arXiv:2611.29917v1 [stat.ME] 23 Nov 2026

用于可解释和预测性表征的张量神经网络

Shuoli Chen[1],Zhiyao Tan[1], Li Liu[2] 和 Huazhen Lin[1,3]

[1] 统计研究中心与统计与数据科学学院，西南财经大学，中国成都[2] 数学与统计学院，武汉大学，中国武汉[3] 新基石科学实验室，中国成都

2026年11月24日

摘要

张量回归面临张量值协变量的指数级维度及其复杂多模态依赖所带来的挑战。因此，现有方法通常依赖于线性张量回归和张量加性模型等简单模型。这些模型在计算上可处理，但会施加严格的结构假设，并可能限制预测准确率。为应对这些局限，我们提出了一种张量神经网络，在保留张量数据多模态结构的同时，学习低维、可解释且足以进行预测的表示。该模型能够捕捉非线性效应，从而实现准确的响应预测，同时保留经典线性张量回归的可解释性。从理论上看，我们建立了对预测而言必要的最小表示维度 K^* 的选择一致性，并证明估计量达到了表示自适应的非渐近误差率。具体而言，该误差率从环境维度阶数 $n^{-\frac{\beta}{\mathcal{D}+2\beta}}$ 提升至 $n^{-\frac{\beta}{K^*+2\beta}}$，其中$\mathcal{D}=\prod_{j=1}^{M} P_j$ 为环境维度。这表明，恢复低维表示对于缓解张量回归中的维数灾难至关重要。此外，最小化所推导的泛化误差界，可以为计算开销较低的调参提供明确的调节参数选择。大量模拟和真实数据分析表明，该方法在不同复杂度的设置下均表现稳定。不同于黑箱深度学习模型，该框架支持直接可视化和解释所学习的张量表示，无需借助外部事后分析工具。

关键词：张量　或特征、加权惩罚、可解释性、预测误差、选择一致性、轻量级调优。

1 引言

在大数据时代，多维数组数据（如磁共振成像扫描、文本、图像、视频、交通数据以及社交网络中的用户行为数据）日益普遍，通常以张量形式表示（Bhargava 等人，2015年；Ma 等人，2019年；Panagakis 等人，2021年）。然而，由于维度呈指数增长、复杂的多模态依赖，以及同时实现灵活性和可解释性的困难，张量数据分析颇具挑战性。

早期的张量数据分析研究通常将每个张量元素视为一个独立变量（Worsley 等人，2004；Lazar，2008 年；Penny 等人，2011年）。尽管这种策略简单直接，却可能忽略张量元素之间的结构相关性，导致有用信息损失。另一种常见方法是将张量向量化，然后应用高维回归技术（Cox 和 Savoy，2003；Craddock 等人，2009）。然而，向量化会破坏多模态张量结构，通常还会导致超高维回归问题；在样本量有限且生成的协变量之间存在强依赖的情况下，估计会变得困难。

另一类研究将张量值协变量视为连续域上潜在函数的离散化观测。例如，Reiss 和 Ogden(2010)；Wang 等人(2014年)针对基于图像的预测提出了函数型线性模型。这些方法利用空间或时间平滑性，从而保留了类图像数据中的重要结构信息。然而，它们通常要求存在有意义的连续域，并预先设定排序、邻域或平滑性的概念。因此，对于模态代表异质或离散特征的张量协变量，这些方法可能不太适用，例如用户行为记录，因为其中可能不存在自然的邻接关系。

为了取得更好的效果 有效利用一般张量协变量中的结构信息，张量

线性回归模型已得到广泛采用，通常形式为

$$f(\mathbf{X}) = \langle \mathbf{X}, \mathbf{B} \rangle := \sum_{j_1, \cdots, j_M} X_{j_1, \cdots, j_M} B_{j_1, \cdots, j_M},$$

其中，B $\in \mathbb{R}^{P_1 \times \cdots \times P_M}$是一个未知的系数张量，$B_{j_1,\ldots,j_M}$表示其第（$j_1, \ldots, j_M$）个元素。为捕捉多模态依赖关系，CANDECOMP/PARAFAC（CP）（Harshman，1970年）和 Tucker 分解（Tucker，1966年）等张量分解技术已广泛应用于 B（Guo 等，2011年；Zhou 等，2013年；Guhaniyogi 等，2017年；Papadogeorgou 等，2021年；Li 等，2018年）。通过对 B 施加低秩结构，这些方法在保留张量模态间内在依赖关系的同时，实现了大幅降维。为避免预先指定张量秩，研究者开发了基于正则化的方法（Zhou 和 Li，2014年；Raskutti 等，2019年），以自适应地学习低维表示。尽管取得了这些进展，上述方法仍依赖线性建模假设。

线性假设通常限制性过强，并且在现实应用中经常不成立，从而导致较低的预测准确率。因此，人们发展了经典的半参数模型和非参数模型。例如，Marx 等（2011年）针对张量协变量考虑了单指标框架，而 Hao等人（2021年）；Chen 等（2023年）；Zhou等人（2024年）研究了应用于张量协变量各元素的加性模型。人们还提出了乘积核方法（Signoretto 等，2013年；Suzuki 等，2016年；Kanagawa 等，2016年；I maizumi 和 Hayashi，2016年），以利用张量 X 的结构信息。这些方法考虑的模型范围从秩一的 $f(\mathbf{X}) = \sum_v \prod_{m=1}^{M} f_v^{(m)}(\mathbf{x}_m)$，到更灵活的秩-$D$模型，例如 $f(\mathbf{X}) = \sum_v \sum_{d=1}^{D} \prod_{m=1}^{M} f_v^{(m)}(\mathbf{x}_{m,d})$，其中 $\mathbf{X} = (\mathbf{x}1, \cdots, \mathbf{x}_M)$ 或 $\mathbf{X} = \sum_{d=1}^{D} \mathbf{x}1,_d \circ \cdots \circ \mathbf{x}_{M,d}$。尽管这些回归模型缓解了高维非参数情境下的维数灾难，但其预测性能通常受到较强模型假设或不恰当的秩和核函数选择的限制。

近年来，深度神经网络（深度神经网络）在各个领域都取得了显著成功。一些研究（Chien 和 Bao，2017年；Kossaifi 等人，2020年；Kolbeinsson 等人，2021年；da Costa 等人，2025年）已将基于张量的特征提取融入神经网络架构，从而在图像、视频和时间序列数据上实现了出色的预测性能。然而，这些架构通常是针对具有特定结构的张量数据设计的，例如空间依赖或时间依赖，因此可能无法直接应用于具有异质或无序模态的一般张量协变量。此外，深度神经网络通常需要大量训练样本，并且通常具有有限的可解释性，这可能限制其在生物学和医学等科学应用中的使用；在这些领域，数据有限且可解释性至关重要。据我们所知，目前尚不存在能够同时实现高预测准确率和可解释性的张量模型。本文旨在开发一个计算上可处理、具有高预测性能且可解释的框架，明确建模 M 阶张量协变量 $\mathbf{X}$ 与响应变量 Y之间的关系。

我们首先考虑一个一般的张量模型：

$$Y \sim g(\mathbf{X}), \tag{1}$$

其中，$\mathbf{X} \in \mathbb{R}^{P_1 \times \cdots \times P_M}$ 是一个 M阶张量预测变量，Y 是标量响应，$\mathcal{D} := \prod_{m=1}^{M} P_m$ 表示张量协变量的环境维度。函数 $g(\cdot)$刻画了张量协变量 $\mathbf{X}$ 与响应变量 Y之间的关系。例如，当 Y为连续型时，该模型可化简为 $Y = g(\mathbf{X}) + \varepsilon$，其中存在测量误差 ε；当 Y为二元型时，模型则表示为 $P(Y = 1) = \text{Sigmoid}\{g(\mathbf{X})\}$。为了构建兼具高预测性能和内在可解释性的张量模型，我们将一组张量特征提取器（记为 $\mathbf{B}_k$、$k = 1$、...、K）引入一般模型（1）中。具体而言，我们

考虑以下张量神经网络（TensorNN）：

$$Y \sim g\big(\langle \mathbf{X}, \mathbf{B}_1 \rangle, \cdots, \langle \mathbf{X}, \mathbf{B}_K \rangle\big), \tag{2}$$

其中， g 由深度神经网络进行逼近。特征数量 K 可以大到环境维度 $\mathcal{D}$。当 $K = \mathcal{D} := \prod_{m=1}^{M} P_m$ 且每个 $\mathbf{B}_k$ 都选为提取 X 中一个元素的典范指标张量时，向量（$\langle$X，$\mathbf{B}_1\rangle$，...，$\langle$X，$\mathbf{B}_K\rangle$）包含 X 的所有元素。在这种情况下，模型（2）退化为一般模型（1），表明 TensorNN 保留了一般张量回归框架的完整预测能力。重要的是，引入 $\mathbf{B}_k$、$k \geq 1$ 为识别 X 影响 Y的模式提供了机会。具体而言，每个 $\mathbf{B}_k$ 都充当一个特征提取器。通过确定 $\mathbf{B}_k$ 为零还是非零，并仅保留相应的非零特征，我们可以确定 X 中对 Y 有贡献的有效特征数量。

此外，排除与零值相关的特征$\mathbf{B}_k$ 会将输入维数从模型 $\mathcal{D}$ （1）中的 K 降低到模型（2）中的 g，从而形成一个更为简约的网络架构来估计 g，该架构通过所提出的架构惩罚项自动确定。通过自适应地同时降低输入维数和网络规模，所提出的方法提高了估计稳定性并减少了随机误差，同时与完整模型（1）相比保留了逼近能力。这些作用进一步降低了预测误差。此外，保留的特征 $\langle \mathbf{X}, \mathbf{B}_k \rangle$ 提供了与线性张量回归模型相同的可解释性，能够明确表示张量协变量如何影响 Y。

我们的贡献可总结如下。(1) 据我们所知，我们首次提出了一个同时考虑预测准确性、可解释性和张量协变量结构信息的框架。(2) 通过对特征提取器 $\mathbf{B}_1, \mathbf{B}_2, ..., \mathbf{B}_K$施加组惩罚，我们能够识别 K^*，即确保预测所需逼近能力的最小表示维度。

(3)

在理论方面，我们建立了 K^* 的选择一致性，并证明所得 TensorNN 估计量达到了 $C_* n^{-2\xi/(2\xi+K^*)}$阶的非渐近预测误差界（略去对数因子），其中 ξ 是低维函数g的平滑度。因此，收敛速率由 K^*决定，而不是由环境张量维度 $\mathcal{D}$决定。对于由深度神经网络估计的一般 d维函数 f(X)，近期研究（Bauer and Kohler，2019年；Chen 等，2019年；Nakada and Imaizumi，2020年；Jiao 等，2023年）已建立了 $C_d n^{-\frac{2\xi}{2\xi+d^*}}$的收敛速率，其中 d^* 表示低维内在维度。然而，前因子 C_d 通常会随环境维度 d呈多项式甚至指数增长，这严重限制了其在高维环境中的可扩展性。相比之下，我们的收敛速率包含前因子 C_*，其取决于 K^*，且仅随环境张量维度 $\mathcal{D}$对数增长。这显著缓解了维数灾难，使我们的方法能够有效处理 $\mathcal{D} \gg n$的超高维非参数问题。(4) 通过最小化定理1中的预测误差上界，可以明确指定相对惩罚权重 $\boldsymbol{\tau} := (\tau_1、\ \tau_2、\ \tau_3)$，它们分别控制对特征指示变量、特征分量和网络规模的惩罚。因此，在选择足够丰富的工作架构后，主要的正则化调参可简化为单个参数 λ，该参数控制整体惩罚强度。这带来了较低的计算复杂度和便于用户使用的实际实现。TensorNN-op 展示了这种轻量级调参策略的有效性：它采用理论驱动的相对权重，并在表2、表3和表4的模拟研究与真实数据研究中均取得了出色表现。(5) 我们的模型通过学习到的张量特征提取器提供了模型层面的可解释性，这些提取器充当协变量 X 上合成的张量模式。在模拟研究中，图2中突出显示的区域与图1中预先定义的活跃区域高度一致，表明所提出的方法能够恢复具有实际意义的潜在张量模式。在真实数据

