arXiv:2601.05993v1 [math.ST] 9 Jan 2026

# 圆形数据中植入结构的检测

Taha Ameen 和 Bruce Hajek *

**摘要**

本文构建了圆形数据的假设检验问题，其中观测值取自单位圆，且可能包含隐藏的相位相干结构。在零假设下，数据是在单位圆上独立均匀分布的；在备择假设下，或者（i）一个大小为 $K$ 的植入子集集中在某个未知相位附近（平坦情形），或者（ii）一个大小为 $k$ 的植入社群在完全图的边之间诱导出相干性（社群情形）。在这两种情形中，本文均考虑两种圆形信号分布：硬聚类分布，其中相关的植入观测值位于一个长度已知、位置未知的弧上；以及冯·米塞斯分布，其中相关的植入观测值服从具有共同未知位置参数的冯·米塞斯分布。对于由此得到的四种模型，本文推导出几乎匹配的可检测性充要条件（相差仅为常数以及偶尔出现的对数因子），从而确立了信息论相变。

## 目录

**1 引言** **2**

1.1 模型、检验与目标 .3 1.2 相关工作 .5 1.3 主要结果总结与讨论 .6 **2 硬聚类平坦模型** 8 2.1 硬聚类平坦模型的可实现性——区间检验 .9 2.2 硬聚类平坦模型的逆命题 .10 **3 冯·米塞斯平坦模型** 13 3.1 冯·米塞斯平坦模型的可实现性——区间检验 .14 3.2 冯·米塞斯平坦模型的逆命题 .15 **4 硬簇社区模型** 17 4.1 硬簇社区模型的可实现性——区间检验 .18 4.2 硬簇社区模型的逆命题 .19 **5 冯·米塞斯平坦模型**

**es 社区模型** 20

5.1 冯·米塞斯社区模型的可实现性——区间检验 .22 5.2 冯·米塞斯社区模型的可实现性——相干性检验 .22 5.3 冯·米塞斯社区模型的逆命题 .24

*作者供职于美国伊利诺伊州厄巴纳市 61801 邮编的伊利诺伊大学电气与计算机工程系和协同科学实验室。电子邮件：{ tahaa3, b-hajek } @illinois.edu

**A 贝塞尔函数的界与近似 25B 广义似然比检验 26**B.1 硬聚类平坦模型的区间检验 .26B.2 冯·米塞斯社区模型的相干性检验 .27**C 关于多项式时间检验 28D 关于区间检验与相干性检验之间的关系 29E 关于已知 $\Theta^*$ 29F 社区恢复及其与随机块模型的联系 30**

# 1 引言

许多现代数据集本质上具有方向性，其取值位于单位圆上，而非实数轴上。例如，振荡信号的相位、一天或一年中的时刻（分别视为以 24 小时或一年为周期）、运动物体的朝向，以及时间序列两两之间的相位差。在这类情形中，一个自然的统计问题是：观测到的角度是否没有结构（与圆上的均匀分布相容），还是存在一个隐蔽子集，其角度呈现出某种偏好的方向或相位。

本文研究存在植入结构时的圆形数据。在零假设下，所有观测值均为圆上的独立同分布均匀分布；而在备择假设下，存在一个规模较小、未知的指标子集（或网络中的顶点社区），其角度集中在某个未知相位附近。这些假设检验问题结合了两个经典主题：(i) 圆形统计学，其中冯·米塞斯分布可视为“圆上的高斯分布” [MJ09], ；以及 (ii) 高维检测中的植入结构与相变，例如植入团和随机块模型中的社群检测。

我们的目标是刻画在方向型植入模型中，检测在信息论意义上何时可行。我们介绍这类问题出现的三个应用领域。

1. 事件数据中的季节性和昼夜效应。考虑带时间戳的事件（交易、故障、攻击、消息），其中 $X_i \in [0, 2\pi$ ）表示第 $i$ 个事件发生的时间（以 24 小时或一年为周期）。许多应用中，背景活动均匀分布于所有时段，同时还有一小部分事件由季节性或昼夜机制驱动（例如故障集中发生在炎热的夏日，或网络攻击集中发生在夜间）。在一种自然的零假设模型下，所有事件时间在 $[0, 2\pi$ 上独立同分布均匀；在备择假设下，大小为 $K$ 的未知子群体，其事件时间集中在长度为 $2\pi\tau$ 的未知弧内（对于 $\tau < 1$），或集中在集中度为 $\kappa$ 的未知优选相位附近。这里，$\tau$ 或 $\kappa$ 量化集中程度的强弱，而 $K$ 表示受影响群体的规模。

2. 锁相振荡器。假设我们测量了 $N$ 个振荡器的瞬时相位 $X_1, \ldots, X_N \in [0, 2\pi)$（神经元群体中的神经元、电网中的设备，或用于建模同步现象的 Kuramoto 振荡器）。当振荡器完全失同步时，其相位在圆上接近均匀分布。然而，当一部分 $K$ 个振荡器受到同一信号源驱动时，其相位会局部化在某个共同方向附近。这种局部化的强度可用集中参数 $\kappa$ 建模，其中冯·

米塞斯分布提供了均匀性的典型备择假设。

3. 网络中的相位相干社区。考虑一个由 $n$ 个时间序列（脑电图通道、金融资产或传感器读数）构成的网络。在给定频率下，序列 $i$ 和 $j$ 之间的成对关系可用其交叉谱相干性来概括。这是一种复值量，其相位 $X_{ij} \in [0, 2\pi)$编码了两个信号之间的相位偏移。由此得到一个矩阵 $(X_{ij})_{1\le i<j\le n}$，其中包含完全图各条边上的角度测量。在无结构的相互作用下，这些边的相位近似均匀分布。然而，如果节点数为 $k$ 的某个社区 $C^*$ 由一个共同的振荡源驱动（例如传感器网络中的共享节律），那么 $i$的相位 $X_{ij}$ 会围绕某个偏移量 $\Theta^*$对齐，并表现出明显高于背景边的相干性，$j \in C^*$ 。

任务是确定何时能够从观测到的圆形数据中检测出一个相位相干子集（或社区），并考察这一条件如何随总体规模（$N$ 或 $n$）、子集大小（$K$ 或 $k$）以及集中强度（$\tau$ 或 $\kappa$）变化。

## 1.1 模型、检验与目标

若随机变量 $X$ 服从位置参数为 $\Theta^* \in [0, 2\pi)$、集中参数为 $\kappa > 0$ 的冯·米塞斯分布，记作 $X \sim$ 冯米塞斯$(\Theta^*, \kappa)$，则其概率密度为

$$f(\theta; \Theta^*, \kappa) = \frac{1}{2\pi I_0(\kappa)} \exp\big(\kappa \cos(\theta - \Theta^*)\big), \qquad \theta \in [0, 2\pi),$$

其中，$I_0$ 是阶数为 0 的第一类修正贝塞尔函数，$I_0(\kappa) = \frac{1}{2\pi}\int_0^{2\pi} \exp\big(\kappa \cos u\big)\, \mathrm{d}u$.

下面介绍四个假设检验问题，并在模型名称后列出模型参数。在所有模型中，零假设 $H_0$ 假定方向同质性（圆上的数据独立同分布且服从均匀分布），而备择假设 $H_1$ 则引入一个结构化成分。

1. **硬聚类平坦模型**（$N$，$K$，$\tau$）。观测到角度 $X_1$、……、$X_N \in [0, 2\pi)$。在 $H_0$下，$X_i$ 相互独立且在 $[0, 2\pi)$ 上均匀分布。在 $H_1$下，存在一个大小为 $K$ 的未知植入子集$S^* \subset [N]$ 以及一个未知弧位置 $\Theta^* \in [0, 2\pi)$，使得

$$X_i \sim \mathrm{Uniform}([\Theta^*, \Theta^* + 2\pi\tau]) \quad \text{for } i \in S^*, \qquad X_i \sim \mathrm{Uniform}([0, 2\pi]) \quad \text{for } i \notin S^*,$$

相互独立。这描述了圆上一个未知宽度为 $2\pi\tau$ 的窗口内聚集的事件或振荡器硬聚类。

2. **冯·米塞斯平坦模型**（$N$，$K$，$\kappa$）。观测到角度 $X_1$，……，$X_N \in [0, 2\pi)$，其零假设与硬聚类平坦模型相同。在 $H_1$条件下，存在一个大小为 $K$的植入子集 $S^*$ 和一个优选相位 $\Theta^*$，使得

$$X_i \sim \mathrm{vonMises}(\Theta^*, \kappa) \quad \text{for } i \in S^*, \qquad X_i \sim \mathrm{Uniform}([0, 2\pi]) \quad \text{for } i \notin S^*,$$

相互独立。

3. **硬簇社区模型**（$n$，$k$，$\tau$）。观测到一个对称角度矩阵 $(X_{ij})_{1\le i,\, j\le n}$ 可将其视为含有 $n$ 个顶点的完全图各边上的角度标签。在$H_0$下，$X_{ij}$ 的 $i < j$ 在 $[0, 2\pi$上独立均匀分布）。在 $H_1$下，存在一个大小为 $k$ 的未知社区 $C^* \subset [n]$ 以及一个未知相位 $\Theta^*$，使得

$$X_e \sim \mathrm{Uniform}([\Theta^*, \Theta^* + 2\pi\tau]) \quad \text{for } e \in E(C^*),$$

其中 $E(C) \triangleq \{e = \{i, j\} : \ i, \ j \in C, \ \ i < j\}$ 表示两个端点都位于 $C$中的所有边的集合，而至少有一个端点位于 $C^*$ 之外的所有边的角度仍在 $[0, 2\pi$上均匀分布）。

**4. 冯·米塞斯社区模型 $(n, k, \kappa)$。**观测到一个对称角度矩阵 $(X_{ij})_{1 \le i,j \le n}$ 。在 $H_0$ 下， $X_{ij}$ 中的 $i < j$ 相互独立，且服从 $[0, 2\pi$ 上的均匀分布。在 $H_1$ 下，存在一个大小为 $k$ 的未知社区 $C^* \subset [n]$ ，以及一个未知相位 $\Theta^*$ ，使得

$$X_e \sim \text{vonMises}(\Theta^*, \kappa) \quad \text{for } e \in E(C^*)\,,$$

而所有其他边仍服从 $[0, 2\pi$ 上的均匀分布。

在每个模型中，假设在 $H_1$ 下，植入集合 $S^*$ （分别为 $C^*$）是在所有大小为 $K$ （分别为 $k$）的子集中均匀随机抽取的，并且植入相位 $\Theta^*$ 是从 $[0, 2\pi$ 上的均匀分布中独立抽取的。于是， $H_1$ 下观测值的分布记为 $\mathbb{P}$，它是对 $(S^*, \Theta^*)$ （分别为 $(C^*, \Theta^*)$ ）的混合分布，而 $\mathbb{Q}$ 表示相应的 $H_0$ 下的分布。

我们并列研究上述四种模型，因为它们体现了方向结构的相关表现形式。硬聚类模型和冯·米塞斯模型对应于圆上集中建模的两种典型方式；在窗口宽度 $\tau$ 较小 $\kappa$或集中参数 $\kappa$较大时，它们密切相关，其粗略对应关系为 $\tau \approx \kappa^{-1/2}$。同样，平坦模型和社区模型的主要区别在于它们的组合结构，但二者共享许多相同的概率机制和界定技术。我们并列研究这些模型，以突出哪些现象是特定几何所独有的，哪些则是更广泛的方向性植入结构所固有的。

现在我们将检测问题形式化。检验是一个可测函数 $\phi$ ，它将数据映射到 $\{0, 1\}$，其中 $\phi(X) = 1$ 表示拒绝 $H_0$ 而支持 $H_1$。虚警和漏检概率定义如下：

$$p_{\text{fa}}(\phi) \triangleq \mathbb{Q}\{\phi(X) = 1\}\,, \qquad p_{\text{miss}}(\phi) \triangleq \mathbb{P}\{\phi(X) = 0\}\,.$$

当 $n$ （或 $N$）趋于无穷大时，我们将关注以下检测区域。

**定义 1** （强检测与弱检测）。

- 可以实现强检测 当且仅当存在一列检验 $(\phi_n)$ 使得 $p_{\text{fa}}(\phi_n) \to 0$ 且$p_{\text{miss}}(\phi_n) \to 0$。
- 可以实现弱检测 当存在一列检验 $(\phi_n)$ 和常数 $\delta > 0$ ，使得 $p_{\text{fa}}(\phi_n) + p_{\text{miss}}(\phi_n) \le 1 - \delta$ 对所有充分大的 $n$均成立。
- 若对于每一列检验 $(\phi_n)$，$p_{\text{fa}}(\phi_n) + p_{\text{miss}}(\phi_n) \ge 1 - o(1)$，则弱检测不可能实现。等价地，若 $\mathbb{P}$ 与 $\mathbb{Q}$ 之间的全变差距离趋于零，则弱检测不可能实现。

**区间与相干性检验** 我们提出并分析了针对方向性植入结构量身定制的自然检验程序。对于平坦模型，我们研究一种区间检验，其窗口长度为$2\pi\tau$，阈值为 $\gamma$。给定观测值 $X_1, \dots, X_N \in [0, 2\pi)$，定义统计量

$$T_{\text{int}}(X) \triangleq \sup_{\theta \in [0, 2\pi)} \sum_{i=1}^{N} \mathbf{1}\{X_i \in [\theta, \theta + 2\pi\tau]\}. \tag{1}$$

该 参数为 $(\tau, \gamma)$ 的平坦模型的区间检验在 $T_{\text{int}}(X) \ge \gamma$ 时拒绝 $H_0$ .

同样地，对于社区模型，我们研究窗口长度为 $2\pi\tau$ 的区间检验。给定观测值 $(X_{ij})_{1\le i<j\le n}$，考虑统计量

$$T_{\text{int}}^{\text{comm}}(X) \triangleq \sup_{\theta\in[0,2\pi)} \sup_{C\subseteq[n]:|C|=k} \mathbf{1}\left\{X_e \in [\theta,\theta+2\pi\tau] \text{ for all } e\in E(C)\right\}. \tag{2}$$

参数为 $(k,\ \tau)$ 的社区区间检验在 $T_{\text{commint}}(X)$ = 时拒绝 $H_0$ 1.

