arXiv:2603.01975v1 [stat.ML] 2 Mar 2026

# 要点

**分类数据的密度矩阵谱嵌入：算子结构与稳定性**

Raquel Bosch-Romeu、Antonio Falcó、José-Antonio Rodríguez-Gallego

- 用于分类数据监督表示的半正定密度算子
- 内在秩受类别数限制
- 从主导谱子空间获得的低维嵌入
- 通过类别条件核密度估计构建潜在空间分类器
- Davis–Kahan 扰动与子空间稳定性分析及合成数据验证

# 分类数据的密度矩阵谱嵌入：算子结构与稳定性

Raquel Bosch-Romeu[a]、Antonio Falcó[a]、José-Antonio Rodríguez-Gallego[b,*]

[a] 数学、物理与技术科学系，卡德纳尔·埃雷拉-*CEU*大学，*CEU*大学联盟，圣巴托洛梅*55*号，阿尔法拉德尔帕特里亚尔卡（瓦伦西亚），邮编*46115*，西班牙[b]微分方程与数值分析系，梅赛德斯王后大道（无门牌号），塞维利亚，邮编*41012*，西班牙

**摘要**

我们提出了一种针对分类数据（以及离散化混合类型数据）的监督式降维方法，该方法基于由类别条件频率诱导的密度矩阵构造。给定一个在独热调查空间中编码的带标签数据集，我们构建一个频率矩阵，其列汇总各类别中的特征出现次数，并定义一个归一化的 Gram 型算子，使其满足密度矩阵的公理。由此得到的表示具有由类别数决定的内在秩界，可通过主导特征模态实现低维谱嵌入。在降维空间中，我们通过类别条件核密度估计和最大似然决策规则进行分类。我们确立了结构不变性，给出了复杂度估计，并在合成基准测试中验证了该方法，测试涵盖高基数、稀疏性、噪声和类别不平衡。

关键词：密度算子，分类数据，半正定矩阵，谱嵌入，低秩表示，核密度估计

*2000*年数学主题分类：15A18, 15A83, 65F15, 62H30, 62G07

## 1. 引言

高维分类数据自然出现在许多场景中，包括调查、事件日志、交易记录，以及更一般地，定性描述符主导特征空间的问题。一种标准编码将每个分类变量展开为一个独热编码块，并将所有块拼接成一个稀疏二进制向量。尽管这种表示是忠实的，但当变量具有许多模态、存在稀有类别或定量协变量被离散化时，它可能导致非常大的环境维数。所得数据矩阵是稀疏且高度结构化的。从线性代数的观点来看，此类数据集自然地由列联型矩阵及相关的格拉姆算子加以概括，从而使谱结构和低秩近似成为分析的核心。

一种经典且影响深远的研究路径，是通过对适当归一化的列联表进行矩阵分解来处理分类数据。对应分析（CA）及其扩展方法，包括多重对应分析（MCA）和判别对应分析（DCA），通过奇异值分解和加权内积构建几何表示，从而从分类表中提取低维结构[8, 7]。更广泛地说，谱方法和核方法通过半正定算子表示数据，其特征结构编码了相关几何信息；代表性例子包括核主成分分析 [15], 、拉普拉斯特征映射 [2], 和谱聚类 [11]。这些视角表明，稀疏独热类别数据可能会受益于精心设

*通讯作者电子邮箱地址：gallego@us.es (José-Antonio Rodríguez-Gallego)

计的半正定算子，其主导谱方向能够捕捉类别相关结构。

在本文中，我们提出一种面向带标签独热分类数据的监督算子构造。从类别条件频率信息出发，我们构造一个对称半正定的单位迹矩阵，并将其解释为量子概率与量子信息论意义上的密度算子 [12, 10]。这种归一化并非仅具形式意义：在对类别条件轮廓进行逐项平方根（振幅）变换之后，该构造诱导出一种与离散分布之间的Bhattacharyya系数和Hellinger距离密切相关的几何 [3, 1]。一个关键的线性代数推论是内在的低秩界：所得算子的秩至多为类别数，因此其非零谱完全由一个 $k \times k$ 格拉姆矩阵决定。由此得到一种紧凑的谱嵌入，其有效维数由类别数而非环境独热维度控制。

除其基本结构性质（对称性、半正定性、单位迹以及秩控制）之外，算子视角还使矩阵理论中居于核心地位的扰动分析成为可能。经验类别频率的微小偏差会引致算子的受控扰动，并且在谱隙条件下，主不变子空间（因而相关联的嵌入）是稳定的。我们通过Davis–Kahan型 $\sin\Theta$ 界 [6, 19]对此加以形式化，并在分块分类模型下的多项式抽样情形中，以显式的高概率扰动估计加以补充。这些结果为将所提出的嵌入与相关的谱构造及低秩构造进行比较提供了有原则的基础。

作为谱坐标的一种应用，我们将样本映射到主特征空间，并在降维空间中通过类条件密度估计进行类别分配。我们在 Parzen–Rosenblatt 框架 [14, 13], 中采用核密度估计（KDE），并遵循标准的带宽选择原则 [17, 16, 18]。虽然下游分类器可以替换为其他决策规则，但这种选择突出了潜在表示的概率解释，并使计算成本取决于降维后的维数。

### *1.1.* 主要贡献

本文的主要贡献如下:

- **一种适用于独热类别数据的监督式半正定算子**。 我们定义了一个从带标签独热数据集到归一化半正定单位迹矩阵（密度算子）的映射，为监督式分类学习[12, 10]提供结构化的矩阵表示。

- **内在低秩谱结构与格拉姆约化**。 我们证明，该算子的秩受类别数限制，并表明其非零谱可通过 $k \times k$ 格拉姆矩阵恢复，从而实现高效且直观的谱计算。

- **海灵格／巴氏几何解释**。我们将这一构造与类别条件分类分布上的 Bhattacharyya亲和度和 Hellinger型几何联系起来，从而将所提出的算子与经典的分布相似性框架相衔接 [3, 1]。

- **扰动与子空间稳定性分析**。我们利用 Davis–Kahan 类型的扰动界证明谱子空间稳定性，并在分块类别模型的多项式抽样下，推导出明确的高概率算子扰动估计 [6, 19]。

- **说明谱行为的合成实验**。 我们设计受控的合成实验（分离度、稀疏性、无关变量和类别不平衡），以说明所提出谱嵌入的经验行为及其稳定性趋势。

*1.2.* 论文结构

第 2 节介绍分类调查空间和标签编码。第 3 节给出密度算子构造及其基本性质。第 4 节发展谱嵌入与低维代理表示。第 5 节提供理论分析，包括海灵格／巴塔恰里亚解释、算子关系、扰动界以及子空间稳定性估计。第 6 节介绍潜在空间密度估计及由此得到的分类规则。第 7 节总结计算流程并分析其复杂度。第 8 节报告合成实验。论文最后讨论局限性与未来工作方向。

## 2. 分类数据表示

我们使用分块独热编码来形式化分类输入空间。这种编码在实际应用中十分常见，也特别便于进行算子理论处理。每个观测由 $q$ 个分类变量（调查问题、字段或事件）描述，其中变量 $i$ 有 $m_i \geq 2$ 个互斥模态。由此得到的环境维度为

$$d = \sum_{i=1}^{q} m_i,$$

每个样本都是 $\{0,1\}^d$ 中的一个稀疏向量，每个分块恰有一个活跃坐标。这种表示使我们能够将数据集视为有限乘积空间上的基本事件所构成的结构化集合；更重要的是，它可以将类别条件频率信息汇总为矩阵，并利用线性代数中的标准工具分析其谱性质。

我们还对输出标签采用独热编码，因此每个带标签样本都可以表示为 $\mathbb{R}^{d\times k}$ 中的一个秩一张量（或外积）。采用这种张量视角并非仅仅为了计算上的便利：它为列联表与由类别条件计数构造的半正定算子搭建了一座自然的桥梁，而这些算子将成为下一节所阐述的密度算子方法的基础。

**记号 2.1**（调查空间与维度）。我们考虑总体 Ω 以及一项由 $q$ 个类别问题组成的调查。第 $i$ 个问题由分类变量 $X_i$ 编码，该变量具有$m_i \geq 2$个互斥模态 $\mathcal{O}_i = \{O_1^{(i)}, \ldots, O_{m_i}^{(i)}\}$。我们用 $d = \sum_{i=1}^{q} m_i$ 表示其所在的独热维度。

每个 i ndividual $\omega \in \Omega$ 对问题 $i$ 给出一个结果，该结果由独热向量编码

$$X_i(\omega) = \sum_{k=1}^{m_i} \alpha_k^{(i)}(\omega)\, \mathbf{e}_k^{(i)} \in \{0,1\}^{m_i},$$

其中，若 $\omega$ 选中模态 $O_k^{(i)}$，则 $\alpha_k^{(i)}(\omega) = 1$；否则 $\alpha_k^{(i)}(\omega) = 0$。这里 $\mathbf{e}_k^{(i)}$ 表示 $\mathbb{R}^{m_i}$的第 $k$ 个标准基向量。

**定义 2.2**（分类状态空间）。对于每个问题 $i$，允许的结果集合为

$$\mathcal{X}_{m_i} := \{\mathbf{e}_1^{(i)}, \ldots, \mathbf{e}_{m_i}^{(i)}\} \subset \{0,1\}^{m_i}.$$

完整的调查结果空间是笛卡尔积

$$\mathcal{X}_d := \mathcal{X}_{m_1} \times \cdots \times \mathcal{X}_{m_q} \subset \{0,1\}^d.$$

**例 2.3 (**二元情形)。若 $m_i = 2$，则 $X_i(\omega) \in \{(1,\ 0)^T,\ (0,\ 1)^T\} \subset \{0,1\}^2$。

**定义 2.4 (**调查向量（分块拼接）)。 $\omega$ 的调查响应由独热编码块拼接表示，

$$x(\omega) := \big(X_1(\omega), \ldots, X_q(\omega)\big) \in \{0,1\}^d.$$

等价地， $x(\omega)$ 恰好有 $q$ 个值为 1 的条目，每个块中各有一个。

*2.1.* 标签编码和张量表示

我们考虑一个 $k$ 类分类问题。输出变量 $Y$ 取值于

$$\mathcal{Y}_k := \{\mathbf{e}_1, \dots, \mathbf{e}_k\} \subset \{0,1\}^k, \qquad k \geq 2.$$

全文中，我们聚焦于 $d \gg k$ 这一情形。

**定义 2.5** (带标签样本的张量编码)。 给定 $x \in \mathcal{X}_d$ 和一个标签 $\mathbf{e}_y \in \mathcal{Y}_k$，定义

$$x \otimes \mathbf{e}_y := x\, \mathbf{e}_y^\top \in \mathbb{R}^{d \times k}.$$

该矩阵除第 $y$ 列外，其余各列均为零；第 $y$ 列等于 $x$。相应的张量状态空间为

$$\mathcal{X}_d \otimes \mathcal{Y}_k := \{x\, \mathbf{e}_y^\top : x \in \mathcal{X}_d,\ \mathbf{e}_y \in \mathcal{Y}_k\}.$$