在这些应用中，预测特征往往不规则且分散于空间各处；如图3所示，使用多个张量特征提取器使 TensorNN 能够捕捉复杂的结构化信号。例如，在微笑面部图像实验中，TensorNN 将嘴部、脸颊和眼睛周围区域突出显示为重要的预测特征。相比之下，在闭嘴分类任务中，TensorNN 主要突出显示一个局部结构化的嘴部区域。这些结果为学习到的表示提供了可解释的张量级摘要。

所提出的方法在模拟研究和真实数据分析中均展现出强大的预测性能。在广泛的实验设置中，包括简单和复杂的数据生成机制，以及表格数据和图像数据等多种张量类型，我们的方法始终优于现有方法。在四个真实世界数据集上的应用结果表明，与卷积神经网络（CNN）和传统张量回归方法相比，该方法取得了明显的改进，其中包括 STAR（Hao等人，2021年）和 TLR（Zhou等人，2013年），如表2、3和4所示。

本文其余部分安排如下。在第2节中，我们提出一种惩罚估计方法，同时学习特征提取器$\mathbf{B}_k$、其数量 K以及网络，以近似模型（中的函数2）。第3 节建立估计量的理论性质，包括特征维数的一致性以及泛化误差的上界。第4 节和第5 节介绍一系列模拟研究和四项现实应用。为简洁起见，省略所有技术性证明。

2 张量神经网络与估计

记 $[K]=\{1,\cdots,K\}$、$\mathbf{B}_{[K]}=\{\mathbf{B}_k:\ k\in[K]\}$，以及 $\langle\mathbf{X},\mathbf{B}_{[K]}\rangle=\big(\langle\mathbf{X},\mathbf{B}_1\rangle,\ \cdots,\ \langle\mathbf{X},\mathbf{B}_K\rangle\big)$。如第1节所述，构建具有强预测性且可解释的

张量回归模型的一项关键策略，是将$\mathbf{B}_k$纳入模型（1），从而得到（2）中的框架。然而，张量$\mathbf{B}^{[K]}$包含 KD 个未知参数，通常远超样本量 n。这种超高维模型带来了巨大的计算负担，并削弱了模型的泛化能力（Kawaguchi 等人，2022年）。为减少参数数量，同时捕捉张量协变量之间的内在依赖关系，我们采用张量回归中的标准方法，对每个Zhou等人，2013年；Hao等人，2021年；Zhou等人，2024年）执行CP分解$\mathbf{B}_k$ 如下：

$$\mathbf{B}_k = \sum_{r=1}^{R} \gamma_{kr} \boldsymbol{\beta}_{kr1\bullet} \circ \cdots \circ \boldsymbol{\beta}_{krM\bullet}, \tag{3}$$

其中 $\boldsymbol{\beta}_{krm\bullet} = (\beta_{krmj}, j \in [P_m])^T$ 沿第四个（位置）维度收集元素，‘$\circ$’表示向量外积，秩 R 通常远小于 $\min_j\{P_j\}$。在分解（3）下，$\mathbf{B}^{[K]}$ 中的参数数量从 KD 减少到 $KR\big(1 + \sum_{m=1}^{M} P_m\big)$，同时尽可能保留张量的内在结构信息（Kolda 和 Bader，2009）。随后，特征提取器$\mathbf{B}^{[K]}$ 的参数定义为

$$\boldsymbol{\theta}_{fea} := \{\boldsymbol{\gamma}_{1\bullet}, \cdots, \boldsymbol{\gamma}_{K\bullet}, \boldsymbol{\beta}_{111\bullet}, \cdots, \boldsymbol{\beta}_{KRM\bullet}\} \in \mathbb{R}^{S_{fea}}, \tag{4}$$

其中 $\boldsymbol{\gamma}_{k\bullet} = \{\gamma_{kj}, j \in [R]\}$, $S_{fea} = KR + KR\sum_{m=1}^{M} P_m$ 是特征提取器的参数总数。

我们使用具有 ReLU激活的前馈神经网络（FNN）来近似目标函数 g。具有 L 个隐藏层的 ReLU前馈神经网络是一种映射 g： $\mathbb{R}^{H_0} \to \mathbb{R}^{H_{L+1}}$，其形式为$g$(x) $= T_L \circ \sigma \circ T_{L-1} \circ \cdots \circ \sigma \circ T_0$(x)，其中 $g_1 \circ g_2$(x)：$= g_1(g_2$(x)$)$，ReLU函数σ(x)：$=$ max(x, 0)按元素应用，而 T_l(x)：$=$ W$_l$x $+$ a$_l$ 是一个仿射变换，其中 W$_l \in \mathbb{R}^{H_{l+1} \times H_l}$，a$_l \in \mathbb{R}^{H_{l+1}}$，对于 $l = 0$，1，...，L，H_l 表示第l层中的神经元数量。我们将 $H =$ max $\{H_1, H_2, ..., H_L\}$ 和 L 分别称为网络的宽度和深度。大小$S_{net} = \sum_{l=0}^{L} H_{l+1}H_l + \sum_{l=0}^{L} H_{l+1}$ 是网络中的参数总数。令

$\mathcal{G}(K,\ H,\ L)$表示输入维度为 K、宽度为 H且深度为 L的 ReLU 网络类。因此，与网络相关的参数为

$$\boldsymbol{\theta}_{net} := \left\{\mathbf{W}_0, \cdots, \mathbf{W}_L, \mathbf{a}_0, \cdots, \mathbf{a}_L\right\} \in \mathbb{R}^{S_{net}}. \tag{5}$$

神经网络的收敛速率在很大程度上取决于输入维度（BauerandKohler，2019年）；在高维情形下，这种依赖可能导致维数灾难。在我们的框架中，表示维度 K决定了 g 的输入维度，因此也控制着特征提取器 $\mathbf{B}^{[K]}$ 和后续网络的复杂度。如果 K 设定过大，冗余成分和网络参数会增大估计方差，并且需要更大的样本量才能准确估计。如果 K 设定过小，模型就缺乏足够的表示能力，导致逼近误差增大。因此，选择能够确保表示能力充足、同时将随机误差降至最低的最小 K 至关重要。

此外，关于神经网络的理论分析（Schmidt-Hieber，2020年；Tan 等人，2024年）表明，网络 g 的规模应根据特征维度 K 进行调整，才能达到最优收敛速率。一种直接的策略是在候选值 K 和网络架构组成的网格中进行搜索，然后选择预测误差最小的组合。然而，这种策略有两个明显的缺点。首先，它的计算成本很高，因为每个候选组合都需要单独训练和验证。其次，选择完全由实证表现决定，无法直接提供理论保证。为了解决这些局限，我们提出一种正则化方法，在保留理论保证的同时，联合学习特征维度、特征提取器，以及针对所学维度进行调整的网络。

特别地，记作 $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{net}, \boldsymbol{\theta}_{fea}\}$。给定独立同分布样本 $\mathbf{S}^n \equiv (\mathbf{X}_i,\ Y_i)_{i=1}^n$ ，其中

总体 $\mathbf{S} = (\mathbf{X}\ Y)$，样本量为 n，我们通过以下方式估计 $\boldsymbol{\theta}$

$$\hat{\boldsymbol{\theta}} \in \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \frac{1}{n} \sum_{i=1}^{n} \ell\big(g_{\boldsymbol{\theta}_{net}}\big(\langle \mathbf{X}_i, \mathbf{B}_{[K]} \rangle\big), Y_i\big) + \mathcal{P}(\boldsymbol{\theta}), \tag{6}$$

其中，$g_{\boldsymbol{\theta}_{net}}(\cdot) \in \mathcal{G}(K、H、L)$ 由 $\boldsymbol{\theta}_{net}$ 指定，$\ell(\cdot、\cdot)$ 是损失函数，例如用于连续响应的平方损失和用于二元响应的逻辑损失，

$$\mathcal{P}(\boldsymbol{\theta}) = \lambda \|\boldsymbol{\theta}\|_{\boldsymbol{\tau}} := \lambda\big(\tau_1 \sum_{k \in [K]} \|\boldsymbol{\gamma}_{k\bullet}\|_2 + \tau_2 \sum_{(k,m,j) \in [K,M]} \|\boldsymbol{\beta}_{k\bullet mj}\|_2 + \tau_3 \|\boldsymbol{\theta}_{net}\|_1\big), \tag{7}$$

$\boldsymbol{\beta}_{k\bullet mj} = (\beta_{krmj}, r \in [R])^{\mathrm{T}}$ 沿秩维度收集元素，$[K, M] := \{(k, m, j)\ |k \in [K]、\ j \in [P_m]、\ m \in [M]\}$，而 $\|\cdot\|_v$ 是向量或矩阵的 v-范数。第一个惩罚项通过检测并排除 $\{\mathbf{B}_k$ 中的零特征提取器来确定特征数量：$k \in [K]\}$ 因为 $\boldsymbol{\gamma}_{k\bullet} = \mathbf{0}$ 意味着$\mathbf{B}_k = \mathbf{0}$。结合式（3）中的张量表达式，第二项对 $\boldsymbol{\beta}_{k\bullet mj}$施加组惩罚，即$\mathbf{B}_k$ 沿模式 m 的第 j个切片在所有CP 分量上的集合。当 $\boldsymbol{\beta}_{k\bullet mj} = \mathbf{0}$时，X沿模式 m的相应第 j个切片与特征 k无关，从而能够识别出与特征 k相关的张量协变量。这进一步阐明了特征 k的含义，从而增强可解释性并降低模型复杂度。第三项将冗余网络权重收缩至0附近，使网络规模与有效维度（即选定特征的数量）相匹配。因而，通过适当调节参数 λ、τ_1、τ_2 和 τ_3，即可自动确定特征数量和网络规模。在下一节中，我们通过最小化定理1所建立的预测误差可达上界，推导相对权重 $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$的显式表达式。因此，只需调节单个超参数 λ，即可控制整体正则化强度，从而大幅降低计算负担。

3 理论分析

考虑的真实函数 g 在（2）中的H¨older类 $\mathcal{H}^{\xi}\left([-D,D]^{K},\ B_{0}\right)$ 定义为

$$\left\{g:[-D,D]^{K}\to\mathbb{R},\max_{\|\boldsymbol{\alpha}\|_{1}\leq\lfloor\xi\rfloor}\|\partial^{\boldsymbol{\alpha}}g\|_{\infty}\leq B_{0},\max_{\|\boldsymbol{\alpha}\|_{1}=\lfloor\xi\rfloor}\sup_{\mathbf{x}\neq\mathbf{x}^{\prime}}\frac{|\partial^{\boldsymbol{\alpha}}g(\mathbf{x})-\partial^{\boldsymbol{\alpha}}g(\mathbf{x}^{\prime})|}{\|\mathbf{x}-\mathbf{x}^{\prime}\|_{2}^{\xi-\lfloor\xi\rfloor}}\leq B_{0}\right\},$$

其中 $\lfloor\xi\rfloor$ 表示严格小于 ξ 的最大整数，$\partial^{\boldsymbol{\alpha}}g=\partial^{|\boldsymbol{\alpha}|}g/\partial x_{1}^{\alpha_{1}}\cdots\partial x_{K}^{\alpha_{K}}$ 表示 $\boldsymbol{\alpha}=$ 阶混合偏导数（α_{1}，.....，α_{K}）$^{T}\in\mathbb{N}_{0}^{K}$，其中 $|\boldsymbol{\alpha}|=\sum_{i=1}^{K}\alpha_{i}$。

