arXiv:2304.07837v1 [math.ST] 16 Apr 2023

二阶马尔可夫多状态模型

Mireia Besal´u[1], Guadalupe G´omez Melis[2]

摘要

在一阶马尔可夫假设下，连续时间和离散时间的多状态模型（MSM）已得到充分发展。受一个 COVID-19患者队列的启发，我们基于患者的 7 个状态之间的 14 种转移设计了一个多状态模型。初步分析显示，部分转移不满足一阶马尔可夫条件，因此我们建立了一个二阶马尔可夫模型，其中未来演化不仅取决于当前状态，还取决于前一状态。在离散时间分析中，我们假设齐性，且过去的信息仅限于连续的两个时间点，据此扩展了转移概率矩阵，并提出了 Chapman-Kolmogorov 方程的扩展形式。

数学主题分类：*62M09*、*62N02*、*60J10*

关键词：多状态模型、非马尔可夫、*COVID-19*

1. 引言

多状态模型（MSM）提供了一种非常方便的方法，用于描述个体的生命历程；个体在任何时刻都处于若干可能状态之一。尤其是，这类模型适用于描述疾病的临床病程，并常用于研究不同状态之间患者的进展情况。

[1]遗传学、微生物学与统计学系，巴塞罗那大学，巴塞罗那，西班牙[2]统计学与运筹学系，加泰罗尼亚理工大学，Jordi Girona 1-3，08034 巴塞罗那，西班牙

多状态模型的理论依据建立在随机过程理论之上，即建立在表示某一过程随时间演变的随机变量集合之上。时间可以设为离散或连续；离散时间假定过程按步骤推进，连续步骤之间的固定时间不属于模型的一部分，而连续时间模型则允许状态在任意时刻发生变化。这类模型具有极强的灵活性，几乎可以对任何类型的纵向失效时间数据进行建模。这对于对不同事件进行建模尤为重要，因为这些事件之间存在事件相关依赖，例如疾病的发生会改变死亡风险（Hougaard (1999)).

一阶马尔可夫条件指出，随机过程的未来演化仅取决于当前状态，这一假设在多状态模型中十分常见。然而，这一条件往往不太现实，难以描述临床情境。为检验这一条件，Titman 和 Putter(2020) 开发了一般对数秩检验，可用于右删失条件下的一般多状态模型。

一种可能的做法是考虑高阶马尔可夫过程，以减弱一阶马尔可夫假设。k 阶马尔可夫过程是指该过程对全部历史的依赖仅通过此前占据的k 个状态体现。尽管常常发现高阶马尔可夫过程能够更好地拟合数据，但高阶马尔可夫过程的模型在实践中很少使用，因为它们依赖大量参数，从而导致计算困难（Ching 等人 (2003)；Logan (1981)）。迄今为止，大多数已使用的高阶马尔可夫模型实例都涉及离散时间模型（即马尔可夫链）。Tong (1975)将k阶马尔可夫链 $\{X1,\cdots,X_n,\cdots\}$定义为满足以下条件概率关系的马尔可夫链

$$P(X_n|X_{n-1},X_{n-2},\cdots)=P(X_n|X_{n-1},X_{n-2},\cdots,X_{n-k}) \tag{1}$$

对所有n均成立，其中$k > 0$是满足上述条件的最小整数。

试图兼顾现实性与简约性，Raftery (1985) 提出了一个阶数为 k、具有 m 个状态的马尔可夫链模型，其中（1）表示为来自各个 X_{n-1}、X_{n-2}、$\cdots$、X_{n-k}的贡献的线性组合，即：

$$P(X_n = x_n | X_{n-1} = x_{n-1}, \cdots, X_{n-k} = x_{n-k}) = \sum_{i=1}^{k} \lambda_i q_{x_n, x_{n-i}} \tag{2}$$

其中 $\lambda_1 + \cdots + \lambda_k = 1$且$Q = \{q_{ij}\}$是一个非负的$m \times m$矩阵，其各列之和为 1，且满足 $0 \leq \sum_{i=1}^{k} \lambda_i q_{x_n,\ x_{n-i}} \leq 1$。由于独立参数的数量为$(m-1) \times m^k$，模型（2）大幅减少了这一数量。例如，具有 5 个状态的二阶过程通常需要 100 个参数，而在（2）的约束下只需 21 个。其他作者，如Ching 等人(2003)；Islam 和 Chowdhury (2006)，也采用了不同的参数间关系，使参数估计切实可行且易于明确解释。

本文提出二阶马尔可夫多状态模型，以丰富路径信息，同时控制参数数量，并保持转移概率的可解释性。采用二阶马尔可夫模型进行分析的研究较为少见。其中，Shorrocks (1976)研究了收入流动性建模中的马尔可夫假设，并得出结论：转移率应同时取决于当前收入和紧接着的过去历史，因此采用了二阶马尔可夫模型。Shamshad 等人(2005)使用二阶马尔可夫模型合成生成风速时间序列数据。

二阶马尔可夫模型假设，个体的病程进展不仅取决于当前状态，也取决于前一状态。二阶马尔可夫多状态模型通过一个$M \times M \times M$张量来表征，其中M是状态数。在这项研究中，我们将扩展转移概率矩阵定义为M个阶数为$M \times M$的不同矩阵。为计算n步转移概率，我们扩展

一阶查普曼-科尔莫戈洛夫方程。论文最后以一项实例研究作结，该研究涉及一个由巴塞罗那大都市区五家医院的2000多名COVID-19患者组成的队列，这些患者在冠状病毒大流行第一波期间（2020年3月至4月）入院。针对这组数据，我们构建了一个多状态模型，其中包含患者入院后可能处于的七种状态之间的14种可能转移。由于观察发现，并非所有转移都满足一阶马尔可夫条件，我们拟合了一个二阶马尔可夫链。我们估计了二阶转移概率，并据此计算了多项指标，包括住院一定天数后从一种状态转移到另一种状态的转移概率，同时区分入院时已患病的患者与在医院内发病的患者。论文最后讨论了研究的局限性，并为未来研究指明方向。

2. 一阶马尔可夫多状态过程的刻画

多状态过程是一个连续时间随机过程$X = \{X_t,\ t \geq 0\}$，取值于离散状态空间$\mathcal{S} = \{1, \cdots, M\}$。我们用 $\mathcal{F}_t := \sigma\{X_s : s \leq t\}$ 表示一个 σ-代数，它由区间 $[0,t]$ 内对过程的观测构成，并称其为滤过。我们可以将滤过视为过程截至时间t的历史，其中包含此前所处状态的信息，直到时间 t 为止。

多状态过程的概率律由其有限维分布定义，并可通过以下 3 个函数中的任意一个完全刻画：转移概率、转移强度或累积转移强度。状态**h** 与j 之间在时间s 和t（s $<t$ </style}）的转移概率 定义如下：

$$P_{hj}(s,t;\mathcal{F}_{s-}) = P(X_t = j \mid X_s = h; \mathcal{F}_{s-}) \qquad \text{for} \qquad h, j \in \mathcal{S} = \{1, \cdots, M\}$$

并表示在已知过程在时间t 处于状态j 的概率，且已知其 s

在时间s 处于状态h ，并已知其在s之前的全部历史轨迹。状态h 与j之间在时间t 的**转移强度** 定义如下:

$$\alpha_{hj}(t;\mathcal{F}_{t-}) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}P_{hj}(t,t+\Delta t;\mathcal{F}_{t-}) \qquad \text{for} \qquad h,j\in\mathcal{S}=\{1,\cdots,M\}$$

并表示在时间t从状态h 转移到状态j 的倾向。状态h 和j 之间在时间t 的**累积（积分）转移强度**定义如下:

$$A_{hj}(t;\mathcal{F}_{t-}) = \int_0^t \alpha_{hj}(u;\mathcal{F}_{t-})du \qquad \text{for} \qquad h,j\in\mathcal{S}=\{1,\cdots,M\}$$

转移概率、转移强度和累积转移强度通过$M\times M$矩阵进行汇总。具体而言，对于在 $\mathcal{F}_{t-}$ 中收集的每条轨迹，以及满足$s<t$的每一组s,t，我们用$\mathbf{P}$ 表示转移概率矩阵

$$\mathbf{P}(s,t;\mathcal{F}_{s-}) = \{P_{hj}(s,t;\mathcal{F}_{s-});h,j\in\mathcal{S}=\{1,\cdots,M\}\}.$$

最后，我们将**状态占有概率**的行向量定义为t时的 $\pi(t)=(\pi_j(t))_{j\in\mathcal{S}}$，其中 $\pi_j(t)=P(X_t=j)$表示在t时处于状态j 的概率。

2.1. 一阶马尔可夫与齐次性假设

显然，为了估计每一对状态h和j、每一对时间s和t以及s之前所有可能轨迹对应的转移概率$P^{hj}(s,t;\mathcal{F}_{s-})$，必须施加某些限制。马尔可夫假设和齐次性假设是使推断成为可能的关键。

