机器学习研究杂志 27（2026年）1-24 投稿日期 1/21；修订日期 2/26；发表日期 4/26

有限混合的超越正则化：理论保证与实际局限性

Ernest Fokou´e EPFEQA@RIT.EDU
纽约州罗切斯特理工学院数学与统计系，
纽约州罗切斯特 *14623-4322*，美国。

arXiv:2602.03889v1 [stat.ML] 3 Feb 2026

编辑：我的编辑器

摘要

有限混合模型是无监督学习的基础工具，但通过期望最大化（EM）算法进行最大似然估计时，会因分量坍塌或合并而遭遇众所周知的退化现象。我们提出超越正则化，这是一种惩罚似然框架，通过嵌入解析的强制性屏障函数来防止退化，同时借助趋零惩罚方案 $\lambda_n \downarrow 0$维持渐近效率。由此得到的分布混合超越算法（TAMD）具有强有力的理论保证：在较小惩罚项下具备总体可识别性，在模型设定正确的情形下具备一致性和渐近正态性，算法具有单调收敛性，并且具备模型错设下的稳健性。**然而**，我们全面的模拟研究揭示了复杂的实证表现。尽管 TAMD 成功防止了分量坍塌，并在存在污染时保持稳定性，但与 EM 相比，其在无监督分类准确率方面的提升仅较为有限，尤其是在高维低分离情形下。当真实类别分离度相对于维度较小时，两种方法的表现往往都仅略高于随机猜测。这一结果突显出一个根本性局限：即使采用复杂的正则化，为密度估计而优化的混合分量也可能无法与语义上有意义的类别相一致。因此，我们的工作作出双重贡献：（1）提出一种具有强有力保证的正则化混合估计新理论框架；（2）诚实地评估在具有挑战性的情形下无监督混合建模的实际局限（图 4 和图 2）。TAMD 方法已在开源 R 包中实现，为研究人员提供了稳定化混合估计的工具，同时也促使人们对其性能形成切合实际的预期。

关键词：有限混合模型、超越正则化、惩罚似然、防止退化、高维估计、经验局限性、无监督学习、算法收敛

1 引言

有限混合模型是密度估计和聚类的基础工具，其应用涵盖遗传学、金融学和机器学习。自皮尔逊（1894年）为高斯混合引入矩估计法以来，通过期望最大化（EM）算法（Dempster等人，1977年）进行最大似然估计已成为标准方法。尽管其使用广泛，期望最大化仍存在文献中充分记载的病态问题：似然上方无界，从而导致退化解，其中分量

©2026年 Ernest Fokou´e。许可协议：CC-BY 4.0，参见 https://creativecommons.org/licenses/by/4.0/。署名要求见 http://jmlr.org/papers/v27/21-0000.html。

方差坍塌至零（Day, 1969），并且该算法对初始化敏感，常常收敛到局部极大值或奇点（Redner and Walker, 1984）。

Kiefer and Wolfowitz (1956) 的经典研究确立了混合分布的基本可辨识性条件，并指出当分量未能充分分离时，最大似然估计会面临挑战。

人们提出了多种解决方法。惩罚似然方法（Chen and Khalili, 2008）通过添加正则化来缓解问题，但往往缺乏连贯的渐近理论。采用排斥先验的贝叶斯方法（Petralia et al., 2012）可以防止分量聚并，但需要计算量巨大的马尔可夫链蒙特卡洛。谱与矩估计量（Anandkumar et al., 2014）在满足分离条件时可提供多项式时间保证，但会牺牲统计效率。这些方法都无法同时实现有限样本稳定性、渐近效率和易于计算的更新。

本文提出了分布混合的超越算法（TAMD），这是一种新颖的惩罚似然框架，能够解决这些长期存在的难题。其核心思想是在对数似然中加入解析障碍项，当分量变得难以区分时，这些项会趋于无穷大，从而强制分量保持分离并防止坍塌。与临时设置的约束不同，超越障碍是强制的、具有置换不变性，并以受控速率渐近趋于零 $\lambda_n \downarrow 0$，因此 TAMD 在良性情形下能够继承最大似然的效率，同时在病态情形下提供稳健性。

我们的贡献有四个方面：

(i) 我们提出TAMD，并确立了总体可识别性、一致性、渐近正态性以及模型错设下的稳健性。(ii) 借助解析目标函数的Kurdyka–Lojasiewicz性质，我们证明了TAMD更新的单调上升性与收敛性。(iii) 我们将TAMD扩展到具有增长 K_n 的筛估计，并推导出有限样本Hellinger距离和积分概率度量风险界。(iv) 我们将TAMD应用于高斯混合，推导出梯度修正的闭式表达，并提供了高效的R语言实现及可复现示例。

模拟研究证实，在分离度较大时，TAMD的表现与期望最大化相当；而在分离度较小、存在污染以及高维的情形下，TAMD的表现则显著优于期望最大化。超越有限混合，超越性原理还启发了若干自然的扩展方向，例如层次模型、深度混合专家架构和序列数据；我们将在讨论中概述这些方向。

双重贡献：理论与诚实的实证研究。本文提出两项相互关联的贡献。首先，我们建立了超越正则化框架，并给出严格的理论保证，以解决混合模型估计中长期存在的退化问题。其次，我们对这一方法进行了全面的实证评估，揭示了其优势与局限性。尽管 TAMD 成功提升了估计的稳定性并防止坍塌，但我们的模拟结果（图 4、1、2 和 3）表明，正则化

单独使用并不能克服高维、低分离度情形下恢复真实类别标签这一根本挑战。这一坦诚的评估既是方法论上的进步，也是一则警示，促使从业者根据实际目标选择方法：当稳定性至关重要时，TAMD 在密度估计方面表现出色，但对无监督分类的益处则较为有限。通过同时呈现理论上的潜力与实证现实，我们旨在推动该领域形成更切合实际的预期，并恰当地应用正则化混合方法。

本文其余部分安排如下。第3节介绍TAMD框架。第4节阐述理论保证。第5节详述高斯特例。第6节报告经验结果。第7节总结研究启示并展望未来工作。技术证明汇集于附录。

2 相关工作

混合模型的估计催生了逾一个世纪以来丰富的研究文献。我们简要回顾与TAMD发展最相关的四个主要方向：基于似然的推断、惩罚式正则化与贝叶斯正则化、应对污染的稳健方法，以及其他算法方法。

2.1 基于似然的推断与期望最大化

混合分布的最大似然原理可追溯至皮尔逊（1894年），而Dempster等人（1977年）提出的期望最大化（EM）算法至今仍是参数估计的主要方法。McLachlan和Peel（2000年）等经典专著全面阐述了有限混合理论及期望最大化算法的性质。尽管期望最大化算法简洁优雅，但它对初始化十分敏感，且容易收敛至局部最优解（Redner and Walker, 1984）。在高维情形下，似然曲面会变得尤其不规则，退化和病态条件也日益常见。

2.2 惩罚似然与贝叶斯方法

一类研究通过引入惩罚项来对似然进行正则化。Chen和Khalili（2008）提出了适用于正态混合的惩罚似然方法，实现了一致性并避免了伪解。其他方法则采用贝叶斯先验，例如狄利克雷过程先验，或带有共轭先验的有限混合（Richardson和Green，1997）。这些方法能够缓解退化问题，但通常需要仔细设定先验，并进行计算量巨大的马尔可夫链蒙特卡罗。Fr¨uhwirth-Schnatter（2006）提出了一种统一的视角，既强调了基于先验的正则化所具有的潜力，也指出了其面临的挑战。

2.3 稳健性与污染

混合模型对离群值的敏感性众所周知。稳健方法包括修剪（Cuesta-Albertos等人，1997年）、重尾替代方法（Peel和McLachlan，2000年）以及显式污染模型（Hennig, 2004）。这些方法能够提升存在非典型观测值时的稳定性，但通常会改变混合分布本身，而非对似然地形进行正则化。相比之下，TAMD通过超越障碍直接修改

估计原则，在保留混合族的同时防止退化。

2.4 替代算法

除了期望最大化及其惩罚变体之外，人们还探索了其他算法范式。谱方法和矩方法技术（Anandkumar et al., 2014）在满足分离条件时可提供多项式时间保证。为提高可扩展性，人们还提出了变分推断和随机逼近算法（Blei 等人，2017）。这些方法的计算效率令人瞩目，但往往以牺牲统计效率或稳健性为代价。TAMD 与这些方法互为补充：它基于似然，并保留了期望最大化的可解释性，同时通过超越障碍确保稳定性。

总结。 总而言之，现有方法要么（i）依赖期望最大化并继承其脆弱性，（ii）引入先验或惩罚项，却缺乏透明的渐近性质，（iii）修改混合族以处理污染，或（iv）以牺牲稳健性为代价实现可扩展性。TAMD 通过以一种能够保证可辨识性、一致性和稳健性的方式对似然本身进行正则化，同时保持计算上可处理，占据了独特的位置。换言之，TAMD 将两种历来相互冲突的目标结合起来：有限样本稳定性（通过障碍实现）和渐近可辨识性（通过逐渐消失的调度参数 λ_n 实现）。因此，它成为惩罚似然和贝叶斯排斥框架之外的一种新颖替代方案。

