基于香农熵的 Kullback–Leibler 散度新估计量

Mehmet Sıddık Çadırcı[1] 和 Martin Singull[2*]

[1]理学院，统计学系，Cumhuriyet大学，Sivas，土耳其[2]数学系，Linköping大学，Linköping，瑞典[*]通讯作者：martin.singull@liu.se

arXiv:2603.06759v1 [math.ST] 6 Mar 2026

摘要

我们研究 Kullback–Leibler（KL）散度的估计，以及拟合优度检验在多元连续分布中的应用。我们的出发点是香农熵的最大熵原理：在所有均值向量和协方差矩阵固定的分布中，多元高斯分布能够唯一地使熵达到最大。因此，矩匹配高斯分布到未知密度的 KL 散度可写为熵差，这是一种适用于衡量与高斯分布偏离程度的信息论度量。

为进行估计，我们采用基于香农熵和 KL 散度的 k近邻（kNN）估计量。这些估计量源自 Kozachenko-Leonenko方法及其后续改进，并具有已确立的一致性和 L^2收敛结果。受此前针对广义高斯模型和 Student型模型中的 Rényi型泛函所提出的基于熵的拟合优度思想启发，我们将一种基于 KL 的检验统计量定义为以下两者之差：(i) 根据样本均值和协方差拟合的高斯模型的熵；(ii) 未知熵与 kNN 估计量之间的 KL 散度。在多元正态性成立时，该统计量收敛于零；在非高斯备择下，则收敛于一个严格为正的界。

针对不同维度和样本量的蒙特卡罗模拟结果表明，所提出的方法能够实现准确的第一类错误控制，并且与传统的多元正态性检验相比，具有准确且总体上更优的功效，尤其是在中等维度和高维情况下。

1 引言

Kullback-Leibler（KL）散度，也称为相对熵，是信息论与统计学中的一项基本度量，用于量化两个概率分布之间的差异。它由Kullback和Leibler提出 [13] ，衡量使用参考模型 g而非真实分布 f时所产生的预期对数似然损失（或过量对数损失）。由于这种决策理论解释，KL 散度自然地出现在基于概率的推断、模型选择、

拟合优度检验、密度比较，以及各种学习和信号处理任务中。

对于两个连续概率密度函数（pdf） f， g， f 到 g 的 KL 散度定义如下： $\mathbb{R}^m \to \mathbb{R}$:

$$D_{KL}(f\|g) = \int_{\mathbb{R}^m} f(x) \log \frac{f(x)}{g(x)}\, dx. \tag{1}$$

由于散度 $D_{KL}(f\|g) \geq 0$，且仅当 $f = g$ 几乎处处成立时才等于零，因此尽管它不具对称性，也不是度量，但仍能以一种“有方向性”的方式衡量分布之间的差异。

KL 散度与香农熵密切相关，后者用于衡量单个分布的不确定性。对于密度 $H(f)$ [22] f，微分香农熵定义为

$$H(f) = -\int_{\mathbb{R}^m} f(x) \log f(x)\, dx, \tag{2}$$

经过简单推导，可得以下结果

$$D_{KL}(f\|g) = -H(f) - \int_{\mathbb{R}^m} f(x) \log g(x)\, dx. \tag{3}$$

等价地， $D_{KL}(f\|g)$定义为交叉熵 $-\mathbf{E}_f[\log\ g(X)]$ 减去熵$H(f)$，因此可视为使用预言机模型 g 代替 f所产生的额外交叉熵（或额外期望对数损失）。

公式（3）也阐明了为何在基于 KL 散度的流程中，某些比较分布尤为自然。在均值和协方差固定的约束条件下，多元高斯分布的香农熵最大；例如参见 [4, 8, 12, 9]。因此，在所有具有相同一阶和二阶矩的分布中，高斯分布是唯一使熵最大化的分布。同样，$D_{KL}(f\|\phi_{\mu,\ \Sigma})$是熵 $H(\phi_{\mu,\ \Sigma})$与 $-\ H(f)$之差，并提供了一种有原则的度量，用于衡量与高斯分布的偏离；其中，$\phi_{\mu,\ \Sigma}$ 是与 f的矩相匹配的适配高斯分布。这一观察为拟合优度检验中的高斯基准提供了依据，也构成了本文所提出的基于 KL 散度流程的基础。

从数据中估计 $D_{KL}(f\|g)$ 对假设检验、密度比较和异常检测等实际应用至关重要。经典方法依赖参数建模，或使用基于直方图和核密度估计器的附加估计量，但这些方法在高维情况下很容易变得不稳定。最近邻（NN）方法提供了一种颇具吸引力的替代方案：它们利用样本云的局部几何结构，而非显式地进行密度重构，并且能够自然地推广到多元情形。熵估计中的最近邻思想可追溯至 Kozachenko 和 Leonenko [11]，此后得到了广泛研究 [7, 14, 20, 19, 1, 18]。至于偏差估计，包括基于 k 近邻的估计器及其一致性性质在内的相关方法，已在 [21, 25, 5, 2]，等文献中得到发展。

本文在这项研究的基础上，聚焦于多元连续情形下的 KL 散度最近邻估计量，并着重探讨拟合优度检验。主要贡献概括如下：

i. 我们通过在均值与协方差约束条件下相对于 KL 散度重新表述最大熵原理，为高斯基准提供了信息论依据。

ii. 然后，在标准正则性与矩条件下，我们回顾并更新了最近邻估计器的 Shannon 熵与 KL 散度的渐近性质（一致性、渐近无偏性以及均方收敛）。

iii. 基于这些估计量，我们构造了一个基于 KL 的检验统计量，并通过大量蒙特卡洛实验考察其有限样本行为，包括经验尺寸、功效，以及维度、样本量和最近邻参数 k 的影响。

本文其余部分安排如下。第 2 节重述最大熵原理，并介绍其在 KL 散度和高斯基准中的应用。第 3 节介绍香农熵的最近邻估计器以及由此得到的 KL 散度估计量。第 4 节提供假设检验的框架。第5节介绍数值实验，第 6节简要讨论可能的扩展，并在最后作出总结。

2 最大熵原理

Jaynes 提出的最大熵原理 [8], 指出，在所有满足特定约束条件（通常为矩条件）的概率分布中，熵最大的分布是信息量最少的选择。由于 KL 散度与香农熵密切相关，最大熵的表达式通常可以改写为相对于适当比较分布的最小 KL 散度。

首先，在标准均值-协方差设定下，对可接受解的类别进行形式化定义是有益的。

定义 2.1（类别 $\mathcal{K}$）。定义在 $\mathbb{R}^m$ 上、具有有限且满足矩约束的密度函数 f 的分布类别 $H(f)$

$$\int_{\mathbb{R}^m} x f(x)\,dx = \mu,$$
$$\int_{\mathbb{R}^m} (x-\mu)(x-\mu)^\top f(x)\,dx = \Sigma,$$

其中，$\mu \in \mathbb{R}^m$ 表示均值向量，Σ 表示大小为 $m \times m$ 的对称正定协方差矩阵。

在类别 $\mathcal{K}$ 中，众所周知，均值为 μ、协方差为 Σ 的多元高斯分布可使香农熵最大化；例如，参见 [4, 12, 9]。具体而言，对于任意 $f \in \mathcal{K}$,

$$H(f) \le \frac{1}{2}\log\left[(2\pi e)^m \det(\Sigma)\right], \tag{4}$$

且当且仅当 f 是高斯密度时，等号成立。

$$\phi_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{m/2}\det(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^\top \Sigma^{-1}(x-\mu)\right). \tag{5}$$

下面的命题用 KL 散度的术语重新表述了这一经典的最大熵结果，并仅作为我们检验程序的基准。

命题 2.1 (KL 散度高斯基准)。 假设 $\mathcal{K}$ 的定义如定义 2.1 所示，且 $\phi_{\mu,\,\Sigma}$ 表示一个高斯密度 (5)，其均值为 μ ，协方差为 Σ。那么，对每个 $f \in \mathcal{K}$,

$$D_{KL}\big(f \,\|\, \phi_{\mu,\Sigma}\big) = H(\phi_{\mu,\Sigma}) - H(f) \;\geq\; 0, \tag{6}$$

