arXiv:2601.07003v1 [stat.ME] 11 Jan 2026

Unity 森林：提升随机森林中的交互建模与可解释性

RomanHornung [1,2]* 和 Alexander Hapfelmeier [3,4]

[1]慕尼黑路德维希-马克西米利安大学医学院医学信息处理、生物统计与流行病学研究所（LMU），德国慕尼黑[2]慕尼黑机器学习中心（MCML），德国慕尼黑[3]慕尼黑工业大学医学院与健康学院临床医学系全科医学与卫生服务研究所（TUM），德国慕尼黑[4]慕尼黑工业大学医学院与健康学院医学人工智能与信息学研究所（TUM），德国慕尼黑*通讯作者：Roman Hornung，hornung@ibe.med.uni-muenchen.de

2026年1月13日

摘要

随机森林（RF）广泛用于预测和变量重要性分析，人们通常认为它们能够通过递归分裂捕捉各种类型的交互作用。然而，由于分裂是在局部选择的，只有当参与交互的至少一个协变量具有边际效应时，交互作用才能得到可靠捕捉。我们提出 Unity 森林（UFO），这是一种随机森林变体，旨在更好地利用涉及无边际效应协变量的交互作用。在 UFO 中，每棵树最初的若干次分裂会针对一个随机协变量子集进行联合优化，以形成能够捕捉此类交互作用的“树根”；其余部分则按常规方式生长。我们还提出统一变量重要性度量（VIM），该度量基于树根的袋外分裂准则值。对于每个协变量，仅考虑少量袋内准则值最高的树根分裂；这反映了纯粹具有交互效应的协变量，只有在树中较早出现了某个交互协变量的分裂时，才具有区分能力。最后，我们提出协变量代表性树根（CRTR），为每个协变量选取具有代表性的树根，并以可解释的方式揭示各协变量在何种条件（边际或交互）下具有最强效应。在模拟研究中，统一 VIM 能够可靠地识别无边际效应的交互协变量，而传统的基于随机森林的变量重要性度量则不能。在大规模真实数据比较中，UFO 的区分能力和预测准确度均高于标准随机森林，且校准表现相当。CRTR 在模拟数据中可靠地还原了协变量的真实效应类型，并在真实数据分析中提供了有意义的见解。

1 引言

随机森林 [Breiman, 2001] 是应用最广泛的基于协变量的预测方法之一。随机森林广受欢迎的重要原因包括其通常出色的预测性能，以及能够利用变量重要性度量（VIMs）按预测相关性对协变量进行排序。其中，基于置换的变量重要性度量 [Breiman, 2001] 可能是应用最广泛的方法。长期以来，平均不纯度下降（MDI） [Breiman等人, 1984] 也很受欢迎，这可能是因为其计算成本较低。然而，Strobl 等人 [2007]证明，MDI

会系统性地赋予具有许多可能分裂点的协变量更高的重要性。为了解决这一问题，人们提出了经过偏差校正的 MDI 版本：一种是基于协变量的置换版本 [Nembrini 等人， 2018] 另一种是利用树的袋外（OOB）数据 [Li 等人，2019 年，Loecher，2020，Zhou 和 Hooker， 2021]。

随机森林预测性能强大的一个关键原因，在于其树在对结果变量与协变量之间的关联进行建模时偏差较低。具体而言，随机森林能够通过使用不同协变量递归地将数据分割为越来越同质的子节点，从而捕捉交互效应。然而，由于分割选择是在每个节点局部进行的，且不考虑后续分割，因此只有当至少一个相关协变量具有足够强的边际效应、因而能够被选作分割协变量时，这些交互作用才能得到充分表示[Wright 等人，2016 年，Hornung 和 Boulesteix， 2022]。

本文提出一种替代性的树构建算法，其中每棵树的前 MAX DEPTH ROOT层（默认值：3）是联合确定的，而非依次确定。我们将该算法称为*Unity* 森林算法，并将由此生成的森林称为*Unity* 森林（UFO）。UFO 算法允许采用非立即具有区分性的分割——也就是说，分割后子节点中的结果变量分布并不比父节点更集中——但这类分割能够使后续层中其他协变量上的分割具有区分性。这类分割组合可以表示一种交互效应，其中所有相关协变量均不表现出边际效应。尽管这类纯交互效应通常不如包含边际效应的交互效应常见，但它们往往具有实质性意义；若不进行显式指定，加性模型（如广义线性模型）便无法捕捉其影响。

以下我们将介绍几个纯粹用于说明的例子，展示在没有边际效应的情况下，交互效应如何出现在实际应用场景中。这类效应通常是定性交互效应，即一个协变量的影响方向取决于另一个协变量的取值。治疗效应的方向可能取决于 BMI 等调节变量：对 BMI 较高的个体有正向影响，而对 BMI 较低的个体则有负向影响。心理干预在某种情境中可能有益，例如平静的工作环境；但在另一种情境中则可能适得其反，例如压力极大的环境。在农学中，由于土壤化学或生物性质存在差异，某种肥料可能促进某一土壤类型中的植物生长，却抑制另一种土壤类型中的植物生长。案例式学习等教学方法可以提升先验知识丰富的学生的学习成效，却可能妨碍先验知识不足的学生学习。同样，颜色编码按钮等界面元素可能提高年轻用户的可用性，却降低年长用户的可用性，例如，这是由于年龄相关的颜色感知变化或对数字惯例的熟悉程度不同所致。

除了 UFO 算法之外，我们还提出了一种新颖的变量重要性度量（VIM），即统一变量重要性度量。该度量旨在量化每个协变量在其影响最强的条件下所产生影响的强度。本文及下文中，为行文简洁起见，使用“影响”一词指代与结果变量之间观察到的关联，而不对因果关系作出任何主张。对于不与其他变量发生交互作用的协变量，不存在使其效应特别强的特定条件，因此，统一变量重要性度量反映的是该协变量对结果变量的边际影响强度。对于与其他变量发生交互作用的协变量，相关条件对应于由这些交互协变量取值范围界定的子组。如果某个协变量与多个其他变量发生交互作用，则只考虑涉及其最强交互作用的子组。这确保只有显著的交互效应才会对统一变量重要性度量产生实质性贡献。相反，如果某个协变量仅与多个其他变量存在较弱的交互作用，就不应期望它获得较高的统一变量重要性度量值——这是有意为之，因为较弱的交互作用难以检测，在旨在刻画协变量效应性质的应用中也很少具有实际意义。此外，一个参与多种交互作用的协变量完全不对结果变量产生直接效应，这种情况通常也不太可能。

如上一段所述，统一变量重要性度量旨在识别具有显著边际影响或由交互作用产生的条件影响的协变量。然而，统一变量重要性度量值只表明哪些协变量具有影响力，而不说明它们如何影响结果变量。

后者对于影响完全基于交互作用的协变量尤为重要。为了更好地了解此类协变量如何影响结果变量，我们提出一种补充方法，旨在为每个协变量选出一棵最能代表其最强影响的树。该过程会从协变量表现出最强袋外影响的树子集中，为每个协变量确定一棵代表性树 [Laabs 等人， 2024], 即协变量代表树根（CRTR）。

完整的UFO框架——由UFO 算法、统一变量重要性度量以及CRTR选择与可视化流程组成——已在R软件包`unityForest`中实现，该软件包可在CRAN上获取。该软件包以广泛使用的RF实现`ranger` [Wright 和 Ziegler的实现为基础， 2017] 并使用快速C++ 代码执行所有计算密集型操作，默认使用所有可用的CPU核心。

本文其余部分的组织结构如下。第2 节回顾相关方法。第3节介绍构成UFO框架的算法。第4 节通过一项大规模真实数据研究，对关键超参数进行实证评估，并将UFO的预测性能与传统随机森林进行比较。第5 节报告一项模拟研究，评估统一变量重要性度量相较于传统变量重要性度量区分信息性协变量（尤其是交互协变量）与非信息性协变量的能力。在第6节中，我们使用模拟数据和真实数据评估CRTR算法。最后，第7 节讨论研究结果、局限性及潜在扩展，第8 节总结主要发现。

2 相关工作

UFO 是一种随机森林变体，旨在改进标准森林中的树结构。若干方法旨在将性能更优的树纳入树集成，但这些方法通常依赖传统的分类与回归树流程 [Breiman等人， 1984] 用于树构建。例如，Bernard等人 [2009] 和Khan等人 [2020] 通过依据性能标准选择树的子集，在寻求提升预测性能和可解释性的同时缩小集成规模。相关方法综述见Manzali和Elfar [2023]。相比之下，UFO 算法并不旨在挑选最佳的传统树，而是生成替代性的树结构，以更全面地反映协变量与结果变量的依赖关系，尤其是没有边际效应的交互效应。其两个主要目标是：（i）利用这些改进的树结构，实现能更好捕捉此类交互作用的变量重要性；（ii）提升生成的树的可解释性。

据我们所知，唯一一种专为更好地捕捉交互效应而设计、且在实践中可用的基于随机森林的变量重要性度量实现，是B´enard等人提出的[2022], 基于Shapley值的SHAFF方法；不过，该方法仅适用于连续结果。

另外两个相关研究领域是规则提取方法和代表性树的选择方法；从概念上说，这两者都与本文提出的CRTR算法相关。RuleFit等规则提取方法 [Friedman和Popescu、 2008], 节点收获 [Meinshausen、 2010], 以及SIRUS[B´enard等人， 2021]旨在将森林概括为稀疏的决策规则集合，从而在保留预测能力的同时提升可解释性。这些方法在规则的生成、截断和组合方式上各不相同，但都秉持同一理念：让森林中隐含的结构更加透明。

代表性树方法旨在找出最能代表集成的树，以供检查。它们定义树之间的成对距离，然后选取与所有其他树的平均距离最小的树作为代表树。 Banerjee 等人 [2012]提出了这一想法，Laabs 等人 [2024] 则通过在距离度量中赋予高层分裂更大的权重对其加以改进。单棵树能够突出具有代表性的结构，但不足以体现森林的复杂性。为解决这一问题，Szepannek 和 Holt [2024] 提出使用少量代表性树或代理树，在可解释性与更全面地覆盖森林所捕获的关系之间取得平衡。

3 UnityForest 框架

令（$\boldsymbol{x}_i$，y_i）表示第 i个观测，其中 $\boldsymbol{x}_i = (x_{i1}, \ldots, x_{ip})^\top$ 是一个 p维协变量向量，而 y_i 是其对应的结果变量。我们考虑分类和回归；在分类情况下，$y_i \in \{1, \ldots, K\}$ 表示类别标签，而在回归情况下，$y_i \in \mathbb{R}$ 表示连续结果。对于分类，我们同时考虑类别标签预测和类别概率估计。类别标签通过多数投票进行预测，与传统随机森林中的做法相同；类别概率则通过对包含待预测观测值的叶节点中的经验类别分布取平均获得，具体如Malley 等人 [2012]所述。

本节将介绍用于树构建的 UFO 算法、用于计算统一变量重要性度量的算法，以及用于选择 CRTR 的算法。

3.1 树构建

3.1.1 算法及其原理

为便于初步了解，算法 1 以高层次概述了 UFO 中用于构建每棵独立树的过程。详细的分步说明，包括具体的停止规则和其他技术细节，可参见补充材料的 A.1 节。

算法 1 Unity 森林中树构建过程示意

1: 无放回抽取比例为 FRACT N （默认值：0.7）的观测值——即树样本——并抽取较小比例 PROP VAR 的协变量。PROP VAR 的默认值设为：对于 $p \leq 100$，设为 $\sqrt{p}/p$；对于 $p > 100$，设为 0.1，其中 p 表示协变量的数量。

2: 使用选定的协变量，在树样本上生长 N CAND TREES （默认值：500）棵深度有限的候选树根，深度为 MAX DEPTH ROOT （默认值：3），其中分裂协变量和分裂点均随机选取。

3: 对于每个候选树根，计算划分准则，以衡量完整根叶节点集合（即树根的叶节点）相较于整个树样本的不纯度所实现的不纯度降低；分类使用基尼不纯度，回归使用方差。

4: 选择划分准则值最优的候选树根；由于树样本的不纯度是恒定的，这相当于选择根叶节点集合最纯的树根。

5: 使用树样本和传统的递归分类与回归树式分割，将该树根扩展为完整的树；在每次分割时，从所有协变量中随机选择 MTRY （默认值：$\lfloor \sqrt{p} \rfloor$）个候选协变量，方式与传统随机森林相同。

尽管如今“随机森林”（随机森林）一词通常指一种特定的过程，但原论文中由 Breiman [2001] 给出的定义要宽泛得多。在该论文中，随机森林被定义为由多棵树组成的集合，其中每棵树都依赖于一个随机向量 Θ，该随机向量在各棵树之间独立同分布。由于算法 1 符合这一定义，因此 UFO 构成了随机森林的一种变体。

在 UFO 算法中，树根处的分裂并未得到充分优化。相反，对于每棵树，都会从许多候选树根中选出能够生成最纯净根叶节点的树根。因此，所选树根通常无法利用算法 1 第 1 步中随机抽取的协变量实现可能的最佳划分，不过，增加候选树根的数量可以获得越来越好的近似结果。但这并非必要。随机森林的预测性能往往会随着单棵树性能的提升而提高，但当不同树的预测相关性增加时，预测性能则会下降 [Breiman, 2001]。在传统随机森林中，每次分割都是根据

随机选取的 MTRY 个协变量子集确定的，并且每次分割都会重新抽样，从而确保较高的树多样性。相比之下，如果在 UFO 算法中对树根处的分裂进行优化，树预测之间的相关性就会增加，进而可能降低森林的整体预测性能。

如果能够找到最优划分，所得树根将对应于所谓的最优树 [Bertsimas 和 Dunn， 2017], 其构建如今已切实可行，但计算开销仍然很大。与最优树类似，UFO 算法选择最佳树根，旨在对由随机选取的协变量所张成的协变量空间进行最有效的划分，而不是像传统 CART 树那样，在各个节点进行局部判别性分割。这种方法使 UFO 算法能够纳入一些自身可能没有边际效应、但可通过后续对其他协变量进行分割而促成有效划分的协变量，从而更全面地捕捉协变量之间的交互效应。

并非所有协变量都会用于构建树根。为提高树多样性（其效果类似于对观测值进行子抽样并使用未经优化的分割），同时由于这一限制对统一变量重要性度量也很重要（将在第 3.2 节讨论），因此只使用其中的一个子集。

随后将树根扩展为完整的树，因为深树与较低的偏差相关。尽管深树的方差较高，但这种方差不会传递到森林的预测中，因为预测结果是对许多棵树的结果取平均值 [Hastie 等人，2009 年，第 588]页。

3.1.2 名义协变量的处理

UFO框架中的树构建过程不能直接扩展到名义协变量，因为该过程要求协变量值有序。为了仍能纳入名义协变量，我们按照 R软件包`ranger` （0.17.0 版）中通过选项 `respect.unordered.factors = "order"` [Wright 和 K¨onig，2019 年，Coppersmith 等人， 1999]所实现的方法，对其类别进行排序。该选项的基本思路是对名义协变量的类别进行排列，使得沿类别顺序递进时，相应的结果值倾向于朝同一方向变化。补充材料 A.2 节详细介绍了针对分类变量和连续结果所采用的具体过程。根据所得顺序，每个名义协变量的类别被赋予数值 $1, \ldots, J$，其中 J 表示类别数量；随后，这些类别在 UFO框架中作为有序协变量处理。