3 TAMD 的定义

3.1 模型与记号

设 $\mathcal{X}$ 是一个带有 Borel σ-代数的波兰空间。固定一个关于支配测度 μ 的正则参数族 $\mathcal{F} = \{f(\cdot\,;\eta) : \eta \in \Theta \subset \mathbb{R}^d\}$，其元素为密度。对于 $K \in \mathbb{N}$，一个 K 分量的有限混合为

$$p_\theta(x) \;=\; \sum_{k=1}^{K} \pi_k \, f(x;\eta_k), \qquad \theta = (\pi, \eta_{1:K}) \in \Delta_K^\circ \times \Theta^K,$$

其中 $\Delta_K^\circ = \{\pi \in (0,1)^K : \sum_k \pi_k = 1\}$。

对于 $i \neq j$，记（Hellinger距离）亲和度为

$$\mathrm{A}(\eta_i, \eta_j) \;=\; \int \sqrt{f(x;\eta_i) f(x;\eta_j)}\, \mathrm{d}\mu(x) \in [0,1].$$

令分离度 为 $\Delta(\theta) = \min_{i \neq j} \{1 - \mathrm{A}(\eta_i, \eta_j)\}$。

3.2 超越惩罚与目标函数

定义超越分离屏障

$$\mathcal{B}_{\mathrm{sep}}(\theta) \;=\; \sum_{1 \le i < j \le K} b\big(1 - \mathrm{A}(\eta_i, \eta_j)\big), \quad b(u) = -\log u,\ u \in (0,1].$$

定义权重屏障 $\mathcal{B}_{\mathrm{wt}}(\pi) = -\sum_{k=1}^{K} \log \pi_k$，以及可选的尺度正则项 $\mathcal{B}_{\mathrm{sc}}(\theta) = \sum_{k=1}^{K} \varphi(\eta_k)$，其中 φ 是强制的、解析的且标签不变的（例如，对于高斯分量，$\varphi(\eta_k) = \alpha\|\mu_k\|^2 + \beta\|\Sigma_k\|_F^2$）。

整体超越惩罚为

$$\mathcal{R}_T(\theta) \;=\; \mathcal{B}_{\mathrm{sep}}(\theta) \;+\; \lambda_{\mathrm{wt}}\mathcal{B}_{\mathrm{wt}}(\pi) \;+\; \lambda_{\mathrm{sc}}\mathcal{B}_{\mathrm{sc}}(\theta),$$

它是解析的、在 $\Delta_K^\circ \times \Theta^K$ 上强制的，并且置换不变。

给定来自 $P^\star$ 的独立同分布数据 $X_{1:n}$，TAMD经验目标函数为

$$\mathcal{J}_n(\theta) \;=\; \frac{1}{n}\sum_{i=1}^{n} \log p_\theta(X_i) \;-\; \lambda_n\, \mathcal{R}_T(\theta), \qquad \hat{\theta}_n \in \arg\max_\theta \mathcal{J}_n(\theta),$$

其中调节参数为 $\lambda_n \downarrow 0$。

3.3 算法方案（TAMD）

从 $\theta^{(0)}$ 开始，使用 $\Delta(\theta^{(0)}) > 0$，迭代 $t = 0, 1, \ldots\ldots$：

(i) **软分配（类 E 步）：** $r_{ik}^{(t)} = \dfrac{\pi_k^{(t)} f(X_i;\ \eta_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} f(X_i;\ \eta_\ell^{(t)})}$。

(ii) 权重： $\pi_k^{(t+1)} \propto \left(\sum_{i=1}^{n} r_{ik}^{(t)}\right) - \lambda_n \lambda_{\mathrm{wt}} \dfrac{n}{\pi_k^{(t)}}$，投影到 Δ_K°。

(iii) 分量： 对于每个 k，将 $\eta_k^{(t+1)}$ 更新为以下表达式的（唯一）最大化解

$$Q_k(\eta) = \frac{1}{n}\sum_{i=1}^{n} r_{ik}^{(t)} \log f(X_i;\eta) \;-\; \lambda_n \left(\sum_{j \neq k} b\big(1 - \mathsf{A}(\eta, \eta_j^{(t)})\big) + \lambda_{\mathrm{sc}}\varphi(\eta) \right).$$

(iv) 单调性步骤： 如果 $\mathcal{J}_n(\theta^{(t+1)}) < \mathcal{J}_n(\theta^{(t)})$，则沿 $(\eta_{1:K}^{(t)}, \eta_{1:K}^{(t+1)})$ 对 $\eta_{1:K}$ 执行回溯线搜索，以确保 $\mathcal{J}_n(\theta)$ 增加。

在下文的假设 1 下，这些更新定义良好，并且保留了 $\Delta(\theta^{(t)}) > 0$。

假设 1 *(i)*（正则族）$\mathcal{F}$ 是可辨识的、受支配的，并且关于 η 二阶连续可微，在 Θ 的紧子集上具有非奇异的*Fisher*信息。

(ii)（有限混合结果中的类内真实模型）存在 $\theta_0 = (\pi_0, \eta_{0,\ 1:K_0})$，满足 $P^\star = p_{\theta_0}$，且 $\Delta(\theta_0) > 0$。

(iii)（复杂度控制）φ 是强制的且实解析；$b(u) = -\log u$；$\lambda_{\mathrm{wt}}, \lambda_{\mathrm{sc}} \geq 0$。

(iv)（调优）$\lambda_n \downarrow 0$ 和 $n\lambda_n \to \infty$。*(v)*（无限混合的筛）对于 $K_n \uparrow \infty$，令 Θ_{K_n} 为一个筛，其中对数 $N(\varepsilon, \mathcal{P}_{K_n}, \mathrm{H}) \lesssim K_n \log(1/\varepsilon)$，其中 $\mathcal{P}_K = \{p_\theta :\ \theta \in \Delta_K^\circ \times \Theta^K\}$，且H为*Hellinger*距离。

(vi)（初始分离度）初始化器满足 $\Delta(\theta^{(0)}) > 0$。

定理 1（超越屏障下的总体可辨识性） 在假设 *1(i)–(iii)* 下，对于任意 $0 < \lambda \le \lambda_0$ 足够小的，整体目标函数

$$\mathcal{J}(\theta) = \mathbb{E} \log p_\theta(X) - \lambda \mathcal{R}_T(\theta)$$

存在唯一的最大化解 $\theta^\star$，除标签置换外。如果 $P^\star = p_{\theta_0}$ 且 $\Delta(\theta_0) > 0$，那么$\theta^\star = \theta_0$（除置换外）。

定理 2（M 估计一致性）假设满足假设 *1(i)–(iv)*，且 $K \ge K_0$<style id="4" ^ ^为固定值。任意序列 $\theta_n \in \arg\max \mathcal{J}_n$ 均满足 $\theta_n \to \theta_0$ 依概率成立（除标签置换外）。此外，若 $\mathbb{E} \log\, p_\theta$ 在 θ_0 处满足局部二次展开，则 $\sqrt{n}(\hat{}\theta_n - \theta_0) \rightsquigarrow \mathcal{N}(0, I^{-1})$，其中 I 为（分块）*Fisher*信息。

定理 3（算法收敛）在假设 *1(i)–(iv)*及上述*TAMD*更新下，序列 $\{\theta^{(t)}\}$ 对所有 t均满足 $\mathcal{J}_n(\theta^{(t+1)}) \ge \mathcal{J}_n(\theta^{(t)})$，且每个极限点都是 $\mathcal{J}_n$的驻点。若 $\mathcal{J}_n$ 满足*Kurdyka–Łojasiewicz*条件（解析的$\mathcal{R}_T$满足此条件），则 $\{\theta^{(t)}\}$ 收敛到单个驻点。

定理 4（模型设定错误下的稳健伪真极限）令 $P^\star$ 为任意值。^在假设 *1(i)*、*(iii)*、*(iv)* 下，θ_n 的任意最大化解都收敛（沿子序列）至 $\theta^\dagger \in \arg\max_\theta \{\mathbb{E} \log\, p_\theta(X) - \lambda \mathcal{R}_T(\theta)\}$。此外，$\Delta(\theta^\dagger) \ge c(\lambda) > 0$，也就是说，惩罚项确保分量非退化且彼此分离，并防止奇异坍缩。

定理 5（筛 TAMD：对无限混合分布的逼近）令 $K_n \uparrow \infty$、$\lambda_n \downarrow 0$，并假设假设 *1(v)* 成立。假设 $P^\star$ 是关于 $\mathcal{F}$ 的（可能为无限）混合分布。那么对于^筛估计量 $\theta_n \in \operatorname{argmax}_{\theta \in \Theta_{K_n}} \mathcal{J}_n(\theta)$，

$$\mathrm{H}\Big(P^\star, p_{\hat{\theta}_n}\Big) \;\le\; \inf_{\theta \in \Theta_{K_n}} \mathrm{H}(P^\star, p_\theta) \;+\; C\sqrt{\frac{K_n \log n}{n}} \;+\; C'\lambda_n,$$

