arXiv:2304.05053v1 [stat.ME] 11 Apr 2023

基于变换傅里叶-沃尔什对角化的二元超立方体密度估计

预印本

Arthur C. Campello斯坦福大学
应用物理系，加利福尼亚州斯坦福市
94305
arthur.campello@stanford.edu

2023年4月12日

摘要

本文聚焦于如何根据多元二元数据，估计高维二元超立方体上的分布元素。针对这一问题，一种常见方法是优化沃尔什基系数；将其改用“傅里叶-沃尔什”对角化表示后，可提升其可解释性。对所得矩阵元素进行单调变换，可得到一种用途广泛的二元密度估计器，这是本文的主要贡献。本文证明，艾奇逊-艾特肯核可由该估计器的一种受约束指数形式导出；放宽这些约束，则可得到该核的灵活可变权重版本，同时仍保持正定性。在这一统一框架下的估计量能够良好地组合，覆盖速度与灵活性权衡的两端，因此可应用于广泛的统计推断与学习问题。

关键词 密度估计 · 二元超立方体 · 沃尔什基 · Aitchison 和 Aitken 核 · 对角化 · 正定核

1 引言

尽管不如连续数据直观，高维二元数据在现代统计学习和人工智能中无处不在。在医学领域，从眼部特征 [1]到药物试验数据，再到基因表达 [2]，许多关键信息都以二元形式呈现。其他常见的二元数据类型包括社交网络和流行病学网络中的连接关系与节点激活、调查数据，以及多变量二元时间序列 [3]。在机器学习和人工智能（AI）应用中，二元数据通常包含从更复杂数据集中提取的重要特征信息。二元特征长期以来一直用于词图像检索 [4]，也用于面部识别 [5] 和跌倒检测 [6]的学习模型。从二元数据中提取洞见的进展，将推动直接处理二元信息和中间形式二元信息的两个领域发展。

本文讨论如何根据 n维二元变量的独立观测值，估计其支持集为 $\{-1,1\}^n$ 超立方体的多变量二元密度。这些密度揭示变量之间的复杂依赖关系，并提供条件概率，从而直接服务于统计学习应用。尽管对于较大的 n，估计超立方体上的全部2^n 概率的成本会高得难以承受，但较小的子集仍能为数据集提供有力洞见。例如，仅根据两个密度元素，就可以求出在多个回归变量给定特定输入条件下，二元响应变量的期望值。

在实际应用中，可以通过三个指标评估密度估计方案的有效性：速度、可解释性和灵活性。尽管这些特性往往需要权衡，实用的估计量仍应能在合理时间内运行，以易于理解的方式进行推断，并能良好适应规模、维度和稀疏程度各异的数据集。具备这些特性的二元密度估计早期成果之一，是 Aitchison 和 Aitken 提出的 AA 核 [1]。AA 核通过对两个二元向量中对应位置相同的数量进行缩放并取幂，来衡量它们之间的接近程度。对于维度为 n 、观测数为 N 的情形，该方法仅需 $O(nN)$ 时间来

估计一个密度分量。此外，AA 核函数已被证明是定义在二元空间上的正定核，这意味着它是一种再生核希尔伯特空间（RKHS）方法 [7]，因此可以利用表示定理。研究表明，该估计量在稀疏数据上表现良好 [8], ，但灵活性有限，因为所有超立方体维度共用一个平滑参数。当密度对某些维度上的相同程度比对其他维度更为敏感时，这种僵化会削弱估计量的表现。
另一种二元密度估计方法采用正交函数（通常为沃尔什函数）的加权和来寻找密度分量 [9, 10]。该方法通过估计密度函数傅里叶–沃尔什展开的系数来进行估计。由于在 n 维空间中存在 2^n 个函数，当 n 较大时，优化每个沃尔什系数便不再可行。为了实现该方法的高灵活性，可以稀疏地或分组估计系数。后一种方法的一个典型例子是递归分块阈值法，它能在满足特定稀疏性条件时，于概率多项式时间内找到系数 [11]。即使在这种权衡的极端情况下，傅里叶-沃尔什方法所需的计算量仍高于 AA 核估计量，但由于其参数数量更多，因而更为灵活。

尽管傅里叶-沃尔什估计量与 AA 核估计量看起来截然不同，本文表明，对前者的一种受限形式进行简单变换即可得到后者。本文进一步推广了这类变换，并保证其归一化，从而得到一种强大的二元密度估计器；通过不同的参数化，该估计器可处于速度与灵活性权衡关系中的不同位置。所得估计器采用一种可解释的核，通过单调函数作用于若干 2^n 个变量乘积的带符号加权和，来衡量两个长度为n的二元向量之间的相似性。傅里叶-沃尔什和AA核估计量的实际应用方式也变得更加明晰。通过将傅里叶-沃尔什系数透明地匹配到相应的变量乘积，该构造使采用可解释性优先方法来确定沃尔什系数优化的优先级成为可能。对于AA核，一般估计量的形式阐明了将该方法扩展为更灵活的维度加权形式，同时不会损害其归一化或正定性。

本文结构如下：首先，文章直观地推导自然排序的沃尔什矩阵，说明它们如何将超立方体上的概率（映射为长度为 2^n 的概率向量）转换为二元变量及其乘积的期望值。文章论证了从数据估计这些期望值时使用收缩系数的合理性，并展示了这些系数与沃尔什系数成正比。由此得到傅里叶-沃尔什估计量的矩阵对角化形式，其中沃尔什系数可解释为特征值。接着，文章介绍了对这一对角化形式进行逐元素单调变换的方法，并利用沃尔什矩阵的特殊性质，确保所得估计量满足归一化。该广义形式使估计量能够采用机器学习中常见的多种激活函数，无论单独使用还是混合使用。随后，文章证明，在沃尔什系数受限时，指数变换对应于一个 AA 核矩阵。放宽对变换前沃尔什系数的约束，则可引入 AA 核的变量加权扩展，同时保留非负性和正定性。此后，文章比较了在一般估计量的不同变体下，计算各类留一交叉验证风险函数所需的时间。最后，文章讨论了一般估计量各变体适用的情形，并展望了本文提出的估计方法所启发的后续研究。