当且仅当 $f = \phi_{\mu,\,\Sigma}$ 几乎处处（*a.e.*）成立时，等号成立。

证明。B 利用 KL 散度的表达式和高斯密度的表达式 y,

$$D_{KL}\big(f \,\|\, \phi_{\mu,\Sigma}\big) = \int_{\mathbb{R}^m} f(x) \log \frac{f(x)}{\phi_{\mu,\Sigma}(x)}\, dx = -H(f) - \int_{\mathbb{R}^m} f(x) \log \phi_{\mu,\Sigma}(x)\, dx.$$

直接计算可得，对于任意 $f \in \mathcal{K}$,

$$\int_{\mathbb{R}^m} f(x) \log \phi_{\mu,\Sigma}(x)\, dx = -\frac{1}{2} \log\big((2\pi)^m \det(\Sigma)\big) - \frac{m}{2},$$

因为相应的一阶和二阶矩与 $\phi_{\mu,\,\Sigma}$的矩相匹配。因此 re

$$D_{KL}\big(f \,\|\, \phi_{\mu,\Sigma}\big) = -H(f) + \frac{1}{2} \log\big((2\pi)^m \det(\Sigma)\big) + \frac{m}{2} = H(\phi_{\mu,\Sigma}) - H(f),$$

(4) 是一个非负值。仅当且唯有在以下条件成立时，等号才成立： $H(f) = H(\phi_{\mu,\,\Sigma})$, 因此 f 必须几乎处处与 $\phi_{\mu,\,\Sigma}$ 重合。

□

命题 2.1 表明，最大熵原理可以解释为约束类 $\mathcal{K}$中的最小 KL 散度原理：在所有均值为 μ 且协方差为 Σ 的密度$f \in \mathcal{K}$ 中，高斯ϕ_μ的熵具有唯一最大值. Σ 等价地，KL 散度 $D_{KL}(f\|\phi_{\mu,\,\Sigma})$ 衡量了 f 相对于该最大值的熵差 。具体而言，对于任意 $f \in \mathcal{K}$,

$$D_{KL}\big(f \,\|\, \phi_{\mu,\Sigma}\big) = 0 \quad\Longleftrightarrow\quad f = \phi_{\mu,\Sigma} \text{ a.e.},$$

这证明了可以使用高斯模型作为检验拟合优度和模型验证的基准。

注 2.2。 对于任意密度对 $(f,\ g)$，吉布斯不等式指出 $D_{KL}(f\|g) \geq 0$，且当且仅当 $f = g$ 几乎处处成立时，等号才成立。命题 2.1 用参考分布 $\phi_{\mu,\,\Sigma}$ 表示 KL 散度 $D_{KL}(f\|\phi_{\mu,\,\Sigma})$，该参考分布的选取方式使得散度恰好等于熵差$H(\phi_{\mu,\,\Sigma}) - H(f)$。

多元广义高斯（或指数功效）族也会得出类似的表格。对于各向同性情形，设 f 是 $X \in \mathbb{R}^m$ 在 $\mathbb{R}^m$ 上的密度，并假设对于某个 $s > 0$ ，矩 $\mathbf{E}\|X\|^s$ 是有限的。诸如 [10, 23, 1] 这样的结果会得到如下形式的熵不等式

$$H(f) \leq m \log\big(\alpha(m,s)\, \mathbf{E}\|X\|^s\big),$$

对与 m 和 s相关联的显式常数 $\alpha(m,s)$成立，当且仅当 X 对相应尺度满足各向同性广义高斯分布 $GG_\tau(m,s)$

参数 τ。同样的结论对具有一般位置参数和方差矩阵的多元指数幂分布成立；参见例如 [23, 1]。对于这些类别，相应约束以径向矩而非二阶矩的形式表述。如同命题2.1，这些熵界可以改写为与相应广义高斯基准相关联的 KL 不等式 g_{GG}。对于 f 与 g_{GG} 共享同一约束的情形（例如，相同的$\mathbf{E}\|X\|^s$ 在各向同性介质中），我们有如下结果

$$D_{KL}\big(f \,\|\, g_{GG}\big) = H(g_{GG}) - H(f) \;\geq\; 0,$$

且当且仅当 $f = g_{GG}$ 几乎处处成立时，等号成立。从这一角度来看，也可以采用非高斯最大熵基准构造基于 KL 的检验；不过，本文将注意力限定在经典的高斯均值—协方差情形。

3 香农熵和 KL 散度的最近邻估计量

估计香农熵和 KL 散度是许多非参数推断问题中的关键步骤。对于多元情形，最近邻方法为直方图估计量和基于核的估计量提供了一种实用的替代方案，尤其是在维度 m 适中且显式的多元密度估计已变得不稳定时。

我们回顾香农熵的 k近邻（kNN）估计量，说明其如何导出 KL 散度估计量，并概述构成我们检验程序基础的渐近性质。

3.1 香农熵的最近邻估计

设 X_1，……，X_N是取值于 $\mathbb{R}^m$、密度为 f的独立同分布随机向量，且 Y_1，……，Y_M是取值于 $\mathbb{R}^m$、密度为 g的独立同分布随机向量，并且与 X 样本相互独立。假设 f 和 g 是（Lebesgue）密度，在其支集上几乎处处连续且几乎处处为正，并满足第 3 节在说明渐近结果时所规定的正则性及尾部/矩条件。

香农熵的最近邻估计量由 Kozachenko 和 Leonenko 提出 [11], ，详见 [7, 14], ，该方法基于局部密度的几何方法：样本 X_i 在 X_i 的第 k个最近邻距离较远时会变得稀疏，主密度随之降低；当邻居较近时，主密度则会增加。

我们假设 $\rho_{i,k,N}$ 表示 X_i 在 $\{X_j\}_{j\neq i}$中到其第 k近邻的欧几里得距离。香农熵的 Kozachenko–Leonenko（kNN）估计量的标准形式如下

$$\widehat{H}_{N,k}(f) = \psi(N) - \psi(k) + \log V_m + \frac{m}{N}\sum_{i=1}^{N} \log \rho_{i,k,N}, \tag{7}$$

其中， $V_m = \pi^{m/2}/\Gamma(\frac{m}{2}+1)$ 是 $\mathbb{R}^m$ 中的单位球体积， $\psi(\cdot)$ 是双伽马函数。

该估计量可视为以下陈述的扩展：

$$H(f) = -\mathbf{E}[\log f(X_1)],$$

whe我们将 $f(X_i)$ 替换为由以下内容指导的局部 kNN 密度近似： $\rho_{i,k,N}$.

