arXiv:2304.00305v1 [cs.LG] 1 Apr 2023

预测异质性：测度与应用

Jiashuo Liu* LIUJIASHUO77@GMAIL.COM
清华大学计算机科学与技术系

Jiayun Wu* JIAYUN.WU.WORK@GMAIL.COM
清华大学计算机科学与技术系

Bo Li LIBO@SEM.TSINGHUA.EDU.CN
清华大学经济管理学院

Peng Cui† CUIP@TSINGHUA.EDU.CN
清华大学计算机科学与技术系

摘要

作为大数据的一项内在且基础的属性，数据异质性存在于各种现实应用中，例如精准医疗、自动驾驶和金融应用等。对于机器学习算法而言，忽视数据异质性会严重损害其泛化性能和算法公平性，因为不同子群体之间的预测机制可能彼此不同。在本研究中，我们聚焦于影响机器学习模型预测的数据异质性，并首次提出可用预测异质性这一概念，它将模型容量和计算约束纳入考量。我们证明，可以利用有限数据对其进行可靠估计，并给出可能近似正确（PAC）界。此外，我们设计了一种双层优化算法，用于从数据中探索可用预测异质性。实证结果表明，所发现的异质性有助于深入理解收入预测、作物产量预测和图像分类任务中的子群体划分，而利用这种异质性能够提升分布外泛化性能。

关键词： 预测异质性、分布外泛化、计算约束

1 引言

大数据为人工智能（AI）系统的发展与进步提供了巨大机遇。如今，人工智能已成为一种无处不在的工具，渗透到当代技术领域的几乎方方面面，在科学发现、政策制定、医疗保健、药物发现等众多领域和行业中发挥着不可或缺的作用。然而，随着人工智能系统的广泛部署，人们日益质疑人工智能算法的可靠性、公平性和稳定性。例如，在社会学研究中（Tipton 等，2020），研究表明，即使是经过精心设计的随机试验，也存在巨大的选择偏差，使得

*。共同贡献。
†。通讯作者。

©2022年 Jiashuo Liu、Jiayun Wu、Bo Li 和 Peng Cui。许可：知识共享署名 4.0，参见 https://creativecommons.org/licenses/by/4.0/。

科学发现变得不可靠；在疾病诊断中，研究（Wynants 等，2020；Roberts 等，2021）发现，数百种现有的人工智能算法无法利用胸部X线片和CT扫描检测新冠肺炎并进行预后预测；在社会福利领域，研究发现，用于信贷申请的人工智能决策支持系统对某些人口群体存在偏见（Hardt 等，2016；Verma，2019）；在各种机器学习任务中，算法在分布偏移下普遍面临严重的泛化性能不佳问题（Shen 等，2021），等等。另一个广为人知的例子是辛普森悖论，它给社会研究带来了错误的发现（Wagner，1982；Hern´an 等，2011）。

为了缓解阻碍人工智能系统在关键、高风险应用中部署的障碍，许多研究人员转而采用成熟的以模型为中心的人工智能研究范式，致力于开发创新算法来应对这些挑战。然而，在当代机器学习讨论中，人们日益认识到，算法面临的挑战不仅源于其内在属性，也与用于训练这些模型的数据性质有关。具体而言，所用数据的异质性已成为导致这些问题的关键因素。数据异质性是指数据中存在的多样性，包括数据来源、生成机制、子群体和数据结构方面的差异。人工智能系统若未能考虑这种多样性，就可能过度关注仅存在于占主导地位的子群体或群体中的模式，进而导致错误的科学发现、不可靠且不公平的决策，以及面对新数据时泛化能力不佳。鉴于这些需要可信人工智能的高风险场景，数据异质性这一大数据固有属性应得到更多关注。此外，在大模型时代，模型开发日益接近极限，研究人员有巨大的机会探索大数据的复杂性，从而推动人工智能的发展与人工智能模型和算法的进步并行。

尽管数据异质性广泛存在，但由于其复杂性，迄今尚未形成统一表述，并且在不同领域具有不同含义。Li 和 Reynolds（1995）基于系统属性以及复杂性或变异性定义了生态学中的异质性。Rosenbaum（2005）将观察性研究中潜在结果的不确定性视为经济学中的个体异质性。近年来，在机器学习领域，若干因果学习研究（Peters 等，2016；Arjovsky 等，2019；Koyama 和 Yamaguchi，2020；Liu et al., 2021a；Creager 等，2021）和鲁棒学习研究（Sagawa 等，2019；Liu et al., 2022）利用来自多个环境的异质数据，以提升分布外泛化能力。然而，已有研究尚未给出精确定义或可靠的量化方法。在本文中，针对机器学习中的预测任务，我们从预测能力的角度出发，提出了预测异质性这一新类型的数据异质性。

从机器学习的角度来看，一个主要问题是数据异质性可能对预测准确率产生不利影响。在本研究中，我们提出了预测异质性的概念，指会影响机器学习模型性能的数据异质性。我们的目标是通过解决这一问题，促进机器学习系统的发展。为此，我们引入了预测异质性的精确定义，用于量化将整个数据分布划分为子群体后所能获得的最大额外预测信息。该度量考虑了

模型容量和计算约束，并且可以利用有限样本通过可能近似正确（PAC）界进行准确估计。我们对该度量的性质进行了理论分析，并在典型的数据异质性场景下对其进行了考察。此外，我们提出了信息最大化（IM）算法，以实证探索数据中的预测异质性。通过实证研究，我们发现，所探索的异质性具有可解释性，并为农业、社会学、物体识别和医疗保健等多个领域的子群体划分提供了有价值的见解。此外，所识别出的子群体可用于确定与新冠肺炎死亡率相关的特征，并提升机器学习模型的分布外泛化性能。这一点已通过模拟数据和真实世界数据上的实验得到验证。总之，我们的研究通过提供预测异质性的精确定义及其估计的可靠度量，为机器学习系统的发展作出了贡献。研究结果表明，IM 算法具有探索预测异质性的潜力，能够助力科学发现，并改善机器学习模型在真实世界应用中的泛化性能。

2 互信息与预测 $\mathcal{V}$-信息的预备知识

本节简要介绍互信息和预测 $\mathcal{V}$-信息（Xu 等人，2020），它们是我们提出的预测异质性的预备知识。

记号。对于概率三元组 $(\mathbb{S}, \mathcal{F}, \mathbb{P})$，定义随机变量 $X\colon \mathbb{S} \to \mathcal{X}$和$Y\colon \mathbb{S} \to \mathcal{Y}$，其中 $\mathcal{X}$ 为协变量空间，$\mathcal{Y}$ 为目标空间。因此，$x \in \mathcal{X}$表示协变量，$y \in \mathcal{Y}$ 表示目标变量。将随机分类变量集合记为 $\mathcal{C} = \{C\colon \mathbb{S} \to \mathbb{N} | \text{supp}(C) \text{ 有限}\}$。此外，$\mathcal{P}(\mathcal{X})$ $\mathcal{P}(\mathcal{Y})$分别表示空间 $\mathcal{X}, \mathcal{Y}$上博雷尔代数的所有概率测度集合。$H(\cdot)$表示离散随机变量的香农熵以及连续随机变量的微分熵，$H(\cdot|\cdot)$表示两个随机变量的条件熵。

在信息论中，两个随机变量 X、Y 的互信息衡量这两个变量之间的依赖关系，它量化了在观测其中一个变量后另一个变量熵的减少：

$$\mathbb{I}(X;Y) = H(Y) - H(Y|X). \tag{1}$$

众所周知，互信息与 Y 的可预测性相关（Cover Thomas 和 Thomas Joy，1991）。互信息的标准定义不切实际地假设预测器具有无限的计算能力，因此尤其是在高维情况下难以估计。为缓解这一问题，Xu 等人（2020）在现实的计算约束下提出了预测 $\mathcal{V}$-信息，其中预测器只能使用预测族 $\mathcal{V}$ 中的模型来预测目标变量 Y。

定义 1（预测族（Xu 等人，2020年）） 设 $\Omega = \{f\colon \mathcal{X} \cup \{\emptyset\} \to \mathcal{P}(\mathcal{Y})\}$。若 $\mathcal{V} \subseteq \Omega$ 满足以下条件，则称其为预测族：

$$\forall f \in \mathcal{V},\ \ \forall P \in \text{range}(f),\ \ \exists f' \in \mathcal{V},\quad s.t.\ \forall x \in \mathcal{X}, f'[x] = P, f'[\emptyset] = P. \tag{2}$$

预测族包含所有允许使用的预测模型，由此形成计算或统计约束。公式 2 中的附加条件意味着，预测器可以随时选择忽略输入协变量（x），而只使用 $\emptyset$。

定义 2（预测 $\mathcal{V}$信息（Xu 等人，2020））设 X、Y 是取值于 $\mathcal{X}\times\mathcal{Y}$ 的两个随机变量，且 $\mathcal{V}$ 是一个预测族。从 X 到 Y 的预测 $\mathcal{V}$-信息定义如下：

$$\mathbb{I}_{\mathcal{V}}(X\to Y)=H_{\mathcal{V}}(Y|\emptyset)-H_{\mathcal{V}}(Y|X), \tag{3}$$

其中，$H_{\mathcal{V}}(Y|\emptyset)$、$H_{\mathcal{V}}(Y|X)$ 是定义如下的预测条件 $\mathcal{V}$-熵：

$$H_{\mathcal{V}}(Y|X)=\inf_{f\in\mathcal{V}}\mathbb{E}_{x,y\sim X,Y}[-\log f[x](y)]. \tag{4}$$

$$H_{\mathcal{V}}(Y|\emptyset)=\inf_{f\in\mathcal{V}}\mathbb{E}_{y\sim Y}[-\log f[\emptyset](y)]. \tag{5}$$

值得注意的是，$f\in\mathcal{V}$ 是一个映射：$\mathcal{X}\cup\{\emptyset\}\to\mathcal{P}(\mathcal{Y})$，因此 $f[x]\in\mathcal{P}(\mathcal{Y})$ 是定义在 $\mathcal{Y}$ 上的概率测度，而 $f[x](y)\in\mathbb{R}$ 是在 $y\in\mathcal{Y}$ 上取值的密度。$H_{\mathcal{V}}(Y|\emptyset)$ 也记作 $H_{\mathcal{V}}(Y)$。

与互信息相比，预测 $\mathcal{V}$-信息对计算能力有所限制，因此在高维情形下更容易估计。当预测族 $\mathcal{V}$ 包含所有可能的模型，即 $\mathcal{V}=\Omega$ 时，已证明 $\mathbb{I}_{\mathcal{V}}(X\to Y)=\mathbb{I}(X;Y)$（Xu 等人，2020）。

3 预测异质性

本文从机器学习的视角出发，对影响决策的 数据异质性进行量化，并将其命名为预测异质性。它易于与机器学习算法相结合，有助于分析大数据并构建更合理的算法。

3.1 交互异质性

为正式定义预测异质性，我们先对交互异质性进行表述。交互异质性定义如下：

定义 3（交互异质性）设 X、Y 为取值属于 $\mathcal{X}\times\mathcal{Y}$的随机变量。将随机分类变量的集合记为 $\mathcal{C}$，并取其子集 $\mathscr{E}\subseteq\mathcal{C}$。那么，$\mathscr{E}$ 是一个环境集合，当且仅当存在 $\mathcal{E}\in\mathscr{E}$ 使得 X、$Y\perp\!\!\!\perp\mathcal{E}$。$\mathcal{E}\in\mathscr{E}$ 称为环境变量。关于环境集合 $\mathscr{E}$，X 与 Y 之间的交互异质性定义如下：

$$\mathcal{H}^{\mathscr{E}}(X,Y)=\sup_{\mathcal{E}\in\mathscr{E}}\mathbb{I}(Y;X|\mathcal{E})-\mathbb{I}(Y;X). \tag{6}$$

每个环境变量 $\mathcal{E}$ 都表示对 $\mathcal{X}\times\mathcal{Y}$的一种随机“划分”，而环境集合的条件意味着，存在这样一种随机划分，使得 X、Y 的联合分布在每个环境中都得以保持。在信息论中，$\mathbb{I}(Y;X|\mathcal{E})-\mathbb{I}(Y;X)$被称为交互信息，它衡量环境变量 $\mathcal{E}$ 对目标变量 Y 与

协变量 X之间共享的信息量所产生的影响。而公式6中定义的交互异质性 量化了通过纳入或揭示环境变量 $\mathcal{E}$所能获得的最大 额外信息。直观地说，较大的 $\mathcal{H}^{\mathscr{E}}(X、Y)$表明，从 X 到 Y 的预测能力因 $\mathcal{E}$而增强，这意味着揭示与环境划分 $\mathcal{E}$ 相关联的潜在子群体将有助于 $X \to Y$ 预测。

3.2 预测异质性

基于互信息计算交互异质性相当困难，因为标准互信息众所周知难以估计，尤其是在大数据场景下。此外，即使能够准确估计互信息，预测模型也未必能充分利用它。

受 Xu 等人（2020）启发，我们提出预测异质性，用于衡量在计算约束下可捕获的交互异质性，以及这种异质性对指定预测族内模型预测的影响。首先，我们提出条件预测 $\mathcal{V}$-信息，它是预测 $\mathcal{V}$-信息的推广。

定义 4（条件预测 $\mathcal{V}$-信息）设 X、Y 是分别取值于 $\mathcal{X} \times \mathcal{Y}$ 和 $\mathcal{E}$ 的两个随机变量，其中 $\mathcal{E}$ 是一个环境变量。对于预测族 $\mathcal{V}$，条件预测 $\mathcal{V}$-信息定义为：

$$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) = H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) - H_{\mathcal{V}}(Y|X, \mathcal{E}), \tag{7}$$

其中， $H_{\mathcal{V}}(Y|\emptyset、\mathcal{E})$ 和 $H_{\mathcal{V}}(Y|X、\mathcal{E})$ 定义为：

$$H_{\mathcal{V}}(Y|X, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{x,y\sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right]. \tag{8}$$

$$H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}\left[\inf_{f\in\mathcal{V}} \mathbb{E}_{y\sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right]. \tag{9}$$

直观而言，条件预测 $\mathcal{V}$-信息度量了各环境中预测 $\mathcal{V}$-信息的加权平均。现在，我们可以正式定义预测异质性度量。

定义 5（预测异质性）设 X, Y 为取值于 $\mathcal{X}\times\mathcal{Y}$ 的随机变量， $\mathscr{E}$ 为环境集合。对于预测族 $\mathcal{V}$，关于 $\mathscr{E}$ 的预测 $X \to Y$ 的预测异质性定义如下：

$$\mathcal{H}^{\mathscr{E}}_{\mathcal{V}}(X \to Y) = \sup_{\mathcal{E}\in\mathscr{E}} \mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(X \to Y), \tag{10}$$

其中， $\mathbb{I}_{\mathcal{V}}(X \to Y)$ 是根据定义 *2* 得出的预测 $\mathcal{V}$-信息。

利用预测 $\mathcal{V}$-信息，方程 10 中定义的预测异质性刻画了预测模型在涉及环境变量 $\mathcal{E}$ 时可以利用的最大额外信息。它限制了 $\mathcal{V}$ 中的预测模型，并且所探索的额外信息可能有助于提升模型的预测性能 $f \in \mathcal{V}$，因此被命名为预测异质性。下面，我们介绍交互异质性和预测异质性的一些基本性质。

命题 6（预测异质性的基本性质）设 X、Y 为取值于 $\mathcal{X}\times\mathcal{Y}$ 的随机变量，$\mathcal{V}$ 为函数族，$\mathscr{E}$、$\mathscr{E}_1$、$\mathscr{E}_2$ 为环境集合。

