使用多项逻辑回归估计高差异分布之间的密度比

Akash Srivastava* *akashsri@mit.edu*
MIT-IBM Watson 人工智能实验室和 IBM 研究院

Seungwook Han* *swhan@mit.edu*
MIT

Kai Xu *xuk@amazon.com*
亚马逊

Benjamin Rhodes *ben.rhodes@ed.ac.uk*
信息学院，爱丁堡大学

Michael U. Gutmann* *michael.gutmann@ed.ac.uk*
信息学院，爱丁堡大学

已在 OpenReview 上评审： *https://openreview.net/forum?id=jM8nzUzBWr*

arXiv:2305.00869v1 [stat.ML] 1 May 2023

摘要

密度比的函数 p/q 广泛应用于机器学习中，用于量化两个分布 p 和 q之间的差异。对于高维分布，基于二元分类的密度比估计器展现出了巨大潜力。然而，当密度充分分离时，使用二元分类器估计密度比颇具挑战。在本研究中，我们表明，最先进的密度比估计器在充分分离 的情况下表现不佳，并证明这是由于训练时与评估时之间存在分布偏移所导致的。我们提出了一种利用多分类进行密度比估计的替代方法，且不受分布偏移问题的影响。该方法使用一组辅助密度 $\{m_k\}_{k=1}^K$ ，并训练多分类逻辑回归，将来自 p, q 和 $\{m_k\}_{k=1}^K$ 的样本分为 $K+2$ 类。我们表明，如果构造这些辅助密度，使其分别与 p 和 q重叠，那么多分类逻辑回归便可在任意一个 $K+2$ 分布的定义域上估计 $\log p/q$ ，从而解决当前最先进方法中的分布偏移问题。我们在合成数据集和真实数据集上将该方法与最先进的密度比估计器进行了比较，并证明了该方法在密度比估计、互信息估计和表征学习任务上的优越性能。代码：https://www.blackswhan.com/mdre/

1 引言

量化两个分布之间的差异，是许多机器学习技术的基础。例如，称为 f-散度（Csiszár，1964）的分布差异度量，被定义为两个概率密度之比的凸函数的期望，在监督和无监督机器学习的许多领域中都十分常见。因此，密度比估计往往是生成建模、互信息与散度估计以及表征学习中的核心任务（Sugiyama 等人，2012；Gutmann 和 Hyvärinen，2010；Goodfellow 等人，2014；Nowozin 等人，2016；Srivastava 等人，2017；Belghazi 等人，2018；Oord 等人，2018；Srivastava 等人，2020 年）。然而，在大多数问题中

*同等贡献

对于所关注的问题，在高维密度情形下，分别对每个概率密度建模来估计密度比，远比直接估计其比值更具挑战性（Sugiyama 等人，2012）。因此，实践中通常采用直接密度比估计器。

最常用的密度比估计器（DRE）之一是通过逻辑回归进行二元分类（BDRE）。在训练其区分来自两个概率密度的样本后，已有研究表明，BDRE估计器能够估计这两个概率密度之间的真实密度比（例如，Gutmann 和 Hyväri- nen，2010；Gutmann 和 Hirayama，2011；Sugiyama 等人，2012；Menon 和 Ong，2016 ）。即使在高维场景中，BDRE 估计器在涉及最小化基于密度比的估计器、以衡量数据分布与模型分布之间差异的问题上也取得了极大成功（Nowozin 等人，2016；Radford 等人，2015）。然而，当用于估计彼此相距甚远或易于区分的两个分布之间的差异时，它们的表现就不尽如人意。Rhodes 等人（2020）最近将这一问题称为密度鸿沟问题。我们在图1中展示了这一现象：使用 BDRE 估计图（a）所示两个一维分布 $p = \mathrm{N}(-1, 0.1)$ 和 $q = \mathrm{N}(1, 0.2)$ 之间的密度比。由于 p 和 q 相距很远，如 BDRE 决策边界的可视化结果所示，分类问题相对容易解决。然而，如图（b）所示，即使在这个简单的设置中，BDRE 也完全无法估计该比率。Kato 和 Teshima（2021）也证实，大多数密度比估计器，尤其是采用深度神经网络实现的估计器，在面对密度鸿沟问题时往往会以某种方式对训练数据过拟合。由于基于 BDRE 的插件估计器常用于许多高维任务，例如互信息估计、表征学习、基于能量的建模、协变量偏移校正和重要性采样（Rhodes 等人，2020；Choi 等人，2021b;a；Sugiyama 等人，2012），解决密度鸿沟问题具有重要的实际意义。

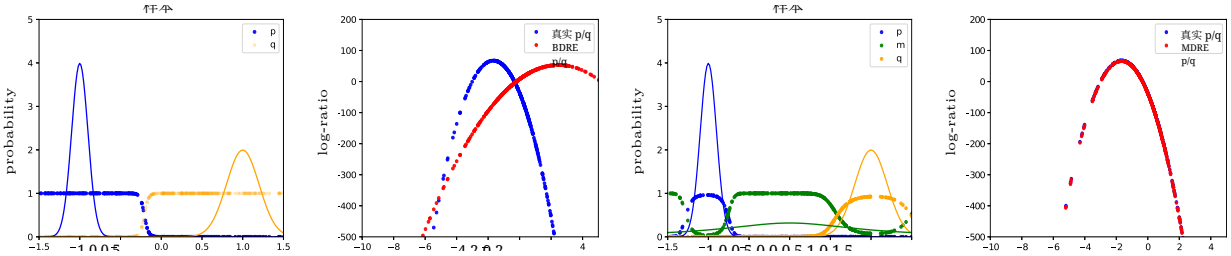
近期提出的一种解决密度鸿沟问题的方法——望远镜式密度比估计（TRE；Rhodes 等人，2020 年）通过将原本较易分类的逻辑回归问题替换为一组较难分类的逻辑回归问题来应对这一问题。简而言之，TRE 构造一组 K 辅助分布（$\{m_k\}_{k=1}^K$），以衔接两个感兴趣的目标分布（$p =: m_0$ 和 $q =: m_{K+1}$），然后在每对 $K+1$ 连续分布（）（m_{k-1}和 m_k之间（ $k = 1$、.....、 K）训练一个 BDRE估计器；假设这些分布足够接近（即不易区分），从而使 BDRE估计器能够发挥良好效果。之后，将所有单独估计值累乘（望远镜式乘积），即可得到总体概率密度比估计值。

在这项工作中，我们指出，上述解决密度鸿沟问题的方法存在分布偏移这一固有问题，这可能导致最终的密度比估计出现显著误差。注意，TRE 构造的 BDRE 链中第 i个 BDRE 仅使用来自分布m_i 和 m_{i+1}的样本进行训练。然而，训练完成后，它通常会在原始密度比估计问题中的分布（即 p 和 q）具有不可忽略概率质量的区域上进行评估。如果 p、 q 以及辅助分布 m_i 的高概率区域互不重叠，那么第 i个 BDRE 的训练分布与评估分布就会不同。由于训练与评估之间存在这种分布偏移，总体密度比估计可能会出现误差（更多详情见图2和第2.1节）。在此，我们提出另一种解决密度鸿沟问题的方法，以避免这种分布偏移。

我们提出了基于多项逻辑回归的密度比估计器（MDRE），这是一种用于密度比估计的新方法，能够解决密度鸿沟问题，且不会受到分布偏移的影响。该方法通过使用辅助分布和多分类实现。MDRE 将简单的二元分类问题替换为单个更难的多分类问题。MDRE 首先构造一组与 p 和 q 重叠的 K 辅助分布$\{m_k\}_{k=1}^K$，然后在这些 $K+2$ 分布上使用多分类逻辑回归，以获得 $\log p/q$ 的密度比估计器。我们将证明，多分类问题的建模方式能够避免 TRE 的分布偏移问题。

本研究的主要贡献如下：

1. 我们研究了密度鸿沟问题的最先进解决方案（TRE；Rhodes 等人，2020 年），并指出其由分布偏移导致的局限性。我们说明，这一固有问题会显著降低其密度比估计性能。



(a) BDRE 类别概率　(b) BDRE 对数密度比　(c) MDRE 类别概率　（d）MDRE 对数密度比

图1：BDRE 与所提出的 MDRE 在密度比对数估计上的比较，其中 $p = \mathcal{N}(-1, 0, 1)$ 和 $q = \mathcal{N}(1, 0, 2)$，对于 MDRE，辅助分布 m_1 为柯西分布 $C(0, 1)$，图 (a) 和 (c) 分别展示了 BDRE 和 MDRE 学习得到的类别概率 $P(Y|x)$，并将其叠加在 p、q 和 m_1 的图上。图 (b) 和 (d) 分别展示了 BDRE 和 MDRE 估计的对数密度比。使用辅助分布 m_1 使 MDRE 能够更准确地估计对数密度比。

2. 我们正式建立了多项逻辑回归与密度比估计之间的联系，并提出了一种新方法（MDRE），利用辅助分布训练多分类分类器，以进行密度比估计。MDRE 通过设计解决了上述分布偏移问题，并有效应对了密度鸿沟问题。

3. 我们构建了一个全面的评估方案，在很大程度上扩展了以往工作中使用的基准。我们对所提出的方法进行了系统性的实证评估，并在多个合成数据集和真实数据集上证明了我们方法的优越性能。结果表明，在 f-散度估计、互信息估计以及高维环境下的表示学习等任务中，MDRE 通常显著优于当前最先进水平的密度比估计方法。

2 相关工作

望远镜式密度比估计（TRE，Rhodes 等人，2020 年）采用两步分而治之策略来解决密度鸿沟问题。第一步，他们通过逐步将来自 p 的样本迁移至来自 q 的样本，构造 K 路径标记分布 $\{m_k\}_{k=1}^{K}$。然后，他们针对每一对相邻分布分别训练 K BDRE估计器。这样便可以将比率 $r_{p/q}$ 估计为 $K+1$BDRE估计器的乘积，$r_{p/q} := \frac{p}{q} = \frac{p}{m_1} \times \cdots \times \frac{m_K}{q}$。Rhodes 等人（2020）提出了两种构造路径标记分布的方案，以确保相邻分布足够接近，从而避免任何 $K+1$ BDRE估计器遭遇密度鸿沟问题。因此，TRE 通过将 p 与 q 之间的比率替换为 $K+1$ 中间密度比的乘积来解决密度鸿沟问题；根据路径标记分布的设计，这些中间密度比不应受到密度鸿沟问题的影响。在一项新研究中，Choi 等人（2021b）提出了 DRE-∞，该方法将 TRE 中路径标记分布的数量取至无穷，并推导出一个极限目标函数，从而得到一个可扩展性更强的 TRE 版本。

F-DRE 是另一项有趣的相关工作，源自 Choi 等人 (2021a)。F-DRE 使用一种基于流的模型（Rezende 和 Mohamed，2015），该模型经过训练，可将来自两个分布混合的样本投影到标准高斯分布上。随后，他们训练一个块对角密度比估计器（BDRE）。不难证明，任何双射映射都会保留原始密度比$r_{p/q}$ 在特征空间中的值，因为雅可比校正项会直接抵消。然而，由于流映射具有双射性，这种方法无法使投影后的分布比原始分布之间的差异更接近。该方法至多只能在投影后，将原始分布之间的差异转移到不同的矩上。由于这一问题，我们发现 F-DRE 在我们所考虑的问题上效果不佳（见第 4 节的实验结果）。最近，Liu 等人 (2021) 通过 (a) 使用归一化梯度下降，以及 (b) 用指数损失替代逻辑损失，提出了一种基于优化的方法来解决指数族分布中的密度鸿沟问题。最后，虽然块对角密度比估计器（BDRE）仍然是密度比估计的主流方法

