arXiv:2303.04874v1 [stat.ML] 8 Mar 2023

原始论文

多元结果的贝叶斯因果森林：应用于国际大规模教育评估中的爱尔兰数据。

Nathan McJames[1,2] | Andrew Parnell[1,2] | Yong Chen Goh[1,2] | Ann O'Shea[2]

[1]梅努斯大学 Hamilton研究所[2]梅努斯大学数学与统计系

通讯作者Nathan McJames, Hamilton研究所及梅努斯大学数学与统计系，爱尔兰基尔代尔郡梅努斯。电子邮箱: nathan.mcjames.2016@mumail.ie

贝叶斯因果森林（BCF）是一种因果推断机器学习模型，基于一种名为贝叶斯加性回归树（BART）的高度灵活的非参数回归与分类工具。受国际数学与科学趋势研究（TIMSS）数据的启发——该研究包含学生数学和科学学业成就的数据——我们提出了BCF算法的多元扩展。通过模拟研究，我们表明，该方法能够准确估计接受相同处理的多个结果的因果效应。我们还将模型应用于TIMSS 2019的爱尔兰数据。研究结果显示，在家中有学习桌具有积极影响（数学平均处理效应95%置信区间：[0.20，11.67]）；同时也凸显了学生在学校经常感到饥饿（数学平均处理效应95%置信区间：[-11.15，-2.78]，；科学平均处理效应95%置信区间：[-10.82，-1.72]）或经常缺勤（数学平均处理效应95%置信区间：[-12.47，-1.55]）所带来的负面影响。

关键词 - 机器学习、因果推断、贝叶斯加性回归树、贝叶斯因果森林、多变量

缩写： TIMSS：国际数学与科学趋势研究，BART：贝叶斯加性回归树，BCF：贝叶斯因果森林

1 | 引言

估计干预对感兴趣的结果变量产生的因果效应，是一项重要但困难的任务。在尝试估计时，我们面临的关键挑战是将相关性与可能存在的任何直接因果影响区分开来。理想情况下，可以通过设计实验来避免这一问题：将个体分配到治疗组的过程完全随机化（Sibbald and Roland，1998）。然而，由于经济、伦理或其他方面的考虑，这类研究往往无法开展（West et al.，2008）。因此，研究人员通常需要基于观察性数据进行分析，而这类数据无法保证个体被随机分配到治疗组。使用观察性数据时还会遇到另一项难题：可能存在混杂变量，同时影响个体被分配到治疗组的概率和感兴趣的结果。混杂变量可能使分析结果产生偏差，让人误以为两个变量之间存在直接因果关系，而事实并非如此（Greenland et al.，1999）。因此，分析观察性数据时，需要采用专门的方法来消除这种偏差。

因果推断文献中解决这一问题的常见方法包括匹配、倾向评分方法和贝叶斯因果网络（综述见 Yao 等人，2021）。匹配方法试图通过平衡处理组单位和对照组单位之间协变量的分布，重现随机对照试验中可能出现的条件（Stuart，2010）。实现这种平衡的方式多种多样，但无一例外都需要使用某种相似性度量，将具有相似特征的个体匹配到对照组和治疗组中。倾向评分方法涵盖了多种广泛使用的技术，用于估计个体接受干预的概率（Pan 和 Bai，2018）。倾向得分加权就是其中一种方法，它根据观测值的倾向得分进行加权，从而平衡对照组和治疗组（Kurz，2022）。最后，由 Pearl（1995）推广的因果图和贝叶斯因果网络，利用有向无环图表示观测变量之间的因果关系，从而识别因果效应。近年来，将先进的机器学习模型应用于这一领域引发了广泛关注（Caron 等人，2020）。我们的方法（将在下文介绍）最适合归入最后一类基于机器学习的方法。

机器学习方法这一大类包括各种各样的元学习器技术，这些技术利用机器学习模型的预测能力来估计异质性处理效应。这类方法的一个主要优势是具有很强的灵活性，许多相关技术几乎可以与任何机器学习模型结合使用。该领域最重要的贡献之一来自 Hill（2011）：他证明，通过使用足够灵活的回归模型，例如贝叶斯加性回归树（BART，Chipman 等人，2010），可以准确估计处理效应。另一项具有重大影响力的贡献来自 Hahn 等人（2020）：他们在 Hill 的工作基础上，采用 Robinson（1988）的治疗效应参数化，将Y的估计分解为预后效应μ和处理效应 τ。这种方法被称为贝叶斯因果森林（BCF），相较于 Hill 的方法具有诸多优势，因为它允许分别为模型的μ 和 τ 组件设置先验分布，并能够直接根据数据估计个体层面治疗效应。

包括贝叶斯因果森林在内的许多因果机器学习方法有一个重要局限：它们仅适用于受二元处理Z影响的单个结果变量。因此，受国际数学与科学趋势研究（Mullis 等人，2020年）的数据启发——该研究包含八年级中学生（年龄约为 14 至 15 岁）的数学和科学学业成就数据——我们提出了 BCF 的一种多元扩展，可用于估计某项干预对

多个结果的因果效应。利用这一新方法，我们考察若干与家庭相关的因素对学生学业成就的影响。具体而言，我们尝试回答以下三个研究问题：1）“在家中有学习桌，对学生的数学和科学学业成就有何影响？”；2）“经常饿着肚子到学校，会对学生的数学和科学学业成就产生什么影响？”；3）“经常缺课，对学生的数学和科学学业成就有何影响？”我们研究这些因素，是因为它们对以学生为重点的举措具有重要意义，例如免费校餐计划和返校补助；这些举措旨在帮助来自弱势背景的学生（Taras，2005年；Kennedy，2013年）。

我们采用多元方法的主要优势在于，由于模型现在能够通过其他处理变量获取额外信息，因此与因果参数相关的不确定性可能会大幅降低。我们的方法与近期一种将贝叶斯因子分析模型扩展为多元因果推断模型的方法有相似之处，该方法展示了多元方法在因果推断中的潜力（Samartsidis 等人，2020年）。我们的研究也与 Segal 和 Xiao（2011年）的工作有相似之处，他们开发了随机森林的多元扩展，并展示了多元树模型的潜力。我们的研究与这两项研究不同：首先，就 Segal 和 Xiao 的研究而言，我们关注的并非对结果变量本身进行建模，而是了解这些变量如何响应特定处理。其次，Samartsidis 等人开发的多元因果因子分析模型与我们基于 BART 的模型采用了截然不同的结构。我们认为，我们的方法具有更大的灵活性，可用于范围广泛得多的情境。

本文其余部分安排如下：第 2 节介绍国际数学与科学趋势研究（Trends in International Mathematics and Science Study），这是激发我们采用多元方法的数据集。第 3.1 节介绍贝叶斯加性回归树，该模型为贝叶斯因果森林奠定了基础。第 3.2 节说明 BCF 如何利用 BART 出色的预测能力来估计异质性处理效应，第 3.3 节则详细介绍将 BCF 扩展至多元情形所需的修改。第 4 节呈现一项模拟研究的结果，我们在其中展示了联合建模所有可用结果变量的显著优势。第 5 节将 BCF 的多元扩展应用于激发本研究的数据集 TIMSS 2019。在此，我们考察若干处理因素对学生数学和科学学业成就的影响，包括家庭学习支持、在校挨饿和缺勤。最后，我们讨论研究结果、研究局限以及未来研究的潜在方向。

2 | 国际数学与科学趋势研究

国际数学与科学趋势研究（TIMSS）是一项由国际教育成就评价协会（IEA）组织的大规模国际研究。自 1995 年以来，该研究每四年在世界各地的许多国家开展一次，TIMSS 2019 有 64 个国家参与。作为研究的一部分，中学四年级和八年级学生（通常分别约为 10 至 11 岁和 14 至 15 岁）需要完成一项简短的数学和科学评估，用于估算其总体成就水平。八年级学生还需填写一份简短的背景问卷，内容涉及家中和课堂环境，以及他们对这些学科的喜爱程度和自信程度。这些学生的教师和校长也需填写简短问卷，内容涉及他们的教育背景、教学实践以及学校获取学习资源的情况，从而为我们提供大量协变量，以便将其作为潜在混杂变量加以控制。因此，对于研究与学生数学和科学自信心及学业成就相关因素的研究人员来说，TIMSS 是极佳的信息来源

与学生自信心以及数学和科学学业成就的关系。

自 1995 年创立以来，TIMSS 数据因其规模庞大且内容全面，已成为教育领域众多研究的对象。近期利用 TIMSS 数据开展的研究包括 Tang 等人（2022），他们探讨了科学教师持续专业发展对学生科学学业成就的影响；以及 Chen（2022），他考察了课堂成就水平与个人成就水平之间的交互作用对学生数学信心的影响。然而，本文将重点关注三种特定处理因素对学生数学和科学学业成就的影响。与现有文献中通常只关注其中一种结果的许多研究不同，我们将对这两个学科的学业成就进行联合建模。

本研究关注的因果因素与学生的家庭环境有关：家庭学习支持、在学校经常感到饥饿以及缺勤。家庭学习支持和教育资源被认为是学生学业成就的重要预测因素（Tsai 和 Yang，2015年）。这一发现已在多项不同研究中得到证实；然而，鉴于家庭教育资源可能与社会经济地位相关，因此有必要在因果框架下确定这种关系的真实性质。学生早晨无法享用健康早餐所造成的负面后果也有充分记载（Vik 等人，2022）。饥饿的学生可能难以在课堂上集中注意力，也可能因无法满足每日营养需求而处于不利地位。因此，有时会提供免费学校供餐计划，以确保所有学生全天都能做好学习准备（Taras，2005年）。在学校挨饿可能与学生的社会经济背景有关，因此必须将其作为混杂因素加以考虑。

