arXiv:2303.10221v1 [stat.ME] 17 Mar 2023

利用汇总统计量对高维表型进行全基因组关联研究的统计框架及其在代谢物全基因组关联研究中的应用

Weiqiong Huang[1], Emily C.Hector[2], Joshua Cape[3], Chris McKennan[1],
*[1]匹兹堡大学统计系[2]北卡罗来纳州立大学统计系[3]威斯康星大学统计系

2023年3月21日

摘要

近期遗传、高维生物样本库及组学数据的迅猛增长，为研究人员提供了机会，使其能够探究数百乃至数千种相关表型的共同遗传起源（多效性）。然而，现有的多表型全基因组关联研究方法无法对多效性建模，仅适用于少量表型，或无法进行推断。更复杂的是，原始遗传数据和表型数据鲜有观测到，这意味着分析必须基于全基因组关联研究汇总统计量开展，而这些统计量在高维情况下的统计性质尚未得到充分理解。因此，我们开发了一种新模型、理论框架和一套方法，可利用汇总统计量对高维表型开展贝叶斯推断；这些方法能够明确地对多效性建模、实现快速计算，并便于使用生物学信息先验。我们通过将该方法应用于代谢物全基因组关联研究，展示了其效用；在此过程中，我们针对代谢物水平的遗传效应开发了新的非参数先验分布，这些先验利用已知的代谢通路信息，并有助于在通路层面进行可解释的推断。

关键词：高维因子模型；贝叶斯非参数方法；全基因组关联研究；汇总统计量；多效性

1 引言

全基因组关联研究（GWAS）旨在探究单核苷酸多态性（SNP）位点的基因型与表型之间的关系，已成为阐明人类变异遗传基础的重要工具（Uffelmann 等，2021）。近年来，高维生物样本库数据和组学数据激增，使生物学家得以探索数百乃至数千种相关表型的共同遗传起源（多效性）（Tanigawa 等，2019；Elliott 等，2018；Yin 等，2022a；Kurilshikov 等，2021）。其中尤其值得关注的是针对代谢物的全基因组关联研究：代谢物是细胞代谢的副产物或终产物，也是小分子。代谢物对于建立基因型与疾病之间的功能联系至关重要（Yin 等，2022b）；因此，代谢物全基因组关联研究（mtGWAS）有望改变我们对疾病病因学的认识，并发现可用于靶向治疗的新型生物标志物（Karjalainen 等，2022）。

*如需联系，请 联系地址为（chm195@pitt .edu）。

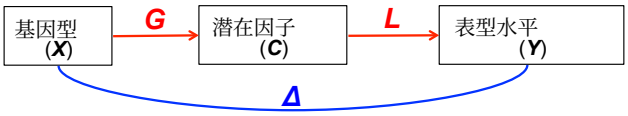

图1：基因型对表型水平影响的图模型。红色：间接效应；蓝色：直接效应。箭头上方的参数是感兴趣参数。

mtGWAS 中许多尚未解决的统计学和方法学挑战，凸显了开展高维表型全基因组关联研究的难点。首先，代谢物水平存在广泛的遗传相关和非遗传相关（McKennan 等，2020；Yin 等，2022a）。然而，现有纳入表型间相关性的多表型全基因组关联研究方法，只能处理少量表型（Qi 等，2018；Dai 等，2019；Liu 等，2017；Lu 等，2022），只能检验某个单核苷酸多态性是否与至少一种表型相关（C. Wu，2020；M. Wang 等，2023），或无法提供推断保证（Tanigawa 等，2019；Petti 等，2022）。因此，几乎所有 mtGWAS 研究都忽略多效性，并进行成对边际回归，同时采用严格的 Bonferroni 校正，以校正 #SNPs $\times$ #代谢物 $\gtrsim 10^8$项检验。其次，出于隐私顾虑，原始代谢物数据和遗传数据很少能够获取。研究人员通常只能取得代谢物水平对 SNP基因型进行边际回归后得到的 #SNPs $\times$ #代谢物汇总统计量。尽管已有方法将单核苷酸多态性视为类似生物样本的对象，并运用标准因子分析工具分析汇总统计量矩阵（Tanigawa 等，2019；Petti 等，2022），但我们表明，这些方法会导致估计和推断出现错误，因此需要新的方法和理论结果。最后，应充分利用有关代谢物归入不同生物通路的丰富生物学信息，以提高统计功效和可解释性。这一点至关重要，因为解读涉及数千种代谢物的 mtGWAS 结果极具挑战，而解读功能通路则容易得多。

为应对这些关键障碍，我们针对高维表型的全基因组关联研究开发了一种新的统计模型、理论框架和方法。图1以图示方式展示了我们的模型，该模型将总遗传效应分解为由一组潜在因子介导的间接效应和直接效应。因子代表可能由遗传决定的生物过程，例如代谢组学中的葡萄糖代谢或微生物组学中的肠道结构（Hall 等，2017）；这些过程调节许多表型的水平，因此有助于对多效性和非遗传相关进行建模。直接效应则捕捉剩余的特异性遗传变异。

我们的理论框架至关重要，因为我们的估计依赖于使用全基因组关联研究汇总统计量进行因子分析，而据我们所知，其统计性质目前尚不清楚。因此，据我们所知，我们首次为基于汇总统计量的因子分析提供了一套理论保证。这些成果不仅仅是技术上的进展，因为我们表明，SNP基因型之间看似微不足道的依赖关系，会使估计量的渐近分布与基于原始数据推导出的渐近分布大相径庭。

最后，我们开发了两种新方法，用于估计图1中的间接遗传效应和直接遗传效应，以应对上述方法学挑战。第一种方法用于估计潜在因子的数量，我们称之为 dBEMA（依赖型总体特征值匹配）；该方法考虑了 SNP 基因型之间未观测到的依赖关系。这些依赖关系会导致原本为原始数据设计的现有方法在应用于全基因组关联研究汇总统计量时，严重误估因子数量。第二种方法是一种通用的贝叶斯因子分析方法，称为 HiGSS（基于汇总统计量的高维全基因组关联研究），可对模型参数进行理论依据充分的贝叶斯推断。不同于计算开销大的传统后验抽样过程和次优的变分

近似方法，HiGSS 利用上述理论结果，通过经验贝叶斯高效确定后验分布，同时不牺牲统计保真度。至关重要的是，HiGSS 支持使用融入生物学信息的先验分布；在传统贝叶斯因子分析流程中，这类先验在计算上难以处理，例如微生物组和脑影像数据中的系统发育先验和空间先验（L. Zhang 等，2021年；Sidén 等，2017年）。虽然我们设计 HiGSS 时使其兼容通用先验分布，但我们投入了大量精力，为代谢物全基因组关联研究中的直接效应和间接效应构建恰当的非参数先验。这些先验利用已知的代谢通路信息，促进通路层面上高效且易于解释的推断。

本文其余部分安排如下。我们在第 2 节介绍汇总统计量模型。第3节介绍所提出的估计算法及其理论保证。在第4节中，我们为代谢物全基因组关联研究中的间接效应和直接效应构建非参数先验。第 5 节考察所提出估计方法的有限样本性能。我们在第6节开展一项代谢物全基因组关联研究，以展示该方法的应用。第7节总结全文。

2 数据模型

2.1 问题符号

对于 $\boldsymbol{Z} \in \mathbb{R}^{n\times m}$，我们令 $\boldsymbol{Z}_{i*} \in \mathbb{R}^m$、 $\boldsymbol{Z}_{*j} \in \mathbb{R}^n$ 和 $\boldsymbol{Z}_{ij} \in \mathbb{R}$ 分别表示 i 行、 j 列以及 $\boldsymbol{Z}$ 的第（i，j）个元素。我们使用标准的矩阵正态分布记号，即对于 $\boldsymbol{M} \in \mathbb{R}^{n\times m}$、 $\boldsymbol{U} \in \mathbb{R}^{n\times n}$ 和$\boldsymbol{V} \in \mathbb{R}^{m\times m}$， $\boldsymbol{Z} \sim MN(\boldsymbol{M}, \boldsymbol{U}, \boldsymbol{V})$，若 $\text{vec}(\boldsymbol{Z}) \sim N(\text{vec}(\boldsymbol{M})\ \boldsymbol{V} \otimes \boldsymbol{U})$。向量 $\boldsymbol{x} \sim (\boldsymbol{\mu}, \boldsymbol{A})$，若 $\mathbb{E}(\boldsymbol{x}) = \boldsymbol{\mu}$且 $\text{Var}(\boldsymbol{x}) = \boldsymbol{A}$。

2.2 未观测基因型和表型数据的模型

设 $\boldsymbol{Y} \in \mathbb{R}^{N\times M}$ 表示在 N 名无亲缘关系的个体中测得的 M 种表型（代谢物）的未观测水平，$\boldsymbol{X} \in \mathbb{R}^{N\times S}$ 表示这些个体在 S 个单核苷酸多态性位点上的未观测基因型。假设

假设 2。1 $\boldsymbol{X}$ 是一个随机矩阵，其元素有界、均值为零且相互独立，行具有相同的分布，并且对于某个常数 $\epsilon > 0$ 以及所有 $i \in \{1, \ldots, n\} = [n]$和 $s \in [S]$，$\text{Var}(\boldsymbol{X}_{is}) \geq \epsilon$。

均值为零和独立性假设等价于对 $\boldsymbol{X}$ 的各列进行中心化，并对存在连锁不平衡的单核苷酸多态性位点进行剪枝；剪枝有助于避免推断出冗余的遗传效应。图1以图示方式描述了我们的基因型对表型影响的模型，该模型将总效应分解为直接效应和由 K 个潜在因子 $\boldsymbol{C} \in \mathbb{R}^{N\times K}$介导的间接效应，等价于以下模型：

$$\begin{aligned} \boldsymbol{Y}_{N\times M} =& \boldsymbol{C}_{N\times K}\boldsymbol{L}^\top_{M\times K} + \boldsymbol{X}_{N\times S}\boldsymbol{\Delta}_{S\times M} + \boldsymbol{E}_{N\times M}, \quad \boldsymbol{E} \sim MN(0, I_N, \text{diag}(\sigma_1^2, \ldots, \sigma_M^2)) \\ \boldsymbol{C}_{N\times K} =& \boldsymbol{X}_{N\times S}\boldsymbol{G}_{S\times K} + \boldsymbol{\Xi}_{N\times K}, \quad \boldsymbol{\Xi}_{1*}, \ldots, \boldsymbol{\Xi}_{N*} \overset{\text{i.i.d}}{\sim} (0, \boldsymbol{\Psi}_{K\times K}), \end{aligned} \tag{2.1}$$

其中， $\boldsymbol{L}_{mk}$ 是第 k个潜在因子对第 m个表型的效应，而 $\boldsymbol{G}_{sk}$ 是第s个单核苷酸多态性对第 k个因子的效应。在该模型下， $\boldsymbol{G}$、 $\boldsymbol{L}$和 $\boldsymbol{\Delta}$ 是感兴趣参数，$\mathbb{E}$（$\boldsymbol{Y} \mid \boldsymbol{X}$、$\boldsymbol{G}$、$\boldsymbol{L}$、$\boldsymbol{\Delta}$）$= \boldsymbol{X}$（$\boldsymbol{G}\boldsymbol{L}^\top + \boldsymbol{\Delta}$）是总遗传效应，其中 $\boldsymbol{G}^\top_{s*}\boldsymbol{L}_{m*}$和 $\boldsymbol{\Delta}_{sm}$ 分别是第 s个单核苷酸多态性对第 m个表型的间接效应和直接效应。表型间的非遗传协方差 $\text{Var}(\boldsymbol{Y}_{i*} \mid \boldsymbol{X}$、$\boldsymbol{G}$、$\boldsymbol{L}$、$\boldsymbol{\Delta}) = \boldsymbol{L}\boldsymbol{\Psi}\boldsymbol{L}^\top + \text{diag}(\sigma_1^2$，……，$\sigma_M^2)$具有低秩加对角结构，这是代谢物、微生物组及其他多表型数据中的一项标准假设（Sankaran 等，2018；McKennan 等，2020；Tanigawa 等，2019）。由于我们提出了一种贝叶斯方法来推断

$\boldsymbol{G}$、 $\boldsymbol{L}$和 $\boldsymbol{\Delta}$，为避免混淆，我们在上述和下述概率表达式中明确以它们为条件。

式（2.1）的两个特点有助于提出关于 $\boldsymbol{G}$ 和 $\boldsymbol{\Delta}$的假设。首先，总遗传系数 $\boldsymbol{GL}^\top + \boldsymbol{\Delta}$ 中的第一项刻画了表型间和 SNP 间效应共享的模式，这表明 $\boldsymbol{\Delta}$ 应反映特异性遗传变异。其次， $\boldsymbol{G}$ 和 $\boldsymbol{\Delta}$ 都表示基因型对一组表型的直接效应，因此预计它们是稀疏的，且非零元素较小（Y. Zhang 等，2018）。因此，我们始终作如下假设：

假设 2.2 对于某个正整数 A 和常数 $c, r \geq 0$

(a) $\boldsymbol{\Delta}$是一个具有独立元素的随机矩阵，这些元素满足 $\mathbb{P}(\boldsymbol{\Delta}_{sm} \neq 0) = \pi_{sm}$和 $\boldsymbol{\Delta}_{sm} \,|(\boldsymbol{\Delta}_{sm} \neq 0) \sim \sum_{a=1}^{A} \gamma_{sma} N(0, \tau_{sma}^2)$，其中 $\pi_{sm}, \gamma_{sma} \in [0,1]$和 $\sum_{a=1}^{A} \gamma_{sma} = 1$。此外，$\max^{s\in[S];m\in[M];a\in[A]}(N^{1/2}\tau_{sma}) \leq c$且$\max^{s\in[S];m\in[M]} \pi_{sm} = o(1)$当 $S, M \to \infty$时成立。

(b) $\max_{s\in[S]} \|N^{1/2}\boldsymbol{G}_{s*}\|_2 = O_P(\log^r N)$ *and* $S^{-1}\sum_{s=1}^{S} \|N^{1/2}\boldsymbol{G}_{s*}\|_2^2 = o_P(1)$ *as* $S, N \to \infty$.

$\boldsymbol{\Delta}$的逐元素独立性意味着它能够捕捉特异性遗传变异；再加上其各元素服从对称的尖峰与板块先验分布，这些都是全基因组关联研究中针对单个或多个表型的遗传效应所采用的经典假设（Zhou 等人，2013；Zhu 等人，2017；Ruffieux 等人，2017；Y. Zhang 等，2018；Turley 等人，2018；Majumdar 等人，2018）。混合正态 slab 具有相当广泛的适用性，可以近似一大类对称分布（Stephens，2017）。参数 τ_{sma} 决定了 $\boldsymbol{\Delta}$非零效应的大小，其中对 $N^{1/2}\tau_{sma}$和 $N^{1/2}\boldsymbol{G}_{s*}$ 的条件反映了这样一种观察：在真实数据中，全基因组关联研究 z 分数通常较小至中等，其大小随 $N^{1/2}$ 乘以单核苷酸多态性的效应大小而变化（Zhu 等人，2017；Y. Zhang 等，2018）。我们在 (a) 中对 π_{sm} 施加的条件以及 (b) 中的平方和约束意味着稀疏性。需要注意的是，虽然我们假设 $\boldsymbol{\Delta}$ 是随机的，但第 3.4 节中所有频率学派估计量的弱收敛结果都是以 $\boldsymbol{\Delta}$为条件给出的。假设 $\boldsymbol{\Delta}$ 是随机的，可以理解为只考虑遗传数据中可能出现的 $\boldsymbol{\Delta}$，而忽略不现实的"边界情形"。关于 (2.1) 中其余参数的假设，我们将在第 3.4 节讨论。

2.3 观测到的全基因组关联研究汇总统计量

我们假设用户观测到了将第 m 个表型对第 s 个单核苷酸多态性的基因型进行回归所得的全基因组关联研究汇总统计量 $\hat{}\,\beta_{sm} = (\boldsymbol{X}_{*s}^\top \boldsymbol{X}_{*s})^{-1}\boldsymbol{X}_{*s}^\top \boldsymbol{Y}_{*m}$。我们还假设 $\boldsymbol{X}_{*s}^\top \boldsymbol{X}_{*s}$ 已知，但若无法获得，也可以根据公共数据估算（Auton 等，2015）。我们将标准化全基因组关联研究统计量的可观测矩阵 $\hat{}\boldsymbol{B} \in \mathbb{R}^{S\times M}$ 定义为 $\hat{}\,\boldsymbol{B}_{sm} = (\boldsymbol{X}_{*s}^\top \boldsymbol{X}_{*s})^{1/2}\hat{}\beta_{sm}$；对于 $\boldsymbol{D} = \text{diag}\{(\boldsymbol{X}_{*1}^\top \boldsymbol{X}_{*1}) \ldots, (\boldsymbol{X}_{*S}^\top \boldsymbol{X}_{*S})\}$ 和 $\boldsymbol{R} = \boldsymbol{D}^{-1/2}\boldsymbol{X}^\top \boldsymbol{X}\boldsymbol{D}^{-1/2}$（单核苷酸多态性之间的样本内相关性），该矩阵可表示为

$$\begin{aligned} \hat{\boldsymbol{B}} &= \boldsymbol{D}^{-1/2}\boldsymbol{X}^\top \boldsymbol{Y} = \tilde{\boldsymbol{G}}\boldsymbol{L}^\top + \boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{\Delta}) + \tilde{\boldsymbol{E}} \\ \tilde{\boldsymbol{G}} &= \boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{G}) + \boldsymbol{D}^{-1/2}\boldsymbol{X}^\top \boldsymbol{\Xi}, \quad \tilde{\boldsymbol{E}} \mid \boldsymbol{X} \sim MN(0, \boldsymbol{R}, \text{diag}(\sigma_1^2, \ldots, \sigma_M^2)). \end{aligned} \tag{2.2}$$

