arXiv:2601.07446v1 [stat.CO] 12 Jan 2026

用于具有组级异质性的事件发生时间数据自适应邻域聚类的惩罚似然优化

AlessandraRagni[1], Lara Cavinato[1], Francesca Ieva[1,2][1]米兰理工大学数学系，MOX，意大利米兰达芬奇广场32号，邮编20133[2]Human Technopole健康数据科学中心，意大利米兰Rita Levi–Montalcini大道1号，邮编20157

摘要

识别事件风险动态相近的患者亚组，对于支持临床研究中的循证决策至关重要。在此类情境下，必须考虑个体嵌套于医院等较高层级单元时产生的内在依赖性。现有生存模型通过脆弱性项处理组级异质性，但无法发现潜在患者亚组；而大多数聚类方法忽略层级结构，且不会与生存结局联合估计。在本研究中，我们提出一个新框架，通过惩罚似然方法同时进行患者聚类和共享脆弱性生存建模。所提出的方法通过谱聚类的一种改进版本，自适应地学习患者间相似度矩阵，从而在考虑组成员关系的同时，直接根据估计的风险特征形成聚类。模拟研究表明，该模型能够恢复潜在聚类并正确估计风险参数。我们将该方法应用于2020至2021年间在伦巴第大区（意大利）住院治疗的一大批COVID-19心力衰竭患者，识别出具有不同风险特征且临床意义显著的亚组，并揭示了呼吸系统合并症和医院层面变异性对死亡结局的影响。该框架为层级数据情境下基于风险的患者分层提供了一种灵活且易于解释的工具。

关键词：脆弱性模型、聚类、惩罚、异质性、生存分析

1 引言

在临床和流行病学研究中产生的真实世界事件发生时间研究里，识别具有相似风险轨迹的患者，有助于将其划分为具有共同关键结局（如死亡、复发、治疗失败）发生可能性的亚组或簇。这种分层方法适用于多种应用领域，在此情境下可支持临床决策、促进疾病亚型划分，并推动精准医疗计划（Collins 和 Varmus，2015）。然而，患者之间往往存在内在依赖关系，因为他们可能属于同一组——例如在同一家医院接受治疗——从而导致结局相关性以及额外的变异性。在分析中必须考虑这些因素，因为它们可能以复杂的方式影响患者的风险特征。

这种情形在医疗保健应用中很常见，我们的研究正是受到 *Enhance-Heart* 项目的启发。该研究在意大利北部的区域行政区（伦巴第大区）开展。新冠肺炎疫情早期，SARS-CoV-2 感染失控传播，导致该地区受到严重影响（Alteri 等，2021）。研究聚焦于心力衰竭（HF）患者，旨在探究患者临床特征与 COVID-19 相关死亡率之间的复杂关系，同时考量不同医院之间的差异。事实上，近期研究指出，疫情期间心力衰竭与 COVID-19 之间存在显著的相互作用：既有心力衰竭会使 COVID-19 的管理和预后更加复杂，并可能引发并发症（Bader 等，2021；Adeghate 等，2021）。因此，在评估呼吸系统疾病和整体健康状况如何影响患者各簇的死亡风险时，识别因 COVID-19 住院的心力衰竭患者中的潜在亚组，是在事件发生时间框架下刻画结局的关键一步。另一方面，医院的资源可用情况，例如重症监护病房床位以及内科/外科普通病床，被发现与

COVID-19 死亡发生率显著相关（Janke 等，2021）。因此，要准确评估这一背景下的生存结局，需要采用能够对患者群体进行分层、同时考量不同医院之间差异的建模方法。

传统生存模型通常通过脆弱性项处理由组级依赖产生的异质性。脆弱性项是风险函数中的一个乘法因子；风险函数表示在已存活至某一时点的条件下，于该时点发生事件的风险（Vaupel 等，1979；Clayton，1978）。该项可采用不同方式设定，既可以采用参数方法（最常见的是使用伽马分布或对数正态分布），也可以采用半参数方法（例如，参见 Abrahantes 等（2007）；Austin（2017）；Balan 和 Putter（2020））。然而，这些模型通常假设组内同质，并忽略了患者之间由其在特定时点发生目标事件的风险所导致的潜在相似模式。相反，无监督分层方法可以利用个体间的相似性揭示隐藏的患者亚组；但由于这类方法与事件发生时间结局的整合有限，其预测性能往往并非最优。例如，Ahlqvist 等（2018）基于六项标准化临床指标进行了聚类分析，识别出具有不同预后和并发症特征、且可重复识别的糖尿病亚组。为推进生存信息引导聚类，Tosado 等（2020）提出先根据各特征与生存的关联（通过鞅残差）变换特征空间，再在变换后的空间中进行聚类。类似地，Chen 等（2016）提出了一种递归分割算法，通过最大化患者组之间的生存差异，构建癌症预后分期系统。转向预测这一重点，Chapfuwa 等（2020）开发了一种贝叶斯非参数框架，将受试者嵌入聚类潜在空间，从而实现准确的事件发生时间预测，并识别具有不同风险特征的亚组。沿着相关思路，Manduchi 等（2021）提出了一种基于变分自编码器的生存聚类方法，同时利用患者特征和事件发生时间结局；而 Buginga 和 de Souza e Silva（2024）提出了一种非参数专家混合模型，其中每个专家刻画一个不同亚组的生存分布，从而揭示异质风险模式。据我们所知，同时解决患者聚类与事件发生时间结局建模、并纳入层级结构的统计模型相对较少。其中，Caldera 等（2025）的聚类加权模型是较新的研究成果：该模型在每个簇内拟合共享脆弱性模型，并假设协变量在不同簇中服从不同的参数分布，同时通过聚类特异性混合权重对整体联合密度进行因子分解。尽管这一方法颇具前景，但其对协变量强分布假设的依赖可能会限制模型的灵活性。

为弥补这一空白，我们提出了一种新模型，将参数化共享脆弱性模型扩展到能够考虑组成员关系（例如，在同一家医院接受治疗的患者），同时通过惩罚似然优化估计患者聚类成员关系（由相似的风险轨迹定义）和风险函数参数，并通过迭代过程获得最终解。患者分层采用修改后的谱聚类方法实现，遵循 Nie 等人（2014，2016）和 Guo（2015）提出的策略：自适应地估计相似度矩阵，以捕捉患者间关系，并表示基于估计的风险函数风险（结合脆弱性和协变量效应）判断患者属于 $neighbors$ 的概率。该过程最后依据图论原理，确保所得图（由学习得到的相似度矩阵表示）的连通分量恰好对应簇数。

我们的模型建立在 Liu 等人（2020）提出的框架之上。该框架最初用于癌症亚型划分，通过将惩罚项嵌入 $partial$ 基于似然的优化方案，将有监督图聚类整合到多组学肿瘤数据建模中。Cavinato 等人（2021，2023）已在影像组学领域探索了这一方法的相关扩展与应用。

该框架的一项关键优势在于，患者聚类直接嵌入优化过程中，因此簇的分配由潜在的风险动态引导，并反映出不同的事件风险特征，从而可从事件风险的角度加以解释。基于这一框架，我们的贡献主要有两方面：

（i）我们在统一的生存建模框架下，将组层面（医院）的依赖性与患者层面异质性纳入模型；（ii）我们提出一种惩罚 $full$似然方法，将患者聚类直接嵌入风险模型，从而能够以风险驱动方式学习相似性结构，并生成在事件风险方面具有可解释性的簇。

本文的组织结构如下。第2节概述所提出的方法论，详细介绍符号和估计程序。第3节介绍一项模拟研究，用于评估所提方法论的性能。第4节描述促成本研究的 *Enhance-Heart* 行政数据库，并报告使用我们的方法获得的主要结果。最后，第5节给出结论性意见，讨论所提方法的优势与局限，并概述未来工作的方向。

该方法通过 R 软件实现（R核心团队（2022年）），可在 https://github.com/alessandragni/HazClust 免费获取。

2 方法论

本节按以下四个连续步骤概述方法论：回顾参数化脆弱性模型的符号表示（第2.1小节）、模型描述（第2.2节）、模型优化（第2.3节），以及估计程序的相关细节（第2.4节）。

2.1 符号表示

设观测数据为 $\mathcal{Z} = \{(y_{gi},\ \delta_{gi},\ \boldsymbol{x}_{gi}) \mid i = 1,\ \ldots,\ n_g,\quad g = 1,\ \ldots, M\}$，其中个体 $i = 1,\ \ldots,\ n_g$ 嵌套于组 $g = 1,\ \ldots,\ M$ 中，个体总数为 $N = \sum_{g=1}^{M} n_g$。对于每个个体，$y_{gi} = \min(t_{gi},\ C^*_{gi})$ 表示观测到的事件发生时间或删失时间（以先发生者为准），而 $\delta_{gi} = \mathbb{I}(t_{gi} \le C^*_{gi})$ 表示事件是否被观测到（$\delta_{gi} = 1$）或被删失（$\delta_{gi} = 0$）。协变量向量 $\boldsymbol{x}_{gi} \in \mathbb{R}^{d\times 1}$ 可包含组级和个体特异的组成部分。我们假设删失独立且为非信息性删失（Munda and Legrand，2014）。为对组 g 中个体 i 在时间 t 的风险率进行建模，并考虑组间存在异质性，在多种可能的方法（例如固定效应、分层）中，我们采用共享脆弱性模型（Duchateau和Janssen，2008）

$$h_{gi}(t \mid u_g) = h_0(t;\boldsymbol{\psi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \tag{1}$$

