低次　方法无法预测鲁棒子空间恢复

HeJia AravindanVijayaraghavan 西北大学 西北大学
he.jia@northwestern.edu aravindv@northwestern.edu

arXiv:2603.02594v1 [stat.ML] 3 Mar 2026

摘要

低次多项式框架在预测平均情形分析和机器学习中高维问题的计算与统计差距方面取得了巨大成功。这一成功催生了低次猜想，该猜想认为，对于广泛的高维统计问题，这种方法能够刻画高效算法的能力与局限。

我们在 $\mathbb{R}^n$ 中发现了一个自然且基本的假设检验问题：该问题可在多项式时间内求解，但低次多项式方法却无法预测其计算可处理性，甚至在次数达到 $k = n^{\Omega(1)}$时也不行。此外，低次矩直到次数 $k = O(\sqrt{\log n/\log\log n})$为止都完全匹配。

我们的问题是一个经过深入研究的鲁棒子空间恢复问题的特例。下界表明，该问题不存在多项式时间算法。与此相反，我们利用分布的反集中性质，给出了一种简单且稳健的多项式时间算法来解决该问题（以及它的含噪变体）。我们的结果表明，低次方法和低次矩无法捕捉基于反集中的算法，这对它们作为计算障碍预测工具的普适性提出了挑战。

1 引言

高维统计和机器学习中的若干问题呈现出显著的统计与计算差距——在这些问题中，只需多项式数量的样本便有可能恢复隐藏结构或信号，但目前尚无已知的多项式时间算法能够做到这一点。此类统计与计算差距已在广泛的问题中得到观察，包括植入团、稀疏主成分分析、社群检测、低秩矩阵与张量分解及估计问题、高斯混合模型、稳健估计等 [Bar+19; BR13; HS17; DKS17; Hop18; KWB19; Wei25]。平均情形分析理论的一个核心目标，是预测此类统计与计算差距何时会出现，并理解高效算法在这类统计推断问题上的极限。不同于最坏情形，在最坏情形中，归约和困难性猜想（例如 $P \neq NP$）构成了我们理解问题的框架；而在平均情形领域，除了少数值得注意的例外 [Reg09; Bru+21; BB19; BBH18]，人们在基于少数核心假设确立强计算困难性方面进展较少。占主导地位的方法是针对特定算法族建立下界，旨在通过针对统计查询算法、低次多项式、凸松弛层级、局部算法和稳定算法的下界，刻画广泛算法范式的极限 [Kea98; Fel+17; Dia+23; Wei25; Hop+17; Gam21]。

低次多项式方法已成为理解统计与计算差距最常用的方法之一。考虑一个假设检验问题，其目标是在 *PLANTED* 模型 P 与 *NULL* 或参考 模型 Q 之间进行区分。在本文中

我们考虑这样一个问题：给定 m 独立同分布样本 $X_1, \ldots\ldots, X_m \in \mathbb{R}^n$ 其中要么是$X_1, \ldots\ldots, X_m \sim P$，要么是 $X_1, \ldots\ldots, X_m \sim Q$。次数为 k 的低次优势为

$$\mathrm{LDA}_{\leq k}^{(m)}(P,Q) = \max_{f:\text{degree-}k \text{ polynomial}} \frac{\left|\mathbb{E}_{P^{\otimes m}}[f(X_1,\ldots,X_m)] - \mathbb{E}_{Q^{\otimes m}}[f(X_1,\ldots,X_m)]\right|}{\sqrt{\mathrm{Var}_{Q^{\otimes m}}[f(X_1,\ldots,X_m)]}} \tag{1}$$

直观地说，这个量衡量了样本的次数为-k 的多项式区分这两个分布的能力。低次优势还通过限制为低次统计量，来近似 P 与 Q之间的似然比。

当低次优势很小（例如，低次优势$_{\leq k}^{(m)}(P, Q) = O(1)$，甚至更好的情况 $o(1)$时，这被视为平均情形困难性的证据 [KWB19；Wei25；Kun25]。在这种情况下，所有次数为k 的多项式都无法区分 P 和 Q，这表明具有相当计算能力的算法也无法成功。虽然低次方法无法涵盖高斯消元和某些基于格的算法 [DK22；Zad+22]，但这些算法利用了丰富的代数结构，且对少量随机噪声十分脆弱。这与低次多项式不同，后者天生能够抵抗随机噪声——这一特性在统计估计问题中尤为相关。

大量相关研究 [Bar+19；HS17；Hop+17] 促成了低次猜想（假设 2.1.5 和猜想 2.2.4）的提出，该猜想出自 Hopkins 的杰出博士论文 [Hop18]（另见 Ding 等人的猜想 1.6 [Din+21]、Moitra 和 Wein 的猜想 2.3 [MW23]），在我们的设定中可以表述如下：

非正式猜想 1。对于由 m 来自 P 的独立同分布样本所定义的“自然”高维检验问题 Q，如果低次优势 $\mathrm{LDA}_{\leq k}^{(m)}$ 在 $m \to \infty$ 时仍有界，那么不存在能解决区分问题、且运行时间为 $\exp(O(k/\mathrm{polylog}(m)))$的（耐噪）算法。

上述猜想通常在更一般的设定下表述，其中检验问题由分布 P_N 和 Q_N 以及输入规模 N指定；在我们的设定中，给定来自分布（ m 的独立同分布样本P 或 Q）且该分布定义在 $\mathbb{R}^n$ 上，对应的输入规模为 $N = mn$。这个猜想是非正式的，因为它没有明确指定可能适用的检验问题类别。尽管最初的猜想针对离散域提出（精确表述见 Hopkins[Hop18] 中的猜想 2.2.4），低次猜想已被用于研究涵盖离散域和连续域的更广泛的高维统计问题 [，例如参见 Din+21；MW23；AV23；Kun25]。

反过来，在实践中通常成立的是：如果 $\mathrm{LDA}_{\leq k}^{(m)} \to \infty$ as $m, n \to \infty$，那么往往存在一个可容忍噪声的判别算法，其运行时间大约为 $(nm)^{O(k/\mathrm{polylog}(mn))}$。低次优势似乎涵盖了现有的大多数算法技术，包括矩方法、谱方法、近似消息传递（AMP）等局部方法、置信传播、多项式阈值函数，甚至某些平方和（SoS）松弛方法 [KWB19; Kun25; Wei25; Hsi+26]。这一猜想在高维统计学的一系列典型问题中展现出显著的预测能力，包括植入团、稀疏主成分分析、张量主成分分析、聚类、尖峰 Wigner 和 Wishart 模型、张量分解、社群检测、高斯混合模型、认证限制等距性质，以及许多其他问题 [例如参见 Hop+17; Din+21; MW23; BKS23; AV23; Dia+23]。另请参阅 Wein [Wei25]和 Kunisky [Kun25]撰写的优秀综述。

然而，Buhai、Hsieh、Jain 和 Kothari 最近的一项令人振奋的成果 [Buh+25] 通过将列表译码纠错码中的一个问题编码为统计估计任务，为一个具体低次猜想的拟多项式时间版本提供了一个反例 [Hop18；HW20；Wei25]，。Buhai 等人 [Buh+25]证明，即使次数为$k = n^{\Omega(1)}$，也不存在低次优势；然而，该问题存在一个拟多项式时间算法。他们还给出了一个

特定检验问题的例子（以单个 $n \times n$ 矩阵为输入）：该问题可以通过特征值计算求解，但对于 $k = \widetilde{O}(n^{1/3})$ 却表现出消失的低次优势。尽管存在这一反例，低次方法仍然是理解高维推断中高效算法极限的最有力框架之一。这引出了以下值得探讨的问题：

问题： 低次优势能否准确预测高维统计问题中自然算法方法的能力范围？是否存在低次优势方法无法捕捉的算法技术？

1.1 我们的结果

我们的主要贡献是提出了一个自然的高维统计问题，在这个问题中，低次优势无法预测计算可处理性。事实上，我们的问题是广为研究的鲁棒子空间恢复问题的一个特例 [HM13; LM18; Bha+19; BK21; Gao+26]。在该问题的一般版本中，存在一个（未知的） d 维子空间 $S \subset \mathbb{R}^n$，我们希望对其进行恢复。给定样本 $X_1, \ldots, X_m \in \mathbb{R}^n$，其中有 α 比例的样本来自一个完全支撑于 S上的分布，其余样本则来自一个完全支撑于 $\mathbb{R}^n$上的分布。当 d 足够小时，鲁棒子空间恢复问题有一种优雅的多项式时间算法，该算法利用了零假设分布 [HM13; Bha+19]的充分反集中性质。相比之下，我们找到了自然分布P（植入分布）和 Q （零假设分布），它们使低次方法在次数达到 $k = n^{\Omega(1)}$ 之前都无法奏效，基于矩的其他方法也同样如此。我们还针对这一设定给出了一种初等的多项式时间算法，并实现了高噪声容忍度。

我们先描述这个假设检验问题。给定来自某个分布 $\mathcal{D}$ 的独立同分布样本。我们的目标是区分以下两个假设：

- NULL: $\mathcal{D}$ 是旋转不变分布 Q_{rot}，由球对称高斯分布的尺度混合给出。形式上，样本 $X \in \mathbb{R}^n$ 的抽取过程如下：首先 $\lambda \sim N(0,1)$，然后 $X \sim N(0, \lambda^2 I)$。
- PLANTED: $\mathcal{D}$ 是一个分布，它在维度至多为 $d = O(1)$的子空间 $S \subset \mathbb{R}^n$ 上至少放置 $\alpha = 1/\mathrm{poly}(n)$的概率质量。

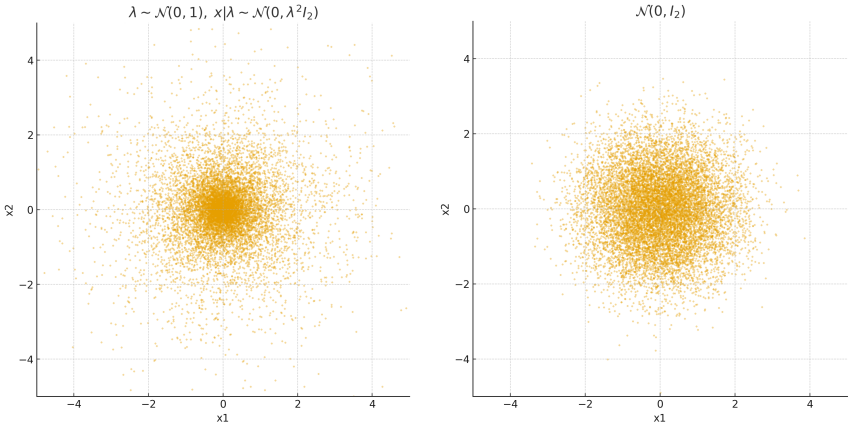


图 1：二维高斯尺度混合与二维标准高斯分布的比较。尺度混合下的径向分布（范数）反集中性更强。

我们的植入分布在一个 d 维子空间 S上具有显著的质量；该设定在 $d = O(1)$时对应于鲁棒子空间恢复问题的最简单设定。事实上，该下界结果甚至在子空间维度 $d = 0$ 即 $S = \{0\}$时仍然成立。零假设分布 Q_{rot} 是旋转不变的：不存在具有非零概率质量的低维子空间。此外，存在一个基本算法可以区分这两种分布：取 $O(d/\alpha)$个样本，并检查是否有 $d+1$ 个点位于一个d维子空间上！

我们的第一个结果表明，分布 NULL 和 PLANTED 的矩在次数不超过 $k = O(\sqrt{\log\ n}/\log\log\ n)$时完全匹配。这也意味着，在相同次数范围内存在低次多项式下界。

定理2 （矩匹配至次数 $k = O(\sqrt{\log n})$）。 由 NULL 模型给出的 $\mathbb{R}^n$ 上的两个分布，以及由上述单位向量 $v \in \mathbb{S}^{n-1}$ 参数化的植入分布族 $\{\mathcal{P}_v : \ v \in \mathbb{S}^{n-1}\}$，满足以下性质：

1. Q_{rot} 是旋转不变的，而 $\mathcal{P}_v$ 是一个混合分布：以概率 $\alpha = 1/\text{poly}(n)$，从一个支撑集位于沿着 v的 $d = 1$ 维子空间上的分布中抽取 X；以概率$1-\alpha$，从一个全维分布中抽取 X。

2. 对于 $k = O(\sqrt{\log\ n/\text{loglog}\ n})$， $\mathcal{P}_v$和 Q_{rot} 的前 k 个矩匹配。因此，对于任意次数为k 的多项式 f： $\mathbb{R}^n \to \mathbb{R}$,

$$|\mathbb{E}_{X\sim\mathcal{P}_v}[f(X)] - \mathbb{E}_{X\sim Q}[f(X)]| = 0.$$

3. 存在一种算法，其运行时间为多项式时间（关于 n），并且以至少 $1 - o(1)$ 的概率区分 P_v 和Q_{rot}。此外，该算法具有鲁棒性，既能应对点的对抗性扰动，也能应对将来自 P_v 的样本中 $p \in [0, 1)$ 比例重新随机化为从 Q_{rot} 中抽取。

