}}\|_F^2\geqslant C_\alpha\frac{rnp(n\vee p)}{N}\right)\geqslant\delta.$$

通过比较以上两式和 (2.16) 可以看出, 我们的方法在相差一个对数因子的意义下是最优的.

Cao 等[37] 考虑了类似的成分矩阵估计问题. 与我们的假设不同, Cao 等假设成分矩阵 $\boldsymbol{\Pi}$ 本身是一近似的低秩矩阵, 并提出了与 (2.15) 类似的核范数惩罚极大似然方法, 推导了其收敛速度和极小极大下界. 由于在迭代过程中对矩阵做奇异值阈值法 (singular value thresholding) 会导致低秩矩阵不在单形空间中, Cao 等提出了广义加速近似点梯度法求解其优化问题. 但需要指出, 成分矩阵中每行的元素位于 $p-1$ 维单形上而非传统的欧式空间, 假定成分矩阵低秩既无非负矩阵分解[38] 的可解释性, 也失去了欧式空间中计算的简便性.

除了以上介绍的非参数 (或 Bayes) 和大维低秩矩阵恢复的方法估计成分数据, 前文介绍的分层模型也可以在模型求解后自动得到各样本的微生物组分. 不过, 由于已经在计数数据模型中对感兴趣的科学问题进行了建模, 因此用这些模型将计数数据转化为成分数据后再进行成分数据分析似无必要. 但是, 分层模型中广泛存在的问题是中间层参数需要通过积分才能写出似然函数: 当使用最简单的 DM 模型时, Dirichlet 分布与多项分布共轭会使似然函数比较简单, 但不能描述复杂的相关性结构;

当使用其他较为复杂的模型时, 又导致似然函数积分后没有显式表达, 从而不可能对高维数据 (菌种较多时) 进行建模.

为此, Zhang 和 Lin[2)] 从算法角度讨论了 ALNM 模型在维度相对较高时的求解方法. Xia 等[24] 使用了 MCEM 算法进行优化, 在其数值实验中最多考虑了 5 个菌种, 因此只能对系统发生树中处于很高层级的微生物种类进行建模. Zhang 等提出了 SAEM (stochastic approximation EM) 算法, 在期望步使用了 HMC (Hamiltonian Monte Carlo) 对 EM 算法中的 Q 函数进行随机近似, 并证明了提出的 SAEM 算法可以收敛到局部极值. 当菌种数量较多时, ALNM 模型中正态变量的协方差矩阵维度过高, 会使计算结果十分不稳定. 为了解决这个问题, Zhang 和 Lin 还提出了带惩罚的 SAEM 算法 R-SAEM, 对协方差矩阵的条件数进行了惩罚. 这里需要注意的是, 为了保证 ALNM 模型中不同基准变量的选择不影响相应的上层正态分布的参数, Zhang 和 Lin 没有直接惩罚协方差矩阵 $\boldsymbol{\Sigma}$, 而是考虑了在基准变量变换下的不变量[33]

$$\boldsymbol{H}^{-\frac{1}{2}}\boldsymbol{\Sigma}\boldsymbol{H}^{-\frac{1}{2}},$$

其中 $\boldsymbol{H}=\boldsymbol{I}_{p-1}+\boldsymbol{J}_{p-1}$, $\boldsymbol{I}_{p-1}$ 是 $p-1$ 阶单位阵, $\boldsymbol{J}_{p-1}$ 是 $p-1$ 阶元素全为 1 的矩阵. 最后, Zhang 等证明了当真值协方差矩阵满足条件数较小的限制时, R-SAEM 算法给出的估计量比不惩罚的 SAEM 方法具有更小的风险 (risk). 在数值模拟中, R-SAEM 算法可以对 $p=150$ 进行计算, 基本满足了微生物组数据中以属 (genus) 为单位时的统计分析.

3 高维成分数据分析

上一小节中, 我们介绍了几种从计数数据估计成分数据的方法. 本节假设已经得到了微生物组成分数据, 并回顾近年来针对高维成分数据的统计方法. 成分数据落在单形中, 满足分量求和为定值的限制, 常规的统计分析方法可能导致错误的结论甚至不适用[33]. 此外, 如前文所述, 微生物组数据中的高维属性给统计分析带来了新的挑战, 我们往往需要对高维数据做进一步的结构性假设, 例如, (2.9) 中的稀疏性假设或 (2.15) 中的低秩性假设. 下面将从成分数据回归模型和基于对数基底 (log basis) 的成分数据统计推断两个方面进行本节的综述.

3.1 高维成分数据回归模型

当我们想研究肠道微生物成分组成对人体健康状况的影响时, 一个自然的做法是, 使用微生物成分数据作为自变量对感兴趣的临床指标 (如衡量肥胖的 BMI 指数) 进行回归分析. 我们仍考虑 n 个样本, 每个样本的成分组成为 $\boldsymbol{x}_i=(x_{i1},\ldots,x_{ip})^{\mathrm{T}}\in\mathcal{S}^{p-1}$, 响应变量为 $y_i\in\mathbb{R}$, 记 $\boldsymbol{y}=(y_1,\ldots,y_n)^{\mathrm{T}}$, $\boldsymbol{X}=(\boldsymbol{x}_1,\ldots,\boldsymbol{x}_n)^{\mathrm{T}}$. 假如我们考虑常规的回归模型

$$\boldsymbol{y}=\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\varepsilon},$$

其中 $\boldsymbol{\beta}$ 是回归系数, $\boldsymbol{\varepsilon}$ 是分量相互独立的白噪声, 满足 $\varepsilon_i\sim\mathcal{N}(0,\sigma^2)$, 那么对于成分数据的协变量矩阵 $\boldsymbol{X}$, 我们会遇到如下问题. 首先, 回归系数的含义难以解释, 因为成分数据的定和限制使得我们不可能只变动成分向量中的一个分量, 所以, β_j 不能简单地解释为菌种 j 对响应变量的效应. 直接使用 Lasso 等[18] 稀疏回归的方法甚至可能只选择出一个变量, 导致选出的子模型完全无法阐释. 其次, 简单地去掉回归中一个分量的做法也不可取. 一是回归结果可能依赖于去掉的是哪个分量, 而参数的结构

2)Zhang J, Lin W. Logistic normal multinomial model for sparse contingency table analysis. Technical Report, 2017.

性假设可能对分量的选取不具有不变性. 二是成分数据去掉一个分量后仍然需满足非负及加和小于 1 的限制, 回归系数不能反映自变量靠近边界时对因变量的作用.

基于以上因素, Aitchison 和 Bacon-Shone [39] 提出了线性对数比模型 (linear log-contrast model), 定义为

$$\boldsymbol{y} = \boldsymbol{Z}^p \boldsymbol{\beta}_{\setminus p} + \boldsymbol{\varepsilon}, \tag{3.1}$$

其中 $\boldsymbol{Z}^p = \{\log(x_{ij}/x_{ip})\}$ 是 $n\times(p-1)$ 维以菌种 p 为基准的对数比设计矩阵, $\boldsymbol{\beta}_{\setminus p} = (\beta_1, \ldots, \beta_{p-1})$ 为回归系数. 为了使高维回归方法 (如 Lasso) 可以应用到成分数据回归模型中, 保证稀疏性假设和相应统计方法对基准变量的选择不变性, Lin 等[40] 把菌种 p 也引入到回归模型中, 令 $\beta_p = -\sum_{j=1}^{p-1}\beta_j$, 则 (3.1) 可写为

$$\boldsymbol{y} = \boldsymbol{Z}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \quad \mathbf{1}_p^{\mathrm{T}}\boldsymbol{\beta} = 0, \tag{3.2}$$

其中 $\boldsymbol{Z} = \{\log(x_{ij}/x_{ip})\} \in \mathbb{R}^{n\times p}$. Lin 等提出了 ℓ_1 惩罚方法估计回归系数, 即

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right) \quad \text{s.t.} \quad \sum_{j=1}^{p}\beta_j = 0, \tag{3.3}$$

其中 λ 为惩罚参数. Lin 等提出了将增广 Lagrange 法 (augmented Lagrange method) 和坐标下降 (coordinate descent) 结合的算法求解上式, 并证明了在稀疏模型和一定的不可表示条件 (irrepresentable condition)[41] 及最小信号强度条件下, (3.3) 定义的估计量满足

$$\|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_1 = O_p\left(s\sqrt{\frac{\log p}{n}}\right) \quad \text{和} \quad \operatorname{sign}(\hat{\boldsymbol{\beta}}) = \operatorname{sign}(\boldsymbol{\beta}),$$

其中 s 为 $\boldsymbol{\beta}$ 的非零元的个数. 该结果与一般的高维线性回归 Lasso 的理论结果类似[41], 但是是在回归系数有零和限制的情况下推导的.

