位置情報履歴の欠損と周期性を考慮したパターン抽出手法

林 亜紀† 亀岡 弘和†† 松林 達史† 澤田 宏†

† 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

†† 日本電信電話株式会社 NTT コミュニケーション科学基礎研究所

{*hayashi.aki, kameoka.hirokazu, matsubayashi.tatsushi, sawada.hiroshi*}*@lab.ntt.co.jp*

概要 情報配信，都市計画，人流制御など多くの目的において，行動パターンの把握が重要視されている．本論文では，可変基底 NMF を位置情報履歴の特性を考慮して拡張することにより，従来よりも解釈が容易な行動パターンの抽出を試みる．行列分解などの従来技術では，行動パターンを階層的に抽出できず，抽出されたパターンの解釈が難解な場合があった．提案法では，行動パターンを，いくつかの類似した行動系列で表現することにより，課題を解決する．この際，可変基底 NMF を拡張し，位置情報履歴に多く見られる履歴の欠損を考慮したパターンを抽出する．また，位置情報履歴が持つ周期性も考慮できるように可変基底 NMF を拡張し，24 時間，1 週間などの異なる周期を持つパターンの抽出を促進する．例えば出勤日と休日におけるいくつかの 1 日の行動パターン推移をそれぞれ独立して抽出した上で，平日は会社，休日はその他といった 1 週間周期の行動パターンも抽出することが可能になる．

キーワード 位置情報，パターン抽出，可変基底 NMF，欠損値補間，周期性

1 はじめに

近年，モバイルデバイスやソーシャルネットワーキングサービスが普及し，ユーザ毎の位置情報履歴の観察が可能になった．位置情報履歴に頻出する，例えば平日は朝自宅を出て会社に出勤し，昼には会社近くのいくつかの飲食店のうち 1 つで昼食を取り，夜になると会社を退社して自宅に帰る，といった連続行動パターン（以下パターン）を抽出する取り組みは，情報配信，都市計画，人流制御による混雑緩和など多くの場面で有効である．

時系列データから，パターンを抽出する方法として，非負値行列分解（Nonnegative Matrix Factorization, NMF) [5] の有効性が示されている．NMF は非負値ベクトルの時系列を並べたデータ行列を基底行列と係数行列の積に分解する手法である．位置情報履歴におけるパターン抽出においても，NMF を行列ではなくテンソルに拡張した非負値テンソル分解（Nonnegative Tensor Factorization, NTF) [6] の有効性が示されている [1]．上記手法では，位置情報履歴をユーザ × 場所 × 時間の 3 次元テンソルで表現した上で，そのテンソルを 3 つの行列の積の形で近似することにより，各時間の行動をいくつかの訪問傾向に分類し，行動パターンを抽出する．しかしながら，抽出する訪問傾向の種類数（基底数）は観察者により手動で設定され，多すぎると，類似した訪問傾向がいくつもの基底で表現されてしまい，解釈が難しくなる．一方で，基底数が少ないと，異なる訪問傾向が 1 つの基底に混在してしまい，解釈が難しくなる．

NMF や NTF において基底を階層的に抽出するために，Factorial Hidden Markov Model (FHMM) [2] の概念を導入した可変基底 NMF [8] が音楽音響信号の解析を題材として提案されている．従来の NMF では，各基底に時間変化する係数を乗じてデータに混在する各成分を表現する．音楽音響信号解析において，各楽音のスペクトルがスケール（音量）以外に時間変化しないと仮定できれば，各分解成分は楽音のスペクトル系列に対応する．しかしながら，実際にはアタック，サスタイン，ディケイなどの「状態」の変化に伴って楽音のスペクトルの形状は時間変化する．そこで，可変基底 NMF では，各基底が状態遷移に応じて変化することを許容したモデルを採用している．これは，スペクトル系列を，複数の HMM からの出力系列の和でモデル化していることに相当するので，FHMM の一種と見なせる．