调整参数 k 在偏差与方差之间取得平衡：较小的 k 会导致较低的偏差，但方差较高；而较大的 k 会增加偏差，从而稳定估计结果。尽管在实践中， k 通常固定为整数（例如 $k \in \{3, 5, 10\}$ ），这与 k 常数渐近理论一致，并且对于我们实验中所考虑的维度而言已足够。

备注 3.1。 一些等价的表述会将常数包含在单个对数中，例如

$$\widehat{H}_{N,k}(f) = \frac{1}{N}\sum_{i=1}^{N} \log\left(\rho_{i,k,N}^{m} V_m e^{\psi(N)-\psi(k)}\right),$$

并使用 $\psi(N) \approx \log(N-1)$ 对于较大的 N。由于这在理论文献中属于标准做法，且能够明确写出有限样本常数，因此我们保留（7）中的 digamma 形式。

3.2 KL 散度的最近邻估计

为估计 $D_{KL}(f\|g)$，一种直接的加法方法是分别估计 f 和 g ，然后使用（1）将这些估计值组合起来。在中等维度或高维情况下，这种策略往往会继承多元密度估计的不稳定性。设 $X_1, \ldots, X_N \sim f$ 和 $Y_1, \ldots, Y_M \sim g$ 为独立样本，并设 $\nu_{i,k}$ 表示从 X_i 到 k 样本中第 Y 个最近邻的欧几里得距离。根据 [21, 14, 25], ，常用的 $D_{KL}(f\|g)$ 的 kNN 估计量 为

$$\widehat{D}_{KL}(f\|g) = \frac{m}{N}\sum_{i=1}^{N} \log\frac{\nu_{i,k}}{\rho_{i,k,N}} + \psi(M) - \psi(N-1). \tag{8}$$

统计量比较每个 X_i 周围的典型邻居距离：邻居分别来自同一分布（通过 $\rho_{i,k,N}$ 获取）和第二个分布（通过 $\nu_{i,k}$ 获取）。当 f 和 g 密切相关时，这些半径相近，估计的偏差较小；系统性差异则会产生正估计值。

直观而言， （8）可由以下恒等式推出

$$D_{KL}(f\|g) = \mathbf{E}[\log f(X_1)] - \mathbf{E}[\log g(X_1)],$$

其中，每个对数密度都由基于相应样本构造的 kNN 近似值替代。在差值中，体积项 V_m 相互抵消，这解释了它为何没有出现在（8）中。对于较大的 M， N 值， $\psi(M) - \psi(N-1) \approx \log\big(M/(N-1)\big)$ 可以使用，但我们保留 digamma 函数形式，以遵循标准的偏差校正方法。

3.3 渐近性质

关于 kNN 熵和散度估计量，已有相当完善的渐近理论；例如，参见 [11, 7, 14, 20, 19, 25, 1, 3]。对于我们的目的，只需给出

在标准光滑性和尾部条件下，针对固定 k 的一个具有代表性的一致性结论。

定理 3.1（k近邻估计量的一致性）。 假设 f 和 g 是定义在$\mathbb{R}^m$ 上的密度，且 H（f） 和 H（g） 有限。假设 f 和 g 几乎处处连续、有上界，并满足以下矩/尾条件：我们假设它们满足如下形式的矩/尾条件：$\mathbf{E}\|X\|^{m+\delta} < \infty$ 和 $\mathbf{E}\|Y\|^{m+\delta} < \infty$，其中 $X \sim f$ 和 $Y \sim g$。令 $k \geq 1$ 为常数。那么，当 $N, M \to \infty$时，

$$\widehat{H}_{N,k}(f) \xrightarrow{a.s.} H(f), \qquad \widehat{D}_{KL}(f\|g) \xrightarrow{a.s.} D_{KL}(f\|g).$$

在相同条件下，这些估计量在均方误差 $\widehat{}$ $\widehat{}$方面也具有一致性，即 $\mathbf{Var}$（$H_{N,\,k}$（f）） $\to 0$和 $\mathbf{Var}$（D_{KL}（$f\|g$）） $\to 0$，且渐近无偏。

上述陈述符合 L^2-一致性框架，该框架由 [20,19]提出。具体而言，Cadirci等人 [3] 在基于熵的检验背景下，采用泊松化技术来估计随机固定 $k \geq 1$ 情形下Kozachenko-Leonenko熵估计量的均方收敛。类似的一组技术结合无偏形式（8），为本文使用的k近邻KL估计量提供了相应均方一致性结论的依据。

注 3.2。 根据支撑是有界还是无界，不同文献给出的充分条件略有差异。对于有界支撑，通常假设支撑内的密度有一个正的下界；对于无界支撑（包括高斯分布及相关分布族），则以关于矩和尾部的条件取代全局下界。在我们的数值实验中，我们选取了定理 3.1 中的条件，使其涵盖所用的分布，并且与所引文献中的标准假设相近。

总之，最近邻方法为多元问题中的香农熵和 KL 散度估计提供了一种简单且有理论依据的途径。这些方法无需显式进行多元密度重构，计算效率高，并构成了本文后续讨论的基于 KL 的拟合优度检验的基础。

4 基于 KL 散度的拟合优度检验

基于 KL 散度，我们对多元正态性进行拟合优度检验。设 $X_1, \ldots, X_N$ 是 $\mathbb{R}^m$ 中密度为 f、均值有限 $\mu = \mathbf{E}[X]$ 且协方差为 $\Sigma = \mathbf{Var}(X)$ 的独立同分布随机向量，均为 $\mathbb{R}^m$ 中的独立同分布随机向量。我们检验复合原假设

H_0： $f \in \mathcal{F}_{\mathcal{N}}$（无条件均值和协方差的 m 变量高斯分布族）

针对一般备择假设 H_1： $f \notin \mathcal{F}_{\mathcal{N}}$。

Let

$$\bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i, \qquad S_N = \frac{1}{N-1}\sum_{i=1}^{N}(X_i - \bar{X}_N)(X_i - \bar{X}_N)^\top$$

分别为样本均值和样本协方差矩阵，并假设 $\phi_{\bar{X}_N,\ S_N}$是具有该均值和协方差的高斯密度。我们回顾命题 2.1：在均值 μ和协方差 Σ固定的分布类 $\mathcal{K}$ 中，多元高斯密度 $\phi_{\mu,\ \Sigma}$ 唯一地使香农熵最大，并满足

$$D_{KL}(f\|\phi_{\mu,\Sigma}) = H(\phi_{\mu,\Sigma}) - H(f) \geq 0,$$

当且仅当 $f = \phi_{\mu,\Sigma}$ 几乎处处成立时取等号，并且

$$H(\phi_{\mu,\Sigma}) = \frac{1}{2}\log\big[(2\pi e)^m \det(\Sigma)\big].$$

基于这一思路，我们定义基于 KL 的检验统计量

$$T^{\mathrm{KL}}_{N,k} := H\big(\phi_{\bar{X}_N,S_N}\big) - \widehat{H}_{N,k}(f) = \frac{1}{2}\log\big[(2\pi e)^m \det(S_N)\big] - \widehat{H}_{N,k}(f), \tag{9}$$

其中 $\widehat{H}_{N,k}(f)$ 是使用第 k 个最近邻得到的香农熵估计量，如式（7）所示。按照构造，$T_{\mathrm{KLN},k}$利用恒等式 $D_{KL}(f\|g) = H(g) - H(f)$，给出了 KL 散度 $D_{KL}(f\|\phi_{\mu,\Sigma})$ 的估计。

定理 4.1。 假设 f 是定义在 $\mathbb{R}^m$ 上的连续密度，具有有限二阶矩，且满足定理 *3.1* 的条件。对于任意固定的 $k \geq 1$,

$$T^{\mathrm{KL}}_{N,k} \xrightarrow{p} H(\phi_{\mu,\Sigma}) - H(f) = D_{KL}(f\|\phi_{\mu,\Sigma}) \quad as\ N \to \infty.$$

特别是,

$$T^{\mathrm{KL}}_{N,k} \xrightarrow{p} 0 \quad if\ X \sim N_m(\mu,\Sigma),$$

并且，对于任何具有相同均值 μ 和协方差Σ的非高斯 $f \in \mathcal{K}$， $T_{\mathrm{KLN},k} \xrightarrow{p} \xi(f) > 0$

证明思路。我们分解

$$T^{\mathrm{KL}}_{N,k} = \underbrace{\big(H(\phi_{\bar{X}_N,S_N}) - H(\phi_{\mu,\Sigma})\big)}_{\text{(I)}} + \underbrace{\big(H(\phi_{\mu,\Sigma}) - H(f)\big)}_{\text{(II)}} + \underbrace{\big(H(f) - \widehat{H}_{N,k}(f)\big)}_{\text{(III)}}.$$

根据大数定律， $S_N \to \Sigma$ 几乎必然成立，因此 (I) $\to 0$，因为$N \to \infty$ 由于 $\mathrm{logdet}(\cdot)$ 在正定矩阵集合上连续。根据命题 2.1，项 (II) 等于 $D_{KL}(f\|\phi_{\mu,\ \Sigma})$ 并且在定理 3.1的假设下依概率趋于零。 □

