Counterfactual Prediction with Continuous Treatments
中文问答导读

独立导读版

目录

1 摘要导读

问：连续 treatment 比二元 treatment 真正多出的难题是什么？

答：不只是把两个潜在结果扩展成一条曲线，而是要同时面对：每个 dose 的局部 support 与 positivity、曲线的平滑结构、以及“每个 dose 的边缘反应分布”与“跨 dose 的完整联合反事实过程”之间的识别鸿沟。平滑预测不等于被数据识别。

问：HCGM-Dose 的核心产物是什么？

答：世界里实际 token 已由 $U = u^\star$ 固定；它从处理前协变量得到对同一 token 的 abduction result，再固定这份结果让可微 dose-indexed 机制 $Y_u(a) = f_\theta(a, E; u)$ 作用。Cauchy 仿射闭包给出解析的条件预测分布、分位数曲面、有限 dose 位置差及局部 dose 导数；这些首先是 dose-by-dose 的边缘对象。

问：论文最重要的正面结论与否定结论各是什么？

答：正面结论是：在一致性、弱条件可交换性、dose-density positivity 与正确设定下，事实总体风险可恢复有 support 的边缘预测分布。否定结论是：它既不识别无 support 的外推，也不识别完整联合反事实过程；个体跨 dose 差值分布还必须额外指定 noise coupling。官方 protocol 结果也分裂，而非某方法全面胜出。

2 引言导读

问：为什么把 dose 当作普通连续特征还不够？

答：因为目标不是相关性插值，而是 intervention-indexed worlds。模型必须明确哪些 dose 有因果 support、如何沿 dose 连续变化，以及哪些输出只是结构先验驱动的外推。

问：HCGM-Dose 的结构承诺是什么？

答：先由 X abduce 潜在 unit U，再由共享机制生成 $Y(a) = c(a) + w(a)^\top U + \sigma(a)E_a$。dose 不进入 abduction，因而把“unit 是谁”与“施加何种 dose”结构化分开。

问：作者刻意没有声称什么？

答：没有声称潜在坐标具有物理识别性，没有声称 abduction 能消除隐藏混杂，也没有声称事实数据决定了所有 dose 上反事实结果的联合随机过程。

3 相关工作导读

问：本文与 generalized propensity score 的关系是什么？

答：它接受同一识别边界：弱无混杂逐 dose 识别潜在结果边缘分布，条件 treatment density 提供 positivity；结果模型不能替代这些因果假设。

问：与 DRNet、SCIGAN、VCNet 相比，新意在哪里？

答： 共同目标是连续 dose-response；差异在显式 latent abduction、解析分布约化，以及对条件位置/分位数目标和识别边界的强调。直接平滑 baseline 用来检验收益是否只是 robust loss 或 dose features 带来的。

问：Cauchy 鲁棒性是否等于抗混杂？

答： 否。Cauchy likelihood 的有界残差 score 可缓和 gross outcome contamination，但 assignment bias 尤其隐藏混杂仍会破坏因果解释。

4 设定与估计目标导读

问：数据下标 i 是否等于 unit ontology？

答： 不是。i 只给 factual records 编号。模型层 primitive 是 U，一次 realization $U = u^\star$ 才是对应的实际 token/unit embedding。

问：positivity 在连续 dose 下应如何理解？

答： 点态要求是目标 (x, a) 上 $f_{A|X}(a \mid x) > 0$；稳定估计还需更强条件。论文定义 strong-support 区域 $\mathcal{S}_\epsilon = \{(x, a) : f_{A|X}(a \mid x) \geq \epsilon\}$，明确把因果主张限制在有 support 的 dose。

问：主要 estimands 是什么？

答： 对每个 dose，先定义条件 CDF F_a 与分位数 Q_a；主要曲面是中位位置 $m(a, x) = Q_a(1/2 \mid x)$，有限 dose contrast 为 $m(a, x) - m(a', x)$，局部效应为 $\partial_a m(a, x)$。

问：导数存在就代表局部因果效应吗？

答： 不代表。除了可微性，还要求邻近 dose 都有局部 support；跨越 support gap 的光滑导数只是模型外推。

问：policy 为什么必须 support-aware？

答： 优化应限制在 unit 的可行 supported dose 集内。否则“最优 dose”可能只是低密度区或无数据区的函数外推；文中的 policy regret 只是合成诊断，不是临床建议。

5 方法导读

问：从 $X = x$ 到 dose-response prediction，中间缺少的 ontology 是什么？

答：世界里先有一个固定 token $U = u^\star$。Learner 把 pretreatment x 当作 factual evidence，对同一个 $u^\star$ 做 abduction；然后固定这份结果，把 factual 或 alternative dose a 作为 mechanism query 输入 $Y_u(a) = f_\theta(a, E; u)$。不能用 alternative dose 重新 abduce unit。

