arXiv:2303.00715v1 [stat.AP] 1 Mar 2023

利用带检测限的阿尔茨海默病生物标志物响应变量，发现亚型的多变量回归混合模型

Ganzhong Tian[a]、John Hanfelt[a]、James Lah[b]、Benjamin B. Risk[a]

[a]埃默里大学生物统计学与生物信息学系[b]埃默里大学医学院神经病学系

文章历史
编于2023年3月2日

摘要
除尸检外，阿尔茨海默病（AD）的诊断没有金标准。无监督学习有助于深入了解 AD 的病理生理学。回归混合模型能够在考虑簇内人口统计学效应的同时，根据多种生物标志物识别簇。AD 的脑脊液（CSF）生物标志物存在检测限，这带来了额外的挑战。我们将带有截断多元高斯分布的回归混合模型（也称为删失多元高斯回归混合模型或多元托比特回归混合模型）应用于埃默里大学戈伊祖埃塔阿尔茨海默病研究中心和埃默里健康大脑研究的 3，000 多名参与者，研究具有已知检测限的脑脊液 β-淀粉样蛋白 1-42（Abeta42）、总 tau 蛋白和磷酸化 tau 蛋白。我们针对采用截断多元高斯分布的回归混合模型研究中存在的三个空白展开工作：软件可用性、推断和聚类准确率。我们发现了三个簇，分别倾向于对应阿尔茨海默病组、正常对照特征和非阿尔茨海默病病理。脑脊液特征因种族、性别和遗传标志物载脂蛋白 E4 而异，这凸显了在存在检测限的无监督学习中考虑人口统计学因素的重要性。值得注意的是，类 AD 组中的非裔美国人参与者的 tau 负担显著较低。

关键词
阿尔茨海默病；删失值高斯回归混合模型；聚类；有限混合模型；潜在类别分析；Tobit 模型；截断正态分布；无监督学习。

1. 引言

只有通过尸检检查脑组织，才能确诊阿尔茨海默病（阿尔茨海默病）（Dubois 等人，2007）。临床诊断历来依赖于以欧洲血统人群为主的研究，且使用了生物标志物，这一事实使问题更加复杂（Blennow 等人，2015年）。非裔美国人在阿尔茨海默病生物标志物研究和临床试验中的代表性严重不足（Shin 和 Doraiswamy，2016），且脑脊液生物标志物水平因种族而异（Garrett 等人，2019年）。研究者将无监督学习应用于脑脊液生物标志物，以揭示

联系人：B. B. Risk。电子邮箱：brisk@emory.edu

有关阿尔茨海默病的见解，但未考虑种族及其他人口统计学因素（Meyer 等人，2010）。分析脑脊液阿尔茨海默病生物标志物数据至少面临三项挑战：1）多变量生物标志物存在检测限，导致数据删失；2）由于没有金标准，疾病状态未知；3）人口统计学效应可能取决于未知亚型。现有统计软件无法同时处理这些问题（表 1）。我们的目标有两个：1）由于没有金标准，且现行标准可能受到欧洲血统等因素的限制，因此使用无监督多变量方法将参与者聚类分组；2）通过估计各簇内人口统计学变量（种族、性别、遗传标志物载脂蛋白 E4、年龄和受教育程度）的效应，加深对病理生理学的认识。

我们的研究受到埃默里大学戈伊祖埃塔阿尔茨海默病研究中心和埃默里健康大脑研究（以下简称埃默里大学ADRC/HBS数据集）的启发。该数据集包含来自 342 [Abeta42]，总 tau 蛋白 [tTau] 和磷酸化 tau 蛋白 [pTau]000 多名个体腰椎穿刺所得的三种脑脊液生物标志物（β淀粉样蛋白 1-,000 名个体（Goetz 等人，2019年）。该数据集中非裔美国人参与者占 16.5%（495 人），显著高于阿尔茨海默病神经影像学计划（< 5%）。检测方法的一项重要局限是，埃默里大学 ADRC/HBS数据集中约 15% 的参与者，其三种生物标志物中至少有一种的水平由该检测方法的检测限所确定。本文对删失响应变量的定义与其他混合建模论文相同（Jedidi 等人，1993年；Lee 和 Scott，2012年）：当响应变量的真实值超出阈值时，其观测值被设为检测限，此时即发生删失；而所有观测值的预测变量均可获得（例如，Abeta42 高于 1,700 的参与者，其 Abeta42 值被设为 1,700）。这与截断不同，截断通常指对总体分布进行限制性抽样（例如，若 Abeta42 高于 1,700 的患者未被记录，则这些数据就会被截断）。

无监督学习（例如高斯混合模型（高斯混合模型））是在缺乏金标准时定义疾病亚型的常用工具（Collins 和 Huynh，2014年）。基于高斯混合模型发展而来的基于模型的聚类方法，相较于基于距离的聚类算法（如 K 均值）具有优势。高斯混合模型会估计每个数据点属于各组的后验概率，而非进行硬聚类。高斯混合模型采用的统计模型能够考虑相关

性。从概率论角度来看，K 均值假设簇呈球形，因此当特征之间存在相关性时，可能导致效果不佳（Coates 和 Ng，2012年）。在我们的应用中，阿尔茨海默病的脑脊液生物标志物高度相关。高斯回归混合模型（GMR），在计量经济学中也称为“切换回归”，将高斯混合模型扩展到包含预测变量的数据集，并通过回归对各组的均值结构进行建模（Goldfeld 和 Quandt，1973年；Quandt 和 Ramsey，1978年）。这类模型允许潜在组改变预测变量的影响。这类模型也在机器学习文献中得到了广泛讨论，被称为“专家混合模型”（Yuksel 等人，2012年）。

删失多元高斯回归混合模型（删失高斯回归混合模型），也称为采用截断多元高斯分布的回归混合模型或多元托比特回归混合模型，已在文献中有所研究。Lee 和 Scott(2012年) 推导了用于将多元高斯混合模型拟合到删失数据的 EM算法。为对预测变量建模，Jedidi 等人(1993年) 推导了针对单变量删失响应的托比特回归混合模型的 EM算法。近期研究则扩展了托比特回归，用有限个高斯和/或非高斯分布的混合模型来描述单变量响应模型的误差（Hanson 和 Johnson，2002年；Caudill，2012年；Karlsson 和 Laitila，2014年；Garay 等人，2017年；Zeller 等人，2019年）。Wang 等人(2019年) 提出了一种面向多元数据的因子分析混合模型，可同时实现聚类和降维；Wang 等人(2021年) 将其扩展到预测变量和删失情形。

我们的主要贡献是使用删失多元高斯回归混合模型分析 Emory ADRC/HBS 数据集。我们实现了 EM算法，以弥补目前尚无适用于删失多元高斯回归混合模型的软件这一重要空白。我们还针对现有文献中关于在这一背景下使用 Wald 检验评估预测变量显著性的不足进行了研究；在此过程中，我们利用经验完全数据得分函数近似信息矩阵。此外，我们还开展了一项模拟研究，以弥补关于预测变量和删失对聚类准确性影响的研究空白。

本文其余部分安排如下。第2节回顾多元 Tobit 模型，并介绍其向删失高斯混合模型的扩展。随后，我们介绍 EM算法和推断方法。第3节通过模拟说明，相较于忽略或删除删失观测的方法，删失多元高斯混合回归具有哪些优势，并且

考察聚类数的选择。第4节对 Emory ADRC/HBSDataset 进行分析。最后，我们在第5节讨论研究发现和未来研究方向。

2. 建模方法与估计

本节首先回顾多元 Tobit 模型及其基于期望最大化（EM）算法的估计方法。随后，我们在此框架基础上推导用于删失多元高斯混合回归的 EM 算法。

2.1. 多元删失回归（***Tobit*** 模型）

令$\mathbf{y}_i$为第 i个受试者的 p维随机向量， $i = 1, \ldots, N$，可将其划分为两部分，

$$\mathbf{y}_i = \begin{pmatrix} \mathbf{y}_{io_i} \\ \mathbf{y}_{ic_i} \end{pmatrix},\ i \in \{1, 2, ..., N\},$$

其中$\mathbf{y}_{io_i}$ 和$\mathbf{y}_{ic_i}$ 分别表示$\mathbf{y}_i$的未删失维度和删失维度。我们用删失指示变量向量 $\boldsymbol{c}_i$ 表示$\mathbf{y}_i$ 的某个维度是否为删失值，其删失方向通过 $\boldsymbol{c}_i = (c_{i1}, ..., c_{ij}, ..., c_{ip})^\top$观测，其中 $\top$ 表示转置，因此：

$$c_{ij} = \begin{cases} 1, & \text{Right} - \text{censored}, \\ 0, & \text{Uncensored}, \\ -1, & \text{Left} - \text{censored}. \end{cases}$$

尽管真实值在被设为删失值之前无法观测到，但我们仍可进一步假设$\mathbf{y}_i$ 是由未观测真实值生成的，作为潜在随机向量$\mathbf{y}_i^* =$

$(y_{i1}^*, \ldots, y_{ij}^*, \ldots, y_{ip}^*)^\top$，其具有一些已知的检测下限和检测上限：

$$y_{ij} = \begin{cases} L_j & \text{if} \quad y_{ij}^* \leq L_j, \\ y_{ij}^* & \text{if} \quad L_j < y_{ij}^* < U_j, \\ U_j & \text{if} \quad y_{ij}^* \geq U_j. \end{cases}$$

同样，我们可以将$\mathbf{y}_i^*$划分为观测部分和删失部分：

$$\mathbf{y}_i^* = \begin{pmatrix} \mathbf{y}_{io_i} \\ \mathbf{y}_{ic_i}^* \end{pmatrix}, \ i \in \{1, 2, ..., N\},$$

其中，$\mathbf{y}_{ic_i}^*$未被观测到，且被删失为$\mathbf{y}^{ic_i}$。在多元回归模型设定中，我们还假设 $\boldsymbol{x}_i$ 是一个 d维向量，表示第 i个受试者的观测预测变量。于是，模型可写为

$$\mathbf{y}_i^* = \boldsymbol{\beta}^\top \boldsymbol{x}_i + \boldsymbol{\epsilon}_i, \tag{1}$$

其中， $\boldsymbol{\beta}$ 是一个 $d \times p$ 系数矩阵，也是主要关注的参数； $\boldsymbol{\epsilon}_i$ 是一个 p维随机噪声向量，且 $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}$（$\mathbf{0}$， $\boldsymbol{\Sigma}$）。

令 $\boldsymbol{\psi}$ 表示参数 $\boldsymbol{\beta}$ 和 $\boldsymbol{\Sigma}$的集合。令$\mathbf{Y}$ 为由观测值$\mathbf{y}_i^\top$堆叠而成的 $N \times p$ 矩阵，$\mathbf{C}$为由删失方向 $\boldsymbol{c}_i^\top$堆叠而成的 $N \times p$矩阵；$\mathbf{X}$为由预测变量向量 $\boldsymbol{x}_i^\top$堆叠而成的 $N \times d$矩阵。然后，可通过最大化不完全数据似然函数来估计 $\boldsymbol{\psi}$ ：

$$\mathcal{L}(\mathbf{Y}; \mathbf{C}, \mathbf{X}, \boldsymbol{\psi}) = \prod_{i=1}^{N} f_{\mathbf{y}_{io_i}}(\mathbf{y}_{io_i}; \boldsymbol{x}_i, \boldsymbol{\psi}) \int_{\mathcal{D}(\boldsymbol{c}_i)} f_{\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}}(\mathbf{y}_{ic_i}; \boldsymbol{x}_i, \boldsymbol{\psi}), \tag{2}$$

其中， $\mathcal{D}(\boldsymbol{c}_i)$ 是 $\mathbb{R}^p$中的一个区域，取决于由 $\boldsymbol{c}_i$表示的删失模式，而 $f_{\mathbf{y}_{ic_i}^*|\mathbf{y}^{io_i}}$是未观测响应变量的条件高斯密度。这些响应变量$\mathbf{y}_{ic_i}^*$ 在给定未发生删失的观测响应变量 $\mathbf{y}^{io_i}$的条件下经历删失。通常，(2) 通过数值方法最大化，具体采用

牛顿-拉夫森法（Amemiya，1973年）。这里，我们概述一种与（Fair，1977；Ruud，1991）类似的 EM算法，并将在第 2.2节将其扩展到高斯混合。设 $\mathbf{Y}^*$ 表示由堆叠 $N \times p$ 而成的真实观测值矩阵$\mathbf{y}_i^{*\top}$。假设观测到 $\mathbf{y}_i^*$、 $i = 1$、.....、 N，定义完整数据似然：

$$\mathcal{L}_c(\mathbf{Y}^*; \mathbf{C}, \mathbf{X}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2^p \pi^p |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{y}_i^* - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i^* - \boldsymbol{\beta}^\top \boldsymbol{x}_i)}.$$

EM算法步骤通过最大化该完整数据对数似然函数的条件期望推导得到，具体细节见附录A.1。

2.2. 删失多元高斯回归混合模型

令 G 表示簇的数量。随后，我们使用信息准则考察 G 的选择。于是，我们可以将式（1）中的模型扩展为混合模型：

$$\mathbf{y}_i^* | g = \boldsymbol{\beta}_g^\top \boldsymbol{x}_i + \boldsymbol{\epsilon}_{ig}, \ g \in \{1, .., G\}. \tag{3}$$