Shi 等[42] 在以上工作的基础上进一步结合了树结构的信息, 把子成分 (subcomposition) 选择的一致性融入到模型之中, 并提出了去偏差 (de-biased) 估计量对回归系数进行区间估计. 从系统发生树的结构我们知道, 所有微生物菌种一起构成一个成分向量, 而仅看一个高层水平分类单元下的菌种 (如同一科下的不同属), 也可以在局部构成一个子成分, 满足子成分的分量之和为 1. 因此, 如果在回归模型中考虑到较高水平分类单元下的子成分也具有成分属性的特征时, 亦应在系数上增加相应的零和限制. 特别地, 假设不同菌种被分为了互不相交的 G 类, 第 g 类中有 p_g 个菌种, 则在此类下的子成分满足

$$\sum_{j=1}^{p_g} x_{i,gj} = 1, \quad g = 1, \ldots, G,$$

其中 $x_{i,gj}$ 表示样本 i 在第 g 类中的第 j 个菌种的相对成分, 且 $\sum_{g=1}^{G} p_g = p$. 此时参考 (3.2), 我们同样以 $\log(x_{i,gj})$ 作为协变量, 此时系数的限制条件应满足

$$\sum_{j=1}^{p_g} \beta_{gj} = 0, \quad g = 1, \ldots, G.$$

因此, Shi 等提出了以下估计量:

$$\hat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \left(\frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{Z}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\right) \quad \text{s.t.} \quad \boldsymbol{C}^{\mathrm{T}}\boldsymbol{\beta} = \mathbf{0},$$

其中

$$\boldsymbol{C}^{\mathrm{T}}=\begin{pmatrix}1&\cdots&1&0&\cdots&0&0&\cdots&0\\0&\cdots&0&1&\cdots&1&0&\cdots&0\\\vdots&\ddots&\vdots&\vdots&\ddots&\vdots&\vdots&\ddots&\vdots\\0&\cdots&0&0&\cdots&0&1&\cdots&1\end{pmatrix}.$$

除了对参数进行估计, Shi 等还沿袭了高维线性回归中对参数构建置信区间和做假设检验的方法[43], 在成分数据回归中提出了去偏差估计量以进行统计推断. 其核心思想是, ℓ_1 惩罚会将所有回归系数的估计值向 0 压缩, 以增大偏差减少方差的方式产生稀疏的解, 故将估计量去除偏差后, 即可证明去偏差估计量在真值附近满足正态分布. 而偏差虽然不能精确计算, 但在一定条件下可以证明偏差的估计误差比参数的估计误差小一个 $\frac{s\log p}{\sqrt{n}}$ 的阶, 从而, 当 $s\log p=o(\sqrt{n})$ 时, 偏差的估计误差渐近可以忽略, 进而可对参数构建置信区间或进行假设检验. 去偏差估计量可以表示为

$$\hat{\boldsymbol{\beta}}^{u}=\hat{\boldsymbol{\beta}}+\frac{1}{n}\tilde{\boldsymbol{M}}\tilde{\boldsymbol{Z}}^{\mathrm{T}}(\boldsymbol{y}-\tilde{\boldsymbol{Z}}\hat{\boldsymbol{\beta}}),$$

其中 $\tilde{\boldsymbol{Z}}=\boldsymbol{Z}(\boldsymbol{I}_p-\boldsymbol{C}\boldsymbol{C}^{\mathrm{T}})$, 而 $\tilde{\boldsymbol{M}}$ 的具体定义可以参见文献 [42, 算法 2 和定理 1].

Wang 和 Zhao[44] 也考虑了在成分数据回归模型中引入树结构信息的问题. 除了回归系数的稀疏性, 他们还在子成分选择 (subcomposition selection) 的水平上做了稀疏性假设, 即在所有的可能的子成分 (系统发生树的任一内点诱导的最底层菌种的组合为一种子成分) 中, 只有某些子成分以成分数据的形式 (即相应回归系数之和为零) 对响应变量有作用. 回顾前文使用的记号, 令 $\mathcal{V}$ 为内部结点的集合, $\mathcal{L}$ 为叶结点的集合, 对 $v\in\mathcal{V}$, 令

$$\boldsymbol{f}_v=\sum_{j\in\mathcal{L}_v}\boldsymbol{e}_j,$$

其中 $\boldsymbol{e}_j\in\mathbb{R}^p$ 是第 j 位置为 1 其他位置为 0 的标准正交基向量, $\mathcal{L}_v\subset\mathcal{L}$ 是由 v 延伸出的叶结点的集合. Wang 和 Zhao 考虑了如下名为 TASSO (tree-guided automatic subcomposition selection operator) 的估计方法:

$$\min_{\boldsymbol{\beta}}\frac{1}{2n}\|\boldsymbol{y}-\tilde{\boldsymbol{Z}}\boldsymbol{\beta}\|_2^2+\lambda_1\|\boldsymbol{\beta}\|_1+\lambda_2\sum_{v\in\mathcal{V}}|\boldsymbol{f}_v^{\mathrm{T}}\boldsymbol{\beta}|,\tag{3.4}$$

其中 $\tilde{\boldsymbol{Z}}$ 为将成分数据 $\boldsymbol{X}$ 每行做中心对数比变换后的设计矩阵, 惩罚项的第一部分是最底层菌种的选择, 由 λ_1 控制惩罚强度, 第二部分是针对子成分的选择, 由 λ_2 控制惩罚强度. 由于 $\tilde{\boldsymbol{Z}}$ 的引入, 以及子成分不是预先给定而是通过变量选择的方法自动挑选的特点, TASSO 估计量不需要对 $\boldsymbol{\beta}$ 的任一部分做零和假设. 但此时由于 $\tilde{\boldsymbol{Z}}$ 是退化的, 为计算简便, Wang 和 Zhao 又引入了 $\frac{\gamma}{2n}\|\boldsymbol{\beta}\|_2^2$ 惩罚项进 (3.4) 中, 变换后等价于一个广义 Lasso 问题的求解, 从而可以使用已有的软件包进行优化. 但是, 由于惩罚项较为复杂, TASSO 算法的理论性质难以推导.

Wang 和 Zhao [45] 没有使用线性对数比模型, 而是直接把成分数据当作协变量, 并提出了类似的树结构惩罚函数. 此时为了考虑到协变量的成分属性, 他们使用了融合 Lasso (fused Lasso) 的想法, 把回归系数间所有可能的两两的差的绝对值和作为惩罚项, 以保证模型的可识别性. 不过, 该方法并未在思想或理论上超出前文所述的高维成分数据回归的方法, 这里不再详细介绍了, 感兴趣的读者可参见文献 [45].