2 估计理论

二元密度估计量以 n 维观测二元数据点作为输入，这些数据点被假定为二元随机变量 $X = (X_1, \ldots, X_n)$的独立同分布样本， $X_i \in \{-1, 1\}$。这些数据用于估计 X分布中的各个元素，该分布由 2^n 个非负概率值组成，且这些概率值之和为一。即使无需计算所有密度元素的估计值，有效的估计量也应确保其完整输出非负且经过归一化。

这个问题的朴素处理方法是，直接根据数据中的相对频率来估计概率。这些估计通常会严重过拟合数据，尤其在高维情况下往往效果不佳。实际应用中的估计量则利用数据，对超出与观测相对应的密度元素进行外推推断。为确保合理的密度外推，可以施加中立性和对称性等典型约束。中立性意味着，新数据对密度估计的影响与现有数据无关。例如，给定两个数据集 D andD'，若一个估计量 $\hat{} f_X^D$with 具有以下性质

$$\hat{f}_X^{D \cup D'} = \frac{1}{|D| + |D'|} \left(|D| \hat{f}_X^D + |D'| \hat{f}_X^{D'} \right) \tag{1}$$

满足中立性。注意到数据集 D 可能只包含一个观测值，因此很明显，这一性质将估计量限制为离散核的形式。对称性约束意味着，对于超立方体 $\{-1, 1\}^n$中的两个不同点，一个点上的观测值对另一点密度估计的影响，应与交换这两个点后相同。本文提出的所有估计量均满足这两项约束。

2.1 矩阵形式的傅里叶-沃尔什估计

一百多年前，Joseph L. Walsh 巧妙地构造了一组完备的正交函数 ϕ_k，用于 $k \in \mathbb{N}$，从而在大小为 2^n，$n \in \mathbb{N}$ [12]的离散空间中得到基函数。具体而言，沃尔什函数可以表示定义在 $x \in \{0,1\}^n$ 上的二元密度 f，形式如下：

$$f(x) = \sum_{k \in \{0,1\}^n} c_k \phi_k(x), \qquad \phi_k(x) = (-1)^{\sum_i x_i k_i}. \tag{2}$$

在这里，系数 c_k 的作用与傅里叶级数中的系数类似，每个系数都编码了与整个分布相关的信息，而非单个元素的信息。这意味着，只需估计所涉及的 2^n 沃尔什系数中的一部分，就能在整个二元超立方体上得到有意义的密度。因此，早期和近期关于二元密度的研究都采用了基于沃尔什基 [11, 9, 10] 的估计方法。本文将这类估计量称为“傅里叶-沃尔什”估计量。

尽管文献中常将沃尔什系数视为内在重要性相同的抽象参数 [11]，但当 X 的各个分量本身具有实际意义时，这些系数也承载着可解释的信息。通过对沃尔什分解进行直观的重新推导可以证明，在使用二元支撑集 $X_i \in \{-1, 1\}$ 时，每个系数都对应于 $\{X_1, \ldots, X_n\}$ 中某个唯一子集内元素的乘积。因此，一些沃尔什系数所表达的含义与输入特征密切对应，而另一些则承载着更为复杂的信息，编码了可能涉及多个特征的乘积。设随机向量$\mathbf{r}_X^{(n)} \in \{-1, 1\}^{2^n}$ 包含$\{X_1, \ldots, X_n\}$的所有 2^n 子集（包括空集）的乘积，其递归生成方式如下：

$$\mathbf{r}_X^{(0)} = [1], \qquad \mathbf{r}_X^{(n+1)} = \begin{bmatrix} \mathbf{r}_X^{(n)} \\ X_{n+1}\mathbf{r}_X^{(n)} \end{bmatrix}. \tag{3}$$

由此构造还可以递归生成矩阵 $W^{(n)} \in \{-1, 1\}^{2^n \times 2^n}$，其列包含$\mathbf{r}_X^{(n)}$ 的支集。其形式为

$$W^{(0)} = [1], \qquad W^{(n+1)} = \begin{bmatrix} W^{(n)} & W^{(n)} \\ W^{(n)} & -W^{(n)} \end{bmatrix} \implies W^{(n)} = \bigotimes_{i=1}^{n} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}. \tag{4}$$

有趣的是，该矩阵是一个自然排序的（或称 Haramard）沃尔什矩阵。它是对称的，并具有重要性质 $W_{(n)}$ $\left[W^{(n)}\right] = \left[W^{(n)}\right]^2 = 2^n I_{2^n}$。根据这些定义可知，$\mathbb{E}\left[\mathbf{r}_X^{(n)}\right] = W^{(n)}\mathbf{p}$，其中

$$p_j = \mathbb{P}\left\{\mathbf{r}_X^{(n)} = W_{:,j}^{(n)}\right\}. \tag{5}$$

注意到$\mathbf{r}_X^{(n)}$ 的构造意味着 $\left[\mathbf{r}_X^{(n)}\right]_{2^{k-1}+1} = X_k$，因此也可以写成

$$p_j = \mathbb{P}\left\{\bigcap_{k=1}^{n}\left(X_k = W_{2^{k-1}+1,j}^{(n)}\right)\right\}. \tag{6}$$