本文探讨的最后一个因素是缺勤，缺勤与较低的成就水平相关（Vesić 等人，2021）。与前文一样，这里需要谨慎。研究发现，例如，欺凌可能是缺勤率上升的重要原因（Bennour，2021）。然而，众所周知，欺凌本身会导致学业成就水平下降（Ladd等人，2017），因此这是一个混杂因素的例子。这凸显了在考察两个变量之间的关系时，考虑其他影响学业成就的重要因素的必要性。当然，学生缺勤还可能有许多其他原因，例如疾病和未经批准的缺勤，但本研究无法对这些原因作出细致区分，因为学生问卷未提供相关信息。

总之，TIMSS 是教育领域研究人员获取信息的绝佳来源。多年来，研究人员广泛利用 TIMSS数据回答了许多重要的研究问题，但如上文简要讨论所示，这些分析往往存在多层复杂因素，可能导致估计结果产生偏差。此外，现有研究大多只关注单一学科的学业成就，并采用多元线性回归模型等传统方法，而这类方法并不适合回答因果性问题。因此，我们提出，多变量因果方法能够灵活地处理可能存在的诸多混杂变量，非常适合用于分析这些数据。

3 | 异质性处理效应的贝叶斯非参数估计

因果推断机器学习文献中发展最快的研究领域之一，是运用贝叶斯非参数机器学习算法来估计异质性处理效应。然而，在详细讨论这些方法之前，我们必须先介绍一些符号约定。本文将采用内曼-鲁宾因果模型（Splawa-Neyman 等，1990；Rubin，1974；Sekhon，2008）。该模型适用于我们关注处理Z对结果Y影响的情形。内曼-鲁宾因果模型

基于潜在结果这一概念。该概念指出，对于每个观测值i，都有两个潜在结果：一个是在接受处理时会观测到的结果$y_i(Z_i = 1)$，另一个是在处于对照状态时会观测到的结果，$y_i(Z_i = 0)$。如果知道$y_i(Z_i = 0)$和$y_i(Z_i = 1)$，我们就能计算单位i的个体处理效应：$\tau_i = y_i(Z_i = 1) - y_i(Z_i = 0)$。当然，这是不可能的，因为我们只能观测到其中一个潜在结果，这就是因果推断的根本问题。

虽然我们可能无法直接观察到两种潜在结果，但可以用$\hat{}y_i(Z_i = 0)$和$\hat{}y_i(Z_i = 1)$对它们进行估计。然后，在满足正确条件的情况下，我们可以用$\hat{\tau}_i = \hat{}y_i(Z_i = 1) - \hat{}y_i(Z_i = 0)$估计$\tau_i$。根据 Kurz（2022），要使这种方法可靠，必须满足以下条件：

1. 稳定单元处理值假设（SUTVA）。该假设要求，任何个体i的潜在结果都不得受到其他任何个体j的治疗状态影响。例如，如果学生j经常缺勤，那么这不得影响任何其他学生i的成就水平。**2.** 可忽略性假设。也称为无混杂假设，该假设要求不存在我们无法控制或未作为研究的一部分收集的混杂变量：$y_i(Z_i = 1), y_i(Z_i = 0) \perp\!\!\!\perp Z_i | x_i$.**3.** 重叠假设。该假设要求任何个体$i$的倾向得分都必须远离零和一：$0 < P(Z_i = 1|x_i) < 1$。例如，如果弱势背景的学生注定永远无法拥有学习桌，这就违反了重叠假设。

上述内容为我们提供了一种非常灵活的方法来估计个体处理效应，因为可以使用任何精度足够高的模型$y_i(Z_i = 1)$和$y_i(Z_i = 0)$，即f，进行估计。出于多种原因，贝叶斯加性回归树是i的一个理想选择，接下来我们将对此进行讨论。

3.1 | 贝叶斯加性回归树

贝叶斯加性回归树（BART）是一种贝叶斯非参数机器学习算法，非常适用于各种回归和分类任务（Chipman 等人，2010）。BART 可描述为一种基于树的集成方法，它通过许多浅树的贡献来预测未知函数$f(X)$。每棵树单独来看都是弱学习器，只能解释未知函数的一小部分；但将它们组合起来后，就能够捕捉数据中变量之间非常复杂的关系和交互作用。凭借出色的预测性能，BART 已受到许多学科研究人员的青睐，并被用于医学、经济学和教育等多个领域的各种应用（Pierdzioch 等人，2016；Sparapani 等人，2016；McJames 等人，2022）。BART 是一种非常灵活的模型，使研究人员能够针对基因组学、具有局部线性的问题，当然还有因果推断等各种专门应用场景，对底层算法进行调整或修改（Sarti 等人，2022；Prado 等人，2021；Hill 等人，2020）。

给定一个长度为y的结果变量n，以及一个由n个d维变量观测值构成的协变量矩阵X，BART模型可表示如下：

$$y_i = \sum_{j=1}^{J} g(T_j, M_j, x_i) + \epsilon_i, \quad \epsilon_i \sim N(0, \sigma^2)$$

其中$g()$是一个函数 据此计算j棵J总树中每棵树的个体贡献。M_j指定了

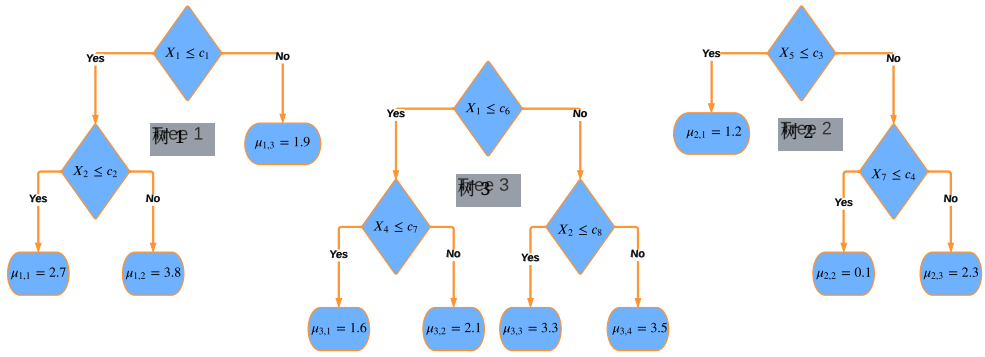


图 1 由三棵决策树构成的 BART模型集成示意图。观测值的预测值通过沿着树中从根节点到终端节点的决策规则，并将各个体贡献相加得到。例如，对于一个满足 $X_1 > c_1, X_5 < c_3, X_1 < c_6$ 且 $X_4 < c_7$,的观测值，最终预测值为 1.9 + 1.2 + 1.6 = 4.7。

与j^{th}棵树Tj相关的终端节点参数。残差 ϵ_i 假设服从均值为 0、方差为 σ^2 的正态分布。作为贝叶斯模型，需要为Tj、Mj和 σ^2 指定适当的先验分布。

BART模型采用马尔可夫链蒙特卡罗方法拟合，并结合吉布斯抽样和 Metropolis-Hastings 步骤。在每次迭代中，通过随机选择四种可能操作之一来提出一棵更新后的树，从而更新j^{th} 棵树的结构：生长、剪枝、更改或交换。如果选择生长，则为随机选定的终端节点分配一条分裂规则，使其成为两个子节点的父节点。如果选择剪枝，则随机选取一个拥有两个终端节点的父节点，并将其子节点从树中移除。在更改操作中，随机选取一个内部节点，并用一条新随机选取的分裂规则替换其原有规则。最后，交换操作会选取一对父子节点（两者均为内部节点），并交换它们各自的分裂规则。

为防止集成中的任何成员生长得过大，我们对j_{th}棵树的结构设定先验分布$P(Tj)$。该先验分布规定，深度为d的任意节点为非终端节点的概率为 $\alpha(1+d)^{-\beta}$。因此，对于一棵Tj树，其终端节点为$h_1...h_K$，非终端节点为$b_1...b_L$，有:

$$P(T_j) = \prod_{k=1}^{K} \alpha(1 + d(h_i))^{-\beta} \prod_{l=1}^{L} [1 - \alpha(1 + d(b_l))^{-\beta}]$$

其中$d()$是一个用于返回任意节点深度的函数， α 和 β 是可调节的超参数，可用于提高所偏好的树深度对应的概率。该先验分布旨在促使树保持较浅，从而限制单棵树能够解释的方差，并有助于避免过拟合。

定义好j^{th} 树的结构后，每个节点上的决策规则便构成一条路径，将观测值引导至树的叶节点。示例见图 1。现在为<style id='j'>j th 树的K 个叶节点分别指定终端节点参数$\mu_{j,k}$，这些参数对模型给出的最终预测值贡献虽小，却十分重要。为避免任何单棵树在$g()$ 中产生过大的影响，并确保μ 参数的尺度足以涵盖全部观测数据，采用先验分布 $\mu_{j,k} \sim N(0, \sigma_\mu^2)$。在数据预处理中将$y$缩放为服从标准正态分布后，超参数 σ_μ^2 的合理取值为

$1/J$，这使所有观测到的y值的范围都具有较高的先验概率。上述先验分布的组合使得 Metropolis-Hastings 步骤中使用的似然可以解析地计算为一个跨终端节点求和的多元正态分布。数据以部分残差的形式进入该多元正态分布；部分残差由响应值减去其他未更新树的预测值计算得到。