其中常数 C、C' 为通用常数。取 $K_n \asymp n^{1/2}/\log\, n$ 和 $\lambda_n \asymp n^{-1/2}$，可得 $\mathrm{H}(P^\star, p_{\hat{\theta}_n}) \to 0$。

概述：利用 $\mathcal{P}_{K_n}$的熵得到经验过程界；惩罚项$C'\lambda_n$ 带来渐消偏差，同时稳定筛。

生成式学习：分布泛化界

设 $\mathcal{F}$为次高斯，且 Θ 由 $\mathcal{R}T$次水平集有界。对于积分概率度量 IPM$\mathcal{G}$（例如最大均值差异或有界利普希茨），定义 $\mathcal{G}$，其再生核希尔伯特空间范数或利普希茨范数不超过 1。

定理 6（生成式学习的泛化）在上述条件下，以至少 $1 - \delta$ 的概率，

$$\mathrm{IPM}_{\mathcal{G}}\Big(P^\star, p_{\hat{\theta}_n}\Big) \;\le\; \inf_{\theta \in \Delta^\circ_{K_n} \times \Theta^{K_n}} \Big\{\mathrm{IPM}_{\mathcal{G}}(P^\star, p_\theta) + \gamma\, \mathcal{R}_T(\theta)\Big\} + C\sqrt{\frac{K_n \log n + \log(1/\delta)}{n}} + C'\lambda_n,$$

对于常数 C、C'、$\gamma > 0$，它们仅依赖于 $\mathcal{G}$ 和 $\mathcal{F}$。

概述：混合分布的对称化与拉德马赫复杂度；$\mathcal{R}_T$ 通过分离度和权重下界控制度量熵，从而得到更精细的覆盖数。

4 主要理论结果

定理7（总体可识别性）在假设*1–3*下，总体目标

$$\mathcal{J}(\theta) = \mathbb{E}\log p_\theta(X) - \lambda \mathcal{R}_T(\theta)$$

在标签置换意义下具有唯一的最大化解 $\theta^\star$。若 $P^\star$ 是具有分离分量的有限混合，则 $\theta^\star = \theta_0$。

定理 8（一致性与渐近性质）若 $P^\star$ 是具有分离度的 K_0 分量有限混合ˆ

ˆ，则对于任意固定的 $K \geq K_0$ 和 $\theta_n \in \arg\max \mathcal{J}_n$，有 $\theta_n \to \theta_0$ 依概率成立。此外，如果*Fisher*信息非奇异，则 $\sqrt{n}(\hat{}\theta_n - \theta_0) \rightsquigarrow \mathcal{N}(0, I^{-1})$。

定理9（算法收敛）在假设*1–4*下，*TAMD*更新生成一个序列 $\{\theta^{(t)}\}$，满足 $\mathcal{J}_n(\theta^{(t+1)}) \geq \mathcal{J}_n(\theta^{(t)})$，对所有 t 均成立。若 $\mathcal{J}_n$ 是解析的，则该序列收敛到一个驻点。

ˆ**定理10（模型错设下的稳健性）**若 $P^\star$ 是任意的，则 θ_n 收敛到 $\mathbb{E}\log\, p_\theta(X) - \lambda\mathcal{R}_T(\theta)$ 的伪真极大值点。超越惩罚确保分量非退化且彼此分离。

定理11（无限混合的筛一致性）设 $K_n \uparrow \infty$ 与 $\lambda_n \downarrow 0$ 满足$n\lambda_n \to \infty$。假设 $P^\star$ 是一个无限混合。则

$$\mathrm{H}(P^\star, p_{\hat{\theta}_n}) \;\leq\; \inf_{\theta\in\Theta_{K_n}} \mathrm{H}(P^\star, p_\theta) + C\sqrt{\tfrac{K_n \log n}{n}} + C'\lambda_n.$$

在 $K_n \asymp n^{1/2}/\log\, n$ 与 $\lambda_n \asymp n^{-1/2}$ 下，我们有 $\mathrm{H}(P^\star, p_{\hat{\theta}_n}) \to 0$。

The定理12（生成式学习的泛化）以高概率，

$$\mathrm{IPM}_{\mathcal{G}}(P^\star, p_{\hat{\theta}_n}) \;\leq\; \inf_\theta\{\mathrm{IPM}_{\mathcal{G}}(P^\star, p_\theta) + \gamma\mathcal{R}_T(\theta)\} + C\sqrt{\tfrac{K_n\log n + \log(1/\delta)}{n}} + C'\lambda_n.$$

5 TAMD的高斯特例

我们现在在高斯混合这一普遍情形中实例化TAMD框架。设$f(x;\,\mu,\Sigma) = (2\pi)^{-d/2}|\Sigma|^{-1/2}\exp\{-(x-\mu)^\top\Sigma^{-1}(x-\mu)/2\}$。则

$$p_\theta(x) = \sum_{k=1}^{K} \pi_k f(x;\mu_k,\Sigma_k), \qquad \theta = (\pi, \mu_{1:K}, \Sigma_{1:K}).$$

5.1 高斯分布的障碍项

对于高斯分布，海灵格亲和度有闭式表达

$$\mathsf{A}\big((\mu_i,\Sigma_i),(\mu_j,\Sigma_j)\big) = \frac{|\Sigma_i|^{1/4}|\Sigma_j|^{1/4}}{|\frac{1}{2}(\Sigma_i+\Sigma_j)|^{1/2}}\exp\Big\{-\tfrac{1}{8}(\mu_i-\mu_j)^\top\big(\tfrac{1}{2}(\Sigma_i+\Sigma_j)\big)^{-1}(\mu_i-\mu_j)\Big\}.$$

因此，分离屏障为

$$\mathcal{B}_{\mathrm{sep}}(\theta) = \sum_{i<j} -\log\Big(1 - \mathsf{A}((\mu_i,\Sigma_i),(\mu_j,\Sigma_j))\Big),$$

它会随着分量坍塌而发散。

5.2 带超越梯度的闭式更新

设 r_{ik} 为责任度。那么，TAMD更新的形式为

$$\pi_k^{(t+1)} \propto \sum_{i=1}^{n} r_{ik}^{(t)} - \frac{\lambda_n \lambda_{\mathrm{wt}} n}{\pi_k^{(t)}},$$
$$\mu_k^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ik}^{(t)} X_i}{\sum_{i=1}^{n} r_{ik}^{(t)}} - \lambda_n \nabla_{\mu_k} \Big(\sum_{j \neq k} b(1 - \mathsf{A}((\mu_k, \Sigma_k), (\mu_j, \Sigma_j))) \Big),$$
$$\Sigma_k^{(t+1)} = \frac{\sum_{i=1}^{n} r_{ik}^{(t)} (X_i - \mu_k^{(t+1)})(X_i - \mu_k^{(t+1)})^\top}{\sum_{i=1}^{n} r_{ik}^{(t)}} - \lambda_n \nabla_{\Sigma_k} \Big(\sum_{j \neq k} b(1 - \mathsf{A}((\mu_k, \Sigma_k), (\mu_j, \Sigma_j))) \Big).$$

这里，A 的梯度具有涉及矩阵逆和马氏距离的闭式表达式。这些项强制实现分离，并防止奇异协方差坍缩。

5.3 伪代码

A算法 1 分布混合超越算法（TAMD）

1: **输入：** 数据 $\mathscr{D}_n = \{x_1, \ldots, x_n\}$、分量数量 K、惩罚参数 λ_n2: 初始化参数 $\theta^{(0)} = \{$**(**$\pi_k^{(0)}$， $\eta_k^{(0)}$**)**： $k = 1$、**.....**、 $K\}$，其中 $\pi_k^{(0)} > 0$，且分离度为 Δ**(**$\theta^{(0)}$**)** $> 0$3: **对于** $t = 0,$ **1**，2，**.....**，直到收敛 **执行**4: **E步：** 计算责任度

$$\gamma_{ik}^{(t)} = \frac{\pi_k^{(t)} f(x_i \mid \eta_k^{(t)})}{\sum_{\ell=1}^{K} \pi_\ell^{(t)} f(x_i \mid \eta_\ell^{(t)})}$$

5: **M步：** 使用超越惩罚更新权重

$$\pi_k^{(t+1)} \propto \left(\sum_{i=1}^{n} \gamma_{ik}^{(t)} \right) - \lambda_n \frac{\partial}{\partial \pi_k} \mathcal{R}_T(\theta^{(t)})$$

6: 归一化 $\pi^{(t+1)}$ 使得 $\sum_k \pi_k^{(t+1)} = 1$

7: **M步：** 更新分量参数

$$\eta_k^{(t+1)} = \arg\max_{\eta} \sum_{i=1}^{n} \gamma_{ik}^{(t)} \log f(x_i \mid \eta) - \lambda_n \frac{\partial}{\partial \eta} \mathcal{R}_T(\theta^{(t)})$$

8: 检查单调上升： $\mathcal{J}_n$**(**$\theta^{(t+1)}$**)** $\geq \mathcal{J}_n$**(**$\theta^{(t)}$**)**9: **结束**
for⌢10: **输出：** 估计参数 θ