3.2 基于对数基底的成分数据统计推断

由于成分数据仅包含了各组分的相对含量信息, 一个自然的想法是, 成分数据是以如下形式生成的:

$$x_{ij} = \frac{w_{ij}}{\sum_j w_{ij}}, \quad j = 1, \ldots, p, \tag{3.5}$$

其中 w_{ij} 代表样本 i 中菌种 j 的绝对丰度, 也被称为基底 (basis), 是我们无法直接观测到的. 从 (3.5) 可以看出, 取对数并对不同成分做差后, 有

$$\log x_{ij} - \log x_{ik} = \log w_{ij} - \log w_{ik},$$

即两菌种成分的对数做差与其对数基底之差相等, 这一关系式使得我们在没有观测到绝对量的情况下依然有可能对其进行统计分析, 即我们感兴趣的是对数基底的统计性质,

$$\boldsymbol{u}_i = (u_{i1}, \ldots, u_{ip})^{\mathrm{T}} = (\log w_{i1}, \ldots, \log w_{ip})^{\mathrm{T}},$$

但我们只有成分数据的信息. 下面从三个方面介绍基于对数基底的成分数据统计推断方法.

3.2.1 两样本均值检验

第 2 节回顾了几种对计数数据进行两样本成分检验的模型和方法. 本小节介绍由 Cao 等[3)] 提出的直接针对高维成分数据的两样本均值检验方法, 亦可参见文献 [46]. 对于成分数据的两样本均值检验, 已有的方法一般是针对固定维度的情况, 即 $p \ll n$, 例如, Aitchson 在专著 [33, 第 7.5 小节] 中提到的广义似然比检验. 然而当变量个数与样本量相当甚至大于样本量时, 经典的检验方法可能功效不高, 甚至无法使用. 对于没有成分属性的高维两样本均值检验, 近年来已发展出一些相应的统计方法, 参见文献 [47] 及其中的引用. 在 Cai 等[48] 介绍的基于极大值检验统计量的基础上, Cao 等提出了针对成分数据的高维两样本均值检验方法.

具体地, 假设两总体的对数基底分别是 $\boldsymbol{u}^{(1)}$ 和 $\boldsymbol{u}^{(2)}$, 其均值分别为 $\boldsymbol{\mu}^{(1)}$ 和 $\boldsymbol{\mu}^{(2)}$, 我们想检验

$$H_0 : \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} \quad \text{vs.} \quad H_1 : \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)}.$$

然而, $\boldsymbol{\mu}^{(k)}$ $(k = 1, 2)$ 的取值在只观测到成分数据的情况下是不可识别的. 例如, 当 $\boldsymbol{u}^*$ 和 $\boldsymbol{x}^*$ 满足基底和成分的关系式 (3.5) 时, 考虑等价类

$$\mathcal{U}(\boldsymbol{u}^*, \boldsymbol{x}^*) = \{\boldsymbol{u} \in \mathbb{R}^p : \boldsymbol{u} = \boldsymbol{u}^* + c\mathbf{1}_p\},$$

其中 c 为任意常数, 则任意等价类中的元素 $\boldsymbol{u}' \in \mathcal{U}(\boldsymbol{u}^*, \boldsymbol{x}^*)$ 都能依关系式 (3.5) 生成相同的成分向量 $\boldsymbol{x}^*$. 因此, 一个可检验的假设是

$$H_0 : \exists\, c \in \mathbb{R} \text{ s.t. } \boldsymbol{\mu}^{(1)} = \boldsymbol{\mu}^{(2)} + c\mathbf{1}_p \quad \text{vs.} \quad H_1 : \forall\, c \in \mathbb{R}, \boldsymbol{\mu}^{(1)} \neq \boldsymbol{\mu}^{(2)} + c\mathbf{1}_p. \tag{3.6}$$

令 $\boldsymbol{y}_i^{(k)} = \mathrm{clr}(\boldsymbol{x}_i^{(k)})$ 为中心对数比变换后的数据, 其中 $\mathrm{clr}(\cdot)$ 的定义如 (2.13) 所示, 则

$$\boldsymbol{y}_i^{(k)} = \boldsymbol{G}\log(\boldsymbol{x}_i^{(k)}) = \boldsymbol{G}\boldsymbol{u}_i^{(k)},$$

3) Cao Y, Lin W, Li H. Two sample mean tests for high-dimensional compositional data. Technical Report, 2017.

其中 $\boldsymbol{G}=\boldsymbol{I}_p-\frac{1}{p}\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}$. 从而设 $\boldsymbol{\nu}^{(k)}=\mathrm{E}(\boldsymbol{y}_i^{(k)})$, 则有 $\boldsymbol{\nu}^{(k)}=\boldsymbol{G}\boldsymbol{\mu}^{(k)}$, 且因 $\boldsymbol{G}\mathbf{1}_p=\mathbf{0}$, 检验 (3.6) 等价于

$$H_0:\boldsymbol{\nu}^{(1)}=\boldsymbol{\nu}^{(2)}\quad\text{vs.}\quad H_1:\boldsymbol{\nu}^{(1)}\neq\boldsymbol{\nu}^{(2)}. \tag{3.7}$$

为进行假设检验 (3.7), Cao 等提出了如下极大值形式的检验统计量:

$$M_n=\frac{n_1n_2}{n_1+n_2}\max_{1\leqslant j\leqslant p}\frac{(\bar{y}_j^{(1)}-\bar{y}_j^{(2)})^2}{\hat{\gamma}_j}, \tag{3.8}$$

其中 n_1 和 n_2 分别是两总体各自的样本数, $\bar{y}_j^{(k)}=\sum_{i=1}^{n_k}y_{ij}^{(k)}/n_k$ 是变量 j 在各自总体内样本均值, $\hat{\gamma}_j=\sum_{k=1}^{2}\sum_{i=1}^{n_k}(y_{ij}^{(k)}-\bar{y}_j^{(k)})^2/(n_1+n_2)$ 是混合总体中变量 j 的方差. Cao 等证明了在适当的关于 $\boldsymbol{u}^{(k)}$ 的协方差矩阵稀疏性假设下, (3.8) 中不同的 $(\bar{y}_j^{(1)}-\bar{y}_j^{(2)})^2$ 仅仅是弱相关的, 从而当 $(n_1,n_2,p)\to\infty$ 时, 我们有 $M_n-2\log p+\log\log p$ 渐近服从第一类极值分布, 故可据此结论选择临界值, 得到指定水平下的检验. 此外, Cao 等给出了稀疏备择假设下统计量 M_n 的功效. 亦即, 当 $\boldsymbol{\nu}^{(1)}-\boldsymbol{\nu}^{(2)}$ 仅在少数位置上非零时, Cao 等推导了检验统计量拒绝零假设的概率, 参见文献 [46, 定理 9].

Cao 等提出的方法是第一个针对高维成分数据的两样本均值检验方法. 他们使用了极大值类型的检验统计量, 对稀疏的备择假设功效较高. 对于非成分属性的高维两样本均值检验, 还有基于平方和类型的检验统计量[49], 这种方法对于稠密的备择假设 ($\boldsymbol{\nu}^{(1)}-\boldsymbol{\nu}^{(2)}$ 在多位置均不等, 但单独位置的信号较弱) 比较有效. 此外, 还有学者考虑了自适应的假设检验方法[50], 综合极大值及平方和等类型的统计量, 使得其方法对不同类型的备择假设都有较高功效. 这些方法都有可能拓展到高维成分数据中. 最后, 上面提到的几种统计量都是针对全局的均值检验, 即只检验两总体均值是否完全相等, 不能给出当其不相等时有差异的具体位置. 如果要分别考虑每个位置上两总体均值是否相等, 则需同时考察 p 个假设, 如何在成分数据中同时进行多个这种局部的检验并控制错误发现率 (false discovery rate, FDR) 是未来值得研究的一个问题.

3.2.2 协方差矩阵统计推断

除了对均值的估计和检验, 统计学中另一个重要的问题是协方差矩阵相关的统计推断. 它刻画了变量间的相关性, 也是很多统计工具的基础 (如分类、聚类和主成分分析等). 由于成分数据具有定和的限制, 直接研究成分间的相关性会额外引进带有欺骗性的负相关[51]; 而且我们更感兴趣的是对数基底 (即绝对量) 之间的相关. 本小节介绍近年来一些关于 $\boldsymbol{\Sigma}=\mathrm{Cov}(\boldsymbol{u}_i)$ 的统计推断方法.

