具有序列相关结构的混合增长曲线模型

献给方开泰教授 80 华诞

潘雅婷¹, 费宇¹, 倪明明¹, 潘建新^{2*}

1. 云南财经大学统计与数学学院, 昆明 650221;

2. Department of Mathematics, The University of Manchester, Manchester M13 9PL, UK

E-mail: Yating-Pan@hotmail.com, feiyukm@aliyun.com, mingming.ni90@gmail.com, Jianxin.Pan@manchester.ac.uk

收稿日期: 2019-05-23; 接受日期: 2019-09-06; 网络出版日期: 2020-05-08; * 通信作者

国家自然科学基金 (批准号: 11871357, 11561071 和 11971421) 和国家重点研发计划 (批准号: 2018YFC0831900) 资助项目

摘要 增长曲线模型是纵向数据分析的重要方法, 它的一个重要假设是组别阵已知, 但这在现实中往往难以满足. 因此, 本文提出全新的混合增长曲线模型处理该问题, 随之将数据分布假设从正态扩展到混合正态. 此外, 新模型中加入了序列相关结构, 为常伴有序列相关的纵向数据提供了合理的分析方法. 本文以数据驱动的方法来搜索最佳模型, 在获得模型参数估计的同时实现了序列相关数据的聚类与拟合. 模拟研究和实际数据分析都说明了所提出的方法具有合理性和有效性.

关键词 增长曲线模型 Gauss 混合模型 序列相关结构 最大期望算法 聚类

MSC (2010) 主题分类 62F99, 62H30, 62J99

1 引言

增长曲线模型 (growth curve model, GCM)^[1] 是纵向数据分析的一种重要方法, 它在经济学、生物学和医学等领域有着广泛的应用. 该模型在一般的多元线性模型基础上引入组别设计阵, 实现了对观测个体纵向趋势的组别控制, 其一般定义如下 (参见文献 [2]):

$$\begin{cases} \mathbf{Y}_{p \times n} = \mathbf{X}_{p \times m} \mathbf{B}_{m \times r} \mathbf{Z}_{r \times n} + \boldsymbol{\varepsilon}_{p \times n}, \\ \boldsymbol{\varepsilon}_{p \times n} \sim N_{p,n}(\mathbf{0}, \boldsymbol{\Sigma}, \mathbf{I}_n), \end{cases} \quad (1.1)$$

其中 $\mathbf{X}_{p \times m}$ 和 $\mathbf{Z}_{r \times n}$ 都是已知设计矩阵, 它们的秩分别为 m 和 r ($m < p$ 且 $r < n$), 回归系数矩阵 $\mathbf{B}_{m \times r}$ 未知, 随机误差矩阵 $\boldsymbol{\varepsilon}_{p \times n}$ 的各列是相互独立的 p 元正态向量, 均值向量为 $\mathbf{0}$, 未知的 p 维协方差矩阵满足 $\boldsymbol{\Sigma} > \mathbf{0}$. 因此, 响应矩阵满足 $\mathbf{Y}_{p \times n} \sim N_{p,n}(\mathbf{X} \mathbf{B} \mathbf{Z}, \boldsymbol{\Sigma}, \mathbf{I}_n)$, 即 $\mathbf{Y}_{p \times n}$ 服从矩阵正态分布, 均值为 $\mathbf{X} \mathbf{B} \mathbf{Z}$, 个体间的协方差为 n 维单位阵 $\mathbf{I}_n$, 即个体相互独立, 而每个个体内 p 次观测的相关关系由正定协方差矩阵 $\boldsymbol{\Sigma}$ 描述.

英文引用格式: Pan Y T, Fei Y, Ni M M, et al. Growth curves mixture model with serial covariance structure (in Chinese). Sci Sin Math, 2020, 50: 645–666, doi: 10.1360/N012019-00145

模型 (1.1) 具有良好的统计性质, 它的参数估计 (广义最小二乘估计和极大似然估计) 都有解析表达式, 这非常有利于统计推断与统计诊断的研究, 并使得它在实际应用中简单、易操作. 但是, 模型 (1.1) 也存在着局限性: 第一, 模型假设组别阵 $\mathbf{Z}_{r \times n}$ 已知, 然而现实中它常常是未知的, 表现为一些观测个体接受过什么处理或具有哪些特征并不清楚, 甚至连组别数 r 都无法确定, 故这样明确的设定限制了模型在这类问题中的应用. 第二, 模型假设协方差阵 Σ 没有特定结构, 但是它可能具有某种特殊结构, 在具体分析中体现为重复观测间明显地存在着某种相关关系, 故这样粗略的设定使模型忽视了纵向数据中重要的特征信息. 第三, 模型假设响应阵 $\mathbf{Y}$ 服从正态分布, 可是真实数据往往难以严格地服从正态分布, 故这样严苛的设定使模型在应用中无法保证估计的有效性.

目前已经有研究讨论了这些局限性: 针对特定协方差结构, 文献 [3] 通过 Rao 的简单协方差结构下 GCM 的极大似然估计和剖面似然函数求解, 分别在均匀相关结构 (uniform covariance structure, UCS) 和序列相关结构 (serial covariance structure, SCS) 下得到了模型 (1.1) 的有效参数估计; 针对组别阵未知的情形, 文献 [4] 借用 Gauss 混合模型中最大期望 (expectation maximization, EM) 算法的思想^[5] 成功地对协方差无特定结构 (unstructured covariance, UC) 且组别阵未知的模型 (1.1) 进行了参数估计. 然而, 在组别阵未知的前提下, 探讨 GCM 的协方差结构也十分重要, 文献 [6] 研究了该前提下具有 UCS 的模型 (1.1), 本文进一步地讨论该前提下具有 SCS 的模型 (1.1). 我们首次提出混合增长曲线模型, 把响应阵的分布假设从正态扩展为混合正态, 再引入 SCS 刻画数据特征, 为具有序列相关性且组别信息未知的纵向数据提供合理的分析方法. 在新模型中, 把随机组别阵看作缺失信息, 用 EM 算法^[7] 求解极大似然估计, 同时实现了序列相关数据的聚类与拟合. 此外, 采用数据驱动的方法来搜索最优模型的多项式阶数和组别数. 模拟研究说明忽视数据自有的相关性会导致分析结果出错, 而本方法能对序列相关数据进行有效的分析. 最后, 将该方法应用于识别酵母基因表达样式的实际问题, 获得了独特的分析结果.

本文第 2 节将定义混合增长曲线模型, 分析其特性, 并引入 SCS 来刻画自相关性. 第 3 节提出用 EM 算法计算新模型中参数的极大似然估计, 用最大后验概率预测随机组别阵 $\mathbf{Z}$, 再用 Akaike 信息准则 (Akaike information criterion, AIC) 或 Bayes 信息准则 (Bayesian information criterion, BIC) 确定未知回归系数阵 $\mathbf{B}_{m \times r}$ 的维数. 第 4 和 5 节分别通过 4 个模拟实验和 1 个实际数据分析展示了新模型的合理性和有效性, 并验证了正确设定协方差结构的重要性. 第 6 节总结全文的研究内容与主要结果, 并拓展讨论方向.

2 具有序列相关结构的混合增长曲线模型

本文在组别阵未知的情形下来探讨模型 (1.1), 提出的新模型称为混合增长曲线模型 (growth curves mixture model, GCMM), 其定义如下:

$$\begin{cases} \mathbf{Y}_{p \times n} = \mathbf{X}_{p \times m} \mathbf{B}_{m \times r} \mathbf{Z}_{r \times n} + \boldsymbol{\varepsilon}_{p \times n}, \\ \mathbf{z}_j \sim \text{Cat}\left(\pi_1, \dots, \pi_{r-1}, 1 - \sum_{k=1}^{r-1} \pi_k\right), \quad j = 1, \dots, n, \\ \boldsymbol{\varepsilon}_{p \times n} \sim N_{p,n}(\mathbf{0}, \Sigma, \mathbf{I}_n), \end{cases} \quad (2.1)$$

其中 $\mathbf{Z}_{r \times n}$ 是随机组别矩阵, 秩为 r ($r < n$), 它的列向量 $\mathbf{z}_j$ ($j = 1, \dots, n$) 服从参数为 $\boldsymbol{\pi} = (\pi_1, \dots, \pi_{r-1}, 1 - \sum_{k=1}^{r-1} \pi_k)$ 的分类分布; $\mathbf{X}_{p \times m}$ 是已知设计矩阵, 秩为 m ($m < p$); 回归系数矩阵 $\mathbf{B}_{m \times r}$ 未知; 随机

误差矩阵 $\epsilon_{p \times n}$ 的各列是相互独立的 p 元正态向量, 均值为 $\mathbf{0}$; 未知的 p 维协方差矩阵 $\Sigma > \mathbf{0}$. 个体间的协方差矩阵为 n 维单位方阵 $\mathbf{I}_n$, 即个体彼此独立; 个体内 p 次观测的相关关系由正定协方差阵 Σ 描述.

新模型 (2.1) 命名为混合增长曲线模型, 是因为模型 (1.1) 转化为模型 (2.1) 的过程中实现了纵向数据的分布混合. 模型 (1.1) 中的待估参数为 $\Theta = (\mathbf{B}, \Sigma)$, 当组别阵不可知时, 模型的未知元素增加了 $\mathbf{Z}$ 这个潜在的随机矩阵变量, 模型变得复杂了. 另外, 模型 (1.1) 中的每个个体必属于且仅属于 r 个组别中的一类, 表现为组别阵 $\mathbf{Z}$ 具有元素非 0 即 1 且元素按列加和为 1 的特性. 因此, 假设每个个体出现在第 k 个组别的概率为 π_k ($k = 1, \dots, r$, $0 \leq \pi_k \leq 1$ 且 $\sum_{k=1}^r \pi_k = 1$), 这时 $\mathbf{Z}_{r \times n}$ 中的 r 维列向量 $\mathbf{z}_j$ ($j = 1, \dots, n$) 服从参数为 $n = 1$ 和 $\boldsymbol{\pi} = (\pi_1, \dots, \pi_r)$ 的多项分布, 这个特殊的多项分布又称为分类分布 (categorical distribution) [8]. 在充分利用模型的先验信息后, $\mathbf{Z}$ 的每个列向量都服从分类分布, 即 $\mathbf{z}_j \sim \text{Cat}(\pi_1, \dots, \pi_{r-1}, 1 - \sum_{k=1}^{r-1} \pi_k)$, $j = 1, \dots, n$. 此时, $\mathbf{Z}$ 实际上是一个随机矩阵, 而待估参数转化为 $\Theta = (\boldsymbol{\pi}, \mathbf{B}, \Sigma)$. 与此同时, 响应阵 $\mathbf{Y}$ 服从的分布也发生了变化, 它的列向量 $\mathbf{y}_j$ ($j = 1, \dots, n$) 从只属于 r 个均值不同而协方差相同的 p 元正态分布中的一个, 转变为服从它们的有限混合分布. 该混合过程是利用随机组别阵的后验概率 $E(\mathbf{Z} | \mathbf{Y}, \Theta)$ 对这些正态分布加权来实现的, 故每个个体的真实分布都可以使用 r 个正态分布的混合去近似. 这样的转变很大程度上放宽了响应阵 $\mathbf{Y}$ 的严苛分布假设, 扩展了模型 (1.1) 的适用范围. 至此, 本文通过提取组别阵的先验信息将模型 (1.1) 转化成了一种特殊的 Gauss 混合模型 (Gaussian mixture model), 所以将模型 (2.1) 命名为混合增长曲线模型.

纵向数据是对相同个体进行重复观测, 而来自同一个体的数据难免具有相关关系, 所以, 相关性的讨论对纵向数据分析十分重要. 模型 (2.1) 中的协方差阵 Σ 被用来描绘纵向观测间的相关关系, 可以用均匀相关和序列相关等结构来刻画. 其中序列相关结构最为常见, 因为纵向数据常是短时间序列上的重复观测, 所以, 本文针对该结构进行更为深入的研究, 后文提及的模型 (2.1) 都是指具有序列相关结构的混合增长曲线模型.

序列相关结构 (SCS) 的具体形式为 $\Sigma = \sigma^2 \Gamma_\rho = \sigma^2 (\rho^{|i-j|})_{1 \leq i, j \leq p}$, 其中 $\sigma^2 > 0$, $-1 < \rho < 1$. Γ_ρ 刻画了相关结构, 而 Γ_ρ^{-1} 可以利用平稳自回归过程中观测变量的分布和 SCS 的广义对称性 (关于两条对角线都对称) 计算, 把 Γ_ρ^{-1} 分解为几个线性无关的矩阵相加, 则有 (参见文献 [9])

$$\Gamma_\rho^{-1} = \frac{1 + \rho^2}{1 - \rho^2} \mathbf{I} - \frac{\rho}{1 - \rho^2} \mathbf{M} - \frac{\rho^2}{1 - \rho^2} \mathbf{K}, \quad (2.2)$$

其中 $\mathbf{I}$ 为 p 维单位方阵, $\mathbf{M}$ 为一个两次对角线为 1、其余元素都为 0 的 p 维方阵, $\mathbf{K}$ 为位置 (1, 1) 与 (p, p) 的元素为 1、其他元素都为 0 的 p 维方阵. 容易看出, 当模型 (2.1) 具有 SCS 时参数估计变得更加困难, 因为这相当于在求解过程中对协方差增加了特定结构的约束条件, 在 SCS 下待估参数进一步地精简为 $\Theta = (\boldsymbol{\pi}, \mathbf{B}, \sigma^2, \rho)$.