其中，$\boldsymbol{\beta} \in \mathbb{R}^{d\times 1}$ 是固定效应系数向量，u_g 是随机变量的实际取值，其概率密度为 $f_U(\cdot \mid \theta)$，即脆弱性分布；$h_0(t;\boldsymbol{\psi})$ 是基线风险，假设为参数形式，并依赖参数向量 $\boldsymbol{\psi}$；$'$ 表示转置。由此可得累积风险 $H_{gi}(t \mid u_g) = u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) H_0(t;\boldsymbol{\psi})$，其中$H_0(t;\boldsymbol{\psi}) = \int_0^t h_0(s;\boldsymbol{\psi})ds$。根据附录A中的推导，边际对数似然可表示为

$$\ell(\boldsymbol{\psi},\boldsymbol{\beta},\theta,\mathcal{Z}) = \sum_{g=1}^{M} \left[\left[\sum_{i=1}^{n_g} \delta_{gi} \left(\log(h_0(y_{gi};\boldsymbol{\psi})) + \boldsymbol{x}'_{gi}\boldsymbol{\beta} \right) \right] + \log \left[(-1)^{d_g} \mathcal{L}^{(d_g)} \left(\sum_{i=1}^{n_g} H_0(y_{gi};\boldsymbol{\psi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \right) \right] \right] \tag{2}$$

其中，$d_g := \sum_{i=1}^{n_g} \delta_{gi}$ 是组 g 中的事件数，$(-1)^k \mathcal{L}^{(k)}(s) := (-1)^k \frac{d^k}{ds^k} \mathcal{L}(s) = \mathbb{E}[U^k \exp(-Us)]$，为 $\mathcal{L}(s)$ 拉普拉斯变换，其中 $s \ge 0$（Munda and Legrand, 2014；Klein, 1992）。

最后，已经能够对脆弱性 $u_g = \mathbb{E}(U \mid \mathcal{Z}_g$ 进行预测；$\hat{\boldsymbol{\psi}}, \hat{\boldsymbol{\beta}}, \hat{}$ 通过及观测数据，

$$\mathbb{E}(U \mid \mathcal{Z}_g; \boldsymbol{\beta},\boldsymbol{\psi},\theta) = -\frac{\mathcal{L}^{(d_g+1)}\left(\sum_{i=1}^{n_g} H_0(y_{gi};\boldsymbol{\psi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \right)}{\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g} H_0(y_{gi};\boldsymbol{\psi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) \right)}. \tag{3}$$

2.2 模型

在参数化设定下，例如 Munda 等人（2012）所研究的情形，其中脆弱性和基线风险均服从指定的参数分布，模型参数 $\boldsymbol{\beta}$，$\boldsymbol{\psi}$，θ 可通过最大化式（2）中的对数似然，或最小化负对数似然来估计，如下所示

$$\min_{\boldsymbol{\beta},\boldsymbol{\psi},\theta} \quad -\ell(\boldsymbol{\psi},\boldsymbol{\beta},\theta,\mathcal{Z}). \tag{4}$$

为了在学习 $\boldsymbol{\beta}$、$\boldsymbol{\psi}$ 和 θ 的同时对个体进行聚类，令 $\mathbf{S} \in \mathbb{R}^{N\times N}$ 为亲和矩阵，其元素 $\{s_{gi,hj}\}$ 表示组 g 中的个体 i 与组 h 中的个体 j 之间的相似性。具体而言，我们将每个元素 $\{s_{gi,hj}\}$ 解释为组 g 中的个体 i 以组 h 中的个体 j 为邻居的概率，并施加约束 $\sum_{h=1}^{M}\sum_{j=1}^{n_h} s_{gi,\,hj} = 1,\ \forall g、\ i$，以确保

每个个体的连接概率之和为一。假设个体之间越相似，成为邻居的概率越高，或者等价地说，距离越小 $d_{gi,hj}(\boldsymbol{\psi}、\boldsymbol{\beta}、\theta、\mathcal{Z})$，我们通过进一步对联合估计 $\boldsymbol{\beta}、\boldsymbol{\psi}、\theta$ 和 $\mathbf{S}$ 的对数似然施加惩罚，重写式（4）中的优化问题：

$$\begin{aligned}\min_{\boldsymbol{\beta},\boldsymbol{\psi},\theta,\mathbf{S}} \quad & -\ell(\boldsymbol{\psi},\boldsymbol{\beta},\theta,\mathcal{Z})+\gamma\sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left(d_{gi,hj}(\boldsymbol{\beta},\theta,\mathcal{Z})\,s_{gi,hj}+\mu\,s_{gi,hj}^2\right)\\ \text{s.t.} \quad & \sum_{h=1}^{M}\sum_{j=1}^{n_h}s_{gi,hj}=1,\quad \mathbf{s}_{gi}\geq 0,\quad \forall\, i=1,\ldots,n_g,\ g=1,\ldots,M\end{aligned} \tag{5}$$

其中，$\mathbf{s}_{gi}\in\mathbb{R}^{1\times N}$ 是$\mathbf{S}$ 中与组 g内个体 i 对应的行，所报告的不等式是逐元素成立的。式（5）中的最后两项遵循自适应邻居聚类的相关文献，例如 Nie 等（2014）。这两项用于调节$\mathbf{S}$中邻域概率的结构：第一项鼓励邻近个体具有较高的概率；若仅有这一项，则会趋向于极限情形，即最近邻的概率为 1，其余个体的概率为 0。二次项则促进更平滑、分布更均匀的分配，其极限情形是为所有个体赋予相等的概率 $1/N$（Nie 等人，2014）。超参数 γ控制惩罚的强度；当 $\gamma=0$时，不施加惩罚，因此不会诱导出聚类，因为惩罚项及其相关约束均消失。相比之下，μ 是一个可调参数，用于调节相似性结构对整体优化的影响。其选择将在第 2.3 节和第 2.4 节中讨论。

由于我们的目标是学习$\mathbf{S}$，其依据是患者 i 在组 g中发生目标事件的风险与患者 j 在组 h中的风险之间的相似性，我们在对数风险率尺度上将其量化为 $\log\, h_{gi}(t\mid u_g)=\log\, h_0(t;\boldsymbol{\psi})+\log\, u_g+\boldsymbol{x}_{gi}'\boldsymbol{\beta}$和 $\log\, h_{hj}(t\mid u_h)$，其中各个体共享的 $\log\, h_0(t;\boldsymbol{\psi})$ 在成对比较中会相互抵消。因此，我们定义

$$d_{gi,hj}(\boldsymbol{\beta},\theta,\mathcal{Z}):=\left\|(\boldsymbol{x}_{gi}'\boldsymbol{\beta}+\log u_g)-(\boldsymbol{x}_{hj}'\boldsymbol{\beta}+\log u_h)\right\|^2, \tag{6}$$

其中，$\|\cdot\|$ 为欧氏距离，u_g 和 u_h 为式（3）中定义的脆弱性项。由此得到的距离能够促使相似性图反映经脆弱性调整后的风险一致性。

为将数据聚类为 C 个簇，我们将$\mathbf{S}\in\mathbb{R}^{N\times N}$视为一个图的邻接矩阵，其中 N 个单元作为节点。在此设定下，我们定义拉普拉斯矩阵$\mathbf{L_S}:=\mathbf{D_S}-(\mathbf{S}'+\mathbf{S})/2$，其中$\mathbf{D_S}\in\mathbb{R}^{N\times N}$为对角度矩阵，其第 gi个对角元素为 $\sum_{hj}(s_{gi,hj}+s_{hj,gi})/2$。根据 Mohar 等（1991），拉普拉斯矩阵$\mathbf{L_S}$的零特征值的重数 C对应于相似度矩阵$\mathbf{S}$所定义图中的连通分量数量。因此，施加 $rank(\mathbf{L_S})=N-C$会促使图划分为 C 个簇。此外，正如 Nie 等（2014）和 Fan（1949）所示，该秩约束可以通过引入矩阵$\mathbf{F}\in\mathbb{R}^{N\times C}$以等价且更简便的方式处理，其中

$$2Tr(\mathbf{F}'\mathbf{L_S}\mathbf{F})=\sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\|\mathbf{f}_{gi}-\mathbf{f}_{hj}\|^2 s_{gi,hj}, \tag{7}$$

即$\mathbf{f}_{gi}\in\mathbb{R}^{C\times 1}$ 中与个体 i 所属$\mathbf{F}$ 的簇 g相对应的行。

结合 (2)、(5) 和 (6)，可得如下优化问题

$$\begin{aligned}\min_{\boldsymbol{\beta},\boldsymbol{\psi},\theta,\mathbf{S},\mathbf{F}} \quad & -\left\{\sum_{g=1}^{M}\left[\left[\sum_{i=1}^{n_g}\delta_{gi}\left(\log(h_0(y_{gi};\boldsymbol{\psi}))+\boldsymbol{x}_{gi}'\boldsymbol{\beta}\right)\right]+\log\left[(-1)^{d_g}\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\boldsymbol{\psi})\exp(\boldsymbol{x}_{gi}'\boldsymbol{\beta})\right)\right]\right]\right\}\\ & +\gamma\left\{\sum_{g,h=1}^{M}\sum_{i=1}^{n_g}\sum_{j=1}^{n_h}\left[\left\|(\boldsymbol{x}_{gi}'\boldsymbol{\beta}+\log u_g)-(\boldsymbol{x}_{hj}'\boldsymbol{\beta}+\log u_h)\right\|^2 s_{gi,hj}+\mu\,s_{gi,hj}^2\right]\right.\\ & \left.+2\lambda Tr(\mathbf{F}'\mathbf{L_S}\mathbf{F})\right\}\\ \text{s.t.} \quad & \sum_{h=1}^{M}\sum_{j=1}^{n_h}s_{gi,hj}=1,\quad \mathbf{s}_{gi}\geq 0,\quad \forall\, i=1,\ldots,n_g,\ g=1,\ldots,M,\quad \mathbf{F}'\mathbf{F}=\mathbf{I},\ \mathbf{F}\in\mathbb{R}^{N\times C}\end{aligned} \tag{8}$$

其中

$$\log u_g=\log\left[(-1)^{d_g+1}\mathcal{L}^{(d_g+1)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\boldsymbol{\psi})\exp(\boldsymbol{x}_{gi}'\boldsymbol{\beta})\right)\right]-\log\left[(-1)^{d_g}\mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g}H_0(y_{gi};\boldsymbol{\psi})\exp(\boldsymbol{x}_{gi}'\boldsymbol{\beta})\right)\right],$$

whi通过注意到 $-1=(-1)^{d_g+1}/(-1)^{d_g}$ 并应用对数性质，可由式 (3) 得出 ch y.