我们使用 ReLU前馈神经网络估计 g 。相应的函数类定义为 $\mathcal{F}(K,H,L)$: $=\left\{g_{\boldsymbol{\theta}_{\mathrm{net}}}\left(\langle\cdot,\mathbf{B}_{[K]}\rangle\right):\ g_{\boldsymbol{\theta}_{\mathrm{net}}}\in\mathcal{G}(K,H,L)\right\}$，本文也将其称为 TNN类。对于任意函数 f，记 $\mathbb{P}\ell_{f}:=\mathbb{E}\{\ell(f(\mathbf{X})\ Y)\}$，以及 $\mathbb{P}\mathcal{L}_{f}:=\mathbb{P}(\ell_{f}-\ell_{f_{0}})$，其中 $\ell_{f}:=\ell(f(\mathbf{X})\ Y)$，而 f_{0} 表示定义 1 中指定的真实函数。进一步记 $\|f\|_{L_{2}(\mathbb{P}_{\mathbf{X}})}:=\{\mathbb{E}[f(\mathbf{X})^{2}]\}^{1/2}$，$\|f\|_{\infty}:=\sup_{\mathbf{X}}|f(\mathbf{X})|$，$\|\mathbf{C}\|_{1,\ \infty}:=\max_{j}\sum_{i}|C_{ij}|$，并用 $|\cdot|$ 表示逐元素绝对值运算。由于通过将每层的输入扩展为常数 1，可以将偏置项吸收到权重矩阵中，因此不失一般性地假设所有层 l 的 $\mathbf{a}_{l}=0$ 。

在全文中，下标“0”表示真实值，而上标“$*$”表示相应的最小量。由于 K 和 $\mathbf{B}^{[K]}$ 的选择具有灵活性，且神经网络参数化存在不可辨识性，模型（2）具有很强的不可辨识性。因此，我们定义以下真实函数类。

定义 1（真实函数类）。我们将真实函数类定义为 $\mathcal{F}_{0}:=\{f_{0}:\ f_{0}\in\operatorname{argmin}_{h\in\mathcal{F}}\mathbb{E}\{\ell(h(\mathbf{X}),Y)\}\}$，其中 $\mathcal{F}=\bigcup_{K\leq\mathcal{D}}\left\{g\left(\langle\mathbf{X}、\ \mathbf{B}_{[K]}\rangle\right)\mid g\in\mathcal{H}^{\xi}\left([-D,D]^{K},B_{0}\right)\right\}$。

由于真实函数可能对应不同数量的提取特征，我们引入以下最小特征维度的概念。

定义2 (最小特征维度**)**。 最小特征维度定义为$K^{*}:=$最小值$\left\{K:\ \ g(\langle\cdot,\mathbf{B}_{[K]}\rangle)\in\mathcal{F}_{0}\right\}$.

那么，已知 K^* 时，真实的预言机函数类定义为

$$\mathcal{F}_0^{or} := \{g(\langle \cdot, \mathbf{B}_{[K^*]} \rangle) \in \mathcal{F}_0\},$$

相应的已知 K^* 的预言机参数空间记为

$$\mathbf{\Theta}^{or}(K^*, H, L) := \{\boldsymbol{\theta} : \|f_{\boldsymbol{\theta}} - f_0\|_\infty \leq \epsilon_n(f_0), f_{\boldsymbol{\theta}} \in \mathcal{F}(K^*, H, L), f_0 \in \mathcal{F}_0^{or}\}, \tag{8}$$

其中，$\epsilon_n(f_0) := \min_{f \in \mathcal{F}(K^*,H,L)} \|f - f_0\|_\infty$. 那么，$\epsilon_n := \max^{f_0 \in \mathcal{F}_0^{or}} {}^{\epsilon_n}(f_0)$ 是真实函数类与具有最少特征数 K^*的网络函数类之间的逼近误差。

3.1 非渐近高概率界

为建立估计量的性质，我们引入以下假设。

假设 1。（有界性）。协变量满足 $\mathbf{X} \in [-1,1]^{P_1 \times \cdots \times P_M}$。存在常数 C_x、B、$F > 0$，使得 $\|vec(\mathbf{X})\|_1 \leq C_x$、$\|\boldsymbol{\theta}\|_\infty \leq B$、$\|f_0\|_\infty \leq F$，并且对于所有 $f \in \mathcal{F}(K$、H、$L)$，$\|f\|_\infty \leq 2F$。

假设2.（*Lipschitz*损失函数）。对于任意有界函数 f_1 和 f_2, $\big|\ell\big(f_1(\boldsymbol{x})\ y\big) - \ell\big(f_2(\boldsymbol{x})\ y\big)\big| \leq \big|f_1(\boldsymbol{x}) - f_2(\boldsymbol{x})\big|$.

假设3.（曲率）。存在常数 $c_f, C_f > 0$ 使得对于每个 $f \in \mathcal{F}(K, H, L)$ $c_f \|f - f_0\|^2_{L_2(\mathbb{P}_{\mathbf{X}})} \leq \mathbb{P}\mathcal{L}_f \leq C_f \|f - f_0\|^2_{L_2(\mathbb{P}_{\mathbf{X}})}$.

假设 1 中的有界性条件是深度神经网络理论中的标准条件（Schmidt-Hieber，2020年；Ohn andKim，2022；Mengand Li，2026）。假设 2 要求损失函数满足 Lipschitz 条件，而假设 3 则施加 Bernstein 型曲率条件，将超额风险与预测误差联系起来。经验风险分析中广泛采用与假设 2和3 类似的条件；例如参见Bartlett等人(2005)；

Koltchinskii (2006 年)；Alquier et al. (2019年)，关于一般非参数估计理论；以及 Farrell et al. (2021年)，关于神经网络估计。将 Lipschitz 常数归一化为 1 不失一般性，因为对损失函数进行缩放只会改变后续界中的常数。在这些正则性条件下，我们接下来确定一个能够达到所需逼近误差的模型稀疏子类。

引理 *1.* 在假设 *1* 下，当 K^* 已知时，对于任意正整数 l 和 $w_n = \lceil n^{K^*/\{2(2\xi+K^*)\}} \rceil$，深度为 $L \geq L^* := 21(\lfloor \xi \rfloor + 1)^2 l \lceil \log_2(8l) \rceil + 2K^*$、宽度为 $H = H_1 = \cdots = H_L \geq H^* := 38(\lfloor \xi \rfloor + 1)^2 3^{K^*} (K^*)^{\lfloor \xi \rfloor + 1} w_n \lceil \log_2(8w_n) \rceil$ 的 *TNN*类 $\mathcal{F}(K^*, H, L)$ 满足 $\epsilon_n \leq c_\epsilon n^{-\frac{\xi}{2\xi+K^*}}$，其中 $c_\epsilon = 19(2D)^\xi B_0(\lfloor \xi \rfloor + 1)^2 (K^*)^{\lfloor \xi \rfloor + \max(\xi,1)/2} l^{-2\xi/K^*}$。

引理 1表明，对于任意 $f_0 \in \mathcal{F}_0^{or}$，存在一个张量神经网络函数 $f \in \mathcal{F}(K^*, H, L)$，其网络规模为 S_{net}，满足 $S_{net} \geq S_{net}^* \asymp L^*(H^*)^2 = O(n^{\frac{K^*}{2\xi+K^*}} \log_2^2 n)$，使得 $\|f - f_0\|_\infty = O(n^{-\frac{\xi}{2\xi+K^*}})$。

记 $J^* := \min\left\{ \sum_{(k,\ m,\ j)\in[K,M]} 1_{\{\beta_{k\bullet mj} \neq \mathbf{0}\}} \mid g(\langle \cdot,\ \mathbf{B}_{[K]} \rangle) \in \mathcal{F}_0,\ K = K^* \right\}$ 为特征分量的最小数量；$\mathcal{F}^*$ 为具有 $K \leq K^*$、$J \leq J^*$ 和 $S_{net} \leq S_{net}^*$ 的模型的稀疏子类；以及 $\|\boldsymbol{\theta} - \boldsymbol{\theta}\|_{\boldsymbol{\tau},\ \tilde{\mathcal{I}}} := \tau_1 \sum_{\tilde{\gamma}_{k\bullet} \neq \mathbf{0}} \|\boldsymbol{\gamma}_{k\bullet} - \tilde{\boldsymbol{\gamma}}_{k\bullet}\|_2 + \tau_2 \sum_{\tilde{\boldsymbol{\beta}}_{k\bullet mj} \neq \mathbf{0}} \|\boldsymbol{\beta}_{k\bullet mj} - \boldsymbol{\beta}_{k\bullet mj}\|_2 + \tau_3 \sum_{(\tilde{\theta}_{net})_j \neq 0} |(\theta_{net})_j - (\theta_{net})_j|$。

假设*4*。（受限下*Lipschitz*条件）。对于满足 $\tau_i^{-1} = O(1)$、$i = 1$、*2*、*3*的任意调节向量 $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$，存在一个常数 $\kappa_{\boldsymbol{\tau}} > 0$，使得对于任意 $f_{\boldsymbol{\theta}} \in \mathcal{F}(K, H, L)$以及 $f_{\tilde{\boldsymbol{\theta}}} \in \mathcal{F}^*$，只要其代表元 $\boldsymbol{\theta} \in \mathrm{op}_{\boldsymbol{\tau}}(0, f_{\boldsymbol{\theta}})$ $\boldsymbol{\theta} \in \mathrm{op}_{\boldsymbol{\tau}}(\boldsymbol{\theta}, f_{\tilde{\boldsymbol{\theta}}})$满足锥条件$\|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\boldsymbol{\tau}} \leq \frac{2}{1-\eta} \|\boldsymbol{\theta} - \tilde{\boldsymbol{\theta}}\|_{\boldsymbol{\tau},\ \tilde{\mathcal{I}}}$、$\eta \in (0, 1)$，则有 $\|f_{\boldsymbol{\theta}} - f_{\tilde{\boldsymbol{\theta}}}\|_{L_2(\mathbb{P}_{\mathbf{X}})} \geq \kappa_{\boldsymbol{\tau}} \|\boldsymbol{\theta} - \boldsymbol{\theta}\|_{\boldsymbol{\tau},\tilde{\mathcal{I}}}$。

这里，$\mathrm{op}_{\boldsymbol{\tau}}(\boldsymbol{\theta},\ f)$ 表示在加权范数 $\|\cdot\|_{\boldsymbol{\tau}}$下，与参考参数 $\boldsymbol{\theta}$ 距离最近的函数 f的参数化集合。由于缩放、置换及其他不变性，CP 张量分解和神经网络都可能存在同一函数的多种参数化。因此，全局下 Lipschitz 条件无法成立，该条件要求将原始参数空间中的距离与诱导函数空间中的距离联系起来，

这通常是不可能的。因此，假设 4可解释为：在考虑 TensorNN 函数的非唯一参数化后，施加受限的局部可辨识性条件。具体而言，该条件仅在通过 $\mathrm{op}_{\tau}(\boldsymbol{\theta}, f)$ 选出最近的稀疏表示后，且仅沿稀疏锥方向，才要求满足下 Lipschitz 型要求。从这个意义上说，假设 4是高维惩罚回归中受限特征值（RE）条件（Bickel 等，2009年；Bellec 等，2018年）的非线性类比。RE 条件是惩罚估计理论中的标准条件：它排除了那些几乎不改变或完全不改变拟合函数的非平凡稀疏扰动。同样，假设 4确保在稀疏预言表示附近，函数差异能够充分反映相关的稀疏参数方向。函数估计与参数惩罚项之间的这种协调，使惩罚项能够推动估计量趋向稀疏预言表示。

