arXiv:2611.28414v1 [physics.data-an] 20 Nov 2026

基于神经网络的预测性拟合优度新物理搜索

Jing Lei 统计与数据科学系 卡内基
梅隆大学jinglei@andrew.cmu.edu

摘要

新物理的模型无关搜索旨在检验数据是否与含有未知干扰参数的参考模型相符。由于信号的形状和位置均未指定，这类检验需要灵活的学习器（如神经网络等拟合程序）；而对神经网络取最大值的似然比不属于标准渐近理论的适用范围，因此校准面临重大挑战。我们提出一种预测性备择方案：利用部分数据拟合相对于参考模型的偏离，再用其余数据评估该偏离相对于拟合参考模型带来的似然增益，并在多个折之间进行交叉拟合，以保持样本效率。我们按照 Neyman 的 $C(\alpha)$ 检验的思路，将干扰参数估计误差的一阶影响纳入检验统计量的方差，使检验无需针对每个干扰参数值重新校准。对于拟合偏离中的随机性，我们在算法稳定性条件下使用交叉验证统计量的中心极限定理处理；若未达到稳定性，则进行方差校正。校准采用解析方法，无需进行伪实验。在已确立的一维和五维（双轻子）基准测试中，对于不同的干扰参数和学习器调优设置，假阳性率始终处于名义水平的蒙特卡罗误差范围内。与现有的学习型检验相比，该检验的功效表现不一：对于某些信号更高，对于另一些则更低，具体取决于拟合偏离对信号的捕捉程度。

1 引言

现代寻找标准模型（SM）之外物理现象的研究会提出一个宽泛的问题：观测到的数据集是否与标准模型预测相容？从统计学角度看，这是针对参考模型的拟合优度检验，而该模型仅在以下条件下已知

：干扰参数受到带噪声的辅助测量约束。对此任务，一个有用的检验必须满足两个要求：(i) 能够探测位置和形状均未预先指定的偏离，这在实践中意味着要拟合一个灵活的偏离模型；(ii) 在估计干扰参数时，仍能控制假阳性率。

New Physics Learning Machine（NPLM） [1–3] 通过使用神经网络拟合偏离来满足第一个要求，其损失函数是似然比的负对数，而检验统计量 t 是最大化对数似然比的两倍，并根据干扰参数各自的辅助似然对其进行剖面化处理。第二个要求通过 Wilks定理得到满足：t 预计服从 χ^2 分布，其自由度数量等于网络参数的数量。我们将其与包含干扰参数的实现进行比较，参见文献 [3]。

这种校准存在两个弱点。第一，Wilks定理的条件很可能不适用于神经网络拟合。该定理描述的是正则参数族的最大化似然比：参数必须是可辨识的，真实值必须位于参数空间的内部，似然在其最大值附近必须能够被二次函数良好近似，并且参数数量必须相对于数据中的信息量足够小。神经网络违反了这些条件中的每一项：例如，在零假设下，其隐藏层参数对似然没有影响；在这种情况下，已知似然比不会服从 Wilks χ^2 [4, 5]，而且权重裁剪会将参数置于允许区域的边界上 [6, 7]。从一开始就有人注意到偏离 χ^2 定律的情况：最初的 NPLM 论文 [1] 在一个二维示例中报告了这种偏离，将其中一个原因归因于“训练不足”，并指出它所观察到的一致性“并不是一致性检验”。 χ^2 定律通过经验方式强制实施：在仅含参考分布的伪实验（玩具实验）上选择权重裁剪界，使得 t的分布与目标相匹配。这种校准同样很脆弱。在文献 [3], 的多变量基准测试中，与所选值在第二位小数上不同的裁剪值，会产生与目标明显不同的 t 分布。这就引出了第二个弱点：学习器的正则化同时也是其校准方式。同一个调节旋钮（例如权重裁剪）既决定网络描述信号的能力，也决定所报告的 p-值是否正确。这限制了为获得最佳检验功效而选择学习器的自由度。

在本文中，我们通过比较在留出数据上的预测来处理这两个问题。偏离仍然使用灵活学习器进行拟合（任何拟合程序，例如惩罚回归或神经网络），但它是在数据的一部分上拟合的

，并在另一部分上进行评估，因此该统计量衡量的是：相比拟合零假设，拟合的备择假设对其未见过的事件的预测好多少。这样，过拟合会降低统计量，而不是使其膨胀；其零分布也不再取决于学习器是否处于 χ^2 状态。此类分样本比较此前已有使用 [8, 9], ，但按原样使用时会在两个方面损失检验功效。我们的两项主要方法学贡献消除了这些损失。

- 交叉拟合。 单次拆分会将部分事件用于拟合，而这些事件随后不会用于评估。参考文献 [10]指出，“拟合（即训练）必须使用与检验完全相同的数据”，并发现，50/50 拆分会使 NPLM 的检验功效降低到仅使用一半数据时的水平。我们采用 K 折进行交叉拟合，使每个事件都同时用于拟合和评估，而不丢弃任何事件。

- 自适应零假设中心化。 与固定阈值比较的留出统计量（如参考文献 [8], 中的分样本似然比检验）忽略了这样一个事实：对于任何拟合的备择假设，其零假设均值都为负；参考文献 [10] 在使用留出数据评估 NPLM 时也观察到了这一点。我们将该统计量中心化到其在拟合零假设下的期望值，从而适应未知的零分布，并对其进行学生化。

所得统计量通过解析方法进行校准；在零假设下，其极限分布为正态分布：其中心化项和方差根据拟合零假设计算，并针对滋扰参数的估计、有限的参考样本以及拟合本身的随机性进行明确修正。

这种校准背后的条件也比 Wilks定理的条件更弱。出发点是关于交叉验证统计量 [11–13], 的中心极限定理，其对学习器的主要要求是稳定性：移动一个训练事件，应该只会使拟合偏离发生小幅变化 [14, 15]。这是对学习器的实质性要求，但并不涉及学习器如何参数化以及如何进行优化。其参数无需具有可辨识性，也无需位于参数空间的内部；其拟合无需达到似然的最大值；参数数量也不计入其中，因此裁剪、惩罚和早停拟合都涵盖在内。正则化对于校准仍然重要，但它是通过改善稳定性来发挥作用的，而不是通过将学习器限制在狭窄的范围内。见证函数是我们对拟合偏离的称呼，随后可以根据检验功效进行选择；我们使用了两种截然不同的见证函数：由光滑函数和高斯隆起函数构成的惩罚字典，以及

与 NPLM 相同的小型神经网络。给定学习器是否足够稳定，取决于学习器本身和数据。具体条件在第3.5节中给出。在我们的实验中，稳定性条件在零假设下并不成立：由于没有需要学习的偏离，学习器只能拟合噪声，因此拟合见证函数在不同拟合之间的变化幅度，可能达到其自身的大小。为了解决这一问题，我们针对拟合随机性加入了一阶方差校正，并从拟合结果本身估计该校正。针对这一情形的渐近理论尚待建立，而校正检验的有效性则通过数值实验得到验证。

我们在参考文献 [3] 中的基准上检验该方法：一个包含三种信号类型的一维指数参考模型，以及一个五维双轻子末态，其中包含 Z' 共振和有效场论（EFT）形变；两者均具有归一化和尺度干扰参数。研究结果主要有两点。第一，当真实干扰参数值偏离中心值一个标准差时，无论学习器如何调优（权重裁剪、训练时长、惩罚强度），也无论采用哪一种见证函数，假阳性率都接近其名义水平。第二，检验功效取决于见证函数捕捉信号的能力：在双轻子基准测试中，我们的网络对窄Z' 共振的灵敏度低于 NPLM，而对宽EFT形变的灵敏度更高。

本文结构如下。第2节使用 NPLM 的记号阐述问题。第3节介绍检验及其校准，并概述其保证和假设。第4节介绍两种见证函数。第5节和第6节展示基准测试，第7节讨论局限性和扩展。附录汇集了校准公式、极限定理的陈述及其证明概要、实现细节和补充结果。

2 统计设定

我们尽可能采用文献 [3] 中的记号；对应关系见表1。

数据。 数据集 $\mathcal{D} = \{X_1, \ldots, X_N\}$ 由样本空间 $\mathcal{X}$中具有特征 x的事件组成，这些事件被建模为强度为 $n(x)$ 的泊松过程：事件数 N服从均值为 $\int n$ 的泊松分布，并且在给定 N的情况下，各事件相互独立，密度为 $n/\int n$。参考（标准模型）假设将强度 $n(x \,|\, \mathrm{R}\nu)$ 预测为干扰参数向量 ν的函数，其中 $\nu = 0$ 为中心值， $N_0 = \int n(\cdot \,|\mathrm{R}_0)$ 为预期的事件数

。该构造适用于任意有限个干扰参数集合；为具体起见，这里写出最常见的两个干扰参数 $\nu = (\nu_n,\ \nu_s)$： ν_n表示所产生事件数的对数归一化， ν_s表示测量特征的对数尺度。

辅助测量。 干扰参数由独立的辅助测量 A 约束，该测量可概括为估计量 $\hat{\nu}_A \sim N(\nu^*,\ \Sigma_A)$，其中心位于真实值 ν^*，且 $\Sigma_A = \mathrm{diag}(\sigma_n^2,\ \sigma_s^2)$。其对数似然为

$$\ell_A(\nu) = -\tfrac{1}{2}(\nu - \hat{\nu}_A)^\top \Sigma_A^{-1}(\nu - \hat{\nu}_A). \tag{1}$$

Σ_A 的检验尺寸决定了必须在多大程度上利用数据本身来确定ν：当 $\Sigma_A \to 0$ 参考强度已知且完全确定时，而当 $\Sigma_A \to \infty$ 干扰参数完全由数据确定时。

参考模型的访问方式。 我们区分参考模型可用的两种方式，而检验的构造只需要其中较弱的一种。

- 预言机访问：强度 $n(x \mid \mathrm{R}\nu)$ 对每个 ν 都以解析形式已知，并且零假设下的每个期望都可以通过数值积分计算。

- 样本访问：强度由模拟链定义，只能通过在 $\nu = 0$ 处获得的 $\mathcal{R}$ M 模拟事件 X'_e的参考样本获取；每个事件携带权重 $r = N_0/M$，使得 $\mathcal{R}$ 重现期望计数 N_0。尺度干扰参数通过已知变换 T_{ν_s} 作用于模拟事件（例如对测量动量进行重新缩放），而归一化则将权重乘以 e^{ν_n}，因此可以获得任意 ν 下的样本。

预言机访问是参考样本无限大时样本访问的极限，下面的构造经过专门设计，使得同样的公式能够涵盖这两种情况。

假设与有效性。 零假设是：数据在干扰参数取某个值时遵循参考模型，

$$H_0 : \quad n(\cdot) = n(\cdot \mid \mathrm{R}_{\nu^*}) \quad \text{for some } \nu^*, \tag{2}$$

其中 ν^* 也是$\hat{\nu}_A$ 的分布中心值。备择假设可以是任何其他强度；不对偏离的形式作任何假设。名义水平为

表1：本文及文献 [3]中使用的符号。

量	本文	文献 [3]
数据集 , 事件数	$\mathcal{D}$, N	$\mathcal{D}$, $\mathcal{N}_{\mathcal{D}}$
在	$n(x \mid \mathrm{R}_\nu)$	$n(x \mid \mathrm{R}_\nu)$
干扰参数 ν		
在该处的期望事件数	N_0	$N(\mathrm{R}_0)$
中心值		
参考样本、检验尺寸、	$\mathcal{R}$、 M、 $r = N_0/M$	$\mathcal{R}$、 $N_{\mathcal{R}}$、 w_e
事件权重		
干扰参数（归一化、	$\nu = (\nu_n, \ \nu_s)$	$\nu = (\nu_{\mathrm{N}}, \ \nu_{\mathrm{S}})$
尺度)		
辅助数据、估计值,	A, $\hat{\nu}_A$, $\sigma_{n,s}$	$\mathcal{A}$, $\hat{\nu}$, $\sigma_{\mathrm{N,S}}$
不确定性		
干扰对数比（学习得到的	$h(x;\nu)$（ $\delta_{1,2}$，式（5））	$\log r(x;\nu)$ ($\log \hat{r}$)
近似）		
拟合偏离	见证函数 ϕ_k，在所有折上拟合，但 k	网络 $f(x;\mathbf{w})$，在 $\mathcal{D}$ 上拟合
检验统计量	标准化 Z	$t = \tau - \Delta$

$\alpha \in (0,1)$ 的检验若在 H_0 下，针对数据、参考样本和辅助测量进行重复时，对 ν^* 的每个取值，其拒绝率至多为 α ，则称其有效。

3 检验

3.1 交叉拟合与训练折上的拟合

数据事件被随机分配到 K 折 $\mathcal{D}_1$、……、$\mathcal{D}_K$ 中，各折大小最多相差一；在具有样本访问权限的情况下，参考样本也被独立划分为折 $\mathcal{R}_1$、……、$\mathcal{R}_K$。为简便起见，在推导中我们假设各折大小相等。对于关于 x 的函数 u 以及干扰参数值 ν ，我们记作

$$\mathbb{E}_\nu[u] = \int n(\cdot \mid \mathrm{R}_\nu)\, u, \qquad \mathbb{E}_{k,\nu}[u] = \frac{1}{K}\,\mathbb{E}_\nu[u], \qquad \mathbb{E}^{\mathrm{tr}}_{k,\nu}[u] = \frac{K-1}{K}\,\mathbb{E}_\nu[u]$$

对于在干扰参数取值为 ν 时，对所有数据事件、折 k 的留出事件以及其训练事件上的 u 期望求和。借助预言机访问，可通过数值积分计算；借助样本访问，则通过对经 T_{ν_s} 变换的参考事件、对 $\mathcal{R}_k$ （针对 $\mathbb{E}_{k,\,\nu}$ ）以及对 $\mathcal{R}_{-k}$ （针对 $\mathbb{E}^{\mathrm{tr}}_{k,\,\nu}$ ）进行加权求和来估计（附录 A.4）。

对于每一折 k，在其余 $K-1$ 折上进行两次拟合，$\mathcal{D}_{-k}$ 和$\mathcal{R}_{-k}$，并结合辅助测量：首先拟合干扰参数 $\hat{\nu}^{(-k)}$，其确定拟合的零假设强度 $\hat{n}_k(x) = n(x \mid \mathrm{R}, \hat{\nu}^{(-k)})$；然后拟合对该零假设的偏离，即见证函数。折 k 的见证函数是拟合的备择假设与拟合零假设之间的对数强度比：

$$\phi_k(x) = \log \frac{n_k^{\mathrm{alt}}(x)}{\hat{n}_k(x)} \,. \tag{3}$$

随后，留出折仅用于评估见证函数（第3.2节）。在全部 K 折上进行交叉拟合时，每个事件既用于拟合，也用于评估；当信号稀少时，这一点很重要。

见证函数拟合。 任何学习器都可以生成 ϕ_k，其形式和调优方式可以根据训练折和辅助测量来选择。第3.3节的校准以实际得到的见证函数为条件，并仅将其用作固定函数；例外是拟合随机性项 V_{fit}，它需要学习器对其训练事件的一阶响应。第4节详细介绍了训练损失以及我们使用的两种学习器。

零假设拟合。 设 ψ 为一个估计函数向量，每个干扰参数对应一个分量。我们的干扰参数拟合采用估计方程（或 M）估计量：$\hat{\nu}^{(-k)}$ 满足以下估计方程

$$\sum_{i \in \mathcal{D}_{-k}} \psi(X_i) = \mathbb{E}^{\mathrm{tr}}_{k,\nu}[\psi] + \Sigma_A^{-1}(\nu - \hat{\nu}_A) \,, \tag{4}$$

这些方程将 ψ 的训练总和设为其在 ν处的零假设模型下的期望值，并由辅助项进行校正。各拟合之间的差异仅在于 ψ的选择；附录A通过这种形式（式（22））对它们一并处理。

- 似然拟合：$\psi = s := \partial_\nu \log n(\cdot \mid \mathrm{R}, \nu)$这一选择给出了干扰参数的最大似然估计（MLE）：(4) 随后令训练事件的扩展对数似然加上 ℓ_A 的导数为零。精确的最大似然估计需要得分函数 s，而只有通过预言机访问才能获得该函数。通过样本访问，无法以闭式形式得知对数比 $h(x;\nu) = \log[n(x \mid \mathrm{R}, \nu)/n(x \mid \mathrm{R}, 0)]$,，因而也无法得知得分函数 $s = \partial_\nu h$。文献 [3] 通过局部表示近似 h

$$h(x;\nu) \simeq \nu_n + \nu_s \, \hat{\delta}_1(x) + \tfrac{1}{2} \, \nu_s^2 \, \hat{\delta}_2(x), \tag{5}$$