正式表述见定理 8 。上述下界表明，依赖于样本上次数为 $k = \widetilde{O}(\sqrt{\log n})$[1] 的低次多项式统计量的算法，无法区分这两个分布。事实上，低次优势$\leq^{(m)}_k = 0$，且与样本数量无关m。这排除了许多基于矩方法的多项式时间算法——对于 n 维的统计问题，基于 k 次矩的算法通常能够在样本复杂度和运行时间方面提供$n^{O(k)}$ 保证。然而，某些特殊的 $\Theta(\log n)$ 次多项式可以在多项式时间内求值；这尤其包括基于对随机矩阵最大特征值进行良好近似的算法。

我们的下一个结果表明，即使次数为 $k = n^{\Omega(1)}$，低次优势也很小。这意味着可能不存在任何多项式时间（甚至拟多项式时间）算法，因此我们的问题在多项式时间内可解这一点尤为有趣。在下文中， $c > 0$ 可以取任意常数，例如 $c = 1/2$。

非正式定理3 (LDA下界，适用于次数不超过 $k = n^{\Omega(1)}$)。 设 $c > 0$ 为任意常数。存在绝对常数 $c_1 > 0$、$c_2 > 0$ ，使得对于 $\alpha = 1/n^c$ 和$m = n^c$ 对数 n，以下结论成立：存在定义在 $\mathbb{R}^n$ 上的旋转不变分布 Q_{rot} 和 P，使得

1. P 在 $X = 0$ 处具有 α 概率质量，而 Q_{rot} 在维度至多为 $n-1$ 的任何子空间上都具有零概率质量。

[1]我们用记号 $\widetilde{O}(f(n))$ 表示 $O(f(n)\text{poly}\log(f(n)))$，省略了自变量的多对数因子。

2. 给定来自该分布的 m 独立同分布样本，对于任意次数 $k \le c_1 n^c$/对数 n，在 m 样本上的低次优势 $LDA_{\le k}^{(m)}(P, Q_{\text{rot}}) \le c_2$。

3. 存在一种算法，可在多项式时间内运行（运行时间为 $O(mn)$），并以至少 $1 - o(1)$ 的概率区分 P和 Q_{rot}。此外，该算法既能抵御点的对抗性扰动，也能应对将比例为 ϵ 的样本重新随机化，使其从 P 改为从 Q_{rot} 抽取。

有界低次优势的正式表述见定理 17；事实上，定理 17 给出了参数 m，α，k 的一个更一般的设定，在该设定下低次优势有界。在上述参数范围 $m\alpha = \Omega(\log n)$ 内，因此，以高概率存在许多位于 $X = 0$ （在鲁棒版本中，则位于其附近）的点，可以通过多项式时间轻松检验。输入大小为 $N = mn = n^{1+c}$。因此，这个鲁棒子空间恢复的特殊情形代表了一个自然的统计问题，而低次方法对它的预测会失效。

我们指出，Hopkins 对离散问题提出的原始猜想 [Hop18] 对对称性（置换不变性）以及零假设分布的乘积形式施加了额外限制 Q_N。在我们的设定中，零假设分布由 m 个来自 Q_{rot}的独立同分布样本构成；其本身具有旋转不变性，这是我们的问题所满足的一种自然对称性条件。尽管 Q_{rot} 并非乘积分布，但我们的输入的零假设分布在各m 样本之间具有乘积形式（事实上，它们是独立同分布的）。因此，当 $m \gg n$时，联合分布会更接近乘积分布。

算法与噪声容忍度。上述定理中 $d = O(1)$ 的互补算法下界3 由第5节中的简单算法达到。该算法抽取$m = O(1/\alpha$个点，并检查是否存在 $d+1$ 个样本近似线性相关。请注意，我们算法结果所针对的植入分布，比下界（定理2 和3）中的特定植入分布更一般。形式化表述见定理20 和24 。

此外，该算法还具有噪声容忍性：它在几种适用于鲁棒子空间恢复问题的自然噪声模型下都具有鲁棒性。首先，从 $p \in [0,1]$ 分布中抽取的样本里，有 PLANTED 比例可以重新随机化，使其改为从 NULL 分布中抽取Q_{rot}，这与许多其他问题中的随机噪声模型类似。 [HW20] 的工作基于纠错码给出了一个反例，该反例采用了这种噪声模型；他们还提出，应考虑对实数域中每个点施加扰动时的鲁棒性，例如，当 NULL 分布为标准球面高斯分布时，对每个元素施加微小的高斯扰动 $N(0, I)$。

对于我们的 NULL 分布 Q_{rot}，由于各点的长度尺度各不相同，因此没有明显且自然的平均情况噪声模型可供选择。我们转而证明，该算法能够抵御~针对每个点的两种自然的对抗性扰动 X X。第一类扰动~的相对误差幅度有一个常数上界，即 $\|X - X\|_2 \le \epsilon \|X\|_2$ 对于$\epsilon = O_d(1)$。第二类扰动（最坏情况加性误差扰动）对~$\|X - X\| \le \eta$的幅度设定绝对上界（当 $S = \{0\}$时，这种方式更有意义）。在此模型中，即使与低效算法相比，该算法所允许的 ϵ，η范围也与最佳可能范围相差不超过一个常数因子。形式化表述请见第20 条定理（相对误差最坏情况扰动）和第24条定理（加性误差最坏情况扰动），均见第5节。

低次多项式无法捕捉反集中性。 我们的结果共同指出了一种自然的候选情形（鲁棒子空间恢复），在这种情形下，低次多项式方法会失效。或许更有趣的是，这表明低次方法无法充分

捕捉基于分布反集中性质的算法。这一点不同于集中性质和尾界，后两者通常可由分布的高阶矩捕捉。鲁棒子空间恢复算法 [HM13；Bha+19] 确实依赖于识别出一小部分但非平凡的 α 样本，这些样本位于一个非常小的区域内——反集中性良好、分布广泛的分布不太可能存在这样的高密度区域。事实上， [Bak+24] 最近的工作研究了反集中性认证问题，旨在为一般分布下的聚类问题提供保证。他们构造低次多项式，以认证不存在包含 α 概率质量 $> C\alpha$ 的小宽度小半空间。然而，认证多项式的次数会增长至 $\Omega(1/\alpha^2)$。对于鲁棒子空间恢复，我们处于 $\alpha = 1/\mathrm{poly}(n)$的情形。这似乎对这类技术构成了挑战，包括 [BK21；Bak+24]。

用于分离的新的候选实例 本文的主要贡献是提出了一个自然的高维问题（鲁棒子空间恢复的一个平均情形设定），它可作为证明不同算法类别之间分离结果的候选问题。我们认为，我们的候选实例（及其变体）可能会给其他算法类别带来挑战，包括统计查询（SQ）模型和平方和（SoS）。具体而言，我们猜想，本文研究的平均情形鲁棒子空间恢复问题（例如定理 2中的问题，或在计算上可处理的范围内，在一个 α质量的 d维子空间上的推广问题）对于符合 SQ 框架的多项式时间算法而言，也是一个困难实例。

这一猜想基于两个主要观察。首先，我们的低次下界（定理 2）表明，单样本分布的前 $k = O(\sqrt{\log\, n/\mathrm{loglog}\, n})$ 阶矩完全匹配。这种完美的矩匹配强有力地表明，任何依赖低次统计量的算法都无法区分这些分布。在某些设定下，低次方法下界与此的联系也已为人所知 [Bre+21]。其次，我们简单的算法以及现有的鲁棒子空间恢复算法 [HM13；Bha+19] 在 α 为次常数时，都依赖于在采样点的一个相对较小子集中识别近似线性相关。人们认为，SQ 模型无法捕捉这一过程，因为该模型仅限于估计整个总体的聚合统计性质。此外，即便对于平方和这类强大的算法类别，要获得此类反集中证书似乎也颇具挑战；目前的技术似乎要求次数界为 $\mathrm{poly}(1/\alpha)$ [Bak+24]。

与非高斯分量分析（NGCA）的比较。 我们的结果不同于 NGCA 等设定中典型的困难性结果 [Bla+06; DKS17; Dia+23]。NGCA 的目标通常是在高维中找到单个隐藏非高斯方向。基于 NGCA 问题及其变体的下界构造，在为若干平均情形问题证明困难性证据方面发挥了重要作用 [DKS17; Bru+21]。例如，近期工作[Dia+23] 表明，当非高斯信号被限制在余维为 1 的子空间（即一个（$n-1$）维子空间）中时，子空间恢复问题对于低次多项式和统计查询算法是困难的。然而，在我们的构造中，我们在 1 维或 0 维子空间上有 α 比例的质量；这使得整个余维为 $\geq n-1$的子空间成为非高斯的。事实上，这种不同的特点是意料之中的——尽管 NGCA 被认为在计算上是困难的，但我们这种子空间为低维的构造在算法上是容易的。这使得设计一个与参考分布的所有混合矩都匹配的植入分布变得显著更加困难。我们希望这一与 NGCA 实例感觉显著不同的新实例，将有助于理解不同算法框架的相对能力。

1.2 技术概述

我们的主要结果是一种分离，我们分两部分加以建立。首先，我们构造了一个假设检验问题，其中一族植入分布 P 与一个旋转不变的零假设分布 Q_{rot} 对于低次多项式是不可区分的。我们以两种强形式建立了这种不可区分性：直至次数 $k = O(\sqrt{\log n})$ 的精确矩匹配，以及对次数 $k = n^{\Omega(1)}$ 而言趋于消失的低次优势。这两项结果都利用了分布 Q_{rot} 良好的结构性质。其次，我们给出了一个简单、高效的算法，能够稳健地解决同一问题。

矩匹配构造。 我们第一个下界的核心是矩匹配这一主张。我们首先将零假设分布 Q 构造为高斯尺度混合。这一选择至关重要。标准高斯分布具有旋转不变性，但其长度（$\|X\|_2$）高度集中。我们要求零假设分布 Q 满足两个性质：(1) 旋转不变性，以及 (2) 其长度（尺度）的反集中性。

我们选择的 Q（一种尺度为 $\lambda \sim \mathcal{N}(0,1)$ 的尺度混合）同时满足两者。由于 P 和 Q 都被设计为在与信号无关的 $(n-1)$ 维子空间中旋转不变，由此得出两个结论：1）零假设分布 Q 不提供关于隐藏子空间的任何信息。2）矩匹配问题大幅简化。我们可以沿无关子空间中的任意向量分析矩，从而将 n 维问题简化为二维问题（一个信号坐标，一个无关坐标）。

接下来，我们着重匹配 P 和 Q的二维矩向量。我们的策略是通过有效扰动 Q的矩，给出一个非构造性证明，以证明 P 的存在。主要的技术难点在于证明这组经过扰动的新矩是有效的，也就是说，确实存在一个分布 μ_2 由这些矩定义。换言之，我们要证明 Q 的矩与矩多胞体的边界相距足够远。为证明其存在性，我们采用 Carath´eodory 风格的几何界。这表明，只要经过扰动的目标矩向量位于所有可能矩向量构成的凸包内，它就是有效的。我们通过证明它具有正 Tukey 深度来说明这一点。此时， Q 的尺度/长度的反集中性质至关重要，因为这一性质会由 Q尺度的二维分布继承。

我们首先建立一个强的双侧反集中界，这一结果依赖于 Carbery-Wright 不等式。接着，关键的一步是将这个双侧界转化为所需的单侧界，以便对 Tukey 深度给出下界。以下引理概括了这一结果。

引理 4 （与引理14相同）。 假设 ν 是定义在 $\mathbb{R}^n$ 上的概率测度，满足以下反集中条件：对于任意多元多项式 p： $\mathbb{R}^n \to \mathbb{R}$ 其次数至多为 k,

$$\forall \delta > 0,\ \mathop{\mathbb{P}}_{x\sim\nu}\left[|p(x) - \mathbb{E}_\nu[p]| \le \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \le \eta_k(\delta).$$

那么对于任意满足 $\eta_k(\delta_1) \le 1/2$ 的 $\delta_1 > 0$，我们有

$$\mathop{\mathbb{P}}_{x\sim\nu}\left[p(x) \ge \mathbb{E}_\nu[p(x)] + \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \ge \frac{\delta_1^2}{8} - \eta_k(\delta).$$

这一正 Tukey 深度保证扰动矩向量位于凸包的内部，从而证明了我们的矩匹配实例存在，其次数最高可达$k = O(\sqrt{\log n/\log\log n})$。零假设分布 Q_{rot} 的这些性质也有助于证明当多项式次数为 k时，有界低次优势成立。

将低次方法的困难性推广到多项式次数。虽然我们的矩-~匹配构造能完全迷惑次数最高为 $k = O(\sqrt{\log n})$的多项式，但标准启发式表明，如果低次下界能够排除次数为 $k = \overline{\omega(\log(n))}$的多项式，那么它们就是困难性的有力证据。为了有力证明低次框架无法预测这个问题的可处理性，我们证明了对于更高次数（即 $k = n^{\Omega(1)}$），低次优势仍有上界。我们使用与矩匹配结果中相同的零假设分布 Q_{rot}。其关键结构性质——旋转不变性和长度的强反集中性——对于证明低次优势的上界仍至关重要。