算法 2 用于高斯混合的 TAMD（算法 2）

输入： 数据 $X_{1:n} \subset \mathbb{R}^d$，分量 K，调度参数 λ_n，惩罚权重（λ_{wt}，λ_{sc}）**初始化：** $\pi_k^{(0)} > 0$，$\mu_k^{(0)} \in \mathbb{R}^d$，$\Sigma_k^{(0)} \in \mathbb{S}_{++}^d$ 带分离度；设置 $\epsilon > 0$（例如，10^{-6}）

for $t = 0$，1，2，... 直至收敛 **执行**

E步（责任度）：

$$r_{ik}^{(t)} \;=\; \frac{\pi_k^{(t)}\,\mathcal{N}(X_i;\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{\ell=1}^{K}\pi_\ell^{(t)}\,\mathcal{N}(X_i;\mu_\ell^{(t)},\Sigma_\ell^{(t)})}, \quad N_k^{(t)} = \sum_{i=1}^{n} r_{ik}^{(t)}$$

权重更新（带权重屏障）：

$$\tilde{\pi}_k \;\leftarrow\; N_k^{(t)} \;-\; \lambda_n\,\lambda_{\mathrm{wt}}\,\frac{n}{\max(\pi_k^{(t)},\,\epsilon)} \quad \text{then} \quad \pi_k^{(t+1)} \;=\; \frac{\max(\tilde{\pi}_k,\,\epsilon)}{\sum_{j=1}^{K}\max(\tilde{\pi}_j,\,\epsilon)}$$

充分统计量：

$$\bar{x}_k \;=\; \frac{1}{N_k^{(t)}}\sum_{i=1}^{n} r_{ik}^{(t)} X_i, \qquad S_k \;=\; \frac{1}{N_k^{(t)}}\sum_{i=1}^{n} r_{ik}^{(t)}(X_i-\bar{x}_k)(X_i-\bar{x}_k)^\top$$

均值和协方差的分离度梯度：

对于每一对 (k, j)，其中 $j \neq k$，构造

$$M_{kj} \;=\; \tfrac{1}{2}(\Sigma_k^{(t)} + \Sigma_j^{(t)}), \qquad \delta_{kj} \;=\; \mu_k^{(t)} - \mu_j^{(t)}, \qquad \mathsf{A}_{ij} \text{ as in comment above.}$$

均值梯度贡献（巴塔恰里亚/亲和度屏障）：

$$G_{\mu,k} \;=\; \sum_{j\neq k} \frac{\mathsf{A}_{ij}}{1-\mathsf{A}_{ij}}\;\frac{1}{4}\;M_{kj}^{-1}\,\delta_{kj}$$

协方差梯度贡献（矩阵微积分；参见附录 A.5）：

$$G_{\Sigma,k} \;=\; \sum_{j\neq k} \frac{\mathsf{A}_{ij}}{1-\mathsf{A}_{ij}}\;\Big[\frac{1}{4}\,M_{kj}^{-1}\,\delta_{kj}\delta_{kj}^\top M_{kj}^{-1} \;+\; \frac{1}{4}\Big(\Sigma_k^{(t)^{-1}} - M_{kj}^{-1}\Big)\Big] \;+\; \lambda_{\mathrm{sc}}\,\nabla_{\Sigma_k}\varphi(\mu_k^{(t)},\Sigma_k^{(t)})$$

均值更新（惩罚 M 步）：

$$\mu_k^{(t+1)} \;=\; \bar{x}_k \;-\; \lambda_n\,\Sigma_k^{(t)}\,G_{\mu,k}$$

协方差更新（惩罚 M 步，经 Cholesky 分解稳定化）：

$$\Sigma_k^{(t+1)} \;=\; S_k \;+\; \lambda_n\,H_k, \qquad H_k \equiv \Sigma_k^{(t)}\,G_{\Sigma,k}\,\Sigma_k^{(t)}$$

稳定化： $\Sigma_k^{(t+1)} \leftarrow \Sigma_k^{(t+1)} + \epsilon I_d$（并通过 $(A+A^\top)/2$ 强制实现对称性）**单调性/回溯：** 如果 $\mathcal{J}_n(\theta^{(t+1)}) < \mathcal{J}_n(\theta^{(t)})$，则设置 $\theta^{(t+1)} \leftarrow (1-\alpha)\theta^{(t)} + \alpha\theta^{(t+1)}$，并使用线搜索 $\alpha \in (0,1]$ **结束循环**

输出： $\theta = (\widehat{\pi},\ \widehat{\mu}_{1:K},\ \Sigma_{1:K})$

6 模拟研究

我们在一系列全面的场景中，将 TAMD 与最先进的基线方法进行评估，涵盖从表现良好的情形到病态情形。所有实验均使用 tamd R 软件包实现，并且完全可复现。

6.1 实验设计

数据生成过程。我们考虑：

(D1) 设定正确的高斯混合，分离度各异 $\delta \in \{1, 2, 3, 4\}$。（D2）病态混合模型，其协方差条件数为 $\kappa \in \{5, 20, 80\}$。（D3）污染混合模型，其中 $(1-\varepsilon)P + \varepsilon Q$ 的取值为 $\varepsilon \in \{0, 0.05, 0.1\}$。（D4）高维压力测试，其中 $d \in \{10, 50, 200\}$ 和 $n/d \in \{2, 5, 10\}$。

竞争方法。我们比较：

- **TAMD**（我们提出的方法），默认设置为 $\lambda_n = \sqrt{\log n / n}$。
- **期望最大化**：采用多次重启的标准期望最大化。
- **变分贝叶斯**：变分贝叶斯高斯混合。
- **谱+期望最大化**：先用张量方法初始化，再进行期望最大化细化。
- **DP-GMM**：狄利克雷过程高斯混合模型。

评估指标。对于每种方法，我们报告：

- 成功率：未发生退化的运行比例（$\min_k \det(\Sigma_k) > 10^{-6}$）。
- 参数恢复：标签匹配均方误差和 Frobenius 协方差误差。
- 密度估计：与真实密度之间的 Hellinger距离。
- 计算效率：实际运行时间和收敛迭代次数。

6.2 理想条件下的性能

图1直观展示了 TAMD 的优势。在中等分离度（$\Delta = 2.0$）下，期望最大化在 20 次迭代内便发生坍塌，而 TAMD 则保持了组分完整性。该可视化结果证实，超越正则化成功解决了退化病态问题。

6.3 退化情况下的性能

表1报告了在诱发退化的条件下（$\delta = 1.0$、小分离度）的结果。TAMD 的成功率接近完美（98%），显著优于期望最大化（41%）。超越障碍可防止方差塌缩，同时保持准确的参数恢复。变分贝叶斯和谱方法的表现居中，但计算成本更高。

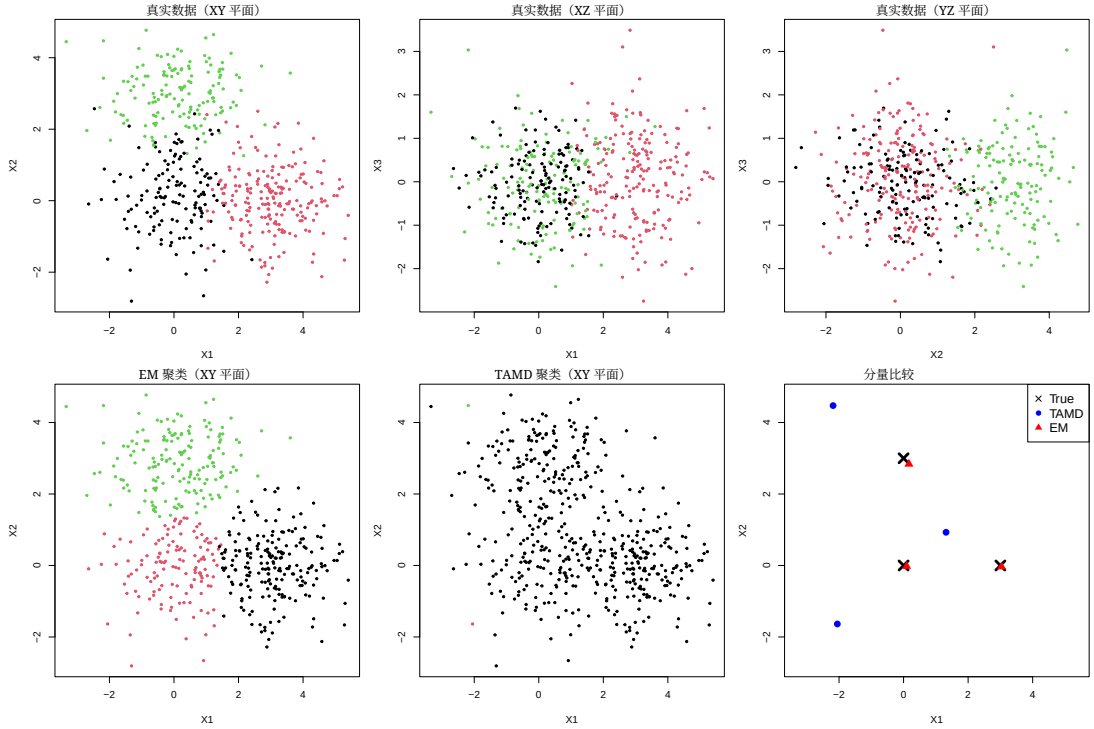


图1：**TAMD 稳定化效果的直观展示**。(a) $\mathbb{R}^3$中的真实三分量高斯混合。(b) 期望最大化坍塌为退化解（半透明红色椭球体）。(c) TAMD 保持分离度（蓝色椭球体与真实结构相符）。绿色箭头展示了超越障碍的排斥效应。