1. 单调性：若 $\mathscr{E}_1 \subseteq \mathscr{E}_2$, $\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_1}(X\to Y) \le \mathcal{H}_{\mathcal{V}}^{\mathscr{E}_2}(X\to Y)$。*2.* 非负性：$\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X\to Y)$ *{v24>*。*3.* 有界性：对于离散的 Y, $\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X\to Y) \le H_{\mathcal{V}}(Y|X)$。*4.* 边界情况：如果预测族 $\mathcal{V}$ 是包含所有可能模型的最大预测族，即 $\mathcal{V}=\Omega$，则有 $\mathcal{H}^{\mathscr{E}}(X,Y)=\mathcal{H}_{\Omega}^{\mathscr{E}}(X\to Y)$。

证明见附录 A。关于预测异质性的更多理论性质，我们将在第 3.3 节中推导其在内生性条件下的显式形式；内生性是数据异质性的一种常见体现。我们还将在第 3.4 节中证明，如果 $\mathcal{V}$ 的复杂度有界（例如其拉德马赫复杂度），则可以在保证条件下对我们提出的预测异质性进行实证估计。

3.3 线性情形下的理论性质

本节对多种线性设置中的预测异质性进行理论分析。具体而言，我们考虑两种情形：(1) 带独立噪声的齐性情形；(2) 由选择偏差和隐藏变量导致内生性的异质情形。通过考察这些典型设置，我们近似求得所提出度量的解析形式，并得出可推广至更复杂情形的深刻结论。

首先，在不存在数据异质性的齐性情形下，定理 7 证明，我们的度量受标签噪声的尺度（通常较小）所限，并且在温和假设下，在线性情形中会降至 0。这表明预测异质性对独立噪声不敏感。值得注意的是，在线性情形中，我们只处理满足 $X\perp\epsilon|\mathcal{E}$ 的环境变量，因为在常见的预测任务中，独立噪声是未知的，也不可能将其用于预测。

定理 *7*（带独立噪声的齐性情形）对于预测任务 $X\to Y$，其中 X、Y 是取值于 $\mathbb{R}^n\times\mathbb{R}$ 的随机变量，考虑数据生成过程 $Y=g(X)+\epsilon, \epsilon\sim\mathcal{N}(0,\sigma^2)$，其中 $g:\mathbb{R}^n\to\mathbb{R}$ 是可测函数。*1)* 对于满足 $g\in\mathcal{G}$ 的函数类 $\mathcal{G}$，将函数族定义为 $\mathcal{V}_{\mathcal{G}}=\{f|f[x]=\mathcal{N}(\phi(x),\sigma_V^2)\ \phi\in\mathcal{G},\ \sigma_V\in\mathbb{R}^+\}$。对于环境集合 $\mathscr{E}$，有 $\mathcal{H}_{\mathcal{V}_{\mathcal{G}}}^{\mathscr{E}}(X\to Y)\le\pi\sigma^2$。*2)* 取 $n=1$，且 $g(x)=\beta x$, $\beta\in\mathbb{R}$。不失一般性，假设 $\mathbb{E}[X]=0$ 和 $\mathbb{E}[X^2]$ 存在。给定函数族 $\mathcal{V}_\sigma=\{f|f[x]=\mathcal{N}(\theta x,\sigma^2)\ \theta\in\mathbb{R},\sigma$，固定 $\}$，且环境集合 $\mathscr{E}=\{\mathcal{E}|\mathcal{E}\in\mathcal{C},|supp(\mathcal{E})|=2, X\perp\epsilon|\mathcal{E}\}$。有 $\mathcal{H}_{\mathcal{V}_\sigma}^{\mathscr{E}}(X\to Y)=0$。证明见附录 *B*。

其次，我们考察所提出的度量在数据异质性的两种典型情形（Fan 等，2014）下的表现，分别称为选择偏差导致的内生性（Heckman，1979；Winship 和 Mare，1992；Cui 和 Athey，2022年）和隐藏变量下的内生性（Fan 等，2014；Arjovsky 等人，2019年）。

首先，在定理 8 中，我们考虑预测任务 $X\to Y$，其中 X 和 Y 取值于 $\mathbb{R}^2\times\mathbb{R}$。令 $X=[S,V]^T$。预测族定义如下：

$$\mathcal{V}=\{f|f[x]=\mathcal{N}(\theta_S S+\theta_V V,\sigma^2),\quad \theta_S,\theta_V\in\mathbb{R},\sigma=1\}. \tag{11}$$

此外，数据分布 $P(X, Y)$ 是潜在子群体的混合分布，可以通过环境变量 $\mathcal{E}^* \in \mathcal{C}$ $P(X, Y) = \sum_{e \in \mathcal{E}^*} P(\mathcal{E}^* = e) P(X, Y|\mathcal{E}^* = e)$ 来表示，使得 $\mathcal{E}^* = \text{supp}(P(X, Y))$。对于每个 $e \in \text{supp}(\mathcal{E}^*)$，$P(X, Y|\mathcal{E}^* = e)$ 是一个同质子群体的分布。注意，预测任务是使用协变量 X 预测 Y，而子群体结构是潜在的。也就是说，对于模型而言，$P(\mathcal{E}^*|X, Y)$ 是 未知 的。下面，我们将在一个典型情形下推导该度量的解析形式。

定理 8（选择偏差下的内生性）对于预测任务 $X = [S, V]^T \to Y$，当存在潜在环境变量 $\mathcal{E}^*$ 时，带有选择偏差的数据生成过程定义如下：

$$Y = \beta S + f(S) + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2); \quad V = r(\mathcal{E}^*) f(S) + \sigma(\mathcal{E}^*) \cdot \epsilon_V, \epsilon_V \sim \mathcal{N}(0, 1), \tag{12}$$

其中 $f: \mathbb{R} \to \mathbb{R}$ 和 $r, \sigma: \mathit{supp}(\mathcal{E}^*) \to \mathbb{R}$ 是可测函数。$\beta \in \mathbb{R}$。假设 $\mathbb{E}[S^2]$ 有限，$\mathbb{E}[f(S)S] = 0$，且存在 $L > 1$ 使得 $L\sigma^2(\mathcal{E}^*) < r^2(\mathcal{E}^*)\mathbb{E}[f^2]$。对于方程 *11* 中定义的预测族和环境集合 $\mathscr{E} = \mathcal{C}$，预测任务 $[S, V]^T \to Y$ 的预测异质性近似为：

$$\mathcal{H}_{\mathcal{V}}^{\mathcal{C}}(X \to Y) \approx \frac{Var(r_e)\mathbb{E}[f^2] + \mathbb{E}[\sigma^2(\mathcal{E}^*)]}{\mathbb{E}[r_e^2]\mathbb{E}[f^2] + \mathbb{E}[\sigma^2(\mathcal{E}^*)]} \mathbb{E}[f^2(S)], \textit{error bounded by } \frac{1}{2}\max(\sigma_Y^2, R(r, \sigma, f)). \tag{13}$$

进一步，我们有

$$\begin{aligned} R(r(\mathcal{E}^*), \sigma(\mathcal{E}^*), f) &= \mathbb{E}[(\frac{1}{\frac{r^2\mathbb{E}[f^2]}{\sigma^2} + 1})^2]\mathbb{E}[f^2] + \mathbb{E}_{\mathcal{E}^*}[(\frac{1}{\frac{r}{\sigma} + \frac{\sigma}{r\mathbb{E}[f^2]}})^2] \\ &< \mathbb{E}[f^2](\frac{1}{(L+1)^2} + \frac{1}{L + 2 + \frac{1}{L}}). \end{aligned} \tag{14}$$

证明见附录 *C*。

直观而言，定理 8 中的数据生成过程在伪特征 V 与目标变量 Y之间引入了伪相关，该相关在不同子群体间有所变化（即 $r(\mathcal{E}^*)$ 和 $\sigma(\mathcal{E}^*)$ 会变化），从而导致数据异质性。这里，$\mathbb{E}[f(S)S] = 0$表示模型设定错误，因为存在一个非线性项 $f(S)$，线性预测族无法利用稳定特征 S推断该项。常数L 刻画了 V与 Y之间伪相关的强度。L 越大，V 就能为预测提供越多信息。

从公式 13 的近似结果可以看出，我们提出的预测异质性主要由两个项决定：(1) $\text{Var}[r(\mathcal{E}^*)]/\mathbb{E}[r^2(\mathcal{E}^*)]$ 刻画了$r(\mathcal{E}^*)$在子群体之间的方差；(2) $\mathbb{E}[f^2(S)]$ 反映了模型设定错误的强度。这两个组成部分对应了选择偏差下数据异质性的两个来源，验证了我们所提出测度的合理性。根据该定理可以推断，$r(\mathcal{E}^*)$在子群体之间的变异性越大，且模型设定错误越严重，预测异质性的程度就越高。换言之，当各子群体之间存在显著差异且模型设定不准确时，预测异质性往往更大。

此外，在定理 9 中，我们分析了隐藏变量下存在内生性时的度量。在定理 9 中，反因果协变量 V 由因果图生成

如 $Y \to V \leftarrow \mathcal{E}^*$ 所示，其中包含一个隐藏环境变量 $\mathcal{E}^*$。然而，由于 $\mathcal{E}^*$ 未被纳入预测模型， V与 Y之间的关系会产生偏差，从而削弱预测能力。

定理 9（隐藏变量下的内生性）对于预测任务 $[S,V]^T \to Y$，在存在潜在环境变量 $\mathcal{E}^*$ 时，含隐藏变量的数据生成过程定义如下：

$$Y = \beta S + f(S) + \epsilon_Y, \epsilon_Y \sim \mathcal{N}(0, \sigma_Y^2); \quad V = r(\mathcal{E}^*)(f(S) + \epsilon_Y) + \sigma(\mathcal{E}^*)\epsilon_V, \epsilon_V \sim \mathcal{N}(0,1), \tag{15}$$

其中 f： $\mathbb{R} \to \mathbb{R}$和 r, σ： $supp(\mathcal{E}^*) \to \mathbb{R}$是可测函数。 $\beta \in \mathbb{R}$。假设 $\mathbb{E}[f(S)S] = 0$，且存在 $L > 1$，使得 $L\sigma^2(\mathcal{E}^*) < r^2(\mathcal{E}^*)(\mathbb{E}[f^2] + \sigma_Y^2)$。对于方程 *11* 中定义的预测族和环境集合 $\mathscr{E} = \mathcal{C}$，预测任务 $[S,V]^T \to Y$的预测异质性可近似为：

$$\begin{aligned} &\mathcal{H}_{\mathcal{V}}^{\mathcal{C}}(X \to Y) \approx \frac{Var(r_e)(\mathbb{E}[f^2] + \sigma_Y^2) + \mathbb{E}[\sigma^2(\mathcal{E}^*)]}{\mathbb{E}[r_e^2](\mathbb{E}[f^2] + \sigma_Y^2) + \mathbb{E}[\sigma^2(\mathcal{E}^*)]}(\mathbb{E}[f^2(S)] + \sigma_Y^2), \\ &\textit{error bounded by } \frac{1}{2}\max(\sigma_Y^2, R(r,\sigma,f)). \end{aligned} \tag{16}$$

此外，我们还可以得到：

$$\begin{aligned} R(r(\mathcal{E}^*), \sigma(\mathcal{E}^*), f) &= \mathbb{E}[(\frac{1}{\frac{r^2(\mathbb{E}[f^2]+\sigma_Y^2)}{\sigma^2} + 1})^2](\mathbb{E}[f^2] + \sigma_Y^2) + \mathbb{E}_{\mathcal{E}^*}[(\frac{1}{\frac{r}{\sigma} + \frac{\sigma}{r(\mathbb{E}[f^2]+\sigma_Y^2)}})^2] \\ &< (\mathbb{E}[f^2] + \sigma_Y^2)(\frac{1}{(L+1)^2} + \frac{1}{L + 2 + \frac{1}{L}}). \end{aligned} \tag{17}$$

证明见附录 *C*。

直观而言，定理 9 中的数据生成过程引入了伪特征 V与目标变量 Y之间的有偏反因果关系，该关系会随不同子群体而变化（即 $r(\mathcal{E}^*)$ 和 $\sigma(\mathcal{E}^*)$ 会变化），从而导致数据异质性。与定理 8 类似， $\mathbb{E}[f(S)S] = 0$ 表示模型设定错误，常数 L则刻画了 V与 Y之间有偏反因果关系的强度，其中 L 越大，意味着当 $\mathcal{E}^*$ 缺失时， V对于预测 Y所能提供的信息越多。从公式 16 的近似结果可以看出，我们提出的预测异质性由两项主导：（1）$\mathrm{Var}[r(\mathcal{E}^*)]/\mathbb{E}[r^2(\mathcal{E}^*)]$ 刻画了 $r(\mathcal{E}^*)$在子群体间的方差；（2） $\mathbb{E}[f^2(S)]+\sigma_Y^2$反映了 V所能提供的额外信息的最大值。

从更广泛的角度来看，定理 1、2 和 3 表明，我们提出的预测异质性度量具有显著的特性：对同质情形不敏感，并能够解释由典型数据异质性来源所产生的潜在异质性。这些发现凸显了该度量在各种机器学习任务中准确刻画预测异质性方面的有效性。

3.4 预测异质性估计的 PAC 保证

在明确的计算约束下定义的预测异质性，若模型族 $\mathcal{V}$ 的复杂度有界，则可以进行经验估计并获得保证。在

本研究利用拉德马赫复杂度给出了有限样本泛化界。首先，我们介绍经验预测异质性的定义，其显式公式见定义10。

数据集 $\mathcal{D}=\{(x_i,\ y_i)\}_{i=1}^{|\mathcal{D}|}$ 独立同分布地抽取自总体 X，Y。给定函数族 $\mathcal{V}$和环境集合 $\mathscr{E}_K$，使得对于 $\mathcal{E}\in\mathscr{E}_K$，$\text{supp}(\mathcal{E})=\{(e_k)_{k=1}^K\}$。，令 $\mathcal{Q}$为 X、Y、$\mathcal{E}$的所有概率分布构成的集合，其中$\mathcal{E}\in\mathscr{E}_K$。经验预测异质性 $\hat{}\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\ \mathcal{D})$ 定义如下：

$$\hat{\mathcal{H}}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})=\sup_{\mathcal{E}\in\mathscr{E}_K}\hat{\mathbb{I}}_{\mathcal{V}}(X\to Y|\mathcal{E};\mathcal{D})-\hat{\mathbb{I}}_{\mathcal{V}}(X\to Y;\mathcal{D}) \tag{18}$$

$$=\sup_{\hat{Q}\in\mathcal{Q}}\sum_{k=1}^{K}\Big[\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|\mathcal{E}=e_k;\mathcal{D})-\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|X^{\mathcal{E}=e_k};\ \mathcal{D})\Big] \tag{19}$$

$$-[\hat{H}_{\mathcal{V}}(Y;\mathcal{D})-\hat{H}_{\mathcal{V}}(Y|X;\mathcal{D})]. \tag{20}$$