当树T_j的所有终端节点参数都已抽样后，对树T_{j+1}应用新选定的生长/剪枝/更改/交换操作，并重复此过程，直到集成中的全部J棵树都已更新其终端节点参数。每次迭代结束时，计算全部J棵树的合并贡献，并从与$1/\sigma^2 \sim Ga\left(\frac{\nu}{2}, \frac{\nu\lambda}{2}\right)$共轭的伽马后验分布中抽取精度参数$1/\sigma^2$，其中$\nu$和$\lambda$分别为先验形状参数和速率超参数。上述过程重复预先指定的迭代次数，最终得到树、终端节点参数和σ^2参数的后验分布。整个过程可概括为算法 1。

算法 1： BART 贝叶斯回代 MCMC 算法

数据： 长度为n的目标变量y（均值为 0，标准差为 1），协变量矩阵X（n行、d列）

结果： 树T的后验分布，σ的样本，$\hat{}y$的样本

初始化；

设置α、β、σ_μ、ν、λ的超参数值；

设置树的数量J、迭代次数N_{iter}；

设置初始值$\sigma^2 = 1$，并将树T_j；$j = 1$、……、M设为树桩，终端节点参数设为0；

对于从1到N_{iter}的迭代i**执行**

  对于从 1 到 J 的树 j **执行**

    计算部分残差 R_j，即y减去除树j以外所有树的预测值；

    生长一棵新树T_{new_j}　基于生长/剪枝/变更/交换；

    使用 Metropolis-Hastings 步骤，接受/拒绝新提出的树结构，依据

    $P(T_j|R_j,\sigma) \propto P(T_j)P(R_j|T_j,\sigma)$；

    使用 μ 值进行采样，服从 $P(M_j|T_j,R_j,\sigma)$；

  end

  合并所有树的预测值，得到 $\hat{}y$；

  更新　$\frac{1}{\sigma^2}$ 使用 $P(\frac{1}{\sigma^2}|\hat{}y)$；

end

3.2 | 贝叶斯因果森林

贝叶斯因果森林是一种先进的因果推断机器学习算法（Hahn 等人，2020）。BCF 以 BART 为基础来估计因果效应，并具有同样的优势，例如出色的预测表现、通过使用贝叶斯先验进行审慎的正则化，以及不确定性量化。不过，与 BART 相比，BCF 在估计异质性处理效应方面有若干优势，而这得益于采用 Robinson 参数化，该参数化将结果 y 表示为：

$$y_i = \mu(x_i,\hat{\pi}_i) + \tau(x_i)Z_i + \epsilon_i$$

其中，$\mu()$ 和 $\tau()$ 均为 BART 集成，它们协同工作，以估计 y的两个不同组成部分：预后效应 μ，可理解为处理变量 Z 取 1 表示接受处理、取 0 表示对照时的预期结果；以及处理效应 τ，可解释为接受处理对 y 的影响。模型的 $\mu()$ 部分还纳入了额外协变量 $\hat{\pi}_i$，即前文所述的倾向得分，它只是个体 i 接受处理的估计概率：$\hat{\pi}_i = P(Z_i = 1)$。在 $\mu()$ 中纳入倾向得分，对于避免一种称为正则化诱导的混杂的现象十分重要；尤其是在接受处理的可能性与控制条件下的预期结果存在某种关联时，这一做法很有用（Hahn 等人，2020）。因此，在本研究的所有实验中，我们都将纳入使用 BART模型得到的$\hat{\pi}_i$ 估计值。

上述参数化有许多重要优点。首先，它允许对μ 和 τ 部分分别施加不同程度的正则化。通常对τ 施加比对μ 更强的正则化，因为我们预计，与y 本身相比，处理效应的异质性相对简单。这种先验信念可以通过在 τ 集成中树的先验$P(T_i)$ 中采用较小的 α 和较大的 β 纳入模型。通过使用比μ更少的树来估计 τ，还可以进一步降低模型中 τ 部分的复杂性。由于我们预计，与μ相比，τ 的幅度相对较小，因此通常也会将 τ 树中终端节点参数的尺度先验 σ_τ^2设为小于 σ_μ^2。其次，如果已知X 中的变量只有一部分会调节Z 对y的影响，那么$\mu()$ 和 $\tau()$ 就可以使用不同的协变量集。最后，由于 τ现在是模型中的显式组成部分，因此可以使用 BCF 直接对处理效应进行推断，从而更直观地解释模型。

拟合 BART模型的流程只需稍作调整，即可用于 BCF。所需的修改见算法 2。

3.3 | 多变量贝叶斯因果森林

受 TIMSS 数据的启发（该数据包含学生在数学和科学两方面的学业成就信息），我们现在将贝叶斯因果森林算法扩展到多变量情境。通过这一扩展，我们可以联合估计给定干预对两个或更多结果的因果效应，从而利用各结果变量之间的相关性及其共享信息来改进预测值。在这一新设定下，BCF 模型设定如下：

$$\boldsymbol{Y}_i = \boldsymbol{\mu}_i + \boldsymbol{\tau}_i \circ \boldsymbol{Z}_i + \boldsymbol{\epsilon}_i$$

其中，$\boldsymbol{Y}_i$是一个长度为p 的向量，表示i^{th} 个观测值对应的p 维结果变量 $\boldsymbol{Y}$；$\boldsymbol{\mu}_i$和 $\boldsymbol{\tau}_i$分别表示函数 $\boldsymbol{\mu}(x_i)$和 $\boldsymbol{\tau}(x_i)$给出的第i个预测值；$\boldsymbol{\epsilon}_i$是第i^{th}个残差，$\circ$表示哈达玛积算子。请注意，在这一设定下，我们可以允许 $\boldsymbol{Z}_i$的各个分量有所不同；这些分量表示给定处理是否适用于结果变量的相应分量。这一特性十分重要，因为它使我们能够将模型应用于这样的情形：处理或干预可能作用于结果的第一个维度$\boldsymbol{Y}$（例如数学成绩），但不作用于第二个维度（例如科学成绩），反之亦然。

用于获取后验样本的马尔可夫链蒙特卡罗算法与单变量贝叶斯因果森林模型的算法有相似之处，但也存在一些重要差异。树先验分布保持不变，因为我们允许所有结果变量共用相同的树结构。动机数据集使这一设定合理，因为我们预计所选协变量

算法 2： BCF 贝叶斯回代 MCMC 算法

数据： 目标变量 y（长度 n；已标准化）、特征矩阵 X（n 行、d 列）、处理变量 Z（长度n；处理为 1；对照为 0）

结果： 树的后验列表T、 σ 的值、拟合值 $\hat{\mu}$、拟合值 $\hat{\tau}$

初始化；

设置超参数值：α_μ、 β_μ、 σ_μ、 σ_τ、 α_τ、 β_τ、 ν、 λ；

设置μ 树的数量$N\mu$、 τ 树的数量$N\tau$、迭代次数N;

设置初始值 $\sigma = 1$，并将所有μ 树和 τ 树设为树桩，终端节点参数设为0；

对于从*1*到N的迭代i，**执行**

  对于从*1*到$N\mu$ 的μ树j，**执行**

    根据y 减去除μ 树j以外所有树的预测值，计算部分残差$R_{\mu,j}$；

    生长一棵新树$T\,new_{\mu,j}$　基于生长/剪枝/变更/交换；

    使用梅特罗波利斯-黑斯廷斯步骤接受/拒绝新提出的树结构

    $P(T_{\mu,j}|R_{\mu,j},\sigma) \propto P(T_{\mu,j})P(R_{\mu,j}|T_{\mu,j},\sigma)$;

    使用 μ 值进行抽样，使用 $P(M_{\mu,j}|T_{\mu,j},R_{\mu,j},\sigma)$;

  end

  对于 τ 棵树 k 从 *1* 到 $N\tau$ **执行**

    计算部分残差$R_{\tau,k}$，即y 减去除 τ 树 k以外所有树的预测值；

    生长一棵新树$T\,new_{\tau,k}$　基于生长/剪枝/变更/交换；

    使用 Metropolis-Hastings 步骤接受/拒绝新提出的树结构，依据

    $P(T_{\tau,k}|R_{\tau,k},\sigma) \propto P(T_{\tau,k})P(R_{\tau,k}|T_{\tau,k},\sigma)$;

    使用 $P(M\tau$ 个样本值 $_{,k}|T\tau_{,}^{k},{}^{R\tau}{}_{,k},\sigma)$;

  end

  将所有树的预测值合并，得到 $\hat{y} = \hat{\mu} + Z\hat{\tau}$;

  使用 $P(\sigma|\hat{y})$ 更新 σ；

end

将以相似的方式预测这两个结果。因此，该算法会优先选择有助于准确估计 $\boldsymbol{Y}$ 所有组成部分的决策规则。这有助于避免过拟合，并起到正则化的作用，从而提升预测性能。对于终端节点参数，我们现在对 $\boldsymbol{\mu}$ 和 $\boldsymbol{\tau}$ 施加多元正态先验分布：

$$\boldsymbol{\mu}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_\mu = \sigma_\mu^2 \boldsymbol{I}\right), \quad \boldsymbol{\tau}_{j,k} \sim MVN\left(\mathbf{0}, \boldsymbol{\Sigma}_\tau = \sigma_\tau^2 \boldsymbol{I}\right)$$

最后，残差协方差矩阵的共轭先验现在是逆威沙特分布：

$$\boldsymbol{\Sigma} \sim \mathcal{W}^{-1}(\nu, \boldsymbol{\Sigma}_\mathbf{0})$$

