arXiv:2602.02931v1 [stat.ME] 3 Feb 2026

聚类数据的加权树和模型

arXiv 预印本

Kevin McCoy
统计学系
莱斯大学
休斯敦，德克萨斯州 77005
kmm12@rice.edu

Zachary Wooten
生物统计学系
圣裘德儿童研究医院
孟菲斯，TN 38105
zachary.wooten@stjude.org

Katarzyna Tomczak
转化分子病理学系
德克萨斯大学MD安德森癌症中心
休斯敦，德克萨斯州 77030
kjtomczak@mdanderson.org

Christine B. Peterson*
统计学系
莱斯大学
休斯敦，德克萨斯州 77005
cbpeterson@rice.edu

2026年2月4日

摘要

当观测值嵌套于各组之中时，便会产生聚类数据，这种数据在临床、教育和社会科学研究中极为常见。传统上，人们会使用包含随机效应以处理组内相关性的线性混合模型，对观测数据进行建模，并对未见过的数据作出新的预测。已有研究尝试将混合模型方法从线性回归扩展到更复杂的非参数模型，例如决策树和随机森林。然而，现有方法在对来自样本外组的数据进行预测时，只能使用全局固定效应，实际上假设所有聚类共享同一个结果模型。我们提出一种轻量级树之和模型，为每个样本组学习一棵决策树。我们对这些树的预测加权组合，使样本外组的预测更贴近训练数据中最相似的组。这一策略还支持推断各组在结果预测模型方面的相似性，因为可以直接比较每个组独有的树结构和变量重要性。我们表明，在多种模拟设置下，我们的模型优于传统决策树和随机森林。最后，我们将该方法应用于癌症基因组图谱的肉瘤队列真实数据，其中患者样本按肉瘤亚型分组。

关键词 决策树 · 随机森林 · 混合效应建模 · 加权集成 · 聚类数据 ta

1 引言

决策树是使用最广泛的机器学习方法之一。其最流行的实现，即分类与回归树（CART）算法，由Breiman等人提出 [2017]。决策树通过递归地将数据空间划分为矩形分区来构建。尽管决策树在众多场景中表现良好，但它们往往会变得过大，并开始过拟合训练数据。为避免这一问题，人们采用集成树方法，例如随机森林（随机森林）或提升模型。随机森林使用自助法并行拟合树，以降低总体模型方差 [Breiman, 2001],而提升方法则对残差依次拟合树，从而迭代改进模型并降低偏差 [Hastie等人, 2001]。决策树及其扩展方法还能容忍缺失数据，具有可解释性，对共线性较为稳健，并内置特征选择。

在临床、教育和社会科学研究中，分组数据十分常见。例如，在教育研究中，学生可能按教室或学校分组。在生物医学研究中，患者可能按

*应将通信寄给的作者 ed.

医院或疾病亚型分组。在这些情况下，由于组内相关性（或称组内类相关性），观测值不再能被视为相互独立 [Jiang and Nguyen, 2021]。尽管观测值之间的独立性并非决策树和其他机器学习模型始终必须满足的严格假设，但在模型构建过程中考虑分组结构可以提升预测性能。当需要预测来自训练样本之外的组的观测值结果时，妥善处理分组观测值尤为重要。例如，我们可能希望对就读于未纳入训练数据集的学校的学生进行预测。

聚类数据只是非独立数据结构的一个例子，这类数据通常使用混合效应模型处理。混合效应指的是这样一种概念：某些关联（称为固定效应）在总体中呈现一致的模式，而数据的其他属性则对结果产生随机效应，导致同一单位的多个观测值之间存在相关性。例如，对于对个别患者进行重复测量的情形，混合模型可能会为每位患者设置一个随机效应；对于患者按医院或临床医生聚类的情形，这些因素则会由随机效应表示。已有研究尝试将混合效应纳入用于聚类数据或纵向数据的基于树的模型中 [Hajjem 等人，2011，Spanbauer and Sparapani， 2021]。然而，现有方法存在混合效应模型本身固有的局限。具体而言，混合模型的样本外预测仅使用固定效应，因为没有来自这些组的训练数据时，无法学习随机效应。许多方法还采用僵化的加法结构来组合固定效应和随机效应，类似于线性混合模型，这削弱了决策树的灵活性。

在第 1.1 节中，我们讨论现有的基于树的聚类数据模型及其局限性。在第 2 节中，我们介绍一种轻量级方法，该方法为每个组学习一棵独特的决策树。随后，根据从训练数据中的各组学习到的决策树所产生的预测的加权线性组合，计算样本外组的预测。在第 3.1 节中，我们重点介绍模拟研究；这些研究表明，在一系列情形下，我们的方法优于标准决策树、随机森林和线性混合模型。最后，在第 3.2 节中，我们展示了该方法在一项真实数据应用中的竞争力：预测训练数据中未出现的肉瘤亚型的 T 细胞丰度。

1.1 现有方法的背景

广义线性混合模型（GLMM）是处理聚类数据情形最常用的方法。在 GLMM 中，我们拟合模型 $g(\mathbb{E}[y|X]) = X\beta + Z\alpha$，其中 $g(\cdot)$ 表示链接函数，y 对应结果变量，X 表示固定效应的设计矩阵，Z 则是随机效应的设计矩阵 [Breslow and Clayton， 1993]。与标准线性模型一样，GLMM 假设线性效应具有加性；这一假设可能过于简单，导致模型在复杂数据集上表现不佳，因为这类数据中可能存在取值极端的预测变量、非线性效应、交互作用，或其他偏离线性模型结构假设的情况。

混合效应回归树（MERT） [Hajjem 等人， 2011] 采用期望最大化（EM）方法迭代拟合模型 $y = f（X） + Z\alpha + \varepsilon$。固定效应 $f（X）$ 通过回归树进行拟合，随后使用标准方法更新并拟合随机效应。Hajjem 等人 [（ 2014]）通过自助法将 MERT 扩展为森林。目前，尚无针对分类结局数据的 MERT 扩展方法。

MERT 方法存在若干局限性。首先，与标准广义线性混合模型（GLMM）一样，其随机部分 $Z\alpha$ 严格呈线性且具有可加性。MERT 还假设不同聚类中的观测值相互独立，但实际情况往往并非如此。此外，固定效应和随机效应是分别估计的，这意味着固定效应与随机效应之间的任何交互作用都无法纳入模型。最后，与标准混合模型一样，MERT 只能利用固定效应进行样本外预测。

个案特异性随机森林（CSRFs） [Xu等人 2016] 为测试数据集中的每个新点构建一个独特的随机森林。与传统自助法对每个观测值采用均匀选择概率不同，距离测试个案更近的点会被赋予更大的权重。该方法的一个显著缺点是，为测试数据集中的每个样本构建新的随机森林所需的计算复杂度与时间。该方法也不能直接考虑混合效应，但在存在混合效应的情形下仍可能有用。

DeJong等人 [2021]提出内部-外部交叉验证，该方法在抽取训练集、验证集与测试数据集时纳入聚类结构，并选择在最多组外样本上表现最佳的模型。然而，最终只选出一个输出模型。

在贝叶斯框架下，Chipman 等人最初提出了贝叶斯加性回归树（BART）[2010]。Spanbauer and Sparapani [2021] 提出了 mixedBART，作为 BART 的扩展，可处理混合效应。与广义线性混合模型（GLMM）和 MERT 类似，mixedBART 拟合一个可加模型，其中固定效应和随机效应分别通过 BART 和贝叶斯混合模型依次拟合。这里，贝叶斯框架