我们分析一种简化的植入分布 $P = (1-\alpha)Q + \alpha\Delta_0$，它是零假设分布与原点 0 处的点质量的混合。利用 Q长度的反集中性质，我们首先证明单样本优势 $\mathrm{LDA}_{\leq k}^{(1)}$ 为 $O(1)$。以下命题是证明单样本优势上界的关键：

命题5 ．考虑任意一个常数项为0的次数为 k 的多项式 f。那么我们有 ve

$$\frac{|\mathbb{E}_Q[f(X)]|}{\sqrt{\mathrm{Var}_Q(f(X))}} \leq O(\sqrt{k}),$$

其中 $Q = Q_{\mathrm{rot}}$ 是旋转不变高斯尺度混合分布。

我们将此命题与张量化论证（引理 18）结合，将结论推广到多样本情形，证明多样本优势仍受一个较小常数所限。

上述命题中的上界（针对单个样本）与维度 n 无关。值得注意的是，对于球形高斯分布 $N(0, I)$，尽管它是旋转不变的，该命题仍不成立；例如，二次多项式 $f(X) = X^\top X$ 集中在 $n \pm O(\sqrt{n})$ 附近。我们利用 Q_{rot} 的一个重要性质：可以将 $X \sim Q_{\mathrm{rot}}$ 写成两个相互独立的随机变量 $X = \lambda g$ 的乘积，其中 $g \in \mathbb{R}^n$ 是一个随机向量（取自旋转不变分布），而 λ 是表示尺度的随机变量；在我们的情形中，它也取自具有良好反集中性质的正态分布。这使我们能够证明上述命题（见定理17的证明）。值得注意的是，低次下界构造也适用于 $n = 1$ 维度（此时可以将 $g \sim_u \{-1, +1\}$ 视为 $n = 1$ ）。尺度分布 λ 中采用正态随机变量 λ 并非关键；也可以使用满足命题5的其他尺度分布选择（可能对 k 有不同的依赖关系），以获得其他理想性质。

简单鲁棒算法。与低次数困难性相反，我们通过提出两种简单算法，证明该问题在计算上是可处理的。这些算法对点进行采样并检查其是否近似线性相关。对于 d 维隐藏子空间，第一种算法通过对样本归一化，并检测任意大小为 $d+1$ 的子集的第 $(d+1)$ 个奇异值是否低于常数阈值，来处理对抗性相对扰动。第二种算法将算法的保证扩展到最坏情况加性扰动的情形。两种算法都依赖于 Q_{rot} 的反集中性质，以确保随机元组鲁棒线性无关（可靠性）。

2 预备知识

2.1 将提升矩匹配推广至多样本情形

矩匹配。我们称定义在 $\mathbb{R}^n$ 上的两个分布 P 和 Q 在次数不超过 k 时具有匹配的矩，如果对于任意总次数至多为 k 的多元单项式 $M\colon \mathbb{R}^n \to \mathbb{R}$，都有

$$\mathbb{E}_{X\sim P}[M(X)] = \mathbb{E}_{X\sim Q}[M(X)].$$

低次优势的定义涉及关于来自某个分布的多个独立同分布样本的多项式。下面的标准引理表明，上文定义的单样本矩匹配性质，直接意味着相应的多样本分布下任意低次多项式的期望也相等。这使我们能够将低次下界的构造聚焦于单样本分布。

引理 6. 设 P 和 Q 是定义在 $\mathbb{R}^n$ 上的两个分布，其矩在次数不超过 k时相匹配。设 m为样本数， $P^{\otimes m}$ 和 $Q^{\otimes m}$ 分别为 m 个独立同分布样本对应的乘积分布。那么，对于任意总次数至多为 k的多项式 $f\colon \mathbb{R}^{m\times n} \to \mathbb{R}$，有

$$\mathbb{E}_{(X_1,\dots,X_m)\sim P^{\otimes m}}[f(X_1,\dots,X_m)] = \mathbb{E}_{(Y_1,\dots,Y_m)\sim Q^{\otimes m}}[f(Y_1,\dots,Y_m)].$$

证明。 根据期望的线性性，只需证明任意总次数至多为 k的单项式 $M(X_1, \dots..., X_m)$都满足该等式。我们可以将任意这样的单项式 M 写成各样本对应的单项式之积:

$$M(X_1,\dots,X_m) = \prod_{i=1}^{m} M_i(X_i)$$

其中总次数为 $\sum_{i=1}^{m}\deg(M_i) \le k$，因此每个 $\deg(M_i) \le k$。由样本的独立性,

$$\mathbb{E}_{P^{\otimes m}}[M] = \mathbb{E}_{P^{\otimes m}}\left[\prod_{i=1}^{m} M_i(X_i)\right] = \prod_{i=1}^{m}\mathbb{E}_{X_i\sim P}[M_i(X_i)].$$

根据矩匹配假设,

$$\prod_{i=1}^{m}\mathbb{E}_{X_i\sim P}[M_i(X_i)] = \prod_{i=1}^{m}\mathbb{E}_{Y_i\sim Q}[M_i(Y_i)] = \mathbb{E}_{Q^{\otimes m}}[M].$$

这证明了该结论。 ■

2.2 埃尔米特多项式

我们使用标准的概率论中的埃尔米特多项式，记作 $\mathrm{He}_m(x)$。在分析中，我们将使用归一化埃尔米特多项式，其定义如下:

$$h_m(x) = \frac{\mathrm{He}_m(x)}{\sqrt{m!}}.$$

这些归一化多项式最重要的性质是，它们构成了关于标准高斯测度 $\mathcal{N}(0, 1)$ 的平方可积函数空间的一组完备正交归一基，也就是说，对于任意 $i, j \ge 0$,

$$\mathbb{E}_{x\sim\mathcal{N}(0,1)}[h_i(x)h_j(x)] = \begin{cases} 1, & i = j, \\ 0, & i \ne j. \end{cases}$$

这种正交归一性也适用于更高维度。作为参考，这些归一化埃尔米特多项式还具有以下标准恒等式，它将它们与单项式基联系起来：

$$h_m(x) = \sum_{j=0}^{\lfloor \frac{m}{2} \rfloor} a_{m,j} x^{m-2j}, \quad a_{m,j} = \frac{(-1)^j \sqrt{m!}}{2^j j!\,(m-2j)!}. \tag{2}$$

3 矩匹配实例

在本节中，我们的目标是构造一个植入分布，使其前 k 阶矩与旋转不变的零假设分布的前阶矩一致 Q，同时仍然隐藏一个一维信号方向。具体而言，我们使用尺度参数 λ 来控制整体的球面扩散，并使用标量 z 在一个坐标中承载信号；该构造使我们能够调节一小部分质量 α_k，使其位于一个一维子空间上，同时匹配截至次数 k的所有混合矩。我们还利用第 3.1节介绍的工具，证明植入分布与零假设分布的矩向量完全匹配。

定义 7（高斯尺度混合）。一个 n 维的高斯尺度混合是定义在 $\mathbb{R}^n$ 上的 X 分布，其中其尺度参数 λ 由分布 μ 控制：

$$\lambda \sim \mu;\ X \mid \lambda \sim \mathcal{N}(0, \lambda^2 I_n).$$

定理 8。设 Q 是一个高斯尺度混合，其高斯尺度 $\mu = \mathcal{N}$（0、1）如定义 7 所示。对于任意正整数 k，存在 $\alpha_k \in$（0、1），满足 $\alpha_k \geq k^{-O(k^2)}$，以及分布（$z$） μ_2（λ、z），它们定义了一个新分布 P，其前 k 阶矩与 Q 的前 k 阶矩匹配。对于 $X =$（$\bar{X}$、X_n），P 在 $\mathbb{R}^n$ 上定义如下，其中 $\bar{X} \in \mathbb{R}^{n-1}$、$X_n \in \mathbb{R}$:

- 以概率 α_k，$X = 0$，$\bar{} X_n \sim \mu_1$，
- 以概率 $1 - \alpha_k$，$(\lambda,\ z) \sim \mu_2$；$\bar{X} \mid \lambda \sim \mathcal{N}(0,\ \lambda^2 I_{n-1})$ $X_n = z$。

我们的证明策略是首先建立矩匹配条件；借助旋转不变性，这将任务简化为只匹配两个坐标的混合矩。我们将证明，植入分布 P 的矩可以表示为两个分量的凸组合。主要技术难点在于证明：由分布 μ_1 和 μ_2定义的植入分布可以被构造为满足所得的矩方程组。为此，我们将该问题重新表述为寻找目标矩向量 θ，使其位于相关矩向量 Φ的支撑集的凸包之内。随后，我们利用与反集中和矩空间的几何相关的有力工具，这些工具发展于第 3.1 节，以证明这样的分布 μ_2 必定存在。

定理 8 的证明。我们首先建立矩匹配条件。我们要证明，对于任意单位向量 $v = (\bar{v},\ v_n)$ $\bar{v} \in \mathbb{R}^{n-1}$，$v_n \in \mathbb{R}$ 以及任意 $\ell \in [k]$，

$$\mathbb{E}_{X \sim P}[\langle X, v \rangle^\ell] = \mathbb{E}_{X \sim Q}[\langle X, v \rangle^\ell].$$

展开内积可得：

$$\mathbb{E}[\langle X, v \rangle^\ell] = \mathbb{E}\left[(\bar{X} \cdot \bar{v} + X_n v_n)^\ell\right] = \sum_{\ell_1, \ell_2 \geq 0, \ell_1 + \ell_2 = \ell} \binom{\ell}{\ell_1} \mathbb{E}[(\bar{X} \cdot \bar{v})^{\ell_1} (X_n v_n)^{\ell_2}].$$

根据定义， Q是零均值高斯分布的混合，因此它是旋转不变的。 P的前 $n-1$ 个坐标的边缘分布也是旋转不变的，因此只需对任意 v证明该等式。不失一般性，我们可以取 $\bar{v}=e_1$。因此，这等价于证明：对任意满足 $\ell_1+\ell_2\le k$的 $\ell_1,\ell_2\ge 0$，

$$\mathbb{E}_P[x_1^{\ell_1}x_n^{\ell_2}]=\mathbb{E}_Q[x_1^{\ell_1}x_n^{\ell_2}]. \tag{3}$$

此外，这两个分布都是中心对称的。这意味着任何奇数阶矩都为零。因此，我们只需匹配两个 ℓ_1 和 ℓ_2 均为偶数的矩。(3) 的右端为

$$\mathbb{E}_Q[x_1^{\ell_1}x_n^{\ell_2}]=\mathbb{E}_{\lambda\sim\mu}\left[\mathbb{E}_{x\sim\mathcal{N}(0,\lambda^2 I)}[x_1^{\ell_1}x_n^{\ell_2}]\right]=\mathbb{E}_{\lambda\sim\mu}\left[c_{\ell_1}c_{\ell_2}\lambda^{\ell_1+\ell_2}\right]=c_{\ell_1}c_{\ell_2}c_{\ell_1+\ell_2}$$

其中， c_ℓ是标准高斯分布 ℓ的阶矩， $\mathcal{N}(0,\ 1)$。当 $\ell_1=0$时，(3) 的左端为

$$\mathbb{E}_P[x_1^{\ell_1}x_n^{\ell_2}]=\alpha_k\mathbb{E}_{z\sim\mu_1}[z^{\ell_2}]+(1-\alpha_k)\mathbb{E}_{z\sim\mu_2}[z^{\ell_2}].$$

当 $\ell_1\neq 0$ 时，我们有

$$\mathbb{E}_P[x_1^{\ell_1}x_n^{\ell_2}]=(1-\alpha_k)\mathbb{E}_{(\lambda,z)\sim\mu_2}\left[\mathbb{E}_{x_1\sim\mathcal{N}(0,\lambda^2)}[x_1^{\ell_1}z^{\ell_2}]\right]=(1-\alpha_k)c_{\ell_1}\mathbb{E}_{(\lambda,z)\sim\mu_2}[\lambda^{\ell_1}z^{\ell_2}].$$

令 $\mathcal{I}=\{(\ell_1,\ \ell_2)$: $\ell_1,\ \ell_2$为偶数 $0<\ell_1+\ell_2\le k\}$且 $D=|\mathcal{I}|$。令 $\nu(\lambda,\ z)$为第3.1节中定义的分布。定义随机变量 $\Phi(\lambda,\ z)\in\mathbb{R}^D$，其坐标为

$$\Phi_{\ell_1,\ell_2}(\lambda,z)=\lambda^{\ell_1}z^{\ell_2}.$$

那么 $\mathbb{E}_\nu[\Phi_{\ell_1,\ell_2}(\lambda,\ z)]=c_{\ell_2}c_{\ell_1+\ell_2}$。定义我们的目标向量 $\theta\in\mathbb{R}^D$

$$\theta_{\ell_1,\ell_2}=\mathbb{E}_{(\lambda,z)\sim\mu_2}[\lambda^{\ell_1}z^{\ell_2}]=\begin{cases}\frac{c_{\ell_2}^2-\alpha_k\mathbb{E}_{z\sim\mu_1}[z^{\ell_2}]}{1-\alpha_k}, & \ell_1=0,\\ \frac{c_{\ell_2}c_{\ell_1+\ell_2}}{1-\alpha_k}, & \ell_1\ge 1.\end{cases}$$