  $\hat{}$ 其系数函数 $\delta_{1,\ 2}$ 是在一个与其他所有内容相互独立的模拟样本上学习得到的，之后保持不变；我们以同样的方式学习它们

。由此得到的似然拟合，即参考文献 [3]，中使用的拟合，继承了它们的误差。因此，我们采用下面的“推前”拟合。

• 推前拟合： $\psi = (1,\ \delta_1)$，即在 $\nu = 0$处学习得到的得分函数。为避免学习得分函数的误差影响估计值，我们将其用作固定估计函数，并通过将 T_{ν_s} 应用于参考事件来计算其期望值，因此该拟合是一个矩估计量。使用第2节的对角 Σ_A ，式（4）可写为

$$
\begin{aligned}
N_{-k} &= \mathbb{E}^{\mathrm{tr}}_{k,\nu}[1] + \frac{\nu_n - \hat{\nu}_{A,n}}{\sigma_n^2}, \\
\sum_{i \in \mathcal{D}_{-k}} \hat{\delta}_1(X_i) &= \mathbb{E}^{\mathrm{tr}}_{k,\nu}[\hat{\delta}_1] + \frac{\nu_s - \hat{\nu}_{A,s}}{\sigma_s^2}.
\end{aligned}
\tag{6}
$$

ˆ 左侧分别是 δ_1的训练计数和训练总和，右侧则是零假设下在 ν 处的相应期望值，并经过辅助项校正。对于任何固定函数 δ_1，两侧在真实 ν^* 处的期望值相等ˆ，因此估计值是一致的ˆ，与 δ_1中的误差无关；该误差只影响其方差。

3.2 留出统计量及其中心化

见证函数在留出折 $\mathcal{D}_k$上进行评估，而它并未见过该折。每个事件以概率 $1/K$落入 $\mathcal{D}_k$ ，因此，数据的强度 n 也是该折的强度 n/K ，其扩展对数似然为 $\ell_k(n) = \sum_{i \in \mathcal{D}_k} \log\{n(X_i)/K\} - \frac{1}{K}\int n$。折 k 的留出统计量是拟合的备择假设相对于拟合零假设的对数似然比，

$$
\begin{aligned}
\hat{T}_k &= \ell_k\big(n_k^{\mathrm{alt}}\big) - \ell_k\big(\hat{n}_k\big) \\
&= \Big[\sum_{i \in \mathcal{D}_k} \log \frac{n_k^{\mathrm{alt}}(X_i)}{K} - \frac{1}{K}\int n_k^{\mathrm{alt}}\Big] - \Big[\sum_{i \in \mathcal{D}_k} \log \frac{\hat{n}_k(X_i)}{K} - \frac{1}{K}\int \hat{n}_k\Big] \\
&= \sum_{i \in \mathcal{D}_k} \log \frac{n_k^{\mathrm{alt}}(X_i)}{\hat{n}_k(X_i)} - \frac{1}{K}\int \big(n_k^{\mathrm{alt}} - \hat{n}_k\big) \\
&= \sum_{i \in \mathcal{D}_k} \phi_k(X_i) - \mathbb{E}_{k,\hat{\nu}^{(-k)}}\big[e^{\phi_k} - 1\big].
\end{aligned}
\tag{7}
$$

此类留出似然比是现有分样本检验的基础，例如，普适推断的分样本似然比检验 [8] 以及文献 [9]中用于新物理搜索的基于分类器的检验。这类

ˆ方法在实践中往往具有次优检验功效，因为它们将 T_kˆ与通过在 0 处进行零假设中心化校准的固定阈值进行比较，而忽略了 T_k在零假设下往往为负。为了解释其原因，假设留出事件ˆ服从拟合零假设；则给定训练折， T_k 的期望为

$$m_k = -\mathbb{E}_{k,\hat{\nu}^{(-k)}}\left[e^{\phi_k} - 1 - \phi_k\right] \;\le\; 0, \tag{8}$$

仅当 $\phi_k \equiv 0$ 时等号成立。在训练折中拟合了噪声的见证函数，其对留出折的预测比零假设的预测更差，因此过拟合ˆ会降低留出统计量，而不是提高它。相反，我们将 T_k 中心化于给定训练折时真实零假设下的期望，

$$m_k^{*} = \mathbb{E}_{k,\nu^{*}}[\phi_k] - \mathbb{E}_{k,\hat{\nu}^{(-k)}}[e^{\phi_k} - 1]\,.$$

由于 ν^{*} 未知，这种中心化无法直接实现，但将 ν^{*} 替换为其训练折估计值 ˆ$\nu^{(-k)}$ 可得 m_k，而第 k 折贡献

$$S_k = \hat{T}_k - m_k = \sum_{i \in \mathcal{D}_k} \phi_k(X_i) - \mathbb{E}_{k,\hat{\nu}^{(-k)}}[\phi_k], \tag{9}$$

见证函数的留出和减去其在拟合零假设下的期望值。交叉拟合检验统计量为

$$S = \sum_{k=1}^{K} S_k\,.$$

以拟合零假设而非真实零假设为中心，会通过 ˆ$\nu^{(-k)}$ 改变 S 的方差；第3.3节将干扰参数估计传播到方差中，并保留一阶项。

3.3 解析校准

在 H_0 下， S 的均值在检验所涉及的阶数上为零，而校准给出其方差。我们报告

$$Z = \frac{S}{\sqrt{\hat{V}}}, \qquad \hat{V} = \sum_{k=1}^{K}\left(\hat{V}_k^{D} + \hat{V}_k^{R}\right) + \hat{V}_A + \hat{V}_{\mathrm{fit}}, \tag{10}$$

并在 $Z > z_{1-\alpha}$ 时以显著性水平 α 拒绝零假设，其中 $Z > z_{1-\alpha}$ 为单侧高斯分位数。这四个方差项对应 S 中的四种随机性来源：留出事件、留出参考事件、辅助测量以及拟合本身。它们的完整表达式汇总于附录A；此处我们解释每一项所涵盖的内容。

留出事件与估计的干扰参数。给定训练折后，式（9）中的留出和是泊松过程上的求和，其方差为$\mathbb{E}_{k,\ \hat{\nu}^{(-k)}}[\phi_k^2]$。如果零假设已知，这便是正确的方差。但实际情况并非如此：$\hat{\nu}^{(-k)}$ 是估计得到的，其误差会使式（9）中被减去的期望值偏移约 $g_k^\top(\hat{\nu}^{(-k)} - \nu^*)$，其中 $g_k = \partial_\nu \mathbb{E}_{k,\ \nu}[\phi_k]$ 在 $\nu = \hat{\nu}^{(-k)}$ 处表示见证函数的零假设期望对干扰参数的敏感度。由于每个事件都参与 $K-1$ 次干扰参数拟合，这些偏移会在各折间累积，其量级与 S 本身相同。我们通过将每个事件产生的偏移与其所在的留出折归为一组，在方差中计入这些影响，从而得到正交化见证函数（推导见附录 A.2）：

$$\tilde{\phi}_k = \phi_k - (K-1)\, g_k^\top J_k^{-1} \psi\,, \tag{11}$$

其中，ψ 是干扰参数估计的估计函数（对于似然拟合，是得分函数 s；对于推前拟合，是 $(1, \delta_1)$），而 J_k 是相应估计方程的雅可比矩阵，其中包括辅助项 Σ_A^{-1}。然后，$\hat{V}_k^D$是 $\tilde{\phi}_k$的泊松方差：

$$\hat{V}_k^D = \mathbb{E}_{k,\hat{\nu}^{(-k)}}\left[\tilde{\phi}_k^2\right]. \tag{12}$$

这里将内曼的 $C(\alpha)$检验 [16], 应用于方差，这正是其核心思想：只有见证函数中干扰参数无法模仿的部分，才会对零假设波动有所贡献。辅助测量本身会在不同实验之间波动，并使每个 $\hat{\nu}^{(-k)}$ 同步变化；其贡献为 $\hat{V}_A = c^\top \Sigma_A^{-1} c$，其中 $c = \sum_k J_k^{-\top} g_k$。

留出参考事件。在能够访问样本的情况下，(9) 中的期望值本身是对有限参考折求和，因此会增加该求和的方差。每个留出的参考事件 $e \in \mathcal{R}_k$ 贡献 $\omega_k \phi_{k,\ e}$，其中

$$\tilde{\phi}_{k,e} = \tilde{\phi}_k\big(T_{\hat{\nu}_s^{(-k)}} X'_e\big) \tag{13}$$

是变换后的事件处的正交化见证函数，$\omega_k = N_0\, e^{\hat{\nu}_n^{(-k)}}/(K M_k)$ 是留出参考分布事件的权重，M_k 是 $\mathcal{R}_k$中的事件数。根据参考折的检验尺寸，其和的方差通过中心化平方和进行估计

$$\hat{V}_k^R = \omega_k^2 \sum_{e \in \mathcal{R}_k} \big(\tilde{\phi}_{k,e} - \bar{\tilde{\phi}}_k\big)^2, \tag{14}$$

$\bar{\tilde{\phi}}_k$ 与 ϕ_k 的均值 $\phi_{k,\ e}$。方差以中心为基准，这与泊松数据项不同，因为参考折的检验尺寸是固定的，不携带任何

关于数据的信息。在中心值附近，$\hat{}V_k^R$ 约为数据项的 $r = N_0/M$ 倍，因此在 $r = 0.2$ 时忽略它，会使名义水平为 5% 的检验的假阳性率升至约 7%。通过预言机访问，$\hat{}V_k^R = 0$。

见证函数拟合的随机性。 上述各项将每个见证函数视为固定不变。然而，见证函数是随机的，并且共享训练事件：在第 k 折中留出的事件同时也位于 ϕ_j、$j \neq k$ 的训练集内，因此第 j 折和第 k 折的贡献彼此相关。当学习器稳定时（替换一个事件不会使其输出发生太大变化），这些相关性 $\hat{}$ 相对于条件方差可以忽略，且 $V_{\mathrm{fit}} = 0$ 在渐近意义下是正确的。在我们的实验中，在零假设下，拟合得到的见证函数主要是估计噪声，稳定性不足，因此必须将见证函数的跨折协方差纳入考虑。我们使用附录 A.3 中推导的一阶表达式对其进行估计，

$$\hat{V}_{\mathrm{fit}} = \sum_{j \neq k} E_j E_k \,\mathrm{tr}\big(D_k H_k^{-1} \tilde{\Omega}_{kj}\, D_j H_j^{-1} \tilde{\Omega}_{jk}\big), \tag{15}$$

其中

- $E_k = \mathbb{E}_{k,\ \hat{\nu}\,(-k)}\,[1]$ 是第 k 折中预期的留出事件数；
- H_k 是第 k 折见证函数拟合在其自由方向上的海森矩阵——即字典中的活跃原子或网络中未被裁剪的参数；$\tilde{}$
- Ω_{kj} 是在归一化拟合零假设下，折 k 和 j 的见证函数在其自由方向上的梯度的二阶矩；
- D_k 是一个对角矩阵，用于阻尼与零假设拟合共享的方向（附录 C.7）。

仅根据各折的 $\hat{}$ 统计量计算的 K 折统计量方差的任何估计量通常都不可能做到无偏 [17]；V_{fit} 则避免了这一点，因为它利用了拟合学习器对其训练事件的一阶响应，而不仅仅是各折统计量；它仅在一阶上无偏。

3.4 算法与成本

1. 将数据事件以及在具备样本访问权限时的参考事件，随机分配到 K 折中。

2. 对每一折 k：在其他折和辅助测量（公式（4）或公式（6））上拟合干扰参数；在其他折上拟合见证函数 ϕ_k；计算 S_k（公式（9））以及该折对应的方差项（第3.3节）。

ˆ 3. 将各折合并为 V 和 Z （公式（10）），并报告单侧 p-值 $1-\Phi$（Z），其中 Φ 是标准正态累积分布函数。

成本为每个数据集对见证函数进行 K 次拟合，每折一次；其余所有量都根据拟合已提供的量计算，校准无需伪实验。我们始终使用 $K = 10$ 。

3.5 哪些内容有保证，又在哪些情况下可能失效

式（10）的依据是针对交叉验证统计量 [11], 的中心极限定理，并将其推广到包含参考样本和受约束干扰参数的泊松数据集。我们在此非正式地陈述该定理；精确的假设和证明概要见附录 B。

结果。 在 H_0条件下，当事件的期望数随 K 增长且 r固定时， Z 依分布收敛于标准正态分布，因此只要满足以下条件，该检验的渐近检验水平为 α。

- 数据 (D)。数据、参考样本和辅助测量相互独立；给定参考事件数后，参考事件是从参考模型独立抽取的样本。检验以参考折大小为条件；这些大小可以是随机的，但不得依赖事件本身。

- 干扰参数估计 (N)。干扰参数估计 $\hat{\nu}_{(-k)}$ 以通常的速率收敛到 ν^* ，即 $N_0^{-1/2}$ 。

- 见证函数 (W1, W2)。在参考事件所在的区域内，拟合偏离保持有界（W1）。相对于自身输出的检验尺寸，学习器是稳定的（W2）：替换单个训练事件后，在典型点处，见证函数的变化量与$\sigma_\phi/\sqrt{N_0}$相比很小，其中 σ_ϕ 是见证函数的典型尺度（一个确定性尺度，例如其在参考模型下标准差的均方根）。见证函数不必收敛到非零极限：随着数据集增大，σ_ϕ 可以缩小，只要相对于 σ_ϕ测量的单个事件影响，以快于 $N_0^{-1/2}$ 的速度趋于零即可。

ˆ 在这些条件下， V_{fit} 可以忽略，因此可以省略。

该结果有三个特点对其应用至关重要。首先，(W2) 针对的是学习器的输出，而非其参数：经过裁剪、惩罚、提前停止的

模型，或不可辨识模型，只要没有任何单个事件能显著改变其拟合结果，就满足这一条件；正则化也有帮助。其次，留出构造意味着检验以实际得到的见证函数为条件：较差的见证函数会降低检验功效，却不会影响有效性。第三，学习函数 δ_1 在我们的检验中仅作为推前方程 (6) 中估计函数 $\psi = (1,\ \delta_1)$ 的一个组成部分；零假设下的每个期望值，都是通过对参考事件应用精确的干扰变换来计算的，因此无论 δ_1的误差如何，条件 (N) 都成立；该误差只会影响干扰参数估计值的渐近方差。相比之下，在参考文献 [3], 中，学习到的表示同时进入 $t = \tau - \Delta$的两个项，其误差如何影响检验的有效性尚不明确。

退化区域。 在零假设下，只要学习器没有偏离可供学习，却返回了一个纯粹由估计噪声构成的见证函数，其幅度与拟合噪声本身一样小，条件（W2）就可能失效，此时单个事件的影响相对于该见证函数不再足够小。我们的实验中，每个见证函数都出现了这种情况。因此， S 的方差包含上述跨折ˆ相关性，并且需要 V_{fit} 。在此区域内，定理ˆ不适用，我们也没有相应的极限定理。 V_{fit} 校正了条件方差作为 Var(S) 估计值时的一阶偏置。在第5节和第6节的零假设研究中，方差校正统计量 Z 足够接近标准正态分布，因此名义5%和1%水平下的假阳性率仍接近名义水平。

4 个见证函数

第3节的校准将见证函数视为任意学习器的输出。本节说明该学习器必须提供什么，并介绍我们所使用的两种学习器。我们选择了差异尽可能大的两种：一种是带惩罚的线性展开，其拟合属于凸优化问题并具有精确解；另一种是参考文献 [3], 中的小型神经网络，通过随机优化训练固定数量的训练轮次。两者采用相同的校准规则。

4.1 学习器的要求

折 k 的见证函数必须是一个函数 $\phi_k(x)$，能够在任意点求值，并且只能根据训练折和辅助测量计算得出。可以采用基于这些输入的任何选择，包括函数形式、调优方式以及对输出的任何修改，因为校准以实际得到的见证函数为条件，

理论随后要求满足第3.5节中的稳定性和有界性条件（W1）与（W2）。这两项性质针对的是拟合函数，而非参数化；任何形式的正则化（惩罚项、权重裁剪、提前停止）都有助于满足这些条件。

拟合随机性ˆ项 V_{fit} （公式（15））还需要两项信息，它描述见证函数对训练事件的一阶响应：拟合可以自由移动的方向，以及训练损失在这些方向上的曲率。对于最大化扩展似然损失的拟合，这两项都是标准的副产物；第3.3节和附录 C 给出了我们所用见证函数的相关信息。若学习器无法提供这些信息，那么在拟合随机性可以忽略时仍可使用；否则，其ˆ零假设校准还需要另一项 V_{fit}估计值。

我们所有的见证函数都拟合相同的损失函数，即训练折上拟合零假设倾斜后的扩展对数似然：

$$\mathcal{L}_k[f] = \sum_{i \in \mathcal{D}_{-k}} f(X_i) - \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}\left[e^f - 1\right], \tag{16}$$

ˆ在函数类 f 上最大化（可能带有惩罚项），其中 $\phi_k = f$。这是 NPLM 的损失函数，只不过参考分布取代中心值，设在拟合的干扰参数处。

4.2 字典见证函数

第一个见证函数是惩罚展开 $f = \sum_j \beta_j A_j$，它基于预先固定的函数字典（原子函数） $\{A_j\}$ ，并通过惩罚泊松回归拟合；这是一个通过数值计算求解至收敛的凸优化问题。字典包含三个块：切向块，即零假设自身变化的方向（常数项和干扰参数得分函数）；用于描述宽泛偏离的平滑块；这两者均不施加惩罚；以及由高斯凸起函数组成、用于描述局部偏离的隆起块，并施加 ℓ_1 （套索回归）惩罚项 [18]。它可作为黑箱学习器的快速替代方案。字典的选择只影响检验功效，不影响校准。我们在第 5 节的基准测试中报告了该方法，并说明了其原子函数和惩罚项。