位置情報ログにおいても可変基底 NMF が適用できれば，1 週間周期で平日は仕事・休日は買物，24 時間周期で自宅→会社→食堂→会社→自宅などの階層的な周期を持つパターンを，それぞれ別基底で抽出可能になると期待される．本論文では位置情報ログの特性を考慮して可変基底 NMF を拡張し，位置情報ログにおける階層的パターンの抽出を実現する．位置情報ログへの適用で考慮すべき特徴として，欠損の多さと，曜日・時刻など絶対時間による周期性が挙げられる．

欠損の多さについて，可変基底 NMF の有効性が示されている音楽音響信号に比べ，位置情報ログは欠損が多い．例えば，位置情報履歴の中でもチェックインサービスのログなど，ユーザによる能動的な作業により訪問履歴が残される形式のログでは，実際にどこかを訪問していても，訪問が記録されない場合や，1 箇所に数時間滞在していても，到着時にしかログが残らない場合があり，

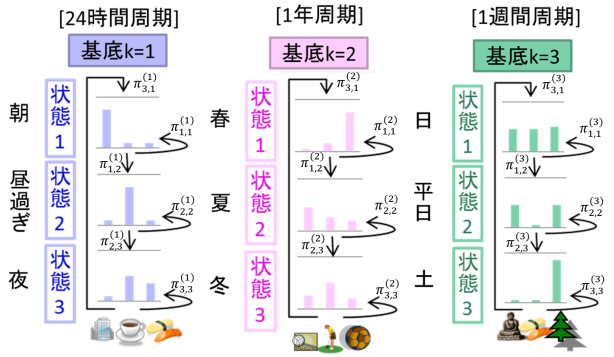

図 1 階層的な周期パターン抽出のイメージ

データが欠損している．ユーザの能動的な記録作業を必要としない GPS 等の移動軌跡ログにおいても，緯度経度の測定誤差や滞留点の検出精度によりデータの欠損が発生する．データの欠損が多いと，全ての場所への訪問確率が 0 となる有用性の低い頻度分布が多くの時刻で抽出される可能性や，一度だけ訪問した場所が抽出パターンに強く作用する可能性がある．提案手法では，パターン抽出時の繰り返し計算中に，更新前のパターンを用いた補間ステップを導入することにより，データの傾向を考慮した欠損値補間を行い，有用性の高いパターンの検出を促進する．位置情報ログの欠損に対応する方法として，同一時間帯の同ユーザの訪問傾向や，他ユーザの同場所への訪問傾向を使ってデータを補間しながら NTF を適用する手法が提案されている [7] が，同一時間帯や同一場所の訪問傾向には必ず類似性があるという強い仮定が入っており，異なる時間帯に起こる類似した傾向等の考慮や，同一時間帯でも複数の傾向が混在するなどの状況にも対応した補間は実現できていない．

周期性について，従来の可変基底 NMF では，各基底における状態推定時に，データの絶対時間軸上の周期性の明示的考慮はなく，1 週間周期や 24 時間周期が想定される位置情報データにおいては，想定される周期性と合致しない，解釈が難しいパターンが抽出される可能性がある．音響音楽信号の解析時に考慮した微妙なスペクトル形状の時間変化に比べ，暦や時間に応じた多様な周期に基づく訪問頻度分布の時間変化が想定される位置情報データでは，周期性の明示的な考慮が有効であると考えられる．提案手法では，1 週間周期，24 時間周期などの外部情報を与え，可変基底 NMF における状態推定に制約を与えることにより，絶対時間上の周期性を考慮できるように可変基底 NMF を拡張する．

図 1 に，提案手法によるパターン抽出イメージを示す．可変基底 NMF の枠組みを導入しているため，従来の NMF における各基底が，それぞれ独立な「状態」と「遷移確率」をもつ left-to-right の HMM に基づいて時間変化する．図中の $\pi_{i,j}^{(k)}$ は基底 k における状態 i から j への遷移確率を示す．ここで，「状態」は場所群への頻度分布を表す．例えば基底 $k=1$ の状態 1 は，会社（ビル）に行きやすい状態，状態 2 はカフェに行きやすい状態，状態 3 はカフェと寿司屋の両方に行きやすいが，ややカフェの方が確率が高い状態を示す．時間変化する状態の系列と係数行列との積が，前述の連続行動パターンを形成する．基底 1 では 24 時間周期，基底 2 では 1 年周期，基底 3 では 1 週間周期など様々な階層的周期性を持った訪問パターンの抽出を実現する．