我们首先简单介绍成分数据相关结构与感兴趣的参数 $\boldsymbol{\Sigma}=(\sigma_{ij})_{p\times p}$ 间的关系, 细节可参见文献 [33, 第 4 和 5 章]. 由于 $\boldsymbol{\Sigma}$ 是潜在的对数基底的协方差矩阵, 在没有观测到绝对量的情况下无法直接估计. 从成分数据出发, 我们介绍两种刻画相关性的矩阵. 第一个是变差矩阵 (variation matrix), 定义为 $\boldsymbol{T}=(\tau_{jk})_{p\times p}$, 其中

$$\tau_{jk}=\mathrm{Var}\left(\log\left(\frac{x_{ij}}{x_{ik}}\right)\right)=\mathrm{Var}\left(\log\left(\frac{u_{ij}}{u_{ik}}\right)\right)=\sigma_{ii}+\sigma_{jj}-2\sigma_{ij}.$$

写成矩阵形式, 有

$$\boldsymbol{T}=\boldsymbol{\sigma}\mathbf{1}^{\mathrm{T}}+\mathbf{1}\boldsymbol{\sigma}^{\mathrm{T}}-2\boldsymbol{\Sigma}, \tag{3.9}$$

其中 $\boldsymbol{\sigma}=(\sigma_{11},\ldots,\sigma_{pp})$. 由上式可以看出, $\boldsymbol{T}$ 是可以从成分数据样本中估计得到的, 但 $\boldsymbol{\Sigma}$ 与 $\boldsymbol{T}$ 之间存在着多对一的关系, 因右边包含了 $p(p+1)/2$ 个自由参数, 而左边受 τ_{jk} 定义的限制只有 $p(p-1)/2$ 个

自由参数. 第二个是中心对数比协方差矩阵, 即中心对数比变换后 $\boldsymbol{y}_i = \mathrm{clr}(\boldsymbol{x}_i)$ 的协方差矩阵, 定义为 $\boldsymbol{\Gamma} = (\gamma_{jk})_{p\times p}$, 满足

$$\gamma_{jk} = \mathrm{Cov}(y_{ij}, y_{ik}) = \sigma_{jk} - \sigma_{j\cdot} - \sigma_{\cdot k} + \sigma_{\cdot\cdot}, \tag{3.10}$$

其中

$$\sigma_{j\cdot} = \frac{1}{p}\sum_{k=1}^{p}\sigma_{jk}, \quad \sigma_{\cdot k} = \frac{1}{p}\sum_{j=1}^{p}\sigma_{jk}, \quad \sigma_{\cdot\cdot} = \frac{1}{p^2}\sum_{j,k=1}^{p}\sigma_{jk}$$

分别代表了 $\boldsymbol{\Sigma}$ 的第 j 行均值、第 k 列均值和总均值. 写成矩阵形式, 则有

$$\boldsymbol{\Gamma} = \boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G}. \tag{3.11}$$

类似地, $\boldsymbol{\Gamma}$ 是可以估计的, 但 $\boldsymbol{\Sigma}$ 不能由 $\boldsymbol{\Gamma}$ 唯一确定.

由于不可能从 $\boldsymbol{T}$ 或 $\boldsymbol{\Gamma}$ 直接估计 $\boldsymbol{\Sigma}$, 学者们提出了基于稀疏假设的方法估计对数基底的协方差矩阵, 即假设 $\boldsymbol{\Sigma}$ 是稀疏的, 从而减少 (3.9) 或 (3.11) 右边自由变量的个数. 在这种情况下, 有可能使得仅有一个满足适当稀疏性要求的协方差阵满足以上两式, 我们将在后面具体讨论参数的可识别性问题. 此外, 稀疏的协方差矩阵在高维数据分析中也是一个常见假设, 因它具有一定生物学意义, 并解决了高维数据带来的问题. 因此, 稀疏性假设在成分数据协方差矩阵估计的框架内起到了一石二鸟的作用.

Friedman 和 Alm [52] 首先提出了 SparCC (sparse correlations for compositional data) 方法, 其目标是估计稀疏的成分数据协方差矩阵 $\boldsymbol{\Sigma}$, 但没有直接考虑稀疏性对可识别性的帮助. SparCC 的主要思想是添加了 p 个额外的假设

$$\sum_{j=1}^{p}\sigma_{ij} = o\Big(\sum_{j}\sigma_{jj}^2\Big), \quad i = 1, \ldots, p, \tag{3.12}$$

然后在 (3.9) 中去除上式的无穷小量后估计 $\boldsymbol{\Sigma}$. 从计算的角度出发, Friedman 和 Alm 又提出了迭代的算法: 如果可以在每次估计后去除相关性较强元素对应的行和列, 然后对剩余变量重复以上算法, 自然 (3.9) 对剩余变量组成的子矩阵仍然成立, 那么多次迭代后剩余变量之间的相关性可能较小, 从而满足 (3.12) 的假设. SparCC 算法是估计 $\boldsymbol{\Sigma}$ 的一个初步的尝试, 在微生物组学研究中成为了一个标杆式的算法. 当然, 它也有一些问题尚待解决. 例如, 当假设 (3.12) 不成立时, 不能保证算法一定能找到真实的强相关; 即使找到的变量 i 和 j 确实是强相关, 同时去掉 i 和 j 所在的整行和整列意味着该行/列中其他的位置只能使用这一步中的估计, 可能并不准确; 描述性的迭代算法在操作上可行, 但难以从理论上证明算法的有效性, 等等.

随后, Fang 等 [53] 和 Ban 等 [54] 分别提出了基于 ℓ_1 惩罚的方法估计稀疏的对数基底协方差矩阵 $\boldsymbol{\Sigma}$, 并在稀疏假设下对参数的可识别性进行了一定探讨. Fang 等指出, 在总体稀疏度的意义下 (用 $\boldsymbol{\Sigma}$ 上三角非对角的非零个数衡量), 则仅当 $\boldsymbol{\Sigma}$ 真值的稀疏度严格小于 $(p-1)/2$ 时, 才不存在另一总体意义上更稀疏的矩阵也满足关系式 (3.9) 或 (3.11). 同时, 确实存在两个总体稀疏度均为 $(p-1)/2$ 的矩阵同时满足 (3.9) 或 (3.11). Fang 等提出的 CCLasso (correlation inference for compositional data through Lasso) 方法建立在 (3.11) 的基础上, 对 $\boldsymbol{\Sigma}$ 的估计由下式给出:

$$\hat{\boldsymbol{\Sigma}} = \mathop{\arg\min}_{\boldsymbol{\Sigma}=\boldsymbol{\Sigma}^{\mathrm{T}}}\left\{\frac{1}{2}\|\boldsymbol{G}\boldsymbol{\Sigma}\boldsymbol{G} - \hat{\boldsymbol{\Gamma}}\|_{\boldsymbol{V}}^2 + \lambda\|\boldsymbol{\Sigma}\|_{1,\mathrm{off}}\right\}, \tag{3.13}$$

其中 $\hat{\boldsymbol{\Gamma}}$ 是中心对数比变换后的样本协方差阵, λ 是针对矩阵稀疏性的惩罚 (因对角元一定大于零, 故不需要惩罚), $\|\cdot\|_{\boldsymbol{V}}^2$ 是以 $\boldsymbol{V}$ 为权的加权 Frobenious 范数的平方, 定义为

$$\|\boldsymbol{A}\|_{\boldsymbol{V}}^2 = \mathrm{tr}(\boldsymbol{A}^{\mathrm{T}}\boldsymbol{V}\boldsymbol{A}).$$

在 CCLasso 算法中, Fang 等令 $\boldsymbol{V}=(\mathrm{diag}(\hat{\boldsymbol{\Gamma}}))^{-1}$, 并提出了一个 ADMM (alternating direction method of multipliers) 算法求解优化问题 (3.13).