3.1.3 超参数

UFO 算法涉及若干超参数。我们建议使用这些超参数的默认值，而非进行优化。此建议基于两点考虑：(i) 超参数调优通常只能小幅提升随机森林的预测性能 [Probst 等人， 2019], 以及 (ii) 某些参数的最优值取决于其他参数，因此单独优化可能会带来问题。

许多超参数与传统随机森林中的超参数相对应，其默认值也基本遵循 `ranger` [Wright 和 Ziegler 的实现 2017]。在附录4.2节中，我们使用 30 个真实世界数据集评估了三个关键超参数的默认值是否合适。

影响可解释性的一个超参数是 MAX DEPTH ROOT，其默认值为 3。此设置允许树根捕捉最多涉及三个协变量的交互效应。

完整的超参数列表及其默认值和详细说明见补充材料附录A.3节。

3.2 统一变量重要性度量

算法 2 概述了计算统一变量重要性度量值的流程，以便初步了解。

算法 2 统一变量重要性度量的计算示意

1: 对于树根中每个不是根叶节点的节点，使用袋内观测值计算一个分裂得分，以衡量相应分割的区分能力。

2: 对于每个协变量，选取得分最高的分裂集合。

3: 对于每个协变量，在得分最高的分裂处，根据在节点内置换该协变量前后基于袋外数据的分裂得分之差，计算统一变量重要性度量。

统一变量重要性度量聚焦于每个协变量最具区分性的分裂，从而实现其预期目的：仅在协变量影响最强的条件下量化其影响。对于边际效应较强的协变量 A，其绝大多数分裂都具有区分性。相比之下，对于边际效应很弱或没有边际效应、但与另一个协变量存在交互作用的协变量 B，只有当同一棵树中先前发生了与 B 交互的协变量分裂时，B 的分裂才具有区分性。因此，B 的大多数分裂都不具区分性；考虑到每个树根只会选取一小部分协变量（见第 PROP VAR 节 3.1.1），这一影响会进一步加剧。对于此类协变量，因此只有最具区分性的分裂才对统一变量重要性度量有所贡献，这一点至关重要。

除了有助于提高树多样性（参见第3.1.1节）之外，将每棵树的树根限制在较小的协变量比例范围内，也能减轻同一批影响力较强的协变量在整个森林中的主导作用。进而，这使得许多树能够捕捉到没有边际效应的交互效应。

如算法2所述，分裂得分的计算以及随后对最优分裂的选择均使用袋内观测值，而基于最优分裂计算的统一变量重要性度量贡献则使用袋外数据。将不同的数据子集用于分裂选择和评估所选分裂的区分能力，可以避免统一变量重要性度量值出现向上偏差；如果在同一数据上进行最优选择，则会产生这种偏差。

统一变量重要性度量的详细计算算法如下：

1. 计算用于量化分裂区分能力的分裂得分。使用袋内观测值，针对所有树的树根内部节点（即不包括根叶节点）上的每个分裂，计算基尼不纯度（用于分类）或方差（用于回归）的降低量，并将该值乘以节点大小。将所得值记为分裂得分。

2. 按协变量选取得分最高的分裂。对于每个协变量，确定树根中分裂所占的比例 PROP BEST SPLITS（默认值：0.01），这些分裂在树根中所有使用该协变量的分裂里具有最大的分裂得分。如果按此比例选出的分裂少于五个，则选择得分最高的五个分裂；如果该协变量可用的分裂少于五个，则选择全部可用分裂。将与所选分裂对应的节点索引记为 j，记作 $\mathcal{B}_j$。

3. 使用得分最高的分裂处基于袋外数据的分裂得分计算统一变量重要性度量。对于每个协变- 量 j，$j \in 1, \ldots, p$，按如下方式计算统一变量重要性度量：

$$\sum_{l \in \mathcal{B}_j} N_l(\text{OOB_SC}_l - \text{OOB_SC_PERM}_l). \tag{1}$$

这里，N_l 表示节点 l中的袋内观测值数量。OOB SC_l 是节点 l处基尼不纯度（分类）或方差（回归）的减少量，仅使用经过该节点的袋外观测计算。OOB SC PERM_l 是相同的量，但在经过节点 l的袋外观测中随机置换协变量 j 的值后计算。 - -

在基于比例的规则（PROP BEST SPLITS）为某个协变量选出的分裂少于五个时，系统会改为选取至多五个分裂，以稳定统一变量重要性度量。若仅

依据数量过少的分裂计算变量重要性度量，其结果会更易变，也更容易受到随机波动的影响。因此，五个分裂的下限可确保统一变量重要性度量基于足够多的有效分裂计算，同时仍聚焦于区分能力最强的分裂。在我们的模拟研究和真实数据分析中，从未触发这一备用机制。

关于该算法的进一步说明以及默认超参数值的讨论，请参见补充材料的 A.4节。

3.3 协变量代表性树根（CRTR）

如前所述，原则上可以为任意协变量选择 CRTR。不过，由于只有统一变量重要性度量值足够高的协变量，才可认为具有相关影响——无论这种影响是边际影响，还是通过与其他协变量的交互作用产生的影响，因此建议仅为此类协变量选择 CRTR。

下面将详细介绍选择 CRTR 的算法：

1. 使用袋外数据计算分裂得分。按照计算统一变量重要性度量值时相同的方法，为树根中的每个内部节点计算分裂得分，但使用袋外观测而非袋内观测值。2. 针对每个协变量选择得分最高的分裂。此步骤与计算统一变量重要性度量值时完全相同，使用相同比例 PROP BEST SPLITS。3. 选择最佳树根。对于每个协变量，找出包含该协变量至少一个得分最高分裂的树根集合。将选出的树根集合称为相应协变量的最佳树根。4. 选择 *CRTR*。对于每个协变量，使用Laabs 等人 [2024]提出的程序，从最佳树根集合中确定 CRTR。5. 为可视化分配协变量得分。在选出的 CRTR 中，为树根的内部节点分配一个协变量得分，其计算方式如下：

$$\frac{\#freq_best_j}{\#freq_best_j + \#freq_all_j}, \tag{2}$$

其中，协变量 j 是相应节点上的分割协变量。这里，$\#freq\ best\ j$ 表示协变量 j_在协变量 j 的最佳树根的内部节点上被用于分割的相对频率，而 $\#freq\ all\ j$ 表示所有树根的内部节点中基于协变量 j 进行分割的相应相对频率。这些协变量得分用于可视化代表性树根；详情请参见第 6.1 节。

通过选择相应协变量能够产生最具区分性的一个或多个分割的树根，得到的最佳树根主要代表该协变量影响最强的条件。当然，这些集合中仍存在一定的变异，因为并非所有树根都能同样准确地捕捉这些条件。Laabs 等人 [2024] 的方法会根据反映协变量使用情况相似性的距离度量，从最佳树根集合中选出与集合内所有其他树根的平均距离最小的树根。较高层级的分割比低层级的分割具有更高权重，因为前者对树结构及由此产生的协变量空间划分影响更大，因此选出的树根不太可能无法充分反映协变量影响最强的条件。

与统一变量重要性度量的计算不同，在上述算法中，得分最高的分裂——因此也是最佳树根——是依据袋外数据而非袋内数据选出的。其理由是，在 CRTR 可视化过程中，用于解释的节点级结果分布或汇总统计量是根据袋内数据计算得出的。如果最佳树根的选择

也基于同一袋内数据，那么由于经过了最优选择，基于 CRTR 协变量的分裂会表现出人为偏高的判别能力。使用袋外数据进行选择可以避免这种乐观偏差，因为可视化中基于袋内数据的注释仍然是无偏的。

在第 6 节中，我们展示了 CRTR 在模拟数据和真实数据中的应用，并详细介绍可视化流程。

4 UFO 在真实数据上的大规模预测性能评估

本节中，我们使用大量具有二元结局的真实世界数据集评估 UFO 的预测性能（其他结果类型的讨论见第7 节）。第 4.1节介绍本研究所用的数据集池。第 4.2节中，我们使用相对较小的数据集子集进行初步分析，以评估为若干关键超参数选取的默认值是否合适。最后，第4.3 节基于其余数据集，对 UFO 与传统随机森林的预测性能进行大规模比较。

4.1 数据

我们使用了 198 个具有二元结局的真实数据集。这些数据集是对Couronn´e 等人 [2018]在大规模基准研究中使用的 243 个数据集进行预处理后得到的子集。为尽量减少数据来源之间的依赖，我们排除了冗余、高度相似或存在统计依赖的数据集。为降低计算开销，对于观测值超过 10,000 个或协变量超过 1,000 个的数据集，我们分别随机抽取子集，使其包含 10,000 个观测值或 1,000 个协变量。补充材料中的表 S1–S5 概述了所用数据集。

为确保完全可复现，本文及补充材料中所有分析所需的 R 代码和数据均已发布在 GitHub 上（https://github.com/RomanHornung/UnityForests_代码_和_数据，提交版本：08edff9）。此处通过专门的 R 脚本详细说明了从原始 243 个数据集中排除 45 个数据集的原因。

4.2 关于关键默认超参数值适当性的初步研究

UFO框架包含若干超参数。在第 3 节以及补充材料的 A.3 和 A.4 节中，我们说明并论证了所选默认值。正如前文所述，随机森林性能通常对默认超参数值的变化相当稳健；不过，我们针对三个可能对预测性能产生较大影响的超参数进行了评估：候选树的数量（n cand trees）、树根的深度（max depth root），以及每个树根采样的协变量比例（prop var）。

本次初步分析的目的并非确定联合最优的超参数组合，而是评估所选默认值所带来的预测性能是否接近在合理范围内可达到的最佳性能，并评估预测性能对各个超参数变化的敏感程度。此外，默认值是基于概念性考量而选择的，本次分析并非旨在推翻这些设计选择。

在本次初步分析中，我们从总计198个数据集中选取了30个数据集。三个超参数分别在由合理值构成的网格中单独变动，其中包括各自的默认值，同时将其他所有超参数固定为其默认值。我们采用重复五折交叉验证，并使用三项性能指标评估UFO的预测性能：布里尔分数、ROC曲线下面积（AUC）和准确率。补充材料B.2节详细介绍了研究设计和所得结果。

总之，所考察的超参数对预测性能的影响可以忽略不计。对于绝大多数数据集，在所考察的取值之间未观察到明显的性能差异

。在所考察的超参数值之间存在细微差异的情况下，采用默认值所获得的性能等于或非常接近观察到的最高性能。这些发现表明，为这三个超参数选定的默认值是合适的，并且在大多数应用中，调整这些参数不太可能带来有意义的性能提升。

4.3 UFO与传统随机森林的预测性能比较

UFO是对传统随机森林的一种改进，而传统随机森林是一种应用广泛且成熟的方法，其预测性能已得到广泛研究。因此，评估这种改进是否能提升或降低预测性能尤为重要。为此，我们使用 198 个可用数据集中的 168 个（排除初步研究中使用的数据集），比较了 UFO 与传统随机森林的预测性能。

UFO的所有超参数均设为默认值。为确保在可比的默认设置下进行公平比较，并避免偏袒任一方法，我们尽可能使传统随机森林的配置与UFO保持一致，将所有随机森林超参数设为相应的UFO默认值。具体而言，我们采用子采样 $\text{FRACT N} = 0.7$ 而非自助抽样，每个森林包含 20,000 棵树，并将分割所需的最小节点大小设为 5，这与我们进行概率预测相一致。[-] 与初步研究一样，我们采用重复五次的五折分层交叉验证作为重采样方案，并使用布里尔分数（Brier）、ROC曲线下面积（AUC）和准确率（ACC）作为性能指标。

表 1 按数据集汇总了在不同性能指标下，UFO 在各数据集上的表现优于、持平（即性能值相同）或逊于传统随机森林的频次。就 布里尔分数而言，UFO 在略多于半数的数据集上逊于随机森林。相比之下，对于 曲线下面积，UFO 在约三分之二的数据集上优于随机森林，在约四分之一的数据集上逊于随机森林。在 ACC方面，UFO 同样优于随机森林：约 60% 的情况下表现更好，约三分之一的情况下表现更差。

表1：UFO 与传统随机森林在各数据集上的比较。每个条目显示 UFO 表现优于、持平或逊于随机森林的数据集数量；括号中的百分比表示相应比例

指标	UFO 更优	UFO 持平	UFO 更差
布里尔分数	78 (46.4%)	0 (0.0%)	90 (53.6%)
AUC	114 (67.9%)	9 (5.4%)	45 (26.8%)
ACC	101 (60.1%)	10 (6.0%)	57 (33.9%)

为评估 UFO 的表现是否在统计学上显著优于或劣于随机森林，我们进行了二项检验，并将数据集视为相互独立的观测。零假设认为，UFO 的预测性能高于随机森林的概率等于 0.5。对于每项检验，均剔除了两种方法结果相同的数据集，因为这些情况无法提供有关相对性能的信息。由此得到的 p 值如下：布里尔分数：0.396；曲线下面积：< 0.001；ACC：< 0.001。采用传统的显著性水平 $\alpha = 0.05$，我们得出结论：在曲线下面积和 ACC 方面，UFO 的表现显著优于随机森林，而布里尔分数方面的差异在统计学上并不显著。

总之，在曲线下面积和 ACC 方面，UFO 相较于随机森林表现出统计学上显著的提升；而在布里尔分数方面，随机森林略微更常表现出优势，但差异并不显著。对数据集特定性能度量值的补充分析（见补充材料 B.3 节）显示，在绝大多数情况下，两种方法之间的差异都很小。

5 模拟研究：比较统一变量重要性度量与传统变量重要性度量

统一变量重要性度量旨在识别具有纯边际效应的协变量，以及主要通过与其他协变量的交互作用影响结果变量的协变量——这些交互协变量本身可能具有边际效应，也可能没有。在此背景下，UFO框架的两个核心组成部分尤为重要：1）同时优化树根处的分割，以避免低估没有边际效应的交互协变量；2）在统一变量重要性度量的计算中，仅纳入每个协变量最具区分性的分割，从而确保充分考虑那些仅在由其他协变量定义的特定子群体中发挥作用的协变量的影响。

相比之下，传统的基于随机森林的变量重要性度量源自标准树，在标准树中，每次分割都只根据当前节点的局部判别能力来选择。此外，这些变量重要性度量会汇总每个协变量在所有分裂中的信息，这可能会低估仅在特定交互条件下发挥作用的协变量的重要性。

鉴于这些概念上的差异，本次模拟研究的主要目标是比较统一变量重要性度量（Unity vim）与两种传统的基于随机森林的变量重要性度量——置换变量重要性度量（Perm）和校正基尼重要性（Gini）——区分信息性协变量（具有边际效应或交互效应）与非信息性协变量的能力。此外，我们还考察了这三种变量重要性度量对具有不同效应类型的协变量进行排序时，排序结果有何差异。

