

基于有限混合回归模型的各国基尼指数纵向聚类研究

史丁伊*, 余 静

重庆理工大学数学科学学院, 重庆

收稿日期: 2026年6月30日; 录用日期: 2026年7月21日; 发布日期: 2026年7月31日

摘 要

纵向数据的异质性特征在经济学与社会科学领域十分常见, 比如不同国家收入不平等的演化轨迹具有明显异质性, 对其进行聚类分析有助于揭示不同国家经济发展的潜在模式与异质性规律。在处理纵向数据的聚类分析中, 有限混合回归模型(Finite Mixture Regression Model, FMR)是一类核心工具。现有的混合建模主要包含两种思路: 传统有限混合回归模型(FMR-1)通常假设具有个体内部观测值相互独立, 仅对各子总体的均值进行建模。然而, 这种忽略组内协方差信息的做法易导致聚类偏差。针对此局限, 有学者提出了基于均值-协方差联合建模的混合回归模型(FMR-3), 其在正态混合框架下同时对均值与协方差结构进行建模。本文运用上述两种方法对36个国家20年间的基尼系数演化轨迹进行聚类分析, 两者均采用EM算法进行参数估计。结果表明, 基于均值-协方差联合建模的混合回归方法在聚类一致性、组内轨迹区分度以及模型拟合优度(BIC)上均显著优于传统均值模型; 而忽略组内相关性不仅容易导致聚类数目的高估, 还会造成不同演化轨迹的相互混杂。本研究证实了在实际宏观经济纵向数据聚类中纳入协方差结构的必要性, 为跨国收入不平等的动态演化研究提供了更稳健的分析工具。

关键词

纵向数据, 有限混合回归模型, 均值-协方差联合建模, 基尼系数

Longitudinal Clustering Analysis of National Gini Coefficients Based on Finite Mixture Regression Model

Dingyi Shi*, Jing Yu

School of Mathematical Sciences, Chongqing University of Technology, Chongqing

Received: June 30, 2026; accepted: July 21, 2026; published: July 31, 2026

*通讯作者。

文章引用: 史丁伊, 余静. 基于有限混合回归模型的各国基尼指数纵向聚类研究[J]. 统计学与应用, 2026, 15(8): 1-10.
DOI: 10.12677/sa.2026.158174

Abstract

Heterogeneity is a pervasive characteristic of longitudinal data in economics and social sciences. Specifically, the evolutionary trajectories of income inequality display pronounced cross-national heterogeneity, and cluster analysis of such trajectories facilitates the identification of underlying patterns of economic development and heterogeneous dynamics among nations. As a core instrument for longitudinal data clustering, finite mixture regression (FMR) models are widely adopted in the literature. Existing mixture modeling frameworks fall into two main categories: the conventional finite mixture regression model (denoted FMR-1) generally assumes independence among observations sharing the same trajectory pattern and only specifies the mean structure for each sub-population. Nevertheless, overlooking within-group covariance information in this manner is prone to clustering bias. To address this limitation, a mixture regression model with joint mean-covariance modeling (denoted FMR-3) has been developed, which jointly models both the mean and covariance structures within the finite normal mixture framework. Accordingly, this paper employs both approaches to conduct cluster analysis on the Gini coefficient trajectories of 36 countries over a 20-year horizon, with parameter estimation for both models implemented via the Expectation-Maximization (EM) algorithm. Empirical results indicate that the joint mean-covariance mixture regression model substantially outperforms the conventional mean-only model across three dimensions: clustering consistency, within-group trajectory discriminability, and model goodness-of-fit as measured by the Bayesian Information Criterion (BIC). Meanwhile, neglecting within-group correlation not only tends to overestimate the optimal number of clusters, but also causes the conflation of distinct evolutionary trajectories. This study corroborates the necessity of incorporating covariance structures when clustering real-world macroeconomic longitudinal data, and provides a more robust analytical framework for investigating the dynamic evolution of cross-national income inequality.

Keywords

Longitudinal Data, Finite Mixture Regression Model, Joint Mean-Covariance Modeling, Gini Index

Copyright © 2026 by author(s) and Hans Publishers Inc.

This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

1. 引言

收入不平等问题及其对宏观经济的深远影响, 始终是发展经济学者与政策制定者关注的核心议题。基尼系数(Gini index)作为衡量收入不平等程度的关键指标, 在跨国比较研究中被广泛应用。随着全球基尼系数追踪数据的不断积累, 学术界的关注点逐渐从单一时间点的截面比较转向长期纵向演化轨迹的刻画。由于不同国家在经济发展阶段、政策导向与社会结构上存在显著差异, 其基尼系数的动态演化模式呈现出高度的异质性。例如, 部分国家的收入不平等程度持续恶化, 部分国家趋于稳定或缓解, 还有国家则表现出非单调的波动。这种异质性表明, 将所有国家视为同质总体进行分析将掩盖重要的局部群体特征。因此, 对各国基尼系数的纵向演化轨迹进行聚类分析, 识别具备相似发展模式的国家群体, 是探究收入不平等内在机制与制定针对性干预政策的重要前提。

针对跨国收入不平等的异质性特征, 现有文献已进行了多维度的探讨。例如, Espoir 等人突破了传统收敛检验的局限, 采用基于回归的俱乐部收敛检验方法, 利用全球 142 个国家的纵向数据揭示了跨国

