皇家天文学会月刊000，1–15
(2023)

预印本 2026年1月28日

使用 MNRAS LaTEX 样式文件 v3.0 编译

使用因子分析器高斯混合模型和基于模型估计重叠的合胞体聚类对完整 BATSE 伽马射线暴目录进行潜在特征描述

Fan Dai, [1]★ [2]† 和 Ranjan Maitra, 数学

科学系，密歇根理工大学，美国密歇根州霍顿市 *Townsend Dr 1400*号，邮编*49931*[2] 统计学系，爱荷华州立大学，美国爱荷华州艾姆斯市 *Osborn Drive 2438*号，邮编*50011-1090*

2024年11月11日录用。2024年11月11日收到；原稿提交于2024年9月15日

arXiv:2601.19140v1 [astro-ph.HE] 27 Jan 2026

摘要

对伽马射线暴（GRBs）进行表征和区分，多年来一直令天文学家感兴趣。虽然一些作者通过仅分析少数参数发现了两组或三组伽马射线暴，但近期研究在考虑九个参数 T_{50}、T_{90}、F_1、F_2、F_3、F_4、P_{64}、P_{256}、P_{1024} 后，识别出五个椭球形群组。另一些研究则提出，早先发现的两组或三组中还存在子类。我们使用高斯因子分析器混合模型，对当前爆发与瞬变源实验（BATSE）目录中观测到九个参数的1150个伽马射线暴进行了分析，并再次建立了用于描述伽马射线暴的五个椭球形群组。这五个群组根据其平均持续时间、通量和谱分别表征为较短-较暗-较硬、长-中等-中等、长-中等-中等、长-亮-中等和短-暗-硬。使用因子分析器描述各个群组的密度，可以基于少数潜在特征，对参数进行更深入的分组表征。然而，鉴于这一结果与许多主张存在两组或三组的其他研究之间存在差异，我们还进行了基于模型估计重叠的合胞体聚类（MOBSynC），该方法会逐步合并分离程度较低的群组。五个椭球形群组先合并为三组，随后又合并为两组：一组由持续时间较短的伽马射线暴组成，另一组则包含持续时间较长的伽马射线暴。这些群组同样可通过由九个参数构成的少数潜在因子进行表征。我们的分析为这三类结果提供了背景，并在此基础上详细说明了BATSE伽马射线暴的多层次表征，同时解释了其变异性的结构。

关键词：方法：统计 - 天文仪器、方法与技术，EM 算法、MixFAD、初始化、测量误差、多层特征描述、SynClustR

1 引言

天体物理学界长期以来一直对伽马射线暴（伽马射线暴）现象深感兴趣，它们是源自太空中大质量恒星的能量最高、强度最大的电磁辐射，因为人们认为它们能够为恒星形成以及早期宇宙的诞生提供洞见（Piran 1992；Piran 2005；Ioka 等 2016）。一般而言，关于伽马射线暴的研究主要集中于通过刻画这些已观测宇宙爆炸的性质与特征来揭示其起源，许多基础性结果表明伽马射线暴内部存在亚群体（Mazets 等 1981；Norris 等 1984；Dezalay 等 1992；Yang 等 2022；Steinhardt 等 2023；Zhu 等人 2024年）。Kouveliotou 等人（1993年）分析了从爆发与暂现源探测器（BATSE）1B目录中观测到的222个伽马射线暴，并发现持续时间变量的分布呈双峰性（具体而言，T_{90} 在对数$_{10}$ 尺度上）。参数 T_{90}，即90%的流量到达所需的时间，是伽马射线暴最显著的单一属性，并被用于将伽马射线暴分为两类：短持续时间（$T_{90} < 2s$）和长持续时间（$T_{90} \geq 2s$）。这种分类表明，伽马射线暴具有不同的天体物理起源：人们认为短时暴与中子星并合有关，可能是中子星彼此并合，也可能是中子星与黑洞并合（Nakar 2007；Berger 等人 2013年；Ghirlanda 等人 2017年）；而较长持续时间的暴则被推测与大质量恒星坍缩有关（Paczyński 1998；Woosley 与 Bloom 2006；Stanek等 2008；Pendleton 等人 1997年）。后续研究（Horváth1998）指出，BATSE 3B 数据中可能存在一个中等持续时间类别

★ 电子邮箱：fand@mtu.edu（FD）† 电子邮箱：maitra@iastate.edu（RM）

© 2023 作者

这一类别还得到后续研究的支持，这些研究提出持续时间分布由三个高斯分量构成（Horváth 2002；Hakkila 等人 2003；Horváth 等人 2004；Horváth 2009；Huja, D. 等人 2009；Tarnopolski 2015；Zitouni 等人 2015；Horváth 与 Tóth 2016）。Mukherjee 等人（1998）和 Chattopadhyay 等人（2007）通过多变量分析、非参数层次聚类以及 k-均值聚类，也支持存在第三个伽马射线暴组。他们的研究使用了三个观测参数和三个复合参数：观测参数为 T_{50}（即50%的流量到达所需的时间）、T_{90}、P_{256}（即256毫秒时间 bin 内的峰值流量）；三个复合参数则为总通量，即 F_t 四个时间积分通道通量之和（F_1、F_2、F_3、F_4），对应20–50、50–100、100–300和 >300 keV 能谱通道；$H_{32} = F_3/F_2$（即使用 F_2 和 F_3 计算的能谱硬度比）；以及 $H_{321} = F_3/(F_1 + F_2)$，即根据通道通量 F_1、F_2、F_3之比得出的谱硬度。在后续研究中，Chattopadhyay 与 Maitra（2017）也强调了在天体物理学中正确初始化并谨慎应用聚类算法的重要性，并证明，对 Mukherjee 等人（1998）和 Chattopadhyay 等人（2007）所用六个参数的分析表明，最优的高斯混合模型（GMM）基于模型的聚类（MBC）方案，在 BATSE 4Br 目录中得到五个椭球形分布的伽马射线暴群。进一步的详细研究（Chattopadhyay与Maitra（2018））显示，全部九个原始参数均直接参与聚类，并通过采用多变量-t 分布进行 MBC，进一步证实了存在五个椭球形群组；这种分布能够刻画重尾分布。（然而，Tóth 等人 2019年指出，这些群组可能只是此前广为人知的三类伽马射线暴组经过切分后形成的群组。）随后开展的其他分析（Berry & Maitra (2019); Almodovar-Rivera & Maitra 2020）证实了这些发现，即 BATSE 4Br 目录中存在五个不同的伽马射线暴组。尤其是，Berry & Maitra (2019) 展示了为何仅选取少数几个参数（实际上只使用持续时间 T_{90}）会得到两组或三组的解。事实上，后续研究（Tarnopolski 2022；Bhave 等人 2022；Tarnopolski 2019）仅使用持续时间和硬度参数进行分析，继续识别出两组；至于 BATSE 数据集中是否还存在第三组伽马射线暴，可能性仍不明朗。

完整的 BATSE 伽马射线暴目录可从 https://heasarc.gsfc.nasa.gov公开获取，其中包含康普顿伽马射线天文台（CGRO）上的 BATSE仪器探测到的伽马射线暴。目录收录了 BATSE 4B目录中的 1635 次伽马射线暴（触发编号 105 至 5586，观测时间为1991年4月19日至1996年8月29日），以及 BATSE 4B目录发布后触发的 1067 次暴发。这类记录共有 2702 条，其中只有 1971 条记录包含本文关注的九个参数。（BATSE 4B目录中曾收录的触发编号为 2463 和 2464 的伽马射线暴，现已不再包含于当前的完整目录中。）与早期版本的目录一样，其中一些参数被记录为零；BATSE伽马射线暴团队的 Charles A. Meegan 在Chattopadhyay 与 Maitra(2017年)的私人通信中指出，这些零值并非数值，而应视为缺失值。因此，在这 1971 次伽马射线暴中，有 373 次记录不完整（主要缺少通量参数）；剔除这些记录后，剩下 1598 次伽马射线暴，其全部九个参数均有观测值。

值得注意的是，这九个原始参数，即上文提到的 T_{50}、T_{90}、F_1、F_2、F_3、F_4，以及 P_{64}、P_{256}

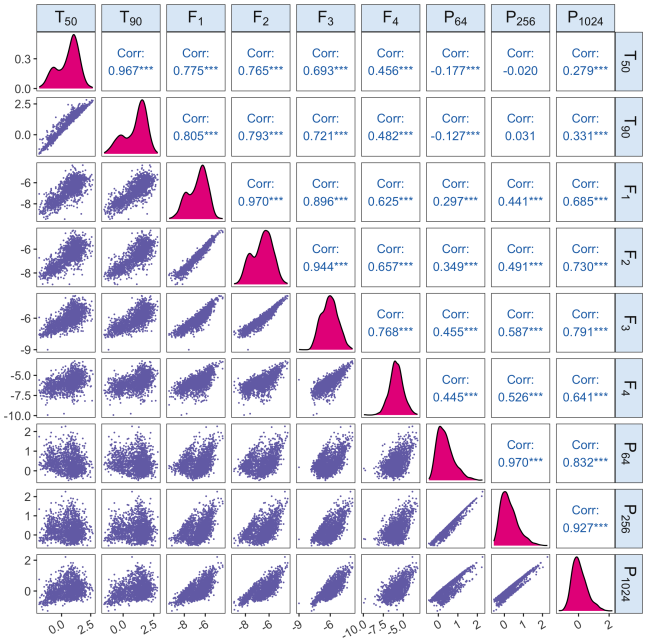
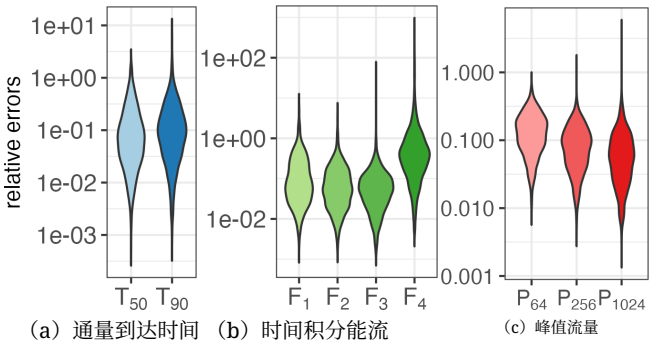


图1. 1598个完整BATSE目录伽马射线暴的九个特征的密度分布和散点图（经 $\log_{10}$ 转换后）。特征之间的相关性显示在上方面板中。



（a）通量到达时间 （b）时间积分能流 （c）峰值流量

相对误差的小提琴图（以 log 为尺度），涵盖全部九个参数：（a）两个通量到达时间；（b）四个时间积分流量；（c）三个峰值流量。数据来自1598个完整BATSE目录伽马射线暴。

并且，64、256 和 1024 毫秒时间区间内的三个峰值流量呈现严重的右偏，而为减轻各自偏度进行对数10变换后的结果则表现出高度相关性（图1）。此外，对这1598个伽马射线暴分别估计的测量误差也呈现严重偏态，其范围为：$0.001-27.456s$ 对于 T_{50}，$0.001-61.576s$ 对于 T_{90} $7.439 \times 10^{-10}-3.347 \times 10^{-7}$ $/cm^2$ F_1 $8.808 \times$，对于 $10^{-10}-2.932 \times 10^{-7}$ $/cm^2$ F_2 为 erg，对于为 erg，$2.74 \times 10^{-9}-1.159 \times$ 对于为 10^{-6} $/cm^2$ F_3 $2.05\times10^{-8}-5.474\times10^{-6}$ $/cm^2$ F_4 erg，对于为 erg，$0.158-2.63$ $/cm^2$ P_{64} $0.076-1.292$ $/cm^2$ 对于为 photons，对于为 photons P_{256}，以及 $0.035-1.29$ 对于 $/cm^2$ photons P_{1024}。图2展示了相对误差的分布（以对数10 尺度表示），表1列出了针对九个参数，各观测值大于 1σ、2σ 和 3σ 的伽马射线暴数量。为了使观测值相对于其测量误差不可忽略，同时确保聚类所需的样本量充足，我们将 2σ 确定为接受阈值，用于

两个通量到达时间，以及 1σ 四个时间积分通量和三个峰值流量的接受阈值；这使伽马射线暴数据集从1598条记录缩减至1150个暴，构成了我们分析所用的稀疏样本。（这些测量误差为观测参数提供了重要信息，理想情况下应纳入任何分析中；但目前我们尚无经过验证的统计方法可用于此，因此分析中仅考虑稀疏样本。）回到图1，我们注意到，尽管九个参数在对数$_{10}$ 尺度上的高度相关性可能让人觉得在分析中删除其中大部分参数是恰当且方便的，Chattopadhyay & Maitra (2018)证明，所有九个参数都包含聚类信息，因此都应纳入 MBC 解中。不过，也有可能存在少数潜变量，它们是这些参数的线性组合，且能更好地刻画这些参数的变异性。这些潜变量称为因子，我们将探究如何利用这些潜变量描述模型。