5.1 研究设计

用于计算三种待比较变量重要性度量的 UFO（统一变量重要性度量）和传统随机森林（置换变量重要性度量、校正基尼重要性）的配置方式，与第 4.3 节所述真实数据分析中的配置相同，以确保结果具有可比性。

5.1.1 数据生成过程

我们考察了两种数据生成过程（DGP），两者的二元结局均为平衡的。第一种 DGP 仅包含连续协变量，第二种则同时包含连续协变量和分类协变量。

对于每种 DGP，我们考察了100、300、500 和 1,000 的样本量，并针对每种样本量生成了 1,000 个数据集。

数据生成过程 *1*：仅含连续协变量

这种 DGP（下文称为 DGP 1）改编自 Hornung 和 Boulesteix [2022], ，他们在一项比较基于随机森林的方法对两两交互效应进行排序的研究中采用了这一过程。该过程包含 18 个信息性协变量和 50 个非信息性协变量。

图 1展示了一个模拟数据集中各类信息性效应的代表性协变量取值。左侧面板显示了两个仅具有边际效应的协变量，中间面板展示了一对存在定量交互作用的协变量（两者均具有边际效应），右侧面板则展示了一对存在定性交互作用、但不具有边际效应的协变量。定量（双变量）交互是指一种情况：一个协变量效应的强度取决于另一个协变量，但其方向不受影响。相比之下，如前所述，定性交互则同时涉及强度和方向的变化。补充材料 C.1.1 节详细描述了图中所示协变量的效应机制。1。

协变量类型 $X_{mrg\ s}$、$X_{qn\ s}$和 $X_{ql\ s}$ 如图 1 所示，代表强效应。对于其中每一种类型，我们还在数据生成过程（DGP）中加入了具有中等和弱效应强度的变体，这些变体具有相同的机制，但关联强度不同。这些变体通过下标“ m”（中等）和“ w”（弱）表示，用于替换变量名称中的“ s”（强）。

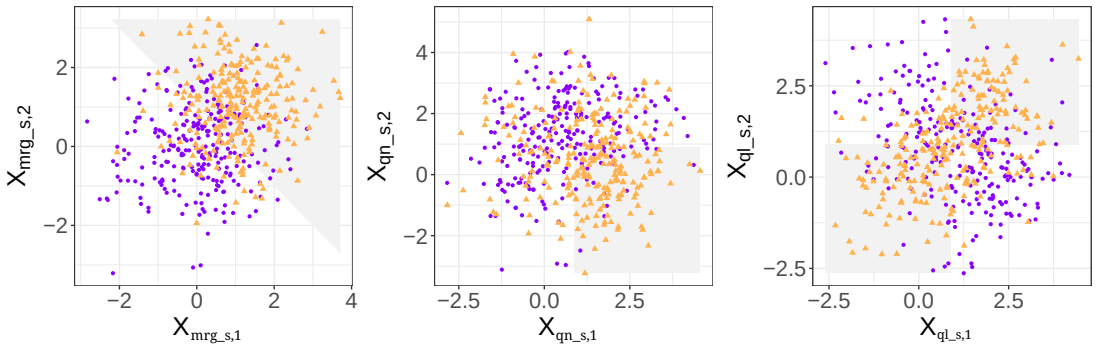

图1：从 DGP 1（$n = 500$）生成的模拟数据集中，具有强效应的变量对示例。圆形和三角形分别表示类别1和类别2的观测值。阴影区域表示协变量空间中类别2观测值浓度高于平均水平的区域

DGP 1 中的 18 个信息性协变量包括：三种纯边际效应类型各对应两个协变量，以及六种交互类型各对应一对协变量。

该数据生成过程的详细设定见补充材料的 C.2.1 节。

数据生成过程 *2*：连续协变量和分类协变量

第二个数据生成过程（简称 DGP 2）包含 19 个信息性协变量，其中四个为分类变量；此外还包含 50 个连续的非信息性协变量。

三个信息性协变量 $X_{mrg\ s}$、$X_{mrg\ m}$ 和 $X_{mrg\ w}$ 仅具有与 DGP 1 中相同的边际效应，因此此处不再进一步讨论。图2 展示了一个模拟数据集中其余所有信息性协变量的经验效应。每一行对应两个协变量之间的一种成对交互。补充材料的 C.1.2 节详细说明了这些效应机制。

与 DGP 1 一样，我们还纳入了图 2 中所示强效应协变量类型的中等和弱效应大小变体。具体而言，除 $X_{ql\ bne\ s}$、$X_{ql\ be\ s}$、$X_{qn\ bne\ s}$ 和 $X_{ql\ cne\ s}$ 外，我们还纳入了相应的中等（“ m”）和弱 -(“ w”）效应变体。其中，缩写“bne”、“be”和“cne”表示交互作用中所涉及的分类协变量类型：“bne”指无边际效应的二元协变量，“be”指具有边际效应的二元协变量，而“cne”指具有三个水平且无边际效应的分类协变量（见图2）。DGP 2 中的 19 个信息性协变量来自为每种所考虑的效应类型纳入一个代表性协变量。

详细　描述该数据生成过程的内容见补充材料 C.2.2 节　　.

5.1.2 评估

我们使用两个指标来评估结果：AUC值和中位排名。

AUC用于量化每种VIM区分信息性协变量与非信息性协变量的能力。对于每种信息性协变量类型，我们计算了一个AUC值，用以表示VIM将该类型的协变量排在非信息性协变量之前的概率。因此，AUC值越高，表明VIM区分信息性协变量与非信息性协变量的能力越强。为计算这些AUC值，我们针对每个模拟数据集中的每种信息性协变量类型分别计算AUC，并以VIM得分作为评分变量。在计算中，第一组包含数据生成过程中的相应协变量类型的代表（或多个代表），第二组包含50个非信息性协变量。由于第一组在DGP 1中仅包含两个观测值，在DGP 2中仅包含一个观测值，因此所得AUC值自然呈现较高的方差。然而，

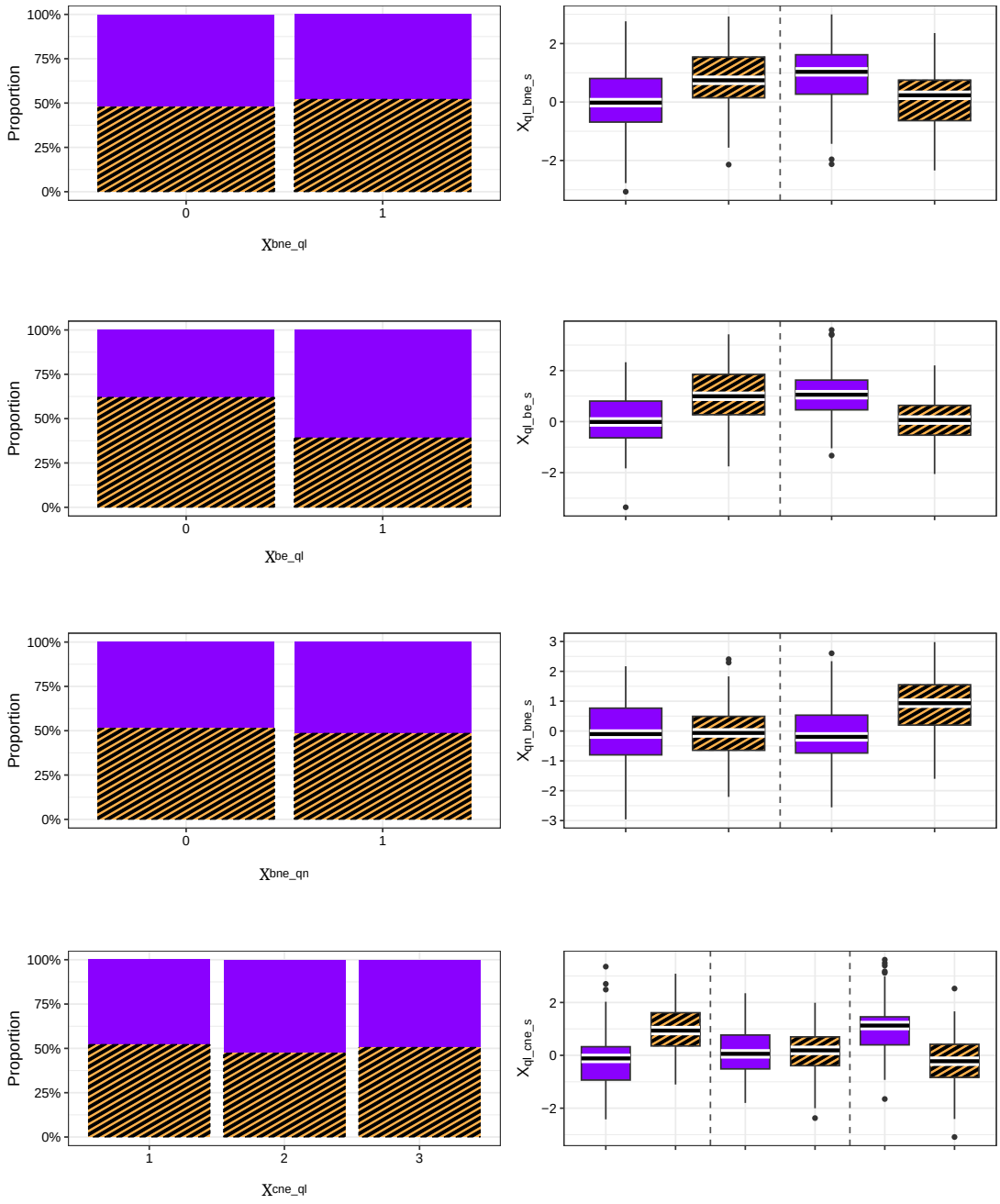

图 2：由 DGP 2（$n = 500$）生成的模拟数据集中交互协变量的经验效应。条形图中，实色区域和斜线填充区域分别表示类别 1 和类别 2 的比例。箱线图中，实色箱体和斜线填充箱体分别对应类别 1 和类别 2 的协变量值

对这些求平均 所有1,000个模拟数据集（按样本量分别统计）中的AUC值大幅

降低了这一方差。为量化剩余的不确定性，利用均值标准误并基于样本均值的渐近正态性计算了95%置信区间。

为评估不同效应类型的协变量在不同变量重要性度量下的排名差异，我们还分析了这些协变量在所有可用协变量中的排名。具体而言，对于每个模拟数据集，我们根据变量重要性度量值确定所有协变量的排名。随后，对于每种信息性协变量类型，我们计算了所有数据集中的中位排名（按样本量分层）。为评估这些排名的变异性，我们还计算了各数据集排名的第一和第三四分位数。排名越低，表示相应变量重要性度量赋予的协变量重要性越高。

5.2 结果

下文将同时报告曲线下面积值和中位排名方面的结果。这两项指标反映了性能的不同方面，但在我们的模拟中呈现出总体一致的模式：较高的曲线下面积值通常与较低的中位排名相关，反之亦然。因此，我们主要关注曲线下面积结果，仅在基于排名的结果能够提供额外见解时加以引用。

下面将详细描述结果，而第5.3 节给出简要总结。

5.2.1 DGP 1

图3 显示了曲线下面积值，图4 展示了相应的排名。数值结果见补充材料中的表S6和S7。

总体而言，对于所考虑的最小样本量（$n = 100$），统一变量重要性度量没有将任何信息性协变量排在非信息性协变量之前。使用统一变量重要性度量得到的大多数曲线下面积值都接近0.5，最大值为0.71。尽管使用`Perm` 和`Gini` 对 $n = 100$ 得到的曲线下面积值几乎无一例外都高于使用统一变量重要性度量得到的值，但即使是这些变量重要性度量，也只有在协变量具有纯粹的边际效应且效应同时较强或中等时，才能清楚地区分信息性协变量和非信息性协变量。对于 $n = 100$ ，统一变量重要性度量未能（可靠地）将任何信息性协变量排在非信息性协变量之前，这一事实表明，小型数据集可能不适合应用统一变量重要性度量。较高的算法复杂度可能导致小样本量下的结果存在较大变异；详见第7节的详细讨论。因此，以下对DGP 1结果的描述仅涉及样本量大于 $n = 100$的情形。

具有纯边际效应的协变量（$X_{mrg\ s}$、 $X_{mrg\ m}$和 $X_{mrg\ w}$）被 Perm 和 Gini 明确地区分于非信息性协变量。统一变量重要性度量也可靠地将这些协变量排在非信息性协变量之前，但在样本量较小的设置中，弱效应和中等效应的情况除外。

对于表现出定量交互效应和边际效应的协变量（$X_{qn\ s}$、$X_{qn\ m}$和 $X_{qn\ w}$），统一变量重要性度量的表现与传统变量重要性度量相当或更差，具体取决于效应大小和样本量。除样本量较小时的弱效应外，传统变量重要性度量对这类协变量的取值始终高于对非信息性协变量的取值。相比之下，统一变量重要性度量只有在样本量最大且效应为中等或强时，才能稳定地将这类协变量排在非信息性协变量之前。对于第二大样本量，统一变量重要性度量仅在强效应的情况下实现了明显区分。在其他所有情况下，统一变量重要性度量都未能稳定地区分这类协变量与非信息性协变量。基于排名的结果也基本反映了这些发现（图4）。值得注意的是，弱效应协变量的排名在统一变量重要性度量下非常靠前，这证实它无法将此类弱定量交互效应与噪声协变量区分开来。对于中等效应和第二大样本量，统一变量重要性度量取得的中位排名为 13，低于