收入不平等的收敛规律,并实证检验了俱乐部形成的驱动因素[1]。Ogundari 等人针对非洲 34 国的研究同样证实了“俱乐部收敛”的存在,表明收入不平等演化缺乏统一的长期路径[2]。此外,Chrisendo 等人基于全球数据的最新研究指出,过去三十年间全球近半数人口所在国家的收入不平等程度加剧,而约三分之一则经历下降[3]。这些发现充分证实,各国基尼系数的演化轨迹绝非同质,需要更细致的纵向聚类分析。

有限混合模型常常用于异质性纵向数据的聚类。McLachlan 等人对有限混合模型进行了全面阐述,其基本假设是:数据来自若干具有不同概率分布的混合成分,每个成分对应一个聚类[4]。有限混合模型采用概率模型来描述异质性数据,并通过最大后验概率识别聚类。比如,对于连续的纵向数据,有限混合模型假设各成分来自正态分布,构建了用于连续数据聚类的高斯混合模型[5];对于计数数据,可假设各成分来自泊松分布,可构建用于计数数据聚类的混合模型,Zamani 等人系统地比较了泊松、混合泊松、广义泊松和有限泊松混合回归模型在处理离散计数数据时的表现,并指出有限泊松混合回归模型在拟合优度上可能最优[6]。针对分组清晰的异质性识别问题,Nagin 等人对群组轨迹模型进行了梳理,该模型假设同一聚类内个体具有相同变化轨迹,突破了传统纵向模型所有个体共享同一轨迹的约束[7]。Lyrval 等人对潜类别分析的前沿方法论进行了系统梳理[8],潜类别分析的核心在于通过可观测的数据特征,去推断并识别总体中存在的、无法直接观测的潜在异质性子群[9]。上述群组轨迹模型和潜在类别分析均需要对数据作分布假设,和有限混合模型具有相同的原理。

然而,近期的纵向数据聚类方法综述(如 Van der Nest 等人[10])指出,上述经典的纵向混合回归模型通常将关注焦点集中于各潜在子群体的均值结构。在实际建模中,这类方法往往隐含地假设组内观测值相互独立,或各混合成分共享同质且恒定的方差-协方差矩阵。Wu 等人指出,当真实的纵向数据存在显著的异方差性或复杂的自相关性时,忽略协方差结构的异质性极易引发均值参数的估计偏误,并导致聚类错误率上升[11]。为了克服仅关注均值结构的局限性,近年来学者们开始探索均值与协方差的联合建模。Yu 等人创新性地提出了基于均值-协方差联合建模的混合回归模型,在有限正态混合回归框架下同步对均值轨迹和协方差矩阵进行动态参数化建模,有效剥离了均值差异与波动差异[12][13]。这为进一步提升高波动纵向数据聚类的准确性与稳健性奠定了方法学基础。

本文将有限混合回归模型应用于跨国基尼系数的动态演化研究中,寻找其中的异质性。具体而言,本文利用 36 个国家跨越 20 年的纵向数据,实证对比了两种不同的混合回归方法:仅关注均值结构的传统有限混合回归模型(FMR-1),以及同步对均值与协方差结构进行动态建模的联合混合回归模型(FMR-3)。两种方法均采用期望最大化(EM)算法进行参数估计,并基于贝叶斯信息准则(BIC)确定最佳聚类数。对比结果表明,相较于传统均值模型,均值-协方差联合混合回归模型能够更精准地捕捉纵向数据的波动特征,有效克服了因忽略组内相关性而导致的聚类数目高估与轨迹模式混杂问题。本文的实证工作不仅清晰勾勒出各国收入不平等演化的异质性模式,也为宏观经济纵向数据的精准聚类提供了更为稳健的工具验证。

2. 模型与方法

2.1. 两种有限混合回归模型

假设 $\mathbf{y}_i = (y_{i1}, \dots, y_{im_i})^\top$ 表示个体 i 在 m_i 个时刻进行测量得到的测量值向量,其中 $i=1, \dots, N$ 。 y_{ij} 为第 i 个个体的第 j 个观测值,其中 $j=1, \dots, m_i$ 。令 $f(\mathbf{y}_i)$ 表示响应向量 $\mathbf{y}_i$ 的概率密度函数,假设它们来自包含 K 个群体的混合分布,则有 $f(\mathbf{y}_i) = \sum_{k=1}^K \pi_k f_k(\mathbf{y}_i)$,其中 π_k 为个体属于第 k 个子总体的混合比例,满

足 $\sum_{k=1}^K \pi_k = 1$ 且 $\pi_k \geq 0$; $f_k(y_i | \theta_k)$ 为第 k 类的概率密度函数, $k=1, \dots, K$; 令 Ψ 为所有待估参数构成的集合。