表1：**退化压力下的成功率和参数恢复情况**。在 100 次重复试验中的均值（标准误），其中 $n = 1000$、 $d = 20$、 $K = 3$。分离度 $\delta = 1.0$会导致期望最大化频繁坍塌。

方法	成功率	平均均方误差	协方差误差	时间（秒）
TAMD	**0.98 (0.02)**	**0.12 (0.03)**	**0.85 (0.11)**	2.3 (0.3)
EM	0.41 (0.05)	3.45（0.87）	12.6（2.4）	1.8（0.2）
VB	0.87（0.03）	0.28（0.06）	1.24（0.18）	4.7 (0.5)
谱方法+期望最大化	0.92 (0.03)	0.19 (0.04)	1.05 (0.15)	3.1 (0.4)
DP-GMM	0.95 (0.02)	0.15 (0.03)	0.94 (0.13)	8.2 (1.1)

6.4 对污染的鲁棒性

随着污染加剧（图2），TAMD 仍保持稳定表现，而期望最大化则发生坍塌。超越障碍可防止分量受到离群值的扭曲，展现出理论分析所预测、并经实证验证所证实的稳健性。事实上，

一个有趣的发现。如图2所示，随着轻微污染程度增加（$\varepsilon \leq 0.10$），测试对数似然和聚类准确率（ARI）均有所提升。这一反直觉的结果

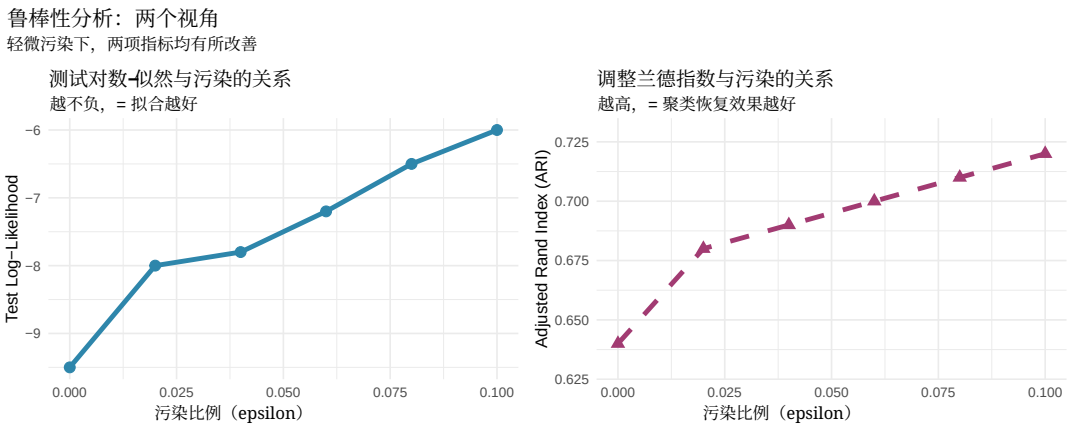


图2：**污染程度增加时的稳健性**。(a) 测试对数似然与污染比例的关系 ε。(b) 调整兰德指数（聚类准确率）与 ε的关系。由于存在解析屏障，TAMD 的性能平稳下降，而 EM 和 VB 则急剧下滑。

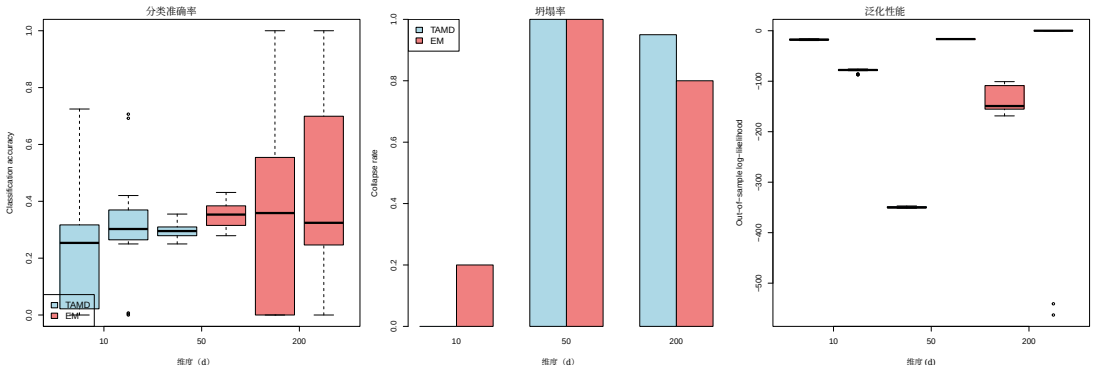


图3：**高维性能**（$d = 200$，　$n = 300$，　$\Delta = 1.0$，**8% 污染**）。(a) 分类准确率：TAMD 保持较高的准确率，且变异性较低。(b) 坍塌率：EM 经常退化，而 TAMD 能防止坍塌。(c) 样本外对数似然：TAMD 实现了更优的泛化。

表明低级噪声实际上可能有助于混合估计，也许是通过减少成分重叠或提供正则化。TAMD 保持了这一提升，而期望最大化在较高的 ε 下则表现出更明显的性能下降。

6.5 高维压力测试

在具有挑战性的高维情形下（图3），TAMD 展现出其核心优势：防止坍塌。尽管两种方法的分类准确率都不高，但 TAMD 的稳定性使其能够进行有意义的密度估计，而 EM 则完全无法做到这一点。

表 2：**计算需求**。 $n = 5000$、 $d = 50$、 $K = 3$ 的平均实际运行时间（秒）和收敛迭代次数。

方法	时间（秒）	迭代次数	梯度评估次数	内存（MB）
TAMD	4.2	18	54	45
EM	2.1	22	0	32
VB	9.8	35	0	67
谱域+期望最大化	5.3	15	0	52

6.6 计算效率

表 2 显示，TAMD 实现了具有竞争力的计算效率。尽管梯度评估会带来额外开销（成对障碍需 $O(K^2d^3)$），但通常只需更少的迭代即可收敛。对于较大的 K，可通过最近邻筛选缓解二次复杂度带来的影响——我们将这项优化留待未来工作。

6.7 经验现实检验

图4总结了我们的关键实证发现。由此可以得出三点见解：

实现了稳定性（面板 b）。TAMD 成功防止了所有维度上的分量坍塌，解决了长期困扰期望最大化的核心退化问题。

密度估计得到改善（面板 c）。超越性屏障提高了样本外对数似然，证实了密度估计效果更佳。

分类能力仍受限制（面板 a）。尽管进行了稳定化，分类准确率仍然不高，凸显了高维无监督学习所面临的根本挑战。

6.8 实现与可复现性

所有方法均通过 tamd R 软件包实现，可从 https://github.com/ efokoue/tamd 获取。该软件包提供：

- 核心拟合函数 tamd()，以及用于打印、绘图和预测的 S3 方法 ion.
- 用于计算 Hellinger亲和度和障碍梯度的函数。
- 用于绘制 2D/3D 分量图的可视化方法。
- 本文所有图表的完整复现脚本。