因此， $\mathbf{y}_i^* | \{g, \boldsymbol{x}_i\} \sim \mathcal{N}(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i, \ \boldsymbol{\Sigma}_g)$。假设观测值独立同分布，定义不完全数据似然函数：

$$\mathcal{L}(\mathbf{Y}; \mathbf{C}, \mathbf{X}, \boldsymbol{\Psi}) = \prod_{i=1}^{N} \sum_{g=1}^{G} \omega_g f_{\mathbf{y}_{io_i}}(\mathbf{y}_{io_i}; \boldsymbol{x}_i, \boldsymbol{\psi}_g) \int_{\mathcal{D}(\boldsymbol{c}_i)} f_{\mathbf{y}_{ic_i}^* | \mathbf{y}_{io_i}}(\mathbf{y}_{ic_i}; \boldsymbol{x}_i, \boldsymbol{\psi}_g), \tag{4}$$

其中， $\boldsymbol{\psi}_g$ 是各成分的向量化参数； $\omega_g \in (0, \ 1)$ 是各混合成分的混合比例，并满足约束：
$\sum_{g=1}^{G} \omega_g = 1$； $\boldsymbol{\Psi}$ 是总体参数向量，使得 $\boldsymbol{\Psi} = (\omega_1$、.....、 ω_{G-1}， $\boldsymbol{\psi}_1$、.....、 $\boldsymbol{\psi}_G)$。与前面介绍的回归模型不同，在混合模型设定中，无法直接最大化上述似然函数。因此，我们使用 EM算法进行参数估计。令 z_{ig} 表示一个指示变量：若第 i 个受试者属于第 g 组，则其值为 1，否则为 0；令 $\mathbf{Z}$ 表示通过堆叠 $[z_{i1}, \ldots, z_{iG}]^\top$形成的 $N \times G$矩阵。假设$\mathbf{Z}$ 和$\mathbf{Y}^*$ 均为观测值，我们可以写出

完整数据似然函数：

$$\mathcal{L}_c(\mathbf{Y}^*, \mathbf{Z}; \mathbf{C}, \mathbf{X}, \mathbf{\Psi}) = \prod_{i=1}^{N}\prod_{g=1}^{G}\left\{\frac{\omega_g}{\sqrt{2^p\pi^p|\mathbf{\Sigma}_g|}}e^{-\frac{1}{2}(\mathbf{y}_i^*-\boldsymbol{\beta}_g^\top\boldsymbol{x}_i)^\top\mathbf{\Sigma}_g^{-1}(\mathbf{y}_i^*-\boldsymbol{\beta}_g^\top\boldsymbol{x}_i)}\right\}^{z_{ig}}. \tag{5}$$

EM算法以 $\mathbf{\Psi}^{(0)}$ 作为初始值，随后在 E 步和 M 步之间迭代，直至收敛，具体如下。令 $\mathbf{\Psi}^{(k)}$ 表示前一次迭代中的参数值。令 Z_{ig} 表示一个随机指示变量：若第 i 个受试者属于第g组，则其值为 1，否则为 0。定义 $\langle z_{ig}\rangle \equiv \mathrm{E}_{\mathbf{\Psi}^{(k)}}(Z_{ig}|\mathbf{y}_i)$，其中 $\mathrm{E}_{\mathbf{\Psi}^{(k)}}$ 表示使用 $\mathbf{\Psi}^{(k)}$计算的期望。令 $\langle\mathbf{y}_i^*\rangle_g \equiv \mathrm{E}_{\mathbf{\Psi}^{(k)}}(\mathbf{y}_i^*|\mathbf{y}_i,\ Z_{ig}=1)$，并令 $\langle\mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g \equiv \mathrm{E}_{\mathbf{\Psi}^{(k)}}(\mathbf{y}_i^*\mathbf{y}_i^{*\top}|\mathbf{y}_i,\ Z_{ig}=1)$。对完整数据对数似然函数取条件期望：

$$\begin{aligned}\mathcal{Q}_c(\mathbf{\Psi};\mathbf{\Psi}^{(k)}) \propto \sum_{i=1}^{N}\sum_{g=1}^{G}\langle z_{ig}\rangle\{&\ln\omega_g - \frac{1}{2}\ln|\mathbf{\Sigma}_g| - \frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_g^{-1}\langle\mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g\right)\\ &+ \mathrm{tr}\left(\mathbf{\Sigma}_g^{-1}\boldsymbol{\beta}_g^\top\boldsymbol{x}_i\langle\mathbf{y}_i^*\rangle_g^\top\right) - \frac{1}{2}\mathrm{tr}\left(\mathbf{\Sigma}_g^{-1}\boldsymbol{\beta}_g^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\beta}_g\right)\}\end{aligned} \tag{6}$$

然后，EM算法的步骤如下：

- **E步**：

$$\langle z_{ig}\rangle \equiv E_{\mathbf{\Psi}^{(k)}}(Z_{ig}|\mathbf{y}_i) = \frac{\omega_g^{(k)}f_g(\mathbf{y}_i;\boldsymbol{x}_i,\boldsymbol{\psi}_g^{(k)})}{\sum_{h=1}^{G}\omega_h^{(k)}f_h(\mathbf{y}_i;\boldsymbol{x}_i,\boldsymbol{\psi}_h^{(k)})} \tag{7}$$

$$\langle\mathbf{y}_i^*\rangle_g \equiv E_{\mathbf{\Psi}^{(k)}}(\mathbf{y}_i^*|\mathbf{y}_i, Z_{ig}=1) = \begin{pmatrix}\mathbf{y}_{io_i}\\ \langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g\end{pmatrix} \tag{8}$$

$$\langle\mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g \equiv E_{\mathbf{\Psi}^{(k)}}(\mathbf{y}_i^*\mathbf{y}_i^{*\top}|\mathbf{y}_i, Z_{ig}=1) = \begin{pmatrix}\mathbf{y}_{io_i}\mathbf{y}_{io_i}^\top & \mathbf{y}_{io_i}\langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g^\top\\ \langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g\mathbf{y}_{io_i}^\top & \langle\mathbf{y}_{ic_i}^*\mathbf{y}_{ic_i}^{*\top}|\mathbf{y}_{io_i}\rangle_g\end{pmatrix}, \tag{9}$$

其中，$\langle\mathbf{y}_{ic_i}^*\mathbf{y}_{ic_i}^{*\top}|\mathbf{y}_{io_i}\rangle_g$ 和 $\langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g$ 分别是截断条件分布的第一阶矩和第二阶矩

高斯分布。这些矩使用 R 包 MomTrunc（Galarza 等，2021年）计算。

- **M 步**：

$$\widetilde{\omega}_g = \sum_{i=1}^{N} \langle z_{ig} \rangle / N \tag{10}$$

$$\widetilde{\boldsymbol{\beta}}_g = (\mathbf{X}^\top \boldsymbol{\mathcal{Z}}_g \mathbf{X})^{-1} \mathbf{X}^\top \boldsymbol{\mathcal{Z}}_g \langle \mathbf{Y}^* \rangle_g \tag{11}$$

$$\widetilde{\boldsymbol{\Sigma}}_g = \frac{\sum_{i=1}^{N} \langle z_{ig} \rangle [\langle \mathbf{y}_i^* \mathbf{y}_i^{*\top} \rangle_g - \widetilde{\boldsymbol{\beta}}_g^\top \boldsymbol{x}_i \langle \mathbf{y}_i^* \rangle_g^\top - \langle \mathbf{y}_i^* \rangle_g \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_g + \widetilde{\boldsymbol{\beta}}_g^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}_g]}{\sum_{i=1}^{N} \langle z_{ig} \rangle} \tag{12}$$

其中，$\langle \mathbf{Y}^* \rangle_g$ 是 $N \times p$ 条件期望堆叠而成的矩阵 $\langle \mathbf{y}_i^* \rangle_g$，且$\boldsymbol{\mathcal{Z}}_g$ $=\text{diag}(\langle z_{1g} \rangle, ..., \langle z_{Ng} \rangle)$。更多详情请参见附录A.2。

2.3. 簇内效应的假设检验

与某些会自动提取信息矩阵的极大似然估计算法（例如牛顿-拉夫森更新）不同，EM算法不会直接计算信息矩阵。作为替代，一些作者采用自助法（McLachlan 和 Peel，2000年；O'Hagan 等人，2019年），这种方法通常可靠，但计算成本高昂。相反，我们使用经验完全数据得分函数近似观测信息矩阵。

在满足温和的正则条件，且作为参数空间 $\widehat{\boldsymbol{\Psi}} \in \text{int}(\boldsymbol{\Theta})$ 内部全局最大值点的极大似然估计量弱相合的情况下，使得 $\widehat{\boldsymbol{\Psi}} \xrightarrow{p} \boldsymbol{\Psi}_0 \in \boldsymbol{\Theta}$，则：

$$\frac{\sum_{i=1}^{N} \boldsymbol{s}_c(\widehat{\boldsymbol{\Psi}}; \boldsymbol{\Psi}^{(k)} = \widehat{\boldsymbol{\Psi}}, \mathbf{y}_i) \boldsymbol{s}_c^\top(\widehat{\boldsymbol{\Psi}}; \boldsymbol{\Psi}^{(k)} = \widehat{\boldsymbol{\Psi}}, \mathbf{y}_i)}{N} \xrightarrow{p} \mathcal{I}(\boldsymbol{\Psi}_0) \tag{13}$$

其中 $\boldsymbol{\Psi}_0$ 是真实参数向量； $\boldsymbol{\Theta}$ 是参数空间； $\boldsymbol{s}_c(\boldsymbol{\Psi};\ \boldsymbol{\Psi}^{(k)},\ \mathbf{y}_i) \equiv \frac{\partial Q_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)},\ \mathbf{y}_i)}{\partial \boldsymbol{\Psi}}$ $=$

$\frac{\partial E_{\Psi^{(k)}}(\log \mathcal{L}_{c_i}|\mathbf{y}_i)}{\partial \Psi}$ 是完整数据对数似然的个体条件期望相对于感兴趣参数的一阶导数。详情请参见McLachlan 和 Peel (2000年)中的公式2.60。随后，我们对簇内效应进行 Wald 检验。这种方法避免了计算二阶偏导数，且在计算上可行。McLachlan 和 Peel (2000年)指出，混合模型要进行有效推断，样本量必须足够大。我们的数据应用包含 $N > 3,000$。在第3节中，我们表明，在 $N = 1,000$的情形下，第一类错误率通常接近其名义水平。

3. 模拟

3.1. 模拟设计

我们通过两种模拟情景——轻度删失和重度删失——考察删失多元高斯混合回归及其估计量。在这两种情景中，首先从具有双变量响应的三簇回归混合模型中生成 $N = 1{,}000$ 个未观测数据（$\mathbf{Y}_1$,$\mathbf{Y}_2$），这些响应服从高斯分布，其均值是截距$\boldsymbol{\beta}_{g,0}$ 和三个预测变量（ $g \in \{1, 2, 3\}$，其中$\mathbf{X}_1$, $\mathbf{X}_2, \mathbf{X}_3$）的线性变换。模拟设计详见表2，并概述如下。预测变量由均值为零的多元高斯分布生成，其协方差矩阵等于基于 Emory ADRC/HBS 三项连续人口统计学特征（年龄、教育程度和蒙特利尔认知评估）计算的样本相关矩阵。真实参数见表2。

在情景 I 中，对第一个响应变量 $\mathbf{Y}_1$,采用下限为 0 的检测限，$\mathbf{Y}_1$, 约有 4.1% 的观测值为左删失；对 $\mathbf{Y}_2$ 采用上限为 30 的检测限，$\mathbf{Y}_2$约有 13.7% 的观测值为右删失。这使删失程度与真实数据集（第 4节）中的情况相近。在情景 II 中，对 $\mathbf{Y}_1$ 采用下限为 2.5 的检测限，对 $\mathbf{Y}_2$ 采用上限为 26.5 的检测限，分别约有 40.2% 的观测值为左删失，37.2% 的观测值为右删失。每种情景下，我们进行了 101 次模拟，并在图中使用中位数误差结果。附录图 C1展示了一次示例模拟中未观测真实值与删失数据的散点图。

我们将删失多元高斯混合回归与三种方法进行比较：忽略删失的多元回归混合模型，即以与未删失观测值相同的方式处理删失观测值，并使用所有观测值（忽略删失 GMR）；删除删失观测值的多元高斯混合模型（删除删失的GMR）；以及忽略预测变量的删失多元高斯混合模型（删失GMM）（Lee 和 Scott，2012年）。我们报告模拟中混合比例 ω_1、ω_2 和 ω_3 的均值和标准差（SD）。其他参数则报告弗罗贝尼乌斯误差的均值和标准差。我们使用调整兰德指数（ARI）评估总体聚类准确率，即比较未观测的真实标签与模型标签；每个观测值均被分配到 $\mathrm{E}\hat{\Psi}(Z_{ig}|\mathbf{y}_i)$ 最高的类别。