**备注 2.6** (数值变量的离散化)。 定量变量可以通过离散化划分为区间，再进行独热编码，从而纳入同一调查空间形式体系。

**定义 2.7** (训练数据集)。 我们考虑一个带标签训练集

$$\mathcal{D} := \{x^{(j)} \otimes \mathbf{e}_{y^{(j)}} : 1 \leq j \leq n\} \subset \mathcal{X}_d \otimes \mathcal{Y}_k.$$

目标是构造一个经验分类映射

$$\ell_{\mathcal{D}} : \mathcal{X}_d \to \mathcal{Y}_k$$

以逼近理想的（未知）分类器 $\ell$。

**3. 根据训练集构造密度矩阵**

**定义 3.1** （类条件频率向量）。 对于每个类别标签 $\mathbf{e}_y \in \mathcal{Y}_k$，定义频率向量

$$\mathbf{f}_y := \sum_{x \otimes \mathbf{e}_y \in \mathcal{D}} x \in \mathbb{R}^d_{\geq 0}.$$

等价地，$(\mathbf{f}_y)_i$ 统计属性 $i$在类别 $y$的样本中出现的次数。

**定义 3.2** （频率矩阵）。 定义频率矩阵

$$F := [\mathbf{f}_1\, \mathbf{f}_2\, \cdots\, \mathbf{f}_k] \in \mathbb{R}^{d \times k}.$$

**定义 3.3** (振幅提升)。 定义逐元素平方根提升

$$X := \sqrt{F} \in \mathbb{R}^{d \times k}, \qquad X_{i,y} := \sqrt{F_{i,y}}.$$

**定义 3.4** (密度矩阵)。 由 $\mathcal{D}$ 诱导的密度矩阵为

$$\rho_{\mathcal{D}} := \frac{XX^\top}{\operatorname{tr}(XX^\top)} \in \mathbb{R}^{d \times d}.$$

**命题 3.5** (密度矩阵公理)。 $\rho_{\mathcal{D}}$ 是对称的、半正定的，并且满足$\operatorname{tr}(\rho_{\mathcal{D}}) = 1$。

证明。 由于 $XX^\top$ 是对称的，因此对称性显然成立。对于任意 $v \in \mathbb{R}^d$, $v^\top XX^\top v = \|X^\top v\|_2^2 \geq 0$，因此 $XX^\top \succeq 0$ 且 $\rho_{\mathcal{D}} \succeq 0$。最后，$\operatorname{tr}(\rho_{\mathcal{D}}) = \operatorname{tr}(XX^\top)/\operatorname{tr}(XX^\top) = 1$。 □

**命题 3.6** (内在秩界)。 $\operatorname{rank}(\rho_{\mathcal{D}}) \leq k$。

证明。 由于 $\rho_{\mathcal{D}}$ 是 $XX^\top$的正标量倍数，因此有 $\operatorname{rank}(\rho_{\mathcal{D}}) = \operatorname{rank}(XX^\top)$ =$\operatorname{rank}(X)$。因为 $X \in \mathbb{R}^{d \times k}$，所以 $\operatorname{rank}(X) \leq k$。 □

**备注 3.7** (对混合态的解释)。 在量子概率中，对称半正定（*PSD*）单位迹矩阵即为密度矩阵。$\rho_{\mathcal{D}}$的谱分解将其表示为秩一投影算子的凸组合，从而以算子理论的方式表征类条件属性结构。

**4. 通过谱坐标构建低维代理表示**

由类别条件频率构造的密度算子 $\rho_{\mathcal{D}}$ 是一个对称半正定且具有单位迹的矩阵，并且根据命题 3.6，其秩至多为类别数。因此，其谱分解提供了一种自然机制，可将环境独热维度 $d$ 缩减为少量自由度，同时保留数据所编码的主不变子空间。本节将介绍相应的谱坐标，并定义由此得到的代理（训练）表示，以用于后续的估计和分类步骤。

令 $\rho_{\mathcal{D}} = \sum_{i=1}^{r_\star} \sigma_i \, \mathbf{U}_i \mathbf{U}_i^\top$为 $\rho_{\mathcal{D}}$的谱分解，其中 $r_\star = 秩(\rho_{\mathcal{D}}) \le k$,$\sigma_i > 0$，且 $\{\mathbf{U}_i\}$ 是 $\mathbb{R}^d$中的正交归一族。每一对 $(\sigma_i,\ \mathbf{U}_i)$ 表示一个非零特征值及其对应的特征向量，而截断阶数 $r \le r_\star$ 选取主导的$r$维不变子空间，用于构建低维代理坐标。

**定义 4.1 定义** (归一化调查向量)。 对于 $x \in \mathcal{X}_d$，定义欧几里得归一化 $n$

$$\widetilde{x} := \frac{x}{\|x\|_2} = \frac{x}{\sqrt{q}} \in \mathbb{R}^d,$$

因为独热调查向量满足 $\|x\|_2^2 = q$。

**定义 4.2** (谱坐标与截断)。 将归一化调查向量 $\widetilde{x}$ 的谱坐标定义为 $\lambda_i(\widetilde{x})\colon = \mathbf{U}_i^\top \widetilde{x}$。对于 $r \le r_\star$，定义潜在坐标映射

$$\pi_r(x) := \big(\lambda_1(\widetilde{x}), \dots, \lambda_r(\widetilde{x})\big) \in \mathbb{R}^r.$$

**命题 4.3** (支撑约束)。 对于任意 $x \in \mathcal{X}_d$，潜在坐标满足 $\|\pi_r(x)\|_2 \le 1$。

证明。 由于 $\{\mathbf{U}_i\}$ 是正交归一的，且 $\|\widetilde{x}\|_2 = 1$,

$$\|\pi_r(x)\|_2^2 = \sum_{i=1}^{r} (\mathbf{U}_i^\top \widetilde{x})^2 \le \sum_{i=1}^{d} (\mathbf{U}_i^\top \widetilde{x})^2 = \|\mathbf{U}^\top \widetilde{x}\|_2^2 = \|\widetilde{x}\|_2^2 = 1.$$

□

**定义 4.4** (代理训练集)。 代理（嵌入）训练数据集 $i$ $s$

$$\widetilde{\mathcal{D}}_r := \{\pi_r(x^{(j)}) \otimes \mathbf{e}_{y^{(j)}} : 1 \le j \le n\} \subset \mathbb{R}^r \otimes \mathcal{Y}_k.$$

**注释 4.5** (极坐标/球坐标参数化)。 由于 $\pi_r(x) \in B_1^r(0)$，可以使用极坐标（针对$r = 2$）或球坐标（针对 $r \ge 3$），获得便于解释的低维参数化。

**5. 密度算子构造的理论分析**

本节阐述所提出的密度算子构造背后的线性代数解释与稳定性理论。我们首先引入一种类别归一化算子，将类条件分类分布的几何结构与类别流行率分离，从而提供一种便捷的总体视角，并能得出更精确的扰动结论。随后我们将证明，由此得到的类别间格拉姆矩阵具有 Hellinger/Bhattacharyya 核解释，将该算子与一种广为人知的分布几何联系起来，并阐明其与谱核方法及对应型分解之间的联系。

从矩阵理论的角度来看，关键问题在于，当底层频率剖面出现有限样本波动所引发的扰动时，经验算子的主不变子空间会如何变化。

因此，我们将 Davis–Kahan 正弦定理 $\Theta$ 子空间界与对算子扰动的显式高概率控制相结合，从而为多项式抽样下谱嵌入的稳定性提供定量保证。最后，我们将基于计数的构造与类归一化构造联系起来，并表明二者的差异仅在于类别质量加权，由此得到明确的范数界以及两种算子产生本质相同嵌入的条件。

**假设 5.1**。 为便于说明，我们总结理论分析中所用的假设。假设 *5.9* 规定了分块类别抽样模型，并引入谱间隙条件 $\lambda_r(\rho_\star) > \lambda_{r+1}(\rho_\star)$ ，以便通过 *Davis–Kahan* 类型的界控制主不变子空间的稳定性。假设 *5.12* 是一项技术性正则条件，用于确保类剖面处处为正；它保证逐元平方根映射在某个正区间上满足利普希茨条件，并使我们能够为扰动范数 $\|\widehat{\rho}_\star - \rho_\star\|_2$ 推导出显式的高概率界。重要的是，这两个假设仅用于获得定量的稳定性估计，并非定义算子构造或嵌入本身所必需。

**记号 5.2**。 为避免歧义，我们用 $\rho_{\mathrm{CN}}$ 表示直接由计数矩阵构造的有限样本类别归一化算子（定义 *5.3*），用 $\rho_\star$ 表示其在分块分类模型下的总体对应形式，并用 $\widehat{\rho}_\star$ 表示在扰动分析中对 $\rho_\star$ 的经验估计量。类似地， $\Psi$ 表示总体振幅矩阵， $\Psi_n$ 表示其在概率界中的经验对应形式。

*5.1.* 归一化总体模型

为了建立一个数学上简洁的比较框架，我们引入频率矩阵的一种类别归一化变体。这种归一化将类条件分布特征的几何结构与类规模分离开来，尤其便于理论分析。

**定义 5.3**（类归一化分布、振幅矩阵与类归一化密度算子）。 令 $F = (\mathbf{f}_1 \,|\mathbf{f}_2\,| \cdots |\, \mathbf{f}_k) \in \mathbb{R}^{d\times k}$ 为类别条件计数矩阵，其中$\mathbf{f}_y \in \mathbb{R}^d_{\geq 0}$汇集了类别 $y$ 的属性计数。假设对所有 $1 \leq y \leq k$ ， $\mathbf{1}^\top \mathbf{f}_y > 0$ ，并定义类归一化分布

$$\widehat{p}_y \in \mathbb{R}^d, \qquad \widehat{p}_y(i) := \frac{F_{i,y}}{\sum_{j=1}^d F_{j,y}} \quad (1 \leq i \leq d,\ 1 \leq y \leq k).$$

定义相应的振幅剖面 $\widehat{\psi}_y := \sqrt{\widehat{p}_y} \in \mathbb{R}^d$ （逐元素平方根）以及振幅矩阵

$$\widehat{\Psi} := [\widehat{\psi}_1 \ \cdots \ \widehat{\psi}_k] \in \mathbb{R}^{d\times k}.$$

最后，定义相应的类归一化密度算子

$$\rho_{\mathrm{CN}} := \frac{\widehat{\Psi}\widehat{\Psi}^\top}{\mathrm{tr}(\widehat{\Psi}\widehat{\Psi}^\top)} \in \mathbb{R}^{d\times d}.$$

^等价地，由于对每个 $y$ 都有 $\|\psi_y\|_2 = 1$ ，因此有

$$\mathrm{tr}(\widehat{\Psi}\widehat{\Psi}^\top) = k, \qquad \textit{hence} \qquad \rho_{\mathrm{CN}} = \frac{1}{k}\widehat{\Psi}\widehat{\Psi}^\top.$$

在后续的概率性陈述（总体/经验扰动分析）中，我们用记号 $\widehat{\rho}_\star$ 表示总体类归一化算子 $\rho_\star$ 的经验估计量。

与基于计数的构造的关系及另一种统计视角。该算子 $\rho_{\mathrm{CN}}$ 是

$$\rho_{\mathcal{D}} = XX^{\top}/\operatorname{tr}(XX^{\top}), \qquad X = \sqrt{F},$$