4.3 神经网络见证函数

第二个见证函数是参考文献 [3]中的网络：一个带有 Sigmoid 激活函数和线性输出的全连接网络，通过使用 Adam 优化器 [19] 并采用 NPLM 的设置（附录 C.8）最大化公式（16）进行训练，训练

固定数量的训练轮次，并在每一步之后将每个权重裁剪到 $[-c, c]$ 。我们使用与参考文献 [3]相同的架构。

网络见证量可以在参考分布中样本稀疏的尾部自由外推，此时它在罕见留出事件上的取值可能对检验统计量贡献很大。如果对此有所顾虑，我们可以将网络输出乘以渐变窗 $\tau_k(x)$；当拟合零假设在 x 之外有充足样本时，渐变窗的值为 1，并随着 x 增大而降至 0（附录 C.8）。渐变窗仅取决于拟合零假设，因此经渐变窗处理后的输出仍是有效的见证函数。我们将其用于第 5 节的网络；第 6 节的网络则未使用渐变窗，其零假设研究（第 6.3 节）表明那里无需使用渐变窗。

裁剪界、训练时长，以及在使用渐变窗时的渐变窗，是仅有的调优选项，并在第 5 节和第 6 节中与各项基准测试一并列出。与参考文献 [3],不同，这些选项并非为了使统计量服从χ^2 分布而选择，而且所需的训练时长要短得多。

5 一维基准测试

5.1 设置

该基准测试是参考文献 [3], 中的单变量问题，其中参考强度服从指数分布：

$$n(x \,|\, \mathrm{R}_\nu) = N_0 \, \exp\big[- x\, e^{-\nu_s} - \nu_s + \nu_n \big], \qquad x \geq 0, \tag{17}$$

期望事件数为 $N_0 = 2000$ ，归一化干扰参数为 ν_n ，尺度干扰参数为 ν_s。参考文献 [3]中的三个信号叠加在参考分布之上：

NP1：在 $x = 6.4$ 处的高斯峰中包含 10 个期望事件，宽度为 0。16；

NP2：包含 180 个期望事件，具有宽广的形状 $\frac{1}{2}x^2 e^{-x}$；

NP3：在 $x = 1.6$ 处宽度为 0.16 的高斯峰中有 90 个期望事件。

数据集的生成方式如下：背景事件数和信号事件数从泊松分布中抽取，辅助估计值 $\hat{}\nu_A$ 从 $N(\nu^*, \ \Sigma_A)$ 中抽取，其中 Σ_A 为对角矩阵，对角元素为 $\sigma_n^2 = \sigma_s^2 = \sigma_A^2$。与 Ref. [3], 用 2×10^5 个事件的样本表示参考分布不同，我们使用精确密度 (17)，因此这是一个预言机访问的情形。

字典见证函数。在第 k 折中，令 $\hat{}p$ 为拟合零假设 $n(\cdot\,|\,\mathrm{R}\hat{\nu}^{(-k)})/(N_0 e^{\hat{\nu}_n^{(-k)}})$ 的密度，$\hat{}F$ 为其分布函数，$\hat{}q_{0.999}$ 为其 0。999 分位数，$(\hat{}\mu,\ \hat{}\sigma)$ 为其均值和标准差。对于指数分布 (17)，$\hat{}\mu = \hat{}\sigma = e^{\hat{\nu}_s^{(-k)}}$。字典见证函数为

$$f(x) = \beta_0 + \beta_s\,\tilde{s}(x) + \sum_{l=1}^{5} \gamma_l\,P_l(x) + \sum_{j=1}^{J} \beta_j\,B_j(x), \tag{18}$$

具有以下原子函数。

- $\tilde{s}$ 是尺度干扰参数 ν_s的得分函数，经中心化和归一化；其中 ν_s的得分函数为 $\partial_{\nu_s} \log\, n(x\,|\,\mathrm{R}\nu) \;=\; x\,e^{-\hat{\nu}_{(-k)^s}} \;\; -1$；加上常数后，$\tilde{s}$ 张成切向块。

- $P_1, \ldots, P_5$ 构成 $2F(x) - 1$ 中次数至多为 6、且与 1 和 $\tilde{s}$正交的多项式的正交归一基，排列顺序使得P_l 的次数至多为 $l + 1$。

- B_j 是高斯凸起函数 $b_j(x) = \exp\{-(x - c_j)^2/2w^2\}$，其宽度为$w = 0.2\,\hat{}\sigma$，中心为 $c_j = 0.1\,\hat{}\sigma\,(j - 1)$、$j = 1, \ldots, J$，直到 $\hat{}q_{0.999}(J = 70)$；其在 $\{1, \tilde{s}, P_1, \ldots, P_5\}$ 上的投影已移除，并缩放至 $\int B_j^2\,\hat{}p = 1$。

这三个区块在$\hat{}p$下彼此正交，且平滑原子函数和隆起原子函数均有界。系数 β_0、β_s、γ_l 不受惩罚，而 β_j带有等级为 1。3 的 ℓ_1 惩罚项。该等级在任何检验之前，根据零假设模拟预先设定：它是这样一个最小等级的四舍五入值，即在该等级下，平均而言进入零假设拟合的隆起不超过 5%。在每个 ν^* 和 σ_A中均不作任何更改地使用该等级。

网络见证量。该网络采用了参考文献 [3]中的架构：

$$f(x) = b_2 + \sum_{h=1}^{4} W_{2,h}\,\mathrm{sigmoid}(W_{1,h}x + b_{1,h}),$$

其八个权重 $W_{1,\,h}$、$W_{2,\,h}$ 在 $c = 4$ 处进行裁剪，五个偏置均不受限制；训练了 3×10^4 个训练轮次，并使用 $\kappa = 2$ 进行渐变加窗（式（50））。

表2：一维基准测试：名义5%和1%下的假阳性率，每项包含400个数据集（蒙特卡罗标准误约为0.011和0.005）。 ν^* 是干扰参数的真实值：0，或一个干扰参数偏移± 1 个标准差。在 $\sigma_A = 0.10$ 处，仅运行了中心值。

		网络		字典	
σ_A	ν^*	5%	1%	5%	1%
0.05	0	0.0375	0.0025	0.0400	0.0125
0.05	$+\sigma_n$	0.0425	0.0025	0.0475	0.0175
0.05	$-\sigma_n$	0.0400	0.0100	0.0425	0.0100
0.05	$+\sigma_s$	0.0475	0.0025	0.0475	0.0075
0.05	$-\sigma_s$	0.0475	0.0100	0.0475	0.0125
0.10	0	0.0700	0.0200	0.0425	0.0100
0.15	0	0.0500	0.0050	0.0625	0.0200
0.15	$+\sigma_n$	0.0500	0.0200	0.0300	0.0025
0.15	$-\sigma_n$	0.0325	0.0025	0.0650	0.0175
0.15	$+\sigma_s$	0.0475	0.0075	0.0450	0.0075
0.15	$-\sigma_s$	0.0450	0.0100	0.0325	0.0025

5.2 有效性

表2给出了 $\sigma_A \in \{0.05, 0.10, 0.15\}$ 下两个见证函数在名义5%和1%水平上的假阳性率，包括中心值处以及每个干扰参数偏移一个标准差时的结果。对于 $\sigma_A = 0.05$，两个见证函数在全部五个干扰参数点上的假阳性率均不高于其名义水平（考虑蒙特卡罗误差后），5%水平下介于 0.0375 至 0.0475 之间，1%水平下介于 0.0025 至 0.0175 之间。对于 $\sigma_A = 0.15$，两个见证函数在全部五个点上的结果均在名义水平上下约两个标准误差范围内，但波动范围更大：5%水平下介于 0.0300 至 0.0650 之间，1%水平下介于 0.0025 至 0.0200 之间。5%水平下的最大偏差出现在 $\sigma_A = 0.10$，网络的结果为 0.0700（1%水平下为 0.0200），约高于名义水平两个标准误差；在 $\sigma_A = 0.15$，同一网络在每个点的5%水平结果均不高于名义水平，因此该偏差不会随 σ_A 增大。总体而言，检验略显保守：零假设下 Z 的均值在每种配置中均介于 -0.22 和 0.00 之间。

图1展示了网络在五个干扰参数点处的 Z 分布，以及未加入拟合随机性项时得到的统计量。经过校正后，分布符合标准正态分布；未经校正时，分布明显过宽，名义5%水平下的发生率是其名义值的1.5到2倍，这正是第3.5节所述的退化区域。

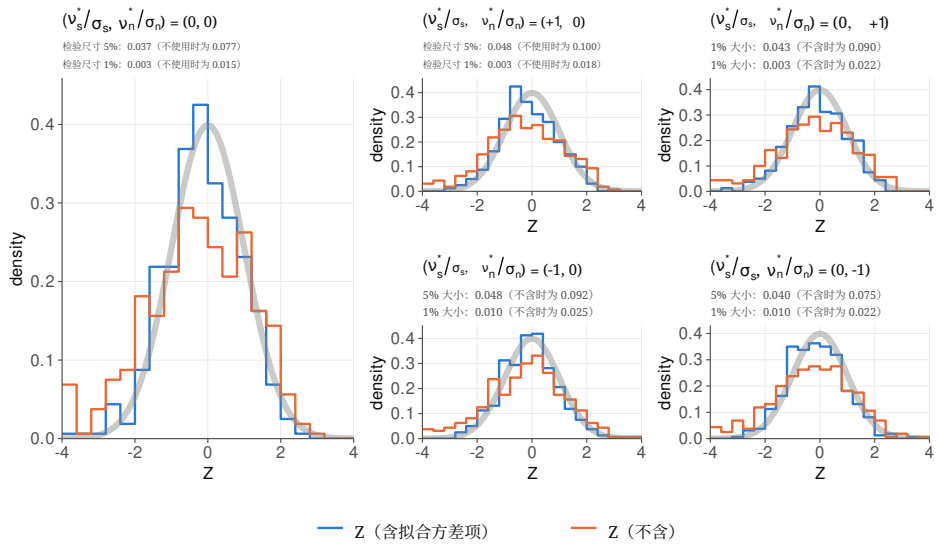

图1：一维基准测试、网络见证量， $\sigma_A = 0.05$：在中心值处（左）以及每个干扰参数偏移 ± 1 个标准差时（右），在零假设下 Z 的分布；分别包含（蓝色）和不包含（橙色）拟合随机性项 V_{fit}，并与标准正态分布的概率密度（灰色）对比。每个面板上方给出了名义5%和1%水平下的拒绝率。每个面板包含400个数据集。

字典对应的图见附录D。

文献 [3] 在相同的干扰参数取值点上，使用100个数据集检验其统计量整体分布的柯尔莫哥洛夫–斯米尔诺夫检验来验证校准；而我们则使用400个数据集直接测量尾部的拒绝率。这两种检查测量的是不同的内容。柯尔莫哥洛夫–斯米尔诺夫检验对分布中心附近最为敏感，而使用100个数据集几乎无法观察尾部，而检验的有效性正是由尾部决定的。

5.3 检验功效

表3给出了中位数 Z、文献 [3], 中中位数 Z-得分 Z 的对应值，以及在中心值和$\sigma_A = 0.05$条件下名义5%和1%时的检验功效。两个见证函数在宽阔的 NP2 和主体峰 NP3 上具有相近的检验功效；而在窄尾峰NP1上，字典的隆起

表3：一维基准测试： Z 的中位数（及其自助法标准误）以及在名义5%和1%水平下的检验功效（拒绝率）；在 $\nu^{*}=0$ 和$\sigma_A=0.05$处，每个条目包含400个数据集。

	网络			字典		
信号	Z的中位数	5%	1%	Z的中位数	5%	1%
NP1	2.41 ± 0.15	0.6450	0.5150	3.08 ± 0.09	0.7600	0.6650
NP2	1.19 ± 0.09	0.3575	0.1525	1.16 ± 0.08	0.3325	0.1700
NP3	3.62 ± 0.11	0.9275	0.8150	3.43 ± 0.07	0.9125	0.7875

宽度与该峰相当，因此表现更佳。文献 [3] 报告称，NP1 和 NP3 的中位数 Z-得分大小相近，而 NP2 的中位数Z-得分更大（1.9 ± 0.12）。更严格的裁剪（例如 $c=3$）会将 NP2 的中位数 Z 提高到约 1.27（第 5.4 节）。

图 2（相当于参考文献 [3]，的图 8）展示了检验功效如何随干扰参数变化。当 σ_A 从 0.05 增至 0.15 时，NP2 的中位数 Z 对两种见证函数都从约 1.2 降至 0.7 到 0.9 之间。共振型 NP1 的中位数 Z 降幅小得多，最多仅下降 0.5；NP3 的中位数则几乎没有变化。参考文献 [3] 得出了相同的趋势和原因：尺度变化可以在尾部模拟宽广超额，因此约束较弱的尺度干扰参数会吸收 NP2 的一部分，但尺度或归一化的任何变化都无法产生窄峰。右面板在固定 σ_A 时改变干扰参数的真实值。归一化的偏移几乎不影响敏感度；尺度的偏移则会使网络（NP1）的敏感度在 1.9 到 3.5 之间变化，因为它改变了信号下方的背景量。由于图中展示的是中位数 Z 本身，而非其与依赖模型的理想见证统计量之比，因此这种对真实尺度的依赖性得以直接呈现。

5.4 对调优的敏感度

图 3 展示了网络的两个调优选项——裁剪界$c \in \{3,4,6\}$ 以及从 2×10^{4} 到 5×10^{4} 个训练轮次的训练时长——在其他条件均固定时的变化；附录D中的表 7 列出了两种见证函数的所有配置。在全部九种配置下，检验尺寸均不高于名义水平，在 5% 水平下最高为 0.045，在 1% 水平下最高为 0.010。检验功效的变化方向符合各信号的形状：更严格的裁剪界有利于宽广的 NP2（中位数 Z 在 $c=3$ 时约为 1.27，而在 $c=6$时为 1.15），更宽松的裁剪界有利于主体峰 NP3（约为 3.4，而另一情况下为 3.9），而 NP1 则受益于更长的

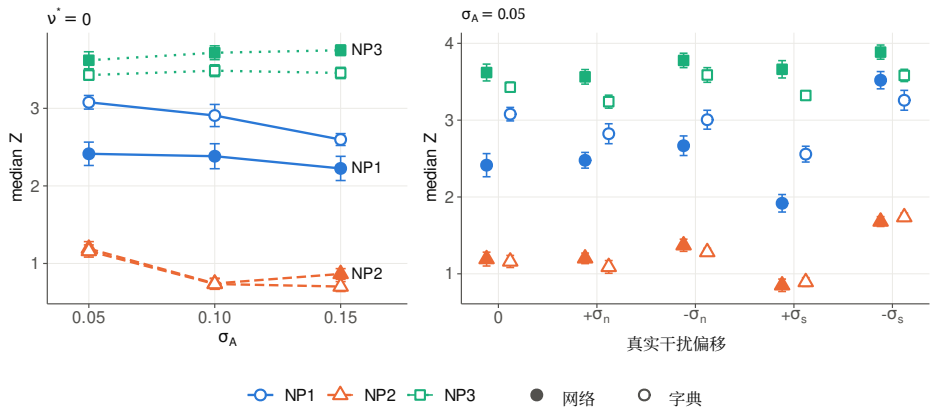

图 2：一维基准测试：NP1–NP3 的中位数 Z （条形：自助法标准误，每个点对应 400 个数据集）。左：中心值$\nu^* = 0$ 随干扰参数不确定性 σ_A的变化。右： $\sigma_A = 0.05$，取中心值，并将每个干扰参数偏移 ± 1 个标准差。不同信号以标记形状区分（NP1 圆形、NP2 三角形、NP3 方形）；左图中还以线型区分（实线、虚线、点线）；实心标记：网络，空心标记：字典。

训练时长。字典在其惩罚水平$\lambda \in \{1.1, 1.3, 1.5\}$下表现出相同的规律，并且使用相同的数据集：检验尺寸在 5% 水平下保持在 0.038 至 0.045 之间；较大的惩罚项使 NP2 的中位数 Z 从 1.09 提高到 1.21，而 NP1 和 NP3 几乎不变。校准不受调优影响，检验功效则会根据见证函数的性质对此作出响应。

6 五维双轻子基准测试

6.1 数据

变量。 第二个基准测试是文献 [3],中作为样本访问案例的双轻子末态。每个事件包含两个轻子，并由以下五个特征 $x = (p_{T,1}, p_{T,2}, \eta_1, \eta_2, \Delta\phi_{12})$ 描述：领头轻子和次领头轻子的横向动量$p_{T,1} \geq p_{T,2}$，即它们动量中垂直于束流轴的分量；它们的赝快度$\eta_{1,2} = -\ln \tan(\theta_{1,2}/2)$，其中 θ 是相对于束流轴的极角；以及它们方位角之差 $\Delta\phi_{12} = \phi_1 - \phi_2$，即它们绕该轴的方位角之差

