

観測範囲に制限のある条件下で収集された移動履歴データを用いたマルコフ連鎖パラメタの推定手法

幸島 匡宏^{1,a)} 倉島 健¹ 松林 達史¹ 戸田 浩之¹

概要：GPS 有効範囲の制限や競合他社の存在などにより、人や車などの移動を表すデータの多くは観測可能な範囲に制限のある条件下で収集されている。本研究では、観測可能な地点（状態）間の遷移のみを記録したデータからマルコフ連鎖のパラメタを推定し、観測不可能な地点を含む全地点間の遷移確率を復元する手法を提案する。提案手法の有効性を実データを用いた実験により検証する。

1. はじめに

マルコフ連鎖 (MC: Markov Chain) は確率的に変動する対象をモデル化する汎用的な手法である [1]。古くからの代表例でもある待ち行列 [2] やマーケティング [3]、生物学 [4] への適用だけでなく、近年では、大都市の車や人の流れの分析にも広く用いられている、e.g. [5], [6]。

マルコフ連鎖のパラメタである遷移確率と初期状態確率は一般に未知であることから、観測された移動履歴データから推定する必要がある。図 1a に示すように全ての状態（地点）が観測可能な理想的な設定であれば、パラメタはこれら状態間の遷移の回数を元に容易に推定することができる [7]。しかしながら、GPS 有効範囲の制限や競合他社の存在などの理由により網羅的なデータの収集は困難である。現実利用可能なデータは前述の理想的な設定とは異なり、観測可能な状態（地点）間の遷移のみを記録したデータ（以後センサー遷移データと呼ぶ）として表現される。

センサー遷移データの例はタクシー会社や移動通信事業者から提供される車や人の移動履歴である。図 1b は、GPS 移動履歴からマップマッチング [8] などを利用して作られた、都市における道路またはエリア間の移動履歴を表すデータの例である。このデータには例えば東京都、神奈川のようにデータが提供される地域の移動履歴のみが含まれており、さらにトンネルや高層ビルの影など GPS が受信不可能なエリアの移動履歴は含まれない。よって提供地域外や GPS の受信不可能な場所に位置する観測不可能な状態の移動履歴は利用できず、観測可能な状態間の移動履歴のみが記録されたセンサー遷移データとして表現される。

センサー遷移データはこの例に限定されず他にも様々な例が挙げられる。例えば、都市に設置した複数のカメラ画像から人物照合 [9] して作成する移動履歴は、カメラの設置してある地点間の移動履歴として表現されるためセンサー遷移データになると考えられる。鉄道会社から提供される乗客の移動履歴も、都市全体の中で鉄道駅間の移動履歴のみが記録され、鉄道駅のないエリア間の人の動きは含まれないことから鉄道駅のある観測可能な地点間の移動履歴を表すセンサー遷移データとして表現される。

本研究では、センサー遷移データを用いて元のマルコフ連鎖のパラメタを推定し、観測不可能な状態を含む全状態間の遷移確率を復元する手法を提案する。手法構築の鍵となるのはセンサーマルコフ連鎖 (CMC: Censored Markov Chain) と呼ばれる、観測不可能な状態を持つマルコフ連鎖の理論である。元の MC と CMC の間のパラメタに成り立つ関係式を利用することで、確率モデルのパラメタ推定における損失関数（尤度関数）を自然に導くことができる。提案手法により観測可能な状態間だけでなく、観測不可能な状態を含む任意の状態間の遷移確率を知ることができる。我々の取組む問題は、元となる MC と観測可能な状態の集合から CMC を構成する、CMC の構成法の逆問題と見なすこともできる。

本研究の貢献は以下の通りである。

- センサー遷移データを用いて元のマルコフ連鎖のパラメタを推定するという新しい問題に取り組んだ。
- 上記の問題を CMC の逆問題として定式化した。
- CMC の理論に基づく損失関数を用いた新しい手法を提案した。
- 実データを用いた実験により手法の有効性を検証した。

¹ 日本電信電話株式会社 NTT サービスエボリューション研究所
〒239-0847 神奈川県横浜須賀町光の丘 1-1

^{a)} masahiro.kohjima.ev@hco.ntt.co.jp

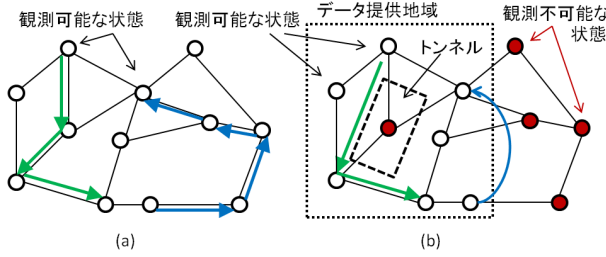


図 1: (a) 理想的な設定と (b) 実環境の設定における移動履歴の例。緑と青の矢印が遷移を表す。(b) の設定では、観測不可能な状態 (赤丸) の訪問は記録されないことから、(b) の遷移データをセンサー遷移データと呼ぶ。

2. 関連研究

近年 MC のパラメタを推定する新しい問題が多数提案され、手法の構築が進められている。例えば、文献 [10], [11] は定常状態確率から MC パラメタを推定する問題、文献 [12], [13] は人口統計情報から推定する問題に取り組んでいる。しかし、これまでにセンサー遷移データからパラメタを推定する問題は取組まれてこなかった。そこで本研究ではこの新しい問題について考える。