定义 2.1。对于所有 h、$j\in\mathcal{S}=\{1,\cdots,M\}$ 和 s、t（满足 $s<t$），多状态过程满足一阶马尔可夫假设

$$P_{hj}(s,t;\mathcal{F}_{s-}) = P(X_t=j\mid X_s=h;\mathcal{F}_{s-}) = P(X_t=j\mid X_s=h) = P_{hj}(s,t).$$

也就是说，在一阶马尔可夫假设下，s 之前的不同轨迹不会改变转移概率。在一阶马尔可夫假设下，需要一个 $M\times M$

矩阵 P（s,t），用于每个（s,t）

$$\mathbf{P}(s,t) = \{P_{hj}(s,t);\ h,j \in \mathcal{S} = \{1,\cdots,M\}\}$$

定义 2.2。 如果在给定时刻 t、s (s $<t$) 任意状态之间的转移概率仅取决于这两个时刻之差（t $-s$），则称一阶马尔可夫多状态过程为齐次，即

$$P_{hj}(s,t) = P_{hj}(0,t-s) = P_{hj}(t-s).$$

在这种情况下，每个时刻 t 只需一个 M $\times M$ 矩阵$\mathbf{P}(t)$。

2.2. 马尔可夫检验

若要在这一假设下继续分析数据，我们应验证马尔可夫条件。 Titman 和 Putter (2020) 讨论了几种用于检验马尔可夫假设的方法，并提出了一种新方法。一种选择是将进入每个状态的时间作为协变量纳入 Cox 模型，并通过似然比检验检验其显著性（Kay(1986））。另一种方法是使用 Commenges-Andersen 检验的分层版本来检测共享脆弱性。其他作者´(Rodr´ıguez-Girondo 和 de U˜na Alvarez (2012）基于观测到的 Kendall's τ为渐进型三状态疾病-死亡模型提出了局部和全局的马尔可夫条件检验。

在本文以及后续的 COVID-19 分析中，我们将使用 Titman 和 Putter（2020）的检验来验证每个转移是否满足马尔可夫假设，下面将对该检验作简要介绍。该检验的主要思想是：在一阶马尔可夫假设下，时刻 t $>$s 的转移率不会受到时刻 s 所处状态的影响。若要检验状态 l 和 m（l, m $\in \mathcal{S}$）之间的转移是否满足马尔可夫假设，我们将受试者分为两组：一组是在时刻 s 处于某个固定状态 j $\in \mathcal{S}$ 的受试者，另一组则是在时刻 s 不处于该状态的受试者。然后，对于每种转移（l $\rightarrow$ m），给定固定状态 j 和固定时间 s（s $\in [t_0,t_{max}] \subset [0,\tau]$， τ总随访时间），原假设表述如下：

表述为:

$$\mathrm{H}_{0s}^{(j)}(l,m): \alpha_{lm}(t \mid X(s)=j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \text{for any} \quad t \in [s,\tau]$$

并且可以使用对数秩统计量进行检验

$$U_s^{(j)}(l,m) = \sum_{i=1}^{n} \int_s^{\tau} \left\{ \delta_i^{(j)}(s) - \frac{\sum_{k=1}^{n} \delta_k^{(j)}(s) Y_{kl}(t)}{\sum_{k=1}^{n} Y_{kl}(t)} \right\} dN_i^{(lm)}(t),$$

其中, $\delta(j)\, i\,(s) = \mathbb{1}\{X_i(s) = j\}$ 表示个体 i 在时间 s 是否处于状态 j, $Y_i(t)$ 是过程 $X_i(t)$ 的风险集指示变量, $Y_{il}(t) = \mathbb{1}\{X_i(t^-) = l\} Y_i(t)$ 是个体 i 对转移 $l \to m$ 的风险集指示变量, 而 $N^{(lm)}\, i\,(t)$ 是计数过程, 记录截至时间 t 转移 $l \to m$ 发生的次数。

标准化统计量

$$\overline{U}_s^{(j)}(l,m) = \frac{U_s^{(j)}(l,m)}{\sqrt{\widehat{Var}(U_s^{(j)})(l,m)}}.$$

可以与 $N(0,1)$. 进行比较。此外, $\left\{ U^{(}{}_{j)\,s}(l,m), s \in [t_0, t_{\max}] \right\}$ 收敛到均值为零的高斯过程, 其协方差函数可以被一致地估计。

给定固定状态 j 的原假设

$$\mathrm{H}_0^{(j)}(l,m): \alpha_{lm}(t \mid X(s)=j) = \alpha_{lm}(t \mid X(s) \neq j) \quad \forall s \in [t_0, t_{\max}] \subset [0,\tau] \text{ and } t \in [s,\tau],$$

可以根据 $\{U^{(}j)\, s\,(l,m), s \in [t_0, t_{\max}]\}$ 的汇总统计量定义总体检验统计量, 例如 $\int_{t_0}^{t_{\max}} \left| t\, U^{(}j)\, s\,(l,m) \right| ds$、$\sup_{s \in [t_0, t_{\max}]} |U(j)\, s\,(l,m)|$, 或者对某个权函数 w(s), 使用 $\int_{t_{\max}}^{t_0} w(s) |U(j)\, s\,(l,m)| ds$。

最后, 对于任意可能的 j 以及所有 $s \in [t_0, t_{\max}] \subset [0,\tau]$ 和 $t \in [s\ \tau]$, 可以构造针对原假设的总体检验统计量。

$\mathrm{H}_0(l,m)$: $\alpha_{lm}(t \mid X(s) = j) = \alpha_{lm}(t \mid X(s) \neq j) \forall j\ \forall s \in [t_0, t_{\max}] \subset [0,\tau]$ 以及 $\forall t \in [s\ \tau]$

可以根据全局检验统计量来定义，例如取它们的均值、最大值或加权均值。这些检验通过 R 中的 MarkovTest函数实现，该函数属于 mstate 包（de Wreede 等人，2011）。关于具体实现方式的细节，将在第 6.3节的示例中介绍。

3. 二阶马尔可夫多状态过程的特征

由于一阶马尔可夫假设是一个强假设，我们通过定义二阶马尔可夫假设来放宽这一条件。其主要思想是，随机过程的未来演化不仅取决于当前状态，还取决于前一个状态。

3.1. 二阶马尔可夫转移概率

我们首先定义二阶马尔可夫转移概率，并说明如何将其概括为一组矩阵，其中矩阵的数量与状态数相同。

定义3.1. 对于时刻(s,t,u)，$s<t<u$，以及相邻状态 h、j、k（即它们之间存在一条直接路径），概率 $P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j;\mathcal{F}_{s-})$满足二阶马尔可夫假设，当且仅当

$$\begin{aligned} P_{hjk}(s,t,u;\mathcal{F}_{s-}) &= P(X_u = k \mid X_s = h, X_t = j|\mathcal{F}_{s-}) \\ &= P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u). \end{aligned}$$

在二阶马尔可夫假设下，对于任意三个时间点（s,t,u），$s<t<u$，转移概率可由一个$M \times M \times M$张量$\mathbf{P}(s,t,u)$概括

$$\mathbf{P}(s,t,u) = \{P_{hjk}(s,t,u); h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

为了得到一个更易于处理的数学对象，我们将每个状态 $h \in \mathcal{S}$,对应的维度为 M 的矩阵 $\mathbf{P^{(h)}}(s,t,u)$ 定义如下：

$$\mathbf{P}_{(\mathbf{h})}(s,t,u) = \left(P_{hjk}(s,t,u)\right)_{j,k\in\mathcal{S}},$$

h因此，转移概率张量 $\mathbf{P}(s,t,u)$ 可以等价地表示为 as

M 阶矩阵，对每个 $s < t < u$。

注释 3.2。 矩阵 $\mathbf{P}^{(\mathbf{h})}(s,t,u)$ 不总是随机矩阵，因为

$$\sum_{k\in\mathcal{S}} P_{hjk}(s,t,u) = \begin{cases} 0 & \text{if } \forall j,k, \text{ the transitions } h \to j \text{ or } j \to k \text{ are not possible} \\ 1 & \text{otherwise.} \end{cases}$$

例如，如果 h 是吸收态，则除元素 $P_{hhh} = 1$ 外，所有矩阵都将为 0。

定义 3.3。 如果在给定时间 (s,t,u) 下任意三个状态之间的转移概率仅取决于两个相邻时间之间的间隔 $t-s$ 和 $u-t$，而与时间本身无关，则称二阶马尔可夫多状态过程是齐次的，即

$$P_{hjk}(s,t,u) = P_{hjk}(t-s, u-t)$$

在这种情况下，对于每一对时间 $(s,t)(s<t)$，只需要一个 $M \times M \times M$ 张量 $\mathbf{P}(s,t)$

$$\mathbf{P}(s,t) = \{P_{hjk}(s,t); h,j,k \in \mathcal{S} = \{1,\cdots,M\}\}.$$

沿用前述记号，将每个状态 $h \in \mathcal{S}$ 以及每一对时间 (s,t) $(s<t)$对应的 M 维矩阵记为$\mathbf{P}^{(\mathbf{h})}(s,t)$，即

$$\mathbf{P}_{(\mathbf{h})}(s,t) = \left(P_{hjk}(s,t)\right)_{j,k\in\mathcal{S}}.$$

