On construction of Markov chains with given dependence and marginal stationary distributions

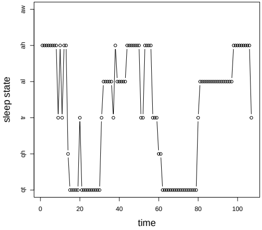
Tomonari Sei

The University of Tokyo

Fall School "Time series, random fields and beyond"
at Ulm University, Sep 23, 2024.

Abstract

- Consider discrete-time processes on a finite state space.



Example: Infant sleep states

(Stoffer et al. 2000)

- We construct Markov models by specifying dependence and marginal distributions separately.

arXiv:2407.17682

Introduction (1/3)

- A Markov model is determined by the Markov kernel (= transition probability matrix), which is designed to have specific dependence relations between the present state x and the future state y.
- For example, consider a Markov kernel

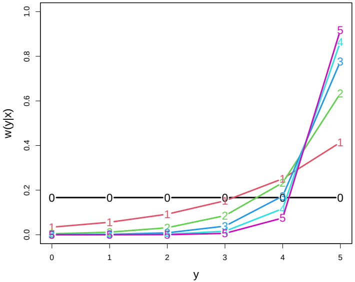
$$w(y|x) = \frac{\exp(\theta xy)}{\sum_{z=0}^{5} \exp(\theta xz)}, \quad x, y \in \{0, \dots, 5\},$$

where $\theta \in \mathbb{R}$ controls the correlation between x and y.
- Problem: the stationary distribution is not directly specified.

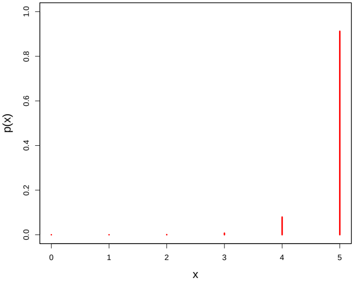
$$\sum_x w(y|x)p(x) = p(y)$$

Introduction (2/3)

Markov kernel
$w(y|x) \propto e^{\theta xy}, \ \ \theta = 0.5$



stationary distribution
$p(x)$



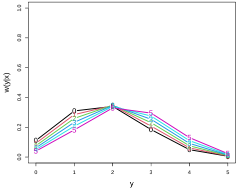
The stationary distribution highly depends on θ.

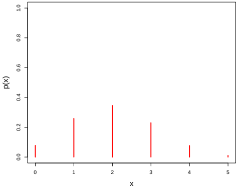
Introduction (3/3)

- It would be convenient if we could design the dependence and stationary distribution separately. It is indeed possible.

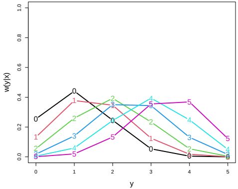
Markov kernel stationary distribution

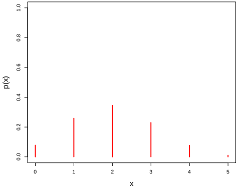
$\theta = 0.1$





$\theta = 0.5$





Preliminaries: Markov kernel

To state the method, we define some symbols and terminology.

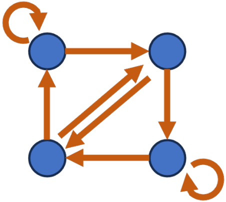
- Let $\mathcal{X}$ be a finite set, which represents the state space.
- Let $\mathbb{R}_+$ and $\mathbb{R}_{\geq 0}$ be the set of positive and non-negative numbers, respectively.
- A Markov kernel on $\mathcal{X}$ is a function $w : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ such that

$$\sum_{y \in \mathcal{X}} w(y|x) = 1$$

  for any $x \in \mathcal{X}$.

Preliminaries: irreducibility

- A graph $(\mathcal{X}, \mathcal{E})$ is said to be strongly connected if for any pair $(x, y) \in \mathcal{X}^2$ there exists a path from x to y.
- A nonnegative matrix $f : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ is said to be irreducible if $\mathrm{supp}(f) = \{(x, y) \in \mathcal{X}^2 \mid f(x, y) > 0\}$ is strongly connected.



A strongly connected graph

Preliminaries: Perron–Frobenius theorem

- Let $\mathcal{P}_+(\mathcal{X})$ denote the set of strictly positive probability distributions on $\mathcal{X}$.

Perron–Frobenius Theorem

If $f : \mathcal{X}^2 \to \mathbb{R}_{\geq 0}$ is irreducible, f has a simple eigenvalue $Z > 0$ and an eigenvector $\gamma \in \mathcal{P}_+(\mathcal{X})$.