2.3 优化过程

为求解式 (8) 中的优化问题，我们采用块坐标下降方法：在每次迭代中，更新一个变量块，同时固定其他变量块，直至满足收敛准则（详见第 2.4 小节）。该过程在以下三个步骤之间交替进行：

1. **固定 S、 $\boldsymbol{\beta}, \boldsymbol{\psi}, \theta$ 并更新 F**
   固定**S**、 $\boldsymbol{\beta}$、 $\boldsymbol{\psi}$和 θ 后，更新**F** 可归结为以下谱优化问题（Nie 等人，2014）：

$$\begin{aligned} \min_{\mathbf{F}} \quad & 2\lambda Tr(\mathbf{F}^T \mathbf{L}_{\mathbf{S}} \mathbf{F}) \\ \text{s.t.} \quad & \mathbf{F}^T \mathbf{F} = \mathbf{I}, \mathbf{F} \in \mathbb{R}^{N \times C} \end{aligned}$$

   因子 2λ 不会影响解，因此可以省略。最优**F** 由拉普拉斯$\mathbf{L}_{\mathbf{S}}$ 对应的 C 个最小特征值的特征向量构成。此步骤需要簇数 C 以及**S**的当前估计值。

2. **固定 F**,**S 并更新** $\boldsymbol{\beta}, \boldsymbol{\psi}, \theta$
   固定**F** 和**S** 后，对（$\boldsymbol{\beta}, \boldsymbol{\psi}, \theta$）进行优化可归结为一个带有额外基于相似性的正则化项的生存建模问题。该优化问题可表示为：

$$\min_{\boldsymbol{\beta}, \boldsymbol{\psi}, \theta} \quad -\ell(\boldsymbol{\beta}, \boldsymbol{w}, \theta, \mathcal{Z}) + \gamma \sum_{g,h=1}^{M} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \left(\left\| \boldsymbol{x}'_{gi} \boldsymbol{\beta} + \log u_g - \boldsymbol{x}'_{hj} \boldsymbol{\beta} - \log u_h \right\|^2 \right) s_{gi,hj}$$

   我们遵循 Munda 等人（2012）的方法，通过采用参数化脆弱性的最大似然优化来估计生存模型参数，并将这一框架扩展为加入基于相似性的惩罚项。此步骤需要 γ、当前的相似度矩阵 **S**、数据集和模型设定。

3. **固定 F** $\boldsymbol{\beta}, \boldsymbol{\psi}, \theta$ **并更新 S**
   固定 **F**、 $\boldsymbol{\beta}$、 $\boldsymbol{\psi}$ 和 θ 后，通过求解以下问题来更新**S** ：

$$\begin{aligned} \min_{\mathbf{S}} \quad & \gamma \Bigg\{ \sum_{g,h=1}^{M} \sum_{i=1}^{n_g} \sum_{j=1}^{n_h} \left[\left\| \boldsymbol{x}'_{gi} \boldsymbol{\beta} + \log u_g - \boldsymbol{x}'_{hj} \boldsymbol{\beta} - \log u_h \right\|^2 s_{gi,hj} + \mu s_{gi,hj}^2 \right] \\ & + 2\lambda Tr(\mathbf{F}' \mathbf{L}_{\mathbf{S}} \mathbf{F}) \Bigg\} \\ \text{s.t.} \quad & \sum_{h=1}^{M} \sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad 0 \leq \mathbf{s}_{gi} \leq 1, \quad \forall\, i = 1, \ldots, n_g,\ g = 1, \ldots, M \end{aligned}$$

   利用关系式 (7) 以及问题在各个体间的可分性，这可简化为针对每个$\mathbf{s}_{gi}$分别进行独立优化：

$$\begin{aligned} \min_{\mathbf{s}_{gi}} \quad & \sum_{h=1}^{M} \sum_{j=1}^{n_h} \left(\left\| \boldsymbol{x}'_{gi} \boldsymbol{\beta} + \log u_g - \boldsymbol{x}'_{hj} \boldsymbol{\beta} - \log u_h \right\|^2 + \mu s_{gi,hj} + \lambda \| \mathbf{f}_{gi} - \mathbf{f}_{hj} \|^2 \right) s_{gi,hj} \\ \text{s.t.} \quad & \sum_{h=1}^{M} \sum_{j=1}^{n_h} s_{gi,hj} = 1, \quad \mathbf{s}_{gi} \geq 0 \end{aligned} \tag{9}$$

   如附录 B 所示，对于每一行$\mathbf{s}_{gi}$，该问题都有闭式解：

$$s_{gi,hj} = \max\left\{0, \alpha_{gi} - w_{gi,hj}/(2\mu_{gi})\right\}$$

   以$\mathbf{s}'_{gi}\mathbf{1} = 1$为约束，其中 $\alpha_{gi} \geq 0$ and 和 μ_{gi} 可能因个体而异。这里， $\boldsymbol{w}_{gi} \in \mathbb{R}^{N \times 1}$ and

$$w_{gi,hj} := \left\| \boldsymbol{x}'_{gi} \boldsymbol{\beta} + \log u_g - \boldsymbol{x}'_{hj} \boldsymbol{\beta} - \log u_h \right\|^2 + \lambda \| \mathbf{f}_{gi} - \mathbf{f}_{hj} \|^2.$$

   通过对$\mathbf{s}_{gi}$施加稀疏性约束，即仅允许 k 个最近邻在组g中与个体 i 相连，按照 Nie 等（2014）；Guo（2015）的方法，我们施加额外约束，使问题适定，同时降低计算成本。令 $\hat{}\boldsymbol{w}_{gi} = (\hat{}w_{gi,1}, \ldots, \hat{}w_{gi,N-1})$表示排序后的距离（不包括自身距离$w_{gi,gi}$）。则最优的有序相似度向量 $\hat{}\mathbf{s}_{gi}$被约束为恰好具有 k 个非零项：

(i) $\hat{s}_{gi,p} > 0$，对于 $p = 1,\ldots,k$，对应于 k 个最近邻；(ii) $\hat{s}_{gi,p} = 0$，对于 $p = k+1,\ldots,N$，包括自相似性。

在这些条件和约束 $\mathbf{s}_{gi}'\mathbf{1} = 1$ 下，如附录 C 所示，我们得到

$$\alpha_{gi} = \frac{1}{k} + \frac{1}{2k\mu_{gi}} \sum_{p=1}^{k} \hat{w}_{gi,p} \quad \text{and} \quad \mu_{gi} = \frac{k}{2}\hat{w}_{gi,(k+1)} - \frac{1}{2}\sum_{p=1}^{k} \hat{w}_{gi,p}. \tag{10}$$

最后，所有行更新完毕后，我们设置 $\mu = \sum_{g,\ i} \mu_{gi}/N$。

此步骤需要 λ, k、$\mathbf{F}$ 的当前估计、数据集和模型设定。

2.4 估计细节

本节介绍所提出算法在估计和实现方面的实践细节与技术要点。

初始化。如果未提供先验信息，则将相似度矩阵 $\mathbf{S}$ 初始化为对角线元素为零，所有非对角线元素为 $1/(N-1)$，使每一行的和为 1。也可以通过 S0 提供初始相似度矩阵。默认将惩罚参数 λ 设为 $\lambda_0 = 1000$；文献建议将其设得足够大（Nie 等人，2014），但用户也可以通过参数 lambda0 指定。首次更新 $\mathbf{S}$ 后，将 λ 重置为通过闭式解得到的 μ 值，然后进行自适应调整：如果 $\mathbf{S}$ 中的连通分量数小于 C，则将其增大（乘以 1.2）；否则将其减小（除以 1.2）（Nie 等人，2017）。

回归参数和脆弱性参数可以通过 inip 和 iniFpar 选项根据用户输入进行初始化，也可以采用 parfm R 包中的默认值（Munda 等人，2012）。与该软件包一样，我们的软件支持多种参数基线风险分布（威布尔分布、指数分布、Gompertz 分布、对数正态分布等），以及多种脆弱性分布（伽马分布、对数正态分布、正稳定分布等）。

收敛。当算法同时稳定相似度矩阵和模型拟合时，即认为达到收敛。具体而言，当以下所有条件均得到满足时，过程停止：(i) $\mathbf{S}$的平均绝对变化小于 tolS，表明图结构已经稳定；(ii) 对数似然的改进小于 t olll，表明统计拟合已趋于平稳；(iii) $\mathbf{S}$ 中的连通分量数量等于目标值 C。这些标准将联合进行评估；当全部条件均满足，或达到 maxit 次迭代的最大次数时，算法终止。此外，第 2 步中的内部优化依赖于 parfm，最多可进行 maxitparfm 次迭代。默认值设为 tolS $= 10^{-4}$、tolll $= 10^{-3}$、maxit $= 500$、maxitparfm $= 500$。

选择 k 及其与 C的关系。最近邻的数量 k 控制初始相似度矩阵 $\mathbf{S}$ 的稀疏程度，从而决定相应图的密度。具体而言，每一行都恰好包含 k 个非零项，由此强制形成固定的局部邻域结构，并排除自相似性（将$\mathbf{S}$的对角线设为零）。