向量 $\mathbf{p}$ 将 X 在超立方体上的概率分布编码为一个向量，并且必须满足约束$\mathbf{1}\,\mathbf{p} = 1$ 和 $\mathbf{p} \succeq \mathbf{0}$。从现在起，$\hat{\mathbf{p}}$ 表示 $\mathbf{p}$ 的估计量。这种向量映射也定义了一个归一化的“计数”向量$\mathbf{p}_k$，用于编码 X 的观测实例。

根据这些定义，$\left[W_{(n)}\mathbf{p}_k\right]_j$编码了乘积 $\left[\mathbf{r}_X^{(n)}\right]_j$ 的样本均值，而 $\left[W_{(n)}\,\hat{\mathbf{p}}\right]_j$则是与二元密度估计相关的$\left[\mathbf{r}_X^{(n)}\right]_j$的期望。目前，令 $\left[W_{(n)}\,\hat{\mathbf{p}}\right]_j$等于 $\left[W_{(n)}\mathbf{p}_k\right]_j$乘以收缩因子 $b_j \in [0, 1]$。这意味着 $W_{(n)}\,\hat{\mathbf{p}} = \mathrm{diag}(\mathbf{b})W_{(n)}\mathbf{p}_k$，其中 $\mathbf{b} \in [0, 1]^{2^n}$ 现在是一个收缩向量，并产生估计量

$$\hat{\mathbf{p}} = \frac{1}{2^n} W^{(n)} \mathrm{diag}(\mathbf{b}) W^{(n)} \mathbf{p}_k. \tag{7}$$

这种形式要求对 $\mathbf{b}$ 施加约束，以确保 $\hat{\mathbf{p}}$ 是真正的密度。重排公式 7 得到 $\mathbf{b} = \mathrm{diag}^{-1}(W_{(n)}\mathbf{p}_k)W_{(n)}\hat{\mathbf{p}}$. 由于该矩阵的第一行显然为 $\mathbf{1}$，此等式意味着 $\mathbf{1}\,\hat{\mathbf{p}} = 1 \iff b_1 = 1$。要使 $\hat{\mathbf{p}} \succeq \mathbf{0}$ 成立，$\mathbf{b}$ 必须是 $\left[\mathrm{diag}^{-1}(W^{(n)}\mathbf{p}_k)W^{(n)}\right]$ 各列的凸组合。从几何上看，这意味着 $\mathbf{b}_{2:2^n}$ 必须位于由 $\left[\mathrm{diag}^{-1}(W^{(n)}\mathbf{p}_k)W^{(n)}\right]$ 去掉第一行后各列所形成的单纯形内。

公式 7 等价于一个傅里叶-沃尔什估计量，其系数与元素 $b_j \left[W^{(n)}\mathbf{p}_k\right]_j$ 成正比。这里，除了前述对 $\mathbf{b}$ 的约束外，约束 $b_j \in [0,1]$ 也有其合理性。所选的二元基$X_i \in \{-1,1\}$意味着 $\left[\mathbf{r}_X^{(n)}\right]_j \in \{-1,1\}$，而 $\left[W^{(n)}\mathbf{p}_k\right]_j$表示其二元样本的均值。假设有一个类似的单变量随机变量 $Y \in \{-1,1\}$，其 $\mathbb{E}[Y] = q$，并观测到 N 个结果 Y，样本均值为$\bar{y}$。如果使用因子 b 将 q 估计为 $\hat{q} = b\bar{y}$，那么在真实值为 q 的条件下，使平方误差最小的 b^* 为

$$b^* = \operatorname*{argmin}_b \mathbb{E}_q\left[\left(b\bar{y} - q\right)^2\right] = q\frac{\mathbb{E}\left[\bar{y}\right]}{\mathbb{E}\left[\bar{y}^2\right]} = \frac{q^2}{\mathbb{E}\left[\bar{y}^2\right]} = \frac{Nq^2}{(N-1)q^2+1}. \tag{8}$$

此公式约束 $b^* \in [0,1]\ \forall\, q \in [-1,1], N \in \mathbb{N}$。正如预期，$b^*(-1) = b^*(1) = 1$ 和 $b^*(0) = 0$；请注意，即使 N 取较大值，最优 b^* 在 q 上也会有显著变化。这种收缩技术与应用统计学中的其他方法类似，例如詹姆斯–斯坦估计量 [13]以及套索回归和岭回归。请注意，$\mathbf{b} = \mathbf{1}$的情形对应于不进行正则化，并将公式 7简化为 $\hat{\mathbf{p}} = \mathbf{p}_k$，即数据频率估计。$\mathbf{b} = [1\ \mathbf{0}\ ]$ 的完全正则化情形得到均匀估计$\hat{\mathbf{p}} = \mathbf{1}/2^n$。后文将讨论如何使用交叉验证在这两个极端之间优化 $\mathbf{b}$。

公式 7 的推导表明，b_j对估计量中使用的 $\left[\mathbf{r}_X^{(n)}\right]_j$ 的期望进行了正则化。这意味着，$\mathbf{b}-$中的 $\binom{n}{k}$ 个元素以及沃尔什系数，都对应于 $\{X_i,\ldots,X_n\}$ 的大小为 $k \in \mathbb{N}$ 的子集内元素的乘积。显然，b_1对应于空集，其乘积为 1，这进一步说明了设定 $b_1 = 1$ 的合理性。因此，在变量 X_i 反映可解释信息的应用中，与 $\{X_i,\ldots,X_n\}$ 的小型子集相关联的$\mathbf{b}$ 元素乘积更具意义。这形成了沃尔什系数的直觉层级，可在 n 较大时为优化选择提供参考。为明确这一层级，定义 $S_k^{(n)} \subset \mathbb{N}$ 为$\mathbf{r}_X^{(n)}$ 中对应于 k 个变量乘积的索引集合。根据 $\mathbf{r}_X^{(n)}$ 的构造，可知

$$S_0^{(n)} = \{1\}, \qquad S_k^{(n+1)} = S_k^{(n)} \cap \left\{x + 2^n \middle|\, x \in S_{k-1}^{(n)}\right\}. \tag{9}$$