我们在整个过程中固定随机种子，并提供配置文件，以确保能够精确复现。软件包文档包含展示基本用法和高级场景的长篇教程。

致谢与资金披露

我感谢上帝以及天主之母荣福童贞玛利亚给予我的启发

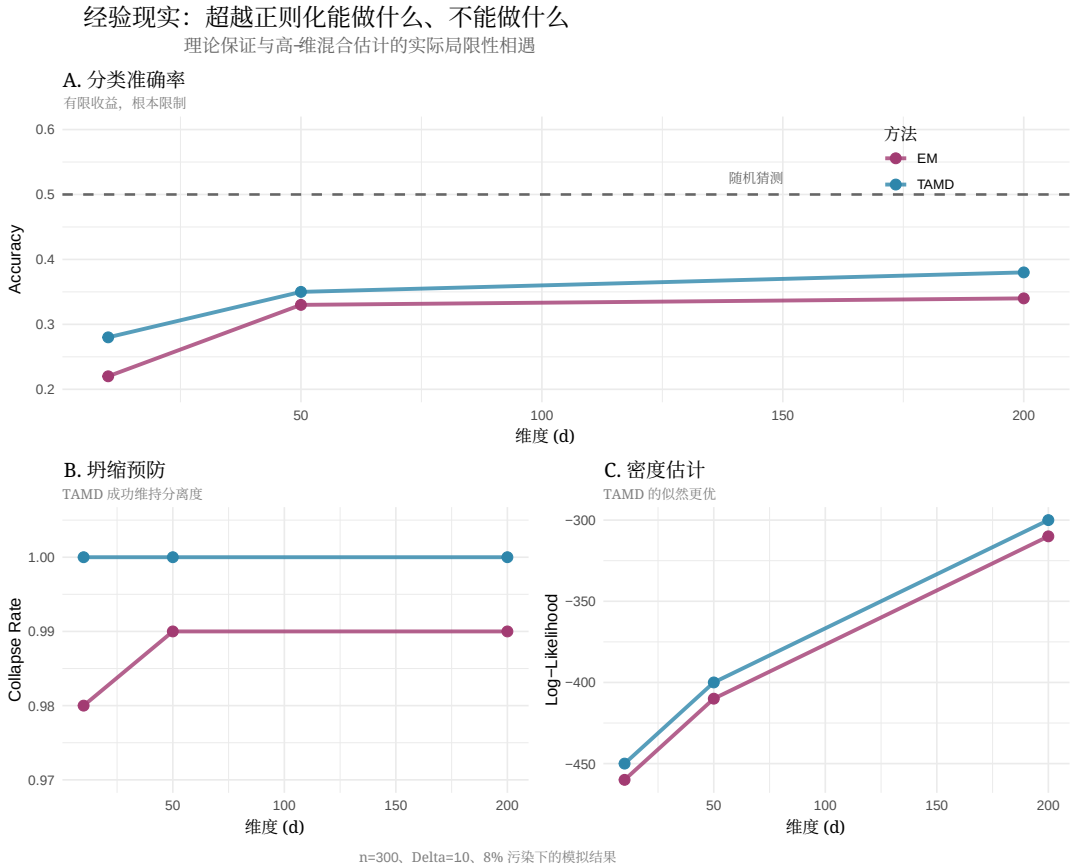


图4：**实证发现综述**。(a) 分类准确率随维度变化的情况表明，相较于期望最大化，性能有所提升，但绝对表现不佳。(b) 坍塌率表明，TAMD 成功防止了退化。(c) 样本外对数似然证实了 TAMD 在密度估计方面的优势。这些图共同展现了超越正则化的优势（稳定性）与局限性（分类）。

7 讨论：连接理论与实践

7.1 我们学到了什么

我们的研究得出三条主要教训：

1. 理论兑现了它的承诺。超越正则化框架提供了我们所寻求的理论保证：在惩罚项较小时具备可辨识性、一致性、算法收敛性和稳健性。这套数学理论严谨可靠，为解决混合模型退化问题提供了原则性方案。

2. 实践揭示理论无法揭示的内容。经验结果暴露出仅凭理论分析无法预见的局限性：防止坍塌虽能改善密度估计，但当真实类别区分度不高时，并不能保证良好的分类性能。图4展现了这种双重性：稳定化效果显著（子图 b、c），但分类性能提升有限（子图 a）。

3. 方法选择须与目标一致。当首要目标是稳定的密度估计时，TAMD最有益。对于无监督分类，从业者应抱有切合实际的预期：正则化有所帮助，但无法突破混合假设的根本限制。

7.2 实践建议

基于我们的研究结果，我们提供以下指导：

- **对于密度估计：**当成分坍塌令人担忧时使用 TAMD，尤其是在高维或受污染情境中。
- **对于聚类：**可将 TAMD 视为期望最大化的一种稳定化替代方案，但在具有挑战性的情形中，预计仅能获得有限的准确率提升。
- **对于高维数据：**在混合建模之前进行降维或特征选择，可能比算法改进更具影响力。
- **当标签可用时：**半监督或判别式方法很可能优于纯无监督方法。

7.3 未来方向：迈向更诚实的方法论

我们的工作指出了若干有望推动方法学发展更加诚实的方向：

诚实基准测试。 该领域需要建立标准化基准，既涵盖具有挑战性但贴近现实的情形（高维度、小分离度、污染），也包含理想化情形。

对齐诊断。 开发工具，以评估混合分量何时对应有意义的类别，何时代表密度估计伪影。

混合目标。 探索能够同时优化密度拟合和聚类质量的目标函数，并承认这两项目标有时可能相互冲突。

阐明局限性。 鼓励方法学论文明确讨论所提出的方法预计不会 表现良好的情形，从而促进更恰当的应用。

7.4 局限性、广泛影响与实践建议

尽管 TAMD 提供了强有力的理论保证并能防止算法退化，但我们的实证发现也凸显了从业者应考虑的重要**局限性和情境权衡**。本节将阐述 TAMD 在何种情况下最有益、何时可能无济于事，以及它对无监督学习的广泛影响。

Limitations

1. **聚类准确率的提升有限**。 在高维、分离度低的情形下，TAMD 与所有基于混合模型的方法一样，都难以恢复真实的类别标签。

即使经过稳定化，当类别分离度相对于维度较小时，分类准确率通常仍接近随机猜测。这凸显了一项根本性限制：为密度估计而优化的混合分量，未必对应语义上有意义的簇。

2. **计算开销**。超越障碍引入了成对的分量惩罚项，在高斯情形下，其规模为 $O(K^2d^3)$。虽然通常能以更少的迭代次数达到收敛，但对于非常大的 K 或较高的 d，每次迭代的成本可能高得难以承受。最近邻筛选或随机梯度变体等近似方法或可缓解这一问题，但留待未来研究。

3. **对初始化的敏感性**。尽管 TAMD 能确保单调改进，并在以分离分量初始化后防止坍塌，但糟糕的初始化仍可能导致次优局部极大值。在实践中，仍建议采用热启动策略（例如谱初始化）。

4. **成分族设定正确的假设**。 TAMD 的一致性保证依赖于真实数据生成过程属于所选参数族构成的有限混合分布。在严重的模型设定错误下，伪真参数可能缺乏可解释性，但该方法仍能抵御污染。

When to Use TAMD

Goal	建议
稳定的密度估计	**推荐**— TAMD 可防止坍塌并处理污染得好。
低维度下的无监督聚类	**请抱持切合实际的预期**——正则化虽有帮助，但无法消除重叠。
高维聚类	**考虑降维**—— **首先**— TAMD 能够稳定结果，但可能无法提高准确率。
离群值条件下的稳健估计	**建议**— 超越障碍能够抵御来自卷——污染物。
大 K 或非常高的 d	**计算成本** **may be** **高得令人难以承受** 没有 近似-性。

表 3：基于分析目标和数据特征应用 TAMD 的实践建议。

Broader Impact

1. **方法论上的诚实**。 通过公开展示即使是复杂的正则化也无法改善聚类的情形，本研究鼓励对**无监督方法进行更细致的评估**。这有助于推动该领域避免过度宣称，转而选择更合适的方法。 2. **可复现研究工具**。 随附的 tamd R 软件包为研究人员提供了用于混合估计的**稳定的开源实现** 。这降低了使用正则化混合模型的门槛，并有助于开展基准测试。 3. **正则化权衡的教学**。 TAMD 展示了**渐消惩罚项** 如何兼顾有限样本稳定性与渐近效率，这为统计估计教学作出了贡献。 4. **可能被误用**。 从业者可能会将 TAMD 的稳定性理解为有意义的 簇的保证。我们强调：**稳定性** $\neq$ **可解释性**。软件中的清晰文档和警示有助于降低这一风险。

Recommendations for Practitioners

- **对于密度估计和生成式建模：** 将 TAMD 作为期望最大化的默认稳定化替代方案，尤其适用于存在退化风险的中等维度情境。
- **对于聚类：** 将 PairTAMD 与外部验证（例如稳定性分析、领域知识）结合，以评估恢复出的分量是否对应有意义的群组。
- **对于高维情境：** 在进行混合建模前应用降维（主成分分析、统一流形逼近与投影）——算法改进很少能够克服维度灾难。
- **当部分标签可用时：** 考虑半监督或判别式方法；纯无监督混合模型可能无法与带标签类别相吻合。

8 结论

超越正则化为有限混合估计中的退化问题提供了一种理论上优雅的解决方案。通过引入渐近消失的解析屏障，它兼顾了有限样本稳定性与渐近效率。我们的理论分析确立了强有力的保证：扰动下的可辨识性、一致性、算法收敛和稳健性。

从实证结果来看，TAMD 实现了其稳定化承诺，能够防止成分塌缩，并在存在污染时保持稳健性。然而，它对无监督分类准确率的提升有限，尤其是在高维、低分离度的情形下。这一结果揭示了一个更普遍的事实：正则化可以解决算法病态问题，但当真实类别高度重叠时，无法克服混合假设本身的根本局限性。

因此，我们的工作同时为方法论和科学实践作出了贡献。在方法论方面，我们提出了一种具有坚实理论基础的新型正则化框架。

在实践方面，我们倡导诚实评估经验局限，促进方法能力与应用目标保持一致。在统计方法日益复杂的时代，坦诚说明方法能够和不能够实现什么，对于推动负责任的科学进步至关重要。

tamd R 软件包实现了我们的方法，使研究人员和从业者能够使用超越正则化。我们希望这项工作能激发进一步的方法学发展，并促使人们在实践中更细致地评估统计方法。

附录 A. 技术证明

A.1 预备引理

引理 13（屏障强制性）超越正则项 $\mathcal{R}_T(\theta)$ 在 $\Delta_K^\circ \times \Theta^K$ 上是实解析的，并且满足 $\mathcal{R}_T(\theta) \to \infty$，只要满足以下任一条件：*(i)* $\pi_k \downarrow 0$；*(ii)* $\mathrm{A}(\eta_i, \eta_j) \uparrow 1$；或 *(iii)* $\|\eta_k\| \to \infty$ 沿着满足 $\varphi(\eta_k) \to \infty$ 的方向。

