arXiv:2602.13888v1 [stat.ME] 14 Feb 2026

用于协方差矩阵的混合专家 Wishart 模型及其在癌症药物筛选中的应用

天脉[1] ,Zhi Zhao[2]

[1]挪威公共卫生研究所，挪威奥斯陆 0456。[2]奥斯陆大学，
挪威奥斯陆 0372。电子邮件：the.tien.mai@fhi.no

摘要

协方差矩阵自然存在于金融、基因组学和神经科学等不同科学领域，可用于编码相依结构并揭示复杂多变量系统的重要特征。在本研究中，我们提出一个全面的贝叶斯框架，通过经典混合模型和一种新颖的专家混合Wishart（MoE–Wishart）模型分析异质协方差数据。所提出的MoE–Wishart模型扩展了标准Wishart混合模型，使混合权重能够通过多项逻辑门控网络依赖于预测变量。该模型形式使模型能够捕捉协方差结构中复杂的非线性异质性，并根据协变量相关模式调整子总体隶属概率。为进行推断，我们开发了一种高效的Gibbs内Metropolis–Hastings抽样算法，专为Wishart似然和门控网络的几何结构设计。此外，我们还推导出一种期望最大化算法，用于专家混合框架下的最大似然估计。大量模拟研究表明，在一系列异质协方差情境下，所提出的贝叶斯估计量和最大似然估计量均能准确恢复子总体并进行估计。最后，我们将该方法创新性地应用于一个具有挑战性的数据集——癌症药物敏感性谱，展示了MoE–Wishart模型利用不同药物剂量及重复测量之间协方差的能力。

我们的方法已在R 软件包`moewishart`中实现，可从https://github.com/zhizuio/moewishart获取。

关键词：贝叶 叶斯推断；马尔可夫链蒙特卡洛；最大似然估计；EM 算法；协方差矩阵；混合模 el.

1 引言

在传统统计建模中，人们通常假设整个总体中的变量之间存在单一且普适的关系。偏离这一关系的观测值通常被视为离群值。然而在实践中，现实世界的系统很少是同质的。相反，数据往往来自多个潜在亚群体，每个亚群体都具有独特的行为模式或结构关系。随着现代数据集的规模和复杂性不断增加，恰当地建模这种潜在异质性变得至关重要。混合模型 [24] 为将经典聚类和回归方法推广到总体结构本质上具有异质性的情形，提供了强大且有理论依据的框架 [5]。

总的来说，协方差矩阵在科学研究中发挥着核心作用，因为它们能够编码变量之间的线性相依模式，并揭示多元现象的内在结构。其重要性涵盖广泛的应用领域。在金融领域，协方差矩阵是投资组合配置、风险评估和投资策略的基础 [30, 11]。在成像/《模式识别》领域，像素/体素强度或区域之间的协方差能够捕捉空间和时空依赖关系，支持聚类和特征提取等任务 [32]。在基因组学中，量化基因-基因相互作用可能有助于

识别具有生物学意义的性状关联 [27]。在神经科学中，协方差矩阵用于描述脑区之间的功能连接和网络层面相互作用 [21]。此外，协方差结构构成了重要多元技术的计算基础，例如主成分分析和因子分析 [26, 7]。

尽管协方差结构无处不在，但在标准聚类技术中往往被忽视。广泛使用的算法（如 k 均值和层次聚类）主要依赖均值结构的差异，从而忽略了协方差模式中蕴含的潜在丰富信息。为解决这一局限，近期研究日益关注基于协方差异质性的聚类。 [8], 提出了一个重要的概率框架，该框架采用具有 Wishart分量密度的混合模型，并使用期望最大化（EM）算法进行参数估计。后续扩展包括针对高维协方差估计的稀疏惩罚 [3], 、变分贝叶斯推断 [21], 以及在判别分析中的应用 [4]。相关地， [6] 研究了偏斜矩阵变量分布的有限混合模型，其中包括作为 Wishart 分布推广的矩阵变元方差–伽马分布，并通过 EM算法进行推断。尽管取得了这些进展，对 Wishart混合模型进行全面的完全贝叶斯处理仍相对少见。

本文针对这一空白，首先构建经典 Wishart分布混合模型的完全贝叶斯表述。在此基础上，我们提出一种新颖的专家混合Wishart（MoE–Wishart）模型，将专家混合架构的灵活性与 Wishart 分布协方差数据的可解释性和结构相结合。不同于具有固定混合比例的标准混合模型，专家混合模型 [12] 允许簇成员概率显式依赖于协变量，从而使模型能够学习潜在亚群体如何随预测变量空间而变化。通过由 softmax链接函数控制的概率门控网络 [23], ，专家混合框架划分输入空间，并将观测值动态分配给专家分量。这种结构使所提出的模型能够捕捉协方差模式中复杂的非线性分段行为，在高维或高度异质的情形下尤其具有价值。关于专家混合模型的概述，参见 [20]。

对于混合模型和所提出的 MoE–Wishart 扩展模型的后验推断，我们采用高效的马尔可夫链蒙特卡罗（MCMC）算法。具体而言，我们构建了一种针对对称正定协方差矩阵所诱导的后验分布几何特征量身定制的 Gibbs-within–Metropolis–Hastings 采样器。此外，我们还针对专家混合框架推导出一种用于最大似然估计的期望最大化算法，因为在该框架下无法直接获得闭式更新。

因此，本文的主要贡献有三点：（i）对 Wishart 分布混合模型进行完整的贝叶斯表述；（ii）提出一种用于协方差矩阵数据的新型专家混合模型（MoE–Wishart模型）；（iii）为 MoE–Wishart模型量身定制推断算法，包括 MCMC 和期望最大化程序。为便于复现和实际应用，我们还提供了一个 R包 `moewishart`，其中实现了所有提出的方法（可通过以下链接获取：
https://github.com/zhizuio/moewishart）。

我们开展了大量模拟研究，以评估所提出方法的统计准确性和计算性能。最后，我们通过一项真实数据应用展示这些方法的实际效用，研究癌症药物发现和药物再利用中产生的协方差模式。在生物医学领域，药物再定位是指为已获批药物和临床前药物寻找新的用途 [1]。随着大规模药物筛选数据和相关分子信息日益丰富，数据驱动的药物再定位有望加速临床进展。具体而言，利用基于共同分子靶点、药理学特性或化学结构的药物间相似性或簇，有助于实现药物再利用[34, 9]。药物筛选数据通常来自这样的实验：将单种药物以多种剂量/浓度施用于患者组织样本或细胞系。因此，通过测量多个剂量下重复样本的细胞活力来评估每种药物的疗效。目前，大多数模型使用标量汇总描述符（例如，IC_{50} 或药物剂量-反应曲线下面积）

来表示每种药物的疗效 [28, 19, 29, 36], ，但这种方法会损失大量信息。我们提出，对于每种药物，使用其响应数据的协方差描述符（即多个剂量下重复样本的细胞活力协方差矩阵），然后根据多种药物的协方差矩阵对潜在簇进行建模。归为同一组的药物可能具有相似的功能，并提示某些药物的新用途。

本文其余部分安排如下。第 2 节介绍用于协方差矩阵数据的 Wishart 混合模型和专家混合模型。第3节为混合模型和专家混合模型建立贝叶斯框架，并介绍相应的 MCMC 抽样算法。第4节介绍用于求解专家混合模型最大似然估计的 EM 算法。第5 节报告模拟研究，第6节介绍癌症药物筛选数据的应用。我们在第 7节进行讨论并作总结。

2 协方差矩阵混合模型

我们首先介绍混合模型，然后介绍其向专家混合模型的扩展。

2.1 混合模型

设 S_1、……、S_n 为观测到的对称正定（SPD） $p \times p$ 矩阵。我们将其建模为 Wishart 分布的有限混合：

$$S_i \mid z_i = k, \Sigma_k, \nu_k \overset{\text{ind}}{\sim} \mathcal{W}_p(\nu_k, \Sigma_k), \qquad i = 1, \ldots, n, \tag{1}$$

其中， $z_i \in \{1, \ldots, K\}$ 表示观测 i的潜在（未观测）簇标签。这里， $\nu_k > p - 1$ 是 Wishart 分布的自由度， Σ_k 是簇k特有的尺度（协方差）矩阵。Wishart 密度（采用本文的参数化方式）为

$$f_{\mathcal{W}}(S \mid \nu, \Sigma) = \frac{|S|^{(\nu-p-1)/2} \exp\left(-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}S\right)\right)}{2^{\nu p/2}|\Sigma|^{\nu/2}\,\Gamma_p\left(\frac{\nu}{2}\right)}, \tag{2}$$

其中 $\Gamma_p(\cdot)$是多元伽马函数。

对潜在簇标签 z_i进行边缘化后，每个观测矩阵 S_i 的分布可写成混合密度形式。令 $\boldsymbol{\pi} = (\pi_1$、……、$\pi_K)$表示（固定的）混合权重，其中$\pi_k = \Pr(z_i = k)$、 $\pi_k \geq 0$和 $\sum_{k=1}^{K} \pi_k = 1$。那么，协方差观测 S_i的条件密度（对潜在簇标签 z_i进行边缘化）为

$$f(S_i \mid \boldsymbol{\pi}, \{\Sigma_k, \nu_k\}_{k=1}^{K}) = \sum_{k=1}^{K} \pi_k f_{\mathcal{W}}\left(S_i \mid \nu_k, \Sigma_k\right), \qquad i = 1, \ldots, n, \tag{3}$$

其中 $f_{\mathcal{W}}(\cdot \mid \nu_k, \Sigma_k)$ 是式（2）中定义的 Wishart 密度。该表述强调，SPD 矩阵的总体分布可表示为 K Wishart 分量的凸组合，每个分量对应一个具有自身尺度矩阵和自由度的不同潜在簇。这样的 Wishart 分布有限混合为对协方差矩阵异质集合进行建模提供了连贯且易于解释的框架，同时明确遵循其几何约束和正定性约束。通过允许不同簇的尺度矩阵和自由度各不相同，该模型能够捕捉协方差结构和离散度方面的差异，从而灵活而简约地表示潜在亚群体。

2.2 混合专家 Wishart 模型

设 $S_1, \ldots, S_n$ 为观测到的、大小为 $p \times p$ 的对称正定矩阵，并设 $X \in \mathbb{R}^{n \times q}$ 为额外的协变量。我们考虑将式（1）中的混合模型扩展为包含 K 个 Wishart 专家的有限专家混合（MoE）模型（MoE–Wishart）：

$$
\begin{aligned}
S_i \mid z_i = k, \Sigma_k, \nu_k &\overset{\text{ind}}{\sim} \mathcal{W}_p(\nu_k, \Sigma_k), \\
\Pr(z_i = k \mid X_i) &= \pi_{ik}, \qquad \text{with } k = 1, \ldots, K,
\end{aligned}
$$

其中，门控概率是 X 的多项逻辑（softmax）函数 $_i$:

$$
\pi_{ik}(X_i; \beta_k) = \frac{\exp(X_i^\top \beta_k)}{\sum_{\ell=1}^{K} \exp(X_i^\top \beta_\ell)}, \qquad \beta_K \equiv 0 \text{ (for identifiability)}. \tag{4}
$$

在此专家混合模型形式下，给定协变量 X_i 时的条件密度（对潜在簇标签 z_i 求边际）为

$$
f(S_i \mid X_i, \{\beta_k, \Sigma_k, \nu_k\}_{k=1}^{K}) = \sum_{k=1}^{K} \pi_{ik}(X_i; \beta)\, f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k)\,. \tag{5}
$$

该 MoE–Wishart 模型在基准有限混合模型的基础上，将协变量信息明确纳入聚类机制，从而使簇成员资格能够随观测预测变量系统性变化，而非由固定的混合比例决定。在相关文献中，X_i 也称为伴随变量，因此，专家混合模型也称为带伴随协变量的混合模型。先前的混合模型假设观测值可交换，且先验簇概率同质；专家混合模型则引入多项逻辑门控网络，使分配概率能够根据个体特异性协变量进行调整，从而以更灵活且更易解释的方式刻画异质性。在这种设定下，Wishart 专家通过 $(\Sigma_k,\ \nu_k)$ 继续发挥建模簇特异性协方差结构和离散度的作用；同时，纳入 X_i 使模型能够捕捉由协变量驱动的潜在状态变化。这一点在结构变异部分由外部因素解释的应用中尤为重要。当 X_i 为**1**列，即截距时，专家混合模型便退化为混合模型。

MoE–Wishart模型的可识别性源于指数族专家混合的一般性结果 [10]。具体而言，只要各簇对应的 Wishart 分布互不相同，且协变量设计矩阵为满秩，模型在潜在簇标签的置换意义下可识别。约束 $\beta_K \equiv 0$解决了多项逻辑门控网络固有的平移不可识别性，从而确保在给定混合成分标签的情况下，参数估计具有唯一性。

3 贝叶斯方法与 MCMC 算法

本节介绍一种贝叶斯框架，以及专为混合模型和 MoE–Wishart 模型设计的 MCMC 抽样算法。据我们所知，这是一个新的贡献。

3.1 贝叶斯混合–Wishart 模型

我们设置 混合模型中参数的以下先验分布（3 ）:

$$
\begin{aligned}
z_i \mid \boldsymbol{\pi} &\sim \text{Categorical}(\pi_1, \ldots, \pi_K), \\
\boldsymbol{\pi} &\sim \text{Dirichlet}(\alpha_1, \ldots, \alpha_K), \\
\Sigma_k &\sim \mathcal{IW}(\nu_0, \Psi_0^{-1}), \qquad k = 1, \ldots, K,
\end{aligned}
$$

其中 $\mathcal{IW}(\nu_0,\ \Psi_0^{-1})$ 表示对 Σ_k 的逆 Wishart 先验，其自由度为 ν_0，尺度矩阵为Ψ_0^{-1}（我们采用常见的共轭参数化方式，使得 Σ_k 的后验分布仍为逆 Wishart 分布，且参数经过更新）。对于簇威沙特自由度 ν_k，我们设置独立的伽马先验：

$$p(\nu_k) \propto \nu_k^{a-1} e^{-b\nu_k}, \qquad \nu_k > 0,$$

超参数为 $a > 0,\ b > 0$。在实际应用中，采样器会强制满足 $\nu_k > p-1$。狄利克雷超参数（α_1、……、α_K）控制混合权重的先验集中度；默认情况下，我们采用对称设定$\alpha_k = \alpha_0/K$，其中较小的 α_0 有利于稀疏的成分使用，而较大的 α_0 则有利于权重近乎均匀。对于逆 Wishart 先验，ν_0 决定先验强度；为获得均值有限的弱信息先验，该值设为 $\nu_0 = p+2$，而 Ψ_0 则取为与单位矩阵成比例。ν_k 的 Gamma 超参数（$a,\ b$）经过选取，使先验均值 a/b 适度高于 $p-1$，从而确保有效性，同时保持弱信息性。

令 $\Theta = \{\boldsymbol{\pi}, z_{1:n}, (\nu_k,\ \Sigma_k)_{k=1}^K\}$。联合后验分布为

$$p(\Theta \mid S_{1:n}) \propto \left[\prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\pi_k f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k)\right\}^{\mathbf{1}_{\{z_i=k\}}}\right] p(\boldsymbol{\pi}) \prod_{k=1}^{K} p(\Sigma_k) p(\nu_k).$$

对潜在标签进行条件化后，可得到因子化后验结构，从而能够高效地进行吉布斯采样，并针对 ν_k执行 Metropolis–Hastings 步骤。
下面给出吉布斯采样器中使用的每个更新步骤，以及针对 ν_k的 Metropolis–Hastings（MH）步骤。整个过程总结于算法1中。

步骤 1：标签 z_i的条件分布。
$z_i = k$ 的后验概率（未归一化）为

$$p(z_i = k \mid S_i, \{\Sigma_\ell, \nu_\ell\}, \boldsymbol{\pi}) \propto \pi_k\ f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k). \tag{6}$$

