分享 d信息支配多视图植入匹配

Ruiqi Wang
RIW010@UCSD.EDU物理系，加利福尼亚大学圣迭戈分校，拉霍亚，*CA 92093*，美国

JingtongHu JHU130
@UCR.EDU数学系，加利福尼亚大学河滨分校，美国加利福尼亚州河滨市 *92521*

arXiv:2611.32892v1 [cs.LG] 24 Nov 2026

摘要

我们研究如何匹配同一批 K 对象的 n 个经过打乱的视图中的行：每个对象在每个视图中均被观测一次，观测结果为具有已知协方差的高斯特征向量。对于任意固定的 K，如果单个对象的 K 次观测所含的共享信息 J 超过$(1+\epsilon)\log\ n$，则所有视图均可匹配，错误数至多为 $o(n)$；如果低于$(1-\epsilon)\log\ n$，则无法做到这一点。即便任意两个视图单独无法匹配，上述结论仍可能成立。低于阈值时，如果某个同时包含这两个视图的视图集合，其共享信息高于 $\log n$，那么这两个视图仍可匹配；否则，即使估计器读取所有视图，它们之间的重叠率也会消失。所有行的匹配由视图中最弱的二分划分决定：在弱相关情形下，跨越该划分的互信息阈值为 $2\log\ n$。对于经由独立噪声观测到的弱共同信号，两个阈值以 $\sum_\ell \rho_\ell^2/\log\ n$为单位分别为 $4/K$和 $4/(K-1)$，这回答了 Wee、Yang、Fan 和 Mao 提出的问题。超出弱相关情形后，二分划分条件并不充分：稀有亚群体要求每种划分为至少三个区块的情形都满足 R´enyi 条件。我们猜想，这两个条件共同刻画了高斯分布下的精确恢复，并证明了对于具有相等相关性的三个视图以及其他若干情形，该猜想成立；其中部分证明借助了计算机辅助。对于相等相关性，当 $K=3$时，分层筛选以$4/K$ 在 $O(Kn^2d)$次操作内实现；而对于一般分布，递归解码器可实现J。近乎精确阈值可推广至已知的乘积分布，以及在随机视图子集中观测对象的情形。

1. 引言

相同的 n 对象通常会以不同且未知的顺序被列出多次：由多个相机观测到的关键点、多个数据库中属于同一人的记录、在多个模态中测量的单元格。每个列表，即视图，为每个对象保存一个特征向量，并且对象的特征在不同视图之间具有相关性。任务是确定哪些行彼此对应。我们研究在什么情况下可以匹配 K 视图中的几乎所有行或全部行；在这一模型中，对象的 K 特征向量服从联合高斯分布，且具有已知

协方差，对象彼此独立。

对于两个视图，答案由 I 对象的两个观测值之间的互信息决定（Cullina 等人，2018年；Dai 等人，2019年、2020年；Shirani等人，2019年；Kunisky 和 Niles-Weed，2022年；Wang等人，2022年；Dai 等人，2023年；Wee 等人，2026年）。对于弱相关的高斯数据库，Dai 等人（2020年，定理 1）证明，除少数

n^λ行之外，对于某个固定的 $\lambda < 1$，当 $I \geq (1+\epsilon)\log n$时可以完成匹配。在高维各向同性模型中，当高于该级别时，除少数 $o(n)$ 行可以完成匹配，而低于 $(1-\epsilon)\log n$ 时，没有估计器能够匹配正比例的行（Wee 等人，2026年，定理 2.1 和推论 2.2）；Ahmad Yarandi 和 Ganassali（2026年，定理 2 和定理 3）关于属性图的结果表明，当图不相关时，也存在相同的转变。匹配每一行需要 I 高于$2\log n$（Dai 等人，2019年，定理 1 和定理 2）。

视图增多时，联合匹配可能在每一对视图都无法匹配的情况下取得成功。假设每个视图都通过独立噪声观测同一个信号，并用 $b = \sum_\ell \rho_\ell^2/\log n$ 衡量信号，其中 ρ_ℓ 是两个视图在坐标 ℓ上的相关性。当相关性较弱时，两个视图的 $I \approx (b/2)\log n$，因此只有在 $b > 2$时，才能仅凭各自的行进行匹配。我们证明，只要满足 $b > 4/K$，所有 K 个视图就都能匹配（图 1）。对于 $4/K < b < 2$，每一对视图几乎都能在所有行上匹配，尽管任何一对视图单独都无法匹配，而且任何先匹配各对视图再将结果合并的解码器都会失败。Wee 等人（2026）以各向同性噪声引入了这一模型，并证明逐次处理一个锚点行、最大化两两内积之和的算法，在 $b > K/(K-1)$ 时可以成功；他们还提出这一级别是否最优的问题（Wee 等人，2026，注 2.6）。Even 等人（2026）也曾将同一模型作为多特征匹配问题进行研究。在我们的高维情形中，他们给出的界（忽略乘法常数）表明，两个恢复阈值均为 b 量级的 $1/K$；他们还提出了一种多项式时间算法，并给出了低次下界，表明视图数量较大时存在统计-计算差距。对于 K 相关高斯图，Vassaux 和 Massouli´e（2026）找到了阈值：它是两图阈值的 $2/K$ 倍。

对于一般的高斯分布，取代 I 的量是共享信息。对于视图的一个划分 $\mathcal{P}$，设 $D_\mathcal{P}$ 为单个对象的元组分布与其各块边缘分布的乘积分布之间的散度，并令 $J = \min_\mathcal{P} D_\mathcal{P}/(|\mathcal{P}|-1)$；对于两个视图$J = I$。

结果。 全文中， K 固定，阈值均以固定裕度给出。结果分为三组。

高斯特征的阈值（第3–5节）。

1. **近乎精确恢复由共享信息决定**。 若满足 $o(n)$ 错误，且 $J \geq (1+\epsilon)\log n$，则所有相对置换都能恢复；若$J \leq (1-\epsilon)\log n$，则任何估计器都无法做到（定理 1）。

2. **哪些配对能够保留**。 低于阈值时，两个视图 a, b 在以下情况下仍可实现近乎精确的匹配：$J_B > \log n$ 对某个视图集合 $B \ni a, b$。否则，每个估计器 —— 即便能够读取所有视图 —— 在 a 与 b 之间的重叠度都会消失（定理 2）。

3. **弱相关下的精确恢复由最弱二分划分决定**。 设I_* 为视图二分划分两侧之间的最小互信息。在弱相关条件下，当 $I_* \geq (2+\epsilon)\log n$时，最大似然估计器可以恢复每一行；而对于任何高斯分布，当 $I_* \leq (2-\epsilon)\log n$ (定理 3)时，不存在任何估计器能够做到这一点。在共同信号模型中，两个阈值分别为 $b = 4/K$ 和 $b = 4/(K-1)$ (推论 5)。这回答了 Wee 等人的问题；他们的逐行规则在 $K/(K-1)$处是精确的，

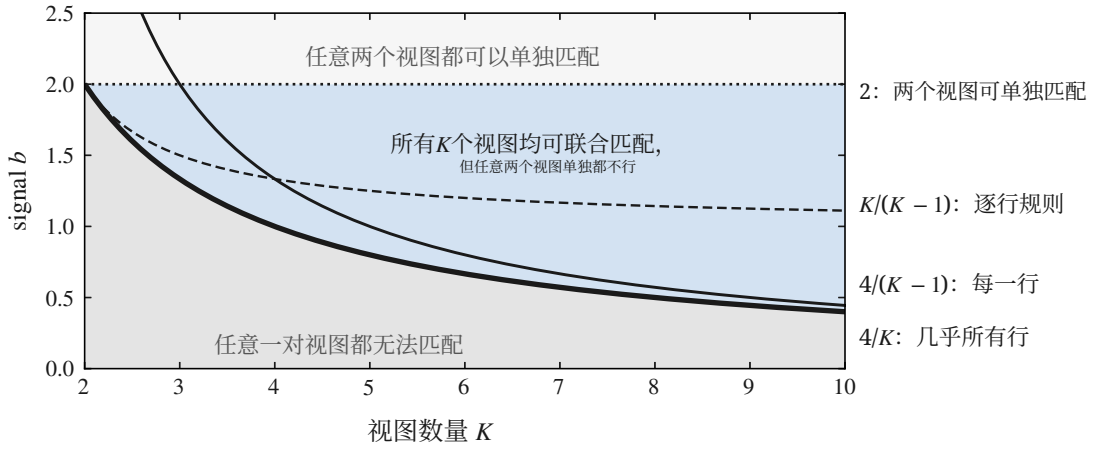

图 1：具有弱相关性的共同信号，$b = \sum_\ell \rho_\ell^2 / \log n$。在 $4/K$ 之上，所有视图中几乎所有行都可以匹配；低于该值时，任意一对视图中都没有正比例的行可以匹配；高于 $4/(K-1)$ 时，每一行都可以匹配，而低于该值时则不行（推论5）。Wee 等人Wee 等人(2026年) 提出的逐行规则需要 $b > K/(K-1)$，而用于匹配成对视图的解码器则需要 $b > 2$ (定理 15)。

因为错误的锚点行可以收集同一对象其他观测值之间的内积（定理15）。

4. 弱相关性之外，稀有亚群体也很重要。对于固定坐标分布，$I_* > 2\log n$ 和 $J > \log n$ 两者合起来仍不足够。含有 n^θ 个非典型对象的亚群体对每个划分为 $q \geq 3$ 个块的情况都施加一个 R´enyi 阶数为 $(q-1)/q$的条件（定理 9）。对于负相关性相等的三个视图，这一条件就是精确阈值；在相关性为 $-1/3$ 时，该阈值为 $d/\log n \approx 10.17$，而另外两个条件从 9.87 起便成立（定理 10）。我们猜想，交换条件和 R´enyi 条件刻画了高斯坐标分布的精确恢复（猜想11），并在其他若干情形中证明了这一点（定理 13）。

解码器及其局限 （第 6节）。分区检验评估 $O_K(n^K)$ 候选元组，因此对于固定的 K，近乎精确阈值可在多项式时间内达到；问题在于指数还能降低多少。

5. **更快的解码器**。 在相关性相等时，分层筛选对于 $4/K$、 $K = 3$、4 和 $b > 1$，对于 $K \geq 5$，可达到 $O(Kn^2d)$ 次白化坐标上的操作（定理 17）。对于一般的高斯分布，沿最弱划分树进行解码可达到 J，其中$K-1$ 当划分为二元时，可实现双视图匹配（定理 18）。

6. **特定快速解码器的局限**。 我们证明，不存在一般性的计算下界。对于 $b < 1$，一种解码器会针对每个锚定行，将两个成对候选列表的乘积分布中的三元组逐一检验；要找到正比例的真实三元组，需要 $n^{3-b-o(1)}$ 次检验，这超过二次复杂度，而 $n^{3-b+o(1)}$ 次检验就足够；对于 $K = 5$，舍入

在 Sinkhorn 软平均的前两轮之后，如果第一个视图上的重叠度消失，则更新该视图的算法也会如此。对于精确恢复，当 $I_* > 3\log n$ 时，用于证明多指标指派的线性规划松弛可行的半边对偶价格，在 $2\log n < I_* < 3\log n$ 时不可行；该处的松弛本身是否精确仍未解决（命题 22）。

其他观测模型 (第7节）。近乎精确阈值在观测模型发生若干变化后依然成立，而精确恢复的两种障碍并不依赖于高斯性。

7. **已知分布、缺失视图、未知协方差**。 对于具有固定坐标分布的所有已知乘积分布（定理23，推论 24），以及在视图的随机子集中观测到的对象（命题25），近乎精确阈值和成对阈值均成立。未知边缘协方差不会改变这些阈值，但未知的正交交叉协方差会改变阈值（命题 26）。

8. **非高斯分布下的精确恢复**。 交换条件和稀有子群体条件仍然是必要条件（定理27）。在弱依赖条件下，精确阈值等于最弱划分的散度除以区块数；三视图奇偶分布表明，该阈值可能由划分为三个区块的划分决定（定理 28）。

技术方法。 定理1 可由对候选元组混合对象的各种方式应用一阶矩界，并结合计数下界论证得出。另有三种论证需要更多技巧。定理 2 的逆命题必须排除这样一类估计器：它们使用的视图本身也可能与 a 或 b相匹配。我们加入共享离散标签，使最弱划分树中的每个节点都可解码，并将结果与乘积分布进行比较。对于定理 3，对错误对齐应用并集界只能得到 $J \geq 2\log n$。破坏相同真实配对的对齐会共享这些配对的得分，因此我们针对每种被破坏配对的模式，仅对其下尾计费一次；Dai 等人 (2023年，附录C.3）对两个数据库采用了类似的分离方法。定理 10 的两个方向都使用达到 R´ényi 速率的倾斜分布。逆命题（定理 9）标记出一个随机子群体，使其行服从该分布，并揭示其余所有信息；相关块模型中社区恢复问题的逆命题也以类似方式揭示除数据依赖集合之外的所有信息（Gaudio 等人，2022年；R´acz 和 Zhang，2024）。上界将最大似然限制在这样的对齐上：其中没有任何一组元组的倾斜程度过大；并通过围绕信念传播的不动点展开，将短环计入对齐数量，从而控制剩余的高斯积分。

证明与计算机辅助。 每节陈述其结果并概述主要论证，证明则见附录。关于弱相关之外的精确恢复的一些结果（定理10 和 13）依赖于有限维不等式，这些不等式通过精确有理数运算或向外舍入区间运算得到验证。文中会说明哪些结论属于这种情况；相关 scripts 及其记录的输出作为辅助文件随论文一同发布。三个结果仅给出了证明概要：命题 22（c）以及附录E.11 和 E.12中的两个结果。

相关工作。 多个数据库。 对于三个数据库，Zeynep K 的学位论文（K，2024年，第 7 章）使用联合典型性解码器分析精确恢复问题；该解码器根据三元组所划分的数据库排除错误三元组，并使用 Fano 界；定理 9则补充了一个来自稀有亚群体的必要条件。在 Even 等人（2026年）的模型中，受约束的 K-均值算法使三元组内两两内积之和最大化，而这正是最大似然估计（备注 16），因此定理 3 给出了其高维精确恢复问题的精确常数。K 和 Nazer 研究了相关高斯数据库的检测（K 和 Nazer （2022年），而 Aleshin-Guendel 和 Sadinle 研究了多文件记录链接（Aleshin-Guendel 和 Sadinle（2023年）。

多个图。Vassaux 和 Massouli´e（2026年）表明，当 K 相关高斯图满足 $n\rho^2/\log n > 8/K$ 时，可以实现精确对齐；而在低于这一级别时，所有 K 排列同时正确的顶点比例趋于零，因此精确恢复和近乎精确恢复共享同一个阈值。Ameen 和 Hajek（2026b）通过揭示除一个对应关系之外的所有对应关系，证明了相应的逆命题；对于图而言，他们的最后匹配界$n\rho^2/\log n > 4/(K-1)$在 $K > 2$ 时并不紧。定理 3(b)给出了同类界；对于具有弱相关的点特征，该界是紧的：全局阈值$4/K$ 低于局部阈值 $4/(K-1)$，因此最弱二分划分决定精确恢-复。Even 和 Ganassali（2026年）、Ameen 和 Hajek（2026a）以及 Ahmad Yarandi 和 Ganassali（2026年）研究了图数量增长、检测和属性图。在点特征模型中，一行的每个坐标都随该行移动；而在图模型中，一个置换同时作用于一条边的两个端点，因此图的证明无法直接迁移。

共享信息。 J 是 Chan 等人（2015，式（4.1））提出的多元互信息，也称为共享信息（Narayan 和 Tyagi，2016）。它是 K 个终端在公开讨论下的密钥容量（Csisz´ar 和 Narayan，2004；Chan，2008；Nitinawarat 和 Narayan，2012）；对于成对独立网络，它等于边信息图的强度（Nitinawarat等人，2010）。信息聚类（Chan 等人，2016）对各子集的共享信息进行阈值化，Raman 和 Varshney（2017）则利用它进行图像聚类和配准。稀疏回归、张量主成分分析和图对齐中也已发现全有或全无转变（Reeves 等人，2019；Niles-Weed 和 Zadik，2020；Wu 等人，2022；Vassaux 和 Massouli´e，2026）；Moharrami 等人（2021）和 Ding 等人（2023）研究了具有一般权重的植入匹配。

2. 模型与记号

固定 $K \geq 2$ 并令 $n \to \infty$；所有其他量都可能依赖于 n，log 表示自然对数。对象 $i \in [n]$ 携带一个元组 $X_i = (X_i^1、\ldots\ldots、X_i^K)$，其中 $X_i^a \in \mathbb{R}^{d_a}$。这些元组独立同分布，服从高斯分布 P，其均值和协方差 Σ 为统计学家所知；我们令均值为零。视图 a 是矩阵 $Y^a \in \mathbb{R}^{n\times d_a}$，其行 $\pi_a(i)$ 为 X_i^a。排列 $\pi_1、\ldots\ldots、\pi_K$ 在 $\mathfrak{S}_n$ 上均匀分布，且与这些元组独立。只有相对位置有意义，因此我们按照锚定视图 1 中的行对对象重新标记；重新标记后 $\pi_1 = \mathrm{id}$，且 $\Pi = (\pi_2、\ldots\ldots、\pi_K)$在 $\mathfrak{S}_n^{K-1}$上均匀分布。

对于 $B \subseteq [K]$, P_B 表示 $X^B = (X^a)_{a\in B}$的分布。对于将 $[K]$ 划分为$|\mathcal{P}| \geq 2$个块的划分 $\mathcal{P}$，令

$$D_{\mathcal{P}} = D\Big(P \,\Big\|\, \bigotimes_{B\in\mathcal{P}} P_B\Big), \qquad J = \min_{|\mathcal{P}|\geq 2} \frac{D_{\mathcal{P}}}{|\mathcal{P}|-1}. \tag{1}$$

若 $\Sigma \succ 0$，则 $D_{\mathcal{P}} = \frac{1}{2}\log\left(\prod_{B\in\mathcal{P}} \det \Sigma_{BB}/\det \Sigma\right)$；对于退化的 Σ，散度可能为无穷大。对于 $K = 2$, J 是互信息 $I(X^1; X^2)$；一般而言，它是 X^1、……、X^K 的共享信息（Chan 等，2015；Narayan 和 Tyagi，2016）。对于满足 $|B| \geq 2$ 的 $B \subseteq [K]$，以 P 替代 P_B，由式（1）定义 J_B；若 $|B| = 1$，则 $J_B = \infty$。对每个视图进行白化后，得到联合协方差 R，其块为 $R_{ab} = \Sigma_{aa}^{-1/2}\Sigma_{ab}\Sigma_{bb}^{-1/2}$；弱相关区域为 $\|R - I\|_{\mathrm{op}} \to 0$。

估计器输出映射 $\widehat{\pi}_a : [n] \to [n]$，这些映射可以是随机的，也不一定是双射。其错误为 $M_a = |\{i : \widehat{\pi}_a(i) \neq \pi_a(i)\}|$。若 $\max_{2\leq a\leq K} M_a/n \to 0$ 以概率趋近于一，则恢复为同时近乎精确；若 $\max_{2\leq a\leq K} M_a = 0$ 以概率趋近于一，则恢复为精确。对于视图 $a \neq b$，从 Y^a 的行到 Y^b 的行的真实对应关系为 $\sigma_{ab} = \pi_b \circ \pi_a^{-1}$；映射 $f : [n] \to [n]$ 的重叠度为 $\mathrm{ov}(f) = n^{-1}|\{j : f(j) = \sigma_{ab}(j)\}|$。所有阈值均以固定裕量表述：类似 $J > \log n$ 的条件表示 $\liminf_n J/\log n > 1$。

3. 近乎精确恢复

定理 1 (尖锐阈值)。 *(a)* 如果 下极限$_n$ J/对数 $n > 1$，则下文定义的划分检验解码器能够同时近乎精确地恢复所有相对置换。

(b) 令 $\mathcal{P}$ 为具有 $m \geq 2$ 个块的任意划分，令 B_1 为包含视图 1 的块，并为每个块选择一个代表元 $r_B \in B$，其中 $r_{B_1} = 1$。对于每个 $n \geq 2$ 以及每个估计器，

$$\frac{1}{(m-1)n}\sum_{B\neq B_1} \mathbb{E}M_{r_B} \;\geq\; 1 - \frac{D_{\mathcal{P}}}{(m-1)\log n} - \frac{2}{\log n}. \tag{2}$$

如果 $D_{\mathcal{P}} = \infty$，则该界是无意义的。特别地，如果 上极限$_n$ J/对数 $n < 1$，那么同时进行近乎精确恢复是不可能的。

划分检验。 固定一个锚定行 i 和一个候选元组 $t = (i, j_2, \ldots, j_K)$。给定 Π，t 的视图根据其来源对象被划分为块。如果该恒等划分 为 $\mathcal{P}$，则 t 的特征服从分布 $Q_{\mathcal{P}} = \bigotimes_{B\in\mathcal{P}} P_B$，因为不同对象彼此独立。该划分检验解码器 接受 t，当且仅当

$$L_{\mathcal{P}}(t) = \log \frac{dP}{dQ_{\mathcal{P}}}(t) > (1+\gamma)(|\mathcal{P}|-1)\log n \qquad \text{for every partition } \mathcal{P},$$

使用一个较小的常数 $\gamma > 0$。每个锚定行都会提出一个接受的元组（如果存在），冲突在每个视图中分别解决，剩余的行则进行匹配

任意地。具有恒等划分的错误元组 $\mathcal{P}$ 以至多 $\mathcal{P}$的概率通过检验， $n^{-(1+\gamma)(|\mathcal{P}|-1)}$这是将马尔可夫不等式应用于 $e^{L_{\mathcal{P}}}$所得的结果。对于给定的锚定行，至多有 $n^{|\mathcal{P}|-1}$ 个元组具有此恒等划分，因为锚点所在的块是固定的，而每个其他块选择一个对象。不需要独立性的并集界给出每个锚定行至多 $o(1)$个错误接受。真实元组能够通过检验，因为 $L_{\mathcal{P}}$ 的均值为 $D_{\mathcal{P}}$ $\geq(|\mathcal{P}|-1)J$，且方差至多为 $2K^2D_{\mathcal{P}}$（引理 30），因此切比雪夫不等式足以完成证明。在每个视图中，解决冲突并完成匹配的代价至多是坏行数量的两倍（附录 A）。

逆命题。 揭示 $\mathcal{P}$ 的每个块内的对应关系。剩下的是 $m-1$ 个独立的均匀置换，它们与数据的互信息至多为$nD_{\mathcal{P}}$。一个与估计结果有 t 个位置不同的置换，可以用大约 $t\log n$ 纳特来描述，由此得到（2）。

阈值以下。 某些对应关系仍可能恢复。哪些对应关系能够恢复，取决于视图子集的共享信息。

定理 2（成对可匹配性）。 对于视图 $a \neq b$ 令 $J^*_{ab} = \max_{B\supseteq\{a,b\}} J_B$。

（*a*）若 $\liminf_n J^*_{ab}/\log n > 1$，将划分检验解码器应用于使 B 最大化的视图，可得到一个置换 $\widehat{\sigma}_{ab}$，其$\mathrm{ov}(\widehat{\sigma}_{ab}) \to 1$ 以概率收敛。

（*b*）若 $\limsup_n J^*_{ab}/\log n < 1$，则 $\sup_f \mathbb{E}\,\mathrm{ov}(f) \to 0$，其中上确界取遍所有估计量 $f\colon [n] \to [n]$ 关于 σ_{ab} 的估计量；这些估计量可以使用全部 K 个视图以及自身的随机性。

特别地，如果 上极限$_n$ J/对数 $n < 1$ 并且，对于所有充分大的 n，存在一个在 (1) 的最小值处取得最小值的划分，将 a 和 b 置于不同的块中，那么每个估计器在 a 与 b之间的重叠度都会消失。

对于阈值 c，满足 $J_B > c$ 的包含极大子集，就是 Chan 等人 (2016)的信息聚类中的簇；定理2 表明，在固定裕度下，级别为 $c =$ 对数 n 的簇，恰好是这样的视图组：其中每一对视图都是可匹配的，即使估计器读取所有视图也是如此。定理的最后一句是一个全有或全无结论。被最弱划分分开的视图对，不属于任何共享信息超过J (Chan 等人，2015，定理5.2–5.3和命题 5.3)的子集，因此 $J^*_{ab} = J$ （附录 J.3给出了适用于一般分布的直接证明），最后一句由 (b) 得出。

逆命题的证明思路。部分 (b)并不由定理 1(b)推出，因为估计器可能使用能够自身与 a 或 b匹配的视图。假设 $J^*_{ab} \leq (1-2\delta)\log n$。从 $[K]$开始，沿着达到 J_B的划分，将每个包含 B 且满足 $J_B \leq (1-\delta)\log n$ 的视图集合进行划分。每个同时包含 a 和 b 的集合都会被划分，因此这两个视图最终位于所得树的不同叶节点中。揭示 r 叶节点内部的对应关系。在每个内部节点 B处，对于 B的每一对子节点，为每个对象赋予一个由两个子节点共同看到的 $[M_B]$ 中的均匀标签，并选择 M_B，使该节点的共享信息上升至 $(1+\eta)\log n$。这样，一旦其子节点完成

对齐，就可以通过划分检验解码每个节点，因此在给定增广数据的条件下，叶节点之间的排列具有条件熵 $o(n\log\, n)$。沿树累加散度，可将增广数据置于距离叶节点边缘分布的乘积分布 Q 的散度 $((r-1)\eta + o(1))\, n \log\, n$ 以内。另一方面，一个在 a 和 b之间具有期望重叠度 $u > 0$ 的估计结果（可视为一个置换）会对行进行配对，使得在它们的最低公共节点 B_0处，a和 b 的至少 un 个标签在期望意义下相同；而在 Q 下，一致的数量服从 $\mathrm{Bin}(n, 1/M_{B_0})$分布。按该数量对 Q 进行倾斜表明，两种分布之间的散度至少为 $(4u\delta/K - o(1))\, n \log\, n$，且与 η 无关；取一个较小的 η 即可得到矛盾（附录 J.3；附录 B 则使用高斯坐标代替标签来完成该论证）。

其他分布。 高斯分布进入定理 1和 2 的过程仅通过真实元组上的集中性$L_{\mathcal{P}}$ 实现。第 7 节表明，在下尾条件下，对于已知分布，两种阈值均成立；特别地，这适用于每个不作矩假设的固定坐标分布，也适用于在视图的随机子集中被观测到的对象。

4. 弱相关性下的精确恢复

恢复每一行所需的不仅仅是 $J > \log n$。只要有一对行错误，就足以造成误差，而代价最低的此类错误，是在视图二分划分的一侧交换两个对象。令

$$I_* = \min_{\varnothing \neq S \subsetneq [K]} I\big(X^S; X^{S^c}\big) \tag{3}$$

表示跨越一个二分划分的最小互信息。在弱相关区域中，阈值由此量决定，而不是由 J 决定。

定理 3（精确恢复）。 固定 $\epsilon \in (0,\ 2)$。

(a) 假设 $h = \|R - I\|_{\mathrm{op}} \to 0$ 第 R 节中的白化联合协方差 *2*满足以下条件：若 $I_* \geq (2+\epsilon)\log\, n$ 对所有充分大的 n均成立，则最大似然估计以趋于 *1* 的概率精确恢复所有相对置换。

(b) 对于任意高斯分布，若 $I_* \leq (2-\epsilon)\log\, n$ 对所有充分大的 n均成立，则对于任何估计量，精确恢复所有相对置换的概率趋于零。更准确地说，对于被最小化二分所分隔的每一对视图 a、b，即使揭示两侧内部的所有对应关系，估计 σ_{ab} 在每一行都正确的概率仍趋于零。

部分 (b) 不需要弱相关。在弱信号区间，两个阈值的大小关系是确定的：依次分离 p 块划分中的每个块，会使每一对被分隔的元素恰好被切割两次；而散度在一阶近似下按被割开的元素对相加（见下文命题 6），因此$D_{\mathcal{P}} \geq (1+o(1))\, pI_*/2$ 和 $J \geq (1+o(1))I_*/2$。

注 4（视图数量增长）。 定理 1 和 3 的逆命题对每个$K = K_n$都成立。对于一般分布，当 $K\log\, K = o(\log n)$ 时，定理 1(a) 成立。在弱信号区间 $h \to 0$，这在下文的共同信号模型中意味着 $(K-1)\rho_{\max} \to 0$,

由于对于正的最小割，割权重至多为 u 倍最小割的划分数量不超过 K^{9u}，因此当 log $K = o(\log\ n)$ 时，定理 1(a) 和定理 3(a) 均成立。关于 K 的条件不能省略。对于 $K = \lceil n^{\beta} \rceil$ 且$\beta > 0$固定的情形，取一个由 $K-1$ 个独立双视图信道构成的星形结构，其中每个信道的互信息为 $(2+\eta)$ log n，且 $\eta < \beta$固定。此时 $h \to 0$ 且 $I_* = (2+\eta)$ log n，但每个叶节点被精确匹配的概率至多为 $1 - n^{-\eta - o(1)}$，因此无法精确恢复全部 n^{β} 个叶节点（附录 P.2）。

共同信号。 假设对每个视图都以相同方式应用可逆线性映射后，坐标 $\ell = 1$、……、d彼此独立，并且对象在坐标 ℓ中的 K观测值服从分布 $N(0,\ R_K(\rho_\ell))$，其中 $R_K(s) = (1-s) I_K + s\mathbf{1}\mathbf{1}^{\top}$且$\rho_\ell \in [0,1)$。当 $X_i^a = Z_i + E_i^a$具有共同中心 $Z_i \sim N(0,\ S)$以及独立同分布的视图噪声 $E_i^a \sim N(0,\ N)$，$N \succ 0$时，该条件成立；此时，对于$N^{-1/2} S N^{-1/2}$的特征值 λ_ℓ，有 $\rho_\ell = \lambda_\ell/(1+\lambda_\ell)$。令 $A = \sum_\ell \rho_\ell^2$和 $\rho_{\max} = \max_\ell \rho_\ell$，则 $h = (K-1)\rho_{\max}$。

推论 *5*（常数 $4/K$ 和 $4/(K-1)$）。在此模型中，单元素划分在 *(1)* 中取得最小值；若 $\rho_{\max} \to 0$，则 $J = (K/4 + O_K(\rho_{\max}))A$ 且 $I_* = ((K-1)/2 + O_K(\rho_{\max}))A$。因此，在 $A/$ *log* n 中具有固定裕度时，同时实现近乎精确恢复在 $4/K$ 以上是可能的，在其以下则不可能；在后一情形下，每一对视图的重叠度都消失；精确恢复在 $4/(K-1)$ 以上是可能的，在其以下则不可能。在 *Wee* 等人（*2026*）的模型中，其中 $X \sim N(0, I_d)$，视图为 $X + \tau Z^a$、$\tau^4 = d/(b$ *log* $n)$ 和 $d/$ *log* $n \to \infty$，有 $A/$ *log* $n \to b$。因此，近乎精确阈值为 $b = 4/K$，严格低于其保证 $K/(K-1)$（对每个 $K \geq 3$ 均如此），而精确阈值为 $b = 4/(K-1)$。

弱信号和生成树。 对于完全图上的非负权重 $w = (w_{ab})_{a<b}$ $[K]$，其强度 为

$$s(w) = \min_{|\mathcal{P}| \geq 2} \frac{\sum_{a<b \text{ in different blocks of } \mathcal{P}} w_{ab}}{|\mathcal{P}| - 1}.$$

根据 Tutte–Nash-Williams 定理（Tutte，1961年；Nash-Williams，1961年），将有理权重缩放为整数并通过连续性推广可得，s（w）等于具有容量 w （Cunningham，1985；Nitinawarat 等，2010）的生成树分数打包的最大总权重。

命题 6 （J 是一阶图强度）。设 $h = \|R - I\|_{\mathrm{op}}$ 和 $I_{ab} = I(X^a;\ X^b) = -\frac{1}{2}$对数行列式$(I_{d_a} - R_{ab}R_{ab}^{\top})$。存在常数 C_K，与维度 d_a 无关，使得只要 $h \leq 1/C_K$，每个划分都满足

$$D_{\mathcal{P}} = (1+\tau_{\mathcal{P}}) \sum_{a<b \text{ in different blocks}} I_{ab}, \qquad |\tau_{\mathcal{P}}| \leq C_K h,$$

因此 $J = (1+\tau)\ s\big((I_{ab})_{a<b}\big)$ 且 $I_* = (1+\tau')$ 最小$S \sum_{a\in S,\ b\notin S} I_{ab}$其中 $|\tau|, |\tau'| \leq C_K h$。

误差是相对于割本身而言的，因此当最小割远弱于其余割时，近似仍然准确。在弱信号区间，近乎精确阈值要求成对信息图中包含 $\log n$ 数量的生成树，而精确阈值要求其最小割承载 $2\log\, n$。沿单棵生成树进行匹配时，树上的每条边都需要承载 $\log$(附录 C)。

证明思路：逆命题。令 $\{S, S^c\}$ 达到 I_*，并揭示 S 内部和 S^c内部的所有对应关系。剩下的是 X^S 与 X^{S^c}之间的双数据库问题。在相关性为 r_j的典范坐标中，交换一侧的两行 i，k 会使对数似然降低 $\Delta_{ik} = \sum_j w_j(U_{ij} - U_{kj})(V_{ij} - V_{kj})$，其中 $w_j = r_j/(1 - r_j^2)$。通过 $e^{-\Delta_{ik}/2}$ 倾斜后，每个典范对都相互独立，且归一化常数为 e^{-I_*}，由此得到 $\mathbb{P}(\Delta_{ik} < 0) \geq \frac{3}{8} e^{-I_* - \sqrt{8I_*}}$；而两个共享一行的交换同时获胜的概率至多为 $e^{-3I_*/2}$。若 $I_* \leq (2 - \epsilon) \log\, n$，二阶矩方法便能找到一个胜过真实分配的交换，因此最大似然（它使精确恢复的概率最大）会失败。

证明思路：可实现性。对数似然是两两得分之和（注16）。当备选方案所产生的 F 错误行配对的得分至少不低于它所破坏的真实配对的得分时，T 该备选方案就会胜出。我们按照备选方案的模式对其分组；该模式记录了对于每个对象，其哪些视图保持在一起。一个模式决定了 T。如果它将$m = n^\alpha$ 个对象划分成 $m + g$ 个组，则此类模式至多有 e^L 个，并且对于给定模式，至多有e^M 个备选方案，其中

$$L = (1 - \alpha)\, m \log n + O_K(m), \qquad M = \alpha\, g \log n,$$

因为一个对象属于 n，但每个离开其对象的组只有 m 个可能的去向。在弱信号区间中，当 $\mu = \mathbb{E}T$时，数对（$T - \mu$，F）是联合次高斯的，其方差代理为 V，满足 $\mu^2/V \geq D := 2$（$1 + \epsilon_0$）（$m + g$）$\log\, n$；对于固定的 $\epsilon_0 > 0$，该结论对所有模式以及 $\mathbb{E}F = 0$均一致成立；这里使用了 $I_* \geq$（$2 + \epsilon$）$\log\, n$，以及这样一个事实：划分成 p 个块的划分至少将最弱二分划分的权重分开 $p/2$ 倍。对于单个备选方案，界 $\mathbb{P}$（$F \geq T$）$\leq e^{-D/4}$ 在当 $M \leq L$时超过其数量 e^{L+M}，但一般情况下并非如此：将其用于所有模式需要 $J \geq 2\log\, n$。当备选方案数量超过模式数量时，我们利用它们共享 T这一点：其下尾每个模式只需计算一次，而对 e^M 个备选方案的并集仅在 T 为典型事件时取并集。对于满足 $a \leq s$的 $a + s = 1$，

$$\mathbb{P}(T < s\mu) \leq e^{-a^2 D/2}, \qquad \mathbb{P}(F \geq T \geq s\mu) \leq e^{-(s^2 + a^2)D/2},$$

且二者 $a^2 D/2 \geq L$ 以及 $s^2 D/2 \geq M$ 都可以满足，并且还有余量，这是因为根据柯西–施瓦茨不等式$(\sqrt{L} + \sqrt{M})^2 \leq (1 + o(1))(m + g) \log\, n \leq D/2$(附录 D)。

5. 弱相关之外的精确恢复

在弱信号区间之外，$I_* > 2\log\, n$ 不足以实现精确恢复。原因之一是，$J > \log n$ 也是必要条件，而且可能是更强的条件。

备注 7（一个近线性约束）。取标量视图，其中 $R = \frac{K}{K-1+\delta}\big(I_K - (1-\delta)uu^\top\big), u = K^{-1/2}\mathbf{1}$ 和 $\delta = n^{-2c}$ 满足 $c > 0$ 固定，这是视图之间的一个近线性约束。此时，每个非平凡划分都有 $D_{\mathcal{P}} = c\log n + O_K(1)$，因此 $I_* = c\log n + O_K(1)$，而 $J = c\log n/(K-1) + O_K(1)$，于是当 $c > \max\{2, K-1\}$ 时可以实现精确恢复，当 $c < \max\{2, K-1\}$ 时则不可能（附录 P.1）。对于 $K \geq 4$ 和 $2 < c < K-1$，最小割超过 $2\log n$，但 $J < \log n$，甚至近乎精确恢复也无法实现。

第二个障碍无法由 I_* 和 J 共同捕捉。它出现在中间尺度上：一组 n^θ 个对象，即 $0 < \theta < 1$，其中特征碰巧呈现非典型状态，而这些对象的视图之间共享的信息过少，因而无法进行匹配。对有限数量对象的任何错误模式都无法检测到这一点。

在本节的其余部分，每个元组由 d 个具有固定密度 p 的独立同分布坐标组成，定义在 $\mathbb{R}^K$ 上；每个视图对应一个标量，而 I_*、J 是逐坐标定义的，因此上述条件写作 $dI_* > 2\log n$ 和 $dJ > \log n$。对于 $0 < \alpha < 1$，令 $D_\alpha(p\|v) = (\alpha-1)^{-1}\log\int p^\alpha v^{1-\alpha}$ 表示阶数为 α 的 R´enyi 散度。

定义 8（稀有子群体速率）。对于 $[K]$ 的一个划分 $\mathcal{P}$，其包含 $q \geq 2$ 个块，令

$$R_{\mathcal{P}} = \frac{1}{q-1}\min_{v=\bigotimes_{B\in\mathcal{P}} v_B} D_{(q-1)/q}\big(p \,\big\|\, v\big),$$

在各块上可分解的密度中取最小值 $\mathcal{P}$。等价地，

$$R_{\mathcal{P}} = \min_Q \Big\{ D(Q\|p) + \frac{1}{q-1} D\Big(Q \,\Big\|\, \bigotimes_{B\in\mathcal{P}} Q_B\Big)\Big\},$$

其中 Q_B 是 Q的块边缘分布。

第二种形式解释了第一种形式。约 $n^\theta = n\,e^{-dD(Q\|p)}$ 个对象的经验坐标分布接近 Q，类比定理 1 在 n^θ尺度下的情形，要在 $\mathcal{P}$的各块之间匹配这些对象，需要 $dD(Q\|\bigotimes_B Q_B)/(q-1) > \theta\log n$。对于使其达到最小值的 Q，该值为 $dR_{\mathcal{P}} > \log n$。这两种形式通过恒等映射

$$D(Q\|p) + \frac{D(Q\|v)}{q-1} = \frac{D(Q\|Q_v)}{\alpha} + \frac{D_\alpha(p\|v)}{q-1}, \qquad \alpha = \frac{q-1}{q}, \quad Q_v \propto p^\alpha v^{1-\alpha},$$

先对 Q 取最小值，再对 v取最小值。在固定参考分布 v下，对Q 取最小值，就是 van Erven 和 Harremo´es 的 Rényi 散度变分公式¨ (2018年，定理 30)。对于有限字母表，Tomamichel 和 Hayashi (2018年，定理 9–10 和命题 15) 表明，针对由 q 因子构成的乘积分布进行检验时，其最优错误指数可由经过同样最小化的 R´enyi 散度（阶数高于 $(q-1)/q$）给出，因此此处采用的阶数是他们所述范围的下端点。取 $Q = p$ 可得 $R_{\mathcal{P}} \leq D_{\mathcal{P}}/(q-1)$，因此下面的条件至少与定理 1中的条件同样强。

定理 9（稀有亚群体）。固定 K，一个密度 p 和一个划分 $\mathcal{P}$，其包含 $q \geq 2$个块。如果 $dR_{\mathcal{P}} \leq (1-\epsilon)\log n$ 对所有充分大的 n 都成立，则任何估计器恢复所有相对置换的概率都趋于零。对于高斯 p 和两个块 S，S^c，$R_{\mathcal{P}} \geq I(X^S; X^{S^c})/2$，因此只有划分为至少三个块时，才会对定理 3（*b*）增加新的内容。

对于两个块，新的条件因此由交换条件蕴含。对于两个视图，这对任何分布都成立，无论是否为高斯分布。如果 a 是核 $\sqrt{p(x,y)}$ 的平方奇异值向量，则一个 m-环的每个对象的切尔诺夫指数为 $-\log \|a\|_m$（Cullina 等人，2018年）。该指数关于 m单调不减。对于一次交换（$m=2$），它是 Cullina 等人（2018）二阶循环互信息的一半，而对于高斯分布，它是互信息的一半。当 $m \to \infty$ 趋于 $-\log \max_i a_i$时，它等于 $R_{1|2}$，即 Lapidoth 和 Pfister（2019）的 $1/2$ 阶依赖度量。对于三个或更多块，导致失败的配置是在 n^θ 个对象之间进行的一般置换。造成失败的是它们的熵，而不是任何固定模式的概率。在稀疏图对齐中，精确恢复同样会受到交集图中孤立顶点的阻碍（Cullina 和 Kiyavash，2017年；Ameen 和 Hajek，2025年）；这里的小组由坐标异常的对象组成。

新条件在三个视图的情况下就已适用，而此时它就是阈值 .

定理 10 (三个视图具有相等的负相关)。 设 $K=3$ 和 $p=N(0,R)$具有单位方差，且所有相关性均等于 $r\in(-1/2,0)$，并设 $x\in(0,1/2)$ 是 $r=-3x/(2+x+2x^2)$的解。那么

$$R_{1|2|3}=-\tfrac{1}{4}\log(1-2x)-\log\left[(1+x)(1-\tfrac{x}{2})\right]<\frac{I_*}{2}.$$

对于每个固定的 $\epsilon>0$，若 $d\geq(1/R_{1|2|3}+\epsilon)\log n$，则最大似然估计器以趋于 *1* 的概率精确恢复所有相对置换；若 $d\leq(1/R_{1|2|3}-\epsilon)\log n$，则任何估计器都会以趋于 *1* 的概率失效。

在 $r=-1/3$ 处，有 $x=2-\sqrt{3}$, $I_*=\frac{1}{2}\log_3 2$, $J=\frac{1}{4}\log_{27}16$ 以及

$$R_{1|2|3}=\tfrac{1}{2}\log\tfrac{16}{27}+\tfrac{3}{4}\log\tfrac{3+2\sqrt{3}}{4}\approx 0.09835.$$

阈值 $1/R_{1|2|3}\approx 10.17$ 超过 $2/I_*\approx 9.87$ 和 $1/J\approx 7.64$：当 $9.9\leq d$/对数 $n\leq$10.1 时，$dI_*>(2+\epsilon)$ 对数 n 和 $dJ>(1+\epsilon)$ 对数 n 均对某个固定的 $\epsilon>0$成立，但精确恢复仍然失败。

下界由定理 9给出。对于 $r=-1/3$ 处的上界，证明只使用显式的十进制不等式，其中最严格的一项裕度不到百分之一。对于一般的r ，证明借助计算机辅助完成：参数范围分为四段，在每一段上，通过精确有理数运算或向外舍入区间运算，对有限个不等式进行验证，并结合对短游走的机器枚举（附录 E.5）。我们猜想，对于高斯分布不存在其他障碍。

猜想11 （高斯坐标分布的精确恢复）。 令 K 为固定值，且 $p=N(0,R)$ 满足 $R\succ 0$。设

$$L(R)=\min\left\{\frac{I_*}{2},\ \min_{|\mathcal{P}|\geq 3} R_{\mathcal{P}}\right\}.$$

在固定裕度下，当 $dL>\log n$ 时，最大似然可精确恢复所有相对置换；而当 $dL<\log n$ 时，没有任何估计器能够做到这一点。

不可能性由定理3（b）和9给出；尚未解决的是充分性。猜想与上述所有已解决的情形一致。在弱信号区间 $L = (1 + o(1))\, I_*/2$，这与定理3一致。对于备注7中的近线性约束，单元素划分给出$R_{\mathcal{P}} = c \log n/(K-1) + O_K(1)$，由此得到阈值最大值$\{2, K-1\}$。对于$K = 2$，只剩下交换条件。定理10 给出了三个视图且负相关相等时的猜想。接下来的两条陈述说明新条件何时可能起约束作用，并列举了猜想已获证明的其他情形。

命题 12（新条件的结构）。固定 K。

（*a*）有界模式。对于每个密度 p，涉及有界数量对象的每个错误模式，其切尔诺夫指数至少为 $\tfrac{1}{2} \min_S I_2^{\circ}(S)$（按每个移动对象计），- 且当交换跨越一个最小化割时取等号。这里的 I_2° 是 *Cullina* 等人提出的二阶循环互信息（*2018*年），对于高斯分布，它等于互信息。

（*b*）高斯参考分布。对于高斯分布 p，定义 *8* 第一种形式中的极小化元唯一，且为中心高斯分布密度。

（*c*）新条件何时起作用。对于具有非负相关性的 $K = 3$，以及当视图为具有正载荷的共同高斯因子加独立噪声时的每个 K，$L = I_*/2$。对于 $K = 4$，存在两两相关性均为正的分布，使得 L 小于 $I_*/2$ 和 J。对于满足$0 < J \le I_*/2$的每个高斯分布，都存在某个划分使得 $R_{\mathcal{P}} < J$。

定理 13（猜想的其他情形）。猜想 *11* 在以下情形下成立 *s.*

(i) $K = 3$ 具有相等相关性 $r \in (0,1)$；此时 $L = I_*/2$。***(ii)*** $K = 4$ 具有相等相关性 $-1/4$；此时 $L = R_{1|2|3|4}$ 和 $9.8511971 < 1/L <$9。8512024，而 $dI_* > 2 \log n$和 $dJ > \log n$只需要 $d/\log n > 8.511$。***(iii)*** $K = 3$ 高斯链，$r_{13} = r_{12} r_{23}$ 视图可重新标记；此时 $L = I_*/2 = \min\{I_{12}, I_{23}\}/2$。***(iv)*** 每个具有单位方差的 R，$I_* > 0$ 和 $\|R - I\|_{\mathrm{op}} \le \gamma_K$，其中常数 $\gamma_K > 0$仅取决于 K。

更一般地，附录 *E.9* 给出了关于 R 的一个条件；通过对大小至多为 K 的矩阵进行有限次区间算术检验即可验证该条件，在此条件下，猜想对 R 成立。该猜想在情形 ***(iv)***、情形 ***(ii)*** 以及定理 $r = -1/3$ *10*中的 部分情形成立；对于接近奇异性的负等相关分布，该猜想不成立，而定理 *10* 则通过其他方法证明。对于 $9/20 \le r \le 7/8$，情形 ***(i)*** 的证明以及情形 ***(ii)*** 的证明都借助了计算机。

命题 14 (临界窗口)。对于 $K = 3$ 和相等相关性 $-1/3$，令 $s_n = dR_{1|2|3} -$ 对数 n。当 $s_n \to +\infty$ 时，精确恢复以趋近于一的概率成功；当 $s_n + \tfrac{3}{4}$对数对数 $n \to -\infty$时，以趋近于一的概率失败。 -

另外两个族，证明概要如下。 附录 E.11 针对一族轴对称三视图协方差陈述了猜想，$r_{12} = r_{13} = r \in [-11/30, -3/10]$并且 $r_{23} = r + \delta$ 对于每个 δ 满足 $R \succ 0$；在 $r = -1/3$ 以下三个项中的每一项$L =$最小值$\{R_{1|2|3}, I_1/2, I_2/2\}$中，$I_i = I(X_i;\ X_{[3]\setminus\{i\}})$ 都是某个区间上的最小值，δ。附录E.12 指出，对于具有相等相关性 $-9/23 \le r < 0$ 或 $-1/2 < r \le -1/2 + 2^{-12}$的三个视图，定理 10 有一个无需区间算术的证明，并且该猜想随后也对 R某个邻域内的每个相关矩阵成立。对于这两个结果，本版本给出了证明的结构和新的要素，但未给出全部细节；我们不将它们计入上述已证明的情形。

证明思路：下界。令 v 达到第一种形式，并令 Q 为密度与 $p^{(q-1)/q} v^{1/q}$ 成正比的倾斜分布，从而达到第二种形式；分别选择的两种形式的最优解未必相容。于是 $D(Q\|p) \le R_{\mathcal{P}}$，因此在定理 9 的假设下，$\theta \ge \epsilon$。在一个 $Q^{\otimes d}$- 概率为 $1 - o(1)$ 的事件 A_d 上，以概率 $e^{-d(D(Q\|p)+\delta)} \prod_{j \le d} (Q/p)(x_j) \le 1$ 独立地标记每个对象。$m \ge n^{\theta - O(\delta)}$ 个标记行随后独立同分布地来自 $Q^{\otimes d}(\cdot \mid A_d)$。揭示标记位置、所有未标记行及其对应关系，以及每个块内部的对应关系。剩下的是 $q - 1$ 个对象之间的均匀双射。若标记行来自在各块上分解的 $v^{\otimes d}$，则它们不会携带关于这些双射的信息，因此任何估计器的成功概率至多为 $(m!)^{-(q-1)}$。每个标记行的换测度代价至多为 $e^{d(D(Q\|v)+\delta)+o(1)}$，当 $dR_{\mathcal{P}} \le (1 - \epsilon) \log n$ (附录 E.1) 时，这不足以弥补损失。揭示数据依赖集合之外的所有信息，也导出了 Gaudio 等（2022）针对相关块模型的精确社区恢复反定理，以及 R´acz 和 Zhang（2024）的相应结果；Ding 等人（2023，附录 E.3）则将植入匹配反向定理限制在植入权重位于两种权重分布共同支撑集内的顶点上；这里，该集合来自似然比标记，因此保留行独立同分布地来自倾斜分布。块乘积参考分布在使用相同的 $|\mathcal{P}| - 1$ 归一化后，也给出了 Tyagi 和 Watanabe（2015）的多终端密钥反向定理。对于命题 14，标记窗口在对数似然比上的宽度为 $O(1)$，这使标记对象的数量减少一个 $d^{-1/2}$ 因子，并产生 3 4 log log n 项。

证明思路：上界为 $r = -1/3$。对错误对齐应用并集界无法达到阈值：其总和由这样的稀有事件主导——在这些事件中，许多元组发生倾斜，并且同时有非常多的对齐超过真实分配。令 $q_3 \propto p^{2/3} (\phi^{\otimes 3})^{1/3}$ 表示单元素划分下的倾斜分布，其中 ϕ 为其最优高斯参考分布。解码器在通过检验的对齐中最大化似然：对于它们的元组组成的任意集合 S，其 $\sum_{i \in S} \log (q_3^{\otimes d} / p^{\otimes d})(x_i)$都不得超过 $\log \binom{n}{|S|} + \eta |S| \log n$。真实对齐以趋近于一的概率通过检验，而该检验将看起来像从 q_3 中抽取的元组数量限制在上述启发式方法给出的规模 n^{θ} 以内。错误对齐与真实分配的差异体现在对象与元组之间的二分图的连通分量上。在一个包含m 个对象、其中 k 个被划分为三个的连通分量中，它超过真实分配且两者

对齐均通过检验的概率至多为

$$\exp\!\left\{\log\tbinom{n}{k}+\eta k\log n-d\left[(m-k)R_2+2kR_3\right]\right\}\,\mathcal{I}_G^{\,d},$$

其中， $R_3=R_{1|2|3}$, $R_2>R_3$ 是划分为两个块的速率，而 $\mathcal{I}_G$ 是对该连通分量的边进行的高斯积分。结合真实似然与错误似然的几何平均数，该检验将每个被划分为三个部分的对象处的 $\sqrt{p}$ 替换为倾斜分布，这正是产生 R´enyi 阶 $2/3$的原因。若 $\mathcal{I}_G\le 1$，则当 $dR_3>\log n$时，该界就会超过约为 $n^m m^k$的分量数。环会提高 $\mathcal{I}_G$。在高斯信念传播的不动点处将其写出，可得 $\mathcal{I}_G\le \det(I-B)^{-1/2}$，其中 B 是一个由非回溯步构成、绝对行和低于 0.556 的矩阵；我们将其闭游走归入闭合短环的边。当连通分量逐边构建时，一条闭合短环的边只有 $O(1)$ 个可能的端点，这使计数减少了 $\log m$。我们可以假设 $d<10.2\log n$。随后，每条这样的边在积分中至多耗费 $0.0488\,d<\frac{1}{2}\log n$，并在 $m\ge\sqrt{n}$时节省 $\log m\ge\frac{1}{2}\log n$。较小的分量则直接通过 Bhattacharyya 指数处理，并用真实行空间与候选行空间之间的主角来表示。

证明思路：其他情形。 对于其他负相关性，采用相同的方案，但短环的代价和边的预算都取决于 r，单个估计无法覆盖整个族。对于中等程度的 $|r|$，通过枚举标记边可能邻域中的短非回溯返回，对上述两步范数上界进行细化。在奇异端点附近 $r\to -1/2$，这些范数上界不再成立，因此改用加权 Ihara–Bass 公式，比较图与其覆盖树之间的有效电阻。对于正等相关性，起约束作用的是交换条件；对于范围 $9/20\le r\le 7/8$，则对真实行的两个平方和进行指数倾斜，此后只涉及主角正弦平方的前两个矩；而对于 $r\ge 7/8$，首先通过双视图最大似然将每个视图与锚点对齐，然后在该对齐附近最大化联合似然。定理 13 的判据将这一方案应用于一般协方差：针对每个划分的局部倾斜因子、信念传播消息的不变盒，以及标记边代价与计数增益之间的预算不等式。

6. 解码器与计算

我们首先比较三种自然的视图组合方式，从而解决 Wee 等人 (2026年) 提出的问题；然后给出两个解码器，其评估的候选元组远少于定理 n^K 1的候选元组。本节最后一部分说明一些自然的快速解码器在哪些地方会失效。表格 1汇总了相关性相等时的级别。

定理 15 (层级)。 考虑 *Wee 等人 (2026年)* 的模型，其中 $d/\log n\to\infty$和$A/\log n\to b$。

(i) 树解码器。 在 $[K]$上固定一棵生成树 $\mathcal{T}$，仅根据Y^a 和 Y^c 估计每个 σ_{ac}、$\{a,c\}\in\mathcal{T}$，并沿 $\mathcal{T}$进行组合。对于 $b>2$，一个合适的双视图估计器

使这一点同时近乎精确。对于 $b < 2$，每一个这样的解码器在 T 的每条边上均具有消失的重叠率。

（*ii*）逐行两两求和（*Wee* 等人，*2026*年，算法 *1*）。对于每个锚定行 i，解码器根据 $j_1 = i$，在 $(j_2, \ldots, j_K) \in [n]^{K-1}$ 上为 $\sum_{1 \le a < c \le K} \langle Y^a_{j_a}, Y^c_{j_c} \rangle$ 提议一个最大化器；随后解决冲突，并将结果补全为排列。如果 $b > K/(K-1)$，则它同时是近乎精确的（*Wee* 等人，*2026*年，定理 *2.5*）。如果 $b < K/(K-1)$，那么对于每个固定的锚定行 i，最大化器对所有 $a \ge 2$ 都满足 $j_a \neq \pi_a(i)$，其概率趋于 *1*。考虑任意一种补全方式：对于每个视图 $a \ge 2$，（*C1*）恰好保留提出每个已提议行的一个锚定行，并且（*C2*）将其余锚定行双射匹配到未被提议的行，其中这两种选择均是原始提议表和独立随机性的函数。那么，每个锚定排列 $\widehat{\pi}_a$ 的期望重叠度都趋于零，两个非锚点视图之间的每个复合估计 $\widehat{\pi}_c \circ \widehat{\pi}_a^{-1}$ 也同样如此（附录 *F*）。

(iii) 划分检验。对于 $b > 4/K$ 定理 *1* 的解码器可同时达到近乎精确；而对于 $b < 4/K$ 不存在任何解码器（推论 *5*）。

注 16（后验分布仅取决于两两得分）。若 $\Omega = \Sigma^{-1}$，则在均匀先验下，$\log p(\Pi \mid Y) = C(Y) - \sum_{a<c} \sum_i (Y^a_{\pi_a(i)})^\top \Omega_{ac} Y^c_{\pi_c(i)}$，因为每个视图的二次项在其行置换下保持不变。在 Wee 等人（2026）的模型中，对于 $a \neq c$，Ω_{ac} 是单位矩阵的负倍数，因此联合最大后验估计最大化与算法 1 相同的总和，但是在所有排列上联合进行。$K/(K-1)$ 与 $4/K$ 之间的差距源于一次只最大化一个锚点行，这使得多个锚点行可以声称属于同一个内部一致的对象。一般而言，不能删除 (ii) 中关于补全的条件：重读数据的补全可以恢复更多内容（备注 37）。

表 1：在弱极限下，具有相等相关性的共同信号所对应的级别，单位为 b。每种解码器在超过其级别时均能成功。在 $4/(K-1)$ 以下，没有估计器能恢复每一行；在 $4/K$ 以下，没有估计器能达到近乎精确；最后两种解码器在其级别以下会以定理 15所述的意义失败。分层筛选需要 $K \ge 3$；运算量按白化坐标计算。

解码器	恢复	操作	级别	$K = 5$
最大似然（定理 3）	精确	—	$4/(K-1)$	1
划分检验（定理 1）	近乎精确	$O_K(n^K + n^2 d)$	$4/K$	0.8
分层筛选（定理 17）	近乎精确	$O(Kn^2 d)$	$4/\min\{K, 4\}$	1
逐行两两求和（定理15）	近乎精确	$O_K(n^K + n^2 d)$	$K/(K-1)$	1.25
树解码器（定理 15）	近乎精确	$O(Kn^2 d)$	2	2

例如，在 $K = 5$和 $b = 1.1$时，最大似然可恢复每一行，分层筛选则能以二次时间恢复除 $o(n)$ 行以外的所有行，而逐行求和和树解码器均会失败（表 1）。

分层筛选。 划分检验无需查看所有 n^K 元组。令 $c(\mathcal{P})$ 为被 $\mathcal{P}$分隔的视图对数量。在推论 5的模型中，前 $\mathcal{P}$个视图的划分 s对应的真实元组的对数似然比均值为 $(b/2 + o(1))\, c(\mathcal{P}) \log n$；而恒等划分为 $\mathcal{P}$的错误元组超过 $\alpha\, c(\mathcal{P})\log n$ 的概率至多为$n^{-\alpha c(\mathcal{P})}$。因此，可以从短列表开始，逐个视图扩展元组，而且存留下来的错误前缀数量很少。

定理 17(分层筛选)。 在推论 5的模型中，假设 $\rho_{\max} \to 0$ 且 $A/\log n \to b < \infty$，并固定 $3 \le m \le K$。若 $b > 4/m$，则下面的解码器能够以近乎精确的方式同时恢复所有相对置换，在白化坐标上需要进行 $O_m(Kn^2d + n^{\gamma_m})$ 次实数运算，其中

$$\gamma_m = \max\left\{2,\ 1 - \frac{2}{m} + \frac{\lfloor (m+1)^2/4 \rfloor}{m}\right\}.$$

在 $m =$ 最小化$\{K, 4\}$ 的情况下，代价为 $O(Kn^2d)$，保证对 $b > 4/3$ 为 $K = 3$，对 $b > 1$ 为 $K \ge 4$；对于 $K = 3,\ 4$ 这就是阈值 $4/K$。

当坐标并非已经独立时，白化是一次单独的预计算，需要进行 $O(d^3 + Knd^2)$ 次操作；在 Wee 等人（2026）的模型中则不需要白化。固定$\alpha \in (2/m, b/2)$和 $\beta \in (1, mb/2)$。解码器：(i) 对于每个锚定行和每个视图 $a \in \{2, \dots, m\}$，列出视图 a 中与该锚定行的双视图对数似然比超过 $\alpha\log n$的各行，并丢弃列表长度超过 $\lceil n^{1-2/m} \rceil$的锚定行；(ii) 对于 $s = 3$、.....、m，用锚点在视图 s 中的列表扩展每个保留下来的前缀，并在其 s 个视图上通过级别为 $\alpha c(\mathcal{P})\log n$的每个分区检验时保留该扩展；如果在某个中间层级上有超过 $2n^{u_s}$ 个前缀存活，则中止，其中 $u_s = \max\{1, s - s(s-1)/m\}$；(iii) 保留一个不相交的 m 元组的极大集合并完成；(iv) 将每个剩余视图与 m 个对齐行的归一化和进行匹配，该归一化和与该视图的互信息为 $(m/2 + o(1))A$，所用方法是在级别 $\beta\log n$ 下进行双视图检验并采用极大匹配（附录 G）。

定理 17 的证明思路。在推论 5 的模型中，一个由 s 个视图组成、其恒等划分有 p 个块的错误前缀，通过针对其自身划分的检验的概率至多为 $n^{-\alpha c(\mathcal{P})}$，而这样的前缀至多有 n^p 个。由于 $c(\mathcal{P}) \ge (p-1)(s - p/2)$，且 $p - \frac{2}{m}(p-1)(s-p/2)$ 关于 p 是凸函数，因此它不超过其在 $p = 1$ 和 $p = s$ 处取值中的较大者，即 u_s；又由于 $\alpha > 2/m$，每个划分还会额外损失一个 $n^{-(\alpha - 2/m)c(\mathcal{P})}$ 因子，因此通过第 s 级的错误前缀的期望数量为 $o(n^{u_s})$。正确前缀都能通过检验，因为每个检验的均值为 $(b/2 + o(1))c(\mathcal{P})\log n$，方差为 $O(\log n)$。由于 $u_m = 1$，最后一级留下 $o(n)$ 个错误的 m 元组，而每个元组在极大不相交选取中至多阻挡 m 个真实元组。列表的期望长度至多为 $1 + n^{1-\alpha}$，因此，丢弃列表长度超过 $n^{1-2/m}$ 的锚定行，会损失其中的 $o(n)$；并将至多

$2n^{u_s}$ 个前缀通过这类列表扩展，可得到指数 γ_m。其余每个视图都通过双视图检验与 m 个对齐行的归一化和相连，该归一化和包含了这些行关于该视图的全部信息。

该保证是渐近的。对于 $n \le 8192$ 间隙 $(b/2-\alpha)\log n$ 与 $O(\sqrt{\log n})$ 分数的波动幅度相当；在我们的模拟中，分层筛选恢复的完整元组少于消息传递或 Sinkhorn 缩放，而后两者没有理论保证（附录 Q）。

沿最弱划分解码。 对于 $U \subseteq [K]$，U 的最弱划分是指能够达到 J_U的划分。选择 $[K]$的最弱划分，再依次选择其每个块的最弱划分，直至单个视图，由此得到一棵树；令 p^* 表示某个节点处块数的最大值。