3 模型参数的极大似然估计

模型 (2.1) 的概率密度函数是 $\mathbf{Y}$ 和 $\mathbf{Z}$ 的联合概率密度函数, 其对数形式如下:

$$\log f(\mathbf{Y}, \mathbf{Z}; \Theta) = \frac{1}{2} \sum_{j=1}^n \left\{ \log |\sigma^{-2} \Gamma_\rho^{-1}| - \frac{1}{\sigma^2} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j)^T \Gamma_\rho^{-1} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j) \right\} + \sum_{j=1}^n \sum_{k=1}^r z_{kj} \log \pi_k, \quad (3.1)$$

其中 z_{kj} 是列向量 $\mathbf{z}_j$ 的第 k 个元素. 可见对数联合似然函数 (3.1) 中, 即使已经求出 Γ_ρ^{-1} 还是无法顺

利进行极大化求解, 因为 (3.1) 中还含有未知组别阵 $\mathbf{Z}$, 其真实的似然函数须在 $f(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\Theta})$ 中将 $\mathbf{Z}$ 积分消除, 而此积分往往无解析表达式, 因此, 本文运用 EM 算法求解参数极大似然估计.

3.1 EM 算法

(1) 期望 (expectation, E) 步 计算辅助函数 $Q \triangleq Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) \triangleq E[\log f(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}]$, 它是在给定已知数据信息 $\mathbf{Y}$ 和参数 $\boldsymbol{\Theta}$ 的第 $t-1$ 步迭代估计值 (即 $\boldsymbol{\Theta}^{(t-1)}$) 时, (3.1) 关于未知组别阵 $\mathbf{Z}$ 的条件期望, 可以推出

$$Q = \frac{1}{2} \sum_{j=1}^n \left\{ \log |\sigma^{-2} \Gamma_{\rho}^{-1}| - \frac{1}{\sigma^2} \mathbf{y}_j^T \Gamma_{\rho}^{-1} \left(\mathbf{y}_j - 2 \sum_{k=1}^r \mathbf{X} \beta_k u_{kj}^{(t)} \right) - \frac{1}{\sigma^2} \text{tr} \left[\sum_{k=1}^r \Gamma_{\rho}^{-1} \mathbf{X} \beta_k \beta_k^T u_{kj}^{(t)} \mathbf{X}^T \right] \right\} + \sum_{j=1}^n \sum_{k=1}^r u_{kj}^{(t)} \log \pi_k, \quad (3.2)$$

其中

$$u_{kj}^{(t)} = E(z_{kj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = \frac{\pi_k^{(t-1)} \mathbf{p}_{kj}^{(t)}}{\sum_{k=1}^r \pi_k^{(t-1)} \mathbf{p}_{kj}^{(t)}}, \quad (3.3)$$

$$\mathbf{p}_{kj}^{(t)} = P(\mathbf{y}_j | z_{kj}^{(t)} = 1, \boldsymbol{\Theta}^{(t-1)}) = (2\pi)^{-\frac{p}{2}} |(\sigma^2 \Gamma_{\rho})^{(t-1)}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{v}_{kj}^{(t-1)})^T ((\sigma^2 \Gamma_{\rho})^{(t-1)})^{-1} \mathbf{v}_{kj}^{(t-1)} \right], \quad (3.4)$$

$$\mathbf{v}_{kj}^{(t-1)} = \mathbf{y}_j - \mathbf{X} \beta_k^{(t-1)}, \quad (3.5)$$

而 β_k 是回归系数阵 $\mathbf{B}_{m \times r}$ 的第 k 列列向量, 详细推导过程见附录 A.

(2) 最大 (maximization, M) 步 对辅助函数 (3.2) 关于待估参数 $\boldsymbol{\Theta} = (\pi, \mathbf{B}, \sigma^2, \rho)$ 极大化求解, 计算参数的极大似然估计.

首先, 用 Lagrange 乘子法极大化加约束惩罚的辅助函数 Q 求解 π ,

$$\begin{cases} \frac{\partial [Q + \lambda(\sum_{k=1}^r \pi_k - 1)]}{\partial \pi_k} = 0, \\ \sum_{k=1}^r \pi_k - 1 = 0, \end{cases} \quad (3.6)$$

可以求出第 t 步的分类分布后验概率为

$$\boldsymbol{\pi}^{(t)} = \frac{1}{n} \mathbf{1}_n^T E((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}), \quad (3.7)$$

其中 $\mathbf{1}_n$ 为元素全为 1 的 n 维列向量, $E((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}) = [E(z_1^{(t)} | \mathbf{y}_1, \boldsymbol{\Theta}^{(t-1)}), \dots, E(z_n^{(t)} | \mathbf{y}_n, \boldsymbol{\Theta}^{(t-1)})]^T$, 而 $E(z_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = [u_{1j}^{(t)}, \dots, u_{rj}^{(t)}]^T, j = 1, \dots, n$, 详细求解过程见附录 B.

然后, 求辅助函数 Q 分别关于 β_k 、 σ^2 和 ρ 的偏导数如下:

$$\frac{\partial Q}{\partial \beta_k} = \mathbf{X}^T \Gamma_{\rho}^{-1} \left[\sum_{j=1}^n \mathbf{y}_j u_{kj}^{(t)} - \sum_{j=1}^n u_{kj}^{(t)} \mathbf{X} \beta_k \right], \quad (3.8)$$

$$\frac{\partial Q}{\partial \sigma^2} = \sum_{j=1}^n \frac{1}{2} \text{tr} \left\{ -\frac{1}{\sigma^2} \mathbf{I} + \frac{1}{\sigma^4} \mathbf{S}_j^{(t)} \Gamma_{\rho}^{-1} \right\}, \quad (3.9)$$

$$\frac{\partial Q}{\partial \rho} = \sum_{j=1}^n \frac{1}{2\sigma^2} \text{tr} \left\{ (\sigma^2 \Gamma_{\rho} - \mathbf{S}_j^{(t)}) \left(\frac{4\rho}{(1-\rho^2)^2} \mathbf{I} - \frac{\rho^2+1}{(1-\rho^2)^2} \mathbf{M} - \frac{2\rho}{(1-\rho^2)^2} \mathbf{K} \right) \right\}, \quad (3.10)$$

其中 $\mathbf{S}_j^{(t)} = \sum_{k=1}^r \mathbf{v}_{kj}^{(t-1)} (\mathbf{v}_{kj}^{(t-1)})^T u_{kj}^{(t)}$.

令偏导数 $\partial Q / \partial \beta_k = 0$, $\partial Q / \partial \sigma^2 = 0$, 可得第 t 步的回归系数阵估计和方差估计为

$$\mathbf{B}^{(t)} = [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{Y} \mathbf{E}((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \mathbf{\Theta}^{(t-1)}) \mathbf{T}^{(t)}, \quad (3.11)$$

$$(\sigma^2)^{(t)} = \frac{[1 + (\rho^{(t-1)})^2] h_1^{(t)} - \rho^{(t-1)} h_2^{(t)} - (\rho^{(t-1)})^2 h_3^{(t)}}{[1 - (\rho^{(t-1)})^2] np}, \quad (3.12)$$

其中 $\mathbf{T}^{(t)} = [\text{diag}(\sum_{j=1}^n u_{1j}^{(t)}, \dots, \sum_{j=1}^n u_{rj}^{(t)})]^{-1}$, $h_1^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{I}$, $h_2^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{M}$, $h_3^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{K}$, 详细求解过程见附录 C 和 D.

令偏导函数 $\partial Q / \partial \rho = 0$, 可以得到包含 ρ 的三次方程

$$a^{(t)} \rho^3 + b^{(t)} \rho^2 + c^{(t)} \rho + d^{(t)} = 0, \quad (3.13)$$

其中 $a^{(t)} = -2n(p-1)(\sigma^2)^{(t-1)}$, $b^{(t)} = d^{(t)} = h_2^{(t)}$, $c^{(t)} = 2n(p-1)(\sigma^2)^{(t-1)} - 4h_1^{(t)} + 2h_3^{(t)}$. 利用三次方程求根公式可得第 t 步的自相关系数估计为

$$\begin{cases} \rho_1^{(t)} = \sqrt[3]{-\frac{\psi^{(t)}}{2} + \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}} + \sqrt[3]{-\frac{\psi^{(t)}}{2} - \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}}, \\ \rho_2^{(t)} = \omega \sqrt[3]{-\frac{\psi^{(t)}}{2} + \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}} + \omega^2 \sqrt[3]{-\frac{\psi^{(t)}}{2} - \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}}, \\ \rho_3^{(t)} = \omega^2 \sqrt[3]{-\frac{\psi^{(t)}}{2} + \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}} + \omega \sqrt[3]{-\frac{\psi^{(t)}}{2} - \sqrt{\left(\frac{\psi^{(t)}}{2}\right)^2 + \left(\frac{\phi^{(t)}}{3}\right)^3}}, \end{cases} \quad (3.14)$$

其中

$$\begin{aligned} \omega &= -1 + 3\sqrt{3}i, \\ \phi^{(t)} &= \frac{c^{(t)}}{a^{(t)}} - \frac{(b^{(t)})^2}{3(a^{(t)})^2}, \\ \psi^{(t)} &= \frac{d^{(t)}}{a^{(t)}} + \frac{2(b^{(t)})^3}{27(a^{(t)})^3} - \frac{b^{(t)}c^{(t)}}{3(a^{(t)})^2}. \end{aligned}$$

令 $\Delta^{(t)} = (\psi^{(t)}/2)^2 + (\phi^{(t)}/3)^3$ 为根的判别式: $\Delta^{(t)} > 0$ 时, 有一个实根两个虚根; $\Delta^{(t)} = 0$ 时, 有三个实根, 且其中至少有两个根相等; $\Delta^{(t)} < 0$ 时, 有三个不等实根. 又因为 $-1 < \rho_i^{(t)} < 1$, $i = 1, 2, 3$, 可计算出 $\rho^{(t)}$ 的解. 通过证明可知 $\rho^{(t)}$ 在 $(-1, 1)$ 区间至少存在一个根, 当 $\rho^{(t)}$ 有多个根时, 取能使对数联合似然函数值达到最大的根, 详细求解和推导过程见附录 E.

(3) 迭代计算 模型 (2.1) 的待估参数为 $\mathbf{\Theta} = (\boldsymbol{\pi}, \mathbf{B}, \sigma^2, \rho)$, 其需要迭代求解, 迭代初值需要设定其中 3 个再计算另外 1 个, 然后才能进入迭代计算.

(1) 设置迭代初值 $\mathbf{\Theta}^{(0)} = (\boldsymbol{\pi}^{(0)}, \mathbf{B}^{(0)}, (\sigma^2)^{(0)}, \rho^{(0)})$. 先通过数据分类分布的先验概率 $\boldsymbol{\pi}^{(0)}$, 常取作等概率向量, 在分类分布中随机产生 $\mathbf{Z}^{(0)}$ 的各个列向量, 构成随机组别阵. 再令 $(\sigma^2)^{(0)} = 0$, $\rho^{(0)} = 1$, 即 $\boldsymbol{\Sigma}^{(0)} = \mathbf{I}$, 将数据观测阵 $\mathbf{Y}$ 、设计阵 $\mathbf{X}$ 、协方差阵 $\mathbf{I}$ 和组别阵 $\mathbf{Z}^{(0)}$ 代入普通 GCM 的极大似然估计得到 $\mathbf{B}^{(0)}$, 此时得到 4 个需要的迭代初值. 需要注意, 由于 EM 算法对迭代初值的敏感性, 计算中随机产生了 s 个 $\mathbf{Z}^{(0)}$, 估计了 s 个 $\mathbf{B}^{(0)}$, 将其分别代回相应的对数联合似然函数, 选取使对数联合似然函数值达到最大的 $\mathbf{Z}^{(0)}$ 和 $\mathbf{B}^{(0)}$, s 可以取任意正整数, 本文设定 $s = 100$.

(2) 进入迭代计算. 迭代计算在 E 步和 M 步不断循环, 用迭代初值 $\Theta^{(0)}$ 更新 Q 函数, 将 $\Theta^{(0)}$ 代入 (3.7)、(3.11)、(3.12) 和 (3.14) 计算 $\Theta^{(1)} = (\pi^{(1)}, B^{(1)}, (\sigma^2)^{(1)}, \rho^{(1)})$. 以此类推, 用上一步的估计值 $\Theta^{(t-1)}$ 更新 Q 函数, 将 $\Theta^{(t-1)}$ 代入此 4 式计算可得

$$\Theta^{(t)} = (\pi^{(t)}, B^{(t)}, (\sigma^2)^{(t)}, \rho^{(t)}).$$

(3) 停止迭代计算. 当计算结果收敛时停止迭代, 即对任意小的阈值 ξ , 当 $\|\Theta^{(t)} - \Theta^{(t-1)}\| < \xi$ 时停止迭代, 将此时得到的 $\Theta^{(t)}$ 作为最终参数估计结果, 本文取 $\xi = 10^{-8}$.