定义 $M_n := \max_{l=0,\cdots,L+1} \left(\left\| |\mathrm{W}|_{L:l} \right\|_1 \cdot \left\| |\mathrm{W}|^{(l-2):-1} \right\|_{1,\infty} \right)$，其中 $|\mathrm{W}|_{l:l'} := |\mathrm{W}_l| \cdots |\mathrm{W}_{l'}|$，$\mathrm{W}_{-1} = \mathrm{B}$，且 $\mathrm{B} = (\mathrm{vec}(\mathrm{B}_1) \ldots, \mathrm{vec}(\mathrm{B}_K))^{\top}$。该量用于控制 $\mathcal{F}(K, H, L)$ 的复杂度。令 π_n 表示一个趋于零的失败概率上界。为简洁起见，省略 π_n 的显式表达式以及常数 c_0、C_{ρ}、τ、C_* 和 $C_{\boldsymbol{\tau}}$。

定理1. 假设假设1–4成立。设 M、R 和 τ_i^{-1}，$i = 1, 2, 3$ 的阶为常数，$\max\{K, \max_{1 \le m \le M} P_m, S_{\mathrm{net}}\} = O\{\mathrm{poly}(n)\}$，$L = O(L^*)$，且 $\boldsymbol{\Theta}^{or}(K^*, H^*, L^*)$ 中的预言参数满足 $\max_l\{\|W_l^{or}\|_{1,\infty}, \|W_l^{or}T\|_{1,\infty}, \|B^{or}\|_{1,\infty}\} \le F$。进一步假设 $M_n = O(n)$ $J^* = O\left(n^{\frac{K^*}{2\xi+K^*}} \log_2^2 n\right)$。若 $n > \max\left\{R(K^* + J^*) + S^*_{net}, \frac{1}{r_{\tau}^2}\max(1, 201F^2 \log_{\sqrt{2}} n, \sqrt{2}^{225F^2} \log n)\right\}$，且调节参数 λ 满足

$$\lambda \in \left(\frac{r_{\boldsymbol{\tau}}}{\eta}(2c_0 + \delta_n),\ r_{\boldsymbol{\tau}}\{2(c_f - c_0) - \delta_n\} \right)$$

对于 $\eta \in (0, 1)$，其中 $r_{\boldsymbol{\tau}} = \widetilde{C}_* n^{-\frac{\xi}{2\xi+K^*}} \log_2 n + \widetilde{C}_{\boldsymbol{\tau}} n^{-1/2} O\{\mathrm{poly}(\log n)\}$，$\delta_n = \{C_f \epsilon_n^2 + \epsilon_n \sqrt{2\tilde{\gamma}/n} +$

$4F\tilde{\gamma}/n\}/(C_{\rho,\,\boldsymbol{\tau}} r_{\boldsymbol{\tau}}^2) = o(1)$ 满足 $\tilde{\gamma} = \widetilde{C}_* n^{\frac{K^*}{2\xi+K^*}}$，则

$$\|f_{\widehat{\boldsymbol{\theta}}} - f_0\|_{L_2(\mathbb{P}_{\mathbf{X}})}^2 \leq 2C_{\rho,\boldsymbol{\tau}} r_{\boldsymbol{\tau}}^2 \tag{9}$$

以至少 $1 - \pi_n$ 的概率。

定理 1 为 TensorNN 提供了预言机型预测保证。尽管估计量是在扩展工作类 $\mathcal{F}(K, H, L)$ 上计算的，但其误差界会适应内在稀疏预言机结构。具体而言， $r_{\boldsymbol{\tau}}^2$ 中的主导项阶为 $n^{-2\xi/(2\xi+K^*)}$（相差对数因子），与 K^* 维 ξ 光滑非参数回归的经典极小极大速率相匹配（Stone，1982）。因此，有效非参数维度是学习到的特征维度 K^*，而非环境张量维度 $\mathcal{D}$。

此外，三部分惩罚项在不同层次上诱导稀疏性：对 $\boldsymbol{\gamma}$ 的惩罚项用于选择活跃张量特征，对 $\boldsymbol{\beta}$ 的惩罚项用于识别每个特征内相关的张量分量，而对 $\boldsymbol{\theta}_{\text{net}}$ 的惩罚项则用于正则化外层神经网络。因此，随机误差取决于预言机稀疏度 $s_n = R(K^* + J^*) + S_{\text{net}}^*$，而非工作模型中的全部参数数量。因此，过度设定的 K、H 和 S_{net} 仅会通过对数因子影响速率，从而可以将工作模型选得足够大以实现近似，同时由惩罚项控制有效模型复杂度。

张量回归的一个关键特征是其超高维特性，即张量环境维度 $\mathcal{D}$ 远大于样本量 n。以往关于神经网络的收敛理论通常假设输入维度固定或适度增长。近期研究，例如 Bauer and Kohler (2019年);Chen 等 (2019年);Nakada andImaizumi (2020年); Jiao 等(2023年)，推导出了如下形式的速率 $C_d n^{-\frac{2\xi}{2\xi+d^*}}$，其取决于内在维度 d^*。然而，前因子 C_d 通常会随着名义维度呈多项式甚至指数增长 d。当名义

输入维度为张量环境维度 $\mathcal{D} \gg n$时，情况便是如此。相比之下，定理1 表明，环境张量维度仅通过对数因子影响随机误差，例如 log K、log($\sum_{m=1}^{M} P_m$) 和 log S_{net}，从而能够有效应用于超高维张量回归。

定理1还为调参提供了实用指导。定理1中的条件$\lambda \in \left(\frac{r_{\boldsymbol{\tau}}}{\eta}(2c_0 + \delta_n)\ r_{\boldsymbol{\tau}}\{2(c_f - c_0) - \delta_n\}\right)$表明，全局惩罚水平$\lambda$ 应选择为与 $r_{\boldsymbol{\tau}}$同一量级，后者反映了 TensorNN 类的内在复杂度。因此，内在复杂度较大的模型需要更强的正则化来控制随机误差。这与其他惩罚估计量中正则化参数的作用一致（Lecu´e 和 Mendelson，2018年；Alquier 等，2019年）。此外，$r_{\boldsymbol{\tau}}$ 的显式上界与 $1/\tau_1$、$1/\tau_2$、$1/\tau_3$呈线性关系。因此，我们可以优化相对权重 $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$，以最小化 $r_{\boldsymbol{\tau}}$，从而得到：

推论 1. 在定理 *1*的条件下，若 $S_{net} \asymp \text{poly}(K)$ 且 $P_m \asymp \text{poly}(K)$ 在 m上统一成立，则速率项 $r_{\boldsymbol{\tau}}$ 在相差一个对数因子 $\log \nu$的意义下，由以下相对权重 $\boldsymbol{\tau} = (\tau_1, \tau_2, \tau_3)$最小化，其中

$$\tau_1 = \frac{\sqrt{\nu_1}}{\sqrt{\nu_1} + \sqrt{\nu_2} + \sqrt{\nu_3}}, \tau_2 = \frac{\sqrt{\nu_2}}{\sqrt{\nu_1} + \sqrt{\nu_2} + \sqrt{\nu_3}}, \tau_3 = \frac{\sqrt{\nu_3}}{\sqrt{\nu_1} + \sqrt{\nu_2} + \sqrt{\nu_3}}, \tag{10}$$

其中 $\nu_1 = \sqrt{R}$对数$\{2RK \cdot \nu\}$, $\nu_2 = \sqrt{R}$对数$\{2RK\ (\sum_{m=1}^{M} P_m)\ \cdot \nu\}$，以及 $\nu_3 =$ 对数$\{2S_{net} \cdot \nu\}$，其中某个 $\nu \geq 1 + C_\nu n M_n$ 满足 C_ν，且其取值取决于 L。

推论 *2.* 在推论 *1* 的条件下，进一步假设 *(10)* 中的 ν 为 $O(\text{poly}(n))$ 阶。则对于任何满足 $\tau_1 + \tau_2 + \tau_3 = 1$、$\tau_i^{-1} = O(1)$、$i = 1$、*2* 、*3* 或 $\tau_1 \geq \frac{\sqrt{\nu_1}}{\sqrt{\nu_1}+\sqrt{\nu_2}+\sqrt{\nu_3}}$ 的正权重 (τ_1, τ_2, τ_3)，K^* 依赖的最优速率 *(9)* 均成立，其中 ν_1、ν_2、ν_3 定义于 *(10)* 中。

定理 1 和推论 1 及 2为论文中提出的轻量级调参策略提供了理论依据：无需联合搜索多个惩罚参数和网络

架构，TensorNN 可以采用理论引导的相对惩罚权重和足够丰富的工作架构进行实现，仅需对网络深度进行粗略选择$L = O(L^*)$，并调节整体正则化强度 λ。第 4 节和 5节中的实证结果支持了这一策略的有效性。具体而言，TensorNN- op 采用推论1所规定的相对惩罚权重，在模拟研究和真实数据分析中均 consistently 表现出色。

3.2 特征选择的一致性

为建立 的选择一致性 K^*，我们进一步引入假设 5 和 6。

假设5。 *(导数正则性)*

- 对于任意 $|\theta| \leq 2F$，损失函数满足 $|\partial_1^2 \ell(\theta, y)| \leq C_\ell$. 此外，给定 $\{\mathbf{X}_i\}_{i=1}^n$后，得分变量 $\partial_1 \ell(f_0(\mathbf{X}_i)\ Y_i)\ i = 1, \ldots, n$ 相互独立、均值为零，且为次高斯变量，其方差代理为 σ_e^2，其中 $\partial_1 \ell(\cdot, \cdot)$和 $\partial_1^2 \ell(\cdot, \cdot)$分别表示损失函数 ℓ 关于其第一个自变量的一阶和二阶偏导数。

- 设 $\boldsymbol{\theta}^*$ 为一个最优稀疏*TensorNN*参数，满足 $f_{\boldsymbol{\theta}^*} \in \mathcal{F}^*$ 和 $f_{\boldsymbol{\theta}^*} = f_{\boldsymbol{\theta}_{\text{或}}}$，其中 $\boldsymbol{\theta}_{\text{或}}$ 定义于 (8)。对于任意满足 $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|_{\boldsymbol{\tau}} = O(r_{\boldsymbol{\tau}})$ 的 $\boldsymbol{\theta}$，$g_{\boldsymbol{\theta}_{\text{网络}}}$ 关于第 k 个特征输入的 ***Clarke*** 次微分满足 $\left\{\frac{1}{n}\sum_{i=1}^n \left|\partial_k g_{\boldsymbol{\theta}^{\text{网络}}}(\langle \mathbf{X}_i, \mathbf{B}_{[K]}\rangle)\right|^2\right\}^{1/2} \leq C_g, k = 1, \ldots, K$. 此外，对于所有满足 $\sum_{k=1}^K |\Delta_k| = O(r_{\boldsymbol{\tau}})$ 的扰动 $\Delta_1, \ldots, \Delta_K$，局部变分条件

$$\sum_{k=1}^K \left\{\frac{1}{n}\sum_{i=1}^n \left|\partial_k g_{\boldsymbol{\theta}_{\text{net}}}(\langle \mathbf{X}_i, \mathbf{B}_{[K]}\rangle + \mathbf{e}_k \Delta_k) - \partial_k g_{\boldsymbol{\theta}_{\text{net}}}(\langle \mathbf{X}_i, \mathbf{B}_{[K]}\rangle)\right|^2\right\}^{1/2} \leq C_g$$