CMC の理論は本研究で重要な役割を果たす。CMC は元となる MC と観測可能な状態の集合から構成される。文献 [14], [15] によればこの構成は 1950 年代に Paul Lévy によって示された [16], [17], [18]。CMC については多様な性質が調べられており [14], [19], [20], 無限個の状態を持つ MC の定常状態確率の計算にも利用されている [21]。我々も CMC と元の MC の間で成り立つ状態遷移確率と初期状態確率の関係式を利用して手法を構築する。構築した我々の手法は CMC (に従う移動履歴データ) から元の MC のパラメタを推定する問題を解くものであり、上記 Paul Lévy による CMC 構成の逆問題を解く手法であるとみなすこともできる。なお、我々の問題で利用するセンサー遷移データには、観測不可能な状態は一度も現れないという設定を考えており、状態がランダムに欠損する設定とは全く異なる問題であることを強調しておく。

3. マルコフ連鎖とセンサーマルコフ連鎖

$\mathcal{X} = \{1, 2, \dots, |\mathcal{X}|\}$ を有限の状態空間とする。 $\mathcal{X}$ 上の離散時間マルコフ連鎖 (MC) は以下のマルコフ性を持つ確率過程 $\{X_t; t = 0, 1, 2, \dots\}$ として定義される。

$$\begin{aligned} Pr(X_{t+1} = x_{t+1} | X_k = x_k; k = 0, \dots, t) \\ = Pr(X_{t+1} = x_{t+1} | X_t = x_t) \quad (\forall x_k \in \mathcal{X}, \forall t \in \mathbb{Z}_{\geq 0}). \end{aligned}$$

MC は 3 つ組 $\{\mathcal{X}, \mathcal{P}, q\}$ で定義できる。 $\mathcal{P} : \mathcal{X} \times \mathcal{X} \rightarrow [0, 1]$ は遷移確率、 $q : \mathcal{X} \rightarrow [0, 1]$ は初期状態確率である。すなわち、 $\mathcal{P}(x_{next}|x) = Pr(X_{t+1} = x_{next} | X_t = x)$,

$q(x_0) = Pr(X_0 = x_0)$ 。本論文を通して対象となる MC は既約^{*1}であるとする。また、各状態から到達可能な状態を表す隣接情報を $\Gamma = \{\Gamma_i\}_{i \in \mathcal{X}}$ と書くこととする。 Γ_i が状態 i から 1 ステップの遷移で到達可能な集合を表す。

次にセンサーマルコフ連鎖 (CMC) を定義する^{*2}。 $\mathcal{O}$ を状態空間 $\mathcal{X}$ の部分集合 $\mathcal{O} \subseteq \mathcal{X}$ であるとする。 $\mathcal{O}$ は観測可能な状態の集合を表すこととし、観測不可能な状態の集合は $\mathcal{U}$ と書く。CMC は元となる MC $\{X_t; t = 0, 1, 2, \dots\}$ と観測可能状態の集合 $\mathcal{O}$ を用いて構築される。CMC を表す確率過程 $\{X_t^c; t = 0, 1, 2, \dots\}$ は、時刻 t における状態 X_t^c が、元の MC で観測可能状態の集合を t 回目訪問した時刻 (これを σ_t と書く) における状態 X_{σ_t} を表すものとして定義される。直感的に言えば、CMC は元の MC の系列から $\mathcal{O}$ に含まれる状態の訪問のみを抜き出して作られたものといえる。フォーマルな定義は以下の通りである。

定義 1 (センサーマルコフ連鎖) $\{\sigma_t; t = 0, 1, 2, \dots\}$ が $X_0 \in \mathcal{O}$ の時 $\sigma_0 = 0$ 、そうでなければ $\sigma_0 = \inf\{m \geq 1 : X_m \in \mathcal{O}\}$ 、さらに $\sigma_t = \inf\{m > \sigma_{t-1} : X_m \in \mathcal{O}\}$ となる系列であるとする。 $X_t^c = X_{\sigma_t}$ となる確率過程 $\{X_t^c; t = 0, 1, 2, \dots\}$ をセンサーマルコフ連鎖と呼ぶ。

これ以降、一般性を失うことなく、状態は並び替えられて状態遷移確率と初期状態確率を表す行列 $\mathbf{P}, (\mathbf{P})_{xx'} = \mathcal{P}(x'|x)$ とベクトル $\mathbf{q}, (\mathbf{q})_x = q(x)$ は以下で表現されることとする。

$$\mathbf{P} = \begin{matrix} & \mathcal{O} & \mathcal{U} \\ \begin{matrix} \mathcal{O} \\ \mathcal{U} \end{matrix} & \begin{pmatrix} \mathbf{P}_{oo} & \mathbf{P}_{ou} \\ \mathbf{P}_{uo} & \mathbf{P}_{uu} \end{pmatrix} \end{matrix}, \quad \mathbf{q} = \begin{pmatrix} \mathbf{q}_o & \mathbf{q}_u \end{pmatrix}. \quad (1)$$