相当方便，因为这两种方法可以轻松整合到马尔可夫链蒙特卡洛（MCMC）算法的交替步骤中。

贝叶斯加性回归树的一个局限在于其对分类变量的处理。通常，贝叶斯加性回归树会对分类变量进行独热编码，将每个唯一类别替换为其自身的二元指示变量。因此，基决策树中的每个分裂都只与该分类变量的一个水平相关联。Deshpande [2025] 最近提出了 flexBART模型，通过允许将分类变量的多个水平分配到分裂规则的任一侧，克服了这一局限。然而，贝叶斯加性回归树和 flexBART都不允许分类变量的某些水平出现在测试数据中、却未出现在训练数据中。

2 方法

2.1 提出的方法

现在，我们将介绍所提出的用于聚类数据灵活预测的方法。在开发该方法时，我们尤其关注如何为不属于训练数据中已出现的组的观测值获取预测这一挑战。

令$\mathbf{X}$ 表示一个 $n \times p$ 矩阵，对应于 p 个特征上的 n 个观测值；令 Y为一个 $n \times 1$ 维结果列向量。在训练数据集中，每个观测值都来自 $j = 1$个已知组或聚类中的一个，2，.....， J，且每个聚类内的观测值彼此相关。我们将组别分配的 $n \times 1$ 维列向量记为 C。测试数据集具有相同的结构，但组的索引现在为 $m = 1$，2，..... M。这里，我们关注测试数据中的组不属于训练数据集中出现过的组的情形。
我们的方法分两个阶段进行。首先，利用训练数据构建分类模型 f： $\mathbf{X} \rightarrow C$。采用这种方法，每个测试观测值都会得到一个预测概率向量，表示其属于 J个组中各组的概率。我们将属于某一类别的输出概率称为权重 $\boldsymbol{w} =$（w_1，.....， w_J）。我们将这些权重视为相似性度量，用于衡量测试数据点 X_t 与训练数据中各组的接近程度，且仅考虑$\mathbf{X}$ 特征空间。如果一个未见组 m中有多个观测值，我们就对该组内的相似性权重取平均，并将其作为 $\boldsymbol{w}$。

我们学习到的权重与因果推断中使用的倾向得分理念相似。不过，我们预测的不是某个观测值是否会被纳入处理组，而是新组与训练数据中已见组之间的“接近程度”。我们的方法也与堆叠集成学习 [Van der Laan et al., 2007], 有几分相似，关键区别在于：基模型是针对输入数据中的不同组分别学习得到的，并且用于组合基模型的权重因受试者而异。
在模型的第二阶段，我们为每个训练组 j学习一棵独特的决策树。随后，测试数据点 X_t 的预测结果 y_t通过对这些 J 树进行线性组合计算得出，每棵树的权重对应于其在 $\boldsymbol{w}$中的取值。

在方法的第一阶段，逻辑回归 [Hosmer Jr et al., 2013] 或朴素贝叶斯 [Hastie et al.,2001] 等模型因其概率结果易于解释而表现良好。在模型的第二阶段，可以使用任何学习器，但我们重点关注决策树，因为它们在可解释性与预测性能之间取得了良好平衡。医疗决策中通常更倾向于使用决策树，因为临床医生更容易理解它们 [Banerjee et al.,2019]。在案例研究中，我们也将随机森林作为第二阶段模型进行考察，因为它们可能带来更稳健的预测性能。**图 1** 展示了我们提出的方法。

2.2 实现

我们的加权树和模型使用 Python 实现。对于基学习器，我们采用 Scikit-learn 对相关方法的实现，并使用默认参数值 [Pedregosa 等人， 2011]。代码已公开发布在 GitHub：https://github.com/kmccoy3/weighted-trees。

3 结果

在本节中，我们比较所提出的加权树和方法与广义线性混合模型、标准决策树和随机森林在模拟数据及真实数据上的表现。我们还考察了随机森林的加权组合，它与所提出的方法相似，但在第二阶段使用随机森林而非决策树。由于没有公开可用的 MERT 实现，我们无法将本方法与 MERT 进行比较。

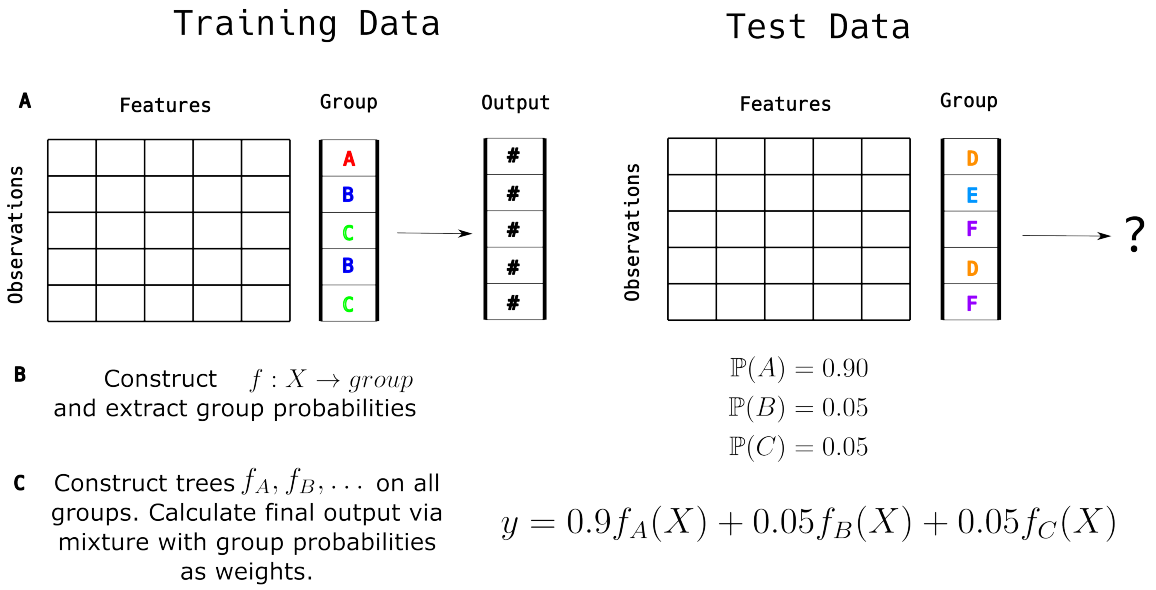


图 1：（**A**）考虑这样的分组数据：每个训练观测都属于某个组（例如，按医院或疾病亚型聚类的患者）。测试数据的结构类似，但来自训练数据中未出现的新组。（**B**）随后，我们构建一个分类器，用于预测每个新观测所属的组，并提取组别隶属的预测概率。（**C**）最后，我们在训练集中出现的每个组上分别构建独立的树或森林。最终输出是所有这些树的预测的线性组合，其中（B）中的概率作为权重。

3.1 模拟数据

3.1.1 数据生成

我们使用以下模型构造连续响应变量 y：

$$y_{i,k} \sim \mathcal{N}(\mu(\boldsymbol{x}_{i,k}), 1)$$
$$i \in \{1, \ldots, n_k\}, \quad k \in \{1, \ldots, K\},$$

其中， μ 是均值函数，索引 i 表示每个组中的第 i个受试者。索引 k 表示所有组，使得 $K = J + M$。我们将 K 中80%的组纳入训练数据，并将其余20%留作测试数据，因此，测试数据中的观测总数 n_{test}为 $0.2 \times n \times K$。