实际上，无法以闭式形式获得 $T^{\mathrm{KL}}_{N,\ k}$ 的零分布。因此，我们在经过调整的高斯模型下使用参数自助法来校准检验。在 $N_m(X_N,\ S_N)$ 下，我们使用参数自助法进行校准。对于选定的邻域大小 k和显著性水平 $\alpha \in (0,\ 1)$ $X_1^{*(b)}:\ N \sim N_m(\bar{X}_N,\ S_N)$，我们模拟自助样本，计算每个样本的 $T_{\mathrm{KLN},\ k}$，并将所得自助值的 $(1-\alpha)$-分位数作为临界阈值 t_α。当 $T^{\mathrm{KL}}_{N,\ k}{}^{\mathrm{obs}}$ $\geq t_\alpha$ 时，我们拒绝 H_0。

备注 4.1 （KL 散度作为期望对数损失）。 该散度

$$D_{KL}(f\|\phi_{\mu,\Sigma}) = \int_{\mathbb{R}^m} f(x)\log\frac{f(x)}{\phi_{\mu,\Sigma}(x)}\,\mathrm{d}x$$

可解释为采用高斯模型$\phi_{\mu,\ \Sigma}$ 替代真实密度 f所产生的平均对数似然损失。事实上，

$$D_{KL}(f\|\phi_{\mu,\Sigma}) = -\,\mathbf{E}_f[\log\phi_{\mu,\Sigma}(X)] + H(f),$$

这意味着它衡量的是相对于了解 f 的预言机模型而言，交叉熵的超额部分（等价地说，是预期对数损失的增加）。

由于 $T_{N}^{\mathrm{KL}}{}_{,k}$ 的解析零分布难以求得，因此我们根据相关的熵和 KL 拟合优度检验 [24, 6, 15, 17, 16]，在 H_0 （算法 1）下使用参数自助法校准临界值。

算法 1 自助法校准 $T_{N}^{\mathrm{KL}}{}_{k,}$

1：输入：数据 $X_{1:N}$、自助法重复次数 B、显著性水平 α、邻域大小 k。 2：计算 X_N 和 S_N。 3：使用（9）根据 $X_{1:N}$ 计算 T^{KL}、$_{N\,k}{}^{\mathrm{obs}}$ 。4：对于 $b = 1$，....., B，执行 5：模拟 $X_1^{(b)}{}_{:N} \sim N_m(\bar{X}_N, S_N)$。6：计算 $T_{(b)\ N,k} = T_{\mathrm{KL}N,k}(X_1^{(b)}{}_{:N})$。7：结束循环。8：设 t_α 为 $\{T_{N,k}^{(b)}\}_{b=1}^{B}$的（$1-\alpha$）分位数。9：判定：如果 $T^{\mathrm{KL}}{}_{,N}{}^{\mathrm{obs}}$，$_{k}$ $\geq t_\alpha$，则拒绝 H_0。

5 数值实验

本节报告基于 KL 的统计量 $T_{N}^{\mathrm{KL}}{}_{,\ k}$ 其定义见（9）。本节有三个目标。首先，从样本量 N、维度 m和邻域大小 k等方面展示该统计量的有限样本行为。其次，考察其针对结构化非高斯备择的经验功效，包括轻尾和重尾偏离。最后，在实际应用中，我们提供经校准的 5% 临界值。如前所述，蒙特卡洛结果基于独立重复，误差条则基于经验标准差。所有实验的 k近邻计算均采用欧几里得距离。

5.1 对 $T_{N}^{\mathrm{KL}}{}_{,\ k}$ 的蒙特卡洛研究：收敛与稳定性

我们首先考察 $T_{N}^{\mathrm{KL}}{}_{,\ k}$ 的收敛行为。命题 2.1 表明，在多元正态性条件下，总体目标为 $D_{KL}(f\|\phi_{\mu,}$ $_{\Sigma}) = 0$，因此随着 N 增大，该统计量应收敛于零。为评估其在高斯模型之外的稳健性

，我们还模拟了来自广义高斯模型的示例，其形状参数为 s，并注意到 $s = 2$ 对应于该族中的高斯参考值。我们针对每个 $s \in \{1, 2, 4, 8\}$ 和 $m \in \{1, 2, 3\}$分别进行了独立重复。图 1 报告了 $N \in \{1000, \ldots, 5000\}$ 的结果。在高斯情形下（$s = 2$），该统计量趋近于零。对于 $s \neq 2$，该统计量稳定在远离零的位置，这与在固定的非高斯备择下收敛于正的 KL 界相一致。

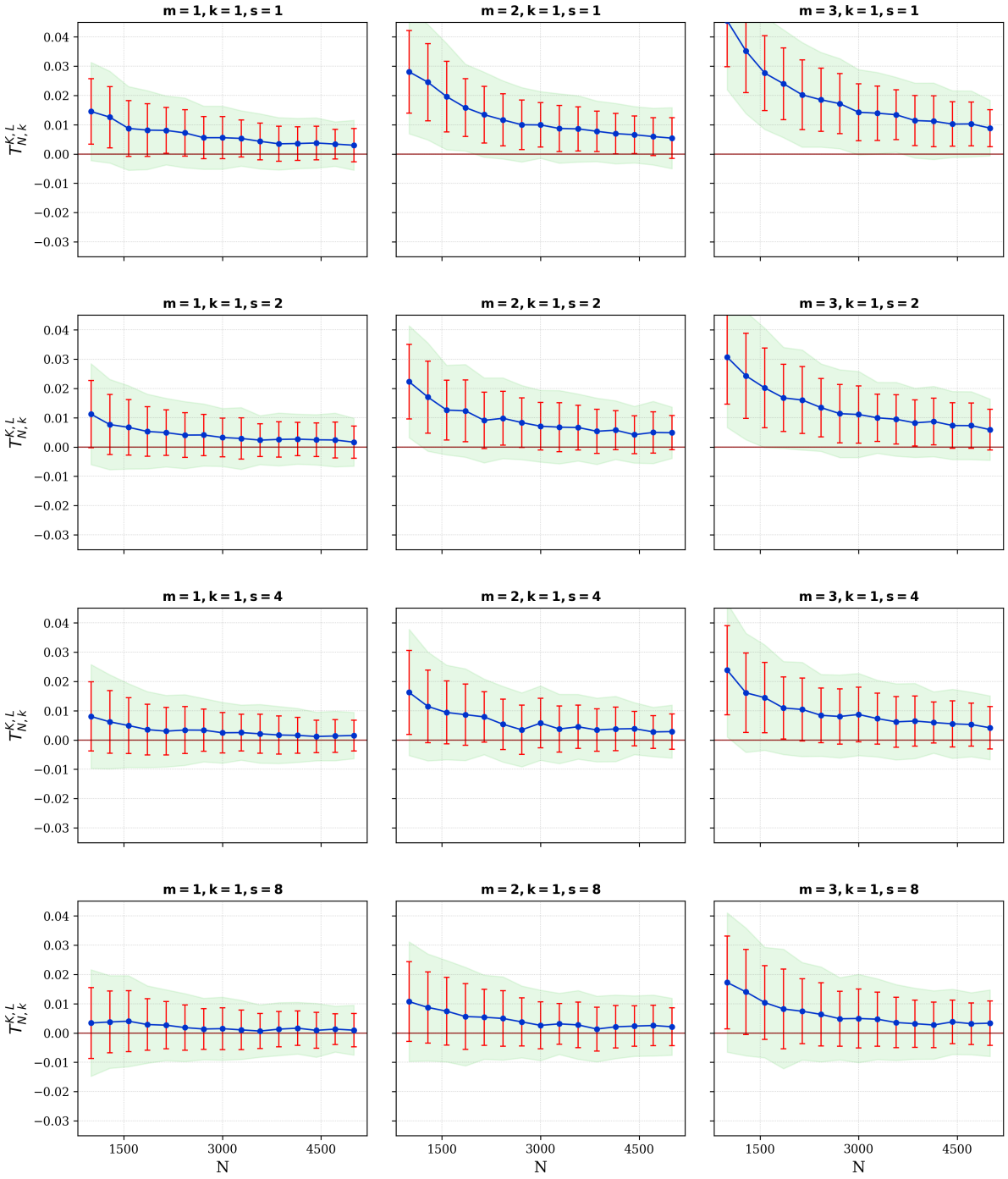


图1：$T_{N,k}^{\mathrm{KL}}$ 在邻域维数为 $k = 1$时的收敛行为。误差条表示经过 $M = 100$ 次迭代后经验得到的一个标准差。这对应于高斯基准 $s = 2$；此处，随着N 增加，该统计量集中在零附近。$T_{N,k}^{\mathrm{KL}}$ 在 $k = 1$时的收敛行为。