Ban 等从另一个稀疏的角度讨论了参数的可识别性问题. 尽管在全局稀疏的意义下, 存在着足够稀疏的两个矩阵 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 满足关系式 (3.9) 或 (3.11), 但 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 的非零元都集中在了同一行内, 即某一个变量同时与许多变量相关, 其他变量间没有相关性 (参见文献 [46, 命题 4]). 因此, 如果从最大度数的角度出发, 其中最大度数定义为

$$\deg_{\max}(\boldsymbol{\Sigma})=\max_i\sum_j I(\sigma_{ij}\neq 0),$$

那么 $\boldsymbol{\Sigma}_1$ 和 $\boldsymbol{\Sigma}_2$ 的最大度数高达 $p/2$. Ban 等证明了, 如果考虑最大度数的稀疏性, 那么当 $\boldsymbol{\Sigma}$ 真值满足 $\deg_{\max}(\boldsymbol{\Sigma})<p/4$ 时, $\boldsymbol{\Sigma}$ 就一定是所有满足关系式 (3.9) 或 (3.11) 的协方差矩阵中最大度数最小的一个. 随后, Ban 等提出了 REBACCA (regularized estimation of the basis covariance based on compositional data) 算法估计 $\boldsymbol{\Sigma}$, 其相当于对 (3.9) 进行线性变换后, 将参数矩阵展开为向量, 得到一线性方程组使用 Lasso 方法对参数进行估计.

Fang 等[53] 和 Ban 等[54] 虽然对成分数据协方差矩阵的参数可识别性进行了一定讨论, 但并没有将估计方法与可识别性联系起来. 亦即, 他们没有证明所得估计量是否在相应的可识别的情况下会收敛到真值. 为此, Cao 等[55] 进一步考虑了模型的近似可识别性 (approximate identifiability), 并在此基础上提出了一个计算上较为简便的 COAT (composition-adjusted thresholding) 方法. 特别地, 从 (3.10) 可以看出, 中心对数比协方差矩阵元素 γ_{jk} 与对数基底协方差矩阵元素 σ_{jk} 间只相差了 $\boldsymbol{\Sigma}$ 的一些均值项. 当 $\boldsymbol{\Sigma}$ 是稀疏矩阵且满足 $\|\boldsymbol{\Sigma}\|_{\ell_1}=o(p)$ 时, γ_{jk} 和 σ_{jk} 只相差了 $o(1)$ 的近似误差 (当 $p\to\infty$). 因此可以使用在常规数据高维协方差矩阵估计的方法[56], 对中心对数比变换后的样本协方差矩阵 $\hat{\boldsymbol{\Gamma}}$ 做自适应阈值法 (adaptive thresholding). COAT 方法在计算简便的同时 (不需求解优化问题), 还可以推导在近似可识别条件下估计量的理论性质, 满足

$$\|\hat{\boldsymbol{\Sigma}}-\boldsymbol{\Sigma}\|_2=O_p\bigg(s(p)\bigg(\sqrt{\frac{\log p}{n}}+\frac{s(p)}{p}\bigg)\bigg),\tag{3.14}$$

其中 $s(p)=\|\boldsymbol{\Sigma}\|_{\ell_1}=o(p)$ 刻画了 $\boldsymbol{\Sigma}$ 的稀疏性. Cao 等的结果是第一个从理论上阐明了对数基底协方差矩阵估计量与真值关系的结果. 但是, 其收敛速率与常规数据高维协方差矩阵估计的极小极大结果[57] 相差了一项近似误差 $\frac{s(p)}{p}$. 这意味着, 当菌种个数 p 较小时, 近似误差会起主导作用, 即使样本量再大也不会减小总的估计误差; 当 p 增大时, 近似误差减小, 高维属性带来的估计误差 $\sqrt{\frac{\log p}{n}}$ 又会成为主导项.

由于以上方法均未能较为完善地解决成分数据协方差矩阵估计中的可识别性问题, Wu 等[4)] 对 (3.11) 带来的的可识别性问题进行了更细致的分析. 容易看出, 对于等价类 $\mathcal{E}(\boldsymbol{\Sigma})=\{\boldsymbol{\Omega}:\boldsymbol{\Omega}=\boldsymbol{\Sigma}+\mathbf{1}_p\boldsymbol{a}^{\mathrm{T}}+\boldsymbol{a}\mathbf{1}_p^{\mathrm{T}},\boldsymbol{a}\in\mathbb{R}^p\}$ 而言, 其中任一元素都可产生相同的 $\boldsymbol{\Gamma}$. 因此, 一个核心问题是, 如何选取合适的稀疏矩阵类使之与 $\mathcal{E}(\boldsymbol{\Sigma})$ 相交仅有 $\boldsymbol{\Sigma}$ 真值这一个元素. Fang 等[53] 考虑的稀疏矩阵类是全局的稀疏性:

$$\mathcal{S}_{\mathrm{g}}(k)=\bigg\{\boldsymbol{\Sigma}:\sum_{i<j}I(\sigma_{ij}\neq 0)\leqslant k\bigg\},$$

4) Wu C, Zhu Y, Deng M, et al. Statistical inference for high-dimensional composition covariance matrix. Technical Report, 2017.

则当 $k < (p-1)/2$ 时两个集合的交唯一. Ban 等[54] 考虑的是最大度数稀疏矩阵类 $\mathcal{S}_{\deg}(k) = \{\mathbf{\Sigma} : \deg_{\max}(\mathbf{\Sigma}) \leqslant k\}$, 则当 $k < p/4$ 时两矩阵类之交唯一. 事实上, 针对 $\mathcal{E}(\mathbf{\Sigma})$ 的结构, $\mathcal{S}_{\deg}(k)$ 是较好的稀疏矩阵类的选择, 因其与矩阵 $\mathbf{1}_p\boldsymbol{\alpha}^{\mathrm{T}} + \boldsymbol{\alpha}\mathbf{1}_p^{\mathrm{T}}$ 具有较好的 "正交性", 且容纳了更丰富的稀疏矩阵. 为了将估计方法与可识别性联系起来, 我们从理论上证明了当 $s(p) = \deg_{\max}(\mathbf{\Sigma}) < p/6$ 时, 如下 ℓ_1 惩罚估计量:

$$\hat{\mathbf{\Sigma}} = \mathop{\arg\min}_{\mathbf{\Sigma}=\mathbf{\Sigma}^{\mathrm{T}}} \left\{ \frac{1}{2}\|\boldsymbol{G}\mathbf{\Sigma}\boldsymbol{G} - \hat{\mathbf{\Gamma}}\|_F^2 + \lambda\|\mathbf{\Sigma}\|_1 \right\} \tag{3.15}$$

满足

$$\|\hat{\mathbf{\Sigma}} - \mathbf{\Sigma}\|_2 = O_p\left(s(p)\sqrt{\frac{\log p}{n}} \right).$$

Fang 等的 CCLasso 估计量 (3.13) 是我们方法的一个加权版本, 而我们首次证明了该方法在最大度数稀疏的情况下可以正确估计真实参数, 并达到了理论上的最优速率[57]. 这里可容许的最大度数从 $p/4$ 加强到 $p/6$, 是为了保证所用的凸优化方法 (从最大度数限制凸松弛到矩阵的元素 ℓ_1 范数) 也能正确估计参数. 我们的理论结果是对 COAT 方法 (3.14) 的改进, 同时放松了假设的条件.

除了对成分数据协方差矩阵进行参数估计, 另一个值得研究的问题是对菌种间是否存在相关性进行假设检验. Faust 等[35] 首先提出了一个非参数的置换重整化自举法 (permutation-renormalization bootstrap, ReBoot), 其思想是直接从菌种成分出发计算其相关性并通过重抽样的方法计算 p 值. 特别地, 由于成分数据的定和限制给成分间带来的额外的负相关, Faust 等建议通过打乱数据等办法计算零假设下相关性的分布, 然后得到单个位置检验的 p 值, 最后使用 Benjamini-Hochberg-Yekutieli 方法[58] 控制 FDR. 不过, 成分数据的特殊属性使得数据在打乱后需要在同一样本内做一次归一化, 这一过程可能影响置换和重抽样方法理论上的有效性. 此外, 重抽样方法也给 p 值的计算和置信区间的构造增加了额外的计算负担.