近期文献中的估计方法、先前的工作，例如 Bickel 等人(2008）和Nocketal.(2016），研究了基于多类分类器的密度比估计，用于估计一组概率密度相对于共同参考分布的比率，以及其在多任务学习中的应用。

2.1 TRE 的性能可能因训练分布与评估分布偏移而下降

在监督学习中，当训练数据（x, y） $\sim p_{\text{train}}$ 和测试数据（x, y） $\sim p_{\text{test}}$ 来自两个不同的分布时，即 $p_{\text{train}} \neq p_{\text{test}}$，就会出现分布偏移（Quiñonero-Candela 等人，2009）。常见的训练方法（例如 BDRE 中使用的方法）只能保证模型在来自与 p_{train} 相同分布的未见数据上表现良好。因此，在测试时发生分布偏移的情况下，模型性能会随偏移程度而下降。现在我们将说明，当分布 p 和 q 之间存在足够大的差异时，TRE 中也会出现类似的分布偏移。回顾一下，在 TRE 中，我们使用 BDRE 估计器来估计 $K+1$ 个密度比 $p/m_1, m_2/m_1, \ldots, m_K/q$，并将其组合成一个望远镜乘积，从而得到总体比率 p/q。我们将这些 $K+1$ 个比率的估计值记作 $\hat{}\eta_1, \ldots, \hat{}\eta_{K+1}$。

根据 BDRE 的理论性质，对于任意 $i \in \{1, \ldots, K+1\}$，$\hat{}\eta_i$ 会在 m_i 的支撑集上估计 r_{m_{i-1}/m_i}（Sugiyama 等人，2012；Gutmann 和 Hyvärinen，2010；Menon 和 Ong，2016）。然而，在 TRE 中，当我们在 p 和 q的支撑集上评估目标比率 p/q 时，实际上是在一些缺乏性能良好保证的定义域上评估各个 $\hat{}\eta_i$。由于 $p/q \approx \hat{}\eta_1 * \cdots * \hat{}\eta_{K+1}$ 的整体估计器由多个比率估计器组合而成，只要其中任一估计器的性能下降，就会受到分布偏移问题的影响。因此，如果 $\{m_i\}_{i=1}^K$、 p和 q 的支撑集不同，或者来自 $\{m_i\}_{i=1}^K$、 p和 q 的样本重叠不够充分，那么 $\hat{}\eta_i$ 的训练定义域与评估定义域就不同，我们预计比率估计 $\hat{}\eta_i$ 以及由此得到的 p/q 整体估计器会表现不佳。下面通过一个玩具示例来说明这一点。

我们考虑估计 $p = \mathrm{N}(-1, 0.1)$ 与 $q = \mathrm{N}(1, 0.2)$ 之间的密度比。由于 p 和 q 相距较远，我们引入三个辅助分布 m_1、m_2、m_3 来连接它们，为 TRE 提供所需的中间分布。辅助分布 m_1、m_2、m_3 是通过线性混合策略构造的，具体将在第3.2节介绍。此设置如图2左上面板所示。我们训练了 4 个 BDRE估计器 $\hat{}\eta_1$、$\hat{}\eta_2$、$\hat{}\eta_3$、$\hat{}\eta_4$，分别估计比率 p/m_1、m_1/m_2、m_2/m_3 和 m_3/q。我们首先展示，每个训练得到的BDRE估计器都能在各自对应的训练分布上准确估计相应的密度比。为此，在图2第一行的第2至第5个面板中，我们在各自分母概率密度 m_1、m_2、m_3、q 的样本上评估 $\hat{}\eta_1$、$\hat{}\eta_2$、$\hat{}\eta_3$、$\hat{}\eta_4$，并通过散点图绘制结果；横轴标示样本所抽取自的分布，纵轴表示对数密度比（红色）。我们以蓝色绘出真实密度比作为比较。显然，红色和蓝色散点图高度重合，表明各个比率估计器在各自对应的分母（训练）分布上都很准确。

接下来，我们在来自 p 和 q 的样本上评估 BDRE $\hat{}\eta_1$、$\hat{}\eta_2$、$\hat{}\eta_3$、$\hat{}\eta_4$，而不是像之前那样在对应的训练分布上评估。分布 p 和 q 展示在图2第二行的第1个面板中。第二行其余面板（2-5）中，估计器 $\hat{}\eta_1$、$\hat{}\eta_2$、$\hat{}\eta_3$、$\hat{}\eta_4$ 与真实对数密度比（蓝色） p/m_1、m_1/m_2、m_2/m_3 和 m_3/q 进行比较；这些真实对数密度比同样是在来自 p 和 q的样本上评估的。与第一行不同，估计的对数密度比与真实值并不吻合。这反映了上文指出的训练评估分布偏移问题。现在我们将说明，单个 BDRE 的准确度下降会导致 TRE 的总体表现恶化。为此，我们首先通过望远镜乘积将分别训练的 BDRE 串联起来，从而得到 TRE 估计器，即 $\hat{}\eta_1 * \hat{}\eta_2 * \hat{}\eta_3 * \hat{}\eta_4$，然后在来自全部 5 个分布 p、m_1、m_2、m_3、q的样本上对其进行评估。结果展示在第三行的面板 1-5 中。估计的对数密度比（红色）与相应的真实对数密度比（蓝色）并不吻合，这表明单个 BDRE 的训练分布与评估分布之间的分布偏移会显著降低 TRE 的总体估计精度。当 p 和 q 都不具有全支撑集时，还会出现其他问题，详见附录G。

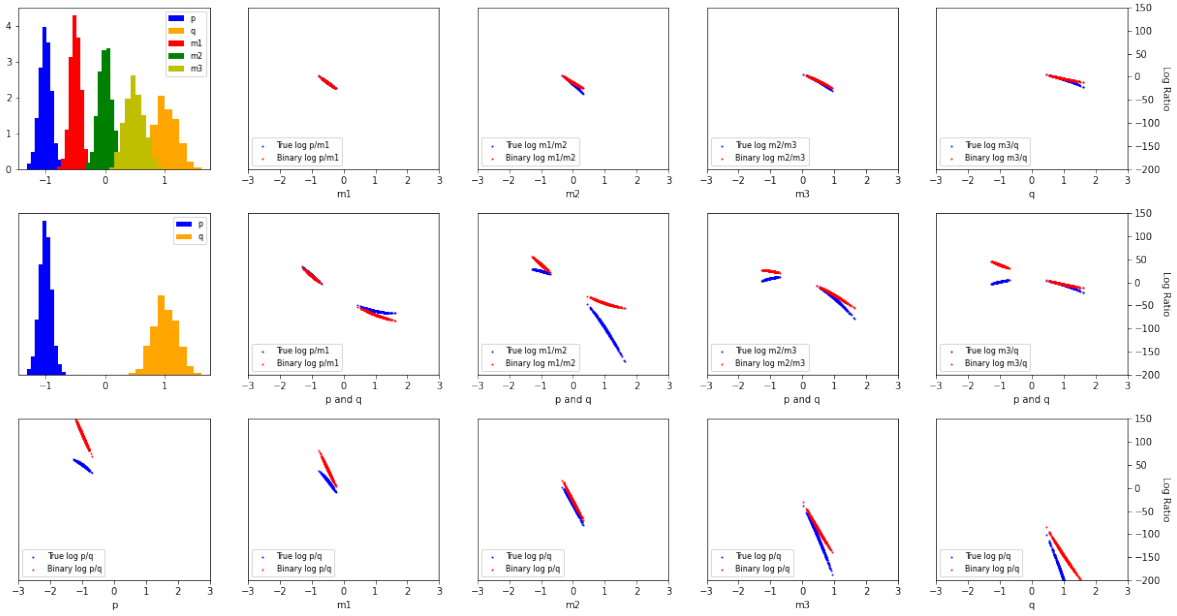


图2：图1中 $p = \mathrm{N}(-1, 0.1)$ 和 $q = \mathrm{N}(1, 0.2)$ 的 TRE。在所有散点图中，x 轴表示抽样分布，y 轴表示对数密度比。第一行的密度图展示了p、 q 和 3 个中间分布；第二行的密度图仅展示了 p 和 q 。第一行的散点图展示了各个密度比估计器在来自相应训练数据（分母概率密度）的样本上的评估结果，表明它们在训练集上能够准确估计。第二行的散点图展示了各个密度比估计器在来自 p 和 q 的样本上的评估结果。由于训练与评估分布偏移，估计精度明显下降。最后一行展示了整体密度比估计器在来自 p、m_1、m_2、m_3、q 的样本上的表现。我们可以看到，整体比率估计受到各个比率估计恶化的显著影响，这表明在分布分离良好的情况下，TRE 对分布偏移问题十分敏感。

3 使用多项逻辑回归进行密度比估计

我们提出基于多项逻辑回归的密度比估计器（MDRE），以解决密度鸿沟问题，同时避免 TRE 的分布偏移问题。与 TRE 一样，我们引入一组 $K \geq 1$ 辅助分布 $\{m_k\}_{k=1}^K$。但是，与 TRE 不同，我们将对数 p/q 的估计问题表述为一个多分类问题，而不是一系列二分类问题。我们证明，这一改变带来的估计器在所有 $K + 2$分布的定义域上都具有较高准确度，因此不会受到分布偏移的影响。

3.1 损失函数

我们在此建立密度比估计与多项逻辑回归之间的正式联系。考虑一组 C 分布 $\{p_c\}_{c=1}^C$，并令 $p_x(x) = \sum_{c=1}^C \pi_c p_c(x)$ 为它们的混合分布，其类别先验概率为 π_c。[1] 此时，多类别分类问题就是根据混合分布 p_x 中的样本，预测其正确类别$Y \in \{1, \ldots\ldots, C\}$。为此，我们考虑模型

$$P(Y = c|x; \theta) = \frac{\pi_c \exp(h_\theta^c(x))}{\sum_{k=1}^C \pi_k \exp(h_\theta^k(x))}, \tag{1}$$

[1]在我们的模拟中，我们将对各类别使用均匀先验。

其中 $h_\theta^c(x)$、 $c = 1, \ldots\ldots, C$ 是由 θ 参数化的未归一化对数概率。我们通过最小化负多项式对数似然（即 softmax 交叉熵损失）$\mathcal{L}(\theta)$ 来估计 θ

$$\mathcal{L}(\theta) = -\sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}[\log P(Y = c|x;\theta)] = \sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}\left[-\log \pi_c - h_\theta^c + \log \sum_{k=1}^{C} \pi_k \exp(h_\theta^k(x))\right], \quad (2)$$

其中，实际应用中会用样本平均替代期望。我们用$\theta^* = \arg\min_\theta$ 损失(θ)表示最优参数。为便于进行理论推导，我们考虑这样一种情形： $h_\theta^c(x)$以足够灵活的方式参数化，因此可以将上述损失函数视为关于 C 函数 $h_1, \ldots, h_C$的泛函，

$$\mathcal{L}(h_1, \ldots, h_C) = \sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}\left[-\log \pi_c - h_c + \log \sum_{k=1}^{C} \pi_k \exp(h_k(x))\right]. \quad (3)$$

以下命题表明，最小化损失$(h_1, \ldots, h_C)$可以估计任意一对 C 分布 p_c之间的对数比。

命题 3.1。 设 $\hat{} h_1, \ldots, \hat{}h_C$ 是方程 损失$(h_1, \ldots, h_C)$ 中的极小值点，其中方程编号为 3。则对于任意 $p_i(x)$ 与 $p_j(x)$ 之间的概率密度比 $i, j \le C$ 可表示为

$$\log \frac{p_i(x)}{p_j(x)} = \hat{h}_i(x) - \hat{h}_j(x) \quad (4)$$

对所有 x ，其中 $p_x(x) = \sum_c \pi_c p_c(x) > 0$。

证明。 首先注意，期望之和 $\sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}$ 在方程 3 中等价于关于混合分布的期望 p_x。将期望写成积分形式，可得

$$\mathcal{L}(h_1, \ldots, h_C) = \sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}[-\log \pi_c - h_c(x)] + \int p_x(x)[\log \sum_{k=1}^{C} \pi_k \exp(h_k(x))]dx. \quad (5)$$