- From the Perron–Frobenius theorem, every irreducible Markov kernel w has a unique stationary distribution $p_w \in \mathcal{P}_+(\mathcal{X})$:

$$\sum_{x \in \mathcal{X}} w(y|x) p_w(x) = p_w(y).$$

Main result 1

We begin with first-order Markov chains.

Theorem 1

- Let $H : \mathcal{X}^2 \to \mathbb{R}$ and $r \in \mathcal{P}_+(\mathcal{X})$ be given.

Then, there exists a unique Markov kernel of the form

$$w(y|x) = \exp(H(x,y) + \kappa(y) - \kappa(x) - \delta(y)), \quad (x,y) \in \mathcal{X}^2,$$

with the stationary distribution

$$p_w(x) = r(x), \quad x \in \mathcal{X}.$$

- $H(x,y)$ controls the dependence between x and y.
- $r(x)$ specifies the stationary distribution.
- κ and δ are unique up to an additive constant.

Minimum information Markov model

- From the theorem, we can construct a Markov kernel

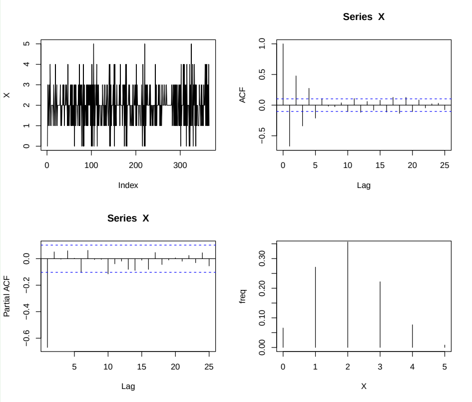
$$\begin{cases} w(y|x) = \exp(H(x,y) + \kappa(y) - \kappa(x) - \delta(y)), \\ p_w(x) = r(x). \end{cases}$$

- We call it the minimum information Markov kernel generated by H and r. This is named after the minimum information copulas (Bedford and Wilson 2014, S. and Yano 2024 etc.).

"The marginal distribution is fixed to $r(x)$."

Example: integer-valued autoregressive process

Let $\mathcal{X} = \{0, 1, \cdots, 5\}$, $H(x, y) = -xy$, $r(x) = \mathrm{Bin}(5, 0.4)$.



sample path	autocorrelation function
partial autocorrelation function	marginal distribution

Remark: Sinkhorn scaling

- Our model is $w(y|x) = e^{H(x,y)+\kappa(y)-\kappa(x)-\delta(y)}$.
- The problem of finding κ and δ is reduced to a system of equations
$$\begin{cases} \sum_y e^{H(x,y)+\alpha(x)+\beta(y)} = r(x), \\ \sum_x e^{H(x,y)+\alpha(x)+\beta(y)} = r(y) \end{cases}$$
with respect to α and β.
- This is the same as Sinkhorn's matrix scaling problem, used in entropic optimal transport: e.g. Nutz (2022).
- In other words, Theorem 1 is just a corollary of the known fact.
- However, this correspondence no longer holds for higher-order Markov chains, as observed below.

Higher-order cases

We next consider d-th-order Markov chains for $d \geq 1$.

- A sequence $(x_s, \ldots, x_t)$ for $s \leq t$ is abbreviated as $x_{s:t}$.
- A d-th-order Markov kernel is a function $w : \mathcal{X}^{d+1} \to \mathbb{R}_{\geq 0}$ such that
$$\sum_{x_{d+1} \in \mathcal{X}} w(x_{d+1} | x_{1:d}) = 1.$$
- Meaning: the future state depends on the past d states.
- The stationary distribution $p_w^{(d)}$ of w is defined by
$$\sum_{x_1} w(x_{d+1} | x_{1:d}) p_w^{(d)}(x_{1:d}) = p_w^{(d)}(x_{2:(d+1)}).$$
- Denote the marginal stationary distribution as
$$p_w^{(1)}(x_1) = \sum_{x_{2:d}} p_w^{(d)}(x_{1:d}).$$

Main result 2

Theorem 2

- Let $H : \mathcal{X}^{d+1} \to \mathbb{R}$ and $r \in \mathcal{P}_+(\mathcal{X})$ be given.