3.2 参数估计的标准误差

EM 算法给出了参数的极大似然估计, 但是并没有给出参数估计的标准误差, 而 Louis 信息准则^[10]或自助抽样 (bootstrap) 法^[11]可以用来解决这个问题. 本文选用后一种方法来近似参数估计的标准误差, 度量估计的有效性, 相较于 Louis 信息准则, 模型 (2.1) 的自助抽样法计算程序简单、易实现. 在计算中, 先针对每个样本集实施有放回的重重复抽样, 产生 1,000 个与样本集等量的样本子集; 再用提出的方法在样本子集上进行 1,000 次参数估计, 获得参数的自助抽样分布; 最后关于每个参数计算 1,000 次估计的均值和标准差, 用均值近似参数估计的期望, 用标准差近似参数估计的标准误差, 因为统计量的自助抽样估计标准误差等于自助抽样分布的标准差.

3.3 组别阵 Z 的预测

EM 算法给出了模型 (2.1) 中参数的极大似然估计 $\hat{\Theta} = (\hat{\pi}, \hat{B}, \hat{\sigma}^2, \hat{\rho})$, 却依然没有给出随机组别阵的预测 $\hat{Z}$, 可以利用分类分布最大后验 (maximum a posteriori, MAP) 概率进行预测. 因为 $E((Z^{(t)})^T | Y, \Theta^{(t-1)}) = [E(z_1^{(t)} | y_1, \Theta^{(t-1)}), \dots, E(z_n^{(t)} | y_n, \Theta^{(t-1)})]^T$, 而 $E(z_j^{(t)} | y_j, \Theta^{(t-1)}) = [u_{1j}^{(t)}, \dots, u_{rj}^{(t)}]^T$, 故在 EM 算法收敛后有 $\hat{Z} = [\hat{z}_1, \dots, \hat{z}_n]$, 而 $\hat{z}_j = [\hat{z}_{1j}, \dots, \hat{z}_{rj}]^T$ ($j = 1, \dots, n$), 对 $\hat{z}_{kj}$ ($k = 1, \dots, r$), 有

$$\hat{z}_{kj} = \begin{cases} 1, & \text{若 } \hat{u}_{kj} > \hat{u}_{\tau j} \text{ } (\tau = 1, \dots, r \text{ 且 } \tau \neq k), \\ 0, & \text{否则,} \end{cases} \quad (3.15)$$

其中 $\hat{u}_{kj}$ 是迭代停止时 (3.3) 的最终计算结果, 即预测组别阵 $\hat{Z}$ 是将每个个体判别给其对应分类分布后验概率最大的类别.

实际数据分析中, 不仅观测个体的所属组别可能无法获知, 甚至连组别数 r 也可能无法确定, 即组别阵 $Z_{r \times n}$ 的维数 r 无法确定. 针对这个具有挑战性的问题, 本文提出一种数据驱动的方法来确定维数 r . 以 AIC 或 BIC 来评价模型, 在 r 的定义范围中进行搜索, 寻找使 AIC 值或 BIC 值达到最小的最优模型, 其对应的 r 值为最佳取值. AIC 和 BIC 的定义如下:

$$\text{AIC} = -2l(\hat{\Theta}) + 2\eta, \quad (3.16)$$

$$\text{BIC} = -2l(\hat{\Theta}) + \eta \log(n), \quad (3.17)$$

其中 $l(\cdot)$ 为对数联合似然函数 (3.1), η 为参数维数, n 为样本量.

同样的方法也可以用来搜索拟合曲线的阶数 m , 即设计阵 $X_{p \times m}$ 的维数 m . 对 m 和 r 最佳取值的搜寻可看作对未知回归系数阵 $B_{m \times r}$ 的维数确定.

4 模拟研究

为研究模型 (2.1) 的性质, 本文共设计了 4 个模拟实验: (1) 模型拟合效果对 B 、 σ^2 和 ρ 的敏感性, (2) 模型拟合效果对 π 的敏感性及参数估计的性质, (3) 误设相关性的影响, (4) 忽视相关性的影响, 分别进行了大量的模拟计算来讨论模型的拟合效果和协方差结构的影响. 实验 (2) 和 (3) 都重复了 1,000 次子实验来计算子实验环境中模型的拟合效果, 而含有大量子实验的实验 (1) 和 (4) 仅重复了 100 次子实验, 其实验结果依旧呈现出了规律性, 并没有因此设定而出现估计偏误和结果混乱.

实验曲线设计为两组特定的增长曲线, 在模型 (2.1) 中设 $p = 7$, 即个体有 7 次观测, 设 $m = 4$, 即用 $t = (1, 2, 3, 4, 5, 6, 7)^T$ 的 3 次多项式来拟合数据. 第一组差异曲线 (Curves.A, Cs.A) 设 $r = 2$, 两类曲线形状差异明显, 一类上凸, 一类下凹, 如图 1(a) 所示. 第二组异阶多项式 (Curves.B, Cs.B) 设 $r = 3$, 三类曲线是截距差甚微的一、二、三阶多项式, 依次为直线、上凸曲线、“S” 形曲线, 如图 1(d) 所示. 两组曲线的多项式回归系数阵分别为

$$B_{Cs.A} = \begin{bmatrix} 0 & 22 & -2.2 & 0 \\ 30 & -28 & 8.8 & -0.6 \end{bmatrix}^T, \quad B_{Cs.B} = \begin{bmatrix} 3.89 & 7.42 & 0 & 0 \\ 0 & 16.13 & -1.34 & 0 \\ 10 & -8 & 6 & -0.6 \end{bmatrix}^T. \quad (4.1)$$

图 1 (网络版彩图) 相关系数与方差的变化尺度. (a) Cs.A: 差异曲线; (b) Cs.A: $\hat{Z}$ 的误判率; (c) Cs.A: $\hat{B}$ 的 Euclid 距离; (d) Cs.B: 异阶多项式; (e) Cs.B: $\hat{Z}$ 的误判率; (f) Cs.B: $\hat{B}$ 的 Euclid 距离

模拟数据通过以下 6 个步骤生成: (1) 设定真实曲线, 即设定 $\mathbf{B}$ 和 $\mathbf{X}$; (2) 生成随机组别阵, 先设定分类分布的参数真值 $\boldsymbol{\pi}$, 再由该特定分布随机产生 $\mathbf{Z}$ 的 n 个列向量; (3) 生成数据均值, 根据模型 (2.1), 用 $\mathbf{XBZ}$ 计算数据集均值; (4) 生成协方差矩阵, 设定真值 ρ 和 σ^2 , 由序列相关结构 $\boldsymbol{\Sigma} = \sigma^2(\rho^{|i-j|})_{1 \leq i, j \leq p}$ 构建协方差; (5) 生成随机扰动, 将第 4 步计算的 $\boldsymbol{\Sigma}$ 代入正态分布 $N_p(0, \boldsymbol{\Sigma})$, 随机产生随机误差 $\boldsymbol{\varepsilon}$ 的 n 个列向量; (6) 生成模拟数据, 依据模型 (2.1) 用 $\mathbf{XBZ} + \boldsymbol{\varepsilon}$ 计算数据集 $\mathbf{Y}$.

估计效果简单地用模型 (2.1) 中组别阵预测 $\hat{\mathbf{Z}}$ 的误判率 $\text{mis.rate}(\hat{\mathbf{Z}})$ 及参数估计 $\hat{\boldsymbol{\pi}}, \hat{\mathbf{B}}, \hat{\boldsymbol{\Sigma}}$ 分别与其真值的 Euclid 距离 $\text{euc.dis.}(\hat{\boldsymbol{\pi}}), \text{euc.dis.}(\hat{\mathbf{B}}), \text{euc.dis.}(\hat{\boldsymbol{\Sigma}})$, 以及参数估计 $\hat{\rho}$ 和 $\hat{\sigma}^2$ 的均方误差 (mean square error, MSE) $\text{MSE}(\hat{\rho}), \text{MSE}(\hat{\sigma}^2)$ 来衡量, 其中 $\text{mis.rate}(\hat{\mathbf{Z}})$ 是 MAP 预测的 $\hat{\mathbf{Z}}$ 与真值 $\mathbf{Z}$ 不一致的个体数占样本量的比率, 而真值 $\mathbf{Z}$ 是模拟数据生成中第 2 步产生的随机组别阵, 因为第 6 步的最终数据由它计算, 所以, 它是模拟数据的真实组别阵. 另外, SCS 形式的 $\text{euc.dis.}(\hat{\boldsymbol{\Sigma}})$ 计算, 先将 $\hat{\rho}, \hat{\sigma}^2$ 和 ρ, σ^2 代回 SCS 的定义式还原协方差的估计与真值, 再来计算 Euclid 距离.

4.1 模型拟合效果对 $\mathbf{B}, \sigma^2$ 和 ρ 的敏感性

此实验用来讨论模型 (2.1) 对 $\mathbf{B}, \sigma^2$ 和 ρ 变化的适应能力, 因为模型的待估参数为 $\boldsymbol{\Theta} = (\boldsymbol{\pi}, \mathbf{B}, \sigma^2, \rho)$, 所以模拟实验依次来探讨这些参数取不同的真值时模型的拟合效果, 即模型 (2.1) 对各种数据情形的适应程度. 此实验先后设定模型 (2.1) 的 $\mathbf{B}$ 为 (4.1) 中的 $\mathbf{B}_{\text{Cs.A}}$ 和 $\mathbf{B}_{\text{Cs.B}}$, 则相应的设计阵 $\mathbf{X}$ 被确定, 即设定实验曲线组为图 1(a) 和 1(d). 再固定 $\boldsymbol{\pi}$ 后, 模型仅随着 σ^2 和 ρ 的变化而变化. 对不断变化的 σ^2 和 ρ 求模型的参数估计, 对比这一系列估计效果的变化. 该实验测试了模型 (2.1) 在不同数据情形中参数估计的概况, 并为更加详细的实验搜寻出较为合适的参数设置.

将生成模拟数据的真值 ρ 与 σ^2 设定在合理的范围开展实验, 根据模型 (2.1) 的设定条件 $-1 < \rho < 1$, 取 ρ 的步长为 0.1, 研究 $\rho = -0.99, -0.9, \dots, 0.9, 0.99$ 的 21 个真值情况. 由变异系数和简单模拟讨论, 在 Cs.A 下, 取 σ^2 的步长为 4, 研究 $\sigma^2 = 4, 8, \dots, 80, 84$ 的 21 个真值情况, 再设 $n = 50$, $\boldsymbol{\pi} = (0.4, 0.6)$, 则上凸曲线有 20 条, 下凹曲线有 30 条; 在 Cs.B 下, 取 σ^2 的步长为 2, 研究 $\sigma^2 = 2, 4, \dots, 40, 42$ 的 21 个真值情况, 再设 $n = 150$, $\boldsymbol{\pi} = (0.200, 0.333, 0.0.467)$, 则一次曲线有 30 条, 二次曲线有 50 条, 三次曲线有 70 条. 在 Cs.A 和 Cs.B 下分别对 ρ 和 σ^2 进行交叉研究, 每组曲线计算了 $21 \times 21 = 441$ 个子实验, 每个子实验均重复实行 100 次数据生成与参数估计, 估计评价 $\text{mis.rate}(\hat{\mathbf{Z}})$ 和 $\text{euc.dis.}(\hat{\mathbf{B}})$ 取 100 次评价的均值.

将 $\text{mis.rate}(\hat{\mathbf{Z}})$ 和 $\text{euc.dis.}(\hat{\mathbf{B}})$ 反映在 σ^2 与 ρ 交叉的热图 (heat map) 上, 如图 1(b)、1(c)、1(e) 和 1(f) 所示, 横轴为 σ^2 , 纵轴为 ρ , mis.rate 和 $\text{euc.dis.}\mathbf{B}$ 的大小在热图中展示为颜色的深浅, 由每张热图上方的图例标度. 热图 1 中可以轻松地比较模型对两组实验曲线的拟合效果, 并看出模型估计效果与 σ^2 和 ρ 的变化关系. 在图 1 中找出颜色变化较为明显的区域, 即是模型对该组实验曲线拟合效果较为敏感的区域, 该区域对应的 σ^2 和 ρ 数值就是模型拟合较为敏感的方差和自回归系数数值.