我们还利用500次模拟，考察在真实系数值均等于0的参数下，删失多元高斯混合回归的第一类错误率。

我们还考察聚类数的选择。我们对以下 $g = 1, \ldots\ldots, 6$拟合删失多元高斯混合回归。对于101次重复试验中的每一次以及每个 g，模型均以随机初始值运行32次，并从所有收敛解中选取似然最高的解。随后，我们计算积分完整似然准则（Biernacki等人，2000年）。

3.2. 模拟结果

在情景 I 中，删失多元高斯混合回归的混合比例无偏且标准差较小；忽略删失的 GMR 得出的混合比例与之相近，而删除删失的GMR的偏差有所增大，高估了第2组的频率、低估了第3组的频率；删失GMM得出的混合比例则极不准确（表 3）。对于回归系数，删失多元高斯混合回归比其他方法更准确。忽略删失的 GMR 和删除删失的GMR在第3组上的估计尤其不准确；该组的回归系数导致其删失程度高于第1组和第2组（表 3）。对于协方差矩阵，删失多元高斯混合回归明显比其他方法更准确。

在情景 II 中，我们观察到类似的模式，但删失多元高斯混合回归的优势更加显著（图 1，表 3）。删失多元高斯混合回归仍能准确估计混合比例，而忽略删失 GMR 会导致严重的误差。删失 GMM

高估了第2组的比例。总体而言，情景 II 的调整兰德指数低于情景 I，但删失多元高斯混合回归的调整兰德指数（0.68）远高于忽略删失 GMR（0.08）和删失GMM（0.17）。删除删失的GMR无法对约 575 个观测值进行聚类。删失多元高斯混合回归仍能准确估计回归系数和协方差矩阵，而其他方法的估计则变得非常不准确。

这些结果凸显了使用删失多元高斯混合回归模型同时对删失和预测变量进行建模的必要性。即使研究并不关注预测变量的影响，为了准确聚类，也需要采用回归混合模型。此外，还必须对删失进行建模，否则聚类结果和回归系数估计都会非常不准确。

在情景 I 和 II 中，所有第一类错误率都接近名义水平，最高的第一类错误率为 0.056（附录表 B1）。

在为删失多元高斯混合回归模型选择成分数时，ICL 在情景 I 和 II 中均能选出正确的成分数（附录图 C2）。

4. 真实数据应用：埃默里大学ADRC/HBS数据集

本分析有两个目的：1）根据患者的脑脊液生物标志物识别患者群组，而不采用可能不准确的临床诊断；2）评估预测变量在各群组内对脑脊液生物标志物的影响。埃默里大学ADRC/HBS数据集包含3,004名个体，其中包括888名阿尔茨海默病患者、661名患有其他认知障碍（其他）的个体，以及1,455名认知正常（CN）个体。诊断基于临床病史、神经心理学测试、血液检查、结构性神经影像学和脑脊液生物标志物的综合结果。纳入我们分析的所有个体均已提供知情同意，同意参与经埃默里大学机构审查委员会批准的研究方案（EHBS IRB00080300、ADRC IRB00079069、NeuCog IRB00078273、Vascular IRB00084414、CRIN IRB00024959）。所有脑脊液样本均采用罗氏Elecsys平台进行检测，以测量β-淀粉样蛋白 1-42（Abeta42）、总 tau 蛋白（tTau）和磷酸化 tau 蛋白（pTau）的水平。由于生物测定的检测限，脑脊液生物标志物存在删失：Abeta42的3,004个观测值中有10个为左删失，349个为右删失；tTau的3,004个观测值中有31个为左删失，4个为右删失；而3,004个

pTau的观测值中有110个为左删失，5个为右删失（图2、表4）。较低的脑脊液 Abeta42对应较高的脑内β-淀粉样蛋白42负荷，因此我们预计阿尔茨海默病患者的Abeta42水平较低。相比之下，较高的脑脊液 tTau 和 pTau与较高的脑内 Tau 蛋白水平相关，因此我们预计阿尔茨海默病患者的tTau和pTau水平较高。一种常见方法是使用Abeta42/tTau或Abeta42/pTau的比值，其中较小的数值表示阿尔茨海默病病理（Hampel 等人、2008年；Meyer 等人、2010）。然而，使用比值会掩盖删失情况，并丢弃多变量方法所能提供的信息。由于在我们的模拟中忽略删失会导致错误的聚类，我们认为，采用删失多元高斯混合回归的多变量方法将更好地反映潜在的生物学机制。

数据包含年龄（十年）、受教育程度（十年）、自我报告的种族、性别和载脂蛋白基因类型。由于美国印第安人或阿拉斯加原住民（$n = 6$）、亚洲人（$n = 36$）以及夏威夷原住民或其他太平洋岛民（$n = 7$）的样本量较小，我们创建了一个二元种族变量：参与者为非裔美国人时取值为 1，否则取值为 0。此外，有 82 名参与者自我报告为西班牙裔或拉丁裔，但由于样本量原因未对其进行分析。我们纳入了四种状态：不携带 $\varepsilon 4$等位基因、ApoE4 杂合子（ApoE4-1）、ApoE4 纯合子（ApoE4-2）以及缺失数据（表 4）。所有参与者均提供了知情同意，且所有程序均经埃默里大学机构审查委员会批准。

为选取最优聚类数，我们计算了聚类数为 G =1、.....、6 时的 ICL。对于每个 G，我们针对三种脑脊液生物标志物，估计一个截距和五个人口统计学变量的回归系数，并估计 3×3 协方差矩阵。模型随机初始化 50 次，并从收敛的解中选取似然最高的解。（当 G=1 为 4 时，50 次迭代均收敛；当 G=5时，45/50 次收敛；当 G=6时，35/50 次收敛。）

ICL 确定的最佳聚类数为三（附录图 C3）。

为了深入理解这些组的含义，我们将它们与三种（可能不正确的）临床诊断之间的关系进行可视化（图 3）。图 A 展示了按诊断组着色的 Abeta42 与 tTau 散点图。图 D 展示了同一散点图，但按删失多元高斯混合回归簇着色。图 B 和图 E 分别展示了阿尔茨海默病诊断和删失多元高斯混合回归中的 Abeta42 与磷酸化 tau 的关系；图 C 和图 F 则分别展示了阿尔茨海默病诊断和删失多元高斯混合回归中的 tTau 与磷酸化 tau 的关系。

我们看到，图 A 和图 B 中的阿尔茨海默病标签（粉色）与图 D 和图 E 中的粉色删失多元高斯混合回归簇相吻合。因此，我们将图 D-F 中的粉色簇称为“阿尔茨海默病样病理”组。

图 D-F 中的绿色组（图3 ）倾向于与图 A-C 中的“正常”组相对应。我们将其称为“对照样”组。

我们将图3 中图 D-F 的蓝色组称为“非阿尔茨海默病病理”组。表5 中的截距表明，与对照样相比，该组的脑脊液 Abeta42 水平较高，即脑内 Abeta 负荷较低，这通常被认为是非阿尔茨海默病病理。然而，该组的脑脊液 tTau 和磷酸化 tau 水平均高于阿尔茨海默病样组和对照样组，表明其脑内 Tau 蛋白负荷较高，这可能与其他类型的痴呆或神经功能损害有关。

第1组：阿尔茨海默病样病理。非裔美国人参与者的 Abeta42 水平与其他种族组成的组相似，而非裔美国人参与者的 tTau 和磷酸化 tau 水平则显著较低（表5，图4）。低 Abeta42/tTau 比值常用于将个体归类为阿尔茨海默病患者。在这个阿尔茨海默病组中，非裔美国参与者的 Abeta42/tTau 比值会高于其他种族，这意味着传统比值可能会错误分类非裔美国患者。非裔美国人的身份对 tTau 和磷酸化 tau 的影响，相当于年龄减少 18 岁（表5）。从临床角度来看，这组非裔美国人参与者很可能患有阿尔茨海默病，但 Tau 蛋白负荷较低，因此被诊断为阿尔茨海默病的可能性可能较小；这表明 Tau 蛋白可能不是非裔美国人群中诊断阿尔茨海默病的良好生物标志物。这可能会对使用脑脊液生物标志物比值分类阿尔茨海默病的传统方法产生重大影响，因为这些传统方法主要基于参与者以欧洲血统为主的研究。

载脂蛋白 E4 对这三种生物标志物均有显著影响：载脂蛋白 E4 会降低脑脊液 Abeta42 水平，并升高磷酸化 tau 和 tTau 水平。载脂蛋白 E4-2 的系数绝对值均大于载脂蛋白 E4-1（表5）。在类 AD 组中，携带两份载脂蛋白 E4 的人，其 tTau 和磷酸化 tau 水平均远高于类对照组中携带两份载脂蛋白 E4 的人。APOE4 缺失（即未收集这些数据的患者）的系数大于 APOE4-1。进一步调查发现，APOE4 的缺失并非随机：约 40% 的患者被诊断为 AD，38% 患有其他病理，这表明 APOE4-2 的频率可能有所升高

与观测到的频率相比，在这一组中。

该组的其他重要发现包括，年龄与 Abeta42 之间没有关联，但与 tTau 和磷酸化 tau 呈正相关。与其他组相比，年龄对 tTau 和磷酸化 tau 的系数较大，这反映了阿尔茨海默病组中 Tau 蛋白病理进展更快。

第2组：类对照组。与类 AD 组相比，在类对照组中，非裔美国人参与者的脑脊液 Abeta42 显著低于其他种族参与者组成的组。非裔美国人参与者的总 Tau 蛋白和磷酸化 tau 也显著降低，但系数小于类 AD 组。这再次表明，非裔美国人参与者的阿尔茨海默病病理有所不同，也凸显了回归混合模型的重要性。此外，类对照组中女性的脑脊液 Abeta42 水平显著高于男性。

载脂蛋白 E4 携带者的脑脊液 Abeta42 显著降低，但与类 AD 组不同，其 Tau 蛋白水平没有变化。此前的研究发现，载脂蛋白 E4 与认知正常老龄化中的 Abeta42 相关，但与 Tau 蛋白无关（Morris 等人，2010）。

与“阿尔茨海默病样病理”组相比，类对照组的脑脊液 Abeta42 会随年龄增长而降低（导致脑内β-淀粉样蛋白42增加）。由于类 AD 组的脑脊液 Abeta42 基线水平低得多，因此在较高年龄时，类对照组的脑脊液 Abeta42 水平仍高于类 AD 组。总 Tau 蛋白和磷酸化 tau 会随年龄增长而升高，但速度低于类 AD 组，这反映了类对照组中随年龄进展的 Tau 蛋白病理。

第3组：非阿尔茨海默病病理。总体而言，我们未发现该组中预测变量与脑脊液生物标志物之间存在显著关系（表 5，图 4）。这可能部分是由于混合比例较小，意味着样本量较小，系数估计不精确。

5. 讨论

我们采用一种带有可行EM算法的删失多元高斯回归混合模型，考察预测变量对阿尔茨海默病脑脊液生物标志物亚组的影响。该方法类似于估计多元回归效应，即所有预测变量均与组变量交互，但在这里我们还会学习组标签。我们的方法能够同时识别

簇，并允许预测变量对具有检测限的多变量结局在不同簇中产生不同影响。不同于耗时的自助法，我们使用经验完全得分函数近似簇内效应 β_g 的渐近协方差矩阵。模拟结果表明，在大样本（$n \approx 1,\ 000$）中，该方法能够充分控制第一类错误率。在删失程度中等（与我们的数据应用相当）和重度的模拟中，我们表明，忽略删失记录、删除删失记录或忽略预测变量都会造成显著误差。我们的方法大幅提升了聚类准确性和回归估计值的准确性。我们的模拟为潜在类别文献增添了证据，表明对删失和预测变量进行建模对于准确聚类都至关重要。

我们使用删失多元高斯混合回归对埃默里大学ADRC/HBS数据集进行分析，揭示了新的见解。我们发现了三个簇，它们分别倾向于对应类 AD 组、类对照组，以及一个病理类型未明确的非阿尔茨海默病病理组。不同簇中的预测变量效应各不相同。类 AD 组中的非裔美国人参与者，其 tTau 和磷酸化 tau 病理程度较轻。脑脊液生物标志物通常采用 Abeta42 与 Tau 蛋白的比值，但以往研究可能依据对非西班牙裔白人的研究来进行此类判定（Meyer 等人，2010）。近期，一些研究人员报告了脑脊液生物标志物可能存在种族差异（Morris 等人，2019年；Garrett 等人，2019年），这与我们的发现一致。此外，载脂蛋白 E4 对脑脊液生物标志物水平的影响在类 AD 组和类对照组之间有所不同；类对照组中的女性脑脊液 Abeta42 水平高于男性；年龄对类 AD 组中的 Abeta42 没有影响，但在类对照组中影响显著；年龄对磷酸化 tau 和 tTau 的影响在类 AD 组中最大。