定理 18 (递归最弱分区解码)。 令 $\Sigma \succ 0$ 和 $J \ge (1+\epsilon)\log n$，其中 $\epsilon > 0$为固定值。从叶节点向上解码。在每个节点，将各子节点的对齐行拼接成一个超视图；对于子节点的每种划分 $\mathcal{R}$，在级别 $(1+\gamma)(|\mathcal{R}|-1)\log n$处进行检验，并接受通过检验的超视图行元组，其中 $\gamma \in (0,\epsilon)$为固定值；保留一个互不相交的已接受元组的极大集合并完成匹配。这样可同时近乎精确地恢复所有相对置换，在 $O_K($ $d^3+nd^2+n^2d)$ 预计算之后，需要 $O_K(n^{p^*})$ 次操作，$d=\sum_a d_a$。如果树中的每个最弱划分都包含两个块，那么只需对超视图进行 $K-1$ 双视图匹配；若这些二分划分还是各自节点唯一的最弱划分，则通过贪心合并当前互信息最大的两个簇的拼接结果即可找到这棵树。

每个节点都继承该裕度，因为 $J_B \ge J_U$ 对于 B 中最弱划分的每个块 U，这些块的超视图恰好具有 J_U的共享信息。错误会加性传播：一个有 p 个子节点的节点至多有 $F+pT+(p+1)\sum_j E_{B_j}$行错误，其中 F 统计未通过检验的真实元组，T 统计通过检验的错误元组，而 E_{B_j}表示子节点中的错误行（附录H）。

例19（聚类视图）。取 $K=4$ 个视图 A、B、C、D（图 2），其中对象的四个观测值的每个坐标都是独立的 $N(0,R)$ 向量，且方差为单位方差；处于弱信号区间 $d/\log n \to \infty$，总信号为 $S=(d/\log n)\sum_{a<b}\rho_{ab}^2$。在均匀设计中，六个相关性均相等且为正。在聚类设计中$\rho_{AB}=\rho_{CD}=s>0$，四个交叉相关性为 $s/4$。S 中的阈值为：

设计	树解码器逐行求和		近乎精确（J）	精确（I_*）
均匀	12	8	6	8
聚类	72	22.5	18	36

在聚类设计中，$\{AB,CD\}$ 同时达到 J 和 I_*。一旦 $S>4.5$，AB 和 CD 这两对便可匹配；而对于 $S<18$，每个跨簇对的重叠度都趋于消失（定理 2）。由于块对数似然比的交叉项是 $\langle A_i+B_{\alpha(i)}$的倍数，$C_j+D_{\delta(j)}\rangle$，阈值 18 也由三种双视图分配达到：$A$ 到 B、C 到 D，以及求和后的行 $A+B$ 到 $C+D$，所需时间为 $O(n^3+n^2d)$（附录 H）。

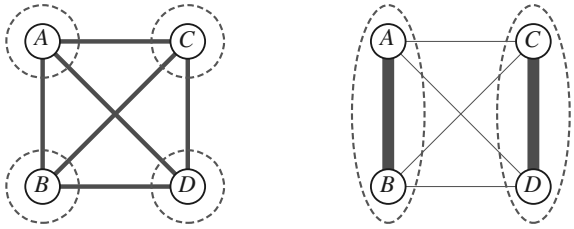

图 2：示例 19 中的两种四视图设计，其总信号相同 S。边的宽度与成对互信息成正比，虚线环标记出达到 J 的划分。

注20（对行求和可能无法达到阈值）。 在例 19 的坐标模型中，设 $\rho_{AB} = \rho_{CD} = 2r$ 和 $\rho_{AC} = r$ 满足 $r \to 0$，令其他三个相关性为零，并令 $dr^2/\log n \to 3$。最弱划分为 $\{AB, CD\}$，满足 $J/\log n \to 3/2$；但求和后的行 $A+B$ 和 $C+D$ 的互信息为 $(3/8 + o(1))\log n$。于是，任何只能观察两个正确对齐簇的求和后行的估计器，其簇间重叠度都会趋于消失；而定理 18中的块似然比则能够成功。

注21（更多计算结果）。(a) 异质分布。分层筛选可推广到成对信息量不等的弱律 w_{ab} log n。对于一个三视图种子，其 (w_{12}, w_{13}, w_{23}) =(0.78, 0.75, 0.72)，以及一个与种子之间权重为 (0.58, 0.48, 0.02) 的第四个视图，每一对都低于双视图阈值，因此没有树解码器能够成功；然而，该种子具有 1.125 log n 的共享信息，第四个视图携带关于它的 1.08 log n 信息，而筛选在经过 $O(n^2 d)$ 的分数预计算后，只需进行 $O(n^2)$ 次候选检查。(b) 多项式时间精确恢复。当 $I_* \geq (2+\epsilon)$ log n 处于弱信号区间时，级别为 $(|\mathcal{P}| + \eta)$ log n 的划分检验只接受正确元组，并留下 $O(n^{1-c})$ 个对象未匹配。多指标指派的线性规划松弛含有 n^K 个变量，当 $I_* \geq (3+\epsilon)$ log n 时是精确的。(c) 低次多项式。当相关性相等、$(\log n)^{41+40\eta} \ll K = n^{o(1)}$ 且 $d \gg K \log n$ 时，次数为 $(\log n)^{1+\eta}$ 的多项式只能以随机猜测的水平（按归一化均方误差衡量）估计两行是否包含同一对象，条件是 $b = (4+\epsilon)/K$，尽管恢复在信息论上可行；这是 Even 等人（2026）的定理 2.2 在我们参数化下的表述。对于固定的 $b > 0$ 和较大的 K，次数为 $O_b(\log n)$ 时，该误差为 $o(1/n)$，因此这类障碍要求 $b_n \to 0$。证明见附录P.3。

自然快速解码器的界限。 对于固定的 $K \geq 5$ 和 $4/K < b < 1$，划分检验可在多项式时间内实现近乎精确恢复，但尚不清楚 $O(n^2 \,\text{poly}(d,\ K))$ 次运算是否足够——

这一问题尚无一般下界。下面的部分 (a) 和部分 (c) 表明，两种自然的快速方法无法覆盖整个窗口，而部分 (b) 表明，备注21中的对偶证书无法达到精确阈值。

命题 *22*（自然解码器的障碍）。*(a)* 显式枚举两两候选项。在推论 *5* 的模型中，所有 ρ_ℓ 均相等，$d/$ 为 *log* $n \to \infty$，且 $A/$ 为 *log* $n \to b \in [0, 1)$，令每个锚定行 i 在另外两个视图中分别有候选列表 $L_{i,2}$ 和 $L_{i,3}$，其中每个列表都是锚点、其自身视图以及独立随机性的函数，长度可以依赖于数据且任意。假设解码器对其扫描的每个锚定行 i，都会访问 $L_{i,2} \times L_{i,3}$ 中的每个元素。若对某个固定的 $\delta > 0$，其访问次数至多为 $n^{3-b-\delta}$ 的概率趋于 *1*，则它扫描到的真实三元组只有 $o_{\mathbb{P}}(n)$。长度为 $\lceil n^{1-b/2+\kappa} \rceil$ 的列表通过原始内积可覆盖 $n - o_{\mathbb{P}}(n)$ 个真实三元组，访问次数为 $O(n^{3-b+2\kappa})$，对任意固定的 $\kappa > 0$ 均成立。

(b) 半边对偶价格。对于高斯分布 $\|R - I\|_{\mathrm{op}} \to 0$ 且 $I_*/$对数 $n \to c \in (2,\ 3)$，当 *21(b)* 中的备注 $I_* \geq (3 + \epsilon)$ 对数 n时，用于证明线性规划松弛成立的对偶价格——即每个行上每对视图真实得分的一半——以趋近于 *1* 的概率违反某个对偶约束。

(c) Sinkhorn 软平均（证明概要）。在 *(a)* 的模型中，取 $K = 5$，考虑附录 *Q* 中的软平均解码器，并在实数运算下进行：对成对表格执行固定次数的 *Sinkhorn* 步骤缩放（*Sinkhorn*，*1964*年；*Sinkhorn* 和 *Knopp*，*1967*），利用场 $S_{a1} + \sum_{c \neq a} S_{ac} Q_c$对每个视图的表格进行 *Gauss–Seidel* 更新，其中 S_{ac} 是视图 a 与 c之间归一化内积构成的矩阵；然后对每个锚点行取排名前 W的候选项并进行对角线补全。在第一轮和第二轮中，第一个更新后的视图的表格给真实行的得分为 $(\sqrt{b} + o_{\mathbb{P}}(1))\sqrt{}$对数 n，而其权重趋近于零；在第二轮中，每个锚点行都有一个错误得分高于 $t\sqrt{}$对数 n 的候选项，对任意固定的 $t < \sqrt{2}$均如此。经过任意正对角缩放后，排名前 $W \leq Cn^{\omega}$ 的候选项 $\omega < 1 - b/2$中，只有 $o_{\mathbb{P}}(n)$ 个锚点行包含真实行，因此在任一轮之后对该视图进行舍入，重叠率都会消失。如果第一次更新改用留出坐标，则任何读取整个留出场以及其他视图训练数据的估计器，其期望重叠度都不会保持非零。

部分 (a) 是针对短列表的测度变换：将一组 ηn锚点行中的真实配对替换为独立行后，在给定列表的条件下，剩余的对应关系变为均匀分布，因此短列表很少包含真实分配，而长列表的代价为 w^2 每个锚点行的访问次数。部分 (b) 是对跨越最小化割的双对象错误元组进行二阶矩计数，其违反概率为 $n^{-(2/3+o(1))c}$；它只涉及这一证书，而松弛本身对于 $2\log n < I_* \leq 3\log n$ 是否精确仍是开放问题。部分 (c) 源于将第一个字段与一个相关性过弱的双视图实验进行比较，也源于对错误评分的矩形界；对于每个非负列随机表，这些界均成立（附录 I）。在我们的模拟中，Sinkhorn软平均是有限规模下表现良好的解码器之一（附录 Q）；其默认的第三次迭代不在涵盖范围内，而在 $b < 1$ 时间内实现恢复仍是开放问题。部分 (a) 和部分 (b) 已在

附录 I；对于部分 (c)，附录陈述了证明所依据的三个引理，并概述了它们的证明。

7. 其他观测模型

迄今为止的结果都假设每个视图中观测到的高斯特征具有已知协方差。本节逐一放宽这些假设。对于非高斯的已知分布，以及在某些视图中缺失的对象，近乎精确恢复和成对阈值仍保持原有形式；未知的边缘协方差无妨，但未知的交叉协方差则不然；对于精确恢复，第 4 节和 5 节的两个必要条件对所有具有密度的分布均成立。

7.1. 已知观测分布

高斯性只通过真实元组信息密度的集中性影响定理1 和2 。设每个视图的特征取值于任意可测空间，且元组分布 P_n 已知。对于一组 B 视图和 B的一个划分 $\mathcal{C}$，令$D_{B,\mathcal{C}}$ 表示边缘分布 $P_{n,B}$ 与其各块边缘分布的乘积分布之间的散度；像之前一样，根据这些散度定义 J_B 和 J^*_{ab}；令 $L_{B,\mathcal{C}}$ 表示这两个分布相对于其和的密度之比的对数。允许散度以及$L_{B,\mathcal{C}}$ 取无穷值。

定理 23 (已知观测分布)。 定理 (2) 的界 *1(b)* 对每个已知分布 P_n均成立。进一步假设，对于每个 B 至少包含两个视图的集合、每个$\mathcal{C}$ 的 B 至少包含两个块的划分，以及每个固定 $\eta > 0$,

$$P_{n,B}\big(L_{B,\mathcal{C}} < \min\{D_{B,\mathcal{C}}, K\log n\} - \eta\log n\big) = o(1). \tag{4}$$

那么定理 *1(a)* 和定理 *2* 对 P_n成立，并且使用相同的解码器。

推论 *24*（固定坐标分布）。设每个元组具有已知分布 $p^{\otimes d}$，其中 p 是在 K 个可测空间的乘积上的固定分布，并且 $d/\log n \to c \in (0,\infty)$，用 p 替代 P_n定义 $J(p)$和 $J^*_{ab}(p)$。当 $cJ(p) > 1$时，可以同时实现近乎精确恢复；当 $cJ(p) < 1$时则不可能。对于视图 $a \neq b$，如果 $cJ^*_{ab}(p) > 1$，则存在一个置换 $\hat{\sigma}_{ab}$，使得 $\mathrm{ov}(\hat{\sigma}_{ab}) \to 1$ 依概率成立；如果 $cJ^*_{ab}(p) < 1$，则 $\sup_f \mathbb{E}\,\mathrm{ov}(f) \to 0$，如定理 *2(b)* 所述。

上限 $K\log n$in(4) 高于解码器进行测试的所有级别 $(1+\gamma)(|\mathcal{C}|-1)\log n$，因此该条件仅要求真实得分很少远低于其均值，且只需考察到该级别为止；高于上限的值（包括无穷大）无关紧要。散度有限时，由 $\mathrm{Var} L_{B,\mathcal{C}} = o(D^2_{B,\mathcal{C}})$ 和切比雪夫不等式即可得出；对于高斯分布，则可由引理 30 的方差界得出。固定 p 时，信息密度是 d 个独立项之和，因此（4）可由大数弱律得出；若散度为无穷大，则对截断后的项应用该定律即可，无需假设信息密度具有任何阶矩。同样的论证只用到 $d \to \infty$，所以若 $d/\log n \to \infty$，则可同时以近乎

当 $J(p) > 0$ 时，可以实现精确恢复；当 $J^*_{ab}(p) > 0$ 时，视图 a, b 可匹配；若 $d/\log n \to 0$ 且 $J(p) < \infty$，根据定理 23 的第一部分，近乎精确恢复是不可能的。精确恢复的情况则有所不同：对于非高斯分布，即使在弱依赖条件下，划分为三个块也可能成为瓶颈（例29），因此不存在以最弱二分划分为依据、类似推论 24 的结论；第 7.4 节给出了适用于一般分布的逆命题，以及弱依赖条件下的阈值。当分布随 n 变化时，近乎精确恢复需要某种类似条件（4）的条件。若每个视图都复制相同的 d 个独立伯努利（$1/n$）比特，那么对于固定的 $\kappa > 1$，只要 $d \geq \kappa n$，J 就会超过 $\log n$；但全零行（约占每个视图的 $e^{-d/n}$）无法匹配，而在这一族中，当且仅当 $d/n \to \infty$ 时，才能实现近乎精确恢复。证明以及对这一族的分析见附录 J。

7.2. 部分观测对象

假设对象 i 仅在一组随机视图 M_i 中被观测到，该视图组独立于特征，且从固定分布 μ 中抽取；每个视图都对其可见行进行均匀随机打乱，并且集合 M_i 是隐藏的；带有 $|M_i| = 1$ 的行是离群点，不属于任何元组。估计器输出若干行组，每组至多包含来自每个视图的一行。损失是指：被估计结果和真实分配恰有一方归入同一组的行对数量，因此离群点每与一行归为一组就产生一次损失。当每个对象都在所有视图中可见时，组估计的损失会限制其所诱导的排列在每个视图中的错位行数量；而若排列在所有视图中均有 M 个错位行，则其诱导的组的损失至多为 $2(K-1)M$，因此这两种准则等价。对于 $M \subseteq [K]$，令 J_M 表示视图 M 上边缘分布的共享信息。

命题 *25*（部分重叠）。在已知分布的高斯模型中，边缘分布固定时，若 $\min\{J_M : \mu(M) > 0, |M| \geq 2\} > \log n$，则损失为 $o_{\mathbb{P}}(n)$；若该最小值低于 $\log n$，则损失不为 $o_{\mathbb{P}}(n)$。对于类型为 M、满足 $\mu(M) > 0$ 且视图为 $a, b \in M$ 的对象，令 $J^*_{ab}(M) = \max\{J_B : \{a, b\} \subseteq B \subseteq M\}$。若 $J^*_{ab}(M) > \log n$，则同一解码器可以恢复对象对 a, b，且类型为 M 的对象中至多有 $o_{\mathbb{P}}(n)$ 个例外；若 $J^*_{ab}(M) < \log n$，则对这些对象，任何估计器都只能以消失的期望比例恢复该对象对。该阈值并非划分散度的 μ平均值：对于三个视图，每个视图以概率 $1/2$ 独立出现，且相关性相等 $1/2$，存在某些维数，使得自然平均值超过 $\log n$，在所有三个视图中都可见的对象能够近乎精确地恢复，而仅在某一对视图中可见的对象与任意一对视图之间的重叠度都趋于零。

解码器会对每一组至少包含两个视图的 K均值 $O_K(n^K)$候选元组执行分区检验，将共同属于某个接受的元组的任意两行连接起来，输出每个每个视图至多包含一行的连通组，并将其他组中的行逐行输出。每个被错误接受的元组最多影响 K 个对象（附录 K）。

7.3. 未知协方差

迄今为止，所有解码器都使用已知协方差。边缘分布的协方差可以通过少量行估计，但在信息阈值处，通常无法从打乱数据中学习未知的交叉协方差。

命题26（未知协方差）。 *(a)* 未知边缘分布。 设各视图的边缘协方差 S_a 未知，并且通过对每个视图使用 $S_a^{-1/2}$进行白化而获得的联合相关矩阵 $R \succ 0$已知。令 $d = \sum_a d_a$, $c_n = \lambda_{\min}$（R）以及 $u_n = (d + \log n)\ \log^2(2d)$。若 $u_n = o(nc_n)$ 且 $\liminf_n J/\log n > 1$，则可以同时实现近乎精确恢复：使用 o（n）保留行的样本协方差对白化每个视图，并运行划分检验。若边缘条件数有界，则 $u_n = d + \log n$ 即足够。

（*b*）未知噪声水平。 在 *Wee* 等人（*2026*年）的模型中，*Corollary* $4/K$ 和 $4/(K-1)$ 的阈值 *5* 在 τ 未知时仍成立。

(c) 未知正交互协方差。 设 $K = 2$，令两个视图的协方差均为恒等映射，交叉协方差为 ρO，其中 $O \in O(d_0)$ 未知，并固定 $\epsilon \in (0,2)$。若$n\rho^4 \le (2-\epsilon)\log n$，则对于任意估计器，对 *Haar*分布的 O取平均后，精确恢复的概率至多为 $\exp[-(\epsilon/4+o(1))n\log n]$。在相同的平均意义下，若 $n\rho^4 = o(\log n)$，则重叠度超过任意固定 $\alpha > 0$ 的概率趋于零。特别地，若 $d_0 \gg \sqrt{n\log n}$ 且 $\rho^2 = 2b\log n/d_0$，其中 $b > 1$ 固定，则$J \sim b\log n$，然而沿着某个矩阵序列 O，每个估计器的期望重叠度都趋于零。

部分 (c) 给出了充分条件中第二项所需的阶数 n

$$\rho^2 \ge C\max\left\{\frac{\log n}{d_0}, \sqrt{\frac{\log n}{n}}\right\}$$

根据Niu 等人（2026年，定理2.3），在完全观测高斯模型中。Huang 等人（2026年，第 1.1 节）报告称，对于观测信息更少的阈值随机几何图模型，也存在同阶的阈值。 (c) 的证明利用 Haar积分给出的似然比二阶矩上界，并通过正交群不变量理论第一基本定理对其进行控制（附录L）。

7.4. 非高斯分布的精确恢复

精确恢复的两种障碍——跨割交换和稀有亚群体——并不依赖高斯性。设每个对象的元组由 $d = d_n$ 个独立同分布坐标组成，其已知分布为 $P = P_n$，相对于 σ个有限测度的乘积分布具有密度 p。对于一个割 $S|S^c$，令 Z_S为核为 $\sqrt{p(x_S, x_{S^c})}$ 的算子，令 $I_2^\circ(S) = -\log \mathrm{tr}[(Z_S^* Z_S)^2]$,，并设 $\chi_* = \min_S I_2^\circ(S)$，对于高斯分布，该值等于 I_*。令 $R_{\mathcal{P}}$ 与定义 8 中相同，设

$$L(P) = \min\left\{\frac{\chi_*}{2}, \min_{|\mathcal{P}|\ge 3} R_{\mathcal{P}}(P)\right\}.$$

定理27 (一般性逆命题)。 设 $P_n = P$ 固定，或设 $d_n \to \infty$。若 $d_n L(P_n) \leq (1-\epsilon)\log n$ 对某个固定的 $\epsilon \in (0,1)$ 以及所有充分大的 n成立，则任意估计器恢复所有相对置换的概率趋于零。对于循环信息为 χ 的割，交换部分仍成立；该割可以随 n变化，只要 $d_n\chi \leq 2(1-\epsilon)\log n$ 且 $d_n(\chi+\chi^2) = o((\log n)^2)$，即使揭示割两侧内部的所有对应关系亦然。

定理28 （弱依赖）。 令 $U = \bigotimes_a P_a$ 为视图边缘分布的乘积分布，假设 $h_n = \|dP_n/dU - 1\|_\infty \to 0$，并令 $B(P_n) =$最小值$_{|\mathcal{P}|\geq 2} D_{\mathcal{P}}/|\mathcal{P}|$。在固定裕度下，若 $d_n B(P_n) >$对数$\,n$，则最大似然估计以趋近于 *1* 的概率恢复所有相对置换；而若 $d_n B(P_n) <$对数$\,n$，则任何估计器都会失败。此外，$L(P_n) = (1+O_K(h_n))\,B(P_n)$。

密度比的均匀界排除了高斯分布，因此定理28 是对定理3 的补充，而非扩展。在弱高斯区域中，$D_{\mathcal{P}}/|\mathcal{P}|$的一阶最小值在二分划分处取得，因为 $D_{\mathcal{P}} \geq (1+o(1))|\mathcal{P}|I_*/2$；对于其他分布则未必如此。

例29 （三视图奇偶性）。 设 $X_a \in \{-1,+1\}$和 $p_h(x_1、x_2、x_3) = 2^{-3}(1 + hx_1x_2x_3)$，其中 $h = h_n \to 0$。每个真边缘分布都是均匀的，因此划分为至少两个块的每个划分都有相同的散度 $D(h) \sim h^2/2$。因此 $B = D/3$，而每个二分划分都具有 D和 $\chi_* \sim D$。如果 $d_n D(h_n)/\log n \to c \in$（2、3），则 $d_n I_*$和 $d_n\chi_*$都以固定裕度超过 $2\log n$，然而精确恢复失败；对于 $c > 3$，最大似然则能够成功。

逆命题结合了定理 3（b）中的交换论证（这里既计算似然相同的情形，也计算严格改进的情形）与定理 9中的稀有子群体论证。弱依赖阈值源于对所有划分散度的二次近似、似然评分的联合集中，以及定理 3（a）中的两层计数（附录 M 和 N）。对于一般分布，是否 $d_n L(P_n) > (1+\epsilon)\log n$ 足够，仍是一个悬而未决的问题。

8. 讨论

阈值。 所有阈值均以固定裕量表述。围绕$J = \log n$和 $I_* = 2\log n$的临界窗口是开放问题，而命题 14 在我们研究的一个窗口中留下了阶为 $\log\log n$的差距。除弱相关性情形外，高斯分布的精确恢复仅在定理10 和13的情形下得到刻画；交换条件与稀有子总体条件是否对每个高斯分布共同构成充分条件（猜想 11）仍是开放问题。对于一般协方差，我们的证明所缺少的是：对含有许多短环的连通分量的高斯积分给出均匀上界。定理 13 的判据将其归结为针对每个协方差的有限个不等式，但这些不等式在接近奇异协方差时失效；而在等相关情形下，该猜想则通过不同的比较论证得到证明。这些证明中的裕量很小，在相关性 $-1/3$处低于百分之一，其中几项证明还借助了计算机。

计算。对于固定的 K，近乎精确恢复在通常意义下不存在统计-计算差距，因为划分检验可在多项式时间内运行；问题在于指数。在相等相关性情形下，分层筛选将其降至 $K = 3$ 的 2、4，但对于更大的 K，它只有在指数随 K 线性增长时才能达到 $4/K$。对于每个协方差，能否通过 J 次操作达到 $O(n^2\,\mathrm{poly}(d, K))$？在 $b = 1$ 以下，命题 22 表明，扫描成对候选列表需要 $n^{3-b-o(1)}$ 次测试；对于 $K = 5$，Sinkhorn软平均的前两轮会失败。这些结论针对的是特定解码器，而不是对所有二次时间算法的下界。当 K 增长时，低次多项式在 $b = (4+\epsilon)/K$ 处失败，但在每个固定的 $b > 0$ 处成功（注21）；因此，$b_n \to 0$ 的计算阈值仍是开放问题。同样开放的还有：当弱信号区间中 $2\log n < I_* \le 3\log n$ 时，$K \ge 3$ 能否在多项式时间内实现精确恢复；此时线性规划的半边对偶证书失效（命题 22(b)）。

其他模型。 未知的边缘协方差可以根据少量行进行估计，但未知的交叉协方差通常无法做到这一点（命题26）；未知正交交叉协方差的临界常数尚未确定。对于已知分布，在满足下尾条件（4）时，几乎精确阈值和成对阈值均成立，而固定坐标分布满足该条件（定理 23）。对于变化分布而言，该条件的必要性，以及对于既非高斯分布也非弱相关分布的分布，以 L 定理 27 中的速率实现精确恢复，均尚待研究。

模拟。 附录 Q 报告了使用 $n \le 8192$的模拟结果。在运行前记录的 33 项预测中，32 项得到了验证；唯一的例外是关于逐行求和的有限尺寸预测，附录对此作了解释。本文的保证是渐近性的：在这些规模下，分层筛选恢复的完整元组少于消息传递或 Sinkhorn 缩放，而我们对后两者没有任何保证。

人工智能助手的使用

本研究使用人工智能助手（Claude、Anthropic；Codex、OpenAI）进行文献检索、起草文本，以及独立检查证明、数值检验和写作。作者负责指导研究工作，核查每一条陈述和证明，并对内容承担全部责任。

参考文献

Mohammad Hassan Ahmad Yarandi 和 Luca Ganassali。带属性高斯图匹配的精确统计阈值，2026年。arXiv：2603.23305v2。

Serge Aleshin-Guendel 和 Mauricio Sadinle。用于记录链接和重复项检测的多文件划分。《美国统计学会会刊》，118(543):1786–1795，2023年。doi: 10.1080/01621459.2021.2013242。

Taha Ameen 和 Bruce Hajek。多个非齐次随机图的对齐：精确恢复的基本极限，2025年。arXiv:2405.12293v2；发表于《运筹学》，doi:10.1287/opre.2025.1808。Taha Ameen 和 Bruce Hajek。多个无标记高斯网络之间相关性的精确检测阈值，2026a。arXiv:2504.16279v2（首次发布于2025年）。Taha Ameen 和 Bruce Hajek。基于最后匹配的多个图对齐和相关性检测的逆命题界，2026b。arXiv:2608.14450v1。Chung Chan。关于多个终端秘密密钥容量的互依赖上界的紧致性，2008年。arXiv:0805.3200v2。Chung Chan、Ali Al-Bashabsheh、Javad B. Ebrahimi、Tarik Kaced 和 Tie Liu。受秘密密钥协商启发的多元互信息。*IEEE*学报，103 (10):1883–1913，2015年。doi: 10.1109/JPROC.2015.2458316。Chung Chan、Ali Al-Bashabsheh、Qiaoqiao Zhou、Tarik Kaced 和 Tie Liu。信息聚类：数据聚类的数学理论。*IEEE*分子、生物与多尺度通信汇刊，2(1):64–91，2016年。doi: 10.1109/TMBMC. 2016.2630054。arXiv:1605.01233。Imre Csisz´ar 和 Prakash Narayan。多个终端的保密容量。*IEEE* 信息论汇刊，50(12):3047–3061，2004年。doi: 10.1109/TIT.2004. 838380。Daniel Cullina 和 Negar Kiyavash。相关 Erd´os–R´enyi 图的精确对齐恢复，2017年。arXiv:1711.06783v2。Daniel Cullina、Prateek Mittal 和 Negar Kiyavash。数据库对齐的基本极限。在*IEEE*信息论国际研讨会（*ISIT*），第651–655页，2018年。doi: 10.1109/ISIT.2018.8437908。arXiv:1805.03829。William H. Cunningham。网络的最优攻击与加固。《*ACM*学报》，32(3):549–561，1985年。doi: 10.1145/3828.3829。Osman Emre Dai、Daniel Cullina 和 Negar Kiyavash。具有高斯特征的数据库对齐。在第*22*届人工智能与统计国际会议（*AISTATS*）论文集中，《机器学习研究论文集》第89卷，第3225–3233页，2019年。arXiv:1903.01422。Osman Emre Dai、Daniel Cullina 和 Negar Kiyavash。相关高斯数据库的近精确对齐的可实现性。在*IEEE*信息论国际研讨会（*ISIT*），第1230–1235页，2020年。doi: 10.1109/ISIT44484.2020.9174507。Osman Emre Dai、Daniel Cullina 和 Negar Kiyavash。高斯数据库对齐与高斯植入匹配，2023年。arXiv:2307.02459v1。

Jian Ding、YihongWu、Jiaming Xu 和 DanaYang。《植入匹配问题：尖锐阈值与无穷阶相变》。《概率论及相关领域》，187（1–2）：1–71，2023年。doi：10.1007/s00440-023-01208-6。arXiv：2103.09383。

Bertrand Even 和 Luca Ganassali。多重高斯图对齐中的统计-计算差距。载于第*29*届国际人工智能与统计学会议（*AISTATS*）论文集，《机器学习研究论文集》第300卷，第3133–3141页，2026年。arXiv：2512.00610。

Bertrand Even、Christophe Giraud 和 Nicolas Verzelen。基于置换的问题的计算障碍与弱相关随机变量的累积量。载于*2026*年*ACM-SIAM*离散算法年度研讨会（*SODA*）论文集，第493–562页，2026年。doi：10.1137/1.9781611978971.22。扩展版：arXiv：2507.07946。

Julia Gaudio、Mikl´os Z. R´acz 和 Anirudh Sridhar。相关随机块模型中的精确社区恢复。载于 第*35*届学习理论会议（*COLT*）论文集，《机器学习研究论文集》第178卷，第2183–2241页，2022年。

Dong Huang、Ankang Jiao、Shuangping Li 和 Zhangsong Li。随机几何图中的图匹配 I：高效算法。arXiv:2610.16675，2026年。

Zeynep K. 数据库对齐：基本极限与多数据库场景。博士论文，波士顿大学，2024年。网址：https://hdl.handle.net/2144/49260。

Zeynep K 和 Bobak Nazer。检测相关高斯数据库。载于*2022*年*IEEE*信息论国际研讨会（*ISIT*），第2064–2069页，2022年。doi：10.1109/ISIT50566.2022.9834731。

Dmitriy Kunisky 和 Jonathan Niles-Weed。几何植入匹配的强恢复。载于《*2022*年*ACM-SIAM*离散算法年度研讨会论文集（*SODA*）》，第834–876页，2022年。doi：10.1137/1.9781611977073.36。arXiv：2107.05567。Amos Lapidoth 和 Christoph Pfister。两种依赖性度量。《熵》，21(8)：778，2019年。Gustav I. Lehrer 和 Ruibin Zhang。正交群不变理论第二基本定理。《数学年刊》，176(3)：2031–2054，2012年。Zhangsong Li。利用低次多项式检测多元置换线性回归的计算转变。arXiv：2504.03097，2025年。Mehrdad Moharrami、Cristopher Moore 和 Jiaming Xu。植入匹配问题：相变与精确结果。《应用概率年刊》，31(6)：2663–2720，2021年。doi：10.1214/20-AAP1660。

Prakash Narayan 和 Himanshu Tyagi。通过公开讨论实现多终端保密性。《通信与信息理论基础与趋势》，13(2–3)：129–275，2016年。doi：10.1561/0100000072。

C. St. J. A. Nash-Williams。有限图的边不相交生成树。《伦敦数学会杂志》，s1-36(1)：445–450，1961年。doi：10.1112/jlms/s1-36.1.445。

Jonathan Niles-Weed 和 Ilias Zadik。稀疏张量主成分分析中的全有或全无现象。载于神经信息处理系统进展，第 33 卷，第 17674–17684 页。Curran Associates, Inc.，2020年。

Sirin Nitinawarat 和 Prakash Narayan。相关高斯源的密钥生成。*IEEE* 信息论汇刊，58(6)：3373–3391，2012年。doi：10.1109/TIT.2012.2184075。

Sirin Nitinawarat、Chunxuan Ye、Alexander Barg、Prakash Narayan 和 Alex Reznik。成对独立网络模型的密钥生成。*IEEE* 信息论汇刊，56(12)：6482–6489，2010年。doi：10.1109/TIT.2010.2081210。arXiv：1007.3518。

Xiaochun Niu、Tselil Schramm 和 Jiaming Xu。通过树计数进行高维Procrustes匹配。arXiv:2607.08538，2026年。

Mikl´os Z. R´acz 和 Jifan Zhang。利用多个相关网络实现精确社区恢复。载于神经信息处理系统进展，第37卷。Curran Associates, Inc.，2024年。

Ravi Kiran Raman 和 Lav R. Varshney。利用划分信息进行通用联合图像聚类与配准。载于*2017*年 *IEEE*信息论国际研讨会（*ISIT*），第2168–2172页，2017年。doi：10.1109/ISIT.2017.8006913。扩展版预印本：arXiv:1701.02776。

Galen Reeves、Jiaming Xu 和 Ilias Zadik。稀疏线性回归中的全有或全无现象。载于第三十二届学习理论会议论文集，《机器学习研究论文集》第99卷，第2652–2663页。PMLR，2019年。

Farhad Shirani、Siddharth Garg 和 Elza Erkip。基于测度集中的数据库去匿名化方法。载于*IEEE*信息论国际研讨会（*ISIT*），第2748–2752页，2019年。doi：10.1109/ISIT.2019.8849392。扩展版本：arXiv:1901.07655。

Richard Sinkhorn。任意正矩阵与双随机矩阵之间的关系。《数学统计年刊》，35(2):876–879，1964年。

Richard Sinkhorn 和 Paul Knopp。关于非负矩阵与双随机矩阵。《太平洋数学杂志》，21(2):343–348，1967年。

Marco Tomamichel 和 Masahito Hayashi。通过针对乘积分布和马尔可夫分布的复合假设检验，对 R´ényi 信息量进行操作性解释。*IEEE* 信息论汇刊，64(2)：1064–1082，2018年。

W. T. Tutte。关于将图分解为 n 连通因子的问题。《伦敦数学会杂志》，s1-36(1)：221–230，1961年。doi：10.1112/jlms/s1-36.1. 221。

Himanshu Tyagi 和 Shun Watanabe。密钥协商与安全计算的逆命题。*IEEE* 信息论汇刊，61(9)：4809–4827，2015年。

Tim van Erven 和 Peter Harremo¨es。R´ényi 散度与 Kullback–Leibler 散度。*IEEE* 信息论汇刊，60(7)：3797–3820，2014年。

Louis Vassaux 和 Laurent Massouli´e。《多图对齐的可行性：贝叶斯方法》。《应用概率进展》，第 1–40 页，2026年。doi：10.1017/ apr.2026.10067。首次发布。arXiv：2502.17142。

Haoyu Wang、Yihong Wu、Jiaming Xu 和 Israel Yolou。几何模型中的随机图匹配：完全图情形。载于 第 *35* 届学习理论会议论文集（*COLT*），第 178 卷，《机器学习研究论文集》，第 3441–3488 页，2022年。

Timothy L. H. Wee、Kaylee Y. Yang、Zhou Fan 和 Cheng Mao。高维中的几何植入匹配：多视图的力量，2026年。arXiv：2607.09026v1。

Hermann Weyl。《经典群：不变量与表示》。普林斯顿大学出版社，1939年。

Yihong Wu、Jiaming Xu 和 Sophie H. Yu。《随机图匹配的精确重构阈值》。*IEEE* 信息论汇刊，68(8):5391–5417，2022年。doi：10.1109/TIT.2022.3169005。

附录 A。定理 1 的证明

A.1. 划分的对数似然比

固定一个具有 m 个 $\mathcal{P}$ 的 $Q_{\mathcal{P}} = \bigotimes_{B\in\mathcal{P}} P_B$，并记作 $Q_{\mathcal{P}} = \bigotimes_{B\in\mathcal{P}} P_B$。将每个块限制在线性支撑 P_B 上，并对其进行白化。于是， $Q_{\mathcal{P}}$ 具有单位协方差，而 P 具有协方差C ，其对角块均为恒等映射。若 P 的支撑集是 $Q_{\mathcal{P}}$的真子空间，则 $D_{\mathcal{P}} = \infty$ 和 $Q_{\mathcal{P}}$ 使 P 的概率为零；随后，解码器使用检验“该元组位于 P的支撑集内”，真实元组几乎必然通过该检验

而具有恒等划分 $\mathcal{P}$ 的元组几乎必然无法通过。否则 $C \succ 0$；令 ξ_1、……、$\xi_N > 0$ 为其特征值。由于 $\operatorname{tr} C$ 等于维数，$\sum_j (\xi_j - 1) = 0$。

引理 30. 设 $L_{\mathcal{P}} = \log(dP/dQ_{\mathcal{P}})$ 且 $D_{\mathcal{P}} < \infty$。则 $\mathbb{E}_{Q_{\mathcal{P}}} e^{L_{\mathcal{P}}} = 1$, $\mathbb{E}_P L_{\mathcal{P}} = D_{\mathcal{P}}$ 以及方差 $_P L_{\mathcal{P}} \le 2m^2 D_{\mathcal{P}} \le 2K^2 D_{\mathcal{P}}$。

证明。 第一个恒等映射成立，因为 $P \ll Q_{\mathcal{P}}$。在白化坐标中

$$L_{\mathcal{P}}(x) = -\tfrac{1}{2} \log\det C - \tfrac{1}{2} x^\top (C^{-1} - I) x.$$

在 P 下，写作 $x = C^{1/2} g$，其中 g 为标准形式，并将 C 对角化，从而

$$L_{\mathcal{P}} = -\tfrac{1}{2} \sum_j \log \xi_j + \tfrac{1}{2} \sum_j (\xi_j - 1) Z_j^2,$$

其中 Z_j 为独立同分布的标准正态变量。由于 $\sum_j (\xi_j - 1) = 0$,

$$\mathbb{E}_P L_{\mathcal{P}} = -\tfrac{1}{2} \sum_j \log \xi_j = \tfrac{1}{2} \sum_j (\xi_j - 1 - \log \xi_j) = D_{\mathcal{P}}, \qquad \operatorname{Var}_P L_{\mathcal{P}} = \tfrac{1}{2} \sum_j (\xi_j - 1)^2.$$

对于任意块向量 z，若 $W = C^{1/2} g$，则由柯西–施瓦茨不等式可得

$$z^\top C z = \mathbb{E}\Big(\sum_B z_B^\top W_B \Big)^2 \le m \sum_B \mathbb{E}(z_B^\top W_B)^2 = m \|z\|^2,$$

因此 $\xi_j \le m$。函数 $u - 1 - \log u$ 在 $u = 1$ 处及其导数均为零，并且在 $(0, m]$ 上的二阶导数为 $u^{-2} \ge m^{-2}$。因此 $\xi - 1 - \log \xi \ge (\xi - 1)^2/(2m^2)$，所以 $\sum_j (\xi_j - 1)^2 \le 4m^2 D_{\mathcal{P}}$，方差的界由此得出。 □

A.2. 可实现性

选取 $\gamma > 0$，使得对于每个划分以及所有充分大的 n，均有 $D_{\mathcal{P}} \ge (1 + 2\gamma)(|\mathcal{P}| - 1) \log n$。对于每个锚定行 i 以及每个 $t = (i, j_2, \ldots, j_K) \in \{i\} \times [n]^{K-1}$，令 $x(t) = (Y_i^1, Y_{j_2}^2, \ldots, Y_{j_K}^K)$。若 $x(t)$ 属于 P 的支撑集，且对于所有满足 $D_{\mathcal{P}} < \infty$ 的 $\mathcal{P}$，均有 $L_{\mathcal{P}}(x(t)) > (1 + \gamma)(|\mathcal{P}| - 1) \log n$，则接受 t。每行提出一个接受的元组（如果有），否则弃权。在每个视图 $a \ge 2$ 中分别处理：对于提出相同列的行，保留其中一行；将未匹配到列的行与未使用的列通过任意双射进行匹配。

真实元组。 根据引理 30，$L_{\mathcal{P}}$ 的均值至少比阈值高 $\gamma D_{\mathcal{P}}/(1 + 2\gamma)$。切比雪夫不等式给出的真实元组未通过 $\mathcal{P}$-检验的概率上界为 $2K^2(1 + 2\gamma)^2/(\gamma^2 D_{\mathcal{P}}) = O(1/\text{对数 } n)$。至多有 B_K 个划分，其中 B_K 是贝尔数。

错误元组。 固定 Π 和一个锚定行。恒等划分为 $\mathcal{P}$ 的元组具有分布 $Q_{\mathcal{P}}$。如果 $D_{\mathcal{P}} = \infty$，则它几乎必然被拒绝；否则，根据引理 30 和马尔可夫不等式可得 $\mathbb{P}(L_{\mathcal{P}} > (1 + \gamma)(|\mathcal{P}| - 1) \log n) \le n^{-(1+\gamma)(|\mathcal{P}|-1)}$。给定 Π，这样的元组由不包含锚点的 $|\mathcal{P}| - 1$ 块中的对象确定，因此其数量至多为 $n^{|\mathcal{P}|-1}$

个。由于 $\mathbf{1}_{\cup E_t} \le \sum \mathbf{1}_{E_t}$ 逐点成立，某个属于该行的错误元组被接受的概率至多为 $\sum_{\mathcal{P}} n^{-\gamma(|\mathcal{P}|-1)} \le B_K n^{-\gamma}$。

补全。如果某一行的真实元组被接受且没有错误元组被接受，则称该行为良好，并令$\mathcal{B}_n$ 表示坏行的数量；则有 $\mathbb{E}\mathcal{B}_n = o(n)$。良好行提出其真实列，而这些列彼此不同。在固定视图中，良好行只有可能被坏行夺走其列，且每个坏行至多占用一列。因此，对于所有 $a \ge 2$ 可同时有 $M_a \le 2\mathcal{B}_n$，由马尔可夫不等式即可完成证明。

裕量。证明仅通过 $D_{\mathcal{P}} - (1+\gamma)(|\mathcal{P}|-1)\log n \ge c\, D_{\mathcal{P}}$和 $B_K n^{-\gamma} \to 0$使用了 γ，其中常数为 $c > 0$。对于任意满足 $\gamma_n \log n \to \infty$的 $\gamma_n \to 0$，两者均成立；例如，当 $\liminf J/\log n > 1$时，$\gamma_n = (\log n)^{-1/2}$即可。因此，解码器无需知道 J 超过 $\log n$多少。按照附录 Q，采用带有固定加性裕度 $\eta > 0$的阈值 $(|\mathcal{P}|-1+\eta)\log n$ 会移动阈值。当对每个 $\mathcal{P}$均有 $D_{\mathcal{P}} \ge (1+\epsilon)(|\mathcal{P}|-1+\eta)\log n$时，上述证明仍然成立。另一方面，若对于某个 $\mathcal{P}$有 $D_{\mathcal{P}} \le (1-\epsilon)(|\mathcal{P}|-1+\eta)\log n$，则引理 30和切比雪夫不等式表明，真实元组未通过 $\mathcal{P}$-检验的概率趋于 1，而每行接受错误元组的概率至多为 $B_K n^{-\eta}$，因此提出提议的锚点行的期望数量为 $o(n)$。

假设在接受的元组中进行选择时，对行重标记具有等变性。例如，该行可以像 G6 中那样最大化单例划分对数似然比，并在需要时均匀打破平局。如果冲突消解和补全仅使用原始提议和独立随机性，如附录 Q中 G4 按指标顺序进行的补全，那么定理 15(ii) 的证明中的补全论证表明，最终排列的期望重叠度趋于零。重读数据的补全可以做得更好（注释 37）。对于相关性相等的三个视图，在弱极限下，双块检验需要 $b > 1+\eta$，而单例检验需要 $b > 2(2+\eta)/3$，因此级别为 $\max\{1+\eta,\, 2(2+\eta)/3\}$，也就是说，对于 $41/30$，有 $\eta = 0.05$。

A.3. 逆命题

令 $\mathcal{P} = \{B_1, \dots, B_m\}, 1 \in B_1$，其代表元分别为 $r_j = r_{B_j}$和 $r_1 = 1$。对于 $a \in B_j$，记 $G_a = \pi_a \circ \pi_{r_j}^{-1}$。从 $(\pi_a)_{a \ge 2}$到 $\big((G_a)_{a \notin \{r_1,\dots,r_m\}},\ (\pi_{r_j})_{j \ge 2}\big)$ 的映射是有限群之间的双射。因此，在均匀先验下，$U = (\pi_{r_2}, \dots, \pi_{r_m})$ 在 $\mathfrak{S}_n^{m-1}$ 上均匀分布，且与 $G = (G_a)$ 独立。将 G 提供给统计学家。已知 G 后，按 G_a^{-1} 重新排列每个视图 $a \in B_j$，使得分块数据库 Y'_j 的第 k 行为 $(Y^a_{G_a(k)})_{a \in B_j}$。如果 $k = \pi_{r_j}(i)$，那么这就是 $X_i^{B_j}$。由于 $(Y',\ G)$ 可以确定 Y，因此在这个扩大的实验中可以运行任何估计器。

设 $Q = \bigotimes_j P_{B_j}^{\otimes n}$，在此条件下，块数据库相互独立且具有独立同分布的行。对于每个 u，按 u 重新排列行后，给定 $U = u$ 和 G 时，Y' 的分布由来自 P 的 n 个独立同分布元组构成，而 Q 在这种重新排列下保持不变。因此 $D(P_{Y'|u,\ G} \| Q) = nD_{\mathcal{P}}$。对 u取平均，

$$I(U; Y' \mid G) + D(P_{Y'|G} \| Q) = nD_{\mathcal{P}}, \qquad \text{so} \qquad I(U; Y' \mid G) \le nD_{\mathcal{P}}.$$

对于一个固定的映射 $\hat{}u : [n] \to [n]$，至多有 $\binom{n}{t} t! \le n^t$ 个排列与 $\hat{}u$ 恰好在 t个位置上不同：先选择这些位置，随后其他位置的像便确定了，并且

剩余的 t 行只能使用剩余的 t 列。先对每个错误计数进行编码，然后再对置换进行编码，

$$H(U \mid \widehat{U}, G) \le (m-1)\log(n+1) + \log n \sum_{j \ge 2} \mathbb{E} M_{r_j}.$$

估计器自身的随机性与其他一切因素相互独立，因此根据数据处理可得 $\leadsto I(U;\, U \mid G)$ $\le I(U;\, Y' \mid G) \le nD_{\mathcal{P}}$。由于 $H(U \mid G) = (m-1)\log n!$，$\log n! \ge n\log n - n$ 且 $\log(n+1) \le n$，我们得到 (2)。若 $\limsup J/\log n < 1$，则对每个 n的最小化划分应用 (2)。此时右侧有一个正的常数下界；与此同时，近乎精确恢复会迫使有界比率 M_a/n 以及它们的期望趋于零。 □

附录 B. 定理2的证明

引理 31 (从映射到排列)。 若某个估计器 $f\colon [n] \to [n]$ 满足 $\mathbb{E}\,\mathrm{ov}(f) = q$，$\tilde{}$ $\tilde{}$，则存在一个随机排列值估计器 f，满足 $\mathbb{E}\,\mathrm{ov}(f) \ge q^2$。

证明。 对于每个列 j 及其 $k_j \ge 1$ 个原像，均匀随机地选取一个原像并保留。然后用任意双射匹配其余行和列。令 c 为 f 正确的行数， H 为这些行所对应列的集合。每个 $j \in H$ 恰有一个正确的原像，且 $|H| = c$。保留的正确行数的期望为 $\sum_{j \in H} k_j^{-1} \ge c^2 / \sum_{j \in H} k_j \ge c^2/n$。取期望并利用 $\mathbb{E}c^2 \ge (\mathbb{E}c)^2$ 即可证明该结论。

□

定理2中全有或全无结论的证明。 假设该结论不成立。根据引理31，存在 $u > 0$，以及 f 的子序列和排列估计器 σ_{ab}，满足 $\mathbb{E}\,\mathrm{ov}(f) \ge u$。划分的数量有限，因此我们可以进一步取一个子序列，使固定的最小化划分 $\mathcal{P}^* = \{B_1, \ldots, B_m\}$ 将 a 与 b 分开，并且 $J/\log n \to j_0 < 1$。令 $g = 1 - j_0$。将视图 a 作为锚点，其中 $a \in B_1$ 和 $b \in B_2$分别作为其所属块的代表元。

折叠。 揭示块内对应关系 G，如逆命题中所述。这会将 K 视图转变为 m 超视图，其超元组分布 P 按 $\mathcal{P}^*$分组，且未知的逐项置换 $\Pi = (\pi_{r_2}, \ldots\ldots, \pi_{r_m})$ 是独立同分布的均匀随机变量。如附录 A 中所述， G 独立于 Π 以及折叠数据；其行仅使用 Π进行打乱。折叠数据和 G 决定了原始观测。将超视图划分为 $\mathcal{C}$ 若干 h 块$^{-}$会诱导出 $\mathcal{C}$ 的粗化 $\mathcal{P}^*$，因此

$$D^{\circ}_{\mathcal{C}} = D_{\bar{\mathcal{C}}} \ge (h-1)J, \qquad D^{\circ}_{\mathrm{sing}} = D_{\mathcal{P}^*} = (m-1)J. \tag{5}$$

这里， D° 指折叠的旧数据，sing 指将超视图划分为单元素集合。由于 $D^{\circ}_{\mathrm{sing}} < \infty$，超元组在每个超视图上选取适当坐标后具有正定协方差。

增广。 固定 $\varepsilon > 0$ 并放置 $H_n = (1+\varepsilon)$ 对数 $n - J > 0$ 和 $L_n = 4H_n/m$。向每个超视图添加新坐标 $\ell = 1,\ 2$、..... ，由相同的行承载。在坐标

ℓ 对象的 m 超视图服从分布 $N(0, R_m(s_\ell))$，在对象、坐标和旧数据之间相互独立。取 $s_\ell = r_n = (\log n)^{-1/2}$ 个坐标，再增加一个坐标吸收余项，使得 $\sum_\ell s_\ell^2 = L_n$ 且 $\max_\ell s_\ell \to 0$。去掉新坐标即可恢复旧实验，因此 f 仍具有 $\mathbb{E}\,\mathrm{ov}(f) \ge u$。对于一个划分 $\mathcal{C}$，其中有 h 个块以及 $e(\mathcal{C})$ 个分离的超视图对，将 $\log\det R_v(s) = (v-1)\log(1-s) + \log(1+(v-1)s)$ 展开可得 $D_{\mathcal{C}}^{+} = \frac{1}{2} e(\mathcal{C}) L_n + O_m(\max_\ell s_\ell L_n)$。将元素从较小的块移至较大的块表明 $e(\mathcal{C}) \ge (h-1)(m - h/2) \ge (h-1)m/2$，当块为单元素集合时取等号。散度可在独立坐标上相加，因此由 (5)

$$\frac{D_{\mathcal{C}}^{\circ} + D_{\mathcal{C}}^{+}}{h-1} \ge (1+\varepsilon)\log n - o(\log n), \qquad D_{\mathrm{sing}}^{\circ} + D_{\mathrm{sing}}^{+} = (m-1)(1+\varepsilon)\log n + o(\log n).$$

根据定理 1（a）应用于 m 超视图，某个估计器从增广数据中恢复 Π，其 Y 满足 $\mathbb{E} M_j = o(n)$ 对每个 j 均成立。

上界。 设 P_{mix} 为 Y 的分布，Q 为这样的分布：超视图数据库相互独立，且其行分别从边缘分布中独立同分布地生成。与逆命题中一样，$I(\Pi; Y) + D(P_{\mathrm{mix}} \| Q) = n(D_{\mathrm{sing}}^{\circ} + D_{\mathrm{sing}}^{+})$。此外，$H(\Pi \mid Y) \le (m-1)\log(n+1) + \log n \sum_j \mathbb{E} M_j = o(n \log n)$，且 $H(\Pi) = (m-1)\log n!$ 因此

$$\limsup D(P_{\mathrm{mix}} \| Q) / (n \log n) \le (m-1)\varepsilon. \tag{6}$$

这一步使用了全部 $m-1$ 个排列的联合熵。仅恢复指定对并不能抵消首项。

下界。 将 P_{mix} 和 Q 分别扩展为加入 G 的独立副本以及估计器的独立种子，并使两边具有相同的分布。这样不会改变它们的散度。设 W^1，W^2 为超视图 1 和 2 的新坐标，并且

$$T = \sum_{\ell} s_\ell \sum_{k} W^1_{k\ell} W^2_{f(k)\ell},$$

其中，f 使用通过 G 及其种子重构的旧数据。给定 Π、旧数据、G 和种子，当 $f(k) = \sigma_{ab}(k)$ 时，每个乘积的均值为 s_ℓ，否则为 0，因此 $\mathbb{E}_{P_{\mathrm{mix}}} T = n\,\mathbb{E}\,\mathrm{ov}(f) \ge n u L_n$。在 Q 下，给定旧数据、G 和种子，新坐标是相互独立的标准正态变量。由于 f 是一个双射，每个坐标中的 n 个乘积相互独立。对于相互独立的标准正态变量 $\mathbb{E} e^{v Z_1 Z_2} = (1-v^2)^{-1/2}$（$Z_1$，$Z_2$ 和 $|v| < 1$），

$$\log \mathbb{E}_Q e^{tT} = -\frac{n}{2}\sum_{\ell} \log(1 - t^2 s_\ell^2) = \frac{n}{2} t^2 L_n + O\Big(n t^4 \sum_{\ell} s_\ell^4\Big) = \frac{n}{2} t^2 L_n + o(n \log n).$$

Donsker–Varadhan 不等式 $D(P_{\mathrm{mix}} \| Q) \ge t\,\mathbb{E}_{P_{\mathrm{mix}}} T - \log \mathbb{E}_Q e^{tT}$，结合 $t = u$ 可得

$$\liminf D(P_{\mathrm{mix}} \| Q) / (n \log n) \ge \tfrac{2u^2}{m}(1 + \varepsilon - j_0) \ge \tfrac{2u^2}{m} g. \tag{7}$$

结论。 常数 u、g 和 m 在选择 ε 之前是固定的。取 $\varepsilon < 2u^2 g / (m(m-1))$ 会使（6）和（7）不相容。

如果 $[K]$ 的单元素划分达到最小值，那么它会分离每一对元素。 □

定理 2 的证明。部分 *(a)*。在一个达到 B 的集合内应用定理 J^{*}_{ab}，并将从其锚点到 a 和 b 的估计复合起来。复合后的置换的错误至多为各自错误之和。如果 B 随 n 变化，则在有限多个最大化集合固定的子序列上应用此论证。

最弱划分。假设 $S \subseteq B$ 与 $v \geq 2$ 个由 $\mathcal{C}_B$ 划分成的最弱划分块相交，其中共有 m_B 个块且 J_B 有限。合并被 v 相交的块。根据链式法则，散度会减少这些完整块之间的全相关。粗化散度至少为 $(m_B - v)J_B$，当 $v = m_B$ 时取值为零，因此减少量至多为 $(v-1)J_B$。将 v 个块投影到它们与 S 的交集，只会降低其全相关。因此， S 的诱导划分给出 $J_S \leq J_B$。特别地，当 $[K]$ 的最弱划分将 a、b 分开，且 J 为亚临界时，任何包含这对元素的集合都有 $J_S \leq J$；部分 (b) 随即给出最终结论。在固定级别 c 上，同一性质表明，对满足 $J_B \leq c$ 的节点进行递归分裂，会生成 c 以上共享信息的包含极大集合，以及任何剩余的单元素集合。

部分 (b)：树与增广。附录 J.3 对一般已知分布采用相同论证，并以共享均匀标签替代此处使用的高斯层。选择$\delta > 0$，使得最终有 $J^{*}_{ab} \leq (1-2\delta)\log n$。从 $[K]$开始，按照最弱划分 $\mathcal{C}_B$，将每个满足 $J_B \leq (1-\delta)\log n$ 的非单元素B划分。对于固定的 K，这样的树只有有限棵，因此在任何重叠度为正的子序列上，我们都可以进一步取一个树固定的子序列。每个同时包含 a和b 的节点都会被划分，因此叶划分 $\mathcal{L}$ 将它们分开。按照优先顺序a， b，选择每个节点的第一个视图，并将其他视图作为其代表。揭示叶内对齐 G，并将每个叶节点与其代表对齐。以 a 所在的叶节点为锚点，其他叶节点代表的 $r-1$ 置换 $\Pi_{\mathcal{L}}$ 是相互独立的均匀置换，且独立于 G。这是因为将原始置换改写为叶内对齐和代表置换，这种改写构成一个双射。对齐后的叶节点与 G 决定了原始观测。

在每个内部节点 B，向$m_B = |\mathcal{C}_B|$ 子节点代表元添加一层独立的高斯坐标，其行排列与原有特征相同。在坐标 ℓ 中，单个对象的元组具有协方差 $R_{m_B}(s_\ell)$，我们选择

$$L_B = \sum_{\ell} s_\ell^2 = \frac{4}{m_B}\big((1+\varepsilon)\log n - J_B\big), \qquad 0 \leq s_\ell \leq (\log n)^{-1/2},$$

对于一个待选定的固定 $\varepsilon > 0$ 。最大相关性的多个副本和最后一个较小的坐标实现了这个和。打乱之前，新层彼此独立，且与旧数据独立。如果子节点内部已对齐，那么按粗化原理，将它们的超视图划分为 v 组时，旧散度至少为 $(v-1)J_B$ 。假设该划分切割了 e 对子节点。附录 C 中使用的等相关行列式展开给出了额外散度 $eL_B/2 + O_K(\max_\ell s_\ell L_B)$。将 v 组分割时，最小割会把 $m_B - v + 1$ 个子节点放在同一组，其余各组各放一个，因此 $e \geq (v-1)m_B/2$。所以，每个归一化散度至少为 $(1+\varepsilon)\log n - o(\log n)$。对于划分为所有 m_B 个单元素集合的情况，合并后的散度为 $(m_B - 1)(1+\varepsilon)\log n + o(\log n)$。由于 K 固定，所有余项都是一致的。

增强实验的熵。令 Y_{aug} 表示打乱的增强叶数据库，且 G。设 P_{mix} 为它们的分布， Q 为叶数据库边缘分布的乘积，并取 G 的一个独立副本。每个增强层的子节点代表分别位于不同的叶节点中。根据 KL 链式法则，旧叶节点散度沿树向下逐级展开，在每个内部节点的贡献为 $(m_B-1)J_B$。特别地，即使旧分布退化，该散度仍然有限。由于有限性意味着联合支撑集等于各叶节点支撑集的乘积，因此可以将每个叶节点投影到其高斯支撑集上。加入新层并使用 $\sum_B (m_B-1) = r-1$，可得单个对象的散度

$$D_{\mathcal{L}}^{\mathrm{aug}} = (r-1)(1+\varepsilon)\log n + o(\log n).$$

以叶节点排列为条件，撤销这些排列后得到 n 个独立的真实元组，且叶节点保持 Q 不变。因此

$$D(P_{\mathrm{mix}}\|Q) + I(\Pi_{\mathcal{L}}; Y_{\mathrm{aug}}) = n D_{\mathcal{L}}^{\mathrm{aug}}. \tag{8}$$

将 $\Pi_{\mathcal{L}}$ 双射地改写为每个内部节点处子节点代表之间的相对置换。按先列出后代的顺序展开其条件熵。对于 B 处的项，在给定较早出现的置换后，可得到其子节点内部的真实对齐。定理 1(a) 适用于它们的旧坐标以及 B 处的新层。解码器忽略所有其他新坐标，包括祖先添加到B中代表元的坐标。由此得到的估计器对每个子节点的期望错误为 $o(n)$。给定一个估计及其错误数 M，至多有 n^M 个可能的真实置换。因此，每个节点对条件熵的贡献至多为 $(m_B-1)\log(n+1) + o(n\log n)$。对各项求和，并在 (8) 中利用 $H(\Pi_{\mathcal{L}}) = (r-1)\log(n!)$，得到

$$\frac{D(P_{\mathrm{mix}}\|Q)}{n\log n} \le (r-1)\varepsilon + o(1). \tag{9}$$

正重叠会导致更大的散度。 如果使用旧观测的估计器在某个子序列上具有至少 $\eta > 0$ 的期望重叠度，引理31 会将其转化为置换估计 $\widehat{f}$，其期望重叠度至少为 $u = \eta^2$。该引理中的柯西–施瓦茨不等式步骤为 $C^2 \le (\sum_{j\in H} k_j)(\sum_{j\in H} 1/k_j) \le n\sum_{j\in H} 1/k_j$，适用于 C 正确的行；将保留数量除以 n后，詹森不等式给出所述的期望重叠度。在同时包含 B_0 和 a 的最低节点 b处，这些视图代表不同的子节点。令 W^a， W^b 其在 B_0处的新坐标，设

$$T = \sum_{\ell} s_\ell \sum_{i} W^a_{i\ell} W^b_{\widehat{f}(i),\ell}.$$

在 P_{mix} 下，给定旧观测、真实排列和估计器的种子后，每对正确匹配的均值为 L_{B_0}。因此 $\mathbb{E}_{P_{\mathrm{mix}}} T \ge n u L_{B_0}$。在 Q 下，给定旧观测和独立种子后，由于 f 是双射，乘积相互独立，因此对于固定的 t，

$$\log \mathbb{E}_Q e^{tT} = -\frac{n}{2}\sum_{\ell} \log(1 - t^2 s_\ell^2) = \tfrac{n}{2} t^2 L_{B_0} + o(n\log n).$$

将相同的独立种子附加到两个实验中，不会改变它们的散度。现在利用 $t=u$ 的熵不等式可得

$$\frac{D(P_{\mathrm{mix}}\|Q)}{n\log n}\geq\frac{u^2L_{B_0}}{2\log n}-o(1)\geq\frac{2u^2\delta}{K}-o(1).$$

选择 $0<\varepsilon<2u^2\delta/[K(K-1)]$ 以反驳（9）。沿子序列选取近乎最优的估计量，即可确立上确界结论。 □

附录 C. 推论的证明5 和命题6

推论 5 的证明。该线性映射可逆，且与行置换可交换，因此不会改变问题。令 $F_v=-\frac12\sum_\ell$ log det $R_v(\rho_\ell)$ 对于 $v\le K$，使得 $F_1=0$和 $D_{\mathcal P}=-F_K-\sum_{B\in\mathcal P}F_{|B|}$。增量 $F_v-F_{v-1}=-\frac12\sum_\ell\big[\log(1-\rho_\ell)+\log{}^{1+}{}_{(v-1)1+\ v-2}{}^{\rho_\ell}(\,)_{\rho_\ell}\big]$ 关于 v单调不减。对于 p 块，将一个~~元素从~~较小的块移至较大的块不会使 $\sum_B F_{|B|}$减小，因此 $\sum_B F_{|B|}\le F_{K-p+1}$。最后 $p-1$ 个增量的平均值至少等于全部 $K-1$ 个增量的平均值，因此 $F_K-F_{K-p+1}\ge(p-1)F_K/(K-1)$。由此，$D_{\mathcal P}/(p-1)\ge F_K/(K-1)$，且在单元素集合的情形下取等号。当 $(K-1)\rho_{\max}\le1/2$时，泰勒展开给出 $F_v=\frac12\binom{v}{2}A+O_K(\rho_{\max}A)$，因此 $J=F_K/(K-1)=(K/4+O_K(\rho_{\max}))A$。阈值和全有或全无结论由定理 1 和定理 2 得出。