注意，$S_1^{(n)} = \{2^{x-1}+1 \mid x \in [n]\}$与公式 6 中 $W^{(n)}$ 个元素的索引相匹配。以直觉为先的优化$\mathbf{b}$ 方法规定，应优先考虑低 k 的集合 $S_k^{(n)}$ 中的索引。例如，如果认为 $\{X_i,\ldots,X_n\}$ 中包含三个以上变量的任何组合都无法解释，那么对于一个可解释的估计量，只需优化 2^n 个元素中的（n^3+5n+6）$/6$。

2.2 傅里叶-沃尔什对角化元素的单调变换

从本质上说，矩阵形式的傅里叶-沃尔什估计量定义了 $\{-1,1\}^n$超立方体上两点之间的相似度量。对于被分配到$\mathbf{p}$中索引 i 和 j 的超立方体点，公式 7 给出了核 $K_{ij} = \mathbf{b}\,(W^{(n)}{}_{:i} \odot W^{(n)}{}_{:j})/2^n$。这个直观的核对$\mathbf{b}$ 的元素求和，并根据两个输入点是否共享相同的 $\left[\mathbf{r}_X^{(n)}\right]_k$，对 b_k 乘以因子（±1）；然后再将其除以元素数量。在许多情况下，这个核的单调变换能够保留其可解释的排序，并对其进行增强。例如，使用非负函数变换元素 K_{ij}，可以保证密度估计为非负，而无需对$\mathbf{b}$施加任何限制。此外，某些变换可以使核成为正定核，而无需像公式 7中那样要求$\mathbf{b} \succ \mathbf{0}$。

使用此类经过变换的傅里叶-沃尔什矩阵估计二元密度，需要保证归一化。本文证明了一个关于傅里叶-沃尔什矩阵的特有事实，以便利用经过单调变换的元素，对公式 7 进行归一化推广。由于沃尔什函数的乘积本身也是沃尔什函数，因此沃尔什矩阵的列和行在逐元素乘法下必然封闭。基于这一事实，假设映射矩阵 $\mathcal{M}^{(n)} \in \mathbb{N}^{2^n \times 2^n}$，其中 $W^{(n)}{}_{:\,\mathcal{M}_{ij}^{(n)}} = W^{(n)}{}_{:\,i} \odot W^{(n)}{}_{:\,j}$。

引理2.1。 $\mathcal{M}^{(n)}$ 存在，且其每一行和每一列在 $[2^n] \subset \mathbb{N}$中均包含唯一元素。

证明。 在基本情形 $n = 0$ 中，显然有 $W^{(0)} = [1] \implies \mathcal{M}^{(0)} = [1]$。根据 $W^{(n)}$ 的递归构造，对于索引 i, $j \in \{1,\ldots,2^n\}$，以下结论成立：

$$W_{:,i}^{(n)} \odot W_{:,j}^{(n)} = W_{:,\mathcal{M}_{ij}^{(n)}}^{(n)} \implies \begin{cases} W_{:,i}^{(n+1)} \odot W_{:,j}^{(n+1)} = W_{:,2^n+i}^{(n+1)} \odot W_{:,2^n+j}^{(n+1)} = W_{:,\mathcal{M}_{ij}^{(n)}}^{(n+1)} \\ W_{:,i}^{(n+1)} \odot W_{:,2^n+j}^{(n+1)} = W_{:,2^n+i}^{(n+1)} \odot W_{:,j}^{(n+1)} = W_{:,2^n+\mathcal{M}_{ij}^{(n)}}^{(n+1)} \end{cases}. \tag{10}$$

因此，

$$\mathcal{M}^{(n+1)} = \begin{bmatrix} \mathcal{M}^{(n)} & 2^n J_{2^n} + \mathcal{M}^{(n)} \\ 2^n J_{2^n} + \mathcal{M}^{(n)} & \mathcal{M}^{(n)} \end{bmatrix}, \tag{11}$$

其中 J_m 是 $m \times m$ 全 1 矩阵。由公式 11 可知，若 $\mathcal{M}^{(n)}$ 的每一行和每一列都包含整数 $[2^n]$，则 $\mathcal{M}^{(n+1)}$ 对整数 $[2^{n+1}]$ 也具有相同性质。由于基例 $\mathcal{M}^{(0)} = [1]$ 满足这一条件，因此该结论对所有 $\mathcal{M}^{(n)}$ 均成立。证毕。

□

定义映射矩阵 $\mathcal{M}^{(n)}$ 后，可以写成

$$\left[W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right]_{ij} = \sum_{k=1}^{2^n} b_k W^{(n)}_{ik} W^{(n)}_{jk} = \mathbf{b}^\intercal \left(W^{(n)}_{:,i} \odot W^{(n)}_{:,j}\right) = \mathbf{b}^\intercal W^{(n)}_{:,\mathcal{M}^{(n)}_{ij}}. \tag{12}$$

根据引理 2.1，$W^{(n)}$**diag(b)**$W^{(n)}$ 的所有行和列都包含相同元素。对 $W^{(n)}$**diag(b)**$W^{(n)}$ 的逐元素变换也同样适用。记 $(Q)^f$ 表示将 $f\colon \mathbb{R} \to \mathbb{R}$ 逐元素作用于矩阵 Q，则这一事实意味着

$$\left(W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right)^f \mathbf{1} = \left[\left(W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right)^f \mathbf{1}\right]_1 \mathbf{1}. \tag{13}$$

由于 $W_{:,1} = \mathbf{1}$，因此

$$\left[\left(W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right)^f \mathbf{1}\right]_1 = \sum_{j=1}^{2^n} f\left(\sum_{k=1}^{2^n} b_k W^{(n)}_{1k} W^{(n)}_{jk}\right) = \left(\mathbf{b}^\intercal W^{(n)}\right)^f \mathbf{1}. \tag{14}$$

这个归一化因子给出了完整的一般形式估计量

$$\hat{\mathbf{p}} = \frac{\left(W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right)^f}{\left(\mathbf{b}^\intercal W^{(n)}\right)^f \mathbf{1}} \mathbf{p}_k. \tag{15}$$

注意，$f(x) = x$ 和 $b_1 = 1$ 得到 $\left(\mathbf{b}_\intercal W_{(n)}\right)^f \mathbf{1} = \mathbf{b}_\intercal W^{(n)}\mathbf{1} = 2^n$，这与公式 7 一致

.