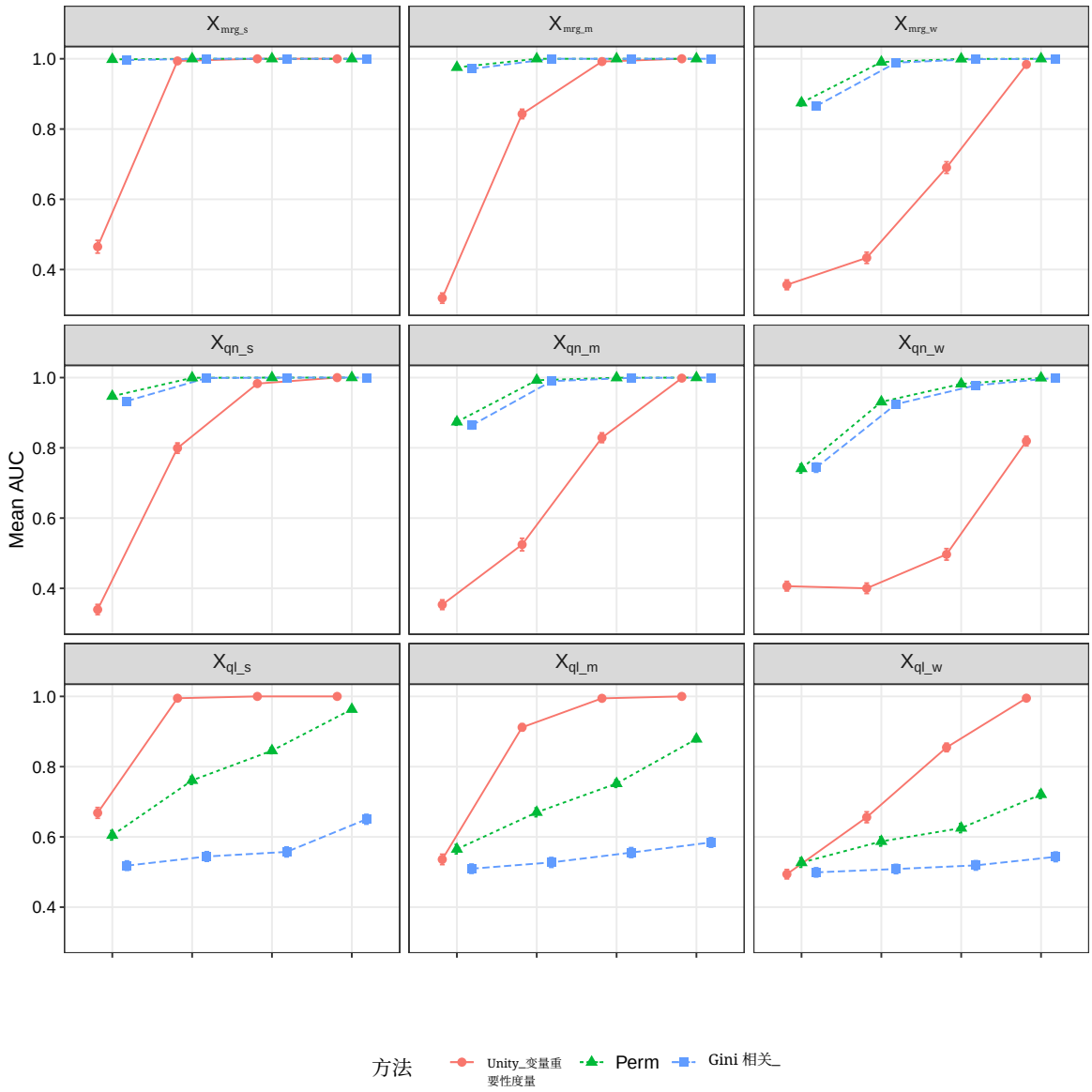

图3：DGP 1中各所考虑样本量和方法对应的平均AUC值及95%置信区间

根据相对适中的曲线下面积值 0.83 所预期的排名；然而，相应排名的变异性相当大。

在检测具有定性交互效应且无边际效应的协变量（$X_{ql\ s}$、$X_{ql\ m}$和 $X_{ql\ w}$）时，统一变量重要性度量始终优于传统变量重要性度量。对于强效应和中等效应，统一变量重要性度量能够可靠地将此类协变量排在非信息性协变量之上；仅在 $n = 300$ 的中等效应情况下存在一个小例外，此时AUC略降至0.91。对于弱效应，只有在样本量最大时，统一变量重要性度量才能将其与非信息性协变量清晰区分开来。

相比之下，Perm 仅在样本量最大且效应较强时，才能对这种协变量类型实现明确区分；对于中等效应，区分效果不完全（AUC $= 0.88$）。在

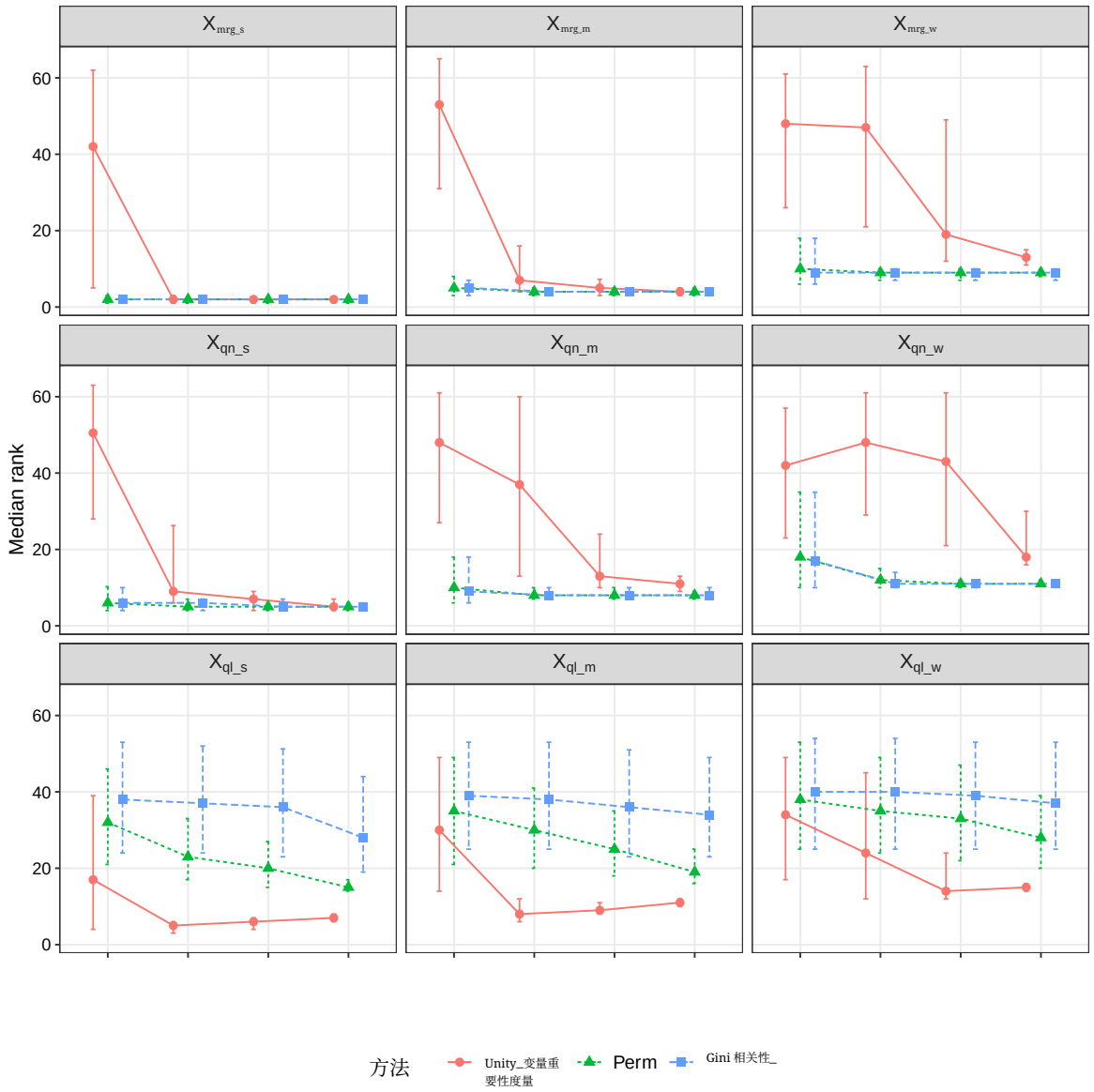

图4：中位数 针对 DGP 1，按所考察的样本量和方法计算的排名及25%和75%分位数

所有其他情况下，`Perm` 都无法进行明确区分。对于`Gini`，这种协变量类型的 AUC 值始终接近 0.5，这表明`Gini`无法将这种协变量类型与非信息性协变量区分开来。第7节对`Perm` 与`Gini` 在检测不存在边际效应的交互效应时性能差异的可能原因作了说明。

尽管 Perm 在识别此类协变量方面成效有限，但其中位排名从未低于 15，这表明如果只关注排名靠前的协变量，此类协变量很可能会被忽略。相比之下，对于强效应和中等效应，使用统一变量重要性度量获得的此类协变量的中位排名均低于 10（有一个例外）。

5.2.2 DGP 2

DGP 2 的结果展示于图5 和6中。底层数值结果见补充材料中的表S8–S11。

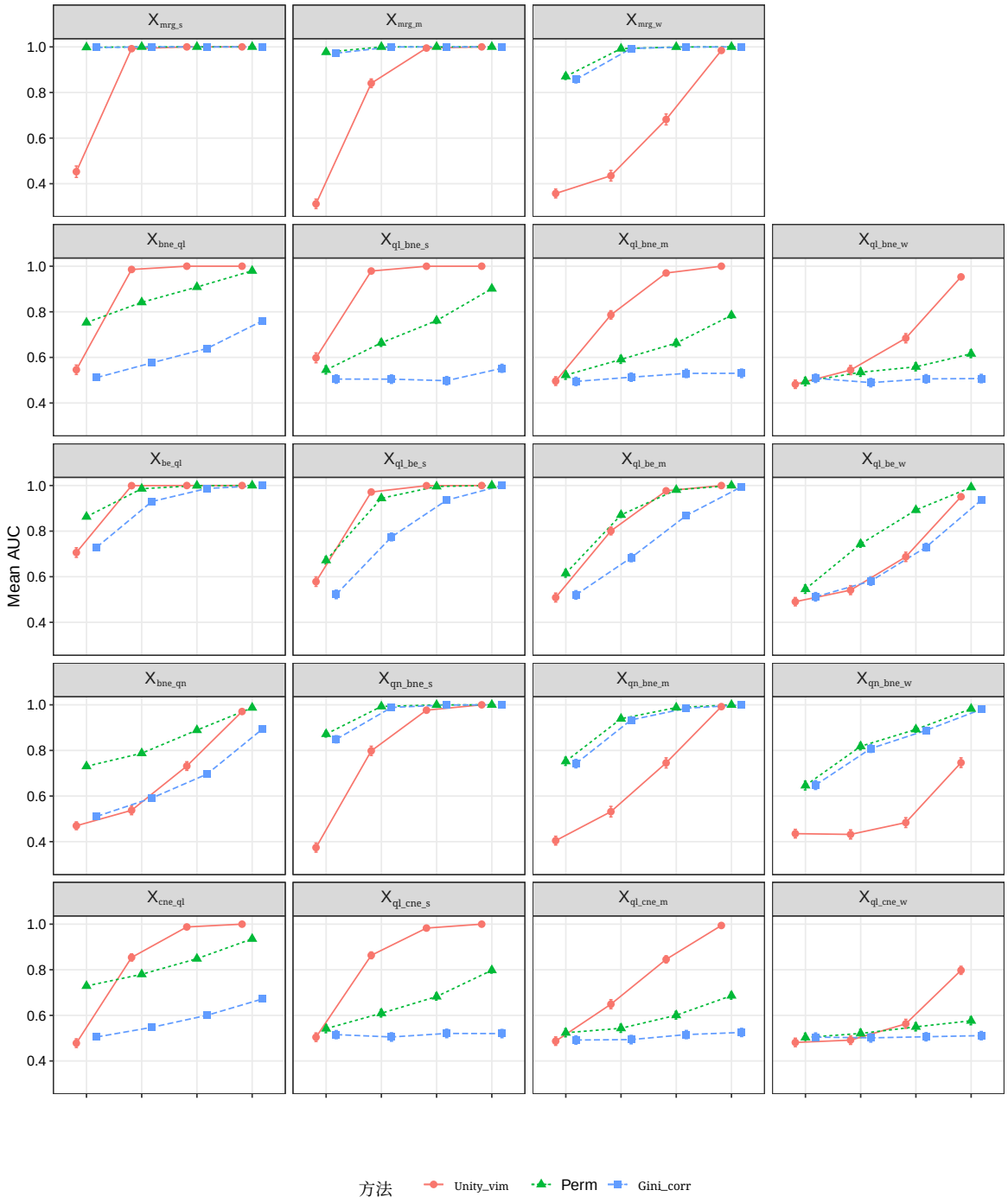

图 5：DGP 2 中各样本量和方法的平均 AUC 值及 95%置信区间

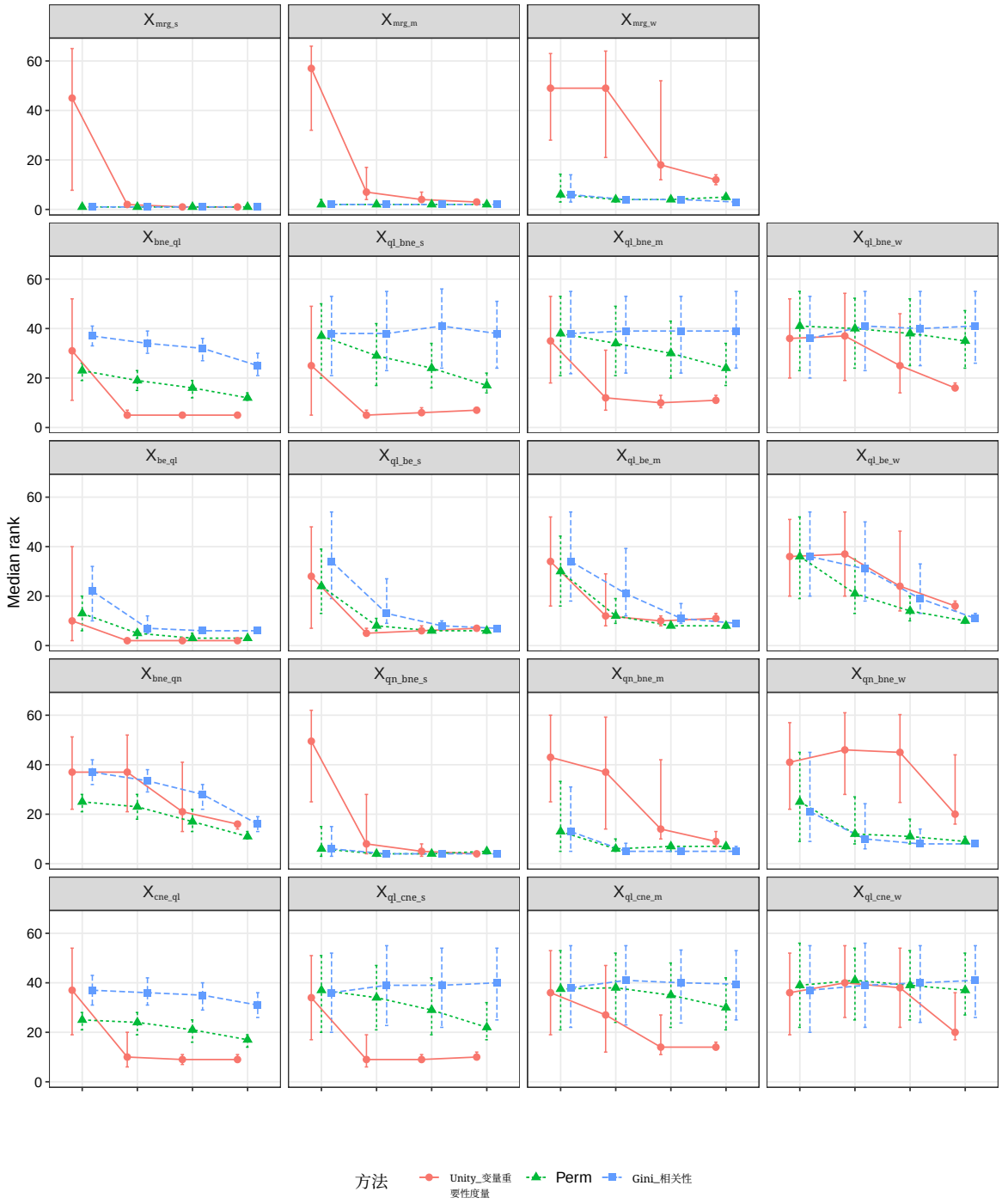

图6：DGP 2 中各个样本量和方法对应的中位排名及其25%和75%分位数

对于最小样本量（$n = 100$），研究结果与 DGP 1 的结果相一致。在这一设置下，统一变量重要性度量未能将任何信息性协变量与非信息性协变量区分开来。相比之下，对于强效应和中等效应，传统变量重要性度量始终将具有纯边际效应的协变量排在非信息性协变量之前，但在其余情况下未能实现明确区分。与上一小节一样，以下讨论