扩展到多变量 BCF 需要对许多参数进行截然不同的更新。为简洁起见，我们将这些内容列在补充材料中，而不是正文中。

4 | 模拟研究

本节将展示新多元方法的优势及其预测性能的提升。具体而言，我们将介绍一项模拟研究的结果，该研究比较了BCF 的多元实现与单变量 BCF 模型，以及使用 Hill（2011）提出的 BART 的单变量 S-Learner 方法的表现。我们比较了模型在样本外针对以下三种不同目标估计量的表现：

1. 控制条件下的预测结果；$\hat{}y_i(Z = 0)$**2.** 观测到的处理状态下的预测结果；$\hat{}y_i$**3.** 个体处理效应估计；$\hat{\tau}_i$

我们还比较了所有受测模型的覆盖率和可信区间宽度。

我们的模拟数据来自第一个 Friedman 数据集的一个修改版本（Friedman，1991）。该数据集在机器学习文献中常用作基准数据集，因为它呈现了复杂的非线性模式和复杂的交互作用。第一个 Friedman 数据集的函数形式为：

$$f(x_i) = 10\sin(\pi x_{1,i} x_{2,i}) + 20(x_{3,i} - 0.5)^2 + 10x_{4,i} + 5x_{5,i}$$

其中，x_1，x_2，x_3，x_4，和x_5 是标准的均匀分布随机变量；此外还有另外五个随机变量x_6 到x_{10}，，它们也服从均匀分布，但不会影响y，而是作为干扰变量。

在我们的多元因果检验中，我们将创建两个结果变量 Y_1 和Y_2，二者均以$f(x)$作为μ的真实潜在模式，即预后效应；我们将其解释为控制条件下的结果。为了确定哪些观测值接受处理，我们将创建两个随机变量Z_1 和Z_2，，其中$P(Z_1 = 1) \propto \mu$，且$P(Z_2 = 1) \propto -\mu$。接下来，我们考虑两种情况：一种是处理效应同质，另一种是处理效应异质。在同质情况下，我们有

$$\begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = \begin{pmatrix} \mu(x_i, \hat{\pi}_i) \\ \mu(x_i, \hat{\pi}_i) \end{pmatrix} + \begin{pmatrix} \tau_1 \\ \tau_2 \end{pmatrix} \circ \begin{pmatrix} Z_{i,1} \\ Z_{i,2} \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}$$

其中，$\epsilon_{i,1}$ 和$\epsilon_{i,2}$ 服从均值为0、协方差矩阵为 $\sigma^2 I_2$, 的多元正态分布，而 σ^2 是随机生成的，并满足一个约束：信噪比必须介于2:1和1:1之间。两个同质处理效应 τ_1 和 τ_2 均为随机生成，其大小小于y标准差的0.3倍，因为我们先验地认为，与y 的总体方差相比，大多数处理的效应相对较小（Kraft，2020）；同时，验证多元BCF能否准确检测出效应量较小的处理也很重要。

在存在异质性的情况下，即处理效应可能受到一个或多个已观测协变量调节时，我们进行模拟

$$\begin{pmatrix} y_{i,1} \\ y_{i,2} \end{pmatrix} = \begin{pmatrix} \mu(x_i, \hat{\pi}_i) \\ \mu(x_i, \hat{\pi}_i) \end{pmatrix} + \begin{pmatrix} \frac{1 + x_{i,6} + x_{i,7}}{2} \tau_1 \\ \frac{2 + x_{i,7}}{3} \tau_2 \end{pmatrix} \circ \begin{pmatrix} Z_{i,1} \\ Z_{i,2} \end{pmatrix} + \begin{pmatrix} \epsilon_{i,1} \\ \epsilon_{i,2} \end{pmatrix}$$

	同质处理效应						异质处理效应					
	MVBCF		BCF		BART		MVBCF		BCF		BART	
指标	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2	Y_1	Y_2
μ 的均方根误差	**1.58**	**1.58**	1.79	1.81	1.69	1.69	**1.60**	**1.60**	1.80	1.80	1.71	1.70
τ 的异质性效应估计精度	**0.34**	**0.34**	0.41	0.40	0.52	0.52	**0.41**	**0.41**	0.49	0.48	0.57	0.58
y 的均方根误差	**3.98**	**3.96**	4.07	4.07	4.02	4.01	**4.01**	**4.02**	4.10	4.10	4.05	4.06
μ 的 CRPS	**0.89**	**0.89**	1.02	1.02	0.96	0.96	**0.89**	**0.89**	1.01	1.02	0.96	0.95
τ 的 CRPS	**0.24**	**0.23**	0.28	0.27	0.34	0.34	**0.27**	**0.26**	0.31	0.29	0.36	0.35
y 的 CRPS	**2.27**	**2.27**	2.33	2.32	2.30	2.29	**2.24**	**2.25**	2.30	2.30	2.27	2.27

表 1 y_1 和y_2的模拟研究结果。最佳结果以粗体标出。

与上文相同，使用 τ_1 和 τ_2，但此时效应与调节效应成正比，其中x_6 和x_7 在生成y时不起作用。

在以下结果中，我们为同质性测试生成了 1000 个合成数据集，为异质性测试也生成了 1000 个合成数据集。每个数据集在每次模拟中包含 500 个训练观测和 500 个测试观测。对于多元 BCF 和单变量 BCF，我们使用集成中的 50 棵树来估计预后效应μ，并使用 20 棵树来估计处理效应 τ。对于 BART 方法，我们总共使用 70 棵树来估计y。模型拟合的预热前和预热后阶段均运行了 500 次迭代。在每次模拟中，我们将多变量贝叶斯因果森林模型拟合到两个结果变量Y_1和Y_2,，并将所有其他方法分别拟合到结果变量Y_1 和Y_2。单变量 BCF 的实现使用了 R软件包BCF（Hahn 等人，2020），而 BART 方法使用了 R 软件包 bartCause（Dorie和Hill，2020）。两种方法均采用作者建议的默认超参数设置。

4.1 结果

图2以图形方式展示了在预测控制条件下的结果 μ_i、个体处理效应 τ_i以及观测结果 y_i时，各研究方法的比较情况。我们采用异质性效应估计精度（PEHE；等价于 τ估计中的均方根误差）来评估估计 τ_i时的预测性能：$PEHE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(\tau_i - \hat{\tau}_i)^2}$。表1中也提供了结果的数值汇总。

如图2所示，在预测μ（即控制条件下的结果）时，多元BCF的表现优于单变量BCF和BART方法。这一性能提升可归因于多元BCF能够联合考虑多个结果变量，从而使模型的预测得以受益于其对这些结果之间共享方差的认识。在均质模拟研究和异质性模拟研究的结果中，均可观察到这一改进。