图 1(b)、1(c)、1(e) 和 1(f) 的颜色都随着 σ^2 的增加而变深, 同时随着 $|\rho - \delta|$ 的减小而变深, 说明随着方差的增大和自回归系数与对称点的靠近, 两组曲线的模型拟合效果越来越差. 从图 1(b) 和 1(e) 可以看出两组曲线中 mis.rate 的 $\delta \approx 0.5$, 存在上偏的情形 (这极可能源于异常大的 σ^2 设定, $\hat{\mathbf{Z}}$ 为最大后验预测, 由计算式可知其会受到 σ^2 波动的影响); 从图 1(c) 和 1(f) 可以看出两组曲线中 $\text{euc.dis.}\mathbf{B}$ 的 $\delta \approx 0$ (易于理解, $\hat{\mathbf{B}}$ 是 EM 算法收敛时的参数估计, 即是极大似然估计^[12], 其具有相合性, 不易受到 σ^2 波动的影响). 相反地, 可以得到, 数据的方差越小, 自相关性越强, 模型 (2.1) 的拟合和估计效果越好.

在图 1 中比较 Cs.A 和 Cs.B 的拟合效果, 通过对比横轴变化范围、图例变化尺度和深色区域面积, 可以推断模型在 Cs.A 中比在 Cs.B 中的误判率更少, 回归系数阵估计偏差更小, 拟合效果更好. 可以简单概括为: 增长曲线的形状差异越大, 模型 (2.1) 的拟合效果越好.

为了明显地展示模型 (2.1) 的拟合和估计效果, 下面三个实验将针对模型拟合效果较差的 Cs.B 曲线的敏感位置进行更为详细的研究. 因为图 1(e) 和 1(f) 的颜色开始有敏感变化的位置为 $\sigma^2 \approx 22$, $\rho \approx \pm 0.7$, 所以下面的实验参数设定在这些位置附近.

4.2 模型拟合效果对 π 的敏感性及参数估计的性质

此实验用来讨论模型 (2.1) 对 π 变化的适应能力, 并详细地讨论模型 (2.1) 中估计的性质, 进而评价本文提出的估计方法的有效性. 因为模型 (2.1) 的待估参数 $\Theta = (\pi, B, \sigma^2, \rho)$, 所以研究模型对样本比例 π 变化的反应需要设定其他 3 个参数真值, 先设定 $B = B_{Cs.B}$, 再设定 $\rho = 0.7$, $\sigma^2 = 9, 16, 25, 36$. 然后, 分别在样本量 $n = 60, 150, 600$ 的情形下研究样本比例依次为 $1:1:1$ 、 $3:5:7$ 和 $2:4:9$ 的参数估计效果, 总共进行了 $4 \times 3 \times 3 = 36$ 个子实验, 每个子实验重复实行 1,000 次数据生成, 从而获得 1,000 组参数估计值和估计评价, 计算它们的均值和标准差, 相应的结果如表 1 所示.

表 1 显示, 第 1 个子实验的模拟真值 $\rho = 0.7$, $\sigma^2 = 9$, 模拟样本量 $n = 60$, 样本比例为 $1:1:1$, 即 Cs.B 中一次、二次、三次曲线的个体数都为 20; 1,000 次重复估计后分类分布后验概率估计的均值为 $(0.334, 0.333, 0.333)$, 标准差为 $(0.004, 0.007, 0.005)$; 自回归系数估计的均值为 0.689, 标准差为 0.034; 方差估计的均值为 8.551, 标准差为 0.922; 回归系数阵估计与真值的 Euclid 距离均值为 4.555, 标准差为 7.200; 组别阵最大后验预测的误判率均值为 3.11%, 标准差为 0.133.

当参数的真值仅有 π 不同时, 观察子实验 1-3、子实验 4-6、...、子实验 31-33、子实验 34-36, 可见参数估计 $\hat{\rho}$ 、 $\hat{\sigma}^2$ 和 $\hat{\pi}$ 的均值和标准差都比较一致, 没有受到样本比例的影响; 而 $\text{euc.dis}(\hat{B})$ 的均值和标准差在样本比例为 $2:4:9$ 时普遍要大一点, 它受到样本比例的影响相对明显, 而随着样本量的增加, 该样本比例的影响逐渐削弱, 易知单类个体数的减少时会导致拟合效果变差; $\text{mis.rate}(\hat{Z})$ 的均值和标准差在样本比例为 $1:1:1$ 时普遍要大一点, 说明误判率主要来自一次和二次曲线, 它们的形状差异不是很明显. 可以归纳为: 参数估计对样本比例的敏感性主要来自于曲线的形状和单类个体数, 模型拟合效果几乎不受样本比例的影响.

当仅有模拟样本量 n 不一样时, 观察子实验 1、4 和 7, 子实验 2、5 和 8, ..., 子实验 29、32 和 35, 子实验 30、33 和 36, 可以看出随着 n 增大, 估计 $\hat{\pi}$ 、 $\hat{\rho}$ 、 $\hat{\sigma}^2$ 、 $\hat{B}$ 和预测 $\hat{Z}$ 的均值逐渐逼近真值, 估计标准差和估计评价标准差也逐渐减小. 这说明模型 (2.1) 的估计 $\hat{\pi}$ 、 $\hat{\rho}$ 、 $\hat{\sigma}^2$ 和 $\hat{B}$ 是渐近无偏估计.

4.3 误设相关性的影响

此实验用来研究误设协方差阵结构对混合增长曲线模型 (GCMM) 的影响, 对比分别具有 SCS 和 UC 的 GCMM 对序列相关数据的拟合和估计效果, 进而讨论当协方差结构被误设为 UC 时对 GCMM 的拟合和估计造成的影响. 设定回归系数真值 $B = B_{Cs.B}$, 样本量 $n = 150$, 样本比例为 $3:5:7$, 即固定分类分布概率真值 $\pi = (0.20, 0.33, 0.47)$, 方差真值 $\sigma^2 = 36$, 取自相关系数真值 ρ 分别等于 0、0.1、0.3、0.5、0.7 和 0.9, 分别对 6 种参数情形重复实行 1,000 次数据生成, 每次生成的数据分别用具有 SCS 和 UC 的 GCMM 进行拟合和估计, 两种模型的估计评价均取 1,000 次重复评价的均值, 总共进行了 $6 \times 2 = 12$ 个子实验.

表 1 SCS-Cs.B: $\rho = 0.7$ 时, 重复 1,000 次的平均估计效果

序号	ρ	σ^2	n	比例	$\hat{\pi}$		$\hat{\rho}$		$\hat{\sigma}^2$		euc.dis.($\hat{B}$)		mis.rate($\hat{Z}$)	
					均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	0.7	9	60	1:1:1	(0.334, 0.333, 0.333)	(0.004, 0.007, 0.005)	0.689	0.034	8.551	0.922	4.555	7.200	3.11%	0.133
2				3:5:7	(0.200, 0.336, 0.464)	(0.004, 0.020, 0.019)	0.688	0.037	8.576	0.943	3.859	4.925	1.82%	0.108
3				2:4:9	(0.140, 0.268, 0.592)	(0.040, 0.034, 0.056)	0.690	0.037	8.678	1.136	4.371	4.440	1.88%	0.105
4			150	1:1:1	(0.334, 0.333, 0.333)	(0.003, 0.004, 0.003)	0.696	0.021	8.839	0.605	2.037	2.402	0.57%	0.042
5				3:5:7	(0.200, 0.333, 0.467)	(0.003, 0.004, 0.004)	0.696	0.021	8.844	0.579	2.012	0.780	0.35%	0.005
6				2:4:9	(0.134, 0.267, 0.600)	(0.012, 0.008, 0.006)	0.696	0.022	8.843	0.629	2.272	1.502	0.32%	0.013
7			600	1:1:1	(0.333, 0.333, 0.334)	(0.001, 0.002, 0.002)	0.700	0.011	8.973	0.296	0.962	0.357	0.29%	0.002
8				3:5:7	(0.200, 0.333, 0.467)	(0.001, 0.002, 0.002)	0.698	0.011	8.939	0.303	1.021	0.399	0.29%	0.002
9				2:4:9	(0.133, 0.267, 0.600)	(0.001, 0.002, 0.002)	0.699	0.011	8.965	0.286	1.140	0.482	0.26%	0.002
10				1:1:1	(0.333, 0.334, 0.333)	(0.015, 0.020, 0.016)	0.690	0.037	15.250	1.634	4.987	5.263	4.00%	0.096
11			60	3:5:7	(0.200, 0.335, 0.465)	(0.014, 0.025, 0.022)	0.687	0.037	15.152	1.706	4.636	3.005	3.12%	0.067
12				2:4:9	(0.136, 0.266, 0.597)	(0.024, 0.024, 0.027)	0.689	0.039	15.211	1.774	5.113	2.651	2.67%	0.047
13				1:1:1	(0.333, 0.333, 0.334)	(0.010, 0.012, 0.010)	0.695	0.023	15.697	1.075	2.599	0.979	2.37%	0.013
14	0.7	16	150	3:5:7	(0.200, 0.333, 0.466)	(0.008, 0.013, 0.011)	0.694	0.023	15.691	1.063	2.787	1.092	2.38%	0.013
15				2:4:9	(0.134, 0.266, 0.600)	(0.007, 0.012, 0.011)	0.696	0.022	15.698	1.077	3.227	1.309	2.16%	0.012
16				1:1:1	(0.334, 0.333, 0.333)	(0.004, 0.006, 0.005)	0.699	0.011	15.945	0.559	1.305	0.488	2.27%	0.006
17			600	3:5:7	(0.200, 0.333, 0.466)	(0.004, 0.006, 0.005)	0.698	0.011	15.916	0.552	1.388	0.575	2.21%	0.006
18				2:4:9	(0.133, 0.267, 0.600)	(0.003, 0.006, 0.005)	0.698	0.011	15.900	0.579	1.544	0.642	2.00%	0.006
19	0.7	25	60	1:1:1	(0.334, 0.333, 0.333)	(0.032, 0.044, 0.036)	0.685	0.041	23.426	2.705	5.814	3.606	8.05%	0.063
20				3:5:7	(0.203, 0.333, 0.464)	(0.030, 0.044, 0.039)	0.687	0.041	23.715	2.817	5.964	2.382	7.55%	0.038
21				2:4:9	(0.139, 0.270, 0.590)	(0.035, 0.051, 0.053)	0.685	0.040	23.547	2.696	7.248	3.925	7.58%	0.064
22			150	1:1:1	(0.333, 0.334, 0.333)	(0.019, 0.027, 0.021)	0.694	0.024	24.391	1.782	3.483	1.325	6.70%	0.021
23				3:5:7	(0.201, 0.335, 0.465)	(0.018, 0.025, 0.021)	0.693	0.025	24.333	1.785	3.695	1.454	6.45%	0.020
24				2:4:9	(0.134, 0.268, 0.598)	(0.016, 0.025, 0.022)	0.692	0.025	24.328	1.816	4.148	1.711	6.12%	0.021
25			600	1:1:1	(0.333, 0.333, 0.334)	(0.010, 0.013, 0.010)	0.698	0.011	24.810	0.878	1.750	0.646	6.25%	0.010
26				3:5:7	(0.200, 0.334, 0.466)	(0.008, 0.012, 0.010)	0.698	0.012	24.846	0.930	1.802	0.669	6.14%	0.010
27				2:4:9	(0.134, 0.267, 0.599)	(0.007, 0.012, 0.010)	0.697	0.012	24.773	0.882	2.070	0.870	5.60%	0.010
28	0.7	36	60	1:1:1	(0.331, 0.336, 0.333)	(0.056, 0.073, 0.062)	0.684	0.041	33.705	3.985	8.098	4.825	14.61%	0.079
29				3:5:7	(0.209, 0.332, 0.459)	(0.054, 0.076, 0.068)	0.680	0.045	33.434	4.174	8.024	4.010	14.06%	0.073
30				2:4:9	(0.152, 0.278, 0.569)	(0.063, 0.085, 0.088)	0.681	0.043	33.373	4.129	9.872	5.239	14.52%	0.092
31			150	1:1:1	(0.333, 0.335, 0.332)	(0.031, 0.042, 0.034)	0.694	0.026	35.225	2.680	4.549	1.803	12.18%	0.029
32				3:5:7	(0.203, 0.332, 0.465)	(0.029, 0.042, 0.037)	0.692	0.026	34.969	2.777	4.904	1.942	11.86%	0.029
33				2:4:9	(0.137, 0.270, 0.593)	(0.030, 0.043, 0.038)	0.694	0.025	34.950	2.610	5.619	2.413	10.94%	0.030
34			600	1:1:1	(0.333, 0.334, 0.333)	(0.015, 0.019, 0.016)	0.698	0.012	35.731	1.357	2.311	0.860	11.20%	0.013
35				3:5:7	(0.200, 0.334, 0.466)	(0.014, 0.021, 0.017)	0.699	0.012	35.803	1.352	2.398	0.944	10.94%	0.013
36				2:4:9	(0.134, 0.267, 0.599)	(0.013, 0.020, 0.018)	0.698	0.013	35.746	1.357	2.773	1.130	9.95%	0.013

在具有 UC 的 GCMM 中, 分类分布后验概率估计 $\pi^{(t)}$ 同 (3.7) 一样, 而回归系数阵估计 $B^{(t)}$ 的形式是将 (3.11) 中的 $\Gamma^{(t-1)}$ 替换成 $\Sigma^{(t-1)}$, 协方差阵的估计如下 (参见文献 [4]):

$$\Sigma^{(t)} = \frac{1}{n} \sum_{j=1}^n \sum_{k=1}^r v_{kj}^{(t-1)} (v_{kj}^{(t-1)})^T u_{kj}^{(t)}, \quad (4.2)$$

其中 $u_{kj}^{(t)}$ 和 $v_{kj}^{(t-1)}$ 分别如 (3.3) 和 (3.5) 所示.

