arXiv:2303.07673v1 [math.ST] 14 Mar 2023

一般隐马尔可夫模型中的库尔贝克-莱布勒散度与赤池信息准则

预印本

Cheng-Der Fuh
统计系 浙江大学城市学院
杭州 310015，中国
cdffuh@gmail.com

Chu-LanMichael Kao统计研究所
国立阳明交通大学 新竹 30010，台湾
chulankao@gmail.com

Tianxiao Pang数学科学学院 浙江大学 杭州 310058，中国 txpang@zju.edu.cn

2023年3月15日

摘要

为刻画一般参数化隐马尔可夫模型中的库尔贝克-莱布勒散度和费舍尔信息，本文首先证明，对数似然及其导数可表示为马尔可夫迭代函数系统的加性泛函，随后通过这一表示对这两个量给出明确刻画。此外，我们证明，库尔贝克-莱布勒散度可由费舍尔信息确定的二次函数在局部近似。文中还给出了与 Cram´er-Rao 下界和 H´ajek-Le Cam 局部渐近极小极大定理相关的结果。作为这些结果的应用，我们从理论上论证了在一般隐马尔可夫模型中使用赤池信息准则（AIC）进行模型选择的合理性。最后，我们研究了三个具体模型：阶数为 $(p,\ q)$ 的高斯向量自回归移动平均模型、循环神经网络和时序受限玻尔兹曼机，以阐释我们的理论。

关键词 赤池信息准则、玻尔兹曼机、Cram´er-Rao 下界、费舍尔信息、H´ajek-Le Cam 定理、隐马尔可夫模型、Kullback-Leibler 散度、马尔可夫迭代函数系统、循环神经网络。

1 引言

Kullback-Leibler（KL）散度，也称为相对熵，已广泛应用于信息论、机器学习、统计学、计量经济学等领域。其应用包括信息论（[1]）、基于深度神经网络的语音识别（[2]）、化学动力学（[3]）、物理学（[4]）、统计学和计量经济学（[5, 6, 7, 8]）。KL散度的理论性质及其与费舍尔信息矩阵的关系已得到充分确立，尤其是对于具有独立同分布（i.i.d.）观测值的模型。

尽管如此，如今许多实际应用都建立在更复杂的有限状态隐马尔可夫模型（HMM）之上，甚至采用具有一般状态的一般隐马尔可夫模型（GHMM）。前者包括语音识别中的机器学习应用（[9]）和计算生物学应用（[10]），以及采用马尔可夫切换模型（[11, 12]）、马尔可夫转换GARCH模型（[13, 14]）等的计量经济学应用。后者包括机器学习中的因子隐马尔可夫模型（[15]）、切换状态空间模型（[16, 17, 18]）和对抗模型（[19]），统计学和计量经济学中的（G）ARCH模型（{v8）及随机波动率模型（SV）（[25, 26, 27]），以及其他学科中的相关应用。众所周知，KL散度在隐马尔可夫模型

和广义隐马尔可夫模型中也发挥着重要作用。例如， [28] 将KL散度用于马尔可夫切换模型中的模型选择。 [29] 和[30] 使用KL散度检测隐马尔可夫模型的变点，而 [31]研究了广义隐马尔可夫模型中的变点检测。 [32] 还通过弗雷德霍姆积分方程，为双状态隐马尔可夫模型提供了一种数值计算方法。另请参见 [33] 和[34], , 以及 [35] 关于马尔可夫模型中更一般的R´ényi熵的研究。这促使我们对广义隐马尔可夫模型中的KL散度进行理论研究。

需要注意的是，关于独立同分布模型已有许多结果和数学机制，但其中许多无法直接应用于广义隐马尔可夫模型，甚至无法直接应用于隐马尔可夫模型。主要原因是，隐马尔可夫模型或广义隐马尔可夫模型的对数似然并非独立同分布随机变量之和，甚至也不是马尔可夫链的函数；因此，无法直接采用经典的大数定律方法。相反， [36]运用金曼次可加遍历定理给出了广义 KL 散度，而 [37]则利用遍历过程来近似有限状态隐马尔可夫模型中的对数似然函数。 [38]运用香农-布雷曼-麦克米伦定理，得到一般状态隐马尔可夫模型中 KL 散度的极限。然而，这些结果要求隐马尔可夫模型具有平稳性，且没有刻画隐马尔可夫模型或广义隐马尔可夫模型中的 KL 散度，费舍尔信息也未得到刻画；因此，隐马尔可夫模型和广义隐马尔可夫模型的许多渐近性质仍未得到研究，包括 KL 散度及其与费舍尔信息的关系。

为正式而详细地解释这一现象，我们首先遵循 [39] 中的定义来定义广义隐马尔可夫模型。设 $\{X_n, n \geq 0\}$ 是一般状态空间 $\mathcal{X}$ 上的马尔可夫链，其转移概率核为 $p_\theta(x, \cdot) = P^\theta\{X_1 \in \cdot | X_0 = x\}$ ，其平稳概率为 $\pi(\cdot) := \pi_\theta(\cdot)$，相对于定义在 $\mathcal{X}$上的 σ-有限测度 Q ，其中 $\theta \in \Theta \subseteq \mathrm{R}^q$表示未知参数。设 $Y_{0:n}$ 为从 Y_0 到 Y_n 的观测值，满足 $Y_n \in \mathrm{R}^d$ ，其分布取决于 X_n和 Y_{n-1}，但与其他观测值相互独立。设 $f(\cdot; \theta|x, y)$ 为给定 $X_n = x$和 $Y_{n-1} = y$时 Y_n的概率密度函数（pdf），相对于定义在 R^d上的 σ-有限测度 $\tilde{Q}$。进一步设 $f(\cdot; \theta|x_0)$ 为给定 $X_0 = x_0$时 Y_0的概率密度函数。请注意，这一设定涵盖了许多有趣的例子，例如马尔可夫切换自回归模型、（G）ARCH模型、随机波动率模型、循环神经网络（RNN）以及时序受限玻尔兹曼机。当 $\mathcal{X}$ 是有限状态空间，且在给定 X_n的条件下 Y_n 相互独立时，这就是经典的隐马尔可夫模型。

对于给定的随机观测 $Y_{0:n}$，完整似然为

$$L(\theta; Y_{0:n}) = \int_{x_0 \in \mathcal{X}} \cdots \int_{x_n \in \mathcal{X}} \pi_\theta(x_0) f(Y_0; \theta|x_0) \times \prod_{t=1}^{n} p_\theta(x_{t-1}, x_t) f(Y_t; \theta|x_t, Y_{t-1}) Q(dx_n) \cdots Q(dx_0). \tag{1}$$

此外，将 $\ell(\theta; Y_{0:n}) := \log L(\theta; Y_{0:n})$ 记为对数似然。注意，在 (1) 中，为方便起见， X_0 的初始分布取为平稳分布 $\pi_\theta(\cdot)$；事实上，任何合适的初始分布 $\bar{\nu}(\cdot)$ 都适用。

于是，对于任意两个参数 θ_0 和 θ_1，KL散度 $K(\theta_1, \theta_0)$ 定义为

$$K(\theta_1, \theta_0) = \lim_{n \to \infty} \frac{1}{n} \left[\ell(\theta_1; Y_{0:n}) - \ell(\theta_0; Y_{0:n})\right], \quad P^{\theta_1}\text{-a.s.}, \tag{2}$$

其中， P^θ 表示当 $(Y_0, \cdots, Y_n)$ 按照 $L(\theta; \cdot)$ 分布时的概率测度。此外，在 P^{θ_0} 下的费舍尔信息可定义为

$$I(\theta_0) = -\lim_{n \to \infty} \frac{1}{n} \frac{\partial^2 \ell(\theta_0; Y_{0:n})}{\partial \theta_0 \partial \theta_0^t}, \quad P^{\theta_0}\text{-a.s.}, \tag{3}$$

其中，上标 t 表示转置；参见 [39] 第 2047 页最后一个公式 .

当 $\{Y_n, n \geq 0\}$ 是概率密度函数为 $f(y; \theta|x, y_0) = f(y; \theta)$ 的独立同分布随机变量时，则

$$\ell(\theta; Y_{0:n}) = \sum_{t=0}^{n} \log f(Y_t; \theta) \tag{4}$$

是独立同分布随机变量 $\{\log f(Y_t; \theta)\ t \geq 0\}$ 之和。因此，在某些正则条件下，由式 (2) 和强大数定律（SLLN），我们有

$$K(\theta_1, \theta_0) = E^{\theta_1}\left[\log f(Y_1; \theta_1) - \log f(Y_1; \theta_0)\right] = \int_{\mathbf{R}^d} \log \frac{f(y; \theta_1)}{f(y; \theta_0)} f(y; \theta_1) \tilde{Q}(dy), \tag{5}$$

其中， E^θ 表示在 P^θ下的期望。类似地，对于 $i, \ j = 1, \ \cdots, \ q$,

$$\frac{\partial^2 \ell(\theta; Y_{0:n})}{\partial \theta_i \partial \theta_j} = \sum_{t=0}^{n} \frac{\partial^2 \log f(Y_t; \theta)}{\partial \theta_i \partial \theta_j} \tag{6}$$

是独立同分布随机变量之和 $\{\frac{\partial^2 \log f(Y_t;\theta)}{\partial\theta_i\partial\theta_j},\ t\geq 0\}$，因此由（3）和强大数定律可得

$$[I(\theta)]_{i,j} = -E^{\theta}\left[\frac{\partial^2 \log f(Y_1;\theta)}{\partial\theta_i\partial\theta_j}\right]. \tag{7}$$

最后，借助 (5) 和 (7)，可知当 $\theta_1 \to \theta_0$时，我们有

$$K(\theta_1,\theta_0) = (\theta_1-\theta_0)^t\frac{I(\theta_0)}{2}(\theta_1-\theta_0) + O(\|\theta_1-\theta_0\|^3), \tag{8}$$

其中，$\|\cdot\|$ 表示欧几里得范数。

然而，对于隐马尔可夫模型和广义隐马尔可夫模型的情形，我们没有 (4)，因此无法通过强大数定律直接得到 (5)。同样，由于我们没有 (6)，也无法用相同的论证推导出 (7)。因此，尽管文献中早已猜想 (8) 成立（例如，参见 [40] 中的注2），但仍缺乏严格证明。

请注意，这些困难都与 (1) 的复杂结构密切相关。因此，本文采用一种创新的表示方法来处理广义隐马尔可夫模型中的对数似然 $\ell(\theta; Y_{0:n})$ 及其导数，从而避开这一复杂性。借此，我们刻画了 KL散度和费舍尔信息，并通过 (8) 中相应的收敛性证明了二者之间的关系。

基于这些新提出的刻画，我们进一步给出了广义隐马尔可夫模型（GHMM）的 Cram´er-Rao 下界和 H´ajek-Le Cam 局部渐近极小极大定理（[41]），表明独立同分布情形下的经典界限对广义隐马尔可夫模型依然成立。我们还证明，与独立同分布情形一样，KL散度满足非负性和可加性。然而，它通常不具有凸性，这与传统的独立同分布或马尔可夫链情形不同；在后两种情形中，KL散度是凸的。作为研究结果的另一项应用，我们进一步为在广义隐马尔可夫模型中使用赤池信息准则（AIC）进行模型选择提供了理论依据。

本文其余部分安排如下。第2节提出条件并陈述主要结果。第3节探讨其在AIC模型选择中的应用。为说明理论结果，第4节讨论了三个具体模型：阶数为（p，q）的高斯向量自回归移动平均模型、循环神经网络和时序受限玻尔兹曼机。第5节总结全文。理论结果的所有证明均列于附录。

2 主要结果

本节分为三个部分。第2.1节定义记号并陈述条件。第2.2节给出预备结果，表明对数似然及其导数可以表示为马尔可夫迭代函数系统（MIFS）的加性泛函。第2.3节陈述主要结果，包括广义隐马尔可夫模型的 KL散度、费 Fisher 信息矩阵的刻画，以及二者之间的关系。此外，我们还给出与 Cram´er-Rao 下界和 H´ajek-Le Cam 局部渐近极小极大定理相关的结果。

2.1 记号与条件

记 E_x^{θ} 为在初始状态为 $X_0 = x$ 时由 P^{θ} 定义的期望，记 $E_{(x,y)}^{\theta}$ 为在初始状态为 $(X_0, Y_0) = (x, y)$ 时由 P^{θ} 定义的期望。对于任意 $1 \leq i \leq q$ 和正整数 k，令 D_i 表示在真值 θ_0 的某个邻域 $N_{\delta}(\theta_0) := \{\theta : \|\theta-\theta_0\| < \delta\}$ 内，关于 θ 的第 i 维的偏导数，并令 $(D_i)^k$ 表示相应的第 k 阶偏导数。此外，对于给定的非负整数向量 $\nu = (\nu^{(1)},\cdots,\nu^{(q)})$，记 $|\nu| = \nu^{(1)}+\cdots+\nu^{(q)}$, $\nu! = \nu^{(1)}!\cdots\nu^{(q)}!$，并令 $D_{\theta}^{\nu} := D^{\nu} = (D_1)^{\nu^{(1)}}\cdots(D_q)^{\nu^{(q)}}$ 表示在 $N_{\delta}(\theta_0)$ 中关于 θ 的第 ν 阶导数。

本文其余部分将始终采用以下条件。

C1. 对于给定的 $\theta \in \Theta$，马尔可夫链 $\{(X_n, Y_n)\ n \geq 0\}$ 是非周期、不可约的，并且满足 s

$$\lim_{n\to\infty}\sup_{x\in\mathcal{X},y\in\mathbf{R}^d,|h|\leq w}\left|\frac{E_{(x,y)}^{\theta}[h(X_n,Y_n)] - \int h(s)\pi(ds)}{w(x,y)}\right| = 0,$$

$$\sup_{(x,y)\in\mathcal{X}\times\mathbf{R}^d}\frac{E_{(x,y)}^{\theta}[w(X_p,Y_p)]}{w(x,y)} < \infty,$$

存在某个权重函数 $w(\cdot,\ \cdot)$，并且 $p \geq 1$。假设

$$0 < p_\theta(x_0, x_1) < \infty \text{ for all } x_0, x_1 \in \mathcal{X}, \tag{9}$$

and

$$0 < \sup_{x \in \mathcal{X}} f(y_1; \theta | x, y_0) < \infty \text{ for all } y_0, y_1 \in \mathbf{R}^d. \tag{10}$$

由于 Q 是 σ-有限的,存在两两不相交的 $\mathcal{X}_n$，使得$\mathcal{X} = \cup_{n=1}^{\infty} \mathcal{X}_n$，且 $0 < Q(\mathcal{X}_n) < \infty$。假设 t

$$E^\theta \left[\sum_{n=1}^{\infty} \frac{1}{2^n} \sup_{x \in \mathcal{X}_n} f(Y_1; \theta | x, y_0) \right] < \infty \quad \text{for all } y_0 \in \mathbf{R}^d. \tag{11}$$

此外，令

$$\tilde{f}_\theta(y_0, y_1) = \sup_{x_0 \in \mathcal{X}} \int_{x \in \mathcal{X}} p_\theta(x_0, x) f(y_1; \theta | x, y_0) Q(dx),$$

并假设存在 $p \geq 1$ 使得

$$\sup_{(x_0, y_0) \in \mathcal{X} \times \mathbf{R}^d} E^\theta_{(x_0, y_0)} \left\{ \log \left((\tilde{f}_\theta(y_0, Y_1))^p \frac{w(X_p, Y_p)}{w(x_0, y_0)} \right) \right\} < 0, \tag{12}$$

$$\sup_{(x_0, y_0) \in \mathcal{X} \times \mathbf{R}^d} E^\theta_{(x_0, y_0)} \left\{ \tilde{f}_\theta(y_0, Y_1) \frac{w(X_1, Y_1)}{w(x_0, y_0)} \right\} < \infty. \tag{13}$$

C2. 真实参数 θ_0 是 Θ的内点。对于所有 $x \in \mathcal{X}$, y_0，$y_1 \in \mathbf{R}^d$, $\theta \in \Theta \subset \mathbf{R}^q$ 以及满足 $|\nu| \leq r$的 ν，偏导数 $D^\nu f(y_0; \theta|x)$和 $D^\nu f(y_1; \theta|x,\ y_0)$均存在。此外，对于所有 $x_0,\ x \in \mathcal{X}$, $\theta \mapsto p_\theta(x_0,\ x)$和$\theta \mapsto \pi_\theta(x_0)$在 θ_0的某个邻域 $N_\delta(\theta_0)$内具有 r阶连续导数。

C3. 对所有满足 ν、 $|\nu| \leq r$ 和 $x_0 \in \mathcal{X}$的

$$\int_{x \in \mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} |D^\nu \pi_\theta(x)| \, Q(dx) < \infty$$

and

$$\int_{x \in \mathcal{X}} \sup_{\theta \in N_\delta(\theta_0)} |D^\nu p_\theta(x_0, x)| \, Q(dx) < \infty.$$

