arXiv:2611.32693v1 [stat.ME] 24 Nov 2026

基于模型的聚类：采用隐半马尔可夫模型的纵向函数型数据

Ba气体 Caesar Suherlan* Emilio Porcu† Antonello Maruot ti‡

摘要

纵向函数型数据的基于模型的聚类必须同时考虑曲线之间的异质性，以及其潜在组成员身份中的时间依赖性。函数型混合模型通常将曲线视为可交换的，而时间序列方法可能会限制状态持续性，或忽略特定于状态的函数变异。我们提出用于多元函数型观测有序序列的隐半马尔可夫模型，其中每个潜在状态都具有简约的、特定于状态的函数主子空间，从而将函数型高维数据聚类（funHDDC）扩展到纵向场景。高斯、污染高斯和多元-t 发射分布能够处理异常观测；显式的特定于状态的驻留时间分布则控制状态持续性，其中隐马尔可夫模型是其几何特例。参数估计采用期望最大化（EM）以及期望/条件最大化（ECM）算法，并结合分段前向—后向递推；同时证明了模型在标签置换和谱不确定性意义下的可识别性。在模拟中，当单条曲线提供的信息较弱时，时间依赖性带来的增益最大；采用稳健发射分布时，特定于状态的子空间优于全局函数主成分表示；当各状态存在重叠且驻留时间分布为非几何分布时，显式持续时间带来的额外增益较小。对空气质量剖面和智能手机活动记录的应用，在留出年份和受试者上进行评估后显示出更好的预测拟合，并说明表面上的持续时间效应可能源于时变转移。

关键词： 污染高斯分布；显式持续时间模型；函数主子空间；状态切换；驻留分布。

*哈利法大学数学系（100066890@ku.ac.ae）。作者声明，本文所述研究未获得任何资助。

†哈利法大学数学系和生物技术中心（emilio.porcu@ku.ac.ae）。

‡哈利法大学公共卫生与流行病学系和生物技术中心，以及 LUMSA 大学 GEPLI 系（antonello.maruotti@ku.ac.ae）。

1 引言

许多当代数据集由多元函数型观测的纵向序列构成。相关示例包括每日污染剖面、重复的生理测量，以及传感器记录中连续的窗口。这类数据呈现出两种不同形式的结构：每个函数型观测内部的变化，以及随时间收集的观测值之间的依赖关系。当底层系统在持续存在的状态之间演化时，聚类必须同时考虑这两种结构。特别是，形状相似的曲线未必属于同一状态，而异常观测也不应被自动解释为状态变化。当相对于单个序列中可获得的信息而言，函数表示具有较高维度时，这些困难会进一步加剧。

通过有限混合、潜在类别模型、混合效应模型和函数方法，对纵向数据进行聚类的研究已十分广泛（Lu 2025、Zeldow 等，2021年、Zhou 等，2023年）。其中一个根本区别在于被聚类的对象。许多纵向聚类方法将个体的完整轨迹分配到一个潜在组中，从而描述个体间异质性，同时允许组内存在变异。函数聚类同样旨在寻找由曲线构成的同质组，通常通过基展开或函数主成分分析实现（Jacques 与 Preda，2014、Bouveyron 与 Jacques，2011）。针对多变量轨迹和纵向轨迹的扩展研究进一步处理了其他异质性来源，包括相位变异和个体特异性函数模式（Park 与 Ahn，2017、Zeldow 等，2021年）。当组成员身份在整个观测期间保持稳定时，这些方法非常有价值。然而，对于同一观测单位可能先后处于不同潜在状态的情形，这些方法并不完全适用。

另一个相关挑战是函数表示的维数。函数型高

维数据聚类（funHDDC）将高维数据聚类（Bouveyron 等，2007）推广到曲线，通过为每个簇关联一个低维函数主子空间和简约残差变异来解决这一问题（Bouveyron 与 Jacques，2011，Schmutz 等，2020）。与全局函数主成分表示不同，状态特定子空间允许不同状态呈现不同的函数变异方向。然而，传统函数型混合模型通常将曲线视为可交换的。因此，它们无法利用重复观测的顺序来区分持续状态与孤立的异常曲线，也无法在状态特定分布重叠时消除分类不确定性。

潜在状态模型在函数聚类与纵向依赖之间建立了自然的概率联系。隐马尔可夫模型（隐马尔可夫模型）通过转移概率表示未观测状态的演化（Zucchini 等，2016年、Bartolucci 等，2012），而扩展到函数型观测的模型已证明其对于建模多变量曲线序列具有实用价值（Martino 等，2020年、Sidrow 等，2022）。相关的因子与子空间隐状态模型还展示了如何将时间依赖性与高维观测的简约表示相结合（Golini 等，2026、Maruotti 等，2017）。这些方法的显著特点是，簇归属可能随时间发生变化，而不是在整个纵向轨迹中保持不变。然而，普通隐马尔可夫模型施加了几何状态持续时间假设：条件于当前状态，离开该状态的概率与已经处于该状态的时间无关。

对于纵向函数观测，这一限制可能会产生重要影响。生理、环境或行为状态的持续时间本身可能包含有关底层过程的信息，而这些信息是个体曲线和连续转移概率所不具备的。隐半马尔可夫模型通过单独指定……来处理这一特征

各状态持续时间的分布，并将其与连续状态之间的转移概率分开（Guédon 2003，Yu 2010）。几何驻留时间分布是隐马尔可夫模型的一个特例。因此，统计上的挑战在于，将显式的持续时间依赖性与简约的函数型发射相结合，同时保留足够的灵活性，以适应异常曲线而不引入虚假状态。

我们提出一类隐半马尔可夫模型，用于对多元函数型观测的纵向序列进行聚类。给定潜在状态后，每条曲线均通过状态特定的函数主子空间表示，其中包含不受约束的均值函数、主要函数型尺度特征值以及简约残差变异。高斯、污染高斯和多元t 发射分布可适应不同的状态内变异模式和异常观测，这一方法建立在稳健函数型混合模型（Amovin-Assagba 等，2022、Anton 和 Smith，2024）的基础之上。嵌入式转移矩阵描述状态间的变化，而移位泊松分布、移位负二项分布和非参数逗留时间分布则允许呈现不同形式的持续性。由此得到的框架在基于似然的模型中结合了函数聚类与纵向状态分段，并将隐马尔可夫模型保留为嵌套基准模型。

参数估计基于分段前向–后向递推，并采用期望最大化（EM）（Dempster 等，1977）或期望/条件最大化（ECM）（Meng 和 Rubin，1993）算法。简约的子空间结构支持对发射密度进行低秩计算，而似然则考虑了末端删失，并在适当情况下纳入初始残余时长。我们给出了函数型发射、状态特定子空间和半马尔可夫动态可识别性的充分条件，同时考虑特征值重数、污染成分配对以及观测时域。

数值研究旨在评估每个建模组

件的贡献，而不是预先假定最灵活规范的优越性。模拟研究在匹配的发射分布和时间结构下，将状态特定函数子空间与全局函数主成分表征进行比较，并区分稳健性、马尔可夫持久性和显式持续时间的影响。针对空气质量剖面和智能手机活动记录的应用，在留出年份和受试者上进行评估。结果表明，时间依赖性和状态特定子空间可以改善聚类和预测拟合，而显式持续时间带来的额外收益取决于状态重叠程度和潜在的持续性结构。特别是，表面上的非几何持续时间可能反映时变转移动态，这凸显了区分持续时间依赖性与非平稳性的重要性。第 2 节介绍模型并给出简明的可识别性陈述，第 3 节阐述似然推断与计算，第4 节和 5 节报告模拟研究和应用，第6 节总结全文。

2 函数表示与隐半马尔可夫聚类

2.1 纵向多变量函数观测值

我们考虑按预定顺序收集的 N 个独立序列 $\mathcal{D}_i = \{\boldsymbol{X}_{i\ell} : \ell = 1, \ldots, L_i\}$，$i = 1, \ldots, N$，每个序列由多元函数型观测构成，其中 L_i 表示序列 i 中函数型观测的数量，且每个观测 $\boldsymbol{X}_{i\ell}(t) = (X_{i\ell}^{(1)}(t) \ldots, X_{i\ell}^{(p)}(t))^\top, t \in [0, T]$ 都属于希尔伯特空间 $\mathcal{H} = L^2([0, T]; \mathbb{R}^p)$。因此，数据涉及两个不同的索引。自变量 t 描述单个函数型观测内部的变化，而 ℓ 则对同一纵向序列中的连续观测进行索引。例如，t 可以表示一天中的小时，ℓ 表示日历日；或者 t 可以表示

传感器窗口内的时间，而 ℓ 表示该窗口在一次记录中的位置。由 i 索引的观测单位可以是受试者、监测位置或独立的记录会话。

每个函数成分仅在有限个时间点上观测，并通过基展开表示，

$$X_{i\ell}^{(s)}(t) = \sum_{r=1}^{R_s} c_{i\ell r}^{(s)} \xi_r^{(s)}(t), \qquad s = 1, \dots, p, \tag{1}$$

其中 $\{\xi_r^{(s)}\}_{r=1}^{R_s}$ 是 $L^2[0,T]$ 中的一组线性无关基函数，$c_{i\ell r}^{(s)}$ 是相应的系数。等价地，

$$\boldsymbol{X}_{i\ell}(t) = \xi(t)\boldsymbol{c}_{i\ell}, \qquad \boldsymbol{c}_{i\ell} \in \mathbb{R}^R, \quad R = \sum_{s=1}^{p} R_s, \tag{2}$$

其中 $\xi(t) \in \mathbb{R}^{p \times R}$ 是块对角基矩阵，$\boldsymbol{c}_{i\ell}$ 将各函数成分对应的系数堆叠在一起。函数表示的几何结构由 Gram 矩阵 $W = \int_0^T \xi(t)^\top \xi(t)\, \mathrm{d}t$ 决定；由于所选基函数线性无关，该矩阵为对称正定矩阵。由于 $\xi(t)$ 是块对角矩阵，$W = \mathrm{diag}(W_1, \dots, W_p)$，其中 $W_s = \int_0^T \xi^{(s)}(t)\, \xi^{(s)}(t)^\top \mathrm{d}t$，且 $\xi^{(s)}(t) = (\xi_1^{(s)}(t) \dots, \xi_{R_s}^{(s)}(t))^\top$。对于分别表示函数 $\boldsymbol{X}$ 和 $\boldsymbol{X}'$ 的两个系数向量 $\boldsymbol{c}$ 和 $\boldsymbol{c}'$，$\langle \boldsymbol{X}, \boldsymbol{X}' \rangle_{\mathcal{H}} = \boldsymbol{c}^\top W \boldsymbol{c}'$，因此，变换后的系数 $W^{1/2}\boldsymbol{c}_{i\ell}$ 在欧几里得坐标中保留了函数内积几何。

我们假设纵向观测值由 K 个潜在状态产生。令 $S_{i\ell} \in \{1, \dots, K\}$ 表示与序列 i 的第 ℓ 个函数型观测对应的未观测状态，并定义 $Z_{i\ell k} = \mathbf{1}\{S_{i\ell} = k\}$。全文中，状态 指随机过程的一个潜在类别，状态 指其作为底层系统持续状态的解释，而簇 则是分配到该状态的函数型观测的集合。给定潜在状态后，假设函数型观测相互独立，其系数密度取决于状态，记为 $\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k \sim f_k(\cdot; \boldsymbol{\theta}_k)$。这种条件独立性并不意味着相邻观测值在边际上相互独立

：纵向依赖是通过潜在状态的联合分布引入的，该分布由下文的半马尔可夫结构规定。不同序列的末尾与开头之间不定义转移。

2.2 状态特定的函数子空间

式（2）中的函数表示可能具有较高维度，尤其是在观测到多个函数分量时，因此，为每个潜在状态估计一个无约束协方差矩阵可能不稳定，或计算开销较大。我们通过函数型高维数据聚类（funHDDC）的简约协方差结构来处理这一问题，使每个状态都对应一个自身的低维函数主子空间。对于状态 k，令 $\mathrm{E}_k \subset \mathrm{H}$ 表示维数为 d_k 的函数子空间，其中 $1 \le d_k \le R-1$，其正交归一基函数表示为$\zeta_{kr}(t) = \xi(t)W^{-1/2}\boldsymbol{u}_{kr}$, $r = 1, \ldots, d_k$，其中 $U_k = [\boldsymbol{u}_{k1}, \ldots, \boldsymbol{u}_{kd_k}]$和 $U_k^\top U_k = I_{d_k}$。事实上，$\zeta_{kj}, \zeta_{kh\mathcal{H}} = \boldsymbol{u}_{kj}^\top \boldsymbol{u}_{kh} = \delta_{jh}$，因此矩阵 U_k 确定了白化系数空间中的状态特定函数主方向。条件于 $S_{i\ell} = k$，我们假设潜在得分表示

$$\boldsymbol{c}_{i\ell} = \boldsymbol{\mu}_k + W^{-1/2}U_k\boldsymbol{\delta}_{i\ell} + \boldsymbol{\varepsilon}_{i\ell}, \tag{3}$$

其中 $\boldsymbol{\mu}_k \in \mathbb{R}^R$不受限制，且相互独立项满足 $\boldsymbol{\delta}_{i\ell} \sim N_{d_k}(\mathbf{0}, \Gamma_k)$ 和 $\boldsymbol{\varepsilon}_{i\ell} \sim N_R(\mathbf{0}, b_k W^{-1})$，其中 $\Gamma_k = \mathrm{diag}(a_{k1} - b_k, \ldots, a_{kd_k} - b_k)$ 且 $a_{k1} \ge \cdots \ge a_{kd_k} > b_k > 0$。由此可得，系数向量的条件高斯分布的均值为 $\boldsymbol{\mu}_k$，协方差矩阵为

$$\Sigma_k = W^{-1/2}\left[U_k A_k U_k^\top + b_k(I_R - U_k U_k^\top)\right]W^{-1/2}, \qquad A_k = \mathrm{diag}(a_{k1}, \ldots, a_{kd_k}). \tag{4}$$

主特征值 a_{kj} 描述沿各状态特定函数主方向的变异，而 b_k 表示其正交补空间中的残差变异。因此，不同状态可能具有不同的均值函数、内在维数、主

方向和函数变异性模式，这比在所有状态间强制采用统一的全局函数主子空间更加灵活。我们将一般协方差结构记为 $[a_{kj}, b_k, U_k, d_k]$。通过对状态内或状态间选定的参数施加相等约束，可得到简约变体，详见附录 A.2。

2.3 函数型发射族

我们考虑三种发射族，它们共享位置参数 $\boldsymbol{\mu}_k$ 和式（4）中的结构化尺度矩阵 Σ_k，但对异常观测及尾部行为的处理方式不同。在高斯设定下，$\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k, \Sigma_k)$，其发射密度为 $f_k(\boldsymbol{c};\, \boldsymbol{\theta}_k) = \phi_R(\boldsymbol{c};\, \boldsymbol{\mu}_k, \Sigma_k) = (2\pi)^{-R/2}|\Sigma_k|^{-1/2}\exp\{-\Delta_k(\boldsymbol{c})/2\}$，其中 $\Delta_k(\boldsymbol{c}) = (\boldsymbol{c}-\boldsymbol{\mu}_k)^\top \Sigma_k^{-1}(\boldsymbol{c}-\boldsymbol{\mu}_k)$ 是马氏距离平方。该设定对应高斯 funHDDC 模型。为适应异常函数型观测，我们考虑污染高斯发射（Punzo & Maruotti 2016年，Punzo & McNicholas 2016年）$f_k(\boldsymbol{c};\, \boldsymbol{\theta}_k) = \alpha_k \phi_R(\boldsymbol{c};\, \boldsymbol{\mu}_k, \Sigma_k) + (1-\alpha_k)\phi_R(\boldsymbol{c}; \boldsymbol{\mu}_k, \eta_k \Sigma_k)$，其中 $0 < \alpha_k < 1$ 是观测值属于常规观测的概率，$\eta_k > 1$ 是异常观测的协方差膨胀因子。等价地，引入潜在指示变量 $\Upsilon_{i\ell} \in \{0, 1\}$，其中 $\Pr(\Upsilon_{i\ell} = 1 \mid S_{i\ell} = k) = \alpha_k$；当 $\Upsilon_{i\ell} = 1$ 时，条件分布为协方差为Σ_k 的高斯分布，否则为 $\eta_k \Sigma_k$。两个分量共享相同的位置参数和主方向，因此无需一定引入额外状态即可处理异常曲线。

作为重尾替代方案（Maruotti 等人 2019），我们考虑多变量t 发射分布（Peel & McLachlan 2000、Bernardi 等人 2017）、$\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k \sim t_R$（$\boldsymbol{\mu}_k$、$\Sigma_k$、$\nu_k$），其密度为

$$f_k(\boldsymbol{c};\boldsymbol{\theta}_k) = \frac{\Gamma((\nu_k + R)/2)}{\Gamma(\nu_k/2)(\pi\nu_k)^{R/2}|\Sigma_k|^{1/2}} \left(1 + \frac{\Delta_k(\boldsymbol{c})}{\nu_k}\right)^{-(\nu_k+R)/2},$$

其中，$\nu_k > 0$ 表示自由度。这里，Σ_k 是尺度矩阵，而非协方差矩阵。当 $\nu_k > 2$时，$\mathrm{Cov}(\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k) = \nu_k\Sigma_k/(\nu_k - 2)$，因此其主

方向与结构化尺度矩阵的方向一致。自由度可以因状态而异，也可以限制为共同值 $\nu_1 = \cdots = \nu_K = \nu$。多元 t 分布还具有潜在尺度表示 $H_{i\ell} \mid S_{i\ell} = k \sim$ 伽马分布 $(\nu_k/2, \nu_k/2)$，采用形状–率参数化，且 $\boldsymbol{c}_{i\ell} \mid H_{i\ell} = h, S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k, \Sigma_k/h)$。辅助变量 $\Upsilon_{i\ell}$ 和 $H_{i\ell}$ 有助于第3节所述的 EM/ECM 估计，详细的发射更新见附录A。

我们分别将这三种规范称为 funHDDC、C-funHDDC 和 T-funHDDC 发射分布。在每种情况下，$\boldsymbol{\theta}_k$ 包含状态特定位置、子空间和尺度参数（$\boldsymbol{\mu}_k$，U_k，A_k，b_k），以及 C-funHDDC 的（α_k，η_k）或 T-funHDDC 的 ν_k。内在维数 d_k 和任何协方差共享约束共同定义所选的发射规范。

2.4 显式持续时间隐半马尔可夫模型

我们通过隐半马尔可夫过程对潜在状态的演变进行建模，其中持续性由显式的、状态特定的驻留时间分布表示。这样便将状态的持续时间与下一状态的选择区分开来。对于序列 i，令M_i 表示与观测窗口相交的潜在片段数。片段j具有状态 $B_{ij} \in \{1, \ldots, K\}$，其完整持续时间为 $V_{ij} \in \{1, 2, \ldots\}$，以连续函数型观测值的个数计量。记作 $\kappa_{i0} = 0$ 和 $\kappa_{ij} = \sum_{j'=1}^{j} V_{ij'}$，观测窗口包含该过程的前 L_i 个时间点，其中 $\kappa_{i,\,M_i-1} < L_i \leq \kappa_{iM_i}$。因此，最后一个片段可能会延续到观测窗口之外。若 $\kappa_{i,\,j-1} < \ell \leq \kappa_{ij}$，则与观测值 ℓ 相关联的状态为 $S_{i\ell} = B_{ij}$。

分段状态遵循一个嵌入马尔可夫链，其初始分布为 $\rho_k = \Pr(B_{i1} = k)$、$\sum_{k=1}^{K} \rho_k = 1$，转移矩阵为 $P^* = (p^*_{hk})$，其中 $p^*_{hk} = \Pr(B_{ij} = k \mid B_{i,\,j-1} = h)$ 对于$j \geq 2$成立。由于持续性通过分段持续时间表示，因此自转移被

排除：$p^*_{hh} = 0$、$p^*_{hk} \geq 0$ 和 $\sum_{k \neq h} p^*_{hk} = 1$。条件于 $B_{ij} = k$，完整持续时间的概率质量函数为 $\varphi_k(v;\, \boldsymbol{\beta}_k) = \Pr(V_{ij} = v \mid B_{ij} = k)$，$v = 1$，2、.....，其中 $\boldsymbol{\beta}_k$ 表示其参数。我们假设，条件于当前分段状态，其持续时间与为下一分段选择的状态以及前序片段历史 $\mathcal{H}_{ij}$相互独立。因此，对于 $h \neq k$，$\Pr(V_{ij} = v,\ B_{i,\,j+1} = h \mid B_{ij} = k,\ \mathcal{H}_{ij}) = \varphi_k(v;\boldsymbol{\beta}_k)\, p^*_{kh}$。

我们考虑三种持续时间设定。移位泊松分布（P）模型假设 $V_{ij} = 1 + V_{ij}$，其中 $V_{ij} \sim \text{Pois}(\lambda_k)$，从而得到 $\mathbb{E}(V_{ij} \mid B_{ij} = k) = 1 + \lambda_k$，且 $\text{Var}(V_{ij} \mid B_{ij} = k) = \lambda_k$。移位负二项分布（NB）模型假设 $V_{ij} = 1 + V_{ij}$，其中 $V_{ij} \sim \text{NB}(m_k, r_k)$；负二项分布以均值 $m_k > 0$ 和大小参数 $r_k > 0$进行参数化，其中

$$\Pr(\tilde{V}_{ij} = v) = \frac{\Gamma(v + r_k)}{\Gamma(r_k)v!} \left(\frac{r_k}{r_k + m_k}\right)^{r_k} \left(\frac{m_k}{r_k + m_k}\right)^{v}.$$

其均值和方差分别为 $\mathbb{E}(V_{ij} \mid B_{ij} = k) = 1 + m_k$ 和 $\text{Var}(V_{ij} \mid B_{ij} = k) = m_k + m_k^2/r_k$。移位泊松分布由 $r_k \to \infty$ 得到，而 $r_k = 1$ 则给出定义在 $\{1, 2, \ldots\}$ 上的几何分布。最后，非参数（NP）规范为有限持续时间支撑集分配概率$\varphi_k(1) \ldots, \varphi_k(V_{\max})$，并满足 $\varphi_k(v) \geq 0$和 $\sum_{v=1}^{V_{\max}} \varphi_k(v) = 1$。将这些持续时间分布和嵌入式转移过程与第 2.3 节中依赖状态的发射分布结合起来，即定义了所提出的函数型隐半马尔可夫模型族。我们分别将其高斯、污染高斯和多元-t版本记为 HSMM-funHDDC、HSMM-C-funHDDC 和 HSMM-T-funHDDC，其简约协方差变体与底层 funHDDC 族相同。当所有逗留时间分布均为几何分布时，即可得到普通隐马尔可夫模型。具体而言，对于具有 $p_{kk} < 1$、$\varphi_k(v) = (1 - p_{kk})p_{kk}^{v-1}$和 $p^*_{kh} = p_{kh}/(1 - p_{kk})$（$h \neq k$）的 HMM 转移矩阵 $P = (p_{hk})$。反过来，具有几何停留时间的半马尔可夫模型会导出一个 HMM，其中 $p_{kk} = 1 - q_k$ 和 $p_{kh} = q_k p^*_{kh}$（$h \neq k$），q_k 为几何退出概率。HMM 的表述及其推断方法见附录B。对于模型选择，半马尔可夫潜在结构包含 $(K - 1) + K(K - 2) + \tau_{\text{dur}}$ 个自由参数，

其中各项分别对应初始分布、嵌入式转移矩阵和持续时间分布；持续时间部分的参数量为 $\tau_{\mathrm{dur}} = K$ （移位泊松分布），$2K$ （平移负二项分布），以及 $K\,(V_{\max} - 1)$ （非参数逗留时间分布）。其余参数来自所选的发射规范；完整的参数量及模型选择准则见附录C.3。

在固定窗口观测约定下，最后一个分段是右删失的。如果其观测长度为 u，则其持续时间贡献为生存概率

$$\bar{\varphi}_k(u;\boldsymbol{\beta}_k) = \Pr(V_{ij} \geq u \mid B_{ij} = k) = \sum_{v \geq u} \varphi_k(v;\boldsymbol{\beta}_k), \tag{5}$$

而非精确持续时间 u (Guédon2003)。假设初始分段始于一个状态进入时间，因此不属于左删失；不过，当它横跨整个观测窗口时，也可能是右删失的。对于始于持续状态过程之中的窗口，需要指定初始剩余持续时间分布，如附录 C.1中所述，并在第5.1节中使用。

2.5 函数型隐半马尔可夫模型的可识别性

对于已知的状态数 K 和固定的函数基，识别问题涉及状态依赖发射密度、它们的函数谱子空间、初始分布 $\boldsymbol{\rho}$、嵌入式转移矩阵 P^*以及持续时间参数$\{\boldsymbol{\beta}_k\}$。附录 D 中的命题 1 给出了从观测系数序列的总体分布中识别这些参数的充分条件，识别结果仅差一个共同的状态标签排列以及标准的谱不确定性。关键在于，该命题假设已知 K、状态发射分布可识别，并且序列从状态进入时刻开始（假设1（A5））；其有限时域条件确保在所选逗留时间族内，可观测的持续时间概率能够确定持续时间参数。对于从正在进行的状态内部开始的窗口，结论并不会自动成立

（注 1），在这种情况下，初始观测到的逗留时间服从剩余持续时间分布。证明还处理了并列主特征值和持续时间族条件。

3 似然推断与计算

˜HSMM 参数向量为 $\boldsymbol{\Theta} = \{\boldsymbol{\rho}, P^*, \boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K\}$。E 步得出后验状态概率 $\gamma_{i\ell k} = \Pr(S_{i\ell} = k \mid \mathcal{D}_i)$、后验片段以及嵌入式转移计数，所有发射更新均使用这些后验状态概率。上标 (m) 表示算法第 m 次迭代时的值。

3.1 发射分布的高效计算

对于系数向量 $\boldsymbol{c}$ 和状态 k，我们定义 $\boldsymbol{z}_k = W^{1/2}(\boldsymbol{c} - \boldsymbol{\mu}_k)$ 和 $\boldsymbol{y}_k = U_k^\top \boldsymbol{z}_k$，其分量为 y_{kj}。funHDDC 结构给出

$$\log(|\Sigma_k|) = -\log(|W|) + \sum_{j=1}^{d_k} \log(a_{kj}) + (R - d_k)\log(b_k), \qquad \Delta_k(\boldsymbol{c}) = \sum_{j=1}^{d_k} \frac{y_{kj}^2}{a_{kj}} + \frac{\|\boldsymbol{z}_k - U_k \boldsymbol{y}_k\|^2}{b_k},$$