表 2 展示了该实验的结果, 当产生模拟数据的自回归系数真值 $\rho = 0$ 时, 相当于数据来自纵向观测彼此独立的情形, 4 个估计评价在 SCS 下都比在 UC 下更小, 即模型 (2.1) 的表现更好. 再在所有自回归系数的取值情形下进行比较, 可见所有参数在 SCS 下的估计效果都要好于 UC 下的估计效果. 容易知道, 当数据纵向观测独立或序列相关时, 模型 (2.1) 的估计表现比具有 UC 的 GCMM 更佳.

表 2 按列观察可见, 无论在 SCS 下还是在 UC 下 $\text{euc.dis.}(\hat{B})$ 和 $\text{euc.dis.}(\hat{\pi})$ 都随着 ρ 的增大而减小, 而 $\text{mis.rate}(\hat{Z})$ 的变化拐点在 $\rho = 0.5$ 左右, 跟第 1 个实验的分析结果一致. $\text{euc.dis.}(\hat{\Sigma})$ 在 SCS 下随着 ρ 的增大而增大; 在 UC 下的变化拐点在 $\rho = 0.5$ 左右, 但都大于 SCS 下的值. 该现象可描述为: 对具有序列相关性的数据, 随着数据序列相关性的增强, 具有 SCS 与 UC 的模型参数估计效果都在改善, 但是具有 SCS 的模型始终要优于具有 UC 的模型. 误设协方差结构会使 GCMM 的拟合和估计效果变糟糕.

4.4 忽视相关性的影响

此实验用来研究考虑相关协方差结构对混合增长曲线模型 (GCMM) 的重要性, 对比分别具有 SCS 和独立协方差结构的 GCMM 对序列相关数据的拟合和估计效果, 进而讨论忽视数据的序列相关性对 GCMM 的拟合和估计具有什么影响. 当 SCS 中的 ρ 为 0 时, $\Gamma_\rho = I$ (I 为 p 维单位方阵), 其退化为独立协方差结构, 所以把模型 (2.1) 中的 ρ 限制为 0 相当于强行用独立协方差 $\sigma^2 I$ 对数据进行拟合, 限定数据在 p 次纵向观测间不相关. 设定回归系数真值 $B = B_{\text{Cs.B}}$, 样本量 $n = 150$, 样本比例为 3:5:7, 即固定真值 $\pi = (0.20, 0.33, 0.47)$, 方差真值 $\sigma^2 = 36$, 在 $\rho = [0, 1)$ 的区间取步长 0.01 的真值, 对 100 种参数情形分别实行 100 次数据生成. 对每次生成的数据都用模型 (2.1) 和将 ρ 限制为 0 的模型 (2.1) 来拟合估计, 两种模型的估计评价均取 100 次重复评价的均值, 共进行了 $100 \times 2 = 200$ 个子实验.

将 5 种模型估计评价绘制于散点图 2 中, 横轴为自回归系数 ρ , 其取值越大代表数据的序列相关性越强; 纵轴为各个参数估计评价, 其取值越大代表估计距离真值越远, 估计效果越差. 在纵轴上红色圆点代表在模型中把 ρ 限制为 0 的相应估计评价, 青色角点代表正常模型的相应估计评价. 可见, 红

表 2 SCS-Cs.B: 协方差无结构的平均效果对比

ρ	$\text{mis.rate}(\hat{Z})$		$\text{euc.dis.}(\hat{B})$		$\text{euc.dis.}(\hat{\pi})$		$\text{euc.dis.}(\hat{\Sigma})$	
	SCS	UC	SCS	UC	SCS	UC	SCS	UC
0.0	10.47%	14.65%	7.222	9.535	0.053	0.087	5.797	29.574
0.1	12.08%	16.68%	7.290	10.049	0.060	0.093	5.810	31.475
0.3	14.22%	20.43%	7.189	10.592	0.069	0.115	7.077	36.310
0.5	14.93%	22.71%	6.301	10.496	0.070	0.130	9.507	41.027
0.7	11.99%	17.21%	4.910	7.338	0.054	0.098	13.627	35.957
0.9	2.03%	3.01%	2.691	3.058	0.014	0.022	18.290	25.140

图 2 (网络版彩图) 实验组 Cs.B 中 ρ 限制为 0 的估计效果对比. (a) $\hat{Z}$ 的误判率对比; (b) $\hat{B}$ 与真值的 Euclid 距离对比; (c) $\hat{\pi}$ 与真值的 Euclid 距离对比; (d) 相关系数 $\hat{\rho}$ 的均方误差对比; (e) 方差 $\hat{\sigma}^2$ 的均方误差对比

色圆点具有随着 ρ 增加而迅速上扬的趋势, 拟合效果越来越差; 相反青色角点具有随着 ρ 增加而下降或稳定的趋势, 拟合效果越来越好. 可以简单概括为: 随着数据序列相关性的增强, 将协方差限制为独立协方差结构的模型估计效果迅速变糟, 而正确选用协方差结构的模型 (2.1) 估计效果越来越优异. 忽视数据带有的相关性会使数据分析得到错误的结果.

综上, 4 个模拟实验都说明了模型 (2.1) 能对序列相关数据进行有效的拟合与估计, 并且 GCM 选用正确的协方差结构对数据分析有着不容忽视的影响.

5 酵母基因表达数据分析

最后, 模型 (2.1) 用来实证分析一个酵母基因表达水平数据, 这个实际数据来自文献 [13] 对酿酒酵母几乎所有基因的微阵列研究. 数据记录了酿酒酵母从厌氧发酵到有氧呼吸的代谢转换过程中的基因表达, 分别在双氧转换的 7 个时刻点上测量了它们的表达水平. 该数据共有 6,400 个基因, 在过滤掉不感兴趣、表达值极低和波动不明显等问题的基因之后, 取出了 614 个基因小样, 其可以在文献 [13] 索引的网络上找到. 将其基因表达水平按测量时刻点绘制成时间序列折线图, 如图 3 所示.

研究酵母基因组中每个基因的功能有助于了解细胞怎样通过全局基因表达模式响应环境变化, 文献 [13] 通过基因编译关键酶的引导与抑制比 (即基因表达水平) 识别了 6 种不一样的表达序列样式, 并基于表达样式的相似程度将 212 个基因进行分组, 利用组内的已知代表性基因探究各组基因的相似功能. 此外, 文献 [8] 还使用 K 均值的方法对该数据进行了聚类分析, 将 614 个基因聚类为 16 类. 本文希望由数据驱动的方式用模型 (2.1) 对该数据进行分析, 实现基因聚类, 并探索基因表达过程中的序

图 3 (网络版彩图) 酿酒酵母基因表达水平序列

列相关性. 同时, 将模型 (2.1) 的分析与 K 均值的分析进行简单对比.

该数据的基因真实组别不可获知, 对酵母基因数据建立模型 (2.1). 根据数据情形, 酵母基因表达水平为观测阵 $\mathbf{Y}_{p \times n}$, 其中 $p = 7$, $n = 614$; 设计阵 $\mathbf{X}_{p \times m}$ 为时间 $t = (1, 2, \dots, 7)^T$ 的多项式, 多项式阶数 m 待选择; 回归系数阵 $\mathbf{B}_{m \times r}$ 待估; 随机组别阵 $\mathbf{Z}_{r \times n}$ 也待预测, 其中组别数 r 同样待选择, 此时, 待估参数 $\Theta = (\mathbf{B}, \pi, \sigma^2, \rho)$, 待预测随机变量为 $\mathbf{Z}$, 待选择维数为 m 和 r .

5.1 与 K 均值结果比较分析

文献 [8] 通过 K 均值对该数据进行了聚类分析, 用 MSE 选出 16 个类别. 为了与此方法进行对比, 同样设定模型 (2.1) 的 $r = 16$, 通过观察图 3 中基因表达序列的样式, 用三阶多项式可以对该数据进行合理的拟合, 故设定 $m = 4$, 用模型 (2.1) 研究该数据, 对比 K 均值的分析结果.

K 均值分析对聚类得到的数据进行了组内均值画像, 将分类数据在 7 个观测点上计算均值, 用均值作出各类的折线图, 如图 4(c) 所示. 进一步, 也将模型 (2.1) 分析结果的每一类拟合值、组内均值分别画像如图 4(a) 和 4(b), 其中 (a) GCMM-SCS 聚类拟合画像就是用时间代入拟合多项式计算得到的预测值绘成的折线图, (b) GCMM-SCS 聚类中心画像就是用聚类后的组别数据在时间点上的组内均值绘成的折线图. 对比图 4(b) 和 4(c), 两种聚类方法得到的类别样式大致一样, 仅有第 3 列中的两个折线图有显著差异, 这是由不同聚类方法的侧重点不一样引起的. 另外, 本文提出的方法可以同时给出多项式回归系数, 即回归系数阵的估计 $\hat{\mathbf{B}}$, 可以进一步地画出拟合曲线 (图 4(a)), 直接描述变量变化关系. 此外, 该观测数据显然是短时间序列, 本文提出的方法就是着重分析纵向观测间的序列相关性, 通过模拟研究可知, 相对于简单地对数据进行聚类, 加入纵向相关性的考虑也不容忽视, 所以本文提出的方法会得到一些独特的分析结果.

然而, 一方面, 在图 4(b) 中注意到几个组别形状比较相似, 发现个别组别中包含的个体数极少, 如第 3 类中仅含有 2 个基因的序列折线, 这样的参数设定使本文提出的方法对该数据的聚类效果尚未达到最优; 另一方面, 在 $m = 4$, $r = 16$ 的设定下, 自相关系数的估计 $\hat{\rho} = 0.035$, 纵向观测之间接近独立, 从图 3 来看这显然不合理. 接下来, 用本文提出的方法搜寻分析该数据的最优模型.

图 4 (网络版彩图) SCS 混合增长曲线模型下基因表达水平聚类画像对比. (a) GCMM-SCS 聚类拟合画像; (b) GCMM-SCS 聚类中心画像; (c) K 均值聚类中心画像

5.2 最优模型分析

搜寻模型 (2.1) 的维数 m 和 r : 第一, 可依据经验预估一个大致范围, 本文先根据纵向观测次数 $p = 7$ 和折线图 3 中基因表达序列的形状, 将多项式阶数 m 的搜索范围设为 $[2, 8]$ 中的整数; 再参见文献 [8, 13] 的研究与分析, 将组别数 r 的搜索范围设为 $[2, 20]$ 中的整数. 第二, 设 $m = 2$, 使 r 在搜索范围逐个取值, 在 r 的每一个取值下都用本文提出的估计方法对模型 (2.1) 进行一次估计, 并计算相应的模型评价值 $AIC_{m=2, r=2, \dots, 20}$ 和 $BIC_{m=2, r=2, \dots, 20}$. 第三, 变换第二步中的 m 值, 令 m 在其搜索范围逐个取值, 重复第二步, 最终得到相应的模型评价值 $AIC_{m=2, \dots, 8, r=2, \dots, 20}$ 和 $BIC_{m=2, \dots, 8, r=2, \dots, 20}$. 第四, 利用 AIC 或 BIC, 在所有的模型评价值中选出最小值, 其对应的 m 和 r 值就是寻找的维数估计值, 相应模型就是最佳模型.

酵母基因数据的 m 和 r 搜寻过程如表 3 所示, 列出了 AIC 值、BIC 值、模型参数维数和负对数联合似然函数值, 比较了不同 m 和 r 的 GMM 对该数据的拟合效果. 在 $m = 2, 3, 5, 6, 8$ 的情形下, 只列出了各 19 个模型中 BIC 最小的模型评价; 在 $m = 4$ 的情形下, 除了讨论 BIC 最小的模型, 还显示了 $r = 16$ 的模型评价, 进一步分析与 K 均值做对比的模型, 可见该模型并不是此多项式阶数下的最优模型, 由其得到的分析结果并不可靠; 在该数据的所有模型搜寻中, 最优模型来自 $m = 7$ 的情形, 所以表 3 选列了该情形下最优的 5 个模型评价. 可见, 最优模型的 $m = 7, r = 14$, 模型参数维数为 114, 负对数联合似然函数值最小为 3,072.013, 相应的 AIC 和 BIC 值分别为 6,372.027 和 6,875.906, 是该模型评价中最小的. 接下来, 用最优模型来拟合分析酵母基因表达数据.