泛函 $h_1\mathcal{L}(h_1, \ldots, h_C)$ 关于 h_i, i=1. . . , C 的泛函导数为

$$\frac{\delta \mathcal{L}}{\delta h_i} = -\pi_i p_i(x) + p_x(x) \frac{\pi_i \exp(h_i(x))}{\sum_{k=1}^{C} \pi_k \exp(h_k(x))} \quad (6)$$

对所有 x ，其中 $p_x(x) > 0$。令导数为零，得到最优解的必要条件

$$\frac{\pi_i p_i(x)}{p_x(x)} = \frac{\pi_i \exp(h_i(x))}{\sum_{k=1}^{C} \pi_k \exp(h_k(x))}, \qquad i = 1, \ldots, C, \text{ and for all } x \text{ where } p_x(x) > 0. \quad (7)$$

方程 7 的左边等于真实条件概率 $P^*(Y = i|x) = {}^{\pi_i p_i}(x)\ {}_{p_x}(x)$。因此，在临界点 ˆ ˆ 处， $h_{1,\ldots,}h_C$ 满足 $P^*(Y|X)$ 得到了正确估计。由方程 7 可知，对于任意两个 i 和 j，有 $(\pi_i p_i)/(\pi_j p_j) = (\pi_i \exp(\hat{}h_i))/(\pi_j \exp(\hat{}h_j))$，即

$$\log \frac{p_i(x)}{p_j(x)} = \hat{h}_i(x) - \hat{h}_j(x) \quad (8)$$

对于所有 x ，其中 $p_x(x) > 0$，至此证明完成。 □

注3.2（可识别性）。尽管我们有 C个未知数 h_1，……， h_C 以及 C个方程，见方程 7，但这些方程存在冗余，因为

$$\sum_{i=1}^{C} \frac{\pi_i p_i(x)}{p_x(x)} = \sum_{i=1}^{C} \frac{\pi_i \exp(h_i(x))}{\sum_{k=1}^{C} \pi_k \exp(h_k(x))} = \frac{\sum_{i=1}^{C} \pi_i \exp(h_i(x))}{\sum_{k=1}^{C} \pi_k \exp(h_k(x))} = 1$$

这意味着，我们无法通过最小化方程 h_i 唯一确定所有 3。然而，根据方程 $h_i - h_j$，对于 $i \neq j$，其差值可以确定，且等于 p_i 与 p_j 之间所需的对数比率，见方程 8。

备注 3.3（参数化和有限样本量的影响）。 在实际中，我们只有有限量的训练数据，而且参数化限制了模型的灵活性。在额外假设下，例如，假设真实密度比 对数 $p_i(x)$ − 对数 $p_j(x)$ 可以表示为 $h^i_\theta(x)$与 $h^j_\theta(x)$之差，我们在附录 A 中证明了比率估计器具有一致性。这里我们不再深入探讨估计器的渐近性质，而是聚焦于方程 8中关键结果的实际应用。

重要的是，方程 8 使我们能够通过将比率估计问题表述为多项式非线性回归问题来估计 $r_{p/q}$，如下述推论所概述。

推论3.4。 设前两个类别的分布分别为 p 和 q，即 $p_1 \equiv p, p_2 \equiv q$，其余 K 个分布均等于辅助分布 m_i，即 $p_3 \equiv m_1, \ldots, p_{K+2} \equiv m_K$。那么

$$\log \hat{r}_{p/q}(x) = \hat{h}_1(x) - \hat{h}_2(x). \tag{9}$$

备注 3.5（不受分布偏移问题影响）。 由于方程8 对所有 x 均成立，其中混合分布 $p_x(x) > 0$，因此方程 $\hat{r}_{p/q}(x)$ 中的估计器 9 对以下范围内的所有 x 均有效：它们属于 p、q、m_1、…、m_K的定义域之并集。这意味着 MDRE 不会遇到分布偏移问题；而在 TRE 中，通过依次求解一系列二项逻辑回归问题时则会出现此类问题。我们将在介绍构造辅助分布的三种方案 3.3 之后，于第 m_1、…、m_K 节对此进行举例说明。

3.2 构造辅助分布

在 MDRE 中，需要构造辅助分布，使其支撑集与 p 和 q的经验密度重叠。这样可以更好地校准多类分类概率，从而实现准确的密度比估计。我们在图1的 (c) 面板中展示了这一点，其中p =服从 N (−1, 0.1)，q =服从 N (1, 0.2)，单个辅助分布 m则设为柯西分布 C(0, 1)，显然它与另外两个分布存在重叠。分类概率以散点图的形式叠加在这些分布的经验密度上。与 (a) 面板中的 BDRE 情形相比，后者在没有任何数据的区域也给出很高的置信度；而多分类器仅在数据的支撑集内为 p 和 q赋予较高的类别概率，而不会在这两个分布几乎没有数据点的区域赋予较高概率。此外，辅助分布很好地覆盖了 p 和 q 的概率密度较低的区域，为该区域中 $h_1(x)$和 $h_2(x)$的取值提供了必要的训练数据，从而得到 (d) 面板所示的准确对数密度比估计。这与 (a) 面板中的 BDRE 形成对比：在 BDRE 中，分类器虽然受到足够约束，能够正确分类，却没有得到充分训练，因而无法准确估计密度比（见面板 b）。与 BDRE 和 TRE 中辅助分布的用法相比，MDRE 中辅助分布的这种细微但重要的区别，使 MDRE 能够推广到域外样本；我们将在第 4.1 节中展示这一点。

接下来，我们简要介绍三种为 MDRE 构造辅助分布的方案，详细内容见附录 B：（1）重叠分布。与 TRE 不同，MDRE 的表述不要求对分布 p 和 q进行“逐步搭桥”，因此我们提出了一种构造辅助分布的新方法。我们将 m_k 定义为任意一种分布，其样本与 p 和 q重叠，且$p << m_k$、$q << m_k$。这包括重尾分布（例如柯西分布、学生 t 分布）、正态分布、均匀分布或它们的混合分布。所有低维模拟均采用此方案。（2）线性混合。在此方案中，m_k 定义为由分布 p 和 q分别生成的样本 $X_p = \{x^i_p\}_{i=1}^N$ 和 $X_q = \{x^i_q\}_{i=1}^N$ 进行线性组合后所得到的样本分布。对于来自 m_k 的单个样本 $x^i_{m_k}$，其生成过程为 $x^i_{m_k} = (1 - \alpha_k)x^i_p + \alpha_k x^i_q$，其中 $x_p \in X_p$、$x^i_q \in X_q$。这一构造与 Rhodes 等人（2020）提出的辅助分布线性组合方案类似，但存在几个关键差异，我们将在附录 B 中详细说明。其中一个差异是，α_k 不局限于 $0 < \alpha_k < 1$，因此可以形成完全包围 p 和 q的非凸混合。高维模拟采用此构造方案。（3）逐维混合。该构造方案由 Rhodes 等人（2020）提出。单个辅助分布m 中的样本通过组合来自 p 和 q的样本的不同维度子集得到。涉及高维图像数据的实验采用此方案。

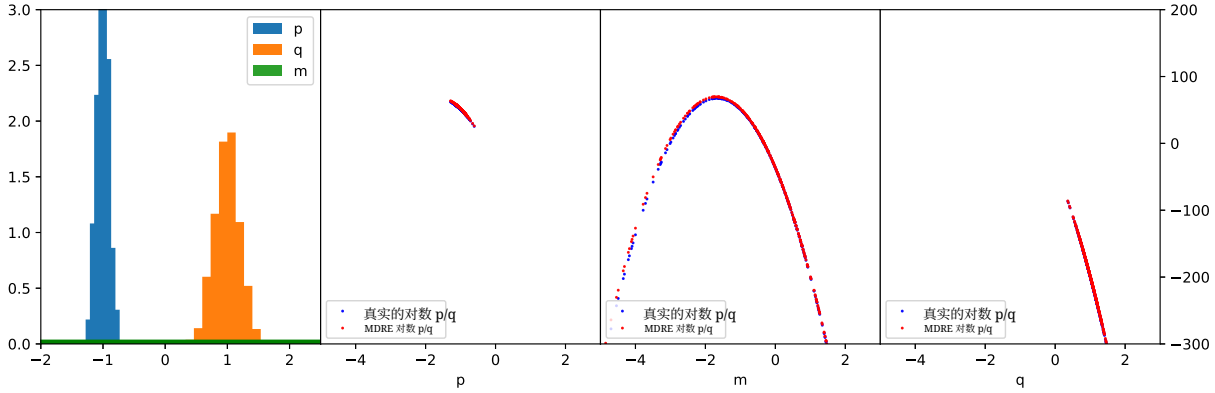


图 3： MDRE 与 $m =$ 柯西（0，1）。第一幅密度图展示了目标密度以及辅助分布（以绿色表示；由于重尾特性和坐标轴范围的原因，较难看清）。三幅散点图展示了在来自 p/q 的样本上评估的估计密度比（红色）和真实密度比（蓝色） p、 m和 q。 MDRE能够在整个输入域内准确估计密度比。与图 2 对比。

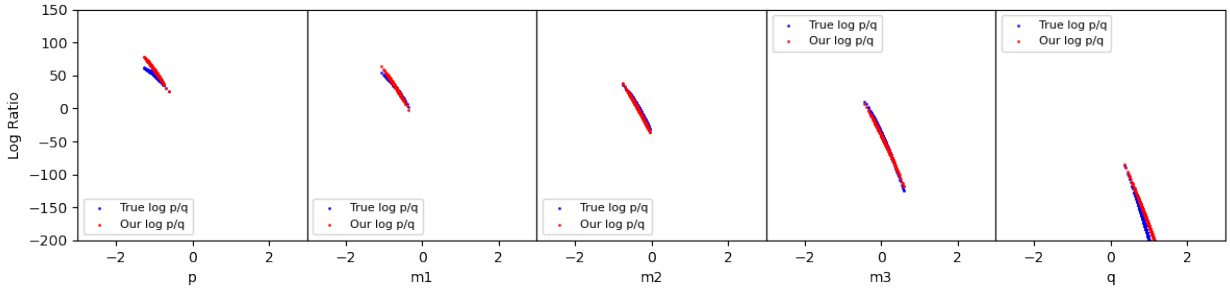


图 4： 使用 TRE 的辅助分布的 MDRE。每幅散点图展示了在横轴所示分布的样本上的总体对数密度比估计值（MDRE 以红色表示，真实比率以蓝色表示）。 MDRE 能够在所有样本上准确估计比率。与图 2 的底行对比。