令 $\beta\in\mathbb{R}^D$ 为任意系数向量，满足 $\|\beta\|=1$。那么

$$p_\beta(\lambda,z):=\langle\beta,\Phi\rangle=\sum_{\mathcal{I}}\beta_{\ell_1,\ell_2}\lambda^{\ell_1}z^{\ell_2}.$$

以及 $\mathbb{E}_\nu[p_\beta]=\sum_{\mathcal{I}}\beta_{\ell_1,\ \ell_2}c_{\ell_2}c_{\ell_1+\ell_2}$。令 $\theta'\in\mathbb{R}^D$ 为

$$\theta'_{\ell_1,\ell_2}=(1+\delta)c_{\ell_2}c_{\ell_1+\ell_2}.$$

因此我们有 $\langle\beta,\ \theta'\rangle=(1+\delta)\ \mathbb{E}_\nu[p_\beta(\lambda,\ z)]$。根据推论 15，我们知道对于任意 $\delta\le 1/k^{Ck^2}$，

$$h_\Phi(\theta')=\inf_{\beta:\|\beta\|=1}\mathbb{P}_\Phi[\langle\beta,\Phi-\theta'\rangle\ge 0]=\inf_{\beta:\|\beta\|=1}\mathbb{P}[p_\beta\ge(1+\delta)\mathbb{E}_\nu[p]]\ge\eta_k(\delta)>0.$$

根据定理 13， θ' 属于随机（$\lambda,\ z$） $\sim\nu$的矩向量 Φ（$\lambda,\ z$）构成的凸集。根据引理 16，取 $r=r$（h_Φ（θ'））>0，使得 B（$\theta',\ r$）$\subseteq$属于 conv(supp(Φ))。令 $\frac{1}{1-\alpha_k}=1+\delta$。对于$\ell_1\ge 1$ 坐标，$\theta'_{\ell_1,\ \ell_2}=\theta_{\ell_1,\ \ell_2}$。对于 $\ell_1=0$， $\theta_0,\ \ell_2=\theta'_0,\ \ell_2-(1+\delta)\ \alpha_k\mathbb{E}_{z\sim\mu_1}[z^{\ell_2}]$。选取 μ_1，使得$\mathbb{E}_{z\sim\mu_1}[z^{\ell_2}]$ 将 $\ell_1=0$ 块移动至多 r，从而 $\theta\subseteq B$（$\theta',\ r$）$\subseteq$属于 conv(supp(Φ))。因此，存在定义在（$\lambda,\ z$）上的有效分布 μ_2，满足 $\mathbb{E}_{\mu_2}[\Phi]=\theta$，也就是说，所有要求的混合矩都与式（3）中的一致。

■

3.1 反集中与 Tukey 深度

如定理 8 的证明中所述，我们的矩匹配构造关键在于证明某个特定分布 μ_2 存在。为此，我们必须证明，某个特定的矩目标向量 θ 位于随机矩向量 $\Phi(\lambda,\ z)$的支撑集的凸包内。本节将介绍证明这一结论所需的反集中工具。

我们先定义一个二维分布 ν，其定义在 (λ, z) 上，并将作为第 $\Phi(\lambda, z)$ 节中随机矩向量 3 的基础。考虑定义在 ν 上$\{(\lambda, z)$ 上的联合分布：$\lambda, z \in \mathbb{R}\}$ 其定义如下：

1. $\lambda \in \mathbb{R}$ 服从标准高斯分布 $\mathcal{N}(0, 1)$
2. 在给定 λ的条件下，$z \in \mathbb{R}$ 服从 $\mathcal{N}(0, \lambda^2)$。

换言之，该联合分布为

$$\nu(\lambda, z) = \frac{1}{\sqrt{2\pi}} e^{-\lambda^2/2} \cdot \frac{1}{|\lambda|\sqrt{2\pi}} e^{-z^2/(2\lambda^2)}. \tag{4}$$

基于上述二维测度 ν，构造定义在 $\mathbb{R}^n$ 上的参考分布（或零假设分布）。参数 λ 将决定该分布的尺度，而 z 将决定沿某个特殊一维投影的分量。

3.2 反集中界

分布的以下反集中性质 ν 对于下界构造至关重要。它确保在参考分布的任意低次多项式下，没有任何单个取值（或狭窄区间）具有过多的概率质量。我们将在第3.3 节中利用这一性质，为我们的矩向量建立正 Tukey 深度。这是证明其均值位于所有有效矩向量的凸包内部的关键。

引理 9。 假设 $k \in \mathbb{N}$。考虑次数为 k 的多项式

$$p(\lambda, z) = \sum_{\ell_1,\ell_2\ even, 0<\ell_1+\ell_2\le k} \beta_{\ell_1,\ell_2} \lambda^{\ell_1} z^{\ell_2}$$

其中 β 是系数向量 $\beta = (\beta_{\ell_1,\ \ell_2}:\ \ \ell_1, \ell_2 \in \{0, 1, \ldots, k\}$ 使得 $\ell_1 + \ell_2 \le k)$。那么存在一个通用常数 C，使得

$$\forall \delta > 0,\ \sup_{t\in\mathbb{R}} \mathop{\mathbb{P}}_{(\lambda,z)\sim\nu} \left[|p_\beta(\lambda, z) - t| \le \delta \cdot |\mathbb{E}[p_\beta]| \right] \le C \cdot k \cdot \left(\sqrt{k}\delta\right)^{\frac{1}{2k}}. \tag{5}$$

引理 9 的证明基于以下针对高斯分布的 Carbery-Wright 反集中不等式。

定理 10 (Carbery-Wright, [CW01])。 设 p: $\mathbb{R}^n \to \mathbb{R}$ 是一个总次数为k的实值多项式，且 ν 是定义在 $\mathbb{R}^n$上的对数凹概率测度。存在一个通用常数 C，使得对于任意 $\delta > 0$,

$$\sup_{t\in\mathbb{R}} \mathop{\mathbb{P}}_{x\sim\nu} \left[|p(x) - t| \le \delta \sqrt{\mathrm{Var}(p(x))} \right] \le C \cdot k \cdot \delta^{1/k}.$$

为了应用 Carbery-Wright 不等式，我们首先证明以下引理，用于比较高斯分布下低次多项式的均值和方差。定理 q 仅为次数 λ 的多项式这一单变量情形是一个特例，也用于界定定理 17 中的低次优势。

引理 11. 令 q 为 (λ, g) 上的一个次数至多为 k 且零常数项的多项式。在$(\lambda, g) \sim N(0, I_2)$下，

$$|\mathbb{E}[q(\lambda, g)| \leq C\sqrt{k\mathrm{Var}(q(\lambda, g))}.$$

证明。令

$$q(\lambda, g) = \sum_{\ell_1,\ell_2:0<\ell_1+\ell_2\leq k} \beta_{\ell_1,\ell_2}\lambda^{\ell_1}g^{\ell_2}.$$

我们也可以使用第 $q(\lambda, g$ 节中定义的归一化厄米基来表示 2.2。设 γ_{i_1,i_2} 是与 $h_{i_1}(\lambda)h_{i_2}(g)$相对应的系数，即，

$$q(\lambda, g) = \sum_{0\leq i_1,i_2\leq k} \gamma_{i_1,i_2}h_{i_1}(\lambda)h_{i_2}(g).$$

由于 $h_{i_1}(\lambda)h_{i_2}(g)$ 在标准高斯分布下正交归一，

$$\mathbb{E}_{N(0,I_2)}[q(\lambda, g)] = \sum_{0\leq i_1,i_2\leq k} \gamma_{i_1,i_2}\mathbb{E}_{N(0,I_2)}[h_{i_1}(\lambda)h_{i_2}(g)] = \gamma_{0,0},$$

$$\mathrm{Var}_{N(0,I_2)}(q(\lambda, g)) = \sum_{0\leq i_1,i_2\leq k} \gamma_{i_1,i_2}\mathbb{E}_{N(0,I_2)}[h_{i_1}(\lambda)^2h_{i_2}(g)^2] - \gamma_{0,0}^2 = \sum_{i_1,i_2\in\mathcal{I}'} \gamma_{i_1,i_2}^2$$

wh其中 $\mathcal{I}' = \{(i_1, i_2) \neq 0 \colon i_1, i_2$即使 $0 \leq i_2 \leq i_1 \leq k\}$。根据公式 (2)，我们有

$$\beta_{0,0} = \sum_{0\leq i_1,i_2\leq k, i_1,i_2 \text{ even}} a_{i_1,\frac{i_1}{2}}a_{i_2,\frac{i_2}{2}}\gamma_{i_1,i_2}, \quad a_{i,\frac{i}{2}} = \left(-\frac{1}{2}\right)^{\frac{i}{2}} \frac{\sqrt{i!}}{\left(\frac{i}{2}\right)!}.$$

因此，

$$\beta_{0,0} = \gamma_{0,0} + \sum_{i_1,i_2\in I^*} a_{i_1,\frac{i_1}{2}}a_{i_2,\frac{i_2}{2}}\gamma_{i_1,i_2}.$$

给定常数项 $\beta_{0,0} = 0$，我们有恒等式

$$\gamma_{0,0} = -\sum_{i_1,i_2\in I^*} a_{i_1,\frac{i_1}{2}}a_{i_2,\frac{i_2}{2}}\gamma_{i_1,i_2}.$$

于是根据柯西-施瓦茨，我们有

$$\frac{\gamma_{0,0}^2}{\sum_{i_1,i_2\in I^*}\gamma_{i_1,i_2}^2} \leq \sum_{i_1,i_2\in I^*} a_{i_1,\frac{i_1}{2}}^2 a_{i_2,\frac{i_2}{2}}^2.$$

根据斯特林近似，

$$a_{i,\frac{i}{2}}^2 = \frac{i!}{2^i(\frac{i}{2})!\,(\frac{i}{2})!} \sim \sqrt{\frac{2}{\pi i}}.$$

因此，我们可以得到如下上界

$$\begin{aligned}
\sum_{i_1,i_2\in I^*} a_{i_1,\frac{i_1}{2}}^2 a_{i_2,\frac{i_2}{2}}^2 &\leq \left(\sum_{1\leq i\leq k, i \text{ even}} a_{i,\frac{i}{2}}^2\right)^2 - a_{0,0}^2 \\
&\leq \left(\sum_{1\leq i\leq k, i \text{ even}} \sqrt{\frac{2}{\pi i}} + 1\right)^2 - 1 \\
&\leq \frac{2k}{\pi} + 2\sqrt{\frac{2k}{\pi}}
\end{aligned}$$

然后，对某个常数 C，$\overline{|\mathbb{E}[q]|} \leq C\sqrt{k\mathrm{Var}(q)}$。 ■

引理的证明 *9.* 考虑联合分布（$\lambda,\ g$） $\sim N(0,\ I_2)$。那么（$\lambda,\ \lambda g$） $\sim \nu$。令$q(\lambda,\ g) = p(\lambda,\ \lambda g)$。要证明以下结论是等价的

$$\forall \delta > 0,\ \sup_{t\in\mathbb{R}} \mathop{\mathbb{P}}_{(\lambda,g)\sim\mathcal{N}(0,1)} \Big[|q(\lambda,g)-t| \le \delta\cdot|\mathbb{E}[q]|\Big] \le C\cdot k\cdot\Big(\sqrt{k}\delta\Big)^{\frac{1}{2k}}.$$

根据引理 11，$|\mathbb{E}[q]| \le C\sqrt{k\mathrm{Var}(q)}$ 对某个常数 C。于是由 Carbery-Wright 可得，

$$\begin{aligned}\forall \delta > 0,\ \sup_{t\in\mathbb{R}} \mathop{\mathbb{P}}_{(\lambda,z)\sim\nu} \Big[|q(\lambda,z)-t| \le \delta\cdot|\mathbb{E}[q]|\Big] &\le \sup_{t\in\mathbb{R}} \mathop{\mathbb{P}}_{(\lambda,z)\sim\nu} \Big[|q(\lambda,z)-t| \le \delta\cdot c\sqrt{k\ \mathrm{Var}(q)}\Big] \\ &\le C\cdot k\cdot(\sqrt{k}\delta)^{\frac{1}{2k}}.\end{aligned}$$

■

3.3 Tukey 深度与矩匹配分布的存在性

接下来，我们将引理 9 中的反集中性质与第3节矩匹配证明所需的凸几何联系起来。我们的目标是证明，目标矩向量 θ属于 $\Phi(\lambda, z)$ 的支撑集的凸包。