为此, Wu 等[4)] 又发展了一个针对 $\mathbf{\Sigma}$ 的元素进行假设检验的方法. 对 $1 \leqslant j < k \leqslant p$, 我们希望同时检验以下 $p(p-1)/2$ 个假设

$$H_{0,jk} : \sigma_{jk} = 0 \quad \text{vs.} \quad H_{1,jk} : \sigma_{jk} \neq 0,$$

并对多重检验的 FDR 进行控制. 我们的方法分为两步, 首先是对每个假设构造检验统计量 T_{jk} 并推导其在零假设下的分布, 然后是发展一个能考虑到 T_{jk} 间相关性的控制 FDR 的方法, 其中的第二步同处理非成分数据的协方差矩阵多重假设检验方法[59] 是类似的. 在第一步中, 易知中心对数比变换后的样本协方差 $\hat{\gamma}_{jk}$ 是以真值 γ_{jk} 为均值的正态分布, 因此, 从关系式 (3.10) 出发, 如果有一个 $\mathbf{\Sigma}$ 的较好的估计 $\hat{\mathbf{\Sigma}}$, 满足 $\|\hat{\sigma}_{j\cdot} + \hat{\sigma}_{\cdot k} - \hat{\sigma}_{\cdot\cdot} - \sigma_{j\cdot} - \sigma_{\cdot k} + \sigma_{\cdot\cdot}\|_1 = o_p(1/\sqrt{n})$, 因 $O(1/\sqrt{n})$ 是 $\hat{\gamma}_{jk}$ 收敛到 γ_{jk} 的速度, 故将 $\hat{\gamma}_{jk}$ 用 (3.15) 的估计量 $\hat{\mathbf{\Sigma}}$ 去偏差后就能得到以 σ_{jk} 为均值的正态分布. 特别地, 令

$$T_{jk} = \frac{\hat{\gamma}_{jk} + \hat{\sigma}_{j\cdot} + \hat{\sigma}_{\cdot k} - \hat{\sigma}_{\cdot\cdot}}{\sqrt{n\hat{\theta}_{jk}}},$$

其中

$$\hat{\theta}_{jk} = \frac{1}{n}\sum_{i=1}^{n}\{(y_{ij} - \bar{y}_j)(y_{ik} - \bar{y}_k) - \hat{\gamma}_{jk}\}^2$$

是 $\theta_{jk} = \mathrm{Var}((y_{ij} - \nu_j)(y_{ik} - \nu_k))$ 的无偏估计, 这里 $\nu_j = \mathrm{E}(y_{ij})$, $\bar{y}_j = \frac{1}{n}\sum_{i=1}^{n} y_{ij}$, 则在一定稀疏性条件下可以证明, 当 $H_{0,jk}$ 成立时, 有 $T_{jk} \to \mathcal{N}(0,1)$. 在第二步中, 记 $\mathcal{H}_0 = \{(j,k) : \sigma_{jk} = 0\}$, 并令 t 为拒

绝假设的临界值, 即当 $|T_{jk}| > t$ 时拒绝零假设 $H_{0,jk}$. 我们的核心目标是确定临界值以保证错误发现比例 (false discovery proportion, FDP) 小于给定的水平, 其中

$$R(t) = \sum_{1\leqslant j<k\leqslant p} I(|T_{jk}| > t), \quad R_0(t) = \sum_{(j,k)\in\mathcal{H}_0} I(|T_{jk}| > t), \quad \mathrm{FDP}(t) = \frac{R_0(t)}{1 \vee R(t)}.$$

为了估计给定临界值时 $R_0(t)$ 的值, 我们利用协方差结构的稀疏性证明了统计量 T_{jk} 间的弱相依性, 从而有 $G(t)|\mathcal{H}_0|$ 在 $t < \sqrt{4\log p - 2\log(\log p)}$ 时是 $R_0(t)$ 的一个较好的近似, 其中 $G(t) = \mathrm{P}(|\mathcal{N}(0,1)| > t)$. 最后, 我们用 $p(p-1)/2$ 当作 $|\mathcal{H}_0|$ 的估计值, 并证明了我们的方法能够控制 FDP 和 FDR. 具体的多重检验过程, 也可参见文献 [59].

3.2.3 精度矩阵的估计

对于对数基底协方差矩阵 $\boldsymbol{\Sigma}$ 的估计和统计推断, 目前的研究已经比较深入. 但是, $\boldsymbol{\Sigma}$ 仅描述了两个菌种间相关性的总效应, 而研究人员往往对菌种间的直接相互作用更感兴趣. 在第 2.1 小节中, 我们曾经提到过精度矩阵 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ 与变量间直接相互作用的关系. 这里介绍两种在 $\boldsymbol{\Omega}$ 的稀疏假设下的估计方法.

Kurtz 等[60] 首先提出了基于中心对数比变换的方法估计 $\boldsymbol{\Omega}$, 其核心思想还是从 (3.10) 得出当 $\boldsymbol{\Sigma}$ 稀疏时, $\boldsymbol{\Gamma}$ 可以较好地近似 $\boldsymbol{\Sigma}$. 在此基础上, Kurtz 等将绝对量数据精度矩阵的两种估计方法引入到成分数据精度矩阵估计的问题中. 第一种是近邻选择法 (neighborhood selection)[61], 利用变量间的回归系数与偏相关系数的等价关系, 对每个变量 j 考虑如下 Lasso 形式的稀疏回归:

$$\min_{\boldsymbol{\beta}_{-j}\in\mathbb{R}^{p-1}} \left\{ \frac{1}{2n}\|\boldsymbol{y}^j - \boldsymbol{Y}^{-j}\boldsymbol{\beta}_{-j}\|_2^2 + \lambda_j\|\boldsymbol{\beta}_{-j}\|_1 \right\},$$

其中

$$\boldsymbol{y}^j = (y_{1j},\ldots,y_{nj})^{\mathrm{T}}, \quad \boldsymbol{Y}^{-j} = (\boldsymbol{y}^1,\ldots,\boldsymbol{y}^{j-1},\boldsymbol{y}^{j+1},\ldots,\boldsymbol{y}^p),$$

而 λ_j 为惩罚参数. 由于 β_{-j} 与 $\boldsymbol{\Omega}$ 的第 j 行存在着对应关系, 每次求解可以得到与变量 j 的 "近邻", 即与其偏相关不为 0 的变量. 对每个变量都考虑上述回归后, 可以估计出偏相关不为 0 的所有变量组合. 第二种是基于惩罚极大似然的图 Lasso 算法 (graphical Lasso)[62]. 特别地, 令 $\boldsymbol{\Omega}$ 的估计为如下优化问题的解:

$$\min_{\boldsymbol{\Omega}}\{-\log\det(\boldsymbol{\Omega}) + \mathrm{tr}(\boldsymbol{\Omega}\hat{\boldsymbol{\Gamma}}) + \lambda\|\boldsymbol{\Omega}\|_1\},$$

其中前两项是 Gauss 分布的负对数似然, 后面是对 $\boldsymbol{\Omega}$ 的稀疏性惩罚项. 这两种方法在绝对数据的精度矩阵估计问题中都研究得较为透彻, 包括理论性质和优化方法, 以及在其他方面的拓展等. 但是, 它们在成分数据中的直接应用还存在着理论上的问题: $\boldsymbol{\Sigma}$ 和 $\boldsymbol{\Gamma}$ 只有在 $\sigma_{j\cdot} \approx 0$ $(j = 1,\ldots,p)$ 时才近似相等, 但加在 $\boldsymbol{\Omega} = \boldsymbol{\Sigma}^{-1}$ 上的稀疏性假设是否与此矛盾, 还没有研究对此进行具体的讨论. 此外, $\boldsymbol{\Omega}$ 的可识别性依然是个困难. 因为所有矩阵 $\boldsymbol{\Omega} \in \mathcal{E}(\boldsymbol{\Sigma})$ 都可诱导相同的中心对数比协方差矩阵 $\boldsymbol{\Gamma}$, 是否存在 $\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2 \in \mathcal{E}(\boldsymbol{\Sigma})$ 同时满足 $\boldsymbol{\Sigma}_1^{-1}$ 和 $\boldsymbol{\Sigma}_2^{-1}$ 都是稀疏矩阵, 仍然是一个难以解决的问题.