将 $B^{(0)}$ 、 $\pi^{(0)}$ 和 $\Sigma^{(0)}$ 代入 EM 算法开始计算极大似然估计, 迭代停止时得到的参数估计值为

$$\hat{B} = \begin{bmatrix} -20.789 & 3.067 & -7.884 & 33.833 & -2.077 & -42.921 & 12.768 & -2.883 & 13.359 & -32.186 & -9.064 & -1.216 & 16.057 & -24.008 \\ 48.752 & -6.514 & 20.314 & -70.785 & 4.632 & 93.328 & -29.939 & 7.781 & -31.892 & 79.598 & 21.410 & 0.442 & -40.367 & 54.586 \\ -42.468 & 5.101 & -21.079 & 50.127 & -3.870 & -74.045 & 26.419 & -9.143 & 28.727 & -75.620 & -18.989 & 1.760 & 32.367 & -45.553 \\ 18.007 & -1.829 & 10.533 & -16.536 & 1.596 & 28.733 & -11.319 & 5.048 & -12.587 & 34.731 & 8.230 & -1.682 & -11.942 & 18.389 \\ -3.967 & 0.305 & -2.664 & 2.756 & -0.340 & -5.877 & 2.513 & -1.367 & 2.861 & -8.191 & -1.856 & 0.585 & 2.230 & -3.867 \\ 0.436 & -0.022 & 0.327 & -0.222 & 0.036 & 0.608 & -0.278 & 0.175 & -0.324 & 0.950 & 0.209 & -0.086 & -0.204 & 0.406 \\ -0.019 & 0.000 & -0.015 & 0.007 & -0.002 & -0.025 & 0.012 & -0.009 & 0.014 & -0.043 & -0.009 & 0.005 & 0.007 & -0.017 \end{bmatrix}, \quad (5.1)$$

$$\hat{\pi} = [0.065, 0.021, 0.043, 0.035, 0.159, 0.014, 0.229, 0.049, 0.034, 0.027, 0.287, 0.009, 0.012, 0.016], \quad (5.2)$$

$$\hat{\sigma}^2 = 0.172, \quad (5.3)$$

$$\hat{\rho} = 0.447. \quad (5.4)$$

利用 MAP 计算模型中组别阵预测 $\hat{Z}$, 即对每个纵向观测个体进行聚类. 将聚类结果绘图展示, 图 5(a) 是最优模型的分折线图; 图 5(b) 是 $m = 7$ 和 $r = 14$ 的最优模型对各类基因序列的拟合曲

表 3 混合增长曲线模型基因聚类维数选取

m	r	AIC	BIC	Par.Num.	$-\loglik.$
2	4	8739.825	8801.705	14	4355.913
3	4	8736.518	8816.078	18	4350.259
4	5	8254.923	8374.262	27	4100.461
4	16	8580.622	8943.062	82	4208.311
5	5	8061.394	8202.834	32	3998.697
6	7	7164.626	7390.046	51	3531.313
7	12	6473.150	6906.310	98	3138.575
7	13	6427.471	6895.990	106	3107.736
7	14	6372.027	6875.906	114	3072.013
7	15	6388.574	6927.813	122	3072.287
7	16	6372.165	6946.765	130	3056.083
8	5	8233.294	8441.033	47	4069.647

图 5 (网络版彩图) GCM-SCS 基因表达水平聚类. (a) 分类折线图; (b) 拟合曲线图 (类别容量: $n_1 = 280$, $n_2 = 84$, $n_3 = 175$, $n_4 = 147$, $n_5 = 679$, $n_6 = 63$, $n_7 = 1001$, $n_8 = 196$, $n_9 = 133$, $n_{10} = 119$, $n_{11} = 1260$, $n_{12} = 42$, $n_{13} = 49$, $n_{14} = 70$)

线图, 比图 4(a) 用 $m = 4$ 和 $r = 16$ 的粗略模型拟合得更为精确, 图 5(b) 很好地拟合了图 5(a) 的类别趋势. 可见, 14 类基因表达水平序列的形状有明显差异, 同类别内的序列无论从形状还是变化尺度都非常相近. 该分类鲜明地描绘了酵母基因在双氧转换中的表达功能样式, 能有效地区分不同功能的酵母基因. 同时, 根据 (5.4) 可知, 酵母在新陈代谢的双氧转换过程中基因表达水平具有一般程度的序列相关性, 自相关系数为 0.447.

6 结论与讨论

本文首先基于一般增长曲线模型中组别阵的特性提出了一种全新的混合增长曲线模型, 进一步在新模型中讨论了协方差阵具有特定结构的情形, 针对协方差阵为纵向数据最为常见的序列相关结构进行了深入的研究. 然后, 用 EM 算法顺利获得了新模型的参数极大似然估计, 用 MAP 实现了序列相

关个体的聚类, 用 AIC 或 BIC 搜寻出最优模型. 接着, 通过大量的模拟实验验证了提出的方法能进行准确有效的拟合和估计. 最后, 由一个酵母基因表达数据的实证分析, 展示了新方法在考虑序列相关性后能得到合理而独特的分析结果.

本文合理地构建了统计模型, 并充分地提取了纵向数据信息: 第一, 混合增长曲线模型为实际数据中组别信息无法获知的问题提供了解决方案, 并将数据分布假设从正态扩展为混合正态, 很大程度上释放了严苛的分布假定. 第二, 于混合增长曲线模型中加入协方差结构的重要性不容忽视, 它能在本就具有序列相关性的纵向数据中捕获更加充分的信息, 得到正确的数据分析结果. 第三, 具有序列相关结构的混合增长曲线模型同时实现了聚类与拟合, 为相关数据聚类提供了新思路.

具有序列相关结构的混合增长曲线模型具有良好的性质, 但是它也存在一定的局限性有待继续探讨, 例如, 如何将模型可处理的数据集从平衡拓展到非平衡情形, 此问题在纵向数据中经常出现; 又如是否可以引入半参数或非参数的方法讨论数据拟合, 线性多项式的拟合对局部信息的捕捉不够灵敏等等, 相信这些问题将在后续研究中陆续被解决.

致谢 非常感谢审稿人对本文提出珍贵的修改意见, 这使修改后的论文更加严谨科学.

参考文献

- 1 Potthoff R F, Roy S N. A generalized multivariate analysis of variance model useful especially for growth curve problems. *Biometrika*, 1964, 51: 313–326
- 2 Pan J X, Fang K T. *Growth Curve Models and Statistical Diagnostics*. New York: Springer, 2002
- 3 Lee J C. Prediction and estimation of growth curves with special covariance structures. *J Amer Statist Assoc*, 1988, 83: 432–440
- 4 倪明明. 组别设计阵未知的增长曲线模型. 硕士学位论文. 昆明: 云南财经大学, 2016
- 5 McLachlan G J, Krishnan T. *The EM Algorithm and Extensions*, 2nd ed. Hoboken: John Wiley & Sons, 2007
- 6 潘雅婷. 纵向数据的混合增长曲线模型. 博士学位论文. 昆明: 云南财经大学, 2018
- 7 Dempster A P, Laird N M, Rubin D B. Maximum likelihood from incomplete data via the EM algorithm. *J R Stat Soc Ser B Stat Methodol*, 1977, 39: 1–22
- 8 Murphy K P. *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012
- 9 Siddiqui M M. On the inversion of the sample covariance matrix in a stationary autoregressive process. *Ann Math Statist*, 1958, 29: 585–588
- 10 Louis T A. Finding the observed information matrix when using the EM algorithm. *J R Stat Soc Ser B Stat Methodol*, 1982, 44: 226–233
- 11 Efron B. Bootstrap methods: Another look at the jackknife. *Ann Statist*, 1979, 7: 1–26
- 12 Wu C F J. On the convergence properties of the EM algorithm. *Ann Statist*, 1983, 11: 95–103
- 13 DeRisi J L, Iyer V R, Brown P O. Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science*, 1997, 278: 680–686

附录 A EM 算法的辅助函数

由模型 (2.1) 可以写出 $\mathbf{Y}$ 和 $\mathbf{Z}$ 的联合概率密度函数, 其对数形式如下:

$$\begin{aligned}
 l(\boldsymbol{\Theta}) &= \log f(\mathbf{Y}, \mathbf{Z}; \boldsymbol{\Theta}) \\
 &= \log \left(\prod_{j=1}^n f(\mathbf{y}_j, \mathbf{z}_j; \boldsymbol{\Theta}) \right) \\
 &= \sum_{j=1}^n \log(f(\mathbf{y}_j | \mathbf{z}_j; \boldsymbol{\Theta}) f(\mathbf{z}_j; \boldsymbol{\Theta}))
 \end{aligned}$$

$$\begin{aligned}
&= \sum_{j=1}^n \log(f(\mathbf{y}_j | \mathbf{z}_j; \boldsymbol{\Theta})) + \sum_{j=1}^n \log f(\mathbf{z}_j; \boldsymbol{\Theta}) \\
&= \sum_{j=1}^n \log \left\{ (2\pi)^{-\frac{p}{2}} |\sigma^2 \boldsymbol{\Gamma}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2\sigma^2} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j)^T \boldsymbol{\Gamma}^{-1} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j) \right] \right\} + \sum_{j=1}^n \log \left(\prod_{k=1}^r \pi_k^{z_{kj}} \right) \\
&= \frac{1}{2} \sum_{j=1}^n \left\{ -p \log(2\pi) + \log |\sigma^{-2} \boldsymbol{\Gamma}^{-1}| - \frac{1}{\sigma^2} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j)^T \boldsymbol{\Gamma}^{-1} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j) \right\} \\
&\quad + \sum_{j=1}^n \sum_{k=1}^r z_{kj} \log \pi_k.
\end{aligned} \tag{A.1}$$

E 步在给定已知信息和上一次迭代得到的参数值条件下, 对数联合似然函数关于缺失信息求期望, 即 $E[L(\boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}]$, 其中 $\boldsymbol{\Theta}^{(t-1)}$ 表示第 $t-1$ 步的参数 $\boldsymbol{\Theta}$ 估计值,

$$\begin{aligned}
&E[L(\boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}] \\
&= \frac{1}{2} E \left\{ \sum_{j=1}^n \left[\log |\sigma^{-2} \boldsymbol{\Gamma}^{-1}| - \frac{1}{\sigma^2} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j)^T \boldsymbol{\Gamma}^{-1} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j) \right] \middle| \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)} \right\} \\
&\quad + \sum_{j=1}^n \sum_{k=1}^r u_{kj}^{(t)} \log \pi_k,
\end{aligned} \tag{A.2}$$

其中 $u_{kj}^{(t)} = E(z_{kj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$, 而

$$\begin{aligned}
&E\{(\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j)^T \boldsymbol{\Gamma}^{-1} (\mathbf{y}_j - \mathbf{X} \mathbf{B} \mathbf{z}_j) | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}\} \\
&= E\{[\mathbf{y}_j - \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) + \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)}]^T \\
&\quad \cdot \boldsymbol{\Gamma}^{-1} [\mathbf{y}_j - \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) + \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)}]\} \\
&= [\mathbf{y}_j - \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})]^T \boldsymbol{\Gamma}^{-1} [\mathbf{y}_j - \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})] \\
&\quad + \text{tr}\{E[(\mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)})^T \boldsymbol{\Gamma}^{-1} (\mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
&\quad - \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)})] | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}\}.
\end{aligned} \tag{A.3}$$

又令

$$\boldsymbol{\Psi}_j = \frac{1}{\sigma^2} \boldsymbol{\nu}_j^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\nu}_j + \frac{1}{\sigma^2} \text{tr}[E(\boldsymbol{\varsigma}_j^T \boldsymbol{\Gamma}^{-1} \boldsymbol{\varsigma}_j) | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}], \tag{A.4}$$

其中 $\boldsymbol{\nu}_j = \mathbf{y}_j - \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$, $\boldsymbol{\varsigma}_j = \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)}$, 而 $(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$ 表示给定 $\mathbf{y}_j$ 和 $\boldsymbol{\Theta}^{(t-1)}$ 后第 t 步 $\mathbf{z}_j$ 的预测. 将回归系数矩阵 $\mathbf{B}$ 按列向量拆分, 再将向量 $E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$ 和 $\mathbf{z}_j^{(t)}$ 按元素拆分, 可得下面乘式:

$$\begin{cases} \mathbf{X} \mathbf{B} E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = \mathbf{X}(\beta_1, \dots, \beta_r) E(\mathbf{z}_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = \sum_{k=1}^r \mathbf{X} \beta_k u_{kj}^{(t)}, \\ \mathbf{X} \mathbf{B} \mathbf{z}_j^{(t)} = \mathbf{X}(\beta_1, \beta_2, \dots, \beta_r) \mathbf{z}_j^{(t)} = \sum_{k=1}^r \mathbf{X} \beta_k z_{kj}^{(t)}. \end{cases} \tag{A.5}$$