对于冯·米塞斯社区模型，我们还研究了一种相干性检验。对于每条边 $e=\{i,j\}$，令$Z_e = e^{\mathrm{i}X_e}$ (其中 i $\triangleq \sqrt{-1}$)，并定义相干性统计量

$$T_{\text{coh}}(X) \triangleq \max_{C\subset[n]:\,|C|=k} \left|\sum_{e\in E(C)} Z_e\right|, \tag{3}$$

参数为 $(k,\ \beta)$ 的相干性检验在 $T_{\text{coh}}(X) \ge \beta$ 时拒绝 $H_0$ 。

方向统计学研究圆或球面上的数据（角度、相位），并拥有成熟的建模与推断工具MJ09，*Fis*。该领域的一个核心问题是检验圆周均匀性：给定独立同分布角度，.....，），判断它们是否与 uniform()) 相符，还是呈现某种系统性偏离。经典检验包括经典的 Rayleigh 检验Ray（与我们的相干性检验相关），以及库伊珀检验Kui和沃森检验Wat（比较经验累积分布函数与均匀分布的累积分布函数），以及 Rao 间距检验 Rao（基于排序后角度之间的间隔）。这些方法（及其贝叶斯对应方法LRM18，MK）在齐次情形下检验拟合优度，即所有观测值均来自同一个共同分布，目标是检测任何整体上的非均匀性。文献Ben研究了一种相关的污染模型，该模型拟合冯·米塞斯混合分布与均匀分布数据，并使用似然比和沃森型拟合优度程序，在均匀分布、冯·米塞斯分布和混合分布拟合之间进行选择。在这里，每个观测值均独立同分布地来自同一个混合分布。

## 1.2 相关工作

**方向统计学**研究圆或球面上的数据（角度、相位），并拥有成熟的建模与推断工具 [MJ09，Fis95]。该领域的一个核心问题是检验圆周均匀性：给定独立同分布角度 $\theta_1$，.....，$\theta_n \in [0,2\pi)$，判断它们是否与 uniform($[0,2\pi)$) 相符，还是呈现某种系统性偏离。经典检验包括经典的 Rayleigh 检验 [Ray80]（与我们的相干性检验相关），以及库伊珀检验 [Kui60] 和沃森检验 [Wat61]（比较经验累积分布函数与均匀分布的累积分布函数），以及 Rao 间距检验 [Rao76]（基于排序后角度之间的间隔）。这些方法（及其贝叶斯对应方法 [LRM18，MK21]）在齐次情形下检验拟合优度，即所有观测值均来自同一个共同分布，目标是检测任何整体上的非均匀性。文献 [Ben06], 研究了一种相关的污染模型，该模型拟合冯·米塞斯混合分布与均匀分布数据，并使用似然比和沃森型拟合优度程序，在均匀分布、冯·米塞斯分布和混合分布拟合之间进行选择。在这里，每个观测值均独立同分布地来自同一个混合分布。

相比之下，我们的设定有两点不同。第一，在备择假设下，只有一小部分未知的观测值（或社区模型中的边）具有结构，而大多数仍服从均匀分布。第二，我们关注问题规模和植入子集规模不断增长的高维情形，并探讨区分可检测性与不可能性的信息论阈值。经典的均匀性检验在这类稀疏植入情形中通常效力有限，因为即使存在相干子集，全局经验分布仍可能接近均匀分布。

除了假设检验之外，扫描统计量也在其他领域中得到研究。在圆周情形下，扫描统计量是弧沿圆周扫过时，任意给定长度的弧内所包含的观测值数量的最大值。Cressie [Cre77，Cre80]研究了这一统计量在圆周均匀性下的分布，专著 [GNW05]回顾了更广泛的发展与扩展。扫描统计量也用于

具有周期结构（例如季节性）的应用中，人们通过围绕一年扫描一个窗口来寻找局部脉冲事件 [WWG89，WNG93]。

非圆形数据中的植入结构 从高维检验的视角来看，我们的模型属于植入结构问题这一广泛类别：低复杂度信号（稀疏子集、社区或低秩扰动）嵌入高维噪声背景中。这类问题的典型例子包括植入团 [Jer92, AKS98, FK00],稀疏主成分分析 [dGJL04, JL09, DM14], 以及稀疏子矩阵检测 [BI13]; 有关此类模型中信息论和计算阈值的最新综述，参见 Wu 和 Xu [WX21]。在图设定中，与本文尤其相关的是稀疏随机块模型中社群检测的可检测性和恢复阈值，例如 [MNS18, Mas14, BMNN 16], ，以及综述 [Abb18]。作为一个特例，植入稠密子图问题研究的是嵌入均匀随机图中的异常稠密顶点子集的检测，并已从统计和计算视角进行了分析 [ACV14, VAC15, DM15, HWX17]。我们的工作可视为这些模型的方向性类比：我们以圆值测量取代标量观测（伯努利或高斯），并刻画植入簇和社区对应的检测与不可能区域。我们的社区模型也不同于植入稠密社区的随机块模型：在我们的模型中，无论哪种假设成立，边标签的边缘分布都相同。

另一类相关研究探讨植入稀疏结构问题中的统计—计算差距：在这类问题中，检测在信息论意义上是可行的，但据猜想，多项式时间算法难以解决。例如，Berthet 和 Rigollet [BR13] 通过从植入团出发进行平均情形归约，为稀疏主成分分析中存在此类差距提供了证据；Brennan、Bresler 和 Huleihel [BBH18]则将这一基于归约的研究计划拓展至多个典型植入模型，并为其得出了紧的条件下界。本文聚焦于方向型植入模型中的信息论可检测性与不可能性；将这些归约框架拓展到圆值观测，是未来研究的一个自然方向。

### 1.3 主要结果总结与讨论

信息论检测阈值 对于四种模型中的每一种，我们都推导出能够区分可检测性与不可能性的精确条件（除常数因子及偶尔出现的对数因子外），这些条件涉及参数组 $(N, K, \tau)$、$(N, K, \kappa)$、$(n, k, \tau)$ 和 $(n, k, \kappa)$。表1总结了特定情形下的这些结果；定理陈述中则列出了更一般的可实现性与不可能性条件。（函数 $A(\kappa)$ 和 $R(\kappa)$ 定义于附录 A。）我们的结果假设决策者不知道植入相位 $\Theta^*$；关于在平坦情形下，获取 $\Theta^*$ 如何提升性能的讨论，见附录 E；关于社群情形下与随机块模型之间联系的讨论，见附录 F。

我们的结果中反复出现一个小规模 $K$ 与大规模 $K$ 之间的二分现象。对于平坦模型，$K = \sqrt{N}$附近会发生定性变化：当 $K = o(\sqrt{N})$ 时，植入子集足够稀疏，因此要实现检测，簇内各方向必须变得越来越集中（等价地，在冯·米塞斯模型中，$\kappa \to \infty$；在硬聚类模型中，$\tau \to 0$）；而当 $K = \omega(\sqrt{N})$ 时，植入观测值的数量足以使检测即使在集中度恒定时也能实现（例如，$\kappa = \Theta(1)$，且窗口宽度 $\tau$ 远离 0 和 1）。相比之下，对于社区模型，相应的转变发生在规模小得多的相关集合处，大约为 $k = \log n$：如果 $k = o(\log n)$，那么强检测要求边级集中度随 $n$ 增强（同样是 $\kappa \to \infty$ 或 $\tau \to 0$）；而如果 $k = \omega(\log n)$，社区贡献的

表1：检测的充要条件总结

| 模型 | 情形 | 不可能性 | 可实现性 |
|---|---|---|---|
| **硬聚类 flat** $(N,\ K,\ \tau)$ | $K$ 常数 | $\tau = \omega\big(N^{-1-\frac{1}{K-1}}\big)$ | $\tau = o\big(N^{-1-\frac{1}{K-1}}\big)$ |
| | $K = N^{\alpha},\ \ \alpha \le 1/2$ | $\tau = \frac{K^2}{cN\,\text{对数}\,N}\ \ c < 1-2\alpha$ | $\tau = \frac{K^2}{cN\,\text{对数}\,N}\ \ c > 2$ |
| | $K = N^{\alpha},\ \ \alpha > 1/2$ | — | $\tau \in (0, 1-\epsilon],\ \epsilon > 0$ |
| **冯·米塞斯 flat** $(N,\ K,\ \kappa)$ | $K = N^{\alpha}, \alpha \le 1/2$ | $\frac{1}{\sqrt{\kappa}} \ge \frac{K^2}{cN\,\text{对数}\,N}\ \ (c < \frac{1-2\alpha}{2\sqrt{\pi}})$ | $\frac{1}{\sqrt{\kappa}} \le \frac{K^2}{cN\,\text{日志}\,N}\ \ (c > 0.5057)$ |
| | $K = N^{\alpha},\ \ \alpha > 1/2$ | — | $\kappa \ge \epsilon,\ \ \epsilon > 0$ |
| **硬聚类 社区** $(n,\ k,\ \tau)$ | $3 \le k \le n$ | — | $\tau \le \big(\frac{k}{ne}\big)^{\frac{(1+\epsilon)k}{\binom{k}{2}-1}}$ |
| | $k = c$对数 $n$ | $\tau > e^{-2/c}$ | $\tau \le e^{-2/c}$ |
| | $k = o(\sqrt{n})$ | $\tau \ge$指数 $\Big(-\frac{(2-\epsilon)\,\text{对数}\big(\frac{n}{k^2}\big)}{k-1}\Big)$ | $\tau \le \big(\frac{k}{ne}\big)^{\frac{(1+\epsilon)k}{\binom{k}{2}-1}}$ |
| | $k = \omega(\log n)$ | $\tau \ge 1 - \frac{2(1-\epsilon)\,k-1\,\text{对数}\big(1+\frac{\epsilon n}{k^2}\big)}{k-1}$ — | $\tau \le 1 - \frac{2(1+\epsilon)\ k-1\ \text{对数}\frac{ne}{k}}{k-1}$ — |
| **冯·米塞斯 社区** $(n,\ k,\ \kappa)$ | $3/\epsilon < k = o(\log n)$ | $\kappa \le \big(\frac{n}{k^2}\big)^{\frac{4-\epsilon}{k-1}}$ | $\kappa \ge \big(\frac{4}{\pi^2}+\epsilon\big)\big(\frac{n}{k}\big)^{\frac{4(1+\epsilon)}{k-1}} \log k$ |
| | $k = c\log\ n$ | $c$对数 $R(\kappa) < 2$ | $c\,A^2(\kappa) > 2$ 和 $c > 2$, 或 $\kappa \ge \frac{2\,\text{对数}\ k}{1-\text{余弦}(\pi\ \text{指数}(-2/c))}$ |
| | $k = \omega($对数 $n)$ | $\kappa \le \sqrt{\frac{4(1-2c-\epsilon)\,\text{对数}\frac{n}{k-1}}{}}$, $k \le n^c$ 和 $c < 1/2$ | $\kappa \ge\ (1+\epsilon)\ \sqrt{8\,\text{对数}\frac{n}{k-1}}$ |

社区内边数量足够多，因此只需有界的集中度，检测便能成功。

在小 $k$ 情形下，硬聚类分布与冯·米塞斯分布非常接近：当$\kappa$ 较大时，冯·米塞斯分布（$\Theta^*,\ \kappa$）可用均值为 $\Theta^*$ 、标准差为 $\frac{1}{\sqrt{\kappa}}$ 的高斯分布很好地近似，由此得到启发式对应关系 $\tau \approx \kappa^{-1/2}$，其中 $\kappa \to \infty$. 。这种对应关系有助于解释为何硬聚类模型与冯·米塞斯模型中会平行地出现相同的基于扫描的过程和阈值。

在大 $k$ 情形下，硬聚类分布与冯·米塞斯分布也相互对应，此时 $\tau \to 1$且 $\kappa \to 0$；也就是说，$[0, 2\pi\tau]$ 上的均匀分布和离散参数为 $\kappa$ 的冯·米塞斯分布都收敛到 $[0, 2\pi]$. 上的均匀分布。比较这两种社区模型的可实现性条件，可以推得对应关系 $1-\tau \approx \kappa^2/4$. 。此情形下的 KL 散度满足

$$\mathrm{D_{KL}}(\text{uniform}([0, 2\pi\tau])||\text{uniform}([0, 2\pi]) = -\ln\tau = 1-\tau+o(1-\tau)$$

and

$$\mathrm{D_{KL}}(\text{vonMises}(0,\kappa)||\text{uniform}([0, 2\pi]) = \kappa^2/4 + o(\kappa^2),$$

这表明存在相同的对应关系。

**方法论** 我们的可达性结果通过分析区间检验和相干性检验，并结合针对虚警概率和漏检概率的精细尾界得出。逆命题结果则依赖二阶矩方法和对似然比的控制，以及超几何分布与二项分布重叠分布之间的凸序比较。在冯·米塞斯模型中，分析采用修正贝塞尔函数的渐近性质，并在大集中度情形下对冯·米塞斯分布进行高斯近似。这些技术展示了如何以统一的方式分析方向几何与植入式组合结构。

## 2 硬聚类平坦模型

本节及接下来的三个章节中，首先以定理的形式给出主要结果，随后给出推论，将充要条件具体化为若干值得关注的情形，最后通过小节给出该定理的证明。

**定理 1.** 考虑参数为 $(N, K, \tau)$ 的硬聚类平坦模型 .

(a) *(*可实现性*)*在以下两种情形中的任意一种下，采用阈值为 $\gamma$ 的区间检验即可实现强检测：

（A1） $\gamma = K$ 和 $\frac{N^K \tau^{K-1}}{(K-1)!} \to 0$.

(A2) 存在序列 $c_N \to \infty$ 和 $\gamma \equiv \gamma_N$，满足

$$\gamma_N \triangleq (N-K)\tau + K - c_N\sqrt{(N-K)\tau}, \qquad \gamma_N \geq 1 + (N-1)\tau \text{ for all sufficiently large } N,$$

使得

$$\frac{(K-1)^2(1-\tau)^2}{(N-K)\tau} \to \infty, \tag{4}$$

$$\frac{\left(K(1-\tau) - c_N\sqrt{(N-K)\tau}\right)^2}{2N\tau + K(1-\tau) - c_N\sqrt{(N-K)\tau}} - \log N \to \infty. \tag{5}$$

(b) （逆命题）若 $\tau \leq 1/2$ 对于充分大的 $N$ 且

$$\frac{2N\tau^2}{K^2}\left(1 + \frac{K}{N\tau}\right)^{K+1} = o(1), \tag{6}$$

则弱检测不可能实现。

**推论 1**（可实现性情形）。 在硬聚类平坦模型下，以下各情形中均可通过区间检验实现强检测。

- $K \geq 2$ 固定且 $\tau = o\left(N^{-1-\frac{1}{K-1}}\right)$ [取 $\gamma = K$，使得 ***(A1)*** 得以满足。]
- $K = N^\alpha$ 其中 $0 < \alpha \leq 1/2$，且 $\tau = \frac{K^2}{(2+\epsilon)N\log N} = \frac{N^{2\alpha-1}}{2+\epsilon}$（）log $N$ 对于某个常数 $\epsilon \in$（0，1/2）。这里，阈值 $\gamma_N$ 的选择使得 $c_N \to \infty$ 和 $c_N = o\left(\sqrt{\log N}\right)$。
- $K = N^\alpha$ 满足 $1/2 < \alpha < 1$，且 $\tau \in (0,\ 1-\epsilon]$ 其中某个常数 $\epsilon \in (0,\ 1)$。这里，阈值 $\gamma_N$ 的选取使得 $c_N \to \infty$ 且 $c_N = o\left(K/\sqrt{N\tau}\right)$。

**推论 2** (不可能性情形)。 在硬聚类平坦模型下，在以下每种情形中，弱检测都是不可能的。

- $K \geq 2$ 固定且 $\tau = \omega\big(N^{-1-\frac{1}{K-1}}\big)$。 [这种缩放意味着定理（6）中的条件1（b）。]
- $K = N^{\alpha}$ 其中 $0 < \alpha < 1/2$，并且对于任意常数 $\epsilon \in (0,\ 1-2\alpha)$，$\tau = \epsilon N^{\frac{K^2}{N}} = \epsilon N^{\frac{N^{2\alpha-1}}{N}}$ 对数 $N$ 对数。 [对于这一对 $(K,\ \tau)$，

$$\frac{2N\tau^2}{K^2} = \frac{2}{\epsilon^2 \log^2 N} N^{2\alpha-1}, \qquad \frac{K(K+1)}{N\tau} = \epsilon \log N\,(1+N^{-\alpha}),$$

因此

$$\frac{2N\tau^2}{K^2}\left(1+\frac{K}{N\tau}\right)^{K+1} \leq \frac{2N\tau^2}{K^2}\exp\left\{\frac{K(K+1)}{N\tau}\right\} = \frac{2+o(1)}{\epsilon^2\log^2 N} N^{2\alpha+\epsilon-1} \to 0,$$

因为 $\epsilon < 1-2\alpha$。]

## 2.1 硬聚类平坦模型的可实现性——区间检验

（a）固定一个整数阈值 $\gamma \geq K$，并考虑对统计量（1）进行阈值判定的区间检验，即如果存在一个长度为 $H_1$ 且至少包含 $2\pi\tau$ $\gamma$ 个 $X_i$，则作出 $H_0$ 判定，否则作出 $\theta$ 判定。该决策规则等价于广义最大似然判决规则：将给定 $\Theta^* = \theta$时条件似然关于 $\theta$ 的最大值与某个阈值进行比较；参见附录B.1。

情形 *(A1)*： $\gamma = K$。此处漏检概率为零：在 $H_1$ 下，区间$[\Theta^*, \Theta^* + 2\pi\tau]$ 包含由 $S^*$索引的所有 $K$个点，因此扫描统计量至少为 $K$和 $p_{\text{miss}} = 0$。剩下的任务是在 $H_0$下界定 $p_{\text{fa}}$。

我们如下界定虚警概率 $p_{\text{fa}}$。令 $B_i$ 表示区间$[X_i, X_i + 2\pi\tau]$ 包含至少 $\gamma$个点（包括其自身）这一事件。那么 $p_{\text{fa}} = \mathbb{Q}(\cup_{i\in[N]} B_i)$，因此根据并集界以及对所有 $i$均有 $\mathbb{Q}(B_i) = \mathbb{Q}(B_1)$这一事实，

$$p_{\text{fa}} \leq N \cdot \mathbb{Q}(B_1).$$

在 $\mathbb{Q}$下， $\{j \in \{2,\dots,N\}:\ X_j \in [X_1, X_1 + 2\pi\tau]\}$的基数服从 $\text{binom}(N-1,\ \tau)$ 二项分布，因此 $[X_1, X_1 + 2\pi\tau]$ 中的观测值总数服从 $1 + \text{binom}(N-1,\ \tau)$ 分布。事件 $B_1$是所有基数为 $\gamma - 1$ 的子集$S \subset \{2,\dots,N\}$ 对应事件 $\{X_j \in [X_1, X_1 + 2\pi\tau]$（其中 $j \in S\}$）的并集。再次应用并集界，

$$\mathbb{Q}(B_1) \leq \binom{N-1}{\gamma-1}\tau^{\gamma-1}.$$

因此，

$$p_{\text{fa}} \leq N\binom{N-1}{\gamma-1}\tau^{\gamma-1} = \binom{N}{\gamma}\gamma\,\tau^{\gamma-1} \leq \frac{N^{\gamma}\,\tau^{\gamma-1}}{(\gamma-1)!}. \tag{7}$$