随后，我们针对较小的样本量重复实验 $N \in \{100, \ldots, 1000\}$和 $k \in \{1, 2, 3\}$，以研究有限样本区间以及 k的影响。图 2显示，增大 k 基本上不会改变集中趋势，却能降低变异性。这些结果与 kNN 泛函中标准的偏差-方差权衡一致：较大的邻域会增加均值，因此方差较小，但通常会以偏差略微增大为代价 [1]。在当前 (m, N) 范围内，方差降低是主要且明显的效应。

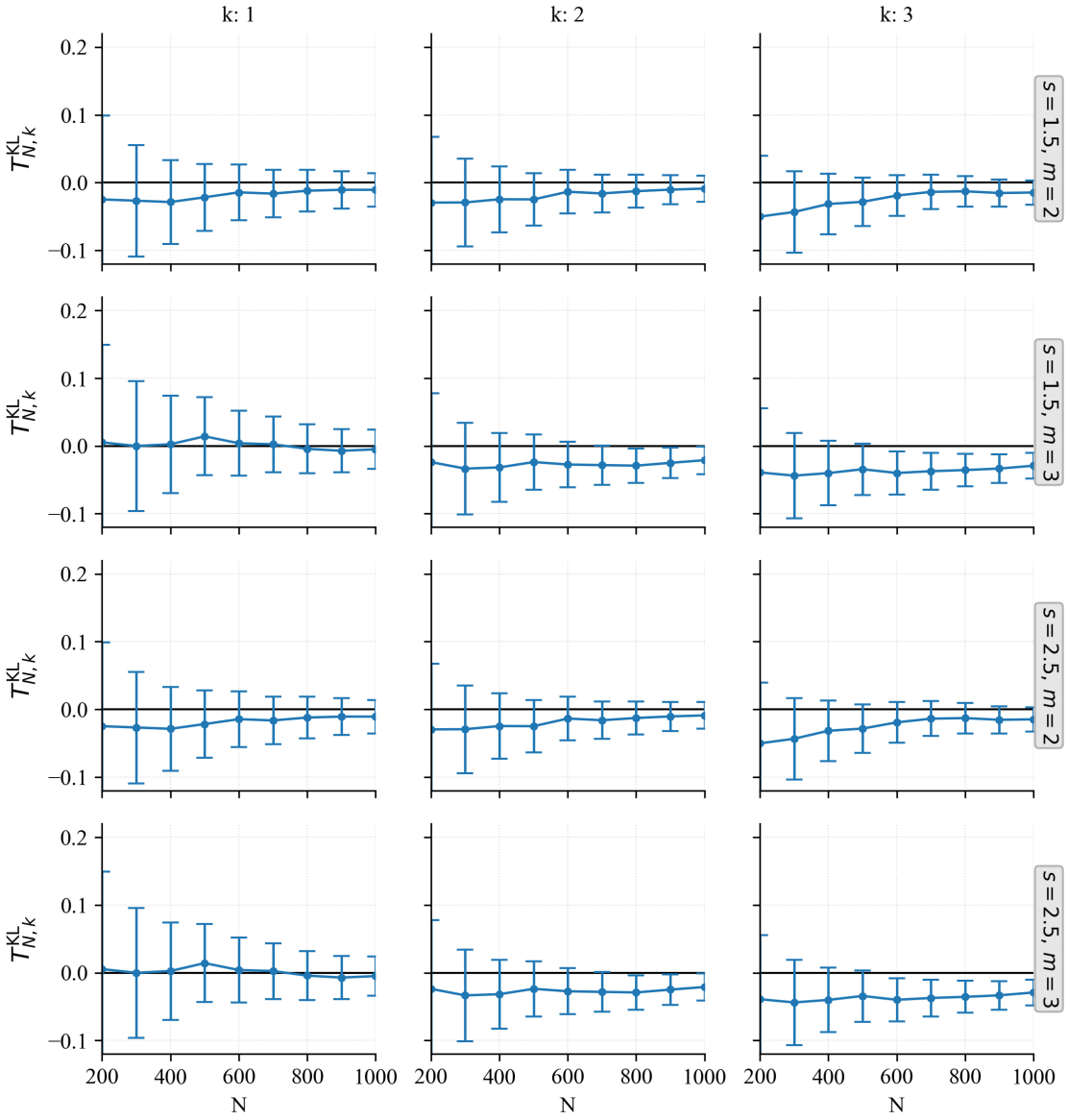


图2：$T^{\mathrm{KL}}_{N,\,k}$ 在不同邻域维数 $k \in \{1, 2, 3\}$下的有限样本稳定性。误差条表示一个经验标准差。变异性随 N 增大而降低，并且随着邻域增大而进一步降低。

5.2 拟合优度检验的功效

对于每种配置，我们利用算法1¯，使用拟合的高斯模型 N_m（X_N， S_N）计算自助法临界值。我们评估经验功效，用以检验

显著性水平 $\alpha = 0.05$下的多元正态性。在此情形下，经验功效是在备择假设下生成的蒙特卡洛重复中，观测到的$T_{N,\,k}^{\mathrm{KL}}$ 超过自助法临界值的比例。

在功效实验中，使用 $M = 1000$ 次重复次数和样本量 $N \in \{500, 1000\}$ 。我们考虑两类互补的备择：**(i)** 广义高斯族内的偏离（受控的尾部行为），以及 **(ii)** 重尾学生型备择（在有限自由度下偏离广义高斯族之外）。

5.2.1 针对广义高斯备择的功效

在第一个实验中，我们通过改变形状参数，使其偏离高斯参考值，从而推广高斯族。具体而言，我们固定位置参数和方差，并将高斯情形与 $X \sim GG(m,\ v)$ 进行比较，其中 $v \neq 2$。图 3 展示了由此得到的功效曲线。

两种模型规模之间存在一致性。首先，功效随偏离高斯分布的程度增大而提高，这反映了 $T_{N,\,k}^{\mathrm{KL}}$ 针对固定备择的一致性。其次，将样本量从 $N = 500$增加到 $N = 1000$会显著提高灵敏度。邻域大小 k 的选择具有次要但系统性的影响：较大的邻域通常会产生更平滑、变异更小的功效曲线，这与图 2中经常观察到的方差降低相一致。