第 2.3 节中的马氏距离平方。因此，发射概率计算既不需要完整的协方差逆矩阵，也不需要残差子空间的基底。

由于 W 固定，$W^{1/2}\boldsymbol{c}_{i\ell}$ 和 $\log(|W|)$ 可以预先计算，而 $W^{1/2}\boldsymbol{\mu}_k$ 和对数行列式则针对每个状态在每次迭代中更新一次。随后，对于所有三种发射族，每个观测值和状态的计算成本为 $O(Rd_k)$。对于膨胀高斯分量，马氏距离平方为 $\Delta_k(\boldsymbol{c})/\eta_k$，对数行列式为 $\log(|\Sigma_k|) + R\log(\eta_k)$。

3.2 分段似然与隐半马尔可夫模型 E 步

我们采用第 2.4 节的记号：B_{ij} 表示序列 j的第 个片段所处的状态i，V_{ij} 表示其持续时间，$\kappa_{ij} = V_{i1} + \cdots + V_{ij}$ 表示累积分段端点，M_i 表示片段的数量

该数量本身是潜变量， P^* 表示嵌入式转移矩阵， $\varphi_k(v;\ \boldsymbol{\beta}_k)$表示逗留时间概率质量函数。对于一段以时刻 r结束、包含 u个连续函数型观测的区块，我们定义 $\Psi_i(r,u;\ k)$ $=\prod_{\ell=r-u+1}^{r} f_k(\boldsymbol{c}_{i\ell};\ \boldsymbol{\theta}_k)$，并将 $\Psi_i^{(m)}(r,u;\ k)$记作在 $\boldsymbol{\theta}_k^{(m)}$ 处计算的同一乘积。每个序列都在固定窗口内进行观测，因此其最后一个片段为右删失。因此，只要某个片段在 L_i 结束，就用生存函数（5）替代持续时间概率质量$\varphi_k(u)$，而内部片段保持不变。将这一替代纳入一个统一符号中会更为方便：对于一个持续时间为 u 、在时刻 r结束的片段，

$$\varphi_k^{\dagger}(u \mid r;\boldsymbol{\beta}_k)=\begin{cases}\varphi_k(u;\boldsymbol{\beta}_k), & r<L_i,\\[2mm] \bar{\varphi}_k(u;\boldsymbol{\beta}_k), & r=L_i,\end{cases} \tag{6}$$

这种依赖仅通过 r 是否为该序列的最终时刻体现。观测长度 u 对于内部片段是完整持续时间，但对于终端片段（包括跨越整个窗口的片段）仅是一个下界 $V_{ij}\geq u$。对于几何持续时间，我们有 $\varphi_k(u)=(1-p_{kk})p_{kk}^{u-1}$以及 $\bar{\varphi}_k(u)=p_{kk}^{u-1}$。采用终端片段的生存函数后，可以精确恢复固定窗口隐马尔可夫模型似然；而持续时间概率质量则会引入额外的终端退出因子 $1-p_{kk}$。

第 i 个序列的似然贡献可以写成对所有潜在分段求和的形式，

$$\mathcal{L}_i(\tilde{\boldsymbol{\Theta}})=\sum_{q=1}^{L_i}\sum_{\substack{b_1,\dots,b_q\\ b_j\neq b_{j+1}}}\sum_{\substack{u_1,\dots,u_q\geq 1\\ u_1+\cdots+u_q=L_i}}\rho_{b_1}\varphi_{b_1}^{\dagger}(u_1\mid u_1;\boldsymbol{\beta}_{b_1})\Psi_i(u_1,u_1;b_1)$$
$$\times\prod_{j=2}^{q}p_{b_{j-1},b_j}^{*}\,\varphi_{b_j}^{\dagger}\Big(u_j\mid\textstyle\sum_{j'=1}^{j}u_{j'};\boldsymbol{\beta}_{b_j}\Big)\,\Psi_i\Big(\sum_{j'=1}^{j}u_{j'},u_j;b_j\Big),$$

其中，各求和分别遍历片段数 q、其状态 b_1，……， b_q及其长度u_1，……， u_q；每种分段方式的最后一个因子带有生存函数，因为该块在 L_i 处结束，完整似然为 $\mathcal{L}(\tilde{\boldsymbol{\Theta}})$ $=\prod_{i=1}^{N}\mathcal{L}_i(\tilde{\boldsymbol{\Theta}})$。完整数据对数似然可~用片段级指示变量 $Z_{ijk}=\mathbf{1}\{B_{ij}=k\}$ 表示，其中 $j=1$，……， M_i。忽略

与参数无关的常数后，我们得到

$$\begin{aligned}\Lambda_c(\tilde{\boldsymbol{\Theta}}) = &\sum_{i=1}^{N}\sum_{k=1}^{K}\widetilde{Z}_{i1k}\log\rho_k + \sum_{i=1}^{N}\sum_{j=2}^{M_i}\sum_{h=1}^{K}\sum_{\substack{k=1\\k\neq h}}^{K}\widetilde{Z}_{i,j-1,h}\widetilde{Z}_{ijk}\log p^*_{hk}\\ &+\sum_{i=1}^{N}\left[\sum_{j=1}^{M_i-1}\sum_{k=1}^{K}\widetilde{Z}_{ijk}\log\varphi_k(V_{ij};\boldsymbol{\beta}_k) + \sum_{k=1}^{K}\widetilde{Z}_{iM_ik}\log\bar{\varphi}_k\Big(L_i-\kappa_{i,M_i-1};\boldsymbol{\beta}_k\Big)\right]\\ &+\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\sum_{k=1}^{K}Z_{i\ell k}\log f_k(\boldsymbol{c}_{i\ell};\boldsymbol{\theta}_k),\end{aligned}$$

通过生存函数进入的终端片段，因为仅知道其观测长度$L_i-\kappa_{i,\ M_i-1}$是已知的。污染高斯和 t 版本会在完整数据对数似然中加入涉及潜变量 $\Upsilon_{i\ell}$ 或 $H_{i\ell}$ 附录A.1中的项。

3.2.1 E 步

在当前参数值 $\tilde{\boldsymbol{\Theta}}^{(m)}$ 下，令 $F_i(r,\ k)$、$1\le r<L_i$ 表示截至时间 r 的观测值与在 r 时刻的一个已完成的状态-k 片段的联合密度。由于 $F_i(0,\ k)$ 未定义（初始片段由 $\boldsymbol{\rho}$ 处理），

$$\begin{aligned}F_i(r,k) = &\sum_{v=1}^{\min(r,V_{\max})}\varphi_k^{(m)}(v)\Psi_i^{(m)}(r,v;k)\\ &\times\left[\mathbf{1}\{r=v\}\rho_k^{(m)} + \mathbf{1}\{r>v\}\sum_{h\neq k}F_i(r-v,h)p_{hk}^{*(m)}\right],\qquad r<L_i.\end{aligned}\tag{7}$$

在观测边界处，没有观测到状态退出。因此，我们使用一个单独的终端量，

$$\begin{aligned}\widetilde{F}_i(L_i,k) = &\sum_{v=1}^{\min(L_i,V_{\max})}\bar{\varphi}_k^{(m)}(v)\Psi_i^{(m)}(L_i,v;k)\\ &\times\left[\mathbf{1}\{L_i=v\}\rho_k^{(m)} + \mathbf{1}\{L_i>v\}\sum_{h\neq k}F_i(L_i-v,h)p_{hk}^{*(m)}\right].\end{aligned}\tag{8}$$

序列似然为 $\mathcal{L}_i(\tilde{\boldsymbol{\Theta}}^{(m)}) = \sum_{k=1}^{K}\widetilde{F}_i(L_i,\ k)$。其中，$V_{\max}$ 是递推中允许的最大观测分段长度；对于 NP 持续时间，它是其有限支撑集的上界。如果持续时间分布不是有限支撑的，则 $V_{\max} = \max_i L_i$ 足以进行似然评估，因为无法观测到更长的已完成持续时间。直接计算时，分段递推的计算成本为每个序列 $O(L_iV_{\max}K^2)$。预先计算入口

权重 $E_i(r,k) = \sum_{h \neq k} F_i(r,h) p_{hk}^{*(m)}$ 在成本为 $O(L_i K^2)$ 的情况下对所有 (r,k) 进行计算，并在 v 中增量累积对数 $\Psi_i^{(m)}(r,v;k)$，可将其降至 $O(L_i K^2 + L_i V_{\max} K)$；相比之下，隐马尔可夫递推的成本为 $O(L_i K^2)$，因此 $V_{\max}$ 决定了半马尔可夫层的额外成本。对于无界参数化持续时间族，计算截断值 $V_{\max} < L_i$会省略可行分段，因此这是一种近似，而不是重新归一化的持续时间分布。可以通过增大截断值来评估其影响；与嵌套隐马尔可夫模型进行似然比较时，需要采用兼容的持续时间支撑集约定。在单核上，对于长度为 10,000 的 $V_{\max} = 640$序列，半马尔可夫递推耗时 0.5 秒；在 $V_{\max} = 50$时，一次半马尔可夫 EM 迭代的耗时约为一次隐马尔可夫迭代的两到三倍（附录 E）。

后向量 $G_i(r,k)$ 表示在状态 k 的一个分段于时间 r 结束的条件下，从时间$r+1$到 L_i 的观测值的条件似然。对于 $r = L_i - 1, \ldots, 1$,

$$G_i(r,h) = \sum_{k \neq h} \sum_{v=1}^{\min(V_{\max}, L_i - r)} p_{hk}^{*(m)} \varphi_k^{\dagger(m)}(v \mid r+v) \Psi_i^{(m)}(r+v, v; k) G_i(r+v, k),$$

其中 $G_i(L_i, k) = 1$，而终端片段从 $r+1\tilde{}$ 开始的项 $v = L_i - r$ 则由生存函数加权。末端前向量 $F_i(L_i, k)$ 仅用于计算似然，而内部量 $F_i(r,k)$ 则用于后验递推。状态 k 下一个长度为 u、于时间 r 结束的分段的后验概率为

$$\chi_i^{(m)}(r,u,k) = \frac{\varphi_k^{\dagger(m)}(u \mid r) \Psi_i^{(m)}(r,u;k) G_i(r,k)}{\mathcal{L}_i(\tilde{\boldsymbol{\Theta}}^{(m)})} \times \left[\mathbf{1}\{r = u\} \rho_k^{(m)} + \mathbf{1}\{r > u\} \sum_{h \neq k} F_i(r-u, h) p_{hk}^{*(m)} \right],$$

因此，$\chi_i^{(m)}(L_i, u, k)$ 是序列在状态 k 的一个分段内部结束，且该分段的观测长度为 u 的后验概率；而 $\chi_{i0}^{(m)}(u,k) = \chi_i^{(m)}(u,u,k)$ 是首个分段的状态为 k 且持续时间为 u 的后验概率，其中情形 $u = L_i$ 对应首个分段同时也是末个分段的序列。

对于 $r > u$，我们 定义从 h 到 k 的嵌入式转移的后验概率，随后

是持续时间为 u 、结束于 r 的分段，记为

$$\tilde{\omega}_i^{(m)}(r,u,h,k) = \frac{F_i(r-u,h)p_{hk}^{*(m)}\varphi_k^{\dagger(m)}(u \mid r)\Psi_i^{(m)}(r,u;k)G_i(r,k)}{\mathcal{L}_i(\tilde{\boldsymbol{\Theta}}^{(m)})}, \qquad h \neq k.$$

时刻 ℓ 的后验状态概率通过对所有覆盖 ℓ 的后验片段求和得到，即

$$\gamma_{i\ell k}^{(m)} = \Pr\left(S_{i\ell} = k \mid \mathcal{D}_i; \tilde{\boldsymbol{\Theta}}^{(m)}\right) = \sum_{r=1}^{L_i} \sum_{\substack{v=1,\ldots,\min(r,V_{\max}) \\ r-v+1 \leq \ell \leq r}} \chi_i^{(m)}(r,v,k).$$

后验片段铺满 1、……、 L_i，因此 $\sum_{k=1}^{K}\gamma_{i\ell k}^{(m)} = 1$ 对每个 ℓ均成立。对于 C-funHDDC 和 T-funHDDC 发射分布，潜变量的后验期望 $v_{i\ell k}^{(m)}$ ， $h_{i\ell k}^{(m)}$ 和 $e_{i\ell k}^{(m)}$ 根据状态条件发射分布计算（附录 A.1）。

3.2.2 M步

初始概率使用第一段的后验分布更新，$\rho_k^{(m+1)} = N^{-1}\sum_{i=1}^{N}\sum_{v=1}^{\min(L_i\ V_{\max})}\chi_{i0}^{(m)}(v,k)$。根据从状态 h 转移到状态 k 的嵌入式转移的后验期望次数，

$$\widehat{N}_{hk}^{(m)} = \sum_{i=1}^{N}\sum_{r=2}^{L_i}\sum_{v=1}^{\min(r-1,V_{\max})}\tilde{\omega}_i^{(m)}(r,v,h,k), \qquad p_{hk}^{*(m+1)} = \frac{\widehat{N}_{hk}^{(m)}}{\sum_{q\neq h}\widehat{N}_{hq}^{(m)}}, \quad h \neq k,$$

其中 $p_{hh}^{*(m+1)} = 0$。持续时间数据在 $r = L_i$ 处分开。令

$$\widehat{M}_k^{(m)}(u) = \sum_{i=1}^{N}\sum_{r=u}^{L_i-1}\chi_i^{(m)}(r,u,k), \qquad \widehat{M}_k^{\mathrm{cens},(m)}(u) = \sum_{i=1}^{N}\chi_i^{(m)}(L_i,u,k)$$

分别为状态 k 中持续时间已知等于 u以及至少为 u的分段数的后验期望。每个序列恰好包含一个终端片段，因此 $\sum_k\sum_u \widehat{M}^{\mathrm{cens}}{}_{,k}^{(m)}(u) = N$。对于参数化持续时间族，通过最大化右删失样本的加权持续时间对数似然来更新持续时间参数，

$$\boldsymbol{\beta}_k^{(m+1)} = \arg\max_{\boldsymbol{\beta}_k}\sum_u\left[\widehat{M}_k^{(m)}(u)\log\varphi_k(u;\boldsymbol{\beta}_k) + \widehat{M}_k^{\mathrm{cens},(m)}(u)\log\bar{\varphi}_k(u;\boldsymbol{\beta}_k)\right].$$

对于支撑集为 $\{1,\ldots,V_{\max}\}$ 的非参数逗留模型，通过为每个右删失终端片段补入其未观测到的完整持续时间，并使用上文定义的期望计数 $\widehat{M}_k^{(m)}(v)$ 和 $\widehat{M}^{\text{cens}}{}_{,\ k}{}^{(m)}(u)$，来更新持续时间概率。给定一个观测长度为 u的删失片段，其完整持续时间等于$v \geq u$ 的概率为 $\Pr(V = v \mid V \geq u,\ \ B = k;\ \boldsymbol{\beta}_k^{(m)}) = \varphi_k^{(m)}(v)/\bar{\varphi}_k^{(m)}(u)$，因此，状态-$k$ 中完整持续时间为 v 的片段的增广期望计数为

$$\widehat{C}_k^{(m)}(v) = \widehat{M}_k^{(m)}(v) + \sum_{u=1}^{v} \widehat{M}_k^{\text{cens},(m)}(u) \frac{\varphi_k^{(m)}(v)}{\bar{\varphi}_k^{(m)}(u)}, \qquad v = 1,\ldots,V_{\max}. \tag{9}$$

在约束条件 $\sum_{v=1}^{V_{\max}} \varphi_k(v) = 1$ 下最大化相应的增广期望完整数据对数似然，得到

$$\varphi_k^{(m+1)}(v) = \frac{\widehat{C}_k^{(m)}(v)}{\sum\limits_{w=1}^{V_{\max}} \widehat{C}_k^{(m)}(w)}, \qquad v = 1,\ldots,V_{\max}. \tag{10}$$

只要相关的生存函数概率为正，这就是增广期望最大化框架下持续时间概率的精确 M步。当不存在末端删失时，更新结果简化为已完成分段持续时间的归一化期望计数。

发射参数更新对于 HSMM 和 HMM 规范是通用的，因为时间依赖性通过后验状态概率 $\gamma_{i\ell k}^{(m)}$ 进入这些更新。对于高斯发射分布，利用这些概率作为权重来估计状态特定位置和散布参数。对于污染高斯和多元-t 发射分布，更新还涉及污染指标或潜在尺度的条件期望，由此产生观测特定的精度权重，从而减小异常函数型观测对估计位置和主子空间的影响。

4 模拟研究

我们研究时间依赖性是否能改善状态恢复，显式持续时间是否能在几何驻留时间之外提供额外信息，以及状态特定函数子空间是否优于全局函数主成分分析。两项实验分别生成马尔可夫和半马尔可夫潜在路径，并通过其逐点最大后验（MAP）标签的调整兰德指数（ARI）（Hubert & Arabie1985）在每个场景的相同100个数据集上评估每种方法。我们比较独立同分布（可交换）混合模型和适当时间模型下的高斯、污染高斯（C）和多元-t （T）发射分布。模型根据时间层、发射分布和逗留时间命名，因此 T-funHDDC 是带有 t 发射分布的独立同分布混合模型，HMM-C-funHDDC 是带有污染高斯发射分布的 HMM，而 HSMM-T-funHDDC（P）是带有 t 发射分布和 P逗留时间的 HSMM。全局-FPCA 竞争方法包括用于高斯、。

4.1 设计与比较

每个数据集由 $N = 25$ 个独立的三元（$p = 3$）曲线序列组成，定义在 $[0, 1]$,上，每个分量展开为 $R_s = 15$ 个正交归一傅里叶函数，因此 $R = 45$且 $W = I_R$。条件于状态 k，典型曲线的系数由（3）生成，

$$\boldsymbol{c}_{i\ell} = \boldsymbol{\mu}_k + U_k \boldsymbol{\delta}_{i\ell} + \boldsymbol{\varepsilon}_{i\ell}, \qquad \boldsymbol{\delta}_{i\ell} \sim N\Big(\mathbf{0}, (a_k - b_k) I_{d_k}\Big), \qquad \boldsymbol{\varepsilon}_{i\ell} \sim N(\mathbf{0}, b_k I_R),$$

使得主尺度特征值在一个状态内相等，$a_{k1} = \cdots = a_{kd_k} = a_k$。载荷矩阵 U_k 由一个公共正交归一框架的前 d_k 列组成；该框架由三个分量中依次选取的最低频非恒定傅里叶模态构成，因此状态子空间既平滑、嵌套，维度又各不相同。在特征值相等的情况下

只有投影算子 $U_k U_k^\top$ 是可识别的（引理 2，附录 D），拟合模型采用一般结构 $[a_{kj}, b_k, U_k, d_k]$，其中包含生成模型。每个状态均值都有三个非零系数，幅度为 $\mu/\sqrt{3}$；每个分量的一个低频模态上各有一个系数，因此 $\|\boldsymbol{\mu}_k\| = \mu$。两个状态均值在两个或三个分量上存在差异，距离为 $2\mu/\sqrt{3}$ 或 $\sqrt{2}\,\mu$，因此不存在能够区分所有状态的单一分量，而分离程度 μ 决定了单条曲线所包含的状态信息量。污染作用于整条曲线。无论状态路径如何，每条曲线以概率 $1-\alpha$成为异常曲线，其系数随后从 $N_R(\boldsymbol{\mu}_k, \eta\Sigma_k)$ 中抽取，从而使其所有标准差扩大 $\sqrt{\eta}$倍。情景 A 为无污染（$\alpha = 1$）。情景 B 存在轻度污染（$\alpha = 0.95$、 $\eta = 5$），情景 C 存在严重污染（$\alpha = 0.85$、$\eta = 10$），因此异常曲线分别占 5% 和 15%。因此，C-funHDDC 发射分布设定正确，T-funHDDC 发射分布是重尾近似，而高斯发射分布设定错误。

全局-FPCA 竞争方法将合并后的白化系数投影到其前 q个主成分上，其中 q 所选取的数量最少，但能够解释至少 90% 的方差（模拟 I 中为 $q = 30$–34 ，模拟 II 中为 32–35 ）。该方法使用与配对 funHDDC 模型相同的发射分布和时间层对得分进行建模，但每个状态采用无约束协方差；因此，两个模型在整个降维步骤上存在差异：前者对所有 R 系数使用状态特定的低秩加各向同性尺度矩阵，而后者采用全局 q维投影，舍弃末尾的 $R-q$ 维度，并使用无约束的状态协方差。由于数据是根据 funHDDC 结构生成的，这一比较衡量的是该结构相对于全局 FPCA 表示的优势，而非仅仅是状态特定子空间的影响。其似然定义在得分上，因此其贝叶斯信息准则（BIC；Schwarz 1978年）不与 funHDDC 的贝叶斯信息准则进行比较。所有方法均采用多起点方案，包括针对 t 以及

表 1：模拟 I：100 次重复试验的平均调整兰德指数（标准差）。†非退化拟合的均值（按阅读顺序分别为 86、70 和 69 次重复试验）。

方法	情景 A	场景B	场景 C
funHDDC	0.768 (0.026)	0.549 (0.056)	0.349 (0.089)
T-funHDDC	0.766 (0.026)	0.686 (0.036)	0.539 (0.071)
C-funHDDC	0.767 (0.027)	0.761 (0.031)	0.719 (0.031)
HMM-funHDDC	0.969 (0.012)	0.676 (0.051)	0.440 (0.103)
HMM-T-funHDDC	0.969 (0.011)	0.941 (0.025)	0.886 (0.076)
HMM-C-funHDDC	0.969 (0.011)	**0.970 (0.011)**	**0.961 (0.014)**
FPCA-GMM	0.669 (0.060)	0.312 (0.135)†	0.292 (0.076)†
函数主成分分析-t MM	0.646 (0.053)	0.540 (0.071)	0.348 (0.089)
FPCA-CMM	0.648 (0.062)	0.634 (0.074)	0.592 (0.039)
FPCA-HMM	0.954 (0.019)	0.880 (0.095)	0.402 (0.116)†
FPCA-HMM-t	0.949 (0.022)	0.927 (0.028)	0.856 (0.083)
函数主成分分析-隐马尔可夫模型-C	0.951 (0.020)	0.948 (0.019)	0.935 (0.021)

污染发射分布（附录 C.2）、细化迭代，以及从拟合的隐马尔可夫模型开始对每个 HSMM 进行热启动；其中 K 和 d_k 固定为其真实值，以分离时间效应和排放效应（附录 F.1）。调整兰德指数根据真实状态计算，并在 N 个序列上合并汇总（附录 C.4）。由于所有模型都在相同数据集上进行拟合，因此两个模型之间的差异通过重复实验上的配对 95% t区间进行评估；除非下文将其描述为可忽略，否则所讨论的每个差异的区间均不包含零。表 *1*中标记为 † 的均值来自不同的非退化重复实验子集，因此不定义配对区间，仅进行描述性比较。

4.2 模拟 I：马尔可夫持久性和稳健发射分布

状态路径是一个具有 $K = 4$ 个状态、序列长度为 L =40、间隔为 $\mu = 2$ 的齐次马尔可夫链，其中 $\boldsymbol{d} = (2, 4, 6, 8)$、 $\boldsymbol{a} = (6, 4, 3, 2)$ 和 $\boldsymbol{b} = (0.5, 0.35, 0.25, 0.2)$。每个状态以概率 $p_{kk} = 0.92$ 持续，离开概率质量向循环前向邻居倾斜，其中 $p_{k,\,k+1} \approx 0.053$ （索引模 K），而其他两个状态各约为 0.013。转移矩阵是双随机的，且该链从其均匀平稳分布开始，因此游程的平均长度为 12.5，每个序列平均包含 3.1 次状态变化（附录F.2中的图3展示了这些曲线）。本实验旨在量化真实持续时间分布为几何分布时依赖性的收益，而非展示隐半马尔可夫模型的优势。表 1 报告了主要比较结果。在干净数据上，隐马尔可夫模型将调整兰德指数从约 0.77 提高到 0.97。在严重污染下，高斯隐马尔可夫模型降至 0.440，而 t 和污染高斯隐马尔可夫模型分别达到 0.886 和 0.961。因此，时间依赖性无法弥补严重的发射分布设定错误。在所有情形下，对于两种稳健隐马尔可夫模型，状态特定子空间均优于全局函数主成分分析，尽管这种差异小于可交换模型与时间依赖模型之间的差异。设定错误的高斯比较并非始终有利，并且某些全局函数主成分分析高斯拟合是退化的（附录F.2）。

图 5 展示了场景B中重复试验1的前3个序列的最大后验标签。隐马尔可夫拟合能够重现真实的状态游程，而独立同分布拟合则将其切割成零散片段。在整个重复试验中，T-funHDDC和C-funHDDC分别将119 和92 条1000 条曲线误分类，而相应的隐马尔可夫拟合仅误分类15和5 条。

4.3 模拟II：非几何持续时间

状态路径按照第2.4节所述逐段生成，包含 $K = 3$ 个状态，序列长度为 $L = 70$， $\boldsymbol{d} =$ (2, 4, 6)， $\boldsymbol{a} =$ (6, 4, 2)， $\boldsymbol{b} =$ (0.5, 0.35, 0.2)。第一个分段从时间点1开始，状态从均匀分布中抽取；对于 $h \neq k$，嵌入链的 $p^*_{hk} = 1/2$，持续时间服从移位泊松分布 $V_{ij} \mid B_{ij} = k \sim 1 + \text{Pois}(\lambda_k)$。由 $L = 70$ 条曲线组成的窗口会对最后一个分段进行右删失。在场景 A–C 中， $\mu = 4$ 和$\boldsymbol{\lambda} =$ (6, 10, 17)，对应的平均持续时间为7、11和18，变异系数分别为0.35、0.29和0.23；相比之下，具有相同均值的几何持续时间的变异系数为0.93–0.97。在 $\mu = 4$ ，大多数干净曲线都能根据自身形状进行分类。因此，作为显式持续时间建模最具信息量的检验，场景 D 将干净重叠发射分布（$\mu = 1.5$）与较短的持续时间相结合， $\boldsymbol{\lambda} =$ (3, 5, 8)，平均持续时间为4、6和9，每个序列约有11个分段（附录F.3中的图 4展示了这些曲线）。每种发射分布都分别采用独立同分布混合模型、隐马尔可夫模型和隐半马尔可夫模型进行拟合，并使用三种逗留时间分布族：设定正确的移位泊松分布（P）、大小参数为 $r_k \in [0.05, 50]$的移位负二项分布（NB），以及定义在{1, . . . , $V_{\max}$}上的非参数分布（NP）。我们取 $V_{\max} = 32$，既作为NP逗留时间的支撑集上界，也作为P和NB逗留时间的计算截断值（第3.2.1节）；它会排除场景 A 中最长状态的持续时间的一部分 7.5×10^{-4}，且满足$V_{\max} \leq L$，符合命题 1（附录 D）中的有限时域假设。表2报告场景 D 和 C 的结果，附录F.3中的表 5则给出完整比较。