对于第 k 类中的个体 y_i , 传统均值有限混合回归模型(FMR-1) [14]假设如下线性结构:

$$y_i = X_i \beta_k + \varepsilon_{ik} \quad (1)$$

其中 X_i 为 $m_i \times p$ 的设计矩阵, β_k 为第 k 类的均值参数向量。FMR 模型的关键局限在于它对随机误差项的强假设: 通常假定 $\varepsilon_{ik} \sim N(0, \sigma_k^2 I_{m_i})$, 即给定协变量后, 个体的纵向观测值条件独立, 且协方差矩阵仅为对角阵 $\sigma_k^2 I_{m_i}$ 。这类仅考虑均值结构的 FMR-1 模型忽略了纵向数据内部普遍存在的异方差性和自相关性。此时, 第 k 类待估参数集合为 $\Psi_k = (\pi_k, \beta_k^\top, \sigma_k^2)^\top$ 。

为了更准确地捕捉纵向数据中的波动特征与组内相关性, 本文参考 Yu 等人的研究, 引入均值-协方差联合建模的混合回归模型(FMR-3)。该模型在保留式(1)均值建模的基础上, 对误差项 ε_{ik} 的协方差结构进行重新设定, 即假设 $\varepsilon_{ik} \sim N(0, \Sigma_{ik})$ 。此时, 第 k 类的多元正态密度函数为

$$f_k(y_i | \Psi_k) = (2\pi)^{-\frac{m_i}{2}} |\Sigma_{ik}|^{-\frac{1}{2}} \exp \left\{ -\frac{1}{2} (y_i - X_i \beta_k)^\top \Sigma_{ik}^{-1} (y_i - X_i \beta_k) \right\}$$

其中协方差矩阵 Σ_{ik} 不再是固定结构的对角阵, 而是依赖于时间等协变量的动态矩阵。为保证 Σ_{ik} 的正定性并实现参数降维, 采用 Pourahmadi [15]提出的修正 Cholesky 分解(MCD)对其进行参数化:

$$T_{ik} \Sigma_{ik} T_{ik}^\top = D_{ik}^2 \quad (2)$$

其中 T_{ik} 为对角线元素为 1 的下三角矩阵, 其非对角线元素为 $-\phi_{ikjl}$ ($1 \leq j, l \leq m_i$); D_{ik}^2 为对角线元素为 $\sigma_{ikj}^2 > 0$ 的对角矩阵。该分解具有清晰的统计解释: ϕ_{ikjl} 是自回归系数, 刻画了当前观测值对过去观测值的线性依赖关系; σ_{ikj}^2 是新息方差, 表示给定过去信息后当前观测值的条件方差。具体而言, 在给定前 $j-1$ 个观测值的条件下, 第 j 个观测值的条件预测为 $\hat{y}_{ikj} = \mu_{ikj} + \sum_{l=1}^{j-1} \phi_{ikjl} (y_{ikl} - \mu_{ikl})$, 且 $\sigma_{ikj}^2 = \text{Var}(y_{ikj} - \hat{y}_{ikj})$ 。进一步, 参考 Cepeda 和 Gamerman [16]以及 Pourahmadi [17]的建模思路, 将 μ_{ikj} 、 $\ln(\sigma_{ikj}^2)$ 和 ϕ_{ikjl} 分别建模为协变量的线性函数:

$$\mu_{ikj} = x_{ij}^\top \beta_k, \ln(\sigma_{ikj}^2) = z_{ij}^\top \lambda_k, \phi_{ikjl} = w_{ijl}^\top \gamma_k \quad (3)$$

其中 $\beta_k, \lambda_k, \gamma_k$ 分别为 p 维、 d 维和 q 维的参数向量, x_{ij}, z_{ij}, w_{ijl} 对应的协变量行向量。这里协变量可能包含一些基础协变量(Baseline Covariates), 时间的多项式和一些相关的交互项。比如当我们使用时间的多项式来对均值、对数新息方差和自回归系数建模时, 其协变量的形式为 $x_{ij} = (1, t_{ij}, t_{ij}^2, \dots, t_{ij}^{p-1})^\top$, $z_{ij} = (1, t_{ij}, t_{ij}^2, \dots, t_{ij}^{d-1})^\top$, $w_{ijl} = (1, (t_{ij} - t_{il}), (t_{ij} - t_{il})^2, \dots, (t_{ij} - t_{il})^{q-1})^\top$ 前提是个体的组内相关性仅依赖于经过的时间。协变量对应的多项式阶数 p, d, q 无需预先指定固定值, 可结合实际数据的时序变化趋势灵活选取。此时, 第 k 类待估参数集合扩展为 $\Psi_k = (\pi_k, \beta_k^\top, \lambda_k^\top, \gamma_k^\top)^\top$ 。参数估计可通过极大似然方法实现。

2.2. 基于 EM 算法的参数估计

本文采用期望最大化(EM)算法对上述模型进行参数估计。引入指示变量 $\{z_{ik}\}$, z_{ik} 为每个个体归类的示性函数, 若 $z_{ik}=1$ 时, 表示 y_i 属于第 k 类, 若 $z_{ik}=0$ 时, 表示 y_i 不属于第 k 类。基于指示变量 $\{z_{ik}\}$, 构建完全数据对数似然函数, 完全数据由观测数据 $\mathbf{Y}$ 和缺失数据 $\{z_{ik}\}$ 组成, 其可以表示为:

$$L_c(\Psi) = \sum_{i=1}^N \sum_{k=1}^K z_{ik} [\ln \pi_k + \ln f_k(y_i | \Psi_k)] \quad (4)$$