Then, there exists a unique Markov kernel of the form

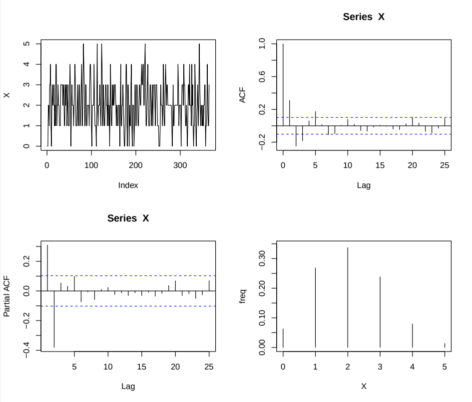
$$\begin{aligned} & w(x_{d+1}|x_{1:d}) \\ & = \exp\left(H(x_{1:(d+1)}) + \kappa(x_{2:(d+1)}) - \kappa(x_{1:d}) - \delta(x_{d+1})\right) \end{aligned}$$

with its marginal stationary distribution

$$p_w^{(1)}(x_1) = r(x_1).$$

Example: integer-valued AR process of order 2

$\mathcal{X} = \{0, 1, \cdots, 5\}$, $H(x, y, z) = 0.6yz - 0.3xz$, $r(x) = \mathrm{Bin}(5, 0.4)$.



sample path	autocorrelation function
partial autocorrelation function	marginal distribution

Exponential family of Markov chains

For proof of the main theorem, we recall information geometry.

Definition (Nagaoka 2005, Hayashi and Watanabe 2016)

- Let $(\mathcal{X}, \mathcal{E})$ be a strongly connected graph.
- Let $C, F_1, \ldots, F_K : \mathcal{E} \to \mathbb{R}$ be given functions.

Then, a family of Markov kernels

$$w_\theta(y|x) = \exp\left(C(x,y) + \sum_{i=1}^{K} \theta_k F_k(x,y) + \kappa_\theta(y) - \kappa_\theta(x) - \psi_\theta\right).$$

supported on $\mathcal{E}$ is called the exponential family generated by $C, F_1, \ldots, F_K$.

Existence of κ_θ and ψ_θ follows from the Perron–Frobenius theorem.

An existence theorem for Markov chains

Theorem (Csiszár et al. 1987)

- Let E be an exponential family generated by $C, F_1, \ldots, F_K$.
- Let M be the set of all Markov kernels w satisfying

$$\sum_{(x,y)\in\mathcal{E}} p_w^{(2)}(x,y)F_k(x,y) = \mu_k, \quad k = 1, \ldots, K$$

for given $\mu_1, \ldots, \mu_K \in \mathbb{R}$.

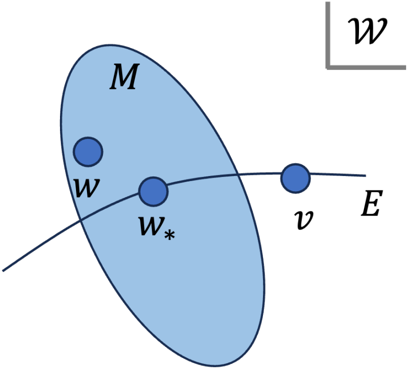
If $M \neq \emptyset$, then there exists a unique $w_* \in M \cap E$.

- Furthermore, we have generalized Pythagorean theorem:

$$D(w|w_*) + D(w_*|v) = D(w|v), \quad w \in M, \quad v \in E$$

for the divergence rate $D(w|v)$. Details are omitted.

Picture



$$M = M_{\mu_1,\dots,\mu_K} = \{w \in \mathcal{W} \mid \sum_{(x,y)} p_w^{(2)}(x,y) F_k(x,y) = \mu_k \ (\forall k)\}$$

$$E = \{v(x,y) = e^{C(x,y)+\sum_k \theta_k F_k(x,y) + \kappa_\theta(y) - \kappa_\theta(x) - \psi_\theta} \mid \theta \in \mathbb{R}^K\}$$

$\mathcal{W}$: the set of all Markov kernels.

Proof of Theorem 1

- Denote $\mathcal{X} = \{\xi_1, \ldots, \xi_m\}$.
- Let $K = m - 1$ and

$$C(x, y) = H(x, y), \quad F_k(x, y) = -I_{\{\xi_k\}}(y), \quad \mu_k = -r(\xi_k).$$

- Then, the generalized Pythagorean theorem

$$\begin{cases} w(y|x) = e^{C(x,y)+\sum_{k=1}^{K} \theta_k F_k(x,y)+\kappa_\theta(y)-\kappa_\theta(x)-\psi_\theta}, \\ \sum_{x,y} p_w^{(2)}(x, y) F_k(x, y) = \mu_k. \end{cases}$$

is read as

$$\begin{cases} w(y|x) = e^{H(x,y)-\delta(y)+\kappa(y)-\kappa(x)}, \\ \sum_{y} p_w^{(1)}(y) = r(y), \end{cases}$$

where $\kappa(y) = \kappa_\theta(y)$ and $\delta(y) = \psi_\theta + \sum_{i=1}^{m-1} \theta_i I_{\{\xi_i\}}(y)$.