公式 15 分子中的矩阵元素，其计算时间与**b** 中非零元素的数量成正比，而与 f无关。然而，对于非平凡的**b** 和非线性 f，使用快速沃尔什变换算法 [14], 计算归一化因子通常需要 $O(n2^n)$ 时间，而与**b**中非零元素的数量无关。对于逻辑斯蒂变换和指数变换，若**b** 满足特定限制，则归一化所需时间可大幅缩短，分别降至 $O(1)$ 和 $O(n)$。这些改进见表 1。

定义向量$\mathbf{w} \in [0,1]^n$ 及其对应的$\mathbf{b_w}$，使得$\mathbf{b_w} = \sum_{k=1}^n w_k \hat{e}_{2^{k-1}+1}$；这意味着$\mathbf{b_w}$ 仅在 $S_1^{(n)}$ 中的索引处具有非零元素。由于 $\{W^{(n)}{}_{:k} \mid k \in S_1^{(n)}\}$ 中的所有列向量都是反对称的——其含义是$\mathbf{v} = -$**flip(v)**——因此 $W^{(n)}\mathbf{b_w}$ 也具有反对称性。由此，sigmoid/logistic 函数的性质$f(x) + f(-x) = 1$ 允许我们写出

$$f = \frac{1}{1+\gamma^{-x}} \implies \mathbf{1}^\intercal \left(W^{(n)}\mathbf{b_w}\right)^f = \left(\mathbf{b}_\mathbf{w}^\intercal W^{(n)}\right)^f \mathbf{1} = 2^n/2, \tag{16}$$

$\forall\, \gamma \in \mathbb{R}_+$。这在 $O(1)$ 时间内求得。转而考虑指数情形，可以注意到

$$W^{(n)}\mathbf{b_w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix} \otimes W^{(n-1)}\left[\mathbf{b_w}\right]_{1:2^{n-1}} W^{(n-1)} + w_n \begin{bmatrix} \mathbf{1} \\ -\mathbf{1} \end{bmatrix}. \tag{17}$$

这意味着

$$f(x) = \gamma^x \implies \left(W^{(n)}\mathbf{b_w}\right)^f = \begin{bmatrix} \gamma^{w_n} \\ \gamma^{-w_n} \end{bmatrix} \otimes \left(W^{(n-1)}\left[\mathbf{b_w}\right]_{1:2^{n-1}} W^{(n-1)}\right)^f = \bigotimes_{i=0}^{n-1} \begin{bmatrix} \gamma^{w_{n-i}} \\ \gamma^{-w_{n-i}} \end{bmatrix} \tag{18}$$

最后，

$$f(x) = \gamma^x \implies = \left(\mathbf{b}_\mathbf{w}^\intercal W^{(n)}\right)^f \mathbf{1} = \prod_{i=1}^{n} \left(\gamma^{w_i} + \gamma^{-w_i}\right). \tag{19}$$

公式 19 的计算时间为 $O(n)$。

核变换具有强大的灵活性，使该二元密度估计器能够采用应用机器学习中广泛使用的各种激活函数，包括指数函数、逻辑函数/ sigmoid 函数、阶跃函数、ReLU 函数、tanh 函数、ELU 函数等。所列前四种函数均为非负函数，因此在保证密度估计非负方面尤其有用。具体选择哪种函数，取决于交叉验证的表现、应用场景以及所需的计算速度。

所提出的矩阵表达形式还有一项额外优势：可以使用归一化矩阵的凸组合，从而构成变换核估计量的混合。这样的混合估计量形式如下：

$$\hat{\mathbf{p}} = \left[\sum_{i=1}^{m} c_i \frac{\left(W^{(n)}\mathrm{diag}(\mathbf{b}_i)W^{(n)}\right)^{f_i}}{\left(\mathbf{b}_i^{\intercal}W^{(n)}\right)^{f_i}\mathbf{1}}\right]\mathbf{p}_k, \tag{20}$$

其中 $\sum_{i=1}^{m} c_i = 1$、 $c_i > 0$。

2.3 基于指数傅里叶-沃尔什矩阵的 Aitchison-Aitken 核

本节将说明，在公式 15 中对指数函数施加$\mathbf{b} = \mathbf{b_w}$ 型约束，会得到 AA 核估计量；尽管 AA 方法与 Fourier-Walsh 估计方法看似存在根本差异。对于参数 $\lambda \in [1/2, 1]$，AA 核为两个长度为 n 的二元向量 $\mathbf{x}_i$ 和$\mathbf{x}j$ 提供相似度度量，形式如下

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = \lambda^{n-d(\mathbf{x}_i,\mathbf{x}_j)}(1-\lambda)^{d(\mathbf{x}_i,\mathbf{x}_j)},$$