成立，其中 $\mathbf{e}_k$ 表示第 k 个标准基向量。

假设5包含两个正则性要求。第一项要求损失函数具有局部光滑性，并且在真实值处满足次高斯得分条件。这些是高维广义非参数回归模型中的标准假设（Haris等人，2022；Yang和Maiti，2022）。例如，对于具有$\ell(\theta, y) = -\{y\theta - \psi(\theta)\}$、$\psi'(f_0(\mathbf{X}_i)) = E(Y_i \mid \mathbf{X}_i)$的广义非参数模型，得分函数$\partial_1 \ell(f_0(\mathbf{X}_i)\ Y_i)$的条件均值为零，并且对于常见模型，曲率项$\partial_1^2 \ell(\theta, y) = \psi''(\theta)$在紧致域上是有界的。假设5的第二部分控制外层神经网络关于每个张量特征的敏感性。由于 ReLU 网络是局部Lipschitz的，但并非处处可微，因此Clarke次微分提供了适当的广义导数（Clarke，1990）。这些次微分的有界性和局部变分条件确保张量特征的小扰动不会导致特征层面得分发生不稳定的变化。这些是稀疏预言表示附近的局部条件，并且随着$r_{\boldsymbol{\tau}} \to 0$而变得不那么严格。

假设6。（特征信号强度）存在一个足够大的常数$C_0 \geq \widetilde{0}$，使得对于任意$K < K^*$、任意张量特征参数$\mathbf{B}^{[\widetilde{K}]}$以及任意可测函数$g$：$\mathbb{R}^{\widetilde{K}} \to \mathbb{R}$，$\left\{ \frac{1}{n} \sum_{i=1}^{n} \left[f_0(\mathbf{X}_i) - g\big(\langle \mathbf{X}_i, \mathbf{B}_{[\widetilde{K}]} \rangle\big) \right]^2 \right\}^{1/2} \geq C_0 r_{\boldsymbol{\tau}}$.

假设 6要求真实回归函数即使在经验意义下，也不能由使用少于K^*个张量特征的任何模型近似。因此，K^*个真特征中的每一个都贡献了不可冗余的预测信息。该条件类似于稀疏回归中为保证模型选择一致性而通常施加的 beta-min 条件或信号分离假设（Zhao 和 Yu，2006 年；Huang等人，2010）。如果没有这样的条件，特征选择一致性将无法识别：如果f_0可以由少于K^*个特征同样良好地表示，那么任何方法都无法始终如一地区分K^*与数量更少的有效特征。

定理 2。 假设推论 1的条件以及假设 5 和6均成立。则存在一个常数 $C = C(\eta,\ B,\ M,\ R,\ C_\ell,\ \sigma_e) > 0$，使得若 $\tau_1 \geq CC_g^2$，$\widehat{}$则 $\mathbb{P}(K = K^*) \to 1$ 当 $n \to \infty$时成立。

由于 $\boldsymbol{\beta}_{k\bullet mj_m}$ 和 θ_j 中非零分量的选择取决于 K的选择，因此要实现选择一致性，特征级惩罚必须强于张量分量和外层神经网络参数所受的惩罚。也就是说，$\tau_1 \geq CC_g^2$ 或等价地，τ_1/τ_2 和 τ_1/τ_3 应足够大。此外，如果它们保持为常数阶（这是定理 1的要求），定理 2 便证明，对于 K^*，TensorNN 同时具备预言机速率预测性能和选择一致性。定理 2 还揭示了 TensorNN 的一项重要实际优势。在标准稀疏参数模型中，为提高选择一致性而增大惩罚强度，往往会带来额外的估计偏差，从而导致预测准确性与变量选择之间的权衡（Zhao 和 Yu，2006 年；van deGeer，2008 年）。相比之下，TensorNN 得益于其非唯一参数化：特征指示变量 $\boldsymbol{\gamma}_{k\bullet}$ 的尺度可以由张量分量和外层神经网络参数进行补偿。因此，只要相对惩罚权重保持为常数阶且满足所需的有界性条件，就可以增大 τ_1 以促进特征选择，而不损害预言机预测速率。因此，定理 2 为 TensorNN 的特征选择版本提供了理论依据，表明它能够同时恢复真实的活跃特征数 K^* 并达到最优收敛速率。

4 模拟研究

本节在若干张量回归和分类任务上评估所提出的 TensorNN。我们考察该方法的两个变体。TensorNN-op 使用

推论 1所建议的、具有理论依据的相对惩罚权重，侧重于预测性能。TensorNN-sc 进一步调整特征选择权重 τ_1，其依据是推论 2 和定理 2，以确保选择一致性。若无需区分这两种变体，我们便统称为 TensorNN。我们将 TensorNN 与三种竞争方法进行比较：卷积神经网络（CNN，Li 等人，2021年）、采用组套索惩罚的非线性稀疏张量加性回归（STAR，Hao等人，2021年），以及采用套索惩罚项的张量线性回归（TLR，Zhou等人，2013年）。

4.1 模拟设置

场景1。我们生成多模张量协变量 $\mathbf{X} \in \mathbb{R}^{P_1 \times \cdots \times P_M}$，其 $M = 2$ 或 3。真实回归函数定义为 $f_0(\mathbf{X}) = \{f_{\text{raw}}(\mathbf{X}) - \mu_0\}/\sigma_0$，其中 $f_{\text{raw}}(\mathbf{X}) = h_0\{h_1(\mathbf{X}^{\mathcal{A}_1}) \ldots, h_{K^*}(\mathbf{X}^{\mathcal{A}_{K^*}})\}$, $h_k(\mathbf{X}^{\mathcal{A}_k}) = \langle \mathbf{B}_k, \mathbf{X} \rangle$，系数张量 $\mathbf{B}_k$的支集为$\mathcal{A}_k$，且 $\|\mathbf{B}_k\|_F = 1$。因此，h_k 从活跃子区域 $\mathcal{A}_k$ 中提取局部信息，并h_0 整合这些局部特征。对于最小二乘回归，我们设定 $\mu_0 = 0$ 和 $\sigma_0 = 1$。对于逻辑回归，其中 $\Pr(Y = 1 \mid \mathbf{X}) = \text{Sigmoid}$ 函数$\{f_0(\mathbf{X})\}$，常数 μ_0 和 σ_0 经过选择，使所得条件概率覆盖 $[0, 1]$的广泛范围。

六种模拟设置包括三种活跃区域几何形状：如图1所示，在 50×50 张量上设置复合模式，在 32×32 张量上设置马形区域，以及在$20 \times 20 \times 10$ 张量上设置椭球区域。对于每种几何形状，我们分别考虑加性设置和交互设置，其中 h_0 分别以加性方式或通过交互作用组合提取的局部特征。模拟规格汇总于表1。

场景2 与场景1类似，不同之处在于特征提取器是非线性的，即 $h_k(\mathbf{X}^{\mathcal{A}_k}) = \langle \mathbf{B}_k, \sin(\pi \mathbf{X}/2) \rangle$，其中正弦变换是逐元素应用的。与（2）中的精确 TensorNN 表示相比，该场景允许我们考察更广泛的模型类。

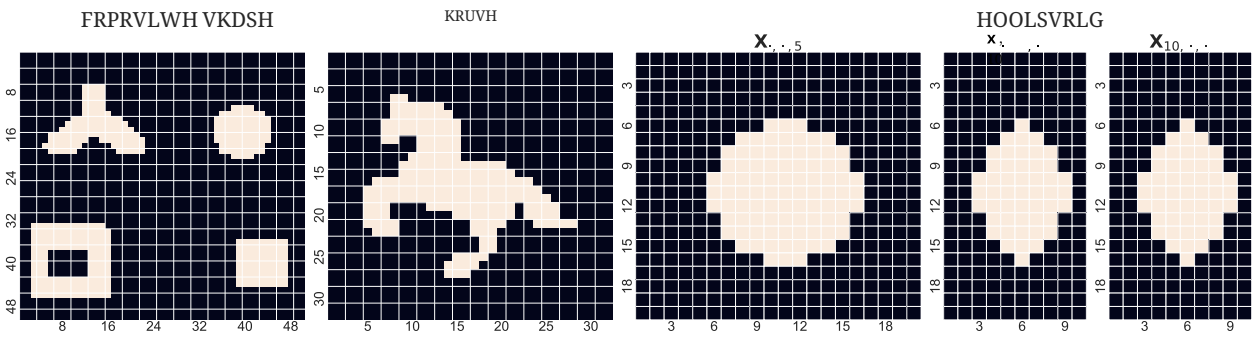

图1：三组模拟设置的活跃区域。前两张热图分别对应复合形和马形活跃区域。最右侧的三张热图展示了椭球体活跃区域的三个切片，分别对应$\mathbf{X}_{\bullet,\bullet,5}$、$\mathbf{X}_{\bullet,10,\bullet}$和$\mathbf{X}_{10,\bullet,\bullet}$。亮色单元格表示活跃的张量元素。

表1：模拟设置的规格。根据活跃区域的形状，这六种设置分为三组。

设置 ID	1	2	3	4	5	6
活动区域模式	复合形状		马		椭球体	
回归类型	最小二乘法				逻辑回归	
模型类型	加性	交互作用	加性	交互作用	加性	交互作用
真实函数 f_0	$f_0(\mathbf{X}) = \{h_0(h_1(\mathbf{X}^{\mathcal{A}_1}) \cdots, h_{K^*}(\mathbf{X}^{\mathcal{A}_{K^*}})) - \mu_0\}/\sigma_0$					
h_k	$h_k(\mathbf{X}^{\mathcal{A}_k}) = \langle \mathbf{B}_k, \mathbf{X} \rangle$					
μ_0	0				校准	
σ_0	1				校准	
K^*	5	5	6	6	4	4

对于每种设定，$\mathbf{X}$ 的各个元素均独立地从均匀分布 Unif$(-1,1)$ 中生成。在最小二乘法设定中，$Y = f_0(\mathbf{X}) + \epsilon$，其中 $\epsilon \sim N(0, \sigma_\epsilon^2)$，并选择 σ_ϵ^2使得 Var$\{f_0(\mathbf{X})\}$/Var$(\epsilon) = 50$。在逻辑回归设定中，$Y \sim$ 伯努利分布 [Sigmoid 函数 $\{f_0(\mathbf{X})\}$]。我们进行了50次相互独立的重复试验，每次包含2,000个训练观测、400个验证观测和1,600个测试观测。验证集用于选择调优参数 λ 和 τ_1。初始特征维度 K、CP秩 R、网络架构和相对权重 $\boldsymbol{\tau}$ 均按照理论指导或基于试验的程序进行选择。

我们使用最小二乘回归的相对预测误差（RPE）和分类的误分类率（MCR）来评估预测：RPE $= \frac{\sum_{i \in S}\{y_i - \widehat{f}(\mathbf{x}_{i})\}^2}{\sum_{i \in S} y_i^2}$ 和 MCR $=$ ———————

$\frac{1}{|S|}\sum_{i\in S}|y_i-\widehat{y}_i|$，其中 S 表示测试集，$\widehat{y}_i = I\{\text{Sigmoid 函数}(\widehat{f}(\mathbf{X}_i)) \geq 0.5\}$. 我们还报告模型估计误差（MEE）。对于最小二乘回归，MEE 是相对$L_2(\mathbb{P}_{\mathbf{X}})$误差 $\|\widehat{f}-f_0\|^2_{L_2(\mathbb{P}_{\mathbf{X}})}/\|f_0\|^2_{L_2(\mathbb{P}_{\mathbf{X}})}$ · 对于分类，MEE 定义为 $\mathbb{E}\big[\{\widehat{\mathbb{E}}(Y\mid\mathbf{X}) - \mathbb{E}(Y\mid\mathbf{X})\}^2\big]$，其中 $\widehat{\mathbb{E}}(Y\mid\mathbf{X}) =$ Sigmoid 函数$\{\widehat{f}(\mathbf{X})\}$。