对于两侧大小分别为 s 和 $K-s$ 的一个割，

$$F_K-F_s-F_{K-s}=\tfrac12 s(K-s)A+O_K(\rho_{\max}A).$$

对 $1\le s<K$ 取最小值得 $I_*=((K-1)/2+O_K(\rho_{\max}))A$。在这些坐标下，非对角联合相关矩阵的算子范数为 $(K-1)\rho_{\max}$，并趋于零。因此，定理 3 给出了固定裕度下的精确阈值 $A/\log n=4/(K-1)$。

在 Wee 等人（2026）的模型中，将每个视图除以 $(1+\tau^2)^{1/2}$ 后，每个坐标都有 $\rho=(1+\tau^2)^{-1}$。如果 $\tau^4=d/(b\log n)$ 且 $d/\log n\to\infty$，那么 $\tau\to\infty$ 和 $A/\log n=d\rho^2/\log n=b(\tau^2/(1+\tau^2))^2\to b$。 □

命题 6 的证明。写作 $R=I+E$，其中 E 具有零对角块，且 $\|E\|_{\mathrm{op}}=h$。对于一个划分 $\mathcal P$，令 $B_{\mathcal P}$ 为 R 沿 $\mathcal P$、$F=R-B_{\mathcal P}$ 和 $W=B_{\mathcal P}^{-1/2}FB_{\mathcal P}^{-1/2}$的分块对角部分。则 $D_{\mathcal P}=-\frac12\log\det(I+W)$、tr $W=0$和 $\|W\|_{\mathrm{op}}=O(h)$，因为 $B_{\mathcal P}$的谱包含于 $[1-h,1+h]$。将 $-\log(1+x)+x=x^2/2+O(x^3)$ 按 W的特征值展开，并利用 $|\mathrm{tr}\ W^j|\le\|W\|_{\mathrm{op}}^{j-2}$tr W^2，可得 $D_{\mathcal P}=\frac14(1+O(h))\|W\|_F^2$。由 $B_{\mathcal P}^{-1/2}$奇异值的界可得$\|W\|_F^2=(1+O(h))\|F\|_F^2=(2+O(h))\sum_{a<b}$分离的 $\|R_{ab}\|_F^2$。最后，$I_{ab}=-\frac12\log\det(I-R_{ab}R_{ab}^{\top})=\frac12(1+O(h^2))\|R_{ab}\|_F^2$。所有常数仅取决于 K，误差是相对于所讨论的割而言的。除以 $|\mathcal P|-1$ 并取最小值，得到 $J=(1+O(h))s(I)$。零权重的割在两侧均有 $D_{\mathcal P}=0$。

对于二分划分 $\{S, S^c\}$, $D_{\{S,S^c\}} = I(X^S;\, X^{S^c})$。对每个二分划分应用相同的相对近似，可得

$$(1 - C_K h)\min_S w(S, S^c) \;\le\; I_* \;\le\; (1 + C_K h)\min_S w(S, S^c), \qquad w(S, S^c) = \sum_{a\in S,\, b\notin S} I_{ab}.$$

这也涵盖最小割为零的情形。对于下界，在所有割上取下界估计的最小值。对于上界，在最小权割处计算上界估计。

Tutte–Nash-Williams 定理将强度与分数树打包等同起来。对于整数权重，存在 k 棵边不相交生成树，当且仅当每个划分为 p 个块的划分都至少被 $k(p-1)$ 条边穿过。将此结论应用于乘以公分母后的有理权重，让缩放因子趋于无穷大，再通过单调近似推广到实权重。

□

附录 D. 定理证明3

D.1. 弱相关性情形下的可实现性

按丢失的真实边对候选项分组。对每个视图进行白化，并以真实排列为条件。重新标记各行，使得每个视图中的真实分配均为恒等映射。将候选项写作 $\tau = (\mathrm{id}\ \tau_2,\ \ldots\ldots,\ \tau_K)$，其中 $\tau_a(i)$ 是候选元组 i 中视图 a 所使用的源行。根据注 16，高斯对数似然中依赖于 τ 的部分为

$$\sum_{i=1}^{n}\sum_{a<b}(X^a_{\tau_a(i)})^\top W_{ab} X^b_{\tau_b(i)}, \qquad W_{ab} = -(R^{-1})_{ab}.$$

此处及下文中，X 表示白化特征。对于所有充分大的 n，协方差 R 均为正定，因为 $h = \|R - I\|_{\mathrm{op}} \to 0$。

对于每个源行 i，若两个视图中该行的副本都对应同一个候选元组，则将这两个视图归为一组；等价地说，它们的 $\tau_a^{-1}(i)$ 值相同。称 $H = (\mathcal{P}_i)_{i=1}^n$ 为 τ的模式。令

$$U = \{i : |\mathcal{P}_i| \ge 2\}, \qquad m = |U|, \qquad Q = \sum_{i\in U}(|\mathcal{P}_i| - 1).$$

非恒等置换至少移动两个源行，而这些行必须与锚定视图划分开，因此错误候选项具有 $m \ge 2$。此外，$m \le Q \le (K-1)m$。具有 m 个拆分行的模式数量至多为 $\binom{n}{m} B_K^m$，其中 B_K为贝尔数。对于固定模式，每个逆置换都固定 U 之外的每一行，因此将 U 映射到其自身。在一个拆分行 i中，包含视图 1 的块对应元组 i，而其他每个块至多可对应 m 个元组。这些选择确定了逆置换，因此模式 H至多有m^Q 个候选项。

令 T_H 表示候选项舍弃的真实边得分之和，即对于所有 $i \in U$，$\mathcal{P}_i$ 中不同块的视图之间的得分。令 F_τ 表示虚假

边得分之和，这些边由 τ 引入。与真实分配相比，得分差为 $F_\tau - T_H$。对于每对视图，丢失的真实边数量等于新增的虚假边数量，因为两种匹配方案都恰好使用 n 条边。每条虚假边连接的是独立源行，因此 $\mathbb{E}F_\tau = 0$。

联合矩界。对于划分 $\mathcal{P}$，令 $c_R(\mathcal{P}) = \sum_{a<b \text{ 由 } \mathcal{P}} \|R_{ab}\|_F^2$ 分隔。对于模式 H，定义

$$\begin{aligned} A_H &= \sum_{i \in U} c_R(\mathcal{P}_i), \\ V_H^0 &= \sum_{i \in U} \sum_{a<b \text{ separated by } \mathcal{P}_i} \|W_{ab}\|_F^2, \\ \mu_H &= \mathbb{E}T_H = \sum_{i \in U} \sum_{a<b \text{ separated by } \mathcal{P}_i} \langle W_{ab}, R_{ab} \rangle_F. \end{aligned}$$

引理 32 (联合矩界**)**。存在仅依赖于 K 的 $h_K > 0$ 和 $C_K > 0$，使得若 $h \le h_K$，则

$$|\mu_H - A_H| + |V_H^0 - A_H| \le C_K h A_H.$$

令 $V_H = (1 + C_K h) V_H^0$，则对于每个模式 H、具有该模式的每个候选项 τ 以及所有 $|x|$、$|y| \le 2$，

$$\log \mathbb{E} e^{x F_\tau + y(T_H - \mu_H)} \le \tfrac{1}{2} V_H (x^2 + y^2). \tag{10}$$

证明。对于划分 $\mathcal{P}$，令 D 为 R 在该划分下的分块对角部分，$E = R - D$，并令 $\mathcal{O}$ 为投影到不同块之间元素的弗罗贝尼乌斯正交投影。设 $\Omega = R^{-1}$。逆矩阵恒等式给出

$$\begin{aligned} \mathcal{O}(-\Omega) - E &= \mathcal{O}(D^{-1} E \Omega - E) \\ &= \mathcal{O}\big((D^{-1} - I) E \Omega + E(\Omega - I)\big). \end{aligned}$$

由于 $\|D - I\|_{\mathrm{op}} \le h$，当 h 较小时，D 和 R 的逆矩阵与恒等映射之差的算子范数为 $O(h)$。因此，所示差的弗罗贝尼乌斯范数为 $O(h)\|E\|_F$。对标量积应用柯西 – 施瓦茨不等式，对范数应用三角不等式，并对拆分行求和，可得 μ_H 和 V_H^0 的所述相对误差。

将 T_H 和 F_τ 表示为堆叠高斯特征中的二次型，其对称矩阵分别为 M_T 和 M_F。视图 a 与 b 之间的每条边都会在一个块中贡献 $W_{ab}/2$，并在相对的块中贡献其转置。由于每对视图的新增边数与丢失边数相等，两个弗罗贝尼乌斯范数都等于 $V_H^0/2$。由于 M_T 的边连接相同的源行，而 M_F 的边连接不同的源行，因此这两个矩阵在弗罗贝尼乌斯意义下正交。每个矩阵的每个块行中至多有 $K - 1$ 个非零块，其分块算子范数为 $O(h)$。由块范数组成的标量对称矩阵可以界定原矩阵的算子范数，其最大行和为 $O_K(h)$。因此 $\|M_T\|_{\mathrm{op}} + \|M_F\|_{\mathrm{op}} = O_K(h)$。

设 $\mathcal{R}=I_n\otimes R$ 和 $C=\mathcal{R}^{1/2}(xM_F+yM_T)\mathcal{R}^{1/2}$。对于给定的 x， y范围，选择h_K 使得 $\|C\|_{\mathrm{op}}\le 1/4$。中心高斯二次型公式给出

$$\begin{aligned}\log \mathbb{E}e^{xF_\tau+y(T_H-\mu_H)} &= -\tfrac{1}{2}\log\det(I-2C)-\operatorname{tr}C\\ &\le (1+O_K(h))\|C\|_F^2\\ &\le \tfrac{1}{2}(1+C_Kh)V_H^0(x^2+y^2).\end{aligned}$$

第一个不等式对每个特征值 z 展开 $-\frac{1}{2}\log(1-2z)-z=z^2+O(|z|^3)$。第二个不等式用到了 $\|\mathcal{R}\|_{\mathrm{op}}\le 1+h$ 以及 M_T 和 M_F 的正交性。两个结论都适用同一个 C_K，且该值与 n、特征维度和模式无关。 □

最小割控制每一种模式。令 $c_*=\min_{\varnothing\neq S\subsetneq[K]} c_R(\{S,S^c\})$。将一个 p 块划分 $\mathcal{P}$ 中的每个块与其余部分分开，会使每条割边都被计数两次。因此

$$c_R(\mathcal{P})\ge \tfrac{p}{2}c_*,\qquad A_H\ge \tfrac{1}{2}(m+Q)c_*. \tag{11}$$

命题 6 和 $I_{ab}=(1+O(h^2))\|R_{ab}\|_F^2/2$给出 $D_{\mathcal{P}}=(1+O_K(h))c_R(\mathcal{P})/2$，且对 $\mathcal{P}$ 和 $I_*=(1+O_K(h))c_*/2$一致成立。由此可得 $A_H\ge(1-O_K(h))(m+Q)I_*$，并且对于$p=|\mathcal{P}|\ge 2$,

$$D_{\mathcal{P}}\ge(1-O_K(h))\tfrac{p}{2}I_*,\qquad J\ge(1-O_K(h))\tfrac{1}{2}I_*.$$

在上侧假设下， $c_*>0$。引理32 给出 $D_H:=\mu_H^2/V_H=(1+O_K(h))A_H$。利用（11）和 $I_*\ge(2+\epsilon)$ 对数n，存在一个固定的 $\epsilon_0>0$，使得对所有充分大的 n以及每个错误模式,

$$D_H\ge 2(1+\epsilon_0)(m+Q)\log n. \tag{12}$$

当 n、 $\mu_H>0$ 和 $\mu_H/V_H=1+O_K(h)$ 足够大时， ≤ 2一致成立。下面所有矩参数的绝对值均不超过 μ_H/V_H。

两级并集界。令 $w=m$对数 n、 $\alpha=$对数 m/对数 n，以及 $q=Q/m$。用$L=(1-\alpha)w$表示模式数量对数中的主导项，并用 $M=\alpha qw$表示每个模式的候选项数量上界的对数。事实上， $\binom{n}{m}B_K^m\le\exp(L+O_K(m))$ 以及 $m^Q=e^M$。固定 $0<\delta\le\epsilon_0/2$。我们将某一给定模式下任意候选项与真实分配平局或胜过真实分配的概率界定为 $2e^{-L-\delta w}$。

若 $M\le L$，应用（10）和 $x=-y=\mu_H/(2V_H)$ 可得 $\mathbb{P}(F_\tau\ge T_H)\le e^{-D_H/4}$。由于 $\alpha\le 1/(1+q)$ 且 $q\ge 1$,

$$w\le L+M=[1+\alpha(q-1)]w\le\frac{2q}{1+q}w\le\frac{1+q}{2}w.$$

根据（12）， $D_H/4\ge(1+\epsilon_0)(L+M)$。对 e^M候选项应用并集界，至多得到$e^{-L-\delta w}$。

若 $M>L$，由柯西–施瓦茨不等式和 $L+M/q=w$ 可得

$$\begin{aligned}(\sqrt{L+\delta w}+\sqrt{M+\delta w})^2 &\le (1+q)(1+\delta+\delta/q)w\\ &\le D_H/2.\end{aligned}$$

定义 $a = \sqrt{2(L+\delta w)/D_H}$ 以及 $s = 1-a$。这里，$M > L$ 意味着 $L \le qw/(1+q)$，因此

$$a^2 \le \frac{q/(1+q)+\delta}{(1+\epsilon_0)(1+q)} \le \frac{1/4+\delta/2}{1+\epsilon_0} \le \tfrac{1}{4}.$$

因此 $s \ge a$ 和 $s\sqrt{D_H/2} \ge \sqrt{M+\delta w}$。先将 (10) 应用于 $x = 0$ 和 $y = -a\mu_H/V_H$。公共事件 $T_H < s\mu_H$ 的概率至多为 $e^{-D_H a^2/2} = e^{-L-\delta w}$。在事件 $F_\tau \ge T_H \ge s\mu_H$ 上，我们有

$$sF_\tau - a(T_H - \mu_H) \ge (s^2+a^2)\mu_H.$$

事实上，左侧至少为 $(s-a)T_H + a\mu_H$。在 (10) 中使用 $x = s\mu_H/V_H$ 和 $y = -a\mu_H/V_H$，可将单个候选项发生该事件的概率界定为 $e^{-D_H(s^2+a^2)/2} \le e^{-L-M-2\delta w}$。对至多 e^M 个候选项求和，并将公共下尾计数一次，即可得到所述 $2e^{-L-\delta w}$。

对模式和 m 求和，得到

$$_n\mathbb{P}(\text{某个错误候选项与真实分配打平或胜过真实分配}) \le 2\sum_{m=2} \exp\{O_K(m) - \delta m \log n\} = o(1)。$$

该界限对真实排列是均匀的，从而证明了部分 (a)。Dai 等人Dai 等人(2023年，附录C.3) 在分析两个数据库的部分恢复时，也采用了将真实得分的下尾与错误得分获胜的概率分离的相关方法。

D.2. 二分划分两侧一对对应关系的逆命题

揭示侧内对应关系。 取任意非平凡二分划分 $\{S, S^c\}$，通过重新标记其两侧使 $1 \in S$，并令 $I = I(X^S; X^{S^c})$。令 $b_0 \in S^c$ 和 $U = \pi_{b_0}$。揭示 $G = (G_a)_{a\in[K]}$，其中对于 $a \in S$ 有 $G_a = \pi_a$，对于 $a \in S^c$ 有 $G_a = \pi_a \circ U^{-1}$。原始排列向量与 (G, U) 构成双射，因此在给定 G 的条件下，U 服从均匀分布，并且独立于特征元组。利用 G 串联每一侧的特征，可得到真实对应关系为 U 的两个数据库。借助这两个数据库和 G，可以重构原始观测。

对于 $a \in S$ 和 $b \in S^c$，真实对应关系是 $G_b \circ U \circ G_a^{-1}$。因此，通过与 G_b^{-1} 和 G_a 进行复合，返回正确对应关系的估计器即可得到正确的 U。对于逆向过程，首先对任何取值为置换的输出求逆，并任意替换非双射输出；这不会改变任何精确成功事件。只需在这个信息更充分的问题中界定恢复 U 的概率即可。

典型相关性与获胜交换。 在各自的边缘支撑集上对两侧进行中心化和白化。通过奇异值分解对交叉协方差进行分解，并分别进行正交坐标变换，随后得到具有相关性 $r_j \in [0,1]$ 的独立配对标准高斯坐标，以及独立的残差坐标。

这些支撑集上的可逆变换与行置换可交换。I 的有限性排除了 $r_j = 1$，因为在其乘积分布边缘下，这样的配对支撑在线性测度为零的直线上。因此

$$I = -\tfrac{1}{2}\sum_{j}\log(1-r_j^2), \qquad w_j = \frac{r_j}{1-r_j^2}.$$

如果即使给定 G，$I = 0$、U 也独立于数据库，并且其精确恢复的概率至多为 $1/n!$ 现在假设 $I > 0$。由于给定 G时先验分布是均匀的，最大后验估计就是最大似然估计，并且对于精确恢复是最优的。给定任意真实 U，并重新标记第二个数据库，使其成为恒等映射。记对象 i 的典范坐标为（ξ_{ij}，η_{ij}）。真实分配的对数似然减去在第二个数据库中交换行 i，k 后的对数似然为

$$\Delta_{ik} = \sum_{j} w_j(\xi_{ij}-\xi_{kj})(\eta_{ij}-\eta_{kj}).$$

任何满足 $\Delta_{ik} < 0$ 的交换都会使真实分配严格处于次优状态。

令 $A_j = (\xi_{1j} - \xi_{2j})/\sqrt{2}$ 和 $B_j = (\eta_{1j} - \eta_{2j})/\sqrt{2}$。这些配对在 j之间相互独立，每一对都具有标准高斯边缘分布，相关性为 r_j。令 $S_{12} = \Delta_{12}/2 = \sum_j w_j A_j B_j$。将其联合密度乘以 $e^{-S_{12}}$ 后，交叉项相互抵消。总质量为 $\prod_j \sqrt{1-r_j^2} = e^{-I}$，在归一化倾斜分布下，每一对中的两个元素相互独立，且服从 $N(0,\ 1-r_j^2)$。在该倾斜分布下，S_{12}具有连续对称分布，方差为 $A_S := \sum_j r_j^2 \le 2I$。由切比雪夫不等式和对称性可知，$-2\sqrt{A_S} \le S_{12} < 0$的倾斜概率至少为 $3/8$。消除倾斜后得到

$$p := \mathbb{P}(\Delta_{12} < 0) \ge \tfrac{3}{8} e^{-I-2\sqrt{A_S}} \ge \tfrac{3}{8} e^{-I-\sqrt{8I}}. \tag{13}$$

重叠交换与二阶矩。对于交换 (1，2) 和 (1，3)，坐标 j 对 $(\Delta_{12}+\Delta_{13})/2$ 的贡献是三行与路径拉普拉斯矩阵的双线性型的 $w_j/2$ 倍。该拉普拉斯矩阵的特征值为 0，1，3。对两个数据库的行向量应用相同的对角化正交变换后，得到相互独立的高斯对 $(A,\ B)$，每对的相关性为 r_j。对于 $t \in \{1/2, 3/2\}$ 和 $\lambda_j = r_j^2/(1-r_j^2)$，直接积分可得

$$\mathbb{E}e^{-t w_j AB} = (1+\lambda_j(2t-t^2))^{-1/2}.$$

事实上，非对角元精度矩阵元素从 $-w_j$ 变为 $(t-1)w_j$，而显示的因子是两个精度矩阵行列式之比的平方根。两个参数都给出 $2t - t^2 = 3/4$。由马尔可夫不等式和对数函数的凹性可得

$$\begin{aligned} p_{\mathrm{both}} &:= \mathbb{P}(\Delta_{12} < 0, \Delta_{13} < 0) \\ &\le \mathbb{E}e^{-(\Delta_{12}+\Delta_{13})/2} \\ &= \prod_j (1+3\lambda_j/4)^{-1} \le e^{-3I/2}, \end{aligned} \tag{14}$$

其中 $\log(1+3\lambda/4) \geq (3/4)\log(1+\lambda)$，且 $\sum_j \log(1+\lambda_j) = 2I$。这些界对所有有限的高斯割信息均成立。

令 $N = \sum_{i<k} \mathbf{1}\{\Delta_{ik} < 0\}$。不相交行上的交换彼此独立，共享一行的不同交换有序对数量为 $6\binom{n}{3}$。因此 $\mathbb{E}N = \binom{n}{2}p$ 且 方差 $N \leq \binom{n}{2}p + 6\binom{n}{3}p_{\text{both}}$。方程（13）和（14）给出

$$\begin{aligned}\mathbb{P}(N=0) &\leq \frac{\operatorname{Var} N}{(\mathbb{E}N)^2} \leq O\left(\frac{1}{n^2 p} + \frac{p_{\text{both}}}{np^2}\right) \\ &\leq O\left(n^{-2}e^{I+\sqrt{8I}} + n^{-1}e^{I/2+2\sqrt{8I}}\right).\end{aligned}$$

如果 $I \leq (2-\epsilon)\log n$，则这两项分别为 $n^{-\epsilon+o(1)}$ 和 $n^{-\epsilon/2+o(1)}$，作为上界。所有上界均关于 G、真实的 U、规范谱以及割两侧的点对一致。因而，即使给定预言机信息，最大似然仍以趋近于一的概率失败。选择一个最小化割即可得到部分 (b)，包括其关于点对的结论。

附录 E. 弱相关之外精确恢复的证明

附录E.1–E.10 证明了第5节的结果。其中若干步骤是有限的计算机辅助检查，采用精确有理数运算或向外舍入区间运算完成。本文将 scripts 及其输出的表格作为辅助文件提供，且本附录中所有文件路径均相对于该目录。scripts 仅使用 Python 标准库，该目录中的README 文件列出了重新生成每张表格所需的命令。附录E.11和E.12陈述了另外两个结果，并简要概述了证明。

E.1. 稀有子总体逆命题与相关性为 $-1/3$时的常数

定义8 及其变分形式。设 p 为一个密度，并设 $\mathcal{P}$ 具有 $q \geq 2$ 个块。对于$\alpha = (q-1)/q$，稀有子群体速率有以下等价形式

$$\begin{aligned}R_{\mathcal{P}} &= \frac{1}{q-1}\min_{v=\bigotimes_B v_B} D_\alpha(p\|v) \\ &= \min_Q \left\{ D(Q\|p) + \frac{1}{q-1} D\Big(Q \,\Big\|\, \bigotimes_B Q_B\Big) \right\}.\end{aligned} \tag{15}$$

这里， Q_B 表示 Q的块边缘分布。

证明。 将 $\beta = 1-\alpha = 1/q$, $A(v) = \int p^\alpha v^\beta$ 和 $q_v = p^\alpha v^\beta / A(v)$ 记为 $A(v) > 0$。对于任意相对于 p, v 具有有限目标函数的 Q，代入倾斜密度可得

$$D(Q\|p) + \frac{D(Q\|v)}{q-1} = \frac{1}{\alpha} D(Q\|Q_v) + \frac{D_\alpha(p\|v)}{q-1}. \tag{16}$$

这由共同支撑集上的 $\log q_v = \alpha \log p + \beta \log v - \log A(v)$ 以及 $D_\alpha(p\|v) = -q \log A(v)$ 得出。若参考分布满足 $A(v) = 0$，则对于每个 Q，目标函数均为无穷大。当 $A(v) > 0$ 时，左侧的两个散度在 Q_v 处均有限。例如，对于 $0 < \eta < \alpha$，

$$\int q_v \, (q_v/p)^{\eta/\beta} = A(v)^{-1-\eta/\beta} \int p^{\alpha-\eta} v^{\beta+\eta} \le A(v)^{-1-\eta/\beta}.$$

这控制了 $\log(q_v/p)$ 在 Q_v 下的正部。交换 p、v 和 α，β 即可得到 v 的相应上界。负部由 $u \log(1/u) \le 1/e$ 在 $0 < u \le 1$ 上的可积性保证，其中积分是相对于参考密度进行的。因此，对于固定的 Q，v 关于 Q_v 的最小值由 ´实现。这是 van Erven 和 Harremo¨es (2014，定理 30) 给出的 Rényi 散度的变分形式。KL链式法则给出

$$D(Q\|v) = D\Big(Q \Big\| \bigotimes_B Q_B\Big) + \sum_B D(Q_B\|v_B).$$

特别地，对于固定的 v，Q 的极小化解是其块边缘分布的乘积分布。这两次极小化给出（15），其中取下确界。

为证明极小值可以达到，在变量 $f_B = v_B^{1/q}$ 中最大化 $A(v)$。每个 f_B 都属于 L^q 的非负单位球。由自反性和正锥的弱闭性可知，它具有弱紧性。泛函 $\int p^\alpha \prod_B f_B$ 在这些球的乘积上联合弱连续。事实上，$p^\alpha \in L^{1/\alpha}$ 可以在其中由有限测度矩形指示函数的有限和逼近。对于每个这样的和式，其积分只依赖于 f_B 的有限个弱坐标，因此连续。H¨older 不等式以 $L^{1/\alpha}$ 误差一致地控制逼近误差，因为在乘积空间上有 $\|\prod_B f_B\|_{L^q} = \prod_B \|f_B\|_{L^q} \le 1$。取正的块密度即可看出最大值为正。正齐次性随后迫使最大值处对每个 B 都有 $\|f_B\|_{L^q} = 1$。相应的 v 和 Q_v 达到联合最小值，因此由链式法则有 $v = \bigotimes_B (Q_v)_B$。在第二种形式中取 $Q = p$，得到 $R_{\mathcal{P}} \le D_{\mathcal{P}}/(q-1)$。 □

定理 9。固定 K，一个密度 p 以及一个划分 $\mathcal{P}$，其包含 $q \ge 2$ 个块。若对所有足够大的 $d R_{\mathcal{P}} \le, \ 1-\epsilon) \log n$，则任意估计器恢复所有相对置换的概率趋于零。对于高斯 p 和 $q = 2$，必要的交换界已经蕴含了精确恢复所需的稀有子总体界。

证明。令 v 和 $Q = Q_v$ 达到（15），并令 $D = D\,(Q\|p)$ 以及 $T = D\,(Q\|v)$。这些量是有限的，并且 $R_{\mathcal{P}} = D + T/\,(q-1)$。若 $R_{\mathcal{P}} = 0$，则二者都趋于零，从而 $p = Q = v$。揭示块内对齐后，剩下 $q-1$ 个独立的均匀置换，块之间没有任何信息，而将它们全部猜中的概率为 $(n!\,)^{-(q-1)}$。

假设 $R_{\mathcal{P}} > 0$。如有必要，通过减小固定裕量，我们可以取 $0 < \epsilon < 1$。将 d 增大到 $\max\{d, \lfloor \sqrt{\log n} \rfloor\}$，对于较大的 n 仍保持该假设，并且只会提高最优成功概率。因此，我们可以假设 $d \to \infty$，并且有 $d = O(\log n)$。固定一个足够小的 $\delta > 0$，使得对于所有足够大的 n，$qd\delta \le (q-1)\epsilon \log n/2$。对于 d 坐标的一行，令 A_d 为以下事件的交集

$$\log \frac{dQ^{\otimes d}}{dP^{\otimes d}} \le d(D+\delta), \qquad \log \frac{dQ^{\otimes d}}{dV^{\otimes d}} \le d(T+\delta).$$

这里 P, V具有密度 p, v，并且密度比可以限制在 $Q^{\otimes d}$的支撑集上。大数定律给出 $Q^{\otimes d}(A_d) \to 1$。对于每个对象独立地以如下概率对其进行标记

$$e^{-d(D+\delta)} \frac{dQ^{\otimes d}}{dP^{\otimes d}} \mathbf{1}_{A_d}.$$

根据第一个定义不等式，这属于 $[0,1]$ 。给定一行被标记，其分布为 $Q_A = Q^{\otimes d}(\cdot \mid A_d)$。标记行的数量 m 服从均值为$ne^{-d(D+\delta)}Q^{\otimes d}(A_d)$的二项分布；由于 $D \le R_{\mathcal{P}}$，该均值至少以 n 的正幂增长。因此，二项分布的集中性表明，以趋近于 1 的概率，

$$\log m = \log n - d(D+\delta) + O(1). \tag{17}$$

对每个块选择一个代表视图，并为其自身所在的块选择视图 1 。揭示所有块内对应关系、代表视图中标记位置的集合，以及代表视图之间所有未标记位置的对应关系。为说明在揭示之前，块内映射与代表置换相互独立，可将每个非代表置换与代表置换的逆矩阵复合。这是有限置换匹配方案集合上的一个双射。给定锚点的标记集、其未标记位置的像以及各标记像的集合后，其他每个代表都在锚点的标记集与其自身标记集之间具有均匀双射，且彼此独立。标记由置换前的行通过独立的随机硬币生成，因此这些双射与洗牌前的标记行相互独立。给定锚点的标记位置以及未标记数据后，标记行仍然相互独立，且服从 Q_A。因此，剩下的问题包含m 这样的行以及 $q-1$ 个相互独立的均匀双射。Gaudio 等人 (2022年) 和R´acz 与 Zhang (2024) 对相关块模型的精确社区恢复反定理，也采用了在数据依赖的顶点集合之外进行类似揭示的方法；此处，该集合由似然比标记确定，使保留行成为来自倾斜分布的独立同分布样本。

在固定已揭示信息后，将其条件分布与以下分布进行比较：对这些双射采用相同的先验，并在打乱前从 $V^{\otimes d}$ 中独立抽取各行。由于 v 按块分解，参考分布下的打乱数据与这些双射相互独立。任意估计器的参考成功概率至多为 $(m!)^{-(q-1)}$，对其随机化取平均后亦然。 A_d 的第二个定义不等式给出

$$\frac{dQ_A}{dV^{\otimes d}} \le \frac{e^{d(T+\delta)}}{Q^{\otimes d}(A_d)}.$$

对各行相乘，可将植入分布下的条件成功概率上界为以下内容的指数

$$m\{d(T+\delta) - \log Q^{\otimes d}(A_d)\} - (q-1)\log(m!).$$

根据（17），使用 $\log(m!) \ge m\log\, m - m$ 将该指数上界为

$$m\{(q-1)(dR_{\mathcal{P}} - \log n) + qd\delta + O(1)\} \;\le\; -\tfrac{1}{4}(q-1)\epsilon m \log n$$

当 n 较大时。因此，在已揭示信息的条件下，成功概率趋于零，原问题中也同样如此。

对于最后一个断言，考虑任意一个双块割 $S|S^c$。非负的希尔伯特–施密特核 $z=\sqrt{p}$，视为从一个块的 L^2 到另一个块的算子，其奇异值为 σ_i，满足 $a_i=\sigma_i^2$ 和 $\sum_i a_i=1$。在块乘积参考分布上最大化 $\int z(x,\,y)\sqrt{v_S(x)}\sqrt{v_{S^c}(y)}\,dx\,dy$，得到 $\|z\|_{\mathrm{op}}=\sqrt{\max_i a_i}$。事实上，算子范数由一对单位奇异向量取得，而取绝对值不会降低它们与非负核的配对值。由此得到的非负单位向量是概率密度的平方根。因此

$$R_{S|S^c}=-\log\max_i a_i=H_\infty(a)\;\ge\;\tfrac{1}{2}H_2(a),\qquad H_2(a)=-\log\sum_i a_i^2.$$

该不等式使用了 $\sum_i a_i^2\ge(\max_i a_i)^2$。跨越该割的交换指数为 $H_2(a)=I_2^{\circ}(S)$；对于高斯 p，根据附录E.2，该值为 $I(X_S;\,X_{S^c})$。因此，在高斯情形下有 $R_{S|S^c}\ge I_*/2$。沿一个 m-环积分，得到中点系数 $\mathrm{tr}[(z^*z)^m]=\sum_i a_i^m$。通过对对象重新标记，该环与其逆环共轭，因此，依据对数凸性，Chernoff系数关于参数对称，并在中点处取最小值。每个移动对象对应的指数为 $-\log\|a\|_m$，其递增至 $H_\infty(a)$。

□

$r=-1/3$ 处的常数以及定理 10 中的失败窗口。令 $K=3$ 和 $p=N(0,R)$具有单位方差，且三者的相关性均等于 $-1/3$。逐坐标速率分别为 $I_*=\tfrac{1}{2}\log(3/2)$、$J=\tfrac{1}{4}\log(27/16)$ 和 $\bar{R}_{1|2|3}\approx0.0983\bar{5}$。如果 $d=\lceil\kappa\log\,n\rceil$ with$9.866<\kappa<10.167$，则总信息量 dI_* 和 dJ 分别超过 $(2+\epsilon)\log\,n$ 和 $(1+\epsilon)\log\,n$，其中某个固定的 $\epsilon>0$。精确恢复仍以趋近于 1 的概率失败。

证明。 R 的特征值为 $4/3$、$4/3$、$1/3$，因此 $\det R=16/27$。任意两视图主子式均为 $8/9$。高斯行列式公式给出 $I_*=\tfrac{1}{2}\log[(8/9)/(16/27)]$，而单元素划分的归一化信息为 $-\tfrac{1}{4}\log\det\,R$。后者小于I_*，因此它等于 J。

对于单例参考，取 $V=vI$ 其中 $v=(1+\sqrt{3})/3$。倾斜协方差为$S=[2R^{-1}/3+I/(3v)]^{-1}$。其对角元素相等，并且 $\mathrm{tr}\,S=3v$ 等价于

$$1=\frac{1}{6v+1}+\frac{4}{3v+2},$$

其满足所述条件 v。该边际方程对于最优性的充分性已在附录 E.3中得到证明。S^{-1} 的特征值为 $\sqrt{3}/2$，重数为二；以及$(3+\sqrt{3})/2$，重数为一。因此 $v\det(S^{-1})=(3+2\sqrt{3})/4$。现在由亲和度公式（ 19 ）可得

$$R_{1|2|3}=\frac{1}{2}\log\frac{16}{27}+\frac{3}{4}\log\frac{3+2\sqrt{3}}{4}.\tag{18}$$

数值证书。scripts/check_d1_k3_sharp.py 中的有理区间检验给出了 (18) 及其倒数的范围。d1_k3_sharp/rows.csv 中的 R3 行和逆 R3 行给出 $0.098353188221 < R_{1|2|3} < 0.098353188222$ 和 $10.167438575994 < R_{1|2|3}^{-1} < 10.167438575995$。下不等式 $\kappa I_* > 2$ 由 $2/I_* < 9.866$得出，而后者可通过保留级数 $\log(3/2) = 2\sum_{j\ge 0}(1/5)^{2j+1}/(2j+1)$ 中前三个正项并比较有理数得到。此外 $J > I_*/2$，因为 $27/16 > 3/2$。上界 $\kappa < 10.167 < R_{1|2|3}^{-1}$ 为定理 9 提供了一个固定裕量，对所有足够大的 n均成立，从而证明了该失败结论。 □

E.2. 有界模式

命题 12(a). 对于每个密度 p，任意涉及有界数量对象的错误模式，其切尔诺夫指数至少为 $\min_S I_2^\circ(S)$ 每个移动对象对应的该值；当交换发生在一个最小化割两侧时取等号。这里的 I_2° 是 Cullina 等人 Cullinaet al.(2018年定义的二阶循环互信息。对于高斯分布，它等于跨越该割的互信息。

证明。 在视图和行上使用乘积基准测度。将第一个视图设为锚点，并令$\pi = (\pi_1, \dots, \pi_K)$满足 $\pi_1 = \mathrm{id}$。保留至少在一个视图中被移动的 k 行，这一集合在每个 π_a 作用下保持不变。对于一个坐标，记作 $x_i = (x_{ia})_{a=1}^K$, $\underline{y}_i = (x_{\pi_a(i)\,a})_{a=1}^K$, $p_0 = \prod_i p(x_i)$，以及 $p_\pi = \prod_i p(y_i)$。给定 $z = \sqrt{p}$，令

$$H_\pi(t) = \int p_0^{1-t} p_\pi^t, \qquad C_\pi = -\log \inf_{0<t<1} H_\pi(t).$$

中点积分的被积函数为 $\prod_i z(x_i)z(y_i)$。对于每一行 i，令 $S_i = \{a : \pi_a(i) = i\}$，这是 $[K]$的一个非空真子集。S_i 中的变量出现在同一对因子 $z(x_i)z(y_i)$ 中，因此可以作为一组一起积分。所得因子为

$$G_i(u, v) = \int z(s, u)z(s, v)\, ds, \qquad (u, v) = (x_{i,S_i^c}, y_{i,S_i^c}).$$

这是 $Z_{S_i}^* Z_{S_i}$ 的核，其中 Z_{S_i} 是希尔伯特–施密特算子，其核为 z，对应割 $S_i|S_i^c$。因此 $\|G_i\|_2^2 = \mathrm{tr}[(Z_{S_i}^* Z_{S_i})^2] = e^{-I_2^\circ(S_i)}$。

每个剩余变量都出现在两个不同的因子中，分别位于 G_j 的第一个自变量和 $G_{\pi_a^{-1}(j)}$ 的第二个自变量中；对变量 x_{ja} 也是如此。逐个变量积分。如果变量 x 出现在因子 F_i, F_j 中，那么柯西–施瓦茨不等式会将 $\int F_i F_j\, dx$ 替换为上界 $(\int F_i^2\, dx)^{1/2}(\int F_j^2\, dx)^{1/2}$，其余变量均固定不变。经过此替换，每个因子都保留其完整的 L^2 范数。托内利定理适用于这些非负被积函数。对所有变量积分后，得到

$$C_\pi \;\ge\; -\log H_\pi(1/2) \;\ge\; \tfrac{1}{2}\sum_{i=1}^{k} I_2^\circ(S_i) \;\ge\; \tfrac{k}{2}\min_S I_2^\circ(S).$$

对于跨越任意割的两个对象进行交换时，交换不包含锚点一侧的视图。该置换是一个对合，因此变量替换可得 $H_\pi(t) =$

$H_\pi(1-t)$。H¨older 给出对数凸性，且最小值在 $1/2$ 处。这里的中点积分恰为 $\int G(u,\ v)G(v,\ u) = \|G\|_2^2$，得出两个移动对象对应的指数 $I_2^\circ(S)$。对 S 取最小值即可达到下界。

对于 $N(0,\ R)$，对交换中的两行应用正交和差变换。若 D_S 为对角矩阵，交换侧的对角元为 -1，另一侧为 $+1$，则在这些坐标下，真实协方差矩阵和交换后的协方差矩阵分别为 $\Sigma_0 = \mathrm{diag}(R,\ R)$ 和 $\Sigma_1 = \mathrm{diag}(R,\ D_S R D_S)$。它们的行列式相等。它们的高斯中点系数为

$$(\det\Sigma_0 \det\Sigma_1)^{1/4} \det[(\Sigma_0+\Sigma_1)/2]^{-1/2}.$$

差分坐标上的平均值为 $(R + D_S R D_S)/2 = R_{SS} \oplus R_{S^cS^c}$，因此中点系数的负对数为

$$\tfrac{1}{2}\log\frac{\det R_{SS}\det R_{S^cS^c}}{\det R} = I(X_S; X_{S^c}).$$

对于 d 个独立坐标，某个特定候选项的中点误差界为$H_\pi(1/2)^d$。在候选似然不低于真实似然的事件上，真实密度不超过二者的几何平均数。对每个固定的 k，候选项至多有$\binom{n}{k}(k!)^{K-1}$ 个。因此，若 $d\min_S I_2^\circ(S) \geq (2+\epsilon)\log n$，则对 $2 \leq k \leq k_0$ 作并集界；对任意固定的 k_0，以趋于 1 的概率排除所有此类竞争者。□

E.3. 高斯乘积参考分布

命题 12(b). 对于具有正定协方差的中心高斯分布 p，定义 8 中的极小化解唯一，且为中心高斯块乘积密度。

证明。取 $\beta = 1/q$、$\alpha = 1-\beta$ 和 $p = N(0, R)$。对于精度矩阵为块对角形式 $W = \bigoplus_B W_B \succ 0$ 的中心高斯参考分布，令 $A(W) = \int p^\alpha v^\beta$。写作 $C = \alpha R^{-1}$，通过高斯积分可得

$$A(W)^2 = (\det R)^{-\alpha}\frac{(\det W)^\beta}{\det(C+\beta W)}. \tag{19}$$

若 $c = \lambda_{\min}(C) > 0$，则行列式比有界于 $\prod_i \lambda_i(W)^\beta/(c+\beta\lambda_i(W))$，因为 $C + \beta W \succeq cI + \beta W$。每个因子都有界，且在 $(0,\infty)$ 的两端都趋于零。因此，$A(W)$ 的非空正超水平集在有限维定义域中是紧集，故最大值可在一个内点 W_*处取得。

令 $S = (C+\beta W_*)^{-1}$。在每个块中沿任意对称方向对 $\log A(W)^2$ 求导可得

$$S_{BB} = W_{*,B}^{-1} \quad \text{for every } B. \tag{20}$$

设 g_B为精度矩阵为$W_{*\text{的高斯密度，}B}$且 $Q_* = N(0,S)$是相对于 p 的归一化倾斜分布 $g = \bigotimes_B g_{B\circ}$ 根据（20），其块边缘分布等于 g_B。对于任意乘积密度 v，令 $h_B = v_B/g_{B\circ}$ 由算术—几何平均不等式可得

$$\frac{A(v)}{A(g)} = \mathbb{E}_{Q_*}\prod_B h_B(X_B)^{1/q} \leq \frac{1}{q}\sum_B \mathbb{E}_{Q_*} h_B(X_B) = 1.$$

等号成立要求 $h_B(X_B)$ 几乎处处与 Q_*一致。该倾斜高斯分布在整个乘积空间上具有严格为正的密度，因此根据 Fubini定理，这些互不相交变量的函数必须几乎处处等于同一个常数。对 v_B 进行归一化可知该常数为一。这证明了乘积密度中的全局最优性与唯一性。对于满足（ W 的任意正定 20），同样的论证也能证明其最优性。根据（16），联合倾斜分布 Q_* 也达到了 KL最小值。 □

E.4. 当新界生效时

命题 12(c). 对于具有非负相关性的三个高斯视图，以及由一个具有正载荷和独立高斯噪声的共同高斯因子生成的任意固定数量的视图， $L = I_*/2$。存在所有两两相关性均为正的四视图高斯分布，使得 L 同时小于 $I_*/2$ 与 J。对于每一个具有 $0 < J \le I_*/2$的高斯分布，某个划分满足 $R_{\mathcal{P}} < J$。

证明。 对于固定高斯协方差矩阵 $C \succ 0$ 与分块对角参考协方差$V \succ 0$，用协方差行列式改写后的方程（19）给出

$$2(q-1)R_{\mathcal{P}} = \min_{V} \Big\{ q \log\det[(C + (q-1)V)/q] - \log\det C - (q-1)\log\det V \Big\}. \tag{21}$$

事实上 $\alpha C^{-1} + (1-\alpha)V^{-1} = C^{-1}(\alpha V + (1-\alpha)C)V^{-1}$，并且附录 E.3 为在高斯参考分布上最小化提供了依据。对于标量块，记 $V = \mathrm{diag}(v_1, \dots, v_q)$ 并选取 S，即对每个坐标独立地以概率 $1/q$予以纳入。主子式展开给出

$$\det[(C + (q-1)V)/q] = \mathbb{E}\Big[\det C_{SS} \prod_{i \notin S} v_i \Big],$$

使用$\det C_{\varnothing\varnothing} = 1$。对数的詹森不等式抵消了 v_i 式（21）中的项，并得到

$$2(q-1)R_{\mathcal{P}} \ge q\,\mathbb{E}\log\det C_{SS} - \log\det C. \tag{22}$$

三个非负相关的视图。将边缘分布方差重新缩放为 1 后，令 $m = \min_{i<j}(1 - r_{ij}^2)$ 和 $\Delta = \det R > 0$。只有一种三块划分。最小割为 $I_* = \frac{1}{2}\log(m/\Delta)$，因此只需证明

$$g = \frac{\det[(R + 2V)/3]}{(\det V)^{2/3}} \;\ge\; m^{1/3}$$

对所有正对角元 V均成立。令 $s_i = (2v_i)^{-1}$ 和 $s = (s_1 s_2 s_3)^{1/3}$。由主子式展开和算术—几何平均不等式可得

$$\begin{aligned} g &= \frac{4}{27 s}\left[1 + \sum_i s_i + \sum_{i<j} (1 - r_{ij}^2) s_i s_j + \Delta s^3 \right] \\ &\ge \frac{4}{27}\big[s^{-1} + 3 + 3ms + \Delta s^2 \big]. \end{aligned} \tag{23}$$

行列式公式分别关于每个相关性都是凹函数，因此其在立方体 $[0,\sqrt{1-m}]^3$ 上的最小值在某个顶点处取得。代入这些顶点可得

$$\Delta \geq \min\{2m-1,\ 3m-2+2(1-m)^{3/2}\}.$$

对于 $0<m\leq 1/2$，去掉 Δs^2 并最小化 $s^{-1}+3ms$，得到 $g\geq(4/27)(3+2\sqrt{3m})$。通过求导，以下界与 $m^{1/3}$ 之比在 $(0,\ 1/2]$ 上递减。在 $1/2$ 处，下界超过 $4/5$，而其立方大于 $1/2$。

对于 $1/2\leq m\leq 5/8$，两个顶点界中较小的是 $2m-1$，因为第二个表达式比第一个大 $(1-m)(2\sqrt{1-m}-1)\geq 0$。函数 $H(m)=8/15+25m/48$ 是 $m^{1/3}$ 在 $64/125$处的上切线。固定 s，从将 Δ 替换为 $2m-1$后的式 (23) 中的下界减去 $H(m)$。所得结果关于 m是仿射的。在左端点处，下界超过 $4/5>H(1/2)$。在右端点处使用 $s^2\geq 4s/3-4/9$，然后对 s取最小值。这给出下界

$$(4/27)(26/9+2\sqrt{53/24})>209/243>H(5/8),$$

其中 $(35/24)^2<53/24$ 给出了第一个不等式。两个端点界对 s 均一致成立，因此在此区间上$g\geq H(m)$。

对于 $5/8\leq m\leq 1$，两个顶点界均至少为 m^3。第一个差为 $(1-m)(m^2+m-1)$，第二个差为 $z^3(1-z)^2(z+2)$，其中 $z=\sqrt{1-m}$。将詹森不等式应用于 g 的主子式展开，可得

$$\log g\geq \frac{2}{27}\sum_{i<j}\log(1-r_{ij}^2)+\frac{1}{27}\log\Delta\geq \tfrac{1}{3}\log m.$$

这证明了所声称的界对所有 m 均成立，因而 $R_{1|2|3}\geq I_*/2$。对于 $m<1$，该论证是严格的。

一个高斯因子。令 $X_j=\lambda_j Z+\epsilon_j$，其中 Z 为标准高斯变量，而ϵ_j 为方差为 $\Delta_{jj}>0$的独立高斯残差，且与 Z独立。对于任意划分，先对白化每个块内的残差，再将白化后的载荷向量旋转至第一坐标轴。于是每个块 B_i 都包含一个载荷为 $\sqrt{x_i}$的标量，其中$x_i=\sum_{j\in B_i}\lambda_j^2/\Delta_{jj}$，以及独立的标准高斯噪声坐标。对这些标量进行逐块投影，以及追加独立噪声的反向信道，都会保持块乘积参照测度不变。根据各方向上 R´enyi 散度的数据处理不等式，定义8 中的最小值不变。

设 $w_i=\sqrt{x_i}$保留协方差为 $C=I+ww^{\top}$。令 $T=\sum_i x_i$。矩阵行列式引理给出 $\det C_{SS}=1+\sum_{i\in S}x_i$。将 $\log(1+x)=\int_0^\infty e^{-t}(1-e^{-tx})\,dt/t$ 代入（22）。对于每个 $t>0$，以概率 $u_i=1-e^{-tx_i}$ 独立地激活每个块，并令 N 表示激活的块数。给定激活集合后，它满足 S的概率为 $1-(1-1/q)^N$。因此，提出因子 e^{-t}/t后，被积函数等于以下量的期望：

$$q[1-(1-1/q)^N]-\mathbf{1}_{N>0}.$$

The 当 $N=0$，1 时，该表达式为零。对于 $2\le N\le q$，Bonferroni不等式给出

$$q[1-(1-1/q)^N]-1\ \ge\ N-1-\binom{N}{2}/q\ \ge\ (q-1)N/(2q).$$

最终的差乘以 $2q$ 后为 $(N-2)(q-N)$。选择一个满足 x_b 最小的指标 b。b 被激活且另一个块也被激活的概率至多为 $\mathbb{E}[N\mathbf{1}_{N\ge 2}]/q$。对于任意 $i\neq b$，i 对应概率与 b 对应概率之差为 $(u_i-u_b)(1-\prod_{j\neq i,\,b}(1-u_j))\ge 0$，因此 b 达到了这些 q 概率中的最小值。

b 的概率积分根据同一对数恒等式为 $\log[(1+x_b)(1+T-x_b)/(1+T)]$。由此可得

$$2(q-1)R_{\mathcal{P}}\ge \tfrac{q-1}{2}\log\frac{(1+x_b)(1+T-x_b)}{1+T}=(q-1)I(X_{B_b};X_{B_b^c})\ge (q-1)I_*.$$

该等式用到了单因子协方差行列式以及被舍弃的噪声坐标的独立性。这证明了该结论对每一种划分都成立。如果某个残差方差为零，而完整协方差是正定的，则在每个视图中加入方差为 $\eta>0$ 的独立噪声。数据处理原理给出的原始 $R_{\mathcal{P}}$ 下界不高于扰动后的 $R_{\mathcal{P}}$，且有限个割行列式在 $\eta\downarrow 0$ 时收敛。

四个严格正相关视图。考虑

$$R=\begin{pmatrix}1 & a & b & b\\ a & 1 & b & b\\ b & b & 1 & a\\ b & b & a & 1\end{pmatrix},\qquad a=1/5,\quad b=59/100.$$

其特征值为 $1-a$、$1-a$、$1+a-2b$、$1+a+2b$，最小值为 $1/50$。记 $D=$行列式R和 $M=$行列式$[(R+3I)/4]$。这些行列式为

$$D=(1-a)^2[(1+a)^2-4b^2]=476/15625,\qquad M=14663459/16000000.$$

每个三视图主子式均为 $(1-a)(1+a-2b^2)$。两区块划分的块主子式乘积属于以下数值：$(1-a^2)^2$,和 $(1-b^2)^2$。其最小值为$p_2=2519/6250$。对于三个区块，唯一的非平凡块大小为二，因此最小乘积为 $p_3=1-b^2=6519/10000$。四个单元素区块的乘积为一。因此

$$I_*/2=\tfrac14\log(p_2/D),\qquad J=\min\{I_*,\tfrac14\log(p_3/D),-\tfrac16\log D\}.$$

可行单例参考 $V=I$ 在（21）中给出 $R_{1|2|3|4}\le U=\frac16\log(M^4/D)$。比较所显示的有理数得到 $p_2>9D$，$p_3>9D$，以及 $M^4<27D<1$。前两个不等式和 $27D<1$ 蕴含 $I_*/2$，$J>$（log 3）/2，而 $M^4<27D$ 蕴含$U<$（log 3）/2。因此 $L\le U<\min\{I_*/2,J\}$。

数值证书。 scripts/check_d3_meso.py 使用分数验证正定性以及比较论证 $p_2 > 9D$、$p_3 > 9D$ 和 $M^4 < 27D < 1$。d3_meso/rows.csv 中的正证书行记录了 $I_*/2$、J 和 U 的相应值。

一种确保严格优于 J 的条件。若 $0 < J \le I_*/2$，则达到J 的划分满足 $q \ge 3$，因为两块的信息率至少为 $I_* \ge 2J > J$。对于一个最小化划分，取 $v = \bigotimes_B p_B$。其对数似然比 $\log(p/v)$ 在 p 下的均值为$D_{\mathcal{P}} = (q-1)J > 0$。该比值不是常数，因为两个高斯密度处处为正，而比值若为常数，则会得到 $p = v$。对 $(\alpha-1)\log(p/v)$ 的指数应用严格詹森不等式，得到 $D_\alpha(p\|v) < D(p\|v)$。因此 $R_{\mathcal{P}} \le D_\alpha(p\|v)/(q-1) < J$。 □

E.5. 三个负相关等相关视图的阈值

定理 10。对于相关性相等的三个高斯视图 $r \in (-1/2,\ 0)$，$L(R) = R_{1|2|3}(r)$。对于任意固定的 $\epsilon > 0$，若 $d \ge (R_{1|2|3}^{-1} + \epsilon)\log n$，则最大似然估计器以趋于一的概率恢复所有置换。若 $d \le (R_{1|2|3}^{-1} - \epsilon)\log n$，则任何估计器的成功概率都趋于零。在 $r = -1/3$，$R_{1|2|3} \approx 0.09835$。

证明。设 $R_3 = R_{1|2|3}$，并令 ϕ_v 表示方差为v的中心化标量高斯密度。我们先证明在 $r = -1/3$处的充分性，然后将证明推广到整个族。

在 $r = -1/3$ 处的倾斜与双侧检验。其最优倾斜及其归一化常数为

$$q_3 = Z_3^{-1} p^{2/3} (\phi_{v_3}^{\otimes 3})^{1/3}, \qquad v_3 = (1+\sqrt{3})/3, \qquad -\tfrac{3}{2}\log Z_3 = R_3.$$

事实上，精度矩阵 $2R^{-1}/3 + I/(3v_3)$ 的逆矩阵的对角元为 v_3，相关性为 $-(2-\sqrt{3})$。将其边缘分布与乘积参考分布匹配后，根据命题 ¨12 的证明所示，由 H12(b) 可得最优性。高斯归一化因子给出 $R_3 = \frac{1}{2}\log(16/27) + \frac{3}{4}\log[(3+2\sqrt{3})/4]$。我们可以取 $d = \lceil \bar{\kappa}\log n \rceil$，对于 $R_3^{-1} < \kappa < \min\{10.2, R_3^{-1} + \epsilon\}$，在解码器中仅使用这么多个坐标即可。

在所有满足每个行子集 S 均符合以下条件的对齐中，选择似然最高的对齐

$$\sum_{i\in S}\sum_{t=1}^{d} \log \frac{q_3(X_{it})}{p(X_{it})} \le t_{|S|}, \qquad t_j = \log\binom{n}{j} + \eta j \log n, \qquad t_0 = 0. \tag{24}$$

如果没有符合条件的对齐，则任意选择一个输出。在真实表上，固定子集对应统计量的指数的期望为一。根据马尔可夫不等式和并集界，对于任意固定的 $\eta > 0$，拒绝概率至多为 $\sum_{j\ge 1} n^{-\eta j} = o(1)$。

重新标记，使真实排列成为恒等置换。用真实行与提议行之间的二分图表示错误对齐。将同一对行共享的所有视图合并为一条边。锚边构成完美匹配，因此非平凡连通分量在两侧各有 m 个顶点。其顶点的度数为二或三。度数之和

相等，因此两侧的三度顶点数量相同。 $k = fm$ 得分差非负、且与合格真实表持平或优于它的错误合格对齐，至少包含一个连通分量。其三度顶点子集在两侧均满足 (24)，因为每一侧的子集都属于各自的整张表。

对于一个固定的此类连通分量，得分差非负这一事件使我们可以用真实密度与提议密度的几何平均数来界定真实密度。在对三度顶点子集进行的两项检验中，我们可以将这个界乘以 $\exp(t_k)$，再乘以这些顶点和坐标上 $(p/q_3)^{1/2}$ 的乘积。在每个三度顶点处，得到的因子为 $Z_3^{3/2} q_3 / \prod_e \sqrt{\phi_{v_3}}$。在二度顶点处，其双视图边上的两个视图可以经过旋转和缩放，变换为标准化和与差。差与和及第三个视图相互独立。同一条边上的差的两个平方根密度积分为一。因此，只剩下标准化和与第三个视图，其相关性为 $\rho = r\sqrt{2/(1+r)}$。令

$$v_2 = \sqrt{1-\rho^2}, \quad y = \frac{|\rho|}{1+v_2}, \quad R_2 = \log\frac{1+v_2}{2v_2}, \quad Z_2 = e^{-R_2/2}.$$

对于其密度 p_2，高斯 $q_2 = Z_2^{-1}\sqrt{p_2\phi_{v_2}\phi_{v_2}}$ 的方差为 v_2，相关性为 $-y$，这可由对二维精度求逆得到。此类顶点处的因子为 $Z_2 q_2 / \prod_e \sqrt{\phi_{v_2}}$。现在从积分中去掉该事件，可将其概率界定为

$$\exp\{t_k - d[(m-k)R_2 + 2kR_3]\}\,\mathcal{I}_G^d, \qquad \mathcal{I}_G = \int \prod_{v\in G} \frac{q_v(x_{\partial v})}{\prod_{e\ni v}\sqrt{\phi_{\nu_v}(x_e)}} \prod_e dx_e, \tag{25}$$

其中 $q_v = q_j$ 和 $\nu_v = v_j$ 位于度为 j 的顶点。

具有匹配边缘分布的高斯积分。对于（25）中的积分，考虑在有向边上与 $\sqrt{\phi_s}$ 成正比的正消息。要传递一条消息，将顶点因子与其他边上的传入消息相乘，并对其他变量进行积分。在 ν_v 方差单位中，方差为 s 的传入消息增加 $H = (\nu_v/s - 1)/2$ 的精度矩阵。如果输出变量随后在这些单位下的方差为 T，则除以剩余的平方根边缘分布密度，得到输出精度比 $2/T - 1$。

在 $r = -1/3$ 处，盒 $39/40 \le \nu_v/s_{v\to w} \le 41/40$ 是不变的。为说明这一点，将 C_v^0 写作 v 处的原始相关矩阵。如果其余输入在除输出 e 之外的边集合 O 上具有精度增量 $D = \mathrm{diag}(H_i)$，那么

$$T = 1 - C_{eO}^0 (I + DC_{OO}^0)^{-1} DC_{Oe}^0.$$

v_2、v_3 的闭式给出 $|H_i| < 0.073$ 和 $H_i > -0.063$。任何三度端口或双视图端口上的输入都满足 $H_i \ge -1/80$，这是由 $v_3 \ge v_2$ 以及双视图端口上的公共度数得到的。利用 $\|C_{OO}^0\| \le 1.32$ 和 $\|C_{Oe}^0\|^2 \le 2(0.268)^2$，可得 $|T-1|$ 不超过 $0.073 \cdot 2(0.268)^2/(1 - 1.32\cdot 0.073) < 0.0117$。因此，$2/T - 1$ 位于该盒内。所有积分精度矩阵均为正，因此更新是连续的，而布劳威尔不动点定理保证存在一个不动点。

在任何这样的不动点处，将边 e 上的两条消息的乘积分布归一化，得到b_e；将顶点因子与所有传入消息相乘并归一化，得到 b_v。更新意味着 $(b_v)_e = b_e$。记 C_v 为所得的顶点相关矩阵，则在 $r = -1/3$ 处，三度时有$|(C_v)_{ij}| < 0.278$，二度时有 < 0.330。

对于三度情形，令 $x = 2-\sqrt{3}$。原相关精度矩阵为 $aI + b\mathbf{1}\mathbf{1}^{\top}$，其中 $a = 1/(1+x)$，且 $b = x/[(1+x)(1-2x)]$。加入消息增量后消去第三个坐标 l，得到对角元为 $a + H_i + b'$、非对角元为 $b' = b(a + H_l)/(a + H_l + b)$ 的二维精度矩阵。绝对相关性至多为 $b/(a - 1/80 + b)$，其中 b 在 $H_l = 0.073$处取值为 b'。二度时，绝对相关性为 $y/\sqrt{[1 + (1 - y^2)H_i][1 + (1 - y^2)H_j]}$。双视图端口满足 $H_i \geq -1/80$，由此得到所述界。

以下是这些不动点的两个一般恒等式。将定义 b_v 的恒等式相乘，并用 b_e 抵消消息后，被积函数等于 $Z_{\mathrm{B}} \prod_v b_v / \prod_e b_e$。对局部归一化的对数取期望，并利用 $(b_v)_e = b_e$ 抵消消息项，可得

$$\log Z_{\mathrm{B}} = \sum_{v} \left[-D(b_v \| q_v) + \tfrac{1}{2} \sum_{e \ni v} D(b_e \| \phi_{\nu_v}) \right] \leq 0.$$

对于该不等式，令 Q，H 表示 q_v、b_v、$M = Q^{-1/2} H Q^{-1/2}$ 和 $u_i = Q^{1/2} e_i / \sqrt{Q_{ii}}$ 的局部协方差矩阵。令 $g_0(t) = (t - 1 - \log t)/2$，则联合熵为 $\operatorname{tr} g_0(M)$，而边缘熵为 $g_0(u_i^{\top} M u_i) \leq u_i^{\top} g_0(M) u_i$，这是由 M 的特征基中的标量詹森不等式得到的。由于 $\lambda_{\max}(C_v^0) < 2$，我们有 $\sum_i u_i u_i^{\top} \preceq 2I$。求和界由 $g_0(M) \succeq 0$ 得出。

对每条边用 b_e 进行标准化，令 $C = \operatorname{diag}_v C_v$ 作用于半边，令 W 交换每条边的两个半边，并令 D_0 将边变量复制到对应的两个半边。后者满足 $D_0 D_0^{\top} = I + \mathsf{W}$ 和 $D_0^{\top} D_0 = 2I$。每当 $C \prec 2I$时，$\prod_v b_v / \prod_e b_e$的积分为

$$[\det C \, \det(D_0^{\top} C^{-1} D_0 - I)]^{-1/2} = \det(I - (C - I)\mathsf{W})^{-1/2}. \tag{26}$$

第一个表达式是对精度矩阵为 $D_0^{\top} C^{-1} D_0 - I \succ 0$的高斯分布积分所得。对于第二个表达式，将 $I - (C - I)\mathsf{W}$ 右乘 W，然后对 $D_0 D_0^{\top} - C$ 应用西尔维斯特行列式恒等式，并利用 $\det \mathsf{W} = (-1)^{|E(G)|}$。

计数与循环收费，适用于 $r = -1/3$。选取一个固定偶数截断值 $\Gamma \geq 4$。给定支撑集和所有真实行划分后，先插入锚定块，再按固定的左侧标签和视图顺序插入其他块。若插入边的端点之间的距离已不超过 $\Gamma - 1$，则将该边标记。由于度至多为三，一条标记边有 $O_{\Gamma}(1)$ 个可能的目标。每个短环都包含一条标记边，这可由考察最后一次插入得出。每个其余的非锚定块至多有 m 个目标，且这样的块共有 $m + k$ 个。有 h 个标记，且 $\theta = \log m / \log n$时，支撑选择 $\binom{n}{m}$，而划分和标记位置有 $\exp[O_{\Gamma}(m)]$ 种可能，因此

$$\log N(m, k, h) \leq (1 + f\theta) m \log n - h \log m + O_{\Gamma}(m). \tag{27}$$

令 $B = (C - I)\mathsf{W}$。它给出一次非回溯转向，随后穿越一条边。在固定分布下，其绝对行和至多为 $a = 0.556$，其元素至多为 $b = 0.330$。

特别地 $C \prec 2I$，并且 –对数行列式$(I-B) = \sum_{j\ge 1}$迹$(B^j)/j$绝对收敛。对于经过给定标记的游走，将其起点移至该标记的一次有向穿越，代价至多为j次移位，相对于因子 $1/j$。从该有向状态出发，当前两步返回在前 $j-2$ 步固定后至多有一种实现，因为该图是简单图。因此，两种定向中的每一种的返回权重均有上界 $a^{j-2}b^2$。该图是二分图，因此正长度均为偶数且至少为四。计入因子 $1/2$ （26），我们得到

$$C_{\mathrm{edge}} = \frac{b^2 a^2}{1-a^2} < 0.04873, \qquad 10.2 C_{\mathrm{edge}} < \tfrac{1}{2}. \tag{28}$$