我们发现，类 AD 组中患者的比例高于被诊断为阿尔茨海默病的患者比例。这是意料之中的，因为相当一部分认知正常者患有无症状阿尔茨海默病（Jansen 等，2022年，2015年）。同样，临床诊断为“其他”的绝大多数人预计会归入非阿尔茨海默病病理或对照样的多变量脑脊液分布。删失多元高斯混合回归的一个优点是，它能够生成个体属于各个聚类的概率。未来的研究可以开发交互式工具，让临床医生输入患者数据，并获得患者归属于类 AD 组、对照样组和非阿尔茨海默病病理组的概率，从而补充现有的阿尔茨海默病诊断方法。

我们的方法存在若干局限。随着响应变量、预测变量和组的数量增加，模型选择和解释可能会变得困难。在高维情形下，惩罚方法可能有所帮助（Khalili 和 Lin，2013；Xie 等，2010）。未来研究的另一个方向是考虑一种替代方法，使用多项回归将潜在类别成员资格的概率建模为协变量的函数（Jacobs 等，1991）。在我们的方法中，预测变量会间接影响后验概率。因此，所得效应更易于解释，有助于加深我们对人口统计特征、行为和遗传学对复杂神经系统疾病中生物标志物影响的理解。

6. 软件

第 3 节中使用的代码可在`https://github.com/GanzhongTian/CensGMR`获取。

利益冲突声明

J. Lah 是罗氏诊断的顾问。

资助

G.T. 获得了 R01 AG055634 和 R01 AG070937 的资助。B.B.R. 获得了 R21 AG066970 的资助。J.H. 获得了 R01 AG055634 和 P50 AG025688 的资助。J.L. 获得了 R01 AG070937 和 P50 AG025688 的资助。用于罗氏 Elecsys 检测的材料经费由 J.L. 获得的罗氏 IIS 资助 RD004723 提供。本研究的额外经费来自 Goizueta 基金会的慷慨捐赠。

表1

软件包	删失值结果	多元结果	混合回归	截断多元正态
mclust	✗	✓	✗	✗
mixtools	✗	✓	✓	✗
FlexMix	✗	✗	✓	✗
SMNCensReg	✓	✗	✗	✗
CensMFM	✓	✓	✗	✗
poLCA	✗	✗	✓	✗
CensMixReg	✓	✗	✓	✗
有限混合模型（Stata）	✓	✗	✓	✗
proc fmm（SAS）	✓	✗	✓	✗
Latent GOLD 6.0	✓	✓	✓	✗

表1：有限混合建模软件列表。mclust：Scrucca 等人(2016）；mixtools：Benaglia 等人(2010）；FlexMix：Grun 和 Leisch(2008年）；SMNCensReg：Garay 等人(2013）；CensMFM：De Alencar 等人(2020）；poLCA：Linzer 和 Lewis(2011）；CensMixReg：Sanchez 等人(2015年）；Latent GOLD 6.0：Vermunt 和 Magidson(2021年）。

表 2

模拟情形，$N = 1{,}000$, $\mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$		
参数	真实值	
$\boldsymbol{\pi} = (\omega_1,\ \omega_2,\ \omega_3)$	$(0.1, 0.7, 0.2)$	
$\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \boldsymbol{\beta}_3)$	$\begin{bmatrix} \begin{pmatrix} 2 & 20 \\ 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 3 & 25 \\ 1 & -3 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 3.5 & 30 \\ 2 & -5 \\ 0 & 0 \\ 0 & 0 \end{pmatrix} \end{bmatrix}$	
$\boldsymbol{\Sigma} = (\boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2, \boldsymbol{\Sigma}_3)$	$\begin{bmatrix} \begin{pmatrix} 1 & 0.1 & \\ 0.1 & & \\ & & 1 \end{pmatrix}, \begin{pmatrix} 2 & & 0.2 \\ & & \\ 0.2 & & 0.5 \end{pmatrix}, \begin{pmatrix} 0.5 & & 0.3 \\ & & \\ 0.3 & & 2 \end{pmatrix} \end{bmatrix}$	
	情景 I	情景 II
检测限	$\mathbf{Y}_1 \in (0, \infty)$, $\mathbf{Y}_2 \in (-\infty, 30)$	$\mathbf{Y}_1 \in (2.5, \infty)$, $\mathbf{Y}_2 \in (-\infty, 26.5)$
$\mathbf{Y}_1$ 删失值 %	左删失 4.1%	左删失 40.2%
$\mathbf{Y}_2$ 删失值 %	右删失 13.7%	右删失 37.2%

表 2：模拟情景摘要。

表3

情景 I					
参数	**真实值**	**删失 GMR**	**忽略删失 GMR**	**删除删失的GMR**	**删失 GMM**
N	-	1000	1000	852	1000
ω_1	0.10	**0.10(0.00)**	0.10(0.00)	0.12(0.00)	0.16(0.11)
ω_2	0.70	**0.70(0.01)**	0.72(0.02)	0.77(0.01)	0.56(0.18)
ω_3	0.20	**0.20(0.01)**	0.18(0.02)	0.11(0.02)	0.29(0.17)
$\|\|\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1\|\|_F$	-	**0.44(0.14)**	0.53(0.19)	0.44(0.15)	-
$\|\|\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_2\|\|_F$	-	**0.19(0.09)**	0.26(0.07)	0.23(0.08)	-
$\|\|\boldsymbol{\beta}_3 - \widehat{\boldsymbol{\beta}}_3\|\|_F$	-	**0.53(0.26)**	3.32(0.37)	1.61(0.83)	-
$\|\|\boldsymbol{\Sigma}_1 - \widehat{\boldsymbol{\Sigma}}_1\|\|_F$	-	**0.24(0.09)**	0.42(0.24)	0.28(0.10)	7.24(10.32)
$\|\|\boldsymbol{\Sigma}_2 - \widehat{\boldsymbol{\Sigma}}_2\|\|_F$	-	**0.11(0.06)**	0.16(0.06)	0.21(0.08)	8.34(2.74)
$\|\|\boldsymbol{\Sigma}_3 - \widehat{\boldsymbol{\Sigma}}_3\|\|_F$	-	**0.35(0.22)**	0.71(0.28)	0.70(0.39)	18.20(13.16)
$\|\|\boldsymbol{\Psi} - \widehat{\boldsymbol{\Psi}}\|\|_F$	-	**0.88(0.23)**	3.48(0.34)	1.91(0.81)	-
ARI	1	**0.89(0.02)**	0.82(0.03)	-	0.31(0.15)

情景 II					
参数	**真实值**	**删失 GMR**	**忽略删失 GMR**	**删除删失的GMR**	**删失 GMM**
N	-	**1000**	1000	418	1000
ω_1	0.10	**0.10(0.00)**	0.39(0.06)	0.08(0.02)	0.15(0.08)
ω_2	0.70	**0.70(0.02)**	0.42(0.04)	0.73(0.06)	0.53(0.18)
ω_3	0.20	**0.20(0.02)**	0.19(0.07)	0.19(0.06)	0.32(0.18)
$\|\|\boldsymbol{\beta}_1 - \widehat{\boldsymbol{\beta}}_1\|\|_F$	-	**0.52(0.18)**	3.51(0.32)	1.52(0.42)	-
$\|\|\boldsymbol{\beta}_2 - \widehat{\boldsymbol{\beta}}_2\|\|_F$	-	**0.23(0.09)**	1.21(0.26)	0.92(0.14)	-
$\|\|\boldsymbol{\beta}_3 - \widehat{\boldsymbol{\beta}}_3\|\|_F$	-	**0.84(0.43)**	6.03(0.71)	4.81(1.20)	-
$\|\|\boldsymbol{\Sigma}_1 - \widehat{\boldsymbol{\Sigma}}_1\|\|_F$	-	**0.40(0.21)**	3.06(0.47)	1.11(0.65)	23.75(94.17)
$\|\|\boldsymbol{\Sigma}_2 - \widehat{\boldsymbol{\Sigma}}_2\|\|_F$	-	**0.14(0.08)**	1.80(0.53)	0.95(0.18)	6.53(2.95)
$\|\|\boldsymbol{\Sigma}_3 - \widehat{\boldsymbol{\Sigma}}_3\|\|_F$	-	**0.59(0.40)**	2.40(1.30)	0.79(0.40)	4.57(4.13)
$\|\|\boldsymbol{\Psi} - \widehat{\boldsymbol{\Psi}}\|\|_F$	-	**1.32(0.43)**	8.42(0.74)	5.49(1.01)	-
ARI	1	**0.68(0.03)**	0.12(0.04)	-	0.17(0.07)

表3：删失多元高斯混合回归模型以及另外三种方法在情景I和II中的模拟结果。忽略删失的多元高斯混合回归采用高斯回归混合，并将删失数据按与未删失数据相同的方式处理。删除删失观测的多元高斯混合回归采用高斯回归混合，但删除删失观测。删失多元高斯混合模型采用删失高斯混合，而不考虑预测变量的影响，因此只有ω_g、 $\boldsymbol{\Sigma}_g$ 和调整兰德指数可与其他方法进行比较。报告的是跨模拟的ω_g均值（标准差）估计值、弗罗贝尼乌斯误差的均值（标准差）以及来自101次重复试验的调整兰德指数（调整兰德指数）。与中位数误差对应的重复试验的聚类结果见图1。由于删除删失观测的多元高斯混合回归无法对删失观测进行聚类，因此不报告其调整兰德指数。19

表4

变量		N=3,004
Abeta42:		
	# < 200:	10
	未删失:	869.4（587.2，1291.0）
	# > 1，700:	349
tTau:		
	# < 80:	31
	未删失:	206.2（153.7，288.3）
	# > 1，300:	4
磷酸化 tau:		
	# < 8:	110
	未删失:	17.84（13.10，26.78）
	# > 120:	5
临床诊断:		
	AD:	888 (0.30)
	其他:	661 (0.22)
	CN:	1,455 (0.48)
年龄（十年）		6.61 (5.94,7.18)
教育程度（十年）		1.60 (1.40,1.80)
种族:		
	美国印第安人或阿拉斯加原住民:	6
	亚洲人:	36
	黑人或非裔美国人:	495
	白人:	2,454
	夏威夷原住民或其他太平洋岛民:	7
	其他:	6
性别:		
	女性:	1,788
	男性:	1,216
载脂蛋白 E4:		
	$\varepsilon4/\varepsilon4$:	193
	$\varepsilon3/\varepsilon4$或 $\varepsilon2/\varepsilon4$:	900
	$\varepsilon4$ 阴性:	1,520
	缺失数据:	391

表4：埃默里大学戈伊祖埃塔阿尔茨海默病研究中心和埃默里健康大脑研究数据集的患者人口统计特征。

表5

	Abeta42	tTau	pTau
第1组：类阿尔茨海默病		$\omega_1 = 0.370$	
截距	632.55***	260.22***	24.57***
年龄（十年）	3.79	29.65***	3.35***
教育年限（十年）	18.84	-18.77	-1.7
女性	31.6*	18.26*	1.29
非裔美国人	-59.1*	-63.95***	-6.53***
Apoe4-2	-140.69***	87.35***	9.51***
Apoe4-1	-23.41	53.68***	6.19***
载脂蛋白 E4 缺失	-72.22***	29.54*	3.83**
第2组：对照样		$\omega_2 = 0.577$	
截距	1284.74***	179.31***	15.43***
年龄（十年）	-58.96***	11.76***	1***
教育年限（十年）	9.84	-2.28	-0.07
女性	111.74***	7.45**	0.66*
非裔美国人	-190.45***	-26.57***	-2.24***
Apoe4-2	-684.8***	0.32	0.55
Apoe4-1	-260.1***	-0.56	-0.07
APOE4 缺失	-222.31***	-1.33	-0.45
第3组：非阿尔茨海默病病理		$\omega_3 = 0.053$	
截距	1116.67***	556.61***	53.79***
年龄（十年）	28.33	4.54	1.98
教育年限（十年）	260.32	-47.64	-4.97
女性	-9.66	22.83	0.78
非裔美国人	350.04	-245.07	-23.48
Apoe4-2	-627.41	281.46*	20.89
Apoe4-1	-156.82	32.89	2.64
Apoe4 缺失	62.93	66.62	-5.24

表5：估计值　　d 系数矩阵。*：$p < 0.05$；**：$p < 0.01$；***：$p < 0.001$，未经校正　　.