使用因子分析描述GRB的想法并非新鲜事。事实上，Bagoly 等人 (2009) 对197个长时暴 BATSE 伽马射线暴进行了因子分析，并证明这些暴中的九个参数可以用三个潜变量来解释。在本文的分析中，我们将因子结构纳入基于混合的模型，用于描述聚类中的每个簇。我们的研究结果还确定了五个椭球形群组，每个群组中的九个参数分别由五个因子，即五个潜变量来表征。我们在此研究的第二个方面是，几乎所有应用于GRB数据的聚类算法都假设并因此得出椭球形分布的簇。例如，Chattopadhyay 与 Maitra (2017年,2018); Berry 与 Maitra (2019); Almodovar-Rivera & Maitra (2020) 或本文中发现的五个簇，实际上可能是数量更少、形状更一般的簇的子类，正如文献中所推测的那样（Bhardwaj 等人 2023）。我们通过将Almodovar-Rivera & Maitra (2020) 的合生聚类框架改编到我们的高斯混合模型框架中，研究这种可能性，从而实现基于模型、基于重叠的聚类（MOBSynC）。事实上，我们发现，BATSE伽马射线暴可以通过多层次表征来刻画：在粗粒度层次上分为两个簇，而在细粒度层次上，这两个簇还可进一步划分为椭球形子群（第一个簇分为两个，第二个簇分为三个）。Chattopadhyay 等人(2022) 将这种多层次表征引入天体物理学应用，用于刻画热恒星系统，但此前尚未应用于GRB，也从未与因子分析器混合模型结合使用。

本文其余部分安排如下。第 2节详细说明如何构建因子分析器高斯混合模型（第 2.1.1节和2.1.2节），如何实施剖面似然最大化以估计协方差参数，以及如何利用重叠度量确定簇的合并（第 2.2节）。第 3 节将我们的方法应用于稀疏化完整BATSE目录中完全观测到的伽马射线暴，以识别用于伽马射线暴分类的最优组别，并找出潜在因子，从而展示伽马射线暴的原始特征。第4节总结全文并讨论后续工作。本文还附有附录，详细介绍高斯因子分析混合模型中快速参数估计的步骤。

2 聚类的统计框架

本节详细介绍本文所用聚类算法的统计学背景。该方法首先使用因子分析器高斯混合模型对数据进行 MBC 聚类，同时开发快速计算方法来估计参数。根据初步聚类结果，我们调整了 Chattopadhyay 等人(2022) 中描述的流程，根据广义重叠度量和成对重叠度量进一步合并已识别的簇。我们的方法通过 `Mix-FAD`(数据因子分析器混合模型)程序和 `MOB-SynC`函数实现（可在 https://github.com/fanstats/MixFAD-GRB获取）；两者均使用开源统计软件R语言(R 核心团队 2024)及其相关软件包编写。

2.1 基于高斯混合因子分析器的聚类

2.1.1 基于高斯混合模型的聚类

基于模型的聚类方法（Anderson, 2003; McLachlan与Peel, 2000; Mardia等, 2006; Melnykov与Maitra, 2010）为在无监督情境下对观测值进行分组提供了一种有原则且形式化的方法。该方法假设数据源自有限混合模型，即由若干成分分布构成的混合，例如多元高斯分布；在这种情况下，我们得到高斯混合模型（Chattopadhyay与Maitra, 2017）。形式上，我们的建模与估计框架假设有 p维观测值$\boldsymbol{x}_i$, $i = 1, 2, \ldots\ldots, n$来自一个包含 K 个成分的高斯混合模型，其中 $\boldsymbol{x}_i$以概率 η_k k =由第个成分生成，$k = 1, 2, \ldots, K$且，。那么，对数似然函数为（Day 1969）

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}) = \sum_{i=1}^{n} \log \Big\{ \sum_{k=1}^{K} \eta_k \boldsymbol{\phi}_(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\}, \tag{1}$$

对于数据矩阵 $\boldsymbol{X} = [\boldsymbol{x}_1^\top, \boldsymbol{x}_2^\top, \ldots, \boldsymbol{x}_n^\top]^\top$、参数空间 $\boldsymbol{\Theta} = (\eta_1, \eta_2, \ldots, \eta_K, \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \ldots, \boldsymbol{\mu}_K, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \ldots, \boldsymbol{\Sigma}_K)$，以及第 k 个混合成分的高斯概率密度函数（PDF） $\boldsymbol{\phi}(\cdot)$，通过优化式（1）进行参数估计几乎总是难以实现，但期望最大化（EM）算法（Dempster等，1977；Rubin与Thayer，1982；McLachlan与Krishnan，2008）使之成为可能；具体而言，该算法为每个 $\boldsymbol{x}_i$ 引入一个潜在组指示变量 z_i，使得给定 $z_i = k$ 时 $\boldsymbol{x}_i$ 的条件分布为 $\mathcal{N}_k(\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)$，且 $\mathbb{P}\mathrm{r}(z_i = k) = \eta_k$。于是，对于 $(\boldsymbol{X}, \boldsymbol{Z})$，其中 $\boldsymbol{Z} = (z_1, z_2, \ldots, z_n)$，并用 $\mathbb{1}(\cdot)$ 表示指示函数，我们可以得到完整数据（包括观测数据和增广数据）的对数似然函数（McLachlan与Peel，2000）：

$$\ell(\boldsymbol{\Theta}; \boldsymbol{X}, \boldsymbol{Z}) = \sum_{i=1}^{n} \sum_{k=1}^{K} \mathbb{1}(z_i = k) \Big\{ \log \eta_k + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k) \Big\}. \tag{2}$$

EM 算法从 $\boldsymbol{\Theta}$ 的一些初始值开始，然后在 E 步（即期望步）和 M 步（即最大化步）之间迭代，直到更新收敛到一个解，并证明该解是最大似然估计。具体而言，E 步和 M 步如下：

E 步骤： 对于未知项$\mathbb{1}(z_i = k)$ 在（2）中，计算其条件期望 $\mathbb{E}[\mathbb{1}(z_i = k)|\mathbf{X}; \boldsymbol{\Theta}]$，并求得期望的

表1。根据九个参数，观测值高于接受水平的伽马射线暴数量。

接受阈值 \ 特征	T_{50}	T_{90}	F_1	F_2	F_3	F_4	P_{64}	P_{256}	P_{1024}
$\mathbf{1\sigma}$	1577	1571	1555	1584	1593	1276	1598	1597	1597
$\mathbf{2\sigma}$	1519	1501	1467	1563	1581	992	1587	1597	1594
$\mathbf{3\sigma}$	1425	1387	1351	1512	1558	777	1521	1587	1583

完整对数似然（所谓的 Q-函数）为

$$\mathrm{Q}(\boldsymbol{\Theta}^*;\mathbf{X},\boldsymbol{\Theta}) = \sum_{i=1}^{n}\sum_{k=1}^{K} \mathrm{E}[\mathbb{1}(z_i=k)|\mathbf{X};\boldsymbol{\Theta}]\Big\{\log\eta_k + \log\phi(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)\Big\}, \tag{3}$$

其中

$$\mathbb{E}[\mathbb{1}(z_i=k)|\mathbf{X};\boldsymbol{\Theta}] = \frac{\eta_k\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K}\eta_k\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Sigma}_k)}. \tag{4}$$

M 步： 通过最大化 $\boldsymbol{\Theta}^*$ 来计算Q-函数（3）的更新估计值。具体而言，我们得到

$$\begin{aligned}
\widehat{\eta}_k^* &= \frac{n_k}{n},\\
\widehat{\boldsymbol{\mu}}_k^* &= \frac{\sum_{i=1}^{n}\mathbb{P}\mathrm{r}(z_i=k|\boldsymbol{x}_i;\boldsymbol{\Theta})\boldsymbol{x}_i}{n_k},\\
\widehat{\boldsymbol{\Sigma}}_k^* &= \frac{\sum_{i=1}^{n}\mathbb{P}\mathrm{r}(z_i=k|\boldsymbol{x}_i;\boldsymbol{\Theta})(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)(\boldsymbol{x}_i-\boldsymbol{\mu}_k^*)^\top}{n_k},
\end{aligned} \tag{5}$$

其中 $n_k = \sum_{i=1}^{n}\mathbb{E}[\mathbb{1}(z_i=k)|\mathbf{X};\ \boldsymbol{\Theta}]$。

正如Maitra(2009年）以及Chattopad-hyay & Maitra(2017年）所讨论和展示的，初始化是影响 EM 算法优化性能的重要因素，因为算法会收敛到初始值附近的一个（局部）最大值。我们采用 Maitra (2013年）或Goren &Maitra (2022）的随机初始化方法：从 m_1随机选取的初始值开始运行 EM 算法一段时间，然后从中选取 m_2 经过几次迭代后观测对数似然函数值最高的候选值，再用这些候选值最终初始化 EM 算法，并使其运行至收敛。最优初始化是指在 m_2 次运行中，最终得到的观测对数似然值最大的那一次。在第3节对伽马射线暴的分析中，我们使用了 $m_1 = 1000,\ m_2 = 10$。

通过获得 Θ 的 EM 估计最大似然估计量，我们将每个观测值 $\boldsymbol{x}_i$ 归入使 k （4）取最大值的簇。既已介绍 GMM 和 MBC，接下来我们将上述框架扩展到高斯混合因子分析器框架。

2.1.2 高斯混合因子分析器 *MBC*

因子模型可以用少数几个（p）潜在因子来解释一个$q < p$维数据集中的变异性。一般而言，针对的因子模型（Thurstone 1931, 1935；Anderson, 2003） $\boldsymbol{x}_i$ 可以表示为以下线性方程

$$\boldsymbol{x}_i = \boldsymbol{\mu} + \boldsymbol{\Lambda}\boldsymbol{F}_i + \boldsymbol{\epsilon}_i, \tag{6}$$

其中， $\boldsymbol{F}_i \sim \mathcal{N}(\mathbf{0},\ \boldsymbol{I}_q)$表示 q潜在因子，而 $\boldsymbol{\epsilon}_i \sim\mathcal{N}(\mathbf{0},\ \boldsymbol{\Psi})$则表示所谓的唯一性矩阵 $\boldsymbol{\Psi}$，该矩阵为对角矩阵，存储各个变量所特有的方差。此外， $\boldsymbol{\Lambda}$ 是一个 $p\times q$ 因子载荷矩阵，用于解释 k在 $\boldsymbol{x}_i$中的特征，即由第 j个潜变量或因子对其贡献的权重或比例，其中 $k = 1$、2、.....、 p，且$j = 1$、2、.....、 q。为确保模型的可识别性，同时由于$\boldsymbol{\Sigma}$的非奇异性，我们约束 $q < \min(n,\ p)$以及$(p-q)^2 > p+q$。根据构造， $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$，其中 $\boldsymbol{\Sigma} = \boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top + \boldsymbol{\Psi}$。如前所述，Bagoly 等人 (2009) 将因子分析引入天体物理学，用于刻画BATSE目录中的197个长暴伽马射线暴。不同于主成分（Bagoly et al. 1998），主成分为数据提供了低维表示，因此假设一个奇异的 $\boldsymbol{\Sigma}$ （秩等于主成分的数量）；因子模型具有非奇异的 $\boldsymbol{\Sigma}$，其所需参数数量更少（$pq+q$），少于无约束的 $\boldsymbol{\Sigma}$，后者具有 $p(p+1)/2$ 个参数。

在聚类情境下，我们将因子模型扩展为表示各个组。具体而言，对于第2.1.1 节定义的 GMM，我们通过将每个高斯分量 ϕ（$\boldsymbol{x}$；$\boldsymbol{\mu}_k$，$\boldsymbol{\Sigma}_k$）表示为具有共同q 因子模型的形式，从而推广因子模型的定义。由此，我们得到高斯混合因子分析器（McLachlan& Peel 2000），其中第 k 个组的协方差矩阵具有受约束的结构 $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k$，其中 $\boldsymbol{\Lambda}_k$ 是所有分量具有共同秩 q的因子载荷矩阵。因此，各组的因子模型为

$$\boldsymbol{x}_i = \boldsymbol{\mu}_k + \boldsymbol{\Lambda}_k\boldsymbol{F}_{ik} + \boldsymbol{\epsilon}_{ik}, \quad \text{when } z_i = k, \tag{7}$$

其中 $\boldsymbol{F}_{ik} \sim \mathcal{N}(\mathbf{0},\boldsymbol{I}_q)$和 $\boldsymbol{\epsilon}_{ik} \sim \mathcal{N}(\mathbf{0},\boldsymbol{\Psi}_k)$。因此，对于包含因子模型成分的高斯混合模型，我们的设定如（1）所示，进一步将每个成分中的 $\boldsymbol{\Sigma}_k$ 分解为 $\boldsymbol{\Sigma}_k = \boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k$，从而得到观测数据的对数似然函数

$$\ell(\boldsymbol{\Theta};\boldsymbol{X}) = \sum_{i=1}^{n}\log\Big\{\sum_{k=1}^{K}\eta_k\phi_(\boldsymbol{x}_i;\boldsymbol{\mu}_k,\boldsymbol{\Lambda}_k\boldsymbol{\Lambda}_k^\top + \boldsymbol{\Psi}_k)\Big\}, \tag{8}$$

表示式（7）在估计时存在可识别性问题：具体而言，对于任意正交矩阵 $\mathbf{T}_k$，因子载荷矩阵$\boldsymbol{\Lambda}_k$ 及其正交变换 $\boldsymbol{\Lambda}_k\mathbf{T}_k$ 会得到相同的似然值，因此需要对参数施加约束，以确保 $\boldsymbol{\Lambda}_k$有唯一解。常见做法是采用我们这里使用的尺度不变约束，使所谓的 信号矩阵 $\boldsymbol{\Lambda}_k^\top\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k$ 为对角矩阵，且对角元素按递减顺序排列。对于给定的 $\boldsymbol{\Lambda}_k$ ，可以将该约束剖面化消去，从而便于计算（Lawley 1940；Dai 等人 2020）。