在所有数值实验和真实数据分析中，优化和调参流程均保持可比。对于 TensorNN，我们使用 RMSProp算法（6），通过 RMSProp算法（Hinton 等，2012）求解其中的惩罚经验风险问题，同时更新张量特征提取器 $\{\mathbf{B}_k\}_{k=1}^K$ 和外层神经网络 $g_{\boldsymbol{\theta}_{net}}$。在适用的情况下，各竞争方法均采用相同的优化方案进行训练。批量大小设为 $0.2N_{training}$，其中 $N_{training}$ 为训练集的样本量。使用默认的 RMSProp超参数，其中 $\alpha = 0.99$，$\epsilon = 10^{-8}$，权重衰减和动量均为零，并将 centered 设置为 False。根据收敛情况，从 10^{-5} 到 10^{-3} 中选择学习率。对于每个候选调参设置，在 10 个等间隔检查点对训练过程进行监控，并选择验证误差最小的检查点。所有计算均使用 Python 实现。

4.2 模拟结果

场景1。 表 2 报告了50次重复试验中平均RPE/MCR和MEE的标准差，以及$\widehat{\ }$所选特征数量 K 的中位数绝对偏差。TensorNN在除设置1之外的所有设置中均取得了最佳表现；在设置1中，TLR略胜一筹。这可能是因为真实函数 $h_0(\mathbf{x}) = \sum_{k=1}^5 x_k + \ln(x_k^2+2)/k$ 主要由其线性成分主导，因此更接近TLR结构。将交互设置（设置2、4和6）与其加性对应设置（设置1、3和5）进行比较后，我们观察到，随着真实函数变得更加复杂，TensorNN的优势愈发明显。此外，TensorNN-sc通常能够很好地恢复 K^*，而TensorNN-op往往会

表2：场景1的平均RPE/MCR和MEE，以及选定特征数量的中位数$\frown K$。括号内的数字分别表示50次重复试验中RPE/MCR和MEE的标准差$\widehat{}$以及 K的中位数绝对偏差。粗体表示各设置下最佳的预测或估计性能。

设置 ID	1	2	3	4	5	6
活跃区域类型	复合形状		马		椭球体	
回归类型	最小二乘法				逻辑回归	
模型类型	加性	交互作用	加性	交互作用	加性	交互作用
	RPE				MCR	
TensorNN-op	0.0483(0.0005)	0.0690(0.0017)	**0.0420**(0.0003)	**0.0536**(0.0006)	0.1705(0.0013)	**0.2069**(0.0017)
TensorNN-sc	0.0476(0.0004)	**0.0647**(0.0012)	0。0421(0。0003)	**0.0537**(0.0005)	**0.1680**(0.0015)	**0.2069**(0.0018)
CNN	0.1156(0.0019)	0.1348(0.0021)	0.1122(0.0007)	0.1349(0.0009)	0.1843(0.0019)	0.2276(0.0022)
STAR	0.0549(0.0004)	0.2054(0.0017)	0.0743(0.0005)	0.1291(0.0009)	0.1748(0.0022)	0.2223(0.0019)
TLR	**0.0457**(0.0003)	0.1524(0.0010)	0.0648(0.0005)	0.1061(0.0007)	0.1781(0.0015)	0.2257(0.0019)
	MEE					
TensorNN-op	0.0128(0.0001)	0.0476(0.0015)	**0.0227**(0.0002)	**0.0157**(0.0002)	0.0631(0.0016)	**0.0948**(0.0022)
TensorNN-sc	0.0125(0.0001)	**0.0435**(0.0010)	**0.0227**(0.0002)	0。0158(0。0002)	**0.0591**(0.0019)	**0.0944**(0.0020)
CNN	0.0425(0.0007)	0.1096(0.0017)	0.0937(0.0004)	0.0540(0.0004)	0.0981(0.0024)	0.1086(0.0017)
STAR	0.0157(0.0001)	0.1773(0.0009)	0.0554(0.0003)	0.0515(0.0002)	**0.0588**(0.0022)	0.1131(0.0008)
TLR	**0.0116**(0.0001)	0.1260(0.0005)	0.0458(0.0003)	0.0410(0.0002)	**0.0585**(0.0007)	0.1145(0.0004)
	中位数 $\widehat{K}$(中位数绝对偏差)					
TensorNN-op	5.0(1.0)	5.0(1.0)	7.0(2.0)	6.0(1.0)	12.5(8.5)	8.0(4.0)
TensorNN-sc	4.0(1.0)	4.0(1.0)	7.0(2.0)	6.0(1.0)	6.0(4.0)	5.0(3.5)

略微高估它，尤其是在设置5和设置6中，此时与不同局部特征相关联的活跃区域开始发生重叠。

与 CNN 的比较进一步说明了局部卷积特征学习与自适应张量特征提取之间的区别。如表2所示，TensorNN 在各种模拟设置下均取得了最佳的总体性能，而 CNN 的结果通常最差或接近最差。

这种性能差距可能归因于：这些设置需要汇集来自空间范围较广或分散的活跃区域的信息，而本文考察的卷积神经网络主要依赖局部感受野。因此，当样本量有限时，这类卷积神经网络在组合空间分离的张量信号方面可能效果较差。相比之下，TensorNN 通过

结构化惩罚项直接学习稀疏张量特征提取器，使其能够适应分散的加性结构、空间范围较广的活跃区域以及非线性交互模式。

TensorNN-op 和 TensorNN-sc 在表2中的 RPE/MCR 和 MEE 值相近，这表明 TensorNN-op 的理论指导的轻量级调优方案已经实现了强大的泛化性能。主要差异在于特征选择。通过进一步调优 τ_1，TensorNN-sc 施加了更强的特征级稀疏性，并选择了更简约的表示，其 $\widehat{\ }$更接近预言特征维数 K 。通过剪除冗余的特征提取器，TensorNN-sc 小幅提升了 TensorNN-op 的预测和估计准确度。实证结果显示，TensorNN-sc 选择的 K^*值明显高于 TensorNN-op，这与定理2一致：适当增大特征级惩罚项有助于提高张量特征的选择一致性。

最后，我们对估计得到的活跃区域进行可视化，以评估 Ten-$\widehat{\ }$sorNN 的可解释性。我们将估计得到的特征提取器 $\mathbf{B}_k$按外层神经网络参数加权，然后将这些加权张量的逐元素平方求和，得到可解释张量 $\widehat{\ }\mathbf{V}^2$。图 2 展示了交互设置下估计得到的活跃区域，每个热图均选自 RPE 最小的重复实验。TensorNN-sc 准确恢复了设置2中的复合活跃模式。在更复杂的马匹设置（设置4）中，突出显示的区域与图 1所示的真实活跃区域大体一致。对于椭球设置（设置6），TensorNN-sc 捕捉到了预期模式：中心区域强烈激活，并逐渐向边界减弱。这些可视化结果表明，TensorNN-sc 能够恢复有意义的活跃张量区域，同时保持强大的预测性能。设置1、3和5的热图（因篇幅所限未展示）也得出了类似结论。

我们还针对网络宽度开展了稳健性实验；结果（为简洁起见略去）表明，网络宽度对模型估计的影响仅处于边际水平，这与定理

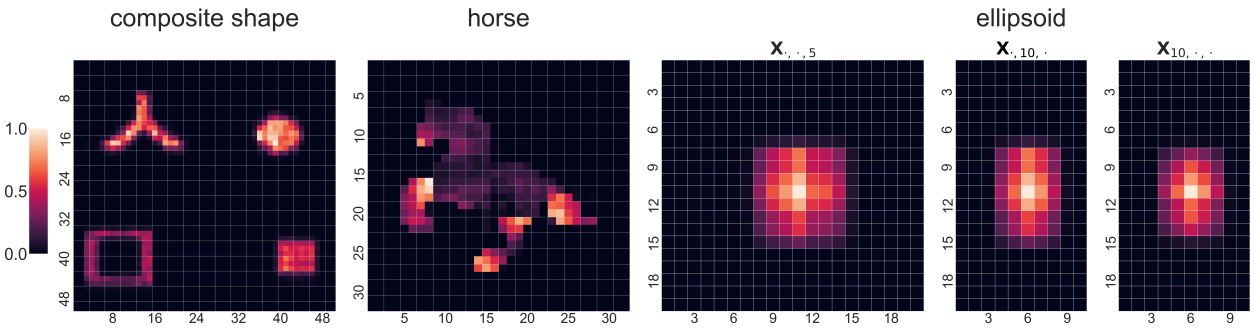

图2：TensorNN-sc在交互设置（设置2、4和6）下得到的热图。第1列和第2列分别对应复合模式（设置2）和马形模式（设置4）。第3–5列展示了设置6下与$\mathbf{X}_{\bullet,\bullet,5}$、$\mathbf{X}_{\bullet,10,\bullet}$和$\mathbf{X}_{10,\bullet,\bullet}$对应的估计活跃区域的三个切片。

所提示的对数依赖关系1一致。

场景2。 表 3 报告了场景2的结果，其中局部特征提取器关于张量协变量呈非线性。总体结论与场景1的结论一致：TensorNN在六种设置下均保持了出色的预测和估计性能。这些结果进一步展现了TensorNN的通用性和自适应能力，表明所提出的框架能够适用于广泛的张量结构和底层回归函数，而不局限于式（2）中精确的模型表示。

5 真实数据实验

我们考虑来自两类张量数据的四个真实数据集：荧光激发-发射矩阵（EEM）数据和人脸图像数据。对于荧光实验，我们使用糖荧光数据（Bro，1999）以及合成荧光数据（Riu和Bro，2003）。糖数据包含268个样本，每个样本由一个 571×7 EEM表示，响应变量为灰分含量。合成荧光数据包含对苯二酚测量值，经过预处理后，包含26个带有 69×24 EEM协变量的样本。

对于图像实验，我们使用 Labeled Faces in the Wild (LFW) 数据集（Huang

表3：场景2的平均RPE/MCR和MEE。括号中的数字为50次重复试验的标准差。粗体表示各设置下最佳的预测或估计性能。

设置 ID	1	2	3	4	5	6
活跃区域类型	复合形状		马		椭球体	
回归类型	最小二乘法				逻辑回归	
模型类型	加性	交互作用	加性	交互作用	加性	交互作用
	RPE				MCR	
TensorNN-op	0.0736(0.0008)	**0.1153**(0.0024)	**0.0658**(0.0005)	**0.0874**(0.0009)	**0.1785**(0.0015)	**0.2229**(0.0020)
CNN	0.1335(0.0016)	0.1590(0.0021)	0.1296(0.0008)	0.1573(0.0010)	0.1890(0.0017)	0.2681(0.0087)
STAR	0.0663(0.0005)	0.2410(0.0019)	0.0841(0.0006)	0.1571(0.0010)	0.1807(0.0015)	0.2444(0.0021)
TLR	**0.0653**(0.0005)	0.1866(0.0013)	0.0855(0.0006)	0.1363(0.0012)	0.1883(0.0014)	0.2495(0.0020)
	MEE					
TensorNN-op	0.0279(0.0003)	**0.0914**(0.0021)	**0.0469**(0.0004)	**0.0366**(0.0004)	0.0694(0.0019)	**0.1063**(0.0024)
CNN	0.0586(0.0007)	0.1310(0.0016)	0。1113(0。0005)	0.0747(0.0005)	0.1070(0.0023)	0.1356(0.0060)
STAR	0.0241(0.0001)	0.2090(0.0010)	0.0653(0.0004)	0.0747(0.0004)	**0.0633**(0.0009)	0.1281(0.0006)
TLR	0。0234(0。0001)	0.1576(0.0008)	0.0668(0.0004)	0.0636(0.0004)	0.0684(0.0012)	0.1286(0.0004)