将其具体化为 $\gamma = K$，这表明只要 $\frac{N^K \tau^{K-1}}{(K-1)!} \to 0$。至此，情形（A1）得证。

情形（*A2*）： $\gamma = \gamma_N$。如果 $K$ 随着 $N$ 增大，那么 $\tau$ 的临界值可能大到在 $\mathbb{P}$ 条件下，区间 $[\Theta^*, \Theta^* + 2\pi\tau]$ 不仅包含由 $S^*$ 索引的 $K$ 点，而且，

以很高的概率还包含数量显著的其他点。这促使我们使用大于 $K$ 的阈值$\gamma$。虚警概率仅取决于 $N$, $\tau$, 和 $\gamma$, 而与 $K$ 无关。

我们首先利用切尔诺夫界对 $p_{\mathrm{fa}}$ 进行界定，该界适用于 $\gamma$ 至少为$\mathbb{Q}$下的均值的情况。如上所述， $p_{\mathrm{fa}} \le N\,\mathbb{Q}(B_1)$。令 $Y \triangleq 1+\mathrm{binom}(N-1,\tau)$ 表示在 $\mathbb{Q}$下落入 $[X_1, X_1+2\pi\tau]$的点数，其均值为 $\mu = 1+(N-1)\tau$。由标准切尔诺夫界可得$P\{Y \ge (1+\delta)\mu\} \le \exp\left(-\mu\delta^2/(2+\delta)\right)$，因此，对于 $\gamma \ge 1+(N-1)\tau$,

$$\mathbb{Q}(B_1) \;=\; P\{1+\mathrm{binom}(N-1,\tau) \ge \gamma\} \;\le\; \exp\left\{-\frac{(\gamma-1-(N-1)\tau)^2}{1+(N-1)\tau+\gamma}\right\}.$$

因此，对于 $\gamma \ge 1+(N-1)\tau$,

$$p_{\mathrm{fa}} \;\le\; N\cdot\exp\left\{-\frac{(\gamma-1-(N-1)\tau)^2}{1+(N-1)\tau+\gamma}\right\}. \tag{8}$$

接下来，我们对漏检概率进行上界估计。假设 $H_1$ 成立，因此点 $(X_i:\ i\in S^*)$ 全部落在区间 $[\Theta^*, \Theta^*+2\pi\tau]$内。如果至少有 $\gamma-K$ 个点 $(X_i:\ i\in[N]\setminus S^*)$ 也落在该区间内，就会检测到。此类点的数量服从 $\mathrm{binom}(N-K,\tau)$ 分布，因此假设 $(N-K)\tau \ge \gamma-K$,

$$p_{\mathrm{miss}} \;\le\; P\{\mathrm{binom}(N-K,\tau) < \gamma-K\} \;\le\; \exp\left\{-\frac{((N-K)\tau-\gamma+K)^2}{2(N-K)\tau}\right\}, \tag{9}$$

其中我们使用了切尔诺夫界 $P\{$二项分布$(\bar{n},\ p)\ \le(1-\delta)\bar{n}p\} \le e^{-\bar{n}p\delta^2/2}$，其中 $\bar{n}=N-K$、$p=\tau$，以及 $\delta = \frac{(N-K)\tau-\gamma+K}{(N-K)\tau}$。现在假设存在一个序列 $c_N\to\infty$ 使得

$$\frac{(K-1)^2(1-\tau)^2}{(N-K)\tau} \to \infty, \tag{10}$$

$$\frac{\left(K(1-\tau)-c_N\sqrt{(N-K)\tau}\right)^2}{2N\tau+K(1-\tau)-c_N\sqrt{(N-K)\tau}} - \log N \to \infty. \tag{11}$$

选择阈值

$$\gamma = \gamma_N \triangleq (N-K)\tau + K - c_N\sqrt{(N-K)\tau},$$

并假设（如定理陈述中所述） $\gamma_N \ge 1+$ （$N-1$）$\tau$对于所有充分大的 $N$，从而（8）适用。作此选择后，对于 $\gamma$，我们有（$N-K$）$\tau-\gamma+K = c_N\sqrt{}$（$N-K$）$\tau$，而（9）推出

$$p_{\mathrm{miss}} \le \exp\left(-c_N^2/2\right) \to 0\,.$$

此外，在（8）中使用 $\gamma=\gamma_N$，并注意到 $\gamma_N-1-(N-1)\tau = K(1-\tau)-c_N\sqrt{(N-K)\tau}-(1-\tau)$ 以及 $1+(N-1)\tau+\gamma_N = 2N\tau+K(1-\tau)-c_N\sqrt{(N-K)\tau}+O(1)$，条件（11）给出 $p_{\mathrm{fa}}\to 0$。至此，情形（A2）得证，从而定理1(a)得证。

## 2.2 硬聚类平坦模型的逆命题

我们将采用二阶矩方法来确定满足 $\mathrm{TV}(\mathbb{P},\ \mathbb{Q})\to 0$ 的条件，其中 $\mathrm{TV}(\mathbb{P},\ \mathbb{Q})$ 是 $\mathbb{P}$ 与 $\mathbb{Q}$之间的全变差距离，从而证明弱检测的不可能性。该方法利用了这样一个事实： $\mathbb{P}$ 是一个分布混合

它们各自相对于 $\mathbb{Q}$. 都有一个简单的似然比。该方法见 [WX21]，为方便读者，我们在此详细说明。关键在于对 $\mathrm{Var}_{\mathbb{Q}}(L)$进行界定，其中$L(X) = \mathrm{d}\mathbb{P}\,\mathrm{d}\mathbb{Q}(X)$，这也称为 $\chi^2$ 与 $\mathbb{P}$ 之间的散度，即 $\mathbb{Q}$ [Tsy09]。由于 $\mathbb{P} \ll \mathbb{Q}$，因此可得

$$\mathrm{TV}(\mathbb{P},\mathbb{Q}) = \frac{1}{2}\,\mathbb{E}_{\mathbb{Q}}|L-1| \le \frac{1}{2}\sqrt{\mathrm{Var}_{\mathbb{Q}}(L)}\,.$$

因此，如果 $\mathrm{Var}_{\mathbb{Q}}(L) \to 0$ as $N \to \infty$，那么 $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \to 0$，且弱检测不可能实现。由于 $E_{\mathbb{Q}}[L] = 1$，可得 $\mathrm{Var}_{\mathbb{Q}}(L) = \mathbb{E}_{\mathbb{Q}}[L^2] - 1$.
对 $(S^*, \Theta^*)$ 的可能取值取平均后，可写为

$$L(X) = \frac{1}{2\pi}\int_0^{2\pi} \frac{1}{\binom{N}{K}} \sum_{S\subset[N]\,:\,|S|=K} L_{S,\theta}(X)\,\mathrm{d}\theta\,, \tag{12}$$

其中

$$L_{S,\theta}(X) \triangleq \tau^{-K}\,\mathbf{1}\big\{X_i \in [\theta, \theta+2\pi\tau] \text{ for all } i \in S\big\}$$

是给定 $S^* = S$ 和 $\Theta^* = \theta$时的条件似然比。如果我们将 $S$ 和 $\Theta$ 视为与 $X$ 独立的随机变量，使得 $S$ 在 $[N]$ 的大小为 $|S| = K$ 的子集上均匀分布，$\Theta$ 在 $[0, 2\pi]$上均匀分布，并且 $S$ 和 $\Theta$ 相互独立，那么式（12）可以简洁地改写为 $L(X) = \mathbb{E}_{S,\,\Theta}[L_{S,\Theta}(X)]$. 此外，若 $(S', \Theta')$ 是 $(S, \Theta)$ 的一个独立副本，则

$$L^2(X) = \mathbb{E}_{S,\Theta}[L_{S,\Theta}(X)]\mathbb{E}_{S',\Theta'}[L_{S',\Theta'}(X)] = \mathbb{E}_{S,\Theta,S',\Theta'}[L_{S,\Theta}(X)L_{S',\Theta'}(X)]$$

然后，对服从分布 $\mathbb{Q}$ 的 $X$ 取期望，并交换积分次序，可得:

$$\mathbb{E}_{\mathbb{Q}}[L^2] = \mathbb{E}_{S,\Theta,S',\Theta'}\left[\mathbb{E}_{Q}[L_{S,\Theta}(X)L_{S',\Theta'}(X)]\right]. \tag{13}$$

方程（13）是本文本节及其他章节中用于计算 $\mathbb{E}_{\mathbb{Q}}[L^2]$和 $\mathrm{Var}_{\mathbb{Q}}(L) = \mathbb{E}_{\mathbb{Q}}[L^2] - 1$.

固定 $S, S', \theta, \theta'$，并令 $j = |S \cap S'|$。令 $2\pi\delta$ 表示两条弧$[\theta, \theta+2\pi\tau]$和 $[\theta', \theta'+2\pi\tau]$（模 $2\pi$）的交集长度。在 $\mathbb{Q}$ 条件下，坐标 $X_i$ 独立同分布均匀地取自 $[0, 2\pi)$，因此

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}\big[L_{S,\theta}(X)\cdot L_{S',\theta'}(X)\big] &= \tau^{-2K}\cdot\mathbb{Q}\left\{\begin{array}{rl} X_i \in [\theta,\theta+2\pi\tau]\cap[\theta',\theta'+2\pi\tau] & \text{for } i \in S\cap S',\\ X_i \in [\theta,\theta+2\pi\tau] & \text{for } i \in S\backslash S',\\ X_i \in [\theta',\theta'+2\pi\tau] & \text{for } i \in S'\backslash S\end{array}\right\}\\ &= \tau^{-2K}\,\delta^{\,j}\,\tau^{2(K-j)} \;=\; \delta^{\,j}\,\tau^{-2j}\,.\end{aligned}$$

接下来，我们对 $\theta, \theta'$ 求平均。令 $u \in [0, 1/2]$ 表示 $\theta$ 与 $\theta'$ 之间的圆周距离，并以 $2\pi$归一化，即

$$u \triangleq \frac{1}{2\pi}\min\big\{|\theta-\theta'|,\, 2\pi - |\theta-\theta'|\big\}, \qquad u \sim \mathrm{uniform}\left(\left[0, \frac{1}{2}\right]\right).$$

因此，归一化重叠 $\delta$ 是由下式给出的函数 $\delta = \delta_\tau(u)$:

$$\delta_\tau(u) = \begin{cases}(\tau-u)_+, & 0 < \tau \le 1/2\,,\\ (2\tau-1) + (1-\tau-u)_+, & 1/2 < \tau < 1\,,\end{cases} \tag{14}$$

因此，对于每个整数 $j \geq 1$,

$$\mathbb{E}_{\theta,\theta'}[\delta^j] = 2\int_0^{1/2} \delta_\tau(u)^j \,\mathrm{d}u = \begin{cases} \frac{2}{j+1}\,\tau^{j+1}, & 0 < \tau \leq \frac{1}{2} \\ \frac{2}{j+1}\Big(\tau^{j+1} - (2\tau-1)^{j+1}\Big) + 2\big(\tau - \frac{1}{2}\big)(2\tau-1)^j, & \frac{1}{2} < \tau < 1 \end{cases} \tag{15}$$

现在也对 $S, S'$ 取平均。记 $J \triangleq |S \cap S'|$，得到

$$\mathbb{E}_{\mathbb{Q}}[L^2] = \mathbb{P}\{J=0\} + \sum_{j=1}^{K} \mathbb{P}\{J=j\} \cdot \tau^{-2j} \cdot \mathbb{E}_{\theta,\theta'}\left[\delta^j\right], \tag{16}$$

且 $J$服从参数为（$N, K, K$）的超几何分布。利用 $\mathrm{Var}_{\mathbb{Q}}(L) = \mathbb{E}_{\mathbb{Q}}[L^2] - 1$可得

$$\mathrm{Var}_{\mathbb{Q}}(L) = \sum_{j=1}^{K} \mathbb{P}\{J=j\}\left(\tau^{-2j} \cdot \mathbb{E}_{\theta,\theta'}[\delta^j] - 1\right) \leq \sum_{j=1}^{K} \mathbb{P}\{J=j\} \cdot \tau^{-2j} \cdot \mathbb{E}_{\theta,\theta'}\left[\delta^j\right].$$

在证明的剩余部分，我们将讨论特例 $0 < \tau \leq 1/2$ 。将（15）代入（16），

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[L^2] &= \mathbb{P}\{J=0\} + \sum_{j=1}^{K} \mathbb{P}\{J=j\}\left(\tau^{-2j} \cdot \frac{2}{j+1}\tau^{j+1}\right) \\ &= \mathbb{P}\{J=0\} + \sum_{j=1}^{K} \mathbb{P}\{J=j\}\,\frac{2}{(j+1)\tau^{j-1}}, \end{aligned}$$

因此

$$\mathrm{Var}_{\mathbb{Q}}(L) \leq \sum_{j=1}^{K} \mathbb{P}\{J=j\}\,\frac{2}{(j+1)\tau^{j-1}}. \tag{17}$$

现在我们将重叠分布与二项分布进行比较。超几何分布（$N, K, K$）的离散程度在二阶随机意义下小于二项分布 $\left(K, \frac{K}{N}\right)$；参见 Hoeffding [Hoe63]。由于$y \mapsto z^y$ 对每个固定的 $z \geq 0$都是凸函数，因此这意味着对所有 $z \geq 0$,

$$\sum_{j=0}^{K} \mathbb{P}\{J=j\} \cdot z^j \leq \left(1 - \frac{K}{N} + \frac{Kz}{N}\right)^K.$$

对 $z \in [0, r]$ 关于 $r > 0$ 积分，得到

$$\begin{aligned} \sum_{j=0}^{K} \mathbb{P}\{J=j\}\,\frac{r^{j+1}}{j+1} &\leq \int_0^r \left(1 - \frac{K}{N} + \frac{Kz}{N}\right)^K \mathrm{d}z = \frac{N}{K(K+1)}\left(1 - \frac{K}{N} + \frac{Kz}{N}\right)^{K+1}\Bigg|_{z=0}^{r} \\ &\leq \frac{N}{K^2}\left(1 + \frac{Kr}{N}\right)^{K+1}. \end{aligned}$$

去掉左侧的 $j=0$项，令 $r = 1/\tau$，两边乘以 $2\tau^2$，并利用（17）可得

$$\mathrm{Var}_{\mathbb{Q}}(L) \leq \sum_{j=1}^{K} \mathbb{P}\{J=j\}\,\frac{2}{(j+1)\tau^{j-1}} \leq \frac{2N\tau^2}{K^2}\left(1 + \frac{K}{N\tau}\right)^{K+1} \tag{18}$$

因此，只要 $(\mathbb{P}, \mathbb{Q}) \leq \frac{1}{2}\sqrt{\mathrm{Var}_{\mathbb{Q}}(L)} \to 0$的右侧趋于零，弱检测在该情形下便不可能。这证明了定理1(b)。

## 3 von米塞斯平坦模型

**定理 2**。考虑参数为$(N, K, \kappa)$的冯·米塞斯平坦模型。对于 $\tau \in (0,\ 1)$，定义

$$p_\kappa(\tau) \triangleq \mathbb{P}\{\mathrm{vonMises}(0,\kappa) \in [-\pi\tau, \pi\tau]\}\,, \qquad g \triangleq K\big(p_\kappa(\tau) - \tau\big)\,.$$

(a) （可实现性）假设存在序列 $c_N \to \infty$ 和 $\gamma_N$，满足

$$\gamma_N \triangleq N\tau + g - c_N\sqrt{N\tau + g}, \qquad \gamma_N \geq 1 + (N-1)\,\tau \ \textit{for all sufficiently large } N,$$

使得

$$\frac{g^2}{N\tau + g} \to \infty\,, \tag{19}$$

$$\frac{\big(g - c_N\sqrt{N\tau+g}\big)^2}{2N\tau + g - c_N\sqrt{N\tau+g}} - \log N \ \to\ \infty\,. \tag{20}$$

那么，区间检验 采用窗口长度 $2\pi\tau$ 和阈值 $\gamma_N$ 可实现强检测 .