EM 算法通过交替执行 E 步和 M 步, 直至参数估计收敛。

E 步(期望步): 在给定当前参数估计值 $\tilde{\Psi}$ 的条件下, 计算 z_{ik} 的条件期望, 即个体 i 属于第 k 类的后验概率:

$$\hat{z}_{ik} = E(z_{ik} | y_i, \tilde{\Psi}) = \frac{\tilde{\pi}_k f_k(y_i | \tilde{\Psi}_k)}{\sum_{l=1}^K \tilde{\pi}_l f_l(y_i | \tilde{\Psi}_l)} \quad (5)$$

M 步(最大化步): 通过最大化完全数据的期望对数似然函数来更新模型参数 Ψ 。混合概率 π_k 可更新为 $\hat{\pi}_k = \sum_{i=1}^N \hat{z}_{ik} / N$ 。

对于参数 Ψ_k , 通过最大化完全数据的期望对数似然进行更新:

$$Q(\Psi | \tilde{\Psi}) = \sum_{i=1}^N \sum_{k=1}^K \hat{z}_{ik} [\ln \pi_k + \ln f_k(y_i | \Psi_k)] \quad (6)$$

在传统仅考虑均值的 FMR 模型中, 这一步等价于加权最小二乘法; 而在本文引入的联合均值 - 协方差混合模型中, 由于均值参数与协方差参数 Σ_k 相互耦合, 通常需采用 Fisher 得分法或 Newton-Raphson 迭代算法同步求解 β_k 及协方差相关参数的最优估计。具体实现时, 可参照 Yu 等人提出的迭代加权最小二乘 EM (IRLS-EM) 算法, 最终可以求得参数值 $\hat{\Psi}$ 。

3. 聚类数目的确定方法

在对纵向混合模型进行参数估计后, 需要进一步确定数据中潜在的最佳聚类类别数 K , 并对不同结构假设的模型进行性能对比。本文采用贝叶斯信息准则(Bayesian Information Criterion, BIC)作为模型选择与评价的核心依据。BIC 综合权衡了模型的拟合优度与模型复杂度, 通过引入对自由参数数量的惩罚项来有效避免过拟合问题。其计算公式如下:

$$\text{BIC} = -2L_c(\hat{\Psi}) + d \ln(N) \quad (7)$$

其中, $L_c(\hat{\Psi})$ 表示模型在达到收敛时的最大化对数似然函数值; d 表示模型中独立待估参数的总个数; N 为样本容量。根据判别准则, BIC 值越小, 表明模型在拟合数据的同时兼具更好的简约性, 即模型表现越优。

4. 跨国收入不平等的实证研究

4.1. 数据来源与样本选择

本文的实证数据来源于世界银行数据库(World Bank DataBank)。考虑到纵向数据建模对时间序列连续性的要求, 本文对原始跨国数据进行了筛选, 剔除了基尼系数(Gini index)缺失较严重或时间跨度不足的样本。最终, 本研究选取了 36 个具有代表性的国家作为研究对象, 研究区间设定为 2004 年至 2023 年。这构成了一个包含 720 个观测值(36 个国家 $\times$ 20 年)的平衡纵向数据。

基尼系数作为衡量收入不平等的经典核心指标, 能够直观反映一国财富分配的均衡程度。为了初步探索这 36 个国家在研究区间内的动态演化特征, 本文绘制了各国基尼系数随时间变化的纵向轨迹(如图 1 所示)。

从图 1 中可以直观地观察到, 各国收入不平等的演化轨迹呈现出高度的异质性。首先, 在整体截距与均值水平上, 各国起点差异巨大, 部分国家基尼系数常年处于高位, 而部分国家则始终保持在较低水平。其次, 在随时间演化的波动特征上, 各国的表现也大相径庭: 部分国家的基尼系数轨迹相对平稳, 而另一些国家则经历了剧烈的上下波动甚至出现了明显的轨迹交叉。这种直观存在的“均值轨迹”与“波

动方差”的双重异质性特征表明,简单的整体回归分析会掩盖不同国家内部真实的发展模式,迫切需要对其进行模式识别与子群划分。同时,传统仅关注均值结构的聚类模型难以充分捕捉图中展现的复杂波动与交叉特征,这也为本文引入基于均值-协方差联合混合回归模型进行精细化聚类提供了直观的数据事实支撑。