具体而言，

$$\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|X,\mathcal{E}=e_k;\mathcal{D}) \tag{21}$$

$$=\inf_{f\in\mathcal{V}}\hat{Q}(\mathcal{E}=e_k)\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i)\frac{\hat{Q}(x_i,y_i|\mathcal{E}=e_k)}{\sum_{x_j,y_j\in\mathcal{D}}\hat{Q}(x_j,y_j|\mathcal{E}=e_k)} \tag{22}$$

$$=\inf_{f\in\mathcal{V}}\hat{Q}(\mathcal{E}=e_k)\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i)\frac{\hat{Q}(\mathcal{E}=e_k|x_i,y_i)\hat{Q}(x_i,y_i)}{\sum_{x_j,y_j\in\mathcal{D}}\hat{Q}(\mathcal{E}=e_k|x_j,y_j)\hat{Q}(x_j,y_j)} \tag{23}$$

$$=\inf_{f\in\mathcal{V}}\hat{Q}(\mathcal{E}=e_k)\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i)\frac{\hat{Q}(\mathcal{E}=e_k|x_i,y_i)\hat{Q}(x_i,y_i)}{\hat{Q}(\mathcal{E}=e_k)} \tag{24}$$

$$=\text{下确界}_{f\in\mathcal{V}}\sum_{x_i,\ y_i\in\mathcal{D}}\hat{}\hat{}-\text{对数}\ f[x_i](y_i)Q(\mathcal{E}=e_k|x_i,\ y_i)Q(x_i,\ y_i) \tag{25}$$

$$=\inf_{f\in\mathcal{V}}\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i)\hat{Q}(\mathcal{E}=e_k|x_i,y_i). \tag{26}$$

$\hat{}\hat{}\hat{}\hat{}\ Q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|\mathcal{E}=e_k;\ \mathcal{D})$、$H_{\mathcal{V}}(Y|X;\ \mathcal{D})$ 和 $H_{\mathcal{V}}(Y;\ \mathcal{D})$ 的显式公式也可以类似地推导得出。现在，我们可以正式定义经验预测异质性。

定义*10*（经验预测异质性）对于预测任务 $X\to Y$（其中 X，Y取值于 $\mathcal{X}\times\mathcal{Y}$），数据集 $\mathcal{D}$ 独立同分布地抽取自总体，满足 $\mathcal{D}=\{\ (x_i,\ y_i)\ _{i=1}^{N}\sim X,\ Y\}$。给定预测族 $\mathcal{V}$ 和环境集合 $\mathscr{E}_K=\{\mathcal{E}|\mathcal{E}\in\mathcal{C},|\textit{supp}(\mathcal{E})|=K\}$，其中 $K\in\mathbb{N}$。不失一般性，我们规定 $\textit{supp}(\mathcal{E})=\{\ (e_k)\ _{k=1}^{K}\}$，其中 e_k 表示单个环境。令 $\mathcal{Q}$ 为 X、Y、$\mathcal{E}$ 的所有概率分布的集合，其中 $\mathcal{E}\in\mathscr{E}_K$。相对于 $\mathcal{D}$ 的经验预测异质性 $\hat{}\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})$ 定义如下：

$$\hat{\mathcal{H}}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})=\sup_{\hat{Q}\in\mathcal{Q}}\sum_{k=1}^{K}\Big[\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|\mathcal{E}=e_k;\mathcal{D})-\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|X,\mathcal{E}=e_k;\mathcal{D})\Big]$$
$$-[\hat{H}_{\mathcal{V}}(Y;\mathcal{D})-\hat{H}_{\mathcal{V}}(Y|X;\mathcal{D})], \tag{27}$$

其中

$$\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|X,\mathcal{E}=e_k;\mathcal{D})=\inf_{f\in\mathcal{V}}\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i)\hat{Q}(\mathcal{E}=e_k|x_i,y_i). \tag{28}$$

$$\hat{Q}(\mathcal{E}=e_k)\hat{H}_{\mathcal{V}}(Y|\mathcal{E}=e_k;\mathcal{D})=\inf_{f\in\mathcal{V}}\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}-\log f[\emptyset](y_i)\hat{Q}(\mathcal{E}=e_k|x_i,y_i). \tag{29}$$

$$\hat{H}_{\mathcal{V}}(Y|X;\mathcal{D})=\inf_{f\in\mathcal{V}}\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}-\log f[x_i](y_i). \tag{30}$$

$$\hat{H}_{\mathcal{V}}(Y;\mathcal{D})=\inf_{f\in\mathcal{V}}\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}}-\log f[\emptyset](y_i). \tag{31}$$

接下来，我们在定理 11 中给出经验可用预测异质性的 PAC界。

定理 *11*（*PAC*界）考虑预测任务 $X\to Y$，其中 X、Y 是取值于 $\mathcal{X}\times\mathcal{Y}$ 的随机变量。假设预测族 $\mathcal{V}$ 满足 $\forall x\in\mathcal{X}$、$\forall y\in\mathcal{Y}$、$\forall f\in\mathcal{V}$，且 $\log f[x](y)\in[-B,B]$，其中 $B>0$。对于给定的 $K\in\mathbb{N}$，环境集合定义为 $\mathscr{E}_K=\{\mathcal{E}|\mathcal{E}\in\mathcal{C},|supp(\mathcal{E})|=K\}$，其中 $K\in\mathbb{N}$。不失一般性，我们规定 $supp(\mathcal{E})=\{(e_k)_{k=1}^K\}$，其中 e_k 表示单个环境。令 $\mathcal{Q}$ 为 X、Y、$\mathcal{E}$ 的所有概率分布构成的集合，其中 $\mathcal{E}\in\mathscr{E}_K$。取一个 $e\in supp(\mathcal{E})$，并定义函数类 $\mathcal{G}_{\mathcal{V}}=\{g|g(x,y)=\log f[x](y)\ Q(\mathcal{E}=e|x,y)\ f\in\mathcal{V},Q\in\mathcal{Q}\}$。记 $\mathcal{G}$ 在 N 个样本下的拉德马赫复杂度为 $\mathscr{R}_N(\mathcal{G})$。那么，对于任意 $\delta\in(0,1/(2K+2))$，以关于 $1-2(K+1)\delta$ 的概率，对于从 X 独立同分布抽取的数据集 $\mathcal{D}$，有：

$$|\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y)-\hat{\mathcal{H}}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})|\le 4(K+1)\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}})+2(K+1)B\sqrt{2\log\frac{1}{\delta}/|\mathcal{D}|}, \tag{32}$$

其中 $\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}})=\mathcal{O}(|\mathcal{D}|^{-\frac{1}{2}})$（*Bartlett*和*Mendelson*，*2002*）。证明见附录 *D*。

4 算法

为了根据定义10进行预测异质性的经验估计，我们从公式27中的形式化定义推导出信息最大化（IM）算法，以推断使经验预测异质性 $\hat{\mathcal{H}}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})$ 最大化的 $\mathcal{E}$ 分布。

目标函数。给定数据集 $\mathcal{D}=\{X_N,Y_N\}=\{(x_i,\ y_i)\}_{i=1}^N$，记 $\text{supp}(\mathcal{E})=\{e_1,\dots,e_K\}$，我们使用权重矩阵 $W\in\mathcal{W}_K$对 $\mathcal{E}|(X_N,\ Y_N)$ 的分布进行参数化，其中 K $\mathcal{W}_K=\{W\ \ W\in\mathbb{R}_+^{N\times K}\ \ W\mathbf{1}_K=$是预$\mathbf{1}_N\}$ w_{ij} W $P\ \mathcal{E}=e_j|x_i\ y_i$ 先定义的环境数量，且 : 是允许的权重空间。中的每个元素表示 (，) （ i数据点属于 j子群体的概率）。对于预测族 $\mathcal{V}$，定义10中的上确界问题的解等价于以下目标函数：

$$\begin{aligned}\min_{W\in\mathcal{W}_K}\mathcal{R}_{\mathcal{V}}(W,\theta_1^*(W),\dots,\theta_K^*(W))&=\left\{\frac{1}{N}\sum_{i=1}^N\sum_{j=1}^K w_{ij}\ell_{\mathcal{V}}(f_{\theta_j^*}(x_i),y_i)+U_{\mathcal{V}}(W,Y_N)\right\},\\ \text{s.t.}\quad\theta_j^*(W)&\in\arg\min_{\theta}\left\{\mathcal{L}_{\mathcal{V}}^j(W,\theta)=\sum_{i=1}^N w_{ij}\ell_{\mathcal{V}}(f_\theta(x_i),y_i)\right\},\quad\text{for } j=1,\dots,K,\end{aligned} \tag{33}$$

其中， f_θ： $\mathcal{X} \to \mathcal{Y}$表示由 θ 参数化的预测函数， $\ell_\mathcal{V}(\cdot,\ \cdot)$： $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$表示损失函数， $U_\mathcal{V}(W,\ Y_N)$ 是正则化项。具体而言， f_θ、 $\ell_\mathcal{V}$和 $U_\mathcal{V}$由预测族 $\mathcal{V}$决定。下面给出两个典型且通用的机器学习任务，即回归和分类的实现。

4.1 回归

对于回归任务，预测族通常建模如下：

$$\mathcal{V}_1 = \{g\colon\ g[x] = \mathcal{N}(f_\theta(x)\ \sigma^2)\ f\text{是预测函数，}\ \theta\text{是可学习的，}\ \sigma\text{是常数}\}. \tag{34}$$

相应的损失函数为 $\ell_{\mathcal{V}_1}(f_\theta(X)\ Y)\ =(f_\theta(X)-Y)^2,$，而 $U_{\mathcal{V}_1}(W, Y_N)$ 变为

$$U_{\mathcal{V}_1}(W, Y_N) = \mathrm{Var}_{j\in[K]}(\overline{Y_N^j}) = \sum_{j=1}^{K}\left(\sum_{i=1}^{N} w_{ij}y_i\right)^2 \frac{1}{N\sum_{i=1}^{N} w_{ij}} - \left(\frac{1}{N}\sum_{i=1}^{N} y_i\right)^2 \tag{35}$$

其中， $\overline{Y_N^j}$表示给定 $\mathcal{E} = e_j$时标签 Y的均值， $U(W, Y_N)$计算子群体 $e_1 \sim e_K$中 Y_N^j的方差。

4.2 分类

对于分类任务，预测族通常建模为：

$$\mathcal{V}_2 = \{g : g[x] = f_\theta(x) \in \Delta_c, f \text{ is the classification model and } \theta \text{ is learnable}\}, \tag{36}$$

其中， c 是类别数， Δ_c 表示 c维单纯形。这里，预测族 $\mathcal{V}_2$ 中的每个模型都会输出一个 c维单纯形形式的离散分布。在这种情况下，相应的损失函数$\ell_{\mathcal{V}_2}(\cdot,\cdot)$是交叉熵损失，正则化项则变为$U_{\mathcal{V}_2}(W, Y_N)\ =\ -\sum_{j=1}^{K}\frac{1}{N}(\sum_{i=1}^{N} w_{ij})\overline{H(Y_N^j)}$，其中 $H(Y_N^j)$是给定 $\mathcal{E} = e_j$时 Y的熵。

4.3 优化。

方程 33 中的双层优化可以通过对 W执行投影梯度下降来求解。 W 的梯度可通过以下方式计算：（此处省略下标 $\mathcal{V}$）

$$\nabla_W\mathcal{R} = \nabla_W U + \left[\ell(f_{\theta_j}(x_i), y_i)\right]_{i,j}^{N\times K} + \sum_{j=1}^{K}\boxed{\nabla_{\theta_j}\mathcal{R}|_{\theta_j^*}\nabla_W\theta_j^*}, \tag{37}$$

$$\text{where } \boxed{\nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^*}\nabla_W\theta_j^*} \approx \nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^t}\sum_{h\le t}\left[\prod_{k<h}(I - \frac{\partial^2\mathcal{L}^j}{\partial\theta_j\partial\theta_j^{\mathrm{T}}}\Big|_{\theta_j^{t-k-1}})\right]\frac{\partial^2\mathcal{L}^j}{\partial\theta_j\partial W^{\mathrm{T}}}\Big|_{\theta_j^{t-h-1}} \tag{38}$$

$$\approx \nabla_{\theta_j}\mathcal{R}\big|_{\theta_j^t}\frac{\partial^2\mathcal{L}^j}{\partial\theta_j\partial W^{\mathrm{T}}}\Big|_{\theta_j^{t-1}} \quad, \text{for } j = 1, \ldots, K. \tag{39}$$

其中， $\ell(f_{\theta_j}(x_i)\ y_i)]_i^{N\times K}$，$_j$是一个 $N \times K$矩阵，其（i， j）-th 个元素为 $\ell(f_{\theta_j}(x_i)\ y_i)$。这里，方程 38 使用内循环梯度下降的 t 步将 θ_j^* 近似为 θ_j^t ，而方程 39 获取 $t\ =\ 1$ 并执行*1* 步截断反向传播（Shaban等，

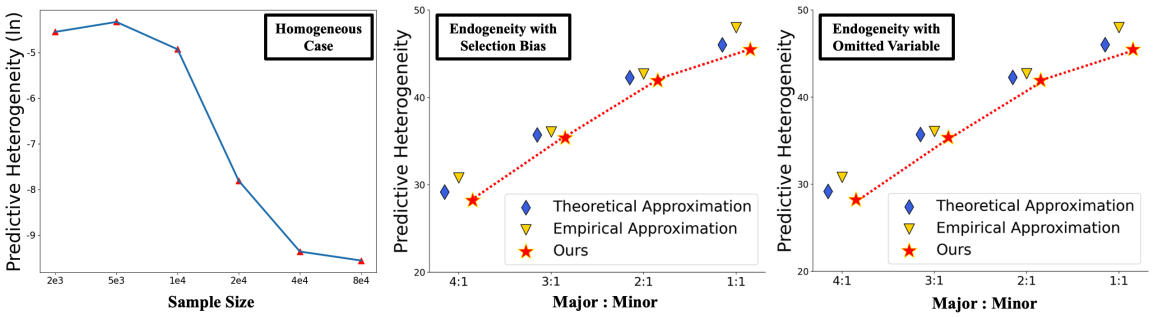


图1：第 3.3 节玩具示例的数值结果。左图绘制了定理 7 所设情形下估计得到的预测异质性；中图绘制了定理 8 所设情形下的理论近似、经验近似和我们的结果；右图对应定理 9 所设情形。

2019年；Zhou等，2022年）。我们的信息最大化算法通过投影梯度下降更新 W，如下所示：

$$W \leftarrow \mathrm{Proj}_{\mathcal{W}_K}\left(W - \eta \nabla_W \mathcal{R}\right), \quad \eta \text{ is the learning rate of } W. \tag{40}$$

接下来，我们证明，最小化公式 33 能够精确求得定义10（正式定义）中关于 $\mathcal{E}$ 的上确界，即经验预测异质性的上确界。

定理12（信息最大化算法的理论依据）对于具有预测族 $\mathcal{V}_1$ 的回归任务和具有 $\mathcal{V}_2$ 的分类任务，公式 *33* 的优化分别等价于公式*27*中预定义环境数量为 K（即 $|supp(\mathcal{E})| = K$）时经验预测异质性$\hat{}\mathcal{H}_{\mathcal{V}_1}^{\mathcal{E}_K}(X \to Y; \mathcal{D})$、$\hat{}\mathcal{H}_{\mathcal{V}_2}^{\mathcal{E}_K}(X \to Y; \mathcal{D})$ 的上确界问题。证明见附录*E*。

注13（与期望最大化的区别）期望最大化（*EM*）算法旨在推断统计模型中的潜变量，以实现最大似然。我们提出的信息最大化（*IM*）算法旨在推断潜变量 W，从而获得与最大信息量相关的最大预测异质性。由于目标函数中的正则化项 $U_{\mathcal{V}}$，*EM* 算法无法高效地解决我们的问题，因此我们采用双层优化技术。

4.4 近似精度

这里我们给出第 3.3 节线性示例的一些补充数值结果。在图1左侧的子图中，我们绘制了定理 7 所述设定下估计的预测异质性，此时解析解等于 0。从结果中可以看出，随着样本量增加，我们的 IM 算法的估计值趋近于 0（注意，y-轴表示 ln（估计值））。在中间的子图中，对于定理 8 中的设定，我们绘制了理论近似、经验近似（有限样本情形）以及不同多数类与少数类比例下预测异质性的估计值（该比例控制了公式 13 中的 $\mathrm{Var}[r(\mathcal{E}^*)]$ ）。并且