请注意，齐性条件下的转移概率张量$\mathbf{P}(s,t)$也可以等价地表示为 M 个 M 维矩阵，对应每一对时间(s,t);$s<t$，其中 s 表示从 h 到 j 的时间，t 表示从 j 到 k 的时间。

3.2. 二阶马尔可夫转移强度

A除了定义二阶转移概率外，我们还可以定义 ne

二阶转移强度。

定义3.4。 对于时间(s,t) $s<t$以及相邻状态 h、j、k，二阶转移强度定义为

$$\alpha_{hjk}(s,t) = \lim_{\Delta t\to 0} \frac{1}{\Delta t} P_{hjk}(s,t,t+\Delta t)$$

也就是说， $\alpha_{hjk}(s,t)$ 表示在已知过程前一瞬间处于状态 j，且在时间 s 处于状态 h 的情况下，过程在时间 t 到达状态 k 的倾向。

如同对转移概率所做的那样，我们可以用维度为 M 的矩阵表示转移强度，即对于每个 h 以及每个 $(s,t), s < t$，有 $\alpha_{(h)}(s,t) = \left(\alpha_{hjk}(s,t)\right)_{j,k\in\mathcal{S}}$。在齐次性假设下，我们有：

$$\alpha_{hjk}(s,t) = \lim_{\Delta t\to 0}\frac{1}{\Delta t}P_{hjk}(t-s,\Delta t).$$

4. 扩展的 Chapman-Kolmogorov 方程

鉴于临床结局通常以天为单位收集，并且旨在计算经过一定天数后处于给定状态的概率，本节考虑离散时间多状态过程，而不是定义在 $t \in [0,T]$ 上的连续时间随机过程。离散时间多状态过程的其他实例已被用于对 COVID-19 疾病进展和临床结局进行建模（Chakladar et al. (2022)）。我们首先回顾一阶马尔可夫链的查普曼-科尔莫戈洛夫方程，随后将其扩展到二阶马尔可夫链。

4.1. 一阶马尔可夫链的查普曼-科尔莫戈洛夫方程

一阶离散时间多状态模型（称为马尔可夫链）取值于离散状态空间 $\mathcal{S} = \{1,\cdots, M\}$，是连续时间一阶马尔可夫过程的离散版本。因此，马尔可夫链是一种随机模型，用于描述一系列可能发生的事件，这些事件发生在离散时间点上，且每个事件发生的概率仅取决于前一事件所达到的状态。查普曼－科尔莫戈罗夫关系是（**离散**）马尔可夫链理论中的重要结论，因为它提供了一种计算n步转移概率的方法。查普曼－科尔莫戈罗夫

方程，对于任意 s，t，$u \in \mathrm{N}$ $(s < u < t)$ 可表示为：

$$P_{hj}(s,t) = P_{hj}(s,u)\,P_{hj}(u,t) \tag{3}$$

并且是马尔可夫条件的推论。查普曼-科尔莫戈洛夫方程使我们能够将一般情况下对$P^{hj}(s,\ t)$的计算（对于任意$s < t$，$(s,\ t \in \mathrm{N})$）简化为对单步转移概率$P^{hj}(s,\ s+1)$ 的计算。在马尔可夫假设下，用 $\mathbf{P}(s)$ 表示单步转移概率矩阵，即

$$\mathbf{P}(s) = \{P_{hj}(s);\, h, j \in \mathcal{S} = \{1, \cdots, M\}\}$$

其中$P^{hj}(s)$表示$P^{hj}(s,\ \ s+1)$。矩阵$\mathbf{P}(s)$的集合可简化为转移概率矩阵$\mathbf{P}$，其定义为

$$\mathbf{P} = \{P_{hj} = P_{hj}(1);\, h, j \in \mathcal{S} = \{1, \cdots, M\}\}$$

在齐次性假设下（参见定义2.2）。因此，要研究过程经过多个时间步后的演变，并借助查普曼-科尔莫戈洛夫方程（3），只需计算单步初始转移概率。下一节将把这一结果推广到二阶马尔可夫链。

4.2. 二阶马尔可夫链的查普曼-科尔莫戈洛夫方程

在二阶马尔可夫假设和齐次性假设下，第 3.1 节定义的转移概率矩阵对于时间 $s < t < u$ 以及相邻状态 h、j、k满足：

$$P_{hjk}(s,t,u;\mathcal{F}_{s-}) = P(X_u = k \mid X_s = h, X_t = j) = P_{hjk}(s,t,u) = P_{hjk}(t-s, u-t)$$

特别地，对于任意 $s \in \mathrm{N}$，$s > 1$ 以及连续时间点 s、$s+1$、$s+2$，我们有

$$\begin{aligned} P_{hjk}(s, s+1, s+2; \mathcal{F}_{s-}) &= P(X_{s+2} = k \mid X_s = h, X_{s+1} = j) \\ &= P_{hjk}(s, s+1, s+2) = P_{hjk}(1,1). \end{aligned}$$

下一个定理给出了计算n步转移概率的方程，例如

$$P(X_{s+n+1} = l | X_{s+1} = j, X_s = h) = P_{hjl}(1, n), \tag{4}$$

对于s，$n \in \mathrm{N}$，$n > 1$，以及l，j，$k \in \mathcal{S}$，仅使用初始转移概率$P(X_3 = l | X_2 = j, X_1 = h) = P^{hjl}(1, 1) = P^{hjl}$。

由于我们所求的概率仅取决于n，不失一般性，我们可以假设$s = 1$，且（4）中的概率可等价地写为

$$P(X_{n+2} = l | X_2 = j, X_1 = h), \quad \text{for} \quad n \in \mathbb{N},\ l, j, k \in \mathcal{S}.$$

回想一下，这些转移概率可以汇总为M 个M 维矩阵，每个$n \in \mathrm{N}$对应一个矩阵。因此，对于每个状态$h \in \mathcal{S}$ 和$n > 1$

$$\mathbf{P}_{(\mathbf{h})}(1, 2, n+2) = \left(P_{hjl}(1, n)\right)_{j,k \in \mathcal{S}} = \left(P_{hjl}(n)\right)_{j,l \in \mathcal{S}},$$

当$n = 1$时，我们将在前面的记号中省略n。

记号 4.1。在介绍主要结果之前，我们先给出所用的矩阵记法，以便于阅读。

- 矩阵 h 的第 j 行：$Pj\cdot(h) = (Phjk)k \in \mathcal{S}$。
- 矩阵 h 的第 k 列：$P\cdot k(h) = (Ph\,jk)j \in \mathcal{S}$。
- 我们将使用 $*$ 符号表示行中的元素与矩阵的每一行相乘。例如，$Phj\cdot * P(h)$ 表示元素 $Phjk$ 与 $P(h)$ 的第 k 行中的所有元素相乘。

- $\mathbf{P}(l)$ 是由各个 Ph 矩阵的第 l 列组成的矩阵 $h \in \mathcal{S}$。
- $\mathrm{Tr}()$表示矩阵的迹。

定理 4.2. 假设 $(X_n)_{n\in\mathbb{N}}$ 是一个齐次二阶马尔可夫链。对于任意状态 $h,j,l \in S$ 以及记号定义见记号 *4.1*，有

$$P(X_4 = l \mid X_2 = j, X_1 = h) = \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \tag{5}$$

$$P(X_5 = l \mid X_2 = j, X_1 = h) = \mathrm{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)}\right) \tag{6}$$

$$P(X_6 = l \mid X_2 = j, X_1 = h) = \sum_{k_1=1}^{M} P_{hjk_1} \cdot \mathrm{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{7}$$

$$P(X_7 = l \mid X_2 = j, X_1 = h) = \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_2} P_{jk_2k_1} \cdot \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \tag{8}$$

一般情形 $n \geq 7$

$$P(X_{n+1} = l \mid X_2 = j, X_1 = h) = \sum_{k_{n-4}=1}^{M} \cdots \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_{n-4}} P_{jk_{n-4}k_{n-3}} \cdots P_{k_3k_2k_1} \times \mathrm{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right)$$