在场景 D 中，逗留时间分布设定正确的泊松分布 HSMM，相较于 HMM 将调整兰德指数提高了 0.016–0.017，且这一结果在不同发射分布下均一致。虽然这一提升小于引入时间依赖性带来的增益，但当个体曲线难以判别时，这一提升仍不可忽视。图 6 展示了场景 D 中重复试验1的前两个序列的最大后验标签。在整个重复试验中，T-funHDDC 将 181 条 1750 条曲线错分，HMM- T-funHDDC 错分 46 条，而采用 P 、 NB 和 NP 逗留时间的 HSMM-T-funHDDC 分别错分 33、34 和 38条。在

表2：模拟 II：100 次重复试验的选定平均调整兰德指数（标准差）。各行表示时间层，Mixture 对应独立同分布混合模型，t (T) 或污染高斯（C）发射分布，以及 P逗留时间。

	场景 D（干净重叠）		场景 C（严重污染）	
方法	funHDDC	全局函数主成分分析	funHDDC	全局函数主成分分析
混合-T	0.755 (0.019)	0.609 (0.030)	0.770 (0.022)	0.522 (0.045)
HMM-T	0.925 (0.012)	0.858 (0.020)	0.975 (0.015)	0.958 (0.053)
HSMM-T（P）	0.941（0.010）	0.894（0.016）	0.983（0.006）	0.977（0.007）
混合-C	0.762 (0.018)	0.621 (0.032)	0.913 (0.013)	0.869 (0.016)
HMM-C	0.926 (0.012)	0.865 (0.020)	0.991 (0.004)	0.984 (0.005)
HSMM-C（P）	0.942 (0.009)	0.897 (0.016)	0.993（0.004）	0.989（0.005）

污染情况下，采用稳健发射分布带来的额外增益，对于 t 至多为 0.009；对于污染高斯则仅为可忽略的 0.002，因为其 HMM 已几乎实现完美分类。在场景 D 中，即使采用 HSMM 推断，状态特定子空间相较于全局函数主成分分析的优势仍然存在（调整兰德指数约高出 0.04–0.05），且大于持续时间带来的增量收益。与非参数族相比，设定正确的移位泊松族能更准确地恢复持续时间分布；非参数族虽更灵活，却要付出显著的方差代价：在场景 D 中，HSMM-T-funHDDC 拟合平均持续时间的绝对误差为 0.208 条曲线，拟合概率质量函数与真实值之间的全变差距离为 0.036（采用 P 逗留时间时，两者均为对状态和重复试验取平均的结果）；采用 NP 逗留时间时，这两个数值分别为 0.222 和 0.168（图 7）。所有持续时间恢复估计（表 6）和负二项比较见附录 F.3。

只要发射分布能够容纳异常曲线，贝叶斯信息准则就能恢复生成结构。在发射分布相同的 funHDDC 系列模型中，它会选择隐半马尔可夫模型

在干净场景A和D的每个重复试验中，无论采用何种发射分布，都会选择具有 P逗留时间的模型；在污染场景B和C中，采用 t 和污染高斯发射分布时也会如此。它从不选择独立同分布混合模型或 NP逗留时间；在模拟 I 中，每个重复试验都选择隐马尔可夫模型而非独立同分布混合模型。然而，在污染情况下使用设定错误的高斯发射分布时，贝叶斯信息准则在每个重复试验中都优先选择隐马尔可夫模型，而不是任何隐半马尔可夫模型。由于截断值 $V_{\max} = 32$可能使隐半马尔可夫模型的似然低于嵌套隐马尔可夫模型的似然（附录F.3），这种偏好并不能说明逗留时间分布。在隐半马尔可夫模型中，贝叶斯信息准则更倾向于负二项分布，而非 P逗留时间，尽管真实持续时间服从 P分布。额外的离散参数吸收了异常曲线造成的部分失配，因此，过度离散的逗留时间并不一定意味着持续时间为非几何分布；第5.1节的空气质量数据中也再次出现了这一警示。边际熵积分完备似然（ICL，附录C.3）作出的选择相同，但在这200 个高斯重复试验中的三个例外中，面对相差一两个单位的贝叶斯信息准则数值，ICL 更倾向于选择隐半马尔可夫模型（NB），而非隐马尔可夫模型。

5 应用

我们分析米兰地区的每日污染剖面和连续智能手机活动记录，这两类有序函数型数据集具有不同的持续性机制；在留出序列上进行分析，其中 K 为保证可解释性而固定。我们比较可交换混合、采用稳健发射分布的隐马尔可夫模型和隐半马尔可夫模型，以及它们对应的全局函数主成分分析基准；其中 q 根据第90% 规则选取，见第 4.1节；我们使用训练序列上的贝叶斯信息准则和留出对数似然进行比较，即在拟合参数下留出序列的对数似然。内在维数通过 Cattell 碎石检验（Cattell 1966）选取，阈值为0。2 ，在每次拟合的前十次迭代中进行（附录 C.3）。外部标签仅用于描述性诊断。作为时间层的基准，我们还对独立同分布混合模型的最大后验标签应用滑动众数，将每个标签替换为

中心窗口内出现频率最高的标签；窗口长度为15 天（米兰）或5 个窗口（HAR）；这些宽度事先固定，未根据留出数据进行调整。预处理细节以及站点、受试者和基底层面的结果见附录 G。

5.1 米兰地区的空气质量状态

每日 NO_2 和 O_3 四个监测站的浓度曲线以三次 B 样条表示（每种污染物使用 $R_s = 11$ 条）。训练期为 2016–2022，2023–2026 年的数据留作测试。对于任一污染物，有效小时数少于 18 小时的日期仍保留在日历序列中，并作为缺失曲线处理，不贡献发射概率（附录 C.1）。每个站点序列的首个游程为左删失，并遵循附录 C.1中的剩余持续时间分布。P 和 NB 停留时间均以 $V_{\max} = 366$ 天为截断值进行评估（第 3.2.1节），该值高于最长的已完成 MAP 运行段（227 天；附录 G.1）；将米兰地区表 3 和 7 中的 HSMM 拟合重新评估，并令 $V_{\max}$ 等于序列长度后，其训练期和留出对数似然保持不变。两个状态（$K = 2$）分别描述一种以 NO_2为主、早晚峰值明显的模式，以及一种以 O_3为主、下午峰值明显的模式（图 1）。

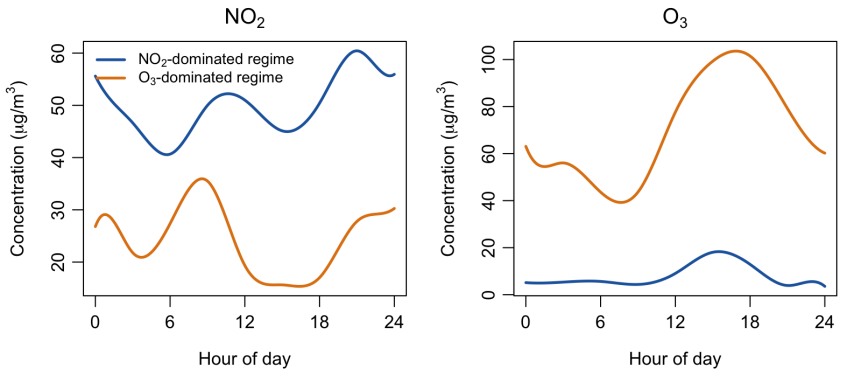

图 1：米兰：HSMM-T-funHDDC（负二项分布）下 NO_2 和 O_3 ($\mu g/m^3$) 的状态均值曲线。隐马尔可夫模型曲线无法区分。

在同质　　时间模型中，HSMM-T-funHDDC（NB）提高了留出

表3：米兰空气质量（$K = 2$）：2016–2022年的贝叶斯信息准则，以及留出年份2023–2026年的对数似然。Cattell检验选择 $d_k = (2,\ 2)$，全局函数主成分分析使用 $q = 5$ 得分。贝叶斯信息准则⌢报告为 $\Lambda(\Theta)\ -(\tau/2)$ 对数 $n_{\bullet}$（越大越好）。最佳值以粗体显示。

	BIC		留出对数似然	
时间层	t	污染	t	污染
独立同分布混合模型	−238845	−247900	−123873	−129443
HMM	−235451	−245099	−121895	−127813
半马尔可夫模型（P）	−240841	−247820	−125061	−130114
半马尔可夫模型（NB）	−235069	−244604	−121656	−127511
隐马尔可夫模型，供暖季转移	−235234	−244805	−121807	−127696
隐马尔可夫模型，月度转移	**−234876**	**−244376**	**−121570**	**−127495**

与供暖季日历的一致性（调整兰德指数）

	funHDDC		全局函数主成分分析	
	t	污染	t	污染
独立同分布混合模型	0.304	0.269	0.268	0.256
独立同分布混合模型，滑动众数（15 个日期）	**0.386**	0.353	0.340	0.318
HMM	0.344	0.341	0.345	0.329
半马尔可夫模型（P）	0.370	**0.401**	**0.410**	**0.412**
半马尔可夫模型（NB）	0.340	0.343	0.340	0.327

HMM-T-funHDDC 的对数似然从 $-121,\ 895$ 增至 $-121,\ 656$，而采用 P逗留时间的 HSMM 表现更差（表3）；未拟合 NP逗留时间，因为尚未实现其左删失版本（附录 C.1）。拟合得到的负二项分布大小参数，$r_k = 0.31$ 和0。15，远低于几何分布的大小参数 $r_k = 1$，表明存在显著的过度离散。然而，这并不能作为内在非几何持续性的决定性证据：负二项分布

大小参数小于 1 时，逗留时间是几何停留时间的混合分布，且风险率递减（附录 4 中的注释 G.1），因此，这些大小参数同样符合这样的解释：离开概率会因访问次数而异，这可能是季节性转移造成的，也可能是选择 $K = 2$后合并子状态造成的。表 3 还报告了转移矩阵取决于已知日历分组（供暖季，即 10 月 15 日至 4 月 15 日，或月份）的隐马尔可夫模型；而月份相关隐马尔可夫模型的拟合效果更好（$-121,570$ 留出对数似然），这表明齐次隐半马尔可夫模型在一定程度上吸收了季节性非平稳性。较长的暖季运行段与许多短暂片段并存，而简单的参数化逗留时间族都无法很好地描述这一现象（附录 G.1）。与供暖季日历的一致性较弱。

5.2 人体活动状态

我们使用人体活动与姿势转换（HAPT）数据的连续记录（Reyes-Ortiz 等人，2016），而非原始人体活动识别（HAR）数据集发布版本中按活动分组、采用 50% 重叠窗口的数据（Anguita 等人，2013），因为其文件顺序不适合估计状态持续时间。来自 30 名志愿者的 61 个连续会话被划分为互不重叠的 2.56 秒窗口，并表示为六分量函数型观测（三个加速度通道和三个角速度通道）。每个会话均按其标注实验方案截取，因此每个序列都从一段活动的起始时刻开始，并适用假设 1（A5）的边界起始约定。三种逗留时间均使用 $V_{\max} = 161$ 个窗口，即最长会话的长度，因此对于 P 和 NB 逗留时间，每个序列使用 $\min(V_{\max}, L_i) = L_i$，其似然不涉及截断。NP 支持集 $\{1, \dots, 161\}$ 超过最长训练会话的长度（155 个窗口），因此引理 3 中“不超过最长序列”这一条件在训练数据上不成立，且大于 $V_{\max}$ 的概率质量仅通过其总和进入训练似然。没有任何拟合运行段接近此上界：初始标签序列和拟合后标签序列中的最长运行段分别持续 99 和 89 个窗口（附录 C.2），因此这些概率质量仅在经过删失的末尾处获得权重

表 4：人类活动（$K = 3$）：42 个训练会话上的贝叶斯信息准则（BIC）和 19 个留出会话的对数似然。Cattell检验为运动、直立和躺卧分别选择 $d_k =$ (6，2，2) (t) 和 (8，2，2)（污染），全局函数主成分分析使用 $q = 25$ 个得分。BIC 报告为$\frown\Lambda(\Theta) - (\tau/2)$ 对数 $n_\bullet$（越大越好）。最佳值以粗体标出。

	BIC		留出对数似然	
时间层	t	污染	t	污染
独立同分布混合模型	-366503	-451985	-159925	-193116
HMM	-362960	-450283	-158609	-192362
半马尔可夫模型（P）	-364897	-451288	-159388	-192865
半马尔可夫模型（NB）	$\mathbf{-362639}$	$\mathbf{-450126}$	-158474	-192297
半马尔可夫模型（NP）	-364421	-451887	$\mathbf{-158452}$	$\mathbf{-192257}$

活动类别的留出集一致性（调整兰德指数）

	funHDDC		全局函数主成分分析	
	t	污染	t	污染
独立同分布混合模型	0.842	0.891	0.470	0.654
独立同分布混合模型，滑动众数（5 个窗口）	**0.889**	0.902	0.489	0.686
HMM	0.858	0.906	0.669	0.682
半马尔可夫模型（P）	0.839	0.871	**0.685**	**0.694**
半马尔可夫模型（NB）	0.855	0.906	0.683	0.690
半马尔可夫模型（NP）	0.846	**0.909**	**0.685**	0.692

游程。按受试者进行的标准划分得到来自21名受试者的42个训练会话，以及来自9名受试者的19个留出会话。我们设定三个可解释状态：运动、直立静态姿势和躺卧。跨越状态转换或未标注时段的窗口仍保留在序列中，但不纳入分类评分。

状态特定子空间表示提供了最清晰的实证优势：全局函数主成分分析模型的留出调整兰德指数约为0.47–0.69，而稳健函数型高维数据聚类模型达到0.84–0.91（表4）。时间依赖性改善了预测拟合，但分类增益有限。在 t 发射分布下，隐马尔可夫模型相对于独立同分布混合模型将留出对数似然提高了1,316，采用负二项分布和NP停留时间分布的隐半马尔可夫模型分别再增加135和157，然而五个窗口上独立同分布标签的滑动众数分类效果优于每一个时间模型（调整兰德指数为0.889，而后者至多为0.858）。拟合持续时间刻画的是实验方案，而非自然活动行为。分类结果因受试者而异，且首选的发射分布族取决于基函数维度（附录G.2 和G.3）。

5.3 稳健性与解释

敏感性分析（附录 G.3，表 7 和 8）区分了稳健结论与依赖于预处理的分类。在米兰，各站点单独来看，在留出似然方面均更倾向于半马尔可夫模型（NB）而非齐次隐马尔可夫模型，不过总体而言，日历依赖隐马尔可夫模型更优。在 HAR 中，重叠窗口夸大了表面上的马尔可夫贡献；较大的基函数维度可能导致污染高斯似然偏向与活动标签不太吻合的分类，而在所考虑的所有基底维度下，半马尔可夫模型（NB）的留出对数似然增益仍略小但为正，优于隐马尔可夫模型。

6 结论

我们构建了一系列用于纵向多元函数型观测聚类的隐半马尔可夫模型，在基于似然的方法框架内，将状态特定函数主子空间、高斯发射分布与稳健发射分布以及显式逗留时间分布结合起来，并将隐马尔可夫模型作为几何持续时间特例。参数估计利用低秩函数协方差表示和分段前向–后向

递推算法，而可辨识性结果（附录D）阐明了在何种条件下可以恢复函数参数和半马尔可夫参数。对潜在时间依赖进行建模，通常能比可交换函数聚类带来最大的改进，尤其是在个体曲线所包含的信息不足时。稳健发射分布能够保护分析免受异常观测的影响；采用稳健发射分布时，状态特定函数子空间保留了在全局降维下可能丢失的判别性变异；而在存在污染时，使用高斯发射分布和全局 FPCA 表示可能获得更好的分类效果。当状态重叠且其持续性偏离几何性时，显式持续时间分布能够带来额外收益，但这种改进并不普遍。应用结果证实了时间建模对预测拟合的价值，并说明了一个重要的限定：表面上的持续时间依赖性（例如空气质量状态中过度离散的逗留时间）可能反映了转移机制中未建模的时间变化，或反映了因 K状态选择而合并的状态。

在模拟中，状态数和内在维度均已固定；在应用中，则出于可解释性的考虑选择状态数。对于基于 funHDDC 结构生成的数据，全局 FPCA 比较并未单独分离状态特定子空间的影响（第 4.1节）。模拟 I 未拟合隐半马尔可夫模型，因此未评估在几何停留时间下选择显式持续时间模型的比率；模拟 II 的停留时间比较采用了计算截断值 $V_{\max} = 32$。污染高斯分类可能对基函数维度和异常观测加权较为敏感；在干净数据上，许多污染高斯拟合达到迭代上限，此时污染参数的可识别性较弱。有限支撑非参数停留时间分布未与左删失初始化结合使用，且可识别性结果不涵盖左删失窗口。报告的 ICL 使用边际状态熵，相较于独立同分布混合模型，这使其对时间

模型而言较为保守。最后，竞争模型集合并未涵盖现有的所有函数型隐状态表述。

仍有若干扩展方向值得研究。依赖日历的半马尔可夫动态将有助于区分非几何持续时间与季节性非平稳性。更灵活但保持简约的持续时间分布、异方差残差协方差结构、函数型子空间参数化之间的自动选择、观测单位之间的依赖关系、函数型重构中的不确定性，以及面向大型纵向数据集的可扩展推断，也都值得进一步研究。

披露声明

作者声明不存在需要披露的利益冲突。

参考文献

Allman, E. S.、Matias, C. 与 Rhodes, J. A.（2009年），‘具有多个观测变量的潜在结构模型中参数的可识别性’，*The Annals of Statistics***37**（6A），3099–3132。

Amovin-Assagba, M.、Gannaz, I. 与 Jacques, J.（2022），‘通过污染混合模型检测多元函数型数据中的异常值’，《计算统计与数据分析》**174**，107496。

Anguita, D., Ghio, A., Oneto, L., Parra, X. & Reyes-Ortiz, J. L.（2013年），《使用智能手机进行人体活动识别的公共领域数据集》，载于《第21届欧洲人工神经网络、计算智能与机器学习研讨会论文集（ESANN 2013）》，比利时布鲁日，第437–442页。

Anton, C. & Smith, I.（2024）, 《通过混合 t分布对函数型数据进行基于模型的聚类》, 《数据分析与分类进展》18（3）, 563–595。

Bartolucci,F., Farcomeni, A.& Pennoni, F. (2012), 纵向数据的隐马尔可夫模型, CRC出版社.

Bernardi, M., Maruotti, A. & Petrella, L. (2017), '多元动态重尾模型的多种风险度量', 《实证金融学杂志》43, 1–32.

Biernacki, C., Celeux, G.& Govaert, G. (2000), '使用积分完备似然评估用于聚类的混合模型', 《*IEEE*模式分析与机器智能汇刊》22(7), 719–725.

Biernacki, C., Celeux, G. & Govaert, G. (2003), '为期望最大化算法选择初始值, 以获得多元高斯混合模型的最高似然',《计算统计与数据分析》41(3-4), 561–575.

Bigi, A. & Ghermandi, G. (2014), '波河谷大气 PM_{10}浓度的长期趋势与变异性', 《大气化学与物理》14(10), 4895–4907.

Böhning, D., Dietz, E., Schaub, R., Schlattmann, P. & Lindsay, B. G. (1994), '来自单参数指数族的密度混合的似然比的分布', 《统计数学研究所年刊》46(2),373–388.

Bouveyron, C., Girard, S. & Schmid, C. (2007), '高维数据聚类', 计算统计与数据分析 52(1), 502–519.

Bouveyron, C. &Jacques, J. (2011), '组特异函数子空间中的时间序列基于模型的聚类', 《数据分析与分类进展》5(4), 281–300.

Cattell, R. B. (1966), '因子数量的碎石检验', 《多元行为研究》1(2), 245–276。

Cuesta-Albertos, J. A., Gordaliza, A. & Matrán, C. (1997), ‘修剪后的 k-均值：尝试使量化器更稳健’, *The Annals of Statistics* **25**(2), 553–576。

Dempster, A. P., Laird, N.M. & Rubin, D. B.(1977), ‘通过期望最大化算法从不完整数据中求最大似然’,英国皇家统计学会会刊 *B*辑：统计方法学**39**(1), 1–22.

Golini, N., Martella, F. & Maruotti, A. (2026), ‘公平且可持续的福祉动态分析：使用多变量简约隐马尔可夫模型对意大利 NUTS3 纵向数据进行聚类’, *AStA* 统计分析进展 第 1–75 页.

Guédon, Y. (2003), ‘根据离散序列估计隐半马尔可夫链’, *Journalof Computational and Graphical Statistics***12**(3), 604–639.

Holzmann, H., Munk, A. &Gneiting, T.(2006), ‘椭圆分布有限混合的可识别性’,斯堪的纳维亚统计学杂志 **33**(4), 753–763.

Hubert, L. & Arabie, P. (1985), ‘比较划分’, 分类学杂志 **2**(1), 193–218.

Jacques, J. & Preda, C. (2014), ‘函数型数据聚类：综述’, 《数据分析与分类进展》 8(3), 231–255.

Lu, Z. (2025), ‘纵向数据聚类：方法与软件包综述’,国际统计评论**93**(3), 425–458.

Maranzano, P. (2022), ‘意大利伦巴第的空气质量：ARPA Lombardia 环境监测系统概述’, 地球 **3**(1), 172–203.

Maranzano, P. & Algieri, A. (2024), 《ARPALData：用于获取和分析来自 ARPA Lombardia（意大利）的空气质量和天气数据的 R 软件包》, 环境与生态统计学 **31**(2), 187–218。

Martino, A., Guatteri, G. & Paganoni, A. M. (2020), 《用于多元函数型数据的隐马尔可夫模型》，统计与概率快报 **167**, 108917。

Maruotti, A.、Bulla, J.、Lagona, F.、Picone, M. 与 Martella, F.（2017），“用于刻画多元空气污染物暴露的因子分析器动态混合模型”，*The Annals ofApplied Statistics* **11**(3)，1617–1648。

Maruotti, A.、Punzo, A. 与 Bagnato, L.（2019），“含多元高峰态正态分量的隐马尔可夫和半马尔可夫模型：日收益率序列的稳健建模”，*Journal of Financial Econometrics* **17**(1)，91–117。

McNicholas, P. D.、Murphy, T. B.、McDaid, A. F. 与 Frost, D.（2010），“通过简约高斯混合模型实现基于模型的聚类的串行与并行实现”，《计算统计与数据分析》**54**(3), 711–723。

Meng, X.-L. 与 Rubin, D. B.（1993），“通过 ECM 算法进行最大似然估计：一个通用框架”，*Biometrika***80**(2)，267–278。

Park, J. & Ahn, J.（2017年），《具有相位变异的多元函数型数据聚类》，*Biometrics***73**（1），324–333。

Peel, D. & McLachlan, G. J.（2000），《使用 t 分布进行稳健混合建模》，*Statistics and Computing***10**（4），339–348。

Punzo, A. & Maruotti, A.（2016年），《多元纵向观测的聚类：污染高斯隐马尔可夫模型》，*Journal of Computational and GraphicalStatistics***25**（4），1097–1116。

Punzo, A. & McNicholas, P. D.（2016年），“多元污染正态分布的简约混合”，*Biometrical Journal***58**(6)，1506–1537。

Reyes-Ortiz, J.-L.,Oneto,L., Samà,A.,Parra,X.& Anguita, D.（2016年），“使用智能手机进行过渡感知人类活动识别”，*Neurocomputing***171**，754–767。

Schmutz, A., Jacques, J., Bouveyron, C., Chèze, L. & Martin, P. (2020年), ‘在组特定函数子空间中对多变量函数型数据进行聚类’, *Computational Statistics***35**(3), 1101–1131.

Schwarz, G. (1978年), ‘估计模型的维度’, *The Annals of Statistics***6**(2), 461–464.

Seinfeld, J. H. & Pandis, S. N. (2016年), 《大气化学与物理》, 第3版, Wiley-Blackwell, Hoboken，NJ.

Sidrow, E., Heckman, N., Fortune, S. M. E., Trites, A. W., Murphy, I. & Auger-Méthé, M. (2022), ‘利用隐马尔可夫模型对多尺度状态切换函数型数据建模’, *The Canadian Journal of Statistics* **50**(1), 327–356.