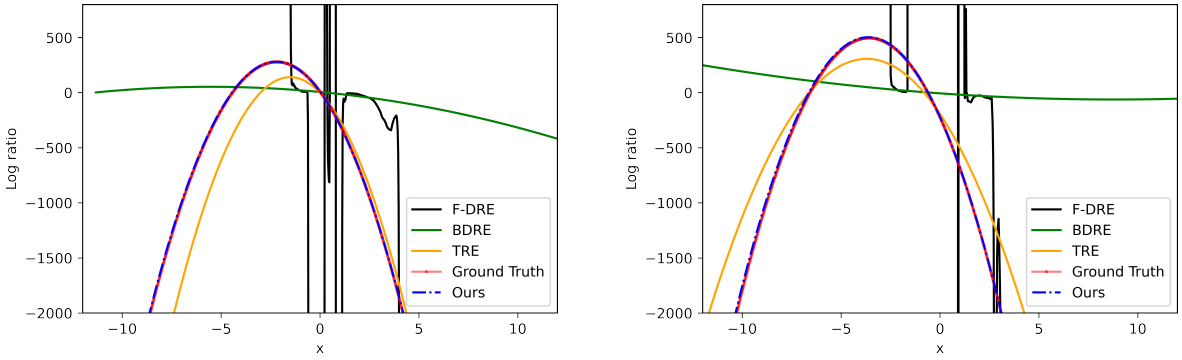
3.3 不受分布偏移问题影响

我们继续使用估计 p = N (-1, 0.1) 与 q = N (1, 0.2) 之间密度比的示例任务，并在此说明备注3.5：MDRE 不会受到第 2.1节所指出的分布偏移问题影响。我们用两类辅助分布测试MDRE 。第一类是在重叠分布方案下使用重尾分布（$m =$ 柯西分布$(0,\ 1)$）；第二类是使用路径标记分布m_1、 m_2、 m_3 ，如图 2 中的 TRE 所用，采用其线性混合构造方案。

图 3 展示了重尾 $\mathcal{C}$auchy$(0,\ 1)$ 辅助分布的结果（绿色，显示在最左侧的图中）。我们可以看到， p 与 q的对数比率，即使超出 p 的经验支撑集之外，仍能被 MDRE 准确地学习。其原因在于，MDRE 使用来自 p、q 和 m 的混合分布的样本进行训练，因此，根据备注3.5，在混合分布的支撑集上不会遇到分布偏移。图4展示了使用我们在图2中采用的 TRE 辅助分布时的结果。我们看到，对于来自 p 和 q的样本，以及来自辅助分布的样本，所学习的对数比率都与真实对数比率很好地吻合。这可以直接与图2的第三行进行比较，在该行中，TRE 受到分布偏移问题的影响，无法得到良好估计的对数比率。注意，我们没有展示与2 的第二行对应的结果，因为MDRE 中对数比率的估计不依赖任何中间密度比。

p	q	真实 KL	BDRE	TRE	F-DRE	MDRE（我们的）
$\mathcal{N}(-1,\ 0。08)$	$\mathcal{N}(2, 0.15)$	200.27	21.74 ± 4.10	136.05 ± 5.91	14.87 ± 1.72	203.32 ± 2.01
$\mathcal{N}(-2,\ 0。08)$	$\mathcal{N}(2, 0.15)$	355.82	20.22 ± 3.64	208.11 ± 18.31	14.22 ± 5.30	360.35 ± 1.37

表1： 针对具有较大一阶和高阶差异的 p 和 q 的一维密度比估计任务。在所有情况下，MDRE 的表现均优于所有基线方法。



(a) $p = \mathcal{N}(-1,\ 0。08)$ $q = \mathcal{N}(2,\ 0。15)$

(b) $p = \mathcal{N}(-2, 0.08)$ $q = \mathcal{N}(2, 0.15)$

图5：与表1中报告的数值对应的对数密度比估计。请注意，真实值曲线与MDRE 曲线重叠，而所有其他估计器的效果都显著较差。

4 实验

我们在此对 MDRE 在合成数据和真实数据上的表现进行了实证分析，结果表明，在三种不同的密度比估计任务中，它的表现优于以往方法——BDRE、TRE 和 F-DRE。我们考察了分子和分母概率密度因均值不同而存在差异的情况，即 p和 q 表现出一阶差异（FOD）；同时也考察了差异源于不同高阶矩的情况（高阶差异，HOD）。比率估计与 KL散度和互信息估计密切相关，因为 KL散度是关于 p下对数比率的期望，而互信息可以表示为联合分布与边缘分布乘积之间的 KL散度。由于这些量是机器学习中核心关注的对象，我们将使用它们来评估比率估计方法。

4.1 大KL散度的一维高斯实验

在以下一维实验中，我们考虑两种情形：一种是 $p = \mathrm{N}(-1, 0.08)$，且 $q = \mathrm{N}(2, 0.15)$；另一种是将 p 的均值移至 -2，以增大两个分布之间的分离程度。在这两种情形下，MDRE 的辅助分布 m 均为柯西分布(0,1)，因此我们得到一个三分类问题（p、 q、 m），以及三个用于参数化 MDRE 分类器的函数 h_θ^i。这三个函数是形式为 $w_1x^2 + w_2x + b$ 的二次多项式。对于所有方法，我们将样本总数设为 100K。[2] 我们在附录C的表5中提供了 MDRE 和其他基线方法的具体超参数设置。

表1 展示了结果。我们可以看到，MDRE 对 KL 散度的估计比基线方法更准确，而基线方法的误差则明显更大。

我们注意到，KL 估计只需在来自分子分布的样本上计算对数比率 $p_。$因此，在图5中，我们展示了所有方法的结果：在宽区间（-12, 12）上计算估计的对数比率。图中显示，在整个区间内，所有基线方法都无法准确估计比率，而MDRE 总体表现良好。这一点很重要，因为

[2]我们发现，即使将样本量缩小至 1K 或 10K，MDRE的结果也保持不变，参见附录4 中的表C。

Dim	μ_1, μ_2	真实互信息	BDRE	TRE	F-DRE	MDRE（我们的方法）
40	0, 0	20	10.90 ± 0.04	14.52 ± 2.07	14.87 ± 0.33	18.81 ± 0.15
	-1, 1	100	29.03 ± 0.09	33.95 ± 0.14	13.86 ± 0.26	119.96 ± 0.94
160	0, 0	40	21.47 ± 2.62	34.09 ± 0.21	12.89 ± 0.87	38.71 ± 0.73
	-0.5, 0.6	136	24.88 ± 8.93	69.27 ± 0.24	13.74 ± 0.13	133.64 ± 3.70
320	0, 0	80	23.47 ± 9.64	72.85 ± 3.93	9.17 ± 0.60	87.76 ± 0.77
	-0.5, 0.5	240	24.86 ± 4.07	100.18 ± 0.29	10.53 ± 0.03	217.14 ± 6.02

表2： 高维互信息估计任务。MDRE 通常能够以大幅优势准确估计互信息。

这意味着，在 p 和 q 概率质量较小的区域内，比率仍能得到良好估计。这些结果证明了MDRE 的有效性：只需使用一个辅助分布，其样本与 p 和 q的样本均有重叠，而不必像 TRE 那样使用由最多 $K = 28$ 个彼此紧密排列的辅助分布组成的 BDRE 链。更多结果和详细信息请参见附录C 。

为了进一步阐明 MDRE 的密度比估计行为，我们使用贝叶斯分析来分析其对数比值估计的不确定性。我们对分类器参数采用标准正态先验，并通过哈密顿蒙特卡罗获得后验样本。这些后验样本继而产生密度比估计的样本。图 6 显示，MDRE 的 KL 散度估计精度很高，这可归因于MDRE 在两个 p 和 q的高密度区域并集附近能够有把握地给出准确估计。更详细的分析见附录D。

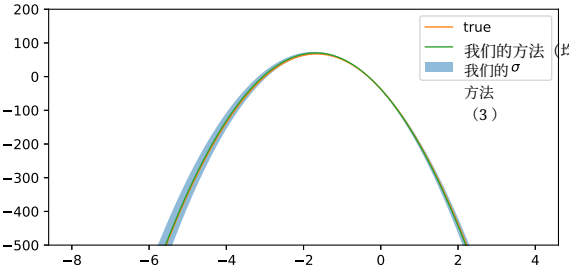


图6： 对 $p = \mathcal{N}(-1.0,\ 0.1)$ $q = \mathcal{N}(1.0,\ 0.2)$ 的 MDRE 进行贝叶斯分析。

4.2 互信息较大的高维实验

遵循 Rhodes 等人(2020 年) 的做法，我们采用 Belghazi 等人(2018) 的互信息估计基准；并根据 Poole 等人(2019) 的方法，在一个更具挑战性的高维问题上评估 MDRE。该任务的目标是估计标准正态分布与具有块对角协方差矩阵的高斯随机变量之间的互信息$x \in \mathbb{R}^{2d}$，其中每个块为 2×2，对角线上为 1，ρ 非对角线上的值为。相关系数 ρ 根据维度数量和目标互信息 $I = -d/2\log(1-\rho^2)$计算。由于这一问题构造仅产生高阶差异（HOD），我们通过移动两个分布的均值增加了额外挑战，从而也引入了一阶差异（FOD）。

对于MDRE，我们使用如下形式的二次函数 h_θ^i对 $x^T W_1 x + W_2 x + b$进行建模。我们使用线性混合来构造每个 m_k，其中 $K = 3$ 或 $K = 5$。在附录E中，我们在表MDRE 的6 中给出了 m 和 K 的确切配置，并说明了如何在实践中选择它们。

表2 展示了每个 MI 任务的结果，这些结果是对使用不同随机种子的 3 次运行取平均得到的。MDRE在原始 MI 任务（分布的均值相同）中优于所有基线方法。当均值允许非零时，MDRE 与基线方法之间的性能差距尤为显著。只有MDRE能够较为准确地估计 MI，而所有基线方法都严重低估了它。我们还注意到，MDRE 最多只使用5 个辅助分布，计算需求低于表现次佳的 TRE；后者最多使用15个辅助分布来构建其望远镜链。

我们发现，Kato 与 Teshima (2021) 提出的解决方案旨在克服基于 Bregman 散度最小化的 DRE 中的过拟合问题，但在实践中效果不佳。在高维

表 2 第 2 行的设置中，真实值 MI 为 100，而 MDRE 估计为 119 ± 0，94，Kato 与 Teshima (2021) 提出的最佳模型得到 1.60，严重低估真实值，且比基于分类器的 DRE 基线方法小一个数量级。更多结果（例如估计对数比率与真实对数比率的图表、训练曲线等），请参见附录 E。

如前所述，遵循先前的研究，我们在 p 和 q 均为正态分布的问题上评估了这些方法。为进一步完善分析，我们还在以下三种新的实验设置中评估了MDRE 。结果汇总于表3。

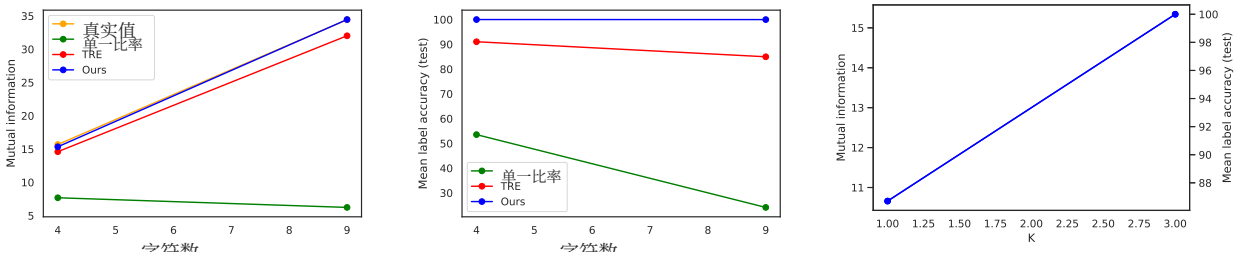
打破对称性在表2所报告的高维实验中，多数情况下，高斯分布 p 和 q 的均值关于零对称。为确保这种对称性并未使MDRE占得优势，我们还在均值经过随机化的高斯分布 p 和 q 上对其进行了评估。结果见表3的第2行和第6行。可以看出，MDRE仍能准确估计真实的 KL散度，表明它并未不公平地受益于分布关于零的对称性。