在预测 τ （个体层面治疗效应）时，我们再次看到了多元方法的优势：在这项检验中，多元BCF的表现也优于单变量 BCF 和 BART 方法。这一点

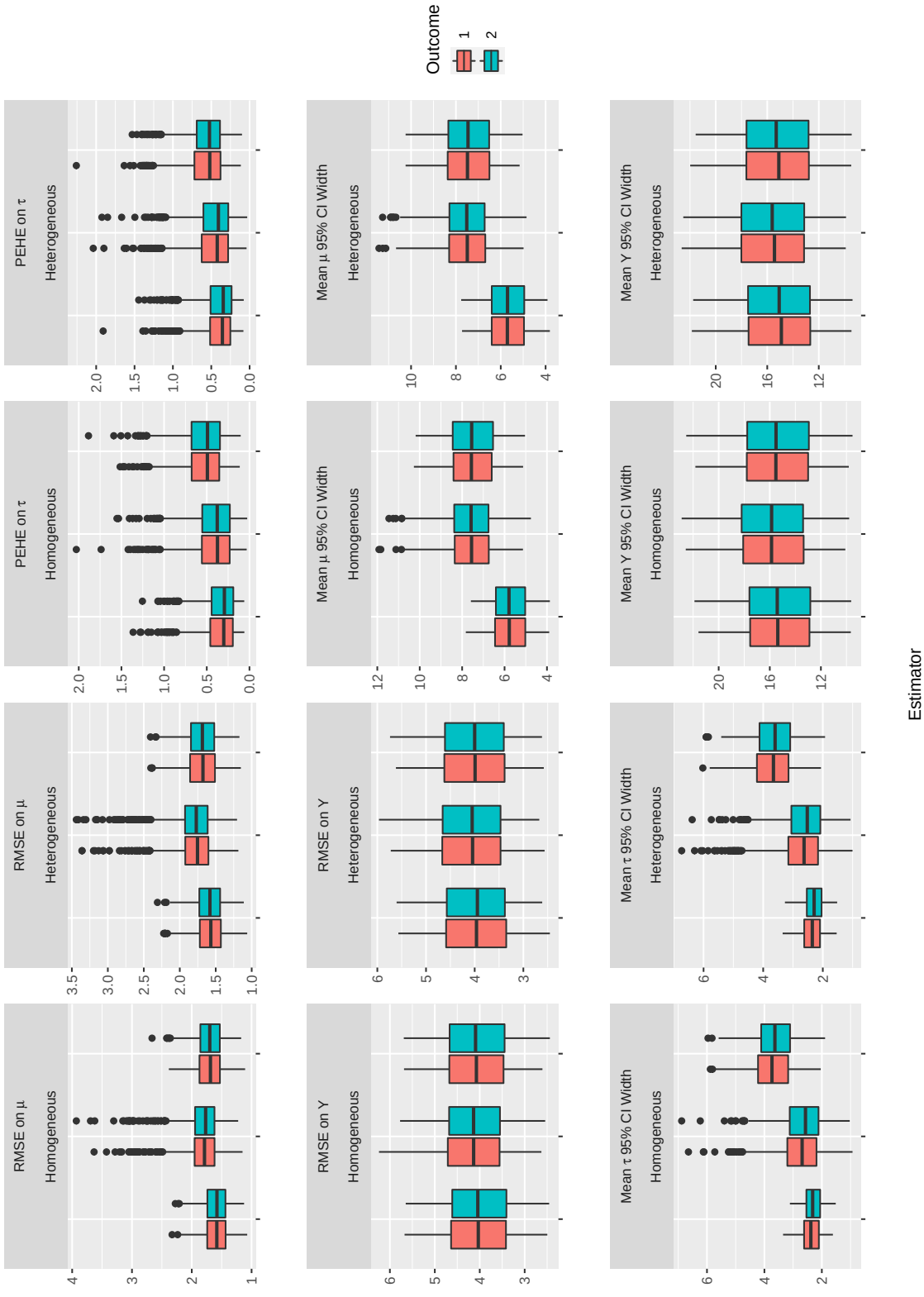

Outcome
1
2
PEHE on τ
Heterogeneous
PEHE on τ
Homogeneous
Mean μ 95% CI Width
Heterogeneous
Mean μ 95% CI Width
Homogeneous
Mean Y 95% CI Width
Heterogeneous
Mean Y 95% CI Width
Homogeneous
RMSE on μ
Heterogeneous
RMSE on μ
Homogeneous
RMSE on Y
Heterogeneous
RMSE on Y
Homogeneous
Mean τ 95% CI Width
Heterogeneous
Mean τ 95% CI Width
Homogeneous
Estimator

图 2 模拟研究结果

这是意料之中的，因为如果能够更准确地估计 μ，那么与 μ直接相关的 τ 估计值也应受益于精度的提升。值得注意的是，即使我们允许处理变量 $Z_{i,1}$ 和 $Z_{i,2}$ 有所不同，仍取得了这样的表现。这种额外的灵活性使我们的模型能够应用于这样的情形：干预可能针对结果变量 Y的一个组成部分，而非其他组成部分。例如，学生可能参加数学课的额外课程，但不参加科学课的额外课程，反之亦然。

在预测y（某个单位在其观测到的处理状态下的结果）时，三种方法的表现大致相当。对此我们颇感意外，因为这表明，在更适合标准 BART 的情形（预测结果y）下，两种 BCF 方法也具有竞争力。这充分体现了 BCF 出色的表现：它能够准确估计y的两个独立组成部分，从而提升结果的可解释性，同时不牺牲任何准确性。

当我们考察μ、τ和y的95%可信区间宽度时，可以看到，多元BCF的区间宽度比单变量 BCF和BART方法更窄。这意味着，与其他方法相比，多元BCF对其预测值更有信心，这也与该模型展现出的更高预测准确性相一致。对于三个目标估计量μ、τ、y以及两种模拟类型，均是如此。

表1还提供了每个估计量和模拟类型对应的连续排序概率评分（CRPS）（Matheson 和 Winkler，1976）指标均值。CRPS会考虑每个预测值的完整后验分布，因此能够很好地反映后验分布的校准程度。对于CRPS而言，得分越低越好；多元BCF再次优于两种单变量方法，不仅展现出更优的预测性能，还为μ、τ和y估计提供了校准良好的后验分布。

总而言之，本节的结果表明，在估计干预对两个相关结果的因果效应时，采用多元方法具有实际优势。这一点体现在多元BCF表现的提升上：在三项任务——预测y、μ和处理效应 τ——中，多元BCF的表现均优于其单变量版本。在两个模拟数据集中，结果均如此：一个数据集具有同质处理效应，另一个数据集的异质处理效应由两个协变量调节。在每项测试中，覆盖率也都令人满意。多元BCF表现出色，令我们备受鼓舞；下一节中，我们将把模型应用于教育领域的真实数据集。

5 | TIMSS 2019 的应用

本节将介绍所使用的 TIMSS 2019 数据，并应用多变量贝叶斯因果森林模型，考察三种不同处理因素对学生数学和科学学业成就的影响。

5.1 | 数据描述与程序

TIMSS 2019 是迄今举办的第七轮 TIMSS 评估。该研究的四年级和八年级部分共有 64 个国家参与，使 TIMSS 2019 成为该项目迄今规模最大的几轮评估之一（Mullis 等人，2020）。不过，本研究仅关注爱尔兰的八年级数据子集。该子集为我们提供了具有代表性的样本，共包含 4118 名八年级中学生，其中 2118 名为男性、1948 名为女性，另有 52 名未报告性别，平均年龄为 14.42 岁。此外，这些学生的数学教师和科学教师也参与了研究，为我们提供了 565 名数学教师和 409 名科学教师的数据。学生所在学校的校长也参与其中，共提供了 149 份校长问卷观测值。

将每名学生的数据与其数学教师、科学教师及校长的数据合并后，最终得到一个综合数据集，共有 4118 个观测值，每个观测值包含 50 个变量，用以描述学生、教师和学校的各种特征。重要的学生层面特征包括性别、年龄、对数学和科学的态度及学习动机、家中藏书量，以及父母双方取得的最高学历。重要的教师层面特征包括教学年限、学位学习领域、对学校重视学业成功程度的认知，以及课堂教学实践。校长数据还提供了学校学生人数、对学生平均社会经济背景的描述，以及对学校总体资源充足程度的概述等信息。在研究第 2 节所述的三个因素时，我们将这些变量作为潜在混杂因素进行控制。所有使用变量的完整列表见补充材料。

TIMSS 2019采用分层两阶段整群抽样设计，以确保收集的数据可作为该国八年级学生总体的全国代表性样本（Martin 等人，2020）。作为这一复杂调查设计的一部分，参与研究的学生会被分配一个抽样权重，用以表示他们代表总体中的多少名学生。本研究通过对个别学生的处理效应估计进行适当加权，在计算学生总体的平均处理效应时纳入了这些权重。

使用 TIMSS 数据时，另一个必须应对的复杂因素是使用可能值来衡量学生学业成就。由于数学和科学题目数量有限，TIMSS 难以准确估计学生学业成就；并且，在 TIMSS 研究期间，并非所有学生回答的都是同一套题目，这使情况更加复杂。因此，TIMSS 研究组织者没有为每位学生提供单一的成就估计值，而是从每位学生成就水平的后验分布中抽取五个可能值，以更好地体现其中存在的不确定性。本研究充分纳入了公开数据中提供的这五个可能值。每个模型均运行了五条 MCMC 链，分别对应五个可能值之一，并在预热期后将这些链合并。

为获得下一节中的结果，每条合理值链均运行了 3500 次预热期迭代和 1500 次预热期后迭代。预后成分 $\mu()$ 使用了 50 棵 BART 树，而估计处理效应 $\tau()$ 则使用了较少的 20 棵 BART 树。我们通过目视检查数据集中随机抽取的一组个体的方差协方差矩阵 Σ、预测值 $\hat{\mu}$ 和 $\hat{\tau}$ 的样本，以及平均处理效应估计本身，来评估模型是否达到令人满意的收敛。

5.2 结果

图 3 展示了我们所研究的各项处理因素的平均处理效应后验分布密度图。散点图中的每个点都对应于五条合理值 MCMC 链之一在单次迭代中所得个体处理效应估计的加权平均值。表2列出了所有这些处理因素的可信区间，并提供了所研究各项干预的对照组和处理组样本量。“有书桌”处理组的样本量为 3672，这表明样本中有 89% 的学生表示家中有书桌，其余 11% 则表示没有。经常在到校时感到饥饿或经常缺勤的对照组和处理组样本量也可作类似解读。下一节将进一步讨论这些结果的更广泛背景，但在此我们暂且只关注结果概述。

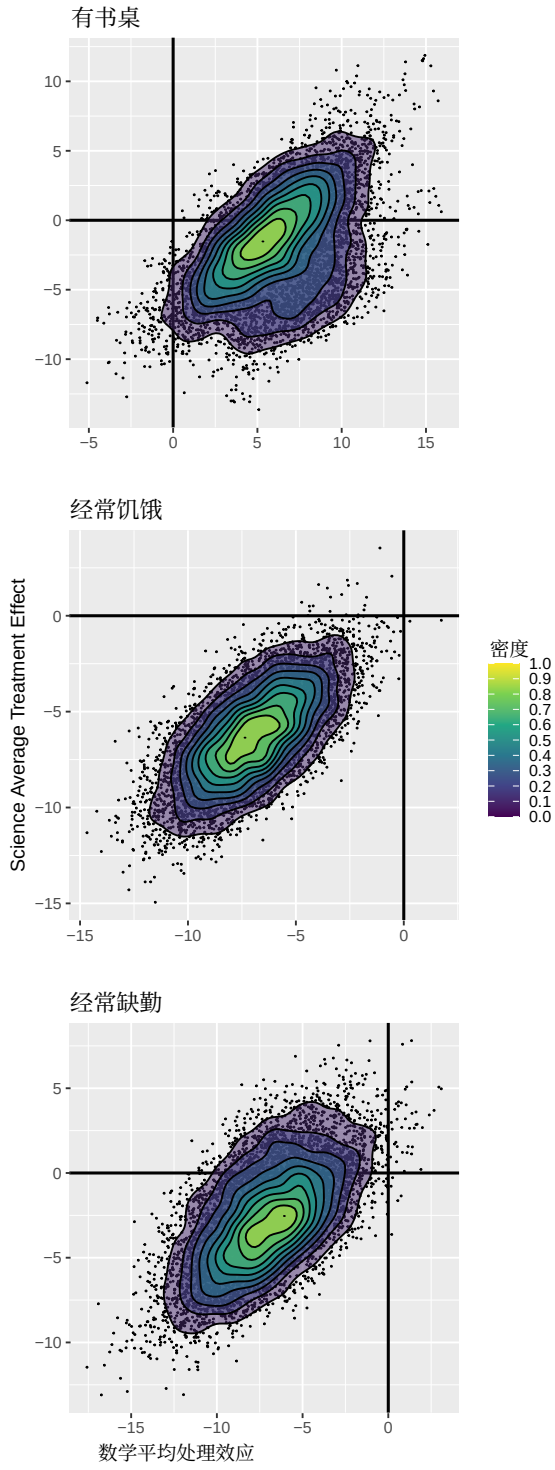


图 3 平均处理效应结果图。每个部分都显示数学（x 轴）和科学（y 轴）平均处理效应抽样后验分布的密度图。

TIMSS 结果			
	有书桌	经常在学校挨饿	经常缺勤
处理组规模	3672（89%）	954（23%）	503（12%）
对照组规模	446（11%）	3164（77%）	3615（88%）
数学平均处理效应	6.10	-6.98	-7.08
科学平均处理效应	-1.93	-6.24	-2.74
数学95%置信区间	（0.20，11.67）	（-11.15，-2.78）	(-12.47, -1.55)
科学 95%置信区间	(-8.57, 5.34)	(-10.82, -1.72)	(-8.44, 3.42)