因此，计算对数未归一化权重

$$\ell_{ik} = \log \pi_k + \frac{\nu_k - p - 1}{2}\log|S_i| - \frac{1}{2}\operatorname{tr}\left(\Sigma_k^{-1} S_i\right) - \frac{\nu_k p}{2}\log 2 - \frac{\nu_k}{2}\log|\Sigma_k| - \log\Gamma_p\left(\frac{\nu_k}{2}\right), \tag{7}$$

并从与 z_i 成比例的分类概率中采样 ℓ_{ik}，为保证数值稳定性，改用 $\exp(\ell_{ik} - \max_k \ell_{ik})$。这些 ℓ_{ik} 既用于抽取标签，也用于计算观测数据对数似然以进行监测。请注意，通过对 $\boldsymbol{\pi}$ 积分，我们得到 $p(z_i = k \mid S_i, \{\Sigma_\ell, \nu_\ell\}, \boldsymbol{\pi})$ $\propto f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot (\alpha_k + n_{-i,k})$，详情见附录 A，这可能有助于改善吉布斯采样的混合。

步骤 2：权重 $\boldsymbol{\pi}$的条件分布。
给定计数 n_k，$\boldsymbol{\pi}$ 的后验分布为狄利克雷分布：

$$\boldsymbol{\pi} \mid \mathbf{z} \sim \text{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

步骤 3：簇尺度矩阵 Σ_k的条件分布。
给定分配 $\{z_i\}$ 与簇自由度 ν_k，先验 $\Sigma_k \sim \mathcal{IW}(\nu_0, \Psi_0^{-1})$与Wishart似然共轭。后验分布为逆 Wishart：

$$\Sigma_k \mid \{S_i : z_i = k\}, \nu_k \sim \mathcal{IW}\Big(\nu_0 + n_k\nu_k,\ \Psi_0^{-1} + S_{(k)}\Big),$$

即自由度 $\nu_0 + n_k\nu_k$ 与尺度矩阵 $\Psi_0^{-1} + S_{(k)}$，其中 $S_{(k)} = \sum_{i\in I_k} S_i$ 与 $I_k = \{i :\ z_i = k\}$。注：当 $n_k = 0$（即簇 k 消失）时，我们从先验 $\mathcal{IW}(\nu_0, \Psi_0^{-1})$中抽取 Σ_k。

步骤 4：ν_k 的更新（对数尺度上的Metropolis–Hastings）。
自由度 ν_k 不适用方便的共轭更新；我们通过在对数尺度上提出候选值，使用 MH 步骤逐一更新每个 ν_k。设当前值为 ν，并提出

$$\log \nu^* \sim \mathcal{N}(\log \nu, \sigma^2_{\text{MH}}), \qquad \nu^* = \exp(\log \nu^*).$$

ν 的对数后验（针对簇 k），利用簇特定汇总量 n_k、 $S_{(k)}$ 和 $L_{(k)} = \sum_{i \in I_k} \log |S_i|$，忽略加法常数后为

$$\begin{aligned}
&\log p(\nu \mid \{S_i\}_{i \in I_k}, \Sigma_k) \\
&= \sum_{i \in I_k} \log f_{\mathcal{W}}(S_i \mid \nu, \Sigma_k) + \log p(\nu) \\
&= \frac{\nu - p - 1}{2} L_{(k)} - \frac{1}{2} \operatorname{tr}\big(\Sigma_k^{-1} S_{(k)}\big) - n_k \Big(\frac{\nu p}{2} \log 2 + \frac{\nu}{2} \log |\Sigma_k| + \log \Gamma_p\big(\tfrac{\nu}{2}\big) \Big) \\
&\quad + (a - 1) \log \nu - \nu b.
\end{aligned}$$

如果记 b y $\ell(\nu)$ 为对数后验分布，那么提议 ν^* 的接受概率为

$$\alpha = \min\left\{1,\ \exp\left(\ell(\nu^*) + \log \nu^* - \ell(\nu) - \log \nu\right)\right\},$$

其中额外的 $\log \nu$ 项是雅可比校正。我们仅对满足 $\nu^* > p - 1$ 的提议尝试 MH 更新。

算法 1 通过 MH 处理 ν_k 的 Wishart 混合模型吉布斯采样器

1: **输入：** S_1、.....、S_n，聚类数 K，超参数 α，ν_0，Ψ_0，MH 提议尺度 σ_{MH}，迭代次数 N、预热期、抽稀。2: **初始化：** 将 S_i 向量化，得到 S_{mat}（大小为 $n \times p^2$）；计算所有 i 的 $\log |S_i|$。初始化标签 z_i（建议对 vec(S_i) 使用 k 均值）、 $\pi_k = n_k/n$、 ν_k（例如 $p + 2$）以及 Σ_k（簇样本 / 先验）。3: **对于** 迭代 $= 1$，.....，N **执行**4: ▷ **步骤 1（标签）：** 对于每个簇 k，计算 Σ_k^{-1}、$\log |\Sigma_k|$ 和 tr($\Sigma_k^{-1} S_i$)，并从以下分布中抽取标签

$$p(z_i = k \mid S_i, \{\Sigma_\ell, \nu_\ell\}, \boldsymbol{\pi}) \propto f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot (\alpha_k + n_{-i,k})$$

5: ▷ **步骤 2（权重）：** 设置 $n_k = \#\{i : \ z_i = k\}$ 并抽取

$$\boldsymbol{\pi} \sim \mathrm{Dirichlet}(\alpha_1 + n_1, \ldots, \alpha_K + n_K).$$

6: ▷**步骤 3（尺度矩阵）：** 对每个 k，计算 $S_{(k)}$ 并抽取

$$\Sigma_k \sim \mathcal{IW}(\nu_0 + n_k \nu_k,\ \Psi_0^{-1} + S_{(k)}).$$

若 $n_k = 0$，则从先验分布 $\mathcal{IW}(\nu_0, \Psi_0^{-1})$ 中抽取。

7: ▷**步骤 4（自由度 ν_k）：**

8: 对于每个 k，提出 log $\nu^* \sim \mathcal{N}$(log $\nu_k, \sigma_{\mathrm{MH}}^2$)，并以概率接受

$$\min\left\{1,\ \exp\left(\ell(\nu^*) + \log \nu^* - \ell(\nu_k) - \log \nu_k\right)\right\},$$

wh此处 $\ell(\nu)$ 是使用 $n_k, L_{(k)}, S_{(k)}$ 和 Σ_k 计算得到的簇 k 的对数后验。

9: ▷**记录/监测：** 计算对数似然；保存样本。10: **结束 for**11: **返回：** $\boldsymbol{\pi}, \{\nu_k\}$, $\{\Sigma_k\}, \{z_i\}$ 的后验抽样以及监测诊断。

实际实现： 迹项的计算进一步通过向量化恒等式 tr(AB) $=$vec(A)$^\top$ vec(B) 进行优化，从而通过一个由 vec(S_i)$^\top$ 的各行构成的 $n \times p^2$ 矩阵与向量 vec(Σ_k^{-1}) 的乘积，同时计算所有 i 的 tr($\Sigma_k^{-1} S_i$)。通过使用 Cholesky分解进行矩阵求逆和行列式计算，并加入极小的对角抖动（例如 $10^{-6} I_p$）以改善病态矩阵的条件数，以及采用 Bartlett分解进行稳定的逆Wishart抽样，从而保持数值精度。最后，该算法采用

弱信息先验（例如 $\nu_0 = p+2$、$\Psi_0 = I_p$），并采用尺度为 σ^2_{MH}（可以因簇而异）的对数空间 MH 提议，将接受率调节至 20–40%；同时使用对数和指数运算，确保密度计算的精度。

3.2 贝叶斯 MoE–Wishart 模型

我们为式（5）中的 MoE–Wishart 模型参数设定如下先验分布：

$$\Sigma_k \sim \mathcal{IW}(\nu_0, \Psi_0), \quad \nu_k \sim \mathrm{Gamma}(a_\nu, b_\nu), \quad \beta_{1:(K-1)} \overset{\text{iid for columns}}{\sim} \mathcal{N}_q(0, \sigma^2_\beta I_q).$$

设 $\Theta = \{z_{1:n}, (\nu_k, \Sigma_k)_{k=1}^K, \beta_{1:(K-1)}\}$。联合后验为

$$p(\Theta \mid S_{1:n}, X_{1:n}) \propto \prod_{i=1}^n \prod_{k=1}^K \Big[\pi_{ik}(X_i;\beta), f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k)\Big]^{\mathbf{1}_{\{z_i=k\}}} \prod_{k=1}^K p(\Sigma_k)p(\nu_k) \prod_{k=1}^{K-1} p(\beta_k).$$

给定潜在标签后，可得到因子化后验结构，从而能够高效地进行吉布斯采样，并对 ν_k 执行 MH 步骤。

整　　个 MoE–Wishart 模型的 Metropolis-within-Gibbs 采样过程如下。

步骤 1：更新标签 z_i（类别采样）。

$z_i = k$ 的后验概率为

$$\Pr(z_i = k \mid S_i, X_i, \Theta) \;=\; \frac{\exp(\ell_{ik})}{\sum_{\ell=1}^K \exp(\ell_{i\ell})},$$

其中 Θ 汇集了所有系数 β_1,..., β_{K-1}，且 z_i 从该类别分布中采样，其中 ℓ_{ik}由式（7）给出。

步骤 2：更新门控系数 $\beta_{1:(K-1)}$（Metropolis–Hastings）。

我们使用 MH 采样，按簇分块更新系数。具体而言，对每个 k =1, . . . , $K-1$，我们提出

$$\beta_k^* = \beta_k + \varepsilon_k, \qquad \varepsilon_k \sim \mathcal{N}\left(0, \sigma^2_{\beta_k} I\right),$$

同时固定 $\beta_K \equiv 0$ 以确保可识别性。令 π^*_{ik} 表示在提议的系数 β^*下计算得到的门控概率。在均值为零、方差为 σ^2_β的高斯先验下，对数先验差为$\Delta \log\, p(\beta) = -\frac{1}{2\sigma^2_\beta} \sum_{k=1}^{K-1} \left(\|\beta_k^*\|_2^2 - \|\beta_k\|_2^2\right)$. 给定潜在簇标签 z后，门控似然为 $\log\, p(\mathbf{z} \mid X, \beta) = \sum_{i=1}^n \log \pi_{i,z_i}$. 因此，MH 对数目标比为

$$\log \alpha = \sum_{i=1}^n \left(\log \pi^*_{i,z_i} - \log \pi_{i,z_i}\right) + \Delta \log p(\beta).$$

该提议以概率 min{1,exp(log α)} 被接受。

步骤 3：更新 Σ_k（共轭逆 Wishart）。

每个 Σ_k 均按其逆 Wishart 后验分布依次更新：

$$\Sigma_k \mid \text{rest} \sim \mathcal{IW}\Big(\nu_0 + n_k\nu_k,\; \Psi_0^{-1} + S_{(k)}\Big),$$

其中 $n_k = |I_k|$、$I_k = \{i : \; z_i = k\}$，以及 $S_{(k)} = \sum_{i \in I_k} S_i$。

步骤 4：更新自由度 ν_k（在对数尺度上使用 Metropolis–Hastings）。

每个 ν_k 均通过对数正态随机游走提议依次更新：

$$\log \nu_k^* \sim \mathcal{N}(\log \nu_k, \sigma^2_\nu), \qquad \nu_k^* = \exp(\log \nu_k^*).$$

使用汇总量 $S_{(k)}$ 和 $L_{(k)} = \sum_{i \in I_k} \log |S_i|$，依赖于 ν 的对数联合密度（数据 + 先验）为

$$\ell(\nu) = \sum_{i \in I_k} \Big\{ \frac{\nu - p - 1}{2} \log |S_i| - \frac{\nu p}{2} \log 2 - \frac{\nu}{2} \log |\Sigma_k| - \log \Gamma_p\big(\tfrac{\nu}{2}\big) \Big\} + (a_\nu - 1) \log \nu - \nu b_\nu + C,$$

其中，C 汇集了与 ν 无关的项。利用 $L_{(k)}$ 和 $\operatorname{tr}(\Sigma_k^{-1} S_{(k)})$，我们可以高效地计算差值。MH 对数目标比为

$$\log \alpha = \big[\ell(\nu^*) + \log \nu^*\big] - \big[\ell(\nu) + \log \nu\big],$$

其中，额外的 $\log \nu$ 项是对数提议所产生的变量变换雅可比行列式。以概率 $\min(1, \exp(\log \alpha))$ 接受 ν^*。

3.3 K的选择

令 K 表示式（5）中定义的专家混合（MoE）模型里的专家数量。由于增加 K 总能使样本内似然至少不降低，模型选择必须在拟合优度与模型复杂度之间进行权衡。因此，我们使用适用于潜在混合模型的信息准则和贝叶斯预测度量来选择 K。

3.3.1 模型维数

令 p 表示 Wishart 随机矩阵 S_i 的维数，并令 q 表示协变量向量 X_i 的维数（包括截距）。每个专家 k 贡献一个 Wishart 尺度矩阵 $\Sigma_k \in \mathbb{S}_+^p$，包含 $\frac{p(p+1)}{2}$ 个自由参数，以及一个自由度参数 $\nu_k > p - 1$。因此，每个专家贡献 $\frac{p(p+1)}{2} + 1$ 个自由参数。softmax 门控网络只有在设定基准类别后才可识别；令 $\beta_K = 0$，即可得到 $(K-1)\,q$ 个自由门控参数。因此，模型的总维数为

$$\mathcal{D}_K = K\Big(\frac{p(p+1)}{2} + 1\Big) + (K-1)q.$$

3.3.2 贝叶斯信息准则

令 Θ_K 表示在包含 K 个组分的模型下的最大似然估计量或后验众数估计量。贝叶斯信息准则（BIC）定义为

$$\mathrm{BIC}(K) = -2\ell(\hat{\Theta}_K) + \mathcal{D}_K \log n,$$

其中 $\ell(\Theta)$ 是（8）中的观测数据对数似然。

3.3.3 积分完备似然

由于混合模型可能通过拆分现有组分来人为提高 K，我们还考虑积分完备似然（ICL）[2]，该准则会对类别分配中的后验不确定性进行惩罚。令 $r_{ik} = \Pr(z_i = k \mid S_i, X_i, \Theta_K)$ 表示后验责任度。ICL 准则为

$$\mathrm{ICL}(K) = \mathrm{BIC}(K) - 2 \sum_{i=1}^{n} \sum_{k=1}^{K} r_{ik} \log r_{ik}.$$

熵项偏好专家之间分隔明显的模型，因此能够选出实质上不同的状态数量。

3.3.4 贝叶斯留一法

要比较贝叶斯预测模型，可以使用留一交叉验证（LOO）或广泛适用信息准则（WAIC）来估计样本外逐点预测准确度。这两种方法都基于在后验抽样处评估对数似然，而 WAIC 与 LOO 渐近等价。然而，[33] 研究表明，通过帕累托平滑重要性抽样（PSIS）实现的 LOO 在有限样本中更稳健，尤其是在存在弱先验或影响性观测时。

LOO 下的期望逐点对数预测密度（elpd）定义为

$$\mathrm{elpd}_{\mathrm{loo}} = \sum_{i=1}^{n} \log p(S_i \mid S_{-i}), \qquad p(S_i \mid S_{-i}) \approx \left(\frac{1}{N} \sum_{t=1}^{N} \frac{1}{p(S_i \mid \Theta^{(t)})} \right)^{-1}.$$

其中，S_{-i} 表示移除第 i 个观测后的数据；$p(S_i \mid S_{-i})$ 可通过对来自 $p(\Theta \mid S)$ 的后验抽样 $\{\Theta^{(t)}\}_{t=1}^{N}$ 进行重要性抽样来近似估计，其中 $\Theta^{(t)}$ 表示第 t 次 MCMC 迭代时的完整参数向量。由于原始重要性比率可能具有厚尾，PSIS 会用拟合的广义帕累托分布替换其上尾，从而得到稳定化的重要性权重和一个诊断形状参数，该参数用于指示近似结果的可靠性。由此得到的预测准确度 LOO 估计量为

$$\widehat{\mathrm{elpd}}_{\mathrm{loo}} = \sum_{i=1}^{n} \widehat{\mathrm{elpd}}_{\mathrm{loo},i} = \sum_{i=1}^{n} \log p(S_i \mid S_{-i}),$$