为了生成具有结构化相关性的 X 数据，我们采用矩阵正态分布：

$$\mathbf{X} \sim \mathcal{MN}_{K \times p}(\mathbf{M}, \mathbf{U}, \mathbf{V}),$$

其中，$\mathbf{M}$ 是一个（$K \times p$）位置矩阵，$\mathbf{U}$ 是（$K \times K$）组协方差矩阵，而$\mathbf{V}$ 是（$p \times p$）特征协方差矩阵。矩阵正态分布使我们能够自由指定不仅存在于特征之间、还存在于聚类数据设定下各组之间的多种相关结构。我们从该分布中生成 n 个样本，总计得到 $n \times K$ 个观测值。

我们考虑三种不同的模拟设置，每种设置都有其自身的均值函数 μ。第一种设置考虑具有独立线性随机效应的非线性固定效应，第二种设置考虑具有相依线性随机效应的非线性固定效应，第三种设置则考虑生成数据的底层函数随组而变化的情况。下面分别讨论每种设置。

模拟设置 *1*：在本次模拟中，我们设定

$$\mu_0 = f(\boldsymbol{x}_{i,k}) + \mathbf{Z}\boldsymbol{\alpha}_k$$

其中， f 为固定效应函数，$\mathbf{Z}$ 为随机效应设计矩阵， $\boldsymbol{\alpha}_k$ 为所有观测值共享的随机斜率和截距，这些观测值属于第 k 组。我们采用 Friedman 五维测试函数 [Friedman， 1991]作为固定效应函数 f:

$$f(\boldsymbol{x}_{i,k}) = \sin(\pi x_1 x_2) + 2(x_3 - 0.5)^2 + x_4 + 0.5x_5.$$

该函数是机器学习方法的常用基准，因为它混合了线性项、非线性项和交互项。我们使用 Unif(−1,1) 分布生成 $\mathbf{M}$ 的元素，其中 $\mathbf{U}$ 和 $\mathbf{V}$ 为单位矩阵。随后，我们使用具有相同行协方差矩阵 $\mathbf{U}$ 的同一矩阵正态分布生成随机效应 $\boldsymbol{\alpha}_k$ 。综合来看，这等价于从矩阵正态分布$\mathcal{MN}_{K\times(p+1)}(\mathbf{0}, \mathbf{U}, \mathbf{V}^*)$ 中生成一个样本，其中 $\mathbf{V}^*$ 是指定随机斜率与截距之间方差结构的列协方差矩阵。我们使用 $\mathbf{V}^* = \sigma_\alpha^2 \times \mathbf{I}$，其中 $\mathbf{I}$ 是 $(p+1) \times (p+1)$ 单位矩阵。该数据生成策略使得在 X 空间中彼此接近的组具有相似的随机效应。

模拟设置 *2*：接下来，我们进行了一项类似的实验：不再使用 $\mathbf{U} = \mathbf{I}$，而是从逆 Wishart 分布 $\mathcal{W}^{-1}(\mathbf{I}\ K+1)$ 生成一个独特的 $\mathbf{U}$ 矩阵，其中 $K+1$ 为自由度，尺度矩阵取为 $K \times K$ 单位矩阵 $\mathbf{I}$。我们这样做是为了模拟特征数据与随机效应具有相似组协方差结构的情形。

模拟设置 *3*：在此情景中，我们考虑多个与 [Deshpande,2025] 中类似的数据生成过程。此处， X 的生成方式与 模拟设置 *1* 完全相同，而 $\mathbf{M}$ 的各项则改为从 Unif(0,2) 分布中生成。本模拟研究旨在检验以下情景：数据由多个基函数支配，且数据中的每个组仅受其中部分基函数支配。与 Deshpande[2025], 类似，我们还生成 5 个额外的噪声变量，使得 $p = 10$。我们有以下基函数：

$$\begin{aligned}
f_0(x) &= 10\text{正弦}(\pi x_1 x_2) \\
f_1(x) &= 10(x_3 - 0.5)^2 \\
f_2(x) &= 10(x_1 - 0.5)^2 + 10x_2 + 5x_3 \\
f_3(x) &= 6x_1 + (4 - 10 * \mathbb{1}(x_2 > 0.5)) * \sin(\pi x_1) - 4 * \mathbb{1}(x_2 > 0.5) + 15
\end{aligned}$$

这些函数随后在以下数据生成过程中组合：

$$\begin{aligned}
\mu_1(x,k) = (f_0(x) + f_1(x) + f_2(x) - 0.75) \cdot \mathbb{1}\{k \in C_0, C_2, C_4, \dots\} + \\
f_3(x) \cdot \mathbb{1}\{k \in C_1, C_3, C_5, \dots\} \\
\mu_2(x,k) = f_0(x) \cdot \mathbb{1}\{k \in C_0, C_3, C_6, \dots\} + f_1(x) \cdot \mathbb{1}\{k \in C_1, C_4, C_7, \dots\} + \\
f_2(x) \cdot \mathbb{1}\{k \in C_2, C_5, C_8, \dots\} + f_3(x) \cdot \mathbb{1}\{k \in C_0, C_2, C_4, \dots\} \\
\mu_3(x,k) = f_0(x) \cdot \mathbb{1}\{k \in C_0, C_1, C_3, C_4, \dots\} + f_1(x) \cdot \mathbb{1}\{k \in C_1, C_2, C_4, C_5, \dots\} + \\
f_2(x) \cdot \mathbb{1}\{k \in C_0, C_2, C_3, C_5, \dots\} + f_3(x) \cdot \mathbb{1}\{k \in C_0, C_2, C_4, \dots\}
\end{aligned}$$

为应用这些竞争方法，我们借助了 Python 包 sklearn 和 statsmodels。除随机森林外，我们对每种方法均使用默认参数值；随机森林使用 J 棵树实现，因为我们的加权树和方法同样基于 J 棵树进行训练。随机森林的加权组合使用 J 个森林，每个森林含 J 棵树。

3.1.2 比较性能

图 2 展示了模拟设置 *1*的均方误差（均方误差）。均方误差在测试数据上按 $\frac{1}{n_{\text{test}}} \sum_{i=1}^{n_{\text{test}}} (y_i - \hat{y}_i)^2$，计算，其中$\hat{y}_i$ 为观测值 i 的预测结果值。我们考虑了组数 K 与每组观测值数 n 各不相同的模拟设置。我们还改变了 σ_α^2，它用于缩放随机效应的列协方差矩阵，因此较大的 σ_α^2 值会使随机效应具有更大的组间差异。

基于这些模拟结果，我们发现，随着随机效应的幅度 σ_α^2 增大，我们的加权树和方法优于标准决策树，并逐渐具备与随机森林相媲美的表现。鉴于该方法既不采用数据自助抽样，也不采用特征随机选择，这一结果令人印象深刻。此外，该方法仅使用 $J = 0.8 \times K$ 棵树进行拟合。有趣的是，随着随机噪声增加，决策树的表现会下降，而线性混合模型的表现则在不同 σ_α^2 值下更加稳定。我们还发现，在模型的第二阶段使用随机森林，并不会显著改变整体方法的表现。

如**图 2**所示，模拟设置 *2* 的结果比$\mathbf{U} = \mathbf{I}$设定下的结果更加一致，不过总体趋势大致相似。在线性混合模型中， σ_α 值在整个范围内的表现都尚可，而