证明。 证明分为三个步骤。

步骤 1。 我们证明第一种情形（5）。

利用全概率定理和二阶马尔可夫性质，

$$\begin{aligned} P(X_4 = l \mid X_2 = j, X_1 = h) &= \sum_{k=1}^{M} P(X_4 = l, X_3 = k \mid X_2 = j, X_1 = h) \\ &= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j, X_1 = h) \\ &= \sum_{k=1}^{M} P(X_3 = k \mid X_2 = j, X_1 = h) P(X_4 = l \mid X_3 = k, X_2 = j) \\ &= \sum_{k=1}^{M} P_{hjk} \cdot P_{jkl} \end{aligned} \tag{9}$$

回顾一下，$P(X_4 = l \mid X_3 = k,\ X_2 = j) = P^{jkl}(1,\ 1) = P^{jkl}$因为所有这些假设下，决定转移概率的只有时间差。从这里我们可以像（5）中那样将其写成矩阵形式。

步骤 2. 现在我们聚焦于第二种情况（6）。

运用与前一种情况相同的论证，并结合所得结果（9）

$$
\begin{aligned}
P(X_5 = l \mid X_2 = j, X_1 = h) &= \sum_{k_1=1}^{M} P(X_5 = l, X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j, X_1 = i) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} P(X_5 = l \mid X_3 = k_1, X_2 = j) P(X_3 = k_1 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_1=1}^{M} \left(\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2} \right) \cdot P_{h j k_1}
\end{aligned}
$$

现在，如果我们想用矩阵形式来表示，可以观察到，$\sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{j k_1 k_2}$对应于每个矩阵的$l$ 列与矩阵$\mathbf{P}^{(j)}$ 的乘积，然后将该矩阵乘积的每一行分别乘以概率P_{hjk_1}，这些概率是矩阵j 行中的元素$\mathbf{P}^{(h)}$。由此可得公式（6）。

步骤 3。 接下来证明（7）。

这里重复前面步骤中的论证，并应用此前的结果。

$$
\begin{aligned}
P(X_6 = l \mid X_2 = j, X_1 = h) &= \sum_{k_3=1}^{M} P(X_6 = l, X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j, X_1 = h) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P(X_6 = l \mid X_3 = k_3, X_2 = j) P(X_3 = k_3 \mid X_2 = j, X_1 = h) \\
&= \sum_{k_3=1}^{M} P_{h j k_3} \left(\sum_{k_1=1}^{M} P_{j k_3 k_1} \sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{k_3 k_1 k_2} \right) \\
&= P_{hj1} \left(\sum_{k_1=1}^{M} P_{j k_3 k_1} \sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{k_3 k_1 k_2} \right) + \ldots + P_{hjM} \left(\sum_{k_1=1}^{M} P_{j k_3 k_1} \sum_{k_2=1}^{M} P_{k_1 k_2 l} \cdot P_{k_3 k_1 k_2} \right)
\end{aligned}
$$

我们注意到，所得表达式与前一步类似，但乘以了概率$P_{hj\cdot}$，这些概率对应于矩阵$\mathbf{P}^{(h)}$的j 行。因此，这种情况下的矩阵表达式显而易见。

接下来，为了证明方程（8），以及一般情形，我们只需重复上一种情形中的相同论证，即可通过归纳法轻松得到一般公式。

在这两种情况下，对矩阵形式的主要修改是在每一步中添加一个求和项。■

推论 4.2.1。 假设 $(X_n)_{n\in\mathbb{N}}$ 是齐次二阶马尔可夫链。对于时间 $s > 0$，任意状态 h、j、$l \in \mathcal{S}$ 以及注记 *4.1* 中定义的记号，有

$$
\begin{aligned}
P(X_{s+3} = l \mid X_{s+1} = j, X_s = h) &= \mathbf{P}_{j\cdot(h)} \cdot \mathbf{P}_{\cdot l(j)} \\
P(X_{s+4} = l \mid X_{s+1} = j, X_s = h) &= \operatorname{Tr}\left(P_{hj\cdot} * \mathbf{P}_{(j)} \cdot \mathbf{P}^{(l)}\right) \\
P(X_{s+5} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_1=1}^{M} P_{hjk_1} \cdot \operatorname{Tr}\left(P_{jk_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right) \\
P(X_{s+6} = l \mid X_{s+1} = j, X_s = h) &= \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} P_{hjk_2} P_{jk_2k_1} \cdot \operatorname{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right)
\end{aligned}
$$

一般情形 $n \geq 7$

$$
\begin{aligned}
P(X_{s+n} = l \mid X_{s+1} = j, X_s = h) = \sum_{k_{n-4}=1}^{M} \cdots \sum_{k_2=1}^{M} \sum_{k_1=1}^{M} & P_{hjk_{n-4}} P_{jk_{n-4}k_{n-3}} \cdots P_{k_3k_2k_1} \\
& \times \operatorname{Tr}\left(P_{k_2k_1\cdot} * \mathbf{P}_{(k_1)} \cdot \mathbf{P}^{(l)}\right)
\end{aligned}
$$

注 4.3。 我们注意到，扩展的 *Chapman-Kolmogorov* 方程仅考虑两个过去时刻相邻的情形。如果它们不相邻，由于无法确定哪个是中间状态，因此无法给出任何计算公式。对于某些模型以及存在非返回状态的特定情形，可以对此进行研究。

5. 二阶马尔可夫假设下的估计与推断

给定三个不同的相邻状态 h、j、l $\in \mathcal{S}$ (h $\to$ j $\to$ l)，其中 h、j 均非吸收态，本节旨在估计任意 s、r $\in$ N、s、r $>$ 1下的 r 步转移概率 P(X_{s+r} = l $|X_{s-1}$ = j, X_{s-2} = h)。在齐性假设下，我们有 P(X_s = l $|X_{s-1}$ = j, X_{s-2} = h) = P_{hjl}(1, 1) = P_{hjl}；并且根据推论 4.2.1，要估计 P(X_{s+r} = l $|X_{s-1}$ = j, X_{s-2} = h)，只需估计初始转移概率 P_{hjl}(1, 1) = P_{hjl}。

我们假设对个体的随访时间最长为 T 个时间单位（如示例所示，以天为单位）。令 $\{X_s^i, s=0,1,\cdots,T\}$ 表示受试者 $i=1,\ldots,n$的非可逆多状态过程，其中 $X_s^i \in \mathcal{S}$。对于 $i=1,\ldots,n$、$s=2,\cdots,T$以及 $h,j,l \in \mathcal{S}$，我们定义计数过程

$$N_{hjl}^i(s) = \mathbb{1}\{X_{s-2}^i = h, X_{s-1}^i = j, X_s^i = l\}$$

若受试者 i 在时间 s -2、s -1、s 分别从状态 h 转移到状态 j，再转移到状态 l，则计数为 1；否则为 0。在时间 s -2、s -1、s 沿路径 h $\to$j $\to$l 行进的个体总数由总和 $\tilde{\mathrm{N}}^{hjl}(\mathrm{s}) = \sum_{i=1}^n \mathrm{N}_{hjl}^i(\mathrm{s})$ 给出。$\tilde{\mathrm{N}}^{hjl}(\mathrm{s})$ 是参数为 $(\mathrm{n}\ \pi_{hjl}(\mathrm{s}))$ 的二项随机变量，其中概率$\pi_{hjl}(\mathrm{s})$ 对应于 $\pi_{hjl}(\mathrm{s}) = \mathrm{P}(\mathrm{X}_{s-2} = \mathrm{h}, \mathrm{X}_{s-1} = \mathrm{j}, \mathrm{X}_s = \mathrm{l})$。我们还定义受试者 i 在时间 s $= 2,\cdots,$ T 对应于状态 h 和 j 的风险集过程，

$$Y_{hj}^i(s-1) = \mathbb{1}\{X_{s-2}^i = h, X_{s-1}^i = j\}$$

若受试者 i 在时间 $s-2$ 处于风险中，可能转移到相邻状态 j，且已知其在时间 $s-2$ 和 $s-1$分别处于状态 h 和 j，则计数为 1。时间 s 处于风险中的个体总数由 $\tilde{Y}_{hj}(s-1) = \sum_{i=1}^n Y_{ihj}(s-1)$ 给出，对应于参数为 $(n\ \pi_{hj}(s-1))$ 的二项随机变量，其中 $\pi_{hj}(s-1) = P(X_{s-2}=h, X_{s-1}=j)$。

对于给定的 $s \geq 2$，关于转移概率 $P^{hjl}(s) = P(X_s = l \mid X_{s-1}=j, X_{s-2}=h)$ 的估计，我们将采用两种不同的方法。第一种方法利用两个概率 $P^{hjl}(s) = \pi_{hjl}(s) / \pi_{hj}(s-1)$ 的比值，而第二种方法则直接利用条件概率 $P^{hjl}(s)$ 的估计。

对于给定的相邻状态（h，j），我们必须考虑哪些时间 s 有个体处于风险中。因此，我们在此引入记号 $R^{hj} = \inf\{s \geq 2 : \tilde{Y}^{hj}(s-1) > 0\}$ 和 $T^{hj} = \sup\{s \geq 2 : Y^{hj}(s-1) > 0\}$，表示对于任意满足以下条件的 s