最大似然参数估计可以像第 2.1.1 节所述，通过采用 EM 算法实现。E 步骤基本与该节所述相同，而 M 步骤除少量修改外，对于 η_k 和 $\boldsymbol{\mu}_k$ 也大体相同。然而，要高效地估计 $\boldsymbol{\Lambda}_k$，$\boldsymbol{\Psi}_k$ 则需要特别谨慎处理；一些技术细节将在附录中说明。推导最终得到数据因子分析器混合（MixFAD）算法，我们已将其实现于同名的 R语言（R 核心团队 2024）软件包中。

2.1.3 混合成分和潜在因子数量的选择

前几节的推导针对的是固定数量的混合成分 K 和固定数量的潜在因子 q；为方便起见，我们假设所有混合成分的因子数量相同。我们采用 Schwarz (1978) 的贝叶斯信息准则（BIC）：$-2\widehat{l}+\nu\log(n)$，其中 $\widehat{l}$ 是根据拟合模型计算出的观测数据对数似然，而 ν 是模型参数的总数。由于较小的 BIC 值意味着较大的对数似然，而较低的惩罚项对应较简单的模型，因此模型对数据的拟合效果更好，我们选择使 BIC 最小的 (K, q) 对。根据这些最优值，我们得到相应的参数估计，然后为每个观测值分配组别归属。

本节介绍了一种方法，该方法既能识别椭球形簇，也能通过少数组级潜在因子描述这些簇的变异性。第3节将说明如何运用此方法，厘清经过抽样稀疏化且完全观测到的 BATSE目录伽马射线暴中的椭球形簇。接下来，我们将讨论如何寻找更具复杂特征的群组来解释这些伽马射线暴。

2.2 基于模型估计重叠的合胞体聚类

MBC 等聚类算法建立在混合成分与群组一一对应的前提之上。（例如，在高斯混合模型中，这会产生椭球形群组。）这类假设未必总是合理；事实上，其中一些簇可能只是更大、结构更宽泛的群组中的亚组。因此，一些作者（Baudry 等人，2010；Hennig，2010；Melnykov，2016；Peterson 等人，2018；Almodovar-Rivera与Maitra（2020））提出了 Almodovar-Rivera与Maitra(2020)所命名的合胞体聚类方法，将分离度较低的群组合并为更大的超簇。Almodovar-Rivera与Maitra(2020)开发了一种数据驱动算法，可合并k均值聚类算法的输出群组。后续研究（Chattopadhyay 等人，2022）将该方法扩展至 MBC；本文也采用并改进了这一方法，用于伽马射线暴的分类。我们的方法首先使用 MixFAD 得到具有因子结构协方差矩阵的高斯分布簇，亦即简单群组（按照 Almodovar-Rivera与Maitra(2020)的定义）；随后使用成对重叠度和广义重叠度（Maitra & Melnykov，2010；Melnykov & Maitra，2011；Melnykov 等人，2012），依次合并 MixFAD 得到的分离度较低的群组，从而形成复合或组合群组。下文将概述这一算法。

2.2.1 简单簇之间的重叠度量

假设初始聚类产生了 K 个简单组。那么，数据向量 $\boldsymbol{x}_i|z_i = k_1$ 被误分类到第 k_2个簇的概率由 Maitra & Melnykov (2010) 给出。

$$\omega_{k_2|k_1} = \mathbb{P}\mathrm{r}\left(\frac{\eta_{k_2}\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_{k_2},\boldsymbol{\Sigma}_{k_2})}{\eta_{k_1}\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_{k_1},\boldsymbol{\Sigma}_{k_1})} > 1\right). \tag{9}$$

接下来，第 k_1个簇与第 k_2个簇之间的成对重叠度为$\omega_{k_1,k_2} = \omega_{k_1|k_2} + \omega_{k_2|k_1}$ (Maitra & Melnykov 2010)。成对重叠度量的取值范围为 0 至 1，可反映两个估计簇之间的差异程度。具体而言，重叠值越小，簇间分离程度越高，反之亦然；重叠值为 1 表示两组之间毫无差异。此外，我们可以得到一个广义重叠度 $\ddot{\omega}$，它由成对 $K \times K$ 重叠度的**矩阵** $\boldsymbol{\Omega}$ 计算得出 $\ddot{\omega} = (\lambda^*_{\boldsymbol{\Omega}} - 1)/(K-1)$，其中 $\lambda^*_{\boldsymbol{\Omega}}$ 是$\boldsymbol{\Omega}$ 的最大特征值(Melnykov & Maitra 2011)，此方法借鉴了Maitra (2010)的思路。广义重叠指数用一个指标衡量整体聚类复杂度，数值越小，K 各组之间的整体分离程度越高。针对高斯混合模型（GMM），这些重叠度量的解析计算已在 R语言（R 核心团队 2024）*MixSim* 软件包（Melnykov 等人 2012）中实现。

2.2.2 复合簇之间的重叠度量

复合簇（或组合簇）由合并简单簇得到，数据向量 $\boldsymbol{x}_i$ 来自第 m_1 个复合簇 $\mathcal{C}_{m_1}$，被误分类到第 m_2 个复合簇 $\mathcal{C}_{m_2}$ 的概率在本例中为

$$\omega_{m_2|m_1} = \mathbb{P}\mathrm{r}\left(\frac{\sum_{j\in\mathcal{C}_{m_2}}\eta_j\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{l\in\mathcal{C}_{m_1}}\eta_l\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)} > 1\right), \tag{10}$$

这可以像Chattopadhyay 等人 (2022) 一样，通过蒙特卡罗方法近似计算如下：

(i) 从第 m_1 个复合簇 $\mathcal{C}_{m_1}$ 的混合分布中生成 M（在第 节中设为 100,000 3 $\boldsymbol{x}_i$ $i=$）个 ,,,, 样本，$1,2,\ldots,M$ 该混合分布定义为 $\sum_{l\in\mathcal{C}_{m_1}}\eta_l^*\phi(\cdot;\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)$，其中 $\eta_l^* = \eta_l/\sum_{h\in\mathcal{C}_{m_1}}\eta_h$。

(ii) 计算两个复合簇 $\mathcal{C}_{m_1}$ 和 $\mathcal{C}_{m_2}$ 之间的成对重叠度，记为 $\omega_{m_1,\,m_2} = \omega_{m_1|m_2} + \omega_{m_2|m_1}$，其中我们估计

$$\widehat{\omega}_{m_2|m_1} = \frac{1}{M}\sum_{i=1}^{M}\mathbb{1}\left\{\frac{\sum_{j\in\mathcal{C}_{m_2}}\eta_j\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_j,\boldsymbol{\Sigma}_j)}{\sum_{l\in\mathcal{C}_{m_1}}\eta_l\boldsymbol{\phi}(\boldsymbol{x}_i;\boldsymbol{\mu}_l,\boldsymbol{\Sigma}_l)} > 1\right\}. \tag{11}$$

因此，可以使用这些两两重叠度，以与第2.2.1节中所述类似的方式计算广义重叠度。

2.2.3 *MOBSynC* 过程

现在给出用于合并 MBC 得到的 K 组的具体 MOBSynC 过程：

(i) 对 $K(K-1)/2$ 对简单簇，计算成对重叠度$\omega_{k_1,\,k_2}$ 以及广义重叠度 $\ddot{\omega}$，具体参见第2.2.1节。

(ii) 若满足 k_1第和 k_2个簇，则将其合并$\omega_{k_1,\,k_2} > \kappa\ddot{\omega}$，其中参数 κ 表示合并阻力，值越大，表示在此步骤中合并的簇对越少。参数 κ的选择将在注释1中讨论。

（iii）对于合并步骤（ii）得到的复合簇，计算所有两两重叠 $\omega_{k_1,\,k_2}$ 以及广义重叠度$\ddot{\omega}$，具体参见第2.2.2节，然后重复步骤（ii）。

（iv）如果当前阶段的广义重叠度，或其相较于上一阶段的变化，小于 ϵ，则停止合并。通常将其设为一个较小的值，此处设为 10^{-3}。

注释 1. Almodovar-Rivera & Maitra (2020) 提供了一种数据驱动的方法：通过使用 $\kappa = 1$、2、……、κ_{max}运行合并过程，确定 κ 的整数值，其中 $\omega_{k_1,\,k_2} < \kappa_{max}\ddot{\omega}$ 适用于所有成对簇。选择能使最终划分的广义重叠度最小的 κ 作为最优值。

注释 2. 算法中的步骤 (iv) 是 Almodovar-Rivera &Maitra (2020) 用来确定一般结构群组数量的步骤。在我们的语境中，我们尤其关注的并非一般形状群组的数量，而是运用 MOBSynC 描述 BATSE伽马射线暴的层次子群结构，从而为近期研究中反复发现的椭球形群组提供更多背景。

2.2.3.1 说明性示例 我们通过一个模拟示例来说明 MOBSynC：从一个五分量高斯混合模型中随机生成一个包含 $n = 1000$ 个观测值的三维数据集。该模型包含两个彼此分离的复合组，分别呈“U”形和“V”形，如图3a中以深浅不同的颜色所示。“U”形和“V”形的各个部分都可以用单独的椭球簇拟合；图3a中用颜色（深色或浅色）表示这五个簇。我们使用最多包含 9 个分量的高斯混合模型拟合数据，并根据贝叶斯信息准则得出最佳分量数为 $K = 5$。拟合结果见图 3b，其中高斯混合模型识别出了这五个独立的椭球簇：这五个简单簇中有 19 个观测值被错误分类，分类错误率为 0.019。

我们使用 MOBSynC 来考察这五个椭球形（高斯）组是否属于某些复合簇。图3c 展示了根据对数据拟合高斯混合模型而得到的五个简单簇之间计算出的两两重叠图：广义重叠度计算结果为 $\ddot{\omega} = 0.011$。根据备注1，我们确定了 $\kappa = 1$ 并得到$\omega_{1,\,2}$，$\omega_{1,\,3}$，$\omega_{4,\,5} > \kappa\ddot{\omega}$，这表明簇 1、2 和 3 合并，簇 4 和 5 也合并。随后，我们计算了复合簇（ 1, 2, 3 4, 5 $\ddot{\omega} < 0.001$ ）和（ ）的重叠度量，得到的广义重叠度为，因此我们在当前阶段停止合并，最终得到图3d所示的两个复合簇，其中我们对复合簇进行了映射，以便更清晰地展示。我们可以看到，合并过程能够忠实地描述原始数据集的多层分组特征。

本节概述了 MixFAD 和 MOBSynC 所需的方法。现在，我们将其应用于稀疏化完整BATSE目录中观测到的伽马射线暴。

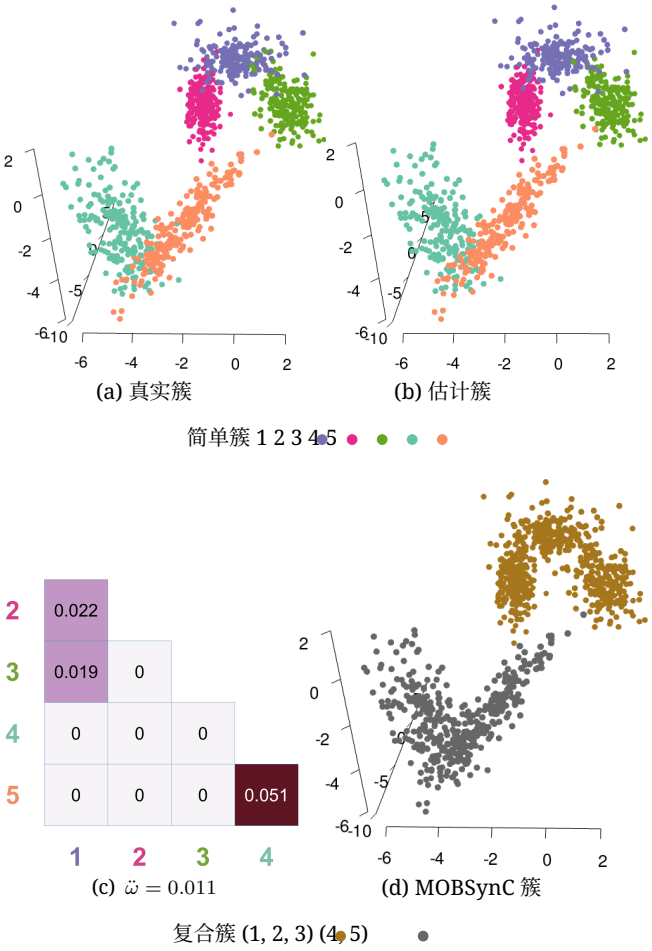


图 3. 模拟数据集：(a) 三维散点图展示真实的聚类结构，不同颜色分别表示椭球形高斯簇，较浅和较深的颜色则分别表示复杂形状的“V”簇和类似的“U”簇；(b) GMMBC 得出的估计簇，识别出五个椭球形群组，并显示 (c) 成对重叠度；(d) MOBSynC 得出的复合簇，重新着色以突出算法所揭示的复合结构和组合结构。

3 伽马射线暴的统计分析

3.1 使用 MixFAD 进行简单椭球簇分析

在初始聚类阶段，我们对稀疏化完整BATSE伽马射线暴目录中的1150条记录拟合 MixFAD 算法，最多使用七个混合成分和五个因子（可拟合的最大数量q 为 $p = 9$）。根据贝叶斯信息准则确定的最优簇数和因子数为 $K = 5$，$q = 5$（见图 4）。

我们研究了所识别群组的性质，并根据九个特征，探讨各个簇中伽马射线暴的天体物理特征。表2给出了因子分析器高斯混合模型得到的五个简单簇各自的样本量。可以看出，第二个簇包含的伽马射线暴数量最多，而第一个和第五个簇包含的暴最少。表3 总结了各个簇中九个参数的均值和标准差。可以看出，两个持续时间参数体现了簇间最显著的差异：两个规模最小的简单组

表 2.五个简单簇中各自的伽马射线暴数量。

簇	1	2	3	4	5
Size	156	392	235	248	119

的 $\log_{10} T_{50}$ 和 $\log_{10} T_{90}$ 平均值低于其他三个简单簇。四个时间积分流量参数也呈现出类似的模式。对于在三个时间段内测得的峰值流量，我们发现第四组的平均水平明显最高，而第三簇的平均水平最低。另一方面，对于 $\log_{10} P_{1024}$，尽管第五簇在 64 和 256 毫秒时间段内的峰值流量仍相对较高，其平均值也较小。正是这些差异界定了在经过稀疏化处理的完整 BATSE目录中，由 Mix-FAD算法识别出的五个简单组。

简单*GRB*组中的潜在结构特征

正如Bagoly 等人(2009针对 GRB 数据所展示的那样，因子分析的一大优势在于，它能够用少数几个参数的线性组合来解释数据集中的变

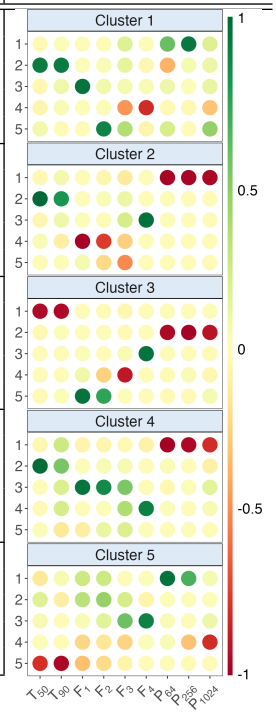
表3。估计　各簇中九个参数（经过 $\log_{10}$ 转换后）的组均值和标准差（括号内）　n).