ただし、 $\mathbf{P}_{oo}, \mathbf{P}_{ou}, \mathbf{P}_{uo}, \mathbf{P}_{uu}$ はそれぞれサイズが $|\mathcal{O}| \times |\mathcal{O}|$, $|\mathcal{O}| \times |\mathcal{U}|$, $|\mathcal{U}| \times |\mathcal{O}|$, $|\mathcal{U}| \times |\mathcal{U}|$ の行列である。

センサーマルコフ連鎖について以下の定理が成り立つ。

定理 1 (e.g., Lemma 6-6 [19]) センサーマルコフ連鎖は以下の状態遷移確率 $\mathbf{R}$ で定義されるマルコフ連鎖である。 $\mathbf{R} = \mathbf{P}_{oo} + \mathbf{P}_{ou}(\mathbf{I} - \mathbf{P}_{uu})^{-1}\mathbf{P}_{uo}$ 。

式中の逆行列の項 $(\mathbf{I} - \mathbf{P}_{uu})^{-1}$ は $\mathbf{P}_{uu}$ のべき乗の無限和 $\sum_{\ell=0}^{\infty} (\mathbf{P}_{uu})^{\ell}$ から得られた項であり、既約な MC においては逆行列の存在は保証される。 $\mathbf{R}$ の式は次のように直感的に解釈することができる。CMC の観測可能な状態間の遷移確率 $\mathbf{R}$ は、元となる MC で (i) 同じ観測可能な状態間の遷移を行う確率 $\mathbf{P}_{oo}$ と (ii) 一度観測不可能な状態に遷移してから $(\mathbf{P}_{ou})$ 、観測不可能な状態間で任意の長さの遷移を行なった後 $(\sum_{\ell=0}^{\infty} (\mathbf{P}_{uu})^{\ell})$ 、観測可能な状態に遷移する $(\mathbf{P}_{uo})$ 確率の和として表現されている。上記の定理の導出とほぼ

^{*1} 任意の状態から任意の状態へ到達できることを表す

^{*2} 他にも *watched Markov chain* や *induced chain* と呼ばれることもある [15], [19], [22]

同様に初期状態確率に関する結果も得ることができる。

定理 2 センサーマルコフ連鎖の初期状態確率は $\mathbf{s} = \mathbf{q}_o + \mathbf{q}_u(\mathbf{I} - \mathbf{P}_{uu})^{-1}\mathbf{P}_{uo}$ で与えられる。

定理 1 と 2 より, マルコフ連鎖 $\{\mathcal{X}, \mathcal{P}, q\}$ と観測可能状態の集合 $\mathcal{O}$ から作られるセンサーマルコフ連鎖が, マルコフ連鎖 $\{\mathcal{O}, \mathcal{R}, s\}$ であることがわかる。

4. センサー遷移データを用いたマルコフ連鎖パラメタの推定

この章でセンサー遷移データを用いてマルコフ連鎖パラメタを推定する問題の定式化とそれを解く手法について示す。図 2 に定式化の全体像を示す。我々の目的は CMC(に従う移動履歴データ) から元の MC のパラメタを復元するものであるから, 前の章で示した MC から CMC を構成する方法の逆問題を解く手法とみなすこともできる。

4.1 問題の定式化

入力データ: 本問題の入力データは (i) 元のマルコフ連鎖の状態集合 $\mathcal{X}$, (ii) 隣接情報 $\Gamma = \{\Gamma_i\}_{i \in \mathcal{X}}$, (iii) 観測可能な状態の集合 $\mathcal{O}$, (iv) センサー遷移データの 4 つから成る。センサー遷移データは $\mathcal{D} = \{N_{ij}\}_{ij \in \mathcal{O}} \cup \{N_k^{ini}\}_{k \in \mathcal{O}}$ としてまとめて表現されているとする。 N_{ij} は (観測可能な) 状態 $i \in \mathcal{O}$ から状態 $j \in \mathcal{O}$ へ遷移した回数, N_k^{ini} は初期状態が状態 $k \in \mathcal{O}$ である回数を表す。

図 1b で示した例の場合, 状態集合 $\mathcal{X}$ と隣接情報 Γ は公開された道路ネットワークや交通ネットワークから作ることができる。観測可能状態の集合 $\mathcal{O}$ はセンサー遷移データ中に存在する状態を抽出して作成しても良い。また, 上記の $\mathcal{D}$ の定義より, 個人レベルの移動履歴データは利用できなくても集計された移動履歴データが使えれば良い^{*3}。

入力モデル: 入力モデルには, 元の MC における状態遷移確率と初期状態確率を表現できる (任意の) モデルが利用できる。ここでは $\theta = (\nu, \lambda)$ をモデルのパラメタとして, パラメタへの依存性を明示するために, 状態遷移確率と初期状態確率のモデルを P^ν, q^λ と書くこととする。モデルとパラメタが与えられれば, 状態遷移と初期状態の確率は, $Pr(X_{t+1} = j | X_t = i, \theta) = (P^\nu)_{ij}, Pr(X_0 = k | \theta) = (q^\lambda)_k$ で与えられる。利用するモデルの例には下記の表形式モデルや対数線形モデルが挙げられる。