$R_{hj} \le s \le T^{hj}$，估计 P^{hjl}（其中 l 与 j相邻）是合理的。从实际角度来看，我们必须确保处于风险中的个体数量$\widetilde{Y}^{hj}(s-1)$足够多，才能对 P^{hjl}进行有意义的估计。若考虑每对状态之间的天数，效果等同。在这种情况下，$R'_{hj}=2$ 和$T'^{hj} = T^{hj} - R^{hj}$。

5.1. 通过伯努利概率 π_{hjl}和 π_{hj}估计 P^{hjl}

鉴于对于所有 $s=2,\cdots,T$

$$\begin{aligned} P_{hjl} &= P_{hjl}(s) = P(X_s = l \mid X_{s-1} = j, X_{s-2} = h) = \frac{P(X_{s-2}=h, X_{s-1}=j, X_s = l)}{P(X_{s-2}=h, X_{s-1}=j)} \\ &= \frac{\pi_{hjl}(s)}{\pi_{hj}(s-1)} \end{aligned}$$

可通过分别估计分子和分母，构造 P^{hjl} 的自然估计量，分别为 $\sum_{=}^{T_{hj}} s\, R_{hj}\ \widetilde{N}^{hjl}(s)/n$ 和 $\sum_{=}^{T_{hj}} s\, R_{hj}\ \widetilde{Y}^{hj}(s-1)/n$。注意，$\sum_{=}^{T_{hj}} s\, R_{hj}\ \widetilde{N}^{hjl}(s)$ 对应于在任意三个时间点 $(s-2, s-1, s)$ 均沿路径 $h \to j \to l$ 行进的个体总数；$\sum_{=}^{T_{hj}} s\, R_{hj}\ \widetilde{Y}^{hj}(s-1)$ 则是在任意两个时间点 $(s-2, s-1)$ 连续沿路径 $h \to j$ 行进的个体总数。

定理 5.1。 对于给定的相邻状态(h, j, l)，统计量

$$\widetilde{P}_{hjl} = \frac{\sum_{s=R_{hj}}^{T_{hj}} \widetilde{N}_{hjl}(s)}{\sum_{s=R_{hj}}^{T_{hj}} \widetilde{Y}_{hj}(s-1)} \tag{10}$$

对 P^{hjl} 进行一致估计。

证明。 注意，根据大数定律，以下两种估计量依概率收敛 o f

以下两种估计量:

$$\frac{1}{n}\sum_{s=R_{hj}}^{T_{hj}}\widetilde{N}_{hjl}(s)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=R_{hj}}^{T_{hj}}N_{hjl}^{i}(s)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=R_{hj}}^{T_{hj}}N_{hjl}^{i}(s)\right)=\sum_{s=R_{hj}}^{T_{hj}}E\left(N_{jhl}^{i}(s)\right)$$
$$=\sum_{s=R_{hj}}^{T_{hj}}\pi_{hjl}(s)$$
$$\frac{1}{n}\sum_{s=R_{hj}}^{T_{hj}}\widetilde{Y}_{hj}(s-1)=\frac{1}{n}\sum_{i=1}^{n}\sum_{s=R_{hj}}^{T_{hj}}Y_{hj}^{i}(s-1)\xrightarrow[n\to\infty]{\mathrm{P}}E\left(\sum_{s=R_{hj}}^{T_{hj}}Y_{hj}^{i}(s-1)\right)$$
$$=\sum_{s=R_{hj}}^{T_{hj}}E\left(Y_{jh}^{i}(s-1)\right)=\sum_{s=R_{hj}}^{T_{hj}}\pi_{hj}(s-1)$$

二阶齐性意味着

$$\begin{aligned}\pi_{hjl}(s) &= P\{X_{s-2}=h,X_{s-1}=j,X_s=l\}\\ &= P\{X_s=l\mid X_{s-2}=h,X_{s-1}=j\}P\{X_{s-1}=j,X_{s-2}=h\}\\ &= P_{hjl}(1,1)P\{X_{s-1}=j,X_{s-2}=h\}=P_{hjl}(1,1)\pi_{hj}(s-1),\end{aligned}$$

由此我们得出，P^{jhl} 依概率收敛于P^{hjl}:

$$\begin{aligned}\widetilde{P}_{jhl} &= \frac{\sum_{s=R_{hj}}^{T_{hj}}\tilde{N}_{hjl}(s)}{\sum_{s=R_{hj}}^{T_{hj}}\tilde{Y}_{hj}(s-1)}=\frac{\sum_{s=R_{hj}}^{T_{hj}}\tilde{N}_{hjl}(s)/n}{\sum_{s=R_{hj}}^{T_{hj}}\tilde{Y}_{hj}(s-1)/n}\xrightarrow[n\to\infty]{\mathrm{P}}\frac{\sum_{s=R_{hj}}^{T_{hj}}\pi_{hjl}(s)}{\sum_{s=R_{hj}}^{T_{hj}}\pi_{hj}(s-1)}\\ &= \frac{\sum_{s=R_{hj}}^{T_{hj}}P_{hjl}(1,1)\pi_{hj}(s-1)}{\sum_{s=R_{hj}}^{T_{hj}}\pi_{hj}(s-1)}=P_{hjl}\end{aligned}$$

■

5.2. 通过条件概率估计P^{hjl}

对于所有满足 s 使得 $R_{hj}\le s\le T^{hj}$ 的情形，由下列比率给出的相对频率

$$\widehat{P}_{hjl}(s)=\frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}$$

是 $P^{hjl}(s)$ 的一个直接估计量。由于对所有 $s = 2,\cdots,T$，$P^{hjl}(s) = P^{hjl}$，P^{hjl} 的估计量可以通过对 $\widehat{P}^{hjl}(s)$ 取平均得到，即

$$\widehat{P}_{hjl} = \frac{1}{T_{hj} - R_{hj} + 1} \sum_{s=R_{hj}}^{T_{hj}} \widehat{P}_{hjl}(s) = \frac{1}{T_{hj} - R_{hj} + 1} \sum_{s=R_{hj}}^{T_{hj}} \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}. \tag{11}$$

定理 5.2. 对于给定的相邻状态 (h,j,l) 统计量 P^{hjl} 由 (11) 给出，是 P 的无偏估计量^{hjl}。

证明。 我们可以将 $P^{hjl}(s)$ 写为

$$\widehat{P}_{hjl}(s) = \frac{\widetilde{N}_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)} = \sum_{\{i:X^i_{s-2}=h,X^i_{s-1}=j\}} \frac{N^i_{hjl}(s)}{\widetilde{Y}_{hj}(s-1)}.$$

$N^i_{hjl}(s)$ 在这些个体于时间 $s-2$,~$s-1$ 连续处于状态 h、j 的条件下，是成功概率为 $P^{hjl}(s)$ 的伯努利随机变量；$N^{hjl}(s)$ 是一个二项~随机变量，其试验次数等于 $Y^{hj}(s-1)$，成功概率^ ~等于 $P^{hjl}(s) = P^{hjl}$。因此，$E(P^{hjl}(s)|Y^{hj}(s-1)) = P^{hjl}$。注意，对于所有 s,

$$E\left(\widehat{P}_{hjl}(s)\right) = E\left\{E\left(\widehat{P}_{hjl}(s)|\widetilde{Y}_{hj}(s-1)\right)\right\} = E\left\{P_{hjl}(s)\right\} = P_{hjl}$$

以某种启发式的方式，将 R^{hj} 和 T^{hj} 视为固定而非随机值，

$$E(\widehat{P}_{hjl}) = E\left(\frac{1}{T_{hj} - R_{hj} + 1} \sum_{s=R_{hj}}^{T_{hj}} \widehat{P}_{hjl}(s)\right) = \frac{1}{T_{hj} - R_{hj} + 1} \sum_{s=R_{hj}}^{T_{hj}} E\left(\widehat{P}_{hjl}(s)\right) = P_{hjl}.$$

■

6. DIVINE 模型

6.1. 描述

我们用于说明的数据集对应一个队列，共有 2076 名 COVID-19 住院患者，他们在大流行病第一波期间（2020年3月至4月）入住西班牙巴塞罗那大都市区南部的五家医院。由于所有患者都接受监测，直至出院或死亡，因此所有受试者的转移时间（以天计）均可准确获知，且不存在因失访造成的不完整数据。

这些数据属于 DIVINE 项目

(https://grbio.upc.edu/en/research/highlighted-projects)