簇 \ 特征	$\log_{10} T_{50}$	$\log_{10} T_{90}$	$\log_{10} F_1$	$\log_{10} F_2$	$\log_{10} F_3$	$\log_{10} F_4$	$\log_{10} P_{64}$	$\log_{10} P_{256}$	$\log_{10} P_{1024}$
1	−0.75(0。36)	−0.39(0.35)	−7.84(0.38)	−7.58(0.38)	−6.74(0.34)	−6.21(0.51)	0.53(0.33)	0。35(0。31)	−0.07(0.33)
2	1。17(0。45)	1.64(0.39)	−6.11(0.39)	−5.99(0.39)	−5.55(0。41)	−5.46(0。52)	0.29(0.24)	0.24(0.26)	0。19(0。27)
3	1.14(0.56)	1.49(0.50)	−6.69(0。54)	−6.60(0.48)	−6.17(0.46)	−5.68(0.44)	0.04(0.24)	−0.07(0.26)	−0.19(0.26)
4	0.87(0.48)	1。43(0。49)	−5.84(0。53)	−5.68(0。52)	−5.16(0.57)	−5.00(0.77)	0.93(0.40)	0.89(0.41)	0.79(0.42)
5	−0.08(0.68)	0.46(0.69)	−7.32(0.68)	−7.14(0.66)	−6.44(0。62)	−5.89(0.63)	0.56(0.43)	0.39(0.45)	0.07(0.48)

表4。 估计的因子载荷（以相关尺度表示），并附有热图，便于快速查阅。为使呈现更清晰，表中省略了区间 (-0.1,0.1) 内的数值，但在热图中以浅色显示。

k	q	$\log_{10} T_{50}$	$\log_{10} T_{90}$	$\log_{10} F_1$	$\log_{10} F_2$	$\log_{10} F_3$	$\log_{10} F_4$	$\log_{10} P_{64}$	$\log_{10} P_{256}$	$\log_{10} P_{1024}$
1	1					0.183		0.615	0.970	0.187
	2	0.952	0.953			0.144		-0.381		0.111
	3			0.992						0.129
	4					-0.481	-0.834			-0.305
	5				0.916	0.370		0.218		0.479
2	1					-0.124		-0.991	-0.992	-0.988
	2	0.996	0.842			0.112				
	3					0.222	0.991			
	4			-0.993	-0.776	-0.262				
	5				-0.217	-0.516				
3	1	-0.979	-0.984							
	2							-0.986	-0.995	-0.937
	3						0.993			
	4				-0.252	-0.914				
	5			0.992	0.779					
4	1		0.267					-0.994	-0.985	-0.831
	2	0.994	0.566							
	3		0.205	0.986	0.891	0.560				0.159
	4		0.224			0.379	0.936			
	5		-0.138	-0.105	0.122	0.194				
5	1	-0.139		0.268	0.264			0.995	0.702	
	2	0.178		0.336	0.471	0.225				
	3				0.161	0.607	0.926			0.105
	4			-0.229	-0.150	-0.238			-0.314	-0.830
	5	-0.822	-0.987	-0.327	-0.208					

热图表示



异性。Bagoly 等人(2009) 仅对 197 个持续时间较长的 BATSE伽马射线暴进行了分析。我们的方法使用 MixFAD 在进行聚类分析的同时，对所得分组中的变异结构进行综合表征，这是其一大优势。现在我们将讨论 MixFAD 得出的结果，这些结果有助于进一步表征简单组，超出表 3 中已讨论的均值和标准差所能提供的信息。为清楚起见，我们再次强调，此处讨论的所有参数均经过对数$_{10}$ 变换。我们再次指出，正如稍后在 (7) 后不久所讨论的，由于正交变换的不变性，因子载荷的正负号并不重要；但是，两个参数对因子载荷的贡献符号相同还是相反，则很重要。

对于估计得到的因子模型，表4以数值形式展示了斜交旋转后的拟合因子载荷，并为便于查阅，同时以图示呈现（Costello和Osborne（2005）），以简化对五个简单簇的解读。这些因子载荷是 k组 $p \times q$ 在（$\mathbf{\Lambda}_k$）中的7矩阵的五列。五个因子所解释的总方差比例依次为：对于编号为1至5的五个簇，分别为74.6%、87.9%、92.5% 83.2% 和67.6%，这表明因子能够相当充分地概括每个估计得到的简单伽马射线暴组。此外，我们在表4 中仅列出载荷中不可忽略的数值，也就是说，这些数值均落在区间$(-0.1, 0.1)$之外。现在，我们将分别讨论并区分每个简单组所得到的五个因子载荷。

对于第一个簇，我们看到，第一个因子是 F_3 的加权平均值，其中 F_3 和 P_{1024}、 P_{64}及 P_{256} 依次贡献较小、较大和显著；第二个因子主要由两个通量到达时间解释，同时也有少量 F_3 和 P_{1024}的贡献，以及少量方向相反（负号）的 P_{64} 贡献。第三个因子载荷向量主要由20–50千电子伏能段的时间积分能流（F_1）决定，另有少量 P_{1024}。第四个因子载荷向量是 P_{1024}、F_3和 F_4 依次贡献较小、中等和显著的线性组合。最后，该组的最后一个因子主要由 F_2 决定，同时P_{1024} 贡献中等， F_3 和 P_{64}贡献较小。

第二个簇中的第一个因子几乎完全是三个峰值流量的简单平均，并包含少量 F_3；第二个因子则基本上是两个通量到达时间的加权平均，其中 T_{50} 比 T_{90}略占主导， F_3的贡献较小。对于这个第二组，第三个

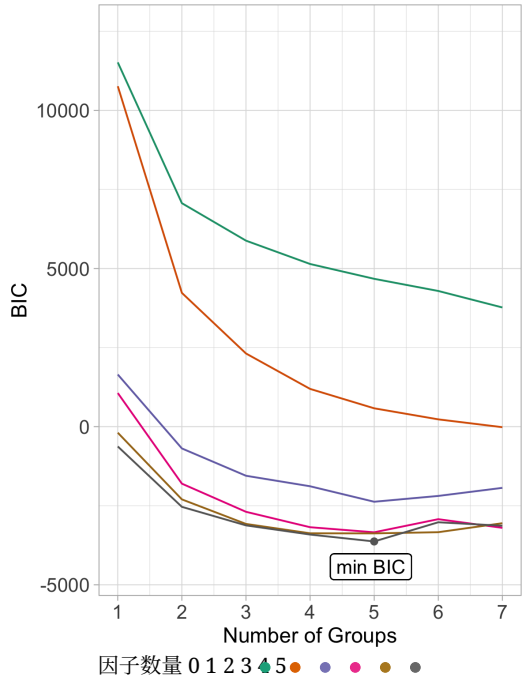


图 4. 使用 1150 个稀疏化后的完整 BATSE目录伽马射线暴拟合因子分析器高斯混合模型时，每个 K和 q对应的贝叶斯信息准则。最小贝叶斯信息准则出现在（K，q）=（5，5）。

因子载荷向量基本上由 F_4 构成，另有少量F_3；第四个因子则是前三个时间积分流量的加权平均，其贡献分别从极大到显著再到较小。最后，该簇的第五个因子是 F_2 和 F_3 的加权平均，其中第三种时间积分流量的贡献幅度是另一种的两倍多。

第三个簇的第一个因子载荷向量是一个加权平均，主要由两个通量到达时间决定。第二个载荷向量同样是加权平均，主要由三个峰值流量构成。第三个因子本质上是 F_4。第四个因子是加权平均，F_3和 F_2的贡献依次从非常高到较低。最后，第五个因子载荷向量是 F_1 和 F_2的加权平均，前者的贡献比后者约高 30%。

第四个组的第一个因子载荷向量体现了少量 T_{90} 与三个峰值流量的加权平均之间的对比（P_{64}、P_{256}的贡献非常高，P_{1024}的贡献较低）。第二个因子载荷向量主要由两个通量到达时间决定，其中 T_{50} 的贡献几乎是 T_{90}的两倍；第三个因子则是少量 T_{90} 与前三个时间积分通量的加权平均（贡献依次从非常高到较高再到中等）。第四个因子载荷向量主要由 F_4解释，另有少量 T_{90} 和 F_3。最后，第五个因子体现了两组加权平均之间的对比：一组为 T_{90} 和 F_1，另一组为 F_2 和 F_3（不过所有这些变量的贡献都相当小）。

我们的第五个也是最后一个简单簇，其第一载荷向量体现了少量 T_{50}与加权平均值之间的对比；后者从相反方向依次包含来自 F_1和 F_2、P_{256} 和 P_{64} 的少量、显著和极大量贡献，而第二因子则是少量 T_{50} 与前三个时间积分流量的线性组合，其中 F_2 的贡献大于 F_1和 F_3。第三因子是一个加权平均值，

表5。五个简单簇的平均因子得分。

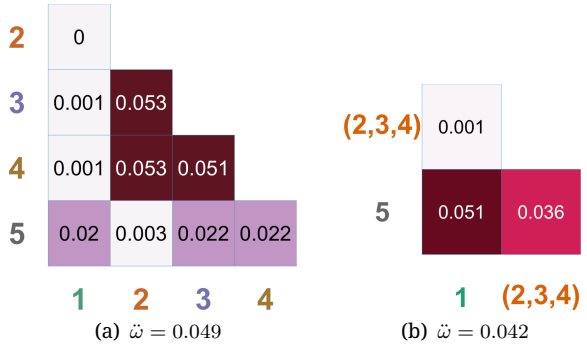
簇	因子1	因子2	因子3	因子4	因子5
1	0.008	0.012	0.007	0.004	-0.002
2	0.019	-0.003	0.017	0.023	-0.034
3	0.091	0.021	0.023	0.064	-0.053
4	-0.086	-0.097	-0.059	-0.015	-0.009
5	-0.069	-0.102	-0.008	0.071	0.001

包含少量 P_{1024} 以及最后三个时间积分流量，贡献从低到中等再到极高。此外，第四因子同样是一个加权平均值，其中 P_{1024} 贡献较大，P_{64}和前三个时间积分流量的贡献较小。最后，第五因子主要由两个通量到达时间解释，其中 T_{90} 比 T_{50}略占主导，同时还有少量来自 F_1 和 F_2的贡献。总而言之，五个简单组的因子载荷各不相同，并进一步刻画了各组中伽马射线暴的离散程度。

我们还估算了因子得分 $\boldsymbol{F}_{ik}$（见第2.1.2节），以进一步刻画这些简单簇。我们采用 Thurstone（1935）常用的方法：对于 $\boldsymbol{x}_i$，若其被归类到第 k个简单组，则最小化加权平方误差损失 $\|\boldsymbol{\Psi}_k^{-1/2}(\boldsymbol{x}_i-\boldsymbol{\mu}_k-\boldsymbol{\Lambda}_k\boldsymbol{F}_{ik})\|^2$。相应的解是（Bartlett 1937）得分，其形式为$\widehat{\boldsymbol{F}}_{ik}=(\boldsymbol{\Lambda}_k^\top\boldsymbol{\Psi}_k^{-1}\boldsymbol{\Lambda}_k)^{-1}\boldsymbol{\Lambda}_k^\top\boldsymbol{\Psi}_k^{-1}(\boldsymbol{x}_i-\boldsymbol{\mu}_k)$，并且可以无偏地估计真实因子得分（Hershberger 2005；Distefano 等，2009）。因此，$\boldsymbol{x}_i$ 的因子得分代表第 i个观测值对 q 个潜在因子重要性的数值权重，即其“评分”（Distefano 等，2009）。表5列出了五个简单簇的平均因子得分。我们发现，对于第一个简单组，其伽马射线暴成员“评定”的潜在因子平均重要性依次为：因子 2 >、因子 1 >、因子 3 >、因子 4 >、因子5；对于第二组，这些重要性评分依次为：因子 4 >、因子 1 >、因子 3 >、因子 2 >、因子5。同样，第三组中各因子重要性的平均得分排序为：因子 1 >、因子 4 >、因子 3 >、因子 2 >、因子5；第四组的排序为：因子 5 >、因子 4 >、因子 3 >、因子 1 >、因子2；第五组的伽马射线暴则将重要因子的顺序“评定”为：因子4 >、因子 5 >、因子 3 >、因子 1 >、因子2。在这项分析中，我们同样看到，平均因子得分清晰地勾勒出各因子对不同简单组的重要性，且各组之间存在差异。接下来，我们将探究，经过稀疏化处理的完整 BATSE目录伽马射线暴中的五个简单组，能否解释为某种复杂超组结构的一部分。