右侧子图绘制了定理 9 所述设定下的相同数值。从这两幅图中可以看出：(1) 有限样本下的经验近似与理论近似非常接近，这与我们在定理 11 中的泛化界一致；(2) 我们的 IM 算法的估计值与理论近似非常接近，这证明了公式 38 和 39 中近似算法的准确性。此外，随着比例从 4 : 1 变为 1 : 1，数据异质性不断增大，预测异质性也随之增大，这由公式 13 和 16 中的 Var($r(\mathcal{E}^*)$) 项控制。

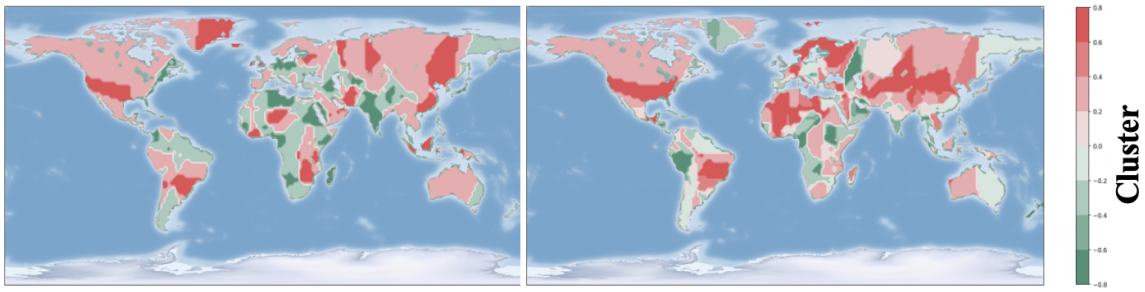
5 实验

5.1 揭示可解释的子群体结构

预测异质性可为子群体划分提供宝贵见解，并支持农业、社会学研究及物体识别等多个领域的决策。下面的示例表明，学习得到的子群体划分具有很强的可解释性，并且与决策需求密切相关。

示例：农业 众所周知，气候会影响作物产量和作物适宜性（Lobell 等，2008）。我们使用美国国家海洋和大气管理局数据库中的数据，该数据库包含世界各地气象站的每日天气数据。遵循 Zhao 等（2021）的方法，我们从 2018 年的天气序列中提取了汇总统计量，包括年平均气温、湿度、风速和降雨天数。任务是利用该地点的天气汇总统计量 和位置协变量（即经度和纬度），预测各地的作物产量 。为便于说明，我们重点关注作物类型为小麦或水稻的地点。值得注意的是，我们的输入协变量不 包含作物类型信息。我们在此任务中使用多层感知机模型，并为 IM 算法设置 $K = 2$ 。

鉴于作物产量预测机制与农作物类型密切相关，而预测任务中并不知道农作物类型，我们认为这会导致整个数据集存在数据异质性，而识别出的预测异质性应当与之相关。为展示我们度量方法的合理性，我们在图 2(a) 中绘制了小麦和水稻种植区域的真实分布图，并在图



(a) Division of wheat and rice cultivation areas **(b) Division learned by our algorithm**

图 2：作物产量数据上的结果。我们根据每个区域的主要作物类型进行着色，色调表示通过k-均值（$k = 3$）平滑处理后主要作物类型所占的比例。

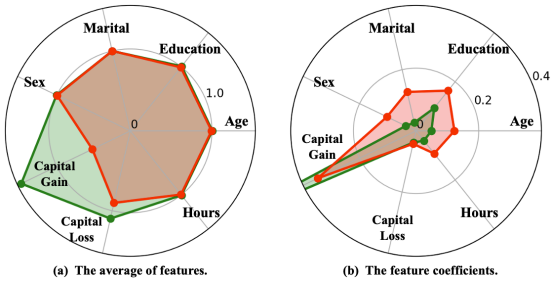


图3：Adults 数据的结果。这里展示两个已学习子群体的特征均值和特征系数。

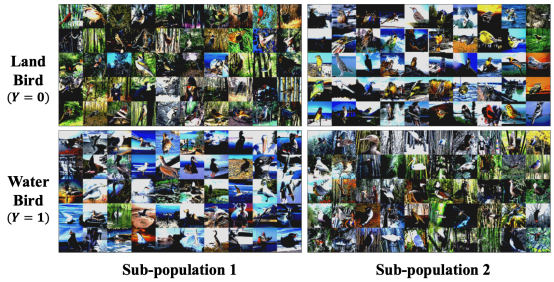


图4：Waterbird 数据的结果。这里为每个类别和每个已学习的子群体随机抽取 50 张图像。

2(b) 中展示了通过我们的 IM 算法学习得到的两个子群体。我们的算法给出的划分与两种农作物的真实划分非常相似，表明了该度量方法的合理性。我们观察到，在亚洲的青藏高原等区域存在一些差异，这可能是因为缺少人口密度和海拔等会显著影响作物产量的重要特征。

示例：社会学 我们使用 UCI Adult 数据集（Kohavi 和 Becker，1996年）。该数据集广泛用于算法公平性研究，源自美国人口普查局开展的1994年当前人口调查。任务是根据个人特征，预测一个人的收入高于还是低于 50,000 美元。在此任务中，我们使用线性模型，并设置 $K = 2$。在本例中，我们旨在探究数据中的子群体结构 是否会影响机器学习模型的学习。

在图3（a）中，我们绘制了两个子群体的汇总统计数据，揭示了资本利得方面的一个关键差异。在图3（b）中，我们展示了线性模型针对这两个子群体给出的特征重要性。我们发现，对于资本利得较高的个体，预测模型主要依赖资本利得，这是公平的。然而，对于资本利得较低的个体，模型还会考虑性别和婚姻状况等敏感属性，而这些属性已知可能导致歧视。我们的结果与 Zhao 等人（2021年）的发现一致，并有助于识别决策中的潜在不平等。例如，我们的发现表明，资本利得较低的个体可能面临歧视，这或可推动算法设计并提升政策公平性。

示例：物体识别 最后，我们使用 Waterbird 数据集（Sagawa 等，2019年）来研究虚假相关性对机器学习模型的影响。该数据集被广泛用作鲁棒学习领域的基准。任务是识别水鸟或陆鸟，但图像中的背景与目标标签之间存在虚假相关性。在多数图像中，水鸟位于水域，陆鸟位于陆地；而在少数图像中，这种相关性则相反。因此，虚假相关性导致该数据集出现预测异质性，进而可能显著影响机器学习模型的性能。在此示例中，我们使用 ResNet18，并在 IM 算法中将 $K = 2$ 设为 。

我们的方法成功捕捉到了虚假相关性，并识别出两个子群体，其图像中的目标与背景之间存在相反的相关性。为展示我们方法的有效性，我们从每个类别和每个所学习的子群体中随机抽取50张图像，并将其绘制在图4中。在子群体1中，多数

陆鸟位于地面上，而水鸟位于水中；在子群体 2 中，多数陆鸟位于水中，而水鸟位于地面上。我们的研究结果表明，稳健学习方法（Sagawa 等，2019；Koyama 和 Yamaguchi，2020）可以利用所提出的方法来提升机器学习模型的泛化能力。通过消除虚假相关性的影响，我们的方法能够显著增强机器学习模型的鲁棒性和可靠性。总体而言，我们的研究凸显了在图像分类任务中应对预测异质性的重要性，并为实现稳健学习性能提供了一种实用解决方案。

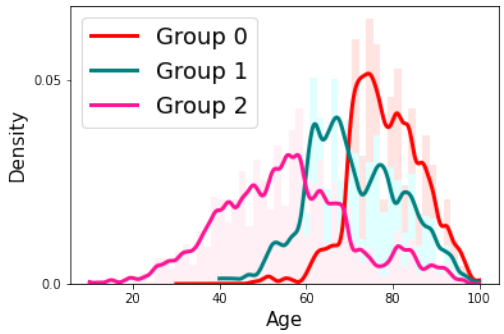


图 5：COVID-19 数据的结果。我们绘制了每个学习得到的亚组中死亡者（$Y = 1$）的年龄分布。

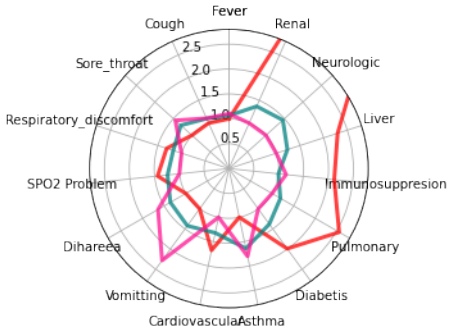


图 6：COVID-19 数据的结果。我们展示了每个学习得到的亚组中死亡者（$Y = 1$）的典型特征均值。

5.2 辅助科学发现：揭示与死亡率相关的因素

为了充分展示预测异质性的有效性，我们聚焦于医疗保健领域的应用，使用巴西患者的 COVID-19 数据集。该数据集包含来自巴西的 6882 名 COVID-19 检测呈阳性的患者，其数据记录时间为2020年2月27日至5月4日。数据集涵盖了广泛的风险因素，包括合并症、症状和人口统计学特征。二元标签对应于由 COVID-19 导致的死亡。我们的目标是验证通过该方法在此数据集中学习得到的子群体，全面解释每个群体 并展示如何利用预测异质性揭示其他方式难以检测到的与死亡率相关的特征。

5.2.1 Learned Sub-populations.

根据风险因素预测死亡率时，重要的是要考虑到，患有不同基础疾病且人口统计学特征各异（如年龄和性别）的患者，其死亡概率可能有所不同。此外，不同个体的死亡率可能由不同因素导致。鉴于这些因素，该数据集中的预测异质性源于不同子群体中导致死亡率的机制多种多样。

在本实验中，我们使用线性模型，损失函数为二元交叉熵损失。我们选择能使经验预测异质性达到最大值的子群体数量 $K \in \{2, 3, 4, 5, 6\}$

$\hat{\mathcal{H}}_V^{\mathcal{E}^K}(X \to Y; \mathcal{D})$，由此得到三个不同的亚组（最优的 $K = 3$）。此外，我们通过实证观察发现，当 $K > 3$ 时，学习得到的子群体会缩减为 3 个子群体。在图 5 和 6 中，为了更全面地考察学习得到的亚组，我们分析了各组的年龄分布以及对应风险因素的平均值。我们的分析揭示了若干值得注意的发现：

1. 我们观察到，学习得到的亚组在年龄分布上存在明显差异。具体而言，组0主要由70岁以上的人群构成，而组1由年龄约为60岁的人群构成。另一方面，第2组由跨越多个年龄段的中年人群构成。

2. 风险因素的平均值显示，各亚组之间存在显著差异，表明导致死亡率的原因各不相同。具体而言，与其他组相比，组0中患有基础疾病的人群比例明显更高，例如肾脏、神经系统、肝脏疾病以及免疫抑制。相比之下，组1的基础疾病水平明显较低。有趣的是，第2组没有任何基础疾病，但腹泻和呕吐的发生率明显更高。这些发现表明，学习得到的亚组可用于识别与死亡率相关的特定风险因素，从而为具有不同风险特征的个体制定有针对性的干预措施提供依据。

在识别出各亚组之间的不同模式后，我们进一步探究与死亡率相关的具体风险因素。为进一步验证研究结果，我们纳入了领域专家的专业知识。借助他们的见解，我们得以确认所识别风险因素的可靠性，以及亚组分析的重要性。

5.2.2 Scientific Findings

根据学习得到的分组，我们针对每个组分别拟合逻辑回归模型，并选取系数最大的前 6 个特征，如表1所示。

首先，我们的分析发现，在组0和组1中，与死亡率相关性最高的特征主要是血氧饱和度和基础疾病，这与老年人常见的风险因素一致。相比之下，第2组呈现出一组不同的关键特征，包括发热、咳嗽和呕吐等新冠肺炎症状。值得注意的是，第2组由年龄跨度较大的中年人组成。我们的研究结果表明，无论年龄如何，严重的新冠肺炎症状都可能导致死亡。

其次，为了进一步开展分析，我们针对整个数据集拟合了一个模型，并观察到排名靠前的特征仍然是SPO2和基础疾病，这与针对老年人发现的排名靠前的特征一致。然而，对于面向通常没有严重基础疾病的年轻人或中年人的干预措施而言，这可能并无益处，甚至可能造成伤害。例如，医生可能倾向于对出现严重新冠肺炎症状的年轻患者持乐观态度，并低估其死亡风险，因为他们通常没有基础疾病。因此，探索并利用数据中的预测异质性，有助于获得更可靠的科学发现，同时避免潜在异质性可能造成的伤害。

第三，我们的分析揭示了第2组中的两个重要特征，即呕吐和腹泻，而传统分析很少考虑这些因素。我们查阅了有关新冠肺炎的相关文献，发现多项研究已确认，这两种症状是新冠肺炎导致死亡率升高的重要指标。Zhong et al.（2020）强调了新冠肺炎中胃肠道和肝损伤的潜在机制，以提高人们对新冠肺炎消化系统损伤的认识。Liu et al.（2021b）分析了2020年3月21日之前在中国大陆武汉以外城市确诊的29,393名实验室确诊的COVID-19患者，发现同时出现胃肠道症状和发热的患者，以及仅发热的患者，其死亡风险均显著更高；其中，胃肠道症状指以下症状之一：恶心、呕吐、腹泻或腹痛。Zeng et al.（2021）还发现，胃肠道症状与新冠肺炎的严重程度相关，在伴有胃肠道症状的新冠肺炎患者中，重症率超过40%。Ghimire et al.（2021）证明，腹泻作为首发症状出现，与疾病严重程度增加以及可能更差的预后相关。Chan et al.（2022）呼吁，对于出现腹痛和胃肠道症状、且这些症状符合胃肠炎或外科急腹症特征的患者，即使其缺乏新冠肺炎的呼吸道症状，也应将新冠肺炎纳入鉴别诊断。这些研究验证了我们研究结果的可靠性，并表明利用所提出的预测异质性开展研究，可以发现整体数据集分析中未出现的非典型风险因素。

本例旨在说明我们的预测异质性能够为众多科学领域带来的潜在益处。通过利用数据集中的异质性，我们的方法能够揭示传统分析可能忽略的新模式和关联，从而推动更可靠、更全面的科学发现

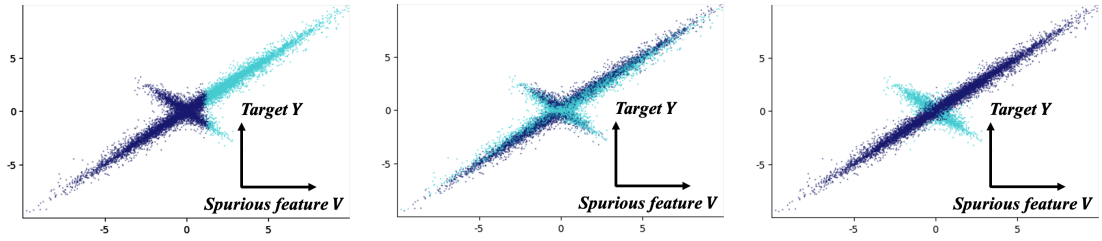
表1：COVID-19 数据集中每个学习得到的亚组及整体数据的主要特征。

组别 ID	主要特征					
0	SPO2	糖尿病	肾脏	神经系统	肺部	心血管
1	糖尿病	SPO2	神经系统	心血管	肺部	肾脏
2	**发热**	**咳嗽**	肾脏	**呕吐**	**呼吸短促**	**腹泻**
All	SPO2	肾脏	神经系统	糖尿病	肺部	心血管