其中，$\widehat{\mathrm{elpd}}_{\mathrm{loo},\,i}$ 表示观测 i 的贡献。

3.3.5 模型选择规则

我们结合这些互补准则来选择专家数量 K。贝叶斯信息准则在大样本下具有一致性，ICL 用于确定具有经济意义的状态数量，而 $\mathrm{elpd}_{\mathrm{loo}}$ 尤其适用于评估贝叶斯方法的预测性能。实际应用中，我们选择使贝叶斯信息准则和 ICL 最小，或使贝叶斯方法的 $\mathrm{elpd}_{\mathrm{loo}}$ 最大时对应的最小K。

4 最大似然与 EM 算法

4.1 最大似然估计

经典 Wishart 混合模型的推断是通过期望最大化（EM）算法完成的 [8]。这里，我们使用 EM 算法通过最大似然法估计式（5）中的专家混合模型（MoE）的参数。由于似然函数包含由未观测成分标签诱导的潜在混合结构，因此 EM 框架尤其适用于此处。

令 $\Theta = \{(\beta_k,\ \nu_k,\ \Sigma_k)\ _{k=1}^{K}\}$ 表示专家混合模型设定下的模型参数。专家混合模型（5）的观测数据对数似然为：

$$\ell(\Theta) = \sum_{i=1}^{n} \log \left(\sum_{k=1}^{K} \pi_{ik}(X_i;\beta) \cdot f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \right). \tag{8}$$

由于对数函数内部存在求和项，直接最大化在解析上难以处理。EM 算法通过引入潜在类别标签 z_i 并迭代最大化一个易于处理的替代目标函数，克服了这一困难。

如果观测到了潜在簇标签 z_i，完全数据对数似然（忽略加性常数）为：

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\Big[\log \pi_{ik}(X_i;\beta) + \log f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k)\Big],$$

将 Wishart 密度代入后，可得：

$$\ell_c(\Theta) = \sum_{i=1}^{n}\sum_{k=1}^{K} z_{ik}\left[\frac{\nu_k - p - 1}{2}\log|S_i| - \frac{\nu_k}{2}\log|\Sigma_k| - \frac{1}{2}\operatorname{tr}(\Sigma_k^{-1}S_i) - \log\Gamma_p\left(\frac{\nu_k}{2}\right) - \frac{\nu_k p}{2}\log 2 + \log \pi_{ik}(X_i;\beta)\right]. \quad (9)$$

这种表示揭示了门控参数 β 和专家参数 $(\nu_k,\ \Sigma_k)$ 在似然中分别以独立的模块形式出现，因此可以方便地执行条件最大化步骤。

4.2 EM 算法

令 $z_i \in \{1,\dots,K\}$ 表示观测值 i的潜在簇标签，并定义后验责任度：

$$r_{ik} := \Pr(z_i = k \mid S_i, X_i, \Theta).$$

我们引入一些汇总量以简化符号表示：

$$n_k = \sum_{i=1}^{n} r_{ik}, \qquad M_k = \sum_{i=1}^{n} r_{ik} S_i, \qquad \overline{\log|S|}_k = \frac{1}{n_k}\sum_{i=1}^{n} r_{ik}\log|S_i|, \qquad \overline{S}_k = \frac{M_k}{n_k}.$$

EM 算法在以下两步间交替进行：通过门控网络计算观测值分配给各专家的“软”分配（责任度）（E 步），以及通过加权多项逻辑回归更新门控系数 β，并通过加权最大似然更新专家参数 $(\nu_k,\ \Sigma_k)$（M 步）。这种交替优化通常会使观测数据对数似然在每次迭代中单调不减，并为 MoE–Wishart 模型提供稳定且具有统计学依据的估计量。

E 步。 在第 (t) 次迭代中，我们计算后验责任度：

$$r_{ik}^{(t)} = \Pr(z_i = k \mid S_i, X_i, \Theta^{(t)}) = \frac{\pi_{ik}(X_i;\beta^{(t)}) f_{\mathcal{W}}(S_i \mid \nu_k^{(t)}, \Sigma_k^{(t)})}{\sum_{j=1}^{K} \pi_{ij}(X_i;\beta^{(t)}) f_{\mathcal{W}}(S_i \mid \nu_j^{(t)}, \Sigma_j^{(t)})}.$$

这些权重表示在当前参数估计值下，观测 i 来自专家 k 的概率。

M 步 。 最大化完全数据对数似然的条件期望 ：

$$Q(\Theta \mid \Theta^{(t)}) = \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t)}\Big[\log \pi_{ik}(X_i;\beta) + \log f_{\mathcal{W}}(S_i \mid \nu, \Sigma)\Big].$$

- 首先，更新门控网络 (β)： β 的更新对应于加权多项逻辑回归：

$$\beta^{(t+1)} = \arg\max_{\beta} \sum_{i=1}^{n}\sum_{k=1}^{K} r_{ik}^{(t)} \log \pi_{ik}(X_i;\beta).$$

这个优化问题没有闭式解，需通过数值方法求解，例如使用 BFGS 或牛顿-拉夫逊法。

- 第二，更新尺度矩阵（Σ_k）：固定 ν_k 时，关于 Σ_k 的 $Q(\Theta \mid \Theta^{(t)})$ 的最大化解具有闭式表达式：

$$\Sigma_k^{(t+1)} = \frac{1}{n_k \nu_k^{(t+1)}} \sum_{i=1}^{n} r_{ik}^{(t)} S_i = \frac{\overline{S}_k}{\nu_k^{(t+1)}}.$$

- 第三，更新自由度（ν_k）：通过求解得分方程，获得 ν_k 的更新值：

$$\psi_p\left(\frac{\nu_k}{2}\right) - \log \nu_k + \log\left(\frac{|M_k|}{n_k}\right) - \overline{\log |S|}_k + p \log 2 - p \log n_k = 0,$$

其中，$\psi_p(\cdot)$ 表示多元 digamma 函数。该非线性方程通过牛顿–拉夫森法或类似的求根方法进行数值求解。

5 模拟研究

用于复现模拟结果的代码可在https://github.com/zhizuio/moewishart/tree/main/numerical_研究中获取。

5.1 混合模型模拟

5.1.1 设置

我们考虑具有 $K = 3$ 个簇的 Wishart 分布有限混合。混合权重固定为 $\boldsymbol{\pi} = (\pi_1, \pi_2, \pi_3) = (0.35, 0.40, 0.25)$. 我们改变样本量 $n \in \{200, 500, 1000\}$ 和矩阵维度$p \in \{2, 8\}$。每个正定观测协方差矩阵 S_i、$i = 1, \ldots, n$均独立地由该混合模型生成

$$S_i \sim \sum_{k=1}^{3} \pi_k \mathcal{W}_p(\nu_k, \Sigma_k),$$

其中 $\mathcal{W}_p(\nu_k, \Sigma_k)$ 表示自由度为 ν_k 、尺度矩阵为 Σ_k的 p维威沙特分布。自由度按照对应于矩阵维度 p 的两种选项的两种情形加以设定：

$$(\nu_1, \nu_2, \nu_3) = \begin{cases} (8, 12, 3), & \text{if } p = 2, \\ (9, 20, 14), & \text{if } p = 8. \end{cases}$$

类似地，尺度矩阵 Σ_k按对应于矩阵维度的情形加以设定：

- If $p = 2$, we set $\Sigma_1 = \begin{pmatrix} 0.5 & 0.2 \\ 0.2 & 0.7 \end{pmatrix}$, $\Sigma_2 = \begin{pmatrix} 2.0 & 0.6 \\ 0.6 & 1.5 \end{pmatrix}$, $\Sigma_3 = \begin{pmatrix} 4.0 & 0.2 \\ 0.2 & 3.0 \end{pmatrix}$.
- If $p = 8$, we take $\Sigma_1 = \{0.5^{|j-j'|}\}_{jj'}$, $\Sigma_2 = \{0.2^{|j-j'|}\}_{jj'}$, $\Sigma_3 = \{0.8^{|j-j'|}\}_{jj'}$, $j, j' \in \{1, ..., p\}$.

所考虑的方法：对于每种模拟设置，我们评估四种相互竞争的估计方法：

- 其中两种基于有限混合模型：完整贝叶斯方法（记为**贝叶斯**）以及通过 EM 算法实现的极大似然估计量（记为**期望最大化**）。
- 另外两种方法基于专家混合框架：完整贝叶斯方法（记为**Bayes-MoE**）和基于EM的极大似然方法（记为**EM-MoE**）。具体而言，**Bayes-MoE** 和**EM-MoE** 均仅将截距作为伴随协变量，门控概率 π_{ik} 则简化为 π_k。

对于两个完整贝叶斯版本（即**贝叶斯** 和**Bayes-MoE**），我们运行马尔可夫链蒙特卡洛，共进行 20000 次迭代，其中前 5000 次迭代作为预热期舍弃。
模型性能评估：我们使用所有组分误差的平均值来评估所考虑方法的性能。具体而言，我们考虑以下指标：

$$\frac{1}{K}\|\hat{\boldsymbol{\pi}}-\boldsymbol{\pi}\|_1;\quad \frac{1}{K}\|\hat{\boldsymbol{\nu}}-\boldsymbol{\nu}\|_1;\quad \frac{1}{K}\sum_{k=1}^{K}\|\hat{\Sigma}_k-\Sigma_k\|_2^2.$$

对于两个完整贝叶斯版本（即**贝叶斯**和**Bayes-MoE**），使用后验均值估计量计算评估指标。我们针对每种模拟设置生成100个数据集，并在图1中报告结果。

5.1.2 模拟结果

图1表明，对于**贝叶斯、期望最大化、Bayes-MoE**和**EM-MoE**这四种方法，参数 $\boldsymbol{\nu}$、 Σ_k和 $\boldsymbol{\pi}$ 的估计精度均随着样本量增加而稳步提高，数据由混合模型生成。

对于 $p=2$，图1A显示，在最小样本量（$n=200$）下，设定正确的混合模型EM方法比其他三种方法产生更大的估计误差；在样本量为 $n\in\{500,1000\}$时则出现较大的离群值。这些离群值是因为期望最大化得到的估计值 $\boldsymbol{\nu}$、 $\Sigma_{1:K}$ 和 $\boldsymbol{\pi}$ 在这些模拟数据集中未能收敛。设定正确的混合模型贝叶斯方法和专家混合方法（Bayes-MoE 和 EM-MoE）在所有考察的样本量（$n\in 200$、500、1000）下，对 $\boldsymbol{\nu}$、$\Sigma_{1:K}$ 和 $\boldsymbol{\pi}$ 的估计误差相近。对于 $p=8$，图1B表明，四种方法的表现相当；但由于某些模拟数据集中 $\boldsymbol{\nu}$、 $\Sigma_{1:K}$ 和 $\boldsymbol{\pi}$ 存在收敛问题，期望最大化方法始终会产生少量离群值。

补充图S7（$p=2$）和S8（$p=8$）展示了贝叶斯方法（贝叶斯和Bayes-MoE）的马尔可夫链蒙特卡洛（MCMC）行为稳定，这一点可从各工作模型的对数似然轨迹图中得到验证。此外，簇特定参数——贝叶斯方法中的π_k、 ν_k和log $|\Sigma_k|$ ，以及专家混合方法中的 ν_k和log $|\Sigma_k|$ （$k\in$ 1、2、3）——均表现出令人满意的混合效果。与 $n=500$ 和 $n=1000$相比， $n=200$ 的轨迹图如预期呈现出更大的波动。表1报告了所选参数的有效样本量（ESS），结果显示，对于两种贝叶斯方法，与较大的样本量相比，$n=200$ 下的ESS值系统性地较小且变化更大。最后，补充图S11A–B证实，对于混合工作模型和专家混合工作模型，基于一组模拟数据，期望最大化算法在所有模拟情景下均能就对数似然可靠收敛。

5.2 带协变量专家模型的混合

5.2.1 设置

我们将 Wishart 分布有限混合扩展到具有依赖协变量的混合比例的专家混合模型设置中。与之前一样，我们取 $K=3$ 个潜在簇，并考察样本量 $n\in\{200,500,1000\}$ 和矩阵维度 $p\in\{2,8\}$。

给定 $q=3$，令 $X\in\mathbb{R}^{n\times q}$ 表示协变量矩阵，其中各元素独立生成为 $x_{ij}\sim\mathcal{N}(0,\ 1)$。混合比例通过多项逻辑回归（softmax）函数获得

$$\pi_{ik}(X_i;\beta)\ =\ \frac{\exp\{X_i^\top\beta_k\}}{\sum_{l=1}^{K}\exp\{X_i^\top\beta_l\}},\quad i=1,\ldots,n,\ k=1,\ldots,K,$$

其中协变量效应为 $\beta_{1:K}\in\mathbb{R}^{q\times K}$，并采用参考约束 $\beta_K=\mathbf{0}$ 以确保可识别性。效应 β 的每个坐标均独立地从均匀分布 $\beta_{jk}\sim\mathcal{U}(-2,\ 2)$ 中模拟生成，并且 $\beta_{1:K}$ 对每次模拟都相同。

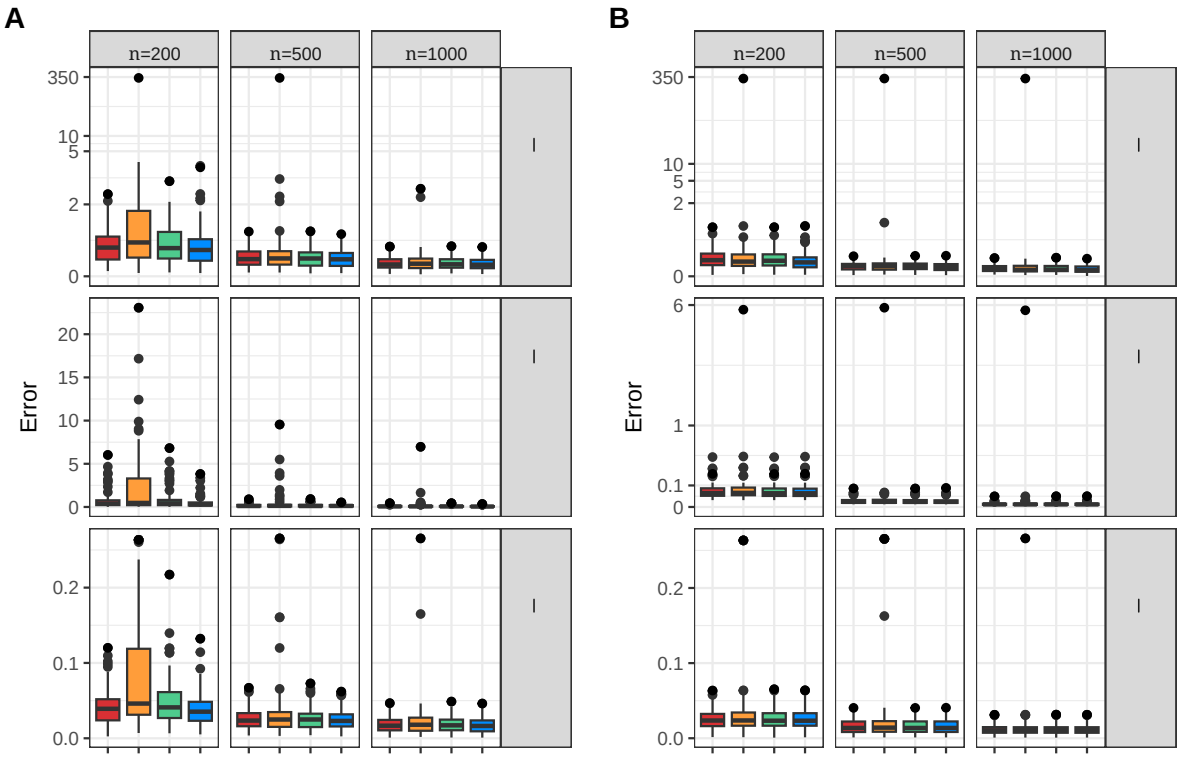


图1：模拟结果。数据来自有限混合模型，共进行了100次蒙特卡洛重复。报告了样本量为 $n \in \{200, 500, 1000\}$ 时两种维度设置下的结果：(A) $p = 2$ 和(B) $p = 8$。比较的四种方法包括：(i) 贝叶斯混合模型（**贝叶斯**）；(ii) 基于期望最大化的混合模型（**期望最大化**）；(iii) 贝叶斯专家混合模型（**Bayes-MoE**）；以及(iv) 基于期望最大化的专家混合模型（**EM-MoE**）。采用分量平均误差评估估计性能：$\frac{1}{K}\|\hat{\boldsymbol{\pi}} - \boldsymbol{\pi}\|_1$；$\frac{1}{K}\|\hat{\boldsymbol{\nu}} = \boldsymbol{\nu}\|_1$；$\frac{1}{K}\sum_{k=1}^{K}\|\hat{\Sigma}_k - \Sigma_k\|_2^2$.