3.2 识别群组中的复杂结构

我们应用第 2.2 节中的 MOBSynC 算法，研究合并倾向以及高阶复合簇或复合簇的存在情况。图 5a 展示了任意两个簇之间的成对重叠度，以及广义重叠度 $\ddot{\omega}$ 为 0.049。对于所选的 $\kappa=1$，簇 2、3 和 4 合并为一个组。对于首次合并后的簇，图 5b 展示了成对重叠度和广义重叠度 $\ddot{\omega}$ 为 0.042，这表明简单组 1 和 5 进一步合并。最后，我们得到两个复合组（简单簇 1、5 与简单簇 2、3、4），其广义重叠度为 $\ddot{\omega}$ 0.032。因此，我们的 MOBSynC 流程表明，经过稀疏化的复合



(a) $\ddot{\omega} = 0.049$ (b) $\ddot{\omega} = 0.042$

图 5. (a) MixFAD 算法结束时简单簇之间的成对重叠度量和广义重叠度量，以及 (b) 第一阶段合并结束时各组之间的重叠度量和广义重叠度量。

完整的 BATSE目录伽马射线暴在最高分辨率层级上被划分为五个椭球簇，在中间层级上被划分为三个簇，其中一个是由原先的椭球簇 2、3 和 4 构成的复合组，其余两个则是另外两个椭球簇。在最终层级上，我们得到两个复合组，分别由两个和三个椭球组构成。我们看到，MOBSynC 表征能够揭示伽马射线暴中的复杂结构和特征。现在我们来分析这些复合组。

描述复合组中的潜在结构

与第 3.1 节获得的简单组不同，其潜在结构是 MixFAD 的副产品；复合组中隐藏的潜在结构则只有经过一些额外的处理才能揭示。这是因为我们的方法找到的任何复合簇都不是高斯分布，而是高斯混合分布；然而，因子分析通常假设观测值至少服从椭球对称分布，而且往往假设其服从高斯分布。我们的方法将考虑分配给每个复合（合成）簇的观测值，对其中的数据进行高斯化，然后采用与 Bagoly 等人 (2006年) 类似的因子分析，但仅限于所考察复合簇中的数据。现在，我们将讨论一种使用高斯分布变换（GDT）对数据集进行高斯化的通用方法（Zhu 等人 2021），该方法在引入天体物理学领域的 Copula 模型通用框架（Nelsen 2006）中得到规定，相关工作由 Yuan 等人(2018) 引入。

GDT 的基本思想，事实上也是 copula 模型的基本思想，源于这样一个事实：对于任意具有连续累积分布函数（CDF）的单变量随机变量（RV） X，其累积分布函数为 $\mathcal{F}(x)$，则 $\mathcal{F}(X)$ 的分布是标准均匀分布。因此，变换后的随机变量 Y 服从标准正态分布，其中 $Y = \varrho(X) \doteq \Phi_0^{-1},_1[\mathcal{F}(X)],$，且 $\Phi_{\mu,\sigma^2}(\cdot)$ 是一个 $\mathcal{N}(\mu, \sigma^2)$ 随机变量的累积分布函数。因此，我们将单变量随机变量 X 高斯化为标准正态随机变量，而复合函数 $\varrho(\cdot) = \Phi_0^{-1},_1 \circ \mathcal{F}(\cdot)$ 是一个高斯化函数。在多元情形下，我们对随机向量 $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$ 的每个分量分别应用高斯化变换。也就是说，我们得到 $Y_j = \varrho_j(X) \doteq \Phi_0^{-1},_1[\mathcal{F}_j(X_j)],$，其中 $\mathcal{F}_j(\cdot)$ 是 X_j 的边缘累积分布函数。随后，变换后的随机向量 $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_p)$ 具有标准正态边缘分布，但各分量之间的关联在 $\boldsymbol{Y}$ 中得以保留；事实上，$\boldsymbol{Y}$ 的方差协方差矩阵是一个相关矩阵。我们对这个相关矩阵进行因子分析，使用Dai 等人（2021）提出的计算效率较高的 FAD算法。该相关矩阵由对分配给每个复合簇的数据进行高斯化后得到；此时还需确定复合簇的边缘累积分布函数 $\mathcal{F}_j(\cdot)$。根据第2.2.2节的讨论，复合簇中随机向量 $\boldsymbol{X}$ 的累积分布函数为 $\sum_{l\in\mathcal{C}_m} \eta_l^* \boldsymbol{\Phi}(\cdot; \boldsymbol{\mu}_l, \boldsymbol{\Sigma}_l)$，其中 $\eta_l^* = \eta_l / \sum_{h\in\mathcal{C}_m} \eta_h$。因此，边缘累积分布函数 $\mathcal{F}_j(\cdot) = \sum_{l\in\mathcal{C}_m} \eta_l^* \Phi_{\boldsymbol{\mu}_{lj}\ \boldsymbol{\sigma}_{ij}^2}$. $\eta_l^* = \eta_l / \sum_{h\in\mathcal{C}_m} \eta_h$ $\boldsymbol{\mu}_{lj}$,

( ) with , 其中 是向量 $\boldsymbol{\mu}_l$ 的第j个分量，而 $\boldsymbol{\sigma}_{lj}^2$ 是 $\boldsymbol{\Sigma}_l$ 的第 (j, j) 个对角元素。定义这些量后，我们便可以研究并刻画复合簇中的潜在结构。不过，我们也在此指出，潜在因子之间的线性关系位于高斯化空间中；与简单组的情形不同，它不在原始数据空间中。尽管如此，这种方法仍能概括九个原始参数在潜在因子中的相对强度，以解释复合组中的变异性。

采用上述方法，我们总结了第一次合并后形成的唯一复合簇（2、3、4）。（其余簇1和5，以及它们的潜在特征描述，均与第 3.1 节相比没有变化。）对于被分配到该簇的伽马射线暴，$\log_{10} T_{50}$ (1.076)、$\log_{10} T_{90}$ (1.537)和$\log_{10} P_{1024}$ (0.258)的平均水平明显高于另外两个简单簇1和5（表3 2、3、4 93.3 中的平均值）。此外，合并簇（ ）中的高斯化数据可以用五个因子进行最佳拟合，解释总方差的约 %。表6列出了四次方最大旋转后的估计因子载荷（Costello和Osborne（2005））， 其中，复合簇（2、3、4）的第一个因子载荷向量在高斯化空间中是一个加权平均值，T_{90}、F_4、F_1、F_2、F_3、P_{64}、P_{256} 和 P_{1024} 的贡献依次由低到高。第二个因子主要由两个通量到达时间决定，同时包含来自四个时间积分能流的少量贡献（均处于高斯化空间中），以及来自三个峰值流量的少量反向贡献。第三个因子载荷向量是在高斯化空间中，一方面由前三个时间积分能流的加权平均值构成（其中 F_3 的贡献比 F_1 和 F_2低约60%），另一方面则是三个峰值流量的少量贡献的简单平均值，二者形成对比。第四个因子同样是在高斯化空间中，由 F_3 和 F_4 的小幅贡献的加权平均值，与P_{64}、P_{256} 和 P_{1024}的小幅贡献的简单平均值形成对比。最后，第五个因子由 P_{64} 和 P_{1024}贡献少量但不可忽略的成分（符号相反）。此外，表7给出了从复合簇（2、3、4）估计的平均因子得分，其中得分值的大小表明，对于来自该复合簇的伽马射线暴，第二和第三因子的重要性评分接近总体平均评分水平，而第一个和最后一个因子的评分明显更高，具体而言，因子 1 > 因子5> 因子 4 > 因子 2 > 因子3。我们可以看到，复合簇（2、3、4）呈现出不同于簇1和5的鲜明特征。

解释完复合簇（2, 3, 4）后，我们现在来解释第二阶段合并后形成的复合簇（1, 5）。我们对高斯化空间中来自复合簇（1, 5）的伽马射线暴进行了因子分析，得到了一个最优四因子模型，可解释总方差的 91.9%。在

表6。 复合簇 (1, 5) 和 (2, 3, 4) 在高斯化空间中的估计因子载荷。为清晰起见，表中未列出区间 (-0.1,0.1) 内的数值。与之前一样，热图提供了快速、直观的参考。

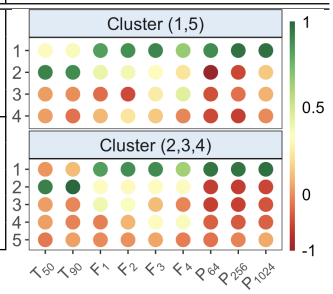
k	q	$\log_{10} T_{50}$	$\log_{10} T_{90}$	$\log_{10} F_1$	$\log_{10} F_2$	$\log_{10} F_3$	$\log_{10} F_4$	$\log_{10} P_{64}$	$\log_{10} P_{256}$	$\log_{10} P_{1024}$	热图表示
(1, 5)	1	0.331	0.362	0.824	0.877	0.901	0.657	0.880	0.962	0.963	
	2	0.912	0.885	0.409	0.388	0.320	0.219	-0.322	-0.172	0.140	
	3				-0.162	0.248	0.450	-0.148			
	4				0.223	0.138		-0.167	-0.191		
(2, 3, 4)	1		0.106	0.827	0.879	0.882	0.630	0.951	0.952	0.965	
	2	0.917	0.990	0.323	0.312	0.316	0.331	-0.196	-0.197	-0.175	
	3			0.395	0.342	0.102		-0.193	-0.186	-0.136	
	4					0.322	0.358	-0.108	-0.123	-0.105	
	5							-0.100		0.100	

表7. 两个复合簇（1，5）和（2，3，4）的平均因子得分。

簇	因子1	因子2	因子3	因子4	因子5	
(1, 5)	0.167	0.111	0.118	0.112	–	
(2, 3, 4)	0.028	0.006	-0.003	0.019	0.025	

四次方最大旋转下估计的组别因子载荷列于表6，其中显示了不可忽略的数值。可以看出，第一因子受到两个通量到达时间的少量影响，受到 F_4 的中等影响，受到 F_1、 F_2、 F_3 和 P_{64} 的显著影响，并受到 P_{256} 和 P_{1024} 的较大影响（均处于高斯化空间）。第二因子载荷主要由两个通量到达时间主导，同时受到四个时间积分流量和 P_{1024} 的中等至较小影响，并受到 P_{64} 和 P_{256} 方向相反的中等至微弱影响。高斯化空间中的第三因子载荷向量体现了最后三个时间积分流量与 P_{64} 之间较小的对比；一方面是 F_3 和 F_4 的加权平均，其中后者的贡献幅度几乎是前者的两倍；另一方面则是 F_2 和 P_{64} 的少量贡献。最后一个，即第四因子，在高斯化空间中体现为 P_{64} 和 P_{256} 的少量贡献与 F_2 和 F_3 的少量贡献简单平均值之间的对比。表7再次列出了复合簇（1, 5）的平均因子得分，其中四个潜在因子变量的重要性依次评为因子 1 > 因子 3 >因子 4 > 因子 2，第一因子的评级比最后一个因子高 50%。

3.3 对我们结果的进一步解读

Mukherjee 等人（1998年）提出了一种持续时间-流量-能谱方案，利用一个原始特征和两个复合特征来表征 BATSE伽马射线暴。这三个汇总特征分别是持续时间（T_{90}）、总流量（$F_1+F_2+F_3+F_4$）和谱硬度（$H_{321}=F_3/(F_1+F_2)$）。Chattopadhyay 等人也（2007年）以及之后的多项研究也采用了这种分类与表征方案（Chattopadhyay与Maitra（2017），2018；Berry与Maitra（2019）；Almodovar-Rivera 与 Maitra 2020；Goren与Maitra（2022））。我们在图6中通过密度图和成对散点图展示了每个简单组的持续时间、总流量和谱硬度分布。虽然成对散点图和单变量密度图只能让我们部分了解各项关系，但我们可以清楚地看到五个简单组各自的特点，以及其中一些组之间的相似之处；这些相似之处使它们在更高层级上合并，下面将对此进行说明。对于簇 2、3、4 和 5，持续时间与总流量的相关性强于簇 1；在簇 1 中，总流量与持续时间的相关性较弱，且与谱硬度呈负相关。持续时间与谱硬度在簇 2 中几乎不相关。比较五个簇的三项特征分布可见，簇 2、3 和 4 的持续时间与总流量水平高于簇 1 和 5；簇 5 的持续时间变异最大。与此同时，簇 2、3 和 4 的谱硬度水平低于簇 1 和 5。这有助于理解第