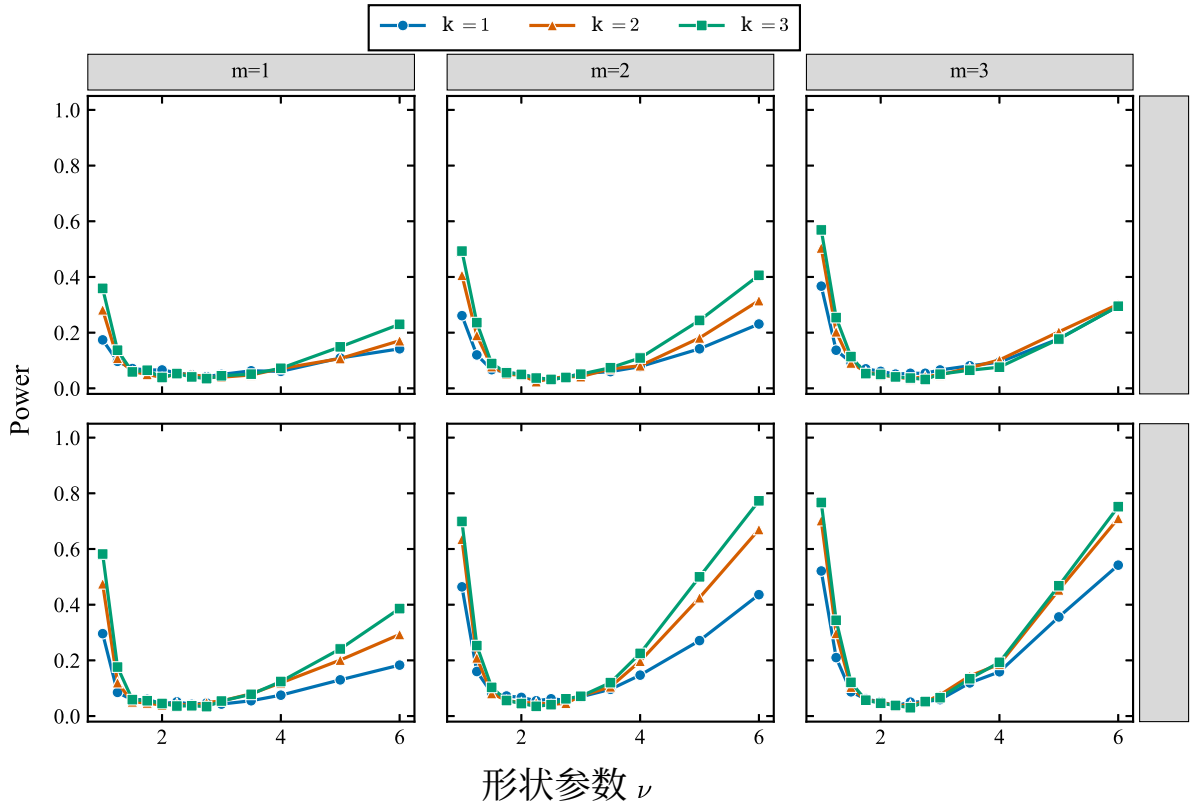


图 3：曲线比较了不同维度下的 $k \in \{1, 2, 3\}$ ；各行分别对应 N =500 和 $N = 1000$，重复次数为 $M = 1000$ 。曲线比较了不同维度下的 $k \in \{1, 2, 3\}$ ；各行分别对应 $N = 500$ 和 $N = 1000$，重复次数为 $M = 1000$ 。

5.2.2 对 Student 型备择的功效

接下来，我们将考虑由自由度 ν参数化的重尾学生 t 型备择。此处，较小的 ν 值对应更重的尾部。图 4 表明，在重尾情形（较小的 ν）下，功效接近 1，并随 ν 增大而下降。可以预期会出现这一趋势，因为较大的 ν 值会产生越来越接近高斯参照分布的分布，从而使备择更难被区分。对于 $N = 500$ 和 $N = 1000$，曲线之间的差异再次证实，较大的样本量会提高灵敏度。在大多数配置下，中等邻域大小会产生略为更稳定的功效曲线，这与收敛实验中观察到的稳定性效应相一致。

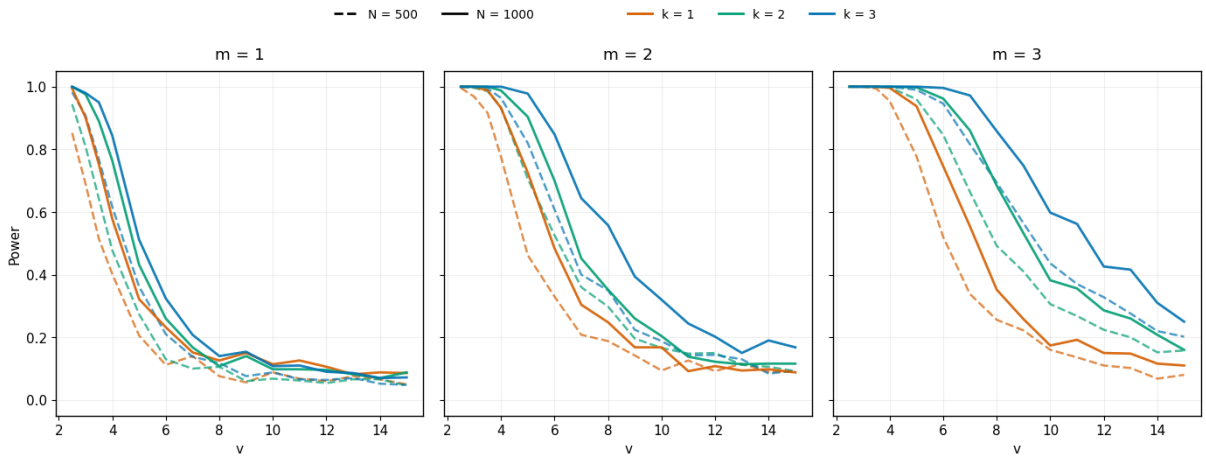


图 4：$T_{N,\,k}^{\mathrm{KL}}$ 与学生化备择相比的经验功效。实线和虚线分别表示 $N = 1000$ 和 $N = 500$；这些线比较的是$k \in \{1, 2, 3\}$。

5.3 标准化统计量的经验分布

尽管使用自助法检验进行了校准，考察统计量标准化后的形状仍可能有所启发。

$$Z_{N,k} = \frac{T_{N,k}^{\mathrm{KL}} - \widehat{\mu}_{N,k}}{\widehat{\sigma}_{N,k}},$$

其中 $\widehat{\mu}_{N,\,k}$ 和 $\widehat{\sigma}_{N,\,k}$ 分别表示使用高斯基准进行蒙特卡洛重复后计算得到的经验均值和标准差。图5 对比了$Z_{N,\,k}$在 $N = 1000$和 $k \in \{1, 2, 3\}$处的核密度估计与标准正态密度。在此情形下，两者拟合良好，表明正态近似可能适用于描述性分析。需要强调的是，这些证据仅具探索性：本文始终使用算法1中的参数自助法计算拒绝阈值，并未采用假设的渐近正态极限。