将汇总统计量除以 $(\boldsymbol{X}_{*s}^\top \boldsymbol{X}_{*s})^{1/2}$ 进行标准化，类似于使用 z 分数，并由此得到协方差相同的行（单核苷酸多态性），其中 $\boldsymbol{D}^{1/2}\boldsymbol{G}$ 和 $\boldsymbol{D}^{1/2}\boldsymbol{\Delta}$ 是 $\boldsymbol{G}$和 $\boldsymbol{\Delta}$的标准化对应量。虽然 $\boldsymbol{D}$ 的对角元素会随着样本量增大，但假设 2.2 意味着$\boldsymbol{D}^{1/2}\boldsymbol{G}$ 和 $\boldsymbol{D}^{1/2}\boldsymbol{\Delta}$ 都适中。显然，$\hat{}\boldsymbol{B}$ 通过 $\boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{G})$和 $\boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{\Delta})$依赖于 $\boldsymbol{D}^{1/2}\boldsymbol{G}$ 和 $\boldsymbol{D}^{1/2}\boldsymbol{\Delta}$。然而，由于在假设2.1下 $\boldsymbol{R}$ 的非对角元素较小，假设 2.2 中的稀疏性假设意味着 $\boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{G}) \approx \boldsymbol{D}^{1/2}\boldsymbol{G}$ 以及 $\boldsymbol{R}(\boldsymbol{D}^{1/2}\boldsymbol{\Delta}) \approx \boldsymbol{D}^{1/2}\boldsymbol{\Delta}$。

2.4 参数可识别性

式 (2.2) 中的感兴趣参数 $(\boldsymbol{G},\ \boldsymbol{L})$ 不可识别，因为对于任意可逆的 $K \times K$ 矩阵 $\boldsymbol{A}$，$(\boldsymbol{GA},\ \boldsymbol{LA}^{-\top})$ 都会得到相同的似然。我们通过以下命题来处理这一问题。

命题 2.1 若第 *3.4* 节中的假设 *2.1*、*2.2* 和 *3.1* 成立，则对于所有足够大的 N、M、S:

(i) 存在一个$(\boldsymbol{G}$、$\boldsymbol{L})$，使得 $\mathbb{E}(S^{-1}\tilde{\boldsymbol{G}}^{\top}\tilde{\boldsymbol{G}} \mid \boldsymbol{G}) = I_K$、$M^{-1}\boldsymbol{L}^{\top}\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots\ldots, \lambda_K)$ 以及$\lambda_1 > \cdots > \lambda_K > 0$。此外，$\boldsymbol{G}$和 $\boldsymbol{L}$ 的列在符号意义下是唯一的。*(ii)* 若 $c^{-1}I_K \preceq E(S^{-1}\tilde{\boldsymbol{G}}^{\top}\tilde{\boldsymbol{G}} \mid \boldsymbol{G}) \preceq cI_K$，其中某个常数为 $c > 1$，则对于所有 $\epsilon \in (0,\ 1)$，存在~一个 $\delta = \delta(\epsilon) > 0$，使得 $\mathbb{P}(|\boldsymbol{G}_{sk}| \geq \delta \mid \boldsymbol{G}) \geq 1 - \epsilon$ 对所有 $s \in [S]$、$k \in [K]$均成立。

在手稿其余部分中，我们采用 (i) 中的参数化方式，这是识别因子模型成分的一种经典方法（McKennan 等，2019）。~$\boldsymbol{G}$ 各列的正交性也为下文的算法 1 提供了依据，该算法使用奇异值分解来估计 $\boldsymbol{G}$、$\boldsymbol{L}$和 $\boldsymbol{\Delta}$。命题 2.1 的第 (ii) 部分表明，无论采用哪种参数化方式，~$\boldsymbol{G}$ 都不会是稀疏的，即使 $\boldsymbol{G}$ 是稀疏的。这是因为 (2.2) 中的 $\boldsymbol{G}$ 是 $\boldsymbol{G}$的含噪版本，而其误差 $\boldsymbol{D}^{-1/2}\boldsymbol{X}^{\top}\boldsymbol{\Xi}$ 是稠密的。这与现有方法形成鲜明对比：这些方法假设 $\boldsymbol{B}$ 具有稀疏低秩结构（Petti 等，2022），这表明它们强加了实际上并不存在的稀疏性。

3 一种估计模型参数的通用方法

3.1 HiGSS：一种估计间接效应和直接效应的算法

这里我们在下文的算法 1 中介绍一种贝叶斯方法，称为 HiGSS（基于汇总统计量的高维全基因组关联研究），用于推导后验分布 $\mathrm{pr}(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta} \mid \boldsymbol{B}, \boldsymbol{D})$，并假设潜在因子的数量 K 已知。为避免贝叶斯因子分析中出现的计算和统计问题，且正如命题 2.1 所述，HiGSS 首先使用奇异值分解推导感兴趣参数的频率学派估计值，随后利用第 3.4 节建立的渐近理论计算后验分布。HiGSS 假设 $\boldsymbol{G}$、$\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 的先验分布相互独立。

算法 1（HiGSS）输入数据：标准化全基因组关联研究估计值 $\hat{}\boldsymbol{B} \in \mathbb{R}^{S\times M}$, $\boldsymbol{D} = \mathrm{diag}\{(\boldsymbol{X}_{*1}^{\top}\boldsymbol{X}_{*1}), \ldots, (\boldsymbol{X}_{*S}^{\top}\boldsymbol{X}_{*S})\}$，潜在因子的数量 $K \geq 0$以及样本量 N。输入先验分布：$\mathrm{pr}(\boldsymbol{G} \mid \boldsymbol{D}) = \prod_{s=1}^{S} \mathrm{pr}_s(\boldsymbol{G}_{s*} \mid \boldsymbol{D})$，$\mathrm{pr}(\boldsymbol{L} \mid \boldsymbol{D}) = \prod_{m=1}^{M} \mathrm{pr}_m(\boldsymbol{L}_{m*})$，以及 $\boldsymbol{\Delta}$特异的超参数 $\{\{\pi_{sm}, \{(\gamma_{sma}, \tau^2_{sma})\}_{a\in[A]}\}\}_{s\in[S];m\in[M]}$。^输出：后验分布 $\mathrm{pr}(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta} \mid \boldsymbol{B}, \boldsymbol{D})$

1. 设 $\boldsymbol{\Gamma} = \mathrm{diag}(\gamma_1, \ldots, \gamma_K)$、$\boldsymbol{U} \in \mathbb{R}^{S\times K}$ 和 $\boldsymbol{V} \in \mathbb{R}^{M\times K}$ 分别为 $\hat{}\boldsymbol{B}$ 的前 K 个奇异值以及左、右奇异向量。定义 $\hat{}\boldsymbol{G} = S^{1/2}\boldsymbol{U}$、$\hat{}\boldsymbol{L} = S^{-1/2}\boldsymbol{V}\boldsymbol{\Gamma}$ 和 $\hat{}\boldsymbol{\Delta} = \hat{}\boldsymbol{B} - \hat{}\boldsymbol{G}\,\hat{}\boldsymbol{L}^{\top}$ 为 $\boldsymbol{G}$、$\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 的估计量。2. 计算 $\mathrm{pr}(\boldsymbol{G} \mid \hat{}\boldsymbol{G}, \boldsymbol{D}) \propto \prod_{s=1}^{S} \mathcal{N}(\hat{}\boldsymbol{G}_{s*};\ \boldsymbol{D}_{ss}^{1/2}\boldsymbol{G}_{s*}, I_K)\ \mathrm{pr}_s(\boldsymbol{G}_{s*} \mid \boldsymbol{D})$，其中 $\mathcal{N}(\boldsymbol{x};\ \boldsymbol{\mu}, \boldsymbol{V})$ 是均值为 $\boldsymbol{\mu}$、方差为 $\boldsymbol{V}$ 的正态分布在 $\boldsymbol{x}$ 处的密度。3. 设 $\hat{}\sigma^2_m = S^{-1}\sum_{s=1}^{S}\hat{}\boldsymbol{\Delta}^2_{sm}$ 为 σ^2_m 的估计值。计算 $\mathrm{pr}(\boldsymbol{L} \mid \hat{}\boldsymbol{L}, \boldsymbol{D}) \propto \prod_{m=1}^{M} \mathcal{N}\{\hat{}\boldsymbol{L}_{m*};\ \boldsymbol{L}_{m*}, (N^{-1} + S^{-1})\hat{}\sigma^2_m I_K\}\,\mathrm{pr}_m(\boldsymbol{L}_{m*})$。4. 设 $\mathrm{pr}(\boldsymbol{\Delta} \mid \hat{}\boldsymbol{\Delta}, \boldsymbol{D}) \propto \prod_{m=1}^{M}\prod_{s=1}^{S} \mathcal{N}(\hat{}\boldsymbol{\Delta}_{sm};\ \boldsymbol{D}_{ss}^{1/2}\boldsymbol{\Delta}_{sm}, \hat{}\sigma^2_m)\ \mathrm{pr}(\boldsymbol{\Delta}_{sm} \mid \pi_{sm}, \tau^2_{sm})$，其中 $\boldsymbol{\Delta}_{sm} \mid (\pi_{sm}, \tau^2_{sm}) \sim (1 - \pi_{sm})\delta_0 + \pi_{sm}\sum_{a=1}^{A}\gamma_{sma}N(0, \tau^2_{sma})$。

ˆ ˆ ˆ ˆ 5. 返回后验分布 $\text{pr}(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta} \mid \boldsymbol{B}, \boldsymbol{D}) = \text{pr}(\boldsymbol{G} \mid \boldsymbol{G}, \boldsymbol{D}) \, \text{pr}(\boldsymbol{L} \mid \boldsymbol{L}, \boldsymbol{D}) \, \text{pr}(\boldsymbol{\Delta} \mid \boldsymbol{\Delta}, \boldsymbol{D})$。

步骤 1 的直观依据是：如果 $\boldsymbol{\Delta}$ 是稀疏的，那么根据 (2.2) 和奇异值分解的精度，$\hat{\boldsymbol{G}} \approx \tilde{\boldsymbol{G}}$、$\hat{\boldsymbol{L}} \approx \boldsymbol{L}$以及残差$\hat{\boldsymbol{\Delta}}$ 应近似为 $\boldsymbol{D}^{1/2}\boldsymbol{\Delta}$ 加上独立噪声（McKennan 等，2019）。 $\boldsymbol{G}$和 $\boldsymbol{L}$的先验分布中，下标“s”和“m”表示这些先验可能取决于特定于单核苷酸多态性 s和表型 m的参数；我们在第 3.4 节中说明了步骤 2–4 所得$\hat{\boldsymbol{G}}$、$\hat{\boldsymbol{L}}$和$\hat{\boldsymbol{\Delta}}$ 的似然依据。由于 $\boldsymbol{\Delta}$ 包含 $S \times M \gtrsim 10^8$个元素，除非选择适当的先验，否则计算其后验分布是不可行的。步骤 4 中的尖峰与混合正态厚尾先验，其参数定义与假设 2.2 中的定义一致，是共轭的，因此便于进行计算。它与假设 2.2 一致纯属巧合；选择该先验是为了符合全基因组关联研究中普遍采用的先验选择（Y. Zhang 等，2018）。

步骤 5 假设总体概率 $\text{pr}(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta})$ 可以分解为前三个步骤中定义的ˆ ˆ ˆ后验分布的乘积。要理解这样做为何合理，请注意，$\{\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}\}$ 是$\hat{\boldsymbol{B}}$ 和$\hat{\boldsymbol{B}} = \hat{\boldsymbol{G}}\hat{\boldsymbol{L}}^\top + \hat{\boldsymbol{\Delta}}$的函数，这意味着总后验分布等价于

$$\text{pr}(\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta} \mid \hat{\boldsymbol{G}}, \hat{\boldsymbol{L}}, \hat{\boldsymbol{\Delta}}, \boldsymbol{D}) \propto \text{pr}(\hat{\boldsymbol{G}}, \hat{\boldsymbol{L}}, \hat{\boldsymbol{\Delta}} \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{D}) \, \text{pr}(\boldsymbol{G} \mid \boldsymbol{D}) \, \text{pr}(\boldsymbol{L} \mid \boldsymbol{D}) \, \text{pr}(\boldsymbol{\Delta} \mid \boldsymbol{D}$$

ˆ ˆ ˆ 因此，只要 $\boldsymbol{G}$、 $\boldsymbol{L}$和 $\boldsymbol{\Delta}$ 在给定 $\{\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{D}\}$的条件下相互独立，步骤 5 就成立；我们将在第 3.4 节证明这种独立性在渐近意义下成立。

ˆ ˆ 算法 1 中或许最出人意料的部分是 $\boldsymbol{L}$在步骤 3 中的似然。由于 $\boldsymbol{B}$ 有ˆS 行，现有的奇异值分解结果表明， $\boldsymbol{L}_{m*}$的渐近方差应为$S^{-1}\hat{}\sigma_m^2 I_K$ ，因为表面上看，ˆ $\boldsymbol{L}_{m*}$ 是在 S 个自由度上估计的（McKennan 等，2019）。这与步骤 3 中的方差大不相同；由于 N 通常小于 S，步骤 3 中的方差更接近 $N^{-1}\hat{}\sigma_m^2 I_K$，而非 $S^{-1}\hat{}\sigma_m^2 I_K$。这是合理的，因为式 (2.2) 中的$\hat{\boldsymbol{B}}$是通过将ˆ$N$ 样本投影到 S 个单核苷酸多态性构成的空间而得到的，这意味着 $\boldsymbol{L}_{m*}$ 实际上是在更接近 N个自由度的条件下估计的。据我们所知，下文第 3.4 节中的定理 3.2 和推论 3.1 首次证明了$\hat{\boldsymbol{L}}$的似然确实成立；正如我们在第 5 节所示，这些结果对于确保关于 $\boldsymbol{L}$的推断准确可靠至关重要。

算法 1具有通用性，因为步骤 2–4 中的似然以及步骤 5 中后验分布的形式，无论分析何种表型都相同。不过，先验分布的选择很可能取决于表型的类型。例如，代谢组学应用中选择的先验分布可能与微生物组数据中使用的不同。因此，我们在第 3.2 节中提供了一般性建议，说明如何为 $\boldsymbol{G}$ 、 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 的超参数选择先验分布。

3.2 先验分布的选择

ˆ 首先，回顾一下， $\boldsymbol{G}$ 表示基因型对少数（K个）潜在表型的影响。由于在算法 1 中， $\boldsymbol{G}$的似然服从正态分布，我们可以将 $\boldsymbol{G}$的各列视为相互独立，并采用许多针对正态数据设计的单表型全基因组关联研究先验（例如 Zhu 等人（2017）和 Y. Zhang 等人（2018））；也可以使用近期为多表型全基因组关联研究设计的方法，对 $\boldsymbol{G}$各列之间的依赖关系进行建模（Turley 等人，2018；Majumdar 等人，2018）。在我们的代谢组学应用中，我们采用前一种方法。

.$\boldsymbol{\Delta}$.的算法 1 中，我们建议采用最简单的方法：将估计 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 视为正态分布均值问题，并分别使用经验贝叶斯推导列特异性先验和矩阵级先验；借助 Stephens（2017）即可高效完成这一过程。另一种方法是将关于 M 表型之间关系的先验知识纳入 $\boldsymbol{L}$ 以及 $\boldsymbol{\Delta}$的超参数先验中。例如，可以使用 L. Zhang 等人（2021）讨论的方法，将系统发育学信息直接纳入 $\boldsymbol{L}$的先验；我们还将在第4节展示如何利用代谢通路信息构建 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$的超参数先验。

ˆˆ 在算法 1 中， $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 的似然可以写成其 K 各列似然的乘积。因此，只要它们的先验也能按列分解，就可以对 $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 的各列并行计算后验分布；传统贝叶斯或近期基于经验贝叶斯的因子分析方法则无法做到这一点（Zhong 等人，2020；W. Wang 等人，2021）。

3.3 确定潜在因子的数量

估计 K 至关重要，因为低估它可能会导致遗漏重要的遗传变异ˆ，而过度高估则可能使统计保证失效。根据 $\boldsymbol{B}$在（2.2）中的模型，人们很容易认为，由于 $\boldsymbol{R}$ （单核苷酸多态性之间的样本内相关性）˜的非对角元素近似为零，因此可以将误差矩阵 $\boldsymbol{E}$ 中的各项视为相互独立，并采用许多为误差相互独立的组学数据设计的估计量之一（Owen等人，2016；Dobriban 等人，2018；Dobriban，2020；Ke 等人，2021）。遗憾的是，如第 5 节所示，忽略 $\boldsymbol{R}$的非对角元素所产生的依赖关系，会导致这些方法失效。

解决上述问题的一种方法是采用 McKennan 等人（2022）的方法，该方法对依赖关系进行了明确建模。然而，在这种情形下应用他们的方法，需要知道 $\boldsymbol{R}$，但它无法观测。第二种选择是采用理论结果允许 $\boldsymbol{E}$各项之间存在依赖关系的方法，例如 Bai 等人（2002）、Ahn 等人（2013）或 Onatski （ 2010 ）。然而，前两种方法要求因子载荷大得不切实际（即满足普遍因子假设），而后一种方法在模拟数据和真实组学数据中显然均会失效（McKennan，2020）。为解决这些缺陷，注意到若为简便起见，假设（2.2）中的 $\boldsymbol{\Delta}$ 为 0，

$$\hat{\boldsymbol{B}} = \underbrace{\tilde{\boldsymbol{G}}\boldsymbol{L}^{\top}}_{\text{rank-}K\text{ signal}} + \underbrace{\tilde{\boldsymbol{E}}}_{\text{noise}}.$$

ˆ因此， $\boldsymbol{B}$中任何超过纯噪声模型对应奇异值的奇异值，都必然源自低秩信号。这表明，如果我们知道 $\boldsymbol{E}$ 的最大奇异值 $s_{\max}$，那么一个自然的估计量是统计 K 中超过 $s_{\max}$ 的 $\boldsymbol{B}$ 的奇异值数量。我们据此提出新方法 dBEMA（依赖型总体特征值匹配）：该方法对 $\boldsymbol{B}$ 的主体部分˜谱进行建模，以预测 $\boldsymbol{E}$ 的最大奇异值。下文将介绍 dBEMA。

˜我们的方法将 Ke 等人（2021）的方法扩展到 $\boldsymbol{E}$ 具有相关行的情形。为与第 2 节保持一致，我们始终假设 (2.1) 中的 $\boldsymbol{E}$ ，因而 (2.2) 中的 $\boldsymbol{E}$ ，服从高斯分布。不过，Onatski（2010）提出的特征值普适性理论表明，我们可以放宽这一假设，只要求 $\boldsymbol{E}$ 的元素满足较弱的矩条件。