模型失配在表3的第4、5、7和8行中，我们通过将一个或两个分布 p 和 q 替换为尺度相同且均值经过随机化的学生 t 分布，来评估MDRE 。对于学生 t 分布，我们将自由度设为5、10 或20。这些实验检验了 MDRE 在模型失配时的表现，也就是当使用与 p 和 q 被设定为具有更轻尾部的高斯分布时相同的二次模型，MDRE 的表现。我们发现，MDRE 在这些情况下仍能准确估计真实的 KL散度。对于其他测试分布（例如高斯混合分布，见表3第3行），我们也得到了相同的结果。

有限支撑 p 和 q 最后，我们在另一个问题上测试MDRE ，其中 p 和 q 是同时具有 FOD 和 HOD 的有限支撑分布。具体做法是将 p 和 q 设为截断正态分布，如表3第 1 行所示。我们还将 m 设为截断正态分布，并将其尺度设为 2，使其与 p 和 q均有重叠。这一设置类似于第3.3 节所示的一维高斯示例，且MDRE 能够准确估计真实KL散度。

Dim	p	q	m	真实 KL	估计 KL
1	截断正态分布 位置=-1 ，尺度=0.1 支撑集=(-1.1,-0.9)	截断正态分布 位置=1，尺度=0.2 支撑集=(-1.1,1.2)	截断正态分布 位置=-1 ，尺度=2 支撑集=(-1.1,1.2)	50.65	52.35
160	正态分布 位置=R(-.5,.5)，协方差=2×2 块对角	正态分布 位置=R(-.5,.5)，协方差=I	线性混合	54.29	54.10
160	正态分布 loc=-1，cov=2×2 块对角	MoG：0.5*正态分布(0.9, I) + 0.5*正态分布(1.1,I)	线性混合	105.60	98.27
160	学生t分布 loc=R(-.5,.5), 尺度=2×2 块对角, 自由度=5	学生t分布 位置=R(-.5,.5)，尺度=I，自由度=5	线性混合	51.26	49.01
320	学生t分布 loc=R(-.5,.5), 尺度=2×2 块对角，自由度=10	学生t分布 位置=R(-.5,.5)，尺度=I，自由度=10	线性混合	53.82	51.03
320	正态分布 位置=R(-1,1)，协方差=2×2 块对角	正态分布 位置=R(-1,1)，协方差=I	线性混合	110.05	102.63
320	学生t分布位置=R(-1,1), 尺度=2×2 块对角，自由度=10	学生t分布 位置=R(-1,1)，尺度=I，自由度=10	线性混合	103.12	113.53
320	正态分布 位置=0，协方差=2×2 块对角	学生t分布 位置=0，尺度=I，自由度=20	线性混合	82.02	83.63

表3： MDRE 的鲁棒性评估。这里，R(a,b) 表示随机均值向量，其中每个维度均从区间（a，b）中均匀采样。在所有情况下，MDRE 都能以很高的精度持续估计真实KL散度。



(a) 互信息估计　(b) 分类准确率　(c) 变化的 K （辅助分布数量）

图 7：SpatialMultiOmniglot 表征学习结果。图 (a) 展示了三种方法估计的互信息，MDRE 能够非常准确地估计真实互信息。图 (b) 展示了由此得到的分类准确率，图 (c) 展示了改变辅助分布数量对 MDRE 互信息估计的影响。

4.3 SpatialMultiOmniglot 表征学习

为了在大规模真实数据上评测 MDRE，我们参照 Rhodes 等人（2020）的设置，将 MDRE 应用于 Spatial-MultiOmniglot 问题（Ozair 等人，2019）中的互信息估计和表征学习任务。目标是估计 u与 v 之间的互信息，其中 u 是由不同 Omniglot 字母表中的字符构成的 $n \times n$ 网格，而 v 是一个 $n \times n$网格，其中包含 u中相应字符的下一个字符的（随机）实现。学习完成后，我们使用标准线性评估协议（Oord 等人，2018）评估编码器生成的表征。对于 MDRE，与 TRE 类似，我们采用互信息表征学习文献中常用的可分离架构，并使用形式为$g(u)^T W f(v)$的函数对未归一化对数分数 h^i_θ 进行建模，其中 g和 f是 14层卷积残差网络（He 等人，2015）。尽管这一模型相当于在 h^i_θ之间共享参数，但我们想强调的是，在前述所有示例中，我们并未在 h^i_θ之间共享参数。我们通过逐维混合构造辅助分布。

这里我们只将MDRE 与单一比率基线和 TRE 进行比较，因为Rhodes 等人(2020 年，图 4）已经证明，TRE 在完全相同的任务上显著优于对比预测编码（CPC）（Oord 等人，2018）和 Wasserstein 预测编码（WPC）（Ozair 等人，2019）。详细的实验设置请参阅附录F 。

如图7a所示，MDRE的表现优于 TRE 和单一比率基线，与真实值互信息完全一致。互信息估计方面的这一改进也体现在表征上。图7b展示了MDRE的编码器学得的表征可实现 ∼100Omniglot 字符分类准确率（两种情况下均为$d = n^2 = 4, 9$）。另一方面，随着任务复杂度增加，单一比率估计器和 TRE（使用完全相同的逐维混合来构造辅助分布）的表现均有所下降；对于 $d = 4$ 和 $d = 9$，TRE 的准确率分别最高仅为 91% 和 85%。为确保公平比较，所有模型均使用相同的编码器架构进行训练。

我们进一步研究了改变 K 对 $d = 4$ 设置的影响。对于 $K = 1$，我们将所有逐维混合样本合并为一个类别；而对于 $K = 3$，则将它们分别归入各自的类别（类别数对应于混合的维度数）。我们在图7c中展示了这一影响。与Ma & Collins (2018)的发现一致，增大 K 不仅有助于MDRE 达到真实值互信息，也能提升表征质量，使测试分类准确率从86.7% 提高到100% 。

5 讨论

在本研究中，我们提出了基于多项逻辑回归的密度比估计器（MDRE）。在我们的模拟中，这种新的密度比估计方法在有限样本（非渐近）条件下的表现优于当前最先进的方法。我们证明，该方法解决了近期 Rhodes 等人 Rhodes 等人(2020 年）提出的方法对潜在分布偏移问题较为敏感这一问题。 MDRE 通过引入辅助

这些分布与比率的分子分布和分母分布具有重叠的支撑集。随后，该方法训练一个多项逻辑回归模型来估计密度比。我们证明，MDRE 既有坚实的理论依据，又具有出色的实证表现，并为高维密度比估计问题确立了新的最先进水平。

不过，这种方法也存在一些局限。首先，尽管我们在实证研究中对比率进行了良好的估计，但没有提供任何估计界限，这意味着估计得到的 KL散度或互信息值可能偏高或偏低。其次，辅助分布 m 的选择是一个重要的考量因素，会显著影响 MDRE 的性能。虽然本文展示了三种构造辅助分布方案的有效性，但从实证角度来看，这绝非一项穷尽性的研究。我们希望在未来的工作中解决这些问题，包括开发基于学习的方法来构造辅助分布。

参考文献

Mohamed Ishmael Belghazi、Aristide Baratin、Sai Rajeswar、Sherjil Ozair、Yoshua Bengio、Aaron Courville 和 R Devon Hjelm。Mine：互信息神经估计。*arXiv* 预印本 *arXiv:1801.04062*，2018年。

Steffen Bickel、Jasmina Bogojeska、Thomas Lengauer 和 Tobias Scheffer。用于HIV疗法筛查的多任务学习。载于第*25*届国际机器学习会议论文集，第56–63页，2008年。

Kristy Choi、Madeline Liao 和 Stefano Ermon。特征化密度比估计。*arXiv*预印本 *arXiv:2107.02212*，2021a。

Kristy Choi、Chenlin Meng、Yang Song 和 Stefano Ermon。基于无穷小分类的密度比估计。*arXiv*预印本 *arXiv:2111.11010*，2021b。

Imre Csiszár。一个信息论不等式及其在证明马尔可夫链遍历性中的应用。*Magyer Tud. Akad. Mat. Kutato Int. Koezl.*，8:85–108，1964年。

Ian J. Goodfellow、Jean Pouget-Abadie、Mehdi Mirza、Bing Xu、David Warde-Farley、Sherjil Ozair、Aaron C. Courville 和 Yoshua Bengio。生成对抗网络。载于神经信息处理系统，2014年。

M. U. Gutmann 和 J. Hirayama。将 Bregman散度作为估计未归一化统计模型的通用框架。载于人工智能不确定性会议论文集（*UAI*），2011 年。

Michael Gutmann 和 Aapo Hyvärinen。噪声对比估计：一种针对未归一化统计模型的新估计原理。载于第十三届人工智能与统计学国际会议论文集，第 297–304 页。JMLR 研讨会与会议论文集，2010 年。

M.U. Gutmann 和 A. Hyvärinen。未归一化统计模型的噪声对比估计及其在自然图像统计中的应用。机器学习研究杂志，13:307–361，2012 年。

何恺明、张祥雨、任少卿和孙剑。用于图像识别的深度残差学习，2015 年。

Masahiro Kato 和 Takeshi Teshima。用于深度直接密度比估计的非负 Bregman散度最小化。载于国际机器学习会议，第 5320–5333 页。PMLR，2021 年。

刘冰彬、Elan Rosenfeld、Pradeep Ravikumar 和 Andrej Risteski。分析并改进噪声对比估计的优化景观。*arXiv* 预印本 *arXiv:2110.11271*，2021 年。

马庄和 MichaelCollins。条件模型中的噪声对比估计与负采样：一致性与统计效率。*arXiv* 预印本 *arXiv:1809.01812*，2018 年。

Aditya Menon 和 Cheng Soon Ong。《用于密度比和类别概率估计的损失函数关联》。载于国际机器学习会议，第 304–313 页。PMLR，2016 年。

Richard Nock、Aditya Menon 和 Cheng SoonOng。一项带有应用的 Ascaledbregmantheorem。神经信息处理系统进展，29:19–27，2016 年。

Sebastian Nowozin、Botond Cseke 和 Ryota Tomioka。f-gan：通过变分散度最小化训练生成式神经采样器。载于第*30*届神经信息处理系统国际会议论文集，第271–279页，2016年。

Aaron van den Oord、Yazhe Li 和 Oriol Vinyals。基于对比预测编码的表征学习。*arXiv*预印本 *arXiv:1807.03748*，2018年。

Sherjil Ozair、Corey Lynch、Yoshua Bengio、Aaron van den Oord、Sergey Levine 和 Pierre Ser manet。用于表征学习的Wasserstein依赖度量，2019年。

Ben Poole、Sherjil Ozair、Aaron van den Oord、Alexander A. Alemi 和 George Tucker。互信息的变分下界，2019年。

Joaquin Quiñonero-Candela、Masashi Sugiyama、Neil D Lawrence 和 Anton Schwaighofer。机器学习中的数据集偏移。麻省理工学院出版社，2009年。

Alec Radford、Luke Metz 和 Soumith Chintala。使用深度卷积生成对抗网络进行无监督表征学习。*arXiv* 预印本 *arXiv:1511.06434*，2015 年。

Danilo Rezende 和 Shakir Mohamed。使用归一化流进行变分推断。载于国际机器学习会议，第 1530–1538 页。PMLR，2015 年。

Benjamin Rhodes、Kai Xu 和 Michael U Gutmann。望远镜式密度比估计。*arXiv* 预印本 *arXiv:2006.12204*，2020 年。

Akash Srivastava、Lazar Valkov、Chris Russell、Michael U Gutmann 和 Charles Sutton。VEEGAN：使用隐式变分学习减少生成对抗网络中的模式崩溃。载于第 *31* 届神经信息处理系统国际会议论文集，第 3310–3320 页，2017 年。

Akash Srivastava、Kai Xu、Michael Gutmann 和 Charles Sutton。生成式比率匹配网络。发表于第八届国际学习表征会议，第 1–18 页，2020 年。