我们首先介绍 Tukey 深度的概念，这是一种衡量数据中心性的几何量。然后，我们使用一个 Carath´eodory 型定理（定理 13），正式建立正 Tukey 深度与集合中随机样本的凸包内存在一个向量之间的联系。

定义 12 (Tukey 深度)。给定一个随机变量 Φ *in* $\mathbb{R}^D$ 和一个点 $\theta \in \mathbb{R}^D$，θ 的 *Tukey* 深度定义为

$$h_\Phi(\theta) = \inf_{u:\|u\|_2=1} \mathop{\mathbb{P}}_{\Phi}\Big[\langle u, \Phi-\theta\rangle \ge 0\Big]. \tag{6}$$

Tukey [Tuk75] 提出的上述概念在统计学中发挥着重要作用。Tukey 深度还与随机多胞体的 Carath´eodory 型界有关，具体如下。要证明 $\theta \in \mathrm{conv}(\{\Phi_1,\dots,\Phi_m\})$，根据超平面分离定理，只需证明，对于每个˜方向 u，都存在 $i\in[m]$ 使得 $\langle u,\Phi_i\rangle \ge \langle u,\theta\rangle$。因此，如果我们从 Φ 的分布中独立同分布地抽取 $m = O(D/h_\Phi(\theta))$ 个随机点 $\Phi_1,\dots,\Phi_m$，就可以结合 Chernoff-Hoeffding 界以及 $\mathbb{R}^D$ 中单位向量上的一个 ϵ-网，论证 $\theta \in \mathrm{conv}\{(\Phi_1,\dots,\Phi_m)\}$ 以高概率成立。[HLO23]中的下一个定理给出了一个更紧的界。

定理 13 (Carath´基于 Tukey 深度的随机多胞体 Carathéodory 界，[HLO23])。考虑一个随机向量 Φ 属于 $\mathbb{R}^D$，并令 $\theta\in\mathbb{R}^D$ 的 *Tukey* 深度为 $h_\Phi(\theta)$。那么，给定 $m = (3D+1)/h_\Phi(\theta)$个独立同分布样本 $\Phi_1,\dots,\Phi_m$，取自 Φ的分布，我们有

$$\mathbb{P}\Big[\theta \in \mathrm{conv}\Big(\{\Phi_1,\dots,\Phi_m\}\Big)\Big] \ge \frac{1}{2}.$$

我们将应用上述定理，其中随机向量 θ 是我们关注的候选矩向量，而 Φ 是来自分布 ν 的随机样本，以证明存在一个有效分布，使其矩与 θ给出的矩相匹配。

我们首先证明以下引理，它将双侧集中界（引理 9）转化为单侧集中界，进而给出目标点 Tukey 深度的下界。

引理 14。 假设 ν 是定义在 $\mathbb{R}^n$ 上的概率测度，并满足以下反集中性质：对于任意次数至多为 k 的多元多项式 $p\colon \mathbb{R}^n \to \mathbb{R}$，

$$\forall \delta > 0,\ \mathop{\mathbb{P}}_{x\sim\nu}\left[\,|p(x) - \mathbb{E}_\nu[p]| \le \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \le \eta_k(\delta).$$

则对任意满足 $\eta_k(\delta_1) \le 1/2$ 的 $\delta_1 > 0$，我们有

$$\mathop{\mathbb{P}}_{x\sim\nu}\left[p(x) \ge \mathbb{E}_\nu[p(x)] + \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \ge \frac{\delta_1^2}{16} - \eta_k(\delta).$$

特别地，存在 $\delta_0 > 0$ 使得

$$\forall \delta < \delta_0,\quad \mathop{\mathbb{P}}_{x\sim\nu}\left[p(x) \ge \mathbb{E}_\nu[p(x)] + \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \ge 3\eta_k(\delta),$$

通过取 $\delta_1 = 8\sqrt{\eta_k(\delta)}$。这里 $\delta_0 > 0$ 是一个选取的常数，使得 $\eta_k(8\sqrt{\eta_k(\delta_0)} < \frac{1}{2}$。-

证明。 考虑随机变量 $Z = p(x) - \mathbb{E}[p(x)]$。我们有 $\mathbb{E}[Z] = 0$。由 p 在测度 ν 下的反集中性质，我们有

$$\mathbb{P}[|Z| \le \delta\sqrt{\mathrm{Var}[Z]}] \le \eta_k(\delta),\ \text{and}\ \mathbb{P}[|Z| \le \delta_1\sqrt{\mathrm{Var}[Z]}] \le \eta_k(\delta_1). \tag{7}$$

此外，由于 $\mathbb{E}[Z] = 0$，由柯西-施瓦茨，

$$\frac{1}{2}\mathbb{E}[|Z|] = \mathbb{E}\Big[Z\ \mathbf{1}[Z \ge 0]\Big] \le \sqrt{\mathrm{Var}[Z]}\sqrt{\mathbb{P}[Z \ge 0]}.$$

Then

$$\mathbb{P}\left[Z \ge \mathbb{E}[Z] + \delta\sqrt{\mathrm{Var}[Z]}\right] \ge \mathbb{P}[Z \ge 0] - \eta_k(\delta) \ge \frac{\mathbb{E}[|Z|]^2}{4\mathrm{Var}[Z]} - \eta_k(\delta). \tag{8}$$

而且

$$\mathbb{E}[|Z|] \ge \mathbb{P}\left[|Z| \ge \delta_1\sqrt{\mathrm{Var}[Z]}\right] \cdot \delta_1\sqrt{\mathrm{Var}[Z]} \ge (1 - \eta_k(\delta_1))\delta_1\sqrt{\mathrm{Var}[Z]}. \tag{9}$$

结合（8）、（9），我们得到

$$\mathbb{P}\left[Z \ge \mathbb{E}[Z] + \delta\sqrt{\mathrm{Var}[Z]}\right] \ge \frac{(1 - \eta_k(\delta_1))^2\delta_1^2}{4} - \eta_k(\delta) \ge \frac{\delta_1^2}{16} - \eta_k(\delta),$$

对于我们所选的参数。 ■

推论 15。 设 $\Phi \in \mathbb{R}^D$ 是一个坐标为 $\Phi_{\ell_1,\ \ell_2} = \lambda^{\ell_1} z^{\ell_2}$ 的随机变量，其中 $(\ell_1, \ell_2)\ \in \mathcal{I} = \{(i_1, i_2) \ne 0 :$ i_1, i_2 为偶数 $0 \le i_2 \le i_1 \le k\}$ 且 $(\lambda, z) \sim \nu$。那么对于任意 $\delta \le 1/k^{Ck^2}$，

$$h_\Phi((1+\delta)\mathbb{E}[\Phi]) \ge 2\eta_k(\sqrt{k}\delta) > 0.$$

证明。 设 $\beta \in \mathbb{R}^D$ 是一个满足 $\|\beta\| = 1$ 的任意系数向量。那么

$$p(\lambda, z) := \langle \beta, \Phi\rangle = \sum_{\mathcal{I}} \beta_{\ell_1,\ell_2}\lambda^{\ell_1} z^{\ell_2}.$$

根据引理 14，我们知道

$$\forall \delta < \delta_0,\quad \mathop{\mathbb{P}}_{x\sim\nu}\left[p(x) \ge \mathbb{E}_\nu[p(x)] + \delta \cdot \sqrt{\mathrm{Var}(p(x))}\right] \ge 3\eta_k(\delta)$$

其中，$\delta_0 > 0$ 的选取满足 $\eta_k(8\sqrt{\eta_k(\delta_0)}) < \frac{1}{2}$。由于 $(\lambda, g) \sim \mathcal{N}(0, I_2)$ 是对数凹的，Carbery-Wright 给出：对于 $q(\lambda, g)$，$\eta_k(\delta) = C_0 k \delta^{1/k}$。因为 $q(\lambda, g) = p(\lambda, \lambda g)$，所以在 $\mathcal{N}(0, I_2)$ 下关于 q 的分布结论可原样移用于 ν 下的 p。因此，对于某个常数 C，我们可以选取 $\delta_0 = 1/k^{Ck^2}$，使其满足引理 14 中的条件。根据引理 9 证明中的论证，对于某个常数 c，我们有 $|\mathbb{E}[p]| \le c\sqrt{k}\mathrm{Var}(p)$。因此

$$\begin{aligned} h_\Phi((1+\delta)\mathbb{E}[\Phi]) &= \inf_{\beta:\|\beta\|_2=1} \mathbb{P}_\nu\left[\langle \beta, \Phi - (1+\delta)\mathbb{E}[\Phi]\rangle \ge 0\right] \\ &= \inf_{\beta:\|\beta\|_2=1} \mathbb{P}_\nu[p \ge (1+\delta)\mathbb{E}[p]] \\ &\ge \inf_{\beta:\|\beta\|_2=1} \mathbb{P}_\nu\left[p \ge \mathbb{E}[p] + c\delta\sqrt{k\mathrm{Var}(p)}\right] \\ &\ge 3\eta_k(c\sqrt{k}\delta) > 0. \end{aligned}$$

■

推论 15 是连接回定理 8 证明的关键环节。它确认，$\Phi(\lambda, z)$ 与 $(\lambda, z) \sim \nu$ 这一特定矩向量由我们的构造得到，并且具有足够的反集中性（通过引理 9 和引理 14），从而保证一个从其均值稍微扰动后得到的点具有非零 Tukey 深度。下面的引理表明，这种非零深度意味着该点位于凸包的内部，为完成定理 8 证明中的矩匹配论证提供了所需的“余地”。

引理 16. 设 Φ 是 $\mathbb{R}^D$ 中的一个随机向量，使得每个仿射超平面的 Φ测度均为零。若 $h_\Phi(\theta) > 0$，则 θ 位于$\mathrm{conv}(\mathrm{supp}(\Phi))$的内部。特别地，存在 $r = r(h_\Phi(\theta)) > 0$，使得 $B(\theta, r) \subseteq \mathrm{conv}(\mathrm{supp}(\Phi))$。

证明。 我们将用反证法证明。假设 θ 位于 conv(supp(Φ)) 的边界上，那么存在一个支撑超平面 u 满足 $\langle u, \phi\rangle \le \langle u, \theta\rangle$ 对支撑集中的所有 ϕ 均成立，由此可推出 $h_\Phi(\theta) = 0$。因此，θ 位于 conv(supp(Φ)) 的内部。由于 conv(supp(Φ)) 的内部是开集，存在 $r > 0$ 使得 $B(\theta, r) \subseteq$ conv(supp(Φ))。

■

4 将困难性扩展至更高次数：通过有界低次优势

在本节中，我们证明定理 3的第一部分。我们将困难性结果扩展到次数为 $k = \mathrm{poly}(n)$ 的多项式，并证明在此情形下低次优势仍然有界。这实际上排除了任何仅依赖低次多项式的高效算法。我们将使用第 3 节中的同一零假设分布 Q，即 Q 是一个高斯尺度混合，其高斯尺度为 $\mu = \mathcal{N}(0, 1)$。定义植入分布 $P = (1-\alpha)Q + \alpha 0$。

定理 17 (LDA 界，最高次数为 $k = n^{\Omega(1)}$)。 对于在 m个样本上、次数至多为 k 的多项式，

$$LDA_{\le k}^{(m)}(P, Q) \le \sqrt{(1 + C^2\alpha^2 k)^m - 1}.$$

特别地，如果 $k = O(1/(\alpha^2 m))$，那么低次优势$_{\le k}^{(m)}(P, Q) = O(1)$。

该定理的证明分为两步。我们首先给出单样本 LDA 界，然后将其提升到 m 样本。以下引理说明，对于任意分布 P 和 Q，低次优势如何从单个样本尺度变化到m独立同分布样本。

引理18. 设 P、Q 为 $\mathbb{R}^n$ 上的分布，其直至次数 $2k$的矩有限。假设单样本低次优势为低次优势$_{\le k}^{(1)}(P,Q) = \delta$。则对任意 m,

$$LDA_{\le k}^{(m)}(P,Q)^2 \le (1+\delta^2)^m - 1.$$

证明。设 $\{\psi_j : j \in \mathcal{J}\}$ 是次数至多为 k 的多项式在 Q 下的一组正交归一基，其中$\psi_0 = 1$。对于多重指标 $\gamma = (\gamma_1, \ldots\ldots, \gamma_m)$，定义

$$\Psi_\gamma(X_1, \ldots, X_m) = \prod_{i=1}^m \psi_{\gamma_i}(X_i).$$

由于 $Q^{\otimes m}$ 是乘积测度， Ψ_γ 关于 $Q^{\otimes m}$ 正交归一。对于任意次数至多为 k 的多项式f ，减去常数 $\mathbb{E}_{Q^{\otimes m}}[f]$ 不会改变低次优势，因此我们可以假设$\mathbb{E}_{Q^{\otimes m}}[f] = 0$。于是 f 可写为

$$f = \sum_{\gamma \ne 0} c_\gamma \Psi_\gamma$$

其中 $\gamma \ne 0$ 表示 γ 的坐标并非全为零。令 $s(\gamma)$ 表示 γ 中非零项的个数。于是 $s(\gamma) \le k$ ，因为 f 的次数至多为 k。由正交归一性,

$$\mathrm{Var}_{Q^{\otimes m}}(f) = \mathbb{E}_{Q^{\otimes m}}[f^2] = \sum_{\gamma \ne 0} c_\gamma^2.$$