为阐明该方法，假设 $S = S(N)$ 和 $M = M(N)$ 满足 $S/N \to \gamma_S \in (0, \infty)$ 且 $M/N \to \gamma_M \in (0, \infty)$，并令 $F_{R,N}$ 和 $F_{\sigma^2,N}$ 分别表示 $\boldsymbol{R}$ 的 S 个特征值及 $\{\sigma_m^2\}_{m\in[M]}$ 的经验分布。若分布 $\lim_{N\to\infty} F_{R,N} = F_R$ 和 $\lim_{N\to\infty} F_{\sigma^2,N} = F_{\sigma^2}$ 存在且支撑于紧区间上，则 Onatski（2010）的引理1和引理3表明， F_R 和 F_{σ^2}˜ 决定了 $\boldsymbol{E}$ 的主体部分和最大奇异值的行为。因此，问题归结为估计 F_R 和 F_{σ^2}，因为这将使我们能够预测 $\boldsymbol{E}$ 的最大奇异值。

首先，假设2.1意味着 F_R 恰好服从马尔可夫–帕斯图分布（Marčenko 等人，1967），当 $x \geq 0$、 $\lambda_- = (1 - \sqrt{\gamma_S})^2$, 且 $\lambda_+ = (1 + \sqrt{\gamma_S})^2$ 时，其形式为

$$F_R(x) = (1 - \gamma_S^{-1})1\{\gamma_S > 1\} + \int_0^x f_R(t)dt, \quad f_R(t) = (2\pi\gamma_S t)^{-1}\sqrt{(\lambda_+ - t)(t - \lambda_-)}1\{t \in [\lambda_-, \lambda_+]\}.$$

与 F_R不同， F_{σ^2} 未知，必须根据数据进行估计。遵循 Ke 等人（2021），我们将 F_{σ^2} 建模为伽马分布，并利用 $\boldsymbol{B}$的主体奇异值来估计其参数。简言之，令$s_{\min(N,M,S)} \leq \cdots \leq s_1$为 $\boldsymbol{B}$的非零奇异值，并定义 $\mathcal{B} =$

$\{s_{\lfloor(\alpha/2)\min(N,\ M,\ S)\rfloor}$，……，$s_{\lfloor(1-\alpha/2)\min(N,M,S)\rfloor}\}$为某个 $\alpha \in \in$(0，1)对应的 $\boldsymbol{B}$主体奇异值。由于对于足够大的 N， $\mathcal{B}$ 会排除 $\boldsymbol{B}$中最大的 K 个奇异值，因此 $\mathcal{B}$的元素近似为 $\boldsymbol{E}$奇异值中居中的˜比例部分。由此，可以根据F_R 以及参数化 F_{σ^2}的速率和尺度参数预测这些值。因此，我们以速率和尺度参数为变量，最小化 $\mathcal{B}$的元素与其预测值之间的差异，以估计 F_{σ^2}。有了 F_R 和对 F_{σ^2}的估计后，我们最后确定 $F_{s_{max}}$，即对˜分布的估计，该分布对应于 $\boldsymbol{E}$的最大奇异值；并将超过 $F_{s_{max}}$的 $1-\beta$ 分位数的 $\boldsymbol{B}$奇异值个数，作为因子数量的估计，其中$\beta \in \in$(0，1)由用户指定。算法见A1节。

如上所述，用于论证 dBEMA 的理论工作要求 F_{σ^2} 具有有界支撑，即 $\min\{x \in \mathbb{R} \cup \{\infty\}:\ F_{\sigma^2}(x) = 1\} < \infty$，这与我们将 F_{σ^2} 建模为伽马分布的决定相矛盾。我们在第 5 节使用模拟数据表明，尽管如此，dBEMA 仍表现良好。

3.4 理论依据

本文为算法 1 中使用的 $\boldsymbol{G}$、 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 的似然提供理论依据。关于第3.3节提出的 K 估计量一致性的理论，将在本手稿的后续版本中给出。我们首先提出关于 (2.1) 中参数的一个假设。

假设3.1 令 $c > 1$ 和 $\epsilon \in (0,1)$ 为常数，并定义 $\lambda_{K+1} = 0$。

(a) 对所有 $m \in [M]$， N/S、$M/S \in [c^{-1}, c]$、 $K \le c$、 $\sigma_m^2 \in [c^{-1}, c]$，且对所有$k \in [K]$， $\boldsymbol{\Xi}_{1k}$ 均为次指数。*(b)* $\mathbb{E}(S^{-1}\tilde{\boldsymbol{G}}^\top\tilde{\boldsymbol{G}} \mid \boldsymbol{G}) = I_K$和 $M^{-1}\boldsymbol{L}^\top\boldsymbol{L} = \mathrm{diag}(\lambda_1, \ldots, \lambda_K) \succ 0$，使得对所有 $k \in [K]$， $\lambda_k/\lambda_{k+1} \ge 1 + c^{-1}$且 $\lambda_k \in [c^{-1}N^{\epsilon-1}, c]$；对所有 $m \in [M]$， $|\boldsymbol{L}_{mk}| \le c\lambda_k^{1/2}$且 $k \in [K]$。

备注 3.1 *(b)* 中的条件隐含地假设 $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 是非随机的，不过，为避免任何混淆，以下所有概率陈述仍以它们为条件。显然，可以将 *(b)* 替换为一个等价假设，允许 $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 为随机变量。

在 (a) 中，我们假设 N、 M和 S具有可比性，以符合组学和生物样本库数据的特点。根据命题 2.1，要求 $\mathbb{E}$（$S^{-1}\tilde{\boldsymbol{G}}^\top\tilde{\boldsymbol{G}} \mid \boldsymbol{G}$） $= I_K$ 和 $M^{-1}\boldsymbol{L}^\top\boldsymbol{L}$ 为对角矩阵且对角元素递减并不失一般性；我们对相邻对角元素之比的条件是标准的特征值间隙假设，可确保 $\boldsymbol{L}$和 $\boldsymbol{G}$ 的各列能够在符号相差的意义下被识别（McKennan 等人，2019）。参数 λ_k 是第k个载荷 $\boldsymbol{L}_{*k}$ 的平方和均值，可解释为第 k个潜在因子的信号强度；当 $\lambda_1, \ldots, \lambda_K \asymp 1$时，满足“普遍因子假设”（Fan 等人，2021）。尽管普遍因子假设在文献中很常见，但它显然不适用于组学数据；在这类数据中，利用该假设的方法往往无法恢复那些非普遍但至关重要的组学变异来源（McKennan，2020）。这一点对我们的数据应用尤其重要，因为其中受遗传调控最强的因子较弱。因此，我们允许 λ_k 较大（$\lambda_k \asymp 1$, ，普遍）或中等至较小（$\lambda_k << 1$，非普遍），以确保能够对尽可能多种类型的因子进行推断。

我们首先论证 $\boldsymbol{G}$ 算法 1 第 2 步中的似然。我们将 $\mathscr{C}_d$ 和 Φ_d 分别定义为所有凸集的集合以及 $\mathbb{R}^d$中的标准正态分布函数，在本节余下部分均沿用此定义。

定理 3.1 假设假设 *2.1*、*2.2* 和 *3.1* 成立。那么，当 $N, M, S \to \infty$ 时，存在$a_1, \ldots, a_K \in \{-1, 1\}$，使得对于任意固定集合 $\mathcal{S} \subset [S]$,

$$\sup_{R_s \in \mathscr{C}_K} \left| \mathbb{P}\left[\bigcap_{s \in \mathcal{S}} \{\hat{\boldsymbol{G}}_{s*} - \boldsymbol{D}_{ss}^{1/2} \operatorname{diag}(a_1, \ldots, a_K) \boldsymbol{G}_{s*}\} \in R_s \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X} \right] - \prod_{s \in \mathcal{S}} \Phi_K(R_s) \right| = o_P(1).$$

由于符号乘数 a_1、.....、a_K对每个单核苷酸多态性 s都相同，这表明我们可以将 $\boldsymbol{G}$ 的每一列 ˆ 估计到一个符号因子。定理 3.1 也证明了在算法 1 中可以将 $\boldsymbol{G}$的各行视为相互独立的标准正态分布。接下来，我们证明算法 1 的第 3 步是合理的。

定理 3.2 假设假设 *2.1*、*2.2* 和 *3.1* 成立，且 $a_1, \ldots, a_K$如定理 *3.1* 中所述。则当 $N, M, S \to \infty$时，以下结论成立：

(i) $\hat{\sigma}_m^2 = \sigma_m^2 + o_P(1)$ 对所有 $m \in [M]$ 均成立。

(ii) $\sup_{R \in \mathscr{C}_1} \left| \mathbb{P}\left\{ \| \hat{\boldsymbol{L}}_{m*} \|_2^{-1} (\hat{\boldsymbol{G}}_{s*}^\top \hat{\boldsymbol{L}}_{m*} - \boldsymbol{D}_{ss}^{1/2} \boldsymbol{G}_{s*}^\top \boldsymbol{L}_{m*}) \in R \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X} \right\} - \Phi_1(R) \right| = o_P(1)$ 对任意 $s \in [S]$ 和 $m \in [M]$ 均成立，条件是 $N^{1/2} \| \boldsymbol{L}_{m*} \|_2 \to \infty$。

(iii) 上确界 $_{R_m \in \mathscr{C}_K} \left| \mathbb{P}\left[\bigcap_{m \in \mathcal{M}} \{(N^{-1} + S^{-1}) \hat{\sigma}_m^2\}^{-1/2} (\hat{\boldsymbol{L}}_{m*} - \boldsymbol{A} \boldsymbol{L}_{m*}) \in R_m \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X} \right] - \prod_{m \in \mathcal{M}} \Phi_K(R_m) \right| = o_P$ (1) 对任意固定集合 $\mathcal{M} \subset [M]$ 以及满足 $\| \boldsymbol{A} - \operatorname{diag}(a_1, \ldots, a_K) \|_2 = O_P(N^{-1/2})$的某个 $\boldsymbol{A} \in \mathbb{R}^{K \times K}$均成立。

结果 (i) 和 (ii) 表明，算法 1 的步骤 3 中的 $\hat{\sigma}_m^2$ 对 σ_m^2 是一致的，因此我们可以推断间接效应 $\boldsymbol{G}_{s*}^\top \boldsymbol{L}_{m*}$。如果表型 m 受到至少一个潜在因子的非平凡影响，则条件 $N^{1/2} \| \boldsymbol{L}_{m*} \|_2 \to \infty$ 将成立；在我们的数据应用中，几乎所有 m 似乎都满足该条件。ˆ 结果 (iii) 表明， $\boldsymbol{L}_{m*}$ 渐近正态，其渐近方差如步骤 3 所示。ˆ然而， $\boldsymbol{L}_{m*}$' 的渐近分布以 $\boldsymbol{A}\boldsymbol{L}_{m*}$ 为中心，而非以 $\operatorname{diag}(a_1, \ldots, a_K)\boldsymbol{L}_{m*}$ 为中心，这是基于奇异值分解的估计量的一项标准性质（J. Wang 等，2017）。尽管二者差异很小，但可能仍不足以保证我们能够推断 $\boldsymbol{L}_{m*}$' 的稀疏模式，因为即使 $\boldsymbol{L}_{mk} = 0$，$(\boldsymbol{A}\boldsymbol{L}_{m*})_k$ 也可能大于其估计值的标准差。当数据同时包含普遍因子和非普遍因子时，这一点尤为重要，因为 $\boldsymbol{A}$ 的非对角元素可能会使较稀疏的非普遍因子载荷受到稠密的普遍因子载荷污染。下文的推论 3.1 表明不会发生这种污染。

推论 3.1 除定理 *3.2* 的条件外，假设对于所有 $k \in [K]$， $\lambda_k \geq \eta$ 或 $\lambda_k \to 0$ 满足 $M \to \infty$，且存在某个常数 $\eta > 0$。定义普遍因子的数量 K_{pf}：若$\lambda_1 \to 0$，则为 *0*；若 $\lambda_{K_{\mathrm{pf}}} \geq \eta$ 但 $\lambda_{K_{\mathrm{pf}}+1} \to 0$ 满足 $M \to \infty$，则为 $K_{\mathrm{pf}} \in [K]$。那么，定理 *3.2* 中的 $\boldsymbol{A}$ 满足

$$\boldsymbol{A} = \begin{cases} \operatorname{diag}(a_1, \ldots, a_K) & \text{if } K_{\mathrm{pf}} = 0 \\ \boldsymbol{A}_{\mathrm{pf}} \oplus \operatorname{diag}(a_{K_{\mathrm{pf}}+1}, \ldots, a_K) & \text{if } K_{\mathrm{pf}} \in [K-1], \quad \| \boldsymbol{A}_{\mathrm{pf}} - \operatorname{diag}(a_1, \ldots, a_{K_{\mathrm{pf}}}) \|_2 = O_P(N^{-1/2}) \\ \boldsymbol{A} & \text{if } K_{\mathrm{pf}} = K \end{cases}$$

并且，对于任意固定的集合 $\mathcal{M} \subset [M]$ 和非普遍因子 $k \in \{K_{\mathrm{pf}} + 1, \ldots, K\}$,

$$\sup_{R_m \in \mathscr{C}_1} \left| \mathbb{P}\left[\bigcap_{m \in \mathcal{M}} \{(N^{-1} + S^{-1}) \hat{\sigma}_m^2\}^{-1/2} (\hat{\boldsymbol{L}}_{mk} - a_k \boldsymbol{L}_{mk}) \in R_m \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X} \right] - \prod_{m \in \mathcal{M}} \Phi_1(R_m) \right| = o_P(1).$$

$\boldsymbol{A}$ 的表达式意味着我们无需担心上述污染；该污染引出了后续结果，即非普遍因子载荷的渐近分布以其真实载荷为中心（至多相差一个符号）。虽然普遍因子载荷不一定如此，但这对推断的影响可以忽略不计，因为污染只会轻微影响其较大的非零项；正如数据示例所示，普遍因子更可能反映非遗传变异。最后，我们考察算法 1 的步骤 4 和 5。

定理 3.3 假设假设 *2.1*、*2.2* 和 *3.1* 成立，且 a_1、$\ldots$、a_K、 $\boldsymbol{A}$ 如定理 *3.1* 和 *3.2* 所述。则对于任意固定集合 $\mathcal{S} \subset [S]$、 $\mathcal{M} \subset [M]$，当 N、 M、 $S \to \infty$ 时，以下结论成立：

$$(i) \quad \sup_{R_{sm} \in \mathscr{C}_1} \left| \mathbb{P}\left\{ \bigcap_{\substack{s \in \mathcal{S} \\ m \in \mathcal{M}}} \hat{\sigma}_m^{-1}(\hat{\boldsymbol{\Delta}}_{sm} - \boldsymbol{D}_{ss}^{1/2}\boldsymbol{\Delta}_{sm}) \in R_{sm} \mid \boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X} \right\} - \prod_{\substack{s \in \mathcal{S} \\ m \in \mathcal{M}}} \Phi_1(R_{sm}) \right| = o_P(1).$$

(ii) 定义 $\boldsymbol{Z}_s^{(\mathrm{G})} = \hat{}\boldsymbol{G}_{s*} - \boldsymbol{D}_{ss}^{1/2}\mathrm{diag}(a_1, \ldots, a_K)\boldsymbol{G}_{s*}$、 $\boldsymbol{Z}_m^{(\mathrm{L})} = \{(N^{-1} + S^{-1})\hat{}\sigma_m^2\}^{-1/2}(\hat{}\boldsymbol{L}_{m*} - \boldsymbol{A}\boldsymbol{L}_{m*})$以及 $z_{sm}^{(\Delta)} = \hat{}\sigma_m^{-1}(\hat{}\boldsymbol{\Delta}_{ms} - \boldsymbol{D}_{ss}^{1/2}\boldsymbol{\Delta}_{ms})$。则随机变量集合 $\{\boldsymbol{Z}_s^{(\mathrm{G})}\}_{s \in \mathcal{S}}$、 $\{\boldsymbol{Z}_m^{(\mathrm{L})}\}_{m \in \mathcal{M}}$ 和 $\{z_{sm}^{(\Delta)}\}_{(s,m) \in \mathcal{S} \times \mathcal{M}}$ 在给定 $\{\boldsymbol{G}, \boldsymbol{L}, \boldsymbol{\Delta}, \boldsymbol{X}\}$ 的条件下渐近独立。

ˆ 第 (i) 部分为算法 1 中 $\boldsymbol{\Delta}$ 的似然提供了依据，该算法假设其各个条目相互独立且服从正态分布。第 (ii) 部分中 z 统计量的渐近独立性——它们是 $\boldsymbol{G}$、 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 的 ˆ ˆ ˆ 标准化等价量——为步骤 5 提供了依据。

4 代谢物全基因组关联研究中以通路为中心的先验和推断

算法 1 提供了一种通用方法，用于计算感兴趣参数的后验分布；其中，步骤 1 中的频率学派估计量、步骤 2–4 中的似然以及步骤 5 中后验分布的形式在不同数据集中均保持一致。不过，先验分布的选择取决于具体领域。本节将介绍一组专用于代谢组学的先验分布，这也凸显了我们的方法能够利用生物学信息先验的优势，而这类先验在标准贝叶斯因子分析中难以处理。

为说明我们的先验分布，图2(a)展示了代谢物可以按层级划分为一组超通路和亚通路；图2(b)所示的明显通路特异性变异表明，同一通路中的代谢物表现相似。因此，我们遵循第 3.2 节的建议，为 $\boldsymbol{L}$ 和 $\boldsymbol{\Delta}$ 设计纳入代谢通路信息的先验分布。

4.1 间接效应的先验

基因型对代谢物水平的间接效应由 $\boldsymbol{G}$ 和 $\boldsymbol{L}$参数化。我们对 $\boldsymbol{G}$的先验遵循第 3.2 节中的讨论，并假设每列的条目均抽取自尖峰与板块先验：

$$\boldsymbol{G}_{sk} \mid (\pi_{G_k}, \tau_{G_k}^2, \boldsymbol{D}) \sim (1 - \pi_{G_k})\delta_0 + \pi_{G_k} N(0, \tau_{G_k}^2 / \boldsymbol{D}_{ss}), \quad s \in [S]; k \in [K], \tag{4.1}$$

ˆ 在算法 1 中，该先验与 $\boldsymbol{G}$的正态似然共轭。按 $\boldsymbol{D}_{ss}$ 标准化意味着标准化效应 $\boldsymbol{D}_{ss}^{1/2}\boldsymbol{G}_{sk}$ 在各单核苷酸多态性 s 之间服从相同分布，这也反映了遗传效应通常与次要等位基因频率成反比这一观察结果（Park等人，2011）。我们通过经验贝叶斯估计 π_{G_k}， $\tau_{G_k}^2$。