其中， $d(\mathbf{x}_i,\mathbf{x}j)$表示$\mathbf{x}_i$ 和$\mathbf{x}j$ 不同的元素个数。该核可保证归一化，并确保密度估计非负 [1]。假设$\mathbf{x}_i,\mathbf{x}_j \in \{-1,1\}^n$，则可写为

$$\mathbf{x}_i^{\intercal}\mathbf{x}_j = [n - d(\mathbf{x}_i,\mathbf{x}_j)] - d(\mathbf{x}_i,\mathbf{x}_j) \implies d(\mathbf{x}_i,\mathbf{x}_j) = \frac{n - \mathbf{x}_i^{\intercal}\mathbf{x}_j}{2}, \quad n - d(\mathbf{x}_i,\mathbf{x}_j) = \frac{n + \mathbf{x}_i^{\intercal}\mathbf{x}_j}{2}. \tag{21}$$

此时，AA 核的公式变为

$$K(\mathbf{x}_i, \mathbf{x}_j; \lambda) = \sqrt{\lambda}^{n+\mathbf{x}_i^{\intercal}\mathbf{x}_j}\sqrt{1-\lambda}^{n-\mathbf{x}_i^{\intercal}\mathbf{x}_j} = \sqrt{\lambda(1-\lambda)}^{n}\sqrt{\frac{\lambda}{1-\lambda}}^{\mathbf{x}_i^{\intercal}\mathbf{x}_j}. \tag{22}$$

利用公式 6 中隐含的从二元超立方体到 2^n-向量的映射，可写为$\mathbf{x}_i\mathbf{x}j = \left[W^{(n)}\mathrm{diag}(\mathbf{b_1})W^{(n)}\right]_{ij}$，其中$\mathbf{b_1}$ 在$\mathbf{w} = \mathbf{1}$的情况下对应$\mathbf{b_w}$。利用公式 19，可得

$$f(x) = \sqrt{\frac{\lambda}{1-\lambda}}^{x} \implies \left[\left(\mathbf{b_1}^{\intercal}W^{(n)}\right)^{f}\mathbf{1}\right]^{-1} = \left(\sqrt{\frac{\lambda}{1-\lambda}} + \sqrt{\frac{1-\lambda}{\lambda}}\right)^{-n} = \sqrt{\lambda(1-\lambda)}^{n}, \tag{23}$$

与 AA核在公式 22 中的归一化一致。因此，AA核估计量是 等价于 to

$$\hat{\mathbf{p}}_{\mathrm{AAK}} = \frac{\left(W^{(n)}\mathrm{diag}(\mathbf{b_1})W^{(n)}\right)^{f}}{\left(\mathbf{b_1}^{\intercal}W^{(n)}\right)^{f}\mathbf{1}}\mathbf{p}_k, \qquad f(x) = \sqrt{\frac{\lambda}{1-\lambda}}^{x}, \tag{24}$$

公式 15 中估计量的一种受限形式。

根据这个方程，可以将$\mathbf{b}$ 从$\mathbf{b_1}$ 放宽为$\mathbf{b_w}$，从而得到AA核的可变加权版本。这使得$\mathbf{w}$ 的各个元素能够对超立方体 n 维度上的不同“平滑”程度进行参数化。可变加权核反映了现实情况：某些索引上的两点相似性对整体相似性的影响，比其他索引上的相似性更大。为方便起见，引入重新参数化 $\gamma = \sqrt{\lambda/(1-\lambda)}$。边界 $\lambda \in [1/2, 1]$ 对应于 $\gamma \in [1, \infty)$。令$\mathbf{b_1} \to \mathbf{b_w}$，即可得到加权AA核估计量

$$\hat{\mathbf{p}}_{\mathrm{WAAK}} = \frac{\left(W^{(n)}\mathrm{diag}(\mathbf{b_w})W^{(n)}\right)^{f}}{\left(\mathbf{b_w}^{\intercal}W^{(n)}\right)^{f}\mathbf{1}}\mathbf{p}_k, \qquad f(x) = \gamma^{x}. \tag{25}$$

现在表明，可以将公式25的分子写成克罗内克积级数，从而简化计算，并阐明关于核的正定性的某些性质。注意到乘积$W_{(n-1)}^{\intercal}\mathrm{diag}\left([a\ \mathbf{0}\ ]\right)\ W^{(n-1)}$给出 $aJ_{2^{n-1}}$， $W_{(n)}$的递归形式意味着

$$W^{(n)}\mathrm{diag}(\mathbf{b_w})W^{(n)} = \begin{bmatrix}1 & 1\\ 1 & 1\end{bmatrix} \otimes W^{(n-1)}\mathrm{diag}\left([\mathbf{b_w}]_{1:2^{n-1}}\right)W^{(n-1)} + \begin{bmatrix}1 & -1\\ -1 & 1\end{bmatrix} w_n J_{2^{n-1}}. \tag{26}$$

这意味着

$$f(x)=\gamma^x \implies \left(W^{(n)}\mathrm{diag}(\mathbf{b_w})W^{(n)}\right)^f=\begin{bmatrix}\gamma^{w_n} & \gamma^{-w_n}\\ \gamma^{-w_n} & \gamma^{w_n}\end{bmatrix}\otimes\left(W^{(n-1)}\mathrm{diag}([\mathbf{b_w}]_{1:2^{n-1}})W^{(n-1)}\right)^f. \tag{27}$$

继续递归展开，可将完整的加权AA核估计量写为

$$\hat{\mathbf{p}}_{\mathrm{WAAK}}=\left[\prod_{j=1}^{n}\left(\gamma^{w_j}+\gamma^{-w_j}\right)\right]^{-1}\bigotimes_{i=0}^{n-1}\begin{bmatrix}\gamma^{w_{n-i}} & \gamma^{-w_{n-i}}\\ \gamma^{-w_{n-i}} & \gamma^{w_{n-i}}\end{bmatrix}\mathbf{p}_k. \tag{28}$$