仅涉及样本量大于 $n = 100$ 的设置。

具有纯边际效应的协变量（$X_{mrg\ s}$、$X_{mrg\ m}$ 和 $X_{mrg\ w}$）的结果与 DGP 1 的结果非常相似。除样本量较小时 Unity vim 对弱效应和中等效应的情况外，所有变量重要性度量都能可靠地将这些协变量排在非信息性协变量之前。

为使讨论更简洁，以下提及特定协变量类型时，我们省略下标“ s”、“ m”和“ w”；例如，$X_{ql\ bne}$ 统称由 $X_{ql\ bne\ s}$、$X_{ql\ bne\ m}$ 和 $X_{ql\ bne\ w}$ 表示的协变量类型。

没有边际效应的二元协变量 $X_{bne\ ql}$——它与连续协变量 $X_{ql\ bne}$ 一同参与了一个不含边际效应的定性交互效应——始终被 Unity vim 排在非信息性协变量之前。相比之下，Perm 只有在样本量最大时才能可靠地区分该协变量与非信息性协变量，而 Gini 在所考察的任何样本量下都未能做到这一点。

同样，对于连续协变量 $X_{ql\ bne}$，统一变量重要性度量的表现优于传统变量重要性度量。在所有样本量下，统一变量重要性度量都能可靠地区分强效应与非信息性协变量；在较大样本量下，也能可靠地区分中等效应；即使是弱效应，在最大样本量下也能以近乎可靠的程度加以区分（曲线下面积 $= 0.95$）。相比之下，只有在样本量最大且效应强的情况下，Perm 才会将此类协变量排在非信息性协变量之前，而且这种情况也并不稳定（曲线下面积 $= 0.90$）。在其他所有情况下，Perm 排列此类协变量的位置都没有显著高于噪声协变量。Gini 得出的曲线下面积值始终接近 0.5，表明该变量重要性度量无法区分此类协变量与非信息性协变量。

二元协变量 $X_{be\ ql}$ 与连续协变量 $X_{ql\ be}$ 之间的定性交互效应与 $X_{bne\ ql}$ 和 $X_{ql\ bne}$ 之间的定性交互效应相似，不同之处在于，$X_{be\ ql}$ 还具有中等边际效应。与对 $X_{bne\ ql}$ 的表现类似，统一变量重要性度量能够可靠地区分 $X_{be\ ql}$ 与非信息性协变量。但与 $X_{bne\ ql}$ 不同，传统变量重要性度量也能清晰地区分 $X_{be\ ql}$——这可能是由于存在边际效应。唯一的例外是 $n = 300$：Gini 将此协变量排在非信息性协变量之前的表现并不稳定，只是频率很高（曲线下面积 $= 0.93$）。从排名（图6）中也能看出 $X_{be\ ql}$ 和 $X_{bne\ ql}$ 在表现上的差异：对于传统变量重要性度量，$X_{bne\ ql}$ 的排名相对较高，而 $X_{be\ ql}$ 在这些变量重要性度量中的排名始终位列前十。统一变量重要性度量将 $X_{bne\ ql}$ 和 $X_{be\ ql}$ 都排在前十，且给予 $X_{be\ ql}$ 的排名甚至高于传统变量重要性度量。

对于连续协变量 $X_{ql\ be}$，Unity vim 的表现与 Perm 相近。样本量较大时，在中等效应或强效应的情况下，这两种方法都能可靠地区分该协变量与非信息性协变量。对于弱效应，Unity vim 的 AUC 值明显低于 Perm。不过，即使是 Perm，在弱效应下也只有样本量最大时才能实现清晰区分（AUC $= 0.99$）；此时 Unity vim 的表现也不错（AUC $= 0.95$）。总体而言，Gini 的表现不如 Unity vim 和 Perm，只有在样本量最大且效应为强效应或中等效应时，才能持续实现清晰区分。

对于与连续协变量 $X_{qn\ bne}$ 存在定量交互作用的二元协变量 $X_{bne\ qn}$，只有 Perm 能始终将其排在非信息性协变量之上，而 Unity vim 也几乎始终如此（AUC $= 0.97$）；这两种方法在样本量最大时均有此表现。样本量较小时，未观察到清晰区分，但 Perm 的 AUC 值显著高于 Unity vim 和 Gini。对于 $X_{qn\ bne}$，在所考察的样本量范围内，Perm 和 Gini 在强效应下均能可靠地区分该协变量与非信息性协变量；在中等效应下，样本量较大时也能做到；在弱效应下，则只有样本量最大时才能做到。Unity vim 也呈现出总体相同的模式，但效果不那么明显：$X_{qn\ bne}$ 只有在强效应且样本量较大时才始终排名靠前；在中等效应下，只有样本量最大时才始终排名靠前；在弱效应下则完全没有始终排名靠前。

具有三个水平的分类协变量 $X_{cne\ ql}$——它没有边际效应，但与同样没有边际效应的连续协变量 $X_{ql\ cne}$ 存在定性交互作用——

只有统一变量重要性度量才能将其与非信息性协变量明确区分开来，而且仅在样本量较大时如此。对于任何样本量，传统变量重要性度量都无法稳定地区分该协变量与非信息性协变量，不过 Perm 的表现略好一些：在样本量最大时，它经常（但并非始终）将该协变量排在较高位置（曲线下面积 $= 0.94$）。

对于连续协变量 $X_{ql\ cne}$，统一变量重要性度量再次优于两种传统变量重要性度量。统一变量重要性度量能够可靠地将该协变量排在非信息性协变量之前：在强效应和中等效应情况下，样本量最大时如此；在强效应情况下，样本量第二大时也如此。如图6所示，$X_{ql\ cne}$ 和 $X_{cne\ ql}$ 的排名往往略低于其他信息性协变量类型。相比之下，对于任何效应大小或样本量，传统变量重要性度量都无法可靠地区分 $X_{ql\ cne}$ 与非信息性协变量，尤其是`Gini`得到的曲线下面积值始终接近 0.5。

5.3 总结与结论

与传统变量重要性度量不同，统一变量重要性度量能够可靠地区分参与交互效应但没有边际效应的协变量与非信息性协变量。对于所考察的任何效应大小或样本量，校正基尼重要性都未能识别此类协变量；而在强效应和大样本量的情况下，置换重要性经常（但并非始终）将它们排在非信息性协变量之前。

所有变量重要性度量都清楚地将那些自身没有边际效应、但与具有边际效应的协变量发生交互的协变量识别为具有影响力的协变量。

对于较小的样本量和较弱的效应强度，统一变量重要性度量在区分具有边际效应的协变量与非信息性协变量方面，不如传统变量重要性度量可靠。一种合理的解释是，统一变量重要性度量的变异性较高，这可能源于 UFO 的算法复杂度更高。从基于排名的结果（图 4和 6）中也能看出这种变异性：对于许多协变量，统一变量重要性度量的四分位距通常大于传统变量重要性度量的四分位距。

对于所考察的最小样本量（$n = 100$），统一变量重要性度量未将任何信息性协变量类型排在非信息性协变量之上。这表明，对于非常小的数据集，解读统一变量重要性度量的结果时应谨慎。

6 CRTR算法的实证评估

在本节中，我们首先将 CRTR算法应用于两个大小为 $n = 500$的模拟数据集，这些数据集由第5节所述的两个数据生成过程生成，以考察算法识别出的关联结构与真实底层结构的吻合程度。随后，我们将该算法应用于著名的葡萄酒数据集（$n = 178$，$p = 13$），以作示例说明。在所有分析中，我们均采用第3 节以及补充材料 A.3 和 A.4 节所述的 UFO框架默认超参数值。

6.1 模拟数据评估

受篇幅所限，正文中仅展示 unity VIM值最高的五个协变量对应的 CRTR。在补充材料 D.1 节中，我们展示并描述了两个模拟数据集各自 unity VIM值和置换变量重要性度量值最高的 20 项，并提供了更多说明。

图7 和8 展示了使用 DGP 1 生成的数据集中 unity VIM值最高的五个协变量对应的 CRTR，以及各 CRTR 得分最高的分裂所对应节点中类别特定协变量分布的核密度估计。[1]

[1]如第 5.1节所述，DGP 1 对每种协变量类型和效应大小均包含两个协变量。为避免重复在一对协变量的两个成员均位列排名最高的五个协变量之列的情况下，统一性较低的协变量，

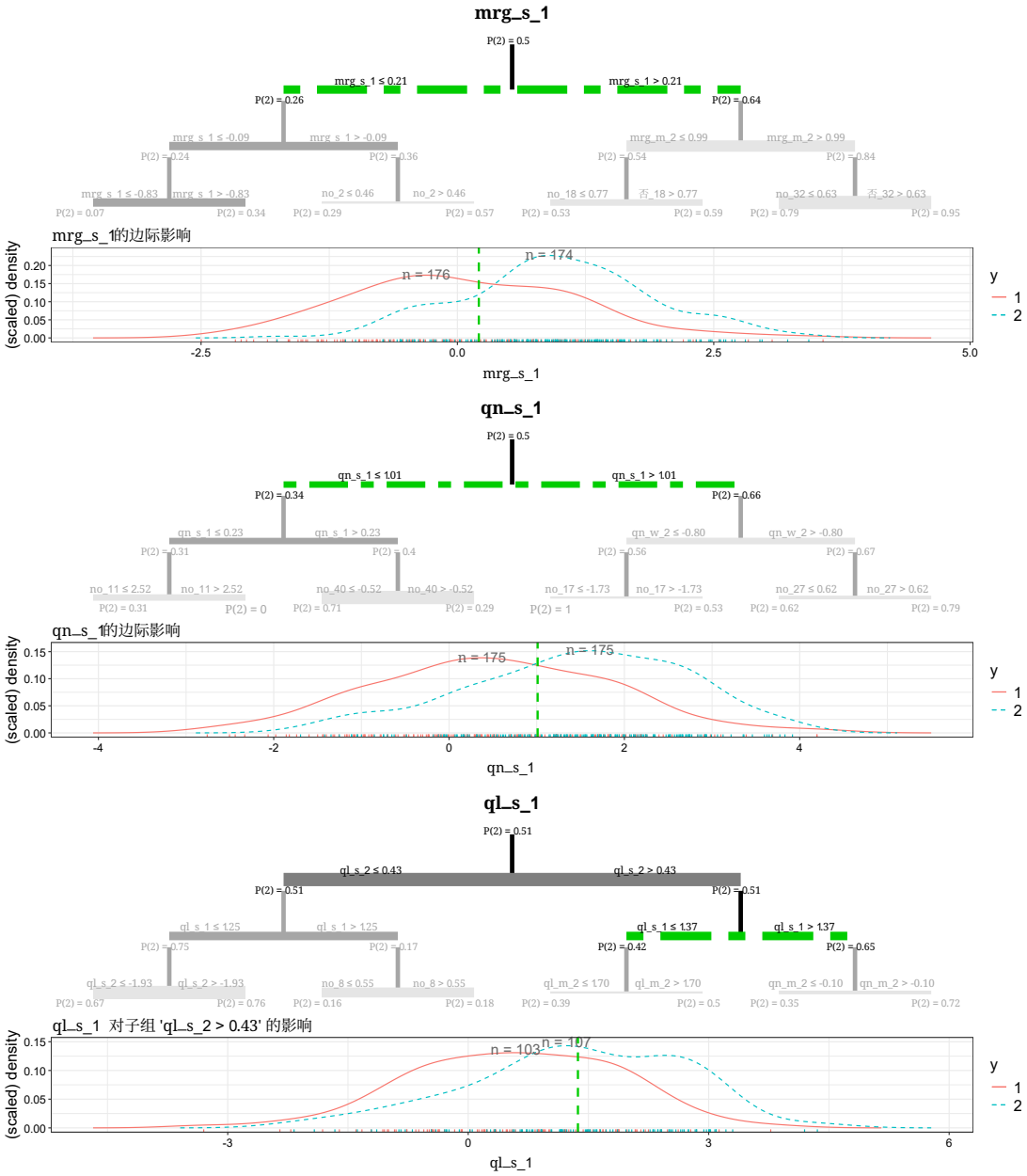

图 7：展示由 DGP 1（$n = 500$）生成的数据集中，统一变量重要性度量值最高的三个协变量对应的 CRTR。水平线的粗细反映协变量得分，虚线表示得分最高的分裂。灰色阴影区域标示不属于得分最高的分裂及其祖先节点的树区域。得分最高的节点中，各类别特定协变量分布的核密度估计按类别比例加权。节点处的类别频数分布和核密度估计均使用袋内数据计算。P(2) 表示相应节点中属于类别 2的观测值的袋内比例。协变量标签采用缩写（例如，mrg s 1 表示 $X_{mrg\ s\ 1}$，不带 * 表示非信息性协变量）