选择 k 会显著影响所构造的图达到 C 个连通分量这一目标的难易程度。如果 k 过小，图可能会碎裂成超过 C 个分量；而较大的 k 会生成稠密图，往往使结构合并为少于 C 个分量。我们的算法通过自适应调整惩罚参数 λ 来应对这种相互作用，从而确保收敛到预设数量 C。不过在实践中，应选择 k 以保证基本连通性，同时不牺牲稀疏性，并根据具体应用仔细调节。常见的启发式做法是让 k 随样本量适度增长。

模型选择与 C的选择。为了评估聚类解的质量，我们采用轮廓系数（Kaufman 和 Rousseeuw, 2009）。给定一个个体 i，令 $a(i)$ 表示其与同一簇中所有其他个体之间的平均距离，令 $b(i)$ 表示从 i 到其他任一簇的最小平均距离。回顾一下，轮廓值定义为

$$s(i) = \frac{b(i) - a(i)}{\max\{a(i),\, b(i)\}}, \quad s(i) \in [-1, 1],$$

数值越高表示分离效果越好。在我们的设定中，个体 i 相对于单元 j 的距离计算为 $\sqrt{d_{gi,\ hj}(\boldsymbol{\beta},\ \theta,\ \mathcal{Z})}$，如公式（6）所示，因为该距离纳入了协变量效应和估计的脆弱性贡献。组索引（例如 i 的 g 和 j 的 h）在此处省略，因为它们与轮廓系数的计算无关。随后，我们使用平均轮廓系数评估总体聚类质量，

并将平均轮廓系数最大的配置选为最优聚类配置，从而相对于拟合的生存模型实现最佳分离效果。

软件。该算法使用 R 实现，可在以下代码仓库中获取：https://github.com/alessandragni/HazClust。核心生存分析和脆弱性估计例程依赖 parfm 软件包（Munda 等人，2012），该软件包提供参数化脆弱性模型的最大似然估计。我们的工作将 parfm 嵌入迭代式基于图的聚类框架中，对其进行了扩展。

3 模拟研究

本节通过模拟研究评估所提出方法的性能。我们主要关注该方法在不同惩罚水平下恢复潜在簇结构的效果。我们还考察不同惩罚强度下参数估计值的表现。此外，当潜在结构包含 C 个簇时，我们评估该方法相较于其他 C 选择识别真实聚类的准确性；平均轮廓系数则可作为聚类质量的综合指标。

3.1 数据生成过程

我们通过构造 $M = 10$ 个组来生成数据，每个组包含 $n_g = 50$（$g = 1$、.....、M），总样本量为 $N = 500$。为引入同一组内个体之间的相关性，我们设置组级脆弱性 u_g，并为每个组独立地从伽马分布 $u_g \sim \text{Gamma}(1/\theta, 1/\theta)$ 中抽取该脆弱性，其中 $\theta = 0.5$，使得 $\mathbb{E}[u_g] = 1$ 且 $\text{Var}(u_g) = \theta$。保持较小的脆弱性方差可确保聚类特异性效应仍是异质性的主要来源。关于这一点的进一步阐述将在最后的讨论中提供。

在每个组 g 内，每个个体 i 都会被随机分配到 $C = 3$ 个潜在聚类中的一个，记为 $z_{gi} \in \{1, 2, 3\}$。每个个体都与一个协变量向量 $\boldsymbol{x}_{gi}$ 和一个线性预测子 $\eta_{gi} = \boldsymbol{x}'_{gi}\boldsymbol{\beta}+\log\ u_g$ 相关，其中 $\boldsymbol{\beta}$ 是回归系数向量；因此，在给定簇分配 z_{gi} 和脆弱性 u_g 的条件下，线性预测子建模为

$$\eta_{gi} \mid z_{gi} = c, u_g \sim \mathcal{N}(\mu_c, \sigma^2), \quad c = 1, 2, 3$$

其中 $[\mu_1, \mu_2, \mu_3] = [-8, 0, 8]$ 和 $\sigma = 1$。该设定生成了跨簇明显分离的 η_{gi} 分布，从而最大化了式（6）所衡量的聚类结构。最后，为保持跨簇预期的模拟结构，我们构造单个观测协变量 x_{gi} $x_{gi} = (\eta_{gi} - u_g)/\beta$ $\beta =$ 2 为 log，并设定 log( )。

生存时间是在威布尔基线风险下生成的，其参数通过 $\boldsymbol{\psi} = [\rho, \psi]$ 至 $H_0(t) = \xi^\rho t^\rho$（ρ，$\xi > 0$）进行参数化，其中 $\rho = 2.5$ 和 $\xi = 0.01$。相应的累积风险函数的逆函数为 $H_0^{-1}(s) = s^{1/\rho}/\xi$。事件时间通过逆变换抽样生成：我们抽取 $V_{gi} \sim \text{Uniform}(0, 1)$，并设定 $t_{gi} = H_0^{-1}\big(-\log(V_{gi})/\exp(\eta_{gi})\big)$。

删失采用以下两种可能的机制之一施加：（i）在固定时间 100 处进行 *administrative* 删失；（ii）采用均值为 130、标准差为 15 的 *Gaussian (normal)* 删失方式，类似于 frailtySurv R 包中 genfrail 函数的默认选项（Monaco 等，2018）。

在这种数据生成机制下，每组生成数据的 Kaplan–Meier 曲线清晰地呈现出三个簇之间的不同模式：一个簇保持较高的生存率，另一个簇逐渐下降，第三个簇则在早期急剧下降。

遵循上述数据生成过程，我们设定

$$\gamma \in \{0, 10^{-6}, 10^{-4}, 10^{-3}, 10^{-2}, 0.1, 0.2, 0.4\}.$$

情形 $\gamma = 0$ 表示没有进行惩罚，因此不执行聚类，只拟合一个与 parfm 中相同的参数化脆弱性模型。当 $\gamma > 0$时，我们设定 $C \in \{2, 3, 4, 5\}$ 和 $k = \{20, 50\}$。对于每种设定，我们将模拟重复 $B = 100$ 次。每次重复模拟中使用相同的模拟数据集，同时改变模型的输入参数，以评估模型在不同配置下的表现。此外，我们设定 tolS $= 10^{-2}$、tolll $= 1$ 和 maxit $= 200$。

3.2 结果

我们首先关注 $C = 3$ 和 $\gamma > 0$的模拟。对于 $k = 20$ 和 $k = 50$，该方法始终能识别出三个簇，性能开始下降时对应的 γ 阈值仅有细微差异。在 *administrative* 删失条件下，对于 $\gamma \leq 0.1$，该方法识别出三个簇的频率为 100%。当 γ较大时，识别率略有下降，在 $\gamma = 0.4$时达到最低值 95%。总体而言，平均经过 17.49 次迭代后即可收敛。在 *normal*删失条件下，对于 $\gamma \leq 10^{-3}$，该方法始终能准确识别簇；此后识别率略微逐渐下降，在 $\gamma = 0.4$时达到最低值 98%。在第二种情形下，该算法平均需要 24.49 次迭代才能收敛。

现在仅考虑正确识别出三个簇的情形，使用准确率和调整兰德指数（ARI）评估聚类性能，其中模拟设计已知每个个体 i 在组 g 中的真实簇成员归属。准确率衡量经过最优标签对齐后，预测簇标签与真实标签相符的观测值比例。调整兰德指数衡量经随机机会校正后的成对一致性，取值范围为 $[-1, 1]$；1 表示完美恢复，0 表示随机聚类，负值则表示一致性低于随机机会。表 1 汇报了不同模型配置下的结果，即改变删失机制、 k和 γ所得的结果。

删失	k	γ	**准确率**			**ARI**		
			Mean	中位数	SD	Mean	中位数	SD
(i) 行政	20	10^{-6}	0.971	1.000	0.089	0.942	1.000	0.161
		10^{-4}	0.971	1.000	0.089	0.942	1.000	0.161
		10^{-3}	**0.975**	**1.000**	**0.081**	**0.950**	**1.000**	**0.147**
		10^{-2}	0.957	1.000	0.082	0.907	1.000	0.172
		0.1	0.950	1.000	0.083	0.888	1.000	0.179
		0.2	0.930	1.000	0.115	0.860	1.000	0.197
		0.4	0.892	0.928	0.141	0.788	0.819	0.242
	50	10^{-6}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		10^{-4}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		10^{-3}	**1.000**	**1.000**	**0.001**	**0.999**	**1.000**	**0.003**
		10^{-2}	1.000	1.000	0.001	0.999	1.000	0.004
		0.1	0.996	1.000	0.023	0.980	1.000	0.054
		0.2	0.945	0.998	0.120	0.897	0.994	0.199
		0.4	0.420	0.390	0.109	0.060	0.019	0.159
(ii) 正态	20	10^{-6}	0.902	0.998	0.142	0.810	0.994	0.246
		10^{-4}	0.904	0.998	0.138	0.810	0.994	0.251
		10^{-3}	0.891	0.970	0.141	0.788	0.918	0.243
		10^{-2}	0.896	0.998	0.149	0.807	0.994	0.239
		0.1	0.913	1.000	0.117	0.820	1.000	0.224
		0.2	**0.922**	0.998	**0.097**	**0.831**	0.994	**0.195**
		0.4	0.870	0.894	0.145	0.752	0.732	0.237
	50	10^{-6}	0.990	1.000	0.040	0.976	1.000	0.094
		10^{-4}	0.990	1.000	0.040	0.977	1.000	0.094
		10^{-3}	0.990	1.000	0.041	0.976	1.000	0.096
		10^{-2}	**0.998**	**1.000**	**0.016**	**0.994**	**1.000**	**0.034**
		0.1	0.994	1.000	0.027	0.986	1.000	0.060
		0.2	0.964	0.998	0.089	0.922	0.994	0.183
		0.4	0.417	0.386	0.102	0.056	0.015	0.148

表 1： *accuracy* 和 *ARI* 在 $B = 100$ 次重复试验中的经验均值、中位数和标准差，在两种删失机制下以及不同的 k 和 γ取值下进行评估，设置为 $C = 3$。每种设置下的最佳估计值以粗体突出显示。