我们对 $\boldsymbol{L}$ 的先验纳入了代谢通路信息，其构建旨在反映图2(b)中的以下观察结果和推断目标。首先，来自同一子通路的代谢物载荷似乎取自同一分布，而同一超通路内不同子通路的分布彼此比不同超通路的分布更相似。

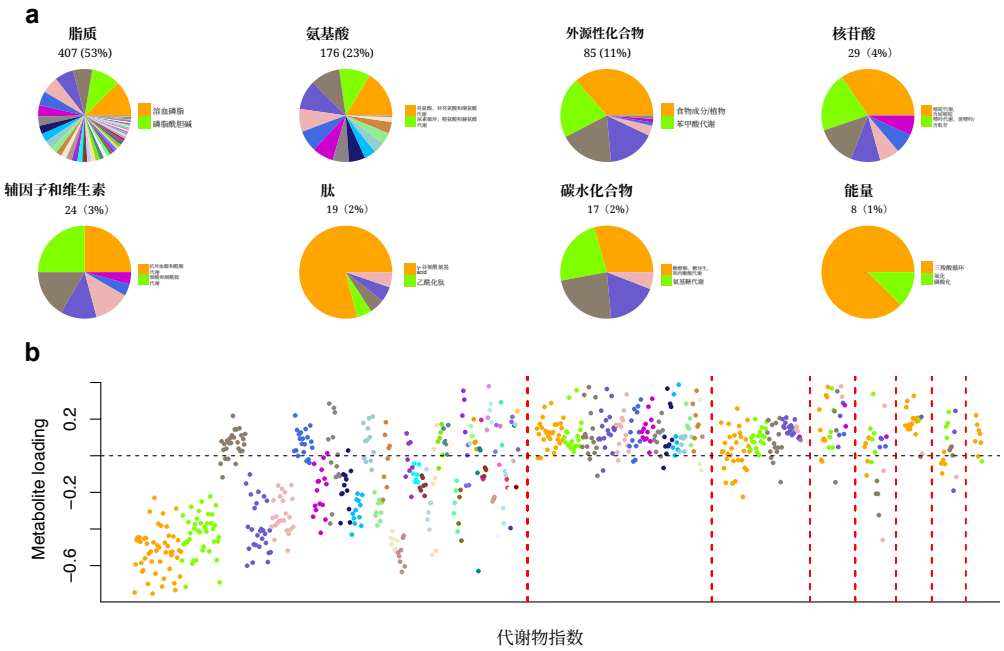

图2：**(a)**：我们数据示例中代谢物的层次划分。颜色表示各超通路所包含的子通路，数字表示各超通路中的代谢物数量。ˆ 标出了每个超通路中最大的两个子通路。**(b)**：我们数据示例中的 $\boldsymbol{L}_{*2}$ 。红线划分超通路，子通路的颜色与**(a)**中的一致。

其次，图2(b)表明，许多子通路具有相同的载荷分布。例如，前两个橙色和绿色脂质子通路可能具有相同的分布。因此，我们的先验应有助于对子通路进行聚类，从而将生物过程关联起来，得到更易于解释的推断。第三，预计有些代谢物会表现为“离群值”，因为它们被错误地划分，或不符合其所属子通路的分布。

为解释这些观察结果，我们使用层次狄利克雷过程（Teh 等，2006）对各子通路的分布参数建模。由此可对子通路进行聚类：来自同一超通路的两个子通路先验上比来自不同超通路的子通路更有可能属于同一个簇。随后，代谢物载荷要么从其所在子通路的分布中抽取，要么从“离群值”分布中抽取，以解释上述第三项观察结果。用数学形式表示，设 $p \in [P]$ 表示超通路的索引，$b \in [B_p]$ 表示其所属子通路的索引，$m \in [M_{pb}]$ 表示其所属代谢物的索引。我们假设 $\boldsymbol{L}$ 的各列相互独立；为简化记号，省略其对因子数量的依赖。因此，在第4.1节余下部分中，令 $\boldsymbol{L}_{pbm} \in \mathbb{R}$ 表示第 m 个代谢物在第 b 个子通路和第 p 个超通路中的一个因子的载荷。对于浓度参数 γ，$\alpha_0 > 0$ 以及定义在 $\mathbb{R} \times \mathbb{R}_{\geq 0}$ 上的概率测度 H，

$$
\begin{aligned}
&F_0 \mid (\gamma, H) \sim \mathrm{DP}(\gamma, H), \quad \pi_{\mathrm{out}} \sim \mathrm{Beta}(c, d) \\
&F_p \mid (\alpha_0, F_0) \sim \mathrm{DP}(\alpha_0, F_0), \quad p \in [P] \\
&(\mu_{pb}, \phi^2_{pb}) \mid F_p \sim F_p, \quad b \in [B_p] \\
&\boldsymbol{L}_{pbm} \mid (\mu_{pb}, \phi^2_{pb}, \pi_{\mathrm{out}}, a_{\mathrm{out}}, b_{\mathrm{out}}) \sim (1 - \pi_{\mathrm{out}}) N(\mu_{pb}, \phi^2_{pb}) + \pi_{\mathrm{out}} U[a_{\mathrm{out}}, b_{\mathrm{out}}], \quad m \in [M_{pb}].
\end{aligned}
\tag{4.2}
$$

图3(a)以图示方式描述了这一先验。暂且假设 $\pi_{\mathrm{out}} = 0$,$\boldsymbol{L}_{pbm}$ 是从层次狄利克雷过程混合模型中抽取的，其中混合成分

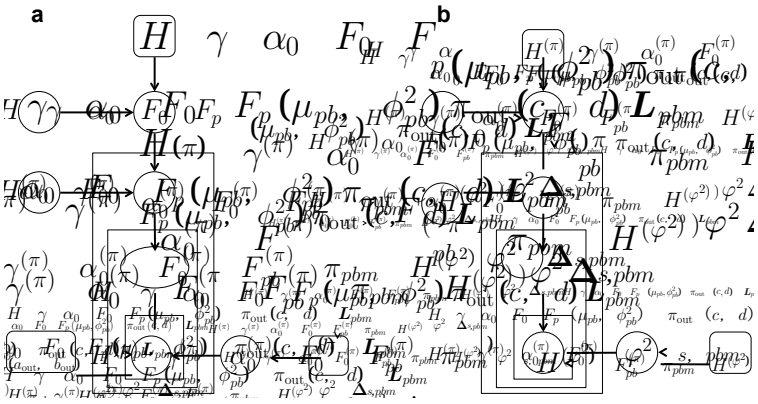

图3： $\boldsymbol{L}$ 的先验的图模型（a）和 $\boldsymbol{\Delta}$ 的先验（b）。方形和圆形节点分别表示固定量和随机量。矩形内的节点表示重复观测。

（簇）由 $\{(\mu_{pb},\ \phi_{pb}^2)\}_{p\in[P];b\in[B_p]}$ 的唯一取值定义（Teh 等，2006）。简而言之，首先从基准测度为 H 的狄利克雷过程中抽取概率测度 F_0，且以概率1，其形式为 $\sum_{j=1}^{\infty}\beta_{0,\ j}\delta_{(\mu_j,\phi_j^2)}$，其中概率权重为 $\beta_{0,\ j}$，且 $(\mu_j,\ \phi_j^2)\in\mathbb{R}\times\mathbb{R}_{\geq 0}$。随后，每个 F_p 都从另一个基准测度为 F_0 的狄利克雷过程中抽取，这意味着 F_p 也可表示为 $\sum_{j=1}^{\infty}\beta_{p,\ j}\delta_{(\mu_j,\phi_j^2)}$，其中概率权重为 $\beta_{p,\ j}$。(4.2)中的第三行表明，超通路内的子通路根据其参数 $(\mu_{pb},\ \phi_{pb}^2)$ 被分配到各个簇中，共享参数的子通路会被分到同一簇。然而，由于 F_p 共享原子，来自不同超通路的子通路也可能被分配到同一簇； α_0 的值越大，这种情况发生的概率就越高。

式 (4.2) 中的参数 π_{out} 表示所有代谢物中“离群值”的比例，即其载荷不符合层次狄利克雷过程混合模型的代谢物。由于我们假设它们的载荷可以以相同概率取任意值，因此将其视为均匀分布，并设定 a_{out}，b_{out}，这样 $[a_{\text{out}},b_{\text{out}}]$ 就包含所有观测到的载荷。我们令 $\pi_{\text{out}}\sim$ 服从 Beta(c, d)，以便让数据帮助确定离群值的频率，并设定 $c=1/5,\ d=999/5$，因此在 1,000 个代谢物中，我们先验地 预期有一个离群值，且离群值不超过五个的概率为 0.95。

与 Teh 等人（2006）类似，我们为 γ 和 α_0 设置弥散先验。我们设置基分布 H，使得 $\mathbb{P}\{(\mu_{pb},\ \phi_{pb}^2)=(0,\ 0)\mid H\}>0$，并且 $\mu_{pb}\perp\!\!\!\perp\phi_{pb}^2\mid\{H,(\mu_{pb},\ \phi_{pb}^2)\neq(0,\ 0)\}$。前者体现了我们的信念：某些子通路的载荷将恰好为 0；我们令 $\mu_{pb}\mid\{H,(\mu_{pb},\ \phi_{pb}^2)\neq(0,\ 0)\}$ 服从尖峰与混合正态厚尾分布，以确保共轭性，其中厚尾部分经过选择，以近似马蹄先验（Carvalho 等人，2009）。我们令 $\phi_{pb}^2\mid\{H,(\mu_{pb},\ \phi_{pb}^2)\neq(0,\ 0)\}$ 服从经过密集离散化的伽马分布；离散化有助于加快计算，因为这样我们可以在吉布斯采样器中对 $\boldsymbol{L}$ 积分消去。A2 节提供了更多细节，并介绍了我们的吉布斯采样器。该采样器利用层次狄利克雷过程的中国餐馆加盟店参数化，从而高效地从后验分布中采样。

4.2 直接效应的先验

与我们对 $\boldsymbol{L}$ 的先验类似，我们在算法 1 中利用层次狄利克雷过程混合模型为 $\boldsymbol{\Delta}$ 的超参数设定先验。不过，它与第4.1节所述的先验有两点重要区别。首先，它只考虑子通路，而忽略超通路标签。虽然这确实舍弃了一些信息，但我们发现，纳入超通路只会使估计值发生微不足道的变化，不值得增加额外的计算量。其次，我们对 $\boldsymbol{\Delta}$ 的超参数所建立的模型

会对代谢物而非子通路进行聚类，从而有助于汇集相关代谢物的信息，改善原本不确定的超参数估计。

我们构建 $\boldsymbol{\Delta}$ 的先验，使其类似于 Ruffieux 等人（2017年）所用的先验；他们将代谢物的遗传效应建模为尖峰与正态板块，并为不同代谢物设置特定的尖峰概率，同时令板块方差保持恒定。为与第4.1节保持一致，我们沿用相同的超通路和子通路记号，并为第4.2节的目的重新索引 $\boldsymbol{\Delta}$ ，使得 $\boldsymbol{\Delta}_{s,\ pbm}$ 表示第 s 个单核苷酸多态性对超通路 $p \in [P]$ 中子通路 $b \in [B_p]$ 的代谢物 $m \in [M_{pb}]$ 的直接效应。于是

$$
\begin{aligned}
&F_0^{(\pi)} \mid \{\gamma^{(\pi)}, H^{(\pi)}\} \sim \mathrm{DP}(\gamma^{(\pi)}, H^{(\pi)}), \quad \varphi^2 \mid H^{(\varphi^2)} \sim H^{(\varphi^2)} \\
&F_{pb}^{(\pi)} \mid \{\alpha_0^{(\pi)}, F_0^{(\pi)}\} \sim \mathrm{DP}(\alpha_0^{(\pi)}, F_0^{(\pi)}), \quad p \in [P]; b \in [B_p] \\
&\pi_{pbm} \mid F_{pb}^{(\pi)} \sim F_{pb}^{(\pi)}, \quad m \in [M_{pb}] \\
&\boldsymbol{\Delta}_{s,pbm} \mid (\pi_{pbm}, \varphi^2, \boldsymbol{D}) \sim (1-\pi_{pbm})\delta_0 + \pi_{pbm} N\left(0, \varphi^2 \sigma_{pbm}^2 / \boldsymbol{D}_{ss}\right), \quad s \in [S],
\end{aligned}
\tag{4.3}
$$

其中， $H^{(\pi)}$ 和 $H^{(\varphi^2)}$ 是定义在 $[0,1]$ 和 $(0,\ \infty)$ 上的概率测度，而 σ_{pbm}^2 是在（2.1）中定义的 σ_m^2 的重新索引版本。实际应用中，我们将其替换为算法 1 第3步中定义的相合估计量。将 $\boldsymbol{\Delta}$ 的板块方差按 σ_{pbm}^2 标准化，可校正代谢物水平的异质性；除以 $\boldsymbol{D}_{ss}$ 则反映了遗传效应通常与次要等位基因频率成反比这一事实。图 3(b) 以图示方式展示了（4.3）；与第4.1节中的做法相同，我们对浓度参数 $\gamma^{(\pi)}$， $\alpha_0^{(\pi)}$ 赋予弥散先验。

对测数据 φ^2 提供信息，因为它在所有单核苷酸多态性和代谢物中共享。然而，π_{pbm} 具有代谢物特异性，考虑到 $\boldsymbol{\Delta}$的稀疏性，其估计值可能不确定。为此，我们对π_{pbm} 采用层次狄利克雷过程混合先验，从而对代谢物进行聚类；同一簇中的代谢物共享π_{pbm}，并且更有可能与同一子通路中的其他代谢物归为一组。这样便能汇集相关代谢物的信息，从而降低我们对π_{pbm}的不确定性。

还需要指定 $H^{(\pi)}$ 和 $H^{(\varphi^2)}$，同时必须确保推断在计算上可行。令$\Theta = \{\{\pi_{pbm}\}_{p\in[P];b\in[B_p];\ m\in[M_{pb}]}, \varphi^2\}$。则 $\boldsymbol{\Delta}$的后验分布为

$$
\mathrm{pr}\{\boldsymbol{\Delta} \mid \hat{\boldsymbol{\Delta}}, \boldsymbol{D}, H^{(\pi)}, H^{(\varphi^2)}\} = \int \mathrm{pr}(\boldsymbol{\Delta} \mid \hat{\boldsymbol{\Delta}}, \boldsymbol{D}, \Theta)\, \mathrm{pr}\{\Theta \mid \hat{\boldsymbol{\Delta}}, \boldsymbol{D}, H^{(\pi)}, H^{(\varphi^2)}\} d\Theta. \tag{4.4}
$$

积分中的第一个后验分布与算法 1 第 4 步中的后验分布相同，是一个具有闭式解的逐项尖峰与正态板块分布。然而， Θ的后验分布通常难以处理，这意味着必须通过马尔可夫链蒙特卡罗近似计算该积分。遗憾的是，若对 $H^{(\varphi^2)}$ 和 $H^{(\pi)}$ 采用标准的共轭逆伽马分布和贝塔分布进行吉布斯采样器，则需要对 $\boldsymbol{\Delta}$ 进行采样，并且由于 $\boldsymbol{\Delta}$维度很高，混合速度会慢到难以处理。我们通过令 $H^{(\pi)}$ 和 $H^{(\varphi^2)}$为离散变量来解决这一问题。简而言之，令 $\{0\} \cup \mathcal{X}^{(\pi)} \subset [0,1)$和 $\mathcal{X}_{(\varphi^2)} \subset (0,\infty)$为有限集合，包含$\pi_{pbm}$和 φ^2的所有可能取值。然后我们设定 $H^{(\pi)}$和 $H^{(\varphi^2)}$，使得

$$
\begin{aligned}
&\mathbb{P}\{\pi_{pbm} = 0 \mid H^{(\pi)}\} = 0.5, \quad \mathbb{P}\{\pi_{pbm} = x \mid H^{(\pi)}\} = 0.5/|\mathcal{X}^{(\pi)}| \text{ for all } x \in \mathcal{X}^{(\pi)} \\
&\mathbb{P}\{\varphi^2 = x \mid H^{(\varphi^2)}\} = 1/|\mathcal{X}^{(\varphi^2)}| \text{ for all } x \in \mathcal{X}^{(\varphi^2)},
\end{aligned}
$$

其中， $H^{(\pi)}$在 0 处的非零概率质量反映了我们的预期：部分代谢物不会产生直接效应。考虑到许多代谢物的水平至少会受到一个单核苷酸多态性的影响，我们软件默认采用的 0.5 可能偏保守（Karjalainen 等，2022）。重要的是，上述离散分布通过允许我们对 $\boldsymbol{\Delta}$ 积分消去，从而加快吉布斯采样器的混合速度，以近似（4.4）；此外，它还使我们能够预先计算所有提议新状态所需的 Θ 对应的似然 $\mathrm{pr}\{\hat{}\boldsymbol{\Delta} \mid \Theta, \boldsymbol{D}, H^{(\pi)}, H^{(\varphi^2)}\} = \mathrm{pr}(\hat{}\boldsymbol{\Delta} \mid \Theta, \boldsymbol{D})$，从而显著缩短计算时间。A3 节详细说明了我们如何设定$\mathcal{X}^{(\pi)}$ 和 $\mathcal{X}^{(\varphi^2)}$。

4.3 直接效应的通路特异性推断

我们在（4.3）中对直接效应的先验使我们能够比较代谢亚通路的“直接”遗传力。为此，令 Θ 如第4.2节所定义，且 $\boldsymbol{Y}_{i,\ pbm}$ 表示个体 i 在子通路和超通路 b 和 p 中代谢物 m 的丰度。那么，在模型（2.1）以及 $\boldsymbol{\Delta}$ 在（4.3）中的先验条件下，给定潜在因子 $\boldsymbol{C}$ 时，代谢物 m 的遗传力为

$$h^2_{pbm} := \frac{N^{-1}\sum_{i=1}^{N}\mathrm{Var}(\sum_{s=1}^{S}\boldsymbol{X}_{is}\boldsymbol{\Delta}_{s,pbm} \mid \boldsymbol{L},\boldsymbol{C},\Theta)}{N^{-1}\sum_{i=1}^{N}\mathrm{Var}(\boldsymbol{Y}_{i,pbm} \mid \boldsymbol{L},\boldsymbol{C},\Theta)} = \frac{\varphi^2\pi_{pbm}}{\varphi^2\pi_{pbm}+1},$$