(b) （逆命题）令 $R(\kappa) \triangleq I_0(2\kappa)/I_0^2(\kappa)$，其中 $I_0$ 是第一类修正贝塞尔函数。如果

$$\frac{K^2}{N}\big(R(\kappa) - 1\big) - \log R(\kappa) \ \to\ -\infty, \tag{21}$$

因此，弱检测是不可能的。

**推论3**（可实现性情形）。 在冯·米塞斯平坦模型下，使用区间检验时，以下每种情形均可实现强检测。

- $K = N^\alpha$ 当 $0 < \alpha \leq 1/2$，且 $\frac{1}{\sqrt{\kappa}} = \frac{K^2}{c_1 N \log N}$ 对于常数 $c_1 > c_0$，其中 $e$

$$c_0 \triangleq \frac{2}{\pi}\min_{c_2>0}\frac{c_2}{\big(1 - 2\,Q(c_2)\big)^2} \approx 0.5057, \qquad c_2^\star \approx 0.7518 \ \textit{(a minimizer)}\,,$$

且 $Q$ 是标准高斯分布的互补累积分布函数。此处，检验区间选为$\tau = c_2^\star/(\pi\sqrt{\kappa})$，阈值 $\gamma_N$的选取满足 $c_N \to \infty$且 $c_N = o\big(\sqrt{\log\ N}\big)$. [对于较大的 $\kappa$，$\mathrm{vonMises}(0,\kappa) \approx \mathcal{N}(0, 1/\kappa)$，因此 $p_\kappa(\tau) \to 1 - 2Q(c_2^\star)$且 $g \sim K(1 - 2Q(c_2^\star))$。结合$\tau = c_2^\star/(\pi\sqrt{\kappa})$和 $1/\sqrt{\kappa} = K^2/(c_1 N \log\ N)$，可得

$$\frac{g^2}{2N\tau\log N} \sim \frac{c_1(1 - 2Q(c_2^\star))^2}{(2/\pi)c_2^\star} > 1\,,$$

因此 (19)–(20)成立。]

- $K = N^\alpha$其中 $1/2 < \alpha < 1$，且 $\kappa = \Omega(1)$。这里，检验区间选取为 $\tau \in [\epsilon, 1-\epsilon]$，其中某个常数 $\epsilon \in (0, 1/2)$，阈值 $\gamma_N$ 的选取使得 $c_N \to \infty$ 以及 $c_N = o\big(K/\sqrt{N}\big)$。 [于是 $p_\kappa(\tau) - \tau$与 0 保持正距离，因此 $g = \Theta(K)$ 以及 (19)–(20) 成立。]

**推论 4** (不可能区域)。 在冯·米塞斯平坦模型下，当 $K = N^\alpha$对某个常数 $0 < \alpha < 1/2$成立，且 $\frac{1}{\sqrt{\kappa}} = \frac{K^2}{c N^{\text{对数}\ N}}$对某个常数 $c < \frac{1-2\alpha}{2\sqrt{\pi}}$. [成立时，弱检测是不可能的。在

这种缩放下，我们有 $\kappa \to \infty$，因此 $R(\kappa) = I_0(2\kappa)/I_0^2(\kappa) \sim \sqrt{\pi\kappa}$ 并且 对数 $R(\kappa) \sim \frac{1}{2}$ 对数 $\kappa$。因此，逆命题条件 (21) 可由以下条件推出

$$\frac{K^2}{N}\sqrt{\pi\kappa} \le \frac{1}{2}\log\kappa - \omega(1).$$

当 $1/\sqrt{\kappa} = K^2/(cN$ 对数 $N)$ 时，左侧等于 $\sqrt{\pi}\, c^{-1}$ 对数 $N$，而右侧满足

$$\frac{1}{2}\log\kappa = \log(cN\log N/K^2) = (1-2\alpha)\log N + O(\log\log N).$$

因此，若 $c < (1-2\alpha)/(2\sqrt{\pi})$，该不等式有 $\omega(1)$ 的余量，且 (21) 得证。 ∎

### 3.1 冯·米塞斯平坦模型的可实现性——区间检验

固定 $\tau \in (0,1)$，并按定理陈述记 $p_\kappa(\tau)$ 和 $g$。我们采用窗口长度为 $2\pi\tau$、阈值为 $\gamma_N \triangleq N\tau + g - c_N\sqrt{N\tau+g}$. 的区间检验

虚警。在 $H_0$ 下，观测值在 $[0, 2\pi$ 上独立同分布均匀分布，与硬聚类平坦模型完全相同。因此，区间检验的虚警概率与该情形相同。特别地，与式（8）一样，只要 $\gamma_N \ge 1 + (N-1)\tau$，

$$p_{\mathrm{fa}} \le N \cdot \exp\left\{-\frac{\big(\gamma_N - 1 - (N-1)\tau\big)^2}{1 + (N-1)\tau + \gamma_N}\right\}. \tag{22}$$

注意到

$$\gamma_N - 1 - (N-1)\,\tau = g - c_N\sqrt{N\tau+g} - (1-\tau), \qquad 1 + (N-1)\,\tau + \gamma_N = 2N\tau + g - c_N\sqrt{N\tau+g} + O(1),$$

我们看到，条件（20）意味着（22）中的指数项大于 log $N$，因此 $p_{\mathrm{fa}} \to 0$。

漏检。在 $H_1$ 下，考虑一个长度为 $2\pi\tau$、以植入相位 $\Theta^*$ 为中心的检验区间，即 $[\Theta^* - \pi\tau, \Theta^* + \pi\tau]$（模 $2\pi$）。令 $Y$ 为落入此窗口的观测值数量。在 $H_1$ 下， $Y$ 分解为

$$Y = Y_0 + Y_1\,, \qquad Y_0 \sim \mathrm{binom}\big(K, p_\kappa(\tau)\big)\,, \qquad Y_1 \sim \mathrm{binom}\big(N-K, \tau\big)\,,$$

其中 $Y_0$ 和 $Y_1$ 相互独立。特别地，

$$\mu_1 \triangleq \mathbb{E}_{\mathbb{P}}[Y] = (N-K)\tau + Kp_\kappa(\tau) = N\tau + g.$$

由于只要 $Y \ge \gamma_N$，区间检验就会拒绝 $H_0$，因此我们有

$$p_{\mathrm{miss}} \le \mathbb{P}\{Y < \gamma_N\} = \mathbb{P}\{Y_0 + Y_1 < \gamma_N\}\,.$$

由于 $Y_0 + Y_1$ 是独立伯努利随机变量之和，我们可以应用标准的切尔诺夫下尾界（泊松二项界）。令 $\gamma_N = (1-\eta)\,\mu_1$，得到

$$p_{\mathrm{miss}} \le \exp\left(-\frac{\eta^2}{2}\mu_1\right) = \exp\left(-\frac{(\mu_1 - \gamma_N)^2}{2\mu_1}\right).$$

代入 $\mu_1 = N\tau + g$和 $\gamma_N = N\tau + g - c_N\sqrt{N\tau+g}$可得

$$p_{\mathrm{miss}} \le \exp\left(-\frac{c_N^2}{2}\right) \to 0\,, \tag{23}$$

因为 $c_N \to \infty$。结合 $p_{\mathrm{fa}} \to 0$ 可证明强检测。这证明了定理 2的（a）部分。

### 3.2 冯·米塞斯平坦模型的逆命题

我们对 $\mathrm{Var}_{\mathbb{Q}}(L)$ 进行界定，其中 $L = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}$。由于 $\mathbb{E}_{\mathbb{Q}}[L] = 1$ 且 $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \le \frac{1}{2}\sqrt{\mathrm{Var}_{\mathbb{Q}}(L)}$，只需证明 $\mathbb{E}_{\mathbb{Q}}[L^2] \to 1$。

对于固定的 $(S, \theta)$，条件似然比为

$$L_{S,\theta}(X) \;=\; I_0(\kappa)^{-K}\,\exp\!\left(\kappa \sum_{i\in S}\cos(X_i-\theta)\right). \tag{24}$$

固定 $S$，$\theta$，$S'$，$\theta'$，其中 $|S| = |S'| = K$，并令 $j = |S \cap S'|$。则

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\big[L_{S,\theta}(X)L_{S',\theta'}(X)\big] &= I_0(\kappa)^{-2K}\cdot \mathbb{E}_{\mathbb{Q}}\Big[\exp\Big(\kappa\sum_{i\in S}\cos(X_i-\theta)+\kappa\sum_{i\in S'}\cos(X_i-\theta')\Big)\Big] \\
&= I_0(\kappa)^{-2K}\cdot \mathbb{E}_{\mathbb{Q}}\Big[\exp\Big(\kappa\sum_{i\in S\cap S'}\big(\cos(X_i-\theta)+\cos(X_i-\theta')\big)\Big)\Big] \qquad (25)\\
&\qquad \times \mathbb{E}_{\mathbb{Q}}\Big[\exp\Big(\kappa\sum_{i\in S\setminus S'}\cos(X_i-\theta)+\kappa\sum_{i\in S'\setminus S}\cos(X_i-\theta')\Big)\Big] \\
&= I_0(\kappa)^{-2j}\cdot \mathbb{E}_{\mathbb{Q}}\left[\exp\big(\kappa\big(\cos(X_1-\theta)+\cos(X_1-\theta')\big)\big)\right]^{j},
\end{aligned}$$

其中，在（25）中我们利用了在 $\mathbb{Q}$下的独立性；在最后一步中，我们利用了 $X_i$ 在 $[0, 2\pi$上服从独立同分布均匀分布，因此每个交集项具有相同的期望。
利用恒等式 $\cos a + \cos b = 2\cos\left(\frac{a-b}{2}\right)\cos\left(\frac{a+b}{2}\right)$，我们得到

$$\cos(X_1-\theta)+\cos(X_1-\theta') = 2\cos\!\left(\frac{\theta-\theta'}{2}\right)\cdot\cos\!\left(X_1-\frac{\theta+\theta'}{2}\right),$$

因此

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\big[\exp\big(\kappa\big(\cos(X_1-\theta)+\cos(X_1-\theta')\big)\big)\big] &= \mathbb{E}_{\mathbb{Q}}\Big[\exp\Big(2\kappa\cos\Big(\frac{\theta-\theta'}{2}\Big)\cdot\cos\Big(X_1-\frac{\theta+\theta'}{2}\Big)\Big)\Big] \\
&= I_0\Big(2\kappa\cos\Big(\frac{\theta-\theta'}{2}\Big)\Big),
\end{aligned}$$

因为在 $\mathbb{Q}$ 下，$X_1 - (\theta+\theta')/2$ 在 $[0, 2\pi)$ 上服从均匀分布。因此

$$\mathbb{E}_{\mathbb{Q}}\Big[L_{S,\theta}(X)L_{S',\theta'}(X)\Big] = \left[\frac{I_0\left(2\kappa\cos\left(\frac{\theta-\theta'}{2}\right)\right)}{I_0^2(\kappa)}\right]^{j}. \tag{26}$$

为了计算 $\mathbb{E}_{\mathbb{Q}}[L^2]$，我们将 $S$、$\theta$、$S'$、$\theta'$ 视为在各自定义域上相互独立且服从均匀分布。令$\vartheta \triangleq (\theta-\theta')/2$，可得

$$\mathbb{E}_{\mathbb{Q}}[L^2] = \frac{1}{2\pi}\int_0^{2\pi}\sum_{j=0}^{K}\mathbb{P}\{|S\cap S'|=j\}\left[\frac{I_0(2\kappa\cos\vartheta)}{I_0^2(\kappa)}\right]^{j}\mathrm{d}\vartheta. \tag{27}$$

我们利用以下事实对总和进行上界估计：$j$ 其中 $|S\cap S'|$ 服从超几何分布$(N, K, K)$，且在二阶意义上受二项分布 $\big(K$支配，$\frac{K}{N}\big)$ (Hoeffding [Hoe63])。由于$z \mapsto z^j$ 在 $z \ge 0$上是凸函数，因此对所有 $z \ge 0$,

$$\sum_{j=0}^{K}\mathbb{P}\{|S\cap S'|=j\}\cdot z^j \;\le\; \sum_{j=0}^{K}\mathbb{P}\left\{\mathrm{binom}\left(K,\frac{K}{N}\right)=j\right\}\cdot z^j \;=\; \left(1-\frac{K}{N}+\frac{Kz}{N}\right)^{K}.$$

将此应用于 $z=\rho_\kappa(\vartheta)$，其中

$$\rho_\kappa(\vartheta) \triangleq \frac{I_0(2\kappa\cos\vartheta)}{I_0^2(\kappa)}\,,$$

然后利用 $1+x\le e^x$ 得到

$$\begin{aligned}\mathbb{E}_{\mathbb{Q}}[L^2] &\le \frac{1}{2\pi}\int_0^{2\pi}\left(1+\frac{K}{N}\big(\rho_\kappa(\vartheta)-1\big)\right)^K \mathrm{d}\vartheta \\ &\le \frac{1}{2\pi}\int_0^{2\pi}\exp\left(\frac{K^2}{N}\big(\rho_\kappa(\vartheta)-1\big)\right)\mathrm{d}\vartheta = e^{-\frac{K^2}{N}}\cdot\frac{1}{2\pi}\int_0^{2\pi}\exp\left(\frac{K^2}{N}\rho_\kappa(\vartheta)\right)\mathrm{d}\vartheta\,. \end{aligned} \tag{28}$$

令 $R(\kappa)\triangleq\max_\vartheta\rho_\kappa(\vartheta)=\rho_\kappa(0)=\frac{I_0(2\kappa)}{I_0^2(\kappa)}$。对于 $x\in[0,R(\kappa)]$ 和 $a>0$，$x\mapsto e^{ax}$ 在 $[0,R(\kappa)]$ 上的凸性意味着

$$e^{ax}\le\left(1-\frac{x}{R(\kappa)}\right)e^0+\frac{x}{R(\kappa)}e^{aR(\kappa)}=1+\frac{e^{aR(\kappa)}-1}{R(\kappa)}\,x\,.$$

将此结果用于 $a=K^2/N$ 和 $x=\rho_\kappa(\vartheta)$ 于 (28) 中，得到

$$\mathbb{E}_{\mathbb{Q}}[L^2]\le e^{-\frac{K^2}{N}}\cdot\frac{1}{2\pi}\int_0^{2\pi}\left(1+\frac{e^{\frac{K^2}{N}R(\kappa)}-1}{R(\kappa)}\,\rho_\kappa(\vartheta)\right)\mathrm{d}\vartheta\,. \tag{29}$$

**引理 1.** 对于每个 $\kappa>0$,

$$\frac{1}{2\pi}\int_0^{2\pi}\rho_\kappa(\vartheta)\,\mathrm{d}\vartheta=1\,.$$

证明。利用 $2\cos(a)\cos(b)=\cos(a+b)+\cos(a-b)$，可得

$$\begin{aligned}\frac{1}{2\pi}\int_0^{2\pi}I_0\big(2\kappa\cos\vartheta\big)\,\mathrm{d}\vartheta &= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi}\exp\Big(2\kappa\cos\vartheta\cos(\varphi-\vartheta)\Big)\,\mathrm{d}\varphi\,\mathrm{d}\vartheta \\ &= \frac{1}{(2\pi)^2}\int_0^{2\pi}\int_0^{2\pi}\exp\Big(\kappa\cos\varphi+\kappa\cos(2\vartheta-\varphi)\Big)\,\mathrm{d}\vartheta\,\mathrm{d}\varphi \\ &= I_0^2(\kappa)\,,\end{aligned}$$

其中，最后一步两次用到了 $I_0(\kappa)$ 的积分表示。两边同时除以 $I_0^2(\kappa)$，即可证明该结论。

□

应用引理1于（29）可得

$$\mathbb{E}_{\mathbb{Q}}[L^2]\;\le\;e^{-\frac{K^2}{N}}\left(1+\frac{e^{\frac{K^2}{N}R(\kappa)}-1}{R(\kappa)}\right)\;=\;\frac{e^{\frac{K^2}{N}(R(\kappa)-1)}}{R(\kappa)}+\left(1-\frac{e^{-\frac{K^2}{N}}}{R(\kappa)}\right).$$

因此，$\mathbb{E}_{\mathbb{Q}}[L^2]\to1$（从而 $\mathrm{Var}_{\mathbb{Q}}(L)\to0$ 以及 $\mathrm{TV}(\mathbb{P},\,\mathbb{Q})\to0$）的一个充分条件是

$$\frac{K^2}{N}\big(R(\kappa)-1\big)-\log R(\kappa)\to-\infty\,,$$

这恰好是（21）。这证明了定理 2 的（b）部分。

## 4 硬簇社区模型

**定理 3**。 考虑参数为 $(n, k, \tau)$ 的硬簇社区模型。

(a) （可实现性 性）当窗口长度为 $2\pi\tau$ 且阈值为 $k$ 时，区间检验得出 $p_{\text{miss}} = 0$ *an* $d$

$$p_{\text{fa}} \le \exp\left(\log\frac{\binom{k}{2}}{\tau} + k\left\{\log\frac{n}{k} + 1 + \frac{k-1}{2}\log\tau\right\}\right). \tag{30}$$

因此，如果 (30) 的右侧趋于零，则可实现强检测。

(b) （逆命题）如果

$$\frac{k^2}{n}\left(\left(\frac{1}{\tau}\right)^{\frac{k-1}{2}} - 1\right) \to 0 \qquad as\ n \to \infty\,, \tag{31}$$

那么弱检测是不可能的。

**推论 5**（可实现性情形）。 在硬簇社区模型下，以下每个条件都意味着区间检验能够实现强检测：

- $k = c\log n$ *for a constant* $c > 0$ *and* $\tau \le e^{-2/c}$.
- $3/\epsilon < k \le n$ *and* $\tau \le \left(\frac{k}{ne}\right)^{\frac{2+\epsilon}{k-1}}$ *for a fixed* $\epsilon \in (0,1)$.
- $k = \omega(\log n)$ *and* $\tau \le 1 - \frac{2(1+\epsilon)}{k-1}\log\frac{ne}{k}$
- $3 \le k \le n$ *and* $\tau = \left(\frac{k}{ne}\right)^{\frac{(1+\epsilon)k}{\binom{k}{2}-1}}$ *for a fixed* $\epsilon > 0$.
- $3/\epsilon < k = o(n)$ *and* $\tau = \tau_\epsilon \triangleq \left(\frac{k}{n}\right)^{\frac{2+\epsilon}{k-1}}$ *for some* $\epsilon \in (0,1)$.