Masashi Sugiyama、Taiji Suzuki 和 Takafumi Kanamori。机器学习中的密度比估计。剑桥大学出版社，2012 年。

L. Wasserman。《*All of Statistics*》。Springer，2004。

A 一致性

在第 3.1 节中，我们重点讨论了方程 3 中损失函数 $\mathrm{L}(h_1, \ldots, h_C)$ 的性质。损失函数的自变量是函数 h_i，而损失函数则根据关于 p_c 的期望来定义。这简化了分析并提供了重要见解，但与实际情形不符。在此，我们放宽这些假设：首先考虑方程 2 中的损失 $\mathrm{L}(\theta)$，其中函数 h_c 由某些参数 θ 参数化。然后考虑用基于 n 个样本的样本平均替代期望的情形。相应的损失函数记作 $\mathrm{L}_n(\theta)$。本节的重点是推导出相应条件，使得最小化 $\mathrm{L}_n(\theta)$ 能够得到第 3.1 节中通过最小化 $\mathrm{L}(h_1, \ldots, h_C)$ 获得的结果。

引理 A.1。 将真实条件分布记为 $P^*(Y|x)$，则有

$$\hat{\theta} = \arg\min_{\theta} \mathcal{L}(\theta) = \arg\min_{\theta} \mathbb{E}_{p_x(x)} KL\left(P^*(Y|x) \,||\, P(Y|x;\theta)\right) \tag{10}$$

其中 $P(Y|x;\ \theta)$ 在公式 *1* 中定义。

证明。 我们从公式$\mathcal{L}(\theta)$ 中的定义开始2:

$$\mathcal{L}(\theta) = -\sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}[\log P(Y=c|x;\theta)] \tag{11}$$

加权期望之和 $\sum_{c=1}^{C} \pi_c \mathbb{E}_{x\sim p_c}$ 对应于关于 $p(Y,x)$ 的联合期望。将联合分布分解为 $p(x)P^*(Y|x)$，由此我们得到

$$\mathcal{L}(\theta) = -\mathbb{E}_{p(x)}\mathbb{E}_{P^*(Y|x)}[\log P(Y|x;\theta)] \tag{12}$$

and

$$\mathcal{L}(\theta) + \mathbb{E}_{p(x)}\mathbb{E}_{P^*(Y|x)} \log P^*(Y|x) = \mathbb{E}_{p(x)}\mathbb{E}_{P^*(Y|x)} \log \frac{P^*(Y|x)}{P(Y|x;\theta)} \tag{13}$$

$$= \mathbb{E}_{p_x(x)}\mathrm{KL}\left(P^*(Y|x) \,||\, P(Y|x;\theta)\right) \tag{14}$$

该结论成立，因为新增项不依赖于 θ。 □

如果真实的条件分布 $P^*(Y|x)$ 属于参数族 $\{P(Y|x;\theta)\}_\theta$，那么 $\hat{\theta}$ 满足 $P(Y|x;\hat{\theta}) = P^*(Y|x)$，对所有满足 $p_x(x) > 0$ 的 x 均成立。因此，正文中公式 7 之后的相同论证可得公式 3.1 的参数形式，我们在以下推论中对其进行概述。

推论 A。 **2.** 如果真实的条件分布 $P^*(Y|x)$ 属于参数族 $\{P(Y|x;\theta)\}_\theta$，那么

$$h^i_{\hat{\theta}}(x) - h^j{}_{\hat{\theta}}(x) = \text{对数}\ \frac{p_i(x)}{p_j(x)} \tag{15}$$

对于所有满足 $p_x(x) = \sum_c \pi_c p_c(x) > 0$ 的 x。

接下来，我们推导使 θ 成为唯一极小值点的条件，这对于证明一致性是必要的。为此，我们围绕 θ 对 $\mathcal{L}(\theta)$ 进行二阶泰勒展开。

引理 A.3。

$$\mathcal{L}(\hat{\theta} + \epsilon\phi) = \mathcal{L}(\theta) + \frac{\epsilon^2}{2}\phi^\top I \phi \tag{16}$$

其中 $\epsilon > 0$ 和 $I = -\mathbb{E}_{p_x(x)}\mathbb{E}_{P^*(Y|x)}[H(Y,x)]$。矩阵 $H(Y,x)$ 包含对数模型的二阶导数，即其(i,j)个元素为

$$[H(Y,x)]_{ij} = \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log P(Y|x;\theta)\Big|_{\theta=\hat{\theta}} \tag{17}$$

其中 θ_i 和 θ_j 分别是 θ 的第 i 个和第 j 个元素。

证明。 在$\mathcal{L}(\hat{\theta})$附近进行二阶泰勒展开，得到

$$\mathcal{L}(\hat{\theta} + \epsilon\phi) = -\mathbb{E}_{p_x(x)} \sum_c P^*(Y=c|x) \log P(Y=c|x, \hat{\theta} + \epsilon\phi) \tag{18}$$

$$= \mathcal{L}(\hat{\theta}) - \nabla_\theta \mathcal{L}(\theta)\Big|_{\theta=\hat{\theta}} - \mathbb{E}_{p_x(x)} \sum_c P^*(Y=c|x) \frac{\epsilon^2}{2} \phi^\top H(Y=c,x)\phi + O(\epsilon^2) \tag{19}$$

$$= \mathcal{L}(\hat{\theta}) - \frac{\epsilon^2}{2}\phi^\top \left[\mathbb{E}_{p_x(x)} \sum_c P^*(Y=c|x) H(Y=c,x)\right] \phi + O(\epsilon^2) \tag{20}$$

这里用到了在极小化器 $\hat{\theta}$ 处，$\mathcal{L}(\theta)$ 的梯度为零这一事实。由于 $\sum_c P^*(Y=c|x)H(Y=c,x) = \mathbb{E}_{P^*(Y|x)}H(Y,x)$，结论得证。 □

注意， $I(x) = -\mathbb{E}_{P^*(Y|x)}[H(Y,x)]$ 是条件费舍尔信息矩阵，而 $I = \mathbb{E}_{p_x(x)}I(x)$ 是关于 $p_x(x)$ 的期望值。

推论 A.4. 如果 I 是正定的，那么 $\hat{\theta}$ 是 $L(\theta)$的唯一极小化器。

证明。 如果 I 是正定的，那么 $\phi^\top I\phi > 0$ 对所有非零的 ϕ，根据引理 A.3， $\mathcal{L}(\hat{\theta} + \epsilon\phi) > \mathcal{L}(\hat{\theta})$只要 $\phi \neq 0$。 □

现在考虑目标函数 $\mathcal{L}_n(\theta)$，其中将 $\mathcal{L}(\theta)$ 中的期望替换为对 n 个样本求得的样本平均。令 $\hat{\theta}_n = \arg\min_\theta \mathcal{L}_n(\theta)$。

Prop命题 A.5. 如果 *(i)* I 是正定的，且 *(ii)* $\sup_\theta |L_n(\theta) - L(\theta)| \xrightarrow{p} 0$，那么 $\hat{\theta}_n \xrightarrow{p} \hat{\theta}$.

证明。根据推论 A.4，条件 (i) 确保 $\hat{\theta} = \arg\min_\theta L(\theta)$ 是唯一的极小化器，因此，将 $\hat{}$改变少许 θ 会使代价函数 $L(\theta)$ 增大。结合关于 $L_n(\theta)$ 一致收敛到 $L(\theta)$ 的技术条件 $\hat{}$ (ii)，可以证明，随着样本量 n 增加， θ_n 依概率收敛到 $\hat{}\ \theta$ ；推导过程与最大似然估计（Wasserman，2004，第 9.13 节）或噪声对比估计（ Gutmann 和 Hyvärinen，2012，附录 A.3.2）一致性的证明（例如）完全相同。

□

推论 A.6。 若 *(i)* I为正定矩阵，*(ii)* 上确界$_\theta$ $|L_n(\theta) - L(\theta)| \xrightarrow{p} 0$，且 *(iii)* 存在一个参数值$\theta^*$ 使得 $P^*(Y|x) = p(Y|x;\ \theta^*)$，则 $\hat{\theta}_n \xrightarrow{p} \theta^*$

证明。 根据命题 A.5，条件 (i) 和 (ii) 保证 $\hat{}\ \theta_n$ 收敛到 $\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$。根据 $\hat{}\ \hat{}$ 引理 A.1， θ 也使 $\mathbb{E}_{p_x(x)}$KL $(P^*(Y|x) \ ||\ P(Y|x;\ \theta))$ 达到最小。因此，若条件 (iii) 成立，则 $\theta = \theta^*$，结论得证。

□

命题 A.7 (比率估计器的一致性)。 若 *(i)* I 为正定矩阵，*(ii)* 上确界$_\theta$ $|L_n(\theta) - L(\theta)| \xrightarrow{p} 0$，*(iii)* 对某个参数值 θ^*，$P^*(Y|x) = p(Y|x;\ \theta^*)$，且 *(iv)* 从 θ 到 h^c_θ 的映射连续，则

$$h^i_{\hat{\theta}_n}(x) - h^j_{\hat{\theta}_n}(x) \xrightarrow{p} \log\frac{p_i(x)}{p_j(x)} \tag{21}$$

对所有 x，其中 $p_x(x) = \sum_c \pi_c p_c(x) > 0$。

证明。根据命题 A.5，条件 (i) 和 (ii) 保证 $\hat{}\ \theta_n$ 收敛到 $\hat{\theta} = \arg\min_\theta L(\theta)$。根据推论 A.2，条件 (iii) 保证，对于所有 x，有 $h^i_{\hat{\theta}}(x) - h^j{}_{\hat{\theta}}(x) =$ 对数 $p_i(x)\ p_j(x)$，其中 $p_x(x) = \sum_c \pi_c p_c(x) > 0$。由于连续函数对加法封闭，若条件 (iv) 成立，则从 θ 到 $h^i_{\hat{\theta}}(x) - h^j{}_{\hat{\theta}}(x)$ 的映射是连续的。于是我们可以应用连续映射定理，得出 $h^i_{\hat{\theta}_n}(x) - h^j{}_{\hat{\theta}_n}(x) \xrightarrow{p} h^i_{\hat{\theta}}(x) - h^j{}_{\hat{\theta}}(x) =$ 对数 $p_i(x)\ p_j(x)$，从而证明了该结论。

——

□

B 构造 M

这里我们将详细说明本研究中使用的三类辅助分布。

重叠分布： 当 $p << m$ 且 $q << m$ 时，MDRE 估计量，对数 $\frac{p}{q}$ = 对数 $\frac{p}{m}$ –对数 $\frac{m}{q}$ 有定义。因此， m 的支撑集必须包含 p 和 q 的支撑集。任何具有全支撑集的分布，例如正态分布，都能轻松满足这一要求。然而，满足这一要求并不能保证有限样本情形下分布 p、 q 与 m 之间存在经验重叠。为确保两对分布的样本之间存在重叠，我们建议如下：

- 重尾分布：与正态分布相比，柯西分布和学生 t 分布等分布是 M 的更好选择。这是因为它们更重的尾部使得在 p 和 q 相距较远时，更容易将两者连接起来，并获得更高的样本重叠度（尤其是在 FOD 的情况下）。

- 混合分布：另一种利用 m 连接 p 和 q 、使它们的样本相互重叠的方法是使用混合分布。在这里，我们先将 p 和 q 分别与标准正态分布进行卷积，再对二者进行等权混合。