Yakowitz, S. J. & Spragins, J. D.（1968年），‘有限混合的可识别性’，*TheAnnals of Mathematical Statistics***39**(1)，209–214。

Yu, S.-Z.（2010），‘隐半马尔可夫模型’，*Artificial Intelligence* **174**(2)，215–243。

Zeldow, B., Flory, J., Stephens-Shields, A., Raebel, M. & Roy, J. A.（2021年），‘纵向数据的函数型聚类方法及其在电子健康记录中的应用’，*Statistical Methods in Medical Research***30**(3)，655–670。

Zhou, J., Zhang, Y. & Tu, W.（2023年），‘clusterMLD：一种用于多元纵向数据的高效层次聚类方法’，*Journal of Computational and GraphicalStatistics***32**(3)，1131–1144。

Zucchini, W., MacDonald, I. L. & Langrock, R.（2016年），《时间序列的隐马尔可夫模型：使用 *R* 入门》，第 2 版，Chapman and Hall/CRC。

A 发射分布与参数更新

A.1 潜在表示与 E 步相关量

第 2.3 节的三种发射分布共用 funHDDC尺度矩阵（4）。以下推导给出了稳健发射分布的潜变量表示、E 步中计算的后验期望，以及适用于隐马尔可夫模型和隐半马尔可夫模型的 M步更新。对于每个函数型观测，引入一个潜变量 $\Upsilon_{i\ell} \mid S_{i\ell} = k \sim$ 伯努利分布(α_k) 即可得到污染高斯发射分布，其中 $\Upsilon_{i\ell} = 1$表示正常观测， $\Upsilon_{i\ell} = 0$ 表示污染观测，且

$$\boldsymbol{c}_{i\ell} \mid \Upsilon_{i\ell} = 1, S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k, \Sigma_k), \qquad \boldsymbol{c}_{i\ell} \mid \Upsilon_{i\ell} = 0, S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k, \eta_k \Sigma_k).$$

这里， Σ_k 是普通成分的协方差矩阵，而完整状态协方差为 $\mathrm{Cov}(\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k) = (\alpha_k + (1 - \alpha_k)\eta_k)\,\Sigma_k$。当 α_k 接近 1 时，会得到近似纯高斯成分，而限制条件 $\eta_k > 1$ 则标记膨胀核。

对于多元-t 发射分布，当 $\nu_k > 1$ 时，位置参数 $\boldsymbol{\mu}_k$ 等于均值；当 $\nu_k > 2$ 时，协方差存在，且等于 $\nu_k \Sigma_k/(\nu_k - 2)$。因此，对于有限的 ν_k， Σ_k 本身并不是协方差矩阵，不过这两个白化矩阵的特征子空间相同。等价地， t 发射分布是一个尺度混合，其潜变量服从形状率参数化的伽马分布 $\mathrm{Gamma}(\nu_k/2,\ \nu_k/2)$，且 $\boldsymbol{c}_{i\ell} \mid H_{i\ell} = h_{i\ell},\ S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k,\ \Sigma_k/h_{i\ell})$。对于污染高斯发射分布，在当前参数值 $\boldsymbol{\Theta}^{(m)}$ 下，E 步会计算任一时间模型的

$$\upsilon_{i\ell k}^{(m)} = \mathbb{E}\left[\Upsilon_{i\ell} \mid S_{i\ell} = k, \boldsymbol{c}_{i\ell}; \boldsymbol{\Theta}^{(m)}\right]$$

as

$$\frac{\alpha_k^{(m)} \phi_R\left(\boldsymbol{c}_{i\ell}; \boldsymbol{\mu}_k^{(m)}, \Sigma_k^{(m)}\right)}{\alpha_k^{(m)} \phi_R\left(\boldsymbol{c}_{i\ell}; \boldsymbol{\mu}_k^{(m)}, \Sigma_k^{(m)}\right) + \left(1 - \alpha_k^{(m)}\right) \phi_R\left(\boldsymbol{c}_{i\ell}; \boldsymbol{\mu}_k^{(m)}, \eta_k^{(m)} \Sigma_k^{(m)}\right)}.$$

对于 t发射分布，它会计算

$$h_{i\ell k}^{(m)} = \mathbb{E}\left[H_{i\ell} \mid S_{i\ell} = k, \boldsymbol{c}_{i\ell}; \boldsymbol{\Theta}^{(m)}\right] = \frac{\nu_k^{(m)} + R}{\nu_k^{(m)} + \Delta_k^{(m)}(\boldsymbol{c}_{i\ell})},$$

and

$$e_{i\ell k}^{(m)} = \mathbb{E}\left[\log H_{i\ell} \mid S_{i\ell} = k, \boldsymbol{c}_{i\ell}; \boldsymbol{\Theta}^{(m)}\right]$$

as

$$\log h_{i\ell k}^{(m)} + \psi\left(\frac{\nu_k^{(m)} + R}{2}\right) - \log\left(\frac{\nu_k^{(m)} + R}{2}\right),$$

其中 $\psi(\cdot)$是双伽马函数。

A.2 M步

给定 $\gamma_{i\ell k}^{(m)}$ 第3.2.1 节 HSMM E步或附录B.2.1中的后验状态概率，定义 $n_k^{(m)} = \sum_i \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)}$。以下更新对两种时间模型均相同，唯一的区别在于生成 $\gamma_{i\ell k}^{(m)}$ 的 E 步不同。对于高斯funHDDC发射分布， $\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k \sim N_R(\boldsymbol{\mu}_k, \Sigma_k)$，均值的更新公式为

$$\boldsymbol{\mu}_k^{(m+1)} = \frac{\sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)} \boldsymbol{c}_{i\ell}}{n_k^{(m)}}.$$

接下来，将加权散布矩阵定义为 $\Omega_k^{(m+1)}$ ，其白化形式为 $\tilde{\Omega}_k^{(m+1)}$ ，其中

$$\Omega_k^{(m+1)} = \frac{1}{n_k^{(m)}} \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)} \left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{(m+1)}\right)\left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{(m+1)}\right)^{\top}.$$

以及 $\tilde{\Omega}_k^{(m+1)} = W^{1/2}\Omega_k^{(m+1)}W^{1/2}$。列 $\boldsymbol{u}_{k1}^{(m+1)}, \boldsymbol{u}_{k2}^{(m+1)}, \ldots, \boldsymbol{u}_{kd_k}^{(m+1)}$ 是对称矩阵 $\tilde{\Omega}_k^{(m+1)}$ 对应于其 d_k 最大特征值$\lambda_{k1}^{(m+1)} \geq \lambda_{k2}^{(m+1)} \geq \cdots \geq \lambda_{kd_k}^{(m+1)}$的标准正交特征向量。对于 $j = 1, \ldots, d_k$， 方差参数按 $a_{kj}^{(m+1)} = \lambda_{kj}^{(m+1)}$ 更新，并且

$$b_k^{(m+1)} = \frac{\operatorname{tr}\left(\tilde{\Omega}_k^{(m+1)}\right) - \sum_{j=1}^{d_k} \lambda_{kj}^{(m+1)}}{R - d_k}.$$

只需要前 d_k 个特征对以及 $\tilde{\Omega}_k^{(m+1)}$ 的迹。第 $O(Rd_k)$ 3.1 节的计算3.1 涉及发射概率计算，不包括散布矩阵构造和特征对计算，并且 $\Sigma_k^{(m+1)}$ 可由（4）得出。

对于 C-funHDDC，精度权重定义为

$$r_{i\ell k}^{C,(m)} = \upsilon_{i\ell k}^{(m)} + \frac{1-\upsilon_{i\ell k}^{(m)}}{\eta_k^{(m)}}.$$

然后，

$$\alpha_k^{(m+1)} = \frac{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)}\upsilon_{i\ell k}^{(m)}}{n_k^{(m)}},$$

在报告的 数值分析中，这些值受到 $\alpha_k \geq \alpha^* = 0.75$ （以及 $\alpha_k < 1$）的约束。

均值更新为

$$\boldsymbol{\mu}_k^{C,(m+1)} = \frac{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)} r_{i\ell k}^{C,(m)} \boldsymbol{c}_{i\ell}}{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)} r_{i\ell k}^{C,(m)}}.$$

散布矩阵随后变为

$$\Omega_k^{C,(m+1)} = \frac{1}{n_k^{(m)}}\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)} r_{i\ell k}^{C,(m)}\left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{C,(m+1)}\right)\left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{C,(m+1)}\right)^{\top}.$$

特征向量 a_{kj} 和 b_k 通过对 $\tilde{\Omega}_k^{C\,(m+1)}$, 应用相同的特征分解步骤获得。在固定 $\boldsymbol{\mu}^C$、$_k{}^{(m+1)}$ 和 $\Sigma_k^{(m+1)}$ 后，膨胀参数更新为

$$\eta_k^{(m+1)} = \max\left\{1+\varepsilon_\eta, \frac{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)}\left(1-\upsilon_{i\ell k}^{(m)}\right)\Delta_k^{(m+1)}(\boldsymbol{c}_{i\ell})}{R\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)}\left(1-\upsilon_{i\ell k}^{(m)}\right)}\right\},$$

其中，$\varepsilon_\eta > 0$ 是一个较小的常数。

对于 T-funHDDC，均值的更新方式为

$$\boldsymbol{\mu}_k^{t,(m+1)} = \frac{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)} h_{i\ell k}^{(m)} \boldsymbol{c}_{i\ell}}{\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\gamma_{i\ell k}^{(m)} h_{i\ell k}^{(m)}}.$$

散布矩阵为

$$\Omega_k^{t,(m+1)} = \frac{1}{n_k^{(m)}} \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)} h_{i\ell k}^{(m)} \left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{t,(m+1)}\right) \left(\boldsymbol{c}_{i\ell} - \boldsymbol{\mu}_k^{t,(m+1)}\right)^{\top}.$$

通过对$\tilde{\mathbf{\Omega}}_k^{t\,(m+1)}$,执行相同的特征分解步骤，可得到特征向量 a_{kj}和 b_k。通过求解以下方程更新自由度:

$$-\psi\left(\frac{\nu_k^{(m+1)}}{2}\right) + \log\left(\frac{\nu_k^{(m+1)}}{2}\right) + 1 + \frac{1}{n_k^{(m)}} \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)} \left(e_{i\ell k}^{(m)} - h_{i\ell k}^{(m)}\right) = 0.$$

在实际应用中，该方程在一个有界区间内通过数值方法求解。如果采用公共自由度参数 $\nu_1 = \cdots = \nu_K = \nu$，则通过求解以下方程得到 $\nu^{(m+1)}$:

$$-\psi\left(\frac{\nu^{(m+1)}}{2}\right) + \log\left(\frac{\nu^{(m+1)}}{2}\right) + 1 + \frac{1}{n_\bullet} \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \sum_{k=1}^{K} \gamma_{i\ell k}^{(m)} \left(e_{i\ell k}^{(m)} - h_{i\ell k}^{(m)}\right) = 0,$$

其中

$$n_\bullet = \sum_{i=1}^{N} L_i.$$

简约协方差变体的更新方式与 funHDDC 族中的相同，只需将普通混合责任度替换为 $\gamma_{i\ell k}^{(m)}$。例如，如果 b 在各状态间为公共参数，则

$$b^{(m+1)} = \frac{\sum_{k=1}^{K} n_k^{(m)} \left[\operatorname{tr}\left(\tilde{\Omega}_k^{(m+1)}\right) - \sum_{j=1}^{d_k} \lambda_{kj}^{(m+1)}\right]}{\sum_{k=1}^{K} n_k^{(m)} (R - d_k)}.$$

如果前 d_k 个方差参数在状态 k内为公共参数，则

$$a_k^{(m+1)} = \frac{1}{d_k} \sum_{j=1}^{d_k} \lambda_{kj}^{(m+1)}.$$

如果前 d_k 个方差参数在所有状态间均为公共参数，则

$$a^{(m+1)} = \frac{\sum_{k=1}^{K} n_k^{(m)} \sum_{j=1}^{d_k} \lambda_{kj}^{(m+1)}}{\sum_{k=1}^{K} n_k^{(m)} d_k}.$$

前述公式给出了无约束的候选更新。当方差参数在各状态间共享时，这些候选更新未必能维持模型所要求的信号-噪声分离

。进行数值估计时，我们施加 $b_k \geq \varepsilon_b$ 和 $a_{kj} \geq (1+\delta) b_k$，其中 $\varepsilon_b > 0$ 和 $\delta > 0$ 在整个拟合过程中保持固定。这些限制确保理论条件 $a_{kj} > b_k > 0$成立。若候选更新满足这些限制以及所选协方差变体所定义的等式，则接受该更新。否则，通过对期望完整数据对数似然中的协方差贡献进行约束最大化，联合更新方差参数。

B Hidd马尔可夫基准：模型与推断

隐马尔可夫模型（HMM）构成了第 2.4 节所介绍的函数隐半马尔可夫模型（HSMM）的几何持续时间特例。我们将描述这一限制及其基于似然的推断，以记录模拟研究和应用中所使用的基准模型。隐马尔可夫模型保留了第 2.1 节的纵向函数表示、第 2.2 节的状态特异性子空间结构，以及第 t 2.3 节的高斯、污染高斯和多元发射分布。差异仅在于支配潜在状态的随机过程。因此，两种时间模型的发射参数更新是相同的，汇总于附录 A 中。

B.1 隐马尔可夫表示

考虑以下 N 纵向函数序列

$$\mathcal{D}_i = \{\boldsymbol{X}_{i\ell} : \ell = 1, \ldots, L_i\}, \qquad i = 1, \ldots, N,$$

及其系数表示 $\boldsymbol{c}_{i\ell} \in \mathbb{R}^R$，如第2.1节所定义。令 $S_{i\ell} \in \{1, \ldots, K\}$表示与序列 ℓ中的第 i个函数型观测相关联的潜在状态。在

根据隐马尔可夫模型的规范，潜在状态构成齐次一阶马尔可夫链，

$$\Pr(S_{i\ell} = k \mid S_{i1}, \ldots, S_{i,\ell-1}) = \Pr(S_{i\ell} = k \mid S_{i,\ell-1}),$$

具有初始分布

$$\rho_k = \Pr(S_{i1} = k), \qquad \rho_k \geq 0, \qquad \sum_{k=1}^{K} \rho_k = 1,$$

以及转移矩阵 $P = (p_{hk})_{h,\,k=1}^{K}$，其中

$$p_{hk} = \Pr(S_{i\ell} = k \mid S_{i,\ell-1} = h), \qquad p_{hk} \geq 0, \qquad \sum_{k=1}^{K} p_{hk} = 1. \tag{B.1}$$

这些序列共享相同的模型参数，但被视为相互独立，序列末尾与另一个序列开头之间不存在转移。

与 HSMM 的联系源于对同一状态的连续访问次数的分布。对于 $p_{kk} < 1$，在发生转移之前处于状态 k 的连续观测次数服从几何分布

$$\varphi_k(v) = (1 - p_{kk}) p_{kk}^{v-1}, \qquad v = 1, 2, \ldots, \tag{B.2}$$

其生存函数为

$$\bar{\varphi}_k(v) = \Pr(V \geq v \mid B = k) = p_{kk}^{v-1}.$$

第 2.4 节中引入的嵌入链的转移概率为

$$p_{hk}^{*} = \frac{p_{hk}}{1 - p_{hh}}, \qquad h \neq k, \qquad p_{hh}^{*} = 0. \tag{B.3}$$

反之，对于具有几何驻留时间分布且各状态具有特定离开概率 $q_k \in (0, 1]$，的 HSMM，其对应的 HMM转移矩阵为

$$p_{kk} = 1 - q_k, \qquad p_{kh} = q_k p_{kh}^{*}, \quad h \neq k.$$

因此，这两种表示描述了相同的潜在状态过程；只要它们具有相同的初始分布和状态依赖发射密度，就会产生相同的观测数据分布。

只要将 HSMM 的终端片段视为右删失，这种等价性在有限观测窗口下也成立。具体而言，内部片段贡献几何质量（B.2），而观测长度为 v 的终端片段贡献生存函数 $\bar{\varphi}_k(v)$，无需在最后一次观测之后再进行转移。因此，相应的似然恰好等于固定窗口隐马尔可夫模型似然。几何分布具有无记忆性：离开状态 k 的概率为 $1-p_{kk}$，且与已处于该状态的时间无关。通过允许采用非几何驻留时间分布，HSMM 恰好放宽了这一限制，同时不改变函数型发射模型。

B.1.1 与可交换函数型混合模型的关系。

隐马尔可夫模型还包含普通可交换函数型混合模型这一特例。如果

$$p_{hk} = \pi_k, \qquad h, k = 1, \dots, K,$$

且 $\rho_k = \pi_k$，其中 $\pi_k \geq 0$ 且 $\sum_k \pi_k = 1$，那么潜在状态是独立同分布的，且

$$\Pr(S_{i\ell} = k) = \pi_k.$$

Cons因此，系数观测值的联合分布可分解为

$$p(\boldsymbol{c}_{i1}, \dots, \boldsymbol{c}_{iL_i}) = \prod_{\ell=1}^{L_i} \left\{ \sum_{k=1}^{K} \pi_k f_k(\boldsymbol{c}_{i\ell}; \boldsymbol{\theta}_k) \right\}.$$

因此，当保留依赖状态的发射分布但移除纵向依赖时，普通 funHDDC 混合模型及其稳健变体可作为可交换基准模型得到。

B.1.2 模型复杂度。

当初始分布和转移矩阵均自由估计时，隐马尔可夫模型时间层包含

$$\tau_{\text{HMM,temporal}} = (K-1) + K(K-1)$$

个自由参数，分别对应于初始概率和转移矩阵的 K 各行。如果将初始分布约束为不可约转移矩阵的唯一平稳分布，则它由 P确定，时间层的贡献便减少为 $K(K-1)$ 个参数。参数总数通过加上相关的发射分布参数数量得到；在适用时，还包括协方差共享约束、污染参数或自由度。这些计数在附录C.3中有所说明。

B.2 似然与期望最大化估计

Let

$$\boldsymbol{\Theta}_{\mathrm{H}} = \{\boldsymbol{\rho}, P, \boldsymbol{\theta}_1, \dots, \boldsymbol{\theta}_K\}$$

表示给定所选状态数、内在维度和发射规范时的隐马尔可夫模型参数。与隐半马尔可夫模型一样，给定其潜在状态后，函数型观测条件独立，且

$$\boldsymbol{c}_{i\ell} \mid S_{i\ell} = k \sim f_k(\cdot; \boldsymbol{\theta}_k).$$

对于潜在状态路径 $\boldsymbol{s}_i = (s_{i1}, \dots, s_{iL_i})$，观测系数与状态路径的联合密度为

$$p(\boldsymbol{c}_i, \boldsymbol{s}_i; \boldsymbol{\Theta}_{\mathrm{H}}) = \rho_{s_{i1}} f_{s_{i1}}(\boldsymbol{c}_{i1}; \boldsymbol{\theta}_{s_{i1}}) \prod_{\ell=2}^{L_i} p_{s_{i,\ell-1}, s_{i\ell}} f_{s_{i\ell}}(\boldsymbol{c}_{i\ell}; \boldsymbol{\theta}_{s_{i\ell}}), \tag{B.4}$$

其中 $\boldsymbol{c}_i = (\boldsymbol{c}_{i1}, \dots, \boldsymbol{c}_{iL_i})$。对所有可能的状态路径进行边缘化，可得序列 i对观测似然的贡献，

$$\mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}) = \sum_{\boldsymbol{s}_i \in \{1,\dots,K\}^{L_i}} p(\boldsymbol{c}_i, \boldsymbol{s}_i; \boldsymbol{\Theta}_{\mathrm{H}}), \tag{B.5}$$

并且，由于序列之间相互独立，

$$\mathcal{L}(\boldsymbol{\Theta}_{\mathrm{H}}) = \prod_{i=1}^{N} \mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}), \qquad \Lambda(\boldsymbol{\Theta}_{\mathrm{H}}) = \sum_{i=1}^{N} \log \mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}).$$

设 $Z_{i\ell k} = \mathbf{1}\{S_{i\ell} = k\}$，与正文中相同。将潜在状态序列视为已观测数据时，完整数据对数似然为

$$\begin{aligned}\Lambda_c(\boldsymbol{\Theta}_{\mathrm{H}}) = &\sum_{i=1}^{N}\sum_{k=1}^{K} Z_{i1k}\log\rho_k \\ &+\sum_{i=1}^{N}\sum_{\ell=2}^{L_i}\sum_{h=1}^{K}\sum_{k=1}^{K} Z_{i,\ell-1,h}Z_{i\ell k}\log p_{hk} \\ &+\sum_{i=1}^{N}\sum_{\ell=1}^{L_i}\sum_{k=1}^{K} Z_{i\ell k}\log f_k(\boldsymbol{c}_{i\ell};\boldsymbol{\theta}_k).\end{aligned} \tag{B.6}$$

对于高斯发射分布，该表达式给出了 EM 算法所使用的完整数据准则。对于污染高斯发射分布和多元t 发射分布，还需额外引入潜在污染指标 $\Upsilon_{i\ell}$ 或尺度参数 $H_{i\ell}$。其增广条件期望和发射更新见附录 A。

B.2.1 E 步：前向–后向递推

在第 m 次迭代中，令 $\boldsymbol{\Theta}^{(m)}_{\mathrm{H}}$ 表示当前的参数估计值。定义前向概率

$$F^{(m)}_{i\ell k} = p(\boldsymbol{c}_{i1},\dots,\boldsymbol{c}_{i\ell}, S_{i\ell}=k;\boldsymbol{\Theta}^{(m)}_{\mathrm{H}}).$$

其初始化为

$$F^{(m)}_{i1k} = \rho^{(m)}_k f_k(\boldsymbol{c}_{i1};\boldsymbol{\theta}^{(m)}_k), \tag{B.7}$$

并且对于 $\ell = 2, \dots, L_i$，满足

$$F^{(m)}_{i\ell k} = f_k(\boldsymbol{c}_{i\ell};\boldsymbol{\theta}^{(m)}_k)\sum_{h=1}^{K} F^{(m)}_{i,\ell-1,h}p^{(m)}_{hk}. \tag{B.8}$$

类似地，定义后向概率

$$B^{(m)}_{i\ell k} = p(\boldsymbol{c}_{i,\ell+1},\dots,\boldsymbol{c}_{iL_i} \mid S_{i\ell}=k;\boldsymbol{\Theta}^{(m)}_{\mathrm{H}}).$$

在初始化 $B^{(m)}_{iL_ik} = 1$ 后，后向递推为

$$B^{(m)}_{i\ell h} = \sum_{k=1}^{K} p^{(m)}_{hk} f_k(\boldsymbol{c}_{i,\ell+1};\boldsymbol{\theta}^{(m)}_k)B^{(m)}_{i,\ell+1,k}, \tag{B.9}$$

对于 $\ell = L_i - 1, \ldots, 1$。序列 i 的观测似然由最终的前向概率得到，

$$\mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}^{(m)}) = \sum_{k=1}^{K} F_{iL_ik}^{(m)}.$$

前向递推和后向递推采用缩放或对数尺度实现，以避免长序列中的数值下溢。

观测 ℓ 属于状态 k 的后验概率为

$$\gamma_{i\ell k}^{(m)} = \Pr(S_{i\ell} = k \mid \mathcal{D}_i; \boldsymbol{\Theta}_{\mathrm{H}}^{(m)}) = \frac{F_{i\ell k}^{(m)} B_{i\ell k}^{(m)}}{\mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}^{(m)})}. \tag{B.10}$$

对于 $\ell = 2, \ldots, L_i$，从状态 h 转移到状态 k 的后验概率为

$$\begin{aligned} \omega_{i\ell hk}^{(m)} &= \Pr(S_{i,\ell-1} = h, S_{i\ell} = k \mid \mathcal{D}_i; \boldsymbol{\Theta}_{\mathrm{H}}^{(m)}) \\ &= \frac{F_{i,\ell-1,h}^{(m)} p_{hk}^{(m)} f_k(\boldsymbol{c}_{i\ell}; \boldsymbol{\theta}_k^{(m)}) B_{i\ell k}^{(m)}}{\mathcal{L}_i(\boldsymbol{\Theta}_{\mathrm{H}}^{(m)})}. \end{aligned} \tag{B.11}$$

这些后验量满足

$$\sum_{k=1}^{K} \gamma_{i\ell k}^{(m)} = 1, \qquad \sum_{h=1}^{K} \sum_{k=1}^{K} \omega_{i\ell hk}^{(m)} = 1,$$

and

$$\sum_{k=1}^{K} \omega_{i\ell hk}^{(m)} = \gamma_{i,\ell-1,h}^{(m)}.$$

后验状态概率 $\gamma_{i\ell k}^{(m)}$ 与第3.2.1节通过 HSMM 前向–后向递推得到的分段对应量发挥相同作用。一旦获得这些概率，两种时间模型都使用相同的发射参数估计公式。对于污染高斯和多元t 发射分布，额外的条件期望 $\upsilon_{i\ell k}^{(m)}$， $h_{i\ell k}^{(m)}$ 以及 $e_{i\ell k}^{(m)}$ 根据附录A中的方法，由当前发射参数计算得出。

B.2.2 M步

分配给状态 k 的观测值期望数量为

$$n_k^{(m)} = \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \gamma_{i\ell k}^{(m)}.$$

相对于初始分布最大化期望完整数据对数似然可得

$$\rho_k^{(m+1)} = \frac{1}{N} \sum_{i=1}^{N} \gamma_{i1k}^{(m)}. \tag{B.12}$$

转移概率通过对后验期望转移计数进行归一化来更新：

$$p_{hk}^{(m+1)} = \frac{\sum_{i=1}^{N} \sum_{\ell=2}^{L_i} \omega_{i\ell hk}^{(m)}}{\sum_{i=1}^{N} \sum_{\ell=1}^{L_i-1} \gamma_{i\ell h}^{(m)}}. \tag{B.13}$$

分子是从 h 到 k 的期望状态转换次数，分母是从状态 h 出发的期望状态转换次数。如果后者为零，则当前 E 步无法识别相应的行，必须按照拟合过程中的数值退化规则进行处理。

对于每个发射分布族，状态特定参数均使用后验权重进行更新 $\gamma_{i\ell k}^{(m)}$，并在适用时使用辅助潜变量期望。由于时间模型仅通过后验状态概率影响发射分布的估计，因此这些公式与函数型隐半马尔可夫模型所用公式相同。完整的更新公式见附录A。如果拟合隐马尔可夫模型时要求其初始分布满足平稳性，则将无约束更新公式（B.12）替换为满足平稳性约束的优化。除非另有说明，以上无约束更新公式对应于所用的基准设定。

B.3 初始化及其与隐半马尔可夫模型拟合的联系

隐马尔可夫模型采用与隐半马尔可夫模型相同的初始函数型划分和发射参数初始化程序，详见附录C.2。两者的区别在于潜在时间参数的初始化。令

$$z_{i\ell k}^{(0)} = \mathbf{1}\left\{S_{i\ell}^{(0)} = k\right\}, \qquad i = 1, \ldots, N, \quad \ell = 1, \ldots, L_i, \quad k = 1, \ldots, K,$$

给定一个正的平滑常数 ε_0，初始状态概率为

$$\rho_k^{(0)} = \frac{\varepsilon_0 + \sum_{i=1}^{N} z_{i1k}^{(0)}}{K\varepsilon_0 + N}. \tag{B.14}$$

初始转移概率由平滑后的经验转移计数得到，

$$p_{hk}^{(0)} = \frac{\varepsilon_P + \sum_{i=1}^{N}\sum_{\ell=2}^{L_i} z_{i,\ell-1,h}^{(0)} z_{i\ell k}^{(0)}}{K\varepsilon_P + \sum_{i=1}^{N}\sum_{\ell=2}^{L_i} z_{i,\ell-1,h}^{(0)}}, \tag{B.15}$$