采用这种形式，加权AA 核矩阵的元素可在 $O(n)$ 时间内计算。另请注意，由于对角元素为 γ^w、非对角元素为 γ^{-w} 的 2×2 矩阵具有特征值 $\gamma^w-\gamma^{-w}$ 和 $\gamma^w+\gamma^{-w}$，因此当 $\mathbf{w}\succ\mathbf{0}$ 且 $\gamma>1$ 时，式 28 中克罗内克积内的矩阵均为正定矩阵。由于两个正定矩阵的克罗内克积也为正定矩阵，因此加权AA核估计量使用的是正定核函数。此分析重新推导了AA核为正定核的已知结论 [7]，并将这一性质推广至其可变权重的一般化形式。

3 交叉验证

所提出的一般形式估计量具有灵活性，这主要得益于其参数数量可变且可能较多。这些参数包括 $\mathbf{b}$ 的元素、定义 f 的参数，以及在混合估计量时应用于矩阵的权重 c_i。对这些参数进行实际优化通常需要采用交叉验证方案。本节概述了使用平方误差（SE）和 Kullback-Leibler（KL）损失函数进行交叉验证的常见留一法。重点讨论一般形式估计量不同变体的风险函数评估耗时。

对于所描述的优化，假设存在一个“真实”的概率分布$\mathbf{p}$ 以及由 λ 参数化的估计量 $\hat{}\mathbf{p}_\lambda$。最优的 λ^* 使损失函数 $L(\mathbf{p},\hat{}\mathbf{p}_\lambda)$ 的期望最小化，即风险最小化。SE 和 KL 损失函数分别定义为

$$L_{\mathrm{SE}}(\mathbf{p},\hat{\mathbf{p}}_\lambda)=||\hat{\mathbf{p}}_\lambda-\mathbf{p}||_2^2,\qquad L_{\mathrm{KL}}(\mathbf{p},\hat{\mathbf{p}}_\lambda)=\sum_{j=1}^{2^n}\mathbf{p}_j\log\left(\mathbf{p}_j/\left[\hat{\mathbf{p}}_\lambda\right]_j\right). \tag{29}$$

为实现留一法，令 $K\in[2^n]$ 为一个对观测值索引进行编码的多重集。此外，定义 $\hat{}\mathbf{p}_\lambda^{(k)}$ 为在不使用与 $k\in K$ 对应的观测数据点时所得的估计值。根据无意识统计学家的定律，

$$\mathbb{E}\left[\hat{\mathbf{p}}_\lambda^\intercal\mathbf{p}\right]=\frac{1}{|K|}\sum_{k\in K}\mathbb{E}\left[\left(\hat{\mathbf{p}}_\lambda^{(k)}\right)^\intercal\hat{e}_k\right]\implies\arg\min_\lambda\mathbb{E}\left[L_{\mathrm{SE}}(\mathbf{p},\hat{\mathbf{p}}_\lambda)\right]=\arg\min_\lambda\left[\hat{\mathbf{p}}_\lambda^\intercal\hat{\mathbf{p}}_\lambda-\frac{2}{|K|}\sum_{k\in K}\left(\hat{\mathbf{p}}_\lambda^{(k)}\right)^\intercal\hat{e}_k\right]. \tag{30}$$

对于 KL 损失，交叉验证优化为

$$\arg\min_\lambda\mathbb{E}\left[L_{\mathrm{KL}}(\mathbf{p},\hat{\mathbf{p}}_\lambda)\right]=\arg\max_\lambda\sum_{k\in K}\log\left[\left(\hat{\mathbf{p}}_\lambda^{(k)}\right)^\intercal\hat{e}_k\right], \tag{31}$$

使用已知的KL交叉验证优化器的离散改编版本，用于连续空间上的密度估计[15]。

现在可以施加熟悉的形式 $\hat{}\mathbf{p}_\lambda=Q_\lambda\mathbf{p}_k=Q_\lambda\left[(1/|K|)\sum_{k\in K}\hat{e}_k\right]$，其中 $Q_\lambda\in\mathbb{R}^{2^n\times 2^n}$ 是某个一般的对称估计量矩阵。这意味着，计算 $\left(\hat{}\mathbf{p}_\lambda^{(k)}\right)\hat{e}_k$需要知道 $|K|-1$ 个 Q_λ的元素。因此，计算 KL 风险函数需要计算 $|K|(|K|-1)/2$ 个 Q_λ的元素。在 SE 风险情形下，第二项也需要知道 $|K|(|K|-1)/2$ 个 Q_λ的元素，而第一项则需要计算 $|K|(|K|+1)/2$ 个 Q_λ^2的元素。对于方程 7 和 28 的分子，平方矩阵元素的计算尤其简单，因为

$$\left(W^{(n)}\mathrm{diag}(\mathbf{b})W^{(n)}\right)^2=2^n\left(W^{(n)}\mathrm{diag}(\mathbf{b}\odot\mathbf{b})W^{(n)}\right) \tag{32}$$

and

$$f(x)=\gamma^x\implies\left[\left(W^{(n)}\mathrm{diag}(\mathbf{b_w})W^{(n)}\right)^f\right]^2=\bigotimes_{i=0}^{n-1}\begin{bmatrix}\gamma^{2w_{n-i}}+\gamma^{-2w_{n-i}} & 2\\ 2 & \gamma^{2w_{n-i}}+\gamma^{-2w_{n-i}}\end{bmatrix}, \tag{33}$$

表1：归一化以及求估计量矩阵中一个元素及其平方所需的时间，它们是 n（超立方体维度）和 b（非恒定对角元素的数量）的函数。表中报告了一般形式估计量（公式 15）在各种单调函数及对$\mathbf{b}$的限制下的结果。