图 1

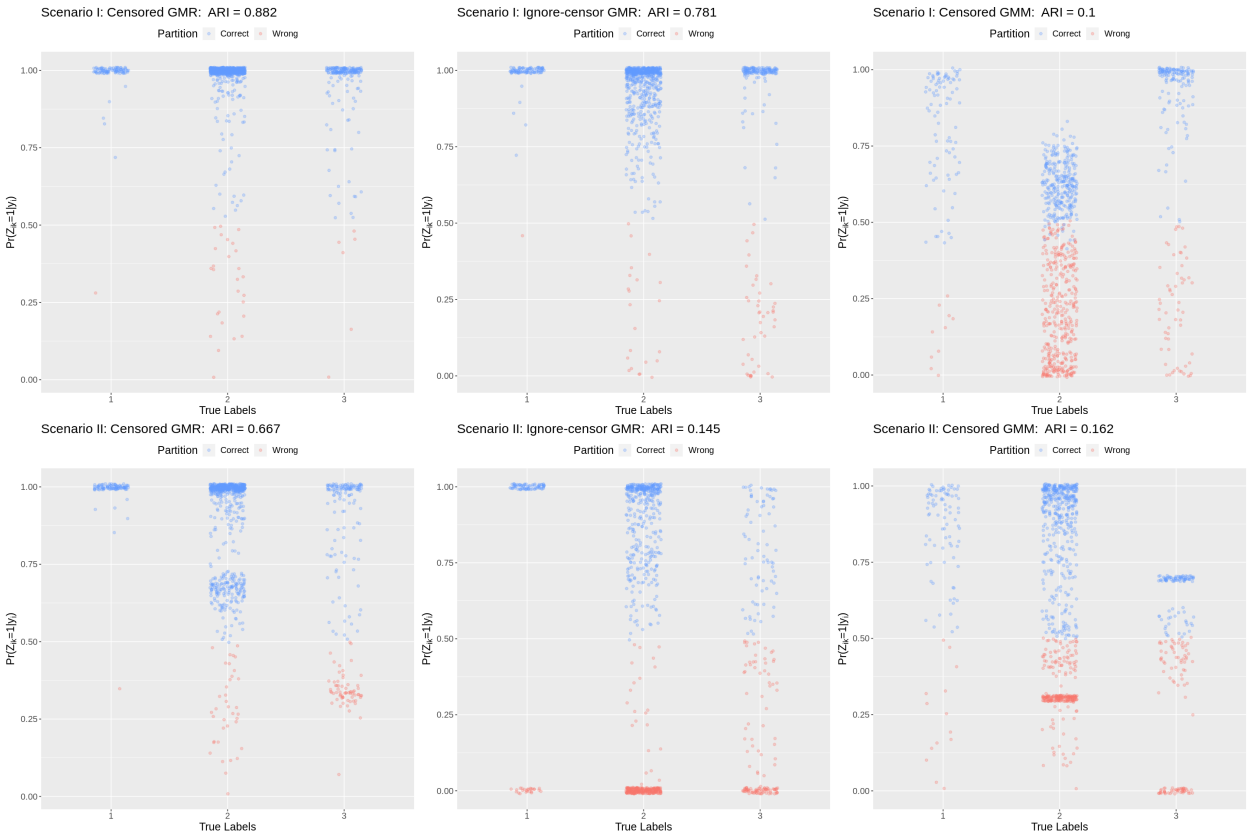


图 1：删失多元高斯混合回归、忽略删失的多元高斯混合回归和删失多元高斯混合模型的聚类准确率。展示的是 101 次模拟中与中位调整兰德指数相关的结果。蓝色表示根据聚类归属的后验概率正确分类的观测，红色表示分类错误。为改善可视化效果，添加了抖动。

图 2

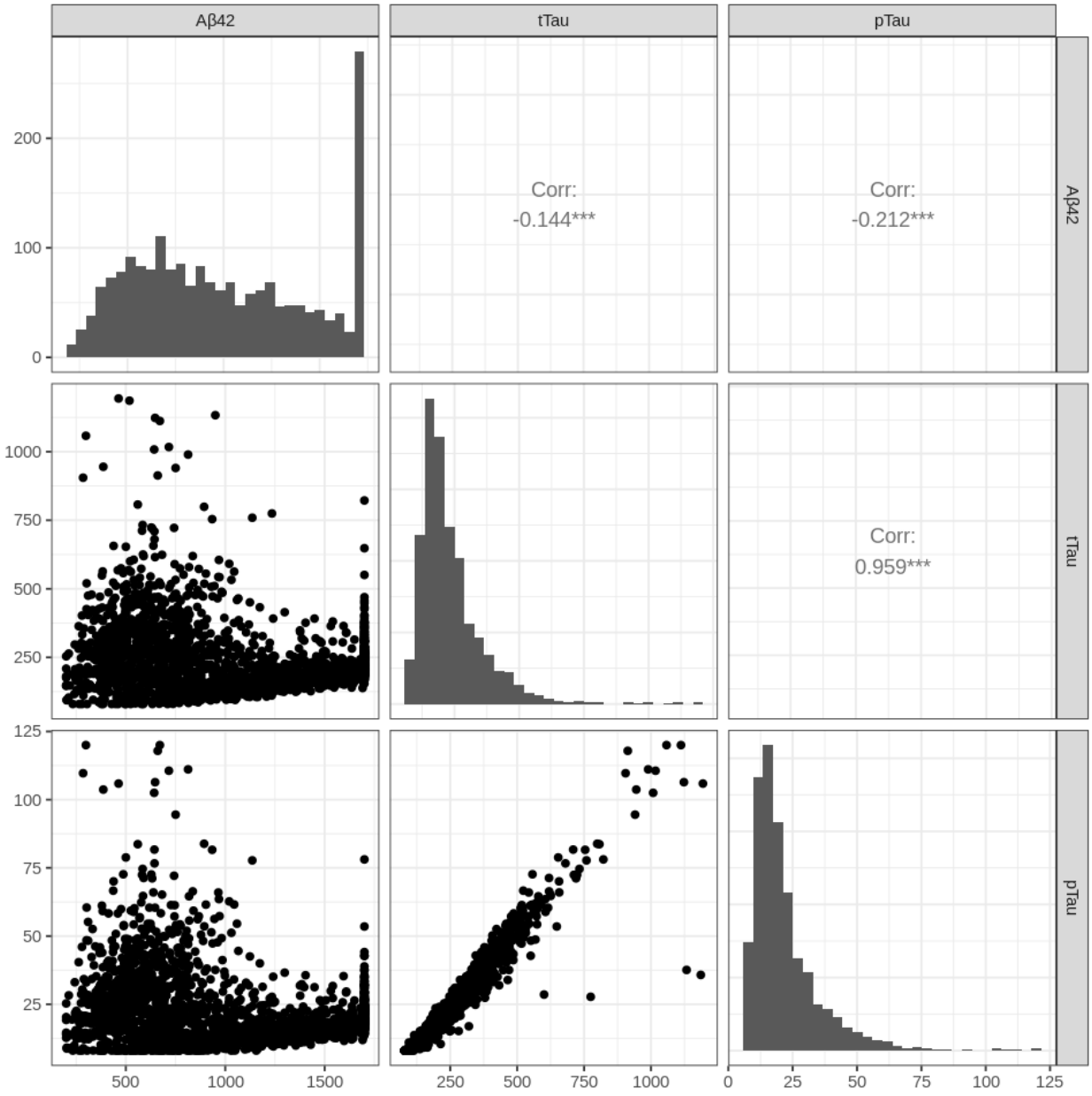


图 2：来自 Emory ADRC/HBS 研究的脑脊液生物标志物数据。

图 3

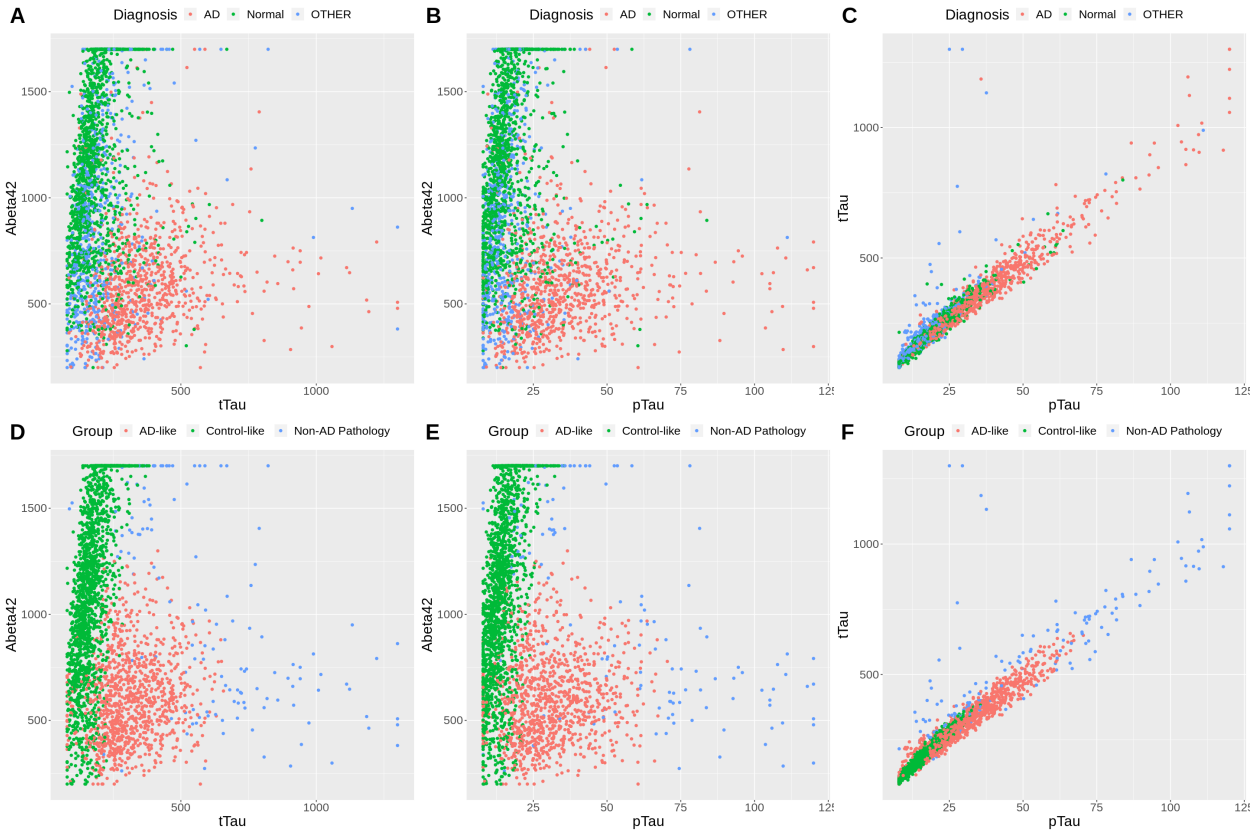


图 3：诊断与使用 $G = 3$ 识别出的簇的散点图。上排的点按诊断标签着色。下排使用删失多元高斯混合回归的输出。簇的描述“阿尔茨海默病样”“对照样”和“非阿尔茨海默病病理”是通过检查各簇的特征确定的；参见正文。

图 4

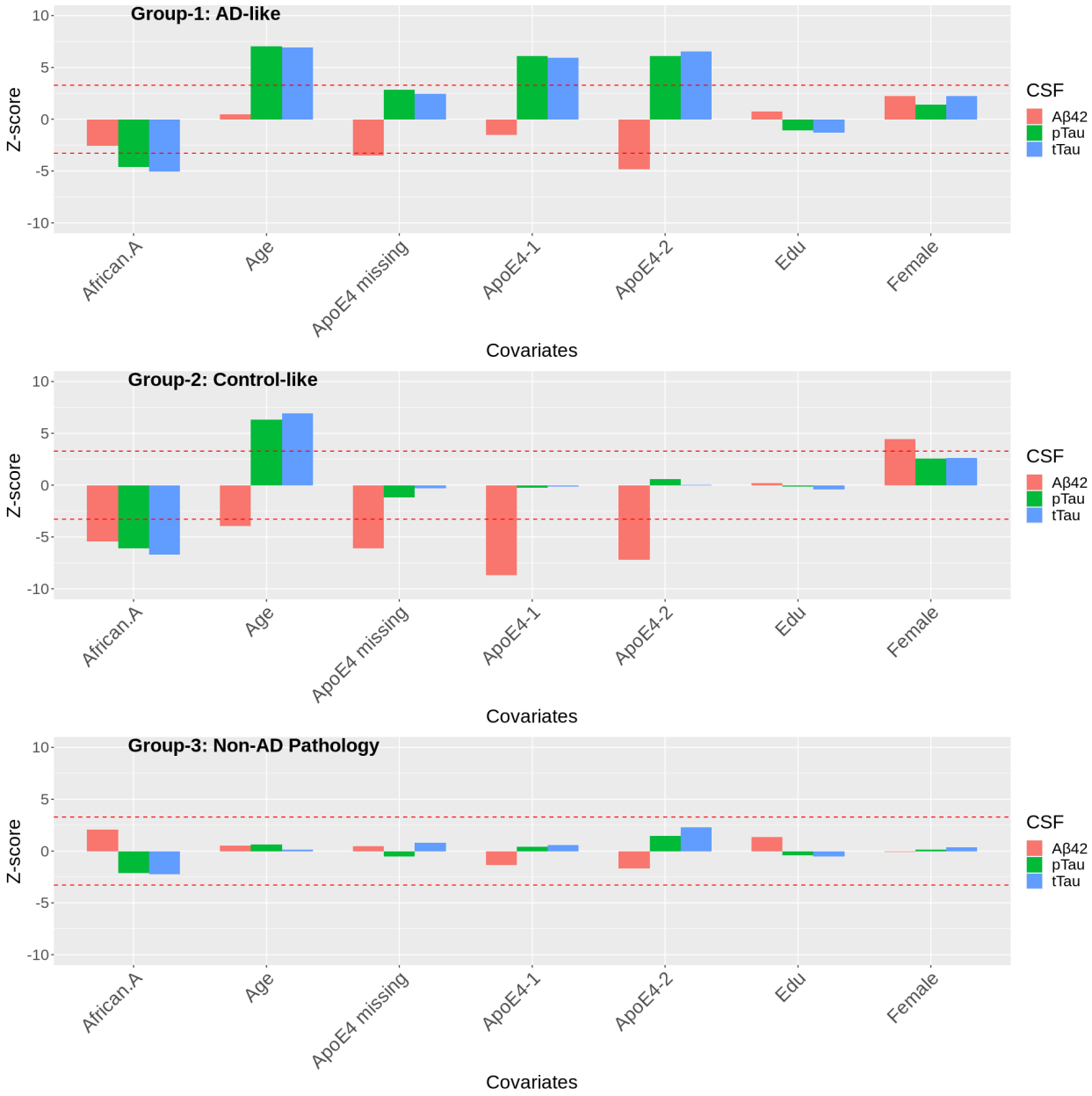


图 4：所选 $G = 3$ 模型中预测变量簇内效应估计值的 Z 分数柱状图。低 Abeta42、高 Tau 蛋白组与阿尔茨海默病诊断的重叠度最高。高 Abeta42、低 Tau 蛋白组与认知正常的重叠度最高。高 Abeta42、高 Tau 蛋白组呈现混合病理，其中包括一些轻度认知障碍和其他诊断。水平虚线表示针对 63 次比较进行邦费罗尼校正时的临界 Z 分数。