令 $\mu_j = \mathbb{E}_P[\psi_j]$。我们有

$$\mathbb{E}_{P^{\otimes m}}[\Psi_\gamma] = \prod_{i=1}^m \mu_{\gamma_i}.$$

由于对于 $j \ne 0$ 有 $\mathbb{E}_Q[\psi_j] = 0$,

$$\mathbb{E}_{P^{\otimes m}}[f] - \mathbb{E}_{Q^{\otimes m}}[f] = \sum_{\gamma \ne 0} c_\gamma \prod_{i=1}^m \mu_{\gamma_i}.$$

根据柯西-施瓦茨不等式,

$$\mathrm{LDA}_{\le k}^{(m)}(P,Q)^2 = \frac{(\mathbb{E}_{P^{\otimes m}}[f] - \mathbb{E}_{Q^{\otimes m}}[f])^2}{\mathrm{Var}_{Q^{\otimes m}}(f)} = \frac{\left(\sum_{\gamma \ne 0} c_\gamma \prod_{i=1}^m \mu_{\gamma_i}\right)^2}{\sum_{\gamma \ne 0} c_\gamma^2} \le \sum_{\gamma \ne 0} \prod_{i=1}^m \mu_{\gamma_i}^2.$$

类似地，当 $m = 1$ 时，我们有

$$\mathrm{LDA}_{\le k}^{(1)}(P,Q)^2 = \frac{\left(\sum_j c_j \mu_j\right)^2}{\sum_j c_j^2} \le \sum_j \mu_j^2.$$

当 c 和 μ 线性相关时，该等式成立。因此，低次优势$_{\le k}^{(1)}(P,\ Q)^2 = \sum_j \mu_j^2 = \delta^2$。固定一个大小为 s 的子集 $S \subseteq [m]$ 。按非零坐标集合 S 对 γ 进行分组,

$$\sum_{\substack{\gamma : \{i : \gamma_i \ne 0\} = S, \\ \sum_{i \in S} \deg(\psi_{\gamma_i}) \le k}} \prod_{i \in S} \mu_{\gamma_i}^2 \le \sum_{j \in \mathcal{J}} \prod_{i \in S} \mu_j^2 = \prod_{i \in S} \left(\sum_{j \in \mathcal{J}} \mu_j^2 \right) = \left(\sum_{j \in \mathcal{J}} \mu_j^2 \right)^s = \delta^{2s}.$$

对所有大小为 $s \le k$的可能 S 求和,

$$\mathrm{LDA}_{\le k}^{(m)}(P,Q)^2 \le \sum_{s=1}^{k} \binom{m}{s} \delta^{2s} \le (1+\delta^2)^m - 1.$$

■

我们现在证明定理 17。

证明定理 *17*。 我们首先给出单样本低次优势的上界。设 f 为任意次数至多为 k 、定义在 $\mathbb{R}^n$ 上的多项式。由于向 f 加上一个常数不会改变低次优势，我们可以假设 f 的常数项为零。于是

$$\mathbb{E}_P[f] - \mathbb{E}_Q[f] = -\alpha \mathbb{E}_Q[f].$$

设 $X \sim Q$。我们可以将 $X = \lambda g$写成相互独立的随机变量 $g \sim \mathcal{N}(0,\ I_n)$ 与 $\lambda \sim \mathcal{N}(0,\ 1)$。定义$h(\lambda) = \mathbb{E}_g[f(\lambda g)]$。于是 $\mathbb{E}_\lambda[h(\lambda)] = \mathbb{E}_Q[f(X)]$ ，并由全方差公式,

$$\mathrm{Var}_Q(f(X)) \ge \mathrm{Var}_\lambda(h(\lambda)).$$

此外， $h(0) = 0$ 即 $h(\lambda)$ 没有常数项。因此，我们已将其归结为关于一维尺度分布 λ 的一个类似断言。现在我们将给出 $|\mathbb{E}_\lambda[h(\lambda)]|/\sqrt{\mathrm{Var}_\lambda[h(\lambda)]}$ 的上界。应用引理 11 于 h ，取 $q(\lambda, g) = h(\lambda)$,

$$|\mathbb{E}_Q[f(X)]| = \mathbb{E}_\lambda[h(\lambda)] \le C\sqrt{k\mathrm{Var}_\lambda(h(\lambda))} \le C\sqrt{k\mathrm{Var}_Q(f(X))}.$$

因此,

$$\mathrm{LDA}_{\le k}^{(1)}(P,Q) \le C\alpha\sqrt{k}.$$

应用引理 18 于 $\delta = C\alpha\sqrt{k}$,

$$\mathrm{LDA}_{\le k}^{(m)}(P,Q) \le \sqrt{(1+C^2\alpha^2 k)^m - 1}.$$

■

在上述证明中，我们关键地用到了如下性质： $X \sim Q$ 可以表示为两个独立随机变量的乘积 $X = \lambda g$ ，其中 $\lambda \in \mathbb{R}$ 表示从 μ中抽取的尺度分布，而 $g \in \mathbb{R}^n$ 抽取自任意旋转不变分布。将 g和 λ 具体取为高斯分布只是为了方便。我们现在精确说明尺度分布 μ 的何种结构性质决定了 LDA 界，以及该论证如何推广。

证明的核心步骤将 n维单样本优势归结为对满足 $h(0) = 0$的一元k 次多项式 $h(\lambda)$的比值 $|E_\mu[h(\lambda)]|/\sqrt{\mathrm{Var}_\mu(h(\lambda))}$的一个界。比值 $|E_\mu[h]|/\sqrt{\mathrm{Var}_\mu(h)}$可通过测度 μ的正交归一多项式 $\{p_j\}_{j\ge 0}$ 得到清晰刻画。将 $h(\lambda) = \sum_{j=0}^{k} c_j p_j(\lambda)$展开，正交归一性给出 $E_\mu[h] = c_0$ 以及$\mathrm{Var}_\mu(h) = \sum_{j=1}^{k} c_j^2$。约束 $h(0) = 0$ 给出

$$c_0 = -\sum_{j=1}^{k} c_j\, p_j(0).$$

由柯西–施瓦茨不等式，

$$c_0^2 \;\le\; \Big(\sum_{j=1}^{k} c_j^2\Big)\Big(\sum_{j=1}^{k} p_j(0)^2\Big) \;=\; \mathrm{Var}_\mu(h)\cdot\sum_{j=1}^{k} p_j(0)^2,$$

从而

$$\frac{|E_\mu[h(\lambda)]|}{\sqrt{\mathrm{Var}_\mu(h(\lambda))}} \;\le\; \sqrt{\sum_{j=1}^{k} p_j(0)^2}, \tag{10}$$

当 $c_j \propto p_j(0)$ 对于 $j \ge 1$ 时取等号。该量

$$\mathcal{K}_k(\mu,0) \;:=\; \sum_{j=0}^{k} p_j(0)^2$$

是 μ 在 0 处的克里斯托费尔和（克里斯托费尔函数 $\lambda_k(\mu,0)^{-1}$ 的倒数）。由于 $p_0 \equiv 1$，界（10）变为 $|E_\mu[h]|/\sqrt{\text{方差}_\mu(h)} \le \sqrt{\mathcal{K}_k(\mu,0)} - 1$。因此，$\mathcal{K}_k(\mu,0)$ 随 k 的增长率，是控制 LDA 界的基本量。特别地，可以选择任意一种尺度分布，使得该量随 k 增长得足够缓慢，从而对单样本低次优势得到同类界，并因此得到相同的提升论证，推广到 m个样本。

5 个简单的抗噪算法

本节将展示，存在一些不基于低次多项式方法、却能高效解决该问题的简单高效算法。考虑以下情形：$\alpha = 1/\mathrm{poly}(n)$ 比例的点位于一个 $d \le 1$ 维子空间上。一方面，即使有 α 比例的点位于该子空间上，低次方法仍会失败。另一方面，我们的简单算法会抽取点，并检查是否存在 $(d+1)$ 个点（当 $d = 0$ 时为一个点）接近某个 d维子空间（当 $d = 0$ 时接近程度为 0）。此外，该算法还能容忍最坏情况误差，并支持重新随机化。

Hardt 与 Moitra [HM13]提出了一种优雅而巧妙的鲁棒子空间恢复算法，它利用了不在子空间上的 $(1-\alpha)$ 个点的反集中性质 S：当 $\alpha \ge \dim(s)\, n$ 时，该算法给出高效算法。随后，Bhaskara、Chen、Perreault 与 Vijayaraghavan [Bha+19]在平滑分析设定下给出了改进算法，其中不在子空间上的点被随机扰动（由平均长度为 $\rho = 1/\mathrm{poly}(n)$ 的随机向量扰动），并在 $\alpha \gtrsim (d/n)^\ell$ 时对任意常数 $\ell > 0$ 给出了 $(nd)^{O(\ell)}$ 时间内的恢复保证。在我们的设定中，$d = O(1)$ 且 $\alpha = 1/n^{O(1)}$，我们可以选取适当的 $\ell = O(1)$ 以获得算法保证。然而，该算法仅具有中等程度的噪声容忍：每个点最多可被对抗性扰动（移动）$1/n^{O(\ell)}$。另见 [BK21；Gao+26]，其中给出了其他参数区间的算法结果。我们将转而给出一种专门针对 $d = O(1)$ 设定的简单算法，它是初等的，并且还允许点被移动 $\epsilon = \Omega(1)$ 的量。

5.1 具有 ϵ 相对误差扰动的鲁棒子空间恢复。

我们在下面定义该问题的假设检验版本。给定来自分布 $\mathcal{D}$（定义在 $\mathbb{R}^n$ 上）的样本，目标是区分以下两个假设：

- **NULL：** $\mathcal{D} = Q_{\mathrm{rot}}$，其中$Q_{\mathrm{rot}}$ 是第3节中定义的旋转不变分布。

• **PLANTED：** $\mathcal{D}$ 是任意分布，其至少有 α 的概率质量支撑在植入子空间 S 上，维度为 d。

请注意， PLANTED 上述情形比第3节下界示例中的植入分布更一般。此外，即使从$\mathcal{D}$ 中抽取的样本随后按以下方式受到对抗性扰动，我们的算法仍然适用。

定义 19 （对抗性（ϵ, p）噪声样本，来自 $\mathcal{D}$）。抽取 m 个样本 x_1、…、 $x_m \in \mathbb{R}^n$，均从 $\mathcal{D}$ 中独立同分布地抽取。然后，对手可以对 $p \in [0,1)$ 执行以下操作：

1. 重新随机抽取，使其来自 Q_{rot}： 对于每个 $i \in [m]$，以概率 p 重新抽取 $x_i \sim Q_{\text{rot}}$，以概率 $(1-p)$ 保持不变。*2.* 对每个点进行对抗性 ϵ扰动：对于每个 $i \in [m]$ ，样本 x_i 受到对抗性扰动，形成 $\tilde{x}_i = x_i + z_i$ ，其中 $\|z_i\|_2 \le \epsilon \|x_i\|_2$。

输入为受损样本 $\tilde{x}_1, \ldots, \tilde{x}_m$。

注意，对手会观察到 m 个样本 $x_1, \ldots, x_m$，而对样本的破坏可以任意相关。我们指出，低次多项式方法在这个模型中仍然会失效。我们分析以下简单算法。

抗噪算法

输入： 给定 m 个样本 $\tilde{x}_1, \ldots, \tilde{x}_m$、维度 d、参数 α的下界 $\alpha_0 \in (0,1)$、参数 p的上界 $p_0 \in [0,1)$**输出：** NULL，或 PLANTED（以及一个维度为 d的子空间 S）。

1. 对所有 $i \in [m]$，将向量 $a_i = \tilde{x}_i / \|\tilde{x}_i\|$ 归一化。
2. 遍历所有大小为 $|T| = d+1$的 $\binom{m}{d+1}$ 种选择 $T \subset [m]$

3. (i) 令 A_T 为由相应的点 $A_T = (a_i : i \in T) \in \mathbb{R}^{n \times (d+1)}$ 构成的矩阵。(ii) 如果 $\sigma_{d+1}(A_T) \le \frac{1}{2}$，则输出 PLANTED 并返回。% （如果需要子空间，则输出 A_T 的前 d 个左奇异空间。）

4. 输出 NULL。

图 2：抗噪算法，用于检测是否有 α 比例的点接近某个 $d = O(1)$ 维子空间

定理 20 （具有相对误差扰动的抗噪算法）。 存在绝对常数 c、$C > 0$，使得当 $m \ge C(d+1)/(\alpha(1-p))$ 且任意 $0 \le \epsilon \le c(1-p)\sqrt{\alpha/(d+1)}$ 时，以下结论成立

存在一个算法（见图 2），其运行时间为 $O(\binom{m}{d+1})$；该算法以含噪样本 $\tilde{x}_1$、…、 $\tilde{x}_m$ 作为输入，按照上述方式设定 $\|\tilde{x}_i - x_i\| \le \epsilon \|x_i\|$，即可高概率地解决该检验问题。

值得注意的是，即使扰动处于最坏情形，当维度 $d = O(1)$ 时，该算法也允许 ϵ 达到常数级。

The 关于 NULL 设定（可靠性）的论证由以下引理得出 .