聚类性能通常很高，尤其是对于较大的 k，准确率和调整兰德指数（ARI）经常接近 1。具体而言，在大多数设置下，中位数都接近 1，这反映了性能指标在上界附近呈偏态分布；相比之下，均值更清楚地显示出逐渐上升和急剧下降的趋势：当 γ 非常小时，性能略低；在所选集合中等程度的 γ 处达到最大值；随后，一旦 γ 达到 0.4，随着惩罚力度显著增强，性能便急剧下降。当准确率和调整兰德指数达到最高时，其标准差极小，表明不同重复试验中的性能非常稳定。

仍以 $C = 3$ 的情形为重点，我们现在将注意范围限定为 $k = 20$ ，以便于可视化并减少展示的情景数量。在此设置下，我们考察共享脆弱性模型中未知参数的估计值， $[\theta, \rho, \xi, \beta]$ （回想一下，在第2节中，我们将 $\psi = [\rho, \xi]$ 命名为威布尔分布，因为它由两个参数确定）。为求完整，我们还纳入 $\gamma = 0$，以便可视化未进行惩罚化时得到的参数估计值。图 1 和图 2 分别展示了行政删失和正态删失情况下，不同 γ 取值对应的参数估计值箱线图。在这些箱线图中，真实参数值以水平虚线标出，从而可以直观评估偏差。正如预期，估计值的偏差通常随着横轴所示的 γ 增大而增大。具体而言，图 (a) 中的参数 θ 表示脆弱性的方差，即使在未惩罚的情况下（$\gamma = 0$，第一个箱线图）也略有低估；这一模式与采用 parfm 软件包实现的既有研究（Munda 等人，2012）一致。此外，在不同 γ 取值下，大多数参数的变异性相对稳定，表明惩罚化对其离散程度影响有限。唯一的例外是 ξ，即威布尔分布的第二个参数；它似乎受影响最大，随着 γ 增大，其变异性明显上升。

为评估惩罚对参数估计值质量的影响，表 D.1 报告了所有 γ 取值及两种删失方案下的方差（Var）和经方差归一化的均方误差（MSE/Var）。附录 D 对这些结果进行了详细解读。

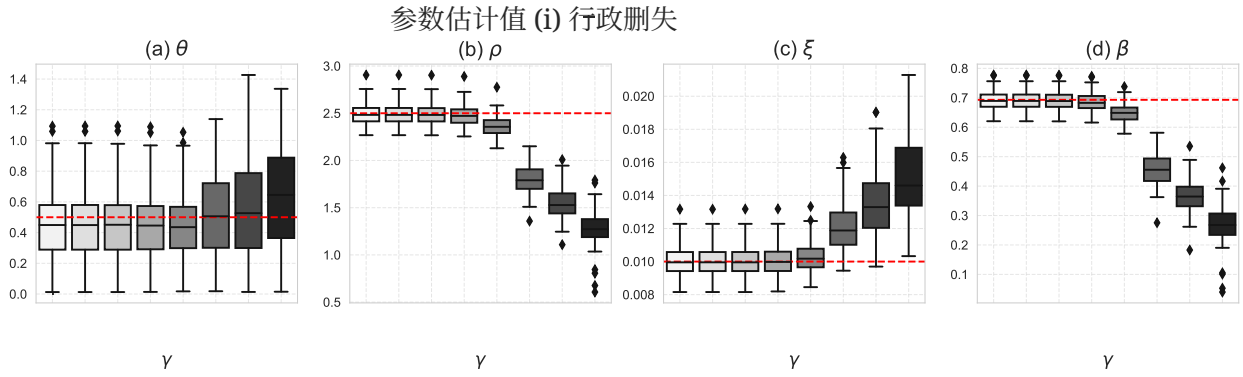

图 1：在 $k = 20$ 和 $C = 3$ 下，行政删失时参数估计值（θ， ρ， ξ， β）随惩罚水平 γ（横轴）变化的箱线图。真实参数值以分段常数水平线表示。

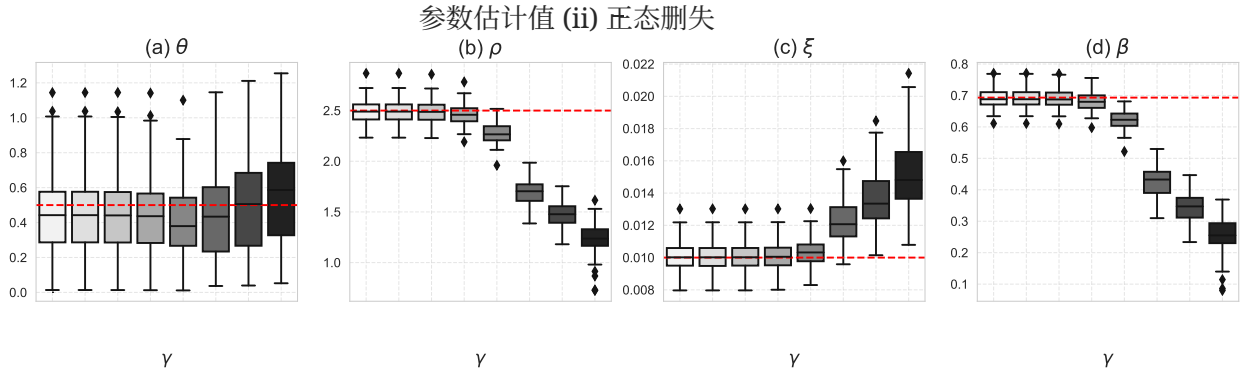

图 2：在 $k = 20$ 和 $C = 3$ 下，正态删失时参数估计值（θ， ρ， ξ， β）随惩罚水平 γ（横轴）变化的箱线图。真实参数值以分段常数水平线表示。

现在，我们转而分析改变 C 所得到的结果。在表 2 中，我们分析了两种删失机制下， C 取不同值时平均轮廓值的经验分布，

以及 $k = 20$ 和 $\gamma = 10^{-4}$ 的特定取值。结果表明，在两种情形下， $C = 3$ 的平均值和中位数平均轮廓值均最高，说明这一选择总体上能实现最佳的簇分离效果。尽管 $C = 3$ 的轮廓值变异性略高，但总体分离效果仍显著更强。此外，在所有重复试验中， $C = 3$ 均能在规定的迭代次数内收敛至选定的簇数，而其他取值并非都能做到这一点。

删失	C	平均轮廓系数			收敛
		Mean	中位数	SD	%
(i) 行政	2	0.660	0.670	0.046	67
	3	**0.792**	**0.832**	**0.116**	**100**
	4	0.661	0.718	0.121	100
	5	0.566	0.612	0.113	100
(ii) 正态	2	0.618	0.649	0.082	93
	3	**0.670**	**0.789**	**0.182**	**100**
	4	0.541	0.577	0.173	99
	5	0.466	0.484	0.145	100

表 2： 在两种删失机制下，针对 $k = 20$ 和 $\gamma = 10^{-4}$， $B = 100$ 次重复试验中平均轮廓值经验分布的汇总。对于 C 的每个取值，表中报告均值、中位数、标准差，以及在 maxit 次迭代内收敛到所选 C 的运行比例。与 $C = 3$ 相关的估计值以粗体突出显示。

4 案例研究

本节将所提出的方法论应用于 *Enhance-Heart* 数据集。该大型行政数据集与伦巴第大区的新冠肺炎登记库[1] 相关联（下文简称 RL）。

4.1 队列选择

我们从与 *Enhance-Heart* 项目相关的行政数据集中，选取在 *January 1, 2018* 至 *December 31, 2020* 期间确诊心力衰竭（HF）的患者，即根据 RL MS-DRG v40 系统，因 DRG 代码 127（“心力衰竭与休克”）住院或前往急诊室就诊的患者，其中包括心力衰竭的主要/次要诊断（ICD-9-CM：428.xx）及相关病症（例如 ICD-9-CM：402.01、402.11、402.91）。在这些患者中，我们保留了根据数据库提供的二元标记，在 *31 January 2020* （意大利全国紧急状态开始）至 *18 June 2021* （实施规范重新开放的绿色通行证规定）期间因 COVID-19 住院的患者。对这些患者，我们在 90 天随访期内评估结局；该时间窗内死亡视为事件，而存活超过该时间窗或之后死亡的患者则视为删失（占样本的 55.25%）。

分析仅限于至少收治过50名住院患者的医院，因为规模较小的医院可能具有不同的组织状况和特征，同时这也有助于确保脆弱性相关估计更加稳健。我们进一步排除了因医疗目的而需要医院转院的患者，以及出现多次感染发作的患者。最终队列包括在RL的 $J = 32$ 家不同医院中因COVID-19住院的 $N = 3086$ 名HF患者。

事件发生时间结局由一对变量 (Time, Status) 定义，其中 Time 表示在 90 天观察窗口内直至死亡或删失的天数，Status 则通过二元变量指示是否发生事件。患者的协变量包括入院年龄、性别和修正多源合并症评分（ModMCS）。该评分用于衡量总体合并症负担，与主要诊断无关，可作为整体健康状况的代理指标；评分越高，表示合并症越多。该评分由 Corrao 等人（2017）提出，并由 Caldera 等（2025）调整，以排除与 COVID-19 相关的呼吸系统疾病。实际上，我们单独考虑了变量 Resp，并将其定义为一个二元指标：当患者患有慢性阻塞性肺疾病、支气管炎、肺炎或呼吸衰竭中的一种以上呼吸系统疾病时，该指标取值为 1。研究变量的详细说明和汇总统计量见表 3。

[1] https://www.dati.lombardia.it/

表 3： RL数据库中所选变量的列表、描述和汇总统计 .