其中，第一个分数的分子和分母分别是代谢物直接遗传效应的平均方差和丰度。由于代谢物的遗传力是 π_{pbm} 的增函数，因此，评估超通路 p 中子通路 b 的遗传力，等价于研究 $\{\pi_{pbm}\}_{m\in[M_{pb}]}$ 的分布。该分布正是式 (4.3) 中定义的 $F^{(\pi)}_{pb}$；当从 $F^{(\pi)}_{pb}$ 中抽取的值趋于较大时，子通路的遗传力就更高。因此，我们将超通路 p 中子通路 b 的直接遗传力评分 DHS_{pb} 定义为

$$DHS_{pb} = \mathbb{P}\{f^{(\pi)}_{pb} > f^{(\pi)}_{0} \mid \hat{\boldsymbol{\Delta}},\boldsymbol{D},H^{(\pi)},H^{(\varphi^2)}\},\quad f^{(\pi)}_{pb} \mid F^{(\pi)}_{pb} \sim F^{(\pi)}_{pb},\quad f^{(\pi)}_{0} \mid F^{(\pi)}_{0} \sim F^{(\pi)}_{0}. \tag{4.5}$$

由于 $F^{(\pi)}_{0} = \mathbb{E}\{F^{(\pi)}_{pb} \mid F^{(\pi)}_{0}\}$，$DHS_{pb}$ 可解释为从子通路 b 中随机选取的代谢物比平均子通路中的代谢物遗传力更高的概率。这些评分可用于按遗传力对子通路进行排序，其中 DHS_{pb} 较大的子通路遗传力更高。我们在第6节中展示了这些评分的实用性。

5 模拟

5.1 模拟设置

我们从 $N = 500$ 个个体中模拟了代谢组学数据，其中 $M = 257$ 种代谢物受 $S = 5{,}000$ 个单核苷酸多态性调控，以评估我们对潜在因子数量的估计量和算法 1。假设存在 $K = 10$个潜在因子，生成未观测到的基因型 $\boldsymbol{X}$ 和代谢物水平 $\boldsymbol{Y}$：

$$\begin{aligned}
&f_s \sim U[0.05, 0.5], \quad \boldsymbol{X}_{is} \sim \mathrm{Bin}(2, f_s), \quad i \in [N]; s \in [S] \\
&\boldsymbol{G}_{sk} \sim (1-10^{-3})\delta_0 + 10^{-3}N(0, 2/N), \quad \boldsymbol{L}_{*k} \sim HDP(\lambda_k), \quad s \in [S]; k \in [K] \\
&\boldsymbol{\Delta}_{sm} \sim (1-\pi_{\Delta_m})\delta_0 + \pi_{\Delta_m}N(0, \tau^2_{\Delta s}), \quad \sigma^2_m \sim \mathrm{Gamma}(1,1), \quad s \in [S]; m \in [M] \\
&\boldsymbol{C} \sim MN(\boldsymbol{X}\boldsymbol{G}, I_N, I_K), \quad \boldsymbol{Y} \sim MN(\boldsymbol{C}\boldsymbol{L}^\top + \boldsymbol{X}\boldsymbol{\Delta}, \mathrm{diag}(\sigma^2_1,\ldots,\sigma^2_M), I_N).
\end{aligned} \tag{5.1}$$

ˆ 对于每个数据集，我们将全基因组关联研究汇总统计量 β_{sm} 定义为回归 $\boldsymbol{Y}_{*m}$ 对 $\boldsymbol{X}_{*s}$ 的斜率估计，并定义观测矩阵 ˆ$\boldsymbol{B}$，使得 ˆ $\boldsymbol{B}_{sm} = (\boldsymbol{X}^\top_{*s}\boldsymbol{P}^{\perp}_{\mathbf{1}_N}\boldsymbol{X}_{*s})^{1/2}$ˆ$\beta_{sm}$，其中$\boldsymbol{P}^{\perp}_{\mathbf{1}_N}$是将向量投影到 $(\mathbf{1},\ldots,\mathbf{1}) \in \mathbb{R}^{1\times N}$的核空间上的正交投影矩阵。

参数 π_{Δ_m} 被模拟为 $\pi_{\Delta_m} \sim 0.3\delta_0 + 0.7F$：$F$ 服从均值为 0.005、取值范围为 0 至 0.01 的截断指数分布，且 $\tau^2_{\Delta s} = 5^2/(\boldsymbol{X}^\top_{*s}\boldsymbol{P}^{\perp}_{\mathbf{1}_N}\boldsymbol{X}_{*s})$。这些参数以及$\boldsymbol{G}$的超参数，都与我们在真实数据示例中估计的值相当。我们将代谢物划分为超通路和亚通路，使每个分区所含代谢物的数量与真实数据示例相符（详情见 A4 节），并从层次狄利克雷过程（HDP）混合模型中独立抽取 $\boldsymbol{L}$ 的各列。参数 $\lambda_k = M^{-1}\mathbb{E}(\boldsymbol{L}^\top_{*k}\boldsymbol{L}_{*k})$决定了每个潜在因子的强度；正如我们在

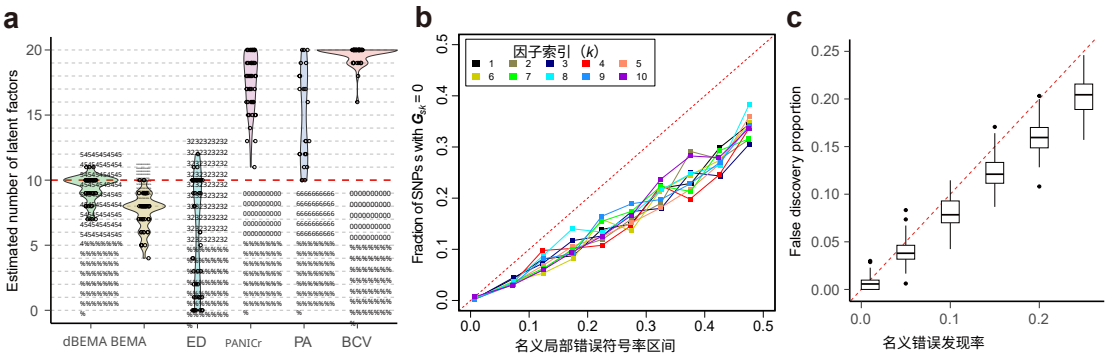

图 4：50 组模拟数据集的模拟结果。（a）： K 的估计值。dBEMA 是我们提出的方法，BCV 的最大 K 设为 20。百分比表示估计值与真实 $K = 10$ 相符的比例。（b）：对 $\boldsymbol{G}$ 进行推断。局部错误符号率分为 10 个区间。（c）：对 $\boldsymbol{\Delta}$ 进行推断。每个点代表一个模拟数据集。

第 3.4 节中所述，λ_k 越大，因子越强，也越容易被恢复。为使模拟更贴近现实，我们让 $\lambda_1 > \cdots > \lambda_K$ 取值范围较广，其中 $\lambda_1 = 3.35$ 对应一个影响广泛的大因子，而 $\lambda_K = 0.25$ 较弱。值得注意的是，这些取值意味着信号与噪声之间的特征值间隙较小，因为 $\boldsymbol{B}$ 的第 K 个奇异值平均只比 $\|\boldsymbol{B} - \mathbb{E}(\boldsymbol{B} \mid \boldsymbol{X}, \boldsymbol{C}, \boldsymbol{L}, \boldsymbol{\Delta})\|_2$（即噪声的大小）大 14.5%。因此，与真实数据中的情况一样，恢复所有因子并非易事。

5.2 模拟结果

我们首先考虑估计潜在因子的数量 K。我们将第3.3节中介绍的方法（dBEMA）与 Ke 等人（2021）提出的整体特征值匹配（BEMA）、Onatski（2010）采用的特征值差分法（ED）、Bai 等人（2002）的信息准则（PANICr）、平行分析（PA）（Leek等人，2007）以及双重交叉验证（BCV）（Owen 等人，2016）进行了比较。图 4(a) 展示了结果，我们的方法明显优于现有方法。值得注意的是，PA 和 BCV 在组学分析流程中应用广泛，却严重高估了 K。这是因为式（2.2）中 $\boldsymbol{B}$ 的表达式所涉及的误差矩阵 $\boldsymbol{E}$ 存在行依赖性（单核苷酸多态性），这与它们假设误差矩阵的元素相互独立相矛盾。因此，它们将误差的行依赖性误判为潜在结构。

假设已知 K，接下来我们考虑估计用于参数化间接效应的 $\boldsymbol{G}$ 和 $\boldsymbol{L}$，以及直接效应 $\boldsymbol{\Delta}$。首先，我们使用算法 1 和第 4.1 节所述的先验分布来估计 $\boldsymbol{G}$ 和 $\boldsymbol{L}$。为解决因子可识别性问题，并使估计的因子与模拟的因子相匹配，我们通过求解普鲁克分析问题 $\tilde{\boldsymbol{Q}} = \operatorname{argmin}_{\boldsymbol{Q}^\top \boldsymbol{Q} = I_K} \|\boldsymbol{G}\boldsymbol{Q} - \boldsymbol{G}\|_F$，对算法 1 中定义的估计量 $\boldsymbol{G}$ 和 $\boldsymbol{L}$ 进行旋转，其中 $\boldsymbol{G}$ 在式 (2.2) 中定义。我们使用局部错误符号率（lfsr）对 $\boldsymbol{G}$ 进行推断：

$$\operatorname{lfsr}(\boldsymbol{G}_{sk}) = 1 - \max\{\mathbb{P}(\boldsymbol{G}_{sk} > 0 \mid \hat{\boldsymbol{B}}, \boldsymbol{D}), \mathbb{P}(\boldsymbol{G}_{sk} < 0 \mid \hat{\boldsymbol{B}}, \boldsymbol{D})\}, \quad s \in [S]; k \in [K]. \tag{5.2}$$

局部错误符号率是指系数推断出的非零符号出错的概率，其值至少与局部错误发现率一样大。此处使用局部错误符号率，是为了避免在零分布与备择分布难以区分时可能出现的局部错误发现率问题（Stephens，2017）。图 4(b) 显示，我们对 $\boldsymbol{G}$ 的局部错误符号率估计能够控制局部错误发现率。A4 节包含 $\boldsymbol{L}$ 的结果，展示了我们在第 3.4 节提出的新理论结果的重要性，这些结果给出了 $\boldsymbol{L}$ 的渐近分布。

a

方法	估计的 K
dBEMA	**36**
BEMA	N/A
ED	1
PANICr	55
PA	N/A
BCV	100*

b

#遗传因子（共 36 个）	32
#唯一单核苷酸多态性 in 遗传因子（共 70,140 个）	418
#独特代谢物遗传因子中的（共 765 个）	765

表1：(a)：估计的因子数量。“N/A”表示该方法的软件无法在这些数据上运行；BCV 考虑的因子数量上限为 100。(b)：间接效应结果。

最后，我们估计了 $\mathbf{\Delta}$。由于第4.2节针对代谢组学数据推荐的 $\mathbf{\Delta}$ 需要大量计算，我们通过以下方式对这些模拟数据进行推断：先使用 $\mathbf{\Delta}$在算法 1 中的似然计算零假设$H_{0,\ sm}$ 的p值，再通过 Benjamini-Hochberg 程序（Benjamini等，1995）控制错误发现率： $\mathbf{\Delta}_{sm} = 0$。图4(c) 显示，这种方法能够控制错误发现率，表明 $\mathbf{\Delta}$在算法 1 中的似然是合适的。

6 真实数据分析

我们使用来自参与 METSIM研究的$N = 6{,}136$ 名芬兰成年人的代谢物汇总统计量（Yin 等，2022a），展示了我们方法的统计功效。简而言之，我们纳入次要等位基因频率为 $\geq 5\%$ 的单核苷酸多态性，随后根据连锁不平衡对其进行剪枝。排除了名称缺失或超过 20% 样本中数据缺失的代谢物，最终得到 $S = 70{,}140$ 个单核苷酸多态性和 $M = 765$ 种代谢物。请注意，尽管部分代谢物存在缺失数据，但这种缺失可能无关紧要，因为超过 80% 的代谢物缺失数据不足 5%。图2(a) 概述了代谢物划分为超通路和亚通路的层次划分。

6.1 间接效应

我们首先估计了潜在因子 K 和间接效应的数量。表1(a)给出了各方法对 K的估计值，其中我们的方法 dBEMA 估计为 36。与 dBEMA 相比，特征值差分法、PANICr 和 BCV 的表现与模拟中的情况相同：特征值差分法可能低估，而^ PANICr 和 BCV 可能高估 K。随后，我们使用算法 1，并以 $K = 36$ 个潜在因子为条件，依据第4.1节的先验分布，确定了 $\boldsymbol{G}$ 和 $\boldsymbol{L}$的后验分布；并使用式 (5.2) 中定义的局部错误符号率（lfsr），确定参与间接效应的单核苷酸多态性和代谢物，其中 lfsr($\boldsymbol{L}_{mk}$) 的定义是在式 (5.2) 中以 $\boldsymbol{L}_{mk}$ 替换 $\boldsymbol{G}_{sk}$。我们将满足以下条件的因子定义为“遗传因子”：对于至少一个单核苷酸多态性 $s \in [S]$，其 lfsr($\boldsymbol{G}_{sk}$) ≤ 0.05；将满足以下条件的单核苷酸多态性定义为“间接单核苷酸多态性”：对于至少一个 $k \in [K]$，其 lfsr($\boldsymbol{G}_{sk}$) ≤ 0.05 ^；将满足以下条件的代谢物定义为“间接代谢物”：对于至少一个遗传因子 k，其 lfsr($\boldsymbol{L}_{mk}$) ≤ 0.05。表1(b) 给出了结果，其中除因子 1、2、3 和 6 外，其余均为遗传因子。由于这些因子的载荷最大，这表明较大的普遍因子反映的是无趣的变异，也进一步凸显了非普遍因子的重要性。

表1(b)表明，所有代谢物的水平都受少数单核苷酸多态性调控，这说明这些单核苷酸多态性具有很强的多效性。Karjalainen等（2022）的研究也支持了我们的结果。他们研究了在136,016个个体中测得的233种血清代谢物，发现所有代谢物都由276个基因组区域调控。值得注意的是，这些区域包含了我们所识别的306个（73%）

间接单核苷酸多态性，这一数量是仅凭偶然因素预期数量的四倍多（p值 $< 10^{-16}$）。

由于所有代谢物似乎都受到间接遗传调控，因此仅报告间接代谢物无助于生物学解释。相反，我们利用 $\boldsymbol{L}$ 的后验分布，研究载荷在代谢超通路和亚通路中的表现，从而解释介导遗传效应的因子。如第4.1节所述， $\boldsymbol{L}$ 的后验分布会产生一种基于子通路的聚类，将生物过程联系起来以改进推断。我们通过研究前三个遗传因子来展示这种方法的效用；其估计的代谢物和单核苷酸多态性载荷绘于图5(a)最后三行中。对于每个因子 $k \in [K]$，我们使用 A2 节详述的吉布斯采样器，从 $\text{pr}[\{(\mu_{pb},\ \phi^2_{pb})\}_{p\in[P];b\in[B_p]} \mid \hat{\boldsymbol{B}},\ \boldsymbol{D}]$ 中抽样，其中参数 $(\mu_{pb},\ \phi^2_{pb})$ 在 (4.2) 中定义，并隐式依赖于因子 k。这些参数刻画了因子k在子通路 b 和超通路 p中代谢物载荷的分布，其中“簇”是共享参数的一组子通路。我们将最大后验 聚类定义为硬聚类；与其他硬聚类方法相比，我们发现这种方法得到的结果最为简约。我们通过富集（$\mathscr{E}$）和符号（$\mathscr{S}$）得分确定簇 $\mathcal{C}$ 的重要性：

$$\mathscr{E}_{\mathcal{C}} = |\mathcal{C}|^{-1}\sum_{(p,b)\in\mathcal{C}} \mathbb{E}(\mu^2_{pb}+\phi^2_{pb} \mid \hat{\boldsymbol{B}}, \boldsymbol{D}), \quad \mathscr{S}_{\mathcal{C}} = \max[2|\mathcal{C}|^{-1}\sum_{(p,b)\in\mathcal{C}} \mathbb{P}\{N(\mu_{pb},\phi^2_{pb})>0 \mid \hat{\boldsymbol{B}},\boldsymbol{D}\} \\ 2|\mathcal{C}|^{-1}\sum_{(p,b)\in\mathcal{C}} \mathbb{P}\{N(\mu_{pb},\phi^2_{pb})<0 \mid \hat{\boldsymbol{B}},\boldsymbol{D}\}] - 1.$$

富集分数是簇的载荷平方幅值的平均值。符号分数介于0和1之间；如果簇的载荷分布远离零，则该分数接近1；如果载荷分布以零为中心对称，则该分数接近0。分数较大的簇具有较大的载荷，且载荷符号一致并非零，因此是帮助推断因子功能的理想候选。

图5(b)简要概述了前三个遗传因子的聚类结果，其中饼图显示了富集分数最高的簇的成员构成。对于因子5，我们只展示一个簇，因为其第二高的富集分数和符号分数都较小。值得注意的是，我们的无监督聚类簇反映了生物学规律。因子5和因子7中的簇1分别包含所有鞘磷脂子通路和性激素的全部三个类别。在因子4和因子7中将鞘磷脂与神经酰胺归为一组是合理的，因为一种普遍存在的反应会由神经酰胺生成鞘磷脂；簇14中包含二酰甘油，很可能是因为该反应会将其作为副产物生成（Villani 等，2008）。

我们利用这些簇及其对应的显著单核苷酸多态性（图5(b)中的红点），推断基因型系统性调节代谢物水平的间接机制。

因子4：葡萄糖稳态。因子4包含四个显著的单核苷酸多态性，它们均定位于GCKR基因。该基因在肝脏中与葡萄糖激酶结合并使其失活，从而抑制肝脏糖酵解及后续的脂肪生成（Brouwers 等，2015）。由于这四个单核苷酸多态性的次要等位基因会降低GCKR与葡萄糖激酶结合的能力（Brouwers 等，2015），我们认为因子4与葡萄糖稳态有关，其特征是肝脏糖酵解和脂肪生成上调（图6）。糖酵解过度会提高酪氨酸及其衍生物的水平，其中包括微生物组来源的苯甲酸盐（Hopper 等，2020）；此外，还可能提高黄嘌呤氧化酶的活性，而该酶与黄嘌呤亚通路中的多种代谢物有关（Hellmuth 等，2016；Hernandez-Hernandez 等，2022）。肝脏脂肪生成会提高鞘磷脂和乳糖基神经酰胺代谢物的水平（Sankella 等，2017）。