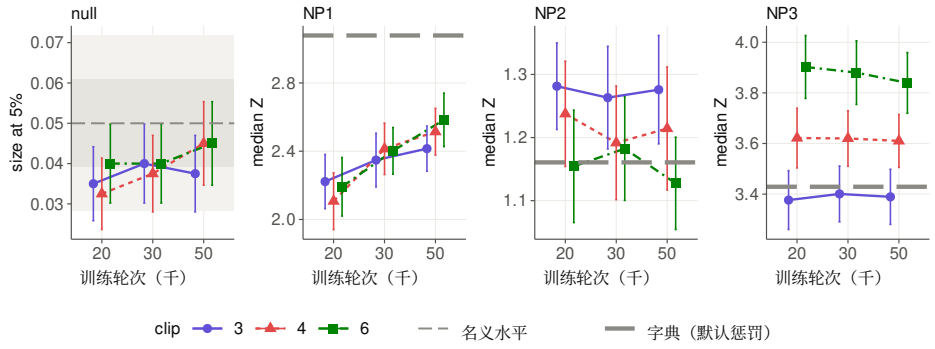

图 3：一维基准测试、网络见证量， $\nu^* = 0, \sigma_A = 0.05$：零假设下名义5%水平的检验率（左），以及针对 NP1– NP3 的中位数 Z 随训练时长的变化，裁剪界为 $c = 3$，4，6（分别以不同颜色、标记和线型表示）。误差棒：蒙特卡罗标准误，每个点使用 400 个数据集。灰色长虚线：名义水平（左）和采用默认惩罚项的字典见证函数（其他面板）。阴影带（左）：围绕有效检验预期名义5%水平的 ± 1 （深色）和 ± 2 （浅色）蒙特卡罗标准误范围。

（取自 $[0, 2\pi)$）。双轻子不变质量是专门的共振搜索会使用的量，它是这些特征的函数，不会单独提供给见证函数。

事件选择。 我们采用参考文献 [3]，的事件选择，即每个事件必须满足以下测量量要求才能纳入分析：$p_T > 20$ 吉电子伏特（两条轻子均须满足），$p_{T,1} < 1$ 太电子伏特（领头轻子须满足），且双轻子不变质量须大于 100 吉电子伏特。中心值为 $\nu = 0$ 时，通过该筛选的标准模型事件预期数量为 $N_0 = 8700$。

干扰参数。 干扰参数包括归一化（ $\sigma_n = 0.025$）以及参考文献 [3]，with$\sigma_s = 3 \times 10^{-3}$ 中类电子区域的轻子动量尺度。该尺度通过第2节的变换 T_{ν_s} 作用于模拟事件：每个轻子的横向动量乘以 e^{ν_s}；对于横动量为 $|\eta| \geq 2.1$ 的轻子，则乘以 $e^{3\nu_s}$ ，然后对变换后的事件进行筛选。只有 ν_s 会改变通过筛选的事件； ν_n 会将预期事件数乘以 e^{ν_n}。

模拟样本。 我们获得了参考文献 [20], 3 735 456 中、在中心干扰参数值下生成的标准模型模拟样本$\mathcal{S}$，共包含若干事件，其生成条件比筛选要求宽松。该样本被随机一次性拆分为两个互不相交的部分，

$$\mathcal{S} = \mathcal{S}_\delta \cup \mathcal{S}_{\text{toy}}, \qquad |\mathcal{S}_\delta| : |\mathcal{S}_{\text{toy}}| = 1 : 3\,. \tag{19}$$

$\hat{\mathcal{S}}_\delta$ 仅用于学习$\delta_{1,\ \text{第6.2节}}$，而 $\mathcal{S}_{\text{toy}}$ 提供参考样本和数据集。对于尺度值 ν_s，令 $\mathcal{P}_{\nu_s}$ 表示变换 $T_{\nu_s} X$ 通过筛选的事件 $X \in \mathcal{S}_{\text{toy}}$，因此 $\mathcal{P}_0$ 是中心值处通过筛选的事件。在 ν 处预期选中的事件数为

$$N_\nu = N_0\, e^{\nu_n}\, \frac{|\mathcal{P}_{\nu_s}|}{|\mathcal{P}_0|}\,.$$

参考样本和零假设数据集。这两类样本发挥着不同的作用。零假设数据集代表观测数据：实验在真实值 ν^*处会记录的数据。它分两步抽取：其事件数服从均值为 N_{ν^*}的泊松分布，然后从 $\mathcal{P}_{\nu_s^*}$中有放回抽取相应数量的事件。参考样本代表检验中的零假设模型：将其加权和变换到任意 ν后，可估计第3.1节中的期望值$\mathbb{E}_{k,\ \nu}$。为了让这些求和能够估计零假设数据的期望值，参考样本也以相同方式抽取，即在中心值处抽取一个零假设数据集，但有两点不同：其样本量固定为 $M = 5N_0 = 43\,500$ 个事件，这些事件来自$\mathcal{P}_0$（因此 $r = 0.2$），并且还包含近阈值事件，其占比与 $\mathcal{S}_{\text{toy}}$ 中的相同（附录 C.1）。信号数据集的构建方式与零假设数据集相同，均基于信号模拟，如下文所述。每个数据集均独立于其参考样本抽取。每种配置下，零假设研究使用 400 个数据集，信号研究使用 160 个。

信号。 这两个信号与参考文献 [3]中的信号相同。对于质量为 300 吉电子伏特的 Z' 共振，数据集由零假设数据集与一个从 Z' 模拟中抽取的样本组成；在中心干扰参数值下，该样本的事件数服从均值为 120 的泊松分布。对于带有$c_W = 1$ 太电子伏特$^{-2}$的有效场论形变，整个数据集均从有效场论模拟中抽取，而不是从标准模型模拟中抽取；在中心干扰参数值下，包含 $1.0216\, N_0$ 期望事件数。在这两种情况下，信号事件都经过与标准模型事件相同的$T_{\nu_s^*}$ 变换，其期望事件数也以相同方式缩放，因为失准的动量尺度会影响每个轻子。与参考文献 [3], 一样，信号数据集是在尺度干扰参数偏离其中心值一个标准差的情况下生成的，即 $\nu_s^* = +\sigma_s$。

6.2 估计

干扰函数。 函数 $\delta_{1,2}$ 是在 $\mathcal{S}_\delta$ 上按照参考文献 [3]中的方法学习得到的。对于尺度干扰参数的训练点 $\nu_i \in \{\pm 1.5, \pm 3, \pm 6, \pm 10\} \times 10^{-3}$，令$\mathcal{S}_1$ 表示 $\mathcal{S}_\delta$ 中在 $\nu = 0$ 下通过筛选的事件，令 $\mathcal{S}_0(\nu_i)$ 表示经过变换 T_{ν_i}后通过筛选的事件。两个网络 δ_1, δ_2 （各包含五个隐藏层，每层十个单元）使下式最小化

$$
\begin{aligned}
&\sum_{i} \Big[\sum_{x \in \mathcal{S}_0(\nu_i)} c_i(x)^2 + \sum_{x \in \mathcal{S}_1} \big(1 - c_i(x)\big)^2 \Big], \\
&c_i(x) = \frac{1}{1 + \exp\{\nu_i\, \delta_1(x) + \tfrac{1}{2}\nu_i^2\, \delta_2(x)\}},
\end{aligned}
\tag{20}
$$

其最小化函数使 $\nu_i \delta_1 + \tfrac{1}{2}\nu_i^2 \delta_2$ 成为偏移事件密度与名义事件密度之比的对数。在 2000 个训练轮次中，留出五分之一的事件，用于选取验证损失最低的网络。在这两个函数中，我们的检验仅使用 δ_1：在推前拟合的估计函数 $\psi = (1,\ \delta_1)$（公式（6））以及正交化见证函数（公式（11））中使用它。

见证函数。 该见证函数是第4.3节中的网络，采用文献 [3], 的架构：三个隐藏层，每层五个单元（96 个参数），所有参数均裁剪至 $c = 0.7$，训练 3000 个训练轮次，使用 $K = 10$ 折；特征按照文献 [2]进行标准化。在这一高维设置中，字典见证函数在探索性研究中明显弱于网络，因此我们不予报告。

6.3 有效性

表4和图4给出了 Z 的零分布：在中心值处，以及每个干扰参数偏移一个标准差时的分布。这五个点是参考文献 [3] 用于验证自身校准的点。在名义5%水平下，发生率介于0.0350和0.0575之间；在名义1%水平下，介于0.0050和0.0150之间，均与名义水平相差约一个半蒙特卡罗标准误以内。Z 的均值与零相差不超过0.08，标准差介于0.93和0.97之间，因此该检验即便有偏，也只是略微保守。若不计入拟合随机性项，名义5%水平下的发生率将升至0.09至0.11之间（图4）：这里和一维情形一样存在退化区域，而修正项$\hat{V}_{\mathrm{fit}}$（包括其参考分量）消除了这种现象。

表4：双轻子基准测试，网络见证量：零假设下 Z 的均值和标准差，以及在中心值和各干扰参数偏移一个标准差时，其在名义5%和1%水平下的比率。每行400个数据集（蒙特卡罗标准误约为0.011和0.005）。

ν^*	Z的均值	Z的标准差	5%水平下的比率	1%水平下的比率
0	-0.012	0.965	0.0575	0.0125
$+\sigma_s$	-0.012	0.928	0.0525	0.0150
$-\sigma_s$	-0.075	0.928	0.0350	0.0125
$+\sigma_n$	0.016	0.942	0.0425	0.0125
$-\sigma_n$	0.005	0.949	0.0375	0.0050

表5：双轻子基准测试：网络见证量的敏感度与参考文献 [3], 中类电子区域 $\nu_s^* = +\sigma_s$的比较。我们的中位数 Z （自助法标准误）和检验功效基于160个数据集；NPLM的中位数 Z-得分 Z 如参考文献 [3] 所示（100个数据集）。

	本工作（神经网络见证量）			NPLM
信号	中位数 Z	检验功效 5%	检验功效 1%	$\overline{Z}$
Z'	2.80 ± 0.13	0.84	0.66	3.9 ± 0.2
EFT	7.00 ± 0.13	1.00	1.00	5.1 ± 0.3

6.4 检验功效： Z' 和有效场论

图5展示了两种信号下 Z 的分布，表5则将其中位数与参考文献 [3]中的相应数值进行了比较。对于窄 Z' 共振，我们的网络不如NPLM敏感（中位数 Z 2.80，而NPLM的中位数 Z-得分为 $Z = 3.9$；两者均由检验自身的p值换算而来）；对于宽EFT形变，我们的网络则更敏感（7.00，对比5.1）。这与第5节所述的模式一致：两种信号的校准相同，而检验功效取决于见证函数——此处是一个经过3000个训练轮次、采用严格裁剪训练的小型网络——能够有效捕捉什么。平滑、宽广的形变容易拟合；而在五维空间的稀疏区域中，窄共振则较难分辨。

6.5 对调优和计算成本的敏感度

图6改变了裁剪界 $c \in \{0.7, 1.0, 1.3\}$和训练时长（从2000到5000个训练轮次）；附录D中的表8列出了每种配置。

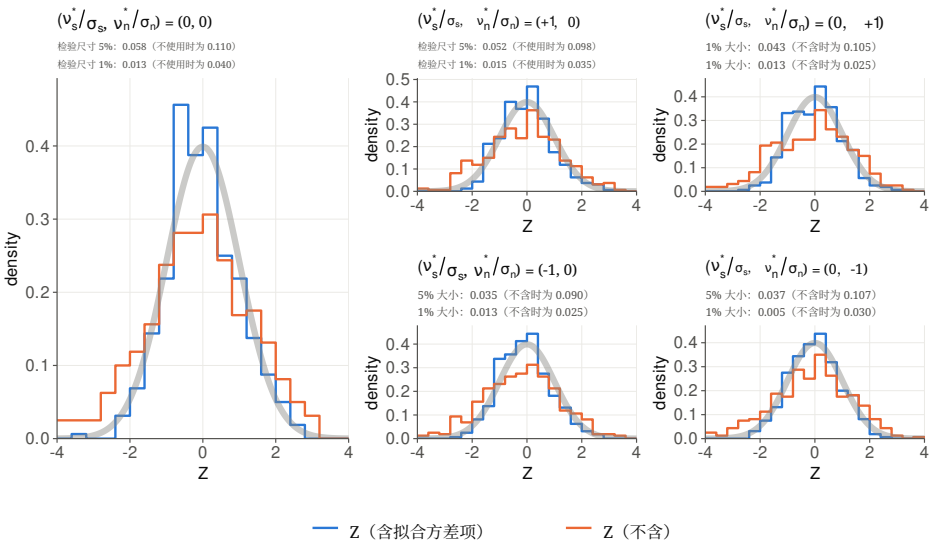

图4：双轻子基准，网络见证函数：在中心值处的零假设下（左）以及每个干扰参数偏移一个标准差时（右），Z 的分布；分别展示含（蓝色）和不含（橙色）拟合随机性项的结果，并与标准正态密度（灰色）进行对比。每个面板包含400个数据集。

在全部九种配置中，名义5%水平下的比率均保持在0.040至0.065之间，与名义水平相差约一个半蒙特卡罗标准误。检验功效再次响应于见证函数：Z' 的中位数 Z 几乎没有变化（介于2.68和2.92之间），而有效场论对应的检验功效则随裁剪界增大，从 $c = 0.7$ 时的约7.0上升到 $c = 1.3$时的8.7至9.5。因而，我们的默认$c = 0.7$ 对于有效场论而言是保守的；放宽裁剪界将进一步提高其敏感度，同时不会影响校准。

网络可以进行短时间训练，而无需达到 χ^2 状态：此处训练3000轮次，而参考文献 [3], 中则需要 5×10^5 或更多轮次，且单次训练需要数小时的CPU时间。在一块GPU上，训练一个双轻子数据集的十个网络约需半分钟。

7 讨论

我们提出了一种用于新物理搜索的拟合优度检验，其中灵活学习器在部分数据上拟合相对于参考模型的偏离

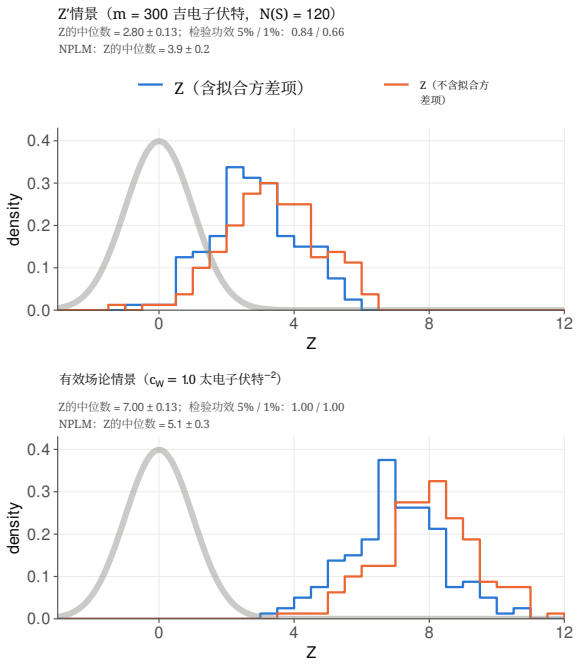

图5：双轻子基准，网络见证函数：在 $\nu_s^* = +\sigma_s$ 处， Z 信号（上）和有效场论（下）信号下，Z' 的分布；包含（蓝色）和不包含（橙色）拟合随机性项；灰色表示零假设下预期的标准正态密度。相同信号下的中位数 Z 和检验功效，以及文献 [3] 的中位数 Z-得分，分别标于各面板上方。每个面板包含160个数据集。

，并在其余数据上进行评估，通过交叉拟合确保每个事件都同时用于这两者；同时采用解析方法进行校准，并明确纳入估计的干扰参数、有限参考样本以及拟合随机性的影响。在文献 [3], 的基准测试中，我们检查的每种配置下，假阳性率都保持在接近名义水平的位置。在两个基准测试中，无论取中心值，还是将每个干扰参数移动一个标准差，这一结论都成立。在一维基准测试中，对于不超过 $\sigma_A = 0.15$的干扰参数不确定性，这一结论同样成立。校准过程中没有使用任何伪实验；本文中的伪实验仅用于检验校准结果。