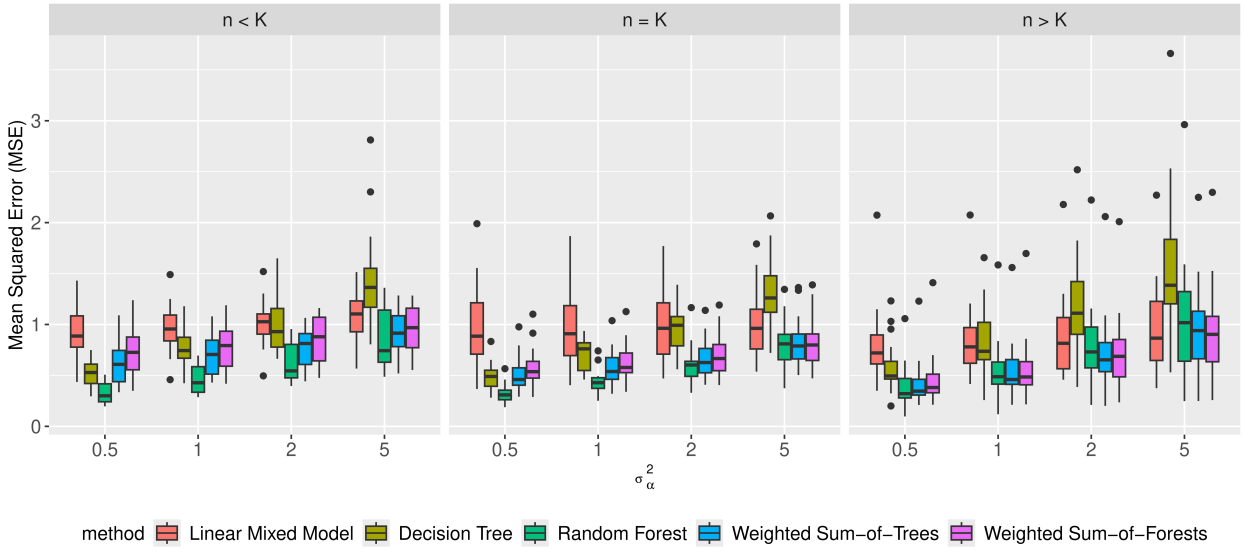


图 2：模拟设置 *1*。在一系列噪声取值下，不同设定的均方误差（MSE） σ_α，其中 $\mathbf{U} = \mathbf{I}$。响应变量 y 的尺度经过标准化。每个箱线图表示 20 个模拟数据集的均方误差值。模拟采用以下设定： $n = 10$、 $K = 40$ （左）， $n = K = 20$ （中），以及 $n = 40$、 $K = 10$ （右）。

决策树和随机森林在较小的 σ_α^2 值设定下表现相对较好，但随着取值增大，性能逐渐变差。随着随机效应增大，与传统决策树相比，我们的加权树和方法优势愈发明显，最终表现得足以与随机森林相竞争。值得注意的是，**图 2** 的尺度大于**图 2**，因为对于某些模拟数据集，每种方法的表现都不佳。

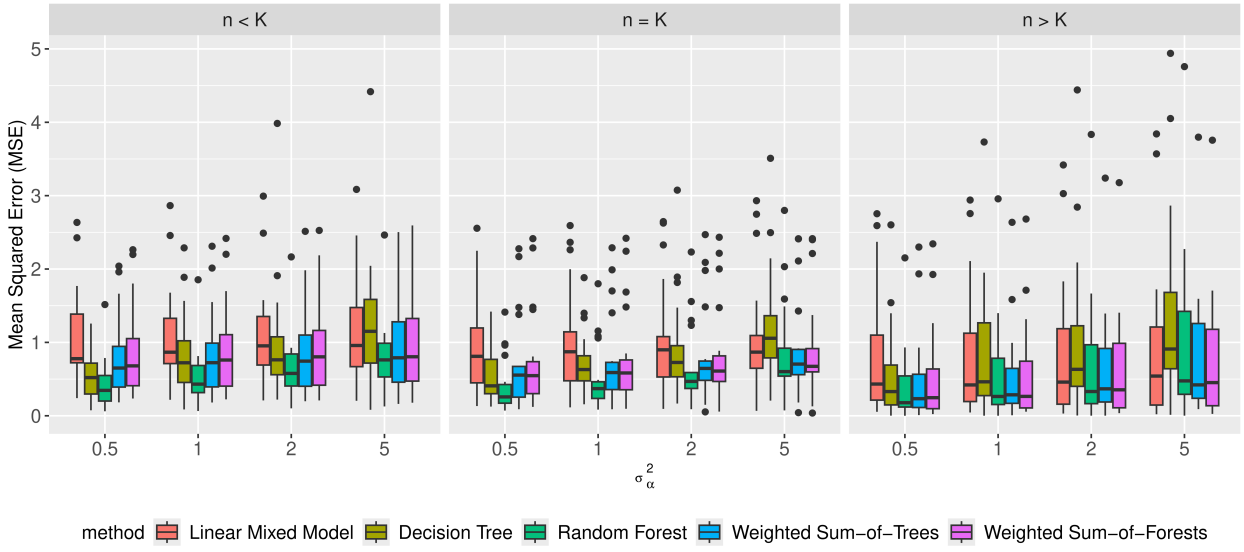


图 3：模拟设置 *2*。对于由逆 Wishart 分布生成 $\mathbf{U}$ 的设定，在一系列随机噪声取值下测试的各种方法的均方误差（MSE） σ_α。响应变量 y的尺度经过标准化。每个箱线图表示 20 个模拟数据集的均方误差值。模拟采用以下设定： $n = 10$、 $K = 40$ （左）， $n = K = 20$ （中），以及 $n = 40$、 $K = 10$ （右）。

最后，我们考虑模拟设置 *3*。结果呈现在**图4**中。对于全部三种数据生成过程，树之和方法、森林和以及随机森林方法的表现均优于线性混合模型和决策树。随着 n 增大，我们的树之和方法在平均表现上超过随机森林，不过

均方误差值的分布仍然有相当大的重叠。值得注意的是，与使用决策树相比，将随机森林用作第二阶段模型并不会显著影响我们方法的表现。对于这一模拟设定，我们还将贝叶斯加性回归树作为比较方法；其表现与随机森林相似（**图8**）。

我们在**表 3**中报告了这些方法的运行时间。对于较小的 n，我们的加权树和模型运行时间与随机森林大致相同；考虑到二者使用的树数量相同，这一点并不意外。然而，随着 n 增大，我们的方法更具优势，运行时间也更接近基决策树，因为用于训练每棵基树的观测值较少。还值得一提的是，我们模型的第一阶段（本例中使用线性回归）并未显著增加运行时间开销。决策树的运行时间最快，而贝叶斯加性回归树和线性混合模型最慢。