本論文では，可変基底 NMF を位置情報ログの特性を考慮して拡張した提案手法の詳細を述べた後，実際のチェックインログを用いて提案手法が階層的な周期性を持つ行動パターンを推定できることを定性的に検証する．

2 関連研究

まず，パターン抽出のための既存手法である可変基底 NMF [8] の概要を説明する．可変基底 NMF は，状態遷移に応じて基底行列が時間変化することを許容した FHMM の一種であり，生成モデルとして次のように定式化できる．$H_{\omega,k,i} \sim \mathrm{Uniform}[0,\infty)$，$U_{k,t} \sim \mathrm{Uniform}[0,\infty)$，$z_t^k | z_{t-1}^k \sim \pi^{(k)}_{z_{t-1}^k, z_t^k}$，$C_{k,\omega,t} \sim \mathrm{Poisson}(C_{k,\omega,t}; H_{\omega,k,z_t^k} U_{k,t})$，$Y_{\omega,t} = \sum_{k=1}^K C_{k,\omega,t}$．$z_t^k$ は時刻 t における基底 k の状態を，K は基底数を表す．Poisson 分布は $\mathrm{Poisson}(y;x) = x^y e^{-x}/y!$ で定義される．ここで，観測行列 $\boldsymbol{Y}$ が与えられた下で未知変数 $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{Z}$ の事後確率

$$P(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}|\boldsymbol{Y}) \propto P(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z})P(\boldsymbol{H})P(\boldsymbol{U})P(\boldsymbol{Z})$$

の対数を最大にする $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{Z}$ を求める問題を考える．観測行列の確率密度関数の対数 $\log P(\boldsymbol{Y}|\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z})$ は，定数項を除き下記の I-divergence の符号を反転させたものと等しくなるため，可変基底 NMF $Y_{\omega,t} \simeq X_{\omega,t} = \sum_{k=1}^K H_{\omega,k,z_t^k} U_{k,t}$ の目的関数は以下のように書ける．

$$\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}) = \sum_{\omega,t} I(Y_{\omega,t}|X_{\omega,t}) - \log p(\boldsymbol{Z})$$

$$I(y|x) = y \log \frac{y}{x} - y + x$$

$$\log p(Z) = \sum_k \left(\log \pi_{z_1^k} + \sum_{t=2}^T \log \pi_{z_{t-1}^k, z_t^k} \right)$$

目的関数 $\mathcal{I}$ を最小化する $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{Z}$ を求めることにより，観測行列 $\boldsymbol{Y}$ を最もよく再現する $\boldsymbol{X}$ が得られる．$\mathcal{I}$ からは，解析的に $\boldsymbol{H}$, $\boldsymbol{U}$, $\boldsymbol{Z}$ の値を求めることはできないが，Jensen の不等式を用いて設定した上界関数

$$\mathcal{I}^+(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z},\boldsymbol{\lambda}) = \sum_{\omega,t} \Big(Y_{\omega,t} \log Y_{\omega,t} - Y_{\omega,t} \sum_k \lambda_{\omega,k,t} \log \frac{H_{\omega,k} U_{k,t}}{\lambda_{\omega,k,t}} - Y_{\omega,t} + X_{\omega,t} \Big) - \log p(\boldsymbol{Z})$$

を最小化することにより，間接的に $\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{Z}$ の値を最適化する補助関数法が有効である．紙面の都合上証明は省略するが，$\mathcal{I}^+$ を $\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{Z}$ に関してそれぞれ最小化するステップと，$\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}) = \mathcal{I}^+(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z},\boldsymbol{\lambda})$ となるように $\boldsymbol{\lambda}$ を更新するステップを繰り返すことで，$\mathcal{I}$ を単調に降下させることができる．$\mathcal{I} = \mathcal{I}^+$ となる $\boldsymbol{\lambda}$ は $\lambda_{\omega,k.t} = \frac{H_{\omega,k,z_t^k} U_{k,t}}{\sum_k H_{\omega,k,z_t^k} U_{k,t}}$ である．