モデル例 1 (表形式モデル) モデルのパラメタ ν, λ が $\nu = \{a_{ij}\}_{j \in \Gamma_i, i \in \mathcal{X}}, \lambda = \{b_i\}_{i \in \mathcal{X}}$ で定義され, $\sum_j a_{ij} = 1$ (for all i), $\sum_i b_i = 1$ を満たすとする。表形式モデルは次

^{*3} プライバシー保護などの理由によりユーザがいつその状態を訪れたかを表す時間情報は提供されない/利用できない状況であっても, 状態を訪問した順番が利用できれば集計してセンサー遷移データを作ることができる。

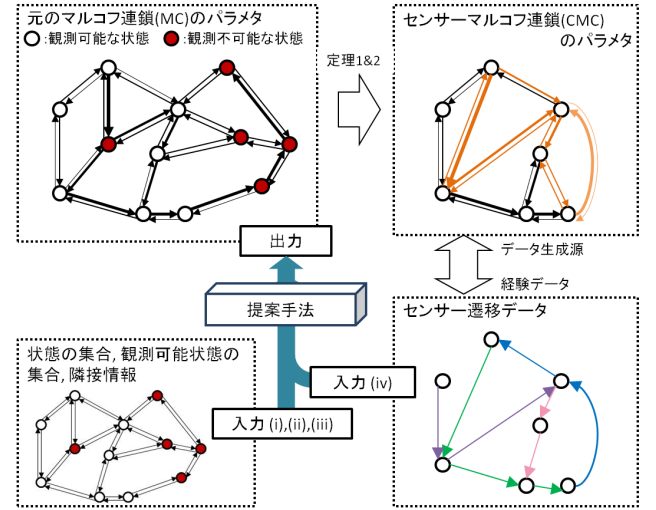


図 2: 本研究の問題設定。提案手法はセンサー遷移データから元のマルコフ連鎖のパラメタを推定する。

の式で定義される。 $(P^\nu)_{ij} = a_{ij}$ if $j \in \Gamma_i$, and $(P^\nu)_{ij} = 0$ otherwise, and $(q^\lambda)_k = b_k$ 。

モデル例 2 (対数線形モデル) モデルのパラメタ ν, λ が $\nu = \{v^{loc}, v^{glo}\}$ and $\lambda = \{w^{loc}, w^{glo}\}$ で定義されているとする。このとき, 対数線形モデルは次の式で定義される。

$$(P^\nu)_{ij} = \begin{cases} \frac{\exp\{v_{ij}^{loc} + \phi(i, j)^T v^{glo}\}}{\sum_{k \in \Gamma_i} \exp\{v_{ik}^{loc} + \phi(i, k)^T v^{glo}\}} & (\text{if } j \in \Gamma_i) \\ 0 & (\text{otherwise}) \end{cases} \quad (2)$$

$$(q^\lambda)_i = \frac{\exp\{w_i^{loc} + \psi(i)^T w^{glo}\}}{\sum_k \exp\{w_k^{loc} + \psi(k)^T w^{glo}\}}. \quad (3)$$

ただし, $\phi(i, j)$, $\psi(k)$ は特徴ベクトルで $\phi(i, j)$ が状態 i と j の逆距離, $\psi(k)$ が状態 k の属性情報などを表すものとする。そのような情報が利用できなければ, 特徴ベクトルとパラメタ v^{glo}, w^{glo} に関する項をモデルから除けば良い^{*4}。

なお, 式 (1) と同様に一般性を失うことなくモデルの行列表現が以下の式で与えられるとする。

$$P^\nu = \begin{matrix} & \mathcal{O} & \mathcal{U} \\ \begin{matrix} \mathcal{O} \\ \mathcal{U} \end{matrix} & \begin{pmatrix} P_{oo}^\nu & P_{ou}^\nu \\ P_{uo}^\nu & P_{uu}^\nu \end{pmatrix} \end{matrix}, \quad q^\lambda = \begin{pmatrix} q_o^\lambda & q_u^\lambda \end{pmatrix}$$

出力: 提案手法の出力はモデルのパラメタ $\theta = (\nu, \lambda)$ である。これによって観測不可能な地点を含む全地点間の遷移確率 P^ν と初期状態確率 q^λ が復元できる。

この問題設定は, 仮に全ての状態が観測可能な状態 $\mathcal{X} = \mathcal{O}$ となる場合, 通常の状態遷移確率推定の設定 [7] と等しい。

^{*4} 対数線形モデルは表形式モデルの特殊ケースでもある。特徴ベクトルに関する項を除き $a_{ij} = \exp\{v_{ij}^{loc}\} / \sum_{k \in \Gamma_i} \exp\{v_{ik}^{loc}\}$ と定義すれば, 表形式モデルを表現できることから明らか。

4.2 提案手法

提案手法は CMC の理論から導かれる損失関数を最小化することでパラメタを推定する。損失関数には L_2 ダイバージェンスや Kullback Leibler (KL) ダイバージェンスなど任意の関数が利用できるが、ここでは以下で定義される KL ダイバージェンスを用いる場合の手法について示す。

$$KL(q||p) = \sum_j q_j \log q_j - q_j \log p_j.$$

ただし、 $q = \{q_j\}$ と $p = \{p_j\}$ は離散の確率分布である。KL ダイバージェンスを用いることで負の対数尤度関数と対応する目的関数が導かれる。