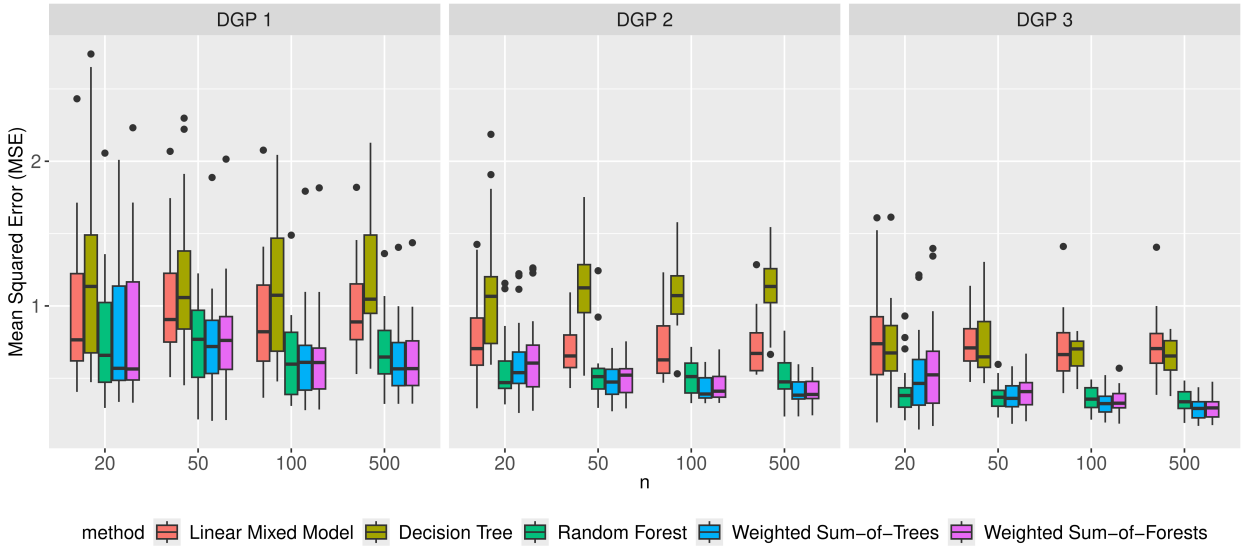


图 4：*模拟设置 3*。在不同每组观测值数量 n 下，对所测试的各种方法计算均方误差（MSE）；这些设定中，数据分别根据 3 种不同的数据生成过程之一生成。在所有模拟中，均使用了 $K = 20$ 组。响应变量 y 的尺度经过标准化。每个箱线图表示 20 个模拟数据集上的 MSE 值。

3.2 真实数据应用

我们将提出的加权树和方法应用于癌症基因组图谱（TCGA）的真实数据，其中患者自然地按癌症亚型分组。

3.2.1 肉瘤数据

肉瘤是一类异质性癌症，起源于人体结缔组织，包括骨、神经、软骨、肌肉、脂肪和血管。由于肉瘤较为罕见（占成人癌症的比例不到 1%），历史上，肉瘤患者在临床试验中一直被归为一组。然而，人们十分关注如何找出可能使特定亚组受益的治疗方法 [Katz et al., 2018]。尤其是免疫治疗药物，虽然已彻底改变许多癌症类型的治疗方式，但在肉瘤治疗中取得的结果喜忧参半；对于免疫微环境有利于产生治疗反应的患者，这类药物可能有所裨益 [Ayodele and Razak, 2020]。

在此，我们分析了 TCGA 中肉瘤队列的数据 [Lazar et al., 2017]。我们尝试预测 T细胞的相对丰度，这是与免疫疗法应答相关的一项关键标志物 [Thorsson et al., 2018]。我们的方法非常适合这一应用，因为 TCGA 数据集中某些类型的肉瘤观测值很少。通过将这些较小组的预测模型与观测值更多的较大组的模型对齐，我们可以提高这些较少见亚型的预测准确性。此外，TCGA 数据集仅涵盖了超过 100 [Grünewald et al., 2020]种肉瘤中最常见的 6 种亚型。我们的方法可用于对现有数据集中表征不足的罕见疾病亚型受试者生成预测。因此，该模型的应用或可解答罕见病精准医学中的一个关键问题：对于一位患有某种亚型的新患者，如果该亚型在临床试验中代表性不足，那么根据其基因组特征，患者从免疫疗法中获益的可能性有多大?

TCGA 数据中包含以下肉瘤亚型的样本：80 例平滑肌肉瘤（LMS）、50 例去分化脂肪肉瘤（DDLPS）、44 例未分化多形性肉瘤（UPS）、17 例黏液纤维肉瘤（MFS）、10 例滑膜肉瘤（SS）和 5 例恶性周围神经鞘瘤（MPNST）。80 例平滑肌肉瘤患者中，有 53 例软组织平滑肌肉瘤（STLMS）和 27 例子宫平滑肌肉瘤（ULMS）。我们使用平滑肌肉瘤、去分化脂肪肉瘤、未分化多形性肉瘤和黏液纤维肉瘤类型的样本作为训练数据，并使用滑膜肉瘤和恶性周围神经鞘瘤的样本作为测试数据。

对于输入特征，我们考虑了可能在免疫活性和肿瘤免疫浸润中发挥作用的临床变量与基因组变量的组合。从临床数据中，我们纳入以下人口统计变量：诊断年龄和性别。我们还纳入了在TCGA肉瘤数据集中被鉴定为频繁异常的以下基因的拷贝数异常 [Lazar等， 2017]：*JUN*,*VGLL3*, *TERT*, *MAP3K5*, *UST*, *CDKN2A*, *YAP1*, *CDKN1B*,*PTPRQ*, *RB1*, *TP53*,*MYOCD*, *NF1*, *CCNE1*, *CEBPA*, *ZNF552*, *ATRX*, *PTEN*, *DDIT3*, *CDK4*, *HMGA2*, *MDM2*, 以及*FRS2*。我们考虑每兆碱基（Mb）的沉默突变和非沉默突变数量；突变数量增加通常与新抗原呈递增加及更高的免疫浸润相关 [Wang等， 2021]。最后，我们还纳入以下基因的mRNA基因表达水平，这些基因先前已被鉴定为与免疫治疗反应相关 [Dulal等， 2023]：*CTLA4*, *TIM3* (*HAVCR2*), *LAG3*, *PD1* (*PDCD1*), *TCF7*, 以及 *TIGIT*。我们使用这些特征来预测肿瘤中T细胞的相对丰度，作为免疫治疗反应的替代指标。

为直观展示数据，我们对输入特征进行主成分分析，并在**图5**中绘制前两个主成分。前两个主成分合计解释了原始数据集29.1%的方差。尽管各亚型内部存在相当程度的异质性，但不同样本组之间仍呈现出明显的模式。

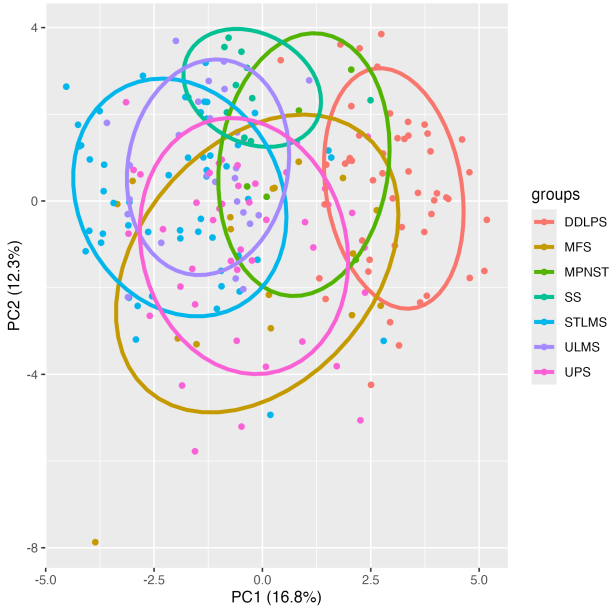


图5：癌症基因组图谱（TCGA）数据集的主成分分析。各点根据相应观测值所属的肉瘤亚型着色。此外，图中绘制的椭圆对应于为每种肉瘤亚型拟合的正态数据曲线，其覆盖概率为0.68。图例：DDLPS = 去分化脂肪肉瘤，MFS =黏液纤维肉瘤，MPNST = 恶性周围神经鞘瘤，SS =滑膜肉瘤，STLMS = 软组织平滑肌肉瘤，ULMS = 子宫平滑肌肉瘤，UPS =未分化多形性肉瘤