此外, Fang 等[63] 提出了 gCoda (conditional dependence network inference for compositional data) 算法, 没有通过中心对数比变换, 而是在假设对数绝对量 $\boldsymbol{u}_i$ 服从正态分布的条件下, 直接写出了观测数据的似然函数, 并加入 ℓ_1 范数惩罚项估计稀疏矩阵 $\boldsymbol{\Omega}$. 由于只有 $\boldsymbol{u}_i$ 的相对信息被观测到, 负对数似然函数在经过积分后可表为

$$-l(\boldsymbol{\Omega}) = -\log\det\boldsymbol{\Omega} + \log(\mathbf{1}_p\boldsymbol{\Omega}\mathbf{1}_p^{\mathrm{T}}) + \mathrm{tr}\left(\hat{\boldsymbol{\Lambda}}\left(\frac{\boldsymbol{\Omega}\mathbf{1}_p\mathbf{1}_p^{\mathrm{T}}\boldsymbol{\Omega}}{\mathbf{1}_p\boldsymbol{\Omega}\mathbf{1}_p}\right)\right),$$

其中 $\hat{\mathbf{\Lambda}}$ 是关于 $\log x_{ij}$ 的样本协方差阵. 由于上式不是关于 $\mathbf{\Omega}$ 的凸函数, Fang 等提出了基于主函数极小化 (majorization minimization) 的方法求解带惩罚的优化问题.

以上这两种方法均从计算的角度给出了估计微生物菌种间直接相互作用的一种方法. 但是, 如前文所述, 对数基底精度矩阵的可识别性问题仍是尚未解决的核心问题. 在没有可识别性条件时, 估计的目标尚不明确, 估计量的相合性等性质更无从谈起, 相应的统计推断 (如假设检验和构造置信区间) 也难以得到有保证的结果. 对这方面的进一步研究可能是未来微生物组学统计分析中的一个重要方向.

4 讨论与展望

高通量测序技术的发展及人们对微生物了解的深入使得统计学和计算生物学工具在微生物组学分析中起到了重要的作用. 本文回顾了近年来微生物组学中高维计数数据和成分数据分析的统计方法, 包括 DM 模型及其拓展, 从计数数据到成分数据的估计、成分回归模型以及基于对数基底的成分数据分析等内容. 这些模型及其背后的假设是依据微生物学测序数据的特点而发展的, 包括计数属性和零计数的存在、成分数据属性、高维属性及相应的系统发生树结构.

关于微生物组学数据分析, 以下问题可能是未来研究中的重点. (1) 进一步发展针对高维计数数据的统计模型和理论, 尤其是具有成分属性的微生物组计数数据. 计数数据的特殊结构使得近年来蓬勃发展的针对连续数据的高维统计方法不能直接应用. 发展新的模型、提出可扩展的算法、建立高维下的统计理论都是未来值得关注的问题. (2) 处理计数中过多的零计数. 本文提到的方法大多认为零计数是菌种含量太低及抽样的样本不够大造成的, 这为后续分析提供了便利. 事实上, 零计数可能既包含因抽样产生的零, 也包含真实的零 (essential zeros). 允许真零的存在会导致模型复杂度指数增加, 已有的方法可能不适用于高维情形[64], 或者未考虑到成分的特点[65]. 如何利用其他信息鉴别计数数据中的真零和抽样的零, 或发展类似于 Lasso 变量选择的方法自动找到零计数中的真零是一个值得研究的课题. (3) 进一步利用系统发生树的结构和信息. 例如, 对于两样本菌种成分的检验问题, 因微生物间存在密切的相互作用, 故同时在不同等级的分类单元上进行检验可能得到更有意义的结果. 本文提到的计数数据建模中考虑到了树结构, 但只进行了全局的假设检验, 不能给出到底哪个等级的哪种微生物在两总体中成分是显著不同的. 因此, 考虑在成分数据两样本检验问题中加入树的信息后同时进行多个层级的假设检验是很有意义的问题. (4) 基于对数基底的成分数据分析中可识别性的研究. 成分数据由基底生成, 没有观测到绝对总量导致的可识别性困难在哪些问题中可以解决, 哪些问题中难以逾越是人们所感兴趣的. 例如, 关于对数基底精度矩阵的估计和统计推断, 可能是这一方向中有望解决的课题. (5) 协同分析环境协变量、微生物组分及宿主的表现型数据. 饮食摄入数据可能同时影响微生物组分及人体的肥胖指数, 而微生物组分也会直接影响人体代谢并导致体重变化, 使用中介分析 (mediation analysis) 或其他方法对相应数据进行因果推断也是未来研究中一个可能的方向.

致谢 感谢林伟教授和张静茹同学提供了部分未发表论文的底稿.

参考文献

1 Turnbaugh P J, Ley R E, Hamady M, et al. The human microbiome project: Exploring the microbial part of ourselves in a changing world. Nature, 2007, 449: 804–810

2 Turnbaugh P J, Hamady M, Yatsunenko T, et al. A core gut microbiome in obese and lean twins. Nature, 2009, 457: 480–484

3 Qin J, Li Y, Cai Z, et al. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature, 2012, 490: 55–60
4 Koeth R A, Wang Z, Levison B S, et al. Intestinal microbiota metabolism of L-carnitine, a nutrient in red meat, promotes atherosclerosis. Nat Med, 2013, 19: 576–585
5 Wu G D, Chen J, Hoffmann C, et al. Linking long-term dietary patterns with gut microbial enterotypes. Science, 2011, 334: 105–108
6 Morgan J L, Darling A E, Eisen J A. Metagenomic sequencing of an in vitro-simulated microbial community. PLoS ONE, 2010, 5: e10209
7 Thomas T, Gilbert J, Meyer F. Metagenomics—A guide from sampling to data analysis. Micro Inform Exp, 2012, 2: 1–12
8 Li H. Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu Rev Stat Appl, 2015, 2: 73–94
9 Li C, Lim K M K, Chng K R, et al. Predicting microbial interactions through computational approaches. Methods, 2016, 102: 12–19
10 Layeghifard M, Hwang D M, Guttman D S. Disentangling interactions in the microbiome: A network perspective. Trends Microbiol, 2016, 25: 217–228
11 Mosimann J E. On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. Biometrika, 1962, 49: 65–82
12 La Rosa P S, Brooks J P, Deych E, et al. Hypothesis testing and power calculations for taxonomic-based human microbiome data. PLoS ONE, 2012, 7: e52078
13 Van der Vaart A W. Asymptotic Statistics. Cambridge: Cambridge University Press, 1998
14 Efron B. The Jackknife, the Bootstrap and Other Resampling Plans. Philadelphia: SIAM, 1982
15 Good P. Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses. New York: Springer, 2013
16 Chen J, Li H. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. Ann Appl Stat, 2013, 7: 418–442
17 Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. J R Stat Soc Ser B Stat Methodol, 2006, 68: 49–67
18 Tibshirani R. Regression shrinkage and selection via the Lasso. J R Stat Soc Ser B Stat Methodol, 1996, 58: 267–288
19 Wardsworth W D, Argiento R, Guindani M, et al. An integrative Bayesian Dirichlet-multinomial regression model for the analysis of taxonomic abundances in microbiome data. BMC Bioinform, 2017, 18: 94–195
20 Yang Y, Chen N, Chen T. Inference of environmental factor-microbe and microbe-microbe associations from metagenomic data using a hierarchical Bayesian statistical model. Cell Syst, 2017, 4: 129–137
21 Biswas S, McDonald M, Lundberg D S, et al. Learning microbial interaction networks from metagenomic count data. J Comput Biol, 2016, 23: 526–535
22 Mandal S, Van Treuren W, White R A, et al. Analysis of composition of microbiomes: A novel method for studying microbial composition. Microb Ecol Health Dis, 2015, 26: 27663
23 Holmes I, Harris K, Quince C. Dirichlet multinomial mixtures: Generative models for microbial metagenomics. PLoS ONE, 2012, 7: e30126
24 Xia F, Chen J, Fung W K, et al. A logistic normal multinomial regression model for microbiome compositional data analysis. Biometrics, 2013, 69: 1053–1063
25 Aitchison J. The Statistical analysis of compositional data. J R Stat Soc Ser B Stat Methodol, 1982, 44: 139–177
26 Billheimer D, Guttorp P, Fagan W F. Statistical interpretation of species composition. J Amer Statist Assoc, 2001, 96: 1205–1214
27 Wang T, Zhao H. A Dirichlet-tree multinomial regression model for associating dietary nutrients with gut microorganisms. Biometrics, 2017, 73: 792–801
28 Dennis S Y. A Bayesian analysis of tree-structured statistical decision problems. J Statist Plann Inference, 1996, 53: 323–344
29 Tang Y, Li M, Dan L N. A phylogenetic scan test on Dirichlet-tree multinomial model for microbiome data. ArXiv:1610.08974, 2016
30 Shi P, Li H. A model for paired-multinomial data and its application to analysis of data on a taxonomic tree. Biometrics, 2017, doi: 10.1111/biom.12681
31 Soriano J, Ma L. Probabilistic multi-resolution scanning for two-sample differences. J R Stat Soc Ser B Stat Methodol, 2017, 79: 547–572