C4. 对所有 $x \in \mathcal{X}$、 $y_0 \in \mathbf{R}^d$ 和 $\theta \in \Theta$,

$$E^\theta_x |D^\nu f(Y_0; \theta|x)|^r < \infty,\ \ E^\theta_{(x, y_0)} |D^\nu f(Y_1; \theta|x, y_0)|^r < \infty$$

对于 $1 \leq |\nu| \leq r$，且

$$E^\theta_x \left(\sup_{\theta \in N_\delta(\theta_0)} |D^\nu f(Y_0; \theta|x)|^r \right) < \infty,$$

$$E^\theta_{(x, y_0)} \left(\sup_{\theta \in N_\delta(\theta_0)} |D^\nu f(Y_1; \theta|x, y_0)|^r \right) < \infty$$

对于 $|\nu| = r + 1$。

C5.

$$E^{\theta_0} \left(\sup_{\|\theta - \theta_0\| < \delta} \sup_{x_0, x_0', x_1, x_1' \in \mathcal{X}} \frac{f(Y_0; \theta|x_0) f(Y_1; \theta|x_1, Y_0)}{f(Y_0; \theta|x_0') f(Y_1; \theta|x_1', Y_0)} \right)^r < \infty.$$

C6. 对于任意的 $\theta \in N_\delta(\theta_0)$ 和 ν，若满足 $|\nu| \leq r$,

$$|D^\nu p_\theta(x_0, x_1)| < \infty \text{ for all } x_0, x_1 \in \mathcal{X},$$

$$\sup_{x \in \mathcal{X}} |D^\nu f(y_1; \theta|x, y_0)| < \infty \text{ for all } y_0, y_1 \in \mathbf{R}^d,$$

$$E^\theta \left[\sum_{n=1}^{\infty} \frac{1}{2^n} \sup_{x \in \mathcal{X}_n} |D^\nu f(Y_1; \theta|x, y_0)| \right] < \infty \quad \text{for all } y_0 \in \mathbf{R}^d.$$

此外，令

$$\tilde{f}_\theta^\nu(y_0,y_1)=\sup_{x_0\in\mathcal{X}}\int_{x\in\mathcal{X}}D^\nu\left\{p_\theta(x_0,x)f(y_1;\theta|x,y_0)\right\}Q(dx),$$

并假设存在 $p\geq 1$ 使得

$$\sup_{(x_0,y_0)\in\mathcal{X}\times\mathbf{R}^d}E^\theta_{(x_0,y_0)}\left\{\log\left(\left|\tilde{f}_\theta^\nu(y_0,Y_1)\right|^p\frac{w(X_p,Y_p)}{w(x_0,y_0)}\right)\right\}<0,$$
$$\sup_{(x_0,y_0)\in\mathcal{X}\times\mathbf{R}^d}E^\theta_{(x_0,y_0)}\left\{\left|\tilde{f}_\theta^\nu(y_0,Y_1)\right|\frac{w(X_1,Y_1)}{w(x_0,y_0)}\right\}<\infty.$$

注释1。 条件C1和C2–C5与 [39], 中的条件C1和C2'–C5'基本相同。另一方面，附加条件C6旨在将C1中的(9)–(13)扩展到 θ_0某个邻域内的高阶导数。许多常用模型都满足这些条件，包括马尔可夫切换模型、自回归滑动平均模型、（G）ARCH模型以及随机波动率模型；详情参见 [39] 。此外，我们还将检验在采用特定分布的循环神经网络和时序受限玻尔兹曼机中，条件C1–C6是否也成立。

2.2 预备结果

[39] 已将对数似然 $\ell(\theta;\cdot)$表示为如下的MIFS加性泛函。具体而言，我们考虑函数空间

$$\mathbf{M}=\left\{h\middle|h:\mathcal{X}\mapsto\mathbf{R}\text{ is }Q\text{-measurable},\int_{x\in\mathcal{X}}|h(x)|Q(dx)<\infty\text{ and }\sup_{x\in\mathcal{X}}|h(x)|<\infty\right\}.$$

此外，对于 $t=1,\ \cdots,\ n$，在 $(\mathcal{X}\times\mathbf{R}^d)\ \times\mathbf{M}$ 上定义随机函数$\mathbf{P}_\theta(Y_0)$和$\mathbf{P}_\theta(Y_j)$，如下所示:

$$\mathbf{P}_\theta(Y_0)h(x)=\int_{x_0\in\mathcal{X}}f(Y_0;\theta|x_0)h(x_0)Q(dx_0),\quad\text{a constant functional,}$$
$$\mathbf{P}_\theta(Y_t)h(x)=\int_{s\in\mathcal{X}}p_\theta(s,x)f(Y_t;\theta|x,Y_{t-1})h(s)Q(ds),$$

并将两个随机函数的复合定义为:

$$\begin{aligned}&\mathbf{P}_\theta(Y_{t+1})\circ\mathbf{P}_\theta(Y_t)h(x)\\=&\int_{z\in\mathcal{X}}p_\theta(z,x)f(Y_{t+1};\theta|x,Y_t)\times\left(\int_{s\in\mathcal{X}}p_\theta(s,z)f(Y_t;\theta|z,Y_{t-1})h(s)Q(ds)\right)Q(dz).\end{aligned}$$

现在，考虑

$$M_n:=\mathbf{P}_\theta(Y_n)\circ\cdots\circ\mathbf{P}_\theta(Y_1)\circ\mathbf{P}_\theta(Y_0).\tag{14}$$

进一步记作 $\langle h\rangle:=\int_{x\in\mathcal{X}}h(x)Q(dx)$。那么，我们有

$$\begin{aligned}\ell(\theta,Y_{0:n})&=\log L(\theta;Y_{0:n})=\log\langle M_n\pi\rangle\\&=\sum_{t=1}^n\log\frac{\langle M_t\pi\rangle}{\langle M_{t-1}\pi\rangle}+\log\langle M_0\pi\rangle\\&=:\sum_{t=1}^n g^0(M_t^0,M_{t-1}^0)+g_0^0(M_0^0),\end{aligned}\tag{15}$$

其中

$$g^0(M_t^0,M_{t-1}^0)=\log\frac{\langle M_t\pi\rangle}{\langle M_{t-1}\pi\rangle},\quad g_0^0(M_0^0)=\log\langle M_0\pi\rangle.\tag{16}$$

换言之，$\ell(\theta)$是 $\{((X_n,Y_n)\ M_n)\ n\geq 0\}$的加性泛函。此外，[39] 表明$\{((X_n,Y_n)\ M_n)\ n\geq 0\}$ 构成一个由基于（14）的 MIFS 诱导、定义在状态空间（$\mathcal{X}\times\mathbf{R}^d)\ \times\mathbf{M}$上的遍历马尔可夫链。 [39] 进一步利用这一结果证明了对数似然的强大数定律。文献 [42]研究了$\{((X_n,Y_n)\ M_n)\ n\geq 0\}$ 收敛到其不变测度的速度。

以下引理将这一思路推广到 $\ell(\theta;\cdot)$的导数。为此，对任意 q维非负整数向量 $\nu=(\nu^{(1)},\cdots,\nu^{(q)})$，定义

$$W_n^{\nu}=D^{\nu}M_n=(D_1)^{\nu^{(1)}}\cdots(D_q)^{\nu^{(q)}}(M_n).$$

现在考虑所有阶数不超过 r 的导数。注意，对于固定整数 $r\geq 1$，恰好有 $K=(r+q)!/(r!q!)$个不同的 ν 满足 $|\nu|\leq r$。将所有此类 ν依次标记为 $\nu_1,\ \nu_2,\ \cdots,\ \nu_K$，并令 $W_n^{(r)}=(W_n^{\nu_1},\ W_n^{\nu_2},\ \cdots,\ W_n^{\nu_K})^t$。

第一个引理表明，我们可以通过 $W_n^{(r)}$构造一个MIFS。

引理1. 假设条件*C1–C6*对某个 $r\geq 1$成立。那么，对于任意 $\theta\in N_\delta(\theta_0)$，

$$\{((X_n,Y_n),W_n^{(r)}),n\geq 0\}$$

是一个非周期、$(\mathcal{X}\times\mathbf{R}^d)\times\mathbf{M}^K$-不可约且*Harris* 常返的马尔可夫链。

证明见补充材料。

第二个引理表明，$\ell(\theta;\cdot)$ 的导数可以表示为这个特定 MIFS 的加性泛函。

引理2。 假设条件*C1–C6*在某个 $r\geq 1$ 下成立。那么，对于任意 $\theta\in N_\delta(\theta_0)$ 以及任意满足 $|\nu|\leq r$ 的 q 维非负整数向量 ν，存在函数 g^{ν} 和 g_0^{ν}，使得

$$D^{\nu}\ell(\theta;Y_{0:n})=\sum_{t=1}^{n}g^{\nu}(W_t^{(|\nu|)},W_{t-1}^{(|\nu|)})+g_0^{\nu}(W_0^{(|\nu|)}). \tag{17}$$

证明见补充材料。

引理1和引理2分别与 [43], 中针对双层隐马尔可夫模型的引理3和引理5几乎相同，为完整起见，我们在此列出。结合引理1和引理2，我们可以应用 MIFS 的大数定律来计算 $D^{\nu}\ell(\theta;\cdot)$。这进一步导出了下一小节的主要结果。

2.3 主要结果

我们将利用引理1和引理2计算费舍尔信息、KL散度及其他性质。然而，由于这些量可能涉及不同的概率测度，以及在不同 θ 处求值的 $\ell(\theta;\cdot)$，因此需要补充一些记号以便明确陈述。对于 $i=0,\ 1$，令 $W_n^{(r)},_{\theta_i}$ 为使用在 $\theta=\theta_i$ 处求值的 $\ell(\theta;\ Y_{0:n})$ 构造的 $W_n^{(r)}$。此外，对于任意 $1\leq j,k\leq q$，令 $I_{jk}(\theta_0)$ 为费舍尔信息矩阵 $I(\theta_0)$ 的第 (j,k) 个分量。进一步记 $\vec{0}=(0,0,\cdots,0)\in\mathrm{R}^q$ 和 $\vec{e}_j=(0,\cdots,0,1,0,\cdots,0)\in\mathrm{R}^q$，其中 1 位于第 j 个分量。

我们的第一个定理表明，广义隐马尔可夫模型的费 Fisher 信息矩阵可以写成类似于式（7）的期望。**定理 1**。 假设条件*C1–C6*在 $r=2$下成立。则有

$$I(\theta_0)=-E^{\theta_0}_{\omega_{\theta_0,2}}\left[G(W_{1,\theta_0}^{(2)},W_{0,\theta_0}^{(2)})\right], \tag{18}$$

其中，$\omega_{\theta,\ r}$是 $\{((X_n,Y_n)\ W_{n\ \theta}^{(r)},\ )\ n\geq 0\}$的平稳分布； E^{θ}_{ω}表示在上述诱导马尔可夫链由 θ 控制且初始分布等于 ω时所取的期望，并且

$$G(w_1,w_0)=\begin{pmatrix}g^{\nu(1,1)}(w_1,w_0) & \cdots & g^{\nu(1,q)}(w_1,w_0)\\ \vdots & \ddots & \vdots\\ g^{\nu(q,1)}(w_1,w_0) & \cdots & g^{\nu(q,q)}(w_1,w_0)\end{pmatrix}, \tag{19}$$

其中 g^{ν} 在引理*2*中定义，且对于所有 $1\leq j$、 $k\leq q$，

$$\nu(j,k)=\vec{0}+\vec{e}_j+\vec{e}_k.$$

注2。 注意，可以将函数 G 与 $\ell(\theta;Y_{0:1})$ 的二阶导数联系起来。详情请参见下文的注9。

我们的第二个定理表明，广义隐马尔可夫模型的 KL散度也可以写成类似于（5）的形式，并且可以像（8）中那样，由费 Fisher 信息矩阵确定的二次函数进行局部近似。

定理2。假设　　条件*C1–C2*在 $r=0$下成立。那么，对于任意 $\theta_1 \in N_\delta(\theta_0)$，$K(\theta_1,\theta_0)$ 是良定义的，且

$$K(\theta_1,\theta_0) = E^{\theta_1}_{\omega_{\theta_1,0}}\left[g^0(W^{(0)}_{1,\theta_1}, W^{(0)}_{0,\theta_1})\right] - E^{\theta_1}_{\omega_{\theta_0,0}}\left[g^0(W^{(0)}_{1,\theta_0}, W^{(0)}_{0,\theta_0})\right], \tag{20}$$

其中，$E^{\theta_1}_{\omega_{\theta,0}}$ 的定义与定理*1*相同，函数 g^0 在引理*2*中定义。此外，若条件*C1–C6*在 $r=3$下成立，则当 $\theta_1 \to \theta_0$时，我们有

$$K(\theta_1,\theta_0) = (\theta_1-\theta_0)^t \frac{I(\theta_0)}{2}(\theta_1-\theta_0) + o(\|\theta_1-\theta_0\|^2). \tag{21}$$

Re注释3。 注意，可以将函数 g^0 与 $\ell(\theta; Y_{0:1})$联系起来。详情见下文注释10。

借助(18)和(20)，我们将证明以下与Cram´er-Rao下界和H´ajek-Le Cam局部渐近极小极大定理相关的结果。有限状态隐马尔可夫模型的H´ajek-Le Cam卷积定理见 [44]。对于任意 n，令 $\mathcal{E}_n$ 为基于 $Y_{0:n}$的 θ 的所有估计量构成的空间，$\mathcal{E}^U_n$为基于 $Y_{0:n}$的 θ 的所有ˆ ˆ无偏估计量构成的空间。记 $\theta_n := \theta_n(Y_{0:n})$为一个基于 $Y_{0:n}$.的估计量

定理3. 假设条件*C1–C6*在 $r=2$下成立。那么，对于任意 $v \in \mathbf{R}^q$ 和 $x \in \mathcal{X}$，

$$\lim_{n\to\infty} \inf_{\hat\theta_n \in \mathcal{E}^U_n} nE^{\theta_0}_x\left[\left(v^t(\hat\theta_n(Y_{0:n}) - \theta_0)\right)^2\right] \geq v^t I^{-1}(\theta_0) v. \tag{22}$$

此外，假设条件*C1–C6*在 $r=3$下成立。那么，对于 $\delta = (nv^tI(\theta_0)v)^{-1/2}$，有

$$\lim_{n\to\infty} \inf_{\hat\theta_n \in \mathcal{E}_n} \max_{\theta\in\{\theta_0,\theta_0+\delta v\}} nE^\theta_x\left[\|\hat\theta_n(Y_{0:n}) - \theta\|^2\right] \geq \frac{1}{16}\frac{\|v\|^2}{v^tI(\theta_0)v}. \tag{23}$$

注4. 当 $q=1$时（即 $\theta \in \mathbf{R}$），(22)化简为

$$\lim_{n\to\infty} \inf_{\hat\theta_n \in \mathcal{E}^U_n} nE^{\theta_0}_x\left[\left(\hat\theta_n(Y_{0:n}) - \theta_0\right)^2\right] \geq \frac{1}{I(\theta_0)}. \tag{24}$$

此外，对于这个一维情形，可以将(23)推广为

$$\lim_{c\to\infty}\lim_{n\to\infty} \inf_{\hat\theta_n\in\mathcal{E}_n} \sup_{\theta:|\theta-\theta_0|\leq c/\sqrt{n}} nE^\theta_x\left[\left(\hat\theta_n(Y_{0:n}) - \theta\right)^2\right] \geq \frac{1}{16}\frac{1}{I(\theta_0)}.$$

关于在紧集上取上确界的具体操作，参见 [45] 第8.7章。

注5. 方程(22)和(23)与 $Y_{0:n}$ 为独立同分布随机变量时的经典情形类似。特别地，(23)表明，只要估计量能够收敛到 θ_0的 $n^{-1/2}$邻域内，无论常数项如何，平方损失都一致地有下界。一个有趣的现象是，(23)中的常数仍与独立同分布情形中的相同，为$1\,16$。然而，对于这个广义隐马尔可夫模型版本，由于我们尚未刻画固定 n 时的费舍尔信息矩阵，因此该论证要求 n 趋于无穷，以便将均方误差与费舍尔信息 $I(\theta_0)$联系起来。

最后，借助 (20)，我们可以证明广义隐马尔可夫模型中 KL散度的以下性质。**推论1.** 假设
条件 *C1–C2* 对以下对象成立 $r=0$。那么，对于任意 $\theta_1 \in N_\delta(\theta_0)$，