其中平方根是逐元素取的。实际应用中，这两个变体都很有用： $\rho_{\mathcal{D}}$ 保留类别规模效应（通过类别质量体现），而 $\rho_{\mathrm{CN}}$ 则隔离了类条件分类分布的几何结构，并能导出更精确的理论结论。

从统计学角度看， $\rho_{\mathrm{CN}}$ 也可以解释为归一化振幅轮廓的经验非中心二阶矩算子。事实上，由于 $\operatorname{tr}(\widehat{\Psi}\widehat{\Psi}^{\top}) = k$，可得

$$\rho_{\mathrm{CN}} = \frac{1}{k}\widehat{\Psi}\widehat{\Psi}^{\top} = \frac{1}{k}\sum_{y=1}^{k}\widehat{\psi}_y\,\widehat{\psi}_y^{\top},$$

这正是向量 $\{\widehat{\psi}_y\}_{y=1}^{k}$的经验非中心二阶矩矩阵。若需要中心化协方差型算子$\widehat{\quad}\quad\widehat{\quad}\quad\bar{}$，则可以将 $\psi_y$ 替换为 $\psi_y - \psi$，其中

$$\bar{\psi} := \frac{1}{k}\sum_{y=1}^{k}\widehat{\psi}_y,$$

然后相应地重新归一化。

*5.2. Hellinger*几何与核主成分分析视角

所提出构造的一个独特之处在于，逐元素平方根提升会自然地在类条件分类分布上诱导出一种 Hellinger型几何。本小节将明确阐述这种联系，并说明由此得到的谱嵌入如何被解释为一种作用于概率向量、而非原始独热观测的核主成分分析过程。其底层相似性度量是经典的 Bhattacharyya系数，这是一种衡量概率分布之间亲和度的标准度量 [3, 1]。

**定义 5.4** （Hellinger距离与Bhattacharyya亲和度）。 设

$$\Delta^{d-1} := \Big\{p \in \mathbb{R}^d \;\Big|\; p_i \geq 0,\ \sum_{i=1}^{d} p_i = 1\Big\}$$

表示$(d-1)$维概率单纯形。设 $p$和 $q \in \Delta^{d-1}$ 是定义在 $\{1,\ldots,d\}$上的离散概率分布。 *Bhattacharyya*亲和度为

$$\mathrm{BC}(p,q) := \sum_{i=1}^{d}\sqrt{p(i)\,q(i)} \in [0,1],$$

且 Hellinger距离为

$$H(p,q) := \frac{1}{\sqrt{2}}\,\|\sqrt{p} - \sqrt{q}\|_2 = \sqrt{1 - \mathrm{BC}(p,q)} \in [0,1].$$

**备注 5.5** （几何意义）。 映射 $\Phi : \Delta^{d-1} \to \mathbb{R}^d$, $\Phi(p) = \sqrt{p}$ （逐元素）将单纯形嵌入单位球面的正卦限，因为 $\|\sqrt{p}\|_2^2 = \sum_i p(i) = 1$。因此， $\mathrm{BC}(p,q) = \langle\Phi(p)\ \bar{\Phi}(q)\rangle$ 是嵌入点的余弦相似度，而 $H(p,q)$ 则成为球面上的（经缩放的）欧氏距离。这一经典观点广泛应用于频数编码数据和分布比较 *[3, 1]*。

**命题 5.6**（类别间的海灵格亲和核）。 设 $\psi_y = \Phi(\widehat{p}_y)$ 为类别 $y$ 的振幅轮廓。则类别间格拉姆矩阵 $G \in \mathbb{R}^{k\times k}$ 定义为

$$G := \widehat{\Psi}^\top \widehat{\Psi}, \qquad G_{yy'} = \langle \widehat{\psi}_y, \widehat{\psi}_{y'} \rangle = \sum_{i=1}^{d} \sqrt{\widehat{p}_y(i)\, \widehat{p}_{y'}(i)},$$

是类别条件离散分布 $\widehat{p}_y$ 和 $\widehat{p}_{y'}$之间的经验 *Bhattacharyya*亲和矩阵。

**命题 5.7**（核结构与半正定性）。 亲和映射

$$K_{\mathrm{Hel}}(p,q) := \sum_{i=1}^{d} \sqrt{p(i)\, q(i)} = \langle \Phi(p), \Phi(q) \rangle$$

在 $\Delta^{d-1}$上定义了一个半正定核。因此， $G$ 是对称半正定的。

证明。 由于 $K_{\mathrm{Hel}}(p,q) = \langle \Phi(p)\ \Phi(q) \rangle$，其中 $\Phi(p) = \sqrt{p} \in \mathbb{R}^d$，因此它是一个内积核。因此，对于任意系数 $(a_y)_{y=1}^k$,

$$\sum_{y,y'=1}^{k} a_y a_{y'} K_{\mathrm{Hel}}(\widehat{p}_y, \widehat{p}_{y'}) = \left\| \sum_{y=1}^{k} a_y \sqrt{\widehat{p}_y} \right\|_2^2 \geq 0,$$

这证明了其半正定性。因此 $G = \widehat{\Psi}^\top \widehat{\Psi} \succeq 0$。 □

与核主成分分析和对应分析的关系。 核主成分分析（核主成分分析）通过对由半正定核 [15]诱导的格拉姆矩阵进行特征值分解，构造低维坐标。命题 5.7 表明， $G = \widehat{\Psi}^\top \widehat{\Psi}$ 恰好是将海林格/巴塔查亚核应用于类别条件轮廓 $(\widehat{p}_y)_{y=1}^k$后得到的格拉姆矩阵。因此，从 $\widehat{\Psi}\widehat{\Psi}^\top$ （或等价地从 $G$）提取的谱方向，可以在由振幅映射 $p \mapsto \sqrt{p}$诱导的特征空间中作核主成分分析解释，其中每个类别由 $\widehat{\psi}_y = \sqrt{\widehat{p}_y}$表示。这也阐明了它与对应分析（CA）和判别对应分析（DCA）的联系：两者都在 $\chi^2$类型的归一化 [8, 7]下，通过奇异值分解从列联表导出低维几何。相比之下，本文的构造以由$\sqrt{\cdot}$诱导的 Hellinger几何取代 $\chi^2$ 几何；这种几何适用于通过 Bhattacharyya亲和度和 Hellinger距离 [3, 1]比较离散分布。因此，所得谱表示天然适用于算子理论稳定性工具_（第 5.6 节），也适用于潜在空间中的下游概率建模。

### *5.3.* 总体算子与子空间稳定性（*Davis–Kahan*）

对任何谱嵌入方法而言，一个核心问题是稳定性：所学算子的主导特征子空间对由抽样噪声、有限数据或建模失配引起的扰动有多敏感。这一问题在矩阵扰动理论中是经典的，并由 Davis–Kahan $\sin\Theta$ 定理精确量化；该定理依据扰动大小与谱间隙，对对称算子不变子空间的旋转给出界。为完整起见，我们回顾一个适用于本设定的标准表述（见 [6, 19]）。

**定理 5.8** (Davis–Kahan 正弦 $\Theta$ 定理)。 设 $A$、 $\widetilde{A} \in \mathbb{R}^{d\times d}$ 为对称矩阵，并令 $\sim E := A - A$。固定 $r \in \{1, \ldots, d-1\}$，并假设存在一个谱间隙

$$\delta := \lambda_r(A) - \lambda_{r+1}(A) > 0,$$

其中 $\lambda_1(A) \geq \lambda_2(A) \geq \cdots \geq \lambda_d(A)$ 为 $A$ 的特征值。令 $U_r \in \mathbb{R}^{d\times r}$（分别）以 $\widetilde{U}_r \in \mathbb{R}^{d\times r}$（分别）以 $A$ 的前 $r$ 个特征向量为列。$\widetilde{A}$（分别）如此。于是

$$\left\|\sin\Theta(\widetilde{U}_r, U_r)\right\|_2 \leq \frac{\|E\|_2}{\delta} = \frac{\|\widetilde{A} - A\|_2}{\lambda_r(A) - \lambda_{r+1}(A)}.$$

定理 5.8 表明，只要扰动 $\|A - A\|_2$ 相较于特征值间隙 $\lambda_r(A) - \lambda_{r+1}(A)$ 足够小，前 $r$ 维特征子空间就具有稳定性。这为在采样波动下论证谱嵌入提供了一种精确且广泛使用的工具。

**假设 5.9** (分块分类模型与谱间隙)。 假设训练样本 $(x^{(j)}, y^{(j)})_{j=1}^n$ 独立同分布。给定 $Y = y$ 后，调查向量 $x \in \{0,1\}^d$ 按照调查模型由相互独立的分类变量块生成，并令 $p_y \in \Delta^{d-1}$ 表示（总体）类条件分类分布。定义总体振幅 $\psi_y := \sqrt{p_y}$ 和 $\Psi := [\psi_1 \cdots \psi_k] \in \mathbb{R}^{d\times k}$。令总体密度算子为

$$\rho_\star := \frac{\Psi\Psi^\top}{\operatorname{tr}(\Psi\Psi^\top)}.$$

固定嵌入维度 $r \leq k$，并假设存在谱间隙

$$\delta_r = \lambda_r(\rho_\star) - \lambda_{r+1}(\rho_\star) > 0,$$

其中，$\lambda_1(\rho_\star) \geq \lambda_2(\rho_\star) \geq \cdots$ 表示 $\rho_\star$ 的特征值。

**记号 5.10** (主导特征子空间)。 令 $U_{\star,r} \in \mathbb{R}^{d\times r}$ 以列向量形式包含 $\rho_\star$ 的前 $r$ 个特征向量，并令 $\widehat{U}_{\star,r} \in \mathbb{R}^{d\times r}$ 以列向量形式包含 $\widehat{\rho}_\star$ 的前 $r$ 个特征向量。我们用 主角正弦值 $\Theta(\widehat{U}_{\star,r}, U_{\star,r})$ 表示相应 $r$ 维子空间之间主角正弦值构成的对角矩阵。

**定理 5.11** (谱子空间稳定性)。 在假设 *5.9* 下，经验算子的主导 $r$ 维谱子空间 $\widehat{\rho}_\star$ 对总体算子 $\rho_\star$ 的扰动具有稳定性，具体而言：

$$\left\|\sin\Theta(\widehat{U}_{\star,r}, U_{\star,r})\right\|_2 \leq \frac{\|\widehat{\rho}_\star - \rho_\star\|_2}{\delta_r}.$$

证明。 对 $A = \rho_\star$ 和 $A = \widehat{\rho}_\star$ 应用定理 5.8。则 $E = \widehat{\rho}_\star - \rho_\star$ 且特征值间隙为 $\delta_r = \lambda_r(\rho_\star) - \lambda_{r+1}(\rho_\star)$，由此得到所述界。 □

*5.4.* 多项式抽样下 $\|\widehat{\rho}_\star - \rho_\star\|_2$ 的扰动界

为补充 Davis–Kahan 稳定性估计（定理 5.11），我们给出分块分类生成器下 $\|\widehat{\rho}_\star - \rho_\star\|_2$ 的一个简单非渐近界。我们用记号 $\widehat{\rho}_\star$ 表示总体类归一化算子 $\rho_\star$ 的经验估计量，以将其与定义 5.3 中引入的有限样本类归一化算子 $\rho_{\mathrm{CN}}$ 区分开来。