已排除变量重要性度量值。因此，所示的每棵条件回归树都代表协变量类型与效应大小的不同组合。

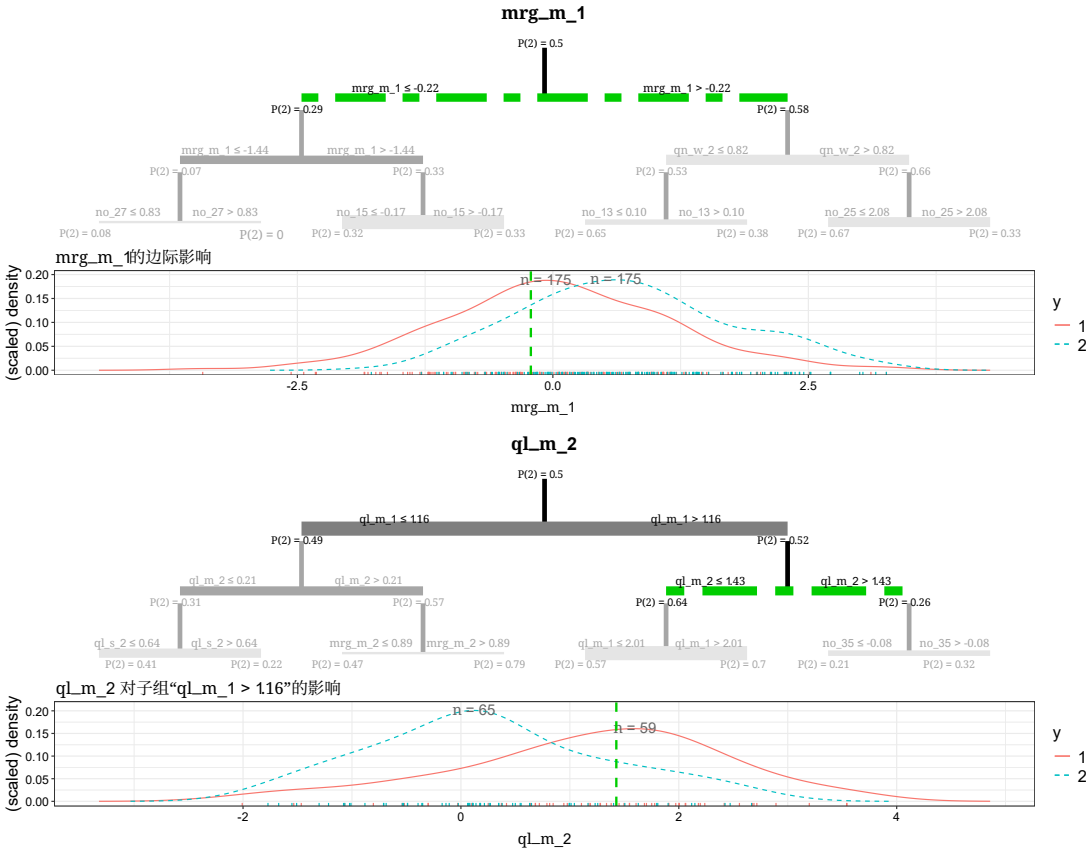

图8：在由 DGP 1（$n = 500$）生成的数据集中，根据统一变量重要性度量，排名第四和第五的协变量对应的 CRTR 的可视化。水平线的粗细反映协变量得分，虚线表示得分最高的分裂，灰色阴影区域标示这些分裂及其祖先节点以外的树区域。节点处的类别频数分布以及类别特定协变量分布的加权核密度估计，均使用袋内数据计算。P(2)表示相应节点中属于类别2的观测值所占的袋内比例。协变量标签采用缩写（例如，mrg m 1 表示 $X_{mrg\ m\ 1}$，不带 * 表示非信息性协变量）

在讨论实质性解释之前，我们先介绍本节使用的 CRTR 可视化方法；经过略微修改的版本也已在 R软件包 unityForest 中实现。在每幅可视化图中，均显示节点处袋内观测值的相对类别频数分布。对于二元结局，如本例所示，只需显示两个类别中其中一个类别的频数即可。

水平线的粗细对应于第 3.3 节算法详细说明中所述的协变量得分。这些得分反映了相较于森林中所有树根处的总体出现频率，相应分裂协变量在所考察协变量的最佳树根处分割时出现的频率高出多少。因此，某个分裂协变量对应的水平线特别粗，意味着该协变量在所获得的条件回归树（CRTR）所针对的协变量具有显著影响的树根处分割时出现得尤为频繁。这表明该协变量可能与我们感兴趣的协变量存在交互作用。

得分最高的分裂以水平虚线标示。为便于理解，最好将 CRTR 中不属于包含得分最高分裂的节点或其

祖先节点的区域用阴影标出，本文所示图例即采用了这种做法。采用这一方法有两个原因。首先，CRTR 的目的是描绘相应协变量产生最强效应的条件，而这些条件由包含得分最高分裂的节点的祖先节点决定。其次，CRTR 中不属于包含得分最高分裂的节点或其祖先节点的区域，解读价值有限。每个 CRTR 都源自相应协变量表现出特别强效应的树根。因此，这些树根中频繁出现的关联模式具有该协变量的特异性。相比之下，CRTR 的阴影区域反映的是数据整体关联结构中任意的方面。此外，这些区域也可能包含非信息性协变量上的分裂，因为每个树根都是从所有可用协变量中随机抽取的一个小子集构建的，而其中许多协变量可能并无信息性。

与此同时，所考察的协变量也可能在阴影区域内发生分割，这些分割对于解释也可能具有参考意义。然而，这些分割并不属于定义 CRTR 的得分最高的分裂，因此通常并不代表该协变量表现出最强效应的条件。

最后，得分最高的分裂所在节点中类别特定协变量分布的核密度估计，有助于评估协变量效应的方向和强度。为提高可解释性，密度按相应类别中观测值的比例进行缩放，因此可将其解读为彼此相对的局部类别特定密度。

图 7 和 8 所示的 CRTR 很好地反映了相应协变量的真实效应类型。协变量 $X_{qn\ s\ 1}$ 与 $X_{qn\ s\ 2}$之间的定量交互效应并未体现在协变量 $X_{qn\ s\ 1}$的 CRTR 中，尽管二者也都表现出边际效应。然而，统一变量重要性度量和 CRTR 的目的并非检测所有可能类型的交互效应，而是识别相应协变量产生最强影响的条件。$X_{qn\ s\ 1}$ 不仅由于交互作用，对 $X_{qn\ s\ 2}$ 值较小的观测值具有较强效应，同时也具有边际效应。再加上得分最高的分裂按节点大小加权，这很可能使 $X_{qn\ s\ 1}$ 中的得分最高分裂成为 CRTR 中的第一个分裂，从而反映出其边际影响，而非基于交互的影响。

$X_{mrg\ s\ 1}$ 和 $X_{mrg\ m\ 1}$的CRTR仅具有边际效应，能够正确表示其影响类型。同样，$X_{ql\ s\ 1}$ 和 $X_{ql\ m\ 2}$ 的CRTR也能正确捕捉其定性交互效应。这些协变量没有边际效应，但分别与$X_{ql\ s\ 2}$ 和 $X_{ql\ m\ 1}$发生交互，因此其效应方向会因 $X_{ql\ s\ 2}$ 或 $X_{ql\ m\ 1}$取值较小还是较大而反转。这种结构在相应的CRTR中清晰可见：在 $X_{ql\ s\ 1}$的CRTR中，首先在 $X_{ql\ s\ 2}$上进行分割，随后在根节点的两个子节点中均在 $X_{ql\ s\ 1}$ 上进行分割。$X_{ql\ m\ 2}$的CRTR也呈现出相同模式。补充材料中的图S6–S14展示了DGP 1数据集中所有信息性协变量的CRTR和得分最高的分割可视化结果。这些结果与正文中的结果一致，并进一步表明，即使是效应较弱的协变量，其CRTR也能正确地刻画它们。

在由 DGP 2 生成的数据集上，统一变量重要性度量排名最高的五个协变量对应的 CRTR 如图 9 和 10所示。这里，CRTR 同样反映了相应协变量的真实效应。对于具有纯边际效应的协变量 $X_{mrg\ s}$ 和$X_{mrg\ m}$，其 CRTR 中得分最高的分裂正确地出现在第一处。

在 $X_{qn\ bne\ s}$的 CRTR 中，它与二元协变量$X_{bne\ s}$具有定量交互效应。与 DGP 1 中定量交互协变量 $X_{qn\ s\ 1}$ 的情况不同，得分最高的分裂并非第一处分裂。相反，第一处分裂发生在$X_{bne\ s}$，随后在 $X_{qn\ bne\ s}$的子节点中出现得分最高的分裂，其中 $X_{bne\ s} = 1$。这种结构与数据生成过程一致。与 DGP 1 中 $X_{qn\ s\ 1}$ 的 CRTR（图 7）不同，此处得分最高的分裂并非第一处分裂，这可归因于两个数据生成过程中的定量交互效应有所不同。在给定 $X_{bne\ s} = 1$ 时（DGP 2），$X_{qn\ bne\ s}$ 的条件效应强于 $X_{qn\ s\ 1}$ 在 $X_{qn\ s\ 2}$ 取值较小时的条件效应（DGP 1）：当 $X_{bne\ s} = 1$时，$X_{qn\ bne\ s}$ 取值较小时一个类别占主导，而取值较大时另一个类别占主导；而当 $X_{qn\ s\ 2}$ 取值较小时，对于较小的 $X_{qn\ s\ 1}$ 取值，两种类别的占比相近，只有当 $X_{qn\ s\ 1}$ 较大时，一个类别才会

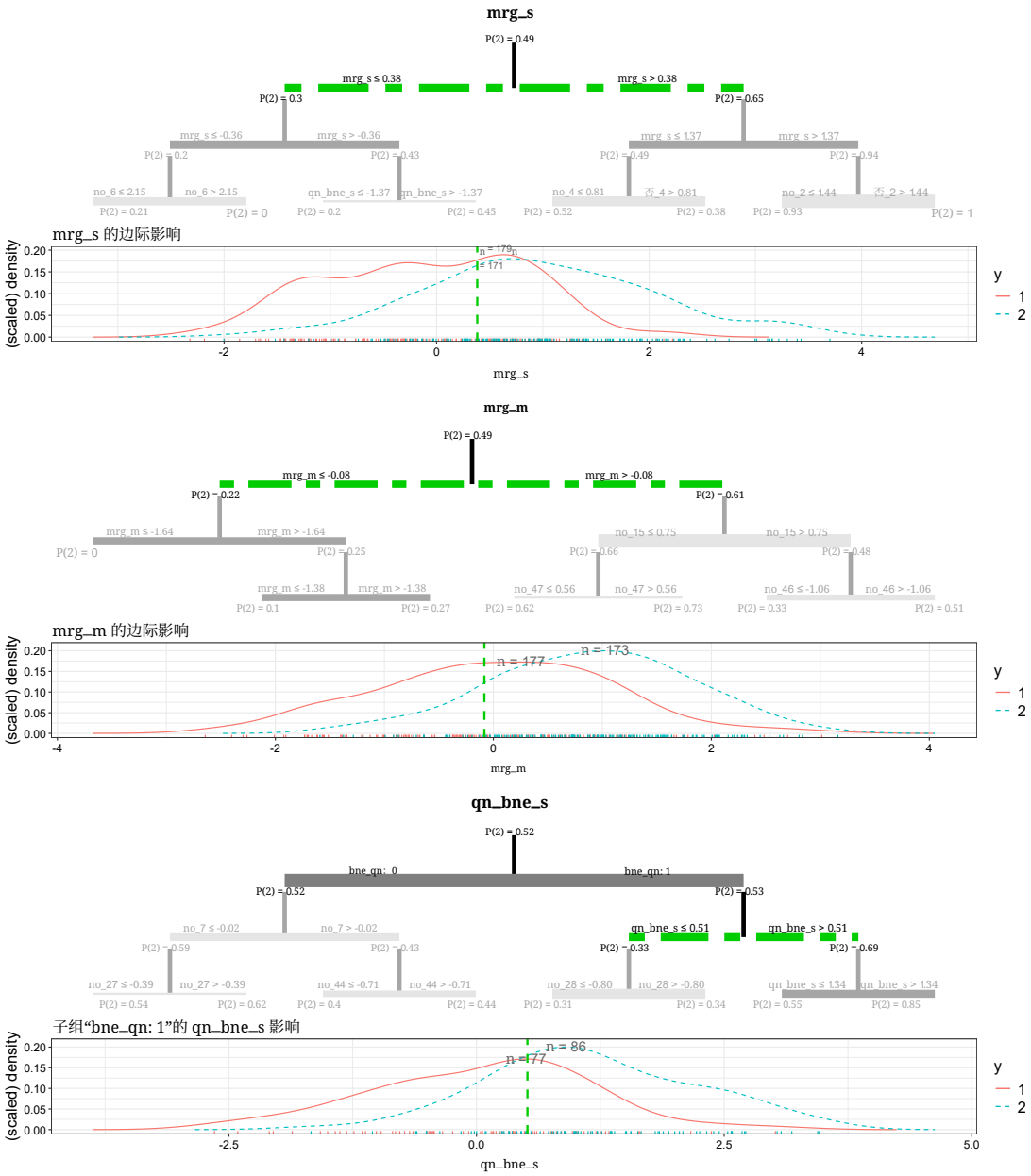

图9：在由 DGP 2 ($n = 500$) 生成的数据集中，统一变量重要性度量值最高的三个协变量对应的 CRTR 的可视化结果。水平线的粗细反映协变量得分，虚线表示得分最高的分裂，灰色阴影区域标示这些分裂及其祖先节点之外的树区域。节点处的类别频数分布以及类别特定的协变量分布的加权核密度估计，均使用袋内数据计算。P(2) 表示相应节点中属于类别2的观测值的袋内比例。协变量标签采用缩写（例如，mrg s 表示 $X_{mrg\ s}$，不带 * 表示非信息性协变量）

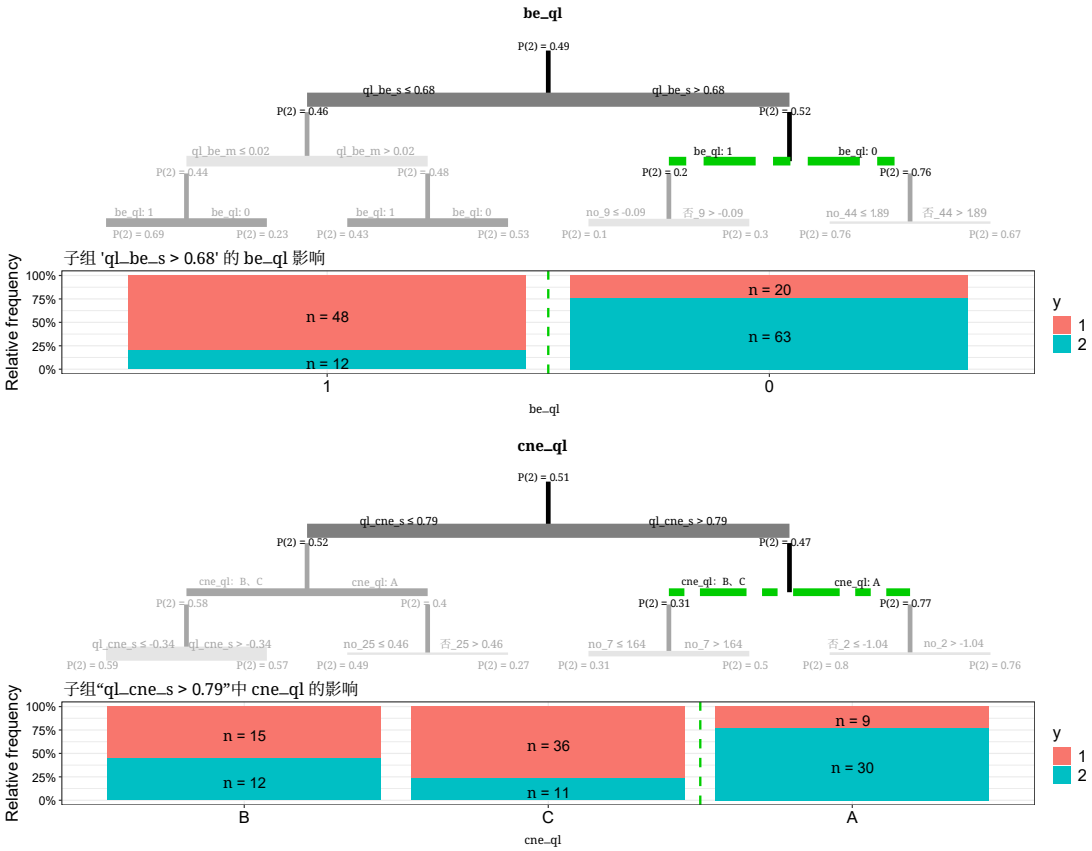

图10：根据统一变量重要性度量，在由 DGP 2 ($n = 500$) 生成的数据集中，对排名第四和第五的协变量对应的 CRTR 进行可视化。水平线的粗细反映协变量得分，虚线表示得分最高的分裂，灰色阴影区域表示这些分裂及其祖先节点之外的树区域。节点处的类别频数分布和条形图是使用袋内数据计算的。P(2) 表示相应节点中属于类别2的观测值的袋内比例。协变量标签采用缩写（例如，be ql 表示 $X_{be\ ql}$ ，而 no * 表示非信息性协变量）