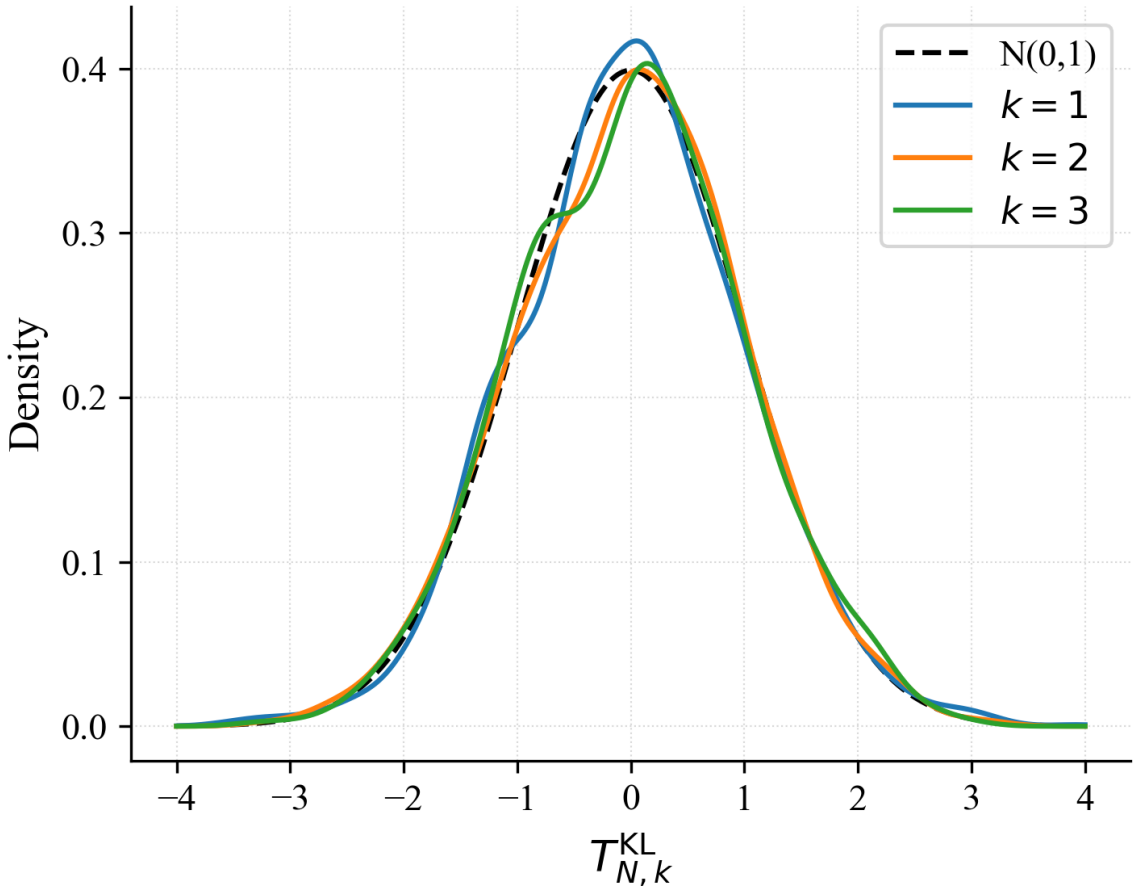


图 5：对于 $N = 1000$ 和 $k \in \{1, 2, 3\}$，对标准化统计量 $Z_{N,\,k}$进行了核密度估计，并与标准高斯密度 $\mathcal{N}(0,\ 1)$进行了比较。

图6 给出了相应的 Q-Q图。与核密度估计的证据一致，经验分位数与参考线十分接近，包括尾部。

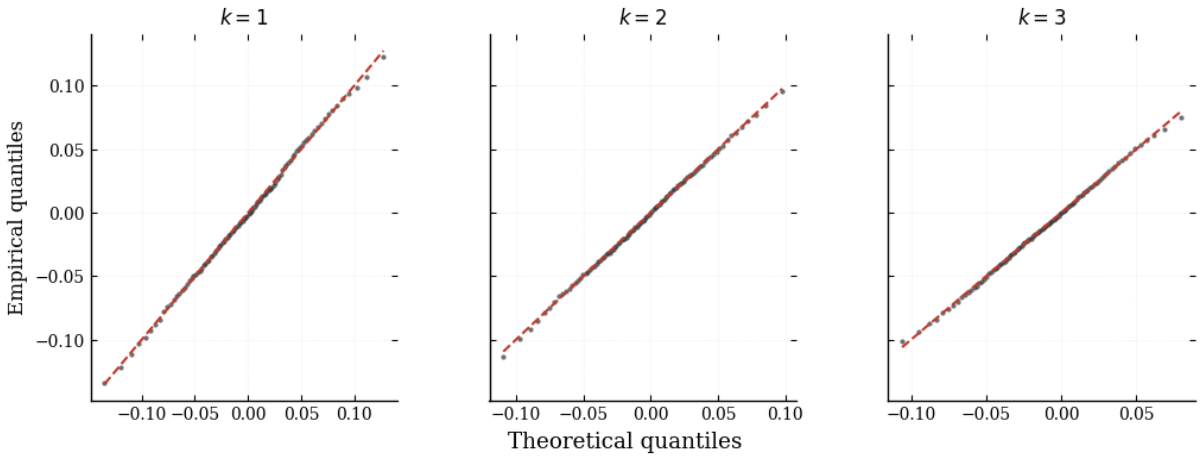


图 6：标准化统计量 $Z_{N,\,k}$ 在 $N = 1000$和 $k \in \{1, 2, 3\}$下的正态分布 Q-Q图。

表1：双对数回归得到的估计斜率系数 β (10) for $\left|\mathbf{E}[T_{\mathrm{KL}N,k}(m,s)]\right|$ 相对于 N。

s	$m=1$			$m=2$			$m=3$		
	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$	$k=1$	$k=2$	$k=3$
1.0	0.0092	0.0085	0.0078	0.0124	0.0110	0.0095	0.0145	0.0131	0.0118
1.5	0.0041	0.0035	0.0029	0.0058	0.0049	0.0042	0.0072	0.0064	0.0056
2.0	0.0004	0.0002	0.0001	0.0008	0.0005	0.0003	0.0012	0.0009	0.0007
2.5	0.0006	0.0004	0.0002	0.0011	0.0009	0.0007	0.0018	0.0015	0.0012
3.0	0.0015	0.0012	0.0009	0.0024	0.0020	0.0017	0.0031	0.0028	0.0024
4.0	0.0032	0.0028	0.0023	0.0045	0.0039	0.0034	0.0056	0.0051	0.0046

5.4 有限样本收敛速率：双对数回归诊断

为概述 $T_{N,k}^{\mathrm{KL}}$ 在高斯基准下的收敛速率，我们对不同样本量下 $\left|\mathbf{E}[T_{N,k}^{\mathrm{KL}}(m,s)]\right|$ 的蒙特卡罗估计值进行了简单的双对数回归：

$$\log\left|\mathbb{E}\left[T_{N,k}^{\mathrm{KL}}(m,s)\right]\right| = \alpha_{m,s,k} + \beta_{m,s,k}\,\log N. \tag{10}$$

对于 kNN熵泛函，在正则情形下，主导偏差项通常为 $\mathcal{O}(N^{-1/2})$ 阶。基于这一直观方法，在高斯基准下（对应于广义高斯族中的 $s=2$），预期有 $\beta_{m,s,k}\approx -\frac{1}{2}$。因此，与该基准的偏离可解释为衰减较慢或有限样本效应较强。

表1 给出了维度 m、邻域大小 k和形状参数 s的斜率估计。这些值在高斯参考值 $s=2$附近达到最小值，此时收敛最快。当我们远离 $s=2$时，斜率略有增加，但仍保持较小，这与该统计量在重尾和轻尾生成器下的稳定行为相一致。此外，当 k 从1 增加到3时，我们注意到 β略有下降，反映出较大邻域中最近邻估计的稳定性有所提高。

5.5 实际实施的临界值

我们通过拟合的高斯模型下的参数自助法估计5%临界值 $t_{0.05}$（算法1），使用 $M=1000$ 次重复。表2 给出了样本量 $N\in\{100,\dots,1000\}$、维度 $m\in\{2,3\}$和邻域大小$k\in\{1,2,3\}$下的这些阈值。（以 s 为索引的行可理解为针对不同生成器的敏感性检验；对于高斯校准，相关行为 $s=2$。）

这些临界值随着 N 增加而缓慢降低，这反映出在高斯基准下，$T_{N,\,k}^{\mathrm{KL}}$ 趋近于零。固定 N时，$m=3$对应的阈值大于 $m=2$，这与更高维度下局部熵估计面临的挑战增大相一致。通常，增大 k 会使阈值略微降低，这与图2所示的方差降低趋势一致。