我们将树之和方法与传统的线性混合模型、标准决策树以及包含5棵树的随机森林进行比较。选择5棵树，是因为这与我们的树之和模型中的树数量相同。最后，我们测试了以随机森林而非决策树作为基学习器的方法，并将结果与包含100棵树的标准随机森林进行比较。我们的方法使用的随机森林包含20棵树，因此总共也使用了100棵树。我们发现，无论以决策树还是随机森林作为基学习器，我们的方法在整个测试集上都能取得更低的均方误差，表现远优于线性混合模型和基础决策树（**表1**）。此外，在树的数量相同的情况下，我们的方法优于标准随机森林。

在**表2**中，我们展示了模型第一阶段学到的权重。这些权重大体上与**图5**所示的模式一致：滑膜肉瘤与子宫平滑肌肉瘤部分重叠，而恶性周围神经鞘瘤则表现出更强的异质性，尽管其样本量较小。

为了展示我们方法的灵活性和可解释性，我们针对每种肉瘤亚型分别构建了一个随机森林。然后，在**图6**中，我们展示了针对所学习的每个随机森林的组特异性变量重要性与变量交互（VIVI）图 [Inglis 等人， 2022]。尽管多种拷贝数变异在第一阶段模型中具有信息量，但我们发现，性别和拷贝数异常特征在结果模型中的变量重要性相对较低，且未表现出显著的交互效应。因此，我们在**图6**中省略了这些变量。

在**图7**中，我们通过将本方法的权重应用于VIVI图本身，展示了仅包含在测试数据中的亚型对应的VIVI图。对于不在训练数据中的组，我们仍能提供组特异性特征重要性，这是我们方法的一项优势。

表1：线性混合模型、决策树、随机森林和我们提出的加权树和方法的均方误差。随机森林方法进行了两次，一次使用5棵树，另一次使用100棵树。我们的树之和方法也进行了两次。决策树之和是5棵决策树的加权组合。随机森林之和是5个随机森林的加权组合，每个森林包含20棵树（总计100棵树）。响应变量（T细胞丰度）的尺度经过标准化。表现最佳的结果以**粗体**标示。图例：LMM = 线性混合模型，DT= 决策树，RF = 随机森林，SS = 滑膜肉瘤，MPNST = 恶性周围神经鞘瘤

	LMM	DT	随机森林（5棵树）	决策树之和	随机森林（100棵树）	随机森林之和
滑膜肉瘤和恶性周围神经鞘瘤（N=15）	2.964	1.786	0.826	0.643	0.694	**0.582**
滑膜肉瘤（N=10）	4.261	2.238	1.084	0.721	0.924	**0.675**
恶性周围神经鞘瘤（N=5）	0.371	0.883	0.309	0.525	**0.236**	0.362

表2：模型第一阶段学习到的权重。键：DDLPS = 去分化脂肪肉瘤，MFS =黏液纤维肉瘤，STLMS =软组织平滑肌肉瘤，ULMS = 子宫平滑肌肉瘤，UPS =未分化多形性肉瘤，SS = 滑膜肉瘤，MPNST = 恶性周围神经鞘瘤

	去分化脂肪肉瘤	黏液纤维肉瘤	软组织平滑肌肉瘤	子宫平滑肌肉瘤	未分化多形性肉瘤
SS	0.099	0.006	0.384	0.4704	0.040
恶性周围神经鞘瘤	0.282	0.012	0.273	0.185	0.247

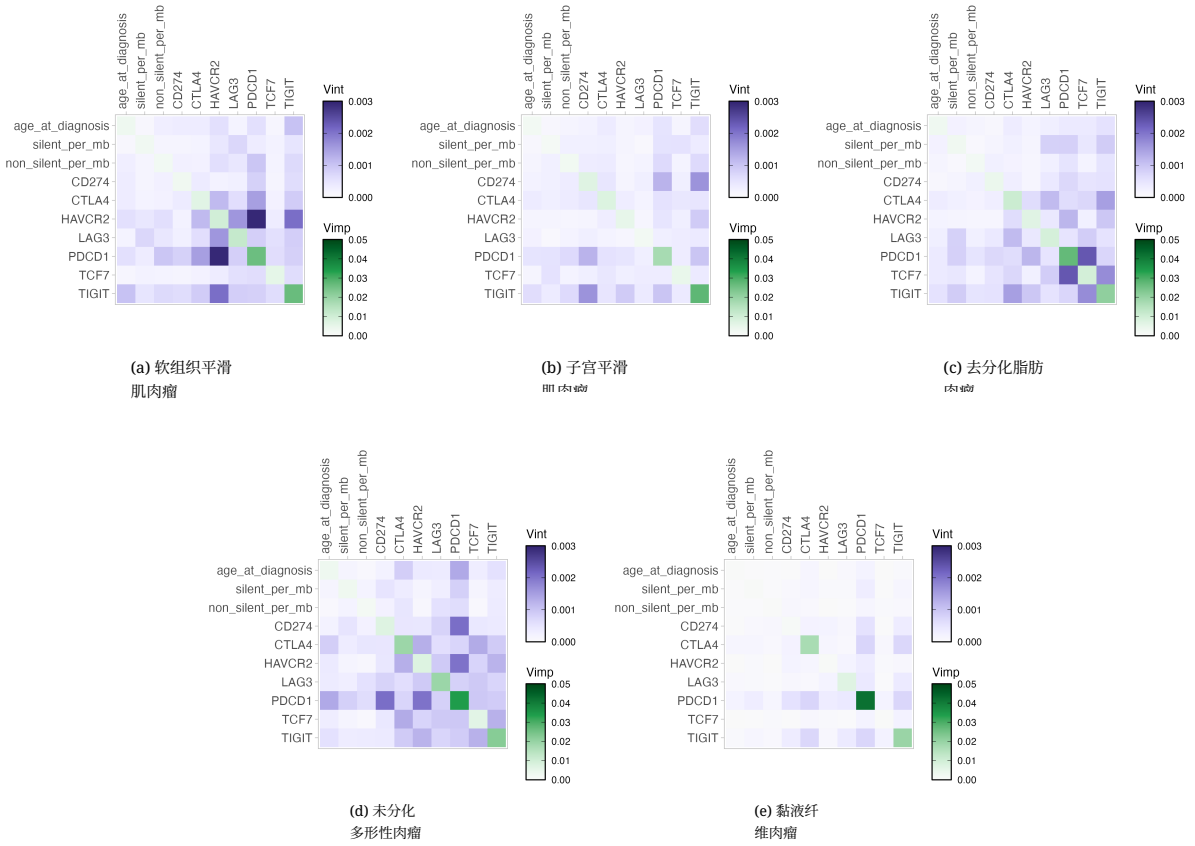
4 讨论

在进行模拟测试时，我们尝试使用不同的学习器来预测组别隶属。我们发现，逻辑回归和朴素贝叶斯通常表现良好。随机森林和决策树分类器也可能表现出色，但有时会得到较差的结果，具体取决于模拟设计参数。我们推荐使用逻辑回归，因为它易于解释，而且可以检查模型系数以识别重要变量。

增加全局噪声通常会降低所有方法的性能，但它们之间的相对排名保持不变。当每组观测值数量超过组数时，我们提出的加权树和方法似乎比其他方法表现更好， $n > K$。这很合理，因为随着 K 增大，逻辑回归（或你选择的其他方法）输出的概率往往会变小，并分散到更多基学习器中，因此输出预测会更像随机森林，但不使用自助法。