表1：贝叶斯和 Bayes-MoE 方法通过马尔可夫链蒙特卡洛得到的有效样本量（ESS）。数据由混合模型生成。报告的是针对所选参数，在 100 次模拟中 ESS 的均值和标准差。

工作模型		$\boldsymbol{\nu}$		Σ		$\boldsymbol{\pi}$	
		ν_1	ν_2	$\Sigma_{1,11}$	$\Sigma_{2,11}$	π_1	π_2
混合模型:	$p=2$						
	$n=200$	675 (250)	164 (64)	1079 (525)	214 (88)	395 (244)	907 (865)
	$n=500$	825 (140)	204 (43)	1339 (349)	259 (61)	500 (155)	1040 (455)
	$n=1000$	794 (110)	183 (31)	1192 (298)	222 (45)	433 (107)	894 (331)
混合模型:	$p=8$						
	$n=200$	1295 (91)	760 (69)	8528 (797)	3303 (514)	14651 (481)	14641 (541)
	$n=500$	1413 (85)	838 (47)	8907 (604)	3709 (338)	14922 (323)	14965 (263)
	$n=1000$	1340 (87)	829 (56)	9179 (651)	3559 (324)	14990 (100)	14992 (99)
混合专家模型:	$p=2$						
	$n=200$	622 (266)	170 (70)	1029 (521)	228 (227)	-	-
	$n=500$	713 (158)	179 (36)	1085 (322)	213 (47)	-	-
	$n=1000$	678 (138)	176 (33)	999 (267)	199 (44)	-	-
混合专家模型:	$p=8$						
	$n=200$	1315（91）	814（66）	8784（871）	3597（462）	-	-
	$n=500$	1387（73）	851（49）	9069 (570)	3610 (301)	-	-
	$n=1000$	1349 (54)	821 (47)	8837 (607)	3565 (316)	-	-

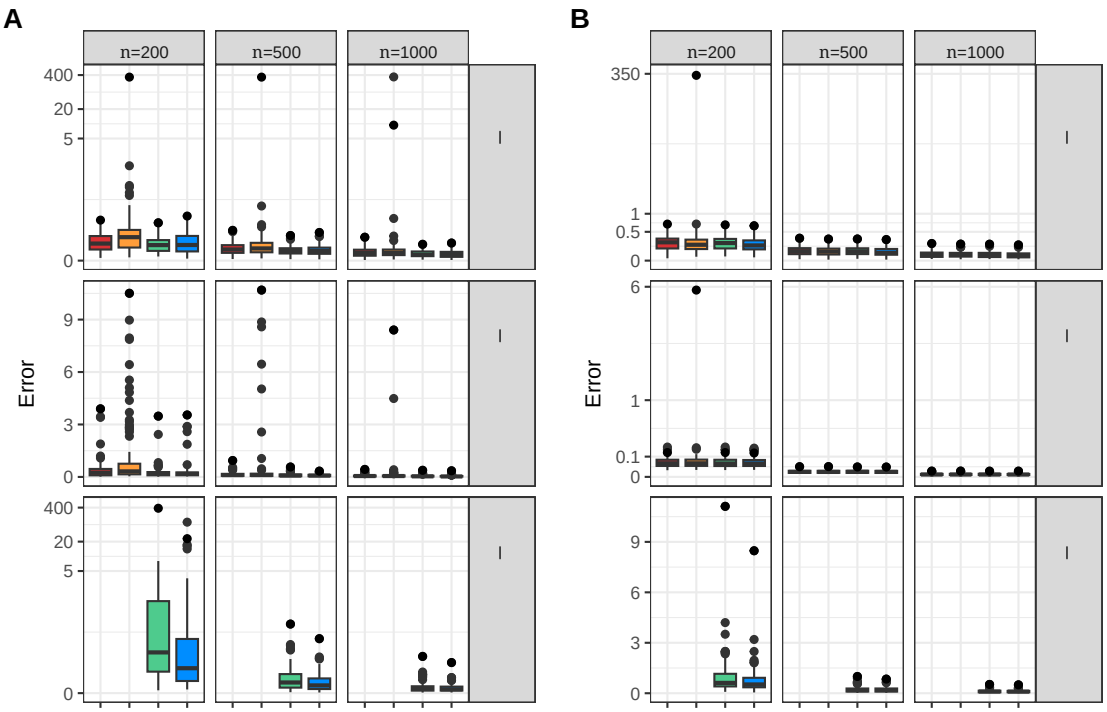


图2：模拟结果。数据由**专家混合模型（MoE）**生成，共包含100个模拟数据集。面板（A）和（B）分别对应设置 $p = 2$ 和 $p = 8$。比较的四种竞争方法为：**(i)** 基于混合模型的贝叶斯推断（**贝叶斯**）；**(ii)** 基于混合模型的基于EM的推断（**期望最大化**）；**(iii)** 基于专家混合模型的贝叶斯推断（**Bayes-MoE**）；以及**(iv)**基于专家混合模型的基于EM的推断（**EM-MoE**）。

给定 X_i，每个正定矩阵 S_i 均独立地由专家混合模型生成

$$S_i \sim \sum_{k=1}^{3} \pi_{ik}(X^\top \beta)\, \mathcal{W}_p(\nu_k, \Sigma_k), \quad i = 1, \ldots, n$$

此处的簇参数（ν_k， Σ_k）与第5.1.1节中有限混合设置的参数相同。
模型性能评估：我们使用所有聚类误差的平均值来评估所考察方法的性能。具体而言，我们考虑以下指标：

$$\frac{1}{K}\|\hat{\boldsymbol{\nu}} - \boldsymbol{\nu}\|_1; \quad \frac{1}{K}\sum_{k=1}^{K}\|\hat{\Sigma}_k - \Sigma_k\|_2^2; \quad \frac{1}{K}\sum_{k=1}^{K}\|\hat{\beta}_k - \beta_k\|_2^2.$$

所有其他设置均与第5.1.1节中的有限混合情形相同。

5.2.2 模拟结果

与混合模型设置中的情况一样，图2表明，随着样本量增加，专家混合模型下的估计精度单调提高（$p = 2$ 对应图2A，$p = 8$对应图2B）。在所有情形下，Bayes-MoE 和 EM-MoE 的表现相当。对于 $p = 2$ 和 $p = 8$，在最小样本量（$n = 200$）下，Bayes-MoE 对 $\boldsymbol{\nu}$ 的估计误差略小于 EM-MoE，但对 $\boldsymbol{\beta}$ 的估计误差略大（左下图）。基于混合模型的方法（贝叶斯和期望最大化）表现略逊于设定正确的专家混合方法；对于 $p = 2$，期望最大化在各项指标上的估计误差始终最大。对于$p = 8$，由于某些模拟数据集中 $\boldsymbol{\nu}$ 和 $\Sigma_{1:K}$ 存在收敛问题，期望最大化也始终会产生少数离群值。

对于由混合专家模型生成的数据，补充图 S9 ($p = 2$) 和 S10 ($p = 8$) 进一步表明，两种贝叶斯方法的马尔可夫链蒙特卡洛（MCMC）表现稳定，证据是各工作模型的对数似然轨迹图和特定成分的参数轨迹。与混合模型结果一致，在 $n = 200$ 时观察到的波动大于 $n = 500$和 $n = 1000$。表 S3 列出了部分参数的有效样本量（ESS）；对于两种贝叶斯方法，有效样本量均随样本量增加而系统性增大。最后，补充图 S11C-D 证实，在两种情景下，无论是混合模型还是专家混合工作模型，期望最大化算法在对数似然方面均已收敛 $p \in \{2, 8\}$。

5.3 识别真实簇数的结果 K

我们使用第 4.3 节介绍的准则，比较了基于信息准则的模型选择在确定混合模型和混合专家模型的聚类数时的表现：贝叶斯工作模型采用 elpd_{loo} 和 ICL，EM 工作模型采用贝叶斯信息准则和 ICL。在真实 $K = 3$ (由混合模型或混合专家模型生成数据时) 的模拟设置下，我们观察到：(i) 贝叶斯混合专家模型 **(Bayes-MoE)**：elpd_{loo} 正确识别了 $K = 3$ (表 2)。(ii) 贝叶斯混合模型 **(Bayes)**：elpd_{loo} 略有倾向于选择比真实值更大的 K。(iii) EM 工作模型 **(EM和 EM-MoE)**：贝叶斯信息准则正确识别了 $K = 3$。在所有设置中，ICL 均比其他准则更为保守，倾向于选择较小的 K (即相对于真实值略有欠分簇)。

对于更高维（$p = 8$），真实 $K = 3$ 下的混合模型和专家混合模型，补充表S4 显示，elpd_{loo} 对于与 K 结果一致的贝叶斯工作模型，略微倾向于选择比真实值更大的 $p = 2$，而 ICL 在所有情景下均能正确识别 $K = 3$。与 $p = 2$ 情况类似，来自 EM 工作模型的 BIC 和 ICL（**期望最大化**和**EM-MoE**）在所有情景下均能正确识别 $K = 3$，但对于由混合专家模型生成的数据，**期望最大化** 会选择 $K = 4$。

表2：由混合模型和混合专家模型生成的模拟数据上的聚类数选择，其中 $n = 200$, $p = 2$ ，且真实的 $K = 3$。报告的是每种信息准则在100次模拟中的均值（标准差）。贝叶斯工作模型通过 elpd_{loo} （越大越好）和 ICL（越小越好）进行评估；EM 工作模型通过贝叶斯信息准则和 ICL（两者均越小越好）进行评估。每种方法的最优值以粗体标出。

K	2	3	4	5	6
由混合模型生成的数据					
贝叶斯					
elpd_{loo}	-802.9 (33.5)	-776.2 (33.5)	-775.7 (33.6)	-775.1 (33.5)	**-774.6** (33.5)
ICL	**1692.0** (70.3)	1753.3 (72.5)	1915.9 (86.5)	2033.3 (85.5)	2143.6 (93.7)
Bayes-MoE					
elpd_{loo}	-1863.5 (33.4)	**-1838.6** (33.5)	-1839.0 (33.5)	-1839.0 (33.5)	-1839.2 (33.6)
ICL	**4030.0** (63.1)	4158.8 (68.0)	4241.9 (79.6)	4303.5 (85.6)	4383.7 (89.9)
EM					
BIC	3751.5 (66.9)	**3725.1** (69.4)	3736.1 (68.2)	3749.7 (68.6)	3766.3 (68.1)
ICL	**3796.1** (70.5)	3829.1 (76.2)	3857.9 (73.5)	3869.1 (75.4)	3889.4 (86.1)
EM-MoE					
BIC	3751.9 (67.8)	**3720.5** (67.3)	3737.7 (67.9)	3755.3 (67.1)	3772.3 (67.1)
ICL	**3797.0** (72.6)	3848.6 (71.8)	3894.5 (80.8)	3932.9 (77.4)	3958.7 (80.3)
由混合专家模型生成的数据					
贝叶斯					
elpd_{loo}	-830.0 (30.7)	-804.3 (29.5)	-803.8 (29.5)	-803.6 (29.5)	**-803.2** (29.5)
ICL	**1774.4** (67.6)	1851.7 (62.0)	2024 (71.7)	2152.1 (73.5)	2253.5 (76.0)
Bayes-MoE					
elpd_{loo}	-1850.5 (31.2)	**-1812.9** (31.2)	-1813.8 (31.3)	-1814.1 (31.5)	-1815.1 (31.5)
ICL	**3902.1** (69.2)	3960.0 (68.7)	4008.5 (70.3)	4061.1 (71.8)	4110.9 (70.2)
EM					
BIC	3821.2 (60.5)	**3812.0** (59.9)	3840.7 (58.5)	3874.0 (59.6)	3904.1 (59.3)
ICL	**3876.0** (66.3)	3939.0 (63.2)	3979.9 (66.3)	4015.3 (75.6)	4047.5 (78.8)
EM-MoE					
BIC	3734.5 (61.4)	**3688.2** (62.4)	3714.9 (62.1)	3741.7 (63.4)	3768.3 (63.7)
ICL	**3769.8** （65.6）	3775.3 (64.7)	3822.6 (73.3)	3861.6 (74.6)	3888.4 (77.2)

6 癌症药物筛选数据的应用

6.1 数据处理

癌症治疗反应门户（CTRP）v2 是一个大规模癌细胞系药物筛选资源，包含 481 种化合物在代表 24 种原发肿瘤类型的 860 个细胞系中的响应数据[28]。这些数据可在 https://portals.broadinstitute.org/ctrp 公开获取。令 $r_i^d(c_j)$ 表示药物 d 在药物浓度/剂量 c_j 下，细胞系 i（$i = 1, ..., 860$）的细胞活力（或抑制率），其中 $j = 1, ..., p$，且 $d = 1, ..., 481$。对于给定药物 d，令 n_d 为测得响应的细胞系数量。令 $\boldsymbol{r}_i^d = (r_i^d(c_1) ..., r_i^d(c_p))^\top \in \mathbb{R}^p$，以及 $\bar{\boldsymbol{r}}^d = (\bar{r}^d(c_1) ..., \bar{r}^d(c_p))^\top \in \mathbb{R}^p$，其中 $\bar{r}^d(c_j)$ 是药物 d 在有测量数据的 n_d 个细胞系中，剂量 c_j 下的平均响应。经验性的逐药物剂量-剂量协方差为

$$S^{(d)} = \frac{1}{n_d - 1} \sum_{i=1}^{n_d} (\boldsymbol{r}_i^d - \bar{\boldsymbol{r}}^d)(\boldsymbol{r}_i^d - \bar{\boldsymbol{r}}^d)^\top \in \mathbb{R}^{p \times p}.$$

这种剂量-反应表示充分利用了剂量-反应信息，而不将其简化为单一参数概括（例如半数抑制浓度或剂量-反应曲线下面积（曲线下面积；图3），并且在曲线拟合不可靠时，避免强行假设特定的 S 形形式。通过对药物剂量之间的协方差建模，$S^{(d)}$ 能够捕捉每种药物在不同细胞系中低、中、高剂量下的抑制作用如何共同变化，从而保留剂量轴上的曲线形状。