校准带来的益处。核心结果是，学习器的选择不再受校准限制。我们的校准以

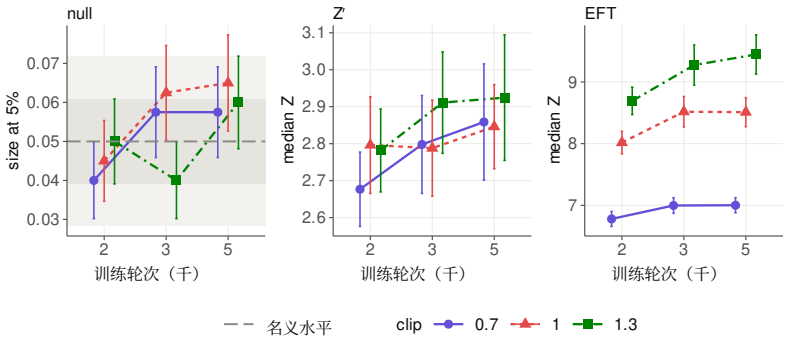

图6：双轻子基准、网络见证量：零假设下名义5%的检验率（左图；每个点400个数据集），以及 Z for Z'和有效场论的中位数（每个点160个数据集）随训练时长的变化；裁剪界为 $c = 0.7$、1。0，1。3，分别以各自的颜色、标记和线型表示。柱状线：蒙特卡罗标准误。阴影带（左图）：围绕有效检验预期的名义5%的 ± 1（较深）和 ± 2（较浅）蒙特卡罗标准误范围。

已实现拟合为条件，因此在下文讨论的限制范围内，学习器本身就是可接受的，其调优可以只考虑检验功效。由此得出两个结果。检验功效取决于学习器捕捉信号的能力，这正是应有的表现；而最佳调优方式因信号而异：在一维情况下，更严格的裁剪界或更大的惩罚项有助于宽广的 NP2，而更宽松的裁剪界有助于主体峰 NP3；在五维情况下，更宽松的裁剪界有助于有效场论形变。整个过程中，校准并未改变。而且拟合所需时间很短，因为没有任何要求规定它们必须达到 χ^2 状态。

可预测性、可计算性与稳定性。我们的预测性拟合优度检验的构建过程，可以通过真实数据科学的可预测性、可计算性与稳定性（PCS）框架来理解 [21, 22]。可预测性是检测的基础：只有当拟合偏离能够改善对未来事件的预测时，它才算作证据。可计算性体现在成本上：我们直接采用正则化和提前停止，而不是不断调参直到统计量达到某种 χ^2 状态；校准过程中也不涉及伪实验或重采样。稳定性 [23] 是校准的出发点，依据的是交叉验证统计量的中心极限定理，但这并非全部：在零假设下，我们的学习器不够稳定，仅靠该理论还不足以完成校准；还需要通过

对拟合随机性的校正。

局限性与展望。 我们列出上述结果所依赖的近似或尚不充分证据之处，以及仍有待进一步研究的方向。

- **退化区域。** 在零假设下，我们的学习器不够稳定，因此第3.5节的定理无法直接适用，校准依赖于跨折校正 $V_{\text{拟合}}$。这种校正是一阶近似。没有极限定理涵盖这一情形，正态参照则基于第5节和第6节的零假设研究。对于网络，其线性化将每个参数视为自由参数或达到裁剪界的参数，并采用一个预先固定的正则化常数（附录 C.8）。在我们使用的配置下，这种线性化足够准确；但当许多参数达到界限时，它可能低估拟合方差。

- **见证构造。** 检验功效取决于见证函数捕捉信号的能力，而我们并未针对检验功效专门设计见证函数。留出评估和较短的训练在某些信号上会带来代价：在我们选定的工作点上，我们的网络对 Z' 共振的灵敏度低于 NPLM；对于宽广的一维过量 NP2，我们的两个见证函数也都不如 NPLM 灵敏。相反，字典中隆起的宽度设定在窄基准信号的尺度上，这可能使其在这些信号上的检验功效更高。由于校准以每一折生成的见证函数为条件，该框架仍允许进一步设计见证函数，而我们尚未探索这方面：使用更大的网络或进行更长时间的训练；构建针对目标信号调整的字典；采用数据驱动的调优，在每一折内仅用该折的训练事件进行验证，以此选择裁剪界或惩罚项；或使用其他学习器，例如已与 NPLM 配合使用的核方法 [24]。NPLM 自身的架构可以保持原样使用，如本文所示，这样两种方法便可共用一个学习器，仅在校准方式上有所不同。最后，无玩具模拟校准使扫描数据集中的多个学习器或多个区域切实可行，前提是对由此产生的检验多重性进行适当处理。

- **两个干扰参数。** 两个基准都包含一个归一化干扰参数和一个尺度干扰参数。该构造适用于任意有限个干扰参数集合，但现实中的分析包含更多干扰参数，而多个干扰参数下推前

拟合和方差项的成本尚未在此进行研究。

数据与代码可用性

五维双轻子样本可公开获取 [20]；一维数据集为模拟数据。本文所有模拟结果、表格和图的复现代码均可在 https://github.com/jinglei/predictive-gof 上公开获取。

A 校准公式详述

本附录汇集了式 (10) 中各项的计算公式，并使用一阶近似推导正交化见证函数 (11) 和拟合随机性项 (15)。A.1–A.3 节假设可进行预言机访问；A.4节列出在样本访问情况下的变化。

A.1 估计方程

每一项都需要计算折的拟合零假设下 x 的函数期望值。对于函数 u，我们将其写作

$$\mathbb{E}_{k,\hat{\nu}^{(-k)}}[u] = \frac{1}{K}\int \hat{n}_k\, u, \qquad \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}[u] = \frac{K-1}{K}\int \hat{n}_k\, u, \tag{21}$$

折 k 的留出事件和训练事件上 u 的期望和，以及在拟合零假设 $\hat{n}_k$ 下留出事件期望数量的 $E_k = \mathbb{E}_{k,\ \hat{\nu}^{(-k)}}[1]$ 。$\mathbb{E}_{k,\ \nu}$ 和 $\mathbb{E}^{\mathrm{tr}}_{k,\ \nu}$ 表示将 $\hat{n}_k$ 替换为 $n(\cdot \mid \mathrm{R}\nu)$ 后相应的期望值。数据事件上的相应经验和为 $\hat{\ }\ \mathbb{E}_k[u] = \sum_{i\in\mathcal{D}_k} u(X_i)$ 和 $\hat{\ }\ \mathbb{E}^{\mathrm{tr}}_k[u] = \sum_{i\in\mathcal{D}_{-k}} u(X_i)$。

第3.1节的两个干扰参数拟合都求解以下形式的估计方程

$$F_{k,\psi}(\nu) := \hat{\mathbb{E}}^{\mathrm{tr}}_k[\psi] - \mathbb{E}^{\mathrm{tr}}_{k,\nu}[\psi] - \Sigma_A^{-1}(\nu - \hat{\nu}_A) = 0, \tag{22}$$

其中，ψ 是一个估计函数向量，对于每个干扰参数各有一个分量；这些函数可能依赖于 ν，也可能不依赖。记

$$J_k = \left\{\partial_\nu\, \mathbb{E}^{\mathrm{tr}}_{k,\nu}[\psi_\nu] - \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}[\partial_\nu \psi_\nu] + \Sigma_A^{-1}\right\}_{\nu=\hat{\nu}^{(-k)}} \approx -\left.\frac{\partial F_{k,\psi}}{\partial \nu}\right|_{\nu=\hat{\nu}^{(-k)}} \tag{23}$$

对于近似的符号反转雅可比矩阵of $F_{k,\ \psi}$ 其中将数据总和替换为拟合零假设下的期望；ψ_ν 中的下标明确表示

ψ对 ν的任何依赖。对强度求导， $\partial_\nu n(\cdot\,|\,\mathrm{R}\nu) \;= n(\cdot\,|\,\mathrm{R}\nu)\; s$其中 $s = \partial_\nu \log\, n(\cdot\,|\,\mathrm{R}\nu)$为得分函数，第一项为$\mathbb{E}^{\mathrm{tr}}_{k,\;\hat{\nu}^{(-k)}}[\psi s^\top] + \mathbb{E}^{\mathrm{tr}}_{k,\;\hat{\nu}^{(-k)}}[\partial_\nu \psi_\nu]$， $\partial_\nu \psi_\nu$中的项相互抵消，并且

$$J_k = \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}\left[\psi\, s^\top\right] + \Sigma_A^{-1}, \tag{24}$$

其中行对应各个方程，列对应干扰参数。两种拟合的差异仅在于 ψ。在具有预言机访问的情况下，似然拟合具有 $\psi = s$，并且

$$J_k = \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}\left[s\, s^\top\right] + \Sigma_A^{-1}, \tag{25}$$

训练拟合的费舍尔信息（包括辅助项），因此 J_k^{-1} 是$\hat{}\nu^{(-k)}$的渐近方差。该 J_k是对称的，因此在下文中 $J_k^{-\top} = J_k^{-1}$；保留转置是因为它们对于使用样本访问的推前拟合（A.4节）很重要。

A.2 正交化见证函数的推导

$\hat{}$按照A.1节的记号，公式（9）写作$S_k \;=\; \mathbb{E}_k[\phi_k] - \mathbb{E}_{k,\;\hat{\nu}^{(-k)}}[\phi_k]$。目前我们将见证函数视为固定的，并将其自身随机性的研究推迟到A.3节。 S_k 随后通过留出数据的求和，以及通过干扰参数估计$\hat{}\nu^{(-k)}$，依赖于数据、训练事件和辅助测量。在 ν^* 附近展开至一阶，

$$S_k = \left\{\hat{\mathbb{E}}_k[\phi_k] - \mathbb{E}_{k,\nu^*}[\phi_k]\right\} - g_k^\top\left(\hat{\nu}^{(-k)} - \nu^*\right) + \ldots, \tag{26}$$

其中

$$g_k = \partial_\nu\, \mathbb{E}_{k,\nu}[\phi_k]\big|_{\nu=\hat{\nu}^{(-k)}} = \mathbb{E}_{k,\hat{\nu}^{(-k)}}[\phi_k\, s]\,.$$

(26) 中的第一项是以其零假设期望为中心的留出和。对于第二项，$F_{k,\;\psi}(\hat{}\nu^{(-k)}) \;= 0$和 $-\partial F_{k,\;\psi}/\partial\nu \approx J_k$，因此围绕 ν^* 将 (22) 线性化，得到

$$\hat{\nu}^{(-k)} - \nu^* = J_k^{-1} F_{k,\psi}(\nu^*) + \cdots = J_k^{-1}\left[\hat{\mathbb{E}}^{\mathrm{tr}}_k[\psi] - \mathbb{E}^{\mathrm{tr}}_{k,\nu^*}[\psi] + \Delta_A\right] + \ldots, \tag{27}$$

其中 $\Delta_A \;=\; \Sigma_A^{-1}\,(\hat{}\nu_A - \nu^*)$。第 k 折的训练总和等于其他各折的留出和之和，即 $\hat{}\;\mathbb{E}^{\mathrm{tr}}_k \;=\; \sum_{j\neq k}\hat{}\mathbb{E}_j$，同样地 $\mathbb{E}^{\mathrm{tr}}_{k,\;\nu^*} \;= \sum_{j\neq k}\mathbb{E}_{j,\;\nu^*}$，因为在共同的 ν 下，期望与折无关。将（27）代入（26），对 k 求和，并按照其事件被留出的折 j 对训练项重新分组，

$$S = \sum_{j=1}^{K}\left\{\hat{\mathbb{E}}_j[\tilde{\phi}_j] - \mathbb{E}_{j,\nu^*}[\tilde{\phi}_j]\right\} - c^\top\Delta_A + \ldots, \tag{28}$$

with

$$\tilde{\phi}_j = \phi_j - \Big(\sum_{k\neq j} J_k^{-\top} g_k\Big)^{\top}\psi, \qquad c = \sum_{k=1}^{K} J_k^{-\top} g_k. \tag{29}$$

在领先阶中，$J_k^{-\top} g_k$ 与 k 无关，因为各折的拟合仅因交换了比例为 $1/K$ 的事件而有所不同，因此 $\sum_{k\neq j} J_k^{-\top} g_k \approx (K-1) J_j^{-\top} g_j = b_j$，而 $\tilde{\phi}_j = \phi_j - b_j^{\top}\psi$ 是公式（11）中的正交化见证函数，即实际计算所采用的形式。

表 6 按照每次一个噪声单位的方式解读式 (28)。在第 j 折中留出的数据事件 X_i，通过其在 S_j 中的评估值 $\phi_j(X_i)$ 产生影响；同时，它还会通过其他 $K-1$ 折的干扰参数拟合产生影响，并使每一折偏移 $J_k^{-1}\psi(X_i)$；这些影响合在一起，得到 $\tilde{\phi}_j(X_i)$。辅助测量从不被留出，因此没有评估值；它会进入所有 K 折的拟合，其偏移由一个随机向量驱动，并相干地累加到 $-c^{\top}\Delta_A$中。

表 6：见证函数固定时，每种噪声来源的一个单位对 S 的贡献。数据事件属于第 j 折，并进入其余 $K-1$ 折 $k \neq j$ 的干扰参数拟合；辅助测量则进入所有K 折的拟合。中间各列表示该单位会进入其拟合的每一折 k；最后一列则如式 (28) 所示，将评估值与所有此类偏移求和。采用样本访问时，会增加第三行（A.4节）。

来源	评估值偏移$\hat{\nu}^{(-k)}$		S的偏移$_k$	总计
数据事件 X_i	$\phi_j(X_i)$	$J_k^{-1}\psi(X_i)$	$-g_k^{\top} J_k^{-1}\psi(X_i)$	$\tilde{\phi}_j(X_i)$
辅助测量	—	$J_k^{-1}\Delta_A$	$-g_k^{\top} J_k^{-1}\Delta_A$	$-c^{\top}\Delta_A$

使用拟合见证函数的方差。写出

$$U_j \coloneqq \hat{\mathbb{E}}_j[\tilde{\phi}_j] - \mathbb{E}_{j,\nu^*}[\tilde{\phi}_j]$$

使得（28）变为

$$S = \sum_j U_j - c^{\top}\Delta_A + \dots .$$

U_j 通过对 $\mathcal{D}_j$ 中的事件进行评估以及其他折的干扰参数拟合，汇总这些事件的贡献。令 $\mathcal{F}_{-j}$ 表示除第 j折中的事件之外的所有内容：其他折和辅助测量。见证函数 ϕ_j、其敏感度g_j、 J_j 以及由此计算得到的 $\tilde{\phi}_j = \phi_j - b_j^{\top}\psi$ 都是函数

关于 $\mathcal{F}_{-j}$的函数，而 $\mathcal{D}_j$ 中的事件独立于 $\mathcal{F}_{-j}$，因此 $\mathbb{E}[U_j \,|\, \mathcal{F}_{-j}] = 0$。将 c 视为固定值， U_j 与 $c^\top \Delta_A$ 的协方差因而为零，并且

$$\mathrm{Var}(S) \approx \sum_j \mathrm{Var}(U_j) + \mathrm{Var}(c^\top \Delta_A) + \sum_{j \neq k} \mathrm{Cov}(U_j, U_k)\,. \tag{30}$$

根据全方差定律，对每一项分别在其自身的条件下应用，可得 $\mathrm{Var}(U_j) = \mathbb{E}\big[\mathrm{Var}(U_j \,|\, \mathcal{F}_{-j})\big]$，其代入估计为 $\hat{V}_j^D$。辅助项为 $\mathrm{Var}(c^\top \Delta_A) = c^\top \Sigma_A^{-1} c = V_A$，因为 $\mathrm{Var}(\Delta_A) = \Sigma_A^{-1}$ 。对于拟合得到的$\tilde{}$见证函数，交叉项不会消失，因为 ϕ_j 是在折 k 的事件 $\tilde{}$上训练的，而 ϕ_k 是在折 j的事件上训练的；它们就是 A.3节中讨论的拟合随机性项。

A.3 拟合随机性项

$\hat{}$ 对于固定的折 k，见证函数为 $\phi_k = f_{\hat{\theta}_k}$，其中 θ_k 在参数向量 θ上最大化训练目标函数 (16)，并且可能带有惩罚项。按照 A.1节的记号，该目标函数写为

$$\mathcal{L}_k(\theta) = \hat{\mathbb{E}}_k^{\mathrm{tr}}[f_\theta] - \mathbb{E}_{k,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[e^{f_\theta} - 1\big], \tag{31}$$

$\hat{}$而 θ_k 满足估计方程

$$G_k(\theta) \coloneqq \partial_\theta \mathcal{L}_k(\theta) = \hat{\mathbb{E}}_k^{\mathrm{tr}}[a_\theta] - \mathbb{E}_{k,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[e^{f_\theta} a_\theta\big] = 0, \tag{32}$$

其中， $a_\theta = \partial_\theta f_\theta$ 被限制在折 k 的自由方向上（字典中的活跃原子、网络的自由参数；附录 C），并且$a_k = a_{\hat{\theta}_k}$。与 (22) 中一样， G_k 是数据和减去期望，此处的期望是在倾斜零假设 $e^{f_\theta}\hat{}\,n_k$下计算的。我们写为

$$\begin{aligned} H_k &= \Big\{ \partial_\theta\, \mathbb{E}_{k,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[e^{f_\theta} a_\theta\big] - \mathbb{E}_{k,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[e^{f_\theta} \partial_\theta a_\theta\big] \Big\}_{\theta=\hat{\theta}_k} \\ &\approx -\frac{\partial G_k}{\partial \theta}\Big|_{\theta=\hat{\theta}_k} \end{aligned} \tag{33}$$