随着随机效应的幅度增大且 σ_α^2 变大，我们的方法表现会更好。基于同样的原因，随机噪声增加时，线性混合模型的性能保持稳定也是合理的。

我们还发现，在模型的第二阶段中采用随机森林而非决策树，并不能显著提升性能。加权树求和方法在保持可解释性和具有竞争力的运行时间的同时，展现出了卓越的性能。与表现相当的随机森林不同，我们的方法不对观测值使用自助法，也不随机抽取特征。因此，我们方法所采用的基学习器仍保留了可解释性。此外，每个样本内组训练一个学习器，因此可以检查各个基模型，从而提高其在临床应用中获得信任的潜力。



(a) 软组织平滑肌肉瘤　(b) 子宫平滑肌肉瘤　(c) 去分化脂肪肉瘤

(d) 未分化多形性肉瘤　(e) 黏液纤维肉瘤

图6：针对每种特定肉瘤亚型训练的随机森林的变量重要性（Vimp）和变量交互作用（Vint）图。性别和拷贝数变异变量被排除，因为它们的重要性低得多，且几乎未显示交互效应。所有图共用相同的标度。

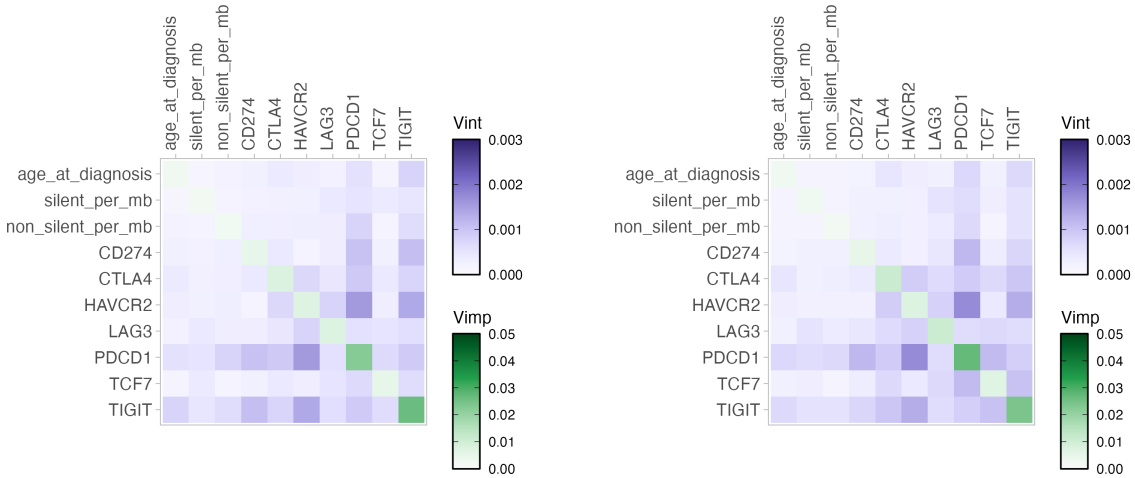
尽管我们仅测试了具有连续输出的数据，但这种方法同样易于用于分类输出，而且第二阶段模型无需使用logit或probit链接函数。

在真实数据结果中，我们看到，与树数量相同的随机森林相比，我们的方法表现更好。考虑到我们的方法既不随机抽样特征，也不对数据进行自助法处理，这一结果更加令人印象深刻。这一发现为如何将这些技巧纳入我们的方法以获得更好的性能打开了大门。

我们还在不同类型肉瘤的VIVI图中看到了有趣的结果。最显著的是，黏液纤维肉瘤显示*PD-1*（*PDCD1*）是黏液纤维肉瘤中高度重要的特征，而所有其他变量和交互作用的重要性均较弱。另一方面，在软组织平滑肌肉瘤和子宫平滑肌肉瘤中，*TIGIT* 作为主要变量也同样重要，并且软组织平滑肌肉瘤中*PD-1*与*TIM-3*（*HAVCR2*）的交互作用尤为显著。这是一个潜在非常有趣的转化医学发现：目前，肿瘤*PD-L1* 表达是免疫治疗反应公认的临床生物标志物 [Li等人， 2022]。我们的发现表明，该标志物可能适用于黏液纤维肉瘤，同时其他基因及其交互作用也可能为其他亚型提供更优的生物标志物。

在案例研究中，我们展示了该方法利用临床和基因组变量预测肉瘤中 T 细胞丰度的实用价值。用于临床风险预测的模型除了为特定患者提供点预测外，还应帮助理解预测的不确定性 [Riley et al., 2025]。已有多种方法用于构建随机森林的预测区间，包括参数方法、分位数随机森林和保形推断[Roy and Larocque, 2020]。Roy 和 Larocque [2020] 推荐采用 Lei 等人提出的保形推断方法的刀切法版本 [2018]。由于该方法不依赖于所使用的预测方法，因此可用于为我们的加权树求和方法构建预测区间。

我们提出的方法存在一些局限性。一个重要的考量是，当每组观测值的数量 n 足够大时，加权树和模型的表现更具竞争力。应用我们提出的方法时还需考虑



（a）滑膜肉瘤（SS） （b）恶性周围神经鞘瘤（MPNST）

图7：两个留出肉瘤类型的预测变量重要性（Vimp）和变量交互作用（Vint）图。这些图通过对其余5种肉瘤类型的VIVI图进行加权求和计算得到。

组别分配被假定为已知。如果事先不知道组别，则可以在应用我们的方法之前，通过无监督聚类学习组标签。最后，如果目标是获得来自样本内组和样本外组的测试观测值的预测，那么对于样本内组，采用线性混合模型或混合效应机器学习模型可能更合适，因为这些模型可以估计随机效应；而对于训练数据中未出现的样本外组，则主要应用我们提出的方法。

5 结论

我们发现，树之和方法在多种模拟设置下均表现可靠。即使树之和方法不对观测值进行自助抽样，也不对特征进行随机抽样，其表现也逐渐可与随机森林相媲美。总体而言，我们的研究表明，在存在聚类或分组观测值时，无需施加严格的模型约束，也能改进决策树和森林的构建。未来，我们计划探究该方法如何应用于分类输出和生存输出。