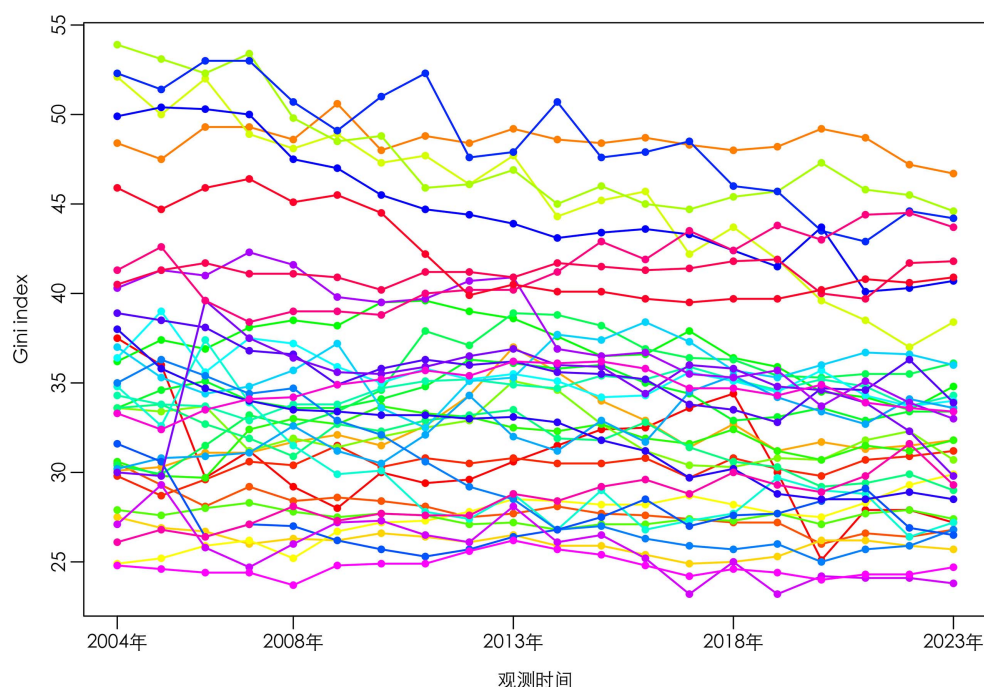


Figure 1. Longitudinal trajectories of Gini coefficients of various countries over time

图 1. 各国基尼系数随时间变化的纵向轨迹

4.2. 模型选择与最优聚类数确定

为了确定基尼系数动态演化的最佳异质性类别数,本文分别计算了传统均值混合回归模型(FMR-1)与均值-协方差联合混合回归模型(FMR-3)在聚类数 K 依次取 1 到 5 时的 BIC 值,其变化趋势如图 2 所示。

从图 2 中可以得出两个核心结论:第一,关于最优聚类数 K 的确定。观察两条折线的走势可以发现,无论是传统的 FMR-1 模型,还是本文引入的 FMR-3 模型,随着聚类数 K 的增加, BIC 值均呈现先下降后上升的“U”型趋势,并且两条曲线均 $K = 4$ 时达到全局最低点。这表明,将这 36 个国家的收入不平等演化轨迹划分为 4 个潜在子群体,能够实现数据拟合与模型复杂度之间的最佳平衡。第二,关于模型结构的优劣对比。通过对比两类模型在相同 K 值下的表现,可以清晰地看到代表均值-协方差联合混合回归模型(FMR-3)的红色曲线始终位于代表传统均值模型(FMR-1)的蓝色曲线下方,且两者之间存在显著的数值落差。这有力地证明了,在放宽了组内独立同分布的强假设后, FMR-3 模型通过对纵向数据的协方差结构进行联合动态建模,极大地提升了模型对复杂跨国纵向数据的解释能力与拟合优度。

4.3. 聚类结果对比与轨迹特征识别

为了直观展示引入协方差结构在纵向数据聚类中的实际优势,本文分别绘制了传统均值模型(FMR-1)与均值-协方差联合混合回归模型(FMR-3)在最佳聚类数 $K = 4$ 时的个体轨迹分类结果,图中 n 代表类别样本国家数量, π 代表类别混合概率。

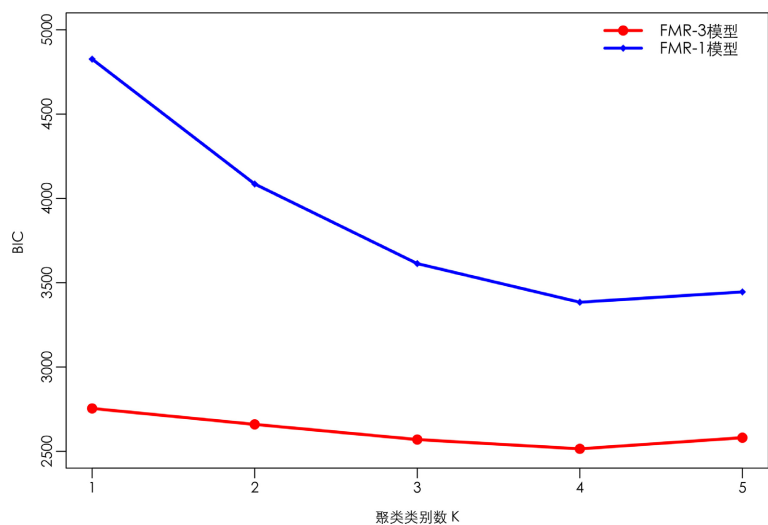


Figure 2. Changes in BIC of the fitted models when $K = 1, \dots, 5$
图 2. $K = 1, \dots, 5$ 时拟合模型的 BIC 变化

观察传统 FMR-1 模型的聚类结果(见图 3)可以发现, 其轨迹区分度较低, 尤其是在基尼系数处于 30 至 40 的中等水平区间内, 模型强制划分出的两个子群体, 在均值轨迹上高度重叠, 且组内个体的波动轨迹发生严重的交叉与混杂。这种现象的根源在于, 传统 FMR 模型严格假设组内观测值条件独立且方差齐性, 导致模型仅能依靠绝对数值的均值差异进行截断式划分, 完全丧失了对纵向数据内部动态波动特征的捕捉能力。