因子7：肝功能。 调控因子7的54个单核苷酸多态性中，有43个可定位到图6所示的三个基因之一。SLCO1B1仅在肝脏中表达，编码负责转运性激素及其相关代谢物的摄取转运蛋白（Cho等人，2014）。我们发现的SLCO1B1基因多态性倾向于提高其活性（X. Wu等人，2018），这可能解释了血清性激素水平的下降。FADS1编码限速酶

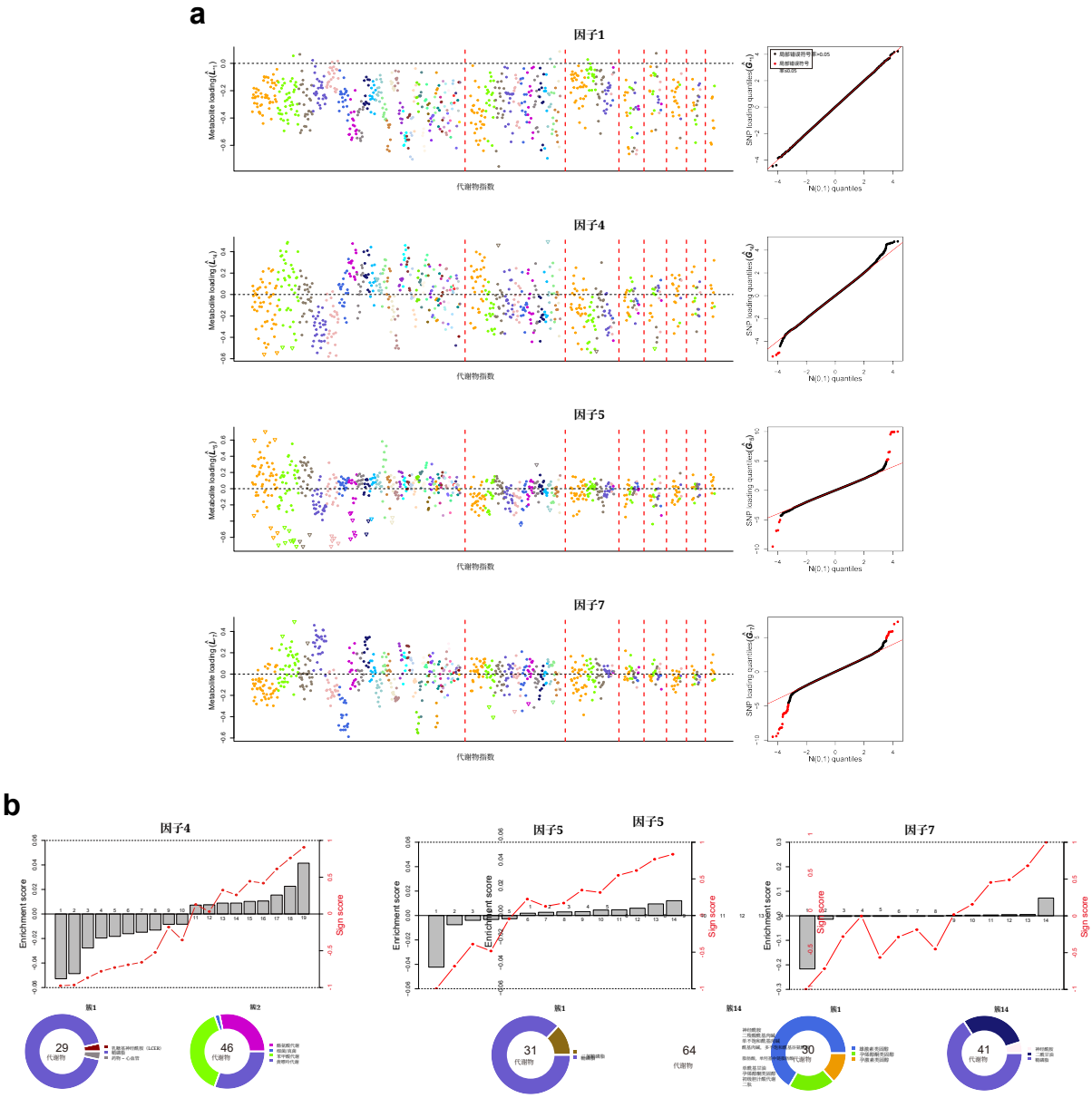

图5：(a)：因子1和前三个遗传因子的代谢物及单核苷酸多态性载荷。红色虚线标示超通路，其名称见图2(a)，每组颜色不同的代谢物代表一个子通路；“▽”标记表示该因子中的离群代谢物，定义为后验概率达到 ≥ 0.95的代谢物。(b)：三个遗传因子中所识别簇的富集和符号ˆ得分。运行吉布斯采样器前，将 $\boldsymbol{L}$ 的各列缩放至具有相同的范数，以便比较不同因子的得分。如果簇中估计载荷为负的数量多于为正的数量，则得分为负。

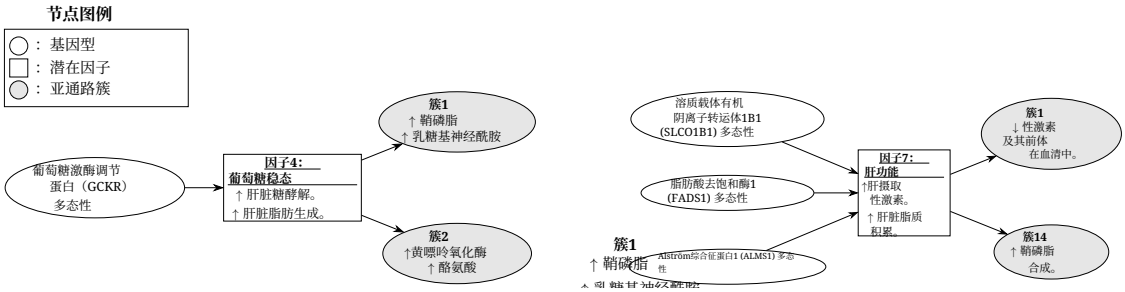

图6：因子4和因子7的解释。

a

	#唯一代谢物	#唯一 SNPs	#代谢物-单核苷酸多态性对
HiGSS 用于推断 Δ (5% 假发现率)	557	7,172	20,581
标准分析 (5% 家族错误率)	511	5,030	13,823

b

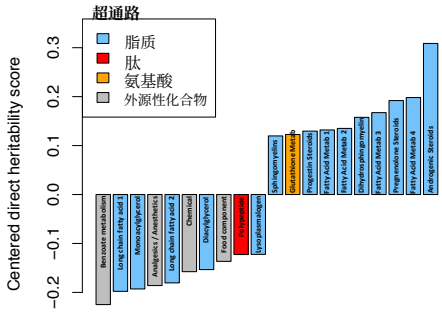

图7：关于 Δ的推断。(a)：假发现率是局部错误符号率的均值。(b)：遗传力最低和最高的10个子通路的均值中心化直接遗传力评分。

在长链多不饱和脂肪酸（长链多不饱和脂肪酸）的合成过程中，我们发现的多态性会降低其表达，从而导致肝脏脂质积累及后续的鞘磷脂合成（Gromovsky 等，2018；Simon 等，2019）。关于 ALMS1，目前所知较少，不过近期研究表明，它可能会促进非酒精性脂肪性肝病的发生（Bettini 等，2021）。

因子5：肝脏脂质代谢。因子5的19个显著单核苷酸多态性位于FADS1/2/3或SLCO1B1基因中。与因子4和因子7不同，因子5的一个特征是存在大量载荷较大且为负值的离群代谢物（图5(a)），这表明这些代谢物的功能相关，可用于帮助推断因子5的功能。值得注意的是，载荷为 < -0.4 的25个离群值中有24个是ω-3长链多不饱和脂肪酸，这与我们观察到的结果一致：我们发现的FADS多态性会降低肝脏中这些抗炎性长链多不饱和脂肪酸的生成（Yang 等，2021）。此外，SLCO1B1的多态性可能是非酒精性脂肪性肝病的标志物（Xiang 等，2009；Puri 等，2017）。

6.2 直接效应

最后，我们使用算法 1 和 4.2 中定义的通路引导先验来推断 $\boldsymbol{\Delta}$。图 7(a) 概述了我们的结果。我们首先将 HiGSS 的结果与标准 mtGWAS 分析的结果进行比较。标准分析使用 $\boldsymbol{B}$ 及其标准误来检验$H_{0,\ sm}$： $\boldsymbol{B}_{sm} = \boldsymbol{G}_{s*}^{\top}\boldsymbol{L}_{m*} + \boldsymbol{\Delta}_{sm} = 0$，随后通过 Bonferroni 校正控制家族错误率。我们识别出的代谢物-SNP配对比标准分析多 49%，这一结果相当显著，因为我们仅推断直接效应，而标准分析会拒绝具有显著直接效应或间接效应的配对。

尽管这些重新 尽管结果凸显了我们方法的统计功效，但要对其进行解释仍然颇具挑战 e

鉴于显著关联的数量，我们因此考虑在通路层面开展更易于解释的分析，以确定直接效应最大的子通路。为此，我们采用了式（4.5）中定义的子通路直接遗传力评分；评分越高，表示该子通路中的代谢物遗传力越高。图 7(b) 展示了遗传力最低和最高的各 10 个子通路的结果。不出所料，通常包含外源性代谢物的外源物质子通路遗传力最低。遗传力最高的子通路是“雄激素类固醇”，这与双生子研究的现有结果一致；这些研究表明，循环雄激素的遗传力可能高达 50%（Jin 等，2012）。

7 结论

在本研究中，我们开发了一种模型、一个理论框架和一套方法，利用汇总统计量对高维表型进行全基因组关联研究；这些方法能够明确建模多效性、实现快速计算，并便于使用生物学信息先验。鉴于代谢物全基因组关联研究的重要性，我们投入了大量精力，为间接和直接遗传效应开发层次狄利克雷先验，以便在代谢通路层面进行可解释的推断。

第6节中的真实数据分析展示了我们方法的实用性：我们使用间接效应的层次狄利克雷过程先验，通过将代谢物子通路聚类为具有生物学意义的组，使其相互关联。我们证明，这些分组有助于解释因子，从而为基因型影响代谢物水平提供可能的作用机制。使用我们的方法所得结果还表明，所有观测到的代谢物都受到遗传调控；这一发现与一项样本量超过我们数据示例20倍的研究结果一致。标准分析无法得出这些结论，它只能识别出67%代谢物的遗传关联，这凸显了我们方法的统计功效。

参考文献

[1] E. Uffelmann, Q. Q. Huang, N. S. Munung, J.de Vries, Y. Okada, A. R. Martin, H. C. Martin, T. Lappalainen, and D. Posthuma. “全基因组关联研究”。载于：*Nature ReviewsMethods Primers* 1.1 (2021年), p. 59.[2] Y. Tanigawa et al. “英国生物样本库 2,138 种表型的遗传关联组成揭示脂肪细胞生物学”。载于：*Nature Communications* 10.1 (2019年), p. 4064.[3] L. T. Elliott, K. Sharp, F. Alfaro-Almagro, S. Shi, K. L. Miller, G. Douaud, J. Marchini, and S. M. Smith. “英国生物样本库脑影像表型的全基因组关联研究”。载于：*Nature* 562.7726 (2018年), pp. 210–216.[4] X. Yin et al. “芬兰男性代谢物的全基因组关联研究发现与疾病相关的位点”。载于：*Nature Communications* 13.1 (2022年), p. 1644.[5] A. Kurilshikov et al. “大规模关联分析发现影响人类肠道微生物组组成的宿主因素”。载于：《自然遗传学》53.2 (2021年), pp. 156–165.[6] X. Yin et al. “整合转录组学、代谢组学和全基因组关联研究，有助于揭示代谢物水平和疾病风险背后的分子机制”。载于：*The American Journal of HumanGenetics* 109.10 (2022年10月), pp. 1727–1741. doi: 10.1016/j.ajhg.2022.08.007.[7] M. K. Karjalainen et al. “循环代谢生物标志物的全基因组特征分析揭示了显著的多效性和新的疾病通路”。载于：(2022年10月). doi: 10.1101/ 2022.10.20.22281089. eprint: medRXiv:10.1101/2022.10.20.22281089.

[8] C. McKennan, C. Ober, and D. Nicolae. “非随机缺失数据和潜在因子条件下代谢组学的估计与推断”。载于：《应用统计年鉴》14.2 (2020年6月). doi: 10.1214/20-aoas1328.[9] G. Qi and N. Chatterjee. “遗传力信息功效优化（HIPO）可增强多种性状遗传关联的检出能力”。载于：《*PLoS*遗传学》14.10 (2018年), e1007549.[10] M. Dai, X. Wan, H. Peng, Y. Wang, Y. Liu, J. Liu, Z. Xu, and C. Yang. “利用多效性联合分析个体水平和汇总水平的全基因组关联研究数据”。载于：《生物信息学》35.10 (2019年), pp. 1729–1736.[11] J. Liu, X. Wan, C. Wang, C. Yang, X. Zhou, and C. Yang. “LLR：一种用于在多个全基因组关联研究中共定位遗传风险变异的潜在低秩方法”。载于：《生物信息学》33 (2017年), pp. 3878– 3886.[12] Y. Lu, M. Oliva, B. Pierce, J. Liu, and L. Chen. “整合跨组学和跨情境分析，阐明复杂性状遗传效应背后的分子联系”。载于：(2022年). doi: 10.21203/rs.3.rs-2137179/v1. eprint: ResearchSquare.[13] C. Wu. “英国生物样本库脑影像表型的多性状全基因组分析”。载于：《遗传学》215.4 (Aug. 2020), pp. 947–958. doi: 10.1534/genetics.120.303242.[14] M. Wang, X. Cao, S. Zhang, and Q. Sha. “基于全基因组关联研究汇总统计量的多表型关联研究聚类线性组合方法”。载于：《科学报告》13.1 (2023年), p. 3389.[15] S. Petti, G. Reddy, and M. M. Desai. “推断基因型-表型图谱中的稀疏结构”。载于：*bioRxiv* (2022年). doi: 10.1101/2022.09.27.509675. eprint: https://www.biorxiv.org/content/early/2022/09/27/2022.09.27.509675.full.pdf.[16] A. B. Hall, A. C. Tolonen, and R. J. Xavier. “疾病中的人类遗传变异与肠道微生物组”。载于：《自然评论：遗传学》18.11 (2017年), pp. 690–699.[17] L. Zhang, Y. Shi, R. R. Jenq, K.-A. Do, and C. B. Peterson. “采用结构化先验进行微生物组特征选择的贝叶斯组成回归”。载于：《生物统计学》77.3 (2021年), pp. 824– 838.[18] P. Sidén, A. Eklund, D. Bolin, and M. Villani. “采用空间三维先验的快速贝叶斯全脑功能磁共振成像分析”。载于：《神经影像》146 (2017年), pp. 211–225.[19] K. Sankaran and S. P. Holmes. “微生物组的潜变量建模”。载于：《生物统计学》20.4 (2018年6月), pp. 599–614. issn: 1465-4644. doi: 10.1093/biostatistics/kxy018.[20] Y. Zhang, G. Qi, J.-H. Park, and N. Chatterjee. “利用全基因组关联研究汇总水平统计量估计32种复杂性状的复杂效应量分布”。载于：《自然遗传学》50.9 (2018年), pp. 1318–1326.[21] X. Zhou, P. Carbonetto, and M. Stephens. “采用贝叶斯稀疏线性混合模型进行多基因建模”。载于：《*PLoS*遗传学》9.2 (Feb. 2013), e1003264–.[22] X. Zhu and M. Stephens. “利用全基因组关联研究汇总统计量进行贝叶斯大规模多元回归”。载于：《应用统计年鉴》11.3 (2017年), p. 1561.[23] H. Ruffieux, A. C. Davison, J. Hager, andI. Irincheeva. “多结局遗传关联研究的高效推断”。载于：《生物统计学》18.4 (Mar. 2017), pp. 618–636. issn: 1465-4644. doi: 10.1093/biostatistics/kxx007.