估计量限制条件	归一化时间	矩阵元素计算时间	矩阵平方元素计算时间
$f(x) = x, b_1 = 1$	$O(1)$	$O(b)$	$O(b)$
$f(x) = \gamma^x$，$\mathbf{b} = \mathbf{b_w}$	$O(n)$	$O(n)$	$O(n)$
$f(x) = (1+\gamma^{-x})^{-1}$，$\mathbf{b} = \mathbf{b_w}$	$O(1)$	$O(n)$	$O(n2^n)$
无限制	$O\ (n2^n)$	$O(b)$	$O\ (n2^n)$

根据克罗内克混合积性质。

本节最后列出了对由公式 15 的限制条件导出的实用估计量的矩阵元素及平方矩阵元素进行归一化和求值所需的计算时间。表1展示了这些计算时间，其中 b 表示$\mathbf{b}$ 的非零（或非恒定）元素数量；对于列出的、计算时间为 $O(n2^n)$的操作，使用快速沃尔什变换 [14] 。

4 总结与讨论

本文提出了一种强大的二元密度估计器，该估计器基于对傅里叶-沃尔什对角化进行逐元素单调变换。为此，本文首先以对角化的形式直观地重新推导了傅里叶-沃尔什分解。在这种形式下，沃尔什系数与构成各个一元二元变量的唯一乘积相关联。随后，本文揭示了沃尔什矩阵的一项特定性质，该性质使得对由任意逐元素变换函数产生的估计量进行归一化成为可能。接着，本文阐明了在上述过程中采用通用指数变换时如何得到 AA 核，并介绍了一种变量加权核扩展，同时保留其理想性质。最后，本文概述了平方误差和 Kullback-Leibler损失函数的留一交叉验证风险函数的实现，并比较了不同估计量的计算时间。这种新估计器在不同约束条件下具备灵活、高效且易于解释等特点，因此非常适合广泛应用于估计和学习任务。

表1的比较表明，在不同约束条件下，所提出估计量的不同变体在数据科学的不同场景中各有优势。对于最多 $n \approx 20$ 维的问题——即二元输入——计算机可以轻松处理 $O(n2^n)$ 次运算，公式 15 中的估计量也可以采用最一般的形式。在这种情形下，学习可能涉及遍历大量变换函数，并探索由若干不同估计矩阵构成的混合模型。文献指出，约 20 维、拥有 40 个数据点的问题，是应用二元密度估计中的典型案例 [8]。另一极端则是考虑高维情形，维数 n 最高可达 $n \approx 10^4$。在这种情况下，除了所引入的加权AA核和未经变换的傅里叶-沃尔什对角化之外，任何其他方法——即使$\mathbf{b}$ 受到严格限制——都会变得极难处理。

这些极端情况不仅展示了通用形式估计量的高度多样性，也让人在面对学习任务时考虑变量选择的可能性。例如，假设有100个变量，其中 5 个编码响应变量，95个是推断重要性各异的回归变量。可以先采用直接的傅里叶-沃尔什对角化估计器，仅优化$\mathbf{b}$ 在 $S_1^{(95)}$、$S_2^{(95)}$ 和 $S_3^{(95)}$ 中的元素。根据这些优化后的量，可以找出与响应变量相关性最高的20 个二元回归变量，然后仅使用这些二元输入，应用正则化程度较低且限制较少的估计器变体。本文所提出的估计器所实现的这类“变量搜索”方法，能够使其在大规模二元空间上的机器学习领域中发挥强大作用。选择用于估计实际密度的超立方体的方法，可能会成为未来令人兴奋的研究重点。

参考文献

[1] J. Aitchison 和 C. G. G. Aitken。通过核方法进行多元二元判别。*Biometrika*，63(3):413–420，1976。[2] I. Shmulevich 和 W. Zhang。基因表达数据的二元分析与基于优化的归一化。*Bioinformatics*，18(4):555–565，2002。[3] C. Jentsch 和 L. Reichmann。广义二元时间序列模型。*Econometrics*，7(4):47–73，2019。[4] B. Zhang、S. N. Srihari 和 C. Huang。使用二元特征进行词图像检索。载于 *SPIE 5296* 论文集，文档识别与检索*XI*，第5296卷，2003年。

[5] J. Lu、V. E. Liong、X. Zhou 和 J. Zhou。学习用于人脸识别的紧凑型二元人脸描述符。*IEEE*《模式分析与机器智能汇刊》，10(1):2041–2056，2015。[6] M. Saleh 和 R. Le Bouquin Jeannes。使用局部二元特征的高效基于机器学习的跌倒检测算法。载于*2018*年第*26*届欧洲信号处理会议（*EUSIPCO*），2018年。[7] H. Mussa。Aitchison-Aitken 核函数再探。《数学研究杂志》，5(1):22–25，2013。[8] B. Grund 和 P. Hall。论核估计量在高维稀疏二元数据上的表现。《多元分析杂志》，44(2):321–344，1993。[9] W. Liang 和 P. R. Krishnaiah。多元二元密度的非参数迭代估计。《多元分析杂志》，16(1):162–172，1985。[10] X. R. Chen、P. R. Krishnaiah 和 W. Liang。利用正交函数估计多元二元密度。《多元分析杂志》，31(2):178–186，1989。[11] M. Raginsky、J. G. Silva、S. Lazebnik 和 R. Willett。二元超立方体上的密度估计递归过程。《电子统计学杂志》，7:820–858，2013。[12] J. L. Walsh。一组闭合的标准正交函数。《美国数学杂志》，45(1):5–24，1923。[13] W. James 和 C. Stein。二次损失下的估计。伯克利数学统计与概率研讨会，4(1):361–379，1961。[14] M. T. Hamood 和 S. Boussakta。快速沃尔什-哈达玛-傅里叶变换算法。*IEEE*《信号处理汇刊》，59(11):5627 – 5631，2011。[15] P. Hall。论 Kullback-Leibler 损失与密度估计。《统计年刊》，15(4):1491–1519，1987。