避开标记的游走长度大于 Γ。最多有 $6m$ 个起始状态，因此其总成本至多为 $m\delta_\Gamma$，其中 $\delta_\Gamma = 3a^{\Gamma+1}/(1-a)$。因此 $\log \mathcal{I}_G \le h C_{\mathrm{edge}} + m\delta_\Gamma$。

对于 $m \ge \sqrt{n}$，我们有 $t_k \le [f(1-\theta) + \eta f] m \log n + O(m)$，使用 $f\log(1/f) \le 1/e$和 $t_0 = 0$。在 (27) 与 (25) 的乘积中，成本 dhC_{edge}由 $h\log m$覆盖。由于$R_2 > R_3$和 $\kappa R_3 > 1$，先取 η 足够小，再取 Γ 足够大，对于每个 m、k、h，概率对数中都会留下 $m\log n$ 的一个均匀负倍数。

对于 $m \le \sqrt[-]{n}$，使用原始的 Bhattacharyya 指数 B_G。以 $\sqrt{1-r}$ 为单位，行均值方向上的方差为 $\delta = (1+2r)/(1-r)$，而垂直于该方向的方差为 1。令 U, V 收集真实行和候选行的正交归一均值方向。于是 $M = U^{\top} V = W/3$，其中 W 记录每条边上的视图数。每一对真实行–候选行的主角（其余弦为 σ）根据二维中点的行列式，对中点协方差行列式比 $1 + (1-\delta)^2(1-\sigma^2)/(4\delta)$ 作出贡献。由此，令 $A = I - MM^{\top}$，

$$B_G = \tfrac{1}{2}\log\det(I + \xi A), \quad \xi = \frac{(1-\delta)^2}{4\delta}, \quad \operatorname{tr} A = m(4+2f)/9. \tag{29}$$

迹的结论成立，因为 M 的平方行范数在二度时为 $5/9$，在三度时为 $1/3$。在 $r = -1/3$ 处，$\xi = 9/16$。对谱 $[0,1]$ 应用凹性可得

$$B_G \ge m(2+f)\log(25/16)/9, \qquad (2/9)\log(25/16) > R_3. \tag{30}$$

式（27）中的未标记计数有上界 $\exp[(1 + f/2) m \log n + O(m)]$。固定裕度也排除了这一范围。每个 m 处的多项式数量的计数可由 $m\log n$的均匀负倍数吸收。因此，合格解码器能够成功。在均匀先验下，完整数据的最大似然（ML）是精确恢复的贝叶斯最优方法，因此也能成功。

固定分布的证书。 文件 `d1_k3_sharp/rows.csv` 中的有理包围由 `scripts/check_d1_k3_sharp.py` 生成，对方差框、两个转向幅度、式（28）中的边电荷以及间隙 $R_2 - R_3$ 和 $(2/9)\log(25/16) - R_3$ 进行界定。本部分的严格不等式使用外向界。

整个族中的参数。 令 $-x$ 表示 q_3 的相关性，则 $0 < x < 1/2$。倾斜协方差的特征值为 $v_3(1-2x)$ 和 $v_3(1+x)$，后者出现两次。求解

$R^{-1}=\frac{3}{2}\,\mathrm{Cov}(q_3)^{-1}-I/(2v_3)$，并令原始方差为单位方差，可得

$$r=-\frac{3x}{2+x+2x^2},\qquad v_3=\frac{(1+x)(2-x)}{2+x+2x^2},\tag{31}$$

$$R_3=-\tfrac{1}{4}\log(1-2x)-\log[(1+x)(1-x/2)],\qquad \delta=\frac{(1-2x)(2-x)}{2(1+x)^2}.\tag{32}$$

R_3 的公式由高斯归一化因子得出。边缘匹配再次给出最优性。映射 $x\mapsto r$ 从零开始严格递减至 $-1/2$。对于 $s=-r$，令 $f_j(r)=\log[(1-r)^{j-1}(1+(j-1)r)]$。乘积参考分布 $N(0,I)$给出 $4R_3\le 3f_3(r/3)-f_3(r)$，且 $4(I_*/2)=f_2(r)-f_3(r)$。两边右侧之差为 $A(s)=f_2(-s)-3f_3(-s/3)$。直接求导得 $A'(s)/(2s)=-1/[(1+s)(1-s)]+1/[(1+s/3)(1-2s/3)]$。它在 $s=3/7$ 之前为正，此后为负。由于 $A(0)=0$，且$A(1/2)=\log[(3/4)/(49/54)^3]>0$，因此 $R_3<I_*/2$ 在整个开区间内成立。或者，将 (32) 代入 $I_*=\frac{1}{2}\log[(1+r)/((1-r)(1+2r))]$，可得恒等式 $e^{4(I_*/2-R_3)}=(2-x)^3(1+x)^2(1-x+x^2)(2+x+2x^2)/16$。其右侧在 $x=0$ 处等于 1，且在 $(0,1/2)$ 上的对数导数为 $9x^2(1-x)(1+2x^2)/[(2-x)(1+x)(1-x+x^2)(2+x+2x^2)]>0$。此外，由 (32) 可得 $R_3>0$，也可由 $R_3'(x)=9x(1-x)/[2(1-2x)(2-x)(1+x)]>0$ 得出。上述标量割公式给出 $I_*=-\log v_2$ 和 $R_2\ge I_*/2$，因为 $(1+v_2)/2\ge\sqrt{v_2}$。这些事实证明 $L(R)=R_3$，并将必要性归结为定理 9。

贯穿整个族的小支撑集。对于此族，令

$$L_0=\log(1+\xi),\quad g=L_0+e^{-L_0}-1,\quad S=L_0/3+2g/27-R_3,$$
$$C_0=\min(R_3,S),\qquad \theta_s=C_0/R_3.$$

其中 ξ 来自（29）。使用原始简单图，并按照（27）标记边，记 h_0 为该标记数。对于一个四环 F，令 $w(F)$ 为其边的视图重数的乘积分布。展开 $81\,\mathrm{tr}[A(I-A)]=81\,\mathrm{tr}(MM^{\top}-(MM^{\top})^2)$ 得到

$$81\,\mathrm{tr}[A(I-A)]=12m-4\sum_{|F|=4}w(F).$$

事实上，对角贡献为 $(20-2f)\,m$，来自 W 的行范数平方 5 或 3。具有公共列的非对角项来自列类型（2，1）和（1，1，1），其贡献为 $(8-2f)\,m$。具有两个不同列索引的项将每个四环计数四次。

对于 $m\ge 4$，任意一条边所参与的四环总权重至多为三。重数为二的边至多产生一个四环。它的对边重数不能为二，否则会与 $m=2$ 一起闭合成一个连通分量。对于与一个二度顶点关联的单位边，取该顶点的另一条边会迫使其为重边，接着下一个顶点处会有唯一一条向外延伸的单位边。这同样至多产生一个四环，权重至多为二。如果两个端点的度数都是三，总权重就是它们其余邻点之间的 2×2 子矩阵的元素之和。该子矩阵的行和与列和至多为二；若总和达到四，就会使全部六个顶点都达到上限，从而与 $m=3$ 构成一个连通分量。因此，$\mathrm{tr}[A(I-$

$A)]/m \ge (4/27)(1-h_0/m)$，因为每个四环都包含一个标记。对于 $0 \le t \le 1$，$\log(1+\xi t) \ge tL_0 + t(1-t)g$。对前两项之差关于 ξ 求导，得到 $t(1-t)\xi/[(1+\xi)(1+\xi t)] \ge t(1-t)\xi/(1+\xi)^2$,，积分后即得所述结论。将此界应用于 (29)，得到

$$B_G/m \ge (2+f)L_0/9 + (2/27)(1-h_0/m)g. \tag{33}$$

对于端点比较，定义 $F = 2L_0/9 + 4g/81 - R_3$和 $H = R_3/2 - 2g/27$。两者均严格为正。为验证符号，令 $u = x/(1-2x)$, $U = 1+2u, V = 1+3u, W_0 = 2+3u$ 和 $D = 4+15u+18u^2$。代入得 $L_0 = 2\log D - 2\log V - \log W_0 - \log 8, R_3 = \frac{9}{4}\log U - \log V - \log(W_0/2)$，以及 $g' = L_0' 81u^2U^2/D^2$，其导数关于 u。按 u 的升幂排列，D^3UVW_0F'和 D^3UVW_0H'的系数分别为

$$(0, 0, 72, 792, 7785/2, 20115/2, 13851, 9234, 2268),$$
$$(0, 144, 1764, 9153, 103275/4, 170829/4, 83349/2, 22599, 5346).$$

两个函数在零点处均为零。在 $\theta = 1/2$ 时，(33) 与 R_3 乘以 (27) 中的计数系数之差有下界

$$(1+f/2)F + [2/27 - (1+f/2)4/81]g + (h_0/m)H > 0.$$

在 $\theta = 0$ 处，命题 12(a) 的高斯指数界给出 $B_G/m \ge I_*/2 > R_3$。其逐点证明对于 m 是均匀的。对于任意固定图，差关于 θ 是仿射的，这涵盖了 $[0, 1/2]$。对于 $\theta \ge 1/2$，标记项增加了 $(h_0/m)(\theta R_3 - 2g/27) \ge 0$。剩余的差关于 f 是仿射的：在 $f = 0$ 处它至少为 F，而在 $f = 1$ 处它为 $S - \theta R_3$。前面对 $\theta = 1/2$ 与 $f = 1$、$h_0 = 0$ 的比较论证也证明了 $\theta_s \ge 1/2$。这涵盖了 $\theta \le \theta_s$。固定维度裕量提供了均匀的节省，因为 $B_G/m \ge I_*/2$。对于 $m = 2$，3，同样的界以及 $O(n^m)$ 计数就足够了。

大支撑集与共同预算。 倾斜、子集检验和分支界（24）–（25）对每个 x 都成立。每条双视图边将两个二度顶点配对。对于 $k > 0$，三度端口之间的每条由二度顶点组成的极大路径都包含 2ℓ 个内部顶点，其中 $\ell \ge 0$。它连接两侧，并且所有此类路径都可压缩为一个连通的三次二部多重图，每侧有 k 个顶点。它们的 ℓ 之和为 $m-k$。每个平行边组至多包含一条直接边，因为原图是简单图。对于 $k > 3$，不会出现三重边，因为其端点将与该连通分量隔离。

在核心计数中，为所有 m 右侧顶点使用任意互不相同的标签。记录端点类型和端口颜色需要 $O(m)$。按左端口对路径排序。长度选择在对数下需要$O(m)$，内部顶点的标签贡献 $((m-k)!)^2$。每个核心端点都用其右侧标签编码；若边带有标记，则由有界长度的端口序列确定。标记位置、端口序列和右端口颜色在对数下需要$O_\Gamma(m)$。除以 $m!$ 以消除右侧重新标记，因为锚点可以识别原始右侧行。数量至多为

$$\binom{n}{m} \exp[O_\Gamma(m)] k^{3k-h} ((m-k)!)^2/m!,$$

由此得到 (27)，其中核心标记计数为 h。以下所有方案都会移除非桥边，因此 $h \le k+1$。当 $k \le 7$ 时，直接使用未标记的原始计数 $\exp[O(m)]n^m m^k$ 和$B_G/m \ge I_*/2$。

对于 $\theta \ge \theta_s$，只需证明

$$\log \mathcal{I}_G \le hC_0 + (m-k)(R_2 - R_3) + \tau_\Gamma m + O_{x,\Gamma}(1), \qquad \tau_\Gamma \longrightarrow 0. \tag{34}$$

事实上，$t_k \le [f(1-\theta) + \eta f]m\log n + O(m)$。当 $d/\log n = 1/R_3$时，(25)中的主导项乘以计数后均为非正，因为 $\theta R_3 \ge C_0$。将 $d/\log n$ 增加一个固定的正量，至少可节省该量乘以 $[m(1+f)R_3 - hC_0]\log n \ge (m-1)R_3 \log n$的大小，然后再考虑(34)中的余项。先取 η 足够小，再取 Γ 足够大。由于 $m \ge \sqrt{n}$，余项可被一致地吸收。对计数求和时使用$m\log n$的一个统一负倍数。预算取等时，此计算同样适用。

非回溯游走的核心界。令 $H(v,t)$表示一个方差为 v、相关性为 t 的二维高斯，并除以两个裕量的平方根。在勒贝格测度下作为核，$H(v,t)H(v,t') = H(v,tt')$：约去公共裕量后，这就是高斯条件转移的复合。因此，非空路径可压缩为 $H(v_2, y^{2\ell})$。直接边确定端点变量。对于 $v_2 \le v' \le v_3$，上述匹配边缘高斯构造适用于三度因子与方差为 v'的二维路径因子的混合。在共同的物理方差框 $[v', v_3]$ 中，三度顶点处的输入精度满足 $0 \le H \le (v_3/v' - 1)/2$。如果右侧的 H 表示这一上端点，则其输出方差比介于 $1 - 2x^2 H/[1 + (1-x)H] \ge 1/(1+H)$与1之间。所显示的下界不等式由 $1 - 2x^2 + (1 - x - 2x^2)H > 0$推出。在绝对相关性为 t的二元顶点处，对于$(v'/v_3 - 1)/2 \le H \le 0$，输出比为 $1 - t^2 H/(1+H)$，并且位于 $[1, 1/(1+H)]$之间。因此，二者都会将输出方差送回该框中。布劳威尔不动点定理以及(26)之前的熵估计适用。若 $b = v'/v_3$，绝对三度转弯至多为

$$P_b = \frac{x + x^2 z}{1 - x^2 z}, \qquad z = \frac{1-b}{1+b}.$$

这是消去第三个坐标后的相关性，此时该坐标的输入增量达到最大可能值，在两个对角元上增加协方差 $-x^2 z$ 并减去 $x^2 z$。另外两种精度矩阵上的非负增量只会减小其绝对相关性。通过对更新后的二维精度矩阵求逆，二元转动幅度至多为 $t/[1 - (1-t^2)(1-b)/2]$。

三度转向块的算子范数也至多为 $2x$。事实上，对于非负对角增量 D，其协方差 S_1 满足 $S_1 \succeq (1-2x)(I + (1-2x)D)^{-1}$ 且$(S_1)_{ii} \le 1/(1 + D_{ii})$（以方差单位计）。因此，其标准化相关矩阵的最小特征值至少为 $1 - 2x$。其非对角元素为负，因此绝对转向块是非负对称矩阵，由佩龙定理可知，其范数至多为 $2x$。结合二元界（当 $t < 1$ 时小于 1），可得 $C \prec 2I$，见式（26）。在行列式中通过舒尔补消去内部有向路径状态，可得它们的乘积分布，作为其

二元转向幅值作为边传输系数。消去块在 $I-B$ 中是单位三角的，因为每条有向链都是无环的。所有闭合核心游走的长度均为偶数，因此去除三度转向符号后，其权重均非负。在未压缩的原始积分中，每条路径上的二元转向次数都是偶数。将每条边传输系数分配到状态的进出两次遍历中，在转向矩阵两侧分别作为平方根处理，这属于对角相似变换。非负核心矩阵 A_0 及其所有删去边后的主子矩阵的范数均至多为 $2x$。取 q_2 为二元顶点因子，也可以在路径压缩之前进行这一论证，此时使用 $v' = v_2$ 和二元幅值 $y/[1-(1-y^2)(1-v_2/v_3)/2]$。

对于一个标记，移除它的两个定向。根据雅可比恒等式，约化行列式与前一个行列式之比，等于 $(I-A_0)^{-1}$的 2×2 定向主子式。其元素非负，因此该比值至多等于两个对角元素的乘积。每个对角元素都是 1 加上指向该定向的正长度回返的总权重。因此，对这些权重中的每一个给出上界 U_0，就会向半对数修正项收取 $\log(1+U_0)$。删除所有标记后，当围长超过 Γ 时，根据迹级数，剩余修正至多为 $3m\,(2x)^{\Gamma+1}/\,(1-2x)$。

从零到 $x = 0.45069846675$。首先标记平行组中的额外边；如果存在直接边，则保留它。然后按固定顺序插入剩余的简单核心，并像之前一样标记短环闭合边。每个平行标记都有一个已指定的伙伴，因此节省一次端点选择。对于简单标记，有界半径论证给出同样的节省。这两组标记留下一个围长大于 Γ的连通图。对于简单标记，这是通过将其插入前缀见证中的每条被标记路径边替换为其自身较早的见证而得到的。令 $u = v_2/v_3$, $P = P_u$,$Q = y/[1-\,(1-y^2)\,\,(1-u)\,/2]$。在未修改图上，每条非空路径的传输系数至多为 Q^2。平行标记有一个可能的两步回返。对于偶数长度 $j \geq 4$，前 $j-2$ 步至多有 2^{j-2} 种选择，倒数第二个连接的总传输系数至多为 $1+Q^2$，而根节点的传输系数至多为 Q^2。对于简单标记，至多有三个四步返回路径，因为四个返回路径会使一个孤立的 $K_{3,\,3}$达到饱和。每个三步终端重数向量的平方范数至多为 12：路径至多有八条，并且至多有两个重合端点，其重数为二。只有当两条不同的首步通向同一个第二步顶点时，才可能出现重合；从这样的每个顶点出发，至多有一条共享的终端弧。在较长回返的中段应用范数上界 $2x$，即可得到相应的费用

$$C_w = \log\left[1+Q^2\left(P^2+(1+Q^2)\frac{4P^4}{1-4P^2}\right)\right], \tag{35}$$

$$C_s = \log\left[1+3P^4+\frac{12P^6}{1-4x^2}\right]. \tag{36}$$

对于 $0 < x \leq 1/16$，P，$Q \leq 2x$。这由 $y \leq 6x/5$ 和 $u > 9/10$ 在此区间内代入（32）以及 v_2的公式可得。因此 $C_s \leq 52x^4$ 和 $C_w \leq 18x^4$ 由$\log(1+t) \leq t$，而 $R_3 \geq x^2$ 由积分 $R_3'(x) \geq 2x$可得。因此，两项费用均小于$R_3/2 \leq C_0$。

第一个核心界的证明证书。在$[1/16, 7/20]$上，由`scripts/check_d1_equicorr_family.py`在`d1_equicorr_family/rows.csv`中生成的`scripts/checkdequicorrfamily.py`所生成的外向有理包络验证了 $1-4P^2>0$ 和 $C_s,\ C_w<\min(S,\ R_3)$。结合此前的解析界，可验证（34）对所有 $x\le 7/20$均成立。

对于下一段，将每条非空路径的方差从 $v=v_2$ 归一化为 v/a，$u\le a\le 1$。令 $T_\ell=(1-y^{2\ell})/(1+y^{2\ell})$。选择新的相关性 t'，使得 $(1-t')/(1+t')=aT_\ell$。在正交归一和差坐标中，核 $H(v,\ t)$ 的精度矩阵元素分别为 $(1-t)/(2v(1+t))$ 及其倒数比 $(1+t)/(2v(1-t))$，前因子为 $[2\pi v(1-t^2)]^{-1/2}$。因此，该替换保持差方向精度不变，将前因子乘以 $(1+T_\ell)/(1+aT_\ell)$ 的逆矩阵，并使和方向精度减少 $(1-a^2)T_\ell/(2v)$。令$_d_3$ 为三度顶点因子的对角精度 $q_3/\prod\sqrt{\phi_{v_3}}$，以及 $j=1/(2d_3)$，因此

$$d_3=\frac{1}{v_3}\left[\frac{1}{1+x}-\frac{1}{2}+\frac{x}{(1+x)(1-2x)}\right].$$

不同路径的和方向彼此正交，因为不同的非空路径使用不同的端口变量。对于每个这样的和方向 e，恢复其原始项之前的归一化高斯精度矩阵 P_0具有 $e^\top P_0 e=d_3+a^2T_\ell/(2v)$。端点位于不同的三度顶点，且先前的恢复与 e正交。恢复缺失的正秩一增量。根据行列式引理和 $e^\top P_0^{-1}e\ge 1/(e^\top P_0 e)$，所得对数积分至少减少 $\tfrac12\log[(1+jT_\ell/v)/(1+ja^2T_\ell/v)]$。因此，每条路径的净代价至多为

$$G_\ell(a)=\log\frac{1+T_\ell}{1+aT_\ell}-\frac12\log\frac{1+jT_\ell/v}{1+ja^2T_\ell/v}.$$

对于 $j/v\le 1$，其至多为 $\ell G_1(a)$。事实上，将 $T=T_\ell$视为变量时，右侧等于从 a 到 1 的积分，积分函数为 $(1-bz)T/[(1+zT)(1+bz^2T)]\,dz$，其中 $b=j/v$。该被积函数关于 $T\in[0,1]$递增且为凹函数，因为其导数的分子 $(1-bz)(1-bz^3T^2)$ 非负且递减，而分母为正且递增。由于 $T_\ell=\tanh(\ell\log(1/y))$ 递增且为凹函数，$G_\ell(a)$ 关于 ℓ 是凹函数，并且在零处取值为零。由于 $T_1=v$，可得

$$G_1(a)=\log\frac{1+v}{1+av}-\tfrac12\log\frac{1+j}{1+ja^2}.$$

选择 a，使 $G_1(a)\le\Delta=R_2-R_3$，同时最小化 $[u,1]$ 中的 a。给定 $H_c=4v^2e^{2R_3}/(1+j)$ 和 $a_c=(H_c-1)/(v+\sqrt{v^2+(H_c-1)(v^2-H_cj)})$，这是$a=\max(u,a_c)$：等式 $G_1(a_c)=\Delta$ 求解 $(1+a_cv)^2=H_c(1+ja_c^2)$，且 G_1 在 a 中递减。记作 $\sigma=\Delta-G_1(a)$。拟合后的转弯具有 $P=P_{u/a}$，且非空路径的传输系数至多

$$W=\frac{t}{1-(1-t^2)(1-u/a)/2},\qquad t=\frac{1-av}{1+av}.$$

这由匹配的二维界和 $T_\ell\ge v$ 得出。每条标记的弱边（包含一条非空路径的边）都使用一对不同的二度顶点。除了C_0之外，它还可以将 σ 计入该顶点对。

下面详细说明如何界定标记的简单边。在每个平行边组中移除一条弱边后，首先移除这样的剩余配对对象：其端点仍位于与原组构成的外部四环中。这样的环不可能包含另一个平行边组：相邻的组会使一个顶点上的边数超限，而相对的组会将四个顶点彼此隔离。每个被移除的配对对象的两个端点，在剩余简单图中的度都至多为二。然后反复删除长度不超过 $\Gamma \geq 14$的简单环上的一条边。优先选择不属于任何四环的边，其次选择恰好属于一个四环的边。如果两者都无法选择，则可以如下所述，在一个缺少一条边的 $K_{2,\,3}$ 或 $K_{3,\,3}$ 中选择一条属于两个四环的边。一条属于三个四环的边，会在其两个端点及其四个邻点上形成不完整的 $K_{3,\,3}$ 。它通向外部的两个端口恰好位于缺失边的两个端点处，而这些位置上的任何外部边都不属于四环。缺失边端点处的一条边属于两个四环。如果没有边属于三个四环，则取四环并集中的一个连通分量。若其中每个顶点的度都是三，那么将四环视为面时，每条边恰好属于两个面，并且在每个顶点处，每一对边都属于一个面。面邻接图是连通的，因此从面到边的关联边界映射在 $\mathbb{F}_2$ 上的秩比面数少一。其像包含于环空间中。写成 $E = 3V/2$ 和$F = E/2$ 可得 $V - E + F \leq 2$，从而得到 $V \leq 8$。这样的饱和连通分量被 $k > 7$ 排除。因此，该并集中存在一个度为二的顶点。其两条关联边属于两个共同的四环，所以它的两个邻点有三个共同邻点，形成 $K_{2,\,3}$。从该顶点引出的任何边都不属于四环，否则一条相邻的内部边将至少属于三个四环。在任一三度中心处选择一条内部边。

对于每次删除，其见证环中的其他边会给出一个长度至多为 $\Gamma - 1$的端口序列。因此，即使采用自适应顺序，（27）中的端点编码仍然有效：从未标记边开始，并恢复任何其见证已可被跟随的边。逆删除顺序保证在核心恢复之前始终至少存在一个这样的见证。被删除的平行边可以根据其现已恢复的配对对象进行解码。

对于一条标记的简单边，令 $(a_4,\ b_6)$ 表示其四环和六环的数量，令 d_8 给出其八步返回数量的上界，并令 K_4 给出其正向和反向四步终端重数范数乘积的上界。长度为六的简单二分图非回溯返回必定是一个环，而长度为八时，它要么是一个八环，要么是两个四环的串联。某个顶点的重复会将两个闭合部分分隔开，且每个部分的长度至少为四。所有长度至少为十的返回中间部分的范数上界给出

$$U_s = a_4 P^4 + b_6 P^6 + d_8 P^8 + K_4 P^8 \frac{4x^2}{1 - 4x^2}.$$

对于所选的两个图，分别取 $a_4 = 2$，使用 $(b_6,\ d_8,\ d_{10}) = (0,\ 6,\ 16)$ 和 $(2,\ 10,\ 26)$，并以尾项 $d_{10}P^{10} + 27P^8(4x^2)^2/(1 - 4x^2)$ 替代最后一项。对于这些计数，外部端口不在任何短环上。若游走经由不同端口返回，则沿图内一条长度至多为三的路径闭合后会形成短环。经由同一端口的游走至少需要六步，其中包括一个外部四圈。长度为十时，它只能插入两个内部根四圈之一。在 $K_{2,\,3}$ 中，每个内部四圈上至多有两个外部端口；在不完整的 $K_{3,\,3}$ 中则有一个。

每个外部四圈最多有四种定向，并可从两个可用外部弧中选择，因此外部返回数分别为 16 和 8。内部计数通过迭代有向非回溯邻接得出。对于一般的一对 $(a_4,\ b_6)$，K_4 和 d_8 的计算枚举二元 2×2 臂邻接、每条臂的共享或私有邻居，以及满足三度条件的剩余中心边。根据第三步入边计数 c_{uv}，v 处的四步终端重数为 $\sum_w c_{wv} - c_{uv}$；缺失的外部邻居可通过补全其端口来计数，并通过对齐不同臂的新端口使平方和有界。八步返回由两个三步前缀和一条中心边串接而成，省去会反转任一相邻弧的连接。若两个前缀端点都是新端口，则计数为 $c,\ d$ 的中心边贡献为 $cd \le (c^2 + d^2)/2$。这些规则给出了有限且穷尽的上界枚举，涵盖所有合法的局部图，也可能包括额外的补全。通过对相应的联合范数界取平方根，可得两个 $a_4 = 2$ 情形的 $K_4 \le 27$。

对于存在外部四环的弱根，记为 $U_w = WP^2 + W(1+W)P^4 + W(8+2W^2)P^6/(1-4x^2)$。前两项使用其平行伙伴和外部四环。在三步前缀中，最多有八条路径。其中两条在内部重复经过弱根，并带有额外因子 W。其余六条最多只能共享一个终端状态，重数为二。事实上，在依次经过平行伙伴和侧端口后，下一种备选方案是到达较近的外部顶点；而首先经过另一侧端口的前缀，只有在两步内到达同一个外部顶点时，才能与这些路径汇合。如果它们的第二步改为使用一对外部平行边，则从其公共端点出发的出弧中至多有一条能够重合，并且由于四个核心顶点届时都已饱和，该端点不可能是较近的外部顶点。因此，对每种定向，平方范数至多为 $W(8 + 2W^2)P^6$。剩下的伙伴有一步初始步骤，至多一次四步返回和一次六步返回（后者经过两个备用外部端口），并且根据此前的三步范数界，其四步终端平方范数至多为 $12P^8$。所以 $U_d = P^4 + P^6 + 12P^8/(1 - 4x^2)$计入该伙伴。

对于一个没有外部四圈的弱根，令其两个侧邻居外侧的两条边在划分时称为 I 型，在并联时称为 II 型。在两步或四步返回路径中，只能使用平行根对。在六步中，两个侧邻居之间还可以有一次三步交叉，或者在其中一个邻居处形成一个二边环并返回根节点。对于 I–I 型，中间交叉至多贡献三条边，因为第四条边会使八顶点连通分量达到饱和。对于 I–II 型，一旦计入平行端口的选择，中间至多有一条边，总权重为 $1 + W$，另加侧向二环的 $2W$。对于 II–II 型，$k > 7$对应的中间边不存在，两个侧向二环至多贡献 $4W$。令 $N_{\mathrm{I}} = 6 + 2W^2$、$N_{\mathrm{II}} = 3 + 3W + 4W^2$。朝向 I 型侧边的定向，其三步范数平方至多为 $WP^6 N_{\mathrm{I}}$：八个终端状态彼此不同，且有两条路径产生重复根因子。在 II 型侧，权重为 $\alpha \le 1$、$\beta \le W$的平行边在提取公因子 $\sqrt{W}P^3$后，给出的终端权重至多为 $\alpha\sqrt{\beta},\ \beta\sqrt{\alpha},\ \alpha + \beta$。其余项的平方和至多为 $2 + 2W^2$，从而得到

WP^6N_{II}。令 $b_{\mathrm{I},\ \mathrm{I}} = 3$、 $b_{\mathrm{I},\ \mathrm{II}} = 1 + 3W$、 $b_{\mathrm{II},\ \mathrm{II}} = 4W$。所得预算使用

$$U_{ij} = WP^2 + W^2P^4 + WP^6 \left[W^2 + b_{ij} + \sqrt{N_iN_j}\frac{4x^2}{1-4x^2} \right].$$

这些不等式计算的是前置删除之前的游走，而删除操作只会使非负权重减小。所需预算为 $\log(1+U_s) \le S$、$\log(1+U_w) + \log(1+U_d) \le 2S+\sigma$ 以及 $\log(1+U_{ij}) \le S+\sigma$。

短环界的有限检查。上述整数枚举记录在 scripts/check_d1_cycle6_motifs.py 中，并记录于 d1_cycle6_motifs/rows.csv 及其 square_free/rows.csv 和 capsules/rows.csv 子路径中。一般表格用于单四环行。该表格将 K_4^2（两个四步平方范数的乘积分布）限制为 36^2。该上限仅对三个局部图生效，且它们均满足 $(a_4, b_6) = (0, 0)$。在每个局部图中，一个或两个定向会到达四个第二层顶点：每个顶点通过一条路径到达，并各自带有两个未探索端口。该脚本通过对齐这些端口来补全未探索端口，使八个单位到达量中的四个进入同一个新顶点，从而得到平方范数 $4^2 + 4^2 + 8 = 40$。一个新顶点至多有三个端口，而同一顶点的两个端口会通向不同的新顶点，因此这些到达量至多划分为 $3 + 3 + 2$，平方范数至多为 $3^2 + 3^2 + 2^2 + 8 = 30$。另一种定向的平方范数至多为 36，因此这些乘积分布至多为 $36 \cdot 30 < 36^2$。scripts/check_d1_short_cycle_budget.py 中 d1_short_cycle_budget/rows.csv 的区间证书检查全部三个预算不等式，$S < R_3$、$j/v < 1$、校准域以及 $W < 1$ on $[7/20, 0.45069846675]$。对于区间的最后部分，使用负导数及其在 0.45069846675 处的正下端点值来界定活跃的 I–I 间隙；所有其他间隙均通过闭子区间界验证。每个额外的 σ 费用仅使用一次。因此，路径与标记费用之和验证了整个区间上的 (34)。

标记边的电学比较。 对于其余区间，使用 $a = u$，因此路径方差为 v_3，三度转弯幅值精确为 x，且非空路径传输至多为 $w = (1 - v_2^2/v_3)/(1 + v_2^2/v_3)$。由于闭游走级数的各项均为非负数，我们可以在上界中增加核心传输系数。所有普通核心边的传输系数均可设为一。在 $[9/20, 239/500]$ 上，对任何并联组标记保留衰减 w。对于 $x \ge 239/500$，我们也可以将这些传输系数增加至一。

从每个平行边组中选取一个代表元；在可能的情况下，优先选择直接边。先插入代表元，再插入没有平行伙伴的边，最后插入额外的平行边；每个阶段内均固定左端口顺序。代表元构成一个匹配，且全部予以保留。保留第一条形成环的边。此后，标记并跳过任何端点间距离不超过 $\Gamma - 1$ 的边（距离在保留前缀中计算）。该过程保持图连通，至少保留一个环，且环长至多为$\le \Gamma$；每个标记在插入时都有一个有界半径目标。编码每个左端口的插入阶段需要 $O(k)$ 位，因此（27）仍然成立。从任一被标记的端点出发，沿最短保留路径到达一个环，再沿该环反复绕行，即可在其腔体中提升得到一条保留射线。每个

端点保留一条或两条边。如果该标记有平行伙伴，则保留该伙伴；由连通性和 $k>7$ 可知，至少一个端点还保留一条额外的边。因此，其端点类型为 (1，2) 或 (2，2)。

为普通边赋予弧权重 $\alpha_e = x$ ，为衰减边赋予弧权重 $\alpha_e = xw$ 。对于无权弧非回溯矩阵 $B = D^{\top}T - \mathsf{W}$，其中 T， D 分别为起点和终点关联矩阵，W 为弧反转矩阵，令 $\mathcal{A} = \mathrm{diag}(\alpha_e)$。将 $I + \mathsf{W}\mathcal{A}$ 从$I - B\mathcal{A}$中因式分解。利用 $(I + \mathsf{W}\mathcal{A})^{-1} = (I - \mathsf{W}\mathcal{A})(I - \mathcal{A}^2)^{-1}$ ，并在剩余行列式中交换矩形矩阵乘积，得到

$$\det(I - B\mathcal{A}) = \prod_{e} (1 - \alpha_e^2)\, \det M_0. \tag{37}$$

这里，$(M_0)_{vv} = 1 + \sum_{e \ni v} \alpha_e^2/(1 - \alpha_e^2)$，而边 uv 使两个非对角元素各增加 $-\alpha_e/(1 - \alpha_e^2)$；并联贡献相加。将 M_0 除以 $x/(1 - x^2)$，得到一个图拉普拉斯矩阵加上对角接地电导。单位边的电导为 1。衰减边的电导为 $c = w(1 - x^2)/(1 - x^2 w^2)$，并为两个端点各增加接地电导 ζ，其中

$$\mu = x^{-1} + 2x - 3 > 0, \qquad \zeta = \frac{(1 - x)(1 - w)}{1 + xw}.$$

每个顶点都具有基图接地电导 μ。这些公式是通过从对角元中减去总入射电导得到的。

为清晰起见，半对数环修正可如下与有限提升进行比较。在 N重提升中，为每条边在其端点的 N 个副本之间赋予一个独立的均匀匹配。揭示任意固定半径的球需要数量有界的匹配，并且在任意根节点处的碰撞概率为 $O(1/N)$。存在一个提升序列，使得每个这样的球在每条边和每个顶点的副本中，有 $1 - o(1)$ 的比例变得类树。对于非回溯迹展开中的固定项，提升后的迹除以N 因而趋于零。其余迹尾项除以 N，利用 $2x < 1$可得到一致有界。因此，归一化的提升环修正趋于零。（37）中的局部边与顶点因子会与有限图因子相消。因此，原始环修正等于提升后的对数行列式除以 N 再减去原始对数行列式所得极限的一半，其中每个行列式均取自接地拉普拉斯矩阵。

当t 在 $[0,1]$, 中变化时，将所有标记电导同时从零插值至其最终值，同时保留所有接地电导和其他边。在 $t = 0$ 处，这个半对数差至多为$\tau_{\Gamma} m + O_{x,\,\Gamma}(1)$，其中 $\tau_{\Gamma} \to 0$。为证明这一点，用次数为 j_0的多项式在公共谱区间 $[\mu, \mu + 6 + \zeta]$ 上对对数函数作一致逼近。该多项式的对角元仅取决于 j_0邻域，其中顶点接地电导作为局部标记。取 $\Gamma \geq 4j_0 + 4$ 可确保在保留图中，除可能存在的一个短环附近的常数距离范围内，这些球邻域都是类树的。在所有类树球邻域上，多项式对角元相同，而每个例外对角元都有界。一致逼近可以做到任意精确，由此得证。对于插值中斜率为 a 的每条边，半对数差的导数等于 $a/2$ 乘以

提升图与有限图的有效电阻之差。事实上，沿 $a(e_u-e_v)(e_u-e_v)^\top$ 方向求 $P(t)$的对数行列式的导数，得 $a(e_u-e_v)^\top P(t)^{-1}(e_u-e_v)$。在任何提升中，有效电阻都不超过 $2/\mu$ ，因为 $P(t)\succeq \mu I$。以下估计适用于任意深度的有限树邻域，只需保留其中的边及其接地电导。基尔霍夫约化，或将单位流延伸至公共接地，都能用这些测试子网络的电阻给整个网络的电阻作上界。先在深度固定时取提升极限，再令深度增大。上界 $2/\mu$ 允许将上极限移入 t积分号内。

上的加权电阻界 $[9/20, 239/500]$。每个顶点至多接触一条衰减边。记 $\mathsf{ser}(c,\ h)\ = ch/(c+h)$, 其中 $\mathsf{ser}(0,\ 0)\ = 0$ 由连续性得到。令 H_* =$(1-x)/x$、 $\xi_0 = 3/500$ 和 $\mu_0 = \mu - \xi_0$。由该区间直接代入可得 $\mu_0 > 0$、 $H_* \geq 109/100$、$w \leq 17/20$ 和 $c \leq 4/5$ 。对于 $0 \leq h \leq H_*$,

$$\zeta + \mathrm{ser}(tc, h+\zeta) \geq \mathrm{ser}(t,h) - \xi_0. \tag{38}$$

事实上，$\mathsf{ser}(t,\ h)$−$\mathsf{ser}(tc,\ h+\zeta)$ 关于 h 单调不减，因为 $c \leq 1$。在 H_* 处，令 $V = H_*+\zeta$，并利用 $\mathsf{ser}(1,\ H_*)\ = 1-x$ 以及 $\mathsf{ser}(c,\ V)\ = 1-x-\zeta$，小于 ζ 的差为

$$\frac{\zeta(1-t)(2ct - H_*V)}{(t+H_*)(tc+V)}.$$

其正部至多为 $\zeta(2c-H_*V)_+^2/(8cH_*V)$ ，这是通过最大化二次分子得到的。若其非零，$w \geq c \geq H_*^2/2 \geq a_* = (109/100)^2/2$。现在 $\zeta \leq (11/20)(1-a_*)/(1+(9/20)a_*)$, $2c - H_*V \leq 8/5 - 2a_*$ 以及 $8cH_*V \geq 16a_*^2$。它们的比值上界严格小于 ξ_0，从而证明了（38）。

令 $H\ =\ \mu_0 + 2\,\mathsf{ser}(t,H)$ 和 $Y\ =\ \mu_0 + \mathsf{ser}(t,H)\ + \mathsf{ser}(1,Y)$ 表示这些正解。它们至多为 H_*，这是在 $t=1$ 处、基图质量为 μ时的解。第一条递推通过比较从叶节点电导 μ_0开始的有限深度树，为一般腔体给出一个下界。若一条衰减边进入腔体，其根节点便有额外的接地电导 ζ，且两条出边均为普通边。若有一条出边是衰减边，其子节点腔体会多出 ζ，而根节点额外的 ζ足以弥补式（38）中的差，所用质量减量至多为 ξ_0。因此，普通入边和衰减入边对应的下界分别为 H 和 $H+\zeta$ 。沿着一条保留射线，指定的出边电导为 1，因此第二条递推分别给出 Y 和 $Y+\zeta$。单调的有限深度迭代收敛到这些正解。这可通过从叶节点质量开始逐步增加深度、利用界 H_*并求解标量二次方程得到。

普通根边端点若有一条保留边，其腔体中的接地电导至少为 $g_1 = Y$；若有两条，则至少为 $g_2 = \mu_0$+串联电导$(1,\ Y)$+串联电导$(1,\ H)$。若它接触到一个衰减标记，其余两条边便是该标记和它的保留伙伴边。该端点是一个保留叶节点，因此删除它后，伙伴边的另一端仍与一个环相连。因此，射线可以越过伙伴边继续前进，而无需回溯。跨过被略过的弱边后，任何保留射线都可用。因此，使用 $g_p = \mu_0 + \zeta +$ 串联电导$(1,\ Y)\ +$ 串联电导$(tc,\ Y+\zeta)$。一个

端点保留计数为 r_0，s_0 的衰减根使用下界 $g_{r_0}+\zeta$，$g_{s_0}+\zeta$。对于腔体下界 g，h，提升后的半电阻乘以电导斜率 a 至多为

$$\frac{a}{2at+2/(g^{-1}+h^{-1})}.$$

经过公共接地的路径，是两个空腔串联后与根边并联。

对于有限普通根，在其端点处分别设置电势 a_0，$-a_0$，在相应的非根邻居集上设置电势 b_0，$-b_0$，其他位置均为零。设 p_0 为与不同衰减标记相连的根端点数，d_0 为两个根端点到其其他邻点的电导之和的一半。普通标记要求存在一个保留环。额外的平行边随后出现，每条边都在距离为一处有其保留伙伴，因此每条边都带有标记。这使非根电导为 $1+t$，或在任一根端点处为 2 或 $1+ct$。邻接顶点至多有 $4-p_0$ 个，剩余邻居端口至多有 $8-3p_0$ 个。两个邻居集之间至多有 j_{p_0} 条边，其中 $j_0=3$ 且 $j_1=j_2=1$。四条跨接边要求每侧各有两个不同的邻点，因为重复的邻点至多有一个未使用端口。这四条边会使六个顶点饱和，并将它们隔离在核心中。每条跨接边至多贡献 $4b_0^2$，其中 $2b_0^2$ 已由其两个端口涵盖。邻点处的接地质量至多为 $\mu+\zeta$。因此测试能量至多为

$$2\big[(\mu+p_0\zeta/2+2t+d_0)a_0^2-2d_0a_0b_0+(d_0+A_{p_0})b_0^2\big],$$
$$A_{p_0}=\frac{(4-p_0)(\mu+\zeta)+8-3p_0+2j_{p_0}}{2}.$$

对 b_0 进行优化，可通过将根节点电势差的平方除以能量，得到有限图半电阻下界 $[\mu+p_0\zeta/2+2t+d_0-d_0^2/(d_0+A_{p_0})]^{-1}$。对于衰减并联根，每个端点都有一个外部邻居，最多有四个总剩余端口和两条交叉边。利用 $d_0=(r_0+s_0-2+(4-r_0-s_0)t)/2$，对电导为 $1+ct$ 的根节点对重复试验，可得到半电阻乘以斜率的下界 c

$$\frac{}{\mu+\zeta+2(1+ct)+d_0-d_0^2/(d_0+\mu+\zeta+4)}.$$

加权比较的证书。 向外有理区间计算脚本 scripts/check_e1_weighted_parallel.py，以及 e1_weighted_parallel/rows.csv 中的行，为全部六种无序普通端点情形和两种并联类型界定了前述电阻界之差的积分。提升后的上界根据 t 的电导单调性而减小。有限下界也会减小，因为 $d_0-d_0^2/(d_0+A)=d_0A/(d_0+A)$ 随 d_0 和 A增加而增加。对前者使用左端点、对后者使用右端点，即可在每个闭参数单元上得到达布上界。这证明了普通电荷 $C_{\mathrm{ord}}<C_0$ and

$$0\le(C_{\parallel}-C_0)_+<R_2-R_3-G_1(u).$$

普通不等式和并联不等式的最小已证间隙分别大于0。0127629695和 0。0002325468。每个组至多有一个并联标记，且每个标记使用一对不同的二度顶点。因此，加权区间上满足预算（34）。

在 $[239/500, 999/2000]$ **上进行单位电导比较。** 将较弱的传输增大为 1，并保持相同的保留规则。移除 ζ 以及腔体计算中的质量减量。现在令 $H = \mu + 2\mathrm{ser}(t,\ H)$ 和 $Y = \mu + \mathrm{ser}(t,\ H) + \mathrm{ser}(1,\ Y)$，并设 $g_1 = Y$, $g_2 = \mu + \mathrm{ser}(1,\ Y) + \mathrm{ser}(1,\ H)$。通过相同的有限深度论证，它们给出了腔体电导的界。对于端点对（$r_0,\ s_0$），定义

$$T_{r_0 s_0} = [2t + 2/(g_{r_0}^{-1} + g_{s_0}^{-1})]^{-1}.$$

对于普通根节点，在有限图检验中使用 $d = (r_0 + s_0 + (4 - r_0 - s_0)t)/2$。即使一对相邻边并联，也至多有四个邻点和八个剩余端口可用。当局部邻点各不相同时，四条交叉边会使一个六顶点分量达到饱和；而重复邻点只有一个剩余端口。这些端口的能量上界为 $14b_0^2$。对于平行根节点，只有两个外部邻点。使用 $d' = (r_0 + s_0 - 2 + (4 - r_0 - s_0)t)/2$，剩余能量的上界为 $8b_0^2$。由此得到的两个有限图半电阻下界为

$$\begin{aligned} F_{r_0 s_0} &= \left[\mu + 2t + d - d^2/(2\mu + d + 7)\right]^{-1}, \\ F^{\parallel}_{r_0 s_0} &= \left[\mu + 2(1 + t) + d' - (d')^2/(\mu + d' + 4)\right]^{-1}. \end{aligned}$$

它们可通过优化与加权情形中相同的试探势得到。普通电荷和平行电荷分别是 $T_{r_0 s_0} - F_{r_0 s_0}$ 和 $T_{r_0 s_0} - F^{\parallel}_{r_0 s_0}$ 在 $t \in [0, 1]$ 上的积分。对 $[239/500, 12/25]$ 上的路径电荷，使用 $G_1(u)$。对于 $x \geq 12/25$，只需使用前因子电荷 $A_1 = \log[(1 + v_2)/(1 + v_2^2/v_3)]$，无需使用秩一递减量。事实上，其一般的 ℓ-对版本 $A_\ell = -\log[(1 + u + (1 - u)y^{2\ell})/2]$ 根据对数和指数函数在 ℓ 上是凹函数，并且在 $\ell = 0$ 处为零。

单位电导比较的证书。scripts/check_e1_pure_trivalent.py 生成的 e1_pure_trivalent/rows.csv 输出验证了普通电荷在整个 C_0 上都小于 $[239/500, 999/2000]$。对于平行电荷 $C_{\parallel}$，同一输出验证了在 $[97/200, 999/2000]$ 上的 $A_1 + (C_{\parallel} - C_0)_+ < R_2 - R_3$。在 $[12/25, 97/200]$ 上，这个不等式由 e1_parallel_endpoint/ rows.csv 得出，该文件由 scripts/check_e1_parallel_endpoint.py 生成。在 $[239/500, 12/25]$ 上，以 $G_1(u)$ 代替 A_1 时的不等式见 e1_path_credit/ rows.csv，该文件由 scripts/check_e1_path_credit.py 生成。其仅路径行在使用到的地方证明了 $G_1(u) < R_2 - R_3$。同样适用达布比较：腔体电导随 t 和 μ 增加，且每个有限分母都增加。脚本以向外舍入的方式，在闭 x 单元格中界定两个端点。此处，通过 $2v_3 d_3 = (1 - x + 2x^2)/[(1 + x)(1 - 2x)] \geq 3/[5(1 - 2x)]$ 和 $v_2/v_3 \geq \sqrt{(1 - 2x)/3}$，保证秩一调整路径电荷凹性的条件 $j/v_2 \leq 1$ 得以满足，因此

$2v_2d_3 \ge 1$ 在 $[9/20, 1/2)$。它同样适用于加权区间。与之前一样，最多$m-k$ 个平行边标记会消耗每条路径剩余的松弛量，得到 (34)。

以奇异边界为终点的区间。对于 $x \ge 499/1000$，保留射线给出与端点类型无关的均匀比较，并且下述路径成本界成立。该证书确定了支撑预算 S 所施加的更严格截断。对于每个 $t>0$，一条所有电导均至少为 t 的侧分支，在深树极限下提供至少为 $t/2$ 的接地电导。事实上，沿深度为 D 的分支等分流量的单位流，其能量至多为 $2/t + 1/(\mu 2^D)$，其中计入了叶节点的接地电导。沿一条保留射线，在每个顶点处使用这些分支，可得到至少为 Z 的腔体电导，并满足 $Z = t/2 + Z/(1+Z)$，因此 $Z = (t/2 + \sqrt{t^2/4 + 2t})/2$。两条射线给出的提升后的半电阻至多为 $1/(Z+2t)$。对于有限图，令 $b = e_u - e_v$ 和 $M_t = \mu I + \mathcal{L}_t$，其中 $\mathcal{L}_t$ 为插值拉普拉斯矩阵。于是 $b^\top M_t b \le 2\mu + 4t + 6$：在剩余的四个根端口中，至多有两个使用同一个平行伙伴，合计贡献至多为四，其余端口各自贡献至多为一。由柯西–施瓦茨不等式可得 $b^\top M_t^{-1} b \ge 4/(b^\top M_t b)$。因此，边电荷被界定为

$$\begin{aligned} C_{\mathrm{end}} &= \int_0^1 \left[\frac{1}{Z + 2t} - \frac{1}{\mu + 3 + 2t} \right] dt \\ &= \tfrac{9}{10}\log 6 - \tfrac{1}{2}\log 2 - \tfrac{1}{2}\log\frac{5+\mu}{3+\mu}. \end{aligned}$$

第一个积分由 $t = 2Z^2/(1+Z)$ 求得，它将 $Z \in [0,1]$ 映射到 $t \in [0,1]$，并在 Z 中给出被积函数 $-1/[2(1+Z)] + 9/[2(1+5Z)]$。

每个二度节点对的路径前因子成本至多为 $\log 2$。记 $e = 1 - 2x$，则在此区间上有 $v_2 = (1-r)\sqrt{\delta/(1+r)} \le (5/4)\sqrt{e}$，且 $(1+x)(1-x/2) > 11/10$。因此，利用 $e \le 1/500$ 和 $500(11/25)^4 > 16$，有 $R_2 - R_3 \ge -\tfrac{1}{4}\log e + \log(11/25) > \log 2$。此外，$\delta < 1/16$ 意味着 $L_0' = [(1-\delta)/(1+\delta)](-\delta'/\delta) > (15/17)2/(1-2x)$。结合 $g' > 0$ 和 $R_3' < 1/[2(1-2x)]$，可得 $S' > 0$。由于 μ 递减，C_{end} 递减，而 R_3 递增。

末端比较的证书。 `d1_equicorr_endpoint/rows.csv` 包含来自 `scripts/` 中 `check_d1_equicorr_endpoint.py` 的有理包围。它们验证了 $R_3 > C_{\mathrm{end}}$ at $499/1000$，并将 $S - C_{\mathrm{end}}$ 的唯一根夹在 (0.49948994483,0.49948994485) 内。特别是，对于 $0.49948994485 \le x < 1/2$，我们有 $C_{\mathrm{end}} \le C_0$。结合这些结果，路径预算覆盖了该范围。

保留地质量的另一项比较。对于 $x \ge 4987/10000$，在覆盖树腔体中保留地质量 μ，使用上文的单位电导递推关系 H和 Y。提升后的半电阻至多为 $1/(Y+2t)$。结合相同的有限图下界，边电荷为

$$C_\mu = \int_0^1 \frac{dt}{Y(\mu,t) + 2t} - \tfrac{1}{2}\log\frac{5+\mu}{3+\mu}.$$

对于 $e \le 13/5000 < (22/75)^4$，，前述路径界给出 $R_2 - R_3 >$对数$(3/2) >$ 对数$(1 + v_2) \ge A_1$。在 $[4987/10000, 49949/100000]$ 上，来自 scripts/ 中 certify_e1_negative_endpoint_mass.py 的 e1_negative_endpoint_mass/rows.csv 中的行，对 certify $C_\mu <$ min$(R_3,\ S)$。在每个 x 单元格中，关于 $t_i = (i/768)^2$ 的左侧求和，根据 Y在 t 和μ中的单调性，在最小的 μ 处界定上方的积分。对被减的对数使用最大的 μ，对递增的预算 $R_3,\ S$ 使用左侧 x 端点。右端点 $49949/100000 > 0.49948994485$ 与末端比较产生额外的重叠度。

额外的重叠度检查。在 $[9/20, 0.46544977415]$ 上，还可以使用 $a = u$ 进行短环检查，转向上限为 $P = x$ 和 $W < 153/200$。对于二分裂根，令 $b = 0$,1,2,3 表示两个外部邻居对之间的边数。在八步计数中，按返回是否留在根节点对内、是否穿越外部图一次，或是否返回各自的外部端点进行分组，可得

$$\begin{aligned} U_b &= Wx^2 + W^2x^4 + Wx^6(W^2 + b) \\ &\quad + Wx^8[W^3 + (1 + 2W)b + E_b] + \frac{Wx^{10}M_b(W)}{1 - 4x^2}. \end{aligned}$$

这里，E_b 给出了五步穿越与四步外部返回联合计数的上界。通过枚举满足每个外部邻居各有两个自由半边条件的公共二乘二邻接重数，得到 $b = 0, 1, 2, 3$ 的取值分别为 20, 22,14, 8。按 $\sqrt{W}x^5$ 缩放的五步终端范数平方由 $M_b(W)$ 给出上界。其计算方法是将自由端口按端点相同或不同分组；对于任何尚未探索的邻域，则按权重递减顺序将剩余到达分成三元组。对于一组三个入射权重 z_i，输出平方范数为 $(\sum_i z_i)^2 + \sum_i z_i^2$。根据主序，排序后的组可使各组总和的平方之和最大。这便得到了最后一项：它来自两个五步终端向量，以及中间步骤中的压缩 $2x$ 。

对应的整数表格位于辅助文件的 d1_weak_eight_fork/、d1_fork_five/ 和 d1_weak_eight_parallel/ 子目录中的 rows.csv。它们由 scripts/ 中的 check_d1_weak_eight_fork.py、check_d1_fork_five.py 和 check_d1_weak_eight_parallel.py 生成。平行侧表还会统计围绕一个平行外部对的八步游走，并界定单平行侧或双平行侧情形下的四步范数。脚本 scripts/check_d1_weak_eight_budget.py 及其 d1_weak_eight_budget/rows.csv 会验证弱根电荷相对于$S + \sigma$的结果，以及配对平方电荷相对于 $2S + \sigma$的结果。对于简单根，脚本 scripts/check_d1_simple_ten.py 会记录截至十步的计数，以及 d1_simple_ten/rows.csv 中的五步范数乘积。这项检查细化了删除规则：在恰好位于一个四环上的边中，首先删除位于最少六环上的边。因此，当某个局部图中的另一条边经证明位于一个短环上且不在任何四环上，或位于一个四环上且位于更少的六环上时，该脚本会省略位于一个四环上的根。对于其余此类根，它会读取

表格 d1_cycle6_motifs/single_square/rows.csv；check_d1_cycle6_motifs.py 会使用选项 --single-square-priority 写入该表。上述覆盖论证既不使用这一细化，也不使用该表。其区间输出 d1_simple_ten/intervals/rows.csv 会验证简单电荷相对于 S 的结果，其中活跃根被限制在 0.46544977415 和 0.46544977424 之间。这些检查与加权电阻论证存在重叠，而后者已经提供了跨越该区域的桥梁。

拼接这些区间。 短环界涵盖 $0 < x \le 0.45069846675$，加权电阻比较涵盖 $9/20 \le x \le 239/500$，单位电导比较涵盖 $239/500 \le x \le 999/2000$。终端比较涵盖 $0.49948994485 \le x < 1/2$，因此最后两个区间在 $[0.49948994485, 999/2000]$上重叠。保质量估计提供了一个独立的重叠。每个比较论证本身都提供（34），并带有各自的标记编码。因此，小支撑集估计和大支撑估计覆盖了每个固定 x的所有候选项。真实表通过子集检验的概率为 $1 - o(1)$，因此合格估计量能够成功。根据贝叶斯最优性，全数据最大似然至少具有同样的成功概率。至此，关于 $(-1/2,\ 0)$ 的固定边缘充分性得证。 □

E.6. 具有相同正相关性的三个视图

定理 13(i). 对于相关性相等的三个视图 $0 < r < 1$，令

$$L = R_* := \frac{I_*}{2} = -\frac{1}{4}\log\frac{(1-r)(1+2r)}{1+r}.$$

若 $\liminf d/\log n > 1/R_*$，则最大似然估计器以趋近于一的概率恢复所有排列。若 $\limsup d/\log n < 1/R_*$，则任何估计器的成功概率都趋近于零。根据命题 R_* 12(c) 中的非负相关性计算，三块速率至少为。下侧由定理 3(b) 得出。

对于每个 $r > 0$ 的小连通分量。 不妨重新标记，使真实排列均为恒等排列。在一个包含 m 条真实行和 m 条候选行的连通分量中，令 W_{ij} 表示两行共有的视图数，$k = fm$ 表示任一侧的三度顶点数，h 表示原始简单图中的短环标记数。两侧的度数计数相同，因为两侧的度数总和相等。标记的指定方式是先插入锚边，然后按固定顺序插入其余边；若一条长度至多为 $\Gamma - 1$ 的路径已连接某条边的两个端点，则将该边标记。固定一个偶数 $\Gamma \ge 4$，并令 $\theta = \log m/\log n$。支撑集的代价至多为 $(1-\theta)m\log n + O(m)$，行划分和标记位置的代价为 $O(m)$，而 $m+k$ 个非锚端点各自的代价至多为 $\log m$；若位于标记处，则代价为 $O_\Gamma(1)$。因此，对于固定的 (m, k, h)，对数计数至多为

$$m\log n(1 + \theta f - \theta h/m) + O_\Gamma(m). \tag{39}$$

未标记图保持连通：递归地用其更早的路径替换每条被移除的边，因此 $h \le k + 1$。

令 $M = W/3$、$S = MM^{\top}$、$A = I - S$ 和 $\delta = (1+2r)/(1-r)$。由于 M 对三个受限置换矩阵取均值，$0 \preceq A \preceq I$。将横向方差缩放为 1 后，真实协方差和虚假协方差分别为 $I + (\delta-1)UU^{\top}$ 和 $I + (\delta-1)VV^{\top}$，其中 U, V 的列是归一化的行和，而 $U^{\top}V = M$。在余弦为 σ 的主角处，两个协方差的算术均值在相应平面上的行列式为 $\delta + (\delta-1)^2(1-\sigma^2)/4$。因此，写作 $\xi = (\delta-1)^2/(4\delta)$ 时，单坐标 Bhattacharyya 指数为 $B_G = \frac{1}{2}\log\det(I + \xi A)$。

我们将使用关于二阶谱矩的两个界。直接展开可得

$$\frac{\operatorname{tr} A}{m} = \frac{4+2f}{9}, \qquad \frac{\operatorname{tr}(A - A^2)}{m} \geq \max\left\{ \frac{4}{27}(1 - h/m), \frac{10 - 6f}{81} \right\} \quad (m \geq 4). \tag{40}$$

事实上，$9S_{ii}$ 根据行度为五或三。在 $81\operatorname{tr}(S - S^2)$ 中，对角项贡献 $(20 - 2f)m$，而同列的非对角项则减去 $(8 - 2f)m$。其余项减去无向四环上 W_e 乘积之和的四倍。每个这样的环都包含一个标记，且每条边承载的环权重至多为三。若一条边的权重为二，或其权重为一且有一个二度端点，则该二度端点会迫使环的另一侧路径确定下来，而权重为四的环会填满一个大小为二的连通分量。对于连接两个三度顶点的边，这个和等于两侧其他邻点之间 2×2 矩阵的权重。权重为四会填满这四个邻点，并形成一个大小为三的连通分量。这就得到了第一个二阶矩界。对于另一个界，二度行中的非对角元素 $9S_{ij}$ 之和为四，且没有元素等于四，因为这同样会隔离出一个大小为二的连通分量。它们的平方和至多为十。在三度行中，它们之和为六，且各自至多为三。其中至多有一个为三。若有两个元素为三，就会迫使另外两行都支撑在相同的三列上。这三行会使这些列饱和，并形成一个孤立的连通分量。平方和至多为十四。将这些界分别从 $9(9S_{ii}) - (9S_{ii})^2$ 中减去，得到十和四。

令 $\Lambda = \log(1+\xi)$ 以及 $g = \Lambda + e^{-\Lambda} - 1$。对于 $0 \leq z \leq 1$，$\log(1+\xi z) \geq z\Lambda + z(1-z)g$，可通过对其中的差 ξ 求导并从零开始积分得到。由（40），

$$\frac{B_G}{m} \geq \frac{2+f}{9}\Lambda + \frac{2}{27}(1 - h/m)g.$$

量 $F = 2\Lambda/9 + 4g/81 - R_*$ 和 $H = R_*/2 - 2g/27$ 均严格为正。为说明这一点，令 $y = r/(1-r)$ 以及 $U_1 = 1+y$、$U_2 = 1+2y$、$U_3 = 1+3y$、$D_0' = 2+3y$。则 $R_* = \frac{1}{4}\log(U_1U_2/U_3)$，且 $\Lambda = \log((D_0')^2/(4U_3))$。对 y 求导可得

$$\begin{aligned} 4(D_0')^3 U_1U_2U_3F' &= 2y^3(4 + 12y + 7y^2), \\ 8(D_0')^3 U_1U_2U_3H' &= 2y(16 + 96y + 192y^2 + 144y^3 + 33y^4). \end{aligned}$$

两个函数在 $y = 0$ 处均为零。在 $\theta = 1/2$ 处，相对于 $R_*(1 + f/2 - h/(2m))$ 的裕度至少为

$$(1 + f/2)F + \left[\frac{2}{27} - (1 + f/2)\frac{4}{81}\right] g + (h/m)H > 0.$$

在 $\theta=0$ 处，命题 12(a) 给出 $B_G \ge mR_*$。进行仿射插值得到与 (39) 对 $0 \le \theta \le 1/2$ 的比较论证。对于 $m=2$，3，使用命题 12(a) 以及计数 $O(n^m)$。由于在所有情况下均有 $B_G \ge mR_*$，$d/\log n$ 中的固定裕量可以吸收 $O_\Gamma(m)$，并使对 m、k、h 的求和消失。这排除了所有具有 $m \le \sqrt{n}$ 的连通分量。

大分量 $0 < r \le 9/20$。令 x 表示三单例最优倾斜中的非对角相关性，v_3 表示其边缘方差。用 ρ 表示一个单例与另外两个视图的标准化和之间的相关性。它们的块内对比相互独立，积分后可消去。将三单例高斯倾斜的平稳性方程沿和方向与对比方向对角化，可得

$$r=\frac{3x}{2-x+2x^2},\qquad v_3=\frac{(1-x)(2+x)}{2-x+2x^2},\qquad \rho=r\sqrt{\frac{2}{1+r}}.$$

在双块倾斜中，边缘方差为 $w=\sqrt{1-\rho^2}$，相关性为 $\overline{e}=\rho/(1+w)$，这可通过求倾斜后的二乘二精度矩阵 $\frac12 P^{-1}+\frac{1}{2w}I$ 的逆得到。双块率 $R_2=\log[(1+w)/(2w)]$ 至少为 $R_*=-\frac12\log w$，这是由算术—几何平均不等式得出的。三单例率也满足同样的不等式 $R_3 \ge R_*$，这可由命题 12(c) 得出。由展示的公式，

$$\beta^2:=(w/v_3)^2=1-\frac{6x^2(1-x)}{(2+x)(2+2x+2x^2)}.$$

映射 $x \mapsto r$ 对 $0<x<1$ 单调递增，且 $r(2/7)=21/46$。因此 $x \le 2/7$、$9/10 \le \beta \le 1$、$w \ge 4/5$ 和 $3x/2 \le r \le 8x/5$。