多学科研究团队由 GRBIO（UPC-UB）、贝尔维奇大学医院和贝尔维奇生物医学研究所的研究人员组成，合作制定了一个以临床医生为明确关注重点的统计框架，旨在更深入地了解由 SARS-CoV-2 病毒引起的重症疾病。基于团队的合作知识，建立了一个包含七种状态和 14 种状态转移的多状态模型（见图 1，括号中的数字表示发生该状态转移的患者人数）。如图1所示，模型考虑了 7 种状态：(1) 非重症肺炎（NSP），(2) 重症肺炎（SP），(3) 重症肺炎康复（Recov），(4) 无创机械通气（NIMV），(5) 有创机械通气（IMV），(6) 出院（Disch）和 (7) 死亡（Death）。

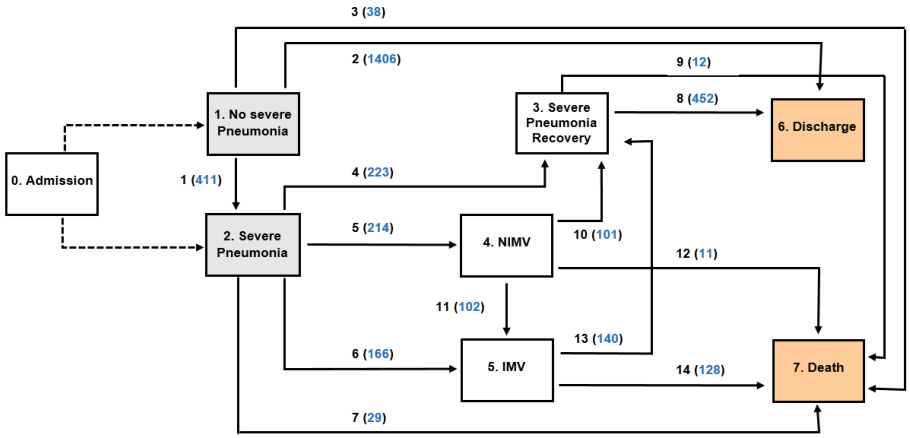

图 1. 用于建模 *COVID-19*住院患者病程轨迹的多状态模型图示。模型考虑了七种状态和 *14* 种可能的状态转移（括号中为各状态转移的样本量）。*NIMV*：无创机械通气，*IMV*：有创机械通气。

以下是相关说明：

1. 患者入院后（状态 0），会立即被分配到两个初始状态之一：非重症肺炎或重症肺炎。假

设该过程在时间$t = 0$开始时处于这两个状态之一。

2. 出院和死亡是吸收状态，这意味着患者一旦出院或死亡，就不能再次入院。

3. 本模型使用的时间尺度是自入院以来的天数。所有状态转移的转移时间（以天为单位）均为已知的确切时间。

4. 患者每天只能跳转到相邻状态。

有关数据和患者临床特征的更多详情，请参见 Pallar`es 等人（2023）；Garmendia 等人（2023）；Piulachs 等人（2023）。

本示例的主要目标是在不受一阶马尔可夫假设限制的情况下，研究患者状态的演变。为此，我们首先验证前述多状态模型中的哪些状态转移符合马尔可夫假设。接下来，我们将估计两个状态之间的转移概率，同时考虑这些概率也可能取决于紧邻的前一状态。最后，我们将比较因非重症肺炎入院的患者与因重症肺炎入院的患者的状态演变。

6.2. 直接转移和两步转移的描述

表1 总结了每种直接转移的患者人数，以及相应相关两步转移（状态连续，但时间不一定连续）的患者人数。例如，进行直接转移 Recov $\rightarrow$ Disch 的个体，可能来自 SP : SP $\rightarrow$ Recov $\rightarrow$ Disch（来自 NIMV : NIMV $\rightarrow$Recov $\rightarrow$ Disch）或 IMV : IMV $\rightarrow$ Recov $\rightarrow$ Disch。注意，我们仅考虑存在至少一个与状态 k 相邻的状态 j ($k \rightarrow j \rightarrow l$) 的直接转移 $j \rightarrow l$ 。

直接转移	样本量	两步转移	样本量	百分比。
重症肺炎 →康复	223	非重症肺炎 →重症肺炎 →康复	171	76.68
		重症肺炎 →康复	52	23.32
重症肺炎 →无创机械通气	214	非重症肺炎 → 重症肺炎 → 无创机械通气	134	62.62
		重症肺炎 → 无创机械通气	80	37.38
重症肺炎 →有创机械通气	166	非重症肺炎 →重症肺炎 →有创机械通气	92	55.42
		重症肺炎 →有创机械通气	74	44.58
重症肺炎 →死亡	29	非重症肺炎 → 重症肺炎 → 死亡	14	48.26
		重症肺炎 → 死亡	15	51.72
康复 → 出院	452	重症肺炎 → 康复 → 出院	223	49.34
		无创机械通气 → 康复 → 出院	96	21.24
		有创机械通气 → 康复 → 出院	133	29.42
康复 → 死亡	12	重症肺炎 → 康复 → 死亡	0	0
		无创机械通气 → 康复 → 死亡	5	41.67
		有创机械通气 → 康复 → 死亡	7	58.33
有创机械通气 → 死亡	128	重症肺炎 → 有创机械通气 → 死亡	71	57.26
		无创机械通气 → 有创机械通气 → 死亡	57	45.97
有创机械通气 → 康复	140	重症肺炎 → 有创机械通气 → 康复	95	67.86
		无创机械通气 → 有创机械通气 → 康复	45	32.14

状态：NSP：非重症肺炎，SP：重症肺炎，Recov：重症肺炎康复，NIMV：无创机械通气，IMV：有创机械通气，Disch：出院，Death：死亡

表1。 根据前一个即时状态汇总每种直接转移的两步路径

将表1 中出现的直接转移样本量与第5节引入的符号联系起来，我们可以看到，直接转移$h \to j$的样本量对应于风险患者人数：$\sum_{s=2}^{T} \tilde{Y}^{hj}$，而两步转移$h \to j \to l$ 的样本量与 $\sum_{s=2}^{T} \tilde{N}^{hjl}$ 并不一致，因为我们纳入了所有经历这一路径的患者，无论这些状态是否出现在连续的时间点上。

表1 显示，对于某一转移（例如，有创机械通气→康复），患者此前处于重症肺炎状态（68%）还是无创机械通气状态（32%），其比例差异显著。我们还可以考察重症肺炎→康复这一转移；若将入院时患有非重症肺炎的患者（76.68%）与入院时患有重症肺炎的患者（23.32%）分开，也会发现显著差异。对以下转移也可以作类似解读：

重症肺炎→无创机械通气。这表明模型可能不满足马尔可夫假设，因此在计算转移概率时，考虑前两个状态可能很重要。

6.3. 马尔可夫假设检验

为了检查哪些两步转移不符合一阶马尔可夫性，我们使用第 2.2 节所述的马尔可夫检验，并遵循 Titman 和 Putter(2020) 关于可进行检验的时间区间 $[t_0, t_{\max}]$的指导原则。基本上，比较仅限于个体数量充足的时间窗，以及具有紧邻前一状态的直接转移。

为了评估对数秩检验，我们在区间 $[1, 11]$ 内按每隔 0.5 天等距取点的网格计算统计量，适用于除转移 7（SP $\rightarrow$ 死亡）和 12（NIMV $\rightarrow$ 死亡）之外的所有转移；这两种转移的区间为 $[1, 7]$，转移 14（IMV $\rightarrow$ 死亡）的区间为 $[1, 16]$。

表 2 汇总了基于 5000 次野生自助法重抽样（Lin 等人 (1993)）得到的对数秩检验 p 值，并采用第 2.2 节所述的三种可能汇总统计量：加权均值、均值和上确界。对于每种转移（行），我们针对所有可能的前一状态以及整体卡方检验进行了检验。部分 p 值对应于全局检验。对于每种转移，转移强度比较先前处于固定状态 j（列）中的受试者与未处于该状态的受试者。

状态转移		NSP	SP	康复	NIMV	IMV	总体
4（重症肺炎→康复）	UM	0.005	0.005				**0.0042**
	WM	0.006	0.006				
	S	0.029	0.029				
5（重症肺炎→无创机械通气）	UM	$< 10^{-16}$	$< 10^{-16}$				**0.0018**
	WM	$< 10^{-16}$	$< 10^{-16}$				
	S	0.026	0.026				
6（重症肺炎→有创机械通气）	UM	0.009	0.009				**0.016**
	WM	0.002	0.002				
	S	0.042	0.042				
7（重症肺炎→死亡）	UM	0.106	0.106				0.196
	WM	0.120	0.120				
	S	0.340	0.340				
8（康复→出院）	UM	0.007	$< 10^{-5}$	0.165	$< 10^{-5}$	$< 10^{-5}$	$< \mathbf{10^{-16}}$
	WM	0.010	$< 10^{-5}$	0.186	$< 10^{-5}$	$< 10^{-5}$	
	S	0.104	$< 10^{-5}$	0.388	$< 10^{-5}$	$< 10^{-5}$	
9（康复→死亡）	UM	0.652	0.298	0.145	0.495	0.143	0.357
	WM	0.644	0.273	0.144	0.464	0.151	
	S	0.656	0.313	0.353	0.639	0.309	
10（无创机械通气→康复）	UM	0.594	0.190		0.694		0.609
	WM	0.588	0.183		0.717		
	S	0.831	0.432		0.892		
11（无创机械通气→有创机械通气）	UM	0.514	0.819		0.728		0.807
	WM	0.501	0.858		0.765		
	S	0.432	0.311		0.304		
12（无创机械通气→死亡）	UM	0.348	0.218		0.649		0.437
	WM	0.378	0.253		0.619		
	S	0.338	0.342		0.719		
13（有创机械通气→康复）	UM	0.564	$< 10^{-3}$		0.514	0.005	0.059
	WM	0.531	$< 10^{-3}$		0.456	0.005	
	S	0.780	0.034		0.376	0.037	
14（有创机械通气→死亡）	UM	0.296	0.663		0.318	0.099	0.305
	WM	0.296	0.674		0.269	0.100	
	S	0.471	0.369		0.264	0.205	