表 2：基于 KL 的检验统计量 $T^{\mathrm{KL}}_{N\,k,}$ $(m,\ \ s)$ 在 5% 水平下的临界值 $t_{0.05}$ 是通过 $M = 1000$ 重复参数自助法估计的。对于高斯拟合优度检验，第 $s = 2$ 行对应高斯参考值。

s	N	$m = 2$			$m = 3$		
		$k = 1$	$k = 2$	$k = 3$	$k = 1$	$k = 2$	$k = 3$
1.5	100	0.1428	0.1315	0.1254	0.1652	0.1540	0.1488
	200	0.0985	0.0912	0.0865	0.1187	0.1095	0.1052
	300	0.0789	0.0724	0.0691	0.0954	0.0881	0.0843
	400	0.0672	0.0615	0.0588	0.0815	0.0748	0.0712
	500	0.0594	0.0541	0.0519	0.0721	0.0665	0.0634
	600	0.0538	0.0489	0.0467	0.0652	0.0598	0.0571
	700	0.0495	0.0448	0.0426	0.0598	0.0549	0.0525
	800	0.0458	0.0415	0.0395	0.0556	0.0508	0.0485
	900	0.0429	0.0388	0.0368	0.0521	0.0475	0.0452
	1000	0.0405	0.0365	0.0346	0.0492	0.0448	0.0426
2.0	100	0.1256	0.1148	0.1085	0.1485	0.1362	0.1295
	200	0.0864	0.0792	0.0745	0.1054	0.0968	0.0915
	300	0.0695	0.0632	0.0598	0.0856	0.0785	0.0742
	400	0.0592	0.0538	0.0505	0.0728	0.0665	0.0628
	500	0.0524	0.0475	0.0448	0.0645	0.0589	0.0556
	600	0.0475	0.0428	0.0402	0.0582	0.0532	0.0501
	700	0.0436	0.0392	0.0368	0.0535	0.0488	0.0462
	800	0.0405	0.0365	0.0342	0.0495	0.0452	0.0428
	900	0.0378	0.0340	0.0318	0.0464	0.0422	0.0398
	1000	0.0356	0.0319	0.0299	0.0438	0.0398	0.0375
2.5	100	0.1345	0.1228	0.1165	0.1568	0.1452	0.1385
	200	0.0925	0.0856	0.0805	0.1124	0.1028	0.0982
	300	0.0742	0.0678	0.0642	0.0905	0.0825	0.0788
	400	0.0635	0.0575	0.0545	0.0772	0.0705	0.0668
	500	0.0562	0.0508	0.0482	0.0685	0.0622	0.0589
	600	0.0508	0.0458	0.0435	0.0618	0.0562	0.0532
	700	0.0468	0.0420	0.0398	0.0568	0.0515	0.0488
	800	0.0435	0.0390	0.0368	0.0526	0.0478	0.0452
	900	0.0408	0.0365	0.0345	0.0492	0.0448	0.0422
	1000	0.0385	0.0342	0.0324	0.0465	0.0422	0.0398

在实际应用中，当观测到的统计量 $T_{N,k}^{\mathrm{KL}} \geq t_{0.05}$ 满足条件时，在 $\alpha = 0.05$ 显著性水平下拒绝多元正态性。在这种情况下， $t_{0.05}$ 通过自助法校准计算（算法 1），或根据此处报告的设置从表 2中获取。

6 结论

本文提出了一种基于信息论的框架，可在似然密度估计不可靠的情形下，以较低的计算复杂度进行多元拟合优度检验。

我们提出了一种基于香农熵的 k 近邻估计量的多元正态性检验 KL 分解方法。其基本结构特征是，在固定的均值—协方差类别中，高斯参考值将 KL 分解转化为熵差，因此统计量 $T_{N,k}^{\mathrm{KL}}$ 仅反映非负的不一致性，并且在多元正态性成立时该不一致性趋于零。在标准正则条件下，k 近邻熵估计量具有一致性，这直接意味着由此得到的基于 KL 的检验统计量也具有一致性。

我们的数值研究得出了三项实用结论。首先，随着 N 增加，高斯基准下的统计量迅速趋近于零。其次，随着偏离强度增加，检验功效平稳上升，这也包括广义高斯模型中的形状变化以及重尾的 Student 型备择。最后，自助法校准能够稳定地控制有限样本误差，并生成其表现符合 N、 m 和 k 预测的临界值。

参考文献

[1] Thomas B. Berrett、Richard J. Samworth 和 Ming Yuan。通过 k 近邻距离实现高效的多元熵估计。统计年鉴，47(1):288– 318，2019。[2] Alexander Bulinski 和 Dimitar Dimitrov。香农熵的统计估计。《数学学报（英文版）》，35(1):17–46，2019。[3]Mehmet Siddik Cadirci、Dafydd Evans、Nikolai Leonenko 和 Vitalii Makogin。广义高斯分布的基于熵的检验。《计算统计与数据分析》，173:107502，2022。[4] Thomas M. Cover 和 Joy A. Thomas。《信息论基础》。John Wiley & Sons，霍博肯，新泽西州，第2版，2006。[5]Weihao Gao、Sewoong Oh 和 Pramod Viswanath。揭开定点 k 近邻信息估计量的神秘面纱。*IEEE* 信息论汇刊，64(8):5629–5661，2018。[6] Valérie Girardin 和 Justine Lequesne。基于熵的拟合优度检验——统一框架：在 DNA复制中的应用。《统计学通讯——理论与方法》，46(24):12080–12110，2017。[7] M. N. Goria、N. N. Leonenko、V. V. Mergel 和 P. L. Novi Inverardi。一类新的随机向量熵估计量及其在统计假设检验中的应用。《非参数统计学杂志》，17(3):277–297，2005。[8] Edwin T. Jaynes。信息论与统计力学。《物理评论》，106(4):620–630，1957。[9] Edwin T. Jaynes。《概率论：科学的逻辑》。剑桥大学出版社，英国剑桥，2003。G. Larry Bretthorst 编。[10] J. N. Kapur。《科学与工程中的最大熵模型》。Wiley，纽约，1989。[11] L. F. Kozachenko 和 N. N. Leonenko。随机向量熵的统计估计。《信息传输问题》，23(2):9–16，1987。[12] Solomon Kullback。《信息论与统计学》。多佛出版社，纽约州米尼奥拉，第2版，1997。[13] Solomon Kullback 和 Richard A. Leibler。论信息与充分性。《数理统计年鉴》，22(1):79–86，1951。[14] N. N. Leonenko、L. Pronzato 和 V. Savani。一类用于多维密度的 Rényi 信息估计量。统计年鉴，36(5):2153–2182，2008。[15] Justine Lequesne 和 Philippe Regnault。vsgoftest：用于基于 Kullback–Leibler 散度进行拟合优度检验的 R软件包。《统计软件杂志》，96(c01):1–39，2020。

[16] Hadi A. Noughabi. 一种新的 Kullback–Leibler 信息估计量及其在拟合优度检验中的应用。统计计算与模拟杂志，89(10):1914–1934，2019。[17] Hadi A. Noughabi 和 Naser Reza Arghami。基于 Kullback–Leibler 信息的拟合优度检验通用处理方法。统计计算与模拟杂志，83(8):1556–1569，2013。[18] Liam Paninski。熵和互信息的估计。神经计算，15(6):1191–1253，2003。[19] M. D. Penrose 和 J. E. Yukich。流形上点过程的极限定理。应用概率年刊，23(6):2161–2211，2013。[20] Mathew D. Penrose 和 Joseph E. Yukich。大数定律与最近邻距离。载于 Peter Jupp、Thomas Wehrly、Barry C. Arnold 和 N. Balakrishnan 编，《方向与线性统计进展》，第189–199页。施普林格，柏林、海德堡，2011。[21] Fernando Pérez-Cruz。连续分布的 Kullback–Leibler 散度估计。载于*2008 IEEE* 国际信息论研讨会（*ISIT*），第1666–1670页。IEEE，2008。[22] Claude Elwood Shannon。通信的数学理论。贝尔系统技术期刊，27(3):379–423，1948。[23] N. Solaro。多元指数幂分布的随机变量生成。统计学与应用，2(2):25–44，2004。[24] Kai-Sheng Song。基于 Kullback–Leibler 判别信息的拟合优度检验。*IEEE* 信息论汇刊，48(5):1103–1117，2002。[25] Qian Wang、Sanjeev R. Kulkarni 和 Sergio Verdú。通过 k 最近邻距离估计多维密度的散度。*IEEE* 信息论汇刊，55(5):2392–2405，2009。