[24] P. Turley 等人。“使用 MTAG 对全基因组关联汇总统计量进行多性状分析”。载于：《自然遗传学》50.2（2018年），第 229–237 页。[25] A. Majumdar、T. Haldar、S. Bhattacharya 和 J. S. Witte。“研究跨表型遗传关联的高效贝叶斯荟萃分析方法”。载于：《*PLoS*遗传学》14.2（2018年2月），e1007139–。[26] M. Stephens。“假发现率：新方案”。载于：《生物统计学》18.2（2017年），第 275–294 页。[27] A. Auton 等人。“人类遗传变异的全球参考”。载于：*Nature* 526.7571（2015年），第 68–74 页。[28] C. McKennan 和 D. Nicolae。“在高维生物数据中处理可估性程度不同的未观测协变量”。载于：《生物计量学》106.4（2019年9月），第 823–840 页。issn：0006-3444。doi：10.1093/biomet/asz037。[29] X. Zhong、C. Su 和 Z. Fan。“高维经验贝叶斯主成分分析”。载于：《英国皇家统计学会会刊：*B*辑（统计方法学）》84（2020年），第 853–878 页。[30] W. Wang 和 M. Stephens。“经验贝叶斯矩阵分解”。载于：《机器学习研究杂志》22.120（2021年），第 1–40 页。网址：http://jmlr.org/papers/v22/20-589.html。[31] A. B. Owen 和 J. Wang。“因子分析的双向交叉验证”。载于：《统计科学》31.1（2016年），第 119–139 页。[32] E. Dobriban 和 A. B. Owen。“确定性平行分析：一种改进的因子和主成分选择方法”。载于：《英国皇家统计学会会刊：*B*辑（统计方法学）》81.1（2018年11月），第 163–183 页。doi：10.1111/rssb.12301。[33] E. Dobriban。“因子分析和主成分分析的置换方法”。载于：《统计年鉴》48.5（2020年10月）。doi：10.1214/19-aos1907。[34] Z. T. Ke、Y. Ma 和 X. Lin。“通过主体部分特征值匹配分析估计协方差矩阵中的尖峰特征值数量”。载于：《美国统计学会会刊》（2021年），第 1–19 页。[35] C. McKennan 和 D. Nicolae。“高维相关数据中未观测协变量的估计与处理”。载于：《美国统计学会会刊》117.537（2022年），第 225–236 页。[36] J. Bai 和 S. Ng。“确定近似因子模型中的因子数量”。载于：《计量经济学》70.1（2002年），第 191–221 页。[37] S. C. Ahn 和 A. R. Horenstein。“因子数量的特征值比率检验”。载于：《计量经济学》81.3（2013年5月），第 1203–1227 页。[38] A. Onatski。“根据特征值经验分布确定因子数量”。载于：《经济学与统计学评论》92.4（2010年），第 1004–1016 页。[39] C. McKennan。具有相关观测的高维生物数据因子分析。2020年。电子版预印本：arXiv:2009.11134。[40] V. A. Marčenko 和 L. A. Pastur。“若干随机矩阵集合的特征值分布”。载于：《苏联数学文集》1.4（1967年4月），第 457–483 页。doi：10.1070/sm1967v001n04abeh001994。[41] J. Fan、K. Wang、Y. Zhong 和 Z. Zhu。“稳健高维因子模型及其在统计机器学习中的应用”。载于：《统计科学》36.2（2021年5月）。doi：10.1214/20-sts785。

[42] J. Wang、Q. Zhao、T. Hastie 和 A. B. Owen。“多重假设检验中的混杂因素调整”。载于：《统计年鉴》45.5（2017年），第 1863–1894 页。[43] J.-H. Park、M. H. Gail、C. R. Weinberg、R. J. Carroll、C. C. Chung、Z. Wang、S. J. Chanock、J. F. Fraumeni 和 N. Chatterjee。“常见遗传易感变异的等位基因频率与效应大小的分布及其相互关系”。载于：《美国国家科学院院刊》108.44（2011年），第 18026–18031 页。[44] Y. W. Teh、M. I. Jordan、M. J. Beal 和 D. M. Blei。“层次狄利克雷过程”。载于：《美国统计学会会刊》101.476（2006年12月），第 1566–1581 页。[45] C. M. Carvalho、N. G. Polson 和 J. G. Scott。“通过马蹄先验处理稀疏性”。载于：人工智能与统计第十二届国际会议论文集。第 5 卷。《机器学习研究论文集》。2009年4月，第 73–80 页。[46] J. T. Leek 和 J. D. Storey。“通过替代变量分析捕捉基因表达研究中的异质性”。载于：《*PLoS*遗传学》3.9（2007年9月），e161–。[47] Y. Benjamini 和 Y. Hochberg。“控制错误发现率：一种实用且有效的多重检验方法”。载于：《英国皇家统计学会会刊：*B*辑》57.1（1995年1月），第 289–300 页。doi：10.1111/j.2517-6161.1995.tb02031.x。[48] M. Villani、M. Subathra、Y.-B. Im、Y. Choi、P. Signorelli、M. Del Poeta 和 C. Luberto。“鞘脂合酶调节高尔基体中的二酰甘油生成”。载于：《生物化学杂志》414.1（2008年7月），第 31–41 页。doi：10.1042/bj20071240。[49] M. C. G. J. Brouwers、C. Jacobs、A. Bast、C. D. A. Stehouwer 和 N. C. Schaper。“葡萄糖激酶调节蛋白的调节：一把双刃剑？”载于：《分子医学趋势》21.10（2015年），第 583–594 页。[50] C. P. Hopper、L. K. De La Cruz、K. V. Lyles、L. K. Wareham、J. A. Gilbert、Z. Eichenbaum、M. Magierowski、R. K. Poole、J. Wollborn 和 B. Wang。“ 一氧化碳在宿主 — 肠道微生物组交流中的作用 ”。载于：《化学评论》 120.24 （2020年12月），第 13273–13311 页。[51] C. Hellmuth、F. F. Kirchberg、N. Lass、U. Harder、W. Peissner、B. Koletzko 和 T. Reinehr。“在肥胖儿童的纵向代谢组学分析中，酪氨酸与胰岛素抵抗相关”。载于：《糖尿病研究杂志》2016（2016年），第 1–10 页。doi：10.1155/2016/2108909。[52] M. E. Hernandez-Hernandez、E. Torres-Rasgado、P. Pulido-Perez、L. Nicolás-Toledo、M. Martínez-Gómez、J. Rodríguez-Antolín、R. Pérez-Fuentes 和 J. R. Romero。“葡萄糖水平紊乱与超重2型糖尿病女性的黄嘌呤氧化酶活性相关”。载于：《国际分子科学杂志》23.19（2022年9月），第 11177 页。doi：10.3390/ijms231911177。[53] S. Sankella、A. Garg 和 A. K. Agarwal。“脂肪营养不良 Agpat2-/- 小鼠肝脏中的鞘脂通路激活”。载于：《内分泌学会杂志》1.7（2017年5月），第 980–993 页。doi：10.1210/js.2017-00157。[54] E. Cho、R. B. Montgomery 和 E. A. Mostaghel。“迷你综述：SLCO和ABC转运蛋白：类固醇转运在前列腺癌进展中的作用”。载于：《内分泌学》155.11（2014年11月），第 4124–4132 页。issn：0013-7227。doi：10.1210/en.2014-1337。[55] X. Wu等人。“SLCO1B1基因多态性与高脂血症患者的肝功能、基线脂质水平及辛伐他汀降脂反应之间的关联”。载于：《临床与应用血栓形成/止血学》24。9_增刊（2018年10月），240S–247S。doi：10.1177/1076029618805863。

[56] A. D. Gromovsky 等：《Δ-5 脂肪酸去饱和酶 FADS1 通过平衡促炎与促炎症消退脂质介质影响代谢性疾病》。载于：动脉硬化、血栓形成与血管生物学 38.1（2018年10月），第 218–231 页。[57] J. Simon、A. Ouro、L. Ala-Ibanibo、N. Presa、T. C. Delgado 和 M. L. Martínez-Chantar：《非酒精性脂肪性肝病与肝细胞癌中的鞘脂：神经酰胺周转》。载于：国际分子科学杂志 21.1（2019年12月），第 40 页。doi：10.3390/ijms21010040。[58] S. Bettini 等：《阿尔斯特伦综合征中的肝纤维化与脂肪变性：代谢综合征的遗传模型》。载于：诊断学 11.5（2021年4月），第797页。doi：10.3390/diagnostics11050797

[59] C. Yang 等人：《美洲印第安人祖源与 FADS 遗传变异对西班牙裔人群 Omega-3 缺乏和心脏代谢性状的影响》。发表于《通讯生物学》4.1（2021年），第918页。[60] X. Xiang、Y. Han、M. Neuvonen、M. K. Pasanen、A. Kalliokoski、J. T. Backman、J. Laitila、P. J. Neuvonen 和 M. Niemi：《SLCO1B1 基因多态性对人体胆汁酸及胆汁酸合成标志物血浆浓度的影响》。发表于《药物遗传学与基因组学》19.6（2009年6月），第447–457页。doi：10.1097/fpc.0b013e32832bcf7b。网址：https://doi.org/10. 1097/fpc.0b013e32832bcf7b。[61] P. Puri 等人：《非酒精性脂肪性肝炎的存在及严重程度与循环胆汁酸的特定变化相关》。发表于肝脏病学67.2（2017年12月），第534–548页。doi：10.1002/hep.29359。网址：https://doi.org/10.1002/hep.29359。[62] G. Jin 等人：《全基因组关联研究发现 10q21 上一个可能影响男性血清雄激素水平的新位点 JMJD1C》。发表于《人类分子遗传学》21.23（2012年8月），第5222–5228页。doi：10.1093/hmg/dds361。

《使用汇总统计量进行高维表型全基因组关联研究的统计框架，并应用于代谢物全基因组关联研究》的附录

A1 dBEMA算法

此处我们介绍算法 dBEMA，该算法已在第3.3节中描述，与 Ke 等人（2021）的算法 2ˆ 相对应。数据矩阵输入 $\boldsymbol{B}$ 的定义见（2.2）。为与 Ke 等人（2021）保持一致，我们将伽马分布参数化为 Gamma(θ, θ/ϕ) $\mathrm{d}=\phi\times$ Gamma(θ, θ)。

假设 $S=S(N)$、$M=M(N)$、$S/N\to\gamma_S\in(0,\infty)$ 和 $M/N\to\gamma_M\in(0,\infty)$，并令 F_R和$F_{\sigma^2,N}$ 如第3.3节所定义。下文中，令 $F_{\theta,\phi,N}$ 表示 $N^{-1}\tilde{\boldsymbol{E}}^\top\tilde{\boldsymbol{E}}$ 的非零特征值的经验分布，并假设 $F_{\sigma^2,N}$ 的 M 个原子服从 Gamma($\theta,\theta/\phi$) 分布。注意，如果假设 $F_{\sigma^2,N}$ 的原子服从截断伽马分布，那么 $F_{\theta,\phi,N}$ 将收敛到一个极限分布，其斯蒂尔切斯变换由 Onatski（2010）的引理1给出。这与 Ke 等人（2021）所作的假设相同。

算法 A1（dBEMA）输入数据：标准化全基因组关联研究估计值 $\hat{\boldsymbol{B}}\in\mathbb{R}^{S\times M}$，样本量 N，参数 $\alpha,\beta\in(0,1)$，网格 $0<\theta_1<\cdots<\theta_G$，以及 $B\in\mathbb{Z}_{>0}$。输出：对 K, K的估计。

1. 定义 $\tilde{N}=\min(N,S,M)$，并令$\hat{s}_1^2\geq\cdots\geq\hat{s}_{\tilde{N}}^2$为 $N^{-1}\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}$ 的非零特征值。按以下步骤估计 θ 和 ϕ：

(i) 对每个 $k\in\{\lfloor(\alpha/2)N\rfloor,\ldots,\lfloor(1-\alpha/2)N\rfloor\}$ 和 $j\in[G]$，确定 q_{kj}，即 $F_{\theta_j,1,N}$ 的 (k/N) 分位数估计值。*(ii)* 对所有 $j\in[G]$，令 $\hat{\phi}_j=(\sum_k q_{kj}\hat{s}_k^2)/(\sum_k q_{kj}^2)$。*(iii)* 定义 $j^*=\operatorname{argmin}_{j\in[G]}\sum_k(\hat{s}_k^2-\hat{\phi}_j q_{kj})^2$ 和 $\hat{\theta}=\theta_{j^*}$, $\hat{\phi}=\hat{\phi}_{j^*}$。

2. 按照以下步骤估计 $N^{-1}\tilde{\boldsymbol{E}}^\top\tilde{\boldsymbol{E}}$ 的最大特征值分布：

(i) 设 $\tilde{\boldsymbol{E}}^{(b)}\sim MN(0,\tilde{\boldsymbol{\Sigma}}_1,\tilde{\boldsymbol{\Sigma}}_2)$，其中 $\tilde{\boldsymbol{\Sigma}}_1\in\mathbb{R}^{S\times S}$ 和 $\tilde{\boldsymbol{\Sigma}}_2\in\mathbb{R}^{M\times M}$ 是对角矩阵$\hat{}$ $\hat{}$ $\hat{}$，其元素分别从 F_R 和 $Gamma(\theta,\theta/\phi)$ 中抽样。*(ii)* 设 $\tilde{\lambda}_1^{(b)}$ 为 $N^{-1}\{\tilde{\boldsymbol{E}}^{(b)}\}^\top\tilde{\boldsymbol{E}}^{(b)}$ 的最大特征值。*(iii)* 重复步骤 *(i)-(ii)*，得到 $b=1,\ldots,B$。

3. 设$\hat{K}$ 为 $N^{-1}\hat{\boldsymbol{B}}^\top\hat{\boldsymbol{B}}$ 中超过 $\{\tilde{\lambda}_1^{(b)}\}_{b=1}^B$ 的 $1-\beta$ 分位数的特征值个数。

在所有模拟和我们的数据应用中，我们令 $\alpha=0.4$、$\beta=0.1$和 $B=500$。我们没有指定网格 $\theta_1,\ldots,\theta_G$，而是使用线搜索估计 θ（随后估计 ϕ），搜索区间的下限和上限分别为 0.1 和 5。最后，我们使用 Ke 等人（2021）算法 3 的一个修改版本来确定分位数 q_{kj} ˆ ˆ ˆ：运行步骤 2，但将 Gamma($\theta,\theta/\phi$) 替换为 Gamma(θ_j,θ_j)，并将 $\tilde{\lambda}_1^{(b)}$ 替换为 $\tilde{\lambda}_k^{(b)}$，即 $N^{-1}\{\tilde{\boldsymbol{E}}^{(b)}\}^\top\tilde{\boldsymbol{E}}^{(b)}$的第 k 大特征值。我们定义了 $q_{kj}=B^{-1}\sum_b\tilde{\lambda}_k^{(b)}$。

使用中国餐馆加盟过程的 A2 Gibbs 抽样

为了对模型（4.2）中的参数进行推断，我们使用基于中国餐馆加盟过程（CRF）的吉布斯采样器。我们从 CRF 过程的一个类比开始。每一对参数 (μ, ϕ^2) 对应一道菜肴，超通路 p 对应一家餐馆，子通路 $b \in [B_p]$ 对应一位顾客。这里，每位顾客都是一组代谢物；除少数离群值外，它们通常对同一餐桌上的菜肴表现出相同的偏好。我们用 n_{pt} 表示餐馆 p 中坐在餐桌 t 的顾客人数，用 m_{pk} 表示餐馆 p 中供应菜肴 k 的餐桌数量。令 $n_{p\cdot} = \sum_t n_{pt}$、$m_{\cdot k} = \sum_p m_{pk}$ 和 $m_{\cdot\cdot} = \sum_k m_{\cdot k}$。我们用 t_{pb} 表示超通路 p 中子通路 b 所坐的餐桌。k_{pt} 表示餐馆 p 中餐桌 t 所供应菜肴的索引。

第4.1节所述的基模型为

$$H = \pi_{0,0}\delta_{(0,0)} + (1 - \pi_{0,0})\,\mathrm{pr}(\mu) \times \mathrm{dIG}(\alpha, \beta) \tag{A2.1}$$

其中 $\pi_{0,\ 0}$ 表示子通路的载荷恰好为 0 的先验概率。pr(μ)表示均值的先验分布，它是尖峰与板块先验的混合

$$\mathrm{pr}(\mu) = \pi_0\delta_0 + \sum_{v=1}^{V} \pi_v N(\mu; 0, \lambda_v^2)$$

其中，非零的 π_0 允许将子通路的均值向 0 收缩。厚尾部分近似于马蹄先验，是多个正态分布的有限混合。标准差 $\{\lambda_v\}$ 取半柯西分布 C^+(0，1) 的 0.2、0.4、0.6、0.8 和 0.9 分位数，其对应权重 $\{\pi_v\}$ 为 $\{\lambda_v\}$的标准化并缩放后的密度，

$$\pi'_v = C^+(\lambda_v; 0, 1), \quad \pi_v = \frac{(1 - \pi_0)\pi'_v}{\sum \pi'_v}$$

dIG(α，β)表示对 ϕ^2 采用的离散化逆伽马先验，其形状参数为 α，尺度参数为 β，

$$\mathrm{dIG}(\alpha, \beta) = \sum_{w=1}^{W} p_w \delta_{\sigma_w^2}$$

其中，$\{\sigma_w^2\}$ 由将逆伽马（α，β）分布的分位数函数应用于数值 $\{1/51, 2/51, ..., 50/51\}$ 得到，并将其标准化密度用作权重 p_w。

对于每个因子 $k = 1, \ldots, K$，在从基分布为 H、浓度参数为 γ 和 α_0的层次狄利克雷过程模型中抽取子通路 b的均值和方差（μ, ϕ^2）后，$\boldsymbol{L}$ 的相应元素按以下方式生成：

$$L_{pbm,k} \sim \pi_{out} U[a_u, b_u] + (1 - \pi_{out})\mathcal{N}(\mu, \phi^2),\ m \in [M_{pb}]$$

$\hat{}\,\boldsymbol{L}$ 的估计值 $\boldsymbol{L}$ 服从

$$\hat{\boldsymbol{L}}_{pbm,k} | \boldsymbol{L}_{pbm,k} \sim \mathcal{N}(L_{pbm,k}, s_{mk}^2)$$

其中，s_{mk}^2 是估计量 $\hat{}\boldsymbol{L}_{pbm,\ k}$ 给定$\boldsymbol{L}_{pbm,\ k}$时的方差。这里我们采用了与算法 1 不同的 $\hat{}\boldsymbol{L}$ 缩放方式；算法 1 中采用的是 $\hat{}\boldsymbol{L} = \sqrt{M}\boldsymbol{V}$，其中 M 为代谢物数量，$\boldsymbol{V}$为算法 1 中定义的前$\hat{}$ K 个右奇异向量。采用新的缩放方式后，$\boldsymbol{L}$ 的每一列的范数均为 $\sqrt{M}$。以这种方式缩放 $\hat{}\boldsymbol{L}$，可以避免根据各因子的强度调整基模型中的超参数

ˆ，之后我们可以重新缩放 $\boldsymbol{L}$ 以及使用吉布斯采样器学习得到的参数。根据定理 3.2 (iii)，当 ˆ $\boldsymbol{L} = \sqrt{M}\boldsymbol{V}$时，有

$$s_{mk}^2 = \frac{M}{\Gamma_{kk}^2}(\frac{1}{N} + \frac{1}{S})\hat{\sigma}_m^2$$

我们对$\pi_{0,0}$ 和 π_0均设置 Beta(1,1) 先验分布，并对 π_{out}设置 Beta(1/5,999/5) 先验分布。后者反映了我们的先验信念：平均每 1,000 个代谢物中有 1 个离群值，且离群值不超过 5 个的概率为 0.95。按照（Teh 等，2006），我们对浓度参数 γ 和 α_0设置 Gamma(1,1) 先验分布。但与（Teh 等，2006）不同，我们对 γ 和 α_0 的伽马先验进行离散化，以避免耗时的辅助变量抽样。

在我们的实现中，我们首先为每个子通路中的非离群代谢物抽样桌索引 t_{pb}，然后更新各餐桌的碟索引 $\{k_{pt}\}$，最后更新参数 $\{(\mu, \phi^2)\}$。其他相关参数的更新过程将在下文中说明。为简化符号，省略因子索引 k，因为吉布斯更新将独立并行地应用于 ˆ$\boldsymbol{L}$ 的每一列。下文将用索引 k 表示参数 θ_k。我们用 $\theta = (\mu, \phi^2)$表示从层次狄利克雷过程模型（4.2）中抽样得到的一对参数。因此，对于子通路 b，其最终使用的参数为 $\theta_{kt_{pb}} = (\mu_{kt_{pb}}, \phi^2_{kt_{pb}})$。