表 8. 根据持续时间对五个简单伽马射线暴簇进行表征 - 通量-谱方案。

簇	1	2	3	4	5
持续时间	较短	long	long	long	短
通量	微弱	中等中等		明亮	微弱
谱	hard	soft	中等中等		hard

3、4 和 5 中，持续时间与总流量的相关性强于簇 1；在簇 1 中，总流量与持续时间的相关性较弱，且与谱硬度呈负相关。持续时间与谱硬度在簇 2 中几乎不相关。比较五个簇的三项特征分布可见，簇 2、3 和 4 的持续时间与总流量水平高于簇 1 和 5；簇 5 的持续时间变异最大。与此同时，簇 2、3 和 4 的谱硬度水平低于簇 1 和 5。这为理解复合簇（2、3、4）以及第 3.2 节第一阶段合并后得到的两个简单簇 1 和 5 提供了线索。我们的观测结果促使我们在表8中对五个简单簇进行分类：均值显示，簇 1 和 5 分别属于较短-较暗-较硬和短-暗-硬，而簇 2（长-中等-软）、簇 3（长-中等-中等）和簇 4（长-亮-中等）表现出相似的特征，且与簇 1 和 5 的差异更为显著；第3.2节的最终合并结果也已表明这一点。同时需要指出，簇 1 和 5 的持续时间参数在偏度和变异方面明显不同：簇 1 的偏度几乎为零，而簇 5 的偏度更大、变异也更高。这也有助于解释为何这两组被分别识别为简单（椭球形）簇。

4 结论

传统上对伽马射线暴进行分类和表征时，主要使用持续时间参数，通常会发现伽马射线暴分为两类或三类；而近期研究（Chattopad- hyay 与 Maitra 2017年，2018年；Berry与Maitra（2019）；Almodovar- Rivera 与 Maitra 2020）则通过纳入更多参数并采用更先进的统计技术，识别并表征了新的伽马射线暴类别。鉴于这些相互矛盾的结果，我们研究了完整 BATSE目录中符合以下条件的 1150 个伽马射线暴：其全部九个天体物理参数T_{50}，T_{90}，F_1，F_2，F_3，F_4，P_{64}，P_{256}，P_{1024} 对于通量到达时间大于2σ，且对于时间积分

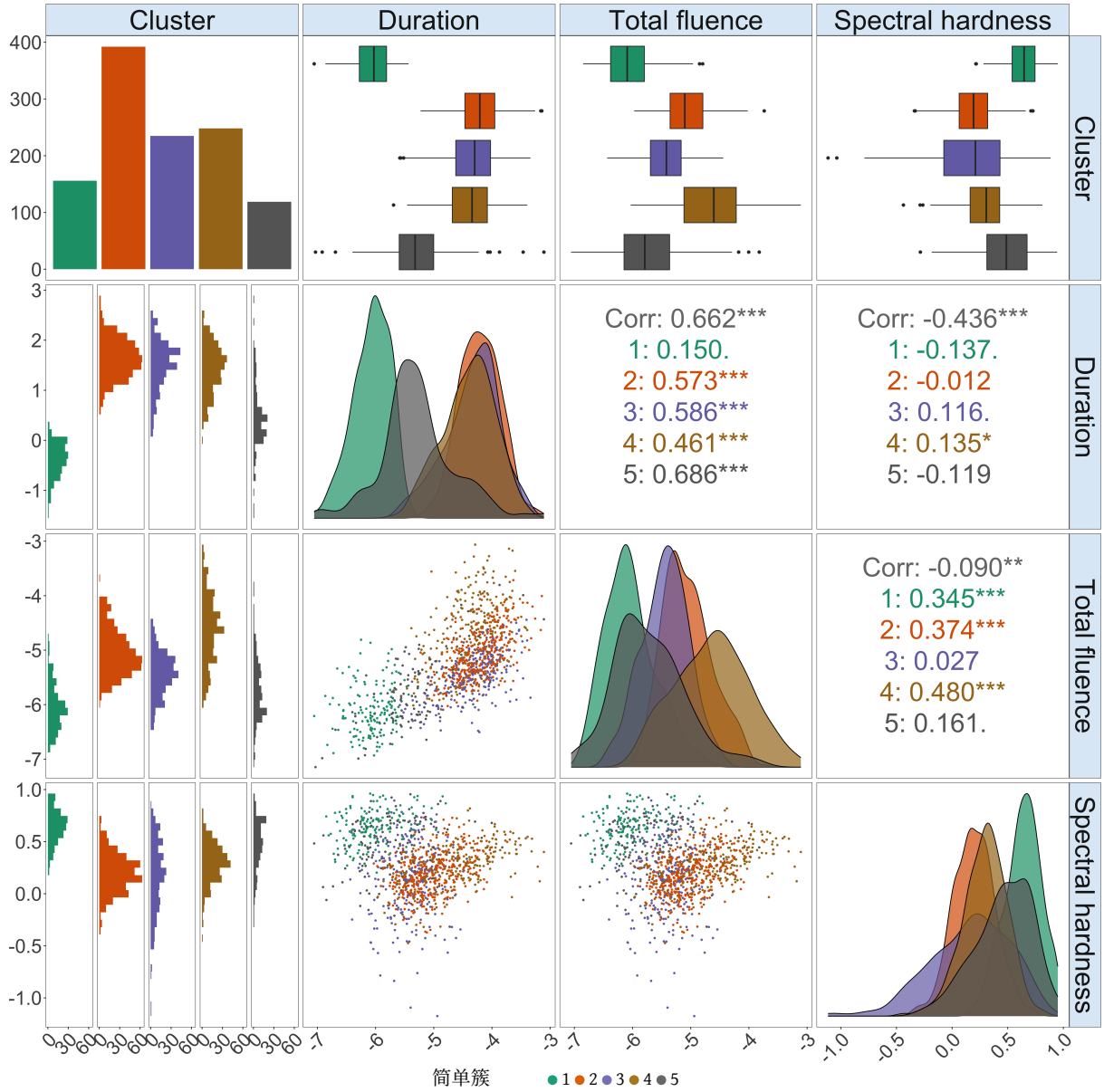


图 6. 对于1150个经稀疏处理的完整BATSE目录伽马射线暴，展示了三个衍生特征（以$\log_{10}$为尺度）的密度和散点图：持续时间、总流量和谱硬度；颜色表示五个简单簇。上方面板展示了特征之间的相关性。请注意，第一个簇的持续时间变量没有偏度，而最后一个簇的偏度极高，这使这两个组得以区分。

通量以及峰值流量参数。我们采用基于因子分析器高斯混合模型的 MBC 分析，将这 1150 个暴划分为五个椭球形群组；该模型通过无矩阵计算实现，可高效进行参数估计。按照持续时间-流量-能谱方案（Mukherjee 等人 1998年；Chattopadhyay 等人 2007年），根据持续时间（T_{90}）、总流量（$F_1 + F_2 + F_3 + F_4$）和能谱硬度比（$H_{321} = F_3/(F_1 + F_2)$，这五个 GRB 组进一步区分为：第 1 组的 156 个较短-较暗-较硬伽马射线暴；第 5 组的 119 个短-暗-硬伽马射线暴，其中第 5 组的持续时间分布偏斜程度更大、变异也更多；第 2 组的 392 个长-中等-软伽马射线暴；第 3 组的 235 个长-中等-中等伽马射线暴；以及第 4 组的 248 个长-亮-中等伽马射线暴。这些群组还通过五个潜在因子得到了鲜明表征这九个变量。我们的研究结果支持了近期的发现，即BATSE伽马射线暴数据集中存在五个椭球形组。对于完整BATSE目录中未进行稀疏处理的1598个伽马射线暴样本，在忽略测量误差的情况下将伽马射线暴纳入分析，也得到了定性上相似的结果。与本文详细研究的稀疏样本类似，我们再次得到了五个简单组，它们先合并为三个复合组，随后再合并为两个复合组。

然而，这五个椭球形组与其他研究结果存在一些不一致；那些研究在分析中使用较少的参数时，发现的伽马射线暴组数量也较少。为了解释这一差异，并受到Almodovar-Rivera & Maitra (2020);Chattopadhyay 与 Maitra(2018); Chattopadhyay 等人 (2022)的合胞体聚类方法启发，我们开发了 MOBSynC，它

根据各组之间的重叠度量，将分离程度较低的组并为复合簇。将 MOBSynC 应用于 MixFAD 发现的五个简单组后，我们发现它首先将第二、第三和第四个简单组合并成一个较大的簇，而另外两个简单组则保持不变。经过第一阶段合并，伽马射线暴被表征为三组，主要可根据持续时间划分为长、短和极短持续时间三类（T_{90}）。下一阶段促使两个简单组合并，从而得到伽马射线暴的两组表征。我们的结果显示，BATSE伽马射线暴具有多层次表征，这也得到了持续时间-通量-谱分类方案的支持。因此，我们的分析表明，存在两种总体形态各异的复合型伽马射线暴，而每一种又分别由两个和三个椭球形子群构成。两个简单组比三个简单组更易区分，并在中间阶段形成三组解。在最深层次上，我们得到由五个椭球形组构成的表征。因此，我们解释了文献中关于伽马射线暴分为两类、三类或五类的发现，并表明这三种解均可在合胞体框架下得到解释。

除了识别不同类型的伽马射线暴外，我们还将伽马射线暴简单组的变异性按组进行了表征，这是 MixFAD 算法自然产生的副产品；我们发现，九个参数的变异性可以用少数几个潜在因子来解释。识别出的伽马射线暴组通过底层特征呈现出参数间不同的依赖结构，并且各因子的重要性评分也各不相同。在更高层次上，复合组的潜在因子表示并不显而易见，因此我们开发了一种新方法，对每个复合簇中的数据进行高斯化，并用四个和五个潜在因子来解释这些复合组中的变异性。

当前工作之外，还有若干值得探讨和进一步拓展的潜在主题。首先，正如审稿人所指出的，每个伽马射线暴的参数都带有测量误差（或系统误差）。虽然这些误差已有数据可用，且应纳入我们的分析，但目前尚无统计方法能够将其纳入 MixFAD 聚类框架，同时不仅对伽马射线暴进行分组，还能用少数几个组别特异的潜在因子来解释其变异性，因此我们无法在分析中纳入这些误差。亟须针对伽马射线暴数据集开发、验证并应用此类统计方法，这将留待未来研究。其次，BATSE目录还包含特征记录不完整的伽马射线暴，聚类方法应考虑缺失值的存在。不过，尽管已有使用一般混合模型对这类数据进行聚类的方法（Goren与Maitra（2022）），使用因子分析器混合模型对这类数据进行聚类的方法仍有待开发和应用。若能将这些观测不完整的伽马射线暴纳入分析，并由此加深对数据集的认识，将会很有意义。此外，我们当前的模型假设所有簇的因子数量相同，但也值得考虑一种更灵活的模型，允许不同组别具有不同数量的因子。最后，本文提出了一种表征和聚类伽马射线暴的方法：我们认为，该方法具有足够的通用性，也可应用于其他类似数据集，以揭示复杂结构和潜在关系。

致谢

作者非常感谢一位匿名审稿人，他迅速而细致地审阅了本文的两个早期版本，并提出了富有见地的意见，极大地改进了稿件内容。我们也感谢科学编辑和助理编辑在审稿过程中及时提供帮助。

数据可用性声明

本文第2.2.3.1节使用的模拟数据集可在 https://github.com/fanstats/ MixFAD-GRB 公开获取。本文使用的伽马射线暴数据集包含全部九个参数的完整数据，这些原始数据取自完整的 BATSE 伽马射线暴目录，网址为 https://heasarc.gsfc.nasa.gov/cgi-bin/W3Browse/w3query.pl?&tablehead=name%3Dheasarc_batsegrb%26description%3DCGRO%2FBATSE+Gamma-Ray+Burst+Catalog%26url%3Dhttp%3A%2F%2Fheasarc.gsfc.nasa.gov%2FW3Browse%2Fcgro%2Fbatsegrb. html%26archive%3DY%26radius%3D300%26mission% 3DCGRO%26priority%3D3&mission=CGRO&Action=More+Options&Action=Parameter+Search&ConeAdd=1。原始和处理后的伽马射线暴数据集也可在 https://github.com/fanstats/MixFAD-GRB 获取。用于执行 MixFAD 的软件将作为名为 MixFAD 的 R语言（R核心团队 2024）软件包公开发布，而 MOBSynC 将作为 R语言中的 SynClustR软件包的一部分公开提供（R核心团队 2024）。

参考文献

Almodovar-Rivera I. A., Maitra R., 2020年, 机器学习研究杂志, 21, 1 Anderson T. W., 2003年,《多元统计分析导论》。Wiley概率与统计丛书，Wiley Bagoly Z., Mészáros A., Horváth I., Balázs L. G., Mészáros P., 1998年，《天体物理学杂志》, 498, 342 Bagoly Z., Borgonovo L., Mészáros A., Balázs L. G., Horvath I., 2009年，《天文学与天体物理学》，493, 51 Bartlett M., 1937年,《英国心理学杂志》。总论部分, 28, 97 Baudry J.-P., Raftery A. E., Celeux G., Lo K., Gottardo R., 2010年，《计算与图形统计学杂志》, 19, 332 Berger E., Fong W., Chornock R., 2013年，《天体物理学杂志快报》,774Berry N., MaitraR., 2019,《统计分析与数据挖掘》——《美国统计学会数据科学杂志》,12, 223Bhardwaj S., Dainotti M. G., Venkatesh S., Narendra A., Kalsi A., Rinaldi E., Pollo A., 2023,《皇家天文学会月刊》, 525,5204Bhave A., KulkarniS., Desai S., Srijith P. K., 2022,《天体物理学与空间科学》, 367 Byrd R. H., P. Lu J. N., Zhu C., 1995,《工业与应用数学学会科学计算杂志》, 16, 1190 Chattopadhyay S., Maitra R., 2017年，《皇家天文学会月刊》, 469, 3374 Chattopadhyay S., Maitra R., 2018年，《皇家天文学会月刊》, 481, 3196 Chattopadhyay T., Misra R., Chattopadhyay A. K., Naskar M., 2007年，《天体物理学杂志》, 667,1017ChattopadhyayS., Kawaler S. D., Maitra R.,2022,《澳大利亚天文学会出版物》, 39, 1 Costello A. B., Osborne J., 2005,《实用评估、研究与评价》，10, 1