其中 $\varepsilon_P > 0$ 是一个较小的平滑常数。初始发射参数根据相应的状态分配计算得出，并按照附录 C.2中的描述，应用针对污染高斯发射分布和多元t 发射分布的稳健初始化步骤。拟合的隐马尔可夫模型还可为具有相同发射分布族的隐半马尔可夫模型提供自然的初始化。具体而言，保留其初始状态分布和发射参数，并通过（B.3）将其转移矩阵转换为嵌入式转移矩阵。每个状态的持续时间分布初始化为由（B.2）所隐含的几何分布，或在所选逗留时间族内采用适当的近似。对于平移负二项设定，将大小参数设为 $r_k = 1$ ，并将均值参数设为

$$m_k = \frac{p_{kk}}{1 - p_{kk}}$$

能够精确复现几何驻留时间分布。

这种初始化方式尤其有用，因为平移负二项隐半马尔可夫模型将隐马尔可夫模型包含为一个特殊情形。在采用相同的观测窗口约定、无限制的持续时间支持集以及两个似然的全局最大值时，最大化后的隐半马尔可夫模型对数似然不可能小于相应隐马尔可夫模型的对数似然。在数值实现中，有限持续时间截断或优化限制可能会影响这种嵌套关系。因此，使用拟合的隐马尔可夫模型作为隐半马尔可夫模型的初始值，既提供了一种计算策略，也为比较时间结构设定提供了有用的诊断方法。

C 观测约定、初始化与模型选择

C.1 左删失的初始片段与缺失曲线

第 2.4 节观测方案在应用中会出现两种偏离。首先，由日历窗口定义的序列始于一个正在进行的状态内部，因此其第一个片段是左删失的。条件于观测起点处的状态，我们对初始剩余持续时间采用平衡前向复发约定。该约定本身并不强制整个半马尔可夫过程服从平稳状态分布。在此约定下，第一个片段的观测长度服从剩余持续时间（前向复发）分布

$$\varphi_k^0(u;\,\boldsymbol{\beta}_k) = \bar{\varphi}_k(u;\,\boldsymbol{\beta}_k)\,E_k(\boldsymbol{\beta}_k) \qquad \bar{\varphi}_k^0(u;\,\boldsymbol{\beta}_k) = \sum_{v\ge u}\varphi_k^0(v;\,\boldsymbol{\beta}_k) = \frac{1}{E_k}(\boldsymbol{\beta}_k)\,\sum_{v\ge u}\bar{\varphi}_k(v;\,\boldsymbol{\beta}_k)$$

$u \ge 1$，其中 $E_k(\boldsymbol{\beta}_k) = \sum_{v\ge 1} v\,\varphi_k(v;\boldsymbol{\beta}_k)$ 是平均持续时间。模型的左删失版本将 $\varphi_k^\dagger(u \mid u)$ 替换为：对于在 $u < L_i$ 结束的首个分段，使用 $\varphi_k^0(u)$；对于覆盖整个窗口的首个分段，使用 $\bar{\varphi}_k^0(L_i)$。其他所有项，以及现在表示首次观测时刻状态 k 的概率的 ρ_k，均保持不变。由于 $\sum_{u\ge 1}\bar{\varphi}_k(u) = E_k$，

剩余时间分布的总和为一，并且$\mathcal{L}_i$ 仍然是窗口内观测值的密度。前向、后向和后验递推通过这一替换即可沿用， $\chi_{i0}^{(m)}\,(u,\ k)$ 现在携带 $\varphi_k^0(u)$ 或 $\bar{\varphi}_k^0(L_i)$。在 M步中，首个片段将从第3.2.2 节的计数中移除，并通过其自身的计数纳入：

$$\widehat{M}_k^{(m)}(u)=\sum_{i=1}^{N}\sum_{r=u+1}^{L_i-1}\chi_i^{(m)}(r,u,k),\qquad \widehat{M}_k^{\mathrm{cens},(m)}(u)=\sum_{i:\,L_i>u}\chi_i^{(m)}(L_i,u,k),$$

$$\widehat{M}_k^{0,(m)}(u)=\sum_{i:\,L_i>u}\chi_{i0}^{(m)}(u,k),\qquad \widehat{M}_k^{0,\mathrm{cens},(m)}(u)=\sum_{i:\,L_i=u}\chi_{i0}^{(m)}(u,k).$$

随后，持续时间参数使下式达到最大值

$$\sum_u\Big[\widehat{M}_k^{(m)}(u)\log\varphi_k(u;\boldsymbol{\beta}_k)+\widehat{M}_k^{\mathrm{cens},(m)}(u)\log\bar{\varphi}_k(u;\boldsymbol{\beta}_k)$$
$$+\widehat{M}_k^{0,(m)}(u)\log\varphi_k^0(u;\boldsymbol{\beta}_k)+\widehat{M}_k^{0,\mathrm{cens},(m)}(u)\log\bar{\varphi}_k^0(u;\boldsymbol{\beta}_k)\Big].$$

因此，第一段分段贡献的是 $\log\varphi_k^0(u)$ 或 $\log\ \bar{\varphi}_k^0(L_i)$，以替代其通常项，而不是在通常项之外再加上该项；之后的每个分段都保留其通常项：如果在L_i 之前结束，则为 $\log\varphi_k(u)$；如果是终端片段，则为 $\log\bar{\varphi}_k(u)$。对于参数化持续时间族，目标函数通过数值方法进行最大化；对于非参数族，我们不使用左删失版本。对于几何持续时间 $\varphi_k^0=\varphi_k$，根据无记忆性，因此隐马尔可夫似然保持不变，并且 HMM 仍嵌套于负二项 HSMM 中。其次，函数型观测可能缺失，例如某个日期的有效小时测量值过少。缺失曲线在每个状态下贡献 $f_k\equiv 1$ ，因此潜在路径会经过该曲线；同时，它会从发射更新以及 $n_\bullet$中排除。

C.2 初始化

ECM 算法需要后验状态概率和模型参数的初始值。按照 funHDDC族中常用的初始化策略，我们通过对汇集后的系数向量应用聚类方法来初始化状态标签

$$\mathcal{C}=\{\boldsymbol{c}_{i\ell}:i=1,\dots,N,\ \ell=1,\dots,L_i\}.$$

一种标准选择是 k-均值，而在数据可能包含离群函数型观测时，采用截尾 k-均值（Cuesta-Albertos等人（1997））作为更稳健的替代方法。我们采用多个随机初始化，并保留最终 BIC 最大的解。令 $S_{i\ell}^{(0)}$ 表示初始状态分配，并定义 $z_{i\ell k}^{(0)} = \mathbf{1}\{S_{i\ell}^{(0)} = k\}$。HMM 的初始化见附录B.3。对于隐半马尔可夫模型，将每个初始化的标签序列压缩为游程。令（$\widehat{b}_{ij}^{(0)}$，$\widehat{V}_{(0)\,ij}$）表示序列 $j = 1$，……，$\widehat{M}_i^{(0)}$ 中第 j个游程的初始状态和观测长度。于是，

$$\rho_k^{(0)} = \frac{\varepsilon_0 + \sum_{i=1}^{N} \mathbf{1}\left\{\widehat{b}_{i1}^{(0)} = k\right\}}{K\varepsilon_0 + N},$$

并且，对于 $h \neq k$,

$$p_{hk}^{*(0)} = \frac{\varepsilon_P + \sum_{i=1}^{N} \sum_{j=2}^{\widehat{M}_i^{(0)}} \mathbf{1}\left\{\widehat{b}_{i,j-1}^{(0)} = h, \widehat{b}_{ij}^{(0)} = k\right\}}{(K-1)\varepsilon_P + \sum_{i=1}^{N} \sum_{j=2}^{\widehat{M}_i^{(0)}} \mathbf{1}\left\{\widehat{b}_{i,j-1}^{(0)} = h\right\}}.$$

持续时间分布根据初始标签序列中的游程进行初始化，并将每个序列的最后一个游程视为右删失，即视为持续时间至少达到其观测长度。对于左删失序列（附录 C.1），第一个游程则遵循剩余持续时间分布；其贡献与右删失的末端游程分开处理。对于非参数持续时间，我们从 $\{1, \ldots, V_{\max}\}$上的均匀分布开始，并按照第 3.2.2节所述执行一次更新，使用初始标签中的游程计数。在归一化之前，我们将 ε_φ 加到每个所得计数上。对于参数化持续时间族，初始参数通过最大化同一小节中的删失目标函数获得。如果不存在有限极大值，则保留预先指定的初始值。对于有限支撑非参数逗留时间，支撑集必须涵盖所有观测到的游程，包括右删失的初始游程或末端游程；长度超过 $V_{\max}$ 的游程与该非参数规范不兼容，不能通过在 $V_{\max}$处进行行政删失来修复。

发射参数通过将高斯、污染高斯或 t发射分布的 M 步公式应用于初始标签进行初始化，并使用预先指定的初始值 $\nu_k^{(0)}$,$\alpha_k^{(0)} \in (\alpha^*,1)$ 和 $\eta_k^{(0)} > 1$。异常曲线会增大初始簇的散布矩阵，而从这种膨胀的散布矩阵开始的污染高斯拟合，即使从良好的划分开始，也可能收敛到伪极大值。因此，对于 t 和污染发射分布，我们还使用稳健初始 M 步。在每个初始簇内，将具有最大马氏距离平方 $\Delta_k(\boldsymbol{c}_{i\ell})$ 的曲线中占比为 $1-\alpha_k^{(0)}$ 的部分暂时剔除，然后根据其余曲线重新计算 M 步。对于 C-funHDDC，随后将 $\eta_k^{(0)}$ 设为被剔除曲线与保留曲线的马氏距离平方均值之比，以估计异常曲线的膨胀程度。

C.3 模型选择

状态数 K、内在维度 d_k、简约协方差结构、在 t-模型中采用状态特定或共用自由度，以及隐半马尔可夫模型中的持续时间设定，均视为超参数。维数 d_k 可通过卡特尔碎石检验来选择：将连续特征值差与阈值 $\varepsilon_{\mathrm{Cat}}$进行比较。或者，也可以对 d_k 的可行取值进行网格搜索，并选择使贝叶斯信息准则达到最大的组合。

Let

$$n_\bullet = \sum_{i=1}^{N} L_i$$

表示函数型观测的总数。我们采用如下形式的贝叶斯信息准则

$$\mathrm{BIC} = \Lambda(\widehat{\boldsymbol{\Theta}}) - \frac{\tau}{2}\log n_\bullet,$$

其中，较大的贝叶斯信息准则值表示更优的模型（等价于最小化传统的 $\widehat{\ }$tional $-2\Lambda + \tau\log n_\bullet$ 准则），而 $\Lambda(\boldsymbol{\Theta})$ 是最大化后的观测对数似然，且

τ 表示自由参数的数量。令 τ_U 表示可识别的方向参数数量，其中

$$\tau_U = \sum_{k=1}^{K} \left[R d_k - \frac{d_k(d_k+1)}{2} - \sum_g \frac{m_{kg}(m_{kg}-1)}{2} \right],$$

其中 m_{kg} 表示在所选协方差模型约束下相等的主特征值块的大小参数，且 $\sum_g m_{kg} = d_k$。当主特征值不受约束时，其贡献为 $R d_k - d_k(d_k+1)/2$。当状态 k 内的所有主特征值都相等时，则简化为 $d_k(R-d_k)$。相应的隐马尔可夫模型参数数量见附录 B.1.2。对于 HSMM-funHDDC 模型，

$$\tau_{\text{HSMM}} = KR + \tau_U + \tau_D + (K-1) + K(K-2) + \tau_{\text{dur}},$$

其中 τ_D 是所选协方差结构中的自由特征值参数数量，$K(K-2)$是嵌入式转移矩阵P^*中的自由参数数量，而 τ_{dur} 是自由持续时间参数的数量。对于定义在$\{1,\dots,V_{\max}\}$上的非参数持续时间分布，$\tau_{\text{dur}} = K(V_{\max}-1)$；而单参数族给出$\tau_{\text{dur}} = K$，双参数族则给出 $\tau_{\text{dur}} = 2K$。对于最一般的高斯模型 HSMM-funHDDC$[a_{kj}, b_k, U_k, d_k;\ \boldsymbol{\beta}_k]$，$\tau_D = \sum_{k=1}^{K} d_k + K$和

$$\tau_{\text{HSMM}} = KR + \sum_{k=1}^{K} d_k \left[R - \frac{d_k+1}{2} \right] + \left(\sum_{k=1}^{K} d_k + K \right) + (K-1) + K(K-2) + \tau_{\text{dur}}.$$

污染高斯版本增加了 $2K$ 参数、α_k 和 η_k，而 t 版本增加了K 自由度；如果这些自由度在各状态之间相同，则增加一个自由度。

除了贝叶斯信息准则之外，我们还报告了对Biernacki 等人(2000)的积分完备似然（ICL）的边际熵近似，

$$\text{ICL} = \text{BIC} + \sum_{i=1}^{N} \sum_{\ell=1}^{L_i} \sum_{k=1}^{K} \gamma_{i\ell k} \log(\gamma_{i\ell k}),$$

其中，后验概率是在拟合参数处计算的，且 $0 \log 0 = 0$。该准则从贝叶斯信息准则中减去边际后验状态熵之和，从而

对不确定的状态分配施加惩罚。此处使用边缘熵之和，代替状态路径的联合后验熵。根据熵的次可加性，该和至少不小于联合熵；对于独立同分布混合模型，两者相等。因此，相较于独立同分布混合模型，该 ICL 对时间模型更为保守。所得准则的值越大越好。

C.4 收敛与分类

初始化后，E 步和 M步交替进行，直至收敛。对于污染高斯模型，该算法属于 ECM 算法，因为膨胀参数 η_k 在条件最大化步中进行更新。令 $\Lambda^{(m)} = \Lambda(\boldsymbol{\Theta}^{(m)})$ 表示迭代 m时的观测对数似然。我们采用艾特肯加速（Böhning 等人 1994，McNicholas 等人 2010）

$$a^{(m+1)} = \frac{\Lambda^{(m+2)} - \Lambda^{(m+1)}}{\Lambda^{(m+1)} - \Lambda^{(m)}},$$

以及渐近估计量

$$\Lambda_\infty^{(m+2)} = \Lambda^{(m+1)} + \frac{\Lambda^{(m+2)} - \Lambda^{(m+1)}}{1 - a^{(m+1)}}.$$

当满足以下收敛条件或达到最大迭代次数 $m_{\max}$时，算法终止。仅当艾特肯准则中的量均为有限值、其分母在数值上与零分离，且满足 $0 \le a^{(m+1)} < 1$时，才使用该准则。当

$$0 \le \Lambda_\infty^{(m+2)} - \Lambda^{(m+1)} \le \varepsilon.$$

否则，停止准则为

$$0 \le \Lambda^{(m+2)} - \Lambda^{(m+1)} \le \varepsilon.$$

这些检查仅适用于已成功完成的迭代。超出数值容差的似然下降会触发数值检查和重新计算；如果问题仍未解决，

则会终止游程，但不会宣告收敛。达到 $m_{\max}$ 同样会被记录为未收敛即终止。

对于逐点状态分类，最大后验概率（MAP）规则分配 s

$$\widehat{S}_{i\ell} = \arg\max_{k=1,\dots,K} \gamma_{i\ell k}.$$

由于观测似然在重新标记状态时保持不变，因此在报告每个拟合模型的参数或跨重启进行汇总之前，都会对其进行后处理，使状态按照规范顺序排列，从而在实践中解决命题 1 的标签切换歧义。对于序列级解码，分段维特比算法给出最可能的状态序列及持续时间，而普通维特比算法给出隐马尔可夫模型最可能的状态路径。对于 C-funHDDC 发射分布，函数型观测 $\boldsymbol{c}_{i\ell}$ 被污染的后验概率为

$$\Pr(\Upsilon_{i\ell} = 0 \mid \mathcal{D}_i) = \sum_{k=1}^{K} \gamma_{i\ell k}(1 - \upsilon_{i\ell k}).$$

D 函数型隐状态模型的可识别性

我们建立了从观测到的纵向序列分布中识别所提出的函数型隐半马尔可夫模型参数的充分条件。结果适用于状态数已知 K 且函数表示的维数固定且有限、Gram矩阵正定的情形 W。可识别性必须考虑三个不同的组成部分：状态依赖发射分布、函数主子空间，以及支配状态转移和持续时间的半马尔可夫动态。为完整起见，此处陈述相关条件和主要结果；正文第2.5节概述了其统计学解释。

D.1 假设和主要结果

假设 1（HSMM可辨识性）。

(A1) 初始状态分布满足 $\rho_k > 0$ 对每个 $k = 1, \ldots, K$。嵌入式转移矩阵 $P^* = (p^*_{hk})$ 是行随机的，且满足 $p^*_{hh} = 0$。条件于当前分段状态，分段持续时间与目标状态相互独立，具体规定见第 2.4 节。

(A2) 状态依赖发射密度 $f_1, \ldots, f_K$ 两两不同。对于污染高斯发射，$2K$ 高斯核

$$\{(\boldsymbol{\mu}_k, \Sigma_k), (\boldsymbol{\mu}_k, \eta_k \Sigma_k)\}_{k=1}^{K}$$

两两不同，其中 $0 < \alpha_k < 1$ 和 $\eta_k > 1$。此外，在污染高斯情形下，状态特定均值 $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_K$ 两两不同。

(A3) 对于每个状态 k，白化尺度矩阵的主特征值满足

$$a_{k1} \geq \cdots \geq a_{kd_k} > b_k > 0, \qquad 1 \leq d_k \leq R - 1.$$

允许存在重复的主特征值，此时识别相应的特征子空间，而非单个特征向量。

(A4) 令 $L_{\max} = \max_i L_i$。对于每个状态 k，驻留分布在 $\{1, \ldots, L_{\max} - 1\}$ 中至少一个持续时间上的概率为正，且映射

$$\boldsymbol{\beta}_k \longmapsto \big(\varphi_k(1; \boldsymbol{\beta}_k), \ldots, \varphi_k(L_{\max} - 1; \boldsymbol{\beta}_k)\big)$$

在指定的持续时间参数空间上是单射。对于支撑集为 $\{1, \ldots, V_{\max}\}$ 的非参数分布，当 $V_{\max} \leq L_{\max}$ 时，该条件成立。

(A5) 每个观测序列都从一个状态进入时刻开始，因此其第一个分段服从状态特定的完整持续时间分布 φ_k。如第 2.4 节所述，观测窗口可能会对最后一个分段造成右删失。

命题 1（函数型 HSMM 的可识别性）。设 $K \geq 2$ 已知，且假设 *1* 成立。那么，初始分布 $\boldsymbol{\rho}$、嵌入式转移矩阵 P^*、状态特定持续时间参数 $\{\boldsymbol{\beta}_k\}_{k=1}^K$以及状态依赖发射分布$\{f_k\}_{k=1}^K$ 均可根据观测系数序列的联合分布识别，但状态标签存在一个共同的置换除外。对于 ***funHDDC*** 发射参数化，状态特定位置参数 $\boldsymbol{\mu}_k$、内在维数 d_k、主尺度特征值 a_{kj}、残差尺度特征值 b_k以及投影到主函数子空间的谱投影算子也均可识别。

当主特征值为单特征值时， U_k 的列向量在符号变换意义下是可识别的。对于重复的主特征值，相应的特征子空间是可识别的，但其标准正交基在等特征值块内的旋转意义下仍然是不确定的。对于相应的发射族，污染参数（α_k、 η_k）或自由度 ν_k 是可识别的。

该结果首先通过从边际混合分布中识别状态依赖发射分布得到。高斯族和多元-t 族具有可识别的有限混合，而污染-高斯发射分布的均值所满足的附加条件确保了其成分核能够在状态内被无歧义地配对。已识别发射密度的线性无关性随后根据观测系数的有限维分布确定潜在状态过程的有限维分布。

由于观测窗口始于一个状态进入时刻，状态 k中长度为 u 的初始连续段、随后进入状态 $h \neq k$的概率满足

$$\Pr(S_{i1} = \cdots = S_{iu} = k, S_{i,u+1} = h) = \rho_k \varphi_k(u; \boldsymbol{\beta}_k) p^*_{kh}.$$

因此，可观测的状态路径分布能够识别观测时域内的持续时间质量和嵌入式转移概率。

假设1（A4）随后能够识别持续时间参数。最后， $W^{1/2}\Sigma_k W^{1/2}$ 的特征分解能够识别内在维度、尺度特征值及相应的谱子空间。以下各小节将给出辅助发射与谱引理、识别证明，以及对实验中所使用持续时间族的明确检验。

D.2 发射分布与谱识别引理

引理 1（发射密度的线性无关性）。在假设 *1*(A2)中的发射条件下，密度 f_1，……，f_K 对于高斯、污染高斯和多元高斯族中的每一种，t 均线性无关。

证明。多元高斯分布有限混合的可识别性（Yakowitz &Spragins 1968）以及位置参数、尺度和自由度各异的多元高斯t 分布的可识别性（Holzmann 等（2006））依据 Yakowitz & Spragins(1968）中有限混合可识别性的刻画，意味着相应核函数族线性无关。因此，不同的高斯或多元高斯t 状态密度线性无关。对于污染高斯发射分布，写作

$$f_k = \alpha_k \phi_R(\cdot; \boldsymbol{\mu}_k, \Sigma_k) + (1-\alpha_k)\phi_R(\cdot; \boldsymbol{\mu}_k, \eta_k \Sigma_k).$$

假设1(A2) 使所有 $2K$ 高斯核两两不同，因而线性无关。设 $\sum_{k=1}^{K} t_k f_k = 0$ 几乎处处成立。在按 $2K$个不同高斯核展开时，与状态 k 相关的两个核的系数分别为$t_k\alpha_k$ 和 $t_k($ $1-\alpha_k)$。由于 $0<\alpha_k<1$，它们均为零意味着 $t_k = 0$ 对每个 k均成立。因此， K污染状态密度线性无关。

□

引理 2 (funHDDC尺度结构的谱唯一性)。 设 Σ_k 形如(4)，并假定假设 *1*(A3)成立。则 Σ_k 确定 b_k， d_k，有序的主

尺度特征值 a_{k1}，……，a_{kd_k}，以及主谱投影算子 $U_kU_k^\top$。对于每个互异的主特征值，其谱投影算子均可被识别。单重主特征值在相差一个符号的意义下确定其单位特征向量；重主特征值仅确定其特征子空间，该特征子空间内的标准正交基在相差一个旋转的意义下等价。

证明。令 $\widetilde{\Sigma}_k = W^{1/2}\Sigma_k W^{1/2}$。根据构造，

$$\widetilde{\Sigma}_k = U_k A_k U_k^\top + b_k(I_R - U_kU_k^\top),$$

其中 $U_k^\top U_k = I_{d_k}$ 和 $a_{kj} > b_k$。因此，b_k 是 $\widetilde{\Sigma}_k$ 的最小特征值，并且

$$d_k = \operatorname{rank}(\widetilde{\Sigma}_k - b_k I_R), \qquad U_kU_k^\top = \mathbf{1}_{(b_k,\infty)}(\widetilde{\Sigma}_k).$$

这里，$\mathbf{1}_{(b_k,\infty)}(\Sigma_k)$ 表示其在特征值大于 b_k的特征子空间直和上的谱投影算子。即使 b_k 的重数为一，这些等式仍然成立；当 $d_k = R-1$ 时便是如此；即使某些主特征值相同，这些等式也依然成立。严格大于 b_k的特征值多重集合可以恢复带有重数的 a_{k1}、……、a_{kd_k}。根据谱定理，每个特征子空间投影算子都是唯一的。一维特征子空间的正交归一基在符号意义下唯一，而高维特征子空间的两个正交归一基之间相差一个正交旋转。

□

D.3 HSMM可辨识性命题的证明

命题证明 *1*。我们分别讨论发射分布的恢复、潜在状态路径概率的恢复、半马尔可夫动态的识别以及函数参数的恢复。步骤 *1*：发射分布与初始状态分布。根据假设 1（A5）的边界起始约定，第一个系数向量具有边际密度

$$q(\boldsymbol{c}) = \sum_{k=1}^{K} \rho_k f_k(\boldsymbol{c}), \qquad \rho_k > 0. \tag{D.1}$$

在高斯和多元t 情形下，有限混合可识别性（Yakowitz&Spragins1968、Holzmann 等（2006））确定了无序集合 $\{(\rho_k,\ f_k)\}_{k=1}^{K}$，其中 q。在污染高斯情形下，

$$q(\boldsymbol{c}) = \sum_{k=1}^{K}\Big[\rho_k\alpha_k\phi_R(\boldsymbol{c};\boldsymbol{\mu}_k,\Sigma_k) + \rho_k(1-\alpha_k)\phi_R(\boldsymbol{c};\boldsymbol{\mu}_k,\eta_k\Sigma_k)\Big].$$

这些 $2K$ 核根据假设1（A2）两两不同，因此其权重和参数在置换意义下可识别。污染情形中的相同假设还要求不同状态具有不同的均值：因此，恰好来自同一状态的两个核共享一个均值，从而确定了它们的配对关系。条件 $\eta_k > 1$ 区分了普通核（尺度矩阵较小）与膨胀核。将它们的权重相加可恢复 ρ_k，而它们的权重比可恢复 α_k。确定的配对随后确定了 f_k。因此，对于全部三个发射族，$\boldsymbol{\rho}$ 以及 f_k 均可在同一个状态标签置换意义下确定。以下证明中固定其中一种标记方式。

步骤 *2*：潜在状态路径概率。对于长度至少为 n的观测序列，给定状态序列的条件独立性得到

$$p^{(n)}(\boldsymbol{c}_1,\dots,\boldsymbol{c}_n) = \sum_{\boldsymbol{s}\in\{1,\dots,K\}^n} w_n(\boldsymbol{s})\prod_{\ell=1}^{n} f_{s_\ell}(\boldsymbol{c}_\ell), \qquad w_n(\boldsymbol{s}) = \Pr(S_{i1}=s_1,\dots,S_{in}=s_n). \tag{D.2}$$

根据引理1，f_1，……，f_K线性无关。它们的 n重张量积同样线性无关：展开一个假设的线性关系，并依次在每个坐标中应用线性无关性，即可迫使每个系数均为零。因此，系数w_n（$\boldsymbol{s}$）在（D.2）中由观测联合密度唯一确定，对于每个 $1 \le n \le L_{\max}$（利用长度至少为 n的序列的边缘分布）。这就识别了潜在状态过程的有限时域分布。这里未使用任何隐马尔可夫模型假设。