**假设 5.12** (类别分布的严格正性). 存在 $p_{\min} \in (0,1)$，使得对于所有类别 $y \in \{1,\ldots,k\}$ 和所有坐标 $i \in \{1,\ldots,d\}$,

$$p_y(i) \geq p_{\min}.$$

**引理 5.13** (从因子扰动到算子扰动). 设 $\Psi, \Psi_n \in \mathbb{R}^{d\times k}$ 并定义

$$\rho_\star := \frac{1}{k}\Psi\Psi^\top, \qquad \widehat{\rho}_\star := \frac{1}{k}\Psi_n\Psi_n^\top.$$

*Then*

$$\|\widehat{\rho}_\star - \rho_\star\|_2 \leq \frac{1}{k}\Big(\|\Psi_n - \Psi\|_2^2 + 2\|\Psi\|_2\,\|\Psi_n - \Psi\|_2\Big).$$

证明。 写作 $\Delta := \Psi_n - \Psi$。可以验证 $\hat{}\rho_\star = \frac{1}{k}(\Psi + \Delta)(\Psi_- + \Delta)^\top$。于是 :

$$\widehat{\rho}_\star - \rho_\star = \frac{1}{k}\big((\Psi+\Delta)(\Psi+\Delta)^\top - \Psi\Psi^\top\big) = \frac{1}{k}\big(\Psi\Delta^\top + \Delta\Psi^\top + \Delta\Delta^\top\big).$$

取谱范数并利用柯西-施瓦茨不等式，可得:

$$\|\widehat{\rho}_\star - \rho_\star\|_2 \le \frac{1}{k}\big(2\|\Psi\|_2\|\Delta\|_2 + \|\Delta\|_2^2\big),$$

此即所证。 □

**定理 5.14** （ $\|\widehat{\rho}_\star - \rho_\star\|_2$的高概率界）。 假设样本按如下方式生成：对每个类别 $y$，我们观测到 $n_y \ge 1$ 个独立同分布的标记样本，其独热调查向量服从分块分类模型，并令 $p_y \in \Delta^{d-1}$ 为独热坐标上的理论类条件分布。令 $\widehat{p}_y$ 表示类别 $y$中的经验分布（相对频率），并定义

$$\psi_y := \sqrt{p_y}, \qquad \widehat{\psi}_y := \sqrt{\widehat{p}_y} \quad \textit{(entrywise)},$$

以及

$$\Psi = [\psi_1, \dots, \psi_k], \qquad \Psi_n = [\widehat{\psi}_1, \dots, \widehat{\psi}_k], \qquad \rho_\star = \frac{1}{k}\Psi\Psi^\top, \qquad \widehat{\rho}_\star = \frac{1}{k}\Psi_n\Psi_n^\top.$$

在假设 *5.12* 下，令

$$n_{\min} := \min_{1\le y\le k} n_y \ge 0.$$

固定 $\delta \in$ （0，1），并进一步假设

$$\sqrt{\frac{\log(4dk/\delta)}{2n_{\min}}} \le \frac{p_{\min}}{2}. \tag{1}$$

那么，以至少 $1-\delta$ 的概率，以下结论同时成立:

$$\max_{1\le y\le k} \|\widehat{p}_y - p_y\|_\infty \le \sqrt{\frac{\log(4dk/\delta)}{2n_{\min}}}, \tag{2}$$

$$\|\widehat{\rho}_\star - \rho_\star\|_2 \le \frac{1}{k}\Big(2\|\Psi\|_2\,\varepsilon + \varepsilon^2\Big), \tag{3}$$

其中

$$\varepsilon :=, \sqrt{\frac{kd\log(4dk/\delta)}{4p_{\min}n_{\min}}}. \tag{4}$$

证明。 我们分三步进行。

步骤 *1*: 经验类别分布的一致集中。 固定一个类别 $y$ 和一个坐标 $i$。给定类别计数 $n_y$ 后，经验频率 $\widehat{p}_y(i)$ 是 $n_y$ 个独立同分布伯努利随机变量的样本均值，其均值为 $p_y(i)$，取值介于 $p_{min}$ 和 1 之间。因此，由霍夫丁不等式 [9, 5] 可得，对每个 $t > 0$,

$$\Pr\Big(\big|\widehat{p}_y(i) - p_y(i)\big| \ge t\Big) \le 2\exp(-2n_y t^2) \le 2\exp(-2n_{\min}t^2).$$

对 $y \in \{1,\dots,k\}$ 和 $i \in \{1,\dots,d\}$ 应用并集界可得

$$\Pr\Big(\max_{y,i}|\widehat{p}_y(i) - p_y(i)| \ge t\Big) \le 2dk\,\exp\Big(\frac{-2t^2}{n_{\min}}\Big).$$

选择

$$t_\delta := \sqrt{\frac{\log(2dk/\delta)}{2n_{\min}}}$$

使右侧至多为 $\delta$。因此，以至少 $1-\delta$ 的概率，

$$\max_{1\le y\le k} \|\widehat{p}_y - p_y\|_\infty \le t_\delta.$$

这就证明了 (2)。

步骤 *2*：正性传递与对 $\|\Psi_n - \Psi\|_2$ 的控制。 在步骤 1 的事件发生时，条件 (1) 表明，对于 $\delta > 0$，以 $1-\delta$ 的概率，

$$\widehat{p}_y(i) \ge p_y(i) - \sqrt{\frac{\log(4dk/\delta)}{2n_{\min}}} \ge p_{\min} - \frac{p_{\min}}{2} = \frac{p_{\min}}{2} \qquad \forall\, y, i.$$

因此，$p_y(i)$ 和 $\widehat{p}_y(i)$ 均属于 $[p_{\min}/2, 1]$，并且已知平方根在该区间内是利普希茨函数 [20]，其常数为 $p_{min} \ge 0$。更确切地说，可以证明：

$$\|\widehat{\psi}_y - \psi_y\|_2 \le \frac{1}{\sqrt{2p_{\min}}} \|\widehat{p}_y - p_y\|_2 \le \frac{\sqrt{d}}{\sqrt{2p_{\min}}} \|\widehat{p}_y - p_y\|_\infty.$$

因此，

$$\|\Psi_n - \Psi\|_2 \le \|\Psi_n - \Psi\|_F = \Big(\sum_{y=1}^{k} \|\widehat{\psi}_y - \psi_y\|_2^2\Big)^{1/2} \le \sqrt{k}\, \frac{\sqrt{d}}{\sqrt{2p_{\min}}} \max_y \|\widehat{p}_y - p_y\|_\infty.$$

结合 (2) 可得，以 $1-\delta$ 的概率，

$$\|\Psi_n - \Psi\|_2 \le \varepsilon,$$

其中 $\varepsilon$ 在（4）中定义。

步骤 *3*：算子扰动界。应用引理 5.13 和估计 $\|\Psi_n - \Psi\|_2 \le \varepsilon$，可得

$$\|\widehat{\rho}_\star - \rho_\star\|_2 \le \frac{1}{k}\Big(2\|\Psi\|_2\, \varepsilon + \varepsilon^2\Big),$$

这就是（3）。 □

**注 5.15**。正性假设用于为逐元平方根映射 $p \mapsto \sqrt{p}$ 得到一个统一的利普希茨常数。更具体地说，需要确保经验分布 $\widehat{p}_y$ 始终处于总体分布剖面的一个正邻域内。条件 （1）是一个充分的样本量要求，可确保这一性质在高概率事件上成立。实践中，也可以通过标准的加性平滑达到同样的效果。

结合定理 *5.14* 和 *Davis–Kahan* 子空间扰动估计（定理 *5.11*），可根据（$d$、$k$、$n_{\min}$、$p_{\min}$）及谱间隙，得到经验 *DMM* 嵌入子空间与总体 *DMM* 嵌入子空间之间主角距离的一个明确（尽管较为保守的）高概率控制。

**推论 5.16**（正交对齐下潜在坐标的稳定性）。 设

$$\pi_{\star,r}(x) := U_{\star,r}^\top \widetilde{x}, \qquad \widehat{\pi}_{\star,r}(x) := \widehat{U}_{\star,r}^\top,$$

为归一化调查向量的总体谱坐标与经验谱坐标， $\widetilde{x}:= x/\sqrt{q}$。则存在绝对常数 $C > 0$，使得对所有 $x \in \mathcal{X}_d$，

$$\inf_{R\in O(r)} \big\|\widehat{\pi}_{\star,r}(x) - R^\top \pi_{\star,r}(x)\big\|_2 \;\le\; C\, \big\|\sin\Theta(\widehat{U}_{\star,r}, U_{\star,r})\big\|_2,$$

其中 $O(r)$ 表示 $\mathbb{R}^{r\times r}$中的正交群。

证明。 固定 $x \in \mathcal{X}_d$，并令 $R \in O(r)$。由于 $\|\widetilde{x}\|_2 = 1$,

$$\|\widehat{\pi}_{\star,r}(x) - R^\top \pi_{\star,r}(x)\|_2 = \|\widehat{U}_{\star,r}^\top \widetilde{x} - R^\top U_{\star,r}^\top \widetilde{x}\|_2 = \|(\widehat{U}_{\star,r} - U_{\star,r}R)^\top \widetilde{x}\|_2 \le \|\widehat{U}_{\star,r} - U_{\star,r}R\|_2.$$

对 $R \in O(r)$ 取下确界，得到

$$\inf_{R\in O(r)} \|\widehat{\pi}_{\star,r}(x) - R^\top \pi_{\star,r}(x)\|_2 \le \inf_{R\in O(r)} \|\widehat{U}_{\star,r} - U_{\star,r}R\|_2.$$

后者可通过 Davis–Kahan $\sin\Theta$ 定理中的主角距离控制（相差一个绝对常数）：

$$\inf_{R\in O(r)} \|\widehat{U}_{\star,r} - U_{\star,r}R\|_2 \le C\,\|\sin\Theta(\widehat{U}_{\star,r}, U_{\star,r})\|_2,$$

由此得证。 □

**注 5.17** （一种有原则的比较手段）。 定理 *5.11* 和推论 *5.16* 提供了一种严谨机制，用于将 *DMM* 型谱嵌入与其他降维方法进行比较：在存在谱间隙时，所学习的嵌入在采样波动和扰动下保持稳定。这在稀疏高基数情形中尤为重要，因为基于距离的方法（例如在环境独热空间中使用 *K* 近邻）可能因集中现象而表现出较大的方差，而 *DMM* 嵌入则受到算子扰动界的控制。

### *5.5.* 最简演算示例

**示例 5.18** （玩具双类别设置中的 Bhattacharyya 亲和度）。 考虑 $k = 2$ 个类别，其经验分布为 $\widehat{p}_1$, $\widehat{p}_2 \in \Delta^{d-1}$。则

$$G_{12} = \sum_{i=1}^{d} \sqrt{\widehat{p}_1(i)\widehat{p}_2(i)} \in [0,1]$$

量化类别相似性： $G_{12} = 1$ 当且仅当 $\widehat{p}_1 = \widehat{p}_2$时成立，而当支撑集不相交时， $G_{12} = 0$ 。在此设置下， $\rho_{\mathrm{CN}}$ 的主导谱方向通常强调对类别分离贡献最大的坐标。