Dai F., Dutta S., Maitra R., 2020, 《计算与图形统计学杂志》, 29, 675 Dai F., Dorman K. S., Dutta S., Maitra R., 2021,《球面数据的探索性因子分析》, doi:10.48550/ARXIV.2111.04940, https://arxiv.org/abs/2111.04940 Day N. E., 1969,《生物计量学》, 56, 463 Dempster A. P., Laird N. M., Rubin D. B., 1977,《皇家统计学会杂志》B辑, 39, 1 Dezalay J.-P., Barat C., Talon R., Syunyaev R., Terekhov O., Kuznetsov A., 1992, 载于 Paciesas W. S., Fishman G. J., 编, 《美国物理学会会议系列》第265卷, 《美国物理学会会议系列》, 第304–309页 Distefano C., Zhu M., Mindrila D., 2009,《实用评估、研究与评价》, 14, 20 Dutta S., Mondal D., 2015,《皇家统计学会杂志：B辑（统计方法）》, 77, 699 Ghirlanda G., 等, 2017年, 《天文学与天体物理学》, 609 Goren E. M., Maitra R., 2022, Stat, 11, 416 Hakkila J., Giblin T. W., Roiger R. J., Haglin D. J., Paciesas W. S., Meegan C. A., 2003, 《天体物理学杂志》, 582, 320 Hennig C., 2010,《数据分析与分类进展》 Hershberger S. L., 2005,《行为科学统计学百科全书》, 第636–644页 Horváth I., 1998,《天体物理学杂志》, 508, 757 Horváth I., 2002,《天文学与天体物理学》, 392, 791 Horváth I., 2009,《天体物理学与空间科学》, 323 Horváth I., Tóth B. G., 2016,《天体物理学与空间科学》, 361, 155 Horváth I., Mészáros A., Balázs L. G., Bagoly Z., 2004,《波罗的海天文学》, 13, 217 Huja, D. Mészáros, A. Řípa, J. 2009,《天文学与天体物理学》, 504, 67 Ioka K., Hotokezaka K., Piran T., 2016,《天体物理学杂志》, 833, 110 Kouveliotou C., Meegan C. A., Fishman G. J., Bhat N. P., Briggs M. S., Koshut T. M., Paciesas W. S., Pendleton G. N., 1993,《天体物理学杂志》, 413, L101 Lawley D. N., 1940,《爱丁堡皇家学会学报》, 60, 64–82 Maitra R., 2009,《IEEE/ACM计算生物学与生物信息学汇刊》, 6, 144 Maitra R., 2010,《神经影像》, 50, 124 Maitra R., 2013, Sankhy¯a:《印度统计学杂志》B辑, 75, 293 Maitra R., Melnykov V., 2010,《计算与图形统计学杂志》, 19, 354 Mardia K. V., Kent J. T., Bibby J. M., 2006,《多变量分析》。Elsevier，阿姆斯特丹 Mazets E. P., 等, 1981,《天体物理学与空间科学》, 80, 3 McLachlan G., Krishnan T., 2008,《EM算法及其扩展》, 第2版。Wiley，纽约，doi:10.2307/2534032 McLachlan G., Peel D., 2000,《有限混合模型》。John Wiley and Sons, Inc.，纽约, doi:10.1002/0471721182 Melnykov V., 2016,《计算与图形统计学杂志》, 25, 66 Melnykov V., Maitra R., 2010, Statist. Surv., 4, 80 Melnykov V., Maitra R., 2011, 机器学习研究杂志, 12, 69 Melnykov V., Chen W.-C., Maitra R., 2012,《统计软件杂志》, 51, 1 Mukherjee S., Feigelson E. D., Jogesh Babu G., Murtagh F., Fraley C., Raftery A., 1998,《天体物理学杂志》, 508, 314 Nakar E., 2007,《物理学报告》, 442, 166 Nelsen R. B., 2006,《Copula导论》, 第2版。Springer，纽约 Norris J. P., Cline T. L., Desai U. D., Teegarden B. J., 1984,《自然》, 308, 434 Paczyński B., 1998,《天体物理学杂志》, 494, L45 Pendleton G. N., 等, 1997,《天体物理学杂志》, 489, 175 Peterson A. D., Ghosh A. P., Maitra R., 2018, Stat, 7, e172 Piran T., 1992, apjl, 389, L45 Piran T., 2005,《现代物理学评论》, 76, 1143 R核心团队 2024, R：统计计算语言与环境。R基金会，维也纳，奥地利， https://www.R-project.org/ Rubin D. B., Thayer D. T., 1982,《心理计量学》, 47, 69 Schwarz G. E., 1978,《统计年鉴》, 6, 461

Sorensen D. C., 1992, SIAM矩阵分析与应用杂志, 13, 357 Stanek K., et al., 2008年,《天体物理学杂志快报》, 591, L17 Steinhardt C., Mann W., Rusakov V., Jespersen C., 2023,《天体物理学杂志》, 945, 67Tarnopolski M., 2015,《天文学与天体物理学》, 581, A29 Tarnopolski M., 2019,《天体物理学杂志》, 887, 97 Tarnopolski M., 2022,《天文学与天体物理学》, 657, A13 Thurstone L., 1931,《心理学评论》, 38, 406 Thurstone L. L., 1935,《心智向量：用于分离基本特质的多因素分析》。。芝加哥大学出版社 Tóth B. G., Rácz I. I., Horváth I., 2019,《皇家天文学会月刊》, 486, 4823 Wilkinson J. H., 1958,《计算机杂志》, 1, 90 Woosley S., Bloom J., 2006年,《天文学与天体物理学年度评论》, 44, 507 Yang H., ShiC., Cai J., Zhou L., Yang Y., ZhaoX., He Y., Hao J., 2022,《皇家天文学会月刊》, 517, 5496 Yuan Z., WangJ., Worrall D. M.,Zhang B.-B., Mao J.,2018,《天体物理学杂志增刊系列》, 239, 33 Zhu Y., Dai F., Maitra R., 2021, 有标签混合特征数据集的可视化, doi:10.48550/arXiv.1904.06366, https://arxiv.org/abs/1904.06366 Zhu S.-Y., Sun W.-P., Ma D.-L., Zhang F.-W., 2024年,《皇家天文学会月刊》, 532, 1434 Zitouni H., Guessoum N., Azzam W. J., Mochkovitch R., 2015,《天体物理学与空间科学》, 357,7

附录A：关于 MixFAD 算法的理论细节

如第 2.1.2 节所述，主要挑战在于对 $\mathbf{\Lambda}_k$ 的和 $\mathbf{\Psi}_k$ 进行计算高效的 M步估计，因此我们在此详细说明一种通过剖面似然法（Dai 等人 2020）联合更新 $\mathbf{\Lambda}_k$ 和 $\mathbf{\Psi}_k$ 的方法，该方法基于在（5）中获得的 $\mathbf{\Sigma}_k^*$。具体而言，可以使用以下方法将 $\mathbf{\Lambda}_k$ 从Q函数（3）中剖面化消去：

结果 1 $\mathbf{\Psi}_k$ $\theta_{1\,k} \geq \theta_{2\,k} \geq \cdots \geq \theta_{q\,k}$ $\boldsymbol{W}_k = \boldsymbol{V}_{q,k}$ 对于正定对角矩阵，设 $\mathbf{\Psi}_k^{-1/2}\mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1/2}$ 为 q 的最大特征值。令 $\boldsymbol{V}_{q,k}$ 的列存储与这些特征值对应的特征向量。然后，关于 $\mathbf{\Lambda}_k$ 的Q函数在 $\widehat{\mathbf{\Lambda}}_k = \mathbf{\Psi}_k^{1/2}\boldsymbol{V}_{q,k}\mathbf{\Delta}_k$ 处达到最大值，其中 是一个 对角矩阵，其 个对角元素为 $max(,_k - 1,0)]^{1/2}$. 剖面Q函数为

$$\mathrm{Q}_p(\mathbf{\Psi}_k) = c - \frac{n}{2}\Big\{ \log\det\mathbf{\Psi}_k + \mathrm{Tr}\,\mathbf{\Psi}_k^{-1}\mathbf{\Sigma}_k^* + \sum_{j=1}^{q}(\log\theta_{j,k} - \theta_{j,k} + 1)\Big\}, \tag{A1}$$

其中 c 是一个与 $\mathbf{\Psi}_k$ 和 $\mathbf{\Sigma}_k^*$ 无关的常数，后者是在（5）中的给定估计值。此外，$\nabla\mathrm{Q}_p(\mathbf{\Psi}_k) = -\frac{1}{2}n\,\mathrm{diag}(\widehat{\mathbf{\Lambda}}_k\widehat{\mathbf{\Lambda}}_k^\top + \mathbf{\Psi}_k - \mathbf{\Sigma}_k^*)$。

证明。 结果 1的证明可由推导Dai 等人 （2020）中的剖面对数似然函数得出。首先，我们得到关于 $\mathbf{\Lambda}_k$、$\mathbf{\Psi}_k$的 Q函数，它是如下的正态对数似然函数：

$$Q(\mathbf{\Lambda}_k, \mathbf{\Psi}_k) = c - \frac{n}{2}\Big\{ \log\det(\mathbf{\Lambda}_k\mathbf{\Lambda}_k^\top + \mathbf{\Psi}_k) + \mathrm{Tr}\,(\mathbf{\Lambda}_k\mathbf{\Lambda}_k^\top + \mathbf{\Psi}_k)^{-1}\mathbf{\Sigma}_k^*\Big\}, \tag{A2}$$

其中 c 是一个与 $\mathbf{\Lambda}_k$、$\mathbf{\Psi}_k$无关的常数。

基于（A2），通过求解得分方程，可得到 $\mathbf{\Lambda}_k$ 和 $\mathbf{\Psi}_k$ 的极大似然估计量

$$\begin{cases} \mathbf{\Lambda}_k(\mathbf{I}_q + \mathbf{\Lambda}_k^\top\mathbf{\Psi}_k^{-1}\mathbf{\Lambda}_k) = \mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1}\mathbf{\Lambda}_k \\ \mathbf{\Psi}_k = \mathrm{diag}(\mathbf{\Sigma}_k^* - \mathbf{\Lambda}_k\mathbf{\Lambda}_k^\top) \end{cases} \tag{A3}$$

From $\mathbf{\Lambda}_k(\mathbf{I}_q + \mathbf{\Lambda}_k^\top \mathbf{\Psi}_k^{-1}\mathbf{\Lambda}_k) = \mathbf{\Sigma}_k^* \mathbf{\Psi}_k^{-1}\mathbf{\Lambda}_k$, we have

$$\begin{aligned}\mathbf{\Psi}_k^{-1/2}\mathbf{\Lambda}_k(\mathbf{I}_q + (\mathbf{\Psi}_k^{-1/2}\mathbf{\Lambda}_k)^\top \mathbf{\Psi}_k^{-1/2}\mathbf{\Lambda}_k) \\ = \mathbf{\Psi}_k^{-1/2}\mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1/2}\mathbf{\Psi}_k^{-1/2}\mathbf{\Lambda}_k.\end{aligned} \tag{A4}$$

假设 $\mathbf{\Psi}_k^{-1/2}\mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1/2} = \boldsymbol{V}_k \boldsymbol{D}_k \boldsymbol{V}_k^\top$，且 $\boldsymbol{D}_k$ 中的对角元素按降序排列，其中 $\theta_{1,k} \geq \theta_{2,k} \geq \cdots \geq \theta_{p,k}$。令 $\boldsymbol{D}_k = \begin{bmatrix} \boldsymbol{D}_q & 0 \\ 0 & \boldsymbol{D}_m \end{bmatrix}$，其中 $m = p - q$，并且 $\boldsymbol{D}_{q,k}$ 包含最大的 q 个特征值 $\theta_{1,k} \geq \theta_{2,k} \geq \cdots \geq \theta_{q,k}$。相应的 q 个特征向量构成矩阵 $\boldsymbol{V}_{q,k}$ 的列，使得 $\boldsymbol{V}_k = [\boldsymbol{V}_{q,k}, \boldsymbol{V}_{m,k}]$。然后，若 $\boldsymbol{D}_{q,k} > \mathrm{I}_q$，则（A4）表明

$$\mathbf{\Lambda}_k = \mathbf{\Psi}_k^{1/2}\boldsymbol{V}_{q,k}(\boldsymbol{D}_{q,k} - \mathbf{I}_q)^{1/2}. \tag{A5}$$

因此，在给定 $\mathbf{\Psi}_k$ 的条件下，$\mathbf{\Lambda}_k$ 在 $\mathbf{\Lambda}_k = \mathbf{\Psi}_k^{1/2}\boldsymbol{V}_{q,k}\mathbf{\Delta}_k$ 处取得最大值，其中 $\mathbf{\Delta}_k$ 是一个对角矩阵，其元素为 $\max(\theta_{j,k} - 1, 0)^{1/2}$，$j = 1, \cdots, q$。