表 2 TIMSS数据应用结果 s

为帮助解读以下结果，请注意，爱尔兰八年级学生的数学成绩近似服从正态分布，均值为524，标准差为73；处于第10和第90百分位数的学生，成绩分别约为432和614。因此，大小为7.3的处理效应相当于学生数学成绩增加或降低0.1个标准差。在教育研究中，这种效应量很常见，可视为“中等”效应（Kraft，2020）。科学成绩的分布非常相似，均值为523，标准差为83。

多元BCF清楚地表明，在家中有学习桌对学生数学成绩有正面影响。然而，有书桌对学生科学成绩的影响不太明确，平均处理效应的中心非常接近0。对于我们研究的第二个处理因素——在学校经常挨饿——结果显示，这一因素与数学和科学成绩均受到显著负面影响有关。所识别出的效应大小在数学成绩和科学成绩上几乎相同。最后，学校缺勤也被发现会对数学和科学成绩产生负面影响。不过，与有书桌的情况一样，其对科学成绩的影响不如对数学成绩的影响明确。对于所研究的三个处理因素，数学和科学成绩的平均处理效应后验分布均呈正相关。这与我们的直觉一致：这些因素中的任何一个对两门学科成绩的影响很可能相似。不过，我们确实观察到效应大小存在一些差异，其中“有书桌”这一处理因素尤为明显。

图 3 所绘制的后验分布有一个缺点，即它们仅概括了平均处理效应。然而，个体感受到的处理效应往往会受到一个或多个调节变量的影响。为考察这一可能性并探究这些变量的调节作用，我们绘制了处理效应的个体条件期望（ICE）（Goldstein 等人，2015年）和部分依赖图（PDP）（Friedman，2001年），以可视化处理效应对这些协变量的依赖关系。图 4 展示了处理“经常挨饿”的结果，其中呈现出一个有趣的趋势：资源较少的学校里的学生似乎受到的负面处理效应较小。爱尔兰（社会保护部，2023年）弱势地区资源较少的学校更有可能获得免费学校供餐计划，因此，这或许表明免费学校供餐计划成功缓解了学生经常饿着肚子到校所造成的负面后果。然而，由于我们不知道哪些学校确实参与了免费学校供餐计划，因此只能推测学校资源在此处的真正调节作用。

图 5 和图 6 显示了另一种模式：父母受教育程度较高、家中资源较多的学生，受频繁缺课的负面影响较小。这些学生可能凭借可获得的物质资源和父母支持，更有条件“补上”落下的学校作业；但同样，我们只能推测家庭教育资源和父母受教育程度对于经常缺课的真正调节作用。

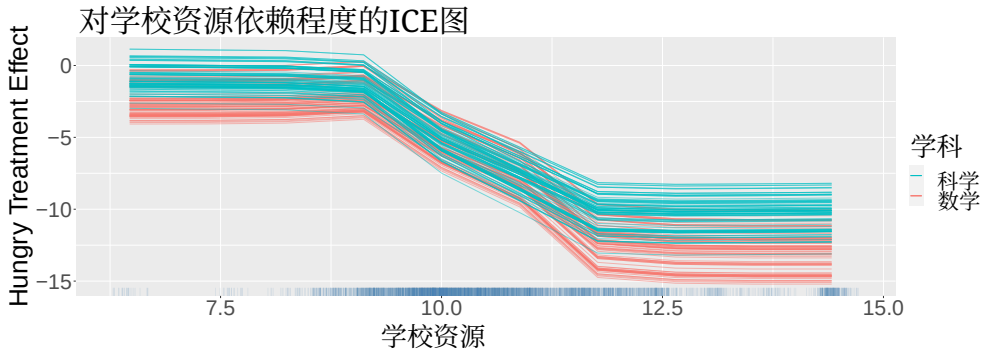


图 4 学校资源对“经常挨饿”处理效应的调节作用的 ICE 图。（为避免图形重叠，随机抽取 100 名学生。）在*x*轴上添加了抖动地毯图，以显示平均学校资源变量的分布。学校资源较少的学校中的学生，因到校时感到饥饿而受到的负面影响似乎较小。

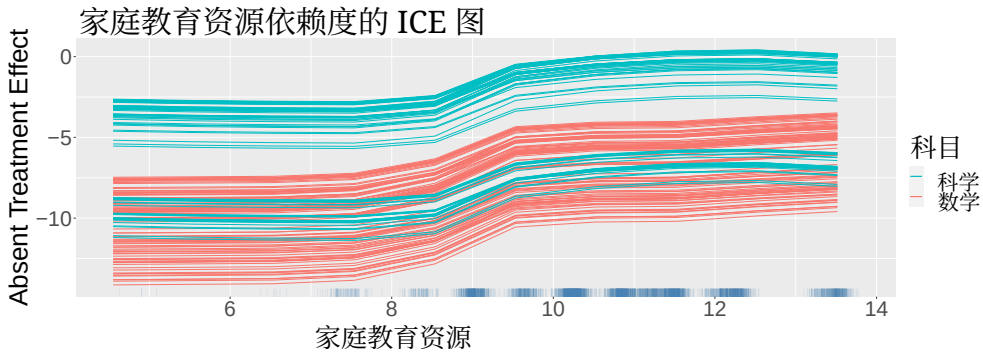


图5 家庭教育资源对“经常缺勤”处理效应调节作用的 ICE图。（随机抽取100名学生，以避免图形过度重叠。）在*x*轴上添加了抖动地毯图，以显示家庭教育资源变量的分布。家中教育资源较多的学生似乎受经常缺课的负面影响较小。请注意两簇蓝色线条，它们分别对应知道和不知道父母受教育程度的学生（见图6）。

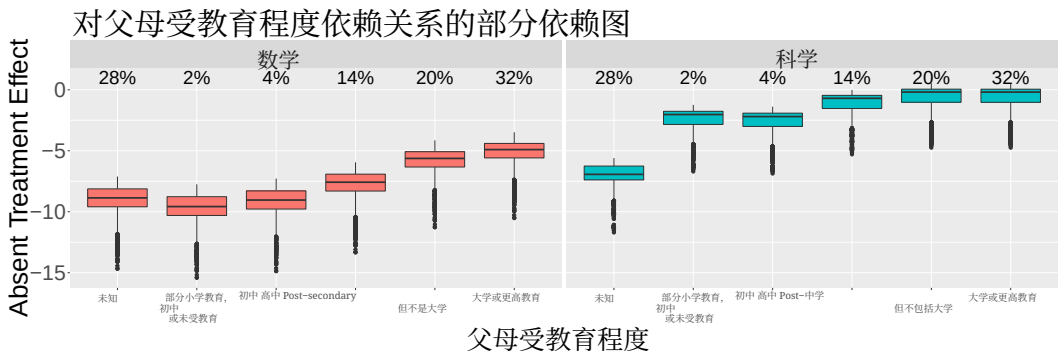


图 6 “经常缺勤”处理效应中父母受教育程度调节作用的部分依赖图（PDP）。父母受教育程度较高的学生似乎受经常缺课的负面影响较小。百分比表示属于各类别的学生比例。

6 | 讨论

本文受到国际数学与科学趋势研究数据的启发。该数据包含学生数学和科学学业成就的信息。我们开发了贝叶斯因果森林的多元扩展，可用于同时估计干预对两个或更多结果变量的因果效应。我们方法的主要优势在于两个结果变量共用同一树结构。这使我们能够利用结果之间的共享方差，从而提升预测表现；模拟研究也证实了这一点。模型的一个有趣特点是，两个结果变量的处理指示变量Z可以不同。当处理能够分别施加于各结果变量时，这一点十分重要，例如，学生可以在一门学科中收到家庭作业，而在另一门学科中没有。这一特点在我们对TIMSS数据的分析中并非必需，因为我们研究的因果因素并不因学科而异；不过，我们在模拟研究中验证了这一特点的实用性。

模拟研究的结果表明，与单变量贝叶斯因果森林模型和 BART S-学习器方法相比，多变量贝叶斯因果森林模型能够以更高的准确度估计因果效应。在预测y时，我们的多变量模型表现甚至与 BART 不相上下；由于预测y是 BART 的唯一目标，它理应在这项任务上表现出色。无论是在同质还是异质模拟数据集中，我们都观察到了这种性能提升。除了预测性能有所提高之外，我们还发现，多元BCF能够实现良好的后验校准，并给出更窄的可信区间，这表明该模型能够以更高的精度和确定性准确估计处理效应。