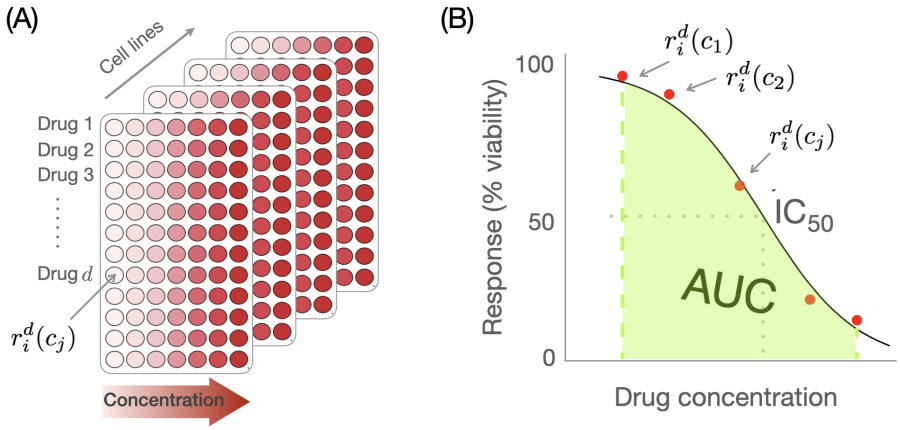


图 3：对接受多种浓度药物处理的患者活检细胞系进行的实验设置与数据。（A）在多孔细胞培养板中进行单个细胞系实验的设置。（B）来自培养板一排孔的数据所呈现的药物剂量/浓度-反应关系。曲线下面积是剂量-反应曲线下面积，通常通过拟合 S 形函数获得。IC_{50}（半数抑制浓度）表示相较于对照组，使细胞活力（存活率）降低 50% 所需的药物浓度。

为估计药物内部的剂量间协方差，我们分析了在相同浓度下进行分析的 $n = 374$ 种药物的一个子集：0.002、0.016、0.130、1.000 和 8.300μM。利用重复细胞活力测量，我们为每种药物计算了一个 $p \times p$ 协方差矩阵 $S^{(d)}$，共得到 $n = 374$ 个矩阵。

我们使用有限混合模型和专家混合模型对每种药物的剂量间协方差矩阵进行建模，以区分潜在药物类别。具有相似作用机制（MoA）的药物预计会在不同细胞系中引发相似的剂量–反应形状关系（例如，低、中、高剂量效应如何共同变化）；这些模式编码在协方差矩阵 $S^{(d)}$ 中。

专家混合模型的设定还纳入了药物特异性或情境特异性协变量（分子结构和/或遗传信息），以强化类别区分，并帮助从分子靶点角度进行解释。对于混合专家模型，我们纳入了两个药物层面协变量。第一个协变量是药物状态，以二元指标编码（0：已获美国食品药品监督管理局批准或用于临床试验；1：实验性

化合物）。第二个协变量根据化合物的 SMILES（简化分子线性输入规范）得出，该规范用于编码化学结构。我们将 SMILES 字符串转换为分子指纹；随后对指纹矩阵进行主成分分析，并将第一个主成分用作混合专家模型中的协变量。

6.2 结果

模型选择结果汇总于补充表S5。对于贝叶斯混合模型（Bayes），$K = 5$ 和 $K = 7$ 的 elpd_{loo} 值最高，且 $K = 5$ 和 $K = 7$ 的表现基本持平。$K = 5$ 的马尔可夫链蒙特卡洛诊断（补充图 S12）表明抽样过程稳定，而 ICL 也更倾向于$K = 5$ 。与这些发现一致，贝叶斯混合专家模型（Bayes-MoE）在 $K = 6$时取得最佳 elpd_{loo}，但 $K = 5$ 的预测性能相近（补充表S5），马尔可夫链蒙特卡洛表现稳定（补充图 S13），且 ICL 更倾向于该模型。因此，在后续分析中，我们对两种贝叶斯方法均采用$K = 5$ 。相比之下，基于期望最大化的混合模型和混合专家模型在 BIC/ICL 下分别选择了 $K = 10$ 和 $K = 11$（见补充表 S5）。

为评估四种方法得到的药物分组是否具有生物学意义，我们将其与针对相同化合物整理并经过充分注释的作用机制（MoA）进行比较验证。作为基准类别，我们还将其与基于剂量反应曲线下面积（AUC）的药物敏感性谱层次聚类结果进行比较；这是文献中的常见做法 [28, 37, 36, 35]。我们将比较范围限定为在370种细胞系中进行测定且具有完整AUC测量值的172种药物。对标准化AUC值生成的树状图进行目视检查（采用完全连接法和欧氏距离；图 4）表明，基于AUC的聚类结果与已确立的作用机制大多不一致，并且倾向于将靶点和药理学特性迥异的药物混在一起。

相比之下，**Bayes-MoE** 模型生成的簇与具有生物学可解释性的作用机制相吻合（纯度= 0.490，图 5 ）。我们使用聚类纯度量化作用机制一致性，其定义为每个簇中与占主导地位的作用机制相符的化合物比例（确切公式见图5 的图注）。值得注意的是，**Bayes-MoE** 模型将多种受体酪氨酸激酶（RTK）抑制剂归入同一簇，包括舒尼替尼、尼洛替尼、利尼法尼、坦度替尼、NVP-ADW742、KW-2449和SU11274。这种一致性涵盖多个RTK类别（例如血管内皮生长因子受体、c-KIT、FLT3），体现了它们共享的靶点类别以及预期的药效相似性。对于细胞毒性DNA损伤剂和烷化剂，**Bayes-MoE** 同样将铂类化合物和经典烷化剂归入符合共同作用机制的连贯簇中，包括替莫唑胺、苯达莫司汀、卡铂和有机铂类药物Platin。尽管贝叶斯混合模型（**贝叶斯**）通常也能得到类似的高层次分组（纯度= 0.480，图 5），但在若干作用机制上存在不一致之处。例如，在**贝叶斯**模型下，组蛋白去乙酰化酶（HDAC）抑制剂被分散到多个簇中：伏立诺他、贝利司他和恩替诺特尽管共享主要靶点类别（HDAC1/2/3/6/8），却未被归为一类。同样，经典抗代谢药和核苷类似物也分布于多个簇中，包括甲氨蝶呤、吉西他滨、阿糖胞苷、氯法拉滨和奈拉滨，导致一个药理学上连贯的作用机制类别被拆散。

基于 EM 的方法（**期望最大化** 和 **EM-MoE**）在 MoA 一致性方面的稳定性更差（纯度分别为 0。334 和 0。295；图 5）。例如，**期望最大化**将 RTK 抑制剂分散到多个簇中，而 **EM-MoE** 的分配则进一步将同一组 RTK 药物拆分到多个组中。这些观察结果凸显了 **Bayes-MoE** 在捕捉与机制指导的生物学类别相一致的模块化结构方面的优势。图 6 展示了对应于五个组分的协方差矩阵后验均值。

7 讨论与结论

本研究构建了一个用于协方差矩阵数据聚类的统一框架。通过建立完全贝叶斯的 Wishart 分布混合模型，并将其扩展为专家混合架构，我们突破了基于均值的聚类范式，直接对协方差模式的变异性进行建模。

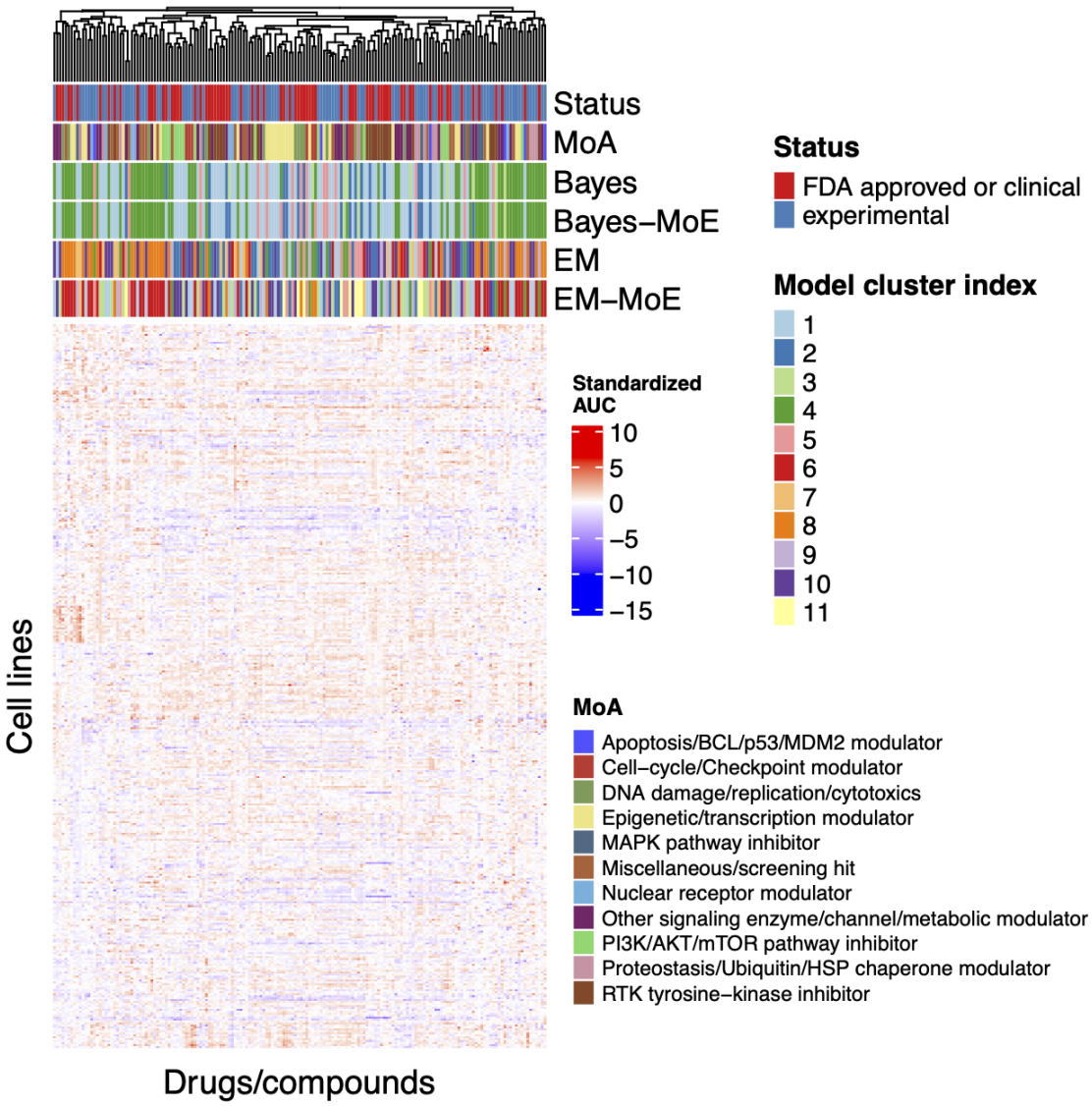


图4：基于敏感性数据的药物聚类。热图显示了370个细胞系经172种药物处理后的AUC值。对AUC值进行标准化后，采用层次聚类（欧氏距离下的完全连接法）来评估化合物之间的相似性。图例“模型聚类索引”展示了四种方法得到的聚类标签。如图所示，基于AUC值的聚类与已知的作用机制（**作用机制**）并不一致，而贝叶斯混合专家模型得到的基于机制的分组则更加合理。

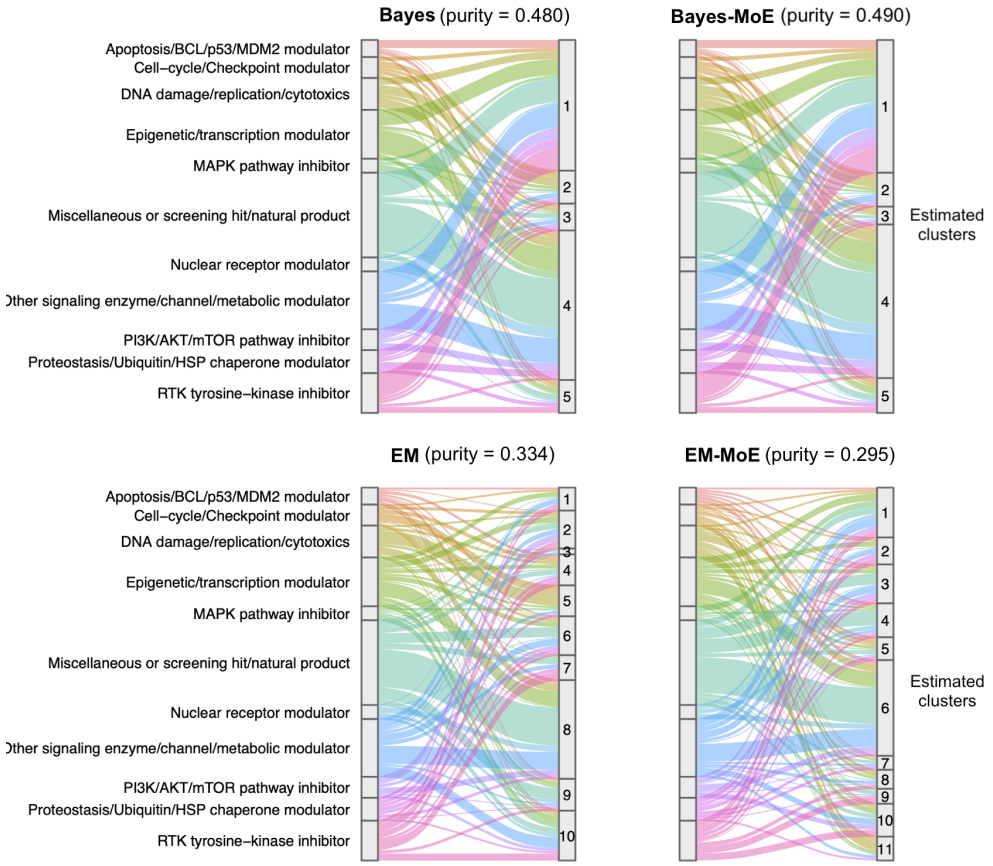


图 5：展示 MoA组流向各个基于模型的聚类的桑基图，用以说明聚类-MoA凝聚性。每个面板中的纯度定义为：purity $= \frac{1}{n}\sum_{k=1}^{K}$ $\max_c \left|\{i : \text{cluster}(i)\ k \wedge$ i $-$ MoA group( ) $= c\}\right|$。

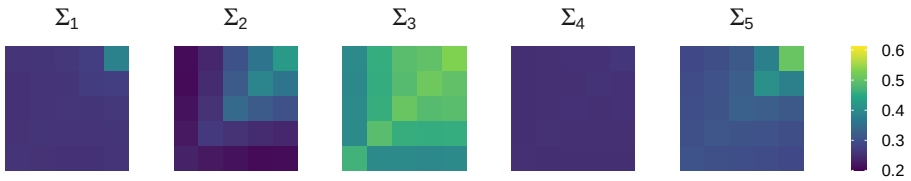


图 6：使用 **Bayes-MoE** 在癌症药物筛选数据上估计的后验均值协方差矩阵热图，其中 $K = 5$。

这种视角尤其适用于相依结构本身是主要研究对象的科学领域，例如金融共同波动、神经科学中的功能连接，以及药理学中的药物剂量-反应关系。

所提出的 MoE–Wishart 模型的一项关键贡献，不仅在于它能够通过概率门控网络实现协变量依赖聚类，还在于开发了量身定制的推断算法，使这类建模在实践中切实可行。通过使混合比例随观测预测变量变化，该模型能够捕捉潜在协方差状态的系统性变化，同时兼具可解释性和灵活性，超越了权重固定的标准混合模型。为支持这一结构，我们开发了专门的 MCMC算法，以适应对称正定矩阵的几何特性并实现稳定的后验推断；此外，我们还提出了一种新颖的期望最大化过程，用于最大

似然估计，适用于无法进行闭式更新的专家混合框架。模拟和真实数据应用的结果表明，与现有的基于EM的方法相比，这种模型设计与算法创新相结合的方案提高了稳定性和推断一致性。