等，2008 年）及其面部属性标注（Kumar 等，2009）。经过预处理后，每张图像表示为一个 $45 \times 45 \times 3$ 张量。在回归任务中，我们使用微笑属性作为响应变量，在分类任务中则使用闭嘴属性。对于分类任务，根据闭嘴属性的符号将其二分化，$y_i = 1_{\{\text{attr}_i > 0\}}$。所有张量协变量均归一化至 $[0, 1]$。

由于荧光数据集的样本量有限，实验重复进行 100 次。对于面部数据集，实验重复进行 50 次。每次重复中，数据都按 5 : 1 : 1 的比例随机划分为训练集、验证集和测试集。

表4 报告了所有方法的预测性能以及 TensorNN-sc 选定的特征数量。在四个真实数据实验中，TensorNN 达到了最佳或接近最佳的性能。对于糖荧光数据，TensorNN 取得了最低的 RPE，相较于 STAR 和 TLR 略有提升，同时显著优于 CNN。对于合成荧光数据，样本量仅为 26，TensorNN 再次取得最佳性能，其次是 CNN，而 TLR 的表现则差得多。由于 STAR 的 RPE 达到 > 1，甚至不如简单的零预测器，因此未报告 STAR。这些

表4：真实数据实验中的数据规模、张量维度、预测性能和选定特征数量。对于回归任务，预测性能以平均 RPE 衡量；对于分类任务，则以平均MCR 衡量。最后一列报告了 TensorNN-sc 的$\widehat{}$选定特征数量的中位数 K 。括号中的数字表示 RPE/MCR 在$\widehat{}$重复随机划分 K 下的标准偏差和中位数绝对偏差。每个实验中最佳的预测性能以粗体显示。

真实数据	数据规模与维度		TensorNN-op	TensorNN-sc	CNN	STAR	TLR		TensorNN-sc
Sugar 荧光	268 & 571×7	RPE	0.1963 (0.0064)	**0.1899** (0.0058)	0.2849 (0.0098)	0.1996 (0.0051)	0.2051 (0.0057)	$\widehat{K}$	11.5 (2.5)
合成 荧光	26 & 69×24		0.1565 (0.0095)	**0.1519** (0.0101)	0.1850 (0。0111)	– –	0.3626 (0.0074)		10.0 (0.0)
微笑	5721 & $45 \times 45 \times 3$		0.1542 (0.0013)	0.1538 (0.0012)	**0.1446** (0.0015)	0.1869 (0.0014)	0.1892 (0.0016)		46.0 (12.0)
闭嘴	5721 & $45 \times 45 \times 3$	MCR	**0.1549** (0.0016)	0.1561 (0.0018)	0.1657 (0.0018)	0.1974 (0.0021)	0.1641 (0.0021)		32.0 (8.0)

结果表明，即使训练数据有限或张量维度高度不平衡，所提出的方法也能自适应地近似回归函数。这种稳健性源于 TensorNN 能够有效学习低维表示，并针对所选维度构建紧凑的外层神经网络，从而减少有效参数数量，进而降低该方法对大样本量的依赖。

对于人脸图像数据，卷积神经网络在微笑预测任务上的 RPE 最低，而 TensorNN 仍具有竞争力，并明显优于 STAR 和 TLR。这是合理的，因为微笑通过空间分布的人脸线索体现出来，而卷积滤波器擅长捕捉空间模式。然而，对于闭嘴分类任务，TensorNN 的误分类率最低，并优于卷积神经网络、STAR 和 TLR。该任务与局部结构化口部区域的关联更为紧密，而基于稀疏 CP 分解的张量特征提取器能够有效捕捉这类区域。这些结果表明，TensorNN 在使用图像张量协变量进行预测时具有竞争力；当预测信号可由低维结构化表示刻画时，它尤其具有优势。

表 4 中报告的选定特征数量也具有参考意义。对于两项面部图像任务，TensorNN-sc 在微笑预测中选取的特征多于闭嘴分类。这个模式与两种属性的特点相符：微笑涉及多个面部区域的协调变化，包括眼睛、眉毛、脸颊和嘴部；而闭嘴属性主要与局部的嘴部$\widehat{}$区域相关。的数值也表明，刻画人类表情可能需要丰富的面部细节。尽管如此，通过在模型选择过程中鼓励模型稀疏性，同时保持具有竞争力的预测准确率，我们可以看到，起主导作用的特征仍只占少数：微笑属性的特征数量中位数从 46 降至 9。5，闭嘴属性的特征数量中位数从 32 降至 3。5，而相对误差仅分别增加 3。64% 和 1。60%。

接下来，我们考察 TensorNN 的可解释性。与第 4.2 节一样，我们通过聚合重要性张量 $\widehat{\mathbf{V}}^{2,}$ 来可视化估计得到的活跃区域，该张量由学习得到的张量特征提取器构建。对于人脸图像，将三个 RGB通道上的重要性值相加，得到二维热图。图 3 展示了 TensorNN-sc 得到的热图，其中较亮的元素表示估计重要性较高的张量元素。对于糖荧光数据，TensorNN-sc 突出了若干重要的激发波长，并在发射波长上呈现垂直结构的模式，这表明学习到的特征捕捉的是有组织的 EEM信号，而非孤立的张量元素。对于合成荧光数据，TensorNN-sc 找到了稀疏且局部集中的 EEM区域，主要对应于约 249–265、293–413 和 455–469 nm 的发射波长，以及约 215–225 和 285–315 nm 的激发波长。对于人脸图像数据，突出显示的区域与相应属性相符：微笑任务涉及较广的人脸区域，包括嘴部、脸颊和眼周区域；而闭嘴任务则主要集中在嘴部周围。这些

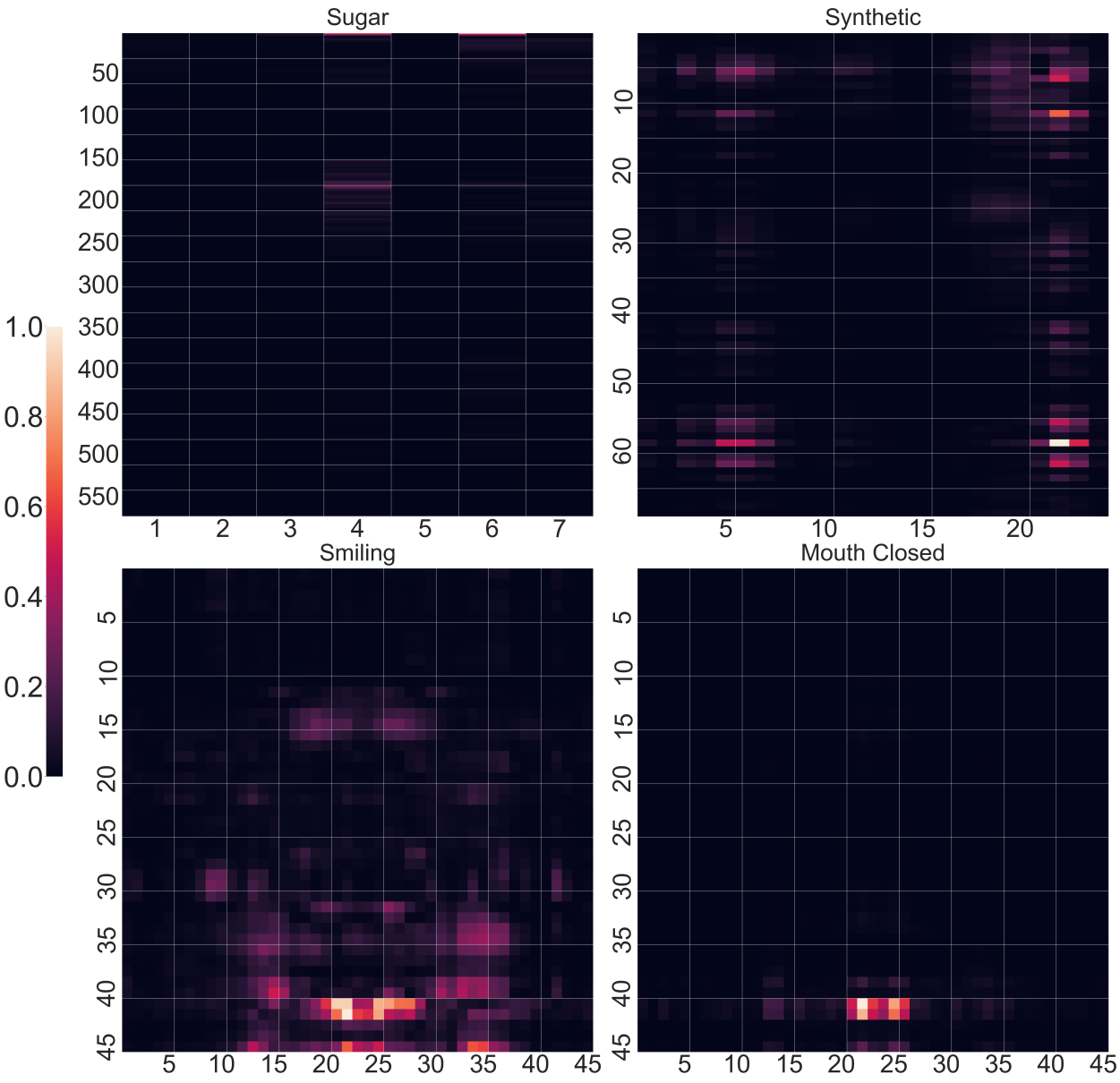

图3：TensorNN-sc 得到的所有真实数据实验的热图。第1–4列分别对应糖荧光、合成荧光、微笑面部和闭口面部实验。热图中方块（i，j）的深浅表示张量协变量 $\mathbf{X}$ 的（i，j）元素的相对重要性。

结果表明，TensorNN 在保持强大预测性能的同时，具备内在的张量级可解释性。

6 讨论

本文提出了一种张量神经网络框架，用于从张量值协变量中学习具有预测性且可解释的表示。其核心思想是识别一组最小的有效张量特征，同时保留张量协变量的多维结构。该方法将张量分解、表示学习和特征及网络组件中的结构化稀疏性相结合，自适应地选择表示维度，并为经典张量回归模型和卷积神经网络提供了一种灵活、具有预测性且可解释的替代方案。

所提出的三部分加权惩罚项是实现这种自适应性的关键。特征级惩罚项用于去除冗余提取器，组件级惩罚项在每个提取器内部引入局部稀疏性，而网络惩罚项则压缩外层神经网络中不必要的权重。因此，可以从一个足够丰富的工作模型开始，采用相对较大的特征维度和网络规模，同时由惩罚项自适应地控制有效表示维度和模型复杂度。根据理论误差界推导出的相对惩罚权重形成了 TensorNN-op 实现；该实现只需对整体正则化强度进行轻量调优，并粗略选择网络深度。当特征恢复是首要目标时，依据选择一致性理论，TensorNN-sc 会进一步调优特征级权重 τ_1。

从理论上看，定理 1中的 Oracle 型预测界表明，主导误差速率由最小特征维度 K^* 决定，而非由环境张量维度 $\mathcal{D}$决定。在所述正则性条件下，TensorNN 达到 K^*维的

非参数速率 $n^{-2\xi/(2\xi+K^*)}$（忽略对数因子），尽管估计是在扩大的工作模型上进行的。定理2 表明，在附加的特征信号强度条件下，提高特征级惩罚项可以实现对K^* 的一致恢复，同时不会牺牲预言机预测速率。