对候选项的两侧应用子集检验 (24)，并使用正相关三单例倾斜。根据 (25)，高斯积分之前的速率为 $(m-k)R_2+2kR_3$，对数检验开销至多为 $(1-\theta+\eta)k\log n+O(m)$。我们如下界定该积分。使用与方差在 $[w,v_3]$ 中的高斯密度平方根成比例的传入消息。在内在标准化坐标中，它们加上精度矩阵 $\Delta_i=(v/\nu_i-1)/2$，其中 $v=w$ 或 v_3 是内在方差，而 ν_i 是消息方差。对除输出端口外的所有端口积分后，将其标准化边缘方差记为 t。其输出方差为 $v/(2/t-1)$。在三度顶点处，另外两个 Δ_i 在 $[0,\Delta_0]$ 中，其中 $\Delta_0=(1/\beta-1)/2$。当两个输入精度矩阵都等于 Δ_0 时，最小的 t 出现，并

$$1 \ge t \ge 1-\frac{2x^2\Delta_0}{1+(1+x)\Delta_0} \ge \frac{1}{1+\Delta_0}.$$

最后一个比较论证由 $1-2x^2+(1+x-2x^2)\Delta_0 \ge 0$ 得出。在二度顶点处，输入精度增量 Δ 属于 $[(\beta-1)/2,0]$，且 $1 \le t=1-e^2\Delta/(1+\Delta) \le 1/(1+\Delta)$。两次更新都保持在盒内。由于内在相关性的最大特征值小于二，且新增的标准化精度大于 $-I/2$，因此完整精度矩阵为正。布劳威尔不动点定理保证存在一个不动点。

仅添加第三个三度精度增量 Δ 后，另外两个端口之间的相关性为 $(x-x^2\tau)/(1-x^2\tau)$，其中 $\tau=\Delta/(1+\Delta)$，且位于 $[0,x]$ 内。额外的

对这些端口添加非负对角精度，只会通过二乘二求逆降低它们的相关性。在二度情况下，同样的求逆将相关性限制为

$$q=\frac{e}{1-(1-e^{2})(1-\beta)/2}<\frac{4x}{3},\qquad e\leq\frac{8\sqrt{2}}{9}x.$$

三度绝对转向矩阵的范数至多为 $2x$。由于内在相关特征值小于二，高斯环积分中的归一化满足 $Z_{\mathrm{B}}\leq 1$根据附录E.5中的熵论证。

在包含三度顶点的连通分量中，压缩每条极大二度路径。该路径上有偶数个二度顶点，由双视图边进行配对。所得核心是一个每侧含有 k 个顶点的二分三次多重图。带有内部配对的压缩边的传输系数至多为q^{2}；直接边的传输系数为一，并且平行边中至多有一条直接边。从环行列式中消去有向内部路径状态，会使局部转向乘以这些传输系数：沿路径方向，消去块是单位三角矩阵。压缩算子的范数至多为 $2x$：其因子包括转向块、边路由置换，以及由不超过一的传输系数组成的对角矩阵。三重平行边类会使两个端点都达到饱和，因此要求 $k=1$。在每个平行边类中标记除一条边以外的所有边；若其中存在直接边，则保留该直接边。按顺序插入剩余的简单核心，标记短环闭合边，并将其保留在前缀中以便计数。随后插入平行边标记，其配对对象已经存在。记 h_{w}， h_{s} 为两种标记的数量，并记 $h=h_{w}+h_{s}$。移除标记不会破坏连通性，并使围长大于 Γ，因此 $h\leq k+1$。

为清晰起见，压缩核心编码使用 m 辅助右侧标签来计算右侧行类型的选择数。配对二度双视图端口至多有（$m-k$）!种选择。将所得配对放入 $3k$ 条定向核心路径中，代价至多再增加（$m-k$）!，以及一个弱合成计数$\exp(O(m)$）。将核心路径发送到右侧端口时，每条未标记边的代价至多为$\log(3k)$，每个标记的代价为 $O_{\Gamma}(1)$。对于辅助标签除以 $m!$ ，因为它们可以通过 $m!$ 种方式分配给每个锚定候选项。计入支撑集和行划分后，得到$\log\ N\ \leq\ m\log\ n+(k-h)\ \log\ m+O_{\Gamma}(m)$。

先删除弱平行边标记。当 $k>3$每条弱标记边至多有一个平行伙伴时，固定方向下，长度为二的返回路径至多耗费 $q^{2}x^{2}$。在长度$j\geq 4$ 中，前 $j-2$ 步的每种选择至多有两个后继，倒数第二步返回起点的总传输系数至多为 $1+q^{2}$，最后一步的弱传输系数至多为q^{2}。接着删除简单核心中的标记。从给定的一条边出发，至多有三条四步返回路径：若有四条，就会用尽一个 $K_{3,\,3}$ 连通分量的所有度。从一条有向边出发的三步路径至多有八条。对于第一步的任一种选择，这些路径的终端有向边互不相同；而两种选择中共同出现的终端状态至多有两个。为证明最后一点，第一步到达的两个邻点必须共享同一个下一顶点，而该顶点除连接这两个邻点外，只剩一条边；这样的公共顶点至多有两个。因此，三步行范数与列范数的平方至多为

$12x^{6}$。由柯西–施瓦茨不等式和核心范数界，长度为 $j\ge 6$ 的返回路径权重至多为 $12x^{6}(2x)^{j-6}$。

对于任一类型，删除操作产生的代价是删除前预解式中两个方向状态的对数行列式的一半。其行列式至多等于对角元素的乘积，而每个对角元都等于 1 加上回返和。对偶数回返长度求和，并利用 $\log(1+t)\le t$，我们得到

$$C_w \le q^2\left[x^2+(1+q^2)\frac{4x^4}{1-4x^2}\right],$$
$$C_s \le 3x^4+\frac{12x^6}{1-4x^2}=\frac{3x^4}{1-4x^2}.$$

剩余的高围长积分的对数至多为 $m\varepsilon_\Gamma$，其中 $\varepsilon_\Gamma\to 0$ 由范数上界给出。简单边系数 C_s/x^2 至多为 $4/11$，弱边系数至多为 $(16/9)(4/49)[1+(1+64/441)(16/33)]$。二者均小于 $45/116$，而 $R_*\ge \rho^2/4\ge (45/58)x^2$。因此，每个标记的代价小于 $R_*/2$。

对于 $m\ge \bar{\sqrt{n}}$，令 $d=\lceil \kappa \log\, n\rceil$，其中固定的 $\kappa>1/R_*$ 位于假定的裕度内；如有必要，舍弃多余的坐标。计数、检验和速率给出的对数概率至多为

$$(m+k-h\theta+\eta k)\log n - d[(m+k-h/2)R_*-m\varepsilon_\Gamma]+O_\Gamma(m).$$

这里是 $m+k-h/2\ge m-1$。对于 $k\le 3$，使用未标记计数：有界核心范数给出一个对数积分 $O_r(1)$；如果 $k=0$二度分量是单个环，则同样具有有界的对数积分。一个较小的固定 η 后接一个较大的固定 Γ，可使指数至多为 $-cm\log\, n$，其中某个 $c>0$对所有充分大的 n均一致。对各计数类别求和，并计入真实表的检验失败概率，即可得到可行似然规则的成功概率。在均匀先验下，完整数据的最大似然至少具有这一成功概率。

$9/20\le r\le 7/8$的大分量。再次使用 $d=\lceil \kappa\log\, n\rceil$，其中 $\kappa>1/R_{*}$。对于一个大小为 $m\ge \sqrt{n}$的给定真实支撑集，令 T_0和 $T_\perp$ 分别表示纵向行子空间与横向行子空间中、在按 $1/\sqrt{1-r}$缩放后沿其 d个坐标的平方和。设 $X=T_0/(md\delta)$、 $Y=T_\perp/(2md)$、 $\psi(t)=(t-1-\log t)/2$，以及 $\Psi=\psi(X)+2\psi(Y)$。在真实分布下， X和 Y是分别除以其自由度 md和 $2md$的独立卡方变量。对于$U\sim \chi^2_p/p$，对上尾和下尾应用Chernoff界可得$\mathbb{P}(p\psi(U)>a)\le 2e^{-a}$。因此， $md\Psi$ 中的每一项在随机序意义下均被一个log 2加上一个独立单位指数变量所支配，并且 $\mathbb{P}(\Psi>z)\le 4(1+mdz)e^{-mdz}$。令 $z=(1-\theta+\eta)/\kappa$。对支撑集以及 $m\ge\sqrt{n}$ 取并集可知，所有此类支撑集以趋近于一的概率通过 $\Psi\le z$。

令 $B=-\min(9r/10,\ 33/50)$、 $C=\min(13r/50,\ 4/25)$、 $b=fB$ 以及 $c=fC$。将真实密度乘以 $\exp[-(bT_0/\delta+cT_\perp)/2]$ 后，其归一化因子为 $[(1+b)(1+c)^2]^{-md/2}$，并将两个方差倾斜为 $v_\parallel=\delta/(1+b)$ 和 $v_\perp=1/(1+c)$。由于似然根据纵向平方和对候选项进行排序，只有当某个候选项的平方和 T_F 满足 $T_F\ge T_0$时，它才会优于真实分配。令 $t=v_\parallel/v_\perp$ 和 $\xi_t=(t-1)^2/(4t)$。在倾斜分布下

乘以 $\exp[s(T_F - T_0)/2]$，其中 $s = (v_\parallel - v_\perp)/(2v_\parallel v_\perp) > 0$。在余弦为 σ的主角平面中，所得的归一化精度行列式等于

$$1 + [s(v_\parallel - v_\perp) - s^2 v_\parallel v_\perp](1 - \sigma^2) = 1 + \xi_t(1 - \sigma^2).$$

横向主子式至少为 $1 - sv_\perp > 0$，因此高斯积分是有限的。两个指数因子之积的负对数矩除以 md为

$$\tfrac{1}{2}[\log(1 + b) + 2\log(1 + c)] + \tfrac{1}{2m}\log\det(I + \xi_t A).$$

在 $T_F \geq T_0$ 上，该乘积至少为$\exp[-md(bX + 2cY)/2]$。由于对 $a' < 1$有 $\psi(u) - a'u/2 \geq \tfrac{1}{2}\log(1 - a')$，故在 $\Psi \leq z$上有$(bX + 2cY)/2 \leq fz - f[\log(1 - B) + 2\log(1 - C)]/2$。因此，与支撑检验相交的候选概率至多为$\exp[-md(G_G - fz)]$,，其中

$$\begin{aligned} G_G &= \frac{\log\det(I + \xi_t A)}{2m} + \Phi, \\ \Phi &= \tfrac{1}{2}\big\{\log(1 + b) + 2\log(1 + c) + f[\log(1 - B) + 2\log(1 - C)]\big\}. \end{aligned}$$

记 $\ell = (4 + 2f)/9$, $u = h/m$, $u_0 = (1 + 3f)/6$, $a = 1 - 4/(27\ell)$以及 $a_u = a + [4/(27\ell)]\min(u, u_0)$。由(40)，$\operatorname{tr} A^2/\operatorname{tr} A \leq a_u$。函数$\log(1 + \xi_t v)/v = \int_0^{\xi_t}(1 + \omega v)^{-1} d\omega$ 递减且为凸函数。对 A 的特征值应用带权重 $\lambda_i/(m\ell)$的詹森不等式，得到

$$\frac{1}{2m}\log\det(I + \xi_t A) \geq \frac{\ell}{2a_u}\log(1 + \xi_t a_u).$$

在 $\min(u, u_0)$ 处的切线下界 $= 0$ 的斜率为 $-D_t$，其中

$$D_t = \frac{2}{27a^2}\left[\log(1 + \xi_t a) - \frac{\xi_t a}{1 + \xi_t a}\right].$$

因此，加入 $R_*\theta u$，其中 $\theta \geq 1/2$ 和 $u \leq f + 1/m$，可得

$$G_G + R_*\theta u \;\geq\; \frac{\ell}{2a}\log(1 + \xi_t a) + \Phi - \min(f, u_0)(D_t - R_*/2)_+ - O_r(1/m). \tag{41}$$

计算机辅助不等式。 脚本 `scripts/certify_e1_正_mid.py` 以及表格`e1_正_mid/rows.csv` 证明了（41）中显式表达式在 $O_r(1/m)$ 项之前，在 $(1+f)R_*$上均匀地超过 $[9/20, 7/8] \times [0, 1]$。该脚本在断点 $8/13, 11/15$ 中以及 r 中 $1/3$ 内的 f处进行划分，然后用闭有理矩形覆盖每个子矩形。它通过区间代入或结合区间导数界的中值定理，检查正部的两个可能分支。该表格记录了每个子域中差值的向外舍入下界以及矩形数量。

将经检验调整的概率与（39）结合，可得每行的对数概率除以 $\log\, n$至多为 $1+f-\theta u-\kappa G_G+f\eta+o(1)$。由于 $1+f-\theta u\ge 1-1/m$，有限的正维度裕度使得当 η足够小时，可得到一个均匀的负界。对 m,k,h 求和趋于零。支撑检验失败的概率也为 $o(1)$，而固定坐标估计器通过最大似然的最优性将结果转移到完整数据上。

高相关性 $7/8\le r<1$。首先，使用双视图最大似然将每个非锚定视图拟合到锚定视图。设 $v=\sqrt{1-r^2}$, $a_0=(1-v)/(1+v)$、$I_{\mathrm{pair}}=-\log\, v$ 和$Q_0=\log[(1+v)/(2v)]$。长度为$\ell\ge 2$的双视图错误循环的 Bhattacharyya 指数为

$$E_\ell=\ell Q_0+\log(1-a_0^{\ell}),\qquad E_2=I_{\mathrm{pair}}.$$

事实上，双块倾斜给出一个具有 $r/(1+v)$恒定转向的 2ℓ 个顶点环。提出 $e^{-\ell Q_0}$ 后，局部归一化为 1，因为传入消息是内在平方根边缘密度。环的两种定向分别对环矩阵的行列式贡献 $1-[r/(1+v)]^{2\ell}=1-a_0^{\ell}$，从而得到该公式。

最大化器中的错误循环必须分别超过或持平于真实得分，否则就可以将其还原。长度为 ℓ 的此类循环最多有 n^{ℓ}/ℓ 个。因此，对于$d\ge\kappa\log\, n$，如果 $Q_0>R_*$，则错误的期望数量至多为

$$\sum_{\ell\ge 2} n^{\ell-\kappa E_\ell}\le\frac{n^{2-\kappa I_{\mathrm{pair}}}}{1-n^{1-\kappa Q_0}},$$

利用 $\log(1-a_0^{\ell})$ 的单调性。为检验速率，令 $j=\sqrt{1+2r/(1+r)}$，于是 $I_*=-\log(jv)$。不等式 $Q_0>R_*$ 等价于 $4v/(1+v)^2<j$。对于 $7/8\le r\le 9/10$，利用 $v(7/8)<97/200$，左侧小于 $22/25$；而利用其在 $9/10$ 处的值，j 大于 $22/25$。对于 $r\ge 9/10$，$v(9/10)<11/25$，且 $4(11/25)/(1+11/25)^2<\sqrt{3/2}\le j$。此外，$I_{\mathrm{pair}}>3I_*/4$ 等价于 $v<j^3$，这由 $\sqrt{15/8}<(\sqrt{3/2})^3$推出。

选择 $\max(0,2-\kappa I_{\mathrm{pair}})<\theta_0<1/2$。由马尔可夫不等式，以趋于 1 的概率，每个真实置换都位于相应双视图拟合结果的汉明距离 n^{θ_0} 以内。在这两个汉明球的乘积分布上最大化联合似然，并向上取整。对于所有充分大的 n，乘积汉明球中的每个候选项与真实分配至多在 $4\lceil n^{\theta_0}\rceil<\sqrt{n}$ 个真实行上不同。小分量论证排除了任何得分差非负的此类错误候选项。因此，约束规则能够成功，贝叶斯最优完全似然规则也同样如此。

E.7. 相关性为 $-1/4$的四个视图

定理 13（ii）。设 $K=4$ 所有非对角相关系数均等于 $-1/4$，并设$\rho=R_{1|2|3|4}$。则 $L=\rho$，且在固定裕量下，$d/\log\, n>1/\rho$可通过最大似然实现精确恢复，而 $d/\log\, n<1/\rho$则使任何估计器的成功概率趋于零。此处 $1/\rho\approx 9.8512$，交换条件与共享信息条件合起来仅要求 $d/\log\, n>8.511$ 达到所述精度。

局部高斯因子。 将每个视图除以 $\sqrt{5/4}$。给定一个块大小为b_i 且包含 q 个块的划分，保留每个块的归一化和。所有块内正交对比均相互独立，服从标准正态分布，且独立于块和向量。约化后的协方差矩阵和精度矩阵为

$$P_b = I - uu^\top/5, \qquad P_b^{-1} = I + uu^\top, \qquad u_i = \sqrt{b_i}, \qquad \det P_b = 1/5.$$

每条边两端都采用相同的坐标变换。独立的对比密度在倾斜图积分中的积分为 1，将这些对比密度与其原始边缘分布结合后，仍能保持参考优化器。事实上，以这种方式构造且满足倾斜边缘一致性的高斯参考分布，根据命题12(b)，是最优的。对于 $\beta = 1/q$， $\alpha = 1 - \beta$ 以及和方向上的参考方差 ν_i，这种一致性与高斯积分给出

$$Q_b^{-1} = \alpha P_b^{-1} + \beta\,\mathrm{diag}(1/\nu_i), \qquad (Q_b)_{ii} = \nu_i.$$

$$R_\pi = \frac{\log\det Q_b^{-1} + \alpha\log(1/5) + \beta\sum_i \log\nu_i}{2\alpha}. \tag{42}$$

共有四种形状，记为 1111、112、13、22。对于形状为 b,$D_\pi = \frac{1}{2}$ 的划分，$[5\prod_i (1 - b_i/5)]$, 由边缘行列式与联合行列式之比给出。两种切割给出 $I_*/2 = \frac{1}{4}\log(8/5)$, 对四种形状上的 $D_\pi/(q-1)$ 取最小值，得到 $J = \frac{1}{6}\log(256/125) > I_*/2$。因此，较早的两个条件要求取 $2/I_*$ 和 $1/J$ 的最大值，等于 $4/\log(8/5)$。下面的证书验证 $\rho < I_*/2$ 和 $R_{112} > \rho$，因此稀有子总体逆命题证明了下侧。

对于上侧，我们应用附录 E.9 中的大分量判据（46），其中$\theta_0 = 29/50$。该理论中顶点处的因子为 $F = q_b/\prod_i \sqrt{\phi_{\nu_i}}$，其中 q_b是协方差为 Q_b 的密度，而 ϕ_ν 是方差为 ν的中心化标量密度。其精度矩阵为 $\mathrm{diag}(a_i) + \alpha uu^\top$，其中 $a_i = \alpha + (\beta - 1/2)/\nu_i$。在本段中，高斯消息的精度以密度单位计量（是附录 E.9中所用平方根密度约定的一半）。舒尔补给出了朝向端口的更新 i

$$t_i' = a_i + \frac{\alpha b_i}{1 + \alpha\sum_{j\neq i} b_j/(a_j + t_j)}. \tag{43}$$

按块大小索引的消息区间为 $[0.662, 0.791]$、$[0.949, 1.119]$ 和$[1.548, 1.582]$。证书检查 $a_i > 0$ 并验证单调更新 (43) 将这些区间映射到自身。第一个不等式还表明，F 的内在标准化协方差小于 $2I$，这是通过将 F 的精度矩阵写为 $Q_b^{-1} - \frac{1}{2}\,\mathrm{diag}(1/\nu_i)$ 得出的。加入所有传入消息后，通过舒尔补移除除 i、j 之外的所有端口。保留的双端口精度的对角部分为 $\mathrm{diag}(a_i + t_i, a_j + t_j)$ 加上 $\lambda u_{ij}u_{ij}^\top$，其中$\lambda = \alpha/(1 + \alpha\sum_{k\notin\{i,j\}} b_k/(a_k + t_k))$。因此，其相关系数平方为

$$\frac{\lambda^2 b_i b_j}{(a_i + t_i + \lambda b_i)(a_j + t_j + \lambda b_j)}.$$

增大 λ 或减小保留的对角元素，都会使该表达式有一个上界，因此证书只使用盒端点。

小分量端点。 对于大小为 m 的连通分量，真实行空间与候选行空间的重叠度为 $M = W/4$，其中 W_{ij} 表示共享视图的数量。其主角度行列式给出 $B_G = \frac{1}{2}$对数行列式$[I + \frac{4}{5}(I - MM^{\top})]$，这与附录 E.6 中的二维计算相同，其中 $\delta = 1/5$。由对数函数的凹性可知，该值至少为

$$\sum_{v\,\mathrm{true}} \left[1 - \sum_{i \in \pi_v} b_i^2/16 \right] \frac{\log(9/5)}{2}.$$

该证书验证了每个求和项都大于 $[1 + \theta_0(q_v - 2)]\rho$。即使没有标记边带来的收益，这也给出了判据的小分量端点。

计算机辅助不等式。脚本 scripts/check_e2_four_view.py 和表格 e2_four_view/rows.csv 验证了四组比较论证。倾斜边缘更新 $\nu \mapsto \mathrm{diag}[(\alpha P_b^{-1} + \beta\,\mathrm{diag}(1/\nu_i))^{-1}]$ 通过单调性将指定的有理方差区间映射到自身。由连续性可知存在不动点。式（42）给出的速率界证实了 ρ 的绑定，并将 $1/\rho$ 限定在 (9.8511971,9.8512024) 内，端点向外取整。上述消息区间和平方相关系数上界给出的绝对转向行和至多为 0.711。将这些上界代入两步端点公式（45），可验证每种形状

$$\frac{1}{2}\sum_{B \in \pi} (C^1(\pi, B) - \theta_0 \rho)_+ \le \frac{q-1}{2}(R_\pi - \rho).$$

最后，刚才所述的小分量裕度不等式对四种形状均成立。对数界使用 2 arctanh t，$t = (z-1)/(z+1)$ 的 48 项级数，绝对误差至多为 $2|t|^{97}/[97(1-t^2)]$。比较论证在显示小数之前使用有理端点进行。由此，该判据证明了上侧。

E.8. 三视图高斯链

定理 13(iii). 对于正定三视图高斯链，重新标记，使得 $r_{13} = r_{12} r_{23}$。则 $L = I_*/2 = \min(I_{12}, I_{23})/2$，且双视图割阈值给出具有固定裕度的精确恢复。

对于高斯向量，相关性恒等式表明，在给定视图二的条件下，视图一与视图三条件独立。因此，其密度为 $p_{12}p_{23}/p_2$。由于均匀相对置换模型在更换锚点时保持不变，我们可以将视图二作为锚点。给定对齐方案的对数似然等于边 12 和边 23 各自的对数似然之和，再加上一个对所有排列都相同的常数。因此，分别最大化每条边的似然即可得到联合最大值；相对于原锚点的相对置换则通过复合得到。根据链式法则，隔离视图一的末端割的信息量为 I_{12}，另一末端割的信息量为 I_{23}，而

条件独立。中间割的信息量至少 I（X_2； X_1， X_3）不小于两者。因此， $I_* = \min(I_{12}, I_{23})$。按需翻转任一末端视图的符号，使两条路径相关系数均为非负。它们的乘积就是末端到末端的相关系数，因此此时三者均非负。这些可逆变换不会改变速率，而命题12（c）给出 $R_{1|2|3} \geq I_*/2$。

当两条路径相关系数均非零时，已有的双视图循环界表明，只要 $dI_{ij} > (2+\epsilon)$ log n 对某个固定的 $\epsilon > 0$成立，则每条边上任意错误最大值出现的概率都趋于零 ℓ 。具体而言，长度为 $\ell I_{ij}/2$ 的循环的 Bhattacharyya 指数，依据命题12(a) 中的循环信息范数界，至少为 n^{ℓ}/ℓ，而循环数量至多为 $1/n$。每个错误最大值都包含一个得分差非负的错误循环。在两条边上对这些界求和即可证明联合成功。定理3(b) 给出了下侧的失败结果。若任一路径相关系数为零，则一个终端视图与另外两个视图独立，因此其相对置换保持均匀，精确成功概率至多为 $L = 0$，与 n 一致 $\epsilon > 0$!

E.9. 有限维充分判据

固定 K 以及一个正定的高斯协方差 R ，满足 $L > 0$，其中 L 如猜想11中所述。以下关于大小至多为 K 的分块矩阵的局部条件，在 $d/\log n > 1/L$时蕴含成功概率趋于一。结合逆命题，这给出了此类 R的猜想阈值。只要 $I_* > 0$，该判据在恒等映射的邻域内成立，这就是定理13(iv)；对于两个负等相关示例也成立：一个是 $K = 3$ ，满足 $r = -1/3$ ；另一个是 $K = 4$ ，满足 $r = -1/4$。但它沿着近奇异负等相关族失效。

局部因子和消息区间。 对于每个将 $\mathcal{P}$ 划分为 $q \geq 2$ 个块的划分，使用两组参数: $\alpha^0 = 1/2$ 和 $\alpha^1 = (q-1)/q$。在任一组参数中记作

$$V = \prod_{B \in \mathcal{P}} v_B, \qquad Z = \int p^{\alpha} V^{1-\alpha}, \qquad Q = Z^{-1} p^{\alpha} V^{1-\alpha}.$$

对于集合 1，取命题 12（b）中的最优乘积高斯参考分布。对于集合 0，只需在中心化分块乘积高斯分布中最大化亲和度即可。该类中的最大值在正定内部取得。事实上，如果 V_0 是分块对角参考协方差，则亲和度为

$$\frac{(detR)^{(1-\alpha)/2}(detV_0)^{\alpha/2}}{\det((1-\alpha)R + \alpha V_0)^{1/2}}.$$

用将 R 替换为 $\lambda_{\min}(R)I$ 后的行列式给分母设定上界，当任一参考特征值趋于零或无穷大时，亲和度趋于零。其他特征值因子均有有限上界。对分块精度矩阵求高斯归一化的导数，在最大值处有 $M_B = S_{BB}$，其中 S 是 Q 的协方差，而 M_B 是 v_B 的协方差。相应的速率为

$$T^0_{\mathcal{P}} = -2\log Z^0, \qquad T^1_{\mathcal{P}} = -q\log Z^1 = (q-1)R_{\mathcal{P}}.$$

对于两区块划分，两种速率相同，且至少为 $I_*/2$ 根据附录E中的两区块比较。

每个可能的块 B 在两个端点都使用同一个坐标系，例如对白化后的 R_{BB}使用坐标。在任一参数集中，局部因子为 $F = Q/\prod_B \sqrt{v_B}$。沿着块-B 边的消息与精度矩阵为 W_B 的高斯密度平方根成正比。选取一个共同的洛文纳区间 $[W_B^-, W_B^+]$，使其包含每个具有该块的划分对应的内在精度矩阵 M_B^{-1}。定义 $\Delta_D = (W_D - M_D^{-1})/2$，并令 H 为包含所有传入消息 $H^{-1} = S^{-1} + \operatorname{diag}_D \Delta_D$时的协方差。令$H_{\setminus B}$ 表示将 Δ_B 设为零时的协方差。对每组参数进行以下检查：

1. H^{-1} 以及每个 $H_{\setminus B}^{-1}$ 在下端点处均为正定。

2. 对于输出块 B，在其余下端点和上端点处的两个 $2[(H_{\setminus B})_{BB}]^{-1} - M_B^{-1}$ 值均属于 $[W_B^-, W_B^+]$。

3. 对 S 的每个边缘分布用其 M_B 白化后得到的内在分块标准化协方差，不超过 $2I$。

检查 2 中的表达式，是通过对局部因子与其传入消息的乘积中所有其他块进行积分得到的。根据求逆的反序性质，该表达式关于每个输入精度矩阵单调。因此，每个有限候选图都存在一个从紧凸盒映射到自身的连续消息更新，而布劳威尔不动点定理保证了不动点的存在。将 F 与顶点 v 处的传入消息相乘并归一化为密度 b_v，并将边 e 上两个相反方向的消息之积归一化为 b_e。更新方程使 b_v 的相应边缘分布相等于 b_e。

对于划分中的任意 $B \neq D$，令 $b(\mathcal{P}; B,\ D) = b(\mathcal{P}; D,\ B)$ 为协方差 H 经其块边缘分布白化后，BD 块的算子范数的一个上界。可通过要求

$$\begin{pmatrix} b(\mathcal{P};B,D)H_{BB} & H_{BD} \\ H_{DB} & b(\mathcal{P};B,D)H_{DD} \end{pmatrix} \succeq 0$$

在整个消息盒中成立。如果 $\ell I \preceq H^{-1} \preceq u I$，则一个实用的较粗上界为

$$b(\mathcal{P}; B, D) = u \| \operatorname{offblock}(S^{-1}) \| / \ell^2 \tag{44}$$

当右侧被用作所选上限时。为说明这一点，从 H 中减去 H^{-1} 的分块对角部分的逆矩阵。所得差是 $-H$、非块对角精度矩阵和该分块对角部分的逆矩阵的乘积。其范数至多为 $\ell^{-2}\|\operatorname{offblock}(S^{-1})\|$。左右边缘白化因子使该上界至多增加 u。

对于集合 $j = 0,\ 1$，根据其帽集计算以下量：

$$a = \max_{\mathcal{P}, B} \sum_{D \neq B} b(\mathcal{P}; B, D), \qquad W(B) = \max_{\mathcal{P} \ni B} \sum_{D \neq B} b(\mathcal{P}; B, D)^2,$$

$$D(\mathcal{P}, B) = \sum_{D \neq B} b(\mathcal{P}; B, D)^2 W(D), \qquad C^j(\mathcal{P}, B) = s_B \log\left(1 + \frac{D(\mathcal{P}, B)}{1 - a^2}\right). \tag{45}$$

此处， s_B 表示边上携带的维数。如果一个块划分为一个相关方向和独立对比，且两端具有相同的边缘分布，则这些对比的积分为 1，在应用这些检查之前可以将其省略。要求每个所用集合满足 $a<1$，并选择 $\theta_0\in(0,\ 1)$，使得对于所有划分，

$$\begin{aligned} T^{0}_{\mathcal{P}} &\geq L[1+\theta_0(q-2)]+\sum_{B\in\mathcal{P}}(C^{0}(\mathcal{P},B)-\theta_0 L)_{+},\\ T^{1}_{\mathcal{P}} &\geq (q-1)L+\sum_{B\in\mathcal{P}}(C^{1}(\mathcal{P},B)-\theta_0 L)_{+}. \end{aligned} \tag{46}$$

集合 0 及其不等式可替换为下文所述的直接 Bhattacharyya 下界。划分和块的数量有限。协方差检验和上限可通过对定阶矩阵进行有限区间界认证，例如使用单调端点检验、显式上限（44）或区间舒尔补。

高斯环积分。 将真实排列重新标记为恒等置换。在一个连通候选项连通分量中，将每对真实行与错误行共有的视图合并为一条边。这是一个简单二分图，每侧有 m 行，且每个顶点的度介于 2 和 K 之间。边携带其视图块。令 $\mathcal{I}_G$ 为其局部因子的乘积分布的积分，每条边对应一个变量。记 C_v 为 b_v的分块标准化协方差，记 C 为其在各顶点上的直和，W 为每条边两条半边的交换。我们断言

$$\mathcal{I}_G = Z_{\mathrm{B}}\det[I-(C-I)\mathsf{W}]^{-1/2},\qquad Z_{\mathrm{B}}\leq 1. \tag{47}$$

不动点会抵消信念乘积与因子乘积之间的消息： $\prod_v F_v = Z_{\mathrm{B}}\prod_v b_v/\prod_e b_e$。对归一化常数取对数，并在一致信念下取期望，得到

$$\log Z_{\mathrm{B}} = \sum_{v}\left[-D(b_v\|Q_v)+\tfrac{1}{2}\sum_{B\in\mathcal{P}_v}D((b_v)_B\|v_B)\right].$$

事实上，每条传入消息都会出现在一个顶点归一化因子和一个边归一化因子中，并且在两者中的期望相同。为界定所显示的和，令 $A=S^{-1/2}\,\mathrm{Cov}(b_v)S^{-1/2}$。相对于 M_B 归一化后的块协方差为 $P_B^{\top}AP_B$，其中P_B 是等距嵌入 $S^{1/2}E_BM_B^{-1/2}$，而 E_B 将块 B的坐标插入其中。由于 $g(t)=(t-1-\log t)/2$，对该压缩的每个特征向量应用詹森不等式可得 $\mathrm{tr}\ g(P_B^{\top}AP_B)\leq \mathrm{tr}(P_B^{\top}g(A)P_B)$。检查 3 表明 $\sum_B P_BP_B^{\top}\preceq 2I$，因为该和与块标准化后的 S具有相同的谱。因此，分块熵之和至多为联合熵的两倍，并且 $Z_{\mathrm{B}}\leq 1$。

白化公共边缘分布 b_e。记 D_E 为将每条边变量复制到其半边的操作，因此 $D_ED_E^{\top}=I+\mathsf{W}$ 和 $D_E^{\top}D_E=2I$。块范数上界给出 $\|C-I\|\leq a<1$，这使得 $D_E^{\top}C^{-1}D_E-I$ 正定。直接进行高斯积分可得 $\det C\det(D_E^{\top}C^{-1}D_E-I)$ 的逆平方根。行列式

恒等映射 $\det(I+XY)\ =\det(I+YX)$ 以及 $\ D_E D_E^{\top}\ =\ I+\mathsf{W}$ 将此乘积分布认定为 $\det[I-(C-I)\mathsf{W}]$, 从而证明（47）。

标记边的代价。插入锚边后，按固定顺序添加非锚边，并将端点此时已由长度至多为$\Gamma\ -\ 1$的路径连接的边标记，其中 $\Gamma\ \geq\ 4$ 为偶数。记 h 为标记数，$\ U\ =\ |E|\ -\ m$，且 $\theta\ =$对数 $\ m/$对数 $\ n$。给定支撑集和真实行划分后，需要发送 U 个非锚端点。每个端点最多有 m 种选择；若位于标记边上，则有$O_{K.\ \mathrm{r}}(1)$种。计入支撑集、划分和标记位置后，得到

$$\log N(m,U,h)\leq[m+\theta(U-m-h)]\log n+O_{K,\Gamma}(m).\tag{48}$$

删除所有标记后，围长大于 Γ，因为残存短环的最后一条边本应被标记。剩余的图是连通的，因为每个标记边都有一条更早出现的替代路径，而该路径上的更早标记边可以递归地替换。因此$h\leq U-m+1$。

为了界定式（47）中的对数行列式，将每个转向的块范数替换为其标量上界。由此得到的非负有向非回溯矩阵 A_+ 是一个对称的局部转向矩阵与一个置换矩阵的乘积，因此 $\|A_+\|\leq a$。对于任意承载 B的有向边 $u\to v$ ，二步路径会通向彼此不同的终端有向边。由于图是简单图，起始边的终点和终端边的起点唯一确定中间边。因此，二步行范数的平方至多为 $D(\mathcal{P}_v,B)$，列范数的平方至多为 $D(\mathcal{P}_u,B)$。柯西–施瓦茨不等式将长度为 $j\ \geq\ 4$ 的返回量界定为$a^{j-4}\sqrt{D(\mathcal{P}_u,B)D(\mathcal{P}_v,B)}$。返回只可能发生在偶数 $j\geq 4$，且删除边状态后，所有界仍然成立。

将一条闭游走在对数行列式展开中归因于其按照固定删除顺序访问的第一条标记边。将其起点旋转到该边的两个状态之一。块迹的绝对值至多为 s_B 乘以标量转移界的乘积。归因于该边的游走的标量和，恰好等于从当前主子矩阵中移除这两个状态时 $-\frac{1}{2}$对数行列式$(I-A_+)$的减少量。雅可比恒等式将该减少量表示为预解式中相应二状态子矩阵的对数行列式的一半。由于诺伊曼级数逐元素非负，其行列式不超过其对角元素的乘积。任一对角元素均至多为$1+\sqrt{D(\mathcal{P}_u,B)D(\mathcal{P}_v,B)}/(1-a^2)$。于是，~~不等式 $1+\sqrt{xy}\leq\sqrt{(1+x)(1+y)}$~~ 将该归因量界定为 $[C^j(\mathcal{P}_u,B)+C^j(\mathcal{P}_v,B)]/2$。未被归因的游走长度大于 Γ，其迹级数由 $O_K(m)\ \sum_{j>\Gamma}a^j$界定。因此

$$\log\mathcal{I}_G\leq\frac{1}{2}\sum_{e=uv\,\text{marked}}[C^j(\mathcal{P}_u,B_e)+C^j(\mathcal{P}_v,B_e)]+m\varepsilon_\Gamma,\qquad\varepsilon_\Gamma\longrightarrow 0.\tag{49}$$

合并两个大小范围。对于集合 $\mathsf{0}$，顶点处原始密度的平方根为 Z^0F，其负对数常数为 $T_{\mathcal{P}}^0/2$。因此，(46) 和 (49) 给出指数$B_G\geq L[m+\theta_0(U-m-h)]-m\varepsilon_\Gamma$，对应大小端点 θ_0。也可以用原始协方差行列式得到的同一上界来替代集合 $\mathsf{0}$。在大小端点为零时，

命题12(a) 给出 $B_G \geq mI_*/2 \geq mL$。对同一连通分量的两个下界进行插值，得到 $B_G \geq L[m+\theta(U-m-h)]-m\varepsilon_\Gamma$，其中 $0 \leq \theta \leq \theta_0$。将其与 (48) 结合，可在固定维度裕量下排除所有这些分量。具体而言，在裕量内为固定的$d' = \lceil \kappa$个对数 $n \rceil$ 坐标，并预留 $\kappa > 1/L$ 个对数 n坐标，使其速率至少为 mL。取充分大的固定 Γ，即可吸收 $m\varepsilon_\Gamma$ 项，并得到阶为负的指数 m对数 n。

对于其他规模，仅当对每个划分 $\mathcal{P}$ with $q \geq 3$ 以及每个由 s 行组成的子集，这些行和所有坐标上的 $\log(Q^1_{\mathcal{P}}/p)$ 之和不超过$t_s = \log \binom{n}{s} + \eta s \log n$时，才接受一张表。真实表以趋近于 1 的概率通过所有检验：固定检验的指数矩为 1，对这些子集求和后，失败界是 $\sum_{s\geq 1} n^{-\eta s}$ 的有限倍数。一个似然至少与真实似然相等的可行竞争者，其得分差至少有一个连通分量非负。按各自的划分对两侧的行进行分组。几何平均数的真实密度与候选密度在该得分事件上相比不小。对于大小为 s with q 个块的一组，检验允许在每个顶点处将 $(p/Q^1)^{(q-2)/2}$ 插入几何平均数，指数代价为 $(q-2)t_s/2$。所得的顶点被积函数为

$$p^{1/2}(p/Q^1)^{(q-2)/2} = (Z^1)^{q/2}F.$$

在每一侧，$\sum_v (q_v - 2) = U - m$。组开销之和至多为 $(1-\theta+\eta)(U-m)\log n + O_K(m)$：使用 $\log \binom{n}{s} \leq s\log(n/m) + s\log(m/s) + s$ 以及 $s\log(m/s) \leq m/e$。由（48）可知，计数和检验开销因此至多为 $(U-\theta h+\eta(U-m))\log n + O_{K,\Gamma}(m)$。与此同时，$-\log(Z^1)^{q/2} = T^1_{\mathcal{P}}/2$。将（46）中的第二个不等式与（49）结合，可得单坐标指数至少为 $L(U-\theta_0 h) - m\varepsilon_\Gamma$。

对于 $\theta \geq \theta_0$，除了检验松弛量和长环尾部之外，计数在阈值处支付，且 $U-\theta_0 h \geq m-1$。固定维度裕量会吸收这些项，只要固定的 η 足够小且固定的 Γ 足够大。对 m, U, h 求和，可得存在可行且获胜竞争者的概率趋于零。因此，可行的最大似然规则能够成功，而对完整数据应用最大似然时，在均匀先验下的成功概率至少达到该值。

弱相关。 将视图方差标准化为 1，并设置 $\gamma = \|R - I\|_{\mathrm{op}}$。对于每个划分，在对白化后的块进行处理后，将协方差写作 $I + E_{\mathcal{P}}$，其中 $E_{\mathcal{P}}$ 的块对角为零，且范数为 $O_K(\gamma)$。对于任一参数集，

$$\begin{aligned} M_B &= I + O_K(\|E_{\mathcal{P}}\|^2), \\ S &= I + \alpha E_{\mathcal{P}} + O_K(\|E_{\mathcal{P}}\|^2), \\ -\log Z &= \frac{\alpha(1-\alpha)}{4}\|E_{\mathcal{P}}\|_F^2(1 + O_K(\gamma)). \end{aligned}$$

对于前两个公式，块一致性方程为 $S_{BB} - M_B = 0$，其中$S^{-1} = \alpha(I + E_{\mathcal{P}})^{-1} + (1-\alpha)\,\mathrm{diag}(M_B^{-1})$。其在恒等映射处关于块对角未知量的导数为 $-\alpha I$，而其关于块外输入的导数在对角块上为零

块。隐函数定理给出了局部分支。根据上述特征值界，高斯亲和度的超水平集在恒等映射附近是一致紧的。因此， $E_{\mathcal{P}} \to 0$ 的任意最大化元序列都趋于唯一的独立性优化解，并且最终位于该分支上。在亲和度对数公式中，线性的块对角项相互抵消。对块外扰动进行二次对数行列式展开，得到第三个公式；其误差是相对 $O_K(\gamma)$。

使用具有足够大固定 A_K 的公共精度矩阵盒 $[1-A_K\gamma^2, 1+A_K\gamma^2]I$。其完整精度矩阵和部分精度矩阵均为正定。对于内禀输入，输出等于其内在精度矩阵。将输入精度矩阵的其他块扰动 $O_K(\gamma^2)$ 后，输出舒尔补改变了 $O_K(\gamma^4)$，因为该扰动位于两个范数为 $O_K(\gamma)$ 的非对角块之间。这证明了当 γ 足够小时的盒不变性。内在标准化协方差为 $I+O_K(\gamma)$，且（44）给出 $b(\mathcal{P};\, B, D) = O_K(\|E_{\mathcal{P}}\|)$，其中 $a < 1$。因此， $C^j(\mathcal{P}, B) = O_K(\gamma^2)T^j_{\mathcal{P}}$，这是一个即使对于极弱切割仍然有效的相对界。

令 c_* 表示某个切割上相关系数平方的最小总和，令 $c_{\mathcal{P}}$ 表示该划分各块上的总和，其中每个切割对仅计数一次。分离各个块可得 $2c_{\mathcal{P}} \geq qc_*$。根据上述展开式和高斯切割行列式， $T^0_{\mathcal{P}} = (1 + O_K(\gamma))c_{\mathcal{P}}/4$，且 $I_*/2 = (1 + O_K(\gamma))c_*/4$。因此，对于 $q \geq 3$，有 $T^0_{\mathcal{P}} \geq (1 - O_K(\gamma))qL/2$。对于 $q = 2$，其至少为 L，且恰好成立。取 $\theta_0 = 1/4$。根据前面对集合 0 的界以及集合 1 中 L 的定义，（46）中的两个无环预算项都不超过其对应的 $T = T^j_{\mathcal{P}}$。若 $T \leq 2KL$，则对于足够小的 γ，每个 C^j 都低于 $\theta_0 L$。若 $T > 2KL$，则无环项至多为 $KL < T/2$，而 C^j 的总和为 $O_K(\gamma^2)T < T/2$。这证明了该判据对最小割的强度均匀成立。如果 $I_* > 0$，则按要求有 $L > 0$：零稀有率将根据所达到的优化解以及亲和度界的等号情形，使 p 在一个非平凡切割上成为乘积分布。

反例。对于等相关性，写作 $\delta = (1 + (K - 1)r)/(1 - r)$ 和 $\xi = (1 - \delta)^2/(4\delta)$。主角公式给出 $M = W/K$。由凹性可得 B_G 的下界，它对真实行求和，求和项为

$$b_{\mathcal{P}} = \tfrac{1}{2}\log(1 + \xi)\left[1 - \sum_{B \in \mathcal{P}} (|B|/K)^2\right].$$

在小分量端点，只需验证 $b_{\mathcal{P}} \geq L[1 + \theta_0(q - 2)]$ 对每个划分都成立。附录 E.7中的检验给出了这一结果，以及 $K = 4, r = -1/4$。

对于 $K = 3, r = -1/3$，取 $\theta_0 = 1/2$。$b_{\mathcal{P}}$ 的两个可能取值为 $(2/9)\log(25/16)$ 和 $(1/3)\log(25/16)$。在大连通分量的计算中，令 $x = 2 - \sqrt{3}$，令 v_3 和 v_2 分别表示附录 E.5 中的内禀三度和二度边际方差，并令 $e = |\rho|/(1 + v_2)$ 表示内禀二度相关性。物理方差区间 $[v_2, v_3]$ 中的消息遵循与附录 E.6 相同的更新。这里，$b = v_2/v_3 \leq 1$, $H_0 = (1/b - 1)/2$，而三度方差下界为 $1 - 2x^2 H_0/[1 + (1 - x)H_0]$。它至少为 $1/(1 + H_0)$，因为

$1-2x^2+(1-x-2x^2)H_0>0$。二度更新保持不变，因此盒具有不变性。内在最大相关特征值为 $1+x$ 和 $1+e$，均小于二。由于 $z=(1-b)/(1+b)$，由对剩余端口进行条件化得到的三度转向上限为 $(x+x^2z)/(1-x^2z)$，而二度上限为 $e/[1-(1-e^2)(1-b)/2]$。

计算机辅助不等式。脚本 scripts/check_e3_three_view. py 和表格 e3_three_view/rows.csv 证明，三度和二度的转向上限分别为 0.274 和 0.334。利用 (45) 和 $a=0.548$，它们证明了上述两个小分量裕度 $L[1+\theta_0(q-2)]$ 以及两个大分量裕度 $L/2-C^1(\mathcal{P},B)$ 均为正。因此，任何度数下都不存在环超额。由于 $T^1_{\mathcal{P}}\geq(q-1)L$，将集合 0 替换为小分量端点后，(46) 仍然成立。四视图数值不等式见附录 E.7。

线性约束附近的失效。固定 $K\geq 3$，并将负等相关律重新缩放为 $I-(1-\delta)uu^\top$，其中 $u=\mathbf{1}/\sqrt{K}$ 且 $\delta\downarrow 0$。在归一化分块和构成的空间中，协方差有一个特征值为 δ，其余特征值均为 1，且每个真主子式都有下界。对于满足 $q\geq 3$ 的划分，利用附录 E.7 的独立对比论证，可将最优乘积参考分布约化为这些和。将其协方差写作 $V_0=\mathrm{diag}(v_i)$，且 $\beta=1/q=1-\alpha$。约化空间中的比值 $\det(\beta R+\alpha V_0)/(\prod_i v_i)^\alpha$ 在正值 v_i范围内具有一致下界。将行列式按 v_i 展开时，所有系数均非负。对于 $\prod_i v_i\leq 1$，保留关于 $\max_i v_i$ 的线性项；根据主子式界，其系数一致为正。然后使用 $\max_i v_i\geq(\prod_i v_i)^{1/q}\geq(\prod_i v_i)^\alpha$。对于 $\prod_i v_i>1$，保留乘积项。选择$V_0=I$ 可为该比值的下确界提供一致上界。亲和度公式给出

$$R_{\mathcal{P}}=\frac{\log(1/\delta)}{2(q-1)}+O_K(1).$$

最小割具有 $I_*/2=\frac{1}{4}\log(1/\delta)+O_K(1)$，因为其分子由真主子式构成。因此，单元素划分最终达到 L 对于 $K>3$。对于$K=3$附录 中的严格比较 E.5，可得出相同结论。

由置换不变性和唯一性，单例参考为 vI。边缘一致性方程为

$$1=\frac{K-1}{K(\alpha v+\beta)}+\frac{\delta}{K(\alpha v+\beta\delta)},\qquad \alpha=1-1/K.$$

第一项意味着 $v\geq(K-2)/(K-1)>0$。取极限可得 $v\to(K-2)/(K-1)$，而倾斜纵向方差的阶为 δ。在公共盒所包含的内在精度处，根据该倾斜的两个特征值，对角元之外的单例倾斜相关性为 $-1/(K-1)+O_K(\delta)$。因此，绝对转向行和为 $1-O_K(\delta)$。任何满足 $a<1$的有效上限都必须有 $1-a^2=O_K(\delta)$，且两步量 $D(\mathcal{P},B)$ 具有正下界。因此，$C^1(\mathcal{P},B)\geq\log(1/\delta)-O_K(1)$。对于此划分，基预算 $(q-1)L$ 已经等于 $T^1_{\mathcal{P}}$，所以（46）式的第二个不等式

不成立。这一对该判据的限制与附录中通过另一种估计为三视图确立的尖锐阈值相容 E.5。

E.10. 相关性 $-1/3$时的临界窗口界

命题 14. 对于相关性相等的三个视图 $-1/3$，令 $R_3 = R_{1|2|3}$ 且 $s_n = dR_3 - \log n$。若 $s_n \to +\infty$，则最大似然以趋于 1 的概率成功。若 $s_n + \frac{3}{4}\log\log n \to -\infty$，则每个估计器的成功概率都趋于零。

上侧。 必要时，仅使用 $\lceil (\log n + \log\log n)/R_3 \rceil$ 个坐标。这使得 $s_n \to +\infty$ 并确保 $d/\log n \to 1/R_3$。将式（24）中的真实子集检验阈值和候选项子集检验阈值替换为 $t_k = \log\binom{n}{k} + a_n k$，其中 $a_n = s_n/4$ 且 $t_0 = 0$。与之前相同，利用一阶矩可知，真实表失效的概率至多为 $\sum_{k\ge 1} e^{-a_n k} = o(1)$。对于所有 $m \le \sqrt{n}$，式（30）中固定的正小连通分量裕度已足够。

对于 $m \ge \sqrt{n}$，取偶数短环截断 $\Gamma = A\log\log n + O(1)$，其中 A 为一个足够大的常数。附录 E.5 的长环界为 $\delta_\Gamma = O(0.556^\Gamma)$，因此$d\delta_\Gamma = o(1)$。在 (27) 中，余项对 Γ 的依赖为 $O(m) + O(\Gamma h)$：提取出显示项后，支撑集、行划分和标记位置的代价为 $O(m)$，而每个标记端点的选择数至多为最大度数为三的图中半径-Γ 球的大小。对于 (28) 中的固定律边代价 C_{edge}，$C_{\mathrm{edge}}/R_3 < 1/2$。因此，对于足够大的 n，每个标记对数概率的贡献至多为 $dC_{\mathrm{edge}} - \log m + O(\Gamma) \le 0$。

复用的计算机辅助不等式。 对于上侧，`scripts/check_d1_k3_sharp.py` 和 `d1_k3_sharp/rows.csv` 证明 $R_2 > R_3$，$(2/9)\log(25/16) > R_3$，以及 $C_{\mathrm{edge}}/R_3 < 1/2$。最后一项由两个向外的比较论证 $1/R_3 < 10.2$和 $10.2 C_{\mathrm{edge}} < 1/2$给出。这些是不等式中使用的固定律不等式，见附录 E.5。

根据（25），$\log\binom{n}{k} \le k(1-\theta)\log n + O(m)$，以及 $R_2 \ge R_3$，剩余的对数概率有界于

$$\begin{aligned}(m+k)(\log n - dR_3) + a_n k + O(m) + md\delta_\Gamma \\ \le -ms_n/2\end{aligned}$$

当 n 足够大时。对 m、k、h 的多项式数量的类别求和，并结合 $m \ge \sqrt{n}$，得到 $o(1)$。可行似然规则能够成功，而根据贝叶斯最优性，全数据最大似然规则也能够成功。

下侧。 令 P、V、Q 分别表示真实分布、最优单例乘积参考分布和 $Q = Z^{-1}p^{2/3}v^{1/3}$。记 $D = D(Q\|P)$、$T = D(Q\|V)$、$\ell = \log(q/p)$ 以及 $\tau = \log(q/v)$。逐点等式

$$\ell + \tau/2 = -\tfrac{3}{2}\log Z = R_3$$

意味着 $D + T/2 = R_3$。此处 $T > 0$，因为倾斜高斯协方差的非对角项非零。取一个足够大的固定 w，令 $A_d = \{dD \le \sum_{j=1}^{d} \ell_j \le dD + w\}$。

在 Q 下，对数比 ℓ 具有正方差和有限的三阶绝对中心矩，因为它是一个非常数高斯二次型。Berry–Esseen定理将其与相应正态区间概率的差界定为一个常数乘以 $d^{-1/2}$。该正态概率渐近等于 $w/[\sqrt{2\pi d}\,\mathrm{sd}_Q(\ell)]$, , 因此选择足够大的 w，即可得到 $Q^{\otimes d}(A_d) \asymp d^{-1/2}$。

对于下侧，我们可以将 d 增大到至少 $\lfloor(\log n - \log\log n)/R_3\rfloor$。下侧前提仍然成立，这只会对估计器有利，并且现在有 $d \sim \log n/R_3$。以条件概率 $e^{-dD-w}q^{\otimes d}/p^{\otimes d}$ 在 A_d 上独立地标记每个真实行，在其外标记概率为零。根据区间的上边界，该概率至多为 1。因此，一个标记行服从 $Q_A = Q^{\otimes d}(\cdot \mid A_d)$，标记行的平均数量为 $ne^{-dD-w}Q^{\otimes d}(A_d)$。由于 $D < R_3$，除去子多项式因子后，该数量按 n 的正幂增长。二项分布集中性给出

$$\log m = \log n - dD - \tfrac{1}{2}\log d + O(1)$$

以趋于 1 的概率成立。

揭示所有视图中的标记集合，以及所有未标记行的完整对应关系。剩余的相对置换是标记行的两个独立均匀置换。在给定辅助信息的条件下，其真实匹配处的三个视图元组相互独立，服从分布 Q_A：独立标记使每一行分别满足条件，而原始均匀置换与未置换的数据及标记相互独立。在 A_d 上的逐点恒等映射给出 $\sum_j \tau_j \le dT$。因此， Q_A 相对于 $V^{\otimes d}$ 的似然比至多为 $\exp(dT + \frac{1}{2}\log d + O(1))$。在参考分布下，两个隐藏置换仍然均匀，并且与观测值相互独立，因为 V 按三个视图进行因子分解。因此，某个估计器成功的概率至多为 $(m!)^{-2}$。进行换测度并使用 Stirling公式，可得条件成功概率上界

$$\min\{1, \exp[m(dT + \tfrac{1}{2}\log d + O(1)) - 2\log(m!)]\}.$$

在集中性事件上，指数为 $2m[s_n + \frac{3}{4}\log d + O(1)]$。由于 $\log d = \log\log n + O(1)$，在给定的下侧前提下，该值趋于负无穷。因此，即使拥有额外信息，成功概率也趋于零。

E.11. 轴对称三视图带：陈述与证明概要

本小节及下一小节将陈述关于猜想11 的另外两个结果，并概述其证明。与附录 E.1–E.10不同，这两节不包含完整证明：我们将介绍各个论证的结构、新要素及其所依据的公式，并省略推导过程。

令 $K = 3$ 具有单位方差， $r_{12} = r_{13} = r$ 和 $r_{23} = s = r + \delta$；正定性条件为 $|s| < 1$ 和 $e = 1 + s - 2r^2 > 0$。记 $u = r^2$, $I_i = I(X_i; X_{[3]\setminus\{i\}})$，于是 $I_1 = -\frac{1}{2}\log\frac{e}{1+s}$ 和 $I_2 = I_3 = -\frac{1}{2}\log\frac{(1-s)}{1-u}^{e}$，并且在乘积参考分布上取 $T_\alpha = \min_V D_\alpha(P\|V)$，因此 $R_{1|2|3} = T_{2/3}/2$ 和 $L = \min\{T_{2/3}/2, I_1/2, I_2/2\}$。最优参考分布为 $\mathrm{diag}(y, z, z)$，其中 z 是由驻点方程得到的三次方程在 $[v_{(2)}^2, 1]$ 中的唯一根；这是命题12(b)中视图 2 和 3 对称的情形。

定理 *33*（轴对称带；证明概要）。对于每个固定的 $r \in [-11/30, -3/10]$ 以及每个满足 $2r^2 - 1 - r < \delta < 1 - r$的 δ，只要 $\liminf d/\log n > 1/L$，最大似然估计就能以趋于 *1* 的概率恢复两个排列。结合定理 *3(b)* 和定理 *9*，$1/L$ 是一阶阈值。在 $r = -1/3$ 处，L 中的最小值为 $I_1/2$，其中 $\delta \le \delta_a$在 $[\delta_a, \delta_b]$上，$R_{1|2|3}$ 在$[\delta_b, 2/3]$上，而 $I_1/2$ 在 $[2/3, 4/3$上），其中 $\delta_a \approx -0.0251207582$ 和 $\delta_b \approx 0.0439919612$。

论证遵循附录E.9的方案：解码器还会运行第5节的子集检验，逐边标记短环，每个标记可节省 $\log m$；在置信传播不动点处计算高斯积分，并在$\theta = \log m / \log n$中进行两层计数。以下介绍新增要素，以及各自适用的参数区域。

两种倾斜与一个共用消息盒。我们使用两种配置：平方根似然（在三单例行处为 $\alpha = 1/2$）和阶-2/3 子集检验；对于划分为两块的行，两者都采用阶-$1/2$ 倾斜。将每个视图按倾斜边际方差重新缩放后，消息仍处于一个共用的方差区间内，转向权重由显式的 p_{ij}，q_i 控制；而一个范数为 $a < 1$ 的非回溯比较矩阵，对所有长度超过两步的返回进行求和。每个被标记的端口需支付 $\log(1 + d/(1 - a^2))$，其中 d 是其两步向量的平方范数。锚点颜色为 1 或 2 时，标记费用由这两种端口成本中的较大者承担，再扣除双块行的超额速率 $P_i - L$, $P_i = \log \frac{1 \mp w_i}{2w_i}$。若 $A = T_{1/2} - L > 0$，平方根成本 $C_0 \le A$ 且 $\max\{A - L,\ A + 2(R_{1|2|3} - L) - C_1\} \ge 0$，则涵盖每种分量大小，因为平方根界处理 $\theta \le \min\{L, A\}/L$，其余部分则由子集检验处理。

有色短环。在 $\delta = -0.32$ 附近，两步电荷过于粗略。对于四种根类型，枚举根节点三步以内的有色邻域，可得到所有经过该根节点的四环和六环的联合权重，以及由三步向量给出的较长回返的一个界；这将平方根论证延伸至 $\delta \approx -0.3586$at $r = -1/3$。

奇异端点：彩色电阻。</style}对于 $s < r < 0$，写作 $x = -s$，并将$R = \mathrm{diag}(D)^{1/2}(I - zz^{\top})\mathrm{diag}(D)^{1/2}$ 表示为，其中 $z_1^2 = u/(u + x)$、$z_2^2 = z_3^2 = x/(1 + x)$ 以及 $\tau = \|z\|^2 < 1$。将候选项中具有相同行对的边视为三种颜色的平行边。于是

$$2B_G = m \log \xi + \log \det(\mu I + \mathcal{L}_G), \qquad \xi = \frac{\tau^2}{4(1 - \tau)}, \quad \mu = \frac{2(1 - \tau)}{\tau},$$

其中，$\mathcal{L}_G$ 是具有彩色电导z_i^2/τ的拉普拉斯矩阵；将其除以轻色电导后，得到 $(1,\ Q,\ Q)$，接地条件为 $M, Q = {}^{x}_{(u+x)\,u(1+x)} \ge 1$ 以及$M = \frac{2xe}{u}(1+x)$。将 G 与局部树状提升进行比较，$2(mT_{1/2} - B_G)$是提升与基图的归一化对数行列式之差的极限。按颜色逐一恢复标记；对数行列式关于某个标记电导的导数就是其有效电阻，

在覆盖树上，该值由最不利的交替射线给出的上界所限制，其电导由 $g_p = \mu + S(q, g_q)$、$g_q = \mu + S(p, g_p)$、$S(a,b) = ab/(a+b)$的不动点确定；而在基图中，则由狄利克雷测试势给出下界。对于带有并联重伙伴的轻标记，两个外部邻点处的电势 $\pm v, v = Q/($ $M+3Q+2)$给出基图中的半电阻界 $(2t+M+3Q-Q^2/(M+3Q+2))^{-1}$。随后，每个标记的计费都小于$T_{1/2}-F, F = I_1/2$，当 $T_{1/2} > 2F$时，计数以裕度$(m-h)(T_{1/2}-2F)$在 $\theta = 1$ 处闭合。接下来，利用 $e = 0$ 单一射线界 $\log(1+2\sqrt{2Q}) < F$即可一次性计入所有标记。这里 $T_{1/2} - 2F = \frac{1}{2}\log \min_{0<z\le 1} \frac{(z+2-D)^2(z+D)(z+e)}{16(2-D)Dz^2}$，其中 $D = 1+s$；该值随 D 递减，随 e递增。

公共标记与中间阶。 当两种配置单独都不足以满足要求时，它们使用相同的标记；对于具有 k 个三单例行和 h 个标记的候选项，$t = \min(h, k)/k$和 $y = L \log m/\log n$，每个三单例行的裕量为

$$M_0 = A - y + t(y - a_0), \qquad M_\alpha = T_\alpha - (1+\lambda)L - (1-\lambda)y + t(y - a_\alpha), \qquad \lambda = \frac{2\alpha - 1}{1-\alpha},$$

第二个来自阶数为 $\alpha \in [1/2, 2/3]$ $\sqrt{p}\ p/q_\alpha{}^{\gamma} = -\ e^{-T_\alpha/2} q_\alpha/\sqrt{v}\ \ 1+\gamma\ \ 1-\alpha\ = 1/2$ $a_0 \le A$ $L - A\ \ a_\alpha - A_+ \le$ 时的子集检验，依据恒等映射 ( ) , ( )( ) . 只需满足 and ( )( ) $(T_\alpha - 2L)(L - a_0)$. 阶数 $11/20$, $3/5$ 和 $13/20$ 用于 $r = -11/30$附近；在三色图上，阶数为 α 的积分仍是一个接地拉普拉斯行列式，其电导为 $(1,\ Qj_1/j_2,\ Qj_1/j_2)$，接地为 $Mj_1, j_i = 1 + (1/(2\alpha) - 1)/\tilde{v}_{i\circ}$

正端。 当 $s^2(1+s) \ge 2r^2$ 时，等价地，当 $I(X_2; X_3) \ge I_1$ 时，先通过双视图最大似然对齐视图 2 和视图 3，再将对齐后的视图对与视图 1 匹配；这两个步骤都至多需要 $2/I_1 \le 1/L_\circ$ 这涵盖了带状区域上的 $\delta \ge 4/5$ 。

证书。 通过向外舍入区间运算，在 $(r,\ \delta)$ 中的有理数闭盒上验证这些有限维不等式。scripts 及其记录的表格位于辅助文件中，其中的 README 列出了这些文件：它们是名称中包含 `d1_axial` 的 scripts，以及 `check_d1_outer_left_short.py` 和 `check_d1_thirtieth_determinant.py`。重新运行这些 scripts 会逐字节复现记录的表格。 r 中的区间端点由证书选定，而该定理是对每个固定协方差成立的逐点陈述。