### *5.6.* 基于计数的算子与类别归一化算子之间的关系

**记号 5.19** （类别质量与归一化分布）。 设 $F \in \mathbb{R}^{d\times k}$ 为类别条件计数矩阵，其列为$\mathbf{f}_y$。定义类别 $y$ 的类别质量（总计数）为

$$s_y := \sum_{i=1}^{d} F_{i,y} = \mathbf{1}^\top \mathbf{f}_y, \qquad S := \sum_{y=1}^{k} s_y.$$

假设对所有 $y$， $s_y > 0$ 均成立。定义类别归一化类别分布

$$p_y := \frac{\mathbf{f}_y}{s_y} \in \Delta^{d-1}, \qquad \psi_y := \sqrt{p_y} \in \mathbb{R}^d, \qquad \Psi := [\psi_1 \cdots \psi_k] \in \mathbb{R}^{d\times k},$$

平方根按元素逐一取值。

**命题 5.20** （精确分解与类别不平衡权重）. 采用定义 *5.3* 与记号 *5.19* 的记号，这两个算子具有显式形式

$$\rho_{\mathcal{D}} \;=\; \frac{1}{S}\,\Psi\,\mathrm{diag}(s_1,\ldots,s_k)\,\Psi^\top \;=\; \sum_{y=1}^{k} w_y\,\psi_y\psi_y^\top, \qquad w_y := \frac{s_y}{S},$$

*and*

$$\rho_{\mathrm{CN}} \;=\; \frac{1}{k}\,\Psi\Psi^{\top} \;=\; \frac{1}{k}\sum_{y=1}^{k}\psi_y\psi_y^{\top}.$$

特别地， $\rho_{\mathcal{D}}$ 是秩一矩阵 $\psi_y\psi_y^{\top}$ 的凸组合，其权重为经验类别质量 $w_y$，而 $\rho_{\mathrm{CN}}$ 则赋予均匀权重 $1/k$。

证明。 由于$\mathbf{f}_y = s_y p_y$，我们逐项有

$$\sqrt{\mathbf{f}_y} = \sqrt{s_y}\,\sqrt{p_y} = \sqrt{s_y}\,\psi_y.$$

因此

$$X = [\sqrt{\mathbf{f}_1}\;\cdots\;\sqrt{\mathbf{f}_k}] = [\sqrt{s_1}\psi_1\;\cdots\;\sqrt{s_k}\psi_k] = \Psi\;\mathrm{diag}(\sqrt{s_1},\ldots,\sqrt{s_k}).$$

因此

$$XX^{\top} = \Psi\;\mathrm{diag}(s_1,\ldots,s_k)\,\Psi^{\top}.$$

此外， $\|\psi_y\|_2^2 = \sum_{i=1}^{d} p_y(i) \;= 1$，所以

$$\mathrm{tr}(XX^{\top}) = \sum_{y=1}^{k} s_y\,\|\psi_y\|_2^2 = \sum_{y=1}^{k} s_y = S.$$

这给出 $\rho_{\mathcal{D}} = \frac{1}{S}\Psi\,\underline{\mathrm{diag}}(s_1,\ldots,s_k)\Psi^{\top} = \sum_y (s_y/S)\psi_y\psi_y^{\top}$。

对于类别归一化算子，注意 $\mathrm{tr}(\Psi\Psi^{\top}) \;= \sum_{y=1}^{k}\|\psi_y\|_2^2 = k$，因此

$$\rho_{\mathrm{CN}} = \frac{1}{k}\Psi\Psi^{\top} = \frac{1}{k}\sum_{y}\psi_y\psi_y^{\top}.$$

□

**命题 5.21** （类别平衡蕴含等价）。 若类别质量平衡，即$s_y = S/k$ 对所有 $y$，则

$$\rho_{\mathcal{D}} = \rho_{\mathrm{CN}}.$$

证明。 若 $s_y = S/k$对所有 $y$，则 $\mathrm{diag}(s_1,\;\ldots\ldots,\;s_k) \;= (S/k)I_k$且

$$\rho_{\mathcal{D}} = \frac{1}{S}\Psi\left(\frac{S}{k}I_k\right)\Psi^{\top} = \frac{1}{k}\Psi\Psi^{\top} = \rho_{\mathrm{CN}}.$$

□

**命题 5.22** （算子偏差受类别不平衡控制）。 设 $w_y \;:=\; s_y/S$ 并定义不平衡程度

$$\delta_w := \max_{1\le y\le k}\left|w_y - \frac{1}{k}\right|.$$

*Then*

$$\|\rho_{\mathcal{D}} - \rho_{\mathrm{CN}}\|_2 \;\le\; \|\Psi\|_2^2\,\delta_w.$$

特别地，由于 $\|\Psi\|_2^2 = \lambda_{\max}\left(\Psi^{\top}\Psi\right)$，该偏差受巴塔恰里亚／海灵格格拉姆矩阵的谱尺度以及类别质量不平衡的控制。

证明。 利用命题 5.20，我们写出

$$\rho_{\mathcal{D}} - \rho_{\mathrm{CN}} = \Psi\Big(\mathrm{diag}(w_1,\ldots,w_k) - \frac{1}{k}I_k\Big)\Psi^{\top}.$$

取谱范数并利用次乘性，

$$\|\rho_{\mathcal{D}} - \rho_{\mathrm{CN}}\|_2 \le \|\Psi\|_2^2\,\left\|\mathrm{diag}(w_1,\ldots,w_k) - \frac{1}{k}I_k\right\|_2.$$

由于最后一个因子是对角的，其谱范数等于最大绝对对角元素，即 $\delta_w$。

□

**备注 5.23**（实际意义）。命题 *5.20* 表明，$\rho_{\mathcal{D}}$ 与 $\rho_{\mathrm{CN}}$ 的差异仅在于它们对类别的加权方式。因此，当旨在消除类别流行率的影响时，$\rho_{\mathrm{CN}}$ 更为可取；而当流行率具有信息性，或目标是拟合完整数据集的经验分布时，$\rho_{\mathcal{D}}$ 则可能更可取。命题 *5.22* 进一步表明，当类别不平衡 $\delta_w$ 较小时，两个算子彼此接近，因而诱导出相似的嵌入。

**推论 5.24 (**轻度不平衡下的嵌入等价性)。令 $\rho_{\mathcal{D}}$ 和 $\rho_{\mathrm{CN}}$ 分别为上文定义的基于计数的算子和类别归一化算子。固定 $r \le k$，并假设 $\rho_{\mathrm{CN}}$ 满足谱间隙条件

$$\lambda_r(\rho_{\mathrm{CN}}) > \lambda_{r+1}(\rho_{\mathrm{CN}}).$$

令 $U_r^{(\mathrm{CN})} \in \mathbb{R}^{d\times r}$ 为 $\rho_{\mathrm{CN}}$的前 $r$ 个特征向量组成的矩阵，令 $U_{(\mathcal{D})r} \in \mathbb{R}^{d\times r}$ 为 $\rho_{\mathcal{D}}$的前 $r$ 个特征向量组成的矩阵。那么

$$\left\|\sin\Theta\big(U_r^{(\mathcal{D})}, U_r^{(\mathrm{CN})}\big)\right\|_2 \;\le\; \frac{\|\rho_{\mathcal{D}} - \rho_{\mathrm{CN}}\|_2}{\lambda_r(\rho_{\mathrm{CN}}) - \lambda_{r+1}(\rho_{\mathrm{CN}})} \;\le\; \frac{\|\Psi\|_2^2\,\delta_w}{\lambda_r(\rho_{\mathrm{CN}}) - \lambda_{r+1}(\rho_{\mathrm{CN}})},$$

其中 $\Psi = [\psi_1 \cdots \psi_k]$ 和 $\delta_w = $最大值$_y \left|w_y - \frac{1}{k}\right|$ 如命题 *5.22* 所示。特别地，如果不平衡程度 $\delta_w$ 较小，那么由 $\rho_{\mathcal{D}}$ 和 $\rho_{\mathrm{CN}}$诱导的主导 $r$维嵌入在主角意义下彼此接近（相差一个旋转）。

证明。应用定理 5.8，并取

$$A = \rho_{\mathrm{CN}}, \qquad \widetilde{A} = \rho_{\mathcal{D}}.$$

由此可得

$$\left\|\sin\Theta\big(U_r^{(\mathcal{D})}, U_r^{(\mathrm{CN})}\big)\right\|_2 \le \frac{\|\rho_{\mathcal{D}} - \rho_{\mathrm{CN}}\|_2}{\lambda_r(\rho_{\mathrm{CN}}) - \lambda_{r+1}(\rho_{\mathrm{CN}})}.$$

第二个不等式由命题 5.22 得出。 □

## 6. 通过类条件密度估计进行潜在空间分类

谱映射 $\pi_r : \mathcal{X}_d \to \mathbb{R}^r$ 将稀疏独热调查向量变换为紧凑的潜在表示。在本节中，我们通过为每个类别估计类条件密度，然后应用最大后验（MAP）规则，在潜在空间中进行分类。我们采用核密度估计（核密度估计），这是一种由 Rosenblatt 和 Parzen [14, 13]提出并在统计学文献 [17, 16, 18]中得到广泛发展的经典非参数方法。

### *6.1.* 类条件潜在点云

**定义 6.1**（类条件潜在点云）**.** 对于每个类别 $y \in \{1, \dots, k\}$，定义潜在样本点云

$$\Pi_y := \{\pi_r(x^{(j)}) : y^{(j)} = y\} \subset \mathbb{R}^r, \qquad n_y := |\Pi_y|.$$

**备注 6.2**（潜在空间中的支撑约束）。根据构造，对于所有 $\pi_r(x) \in B_1^r$（0）$x \in \mathcal{X}_d$ 均成立（第 *4* 节中的性质）。这表明有两种在实践中相关的核选择：（*i*）$\mathbb{R}^r$ 上的平滑核（例如高斯核），这类核忽略边界，但具有较强的稳健性；（*ii*）紧支撑核（例如 *Epanechnikov* 型核），这类核可以强制满足潜在空间中的支撑约束。

### *6.2.* $\mathbb{R}^r$中的核密度估计

下文将采用 [16]提出的记号。

**定义 6.3**（$\mathbb{R}^r$中的核密度估计量）。设 $K : \mathbb{R}^r \to [0, \infty)$ 满足 $\int_{\mathbb{R}^r} K(u)\, du = 1$，并令 $h > 0$ 为带宽。对于每个类别 $y$，定义核密度估计量

$$\widehat{f}_h(z \mid y) = \frac{1}{n_y\, h^r} \sum_{w \in \Pi_y} K\!\left(\frac{z - w}{h}\right), \qquad z \in \mathbb{R}^r.$$

**命题 6.4**（估计量的基本性质）。假设 $K$ 可积且具有单位质量。那么对于每个 $y$,

$$\widehat{f}_h(\cdot \mid y) \geq 0, \qquad \int_{\mathbb{R}^r} \widehat{f}_h(z \mid y)\, dz = 1.$$

证明。 由 $K \geq 0$ 可知非负性显然成立。对于归一化，利用变量替换$u = (z - w)/h$,

$$\int_{\mathbb{R}^r} \widehat{f}_h(z \mid y)\, dz = \frac{1}{n_y\, h^r} \sum_{w \in \Pi_y} \int_{\mathbb{R}^r} K\left(\frac{z-w}{h}\right) dz = \frac{1}{n_y} \sum_{w \in \Pi_y} \int_{\mathbb{R}^r} K(u)\, du = 1.$$