根据 $\mathrm{V}_{q,k}$ 和 $\mathrm{V}_{m,k}$的构造，我们有 $\mathrm{V}_{q,k}^\top \mathrm{V}_{q,k} = \mathrm{I}_q, \mathrm{V}_{m,k}^\top \mathrm{V}_{m,k} = \mathrm{I}_m, \mathrm{V}_{q,k}\mathrm{V}_{q,k}^\top + \mathrm{V}_{m,k}\mathrm{V}_{m,k}^\top = \mathrm{I}_p, \mathrm{V}_{q,k}^\top \mathrm{V}_{m,k} = \mathbf{0}$，因此，$(\mathrm{V}_{q,k}\mathrm{D}_{q,k}\mathrm{V}_{q,k}^\top + \mathrm{V}_{m,k}\mathrm{V}_{m,k}^\top)(\mathrm{V}_{q,k}\mathrm{D}_{q,k}^{-1}\mathrm{V}_{q,k}^\top + \mathrm{V}_{m,k}\mathrm{V}_{m,k}^\top) = \mathrm{I}_p$。

Let $\mathbf{A}_k = \mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top$. Then $\mathbf{A}_k\mathbf{A}_k = \mathbf{V}_{q,k}\mathbf{\Delta}_k^4\mathbf{V}_{q,k}^\top$ and

$$\begin{aligned}|\mathbf{A}_k + \mathbf{I}_p| = |(\mathbf{A}_k + \mathbf{I}_p)\mathbf{A}_k|/|\mathbf{A}_k| &= \frac{|\mathbf{V}_{q,k}(\mathbf{\Delta}_k^4 + \mathbf{\Delta}_k^2)\mathbf{V}_{q,k}^\top|}{|\mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top|} \\ &= |\mathbf{\Delta}_k^2 + \mathbf{I}_{q,k}| = \prod_{j=1}^{q}\theta_{j,k}\end{aligned} \tag{A6}$$

and

$$\begin{aligned}(\mathbf{A}_k + \mathbf{I}_p)^{-1} &= (\mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top + \mathbf{V}_{q,k}\mathbf{V}_{q,k}^\top + \mathbf{V}_{m,k}\mathbf{V}_{m,k}^\top)^{-1} \\ &= (\mathbf{V}_{q,k}(\mathbf{\Delta}_k^2 + \mathbf{I}_q)\mathbf{V}_{q,k}^\top + \mathbf{V}_{m,k}\mathbf{V}_{m,k}^\top)^{-1} \\ &= (\mathbf{V}_{q,k}\mathbf{D}_{q,k}\mathbf{V}_{q,k}^\top + \mathbf{V}_{m,k}\mathbf{V}_{m,k}^\top)^{-1} \\ &= \mathbf{V}_{q,k}\mathbf{D}_{q,k}^{-1}\mathbf{V}_{q,k}^\top + \mathbf{V}_{m,k}\mathbf{V}_{m,k}^\top.\end{aligned} \tag{A7}$$

基于（A2）、（A6）和（A7），剖面对数似然为，

$$\begin{aligned}Q_p(\mathbf{\Psi}_k) &= c - \frac{n}{2}\log|\widehat{\mathbf{\Lambda}}_k\widehat{\mathbf{\Lambda}}_k^\top + \mathbf{\Psi}_k| \\ &\quad - \frac{n}{2}\mathrm{Tr}\,(\widehat{\mathbf{\Lambda}}_k\widehat{\mathbf{\Lambda}}_k^\top + \mathbf{\Psi}_k)^{-1}\mathbf{\Sigma}_k^* \\ &= c - \frac{n}{2}\Big\{\log|\mathbf{\Psi}_k^{1/2}(\mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top + \mathbf{I}_p)\mathbf{\Psi}_k^{1/2}| \\ &\quad + \mathrm{Tr}\,(\mathbf{\Psi}_k^{1/2}(\mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top \\ &\quad + \mathbf{I}_p)\mathbf{\Psi}_k^{1/2})^{-1}\mathbf{\Sigma}_k^*\Big\} \\ &= c - \frac{n}{2}\Big\{\log\det\mathbf{\Psi}_k + \log|\mathbf{V}_{q,k}\mathbf{\Delta}_k^2\mathbf{V}_{q,k}^\top + \mathbf{I}_p| \\ &\quad + \mathrm{Tr}\,(\mathbf{V}_{q,k}\mathbf{D}_{q,k}^{-1}\mathbf{V}_{q,k}^\top \\ &\quad + \mathbf{V}_{m,k}\mathbf{V}_{m,k}^\top)\mathbf{\Psi}_k^{-1/2}\mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1/2}\Big\} \\ &= c - \frac{n}{2}\Big\{\log\det\mathbf{\Psi}_k + \sum_{j=1}^{q}\log\theta_{j,k} \\ &\quad + \mathrm{Tr}\,\mathbf{D}_{q,k}^{-1}\mathbf{V}_{q,k}^\top\mathbf{V}_k\mathbf{D}_k\mathbf{V}_k^\top\mathbf{V}_{q,k} \\ &\quad + \mathrm{Tr}\,\mathbf{V}_{m,k}^\top\mathbf{V}_k\mathbf{D}_k\mathbf{V}_k^\top\mathbf{V}_{m,k}\Big\} \\ &= c - \frac{n}{2}\Big\{\log\det\mathbf{\Psi}_k + \sum_{j=1}^{q}\log\theta_{j,k} + \mathrm{Tr}\,\mathbf{D}_{q,k}^{-1}\mathbf{D}_{q,k} \\ &\quad + \mathrm{Tr}\,\mathbf{D}_{m,k}\Big\} \\ &= c - \frac{n}{2}\Big\{\log\det\mathbf{\Psi}_k + \mathrm{Tr}\,\mathbf{\Psi}_k^{-1}\mathbf{\Sigma}_k^* + \sum_{j=1}^{q}\log\theta_{j,k} \\ &\quad - \sum_{j=1}^{q}\theta_{j,k} + q\Big\},\end{aligned} \tag{A8}$$

其中，c 是一个与 $\mathbf{\Psi}_k$ 无关的常数。 □

根据结果 1，我们需要 $\boldsymbol{W}_k$ 中最大的 q 个特征向量。$\boldsymbol{W}_k$ 的部分特征分解结果可以通过 Lanczos 算法（Sorensen 1992；Dutta& Mon-dal 2015）高效获得，该算法仅需通过矩阵-向量乘积处理 $\boldsymbol{W}_k$，从而大幅减少存储需求和计算时间。（下一节将详细介绍 Lanczos 算法。）然后，我们可以通过带盒约束的有限内存 Broyden-Fletcher-Goldfarb-Shanno 拟牛顿算法（L-BFGS-B）(Byrd 等人 1995)，求得 $\mathbf{\Psi}_k^* = \arg\max Q_p(\mathbf{\Psi}_k)$。该算法利用最近几次迭代中的 $Q_p(\mathbf{\Psi}_k)$ 和 $\nabla Q_p(\mathbf{\Psi}_k)$ 的值来近似精确的海森矩阵，从而将存储开销从 $O(p^2)$ 降至 $O(p)$。给定 $\mathbf{\Psi}_k^*$，载荷矩阵更新为 $\mathbf{\Lambda}_k^* = \mathbf{\Psi}_k^{*1/2}\boldsymbol{V}_{q,k}\mathbf{\Delta}_k$。

针对 $\boldsymbol{W}_k$的部分特征分解的 Lanczos 算法

如 Dai等（2021）所示，使用隐式重启 Lanczos 算法（Sorensen 1992）可以求得 $\boldsymbol{W}_k = \mathbf{\Psi}_k^{-1/2}\mathbf{\Sigma}_k^*\mathbf{\Psi}_k^{-1/2}$ 的 q 个最大特征值及其特征向量。假设 $m = \max\{2q + 1, 20\}$，且 $\mathrm{f1}_{,k} \in \mathbb{R}^p$ 是满足 $\|\mathrm{f1}_{,k}\| = 1$ 的任意向量，并初始化 $\mathrm{F1}_{,k} = \mathrm{f1}_{,k\cdot}$。随后，我们采用 Lanczos 迭代（Dutta 与 Mondal 2015），具体如下。对于 $l = 1, 2, \ldots, m$,

- Compute $\boldsymbol{u}_{l,k} = \boldsymbol{W}_k\boldsymbol{f}_{l,k}$ and $\alpha_{l,k} = \boldsymbol{f}_{l,k}^\top\boldsymbol{u}_{l,k}$.
- 计算$\boldsymbol{r}_{l,k} = \boldsymbol{u}_{l,k} - \alpha_{l,k}\boldsymbol{f}_{l,k} - \beta_{l-1,k}\boldsymbol{f}_{l-1,k}$（假设$\beta_{0,k} = 0$且$\boldsymbol{f}_{0,k} = \mathbf{0}$）。

• 设$\beta_{l,\ k}\ =\ \|\boldsymbol{r}_{l,\ k}\|$，并且如果 $l\ <\ q$ 且$\beta_{l,\ k}\ \neq\ 0$，计算$\boldsymbol{f}_{l+1,\ k}=\boldsymbol{r}_{l,\ k}/\beta_{l,\ k}$并令$\boldsymbol{F}_{l+1,\ k}=[\boldsymbol{F}_{l,k},\boldsymbol{f}_{l+1,k}]$。

假设 $\boldsymbol{T}_{m,k}$ 是对称三对角矩阵 $m\times m$，其对角元素为 $\alpha_{1,k},\alpha_{2,k},\ldots,\alpha_{m,k}$，第 j 条副对角线上的元素为 $\beta_{j,k}$，其中 $j\ =\ 1,\ldots,m\ -\ 1$。通过 Sturm 序列算法（Wilkinson，1958），我们计算 $\boldsymbol{T}_{m,k}$ 的特征值 $e_{1,k}\ >\ e_{2,k}\ >\ \cdots\ >\ e_{m,k}$ 及其特征向量 $\boldsymbol{g}_{1,k},\boldsymbol{g}_{1,k},\ldots,\boldsymbol{g}_{m,k}$。另外，令 $\boldsymbol{v}_{j,k}\ =\ \boldsymbol{F}_{m,k}\boldsymbol{g}_{j,k}$，其中 $1\ \leq\ j\ \leq\ m.$。$e_{j,k}$ 称为 $\boldsymbol{W}_k$ 的 Ritz值，$\boldsymbol{v}_{j,k}$ 称为 Ritz向量。可以证明$\|\boldsymbol{W}_k\boldsymbol{v}_{j,k}-\boldsymbol{v}_{j,k}e_{j,k}\|=\beta_{m,k}|g_{m,j,k}|$，其中 $g_{m,j,k}$ 是向量 $\boldsymbol{g}_{j,k}$ 的第 m 个分量，$j=1$ 为 $2,\ldots,m.$。如果

$$\beta_{m,k}\ \max_{1\leq j\leq m}|g_{m,j,k}|<\delta \tag{A9}$$

对于某个预先指定的容差 δ，$e_{1,\ k}$、$e_{2,\ k}$、.....$e_{q,\ k}$ 以及 $\boldsymbol{v}_{1,\ k}$、$\boldsymbol{v}_{2,\ k}$、..... $\boldsymbol{v}_{q,\ k}$ 是 $\boldsymbol{W}_k$的 q 个最大特征值及其对应特征向量的精确近似。

然而，在实践中，Ritz向量和Ritz值需要经过更多次迭代才能收敛到 $\boldsymbol{W}_k$的特征值和特征向量，因此 Sorensen （1992）建议对 Lanczos 算法进行隐式重启，并迭代地对对称三对角矩阵的谱进行移位，以迫使新的残差$\boldsymbol{r}_{m,\ k}$ 趋于零，从而加快收敛速度。因此，我们˜计算 QR分解：$\boldsymbol{T}_{m,\ k}\ -\ e_{j,\ k}\boldsymbol{I}_m\ =\ \boldsymbol{Q}_{j,\ k}\boldsymbol{R}_{j,\ k}$，对于 $j\ =\ q+1$、.....、m，令$\tilde{\boldsymbol{Q}}_k\ =\ \boldsymbol{Q}_{q+1,\ k}\boldsymbol{Q}_{q+2,\ k}\cdots\boldsymbol{Q}_{m,\ k}$，并重置 $\boldsymbol{F}_{m,\ k}=\boldsymbol{F}_{m,\ k}\tilde{\boldsymbol{Q}}_k$和 $\boldsymbol{T}_{m,\ k}=\tilde{\boldsymbol{Q}}_k^{\top}\boldsymbol{T}_{m,\ k}\tilde{\boldsymbol{Q}}_{k\cdot}$。然后

$$\boldsymbol{W}\boldsymbol{F}_{q,k}=\boldsymbol{F}_{q,k}\boldsymbol{T}_{q,k}+\beta_k^*\boldsymbol{f}_{q+1,k}\boldsymbol{e}_q^{\top} \tag{A10}$$

其中，β_k^*是$\boldsymbol{T}_m$的第（$q+1$，q）个元素，$_k\boldsymbol{e}_q$是 $\mathbb{R}^q$中的第 q个标准基向量，而$\boldsymbol{T}_{qk}$ 是$\boldsymbol{T}_{mk}$ 的第 $q\times q$ 个主子矩阵（Sorensen 1992）。因此，(A10)本身就是 $\boldsymbol{W}_{k\cdot}$的 q阶 Lanczos 因子分解。接下来，我们从$l=q+1$，.....，m开始“重启”Lanczos 迭代，而不是从 1 到m，如果满足 (A9) 则终止，否则重新启动算法。

本文由作者准备的 TEX/LaTEX 文件排版。