E.12. 端点处的解析论证与开邻域：陈述及证明概要

对于负等相关族，参数化 $r = -3x/(2 + x + 2x^2)$，$0 < x < 1/2$，如定理 10所述。附录 E.5 证明了该定理，并由此证明了关于整个族的猜想 11，其中部分论证使用了经认证的数值计算。这里我们介绍在两个端点处不使用经认证数值计算的论证，以及阈值在 R的小扰动下的稳定性。如附录 E.11一样，这里只给出概要。

命题 34 (端点处的解析论证；证明概要)。若 $0 < x \le 1/3$ 或$0 < 1 - 2x \le 2^{-10}$，则 $L = R_{1|2|3} < I_*/2$，且当$\liminf dR_{1|2|3}/\log n > 1$ 时，最大似然以趋于 *1* 的概率恢复两个置换；这一结论可由包含显式有理常数的论证得出。每个这样的等相关 R 在三视图相关矩阵空间中都有一个邻域，对于其中每个固定矩阵，$L = R_{1|2|3}$，且同样的结论成立。

在奇异端点附近 ($0 < 1 - 2x \le 2^{-10}$)，倾斜作用于每一行。在每侧具有 m 行的候选项的三色展开中，带符号非回溯矩阵 B的高斯积分为 $\det(I - xB)^{-1/2}$，而加权 Ihara–Bass 公式给出

$$\det(I - xB) = (1 - x^2)^m x^{2m} \det(\mu I + L), \qquad \mu = \frac{1}{x} + 2x - 3,$$

等价地，$[(1 - x^2)^2(1 - 4x^2)]^m \det(I + \zeta \mathcal{A})$，其中 $\zeta = 9x^2/((1 - x^2)(1 - 4x^2))$。在含有三单例行的连通分量中，种子边构成一片森林；通过腔体电导与大围长提升进行比较论证，可将每个标记的电荷界定为 x的一个显式函数；所得的闭合不等式可通过初等估计对 $1 - 2x \le 2^{-10}$ 予以证明，并且从数值上看，它一直持续到 $1 - 2x \approx 2.5 \times 10^{-3}$。

接近零点时 ($0 < x \le 1/3$)，划分为两个块的行使用平方根因子，而三单例行使用倾斜，因此被积函数混合了两种精度。对混合归一化进行带符号迹展开表明，只有长度至少为四的偶数回路会出现；对于经过三单例行的环上的 M个标记，其代价至多为 $MC_m + h_0 C_h$，对于其他 $C_m < \Delta$ 个标记则至多为 h_0，其中 $C_h < C_0$ 和 Δ，C_0 是关于 x的显式函数；这些结论由在整个范围内证明的多项式不等式给出；例如，在此范围内有 $I_*/2 - R_{1|2|3} > \frac{7}{50} x^3$。

开邻域。最优乘积参考分布关于 R连续，因此所有速率和严格预算在附近仍然成立；扰动矩阵的转向权重逐项受一个具有略大 x 的等相关比较分布所支配，并且上述积分恒等式保持不变，因为每种颜色在一条边的两端使用相同的方差。该论证没有给出显式半径。

方法的局限。在 $x = 7/20$，也就是 $r = -70/173$，这两种估计都失效：围绕平行边对的闭游走迫使全支撑倾斜中的 $\log \mathcal{I} \ge (m - k)P$，而混合代价超过其预算（$C_m \approx 0.0354 > \Delta \approx 0.0261$）。这是这两种估计的局限，并非反例；根据定理 10，该处的猜想仍然成立。

附录 F。定理15的证明

在本附录中，我们采用标准化模型：每个对象在每个坐标 $\ell \in [d]$上的 K向量分布为 $N(0, R_K(\rho))$，其中 $\rho = \rho_n \to 0$, $d/\log n \to \infty$且$b_n = d\rho^2/\log n \to b$。Wee 等人Wee 等人(2026年的模型只需将所有视图乘以同一个常数即可化为此模型，而这不会改变两两求和的最大化元。令 $H = \sqrt{d}\log n$；则 $d\rho = \sqrt{b_n}\, H$。

F.1. 树解码器

对于 $K=2$, $J=I(X^1;\ X^2)=(\frac{1}{2}+o(1))d\rho^2\,J/\log n\to b/2$。若 $b>2$，定理1(a) 给出一个双视图估计器，其在每条树边上的 $o(n)$ 错误率。沿着从视图1 到视图 a 的路径进行复合，误差至多为该路径上误差之和。若 $b<2$，定理2结合 $K=2$ 表明，任何针对 σ_{ac} 的双视图估计器的期望重叠度都趋于零；这适用于每条树边上的估计。

F.2. 类型与成对和的最大值

固定一个承载对象 o 的锚定行 i，并将候选项 t 写作 $T(t)=\sum_{a<c}\langle Y^a_{ja},\ Y^c_{jc}\rangle$。$t$ 的恒等划分 $\mathcal{P}$ 称为其 类型，其中有一个大小为 s 的锚块 $B_0\ni 1$，承载o，以及 $r=|\mathcal{P}|-1$ 个承载不同对象 o_1、.....、$o_r\neq o$ 的自由块。对于一对视图$e=\{a,c\}$，令 S_e 为包含 a 或 c 的自由块集合，因此当且仅当 $e\subseteq B_0$ 时，$S_e=\varnothing$。令 $m_0(\mathcal{P})=\sum_{B\in\mathcal{P}}\binom{|B|}{2}$ 表示块内的对象对数，并令 $M=\binom{K}{2}$。定义

$$V(\mathcal{P})=\max\Big\{\sum_{e:S_e\neq\varnothing}x_e:\ \tfrac{1}{2}\sum_{e:\,\varnothing\neq S_e\subseteq U}x_e^2\le|U|\ \text{ for all }\varnothing\neq U\subseteq\{1,\dots,r\}\Big\},\tag{50}$$

与 $V(\mathcal{P})=0$（如果 $r=0$）。

引理 35。对于每种类型 $\mathcal{P}$，最大值$_{t\text{ 类型为 }\mathcal{P}}T(t)/H=m_0(\mathcal{P})\sqrt{b+V(\mathcal{P})}+o_P(1)$。

证明。对锚点进行条件化。对对象 o 的特征 $z=(z^a)_{a\in[K]}$进行条件化。每个 $\langle z^a,\ z^c\rangle$ 的方差为 $O(d)$，因此以 $1-O((\log n)^{-1/2})$ 的概率，我们有 $|\langle z^a,\ z^c\rangle-d(\mathbf{1}_{a=c}+\rho\mathbf{1}_{a\neq c})|\le\sqrt{d}(\log n)^{1/4}$对所有 a，c均成立；我们称此类 z为典型。由于 $\sqrt{d}(\log n)^{1/4}=o(H)$，$e\subseteq B_0$ 对所述类型的所有候选项而言，相应的成对项随后都对 T贡献 $\binom{s}{2}\sqrt{b_n}H+o(H)$，且该贡献相同。对于其他成对项，令 X_e 为内积减去其条件均值后再除以 H所得的值。对于自由块内的一对，条件均值为 $d\rho$；在其他情况下则为 0。

中等偏差。令 $\mathcal{E}$ 表示在一个候选项中出现的不同内积的集合，或表示在两个候选项中出现的不同内积的集合，其中这两个候选项的自由对象来自不同块的不相交池。我们声称，对于每个固定的 $t\in\mathbb{R}^{\mathcal{E}}$，在典型的 z上均匀成立，

$$\frac{1}{\log n}\log\mathbb{E}\Big[\exp\Big(\log n\sum_{e\in\mathcal{E}}t_eX_e\Big)\,\Big|\,z\Big]\to\frac{1}{2}\sum_{e\in\mathcal{E}}t_e^2.\tag{51}$$

给定 z 时，各坐标相互独立。在坐标 ℓ 中，自由变量构成高斯向量 g_ℓ，其协方差为 $C=I+O(\rho)$，而 $\sum_e t_eX_e$ 限制在坐标 ℓ 上等于$H^{-1}\big(g_\ell^\top Bg_\ell-\mathrm{tr}(BC)+c_\ell^\top g_\ell\big)$。其中，$B$ 是一个固定对称矩阵，在每个二次项对的两个位置上携带 $t_e/2$，而 c_ℓ 关于 z_ℓ 是线性的。令 $v=\log n/H=\sqrt{\log n/d}\to0$、$M_0=C^{1/2}BC^{1/2}$ 和 $\tilde{c}_\ell=C^{1/2}c_\ell$。配方后，坐标 ℓ 的条件对数矩母函数为

$$-v\,\mathrm{tr}\,M_0-\tfrac{1}{2}\log\det(I-2vM_0)+\tfrac{1}{2}v^2\tilde{c}_\ell^\top(I-2vM_0)^{-1}\tilde{c}_\ell=v^2\,\mathrm{tr}\,M_0^2+\tfrac{1}{2}v^2\|\tilde{c}_\ell\|^2+O\big(v^3(1+\|z_\ell\|^2)\big).$$

对 ℓ 求和并除以 $\log n = v^2 d$，得到 $\operatorname{tr} M_0^2 + (2d)^{-1}\sum_\ell \|\tilde{c}_\ell\|^2 + O(v)$。由于 $\rho \to 0$，在二次项对上，$\operatorname{tr} M_0^2 \to \operatorname{tr} B^2 = \frac{1}{2}\sum t_e^2$。在典型的 z 上，$(2d)^{-1}\sum_\ell \|c_\ell\|^2 \to \frac{1}{2}\sum t_e^2$，遍历连接锚块与自由块的二次项对，因为 $\langle z^a, z^{a'}\rangle/d \to \mathbf{1}_{a=a'}$ 以及进入同一自由变量的不同此类二次项对分别来自不同的锚定视图。这就证明了（51）。根据通常的倾斜论证，对于围绕 x、边长为 η 的盒，我们得到 $\mathbb{P}(X \in \text{盒} \mid z) = n^{-\frac{1}{2}\|x\|^2 + O(\eta\|x\|) + o(1)}$，并且对典型的 z 均匀成立。

上界。当 η 较小时，至多有 n^{K-1} 个候选项，且每个 $|X_e|$ 以概率 C 超过常数 $n^{-C^2/2+o(1)}$。因此，以高概率，所有候选项都位于一个固定立方体内。对于每个非空 $U \subseteq [r]$，子向量 $(X_e)_{\varnothing \neq S_e \subseteq U}$ 仅取决于 U中各块的对象，而这些对象至多有 $n^{|U|}$ 种选取方式。将该立方体覆盖为小盒。根据对这些选取方式应用并集界，对于某个 $\frac{1}{2}\sum_{\varnothing \neq S_e \subseteq U} x_e^2 > |U| + \eta$ 满足 U 的盒，以高概率不包含任何候选项。因此，所有候选项都位于（50）的可行集附近，并扩大了 η。令 $\eta \to 0$，并利用值的连续性即可得到上界。

下界。令 x^* 达到 $V(\mathcal{P})$ 和 $x = (1-\alpha)x^*$，从而对每个 U 都有 $I_U = \frac{1}{2}\sum_{\varnothing \neq S_e \subseteq U} x_e^2 < |U|$。为每个自由块分配一个包含 $\lfloor (n-1)/r \rfloor$ 个除 o之外对象的不相交池，并令 Z 统计来自这些池且其向量位于 x周围一个小盒中的候选项。于是 $\mathbb{E}Z = n^{r - I + O(\eta) + o(1)}$，其中 $I = I_{[r]} < r$。恰好在 U 的块上共享对象的两个候选项，恰好共享关于 $S_e \subseteq U$ 的内积；不相交池排除了任何其他重合。将(51)应用于它们的并集可知，这样的一对以概率 $n^{-2I + I_U + O(\eta) + o(1)}$落入该盒中，并且共有 $O(n^{2r-|U|})$ 个这样的候选对。不共享任何对象的候选对在给定 z 的条件下相互独立。因此，当 η 足够小时，$\operatorname{Var} Z/(\mathbb{E}Z)^2 \le o(1) + \sum_{U \neq \varnothing} n^{I_U - |U| + O(\eta) + o(1)} \to 0$，所以 $Z > 0$ 以高概率成立。令 α、$\eta \to 0$，即可得到下界；去除关于典型 z 的条件限制便完成证明。 □

引理36。设 $\mathcal{P}$ 是一种类型，其锚块的大小为 $s \ge 2$；将锚块的 $s-1$ 个非锚视图移入一个新的自由块后得到 $\mathcal{P}'$。于是$m_0(\mathcal{P}') = m_0(\mathcal{P}) - (s-1)$ 以及 $V(\mathcal{P}') \ge V(\mathcal{P}) + \sqrt{s(s-1)}$。

证明。连接锚点与被移动视图的 $s-1$ 个配对离开这些块，而被移动视图之间的 $\binom{s-1}{2}$ 个配对留在一个块内。取一个针对 $\mathcal{P}$ 的最优 x，并保留其在旧配对上的取值。旧锚块的 $\binom{s}{2}$ 个配对现在具有S_e = {新的}；为每个配对赋值 $\sqrt{2/\binom{s}{2}}$，这恰好使用一个单位的预算，并使目标函数增加 $\sqrt{2\binom{s}{2}} = \sqrt{s(s-1)}$。旧配对只能在其$S_e$中获得新块。因而，不涉及新块的约束只会减少项。针对$U \ni$ 新的约束涉及带有 $S_e \subseteq U \setminus \{$新$\}$的旧配对，代价至多为 $|U| - 1$，再加上新配对的代价 1。

□

定理 15(ii) 的证明。在 $K/(K-1)$ 以下。对于具有 $s \ge 2$ 的类型，引理35和36表明，类型 $\mathcal{P}'$ 的最佳候选项至少超过类型 $\mathcal{P}$ 的最佳候选项 $\sqrt{s(s-1)} - (s-1)\sqrt{b} - o_P(1)$，单位为 H。由于 $b < K/(K-1) \le s/(s-1)$，该裕度为正。由于 $\mathcal{P}'$ 具有锚块 $\{1\}$，且类型数量有限，行

最大化器以趋近于 1 的概率具有锚块 $\{1\}$；也就是说，每个非锚点条目都是错误的。

补全。令 F 为原始提议表，$U_a \subseteq [n]$ 为视图 a 中没有任何锚点行提议的行的集合。视图 a 中的每个被提议行恰好保留一个锚点行，因此视图 a 中待补全的行恰好是 U_a。假设冲突消解和补全是 F 及独立随机性的函数。将视图 a 的行按任意一个支撑集为固定集合 U 的置换 τ 重新标记，不会改变数据和真实分配的联合分布。在事件 $U_a = U$ 上，这也不会改变 F，因为原始最大化具有等变性、并列的概率为零，且没有提议指向 U。因此，在给定 F 和随机性的条件下，真实行位于视图 a 中 U_a 内的对象，其行在 U_a上均匀分布。因此，未匹配的 $|U_a|$ 锚点行在期望意义下至多接收一次其真实行。保留的提议正确的次数至多与原始提议一样多，因此，根据锚点行的可交换性，$\mathbb{E}\mathsf{ov}(\widehat{\pi}_a) \le \mathbb{P}$（给定行的最大化器在视图 a中是正确的）$+ 1/n \to 0$。

两个非锚点视图。对于 $a,\ c \ge 2$，统计 $\widehat{\pi}_c \circ \widehat{\pi}_a^{-1}$通过锚定行得到的命中数。在视图 a 或视图 c 中由补全填充的行，依据上述对称性，每个视图的期望命中数至多为一。在两个视图中都保留、且其提议来自不同对象的行没有贡献。在两个视图中都保留、且其提议来自同一个错误对象 o' 的行，需要 a 和 c 位于获胜类型的同一个自由块中。(50)的每个最优解的所有坐标都有某个 $\kappa_K > 0$ 作为下界：如果某个坐标为零，将其增加 δ 并按 $(1+\delta^2/2)^{-1/2}$ 重新缩放，就能保持所有约束，因为 $|U| \ge 1$，同时提高目标函数。根据引理 35 的证明，获胜者与其类型的最优解集合之间的距离在 $o_P(1)$ 以内。因此，除去期望大小为 $o(n)$ 的一组行外，这样的命中会迫使 $\langle Y^a_{\pi_a(o')},\ Y^c_{\pi_c(o')} \rangle \ge d\rho + \kappa_K H/2$。根据切比雪夫不等式，每个对象以 $O(1/\log\ n)$ 的概率满足这一条件，且每个对象至多产生一次命中。因此，复合估计的期望重叠度也趋于零。</output>

以上是 $K/(K-1)$。这可由 Wee 等人（2026）的定理 2.5 得出，也可直接从引理 35 看出。根据柯西–施瓦茨不等式，$V(\mathcal{P}) \le \sqrt{2r(M - \binom{s}{2})} \le \sqrt{2rM}$，而错误类型的均值会损失 $e(\mathcal{P})\sqrt{b}$，其中 $e(\mathcal{P}) \ge r(2K-r-1)/2$ 是其分离对的数量。比值 $8M/(r(2K-r-1)^2)$ 在 $r=1$ 或 $r=K-1$ 时最大，此时分别等于 $K/(K-1)$ 和 $4/K$。因此，对于 $b > K/(K-1)$，每种错误类型以高概率都会造成损失，而坏行的数量为 $o_P(n)$，任何补全都能容忍这一数量。

□

备注 37（重读数据的补全）。取 $K=3$和 $4/3 < b < 3/2$，并用 H归一化。根据引理 35，真实元组的得分为_$3\sqrt{b}$，保留一个正确非锚点视图的元组得分至多为 $\sqrt{b}+2$，使用两个不同错误对象的元组得分至多为 $\sqrt{12}$，而将锚点与其他两个视图中的同一个错误对象配对的最佳元组得分为 $\sqrt{b}+\sqrt{6}$，是这些得分中最大的。由于交叉项 $\langle Y^1_i, Y^2_{\pi_2(j)} + Y^3_{\pi_3(j)} \rangle / H$ 对所有错误对象 $j \ne o$均至多为 $2 + o_P(1)$，因此，典型行的获胜元组在视图 2 和 3 之间的内部内积至少为 $d\rho + cH$，其中某个固定的 $c > 0$。只有渐近消失比例的对象具有如此大的内部内积，因此，在每个非锚点视图中，原始提议仅覆盖 $o(n)$ 个互异行。补全可以先计算

定理 1 在整个表格上的分区检验解，然后通过交换输出行来强制保留 $o(n)$ 提议。每次交换至多会改变两个锚定行，并且不会影响此前已强制保留的提议，因为保留的提议使用互异行。结果是近乎精确的。因此，关于算法 1 最终输出的陈述必须明确指定补全规则。

注38 （各向异性谱）。当 ρ_ℓ 不相等时，未加权的两两求和会得到 $(\sum_\ell \rho_\ell)^2/(d\log n)$，这可能远小于 $A/\log n$。按坐标 ℓ 以 ρ_ℓ 加权，可恢复信号 $A/\log n$。若满足 $h \to 0$ 和 $h\sqrt{\log n/A} \to 0$，则上述分析适用于 $H^2 = A\log n$；当 $A/\log n$ 趋于正数时，后一个条件自动成立。若不满足这类扩散条件，结论可能不成立，例如相关性全部集中在单个坐标上时便是如此。

附录 G. 定理17 的证明

G.1. 筛选种子视图

固定 $3 \le m \le K$ 并记 $A = \sum_\ell \rho_\ell^2$ 和 $h = \max_\ell \rho_\ell$。对于 $2 \le s \le m$，令 P_s 为前 s 个视图的真实元组的分布。对于 $[s]$ 的每个非平凡划分 $\mathcal{P}$，令

$$Q_{s,\mathcal{P}} = \bigotimes_{B\in\mathcal{P}} P_B, \qquad L_{s,\mathcal{P}} = \log \frac{dP_s}{dQ_{s,\mathcal{P}}}, \qquad c(\mathcal{P}) = \sum_{a<b} \mathbf{1}\{a,b \text{ in different blocks of } \mathcal{P}\}.$$

在级别 s，如果每个 $L_{s,\,\mathcal{P}}$ 都超过 $\alpha c(\mathcal{P})\log n$，则元组通过检验，其中 $2/m < \alpha < b/2$。对于每个锚点行和视图 $a = 2,\dots,m$，在视图 $1, a$ 上形成通过相应双视图检验的行列表。如果某个列表的长度大于 $D_{\max} = \lceil n^{1-2/m} \rceil$，则丢弃相应的锚点行。从剩余的锚点开始，用下一个视图的锚点列表扩展每个保留下来的前缀，并对新的长度应用检验。长度为二时的检验已在为视图 2 生成列表时完成。在中间层级 $s < m$，如果超过 $2n^{u_s}$ 个前缀存活，则使用恒等置换中止，其中 $u_s = \max\{1, s - s(s-1)/m\}$。在级别 m，扫描通过检验的元组，并保留其行均未被使用的每个元组。这样便得到一个不相交元组的极大集合。最后，按照索引顺序，将每个视图中未使用的行与未使用的锚点行进行匹配。

真实元组。在坐标 ℓ 处，C_ℓ 和 B_ℓ 分别是 P_s 和 $Q_{s,\,\mathcal{P}}$ 下的协方差矩阵。这两个矩阵在 $\mathcal{P}$ 的对角块上相同，因此 $\operatorname{tr}(B_\ell^{-1}C_\ell) = s$。对高斯散度进行二阶展开可得

$$\begin{aligned}\mathbb{E}_{P_s} L_{s,\mathcal{P}} &= \tfrac{1}{2}\sum_\ell \log(\det B_\ell / \det C_\ell) \\ &= \tfrac{1}{2} c(\mathcal{P}) A + O_m(hA).\end{aligned}$$

事实上，与恒等映射的差的迹为零，其弗罗贝尼乌斯范数的平方在坐标 ℓ 处相差 $2c(\mathcal{P})\rho_\ell^2$。余项为 $O_m(\rho_\ell^3)$，总和为 $O_m(hA)$。根据

引理 30 方差为 $O_m(A)$。由于 A/对数 $n \to b$ 且 $\alpha < b/2$，每个真实检验的均值都比其阈值高出一个固定正倍数的 对数 n。对每个检验（包括所有锚点列表检验）应用切比雪夫不等式和并集界，得到$o(1)$ 每个真实元组的失败概率。根据马尔可夫不等式，未通过这些检验中任一项的真实元组数量为 $o_{\mathbb{P}}(n)$。

锚点列表。给定隐藏置换，错误锚点对服从两个行边缘分布的乘积分布 Q 。其对数似然比 L 具有 $\mathbb{E}_Q e^L = 1$，因此通过的概率至多为 $n^{-\alpha}$。每个列表至多有一个正确元素，且期望错误元素数至多为$n^{1-\alpha}$ 。根据马尔可夫不等式，对每个列表分别应用马尔可夫不等式并结合并集界，锚点被丢弃的概率为 $O_m(n^{-(\alpha-2/m)})$。因此，仅有 $o_{\mathbb{P}}(n)$ 个锚点被丢弃。

前缀计数。给定隐藏置换，具有恒等划分的错误元组 $\mathcal{P}$ 服从分布 $Q_{s,\ \mathcal{P}}$。若 $\mathcal{P}$ 有 $p \geq 2$ 个块，则具有该恒等划分的元组至多有 n^p 个。无论其前缀上的检验结果如何，仅通过 $\mathcal{P}$ 检验的概率至多为 $n^{-\alpha c(\mathcal{P})}$。在将 s 个视图划分为 p 个块的所有划分中，当块大小为 $s-p+1$，1，……，1 时，跨块对数最小，因此

$$c(\mathcal{P}) \geq (p-1)(s-p/2).$$

函数 $p \mapsto p-(2/m)(p-1)(s-p/2)$ 是凸函数，其最大值 u_s 在 $[1,s]$ 处的端点取得。由于 $\alpha > 2/m$，每个非平凡划分贡献 $n^{p-\alpha c(\mathcal{P})} = o(n^{u_s})$ 个错误元组的期望数量。对有限个划分求和，可得最终层级 s错误幸存项的期望数量也满足相同上界。该计数包含在该级别通过测试的所有索引元组，因此也构成算法所达到计数的上界。由于真实元组至多有 n 个，且 $u_s \geq 1$，根据马尔可夫不等式和并集界，触及任何中间上限的概率为 $o(1)$。

选择。 由于 $u_m = 1$，在最终级别只有 $o_{\mathbb{P}}(n)$ 个错误元组通过。随着概率趋于一，不会发生中止，并且除去那些测试失败或锚点被丢弃的真实元组外，所有真实元组都能到达最终选择阶段。一个未被纳入极大不相交集合的通过测试的真实元组，会与某个选中的错误元组相交，因为不同的真实元组彼此不相交。每个选中的错误元组至多可以阻挡 m 个真实元组。因此，前 m 个视图在包含 $n-o_{\mathbb{P}}(n)$ 个锚定行的公共集合上实现对齐。

G.2. 附加剩余视图

对每个锚定行，将对齐后的前 m 个视图中对应的行相加，并将坐标 ℓ 除以 $\sqrt{m}$ $(1+(m-1)\rho_\ell)$。在前 m 个视图已真实对齐的预言机数据中，这会生成一个标准高斯变量行。对于任意固定的剩余视图，其真实配对行的坐标相关性

$$r_\ell = \frac{\sqrt{m}\,\rho_\ell}{\sqrt{1+(m-1)\rho_\ell}}.$$

在真实配对分布下，双视图对数似然比的均值为 I_m，方差为 $\sum_\ell r_\ell^2$，其中

$$I_m = -\tfrac{1}{2}\sum_{\ell} \log(1 - r_\ell^2) = (m/2 + O_m(h))A.$$

在预言机协方差中，目标视图仅在全 **1** 方向上与前 m 个视图相关联。种子对比与总和以及目标相互独立，因此总和保留了它们与该视图之间的互信息。使用固定阈值 $\beta \log n$，其中 $1 < \beta < mb/2$。在预言机数据中，根据切比雪夫不等式和马尔可夫不等式，阈值图会遗漏 $o_{\mathbb{P}}(n)$ 个真实配对。由于错误配对服从乘积分布，其通过阈值的概率至多为 $n^{-\beta}$。因此，预言机图中虚假边的期望数量至多为 $n^{2-\beta} = o(n)$。

对于在前 m 个视图中正确对齐的每个锚定行，实际图与预言机图一致。极大匹配至多可以选择一条与每个受污染锚定行关联的虚假边。其他被选中的虚假边均位于预言机图中。因此，它选择的虚假边数量为 $o_{\mathbb{P}}(n)$。在干净行上，除 $o_{\mathbb{P}}(n)$ 条真实边外，其余真实边均存在，并且每条被选中的虚假边至多会阻断其中两条边。该匹配保留 $n - o_{\mathbb{P}}(n)$ 条真实边，这些边通过补全得以保留。该比较论证逐实例成立，并使用关于预言机图的无条件界。将其应用于有限个剩余视图中的每一个，可得到一组共同的 $n - o_{\mathbb{P}}(n)$ 个正确对齐的锚点。

G.3. 算术计算成本

给定白化行后，固定的高斯检验得分等于一个常数、加权行平方和加权两两内积之和。其系数仅取决于划分和已知的坐标相关性。为前m 个视图上的所有检验（包括锚点列表检验）预计算表格的成本为 $O_m(n^2 d)$。随后，每个候选扩展均可在 $O_m(1)$次操作内完成检验，其中常数项涵盖其级别上的所有划分。级别 $s < m$包含 $O(n^{u_s})$个前缀，其中 $u_1 = 1$，且每个前缀最多可通过 $D_{\max}$种方式扩展。前缀上限为每个输入都给出了这一界限。对于整数$1 \le s < m$， u_s 的最大值等于 $\lfloor (m+1)^2/4 \rfloor / m$，这是通过最大化 $s(m+1-s)/m$得到的。因此，扩展前缀并选择互不相交的最终元组的总成本上界为 $O_m(n^{\gamma_m})$，其中

$$\gamma_m = \max\left\{2,\ 1 - \frac{2}{m} + \frac{\lfloor (m+1)^2/4 \rfloor}{m}\right\}.$$

剩余视图通过扫描其边来进行评分和匹配，耗时为 $O((K-m)n^2 d)$。总计为 $O_m(Kn^2 d + n^{\gamma_m})$。对于 $m = 3,\ 4,\ \gamma_m = 2$，从而得到二次界和所述阈值。计算或应用白化变换不计入此次算术计数。

附录 H. 定理 18 与四视图设计

H.1. 递归最弱划分解码

定理 *18* 的证明。间隔的继承。 对于节点 U 和划分 $\mathcal{P}$ U，记作$D_U(\mathcal{P})$。 $= D($ $P_U \| \bigotimes_{B\in\mathcal{P}} P_B)$。假设 $\mathcal{P}$取得 J_U并具有 p个块。通过将其中一个块用划分 $\mathcal{Q}$细分为 B个 $q \geq 2$块，可得到一个划分 $\mathcal{P}'$为 $p+q-1$个块。散度的链式法则给出

$$D_U(\mathcal{P}') = D_U(\mathcal{P}) + D_B(\mathcal{Q}).$$

左侧至少为 $(p+q-2)J_U$，而 $D_U(\mathcal{P}) = (p-1)J_U$。因此 $D_B(\mathcal{Q}) \geq (q-1)J_U$，对 $\mathcal{Q}$取最小值得到 $J_B \geq J_U$。按照约定，单元素区块具有 $J_B = \infty$。每个内部节点继承根节点的裕度。在 U 上对 p 个超视图进行划分，且其分布为 P_U，这是对 $\mathcal{P}$ 的粗化，具有相同的块数和相同的散度。其归一化散度至少为 J_U，当划分为 p 个独立的超视图时取等号。因此，此解码步骤中的共享信息恰为 J_U。

一次解码步骤。为每个节点固定一个代表视图，并将其同时用作包含该视图的子节点的代表视图。节点的输出以代表视图的行作为指标，每个成员视图中的每个原始行恰好使用一次。叶节点输出其观测到的行。在分析具有 p 个子节点的节点时，将每个子节点的行重新标记为其代表视图中的对象。若子节点 B_j 的任一成员行属于不同对象，则该子节点的一行即为受污染行。令 E_{B_j} 统计这些行，并设 $c = \sum_{j=1}^{p} E_{B_j}$。

对于预言机数据（其中每个子节点的行都已正确拼接），设 F 为未通过分区检验的真实元组数量，T 为通过所有检验的错误元组数量。对于子节点的一个非平凡划分 $\mathcal{R}$，设 $L_{\mathcal{R}}$ 为相应的对数似然比，$D_{\mathcal{R}}$ 为其真实均值。继承而来的裕度给出 $D_{\mathcal{R}} \geq (1+\epsilon)(|\mathcal{R}|-1)\log n$，因此与阈值之间的差距至少为 $(\epsilon-\gamma)D_{\mathcal{R}}/(1+\epsilon)$。由引理 30 可知，$\mathrm{Var}(L_{\mathcal{R}}) \leq 2K^2 D_{\mathcal{R}}$，因此由切比雪夫不等式和并集界可得 $\mathbb{E}F = O_{K,\,\epsilon,\,\gamma}(n/\log n)$。

以真实排列为条件。一个恒等划分为 $\mathcal{R}$ 的错误预言机元组，其乘积分布为 $Q_{\mathcal{R}}$ 相应块边缘分布的乘积分布。在该分布下，$\mathbb{E}_{Q_{\mathcal{R}}} e^{L_{\mathcal{R}}} = 1$，因此通过其自身划分检验的概率至多为 $n^{-(1+\gamma)(|\mathcal{R}|-1)}$。此类元组至多有 $n^{|\mathcal{R}|}$ 个，对各划分求和可得 $\mathbb{E}T = O_K(n^{1-\gamma})$。这两个估计针对原始分布下的预言机数据，无需以任何子节点解码器的输出为条件。

在实际数据中，至少有 $n-F-c$ 个真实元组的所有行均未受污染，并通过了测试。选定元组彼此不相交。在选中的错误元组中，至多有 c 个使用受污染行，其余均由 T 计数。任何未被最大选择选中的、通过检验且行未受污染的真实元组，都会与某个选中的错误元组相交。每个这样的错误元组至多与 p 个真实元组相交。因此，节点输出至少保留 $n-F-c-p(T+c)$ 个行未受污染的真实元组。补全会保留这些元组，从而得到逐实例界

$$E_U \leq F + pT + (p+1)\sum_{j=1}^{p} E_{B_j}.$$

从 $E_{\{a\}} = 0$ 叶节点处的结果和两个预言机计数出发，对固定规模树进行归纳可得 $\mathbb{E}E_{[K]} = O_{K,\ \epsilon,\ \gamma}(n/\text{对数}\ n + n^{1-\gamma})$。马尔可夫不等式给出 $E_{[K]} = o_{\mathbb{P}}(n)$，这意味着同时近乎精确恢复。

算术计算成本。 枚举 $[K]$ 的子集划分并计算其高斯散度，成本为 $O_K(d^3)$，其中 $d = \sum_a d_a$。在一个节点处，每个高斯检验得分都由一个常数项、每个子节点行对应的一个二次项，以及每对子节点行对应的一个双线性项组成。对于系数块 H 和行矩阵 X，Y，通过计算 $(XH)Y^\top$ 来制表记录双线性项。对所有检验和节点求和，构建表格需要 $O_K(nd^2 + n^2 d)$ 次操作。系数由已知协方差确定，而每个节点的得分表使用其可用的子节点行。在一个有 p 个子节点的节点处，扫描 n^p 个元组会检验每个候选项，并利用表格以 $O_K(1)$ 次操作更新最大选择。因此，总扫描成本为 $O_K(n^{p^*})$。当每次划分都是二元划分时，恰有 $K-1$ 个内部节点，每个节点都使用双视图阈值检验和极大匹配。

通过贪心合并恢复树。 对于一个具有唯一最弱二分划分 $U_L \mid U_R$ 的内部节点 U，记作 $\lambda_U = J_U = I(X^{U_L}; X^{U_R})$。若其子节点 W 是内部节点，则用 W 的最弱划分细分 $U_L \mid U_R$，会得到一个包含三个区块的竞争划分。根据链式法则，其归一化散度为 $(\lambda_U + \lambda_W)/2$，并因唯一性而大于 λ_U。因此 $\lambda_W > \lambda_U$。

保持如下不变量：贪心过程当前的簇是目标树中互不相交的节点，并覆盖所有叶节点。对于初始的单元素簇，该不变量成立。对于当前不成兄弟节点对的节点 A，B，令 U 为它们的最近公共祖先。根据投影下的数据处理不等式，

$$I(X^A; X^B) \le I(X^{U_L}; X^{U_R}) = \lambda_U .$$

U 至少有一个子节点仍由两个或更多当前簇覆盖。在该子节点的子树中，存在一个最深的尚未合并的内部节点 W，其两个子节点都已作为当前簇出现。合并它们的得分为 $\lambda_W > \lambda_U$。因此，每一对非兄弟节点的得分都严格低于某一可用兄弟节点对的得分，所以选择得分最高的合并操作会保持该不变量。经过 $K-1$ 步后，目标树的所有内部节点都已合并。 □

命题 39 （沿两区块划分进行解码）。 设 $\Sigma \succ 0$，并令 $\{B_1, B_2\}$ 为划分 $[K]$。假设 J_{B_1}、J_{B_2} 和 $D_{\{B_1,B_2\}}$ 均至少为$(1+\epsilon)\log n$，其中 $\epsilon > 0$ 为固定值，且对所有足够大的 n 都成立。将两个块都近乎精确地对齐，拼接每个块中对齐的行，并保留块对数似然比超过$(1+\gamma)\log n$ 的配对所构成的极大匹配，其中 $0 < \gamma < \epsilon$。经过一对一补全后，所得结果同时达到近乎精确。如果该二分划分是最弱的，则充分条件简化为 $J \ge (1+\epsilon)\log n$。

证明。 根据定理，块内间隔保证解码器1(a)。在上文的单步证明中，针对两个子节点的唯一检验，其均值为 $D_{\{B_1,B_2\}}$，阈值为 $(1+\gamma)\log n$。其预言机计数满足 $\mathbb{E}F = O_{K,\ \epsilon,\ \gamma}(n/\log n)$ 和 $\mathbb{E}T = O(n^{1-\gamma})$。同一个逐实例界现在为 $E_{[K]} \le F + 2T + 3(E_{B_1} + E_{B_2})$，因此对满足这些条件的任意划分，结论均成立。当划分最弱时，$D_{\{B_1,B_2\}} = J$，而间隔的继承给出J_{B_1}，$J_{B_2} \ge J$。 □

H.2. 四视图设计的阈值和解码器

命题 40 (聚类示例的阈值)。 在示例 *19*的均匀设计中，最佳树解码器、逐行两两求和（采用定理*15(ii)* 的证明中所述的逐视图补全）以及划分检验的近乎精确阈值分别为 12、8和6 （以 S为单位）。在聚类设计中，它们的阈值分别为72、22。5和18。相应的精确恢复阈值为8 和36。当 $S > 4.5$时，聚类设计中的簇内对可匹配；而当 $S < 18$ 时，在任何估计器下，每个跨簇对的重叠率都会消失。

证明。 在均匀设计中，每一对都有 $A_{ab}/\log n = S/6$。最佳树需要 $S/6 >$2，两两求和需要 $S/6 > 4/3$，并且 $J/\log n \to (K/4)\, S/6 = S/6$ 在弱极限中。

在聚类设计中，强对具有 $A/\log n = S/2.25$，四个跨簇对$S/36$。每棵生成树都使用一个跨簇对，这需要 $S/36 > 2$，也就是 $S > 72$。根据命题 6， $J/\log n$趋于权重为 $\frac{1}{2}A_{ab}/\log n$的图的强度。割$\{AB, CD\}$ 的权重为 $\frac{1}{2}\cdot 4\cdot S/36 = S/18$。任何其他划分都会切割一个权重为$2S/9$ 的强对，并且至多有四个块，因此其归一化权重至少为 $2S/27 > S/18$。因此，当 n较大时，最小化划分是 $\{AB, CD\}$，且 $J/\log n \to S/18$，阈值为 18。仅由该对 AB 构成的部分具有 $I_{AB}/\log n \to \frac{1}{2}S/2.25$，当 $\underline{S} > 4.5$时超过 1。

对于两两求和，考虑带有锚点 A 的类型 $\{A, B\} \mid \underline{\{C, D\}}$。相对于真实元组，它在四个跨簇对的均值上损失了 $4\cdot\frac{1}{4}\sqrt{S/2.25}$，以 H 为单位。其五个非固定对 AC、 AD、 BC、 BD、 CD 都依赖于一个自由对象，因此由 (50) 得到 $V = \sqrt{10}$。以趋于 1 的概率，当 $\sqrt{S/2.25} < \sqrt{10}$，即 $S < 22.5$ 时，该类型胜过真实元组；当 $S > 22.5$ 时则输给真实元组；这些极限在 $S = 22.5$ 本身处没有说明。引理 35 可推广到这一设计，因为其相关性非负且在各坐标间相同，只需将均值项替换为各块内部各对的 $\sqrt{A_{ab}}/\log n$ 之和。以 $\mu = \sqrt{S/2.25}$ 为单位，强对的均值为 1，跨簇对的均值为 $1/4$；并且以趋于 1 的概率，当 $V > \mu L$ 时某类型胜过真实元组，当 $V < \mu L$ 时输给真实元组，其中 L 是它损失的均值。由柯西–施瓦茨不等式，利用 $U = [r]$， $V^2 \le 2r\,(6 - \binom{s}{2})$。其他两块类型具有 $L \ge 3/2$ 和 $V^2 \le 12$，三块类型具有 $L \ge 2$ 和 $V^2 \le 24$，四块类型具有 $L = 3$ 和 $V^2 \le 36$。它们的比率 V^2/L^2 至多为 $16/3$、6 和 4，均低于 $\{A, B\} \mid \{C, D\}$ 的值 10。因此，当参数高于 22.5 时，真实元组胜过所有类型。当参数低于 22.5 时，一个典型锚点至少会提议一个错误目标，因此期望中完全正确的原始提议只有 $o(n)$。在补全过程中，每个视图为每个被提议的目标保留一个提议者，并仅使用原始提议和独立随机性来填充其余元素。原始最大化在行重标记下是等变的。根据定理 15(ii) 证明中的对称性论证，补全在每个非锚点视图中至多贡献一个期望命中。每个新得到的正确元组至少需要一次这样的命中，因此期望的完整元组重叠趋于零。

根据命题，对精确恢复而言，二分划分的互信息渐近等于其总割权重 6。在均匀设计中，每条边的权重为 $S/12$，单位为对数 n。最小割包含三条边，因此 $I_*/$对数 $n \to S/4$。在聚类设计中，割 $\{AB, CD\}$ 的权重为 $S/18$，而其他每个割都会跨越一条权重为

$2S/9 > S/18$。因此 I_*/对数 $n \to S/18$，且当 n足够大时，此割达到最小值。定理 $I_* = 2$对数 n 中的阈值 3 给出 $S = 8$（均匀设计）和 $S = 36$（聚类设计）。

{/style}可由定理 2推出：由于唯一的最弱划分将各簇分开，因此每个跨簇对都有 $J^*_{ab} \le J$。对于 $S > 4.5$，簇内对均高于双视图近乎精确阈值。

□

命题 41（聚类设计的解码）。 对于 $S > 18$，在例 *19*的聚类设计中，将 A 与 B 对齐，并将 C 与 D 对齐，使用内积得分进行最大权重指派；随后，使用双视图近乎精确解码器匹配每个所得簇中求和后的行。这样便能同时近乎精确地恢复所有相对置换。特别地，只需再进行一次最大权重指派，且三次指派的运行时间为 $O(n^3 + n^2 d)$。对于 $S < 18$，同时近乎精确恢复是不可能的。

证明。 令 s 表示簇内的相关性，因此 $S = \frac{9}{4} d s^2 / \log n$ 且 $s \to 0$，并按对象为各行编制指标。对于 $j \neq i$，在给定 A_i 的条件下，向量 $B_i - B_j$ 服从均值为 sA_i、协方差为 $(2 - s^2) I_d$的高斯分布。因此，对于 $\lambda > 0$，

$$\mathbb{E} e^{-\lambda \langle A_i, B_i - B_j \rangle} = \big(1 - 2c(\lambda)\big)^{-d/2},$$
$$c(\lambda) = -\lambda s + \tfrac{1}{2}\lambda^2 (2 - s^2),$$

且 $\lambda = s/(2 - s^2)$ 给出 $\mathbb{P}(\langle A_i,\ B_i - B_j \rangle \le 0) \le (1 - s^2/2)^{d/2} \le e^{-d s^2/4} = n^{-S/9}$。对每个簇中的 $n(n-1)$ 个有序对应用并集界可知，对于 $S > 18$，以至少 $1 - 2n^{2 - S/9} \to 1$ 的概率，A 中的每一行都更偏好其在 B 中的真实匹配行，而不是其他任何行；对于 C 和 D亦然。因此，真实对齐是最大权重匹配的唯一解，所以步骤 (i) 在两个簇中都是精确的。

考虑这样一个预言机：它接收真实的簇内对齐，并构造 $U_i = A_i + B_i$ 和 $V_i = C_i + D_i$。这些数据库具有独立同分布的行，其坐标的方差为 $2(1 + s)$，协方差为 $4 \cdot s/4 = s$，即相关性为 $\gamma = s/(2(1 + s))$。因此$I(U; V)/\log n = (\frac{1}{2} + o(1))\, d\gamma^2/\log n \to S/18 > 1$，所以根据定理 1(a)，双视图划分检验是近乎精确的。权重为 $\langle U_i, V_j \rangle$ 的最大权重匹配也同样近乎精确。事实上，高斯配对对数似然是该内积的正倍数，加上各行独立的项，而这些项的总和在不同匹配下保持不变。因此，最大权重匹配就是最大似然匹配；由于 $\gamma \to 0$ 且信息量超过 $\log n$ 并具有固定裕量，平衡双视图高斯结果适用（Dai 等，2023，定理 3.1）。对于完整的四视图块似然比，簇内的和坐标与差分坐标将簇内协方差对角化。簇间协方差只耦合求和坐标，因为它与 2×2 全 1 矩阵成正比。差分坐标与求和坐标彼此独立，且差分坐标之间也相互独立，因此划分两侧的似然比只取决于求和后的行。

在前两个匹配方案均精确的事件上，解码器与其预言机对应项相同。三个匹配方案的代价为 $O(n^3)$，其分数计算代价为 $O(n^2 d)$。下界 $S < 18$ 由命题 40 和定理 1(b) 得出。

□

均匀设计中的三个匹配方案。在固定 $S > 12$ 的均匀设计中，附录 Q 中使用的三个最大权重指派同时近乎精确。以对象为行指标，并记 $U_i = A_i + B_i$, $V_j = C_j + D_j$, $W_{ij} = \langle U_i, V_j \rangle$ 和 $\rho = \sqrt{S}\log n/(6d)$。每个簇内对象对的 $I/\log n \to S/12 > 1$，因此最大权重指派 α 和 δ（此处即双视图最大似然指派）在 $o(n)$ 行上出错（Dai 等，2023，定理 3.1 和 3.2）。令 $\mathcal{E}$ 为这样一组对象：其在 A 或 C 中对应的行受到这些错误影响，因此以概率趋于 1 有 $|\mathcal{E}| = o(n)$。第三个匹配方案 σ 最大化 $\sum_i W_{i\sigma(i)}$；除非 $i \in \mathcal{E}$ 或 $j \in \mathcal{E}$，否则 $W_{ij} = W_{ij}$。W 和 W 的每个条目都具有 $\langle A_i + B_k, C_j + D_l \rangle$ 的形式；由伯恩斯坦不等式以及对 n^4 个此类乘积使用并集界可知，随着概率趋于 1，它们的绝对值均不超过常数 C_0 乘以 $\log n$。在给定 U_i 的条件下，对于 $j \neq i$，差 $W_{ii} - W_{ij}$ 服从均值为 $(4 + o(1))\rho d$、方差为 $(8 + o(1))d$ 的高斯分布；当 $\|U_i\|^2 = (2 + o(1))d$ 对所有 i 成立时，此结论一致成立。因此，若 $c > 0$ 足够小，使得 $(\sqrt{S}/6 - c/4)^2 > 2$，则对 n^2 个有序对使用并集界可知，以趋于 1 的概率，对所有 $i \neq j$ 均有 $W_{ii} - W_{ij} \geq c\sqrt{d \log n}$。在这些事件上，将 σ 与恒等映射比较可得

$$
\begin{aligned}
c\sqrt{d \log n}\,\big|\{i : \sigma(i) \neq i\}\big| &\leq \sum_i \big(W_{ii} - W_{i\sigma(i)}\big) \\
&\leq \sum_i \big(\widetilde{W}_{ii} - \widetilde{W}_{i\sigma(i)}\big) + 8M|\mathcal{E}| \\
&\leq 8M|\mathcal{E}|,
\end{aligned}
$$

因为 $\widetilde{W}$ 和 W 仅在由 $\mathcal{E}$ 标记的行和列上有所不同，而这些行和列至多出现在 $2|\mathcal{E}|$ 项中。因此，第三种赋值在 $o(n)$ 行上出错，每个复合映射也是如此。

附录 I. 自然解码器的障碍

下文始终，$\Lambda = \sqrt{\text{取对数}}\, n$。命题 22 的部分 (a) 和部分 (b) 将给出完整证明。对于部分 (c)，我们陈述证明所依据的三个引理，并概述它们的证明及推论；概述给出主要步骤，而非每一个估计。

I.1. 显式枚举

命题 22*(a)* 的证明。 固定一个目标视图 $a \in \{2, 3\}$，令 $\ell_{i,\,a}$ 表示真实对 $(Y_i^1, Y_{\pi_a(i)}^a)$ 相对于独立行的似然比的对数。对每个坐标，它等于 $-\frac{1}{2}\log(1 - \rho^2) + \frac{\rho}{2}(G_+^2 - G_-^2)$，其中 $G_\pm$ 为独立标准变量，因此 $\mathbb{E}\ell_{i,\,a} = (b/2 + o(1))\log n$ 且 $\mathrm{Var}\,\ell_{i,\,a} = d\rho^2 = O(\log n)$。固定 $0 < \varepsilon < \min\{1 - b/2, \delta/2\}$、$w = n^{1-b/2-\varepsilon}$，以及 $u = (b/2 + \varepsilon/2)\log n$。由切比雪夫不等式和马尔可夫不等式，$\#\{i : \ell_{i,\,a} > u\} = o_{\mathbb{P}}(n)$。

固定 $\eta \in (0,\ 1)$ 和一组 D 共计 $m = \lceil \eta n \rceil$ 个 D 锚点行，并在给定 π_a 之外的条件下。在将 D 中的行替换为独立行的参考分布下，

在给定 $(Y^{1},\ \dots,\ Y^{a})$ 和算法随机性的条件下，剩余的双射是均匀的，因为列表不会读取其他内容。因此，事件 E_D 即每个 $i\in D$ 都具有 $|L_{i,\ a}|\le w$ 且 $\pi_a(i)\in L_{i,\ a}$ 的参考概率至多为 $w^m/m!$。真实分布相对于该参考分布的似然比为 $\exp\sum_{i\in D}\ell_{i,\ a}$，因此在 $F_D=\{\ell_{i,a}\le u,\ i\in D\}$ 上，

$$\mathbb{P}(E_D\cap F_D)\le e^{mu}w^m/m!=\exp\Big(-\tfrac{\varepsilon}{2}m\log n+O_\eta(m)\Big).$$

对至多 2^n 个集合取并集 D，可得 $B_{a,n}=\#\{i:|L_{i,a}|\le w,\ \pi_a(i)\in L_{i,a}\}=o_{\mathbb{P}}(n)$。确定性地，对于一个被扫描到的真实三元组，某个视图中的列表长度至多为 w，否则至少需要 w^2 次访问，因此 $H_n\le B_{2,n}+B_{3,n}+T_n/w^2$，且 $T_n/w^2\le n^{1-\delta+2\varepsilon}=o(n)$。

对于满足 $b>0$ 的上侧情形，取 $0<\kappa<b/2$，并使用原始内积最大的 $\lceil n^{1-b/2+\kappa}\rceil$ 个元素组成列表。给定锚定行 x，归一化真实得分为 $N(\rho\|x\|,1-\rho^2)$，而 $n-1$ 个错误评分是相互独立的标准正态变量；当 $t<\sqrt{b}$ 且 $t^2/2>b/2-\kappa/2$ 时，高于真实得分 $\ge t\Lambda$ 的错误评分期望数量至多为 $n^{1-t^2/2}$，因此根据马尔可夫不等式，真实行被排除在列表之外的概率至多为 $n^{-\kappa/2}$。

□

I.2. 半边对偶价格

对于每个元组 $v=(i_1,\ \dots\dots,\ i_K)$，线性规划都有一个变量 $z_v\ge 0$；每个视图中都有单位行约束，目标函数为 $\sum_v z_v\sum_{a<c}s_{ac}(i_a,\ i_c)$，其中 $s_{ac}(i,\ j)=(X_i^a)^{\top}R_{ac}X_j^c$。价格 $u_{a,\ i}=\frac{1}{2}\sum_{c\ne a}s_{ac}(i,\ \tilde{\imath})$ 使每个真实的对偶约束都取等号。

命题 *22*（*b*）的证明。令 $S|S^c$ 为一个最小化割，$W=R_{SS^c}$，$A=\|W\|_F^2$ 且 $h=\|R-I\|_{\mathrm{op}}$。在块内进行白化，并展开 $-\frac{1}{2}\log\det(I-G^{\top}G)$，得到 $A=(2+O(h))I_*=(2c+o(1))\log n$。对于对象 $i\ne j$，在 S 的视图中取对象 i，并在 S 以及 j 的视图中取对象 S^c 所构成的错误元组，其对偶约束违反量为

$$Z_{ij}=F_{ij}-\tfrac{1}{2}D_i-\tfrac{1}{2}D_j,\qquad F_{ij}=(X_i^S)^{\top}WX_j^{S^c},\quad D_i=F_{ii},$$

由于块内得分相互抵消；$\mathbb{E}Z_{ij}=-A$。作为堆叠行的一个二次型，Z_{ij} 的系数矩阵的弗罗贝尼乌斯范数平方为 $\binom{A}{2}(1+\frac{1}{4}+\frac{1}{4})=\frac{3A}{4}$；对于在相同角色上共享一个对象的两对，其和的 $\binom{A}{2}(1+1+1+\frac{1}{4}+\frac{1}{4})=\frac{7A}{4}$，均值为 $-2A$；所有算子范数均为 $O(h)$。因此，对于有界的 t，此类二次型的累积量生成函数一致为 $t\mathbb{E}L+t^2\|M\|_F^2+O(hA)$。利用 $t=2/3$ 和在 $t=2/3+\alpha$ 处的指数倾斜应用切尔诺夫方法，可得 $\mathbb{P}(Z_{ij}>0)=e^{(-1/3+o(1))A}$；利用 $t=4/7$ 应用切尔诺夫方法，可得 $\mathbb{P}(Z_{ij}>0,\ Z_{kl}>0)\le e^{(-4/7+o(1))A}$，适用于共享一个对象的各对。

将对象划分为两半 $I_1,\ I_2$，并令 $N=\sum_{(i,\ j)\in I_1\times I_2}\mathbf{1}\{Z_{ij}>0\}$。不相交对彼此独立，而 $n_1n_2(n_1+n_2-2)$ 个有序对共享一个对象，且该对象在两对中扮演相同角色。然后 $\mathbb{E}N=n^{2-2c/3+o(1)}\to\infty$ 和

$$\frac{\operatorname{Var}N}{(\mathbb{E}N)^2}\le\frac{1}{\mathbb{E}N}+O(n^{-1})e^{(2/3-4/7+o(1))A}=\frac{1}{\mathbb{E}N}+n^{-1+4c/21+o(1)}\to 0$$

对于 $c<3$，因此以趋近于一的概率 $N>0$。 □

I.3. Sinkhorn软平均：证明概要

用 $\mathcal{N}(H)$ 表示对 e^H 进行 T 交替的行归一化和列归一化，并以列归一化结束，从而每个表格都具有单位列和；初始表为 $Q_a^0 = \mathcal{N}(\beta_0 S_{a1})$，视图 a 的一次更新为 $Q_a = \mathcal{N}(\beta F_a)$，其中包含 $F_a = S_{a1} + \sum_{c \neq a} S_{ac} Q_c$，并按 Gauss–Seidel顺序 2，3，4，5 读取，以及 $\beta/\Lambda \to \sqrt{b}$。该字段读取为 $F_a(j,\ i) = \langle X_j^a,\ \bar{P}_i \rangle / \sqrt{d}$，其原型为 $P_i = X_i^1 + \sum_{c \neq 1,\ a} \sum_r \bar{Q}_c(r,\ i) X_r^c$。在分析中，按真实对象重新标记行。

引理 42（初始表的平方质量）。对于每个固定的归一化次数 $\mathbb{E}\|Q_c^0\|_F^2/n \ \leq\ n^{-(1-b)/2+o(1)}$，训练坐标表同样成立，只需将 b 替换为训练信号。

证明概要。对于似然核 $A_{ji} = e^{\ell(x_i, y_j)}/n$，其中 $\mathbb{E}_{\text{ind}} e^{2\ell} = M_2 = (1-\rho^2)^{-d} = n^{b+o(1)}$，行和与列和平均偏离 1 达 $O(\sqrt{M_2/n + M_2/n})$，一次归一化步骤满足 $\|B' - A\|_1 \leq 2\|B - A\|_1 + D(A)$，并且对于元素属于 $[0,1]$ 的表格，$\|Q\|_F^2 \trianglelefteq \|\min(A,\ 1)\|_F^2 + 2\|Q - A\|_1$。原始核与似然核相差一些行因子和列因子，其对数一致为 $o(\log n)$，且每次归一化至多使逐元素比值的对数加倍。 □

引理*43*（矩形中的错误分数）。固定 $0 < \tau < \sqrt{2}$ 和 $\varepsilon \bar{>} 0$。存在 $\xi > 0$，使得对于 $0 < \delta < \xi$ 和固定的 $\eta \in (0, 1/2)$，以趋于 *1* 的概率，对于所有非负列随机表 Q_3, Q_4, Q_5（它们可能依赖于全部数据）以及所有至少包含 ηn 个标签的不相交集合 A, B，某个列 $i \in A$ 中至少有 $c_\eta n^{1-\tau^2/2-\delta}$ 行 $j \in B$ 满足 $F(j, i) > (\tau - 4\varepsilon)\Lambda$。在同一事件中，对于某个常数 $c > 0$，每一列 $i \in [n]$ 中至少有 $c\, n^{1-\tau^2/2-\delta}$ 行 $j \neq i$ 满足 $F(j, i) > (\tau - 4\varepsilon)\Lambda$；这里，行遍历整个 $[n] \setminus \{i\}$，而非某个指定集合。

证明概要。以四个源视图为条件；目标行 $\mu_j + \sigma G_j$ 具有独立标准 G_j。对于固定的（i、c、r），事件 $\{S_{21}(j,\ i) > \tau\Lambda,\ S_{2c}(j,\ r) < -\varepsilon\Lambda\}$ 在 j 中相互独立，且概率至多为 $n^{-\tau^2/2-\varepsilon^2/2+o(1)}$，因此根据 Chernoff界并对 $O(n^2)$ 个三元组取并集，至多有 $n^{1-\tau^2/2-\xi}$ 行同时满足这两个事件。将 A 划分为每块含 $n^{\tau^2/2+\delta}$ 列，Sudakov–Fernique 不等式和高斯 Poincar´e 不等式表明，在给定行中，一个块不存在锚点分数高于 $\tau\Lambda$ 的情况，其概率为 $O(1/\log\ n)$；利用行的独立性，并对至多 4^n 个矩形取并集，可知对于每个（A、B），A 的某一列中都有许多较高的锚点分数。对于单列 i，给定源视图后，行 $j \neq i$ 的锚点分数相互独立，且以至少 $n^{-\tau^2/2-\delta/2}$ 的概率超过 $\tau\Lambda$，因此根据 Chernoff界并对 n 列取并集，可得所有行上的计数。由于每个表格的列和均为单位值，三个表格在这些行中赋予分数低于 $-\varepsilon\Lambda$ 的总权重为其行数的 o；而在其余行中，三个回流项之和至少为 $-4\varepsilon\Lambda$。在给定这两个计数事件后，此论证是确定性的，因此对所有表格都同时成立。 □

引理 44 (真实得分)。在前两轮中的每一轮里，更新后的表格赋予均匀真实标签的权重以概率趋于零，并且第一个更新后的视图的场满足 $F(I,\ I)/\Lambda \to \sqrt{b}$（以概率计），其中均匀标签为 I。

证明概要。这里结合了三个事实。(i) 行扩散。对于某个 $\kappa>0$，在对初始表格以及视图 2 的第一遍表进行每次中间列归一化之后，均匀行的最大项至多为 $n^{-\kappa}$，且该事件发生的概率趋于一；对于初始表格，这是引理 42 的结果，而对于视图 2，则来自一个对所有正行因子同时成立的列分母矩形下界。(ii) 重采样稳定性。如果一个 logit 矩阵中的一行发生变化，并且两个版本在每个中间阶段的行最大值都至多为$\eta<1/2$，则每一步的列势变化至多为 4η；大小为 $\Delta=o(1)$ 的全矩阵变化会使列变化 $O_T(\Delta)$，其中 ℓ_1。因此，对均匀目标行的高斯残差进行重采样，会使场所读取的每个表在列 ℓ_1 中变化 $O_{\mathbb{P}}(n^{-\kappa}(\log n)^3)$，而使用重采样原型对原始残差进行评分可得

$$F(J,I)=\rho\sqrt{d}\Big(\mathbf{1}_{\{J=I\}}+\sum_{c}q_c(J,I)\Big)+O_{\mathbb{P}}(1)+O_{\mathbb{P}}(\rho\Lambda).$$

(iii) 权重比较。若对每个固定的 δ，低于$\bar{}-\delta\Lambda$ 的元素仅有 $o_{\mathbb{P}}(n^2)$ 个，当且仅当 $F(I,I)/\Lambda\to\sqrt{b}$，那么经过任意正行因子和列归一化后，对每个 $\lambda<1-b$，均匀真实权重至多为 $n^{-\lambda}$：按行因子对标签排序后，低因子列有线性多个来自高因子行的竞争者，其得分至少为 $-\delta\Lambda$。从引理 42 出发沿更新顺序归纳，即可得到这些结论；在第二遍中，$F_2^2(I,I)-S_{21}(I,I)=o_{\mathbb{P}}(\Lambda)$。 □

命题 22(c) 的证明提纲。选择 $\sqrt{b}<\tau<\sqrt{2(1-\bar{}\omega)}$、$\varepsilon$，使得 $\overline{\tau-4\varepsilon}>t_0>\sqrt{b}$ 且 $\delta<\bar{}\min\{\xi,1-\tau^2/2-\omega\}$。根据引理44，只有 $o_{\mathbb{P}}(n)$ 个真实得分超过$t_0\Lambda$。若按行因子 r_j 和列因子缩放后，剩余列中有固定比例 γ 的列仍将真实行排在前 W 名之内，则将这些标签划分为低因子列 A 和高因子行 B；引理43 给出一个 A 列，其中有超过 $Cn^{\omega}\ge W$ 行属于B，且它们的得分和行因子均高于真实元素，矛盾。引理43 只用到列随机表，因此适用于两遍运行中的两个阶段。补全边只有在均匀相对置换的不动点处才正确，这类点有 $O_{\mathbb{P}}(1)$ 个，因此舍入重叠趋于零。两遍运行中，第一个目标的表在舍入前不会再次更新；而在默认运行中，它会在第三遍更新，这不在覆盖范围内。第二遍的错误得分界是适用于所有列的引理43。

对于留出陈述，令 G 表示其他视图的训练数据、它们的置换以及算法随机性。给定 G，留出原型的协方差为$V=(1-\rho)(I+\sum_c q_c^{\top}q_c)+\rho\mathcal{A}^{\top}\mathcal{A}$，与目标的交叉协方差为 $\rho\mathcal{A}$，其中 $\mathcal{A}=I+\sum_c q_c$。将交叉协方差替换为 ρI，即可得到一个参考实验，它等价于一个相关性为 $\rho/\sqrt{1-\rho}$、加上独立噪声的干净的双视图问题。通过 $V^{-1/2}$ 进行白化后，两个交叉块的算子范数均为 $O(\sqrt{\rho})$，因此每个坐标的相对熵为 $O(\rho^2\|\mathcal{A}-I\|_F^2)$；根据引理 42，总量为 $o(n\log n)$。在参考实验中，证明部分 (a) 时采用的测度变换计数（列表长度为一）表明，匹配 ηn 行的概率为 $e^{-\Omega_{\eta}(n\log n)}$。数据处理界 $D(\mathbb{P}\|\mathbb{Q})\ge p\log(1/q)-\log 2$ 随即迫使真实分布下匹配 ηn 行的概率趋于零，对每个固定的 η 均成立。 □

Ap附录 J。已知观测分布下的近乎精确匹配 s

沿用第 2 节的采样与置换模型，其中 K 固定。视图的特征可以取值于任意可测空间，且这些空间乘积上的共同元组分布 P_n 已知。记 $P_{n,\ B}$ 为其在 $B \subseteq [K]$ 上的边缘分布。对于将 B 划分为至少两个块的划分 $\mathcal{C}$，令

$$Q_{B,\mathcal{C}} = \bigotimes_{A\in\mathcal{C}} P_{n,A}, \qquad D_{B,\mathcal{C}} = D(P_{n,B}\|Q_{B,\mathcal{C}}), \qquad J_B = \min_{\mathcal{C}} \frac{D_{B,\mathcal{C}}}{|\mathcal{C}|-1}.$$