5.3 效益泛化

本节旨在评估我们的 IM 算法在提升机器学习模型分布外泛化性能方面的效果。为此，我们在模拟数据和真实世界的彩色 MNIST 数据上进行了实验。结果表明，我们的 IM 算法所学习到的子群体结构能够显著提升机器学习模型的分布外泛化能力。

基线方法 首先，我们将其与经验风险最小化 （ERM）以及用于不变学习的环境推断（EIIL，（Creager 等，2021））进行比较，后者能够推断用于学习不变性的环境。随后，我们将其与广为人知的$KMeans$ 算法进行比较，后者是最常用的聚类算法。对于我们的 IM 算法和 KMeans，我们采用三种算法作为骨干网络，以利用所学习的子群体，包括子群体平衡和不变学习方法。子群体平衡



(a) KMeans。 (b) EIIL。 (c) 我们的 IM。

图 7：三种方法在模拟数据上的子群体划分，其中两种颜色表示两个子群体。

方法仅对所学习的子群体赋予相同权重。不变风险最小化 （IRM, （Ar- jovsky 等, 2019））和环境间梯度对齐 （IGA, （Koyama 和 Yam- aguchi, 2020））是不变学习方法中的典型方法，应用于分布外泛化，并将子群体作为输入环境，以学习不变模型。

5.3.1 Simulation Data of Sample Selection Bias

输入特征 $X = [S, T, V]^T \in \mathbb{R}^{10}$ 由稳定特征 $S \in \mathbb{R}^5$、噪声特征 $T \in \mathbb{R}^4$ 和伪特征 $V \in \mathbb{R}$组成:

$$S \sim \mathcal{N}(0, 2\mathbf{I}_5), T \sim \mathcal{N}(0, 2\mathbf{I}_4), Y = \theta_S^T S + h(S) + \mathcal{N}(0, 0.5), V \sim \text{Laplace}(\text{sign}(r) \cdot Y, 1/(5 \ln |r|)) \quad (41)$$

其中， $\theta_S \in \mathbb{R}^5$ 是系数， $h(S) = S_1 S_2 S_3$ 是非线性项。 $|r| > 1$ 是每个子群体的一个因子，这里数据异质性由含有隐藏变量的内生性（Fan 等，2014）引起。 V 是伪特征，它与 Y 的关系在不同子群体间不稳定，并由因子 r控制。直观而言，sign(r)控制 V 与 Y 之间伪相关的正负。 $|r|$ 控制伪相关的强度，即 $|r|$ 越大，伪相关越强。在训练中，我们生成 10000 个点，其中主要群体包含 80% 的数据， $r = 1.9$ （即强正向伪相关）；次要群体包含 20% 的数据， $r = -1.9$（即强负向伪相关）。在测试中，我们分别测试两个群体的表现，并且还设置 $r = -2.3$ 和 $r = -2.7$

表 2：分布外泛化实验结果，包括模拟数据和彩色 MNIST 数据。

方法		1. 模拟数据				2. 彩色 MNIST	
		训练子群体误差		测试误差		训练准确率	测试准确率
		多数群体（$r = 1.9$）	少数群体（$r = -1.9$）	$r = -2.3$	$r = -2.7$		
ERM		0.255(±0.024)	0.740(±0.022)	0.738(±0.035)	0.737(±0.023)	0.998(±0.001)	0.406(±0.019)
EIIL		**0.164**(±0.014)	1.428(±0.035)	1.431(±0.061)	1.431(±0.046)	0.812(±0.006)	0.610(±0.016)
KMeans	平衡	0.231(±0.022)	0.847(±0.024)	0.846(±0.039)	0.845(±0.026)	**0.999**(±0.001)	0.328(±0.021)
	IRM	0.231(±0.022)	0.845(±0.024)	0.844(±0.039)	0.843(±0.026)	0.947(±0.004)	0.259(±0.021)
	IGA	0.235(±0.022)	0.840(±0.023)	0.839(±0.038)	0.838(±0.027)	0.997(±0.001)	0.302(±0.021)
Ours	平衡	0.403(±0.041)	**0.423**(±0.016)	**0.416**(±0.022)	**0.416**(±0.014)	0.749(±0.012)	**0.692**(±0.039)
	IRM	0.391(±0.039)	**0.432**(±0.016)	**0.430**(±0.022)	**0.430**(±0.014)	0.759(±0.014)	**0.727**(±0.047)
	IGA	0.449(±0.037)	**0.426**(±0.017)	**0.417**(±0.022)	**0.417**(±0.014)	0.759(±0.012)	**0.713**(±0.034)

以模拟更强的分布偏移。我们使用线性回归，并为所有方法设置 $K = 2$ ，然后报告所有方法的均方误差（MSE）。

10 次运行的结果如表 2 所示。从表 2 的结果可以看出，对于模拟数据和彩色 MNIST 数据，采用我们 IM 算法的两个骨干网络都取得了最佳的分布外泛化性能。此外，对于模拟数据，学到的预测异质性使骨干算法能够平等对待数据中的多数类和少数类（即“多数群体”与“少数群体”之间的性能差距较小），并显著提升分布外泛化性能。此外，我们在图 7 中绘制了采用我们 IM 算法学到的子群体。从图 7 可以看出，与 KMeans 和 EIIL 相比，我们的预测异质性利用了 V 与 Y之间的伪相关，并使骨干算法能够消除这种伪相关。

5.3.2 Simulation Data of Hidden Variables

此外，我们增加了一项实验来说明：（1）当所选的 K 小于真实值时，我们的方法性能会下降，但仍优于经验风险最小化；（2）当所选的 K 较大时，性能不会受到太大影响。

输入特征 $X = [S, T, V] \in \mathbb{R}^{10}$ 由稳定特征 $S \in \mathbb{R}^5$、噪声特征$T \in \mathbb{R}^4$ 和伪特征 $V \in \mathbb{R}$组成：

$$S \sim \mathcal{N}(2, 2\mathbb{I}_5), \quad T \sim \mathcal{N}(0, 2\mathbb{I}_4), \quad Y = \theta_S^T S + S_1 S_2 S_3 + \mathcal{N}(0, 0.5),$$

并通过以下方式生成伪特征：

$$V = \theta_V^e Y + \mathcal{N}(0, 0.3),$$

其中， θ_V^e 在不同子群体之间变化，并取决于数据点所属的子群体。在训练中，我们从 e_1 中抽取 8000 个数据点，其 $\theta_V^1 = 3.0$；从 e_2 中抽取 1000 个数据点，其 $\theta_V^2 = -1.0$；从 e_3 中抽取 1000 个数据点，其 $\theta_V^3 = -2.0$ ；并从e_4 中抽取 1000 个数据点，其 $\theta_V^4 = -3.0$。因此，真实值中的子群体数量为 4。在测试中，我们在 e_4 上测试性能，其 $\theta_V^4 = -3.0$，与训练数据存在显著的分布偏移。10 次运行的平均 MSE 如图 8 所示。

从结果中可以看出，当 K 小于真实值时，增加K 有利于分布外泛化性能；而当 K 较大时，性能不会受到太大影响。

对于我们的 IM 算法，我们认为主要有两种方式可以选择 K：

- 根据预测异质性指数：当选定的 K 小于真实值时，随着 K增加，我们的度量值往往会快速上升；而当K 大于真实值时，上升速度会减慢，这可以指导人们选择合适的 K。

- 根据预测模型：由于我们的 IM 算法旨在学习子群体 s
  对于具有不同预测机制的情况，可以比较学习得到的模型参数 θ_1，.....， θ_K，以判断 K是否远大于真实值；也就是说，如果两个得到的模型非常相似，那么 K 可能过大了（将一个子群体划分为两个）。对于线性模型，可以直接比较系数。对于深度模型，我们认为可以计算各子群体之间的迁移损失。

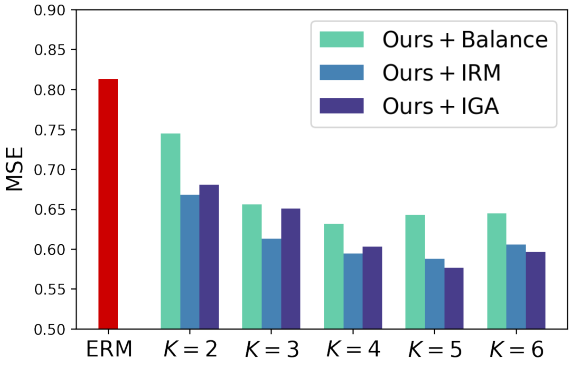


图 8：在补充实验中，以子群体平衡、IRM 和 IGA 为骨干网络时，我们的方法的 OOD 泛化误差。真实子群体数量为 4。



图 9：我们 IM 算法在 MNIST 数据上的子群体划分。

5.3.3 Colored MNIST

按照 Arjovsky 等人（2019年）的做法，我们设计了一个基于 MNIST 数据集构建的二元分类任务。首先，数字 $0 \sim 4$ 被标记为 $Y = 0$，数字 $5 \sim 9$ 被标记为 $Y = 1$。˜其次，以 0.2 的概率随机翻转标签 Y，从而引入噪声标签 $\tilde{Y}$。然后，我们对与 Y 存在虚假相关性的颜色标识 V 进行采样，具体如下：

$$V = \begin{cases} +\tilde{Y}, & \text{with probability } r, \\ -\tilde{Y}, & \text{with probability } 1 - r. \end{cases} \tag{42}$$

˜事实上，r 控制着 Y 与 V 之间的伪相关。在训练中，我们随机抽取 10000 个数据点，并设置 $r = 0.85$，这意味着对于 85% 的数据，V 与 Y 呈正˜相关；对于其余 15% 的数据，伪相关则变为负相关，从而˜导致 V 和 Y 存在数据异质性。在测试中，我们设置 $r = 0$（强负伪相关），使训练集与测试集之间产生强分布偏移。

从表 2 的结果可以看出，对于模拟数据和彩色 MNIST 数据，我们的 IM 算法与两种骨干网络结合时均取得最佳的分布外泛化性能。我们在图 9 中绘制了 IM 算法学习到的子群体。从图 9 可以看出，我们的方法学习到的子群体也反映了数字标签 Y 与颜色（红色或绿色）之间伪相关的不同方向，这有助于骨干网络方法避免利用颜色预测数字。

6 相关工作

据我们所知，数据异质性至今尚未形成统一的表述方式，在不同领域中也具有不同含义。Li 和 Reynolds（1995）根据系统属性以及复杂性或可变性来定义生态学 中的异质性。Rosenbaum（2005）将经济学观察性研究中潜在结果的不确定性视为个体异质性。对于图 数据，异质性指节点和边的多种类型（Wang 等人（2019））。近来，在机器学习中，若干

因果学习（Peters 等，2016；Arjovsky 等，2019；Koyama 和 Yamaguchi，2020；Creager 等，2021）和鲁棒学习（Sagawa 等，2019）的相关研究利用来自多个环境的异质数据来提升分布外泛化能力。具体而言，不变学习方法（Arjovsky 等，2019；Koyama 和 Yamaguchi，2020；Creager 等，2021；Zhou等，2022）利用异质环境来学习在各环境中表现一致的不变预测器。在分布鲁棒优化领域，Sagawa 等（2019）和 Duchi 等（2022）提出优化最差群体预测误差，以保证分布外泛化性能。然而，在机器学习领域，先前的研究尚未对数据异质性给出精确定义或可靠的量化方法，因此如何利用数据异质性来开发更合理的机器学习算法仍令人困惑且难以实现。

至于聚类算法，大多数算法只关注协变量 X，KMeans 和高斯混合模型（GMM，（Reynolds，2009））便是典型例子。然而，我们的实验表明，KMeans 学到的簇只能反映 $P(X)$ 中的异质结构。值得注意的是，我们的预测异质性可以反映 $P(Y|X)$ 中的异质性。期望最大化（EM，（Moon，1996））也可用于聚类。然而，我们的 IM 算法与 EM 存在本质区别：IM 算法推断能够使预测异质性最大化的潜变量，而 EM 最大化的是似然。此外，不变学习领域也有一些用于推断环境的方法（Creager 等，2021）。尽管这些方法有助于分布外泛化，但缺乏理论基础，且只适用于某些设置。

7 关于与子群体发现差异的讨论

子群体发现（SD，（Helal，2016））旨在提取不同变量（X）之间与目标变量 Y 相关的“有趣”关系。这类方法关注每个发现群体的覆盖度和精确率。具体而言，它在$P(X)$上学习一种划分，使得某个目标标签 y 在每个群体中占主导地位。子群体发现与我们的预测异质性之间最显著的区别，在于簇间分布偏移的模式：对于子群体发现，$P(X)$和 $P(Y)$会因亚组而异，但存在一个通用的 $P(Y|X)$。而对于预测异质性，$P(Y|X)$会因子群体而异，这表明预测机制具有多样性。正是预测机制的这种差异，使通用预测模型在异质数据集上的表现受限；这也是分布外问题和群体公平性所关注的重点。

我们认为，子群体发现更适用于分布偏移较小且对可解释性要求较高的场景，因为它能够生成人们易于理解的简化规则。此外，子群体发现方法适用于仅涉及表格数据（通常来自关系数据库）的场景，在这些场景中，输入特征具有明确的语义。我们提出的方法则能够应对一般的机器学习场景，包括涉及表示学习的复杂数据（例如图像数据）。另外，当人们需要处理数据内部存在预测机制异质性的场景时，我们的方法更为适用。不过，这两类方法都可以帮助人们理解数据并做出更合理的决策。

8 关于公平性潜力的讨论

我们发现，将我们的度量方法与算法公平性相结合是一个有趣且很有前景的方向，并且我们认为该度量方法有望应对算法偏见。我们的方法可以生成预测机制可能不同的子群体，这有助于以下方面：

风险特征选择：我们可以根据预测异质性度量来选择特征，以了解哪些特征带来的异质性最大。如果这些特征是敏感特征，人们就应避免其影响；如果不是，则可以据此构建更好的机器学习模型。

考察算法公平性：我们可以利用学得的子群体，通过计算各子群体之间的性能差距，考察给定算法是否公平。

9 结论

我们将预测异质性定义为首个对影响机器学习模型预测的数据异质性进行定量刻画的形式。我们证明了其理论性质，并表明它有助于提升分布外泛化性能。

附录 A. 命题 6 的证明

证明 [命题 6]的证明

1. 单调性：由于 $\mathscr{E}_1 \subseteq \mathscr{E}_2$,

$$\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_1}(X \to Y) = \sup_{\mathcal{E} \in \mathscr{E}_1} \mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(X \to Y) \tag{43}$$

$$\leq \sup_{\mathcal{E} \in \mathscr{E}_2} \mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}) - \mathbb{I}_{\mathcal{V}}(X \to Y) \tag{44}$$

$$= \mathcal{H}_{\mathcal{V}}^{\mathscr{E}_2}(X \to Y). \tag{45}$$

2. 非负性：
根据环境集合的定义，存在 $\mathcal{E}_0 \in \mathscr{E}$，使得对于任意 $e \in \text{supp}(\mathcal{E})$， X， $Y|\mathcal{E} = e$ 与 X， Y同分布。因此，我们有

$$\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X \to Y) = \sup_{\mathcal{E} \in \mathscr{E}} [H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) - H_{\mathcal{V}}(Y|X, \mathcal{E})] - [H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)] \tag{46}$$

$$\geq [H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}_0) - H_{\mathcal{V}}(Y|X, \mathcal{E}_0)] - [H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)]\,. \tag{47}$$