在将多变量贝叶斯因果森林模型应用于作为研究动机的 TIMSS 数据集时，我们发现，在家中有学习桌与数学成绩明显提高相关，但与科学成绩没有明显变化相关。意料之中的是，经常在学校感到饥饿会对数学和科学学业成就造成非常负面的影响。经常缺勤也与类似的负面影响相关。这些结果与教育领域以往研究的发现一致；那些研究指出，家中相关因素对于预测学生在数学和科学方面的学业成就十分重要（例如 Tsai 和 Yang，2015年；Vik 等人，2022年；Vesić 等人，2021年）。因此，本研究通过在因果机器学习框架下验证这些结果，作出了重要贡献。

我们的结果提供 这进一步证明了有针对性的干预（例如免费校餐计划）

可以应对学生早上经常吃不到健康早餐，或在校期间吃不到午餐所带来的负面影响。学生在学校挨饿会产生极为负面的影响，这一点清楚地表明，此类计划或许有助于避免这种情况。拥有学习桌所带来的积极影响，也可能说明有必要让家长了解学生在家中拥有专用学习空间的重要性。最后，学生经常缺勤所造成的明显负面影响，或许提示学校可以调查缺勤原因，并为受影响的学生提供额外支持。

基于多元树的模型的一个局限是，当所关注的结果变量之间相关性较弱甚至完全不相关时，这类模型可能难以应对。在某些情况下，不同结果变量可能适合采用截然不同的树结构；在这种情况下，要求两个结果变量共享同一树结构可能会有很强的限制性，进而降低整体表现。例如，Rahman等人（2017）发现，随机森林机器学习模型的多元扩展在处理高度相关的结果时表现更好。我们的模型可能也存在类似的局限。不过，这不太可能影响我们对TIMSS数据的研究，因为数学和科学学业成就之间存在很强的正相关性，$r = 0.85$，而且与学生家庭社会经济地位或学校相关的变量，很可能会对两个学科的学业成就产生相似的影响，因此采用共享的树结构非常合适。

多元 BCF 方法的潜在改进方向包括：允许 μ() 集成中的一部分 BART 树 M_1独立预测第一个结果，同时允许另一部分 M_2 预测第二个结果。更高的灵活性将使多元 BCF 更容易考虑到不同的树结构可能更适用于相关性较低的结果变量。若有理由认为处理效应对两个结果的影响可能大不相同，并且每个结果的效应可能受到不同变量集的调节，也可以对负责预测 τ 的树集成进行类似的修改。

最后，尽管本研究的动机数据集来自教育领域，但我们的多元方法也可能适用于经济学或医学等其他领域。例如，研究者可能会关注某种药物 D 对患者收缩压和舒张压的影响，这些患者已遵医嘱服用该药物。因此，未来的研究方向还包括将多元 BCF 应用于其他关注多元结果的学科。

披露声明

作者声明不存在利益冲突。

资助

本研究基于在爱尔兰科学基金会资助下开展的研究，资助编号为 18/CRT/6049。此外，Andrew Parnell 的工作还获得了以下资助：爱尔兰科学基金会职业发展奖（17/CDA/4695）和 SFI 研究中心奖（12/RC/2289_P2）。为实现开放获取，作者已将 CC BY 公共版权许可应用于本次投稿所产生的任何作者接受稿版本。

参考文献

Khaled Bennour。多变量 沙特学校中欺凌对缺勤和表现影响的水平建模。*MPRA*，2021年。

Alberto Caron、Gianluca Baio 和 Ioanna Manolopoulou。使用非参数回归模型估计个体处理效应：综述。*arXiv* 预印本 *arXiv:2009.06472*，2020年。Xin Chen。个体和班级层面的学业成就对数学态度的影响：使用 TIMSS 2019 对香港学生的分析。《教育评估研究》，72:101113，2022年。Hugh A Chipman、Edward I George 和 Robert E McCulloch。BART：贝叶斯加性回归树。《应用统计年鉴》，4(1):266–298，2010年。社会保护部。学校膳食计划。*https://www.gov.ie/en/service/29a3ff-school-meals-scheme/*，2023年。Vincent Dorie 和 Jennifer Hill。软件包“bartcause”。*URL:https://cran.r-project.org/web/packages/bartCause/bartCause.pdf*，2020年。Jerome H Friedman。多元自适应回归样条。《统计年鉴》，19(1):1–67，1991年。Jerome H Friedman。贪婪函数逼近：梯度提升机。《统计年鉴》，第1189–1232页，2001年。Alex Goldstein、Adam Kapelner、Justin Bleich 和 Emil Pitkin。窥视黑箱内部：使用个体条件期望图可视化统计学习。《计算与图形统计学杂志》，24(1):44–65，2015年。Sander Greenland、Judea Pearl 和 James M Robins。因果推断中的混杂与可折叠性。《统计科学》，14 (1):29–46，1999年。Richard P Hahn、Jared S Murray 和 Carlos M Carvalho。用于因果推断的贝叶斯回归树模型：正则化、混杂与异质性效应（附讨论）。《贝叶斯分析》，15(3):965–1056，2020年。Jennifer Hill、Antonio Linero 和 Jared Murray。贝叶斯加性回归树：综述与展望。《统计学及其应用年度评论》，7(1)，2020年。Jennifer L Hill。用于因果推断的贝叶斯非参数建模。《计算与图形统计学杂志》，20(1):217–240，2011年。Patricia Kennedy。社会政策的关键主题。Routledge，2013年。Matthew A Kraft。解读教育干预的效应量。《教育研究者》，49(4):241–253，2020年。Christoph F Kurz。增广逆概率加权与双重稳健性。《医学决策》，42 (2):156–167，2022年。Gary W Ladd、Idean Ettekal 和 Becky Kochenderfer-Ladd。从幼儿园到高中的同伴受害轨迹：对儿童学校投入和学业成就的差异化路径？《教育心理学杂志》，109 (6):826，2017年。Michael O Martin、Matthias von Davier 和 Ina VS Mullis。方法与程序：TIMSS 2019技术报告。教育成就评估国际协会，2020年。James E Matheson 和 Robert L Winkler。连续概率分布的评分规则。《管理科学》，22(10):1087–1096，1976年。Nathan McJames、Andrew Parnell 和 Ann O’Shea。影响教师工作满意度和留任的因素：使用 TALIS 2018 数据的因果推断机器学习方法。*EdArXiv*，2022年。Ina VS Mullis、Michael O Martin、Pierre Foy、Dana L Kelly 和 Bethany Fishbein。TIMSS 2019数学和科学国际结果。取自波士顿学院 *TIMSS*与*PIRLS*国际研究中心网站：*https://timssandpirls. bc. edu/timss2019/international-results*，2020年。

Wei Pan 和 Haiyan Bai。用于因果推断的倾向评分方法：综述。*Behaviormetrika*，45(2):317–334，2018。Judea Pearl。从贝叶斯网络到因果网络。载于人工智能中处理部分知识的数学模型，第157–182页。Springer，1995。Christian Pierdzioch、Marian Risse 和 Sebastian Rohloff。贵金属能否对冲汇率变动？基于贝叶斯加性回归树的实证探索。《北美经济与金融杂志》，38:27–38，2016。Estevão B Prado、Rafael A Moral 和 Andrew C Parnell。带模型树的贝叶斯加性回归树。《统计与计算》，31(3):1–13，2021。Raziur Rahman、John Otridge 和 Ranadip Pal。IntegratedMRF：用于整合不同数据类型预测结果的基于随机森林的框架。《生物信息学》，33(9):1407–1410，2017。Peter M Robinson。根号 n 一致半参数回归。《计量经济学：计量经济学会杂志》，第931–954页，1988。Donald B Rubin。估计随机研究和非随机研究中处理因素的因果效应。《教育心理学杂志》，66(5):688，1974。URLhttps://doi.org/10.1037/h0037350。Pantelis Samartsidis、Shaun R Seaman、Silvia Montagna、André Charlett、Matthew Hickman 和 Daniela De Angelis。利用多个结果的观察性时间序列数据评估干预的贝叶斯多变量因子分析模型。《英国皇家统计学会杂志：*A*辑（社会统计学）》，183(4):1437–1459，2020。Danilo A Sarti、Estevão B Prado、Alan N Inglis、Antônia AL Dos Santos、Catherine B Hurley、Rafael A Moral 和 Andrew C Parnell。用于基因型与环境互作模型的贝叶斯加性回归树。*bioRxiv*，第2021–05页，2022。Mark Segal 和 Yuanyuan Xiao。多变量随机森林。*Wiley*跨学科评论：数据挖掘与知识发现，1(1):80–87，2011。Jasjeet S Sekhon。因果推断的 Neyman-Rubin 模型及基于匹配方法的估计。《牛津政治方法学手册》，2:1–32，2008。URL https://doi.org/10.1093/oxfordhb/9780199286546.003.0011。Bonnie Sibbald 和 Martin Roland。理解对照试验。为什么随机对照试验很重要？《英国医学杂志》：英国医学杂志，316(7126):201，1998。Rodney A Sparapani、Brent R Logan、Robert E McCulloch 和 Purushottam W Laud。使用贝叶斯加性回归树（BART）进行非参数生存分析。《医学统计学》，35(16):2741–2753，2016。Jerzy Splawa-Neyman、Dorota M Dabrowska 和 TP Speed。概率论在农业试验中的应用。原理论文。第9节。《统计科学》，5(4):465–472，1990。URLhttps://doi.org/10.1214/ss/1177012031。Elizabeth A Stuart。用于因果推断的匹配方法：综述与展望。《统计科学：数学统计学会评论期刊》，25(1):1，2010。AiBin Tang、WenYe Li 和 Dawei Liu。教师在科学教学法方面的专业发展对学生成绩的影响：来自 TIMSS 2019 的证据。《波罗的海科学教育杂志》，21(2):258–274，2022。Howard Taras。营养与学生在校表现。《学校卫生杂志》，75(6):199–213，2005。Liang-Ting Tsai 和 Chih-Chien Yang。学校、课堂和学生层面因素对台湾八年级学生科学成绩的层级效应。《国际科学教育杂志》，37(8):1166–1181，2015。