步骤 *3*：持续时间和嵌入式转移。由于每个序列都从片段边界开始，因此第一个分段满足 $B_{i1} \sim \boldsymbol{\rho}$ 和 $V_{i1} \mid B_{i1}=k \sim \varphi_k(\cdot;\boldsymbol{\beta}_k)$。在半马尔可夫分解下，其持续时间和下一片段的目的地条件于

给定 B_{i1}时相互独立。嵌入链不存在自转移。因此，对于$h \neq k$ 和 $1 \leq u < L_{\max}$，

$$\Pr(S_{i1} = \cdots = S_{iu} = k,\ S_{i,u+1} = h) = \rho_k \varphi_k(u; \boldsymbol{\beta}_k) p^*_{kh}. \tag{D.3}$$

左侧由步骤 2 确定。对所有目的状态求和可得

$$\varphi_k(u; \boldsymbol{\beta}_k) = \frac{1}{\rho_k} \sum_{h \neq k} \Pr(S_{i1} = \cdots = S_{iu} = k,\ S_{i,u+1} = h), \qquad 1 \leq u < L_{\max}. \tag{D.4}$$

假设 1(A4) 确保至少有一个这样的 u 对每个 k都有正质量。对于任意这样的 u，比值

$$p^*_{kh} = \frac{\Pr(S_{i1} = \cdots = S_{iu} = k,\ S_{i,u+1} = h)}{\sum_{g \neq k} \Pr(S_{i1} = \cdots = S_{iu} = k,\ S_{i,u+1} = g)} \tag{D.5}$$

确定了 k的第 P^*行中的每个元素。概率 $\varphi_k(1) \ldots, \varphi_k(L_{\max} - 1)$ 由 (D.4) 确定；其补数确定生存函数 $\Pr(V_{i1} \geq L_{\max} \mid B_{i1} = k)$。假设 1(A4) 中的单射性现在确定了 $\boldsymbol{\beta}_k$。无需对 P^* 作满秩或不可约性假设：所有 ρ_k 均为正，使得每个初始状态实验都能在总体分布中观测到。

步骤 *4*：函数参数。一旦 f_k 已知，其位置参数 $\boldsymbol{\mu}_k$ 和尺度矩阵 Σ_k 即可确定。对于高斯发射分布，它们分别是均值和协方差。对于多元t发射分布，它们是位置参数和尺度，以及自由度 ν_k；不要求 t 协方差存在。对于污染高斯发射分布，步骤 1 中的核分解可确定 $(\boldsymbol{\mu}_k,\ \Sigma_k,\ \alpha_k,\ \eta_k)$。引理 2 根据 Σ_k确定内在维数、主尺度特征值和残差尺度特征值，以及主谱投影算子；其中存在指定的特征向量/符号不确定性或特征子空间内旋转不确定性。证明至此完成。

□

D.4 持续时间族的可识别性

引理 3（移位泊松分布、负二项分布和有限支撑持续时间）。若 $L_{\max} \geq 2$，则移位泊松分布参数 λ_k 可由第一个持续时间概率质量确定。若 $L_{\max} \geq 3$，则移位负二项分布参数 (m_k, r_k) 可由 m_k、$r_k > 0$对应的前两个持续时间概率质量确定。对于支撑集为 $\{1, \ldots, V_{\max}\}$ 上的非参数持续时间分布，当且仅当 $V_{\max} \leq L_{\max}$ 时，有限时域概率质量向量才能确定整个分布（不考虑模型类中的其他限制，因为这些限制可能另行保证可识别性）。

证明。 条件 (A4) 讨论的是从有限观测时域中实现可识别性；它并不要求观测每一种可能逗留时间的完整实现。我们在假设参数位于其所述参数空间内部的前提下，对稿件中使用的持续时间族验证这一条件。对于平移泊松分布，$\varphi_k(1) = \exp(-\lambda_k)$，因此 $\lambda_k = -\log \varphi_k(1)$。

对于移位负二项分布，省略状态下标并令

$$a = \varphi(1) = \left(\frac{r}{r+m}\right)^r, \qquad h = \frac{\varphi(2)}{\varphi(1)} = \frac{rm}{r+m}.$$

两者 a 和 h 均 由前两个概率质量确定，其中 $0 < h < r$。消去 m 得到

$$a = \left(1 - \frac{h}{r}\right)^r, \qquad r > h, \qquad m = \frac{hr}{r-h}.$$

对于固定的 $h > 0$, t 右侧的对数关于 $r > h$ 严格递增，因为

$$\frac{\mathrm{d}}{\mathrm{d}r}\left[r \log\left(1 - \frac{h}{r}\right)\right] = \log(1-x) + \frac{x}{1-x} > 0, \qquad x = \frac{h}{r} \in (0,1),$$

例如，对 x 中右侧求导并注意到其在 $x = 0$ 处为零，即可得到该不等式。因此，a 和 h 唯一确定一个有限的 r，进而得到 m。泊松极限 $r = \infty$ 是一个边界点，不属于这个双参数内部模型。对于非参数模型，若 $V_{\max} \leq L_{\max}$，则通过可识别的状态分布可以观测到各质量 $\varphi(1) \ldots$，$\varphi(V_{\max} - 1)$，而最后一个质量由归一化得出。若$V_{\max} > L_{\max}$，不同的分布可以在低于 $L_{\max}$ 的持续时间上赋予相同的质量

并将剩余质量以不同方式分配到持续时间 $L_{\max}, \ldots, V_{\max}$ 上；因此，无约束的非参数分布不可识别。

□

D.5 可识别性假设的范围与解释

注 1 (边界起始与左删失序列**)**。 关于 (D.3) 的论证依赖于观测序列从片段进入时间开始。对于在进行中的片段内开始的观测窗口，第一个观测到的游程服从其剩余持续时间分布。在首次运行恒等式中，必须用其概率质量替换 $\varphi_k(u)$，而此前的证明无法直接确立这种抽样约定下的可识别性。如果 ρ_k 定义为观测起点处的状态概率，$\Pr(S_{i1} = k) = \rho_k$ 仍然成立；发生变化的是首次运行的持续时间分布，而不一定是初始状态边际分布。窗口内部已完成的游程可能提供可识别性信息，但要推广这一结论还需单独论证。

注 2 (污染高斯发射分布的可识别性**)**。 对于污染高斯发射分布，边际高斯混合的可识别性可以确定其 $2K$ 成分分布，但这本身并不能确定它们如何配对组成 K 状态依赖发射分布。假设*1*(A2) 中的均值区分条件提供了一种解决这一歧义的充分方法：与每个状态相关联的两个高斯成分具有相同的均值，而不同状态的均值各不相同。约束 $\eta_k > 1$ 则可区分每对中的普通成分和膨胀成分。

该条件对于此处所采用的基于边际分布的可识别性论证是充分的，但并未断言其对完整隐半马尔可夫模型的可识别性是必要的。特别地，相继观测值的联合分布可能借助潜在状态依赖结构，为各分量的配对提供额外信息。在更弱条件下建立可识别性超出了本文分析的范围。

D.6 隐马尔可夫特例的可识别性

命题 2（函数型隐马尔可夫模型的可识别性）。 设 $K \geq 2$ 已知。假设*HMM*转移矩阵 P 满秩，其初始分布满足 $\rho_k > 0$ 对所有k成立，且观测序列长度至少为三。假设发射密度线性无关，其中高斯、污染高斯或多元t 条件取自假设 *1*(A2)，但污染高斯发射分布不要求各状态均值互异。同时假设谱条件 (A3)。则 $\boldsymbol{\rho}$、 P以及状态发射参数均可由观测律识别，至多相差一个共同的状态置换；对于载荷向量，则允许单重主特征值的符号变化，或重特征值块内的旋转。

证明。 隐马尔可夫模型的可识别性也可以通过三观测分解直接确立。为求完整起见，这一独立结果一并列出，但上文的隐半马尔可夫模型证明并未用到它。前三个系数向量的联合密度为

$$p^{(3)}(\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3) = \sum_{h,k,j=1}^{K} \rho_h P_{hk} P_{kj} f_h(\boldsymbol{c}_1) f_k(\boldsymbol{c}_2) f_j(\boldsymbol{c}_3) \tag{D.6}$$

$$= \sum_{k=1}^{K} \varrho_k \, g_k^{(-)}(\boldsymbol{c}_1) f_k(\boldsymbol{c}_2) g_k^{(+)}(\boldsymbol{c}_3), \tag{D.7}$$

其中

$$\varrho_k = \sum_h \rho_h P_{hk}, \qquad g_k^{(-)} = \sum_h \frac{\rho_h P_{hk}}{\varrho_k} f_h, \qquad g_k^{(+)} = \sum_j P_{kj} f_j.$$

由于 $\boldsymbol{\rho}$ 严格为正，且 P 非奇异且为行随机矩阵， $\varrho_k > 0$ 对每个 k均成立。相对于线性无关基 $(f_1, \ldots\ldots, f_K)$，在（D.7）中的三个因子系统的系数矩阵分别为 $\mathrm{diag}(\boldsymbol{\rho}) P$ $\mathrm{diag}(\boldsymbol{\varrho})^{-1}$、 I_K和 $P^\top$，具体取决于方向约定。每个矩阵都具有满列秩，因此其克鲁斯卡尔秩为 K。由于 $3K \geq 2K + 2$ 对于 $K \geq 2$成立，秩-K 三元分解的唯一性（Allman 等人（2009），定理 9）表明，这些因子函数在一个共同的置换和缩放意义下是唯一确定的。将其归一化为概率

密度可消除缩放歧义。特别地，它在相同的状态标签下确定了 f_k 和 $g_k^{(+)}$ 。

线性无关性 f_j 给出唯一展开 $g_k^{(+)} = \sum_j P_{kj} f_j$，因此可识别 P的每一行。单观测边缘分布 $\sum_k \rho_k f_k$ 进而可识别 $\boldsymbol{\rho}$。对于每个已识别的 f_k，高斯、污染高斯或多元 t 参数化均可像隐半马尔可夫模型证明中的步骤 4 那样识别；在污染情形下，可以在恢复出的状态密度中识别其两个核，因此不需要各状态均值两两不同。引理 2 可恢复 funHDDC 的尺度特征值和谱子空间，但存在所述的不确定性。

□

注 3 (参数限制)。 限制 $\eta_k > 1$ 可区分膨胀高斯核与普通高斯核；若施加额外的数值约束 $\alpha_k > 1/2$，识别论证并不需要该约束。对 P 的满秩限制仅适用于命题 *2*；命题 *1*既不需要该限制，也不需要隐马尔可夫模型初始分布的平稳性。

E 计算性能

该实验方案的计算开销适中。综合所有初始值、筛选和细化迭代后，每个模型和数据集的中位耗时在模拟 I 中为 5 至 16 秒，在模拟 II 中为 5 至 44 秒；其中，采用污染发射分布和 NP 逗留时间的 HSMM 开销最大。在 85–100% 的情形下，高斯模型和 t 拟合在 200 次迭代内达到收敛准则。对于无污染重复样本，污染发射分布拟合约有 80% 达到迭代上限；这些样本中没有曲线属于异常值，且污染参数（α_k，η_k）的可识别性较弱。而在污染样本中，达到迭代上限的比例至多为 4%。这不会影响分类：因为随着 $\alpha_k \to 1$ 或 $\eta_k \to 1$，污染高斯密度趋近于高斯密度；并且在干净场景中，高斯模型和污染高斯

funHDDC 系列模型之间的调整兰德指数相差0.001（表1和5）。

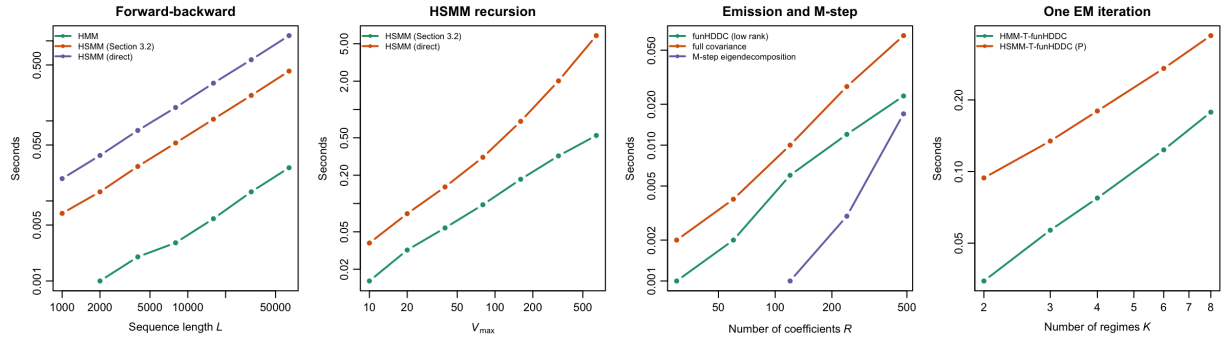

图2：计算扩展性（单核，五次游程的中位数），从左至右：前向–后向时间与 L 的关系、半马尔可夫算法和直接递推与 $V_{\max}$ 的关系、发射分布评估和特征分解与 R 的关系，以及一次 EM 迭代与 K 的关系。

图2报告了主要计算步骤的单核计时结果（在 Apple M4 Max处理器上运行五次的中位数）。隐马尔可夫和隐半马尔可夫的前向—后向递推与序列长度呈线性关系。第3.2.1节中的半马尔可夫递推，计算开销为 $O(L_i K^2 + L_i V_{\max} K)$，比计时的直接实现快约 3 倍（在 $V_{\max} = 50$ 时），在 $V_{\max} = 640$ 时则快约 11 倍；处理包含 10,000 条曲线的序列需要 0.5 秒。直接实现以 $O(L_i V_{\max} K^2)$ 的开销计算相同的递推（第3.2.1节）。由于两种开销之比 $V_{\max} K/(K + V_{\max})$ 小于 $K = 3$，实测加速比更高，是因为直接实现存在块发射和循环开销，而不仅仅是渐近计算开销。对经白化系数计算 funHDDC 发射概率的低秩计算随 R 线性增长，在 $R = 480$ 时比全协方差密度快约 3 倍。在此范围内，M步的完整特征分解大致呈二次增长；当 R 更大时，它将成为主要开销，此时可用部分特征分解替代。一次期望最大化迭代的开销随 K 大致线性增长，直至 $K = 8$；在 $V_{\max} = 50$ 时，半马尔可夫迭代的开销是隐马尔可夫迭代的两到三倍。针对 L 的递推使用 $K = 3$ 和 $V_{\max} = 50$，针对 $V_{\max}$ 的递推则使用包含 $L = 10{,}000$ 条曲线的序列。发射概率基准测试计算 20,000

某一状态下的白化曲线，具有 $d_k = 5$；EM 迭代使用 20 个序列，每个序列包含 1，000 条曲线，并采用 $R = 60$ 和 $V_{\max} = 50$。

F 模拟方案与补充结果

F.1 模拟方案与数值设置

设计方案和全局 FPCA 竞争方法详见第 4.1 节。比场景 C 更极端的污染设置容易出现伪极大值：即使从真实标签开始拟合，一个几乎空置的状态也会吸收少数极端曲线，因此未予采用。全局成分的数量 q 在研究开始前按 90% 规则确定，不作调整；竞争方法中每个状态的无约束协方差采用 funHDDC 结构，并对得分使用 $d_k = q - 1$ 。

附录 A.2 中的数值限制为 $b_k \geq 10^{-3}$ 和 $a_{kj} \geq (1 + 10^{-3})\, b_k$。污染发射分布采用 $\alpha_k \in (0.75, 1)$ 和 $\eta_k \geq 1 + 10^{-4}$估计，基于 $\alpha_k^{(0)} = 0.9$；使用常规初始 M 步时，则采用 $\eta_k^{(0)} = 2$。因此，附录 C.2 中的稳健初始 M 步会从每个初始簇中剔除 10% 的曲线，其 $\eta_k^{(0)}$ 下界为 1.5。 t发射分布的自由度因状态而异，初始值设为 $\nu_k^{(0)} = 50$ ，并基于 $[2, 200]$,更新；马氏距离平方在进入 t密度前会截断为零，权重 $h_{i\ell k}$也会作相同处理，以避免舍入误差。为保证数值稳定性，负二项逗留时间的大小参数限制为 $r_k \in [0.05, 50]$ ，因此拟合分布无法恰好达到泊松极限。附录 C.2 中的平滑常数为 $\varepsilon_0 = \varepsilon_P = \varepsilon_\varphi = 1/2$，附录 C.4 中的停止规则采用 $\varepsilon = 10^{-6}$ 和 $m_{\max} = 200$。由于似然存在许多局部极大值，尤其是在存在污染时，因此所有模型均采用相同的多起点方案进行估计：

1. 共享初始划分。每个数据集仅生成一次合并系数的十六个初始划分

这些划分由固定随机种子生成，并由所有模型共用：八个随机标记、六个k-均值划分（沿每个序列以宽度为五的滑动众数进行平滑），以及两个 k-均值划分。当 d_k 不相等时，哪个状态对应哪个维数是一个离散参数，期望最大化迭代无法改变。因此，对每个初始划分重新标记，使其簇的内在维数秩（通过 Cattell 碎石检验以 0.2 为阈值估算）与 d_k的秩相匹配。由于期望最大化迭代无法纠正将维数分配给错误状态的初始划分，因此需要执行此步骤以及下文的步骤（iii）。

2. 筛选。 每个起始值运行十次迭代，选取贝叶斯信息准则最大的三个起始值继续迭代直至收敛，并保留贝叶斯信息准则最大的非退化拟合（Biernacki 等人 2003）。对于 t和污染发射分布，每个起始值分别从附录C.2的普通初始M步和稳健初始 M 步开始，接受两次筛选。

3. 细化。 随后进行三轮处理，只有在能够提高贝叶斯信息准则时，才会用新的拟合结果替换当前结果。(i) 对每个模型重新启动：筛选过程保持不变，但从针对同一数据集拟合的所有其他模型的最终最大后验划分出发，额外进行两次启动。(ii) 每个隐半马尔可夫模型都从具有相同发射分布的拟合隐马尔可夫模型开始：使用其发射参数和初始分布、嵌入式转移$p^*_{hk} = p_{hk}/(1 - p_{hh})$，以及最接近该隐马尔可夫模型几何驻留时间的驻留时间分布；具体而言，对于负二项分布驻留时间，使用几何分布本身，因为大小参数为 $r_k = 1$ 的负二项驻留时间服从几何分布；对于 NP逗留时间，使用其截断为 $\{1, \ldots\ldots, V_{\max}\}$ 的形式；对于 P逗留时间，则使用具有相同均值的移位泊松分布。精确的、未截断的半马尔可夫模型（NB）似然将隐马尔可夫模型似然作为一个特例。然而，由于持续时间截断、参数约束或数值优化不完整，拟合的隐半马尔可夫模型最终的结果仍可能低于隐马尔可夫模型。(iii) 对于每个 funHDDC 模型，交换每一对状态的维度，并从其

最终划分重新拟合，直到没有任何交换能够提高贝叶斯信息准则。

如果其最小预期状态大小 $\sum_{i,\ell}\gamma_{i\ell k}$低于 $n_{\min}=\max\{0.01\,n_{\bullet},\ 2(\max_k d_k+1)\}$，则将该拟合视为退化并予以舍弃，因为污染高斯和 t似然会产生虚假的极大值，使某个状态塌缩到少数几条曲线上。作为对最大化过程的检查，还使用真实标签对每个模型进行一次拟合。该拟合从不用于模型选择。在 funHDDC 系列模型中，在 7，736 个模型–重复实验配对中，实验方案拟合的贝叶斯信息准则至少达到该拟合的贝叶斯信息准则（允许相差两个单位），共计 7，749 个配对中的 99。8%：包括模拟 II 中的全部 5，998 个配对，以及模拟 I 中 1，738 个配对（共 1，751 个），其中最低比例为场景 C 中 HMM-C-funHDDC 的 97%。另有 51 个配对因其基于真实标签的拟合本身退化而被排除。

通过使拟合状态与真实状态一致性最大的排列对齐后，我们使用平均持续时间的绝对误差以及拟合与真实持续时间概率质量函数之间的总变差（TV）距离，比较 HSMM 的拟合持续时间分布与真实持续时间分布。每个（研究、情景、重复实验）数据集均由各自的固定随机种子生成。我们还记录了附录C.3 中的候选贝叶斯信息准则和边际熵 ICL 如何从发射分布相同的 funHDDC 系列模型中进行选择。

F.2 模拟 I：隐马尔可夫结构

图 3 展示了第 4.2 节设计中场景 A–C 按状态划分的模拟曲线。在无污染场景中，每种隐马尔可夫模型的调整兰德指数均达到 0.969，而其独立同分布对应模型约为0.77（表1）。没有异常曲线时，三种发射分布无法区分，因此这约 0.20 的提升完全归功于时间层。在污染情况下，隐马尔可夫层搭配每种发射分布时仍优于独立同分布混合模型，但其达到的准确率取决于发射分布以及污染的严重程度

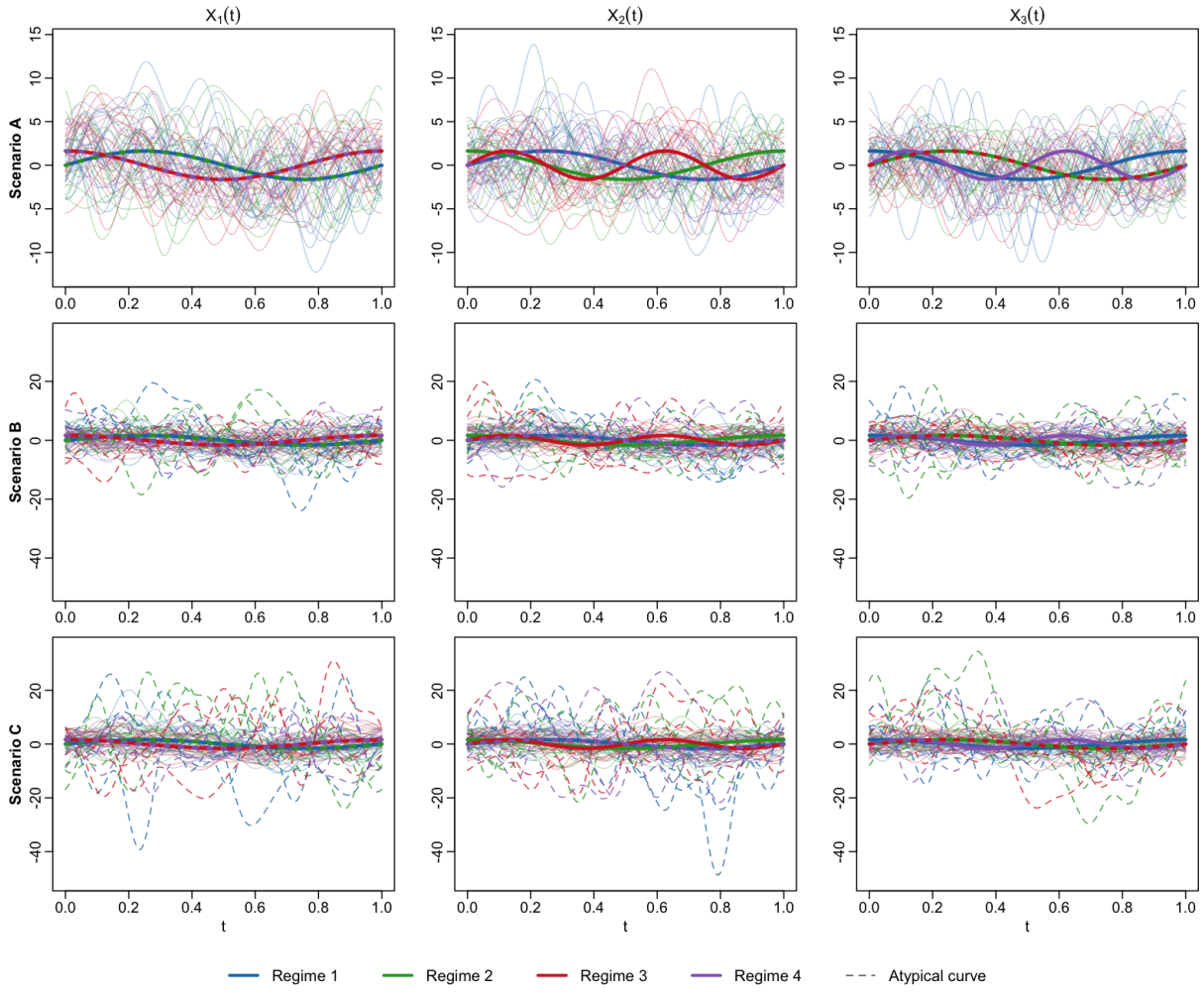

图3：模拟 I：场景 A–C 中按状态划分的模拟曲线，以及真实状态均值（两者重合时以点线表示）。虚线表示非典型曲线。

。轻度污染时，HMM-C-funHDDC 和 HMM-T-funHDDC 的准确率分别达到 0.970 和 0.941，接近其在干净数据上的准确率；而设定错误的高斯隐马尔可夫模型则降至 0.676。严重污染时，三种发射分布的表现明显分化。发射分布设定正确的 HMM-C-funHDDC 保持了 0.961 的调整兰德指数，HMM- T-funHDDC 则降至 0.886，且离散程度更大，而 HMM-funHDDC 降至 0.440。采用稳健发射分布时，时间层带来的提升随污染严重程度增加而增大：污染发射分布下从 0.21 增至 0.24，t 发射分布下从 0.25 增至 0.35；而采用高斯发射分布时，提升则从 0.13 缩小至 0.09。真实模型的拟合结果

标签表明，严重污染下 HMM-T-funHDDC 准确率较低是 t 似然的性质：其似然最高的划分并非最接近真实情况的划分。

与表 1 中全局 FPCA 竞争方法的比较，区分了状态特定子空间与时间层各自的贡献。对于独立同分布混合模型，funHDDC 模型在每种情景和每种发射分布下都比竞争方法更准确：在无污染情景下高出 0.10 至 0.12，在存在污染时高出 0.06 至 0.24。加入隐马尔可夫层后，优势有所缩小，因为时间层能从相邻曲线中恢复出许多状态信息，而这些信息是状态特定子空间从单条曲线中提取的。不过，在每种情景下，采用两种稳健发射分布时优势仍然存在：采用污染发射分布时为 0.018 至 0.026，采用t 发射分布时为 0.014 至 0.030。采用设定错误的高斯发射分布时，比较结果则不稳定。在轻度污染下，竞争方法准确率高得多（0.880 对 0.676）；在严重污染下则较低（0.402 对 0.440），而高斯竞争模型在 14 至 31 次污染重复实验中未能得到非退化拟合。对于每个模型，取其实验方案拟合结果与基于真实标签的拟合结果中表现较好者，并不会改变这些比较结果。