具体而言，

$$H_{\mathcal{V}}(Y|X, \mathcal{E}_0) = \mathbb{E}_{e \sim \mathcal{E}_0}\left[\inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right] \tag{48}$$

$$= \mathbb{E}_{e \sim \mathcal{E}_0}\left[\inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y}[-\log f[x](y)]\right] \tag{49}$$

$$= H_{\mathcal{V}}(Y|X)。 \tag{50}$$

类似地， $H_{\mathcal{V}}(Y|\emptyset,\ \mathcal{E}_0) = H_{\mathcal{V}}(Y|\emptyset)$。因此， $\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X \to Y) \geq 0$。

3. 有界性：
首先，我们有

$$H_{\mathcal{V}}(Y|X, \mathcal{E}) = \mathbb{E}_{e \sim \mathcal{E}}\left[\inf_{f \in \mathcal{V}} \mathbb{E}_{x,y \sim X,Y|\mathcal{E}=e}[-\log f[x](y)]\right] \tag{51}$$

$$= \mathbb{E}_{e \sim \mathcal{E}}\left[\inf_{f \in \mathcal{V}} \mathbb{E}_{x \sim X|\mathcal{E}=e}\left[\mathbb{E}_{y \sim Y|x,e}[-\log f[x](y)]\right]\right] \tag{52}$$

$$\geq 0, \tag{53}$$

注意到 $\mathbb{E}_{y \sim Y|x}[-\log\ f[x](y)]$ 是 $Y|x,\ e$ 与 $f[x]$ 之间的交叉熵。接下来，

$$H_{\mathcal{V}}(Y|\emptyset, \mathcal{E}) = \mathbb{E}_{e \sim \mathcal{E}}\left[\inf_{f \in \mathcal{V}} \mathbb{E}_{y \sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right] \tag{54}$$

$$\leq \inf_{f \in \mathcal{V}} \mathbb{E}_{e \sim \mathcal{E}}\left[\mathbb{E}_{y \sim Y|\mathcal{E}=e}[-\log f[\emptyset](y)]\right] \tag{55}$$

$$= \inf_{f \in \mathcal{V}} \mathbb{E}_{y \sim Y}[-\log f[\emptyset](y)] \tag{56}$$

$$= H_{\mathcal{V}}(Y|\emptyset), \tag{57}$$

其中，公式 55 由 Jensen不等式 得出。

综合以上不平等,

$$\mathcal{H}_{\mathcal{V}}^{\mathscr{E}}(X \to Y) = \sup_{\mathcal{E}\in\mathscr{E}} [H_{\mathcal{V}}(Y|\emptyset,\mathcal{E}) - H_{\mathcal{V}}(Y|X,\mathcal{E})] - [H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)] \tag{58}$$
$$\leq \sup_{\mathcal{E}\in\mathscr{E}} H_{\mathcal{V}}(Y|\emptyset,\mathcal{E}) - [H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)] \tag{59}$$
$$\leq H_{\mathcal{V}}(Y|\emptyset) - [H_{\mathcal{V}}(Y|\emptyset) - H_{\mathcal{V}}(Y|X)] \tag{60}$$
$$= H_{\mathcal{V}}(Y|X)。 \tag{61}$$

4. 边界情况:
根据 Xu 等人（2020）中的命题 2,

$$H_\Omega(Y|\emptyset) = H(Y). \tag{62}$$
$$H_\Omega(Y|X) = H(Y|X). \tag{63}$$

取随机变量 R、S与 X、$Y|\mathcal{E}=e$同分布，且对于 $e\in\text{supp}(\mathcal{E})$，我们有

$$H_\Omega(Y|X,\mathcal{E}=e) = H_\Omega(S|R) = H(S|R) = H(Y|X,\mathcal{E}=e). \tag{64}$$

因此,

$$H_\Omega(Y|X,\mathcal{E}) = \mathbb{E}_{e\sim\mathcal{E}}[H_\Omega(Y|X,\mathcal{E}=e)] = \mathbb{E}_{e\sim\mathcal{E}}[H(Y|X,\mathcal{E}=e)] = H(Y|X,\mathcal{E}). \tag{65}$$

同理，我们有 $H_\Omega(Y|\emptyset,\ \mathcal{E}) = H(Y|\mathcal{E})$。因此,

$$\mathcal{H}_\Omega^{\mathscr{E}}(X \to Y) = \sup_{\mathcal{E}\in\mathscr{E}} [H_\Omega(Y|\emptyset,\mathcal{E}) - H_\Omega(Y|X,\mathcal{E})] - [H_\Omega(Y|\emptyset) - H_\Omega(Y|X)] \tag{66}$$
$$= \sup_{\mathcal{E}\in\mathscr{E}} [H(Y|\mathcal{E}) - H(Y|X,\mathcal{E})] - [H(Y) - H(Y|X)] \tag{67}$$
$$= \sup_{\mathcal{E}\in\mathscr{E}} \mathbb{I}(Y;X|\mathcal{E}) - \mathbb{I}(Y;X) \tag{68}$$
$$= \mathcal{H}^{\mathscr{E}}(X,\ Y)。 \tag{69}$$

■

附录 B. 定理 7 的证明

证明 [定理 7] 的证明
1)

$$H_{\mathcal{V}_\mathcal{G}}(Y|X) = \inf_{f\in\mathcal{V}_\mathcal{G}} \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim Y|x}[-\log f[x](y)]\right] \tag{70}$$
$$\leq \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim Y|x}[-\log \frac{1}{\sqrt{2\pi}\cdot\frac{1}{\sqrt{2\pi}}} \exp\left[-\frac{(y-g(x))^2}{2\cdot\frac{1}{2\pi}}\right]\right] \tag{71}$$
$$= \mathbb{E}_{x\sim X}\left[\mathbb{E}_{y\sim Y|x}[\pi(y-g(x))^2]\right] = \pi\sigma^2. \tag{72}$$

通过取 $f[x] = \mathcal{N}(g(x)\ \frac{1}{2\pi})$，方程 71 成立。 —

2)

给定函数族 $\mathcal{V}_\sigma = \{f|f[x] = \mathcal{N}(\theta x,\ \sigma^2)\ \theta \in \mathbb{R},\ \sigma$ 固定 $\}$，通过展开预测性 $\mathcal{V}$-信息定义中的高斯概率密度函数，可以证明

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y) \propto \min_{k\in\mathbb{R}} -\mathbb{E}[(Y - kX)^2] + \mathrm{Var}(Y), \tag{73}$$

其中，预测性 $\mathcal{V}$-信息与均方误差减去目标变量方差的结果成正比，其系数完全取决于 σ。

该最小化问题的解为

$$k = \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]} = 1. \tag{74}$$

将 k 代入方程 73,

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y) \propto (-\mathbb{E}[\epsilon^2] + \mathrm{Var}(X + \epsilon)) = \mathrm{Var}(X) = \mathbb{E}[X^2]. \tag{75}$$

记 $\mathrm{supp}(\mathcal{E})$ 为 $= \{\mathcal{E}_1, \mathcal{E}_2\}$。令 Q 为 $(X,\ \epsilon,\ \mathcal{E})$ 的联合分布。令 $Q(\mathcal{E}_1)$ 为 $= \alpha,\ \ Q(\mathcal{E}_2)$ 为 $= 1 - \alpha$，二者均为 $\mathcal{E}$ 的边缘分布。将 $Q(X,\ \epsilon|\mathcal{E} = \mathcal{E}_1)$ 简记为 $P_1(X,\ \epsilon)$，将 $Q(X,\ \epsilon|\mathcal{E} = \mathcal{E}_2)$ 简记为 $P_2(X,\ \epsilon)$。

与 73 类似,

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y|\mathcal{E}) \propto \min_{k} -\mathbb{E}[(Y - kX)^2|\mathcal{E}] + \mathrm{Var}(Y|\mathcal{E}). \tag{76}$$

对于 $\mathcal{E} = \mathcal{E}_1$，最小化问题的解为

$$k = \frac{\mathbb{E}_{P_1}[XY]}{\mathbb{E}_{P_1}[X^2]}. \tag{77}$$

因此,

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y|\mathcal{E} = \mathcal{E}_1) \propto -\mathbb{E}_{P_1}\left[\left(Y - \frac{\mathbb{E}_{P_1}[XY]}{\mathbb{E}_{P_1}[X^2]}X\right)^2\right] + \mathrm{Var}_{P_1}(Y) \tag{78}$$

$$= -\mathbb{E}_{P_1}[Y^2] + \frac{\mathbb{E}^2_{P_1}[XY]}{\mathbb{E}_{P_1}[X^2]} + (\mathbb{E}_{P_1}[Y^2] - \mathbb{E}^2_{P_1}[Y]) = -\mathbb{E}^2_{P_1}[Y] + \frac{\mathbb{E}^2_{P_1}[XY]}{\mathbb{E}_{P_1}[X^2]}. \tag{79}$$

同理，我们有

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y|\mathcal{E} = \mathcal{E}_2) \propto -\mathbb{E}^2_{P_2}[Y] + \frac{\mathbb{E}^2_{P_2}[XY]}{\mathbb{E}_{P_2}[X^2]}. \tag{80}$$

值得注意的是，$\mathbb{E}_{P_1}[X^2]$ 和 $\mathbb{E}_{P_2}[X^2]$ 受 α 和 $\mathbb{E}[X^2]$的约束。

$$\mathbb{E}[X^2] = \mathbb{E}[\mathbb{E}[X^2|\mathcal{E}]] = \alpha\mathbb{E}_{P_1}[X^2] + (1 - \alpha)\mathbb{E}_{P_2}[X^2]. \tag{81}$$

同样,

$$\mathbb{E}[X^2] = \mathbb{E}[XY] = \alpha\mathbb{E}_{P_1}[XY] + (1 - \alpha)\mathbb{E}_{P_2}[XY]. \tag{82}$$

$$0 = \mathbb{E}[Y] = \alpha \mathbb{E}_{P_1}[Y] + (1-\alpha)\mathbb{E}_{P_2}[Y]. \tag{83}$$

P_2 的矩此后可由 P_1的矩表示。

$$\mathbb{E}_{P_2}[X^2] = \frac{\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[X^2]}{1-\alpha}, \mathbb{E}_{P_2}[XY] = \frac{\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[XY]}{1-\alpha}, \mathbb{E}_{P_2}[Y] = -\frac{\alpha \mathbb{E}_{P_1}[Y]}{1-\alpha}. \tag{84}$$

代入式 (80),

$$\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y | \mathcal{E} = \mathcal{E}_2) \propto -\frac{\alpha^2}{(1-\alpha)^2} E^2_{P_1}[Y] + \frac{1}{1-\alpha} \frac{\left(\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[XY]\right)^2}{\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[X^2]}. \tag{85}$$

因此,

$$\mathcal{H}^{\mathscr{E}}_{\mathcal{V}_\sigma}(X \to Y) = \sup_{\mathcal{E} \in \mathscr{E}} -\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y) + \alpha \mathbb{I}_{\mathcal{V}_\sigma}(X \to Y | \mathcal{E} = \mathcal{E}_1) + (1-\alpha)\mathbb{I}_{\mathcal{V}_\sigma}(X \to Y | \mathcal{E} = \mathcal{E}_2) \tag{86}$$

$$\propto \sup_{\mathcal{E} \in \mathscr{E}} -\mathbb{E}[X^2] - \alpha \mathbb{E}^2_{P_1}[Y] + \alpha \frac{\mathbb{E}^2_{P_1}[XY]}{\mathbb{E}_{P_1}[X^2]} - \frac{\alpha^2}{1-\alpha}\mathbb{E}^2_{P_1}[Y] + \frac{\left(\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[XY]\right)^2}{\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[X^2]} \tag{87}$$

$$= \sup_{\mathcal{E} \in \mathscr{E}} -\frac{\alpha}{1-\alpha}\mathbb{E}^2_{P_1}[X+\epsilon] + \alpha \frac{\mathbb{E}^2_{P_1}[X\epsilon]}{\mathbb{E}_{P_1}[X^2]\left(\mathbb{E}[X^2] - \alpha \mathbb{E}_{P_1}[X^2]\right)} \mathbb{E}[X^2]. \tag{88}$$

假设 $X \perp \epsilon \mid \mathcal{E}$,

$$\mathcal{H}^{\mathscr{E}}_{\mathcal{V}_\sigma}(X \to Y) \propto \sup_{\mathcal{E} \in \mathscr{E}} -\frac{\alpha}{1-\alpha}\mathbb{E}^2_{P_1}[X+\epsilon] \leq 0. \tag{89}$$

根据命题 6，我们有 $\mathcal{H}^{\mathscr{E}}_{\mathcal{V}_\sigma}(X \to Y) \geq 0$。因此， $\mathcal{H}^{\mathscr{E}}_{\mathcal{V}_\sigma}(X \to Y) = 0$。 ■

附录 C. 线性情形的证明（定理 8 和 9）

证明 [定理 8] 的证明

为方便表述，我们将 $r(\mathcal{E}^*)$ 记为 r_e，将 $\sigma(\mathcal{E}^*)$ 记为 σ_e，并将 $\sigma(\mathcal{E}^*) \cdot \epsilon_v$ 记为 $\epsilon_{e\circ}$ 此外，我们省略 $\mathcal{H}^{\mathcal{C}}_{\mathcal{V}}$ 的上标 $\mathcal{C}$。首先，我们计算 $H_{\mathcal{V}}[Y|\emptyset]$:

$$H_{\mathcal{V}}[Y|\emptyset] = \frac{1}{2\sigma^2}\mathrm{Var}(Y) + \log\sigma + \frac{1}{2}\log 2\pi, \tag{90}$$

$$H_{\mathcal{V}}[Y|\emptyset, \mathcal{E}^*] = \frac{1}{2\sigma^2}\mathbb{E}_{\mathcal{E}^*}[\mathrm{Var}(Y|\mathcal{E}^*)] + \log\sigma + \frac{1}{2}\log 2\pi. \tag{91}$$

因此，我们有

$$H_{\mathcal{V}}[Y|\emptyset, \mathcal{E}^*] - H_{\mathcal{V}}[Y|\emptyset] = -\frac{1}{2\sigma^2}\mathrm{Var}(\mathbb{E}[Y|\mathcal{E}^*]) \leq 0. \tag{92}$$

对于 $H_{\mathcal{V}}[Y|X]$，我们有

$$H_{\mathcal{V}}[Y|X] = \inf_{h_S, h_V} \mathbb{E}_{X,Y}\left[\|Y - (h_S S + h_V V)\|^2\right] \frac{1}{2\sigma^2} \tag{93}$$

$$= \inf_{h_S,h_V} \mathbb{E}_{\mathcal{E}^*}\left[\mathbb{E}[\|f(S)+\epsilon_Y-(h_S S+h_V(r_e f(S)+\epsilon_e))\|^2|\mathcal{E}^*]\right]\frac{1}{2\sigma^2}, \tag{94}$$

其中我们令 $h_S = h_S - \beta$。于是我们有

$$2\sigma^2 H_{\mathcal{V}}[Y|X] = \inf_{h_S,h_V} \mathbb{E}_{\mathcal{E}^*}\left[\mathbb{E}[\|(1-h_V r_e)f(S)+\epsilon_Y-h_S S-h_V\epsilon_e\|^2|\mathcal{E}^*]\right] \tag{95}$$

$$= \inf_{h_S,h_V} \mathbb{E}_{\mathcal{E}^*}\left[\mathbb{E}[\|(1-h_V r_e)f(S)-h_S S\|^2|\mathcal{E}^*]\right] + \sigma_Y^2 + h_V^2\mathbb{E}_{\mathcal{E}^*}[\sigma_e^2], \tag{96}$$