□

**注 6.5**（带宽选择）。 带宽 $h$ 控制核密度估计的偏差—方差权衡，对分类准确率至关重要。实践中，可以通过经验法则、插件估计量或交叉验证程序来选择 $h$ ； 标准指导请参见 *[17, 16, 18]*。

*6.3* 后验得分与决策规则

**定义 6.6**（经验类别先验）。 定义经验类别先验

$$\widehat{\pi}_y := \frac{n_y}{n}, \qquad y \in \{1, \ldots, k\}.$$

更一般地，可以指定先验 $\pi_y \in (0,\ 1)$ , 其中 $\sum_y \pi_y = 1$ 用于反映特定应用成本或类别不平衡。

**定义 6.7**（潜在空间中的 MAP 分类器）。 给定类别先验 $(\pi_y)_{y=1}^{k}$，潜在空间中的诱导分类器为

$$\ell_{\widetilde{\mathcal{D}}_r}(z) = \mathbf{e}_{\widehat{y}}, \qquad \widehat{y} \in \arg\max_{1 \leq y \leq k}\ \pi_y\, \widehat{f}_h(z \mid y).$$

原始调查空间中的最终分类器为

$$\ell_{\mathcal{D}}(x) := \ell_{\widetilde{\mathcal{D}}_r}(\pi_r(x)).$$

决策规则与数值实现。 如果为所有类别分配相等的先验（或忽略先验），则最大后验概率规则会简化为最大似然规则

$$\widehat{y} \in \arg\max_{y}\ \widehat{f}_h(z \mid y).$$

当抽样分布存在类别不平衡时，纳入先验$\pi_y$ （例如经验先验$\widehat{\pi}_y$）会使决策规则与所选先验下的贝叶斯风险保持一致；然而，如实验 S4 所示，当先验高度偏斜时，这可能会降低少数类敏感指标，例如平衡准确率和宏平均 F1。

在数值实现中，通常使用对数得分来计算决策规则,

$$\widehat{y} \in \arg\max_{y}\ \Big(\log \pi_y + \log \widehat{f}_h(z \mid y)\Big),$$

^这在中等维度或高维情况下，当 $f_h(z \mid y)$ 变得非常小时，可以提高稳健性。

*6.4* 一致性视角（可选的理论背景）

贝叶斯规则与插件分类。 令 $f(z \mid y)$ 表示潜在空间中的真实类条件密度，并令 $\pi_y$ 表示类别先验。贝叶斯决策规则是最大后验概率分类器

$$y^\star(z) \in \arg\max_{y}\ \pi_y\, f(z \mid y),$$

它在所有可测决策规则中使误分类风险最小。一种标准做法是将 $f(z \mid y)$ 替换为非参数估计，从而构造插件分类器。在我们的设定中，我们采用核密度估计（KDE），遵循经典的 Parzen–Rosenblatt 框架 [14, 13, 17], ，在降维坐标中得到显式的似然（或后验）得分。

嵌入维度作为谱截断参数。截断阶数 $r$ 是一个纯粹的谱参数：它决定用于嵌入的不变子空间的维度，进而控制保留的自由度数量。从密度估计的角度来看，较小的 $r$ 有助于提高核密度估计的集中性和数值稳定性，而过大的 $r$ 则可能因维数灾难而降低性能。因此，可以依据谱衰减准则（例如密度算子的特征值间隙）或通过基于验证的调优来选择 $r$，具体取决于侧重的是算子理论结构还是预测性能。

前面的定义给出了本文中始终使用的潜空间分类器。接下来，我们概述端到端计算流程并推导其复杂度，其中包括一种基于类别 Gram 矩阵的高效低秩实现。

## 7. 算法实现与复杂度分析

在定义了谱嵌入和潜空间核密度估计/最大后验分类器之后，我们现在概述完整的计算流程及其基于矩阵的实现。关键观察是，密度算子为半正定且本质上是低秩的（命题 3.6），因此可以通过 $k \times k$ 格拉姆矩阵而非完整的 $d \times d$ 特征值分解来恢复其非零谱信息。因此，本节将介绍：（i）构造算子、提取截断不变子空间并进行潜空间分类的端到端流程；（ii）复杂度分析，重点说明 $k \ll d$高基数独热编码中典型的情形。

符号约定。 本节中的算法针对基于计数的 算子 $\rho_{\mathcal{D}}$编写。讨论第 5 节中引入的类别归一化替代方案时，我们使用符号$\rho_{\mathrm{CN}}$ （而非 $\widehat{\rho}$），以避免与扰动分析中使用的经验总体估计量符号冲突。

### *7.1.* 算法总结

算法 1 概述了完整流程。

算法 1 密度矩阵谱嵌入 + 核密度估计分类（基于计数的算子） 要求：带标签数据集 $\mathcal{D} = \{($ $x^{(j)}, y^{(j)})\}_{j=1}^{n}$，类别数 $k$，截断参数 $r \le k$，核$K$，带宽 $h$确保：分类器 $\ell_{\mathcal{D}}(\cdot)$ 1：将每个样本编码为独热调查向量 $x^{(j)} \in \{0,1\}^d$2：构建类别计数 $\mathbf{f}_y \leftarrow \sum_{j:\, y^{(j)}=y} x^{(j)}$，其中 $y = 1, \ldots, k$3：构造 $F \leftarrow [\mathbf{f}_1 | \cdots | \mathbf{f}_k] \in \mathbb{R}^{d \times k}$4：计算振幅 $X \leftarrow \sqrt{F}$ （逐元素） 5：计算基于计数的密度算子 $\rho_{\mathcal{D}} \leftarrow XX^\top / \mathrm{tr}(XX^\top)$ 6：计算前 $r$ 个特征向量 $U_r \in \mathbb{R}^{d \times r}$ of $\rho_{\mathcal{D}}$ （通过 Gram 矩阵路径；参见第 7.2 节） 7：定义嵌入 $\pi_r(x) \leftarrow U_r^\top \left( x / \sqrt{q} \right)$8：对于每个类别 $y$，在 $\Pi_y = \{\pi_r(x^{(j)}) : y^{(j)} = y\}$ 上拟合核密度估计 $\widehat{f}_h(\cdot \mid y)$ 9：预测 $\ell_{\mathcal{D}}(x) \leftarrow \arg\max_y \widehat{f}_h\left(\pi_r(x) \mid y\right)$

–

### *7.2.* 通过类别 *Gram* 矩阵求非零谱和特征向量

算子 $\rho_{\mathcal{D}} = XX^\top / \mathrm{tr}(XX^\top)$ 是一个 $d \times d$ 半正定矩阵，其秩为 $\mathrm{rank}(\rho_{\mathcal{D}}) \le k$ （命题 3.6）。这种低秩结构意味着， $\rho_{\mathcal{D}}$ 的所有非零谱信息都已包含在规模小得多的 $k \times k$ 格拉姆矩阵 $G := X^\top X$ 中。下面的结果将这一等价关系形式化，并给出如何利用 $G$的特征向量显式构造 $XX^\top$ 的特征向量。

以下命题是经典结果（参见 [4]）：

**命题 7.1**（ $XX^\top$ 与 $X^\top X$的非零谱）。 令 $X \in \mathbb{R}^{d\times k}$ 并设

$$A := XX^\top \in \mathbb{R}^{d\times d}, \qquad G := X^\top X \in \mathbb{R}^{k\times k}.$$

则以下结论成立。

1. 矩阵 $A$和 $G$ 均为对称半正定矩阵，且具有相同的非零特征值多重集（按代数重数计）。特别地，

$$\operatorname{rank}(A) = \operatorname{rank}(G) = \operatorname{rank}(X) \le k.$$

2. 如果 $v \in \mathbb{R}^k$ 满足 $Gv = \lambda v$ ，其中 $\lambda > 0$为某个值，那么 $u := Xv \in \mathbb{R}^d$ 非零且满足 $Au = \lambda u$。此外， $\|u\|_2^2 = \lambda\|v\|_2^2$。3. 如果 $u \in \mathbb{R}^d$ 满足 $Au = \lambda u$ ，其中 $\lambda > 0$为某个值，那么 $v := X^\top u \in \mathbb{R}^k$ 非零且满足$Gv = \lambda v$。

**推论 7.2**（由格拉姆矩阵得到的 $\rho_{\mathcal{D}}$ 的特征对）。 设 $\rho_{\mathcal{D}} = XX^\top/\operatorname{tr}(XX^\top)$ 且 $G = X^\top X$。若 $(\lambda_i, v_i)$ 是 $G$ 的特征对，其中 $\lambda_i > 0$ 且 $\|v_i\|_2 = 1$，则 $\rho_{\mathcal{D}}$ 的特征对 $(\sigma_i, u_i)$ 由下式给出：

$$\sigma_i = \frac{\lambda_i}{\operatorname{tr}(XX^\top)} \quad and \quad u_i = \frac{1}{\sqrt{\lambda_i}} X v_i$$

并满足 $\|u_i\|_2 = 1$。

证明。 根据命题 7.1(2)， $XX^\top(Xv_i) = \lambda_i(Xv_i)$。除以 $\operatorname{tr}(XX^\top)$ 后，得到 $\rho_{\mathcal{D}}$ 的一个特征对；根据命题 7.1(2)，归一化 $\|u_i\|_2 = 1$ 也成立。 □

*7.3.* 计算复杂度

设 $n$ 为样本数， $d$ 为独热维度， $q$ 为类别问题（块）的数量， $k$ 为类别数。我们还用 $r \le k$ 表示谱截断后保留的潜在嵌入维度。

步骤 *1*：构建类别条件计数。 构建频率矩阵 $F = [\mathbf{f}_1|\cdots|\mathbf{f}_k] \in \mathbb{R}^{d\times k}$ 只需遍历一次数据集。在最坏情况下（稠密表示），其计算成本为$O(nd)$。然而，独热调查结构意味着每个样本恰有 $q$ 个非零元素，因此每个样本的更新可在 $O(q)$ 时间内完成。因此，利用稀疏性可使总成本

$$O(nq),$$

存储所需内存为 $O(dk)$ $F$ （采用流式／稀疏累积策略时则更少）。

步骤 *2*：振幅提升。振幅矩阵 $X = \sqrt{F} \in \mathbb{R}^{d\times k}$ 通过逐项计算得出，所需运算次数为 $O(dk)$。当 $F$ 为稀疏矩阵时（高基数场景下通常如此），此步骤可简化为 $O(\operatorname{nnz}(F))$，其中 $\operatorname{nnz}(F)$ 表示非零元素的数量。