*1. (*非负性*)* $K(\theta_1,\theta_0) \geq 0$。

*2. (*可加性*)* 假设 $Y_{0:n} = (Y^1_{0:}$ $L\ \theta\ y^1_{0:n}$ $L\ \theta\ y^2_{0:n}$ $\theta=\theta_0\ \theta^n_1$ $y_{0:n} =$ $y_{0:n} =$, $Y^2_{0:})$，并且对于任意 $n\in\mathbb{N}$和$(y^1_{0:n}, y^2_{0:n})$，有 $L(\theta;)\,(;)(;)$对于，。那么，

$$K(\theta_1,\theta_0) = K_1(\theta_1,\theta_0) + K_2(\theta_1,\theta_0),$$

其中，对于 $i=1,2$, K_i 是通过将(2) 中的 $L(\theta; Y_{0:n})$ 分别替换为 $L(\theta; Y^i_{0:n})$所定义的 *KL*散度。

注释6。请注意，当 $\{Y_n, n\geq 0\}$ 是独立同分布的有限混合随机变量序列或马尔可夫链时，还可以证明 $K(\theta_1,\ \theta_0)$ 是一个凸函数。然而，一般而言，这里并非如此。为了解释原因，考虑 $\{Y_n, n\geq 0\}$ 是马尔可夫链的情形。通过使用与 [46], 的定理1 类似的论证，可以证明

$$K(\theta_1,\theta_0) = E^{\theta_1}_\mu\left[\log f(Y_1;\theta_1|Y_0)\right] - E^{\theta_1}_\mu\left[\log f(Y_1;\theta_0|Y_0)\right]$$

$$= E_{\mu}^{\theta_1}\left[\log \frac{f(Y_1;\theta_1|Y_0)}{f(Y_1;\theta_0|Y_0)}\right]$$
$$= E_{\mu}^{\theta_1}\left[\log \frac{f(Y_1;\theta_1|Y_0)\mu(Y_0)}{f(Y_1;\theta_0|Y_0)\mu(Y_0)}\right], \tag{25}$$

其中， $\mu(\cdot)$ 是 $\{Y_n, n \geq 0\}$ 在 θ_1 下的不变测度， $E_{\mu}^{\theta_1}$ 是马尔可夫链由 P^{θ_1} 控制且 Y_0 服从 μ 分布时的期望。由此，应用对数和不等式的经典论证可得 $K(\theta_1, \theta_0)$ 的凸性。

然而，这一论证不适用于 $\{Y_n, n \geq 0\}$ 是隐马尔可夫模型的情况。这是因为，与 (25) 不同，(20) 中的两个期望分别是在不同的不变测度下计算的，因此无法像 (25) 那样将它们合并，也无法应用对数和不等式。这种非凸性是隐马尔可夫模型独有的特征，有别于独立同分布或马尔可夫链的情形。 [47] 中也观察到了 KL散度类似的非凸性。

我们用一个数值示例来结束这条注释。考虑一个三状态隐马尔可夫模型，其中 $\mathcal{X} = \{1, 2, 3\}$ $P\{X_1 =$ ，对所有， 均成立。对于观测值， $Y_n \in \{1, 2\}$ $P\{Y_n = 1|X_n = x\} =$ $x_1|X_0 = x_0\} = \frac{1}{2}$ $x_0\ x_1 \in \mathcal{X}$ 我们假设 q_x^{δ}，其中

$$(q_1^{\delta}, q_2^{\delta}, q_3^{\delta}) = \left(1, \frac{1}{2} + \delta, 0\right).$$

令 θ_0 为对应的概率测度，其中 $\delta = 0$；令 θ_1 为对应的概率测度，其中 δ从 0.1 变化至 0.2。计算得到的 $K(\theta_1, \theta_0)$ 如图1所示。正如预期的那样，随着δ 减小， $K(\theta_1, \theta_0)$也随之减小。然而，如上所述，图中显示 $K(\theta_1, \theta_0)$ 并非凸函数。

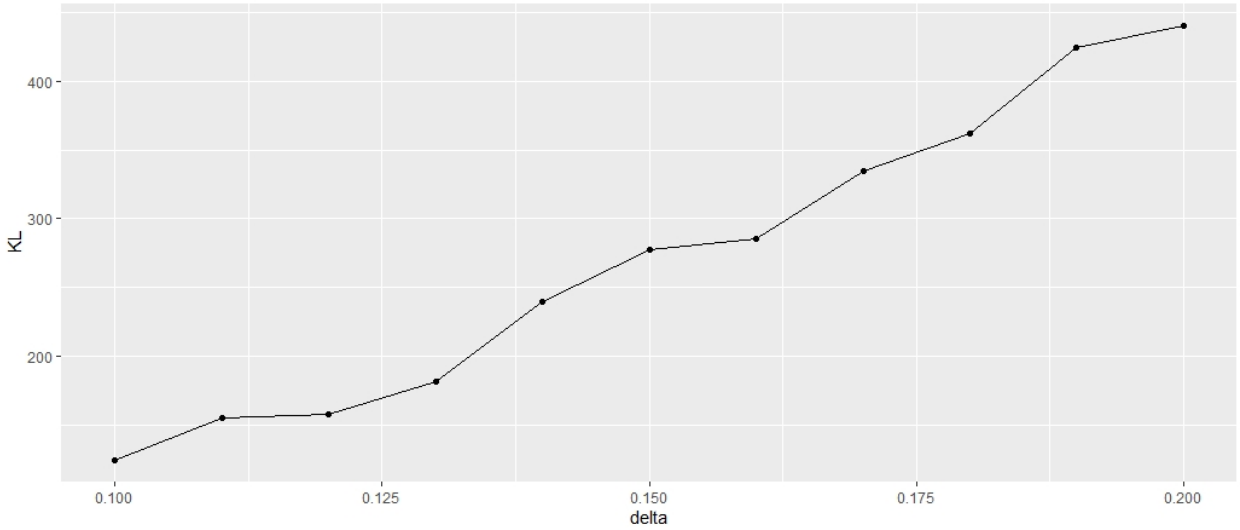

图1：三状态隐马尔可夫模型的 $K(\theta_1, \theta_0)$

该图展示了不同 θ_1 下的 $K(\theta_1, \theta_0)$。模型定义见注释6，其中 θ_0 对应于具有 $\delta = 0$ 的模型，而 θ_1 对应于 δ 从 0.1 变化至 0.2 的模型。对于每个 δ， y 轴表示相应的 $K(\theta_1, \theta_0)$，该值通过蒙特卡洛模拟，使用 (2) 并取 $X_0 = 0$ 和 $n = 50,000$ 计算得到。

注7。 这些条件可以略微放宽。例如，由于定理1–3和推论1仅涉及 θ_0 的邻域，C2 中的可微性假设可以放宽为仅要求 $\theta \in N_{\delta}(\theta_0)$，而不必涵盖所有 θ。此外，正如证明中所示，C4 的第二部分仅用于界定 $\log L(\theta; Y_{0:n})$ 的三阶导数，以便得到式（21）中的小-o 项。因此，对于与式（21）中的小-o 项无关的结果，C4 可以放宽。

3 AIC模型选择

本节将使用赤池信息准则（AIC）来确定广义隐马尔可夫模型（GHMM）的阶数。请注意，隐马尔可夫模型（HMM）的定义采用第1节中的一般定义。为此，我们提出一种客观方法，用于

确定具有有限状态空间的遍历一般隐马尔可夫模型的阶数。该方法利用 [43], 中似然比统计量的渐近性质，以及定理2中定义的 KL散度，以区分两种 GHMM分布。

$\{X_n, n \geq 0\}$ 称为 k阶马尔可夫链，如果 k 是使下式成立的最小非负整数

$$P(X_n|X_{n-1}, X_{n-2}, \cdots) = P(X_n|X_{n-1}, X_{n-2}, \cdots, X_{n-k}) \quad \text{for all } n.$$

接下来，我们假设 $\{X_n, n \geq 0\}$ 是有限状态空间 $\mathcal{D} = \{1, \cdots, l\}$上的 m阶马尔可夫链。已知对于 m阶马尔可夫链 $\{X_n, n \geq 0\}$，则 $\{(X_t, X_{t+1}, \cdots, X_{t+m-1})\ t = 0, \ldots, n-m+1, \cdots\}$构成一个马尔可夫链。根据第1节和式（1）中的定义，$\{Y_n, n \geq 0\}$ 称为 m阶广义隐马尔可夫模型。接下来，在本节中，我们进一步假设该 m阶广义隐马尔可夫模型满足条件C1-C6。

设 $Y_{0:n} = \{Y_0, \cdots, Y_n\}$ 为来自 m阶广义隐马尔可夫模型的观测值，使得 $Y_t \in \mathbf{R}^d$ 的分布依赖于 X_n 和 Y_{n-1}，但与其他变量独立。设 $f(\cdot;\ {}_m\theta|x, y)$为给定 $X_n = x$ 和 $Y_{n-1} = y$时 Y_n 的概率密度函数，关于$\mathbf{R}^d$上的一个 σ-有限测度，其中 ${}_m\theta = ({}_m\theta_1, \cdots, {}_m\theta_q)^t \in \Theta \subset \mathbf{R}^q$ 且 $q \geq m$。此外，设$f(\cdot;\ {}_m\theta|x_0)$为给定 $X_0 = x_0$时 Y_0 的概率密度函数。记 $i_j = (x_j, \cdots, x_{j+m-1})$表示 $j = 0, \cdots, n-m+1$，则该 m阶广义隐马尔可夫模型的完整似然为

$$L({}_m\theta; Y_{0:n}) = \sum_{x_0=1}^{l} \cdots \sum_{x_n=1}^{l} \pi_{{}_m\theta}(i_0) f(Y_0; {}_m\theta|x_0) \prod_{j=1}^{n-m+1} p_{{}_m\theta}(i_{j-1}, i_j) \prod_{j=1}^{n} f(Y_j; {}_m\theta|x_j, Y_{j-1}), \tag{26}$$

其中，$\pi_{{}_m\theta}(\cdot)$ 是马尔可夫链 $\{(X_t, X_{t+1}, \cdots, X_{t+m-1})\ t = 0, \ldots\ldots, n-m+1, \cdots\}$的平稳分布，而 $p_{{}_m\theta}(\cdot, \cdot)$ 是相应的转移概率核。记

$${}_m\hat{\theta} = \arg\max_{{}_m\theta \in \Theta} L({}_m\theta; Y_{0:n}).$$

也就是说，${}_m\theta$ 是基于该 m阶广义隐马尔可夫模型的观测值 $Y_{0:n}$ 得到的 ${}_m\theta$ 的最大似然估计量（MLE）。下文中，我们假设 ${}_m\theta$ 的真值为${}_m\theta_0$。记 $I({}_m\theta)$ 为与该 m阶广义隐马尔可夫模型对应的费舍尔信息矩阵。于是，对于任意 $\theta \in \Theta \subset \mathbf{R}^q$，我们记

$$\|\theta\|_J^2 = \theta^t I({}_m\ddot{\theta})\theta.$$

设 Δ 为上标上的差分算子，$\Delta_t^j = t^j - t^{j-1}$ 对应于 $j \geq 1$. 。与 [48], 中一样，我们考虑$k \leq m$的情形。假设由于马尔可夫链 $\{X_n, n \geq 0\}$的阶数从 m 变为 k，${}_k\theta$ 受限于参数空间 Θ，其中 ${}_k\theta$ 的 $\Delta_l^{m+1} - \Delta_l^{k+1}$ 个分量等于零。不失一般性，假设当马尔可夫链$\{X_n, n \geq 0\}$ 的阶数从 m 变为 k时，${}_k\theta$ 的最后 $\Delta_l^{m+1} - \Delta_l^{k+1}$ 个分量被限制为零。将上述受限参数空间记为 Θ_R，并令${}_k\theta$ 为该受限参数空间中 ${}_k\theta$的最大似然估计量。此外，令 ${}_k\theta \in \Theta_R$ 满足

$$\|{}_k\ddot{\theta} - {}_m\ddot{\theta}\|_J^2 = \min_{\theta \in \Theta_R} \|\theta - {}_m\ddot{\theta}\|_J^2.$$

也就是说，${}_k\theta$ 是依据 $\|\cdot\|_J$定义的度量，在空间 Θ_R 中对 ${}_m\theta$ 的投影。记 $p = q - (\Delta_l^{m+1} - \Delta_l^{k+1})$为受限参数空间 Θ_R的有效维数。于是显然有

$$\sum_{j=1}^{p} I({}_m\ddot{\theta})_{i,j} \times {}_k\ddot{\theta}_j = \sum_{j=1}^{q} I({}_m\ddot{\theta})_{i,j} \times {}_k\ddot{\theta}_j = \sum_{j=1}^{q} I({}_m\ddot{\theta})_{i,j} \times {}_m\ddot{\theta}_j, \quad \text{for all } i = 1, \cdots, q,$$

其中，$I({}_m\theta)_{i,\ j}$ 表示 $I({}_m\theta)$的第（i，j）个元素。这进一步意味着

$$\|{}_k\hat{\theta} - {}_m\ddot{\theta}\|_J^2 = \|{}_k\hat{\theta} - {}_k\ddot{\theta}\|_J^2 + \|{}_k\ddot{\theta} - {}_m\ddot{\theta}\|_J^2. \tag{27}$$

下文中，我们将 $\|{}_k\hat{\theta} - {}_m\ddot{\theta}\|_J^2$作为损失函数，因为它近似等于 $2K({}_m\ddot{\theta},\ {}_k\hat{\theta})$，而后者接近于 $2K({}_m\theta,\ {}_k\theta)$，其中 $K({}_m\theta,\ {}_k\theta)$表示 m阶广义隐马尔可夫模型与 k阶广义隐马尔可夫模型之间的KL散度。

记

$${}_k\lambda_m = \frac{L({}_m\hat{\theta}; Y_{0:n})}{L({}_k\hat{\theta}; Y_{0:n})}$$

为在已知 $Y_{0:n}$ 来自一个 m 阶广义隐马尔可夫模型的条件下所得的最大似然，与在已知 $Y_{0:n}$ 来自一个 k 阶广义隐马尔可夫模型的条件下所得的最大似然之比。那么

$$\begin{aligned}\log({}_k\lambda_m) &= \log L({}_m\hat{\theta}; Y_{0:n}) - \log L({}_k\hat{\theta}; Y_{0:n}) \\ &= \log \frac{L({}_k\ddot{\theta}; Y_{0:n})}{L({}_k\hat{\theta}; Y_{0:n})} - \log \frac{L({}_k\ddot{\theta}; Y_{0:n})}{L({}_m\hat{\theta}; Y_{0:n})}.\end{aligned} \tag{28}$$

考虑到这些关系

$$\frac{\partial \log L({}_m\theta; Y_{0:n})}{\partial {}_m\theta}\Big|_{{}_m\theta={}_m\hat{\theta}} = 0 \quad \text{and} \quad \frac{\partial \log L({}_k\theta; Y_{0:n})}{\partial {}_k\theta}\Big|_{{}_k\theta={}_k\hat{\theta}} = 0,$$

由泰勒展开可得

$$\begin{aligned}&\log L({}_k\ddot{\theta}; Y_{0:n}) = \log L({}_m\hat{\theta}; Y_{0:n}) \\ &\quad + \frac{n}{2}\sum_{i=1}^{q}\sum_{j=1}^{q}({}_k\ddot{\theta}_i - {}_m\hat{\theta}_i)({}_k\ddot{\theta}_j - {}_m\hat{\theta}_j)\frac{1}{n}\frac{\partial^2 \log L({}_m\theta; Y_{0:n})}{\partial {}_m\theta_i \partial {}_m\theta_j}\Big|_{{}_m\theta={}_m\hat{\theta}+\alpha({}_k\ddot{\theta}-{}_m\hat{\theta})}\end{aligned} \tag{29}$$

其中 $0 \le \alpha \le 1$，并且

$$\begin{aligned}&\log L({}_k\ddot{\theta}; Y_{0:n}) = \log L({}_k\hat{\theta}; Y_{0:n}) \\ &\quad + \frac{n}{2}\sum_{i=1}^{p}\sum_{j=1}^{p}({}_k\ddot{\theta}_i - {}_k\hat{\theta}_i)({}_k\ddot{\theta}_j - {}_k\hat{\theta}_j)\frac{1}{n}\frac{\partial^2 \log L({}_k\theta; Y_{0:n})}{\partial {}_k\theta_i \partial {}_k\theta_j}\Big|_{{}_k\theta={}_k\hat{\theta}+\beta({}_k\ddot{\theta}-{}_k\hat{\theta})}\end{aligned}$$

其中 $0 \le \beta \le 1$。应用（15）中的分解，可写为

$$\log L({}_m\theta; Y_{0:n}) = \sum_{t=1}^{n} g^0(M_t^0({}_m\theta), M_{t-1}^0({}_m\theta)) + g_0^0(M_0^0({}_m\theta)),$$