- 截断正态分布：如果 p 和 q 具有有限支撑集，也可以使用截断正态分布，或使用支撑范围至少覆盖 q整个支撑集的均匀分布。这一做法假设$p << q$。

线性混合：在这种构造方案中，分布 M 被定义为由样本构成的经验分布；这些样本通过对分布 p和 q中的样本 $X_p = \{x_p^i\}_{i=1}^N$ 和 $X_q = \{x_q^i\}_{i=1}^N$ 进行线性组合而构造。也就是说， m 是集合 $X_m = \{x_m^i | x_m^i = \alpha x_p^i +(1-\alpha)x_q^i, x_p \in X_p, x_q^i \in X_q\}_{i=1}^N$上的经验分布，其中 α 不受限于形成凸混合。此构造与 Rhodes 等人（2020）的线性组合辅助分布有关。在 TRE 中，辅助分布被定义为集合 $X_m = \{x_m^i | x_m^i = \sqrt{1-\alpha^2} x_p^i + \alpha x_q^i, x_p \in X_p, x_q^i \in X_q\}_{i=1}^N$上的经验分布，其中 $0 \leq \alpha \leq 1$。这种加权方案会使辅助分布中的样本偏向 p。因此，当 p 和 q 是有限支撑分布时，需要谨慎处理，以免辅助分布中的样本落在 q的支撑集之外。

使用任一种加权方案，都可以构造出 K 个不同的辅助分布。MDRE 可以使用 K+2 分类器分别处理这些 K 辅助分布，也可以将它们作为成分分布定义一个单一的混合分布，并训练一个 3 类分类器。我们将这种构造称为线性混合的混合。

逐维混合： 在这种构造方案中，直接沿用了 TRE 的方法， M 被定义为这样一种经验分布：从 p 和 q 的样本中选取不同的维度子集并进行组合，由此生成样本。为完整起见，下面介绍 TRE 中的确切构造方案：

给定一个长度为 d 的向量 x ，并且 d 可被 l 整除，我们可以写出 $x = (x[1], ...x[l])$，其中每个 $x[i]$ 的长度为 d/l。然后，第 k 个辅助分布中的一个样本表示为： $x_k^i = (x_q^i[1], ...x_q^i[j], x_p^i[j+1], ... x_p^i[l])$，其中 $j = 1, ... l)$， $x_p^i \sim p$ 和 $x_q^i \sim q$ 是随机配对的。

C 一维密度比估计任务

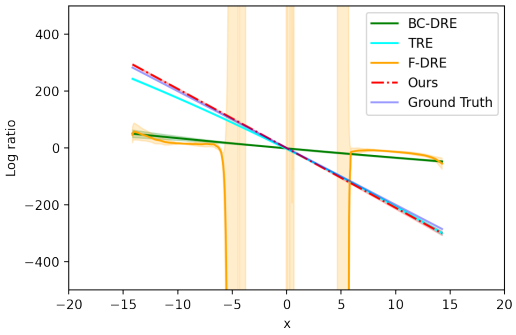
在第 4.1 节中，我们研究了两个分布 p 和 q 被 FOD 和 HOD 同时分开的三种情形。在所有这些一维实验中，所有模型均使用 100,000 个样本进行训练，所有结果均报告了采用不同随机种子的 3 次运行结果。此外，我们发现 MDRE 在使用 1K 和 10K 个样本时效果同样出色。实验配置（包括 MDRE 的辅助分布）详见表 5。

p	q	真实 KL 散度	MDRE @ 1K	MDRE @ 10K	MDRE @ 100K
$\mathcal{N}$(-1, 0.08)	$\mathcal{N}$(2, 0.15)	200.27	195.05	196.50	203.32
$\mathcal{N}$(-2, 0.08)	$\mathcal{N}$(2, 0.15)	355.82	346.92	348.97	360.35

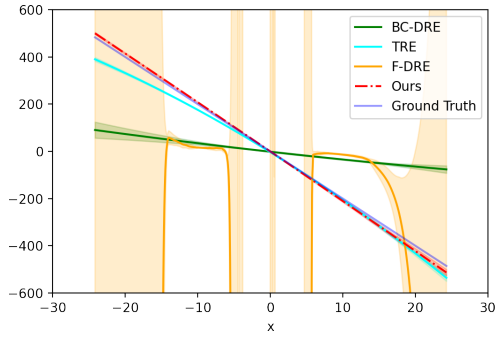
表 4： MDRE 在三种样本量设置下进行一维密度比估计。对于这三种设置，MDRE 都能很好地估计密度比。

p	q	TRE p_k	MDRE m
$\mathcal{N}(-1,\ 0。08)$	$\mathcal{N}(2,\ 0。15)$	线性混合，搭配 $\alpha = [0.053,\ 0.11,\ 0.16,\ 0.21,\ 0.26,\ 0.31,\ 0.37,\ 0.42,\ 0.47,\ 0.53,\ 0.58,\ 0.63,\ 0.68,\ 0.74,\ 0.79,\ 0.84,\ 0.89,\ 0.95]$	$\mathcal{C}(0,\ 1)$
$\mathcal{N}(-2,\ 0。08)$	$\mathcal{N}(2, 0.15)$	线性混合，取值为 $\alpha = [0.03、0.07,\ 0.1,\ 0.14,\ 0.17,\ 0.21,\ 0.24,\ 0.28,\ 0.31,\ 0.34,\ 0.38, 0.41, 0.45, 0.48,\ 0.52,\ 0.55,\ 0.59,\ 0.62,\ 0.66,\ 0.69,\ 0.72,\ 0.76,\ 0.79,\ 0.83,\ 0.86,\ 0.9,\ 0.93, 0.97]$	$\mathcal{C}(0,\ 1)$
$\mathcal{N}(-10,\ 1)$	$\mathcal{N}(10,\ 1)$	线性混合与 $\alpha = [0.11、0.22、0.33,\ 0.44,\ 0.55,\ 0.66,\ 0.77, 0.88]$	$\mathcal{C}(0,\ 2)$

表5：正文表1 和附录表4 的实验配置（见附录C



(a) 域内：$p = \mathcal{N}(-10,\ 1)$ 和 $q = \mathcal{N}(10,\ 1)$

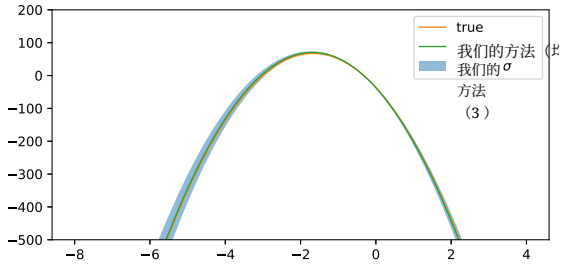


(b) 扩展域：$p = \mathcal{N}(-10,\ 1)$ 和 $q = \mathcal{N}(10, 1)$

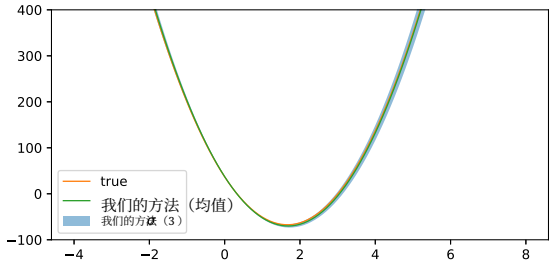
图8：一维密度比估计分析。图8a 评估了相应训练分布定义域内均匀样本的对数密度比。图8b评估了来自训练分布扩展域的均匀样本的对数密度比。阴影表示估计值的1个标准差。

D 使用哈密顿蒙特卡罗对 MDRE 对数比值估计进行不确定性量化

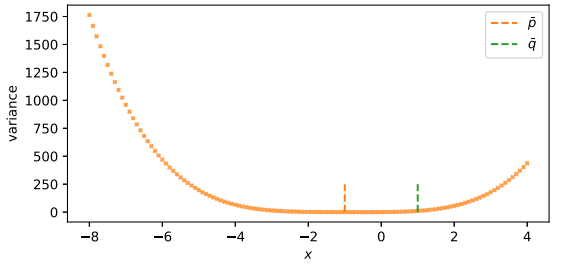
在一维实验中，即使在最先进的方法都无法奏效的困难情形下，MDRE 也始终能给出高度准确的 KL 散度估计。为了解 MDRE 为何能如此准确地估计 KL 散度，我们在贝叶斯框架下分析其对数比值估计的分布，以评估这些估计的可靠性，并研究其对 KL 散度估计的影响。在这项分析中，我们使用一个分类器，并将标准正态分布设为其参数的先验。对数比值估计的分布，就是采用不同



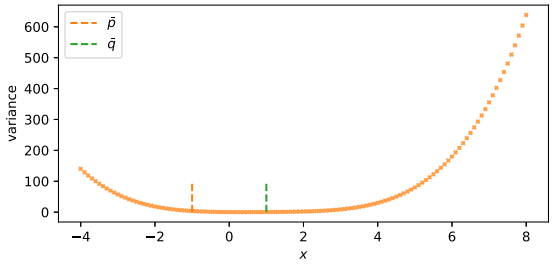
（a）对数$\frac{p}{q}$ 对于 $p = \mathcal{N}(-1.0,\ 0.1)$ $q = \mathcal{N}(1.0,\ 0.2)$



(b) 对数$\frac{p}{q}$ 对于 $p = \mathcal{N}(-1.0,\ 0。2)$ $q = \mathcal{N}(1。0,\ 0。1)$



(c) $\frac{p}{q}$的对数方差作为距 p和 q的众数距离的函数，针对 $p = \mathcal{N}(-1.0, 0.1)$ $q = \mathcal{N}(1.0, 0.2)$



(d) 对数 $\frac{p}{q}$ 的方差随距 p 和 q 的众数的距离变化，适用于 $p = \mathcal{N}(-1.0, 0.2)$ $q = \mathcal{N}(1.0, 0.1)$

图9：MDRE 估计量的不确定性量化。我们以浅蓝色绘制均值周围 3 倍标准差的范围。在图 (c) 和 (d) 中，柱形表示 p 和 q 的均值。

后验参数，通过采样得到。我们考虑两种设置：首先设定 $p = \mathrm{N}(-1.0, 0.1)$ 和$q = \mathrm{N}(1.0, 0.2)$，然后交换它们的尺度，即 $p = \mathrm{N}(-1.0, 0.1)$ 和 $q = \mathrm{N}(1.0, 0.2)$。在这两种情况下，我们都使用以分类器参数的最大似然估计为初始值的哈密顿蒙特卡罗 (HMC) 采样器，从后验分布中抽取样本。随后，我们利用这些样本计算一组 MDRE 对数比值估计的样本，并据此估计均值和标准差。图9 (a) 和 (b) 展示了这些结果。我们发现，MDRE 的估计准确，并且在 p (-1.0) 和 q $(+1.0)$ 的均值之间的区域附近不确定性最低。随着我们远离分布 p 和 q 的众数，不确定性会增加。图 (c) 和 (d) 展示了这一点，其中我们绘制了估计方差随样本位置变化的曲线。

由于 KL散度是来自 p 的样本上对数比率的期望值，且 p的高密度区域与 MDRE 的高置信度区域完全一致，因此即使 p 与 q 相距甚远，它也能稳定而准确地估计 KL散度。

E 高维实验

在第 4.2 节中，我们展示了，当 MDRE 和 p 以及 q 均为高维高斯分布时，其表现优于所有基线模型。此前，Rhodes 等人(2020 年) 仅研究了高维 HOD 情形；而我们还考察了 FOD 和 HOD 情形，以提供更全面的分析。结果表明，随着估计密度比变化，MDRE 在互信息估计任务中的表现优于所有其他方法。值得注意的是，MDRE 最多只使用 5 个通过线性混合方案构造的辅助分布，在同时包含 FOD 和 HOD 的情形下，表现显著优于 TRE；而 TRE 最多使用 15 个同样通过线性混合方案构造的辅助分布。这表明，采用多分类逻辑回归的方案确实能够避免 TRE 在同时存在 FOD 和 HOD 时的分布偏移问题，并有助于更准确地估计密度比。