F.3 模拟 II：隐半马尔可夫结构

图 4展示了第 4.3 节设计中情景 A–D 的模拟曲线。三个发射分布分别在独立同分布混合模型、HMM 和 HSMM 下进行拟合，并采用第 4.3 节中的 P、NB 和 NP 驻留时间分布，由此得到十五个 funHDDC 系列模型和十五个全局 FPCA 竞争方法。所有 HSMM 拟合在整个递推过程以及第 3.2.2 节的删失持续时间更新中均使用终端生存函数（6）。

在情景 A 中，每个时间模型的调整兰德指数均达到 0.994–0.995，而对于

Table 5: Simulation II: mean ARI (SD) over 100 replicates.

	funHDDC				FPCA			
Method	Scenario A	Scenario B	Scenario C	Scenario D	Scenario A	Scenario B	Scenario C	Scenario D
Mixture	0.941 (0.009)	0.625 (0.028)	0.338 (0.032)	0.763 (0.018)	0.912 (0.012)	0.626 (0.026)	0.515 (0.024)	0.633 (0.030)
Mixture-T	0.939 (0.010)	0.862 (0.020)	0.770 (0.022)	0.755 (0.019)	0.909 (0.012)	0.820 (0.025)	0.522 (0.045)	0.609 (0.030)
Mixture-C	0.941 (0.009)	0.935 (0.012)	0.913 (0.013)	0.762 (0.018)	0.911 (0.012)	0.901 (0.014)	0.869 (0.016)	0.621 (0.032)
HMM	0.994 (0.004)	0.683 (0.024)	0.378 (0.087)	0.926 (0.013)	0.990 (0.005)	0.687 (0.023)	0.566 (0.024)	0.866 (0.020)
HMM-T	0.994 (0.003)	0.982 (0.008)	0.975 (0.015)	0.925 (0.012)	0.990 (0.005)	0.977 (0.008)	0.958 (0.053)	0.858 (0.020)
HMM-C	0.994 (0.004)	0.993 (0.004)	0.991 (0.004)	0.926 (0.012)	0.990 (0.005)	0.989 (0.005)	0.984 (0.005)	0.865 (0.020)
HSMM (P)	0.995 (0.003)	0.684 (0.024)	0.375 (0.086)	**0.943 (0.009)**	0.993 (0.004)	0.684 (0.024)	0.560 (0.024)	0.898 (0.016)
HSMM (NB)	0.995 (0.003)	0.683 (0.025)	0.378 (0.087)	0.942 (0.009)	0.993 (0.004)	0.687 (0.024)	0.566 (0.024)	0.898 (0.017)
HSMM (NP)	0.995 (0.003)	0.681 (0.025)	0.385 (0.093)	0.938 (0.010)	0.991 (0.004)	0.686 (0.024)	0.566 (0.024)	0.890 (0.020)
HSMM-T (P)	0.995 (0.003)	0.990 (0.004)	0.983 (0.006)	0.941 (0.010)	0.993 (0.004)	0.988 (0.005)	0.977 (0.007)	0.894 (0.016)
HSMM-T (NB)	0.995 (0.003)	0.991 (0.005)	0.983 (0.006)	0.940 (0.010)	0.993 (0.004)	0.987 (0.005)	0.973 (0.036)	0.893 (0.017)
HSMM-T (NP)	0.995 (0.003)	0.987 (0.005)	0.977 (0.016)	0.934 (0.011)	0.991 (0.004)	0.984 (0.007)	0.968 (0.038)	0.885 (0.020)
HSMM-C (P)	0.995 (0.003)	**0.995 (0.003)**	**0.993 (0.004)**	0.942 (0.009)	0.993 (0.004)	0.991 (0.004)	0.989 (0.005)	0.897 (0.016)
HSMM-C (NB)	0.995 (0.003)	**0.995 (0.003)**	**0.993 (0.004)**	0.942 (0.009)	0.993 (0.004)	0.991 (0.004)	0.989 (0.005)	0.896 (0.016)
HSMM-C (NP)	0.995 (0.003)	0.994 (0.004)	0.991 (0.004)	0.938 (0.010)	0.991 (0.004)	0.990 (0.004)	0.987 (0.005)	0.889 (0.019)

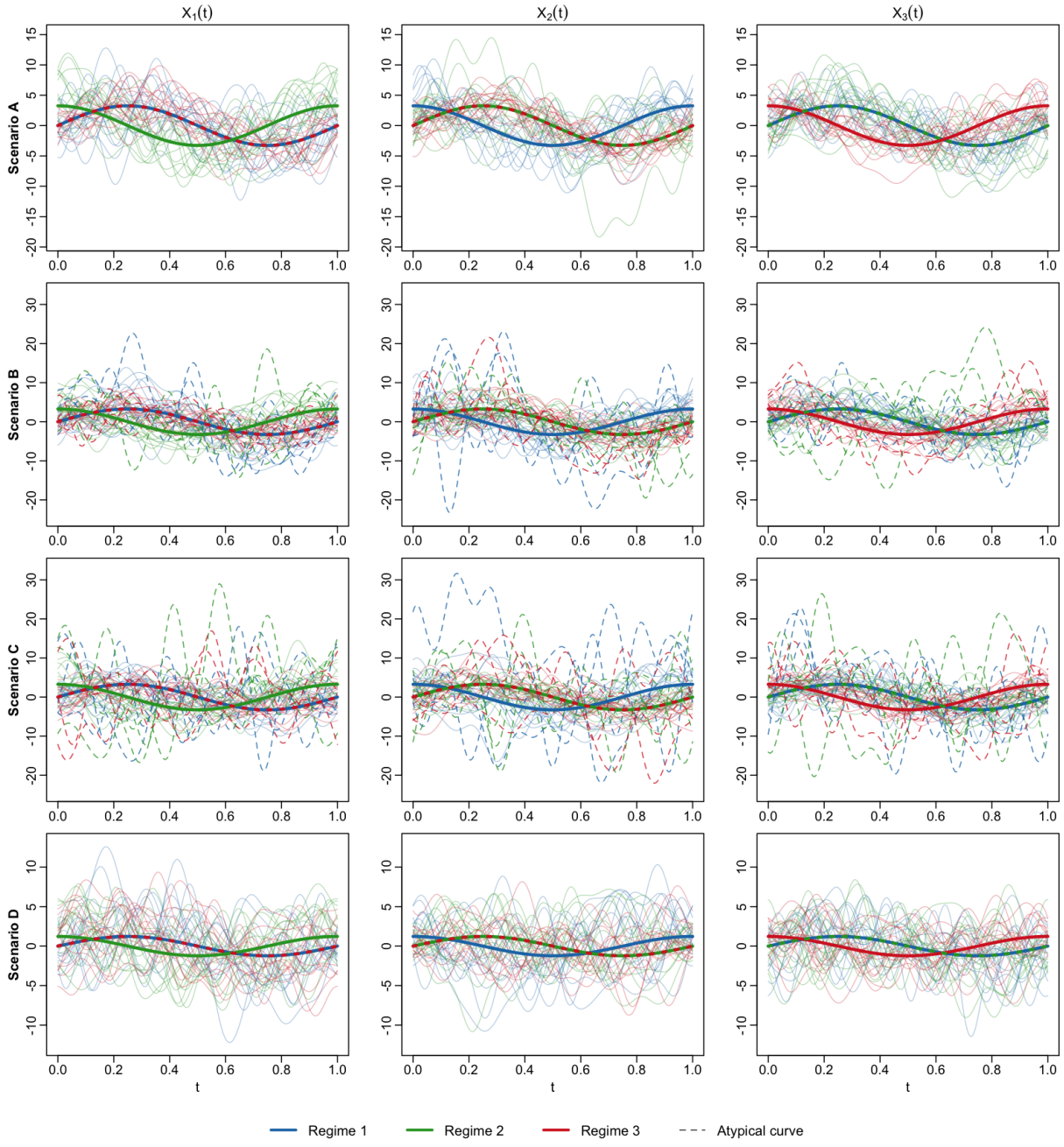

图 4：模拟 II：情景 A–D 中按状态划分的模拟曲线，以及真实状态均值（两个状态均值重合处以点线表示）。虚线曲线为非典型曲线。场景 D 有 $\mu = 1.5$。

独立同分布混合模型（表 5），半马尔可夫层最多仅比0.001 增加到隐马尔可夫层的效果。场景 D 是持续时间信息发挥作用的理想情形。在该场景中，隐马尔可夫层将调整兰德指数从约 0.76 提高到 0.925–0.926，而隐

半马尔可夫层结合 P 逗留时间后，进一步将其提高到 0.941–0.943。相较于隐马尔可夫层，这一相对于 HMM 的增益仅为 0.016 至 0.017，幅度较小，但对于每种发射分布都成立。模型设定错误的 NB 和 NP逗留时间仍保留了其中大部分增益，其调整兰德指数分别为 0.940–0.942 和 0.934–0.938。

在污染情况下，模型的排序取决于发射分布。采用稳健发射分布时，隐马尔可夫层相较于独立同分布混合模型有显著提升，ARI 从 0.06 提高到 0.21。采用 t 发射分布时，隐半马尔可夫层带来的提升较小（在采用 P逗留时间的两种污染场景中均为 0.008，采用任何停留分布时最高为 0.009）；采用污染发射分布时，提升则微乎其微（0.002），此时 HMM 已达到 0.993 和 0.991。在两种污染场景中，采用 P 或负二项分布驻留时间的 HSMM-C-funHDDC 都是准确度最高的模型（分别为 0.995 和 0.993）。对于设定错误的高斯发射分布，任何时间层都无法弥补异常曲线带来的影响：高斯模型在轻度污染下的表现约为 0.68，在严重污染下约为 0.38。因此，半马尔可夫层并非总能优于隐马尔可夫层。当状态重叠或发射分布不理想时，其 ARI 提升约为 0.01 至 0.02；而当仅凭发射分布就能区分状态时，这种提升便不复存在。在各种停留分布族中，所有场景下 P 和负二项分布驻留时间的表现相差不超过 0.003；在干净场景和采用稳健发射分布时，NP逗留时间的准确度最低。

全局 FPCA 竞争方法显示了状态特定子空间的重要性（表 5）。在场景 D 中，无论采用哪种发射分布、时间层和逗留时间族，funHDDC 在每次重复实验中的准确度都高于竞争方法。对于独立同分布混合模型，提升幅度为 0。13至0。15；对于隐马尔可夫模型，为 0。06 至 0。07 ；对于隐半马尔可夫模型，为 0。04至 0。05。因此，无论通过哪一层对时间依赖性建模，状态特定子空间的优势依然存在，且其增益约为持续时间层所带来增益的三倍。例如，HSMM-T-funHDDC (P) 的得分为 0。941 ，而 FPCA-HSMM-t (P) 为 0.894；funHDDC 隐马尔可夫模型

表6：模拟 II，干净场景：拟合平均持续时间的绝对误差（MAE，以曲线计）以及与真实概率质量函数的全变差距离（TV），在各状态和100次重复实验中取平均值。

	funHDDC				FPCA			
	情景 A		场景 D		情景 A		场景 D	
模型	平均绝对误差	TV	平均绝对误差	TV	平均绝对误差	TV	平均绝对误差	TV
半马尔可夫模型（P）	0.401	0.049	0.208	0.036	0.403	0.050	0.226	0.040
半马尔可夫模型（NB）	0.396	0.071	0.209	0.049	0.398	0.071	0.233	0.058
半马尔可夫模型（NP）	0.405	0.234	0.218	0.166	0.400	0.239	0.259	0.194
HSMM-T（P）	0.401	0.049	0.208	0.036	0.403	0.050	0.226	0.040
HSMM-T（NB）	0.396	0.071	0.209	0.049	0.396	0.071	0.233	0.057
HSMM-T（NP）	0.402	0.235	0.222	0.168	0.400	0.239	0.256	0.199
HSMM-C（P）	0.401	0.049	0.208	0.036	0.403	0.050	0.228	0.040
HSMM-C（负二项分布）	0.396	0.071	0.209	0.049	0.398	0.071	0.230	0.057
HSMM-C（NP）	0.405	0.234	0.216	0.166	0.400	0.239	0.255	0.195

（0.925–0.926）的准确度高于所有函数主成分分析隐半马尔可夫模型（0.885–0.898）。在场景 A–C 中，稳健函数型高维数据聚类模型的准确度也高于其竞争方法：独立同分布混合模型高出 0.03 至 0.25，时间模型高出 0.002 至 0.017。在严重污染且高斯发射分布设定错误的情况下，竞争方法的准确度高出 0.18 至 0.19，且两种方法的分类效果都较差。第 4.3 节报告了通过贝叶斯信息准则和 ICL 选择时间层与逗留时间族的结果。

每个隐半马尔可夫模型也都以具有相同发射分布的拟合隐马尔可夫模型为起点（附录实验方案的步骤 (ii) F.1），而这一热启动往往能得到最终保留的拟合结果。由于半马尔可夫模型（NB）包含隐马尔可夫模型，其未截断似然的精确全局最大值

至少与隐马尔可夫模型的最大值一样高。有限持续时间截断和不完全的数值优化可能会影响这一比较。对于 t 和污染发射分布，这一结论在每次重复实验中都成立。在污染情形下使用高斯发射分布时，拟合的隐马尔可夫模型具有很强的持续性，其几何驻留时间分布中有几个百分点的概率质量落在 $V_{\max} = 32$ 之外，因此半马尔可夫模型（NB）的似然可能略低于隐马尔可夫模型；将持续时间支撑集扩展到序列长度后，这种情况便不再出现。

表6总结了干净场景下持续时间分布的恢复情况。三种逗留时间分布族都能恢复平均持续时间，在场景 D 中的绝对误差约为0.21条曲线，在持续时间更长的情景 A 中约为0.40条曲线；但它们在恢复整个分布方面的表现有所不同。设定正确的 P逗留时间达到0.036–0.049的全变差距离，而负二项驻留时间为0.049–0.071。后者部分反映了上界 $r_k \le 50$ 的影响，该上界使拟合分布相较于泊松分布保持过度离散：对于情景 A 中持续时间最长的状态，方差至少为 $17 + 17^2/50 \approx 22.8$，而不是17。NP逗留时间在每个状态具有 $V_{\max} - 1 = 31$ 个自由质量，其全变差距离达到0.166–0.235。发射分布族没有明显影响。基于FPCA的隐马尔可夫模型在情景 A 中对持续时间的恢复效果与基于funHDDC的模型相当，而在场景 D 中略差；在场景 D 中，它们的分类准确率也较低。图7展示了场景 D 中拟合概率质量函数的中位数，以及各重复试验中10–90%的区间带。参数拟合与真实概率质量函数十分接近，而NP拟合的中位数虽无偏，但变异性明显更大。

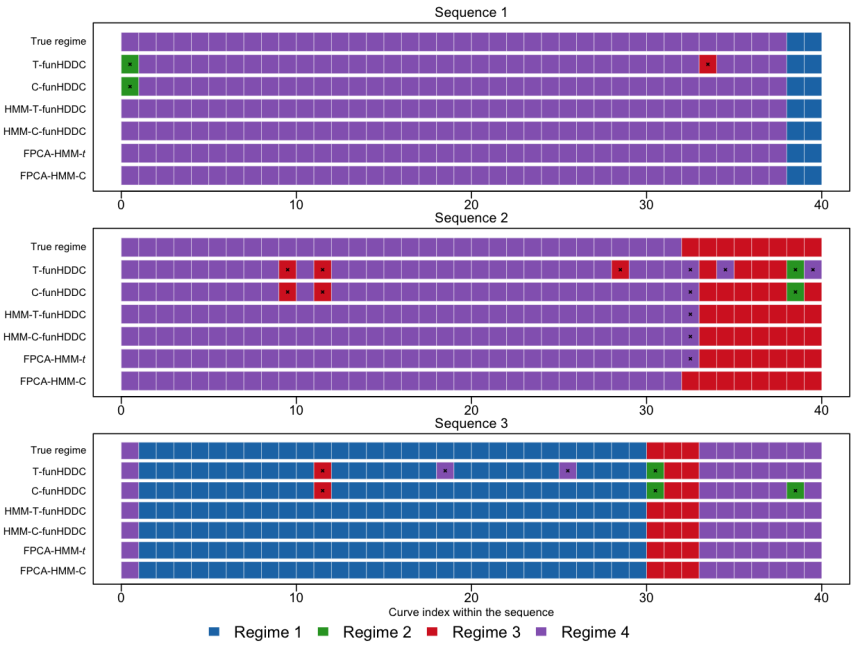

图5：模拟 I，场景B：重复试验1中前三个序列的六个模型的最大后验标签和真实状态。十字标记错误。

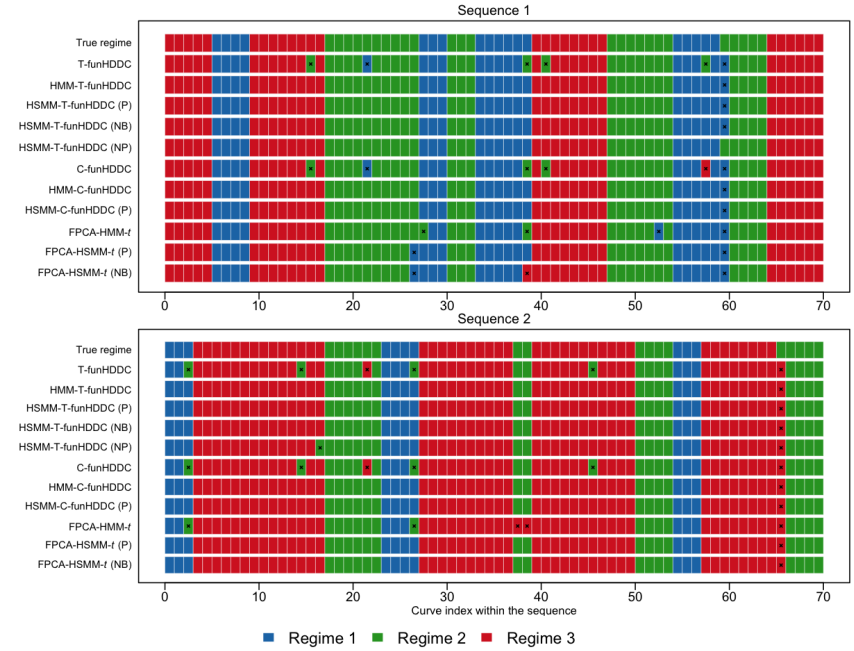

图6：模拟 II，场景 D：重复试验1中前两个序列的十一个模型的最大后验标签和真实状态。十字标记错误。

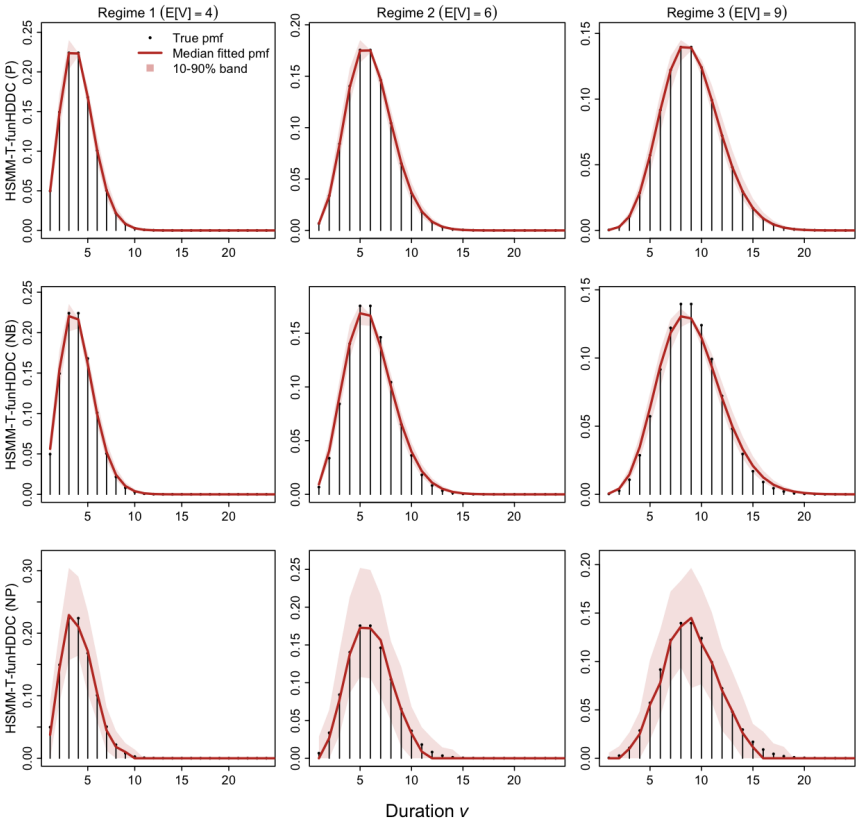

图7：模拟 II，场景 D：按逗留时间族（行）和状态（列）展示的真实持续时间概率质量函数（点）与 HSMM-T-funHDDC 的拟合概率质量函数的中位数（线，10–90% 区间带）。

G 应用方法和敏感性分析

G.1 米兰地区的空气质量

我们使用由 ARPA Lombardia 运营的四个监测站所记录的二氧化氮（NO_2），这是一种受道路交通和家庭供暖驱动的一次污染物；以及臭氧（O_3），这是一种二次光化学污染物。数据取自伦巴第开放数据门户（Maranzano &Algieri2024）。其中两个站点位于米兰市内（Pascal Città Studi，城市背景站点；以及位于市中心的 Verziere），另外两个位于其大都会区（Pioltello Limito 和 Cormano）。每个站点的每个日历日都对应于日周期 $[0, 24]$ 小时上的一条二元曲线。对于每种污染物，均使用训练期内各小时值的均值和标准差进行中心化和尺度化，并将每个

分量投影到 $R_s = 11$ 三次 B 样条（$R = 22$； W_s的条件数为 27）上。当两种污染物均至少有 18 个有效小时数时，保留该日期；否则将其视为缺失曲线。我们不删除缺失曲线，因为这样会使不连续的日期彼此相邻：缺失曲线以无发射的形式保留在序列中（附录 C.1），这占各站点训练日的 7–11%，以及留出日的 3–6%。2016–2022年构成训练序列（四个序列，每个包含 2,557 个日期和 9,247 条观测曲线，因此训练日中有 9.6% 缺失），而2023年1月1日至2026年9月24日构成留出序列（每个站点 1,363 个日期）。每个序列都始于某个季节之内，因此其第一个游程是左删失的；所有隐半马尔可夫模型拟合均使用附录C.1中的残差持续时间分布。四个站点共享相同的气象条件，因此其序列远非相互独立；因此，我们还分别对每个站点进行拟合（附录 G.3）。

我们将 $K = 2$ 设为一种便于解释的选择，而非通过统计方法选出的值，目的是对比波河谷地两种占主导地位的日污染状态：一种以 NO_2, 为主，在寒季稳定、通风不良的条件和供暖排放下累积；另一种以光化学 O_3 为主（Seinfeld & Pandis 2016, Bigi & Ghermandi 2014, Maranzano 2022）。如图 1 所示，拟合得到的状态恰好呈现这种形式。第一种状态的 NO_2 在早晚出现峰值，浓度为 50–$60\,\mu g/m^3$ ，而臭氧几乎为零。第二种状态的 NO_2 浓度较低，臭氧则在下午达到峰值，约为 $100\,\mu g/m^3$。不过，这些状态并不等同于日历季节（图8）。从四月到八月，几乎每天都处于 O_3 主导状态（NO_2 主导状态至多只占这些日期的 1%）；而从十一月到一月，NO_2 主导状态占据主导地位（占 94–98% 的日期），但期间会穿插持续时间较短的 O_3 主导状态。在 4×7 个训练站点-年份中，HSMM-T-funHDDC（NB）共有 686 个完成游程。其中恰有 28 个是持续时间较长的 O_3 主导游程，每个站点-年份各有一个，持续 110 至 227 天，开始时间在二月至六月之间。其余 O_3 主导游程中，87% 不到十

天，而 NO_2 主导游程中也有 74% 同样短暂。

表3比较了各模型。在每种发射分布中，贝叶斯信息准则和留出似然均表明，半马尔可夫模型（NB）是表现最佳的齐次模型，隐马尔可夫模型位居第二，而独立同分布混合模型和半马尔可夫模型（P）则排名垫底。采用 t 发射分布时，各项标准均显示其表现最佳：贝叶斯信息准则从 $-238{,}845$（独立同分布混合模型）升至 $-235{,}451$（隐马尔可夫模型）和 $-235{,}069$（半马尔可夫模型（NB））， 留出对数似然则从 $-123{,}873$ 升至 $-121{,}895$ 和 $-121{,}656$。因此，半马尔可夫模型相较隐马尔可夫模型的提升，约为隐马尔可夫模型相较独立同分布混合模型提升的十分之一；但这一提升在留出年份以及每个站点上都成立（附录 G.3）。拟合得到的负二项分布大小参数为 0.31 和 0.15，远低于产生隐马尔可夫模型几何驻留时间的 1，因此拟合持续时间的离散程度远高于几何分布；而欠离散泊松驻留时间的拟合效果最差（图 9）。对于以 NO_2 为主的状态，负二项驻留时间与经验运行长度十分吻合。对于以 O_3 为主的状态，它则在大量短游程和漫长的暖季运行段之间作出折中；后者在约 200 天处呈现为一个独立的众数，而两个参数族都无法再现这一特征。

注4 （大小参数低于一的负二项驻留时间）。 设 $0 < r_k < 1$ 和 $\theta_k = r_k/\left(r_k + m_k\right)$。第 *2.4* 节 中的平移负二项驻留时间随后是隐马尔可夫模型的几何停留时间的混合，

$$\varphi_k(v;\boldsymbol{\beta}_k) = \int_{(0,\,1-\theta_k)} (1-s)\, s^{v-1}\, G_k(\mathrm{d}s), \qquad v \geq 1,$$