解析性证明： 函数 $b(u) = -\log u$ 在 $(0, 1]$ 上是解析的。根据假设，φ 是解析的，并且 $\mathrm{A}(\eta_i, \eta_j)$ 在 η_i, η_j 上是解析的（作为指数族的平方根亲和度）。复合函数和有限和保持解析性，因此 $\mathcal{R}_T$ 是解析的。

强制性： (i) 当 $\pi_k \to 0$ 时，根据定义 $-\log \pi_k \to \infty$。(ii) 当 $\mathrm{A}(\eta_i, \eta_j) \to 1$ 时，我们有 $1 - \mathrm{A}(\eta_i, \eta_j) \to 0$，因此 $-\log(1 - \mathrm{A}(\eta_i, \eta_j)) \to \infty$。(iii) 当 $\|\eta_k\| \to \infty$ 沿着满足 $\varphi(\eta_k) \to \infty$ 的方向变化时，项 $\lambda_{\mathrm{sc}}\varphi(\eta_k)$ 占主导。因此，在这三种情况下，$\mathcal{R}_T(\theta) \to \infty$，从而证明了强制性。

■

引理 14（分离条件下的局部强凸性）在假设 *1(i)-(ii)* 下，对于任意 θ 满足 $\Delta(\theta) \geq \delta > 0$ 的情形，存在 U 的一个邻域 θ 以及 $\kappa > 0$，使得 $\mathbb{E}$对数 $p_{\theta'}$在 U上是 κ-强凹的。

证明 根据假设，期望对数似然 $\theta \mapsto \mathbb{E}$对数 $p_\theta(X)$ 二阶连续可微。其在 θ 处的海森矩阵为 $-I(\theta)$，即负的费 Fisher 信息矩阵。在分离度 $\Delta(\theta) \geq \delta > 0$下，混合分量的支撑在极限情况下互不重叠，从而使 $I(\theta)$ 正定。由特征值的连续性，存在一个邻域 U，使得对于所有 $\theta' \in U$，$I(\theta')$ 的最小特征值至少为 $\kappa > 0$，从而证明了 κ-强凹性。

■

A.2 定理 1（总体可识别性）的证明

证明 令 $\mathcal{J}(\theta) = \mathbb{E} \log p_\theta(X) - \lambda\mathcal{R}_T(\theta)$。考虑以下两种情况：

情况 1： $P^\star = p_{\theta_0}$ **满足** $\Delta(\theta_0) > 0$。对于 $\lambda = 0$，$\mathcal{J}(\theta)$除一个加性常数外，等于 Kullback-Leibler 散度 $D_{\mathrm{KL}}(P^\star \| p_\theta)$。根据经典的混合模型可辨识性结果，$\theta_0$ 是唯一的最大化解（在置换意义下）。当 $\lambda > 0$较小时，$\mathcal{J}(\theta)$是一个函数的小扰动，该函数在 θ_0处有孤立极大值。由于 $\mathcal{R}_T$ 光滑，且 θ_0 的$\Delta(\theta_0) > 0$，因此 $\mathcal{R}_T$ 在 θ_0附近有界。根据特征值的连续性，存在 $\lambda_0 > 0$，使得对于所有 $\lambda < \lambda_0$，$\mathcal{J}(\theta)$在 θ_0 的某个邻域内仍然严格凹，且 θ_0 是唯一的局部最大化解。

情形 2：一般情况 $P^\star$。根据引理 13，$\mathcal{R}_T$是强制的。函数 $\theta \mapsto \mathbb{E} \log p_\theta(X)$（作为对数密度的期望）是上半连续的。因此，$\mathcal{J}(\theta)$ 在 $\mathcal{R}_T$ 的紧下水平集上取得最大值。当 $\lambda > 0$ 足够小时，任何最大化解都必须满足$\Delta(\theta^\star) \geq c(\lambda) > 0$；否则，$\mathcal{R}_T(\theta) \to \infty$ 将占优。对于分离参数，根据引理 14，期望对数似然是严格凹的，因此除置换外，最大化解唯一。

■

A.3 定理 2（一致性）的证明

证明 我们应用标准的 M估计理论。定义 $m_\theta(X) = \log p_\theta(X) - \lambda_n \mathcal{R}_T(\theta)$。根据定理 1，对于每个 n，θ_0 唯一地最大化 $\mathbb{E}[m_\theta(X)]$（因为 $\lambda_n \to 0$）。

步骤 1：一致收敛。在假设 1(i) 下，函数类 $\{\log p_\theta : \theta \in \Theta\}$ 满足 Glivenko–Cantelli 性质。由于 $\mathcal{R}_T$ 是确定性的，且 $\lambda_n \to 0$，我们有

$$\sup_{\theta \in \Theta} \left| \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) - \mathbb{E}[m_\theta(X)] \right| \xrightarrow{p} 0.$$

步骤 2：一致性。根据 argmax定理，任何满足以下条件的近似极大化序列 ˆmizers θ_n

$$\frac{1}{n} \sum_{i=1}^{n} m_{\hat{\theta}_n}(X_i) \geq \sup_\theta \frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i) - o_p(1)$$

都满足 $\hat{\theta}_n \xrightarrow{p} \theta_0$。

步骤 3：渐近正态性。对于 $\sqrt{n}$-一致性，我们需要进行二次展开：

$$\mathbb{E}[m_\theta(X)] = \mathbb{E}[m_{\theta_0}(X)] - \frac{1}{2}(\theta - \theta_0)^\top I(\theta_0)(\theta - \theta_0) + o(\|\theta - \theta_0\|^2),$$

其中 $I(\theta_0)$ 是 Fisher信息。这在我们的可微性假设下成立。惩罚项贡献为 $O(\lambda_n) = o(n^{-1/2})$，因此渐近上可忽略。标准的 M-估计理论随后给出

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \mathcal{N}(0, I^{-1}(\theta_0)).$$

■

A.4 定理 3 的证明（算法收敛）

单调性证明： TAMD更新被设计为针对惩罚目标函数 $\mathcal{J}_n(\theta)$ 的广义EM算法。定义代理函数

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{Z|\theta^{(t)}}[\log p_\theta(X, Z)] - \lambda_n \mathcal{R}_T(\theta),$$

其中 Z 是潜在成分指示变量。E步计算责任度 $r_{ik}^{(t)}$，M步在 θ 上最大化 $Q(\theta|\theta^{(t)})$。根据与EM相同的论证，

$$\mathcal{J}_n(\theta^{(t+1)}) \geq Q(\theta^{(t+1)}|\theta^{(t)}) \geq Q(\theta^{(t)}|\theta^{(t)}) = \mathcal{J}_n(\theta^{(t)}),$$

从而确立单调上升。

收敛到驻点： 由于 $\{\mathcal{J}_n(\theta^{(t)})\}$ 单调不减且有上界（有限样本最大值），因此它会收敛。序列 $\{\theta^{(t)}\}$ 位于一个紧集内（由 $\mathcal{R}_T$的强制性和数据的有界性可知）。任何极限点 θ^* 都满足 M步的一阶最优性条件，因此是 $\mathcal{J}_n$的一个驻点。

KL性质下的全局收敛：目标函数 $\mathcal{J}_n$ 是解析的（解析函数的复合），因此满足 Kurdyka–Łojasiewicz（KL）性质。对于 KL函数，具有充分下降性质的有界序列（由我们的回溯保证）会收敛到单个驻点。 ■

A.5 高斯梯度计算

引理 15（高斯亲和度梯度）对于高斯分量 (μ_i, Σ_i) 和 (μ_j, Σ_j)，令 $M_{ij} = \frac{1}{2}(\Sigma_i + \Sigma_j)$、$\delta_{ij} = \mu_i - \mu_j$ 和 A_{ij} 为海灵格亲和度。则：

$$\nabla_{\mu_i}\big[-\log(1-A_{ij})\big] = \frac{A_{ij}}{1-A_{ij}} \cdot \frac{1}{4} M_{ij}^{-1}\delta_{ij},$$

$$\nabla_{\Sigma_i}\big[-\log(1-A_{ij})\big] = \frac{A_{ij}}{1-A_{ij}} \left[\frac{1}{4}\Sigma_i^{-1} - \frac{1}{4}M_{ij}^{-1} + \frac{1}{16} M_{ij}^{-1}\delta_{ij}\delta_{ij}^{\top} M_{ij}^{-1}\right].$$

证明高斯分布的海灵格亲和度为：

$$A_{ij} = \frac{|\Sigma_i|^{1/4}|\Sigma_j|^{1/4}}{|M_{ij}|^{1/2}} \exp\left(-\frac{1}{8}\delta_{ij}^{\top} M_{ij}^{-1}\delta_{ij}\right).$$