其中 g^0 和 g_0^0 的定义见（16）。于是，我们有

$$\begin{aligned}&\frac{1}{n}\frac{\partial^2 \log L({}_m\theta; Y_{0:n})}{\partial {}_m\theta_i \partial {}_m\theta_j}\Big|_{{}_m\theta={}_m\hat{\theta}+\alpha({}_k\ddot{\theta}-{}_m\hat{\theta})} \\ =&\frac{1}{n}\sum_{t=1}^{n}\frac{\partial^2 g^0(M_t^0({}_m\theta), M_{t-1}^0({}_m\theta))}{\partial {}_m\theta_i \partial {}_m\theta_j}\Big|_{{}_m\theta={}_m\hat{\theta}+\alpha({}_k\ddot{\theta}-{}_m\hat{\theta})} + \frac{1}{n}\frac{\partial^2 g_0^0(M_0^0({}_m\theta))}{\partial {}_m\theta_i \partial {}_m\theta_j}\Big|_{{}_m\theta={}_m\hat{\theta}+\alpha({}_k\ddot{\theta}-{}_m\hat{\theta})} \\ \to& -I({}_m\ddot{\theta})_{ij} \quad P^{{}_m\ddot{\theta}}\text{-a.s.}\end{aligned}$$

由于 ${}_m\hat{\theta}$ 实际上是 ${}_m\ddot{\theta}$ 的最大似然估计，且该估计渐近有效（参见 [43] 中的定理 2），因此 $\sqrt{n}\|{}_k\ddot{\theta} - {}_m\theta\|_J$ 有界。结合式 (29) 可得

$$\log \frac{L({}_k\ddot{\theta}; Y_{0:n})}{L({}_m\hat{\theta}; Y_{0:n})} = \frac{n}{2}({}_k\ddot{\theta} - {}_m\hat{\theta})I({}_m\ddot{\theta})({}_k\ddot{\theta} - {}_m\hat{\theta})^t + o_{P^{{}_m\ddot{\theta}}}(1) = \frac{n}{2}\|{}_k\ddot{\theta} - {}_m\hat{\theta}\|_J^2 + o_{P^{{}_m\ddot{\theta}}}(1).$$

¨ 同理，可以证明，若 $\sqrt{n}\|{}_k\overline{\theta} - {}_m\theta\|_J$ 有界，则

$$\log \frac{L({}_k\ddot{\theta}; Y_{0:n})}{L({}_k\hat{\theta}; Y_{0:n})} = \frac{n}{2}({}_k\ddot{\theta} - {}_k\hat{\theta})I({}_m\ddot{\theta})({}_k\ddot{\theta} - {}_k\hat{\theta})^t + o_{P^{{}_m\ddot{\theta}}}(1) = \frac{n}{2}\|{}_k\ddot{\theta} - {}_k\hat{\theta}\|_J^2 + o_{P^{{}_m\ddot{\theta}}}(1).$$

因此，由式 (28) 可得

$$\begin{aligned}&-2\log({}_k\lambda_L) \\ =&n\|{}_k\ddot{\theta} - {}_m\hat{\theta}\|_J^2 - n\|{}_k\ddot{\theta} - {}_k\hat{\theta}\|_J^2 + o_{P^{{}_m\ddot{\theta}}}(1) \\ =&n\|{}_k\ddot{\theta} - {}_m\ddot{\theta}\|_J^2 + n\|{}_m\ddot{\theta} - {}_m\hat{\theta}\|_J^2 - n\|{}_k\ddot{\theta} - {}_k\hat{\theta}\|_J^2 - 2n({}_k\ddot{\theta} - {}_m\ddot{\theta}, {}_m\hat{\theta} - {}_m\ddot{\theta})_J + o_{P^{{}_m\ddot{\theta}}}(1),\end{aligned} \tag{30}$$

其中，$({}_k\ddot{\theta} - {}_m\ddot{\theta}, {}_m\hat{\theta} - {}_m\ddot{\theta})_J = ({}_k\ddot{\theta} - {}_m\ddot{\theta})^t I({}_m\ddot{\theta})({}_m\hat{\theta} - {}_m\ddot{\theta})$ 表示由费舍尔信息矩阵 $I({}_m\theta)$¨定义的$({}_k\ddot{\theta} - {}_m\ddot{\theta})$与$({}_m\hat{\theta} - {}_m\ddot{\theta})$的内积。

根据 [43], 中的定理2，我们有

$$n\|_m\hat{\theta} - _m\ddot{\theta}\|_J^2 \to \chi_q^2(0) \text{ in distribution.}$$

注意，从几何上看，$_k\theta - _k\theta$ 近似为 $_m\theta - _m\theta$ 在 Θ_R空间上的投影，因此只要 $\sqrt{n}\|_k\theta - _m\theta\|_J$ 有界，就有

$$n\|_m\hat{\theta} - _m\ddot{\theta}\|_J^2 - n\|_k\hat{\theta} - _k\ddot{\theta}\|_J^2 \to \chi_{q-p}^2(0) \text{ in distribution,}$$
$$n\|_k\hat{\theta} - _k\ddot{\theta}\|_J^2 \to \chi_p^2(0) \text{ in distribution,}$$

且 $n\|_m\hat{\theta} - _m\ddot{\theta}\|_J^2 - n\|_k\hat{\theta} - _k\ddot{\theta}\|_J^2$与 $n\|_k\hat{\theta} - _k\ddot{\theta}\|_J^2$渐近独立。注意，[43]的定理2还表明，$n(_k\theta - _m\theta,\ _m\theta - _m\theta)_J$的渐近分布的标准差等于 $\sqrt{n}\|_k\theta - _m\theta\|_J$。因此，如果后一项足够大，那么与项 $n\|_k\ddot{\theta} - _m\ddot{\theta}\|_J^2$相比，$n(_k\ddot{\theta} - _m\ddot{\theta},\ _m\hat{\theta} - _m\ddot{\theta})_J$可以忽略不计。

那么，根据上述论证以及方程（27）和（30），$\{p + [-2\log(_k\lambda_m) - (q-p)]\}/n = [-2\log(_k\lambda_m) - q + 2p]/n$可作为 $E^{m\ddot{\theta}}\|_k\hat{\theta} - _m\ddot{\theta}\|_J^2$的一个有效估计量。也就是说，可以通过最小化以下赤池信息准则来确定广义隐马尔可夫模型的阶数

$$-2\log(_k\lambda_m) - q + 2p = -2\log(_k\lambda_m) - 2(\Delta_l^{m+1} - \Delta_l^{k+1}) + q \tag{31}$$

回忆 $p = q - (\Delta_l^{m+1} - \Delta_l^{k+1})$。从（31）式右侧去掉与 k 无关的项，我们得到以下广义隐马尔可夫模型的赤池信息准则：

$$\text{AIC}(k) = 2\log L(_k\hat{\theta}; Y_{0:n}) + 2\Delta_l^{k+1}. \tag{32}$$

采用相同的方法可以证明，对于给定的 m阶广义隐马尔可夫模型，用于选择隐状态数量的赤池信息准则为

$$\text{AIC}(k) = 2\log L(_k\hat{\theta}; Y_{0:n}) + 2\Delta_k^{m+1}, \tag{33}$$

其中，k 表示 m阶广义隐马尔可夫模型中的隐状态数，$_k\theta$ 表示该模型中参数的最大似然估计。[49] 研究了 Y_n 仅取决于 X_n 时隐马尔可夫模型的赤池信息准则模型选择。

4 示例

例 1. 高斯向量自回归滑动平均模型。

考虑一个 m 维、阶数为（p，q）的高斯向量自回归滑动平均（VARMA）模型（见 [50]），使得对所有 $n \geq 0$,

$$\sum_{j=0}^{p} \alpha_j Y_{n-j} = \sum_{j=0}^{q} \beta_j \epsilon_{n-j}, \tag{34}$$

其中，α_j 和 β_j 是 $m\times m$ 实值矩阵，且 $\alpha_0 = \beta_0 = I_m$（$m\times m$ 单位矩阵）；$\{\epsilon_n, n \geq 1\}$ 是均值为零、协方差矩阵为 Σ的独立同分布 m维正态随机变量。这里假设p、q 和 Σ 已知，因此未知参数可表示为

$$\theta = vec\{\alpha_1, \cdots, \alpha_p, \beta_1, \cdots, \beta_q\},$$

对于任意矩阵 M，$vec\{M\}$ 表示将 M的所有列依次上下堆叠而成的向量。

众所周知，式 (34) 中的 VARMA模型可以用多种方式表示为线性状态空间模型（LSSM），参见 [51]。例如，令 $h = \max(p, q)$，并假设 $X_n \in \mathbf{R}^{hm}$ 和 $Y_n \in \mathbf{R}^m$ 满足以下 LSSM

$$X_{n+1} = \Phi X_n + F\epsilon_n,$$
$$Y_n = HX_n + \epsilon_n, \tag{35}$$

其中

$$\Phi = \begin{pmatrix} -\alpha_1 & I_m & 0_m & \cdots & 0_m \\ -\alpha_2 & 0_m & I_m & \cdots & 0_m \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ -\alpha_{h-1} & 0_m & 0_m & \cdots & I_m \\ -\alpha_h & 0_m & 0_m & \cdots & 0_m \end{pmatrix}_{hm\times hm}, \quad F = \begin{pmatrix} \beta_1 - \alpha_1 \\ \beta_2 - \alpha_2 \\ \vdots \\ \beta_h - \alpha_h \end{pmatrix}_{hm\times m},$$

且 $H=(I_m,\ 0_m,\ \cdots,\ 0_m)$，其中 0_m 为 $m\times m$ 零矩阵，$\alpha_i=0_m$ 对所有 $i>p$ 均成立，且 $\beta_j=0_m$ 对所有 $j>q$ 均成立。然后，[51] 证明了（35）中的 Y_n 满足（34）中的 VARMA模型；另见 [50], [52] 和 [53]。

由于 (35) 具有线性状态空间模型的形式，考虑通过以下卡尔曼滤波器方程得到的样本创新 $\hat{\epsilon}_n$：

$$\begin{cases} P_{n+1}=\Phi P_n\Phi^t+\Sigma-(\Phi P_nH^t)(HP_nH^t)^{-1}(HP_n\Phi^t), \\ K_n=(\Phi P_nH^t)(HP_nH^t)^{-1}, \\ \hat{X}_{n+1|n}=(\Phi-K_nH)\hat{X}_{n|n-1}+K_nY_n, \\ \hat{Y}_{n|n-1}=H\hat{X}_{n|n-1}, \\ \hat{\epsilon}_n=Y_n-\hat{Y}_{n|n-1}, \end{cases} \tag{36}$$

其中 P_1 由 $P_1=\Phi P_1\Phi^t+FF^t$ 给出。进一步记 $\hat{\Sigma}_n=E[\hat{\epsilon}_n\hat{\epsilon}_n^t]$。于是，VARMA 的对数似然可写为

$$\ell(\theta;Y_{1:n})=\sum_{i=1}^{n}\left\{-\frac{n}{2}\log 2\pi-\frac{1}{2}\log|\hat{\Sigma}_i|-\frac{1}{2}\hat{\epsilon}_i^t\hat{\Sigma}_i^{-1}\hat{\epsilon}_i\right\},$$

这进一步意味着

$$\begin{aligned} &-\frac{1}{n}E^{\theta_0}\left[\frac{\partial^2\ell(\theta_0;Y_{1:n})}{\partial\theta\partial\theta^t}\right] \\ =&\frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}\left(\frac{\partial vec\left\{\hat{\Sigma}_i\right\}}{\partial\theta}\right)^t(\hat{\Sigma}_i\otimes\hat{\Sigma}_i)^{-1}\left(\frac{\partial vec\left\{\hat{\Sigma}_i\right\}}{\partial\theta}\right) \\ &+\frac{1}{n}\sum_{i=1}^{n}E^{\theta_0}\left[\left(\frac{\partial\hat{\epsilon}_i}{\partial\theta}\right)^t\hat{\Sigma}_i^{-1}\left(\frac{\partial\hat{\epsilon}_i}{\partial\theta}\right)\right], \end{aligned} \tag{37}$$

其中 $\otimes$ 表示克罗内克积。详情见 [50] 。

为应用定理1和定理2，我们需要检验条件C1-C6是否成立。[39]的第6.2节检验了条件C1-C6对自回归滑动平均成立。对于线性状态空间模型，[54] 的第16.5.1节表明，C1（ω-一致性）对任意维度的线性状态空间模型均成立，因此，采用相同的论证可知，C1对 VARMA (34) 和 (35) 成立。利用 ϵ_n的正态性，可以直接检验条件C2-C6成立。

现在，[50] 已表明，当 $n\to\infty$ 时，$\hat{\epsilon}_n$ 的极限分布与 ϵ_1 的分布相同。此外，由于 $\hat{\epsilon}_n$ 和 ϵ_1 都服从高斯分布，$\hat{\epsilon}_n$ 的协方差矩阵收敛于 ϵ_1 的协方差矩阵；换言之，$\Sigma_n\to\Sigma$，且它与 θ 独立。因此，当 $n\to\infty$ 时，式（37）中的第一项趋于零，于是，

$$I(\theta_0)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E^{\theta_0}\left[\left(\frac{\partial\hat{\epsilon}_i}{\partial\theta}\right)^t\Sigma^{-1}\left(\frac{\partial\hat{\epsilon}_i}{\partial\theta}\right)\right]. \tag{38}$$

此外，已知当 $n\to\infty$ 时，$K_n\to K_\infty$ 收敛于某个有限常数矩阵 K_∞（取决于 θ）。因此，式（36）的渐近形式为

$$\begin{cases} \hat{X}^{\infty}_{n+1|n}=(\Phi-K_\infty H)\hat{X}^{\infty}_{n|n-1}+K_\infty Y_n, \\ \hat{Y}^{\infty}_{n|n-1}=H\hat{X}^{\infty}_{n|n-1}, \\ \hat{\epsilon}^{\infty}_{n}=Y_n-\hat{Y}^{\infty}_{n|n-1}. \end{cases} \tag{39}$$

由式（38）可得

$$I(\theta_0)=\lim_{n\to\infty}\frac{1}{n}\sum_{i=1}^{n}E^{\theta_0}\left[\left(\frac{\partial\hat{\epsilon}^{\infty}_i}{\partial\theta}\right)^t\Sigma^{-1}\left(\frac{\partial\hat{\epsilon}^{\infty}_i}{\partial\theta}\right)\right]. \tag{40}$$

现在，由（39）可得

$$\frac{\partial\hat{X}^{\infty}_{n+1|n}}{\partial\theta}=\frac{\partial(\Phi-K_\infty H)}{\partial\theta}\hat{X}^{\infty}_{n+1|n}+(\Phi-K_\infty H)\frac{\partial\hat{X}^{\infty}_{n+1|n}}{\partial\theta}+\frac{\partial K_\infty}{\partial\theta}Y_n, \tag{41}$$

这表明 $Z_n = \left(X_n, Y_n, \hat{X}_{n|n-1}, \frac{\partial}{\partial\theta}\hat{X}_{n|n-1}\Big|_{\theta=\theta_0} \right)$ 构成一个马尔可夫链。此外，注意到 $\hat{\epsilon}_n^{\infty} = \hat{Y}_n - Y_{n|n-1}^{\infty} = Y_n - HX_{n|n-1}^{\infty}$，由此可得

$$\frac{\partial\hat{\epsilon}_n^{\infty}}{\partial\theta} = -\frac{\partial H}{\partial\theta}\hat{X}_{n|n-1}^{\infty} - H\frac{\partial\hat{X}_{n|n-1}^{\infty}}{\partial\theta}. \tag{42}$$

换言之，$\frac{\partial}{\partial\theta}\hat{\epsilon}_n^{\infty}$是 Z_n的函数。因此，定理1本质上表明，在$\omega_{\theta_0,\,2}$下，Z_n 是平稳的，因此

$$I(\theta_0) = E_{\omega_{\theta_0},2}^{\theta_0}\left[\left(\frac{\partial\hat{\epsilon}_1^{\infty}}{\partial\theta} \right)^t \Sigma^{-1} \left(\frac{\partial\hat{\epsilon}_1^{\infty}}{\partial\theta} \right) \Big|_{\theta=\theta_0} \right]. \tag{43}$$

结果（43）与 [50], 中的结果相同；他们用 $\frac{\partial\epsilon}{\partial\theta}$ 表示一个随机变量，其分布为 $\frac{\partial\hat{\epsilon}_n}{\partial\theta}$ 在 $n \to \infty$时的极限分布。这里，我们从增广马尔可夫链的视角，根据不变概率测度推导出费舍尔信息（43）。一般线性状态空间模型的费舍尔信息可见 [55]。

请注意，上述结果之所以成立，是因为如果（X_n，Y_n）服从参数为 θ的线性状态空间模型，那么在 P^{θ}下，$\hat{\epsilon}_n$ 的极限分布与 ϵ_1相同。若要得到KL散度的表示，我们需要求出在 $P^{\theta'}$ 下、$\theta' \neq \theta$对应的$\hat{\epsilon}_n$ 的极限分布。这是因为（20）中的第二个期望涉及在参数为 $\theta_0 \neq \theta_1$的 P^{θ_1}下的 $W_{(0)\,1,\,\theta_0}$和 $W_0^{(0)},\,\theta_0$。相较之下，定理2中的（21）以费舍尔信息给出了KL散度的局部近似。此外，我们还从理论上论证了使用（33）中的AIC模型选择准则来选择（34）和（35）中的（p，q）。线性状态空间模型的KL散度及AIC模型选择的计算方法见 [56]。