推论 5的证明。 当 $k = c$对数 $n$ 且 $\tau = e^{-2/c}$时，界限（30）变为

$$\begin{aligned} p_{\text{fa}} &\le \exp\left(\log\frac{\binom{k}{2}}{\tau} + c\log n\left\{\log\frac{n}{k} + 1 + \frac{c\log n - 1}{2}\left(-\frac{2}{c}\right)\right\}\right) \\ &= \frac{1}{\tau}\exp\left(\log\binom{k}{2} + (\log n)(-c\log k + c + 1)\right) \to 0\,. \end{aligned}$$

当 $\tau = \left(\frac{k}{ne}\right)^{\frac{2+\epsilon}{k-1}}$时，界限（30）变为：

$$\begin{aligned} p_{\text{fa}} &\le \exp\left(\frac{2+\epsilon}{k-1}\log\frac{n}{k} + \log\binom{k}{2} + k\left\{-\frac{\epsilon}{2}\log\frac{n}{k} - \frac{\epsilon}{2}\right\}\right) \\ &= \exp\left(\left[-\frac{k\epsilon}{2} + \frac{2+\epsilon}{k-1}\right]\log\frac{n}{k} + \log\binom{k}{2} - \frac{\epsilon k}{2}\right) \\ &\to 0 \text{ if } 3/\epsilon < k \le n\,. \end{aligned}$$

对于 $\tau \le 1 - \frac{2}{(1+\epsilon)k-1}\log\frac{ne}{k}$ 和 $k = \omega(\log n)$，因为

$$\frac{2(1+\epsilon)}{k-1}\log\frac{ne}{k} \to 0\,, \quad \text{and} \quad \frac{2(1+\epsilon)}{k-1}\log\frac{ne}{k} = \left(\frac{1+\epsilon}{1+\epsilon/2}\right)\frac{2+\epsilon}{k-1}\log\frac{ne}{k}\,,$$

由此可知 $1 - {}^{2}_{(1+\epsilon)\,k-1}\log\ \frac{ne}{k} \le \binom{k}{ne}^{\frac{2+\epsilon}{k-1}}$，所以根据前一种情况。

对于 $\tau = \binom{k}{ne}\binom{k}{2}^{-1}$ 以及 $3 \le k \le n$ 界（30）变为：

$$p_{\mathrm{fa}} \le \exp\left(\log\binom{k}{2} - \epsilon k \log\frac{ne}{k}\right) \to 0\,.$$

对于 $\tau = \tau_\epsilon \triangleq \binom{k}{n}^{\frac{2+\epsilon}{k-1}}$ 界（30）变为：

$$\begin{aligned} p_{\mathrm{fa}} &\le \exp\left(\frac{2+\epsilon}{k-1}\log\frac{n}{k} + \log\binom{k}{2} + k\left\{-\frac{\epsilon}{2}\log\frac{n}{k} + 1\right\}\right) \\ &= \exp\left(\left[-\frac{k\epsilon}{2} + \frac{2+\epsilon}{k-1}\right]\log\frac{n}{k} + \log\binom{k}{2} + k\right) \\ &\to 0 \text{ if } k = o(n) \text{ and } k > 3/\epsilon\,. \end{aligned}$$

□

**推论6** (不可能性情形)。 在硬簇社区模型下，以下各条件均意味着弱检测不可能：

- $k = c$对数 $n$ 对于常数 $c > 0$ 和常数 $\tau > e^{-2/c}$;
- $k = o(\sqrt{n})$ 以及 $\tau \ge \exp\left(\frac{-(2-\epsilon)\log(n/k^2)}{k-1}\right)$ 对于常数 $\epsilon > 0$;
- $k = \omega(\log n)$ 和 $\tau \ge 1 - {}^{2}_{(1-\epsilon)\,k-1}\log\left(1 + \frac{\epsilon_n n}{k^2}\right)$ 对于正数常数 $\epsilon$ 和 $\epsilon_n$，且满足 $\epsilon_n \to 0$; [这些条件意味着 $\tau \to 1$，因此对于充分大的 $n$，有 $\tau \ge \exp\left\{{}^{2}_{k-1}\log\left(1 + \frac{\epsilon_n n}{k^2}\right)\right\}$，其中 $\exp\left\{{}^{2}_{k-1}\log\left(1 + \frac{\epsilon_n n}{k^2}\right)\right\}$ 是将（31）左侧设为 $\epsilon_n$.]后得到的 $\tau$值

### 4.1 硬聚类社区模型的区间检验可实现性

在 $H_1$ 下，对于植入对 $(C^*, \Theta^*)$，对所有 $i, j \in C^*$, $i \ne j$，我们有 $X_{ij} \in [\Theta^*, \Theta^* + 2\pi\tau]$。因此，该检验以 $\theta = \Theta^*$ 和 $C = C^*$ 成功，故 $p_{\mathrm{miss}} = 0$.

现在我们计算 $\ell = 2\pi\tau$ 下区间检验的虚警概率。在 $H_0$ 下，边标签 $(X_{ij})_{1 \le i < j \le n}$ 在 $[0, 2\pi)$ 上独立同分布均匀。对于一个固定的子集 $C \subset [n]$，满足 $|C| = k$，令

$$E_C \triangleq \left\{\exists\,\theta \in [0, 2\pi) :\ X_{ij} \in [\theta, \theta + 2\pi\tau]\ \ \forall\, i, j \in C,\ i \ne j\right\}.$$

对所有社区应用并集界，

$$p_{\mathrm{fa}} = \mathbb{Q}\left(\bigcup_{C:|C|=k} E_C\right) \le \binom{n}{k}\mathbb{Q}(E_C), \tag{32}$$

其中 $\mathbb{Q}(E_C)$ 不依赖于 $C$ 的具体选择。

为了在扫描 $\theta$ 时界定 $\mathbb{Q}(E_C)$，注意到如果发生 $E_C$，那么特别地，所有 $\binom{k}{2}$ 角度$\{X_{ij} :\ i < j,\ i,\ j \in C\}$ 都位于某个长度为 $2\pi\tau$ 的弧内。令 $e^\star$ 表示 $C$ 中在这些 $m$ 角度中取得最小角度的一条边。那么 $E_C$ 意味着所有其余的 $\binom{k}{2} - 1$ 角度都位于

弧 $[X_{e^\star}, X_{e^\star} + 2\pi\tau]$ 内。对最小角度边的选择应用并集界，可得$\mathbb{Q}(E_C) \le \binom{k}{2}\tau^{\binom{k}{2}-1}$.因此，

$$\begin{aligned} p_{\mathrm{fa}} &\le \binom{k}{2}\binom{n}{k}\tau^{\binom{k}{2}-1} \le \binom{k}{2}\left(\frac{ne}{k}\right)^k \tau^{\binom{k}{2}-1} \\ &= \exp\left(\log\frac{\binom{k}{2}}{\tau} + k\left\{\log\frac{n}{k} + 1 + \frac{k-1}{2}\log\tau\right\}\right). \end{aligned}$$

这证明了定理3（a）。

### 4.2 硬簇社区模型的逆命题

我们给出 $\mathrm{Var}_{\mathbb{Q}}(L)$ 的上界，其中 $L(X) = \mathrm{d}\mathbb{P}/\mathrm{d}\mathbb{Q}(X)$ 且 $X = (X_{ij})_{1\le i<j\le n}$。由于 $\mathbb{P}$ 和 $\mathbb{Q}$ 绝对连续，$\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \le \frac{1}{2}\sqrt{\mathrm{Var}_{\mathbb{Q}}(L)}$,-因此若 $\mathrm{Var}_{\mathbb{Q}}(L) \to 0$（等价地，$\mathbb{E}_{\mathbb{Q}}[L^2] \to 1$），则 $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \to 0$，且弱检测不可能实现。

对 $(C^*, \Theta^*)$ 的可能取值 $(C, \theta)$ 求平均，我们得到

$$L(X) = \int_0^{2\pi} \frac{1}{\binom{n}{k}} \sum_{C\subset[n]:|C|=k} L_{C,\theta}(X)\,\frac{\mathrm{d}\theta}{2\pi},$$

其中

$$L_{C,\theta}(X) = \begin{cases} \tau^{-\binom{k}{2}} & \text{if } X_{ij} \in [\theta, \theta+2\pi\tau] \text{ for all } i,j \in C,\ i<j, \\ 0 & \text{otherwise.} \end{cases}$$

对于固定的 $C$，$\theta$，$C'$，$\theta'$，令 $S = |C \cap C'|$ 并

$$\delta \triangleq \frac{1}{2\pi}\,\mathrm{length}\big([\theta, \theta+2\pi\tau] \cap [\theta', \theta'+2\pi\tau]\big).$$

Then

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}\big[L_{C,\theta}(X)L_{C',\theta'}(X)\big] &= \frac{1}{\tau^{2\binom{k}{2}}}\,\mathbb{Q}\left\{\begin{array}{l} X_{ij} \in [\theta, \theta+2\pi\tau] \cap [\theta', \theta'+2\pi\tau] \ \text{ for } i,j \in C\cap C',\ i<j, \\ X_{ij} \in [\theta, \theta+2\pi\tau] \ \text{ for } i,j \in C,\ \{i,j\} \not\subset C\cap C',\ i<j, \\ X_{ij} \in [\theta', \theta'+2\pi\tau] \ \text{ for } i,j \in C',\ \{i,j\} \not\subset C\cap C',\ i<j \end{array}\right\} \\ &= \frac{\delta^{\binom{S}{2}}\tau^{k(k-1)-S(S-1)}}{\tau^{k(k-1)}} = \delta^{\binom{S}{2}}\,\tau^{-2\binom{S}{2}}. \end{aligned}$$

因此

$$\begin{aligned} \mathbb{E}_{\mathbb{Q}}[L^2] &= \mathbb{E}_{C,C',\Theta,\Theta'}\big[\mathbb{E}_{\mathbb{Q}}\big[L_{C,\Theta}(X)L_{C',\Theta'}(X)\big]\big] \\ &= \mathbb{E}_S\Big[\mathbb{E}_{\Theta,\Theta'}\Big[\delta^{\binom{S}{2}}\tau^{-2\binom{S}{2}}\Big]\Big] = \mathbb{E}_S\left[\mathbb{E}_{\Theta,\Theta'}\left[\left(\frac{\delta}{\tau^2}\right)^{\binom{S}{2}}\right]\right]. \end{aligned}$$

由于最大重叠长度为 $2\pi\tau$，我们有 $\delta \le \tau$，因此 $\delta/\tau^2 \le 1/\tau$且

$$\mathbb{E}_{\mathbb{Q}}[L^2] \le \mathbb{E}_S\left[\left(\frac{1}{\tau}\right)^{\binom{S}{2}}\right] \overset{\text{(a)}}{=} \mathbb{E}_S\left[\left(\frac{1}{\tau}\right)^{\frac{k-1}{2}\times S}\right],$$

其中 (a) 对 $\binom{S}{2} = \frac{S(S-1)}{2} \le \frac{k-1}{2}S$ 用于 $0 \le S \le k$。

变量 $S$服从超几何分布$(n, k, k)$，其在凸序意义下被二项分布$(k, p)$所支配，其中 $p = k/n$。令 $z = (1/\tau)^{\frac{k-1}{2}}$。利用 $s \mapsto z^s$在 $z \geq 0$上是凸函数这一点，

$$\mathbb{E}_S[z^S] \leq (1 - p + pz)^k = \left(1 + \frac{k}{n}(z-1)\right)^k .$$

最后，由于 $1 + x \leq e^x$，

$$\mathbb{E}_{\mathbb{Q}}[L^2] \leq \exp\left(\frac{k^2}{n}(z-1)\right) = \exp\left(\frac{k^2}{n}\left(\left(\frac{1}{\tau}\right)^{\frac{k-1}{2}} - 1\right)\right).$$

若（31）成立，则指数趋于 0，因此 $\mathbb{E}_{\mathbb{Q}}[L^2] \to 1$ 且 $\mathrm{Var}_{\mathbb{Q}}(L) = \mathbb{E}_{\mathbb{Q}}[L^2] - 1 \to 0$，这意味着 $\mathrm{TV}(\mathbb{P}, \mathbb{Q}) \to 0$ 并完成了定理 3(b) 的证明。

## 5 冯·米塞斯社区模型

回顾函数

$$R(\kappa) = \frac{I_0(2\kappa)}{I_0^2(\kappa)} \qquad A(\kappa) = \frac{I_1(\kappa)}{I_0(\kappa)},$$

其中 $I_0$ 和 $I_1$ 分别为 0 阶和 1 阶第一类修正贝塞尔函数。定义及这些函数的性质参见附录A 。

**Th定理 4.** 考虑参数为 $(n, k, \kappa)$ 的冯·米塞斯社区模型。

(a) （可实现性）当窗口长度为 $2\pi\tau$ 、阈值为 $k$ 时，区间检验满足

$$p_{\mathrm{fa}} \leq \exp\left(\log\frac{\binom{k}{2}}{\tau} + k \cdot \left\{\log\frac{n}{k} + 1 + \frac{k-1}{2}\log\tau\right\}\right), \tag{33}$$

$$p_{\mathrm{miss}} \leq \binom{k}{2} \cdot \frac{\exp\{(\cos(\pi\tau) - 1)\kappa\}}{2\pi^2\, I_0(\kappa)\, e^{-\kappa} \cdot \kappa \cdot |\sin(\pi\tau)|}. \tag{34}$$

因此，若 (33) 和 (34)的右端趋于 0，则区间检验可实现强检测。此外，对于任意整数 $B \geq 3$，$\epsilon \in (0, 1)$ 以及合适的阈值 $\beta$ ，相干性检验满足

$$p_{\mathrm{fa}} \leq B \cdot \exp\left\{k\left(\log\left(\frac{ne}{k}\right) - (1 - \epsilon/4)^2 \cdot (k-1) \cdot A^2(\kappa) \cdot \cos^2(\pi/B)/2\right)\right\}, \tag{35}$$

$$p_{\mathrm{miss}} \leq \exp\left(-\frac{\epsilon^2 K\, A^2(\kappa)}{32}\right). \tag{36}$$

因此，若 (35)和 (36)的右端趋于0，则相干性检验可实现强检测。

(b) （逆命题）若满足以下条件，则弱检测不可能

$$\frac{k^2}{n}\left(\exp\left(\frac{(k-1)\log R(\kappa)}{2}\right) - 1\right) \xrightarrow[n\to\infty]{} 0. \tag{37}$$