32 Tang Z Z, Chen G, Alekseyenko A V, et al. A general framework for association analysis of microbial communities on a taxonomic tree. Bioinformatics, 2017, 33: 1278–1285
33 Aitchison J. The Statistical Analysis of Compositional Data. Caldwell: Blackburn Press, 2003
34 Martín-Fernández J A, Hron K, Templ M, et al. Bayesian-multiplicative treatment of count zeros in compositional data sets. Stat Model, 2015, 15: 134–158
35 Faust K, Sathirapongsasuti J F, Izard J, et al. Microbial co-occurrence relationships in the human microbiome. PLoS Comput Biol, 2012, 8: e1002606
36 Negahban S, Wainwright M J. Estimation of (near) low-rank matrices with noise and high-dimensional scaling. Ann Statist, 2011, 39: 1069–1097
37 Cao Y, Zhang A, Li H. Composition estimation from sparse count data via a regularized likelihood. ArXiv:1706.02380, 2017
38 Lee D D, Seung H S. Learning the parts of objects by non-negative matrix factorization. Nature, 1999, 401: 788–791
39 Aitchison J, Bacon-Shone J. Log contrast models for experiments with mixtures. Biometrika, 1984, 71: 323–330
40 Lin W, Shi P, Feng R, et al. Variable selection in regression with compositional covariates. Biometrika, 2014, 101: 785–797
41 Zhao P, Yu B. On model selection consistency of Lasso. J Mach Learn Res, 2006, 7: 2541–2563
42 Shi P, Zhang A, Li H. Regression analysis for microbiome compositional data. Ann Appl Stat, 2016, 10: 1019–1040
43 Javanmard A, Montanari A. Confidence intervals and hypothesis testing for high-dimensional regression. J Mach Learn Res, 2014, 15: 2869–2909
44 Wang T, Zhao H. Structured subcomposition selection in regression and its application to microbiome data analysis. Ann Appl Stat, 2017, 11: 771–791
45 Wang T, Zhao H. Constructing predictive microbial signatures at multiple taxonomic levels. J Amer Statist Assoc, 2017, 112: 1022–1031
46 Cao Y. Statistical methods for high dimensional count and compositional data with applications to microbiome studies. PhD Thesis. Philadelphia: University of Pennsylvania, 2016
47 Hu J, Bai Z. A review of 20 years of naive tests of significance for high-dimensional mean vectors and covariance matrices. Sci China Math, 2016, 59: 2281–2300
48 Cai T, Liu W, Xia Y. Two-sample test of high dimensional means under dependence. J R Stat Soc Ser B Stat Methodol, 2014, 76: 349–372
49 Chen S, Qin Y. A two-sample test for high-dimensional data with applications to gene-set testing. Ann Statist, 2010, 38: 808–835
50 Xu G, Lin L, Wei P, et al. An adaptive two-sample test for high-dimensional means. Biometrika, 2016, 103: 609–624
51 Pearson K. Mathematical contributions to the theory of evolution—On a form of spurious correlation which may arise when indices are used in the measurement of organs. Proc R Soc Lond, 1896, 60: 489–498
52 Friedman J, Alm E J. Inferring correlation networks from genomic survey data. PLoS Comput Biol, 2012, 8: e1002687
53 Fang H, Huang C, Zhao H, et al. CCLasso: Correlation inference for compositional data through Lasso. Bioinformatics, 2015, 31: 3172–3180
54 Ban Y, An L, Jiang H. Investigating microbial co-occurrence patterns based on metagenomic compositional data. Bioinformatics, 2015, 31: 3322–3329
55 Cao Y, Lin W, Li H. Large covariance estimation for compositional data via composition-adjusted thresholding. ArXiv:1601.04397, 2016
56 Cai T, Liu W. Adaptive thresholding for sparse covariance matrix estimation. J Amer Statist Assoc, 2011, 106: 672–684
57 Cai T, Ren Z, Zhou H. Estimating structured high-dimensional covariance and precision matrices: Optimal rates and adaptive estimation. Electron J Stat, 2016, 10: 1–59
58 Benjamini Y, Yekutieli D. The control of the false discovery rate in multiple testing under dependency. Ann Statist, 2001, 29: 1165–1188
59 Cai T, Liu W. Large-scale multiple testing of correlations. J Amer Statist Assoc, 2016, 111: 229–240
60 Kurtz Z D, Müller C L, Miraldi E R, et al. Sparse and compositionally robust inference of microbial ecological networks. PLoS Comput Biol, 2015, 11: e1004226
61 Meinshausen N, Bühlmann P. High-dimensional graphs and variable selection with the Lasso. Ann Statist, 2006, 34: 1436–1462
62 Friedman J, Hastie T, Tibshirani R. Sparse inverse covariance estimation with the graphical lasso. Biostatistics, 2008, 9: 432–441

63 Fang H, Huang C, Zhao H, et al. gCoda: Conditional dependence network inference for compositional data. J Comput Biol, 2017, 24: 699–708
64 Bear J, Billheimer D. A logistic normal mixture model for compositional data allowing essential zeros. Aust J Stat, 2016, 45: 3–23
65 Kaul A, Davidov O, Peddada S D. Structural zeros in high-dimensional data with applications to microbiome studies. Biostatistics, 2017, 18: 422–433

High-dimensional count and compositional data analysis in microbiome studies

WU ChangJing, HE Shun & DENG MingHua

Abstract The human microbiome plays an important role in human health and disease. The development of high-throughput sequencing technologies makes it possible to quantify all microbes constituting the microbiome. In this paper, we give a review of recent advances in high-dimensional count and compositional data analysis in microbiome studies. It includes the Dirichlet-multinomial model and its extensions, composition estimation from large sparse count matrix, high-dimensional regression with compositional covariates, and statistical inference for log-basis-based compositional data.

Keywords **high-dimensional count, compositional data, Dirichlet-multinomial model, sparsity, identifiability, regression model**

MSC(2010) **62-02, 62F15, 62H12, 62H15, 62J07, 62P10**

doi: 10.1360/N012017-00147