对于关于持续概率 s 的概率测度 G_k，其风险率 $\varphi_k(v;\ \boldsymbol{\beta}_k)/\bar{\varphi}_k(v;\ \boldsymbol{\beta}_k)$ 随 v 严格递减。因此，拟合得到的小于 *1* 的大小参数，在其驻留时间分布中等价于这样一种状态：每次进入时都重新抽取持续概率。仅凭这一点，无法区分内在的持续时间依赖性与各次逗留之间离开概率的异质性，例如季节性转移或由较小的 K 合并而成的子状态所导致的异质性。

~ ~证明。设 $r = r_k$ 和 $\theta = \theta_k$，于是 $V = V - 1$ 的 $\Pr(V = v)$ 为 $= \{\Gamma(v + r)/(\Gamma(r)\ v!)\}\,\theta^r(1-\theta)^v$，其中 $v = 0,\ 1,\ \ldots$。对于 $0 < r < 1$，贝塔积分给出

$$\frac{\Gamma(v+r)}{\Gamma(r)\,v!} = \frac{B(v+r, 1-r)}{\Gamma(r)\,\Gamma(1-r)} = \frac{\sin(\pi r)}{\pi}\int_0^1 t^{\,v+r-1}(1-t)^{-r}\,\mathrm{d}t.$$

乘以 $\theta^r(1-\theta)^v$ 并代入 $s = t(1-\theta)$ 后，$1-\theta$ 的幂相互抵消，

$$\Pr(\widetilde{V} = v) = \frac{\sin(\pi r)}{\pi}\,\theta^r \int_0^{1-\theta} s^{\,v+r-1}(1-\theta-s)^{-r}\,\mathrm{d}s = \int_0^{1-\theta}(1-s)\,s^v\,g(s)\,\mathrm{d}s,$$

其中 $g(s) = \pi^{-1}\sin(\pi r)\,\theta^r s^{r-1}(1-\theta-s)^{-r}(1-s)^{-1} \geq 0$。对 v 求和，并对 $s \in (0,1)$ 使用$\sum_{v\geq 0}(1-s)s^v = 1$，根据托内利定理可知，$g$ 的积分为 1，因此~$G_k(\mathrm{d}s) = g(s)\,\mathrm{d}s$ 是一个概率测度，而 $\varphi_k(v;\,\beta_k) = \Pr(V = v-1)$ 具有所述 ~ ~形式。最后，当 $r < 1$ 时，$\Pr(V = v+1)/\Pr(V = v) = (1-\theta)(v+r)/(v+1)$ 关于 v 严格递增，因此概率质量函数是严格对数凸的，其风险率严格递减。

□

这些齐次模型忽略了状态过程具有季节性这一事实，而季节性正是半马尔可夫层带来大部分优势的原因。因此，我们还拟合了转移矩阵依赖于已知日历分组（供暖季或月份）的隐马尔可夫模型，采用相同的实验方案进行估计，并根据留出年份的日历对这些年份进行解码。对于两个转移矩阵（一个对应供暖季，另一个对应一年中的其余时间），隐马尔可夫模型优于齐次隐马尔可夫模型，但不及半马尔可夫模型（NB）。对于每月一个转移矩阵（比齐次隐马尔可夫模型多 22 个转移参数），它的表现超过了半马尔可夫模型（NB）。采用 t 发射分布时，其贝叶斯信息准则为 $-234,\ 876$，留出对数似然为 $-121,\ 570$，而相应的数值为 $-235,\ 069$ 和 $-121,\ 656$ （表 3）；采用污染发射分布时，排序也相同。其月度持续概率直接描述了季节循环：从四月到八月，状态从一天延续到次日且由 O_3主导的概率至少为 0。99；从十一月到一月，该概率至多为 0。36；而由 NO_2主导的状态则恰好相反。

因此，齐次半马尔可夫模型中过度离散的持续时间，很大程度上是季节性非平稳性的表征：齐次模型必须将冬季的短片段与每年一个持续时间很长的暖季运行段混合起来，而小大小参数的负二项驻留时间比几何分布更能近似这种情况。因此，我们不会将米兰数据解读为非几何持续时间本身的证据。要判断在考虑季节性之后持续时间是否仍为非几何分布，需要采用持续时间依赖季节的半马尔可夫模型；这将留待未来研究。

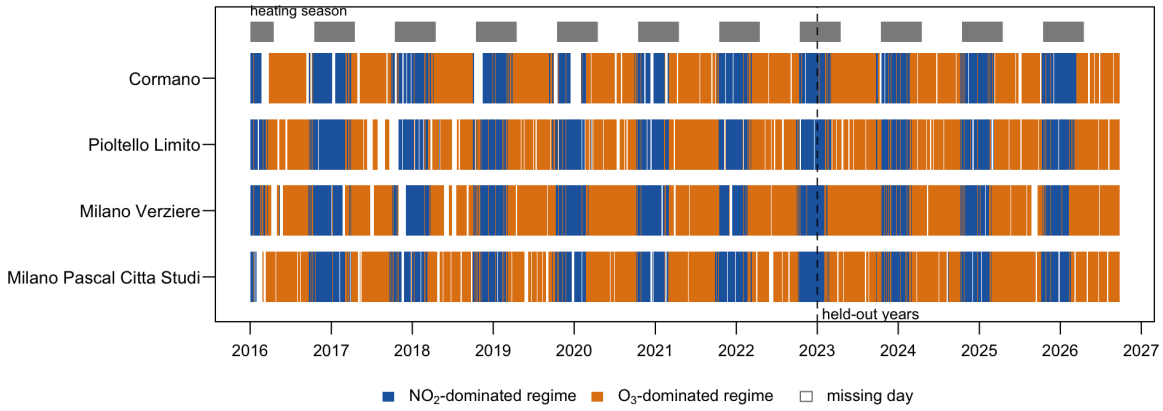

图8：米兰：基于 HSMM-T-funHDDC（NB）拟合 2016–2022 年数据，并对 2023–2026 年进行解码后，各站点和日期的最大后验状态。灰色条表示供暖季（10月15日—4月15日），白色表示缺失日期。

供暖季日历描述了分段情况，但并不能验证分段。在留出年份中，HSMM-T-funHDDC（NB）的最大后验状态与日历的一致性对应的调整兰德指数为 0.34，而独立同分布混合模型为 0.30 （表 3）。对独立同分布模型的标签采用15天滑动众数后，与日历的一致性更高（调整兰德指数为 0.39），超过任何采用t 发射分布的时间模型。采用污染发射分布或基于全局FPCA得分拟合P逗留时间的隐半马尔可夫模型也有类似表现（调整兰德指数为 0.40–0.41）；这些模型将数据分为持续时间均值约为 140 和220 天的季节，但其似然低于相应的隐马尔可夫模型。因此，与日历的一致性衡量的是分段的季节性，而非它对曲线的拟合程度

。相比之下，无标签准则更倾向于funHDDC模型所发现的更精细的状态结构。由于数据涵盖十年和数百个完成运行，对时间层的比较不再局限于单次季节性转变。结果表明，这些状态的持续性主要由日历决定。

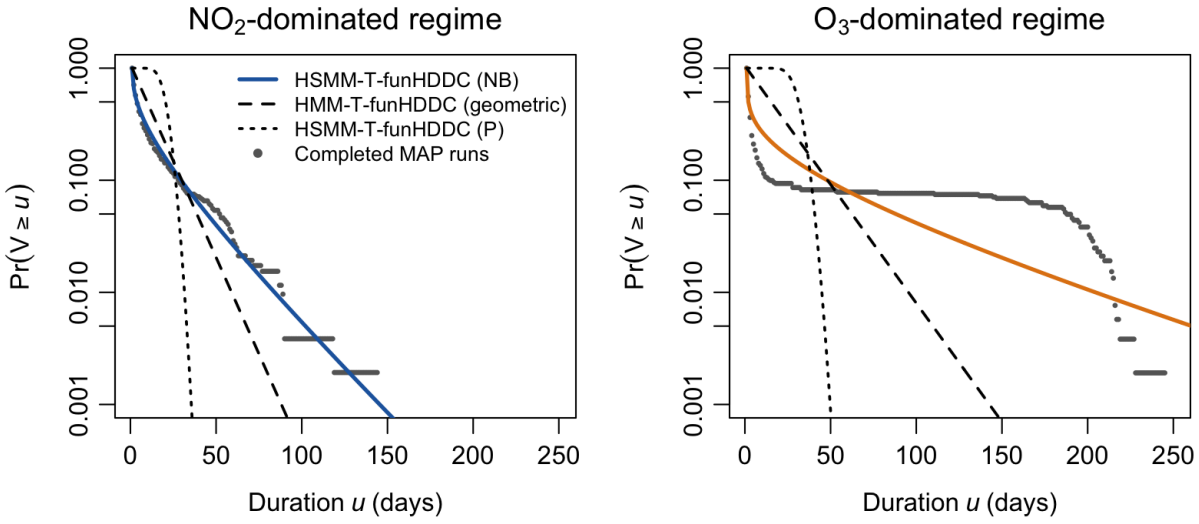

图 9：米兰：HSMM- T-funHDDC（NB）、HSMM-T-funHDDC（P）和 HMM-T-funHDDC 下拟合持续时间的生存函数 Pr($V \geq u$)，以及 NB 模型已完成 MAP 运行段的经验生存函数（点），采用对数尺度。

G.2 人类活动识别

使用智能手机进行人体活动识别数据集（Anguita 等人，2013）包含30 名志愿者在腰间佩戴智能手机，按照实验方案完成六项活动（步行、上楼和下楼、坐姿、站立和躺卧）所采集的惯性信号；数据以2.56 秒的重叠窗口形式发布。该数据集发布版本有两个特点，使其无法进行时间分析。首先，连续窗口重叠50%，因此相邻观测值共享一半样本。其次，每个受试者在每次记录会话中的窗口按活动分组，而非按时间顺序排列。我们通过重叠情况核实了第二点：当一个窗口的后半部分恰好等于下一个窗口的

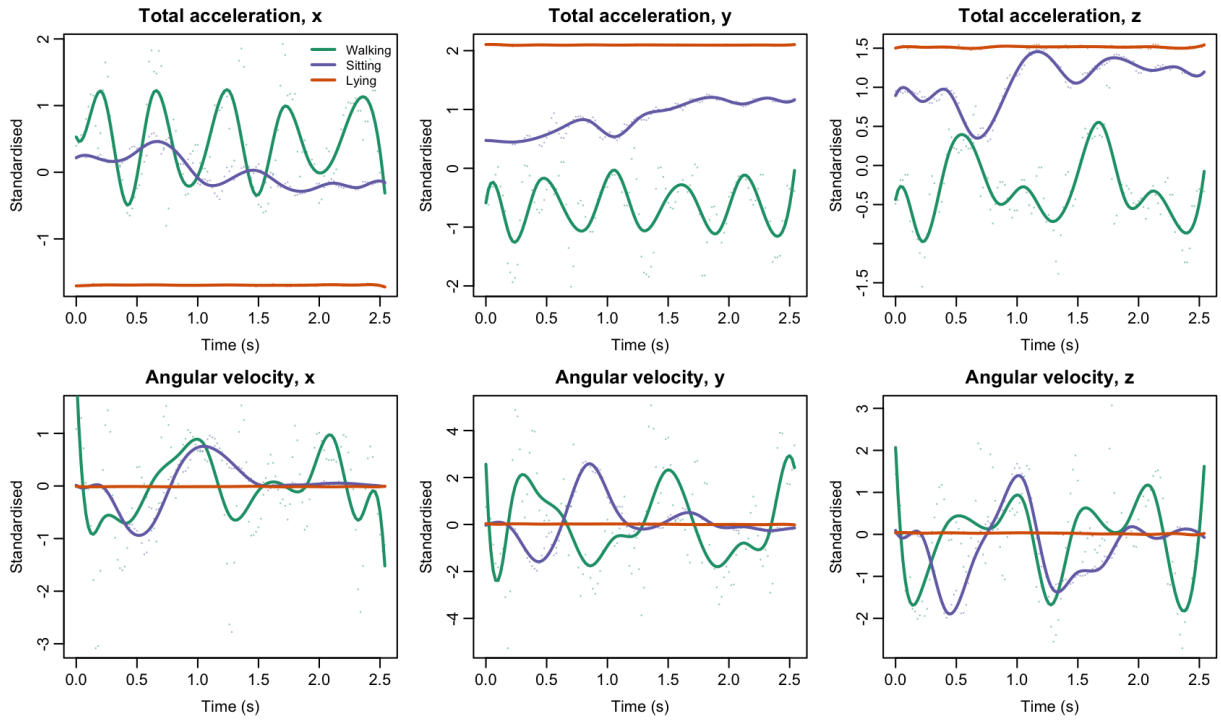

图 10：人类活动：六个通道中步行、坐姿和躺卧的一个窗口（标准化样本和最小二乘B样条拟合）。

前半部分时，两个窗口才是连续的；而在训练文件中，259 次活动变化无一例外都意味着连续性中断，并且每次会话的各个数据块按活动分组排列（先是所有站立段，然后是所有坐姿段，依此类推），而实验方案中的活动则交错进行。因此，文件顺序中的游程长度只是人为造成的。我们改用随 HAPT 数据集发布的同一批30名志愿者的原始记录（Reyes-Ortiz 等人，2016年），其中61 次记录会话均以连续的50 Hz数据流形式提供，并标明每项活动和姿势转换的开始与结束时间。

每个会话裁剪至其标注协议后，构成一个序列，并按时间顺序划分为互不重叠的窗口，每个窗口包含 128 个样本（2.56 秒）。会话的窗口数中位数为 128（范围为 52–161）。每个窗口是一条包含六个分量的曲线：三轴总加速度（其中包含重力方向）和三轴角速度。我们没有加入原始发布版本中的身体加速度通道，因为在一个窗口内，它们等于

总加速度减去一个近乎恒定的重力项。我们使用训练受试者的样本对每个通道进行中心化和尺度标准化，并将其投影到 $R_s = 15$ 三次 B 样条上，使得$R = 90$ 且 W_s 的条件数为 27。我们保留标准划分：21 名训练受试者（42 个会话，共 5,529 个窗口）和 9 名留出受试者（19 个会话，共 2,274 个窗口）。与姿势转换、未标注区段（例如在走廊尽头转弯）或两段活动交界处重叠的窗口仍保留在序列中，以便时间层能够看到整个会话，但这些窗口不计分。已评分窗口占总数的 64%，且完全位于某一活动的一段活动过程中。

我们设定 $K = 3$ 并将各项活动归为运动（步行、上楼和下楼）、直立静态姿势（坐姿和站立）以及躺卧。这是一种便于解释的建模选择，并非传感器所能支持的唯一分类方式。如图10所示，运动能量可区分运动与静态姿势，而总加速度中的重力方向可区分躺卧与直立。不过，也可以研究其他分类方式，或直接研究这六项活动本身。这些数据中的持续时间由实验方案设定，该方案规定了每项活动的执行时长。因此，这些持续时间描述的是实验，而非自然行为，我们也据此对其进行解释。

表 4 报告了训练会话上的拟合结果及其在留出受试者上的表现。在 $R_s = 15$ 下，发射分布族对分类的影响最大。污染-高斯模型对留出窗口的分类效果最佳（调整兰德指数0。87–0。91），其次是 t 模型（0。84–0。86），尽管贝叶斯信息准则和留出似然以较大优势偏好 t 发射分布。与严重污染下的模拟 I 一样，拟合和分类因此偏好不同的发射分布；按分类得到的这一排序对基函数数量并不稳定，而时间层的排序则是稳定的（附录 G.3）。全局-FPCA 竞争方法的分类效果差得多（调整兰德指数 0。47–0。69），尤其会混淆躺卧与直立姿势（图 11）：当使用 $q = 25$ 全局成分时，状态特定子空间的重要性

这里的内容远不止时间层。

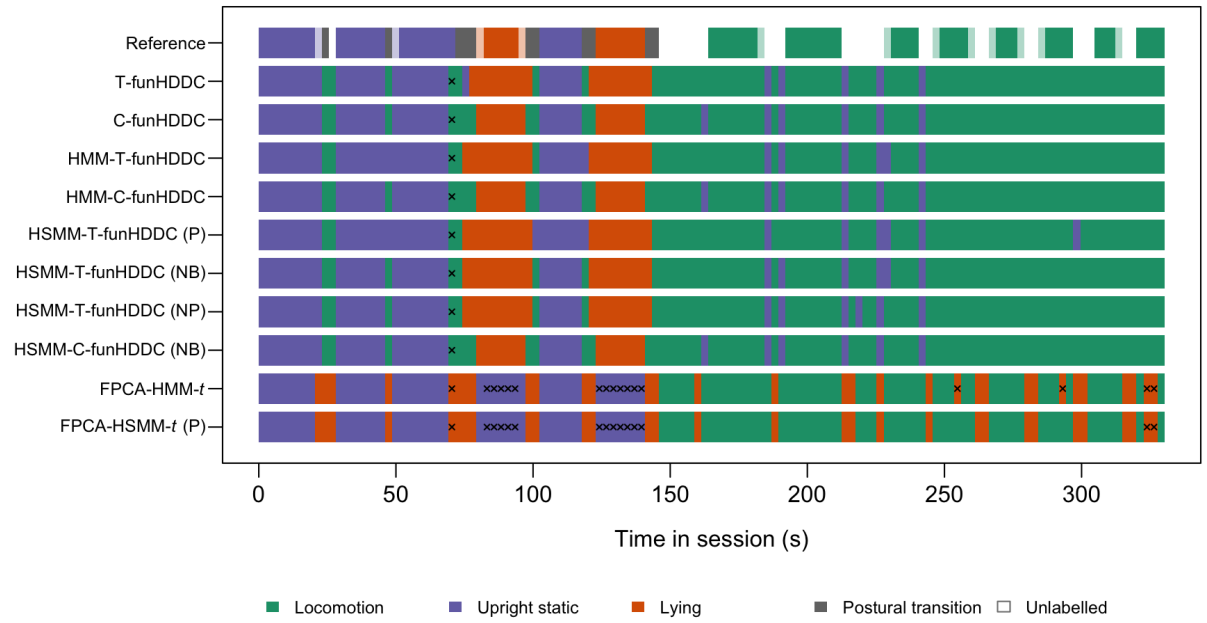

图 11：人类活动：第一次留出会话中所选模型的参考类别和留出集最大后验标签。浅色区域表示未评分窗口，灰色表示姿势转换，白色表示未标记区段。十字标记已评分窗口中的错误。

时间层改善了无标签准则，但对分类的改变很小。使用t 发射分布时，HMM 相比独立同分布混合模型将留出对数似然提高了 1,316（使用污染发射分布时提高了 754），而采用 NB 和 NP 逗留时间的隐半马尔可夫模型又分别将其提高了 135 和 157（65 和 105）。HSMM（P）会降低该值；对于两种稳健发射分布，贝叶斯信息准则和 ICL 都将 HSMM（NB）排在首位。然而，对于两种稳健发射分布，留出调整兰德指数从独立同分布混合模型到 HMM 仅提高了 0.015–0.016；半马尔可夫层几乎没有进一步影响。对于使用 t 发射分布的模型，在独立同分布标签上取五个窗口的滑动众数，其分类效果优于所有时间模型（调整兰德指数为 0.889，而最高仅为 0.858）；对于污染发射分布，其表现也相当（0.902，相比 0.871–0.909）。时长为 2.56 秒的单个窗口已经能够有效区分这三个类别。未计分的姿势转换大多被分配到运动状态（图 11），

已评分窗口上的错误并未集中出现在活动段边界：HMM-C-funHDDC 的错误中，33% 紧邻未评分窗口，而所有已评分窗口中这一比例为 32%。HSMM-T-funHDDC（NB）的拟合平均持续时间分别为：运动状态 61 秒、直立姿势 16 秒、躺卧 27 秒。协议活动段的中位数分别为：步行 19 秒、楼梯 12–13 秒、静态姿势 21–23 秒；运动的 MAP 运行段会跨越步行活动段之间未标注的转身段。各受试者的表现有所不同：在 9 名留出受试者中，HMM-C-funHDDC 的留出调整兰德指数介于 0.69 和 1.00 之间；其中两名受试者占其已评分窗口上 48 个错误中的 39 个，而这 48 个错误中有 36 个是被归为运动状态的躺卧窗口。

这六种活动本身更难以识别。在一项单独分析中，设置 $K = 6$ 个状态，每种活动对应一个状态，结果发现所有模型都会混淆坐姿与站立，以及彼此之间的三种运动活动；时间层和更大的内在维度都无法消除这些混淆。因此，使用这些通道和窗口时，三类聚合也是能够对模型进行有意义评估的层级。

G.3 敏感性分析

表 7 报告了针对 t 和污染高斯发射分布的三项敏感性分析，表 8 报告了米兰数据按站点拟合的结果。按照预期，使用原始发布版本中的 50% 重叠窗口重复进行 HAR 分析，会夸大隐马尔可夫层的表观贡献，因为相邻观测共享一半样本。每 100 个留出窗口中，与独立同分布混合模型相比，其对数似然增益在采用 t 发射分布时由 58 上升至 69，在采用污染发射分布时由 33 上升至 56；而半马尔可夫层带来的额外增益则分别从 5.9 降至 2.8，以及从 2.9 降至 1.1。使用重叠窗口时，似然最高的污染高斯拟合会将约四分之一的窗口归入膨胀成分，其中 α_k 在每个

表 7：敏感性分析：留出调整兰德指数、每 100 条曲线中 HMM 相对于独立同分布混合模型（$\Delta\ell$ HMM）以及 HSMM（NB）相对于 HMM（$\Delta\ell$HSMM）的留出对数似然增益，以及 HSMM（NB）相对于 HMM 的贝叶斯信息准则增益（ΔBIC）。

分析	发射分布	调整兰德指数（独立同分布混合模型 / HMM / HSMM）	$\Delta\ell$ HMM	$\Delta\ell$ HSMM	ΔBIC
HAR，主要分析	t	0.842 / 0.858 / 0.855	$+57.9$	$+5.9$	$+321$
($R_s = 15$)	污染	0.891 / 0.906 / 0.906	$+33.2$	$+2.9$	$+157$
HAR，50% 重叠窗口	t	0.842 / 0.846 / 0.844	$+68.8$	$+2.8$	$+332$
窗口	污染	0.367 / 0.440 / 0.439	$+55.7$	$+1.1$	$+93$
HAR，$R_s = 11$	t	0.861 / 0.865 / 0.858	$+105.6$	$+5.8$	$+355$
	污染	0.821 / 0.817 / 0.819	$+52.0$	$+5.1$	$+268$
HAR，$R_s = 21$	t	0.760 / 0.773 / 0.775	$+56.8$	$+5.8$	$+313$
	污染	0.366 / 0.922 / 0.922	-506.3	$+4.5$	$+243$
米兰，主要分析	t	0.304 / 0.344 / 0.340	$+38.0$	$+4.6$	$+382$
($R_s = 11$)	污染	0.269 / 0.341 / 0.343	$+31.3$	$+5.8$	$+495$
米兰，$R_s = 7$	t	0.310 / 0.373 / 0.363	$+39.9$	$+3.3$	$+307$
	污染	0.295 / 0.370 / 0.352	$+40.5$	$+2.8$	$+327$
米兰，$R_s = 15$	t	0.301 / 0.328 / 0.325	$+36.5$	$+4.5$	$+440$
	污染	0.257 / 0.291 / 0.286	$+34.8$	$+4.4$	$+519$

状态中都处于或接近其 0.75 的下界，并且分类效果较差（留出调整兰德指数为 0.37–0.44）。 t 拟合不受影响（0.84–0.85）。与仅从普通初始 M步开始得到的拟合相比，这些污染拟合的 BIC 高出 800 至 2,100；前者的分类结果为 0.85–0.92。因此，分类效果不佳是这些重尾数据上污染高斯似然的特性，而非优化失败。在本分析中， $V_{\max} = 322$ 窗口，即最长会话的长度；最长训练会话包含 310 个窗口。

对于 HAR 取 $R_s = 11$ ，或米兰取 21、HAR 取 $R_s = 7$ ，关于

表8：米兰，每个站点分别拟合（t 发射分布）：隐半马尔可夫模型的贝叶斯信息准则和留出对数似然减去隐马尔可夫模型相应值的结果，以及 NO_2-和 O_3-主导状态的负二项分布平均持续时间（日期）和大小参数（大小参数为 1 时为几何驻留时间）。

	Δ贝叶斯信息准则（隐半马尔可夫模型）		Δ 留出对数似然（隐半马尔可夫模型）		半马尔可夫模型（NB）	
站点	P	NB	P	NB	平均持续时间	Size
米兰帕斯卡尔·奇塔·斯图迪	−516	+99	−188	+33	10.0/16.4	0.36/0.13
米兰韦尔齐耶雷	−1404	+95	−897	+34	7.6/13.1	0.50/0.15
Pioltello Limito	−1362	+111	−737	+29	9.7/15.8	0.34/0.13
Cormano	−1309	+89	−746	+56	12.3/19.5	0.26/0.13

时间层的结论依然成立，但有一个例外。在八项分析中的七项里，HMM 的留出曲线每百条比独立同分布混合模型多出 35 至 106 的对数似然。在全部八项分析中，半马尔可夫模型（NB）又进一步提高了 2.8 至 5.8，并且根据贝叶斯信息准则，比 HMM 高出 243 至 519，因此更受青睐。例外是 HAR 取 $R_s = 21$ 时的污染发射分布。其似然最高的独立同分布拟合再次将很大一部分窗口归入膨胀成分，分类效果不佳（调整兰德指数为 0.37）；其留出对数似然比 HMM 每百个窗口高出 506。然而，在训练会话上，HMM 拟合的对数似然更高（$-677{,}847$ 对 $-683{,}086$），因此在此情形下，HMM 对训练数据的拟合更好，但泛化能力更差。与活动类别的一致程度受基底影响（HMM 的留出调整兰德指数为 0.77 至 0.92），各发射分布按一致程度排序的结果也会受其影响。采用 $R_s = 11$ 时，t 模型的分类效果最佳（HMM 为 0.87），污染模型最差（0.82）；采用 $R_s = 21$ 时则相反（分别为 0.77 和 0.92）。分别对各站点进行拟合时，米兰的四个站点得出相同的排名：半马尔可夫模型（NB）最佳，其次是 HMM，再次是半马尔可夫模型（P）；无论按贝叶斯信息准则还是留出对数似然，排名均如此（表8）。NO_2占主导的状态，其拟合大小参数为 0.26–0.50；O_3占主导的状态则为 0.13–0.15，平均持续时间分别为 8–12 天和 13–20 天。