对于 G_k 的近似符号反转雅可比矩阵，其中数据和 $\hat{}\,\mathbb{E}_k^{\mathrm{tr}}[\partial_\theta a_\theta]$ 被其在拟合的倾斜零假设下的期望所替代。这是式 (23) 中 J_k 的见证函数对应项。类似的抵消得出

$$H_k = \mathbb{E}_{k,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[e^{\phi_k} a_k a_k^\top\big]\,. \tag{34}$$

对于字典， a_θ 是活跃原子函数的向量，且不依赖于 θ，因此没有需要抵消的项， H_k 就是带惩罚的泊松回归在其活跃列上的海森矩阵。对于网络，抵消项就是高斯–牛顿近似，而 H_k 是带有岭惩罚的高斯–牛顿矩阵（49）。

一阶协方差。在 H_0 下，拟合得到的见证函数接近一个满足 $f_{\theta_0} \equiv 0$的参数 θ_0。如同式（27）对（22）进行线性化一样，对（32）在 θ_0附近进行线性化，$\hat{}\theta_k - \theta_0 = H_k^{-1} G_k(\theta_0) + \dots$。我们以真实值为中心，因此每个求和仅依赖于一折中的事件：

$$\xi_j(u) = \hat{\mathbb{E}}_j[u] - \mathbb{E}_{j,\nu^*}[u] \tag{35}$$

是函数 u 在第 j折的事件上的中心化和；$\xi_j(a_k)$ 是第 k 折自由方向的中心化和，而 $G_k(\theta_0) = \sum_{j \neq k} \xi_j(a_k) - \mathbb{E}_{k,\,\hat{\nu}^{(-k)}}^{\mathrm{tr}}\big[a_k\, s^\top\big]\big(\hat{}\nu^{(-k)} - \nu^*\big) + \dots$。第二项是被干扰参数拟合吸收的训练波动部分，见证函数正是针对这一部分进行拟合。此处我们暂不考虑它，并在计算中通过对干扰方向施加阻尼来处理（附录 C.7）；于是

$$\hat{\theta}_k - \theta_0 \approx H_k^{-1} \sum_{j \neq k} \xi_j(a_k). \tag{36}$$

见证函数在一阶近似下为 $\phi_k = a_k^\top(\hat{}\theta_k - \theta_0)$。从 A.2 节中 U_k 的定义出发，并将 b_k 视为给定 $\mathcal{F}_{-k}$ 时的固定量，

$$\begin{aligned} U_k &= \xi_k(\tilde{\phi}_k) \\ &= \xi_k(\phi_k) - b_k^\top \xi_k(\psi) \\ &= \xi_k(a_k)^\top(\hat{\theta}_k - \theta_0) - b_k^\top \xi_k(\psi) + \dots \\ &= \xi_k(a_k)^\top H_k^{-1} \sum_{j \neq k} \xi_j(a_k) - b_k^\top \xi_k(\psi) + \dots \end{aligned} \tag{37}$$

第一行是按照 (35) 的记号对 U_k 的定义；第二行利用 ξ_k 的线性性质，将 $\tilde{\phi}_k = \phi_k - b_k^\top \psi$ 拆开；第三行代入线性化见证统计量 $\phi_k = a_k^\top(\hat{}\theta_k - \theta_0)$；最后一行使用 (36)。第一项包含拟合噪声：第 k 折的留出事件与训练事件配对。对于 $j \neq k$，U_j 和 U_k 共享两组这样的配对：第 k 折的事件在 U_k 中作为留出事件，在 U_j 中作为训练事件；第 j 折的情况则相反。由于各折是相互独立的泊松过程，且每个中心化和的均值为零，因此只有这些配对有贡献：

$$\begin{aligned} \operatorname{Cov}(U_j, U_k) &= \operatorname{Cov}\big(\xi_j(a_j)^\top H_j^{-1} \xi_k(a_j),\ \xi_k(a_k)^\top H_k^{-1} \xi_j(a_k)\big) \\ &= \operatorname{tr}\Big(H_j^{-1}\, \mathbb{E}\big[\xi_k(a_j)\, \xi_k(a_k)^\top\big]\, H_k^{-1}\, \mathbb{E}\big[\xi_j(a_k)\, \xi_j(a_j)^\top\big]\Big) \\ &= E_j E_k \operatorname{tr}\big(H_k^{-1} \Omega_{kj} H_j^{-1} \Omega_{jk}\big). \end{aligned} \tag{38}$$

第二行将两个标量的乘积写成 $x^\top A y\, z^\top B w = \mathrm{tr}(A\, y z^\top B\, w x^\top)$，并逐折取期望：$\xi_k(a_j)$ 和 $\xi_k(a_k)$ 取决于第 k 折的事件，$\xi_j(a_k)$ 和 $\xi_j(a_j)$ 取决于第 j 折的事件，而各折彼此独立。最后一行使用坎贝尔定理

$$\mathbb{E}\big[\xi_k(a_j)\,\xi_k(a_k)^\top\big] = \mathrm{Cov}_{k,\nu^*}(a_j, a_k) = E_{k,\nu^*}[a_j a_k^\top] = E_{k,jk}\Omega,$$

以及迹的循环性质；在理想情况下，

$$\Omega_{kj} = \int a_k\, a_j^\top\, dQ^*, \qquad Q^* = \frac{n(\cdot \mid \mathrm{R}_{\nu^*})}{\int n(\cdot \mid \mathrm{R}_{\nu^*})}, \tag{39}$$

是归一化真实零假设下各折 k 和 j 自由方向的二阶矩；其计算方式见附录 C.6。将 H_k^{-1} 替换为 $D_k H_k^{-1}$（附录 C.7 中的阻尼）后，这就是公式（15）中的被加项。当所有折共享相同的 p 未惩罚方向，且这些方向在 Q^* 下正交归一时，对 $j \neq k$ 的求和在主阶上等于条件方差$\sum_k \hat{V}_k^D$：在线性情形下，拟合随机性会使方差加倍。对于这种已知干扰参数的线性见证函数，标准线性代数表明，给定拟合的一阶展开，该修正是精确的。

A.4 样本访问：有哪些变化

在样本访问情形下，参考模型只能通过模拟样本获知，而 A.1–A.3 节中每一个被计算（而非仅被写出）的期望 $\mathbb{E}_{k,\,\nu}$ 或 $\mathbb{E}^{\mathrm{tr}}_{k,\,\nu}$ 都会被其参考估计$\bar{\mathbb{E}}_{k,\,\nu}$ 或 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\,\nu}$所替代。差值 $\bar{\mathbb{E}} - \mathbb{E}$ 就是参考噪声。它为表 6 增添了第三个噪声来源，即方差中的项 $\hat{V}_k^R$，以及重叠量 V_{fit}中的参考 $\hat{}$ 部分。

参考期望值。在 T_{ν_s} 依次进行干扰变换、事件筛选和 M_k后，M_{-k} 在 $\nu = 0$ 通过筛选的留出参考事件和训练参考事件数量，

$$\bar{\mathbb{E}}_{k,\nu}[u] = \frac{N_0\, e^{\nu_n}}{K\, M_k} \sum_{e \in \mathcal{R}_k} \mathbf{1}[T_{\nu_s} X'_e \text{ passes}]\, u(T_{\nu_s} X'_e), \tag{40}$$

而 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\,\nu}$ 则是对 $\mathcal{R}_{-k}$ 的相同求和，并采用权重 $(K-1)N_0 e^{\nu_n}/(K\, M_{-k})$。在拟合值 $\nu = \hat{\nu}^{(-k)}$ 处，留出权重为 $\omega_k = N_0\, e^{\hat{\nu}_n^{(-k)}}/(K\, M_k)$，其中 $\hat{\nu}_n^{(-k)}$ 是 $\hat{\nu}^{(-k)}$的归一化分量。两种权重都接近$r\, e^{\hat{\nu}_n^{(-k)}}$，因为每个留出折约占参考样本的 $1/K$。参考折按固定大小抽取（附录 C.5）。

敏感度 g_k通过正交化 b_k将干扰参数估计传递到方差中，它根据这些求和计算得出：对于 ν_n ，其对应项为$g_{k,\ n} = \mathbb{E}_{k,\ \hat{\nu}^{(-k)}}[\phi_k]$，因为权重与 e^{ν_n}成正比；对于 ν_s，其对应项与 J_k中的 ν_s-导数一样，是这些求和的中心差分，并在每个偏移值处重复进行筛选（附录 C.2）。

ˆ推前干扰参数拟合。 默认拟合 (6) 具有 $\psi = (1, \delta_1)$，这是一个固定函数。由于强度与 e^{ν_n} 成正比，其两个训练期望值可分解为 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\ \nu}[1] = e^{\nu_n} C_{0,\ k}(\nu_s)$ 和 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\ \nu}[\hat{}\delta_1] = e^{\nu_n} C_{1,\ k}(\nu_s)$。式 (24) 中的各项是它们的导数：在 ν_n 中是期望值本身，在 ν_s 中是导数$C'_{0,\ k}$ 和 $C'_{1,\ k}$，均在训练参考样本上计算，

$$J_k = e^{\hat{\nu}_n^{(-k)}} \begin{pmatrix} C_{0,k} & C'_{0,k} \\ C_{1,k} & C'_{1,k} \end{pmatrix} + \Sigma_A^{-1}. \tag{41}$$

中的参考噪声 U_j。 现在，被减去的期望值是参考求和，而中心化和(35)变为

$$\xi_j(u) = \hat{\mathbb{E}}_j[u] - \bar{\mathbb{E}}_{j,\nu^*}[u] = \Big\{\hat{\mathbb{E}}_j[u] - \mathbb{E}_{j,\nu^*}[u]\Big\} - \Big\{\bar{\mathbb{E}}_{j,\nu^*}[u] - \mathbb{E}_{j,\nu^*}[u]\Big\},$$

数据噪声减去第 j 折的参考噪声，二者都仅是第 j 折数据事件和参考事件的函数。训练参考事件通过 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\ \nu}[\psi]$ 以及 $\bar{\mathbb{E}}^{\mathrm{tr}}_{k,\ \nu} = \sum_{j \neq k} \bar{\mathbb{E}}_{j,\ \nu}$ 进入干扰参数拟合，其影响一直到折大小的不平衡为止，因此 A.2 节的重新分组保持不变：$S = \sum_j U_j - c^\top \Delta_A + \ldots$其中 $U_j = \xi_j(\tilde{\phi}_j)$，而 $\mathcal{F}_{-j}$ 现在也排除了第 j 折的参考事件。因此，第 j 折的一个参考事件 X'_e 的作用就像一个符号相反、权重为 ω_j 的数据事件：其在表 6 中所在行的每一项都是数据行对应项的 $-\omega_j$ 倍，其中 $\phi_j(e)$ 和 $\psi(e)$ 在变换后的事件 $T_{\hat{\nu}_s^{(-j)}} X'_e$ 处计算；若该事件未通过筛选，则将其设为零，而其总贡献为

$$-\omega_j \, \tilde{\phi}_j(e). \tag{42}$$

ξ_j 的数据部分和参考集部分彼此独立，因此，公式 (30) 中 $U_j \mid \mathcal{F}_{-j}$ 的方差是数据部分（由 $\hat{}V_j^D$ 估计）与参考集部分之和。参考集部分是对 $\mathcal{R}_j$ 上总量 (42) 求和所得结果的方差，由公式 (14) 中居中后的 $\hat{}V_j^R$ 估计（附录 C.5）；它约为 $\omega_j^2 \sum_e \tilde{\phi}^2_{j,\ e} = \omega_j \, \bar{\mathbb{E}}_{j,\ \hat{\nu}^{(-j)}}[\tilde{\phi}_j^2]$，约为数据部分的 r 倍。

拟合随机性。折 j 的参考事件也包含在见证函数 ϕ_k、$k \neq j$ 的训练样本中，因此交叉项 $\mathrm{Cov}(U_j,\ U_k)$ 按照 A.3节中的方法推导，只是更新了 ξ_k 的定义。ξ_k 的数据部分和参考集部分彼此独立，数据部分与之前一样给出 $\mathbb{E}_{k,\nu^*}[a_j a_k^\top] = E_k \Omega_{jk}$。参考集部分是 ω_k 乘以折 k 的 M_k 个参考事件之和，这些事件是从归一化零假设 Q^* 中固定抽取的。此类求和的协方差是单次抽取协方差 $\Omega_{jk} - \mu_j \mu_k^\top$ 的 M_k 倍，而不是泊松计数产生的原始二阶矩 Ω_{jk}：由于抽取次数固定，不存在计数波动，因此无需额外加上 $\mu_j \mu_k^\top$。其中，$\omega_k M_k = E_k$ 为留出事件的预期数量，且 $\omega_k \approx r$ 接近中心值，

$$
\begin{aligned}
\mathbb{E}\big[\xi_k(a_j)\,\xi_k(a_k)^\top\big] &= \mathbb{E}_{k,\nu^*}[a_j a_k^\top] + \omega_k^2 M_k \,\mathrm{Cov}_{Q^*}(a_j, a_k) \\
&= E_k \Omega_{jk} + \omega_k E_k \big(\Omega_{jk} - \mu_j \mu_k^\top\big) \\
&\approx E_k \tilde{\Omega}_{jk}
\end{aligned}
$$

with

$$
\tilde{\Omega}_{kj} = \Omega_{kj} + r\big(\Omega_{kj} - \mu_k \mu_j^\top\big), \qquad \mu_k = \int a_k \, dQ^*. \tag{43}
$$

这里的 Q^* 是公式 (39) 中归一化的真实零假设。因子 r 来自权重：一个参考折的事件数约为数据折的 $1/r$ 倍，每个事件的权重为 r，因此其数据和的方差为数据和方差的 $(1/r)\, r^2 = r$ 倍。如果参考折是作为泊松过程抽取的，那么其 $\tilde{}$ 阶矩就会像数据部分一样不以中心为基准，而 $\tilde{\Omega} = (1 + r)\Omega$。

A.5 总结

$\tilde{}$对于每一折，计算会在留出数据和拟合零假设上返回 S_k、g_k、J_k、ϕ_k，以及 $\hat{}V_k^D$、$\hat{}V_k^R$；对于 $\hat{}\,V_{\mathrm{fit}}$，还会在共同的 $\hat{}$测度 Q（附录 C.6）上返回自由方向，其中 H_k 和 D_k。统计量为

$$
\begin{aligned}
Z &= \frac{\sum_k S_k}{\sqrt{\hat{V}}}, \\
\hat{V} &= \sum_k \big(\mathbb{E}_{k,\hat{\nu}^{(-k)}}[\tilde{\phi}_k^2] + \hat{V}_k^R\big) + c^\top \Sigma_A^{-1} c \\
&\quad + \sum_{j \neq k} E_j E_k \,\mathrm{tr}\big(D_k H_k^{-1} \tilde{\Omega}_{kj} D_j H_j^{-1} \tilde{\Omega}_{jk}\big).
\end{aligned} \tag{44}
$$

通过预言机访问，期望值是精确的，$\hat{}V_k^R = 0$ 以及 $\tilde{}\Omega_{kj} = \Omega_{kj}$；通过$\tilde{}$样本访问，每个计算得到的 $\mathbb{E}$ 都是第 A.4节的参考估计 $\mathbb{E}$。

分子使用原始见证函数 ϕ_k，而方差使用正交化后的 $\tilde{\phi}_k$（第3.3节）。

B 极限定理的陈述与证明概要

我们以本文采用的形式陈述第3.5节的结果，并概述其证明。精确的陈述、证明以及对退化区域的推广将在一篇配套统计论文中给出。

设定与尺度。 事件数 N_0 不断增加，K、$r = N_0/M$ 和 $\Sigma_A N_0$ 保持不变。令 $\sigma_\phi = [\mathbb{E}\mathrm{Var}_{Q^*}(\phi_k)]^{1/2}$ 表示见证函数的典型大小，其中 $\mathrm{Var}_{Q^*}(\phi_k)$ 是在式 (39) 中 Q^* 条件下 ϕ_k 的条件方差。令 $\sigma_S^2 = \sum_k (V_k^D + V_k^R) + V_A$，其中 V_k^D 和 V_k^R 是给定训练折后得到的条件方差，这些训练折上的 $\hat{V}_k^D$ 和 $\hat{V}_k^R$ 分别是它们的估计值（附录 C.4 和 A.4）。第3.5节的条件如下：

(D) $\mathcal{D}$ 是强度为 $n(\cdot\,|\mathrm{R}_{\nu^*})$ 的泊松过程；给定各层的样本量（附录 C.1），参考事件是从参考模型中独立抽取的样本；$\hat{\nu}_A \sim N(\nu^*,\ \Sigma_A)$；三者相互独立，且折分配在各层内随机进行。

(N) 每个 $\hat{\nu}^{(-k)}$ 都可按式（27）线性化，余项为 $o_P(N_0^{-1/2})$，且 $\nu \mapsto \mathbb{E}_{k,\nu}[\phi_k]$ 在 ν^* 附近二阶可微，二阶导数为 $O_P(N_0\sigma_\phi)$。**(W1)** 当样本量趋于无穷时，以趋近于 1 的概率，$\sup_x |\phi_k(x)| \le B\sigma_\phi$ 在参考事件的支撑集上成立，其中常数为 B。