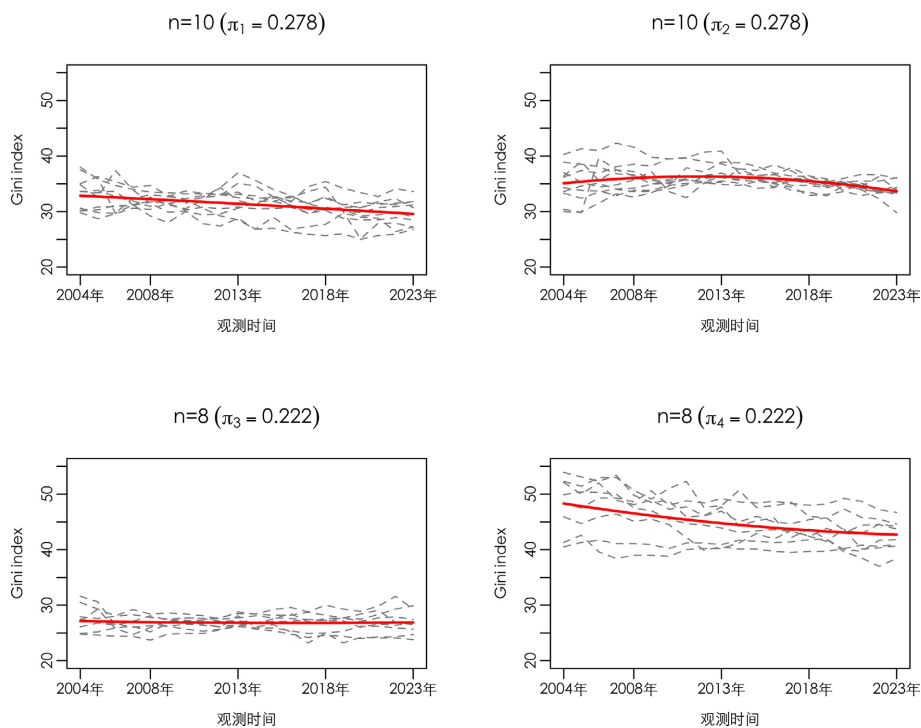


Figure 3. Trajectory plots of observed individuals (gray dashed lines) and fitted mean curves (red solid lines) for each category of FMR-1
图 3. FMR-1 每类观测个体的轨迹图(灰色虚线)及各类拟合均值趋势(红色实线)

相比之下, 均值-协方差联合混合回归模型(FMR-3)的聚类结果展现出了显著的优越性(见图 4)。在放宽了组内独立同分布的强假设后, FMR-3 不仅清晰识别出了处于极端高位与低位的国家群体, 更重要的是, 它能够有效利用数据的动态方差与协方差信息, 将均值水平相近但演化波动模式(如“平稳维持”与“高波动下降”)截然不同的国家精准剥离。这种基于“均值-协方差”双重特征的精细化聚类, 有效克服了传统模型因忽略组内相关性而导致的聚类混杂问题, 为后续深入剖析各类国家的收入分配演化机制提供了更为清晰、可靠的实证基础。

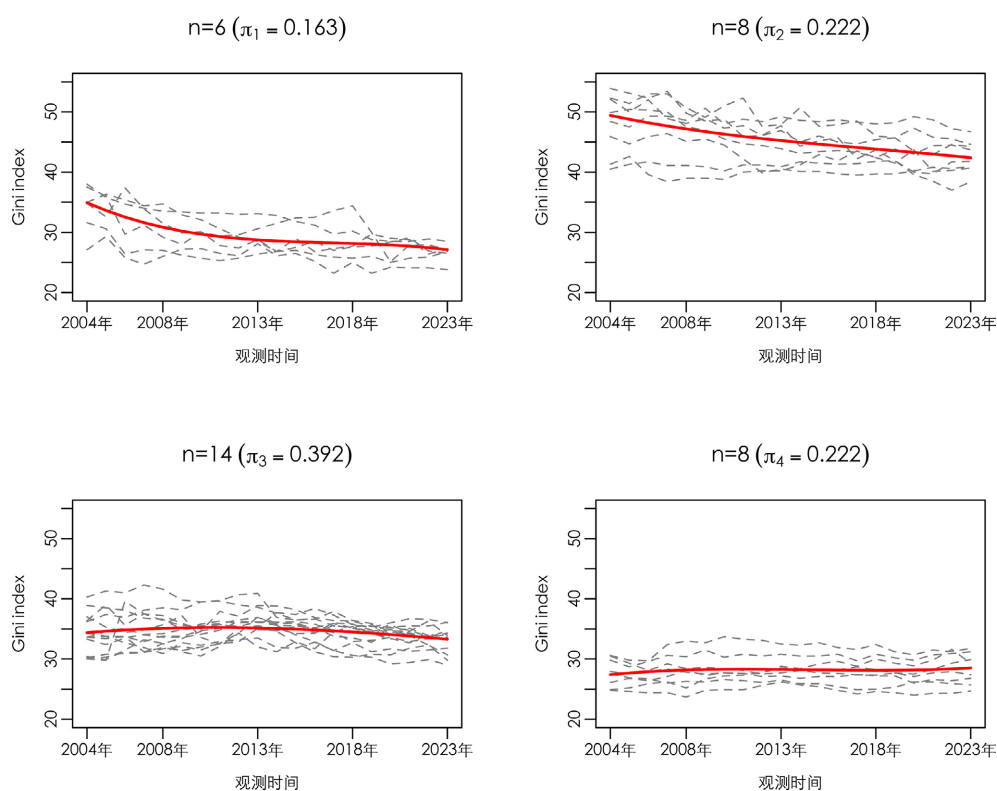


Figure 4. Trajectory plots of observed individuals (gray dashed line) and fitted mean curves (red solid line) for each category of FMR-3