将上式代入 $\boldsymbol{\Psi}_j$ 可得

$$\boldsymbol{\Psi}_j = \frac{1}{\sigma^2} \left[\mathbf{y}_j - \sum_{k=1}^r \mathbf{X} \beta_k u_{kj}^{(t)} \right]^T \boldsymbol{\Gamma}^{-1} \left[\mathbf{y}_j - \sum_{k=1}^r \mathbf{X} \beta_k u_{kj}^{(t)} \right]$$

$$\begin{aligned}
 & + \frac{1}{\sigma^2} \text{tr} \left\{ \Gamma^{-1} \mathbf{X} \left[\sum_{k=1}^r \sum_{l=1}^r \beta_k \beta_l^T \mathbb{E}(u_{kj}^{(t)} - z_{kj}^{(t)})(u_{lj}^{(t)} - z_{lj}^{(t)})^T \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)} \right] \mathbf{X}^T \right\} \\
 & = \frac{1}{\sigma^2} \mathbf{y}_j^T \Gamma^{-1} \left(\mathbf{y}_j - 2 \sum_{k=1}^r \mathbf{X} \beta_k u_{kj}^{(t)} \right) \\
 & + \frac{1}{\sigma^2} \text{tr} \left\{ \sum_{k=1}^r \sum_{l=1}^r \Gamma^{-1} \mathbf{X} \beta_k \beta_l^T \mathbb{E}[(u_{kj}^{(t)} - z_{kj}^{(t)})(u_{lj}^{(t)} - z_{lj}^{(t)})^T \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)} \right. \\
 & \left. + u_{kj}^{(t)} u_{lj}^{(t)} \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}] \mathbf{X}^T \right\}, \tag{A.6}
 \end{aligned}$$

其中令

$$r_{jkl} = \mathbb{E}[(u_{kj}^{(t)} - z_{kj}^{(t)})(u_{lj}^{(t)} - z_{lj}^{(t)})^T \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)} + u_{kj}^{(t)} u_{lj}^{(t)} \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}], \tag{A.7}$$

因此可以得到 (A.2) 的变换形式

$$\mathbb{E}[L(\boldsymbol{\Theta}) \mid \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}] = \frac{1}{2} \sum_{j=1}^n \{\log |\sigma^{-2} \Gamma^{-1}| - \boldsymbol{\Psi}_j\} + \sum_{j=1}^n \sum_{k=1}^r u_{kj}^{(t)} \log \pi_k. \tag{A.8}$$

因为 $z_{kj}^{(t)}$ 为分类分布的元素, 所以 $z_{kj}^{(t)}$ 的取值非 0 即 1, 可得

$$\begin{aligned}
 u_{kj}^{(t)} & = \mathbb{E}(z_{kj}^{(t)} \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
 & = 1 \times \mathbb{P}(z_{kj}^{(t)} = 1 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) + 0 \times \mathbb{P}(z_{kj}^{(t)} = 0 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
 & = \mathbb{P}(z_{kj}^{(t)} = 1 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
 & = \frac{\mathbb{P}(\mathbf{y}_j \mid z_{kj}^{(t)} = 1, \boldsymbol{\Theta}^{(t-1)}) \mathbb{P}(z_{kj}^{(t)} = 1 \mid \boldsymbol{\Theta}^{(t-1)})}{\mathbb{P}(\mathbf{y}_j \mid \boldsymbol{\Theta}^{(t-1)})} \\
 & = \frac{\pi_k^{(t-1)} \mathbb{P}(\mathbf{y}_j \mid z_{kj}^{(t)} = 1, \boldsymbol{\Theta}^{(t-1)})}{\sum_{k=1}^r \pi_k^{(t-1)} \mathbb{P}(\mathbf{y}_j \mid z_{kj}^{(t)} = 1, \boldsymbol{\Theta}^{(t-1)})}. \tag{A.9}
 \end{aligned}$$

同样由 z_{kj} 非 0 即 1, 可知 $z_{kj}^{(t)} = 1$ 等价于列向量 $\mathbf{z}_j^{(t)}$ 的第 k 个元素为 1, 其余元素为 0, 则

$$\mathbb{P}(\mathbf{y}_j \mid z_{kj}^{(t)} = 1, \boldsymbol{\Theta}^{(t-1)}) = (2\pi)^{-\frac{p}{2}} |(\sigma^2 \Gamma)^{(t-1)}|^{-\frac{1}{2}} \exp \left[-\frac{1}{2} (\mathbf{v}_{kj}^{(t-1)})^T ((\sigma^2 \Gamma)^{(t-1)})^{-1} \mathbf{v}_{kj}^{(t-1)} \right], \tag{A.10}$$

其中 $\mathbf{v}_{kj}^{(t-1)} = \mathbf{y}_j - \mathbf{X} \beta_k^{(t-1)}$. 通过 (A.10) 就可以计算出 (A.9) 的具体数值.

同样因为 $z_{kj}^{(t)}$ 非 0 即 1, 于是 (A.7) 分情形讨论如下: 若 $k = l$ 时,

$$\begin{aligned}
 r_{jkl} & = \mathbb{E}((z_{kj}^{(t)})^2 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - (u_{kj}^{(t)})^2 + (u_{kj}^{(t)})^2 \\
 & = 1 \times \mathbb{P}(z_{kj}^{(t)} = 1 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) + 0 \times \mathbb{P}(z_{kj}^{(t)} = 0 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
 & = \mathbb{P}(z_{kj}^{(t)} = 1 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) \\
 & = \mathbb{E}(z_{kj}^{(t)} = 1 \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}); \tag{A.11}
 \end{aligned}$$

若 $k \neq l$ 时,

$$r_{jkl} = \mathbb{E}(z_{kj}^{(t)} z_{lj}^{(t)} \mid \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) - u_{kj}^{(t)} u_{lj}^{(t)} + u_{kj}^{(t)} u_{lj}^{(t)}$$

$$= E(z_{kj}^{(t)} z_{lj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}). \quad (\text{A.12})$$

根据分类分布中第 j 个样本只能属于其中一个类别, 可知 z_{kj} 与 z_{lj} 至少有一个为 0, 所以在 $k \neq l$ 时, $E(z_{kj}^{(t)} z_{lj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = (z_{kj}^{(t)} z_{lj}^{(t)}) P(z_{kj}^{(t)} z_{lj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = 0$.

先将 (A.11) 和 (A.12) 代入 (A.7), 同时将 (A.10) 代入 (A.9), 再将 (A.7) 和 (A.9) 代入 (A.6), 最后将 (A.6) 和 (A.9) 代入 (A.8) 可得

$$E[L(\boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}] = \frac{1}{2} \sum_{j=1}^n \left\{ \log |\sigma^{-2} \boldsymbol{\Gamma}^{-1}| - \frac{1}{\sigma^2} \mathbf{y}_j^T \boldsymbol{\Gamma}^{-1} \left(\mathbf{y}_j - 2 \sum_{k=1}^r \mathbf{X} \boldsymbol{\beta}_k u_{kj}^{(t)} \right) - \frac{1}{\sigma^2} \text{tr} \left[\sum_{k=1}^r \boldsymbol{\Gamma}^{-1} \mathbf{X} \boldsymbol{\beta}_k \boldsymbol{\beta}_k^T u_{kj}^{(t)} \mathbf{X}^T \right] \right\} + \sum_{j=1}^n \sum_{k=1}^r u_{kj}^{(t)} \log \pi_k. \quad (\text{A.13})$$

因为 $E[L(\boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}]$ 可以表示为参数 $\boldsymbol{\Theta}^{(t-1)}$ 给定时关于 $\boldsymbol{\Theta}$ 的函数, 所以接下来记辅助函数 $Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) = E[L(\boldsymbol{\Theta}) | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}]$.

附录 B 求 π 的估计

在给定 $\boldsymbol{\Theta}^{(t-1)}$ 和 $\mathbf{Y}$ 的条件下求 π 的估计, 求 $Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)})$ 关于 π_k 的最值相当于优化求解. 发现直接优化求解 $\pi_r = 1 - \sum_{k=1}^{r-1} \pi_k$ 极为复杂, 而它等价于带约束条件 “ $\sum_{k=1}^r \pi_k - 1 = 0$ ” 的 π , 所以采用 Lagrange 乘子法. 令

$$\begin{cases} \frac{\partial [Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) + \lambda(\sum_{k=1}^r \pi_k - 1)]}{\partial \pi_k} = 0, \\ \sum_{k=1}^r \pi_k - 1 = 0 \end{cases} \Rightarrow \begin{cases} \lambda = -n, \\ \pi_k^{(t)} = \frac{1}{n} \sum_{j=1}^n u_{kj}^{(t)}, \end{cases} \quad (\text{B.1})$$

$k = 1, \dots, r$. 将 (A.9) 代入 (B.1) 第二式, 可以计算得后验概率 $\pi_k^{(t)}$ ($k = 1, \dots, r$). 又由 $u_{kj}^{(t)} = E(z_{kj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$ 可知 $\pi_k^{(t)}$ 为 $\boldsymbol{\Theta}^{(t-1)}$ 的函数, 即 $\pi_k^{(t)} = f(\boldsymbol{\Theta}^{(t-1)})$. 将各组别的后验概率 $\pi_k^{(t)}$ 组合成完整的分类分布参数 $\boldsymbol{\pi}^{(t)}$ 有如下形式:

$$\boldsymbol{\pi}^{(t)} = (\pi_1^{(t)}, \pi_2^{(t)}, \dots, \pi_r^{(t)}) = \frac{1}{n} \mathbf{1}_n^T \begin{bmatrix} u_{11}^{(t)} & u_{21}^{(t)} & \cdots & u_{r1}^{(t)} \\ u_{12}^{(t)} & u_{22}^{(t)} & \cdots & u_{r2}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n}^{(t)} & u_{2n}^{(t)} & \cdots & u_{rn}^{(t)} \end{bmatrix} = \frac{1}{n} \mathbf{1}_n^T E((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}), \quad (\text{B.2})$$

其中 $\mathbf{1}_n$ 为元素全为 1 的 n 维列向量, $E((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}) = [E(z_1^{(t)} | \mathbf{y}_1, \boldsymbol{\Theta}^{(t-1)}), \dots, E(z_n^{(t)} | \mathbf{y}_n, \boldsymbol{\Theta}^{(t-1)})]^T$, 而 $E(z_j^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)}) = [u_{1j}^{(t)}, \dots, u_{rj}^{(t)}]^T$, $j = 1, \dots, n$.

附录 C 求 B 的估计

在给定 $\boldsymbol{\Theta}^{(t-1)}$ 和 $\mathbf{Y}$ 的条件下, 求 $B^{(t)}$ 的估计, 若 $\boldsymbol{\Sigma} > \mathbf{0}$, $\sigma^2 > 0$, $\boldsymbol{\Gamma} > \mathbf{0}$ 和 $\text{rank}(\mathbf{X}^T \mathbf{X}) = \text{rank}(\mathbf{X} \mathbf{X}^T) = \text{rank}(\mathbf{X}) = m$, 则令 $\partial Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)}) / \partial \boldsymbol{\beta}_k = 0$, 可得

$$\mathbf{X}^T \boldsymbol{\Gamma}^{-1} \left[\sum_{j=1}^n \mathbf{y}_j u_{kj}^{(t)} - \sum_{j=1}^n u_{kj}^{(t)} \mathbf{X} \boldsymbol{\beta}_k \right] = 0. \quad (\text{C.1})$$

由 $\text{rank}(\mathbf{X}^T \mathbf{\Gamma}^{-1} \mathbf{X}) = m$, 且 $\mathbf{X}^T \mathbf{\Gamma}^{-1} \mathbf{X}$ 为一个 $m \times m$ 的矩阵, 则 $\mathbf{X}^T \mathbf{\Gamma}^{-1} \mathbf{X}$ 可逆, 可得第 t 步的 β_k 的估计

$$\begin{aligned} \beta_k^{(t)} &= \left[\sum_{j=1}^n u_{kj}^{(t)} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X} \right]^{-1} \left[\sum_{j=1}^n \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{y}_j u_{kj}^{(t)} \right] \\ &= [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \frac{\sum_{j=1}^n \mathbf{y}_j u_{kj}^{(t)}}{\sum_{j=1}^n u_{kj}^{(t)}} \\ &= [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} [\mathbf{y}_1, \mathbf{y}_2, \dots, \mathbf{y}_n] \left[\frac{u_{k1}^{(t)}}{\sum_{j=1}^n u_{kj}^{(t)}}, \frac{u_{k2}^{(t)}}{\sum_{j=1}^n u_{kj}^{(t)}}, \dots, \frac{u_{kn}^{(t)}}{\sum_{j=1}^n u_{kj}^{(t)}} \right]^T \\ &= [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{Y} \boldsymbol{\eta}_k, \end{aligned} \quad (\text{C.2})$$

其中 $\boldsymbol{\eta}_k = [u_{k1}^{(t)} / \sum_{j=1}^n u_{kj}^{(t)}, u_{k2}^{(t)} / \sum_{j=1}^n u_{kj}^{(t)}, \dots, u_{kn}^{(t)} / \sum_{j=1}^n u_{kj}^{(t)}]^T$.