变量	描述	Type	汇总统计
`Status`	死亡的二元指标；如果事件在观察期内发生，则为 1，否则为 0。	分类变量	1（死亡）占44.75%，0（存活）占55.25%。
`Time`	生存时间（以天为单位）或观察期，取决于 `Status`。	连续变量	均值 = 19.20，标准差 = 18.04，中位数 = 12, [Q1；Q3] = [6； 26], [最小值；最大值] = [1； 89]。
`Age`	患者入院年龄。	连续变量	均值 = 81.36，标准差 = 8.27，中位数 = 82.0, [Q1；Q3] = [76； 88], [最小值；最大值] = [60； 106]
`Gender`	患者的生理性别（男/女）。	分类变量	男性占55.83%，女性占44.17%。
`ModMCS`	改良多源共病评分 (Caldera et al., 2025; Corrao et al., 2017) 量化其他疾病的累积疾病负担。	连续变量	均值 = 11.47，标准差 = 7.77，中位数 = 10, [Q1；Q3] = [6； 16], [分钟；最大值] = [0； 57]
`Resp`	*chronic ob*-中的呼吸系统疾病数量 *structive pulmonary disease*、*bronchitis*、*pneumonia*，以及 *respiratory failure*。	连续变量	均值 = 0.97，标准差 = 1.18，中位数 = 1， [四分位数1；四分位数3] = [0； 2], [最小值；最大值] = [0； 4]

4.2 模型设置与结果

我们在多种配置下拟合模型，并按照第 2.4 节所述（Kaufman 和 Rousseeuw，2009），将平均轮廓值最大的配置确定为最优配置。具体而言，我们考虑两种脆弱性分布（伽马分布和逆高斯分布），以及基线风险分布的两种选项（威布尔分布和指数分布）。对于惩罚参数，我们考察 $\gamma = \{10^{-8}, 10^{-5}, 10^{-2}, 10^{-1}\}$ 和 $k \in \{50, 100\}$。最后，我们评估采用 $C \in \{2, 3, 4\}$ 的模型。我们设置 maxit $= 200$、tolS $= 10^{-2}$、t olll $= 1$，其余参数均采用默认设置。上述所有配置的平均轮廓值见表4。

我们选择脆弱性服从逆高斯分布、基线分布为威布尔分布，$\gamma = 10^{-1}$、$k = 150$ 和 $C = 3$ 作为最优聚类配置的模型，因为该模型的平均轮廓系数最高，达到 0.609，因此相对于拟合的生存模型具有最佳的分离效果。采用所选模型后，我们得到以下回归系数：$\hat{}\beta_{\texttt{Age}} = 0.0025$、$\hat{}\beta_{\texttt{Gender=male}} = 0.0244$、$\hat{}\beta_{\texttt{ModMCS}} = 0.0013$ $\beta_{\texttt{Resp}} = 0.0033$ < 0.05 和$\hat{}$。对应的 p 值均为 ，表明这些效应具有统计显著性。估计的脆弱性方差为 $\theta = 0.573$，威布尔形状参数和尺度参数分别为$\rho = 0.618$ 和 $\lambda = \psi^{\rho} = 1.126$。

模型经过 23 次迭代后收敛。得到的三个簇——对应患者图中的非连通分量——分别包含 1077、1054 和 955 名患者，揭示了具有不同风险特征的亚组。具体而言，第一个（也是最大的）簇包含年龄较大的患者（平均年龄 83.5 岁），其改良多源共病评分最高（平均 ModMCS 为 12.6），呼吸系统疾病负担中等（平均 Resp 为 0.99），且以男性为主，平均随访时间为 0.147。簇 2 包含年龄稍小的患者（平均年龄 81.8 岁），其 ModMCS 处于中等水平（11.8），呼吸系统疾病负担最高（1.03），也以男性为主，平均随访时间为 0.158。簇 3 是规模最小的簇，包含年龄最小的患者（平均年龄 78.4 岁），其 ModMCS（9.8）和呼吸系统疾病负担（0.87）均最低，女性多于男性，且平均随访时间最长（0.176）。

各医院的簇成员关系差异显著。属于主导簇的观测值比例中位数为 75.2%（四分位距为 66.9–95.5%），表明总体上存在中等程度的医院内异质性。13 家医院（40.6%）呈现出单一簇明显占主导的情况，其中至少 80% 的观测值被分配到该簇；另有 5 家医院（15.6%）的观测值全部归属于同一个簇。相比之下，仅有 3 家医院（9.4%）呈现混合的簇组成，没有任何一个簇包含超过 60% 的观测值。

簇模式的范围很广——从由单一簇主导的医院到簇组成更加混合的医院——这凸显了显著的医院内异质性，也说明有必要采用专门用于处理这种变异性的方法，而非假定情况一致。估计的 θ 也进一步体现了这种异质性，表明医院层面效应存在有意义的变异，而我们的方法能够对此加以考虑。

参数分布		惩罚参数		C		
u_g	$H_0(t;\ \psi)$	γ	k	2	3	4
伽马分布	威布尔分布	10^{-8}	50	0.544	0.470	0.320
			150	0.535	0.526	0.277
		10^{-5}	50	0.542	0.379	0.416
			150	0.526	0.496	0.293
		10^{-2}	50	0.526	0.508	0.292
			150	0.518	0.464	0.461
		10^{-1}	50	0.474	0.363	0.252
			150	0.591	**0.608**	0.033
伽马分布	指数分布	10^{-8}	50	0.542	0.376	0.243
			150	0.548	0.473	0.196
		10^{-5}	50	0.545	0.290	0.277
			150	0.533	0.514	0.092
		10^{-2}	50	0.538	0.495	0.440
			150	0.525	0.404	0.379
		10^{-1}	50	0.524	0.126	0.441
			150	0.493	0.132	0.402
逆高斯分布	威布尔分布	10^{-8}	50	0.545	0.512	0.399
			150	0.545	0.498	0.419
		10^{-5}	50	0.525	0.462	0.170
			150	0.527	-0.244	0.187
		10^{-2}	50	0.532	0.491	0.426
			150	0.528	0.420	0.412
		10^{-1}	50	0.487	0.481	0.224
			150	0.594	**0.609**	0.001
逆高斯分布	指数分布	10^{-8}	50	0.544	0.452	0.434
			150	0.546	0.485	0.473
		10^{-5}	50	0.543	0.286	0.472
			150	0.525	0.305	0.417
		10^{-2}	50	-0.228	0.472	0.514
			150	0.533	0.437	0.427
		10^{-1}	50	0.536	0.046	0.209
			150	0.502	0.141	0.126

表4：不同模型配置下的平均轮廓值，涵盖不同的脆弱性和基线风险分布、惩罚参数 γ 和 k以及簇数 C。

5 讨论

受识别层级生存时间数据中潜在患者亚组这一需求的启发，我们在本研究中提出了一种框架，可在通过参数化共享脆弱性模型处理组级依赖的同时，联合进行基于患者风险的聚类和生存建模，确保簇分配依据潜在的生存动态，而非仅依据协变量相似性。从方法论角度看，我们的方法填补了现有文献中的一项空白：在单一的基于似然的框架内联合建模生存结局、潜在患者亚组和组级依赖，使患者分层能够依据估计得到的风险特征自然形成，同时妥善处理医院内相关性。

模拟研究表明，只要将脆弱性方差保持在较小水平，使聚类特异性效应仍是异质性的主要来源，所提出的方法论就能在广泛的惩罚设置下可靠地恢复潜在聚类结构。这种选择有助于确保可识别性，避免个体层级随机效应掩盖潜在风险轨迹中的系统性差异。在这些条件下，若以轮廓系数作为最优聚类和参数估计的选择标准，并据此恰当调整惩罚，该方法还能够准确估计参数化脆弱性模型。

将该方法应用于因 COVID-19 住院的心力衰竭患者 Enhance-Heart队列后，识别出三个患者簇，这些簇具有不同的风险特征，并且在年龄、合并症负担、呼吸系统疾病和生存模式方面存在临床上可解释的差异。通过选择使轮廓系数最大化的参数配置和参数分布，得到这些簇，从而确保最佳的分离度和凝聚度。这些簇对应于患者图中的不连通分量，并揭示了具有临床意义的亚组：第一簇为年龄较大、合并症较多但呼吸系统疾病负担中等的患者；第二簇包括年龄略小、呼吸系统疾病负担最高的患者；第三簇则由年龄最小、合并症和呼吸系统疾病负担最低的患者组成。这凸显了该方法论能够揭示事件风险模式中细致的异质性。

除患者层面的分层外，我们的结果还显示医院之间存在显著异质性。各医院的簇组成差异显著，许多机构中单一簇占据明显主导地位，而少数机构的患者构成则更为多样。此外，脆弱性参数的值不可忽略，这表明潜在的机构因素——例如资源可用性、临床方案或新冠肺炎疫情期间的情境压力——可能会影响生存结局，其作用超出了观测协变量和簇成员关系所能解释的范围。这一发现凸显了显式建模层级依赖的重要性，并支持将聚类与脆弱性项结合使用，作为相辅相成的工具，以厘清异质性的多重来源。

有若干局限性值得考虑。首先，相似性图的构建依赖于邻域大小参数 k的选择，该参数会影响图稀疏性和聚类形成。尽管算法中嵌入的惩罚参数 λ 可通过自适应调节在实践中缓解这一问题，但 k 与目标簇数 C 之间的关系仍有待进一步研究。此外，每个个体都必须连接到其 k 个最近邻，无论这些邻居位于聚类核心附近还是边界处。谱聚类文献（Cai等，2022年）也指出了这一局限，并提出剪除弱连接以及通过幂函数渐近修改优化项作为潜在解决方案。然而，这类方法会引入额外的调节参数（包括最大邻域大小），且在我们的研究设置中并未带来性能提升。