在癌症药物筛选分析中，对完整的剂量-剂量协方差描述符进行建模，保留了药物效应如何随浓度共同变化的丰富信息；与曲线下面积等标量汇总描述符相比，这种方法也能实现更有意义的药物分层。在所考察的所有方法中，含有 $K = 5$ 个簇的贝叶斯MoE–Wishart模型拟合稳定，且生成的簇与整理后的药物作用生化机制高度吻合，能够依据共同的药理学特征，将RTK 抑制剂以及DNA损伤剂或烷化剂合理地归为一组。相比之下，基于期望最大化的混合方法产生的分组较为零散，且在生物学上缺乏一致性。这些结果表明，在贝叶斯MoE框架中联合建模药物剂量-反应协方差结构和药物层面协变量，为从大规模筛选数据中发现机制导向的药物类别提供了一种稳健且易于解释的方法。

若干局限性为未来研究指明了有前景的方向。首先，Wishart似然对协方差结构施加了较强的假设，例如单峰性和有限的尾部行为；在存在极端依赖模式或重尾变异的情形下，这些假设可能并不适用。扩展到更灵活的矩阵变元分布或非参数贝叶斯混合模型，或可缓解这些限制。其次，对于超高维协方差矩阵或大量组分而言，可扩展性仍是一项挑战，因此有必要开发更高效的抽样方案或变分近似。最后，更丰富的门控网络（可能纳入非线性或层次化的协变量效应）能够进一步增强MoE–Wishart框架的表达能力，并拓宽其在日益复杂的结构化数据中的适用范围。未来研究还将探讨其理论性质 [15, 18, 17, 16, 14, 13]。

尽管专家混合框架在协方差描述符层面纳入了协变量，却舍弃了个体层面信息（例如患者特异性的临床或基因组协变量）。一种自然的扩展是监督式分层高斯过程（GP）函数回归，其中协变量控制均值，高斯过程则捕捉输入域上的平滑性，并结合残差协方差建模（例如 Wishart）和潜在簇来表征亚群体间的相似性。另一种方法是直接在高斯过程上构建混合模型，例如使用谱混合核 [22], 高斯过程混合模型 [31],或高斯过程专家混合模型 [25], ，以建模具有复杂协变量效应的多模态、异质性函数响应。

致谢

本文所述观点、发现和意见仅代表作者本人，不反映挪威公共卫生研究所的官方立场。Z.Z. 得到了欧盟 Horizon 2020 研究与创新框架计划下 ERA-NET联合资助计划的 ERA PerMed 项目资助（项目“SYMMETRY”，编号 ERAPERMED2021-330）。

利益冲突/竞争性利益

作者声明不存在潜在利益冲突。

参考文献

[1] Ashburn, T. T. 和 Thor, K. B.（2004）。药物再定位：识别并开发现有药物的新用途。《自然综述：药物发现》，3(8):673–683。

[2] Biernacki, C., Celeux, G. 和 Govaert, G.（2002）。利用积分完备似然评估用于聚类的混合模型。《IEEE模式分析与机器智能汇刊》，22(7):719–725。

[3] Cappozzo, A. 和 Casa, A.（2025）。通过惩罚 Wishart 混合模型对协方差矩阵进行基于模型的聚类。《计算统计与数据分析》，212:108232。[4] Chaudhari, N., Mitra, S. K., Chirakkal, S., Mandal, S., Putrevu, D. 和 Misra, A.（2021）。使用 Wishart 混合模型和极化 SAR 数据区分多作物场景。《应用遥感杂志》，15(3):034514–034514。[5] Fruhwirth-Schnatter, S., Celeux, G. 和 Robert, C. P.（2019）。《混合分析手册》。CRC 出版社。[6] Gallaugher, M. P. 和 McNicholas, P. D.（2018）。偏斜矩阵变量分布的有限混合模型。《模式识别》，80:83–93。[7] Gao, H., Zhang, T., Wu, Y., Jiang, L., Zhan, J., Li, J. 和 Yang, R.（2014）。基于主成分分析和残差协方差矩阵的多性状全基因组关联研究。《遗传》，113(6):526–532。[8] Hidot, S. 和 Saint-Jean, C.（2010）。Wishart 混合模型的期望最大化算法：在运动聚类中的应用。《模式识别快报》，31(14):2318–2324。[9] Huang, K., Lalagkas, P. N., Sultan, B. 和 Melamed, R.（2025）。利用药物分子效应的相似性实现药物再利用。《人类基因组学》，19:110。[10] Jiang, W. 和 Tanner, M. A.（1999）。论专家混合模型的可识别性。《神经网络》，12(9):1253–1258。[11] Johansson, K., Ogut, M. G., Pelger, M., Schmelzer, T., Boyd, S. 等（2023）。一种预测金融收益协方差矩阵的简便方法。《计量经济学基础与趋势》®《计量经济学》，12(4):324–407。[12] Jordan, M. I. 和 Jacobs, R. A.（1994）。分层专家混合模型与 EM 算法。《神经计算》，6(2):181–214。[13] Mai, T. T.（2024a）。利用马蹄先验估计高维精度矩阵的稀疏贝叶斯模型集中性质。《统计》，13(4):e70008。[14] Mai, T. T.（2024b）。利用经验铰链损失的指数加权进行高维稀疏分类。《荷兰统计学》，78(4):664–691。[15] Mai, T. T.（2025a）。1 位矩阵补全中分数后验的集中性质。《机器学习》，114(1):7。[16] Mai, T. T.（2025b）。高维有界响应的处理：一种马蹄先验贝叶斯 Beta 回归方法。arXiv 预印本 arXiv:2505.22211。[17] Mai, T. T.（2025c）。采用马蹄先验对删失响应进行高维贝叶斯 Tobit 回归。arXiv 预印本 arXiv:2505.08288。[18] Mai, T. T.（2025d）。广义降秩回归中分数后验的性质。《多元分析杂志》，第105481页。[19] Mai, T. T., Rønneberg, L., Zhao, Z., Zucknick, M. 和 Corander, J.（2019）。利用具有局部低秩结构的多组学数据，从大规模筛选中学习癌症药物敏感性。载于《国际生物信息学与生物统计学计算智能方法会议》，第67–79页。施普林格。

[20] Nguyen, H. D. 和 Chamroukhi, F.（2018年）。专家混合建模的实践与理论：综述。*Wiley*跨学科评论：数据挖掘与知识发现，8(4):e1246。[21] Nielsen, S. F., Madsen, K. H., Schmidt, M. N. 和 Mørup, M.（2017年）。使用 Wishart 混合模型对动态功能连接进行建模。载于*2017*年国际神经影像模式识别研讨会（*PRNI*），第1–4页。IEEE。[22] Parra, G. 和 Tobar, F.（2017年）。多输出高斯过程的谱混合核。载于第*31*届神经信息处理系统国际会议论文集，NIPS'17，第6684–6693页，美国纽约州雷德胡克。Curran Associates Inc. [23] Peng, F., Jacobs, R. A. 和 Tanner, M. A.（1996年）。专家混合模型和分层专家混合模型中的贝叶斯推断及其在语音识别中的应用。美国统计学会杂志，91(435):953–960。[24] Quandt, R. E. 和 Ramsey, J. B.（1978年）。正态分布混合模型和切换回归的估计。美国统计学会杂志，73(364):730–738。[25] Rasmussen, C. 和 Ghahramani, Z.（2001年）。高斯过程专家的无限混合。载于 Dietterich, T., Becker, S. 和 Ghahramani, Z.（编），神经信息处理系统进展，第14卷。麻省理工学院出版社。[26] Runcie, D. E. 和 Mukherjee, S.（2013年）。通过对遗传协方差矩阵进行贝叶斯稀疏因子分析，剖析高维表型。《遗传学》，194(3):753–767。[27] Sch¨afer, J. 和 Strimmer, K.（2005年）。大规模协方差矩阵估计的收缩方法及其对功能基因组学的影响。《遗传学与分子生物学统计应用》，4(1):32。[28] Seashore-Ludlow, B., Rees, M. G., Cheah, J. H., Cokol, M., Price, E. V., Coletti, M. E., Jones, V., Bodycombe, N. E., Soule, C. K., Gould, J., Alexander, B., Li, A., Montgomery, P., Wawer, M. J., Kuru, N., Kotz, J. D., Hon, C. S.-Y., Munoz, B., Liefeld, T., Danˇc´ık, V., Bittker, J. A., Palmer, M., Bradner, J. E., Shamji, A. F., Clemons, P. A. 和 Schreiber, S. L.（2015年）。利用大型小分子敏感性数据集中的连接性。《癌症发现》，5(11):1210–1223。[29] Sharifi-Noghabi, H., Jahangiri-Tazehkand, S., Smirnov, P., Hon, C., Mammoliti, A., Nair, S. K., Mer, A. S., Ester, M. 和 Haibe-Kains, B.（2021年）。基于细胞系的药物基因组学数据进行药物敏感性预测：开发机器学习模型的指南。《生物信息学简报》，22(6):bbab294。[30] Sun, R., Ma, T., Liu, S. 和 Sathye, M.（2019年）。投资组合风险度量中的协方差矩阵估计改进：综述。《风险与金融管理杂志》，12(1):48。[31] Tresp, V.（2000年）。高斯过程混合模型。载于 Leen, T., Dietterich, T. 和 Tresp, V.（编），神经信息处理系统进展，第13卷。麻省理工学院出版社。[32] Tuzel, O., Porikli, F. 和 Meer, P.（2006年）。区域协方差：一种用于检测和分类的快速描述符。载于 Leonardis, A., Bischof, H. 和 Pinz, A.（编），计算机视觉——*ECCV 2006*，第589–600页。Springer Berlin Heidelberg。[33] Vehtari, A., Gelman, A. 和 Gabry, J.（2017年）。使用留一交叉验证和 WAIC 进行实用的贝叶斯模型评估。《统计学与计算》，27(5):1413–1432。

[34] Xue, H., Li, J., Xie, H. 和 Wang, Y.（2018年）。药物再利用方法与资源综述。国际生物科学杂志, 14(10):1232–1244。

[35] 曾, A. G. X., 班萨尔, S., 金, L., 米切尔, A., 陈, W. C., 阿巴斯, H. A., Chan-Seng-Yue, M., Voisin, V., van Galen, P., Tierens, A., Cheok, M., Preudhomme, C., Dombret, H., Daver, N., Futreal, P. A., Minden, M. D., 肯尼迪, J. A., 王, J. C. Y., 和 Dick, J. E. (2022). 用于理解急性髓系白血病异质性并预测药物反应的细胞层级框架。《自然·医学》, 28(6):1212–1223.

[36] 赵, Z., 王, S., Zucknick, M., 和 Aittokallio, T. (2022). 用于泛癌药物反应预测的多组学特征组织特异性识别。 *iScience*, 25(8):104767.

[37] Zhao, Z. 和 Zucknick, M.（2020）。用于药物敏感性预测的结构化惩罚回归。英国皇家统计学会杂志*C*辑：应用统计，69(3)：525–545。

A 算法详情

在第 3.1 节中 对于混合模型的 Gibbs-within-MH采样器，我们可以积分消去 $\boldsymbol{\pi}$ 在公式（6):

$$
\begin{aligned}
& p(z_i = k \mid z_{-i}, S_i, \{\Sigma_\ell, \nu_\ell\}, \boldsymbol{\alpha}) \\
\propto & \int \pi_k f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot \mathrm{Dir}(\boldsymbol{\pi} | z_{-i}, \boldsymbol{\alpha}) d\boldsymbol{\pi} \\
\propto & \int \pi_k f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot \mathrm{Dir}(\boldsymbol{\pi} | \boldsymbol{\alpha}) \cdot \mathrm{Cat}(z_{-i} | \boldsymbol{\alpha}) d\boldsymbol{\pi} \\
\propto & \int \pi_k f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot \mathrm{Dir}(\boldsymbol{\pi} | \boldsymbol{a}) d\boldsymbol{\pi}, \quad \text{where } a_j = \alpha_j + n_{-i,j},\ n_{-i,j} := \#\{l \neq i | z_l = j\} \\
= & f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \int \pi_k \prod_{j=1}^{K} \frac{\pi_j^{a_j - 1}}{B(\boldsymbol{a})} d\boldsymbol{\pi} \\
= & f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \frac{1}{B(\boldsymbol{a})} \int \pi_1^{a_1 - 1} \pi_2^{a_2 - 1} \cdots \pi_k^{a_1} \cdots \pi_K^{a_K - 1} d\boldsymbol{\pi} \\
= & f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \frac{1}{B(\boldsymbol{a})} B(a_1, ..., a_k + 1, ..., a_K) \int d\mathbb{P}_{\boldsymbol{\pi} \sim \mathrm{Dir}(a_1, ..., a_k + 1, ..., a_K)} \\
= & f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot a_k / \sum_{j=1}^{K} a_j \\
\propto & f_{\mathcal{W}}(S_i \mid \nu_k, \Sigma_k) \cdot (\alpha_k + n_{-i,k})
\end{aligned}
$$

与从公式（6）中的条件分布进行采样相比，这会带来更好的 MCMC混合。

B 其他模拟结果

表 S3：混合工作模型和专家混合工作模型的有效样本量（ESS）。数据由专家混合模型生成。报告的是 100 次模拟中所选参数 ESS 的均值（标准差）。 ν_1 和 ν_2 是前两个簇对应的 Wishart 分布的自由度。$\Sigma_{1,11}$ 和$\Sigma_{2,11}$ 是相应尺度矩阵的 (1，1) 元素。 β_{11}和 β_{12} 是前两个簇的药物状态效应。

工作模型	$\boldsymbol{\nu}$		Σ		β	
	ν_1	ν_2	$\Sigma_{1,11}$	$\Sigma_{2,11}$	β_{11}	β_{12}
混合模型：$p=2$						
$n=200$	723 (203)	169 (53)	1179 (435)	216 (78)	-	-
$n=500$	784 (113)	169 (37)	1243 (343)	214 (54)	-	-
$n=1000$	790 (124)	169（28）	1155（317）	200（40）	-	-
混合模型：$p=8$						
$n=200$	1397（92）	722（57）	8885 (743)	3222 (426)	-	-
$n=500$	1484 (72)	814 (54)	8908 (829)	3591 (304)	-	-
$n=1000$	1450 (85)	761 (49)	9495 (639)	3453 (346)	-	-
混合专家模型：$p=2$						
$n=200$	996 (168)	251 (75)	1289 (483)	313 (120)	133 (71)	131 (75)
$n=500$	1023 (103)	223 (43)	1366 (345)	263 (57)	164 (40)	203 (71)
$n=1000$	1009 (82)	233 (38)	1438 (312)	271 (52)	179 (53)	187 (48)
混合专家模型：$p=8$						
$n=200$	1342 (90)	752 (53)	8644 (1009)	3333 (368)	407 (106)	418 (111)
$n=500$	1495 (78)	814 (47)	8843 (449)	3475 (416)	442 (84)	446 (91)
$n=1000$	1458 (82)	796 (44)	9514 (871)	3588 (329)	469 (108)	439 (83)