例2. 循环神经网络。

首先，我们考虑如下线性循环神经网络（RNN）：

$$Y_n = \mu_{y,n} + \sigma_{y,n}\varepsilon_n, \tag{44}$$

其中 $(\mu_{y,n}, \sigma_{y,n}^2) \sim \varphi_{\tau}(X_{n-1})$，其中 φ_{τ}可以是任意高度灵活的函数，例如神经网络。$\sigma_{y,n} > 0$几乎必然成立，$\varepsilon_n \sim N(0,1)$ 是一个独立同分布随机变量序列，并且对于所有 n，ε_n都独立于 $\{Y_{n-k}, k \geq 1\}$。

为说明线性循环神经网络的广义隐马尔可夫模型方法，将（44）作为 Y_n 的输出模型后，线性循环神经网络使用以下递推方程更新其隐状态:

$$X_n = f_{\theta}(\phi_{\tau}(Y_n), X_{n-1}) = \delta + \alpha Y_{n-1} + \beta X_{n-1}, \tag{45}$$

其中 $\delta > 0$, $\alpha > 0$, 和 $\beta > 0$ 为常数。

如 [42] 所述，当（44）中的 $\mu_{y,n} = 0$ 和 $\sigma_{y,n}^2 = X_n$，以及（45）中以 Y_{n-1}^2 替换 Y_{n-1} 时，线性循环神经网络模型（44）和（45）可视为著名的 GARCH(1,1) 模型。然而，一般而言，（44）中定义的 $\mu_{y,n}$ 和 $\sigma_{y,n}$可以是 X_n 的非线性函数。

为明确计算费舍尔信息和KL散度，我们考虑式（44）的一种特定形式，即一个简单的GARCH(1，1)模型（见 [21]），如下:

$$Y_n = \sigma_n\epsilon_n, \quad \text{and } \sigma_n^2 = \delta + \alpha Y_{n-1}^2 + \beta\sigma_{n-1}^2,$$

其中 $\delta > 0$、$\alpha > 0$ 和 $\beta > 0$ 为常数，且 $\alpha + \beta < 1$；$\{\epsilon_n, n \geq 1\}$ 为独立同分布的标准正态随机变量，且 ϵ_n 与 $\{Y_t, t = 1, \cdots, n-1\}$相互独立。

[39] 的第6.3节已验证GARCH(p，q)模型满足条件C1-C6。对于线性循环神经网络，条件C1（ω一致性）可通过状态空间表示在 [54] 的第16.5.1节中找到。由于 ε_n的正态性假设，条件C2-C6均成立。因此，可以应用定理1和定理2。

记为 $\theta = (\delta, \alpha, \beta)^t$。注意，GARCH(1, 1) 模型的对数似然函数可表示为

$$\ell(\theta; Y_{1:n}) = \sum_{t=1}^{n}\left\{ -\frac{1}{2}\log 2\pi - \frac{1}{2}\log\sigma_t^2 - \frac{1}{2}\frac{Y_t^2}{\sigma_t^2} \right\}, \tag{46}$$

这进一步意味着

$$-E^{\theta_0}\left[\frac{\partial^2\ell(\theta; Y_{1:n})}{\partial\theta\partial\theta^t} \right] = \sum_{t=1}^{n} E^{\theta_0}\left[\frac{1}{2\sigma_t^2}\frac{\partial\sigma_t^2}{\partial\theta}\frac{\partial\sigma_t^2}{\partial\theta^t} \right]. \tag{47}$$

详见 [57] 。事实上，定理1基本上表明

$$I(\theta_0) = E^{\theta_0}_{\omega_{\theta_0,2}}\left[\frac{1}{2\sigma_1^2}\frac{\partial\sigma_1^2}{\partial\theta}\frac{\partial\sigma_1^2}{\partial\theta^t}\Big|_{\theta=\theta_0}\right], \tag{48}$$

其中 $\omega_{\theta_0,\,2}$ 是 $\left\{\left(\sigma_n^2,\,\frac{\partial\sigma_n^2}{\partial\theta}\right),\ \ n\geq 0\right\}$ 的平稳分布。

我们还可以将（48）与 [57] 中的公式联系起来。作为说明，我们仅考虑关于 β 的偏导数。直接计算可得

$$\frac{\partial\sigma_n^2}{\partial\beta} = \sigma_{n-1}^2 + \beta\frac{\partial\sigma_{n-1}^2}{\partial\beta} = \cdots = \sum_{k=1}^{n}\beta^{k-1}\sigma_{n-k}^2 + \beta^n\frac{\partial\sigma_0^2}{\partial\beta}.$$

（另见 [58] 中的方程（7）。）此外，在 $\omega_{\theta_0,\,2}$ 条件下，使用将马尔可夫链扩展为双无限平稳序列的经典方法（例如，参见 [37] 的第4节），我们有

$$\frac{1}{2\sigma_n^2}\left(\frac{\partial\sigma_n^2}{\partial\beta}\right)^2 = \frac{1}{2\sigma_n^2}\left(\sum_{k=1}^{\infty}\beta^{k-1}\sigma_{n-k}^2\right)^2, \tag{49}$$

它与 $|\beta|<1$ 中的 $\frac{1}{2\sigma_0^2}\left(\sum_{k=1}^{\infty}\beta^{k-1}\sigma_{-k}^2\right)^2$ 具有相同的分布。结合（47）和（48），我们得到

$$E^{\theta_0}_{\omega_{\theta_0,2}}\left[\frac{1}{\sigma_1^2}\frac{\partial\sigma_1^2}{\partial\beta}\frac{\partial\sigma_1^2}{\partial\beta^t}\right] = \lim_{n\to\infty}E^{\theta_0}\left[\frac{1}{\sigma_n^2}\frac{\partial\sigma_n^2}{\partial\beta}\frac{\partial\sigma_n^2}{\partial\beta^t}\right] = E^{\theta_0}\left[\frac{\left(\sum_{k=1}^{\infty}\beta^{k-1}\sigma_{-k}^2\right)^2}{2\sigma_0^2}\right], \tag{50}$$

这与 [57] 中给出的闭式表达式 (17) 一致。

类似的方法也适用于 KL散度。令 $\theta_i = (\delta_i,\ \alpha_i,\ \beta_i)^t$ 对于 $i=0,\ 1$，并记 $\sigma_{n,\,i}^2$ 为在 θ_i 下计算的 σ_n^2。由于 (46) 已将对数似然写成一个加性泛函，定理2 实质上意味着

$$K(\theta_1,\theta_0) = -\frac{1}{2}E^{\theta_0}_{\omega_{\theta_1,0}}\left[\log\sigma_{1,1}^2 + \frac{Y_1^2}{\sigma_{1,1}^2}\right] + \frac{1}{2}E^{\theta_0}_{\omega_{\theta_0,0}}\left[\log\sigma_{1,0}^2 + \frac{Y_1^2}{\sigma_{1,0}^2}\right]. \tag{51}$$

为了进一步将 (51) 与双无限平稳序列联系起来，注意对于 $i=0,\ 1$，直接计算可得

$$\begin{aligned}
\sigma_{n,i}^2 &= \delta_i + \alpha_i Y_{n-1}^2 + \beta_i\sigma_{n-1,i}^2 \\
&= \delta_i + \alpha_i Y_{n-1}^2 + \beta_i(\delta_i + \alpha_i Y_{n-2}^2 + \beta_i\sigma_{n-2,i}^2) \\
&= \cdots \\
&= \delta_i\sum_{k=0}^{\infty}\beta_i^k + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{n-1-k}^2 \\
&= \frac{\delta_i}{1-\beta_i} + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{n-1-k}^2,
\end{aligned}$$

因此，根据平稳性，

$$\begin{aligned}
&E^{\theta_0}_{\omega_{\theta_1,i}}\left[\log\sigma_{1,i}^2 + \frac{Y_1^2}{\sigma_{1,i}^2}\right] \\
=&\lim_{n\to\infty}E^{\theta_0}\left[\log\left(\frac{\delta_i}{1-\beta_i} + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{n-1-k}^2\right) + \frac{Y_n^2}{\frac{\delta_i}{1-\beta_i} + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{n-1-k}^2}\right] \\
=&E^{\theta_0}\left[\log\left(\frac{\delta_i}{1-\beta_i} + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{-(k+1)}^2\right) + \frac{Y_0^2}{\frac{\delta_i}{1-\beta_i} + \alpha_i\sum_{k=0}^{\infty}\beta_i^k Y_{-(k+1)}^2}\right].
\end{aligned}$$

注8。 如（50）所示，定理1中的不变测度 $\omega_{\theta_0,\,2}$ 实际上包含了整个过去历史的数据。这可以这样理解：不变概率测度 $\omega_{\theta_0,\,2}$ 表示 $\ell(\theta; Y_{0:\infty})$ 的导数的极限行为；当过程平稳时，这等价于 $\ell(\theta;\ Y_{-\infty:0})$ 的导数的极限行为。定理2中的 $\omega_{\theta,\,0}$ 也有类似情况。

为了分析线性循环神经网络（44）和（45），我们按照 [42] 中的相同方法进行如下操作：令 $W_n = (X_{n-1},\ X_n,\ Y_n)^t$ 为定义在 $\mathcal{X}:=(\mathbf{R}\times\mathbf{R}\times\mathbf{R})$ 上的马尔可夫链。记 $\eta_n = X_{n-1}^{-1}Y_n$，并令 $\tau_n = (\alpha+\beta\eta_n) \in \mathbf{R}$。令 A_n 为一个 3×3 矩阵，写作

$$A_n = \begin{pmatrix} 0 & 1 & 0 \\ 0 & \tau_n & 0 \\ 0 & \eta_n & 0 \end{pmatrix}. \tag{52}$$

注意，$\{A_n, n \geq 0\}$ 是由马尔可夫链 $\{W_n, n \geq 0\}$驱动的随机矩阵。

令 $Z_n = (0, \delta, 0)^t \in \mathbf{R}^3$。于是，线性循环神经网络(44)和(45)具有如下状态空间表示：W_n 是由以下机制控制的马尔可夫链

$$W_n = A_n W_{n-1} + Z_n, \tag{53}$$

以及 Y_n：观测到的随机量$= g(X_n)$是 X_n的不可逆函数。

根据 [42], 中的定理1，稳定性的一个充分条件是 $\alpha + \lambda\beta < 1$，其中 $\lambda = E_{\Pi}{}^{\mu_y}{}_{,h_1}{}^{1}$,—且 Π 是马尔可夫链 $\{(X_n,\ A_n \cdots A_1)\ n \geq 0\}$的平稳分布。条件C1（$\omega$-一致性）可通过状态空间表示在 [54]的第16.5.1节中找到。定理1和定理2中的条件C2–C6在(44)的正态性假设下成立。采用与(43)类似的方法，可得到费舍尔信息矩阵的表示。

一般而言，循环神经网络可以将变长序列 $y = (y_1,\ \cdots,\ y_n)$ 作为输入，通过递归处理每个符号，同时维护其内部隐状态 h。在每个时间步 n，循环神经网络读取符号 $Y_n \in \mathbf{R}^q$，并通过以下方式更新其隐状态 $h_n \in \mathbf{R}^p$：

$$h_n = f_\theta(Y_n, h_{n-1}), \tag{54}$$

其中，f_θ 是一个确定性的非线性转移函数，而 θ 是 f_θ的参数。

对于给定的循环神经网络模型序列，通过将联合序列概率分布因式分解为条件概率的乘积来参数化，我们有

$$\begin{aligned} P(Y_1, \cdots, Y_n) &= \prod_{k=1}^{n} P(Y_k | Y_1, \cdots, Y_{k-1}), \\ P(Y_n | Y_1, \cdots, Y_{n-1}) &= g_\theta(h_{n-1}), \end{aligned} \tag{55}$$

其中，g_θ 是一个将循环神经网络隐状态 h_{t-1} 映射为可能输出上的概率分布的函数，而 θ是 g_θ的参数。

如 [59], 中所述，给定一组 N 训练序列 $\{y_1^{(n)}, \cdots, y_{T_n}^{(n)}\}$，可以通过最小化以下代价函数来估计循环神经网络中的参数：

$$J(\theta) = \frac{1}{N} \sum_{n=1}^{N} \sum_{t=1}^{T_n} d(y_t^{(n)}, g_\theta(h_{t-1}^{(n)})), \tag{56}$$

其中，$d(a,\ b)$ 是 a 与 b之间预先定义的散度度量，例如欧几里得距离或 KL散度（或交叉熵）。定理2表明，就扩展马尔可夫链的平稳分布而言，KL散度是定义良好的。这为在循环神经网络中将 KL散度用作代价函数提供了理论基础。对于正则化问题，一种可行的方法是文献 (32) 和 (33) 中著名的赤池信息准则模型选择方法；文中还从理论上论证了在循环神经网络中使用该方法的合理性。

示例 3. 时序受限玻尔兹曼机。

玻尔兹曼机是一种由随机二值单元构成的网络，其中包含一组可见单元 $y \in \{0,1\}^D$和一组隐藏单元 $h \in \{0,1\}^P$。$\{y, h\}$ 的能量定义为

$$E(y, h; \theta) = -\frac{1}{2} y^t L y - \frac{1}{2} h^t J h - \frac{1}{2} y^t W h, \tag{57}$$

其中，$\theta = \{W, L, J\}$ 为参数：W，L，J分别表示可见层与隐藏层、可见层与可见层以及隐藏层与隐藏层之间的对称交互项。L 和 J 的对角元素设为 0。模型赋予可见向量 y 的概率为：

$$p(y; \theta) = \frac{p^*(y; \theta)}{Z(\theta)} = \frac{1}{Z(\theta)} \sum_h \exp(-E(y, h; \theta)), \tag{58}$$

$$Z(\theta) = \sum_{y} \sum_{h} \exp(-E(y, h; \theta)), \tag{59}$$

其中， p^* 表示未归一化概率，而 $Z(\theta)$是配分函数（归一化常数）。

令 $J = 0$ 和 $L = 0$ 均成立，即可得到著名的受限玻尔兹曼机（RBM）模型，参见 [60]。RBM 通过以下方程定义向量对 $Y \in \{0,1\}^D$ 和 $H \in \{0,1\}^P$ 上的概率分布

$$p(y, h; \theta) = P(Y = y, H = h; \theta) = \frac{1}{Z(\theta)} \exp(y^t b_Y + h^t b_H + y^t W h), \tag{60}$$

$$Z(\theta) = \sum_{y} \sum_{h} \exp(y^t b_Y + h^t b_H + y^t W h), \tag{61}$$

其中， b_Y 是可见向量的偏置向量， b_H 是隐藏向量的偏置向量， W 是连接权重矩阵，而 $Z(\theta)$ 是配分函数（归一化常数）。

接下来，我们考虑时序受限玻尔兹曼机（TRBM），参见 [61]。在其最简单的形式下，TRBM 可视为一种隐马尔可夫模型（HMM），其状态空间呈指数级增长，但转移概率和发射概率的参数化极为紧凑。记 $a_{1:n} = (a_1, \cdots, a_n)$。TRBM 通过以下方程定义概率分布 $P(Y_{1:n} = y_{1:n}, H_{0:n} = h_{0:n}) = P((Y_1, \cdots, Y_n) = (y_1, \cdots, y_n), (H_0, \cdots, H_n) = (h_0, \cdots, h_n))$

$$P(y_{1:n}, h_{0:n}) = \prod_{t=1}^{n} P(y_t, h_t | h_{t-1}) \bar{\nu}(h_0), \tag{62}$$

其形式与式（1）中定义的概率相同。这里， $\bar{\nu}$ 可以是 H_0的任意合适的初始分布。条件分布 $P(Y_t, H_t|h_{t-1})$是一个受限玻尔兹曼机的条件分布，其针对 H_t 的偏置是 h_{t-1}的函数。也就是说

$$P(y_t, h_t | h_{t-1}) = \exp(y_t^t b_Y + v_t^t W h_t + h_t^t (b_H + W' h_{t-1}))/Z(h_{t-1}), \tag{63}$$

其中， b_Y， b_H和 W与方程（60）中的相同，而 W'是从 H_{t-1}到 H_t的连接权重矩阵，因此$b_H + W' h_{t-1}$是时刻 t受限玻尔兹曼机的偏置。