(W2) 将一个训练数据事件替换为独立副本，会使 ϕ_k 在一个独立点处改变 $o(\sigma_\phi/\sqrt{N_0})$（按 L_2 度量）；将一个训练参考事件（其损失函数中的权重约为 r）替换后，会使其改变 $o(\sqrt{r}\,\sigma_\phi/\sqrt{N_0})$。

定理。 在 H_0 和 (D)、(N)、(W1)、(W2) 条件下，

$$\frac{S}{\sqrt{\sum_k \left(\hat{V}_k^D + \hat{V}_k^R\right) + \hat{V}_A}} \ \rightsquigarrow\ N(0,1).$$

$\hat{}$ V_{fit} 所估计的跨折协方差是 $o(\sigma_S^2)$，因此只要 Z 在概率意义下成立，式（10）的 $V_{\mathrm{fit}}/\sigma_S^2 \to 0$ 就具有相同的极限。借助预言机访问

ˆ 参考项不存在。没有任何条件涉及$\delta_{1,\ 2}$：干扰变换精确地应用于参考事件，因此学习函数只会影响$\nu^{(-k)}$的精度。

证明思路。 该论证是针对交叉验证的随机中心化中心极限定理 [11], ，应用于两种输入类型和一项辅助测量。

1. 插件法。 在 ν^* 附近展开 S_k 并代入 (27)，得到 $S = \sum_j U_j - c^\top \Delta_A + R$，其中 $U_j = \xi_j(\tilde{\phi}_j)$ 如附录 A.2 所示： ξ_j 是对第 j 折数据事件求和得到的中心化和 (35)，减去其具有样本访问的参考分布对应项（附录 A.4）； Δ_A 如 (27) 所示。由 (N) 和 (W1)，余项 R 关于 ˆ$\nu^{(-k)} - \nu^*$ 为二阶，是 $o_P(\sigma_S)$。

¯ ˜2. 稳定化。 令 $\phi_j(x) = \mathbb{E}[\phi_j(x)]$, 表示在固定 x 时，对第 j 折（数据和参考分布）的训练事件以及 ˆν_A 取期望。由 Efron–Stein 不等式和 (W2)，对两种类型的训练事件求和后，

˜ ¯ ˜ ¯给出 $\phi_j = \phi_j + o_P(\sigma_\phi)$，其阶为 $L_2(Q^*)$。因此，用 ϕ_j 替换 ϕ_j 会使S 改变 $o_P(\sigma_S)$；此外， S_j 对 S_k 的评估事件的依赖， $j \neq k$ 对 Var(S) 的贡献为 $o(\sigma_S^2)$。3.中心极限定理。 主导项是若干独立贡献之和：数据事件上的复合泊松和，其方差未经中心化；参考事件上的和，在各层内中心化；以及高斯分布 $-c^\top \Delta_A$。其方差为 σ_S^2，且 (W1) 给出 Lindeberg 条件。4. 学生化。 给定训练折后， ˆV_k^D 和 ˆV_k^R 分别是 V_k^D 和 V_k^R 的无偏估计；由 (W1)、(W2)，其和对$\sigma_S^2 - V_A$具有比率一致性。由 Slutsky 引理即可得证。

当见证函数纯粹由估计噪声构成时，条件（W2）可能在 H_0下失效（第3.5节）。此时第2步会失效，跨折协方差不再可忽略，因此需要采用附录 A.3中的修正。

C 实现细节

本附录介绍如何计算附录A中的各项量：具有样本访问权限的参考样本、期望值及其干扰参数导数、干扰参数拟合，以及学习函数 $\delta_{1,\ 2}$ 所使用的

样本访问权限、方差项，以及每个见证函数的拟合随机性项。

C.1 具有样本访问权限的参考样本

在第6节的基准测试中，筛选取决于尺度干扰参数，因此，如果参考分布仅由在 $\nu = 0$ 时通过筛选的事件构成，那么经过变换 T_{ν_s} 后，尺度偏移带入筛选范围的事件就会缺失。因此，参考分布是在筛选前抽取的。所考虑的尺度偏移位于 $\mathcal{V}_s = [-\bar{\nu}_s, \bar{\nu}_s]$, $\bar{\nu}_s = 0.065/3 \approx 0.022$ （约为 $7\sigma_s$）范围内，该范围涵盖了拟合和数据集中使用的所有取值。令 $\mathcal{B}$ 表示 $\mathcal{S}_{\text{toy}}$ 中在 $\nu = 0$ 时未通过筛选、但在将其 p_T 和不变质量阈值放宽至原来的 $e^{0.065}$ 倍后通过筛选的事件。由于对于 $\nu_s \in \mathcal{V}_s$，没有任何动量的变化超过这一倍数，因此 $\mathcal{B}$ 包含了此类偏移可能带入筛选范围的所有事件，

$$\bigcup_{\nu_s \in \mathcal{V}_s \setminus \{0\}} \mathcal{P}_{\nu_s} \setminus \mathcal{P}_0 \subseteq \mathcal{B},$$

其中 $\mathcal{P}_{\nu_s}$ 与第6.1节中的定义相同。参考样本由从 $\mathcal{P}_0$ 中有放回抽取的 M 个事件，以及额外从 $\mathcal{B}$ 中有放回抽取的 $n_{\mathcal{B}} = M\,|\mathcal{B}|/|\mathcal{P}_0|$ 个事件构成，使这两类事件在样本中的比例与 $\mathcal{S}_{\text{toy}}$ 中相同。其权重按 M 归一化，即在 $\nu = 0$ 时通过筛选的事件（式（40））；只有当 T_{ν_s} 将某个事件带入筛选范围时，该事件才会对 ν 处的期望值有所贡献。以相同的抽样率从全部 $\mathcal{S}_{\text{toy}} \setminus \mathcal{P}_0$ 中抽取第二层，会得到相同的期望值：将其限制在 $\mathcal{B}$ 中，只会遗漏在任何 $\nu_s \in \mathcal{V}_s$ 下都无法通过筛选、因而贡献为零的事件，从而省去在每个 ν_s 都对它们进行变换的步骤。将参考样本拆分为多折时，两层仍保持分开（附录 C.5）。

C.2 期望值与干扰参数导数

通过预言机访问，使用梯形法则在均匀网格上计算式 (21) 的积分；该网格覆盖拟合零假设的支撑集，直至其$1 - 10^{-9}$ 分位数。字典见证函数使用 16 001 个节点，网络使用 4 001 个节点。

通过样本访问，期望值就是参考分布的求和（40）。它们关于 ν_n 的导数就是求和本身；关于 ν_s 的导数则是中心差分，

$$g_{k,s} = \frac{\bar{\mathbb{E}}_{k,(\hat{\nu}_n^{(-k)},\,\hat{\nu}_s^{(-k)}+\varepsilon)}[\phi_k] - \bar{\mathbb{E}}_{k,(\hat{\nu}_n^{(-k)},\,\hat{\nu}_s^{(-k)}-\varepsilon)}[\phi_k]}{2\varepsilon}, \qquad \varepsilon = 10^{-3}, \tag{45}$$

其中，将留出参考事件转换到偏移后的取值，并重复进行筛选，从而统计跨越选择阈值的事件。式（41）中的导数$C'_{0,\,k}$和$C'_{1,\,k}$，以相同方式在训练参考样本上计算。

C.3 干扰参数拟合

似然拟合。 在具有预言机访问的情况下，似然拟合（式 (4)，其中包含 $\psi = s$）通过牛顿法求解，从无约束估计（$\Sigma_A^{-1} = 0$）开始迭代，直至估计方程的残差低于 10^{-10}。在第 5 节的基准测试中，归一化方程和尺度方程分解为两个凹一维问题，且迭代至多在一次过冲后便呈单调变化。

推前拟合。 在固定 ν_s 时，式 (6) 的计数方程关于 ν_n 严格递减，因此通过牛顿法求解 ν_n，直至达到 10^{-12}。代入该剖面值 ν_n（ν_s）后，剩下一个关于尺度的方程，

$$F_k(\nu_s) = \sum_{i \in \mathcal{D}_{-k}} \hat{\delta}_1(X_i) - e^{\nu_n(\nu_s)} C_{1,k}(\nu_s) - \frac{\nu_s - \hat{\nu}_{A,s}}{\sigma_s^2} = 0,$$

其中 $C_{1,\,k}$ 如附录 A.4 所述。$C_{1,\,k}$ 是对训练参考样本求和，并在每个 ν_s 处重复进行筛选，因此每当参考分布中的事件越过选择阈值时，它就会发生跳变；所以使用无需导数的布伦特法求根，令 ν_s 中的绝对容差为 10^{-6}。搜索范围限定为 $|\nu_s| \le 10^{-2}$，即训练 $\hat{\delta}_{1,\,2}$ 时所用的尺度偏移范围；在此范围内，参考分布保留所有可能因偏移而进入筛选范围的事件（附录 C.1）。首先在此区间内求解全数据拟合。每个折叠拟合都从以全数据估计值为中心、半宽为 10^{-3} 的区间括点开始，并将其宽度扩大四倍，直至 F_k 改变符号或区间括点达到整个区间。若 F_k 在该区间内未改变符号，则将估计值设为边界值并标记该拟合。在第 6 节的研究中（2320 个数据集，23 200 个折拟合），没有拟合达到边界；最大的 $|\hat{\nu}_s^{(-k)}|$ 为 0.0076。

学习函数$\hat{\delta}_{1,\,2}$。损失函数 (20) 的样本 $\mathcal{S}_1$ 和 $\mathcal{S}_0(\nu_i)$ 由 $\mathcal{S}_\delta$ 中相同的事件构建，分别对应 T_{ν_i} 之前和之后。输入按照参考文献 [2] 中的方法进行预处理：赝快度经过标准化，正特征（$p_{T,\,1}, p_{T,\,2}, \Delta\phi_{12}$）除以其均值，所用统计量来自 $\mathcal{S}_1$ 的训练事件。两个网络均采用 ReLU隐藏单元和线性

输出，按照 Glorot 均匀初始化规则进行初始化，并使用 Adam，按照附录 C.8 中的设置，以全批次方式联合训练。每十个训练轮次，都会在留出的五分之一数据上评估验证损失（按事件划分，因此同一事件及其偏移副本会被分到同一侧）。在检验中，通过前向传播评估网络，其输出与训练器的输出之差在10^{-8}以内。

C.4 计算 $\hat{V}_k^D$

$\hat{V}_k^D$是第 k 折中留出的事件各自的总贡献 $\phi_k(X_i)$（见表 6）之和的方差。给定训练折后，留出的事件近似构成强度为 $\hat{n}_k/K$ 的泊松过程：其数量和位置都是随机的。根据坎贝尔定理 [25]，$\hat{\mathbb{E}}_k[\tilde{\phi}_k]$ 的方差是未中心化的二阶矩 $\hat{V}_k^D = \mathbb{E}_{k,\ \hat{\nu}^{(-k)}}[\tilde{\phi}_k^2]$；计数的波动也是 S 的零假设波动的一部分。

C.5 计算 $\hat{V}_k^R$

参考样本分为两层（附录 C.1）：通过 $\nu = 0$ 筛选的事件，以及未通过筛选、但可能 $T_{\hat{\nu}_s^{(-k)}}$
进入接纳范围的近阈值事件；仍未通过 $\hat{\nu}^{(-k)}$ 的事件具有 $\tilde{\phi}_{k,\ e} = 0$。令 $\mathcal{P}_{0,\ k}$ 和 $\mathcal{B}_{0,\ k}$ 分别表示第 k 折中两个层内留出的参考事件，其检验尺寸分别为 $M_k = |\mathcal{P}_{0,\ k}|$ 和 $n_{\mathcal{B},\ k} = |\mathcal{B}_{0,\ k}|$。各层内分别抽取样本，因此 M_k 和$n_{\mathcal{B},\ k}$ 是固定的，权重 ω_k 按M_k归一化，只有抽取了哪些事件是随机的。在每一层内，留出的事件构成简单随机样本，其总和的方差就是层内中心化平方和，

$$\hat{V}_k^R = \omega_k^2 \Big[\sum_{e \in \mathcal{P}_{0,k}} \big(\tilde{\phi}_{k,e} - \bar{\tilde{\phi}}_k^{\mathcal{P}}\big)^2 + \sum_{e \in \mathcal{B}_{0,k}} \big(\tilde{\phi}_{k,e} - \bar{\tilde{\phi}}_k^{\mathcal{B}}\big)^2 \Big], \tag{46}$$

其中 $\bar{\tilde{\phi}}_k^{\mathcal{P}}$和 $\bar{\tilde{\phi}}_k^{\mathcal{B}}$ 是对 $\mathcal{P}_{0,k}$ 和 $\mathcal{B}_{0,k}$求得的均值。与对整个折求均值相比，在各层内进行中心化很重要：$\mathcal{B}_{0,k}$ 主要由满足$\phi_{k,e} = 0$的事件构成；若层大小固定，层均值之间的差异不会波动，否则就会被计入方差。只有一个层时，公式 (46) 化为 (14)。如果该折是按泊松过程抽取的，那么坎贝尔定理同样适用于数据项，并且会给出未中心化的 $\omega_k^2 \sum_e \tilde{\phi}_{k,e}^2$；二者相差 $\omega_k^2\big[M_k(\bar{\tilde{\phi}}_k^{\mathcal{P}})^2 + n_{\mathcal{B},k}(\bar{\tilde{\phi}}_k^{\mathcal{B}})^2\big]$。

ˆC.6 计算 V_{fit}

泊松矩。 在 H_0 中，数据事件构成强度为 $n(\cdot\,|\mathrm{R}_{\nu^*})$ 的泊松过程；将事件随机分配到各折后，该过程被拆分为 K 个相互独立的泊松过程，每个过程的强度为 $n(\cdot\,|\mathrm{R}_{\nu^*})/K$。对于向量函数 b 和 c，坎贝尔定理 [25] 给出了它们在折 l 的事件上求和所得结果的协方差，其形式为未中心化的二阶矩，

$$\mathrm{Cov}\big(\hat{\mathbb{E}}_l[b],\hat{\mathbb{E}}_l[c]\big)=\mathbb{E}_{l,\nu^*}\big[b\,c^{\top}\big],\tag{47}$$

˜其中，附录 C.4 是 $b=c=\phi_l$ 的情形；不同折上的求和相互独立。在固定自由方向的条件下，中心化和（35）因此具有 $\mathbb{E}\big[\xi_l(a_j)\xi_l(a_k)^{\top}\big]=\mathbb{E}_{l,\nu^*}[a_j a_k^{\top}]$ 对于每个折 l 都成立，并且在不同的 l 之间相互独立。

配对。 将（37）代入公式（38）后，只有共享配对会产生贡献：$\xi_j(a_j)^{\top}H_j^{-1}\xi_l(a_j)\,\xi_k(a_k)^{\top}H_k^{-1}\xi_m(a_k)$，其中 $l=k$ 和 $m=j$，因为各折相互独立，且每个 ξ 的均值为零。

公共测度与重叠量。 写作 $\mathbb{E}_{l,\,\nu^*}[\,\cdot\,]=E_l\int\cdot\,dQ_l$，其中 Q_l 是折 l 的归一化零假设，公式（38）需要在第三个折的零假设下计算两个折的自由方向的二阶矩。各折的拟合零假设仅通过其干扰参数估计而不同，其相对阶数为 $N_0^{-1/2}$，因此计算中将它们全部以及 Q^* 替换为一个公共概率 ˆ ˆ 测度 Q。在预言机访问下，Q 是第一个折的归一化拟合零假设，在其积分网格上求值。在样本访问下，它是完整数据干扰参数拟合下的归一化零假设，由参考样本表示。计算得到的重叠量为

$$\Omega_{kj}=\int a_k\,a_j^{\top}\,d\hat{Q},\tag{48}$$

出于与 ˆV_k^D 相同的原因，它未中心化；采用附录 C.7 的阻尼后，ˆ V_{fit} 的被加数为 $E_jE_k\mathrm{tr}\big(D_kH_k^{-1}\,\Omega_{kj}D_jH_j^{-1}\Omega_{jk}\big)$。曲率是在同一测度上计算的，并按第 k 折的训练暴露量重新缩放：$H_k=E_k^{\mathrm{tr}}\int e^{\phi_k}a_ka_k^{\top}\,d\,\hat{}Q$，其中 $E_k^{\mathrm{tr}}=\mathbb{E}^{\mathrm{tr}}_{k\,,\,\hat{\nu}^{(-k)}}[1]$ 表示预期训练事件数。通过样本访问，重叠量会包含式（43）的参考部分。

C.7 阻尼

线性化见证统计量（36）略去了 $G_k(\theta_0)$ 的第二项，即被干扰参数拟合吸收的训练波动部分。沿着一个方向

在该方向上，零假设本身可以移动，训练波动由零假设和见证函数共同分担：干扰参数拟合吸收其中的 ρ，见证函数保留其余部分。辅助测量限制了零假设跟随数据的程度，因此留给见证函数的份额 $1-\rho$ 介于 0 和 1 之间：对于不受约束的干扰参数，其会完全吸收该方向，份额为 0；对于被辅助测量固定的干扰参数，份额为 1。阻尼会按照这一份额缩放此类方向的拟合随机性。