图 4. FMR-3 每类观测个体的轨迹图(灰色虚线)及各类拟合均值趋势(红色实线)

4.4. 样本均值轨迹对比与演化模式捕捉

为了进一步直观揭示两种模型在提取宏观动态趋势上的差异, 本文在 BIC 准则确定的最优聚类数下, 分别提取 FMR-1 与 FMR-3 的组别样本均值轨迹, 绘制如图 5 所示。

观察图 5 左侧的 FMR-1 均值轨迹可以发现, 其划分出的四个子群轨迹近乎平行。这表明传统均值模型主要依据基尼系数的绝对数值水平进行划分, 各组之间的核心差异被固化为静态的截距差异, 难以反映跨国收入不平等在长期发展中可能存在的不同斜率与演化方向。这种隐含的“平行趋势”假设, 可能掩盖数据内部真实的结构性特征。

相比之下, 图 5 右侧 FMR-3 的均值轨迹则呈现出丰富的动态演化特征。最显著的差异体现在红色轨迹上: 该组国家展现出一条特征鲜明的“高位快速下降”路径, 其演化斜率与其他三组明显不同, 甚至在观测期中后段与处于低位平稳的绿色轨迹发生了交叉。这种具有强烈异质性斜率与交叉特征的动态演化模式, 在 FMR-1 模型中未能被识别。上述对比表明, 面对具有复杂非平行甚至交叉特征的纵向数据,

引入协方差结构建模的联合模型(FMR-3)能够更有效地还原数据背后的真实演化轨迹。

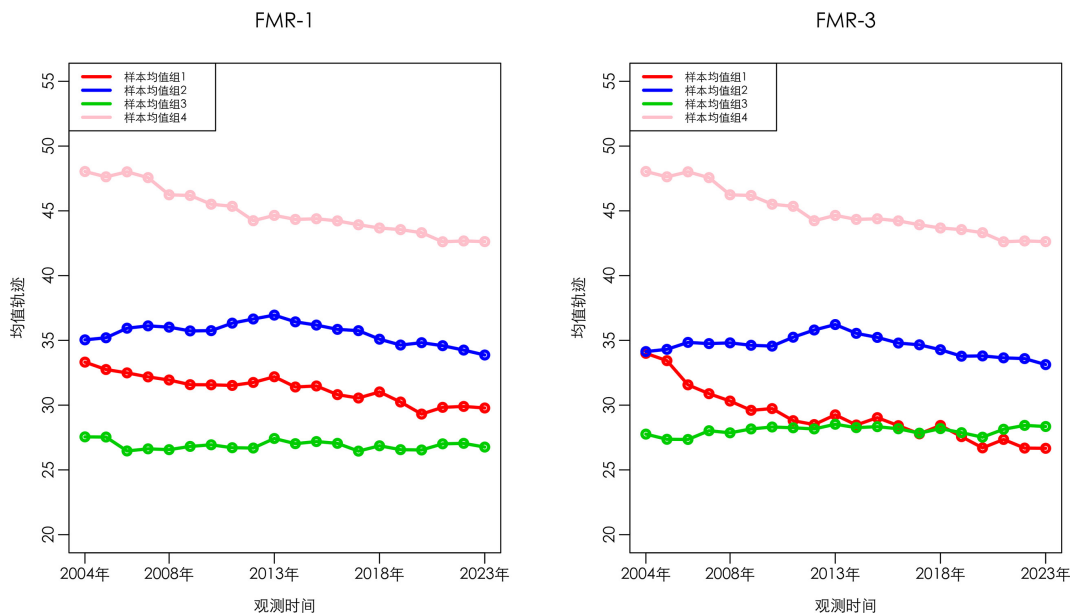


Figure 5. Grouped sample mean trajectories of two mixture models
图 5. 两种模型聚类下组别样本均值轨迹

5. 结论与局限性

5.1. 研究局限性

本研究存在着一定局限性。首先，本研究主要依赖基尼系数这一单一指标，尽管其具有广泛的代表性，但可能无法全面捕捉跨国财富分配或多维度不平等的复杂全貌。其次，受限於高质量增长面板数据的可得性，本文选取的 36 个样本国家在体现全球极其多元的发展模式时，其样本代表性仍有进一步扩展的空间。最后，本文采用的有限混合模型本质上属于探索性数据分析方法，其识别出的演化轨迹反映的是统计学意义上的相关性 与模式归类，并不能直接推断导致这些异质性变化的因果机制。