- This proves Theorem 1. Theorem 2 is similarly proved.

Future work

Summary

- We proved existence of a Markov kernel that satisfies given dependence and marginal conditions, for finite state spaces.
- Information geometry plays a central role in the proof.

Future work

- Infinite state space (ongoing work)
  - In i.i.d. theory, Csiszár (1975) and Nutz (2022) used Pinsker's inequality
$$\|Q - R\|_{\mathrm{TV}} \leq \sqrt{2D(Q|R)}$$
to prove the existence.
  - A Markov analogue called "Marton's inequality" does not work in the present purpose.
- Relation with INAR models (McKenzie 1985 among others)
- Statistical inference

Thank you for your kind attention!

References I

- Amari, S. and Nagaoka, H. (2000). *Methods of Information Geometry*, American Mathematical Society.
- Bedford, T. and Wilson, K. J. (2014). On the construction of minimum information bivariate copula families, *Ann. Inst. Stat. Math.*, **66**, 703–723.
- Csiszár, I. (1975). I-divergence geometry of probability distributions and minimization problems, *The Annals of Probability*, **3** (1), 146–158.
- Csiszár, I., Cover, T. M. and Choi, B. S. (1987). Conditional limit theorems under Markov conditioning, *IEEE Transactions on Information Theory*, **33** (6), 788–801.
- Hayashi, M. and Watanabe, S. (2016). Information geometry approach to parameter estimation in Markov chains, *The Annals of Statistics*, **44** (4), 1495–1535.
- Küchler, U. and Sørensen, M. (1998). On exponential families of Markov processes, *Journal of Statistical Planning and and inference*, **66**, 3–19.
- McKenzie, E. (1985). Some simple models for discrete variate time series, *JARWA*, **21** (4), 645–650.

References II

- Nagaoka, H. (2005). The exponential family of Markov chains and its information geometry, The 28th Symposium on Information Theory and Its Applications (SITA2005) Onna, Okinawa, Japan, Nov. 20–23, 2005. (arXiv:1701.06119)
- Nutz, M. (2022). *Introduction to Entropic Optimal Transport*, Lecture notes, Columbia University.
- Sei, T. and Yano, K. (2024). Minimum information dependence modelling, *Bernoulli*, **30** (4), 2623–2643.
- Stoffer, D. S. Tyler, D. E. and Wendt, D. A. (2000). The spectral envelope and its applications, *Statistical Science*, **15** (3), 224–253.

Appendix: Divergence rate

- Let $\mathcal{W}$ be the set of Markov kernels supported on $\mathcal{E}$.
- Define the divergence rate of Markov chains by

$$D(v|w) = \sum_{(x,y)\in\mathcal{E}} p_v^{(2)}(x,y) \log \frac{v(y|x)}{w(y|x)}, \quad v, w \in \mathcal{W},$$

- $D(v|w) \geq 0$ with equality if and only if $v = w$.
- Property:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{x_{1:n}\in\mathcal{X}^n} p_v^{(n)}(x_{1:n}) \log \frac{p_v^{(n)}(x_{1:n})}{p_w^{(n)}(x_{1:n})} = D(v|w).$$

Appendix: Proof sketch of the Pythagorean theorem

- If $w \in M$, $w_* \in M \cap E$ and $v \in E$, then

$$
\begin{aligned}
&D(w|w_*) + D(w_*|v) - D(w|v) \\
&= \sum_{(x,y)\in\mathcal{E}} (p_w^{(2)}(x,y) - p_{w_*}^{(2)}(x,y)) \quad \underbrace{\log \frac{v(y|x)}{w_*(y|x)}}_{\in \mathrm{span}(F_1,\ldots,F_K,\mathcal{N})} \\
&= 0.
\end{aligned}
$$

- Uniqueness follows from the identity: if $w, w_* \in M \cap E$, then $D(w|w_*) + D(w_*|w) = D(w|w) = 0$ and so $w = w_*$.
- For existence, it is shown that the function $p_w^{(2)} \mapsto D(w|v)$ is continuous, convex and steep.

See the preprint arXiv:2407.17682 for details. $\square$

Appendix: Conference FDIG 2025

For your information...

- We will hold a conference titled

  Further Developments of Information Geometry (FDIG) 2025

  in March 17–21, 2025 at Tokyo.
- https://sites.google.com/view/fdig2025/
- Contributed talks are welcome by Sep 30 (maybe extended).
- If you have geometric ideas in probability and statistics, please consider to apply!