值得注意的是，这里对于 $e_i, e_j \in \text{supp}(\mathcal{E}^*)$，我们假设 $P^{e_i}(S,Y) = P^{e_j}(S,Y)$（我们选择这样的$\mathcal{E}^*$ 作为一种可能的划分）。而 h_S, h_V 的解为

$$h_S = \frac{\text{Var}(r_e)\mathbb{E}[f^2(S)]\mathbb{E}[f(S)S]+\mathbb{E}[\sigma_e^2]\mathbb{E}[f(S)S]}{\mathbb{E}[r_e^2]\mathbb{E}[f^2(S)]\mathbb{E}[S^2]+\mathbb{E}[\sigma_e^2]\mathbb{E}[S^2]-\mathbb{E}^2[r_e]\mathbb{E}^2[f(S)S]}, \tag{97}$$

$$h_V = \frac{\mathbb{E}[r_e](\mathbb{E}[f^2(S)]\mathbb{E}[S^2]-\mathbb{E}^2[f(S)S])}{\mathbb{E}[r_e^2]\mathbb{E}[f^2(S)]\mathbb{E}[S^2]+\mathbb{E}[\sigma_e^2]\mathbb{E}[S^2]-\mathbb{E}^2[r_e]\mathbb{E}^2[f(S)S]}. \tag{98}$$

根据假设 $\mathbb{E}[f(S)S] = 0$，我们有

$$h_S = 0, \quad h_V = \frac{\mathbb{E}[r(\mathcal{E}^*)]\mathbb{E}[f^2]}{\mathbb{E}[r^2(\mathcal{E}^*)]\mathbb{E}[f^2]+\mathbb{E}[\sigma^2(\mathcal{E}^*)]}. \tag{99}$$

因此，我们有

$$2\sigma^2 H_{\mathcal{V}}[Y|X] = \mathbb{E}_{\mathcal{E}^*}[\mathbb{E}[\|(1-h_V r_e)f(S)\|^2|\mathcal{E}^*]] + \sigma_Y^2 + h_V^2\mathbb{E}_{\mathcal{E}^*}[\sigma_e^2] \tag{100}$$

$$= \frac{\text{Var}(r_e)\mathbb{E}[f^2]+\mathbb{E}[\sigma^2(\mathcal{E}^*)]}{\mathbb{E}[r_e^2]\mathbb{E}[f^2]+\mathbb{E}[\sigma^2(\mathcal{E}^*)]}\mathbb{E}[f^2(S)] + \sigma_Y^2, \tag{101}$$

$$2\sigma^2 H_{\mathcal{V}}[Y|X,\mathcal{E}^*] = \sigma_Y^2 + \mathbb{E}[(\frac{1}{\frac{r_e^2\mathbb{E}[f^2]}{\sigma_e^2}+1})^2]\mathbb{E}[f^2] + \mathbb{E}_{\mathcal{E}^*}[(\frac{1}{\frac{r_e}{\sigma_e}+\frac{\sigma_e}{r_e\mathbb{E}[f^2]}})^2]. \tag{102}$$

注意，这里我们在正文中简单地令 $\sigma = 1$。并且有：

$$\mathcal{H}_{\mathcal{V}}(X\to Y) \approx \frac{\text{Var}(r_e)\mathbb{E}[f^2]+\mathbb{E}[\sigma^2(\mathcal{E}^*)]}{\mathbb{E}[r_e^2]\mathbb{E}[f^2]+\mathbb{E}[\sigma^2(\mathcal{E}^*)]}\mathbb{E}[f^2(S)] \tag{103}$$

近似误差由 $\frac{1}{2}\max(\sigma_Y^2, R(r(\mathcal{E}^*), \sigma(\mathcal{E}^*), \mathbb{E}[f^2]))$ 界定，且 $R(r(\mathcal{E}^*), \sigma(\mathcal{E}^*), \mathbb{E}[f^2])$ 定义如下：

$$R(r(\mathcal{E}^*),\sigma(\mathcal{E}^*),\mathbb{E}[f^2]) = \mathbb{E}[(\frac{1}{\frac{r_e^2\mathbb{E}[f^2]}{\sigma_e^2}+1})^2]\mathbb{E}[f^2] + \mathbb{E}_{\mathcal{E}^*}[(\frac{1}{\frac{r_e}{\sigma_e}+\frac{\sigma_e}{r_e\mathbb{E}[f^2]}})^2] \tag{104}$$

■

证明 [定理 9] 的证明。证明思路与上述证明类似。 ■

附录 D. 有限样本估计的误差界证明（定理 11）

本节将证明使用经验预测异质性估计预测异质性时的误差界。在证明受 Xu 等人（2020）启发的定理 11 之前，我们将介绍三个引理。

引理 14 假设 $\forall x \in \mathcal{X}$、$\forall y \in \mathcal{Y}$、$\forall f \in \mathcal{V}$、对数 $f[x](y) \in [-B, B]$，其中 $B > 0$。定义函数类 $\mathcal{G}_{\mathcal{V}}^{k} = \{g | g(x、y) = \text{对数 } f[x](y)\, q(\mathcal{E} = e_k | x、y) \quad f \in \mathcal{V}、q \in \mathcal{Q}\}$。记 $\mathcal{G}$ 在 N 个样本下的拉德马赫复杂度为 $\mathscr{R}_N(\mathcal{G})$。定义

$$\hat{f}_k = \arg\inf_{f} \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i). \tag{105}$$

那么，对于任意 $q \in \mathcal{Q}$、任意 $\delta \in (0、1)$，以关于 $1-\delta$ 的某个概率，我们有

$$\left| q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y|X, \mathcal{E} = e_k) - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log \hat{f}_k[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i) \right| \tag{106}$$

$$\leq 2\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^{k}) + B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}. \tag{107}$$

证明 将 McDiarmid 不等式应用于函数 $\Phi(\mathcal{D})$，其定义如下：

$$\Phi(\mathcal{D}) = \sup_{f \in \mathcal{V}, q \in \mathcal{Q}} \left| q(\mathcal{E} = e_k) \mathbb{E}_q \left[-\log f[x](y) | \mathcal{E} = e_k \right] - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i) \right|. \tag{108}$$

Let $\mathcal{D}$ 和 $\mathcal{D}'$ 是两个仅有一个数据点 $x_j \neq x_j'$不同的相同数据集。我们有 e:

$$\Phi(\mathcal{D}) - \Phi(\mathcal{D}') \tag{109}$$

$$\leq \sup_{f \in \mathcal{V}, q \in \mathcal{Q}} \left[\left| q(\mathcal{E} = e_k) \mathbb{E}_q \left[-\log f[x](y) | \mathcal{E} = e_k \right] - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i) \right| \right. \tag{110}$$

$$\left. - \left| q(\mathcal{E} = e_k) \mathbb{E}_q \left[-\log f[x](y) | \mathcal{E} = e_k \right] - \frac{1}{|\mathcal{D}'|} \sum_{x_i', y_i' \in \mathcal{D}'} -\log f[x_i'](y_i') q(\mathcal{E} = e_k | x_i', y_i') \right| \right] \tag{111}$$

$$\leq \sup_{f \in \mathcal{V}, q \in \mathcal{Q}} \left| \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i) - \frac{1}{|\mathcal{D}'|} \sum_{x_i', y_i' \in \mathcal{D}'} -\log f[x_i'](y_i') q(\mathcal{E} = e_k | x_i', y_i') \right| \tag{112}$$

$$= \sup_{f \in \mathcal{V}, q \in \mathcal{Q}} \frac{1}{|\mathcal{D}|} \left| \log f[x_j](y_j) q(\mathcal{E} = e_k | x_j, y_j) - \log f[x_j'](y_j') q(\mathcal{E} = e_k | x_j', y_j') \right| \tag{113}$$

$$\leq \frac{2B}{|\mathcal{D}|}. \tag{114}$$

根据 McDiarmid 不等式，对于任意 $\delta \in (0, 1)$，在 $1-\delta$ 上以一定概率成立：

$$\Phi(\mathcal{D}) \leq \mathbb{E}_{\mathcal{D}}[\Phi(\mathcal{D})] + B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}. \tag{115}$$

接下来，我们推导 $\mathbb{E}_{\mathcal{D}}[\Phi(\mathcal{D})]$ 的一个界。考虑一个数据集 $\mathcal{D}'$，它独立同分布地抽取自 $q(X, Y) = P(X, Y)$，且其大小与 $\mathcal{D}$ 相同。我们注意到

$$q(\mathcal{E}=e_k)\mathbb{E}_q\left[-\log f[x](y)|\mathcal{E}=e_k\right] = \mathbb{E}_{\mathcal{D}'}\left[-\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'} -\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')\right]. \tag{116}$$

因此，$\mathbb{E}_{\mathcal{D}}[\Phi(\mathcal{D})]$ 可以改写为：

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\Phi(\mathcal{D})] = \mathbb{E}_{\mathcal{D}}\Bigg[\sup_{f\in\mathcal{V},q\in\mathcal{Q}}\Bigg|\mathbb{E}_{\mathcal{D}'}\left[-\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'} -\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')\right] & \tag{117}\\
-\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i)\Bigg|\Bigg] & \tag{118}\\
\leq \mathbb{E}_{\mathcal{D}}\Bigg[\sup_{f\in\mathcal{V},q\in\mathcal{Q}}\mathbb{E}_{\mathcal{D}'}\Bigg|-\frac{1}{|\mathcal{D}'|}\sum_{x_i',y_i'\in\mathcal{D}'} -\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i') & \tag{119}\\
-\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i)\Bigg|\Bigg] & \tag{120}\\
\leq \mathbb{E}_{\mathcal{D},\mathcal{D}'}\Bigg[\sup_{f\in\mathcal{V},q\in\mathcal{Q}}\frac{1}{|\mathcal{D}|}\Bigg|\sum_{x_i,y_i\in\mathcal{D}}\log f[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i) & \tag{121}\\
-\sum_{x_i',y_i'\in\mathcal{D}'}\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')\Bigg|\Bigg] & \tag{122}\\
-\sum_{x_i',y_i'\in\mathcal{D}'}\sigma_i\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')\Bigg|\Bigg] & \tag{123}\\
\leq \mathbb{E}_{\mathcal{D},\sigma}\Bigg[\sup_{f\in\mathcal{V},q\in\mathcal{Q}}\frac{1}{|\mathcal{D}|}\Bigg|\sum_{x_i,y_i\in\mathcal{D}}\sigma_i\log f[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i)\Bigg|\Bigg] & \tag{124}\\
+\mathbb{E}_{\mathcal{D}',\sigma}\Bigg[\sup_{f\in\mathcal{V},q\in\mathcal{Q}}\frac{1}{|\mathcal{D}'|}\Bigg|\sum_{x_i',y_i'\in\mathcal{D}'}\sigma_i\log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')\Bigg|\Bigg] & \tag{125}\\
= 2\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^k), & \tag{126}
\end{aligned}$$

其中，σ_i 是相互独立的拉德马赫变量。根据詹森不等式和 sup 的凸性，可得方程 121。由于 $\log f[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i) - \log f[x_i'](y_i')q(\mathcal{E}=e_k|x_i',y_i')$ 的对称性，以及拉德马赫变量在凸映射下保持对称随机变量期望和的论证，方程 123 成立（Ledoux 和 Talagrand（1991），引理 6.3）。

将方程 126 代入方程 115，对于任意 $\delta \in (0,1)$，关于 $1-\delta$、 $\forall f \in \mathcal{V}$、 $\forall q \in \mathcal{Q}$ 的概率，以下结论成立：

$$\left| q(\mathcal{E}=e_k)\mathbb{E}_q\left[-\log f[x](y)|\mathcal{E}=e_k\right] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E}=e_k|x_i,y_i) \right| \tag{127}$$

$$\leq 2\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^k) + B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}. \tag{128}$$

令 $\tilde{f}_k = \arg\inf_f \{q(\mathcal{E}=e_k)\mathbb{E}_q\left[-\log f[x](y)|\mathcal{E}=e_k\right]\}$。
令 $\hat{f}_k = \arg\inf_f \{\frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E}=e_k|x_i, y_i)\}$。现在我们有

$$q(\mathcal{E}=e_k)\mathbb{E}_q\left[-\log \tilde{f}_k[x](y)|\mathcal{E}=e_k\right] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \tilde{f}_k[x_i](y_i) q(\mathcal{E}=e_k|x_i,y_i) \tag{129}$$

$$\leq q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|X,\mathcal{E}=e_k) - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \hat{f}_k[x_i](y_i) q(\mathcal{E}=e_k|x_i,y_i) \tag{130}$$

$$\leq q(\mathcal{E}=e_k)\mathbb{E}_q\left[-\log \hat{f}_k[x](y)|\mathcal{E}=e_k\right] - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \hat{f}_k[x_i](y_i) q(\mathcal{E}=e_k|x_i,y_i). \tag{131}$$

结合方程 127 和方程 129-131，命题得证。 ■

引理15 假设 $\forall x \in \mathcal{X}, \forall y \in \mathcal{Y}, \forall f \in \mathcal{V}, \log f[\emptyset](y) \in [-B, B]$，其中 $B > 0$。$\mathcal{G}_{\mathcal{V}}^k$ 和 $\mathscr{R}_N(\mathcal{G})$ 的定义由引理 *14* 得出。定义 $\hat{f}_k = \arg\inf_f \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log f[\emptyset](y_i) q(\mathcal{E}=e_k|x_i, y_i)$。

那么，对于任意 $q \in \mathcal{Q}$ 和任意 $\delta \in (0,1)$，在关于 $1-\delta$ 的概率下，我们有

$$\left| q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|\mathcal{E}=e_k) - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \hat{f}_k[\emptyset](y_i) q(\mathcal{E}=e_k|x_i,y_i) \right| \tag{132}$$

$$\leq 2\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^k) + B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}. \tag{133}$$

证明 与引理 14 类似，我们可以证明

$$\left| q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|\mathcal{E}=e_k) - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log \hat{f}_k[\emptyset](y_i) q(\mathcal{E}=e_k|x_i,y_i) \right| \tag{134}$$

$$\leq 2\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}^\emptyset}^k) + B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}, \tag{135}$$

其中 $\mathcal{G}_{\mathcal{V}^\emptyset}^k = \{g | g(x, y) = \log f[\emptyset](y) q(\mathcal{E}=e_k|x, y)\ f \in \mathcal{V},\ q \in \mathcal{Q}\}$。

根据预测族的定义 $\mathcal{V}$ (Xu 等人（2020年），定义 1)，$\forall f \in \mathcal{V}$，存在 $f' \in \mathcal{V}$ 使得 $\forall x \in \mathcal{X}$， $f[\emptyset] = f'[x]$。因此，$\mathcal{G}^k_{\mathcal{V}^\emptyset} \subset \mathcal{G}^k_{\mathcal{V}}$，进而$\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}^k_{\mathcal{V}^\emptyset}) \leq \mathscr{R}_{|\mathcal{D}|}(\mathcal{G}^k_{\mathcal{V}})$。代入公式 134，即证明了该引理。 ■

引理 16（（Xu 等人，2020），定理 1）假设 $\forall x \in \mathcal{X}$,$\forall y \in \mathcal{Y}$,$\forall f \in \mathcal{V}$, 对数 $f[x](y) \in [-B, B]$ 其中 $B > 0$。定义函数类 $\mathcal{G}^*_{\mathcal{V}} = \{g | g(x, y) = \text{对数 } f[x](y) \ f \in \mathcal{V}\}$。 $\mathscr{R}_N(\mathcal{G})$ 的定义见引理 *14*。

那么，对于任意 $\delta \in (0, 0.5)$，以关于 $1 - 2\delta$ 的概率，我们有

$$\left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y) \right| \leq 4\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}^*_{\mathcal{V}}) + 2B\sqrt{\frac{2 \log \frac{1}{\delta}}{|\mathcal{D}|}}. \tag{136}$$