**推论 7** （可实现性情形）。 在冯·米塞斯社区模型下，在以下每种情形中均可以进行强检测。

- $3/\epsilon < k = o$（对于某个 $\epsilon \in$（0, 1）的 $n$以及 $\kappa \geq \left(\frac{4}{\pi^2} + \epsilon\right)$ $(1+\epsilon)$ $\left(\frac{n}{k}\right)^4$（$k$） $^{k-1}$ . [使用区间检验。令 $\tau \equiv \tau_\epsilon \triangleq \left(k/n\right)^{\frac{2+\epsilon}{k-1}}$，并使用窗口长度为 $2\pi\tau$ 、子集大小为 $k$的区间检验，由（33）可得 $pfa \to 0$ （参见推论 5 的证明）。此外，利用 $\tau \to 0$、 $\kappa \to \infty$、cos（$\pi\tau$） $-1 = -$（$\pi\tau$） 2 2（$1 + o$（1））、sin（$\pi\tau$） $= \pi\tau$（$1 + o$（1））以及 $I_0$（$\kappa$） $\sim \frac{e^\kappa}{\sqrt{2\pi\kappa}}$，方程（34）得到

$$\log p_{\text{miss}} \leq (\text{const}) + 2\log k - \frac{1}{2}\log(\kappa\tau^2) - \frac{\pi^2\kappa\tau^2(1+o(1))}{2}, \tag{38}$$

因此，由于 $\kappa\tau^2 \geq (\frac{4}{\pi^2} + \epsilon)(\log k)(\frac{n}{k})^{\frac{2\epsilon}{k-1}}$，(38) 中的最后一项（负项）占主导，因此log $p_{\text{miss}} \to -\infty$。 ]

- $k = c\log n$ 对某个 $c > 2$ 和固定的 $\kappa$，满足 $A^2(\kappa)\, c > 2$. [使用相干性检验。取足够大的 $B$ 和足够小的 $\epsilon$ 。]
- $k = c\log\, n$ 和 $\kappa \geq \frac{2\log\frac{1}{}}{}$ $^k\cos(\pi e^{-2/c})$。 [令 $\tau = e^{-2/c}$，并使用窗口大小为$2\pi\tau$ 的区间检验。注意，对于此处选择的 $k$ 和 $\tau$，pfa $\to 0$ （见推论 5 的证明）。利用 $I_0(\kappa)$ $\sim \frac{e^\kappa}{\sqrt{2\pi\kappa}}$（因为 $\kappa \to \infty$），(34) 得出 log $p$miss $\leq$(const) $+2$log $k - \frac{1}{2}$log($\kappa$) $+$(cos($\pi\tau$) $-1$) $\kappa \to -\infty$.]
- $k = \omega(\log\, n)$ 和 $\kappa \geq (1+\epsilon)\sqrt{^8\log\frac{n}{k-1}}$ 对某个 $\epsilon$，满足 $0 < \epsilon < 1$. [使用相干性检验。(36) 和 (35) 右侧关于 $\kappa$ 单调递减，因此我们可以假设 $\kappa = (1+\epsilon)\sqrt{^8\log\frac{n}{k-1}}$. 这些条件意味着 $\kappa \to 0$，因此 $A(\kappa) \sim \frac{\kappa}{2}$（见附录）。由于 $(1-\epsilon/4)(1+\epsilon) \geq 1+\epsilon/2$，我们可以选取足够大的 $B$，使得 $(1-\epsilon/4)(1+\epsilon)\cos\left(\pi/B\right) \geq 1+\epsilon/4$。那么，(35) 中的指数小于或等于 $k\left\{\log\binom{ne}{k} - (1+\epsilon/4)^2\log\, n\right\}$，因此 $p$fa $\to 0$. 此外， $K\kappa^2 \to \infty$，所以 $p$miss $\to 0$.]

**推论 8 (**不可能区域**)**。 在冯·米塞斯社区模型下，以下每个条件都意味着弱检测不可能。请注意，随着 $k$ 增大，这些条件要求 $\kappa$ 变小；特别是当 $k = \omega$ 时， $\kappa \to 0$ $(\log n)$。

- $k = o(\log\, n)$ 且 $\kappa \leq \left(n/k^2\right)^{\frac{4-\epsilon}{k-1}}$ 对于某个 $\epsilon > 0$. [这些条件意味着 $\kappa \to \infty$ 因此$R(\kappa)$ $\sim \sqrt{\pi\kappa}$. 这些条件还意味着 $\log\binom{k^2}{n} + (k-1)\log(\sqrt{\pi\kappa}(1+o(1))/2) \to -\infty$ . ]
- $k = c\log\, n$ 对于某个 $c > 0$ 且 $c\log\, R(\kappa) < 2$. [这意味着 $\log\binom{k^2}{n} + (k-1)\log(R(2^\kappa \to -\infty$. ])
- $k = \omega(\log\, n)$ 且 $k \leq n^c$ 对于某个 $c$ 满足 $0 < c < 0.5$ 且 $\kappa^2 \leq {}^4(1-2c-\epsilon)\log_{k-1}{}^n$ 对于某个$\epsilon > 0$. [这些条件意味着 $\kappa \to 0$ 因此 $\log R(\kappa) \sim \frac{\kappa^2}{2}$ 而且它们还意味着 $\log\binom{k^2}{n} + (k-1)\kappa^2(1+o(1))/4 \to -\infty$. ]

注 1。推论 8 的第一个条件与推论 7 中强恢复的第一个充分条件相匹配，仅相差 $\log k$ 因子。此外，如果我们将关于 $k$ 的条件进一步加强为 $k = o\left(\log\, n \log\log\, n\right)$，那么 $\log\, k = o\left(\binom{n}{k^2}^\epsilon\right)$；在这种情况下，因子 $\overline{\log\, k}$ 可以并入另一项，因此没有实际意义。推论 8 的第二个条件与推论 7 中强恢复的第二个充分条件也很好地吻合——对于第二个条件以及 $c$ 的较大取值， $\kappa^2$ 相差一个 2 的因子。特别地，当 $\kappa \to 0$ 时，$\log R(\kappa) \sim \frac{\kappa^2}{2}$ 且 $A^2(\kappa) \sim \frac{\kappa^2}{4}$. 。类似地，推论 8 的第三个条件与第四个充分条件很好地吻合

强恢复在推论 7中的条件；这里同样相差一个2因子，对$\kappa^2$ 当 $c$ 接近 0时的取值而言。

### 5.1 冯·米塞斯社区模型的区间检验可实现性

在 $H_0$下，边变量在 $[0, 2\pi)$ 上独立同分布均匀，恰如硬簇社区模型。因此，虚警界（33）与（30）相同，后者已得到证明。

如果所有社区内边的标签都落在以真实值 $\Theta$.为中心的检测窗口内，区间检测器就返回决策 $H_1$ 。因此， $p_{\text{miss}}$ 小于或等于在 $\mathbb{P}$ 下， $\binom{k}{2}$ 个此类标签中至少有一个未落入以 $\Theta$. 的真实值为中心的检测区间的概率。由并集界可得，

$$p_{\text{miss}} \le \binom{k}{2}\left(1 - p_\kappa(\tau)\right),$$

其中 $p_\kappa$ 定义见定理 2。

利用以下事实：若 $\Theta$ 在 $[0, 2\pi]$ 上均匀分布，则 $\cos\,\Theta$ 的概率密度函数为 $\frac{1}{\pi\sqrt{1-t^2}}$，其中 $t \in [-1, 1]$，可得

$$\begin{aligned}
1 - p_\kappa(\tau) &= \frac{1}{2\pi I_0(\kappa)} \int_{\{\theta\in[0,2\pi]:\,\cos\theta \le \cos(\pi\tau)\}} e^{\kappa\cos\theta}\,\mathrm{d}\theta \\
&= \frac{1}{2\pi^2 I_0(\kappa)} \int_{-1}^{\cos(\pi\tau)} \frac{e^{\kappa t}}{\sqrt{1-t^2}}\,\mathrm{d}t \\
&\le \frac{1}{2\pi^2\, I_0(\kappa)\,\sqrt{1-\cos^2(\pi\tau)}} \int_{-\infty}^{\cos(\pi\tau)} e^{\kappa t}\,\mathrm{d}t \\
&= \frac{\exp\{(\cos(\pi\tau)-1)\kappa\}}{2\pi^2\, I_0(\kappa)\, e^{-\kappa}\cdot\kappa\cdot|\sin(\pi\tau)|}
\end{aligned}$$

从而得到（34）。

**注 2**。 考虑区间检验。如果不使用窗口大小 $2\pi\tau_\epsilon$ ，而改用窗口大小 $2\pi\tau_f$ ，其中

$$\tau_f = \left(\frac{k}{ne}\right)^{\frac{(1+\epsilon)k}{\binom{k}{2}-1}}$$

并且如果

$$\kappa \ge \tau_f^{-2}\left(\frac{4}{\pi^2}+\epsilon\right)(\log k)\left(\frac{n}{k}\right)^{\frac{2\epsilon}{k-1}}$$

那么对于 $k = o$，即可实现强恢复$(\log n)$ （即不再需要条件 $k \ge 3/\epsilon$ ）。

### 5.2 冯·米塞斯社区模型的可实现性——相干性检验

我们对 $p_{\text{fa}}$ 进行界定。根据并集界，对于任意满足 $|C| = k$， 的 $C \subset [n]$ ，

$$p_{\text{fa}} \le \binom{n}{k} \cdot \mathbb{Q}\left(\left|\sum_{e\in E(C)} Z_e\right| > \beta\right). \tag{39}$$

设 $B$ 是一个满足 $B \geq 3$ 的整数，并考虑内接于复平面中单位圆的 $B$边形。每条边到原点的距离为 $\cos\left(\frac{\pi}{B}\right)$。因此，利用对称性和并集界，

$$\begin{aligned} p_{\mathrm{fa}} &\leq \binom{n}{k} \cdot B \cdot \mathbb{Q}\left(\Re\left(\sum_{e\in E(C)} Z_e\right) > \beta \cos\left(\frac{\pi}{B}\right)\right) \\ &= \binom{n}{k} \cdot B \cdot \mathbb{Q}\left(S_c > \beta \cos\left(\frac{\pi}{B}\right)\right), \end{aligned} \tag{40}$$

其中 $S_c \triangleq \sum_{e\in E(C)} \cos(X_e)$。对于任意 $s > \mathbb{E}_{\mathbb{Q}}[S_c] = 0$，若满足 $K = \binom{k}{2}$，

$$\mathbb{Q}(S_c \geq s) \overset{(A)}{\leq} \inf_{\theta>0} \exp\left\{-\theta\, s + K \log \mathbb{E}\left[e^{\theta \cos(X_{12})}\right]\right\} \overset{(B)}{\leq} \inf_{\theta>0} e^{-\theta\, s + K\frac{\theta^2}{4}} \overset{(C)}{=} e^{-s^2/K}, \tag{41}$$

其中，(A) 利用了 $X_e$ 在 $\mathbb{Q}$ 下独立同分布这一性质，(B) 利用了

$$\mathbb{E}_{\mathbb{Q}}[e^{\theta\cos(X_i)}] = \frac{1}{2\pi}\int_0^{2\pi} e^{\theta\cos t}\,\mathrm{d}t = I_0(\theta) \leq e^{\frac{1}{4}\theta^2},$$

以及 (C) 利用了 $\theta^* = 2s/K$ 对该表达式进行优化。结合 (40) 和 (41) 可得

$$p_{\mathrm{fa}} \leq \binom{n}{k} \cdot B \cdot \exp\left\{-\frac{\beta^2\cos^2(\pi/B)}{K}\right\} \leq B \cdot \exp\left\{k\log\left(\frac{ne}{k}\right) - \frac{\beta^2\cos^2(\pi/B)}{K}\right\}. \tag{42}$$

令 $\beta = (1-\frac{\epsilon}{4})K A(\kappa)$，其中任意 $\epsilon$ 满足 $0<\epsilon<1$，可得 (35)。

下面来界定漏检概率。对于任意边集 $E$，

$$\left|\sum_{e\in E} Z_e\right| = \max_{\theta} \sum_{e\in E} \cos(X_e - \theta) \geq \sum_{e\in E} \cos(X_e - \Theta^*),$$

其中 $\Theta^*$ 是植入角。因此，

$$p_{\mathrm{miss}} = \mathbb{P}\left(\max_{C:|C|=k}\left|\sum_{e\in E(C)} Z_e\right| \leq \beta\right) \leq \mathbb{P}\left(\left|\sum_{e\in E(C^*)} Z_e\right| \leq \beta\right) \leq \mathbb{P}\left(\sum_{e\in E(C^*)} \cos(X_e - \Theta^*) \leq \beta\right), \tag{43}$$

其中 $C^*$ 是植入集合。容易验证，随机变量 $\{\cos(X_e - \Theta^*)\}_{e\in E(C^*)}$是独立同分布的，且其取值范围为 $[-1,1]$，均值为

$$\int_0^{2\pi} \cos(x-\Theta^*)\cdot\frac{\exp(\kappa\,\cos(x-\Theta^*))}{2\pi I_0(\kappa)}\,\mathrm{d}x \overset{(a)}{=} \frac{1}{I_0(\kappa)}\cdot\left[\frac{1}{2\pi}\int_0^{2\pi}\cos(y)\,e^{\kappa\cos(y)}\,\mathrm{d}y\right] \overset{(b)}{=} \frac{I_1(\kappa)}{I_0(\kappa)} =: A(\kappa),$$

其中 (a) 使用了变量替换 $y = x - \Theta^*$，(b) 使用了修正贝塞尔函数 $I_1(\cdot)$ 的定义。应用 Hoeffding 不等式可得，对于任意 $u>0$,

$$\mathbb{P}\left(\sum_{e\in E(\mathcal{C}^*)} \cos(X_e - \Theta^*) \leq K\,A(\kappa) - u\right) \leq \exp\left(-\frac{u^2}{2K}\right). \tag{44}$$

具体而言，将 (43) 和 (44) 结合起来，并令 $u = \epsilon K\,A(\kappa)/4$ 以及 $\beta = (1-\frac{\epsilon}{4})K\,A(\kappa)$ 可得 (36)。

### 5.3 冯·米塞斯社区模型的逆命题

令 $K \equiv \binom{k}{2}$。与平坦模型中一样，固定候选项 $(C,\theta)$ 的条件似然比为

$$L_{C,\theta}(X) = I_0(\kappa)^{-K} \exp\Big\{\kappa \sum_{e\in E(C)} \cos(X_e - \theta)\Big\},$$

其中，$E(C)$ 是 $C$ 内部的 $\binom{k}{2}$ 无序对集合。无条件似然比为混合分布 $L(X) = \mathbb{E}_{C,\,\Theta}[L_{C,\Theta}(X)]$。

设 $(C,\ \Theta)$ 和 $(C',\ \Theta')$ 是植入参数的独立抽样。按照平坦模型中的做法，利用 $\mathbb{Q}$ 下边的独立性：

$$\mathbb{E}_{\mathbb{Q}}\big[L_{C,\Theta}(X)\, L_{C',\Theta'}(X) \;\big|\; C,\Theta,C',\Theta'\big] = \left[\frac{I_0\big(2\kappa\cos\frac{\Theta-\Theta'}{2}\big)}{I_0^2(\kappa)}\right]^{\binom{S}{2}},$$

其中，$S \triangleq |C\cap C'|$，并且我们利用了以下事实：$|E(C)\cap E(C')| = \binom{S}{2}$ 是重叠的社区内边的数量。于是，对 $(C,\ \Theta)$ 和 $(C',\ \Theta')$ 取平均可得

$$\mathbb{E}_Q[L^2] \;=\; \mathbb{E}_{C,C'}\left[\frac{1}{2\pi}\int_0^{2\pi}\left(\frac{I_0\big(2\kappa\cos\frac{\phi}{2}\big)}{I_0^2(\kappa)}\right)^{\binom{S}{2}} \mathrm{d}\phi\right], \qquad \phi \equiv \Theta - \Theta'. \tag{45}$$

我们进一步通过对 $\phi$ 取最大值来给出上界。由于对所有 $\phi$，$I_0\Big(2\kappa\cos\frac{\phi}{2}\Big) \le I_0(2\kappa)$，

$$\mathbb{E}_{\mathbb{Q}}[L^2] \;\le\; \mathbb{E}_{C,C'}\left[\,R(\kappa)^{\binom{S}{2}}\right],$$

其中 $R(\kappa) \triangleq \frac{I_0(2\kappa)}{I_0^2(\kappa)}$ · 随机变量 $S$ 服从超几何分布 $(n,k,k)$。根据凸序，该分布的 $z$–变换被 Binom$(k,k/n)$ 二项分布的 $z$–变换所支配：

$$\mathbb{E}\big[z^S\big] \;\le\; \big(1-p+pz\big)^k, \qquad \text{for all } z\ge 0.$$

对 $0\le s\le k$ 使用 $\binom{s}{2} \le \frac{k-1}{2}\, s$，

$$R(\kappa)^{\binom{S}{2}} \;=\; \exp\left(\big(\log R(\kappa)\big)\cdot\binom{S}{2}\right) \le \exp\left(\tfrac{(k-1)\log R(\kappa)}{2} S\right). \tag{46}$$