ここから CMC の理論を利用する。図 2 で示すようにセンサー遷移データ $\mathcal{D}$ は $CMC\{\mathcal{O}, \mathbf{R}^*, \mathbf{s}^*\}$ から経験的に得られた遷移データとみなすことができる。ただし、 $\mathbf{R}^*$ と $\mathbf{s}^*$ は元の MC の真のパラメタ $\mathbf{P}^*$ と $\mathbf{q}^*$ と $\mathcal{O}$ から作られる CMC の真の遷移確率と初期状態確率を表す。定理 1 と 2 より、 $MC\{\mathcal{X}, P^\nu, q^\lambda\}$ と観測可能な状態の集合 $\mathcal{O}$ から作られる $CMC\{\mathcal{O}, \mathbf{R}^\nu, \mathbf{s}^{\nu,\lambda}\}$ の $\mathbf{R}^\nu$ と $\mathbf{s}^{\nu,\lambda}$ は以下の式で与えられる。

$$\begin{aligned} \mathbf{R}^\nu &\triangleq \mathbf{P}_{oo}^\nu + \mathbf{P}_{ou}^\nu (\mathbf{I} - \mathbf{P}_{uu}^\nu)^{-1} \mathbf{P}_{uo}^\nu, \\ \mathbf{s}^{\nu,\lambda} &\triangleq \mathbf{q}_o^\lambda + \mathbf{q}_u^\lambda (\mathbf{I} - \mathbf{P}_{uu}^\nu)^{-1} \mathbf{P}_{uo}^\nu. \end{aligned}$$

なお $\mathbf{R}^\nu$ と $\mathbf{s}^{\nu,\lambda}$ の要素はそれぞれ次の遷移確率と初期状態確率に対応する。 $Pr(X_{t+1}^c = j | X_t^c = i, \theta) = (\mathbf{R}^\nu)_{ij}$, $Pr(X_0^c = k | \theta) = (\mathbf{s}^{\nu,\lambda})_k$ 。したがって、 $\mathbf{R}^\nu$ と $\mathbf{R}^*$, $\mathbf{s}^{\nu,\lambda}$ と $\mathbf{s}^*$ の KL ダイバージェンスは以下の式で与えられる。

$$\begin{aligned} L_R^{KL}(\mathbf{R}^* || \mathbf{R}^\nu) &= \sum_{i,j \in \mathcal{O}} (\mathbf{R}^*)_{ij} \log (\mathbf{R}^*)_{ij} - (\mathbf{R}^*)_{ij} \log (\mathbf{R}^\nu)_{ij}, \\ L_s^{KL}(\mathbf{s}^* || \mathbf{s}^{\nu,\lambda}) &= \sum_{i \in \mathcal{O}} (\mathbf{s}^*)_i \log (\mathbf{s}^*)_i - (\mathbf{s}^*)_i \log (\mathbf{s}^{\nu,\lambda})_i. \end{aligned}$$

目的関数を上記の 2 つのダイバージェンスの和に正則化項を加えたもので定義する。CMC の真のパラメタ $\mathbf{R}^*$ と $\mathbf{s}^*$ に関する項を標本平均で置き換えて定数項を除去すれば、以下の目的関数が導かれる。

$$\begin{aligned} \mathcal{L}(\theta) &= -\mathcal{Z}^{-1} \sum_{i,j \in \mathcal{O}} N_{ij} \log (\mathbf{R}^\nu)_{ij} \\ &\quad - \mathcal{Z}_{ini}^{-1} \sum_{k \in \mathcal{O}} N_k^{ini} \log (\mathbf{s}^{\nu,\lambda})_k + \Omega(\theta). \end{aligned} \quad (4)$$

ただし、 $\mathcal{Z}$ と $\mathcal{Z}_{ini}$ は正規化項、 $\Omega(\theta)$ が正則化項である。のちの実験では $\mathcal{Z} = \sum_{i,j} N_{ij}$, $\mathcal{Z}_{ini} = \sum_k N_k^{ini}$ と設定し、 L_2 正則化項 $\Omega(\theta) = \alpha \|\theta\|^2$ を利用した。 α はハイパーパラメタである。パラメタ θ はこの目的関数を最適化で得られる。

$$\hat{\theta} = \arg \min_{\theta} \mathcal{L}(\theta).$$

目的関数は正則化項を含んだ負の対数尤度であると言え、この推定量は (正則化) 最尤推定量であると言える。目的関数 (式 (4)) の最小化には、勾配法やニュートン法など任意の最適化手法が利用できる。実験では L-BFGS 法 [23] を利用する。

5. 実験

この章で提案手法がセンサー遷移データから元の MC のパラメタを推定できていることを確認する。評価のために 2 種類の実験設定のそれぞれで学習データとテスト MC データを用意する。テスト MC データは図 1a のような観測不可能状態のない元の MC から生成されたデータ、学習データは図 1b のようなセンサー遷移データの形式で与えられるデータである。手法のパフォーマンスはテスト MC データを用いて評価する。また、提案手法が元の MC に加えて CMC も表現できていることを確認するため、CMC から生成されたテストデータ (テスト CMC データ) も用意する。