此外，在当前框架中，假设各簇具有共同的基线风险，该假设在惩罚距离中得到简化，而异质性则通过协变量和脆弱性项来刻画。未来的工作将致力于扩展模型，使其能够允许聚类特异性基线风险函数，从而更细致地描述不同患者亚组的风险动态。这一扩展需要重新考虑脆弱性结构，因为不同簇将具有不同的脆弱性，超出了本研究的范围。后续研究还可探索其他相似性度量，以及将模型扩展至半参数或全非参数基线风险函数，从而进一步提高模型的灵活性。

利益冲突

未声明存在利益冲突。

数据可用性

本研究参与者未书面同意公开共享其数据，因此由于研究的敏感性质以及适用于 RL 共享条款的相关规定，无法提供完整的支撑数据。

致谢

本研究是 *Enhance-Heart* 项目的一部分：该项目旨在评估心力衰竭患者治疗护理路径的疗效、医疗服务提供者的影响以及风险分层。作者感谢“Unit`a Organizzativa Osservatorio Epidemiologico Regionale”和 ARIA 股份公司提供数据及技术支持，并感谢意大利大学与研究部通过米兰理工大学数学系的“卓越系 2023—2027”资助项目提供资金支持。

边际似然的推导

第 g个簇的 *conditional likelihood* 表示为

$$L_g^c(\boldsymbol{\psi}, \boldsymbol{\beta} \mid u_g, \mathcal{Z}) = \prod_{i=1}^{n_g} \left(h_0(y_{gi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)^{\delta_{gi}} \cdot \exp\left(-H_0(y_{gi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)$$

而 *marginal likelihood* 则通过在其支撑集 S_U上对脆弱性分布积分消去而得到：

$$\begin{aligned}
L_g(\boldsymbol{\psi}, \boldsymbol{\beta}, \theta, \mathcal{Z}) &= \int_{S_U} L_g^c(\boldsymbol{\psi}, \boldsymbol{\beta} \mid u_g, \mathcal{Z}) \cdot f_U(u_g \mid \theta) du_g \\
&= \int_{S_U} \prod_{i=1}^{n_g} \left(h_0(y_{gi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)^{\delta_{gi}} \cdot \exp\left(-H_0(y_{gi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right) \cdot f_U(u_g \mid \theta) du_g \\
&= \prod_{i=1}^{n_g} \left(h_0(y_{gi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)^{\delta_{gi}} \int_{S_U} \prod_{i=1}^{n_g} (u_g)^{\delta_{gi}} \cdot \exp\left(-H_0(y_{gi}) u_g \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right) \cdot f_U(u_g) du_g \\
&= \prod_{i=1}^{n_g} \left(h_0(y_{gi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)^{\delta_{gi}} \int_{S_U} u_g^{\sum_{i=1}^{n_g} \delta_{gi}} \cdot \exp\left(-\sum_{i=1}^{n_g} H_0(y_{gi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta}) u_g\right) \cdot f_U(u_g) du_g \\
&= \prod_{i=1}^{n_g} \left(h_0(y_{gi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)^{\delta_{gi}} (-1)^{d_g} \mathcal{L}^{(d_g)}\left(\sum_{i=1}^{n_g} H_0(y_{gi}) \exp(\boldsymbol{x}'_{gi}\boldsymbol{\beta})\right)
\end{aligned}$$

其中，在最后一个等式中，我们使用了 $(-1)^k \mathcal{L}^{(k)}(s) := (-1)^k \frac{d^k}{ds^k} \mathcal{L}(s) := \mathbb{E}[U^k \exp(-Us)]$, 为 $\mathcal{L}(s)$，对于 $s \geq 0$, 即拉普拉斯变换（Munda and Legrand, 2014; Klein, 1992）。对簇求对数并求和后，我们得到式 (2)。

B 式 (9) 闭式解的推导

鉴于 $\sum_{h,\ j} \mu s_{gi,\ hj}^2 + w_{gi,\ hj} s_{gi,\ hj} = \mu \|\mathbf{s}_{gi}\|^2 + \boldsymbol{w}'_{gi}\mathbf{s}_{gi} = \mu \|\mathbf{s}_{gi} + \frac{1}{2\mu}\boldsymbol{w}_{gi}\|^2 - \frac{1}{4\mu}\|\boldsymbol{w}_{gi}\|^2$，问题 (9) 可改写为

$$\min_{\mathbf{s}'_{gi}\mathbf{1}=1, \mathbf{s}_{gi} \geq 0} \quad \left\|\mathbf{s}_{gi} + \frac{1}{2\mu}\boldsymbol{w}_{gi}\right\|^2$$

其中，最后一项 $\frac{1}{4\mu}\|\boldsymbol{w}_{gi}\|^2$ 和乘数 μ 已被去除，因为前者与$\mathbf{s}^{gi}$ 无关，而后者可从优化问题中去除。正如我们所述，该问题针对组 g 中的每个个体 i 独立求解，因此从本节开始，我们将 μ 记为 μ_{gi}。该问题可通过 KKT 条件求解。事实上，拉格朗日函数可写为

$$\mathcal{L} = \left\|\mathbf{s}_{gi} + \frac{1}{2\mu_{gi}}\boldsymbol{w}_{gi}\right\|^2 - \alpha_{gi}(\mathbf{s}'_{gi}\mathbf{1} - 1) - \boldsymbol{r}'\mathbf{s}_{gi}$$

其中， $\alpha_{gi} \geq 0$ 和 $\boldsymbol{r} \geq 0$ 是拉格朗日乘子。根据 KKT 条件，我们得到以下解

$$s_{gi,hj} = \max\left\{0, \alpha_{gi} - \frac{w_{gi,hj}}{2\mu_{gi}}\right\} \tag{B1}$$

满足约束 $\mathbf{s}'_{gi}\mathbf{1} = 1$。

C α_{gi} 和 μ_{gi} 在(10)中的推导

最优解由 B1 给出，并满足约束 $\mathbf{s}'_{gi}\mathbf{1} = \sum_{h=1}^{N} s_{gi,\ hj} = 1$。 L

et $\hat{w}_{g_{i,\ 1}} \leq \hat{w}_{gi,\ 2} \leq \cdots \leq \hat{w}_{gi,\ N-1}$ 表示排序后的距离（不包括自身距离）。施加 g

k-最近邻- 稀疏性可得

$$\begin{aligned}
&\hat{s}_{gi,p} > 0, \quad p = 1, \ldots, k, \\
&\hat{s}_{gi,p} = 0, \quad p = k+1, \ldots, N.
\end{aligned}$$

对于 k 非零项,

$$\hat{s}_{gi,p} = \alpha_{gi} - \frac{\hat{w}_{gi,p}}{2\mu_{gi}}, \qquad p = 1, \ldots, k.$$

对于第 $(k+1)$ 个邻居，稀疏性条件意味着

$$\alpha_{gi} - \frac{\hat{w}_{gi,k+1}}{2\mu_{gi}} \leq 0.$$

在最优解处，该约束为紧约束，因此得到

$$\alpha_{gi} = \frac{\hat{w}_{gi,k+1}}{2\mu_{gi}}.$$

对 k 非零项施加单纯形约束，得到

$$\sum_{p=1}^{k} \hat{s}_{gi,p} = \sum_{p=1}^{k} \left(\alpha_{gi} - \frac{\hat{w}_{gi,p}}{2\mu_{gi}}\right) = 1,$$

$$k\alpha_{gi} - \frac{1}{2\mu_{gi}} \sum_{p=1}^{k} \hat{w}_{gi,p} = 1.$$

求解 α_{gi} 可得

$$\alpha_{gi} = \frac{1}{k} + \frac{1}{2k\mu_{gi}} \sum_{p=1}^{k} \hat{w}_{gi,p}. \tag{C1}$$

结合 (C) 和 (C1)，得到

$$\frac{\hat{w}_{gi,k+1}}{2\mu_{gi}} = \frac{1}{k} + \frac{1}{2k\mu_{gi}} \sum_{p=1}^{k} \hat{w}_{gi,p},$$

$$2\mu_{gi} = k\hat{w}_{gi,k+1} - \sum_{p=1}^{k} \hat{w}_{gi,p}.$$

因此,

$$\mu_{gi} = \frac{k}{2}\hat{w}_{gi,k+1} - \frac{1}{2}\sum_{p=1}^{k} \hat{w}_{gi,p}.$$

D 参数估计值：变异性与估计误差

在本附录中，我们解读表 D.1。该表报告了不同惩罚水平 γ 下，在行政删失和正态删失两种情况下参数估计值的归一化均方误差和方差。对于非常小的惩罚项（$\gamma \leq 10^{-4}$），估计量保持稳定，所有参数的 MSE/Var 值都接近 1，表明估计误差几乎完全由方差驱动；图 1 和图 2 也证实了这一观察结果。随着 γ 增至 10^{-3}–10^{-2}这一范围，偏差开始显现，在 ρ 和 β上尤为明显：尽管方差变化相对较小，MSE/Var 却有所上升。对于较大的惩罚项（$\gamma \geq 0.1$），估计准确率急剧下降，MSE/Var 大幅增加，尤其是在 ρ 和 β上。两种删失机制下都观察到了这一模式，但当删失服从正态分布时，这一现象始终更加明显。