问：point、Gaussian 与 Cauchy abduction 分别表达什么？

答：Point 直接输出 $\widehat{u}$；Gaussian 表达围绕局部中心的有限方差 uncertainty；Cauchy 用 location/scale 表达重尾、全局开放的 candidate-unit uncertainty。本文实证 Cauchy 与 Gaussian。公式中的 μ 只是 legacy implementation field，含义是 location 而非 mean；非退化 Cauchy 没有有限 mean/variance，且 full support 不意味着 generator 能产生任意 outcome。

问：解析 response kernel 从何而来？

答：编码器给出关于同一实际 token embedding 的独立坐标 Cauchy abduction result，机制对 candidate u 仿射作用并加入 Cauchy event noise。稳定分布闭包遂给出 $m(a, x) = c(a) + w(a)^\top \mu(x)$ 与 $s(a, x) = \sum_j |w_j(a)|\gamma_j(x) + \sigma(a)$，所以 $Y(a) \mid X = x$ 的预测分布和分位数可闭式计算。

问：模型到底学的是 marginal predictive distributions 还是 full joint process？

答：训练只需每个 factual dose 上的条件边缘预测分布。对 $\{E_a\}$ 的跨 dose 联合关系故意不指定，因此模型并未从数据学到 full joint counterfactual process。

问：为什么 pairwise individual contrast 需要 cross-dose noise coupling？

答：$Y(a) - Y(a')$ 同时依赖两个 dose 的 event noise。若独立重采样，scale 含 $\sigma(a) + \sigma(a')$；若共享同一 realization，则含 $|\sigma(a) - \sigma(a')|$。两者边缘分布相同，却给出不同差值不确定性与受益概率，事实边缘数据无法二选一。

问：连续机制相对分箱的直接优势是什么？

答：共享可微机制保留连续性并产生解析局部导数；分箱是 piecewise constant，箱内导数为零、边界不可导，还引入与 bin width 成正比的近似误差。

6 识别与性质导读

问：Fisher consistency 究竟保证了多少？

答：在真实 factual conditional Cauchy kernel 可由模型实现、交叉熵有限且求得总体无惩罚风险全局最优时，KL 风险差保证在 $P_{X,A}$ 几乎处处恢复 kernel；连续性加区间内正 density 才把结论延伸到该区间每个 dose。它不保证有限样本优化成功，也不识别唯一 latent factorization。

问：为什么无 support 区域永远不能靠 factual risk 恢复？

答：可在无 factual probability 的开集内给位置曲面加一个非零光滑 bump，而不改变任何事实样本处的密度与风险。故 smooth parameterization 只能选择一种外推，不能把它变成识别结论。

问：边缘可识别为何不推出联合可识别？

答：每个 F_a 固定后，跨 dose 仍可有许多 copula/noise coupling；它们共享全部边缘，却有不同 contrast scale 和 benefit probability。更进一步，任意拼出的 pairwise coupling 还可能不满足一个连续随机过程所需的有限维一致性。

问：Cauchy bounded score 的边界在哪里？

答：它界定的是相对于位置和 log-scale 的残差 score；不自动界定网络参数梯度，更不消除混杂导致的 causal bias。

7 实验设计导读

问：合成实验如何避免用 oracle 偷看训练？

答：五个冻结 seed 中，训练与早停只看 factual records；oracle 曲面、导数和全 dose evaluation tensor 只在选模后用于评估。指标分别覆盖位置曲面、导数、0.8 对 0.2 contrast 与网格 policy regret。

问：四种情景分别审查什么？

答：随机 assignment 检查基础恢复；观测选择检查已观测混杂下的性能；5% gross-label contamination 检查 likelihood 鲁棒性；隐藏混杂作为负对照，检验所有 factual learner 是否按理论边界失败。

问：official protocol 是否公平对齐？

答：作者固定公共 VCNet repository commit，运行其 VCNet-TR/DRNet-TR 已发表的 800-epoch 设置；HCGM 接收相同 factual matrices，并在相同 oracle grids 上用 average dose-response MSE 评估。该指标只测 population curve，不覆盖 HCGM 的完整条件分布目标。

问：support sensitivity 如何区分插值与外推？

答：用 assignment 机制诱导的精确条件 density，以 $f_{A|X} \geq 0.10$ 定义 strong support，并把 full-grid、strong-support 与 low-support cells 分开计分；policy 也限制在各 unit 的 strong-support dose 集。

8 结果导读

问：合成结果是否证明 latent factorization 普遍更准？

答：没有。干净 Gaussian outcome 上 direct Gaussian 的平均曲面误差最低；污染下 Cauchy HCGM 稳定，但 Direct-Cauchy 还略准，说明优势很大部分来自 Cauchy likelihood。factorization 的明确价值是解析分布约化和 intervention structure，而非普遍预测统治。