参考文献

Leo Breiman、Jerome Friedman、Richard A Olshen 和 Charles J Stone。《分类与回归树》。Chapman and Hall/CRC，2017。Leo Breiman。随机森林。*Machine Learning*，45(1):5–32，2001。ISSN 0885-6125。Trevor Hastie、Jerome Friedman 和 Robert Tibshirani。《统计学习要素》。Springer统计学系列。Springer New York，纽约州纽约，2001。Jiming Jiang 和 Thuan Nguyen。《线性与广义线性混合模型及其应用》。Springer统计学系列。Springer New York，纽约州纽约，2021。Ahlem Hajjem、François Bellavance 和 Denis Larocque。用于聚类数据的混合效应回归树。*Statistics& ProbabilityLetters*，81(4):451–459，2011年4月。ISSN 01677152。Charles Spanbauer 和 Rodney Sparapani。面向精准医学的非参数机器学习：纵向临床试验、贝叶斯加性回归树与混合模型。*StatisticsinMedicine*，40(11):2665–2691，2021年5月。ISSN 0277-6715, 1097-0258。Norman E Breslow 和 David G Clayton。广义线性混合模型的近似推断。《*Journal of the American Statistical Association*》，88(421):9–25，1993。Ahlem Hajjem、François Bellavance 和 Denis Larocque。用于聚类数据的混合效应随机森林。*Journal ofStatistical Computation and Simulation*，84(6):1313–1328，2014年6月。ISSN 0094-9655, 1563-5163。Ruo Xu、Dan Nettleton 和 Daniel J. Nordman。个案特异性随机森林。*Journalof Computationaland GraphicalStatistics*，25(1):49–65，2016年1月。ISSN 1061-8600, 1537-2715。Valentijn M. De Jong、Karel G. Moons、Marinus J. Eijkemans、Richard D. Riley 和 Thomas P. Debray。如何利用汇总研究和大型聚类数据集开发更具泛化能力的预测模型。*Statistics inMedicine*，40 (15):3533–3559，2021年7月。ISSN 0277-6715, 1097-0258。Hugh A. Chipman、Edward I. George 和 Robert E. McCulloch。贝叶斯加性回归树。*The AnnalsofApplied Statistics*，4(1)，2010年3月。ISSN 1932-6157。Sameer K. Deshpande。flexBART：具有分类预测变量的灵活贝叶斯回归树。*Journal of Computational and Graphical Statistics*，第1–10页，2025年1月。ISSN 1061-8600, 1537-2715。Mark J Van der Laan、Eric C Polley 和 Alan E Hubbard。超级学习器。*Statistical Applications in Genetics andMolecularBiology*，6(1)，2007。David W Hosmer Jr、Stanley Lemeshow 和 Rodney X Sturdivant。《应用逻辑回归》。John Wiley & Sons，霍博肯，新泽西州，2013。Mousumi Banerjee、Evan Reynolds、Hedvig B Andersson 和 Brahmajee K Nallamothu。基于树的分析：创建临床决策工具的实用方法。*Circulation: Cardiovascular Qualityand Outcomes*，12 (5):e004879，2019。Fabian Pedregosa、Gael Varoquaux、Normalesup Org、Alexandre Gramfort、Vincent Michel 等。Scikit-learn：Python机器学习。*Python*机器学习，2011。Jerome H. Friedman。多变量自适应回归样条。*The Annals ofStatistics*，19(1)，1991年3月。ISSN 0090-5364。Daniela Katz、Emanuela Palmerini 和 Seth M Pollack。超过50种软组织肉瘤亚型：为组织学指导的治疗铺平道路。美国临床肿瘤学会教育丛书，38:925–938，2018。O Ayodele 和 AR Abdul Razak。软组织肉瘤的免疫疗法。*Current Oncology*，27(Suppl 1):17，2020。Alexander J. Lazar、Adam Abeshouse、Clement Adebamowo、Sally N. Adebamowo、Rehan Akbani、Teniola Akeredolu 等。成人软组织肉瘤的综合整合基因组特征分析。*Cell*，171(4):950–965.e28，2017年11月。ISSN 00928674。Vésteinn Thorsson、David L. Gibbs、Scott D. Brown、Denise Wolf、Dante S. Bortone 等。癌症的免疫图谱。*Immunity*，48(4):812–830.e14，2018年4月。ISSN 10747613。Thomas Gp Grünewald、Marta Alonso、Sofia Avnet、Ana Banito、Stefan Burdach、Florencia Cidre-Aranaz 等。分子医学时代的肉瘤治疗。*EMBO Molecular Medicine*，12(11):e11131，2020年11月。ISSN 1757-4676, 1757-4684。Peipei Wang、Yueyun Chen 和 Chun Wang。超越肿瘤突变负荷：肿瘤新抗原负荷作为免疫疗法及其他类型疗法的生物标志物。*Frontiers in Oncology*，11:672677，2021。

Dharmindra Dulal、Andrew Boring、David Terrero、Tiffany Johnson、Amit K Tiwari 和 Dayanidhi Raman。通过靶向替代性免疫检查点治疗免疫难治性肿瘤。《*Cancers*》，15(10):2774，2023。Alan Inglis、Andrew Parnell 和 Catherine Hurley。vivid：用于机器学习模型变量重要性与变量交互作用可视化的 R 包，2022。网址：https://arxiv.org/abs/2210.11391。版本号：3。Hanxiao Li、P Anton van der Merwe 和 Shivan Sivakumar。PD-1通路阻断的应答生物标志物。《*British Journal of Cancer*》，126(12):1663–1675，2022。Richard D Riley、Gary S Collins、Laura Kirton、Kym IE Snell、Joie Ensor 等。临床预测模型风险估计的不确定性：依据、挑战与方法。《*BMJ*》，388，2025。Marie-Hélène Roy 和 Denis Larocque。随机森林的预测区间。《*Statistical Methods in Medical Research*》，29(1):205–229，2020。Jing Lei、Max G'Sell、Alessandro Rinaldo、Ryan J Tibshirani 和 Larry Wasserman。回归的无分布假设预测推断。《*Journal of the American Statistical Association*》，113(523):1094–1111，2018。

附加表格和图

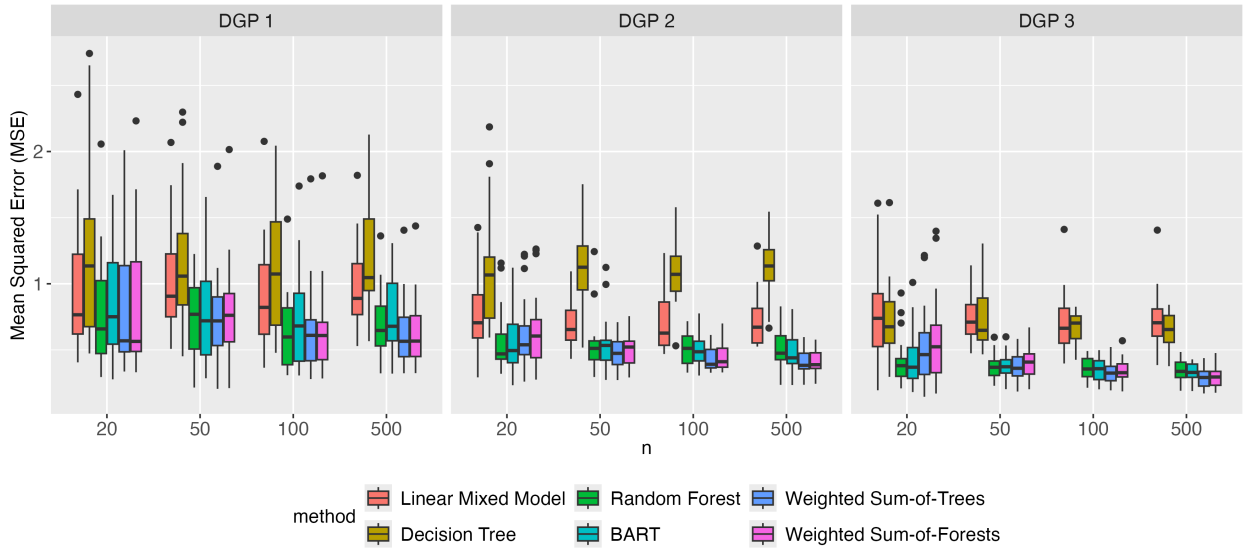


图 8：贝叶斯加性回归树比较

表 3：平均运行时间（秒）。请注意，n 是每组观测值的数量，因此用于训练每种方法的观测值总数等于 n 乘以 J，本例中为 16 (J=$0.8 \times K = 16$)。

n	20	50	100	500
线性混合模型	2.346	2.790	2.898	5.601
决策树	0.004	0.008	0.017	0.100
随机森林	0.030	0.073	0.156	0.981
BART	0.188	0.392	0.807	4.921
加权树和	0.033	0.038	0.048	0.147
加权森林和	0.154	0.171	0.213	0.668