这些发现应结合表 1 一并解读；表中显示，中等程度的惩罚水平最优，体现了收缩与估计误差之间的权衡。

删失	γ	θ		ρ		ψ		β	
		均方误差/方差	Var	均方误差/方差	Var	均方误差/方差	Var	均方误差/方差	Var
(i) 行政	0	1.022	0.051	1.006	0.013	1.001	0.000	1.007	0.001
	10^{-6}	1.022	0.051	1.006	0.013	1.001	0.000	1.007	0.001
	10^{-4}	1.023	0.051	1.008	0.013	1.001	0.000	1.010	0.001
	10^{-3}	1.025	0.050	1.042	0.012	1.003	0.000	1.056	0.001
	10^{-2}	1.046	0.046	2.633	0.011	1.077	0.000	3.190	0.001
	0.1	1.000	0.069	20.407	0.025	3.036	0.000	20.207	0.003
	0.2	1.047	0.091	35.186	0.027	4.269	0.000	33.404	0.003
	0.4	1.205	0.096	44.550	0.034	5.471	0.000	43.941	0.004
(ii) 正态	0	1.022	0.052	1.004	0.012	1.002	0.000	1.007	0.001
	10^{-6}	1.022	0.052	1.004	0.012	1.002	0.000	1.007	0.001
	10^{-4}	1.024	0.052	1.009	0.012	1.003	0.000	1.013	0.001
	10^{-3}	1.036	0.050	1.131	0.011	1.007	0.000	1.169	0.001
	10^{-2}	1.170	0.041	5.808	0.010	1.137	0.000	6.310	0.001
	0.1	1.077	0.054	46.782	0.014	3.922	0.000	39.592	0.002
	0.2	1.001	0.068	77.124	0.014	5.371	0.000	64.944	0.002
	0.4	1.058	0.078	74.959	0.022	6.449	0.000	71.004	0.003

表 D.1： 对于 $k = 20$ 和 $C = 3$，参数估计值的归一化均方误差（均方误差/方差）和方差（方差）。对于每个参数 $p \in \theta,\ \rho,\ \xi,\ \beta$，我们定义 $\text{MSE}(\hat{p}) = \mathbb{E}[\|\hat{p} - p\|_2^2]$ 和 $\text{Var}(\hat{p}) = \mathbb{E}[\|\hat{p} - \mathbb{E}[\hat{p}]\|_2^2]$，其中 p 表示第3.1节中引入的真实参数。该表报告了所有考察的惩罚水平 γ下的行政删失和正态删失结果。

参考文献

Abrahantes, J. C., C. Legrand, T. Burzykowski, P. Janssen, V. Ducrocq, 和 L. Duchateau (2007). 带随机效应比例风险模型的不同估计程序比较。*Computational statistics & data analysis 51*(8), 3913–3930.

Adeghate, E. A., N. Eid, 和 J. Singh (2021). COVID-19诱发心力衰竭的机制：简要综述。 *Heart failure reviews 26*(2), 363–369.

Ahlqvist, E., P. Storm, A. K¨ar¨aj¨am¨aki, M. Martinell, M. Dorkhan, A. Carlsson, P. Vikman, R. B. Prasad, D. M. Aly, P. Almgren, 等 (2018). 成人起病糖尿病的新亚组及其与结局的关联：基于六个变量的数据驱动聚类分析。 *The lancet Diabetes & endocrinology 6*(5), 361–369.

Alteri, C., V. Cento, A. Piralla, V. Costabile, M. Tallarita, L. Colagrossi, S. Renica, F. Giardina, F. No- vazzi, S. Gaiarsa, 等 (2021). SARS-CoV-2的基因组流行病学揭示了多个谱系以及SARS-CoV-2感染在意大利伦巴第的早期传播。 *Nature communications 12*(1), 434.

Austin, P. C. (2017). 多层次生存分析教程：方法、模型与应用。*International Statistical Review 85*(2), 185–203.

Bader, F., Y. Manla, B. Atallah, 和 R. C. Starling (
2021 19 *Heart failure*).
reviews 26 1 1 10
心力衰竭与COVID- . ( ), – . Balan, T. A. 和 H. Putter (
2020 *Statistical methods in medical re-*).
search 29 11 3424 3454
脆弱性模型教程。( ), – . Buginga, G. 和 E. de Souza e Silva (2024). 使用非参数专家混合模型对生存数据进行聚类。 *arXiv preprint arXiv:2405.15934*. Cai, Y., J. Z. Huang, 和 J. Yin (2022年). 构建用于谱聚类的自适应k近邻相似图矩阵的新方法。 *Neurocomputing 493*, 191–203. Caldera, L., A. Cappozzo, C. Masci, M. Forlani, B. Antonelli, O. Leoni, A. M. Paganoni, 和 F. Ieva (2025). 对寿命层级数据进行聚类加权建模，以刻画COVID-19心力衰竭患者的特征。 *arXiv preprint arXiv:2507.12230*. Caldera, L., C. Masci, A. Cappozzo, M. Forlani, B. Antonelli, O. Leoni, 和 F. Ieva (2025, 04). 揭示COVID-19心力衰竭患者的死亡模式与医院效应：一种新型多层逻辑聚类加权建模方法。 *Biometrics 81*(2), ujaf046. Cavinato, L., N. Gozzi, M. Sollini, C. Carlo-Stella, A. Chiti, 和 F. Ieva (2021). 用于霍奇金淋巴瘤影像组学表型亚型划分的复发特异性监督图聚类。载于 *2021 43rd annual international conference of the IEEE engineering in medicine & biology society (EMBC)* ,
第2155–2158页。IEEE。Cavinato, L., N. Gozzi, M. Sollini, M. Kirienko, C. Carlo-Stella, C. Rusconi, A. Chiti, 和 F. Ieva (2023). 通过从霍奇金淋巴瘤回顾性癌症亚型分类中提取基于影像的规则来构建视角迁移模型。 *Authorea Preprints*. Chapfuwa, P., C. Li, N. Mehta, L. Carin, 和 R. Henao (
2020 *Proceedings*).
of the ACM Conference on Health Inference and Learning 60 68
生存聚类分析。载于,,, 第–页。Chen, D., H. Wang, L. Sheng, M. T. Hueman, D. E. Henson, A. M. Schwartz, 和 J. A. Patel (2016). 一种用于构建癌症预后系统的算法。 *Journal of medical systems 40*(7), 160. Clayton, D. G. (1978). 双变量生命表中的关联模型及其在慢性病发病家族倾向流行病学研究中的应用。 *Biometrika 65*(1), 141–151. Collins, F. S. 和 H. Varmus (
2015 *New England journal of*).
medicine 372 9 793 795
精准医学新倡议。( ), – . Corrao, G., F. Rea, M. Di Martino, R. De Palma, S. Scondotto, D. Fusco, A. Lallo, L. M. B. Belotti, M. Ferrante, S. P. Addario, 等 (2017). 基于行政数据开发并验证一种新型多源合并症评分：一项来自意大利的大型基于人群的队列研究。 *BMJ open 7*(12), e019503. Duchateau, L. 和 P. Janssen (2008). *The frailty model*. Springer。Fan, K. (
1949 *Proceedings of*).
the National Academy of Sciences 35 11 652 655
关于韦伊有关线性变换特征值的定理 I。( ), – . Guo, X. (2015). 通过同时学习数据表示及其亲和矩阵进行稳健子空间分割。载于 *IJCAI*，第3547–3553页。Janke, A. T., H. Mei, C. Rothenberg, R. D. Becher, Z. Lin, 和 A. K. Venkatesh (2021). 美国各地医院资源可用性与COVID-
19 *Journal of hospital* 死
medicine 16 4 211 214
亡率分析。( ), – . Kaufman, L. 和 P. J. Rousseeuw (2009).
Finding groups in data: an introduction to cluster analysis. John Wiley & Sons。Klein, J. P. (1992). 基于EM算法使用Cox模型进行随机效应的半参数估计。 *Biometrics*, 795–806.

Liu, C., W. Cao, S. Wu, W. Shen, D. Jiang, Z. Yu, and H.-S. Wong (2020). 基于生存分析和多组学肿瘤数据整合的癌症亚型划分有监督图聚类。 *IEEE/ACM Transactions on Computational Biology and Bioinformatics 19* 2 1193 1202 ( ), – 。Manduchi, L., R. Marcinkevi˘cs, M. C. Massi, T. Weikert, A. Sauter, V. Gotta, T. M¨uller, F. Vasella, M. C. Neidert, M. Pfister, et al. (2021 *arXiv*). *preprint arXiv:2106.05763*
一种用于生存数据聚类的深度变分方法。. Mohar, B., Y. Alavi, G. Chartrand, and O. Oellermann (1991 *Graph*). *theory combinatorics and applications 2* 871 898 12
图的拉普拉斯谱。,,(-), 。Monaco, J. V., M. Gorfine, and L. Hsu (2018). 通用半参数共享脆弱性模型：使用 frailtysurv 进行估计和模拟。 *Journal of statistical software 86*, 1–42. Munda, M. and C. Legrand (2014). 在具有事件发生时间结局的多中心临床试验中调整中心异质性。 *Pharmaceutical statistics 13*(2), 145–152. Munda, M., F. Rotolo, and C. Legrand (2012 *Journal of*). *Statistical Software 51* 1 20
parfm：R 中的参数脆弱性模型。, – 。Nie, F., G. Cai, and X. Li (2017). 使用自适应邻居进行多视图聚类和半监督分类。见 *Proceedings of the AAAI conference on artificial intelligence*，第 31 卷。Nie, F., X. Wang, and H. Huang (2014). 使用自适应邻居进行聚类和投影聚类。见 *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* 977 986 ,
第 – 页。Nie, F., X. Wang, M. Jordan, and H. Huang (2016). 用于基于图的聚类的约束拉普拉斯秩算法。见 *Proceedings of the AAAI conference on artificial intelligence*，第 30 卷。R核心团队 (2022)。 *R: A Language and Environment for Statistical Computing*。奥地利维也纳：R统计计算基金会。
Tosado, J., L. Zdilar, H. Elhalawani, B. Elgohari, D. M. Vock, G. E. Marai, C. Fuller, A. S. Mohamed, and G. Canahuate (2020). 对大量右删失的口咽头颈癌患者进行聚类，以形成判别性分组并改善结局预测。 *Scientific reports 10*(1), 3811。Vaupel, J. W., K. G. Manton, and E. Stallard (1979). 个体脆弱性异质性对死亡率动态的影响。 *Demography 16*(3), 439–454。