开始占据主导地位。

排名第四和第五的协变量 $X_{be\ ql}$ 和 $X_{cne\ ql}$是分类协变量，分别与 $X_{ql\ be\ s}$ 和 $X_{ql\ cne\ s}$呈现定性交互效应。这一点也体现在它们的 CRTR 中：首先在 $X_{ql\ be\ s}$ 和 $X_{ql\ cne\ s}$上进行分割，随后分别在 $X_{be\ ql}$ 和$X_{cne\ ql}$上进行得分最高的分裂。尽管 $X_{be\ ql}$ 也具有边际效应，但在 $X_{ql\ be\ s}$ 值较高的观测值中，其影响要强得多。对于这些分类协变量，使用堆叠条形图而非核密度估计来展示分类协变量与二元结局之间的关系。

补充材料中的图 S15–S23 展示了所有信息性协变量的相应结果，与 DGP 1 的分析类似。这些结果证实了正文中所示的模式：CRTR 正确再现了协变量的真实效应类型。唯一的例外是 $X_{ql\ cne\ w}$ 的 CRTR，它未能捕捉到与分类协变量 $X_{cne\ ql}$之间较弱的交互效应。

6.2 葡萄酒数据集的应用

知名的葡萄酒数据集包含对意大利同一地区用三种葡萄品种（巴罗洛、格里尼奥利诺和巴贝拉）酿造的 178 种葡萄酒进行化学分析的结果。该数据集最初由 Forina [1984] 提出，随后由 Forina 等人 [1986]进行了详细介绍。每个样本测量了 13 种连续化学成分，作为区分不同葡萄品种的协变量。在我们的分析中，我们使用了该数据集的一个版本，其二元结局用于区分格里尼奥利诺（“G”）与另外两个品种（巴罗洛和巴贝拉）。此版本可在 OpenML 上获取，数据 ID 为 973。协变量包括酒精（`Alc`）、苹果酸（`Mal`）、灰分（`Ash`）、灰分碱度（`AlcAsh`）、镁（`Mg`）、总酚（`TP`）、黄酮类化合物（`Fla`）、非黄酮类酚（`NFP`）、原花青素（`ProAn`）、颜色强度（`Col`）、Hue（`Hue`）、稀释葡萄酒的 OD280/OD315（葡萄酒吸光度指数）（`WAI`）以及脯氨酸（`Prol`）。

与模拟数据的分析一样，我们聚焦于统一VIM值最高的五个协变量对应的 CRTR。正如第3.3节所述，CRTR 通常只有在协变量的统一VIM值足够高时才具有信息价值。所得的统一变量重要性度量值和置换VIM值见补充材料中的图 S24。根据这两种变量重要性度量对协变量进行的排序总体上相似，但也存在一些差异，这些差异也反映在下文所示的 CRTR 中。统一VIM值与置换VIM值之间的斯皮尔曼相关系数为 0.85。

统一VIM排名最高的五个协变量对应的 CRTR 如图11 和12所示。对于颜色强度和酒精这两个协变量，得分最高的分裂分别出现在对应 CRTR 的第一个分裂位置，这与相应核密度估计所显示的显著边际效应一致。

相比之下，脯氨酸和葡萄酒吸光度指数的 CRTR 结构各不相同：两种情况下，首次分割都发生在另一个协变量上，随后在由此产生的两个子节点中，分别按感兴趣的协变量进行分割。核密度估计证实，这些协变量的条件影响远强于其边际效应。具体而言，葡萄酒吸光度指数仅表现出微弱的边际效应。该协变量的 CRTR 还表明，葡萄酒吸光度指数与脯氨酸之间存在定性交互作用：当脯氨酸值较小时，葡萄酒吸光度指数似乎会提高格里尼奥利诺的概率；而当脯氨酸值较大时，它似乎会产生相反的影响。脯氨酸的条件效应强于边际效应这一现象，也反映在变量重要性度量中。尽管其统一变量重要性度量值仅略低于颜色强度，但按置换变量重要性度量排名，脯氨酸位居第三，其变量重要性度量值不到颜色强度的一半。

黄酮类化合物的条件回归树显示，其与 Grignolino 的概率之间存在非线性关联。在这里，第一次分割发生在黄酮类化合物上；根节点的左子节点中又出现了第二次分割，同样基于黄酮类化合物，而只有第二次分割的得分最高。黄酮类化合物含量较低或较高时，格里尼奥利诺的频率都较低；含量处于中间水平时，频率则较高。这种模式在条件回归树以及根据核密度图估算的边际效应中都很明显。

6.3 结论

针对模拟数据集得到的 CRTR 在整个过程中都很好地反映了信息性协变量的真实效应。对葡萄酒数据集进行 CRTR 分析，揭示了协变量效应的一些有趣见解，而这些见解在传统分析中很可能会被忽略。具体而言，CRTR 为深入解读 VIM 结果提供了有价值的手段，因为它既能捕捉边际效应，也能捕捉条件协变量效应。

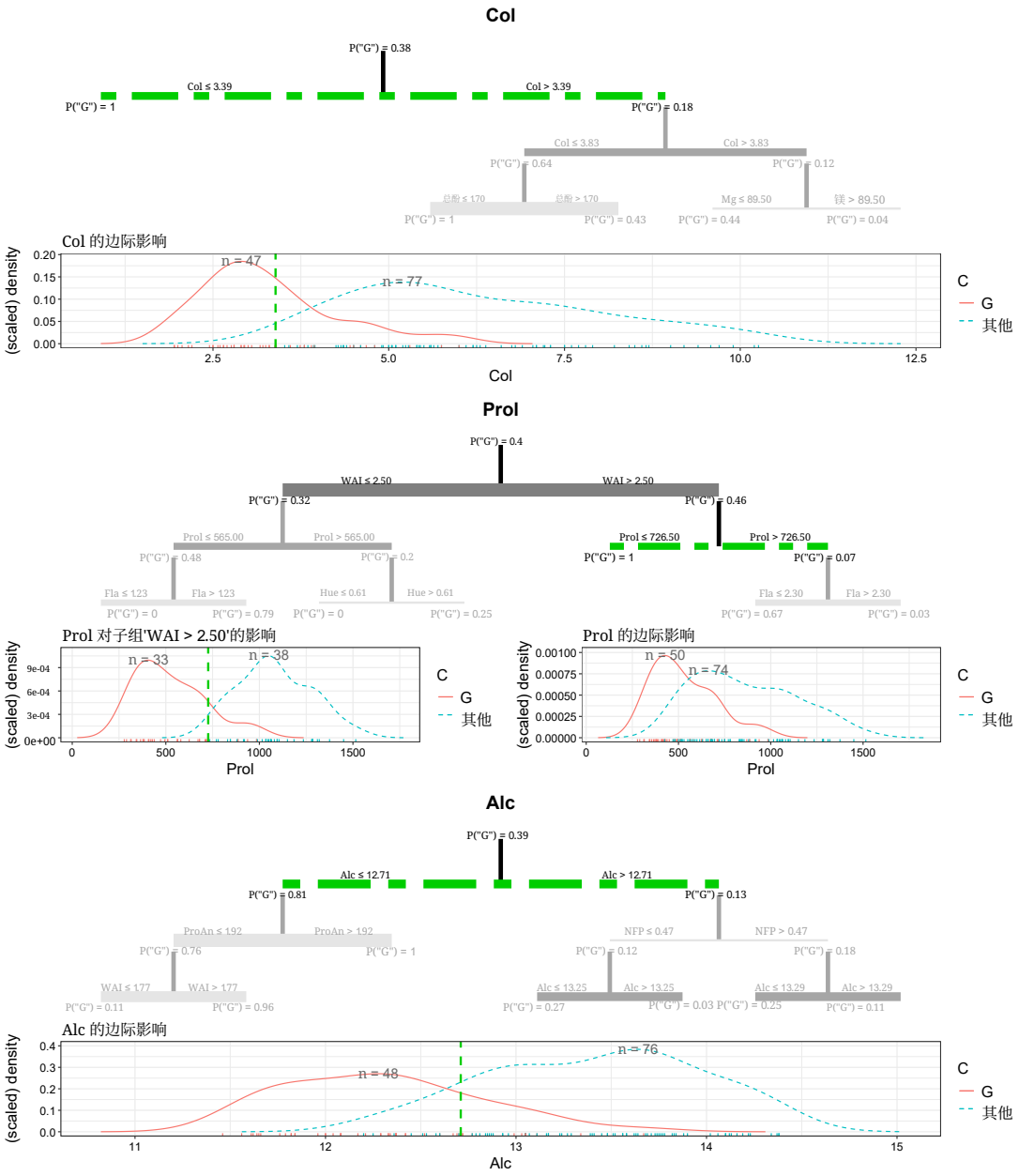

图11：葡萄酒数据集中统一变量重要性度量值最高的三个协变量所对应的 CRTR 的可视化结果（$n = 178$）。水平线的粗细反映协变量得分，虚线表示得分最高的分裂，灰色阴影区域标示这些分裂及其祖先节点之外的树区域。对于第一个分裂并非得分最高分裂的 CRTR，同时显示所有袋内观测值以及得分最高节点中的观测值的核密度估计。节点处的类别频数分布和类别特定协变量分布的加权核密度估计均使用袋内数据计算。P(“G”) 表示相应节点中属于格里尼奥利诺的观测值的袋内比例

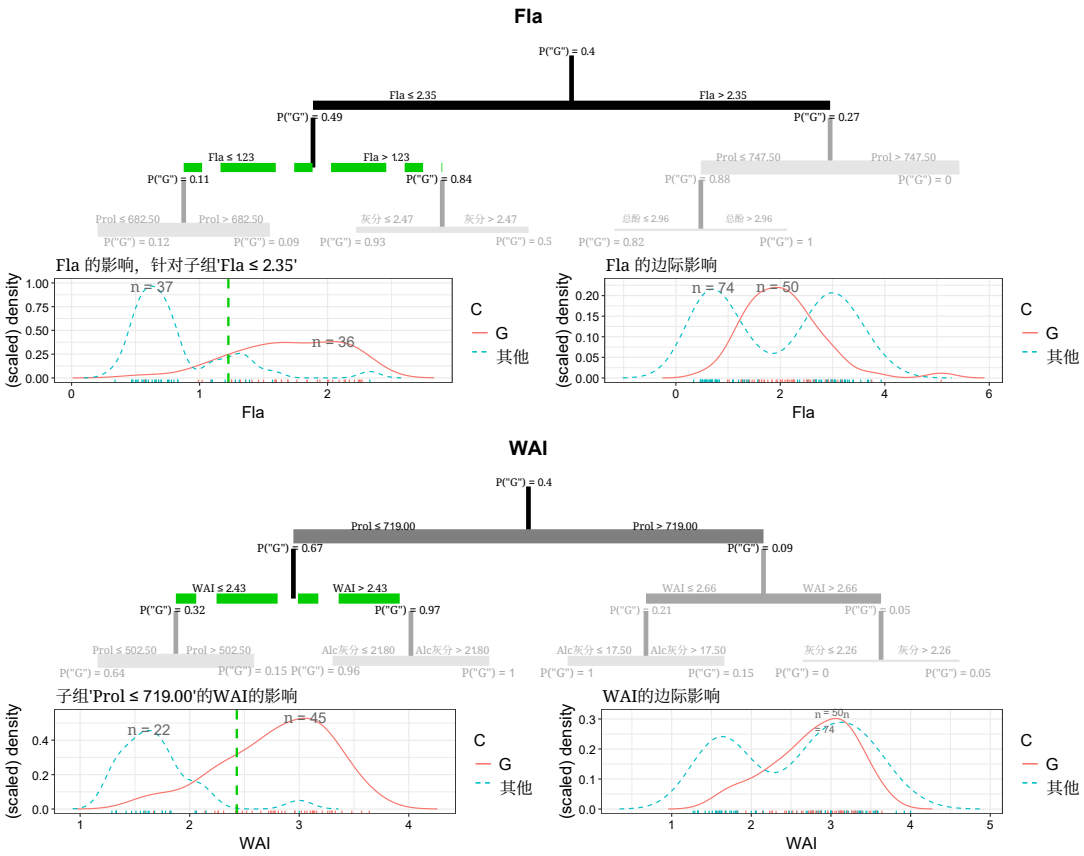

图12：葡萄酒数据集（$n = 178$）中，根据统一变量重要性度量排名第四和第五的协变量对应的 CRTR 可视化。水平线的粗细反映协变量得分，虚线表示得分最高的分裂，灰色阴影区域标示这些分裂及其祖先节点之外的树区域。对于首次分割并非得分最高分裂的 CRTR，同时展示所有袋内观测值以及得分最高节点中的观测值的核密度估计。节点处的类别频数分布以及类别特定协变量分布的加权核密度估计均使用袋内数据计算。P("G")表示相应节点中属于格里尼奥利诺的观测值所占的袋内比例

7 讨论

*CRTR*的动机与核心思想

传统参数方法通常以简化的方式（往往是线性的）对协变量效应进行建模，并且只有在明确指定交互作用时才会捕捉到这些作用。然而在实际中，协变量效应可能更为复杂。UFO框架能够筛选出条件回归树，识别协变量产生最强效应的条件——这些条件可能取决于其他协变量取值的特定范围——并据此衡量这些效应。在我们的应用中，CRTR使我们能够展示此类更复杂效应模式的存在：部分得分最高的分裂位于根节点以下，表明某些强效应是通过交互作用而非边际作用产生的。

更复杂的协变量效应：*CRTR*算法面临的挑战与可能的扩展

尽管具备这种能力，现实世界中的关联结构仍可能更加复杂。例如，单个协变量可能与多个其他协变量发生交互作用，从而形成多种不同条件，在这些条件下该协变量会产生强效应。然而，UFO框架目前仅用于识别每个协变量表现出最强效应的单一条件。即便如此，它也已能够检测出这类效应，而只关注协变量总体影响或边际影响的方法则会忽略这些效应。此外，复杂效应本身也更难解释和传达。另一个实际局限是，复杂的效应结构难以识别：检测这类结构的统计功效较低，而且识别出的模式更有可能只是观测数据中无法泛化的假象，而非真实关联。对于CRTR而言，这种风险相对较低，因为其目标仅限于识别每个协变量的最强效应。