Dragan Vesić、Vladimir Džinović 和 Snežana Mirkov。基于自我概念和动机预测数学成绩时缺勤的作用：塞尔维亚 TIMSS 2015。*Psihologija*，54(1):15–31，2021。

Frøydis Nordgård Vik、Trude Nilsen 和 Nina Cecilie Øverby。营养缺乏与认知结果的若干方面——对 TIMSS 中学生和教师的跨时间与学科领域三角测量。《国际教育发展杂志》，89:102553，2022年。

Stephen G West、Naihua Duan、Willo Pequegnat、Paul Gaist、Don C Des Jarlais、David Holtgrave、José Szapocznik、Martin Fishbein、Bruce Rapkin、Michael Clatts 等人。随机对照试验的替代方案。《美国公共卫生杂志》，98(8):1359–1366，2008年。

Liuyi Yao、Zhixuan Chu、Sheng Li、Yaliang Li、Jing Gao 和 Aidong Zhang。《因果推断综述》。《ACM 数据知识发现汇刊》(TKDD)，15(5):1–46，2021。

A| 多元 BCF 更新

A.1 | μ 树的对数似然

令T_j表示集成中具有部分残差R_j的$j^{th}\mu$ 树。另设树T_j有K个终端节点$h_1...h_K$，以及L 个非终端节点$b_1...b_L$。此外，令$R_{k,1}...R_{k,n_k}$ 表示落入树T_j的k^{th}号终端节点的部分残差。那么，给定残差协方差矩阵 Σ、树先验 α和 β，以及终端节点参数的先验协方差矩阵 Σ_μ，我们有：

$$\ell(T_j|R_j,\Sigma) \propto \ell(R_j|T_j,\Sigma) + \ell(T_j)$$

$$\ell(T_j) = \sum_{k=1}^{K} \log\left(1-\alpha(1+d(h_k))^{-\beta}\right) + \sum_{l=1}^{L} \log(\alpha) - \beta\log(1+d(b_l))$$

$$\begin{aligned}\ell(R_j|T_j,\Sigma) &= \sum_{k=1}^{K} \ell(R_{k,1},\ldots,R_{k,n_k}|T_j,\Sigma) \\ &\propto \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log(|\Sigma|) - \frac{1}{2}\log(|\Sigma_\mu|) + \frac{1}{2}\log(|\Sigma_{k,0}|) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{k,i}^T\Sigma^{-1}R_{k,i} - \mu_{k,0}^T\Sigma_{k,0}^{-1}\mu_{k,0}\right)\right\}\end{aligned}$$

其中

$$\Sigma_{k,0}^{-1} = n_k\Sigma^{-1} + \Sigma_\mu^{-1}$$

and

$$\mu_{k,0} = \Sigma_{k,0}\Sigma^{-1}\left(\sum_{i=1}^{n_k} R_{k,i}\right)$$

A.2 | μ 树中终端节点参数的后验分布

对于任意树 T_j 的k^{th}终端节点，其终端节点参数$\mu_{j,k}$的后验分布（先验均值为μ_0）如下：

$$\mu_k|\ldots \sim N(\mu_n,\Sigma_n)$$

其中

$$\mu_n = \left(\Sigma_\mu^{-1} + n_k\Sigma^{-1}\right)^{-1}\left(\Sigma_\mu^{-1}\mu_0 + n_k\Sigma^{-1}\overline{R}\right)$$

and

$$\Sigma_n = \left(\Sigma_\mu^{-1} + n\Sigma^{-1}\right)^{-1}$$

A.3 | 树的对数似然

给定终端节点参数的先验协方差矩阵 Σ_τ，以及 $k \times p$ 矩阵 Z_k，该矩阵记录了终端节点k中每个观测值的治疗状态，其转置后的i^{th}行记为$Z_{k,\ i}$，我们得到：

$$\ell(T_j|R_j,\Sigma) \propto \ell(R_j|T_j,\Sigma) + \ell(T_j)$$

$$\ell(T_j) = \sum_{k=1}^{K} \log\left(1-\alpha(1+d(h_k))^{-\beta}\right) + \sum_{l=1}^{L} \log(\alpha) - \beta\log(1+d(b_l))$$

$$\begin{aligned}\ell(R_j|T_j,\Sigma) &= \sum_{k=1}^{K} \ell(R_{k,1},\ldots,R_{k,n_k}|T_j,\Sigma) \\ &\propto \sum_{k=1}^{K}\left\{-\frac{n_k}{2}\log(|\Sigma|) - \frac{1}{2}\log(|\Sigma_\tau|) + \frac{1}{2}\log(|\Sigma_{k,0}|) - \frac{1}{2}\sum_{i=1}^{n_k}\left(R_{k,i}^T\Sigma^{-1}R_{k,i} - \tau_{k,0}^T\Sigma_{k,0}^{-1}\tau_{k,0}\right)\right\}\end{aligned}$$

其中

$$\Sigma_{k,0}^{-1} = Z_k^T Z_k \Sigma^{-1} + \Sigma_\tau^{-1}$$

$$\tau_{k,0} = \Sigma_{k,0}\sum_{i=1}^{n_k} Z_{k,i} \circ \Sigma^{-1} R_{k,i}$$

A.4 | 树中终端节点参数的后验分布

与μ 树中的终端节点参数类似，k^{th}叶节点中j_{th} τ树T_j的终端节点参数$\tau_{j,\ k}$在先验均值为 τ_0时的后验分布为：

$$\tau_{j,k}|\ldots \sim N(\tau_n,\Sigma_n)$$

其中

$$\tau_n = \left(\Sigma_\tau^{-1} + Z_k^T Z_k \circ \Sigma^{-1}\right)^{-1} \left(\Sigma_\tau^{-1}\tau_0 + \sum_{i=1}^{n_k} Z_{k,i} \circ [\Sigma^{-1} R_{k,i}]\right)$$

and

$$\Sigma_n = \left(\Sigma_\tau^{-1} + Z_k^T Z_k \circ \Sigma^{-1}\right)^{-1}$$

A.5 | 残差协方差参数 Σ的后验分布

给定观测值和预测值y和$\hat{y}$，对于n个残差，协方差矩阵 Σ 的后验分布在先验尺度矩阵 Σ_0 和 ν_0 个自由度下为：

$$\Sigma|\ldots \sim inv - Wish\left(\nu_0 + n, [S_0 + S_\theta]^{-1}\right)$$

其中

$$S_\theta = \sum_{i=1}^{n} (y_i - \hat{y}_i)^T \Sigma^{-1} (y_i - \hat{y}_i).$$

B | TIMSS 研究中使用的变量

使用的变量		
变量代码	来源	描述
BSDAGE	学生问卷	学生年龄
BSBG01	学生问卷	学生性别
BSBG03	学生问卷	学生在家中说英语的频率
BSBG04	学生问卷	家中的书籍数量
BSBG07	学生问卷	学生预期接受教育的程度
BSBG08A	学生问卷	父节点/监护人 A 是否出生于爱尔兰
BSBG08B	学生问卷	父节点/监护人 B 是否出生于爱尔兰
BSBG09A	学生问卷	学生是否出生于爱尔兰
BSBG10	学生问卷	学生缺勤的频率
BSBG11A	学生问卷	学生到校时感到饥饿的频率
BSBG11B	学生问卷	学生到校时感到疲倦的频率
BSDGEDUP	学生问卷	父母的最高受教育程度
BSBGHER	学生问卷	家庭教育资源数量
BSBGSSB	学生问卷	学校归属感
BSBGSB	学生问卷	校园欺凌
BSBGSCM/BSBGSCS	学生问卷	数学/科学自信心
BSBGSVM/BSBGSVS	学生问卷	学生对数学/科学的重视程度
BSBGICM/BSBGICS	学生问卷	数学/科学教学清晰度
BSBG05A	学生问卷	家中拥有电脑/平板电脑
BSBG05B	学生问卷	家中有学习桌
BSBG05C	学生问卷	有自己的卧室
BSBG05D	学生问卷	有家庭互联网连接
BSBG05E	学生问卷	拥有自己的手机
BSBG05F	学生问卷	拥有游戏设备
BSBG05G	学生问卷	家庭电视可收看“高级”电视频道
BTBG01	教师问卷	任教年限
BTBG02	教师问卷	教师性别
BTBG03	教师问卷	教师年龄
BTBG10	教师问卷	班级学生人数
BTBGTJS	教师问卷	教师工作满意度
BTBGSOS	教师问卷	安全有序的学校
BTBGLSN	教师问卷	教学受到尚未做好学习准备的学生的限制
BTBGEAS	教师问卷	重视学业成功
BTDMME	教师问卷	学位类型
BCBGDAS	校长问卷	学校纪律
BCBGEAS	校长问卷	强调学业成功
BCBGMRS/BCBGSRS	校长问卷	数学/科学资源短缺
BCDGSBC	校长问卷	学校平均社会经济背景

表3 所用控制变量