5.1 実験設定

人工データ: 人工データを用いた実験では、文献 [24] と同様の方法で状態数 100 のマルコフ連鎖を作成し利用した。真の遷移確率は対数線形モデル (式 (2), (3)) をベースに設定した。真のパラメタ ν^* , λ^* と特徴量 ϕ を標準正規分布 (特徴量 ψ は利用しない)、ノイズ成分をパラメタ 0.3 の対称ディリクレ分布を用いて生成し、ノイズ成分に $\beta = 0.1$ を掛けた値と遷移確率に $1 - \beta = 0.9$ を掛けた値の和を真の遷移確率とした。これを用いて 20 ステップのエピソード (初期状態からの遷移の系列) を生成することで学習データと 2 種類のテストデータの元となるデータを用意した。学習データとテストデータのエピソード数はそれぞれ 100, 1000 である。さらに状態集合の中から観測可能な集合とするものをランダムに選び、学習データとテスト CMC データ用のエピソード中で観測不可能状態の訪問した情報を除外することで学習データとテスト CMC データを作成した。観測可能な状態の比率 $|\mathcal{O}|/|\mathcal{X}|$ を変えながらこの手続きを繰り返し、各比率ごとに学習データ、テスト MC データ、テスト CMC から成る計 5 つのデータの組を用意した。

実データ: 実データを用いた実験では、株式会社ナビタイムジャパン提供のデータを利用した。このデータは首都圏におけるカーナビアプリケーション利用者の移動軌跡を記録したものである。2015.4.13 から 2015.4.17 までの平日 5 日間の 6:00 am から 10:59 am までの時間のデータを利用し、各セルのサイズが 5km $\times$ 5km の地域メッシュのセル間の移動履歴に変換して利用した。セルの総数は約 150, 一日当たりの平均ユーザ数は約 8000 となった。なお、一日当たり 20 回より少ない状態 (セル) は取り除き、3 ステップ以上移動しているユーザの遷移のみを利用した。特徴量 ϕ , ψ は利用しなかった。学習データとテストデータは上記のある 1 日のデータとその次の日のデータを用いることとし、学習データとテスト CMC データは人工データ実験と同様に観測不可能状態の訪問した情報を除外して作成した。作成した学習データ、テスト MC データ、テスト CMC から成る計 4 つのデータの組を利用して実験を行った。

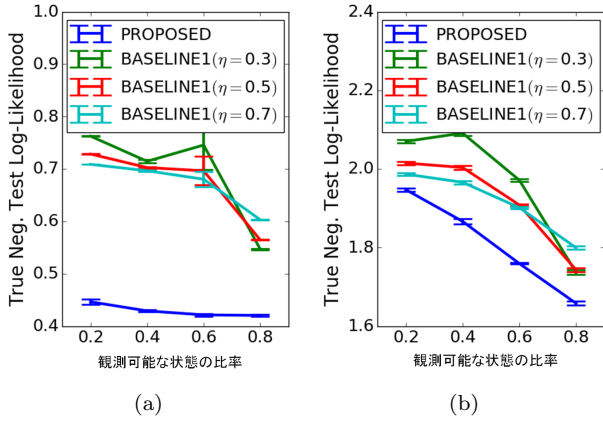


図 3: (主結果) (a) 人工データ実験と (b) 実データ実験における *True negative test log-likelihood* の平均と標準偏差。小さいほど良い。横軸は観測可能な状態の比率 $|O|/|X|$ 。

評価尺度: 評価尺度には、テスト MC データを利用して計算する負の対数尤度 (*true negative test log likelihood*) を利用する。 *True negative test log likelihood* は次のように定義した。 $(1/T_{tt}) \sum_{i,j \in X} -n_{ij}^{tt} \log \hat{p}_{ij}$ 。ただし、 T_{tt} は総遷移数、 n_{ij}^{tt} はテスト MC データで状態 i から状態 j へ遷移した回数を表す。 $\hat{p}_{ij}$ は推定した元の MC の遷移確率である。値が小さいほど提案手法は元の MC のパラメタを精度よく推定できていることを示す。さらに提案手法が MC だけでなく CMC のパラメタも同様に精度よく推定できているのかを確認するため、テスト CMC データを利用して計算する負の対数尤度 (*censored negative test log likelihood*) も合わせて計算した。 *Censored negative test log likelihood* は、 $(1/T_{ct}) \sum_{i,j \in O} -n_{ij}^{ct} \log \hat{p}_{ij}$ と定義され、 T_{ct} は総遷移数、 n_{ij}^{ct} はテスト CMC データで状態 i から状態 j へ遷移した回数を表す。 $\hat{p}_{ij}$ は推定した CMC の遷移確率である。

比較手法: 提案手法はセンサー遷移データから元の MC パラメタを推定できるのはじめての手法であるため、“強い”比較手法は存在しない^{*5}。そこで我々はベースライン手法として、通常のマルコフ連鎖モデルと一様分布を混合させたモデルを利用することにした。これは、観測不可能な状態が少ない時ならば通常のマルコフ連鎖モデルを適用することとはリーズナブルな結果をもたらすと期待できるし、さらに観測不可能な状態が多いのであれば、隣接情報を利用した一様分布を用いることはシンプルな妥協案であると考えられるからである。この方針に基づき、我々は Baseline1 (Base1) と Baseline2 (Base2) の 2 種類のベースラインを利用した。それぞれの遷移確率の推定方法は以下の通りである。

$$\hat{p}_{ij}^{\text{Base1}} = (1-\eta)P_{ij}^{M1} + \eta P_{ij}^{U1}, \quad \hat{p}_{ij}^{\text{Base2}} = (1-\xi)P_{ij}^{M2} + \xi P_{ij}^{U2},$$