步骤 *3*：通过低秩路径进行谱分解。显式计算 $\rho_{\mathcal{D}} = XX^\top/\operatorname{tr}(XX^\top)$ 需要 $O(d^2k)$ 的运算量来构造 $XX^\top$ ，并需要 $O(d^2)$ 的内存，这对于大型 $d$ 来说代价过高。相反，我们利用秩的内在上界 $\operatorname{rank}(\rho_{\mathcal{D}}) \le k$ （命题 3.6），并从较小的格拉姆矩阵中计算谱信息

$$G := X^\top X \in \mathbb{R}^{k\times k}.$$

构造 $G$ 的运算量为 $O(dk^2)$（采用稀疏运算时为 $O(\operatorname{nnz}(X)\, k)$）。若 $G = V\Lambda V^\top$ 的特征分解对应正特征值 $\lambda_1, \ldots, \lambda_s$，则 $XX^\top$ 对应的非零特征向量可表示为

$$U_s = XV_s\Lambda_s^{-1/2}, \qquad XX^\top = U_s\Lambda_s U_s^\top,$$

其中， $V_s$ 收集了与正谱相关的特征向量， $\Lambda_s = \operatorname{diag}(\lambda_1, \ldots, \lambda_s)$。 $G$ 的特征值分解耗时为 $O(k^3)$，因此谱阶段的总复杂度为

$$O(dk^2 + k^3),$$

内存为 $O(dk + k^2)$，当 $k \ll d$ 时可行。

步骤 *4*：计算嵌入。给定前 $r$ 个特征向量 $U_r \in \mathbb{R}^{d\times r}$，嵌入一个新观测值 $x \in \mathcal{X}_d$ 需要

$$\pi_r(x) = U_r^\top \widetilde{x}, \qquad \widetilde{x} = x/\sqrt{q}.$$

在稠密运算中，每次查询的开销为 $O(dr)$。利用独热块稀疏性（恰好有 $q$ 个 1），计算可简化为

$$O(qr),$$

因为 $U_r^\top x$ 只需处理 $x$ 的 $q$ 个活跃坐标。

步骤 *5*：核密度估计训练与预测。构造潜在点云 $\Pi_y$ 需要嵌入所有样本；利用稀疏性时，开销为 $O(nqr)$（稠密运算时则为 $O(ndr)$）。核密度估计模型的总存储开销为 $O(nr)$。对于朴素的核密度估计，单次查询的开销为

$$O\left(\sum_{y=1}^{k} n_y r\right) = O(nr),$$

假设每个类别都计算一次类条件密度，然后进行比较。当 $n$ 较大时，标准加速方法（基于树的方法、快速高斯变换或低秩核近似）可以降低这一开销；但在 $r$ 和 $n$ 适中的典型情形下，朴素计算通常已足够实用。

总体开销总结。在稀疏情形（独热编码块）下，主要开销为

$$\boxed{\text{Training: } O(nq) \;+\; O(dk^2 + k^3) \;+\; O(nqr), \qquad \text{Query: } O(qr) \;+\; O(nr).}$$

谱步骤 $O(dk^2 + k^3)$ 通常是瓶颈，但其随 $k$ 而非 $d^2$ 增长，因此对于 $k \ll d$ 仍然高效。

**备注 7.3**（可扩展性情形）。当 $k$ 适中而$d$ 非常大时，该方法尤其具有吸引力；这在高基数分类学习中很常见。通过 $G = X^\top X$ 实现的低秩路径无需构造任何 $d \times d$ 矩阵，并且利用稀疏性，每次查询都能在 $O$（$qr$）时间内完成嵌入阶段，因此整个流程适用于大规模独热调查表示。

**备注 7.4**（切换到类别归一化算子）。第 *5* 节中的类别归一化算子 $\rho_{\mathrm{CN}}$可以用相同的渐近复杂度特征计算，因为它只在构造格拉姆矩阵之前修改列缩放／归一化。因此，在 $\rho_{\mathcal{D}}$ 和 $\rho_{\mathrm{CN}}$ 之间切换不会改变谱阶段的渐近计算成本。

## 8. 合成实验

本节在专门设计的合成数据集上评估所提方法，以检验与分类学习相关的受控现象。目标并非优化特定应用领域上的基准性能，而是刻画模型在稀疏性增加、高基数、无关变量和类别不平衡等情况下的表现。

### *8.1.* 合成生成器：分类变量块

我们考虑 $q$ 个分类变量，其模态大小为 $m_1$、.....、$m_q$，总维度为 $d = \sum_i m_i$。生成样本时，会在每个块中选择一个模态。对于每个类别 $c \in \{1, \ldots, k\}$ 和每个变量 $i$，我们定义一个分类分布 $p_{i,\,c} \in \Delta^{m_i-1}$，其中 $\Delta^{m_i-1}$表示单纯形。类别 $c$中的样本通过按照 $p_{i,\,c}$ 独立抽取模态，并将相应的独热向量拼接起来生成。

我们通过调整所选信息变量的$(p_{i,c})_c$ 之间的散度来控制类别可分性。

*8.2.* 实验 *S1*：受控分离

本实验在受控的类别分离机制下评估所提出的嵌入分类流程的行为。回顾可知，对于每个分类块 $i \in \{1,\dots,q\}$和类别 $c \in \{1,\dots,k\}$，我们定义分类分布$p_{i,c} \in \Delta^{m_i-1}$。选取信息变量的一个子集 $I \subset \{1,\dots,q\}$，并且对于 $i \in I$，类条件分布特征$(p_{i,c})_{c=1}^{k}$通过分离参数 $\delta \in [0,1]$变得越来越不同。对于非信息块 $i \notin I$，我们令$p_{i,c}$ 在各类别间相同，从而使这些变量不携带判别信号。

定量结果汇总于表 1，其中报告了所有方法随分离参数 $\delta$变化的准确率和宏平均 F1。如下文所述，该表清晰展示了所有有竞争力的基线方法从 $\delta = 0$ 处的随机水平性能过渡到$\delta \geq 0.6$ 处近乎完美的分类，同时突出了在中等分离区间内（尤其是在 $\delta = 0.2$处）DMM+核密度估计相较于 PCA+K 近邻的优势。

表 1：实验 S1（受控分离）：准确率和宏平均 F1 随分离参数变化的简要汇总 $\delta$. 每个条目均以准确率 / 宏平均 F1 的形式报告。

| $\delta$ | DMM+核密度估计 | 主成分分析+K 近邻 | LDA | RF |
|---|---|---|---|---|
| 0.0 | 0.344 / 0.311 | 0.345 / 0.342 | 0.335 / 0.333 | 0.328 / 0.326 |
| 0.2 | 0.586 / 0.585 | 0.534 / 0.532 | 0.578 / 0.578 | 0.586 / 0.586 |
| 0.4 | 0.790 / 0.790 | 0.777 / 0.776 | 0.790 / 0.790 | 0.791 / 0.791 |
| 0.6 | 0.919 / 0.920 | 0.919 / 0.919 | 0.920 / 0.921 | 0.920 / 0.920 |
| 0.8 | 0.983 / 0.983 | 0.981 / 0.981 | 0.982 / 0.983 | 0.983 / 0.983 |
| 1.0 | 1.0 / 1.0 | 1.0 / 1.0 | 0.339 / 0.337 | 1.0 / 1.0 |

表 1 展示了随着 $\delta$ 增加，DMM+核密度估计、PCA+K 近邻、线性判别分析以及随机森林（RF）基线方法的准确率和宏平均 F1。在 $\delta = 0$ 时，信息块上的所有类条件分布都相同，因此 $X$ 在统计上独立于 $Y$，任何分类器的表现都无法优于随机猜测。结果证实了这一点：所有方法的准确率都接近 $1/k$ （此处约为 0.33），宏平均 F1 基本与准确率相当，这符合平衡多分类设置下的预期。

随着 $\delta$ 增加，信息块逐渐呈现出更强的类别特异性，所有方法的准确率和宏平均 F1 均持续提升。在中等分离度区间（$\delta = 0.2$ 和 $\delta = 0.4$），DMM+核密度估计始终具有竞争力，在 $\delta = 0.2$ 时优于 PCA+K 近邻，并在 $\delta = 0.4$ 时达到与最强基线方法（随机森林和线性判别分析）相当的表现。在高分离区间（$\delta \geq 0.6$），除线性判别分析外，所有方法都接近性能上限，实现近乎完美的分类（$\approx 0.92$，在 $\delta = 0.6$ 时；$\approx 0.98$，在 $\delta = 0.8$时），最终 DMM+核密度估计、PCA+K 近邻和随机森林在 $\delta = 1.0$ 时均达到完美准确率。

DMM+KDE 的主要优势体现在中间区间：此时存在可分性，但并不显著。在这种情况下，密度矩阵算子通过振幅提升$\sqrt{\cdot}$聚合类别条件频率结构，从而形成第 5.2 节讨论的海灵格/巴塔恰里亚几何。由此得到的谱坐标会放大稳定且类别一致的类别模式，同时舍弃干扰自由度，这解释了其相较于 $\delta = 0.2$ 处 PCA+KNN 所观察到的性能提升。随着可分性增强，分类问题会变得几乎不言自明，所有灵活基线方法的性能也都趋于饱和。

尽管线性判别分析在 $\delta = 0.8$ 之前呈现出总体提升趋势，但在这一合成设置中，其性能在 $\delta = 1.0$ 时骤降至随机猜测水平。这一现象与以下事实相符：在高维独热表示中，类别分离度极高时，类内协方差估计可能会变得病态或实际上退化，从而导致高斯线性判别方法不稳定，除非引入额外的正则化（例如收缩估计量）。相比之下，DMM+KDE 和随机森林在这一情形下仍保持稳定，因为它们不依赖于对环境空间中的完整协方差算子求逆。

总体而言，实验 S1 验证了所提方法的预期表现：(i) 在不存在信号时，性能会降至随机水平（$\delta = 0$）；(ii) 随着分离度提高，性能平稳提升，并且与强基线方法相比仍具竞争力；(iii) 在完全可分情形下实现完美分类（$\delta = 1$），同时保持稳定的宏平均 F1，与类别均衡性能相符。

### *8.3.* 实验 *S2*：高基数与稀疏性

实验 S2 评估在样本量 $n$固定的情况下，增加类别基数对分类性能的影响。增加模态大小（$m_i$）会扩大环境独热维度 $d = \sum_{i=1}^{q} m_i$，使表示逐渐变得稀疏（每个样本仍恰好包含 $q$ 个激活项）。表 2 报告了准确率和宏平均 F1 随缩放因子及所得维度 $d$的变化。

表 2：实验 S2（高基数与稀疏性）：随缩放因子及所得环境维度 $d$变化的简要汇总。各项数值按准确率 / 宏平均 F1 的格式报告。

| 规模 | $d$ | DMM+核密度估计 | 主成分分析+K 近邻 | RF |
|---|---|---|---|---|
| 1 | 75 | 0.979167 / 0.979039 | 0.978000 / 0.977877 | 0.978000 / 0.977933 |
| 2 | 100 | 0.977333 / 0.977202 | 0.978667 / 0.978554 | 0.976833 / 0.976741 |
| 4 | 150 | 0.977833 / 0.977701 | 0.978833 / 0.978718 | 0.976500 / 0.976411 |
| 8 | 250 | 0.977500 / 0.977350 | 0.978000 / 0.977876 | 0.976167 / 0.976069 |

在所考察的范围内（$d$ 从 75 到 250），所有方法均保持较高的性能，准确率和宏平均 F1 均接近 0.98。值得注意的是，随着 $d$ 增大，DMM+KDE 的变化幅度非常小，这证实了谱嵌入在环境维度不断增加时仍具有稳健性。这种表现与密度算子的内在低秩结构（命题 3.6）一致：尽管独热空间不断扩展，它仍限制了嵌入的有效自由度。