对于单元素集合 B，令集合为 $J_B = \infty$。与正文一样， $J = J_{[K]}$ 且 $J^*_{ab} = \max_{B\supseteq\{a,b\}} J_B$。令 f 和 g 分别为 $P_{n,B}$ 和 $Q_{B,\mathcal{C}}$ 相对于它们之和的密度，并记 $L_{B,\mathcal{C}} = \log(f/g)$。在 $\{f=0\}$ 上取 $L_{B,\mathcal{C}} = -\infty$，并在 $\{f>0,\ g=0\}$上取 $L_{B,\mathcal{C}} = +\infty$。我们对真实元组使用以下条件：对于所指集合 B 的每个划分 $\mathcal{C}$ 以及每个固定的 $\eta > 0$，均需满足该条件：

$$P_{n,B}(L_{B,\mathcal{C}} < \min\{D_{B,\mathcal{C}}, K\log n\} - \eta\log n) = o(1). \tag{52}$$

定理 45（已知观测分布）。 *(a)* 将 $[K]$ 划分为 $m \geq 2$ 个块，记为 $\mathcal{C}$。从每个块 A中选取一个代表 r_A，其中对于包含视图 1的块 A_1，取 $r_{A_1} = 1$。对于每个 $n \geq 2$以及每个估计器，

$$\frac{\sum_{A\neq A_1} \mathbb{E}M_{r_A}}{(m-1)n} \geq 1 - \frac{D_{[K],\mathcal{C}}}{(m-1)\log n} - \frac{2}{\log n}. \tag{53}$$

当 $D_{[K],\mathcal{C}} = \infty$ 时，该界是空泛的。如果 $\limsup_n J/\log\ n < 1$，则不可能同时实现近乎精确恢复。

(b) 对于一个至少包含两个视图的固定集合 B，假设 $\gamma_n > 0$ 满足 $\gamma_n \log n \to \infty$，并且对于将 B 划分为至少两个块的每种划分 $\mathcal{C}$，

$$P_{n,B}(L_{B,\mathcal{C}} \leq (1+\gamma_n)(|\mathcal{C}|-1)\log n) = o(1). \tag{54}$$

那么，采用置换补全的划分检验解码器可在 B上同时实现近乎精确恢复。特别地，$\liminf_n J_B/\log\ n > 1$ 以及(52) 在 B 上足以满足条件。

(c) 假设(52) 对每个至少包含两个视图的子集都成立。如果$\liminf_n J^*_{ab}/\log\ n > 1$，则 σ_{ab} 的置换估计在概率意义下的重叠度趋于一。如果$\limsup_n J^*_{ab}/\log n < 1$，则$\sup_f \mathbb{E}\mathrm{ov}(f) \to 0$，其中上确界取遍使用所有视图的随机映射。

（*d*）假设 上极限$_n$ J/对数 $n < 1$ 并且，对于每个充分大的 n，存在一个将 $[K]$ 与 a 分隔开的 b最小化划分。如果 （52）对于 $[K]$的划分成立，那么上确界$_f$ $\mathbb{E}$ ov（f） $\to 0$ 对于这对对象成立。

（a）–（d）各部分共同给出了定理 23：每个至少包含两个视图的集合都满足条件（52），即为条件（4）。

J.1. 有限样本逆命题与划分检验

定理45（*a*）的证明。假设 $D_{[K],\,c}<\infty$。揭示每个块内的相对置换 $\pi_a\pi_{r_A}^{-1}$ for $a\in A$，统称为 G。使用 G 将每个块与其代表对齐，并将所得数据称为 Y'。令 $U=(\pi_{r_A})_{A\neq A_1}$。原始置换与 $(U,\ G)$ 构成双射，因此 U 在 $\mathfrak{S}_n^{m-1}$上均匀分布，且与 G独立。原始数据可由 $(Y',\ G)$ 重构。

令 Q 为块数据库的乘积分布，其中每个块 A包含分布为 $P_{n,\ A}$的独立行。给定 $(U,\ G)$ 后，撤销各块的行置换会将 Y' 变为 n 个分布为 P_n的独立元组，并保持 Q 不变。因此，对于每个条件取值，$D(\mathcal{L}(Y'\mid U,\ G)\|Q)=nD_{[K],\,c}$。对 $(U,\ G)$ 求平均，并通过 $\mathcal{L}(Y'\mid G)$ 分解密度，得到

$$I(U;Y'\mid G)+\mathbb{E}D(\mathcal{L}(Y'\mid G)\|Q)=nD_{[K],\mathcal{C}}.$$

给定映射 $[n]\to[n]$，固定真实置换与其不同的 t 输入。如果保留的像有重复，则不存在相容置换。否则，t 自由列至多允许 $t!$ 个匹配方案。可能的排列至多有 $\binom{n}{t}t!$ 个 $\le n^t$。对于$\widehat{\ }$向量 U 个估计映射，先编码错误计数，再编码真实排列：

$$H(U\mid\widehat{U},G)\le(m-1)\log(n+1)+(\log n)\sum_{A\neq A_1}\mathbb{E}M_{r_A}. \tag{55}$$

$\widehat{\ }$估计器的随机种子是独立的，因此 $I(U;\,U\mid G)\le I(U;\,Y'\mid G)$。将这些不等式与 $H(U\mid G)=(m-1)\log(n!)$、$\log(n!)\ge n\log n-n$ 以及 $\log(n+1)\le n$ 结合，可得（53）。对于 $\limsup_n J/\log n<1$，最小化划分在（53）中具有一个正的固定间隔。同时近乎精确恢复将使每个有界错误比例的期望趋于零，从而使左侧趋于零。□

定理证明45（*b*）。对于一个固定的划分，简写为 $Q=Q_{B,\,c}$ 和 $L=L_{B,\,c}$。按照上文的记号，f、g 为，

$$\mathbb{E}_Qe^L=\int_{\{g>0\}}f\,d(P_{n,B}+Q)\le1,\qquad Q(L>t)\le e^{-t}.$$

$P_{n,\,B}$ 的奇异部分支撑于 $\{L=+\infty\}$ 上，而该集合的 Q-测度为零。

对于每个锚定行，当候选元组在每个划分中都超过 (54) 中的阈值时，解码器便接受该候选元组。由有限个划分上的并集界可知，真实元组未通过的概率为 $o(1)$。在给定置换的条件下，如果错误元组的各行在 $\mathcal{C}$ 的不同块中来自不同对象，而在每个块内来自同一对象，则其分布为 $Q_{B,\ c}$。每个锚定行至多对应 $n^{|\mathcal{C}|-1}$ 个这样的候选项，因为锚点

对象是固定的，而其他每个块各选一个对象。任一候选项被接受的概率至多为 $n^{-\gamma_n(|\mathcal{C}|-1)}$。对各划分求和可得 $o(1)$。

每个带有接受的元组的锚定行都会提出一个元组。如果某行的真实元组被接受，且所有错误元组都被拒绝，则该行是良好行。坏行的期望数量为 $o(n)$。在每个视图中，每列保留一个提议，并通过对剩余的行和列任意取一个双射来完成匹配。良好行所提议的真实列彼此不同。每个坏行在每个视图中至多会排挤一个良好行。因此，每个视图中错误的数量至多为坏行数量的两倍，由马尔可夫不等式可得同时近乎精确恢复。

对于固定间隙推论，选择一个固定的 $\gamma > 0$，使得 $J_B \geq (1+2\gamma)\log n$ 最终成立，并且 $(1+2\gamma)(K-1) \leq K$。对于 $s = |\mathcal{C}| - 1$，式（52）中的截断散度至少为$(1+2\gamma)s\log n$。取 $\eta = \gamma/2$ 会使阈值 $(1+\gamma)s\log n$ 严格低于（52）中的界，从而证明（54）。

对于有限正的 $D = D_{B,\mathcal{C}}$，方差条件 $\mathrm{方差}_{P_{n,B}} L_{B,\mathcal{C}} = o(D^2)$ 足以推出（52）。下尾事件要求与均值 D 至少相差 $\eta \log n$，当$D \leq 2K\log n$时如此，否则至少相差 $D/2$。在这两种情况下，切比雪夫不等式均给出 $o(1)$。在 $D = 0$处，信息密度几乎必然为零。□

对于散度有限的高斯分布，引理 30 给出 $\mathrm{方差}_{P_{n,B}} L_{B,\mathcal{C}} \leq 2K^2 D_{B,\mathcal{C}}$，因此上述两个切比雪夫界为 $O(\log n)$。当散度无穷大时，附录 A 中的高斯支撑论证给出 $L_{B,\mathcal{C}} = +\infty$ 在 $P_{n\,\text{下几乎必然成立},B}$。因此，（52）对高斯分布成立，定理1(a) 和 2 由定理 45得出。

J.2. 固定坐标分布

推论 *24* 的证明。固定 B，$\mathcal{C}$。记 p_B 为单坐标边缘分布，$q = \bigotimes_{A\in\mathcal{C}} p_A$，并令 ℓ 为它们的扩展对数似然比。负部可积。事实上，记 f 为 p_B相对于 q 的绝对连续部分的密度，得到

$$\mathbb{E}_{p_B}\ell^- = \int_{\{0<f<1\}} f\log(1/f)\,dq \leq e^{-1}.$$

按坐标重新分组后，d 坐标的块乘积参照测度为 $q^{\otimes d}$。以 $(p_B+q)^{\otimes d}$ 作为共同支配测度表明，在真实分布下，对于独立副本，行信息密度为 $\sum_{j=1}^{d}\ell_j$。每个真实求和项都属于 $(-\infty, +\infty]$。

如果 $D(p_B\|q) < \infty$，则 ℓ 的两部分都是可积的。弱大数定律给出 $L_B、_C = dD(p_B\|q) + o_{\mathbb{P}}(d)$。如果 $D(p_B\|q) = \infty$，那么对于固定的 $A > 0$，$\ell \wedge A$ 是可积的，并且 $\mathbb{E}(\ell\wedge A) \uparrow +\infty$，当 $A \to \infty$ 时成立。对于每个固定的有限目标，选择 A 使得这个均值超过该目标，然后将弱大数定律应用于截断和。原始和支配该截断和，从而得到 $L_B、_C/d \to +\infty$ 依概率成立。结合 $d/\log n \to c \in (0,\infty)$，这些极限对任意 B、$\mathcal{C}$都给出（52）。

相同的乘积分布表示给出$D_{B,\mathcal{C}} = dD(p_B\|q)$ 在扩展实数中。因此$J_B(P_n) = dJ_B(p)$ 以及$J^*_{ab}(P_n) = dJ^*_{ab}(p)$。应用定理45（其成对结论已在第J.3节中证明）可得到两个阈值。□

J.3. 成对可匹配性

可实现性。 选择一个达到 J_{ab}^{*} 的集合 $B \supseteq \{a,b\}$，并在B内应用分区检验，其中将视图 a 作为锚点。对于视图 b 的估计是一个重叠度趋于一的置换。由于 B的选择只有有限种，因此即使最大化集合随 n变化，结论仍然成立。

划分粗化。 设 $\mathcal{C}$ 细化 $\mathcal{D}$，二者都是 B的划分。KL链式法则给出

$$D_{B,\mathcal{C}} = D_{B,\mathcal{D}} + \sum_{T\in\mathcal{D}} D\left(P_{n,T}\,\middle\|\,\bigotimes_{A\in\mathcal{C}:A\subseteq T} P_{n,A}\right). \tag{56}$$

令$D_{B,\{B\}} = 0$。为说明左侧有限时该恒等式成立，设 $P = P_{n,B}$, $V = \bigotimes_{T\in\mathcal{D}} P_{n,T}$且 $Q_{\mathcal{C}} = \bigotimes_{A\in\mathcal{C}} P_{n,A}$。投影到 T后，可将每个求和项界定为 $D(P\|Q_{\mathcal{C}})$。因此，$dV/dQ_{\mathcal{C}}$ 是粗块边际密度比的乘积。每个因子的零点集在相应的 $P_{n,T}$下均为零测集，故 $dV/dQ_{\mathcal{C}} > 0$ P-几乎处处成立。将 $dP/dQ_{\mathcal{C}}$ 除以该乘积，得到 dP/dV。根据所述散度有限以及每个对数密度比的负部可积，商中的两个对数在 P下均可积。积分即可证明式（56）。当已知右侧各项有限时，$P \ll V \ll Q_{\mathcal{C}}$ 通过将密度比相乘并对其对数积分，可得到同一恒等式。

假设 $S \subseteq B$ 满足最弱划分 $\mathcal{C}_B$ 的 m_B 个块中的 $v \ge 2$ 个块，且有限J_B。将这 v 个块合并。粗化散度至少为 $(m_B - v)J_B$；当只剩一个块时，其值为零。根据式（56），合并后的 v 个块与其乘积分布之间的联合散度至多为 $(v-1)J_B$。将每个块投影到其与S 的交集上只会降低该散度。因此，所得的 v块划分 S 给出$J_S \le J_B$。特别地，如果 $[K]$ 的有限最弱划分将 a、b分开，则任何包含这对元素的 S 都满足 $J_S \le J$。取 $S = [K]$ 即可得到 $J_{ab}^{*} = J$。这一分离性质也可由 Chan 等人（2015 年，定理5.2–5.3及命题 5.3）关于最优划分的结构推出。

归约与树。 对于逆命题，取一个固定的 $\delta > 0$，使得对所有充分大的 n，$J_{ab}^{*} \le (1-2\delta)\log n$。假设沿某个子序列，使用全部原始观测的估计量具有至少 $\eta > 0$的期望重叠度。它们可以转换为置换估计。给定一个映射 $g:[n]\to[n]$，保留每个所使用列的一个均匀选取的原像，并通过对剩余行和列进行双射来补全部分分配。如果 g 有 C 个正确项，则其列构成一个大小为 C的集合 S_g。令k_j 表示列 j的原像数量，则在给定 g 和真实分配的条件下，所保留正确项的期望数量为

$$\sum_{j\in S_g} \frac{1}{k_j} \ge \frac{C^2}{\sum_{j\in S_g} k_j} \ge \frac{C^2}{n}$$

由 $C > 0$，根据柯西–施瓦茨不等式。对于 $C = 0$，该界为零。詹森不等式给出置换估计 f，其中 $\mathbb{E}\,\mathrm{ov}(f) \ge \eta^2 =: u$。

从 $[K]$ 开始，将每个非单元素 B 按 $J_B \le (1-\delta)\log n$ 划分为最弱划分 $\mathcal{C}_B$ 的 m_B 个块。每个同时包含 a，b 的节点都会被划分，因此 a 和 b 位于不同的叶节点中。采用优先顺序，将 a 排在第一位、b 排在第二位，并选择优先级最高的视图作为每个节点的代表。一个子节点继承其父节点的代表。固定 K 时，树的数量有限，因此可进一步取一个子序列，使树保持不变。

在 a 处锚定实验，并揭示叶节点内的真实对应关系 G 。用于同时恢复逆命题的有限置换双射表明，$r-1$ 个非锚定叶节点的代表置换 $\Pi_{\mathcal{L}}$相互独立且均匀，并且独立于 G 和未打乱元组，其中 r 是叶节点的数量。重新锚定会将每个置换与新的锚定排列的逆矩阵复合，并重新排列独立同分布元组，因此这些分布律得以保持。在每个叶节点内完成对齐后，叶数据库和 G便可确定原始观测。

增广一个节点。固定 $\varepsilon > 0$，其具体取值将在下文确定。在每个内部节点 B处，对于每一对子节点以及每个对象，将同一个均匀标签从 $[M_B]$附加到这对子节点代表上，其中

$$M_B = \left\lceil \exp\left(\frac{2}{m_B}\big((1+\varepsilon)\log n - J_B\big)\right)\right\rceil .$$

这些标签在不同节点、子节点对和对象之间相互独立，且独立于原始元组。每个标签会随其代表所在的行移动。

假设已知 B 的子节点内的真实对齐。 B 处的解码器使用对齐后的原始特征以及附加在 B处的标签。将子节点划分为$v \ge 2$ 组，对应于对 $\mathcal{C}_B$进行粗化。其原始散度满足

$$(v-1)J_B \le D_{\text{old}} \le (m_B - 1)J_B < K\log n.$$

下界由 J_B 的定义得出，上界则由（56）得出。因此，下尾条件（52）给出 $L_{\text{old}} \ge D_{\text{old}} - o_{\mathbb{P}}(\log n)$ ，该结论在真实分布下成立。对于被划分分开的子节点对，其两个标签在真实分布下相同，而在分块乘积分布下是独立的均匀标签。该对子对真实信息密度的贡献为$\log M_B$ 。同一组内的对子贡献为零。因此 e 个分开的子节点对对真实元组上的信息密度和散度均贡献 $e\log\ M_B$。

对于固定的 v，当一个组包含 $m_B - v + 1$ 个子节点而其他组均为单元素集合时，分离对的数量最少。事实上，从任意一个大小至少为二的其他组中移出一个子节点，放入指定的最大组，会增加组内对的数量。因此

$$e \ge \binom{m_B}{2} - \binom{m_B - v + 1}{2} = (v-1)(m_B - v/2) \ge (v-1)m_B/2.$$

合并后的真实信息密度至少为 $(v-1)(1+\varepsilon)\log\ n - o_{\mathbb{P}}(\log\ n)$。因此，将（54）中的接受阈值设为 $\gamma_n = \varepsilon/2$ ，可得期望误差 $o(n)$，适用于每个子节点代表置换。当所有子节点都属于不同的组时，合并后的散度

等于 $(m_B-1)(1+\varepsilon)\log n + O_K(1)$，因为 $e=\binom{m_B}{2}$ 且 $0 \le \log\lceil x\rceil - \log x \le \log 2$ for $x \ge 1$。

增强散度的上界。 设 Y 包含增强叶数据库和 G。记 P_{mix} 为其植入分布，并记 Q 为叶数据库边缘分布的乘积与 G的独立副本。迭代（56）得到 $\sum_B (m_B-1) J_B$，其中这是原始单对象叶节点散度，求和遍历所有内部节点。任一内部节点的子节点代表元位于不同叶节点中。因此，其标签层对叶划分的贡献就是其单点散度。由于 $\sum_B (m_B-1) = r-1$，增强单对象叶节点散度为

$$D_{\mathcal{L}}^{\mathrm{aug}} = (r-1)(1+\varepsilon)\log n + O_K(1).$$

在给定 $\Pi_{\mathcal{L}}$ 的条件下，撤销置换后得到 n 个相互独立的增强元组，并保持 Q。对条件散度取平均可得

$$D(P_{\mathrm{mix}} \| Q) + I(\Pi_{\mathcal{L}}; Y) = n D_{\mathcal{L}}^{\mathrm{aug}}. \tag{57}$$

设 U_B 为从 B的代表元到其其他子节点代表元的 m_B-1 个相对置换组成的向量。这些向量通过沿树进行复合来确定 $\Pi_{\mathcal{L}}$，反之亦然。按子节点先于祖先节点的顺序展开 $H(\Pi_{\mathcal{L}} \mid Y)$。在节点 B处，给定此前的向量以及 Y，即可利用后代的置换和 G得到每个子节点内部的对齐。节点估计器使用原始数据以及在 B处添加的标签。其输入与 U_B 的联合边缘分布就是刚刚分析的节点实验。这可通过在 B 的代表元处重新锚定，并利用置换的独立性和增强^元组的独立同分布性得到。选择确定性的平局处理和补全方式。所得 U_B 仅依赖于 Y及此前的向量，并且在每个连通分量中的期望误差为 $o(n)$。（55）中的汉明^计数将节点条件熵界定为 $H(U_B \mid U_B) = o(n\log n)$。节点数量有限，因此 $H(\Pi_{\mathcal{L}} \mid Y) = o(n\log n)$。在（57）中使用 $H(\Pi_{\mathcal{L}}) = (r-1)\log(n!)$ 可得

$$\frac{D(P_{\mathrm{mix}} \| Q)}{n \log n} \le (r-1)\varepsilon + o(1). \tag{58}$$

正重叠度迫使下界成立。 令 B_0 为同时包含 a 和b的最低节点。根据优先顺序，它们分别代表 B_0 的两个不同子节点及其各自的叶节点。令 T 统计在根据原始数据计算出的置换估计 f 所配对的行中，它们的共享标签在 B_0 处相等的次数。在 P_{mix}下，每个正确配对都会产生相等，因此 $\mathbb{E}_{P_{\mathrm{mix}}} T \ge n u$。在每个叶节点的边缘分布下，原始特征与辅助标签相互独立。给定共同的置换后，这两个独立同分布数组仍保持其分布，且该分布不依赖于置换。在 Q下，各叶节点相互独立。因此，给定原始数据 G 及独立估计器随机种子后，这两个标签数组相互独立，且各行均匀分布。由于 f 是一个置换，各相等指示变量使用互不重叠的行对，因此 $T \sim$ 服从 $\mathrm{Bin}(n, 1/M_{B_0})$。

向两个分布都加入相同的独立种子不会改变它们的散度。相对于通过 $\exp(tT)$ 对 Q 进行倾斜所得到的分布，KL散度非负，因此

$$D(P_{\mathrm{mix}}\|Q) \geq t\mathbb{E}_{P_{\mathrm{mix}}}T - \log \mathbb{E}_Q e^{tT}.$$

对于 $t = \log M_{B_0}$，最后一项为 $n\log(2 - 1/M_{B_0}) \leq n\log 2$。由于 $J_{B_0} \leq J^*_{ab} \leq (1-2\delta)\log n$ and $m_{B_0} \leq K$，我们得到

$$\frac{D(P_{\mathrm{mix}}\|Q)}{n\log n} \geq \frac{4u\delta}{K} - o(1). \tag{59}$$

选择 $0 < \varepsilon < 4u\delta/[K(K-1)]$。这与(58)矛盾，因为 $r \leq K$，这证明期望重叠度的上确界趋于零。

由最弱划分分隔的成对节点。 假设 $J \leq (1-2\delta)\log n$ 最终对某个固定的 $\delta > 0$ 成立，并且某个最弱划分分隔了 a，b 对于每个充分大的 n均如此。将该划分作为唯一的划分，并揭示其各块内的对应关系。在上述证明中，节点所使用的唯一旧信息密度现在是该全局划分的粗化所对应的密度。它们的散度均为有限值，且至多为$(K-1)J$。根节点速率 J对$\log M_B$给出相同的下界。标签增广以及(58)和(59)中的界，在(52)条件下，仅针对全局划分便可得到趋于零的最优期望重叠度。这证明了定理45的最后一部分。

J.4. 具有变化分布的稀疏族

每个视图都观测到同一个由独立 Bernoulli$(1/n)$ 比特组成的向量 $X \in \{0,1\}^d$。由于各视图完全相同，任何划分成 q 个块的方式都有 $D_{\mathcal{P}} = (q-1)H(X)$，因此 $J = H(X) = d\,h(1/n)$，其中 h 为二元熵；并且当 $d \geq n$ 时，$J = (1+o(1))(d/n)\log n$。特别地，对于固定的 $\kappa > 1$，一旦 $d \geq \kappa n$，就有 $J \geq (1+\epsilon)\log n$。

命题46. 在这个族中，若 $d/n \to \infty$，则同时近乎精确恢复成立。若沿着一个满足 $c < \infty$ 固定的子序列有 $d \leq cn$，那么对于每一对视图以及每个估计器，期望重叠度至多为 $1 - e^{-c} + o(1)$。

证明. 给定数据后，具有相同向量的对象是可交换的：给定观测多重集后，在一组 N 个相同行上的相对置换限制是均匀的。对于任意映射 f，该组中正确匹配的像的期望数量至多为 1，因为组中的每个目标行在给定条件下正确匹配的概率为 $1/N$。全零组的大小为 $N_0 \sim \mathrm{Bin}(n,(1-1/n)^d)$，因此沿着该子序列，$\mathbb{E}N_0/n = (1-1/n)^d \geq (1-1/n)^{cn} = e^{-c} + o(1)$，且这些行中正确匹配的行的期望数量至多为 1。这就证明了第二个结论。

如果 $d/n \to \infty$，则将每个视图中向量在 n 对象中具有唯一性的每一行，与该视图中具有相同向量的行进行匹配。固定对象的向量为全零向量的概率为 $(1-1/n)^d \to 0$。给定其向量恰好含有一个 1，另一个对象与其共享该向量的概率至多为 $n \cdot n^{-1}(1-1/n)^{d-1} \to 0$；给定其含有 $w \geq 2$ 个 1，另一个对象与其共享该向量的概率至多为 $n \cdot n^{-w} \leq n^{-1}$。因此，除 $o_{\mathbb{P}}(n)$ 个对象外，所有对象的向量都具有唯一性，并且它们均能被正确匹配。 □

对于 $L_{B,\ c}$ 等于 $(q-1)\log(1/p(X))$ 对于已实现的向量 X，对于全零向量，该值仅为 $(q-1)d\log\ {}^{n}_{n-1} = O(d/n)$，尽管 $J \sim (d/n)\log\, n$；这正是命题背后（52）失效的原因。

附录 K. 部分观测对象

命题 25 的证明。逆命题。向统计学家揭示每一行的类型 M_i，以及固定类型 M 中与 $\mu(M) > 0$以外的所有对应关系。给定掩码和每种类型的行位置后，视图 M 中类型为 M 的行，是由服从分布 P_n 的 N_M 独立同分布元组构成的独立且均匀打乱的副本，并且与已揭示的所有信息独立。任何仅限于这些行的估计器都可以扩展为排列，且不会减少正确配对的数量，因此定理 1(b) 应用于 N_M 个对象后，可对其正确配对的期望数量给出上界。以趋近于 1 的概率，$\mu(M)n/2 \le N_M \le n$，因此在该事件上，一致地有 $\log\ N_M = \log\ n + O(1)$。若 $J_M < (1-\epsilon)\log\ n$，则对 M 的最弱划分，界 (2) 给出 $\Omega(n)$ 个遗漏同对象配对的期望数量；由于归一化损失有界，它不可能是 $o_{\mathbb{P}}(n)$。成对反向结论则是以相同方式应用定理 2(b)，其中 $J^{*}_{ab}(M)$。

解码器。对于每个至少包含 B 个视图的集合，针对来自 B 的每个视图各取一行的每个候选元组，检验其是否符合在级别 $(q-1)(\log n + \sqrt{\log n})$ 下将 B 划分为 $q \ge 2$ 个块的每个划分 $\mathcal{C}$，并对无穷散度使用附录 A 的支撑检验。给定掩码和洗牌，一个行来自 q 个不同对象的候选项，其分布是相应块边缘分布的乘积分布；对于每个恒等划分，此类候选项至多有 n^{q} 个。对恒等划分的似然比应用马尔可夫不等式可得，接受的错误候选的期望数量为 $F\ O_K(ne^{-\sqrt{\log n}}) = o(n)$。如果 $J_B \ge (1+\epsilon)\log\ n$，则由引理 30 的方差界和切比雪夫不等式可知，对于类型包含 B 的对象，其在 B 上的真实子元组未通过某项检验的概率为 $O(1/\log\ n)$。

将属于同一接受的元组的任意两行连接起来，保留每个视图至多包含一行的连通组，并将其他组拆分为单行。包含两个对象的行的组必然含有一个接受的错误候选，因为任何离开某个对象的路径都会经过一个错误候选；因此，至多有 KF 个对象位于混合对象的组中。每行在真实分配和估计中至多有 $K-1$ 个同对象伙伴，因此这些对象对损失的贡献为 $O_K(F)$。对于每个其他对象，如果其在 M_i 上的真实元组被接受，则会被精确输出。对每个正质量类型取 $B = M$ 可得第一个陈述；对 B 取一个达到 $J^{*}_{ab}(M)$ 的集合，则可得成对陈述，其期望为 $o(n)$，从而由马尔可夫不等式可得其概率结论。

平均公式。对于 $[K]$ 的划分 $\mathcal{P}$ 和类型 M，记 $\mathcal{P}|M$ 为 M 的诱导划分，其中包含 $q_{\mathcal{P}}(M)$ 个块。自然平均为

$$J_{\mu} = \min_{\mathcal{P}} \frac{\sum_{M} \mu(M) D_{\mathcal{P}|M}}{\sum_{M} \mu(M)(q_{\mathcal{P}}(M)-1)_{+}},$$

$D_{\mathcal{P}|M}/(q-1) \ge J_M$的比率加权平均，因此 $J_{\mu} \ge \min_{M} J_M$。取 $K = 3$，每个视图以概率 $1/2$ 独立出现，并取具有 d 独立同分布坐标的三个

视图，它们的相关性为 $1/2$，且等相关。对于每个坐标，成对信息为$I = \frac{1}{2}\log_{4}3$，三个视图的全相关为 $T = \frac{1}{2}\log 2 > 2I$。单元素划分给出比率$(3I+T)/5$，而每个二分划分给出 $(I+T)/3$，后者更大，因此$J_\mu/d = (3I+T)/5 \in (I, T/2)$，而 $J_{123}/d = T/2$。对于 $d = \lceil c\log\, n\rceil$，当 $5/(3I+T) < c < 1/I$ 时，平均值超过 $\log\, n$；类型 $\{1,2,3\}$ 具有 $J_{123} > \log\, n$，并且能够被近乎精确地恢复；每个双视图类型都有 $J^*_{ab}(M) = dI < \log\, n$，因此每个估计器在其上的重叠度都消失。 □

附录 L. 未知协方差

L.1. 未知边缘协方差

命题的证明 *26(a).* 校准。保留 $m = \lceil \sqrt{n u_n/c_n}\rceil = o(n)$ 预先确定每个视图中的行，并围绕已知的零均值计算其样本协方差。记 S_a为这些估计量，并记 $F_a = S_a^{-1/2}\widehat{S}_a S_a^{-1/2} - I$。对于固定的单位方向，归一化样本二次型为 $\chi^2_m/m - 1$，且一个 χ^2 尾界作用于一个大小至多为 $1/4$ 的网，得到 最大值$_a \|F_a\|_{\mathrm{op}} \lesssim \sqrt{(d + \text{对数}\, n)/m}$，其概率为 $1 - o(1)$。

对于一个视图，令 $A_t = [S^{1/2}(I + tF)S^{1/2}]^{1/2}$ 和 $T_t = A_t^{-1}S^{1/2}$, $0 \le t \le 1$。则 $T_t^\top T_t = (I + tF)^{-1}$，因此 $\|T_t\|_{\mathrm{op}}$在校准事件上有界。在其特征基下对 A_t^2求导，得到 $(A_t^{-1}A_t')_{ij} = \frac{a_j}{a_i + a_j}(T_t F T_t^\top)_{ij}$，其中$a_i$是 A_t的特征值。该乘子是掩码的逻辑平均 $\mathbf{1}\{\log a_i + z \le \log a_j\}$，每个掩码都是维数为 $2d_a$的上三角截断的压缩，其范数至多为 $1 + \lceil \log_2(2d_a)\rceil$。对$T_t' = -A_t^{-1}A_t'T_t$积分得到 $\|T_1 - I\|_{\mathrm{op}} \lesssim \|F\|_{\mathrm{op}}\log(2d_a)$。当条件数有界时，西尔维斯特方程 $A_1 Z + Z A_0 = S^{1/2}FS^{1/2}$ 的积分表示给出 $\|T_1 - I\|_{\mathrm{op}} \lesssim \kappa(S)\|F\|_{\mathrm{op}}$。因此，分块对角映射 $T = \mathrm{diag}(\widehat{S}_a^{-1/2}S_a^{1/2})$ 满足$e = \|T - I\|_{\mathrm{op}} = O(\sqrt{u_n/m})$，且 $q_n = u_n/(mc_n) = o(1)$。

分区得分。固定一个划分 $\mathcal{P}$ 为 $p \ge 2$ 个块，令 B 为 R的分块对角部分，$U = R - B$、$H = R^{-1} - B^{-1}$、$C = B^{-1/2}RB^{-1/2}$、$C' = R^{1/2}B^{-1}R^{1/2}$ 以及 $D = D_{\mathcal{P}} = -\frac{1}{2}$对数行列式。$R$和 B均介于 $c_n I$和 KI之间，而等谱矩阵 C，C' 的迹为 d，且特征值至多为 p。因此 $\|C - I\|_F^2 \le 4K^2 D$ 和$\|U\|_F \lesssim_K \sqrt{D}$。这些恒等式

$$H = R^{-1/2}(I - C')R^{-1/2} = -B^{-1}UB^{-1} + V, \qquad V = R^{-1/2}(C' - I)^2 R^{-1/2},$$

给出 $\|H\|_F \lesssim_K \sqrt{D}/c_n$ 和 $\mathrm{tr}\, V \lesssim_K D/c_n$。

以隐藏置换和保留行为条件。保留行中至多出现 Km 个对象；将不包含这些对象的元组称为干净元组。具有恒等划分 $\mathcal{P}$ 的干净错误元组在真实白化坐标中的 $x \sim N(0,\ B)$，其代入得分为 $L_{\mathcal{P}}(Tx) = D - (Tx)^\top H(Tx)/2$。由于 $F_B = B^{-1/2}(T - I)B^{1/2}$且 $\Delta = F_B + F_B^\top + F_B F_B^\top$，向量 $B^{-1/2}Tx$的协方差为 $I + \Delta$，其中 Δ为块对角矩阵，且$\|\Delta\|_{\mathrm{op}} \lesssim_K e/\sqrt{c_n} = o(1)$，并且

$$\log \mathbb{E}_B e^{L_{\mathcal{P}}(Tx)} = -\tfrac{1}{2}\log\det[I + (I - C)\Delta] \le C_K D e^2/c_n, \tag{60}$$

由于 $A=(I-C)\Delta$ 的迹为零，$\|A\|_{\mathrm{op}}=o(1)$ 且 $|\operatorname{tr} A^j|\le\|A\|_F^2\|A\|_{\mathrm{op}}^{j-2}$，其中$\|A\|_F\le\|I-C\|_F\|\Delta\|_{\mathrm{op}}$。对于真实元组，$x\sim N(0,\ R)$；均值的线性变化为零，因为 $RH=-UB^{-1}$ 不是块对角矩阵，而 $T-I$ 是块对角矩阵；二次变化至多为 $\operatorname{tr} V\,\|EBE^\top\|_{\mathrm{op}}+\|H\|_F e^2\|U\|_F$，其中 $E=T-I$。因此，真实得分的均值为 $D+O_K(De^2/c_n)$，方差为 $O_K(D)$。

解码。排除保留行后，若对每个划分都有 $L_{\mathcal{P}}(Tx)\ge(1-\eta_n)D_{\mathcal{P}}$，则接受该元组，其中 $\eta_n=\sqrt{q_n+(\log n)^{-1/4}}$。若 $J\ge(1+\epsilon)\log n$，则根据切比雪夫不等式，干净的真实元组被拒绝的概率为 $o(1)$；而由 (60) 和马尔可夫不等式，具有 p 块恒等划分的干净错误元组被接受的概率至多为 $n^{-(1+\epsilon/2)(p-1)}$，因此通过检验的干净错误元组有 $o_{\mathbb{P}}(n)$ 个。随后，互不相交的已接受元组的极大集合包含 $n-o_{\mathbb{P}}(n)$ 个真实元组：至多有 K^2m 个选定元组与保留对象相接触，且每个选中的错误元组至多阻碍 K 个真实元组。补全为排列后，这些匹配仍然保留。 □

L.2. 未知噪声水平

命题 *26(b)* 的证明。 在 Wee 等人 (2026年) 的模型中，Ω_{ac} 是单位矩阵的负倍数，与 $a\neq c$ 无论 τ 为何值都成立（注16），因此最大似然估计最大化 $\sum_{a<c}\sum_i\langle Y^a_{\pi_a(i)},\ Y^c_{\pi_c(i)}\rangle$，且不依赖于 τ；因此，定理 3(a) 可直接适用，而逆命题不使用解码器。

对于划分检验，写作 $v=1+\tau^2$ 和 $w=v-1$。对象的K观测值中的每个坐标的协方差为 $wI_K+\mathbf{1}\mathbf{1}^\top$，且 $v\to\infty$，因为 $\tau^4=d/(b\log n)$ 和$d/\log n\to\infty$。对 $wI_m+\mathbf{1}\mathbf{1}^\top$ 求逆，分别用于 $m=K$和划分 $\mathcal{P}$的各块，可得对于元组 $y=(y^a)_{a\in[K]}$，其块和为 $s_B=\sum_{a\in B}y^a$，

$$L_{\mathcal{P}}(y;v)=\frac{1}{2w}\Big[\frac{\|s_{[K]}\|^2}{w+K}-\sum_{B\in\mathcal{P}}\frac{\|s_B\|^2}{w+|B|}\Big]+\frac{d}{2}\,g_{\mathcal{P}}(w),$$

其中 $g_{\mathcal{P}}(w)=\sum_{B\in\mathcal{P}}\log(w+|B|)-\log(w+K)-(|\mathcal{P}|-1)\log w$。对于 $1\le m\le K$，函数 $w\mapsto 1/(w(w+m))$ 的导数绝对值至多为 $2/w^3$，且$g'_{\mathcal{P}}(w)=-w^{-1}\sum_B|B|(K-|B|)/((w+|B|)(w+K))$，因此 $|g'_{\mathcal{P}}(w)|\le K^2/w^3$。令 E_n表示如下事件：对于每个视图中的每一行，$\|Y_i^a\|^2\le 2vd$。由于$\|Y_i^a\|^2/v\sim\chi^2_d$ 且$d/\log n\to\infty$，通过一个 χ^2 尾界并对 Kn 行取并集，可得 $\mathbb{P}(E_n)\to 1$。在 E_n上，每个候选元组对于视图的每个集合 B 都有 $\|s_B\|^2\le|B|\sum_{a\in B}\|y^a\|^2\le 2K^2vd$。因此，对于每个划分以及每个满足 $v'-1\ge w/2$的 v'，

$$\Big|\frac{\partial}{\partial v}L_{\mathcal{P}}(y;v')\Big|\le C_K\,\frac{vd}{w^3}=O_K\Big(\frac{d}{v^2}\Big).$$

用 $\widehat{v}=(nd)^{-1}\sum_i\|Y_i^1\|^2$ 估计 v。由于 $nd\,\widehat{v}/v\sim\chi^2_{nd}$ 的方差为 $2nd$，切比雪夫不等式给出 $|\widehat{v}/v-1|\le\sqrt{\log n/(nd)}$，其成立概率为 $1-o(1)$。这限制了相对误差；绝对误差 $|\widehat{v}-v|$ 则大了 v 倍，而该因子趋于无穷大。在两个事件的交集上，且当 n 足够大时，介于 v 和 $\widehat{v}$ 之间的每个 v' 都有 $v'-1\ge w/2$，

以及 m　均值定理给出：对所有候选元组和所有划分，均匀地有

$$\left|L_{\mathcal{P}}(y;\widehat{v}) - L_{\mathcal{P}}(y;v)\right| \le C_K \frac{d}{v^2}\cdot v\sqrt{\frac{\log n}{nd}} = C_K\sqrt{\frac{d}{v^2}\cdot\frac{\log n}{n}} \le C_K\sqrt{b}\,\frac{\log n}{\sqrt{n}} = o(1),$$

因为 $d/v^2 \le d/\tau^4 = b$对数 n。这两个事件都不涉及候选项，因此重复使用数据来 $\widehat{v}$并无妨碍，定理 1(a) 的证明对插件评分仍然成立，只需将每个阈值平移 $o(1)$。 □

L.3. 未知正交互协方差

令 Π 在 $\mathfrak{S}_n$ 上均匀分布，并独立地令 O 在 $O(d_0)$上服从哈尔分布，且各行保持固定。给定 O后，匹配对相互独立，协方差为 $\Sigma_O = \left(\begin{smallmatrix} I & \rho O \\ \rho O^\top & I \end{smallmatrix}\right)$。记 ov 为满足 $\Pi(i) = \Pi(i)$ 的行所占比例$\widehat{\ }$。

引理 47. 对于任意估计器（可能是随机的，且已知 ρ）以及满足$0 < \alpha \le 1$的 $r = \lceil \alpha n \rceil$，

$$\mathbb{P}(\mathrm{ov} \ge \alpha) \le \min\left\{1,\ \frac{(1-\rho^4)^{-n(n+1)/4}}{\sqrt{r!}}\right\}.$$

证明。令 Q_0为两个相互独立的标准高斯视图的分布。固定 $\Pi = \pi$，并在 π下对齐各行。令 $A_O = \Sigma_O - I$，并令 ℓ_O为一对视图的密度比。高斯积分中 $\ell_O\ell_{O'}$的精度矩阵$H = \Sigma_O^{-1} + \Sigma_{O'}^{-1} - I$满足 $H \succeq [2/(1+\rho)-1]I \succ 0$和$\Sigma_O H \Sigma_{O'} = I - A_O A_{O'}$，其对角块 $I - xOO'^\top$和 $I - xO^\top O'$, $x = \rho^2$的行列式均为相等的正数。因此，单对积分为 $\det(I - xO^\top O')^{-1}$；对独立的哈尔 O，O'取平均后，

$$1 + \chi^2(P_{Y|\Pi=\pi}\|Q_0) = \int_{O(d_0)} \det(I - xU)^{-n}\, dU,$$

该积分即为 Li（2025，公式 (3.34)）中的积分。令 $\mathcal{H}_k$ 为 $\mathbb{R}^{d_0}$的 n 个向量中次数为 k 的齐次多项式。则 $\det(I - xU)^{-n} = \sum_k x^k \,\mathrm{tr}(U \mid \mathcal{H}_k)$，逐项积分可由 $\sum_k x^k \dim \mathcal{H}_k = (1-x)^{-nd_0}$保证其合理性，而哈尔平均将每个$\mathcal{H}_k$ 投影到其不变量上。根据正交群的不变量理论第一基本定理（Weyl，1939年；Lehrer 和 Zhang，2012年），不变量由$M = n(n+1)/2$ 个内积生成，其次数为二，因此该积分逐系数不超过 $(1-x^2)^{-M}$。

在参考分布 $\mathrm{Unif}(\Pi) \otimes Q_0$ 下，给定数据和估计器的随机性后，指定的 r 行被正确匹配的概率至多为 $(n-r)!/n!$，因此 ov $\ge \alpha$ 的参考概率至多为 $1/r!$。联合分布 (Π, Y) 的似然比平方的参考期望至多为 $(1-x^2)^{-M}$，由柯西–施瓦茨不等式即可得出该引理。 □

命题 26(c) 的证明。取 $\alpha = 1$：由于 $\log n! = n\log n + O(n)$ 且 $-\log(1-\rho^4) = \rho^4 + O(\rho^8)$，条件 $n\rho^4 \le (2-\epsilon)\log n$ 给出了界 $\exp[-(\epsilon/4 + o(1))n\log n]$。

如果 $n\rho^4 = o(\log\ n)$，分子为 $e^{o(n\log n)}$，而 $r! = e^{(\alpha+o(1))n\log n}$，因此对于每个固定的 α，$\mathbb{P}($ ov $\geq \alpha) \to 0$。这两个界都是关于 O 的平均值，因此对于任意估计器，它们也沿着一列固定矩阵 O 成立。在最后一个例子中，$J = -\frac{d_0}{2}\log(1-\rho^2) \sim b\log\ n$且 $n\rho^4 = 4b^2n(\log n)^2/d_0^2 = o($l og $n)$。 □

附录 M. 一般性逆命题的证明

引理 48（几何倾斜下的矩界）。设 f, g 为概率密度，$t \in [1/K, 1-1/K]$，且 $Z = \int f^t g^{1-t} > 0$。在密度为 $f^t g^{1-t}/Z$ 的分布 Q 下，在共同支撑集上令 $Y = \log(f/g)$。记 $b = -\log Z$，则有 $\mathbb{E}_Q Y^2 \leq C_K\,(b + b^2)$。

证明。令 $\beta = 1-t$，直接积分得到 $\mathbb{E}_Q e^{\beta Y} \leq e^b$ 和 $\mathbb{E}_Q e^{-tY} \leq e^b$。积分范围可以限制在共同支撑集上。函数 $te^{\beta y} + \beta e^{-ty}$ 在零点处等于 1，在该处导数为零，并且对所有 y，其二阶导数至少为 $\min(t\beta^2,\ \beta t^2) > 0$。取期望可得：若 $b \leq 1$，则 $\mathbb{E}_Q Y^2 = O_K(b)$。若 $b > 1$，相同的指数矩意味着 $\mathbb{P}_Q(|Y| > s) \leq 2e^{b-s/K}$。从 $s = K(b+\log 2)$ 开始对 Y^2 的尾部进行积分，并用该截断值的平方界定较低部分，可得 $O_K(b^2)$。 □

M.1. 跨割置换

揭示侧内对齐，并将每一侧合并为一个视图。原始置换可双射地表示为侧内相对置换与两个代表视图之间的跨侧置换。揭示后，后者仍然均匀分布，且独立于特征。利用揭示的对齐恢复任意一对跨侧对象，即可确定此置换。

通过重新标记，不妨设真实置换为恒等映射。记作 $A = Z_S^* Z_S$。两个对象上的真实密度与交换后密度平方根的亲和度为 $\mathrm{tr}(A^2) = e^{-\chi}$。事实上，在平方根被积函数中，长度为 k 的一个环会将每个左侧特征与两个连续的右侧特征配对。对左侧特征积分后，得到 k 个核 $A(\overline{y, y'}) = \int \sqrt{p(x,y)p(x,y')}\,dx$的循环积分，即 $\mathrm{tr}(A^k)$。A 的特征值均非负且总和为一，因此对于任意固定密度，χ 均为有限值。

令 f 和 g 分别为单坐标的真实密度和交换后的密度。在归一化中点倾斜下，对数比 $\log(g/f)$ 是对称的，因为交换是一个对合。引理 48将其 d坐标和 Λ 的方差界定为 $C_K d(\chi + \chi^2) = o((\log\ n)^2)$。选择$s_n = o(\log\ n)$ 使得借助切比雪夫不等式，倾斜概率 $[-s_n, s_n]$ 至少为 $3/4$。于是，倾斜概率 $[0, s_n]$ 至少为 $3/8$。变回真实分布后，指定交换与真实分配打平或胜过真实分配的概率至少为

$$a_n \geq \tfrac{3}{8} e^{-d\chi - s_n/2}.$$

在真实分配下，不相交对象上的交换是独立的。对于共享一个对象的两个交换，它们的联合事件的概率至多等于两个交换密度之间的亲和度：在该事件发生时，真实密度受其几何平均数的约束。相对

置换是一个三循环，其单坐标亲和度为 $\mathrm{tr}(A^3) \le (\mathrm{tr}(A^2))^{3/2}$。如果 N 计算与真实分配持平或胜过真实分配的交换数，那么

$$\mathbb{E}N = \binom{n}{2} a_n \to \infty, \qquad \frac{\operatorname{Var} N}{(\mathbb{E}N)^2} \le \frac{1}{\mathbb{E}N} + O\big(n^{-1} e^{d\chi/2 + s_n}\big) = o(1)$$

在固定裕度下。因此 $N \to \infty$以概率收敛。对最大后验概率最大化解进行均匀平局处理是贝叶斯最优的，并且在行重标记下保持等变。给定真实分配和数据，其成功概率至多为 $1/(1+N)$：如果真实分配最大化似然，则所有N 计数的交换也都是最大化元。该上界的期望趋于零。

M.2. 稀有亚群体

对于一个 q-块划分和任意乘积参考分布 V，令 $\alpha = (q-1)/q$、 $\beta = 1/q$，以及 $r = D_\alpha(P\|V)/(q-1)$。我们证明，当

$$dr \le (1-\epsilon)\log n, \qquad d(r + r^2) = o((\log n)^2). \tag{61}$$

给定 $Z = \int p^\alpha v^\beta$，令 Q 的密度为 $p^\alpha v^\beta / Z$。在共同支撑集上，对数比 $\log(dQ/dP)$ 和 $\log(dQ/dV)$ 都是 $\log(p/v)$ 的仿射函数，其斜率分别为 $-\beta$ 和 α，共同常数为 $-\log Z = \alpha r$。因此，对于 $D_0 = D(Q\|P)$ 和 $D_1 = D(Q\|V)$，

$$D_0 + \frac{D_1}{q-1} = r.$$

两个对数比的方差均为 $O_K(r + r^2)$ 根据引理48。

固定 $\eta = \epsilon/16$。令 W 表示这样一个行事件：该行的两个 d-坐标对数比不超过各自的均值 dD_0, dD_1 加上 η 对数 n。由（61）和切比雪夫不等式， $Q^{\otimes d}(W) = 1 - o(1)$。对每一行独立地以条件概率进行标记

$$\mathbf{1}_W \exp\left(\log \frac{dQ^{\otimes d}}{dP^{\otimes d}} - dD_0 - \eta \log n \right).$$

根据 W 中的第一个约束，这一概率至多为一。标记行的条件分布为 $Q^{\otimes d}(\,\cdot \mid W)$。标记计数 m 服从二项分布，其均值为 $ne^{-dD_0 - \eta}$对数$n Q^{\otimes d}(W) \ge n^{\epsilon - \eta}(1 - o(1))$，因为 $D_0 \le r$。因此，以概率 $1 - o(1)$，对数 m =对数 $n - dD_0 - \eta$对数 $n + O(1)$。

揭示锚点中的标记模式、块内对齐、每个代表视图中的标记位置以及所有未标记对应关系。同时，以未标记数据为条件。均匀置换可双射地划分为块内相对置换和代表置换。在给定所披露部分的条件下，标记位置之间剩余的 $q-1$ 双射是独立且均匀的。由于标记是在置换之前按行独立生成的，标记的潜在行仍然独立，服从 $Q^{\otimes d}(\,\cdot \mid W)$，并且与这些双射相互独立。

将这个条件实验与另一种实验进行比较：从$V^{\otimes d}$中抽取带标记的潜在行，并使用相同的均匀双射。在 V 下的块独立性使打乱后的

数据与其余双射独立，因此参考分布下的成功概率至多为$(m!)^{-(q-1)}$。带标记的潜在行和排列上的联合似然比由 $\left[e^{dD_1+\eta_{\text{对数}^n}}/Q^{\otimes d}(W)\right]^m$所界。 W中的第二个约束逐行给出了这个上界。利用$\log(m!) \ge m\log\, m - m$，真实条件成功概率的对数至多为

$$m\big[(q-1)(dr-\log n)+q\eta\log n+O_K(1)\big]$$

在高概率计数事件上。方括号内的量至多为负的常数乘以 $\log n$。例外计数事件对无条件成功概率的贡献为 $o(1)$。

M.3. 结合这两种障碍

若切分项达到 $L(P_n)$，则应用换位界。若某个划分项达到该值，则选择一个乘积参考分布，其速率 r 至多比下确界高 $\min\{1,\epsilon\log\, n/(2d)\}$。这保留了 (61) 中的固定裕度。对于固定的 P，活跃的切分信息或近最优参考速率有界，因为切割和划分的数量有限。它与 d 的乘积为 $O(\log n)$，这也给出了所需的二阶矩条件。对于变化的 P_n 和 $d\to\infty$，任一活跃量 a 均满足$d(a+a^2) = O(\log\, n + (\log n)^2/d) = o((\log\, n)^2)$。这些界适用于相应的指标，并证明定理27。

为完整起见，双块 R´enyi 速率已经至少达到相应的换位速率。在阶数为 $1/2$ 时，乘积参考分布上的亲和度上确界为 $\|Z_S\|_{\mathrm{op}}$：参考密度的平方根是非负的单位向量，而取绝对值不会减小其与 Z_S 的配对。因此 $R_{S|S^c} = -\log\, \lambda_{\max}(Z_S^* Z_S) \ge I_2^\circ(S)/2$。

附录 N. 弱依赖阈值的证明

在本节中 $h = h_n = o(1)$。所有 $O_K(h)$ 界都对切割、划分、候选项以及底层坐标分布一致成立。

N.1. 划分散度及其二次近似

令 $g = \log(dP/dU)$，并依据 U 下的正交 ANOVA 分解对 $g = \sum_{A\subseteq[K]} g_A$ 进行分解。具体而言，

$$g_A = \sum_{C\subseteq A} (-1)^{|A|-|C|}\mathbb{E}_U[g \mid X_C].$$

每个非常数连通分量在其任一变量上的积分均为零，并且其上确界范数为$O_K(h)$。置

$$c(\mathcal{P}) = \sum_{A:\ A \text{ meets at least two blocks of } \mathcal{P}} \|g_A\|_{2,U}^2.$$

写出 $V = V_{\mathcal{P}}^0$、 $t = \log(dP/dV)$ 和 $z = e^t - 1$。在 U 下进行边缘化得到 $dV/dU = 1 + O_K(h)$，因此 $\|t\|_\infty = O_K(h)$。由于 P 和 V具有相同的块边缘分布， $\mathbb{E}_V[z \mid$

$X_B] = 0$ 对每个块 B 都成立。由界 $|t - z| \le C_K h|t|$，再结合条件 Jensen 不等式，可得 $\|\mathbb{E}_V[t \mid X_B]\|_{2,\,U} \le C_K h\|t\|_{2,\,U}$。在 V 下，其余块的条件分布相对于它们在 U 下的分布，其密度为 $1 + O_K(h)$。因此，同一界也适用于$\mathbb{E}_U[t \mid X_B]$。

t 的跨块投影与 g的跨块投影完全相同，因为 $\log(dV/dU)$ 是块函数之和。块内投影和常数投影的平方范数至多为 $\sum_B \|\mathbb{E}_U[t \mid X_B]\|^2_{2,\,U}$。由此可得 $\|t\|^2_{2,\,U} = (1+O_K(h))c(\mathcal{P})$。在 V 下，$\mathbb{E}_V z = 0$。展开 $(1+z)\log(1+z)$ 和 $(1+z)^\alpha$，并利用$z^2 = (1 + O_K(h))t^2$，得到

$$D_{\mathcal{P}} = (1 + O_K(h))\frac{c(\mathcal{P})}{2}, \qquad D_\alpha(P\|V) = (1 + O_K(h))\alpha D_{\mathcal{P}} \tag{62}$$

对于此处用到的有限多个阶数 $\alpha = (q-1)/q$。泰勒余项有界于 $O_K(h)\mathbb{E}_V z^2$。即使某个划分的散度远小于 h^2，这些仍是相对估计。

N.2. 划分阈值处的逆命题

若 $dB(P) \le (1-\epsilon)\log n$，取一个含有 q 个块的最小化划分，并将其边缘乘积分布 $V^0_{\mathcal{P}}$ 用作附录 M.2 中的参考分布。根据 (62)，其中的参考速率满足

$$r = \frac{D_{(q-1)/q}(P\|V^0_{\mathcal{P}})}{q-1} = (1 + O_K(h))\frac{D_{\mathcal{P}}}{q}.$$

特别地，$dr \le (1-\epsilon/2)\log n$最终成立，且 $r = O_K(h^2)$，因此 $d(r + r^2) = O_K(\log n)$。针对特定参考分布的反向结论给出成功概率 $o(1)$，即使 d 有界时也成立。

N.3. 似然评分与联合集中

重新标记各行，使真实分配为恒等映射。在一个候选对齐中，每个源行 i 的视图会按其所属的候选元组划分。将这些划分的列表称为轮廓 H。若某行的划分至少包含两个块，则称该源行为拆分行。错误轮廓含有 $m \ge 2$ 个拆分行。若某个拆分行上的块数为 q_i，则令 $Q_H = \sum_i (q_i - 1)$，对所有拆分行求和。含有 m 个拆分行的轮廓数量至多为 $\binom{n}{m}\mathrm{Bell}(K)^m$。每个未拆分行都必须留在自己的元组中，因为其锚定视图是固定的。在由此得到的 m 个元组中，包含锚点的块有已知的去向，而每个其他块至多有 m 个去向。因此，每个轮廓至多对应 m^{Q_H} 个候选项。

在一个坐标上，候选项 τ与真实分配之间的似然差为 $F_\tau - T_H$。其中，T_H 是候选项遗漏的真实交互项 g_A 之和，而 F_τ 是新增候选项项之和。由于每个视图的行都经过置换，视图边缘似然相互抵消。遗漏的项恰好是跨越某个源行划分的项。对于每个 A,

遗漏项的数量等于新增项的数量，因为两种对齐都包含 n个该类型的项，而且共同项相互抵消。

在独立参考分布 $U^{\otimes n}$ 下，其余各项均已中心化且两两正交。它们的（源行、视图）元素集合彼此不同，因此，仅出现在两个项之一中的变量积分后为零。因此，两个边际得分方差均为$A_H = \sum_i c(\mathcal{P}_i)$，其中求和遍历拆分行，且它们的协方差为零。

在真实分布下，若两个局部项共享一个源行，则将它们相连。这构成一个度以上界由 K 控制的依赖图。每个源行最多为 K 个候选元组提供元素，而每个元组至多涉及 2^K 个交互。相对于独立性，涉及局部项或相邻项对的至多 $2K$ 个源行上的真实密度为 $1 + O_K(h)$。因此，一项的方差相对于其参考方差的变化至多为 $O_K(h)$ 倍，而相邻协方差至多为参考标准差乘积的 $O_K(h)$ 倍。按 $2ab \le a^2 + b^2$ 求和并结合度上界可得

$$\mathrm{Var}(xF_\tau + yT_H) = (1 + O_K(h))(x^2 + y^2)A_H.$$

对于 $|x|,\ |y| \le 2$，该线性组合中的中心化局部项的上确界范数为 $O_K(h)$，其方差之和为 $O_K((x^2 + y^2)A_H)$。因此，引理 49给出了如下双边估计

$$\log \mathbb{E}\exp\{x(F_\tau - \mathbb{E}F_\tau) + y(T_H - \mathbb{E}T_H)\} = \frac{(x^2 + y^2)A_H}{2}\big(1 + O_K(h)\big). \tag{63}$$

真实分布与候选分布相对于独立参考分布都有正密度。因此 $\mathbb{E}e^{F_\tau - T_H} = 1$。在（63）中取（$x, y$） $= (1, -1)$，得到 $\mathbb{E}(T_H - F_\tau) = (1 + O_K(h))A_H$，且对轮廓一致成立。对于 d 坐标得分，令 μ 为轮廓内均值差的最小值。令 T 为真实项的中心化和加上 μ，令 F 为候选项的中心化和。如果候选项与真实分配持平或胜过真实分配，则 $F \ge T$。这种选择使 T 对轮廓中的所有候选项都相同。由各坐标之间的独立性，

$$\log \mathbb{E}e^{xF + y(T-\mu)} \le \tfrac{1}{2}V(x^2 + y^2) \quad (|x|, |y| \le 2), \qquad V = d(1 + C_K h)A_H.$$

在超临界区间 $dB(P) \ge (1 + \epsilon)\log n$，（62）和 $\sum_i q_i = m + Q_H$可推出

$$D_H := \mu^2/V \ge 2(1 + \gamma)(m + Q_H)\log n, \qquad 0 < \mu/V \le 2 \tag{64}$$

对于某个固定的 $\gamma > 0$，以及所有充分大的 n。

N.4. 两层计数

令 $w = m \log n$，$r = Q_H/m \ge 1$，$\theta = \log m/\log n$，$\ell = (1 - \theta)w$，以及 $M = \theta r w$。轮廓的数量至多为 $e^{\ell + O_K(m)}$，且每个轮廓至多有 e^M 个候选项。选择 $0 < \delta \le \gamma/2$。

如果 $M \le \ell$，那么 $\theta \le 1/(1 + r)$，且

$$w \le \ell + M = [1 + \theta(r - 1)]w \le \tfrac{2r}{1+r}w \le \tfrac{1+r}{2}w.$$

对于单个候选项，使用 $x=-y=\mu/(2V)$ 在联合界中得到 $\mathbb{P}(F\ge T)\le e^{-D_H/4}$。根据（64），对其轮廓的并集至多为 $e^{-\ell-\delta w}$。

若 $M>\ell$，则恒等映射 $\ell+M/r=w$ 和柯西–施瓦茨不等式给出

$$(\sqrt{\ell+\delta w}+\sqrt{M+\delta w})^2\le(1+r)(1+\delta+\delta/r)w\le D_H/2.$$

令 $a=\sqrt{2(\ell+\delta w)/D_H}$ 以及 $s=1-a$。由于 $\ell<rw/(1+r)$，（64）给出 $a^2\le(1/4+\delta/2)/(1+\gamma)$，$\le 1/4$因此 $s\ge a$。公共事件 $T<s\mu$ 的概率至多为 $e^{-D_Ha^2/2}=e^{-\ell-\delta w}$，取 $x=0,\quad y=-a\mu/V$。在 $F\ge T\ge s\mu$ 上有

$$sF-a(T-\mu)\ge(s^2+a^2)\mu.$$

取 $x=s\mu/V,\quad y=-a\mu/V$ 将该事件界定为 $e^{-D_H(s^2+a^2)/2}\le e^{-\ell-M-2\delta w}$。最后一个不等式由 $s\sqrt{D_H/2}\ge\sqrt{M+\delta w}$得到 $\overline{s\sqrt{D_H/2}\ge\sqrt{M+}}\,\delta w$。所有参数均处于联合矩界（64）的范围内。对候选项界求和，并将公共下尾只计算一次，可得每个轮廓至多为 $2e^{-\ell-\delta w}$。

在这两种情形下，对所有轮廓及所有 $m\ge 2$ 求和可得

$$2\sum_{m=2}^{n}e^{O_K(m)-\delta m\log n}=o(1).$$

因此，以概率 $1-o(1)$，没有错误的候选项能与真实分配持平或胜过它，这就证明了最大似然上界。

N.5. 两种速率的比较论证

对于一个割 S，令 Y 表示在一个坐标上，对两个对象进行交换相对于真实分配的对数似然比。那么 $\|Y\|_\infty=O_K(h)$。在 $U^{\otimes 2}$下，损失项与新增项相互正交，各自贡献方差 $2c(S|S^c)$。因此，转换到真实分布后可得 $\mathbb{E}Y^2=(1+O_K(h))\,4c(S|S^c)$。由于 $\mathbb{E}e^Y=1$，泰勒展开给出 $\mathbb{E}Y=-\frac12\mathbb{E}Y^2+O_K(h)\mathbb{E}Y^2$ 以及 $\mathbb{E}e^{Y/2}=1-\frac18(1+O_K(h))\mathbb{E}Y^2$。取负对数并利用 (62) 得

$$I_2^\circ(S)=(1+O_K(h))D_{S|S^c}.$$

对于由 q 个块组成的划分以及 $\alpha=(q-1)/q$，我们还需要优化后的 R´enyi 速率的下界。相对于 U进行分析，并令 v_B 为相对于 U_B的块参考密度。固定其他块的密度后，在 v_B 中，取v_B 与以下内容成比例，可使亲和度最大化

$$\left(\int(dP/dU)^\alpha\prod_{C\neq B}v_C^{1-\alpha}\,dU_{-B}\right)^{1/\alpha}.$$

这是 H¨older 不等式中的等号情形。在乘方之前，括号中的表达式等于 $1+O_K(h)$ 乘以一个与 X_B 无关的正常数。因此，每个块更新一次，就会为每个块得到 $v_B=1+O_K(h)$，且亲和度不会降低。因此

只使用这种形式的均匀有界参考分布即可逼近最优值。对于这类 V，$\log(dP/dV)$ 的跨块投影就是 g 的跨块投影。因此，按上述方式展开 $D_\alpha(P\|V)$，得到下界 $(1-O_K(h))\alpha c(\mathcal{P})/2$。根据（62），边缘乘积参考给出了相应的上界。因此 $R_{\mathcal{P}}(P)=(1+O_K(h))D_{\mathcal{P}}/q$。结合割比较论证与划分比较论证，即可证明 $L(P)=(1+O_K(h))B(P)$。

附录 O. 依赖图矩界

引理 49。设中心化实变量 Y_i 具有最大度数为Δ的依赖图，并假设 $|Y_i|\le b$。依赖图的含义是：两个不相交且彼此之间没有边的顶点集所指标记的独立族相互独立。令 $W=\sum_i Y_i$ 和 $S=\sum_i \mathbb{E}Y_i^2$。存在常数 c_Δ、$C_\Delta>0$，使得对于满足 $|t|b\le c_\Delta$的实数 t，

$$\left|\log \mathbb{E}e^{tW}-\tfrac{1}{2}t^2\operatorname{Var}W\right|\le C_\Delta b|t|^3 S.$$