令 $t \equiv (k-1)\log R(\kappa)\,2$ 且令 $z = e^t$，得到

$$\mathbb{E}_{\mathbb{Q}}[L^2] \;\le\; \mathbb{E}\big[e^{tS}\big] \;\le\; \big(1-p+pe^t\big)^k \;\le\; \exp\Big(k\,p\,(e^t-1)\Big) \;=\; \exp\Big(\frac{k^2}{n}\big(e^t-1\big)\Big).$$

这证明了该界

$$\mathbb{E}_{\mathbb{Q}}[L^2] \;\le\; \exp\left\{\frac{k^2}{n}\Big(\exp\big(\tfrac{(k-1)\log R(\kappa)}{2}\big)-1\Big)\right\}. \tag{47}$$

因此，$\mathbb{E}_{\mathbb{Q}}[L^2]\to 1$（因此，当式（47）中的指数趋于 0 时，检测是不可能的。这正是必要条件（37）。定理4（b）的证明到此完成。

## 贝塞尔函数的界与近似

本附录汇总了与第一类修正贝塞尔函数相关的界和渐近值， $I_0(\kappa)$。

- $I_0$ 的幂级数展开与在 $x = \kappa^2/4$ 处求值的 $e^x$的幂级数展开进行比较可得，对所有 $\kappa$,

$$I_0(\kappa) = \sum_{k=0}^{\infty} \frac{1}{(k!)^2} \left(\frac{\kappa}{2}\right)^{2k} \leq e^{\frac{1}{4}\kappa^2}$$

- 设 $\kappa > 1/2$. 对于 $\Theta$ 在上均匀分布 $[0, 2\pi]$，cos的概率密度 $\Theta$ 为 $\frac{1}{\pi\sqrt{1-t^2}}$ 对于 $-1 < t < 1$. 这给出了积分关系（经典上称为 [DLMF]）及其界：

$$I_0(\kappa) = \frac{1}{\pi} \int_{-1}^{1} \frac{e^{\kappa t}}{\sqrt{1-t^2}} \, \mathrm{d}t \geq \frac{1}{\pi} \int_{1-\frac{1}{\kappa}}^{1} \frac{e^{\kappa t}}{\sqrt{1-t^2}} \, \mathrm{d}t$$

  在积分区间上使用 $1-t^2 \leq 2\,（1-t）$ 和 $e^{\kappa t} \geq e^{\kappa}\,（1-（1-t）\,）\,\kappa$并积分，得到

$$I_0(\kappa) \geq \frac{2\sqrt{2}}{3\pi} \frac{e^{\kappa}}{\sqrt{\kappa}} \geq (0.3) \frac{e^{\kappa}}{\sqrt{\kappa}}.$$

  这个下界接近大 $\kappa$ 渐近式：

$$I_0(\kappa) \sim \frac{e^{\kappa}}{\sqrt{2\pi\kappa}} \approx (0.39894) \frac{e^{\kappa}}{\sqrt{2\pi\kappa}}\,.$$

  考虑余项，可得当 $\kappa \to \infty$时，

$$I_0(\kappa) = \frac{e^{\kappa}}{\sqrt{2\pi\kappa}} \left(1 + O(\kappa^{-1})\right). \tag{48}$$

- 令

$$A(\kappa) = \frac{1}{I_0(\kappa)} \cdot \left[\frac{1}{2\pi} \int_0^{2\pi} \cos(y)\, e^{\kappa \cos(y)} \, \mathrm{d}y\right] = \frac{I_1(\kappa)}{I_0(\kappa)}$$

  泰勒展开可得（另见图1）

  - As $\kappa \to 0$, $A(\kappa) = \frac{\kappa}{2} + O(\kappa^3)$.
  - As $\kappa \to \infty$, $A(\kappa) \to 1$.

- 令 $\rho_\kappa(\theta) \triangleq I_0(2\kappa \cos\theta)/I_0^2(\kappa)$。 引理 1 表明，对于每个 $\kappa > 0$， $\frac{1}{2\pi} \int_0^{2\pi} \rho_\kappa(\theta) \, \mathrm{d}\theta = 1$.

- 令 $R(\kappa) \triangleq I_0(2\kappa)/I_0^2(\kappa)$。 $\rho_\kappa(\theta)$ 在 $\theta$ 上的最大值为 $R(\kappa) = \rho_\kappa(0)$。 </style}上述 $I_0$ 的渐近行为意味着 $R$.的以下渐近行为

  - As $\kappa \to 0$, $R(\kappa) = 1 + \frac{\kappa^2}{2} + O(\kappa^4)$.
  - As $\kappa \to \infty$, $R(\kappa) = \sqrt{\pi\kappa}\left(1 + O(\kappa^{-1})\right)$.

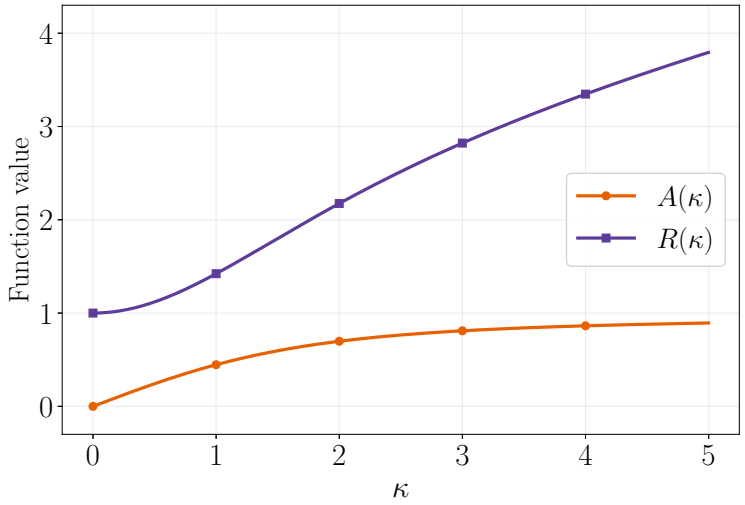


图 1：函数 $A(\kappa)$ 和 $R(\kappa)$

- 当 $\kappa \to 0$ 时，

$$\begin{aligned} \mathrm{D_{KL}}(\mathrm{vonMises}(0,\kappa)||\mathrm{uniform}([0,2\pi]) &= \kappa\frac{I_1(\kappa)}{I_0(\kappa)} - \log I_0(\kappa) \\ &= \frac{\kappa(\kappa/2+o(\kappa))}{I_0(\kappa)} - \log I_0(\kappa) \\ &= \frac{\kappa^2}{2} - \frac{\kappa^2}{4} + o(\kappa^2) = \frac{\kappa^2}{4} + o(\kappa^2) \end{aligned}$$

# B 广义似然比检验

## B.1 硬聚类平坦模型的区间检验

对于 $\theta \in [0,1]$，定义计数

$$N(\theta) \triangleq \sum_{i=1}^{N} \mathbf{1}\left\{X_i \in [\theta, \theta+2\pi\tau] \pmod 1\right\}, \qquad M_\tau(X) \triangleq \sup_{\theta\in[0,2\pi]} N(\theta),$$

即，$M_\tau(X)$ 是长度为 $2\pi\tau$ 的区间内所能包含的最大点数。回想一下，上述窗口为 $2\pi\tau$、阈值为 $\gamma$ 的区间检验为

$$\phi_{\tau,\gamma}(X) \triangleq \mathbf{1}\left\{M_\tau(X) \geq \gamma\right\}, \qquad \gamma \in \{1,\ldots,n\}.$$

**似然比。** 回顾给定 $(S,\ \theta)$ 时的条件似然比，

$$L_{S,\theta}(x) = \prod_{i\in S} \frac{\mathbf{1}\{x_i \in I_\theta\}}{\tau} = \tau^{-K}\,\mathbf{1}\left\{x_i \in I_\theta\ \forall i \in S\right\}. \tag{49}$$

对 $S$ 和 $\Theta$ 取平均，得到（无条件）似然比，

$$
\begin{aligned}
L(x) \triangleq \frac{\mathrm{d}P}{\mathrm{d}Q}(x) &= \frac{1}{\binom{N}{K}\tau^K} \int_0^{2\pi} \sum_{\substack{S\subset[N]\\|S|=K}} \mathbf{1}\big\{\, S \subset \{i : x_i \in I_\theta\}\big\} \frac{\mathrm{d}\theta}{2\pi} \\
&= \frac{1}{\binom{n}{K}\tau^K} \int_0^{2\pi} \binom{N(\theta)}{K} \frac{\mathrm{d}\theta}{2\pi}.
\end{aligned} \tag{50}
$$

最后一步用到了：大小为 $N(\theta)$ 的集合中包含的 $K$-子集数量为 $\binom{N(\theta)}{K}$。根据消去参数 $(S,\theta)$ 的方式，我们考虑两种广义似然比检验，如下：

(A) 对 $S$ 和 $\theta$ 均取最大值。在这种情况下，

$$
T_A(x) \triangleq \sup_{\theta\in[0,1]} \sup_{\substack{S\subset[N]\\|S|=K}} L_{S,\theta}(x) \overset{\text{(a)}}{=} \sup_{\theta} \tau^{-K}\, \mathbf{1}\{\, N(\theta) \ge K \,\} \overset{\text{(b)}}{=} \tau^{-K}\, \mathbf{1}\{\, M_\tau(x) \ge K \,\}, \tag{51}
$$

其中，(a) 用到了如下事实：给定 $\theta$，存在由 $[\theta, \theta+2\pi\tau]$ 中的点构成的某个指标集 $S$，当且仅当 $N(\theta) \ge K$；(b) 则用到了 $M_\tau(x)$ 的定义。

因此，这种情况下的决策规则恰好是将 $T_A$ 与任意 $t \in (0, \tau^{-K}]$ 进行比较：

$$
\text{reject } H_0 \iff M_\tau(X) \ge K,
$$

也就是说，这是阈值为 $\gamma = K$ 的区间检验。

(B) 对 $S$取平均值，然后对 $\theta$取最大值。根据（50），可得

$$
T_B(x) \triangleq \sup_{\theta\in[0,1]} L_\theta(x) = \frac{1}{\binom{N}{K}\tau^K} \sup_{\theta} \binom{N(\theta)}{K}. \tag{52}
$$

由于 $m \mapsto \binom{m}{K}$ 在 $m \ge K$ 上严格递增，因此对 $T_B$ 进行阈值判定等价于对$M_\tau(X)$进行阈值判定：

$$
T_B(x) > t \iff M_\tau(X) \ge \gamma \quad \text{for some integer } \gamma \ge K \text{ depending on } t, \tag{53}
$$

即具有可调参数 $\gamma \ge K$ 的区间检验。

### B.2 冯·米塞斯社区模型的相干性检验

考虑参数为 $n$，$k$，$\kappa$. 的冯·米塞斯社区模型检测问题。在给定 $\Theta^* = \theta$ 和 $C^* = C$ 的情况下，对数似然比为：

$$
\begin{aligned}
\log L_{\theta,C}(X) &= \kappa \sum_{e\in E(C)} \cos(X_e - \theta) - K\log(I_0(\kappa)) \\
&= \kappa\cos(\theta) \sum_{e\in E(C)} \cos(X_e) + \kappa\sin(\theta) \sum_{e\in E(C)} \sin(X_e) - K\log(I_0(\kappa))
\end{aligned}
$$

对于每条边 $e$，令 $Z_e = e^{\mathrm{i}X_e} = \cos(X_e) + \mathrm{i}\sin(X_e)$。对 $\theta$ 取最大值，得到s:

$$
\max_{\theta\in[0,2\pi]} \log L_{\theta,S}(X) = \kappa \left| \sum_{e\in E(C)} Z_e \right| - K\log(I_0(\kappa))
$$

If w进一步对 $C$ 取最大值，并选择阈值 $\beta$，我们得到 GLRT 检验 st

$$\psi_n(\beta) \triangleq \mathbf{1}\left\{ \max_{\substack{C\subseteq[n]\\ |C|=k}} \left| \sum_{e\in E(C)} Z_e \right| \geq \beta \right\}.$$

这是我们针对冯·米塞斯社区模型的相干性检验。

## C 多项式时间检验

下文定义的 *Rayleigh* 检验已成功用于文献中，用于判断一组观测值是来自冯·米塞斯分布还是均匀分布。我们表明，在冯·米塞斯社区模型的植入设定中，该检验的效果并不理想。

设 $Z_{ij} = e^{\mathrm{i}X_{ij}}$，并考虑检验统计量

$$T_{\text{Rayleigh}} \triangleq \left| \sum_{1\leq i<j\leq n} Z_{ij} \right|.$$

我们分析阈值 thr eshold $\beta$ 下的检验 $\mathbf{1}\{T_{\text{Rayleigh}} \geq \beta\}$。令 $N_E = \binom{n}{2}$ 为边的总数 .

界定虚警概率。在 $H_0$下， $Z_{ij}$ 在单位圆上独立同分布均匀。利用集中界（40）和（41）：

$$p_{\text{fa}} = \mathbb{Q}\left(T_{\text{Rayleigh}} \geq \beta\right) \leq 4\exp\left(-\frac{\beta^2}{2N_E}\right). \tag{54}$$

界定漏检概率。在 $H_1$下，令 $\Theta$ 为植入角，并令 $m = \binom{k}{2}$。采用与（43）和（36）相同的思路，可得总和 $S = \sum_{i<j} Z_{ij}$ 在给定 $\Theta$ 时的期望值为$\mathbb{E}_P[S|\Theta] = mA\left(\kappa\right) e^{\mathrm{i}\Theta}$。令 $\mu_1 = m\,A\left(\kappa\right)$。我们通过投影到真实方向 $\Theta$，给统计量取下界：

$$T_{\text{Rayleigh}} = |S| \;\geq\; \Re(Se^{-\mathrm{i}\Theta}) \;=\; \sum_{i<j} \cos(X_{ij} - \Theta) \;\triangleq\; S_\Theta\,.$$

$S_\Theta$ 是由 $N_E$ 个相互独立、取值范围在 $[-1,1]$内且均值为 $\mu_1$的实随机变量构成的和。根据 Hoeffding 不等式：

$$p_{\text{miss}} \leq P(S_\Theta < \beta) \leq \exp\left(-\frac{(\mu_1-\beta)^2}{2N_E}\right),$$

当 $\beta < \mu_1$成立时。
为了确保两种错误概率都趋于零，我们必须能够选择 $\beta$，使得 $\beta^2 = \omega\big(N_E\big)$ （对$p_{\text{fa}}$）且 $\beta < \mu_1$ （对 $p_{\text{miss}}$）。这要求 $\mu_1^2 = \omega\big(N_E\big)$。取 $\beta = \mu_1/2$，于是总错误概率

$$p_{\text{fa}} + p_{\text{miss}} \leq 5\exp\left(-\frac{\mu_1^2}{8N_E}\right) = 5\exp\left(-\frac{m^2A^2(\kappa)}{8N_E}\right).$$

代入 $m \approx k^2/2$ 和 $N_E \approx n^2/2$ 后，该检验成功的充分条件是

$$k^4A^2(\kappa) = \omega\big(n^2\big) \quad \text{or equivalently} \quad k^2A(\kappa) = \omega\big(n\big)\,.$$

与（计算上困难的）相干性检验相比，这项检验在计算上效率较高，但要求更大的社群规模（例如， $k = \omega\big(\sqrt{n}\big)$ if$A(\kappa) = \Theta(1)$）。

## D 关于区间与相干性检验之间关系

我们将讨论社区模型的区间检验与相干性检验之间的关系。在定理 4中，针对冯·米塞斯社区模型，给出了使用区间检验实现强检测的充分条件，其条件为 $k = o(\log n)$；而使用相干性检验时，条件为 $k = c\log n$ 以及 $k = \omega(\log n)$。尽管这两节互为补充，但正如我们将在本节中解释的，在 $k$ 较小（因此 $\kappa$较大）的情形下，估计量本身相当相似。

对于给定的 $C \subset [n]$，具有 $|C| = k$ 的相干性检验所使用的统计量满足

$$\left| \sum_{e\in E(C)} Z_e \right| = \sum_{e\in E(C)} \cos(X_e - \widehat{\Theta}_C) \quad \text{where} \quad \widehat{\Theta}_C = \arg\max_{\theta} \sum_{e\in E(C)} \cos(X_e - \theta).$$

如果阈　值 $\beta$ 的形式为 $K - \epsilon$，其中某个 $\epsilon \in \in(0,1)$，并且超过该阈值:

$$\sum_{e\in E(C)} \cos(X_e - \widehat{\Theta}_C) \geq K - \epsilon \tag{55}$$

那么，对每个 $e.$，$\cos(X_e - \Theta_C) \geq 1 - \epsilon$。如果 $\epsilon$ 接近零，那么根据 cos 的泰勒近似，我们有

$$\left| \sum_{e\in E(C)} Z_e \right| \approx K - \frac{1}{2} \sum_{e\in E(C)} |X_e - \widehat{\Theta}_C|^2$$

$$\widehat{\Theta}_C \approx \arg\min_{\theta} \sum_{e\in E(C)} |X_e - \theta|^2$$