表 2. 通过计算各状态转移及各前一状态的马尔可夫检验所得的 p 值。计算了三种不同的汇总统计量：非加权均值（*UM*）、加权均值（*WM*）和上确界（*S*）。统计学上显著（*0.05*）的状态转移以粗体标示。

考虑总体 p 值，状态转移 4（重症肺炎 → 康复）、5（重症肺炎 → 无创机械通气）、6

（SP → IMV）和 8（康复 → 出院）显示出明显偏离马尔可夫假设，而转移 13（IMV → 康复）具有边缘显著性。此外，任一汇总检验都拒绝了转移 4、5 和 6 中的马尔可夫性。对于转移 8，从 NSP 和康复状态出发的上确界统计量不会拒绝马尔可夫性。最后，转移 13 中为 0.059 的全局 p 值主要源于 SP 和 IMV 状态所体现的非马尔可夫性。

这些发现表明，一旦患者病情危重，例如处于 NIMV 和 IMV 状态，其未来的临床演变与住院时被诊断为 NSP 还是 SP 无关。然而，对于住院时最初被诊断为 NSP 的患者和被诊断为 SP 的患者，其向 NIMV 或 IMV 的临床演变会有所不同。这些结果促使我们考虑二阶马尔可夫多状态模型，以研究 COVID-19患者在疫情第一波期间的临床演变。

6.4. 转移概率矩阵的估计

现在，我们将使用第 5 节中介绍的估计量，对七个可能的转移概率矩阵进行估计。正如之前提到的，我们知道确切的转移时间，因此可以根据处于转移风险中的患者人数以及最终完成转移的患者人数，轻松估计转移概率。对于每个矩阵的每一行，处于风险中的患者人数都各不相同。

这里我们给出矩阵 $\mathbf{P}^{(1)}$ 和 $\mathbf{P}^{(2)}$的估计。其余矩阵的估计方法类似，但矩阵$\mathbf{P}^{(6)}$ 和 $\mathbf{P}^{(7)}$ （状态 6 和 7 是吸收态）除元素 $(6,6)$ 和 $(7,7)$ 等于 1 外，均为零矩阵。

为了估计矩阵$\mathbf{P}^{(1)}$，我们从所有因非重症肺炎（NSP）入院的患者开始。患者住院后的第二天，他/她

仍可能处于非重症肺炎（NSP）状态，也可能转为重症肺炎（SP）、出院或死亡。因此，只有第 1、2、6 和 7 行的概率不为 0。为了估计$\mathbf{P}^{(1)}$ 第 1 行中的概率单元格，对于每个时间s，我们考虑所有在非重症肺炎（NSP）状态下至少连续待了两天的患者，即 $\sum_{s=2}^{43} Y_{11}(s-1) = 10577$（请注意，患者在非重症肺炎（NSP）状态下连续待两天的最大天数为 43）。在这 10577 名处于风险中的患者中，次日仍处于非重症肺炎（NSP）状态的患者人数为 $\sum_{s=2}^{43} N_{111}(s) = 8919$；而 $\sum_{s=2}^{43} N_{112}(s) = 257$ 转为重症肺炎（SP），$\sum_{s=2}^{43} N_{116}(s) = 1369$ 已出院，最后，$\sum_{s=2}^{43} N_{117}(s) = 32$ 已死亡。对于$\mathbf{P}^{(1)}$ 第 2 行中概率单元格的估计，我们采用类似的方法，从那些次日由非重症肺炎（NSP）转为重症肺炎（SP）的 $\sum_{s=2}^{36} Y_{12}(s-1) = 411$ 名患者开始。

$$\widehat{\mathbf{P}}_{(1)} = \begin{pmatrix} \frac{8919}{10577} & \frac{257}{10577} & 0 & 0 & 0 & \frac{1369}{10577} & \frac{32}{10577} \\ 0 & \frac{253}{411} & \frac{3}{411} & \frac{92}{411} & \frac{62}{411} & 0 & \frac{1}{411} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

我们以类似的方式估计 $\mathbf{P}^{(2)}$。在这种情况下，我们从曾在任何时间处于 SP 状态的所有患者开始。次日，这些患者仍可能处于 SP 状态，也可能转移至康复、无创机械通气、有创机械通气或死亡。因此，第 1 行和第 6 行为 0，因为 SP 不会直接转移至 NSP 或出院。对于第 2 行，处于风险中的患者数，也就是连续两个时间点均处于 SP 状态的患者数，为 $\sum_{s=2}^{47} Y_{22}(s-1)$ =2668。第 3 行从已由 SP 转移至康复的患者开始，共有$\sum_{s=2}^{47} Y_{23}(s-1) = 223$。同理，对于第 4 行和第 5 行，$\sum_{s=2}^{14} Y_{24}(s-1) = 214$ 名患者已

立即从 SP 转移至无创机械通气，而 $\sum_{s=2}^{23} Y_{25}(s-1) = 166$ 名患者从 SP 转移至有创机械通气。概率矩阵 $\mathbf{P}^{(3)}$、$\mathbf{P}^{(4)}$ 和 $\mathbf{P}^{(5)}$ 以类似方式估计，分别从处于康复、无创机械通气和有创机械通气状态的患者开始。

$$\widehat{\mathbf{P}}_{(2)} = \begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \frac{2307}{2668} & \frac{220}{2668} & \frac{68}{2668} & \frac{49}{2668} & 0 & \frac{24}{2668} \\ 0 & 0 & \frac{207}{223} & 0 & 0 & \frac{16}{223} & 0 \\ 0 & 0 & \frac{6}{214} & \frac{159}{214} & \frac{45}{214} & 0 & \frac{4}{214} \\ 0 & 0 & \frac{3}{166} & 0 & \frac{160}{166} & 0 & \frac{3}{166} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

6.5. 通过查普曼-科尔莫戈洛夫方程进行预测

第2.2 节中进行的马尔可夫检验显示，对于从重症肺炎（SP）出发的四种转移中的三种，一阶马尔可夫假设不成立：转移至无创机械通气（NIMV）、有创机械通气（IMV）和康复（Recov）。这表明患者是否被诊断为非重症肺炎（NSP）会影响其预后。二阶模型可根据患者住院时的诊断，预测未来事件发生的时间。定理3 中的Chapman-Kolmogorov扩展是计算相应概率的关键。

对于从重症肺炎 $\rightarrow$无创机械通气的转移，我们将计算 $s \in \{0,\ldots,6\}$ 概率 $p_{224} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 2)$，即入院时被诊断为重症肺炎且第二天仍处于重症肺炎状态的患者转为无创机械通气的概率。此外，$p_{124} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 1)$ 表示最初患有非重症肺炎且第二天转为重症肺炎的患者转为无创机械通气的概率。

入院时第二天转为重症肺炎的患者。我们在图 2中绘制了这些概率，可以看到初始状态对于最初几个时间点的重要性。最初被诊断为重症肺炎的患者（$X_1 = 2$, $X_2 = 2$）转为无创机械通气的概率，远低于最初被诊断为非重症肺炎的患者（$X_1 = 1, X_2 = 2$）转为无创机械通气的概率。随着时间推移，这两个概率之间的差距会逐渐缩小。