形式上，将式 (27) 代入 $G_k(\theta_0)$ 的第二项，其数据部分将 a_k 的训练求和替换为残差化方向的训练求和

$$\tilde{a}_k = a_k - \mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}[a_k\, s^\top]\, J_k^{-1}\psi,$$

并移除了被干扰参数拟合吸收的部分。字典中的两个原子函数本身就是干扰参数得分函数，我们称其为得分原子：截距，即归一化的得分函数 1，以及式 (18) 中的原子函数 $\tilde{s}$，它是尺度得分函数 s_s 的倍数。对于似然拟合，$\psi = s$，因此每个得分原子都是某个估计函数 $a = c\,\psi_m$ 的倍数。由于 $\mathbb{E}^{\mathrm{tr}}_{k,\hat{\nu}^{(-k)}}[\psi_m s^\top]$ 根据式 (24) 是 $J_k - \Sigma_A^{-1}$ 的第 m 行，

$$\tilde{a}_k = c\, e_m^\top \Sigma_A^{-1} J_k^{-1}\psi = (1-\rho_{k,m})\, a,$$

其中第二个等式利用了 J_k 为对角矩阵这一事实，因为两个得分原子在拟合零假设下彼此正交。在单位信息参数化中

$$\rho_{k,m} = \frac{E_k^{\mathrm{tr}}}{E_k^{\mathrm{tr}} + \sigma_m^{-2}},$$

其中 E_k^{tr} 是预期的训练事件数（附录 C.6），σ_m 是对干扰参数 m 的辅助测量的不确定性。$\rho_{k,\,m}$ 是受约束干扰参数拟合的收缩：拟合得到的干扰参数从 $\hat{\nu}_A$ 向无约束估计移动了 $\rho_{k,\,m}$ 的比例，而见证函数在得分原子上的系数是相对于无约束零假设时的 $1-\rho_{k,\,m}$ 倍。

对于字典见证函数，公式（15）中的 D_k 在 $1-\rho_{k,\,m}$ 中，干扰参数的得分函数原子对应的值为 m，而光滑原子和隆起对应的值为 1。根据构造，在拟合零假设下，它们与 1 和 s_s 正交，因此不会被阻尼（第 5.1 节）。

对于网络，残差化没有闭式解：参数梯度 a_k 与得分函数并不正交，而干扰参数拟合所吸收的梯度部分并未计算。只有输出偏置会使见证函数沿常数方向移动，因此受到 $1-\rho_{k,n}$ 的阻尼；其他所有自由参数

都保持因子 1，尽管隐藏单元也能产生近似常数的偏移，而且尺度方向没有专门的参数。对只被部分吸收的方向不施加阻尼会高估拟合方差，因此这种选择是保守的。

ˆC.8 V_{fit} 的其他细节

网络采用 Adam 进行训练，学习率为 10^{-3}, $\beta_1 = 0.9$, $\beta_2 =$0.99， $\epsilon = 10^{-7}$ 如参考文献 [2], 所述，使用全批次（每个训练轮次执行一步），并在每一步之后进行裁剪。将 $a_k(x) = \partial f(x;\, \theta)/\partial\theta$ 限定为自由参数后，曲率为式（16）的高斯–牛顿矩阵，

$$H_k = \frac{K-1}{K} \int \hat{n}_k \, e^{\phi_k} \, a_k a_k^{\top} + \epsilon_H \, \overline{\mathrm{diag} H} \, I, \tag{49}$$

其中， I 是单位矩阵，$\mathrm{diag}\overline{H}$ 是第一˜项的对角元素均值，而重叠量 Ω_{kj} 则由 a_k 和 a_j 在公共测度上构造。下面讨论自由参数、岭惩罚 ϵ_H 以及渐变窗见证函数的选择。

自由参数。自由参数是严格位于裁剪界内的参数，以及任何未被裁剪的参数。在精确约束最优点处，处于盒约束边界上的参数不会对数据的微小变化作出响应；而采用固定训练轮次数进行训练则无法达到该点。在零假设研究中，随着更多参数到达边界，这条规则可能会低估拟合方差（表8： Z 的方差随训练时长增加），但对于我们使用的配置而言已经足够。

岭惩罚。 在没有正则化的情况下， H_k 在某些自由方向上几乎是奇异的，而 V_{fit} 平均而言是无偏的，但噪声较大，且偶尔会出现负值。岭惩罚$\epsilon_H = 10^{-4}$ 曾在早期配置下的双轻子基准零假设模拟中确定一次（$c = 0.72$，2800 个训练轮次；288 个数据集），并在留出数据上进行检查；在两个基准测试的每种配置中均保持不变使用。它是唯一以这种方式固定的校准常数。

渐缩。 对于上尾部稀疏填充的特征，第 4.3 节中的渐变窗为

$$\tau_k(x) = \min\Big\{1,\ \frac{E_k^{\mathrm{tr}}\left[1 - F_k(x)\right]}{\kappa}\Big\}, \tag{50}$$

其中 F_k 是拟合零假设的分布函数， E_k^{tr} 是预期训练事件数（附录 C.6），因此分子是超过 x 的预期训练事件数。见证函数 $\tau_k f$ 在至少还预期有 κ 个训练事件超过 x 时，等于网络输出；在更远处则按该数量的比例渐弱。采用渐变窗时（用于第 5 节的网络），由于网络是在未使用渐变窗的情况下训练的，因此它通过未渐缩的梯度 a_k 对训练事件作出响应，而统计量则通过 $\tau_k a_k$ 感知渐缩后的见证函数。因此，重叠量将二者配对：式（39）中的 Ω_{kj} 变为 $\Omega_{kj}(\tau) = \int a_k\,(\tau_j a_j)^{\top} dQ^*$，按照附录 C.6，在ˆ 公共测度 Q 上计算；当 $\tau \equiv 1$ 时，它简化为 Ω_{kj} 。这里忽略了 τ_k 对干扰参数估计的依赖；其贡献是见证函数与干扰参数误差的乘积，在零假设下属于二阶项，因为此时见证函数本身与噪声同量级。

D 其他结果

本附录汇总了第 5 节和第 6 节中提及的结果：字典见证函数的 Z 零分布，以及图 3 和图 6 背后的完整调参网格。

零假设下的字典见证函数。 图 7 是字典见证函数对应的图 1，在相同的五个干扰参数点且 $\sigma_A = 0.05$下的对应图。其结果与神经网络相同。加入拟合随机性项后，名义5%水平下的比率介于 0.0400 和 0.0475 之间；不加入该项时，分布明显过宽，比率介于 0.080 和 0.100 之间，即为名义值的 1.6 至 2 倍。

调参网格。 表格 7 和 8 给出了图 3 和图 6 中绘制的数值；对于一维基准测试，还给出了第 5.4 节所总结的字典惩罚水平扫描结果。对于每个见证函数，所有配置都在相同的数据集上运行，因此其各行之间的差异不会因独立的蒙特卡洛噪声而被夸大。

参考分布

[1] Raffaele Tito D’Agnolo 和 Andrea Wulzer。从机器学习新物理。*Physical Review D*，99(1):015014，2019。doi: 10.1103/PhysRevD. 99.015014。

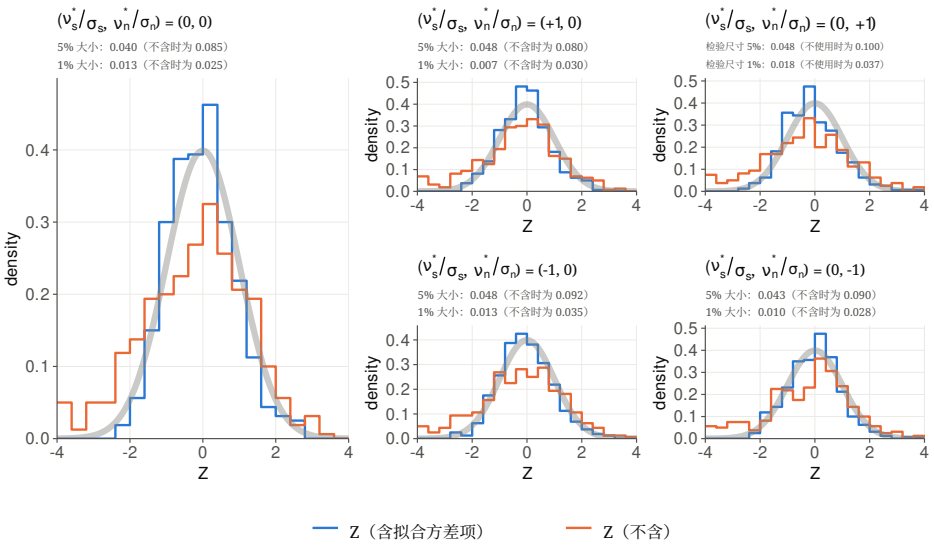

图 7：与图 1 相同，针对字典见证函数。

[2] Raffaele Tito D'Agnolo、Gaia Grosso、Maurizio Pierini、Andrea Wulzer 和 Marco Zanetti。学习多变量新物理。欧洲物理学杂志 *C*，81(1):89，2021。doi: 10.1140/epjc/s10052-021-08853-y。[3] Raffaele Tito D'Agnolo、Gaia Grosso、Maurizio Pierini、Andrea Wulzer 和 Marco Zanetti。从不完美机器学习新物理。欧洲物理学杂志 *C*，82(3):275，2022。doi: 10.1140/epjc/s10052-022-10226-y。[4] Robert B. Davies。当干扰参数仅存在于备择假设下时的假设检验。《生物计量学》，74(1):33–43，1987。doi: 10.1093/biomet/74.1.33。[5] Eilam Gross 和 Ofer Vitells。高能物理中别处效应的试验因子。欧洲物理学杂志 *C*，70:525–530，2010。doi: 10.1140/epjc/s10052-010-1470-8。[6] Herman Chernoff。关于似然比分布。《数学统计年刊》，25(3):573–578，1954。doi: 10.1214/aoms/1177728725。[7] Steven G. Self 和 Kung-Yee Liang。最大似然估计量和似然比检验的渐近性质，在非标准……

表 7：一维基准测试， $\nu^{*}=0$, $\sigma_{A}=0.05$：名义5%和1%水平下的假阳性率，以及 NP1–NP3 的中位数 Z，括号内为其自助法标准误，每项包含 400 个数据集。网络：裁剪界 c 和训练轮数；字典：惩罚水平 λ。 †：第 5 节中使用的配置。

见证函数	设置 (c，训练轮数) / λ	零假设：比率为 5%	零假设：比率为 1%	中位数 Z NP1	中位数 Z NP2	中位数 Z NP3
网络	$(3, 2\times10^{4})$	0.035	0.000	2.22 (0.16)	1.28 (0.07)	3.38 (0.12)
网络	$(3, 3\times10^{4})$	0.040	0.000	2.35 (0.16)	1.26 (0.08)	3.40 (0.11)
网络	$(3, 5\times10^{4})$	0.038	0.000	2.42 (0.13)	1.28 (0.09)	3.39 (0.11)
网络	$(4,\ 2\times10^{4})$	0.033	0.003	2.11 (0.17)	1.24 (0.08)	3.62 (0.12)
网络	$(4,\ 3\times10^{4})^{\dagger}$	0.038	0.003	2.41 (0.15)	1.19 (0.09)	3.62 (0.11)
网络	$(4,\ 5\times10^{4})$	0.045	0.003	2.51 (0.14)	1.21 (0.10)	3.61 (0.10)
网络	$(6,\ 2\times10^{4})$	0.040	0.005	2.19（0.17）	1.15（0.09）	3.90（0.12）
网络	$(6,\ 3\times10^{4})$	0.040	0.008	2.40 (0.14)	1.18 (0.08)	3.88 (0.13)
网络	$(6,\ 5\times10^{4})$	0.045	0.010	2.58 (0.16)	1.13 (0.07)	3.84 (0.12)
字典	1.1	0.045	0.013	2.97 (0.11)	1.09 (0.08)	3.42 (0.07)
字典	$1.3^{\dagger}$	0.040	0.013	3.08 (0.09)	1.16 (0.08)	3.43 (0.07)
字典	1.5	0.038	0.010	3.09 (0.09)	1.21 (0.08)	3.38 (0.08)

dard conditions. *J. Amer. Statist. Assoc.*, 82(398):605–大987. doi：10.1080/01621459.1987.10478472.[8] Larry Wasserman、Aaditya Ramdas 和 Sivaraman Balakrishnan。通用推断。美国国家科学院院刊，117(29):16880–16890，2020。doi: 10.1073/pnas.1922664117.[9] Purvasha Chakravarti、Mikael Kuusela、Jing Lei 和 Larry Wasserman。使用可解释的半监督分类器检验对新物理信号进行模型无关检测。应用统计年刊，17(4):2759–2795，2023。doi: 10.1214/22-AOAS1722.[10] Gaia Grosso、Marco Letizia、Maurizio Pierini 和 Andrea Wulzer。通过 Neyman–Pearson检验评估拟合优度。*SciPost*物理，16(5):123，2024。doi: 10.21468/SciPostPhys.16.5.123.[11] Jing Lei。稳定性视角下的交叉验证现代理论。*Found. Trends Stat.*，1(3-4):391–548，2025。doi: 10.1561/3600000005。

[12] Morgane Austern 和 Wenda Zhou。《交叉验证的渐近理论》，2020。

表8：双轻子基准测试，网络见证量：零假设下 Z 的方差及其在名义5%和1%水平下的比率（$\nu^{*}=0$，400个数据集）；针对 c 裁剪界和训练轮数，Z' 与有效场论的中位数 Z （160个数据集），括号内为其自助法标准误。†：第6节中使用的配置。

		null			中位数 Z	
c	训练轮数	Var(Z)	5%水平比率	1%水平比率	Z'	EFT
0.7	2000	0.83	0.0400	0.0075	2.68 (0.10)	6.78 (0.12)
0.7	3000†	0.93	0.0575	0.0125	2.80 (0.13)	7.00 (0.13)
0.7	5000	0.93	0.0575	0.0075	2.86 (0.16)	7.00 (0.12)
1.0	2000	0.90	0.0450	0.0075	2.80 (0.13)	8.02 (0.18)
1.0	3000	1.06	0.0625	0.0100	2.79（0.13）	8.52（0.25）
1.0	5000	1.11	0.0650	0.0100	2.85（0.11）	8.51（0.24）
1.3	2000	0.92	0.0500	0.0000	2.78（0.11）	8.69（0.22）
1.3	3000	1.08	0.0400	0.0025	2.91（0.14）	9.27（0.32）
1.3	5000	1.16	0.0600	0.0075	2.92（0.17）	9.45（0.32）

[13] Pierre Bay le, Alexandre Bay le、Lucas Janson、and Lester Mac key. C 交叉验证测试误差的置信区间，2020。

[14] Michael Kearns 和 Dana Ron。留一交叉验证的算法稳定性与合理性检验界。*Neural Comput.*，11(6):1427– 1453，1999。doi: 10.1162/089976699300016304。

[15] Olivier Bousquet 和 Andr´e Elisseeff。稳定性与泛化。*J.Mach. Learn. Res.*，2:499–526，2002。

[16] Jerzy Neyman。复合统计假设的最优渐近检验。载于 Ulf Grenander 编，概率与统计：*Harald Cram´er* 纪念文集，第 213–234 页。Almqvist & Wiksell，斯德哥尔摩，1959。

[17] Yoshua Bengio 和 Yves Grandvalet。K 折交叉验证方差的无偏估计量不存在。*J. Mach. Learn. Res.*，5:1089–1105，2004。

[18] Robert Tibshirani。通过套索回归实现回归收缩与变量筛选。《皇家统计学会杂志：*B*辑》，58(1):267–288，1996。doi: 10.1111/j.2517-6161.1996.tb02080.x。

[19] Diederik P. Kingma 和 Jimmy Ba。Adam：一种随机优化方法。载于第 *3* 届国际学习表征会议（*ICLR*），2015。[20] G. Grosso、R. T. D' Agnolo、M. Pierini、A. Wulzer 和 M. Zanetti。NPLM：学习多变量新物理，2021。[21] Bin Yu 和 Karl Kumbier。真实数据科学。美国国家科学院院刊，117(8):3920–3929，2020。doi: 10.1073/pnas.1901326117。[22] Bin Yu 和 Rebecca L. Barter。真实数据科学：负责任的数据分析与决策实践。麻省理工学院出版社，2024。[23] Bin Yu。稳定性。*Bernoulli*，19(4):1484–1500，2013。doi: 10.3150/ 13-BEJSP14。[24] Marco Letizia、Gianvito Losapio、Marco Rando、Gaia Grosso、Andrea Wulzer、Maurizio Pierini、Marco Zanetti 和 Lorenzo Rosasco。利用非参数方法高效学习新物理。欧洲物理学杂志 *C*，82(10):879，2022。doi: 10.1140/epjc/s10052-022-10830-y。[25] J. F. C. Kingman。泊松过程，牛津概率论研究第 3 卷。牛津大学出版社，牛津，1992。ISBN 9780198536932。doi: 10.1093/oso/9780198536932.001.0001。