在此范围内，PCA+KNN 同样保持稳定，这表明即使稀疏性增加，所生成的低维表示仍保留了足够的类别判别结构。相比之下，随着 $d$ 增大，RF 的性能略有单调下降，这与基于树的学习器在极度稀疏的高维独热设置中具有较高方差和特征选择负担相符。最后，在所有配置中，宏平均 F1 都与准确率紧密同步，表明在整个参数扫描范围内，各类别上的表现均衡。

### *8.4.* 实验 *S3*：无关变量与噪声鲁棒性

在实验 S3 中，我们通过向调查空间添加 $q_{\text{noise}}$ 个额外的分类块来评估模型对干扰维度的鲁棒性，这些分类块在各类别中的分布相同。换言之，这些额外变量不包含任何标签信息，仅作为无关协变量。我们将噪声水平参数化为 $\alpha := q_{\text{noise}}/q$，其中 $q$ 是原始调查块的数量。表 3 报告了 DMM+KDE 和两种具有代表性的基线方法的准确率和宏平均 F1。

总体而言，在测试的各类噪声条件下，性能基本保持不变。尤其是，即使噪声块的数量与原始块相当或更多（最多达到 $\alpha = 2.0$），DMM+核密度估计的准确率和宏平均 F1 仍然很高（约为 0。98）。PCA+K 近邻和随机森林也呈现出类似表现，这表明在这一受控生成器下，集中在信息块上的判别信号并未因添加类别无关的分类噪声而显著减弱。

表 3：实验 S3（无关变量和噪声鲁棒性）：随着无关分类变量块数量增加时的准确率和宏平均 F1。每个条目按准确率 / 宏平均 F1 的顺序报告。

| $\alpha$ | $q_{\text{noise}}$ | DMM+核密度估计 | PCA+K 近邻 | RF |
|---|---|---|---|---|
| 0.0 | 0 | 0.983833 / 0.983898 | 0.984667 / 0.984712 | 0.983000 / 0.983022 |
| 0.5 | 8 | 0.983167 / 0.983246 | 0.982333 / 0.982399 | 0.982833 / 0.982909 |
| 1.0 | 15 | 0.979000 / 0.979101 | 0.980000 / 0.980090 | 0.980000 / 0.980060 |
| 2.0 | 30 | 0.983667 / 0.983700 | 0.985833 / 0.985864 | 0.986167 / 0.986194 |

### *8.5.* 实验 *S4*：*DMM*决策层中的类别不平衡

实验 S4 仅专门 考察所提出的 DMM 流程，其设计目的是通过消融实验阐明类别不平衡在 *DMM* 内部 的作用，而非与外部基线方法进行比较。具体而言，我们固定 DMM 谱嵌入，仅通过比较纯粹的最大似然（ML）规则与先验加权最大后验（MAP）规则，来考察二分类先验不平衡程度逐渐增加时的表现。我们报告平衡准确率和宏平均 F1，因为这两项指标都侧重于少数类性能，因此对类别不平衡较为敏感。

表4凸显了一个显著的转变：ML决策规则始终高度稳定，即使在严重类别不平衡（最高达 $95/5$）的情况下，平衡准确率仍接近1；而随着先验分布越来越偏斜，最大后验概率规则的表现迅速恶化，最终在极端的 $95/5$ 情形下，平衡准确率降至近乎随机的水平（$\approx 0.5$）。这一现象符合预期：在抽样分布严重不平衡时，最大后验概率规则在最小化整体误分类风险方面是最优的，但它往往偏向多数类，并可能严重损害少数类召回率，而这正是平衡准确率和宏平均F1所反映的。最后，在本实验中，基于计数的DMM算子与类别归一化DMM算子产生相同的嵌入和结果；为清晰起见，表4报告基于计数的结果。

表4：实验 S4（仅 DMM，类别不平衡）：不同二分类先验下 ML 与 MAP 决策规则的平衡准确率与宏平均 F1。每个条目以平衡准确率 / 宏平均 F1 的形式报告。

| $p_{\text{major}}$ | $p_{\text{minor}}$ | DMM（ML） | DMM（MAP） |
|---|---|---|---|
| 0.50 | 0.50 | 0.995895 / 0.995888 | 0.995094 / 0.995109 |
| 0.80 | 0.20 | 0.994351 / 0.991741 | 0.930661 / 0.953848 |
| 0.90 | 0.10 | 0.993910 / 0.986778 | 0.788744 / 0.853848 |
| 0.95 | 0.05 | 0.997170 / 0.982675 | 0.506990 / 0.500835 |

**备注8.1**（可复现性）。 所有合成生成器均已完整说明，参数为（$q, m_i, k, n$），类条件单纯形为（$p_{i,\ c}$），并可在代码仓库中获取，详见第*10*节“数据可用性”。

这使得能够精确复现并系统地遍历机制参数。

## 9. 讨论与局限性

密度矩阵构造为独热调查空间中的监督式分类学习提供了一种有原则的算子视角。内在秩界 rank($\rho$) $\leq k$ 由此给出了一个随类别数增长的潜在维度，这表明即使在环境维度极高的情况下，也能获得稳定嵌入。

仍存在若干局限和有待探索的方向。首先，平方根提升 $X = \sqrt{F}$ 是一种建模选择，旨在确保使用 Hellinger距离和 Bhattacharyya亲和度；可以探索其他单调变换，并在

生成假设下为其提供理论保证。其次，核密度估计的性能取决于带宽选择，尤其是在类别严重不平衡或存在多模态潜在簇的情况下。第三，当前构造在每个类别内对频数进行线性聚合；纳入成对共现或交互特征的扩展，或许能进一步提升表达能力，但会增加计算成本。

最后，本文侧重于合成数据验证，以突出方法的表现特性；后续研究可系统考察其在各领域真实应用中的表现，例如调查分析、基于事件的预测以及离散化混合类型问题。

## 10. 结论与未来工作

我们提出了一种针对分类数据的监督式降维与分类框架，该框架基于由类别条件频率诱导的类密度矩阵算子。该模型生成一个半正定、单位迹矩阵，其内在秩受类别数量界定，从而能够构建紧凑的谱嵌入。我们通过类别条件核密度估计和最大似然预测，得到一个潜在空间概率分类器。结构不变性与复杂度估计阐明了该方法在数学上的稳健性及计算优势。

未来工作将研究逐元素平方根以外的变换变体，引入感知交互的算子，并针对分类块的显式生成模型建立理论保证。广泛的应用驱动型研究（包括领域特定数据集和可解释性分析）是顺理成章的下一步，最好通过聚焦于应用的研究成果加以探索。

## 作者贡献

概念构思：所有作者；方法：所有作者；软件：所有作者；验证：所有作者；正式分析：所有作者；调查：所有作者；资源：所有作者；数据整理：所有作者；论文初稿撰写：所有作者；论文审阅与编辑：所有作者；可视化：所有作者；监督：A.F；项目管理：A.F；经费获取：A.F。所有作者均已阅读并同意发表该手稿的最终版本。

## 经费

本研究获得西班牙瓦伦西亚自治区政府资助，资助编号为 COMCUANTICA/007；并获得西班牙 CEU卡德纳尔·埃雷拉大学资助，资助编号为 INDI24/17。此外，R. Bosch-Romeu 和 A. Falco 感谢西班牙科学、创新与大学部通过资助项目 大学教师培养资助 *FPU 2022* （FPU 22/02810）提供的支持。此外，本研究还得到 TED2021-129347B–C21 项目的支持，该项目由 MCIN/AEI/10.13039/501100011033 和 “ 欧盟下一代欧盟计划 / 复苏、转型与韧性计划 ” 资助。本研究还获得了富布赖特计划的资助，该计划由美国国务院教育与文化事务局赞助，并由国际教育协会管理。

## 数据可用性

支撑文中报告结果的代码和合成实验方案已在 GitHub 上公开，并存档于 Zenodo：

- 代码仓库：https://github.com/afalco/dmm-synthetic-experiments.
- Zenodo（v1.0.0）：https://doi.org/10.5281/zenodo.18299399.

该存档包含用于生成实验 S1–S4 的 Jupyter 笔记本，以及可复现的 conda 环境规范（environment.yml）。

## 声明 写作过程中生成式 AI 和 AI 辅助技术的应用 s

在准备本文期间，作者使用 Claude Sonnet 4.5 改进稿件的可读性和语言表达。使用该工具/服务后，作者对内容进行了审阅，并根据需要进行了编辑；作者对已发表文章的内容承担全部责任。

## 利益冲突

作者声明不存在利益冲突。

## 参考文献

[1] F. J. Aherne、N. A. Thacker 和 P. I. Rockett。Bhattacharyya度量作为频率编码数据的绝对相似性度量。*Kybernetika*，34(4):363–368，1998。[2] M. Belkin 和 P. Niyogi。用于降维和数据表示的拉普拉斯特征映射。*Neural Computation*，15(6):1373–1396，2003。[3] A. Bhattacharyya。关于由概率分布定义的两个统计总体之间散度的度量。*Bulletin of the Calcutta Mathematical Society*，35:99–109，1943。[4] G. Strang。线性代数及其应用。第4版，Brooks/Cole，Belmont，CA，2006。[5] S. Boucheron、G. Lugosi 和 P. Massart。集中不等式：独立性的非渐近理论。牛津大学出版社，Oxford，2013。[6] C. Davis 和 W. M. Kahan。扰动引起的特征向量旋转。III。*SIAMJournal on Numerical Analysis*，7(1):1–46，1970。[7] Hervé Abdi。判别对应分析。载于测量与统计百科全书。Sage，Thousand Oaks，CA，2007。DOI：10.4135/9781412952644。[8] M. J. Greenacre。对应分析理论与应用。Academic Press，1984。[9] W. Hoeffding。关于有界随机变量之和的概率不等式。*Journal ofthe American Statistical Association*，58(301):13–30，1963。[10] A. S. Holevo。量子理论的统计结构。Springer，2001。[11] A. Y. Ng、M. I. Jordan 和 Y. Weiss。关于谱聚类：分析与算法。载于神经信息处理系统进展（*NeurIPS*），2002。[12] M. A. Nielsen 和 I. L. Chuang。量子计算与量子信息。剑桥大学出版社，2000。[13] E. Parzen。关于概率密度函数及众数的估计。*The Annals ofMathematical Statistics*，33(3):1065–1076，1962。[14] M. Rosenblatt。关于密度函数若干非参数估计的评述。*TheAnnals of Mathematical Statistics*，27(3):832–837，1956。[15] B. Schölkopf、A. J. Smola 和 K.-R. Müller。作为核特征值问题的非线性分量分析。*Neural Computation*，10(5):1299–1319，1998。

[16] D. W. Scott。多元密度估计：理论、实践与可视化。第2版；约翰·威利父子，2015。[17] B. W. Silverman。统计学与数据分析的密度估计。Chapman and Hall，1986。[18] M. P. Wand 和 M. C. Jones。核平滑。Chapman and Hall/CRC，1995。[19] Y. Yu、T. Wang 和 R. J. Samworth。适用于统计学家的 Davis–Kahan定理的一个实用变体。*Biometrika*, 102(2):315–323，2015。[20] Rudin，Walter。数学分析原理 麦格劳-希尔，第3版，1976。978-0070542358