附录A. 拟议模型

A.1. 多元删失回归的*EM*算法

基于主文稿第2.1节中的模型定义，假设观测到$\mathbf{y}_i^*$, $i = 1, \ldots, N$，则完整数据似然为

$$\mathcal{L}_c(\mathbf{Y}^*; \mathbf{C}, \mathbf{X}, \boldsymbol{\psi}) = \prod_{i=1}^{N} \frac{1}{\sqrt{2^p \pi^p |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{y}_i^* - \boldsymbol{\beta}^\top \boldsymbol{x}_i)^\top \boldsymbol{\Sigma}^{-1} (\mathbf{y}_i^* - \boldsymbol{\beta}^\top \boldsymbol{x}_i)}.$$

令 $\langle \cdot \rangle$ 表示在给定观测数据$\mathbf{y} \in \mathbb{R}^{n \times p}$条件下的期望，则完整数据对数似然函数的条件期望形式如下：

$$\begin{aligned}\mathcal{Q}_c(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}, \mathbf{y}) \propto \sum_{i=1}^{N} \{ &-\frac{1}{2} \ln |\boldsymbol{\Sigma}| - \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \langle \mathbf{y}_i^* \mathbf{y}_i^{*\top} \rangle \right) + \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}^\top \boldsymbol{x}_i \langle \mathbf{y}_i^* \rangle^\top \right) \\ &- \frac{1}{2} \mathrm{tr} \left(\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} \right) \}.\end{aligned}$$

EM算法的步骤通过最大化$\mathcal{Q}_c(\boldsymbol{\psi};\ \boldsymbol{\psi}^{(k)};\mathbf{y})$的条件期望推导得出：

- **E步**，根据旧的参数值更新条件期望：

$$\langle \mathbf{y}_i^* \rangle \equiv \mathrm{E}_{\boldsymbol{\psi}^{(k)}}(\mathbf{y}_i^* | \mathbf{y}_i) = \begin{pmatrix} \mathbf{y}_{io_i} \\ \langle \mathbf{y}_{ic_i}^* | \mathbf{y}_{io_i} \rangle \end{pmatrix} \tag{A1}$$

$$\langle \mathbf{y}_i^* \mathbf{y}_i^{*\top} \rangle \equiv \mathrm{E}_{\boldsymbol{\psi}^{(k)}}(\mathbf{y}_i^* \mathbf{y}_i^{*\top} | \mathbf{y}_i) = \begin{pmatrix} \mathbf{y}_{io_i} \mathbf{y}_{io_i}^\top & \mathbf{y}_{io_i} \langle \mathbf{y}_{ic_i}^* | \mathbf{y}_{io_i} \rangle^\top \\ \langle \mathbf{y}_{ic_i}^* | \mathbf{y}_{io_i} \rangle \mathbf{y}_{io_i}^\top & \langle \mathbf{y}_{ic_i}^* \mathbf{y}_{ic_i}^{*\top} | \mathbf{y}_{io_i} \rangle \end{pmatrix} \tag{A2}$$

其中

$$\langle \mathbf{y}_{ic_i}^* \mathbf{y}_{ic_i}^{*\top} | \mathbf{y}_{io_i} \rangle = \boldsymbol{\mu}_{c_i|o_i}^2 (\boldsymbol{\beta}^\top \boldsymbol{x}_i, \boldsymbol{\Sigma}; \mathcal{D}(\mathbf{c}_i))$$

$$\langle \mathbf{y}^*_{ic_i} | \mathbf{y}_{io_i} \rangle = \boldsymbol{\mu}^1_{c_i|o_i}(\boldsymbol{\beta}^\top \boldsymbol{x}_i, \boldsymbol{\Sigma}; \mathcal{D}(\mathbf{c}_i))$$

是受截断区域 $\mathcal{D}(\mathbf{c}_i)$ 约束的多元截断条件高斯分布的一阶矩和二阶矩，可使用准蒙特卡罗积分算法进行数值计算（Genz和Bretz，2002年；Genz，2004年）。

关于感兴趣参数的条件完全数据得分函数：

$$\begin{aligned}
\mathbf{S}_c(\boldsymbol{\beta}; \boldsymbol{\psi}^{(k)}, \mathbf{y}) &\equiv \frac{\partial \mathcal{Q}_c(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}, \mathbf{y})}{\partial \boldsymbol{\beta}} \\
&= \sum_{i=1}^{N} \boldsymbol{x}_i \langle \mathbf{y}^*_i \rangle^\top \boldsymbol{\Sigma}^{-1} - \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1} \\
\mathbf{S}_c(\boldsymbol{\Sigma}; \boldsymbol{\psi}^{(k)}, \mathbf{y}) &\equiv \frac{\partial \mathcal{Q}_c(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}, \mathbf{y})}{\partial \boldsymbol{\Sigma}} \\
&= \sum_{i=1}^{N} -\frac{1}{2}\boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1} \langle \mathbf{y}^*_i \mathbf{y}^{*\top}_i \rangle \boldsymbol{\Sigma}^{-1} - \boldsymbol{\Sigma}^{-1} \langle \mathbf{y}^*_i \rangle \boldsymbol{x}_i^\top \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1} + \frac{1}{2}\boldsymbol{\Sigma}^{-1} \boldsymbol{\beta}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta} \boldsymbol{\Sigma}^{-1}
\end{aligned}$$

- **M 步**，最大化 $Q_c(\boldsymbol{\psi}; \boldsymbol{\psi}^{(k)}, \mathbf{y})$：
  - 在 $\mathbf{0}$ 处求解条件完全数据得分 $\mathbf{S}_c(\boldsymbol{\beta}; \boldsymbol{\psi}^{(k)}, \mathbf{y})$：

$$\begin{aligned}
\widetilde{\boldsymbol{\beta}} &= (\sum_{i=1}^{N} \boldsymbol{x}_i \boldsymbol{x}_i^\top)^{-1} (\sum_{i=1}^{N} \boldsymbol{x}_i \langle \mathbf{y}^*_i \rangle^\top) \\
&= (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \langle \mathbf{Y}^* \rangle
\end{aligned} \tag{A3}$$

  - 在 $\mathbf{0}$ 处求解条件得分 $\mathbf{S}_c(\boldsymbol{\Sigma}; \boldsymbol{\psi}^{(k)}, \mathbf{y})$，并代入 $\widetilde{\boldsymbol{\beta}}$：

$$\begin{aligned}
\widetilde{\boldsymbol{\Sigma}} &= \frac{\sum_{i=1}^{N} \langle \mathbf{y}^*_i \mathbf{y}^{*\top}_i \rangle - \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{x}_i \langle \mathbf{y}^*_i \rangle^\top - \langle \mathbf{y}^*_i \rangle \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}} + \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \widetilde{\boldsymbol{\beta}}}{N} \\
&= \frac{\sum_{i=1}^{N} \langle (\mathbf{y}^*_i - \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{x}_i)(\mathbf{y}^*_i - \widetilde{\boldsymbol{\beta}}^\top \boldsymbol{x}_i)^\top \rangle}{N} \\
&= \frac{\langle (\mathbf{Y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}})^\top (\mathbf{Y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}) \rangle}{N}
\end{aligned} \tag{A4}$$

其中，$\langle \mathbf{Y}^* \rangle_g$ 是由条件期望 $\langle \mathbf{y}^*_i \rangle_g$ 堆叠而成的 $N \times p$ 矩阵。EM算法依次在 **E** 步和 **M** 步之间迭代，直至收敛。

A.2. 删失多元高斯混合回归（托比特回归）的***EM***算法

同样地，根据第2.2节中的模型定义，假设潜在成分标签 $\mathbf{z}^{ig}$ 和潜在数据 $\mathbf{y}_i^*$ 均为观测值时的完整数据似然：

$$\mathcal{L}_c(\mathbf{Y}^*, \mathbf{Z}; \mathbf{C}, \mathbf{X}, \boldsymbol{\Psi}) = \prod_{i=1}^{N}\prod_{g=1}^{G}\left\{\frac{\omega_g}{\sqrt{2^p\pi^p|\boldsymbol{\Sigma}_g|}}e^{-\frac{1}{2}(\mathbf{y}_i^*-\boldsymbol{\beta}_g^\top\boldsymbol{x}_i)^\top\boldsymbol{\Sigma}_g^{-1}(\mathbf{y}_i^*-\boldsymbol{\beta}_g^\top\boldsymbol{x}_i)}\right\}^{z_{ig}}. \tag{A5}$$

再次令 $\langle\cdot\rangle$ 表示以观测数据$\mathbf{y} \in \mathbb{R}^{N\times p}$为条件的期望，则完整数据对数似然函数的条件期望形式如下：

$$\begin{aligned}\mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}, \mathbf{y}) \propto \sum_{i=1}^{N}\sum_{g=1}^{G}\langle z_{ig}\rangle\{\ln\omega_g - \frac{1}{2}\ln|\boldsymbol{\Sigma}_g| - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}\langle\mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g\right)\\ + \mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\beta}_g^\top\boldsymbol{x}_i\langle\mathbf{y}_i^*\rangle_g^\top\right) - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Sigma}_g^{-1}\boldsymbol{\beta}_g^\top\boldsymbol{x}_i\boldsymbol{x}_i^\top\boldsymbol{\beta}_g\right)\}.\end{aligned} \tag{A6}$$

EM算法的步骤通过最大化$\mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}, \mathbf{y})$的条件期望推导得出：

- **E步**，根据旧参数值更新条件期望：

$$\langle z_{ig}\rangle \equiv \mathrm{E}_{\boldsymbol{\Psi}^{(k)}}(Z_{ig}|\mathbf{y}_i) = P_{\boldsymbol{\Psi}^{(k)}}(Z_{ig}=1|\mathbf{y}_i) = \frac{\omega_g^{(k)}f_g(\mathbf{y}_i;\boldsymbol{x}_i,\boldsymbol{\psi}_g^{(k)})}{\sum_{h=1}^{G}\omega_h^{(k)}f_h(\mathbf{y}_i;\boldsymbol{x}_i,\boldsymbol{\psi}_h^{(k)})} \tag{A7}$$

$$\langle\mathbf{y}_i^*\rangle_g \equiv \mathrm{E}_{\boldsymbol{\psi}_g^{(k)}}(\mathbf{y}_i^*|\mathbf{y}_i) = \begin{pmatrix}\mathbf{y}_{io_i}\\ \langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g\end{pmatrix} \tag{A8}$$

$$\langle\mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g \equiv \mathrm{E}_{\boldsymbol{\psi}_g^{(k)}}(\mathbf{y}_i^*\mathbf{y}_i^{*\top}|\mathbf{y}_i) = \begin{pmatrix}\mathbf{y}_{io_i}\mathbf{y}_{io_i}^\top & \mathbf{y}_{io_i}\langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g^\top\\ \langle\mathbf{y}_{ic_i}^*|\mathbf{y}_{io_i}\rangle_g\mathbf{y}_{io_i}^\top & \langle\mathbf{y}_{ic_i}^*\mathbf{y}_{ic_i}^{*\top}|\mathbf{y}_{io_i}\rangle_g\end{pmatrix} \tag{A9}$$

其中

$$\langle \mathbf{y}^*_{ic_i} {\mathbf{y}^*_{ic_i}}^\top | \mathbf{y}_{io_i} \rangle_g = \boldsymbol{\mu}^2_{c_i|o_i}(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i, \boldsymbol{\Sigma}_g; \mathcal{D}(\mathbf{c}_i))$$

$$\langle \mathbf{y}^*_{ic_i} | \mathbf{y}_{io_i} \rangle_g = \boldsymbol{\mu}^1_{c_i|o_i}(\boldsymbol{\beta}_g^\top \boldsymbol{x}_i, \boldsymbol{\Sigma}_g; \mathcal{D}(\mathbf{c}_i))$$