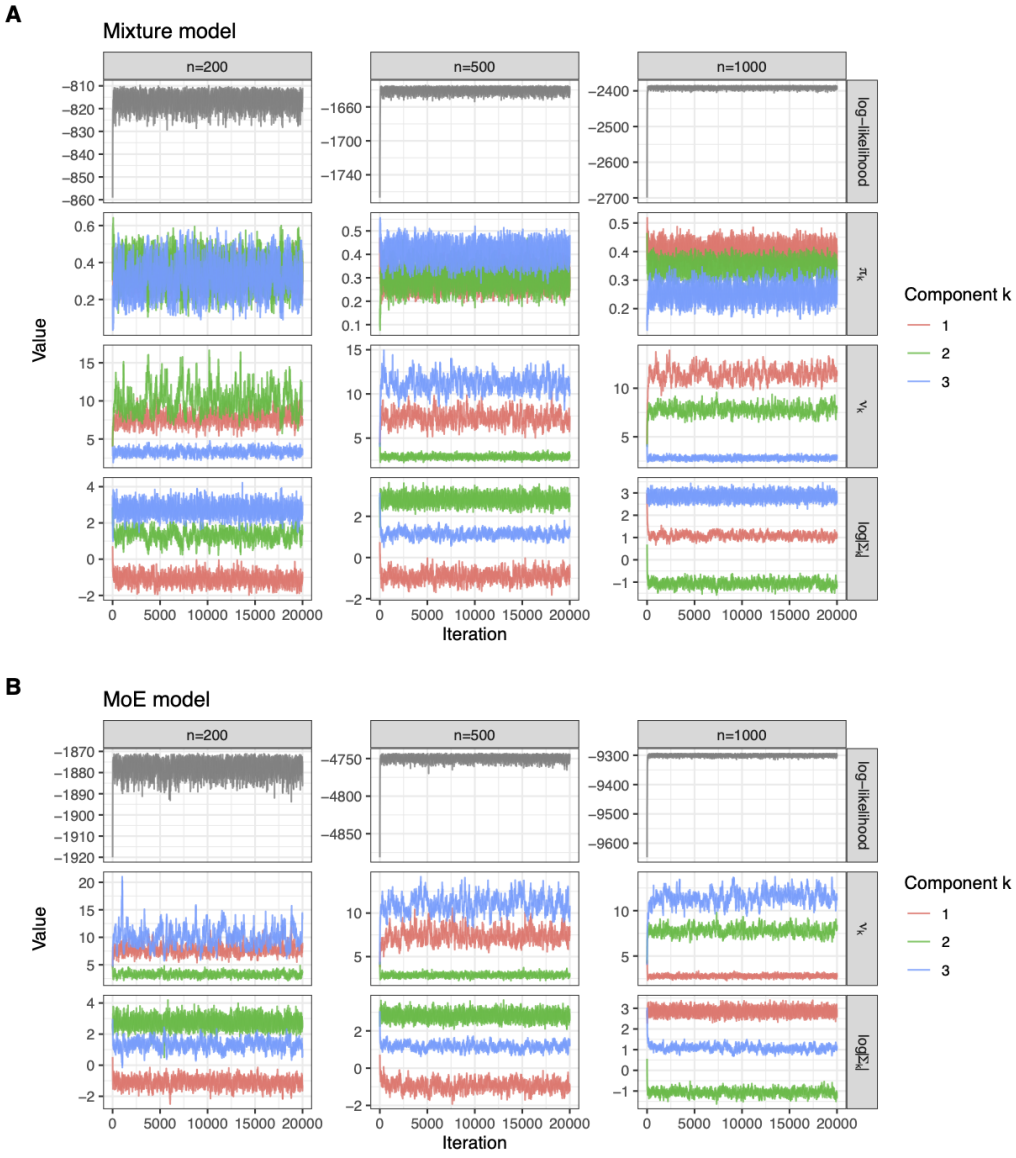


图 S7：由**混合模型** 生成的模拟数据的马尔可夫链蒙特卡洛诊断轨迹图，其中$\boldsymbol{p = 2}$和 $n = (200, 500, 1000)$，分别使用贝叶斯混合模型和专家混合工作模型进行拟合。(A) 混合模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断指标为模型的对数似然、 π_k、 ν_k 和 $\log\ |\Sigma_k|$、 $k \in \{1, 2, 3\}$。(B) 混合专家模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断指标为模型的对数似然、 ν_k 和 $\log\ |\Sigma_k|$、 $k \in \{1, 2, 3\}$。

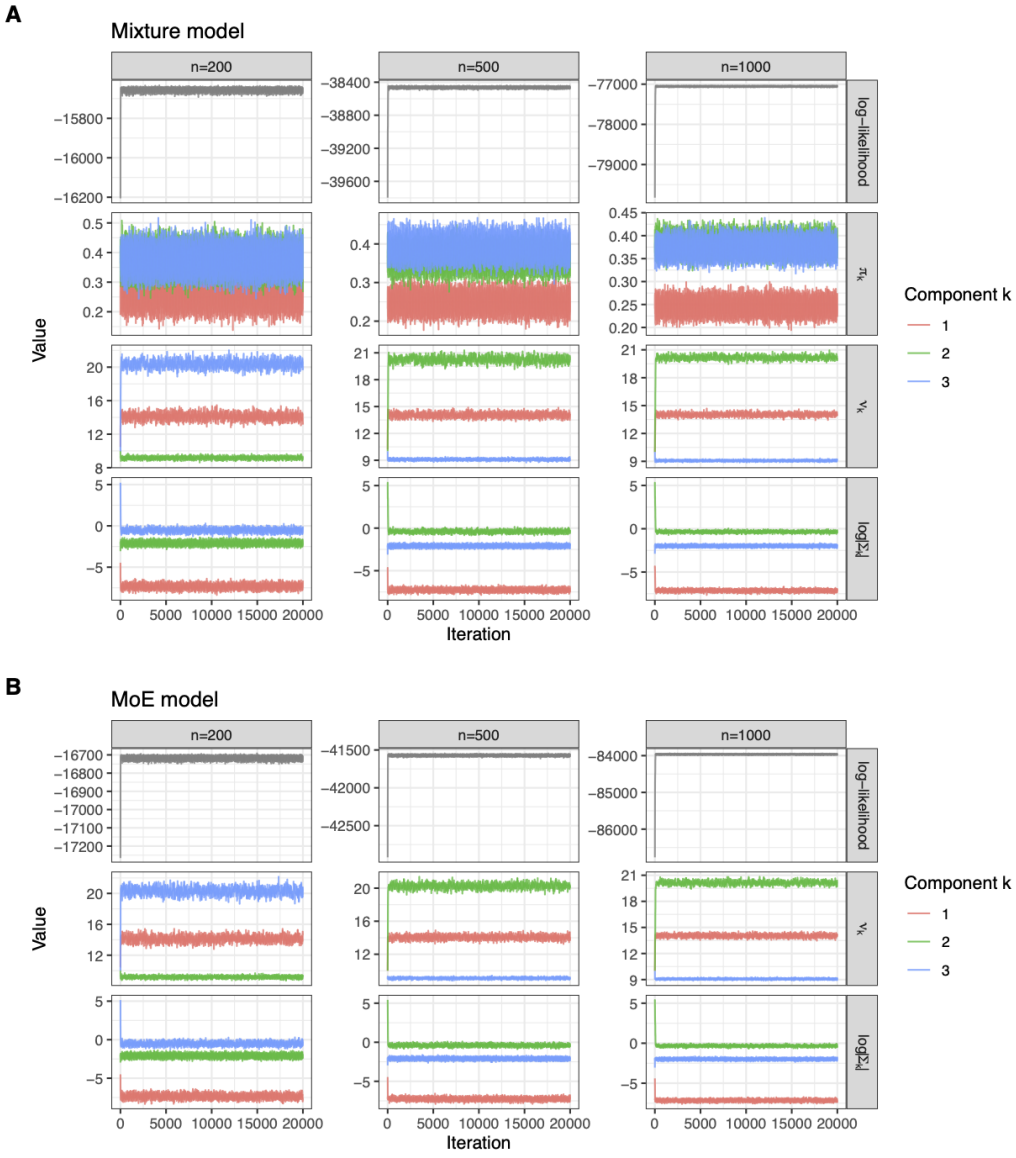


图 S8：由**混合模型**生成的模拟数据的马尔可夫链蒙特卡洛诊断轨迹图，其中$\boldsymbol{p} = \mathbf{8}$和 $n =$(200，500，1000)，分别使用贝叶斯混合模型和专家混合工作模型进行拟合。(A) 混合模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 π_k、 ν_k 和对数 $|\Sigma_k|$、 $k \in \{1, 2, 3\}$。(B) 混合专家模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 ν_k 和对数 $|\Sigma_k|$、 $k \in \{1, 2, 3\}$。

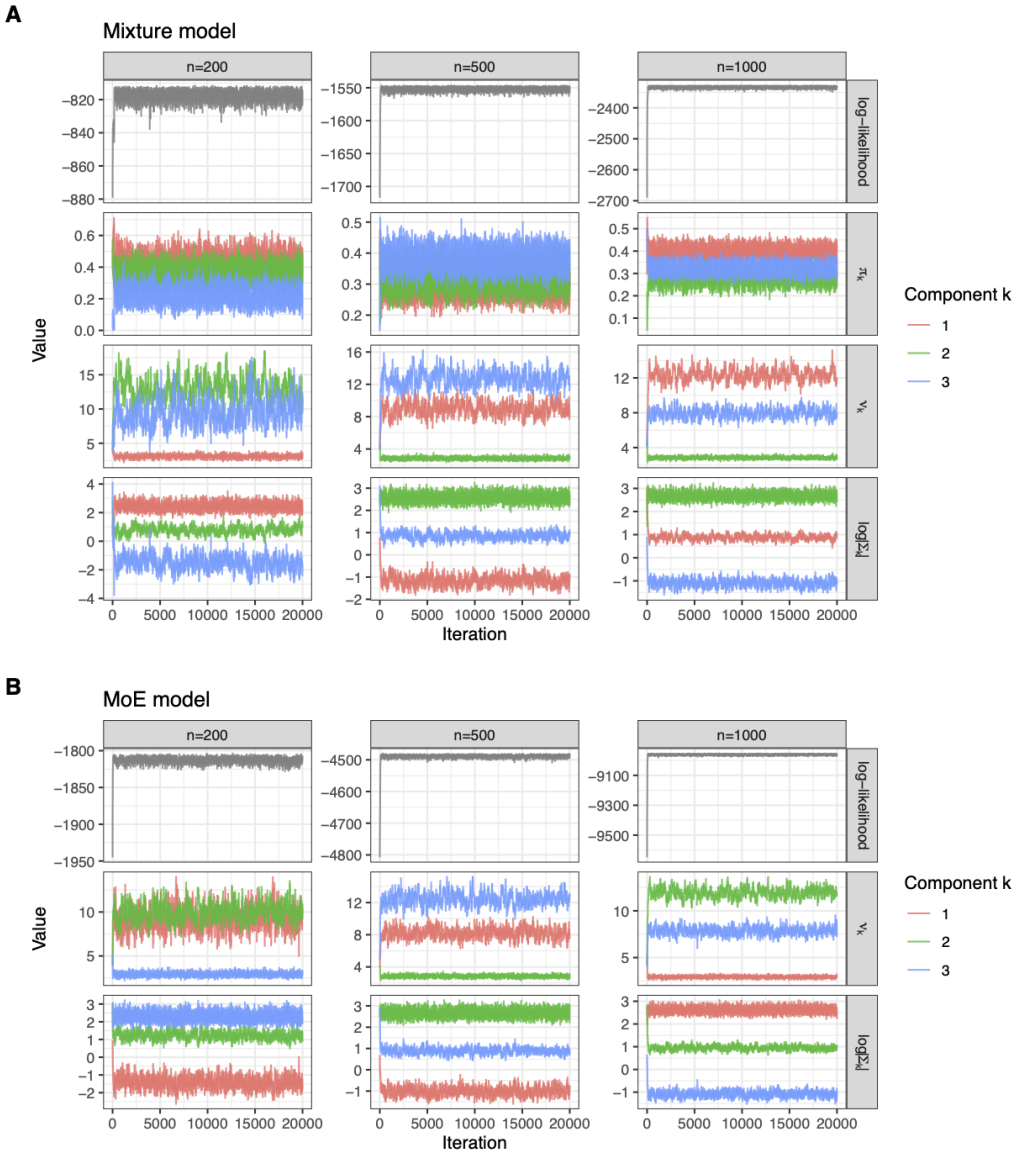


图 S9：由**专家混合模型**生成的模拟数据的马尔可夫链蒙特卡洛诊断轨迹图，其中 $\boldsymbol{p = 2}$和 $n =$(200，500，1000)，使用贝叶斯混合模型和专家混合工作模型进行拟合。(A) 混合模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 π_k, ν_k 和 $\log|\Sigma_k|, k \in \{1, 2, 3\}$。(B) 专家混合模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 ν_k和 $\log\ |\Sigma_k|$, $k \in \{1, 2, 3\}$。

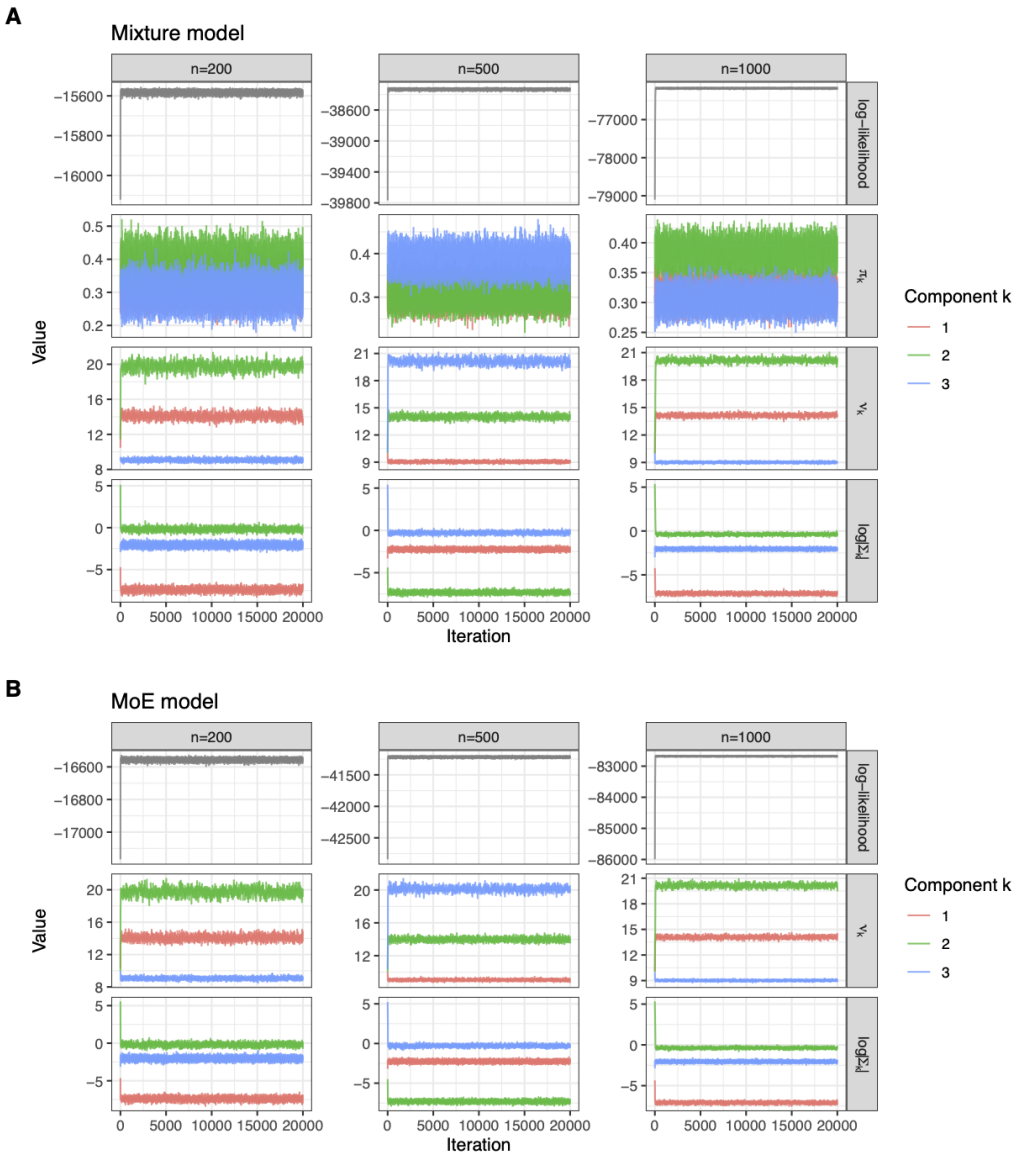
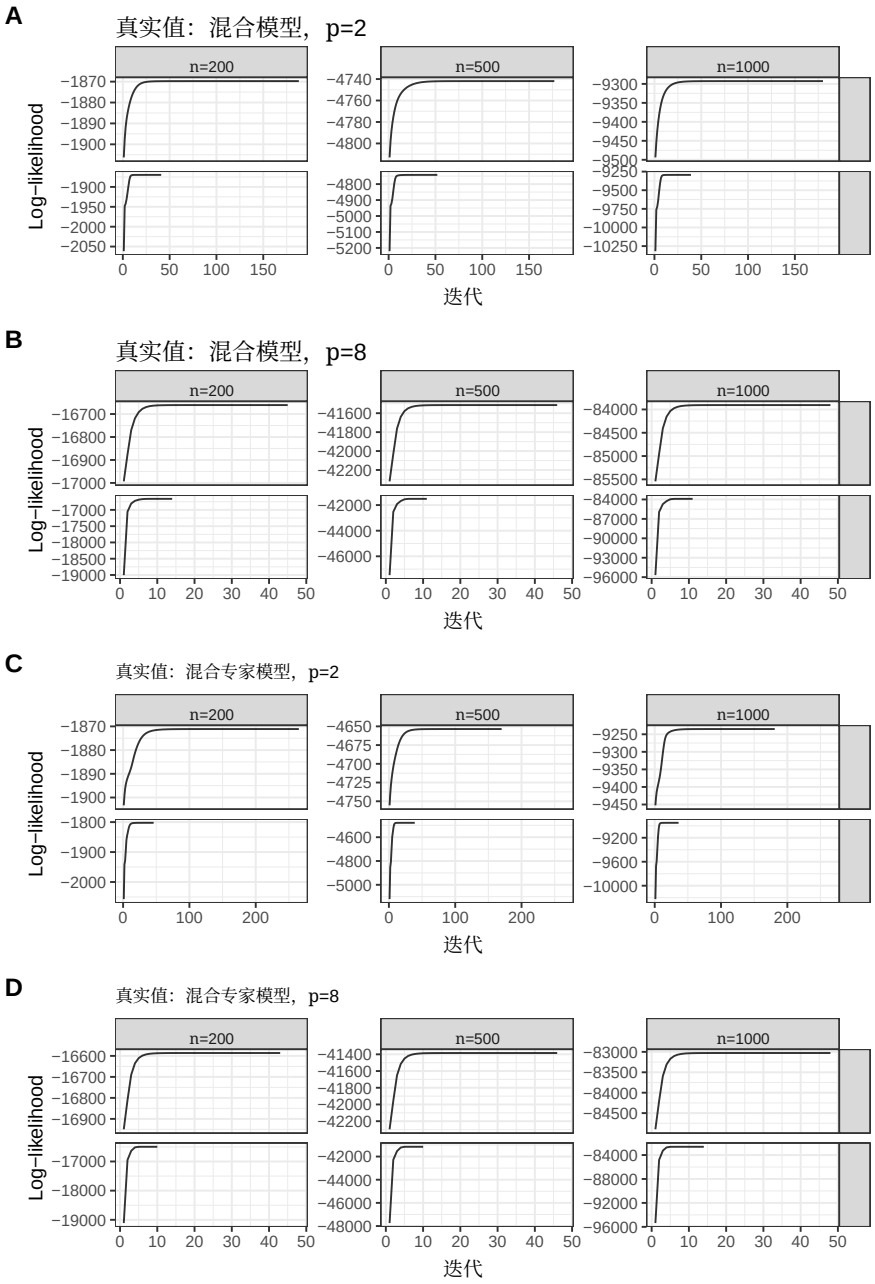