5.2. 结论

本文利用 36 个国家跨越 20 年的基尼系数纵向数据，对传统均值有限混合回归模型与基于均值 - 协方差联合建模的有限混合回归模型在聚类分析中的表现进行了实证对比。研究表明，在面对具有复杂波动特征的宏观经济指标时，传统均值模型易因强加的方差齐性假设而导致聚类边界模糊与轨迹混杂；相反，联合模型通过同步对均值轨迹与动态协方差结构进行参数化建模，在贝叶斯信息准则(BIC)下取得了最优的模型拟合效果。具体而言，基于联合模型的混合回归将样本国家的基尼系数演化轨迹清晰且稳健地划分为四种异质性模式：高位且剧烈波动的“高位震荡型”、呈现长期收敛趋势的“平稳下降型”，以及轨迹相对固化的“中等平稳型”与“低位稳定型”。本研究从实证层面证实了在高度异质性纵向数据聚类中纳入协方差结构的必要性，从而为精准刻画跨国收入不平等的动态演化规律提供了新的统计分析视角，并揭示了未来需要进一步深入探究的异质性模式。

参考文献

[1] Espoir, D.K. (2022) Convergence or Divergence Patterns in Income Distribution across Countries: A New Evidence

- from a Club Clustering Algorithm. *Cogent Economics & Finance*, **10**, Article 2025667. <https://doi.org/10.1080/23322039.2022.2025667>
- [2] Ogundari, K. (2023) Club Convergence in Income Inequality in Africa. *Social Indicators Research*, **167**, 319-337. <https://doi.org/10.1007/s11205-023-03108-7>
 - [3] Chrisendo, D., Niva, V., Hoffmann, R., Masoumzadeh Sayyar, S., Rocha, J., Sandström, V., et al. (2025) Rising Income Inequality across Half of Global Population and Socioecological Implications. *Nature Sustainability*, **8**, 1601-1613. <https://doi.org/10.1038/s41893-025-01689-4>
 - [4] McLachlan, G. and Peel, D. (2000) Finite Mixture Models. Wiley. <https://doi.org/10.1002/0471721182>
 - [5] Bouguila, N. and Fan, W. (2020) Mixture Models and Applications. Springer.
 - [6] Zamani, H., Faroughi, P. and Ismail, N. (2014) Estimation of Count Data Using Mixed Poisson, Generalized Poisson and Finite Poisson Mixture Regression Models. *AIP Conference Proceedings*, Kuala Lumpur, 17-19 December 2013, 1144-1150. <https://doi.org/10.1063/1.4882628>
 - [7] Nagin, D.S. and Odgers, C.L. (2010) Group-Based Trajectory Modeling in Clinical Research. *Annual Review of Clinical Psychology*, **6**, 109-138. <https://doi.org/10.1146/annurev.clinpsy.121208.131413>
 - [8] Lyrvall, J., Di Mari, R., Bakk, Z., Oser, J. and Kuha, J. (2025) Multilevel Latent Class Analysis: State-of-the-Art Methodologies and Their Implementation in the R Package multilevLCA. *Multivariate Behavioral Research*, **60**, 731-747. <https://doi.org/10.1080/00273171.2025.2473935>
 - [9] Proust-Lima, C., Philipps, V. and Lique, B. (2017) Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *Journal of Statistical Software*, **78**, 1-56. <https://doi.org/10.18637/jss.v078.i02>
 - [10] Van der Nest, G., Lima Passos, V., Candel, M.J. and van Breukelen, G.J. (2020) An Overview of Mixture Models for Clustering Longitudinal Data. *Advances in Data Analysis and Classification*, **14**, 361-384.
 - [11] Wu, L., Li, S. and Tao, Y. (2020) Estimation and Variable Selection for Mixture of Joint Mean and Variance Models. *Communications in Statistics-Theory and Methods*, **49**, 3904-3925.
 - [12] Yu, J., Nummi, T. and Pan, J. (2022) Mixture Regression for Longitudinal Data Based on Joint Mean-Covariance Model. *Journal of Multivariate Analysis*, **190**, Article 104956. <https://doi.org/10.1016/j.jmva.2022.104956>
 - [13] Yu, J. and Pan, J. (2025) Variable Selection in Mixture Regression for Longitudinal Data Based on Joint Mean-Covariance Model. *Journal of Multivariate Analysis*, **212**, Article 105548. <https://doi.org/10.1016/j.jmva.2025.105548>
 - [14] De la Cruz-Mesía, R., Quintana, F.A. and Marshall, G. (2007) Model-Based Clustering for Longitudinal Data. *Computational Statistics & Data Analysis*, **51**, 6443-6457.
 - [15] Pourahmadi, M. (1999) Joint Mean-Covariance Models with Applications to Longitudinal Data: Unconstrained Parameterisation. *Biometrika*, **86**, 677-690. <https://doi.org/10.1093/biomet/86.3.677>
 - [16] Cepeda, E. and Gamerman, D. (2001) Bayesian Modeling of Variance Heterogeneity in Normal Regression Models. *Brazilian Journal of Probability and Statistics*, **15**, 139-153.
 - [17] Pourahmadi, M. (2000) Maximum Likelihood Estimation of Generalised Linear Models for Multivariate Normal Covariance Matrix. *Biometrika*, **87**, 425-435. <https://doi.org/10.1093/biomet/87.2.425>