是截断区域 $\mathcal{D}(\boldsymbol{c}_i)$ 下第 g 个混合成分的截断条件高斯分布的一阶矩和二阶矩，可使用由（Genz 和 Bretz，2002年；Genz，2004年）开发的准蒙特卡罗积分算法进行数值计算。

关于感兴趣参数的条件完全数据得分函数为

$$\begin{aligned}
\mathbf{S}_c(\boldsymbol{\beta}_g; \boldsymbol{\Psi}^{(k)}, \mathbf{y}) &\equiv \frac{\partial \mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}, \mathbf{y})}{\partial \boldsymbol{\beta}_g} \\
&= \sum_{i=1}^{N} \langle \boldsymbol{z}_{ig} \rangle \boldsymbol{x}_i \langle \mathbf{y}_i^* \rangle^\top \boldsymbol{\Sigma}_g^{-1} - \langle \boldsymbol{z}_{ig} \rangle \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta}_g \boldsymbol{\Sigma}_g^{-1} \\
\mathbf{S}_c(\boldsymbol{\Sigma}_g; \boldsymbol{\Psi}^{(k)}, \mathbf{y}) &\equiv \frac{\partial \mathcal{Q}_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}, \mathbf{y})}{\partial \boldsymbol{\Sigma}_g} \\
&= \sum_{i=1}^{N} \langle \boldsymbol{z}_{ig} \rangle \left(-\frac{1}{2} \boldsymbol{\Sigma}_g^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_g^{-1} \langle \mathbf{y}_i^* \mathbf{y}_i^{*\top} \rangle \boldsymbol{\Sigma}_g^{-1} - \boldsymbol{\Sigma}_g^{-1} \langle \mathbf{y}_i^* \rangle \boldsymbol{x}_i^\top \boldsymbol{\beta}_g \boldsymbol{\Sigma}_g^{-1} + \frac{1}{2} \boldsymbol{\Sigma}_g^{-1} \boldsymbol{\beta}_g^\top \boldsymbol{x}_i \boldsymbol{x}_i^\top \boldsymbol{\beta}_g \boldsymbol{\Sigma}_g^{-1} \right)
\end{aligned}$$

- **M 步**，最大化 $Q_c(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(k)}, \mathbf{y})$：
  - 通过求解拉格朗日乘子来更新 ω_g：

$$\widetilde{\omega}_g = \sum_{i=1}^{N} \langle z_{ig} \rangle / N \tag{A10}$$

◦ 在 $\mathbf{0}$ 处求解条件完全数据得分 $\mathbf{S}_c(\boldsymbol{\beta}_g; \boldsymbol{\Psi}_{(k)}, \mathbf{y})$:

$$\begin{aligned}\widetilde{\boldsymbol{\beta}}_g &= (\sum_{i=1}^{N}\langle z_{ig}\rangle \boldsymbol{x}_i\boldsymbol{x}_i^\top)^{-1}(\sum_{i=1}^{N}\langle z_{ig}\rangle \boldsymbol{x}_i\langle \mathbf{y}_i^*\rangle_g^\top) \\ &= (\mathbf{X}^\top \boldsymbol{\mathcal{Z}}_g\mathbf{X})^{-1}\mathbf{X}^\top\boldsymbol{\mathcal{Z}}_g\langle \mathbf{Y}^*\rangle_g\end{aligned} \tag{A11}$$

◦ 在$\mathbf{0}$处求解条件得分$\mathbf{S}_c(\boldsymbol{\Sigma}_g;\ \boldsymbol{\Psi}_{(k)},\ \mathbf{y})$，并代入 $\widetilde{\boldsymbol{\beta}}_g$:

$$\begin{aligned}\widetilde{\boldsymbol{\Sigma}}_g &= \frac{\sum_{i=1}^{N}\langle z_{ig}\rangle\left(\langle \mathbf{y}_i^*\mathbf{y}_i^{*\top}\rangle_g - \widetilde{\boldsymbol{\beta}}_g^\top\mathbf{x}_i\langle \mathbf{y}_i^*\rangle_g^\top - \langle \mathbf{y}_i^*\rangle_g\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}}_g + \widetilde{\boldsymbol{\beta}}_g^\top\mathbf{x}_i\mathbf{x}_i^\top\widetilde{\boldsymbol{\beta}}_g\right)}{\sum_{i=1}^{N}\langle z_{ig}\rangle} \\ &= \langle(\mathbf{Y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_g)^\top\boldsymbol{\mathcal{Z}}_g(\mathbf{Y}^* - \mathbf{X}\widetilde{\boldsymbol{\beta}}_g)\rangle_g/(\widetilde{\omega}_g N)\end{aligned} \tag{A12}$$

EM算法依次迭代执行$\mathbf{E}$ 步和$\mathbf{M}$ 步，直至收敛。

附录 B. 表格

500 次重复试验， $N = 1,\ 000,\ \mathbf{Y} = (\mathbf{Y}_1, \mathbf{Y}_2)$	
参数	真实值
$\boldsymbol{\beta} = (\boldsymbol{\beta}_1,\ \boldsymbol{\beta}_2,\ \boldsymbol{\beta}_3)$	$\begin{pmatrix} 2 & 20 \\ 0 & -2 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 3 & 25 \\ 1 & -3 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}, \begin{pmatrix} 3.5 & 30 \\ 2 & -5 \\ 0 & 0 \\ 0 & 0 \end{pmatrix}$
类型 1 错误率 of 情景 I:	$\begin{pmatrix} - & - \\ 0.038 & - \\ 0.024 & 0.044 \\ 0.028 & 0.040 \end{pmatrix}, \begin{pmatrix} - & - \\ - & - \\ 0.040 & 0.048 \\ 0.042 & 0.056 \end{pmatrix}, \begin{pmatrix} - & - \\ - & - \\ 0.048 & 0.046 \\ 0.046 & 0.036 \end{pmatrix}$
第一类错误率 of 情景 II:	$\begin{pmatrix} - & - \\ 0.034 & - \\ 0.022 & 0.042 \\ 0.024 & 0.050 \end{pmatrix}, \begin{pmatrix} - & - \\ - & - \\ 0.038 & 0.042 \\ 0.036 & 0.046 \end{pmatrix}, \begin{pmatrix} - & - \\ - & - \\ 0.038 & 0.038 \\ 0.048 & 0.052 \end{pmatrix}$

表 B1：情景 I（轻度删失）和情景 II（重度删失）中簇内效应的第一类错误率。

附录 C. 图

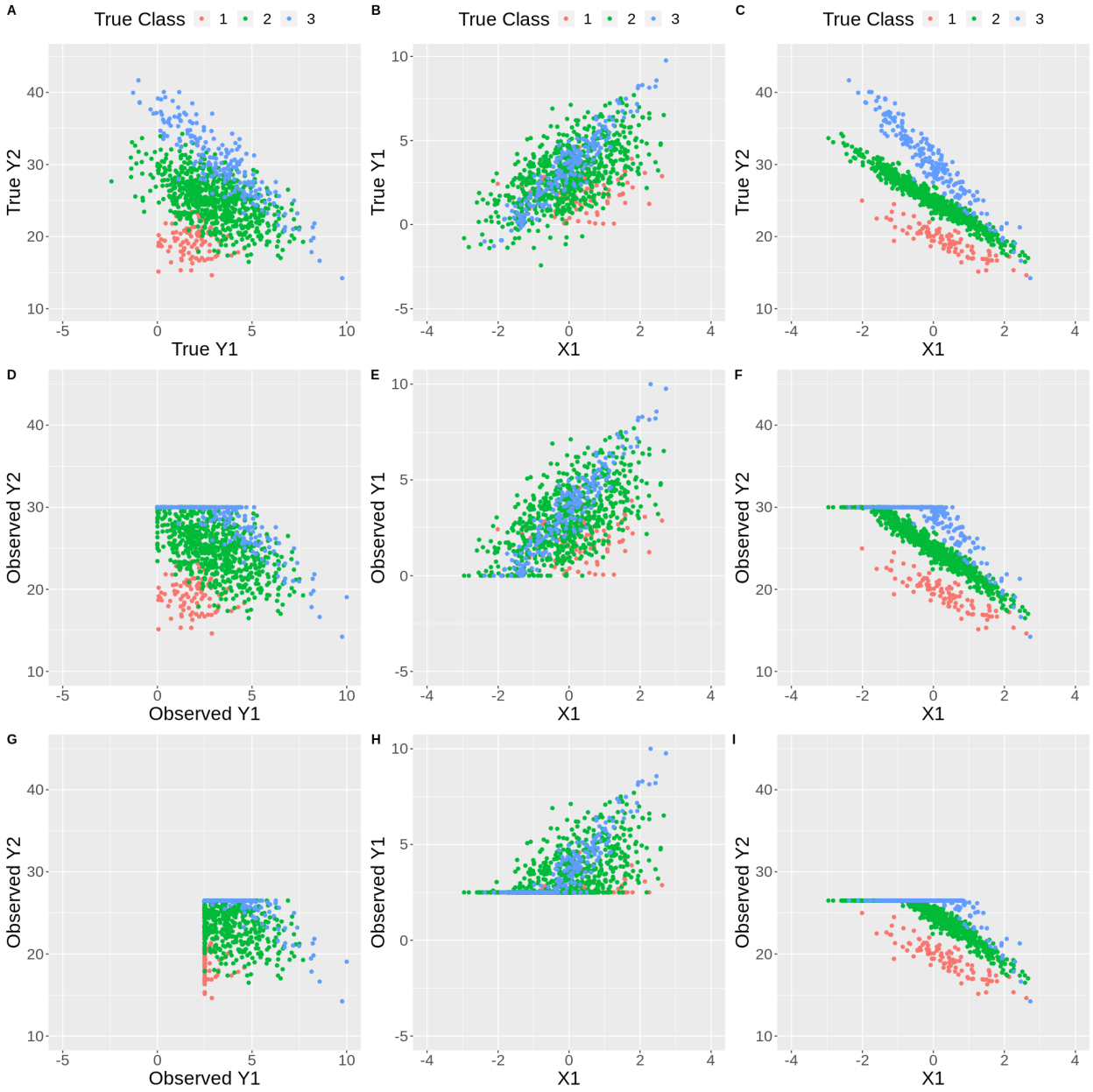


图 C1.：模拟数据集示例及删失机制。第 1 行展示删失前的数据。情景 I（第 2 行）中，$\mathbf{Y}_1$ 在 0 处左删失，$\mathbf{Y}_2$ 在 30 处右删失。情景 II（第 3 行）中，$\mathbf{Y}_1$ 在 2.5 处左删失，$\mathbf{Y}_2$ 在 30 处右删失。

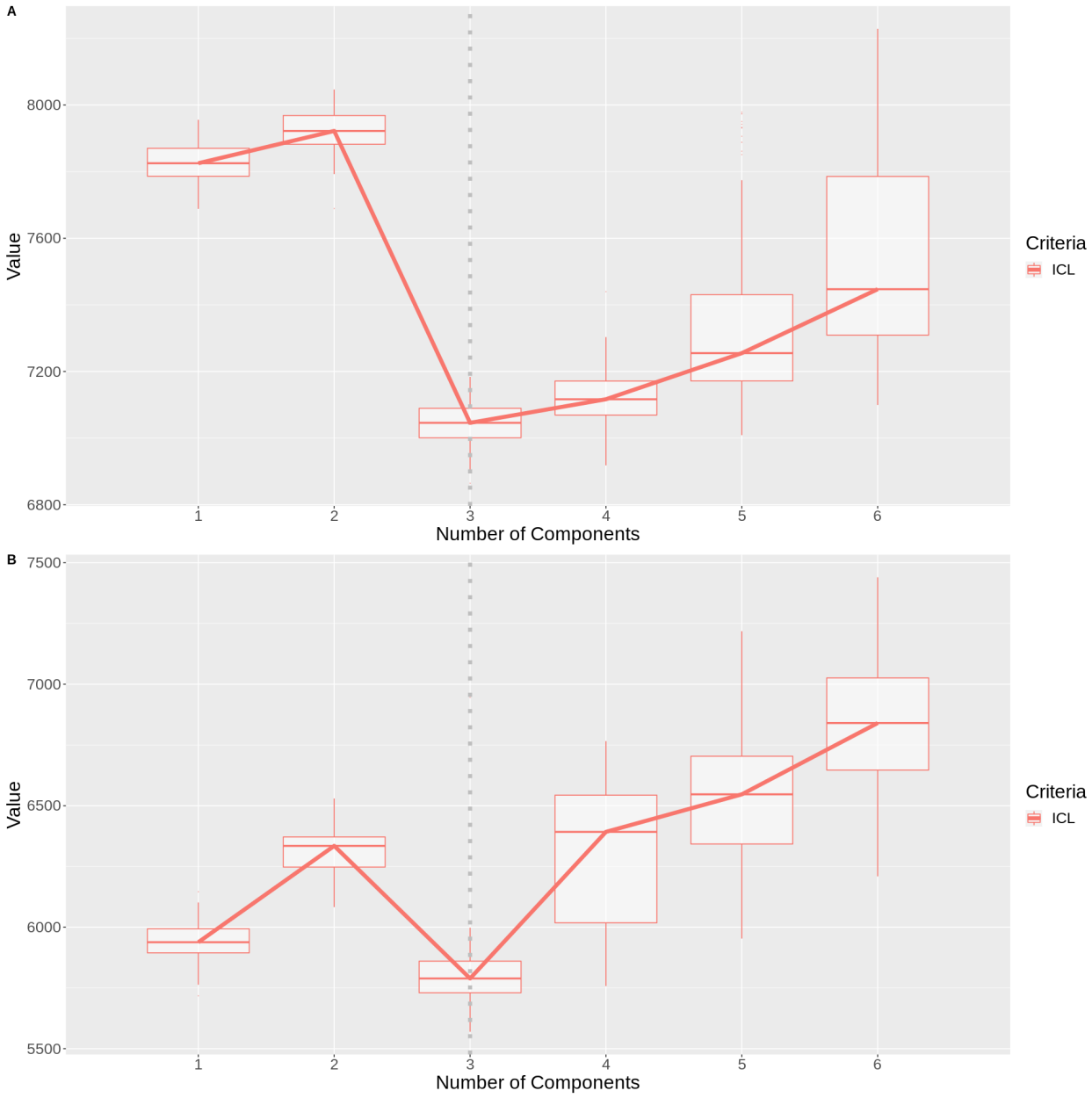


图 C2.：情景 I 和情景 II 下，基于 101 个数据重复样本选择最优的 G 。**A**：情景 I 下选择最优的 G ；**B**：情景 II 下选择最优的 G 。