^{*5} 観測不可能な状態がセンサー遷移データ中に一度も現れないため、隠れマルコフモデルやベイジアンネットワークなどの手法も今回の問題に適用することはできない。

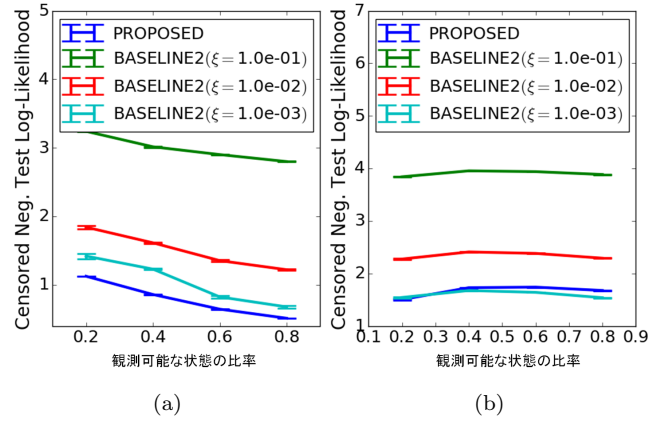


図 4: (a) 人工データ実験と (b) 実データ実験における *Censored negative test log-likelihood* の平均と標準偏差。小さいほど良い。横軸は観測可能な状態の比率 $|O|/|X|$ 。

ただし、 $P_{ij}^{M1} = N_{ij}/Z_i$ if $j \in \Gamma_i$ and $P_{ij}^{M1} = 0$ otherwise, $P_{ij}^{U1} = 1/|\Gamma_i|$ if $j \in \Gamma_i$ and $P_{ij}^{U1} = 0$ otherwise, $P_{ij}^{M2} = N_{ij}/Z'_i$, $P_{ij}^{U2} = 1/|X|$ とした。 Z_i と Z'_i は正規化定数であり、 η と ξ はハイパーパラメタである。ベースラインのハイパーパラメタ η, ξ は $\{10^{-3}, 10^{-2}, 10^{-1}, 0.3, 0.5, 0.7\}$ の中から選び、最も性能の良かった上位 3 つの結果を次節では示す。なお、提案手法のハイパーパラメタは共通で $\alpha = 10^{-3}$ と設定した^{*6}。

5.2 実験結果

図 3 に *true negative test log-likelihood* の結果を示す。観測可能な状態の比率に依らず提案手法がベースラインを上回る性能を示していることがわかる。この結果は元の MC のパラメタを推定する提案手法の有効性を示していると言える。図 4 に *censored negative test log-likelihood* の結果を示す。観測可能な状態の比率に依らず提案手法がベースラインと同等かそれを上回る性能を示していることがわかる。提案手法は元の MC のパラメタだけでなく、CMC のより精度よく推定できることがわかる。図 3, 4 から提案手法は MC の推定精度と CMC の推定精度のどちらかを犠牲にすることなく推定できることが分かる。

図 5 に真の遷移確率と推定結果の遷移確率を示す。提案手法の推定結果はベースラインのそれよりも真の遷移確率に近い結果となっていることが読み取れる。これらの結果から提案手法の有効性を確かめることができた。

^{*6} 我々の問題設定では (センサー遷移データを用いた) 交差検定によってハイパーパラメタを設定することは適切であるとは言えない。なぜなら *censored negative test log likelihood* の意味で良い設定値と *true test log-likelihood* の意味で良い設定値が一致するとは期待できないからである。実際ベースラインでもハイパーパラメタの値が小さいほうが *censored log-likelihood* の値は良くなるが *true log-likelihood performance* の値は悪くなる傾向がある。テスト MC データも実環境では手に入るとも限らない。そこで本実験では、人工データ実験で安定して良い性能を示した $\alpha = 10^{-3}$ を全実験で共通して利用することとした。

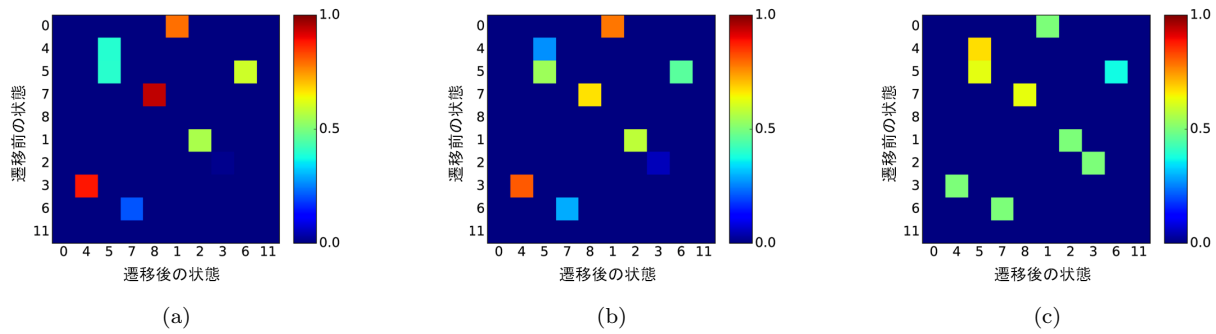


図 5: (a) 真の遷移確率と (b) 提案手法, (c) Baseline1 によって推定した遷移確率. 観測可能な状態の比率が 0.5 の時の人工データ実験の結果. 抽出した 10 状態の間の遷移確率を示す. 最初の 5 状態 (0, 4, 5, 7, 8) が観測可能な状態, 残りが観測不可能な状態である.