引理 21（NULL 设定）。 在定理 *20*的设定下，存在一个常数 $c' > 0$ 使得以至少 $1 - m^2\exp(-c'n)$ 的概率

$$\forall T \subset [m] \ \textit{with}\ |T| = d+1, \quad \sigma_{d+1}(A_T) > \frac{1}{2}.$$

证明。 该证明由矩阵 A 的非相干性 得出。

命题 22。 存在绝对常数 $c, c' > 0$，使得向量集合 $x_1, \ldots, x_m \sim Q_{\mathrm{rot}}$，即（受损的）归一化向量 $\{a_i = (x_i+z_i)/\|x_i+z_i\| : i \in [m]\}$ 以至少$1 - O(m^2\exp(-c'n))$ 的概率满足

$$\forall i \neq j \in [m], \ |\langle a_i, a_j\rangle| \leq \frac{1}{(1-\epsilon)^2}\Big(\frac{c\sqrt{\log m}}{\sqrt{n}} + 2\epsilon + \epsilon^2\Big). \tag{11}$$

命题证明。 根据随机单位向量的标准集中性质，对于某个通用常数 $c > 0$，我们有：

$$\begin{aligned} &\mathop{\mathbb{P}}_{\hat{x},\hat{y}\sim_{unif}\mathbb{S}^{n-1}}\Big[|\langle \hat{x}, \hat{y}\rangle| \geq \frac{c\sqrt{\log m}}{\sqrt{n}}\Big] \leq \frac{1}{m^3}, \quad \text{and hence by a union bound,} \\ \forall i \neq j \in [m], \ &\mathop{\mathbb{P}}_{x_i,x_j\sim Q_{\mathrm{rot}}}\Big[\Big|\Big\langle \frac{x_i}{\|x_i\|}, \frac{x_j}{\|x_j\|}\Big\rangle\Big| \leq \frac{c\sqrt{\log m}}{\sqrt{n}}\Big] \geq 1 - \frac{1}{m}. \end{aligned} \tag{12}$$

我们以（12）中的事件为条件。另请注意 $\|x_i + z_i\| \geq (1-\epsilon)\|x_i\|$ 因此，对于所有$i \neq j \in [m]$

$$\begin{aligned} |\langle a_i, a_j\rangle| = \Big|\Big\langle \frac{x_i+z_i}{\|x_i+z_i\|}, \frac{x_j+z_j}{\|x_j+z_j\|}\Big\rangle\Big| &\leq \frac{|\langle x_i, x_j\rangle|}{(1-\epsilon)^2\|x_i\|\|x_j\|} + \frac{|\langle z_i, x_j\rangle|+|\langle x_i, z_j\rangle|+|\langle z_i, z_j\rangle|}{(1-\epsilon)^2\|x_i\|\|x_j\|} \\ &\leq \frac{c\sqrt{\log m}}{(1-\epsilon)^2\sqrt{n}} + \frac{\|z_i\|\|x_j\| + \|x_i\|\|z_j\| + \|z_i\|\|z_j\|}{\|x_i\|\|x_j\|} \leq \frac{c\sqrt{\log m}}{(1-\epsilon)^2\sqrt{n}} + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}, \end{aligned}$$

其中最后一行用到了（12）和 $\|z_i\| \leq \epsilon\|x_i\|$。

现在，为完成证明的其余部分，考虑任意一个满足 $|T| = d+1$ 的 $T \subset [m]$，以及矩阵 A_T 的测试单位向量 $\beta \in \mathbb{R}^{d+1}$。我们有

$$\begin{aligned} \|A_T\beta\|_2^2 = \Big\|\sum_{i\in T}\beta_i a_i\Big\|_2^2 &\geq \sum_{i\in T}\beta_i^2\|a_i\|_2^2 - \sum_{i\neq j}|\beta_i||\beta_j||\langle a_i, a_j\rangle| \\ &\geq \sum_{i\in T}\beta_i^2 - \sum_{i\neq j\in T}|\beta_i||\beta_j|\Big(\frac{c\sqrt{\log m}}{\sqrt{n}(1-\epsilon)^2} + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\Big) \\ &\geq 1 - \|\beta\|_1^2\Big(\frac{c\sqrt{\log m}}{\sqrt{n}(1-\epsilon)^2} + \frac{2\epsilon+\epsilon^2}{(1-\epsilon)^2}\Big) \geq 1 - \frac{(d+1)}{(1-\epsilon)^2}\Big(\frac{c\sqrt{\log m}}{\sqrt{n}} + 2\epsilon + \epsilon^2\Big) \\ &> \frac{1}{2} \qquad \text{since } \epsilon < \frac{1}{8(d+1)}, \text{ and } m < 2^{o(n)}. \end{aligned}$$

■

证明。 为了证明该算法有效，我们将证明：(a) 在 PLANTED 情形下，从子空间中 α 部分采样得到的点中，$d+1$ 之间存在近似线性相关；(b) 在 NULL 情形下，任意一组 $d+1$ 个点之间都不存在近似线性相关。

PLANTED 情形（完备性）： 令 x_i 为对应于 $\tilde{x}_i$ 的 $\mathcal{D}$ 中的未受污染样本，并令 $\tilde{x}_i = x_i + z_i$ 满足 $\|z_i\| \le \epsilon \|x_i\|$。给定 $m \ge C(d+1)/(\alpha(1-p))$，期望至少有 Cd 个来自 $\mathcal{D}$ 的点位于子空间 S 中；因此，以高概率存在 $T^\star \subset [m]$，满足 $|T| = d+1$，且对应于以下点 $\{x_i : i \in T\}$，这些点属于维度为 d 的子空间 S。回忆一下，$a_i = \tilde{x}_i / \|\tilde{x}_i\|$。

子矩阵 $M_{T^\star} = (\frac{x_i}{\|x_i\|} : i \in T^\star)$ 具有 $\sigma_{d+1}(M_{T^\star}) = 0$。令 $\beta \in \mathbb{R}^{d+1}$ 为满足 $M_{T^\star}\beta = 0$ 的单位向量。则

$$
\begin{aligned}
A_{T^\star}\beta &= \sum_{i \in T^\star} \beta_i a_i = \sum_{i \in T^\star} \beta_i \frac{x_i + z_i}{\|x_i + z_i\|} = \sum_{i \in T^\star} \beta_i \frac{x_i}{\|x_i + z_i\|} + \beta_i \frac{z_i}{\|x_i + z_i\|} \\
&= \sum_{i \in T^\star} \beta_i \Big(\frac{x_i}{\|x_i + z_i\|} - \frac{x_i}{\|x_i\|} \Big) + \sum_{i \in T^\star} \beta_i \frac{z_i}{\|x_i + z_i\|} && (\text{since } M_{T^\star}\beta = 0) \\
\|A_{T^\star}\beta\|_2 &\le \sum_{i \in T^\star} |\beta_i| \Big(\frac{\|x_i\|}{\|x_i + z_i\|} - 1 \Big) + \epsilon \sum_{i \in T^\star} |\beta_i| \frac{\|x_i\|}{\|x_i + z_i\|} && (\text{since } \|z_i\| \le \epsilon \|x_i\|) \\
&\le \Big(\sum_{i \in T^\star} |\beta_i| \Big) \Big(\frac{2\epsilon}{1-\epsilon} \Big) \le \frac{2\epsilon\sqrt{d+1}}{1-\epsilon} < \frac{1}{2} && (\text{由于 } \epsilon < \frac{1}{8\sqrt{(d+1)}}).
\end{aligned}
$$

NULL 情形（可靠性）： 这由引理 21 得出。证明完毕。

■

5.2 具有加性扰动的模型

在本节中，我们给出了一种算法变体，该算法同样适用于幅度有界的最坏情况加性扰动。

我们考虑与之前相同的假设检验问题：给定来自定义在 $\mathbb{R}^n$ 上的分布 $\mathcal{D}$ 的样本。目标是区分以下两个假设：

- **NULL：** $\mathcal{D} = Q_{\text{rot}}$ 其中 Q_{rot} 是第 3 节中定义的旋转不变分布。
- **PLANTED：** $\mathcal{D}$ 是任意 分布，其至少有 α 的概率质量支撑在维度为 d 的植入子空间 S 上。

我们现在考虑如下设定：从 $\mathcal{D}$ 抽取的样本随后按如下方式受到对抗性扰动。

定义 23 （对抗性加性（η, p）-噪声样本，来自 $\mathcal{D}$）。 抽取 m 个样本 $x_1, \dots, x_m \in \mathbb{R}^n$，它们独立同分布地取自 $\mathcal{D}$。然后对手 可以对 $\eta, p \in [0,1)$ 执行以下操作：

1. 重新随机化以从 Q_{rot}：对于每个 $i \in [m]$，以概率 p 重新抽取 $x_i \sim Q_{\text{rot}}$，并以概率 $(1-p)$ 保持不变。

2. 对抗性 η-扰动到每个点：对于每个 $i \in [m]$，样本 x_i 经过对抗性扰动后形成 $\tilde{x}_i = x_i + z_i$，其中 $\|z_i\|_2 \le \eta$。

输入为受损样本 $\tilde{x}_1, \dots, \tilde{x}_m$。

我们分析以下简单算法。

加性扰动下的抗噪算法

输入： 给定 m 个样本 $\tilde{x}_1, \ldots\ldots, \tilde{x}_m$、维度 d、参数 $\alpha \in (0, 1)$、参数 $p \in (0, 1)$ 以及阈值 $\tau > 0$

输出： NULL，或 PLANTED（以及一个维度为 d 的子空间 S）。

1. 对 $J \subseteq [m]$ 进行大小为 $|J| = 2\log(4/\delta)(d+1)/(\alpha(1-p))$ 的子采样。
2. 遍历所有大小为 $|T| = d+1$ 的 $\binom{m}{d+1}$ 选取方案 $T \subset J$

3. (i) 令 A_T 为由对应的点 $A_T = (\tilde{x}_i : i \in T) \in \mathbb{R}^{n \times (d+1)}$ 构成的矩阵。(ii) 如果 $\sigma_{d+1}(A_T) \le \tau$，则输出 PLANTED。%（如果需要该子空间，则输出 A_T 的顶部-d 左奇异空间。）

4. 输出 NULL

图 3：即使存在加性扰动，用于检测是否有 α 比例的点靠近一个 $d = O$ 维子空间的算法

定理 *24*（具有加性误差扰动的抗噪算法）。存在绝对常数 c、$c' > 0$、$C \ge 1$，使得当 $m \ge 2(d+1)\log(4/\delta)/(\alpha(1-p))$ 且任意 $\delta \in (0, 1)$ $0 \le \eta \le \alpha(1-p)\sqrt{n})()\log(4/\delta)^2$ 时，以下结论成立。此时存在一个算法（见图 3，具有 $\tau = \alpha(1-p)\sqrt{n})\log(C(d+1)/\delta))$，其运行时间为 $O(\binom{m}{d+1})$；该算法以 $\tilde{x}_1, \ldots\ldots, \tilde{x}_m$ 作为输入，这些数据按上述方式生成且满足 $\|\widetilde{x}_i - x_i\|_2 \le \eta$，并能以至少 $1-\delta$ 的概率解决该检验问题。

在此模型中，从 Q_{rot} 中抽取的点的平均长度为 $\mathbb{E}[\|X\|_2^2] = \Theta(n)$。此外，从 Q_{rot} 中抽取的点里，有α 的比例长度为 $\|X\|_2 \le c_\star \alpha\sqrt{n}$，其中$c_\star > 0$是某个绝对常数$\eta \ge c_\star\alpha\sqrt{n}$– 因此，当 $\eta \ge c_\star\alpha\sqrt{n}$时，假设检验问题在统计上变得不可能。定理 24 表明，该算法处理的 η 与 $d = O(1)$ 相差一个常数因子，且这是 $c_\star\alpha\sqrt{n}$所能达到的最佳界。

下面的命题表明，对于从 Q_{rot} 中抽取的大多数样本点，其长度至少为2τ。

命题 25。 存在一个通用常数 $c > 0$，使得对于任意 $\gamma \in (0, 1/2)$，我们有

$$\mathop{\mathbb{P}}_{X \sim Q_{\mathrm{rot}}}\left[\|X\|_2 \le \gamma\sqrt{n}\right] \le \min\{c\gamma, 1\}. \tag{13}$$

特别地，令 $\gamma = \alpha(1-p)/(4c\log^2(4/\delta)(d+1))$，可得从 Q_{rot} 中抽取的点里，长度 $\|X\|_2 \le \alpha(1-p)\sqrt{n})\log(4/\delta)^2$ 的比例至多为 $\frac{\alpha(1-p)4\log(4/\delta)^2}{2c}$。