然而，在某些探索性分析中，全面了解协变量影响可能会很有意义。一种可能的扩展方法是针对每个协变量选取若干个异质的 CRTR，以捕捉不同的效应类型。为此，可以利用相应的最佳树根集合（参见第 3.3 节），根据 Laabs 等人 [2024] 提出的方法计算这些树根之间的距离矩阵。随后可以按照 Szepannek 和 Holt [2024] 的方法对该矩阵进行层次聚类。可以使用间隙统计量 [Tibshirani 等人 提出的方法 2001], 确定最优簇数，再从每个簇中选出与该簇内所有其他树根的平均距离最小的 CRTR，作为代表性结构。

树根与最优树及其性质的关系

如第 3.1 节所述，UFO 中的树根在概念上与最优树相关。构造最优树时，会同时选择所有分裂协变量及相应的分裂点，以实现叶节点纯度最大化。因此，这类树可能容易出现过拟合，从而可能损害 CRTR 的可解释性。

然而，Bertsimas 和 Dunn [2017], 提出了一种用于构建最优分类树的算法，其分析结果与这一观点相矛盾。他们的结果表明，最优树并未表现出比分类与回归树更严重的过拟合。此外，过拟合并未随着协变量数量的增加而加剧，而且所有分析中的样本量都较小。在 Dunn 等人 [2021], 开展的一项比较研究中，基于最优树得出的变量重要性度量，将非信息性协变量误判为信息性协变量的可能性低于基于分类与回归树得出的变量重要性度量。然而，Dunn 等人 [2021] 并未说明其分析中使用的变量重要性度量的确切类型。他们的结果还表明，与分类与回归树不同，最优树不会系统性地偏向具有较多潜在分裂点的协变量，而忽视潜在分裂点较少的协变量。

利用树根进行提升的潜在用途

UFO 中用于构建树根的原则——从大量随机生成的候选树中选出最佳树——也有望应用于基于树的梯度提升。

由于梯度提升中使用的树通常较浅，因此无需像 UFO 算法那样整合传统的 CART 式分裂。树根构建原则能够捕捉交互效应而不捕捉边际效应，这一特点进一步表明，它有望提升梯度提升的预测性能。鉴于传统的基于树的梯度提升本已跻身最强大的预测方法之列，而 XGBoost 在实践中的广泛成功也印证了这一点，这将尤为值得关注 [Chen 和 Guestrin, 2016]。

其他结果类型

虽然 UFO框架针对分类变量和连续结果均已定义，但其性能评估仅针对分类结果进行。然而，连续结果版本与分类结果版本的差异仅在于划分准则和分裂得分的计算方式：前者使用方差降低，而非基尼不纯度降低。该框架的所有其他组成部分均保持不变，因此，本文分析所得结论大体上也适用于连续结果版本。

除了分类结果和连续结果外，生存结局也是一种重要的结果类型，在实践中经常遇到，尤其是在（生物）医学应用中。因此，开发适用于生存结局的 UFO框架变体将是可取的。不过，这需要定义适用于生存结局的划分准则。对于分类结果和连续结果，可以将基尼不纯度降低和方差降低从二元情形直接推广至多路情形，因为这些准则旨在最大限度地提高子节点内结果值的纯度。相比之下，生存树中最常用的准则——对数秩检验统计量——旨在最大限度地提高子节点之间结果分布的差异。在划分为两个以上节点时（这是 UFO框架中划分准则的要求），即使只有一个子节点与其他子节点存在显著差异，而其余子节点彼此非常相似，该统计量也可能变得很大。然而，适用于生存结局的划分准则应当奖励末端节点内结果值的同质性，这与用于分类结果和连续结果的准则类似。

局限性与挑战

我们的模拟研究未纳入高维数据。这一决定主要是因为，UFO框架的一个重要目标是利用条件回归树（CRTR）解释最强的边际效应或条件效应。然而，在高维情境下，维数灾难使得识别单个交互效应变得困难，并增加了出现假阳性的风险。此外，随着每棵树的树根所考虑的协变量数量增加，CRTR中识别出的分裂点可能会变得不够精确。基于这些原因，将UFO框架应用于高维数据集所获得的结果应谨慎解读。

在模拟研究中，对于所考察的较小样本量，统一变量重要性度量未能始终将信息性协变量排在非信息性协变量之前，而传统变量重要性度量在这种情况下往往表现更好。尽管无法明确确定统一变量重要性度量在小样本中表现较弱的确切原因，但这很可能与UFO框架相对较高的复杂性有关。与传统随机森林及其相关变量重要性度量相比，UFO框架涉及更多计算步骤和超参数，这可能导致结果的变异性更大。此外，在构建每棵树的树根时，同时选择所有分裂协变量和分裂点，可能会增加小样本量下的不稳定性——尽管Bertsimas 和 Dunn [2017]的模拟研究中，相关最优树并未出现这种不稳定性。不过，也不能排除统一变量重要性度量在其他数据生成过程下，对小型数据集可能表现得更可靠，而这些数据生成过程并未用于我们的模拟研究。例如，将其应用于相对较小的葡萄酒数据集（$n = 178$）时，统一变量重要性度量得出的结果与置换变量重要性度量相似，这支持了其在较小数据集上的适用性。然而，对于此类情况，仍建议将统一变量重要性度量的结果与传统变量重要性度量的结果进行比较，以避免产生误导性解释。

与传统变量重要性度量相比，统一变量重要性度量对弱效应的检测不够可靠。不过，这并不构成重大局限，因为统一变量重要性度量的主要目标是识别相关的强协变量效应——尤其是传统变量重要性度量未能捕捉到的效应——并通过条件回归树（CRTR）刻画这些效应的性质。

交互协变量的置换变量重要性度量与校正基尼重要性

模拟研究中的另一项观察结果是，与校正基尼重要性相比，置换变量重要性度量更能识别交互协变量。一个合理的解释是，校正基尼重要性仅反映针对所考察协变量进行分裂时的判别能力，而置换变量重要性度量通过置换协变量值，也会影响与之存在交互作用的其他协变量的分裂判别能力。

可以通过一个简单示例来说明这一机制。考虑一个与连续协变量 X_2 存在交互作用的二元协变量 $X_1 \in \{0,1\}$：当 $X_1 = 0$ 时，X_2 没有影响；但当 $X_1 = 1$ 时，它会产生显著影响。在这种情况下，随机森林中的一部分树很可能先按 X_1 进行分裂，然后由于交互作用，在对应于 $X_1 = 1$ 的子节点中（或后续节点中）再按 X_2 进行分裂。在计算 X_1 的置换变量重要性度量时，一些具有 $X_1 = 1$ 的观测值在置换后会被赋值为 $X_1 = 0$。在 $\mathcal{T}$ 中的树里，这不仅会削弱按 X_1 分裂的判别能力，还会消除后续按 X_2 分裂带来的益处，因为这些观测值不再参与该次分裂的评估。

一般而言，相似的机制可以解释为何涉及交互效应的协变量往往获得高于校正基尼重要性值的置换变量重要性度量值。

8 总结与结论

UFO框架是一种适用于分类变量和连续型结果变量预测问题的数据分析方法。它能够更好地检测和利用不存在边际效应的交互效应，同时提供可解释的见解，揭示协变量最强的边际效应或条件效应。

我们仔细选择了所有超参数的默认值，并使用30个数据集对关键参数的适用性进行了实证评估。研究发现，这些参数对预测性能的影响很小。

在一项涉及168个数据集的大规模比较研究中，就曲线下面积和准确率而言，UFO的表现显著优于传统随机森林。在布里尔分数方面，UFO在略少于半数的数据集上优于随机森林，但差异在统计学上并不显著。总体而言，这两种方法在预测性能上的差异很小。

在模拟研究中，统一变量重要性度量能够可靠地识别仅表现出交互作用效应的协变量，这与传统度量——置换重要性和校正基尼重要性形成对比。所有变量重要性度量都能一致地检测到其他协变量效应。这可能是由于 UFO 框架的算法复杂度较高，在小样本和弱效应情况下，统一变量重要性度量区分信息性协变量与非信息性协变量的可靠性略低于传统变量重要性度量。因此，对于极小样本量，UFO 的结果应谨慎解读，并与传统方法进行比较。

协变量代表性树根（CRTR）能够可靠地体现模拟数据中的真实协变量效应类型。此外，对真实数据集的应用表明，CRTR 能揭示在传统分析中很可能无法检测到的协变量效应。

总的来说，UFO框架是一种强大的探索性工具，可用于识别现有方法往往会忽视的重要协变量和效应结构，从而有望为领域科学中的依赖结构带来新的见解。

致谢

本研究得到德国科学基金会 [DFG-Einzelf¨orderung HO 6422/1-3 对 Roman Hornung] 的资助。

补充材料和 R 代码

本文的补充材料可在 arXiv 页面获取。用于生成正文及补充材料中所示结果的 R 代码可在 GitHub 上获取（https://github.com/RomanHornung/UnityForests_代码_和_数据，提交版本：08edff9，访问日期：2026年1月11日）。

参考文献

M. Banerjee、Y. Ding 和 A.-M. Noone。 从集成中识别代表性树。《医学统计学》，31(15):1601–1616，2012 年。doi:10.1002/sim.4492。C. B´enard、G. Biau、S. Da Veiga 和 E. Scornet。SIRUS：用于分类的稳定且可解释规则集。《电子统计学杂志》，15(1):427–505，2021年。doi:10.1214/20-EJS1792。C. B´enard、G. Biau、S. Da Veiga 和 E. Scornet。SHAFF：通过随机森林快速且一致地估计 SHApley 效应。《国际人工智能与统计会议论文集》，PMLR，第5563–5582页，2022年。S. Bernard、L. Heutte 和 S. Adam。关于随机森林中决策树的选择。载于《2009年国际神经网络联合会议论文集》，第302–307页。IEEE，2009 年。doi:10.1109/IJCNN.2009.5178693。D. Bertsimas 和 J. Dunn。最优分类树。《机器学习》，106(7):1039–1082，2017年。doi:10.1007/s10994-017-5633-9。L. Breiman。随机森林。《机器学习》，45:5–32，2001年。doi:10.1023/A:1010933404324。L. Breiman、J. H. Friedman、R. A. Olshen 和 C. J. Stone。《分类与回归树》。Chapman and Hall/CRC，佛罗里达州博卡拉顿，第1版，1984年。doi:10.1201/9781315139470。T. Chen 和 C. Guestrin。XGBoost：一种可扩展的树提升系统。载于第22届ACM SIGKDD知识发现与数据挖掘国际会议论文集，第785–794页，2016 年。doi:10.1145/2939672.2939785。D. Coppersmith、S. J. Hong 和 J. R. Hosking。决策树中名义属性的划分。《数据挖掘与知识发现》，3:197–217，1999 年。doi:10.1023/A:1009869804967。R. Couronn´e、P. Probst 和 A.-L. Boulesteix。随机森林与逻辑回归：一项大规模基准实验。《BMC生物信息学》，19:270，2018年。doi:10.1186/s12859-018-2264-5。J. Dunn、L. Mingardi 和 Y. D. Zhuo。比较特征选择中的可解释性与可说明性。arXiv预印本 arXiv:2105.05328，2021年。doi:10.48550/arXiv.2105.05328。M. Forina。Parvus。《TrAC分析化学趋势》，3(2):38–39，1984年。doi:10.1016/0165- 9936(84)87050-8。M. Forina、C. Armanino、M. Castino 和 M. Ubigli。多变量数据分析作为葡萄酒产地的鉴别方法。《Vitis》，25:189–201，1986年。doi:10.5073/vitis.1986.25.189-201。

J. H. Friedman 和 B. E. Popescu。通过规则集成进行预测学习。《应用统计年鉴》，2(3):916–954，2008年。doi:10.1214/07-AOAS148。T. Hastie、R. Tibshirani 和 J. Friedman。《统计学习基础：数据挖掘、推断与预测》。施普林格，纽约，第2版，2009 年。doi:10.1007/978-0-387-84858-7。R. Hornung 和 A.-L. Boulesteix。交互森林：识别并利用可解释的定量与定性交互效应。《计算统计与数据分析》，171:107460，2022年。doi:10.1016/j.csda.2022.107460。Z. Khan、A. Gul、A. Perperoglou、M. Miftahuddin、O. Mahmoud、W. Adler 和 B. Lausen。最优树集成、随机森林与随机投影集成分类。《数据分析与分类进展》，14:97–116，2020年。doi:10.1007/s11634-019-00364-9。B.-H. Laabs、A. Westenberger 和 I. R. K¨onig。基于一种新的基于树的距离度量识别随机森林中的代表性树。《数据分析与分类进展》，18(2): 363–380，2024年。doi:10.1007/s11634-023-00537-7。X. Li、Y. Wang、S. Basu、K. Kumbier 和 B. Yu。一种用于随机森林的去偏MDI特征重要性度量。《神经信息处理系统进展》，32:1–17，2019 年。M. Loecher。随机森林的无偏变量重要性。《统计通讯：理论与方法》，51(5):1413–1425，2020年。doi:10.1080/03610926.2020.1764042。J. D. Malley、J. Kruppa、A. Dasgupta、K. G. Malley 和 A. Ziegler。概率机：使用非参数学习机进行一致的概率估计。《医学信息学方法》，51(1):74–81，2012年。doi:10.3414/ME00-01-0052。Y. Manzali 和 M. Elfar。随机森林剪枝技术：近期综述。《运筹学论坛》，4:43，2023年。doi:10.1007/s43069-023-00223-6。N. Meinshausen。节点收获。《应用统计年鉴》，4(4):2049–2072，2010年。doi:10.1214/10-AOAS367。S. Nembrini、I. R. K¨onig 和 M. N. Wright。基尼重要性的复兴？《生物信息学》，34 (21):3711–3718，2018年。doi:10.1093/bioinformatics/bty373。P. Probst、A.-L. Boulesteix 和 B. Bischl。可调性：机器学习算法超参数的重要性。《机器学习研究杂志》，20:1–32，2019年。C. Strobl、A.-L. Boulesteix、A. Zeileis 和 T. Hothorn。随机森林变量重要性度量中的偏差：示例、来源及解决方案。《BMC生物信息学》，8:25，2007年。doi:10.1186/1471- 2105-8-25。G. Szepannek 和 B.-H. v. Holt。只见树木不见森林：分析树丛以解释随机森林。《行为计量学》，51(1):411–423，2024年。doi:10.1007/s41237-023-00205-2。R. Tibshirani、G. Walther 和 T. Hastie。利用间隙统计量估计数据集中的簇数。《英国皇家统计学会杂志B辑：统计方法学》，63(2):411–423，2001年。doi:10.1111/1467-9868.00293。M. N. Wright 和 I. R. K¨onig。随机森林中分类预测变量的分割。PeerJ，7:e6339，2019年。doi:10.7717/peerj.6339。M. N. Wright 和 A. Ziegler。ranger：一种在 C++和 R 中快速实现高维数据随机森林的方法。《统计软件杂志》，77(1):1–17，2017年。doi:10.18637/jss.v077.i01。

M. N. Wright、A. Ziegler 和 I. R. K¨onig。微小交互作用会在暗随机森林中丢失吗？《*BMC*生物信息学》，17:145，2016 年。doi：10.1186/s12859-016-0995-8。Z. Zhou 和 G. Hooker。树方法中特征重要性的无偏测量。*ACM*数据知识发现汇刊，15(2):1–21，2021年。doi：10.1145/3429445。