6. まとめ

本研究では, センサー遷移データを用いて元のマルコフ連鎖のパラメタを推定する問題に取り組んだ. この問題を CMC の逆問題として定式化し, CMC の理論を利用して手法を構築した. 人工データと実データを用いた実験によって提案手法の有効性を確かめた. 今後の研究には, 観測可能な状態の比率が手法のパフォーマンスに与える影響を理論的に解明することなどが挙げられる.

参考文献

- [1] 森村英典, 高橋幸雄: マルコフ解析, 日科技連出版社 (1979).
- [2] Neuts, M. F.: *Matrix-geometric solutions in stochastic models: an algorithmic approach*, Johns Hopkins University Press (1981).
- [3] Pfeifer, P. E. and Carraway, R. L.: Modeling customer relationships as Markov chains, *Journal of interactive marketing*, Vol. 14, No. 2, pp. 43–55 (2000).
- [4] Nowak, M. A.: *Evolutionary dynamics*, Harvard University Press (2006).
- [5] Crisostomi, E., Kirkland, S. and Shorten, R.: A Google-like model of road network dynamics and its application to regulation and control, *International Journal of Control*, Vol. 84, No. 3, pp. 633–651 (2011).
- [6] Fan, Z., Song, X., Shibasaki, R. and Adachi, R.: CityMomentum: an online approach for crowd behavior prediction at a citywide level, *International Joint Conference on Pervasive and Ubiquitous Computing*, pp. 559–569 (2015).
- [7] Billingsley, P.: *Statistical inference for Markov processes*, University of Chicago Press (1961).
- [8] Zheng, Y. and Zhou, X.: *Computing with spatial trajectories*, Springer Science & Business Media (2011).
- [9] Zheng, L., Yang, Y. and Hauptmann, A. G.: Person re-identification: Past, present and future, *arXiv preprint arXiv:1610.02984* (2016).
- [10] Morimura, T., Osogami, T. and Idé, T.: Solving inverse problem of Markov chain with partial observations, *Advances in Neural Information Processing Systems*, pp. 1655–1663 (2013).
- [11] Kumar, R., Tomkins, A., Vassilvitskii, S. and Vee, E.: Inverting a steady-state, *International Conference on Web Search and Data Mining*, pp. 359–368 (2015).
- [12] Iwata, T., Shimizu, H., Naya, F. and Ueda, N.: Estimating people flow from spatiotemporal population data via collective graphical mixture models, *ACM Transactions on Spatial Algorithms and Systems*, Vol. 3, No. 1, p. 2 (2017).
- [13] Akagi, Y., Nishimura, T., Kurashima, T. and Toda, H.: A Fast and Accurate Method for Estimating People Flow from Spatiotemporal Population Data., *International Joint Conferences on Artificial Intelligence*, pp. 3293–3300 (2018).
- [14] Freedman, D.: *Approximating countable Markov chains*, Springer-Verlag New York (1983).
- [15] Zhao, Y. Q. and Liu, D.: The censored Markov chain and the best augmentation, *Journal of Applied Probability*, Vol. 33, No. 3, pp. 623–629 (1996).
- [16] Lévy, P.: Systèmes markoviens et stationnaires. Cas dénombrable., *Ann. Sci. École Norm. Sup.*, Vol. 68, No. 3, pp. 327–381 (1951).
- [17] Lévy, P.: Complément à l’étude des processus de Markoff, *Ann. Sci. École Norm. Sup.*, Vol. 69, No. 3, pp. 203–212 (1952).
- [18] Lévy, P.: Processus markoviens et stationnaires. Cas dénombrable., *Ann. Inst. H. Poincaré*, Vol. 18, pp. 7–25 (1958).
- [19] Kemeny, J. G., Snell, J. L. and Knapp, A. W.: *Denumerable Markov chains*, Springer-Verlag New York (1976).
- [20] 牧本直樹: 待ち行列アルゴリズム: 行列解析アプローチ, 朝倉書店 (2001).
- [21] Grassmann, W. K. and Heyman, D. P.: Computation of steady-state probabilities for infinite-state Markov chains with repeating rows, *ORSA Journal on Computing*, Vol. 5, No. 3, pp. 292–303 (1993).
- [22] Levin, D. A. and Peres, Y.: *Markov chains and mixing times*, American Mathematical Soc. (2017).
- [23] Liu, D. C. and Nocedal, J.: On the limited memory BFGS method for large scale optimization, *Mathematical programming*, Vol. 45, No. 1-3, pp. 503–528 (1989).
- [24] Morimura, T., Uchibe, E., Yoshimoto, J. and Doya, K.: A generalized natural actor-critic algorithm, *Advances in Neural Information Processing Systems*, pp. 1312–1320 (2009).