证明。这直接由标准高斯分布的小球概率界（即 $\mathbb{P}_{g \sim N(0,1)}[|g| \le \delta/2] \le c\delta/2$）以及标准球面高斯分布的长度集中性得出。 ■

现在我们继续证明定理24。

定理的证明 *24*。为了证明该算法有效，我们将证明：（a）从 $d+1$ 个点中采样得到的许多点组之间存在近似线性相关，这些点来自 α 部分所在的

子空间（ PLANTED 情形）； （b）在 NULL 情形下，样本中少得多的 $(d+1)$ 元组集合之间存在近似线性相关。

植入情形（完备性）： 令 x_i 为来自 $\mathcal{D}$ 、对应于 $\tilde{x}_i$的未受污染样本，并令 $\tilde{x}_i = x_i + z_i$ 满足 $\|z_i\| \le \eta$。给定 $m \ge 2\log(4/\delta)(d+1)/(\alpha(1-p))$，期望有$(1-p)\alpha|J| \ge 2(d+1)\log(4/\delta)$个来自 $\mathcal{D}$ 的点位于子空间 S中。因此，以概率 $1-\delta/4$，至少有 $d+1$ 个样本 $J_S \subseteq J$ 位于子空间 S中。对于任意满足 $|T| = d+1$的元组$T \subseteq J_S$，我们有 $\sigma_{d+1}((x_i : i \in T)) = 0$。由韦尔不等式可得$\sigma_{d+1}(A_T) \le \eta\sqrt{d+1} \le \tau/2$，即得所需结论。

NULL 情形（可靠性）： 可靠性分析与定理20的可靠性分析非常相似。根据引理25可知，对于至少占 $1-\alpha/4$ 比例的样本 $\{x_i : i \in J\}$，这些样本取自 Q_{rot}，并且有 $\|x_i\|_2 \ge 2\tau$。对于这些至少占（$1-\alpha/4$比例的样本，可以将加性扰动视为乘性扰动，并应用引理22得到一个界。

形式化地，在这 m 个样本中，令 $J_{\text{big}} = \{i \in J : \|x_i\|_2 \ge 2\tau\}$。考虑一个修改后的点集 x'_1, ..., x'_m，其中对于每个 $i \in J$，当 $i \in J_{\text{big}}$ 时有 $x'_i = \tilde{x}_i$，当 $i \notin J_{\text{big}}$ 时有 $x'_i = x_i$。首先我们注意到，这些点 $\{x'_i : i \in J\}$ 满足引理 21 的条件，因为扰动满足 $\|x'_i - x_i\|_2 < \frac{1}{8}(d+1)\|x_i\|_2$。令 A' 为一个矩阵，其 i 列为 A'，且等于 $x'_i/\|x'_i\|$。因此，我们现在可以应用引理 21，得出对于某个绝对常数 $c' > 0$，以至少 $1 - O(m^2\exp(-c'n))$ 的概率，有

$$\forall T \subseteq [J] \text{ s.t. } |T| = d+1,\ \sigma_{d+1}(A'_T) > 1/2.$$

此外，由断言 25 可知

$$\mathbb{E}[|J \setminus J_{\text{big}}|] \le \frac{\alpha(1-p)}{4(d+1)\log(4/\delta)^2} \cdot \frac{2\log(4/\delta)(d+1)}{\alpha(1-p)} \le \frac{1}{2\log(4/\delta)}.$$

因此，以至少 $1-\delta/2$ 的概率，我们有 $J_{\text{big}} = J$。换言之，以至少 $1-\delta/2$ 的概率，$x'_i = \tilde{x}_i\ \forall i \in J$。因此，我们以至少 $1-\delta/2-m^2\exp(-c'n)$ 的概率有，

$$\forall T \subseteq J \text{ s.t. } |T| = d+1,\quad \sigma_{d+1}(A_T) \ge \min_{i \in T}\|\tilde{x}_i\|_2 \cdot \sigma_{d+1}(A'_T) \ge \frac{1}{2} \cdot 2\tau \ge \tau.$$

至此，证明完毕。 ■

致谢

我们感谢 Alex Wein、Miklos Racz 和 Sidhanth Mohanty 对早期草稿的反馈，并感谢 Sanchit Kalhan 就矩匹配界进行的有益讨论。作者得到了由美国国家科学基金会资助的数据、计量经济学、算法与学习研究所（IDEAL）通过资助项目 NSF ECCS-2216970 的支持。

参考文献

[AV23] Gabriel Arpino 和 Ramji Venkataramanan。《混合稀疏线性回归中的统计—计算权衡》。载于：第三十六届学习理论会议论文集。由 Gergely Neu 和 Lorenzo Rosasco 编。第 195 卷。机器学习研究论文集。PMLR，2023 年 12 月，第 921–986 页。 URL: https://proceedings.mlr.press/v195/arpino23a.html。

[Bak+24] Ainesh Bakshi 等。《高斯分布之外反集中性的高效证书》。载于：第 65 届 IEEE计算机科学基础会议（即将发表）。2024 年。URL：https://arxiv.org/abs/2405.15084。[Bar+19] Boaz Barak 等。《植入团问题近乎紧致的平方和下界》。载于：SIAM计算期刊 48.2（2019），第 687–735 页。[BB19] Matthew Brennan 和 Guy Bresler。《稀疏主成分分析的最优平均情况归约：从弱假设到强困难性》。载于：学习理论会议。PMLR。2019 年，第 469–470 页。[BBH18] Matthew Brennan、Guy Bresler 和 Wasim Huleihel。《植入稀疏结构问题的可归约性与计算下界》。载于：学习理论会议。PMLR。2018 年，第 48–166 页。[Bha+19] Aditya Bhaskara 等。《通过解耦进行无监督学习中的平滑分析》。载于：第 60 届 IEEE计算机科学基础年会（FOCS）论文集。IEEE。2019 年。[BK21] Ainesh Bakshi 和 Pravesh K Kothari。《列表可解码子空间恢复：多项式时间内与维度无关的误差》。载于：2021 年 ACM-SIAM离散算法研讨会（SODA）论文集。SIAM。2021 年，第 1279–1297 页。[BKS23] Rares-Darius Buhai、Pravesh K. Kothari 和 David Steurer。《逼近半随机植入团阈值的算法》。载于：第 55 届 ACM理论计算年会论文集。STOC 2023。美国佛罗里达州奥兰多：美国计算机协会，2023 年，第 1918–1926 页。ISBN：9781450399135。DOI：10.1145/ 3564246.3585184。URL：https://doi.org/10.1145/3564246.3585184。[Bla+06] Gilles Blanchard 等。《探寻高维分布的非高斯分量》。载于：机器学习研究期刊 7.9（2006），第 247–282 页。URL：http://jmlr.org/papers/v7/blanchard06a.html。[BR13] Quentin Berthet 和 Philippe Rigollet。《稀疏主成分分析的计算下界》。载于：arXiv 预印本 arXiv:1304.0828（2013）。[Bre+21] Matthew S Brennan 等。《统计查询算法与低次检验几乎等价》。载于：第三十四届学习理论会议论文集。由 Mikhail Belkin 和 Samory Kpotufe 编。第 134 卷。机器学习研究论文集。PMLR，2021 年 8 月 15–19 日，第 774–774 页。URL：https://proceedings.mlr.press/v134/brennan21a.html。[Bru+21] Joan Bruna 等。《连续LWE》。载于：第 53 届 ACM SIGACT理论计算年会论文集。2021 年，第 694–707 页。[Buh+25] Rares-Darius Buhai 等。《拟多项式低次猜想是错误的》。载于：arXiv 预印本 arXiv:2505.17360（2025）。[CW01] A Carbery 和 J Wright。《凸体上多项式的分布不等式与 L-q范数不等式》。英文。载于：Mathematical Research Letters 8.3（2001 年 5 月），第 233–248 页。ISSN：1073-2780。[Dia+23] Ilias Diakonikolas 等。《假设较弱时非高斯成分分析的SQ下界》。载于：神经信息处理系统进展 36（2023），第 4199–4212 页。

[Din+21] Yunzi Ding 等人。《认证限制等距性质的平均情形时间复杂度》。载于：IEEE 信息论汇刊 67.11（2021年11月），第7355–7361页。 ISSN：0018-9448。 DOI：10.1109/TIT.2021.3112823。 URL：https://doi.org/10.1109/TIT. 2021.3112823。[DK22] Ilias Diakonikolas 和 Daniel Kane。《通过格基约化进行非高斯成分分析》。载于：学习理论会议。PMLR。2022年，第4535–4547页。[DKS17] Ilias Diakonikolas、Daniel M Kane 和 Alistair Stewart。《高维高斯分布及高斯混合分布稳健估计的统计查询下界》。载于：2017 IEEE 第58届年度计算机科学基础研讨会（FOCS）。IEEE。2017年，第73–84页。[Fel+17] Vitaly Feldman 等人。《统计算法与检测植入团的一个下界》。载于：ACM 会刊（JACM）64.2（2017年），第1–37页。[Gam21] David Gamarnik。《重叠间隙性质：优化随机结构的拓扑障碍》。载于：美国国家科学院院刊 118.41（2021年），e2108492118。 DOI：10.1073/pnas.2108492118。电子版：https://www.pnas.org/doi/ pdf/10.1073/pnas.2108492118。 URL：https://www.pnas.org/doi/abs/10.1073/ pnas.2108492118。[Gao+26] Chao Gao 等人。《学习置信椭球及其在鲁棒子空间恢复中的应用》。2026年。arXiv：2512。16875 [cs.DS]。 URL：https://arxiv.org/abs/2512. 16875。[HLO23] S Hayakawa、T Lyons 和 H Oberhauser。《估计给定向量属于随机样本凸包的概率》。载于：概率论及相关领域 185.3-4（2023年），第705–746页。[HM13] Moritz Hardt 和 Ankur Moitra。《鲁棒子空间恢复的算法与计算困难性》。载于：第26届年度学习理论会议论文集。Shai Shalev-Shwartz 和 Ingo Steinwart 编。第30卷。机器学习研究论文集。美国新泽西州普林斯顿：PMLR，2013年12月，第354–375页。 URL：https://proceedings.mlr. press/v30/Hardt13.html。[Hop+17] Samuel B Hopkins 等人。《平方和方法在检测隐藏结构中的威力》。载于：2017 IEEE 第58届年度计算机科学基础研讨会（FOCS）。IEEE。2017年，第720–731页。[Hop18] Samuel B. Hopkins。《统计推断与平方和方法》。博士论文。康奈尔大学，2018年。[HS17] Samuel B Hopkins 和 David Steurer。《基于少量样本的贝叶斯估计：社群检测及相关问题》。载于：arXiv 预印本 arXiv:1710.00264（2017年）。[Hsi+26] Jun-Ting Hsieh 等人。《低次数启发式的严格推论》。2026年。arXiv：2601。05850 [cs.CC]。 URL：https://arxiv.org/abs/2601.05850。[HW20] Justin Holmgren 和 Alexander S Wein。《低次猜想的反例》。载于：arXiv 预印本 arXiv:2004.08454（2020年）。[Kea98] Michael Kearns。《基于统计查询的高效抗噪学习》。ACM 会刊 45.6（1998年11月），第983–1006页。 ISSN：0004-5411。 DOI：10.1145/293347.293351。 URL：https://doi.org/10.1145/293347.293351。

[Kun25] Dmitriy Kunisky。《低坐标度算法 I：假设检验计算阈值的普适性》。载于：《统计学年刊》53（2025年4月）。 DOI：10.1214/24-AOS2484。[KWB19] Dmitriy Kunisky、Alexander S Wein 和 Afonso S Bandeira。《关于假设检验计算困难性的札记：基于低次似然比的预测》。载于：ISAAC大会（国际分析、应用与计算学会）。施普林格。2019年，第1–50页。[LM18] Gilad Lerman 和 Tyler Maunu。《鲁棒子空间恢复综述》。载于：《IEEE会刊》106.8（2018年8月），第1380–1410页。 ISSN：1558-2256。 DOI：10.1109/jproc.2018.2853141。 URL：http://dx.doi.org/10.1109/JPROC.2018. 2853141。[MW23] Ankur Moitra 和 Alexander S. Wein。《计算高效检验的精确错误率》。载于：《统计学年刊》（2023年）。 URL：https://api.semanticscholar. org/CorpusID:264832918。[Reg09] Oded Regev。《关于格、带误差学习、随机线性码与密码学》。载于：ACM 会刊（JACM）56.6（2009年），第1–40页。[Tuk75] John W Tukey。《数学与数据的图示》。载于：国际数学家大会论文集。第2卷。温哥华。1975年，第523–531页。[Wei25] Alexander S Wein。《统计学的计算复杂性：低次数多项式带来的新见解》。载于：arXiv 预印本 arXiv:2506.10748（2025年）。[Zad+22] Ilias Zadik 等人。《基于格的方法在聚类中胜过平方和方法》。载于：学习理论会议。PMLR。2022年，第1247–1248页。