因此，如果我们定义 $V_C = \min_\theta \frac{1}{K-1} \sum_{e\in E(C)} |X_e - \theta|^2$，事件 (55) 近似等同于 $V_C \leq \frac{2\epsilon}{K-1}$. 如果值 $X_e$ 对于 $e \in E(C)$ 都位于单位圆的同一半圆上，那么 $V_C$就是实值观测值的常用方差无偏估计量。基于这一动机，我们将方差检验定义为

$$\psi_n(\tau) \triangleq \mathbf{1}\left\{ \min_{\substack{C\subseteq[n] \\ |C|=k}} V_C(X) \leq \sigma^2 \right\}.$$

对于某个阈值 $\sigma^2$。

总之，阈值为 $\beta = K - \epsilon$ 的相干性检验与阈值为 $\sigma^2 = \frac{2\epsilon}{K-1}$ 的方差检验大致相同。反过来，方差检验与区间宽度为 $\tau = \sigma$、阈值为 $K$ 的区间检验也相当相似，因为样本方差的平方根与 $(X_e : e \in E(C))$ 的范围宽度，都是衡量这些随机变量离散程度的相似指标。

## E 关于 $\Theta^*$ 的未知性

对于社区模型，我们的可达性结果不依赖于对植入相位$\Theta^*$的了解，而我们的反向界即使在向决策者揭示 $\Theta^*$ 的情况下仍然成立。对于平坦模型，我们可以刻画获取 $\Theta^*$ 后如何通过降低最优错误概率来提升性能，具体如下述命题所述。

**命题1**。 考虑平坦均匀模型下的检测问题，其中 $\Theta^*$ 已知于决策者。若 $K^2 \leq N$ 且 $\tau = o\left(K^2/N\right)$，则可以实现强检测。若 $K^2 \leq N$ 且 $\tau = \omega\left(K^2/N\right)$，则无法实现弱检测。

**备注3**。 比较推论 *1* 和命题 *1*表明，知道 $\Theta$ 的价值在于， $\tau$可以增大 $N^{\frac{1}{K-1}}$ 倍（当 $K$ 为常数时），并且可以增大 $\log\ N$ 倍（当$\overline{K} = N^{\alpha}$ 时）。 $0 < \alpha < 1/2.$

证明。 不失一般性，根据对称性，假设已知 $\Theta^* =$ $0.$
（可实现性）设 $K^2 \leq N$ 且 $\tau = o\left(K^2/N\right)$ . 令 $Y$表示落入区间 $[0,\tau]$ 的观测值数量，并考虑基于 $Y.$、阈值为 $\gamma$的检验。错误概率为:

$$p_{\text{miss}} = P(K + \text{binom}(N-K,\tau) < \gamma) \quad \text{and} \quad p_{\text{fa}} = P(\text{binom}(N,\tau) \geq \gamma).$$

注意， $p_{\text{miss}}$ 在这里等于第2.1节中使用的 $p_{\text{miss}}$ 的上界。因此，我们可以像该节中一样，对 $p_{\text{miss}}$ 应用相同的上界。此外， $p_{\text{fa}}$ 小于 $\mathbb{Q}$（$B_1$），后者小于式（8）的右侧（对于 $\gamma \geq 1 +$ （$N-1$）$\tau$）以及式（7）的右侧，二者都去除了前导因子 $N$。如果 $K$ 为常数，则 $n\tau \to 0$，我们取 $\gamma = K$，从而得到 $p_{\text{miss}} = 0$ 和$p_{\text{fa}} \to 0$，根据式（7）并去除因子 $N$。如果 $K \to \infty$ 使得 $K^2 \leq N$ 且 $\tau = o$（$K^2/N$），那么去除项 $-\log(N)$后的式（4）和式（5）均成立，并且二者结合足以证明$p_{\text{miss}} \to 0$ 和 $p_{\text{fa}} \to 0.$

(逆命题) 假设 $K^2 \leq N$ 且 $\tau = \omega\left(\frac{K^2}{N}\right)$。我们沿用定理 1(b) 的证明中的矩方法。已知 $\Theta^* \equiv 0,$ 只有 $\theta = \theta' = 0$ 这一种情况相关，因此变量 $\delta$ 等于 $\tau.$ 由此可得:

$$E_{\mathbb{Q}}[L^2] = \sum_{j=0}^{K} P\{|S \cap S'| = j\}\,\tau^{-j} \leq \sum_{j=0}^{K} P\left\{\text{binom}\left(K,\frac{K}{N}\right) = j\right\}\tau^{-j} = \left(1 - \frac{K}{N} + \frac{K}{N\tau}\right)^{K}$$
$$\leq \exp\left(\frac{K^2}{N\tau}\right) \to 1.$$

因此，弱检测是不可能的。 □

**备注4**。 命题 *1* 上述证明所用的假设检验是最大似然比检验，因此是帕累托最优的。因此，证明该命题逆命题部分的另一种方法是证明该检验满足 $\min_{\gamma} p_{\text{fa}} + p_{\text{miss}} \to 1/2$ 。

## F 社区 恢复及其与随机块模型的联系

本文中的社区模型与植入稠密社区的随机块模型不同之处在于，在这两种分布下，边标签的边缘分布相同， $\mathbb{P}$和 $\mathbb{Q}.$ 然而，如果预言机揭示参数 $\Theta$ 的值，那么本文中的社区模型便属于具有一般边标签分布的随机块模型范畴$P$和 $Q$，如 [WX21、HWX17]所示。在该模型中，对于给定的 $n$、 $k$、 $P$、 $Q$，假设从 $[n]$的所有基数为 $k$.的子集中均匀随机抽取一个基数为 $k$ 的子集 $C^*$ 。观测到一个对称的、零对角线的 $n \times n$ 矩阵 $A$，使得对于所有 $1 \leq i < j \leq n$， $A_{ij}$相互独立；若 $i$，则 $A_{ij} \sim P$，否则为 $A_{ij} \sim Q$ 。信息论上的必要且充分条件

弱检测和强检测的理论条件见 [WX21]，弱恢复和精确恢复的理论条件见 [HWX17]。由于可以忽略提供 $\Theta$ 的预言机，因此其中的必要条件可转化为此处模型的必要条件。为简单起见，不失一般性地假设 $\Theta = 0$.

为应用 [HWX17]，令 $P$ 表示参数为 $\kappa$、0 的冯·米塞斯分布，并令 Q 为 $[0, 2\pi]$. 上的均匀分布。似然比由 $\mathrm{d}P\mathrm{d}Q(x) = e^{\kappa}\cos(x)\, I_0(\kappa)$ 给出；当 $\kappa$ 有界时，该似然比有界，此时根据 [HWX17]中的引理 1，[HWX17] 的假设 1 成立。因此，根据 [HWX17] 中的定理 1，在 $\kappa$ 有界的情形下，关于 $\kappa$ 实现弱恢复的必要条件是 $K \cdot \mathrm{D}_{\mathrm{KL}}(P||Q) \to \infty$，且 $\liminf (K-1)\, \mathrm{D}_{\mathrm{KL}}(P||Q) \log(n/K) \geq 2$.。注意，

$$\mathrm{D}_{\mathrm{KL}}(P||Q) = \mathbb{E}_P\left[\log \frac{\mathrm{d}P}{\mathrm{d}Q}\right] = \mathbb{E}_P[\kappa \cos(\Theta)] - \log I_0(\kappa) = \kappa\, A(\kappa) - \log I_0(\kappa)$$

对于 $\kappa \to 0$，$\mathrm{D}_{\mathrm{KL}}(P||Q) \simeq \frac{\kappa^2}{4}$. 因此，如果我们考虑 $k$ 随 $n$ 变化，且满足 $\Omega(\log n) \leq k \leq o(n)$，我们发现弱恢复的必要条件是 $\liminf (k-1)\kappa^2 \log(n/k) \geq 2$，例如当 $\kappa^2 \geq \frac{2}{\log(n/k)\, k}$ .［我们还可以尝试使用 [HWX17] 中的定理 2，来得到精确恢复的必要条件。］

如果已知 $\Theta$，我们还可以应用 [HWX18]（与 [DM15] 相关）的结果，在信噪比满足某个条件时，得到一个用于弱恢复的多项式时间消息传递算法，并在经过清理过程后实现精确恢复；条件为 $\Omega(\sqrt{n}) \leq k \leq o(n)$（该条件对于弱恢复而言在信息论上并非紧致，但对于精确恢复或许是紧致的）。这些论文中的分析基于观测值的矩，因此我们进行变量替换，并使用由 $W_{ij} = \sqrt{2}\cos(X_{ij} - \Theta)$ 给出的观测值 $W_{ij}$。新的观测值构成一个充分统计量（假设已知 $\Theta$），因此不会造成信息论损失。我们有：若 $\{i, j\} \not\subset C^*$，则 $\mathbb{E}[W_{ij}] = 0$ 和 $\mathbb{E}[W_{i,j}^2] = 1$；若 $\{i, j\} \subset C^*$，则 $\mu = \mathbb{E}[W_{ij}] = \sqrt{2}A(\kappa) = \kappa/\sqrt{2} + O(\kappa^3)$，因此当 $\lambda = \frac{\kappa^2 K^2}{2n} > \frac{1}{e}$. 时，消息传递算法能够成功。

消息传递算法可以按如下方式调整，从而无需知道 $\Theta$。给定一个大整数 $B$，我们可以并行运行 $B$ 个消息传递算法副本，其中第 $b^{th}$ 个版本应用于观测矩阵 $W^{(b)} = \sqrt{2}\cos(X_{ij} + \frac{b}{2\pi})$。然后对于每个 $b$，$\mu_b = E[W_{ij}|\Theta^* = \theta] = \cos(\theta - \frac{b}{2\pi})A(\kappa)$ 因此，$\max_b \mu_b \geq \cos\left(\frac{1}{4\pi B}\right) A(\kappa)$。

因此，如果 $\lambda$ 在 $\lambda > \frac{1}{e}$ 下固定不变，我们选择一个足够大的正整数 $B$，使得 $\lambda \cos\left(\frac{1}{4\pi B}\right) > \frac{1}{e}$。然后，如果选择 $\kappa$ 使得 $\lambda = \frac{\kappa^2 K^2}{2n}$，组合消息传递算法便可使两种错误概率都趋于零——也就是说，可以在多项式时间内实现强检测。

## 参考文献

[Abb18] Emmanuel Abbe. 社群检测与随机块模型：近期进展。机器学习研究期刊，18(177):1–86，2018年。[ACV14] Ery Arias-Castro 和 Nicolas Verzelen. 稠密随机网络中的社群检测。统计年鉴，页码 940–969，2014年。[AKS98] Noga Alon、Michael Krivelevich 和 Benny Sudakov. 在随机图中寻找大型隐藏团。随机结构与算法，13(3-4):457–466，1998年。

[BBH18] Matthew Brennan、Guy Bresler 和 Wasim Huleihel. 植入稀疏结构问题的可约性与计算下界。载于《学习理论会议》，页码 48–166。PMLR，2018年。[Ben06] John Bentley. 使用冯·米塞斯混合分布和均匀分布对圆形数据建模。硕士论文，2006年。[BI13] Cristina Butucea 和 Yuri I. Ingster. 高维含噪矩阵中稀疏子矩阵的检测。伯努利，19(5B):2652–2688，2013年。[BMNN16] Jess Banks、Cristopher Moore、Joe Neeman 和 Praneeth Netrapalli. 稀疏网络社群检测的信息论阈值。载于《学习理论会议》，页码 383–416。PMLR，2016年。[BR13] Quentin Berthet 和 Philippe Rigollet. 稀疏主成分检测的复杂性理论下界。载于《学习理论会议》，页码 1046–1066。PMLR，2013年。[Cre77] Noel Cressie. 圆和直线上的扫描统计量的一些性质。《应用概率学杂志》，14(2):272–283，1977年。[Cre80] Noel Cressie. 均匀分布下扫描统计量的渐近分布。《概率年刊》，页码 828–840，1980年。[dGJL04] Alexandre d’Aspremont、Laurent Ghaoui、Michael Jordan 和 Gert Lanckriet. 使用半定规划求解稀疏主成分分析的直接形式。《神经信息处理系统进展》，17，2004年。[DLMF] 美国国家标准与技术研究院数学函数数字图书馆。https://dlmf.nist.gov/，2025年3月15日发布的 1.2.4 版。F. W. J. Olver、A. B. Olde Daalhuis、D. W. Lozier、B. I. Schneider、R. F. Boisvert、C. W. Clark、B. R. Miller、B. V. Saunders、H. S. Cohl 和 M. A. McClain 编。[DM14] Yash Deshpande 和 Andrea Montanari. 信息论最优稀疏主成分分析。载于《2014年IEEE国际信息论研讨会》，页码 2197–2201。IEEE，2014年。[DM15] Yash Deshpande 和 Andrea Montanari. 以近线性时间寻找隐藏团。《计算数学基础》，15(4):1069–1128，2015年。[Fis95] Nicholas I. Fisher. 圆形数据的统计分析。剑桥大学出版社，1995年。[FK00] Uriel Feige 和 Robert Krauthgamer. 在半随机图中寻找并认证大型隐藏团。随机结构与算法，16(2):195–208，2000年。[GNW05] Joseph Glaz、Joseph Naus 和 Sylvan Wallenstein. 扫描统计，第1卷。施普林格，2005年。[Hoe63] Wassily Hoeffding. 有界随机变量之和的概率不等式。《美国统计学会杂志》，58(301):13–30，1963年。[HWX17] Bruce Hajek、Yihong Wu 和 Jiaming Xu. 恢复隐藏社群的信息论极限。《IEEE信息论汇刊》，63(8):4729–4745，2017年。

[HWX18] Bruce Hajek、Yihong Wu 和 Jiaming Xu。通过消息传递进行子矩阵定位。机器学习研究期刊，18(186):1–52，2018年。[Jer92] Mark Jerrum。大团难以被 Metropolis过程发现。随机结构与算法，3(4):347–359，1992年。[JL09] Iain M. Johnstone 和 Arthur Yu Lu。关于高维主成分分析的一致性与稀疏性。《美国统计学会杂志》，104(486):682–693，2009年。[Kui60] Nicolaas H. Kuiper。关于圆上随机点的检验。数学研究（论文集），63:38–47，1960年。[LRM18] Lukas Landler、Graeme D Ruxton 和 E Pascal Malkemper。生物学中的圆形数据：有效实施统计程序的建议。行为生态学与社会生物学，72(8):128，2018年。[Mas14] Laurent Massouli´e。社群检测阈值与弱 Ramanujan 性质。载于第四十六届 ACM计算理论研讨会论文集，第694–703页，2014年。[MJ09] Kanti V. Mardia 和 Peter E. Jupp。方向统计学。John Wiley & Sons，2009年。[MK21] Kees Tim Mulder 和 Irene Klugkist。圆周均匀性的贝叶斯检验。统计规划与推断杂志，211:315–325，2021年。[MNS18] Elchanan Mossel、Joe Neeman 和 Allan Sly。块模型阈值猜想的证明。组合数学，38(3):665–708，2018年。[Rao76] JS Rao。基于圆弧长的若干检验。Sankhy¯a：《印度统计学杂志》B辑，第329–338页，1976年。[Ray80] Lord Rayleigh。关于大量具有相同频率和任意相位的振动的合成结果。《伦敦、爱丁堡和都柏林哲学杂志与科学杂志》，10(60):73–78，1880年。[Tsy09] Alexandre B. Tsybakov。非参数估计导论。施普林格统计丛书。施普林格纽约，纽约州，2009年。[VAC15] Nicolas Verzelen 和 Ery Arias-Castro。稀疏随机网络中的社群检测。《应用概率年刊》，25(6):3465–3510，2015年。[Wat61] George S Watson。圆上的拟合优度检验。Biometrika，48(1/2):109–114，1961年。[WNG93] Sylvan Wallenstein、Joseph Naus 和 Joseph Glaz。扫描统计量检测聚类的效能。医学统计，12(19-20):1829–1843，1993年。[WWG89] Sylvan Wallenstein、Clarice R Weinberg 和 Madelyn Gould。季节性事件数据中脉冲的检验。生物统计学，第817–830页，1989年。[WX21] Yihong Wu 和 Jiaming Xu。植入结构的统计问题：信息论与计算极限，第383–424页。剑桥大学出版社，2021年。