证明。情形 $b=0$ 是显然的。对于图顶点的子集 U 和复数 z，令

$$Z_U(z)=\mathbb{E}e^{z\sum_{i\in U}Y_i},\qquad w(A)=\mathbb{E}\prod_{i\in A}(e^{zY_i}-1).$$

展开 Z_U，并按照所选顶点集的连通分量对各项进行因式分解。对于$i\in U$，按包含 i 的连通分量分组可得递推式

$$Z_U=Z_{U\setminus\{i\}}+\sum_{\substack{i\in A\subseteq U\\ A\text{ connected}}} w(A)Z_{U\setminus N[A]},$$

其中 $N[A]$ 是 A的闭邻域。令 $M_0=\Delta+1$，则经过一个根节点的连通k-顶点集的数量至多为 $M_0^{2(k-1)}$。沿着一棵返回根节点的生成树遍历，可以在 $2(k-1)$ 步内记录这样的顶点集。

令 $R=(4\cdot 2^{M_0+1}M_0^2 b)^{-1}$。对于 $|z|\le R$，$|w(A)|\le(2|z|b)^{|A|}$。对 $|U|$ 使用归纳法可知，每个删除比 $Z_U/Z_{U\setminus\{i\}}$ 都在距 1 不超过 $1/2$ 的范围内。事实上，通过逐次删除一个顶点，小集合上的比率给出 $|Z_{U\setminus N[A]}/Z_{U\setminus\{i\}}|\le 2^{M_0|A|}$。递推关系和连通集计数随后将其与 1 的距离界定为

$$M_0^{-2}\sum_{k\ge 1}(M_0^2 2^{M_0+1}|z|b)^k\le 1/3.$$

这就证明了归纳法，并证明圆盘内无零点。

写出 $\sigma_i^2=\mathbb{E}Y_i^2$。中心化将单顶点界改进为 $|w(\{i\})|\le C|z|^2\sigma_i^2$。对于 $|A|\ge 2$，对其中两个因子应用柯西–施瓦茨不等式，并对其余因子取均匀界，得到

$$|w(A)|\le C|z|^2(2|z|b)^{|A|-2}\sum_{i\in A}\sigma_i^2.$$

沿固定的删除顺序，将删除比率的对数相加，并对 $|u| \le 1/2$使用 $|\log(1+u)| \le 2|u|$。每个连通集仅在其第一个顶点被删除时出现。对于给定大小k，所有连通 k-集上的 $\sum_{i\in A}\sigma_i^2$ 之和至多为 $M_0^{2(k-1)}S$。对所得几何级数求和，在圆盘内得到 $|\log Z_U(z)| \le C_\Delta |z|^2 S$，其中所选对数函数在零点取零值。对于完整的顶点集，解析函数 $z^{-2}\log Z_U(z)$在零点取值 $\operatorname{Var} W/2$。对 $|z| = R/2$ 应用柯西系数估计，可得其在 $|t| \le R/4$ 处与该常数之差的界为 $C_\Delta |t| S/R$。乘以 t^2 即证明了该引理。 $\square$

附录 P. 补充说明

P.1. 一个近线性高斯约束

我们证明备注 7中固定 $c > 0$的阈值。记 $a = K/(K-1+\delta)$。大小为 $r < K$的主子矩阵的特征值为 a，重数为 $r-1$，且 $a[1-(1-\delta)r/K]$。其对数行列式为 $O_K(1)$。完整协方差的行列式为 $a^K\delta$。因此，利用第2 节的行列式公式，对于每个非平凡划分，

$$D_{\mathcal{P}} = -\tfrac{1}{2}\log\delta + O_K(1) = c\log n + O_K(1).$$

对二分划分取最小值得到 $I_* = c\log n + O_K(1)$。按块数减一进行归一化后，单元素划分的首阶项最小，因此$J = c\log n/(K-1) + O_K(1)$。若 $c < \max\{2, K-1\}$，则定理 3中的逆命题适用，或者定理 1 排除同时近乎精确恢复。对于后一种情况，根据下面的似然界，还可得出更强的结论：精确成功的概率趋于零。

揭示具有 p 个块的最小化划分的块内对齐。剩余假设在 $N = (n!)^{p-1}$ 个跨块置换上均匀分布。在具有独立块和行的参考分布下，该假设与数据相互独立。相对于此参考分布，真实分布的对数似然比 L 等于 n 个在真实分布下的 $L_{\mathcal{P}}$ 独立副本之和。对于任意 $t > 0$，按照 $L \ge \log N - t$ 对成功事件进行划分，可将其概率界定为 $\mathbb{P}(L \ge \log N - t) + e^{-t}$。引理 30 对 $\operatorname{Var}(L) \le 2K^2 n D_{\mathcal{P}}$ 给出了界。如果 $J < (1-\epsilon)\log n$，则使用 $t = \epsilon(p-1)n\log n/3$ 和 $\log(n!) = n\log n - O(n)$。切比雪夫不等式也使第一个概率趋于零。

对于上侧，将真实对齐重新标记为恒等映射。对于竞争对齐 $\sigma\Pi$，在行索引上构造图：对于每个视图 a，在 i 和 $\widetilde{\pi}_a(i)$ 之间连一条边。似然差在各个连通分量之间具有可加性，因为每个置换都保持每个连通分量不变。一个似然至少与真实分配相同的竞争对齐，必须包含一个大小为 $m \ge 2$ 的移动分量，且该分量上的似然至少与真实分配在该分量上的似然相同。

对于这样的连通分量，令 U, V 分别表示其列向量包含真实约束方向和候选约束方向的矩阵。每列在一个元组中对应的观测值处有 K 个元素等于 $K^{-1/2}$，其余位置均为零。因此 $U^\top U = V^\top V = I_m$

以及 $M = U^\top V = K^{-1}\sum_{a=1}^{K} T_a$，其中 T_a 是该连通分量上的候选置换矩阵，且 $T_1 = I_m$。两个协方差分别为 $C_0 = a[I - (1-\delta)UU^\top]$ 和 $C_1 = a[I - (1-\delta)VV^\top]$，其行列式相等。令 $b_0 = (1-\delta)/2$，并令 $s_1 \ge \cdots \ge s_m$ 为 M 的奇异值。对 $[U\ V]$ 应用西尔维斯特行列式恒等式可得

$$\frac{\det((C_0 + C_1)/2)}{\det C_0} = \delta^{-m} \det \begin{pmatrix} (1-b_0) I_m & -b_0 M \\ -b_0 M^\top & (1-b_0) I_m \end{pmatrix} = \prod_{j=1}^{m} \left[1 + \frac{(1-\delta)^2}{4\delta} (1 - s_j^2) \right].$$

矩阵 $I_m - MM^\top$ 是一个边权重为 $(MM^\top)_{ij}$ 的图的拉普拉斯矩阵，其中 $i \neq j$。其图包含该连通分量的所有边，因为 M 具有正的对角元。每个正边权重至少为 $1/K^2$。因此，加权矩阵树定理给出 $\prod_{j=2}^{m}(1 - s_j^2) \ge mK^{-2(m-1)}$。高斯平方根似然比矩是上述行列式比的逆平方根。根据马尔可夫不等式，该竞争者等于或优于真实分配的概率至多为

$$m^{-1/2} \left(\frac{2K\sqrt{\delta}}{1-\delta} \right)^{m-1}.$$

至多有 $\binom{n}{m}(m!)^{K-1} \le n^m m^{(K-2)m}$ 种选择。对于每个固定的 $m \ge 2$，如果 $c > 2$，则该数量与概率的乘积趋于零，因为 $m - c(m-1) < 0$。如果 $c > K - 1$，取固定的 $\xi > 0$ 和 m_0，使得对于所有 $m \ge m_0$，都有 $c(m-1) \ge (K - 1 + \xi)m$。在计数中使用 $m \le n$，这些大小对应的总和至多为 $\sum_{m \ge m_0} (C_K n^{-\xi})^m = o(1)$。这就证明了上侧。例如，$K = 4, c = 5/2$ 给出 $I_*/\log n \to 5/2$ 和 $J/\log n \to 5/6$。

P.2. 视图数量的增加

有限样本界（2）已经对每个 K 成立。如果最大错误比例依概率趋于零，那么由有界性可知，其期望也趋于零，并且（2）的左侧也趋于零。对于精确恢复，揭示最小化二分中的侧内对齐后，该问题可化为双视图匹配问题。定理 3（b）的证明中的交换界仅取决于 n 和 I_*，因此当 K 增长时，它们同样适用。

划分检验的一致尾界。 对于一个具有 p 个块且散度有限的划分，在各块边缘分布的乘积分布的支撑集内进行白化。联合白化协方差矩阵的特征值为 $0 < \lambda_j \le p$。上界可通过对 p 个块上的线性形式的标准差应用柯西–施瓦茨不等式得到，这些块的边缘协方差矩阵均为单位矩阵。令 $x_j = \lambda_j - 1$，并令 $g \ge 1$ 界定所有 λ_j。引理 30 给出了在真实元组下的结果：

$$L_{\mathcal{P}} - D_{\mathcal{P}} = \tfrac{1}{2} \sum_j x_j (Z_j^2 - 1), \qquad \sum_j x_j^2 \le 4 g D_{\mathcal{P}},$$

其中包含相互独立的标准正态变量 Z_j。第二个上界使用了 $x-\log(1+x)\ge x^2/[2\max\{1,1+x\}]$，对于 $x>-1$。对于 $0<t\le 1/2$、$tx_j\ge -1/2$ 和 $v-\log(1+v)\le v^2$，对于 $v\ge -1/2$，因此 $-t(L_{\mathcal{P}}-D_{\mathcal{P}})$ 的对数矩至多为 $2gt^2D_{\mathcal{P}}$。利用 $t=\kappa/(4g)$，对于任意固定的 $0<\kappa<1$，我们得到

$$\mathbb{P}_P(L_{\mathcal{P}}<(1-\kappa)D_{\mathcal{P}})\le \exp[-\kappa^2 D_{\mathcal{P}}/(8g)].$$

假设 $J\ge(1+\epsilon)\log n$。选择 $\kappa>0$ 足够小，使得$(1-\kappa)D_{\mathcal{P}}-(p-1)\log n\ge c_\epsilon D_{\mathcal{P}}$ 对所有 $\mathcal{P}$ 均成立，其中 $c_\epsilon>0$为固定值。在阈值$(1-\kappa)D_{\mathcal{P}}$下，对于给定锚点，接受恒等划分 $\mathcal{P}$ 中的任一错误元组的概率至多为 $e^{-c_\epsilon D_{\mathcal{P}}}$，这是由乘积分布下的马尔可夫不等式及计数 n^{p-1}得到的。无限散度检验使用支撑集成员关系，如附录A所述。真实元组被拒绝和错误元组被接受的概率均为零。

对于 $g=p$，两个错误界均为 $n^{-\Omega_\epsilon(1)}$，因此当 $K\log K=o(\log n)$ 时，对至多 K^K 个划分取并集界即可。在特别情况下，真实元组未通过检验或某个错误元组通过检验的锚点数为 N_{bad} $o_{\mathbb{P}}(n)$。分别在每个视图中解决提议冲突并完成匹配。每个坏锚点在每个视图中至多会挤掉一个正确提议，因此 $\max_a M_a\le 2N_{\mathrm{bad}}$。

弱律的割权重计数。令 $R_{\mathcal{P}}$ 为 $\mathcal{P}$ 上 R 的块对角部分，并令 $c_R(\mathcal{P})=\sum_{a<b\text{ 遍及 }\mathcal{P}}\|R_{ab}\|_F^2$。$R_{\mathcal{P}}$ 的谱属于 $[1-h,1+h]$。矩阵

$$H_{\mathcal{P}}=R_{\mathcal{P}}^{-1/2}(R-R_{\mathcal{P}})R_{\mathcal{P}}^{-1/2}$$

的迹为零，$\|H_{\mathcal{P}}\|_{\mathrm{op}}\le 2h/(1-h)$，且 $\|H_{\mathcal{P}}\|_F^2=(1+O(h))\,2c_R(\mathcal{P})$。最后一个估计是通过将两个白化因子的奇异值界定在 $(1+h)^{-1/2}$ 和 $(1-h)^{-1/2}$ 之间得到的。展开 $-\log\det(I+H_{\mathcal{P}})/2$ 得到

$$D_{\mathcal{P}}=(1+O(h))\,c_R(\mathcal{P})/2$$

其绝对误差常数在所有划分及 K 上均一致。记 c_* 为最小二分权重。特别地，$c_*=(2+O(h))I_{*0}$。

对于 $u\ge 1$，当 $c_*>0$ 时，权重不超过 uc_* 的划分至多有 K^{9u} 个。为证明这一点，按照边权成比例的概率随机选取边并将其收缩，直到剩下 $t=\lceil 4u\rceil$ 个顶点。若 $t\ge K$，则由划分总数 K^K 可得该界。在有 $r>t$ 个顶点的一步中，总边权至少为 $rc_*/2$：每个收缩后的顶点都对应一个权重至少为 c_* 的原始割。假设某个权重不超过 uc_* 的指定划分至此仍然存活，则下一步将其破坏的条件概率至多为 $2u/r$。其存活概率至少为 $\exp(-4u\sum_{r=t+1}^{K}r^{-1})\ge K^{-4u}$。从最终顶点中均匀选取一个划分，与该指定划分一致的条件概率至少为 K^{-t}。不同原始划分所对应的输出互不相交，且 $4u+t\le 9u$。

如果 $c_*\gtrsim\log n$，则对于任意固定的 $c_0>0$，该计数给出：

$$\sum_{\mathcal{P}:\,|\mathcal{P}|\ge 2}e^{-c_0c_R(\mathcal{P})}\le\sum_{r\ge 1}K^{9(r+1)}e^{-c_0rc_*}=o(1)\tag{65}$$

当 $\log K = o(\log n)$，通过将比率 $c_R(\mathcal{P})/c_*$ 分组到区间 $[r, r+1)$中。对于近乎精确上侧， $I_* \geq J$ 给出 $c_* \gtrsim \log n$。弱协方差界允许$g = 2$ 对所有充分大的 n成立，并且真实错误界和虚假错误界均为 $e^{-\Omega_\epsilon(c_R(\mathcal{P}))}$。公式（65）取代了基于 K^K的并集界。

弱律下的精确恢复。 我们给出所需的均匀估计，以推广附录 D中的两级计数。对于一个移动源行 i的候选项，令 $\mathcal{P}_i$ 按接收对象 i观测值的候选元组划分视图；设 $m = |\{i : |\mathcal{P}_i| \geq 2\}| \geq 2$ 且 $Q = \sum_i (|\mathcal{P}_i| - 1)$。移动集合之外的行，其所有视图均与自身锚点配对，因此在每个视图中都保持不变。对于固定的移动集合和固定的划分模式，候选项在每个非锚定块中至多指定一个来自移动集合的目标。因此，此模式至多有 m^Q 个候选项。记 $A_H = \sum_i c_R(\mathcal{P}_i)$。

令 W 为 $-R^{-1}$ ，并将其对角视图块置零。其算子范数至多为 $2h/(1-h)$，因为从 $I - R^{-1}$ 中删除对角块，相当于减去一个范数至多为 $\|I - R^{-1}\|_{\mathrm{op}}$的分块对角压缩。被舍弃的真实二次评分 T 以及引入的虚假评分 F 的矩阵，是通过从真实评分矩阵和候选评分矩阵中删除相应的分块对角压缩得到的。因此，它们的算子范数均为 $O(h)$，且对 K一致。它们的支撑集不相交，因为 T 中的边连接同一真实对象内部，而 F 中的边连接不同对象。对于一个划分 $\mathcal{P}_i$，恒等式

$$R^{-1} = R_{\mathcal{P}_i}^{-1} - R_{\mathcal{P}_i}^{-1}(R - R_{\mathcal{P}_i})R^{-1}$$

这表明， W和 R 的跨块部分在弗罗贝尼乌斯范数下之差至多为 $O(h)\|R - R_{\mathcal{P}_i}\|_F$。事实上， $R_{\mathcal{P}_i}^{-1}$ 是块对角的，并且第二项中的两个逆矩阵因子在算子范数下距恒等映射至多为 $O(h)$。因此， T 的均值为 $\mu_H = (1 + O(h))A_H$； F 的均值为零。两个二次型矩阵的弗罗贝尼乌斯范数平方均等于$(1 + O(h))A_H/2$。对于 F，这是因为对于每一对视图，虚假边的数量等于被舍弃的真实边数量。

数据协方差的谱位于 $[1-h, 1+h]$中，且与 K无关。由高斯二次型矩公式和对数行列式展开可知，对于 $|x|, |y| \leq 2$，存在一个常数$V_H = (1 + O(h))A_H$，使得

$$\log \mathbb{E} \exp\{xF + y(T - \mu_H)\} \leq V_H(x^2 + y^2)/2.$$

具体而言，在系数矩阵的两侧分别乘以协方差平方根。其算子范数仍为 $O(h)$，并且由于支集不相交，其弗罗贝尼乌斯范数平方至多为$(1+O(h))A_H(x^2+y^2)/2$ 。中心化对数矩至多为该弗罗贝尼乌斯范数平方的 $1 + O(h)$倍。令 $D_H = \mu_H^2/V_H$。通过对各块的切割求和再除以二，一个 p-块划分至少切割 $pc_*/2$ 。因此， $I_* \geq (2 + \epsilon)$对数 n给出，对于某个固定的 $\epsilon_0 > 0$，

$$D_H = (1 + O(h))A_H \geq 2(1 + \epsilon_0)(m + Q)\log n.$$

为汇总模式界，令 $w = m \log n$, $q = Q/m$, $\alpha = \log m / \log n$, $L = (1 - \alpha)w$，并令 $M = \alpha q w$。每个

模式至多有 $F \geq T$ 个移动集合和 $\leq e^{-D_H/4}$ 个候选项。若 $M \leq L$，则 $\alpha \leq 1/(1+q)$、$q \geq 1$ 和 $L+M \leq (1+q)w/2 \leq D_H/[4(1+\epsilon_0)]$。联合矩界给出 $\mathbb{P}(F \geq T) \leq e^{-D_H/4}$。对候选项求和后，对于足够小的固定 $\delta_0 > 0$，至多为 $e^{-L-\delta_0 D_H}$。若 $M \geq L$，由柯西–施瓦茨不等式可得

$$\begin{aligned}\sqrt{L+\delta_0 D_H} + \sqrt{M+\delta_0 D_H} &\leq \sqrt{(1+q)w} + 2\sqrt{\delta_0 D_H} \\ &\leq \sqrt{D_H/2},\end{aligned}$$

对于仅依赖于 ϵ_0 的足够小的 δ_0，结论同样成立。令 $a_0 = \sqrt{2(L+\delta_0 D_H)/D_H}$ 和 $s_0 = 1 - a_0$。那么，对于足够小的 δ_0，有 $a_0 \leq 1/2$，因为 $L \leq qw/(1+q)$ 和 $q/(1+q)^2 \leq 1/4$。关于共享真实得分的一个界给出 $\mathbb{P}(T < s_0\mu_H) \leq e^{-L-\delta_0 D_H}$。在 $F \geq T \geq s_0\mu_H$ 上，$s_0 F - a_0(T-\mu_H) \geq (s_0^2+a_0^2)\mu_H$。参数为 $(s_0\mu_H/V_H, -a_0\mu_H/V_H)$ 的联合矩界随后给出至多为 $\exp[-D_H(s_0^2+a_0^2)/2] \leq e^{-L-M-2\delta_0 D_H}$ 的概率。对于较大的 n，所有使用的矩参数都属于 $[-2,2]$。对候选项求和可知，在任一情况下，模式失效的概率至多为 $2e^{-L-\delta_0 D_H}$。

最后 $D_H \geq c_1 \sum_i c_R(\mathcal{P}_i)$ 对于一个绝对的 $c_1 > 0$。因此，对移动集合和模式求和，模式界至多为 $\sum_{m\geq 2} 2(eZ_n)^m = o(1)$，其中 $Z_n = \sum_{\mathcal{P}:\ |\mathcal{P}|\geq 2} e^{-\delta_0 c_1 c_R(\mathcal{P})} = o(1)$ 由（65）得出。这证明了精确恢复的所述充分条件。

具有多项式数量视图的星形结构。 设 $K = \lceil n^\beta \rceil$ 和 $K-1$ 为相互独立的双视图弱高斯信道，其互信息为 $I = (2+\eta)\log n$，其中 $0 < \eta < \beta$。锚定视图包含拼接后的锚定端，而每个叶节点仅包含其信道的另一端。按特征顺序排列，协方差是这些信道的直和，因此$h = o(1)$，且最小割为 $I_* = I$。在独立先验下，叶节点排列的后验分布按信道分解。因此，最优的整体成功概率等于各个单独成功概率的乘积。

对于一个叶节点，记 p 为指定换位严格提高真实似然的概率。附录 D 中的交换界给出 $e^{-I-O(\sqrt{I})} \lesssim p \leq e^{-I}$，而共享一个指标的两个换位的交集概率至多为 $e^{-3I/2}$。不相交的换位彼此独立。容斥原理的前两项给出失败概率至少为 $\binom{n}{2}p - O(n^4p^2 + n^3e^{-3I/2})$。减去的两项均为 $o(n^2p)$，因此单个叶节点的失败概率至少为 $n^{-\eta-o(1)}$。乘积成功概率趋于零，因为 $\eta < \beta$。

P.3. 进一步的计算结果

异质分层筛查。对于视图集合 S 及其划分 $\mathcal{P}$，令 $C_S(\mathcal{P}) = \sum_{a<b} 在\ _{\mathcal{P}}\, w_{ab}$ 上，其中 $w_{ab} = \lim_n I(X^a; X^b)/\log n$。对于固定的 K，命题 6 表明，划分得分的真实均值为 $[C_S(\mathcal{P}) + o(1)]\log n$。取一个较小的 $\eta > 0$，并在 $\theta_S(\mathcal{P})\log n$ 的级别进行检验，其中 $\theta_S(\mathcal{P}) = (C_S(\mathcal{P}) - \eta)_+$。级别为零的检验跳过。根据引理 30，真实元组未通过某个启用的检验的概率为 $o(1)$，因为方差为 $O_K(\log n)$，且对于足够大的 n，均值与阈值之间至少相差 $(\eta/2)\log n$。恒等划分 $\mathcal{P}$ 中具有 p 个块的元组至多有 n^p 个。每个元组通过其自身的

划分检验的概率至多为 $n^{-\theta_S(\mathcal{P})}$，该概率是在相应乘积分布下计算的；检验跳过时，该界为 1。因此，该类型的错误元组通过所有检验的期望数量至多为 $n^{p-\theta_S(\mathcal{P})}$。在程序中使用已知的有限-$n$权重 $I(X^a;X^b)/\log n$ 时，同样的界仍然适用。

选择并排列一个种子 S_m，使得 $\min_{\mathcal{P}} C_{S_m}(\mathcal{P})/(|\mathcal{P}|-1) > 1$。将其前 s 个视图记为 S_s，其中 a_s 是最后一个视图。在步骤 s，选择使 w_{aa_s} 最大的源视图 $a \in S_{s-1}$。对每个源行，预先计算通过其与 a_s 的双视图检验的列表。当阈值指数为 $\theta = (w_{aa_s} - \eta)_+$ 时，列表期望长度至多为 $1 + n^{1-\theta}$。若列表超过 $\min\{n, \lceil 2n^{(1-\theta)_+ + \eta}\rceil\}$，则将其丢弃。根据马尔可夫不等式，丢弃指定真实对象的源列表的概率为 $o(1)$；若上限等于 n，则不会丢弃。用这些列表扩展保留前缀，并在 S_s 上检验每个扩展。若中间输出超过 $2n^{v_s+\eta}$，则中止，其中 $v_s = \max\{1, \max_{\mathcal{P}}(|\mathcal{P}| - \theta_{S_s}(\mathcal{P}))\}$，最大值取遍 S_s 的所有非平凡划分。保留的错误元组期望数量的上界和马尔可夫不等式表明，中止的概率为 $o(1)$。在最后一级，选择足够小的 η，使得对每个非平凡划分都有 $p - \theta_{S_m}(\mathcal{P}) < 1$。此时有 $n - o_{\mathbb{P}}(n)$ 个保留的真实元组和 $o_{\mathbb{P}}(n)$ 个保留的错误元组。进行极大不相交选取，随后补全，即可在种子上同时达到近乎精确：每个选中的错误元组至多能阻挡 m 个真实元组。

定义

$$u_1 = 1, \qquad u_s = \max\{1, \max_{\mathcal{P}}(|\mathcal{P}| - C_{S_s}(\mathcal{P}))\} \quad (s \ge 2),$$
$$\lambda_s = (1 - \max_{a \in S_{s-1}} w_{aa_s})_+.$$

在步骤 s，前缀上限和列表上限所贡献的扫描指数至多为 $u_{s-1} + \lambda_s + 4\eta$。因此，通过将 η 选得足够小，对于任意 $\xi > 0$，稠密列表和扫描的开销为 $O_K(n^{\Gamma+\xi})$，其中 $\Gamma = \max\{2, \max_{2 \le s \le m}(u_{s-1} + \lambda_s)\}$。如果每个不含松弛量的扫描指数都严格小于 2，则开销为 $O_K(n^2)$。划分似然分数可以由逐行得分和逐对得分组装而成。在 $d = \sum_a d_a$ 的条件下，按照定理 18，构造其权重并预计算得分的开销为 $O_K(d^3 + nd^2 + n^2 d)$。在具有独立坐标的共享坐标模型中，权重构造和逐行变换的开销则为 $O_K(nd)$。

当 $\sum_{a \in S} w_{ab} > 1$ 时，将新视图 b 添加到当前的对齐视图 S 中。在真实对齐下，将 S 与 b 连接后的双视图对数似然的均值为 $[\sum_{a \in S} w_{ab} + o(1)] \log n$。对于固定的 $\gamma > 0$（低于裕度），在 $(1+\gamma)\log n$ 处进行检验，会漏掉 $o_{\mathbb{P}}(n)$ 条真实边，并允许 $o_{\mathbb{P}}(n)$ 条虚假边。在实际连接中，有 $o_{\mathbb{P}}(n)$ 行受到污染。在极大不相交选取中，每个受污染的行至多只能被一条边使用。对于其余行，参考检验保持不变，因此选择和补全仍然只有 $o_{\mathbb{P}}(n)$ 个错误。迭代这一过程即可实现同时恢复。

备注 21中的示例里，三个种子二分权重分别为 1。53、1。50、1。47，而单元素划分的归一化权重为 $2.25/2 = 1.125$。按 1、2、3、 $u_2 = 2 - 0.78 = 1.22$、 $\lambda_2 = 0.22$以及 $\lambda_3 = 0.25$。两次扫描在松弛前的指数分别为 1。22 和 1。47。视图4相对于种子的权重为 $0.58 + 0.48 + 0.02 = 1.08$，因此在一次稠密双视图扫描中将其连接。六条单边的权重均小于1。每条双视图边

根据应用于双视图模型的定理 2，估计的重叠率会消失。为实现这些权重，为六对视图分别分配独立的坐标组，并在每个坐标组的两个视图中设置两个相关坐标。选择每个坐标组中的坐标数量及相关性，使互信息除以 $\log n$ 趋于指定权重，同时最大相关性趋于零。协方差在重新排列坐标后，是弱配对信道的直和。

部分对齐与精确线性规划。 在弱信号区间中，刚才使用的割的双重计数给出 $D_{\mathcal{P}} \ge (1+O(h))\, pI_*/2$，适用于 p-块划分。若 $I_* \ge (2+\epsilon)\log n$，则选取一个固定的 $\eta > 0$，使每个 $D_{\mathcal{P}} \ge (p+2\eta)\log n$。检验所有级别为 $(p+\eta)\log n$ 的元组。对于每种错误类型，接受任意元组的概率至多为 $n^{p} n^{-p-\eta}$。根据上述划分尾界，指定对象的真实漏检概率为 $n^{-\Omega_\epsilon(1)}$。对错误类型使用并集界，并对漏检的真实对象数量使用马尔可夫不等式，可得以高概率获得一个无错误的部分对齐，其中缺失对象数为 $O(n^{1-c_\epsilon})$，对于固定的$c_\epsilon > 0$。

对于更高信号的保证，在白化坐标中使用观测得分 $s_{ab}(i,\ j) = (Y_i^a)^\top R_{ab} Y_j^b$，并设定 $s_{ba}(j,\ i) = s_{ab}(i,\ j)$。对于每个元组 $v = (i_1,\ \ldots\ldots,\ i_K)$，令$s(v) = \sum_{a<b} s_{ab}(i_a,\ i_b)$。使用线性规划

$$\begin{aligned} &\text{maximize } \sum_{v} s(v) z_v \\ &\text{subject to } z_v \ge 0, \qquad \sum_{v:\, i_a = i} z_v = 1 \quad (a \in [K],\, i \in [n]). \end{aligned}$$

为了分析其最优解，重新标记各行，使真实分配成为恒等映射。对偶价格 $u_{a,i} = \frac{1}{2}\sum_{b \ne a} s_{ab}(i,i)$ 使每个真实元组的约束都取等号。对于恒等划分 $\mathcal{P}$ 中的一个错误元组，若其源行 j_B 属于块 B，则对偶松弛量为 $T - F$，其中

$$\begin{aligned} T &= \tfrac{1}{2} \sum_{B \in \mathcal{P}} \sum_{\substack{a \in B \\ b \notin B}} s_{ab}(j_B, j_B), \\ F &= \sum_{\substack{a<b \\ a,b \text{ across } \mathcal{P}}} s_{ab}(i_a, i_b). \end{aligned}$$

块内得分相互抵消，且 $\mathbb{E}T = A_{\mathcal{P}}, \mathbb{E}F = 0$，其中 $A_{\mathcal{P}} = c_R(\mathcal{P})$。作为元组中独立对象的二次型，$T$和 F具有互不相交的矩阵支集。对于每条跨块边，T中的系数减半后分别用于两个对象，而 F则在两个对象之间使用一次。相应系数矩阵的弗罗贝尼乌斯范数平方分别为 $A_{\mathcal{P}}/4$和 $A_{\mathcal{P}}/2$，算子范数为 $O_K(h)$。因此，根据上文所用的高斯二次型界，对于有界的 t，$t(F - T + A_{\mathcal{P}})$的对数矩至多为 $(1 + O_K(h))\, 3t^2 A_{\mathcal{P}}/4$。对于固定的 $0 < r < 1$，利用 $t = 2(1-r)/3$ 可得

$$\mathbb{P}(T - F \le r A_{\mathcal{P}}) \le \exp\big\{ -[(1-r)^2/3 - O_K(h)] A_{\mathcal{P}} \big\}.$$

由于 $A_{\mathcal{P}} \geq (1 - O(h))pI_*$，对于 $I_* \geq (3 + \epsilon)$，对数 n 可取足够小的 $r > 0$，使得该概率至多为 n^{-p-c}，其中 $c > 0$为某个固定值。对每种类型至多 n^p 个错误元组求和，可知以高概率，每个错误元组都具有严格对偶松弛，其量级至少为 $\log n$。因此，根据互补松弛，唯一最优解必为整数真实对齐。

该线性规划有 n^K 个变量和 Kn 个等式约束。对于位复杂度实现，可以将目标函数四舍五入为整数，并将绝对误差限制在有界范围内：将每个真实元组的舍入误差平均分配给其 K 个对偶变量，使任意松弛量的变化仅限于有界范围。上述严格裕度得以保持。高斯行范数尾界表明，对于白化数据，以高概率有 $\max_{a,i} \|Y_i^a\| = O_K(\sqrt{d} + \log n)$，因此舍入后目标函数的系数使用 $O_K(\log(n + d))$ 位。以有界绝对精度计算分数，并求解所得的有理数线性规划；对于固定的 K，该过程是多项式时间的。

低次下界。 考虑具有 d 个独立坐标、共同相关性为 $\rho \geq 0$ 的等相关模型。令 $\Gamma^*_{ai,cj}$ 表示视图 $a \neq c$ 中行 i, j 的同对象指示变量。植入的 D 次风险定义为：对这些元素的估计平方误差之和的期望，除以 $K(K-1)n^2$，其中每个元素均由总次数至多为 D 的多项式估计。常数多项式 $1/n$ 的风险为 $1/n - 1/n^2$。

将观测值除以 $\sqrt{1-\rho}$，这不会改变多项式度。在所得模型中，每个共享中心服从 $N(0, \lambda^2 I_d)$，$\lambda^2 = \rho/(1-\rho)$，每个观测值则加入独立的 $N(0, I_d)$ 噪声。这是 Even 等人（Even 等人(2026)）定义 1 中的模型，其参数(K, M, p) 对应于我们的参数(n, K, d)。固定一个锚点只会选定相同相对对齐的代表元，不会改变观测分布或同对象标签。他们的定理 2.2 将我们的风险上界限制为

$$R_D \geq \frac{1}{n} - \frac{1}{n^2}\left(1 + 64(D+1)D^{36}\frac{\zeta}{1-\sqrt{\zeta}}\right),$$
$$\zeta = 64D^{40}\frac{b\log n}{(1-\rho)^2}\max(1, K/n),$$

给定 $n \geq 2(D+2)^2$ 和 $\zeta < 1$。对于 $D = \lfloor(\log n)^{1+\eta}\rfloor$，$b = (4+\epsilon)/K$，以及 $(\log n)^{41+40\eta} \ll K = n^{o(1)}$，有 $\zeta = o(1)$ 和 $D^{37} = o(n)$。因此 $R_D = (1 - o(1))/n$，这是随机水平估计的一阶风险。根据视图和行重标记的对称性，对于任意指定的一对视图，使用所有视图作为输入时，同样的上界也成立。这是植入模型中多项式输出的平方误差保证。另一方面，$d \gg K \log n$ 意味着在此信号强度下有 $K\rho = o(1)$。在均匀割图中，一个 p-块划分至少切割 $K(p-1)/2$ 对，单元素划分时取等号。因此，均匀散度展开给出 $J = (1 + o(1))Kd\rho^2/4$。根据附录 P.2，在此情形下，近乎精确恢复在信息论上可行。

固定信号下的对数度估计量。假设 $b = d\rho^2/\log n \to b_0 > 0$、$\rho \to 0$ 和 $K \to \infty$。固定一个满足 $q \geq 2$ 的整数 $q > 16/b_0$。对于每个目标视图对，使用包含该视图对的 q 个视图组成的固定子集。对于一个候选元组，令 $L_{\mathcal{P}}$ 表示其在这些视图上的划分对数似然比。刚刚使用的最小归一化割界，其中

将 K 替换为 q 后，得到 $\mathbb{E}_P L_{\mathcal{P}} \geq (4-o(1))(|\mathcal{P}|-1)\log n$。记 $\ell_{\mathcal{P}} = \frac{5}{2}(|\mathcal{P}|-1)$ 和 $u_{\mathcal{P}} = 3(|\mathcal{P}|-1)$。我们用从 $\ell_{\mathcal{P}}$ 到 $u_{\mathcal{P}}$ 的过渡来近似应用于 $L_{\mathcal{P}}/\log n$ 的阈值。

在该元组的有限个恒等划分 $\mathcal{T}$ 上均匀成立，其中包括真实单块划分；对于每个固定的 $t>0$，存在一个常数 C_t 使得 $\mathbb{E}_{Q_{\mathcal{T}}} e^{t|L_{\mathcal{P}}|} \leq n^{C_t}$。事实上，在一个坐标上，由于协方差扰动的迹为零，$L_{\mathcal{P}}$ 含有一个阶为 $O_q(\rho^2)$ 的对数行列式项。其在 $Q_{\mathcal{T}}$ 下白化后的二次型具有算子范数 $O_q(\rho)$、平方弗罗贝尼乌斯范数 $O_q(\rho^2)$ 以及均值 $O_q(\rho^2)$。最后一个界可通过将两个逆矩阵展开至一阶，并利用其一阶差的迹为零得到。高斯矩公式给出，对每个坐标，任一符号下的对数矩均被 $O_{q,t}(\rho^2)$ 所界，且为 $d\rho^2 = O(\text{对数 } n)$。此外，在真实元组上，$L_{\mathcal{P}}/\text{对数 } n$ 的方差为 $O_q(1/\text{对数 } n)$。

固定 $M > \max_{\mathcal{P}} u_{\mathcal{P}}$，并在 $[-M,M]$ 上考虑如下函数：在 $\ell_{\mathcal{P}}$ 以下取 0，在 $[\ell_{\mathcal{P}}, u_{\mathcal{P}}]$ 上为线性函数，在 $u_{\mathcal{P}}$ 以上取 1。令 $x = M\cos\theta$，即可将其变为一个 Lipschitz 常数为 $O_q(M)$ 的偶周期函数。用归一化的非负杰克逊核 $(\sin(j\theta/2)/\sin(\theta/2))^4$ 对其进行卷积。根据在 $[-\pi,\pi]$ 上将核界为 $C\min\{j^4, |\theta|^{-4}\}$，并在 $[-1/j, 1/j]$ 上将其从下方界为 cj^4，可知归一化之前，其积分为 $\Theta(j^3)$，绝对一阶矩为 $O(j^2)$。取 $j = O_q(M)$，其中常数足够大，即可得到一个次数为 $2(j-1)$ 的偶三角多项式，并且其一致误差不超过 $1/8$。作为 x/M 的多项式，它在 $[-M,M]$ 上取值于 $[0,1]$。将此多项式与在 $k = O_q$ 个独立伯努利试验中以该成功参数取得多数的概率（对数 n）复合。二项分布尾界给出一个次数为 $O_q(M \text{ 对数 } n)$ 的多项式 $p_{\mathcal{P}}$；它在该区间上仍被 0 和 1 所界，在 $\ell_{\mathcal{P}}$ 以下至多为 n^{-3q}，在 $u_{\mathcal{P}}$ 以上至少为 $1-n^{-3q}$。

令 $W = \prod_{\mathcal{P}} p_{\mathcal{P}}(L_{\mathcal{P}}/\log n)$ 且令 $U = \max_{\mathcal{P}} |L_{\mathcal{P}}|/\log n$。在 $[-M,M]$ 上有界于 1 的 k 次多项式，通过在重新缩放的区间上展开为切比雪夫多项式，可由 $(8|x|/M)^k$ 界定，其中 $|x| \geq M$。因此在 $U > M$ 上，对于某个与 M 无关的 C，有 $W^2 \leq \exp(CU\log n)$，利用 $\log(8U/M) \leq C'U/M$ 以及上述度上界。将 $t = C+1$ 代入矩估计。必要时增大固定的 M，即可对恒等划分一致地得到

$$\mathbb{E}[W^2 \mathbf{1}_{\{U>M\}}] + \mathbb{P}(U>M) \leq n^{-6q}.$$

对于真实元组，其划分得分的集中性因此给出 $\|W-1\|_{L^2} = o(1)$。对于恒等划分 $\mathcal{P}$ 包含 p 个块的错误元组，根据 $Q_{\mathcal{P}}$ 下的马尔可夫不等式，其自身得分以至多 $n^{-5(p-1)/2}$ 的概率超过 $\ell_{\mathcal{P}}\log n$。区间其余部分的多项式抑制和尾界给出 $\|W\|_{L^2} = O(n^{-5(p-1)/4})$。

为估计目标视图中与固定锚定行相匹配的行的指示向量，对该锚点的所有候选元组求和，其中 We_j 是每个元组的目标行，j 是相应的标准基向量。真实项与真实向量的 o 范数相差 $L^2(1)$。每种错误类型至多有 n^{p-1} 个候选项，因此根据三角不等式，它们的合计范数为 $O(n^{-(p-1)/4})$。对有限种类型求和后，每个锚定行的均方误差为 $o(1)$。每个评分都是观测值的二次多项式。在 q, M 固定时，总次数为 $O_{b_0}(\log n)$。该界对目标视图对均匀成立，因此在此次数下可得 $R_D = o(1/n)$。

附录 Q. 数值细节

Q.1. 设置与分析

表2 列出了规模、特征维度和重复次数。G1、G2、G4 和 G5 使用等相关高斯坐标。G1、G2 和 G4 改变 $b = d\rho^2/\log n$，其中额外的 G2 单元格经过校准，以获得精确的 $J/\log n$。G5 直接校准 $I_*/\log n$。G3 使用聚类、分层和路径协方差。G6 使用三个视图，其中 AB的成对信息固定，同时改变 C 与二者的相关强度。每次重复中的所有方法使用相同的观测值和排列。G6 还在不同设计间共享 AB 观测值。

表 2：完整运行网格。解码器在每次重复试验中共享同一独立抽样。在 G3 中，$n = 8$仅出现在混合分支设计中。

组	K	n	d	每单元重复次数
G1	3	256, 512, 1024, 2048	16384	16
G2	3, 4, 6, 8	1024, 4096, 8192	4096	8
G3	4, 6, 8	8, 256, 1024, 4096, 8192	4096	4
G4	3	256, 1024, 4096, 8192	4096	4
G5	3	256, 512, 1024, 2048, 4096	2048	8
G6	3	256, 1024, 4096, 8192	4096	16

完整元组重叠度是配对对象全部正确的锚点行所占的比例。配对重叠率只统计指定配对中的配对对象。我们先在每次重复试验内计算各项统计量，再按固定的规模、设计和信号对重复试验取平均值。计算增益时，先在每次重复试验内减去基线，再取平均值。我们以重复试验为单位计算标准误。按照记录的规则，若方向性差异的绝对值超过其合并标准误的两倍，则认为该差异已得到判定。如果补充点预测中，超过 5% 的单元格落在记录容差范围之外，或任一单元格的偏差超过 0.10，则该预测不合格。区间预测逐个单元格进行检查。对于 G2，我们将原始运行中的前 616 条记录与两个各含 28 条记录的分片合并。原始运行中的最后 21 条记录与分片记录重复，因此予以省略。

图 3 汇总了与定理 3 和 2 以及注记 20直接相关的三项比较论证；下方表格给出了各单元格的数值。

Q.2. 精确恢复：失败证书与获胜交换

G1 对植入对齐进行评分，并寻找得分严格更高的可行分配。这样的分配可以证明最大似然失败。相反的证书要求一个能够在植入得分处闭合的有效上界。表格3分别统计失败实例和未解决实例。随着实际的 I_*/对数 n 增加，失败计数随之下降。对偶证书在此网格上始终未能闭合，因此，认证失败数为零的单元格并不能为最大似然恢复的概率提供正下界。

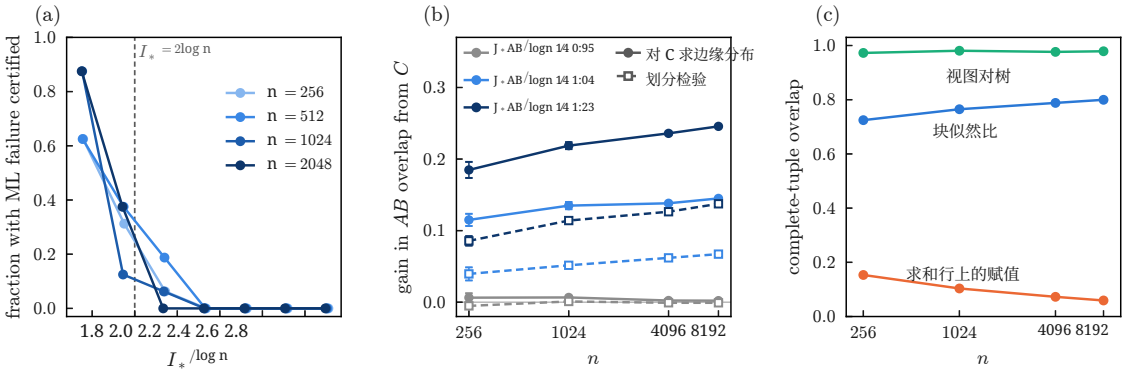

图3：有限样本n 主要阈值检验；点表示重复实验的均值。(a) G1，$K = 3$，$d = 16384$：每个单元格的 16 次重复实验中，找到严格优于植入分配的可行分配所占的比例，随实际实现的I_*/对数 n变化。这种分配可以证明最大似然法会失败。相反的证书始终未能得出，因此数值为零并不意味着成功。(b) G6，$K = 3$，$d = 4096$，I_{AB}/对数 $n = 0.95$：读取视图 A–B 时重叠度的配对变化，误差线为一个标准误。具有 $J^*_{AB} >$ 对数 n 的两种设计增益更大（定理2）。(c) G3 路径设计，$K = 4$，dr^2/对数 $n = 3$，四次重复实验：块似然比使信息通道保持跨越根割，而求和后的行会稀释该通道（注20）。原始视图对树在这种设计下也表现良好。

表3：G1。在 $K = 3$ 和 $d = 16384$处的等相关失败证书。每种规模和信号单元均有16次重复实验，显示存在优于真实分配的可行分配的实例数。$I_*/\log n$ 列给出了不同规模下的范围。在448个实例中，有0个实例将真实分配认证为最大似然分配。

b	$I_*/\log n$	256	512	1024	2048	失败	未解决
1.8	1.751–1.758	10	10	14	14	48	16
2.0	1.943–1.951	5	6	2	6	19	45
2.2	2.134–2.143	1	3	1	0	5	59
2.4	2.325–2.335	0	0	0	0	0	64
2.6	2.516–2.527	0	0	0	0	0	64
2.8	2.706–2.719	0	0	0	0	0	64
3.0	2.896–2.910	0	0	0	0	0	64

G5 统计每个单元素切分中的严格提升似然的双行交换次数，与附录 D 中逆命题所述相同。表4 报告观测到的次数，以及根据高斯交换概率计算出的精确期望，二者采用相同的切分与重复实验聚合方式。在这个各向同性切分模型中，典型相关为 $r = \sqrt{1 - \exp(-2I_*/d)}$。将行差对旋转为独立的和与差，可得期望次数 $\binom{n}{2}\mathbb{P}(B < (1-r)/2)$，其中 $B \sim$ 贝塔分布$(d/2, d/2)$。G1 还可以找到移动两行以上的更优匹配方案。

表 4：G5。$K = 3$和 $d = 2048$的有利换位。各列以切分信息 $I(X^S;\, X^{S^c})/\log\, n$为指标。单元格给出观测次数的均值和精确期望的均值，两者以斜线分隔，并对 8 次独立重复实验中的 3 个单元素切分取平均。最后一行汇总了不同规模、切分和重复实验下的次数及其期望。

n	1.0	1.5	1.8	2.0	2.2
256	14.6/ 14.2	0.917 / 0.741	0.0833 /0.129	0 / 0.0406	0 / 0.0128
512	28.2 / 27	1 / 0.995	0.125 / 0.141	0 / 0.0385	0 / 0.0106
1024	58.1 / 51.6	1.5 / 1.34	0.125 / 0.154	0.0417 / 0.0367	0.0417 / 0.00878
2048	89.3 / 98.9	1.33 / 1.82	0.0833 / 0.17	0 / 0.0352	0 / 0.00732
4096	183 / 190	2.12 / 2.47	0.167 / 0.187	0.0417 / 0.0338	0.0417 / 0.00612
总计	8960/ 9169.1	165 / 176.8	14 / 18.8	2 / 4.4	2 / 1.1

Q.3. 成对可匹配性：第三个视图带来的增益

在 G6 中，直接双视图分配提供了 AB 基线。我们将其与两种读取 C的方法进行比较：在 AB 分配之前，对可能的第三行进行局部边缘化；以及三视图分区检验解码器。对于每个划分 $\mathcal{P}$，每个候选项的对数似然比都必须超过 $(|\mathcal{P}| - 1 + 0.05)\log n$。如果某个元组通过检验，每个锚点都会提出其中单例划分对数似然比最大的通过检验的元组。解码器按照该得分从高到低，采用贪心方式保留不相交三元组，在剩余行上运行最大权重 AB分配，并按照索引顺序将剩余的 C 行与未匹配的锚点进行匹配。表5报告了 AB 重叠度和信息级别 $J^*_{AB}/\log\, n$的成对变化。在最弱的 C 下，两项增益都接近于零。在 $n \geq 1024$下，边缘化增益范围为 0.002 至 0.007，并且每个单元格中的增益都超过两个成对标准误。两个更强的设计使J^*_{AB}高于 $\log\, n$，并在该网格上显示出更大的增益，说明了有限 n下的定理 2。

Q.4. 递归解码与超视图得分的选择

G3 比较了似然比递归、仅使用原始视图对的树结构，以及求和行分配。原始二叉层次结构与更紧凑的二叉层次结构在子节点信息传递的范围上有所不同——

表5：G6. 在有视图 C 和没有视图 C 的情况下，将 A 匹配到 B，使用 $K = 3$、$d = 4096$ 和 $I_{AB}/$ 的 log $n = 0.95$。每次重复实验在不同 C 强度下使用相同的 AB 数据。相对于直接双视图分配，增益为配对的 AB 重叠度差；其中包括对 C 进行局部边缘化以及三视图分区解码。元素给出 16 次重复实验的均值，每项增益附带一个标准误。

n	$J/\log n$	$J^*_{AB}/\log n$	直接 AB	边缘化增益	划分增益
$I_{AC}/$对数 $n = I_{BC}/$对数 $n = 0.2$，最弱划分 $AB \mid C$					
256	0.381	0.950	0.722	$+0.0063 \pm 0.0062$	-0.0051 ± 0.0029
1024	0.379	0.950	0.698	$+0.0067 \pm 0.0023$	$+0.0010 \pm 0.0012$
4096	0.377	0.950	0.678	$+0.0024 \pm 0.0007$	-0.0008 ± 0.0004
8192	0.376	0.950	0.668	$+0.0022 \pm 0.0007$	-0.0008 ± 0.0004
$I_{AC}/$对数 $n = I_{BC}/$对数 $n = 0.6$，最弱划分 $A \mid B \mid C$					
256	1.046	1.046	0.722	$+0.1150 \pm 0.0085$	$+0.0396 \pm 0.0092$
1024	1.043	1.043	0.698	$+0.1349 \pm 0.0050$	$+0.0516 \pm 0.0039$
4096	1.041	1.041	0.678	$+0.1382 \pm 0.0019$	$+0.0620 \pm 0.0011$
8192	1.039	1.039	0.668	$+0.1450 \pm 0.0015$	$+0.0672 \pm 0.0013$
$I_{AC}/$对数 $n = I_{BC}/$对数 $n = 0.8$，最弱划分 $A \mid B \mid C$					
256	1.237	1.237	0.722	$+0.1848 \pm 0.0112$	$+0.0857 \pm 0.0067$
1024	1.233	1.233	0.698	$+0.2188 \pm 0.0048$	$+0.1141 \pm 0.0041$
4096	1.229	1.229	0.678	$+0.2359 \pm 0.0024$	$+0.1264 \pm 0.0021$
8192	1.228	1.228	0.668	$+0.2457 \pm 0.0016$	$+0.1375 \pm 0.0014$

表 6：递归划分检验、原始视图树以及基于块和的最大权重指派的平均完整元组重叠度。对于每种层级和混合分支设计，我们展示测试过的最大 n 对应的所有信号；对于路径，则展示所有测试过的规模。前三种设计的信号为 $I_{\mathrm{root}}/\log n$，路径的信号为 $dr^2/\log n$。破折号表示未测量的方法。

设计	K	n	信号	递归	Tree	求和指派
原始层级	6	8192	0.75	0.204	0.010	0.463
	6	8192	1	0.433	0.022	0.717
	6	8192	1.25	0.626	0.040	0.890
	6	8192	1.5	0.765	0.063	0.976
	8	8192	0.75	0.205	0.002	0.462
	8	8192	1	0.433	0.007	0.720
	8	8192	1.25	0.625	0.017	0.893
	8	8192	1.5	0.762	0.028	0.980
紧密层级	6	8192	0.75	0.019	0.000	0.167
	6	8192	1	0.141	0.003	0.585
	6	8192	1.25	0.365	0.010	0.878
	6	8192	1.5	0.595	0.025	0.977
	8	8192	0.75	0.008	0.000	0.126
	8	8192	1	0.099	0.000	0.570
	8	8192	1.25	0.314	0.000	0.876
	8	8192	1.5	0.557	0.002	0.977
混合分支	6	1024	1.125	0.216	0.152	—
	6	1024	1.2	0.278	0.210	—
Path	4	256	3	0.725	0.973	0.153
	4	1024	3	0.765	0.981	0.104
	4	4096	3	0.788	0.977	0.073
	4	8192	3	0.800	0.979	0.059

信息量超过父节点信息量。混合分支设计同时采用二叉节点和三叉节点，其预言机根节点得分衡量在子节点对齐正确时的根节点匹配。在路径设计中，根块似然利用割两侧的信息通道，而对块中的所有坐标求和会稀释该信息通道。拟合得到的层级结构和路径测量结果见表 6。路径对比涉及两种超视图统计量的选择。原始视图对树在该设计中表现良好。

Q.5. 有限规模下的分层筛选

G2 比较了附录G 中的筛选解码器（B3）、其采用调整后阈值和截断的变体 B3c、Wee 等人（2026年）的逐行两两求和，以及其他多项式时间解码器。表格 7和图 4 报告完整元组重叠度。逐行两两求和产生的原始提案可能重复使用行，因此与其补全后的输出分开报告。列表和前缀上限限制了中等规模下的覆盖率 n。特别是，在渐近标度中固定的裕量可能相较于有限样本分数波动而言很小。Sinkhorn 在此处是一个经验比较方法，在这种设置下不提供恢复保证。

G2 使用 $d = 4,096$，每种设置进行 8 次配对重复试验。我们的结果是整个元组均正确的锚定行所占比例。B3 使用分层筛选的中点阈值，而 B3c 使用有限样本阈值；当中间层超过上限时，保留得分最高的前缀。对于 $K > 4$，所展示的重叠度包括附加到四视图种子上的剩余视图。我们在整个网格中评估了 Pair sparse、消息传递、Sinkhorn，以及初始和细化后的谱方法输出。

表格7 展示了中等规模下的信号扫描。B3c 的均值最终重叠度在 84 种设置中均高于 B3。在 $K = 4$，$b = 2$ 它从 0.712 上升至 0.760，而 n 则从$1,024$ 增至$8,192$。在 $b = 2$，真实种子元组的所有元素均包含在保留的 B3c 列表中的平均比例，对于 $K = 3$，4为 0.688 至 0.805，因此即使在最终测试之前，列表筛选也会限制重叠度。在经过校准的 J/对数 n 网格上，B3c 的最高均值为 0.494。消息传递和 Sinkhorn 分别达到 0.744 和 0.774。

图 4 比较了名义逐行边界下的规模。在 $K = 8$，$b = 8/7$，Sinkhorn 的均值按规模递增顺序分别为 0.992、0.972、0.950，而 B3c 分别为 0.234、0.254、0.258。在较弱的名义 $b = 0.821$下，Sinkhorn 从 0.617 降至 0.376。这些是对实现的有限尺寸描述。

在 $K = 3$下，Wee 逐行规则在 $n = 1,024$、$4,096$这两个网格上进行了评估。其原始完整元组重叠率比完成后的重叠率高 0.058–0.117（跨不同单元格均值）。原始规则为每个锚定行独立选择配对对象。完成规则则按索引顺序解决冲突并填补未匹配行。第 Q.7 节讨论了对原始提议的预测。

Q.6. 真实行排名

在 G4 中，每个候选项的对数似然比必须超过 $(|\mathcal{P}| - 1 + 0.05)\log n$，对每个划分而言 $\mathcal{P}$。如果有通过检验的元组，每个锚点都会提议其中两两求和最大的元组。在每个非锚点视图中，按索引顺序最先提出提议者保留发生争用的行，其余锚点和未使用的行则按索引顺序匹配。在 $K = 3$ 的弱等相关极限下，这一裕度使未经剪枝的划分检验解码器的阈值为 $b = 41/30$ (附录 A)，高于测试值 $b = 4/3$。

在三视图秩剪枝实验中，我们在应用划分检验之前，根据每个非锚点视图中的配对得分保留排名靠前的候选项。列表宽度为 $\lceil n^{1-b/4} \rceil$，以及该宽度的一半向上取整。在采样单元中，真实元组所占的均值比例

表 7： $n = 4$，096 时 G2 的完整元组重叠。每个单元格是 8 次重复实验重叠的均值。名义 b 和校准后的 J/log n 表示不同的网格。破折号表示未运行的方法。PS/SI 分别是 Pair sparse 和谱初始化，在这些行中的重叠度相同。MP 是消息传递，SR 是谱细化。Wee 原始允许行重用。其余所有列均报告在每个视图中强制施加置换后的输出。

K	信号	PS/SI	MP	Sinkhorn	SR	B3	B3c	Wee	
								原始	完成
3	$b = 1.333$	0.160	0.512	0.509	0.349	0.160	0.334	0.491	0.374
	$b = 1.417$	0.197	0.582	0.589	0.407	0.195	0.384	0.545	0.435
	$b = 1.500$	0.236	0.652	0.667	0.463	0.215	0.433	0.589	0.486
	$b = 1.750$	0.373	0.812	0.857	0.628	0.273	0.579	0.709	0.630
	$b = 2.000$	0.528	0.905	0.954	0.764	0.331	0.700	0.800	0.741
	J/log $n = 1.050$	0.209	0.607	0.617	0.426	0.202	0.402	0.561	0.453
	J/日志 $n = 1.150$	0.279	0.713	0.744	0.525	0.234	0.486	0.633	0.539
4	$b = 1.000$	0.015	0.296	0.304	0.132	0.041	0.187	–	–
	$b = 1.167$	0.034	0.457	0.525	0.242	0.081	0.299	–	–
	$b = 1.333$	0.067	0.610	0.739	0.387	0.114	0.414	–	–
	$b = 1.667$	0.195	0.811	0.969	0.680	0.191	0.612	–	–
	$b = 2.000$	0.387	0.904	0.998	0.871	0.269	0.748	–	–
	J/对数 $n = 1.050$	0.027	0.404	0.452	0.204	0.072	0.261	–	–
	J/对数 $n = 1.150$	0.043	0.509	0.598	0.288	0.092	0.337	–	–
6	$b = 0.667$	0.000	0.042	0.068	0.010	0.001	0.024	–	–
	$b = 0.933$	0.001	0.210	0.451	0.090	0.014	0.131	–	–
	$b = 1.200$	0.006	0.428	0.894	0.315	0.059	0.304	–	–
	$b = 1.600$	0.054	0.700	0.998	0.752	0.135	0.548	–	–
	$b = 2.000$	0.215	0.848	1.000	0.944	0.229	0.709	–	–
	J/log $n = 1.050$	0.000	0.093	0.170	0.027	0.003	0.055	–	–
	J/log $n = 1.150$	0.000	0.144	0.288	0.051	0.007	0.085	–	–
8	$b = 0.500$	0.000	0.003	0.011	0.001	0.000	0.001	–	–
	$b = 0.821$	0.000	0.078	0.438	0.040	0.003	0.059	–	–
	$b = 1.143$	0.001	0.278	0.972	0.300	0.031	0.254	–	–
	$b = 1.571$	0.015	0.598	1.000	0.817	0.099	0.516	–	–
	$b = 2.000$	0.126	0.798	1.000	0.975	0.194	0.679	–	–
	J/对数 $n = 1.050$	0.000	0.011	0.050	0.003	0.000	0.007	–	–
	J/日志 $n = 1.150$	0.000	0.022	0.110	0.007	0.000	0.014	–	–

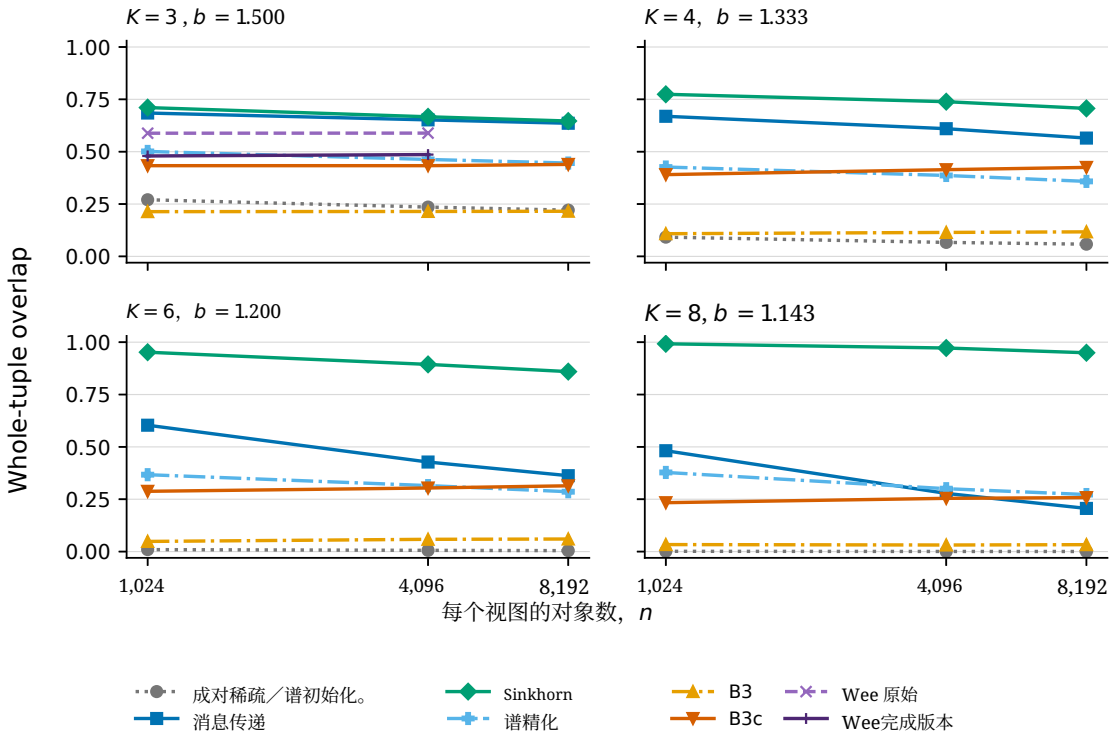

图 4：名义行边界处不同规模下的 G2 完整元组重叠度 $b = K/(K-1)$。各点为 8 次重复实验的均值。两条 Wee 串联仅在进行评估的位置显示。在这些行中，Pair sparse 与谱初始化重合。

两个列表共同保留的比例，在名义宽度下为 92.3%–95.7%，在半宽度下为 82.7%–92.9%。在 $n = 8192$ 和所测试的最弱信号下，两个列表分别生成了全部候选元组的 0.247% 和 0.062%。该实验还在 $n = 256$ 测量了未剪枝枚举。在那里，相对于枚举，最大的均值完整元组重叠损失在名义宽度下为 0.008，在半宽度下为 0.030。

Q.7. 在运行前记录的预测

原始预测文件和补充预测文件均在相应运行之前记录。两个预测文件的 SHA-256 摘要列于辅助文件的 `README` 中。按照记录的规则，33 项预测中有 32 项成立。唯一的例外是关于逐行求和的有限尺寸预测。记录的模拟遗漏了$n - 1$ 个将一个锚点与同一错误对象的另外两个视图配对的候选元组。这两行的平均内积为 $d\rho$，因此，这些冒名元组比两个行分别来自不同错误对象的候选项更具竞争力。在 $n = 1024$ 和 4096 下进行的诊断模拟保留了这些元组。模拟值减去观测值的原始逐锚点准确率范围为 -0.028 至 $+0.025$。冒名元组在模拟锚点的 16%–37% 中胜出。原始预测按照其记录的规则仍然失败。这就是定理 15（ii）中的冒名机制。