下面介绍 MDRE 的配置及其他相关设置细节。

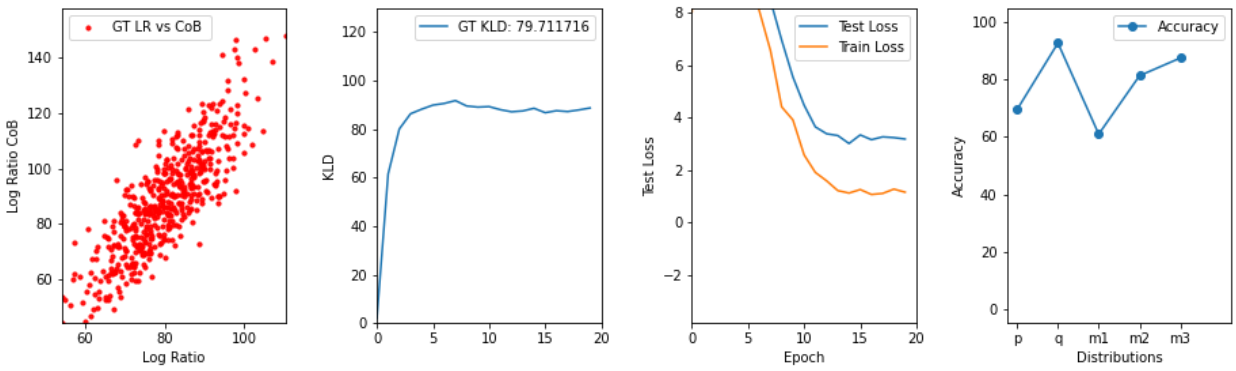


图10： 高维实验的诊断图。

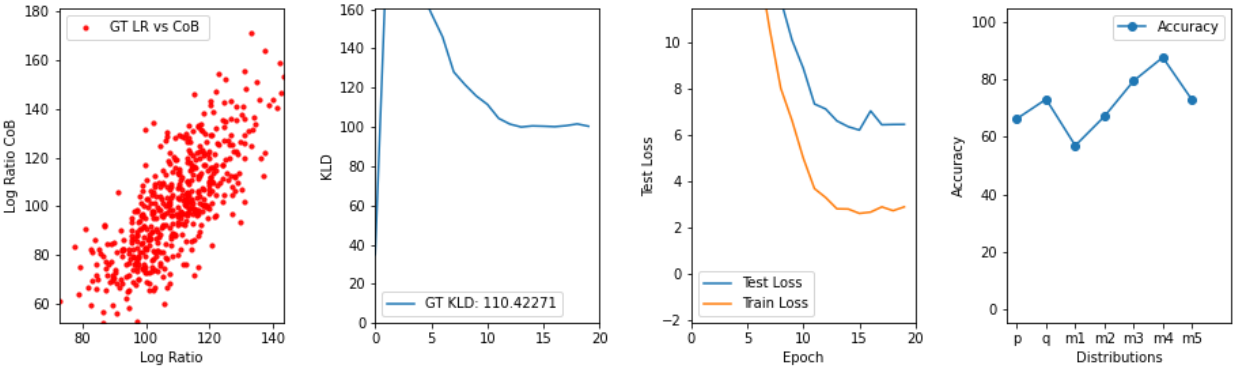


图11： 使用随机化均值的高维实验诊断图。

Dim	MI	m 使用 LM
40	20	[0.25,0.5,.75]
	100	[0.35,0.5,.85]
160	40	[0.25,0.5,.75]
	136	[0.15,0.35,0.5,.75,.95]
320	80	[0.25,0.5,.75]
	240	[0.15,0.35,0.5,.75,.95]

表 6： 高维实验中 MDRE 的配置。LM 表示线性混合

辅助分布： 在本文所有高维实验中，我们均按照附录 m 中所述的线性混合方案构造B。表 6 给出了 K 个辅助分布的数量，以及 6 种设置各自的精确混合权重。

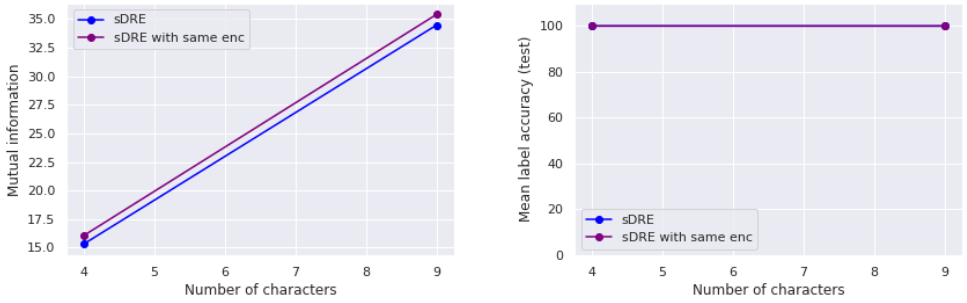
一般而言，我们选择了这三组混合权重，使其累积样本与 p 和 q 的样本重叠，类似于一维情形中重尾分布的作用。请注意，当 p 和 q 具有较高的 FOD 时，重尾分布可以有效地连接二者。然而，如果差异主要来自 HOD，它们的效果就不那么理想。例如，考虑 $p = \mathcal{N}(0,\ 1e-6)$ 和$q = \mathcal{N}(0,\ 1)$。在这种情况下，将 m 设为以零为中心的重尾分布并无帮助。我们需要一种在零附近高度集中的 m，同时又能与 q保持适度重叠。另一方面，线性混合 p 和 q 会混合一阶及更高阶统计量（二阶或更高阶），因此能够生成与 p 和 q均有重叠的样本。在某些情况下，我们发现，线性混合与 $K = 1$ 的混合也可用于估计密度比。不过，这需要使用基于神经网络的分类器，并且需要对超参数进行更多调优。

对于选择 K，我们采用基于网格搜索的方法。我们监测所有$K + 2$ 分布上的分类准确率。如果这一准确率非常高（> 95所有类别均为 %），则说明分类任务很容易，因此 DRE 可能会受到概率密度鸿沟问题的影响。另一方面，如果分类准确率过低（< 50所有类别均为 %），那么 DRE 同样无法进行准确估计。我们发现，以图 10 （最后一个面板）所示的准确率曲线为目标，经验上能够得到准确的密度比估计。该曲线绘制了所有类别的测试准确率；经验上，当准确率处于低、高界限（50%、95%）之间时，DRE 能够较好地估计比率。第一个面板显示，MDRE 能够准确估计来自所有 $K+2$分布中各样本的真实比率；第二个面板显示，MDRE 的 KL 估计接近真实 KL 值；第三个面板则显示，测试损失和训练损失均已收敛。图11 展示了随机化均值情况下的另一个示例。虽然MDRE 也成功得到了正确的真实 KL 值，而且大多数比率估计也很准确，但对于来自 p的一些样本，它对数比率的估计略微偏高。

F SpatialMultiOmniglot 实验

SpatialMultiOmniglot 是一个成对图像数据集 u和 v，其中 u是由来自不同 Omniglot 字母表的 Omniglot 字符组成的 $n \times n$网格，而 v是一个 $n \times n$网格，其中包含 u中对应字符的下一个字符。在此设置中，我们将每个 $n \times n$网格视为由 n^2 个类别随机变量组成的集合，而非单个像素。互信息 $I(u, v)$ 可计算为： $I(u, v) = \sum_{i=1}^{n^2}$对数 l_i，其中l_i 是 u中 i^{th} 字符的字母表大小。这个问题使我们能够轻松控制任务的复杂度，因为增加 n 会增加互信息。

对于模型，与 TRE 一样，我们采用基于互信息的表示学习文献中常用的可分离架构，并使用如下形式的函数对未归一化对数分数建模 $g(u)^T W f(u)$，其中g 和 f 是 14层卷积残差网络（He 等人，2015）。我们通过逐维混合来构造辅助分布——与 TRE 的做法完全相同。



(a) 互信息估计　　(b) 表征学习准确率

图12：使用相同编码器对 f 和 g进行表征学习的 SpatialMultiOmniglot 结果。

为了评估学习后的表示，我们采用标准的线性评估协议，在冻结的编码器 $g(u)$ 的输出上训练线性分类器，以预测网格 u 中每个字符的字母索引。

补充实验

除了正文中的实验外，我们还使用 SpatialMultiOmniglot 进行了额外实验，以测试对 g 和 f 使用相同编码器（即，采用 $g(u)^T W g(v)$ 的形式对未归一化对数分数建模），而不是使用 log $p(u,v) = g(u)^T W f(v)$）的效果。

单编码器设计：我们测试使用两个不同编码器 f 和 g 而非一个编码器的作用。如图12所示，在 $d = 4$、9这两种情况下，两个模型得到的互信息估计值略有不同，但相似；有趣的是，测试分类准确率完全没有差异。根据经验，我们还发现，使用一个编码器有助于模型更快地收敛。总体而言，这项实验表明，使用两个不同的编码器未必对我们有利。

K =1 时有限支撑分布上的 G TRE

对于 $K = 1$，TRE 提出了如下望远镜式分解：log p/q = log p/m + log m/q。因此，要使 TRE 定义良好，$\frac{dM}{dQ}$（即拉东-尼科迪姆导数 (RND)）必须存在。由此可见，只有当 $p << m << q$ 时，TRE 才有定义。然而，例如当 p 和 q 是有限支撑分布的混合分布时，这一条件很容易被破坏；唯一的例外是 m 的支撑集恰好与 q 的支撑集完全相同。

我们现在通过图13中的一个具体示例来演示这一点。此处我们设置 $p = 0.5 \times$ T N (−1, 0.1, low =−1.1, $high = -0.9$) $+ 0.5 \times$ T N (1, 0.1, $low = 0.9, high = 1.1$) 和 $q = 0.5 \times$ T N (−1, 0.2, $low = -1.2, high$ =0.8) $+ 0.5 \times$ T N (1, 0.2, $low = 0.8, high = 1$)，如图13a所示，其中 T N 表示截断正态分布。我们使用所提出的重叠分布构造，将 TRE 中的辅助分布 m 设置为 $m =$ T N (0, 1, $low = -1.2, high = 1.2$)。因此，$p << q << m$，而对于第二项，TRE 未定义，因为对于来自 m 且位于 q支撑集之外的样本，$\frac{m}{q}$ 未定义。从图13b中可以清楚地看到，对于来自 m 且 q 在其上没有任何支撑集的样本，$\frac{m}{q}$ 会飙升至极高值。对于 Rhodes 等人（2020）提出的所有辅助分布构造方案，都可以构造类似示例。

请注意，尽管 dM/dQ 未定义，但与所提出的 m 配合使用时，TRE 能够在来自 p 的样本上准确估计 log$r_{p/q}$。我们推测，这是因为 dM/dQ 的数值估计在 p 的支撑集上是有限的。

(a) 设置　　(b) dM/dQ

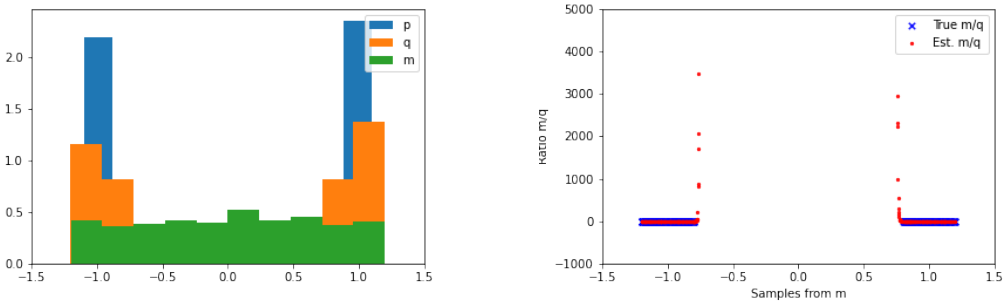


图13： dM/dQ 有限支撑集分布混合下