现在，我们需要检查定理1和定理2中的条件C1–C6是否在式（62）和（63）的模型假设下成立。首先注意到， h_t 的状态空间是 $\mathcal{X} = \{0,1\}^P$，它是有限的（尽管规模呈指数级增长）；这意味着底层马尔可夫链具有一致遍历性，因此条件C1成立。至于其他条件，注意到由于状态空间 $\mathcal{X}$是有限的，条件C2-C6中对 $\mathcal{X}$取上确界或积分的结果都是有限的。此外， y_t 的状态空间是 $\{0,1\}^D$，它是有限的；这意味着 Y的矩生成函数存在。另外，由于逻辑函数的对数在 θ的任意局部邻域内均无限次可微，条件C2-C6中对 $N_\delta(\theta_0)$取上确界的结果是有限的。因此，条件C2-C6成立。

如 [62], 所述，变分学习有一个优点：除了尝试最大化训练数据的对数似然外，它还尝试寻找能够使近似后验分布与真实后验分布之间的 KL散度最小化的参数。定理2表明，根据扩展马尔可夫链的平稳分布，KL散度是定义良好的。这为在 TRBM 的随机梯度下降中使用 KL散度作为代价函数提供了理论依据。对于正则化问题，一种可能的方法是文献（32）和（33）中广为人知的 AIC模型选择方法，我们在其中给出了在 TRBM 中使用该方法的理论依据。

然而，计算 TRBM 和循环神经网络中的 KL散度及费舍尔信息并非易事。这是因为与独立同分布的情形不同，TRBM 中公式（2）和（3）里的极限没有显式形式。传统上，人们通过模拟很长的 TRBM 序列来逼近这些极限，并进行数值计算。得益于定理1和定理2，我们可以改用蒙特卡洛方法和其他计算技术来评估公式（18）和（20）中的期望。值得一提的是，计算这些期望仍非易事；定理1和定理2提供了一种可能的工具，可借助各种统计计算技术进行数值计算。

5 结论

本文给出了广义隐马尔可夫模型中 KL散度和费舍尔信息的明确刻画，并推导了这两个重要量之间的关系。研究结果基于这样一种表示：将对数似然及其导数表示为 MIFS 的加性泛函，从而可以利用强大数定律及其他相关结果研究对数似然的行为。利用这些结果，我们还给出了广义隐马尔可夫模型下的 Cram´er-Rao 下界以及 H´ajek-Le Cam 局部渐近极小极大定理。进一步的刻画还表明，

隐马尔可夫模型中的 KL散度一般并非凸函数，这与传统的独立同分布或马尔可夫链情形不同。此外，我们为在具有有限状态空间的广义隐马尔可夫模型中使用赤池信息准则模型选择提供了理论依据。
预计这种表示方法将有助于进一步研究广义隐马尔可夫模型，例如随机波动率模型中的模型选择和广义矩估计法、广义隐马尔可夫模型中的指数倾斜估计量和拟最大似然估计量，以及以 KL散度（相对熵）为惩罚项的循环神经网络正则化等相关课题。

参考文献

[1] C. Aghamohammadi、S. P. Loomis、J. R. Mahoney 和 J. P. Crutchfield，“罕见事件采样的极端量子记忆优势”，《物理评论X》，第8卷，第1期，页码011025，2018年。[2] D. Yu、K. Yao、H. Su、G. Li 和 F. Seide，“通过KL散度正则化深度神经网络自适应提升大词汇量语音识别性能”，载于2013年IEEE声学、语音与信号处理国际会议。IEEE，2013年，第7893–7897页。[3] K. M. Hangos，“化学反应动力学中的工程模型降阶与基于熵的李雅普诺夫函数”，《熵》，第12卷，第4期，第772–797页，2010年。[4] C. Beck，“物理学中的广义信息与熵度量”，《当代物理学》，第50卷，第4期，第495–510页，2009年。[5] H. White，“错设模型的最大似然估计”，《计量经济学》，第50卷，第1–26页，1982年。[6] C. Gourieroux、A. Monfort 和 A. Trognon，“伪最大似然方法：理论”，《计量经济学》，第52卷，第681–700页，1984年。[7] Y. Kitamura 和 M. Stutzer，“广义矩估计的信息论替代方法”，《计量经济学》，第65卷，第4期，第861–874页，1997年。[8] G. W. Imbens、R. H. Spady 和 P. Johnson，“矩条件模型推断的信息论方法”，《计量经济学》，第66卷，第333–357页，1998年。[9] B. H. Juang 和 L. R. Rabiner，“语音识别中的隐马尔可夫模型”，《技术计量学》，第33卷，第3期，第251–272页，1991年。[10] J. C. Marioni、N. P. Thorne 和 S. Tavaré，“BioHMM：用于分割阵列比较基因组杂交数据的异质隐马尔可夫模型”，《生物信息学》，第22卷，第9期，第1144–1146页，2006年。[11] J. D. Hamilton，“非平稳时间序列与商业周期经济分析的新方法”，《计量经济学》，第57卷，第2期，第357–384页，1989年。[12] L. E. Calvet 和 A. J. Fisher，“多重分形波动率预测”，《计量经济学杂志》，第105卷，第27–58页，2001年。[13] J. Cai，“切换机制ARCH的马尔可夫模型”，《商业与经济统计杂志》，第12卷，第309–316页，1994年。[14] J. D. Hamilton 和 R. Susmel，“自回归条件异方差与机制变化”，《计量经济学杂志》，第64卷，第307–333页，1994年。[15] Z. Ghahramani 和 M. I. Jordan，“因子隐马尔可夫模型”，《机器学习》，第29卷，第2期，第245–273页，1997年。[16] C. J. Kim，“带有马尔可夫机制转换的动态线性模型”，《计量经济学杂志》，第60卷，第1–22页，1994年。[17] C. J. Kim 和 C. R. Nelson，“商业周期转折点、新的同步指标，以及基于带机制转换动态因子模型的持续期依赖检验”，《经济学与统计学评论》，第80卷，第188–201页，1998年。[18] Z. Ghahramani 和 G. E. Hinton，“切换状态空间模型的变分学习”，《神经计算》，第12卷，第4期，第831–864页，2000年。[19] I. Goodfellow、J. Pouget-Abadie、M. Mirza、B. Xu、D. Warde-Farley、S. Ozair、A. Courville 和 Y. Bengio，“生成对抗网络”，载于《神经信息处理系统进展》，2014年，第2672–2680页。[20] R. F. Engle，“利用英国通货膨胀方差估计的自回归条件异方差”，《计量经济学》，第50卷，第4期，第987–1007页，1982年。[21] T. Bollerslev，“广义自回归条件异方差”，《计量经济学杂志》，第31卷，第3期，第307–327页，1986年。[22] J. Fan 和 Q. Yao，《非线性时间序列》。施普林格统计学丛书。施普林格纽约，2003年。

[23] P. Hall 和 Q. Yao，“具有重尾误差的ARCH和GARCH模型中的推断”，《计量经济学》，第71卷，第1期，第285–317页，2003年。[24] C. Francq 和 J.-M. Zako¨ıan，“严格平稳性检验以及对爆炸性和平稳广义自回归条件异方差模型的估计”，《计量经济学》，第80卷，第2期，第821–861页，2012年。[25] P. K. Clark，“一种适用于投机价格且具有有限方差的从属随机过程模型”，《计量经济学》，第41卷，第1期，第135–155页，1973年。[26] S. Taylor，《金融时间序列建模》。John Wiley & Sons，英国，1986年。[27] O. E. Barndorff-Nielsen 和 N. Shephard，“随机波动率与跳跃下的幂变差和双幂变差”，《金融计量经济学杂志》，第2卷，第1期，第1–37页，2004年。[28] A. Smith、P. A. Naik 和 C.-L. Tsai，“使用库尔贝克–莱布勒散度进行马尔可夫转换模型选择”，《计量经济学杂志》，第134卷，第2期，第553–577页，2006年。[29] C. D. Fuh，“隐马尔可夫模型中的序贯概率比检验和累积和”，《统计年鉴》，第31卷，第942–977页，2003年。[30] E. Andreoua 和 E. Ghysels，“结构性信用风险模型的质量控制”，《计量经济学杂志》，第146卷，第364–375页，2008年。[31] C. D. Fuh，“状态空间模型中复合假设的渐近最优变点检测”，《IEEE信息论汇刊》，第67卷，第485–505页，2021年。[32] C. D. Fuh 和 Y. J. Mei，“双状态隐马尔可夫模型的最快变点检测与库尔贝克–莱布勒散度”，《IEEE信号处理汇刊》，第63卷，第18期，第4866–4878页，2015年。[33] A. N. Gorban、P. A. Gorban 和 G. Judge，“熵：马尔可夫排序方法”，《熵》，第12卷，第5期，第1145–1193页，2010年。[34] N. F. Travers，“满足路径可合并条件的隐马尔可夫模型中熵率近似收敛的指数界”，《随机过程及其应用》，第124卷，第12期，第4149–4170页，2014年。[35] M. Obremski 和 M. Skorski，“估计马尔可夫链R´enyi熵的复杂度”，载于2020年IEEE国际信息论研讨会（ISIT）。IEEE，2020年，第2264–2269页。[36] B. G. Leroux，“混合分布的相合估计”，《统计年鉴》，第20卷，第3期，第1350–1360页，1992年。[37] P. J. Bickel、Y. Ritov 和 T. Ryden，“一般隐马尔可夫模型最大似然估计量的渐近正态性”，《统计年鉴》，第26卷，第4期，第1614–1635页，1998年。[38] R. Douc、E. Moulines、J. Olsson 和 R. V. Handel，“一般隐马尔可夫模型最大似然估计量的相合性”，《统计年鉴》，第39卷，第1期，第474–513页，2011年。[39] C. D. Fuh，“状态空间模型中的有效似然估计”，《统计年鉴》，第34卷，第2026–2068页。更正刊于*38*，第1279–1285页，（2010），2006年。[40] ——，“隐马尔可夫模型中最大似然估计量的Bahadur效率”，《华人统计学》，第14卷，第127–154页，2004年。[41] A. Van der Vaart，“Lucien Le Cam的统计学研究”，《统计年鉴》，第30卷，第3期，第631–682页，2002年。[42] C. D. Fuh，“马尔可夫迭代函数系统的渐近行为”，《随机过程及其应用》，第138卷，第186–211页，2021年。[43] C. D. Fuh 和 T. Pang，“一般马尔可夫切换模型最大似然估计量的渐近行为”，《华人统计学》，2022年（即将出版，doi:10.5705/ss.202021.0336）。[44] P. J. Bickel 和 Y. Ritov，“隐马尔可夫模型中的推断 I：平稳情形下的局部渐近正态性”，《伯努利》，第2卷，第3期，第199–228页，1996年。[45] A. Van der Vaart，《渐近统计学》。剑桥大学出版社，2000年，第3卷。[46] Z. Rached、F. Alajaji 和 L. L. Campbell，“马尔可夫源之间的库尔贝克–莱布勒散度率”，《IEEE信息论汇刊》，第50卷，第5期，第917–921页，2004年。[47] M. Vidyasagar，“不同基数集合上概率分布之间的库尔贝克–莱布勒散度率”，载于第49届IEEE决策与控制会议（CDC）。IEEE，2010年，第948–953页。[48] H. Akaike，“信息论与最大似然原理的扩展”，载于《第二届国际信息论研讨会论文集》，1973年，第267–281页。

[49] S. Yonekuraa、A. Beskosa 和 S. S. Singhb，“一般隐马尔可夫模型模型选择准则的渐近分析”，《随机过程及其应用》，第 132 卷，第 164–191 页，2021 年。[50] A. Klein、G. M´elard 和 A. Saidi，“向量 ARMA 过程的渐近与精确费舍尔信息矩阵”，《统计与概率快报》，第 78 卷，第 12 期，第 1430–1433 页，2008 年。[51] A. C. Harvey 和 G. D. Phillips，“具有自回归滑动平均扰动的回归模型的最大似然估计”，《生物统计学》，第 66 卷，第 1 期，第 49–58 页，1979 年。[52] J. Pearlman，“高阶自回归移动平均过程精确似然的一种算法”，《生物统计学》，第 67 卷，第 1 期，第 232–233 页，1980 年。[53] E. J. Hannan 和 M. Deistler，《线性系统统计理论》。SIAM，2012 年。[54] S. P. Meyn 和 R. L. Tweedie，《马尔可夫链与随机稳定性》。施普林格科学与商业媒体，2012 年。[55] A. Klein 和 H. Neudecker，“高斯向量状态空间模型精确费舍尔信息矩阵的直接推导”，《线性代数及其应用》，第 321 卷，第 1–3 期，第 233–238 页，2000 年。[56] T. Bengtsson 和 J. E. Cavanaugh，“ 一种改进的状态空间模型选择赤池信息准则”，《计算统计与数据分析》，第 50 卷，第 10 期，第 2635–2654 页，2006 年。[57] J. Ma，“ GARCH (1, 1) 模型准极大似然估计量的闭式渐近方差协方差矩阵”，可从 SSRN 889461 获取，2008 年。[58] G. Fiorentini、G. Calzolari 和 L. Panattoni，“解析导数与 GARCH 估计值的计算”，《应用计量经济学杂志》，第 11 卷，第 4 期，第 399–417 页，1996 年。[59] R. Pascanu1、C. Gulcehre1、K. Cho 和 Y. Bengio，“如何构建深度循环神经网络”，ICLR，2014 年。[60] G. Hinton 和 R. Salakhutdinov，“用神经网络降低数据维度”，《科学》，第 313 卷，第 5786 期，第 504–507 页，2006 年。[61] H. Sutskever、G. E. Hinton 和T. W.Graham，“循环时序受限玻尔兹曼机”，载于 NIPS，2008 年。[62]R. Salakhutdinov 和 G. E. Hinton，“深度玻尔兹曼机”，第十二届国际人工智能与统计学会议（AISTATS）论文集，2009 年。[63] T. S. Ferguson，《大样本统计学课程》。劳特利奇，博卡拉顿，2017 年。[64] A. B. Tsybakov，《非参数估计导论》。施普林格科学与商业媒体，2008 年。[65] J. L. Jensen，“关于《状态空间模型中的有效似然估计》一文中的若干问题”，《统计年鉴》，第 38 卷，第 2 期，第 1279–1281 页，2010 年。

A 附录：主要结果的证明

由于引理1和引理2分别与 [43] 中关于双层隐马尔可夫模型的引理3和引理5几乎相同，为了完整起见，我们在补充材料中给出这两个引理的证明。利用这两个引理，我们首先分别在A.1节和A.2节证明定理1和定理2。然后，我们基于定理1和定理2，在A.3节证明定理3和推论1。

A.1 定理1的证明

为证明定理1，我们只需证明以下引理：

引理3. 假设条件*C1–C6*在 $r = 2$ 下成立。对于任意 $1 \le j, k \le q$，有

$$I_{jk}(\theta_0) = -E^{\theta_0}_{\omega_{\theta_0,2}}[g^{\nu(j,k)}(W^{(2)}_{1,\theta_0}, W^{(2)}_{0,\theta_0})],$$

其中，$E^{\theta}_{\omega_{\theta_0,2}}$ 和 $g^{\nu(j,k)}$ 与定理*1*中的相同。

因此，定理1可直接由引理3得出。

证明。 对任意 $1 \le j,\ k \le q$，由 (3) 可得

$$I_{jk}(\theta_0) = -\lim_{n\to\infty} \frac{1}{n} D^{\nu(j,k)} \ell(\theta_0; Y_{0:n}) \quad P^{\theta_0}\text{-a.s.} \tag{64}$$

由 (17)、引理1以及 [54], 中的马尔可夫随机游走强大数定律可得

$$\frac{1}{n}D^{\nu(j,k)}\log L(\theta_0;Y_{0:n}) = \frac{1}{n}\left\{\sum_{t=1}^{n} g^{\nu(j,k)}(W^{(2)}_{t,\theta_0}, W^{(2)}_{t-1,\theta_0}) + g_0^{\nu(j,k)}(W^{(2)}_{0,\theta_0})\right\}$$
$$\to E^{\theta_0}_{\omega_{\theta_0,2}}[g^{\nu(j,k)}(W^{(2)}_{1,\theta_0}, W^{(2)}_{0,\theta_0})] \quad P^{\theta_0}\text{-a.s.} \tag{65}$$

因此，引理3直接由式（64）和（65）得出。 □

注9。 还可以将函数 G 与 log $L(\theta_0; Y_{0:1})$ 的二阶导数联系起来。定义

$$G_0(w) = \begin{pmatrix} g_0^{\nu(1,1)}(w) & \cdots & g_0^{\nu(1,q)}(w) \\ \vdots & \ddots & \vdots \\ g_0^{\nu(q,1)}(w) & \cdots & g_0^{\nu(q,q)}(w) \end{pmatrix}. \tag{66}$$

那么，根据式（17），我们有

$$G(W^{(2)}_{1,\theta_0}, W^{(2)}_{0,\theta_0}) = D^2_\theta \log L(\theta_0; Y_{0:1}) - G_0(W^{(2)}_{0,\theta_0}),$$

因此我们有

$$I(\theta_0) = -E^{\theta_0}_{\omega_{\theta_0,2}}\left[D^2_\theta \log L(\theta_0; Y_{0:1}) - G_0(W^{(2)}_{0,\theta_0})\right].$$