表 2、3和4 中的结果总体上支持这些理论结论。在模拟和真实数据应用中，TensorNN 都比竞争方法取得了更强的预测性能，同时生成了具有可解释性的低维张量表示。当底层模型较为复杂时，其优势尤为明显。总体而言，实证证据表明，TensorNN 能够适应不同类型的张量协变量和信号结构。

不同于卷积神经网络（CNNs）通常依赖事后工具进行解释，TensorNN 通过其学习得到的张量特征提取器，在模型层面提供活动区域摘要。图 2和 3 中的模拟数据和真实数据热图表明，TensorNN 能够识别出与响应变量相关的有意义的张量区域。这种内置的可解释性对于非图像张量应用尤其有价值，因为通用的卷积架构和事后可视化工具可能无法提供稳定或具有科学意义的解释。

仍有两项局限。首先，当前框架使用通用的张量特征提取器；纳入更多特定领域的结构可能会进一步提升性能和可解释性。例如，对于图像张量，可以施加空间平滑性；而时间依赖性或上下文依赖性则可通过将张量特征提取与循环架构或基于 Transformer 的架构相结合来纳入。其次，活动区域热图应被视为基于模型的摘要，而非因果解释。这些局限为将 TensorNN 扩展至更广泛的张量学习问题提供了很有前景的方向

参考文献

Alquier, P., V. Cottet, and G. Lecu´e（2019年）。具有 Lipschitz 损失函数的正则化方法的估计界与尖锐预言机不等式。《统计年鉴》47(4)，2117–2144。Bartlett, P., O. Bousquet, and S. Mendelson（2005）。局部拉德马赫复杂度。统计年鉴33(4)，1497–1537。Bauer, B. and M. Kohler（2019年）。论深度学习如何克服非参数回归中的维数灾难。《统计年鉴》47(4)，2261–2285。Bellec, P. C., G. Lecu´e, and A. B. Tsybakov（2018年）。Slope 与 Lasso 的结合：改进的预言机界与最优性。《统计年鉴》46(6B)，3603–3642。Bhargava, P., T. Phan, J. Zhou, and J. Lee（2015）。谁、什么、何时、何地：利用张量分解对稀疏用户生成数据进行多维协同推荐。载于《第24届万维网国际会议论文集》，第130–140页。Bickel, P. J., Y. Ritov, and A. B. Tsybakov（2009）。Lasso 与 Dantzig选择器的同步分析。《统计年鉴》37(4)，1705–1732。Bro, R.（1999）。利用荧光光谱法和多维分析探索糖类生产。化学计量学与智能实验室系统46(2)，133–147。Chen, M., H. Jiang, W. Liao, and T. Zhao（2019年）。低维流形上函数的深度ReLU网络高效逼近。载于《神经信息处理系统进展》第32卷，第8174–8184页。Curran Associates, Inc. Chen, S., K. He, S. He, Y. Ni, and R. K. Wong（2023年）。采用函数型融合弹性网络先验的贝叶斯非线性张量回归。Technometrics 65(4)，524–536。Chien, J.-T. and Y.-T. Bao（2017年）。张量分解神经网络。IEEE神经网络与学习系统汇刊29(5)，1998–2011。Clarke, F. H.（1990）。优化与非光滑分析。SIAM。Cox, D. D. and R. L. Savoy（2003）。功能性磁共振成像（fMRI）“脑解读”：检测并分类人类视觉皮层中的分布式 fMRI 活动模式。《神经影像》19(2)，261–270。Craddock, R. C., P. E. Holtzheimer III, X. P. Hu, and H. S. Mayberg（2009）。基于静息态功能连接的疾病状态预测。Magnetic Resonance in Medicine: An Official Journal of the International Society for Magnetic Resonance in Medicine 62(6)，1619–1628。da Costa, M. N., R. Attux, A. Cichocki, and J. M. Romano（2025年）。用于学习多维数据预测建模的张量列车网络。《神经计算》637，130037。Farrell, M. H., T. Liang, and S. Misra（2021年）。用于估计与推断的深度神经网络。Econometrica 89(1)，181–213。

Guhaniyogi, R., S. Qamar, and D. B. Dunson (2017). 贝叶斯张量回归。《机器学习研究期刊》18(79), 1–31。Guo, W., I. Kotsia, and I. Patras (2011). 用于回归的张量学习。IEEE图像处理汇刊 21(2), 816–827。Hao, B., B. Wang, P. Wang, J. Zhang, J. Yang, and W. W. Sun (2021). 稀疏张量加性回归。《机器学习研究期刊》22(64), 1–43。Haris, A., N. Simon, and A. Shojaie (2022). 广义稀疏加性模型。《机器学习研究杂志》23(1), 3035–3090。Harshman, R. (1970年). PARAFAC过程的基础：解释性多模态因子分析的模型与条件。加州大学洛杉矶分校语音学工作论文 16, 1–84。Hinton, G., N. Srivastava, and K. Swersky (2012). 机器学习神经网络讲座6a：小批量梯度下降概述。引用于14(8), 2。Huang, G. B., M. Mattar, T. Berg, and E. Learned-Miller (2008 年). Labeled Faces in the Wild：用于研究非受限环境下人脸识别的数据库。载于“真实生活”图像中的人脸研讨会：检测、对齐与识别。Huang, J., J. L. Horowitz, and F. Wei (2010). 非参数加性模型中的变量选择。《统计年鉴》38(4), 2282。Imaizumi, M. and K. Hayashi (2016年). 双重分解非参数张量回归。载于国际机器学习大会，第727–736页。机器学习研究论文集。Jiao, Y., G. Shen, Y. Lin, and J. Huang (2023年). 近似流形上的深度非参数回归：带有多项式前因子的非渐近误差界。《统计年鉴》51(2), 691–716。Kanagawa, H., T. Suzuki, H. Kobayashi, N. Shimizu, and Y. Tagami (2016年). 高斯过程非参数张量估计器及其极小极大最优性。载于国际机器学习大会，第1632–1641页。机器学习研究论文集。Kawaguchi, K., Y. Bengio, and L. Kaelbling (2022). 深度学习中的泛化。载于 P. Grohs 和 G. Kutyniok（编），《深度学习的数学方面》，第112–148页。剑桥：剑桥大学出版社。Kolbeinsson, A., J. Kossaifi, Y. Panagakis, A. Bulat, A. Anandkumar, I. Tzoulaki, and P. M. Matthews (2021年). 面向鲁棒学习的张量随机失活。IEEE信号处理专题精选期刊 15(3), 630–640。Kolda, T. G. and B. W. Bader (2009 年). 张量分解及其应用。SIAM评论 51(3), 455–500。Koltchinskii, V. (2006 年). 风险最小化中的局部拉德马赫复杂度与预言机不等式。《统计年鉴》34(6), 2593–2656。Kossaifi, J., Z. C. Lipton, A. Kolbeinsson, A. Khanna, T. Furlanello, and A. Anandkumar (2020年). 张量回归网络。《机器学习研究期刊》21(123), 1–21。

Kumar, N., A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2009年). 用于人脸验证的属性与明喻分类器。载于2009年IEEE第12届国际计算机视觉会议，第365–372页。IEEE。Lazar, N. (2008 年). 功能性磁共振成像数据的统计分析。Springer Science & Business Media。Lecu´e, G. and S. Mendelson (2018年). 正则化与小球方法 I：稀疏恢复。《统计年鉴》46(2), 611–641。Li, X., D. Xu, H. Zhou, and L. Li (2018 年). Tucker 张量回归与神经影像分析。生物科学统计学 10(3), 520–545。Li, Z., F. Liu, W. Yang, S. Peng, and J. Zhou (2021年). 卷积神经网络综述：分析、应用与展望。IEEE神经网络与学习系统汇刊 33(12), 6999–7019。Ma, X., P. Zhang, S. Zhang, N. Duan, Y. Hou, M. Zhou, and D. Song (2019年). 用于语言建模的张量化 Transformer。载于《神经信息处理系统进展》第32卷，第2232–2242页。Marx, B. D., P. H. Eilers, and B. Li (2011年). 多维单指标信号回归。化学计量学与智能实验室系统 109(2), 120–130。Meng, X. and Y. Li (2026). 广义非参数模型中深度神经网络估计量的推断。《美国统计学会杂志》，1–13。Nakada, R. and M. Imaizumi (2020年). 深度神经网络的自适应逼近与泛化及其内在维数。《机器学习研究期刊》21(174), 1–38。Ohn, I. and Y. Kim (2022). 深度神经网络的非凸稀疏正则化及其最优性。神经计算 34(2), 476–517。Panagakis, Y., J. Kossaifi, G. G. Chrysos, J. Oldfield, M. A. Nicolaou, A. Anandkumar, and S. Zafeiriou (2021年). 计算机视觉与深度学习中的张量方法。IEEE论文集 109(5), 863–890。Papadogeorgou, G., Z. Zhang, and D. B. Dunson (2021年). 软张量回归。《机器学习研究期刊》22(219), 1–53。Penny, W. D., K. J. Friston, J. T. Ashburner, S. J. Kiebel, and T. E. Nichols (2011年). 统计参数映射：功能性脑影像分析。爱思唯尔。Raskutti, G., M. Yuan, and H. Chen (2019年). 高维多响应张量回归的凸正则化。《统计年鉴》47(3), 1554–1584。Reiss, P. T. and R. T. Ogden (2010). 以图像为预测变量的函数型广义线性模型。生物统计学 66(1), 61–69。Riu, J. and R. Bro (2003). 用于异常值检测和PARAFAC模型标准误差估计的刀切法。化学计量学与智能实验室系统 65(1), 35–49。

Schmidt-Hieber, J. (2020). 使用 ReLU激活函数的深度神经网络进行非参数回归。《统计年鉴》48(4), 1875–1897. Signoretto, M., L. D. Lathauwer, and J. A. K. Suykens (2013). 在再生核希尔伯特空间中使用多线性谱惩罚学习张量。CoRR abs/1310.4977. Stone, C. J. (1982). 非参数回归的最优全局收敛速率。《统计年鉴》10(4), 1040–1053. Suzuki, T., H. Kanagawa, H. Kobayashi, N. Shimizu, and Y. Tagami (2016). 核非参数张量学习的极小极大最优交替最小化。载于《神经信息处理系统进展》，第29卷，第3783–3791页。Tan, Z., L. Zhou, and H. Lin (2024). 具有最优输入维度的生成对抗学习及其自适应生成器架构。arXiv预印本 arXiv:2405.03723. Tucker, L. R. (1966). 关于三模因子分析的一些数学注记。《心理测量学》31(3), 279–311. van de Geer, S. A. (2008). 高维广义线性模型与 Lasso。《统计年鉴》36(2), 614 – 645. Wang, X., B. Nan, J. Zhu, and R. Koeppe (2014). 基于Haar小波的脑影像数据正则化三维函数回归。《应用统计年鉴》8(2), 1045. Worsley, K. J., J. E. Taylor, F. Tomaiuolo, and J. Lerch (2004). 统一单变量与多变量随机场理论。《神经影像》23, S189–S195. Yang, K. and T. Maiti (2022). 超高维广义加性模型：统一理论与方法。《斯堪的纳维亚统计学杂志》49(3), 917–942. Zhao, P. and B. Yu (2006). 关于 Lasso 的模型选择一致性。《机器学习研究杂志》7, 2541–2563. Zhou, H. and L. Li (2014). 正则化矩阵回归。《皇家统计学会会刊》：B辑（统计方法论）76(2), 463–483. Zhou, H., L. Li, and H. Zhu (2013). 张量回归及其在神经影像数据分析中的应用。《美国统计学会杂志》108(502), 540–552. Zhou, Y., R. K. Wong, and K. He (2024). 广播式非参数张量回归。《皇家统计学会会刊》B辑：统计方法论 86(5), 1197–1220.