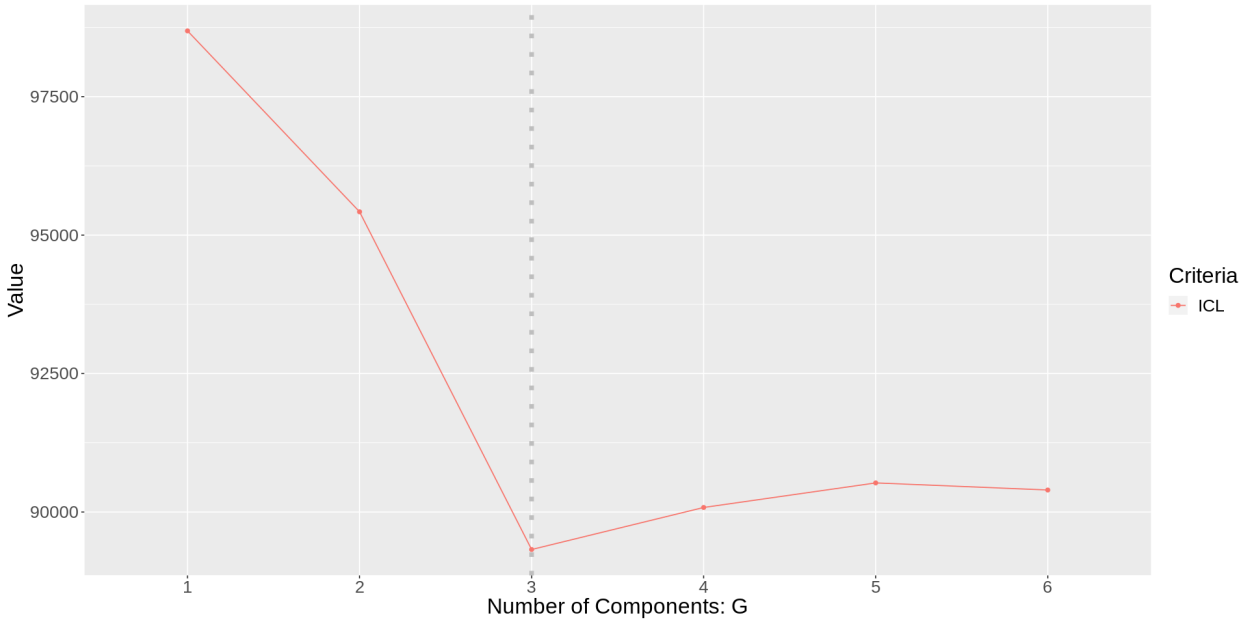


图 C3.：在埃默里认知神经病学生物标志物数据上选择最优的 G 。ICL 表示 3 组。

参考文献

Amemiya, T. (1973年). 因变量服从截断正态分布时的回归分析。《计量经济学》, 41(6):997. Benaglia, T., Chauveau, D., Hunter, D. R., and Young, D. S. (2010). mixtools：用于分析混合模型的 R 软件包。*Journal of Statistical Software*, 32:1–29. Biernacki, C., Celeux, G., and Govaert, G. (2000年). 使用积分完备似然评估用于聚类的混合模型。*IEEE Transactions on Pattern Analysis and Machine In-telligence*, 22(7):719–725. Blennow, K., Dubois, B., Fagan, A. M., Lewczuk, P., De Leon, M. J., and Hampel, H. (2015年). 脑脊液生物标志物在早期阿尔茨海默病诊断中的临床应用价值。《阿尔茨海默病与痴呆》, 11(1):58–69. Caudill, S. B. (2012年). 基于正态分布混合的删失回归模型部分自适应估计量。《统计方法与应用》, 21:121–137. Coates, A. and Ng, A. Y. (2012年). 使用 K 均值学习特征表示。载于 Montavon G., Orr G.B., and M¨uller KR. 编，《神经网络：实践技巧》《神经网络：实践技巧》，第22章，第561–580页。Springer，柏林，海德堡，第2版。 Collins, J. and Huynh, M. (2014年). 无需完全验证的诊断试验准确性估计：潜在类别方法综述。《医学统计学》, 33(24):4141. De Alencar, F., Galarza, C., Matos, L., and Lachos, V. (2020). Censmfm：多元删失/缺失数据的有限混合模型。*R*软件包版本, 2. Dubois, B., Feldman, H. H., Jacova, C., DeKosky, S. T., Barberger-Gateau, P., Cummings, J., et al. (2007). 阿尔茨海默病诊断研究标准：修订NINCDS–ADRDA标准。《柳叶刀·神经病学》, 6(8):734–746. Fair, R. C. (1977). 关于Tobit估计量计算的注记。《计量经济学》, 45(7):1723. Galarza, C. E., Kan, R., and Lachos, V. H. (2021年). MomTrunc：折叠及双重截断多元分布的矩。*R*软件包版本 *5.97*. Garay, A., Lachos, V., and Massuia, M. (2013). Smncensreg：在正态分布尺度混合模型下拟合单变量删失回归模型。*R*软件包版本, 2. Garay, A. M., Lachos, V. H., Bolfarine, H., and Cabral, C. R. (2017年). 线性删失回归

模型与正态分布尺度混合模型。*Stat Papers*, 58:247–278. Garrett, S. L., McDaniel, D., Obideen, M., Trammell, A. R., Shaw, L. M., Goldstein, F. C., and Ha- jjar, I. (2019). 脑脊液淀粉样蛋白和Tau生物标志物的种族差异及其与轻度认知障碍相关的临界值。*JAMA Network Open*, 2(12):e1917363–e1917363. Genz, A. (2004年). 矩形双变量和三变量正态分布及t概率的数值计算。*Statistics and Computing 2004 14:3*, 14(3):251–260. Genz, A. and Bretz, F. (2002年). 多元t概率计算方法的比较。计算与图形统计学杂志, 11(4):950–971. Goetz, M. E., Hanfelt, J. J., John, S. E., Bergquist, S. H., Loring, D. W., Quyyumi, A., Clifford, G. D., Vaccarino, V., Goldstein, F., Johnson, T. M., Kuerston, R., Marcus, M., Levey, A. I., and Lah, J. J. (2019). 埃默里健康老龄化研究与埃默里健康大脑研究的理论依据和设计。神经流行病学, 53(3-4):187–200. Goldfeld, S. M. and Quandt, R. E. (1973年). 切换回归的马尔可夫模型。计量经济学杂志, 1(1):3–15. Grun, B. and Leisch, F. (2008年). Flexmix 2版：包含伴随变量及可变和固定参数的有限混合模型。Hampel, H., B¨urger, K., Teipel, S. J., Bokde, A. L., Zetterberg, H., and Blennow, K. (2008年). 阿尔茨海默病的核心候选神经化学和影像学生物标志物。《阿尔茨海默病与痴呆》, 4(1):38–48. Hanson, T. and Johnson, W. O. (2002年). 使用波利亚树混合模型对回归误差建模。*Journal of the American Statistical Association*, 97(460):1020–1033. Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). 自适应局部专家混合模型。神经计算, 3(1):79–87. Jansen, W. J., Janssen, O., Tijms, B. M., and Others (2022年). 阿尔茨海默病临床谱系中淀粉样蛋白异常的患病率估计。*JAMA Neurology*, 79(3):228–243. Jansen, W. J., Ossenkoppele, R., Knol, D. L., Tijms, B. M., and Others (2015年). 无痴呆人群脑淀粉样蛋白病理的患病率：荟萃分析。*JAMA*, 313(19):1924–1938. Jedidi, K., Ramaswamy, V., and Desarbo, W. S. (1993年). 一种用于潜在

涉及删失因变量的分类回归。*Psychometrika*, 58(3):375–394. Karlsson, M. and Laitila, T. (2014). 删失回归模型的有限混合建模。*StatPapers*, 55:627–642. Khalili, A. and Lin, S. (2013). 发散参数数量下回归模型有限混合的正则化。*Biometrics*, 69(2):436–446. Lee, G. and Scott, C. (2012). 处理截断和删失数据的多元高斯混合模型EM算法。*Computational Statistics & Data Analysis*, 56(9):2816–2829. Linzer, D. A. and Lewis, J. B. (2011). poLCA：用于多分类变量潜类别分析的R软件包。*Journal of Statistical Software*, 42(10):1–29. McLachlan, G. J. and Peel, D. (2000). 有限混合模型. Wiley. Meyer, G. D., Shapiro, F., Vanderstichele, H., Vanmechelen, E., Engelborghs, S., Deyn, P. P. D., Coart, E., Hansson, O., Minthon, L., Zetterberg, H., Blennow, K., Shaw, L., Trojanowski, J. Q., and Initiative, A. D. N. (2010). 认知正常老年人中与诊断无关的阿尔茨海默病生物标志物特征。*Archives of Neurology*, 67(8):949–956. Morris, J. C., Roe, C. M., Xiong, C., Fagan, A. M., Goate, A. M., Holtzman, D. M., and Mintun, M. A. (2010). 在认知正常老龄化过程中，载脂蛋白E可预测β淀粉样蛋白，但无法预测tau蛋白阿尔茨海默病病理。*Annals of Neurology*, 67(1):122–131. Morris, J. C., Schindler, S. E., McCue, L. M., Moulder, K. L., Benzinger, T. L., Cruchaga, C., Fagan, A. M., Grant, E., Gordon, B. A., Holtzman, D. M., and Xiong, C. (2019). 阿尔茨海默病生物标志物中的种族差异评估。*JAMA Neurology*, 76(3):264–273. O'Hagan, A., Murphy, T. B., Scrucca, L., and Gormley, I. C. (2019). 使用高斯混合模型，通过刀切法、自助法和加权似然自助法研究聚类中的参数不确定性。*Computational Statistics 2019 34:4*, 34(4):1779–1813. Quandt, R. E. and Ramsey, J. B. (1978). 正态分布混合与切换回归的估计。*Journal of the American Statistical Association*, 73(364):730–738. Ruud, P. A. (1991). 使用EM算法的估计方法扩展。*Journal of Econo-metrics*, 49(3):305–341. Sanchez, L. B., Lachos, V. H., Moreno, E. J. L., Sanchez, M. L. B., and LazyData, T. (2015). 软件包“censmixreg”。

Scrucca, L., Fop, M., Murphy, T. B., and Raftery, A. E. (2016). mclust 5：使用高斯有限混合模型进行聚类、分类和密度估计。*The R Journal*, 8(1):289–317. Shin, J. and Doraiswamy, P. M. (2016). 阿尔茨海默病试验中非裔美国人代表不足：呼吁采取平权行动。*Frontiers in aging neuroscience*, 8:123. Vermunt, J. K. and Magidson, J. (2021). LG-syntax用户指南：Latent GOLD Syntax 模块6.0版手册。马萨诸塞州阿灵顿：*Statistical Innovations Inc*. Wang, W. L., Castro, L. M., Hsieh, W. C., and Lin, T. I. (2021). 用于多重删失因变量建模的含协变量因子分析器混合模型。*Statistical Papers*, 62(5):2119– 2145. Wang, W. L., Castro, L. M., Lachos, V. H., and Lin, T. I. (2019). 通过因子分析器混合模型对删失数据进行基于模型的聚类。*Computational Statistics & Data Analysis*, 140:104–121. Xie, B., Pan, W., and Shen, X. (2010). 惩罚因子分析器混合模型及其在高维微阵列数据聚类中的应用。*Bioinformatics*, 26(4):501–508. Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). 专家混合模型二十年。*IEEETransactions on Neural Networks and Learning Systems*, 23(8):1177–1193. Zeller, C. B., Cabral, C. R. B., Lachos, V. H., and Benites, L. (2019). 基于正态分布尺度混合模型的删失数据回归模型有限混合。*Advances in Data Analysis and Classification*, 13(1):89–116.