更新离群值。对于每个子通路 b，我们首先更新离群值，因为离群值并非由层次狄利克雷过程模型生成，只有非离群条目会纳入下一轮吉布斯更新。给定 ˆ$\boldsymbol{L}$ 和其他参数时，子通路 b 中代谢物 m 为离群值的后验概率，可通过对 $\boldsymbol{L}$ 积分得到，

$$\text{pr}(m \text{ is an outlier}|\cdot) \propto \frac{\pi_{out}}{b_{out} - a_{out}} \left[\Phi(\frac{b_{out} - \hat{\boldsymbol{L}}_{pbm}}{s_m}) - \Phi(\frac{a_{out} - \hat{\boldsymbol{L}}_{pbm}}{s_m})\right]$$

其中 Φ 是标准正态分布的累积分布函数。均匀分布的参数 ˆa_{out} 和 b_{out} 应覆盖 $\boldsymbol{L}_k$ 的范围。在我们的实现中，a_{out} 和 b_{out} 设置为以 ˆ 零为中心，并覆盖 $\boldsymbol{L}_k$ 范围的 1.25 倍。

代谢物由层次狄利克雷过程模型生成的后验概率取决于子通路 b当前的吉布斯样本。如果 b 中包含其他非离群代谢物，且这些代谢物在上一轮吉布斯更新中已被分配到餐桌 t，那么

$$\text{pr}(m \text{ is not an outlier}|\cdot) \propto (1 - \pi_{out}) \text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_{k_{t_{pb}}}, \sigma_m^2)$$

其中

$$\text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_{k_{t_{pb}}}, s_m) = N(\hat{\boldsymbol{L}}_{pbm}; \mu_{k_{t_{pb}}}, \phi^2_{k_{t_{pb}}} + s_m^2).$$

否则，如果当前 b 中除 m 外没有其他非离群代谢物——当 b 仅包含一个代谢物，或 b 中所有其他代谢物均为离群值时，就可能出现这种情况——则需要在给定当前餐桌和菜肴分配的条件下，对 $\theta_{k_{t_{pb}}}$ 积分，

$$\begin{aligned}\text{pr}(m \text{ is not an outlier}|\cdot) \propto &(1 - \pi_{out}) \left[\sum_t \frac{n_{pt}}{n_p - 1 + \alpha_0} \text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_{k_{pt}}, s_m)\right. \\ &\left. + \frac{\alpha_0}{n_p - 1 + \alpha_0} \left(\sum_k \frac{m_{\cdot k}}{m_{\cdot\cdot} + \gamma} \text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_k, s_m) + \frac{\gamma}{m_{\cdot\cdot} + \gamma} \text{pr}(\hat{\boldsymbol{L}}_{pbm}|H, s_m)\right)\right]\end{aligned}$$

其中

$$\text{pr}(\hat{\boldsymbol{L}}_{pbm}|H, s_m) = \pi_{0,0} N(\hat{\boldsymbol{L}}_{pbm}; 0, s_m^2) + (1 - \pi_{0,0}) \iint N(\hat{\boldsymbol{L}}_{pbm}; \mu, \phi^2 + s_m^2) \text{pr}(\mu) \text{pr}(\phi^2) d\mu d\phi^2$$

上式右侧的积分很容易计算，因为我们对 pr(μ) 和 pr(ϕ^2) 进行了离散化。否则，该积分难以求解，因为由于 s_m^2的存在， $N(\hat{\boldsymbol{L}}_{pbm};\ \mu,\ \phi^2+s_m^2)$ 与逆伽马先验不共轭。对ϕ^2 采用这种离散化先验有助于解决非共轭问题，并避免耗时的蒙特卡罗模拟。

抽样 $\boldsymbol{t}$。 根据可交换性，我们可以将 t_{pb} 视为餐馆 p 中最后抽样的变量。若 t 非空，则给定其他参数时， $t_{pb}=t$ 的似然与坐在 t 的子通路数量成正比；若 t 是新的，则与 α_0 成正比。给定 t_{pb} 时， $\boldsymbol{L}_{pb}$的似然是该子通路中所有非离群代谢物似然的乘积。给定 $t=t_{new}$ 时， $\boldsymbol{L}_{jg}$ 的似然可通过对 k_{pt}积分消去来计算，

$$\text{pr}(t_{pb}=t|\cdot)\propto\begin{cases}n_{pt}^{-b}\prod_{m\in[M_{pb}]}\text{pr}(\hat{L}_{pbm}|\theta_{k_{pt}},s_m) & \text{if } t \text{ previously used}\\ \alpha_0\prod_{m\in[M_{pb}]}\text{pr}(\hat{L}_{pbm}|t=t_{new},\boldsymbol{\theta},\boldsymbol{k},s_m) & \text{if } t=t_{new}\end{cases}\tag{A2.2}$$

and

$$\text{pr}(\hat{L}_{pbm}|t=t_{new},\boldsymbol{\theta},\boldsymbol{k},s_m^2)=\sum_k\frac{m_{\cdot k}}{m_{\cdot\cdot}+\gamma}\prod_{m\in[M_{pb}]}\text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_k,s_m)+\frac{\gamma}{m_{\cdot\cdot}+\gamma}\prod_{m\in[M_{pb}]}\text{pr}(\hat{\boldsymbol{L}}_{pbm}|H,s_m)\tag{A2.3}$$

如果子通路 g 坐在 t_{new}，我们需要通过以数据为条件抽样 $k_{pt_{new}}$，为 t_{new} 分配一道菜，这将在下一步介绍。更新 t_{pb}后，某些餐桌可能会变为空桌；由于 $n_{pt}=0$，它们在后续迭代中再次被抽样的概率为 0。因此，每次更新 t_{pb} 后，我们都会移除空桌。由于移除餐桌以及在下一步更新 k_{pt} 都可能导致某些菜肴变为空菜肴，我们也会在更新 $\boldsymbol{t}$ 和 $\boldsymbol{k}$后删除空菜肴。

对 $\boldsymbol{k}$ 进行采样。 对 k_{pt} 进行采样的过程与对 t_{pb}进行采样类似。首先，我们假设菜肴分配具有可交换性，并将 k_{pt} 视为最后采样的变量。由于一张餐桌可以服务多个子通路， $k_{pt}=k$ 的似然将包含各子通路及其代谢物的生成，

$$\text{pr}(k_{pt}=k|\cdot)\propto\begin{cases}m_{\cdot k}\prod_{b:t_{pb}=t}\prod_{m\in[M_{pb}]}\text{pr}(\hat{\boldsymbol{L}}_{pbm}|\theta_k,s_m) & \text{if } k \text{ is previously used}\\ \gamma\prod_{g:t_{pb}=t}\prod_{m\in[M_{pb}]}\text{pr}(\hat{\boldsymbol{L}}_{pbm}|H,s_m) & \text{if } k=k_{new}\end{cases}\tag{A2.4}$$

如果 $k_t=k_{new}$，我们将从 H 中抽取一对新的参数 $\theta_{k_{new}}=(\mu_{new},\ \sigma_{new})$，其条件是给定坐在餐桌 t上的所有子通路的 $\hat{\boldsymbol{L}}_{pbm}$ 和 s_m，即 $t_{pb}=t$。由于 s_m^2导致非共轭问题，我们将对 ϕ^2 上的逆伽马先验进行离散化。详情请参阅下一步。

更新浓度参数。 更新 γ 和 α_0 不依赖于 $\boldsymbol{\theta}$的取值，而仅取决于菜肴和餐桌的数量；在对 t 和 k 进行采样后，这些数量便固定下来。因此，在更新 $\boldsymbol{\theta}$之前执行了这一步。

为更新浓度参数，我们使用了在给定 K的条件下，(Teh 等，2006) 中 γ 和 α_0 的似然，其中 K是菜肴总数， $\{T_p\}_1^P$是超通路 p中的餐桌数， $\{B_p\}_1^P$是超通路 p中的子通路数。在针对 $\boldsymbol{L}$的生成模型 (4.2) 中，浓度参数 α_0 由所有超通路共享。因此，我们有

$$\text{pr}(T_1,...T_P|\alpha_0,B_1,...,B_P)=\prod_{p=1}^P s(B_p,T_p)\alpha_0^{T_p}\frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+B_p)}$$

and

$$\text{pr}(K|\gamma,\sum_p T_p)=s(\sum_p T_p,K)\gamma^K\frac{\Gamma(\gamma)}{\Gamma(\gamma+\sum_p T_p)}$$

其中 $s(n,\ m)$ 是第一类无符号斯特林数。我们没有使用辅助变量抽样或自适应拒绝抽样来迭代抽取 γ 和 α_0，而是将 γ 和 α_0 的伽马先验离散化，并根据其后验概率从网格中抽样，以更新它们的取值。 γ和 α_0的网格均设为 $\{0.1, 0.2, ..., 0.9, 1, 2, ..., 10\}$。

更新 $\boldsymbol{\theta}$。 </style}不仅抽取新参数，更新旧参数也需要在给定代谢物载荷 $\hat{\boldsymbol{L}}_{pbm}$ 和误差方差 s_m^2的条件下，对子通路共享的均值和方差进行抽样。由于所有餐馆（超通路）共享参数 $\boldsymbol{\theta}$，我们将根据各超通路和子通路中所有消耗菜肴 k 的非离群代谢物，在条件下更新 θ_k。令 $\{k\}$ 表示所有消耗菜肴 k的代谢物，则有

$$\mathrm{pr}\left(\theta_k = (0,0)|\cdot\right) \propto \pi_{0,0} \prod_{m\in\{k\}} N(\hat{\boldsymbol{L}}_{pbm}; 0, s_m^2) \tag{A2.5}$$

and

$$\mathrm{pr}\left(\theta_k \neq (0,0)|\cdot\right) \propto (1-\pi_{0,0}) \iint \prod_{m\in\{k\}} N(\hat{\boldsymbol{L}}_m; \mu, \phi^2 + s_m^2) d\mathcal{P}(\mu) d\mathrm{IG}(\phi^2) \tag{A2.6}$$

如果 $\theta_k \neq (0,\ 0)$，我们将抽取指示变量

$$\mathrm{pr}\left(Z_{v,w} = 1|\cdot\right) \propto p_v p_w \prod_{m\in\{k\}} N(\hat{\boldsymbol{L}}_m; 0, \lambda_v^2 + \sigma_w^2 + s_m^2) \tag{A2.7}$$

如果 $Z_{0,\ r} = 1$，我们直接令 $\theta_k = (0,\ \sigma_w)$。否则，我们在给定 λ_v 和 $Z_{v,\ w}$的条件下，抽取 μ_k 的值，

$$\sigma_k = \sigma_w,\ \ \mu_k \sim N\left(\frac{\sum_{m\in\{k\}} \frac{\hat{\boldsymbol{L}}_m}{\sigma_w^2+s_m^2} + \frac{0}{\lambda_v^2}}{\sum_{m\in\{k\}} \frac{1}{\sigma_w^2+s_m^2} + \frac{1}{\lambda_v^2}}, \frac{1}{\sum_{m\in\{k\}} \frac{1}{\sigma_w^2+s_m^2} + \frac{1}{\lambda_v^2}}\right) \tag{A2.8}$$

更新 π_{out}，$\pi_{0,\ 0}$，π_0。 </style}、，0 </style和 π_0 的更新可通过以下方式获得

$$\pi_{out} \sim \mathrm{Beta}(\frac{1}{5} + \#outliers, \frac{999}{5} + M - \#outliers)$$

$$\pi_{0,0} \sim \mathrm{Beta}(1 + \#spikes, 1 + K - \#spikes)$$

$$\pi_0 \sim \mathrm{Beta}(1 + \#centers, 1 + K - \#spikes - \#centers)$$

其中，$\#outliers$、$\#spikes$ 和 $\#centers$ 分别表示离群值的数量、参数$\theta = (0,\ 0)$的数量，以及均值为 0 但方差非零的参数数量。 M 和 K 分别表示当前吉布斯更新轮次中的代谢物总数和参数数量。

A3 Es代谢物全基因组关联研究中直接效应的估计：设置网格点 s

这里我们介绍如何确定第4.2节中定义的 $\mathcal{X}_{(\pi)}$和 $\mathcal{X}^{(\varphi^2)}$的网格点。简而言之，我们首先通过最大似然估计 $\{\pi_{pbm}\}$ 和 φ^2，即

$$\{\{\hat{\pi}_{pbm}\}, \hat{\varphi}^2\} = \underset{\{\pi_{pbm}\},\varphi^2}{\mathrm{argmax}}\ \mathrm{pr}(\hat{\boldsymbol{\Delta}} \mid \{\pi_{pbm}\}, \varphi^2),$$

假设

$$\hat{\boldsymbol{\Delta}}_{s,pbm} \mid \boldsymbol{\Delta}_{s,pbm} \sim N(\boldsymbol{D}_{ss}^{1/2}\boldsymbol{\Delta}_{s,pbm}, \hat{\sigma}_{pbm}^2)$$

$$\boldsymbol{\Delta}_{s,pbm} \mid \pi_{pbm}, \varphi^2 \sim (1-\pi_{pbm})\delta_0 + \pi_{pbm} N(0, \varphi^2 \hat{\sigma}_{pbm}^2 / \boldsymbol{D}_{ss}).$$

接下来，我们将 $\mathcal{X}_{(\pi)}$和 $\mathcal{X}^{(\varphi^2)}$分别定义为包含 100 个元素的集合，其中 $\mathcal{X}_{(\varphi^2)} = \{\hat{}\varphi^2/2, \ldots, 2\,\hat{}\varphi^2\}$包含等间距元素。我们设定 $\mathcal{X}^{(\pi)}$，使其最小元素为 $10/S$，最大元素为 $\{\hat{}\pi_{pbm}\}$的 0.99 分位数的 1.5 倍，并且在对数尺度上等间距分布。

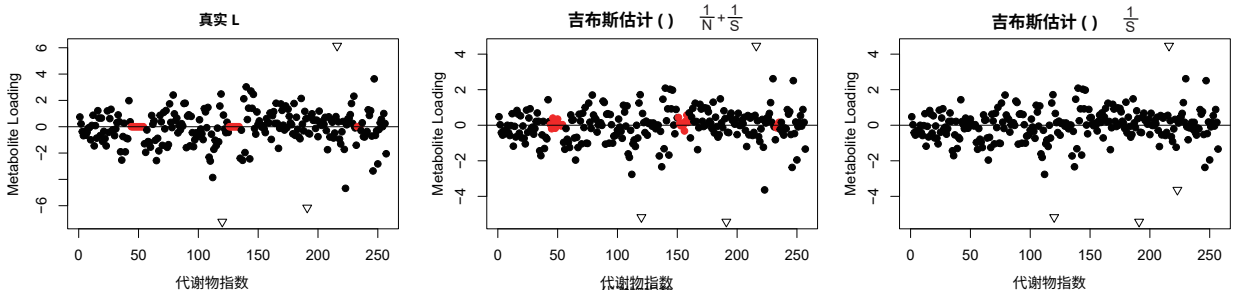

图 A1：左：使用公式 4.2 中的模型模拟的真实 $\boldsymbol{L}$ 。尖峰通路以红色标记。离群值以倒三角形标记。中：使用 $\boldsymbol{L}$ 的正确渐近方差ˆ对尖峰和离群值进行估计。估计的尖峰通路以红色标记。估计的离群值以倒三角形标记。右：通过忽略单核苷酸多态性之间的相关性得到的 $\boldsymbol{L}$ 的不正确方差ˆ下，对尖峰和离群值进行估计。估计的尖峰通路以红色标记。估计的离群值以倒三角形标记。

A4 L 的实验模拟

为了模拟包含 K 个因子的 $\boldsymbol{L}$ ，需要首先指定代谢物的超通路-子通路结构。在我们的实现中，每个超级通路的子通路数量以及各子通路的大小均从泊松分布中随机抽样。总之，我们生成了一个包含 3 个超级通路和 257 个代谢物的代谢物档案。

给定代谢物的通路结构， $\boldsymbol{L}$ 的每一列都是独立生成的。 $\boldsymbol{L}$ 的生成过程严格遵循公式 (A2.1) 中层次狄利克雷过程模型的中国餐馆特许经营构造。马蹄先验的全局收缩参数 τ 设为 0.2，离群值率 π_{out} 设为 0.01，尖峰率 $\pi_{0,\ 0}$ 设为 0.1， π_0 从 Beta(2,1) 分布中随机抽取；用于生成 ϕ^2 的逆伽马分布的尺度参数和速率参数分别设为 2 和 1。为确保离群值能够被识别，离群值条目生成在子通路均值上下 2 个标准差之外，但其数值不会超出 (-8,8)。该模拟过程使我们能够追踪哪些子通路包含尖峰、哪些代谢物是离群值，以及哪些子通路共享相同的参数。最后，将 $\boldsymbol{L}$ 的所有条目乘以 0.4，使信号强度与噪声水平相当。

最后，我们展示，A2 节中描述的吉布斯采样器可以利用定理 3.2 中推导的 $\boldsymbol{L}$ 渐近方差来识别尖峰和离群值。为此，我们模拟了一个包含 257 个代谢物、5000 个单核苷酸多态性且样本量为 500 的表型矩阵$\boldsymbol{Y}$，使比率为 $\frac{S}{N} = 10$。 $\boldsymbol{Y}$包含$K = 1$ 因子，以避免多于一个因子时出现的可识别性问题，从而简化模拟。用于模拟 $\boldsymbol{Y}$的 $\boldsymbol{L}$ 如图 A1 左侧所示，其中包含 3 个尖峰子通路（以红色标示）和 3 个离群值（以倒三角形标示）。我们将 A2 节中描述的吉布斯采样器应用于估计值 ˆ$\boldsymbol{L}$ 两次：一次假设 $\text{Var}(\hat{\boldsymbol{L}}_{mk} \mid \boldsymbol{L}) = (1/N + 1/S)\sigma_m^2$，另一次假设 $\text{Var}(\hat{\boldsymbol{L}}_{mk} \mid \boldsymbol{L}) = (1/S)\sigma_m^2$。第二种是将式 (2.2) 中的 $\boldsymbol{R}$ 视为单位矩阵时所假设的方差（McKennan 等，2019）。结果如图 A1中间和右侧所示。如果某个子通路的参数取值为 (0,0) 的后验概率超过 0.95，则将其判定为尖峰。若某个代谢物属于离群值的后验概率大于 0.95，则将其定义为离群值。可以看出，采用正确的方差时，吉布斯采样器能够正确识别尖峰和离群值；但采用另一种方差时，它未能识别任何尖峰子通路，并且高估了离群值的数量。