图 S10：由**混合专家模型**生成的模拟数据的马尔可夫链蒙特卡洛诊断轨迹图，其中$\boldsymbol{p = 8}$和 $n =$(200，500，1000)，分别使用贝叶斯混合模型和专家混合工作模型进行拟合。(A) 混合模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 π_k、 ν_k 和 $\log\ |\Sigma_k|$、 $k \in \{1, 2, 3\}$。(B) 混合专家模型拟合的马尔可夫链蒙特卡洛诊断。对于每个 n，诊断对象为模型的对数似然、 ν_k和 $\log\ |\Sigma_k|$、 $k \in \{1, 2, 3\}$。



图S11：对于从混合模型或混合专家模型（真实模型）生成的模拟数据，EM算法每次迭代的对数似然。面板 (A)–(D) 中，每个面板的列对应模拟数据 $n =$(200，500，1000)，行对应混合模型和混合专家模型的拟合结果（分别使用 EM 和 EM-MoE）。(A) 真实基础模型是维度为 $p = 2$ 的混合模型。(B) 真实基础模型是维度为 $p = 8$ 的混合模型。(C) 真实基础模型是维度为 $p = 2$ 的混合专家模型。(D) 真实基础模型是维度为 $p = 8$ 的混合专家模型。

odels with $n = 1000$, $p = 8$
Bayesian working models
and ICL (both smaller is

6
27.3 (577.6)
29.8 (1155.3)
20.5 (577.7)
96.3 (1144.5)
14.5 (1171.1)
86.0 (1168.8)
02.9 (1155.2)
96.2 (1191.2)
69.8 (537.0)
21.5 (1074.1)
58.0 (537.2)
51.4 (1072.0)
69.9 (1080.3)
51.0 (1085.3)
52.2 (1073.2)
91.9 (1089.9)

C 药物反应数据分析的其他结果

表 S5：在癌症药物反应数据上选择聚类数 K。使用 elpd_{loo}（越大越好）和 ICL（越小越好）评估贝叶斯方法（**贝叶斯**和**Bayes-MoE**）。使用贝叶斯信息准则和 ICL（越小越好）评估期望最大化方法（**期望最大化** 和**EM-MoE**）。每种方法的最优值以粗体显示。

K	2	3	4	5	6	7	8	9	10	11
贝叶斯										
elpd_{loo}	8269.6	9369.8	9507.2	**9700.2**	9696.9	**9700.4**	9698.7	-	-	-
ICL	-16371	-18461.6	-18628.7	**-18892.5**	-18775.6	-18670.7	-18554.1			
Bayes-MoE										
elpd_{loo}	6054.9	7158.1	7283.5	7463.4	**7466.0**	7462.2	7465.1	-	-	-
ICL	-11560.8	-13324.8	-13397.9	**-13549.6**	-13438.6	-13316.6	-13202.9	-	-	-
EM										
BIC	-12156.4	-15347.2	-16034.7	-16591.1	-16769	-17365.2	-17508.3	-17359.9	**-17742.7**	-17595.2
ICL	-12139.3	-15335.2	-16007.3	-16555.2	-16728.2	-17336.9	-17473.2	-17328.8	**-17704.9**	-17555.8
EM-MoE										
BIC	-12167.3	-15369.4	-15927.1	-16646.1	-17102.4	-17374.4	-17516.8	-17768.2	-17810.3	**-17874.0**
ICL	-12150.3	-15357.5	-15900.4	-16614.8	-17072.1	-17343	-17479.6	-17731.4	-17770.0	**-17839.7**

表 S6：癌症药物敏感性真实数据的补充结果。采用四种竞争方法进行参数估计。两种混合模型为**贝叶斯**（含 $K = 5$ 个簇）和**期望最大化**（含 $K = 10$ 个簇）。两种混合专家模型为**Bayes-MoE**（含 $K = 5$ 个簇）和**EM-MoE**（含 $K = 11$ 个簇），并纳入了两个伴随协变量：(i) 编码为二元变量的药物状态，以及 (ii) 由药物 SMILES 衍生指纹矩阵的主成分分析得到的第一主成分。对于**期望最大化** 和**EM-MoE**方法，报告所选参数的点估计；对于**贝叶斯** 和**Bayes-MoE**方法，则报告后验均值（95% 可信区间）。 ν_1 和 ν_2 分别是前两个簇的 Wishart 分布的自由度。$\Sigma_{1,\ 11}$ 和 $\Sigma_{2,\ 11}$ 分别是前两个簇的 Wishart 分布尺度矩阵的首个元素。π_1 和 π_2 分别是前两个簇的混合权重。 β_{11} 和 β_{12} 分别是前两个簇的药物状态效应。

参数	贝叶斯	EM		Bayes-MoE	EM-MoE
ν_1	10.9 (10.3, 11.4)	31.4		10.8 (10.3, 11.4)	33.6
ν_2	12.4 (11.0, 13.9)	37.2		12.4 (10.9, 14.0)	24.5
$\Sigma_{1,11}$	0.3 (0.24, 0.28)	0.09		0.3 (0.24, 0.28)	0.08
$\Sigma_{2,11}$	0.2 (0.19, 0.28)	0.08		0.2 (0.19, 0.28)	0.11
π_1	0.4（0.32, 0.42）	0.06	β_{11}	-0.2（-0.87, 0.51）	0.38
π_2	0.1（0.05, 0.11）	0.10	β_{12}	-1.0（-2.01, 0.05）	-0.34

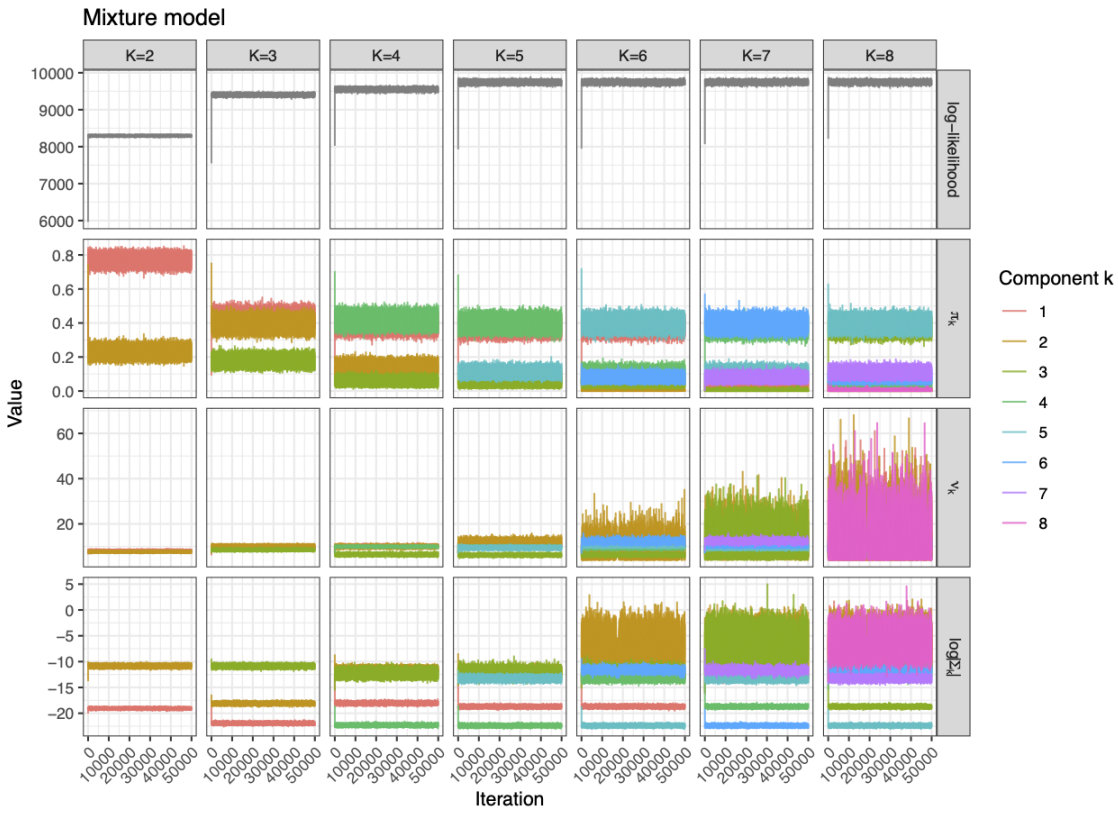
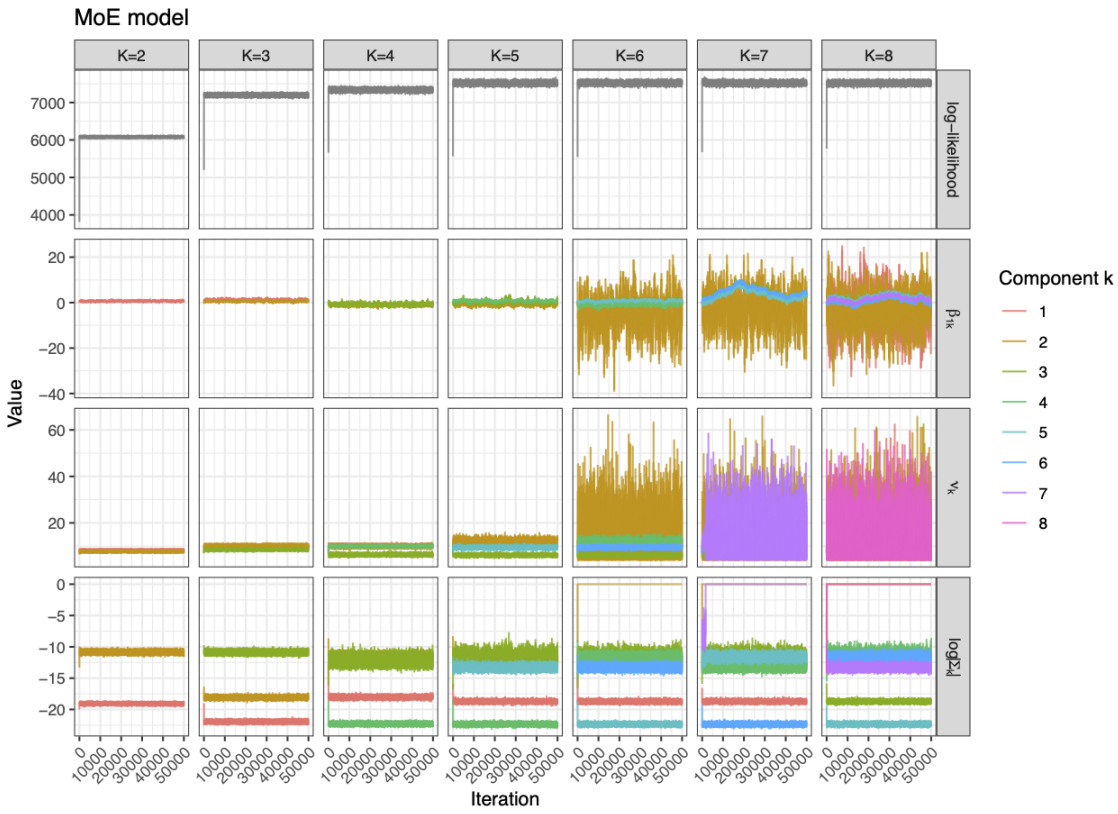


图 S12：癌症药物敏感性真实数据的补充结果。使用指定的 $K \in \{2,3,4,5,6,7,8\}$ 拟合药物敏感性数据所得参数的马尔可夫链蒙特卡洛诊断轨迹图，拟合采用**混合模型**。第一行展示模型对数似然相对于 K 的轨迹图。第二行展示 π_k（混合权重）、$k = 1, 2, \ldots\ldots, K$ 相对于 K 的轨迹图。第三行展示 ν_k（威沙特分布的自由度）相对于 K 的轨迹图。第四行展示 $\log|\Sigma_k|$（Σ_k 为威沙特分布的尺度矩阵）相对于 K 的轨迹图。图例以不同颜色表示组分/聚类索引 k。



图S13：癌症药物敏感性真实数据的附加结果。针对由**专家混合模型** 在指定 $K \in \{2,3,4,5,6,7,8\}$下拟合的药物敏感性数据，用于参数马尔可夫链蒙特卡洛诊断的轨迹图。第一行显示模型对数似然关于 K的轨迹图。第二行显示 β_{1k}（药物状态效应）、 $k = 1$，2，..... K关于 K的轨迹图。第三行显示 ν_k（威沙特分布的自由度）关于 K的轨迹图。第四行显示$\log|\Sigma_k|$（Σ_k 为威沙特分布的尺度矩阵）关于 K的轨迹图。图例以不同颜色表示组分/聚类索引 k。