最后，我们可以证明定理 11。

证明 [定理 11] 的证明我们首先利用引理 14、15、16 中各项之和，界定经验估计的误差。

$$\begin{aligned}
&|\mathcal{H}^{\mathscr{E}_K}_{\mathcal{V}}(X \to Y) - \hat{H}^{\mathscr{E}_K}_{\mathcal{V}}(X \to Y; \mathcal{D})| && (137)\\
\leq &\left| \sup_{\mathcal{E} \in \mathscr{E}_K} \mathbb{I}_{\mathcal{V}}(X \to Y | \mathcal{E}) - \sup_{\mathcal{E} \in \mathscr{E}_K} \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y | \mathcal{E}; \mathcal{D}) \right| + \left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right| && (138)\\
\leq &\sup_{\mathcal{E} \in \mathscr{E}_K} \left| \mathbb{I}_{\mathcal{V}}(X \to Y | \mathcal{E}) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y | \mathcal{E}; \mathcal{D}) \right| + \left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right| && (139)\\
= &\sup_{q \in \mathcal{Q}} \left| \sum_{k=1}^{K} [q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | \mathcal{E} = e_k) - q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | X, \mathcal{E} = e_k)] \right. && (140)\\
&\left. - \sum_{k=1}^{K} \left[q(\mathcal{E} = e_k) \hat{H}_{\mathcal{V}}(Y | \mathcal{E} = e_k; \mathcal{D}) - q(\mathcal{E} = e_k) \hat{H}_{\mathcal{V}}(Y | X, \mathcal{E} = e_k; \mathcal{D}) \right] \right| && (141)\\
&+ \left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right| && (142)\\
\leq &\sum_{k=1}^{K} \sup_{q \in \mathcal{Q}} \left| q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | \mathcal{E} = e_k) - q(\mathcal{E} = e_k) \hat{H}_{\mathcal{V}}(Y | \mathcal{E} = e_k; \mathcal{D}) \right| && (143)\\
&+ \sum_{k=1}^{K} \sup_{q \in \mathcal{Q}} \left| q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | X, \mathcal{E} = e_k) - q(\mathcal{E} = e_k) \hat{H}_{\mathcal{V}}(Y | X, \mathcal{E} = e_k; \mathcal{D}) \right| && (144)\\
&+ \left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right| && (145)\\
= &\sum_{k=1}^{K} \sup_{q \in \mathcal{Q}} \left| q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | \mathcal{E} = e_k) - \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\text{对数 } \hat{f}_k[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i) \right| && (146)\\
&+ \sum_{k=1}^{K} \sup_{q \in \mathcal{Q}} \left| q(\mathcal{E} = e_k) H_{\mathcal{V}}(Y | X, \ \mathcal{E} = e_k) - \frac{1}{|\mathcal{D}|} \sum_{x_i, \ y_i \in \mathcal{D}} -\text{对数 } \hat{f}'_k[\emptyset](y_i) q(\mathcal{E} = e_k | x_i, y_i) \right| && (147)\\
&+ \left| \mathbb{I}_{\mathcal{V}}(X \to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X \to Y; \mathcal{D}) \right|, && (148)
\end{aligned}$$

其中 $\hat{f}_k = \arg\inf_f \frac{1}{|\mathcal{D}|} \sum_{x_i, y_i \in \mathcal{D}} -\log f[x_i](y_i) q(\mathcal{E} = e_k | x_i, y_i)$,

并且 ˆ

$$f'_k = \quad f\ _{|\mathcal{D}|}{}^{1}\overline{\sum}_{x_i\ y_i\in\mathcal{D}} - \quad f[\emptyset]\ y_i\ q\ \mathcal{E}=e_k|x_i\ y_i \qquad q\in\mathcal{Q} \qquad 1\le k\le \arg$$

K
inf,log ( ) (,)，对任意 和 。为简便起见，令

$$\mathrm{Err}_k = \sup_{q\in\mathcal{Q}}\left|q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|X,\mathcal{E}=e_k) - \frac{1}{|\mathcal{D}|}\sum_{x_i,y_i\in\mathcal{D}} -\log\hat{f}_k[x_i](y_i)q(\mathcal{E}=e_k|x_i,y_i)\right|. \tag{149}$$

$$\text{误差}'_k = \text{上确界}_{q\in\mathcal{Q}}\left|q(\mathcal{E}=e_k)H_{\mathcal{V}}(Y|X,\ \ \mathcal{E}=e_k) - \frac{1}{|\mathcal{D}|}\sum_{x_i y_i\in\mathcal{D}} {}^{-\text{对数}}\ \hat{f}'_k[\emptyset](y_i)q(\mathcal{E}=e_k|x_i,\ \ y_i)\right|. \tag{150}$$

$$\text{误差}^* = \left|\mathbb{I}_{\mathcal{V}}(X\to Y) - \hat{\mathbb{I}}_{\mathcal{V}}(X\to Y;\ \mathcal{D})\right|. \tag{151}$$

于是，根据引理 14,15,16,

$$\Pr\left[|\mathcal{H}_K^{\mathcal{V}} - \hat{\mathcal{H}}_K^{\mathcal{V}}(\mathcal{D})| > 4(K+1)\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}) + 2(K+1)B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right] \tag{152}$$

$$\le \Pr\left[\sum_{i=1}^{K}\mathrm{Err}_k + \sum_{i=1}^{K}\mathrm{Err}'_k + \mathrm{Err}^* > 4(K+1)\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}) + 2(K+1)B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right] \tag{153}$$

$$\le \Pr\left[\sum_{i=1}^{K}\mathrm{Err}_k + \sum_{i=1}^{K}\mathrm{Err}'_k + \mathrm{Err}^* > \sum_{k=1}^{K}4\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^k) + 4\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^*) + 2(K+1)B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right] \tag{154}$$

$$+\left(\mathrm{Err}^* > 4\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^*) + 2B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right)\Bigg] \tag{155}$$

$$+\Pr\left[\mathrm{Err}^* > 4\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^*) + 2B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right] \tag{156}$$

$$\le 2(K+1)\delta. \tag{157}$$

公式154是因为 $\mathcal{G}_{\mathcal{V}}^k = \mathcal{G}_{\mathcal{V}}$、 $\mathcal{G}_{\mathcal{V}}^* \subset \mathcal{G}_{\mathcal{V}}$ ，因此 $R_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^k) \le R_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}})$， $R_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}^*)$ $\le R_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}})$。因此，

$$\Pr\left[|\mathcal{H}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y) - \hat{H}_{\mathcal{V}}^{\mathscr{E}_K}(X\to Y;\mathcal{D})| \le 4(K+1)\mathscr{R}_{|\mathcal{D}|}(\mathcal{G}_{\mathcal{V}}) + 2(K+1)B\sqrt{\frac{2\log\frac{1}{\delta}}{|\mathcal{D}|}}\right] \tag{158}$$

$$\ge 1 - 2(K+1)\delta. \tag{159}$$

■

附录E. 定理12的证明

证明 [定理的证明 12]我们的 IM 算法的目标函数直接源自定义 10 中经验预测异质性的定义。对于回归任务，我们将预测族假设为

$$\mathcal{V}_1 = \{g\colon\ g[x] = \mathcal{N}(f_\theta(x)\ \sigma^2)\ f\text{是回归模型，且 }\theta\text{是可学习的，}\ \sigma = 1.0\text{（固定）}\}, \tag{160}$$

其中我们只关注模型的输出，而高斯分布的噪声尺度通常会被忽略，因此我们简单地将 $\sigma = 1.0$ 设置为固定项。然后，对于每个环境 $e \in \text{supp}(\mathcal{E}^*)$，$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}^* = e)$ 变为

$$\mathbb{I}_{\mathcal{V}}(X \to Y|\mathcal{E}^* = e) \propto \min_{\theta} \mathbb{E}^{[}\|Y - f_\theta(X)\|^2|\mathcal{E}^* = e] - \text{Var}(Y|\mathcal{E}^*), \tag{161}$$

这与均方误差损失以及公式 35 中提出的正则化项相对应。对于分类任务，推导过程类似，正则化项变为子群体 Y 中的熵，而损失函数变为交叉熵损失。

■

参考文献

Martin Arjovsky、L´eon Bottou、Ishaan Gulrajani 和 David Lopez-Paz。不变风险最小化。*arXiv* 预印本 *arXiv:1907.02893*，2019年。

Peter L Bartlett 和 Shahar Mendelson。拉德马赫复杂度与高斯复杂度：风险界和结构性结果。*JMLR*，3（11月）：463–482，2002年。SM Chan、LD Ying、JM Morton 和 S Ghiassi。新冠肺炎模拟外科急腹症的胃肠道表现。美国外科医师学会，2022年。M Cover Thomas 和 A Thomas Joy。信息论基础。纽约：*Wiley*，3：37–38，1991年。

Elliot Creager、J¨orn-Henrik Jacobsen 和 Richard Zemel。用于不变学习的环境推断。载于 *ICML*，第 2189–2200 页。PMLR，2021年。

Peng Cui 和 Susan Athey。稳定学习在因果推断与机器学习之间建立了某种共同基础。*Nature Machine Intelligence*，4(2):110–115，2022年。

John Duchi、Tatsunori Hashimoto 和 HongseokNamkoong。针对潜在协变量混合的分布鲁棒损失。*Operations Research*，2022年。

Jianqing Fan、Fang Han 和 Han Liu。大数据分析的挑战。*National sciencereview*，1(2):293–314，2014年。

Subash Ghimire、Sachit Sharma、Achint Patel、Rasmita Budhathoki、Raja Chakinala、Hafiz Khan、Matthew Lincoln 和 Michael Georgeston。腹泻与新冠肺炎患者病情严重程度增加有关：系统性综述与荟萃分析。*SN ComprehensiveClinical Medicine*，第 1–8 页，2021年。

Moritz Hardt、Eric Price 和 Nati Srebro。监督学习中的机会平等。神经信息处理系统进展，29，2016。James J Heckman。将样本选择偏差视为设定误差。*Econometrica*：计量经济学会期刊，第 153–161 页，1979。Sumyea Helal。子群体发现算法：综述与实证评估。*Journal of Computer Science and Technology*，31(3):561–576，2016。

Miguel A Hern´an、David Clayton 和 Niels Keiding。《辛普森悖论揭秘》。《国际流行病学杂志》，40(3):780–785，2011年。R. Kohavi 和 B. Becker。《Adult》，1996年。Masanori Koyama 和 Shoichiro Yamaguchi。利用最大不变预测器实现分布外泛化。2020年。Michel Ledoux 和 Michel Talagrand。《巴拿赫空间中的概率：等周问题与过程》，第23卷。施普林格科学与商业媒体，1991年。H Li 和 JF Reynolds。异质性的定义与量化。《*Oikos*》，第280–284页，1995年。Jiashuo Liu、Zheyuan Hu、Peng Cui、Bo Li 和 Zheyan Shen。异质风险最小化。载于*ICML*，第6804–6814页。PMLR，2021a。Jiashuo Liu、Zheyan Shen、Peng Cui、周林军、匡坤和 Bo Li。采用稳定对抗训练的分布鲁棒学习。《*IEEE*知识与数据工程汇刊》，2022年。刘珏、陶立元、刘霞、姚红艳、余世成、王琪琪、张娇娇、高占成、姜荣猛、景文展和刘敏。胃肠道症状和发热会增加新冠肺炎患者重症和死亡的风险。《*Gut*》，70(2):442–444，2021b。ISSN 0017-5749。doi: 10.1136/gutjnl-2020-321751。URL https://gut.bmj.com/content/70/2/442。David B Lobell、Marshall B Burke、Claudia Tebaldi、Michael D Mastrandrea、Walter P Falcon 和 Rosamond L Naylor。2030年粮食安全气候变化适应需求的优先排序。《科学》，319(5863):607–610，2008年。Todd K Moon。期望最大化算法。《*IEEE*信号处理杂志》，13(6):47–60，1996年。Jonas Peters、Peter B¨uhlmann 和 Nicolai Meinshausen。利用不变预测进行因果推断：识别与置信区间。《英国皇家统计学会会刊*B*辑（统计方法学）》，78(5):947–1012，2016年。Douglas A Reynolds。高斯混合模型。《生物识别百科全书》，741(659-663)，2009年。Michael Roberts、Derek Driggs、Matthew Thorpe、Julian Gilbey、Michael Yeung、Stephan Ursprung、Angelica I Aviles-Rivero、Christian Etmann、Cathal McCague、Lucian Beer 等。使用机器学习通过胸部X线片和CT扫描检测新冠肺炎并预测其预后时的常见陷阱与建议。《自然机器智能》，3(3):199–217，2021年。Paul R Rosenbaum。异质性与因果性：观察性研究中的个体异质性与设计敏感性。《美国统计学家》，59(2):147–152，2005年。

Shiori Sagawa、Pang Wei Koh、Tatsunori B Hashimoto 和 Percy Liang。面向群体偏移的分布鲁棒神经网络：正则化对最坏情况泛化的重要性。*arXiv*预印本 *arXiv:1911.08731*，2019年。Amirreza Shaban、Ching-An Cheng、Nathan Hatch 和 Byron Boots。用于双层优化的截断反向传播。载于第*22*届国际人工智能与统计会议，第1723–1732页。PMLR，2019年。Zheyan Shen、Jiashuo Liu、Yue He、Xingxuan Zhang、Renzhe Xu、Han Yu 和 Peng Cui。迈向分布外泛化：综述。2021年。E. Tipton、J. Spybrook、K. G. Fitzgerald、Q. Wang 和 C. Davidson。迈向人人共享的证据体系：37项随机试验的当前实践与未来机遇。《教育研究者》，第0013189X2096068页，2020年。Shikha Verma。《数学杀伤性武器：大数据如何加剧不平等并威胁民主》。*Vikalpa*，44(2):97–98，2019年。Clifford H Wagner。现实生活中的辛普森悖论。《美国统计学家》，36(1):46–48，1982年。Xiao Wang、Houye Ji、Chuan Shi、Bai Wang、Yanfang Ye、Peng Cui 和 Philip S Yu。异质图注意力网络。载于*WWW*，第2022–2032页，2019年。Christopher Winship 和 Robert D Mare。样本选择偏差模型。《社会学年度评论》，第327–350页，1992年。Laure Wynants、Ben Van Calster、Gary S Collins、Richard D Riley、Georg Heinze、Ewoud Schuit、Marc MJ Bonten、Darren L Dahly、Johanna A Damen、Thomas PA Debray 等。新冠肺炎诊断与预后的预测模型：系统综述与批判性评估。《英国医学杂志》，369，2020年。Yilun Xu、Shengjia Zhao、Jiaming Song、Russell Stewart 和 Stefano Ermon。计算约束下可用信息的理论。载于*ICLR*，2020年。W. Zeng、K. Qi、M. Ye、L. Zheng 和 Y. Wei。胃肠道症状与2019冠状病毒病严重程度相关：系统综述与荟萃分析。《欧洲胃肠病学与肝病学杂志》，网络先行发表，2021年。Shengjia Zhao、Abhishek Sinha、Yutong He、Aidan Perreault、Jiaming Song 和 Stefano Ermon。通过衡量影响决策的差异来比较分布。载于国际学习表征会议，2021年。Peijie Zhong、Jing Xu、Dong Yang、Yue S hen 和 Yangbai Sun。新冠肺炎相关胃肠道和肝脏损伤：临床特征与潜在机制。《信号转导与靶向治疗》，5(1):256，2020年。Xiao Zhou、Yong Lin、Renjie Pi、Weizhong Zhang、Renzhe Xu、Peng Cui 和 Tong Zhang。用于分布外学习的模型无关样本重加权。载于国际机器学习会议，第27203–27221页。PMLR，2022年。