问：official protocol 的 split results 到底怎样？

答：结果明确分裂：Simu1 上 VCNet-TR/DRNet-TR 强于两种 HCGM；五个冻结 IHDP split 的均值由 HCGM-Dose-Gaussian 最低，Cauchy 版第二；News 五个 split 的均值同样是 Gaussian HCGM 最低，但 DRNet-TR 或 VCNet-TR 赢了其中两个单独 split，且 HCGM 标准误更大。因此不能宣称统计显著或普遍优越。

问：strong-support 与 full-grid evidence 为什么会给出不同故事？

答：assignment selection strength 从 0.5 增至 3.0 时，HCGM-Cauchy 的 strong-support fraction 从 0.916 降到 0.545，full-grid error 增至 1.51 倍；但 strong-support error 近乎不变（0.065 对 0.062），强选择下 low-support error 是 strong-support 的 2.52 倍。全网格恶化主要混入了扩大的低 support 外推区。

问：隐藏混杂负对照说明什么？

答：各方法的曲面、导数、contrast 和 policy error 一起上升，直接否定“latent abduction 自动恢复 exchangeability”的误读。

9 讨论与局限导读

问：部署连续 dose 系统时最该附带什么？

答：不应只返回平滑预测，还应返回 treatment density/support 状态；policy 搜索默认限制到 supported doses，若允许外推则必须明确标注为 modeling decision。

问：下一步方法学改进是什么，论文是否已验证？

答：density-aware diagnostics、balancing 或 doubly robust continuous-treatment objectives 都是自然方向，但本文未提供其有效证据，不能把未来工作写成已有贡献。

问：外部有效性还缺什么？

答：现有证据限于 synthetic 与 semi-synthetic。尚缺 direct SCIGAN、TCGA 以及设计上支持连续 dose 因果解释的真实应用；多 treatment、纵向 dosing、time-varying confounding、IV 与隐藏混杂 partial identification 也都未解决。

10 附录合并导读

问：理论附录最关键的反例手法是什么？

答：在无 support 的开集内加入 smooth bump，保持事实 support 上模型完全不变，即构造相同 factual risk、不同反事实曲面的两个模型；这是“平滑不等于识别”的直接证明。

问：coupling 非识别在证明中如何落地？

答：对两个 dose 的结构方程作差：独立 event draws 与共享 realization 都保持单 dose 边缘 Cauchy kernel，却分别产生不同的差值 scale。由此可见，pairwise contrast law 不是边缘 response kernels 的函数。

问：实现与复现实验有哪些关键冻结项？

答：HCGM encoder 为两个 32-unit 隐层、latent dimension 6；dose mechanism 为两个 24-unit 隐层并使用多项式/Fourier dose map。训练的初始化、minibatch、DGP、split 和 contamination 均有确定 seed；主合成实验使用 AdamW、最多 160 epochs 和 early stopping。

问：官方 benchmark 的算力差异会怎样影响解读？

答： Simu1 在本地 CPU；IHDP/News 的 HCGM fits 用 NVIDIA GB10，而官方外部模型因动态层实现留在 CPU。结果可作固定 protocol 下的预测比较，但不能把 wall-clock 或硬件效率混为算法优势。

参考文献

[1] Ioana Bica, James Jordon, and Mihaela van der Schaar. Estimating the effects of continuous-valued interventions using generative adversarial networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 16434–16445, 2020.

[2] Keisuke Hirano and Guido W. Imbens. The propensity score with continuous treatments. In Andrew Gelman and Xiao-Li Meng, editors, *Applied Bayesian Modeling and Causal Inference from Incomplete-Data Perspectives*, pages 73–84. Wiley, 2004.

[3] Kosuke Imai and David A. van Dyk. Causal inference with general treatment regimes: Generalizing the propensity score. *Journal of the American Statistical Association*, 99(467):854–866, 2004.

[4] Myrl G. Marmarelis, Elizabeth Haddad, Andrew Jesson, Neda Jahanshad, Aram Galstyan, and Greg Ver Steeg. Partial identification of dose responses with hidden confounders. In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, pages 1368–1379. PMLR, 2023.

[5] Lizhen Nie, Mao Ye, Qiang Liu, and Dan Nicolae. VCNet and functional targeted regularization for learning causal effects of continuous treatments. *arXiv preprint arXiv:2103.07861*, 2021.

[6] Patrick Schwab, Lorenz Linhardt, and Walter Karlen. Learning counterfactual representations for estimating individual dose-response curves. In *AAAI Conference on Artificial Intelligence*, volume 34, pages 5612–5619, 2020.