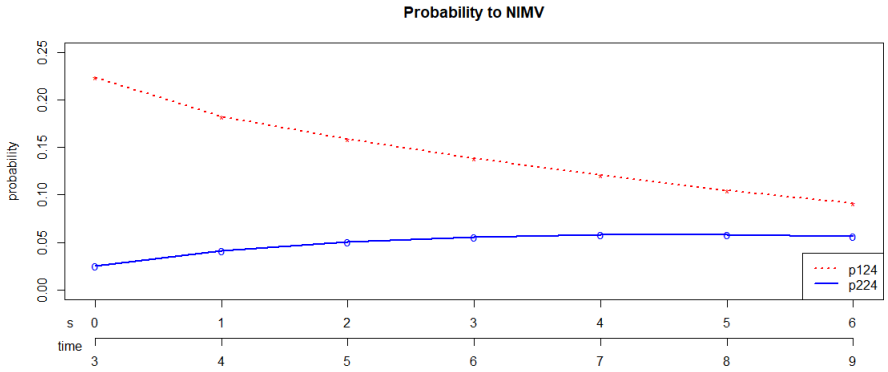

图2。 连续两天处于*SP*状态的患者从*SP*转至*NIMV*的概率（实线）与一天处于*NSP*状态、一天处于*SP*状态的患者（点线）相比。$p_{224} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 2)$ 与 $p_{124} = P(X_{3+s} = 4 | X_2 = 2, X_1 = 1)$

图 3 展示了同类图形，用于研究向 SP $\rightarrow$ IMV 的转移。在这种情况下，患者也根据其初始状态分为 NSP 或 SP。与图 2一样，最初诊断为 SP 的患者转为 IMV 的概率（$X_1 = 2$，$X_2 = 2$）远低于最初诊断为 NSP 的患者转为 IMV 的概率（$X_1 = 1$，$X_2 = 2$）。然而，两种概率都会随时间推移而上升，且在接下来的几天里，两者之间的差异一直存在。这表明，最初诊断为 NSP 的患者与诊断为 SP 的患者在需要呼吸机械通气（IMV）方面的预后不同。

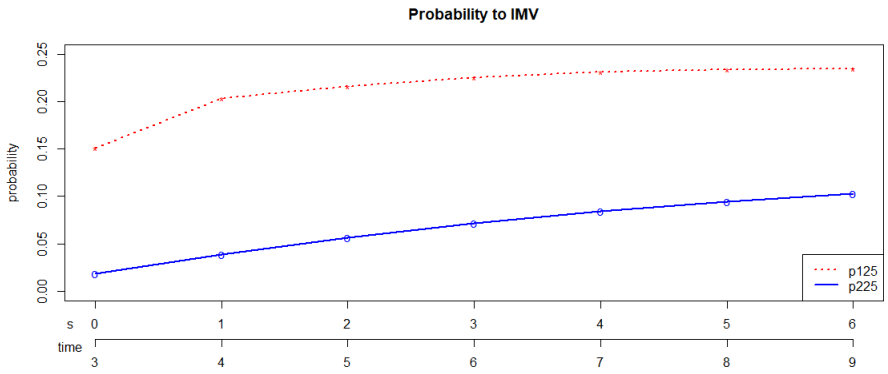

图3. 在重症肺炎（*SP*）中连续停留*2*天的患者，从*SP*转为有创机械通气（*IMV*）的概率（实线）；与在非重症肺炎（*NSP*）中停留*1*天、在*SP*中停留*1*天的患者（点线）相比。$p_{225} = P(X_{3+s} = 5|X_2 = 2, X_1 = 2)$ 与 $p_{125} = P(X_{3+s} = 5|X_2 = 2, X_1 = 1)$

7. 讨论

本文介绍了一种二阶马尔可夫多状态模型，并扩展了查普曼-科尔莫戈洛夫方程，以计算 r 步转移概率。我们使用 DIVINECOVID-19 数据估计转移概率，并根据患者住院前 2 天所处的状态，预测其接受 NIMV 和 IMV 的概率。

需要指出的是，Chapman-Kolmogorov 扩展基于对时间尺度的离散化，且仅在连续两个时间点的条件下计算。若前两个时间点并不连续，则无法计算转移概率，因为两个非连续时间点之间的时间窗口可能包含处于多个状态的受试者。

本文概述了两种估计转移概率的不同方法。第一种使用伯努利概率，这正是COVID示例中用于计算转移概率的方法；第二种使用条件概率。由于DIVINE 项目的数据是在第一波疫情结束一年后收集的，因此我们拥有完整的登记数据，也正因如此，目前我们只

开发了适用于完整（未删失）数据的这两种方法。不过，将这些估计量扩展到适用于右删失数据确实很有意义。第5.2 小节介绍的第二种估计方法为如何处理右删失数据提供了线索。该估计量是对每个时间点的比值取平均值，其中分子为发生特定转移的受试者人数，分母为处于风险中的受试者人数；这与用于估计累积风险函数的Nelson-Aalen估计量类似。若要进行全面的统计分析，还需要推导这些估计量的方差及其渐近分布。此外，针对完整数据和右删失数据的状态占有概率及转移强度估计量，也是值得关注的课题。所有这些问题都有待我们在未来的研究中探讨。

通过重新定义状态空间，可以将二阶马尔可夫模型转换为一阶马尔可夫模型。为此，可以创建由直接的一步转移构成的额外状态。例如，我们可以不只设定一个死亡状态，而是定义三个新状态，分别表示从无创机械通气、有创机械通气或康复状态转入死亡状态的患者。这些新状态的优势显而易见，因为我们可以运用一阶马尔可夫模型的全部知识。然而，新模型的状态和转移数量将大幅增加，解释起来也会变得繁琐。此外，由于需要估计的参数数量增加，估计这些参数所需的样本量也会随之增加，因此二阶马尔可夫方法更可取。

致谢

本研究由科学与创新部（西班牙）资助，项目编号为 PID2019- 104830RB-I00/ DOI(AEI): 10.13039/501100011033] ；并由加泰罗尼亚政府通过项目 2020PANDE00148 和 01421 SGR-Cat 2021 资助。我们感谢 DIVINE 小组的同事们提出的宝贵见解和投入的宝贵时间。

参考文献

Chakladar, S., R. Liao, W. Landau, M. Gamalo, 和 Y. Wang (2022). 用于建模 COVID-19 疾病进展和临床结局的机制转换离散时间多状态模型。生物制药研究统计学 *14*(1), 52–66。

Ching, W. K., E. S. Fung, and M. K. Ng (2003). 用于报童问题的高阶马尔可夫模型。运筹学会期刊 *54*(3), 291–298.

de Wreede, L., M. Fiocco, and H. Putter (2011). mstate：用于竞争风险分析和多状态模型的 R 软件包。统计软件期刊 *38*(7), 1–30.

Garmendia, L., J. Cort´es, andG. G´omez Melis(2023). MSMpred：通过多状态模型交互式建模并预测个体演变。*(*已投稿*)*。

Hougaard, P. (1999). 多状态模型：综述。*LifetimeData Anal. 5*(3), 239–26 4.

Islam, M. A. 和 R. I. Chowdhury（2006）。用于分析协变量依赖的高阶马尔可夫模型。应用数学建模 *30*(6)，477–488。

Kay, R.（1986）。用于生存研究中分析癌症标志物和疾病状态的马尔可夫模型。生物统计学 *42*(4)，855–865。

Lin, D. Y.、L. J. Wei 和 Z. Ying（1993）。使用基于鞅的残差的累积和检验 Cox 模型。*Biometrika80*(3)，557–572。

Logan, J. A.（1981）。纳入逆转效应的高阶马尔可夫过程结构模型。数学社会学杂志 *8*(1)，75–89。

Pallar`es, N., C. Teb´e, G. Abelenda-Alonso, A. Rombauts, I. Oriol, A. F. Simonetti, A. Rodr´ıguez-Molinero, E. Izquierdo, V. D´ıaz-Brito, G. Molist, G. G´omezMelis,

J. Carratal`a, S. Videla 和 M. study groups（2023）。大都市地区大流行病四波期间因 COVID-19 住院患者的护理上限特征与结局：一项多中心队列研究。《传染病与治疗》12(1), 273–289。

Piulachs, X., K. Langhor, M. Besal´u, N. Pallar`es, J. Carratal`a, C. Teb´e, and G. G´omez Melis（2023）。面向多队列事件史数据的半马尔可夫多状态方法。（已投稿）。

Raftery, A. E.（1985）。高阶马尔可夫链模型。《英国皇家统计学会杂志：*B*辑（方法论）》*47*（3），528–539。

´Rodr´ıguez-Girondo, M. and J. de U˜na Alvarez（2012）。疾病-死亡模型中马尔可夫性的非参数检验。《医学统计学》*31*（30），4416–4427。

Shamshad, A., M. Bawadi, W. Wan Hussin, T. Majid, and S. Sanusi（2005）。用于合成生成风速时间序列的一阶和二阶马尔可夫链模型。《能源》*30*（5），693–708。

Shorrocks, A. F.（1976）。收入流动性与马尔可夫假设。《经济学杂志》*86*(343)，566–578。

Titman, A. C. 和 H. Putter（2020）。多状态模型中马尔可夫性质的一般检验。《生物统计学》*23*(2)，380–396。

Tong, H.（1975）。利用赤池信息准则确定马尔可夫链的阶数。《应用概率杂志》*12*(3)，488–497。