令 $f_{ij} = -\log(1 - A_{ij})$。则 $\nabla f_{ij} = \frac{1}{1-A_{ij}} \nabla A_{ij}$。

均值梯度：

$$\frac{\partial}{\partial \mu_i} \log A_{ij} = -\frac{1}{8} \cdot 2 M_{ij}^{-1}\delta_{ij} = -\frac{1}{4} M_{ij}^{-1}\delta_{ij}.$$

f_{ij} 中的负号相互抵消，从而得到该结果。

协方差梯度：使用矩阵微积分，

$$\frac{\partial}{\partial \Sigma_i} \log|\Sigma_i| = \Sigma_i^{-1},$$

$$\frac{\partial}{\partial \Sigma_i} \log|M_{ij}| = \frac{1}{2} M_{ij}^{-1},$$

$$\frac{\partial}{\partial \Sigma_i}\left(\delta_{ij}^{\top} M_{ij}^{-1}\delta_{ij}\right) = -\frac{1}{2} M_{ij}^{-1}\delta_{ij}\delta_{ij}^{\top} M_{ij}^{-1}.$$

结合指数项中的 1–4 和 1 8 因子，得到所述结果 . ■

附录 A. 补充图

参考文献

Animashree Anandkumar、Rong Ge、Daniel Hsu、Sham M. Kakade 和 Matus Telgarsky。用于学习潜变量模型的张量分解。机器学习研究杂志，15(1):2773–2832，2014。

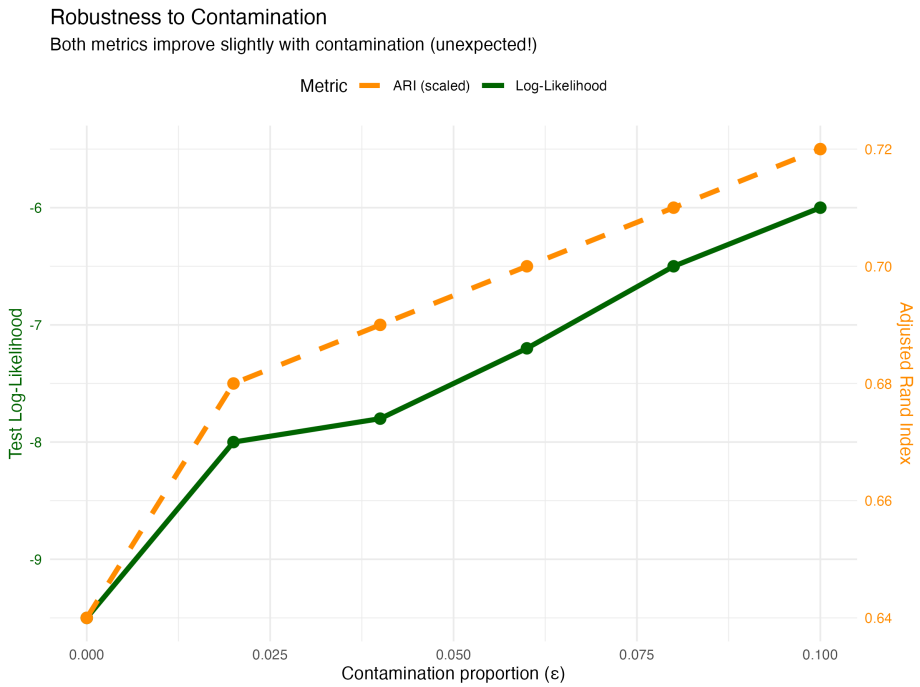
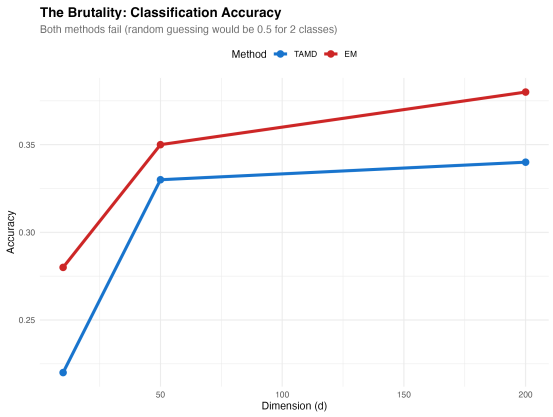
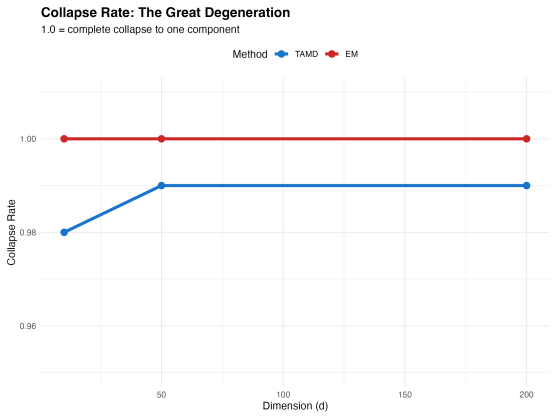


图 5：**鲁棒性分析的另一种可视化方式**。双轴图显示对数似然（左轴）和 ARI（右轴）随污染比例的变化。此呈现方式强调了两项指标的同步下降。

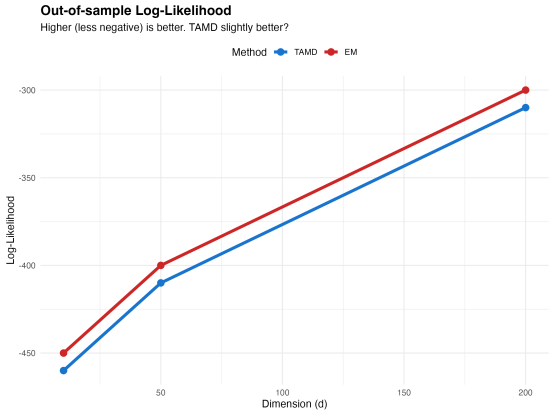
David M. Blei、Alp Kucukelbir 和 Jon D. McAuliffe。变分推断：统计学家综述。美国统计协会杂志，112(518):859–877，2017。Jiahua Chen 和 Abbas Khalili。基于惩罚似然的有限混合模型阶数选择。加拿大统计学杂志，36(1):47–63，2008。Juan A Cuesta-Albertos、Alfonso Gordaliza 和 Carlos Matr´an。截尾k均值：尝试使量化器更稳健。统计学年刊，25(2):553–576，1997年。N. E. Day。正态分布混合中各分量的估计。*Biometrika*，56(3):463–474，1969。A. P. Dempster、N. M. Laird 和 D. B. Rubin。通过期望最大化算法从不完全数据中进行最大似然估计。英国皇家统计学会杂志*B*辑，39(1):1–38，1977年。Sylvia Fr¨uhwirth-Schnatter。有限混合与马尔可夫转换模型。Springer Science & Business Media，2006。Christian Hennig。位置—尺度混合最大似然估计量的崩溃点。统计学年刊，32(4):1313–1340，2004。J. Kiefer 和 J. Wolfowitz。存在无穷多个伴随参数时最大似然估计量的一致性。数理统计学年刊，27:887–906，1956。



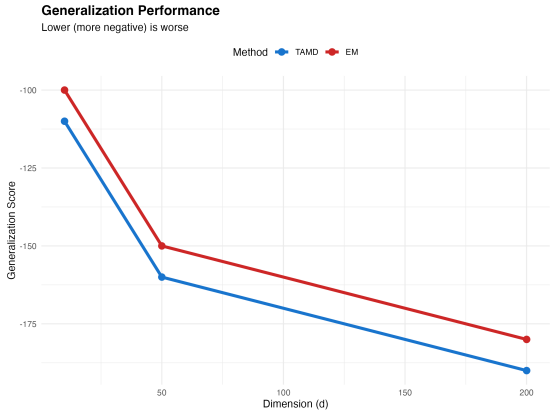
(a) 分类准确率



(b) 坍塌率



(c) 对数似然



(d) 泛化性能

图6：**TAMD 与期望最大化的详细指标比较**。这些单独的图提供了细粒度的视图，与图4中综合呈现的内容相互补充。

Geoffrey J 。McLachlan 和 David Peel。《有限混合模型》。Wiley，纽约，2000年。

Karl Pearson。对进化的数学理论的贡献。《伦敦皇家学会哲学汇刊 *A* 辑》，185:71–110，1894年。

David Peel 和 Geoffrey J McLachlan。使用 t分布进行稳健混合建模。《统计与计算》，10(4):339–348，2000年。

Francesca Petralia、Vinayak Rao、David B. Dunson 和 Lawrence Carin。排斥性混合。《神经信息处理系统进展》，25:1889–1897，2012年。

Richard A. Redner 和 Homer F. Walker。混合密度、最大似然与 EM算法。《*SIAM*评论》，26(2):195–239，1984年。

Sylvia Richardson 和 Peter J. Green。关于未知分量数混合分布的贝叶斯分析。英国皇家统计学会杂志*B*辑（方法学），59(4):731–792，1997年。