A.2 定理2的证明

为证明定理2，我们将 $W^{(r)}_{n\,\theta,}$ 的定义扩展到 $r \geq 0$，其中 $W_{(0)\,n,\theta} = W^0_{n,\theta_0}$。注意，引理1对 $r = 0$ 也成立；见 [39]。因此，我们可以将 $\omega_{\theta,0}$ 定义为诱导马尔可夫链 $\{((X_n, Y_n)\ W^{(0)}_{n\,\theta},)\ n \geq 0\}$ 的平稳分布。

定理2的证明。 为简单起见，我们证明 $q = 1$ 和 $\theta_1 > \theta_0$ 的情形；一般情形的证明类似。

首先，容易验证 (17) 对 $r = 0$ 也成立，因此我们有

$$\frac{1}{n}\left[\log L(\theta_1; Y_{0:n}) - \log L(\theta_0; Y_{0:n})\right]$$
$$= \frac{1}{n}\left\{\sum_{t=1}^{n} g^0(W^{(0)}_{t,\theta_1}, W^{(0)}_{t-1,\theta_1}) + g_0^0(W^{(0)}_{0,\theta_1})\right\} - \frac{1}{n}\left\{\sum_{t=1}^{n} g^0(W^{(0)}_{t,\theta_0}, W^{(0)}_{t-1,\theta_0}) + g_0^0(W^{(0)}_{0,\theta_0})\right\}. \tag{67}$$

对 (67) 两边取 $n \to \infty$，再由引理1、(17) 以及 [54], 中的马尔可夫随机游走强大数定律，我们有

$$K(\theta_1, \theta_0) = E^{\theta_1}_{\omega_{\theta_1,0}}\left[g^0(W^{(0)}_{1,\theta_1}, W^{(0)}_{0,\theta_1})\right] - E^{\theta_1}_{\omega_{\theta_0,0}}\left[g^0(W^{(0)}_{1,\theta_0}, W^{(0)}_{0,\theta_0})\right],$$

由此完成了对 (20) 的证明。

对于 (21)，通过泰勒展开可得

$$\ell(\theta_0; Y_{0:n}) - \ell(\theta_1; Y_{0:n})$$
$$= D^1\ell(\theta_1; Y_{0:n})(\theta_0 - \theta_1) + \frac{1}{2}D^2\ell(\theta_1; Y_{0:n})(\theta_0 - \theta_1)^2 + \frac{1}{6}\frac{D^3\ell(\tilde{\theta}_n; Y_{0:n})}{n}(\theta_1 - \theta_0)^3, \tag{68}$$

˜其中 $\theta_n \in (\theta_0,\ \theta_1)$。将 (68) 两边除以 n，可得

$$\frac{1}{n}\left[\ell(\theta_1; Y_{0:n}) - \ell(\theta_0; Y_{0:n})\right]$$
$$= \frac{D^1\ell(\theta_1; Y_{0:n})}{n}(\theta_1 - \theta_0) - \frac{1}{2}\frac{D^2\ell(\theta_1; Y_{0:n})}{n}(\theta_1 - \theta_0)^2 + \frac{1}{6}\frac{D^3\ell(\tilde{\theta}_n; Y_{0:n})}{n}(\theta_1 - \theta_0)^3. \tag{69}$$

For 根据与定理1的证明类似的论证，(69) 右侧的第一项有

$$\lim_{n\to\infty}\frac{D^1\ell(\theta_1; Y_{0:n})}{n} = E^{\theta_1}_{\omega_{\theta_1,1}}[g^1(W^{(1)}_{1,\theta_1}, W^{(1)}_{0,\theta_1})] \quad P^{\theta_1}\text{-a.s.} \tag{70}$$

与此同时，我们还有 $\frac{1}{n}E_x^{\theta_1}[D^1\ell(\theta_1; Y_{0:n})] = 0$，这是因为 $L(\theta_1; \cdot)$ 是在 θ_1 下的似然。此外，

$$\frac{1}{n}E_x^{\theta_1}[D^1\ell(\theta_1; Y_{0:n})] \to E^{\theta_1}_{\omega_{\theta_1,1}}[g^1(W^{(1)}_{1,\theta_1}, W^{(1)}_{0,\theta_1})]. \tag{71}$$

结合 (70) 和 (71)，我们证明第一项几乎必然趋于零 P^{θ_1}。

对于 (69) 右侧的第二项，由 (3) 可得

$$\lim_{n\to\infty}\frac{D^2\ell(\theta_1; Y_{0:n})}{n} = -I(\theta_1) \quad P^{\theta_1}\text{-a.s.} \tag{72}$$

对于 (69) 右侧的第三项，通过类似于定理1证明的论证，并结合经典一致大数定律（见 [63], 第16章），存在一个常数 $C_3 > 0$，使得

$$\begin{aligned}
&\limsup_{n\to\infty}\left|\frac{1}{6}\frac{D^3\ell(\tilde{\theta}_n; Y_{0:n})}{n}(\theta_1-\theta_0)^3\right| \\
\le&\limsup_{n\to\infty}\sup_{\theta\in[\theta_0,\theta_1]}\left|\frac{1}{6}\frac{D^3\ell(\theta; Y_{0:n})}{n}(\theta_1-\theta_0)^3\right| \\
\le&C_3|\theta_1-\theta_0|^3 \quad P^{\theta_1}\text{-a.s.}
\end{aligned} \tag{73}$$

因此，取 $n \to \infty$ 代入 (69)，并应用 (70)–(73)，结合 (2) 中 $K(\theta_1, \theta_0)$ 的定义，我们有

$$K(\theta_1,\theta_0) = \frac{I(\theta_1)}{2}(\theta_1-\theta_0)^2 + O\left(|\theta_1-\theta_0|^3\right) = \frac{I(\theta_1)}{2}(\theta_1-\theta_0)^2 + o\left((\theta_1-\theta_0)^2\right),$$

进一步注意到当 $\theta_1 \to \theta_0$时，$I(\theta_1) \to I(\theta_0)$，即可得到 (21)。 □

注释 10. 由 (17) 可得，$g^0(W_{(0)\,1,\theta}, W^{(0)}_{0\,\theta},) = \log L(\theta; Y_{0:1}) - g^0_0(W_{(0)\,0,\theta})$，因此有

$$K(\theta_1,\theta_0) = E^{\theta_1}_{\omega_{\theta_1,0}}\left[\log L(\theta_1; Y_{0:1}) - g^0_0(W^{(0)}_{0,\theta_1})\right] - E^{\theta_1}_{\omega_{\theta_0,0}}\left[\log L(\theta_0; Y_{0:1}) - g^0_0(W^{(0)}_{0,\theta_0})\right],$$

这进一步将 KL散度与对数似然联系起来。

A.3 定理3和推论1的证明

定理*3*的证明。 我们先从 (22) 入手。对于任意固定的 n，经典的多元 Cramér–Rao 下界给出

$$E_x^{\theta_0}\left[\left(v^t(\hat{\theta}_n(Y_{0:n}) - \theta_0)\right)^2\right] \ge v^t I_n^{-1}(\theta_0)v, \tag{74}$$

其中，$I_n(\theta_0) = -E_x^{\theta_0}\left[D^2_\theta \log L(\theta_0; Y_{0:n})\right]$ 是基于 $Y_{0:n}$ 的费舍尔信息。采用与定理1的证明相同的论证，我们有

$$\frac{1}{n}I_n(\theta_0) \to I(\theta_0). \tag{75}$$

公式（22）可由（74）和（75）立即推出。

现在转而证明（23）。记 P_n^θ 为 $Y_{0:n}$ 在 P^θ 下的概率分布。于是，根据平方误差损失下的勒卡姆方法（[64], 第 2 章），我们有

$$\inf_{\hat{\theta}_n}\max_{\theta\in\{\theta_0,\theta_0+\delta v\}} E_x^\theta\left[\|\hat{\theta}_n(Y_{0:n}) - \theta\|^2\right] \ge \frac{\delta^2\|v\|^2}{8}\left[1 - 2\|P_n^{\theta_0} - P_n^{\theta_0+\delta v}\|^2_{TV}\right], \tag{76}$$

˜其中，$\|\cdot\|_{TV}$表示全变差距离。此外，对于任意概率分布 P和 P，平斯克不等式（[64],引理 2.5(i)）指出，$2\|P-\tilde{P}\|^2_{TV} \le D_{KL}(P\|\tilde{P})$，其中 $D_{KL}(P\|\tilde{P}) = \int\log\left(\frac{P}{\tilde{P}}\right)dP$是˜ P与 P之间的 KL散度。据此，我们有

$$2\|P_n^{\theta_0} - P_n^{\theta_0+\delta v}\|^2_{TV} \le E_x^{\theta_0+\delta v}\left[\log\frac{L(\theta_0+\delta v; Y_{0:n})}{L(\theta_0; Y_{0:n})}\right]. \tag{77}$$

采用与定理2中式 (20) 证明相同的论证，我们有

$$\begin{aligned}&\frac{1}{n}E_x^{\theta_1}\left[\log\frac{L(\theta_1;Y_{0:n})}{L(\theta_0;Y_{0:n})}\right]=\frac{1}{n}E_x^{\theta_1}\left[\ell(\theta_1;Y_{0:n})\right]-\frac{1}{n}E_x^{\theta_1}\left[\ell(\theta_0;Y_{0:n})\right]\\ \to&E_{\omega_{\theta_1,0}}^{\theta_1}\left[g^0(W_{1,\theta_1}^{(0)},W_{0,\theta_1}^{(0)})\right]-E_{\omega_{\theta_0,0}}^{\theta_1}\left[g^0(W_{1,\theta_0}^{(0)},W_{0,\theta_0}^{(0)})\right]\quad P^{\theta_1}\text{-a.s.}\\ =&K(\theta_1,\theta_0)\end{aligned}\tag{78}$$

对于任意 $\theta_1\in N_\delta(\theta_0)$。此外，与经典一致大数定律类似（见 [63]），(78) 中的收敛在 $N_\delta(\theta_0)$ 的任意紧子空间上一致。

现在，回想一下 $\delta^2=(nv^tI(\theta_0)v)^{-1}$。结合 (21)、(77) 和 (78)，我们得到

$$\begin{aligned}&\lim_{n\to\infty}2\|P_n^{\theta_0}-P_n^{\theta_0+\delta v}\|_{TV}^2\le\lim_{n\to\infty}nK(\theta_0+\delta v,\theta_0)\\ =&\lim_{n\to\infty}n\left\{(\delta v)^t\frac{I(\theta_0)}{2}(\delta v)+o(\|\delta v\|^2)\right\}=\frac{1}{2}.\end{aligned}\tag{79}$$

结合 (76) 和 (79)，即可得到所需的 (23)。 □

推论*1*的证明。 关于非负性，注意到 (20) 是将马尔可夫随机游走强大数定律应用于 (2) 得到的，由此我们还有

$$K(\theta_1,\theta_0)=\lim_{n\to\infty}\frac{1}{n}\left\{E^{\theta_1}\left[\ell(\theta_1;Y_{0:n})\right]-E^{\theta_1}\left[\ell(\theta_0;Y_{0:n})\right]\right\}.$$

然而，根据 Gibb' 根据不等式，我们有 $E^{\theta_1}\left[\ell(\theta_1;Y_{0:n})\right]-E^{\theta_1}\left[\ell(\theta_0;Y_{0:n})\right]\ge 0$，因此非负性得证 s.

另一方面，可加性是式 (2) 以及以下事实的直接推论：对于 $i=1,2$ 和任意 n，$\log L(\theta_i;Y_{0:n})=\log L(\theta_i;Y_{0:n}^1)+\log L(\theta_i;Y_{0:n}^2)$。 □

注11. [38] 的第1.2节也采用另一种方法证明了KL散度的非负性。 [33]对可加性进行了部分研究。

一般隐马尔可夫模型中的补充库
尔贝克-莱布勒散度与赤池信息准则

Cheng-Der Fuh、Chu-Lan Michael Kao 和 Tianxiao Pang

在证明引理1之前，我们需要给出以下定义。由于我们将对 $M_n = \mathbf{P}_\theta(Y_n) \circ \cdots \circ \mathbf{P}_\theta(Y_0)$ 求导，因此需要考察微分算子 D_i 与算子 $\circ$ 的相互作用。回顾一下第2.2节第一段中定义的 $\mathbf{M}$ 和 $\mathbf{P}_\theta$。注意，对于任意两个给定的随机函数 $\mathbf{P}_\theta(Y_{t+1})$ 和 $\mathbf{P}_\theta(Y_t)$，以及任意 $h_\theta \in \mathbf{M}$，根据条件C1–C6和控制收敛定理，我们有

$$\begin{aligned}
&D_i \left\{ \mathbf{P}_\theta(Y_t) h_\theta(x) \right\} \\
=&D_i \left\{ \int_{s \in \mathcal{X}} p_\theta(s,x) f(Y_t;\theta|x,Y_{t-1}) h_\theta(s) Q(ds) \right\} \\
=&\int_{s \in \mathcal{X}} \Big\{ f(Y_t;\theta|x,Y_{t-1}) h_\theta(s) D_i p_\theta(s,x) + p_\theta(s,x) h_\theta(s) D_i f(Y_t;\theta|x,Y_{t-1}) \\
&+ p_\theta(s,x) f(Y_t;\theta|x,Y_{t-1}) D_i h_\theta(s) \Big\} Q(ds)
\end{aligned} \tag{80}$$

and

$$\begin{aligned}
&D_i \left\{ \mathbf{P}_\theta(Y_{t+1}) \circ \mathbf{P}_\theta(Y_t) h_\theta(x) \right\} \\
=&D_i \Big\{ \int_{z \in \mathcal{X}} p_\theta(z,x) f(Y_{t+1};\theta|x,Y_t) \times \\
&\left(\int_{s \in \mathcal{X}} p_\theta(s,z) f(Y_t;\theta|z,Y_{t-1}) h_\theta(s) Q(ds) \right) Q(dz) \Big\} \\
=&\int_{z \in \mathcal{X}} D_i \left\{ p_\theta(z,x) f(Y_{t+1};\theta|x,Y_t) \right\} \times \\
&\left(\int_{s \in \mathcal{X}} p_\theta(s,z) f(Y_t;\theta|z,Y_{t-1}) h_\theta(s) Q(ds) \right) Q(dz) \\
&+ \int_{z \in \mathcal{X}} p_\theta(z,x) f(Y_{t+1};\theta|x,Y_t) \times \\
&\left(\int_{s \in \mathcal{X}} D_i \left\{ p_\theta(s,z) f(Y_t;\theta|z,Y_{t-1}) h_\theta(s) \right\} Q(ds) \right) Q(dz) \\
=&\left\{ D_i \mathbf{P}_\theta(Y_{t+1}) \right\} \circ \mathbf{P}_\theta(Y_t) h_\theta(x) + \mathbf{P}_\theta(Y_{t+1}) \circ \left\{ D_i (\mathbf{P}_\theta(Y_t) h_\theta(x)) \right\}.
\end{aligned}$$

由此，我们有 $D^\nu \langle M_n \pi \rangle = \langle D^\nu(M_n \pi) \rangle$。此外，对于给定的 ν_i 和 ν_j，令 $\nu_i + \nu_j$ 表示这些向量的逐分量加法。于是，类似于式（80），我们有

$$\begin{aligned}
&D^\nu \left\{ \mathbf{P}_\theta(Y_t) h_\theta(x) \right\} \\
=&\sum_{\nu_p + \nu_f + \nu_h = \nu} \Big\{ \int_{s \in \mathcal{X}} D^{\nu_p} p_\theta(s,x) \times D^{\nu_f} f(Y_t;\theta|x,Y_{t-1}) D^{\nu_h} h_\theta(s) Q(ds) \Big\}.
\end{aligned}$$

因此，我们有

$$\begin{aligned}
&\left| D^\nu \left\{ \mathbf{P}_\theta(Y_t) h_\theta(x) \right\} - D^\nu \left\{ \mathbf{P}_\theta(Y_t) g_\theta(x) \right\} \right| \\
=&\Big| \sum_{\nu_p + \nu_f + \nu_h = \nu} \int_{s \in \mathcal{X}} D^{\nu_p} p_\theta(s,x) D^{\nu_f} f(Y_t;\theta|x,Y_{t-1}) D^{\nu_h} h_\theta(s) Q(ds) \\
&- D^{\nu_p} p_\theta(s,x) D^{\nu_f} f(Y_t;\theta|x,Y_{t-1}) D^{\nu_h} g_\theta(s) Q(ds) \Big|.
\end{aligned} \tag{81}$$

因此，若对所有 $|\nu| \leq r$， $D^\nu h_\theta(x) \in \mathbf{M}$，则通过与 [39] 中引理3的证明类似的论证（将其中的条件C1替换为我们的条件C6），我们有对所有 $|\nu| \leq r$， $D^\nu \{\mathbf{P}_\theta(Y_t) h_\theta(x)\} \in \mathbf{M}$。此外，根据C2和C3，我们有对所有 $|\nu| \leq r$， $D^\nu \pi_\theta(x) \in \mathbf{M}$。