又注意到

$$\begin{aligned} \boldsymbol{\eta} &= [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_r] \\ &= \begin{bmatrix} u_{11}^{(t)} & u_{21}^{(t)} & \cdots & u_{r1}^{(t)} \\ u_{12}^{(t)} & u_{22}^{(t)} & \cdots & u_{r2}^{(t)} \\ \vdots & \vdots & \ddots & \vdots \\ u_{1n}^{(t)} & u_{2n}^{(t)} & \cdots & u_{rn}^{(t)} \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n u_{kj}^{(t)} & 0 & \cdots & 0 \\ 0 & \sum_{j=1}^n u_{kj}^{(t)} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sum_{j=1}^n u_{kj}^{(t)} \end{bmatrix}^{-1} \\ &= \mathbf{E}((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}) \mathbf{T}, \end{aligned} \quad (\text{C.3})$$

其中 $\mathbf{T} = [\text{diag}(\sum_{j=1}^n u_{1j}^{(t)}, \sum_{j=1}^n u_{2j}^{(t)}, \dots, \sum_{j=1}^n u_{rj}^{(t)})]^{-1}$. 于是第 t 步的 $\mathbf{B}$ 的估计为

$$\begin{aligned} \mathbf{B}^{(t)} &= (\beta_1^{(t)}, \beta_2^{(t)}, \dots, \beta_r^{(t)}) \\ &= [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{Y} [\boldsymbol{\eta}_1, \boldsymbol{\eta}_2, \dots, \boldsymbol{\eta}_r] \\ &= [\mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{\Gamma}^{(t-1)})^{-1} \mathbf{Y} \mathbf{E}((\mathbf{Z}^{(t)})^T | \mathbf{Y}, \boldsymbol{\Theta}^{(t-1)}) \mathbf{T}, \end{aligned} \quad (\text{C.4})$$

其中 $u_{kj}^{(t)}$ 由 (A.9) 获得, 由 $u_{kj}^{(t)} = \mathbf{E}(z_{kj}^{(t)} | \mathbf{y}_j, \boldsymbol{\Theta}^{(t-1)})$ 和 $\mathbf{\Gamma}^{(t-1)}$ 可知 $\beta_k^{(t)}$ 为 $\boldsymbol{\Theta}^{(t-1)}$ 的函数, 即 $\beta_k^{(t)} = f(\boldsymbol{\Theta}^{(t-1)})$.

附录 D 求 σ^2 的估计

对辅助函数 (A.13) 关于方差 σ^2 优化求解,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\Theta} | \boldsymbol{\Theta}^{(t-1)})}{\partial \sigma^2} &= \frac{\partial \sum_{j=1}^n \frac{1}{2} \{ \log |\frac{1}{\sigma^2} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1}| - \text{tr}[\frac{1}{\sigma^2} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1} \sum_{k=1}^r \mathbf{v}_{kj}^{(t)} (\mathbf{v}_{kj}^{(t)})^T u_{kj}^{(t)}] \}}{\partial \sigma^2} \\ &= \frac{\partial \frac{1}{2} \sum_{j=1}^n \{ \log |\frac{1}{\sigma^2} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1}| - \text{tr}[\frac{1}{\sigma^2} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1} \mathbf{S}_j^{(t)}] \}}{\partial \sigma^2} \\ &= \frac{1}{2} \sum_{j=1}^n \text{tr} \left\{ -\sigma^2 \mathbf{\Gamma}_{\rho^{(t-1)}} \left(\frac{1}{\sigma^4} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1} \right) + \mathbf{S}_j^{(t)} \left(\frac{1}{\sigma^4} \mathbf{\Gamma}_{\rho^{(t-1)}}^{-1} \right) \right\}, \end{aligned} \quad (\text{D.1})$$

其中 $\mathbf{S}_j^{(t)} = \sum_{k=1}^r \mathbf{v}_{kj}^{(t-1)} (\mathbf{v}_{kj}^{(t-1)})^T \mathbf{u}_{kj}^{(t)}$. 令 $\partial Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t-1)}) / \partial \sigma^2 = 0$, 记 $h_1^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{I}$, $h_2^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{M}$, $h_3^{(t)} = \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{K}$, 则有

$$\begin{aligned} (\sigma^2)^{(t)} &= \frac{1}{np} \sum_{j=1}^n \text{tr}(\mathbf{S}_j^{(t)} \boldsymbol{\Gamma}_{\rho^{(t-1)}}^{-1}) \\ &= \frac{1}{np} \sum_{j=1}^n \text{tr} \left(\mathbf{S}_j^{(t)} \left(\frac{1 + (\rho^{(t-1)})^2}{1 - (\rho^{(t-1)})^2} \mathbf{I} - \frac{\rho^{(t-1)}}{1 - (\rho^{(t-1)})^2} \mathbf{M} - \frac{(\rho^{(t-1)})^2}{1 - (\rho^{(t-1)})^2} \mathbf{K} \right) \right) \\ &= \frac{[1 + (\rho^{(t-1)})^2] h_1^{(t)} - \rho^{(t-1)} h_2^{(t)} - (\rho^{(t-1)})^2 h_3^{(t)}}{[1 - (\rho^{(t-1)})^2] np}. \end{aligned} \quad (\text{D.2})$$

由 $h_1^{(t)}$ 、 $h_2^{(t)}$ 、 $h_3^{(t)}$ 和 $\rho^{(t-1)}$ 的表达式可知, $(\sigma^2)^{(t)}$ 为关于 $\boldsymbol{\Theta}^{(t-1)}$ 的函数, 即 $(\sigma^2)^{(t)} = f(\boldsymbol{\Theta}^{(t-1)})$.

附录 E 求 ρ 的估计

对辅助函数 (A.13) 关于自相关系数 ρ 优化求解,

$$\begin{aligned} \frac{\partial Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t-1)})}{\partial \rho} &= \frac{\partial \sum_{j=1}^n \frac{1}{2} \{ \log | \frac{1}{(\sigma^2)^{(t-1)}} \boldsymbol{\Gamma}_{\rho}^{-1} | - \text{tr} [\frac{1}{(\sigma^2)^{(t-1)}} \boldsymbol{\Gamma}_{\rho}^{-1} \mathbf{S}_j^{(t)}] \}}{\partial \rho} \\ &= \sum_{j=1}^n \frac{1}{2} \text{tr} \left\{ \boldsymbol{\Gamma}_{\rho} \frac{\partial \boldsymbol{\Gamma}_{\rho}^{-1}}{\partial \rho} - \frac{1}{(\sigma^2)^{(t-1)}} \mathbf{S}_j^{(t)} \frac{\partial \boldsymbol{\Gamma}_{\rho}^{-1}}{\partial \rho} \right\} \\ &= \sum_{j=1}^n \frac{1}{2(\sigma^2)^{(t-1)}} \text{tr} \left\{ ((\sigma^2)^{(t-1)} \boldsymbol{\Gamma}_{\rho} - \mathbf{S}_j^{(t)}) \frac{\partial (\frac{1+\rho^2}{1-\rho^2} \mathbf{I} - \frac{\rho}{1-\rho^2} \mathbf{M} - \frac{\rho^2}{1-\rho^2} \mathbf{K})}{\partial \rho} \right\} \\ &= \sum_{j=1}^n \frac{1}{2(\sigma^2)^{(t-1)}} \text{tr} \left\{ ((\sigma^2)^{(t-1)} \boldsymbol{\Gamma}_{\rho} - \mathbf{S}_j^{(t)}) \left(\frac{4\rho}{(1-\rho^2)^2} \mathbf{I} - \frac{\rho^2+1}{(1-\rho^2)^2} \mathbf{M} - \frac{2\rho}{(1-\rho^2)^2} \mathbf{K} \right) \right\}. \end{aligned}$$

令 $\partial Q(\boldsymbol{\Theta} \mid \boldsymbol{\Theta}^{(t-1)}) / \partial \rho = 0$, 可得

$$\text{tr}[(\sigma^2)^{(t-1)} \boldsymbol{\Gamma}_{\rho} (4\rho \mathbf{I} - (\rho^2 + 1) \mathbf{M} - 2\rho \mathbf{K})] = \frac{1}{n} \sum_{j=1}^n \text{tr}[\mathbf{S}_j^{(t)} (4\rho \mathbf{I} - (\rho^2 + 1) \mathbf{M} - 2\rho \mathbf{K})], \quad (\text{E.1})$$

$$f(\rho) = -2(n(p-1)(\sigma^2)^{(t-1)})\rho^3 + h_2^{(t)}\rho^2 + (2n(p-1)(\sigma^2)^{(t-1)} - 4h_1^{(t)} + 2h_3^{(t)})\rho + h_2^{(t)} = 0. \quad (\text{E.2})$$

根据 (E.2), 设 a 为三次项系数、 b 为二次项系数、 c 为一次项系数、 d 为常数项, 则

$$\begin{cases} a^{(t)} = -2n(p-1)(\sigma^2)^{(t-1)}, \\ b^{(t)} = d^{(t)} = h_2^{(t)}, \\ c^{(t)} = 2n(p-1)(\sigma^2)^{(t-1)} - 4h_1^{(t)} + 2h_3^{(t)}. \end{cases}$$

又由三次函数求根公式 (3.14) 可计算出 ρ 的解. 由 $h_1^{(t)}$ 、 $h_2^{(t)}$ 、 $h_3^{(t)}$ 和 $(\sigma^2)^{(t-1)}$ 的表达式可知, $\rho^{(t)}$ 为关于 $\boldsymbol{\Theta}^{(t-1)}$ 的函数, 即 $\rho^{(t)} = f(\boldsymbol{\Theta}^{(t-1)})$.

$\hat{\rho}$ 在定义范围内必有根 由 (E.1) 可得

$$\begin{cases} (\sigma^2)^{(t-1)} \text{tr} \Gamma_{\rho^{(t)}} \mathbf{I} = \frac{1}{n} \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{I}, \\ (\sigma^2)^{(t-1)} \text{tr} \Gamma_{\rho^{(t)}} \mathbf{M} = \frac{1}{n} \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{M}, \\ (\sigma^2)^{(t-1)} \text{tr} \Gamma_{\rho^{(t)}} \mathbf{K} = \frac{1}{n} \sum_{j=1}^n \text{tr} \mathbf{S}_j^{(t)} \mathbf{K}, \end{cases} \quad (\text{E.3})$$

由此可得 $h_1^{(t)} = np(\sigma^2)^{(t-1)}$, $h_2^{(t)} = 2n(p-1)(\sigma^2)^{(t-1)}\rho^{(t)}$ 和 $h_3^{(t)} = 2n(\sigma^2)^{(t-1)}$. 接下来, 通过计算 $f(\rho)$ 在 ρ 的定义区间端点的情况来判断 (E.2) 是否有根:

$$\begin{cases} \rho = 1, & f(\rho = 1) = 0, & f'(\rho = 1) = -4n(p-1)(\sigma^2)^{(t-1)} < 0, & \text{若 } p > 1, \\ & \Rightarrow \exists N_1 > 1, & f\left(1 - \frac{1}{N_1}\right) > 0, & \text{若 } p > 1; \\ \rho = -1, & f(\rho = -1) = 0, & f'(\rho = -1) = -12n(p-1)(\sigma^2)^{(t-1)} < 0, & \text{若 } p > 1, \\ & \Rightarrow \exists N_2 > 1, & f\left(-1 + \frac{1}{N_2}\right) < 0, & \text{若 } p > 1. \end{cases} \quad (\text{E.4})$$

由 (E.2) 计算 $f(\rho)$ 关于 ρ 的导数, 可得 $f'(\rho) = -6(n(p-1)(\sigma^2)^{(t-1)})\rho^2 + 2h_2^{(t)}\rho + 2n(p-1)(\sigma^2)^{(t-1)} - 4h_1^{(t)} + 2h_3^{(t)}$. 然后, 在定义区间讨论三次方程解的情形如上.

因为 f 为连续函数, 又由 (E.4) 容易满足条件 $p > 1$, 可得当 $\rho \in (-1, 1)$ 时, f 至少有一次穿过零轴, 即 $f(\rho)$ 在 $(-1, 1)$ 上必有根.

Growth curves mixture model with serial covariance structure

Yating Pan, Yu Fei, Mingming Ni & Jianxin Pan

Abstract Growth curve models (GCMs) play an important role in longitudinal data analysis. However, a very crucial assumption is that the grouping design matrix is assumed to be known, which may not be true in practice. In this paper, we propose a novel mixture model which can handle the problem caused by the unknown grouping matrix. In the meantime, it approximates the true distribution of the response matrix using a mixture of normal distributions. The new model is studied under the serial covariance structure (SCS), which accounts for serial correlation among different observations within the same individual. The maximum likelihood estimation for the proposed model is considered by the use of EM (expectation maximization) algorithm, along with clustering individuals. A data-driven method is used to find a appropriate model. Simulation studies and real data analysis show that the proposed procedure works well in parameter estimation and individual clustering, simultaneously.

Keywords growth curve model, Gaussian mixture model, serial covariance structure, EM algorithm, clustering

MSC(2010) 62F99, 62H30, 62J99

doi: 10.1360/N012019-00145