现在我们可以证明引理1。

引理*1*的证明。 首先，根据上述论证，我们有 $W_n^{(r)} \in \mathbf{M}^K := \{v = (m_1, \cdots, m_K)^t : m_k \in \mathbf{M}\ 1 \le k \le K\}$。这意味着 $\{((X_n, Y_n)\ W_n^{(r)})\ n \ge 0\}$ 是 $(\mathcal{X} \times \mathbf{R}^d) \times \mathbf{M}^K$ 上的一个随机过程。

为证明 $\{((X_n, Y_n)\ W_n^{(r)})\ n \ge 0\}$ 是 MIFS，我们来考察 $W_{(r)n}$ 的动态特性。注意，对于任意 ν_i,

$$\begin{aligned} W_n^{\nu_i} =& D^{\nu_i}\big(\mathbf{P}_\theta(Y_n) \circ \cdots \circ \mathbf{P}_\theta(Y_1) \circ \mathbf{P}_\theta(Y_0)\big) \\ =& \sum_{\substack{1 \le j \le k \le K \\ \nu_i = \nu_j + \nu_k}} \left\{ \frac{(\nu_i)!}{(\nu_j)!(\nu_k)!} D^{\nu_k} \mathbf{P}_\theta(Y_n) \circ D^{\nu_j} \left(\mathbf{P}_\theta(Y_{n-1}) \circ \cdots \circ \mathbf{P}_\theta(Y_0) \right) \right\} \\ =& \sum_{\substack{1 \le j \le k \le K \\ \nu_i = \nu_j + \nu_k}} \frac{(\nu_i)!}{(\nu_j)!(\nu_k)!} \left\{ D^{\nu_k} \mathbf{P}_\theta(Y_n) \circ W_{n-1}^{\nu_j} \right\}. \end{aligned} \tag{82}$$

因此，w e 可以定义一个 $K \times K$ 矩阵形式 $A_n = \left\{a_n^{ij} : 1 \le i, j \le K\right\}$，其中每个 $a_n^{ij} \in \mathbf{M}$ 均定义为 s

$$a_n^{ij} = \begin{cases} \frac{(\nu_i)!}{(\nu_j)!(\nu_k)!} D^{\nu_k} \mathbf{P}_\theta(Y_n) & \text{if } \exists 1 \le k \le K \text{ such that } \nu_i = \nu_j + \nu_k, \\ 0 & \text{otherwise.} \end{cases} \tag{83}$$

此外，对于每个 $K \times K$ $\mathbf{M}$ 值矩阵形式 $B = \{b_{ij} : 1 \le i, j \le K\}$，以及每个 K 维的 $\mathbf{M}$ 值向量 $V = (V_1, V_2, \cdots, V_K) \in \mathbf{M}^K$，我们定义

$$B \circ V := \begin{pmatrix} \sum_{j=1}^K b_{1j} \circ V_j \\ \sum_{j=1}^K b_{2j} \circ V_j \\ \vdots \\ \sum_{j=1}^K b_{Kj} \circ V_j \end{pmatrix}. \tag{84}$$

于是，根据 (82)，我们有 $W_n^{(r)} = A_n \circ W_{n-1}^{(r)}$，从而

$$W_n^{(r)} = A_n \circ A_{n-1} \circ \cdots \circ A_1 \circ W_0^{(r)}, \tag{85}$$

其中 $W_{(r)0} = \{W_0^\nu : |\nu| \le r\}$，其中 $W_0^\nu = D^\nu \mathbf{P}_\theta(Y_0)$。

更重要的是，由于 $W_n^{(r)} = A_n \circ W_{n-1}^{(r)}$，且根据 (83)， $A_n = \left\{a_n^{ij} \text{ 的值} : 1 \le i,\ j \le K\right\}$ 仅由 Y_n 决定，因此我们知道 $W_{(r)n}$ 的值仅由 $(Y_n,\ W_{n-1}^{(r)})$ 决定。此外，由于 Y_n 的分布基于 X_n 和 Y_{n-1}，且 $\{X_n, n \ge 0\}$ 是一个马尔可夫链，$((X_n,\ Y_n)\ W_n^{(r)})$ 是马尔可夫的，正如所需。

最后，对于遍历性，通过类似于 [39]，中引理3的证明过程，可以证明，对于 $\theta \in N_\delta(\theta_0)$，MIFS $\{((X_n,\ Y_n)\ W_n^{(r)})\ n \ge 0\}$ 满足 [39] 中的假设K。此外， [39] 中的引理4对状态空间 $(\mathcal{X} \times \mathbf{R}^d) \times \mathbf{M}^K$ 上的诱导马尔可夫链 $\{((X_n,\ Y_n)\ W_n^{(r)})\ n \ge 0\}$ 成立，由此可直接推出引理1。证明完毕。

□

注12。 为说明（85），令 $q = 1$，即 θ 是一维的。在这种情况下， $\nu \in \mathbf{R}$，我们可以按自然顺序对所有$|\nu| \le r$进行简单标记，使得 $W_{(r)n} = (W_n^0,\ W_n^1,\ \cdots,\ W_n^r)^t$，即前 r 阶导数构成的向量。于是，对于任意$0 \le k \le r$，有

$$\begin{aligned} W_n^k =& D^k(\mathbf{P}_\theta(Y_n) \circ \cdots \circ \mathbf{P}_\theta(Y_1) \circ \mathbf{P}_\theta(Y_0)) \\ =& \sum_{0 \le k_1 \le k} \left\{ \frac{k!}{(k_1)!(k-k_1)!} D^{k_1} \mathbf{P}_\theta(Y_n) \circ D^{k-k_1} \left(\mathbf{P}_\theta(Y_{n-1}) \circ \cdots \circ \mathbf{P}_\theta(Y_0) \right) \right\} \\ =& \sum_{0 \le k_1 \le k} C_{k_1}^k \left\{ D^{k_1} \mathbf{P}_\theta(Y_n) \circ W_{n-1}^{k-k_1} \right\}, \end{aligned}$$

其中 $C_a^b = {}^{b!}\underline{\ a!\ \ (b-a)\ !}$。因此 $W_{(r)n} = A_n \circ W_{(r)n-1}$，其中

$$A_n = \begin{pmatrix} \mathbf{P}_\theta(Y_n) & 0 & \cdots & 0 \\ C_1^1 D^1 \mathbf{P}_\theta(Y_n) & \mathbf{P}_\theta(Y_n) & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ C_r^r D^r \mathbf{P}_\theta(Y_n) & C_{r-1}^r D^{r-1} \mathbf{P}_\theta(Y_n) & \cdots & \mathbf{P}_\theta(Y_n) \end{pmatrix}, \tag{86}$$

其中，0 表示$\mathbf{M}$中的零函数。

备注13。 请注意，(83) 中的 A_n 和 (85) 中的 $W_n^{(r)}$取值于$\mathbf{M}$，而非传统的分别取值于$\mathbf{R}$的矩阵和向量。为说明这一现象，我们考虑一个具有一维参数 θ情形的 D状态隐马尔可夫模型；此时，(86) 中的 A_n 是一个 $K\times K$ 矩阵形式，其中每个元素都是一个 $D\times D$ 矩阵（其中 0 是一个 $D\times D$零矩阵）。同样，尽管 (84) 中定义的算子看起来是传统的矩阵乘法，但二者并不相同，因为其中每个分量内的乘法都被 $\circ$取代。尽管如此，核心思想是为 $W_n^{(r)}$引入一种矩阵形式，从而可以通过 (85) 证明它构成一个遍历马尔可夫链。

备注14。这一构造的关键创新在于，为了获得马尔可夫结构，需要考虑阶数小于或等于 r 的所有导数。原因是，如（82）所示， $W_n^{\nu_i}$的迭代表示涉及所有 $W_{n-1}^{\nu_j}$和 $W_{n-1}^{\nu_k}$，其中 $\nu_i=\nu_j+\nu_k$。因此，需要考虑 $W^{(r)}{}_n$，而不是 W_n^{ν}。

请注意，这使得我们的方法与此前关于隐马尔可夫模型的研究（例如 [37], ）有很大不同；此前的研究在考察费舍尔信息矩阵时，只研究 $\ell(\theta;Y_{0:n})$ 的二阶导数。另一方面，我们在进行此类考察时，研究阶数小于或等于二的所有导数。

值得一提的是，（85）中能够得到简洁形式这一特点，基于（83）中对截至 r阶的所有偏导数的矩阵表示。这在很大程度上有助于我们得到引理1中的结果。

在证明一般情形下的引理2之前，我们先给出对数似然函数的一阶和二阶偏导数的具体形式如下。对于 $|\nu|=1$，注意到我们有

$$\frac{\langle D^\nu(M_t\pi)\rangle}{\langle M_t\pi\rangle}=\frac{\langle (D^\nu M_t)\pi+M_t(D^\nu\pi)\rangle}{\langle M_t\pi\rangle}=\frac{\langle (W_t^\nu)\pi+W_t^0(D^\nu\pi)\rangle}{\langle W_t^0\pi\rangle}$$

对于任意 $t\geq 0$。因此

$$\begin{aligned}
&D^\nu(\log L(\theta;Y_{0:n}))=D^\nu(\log\langle M_n\pi\rangle)=\frac{\langle D^\nu(M_n\pi)\rangle}{\langle M_n\pi\rangle}\\
=&\sum_{t=1}^n\left\{\frac{\langle D^\nu(M_t\pi)\rangle}{\langle M_t\pi\rangle}-\frac{\langle D^\nu(M_{t-1}\pi)\rangle}{\langle M_{t-1}\pi\rangle}\right\}+\frac{\langle D^\nu(M_0\pi)\rangle}{\langle M_0\pi\rangle}\\
=&\sum_{t=1}^n\left\{\frac{\langle (W_t^\nu)\pi+W_t^0(D^\nu\pi)\rangle}{\langle W_t^0\pi\rangle}-\frac{\langle (W_{t-1}^\nu)\pi+W_{t-1}^0(D^\nu\pi)\rangle}{\langle W_{t-1}^0\pi\rangle}\right\}+\frac{\langle (W_0^\nu)\pi+W_0^0(D^\nu\pi)\rangle}{\langle W_0^0\pi\rangle}\\
=:&\sum_{t=1}^n g^\nu(W_t^{(1)},W_{t-1}^{(1)})+g_0^\nu(W_0^{(1)}).
\end{aligned}\tag{87}$$

也就是说，对数似然函数的一阶导数可以重写为马尔可夫链 $\{((X_n,Y_n)\ W_n^{(1)})\ n\geq 0\}$ 的加性泛函。

为表示对数似然的二阶偏导数，对于 $|\nu|=2$，令 $\nu=\nu_1+\nu_2$，使得$|\nu_1|=|\nu_2|=1$。于是，我们有

$$\begin{aligned}
&D^\nu\log L(\theta;Y_{0:n})=D^{\nu_1}\left(D^{\nu_2}\log L(\theta;Y_{0:n})\right)\\
=&D^{\nu_1}\frac{\langle W_n^{\nu_2}\pi+W_n^0(D^{\nu_2}\pi)\rangle}{\langle W_n^0\pi\rangle}\\
=&\frac{\langle W_n^\nu\pi+W_n^{\nu_2}(D^{\nu_1}\pi)+W_n^{\nu_1}(D^{\nu_2}\pi)+W_n^0(D^\nu\pi)\rangle}{\langle W_n^0\pi\rangle}\\
&-\frac{\langle W_n^{\nu_2}\pi+W_n^0(D^{\nu_2}\pi)\rangle\times\langle W_n^{\nu_1}\pi+W_n^0(D^{\nu_1}\pi)\rangle}{\langle W_n^0\pi\rangle^2}\\
=:&\sum_{t=1}^n g^\nu(W_t^{(2)},W_{t-1}^{(2)})+g_0^\nu(W_0^{(2)}),
\end{aligned}\tag{88}$$

其中

$$\begin{aligned}
g^\nu(W_t^{(2)},W_{t-1}^{(2)})=&\left\{\frac{\langle W_t^\nu\pi+W_t^{\nu_2}(D^{\nu_1}\pi)+W_t^{\nu_1}(D^{\nu_2}\pi)+W_t^0(D^\nu\pi)\rangle}{\langle W_t^0\pi\rangle}\right.\\
&\left.-\frac{\langle W_{t-1}^\nu\pi+W_{t-1}^{\nu_2}(D^{\nu_1}\pi)+W_{t-1}^{\nu_1}(D^{\nu_2}\pi)+W_{t-1}^0(D^\nu\pi)\rangle}{\langle W_{t-1}^0\pi\rangle}\right\}
\end{aligned}$$

$$-\left\{\frac{\langle W_t^{\nu_2}\pi + W_t^0(D^{\nu_2}\pi)\rangle \times \langle W_t^{\nu_1}\pi + W_t^0(D^{\nu_1}\pi)\rangle}{\langle W_t^0\pi\rangle^2}\right.$$
$$\left.-\frac{\langle W_{t-1}^{\nu_2}\pi + W_{t-1}^0(D^{\nu_2}\pi)\rangle \times \langle W_{t-1}^{\nu_1}\pi + W_{t-1}^0(D^{\nu_1}\pi)\rangle}{\langle W_{t-1}^0\pi\rangle^2}\right\}, \tag{89}$$

and

$$g_0^{\nu}(W_0^{(2)}) = \frac{\langle W_0^{\nu}\pi + W_0^{\nu_2}(D^{\nu_1}\pi) + W_0^{\nu_1}(D^{\nu_2}\pi) + W_0^0(D^{\nu}\pi)\rangle}{\langle W_0^0\pi\rangle} - \frac{\langle W_0^{\nu_2}\pi + W_0^0(D^{\nu_2}\pi)\rangle \times \langle W_0^{\nu_1}\pi + W_0^0(D^{\nu_1}\pi)\rangle}{\langle W_0^0\pi\rangle^2}.$$

也就是说，对数似然函数的二阶导数也可以改写为马尔可夫链 $\{((X_n, Y_n)\ W_n^{(2)})\ n \geq 0\}$ 的加性泛函。

引理*2*的证明。 我们通过数学归纳法证明该引理如下。如式（87）所述，对于所有 $|\nu| = 1$，这样的 g^{ν} 和 g_0^{ν}都存在。现在假设对于所有 $|\nu| < r$， g^{ν} 和 g_0^{ν} 都存在。那么，当 $|\nu| = r$时，取 ν_1 和 ν_2 ，使得$|\nu_2| = 1$ 和 $\nu_1 + \nu_2 = \nu$。根据归纳假设， g^{ν_1} 和 $g_0^{\nu_1}$存在，因此我们有

$$D^{\nu_1+\nu_2}\ell(\theta; Y_{0:n}) = D^{\nu_2}\left\{\sum_{t=1}^{n} g^{\nu_1}(W_t^{(|\nu_1|)}, W_{t-1}^{(|\nu_1|)}) + g_0^{\nu_1}(W_0^{(|\nu_1|)})\right\}$$
$$= \sum_{t=1}^{n} D^{\nu_2} g^{\nu_1}(W_t^{(|\nu_1|)}, W_{t-1}^{(|\nu_1|)}) + D^{\nu_2} g_0^{\nu_1}(W_0^{(|\nu_1|)}).$$

此外，如式（87）所示，$g^{\nu_1}(W_{\ (|\nu_1|)\ t},\ W^{(|\nu_1|)}_{\ t-1})$ 和 $g_0^{\nu_1}(W_{\ (|\nu_1|)\ 0})$ 仅涉及最高阶数为 $|\nu_1|$的导数，因此 $D^{\nu_2}g^{\nu_1}(W_{\ (|\nu_1|)\ t},\ W^{(|\nu_1|)}_{\ t-1})$ 和 $D^{\nu_2}g_0^{\nu_1}(W_{\ (|\nu_1|)\ 0})$ 仅涉及最高阶数为 $|\nu_1| + |\nu_2| = |\nu|$的导数。换言之，它们是 $\{W_n^{(|\nu|)}, n \geq 0\}$ 的函数，因为它们由最高阶数为 $|\nu|$的所有导数组成。因此，对于所有 $|\nu| \leq r$，这样的g^{ν} 和 g_0^{ν} 都存在，证明完毕。 □

注15。 要通过数学归纳法证明（17），实际上只需知道（87）中$|\nu| = 1$ 对应的 g^{ν} 和g_0^{ν} 的精确形式。然而，由于费舍尔信息的刻画特别涉及二阶导数的表示，我们在（88）中给出了 $|\nu| = 2$对应的 g^{ν} 和 g_0^{ν} 的精确形式。

注16。 $D^{\nu}\ell$（θ； $Y_{0:n}$）的表示也弥补了 [39]中的空缺；也就是说， [65], 中的第2.2节提出了如何处理得分函数等的问题。