観測行列 $Y_{\omega,t}$ に対して，以下のような手順で可変基底 NMF を適用することができる．

(i) NMF の結果の $H_{\omega,k}, U_{k,t}$ を使い，$H_{\omega,k,z_t^k}, U_{k,t}$ の初期値を求める．

(ii) 補助変数 $\boldsymbol{\lambda}$ の初期値を計算する．

(iii) Viterbi Algorithm により z_t^k を更新する．

(iv) H_{ω,k,z_t^k} と $U_{k,t}$ を更新する．

(v) 収束するまで **(iii)** と **(iv)** を繰り返す．

(vi) $\mathcal{I} = \mathcal{I}^+$ となるよう補助変数 $\boldsymbol{\lambda}$ の値を更新する．

(vii) **(iii)** から **(vi)** を収束するまで繰り返す．

(viii) 目的関数 $\mathcal{I}$ の値が十分に小さくなれば，各基底の状態数を順に増加していく．

(i) の初期値を作成する際に，各基底内で類似した状態が設定されるようにするために，$H_{\omega,k,1}$ には $H_{\omega,k}$ の値を入れ，$H_{\omega,k,i}$ (i は $1 \leq i \leq N_k$ を満たす整数) には $H_{\omega,k}$ をランダムに微小変化させた値を入れるなどの方法が考えられる．ここで N_k は基底 k の状態数を表し，N_k の初期値は 2 とする．

(iii) において，観測行列 $Y_{\omega,t}$ に対して $\mathcal{I}^+$ を最小にする最適状態系列 $\{z_t^k\}_{t=1}^T$ を k ごとに動的計画法により効率的に解くことができる．T は時刻 t の最大値である．まず，$\mathcal{J}_{k,t}^+(s) := \sum_\omega (-Y_{\omega,t}\lambda_{\omega,k,t} \log \frac{H_{\omega,k,s}U_{k,t}}{\lambda_{\omega,k,t}} + H_{\omega,k,s}U_{k,t})$ と置き，$\delta_1^k(s) = \mathcal{J}_{k,1}^+(s) - \log \pi_s^{(k)}$ とする．次に $t = 2, ..., T$ について，逐次的に $\delta_t^k(s)$ を $\delta_t^k(s) = \min_{s'}[\delta_{t-1}^k(s') + \mathcal{J}_{k,t}^+(s) - \log \pi_{s',s}^{(k)}]$ により計算していくことができる．各ステップで選択される状態番号を $\psi_t^k(s) = \arg\min_{s'}[\delta_{t-1}^k(s') - \log \pi_{s',s}^{(k)}]$ に記憶しておくことで，$t = T$ まで到達後に $z_{t-1}^k = \psi_t^k(z_t^k)(t = T, \ldots, 2)$ により選択された状態番号を辿っていくことができ，最適経路 $z_1^k, \ldots, z_T^k$ を得ることができる．この際，推定された状態系列における遷移確率を元に $\pi_{i,j}^{(k)}$ を更新する．

上記手順の収束後，基底行列 H_{ω,k,z_t^k} では，基底 k の，各状態 z_t^k における各場所 ω に対する訪問確率分布が表現される．係数行列 $U_{k,t}$ では，各時刻 t 毎にどの基底がアクティベートされるか（有効になっているか）が表現される．例えば，時刻 $t = 3$ において，$k = 1$ のとき $U_{1,3}$ が 1 で，k が他の値の時 $U_{k,3}$ が全て 0 になっていた場合，$H_{\omega,k,z_t^k}U_{k,t}$ の分布は $H_{\omega,1,z_3^1}U_{1,3} = H_{\omega,1,z_3^1}$ となる．ここで z_3^1 は時刻 $t = 3$ における基底 $k = 1$ の状態を示す．同一時刻において，各基底の状態は 1 つの状態に定まるものとする．

3 提案手法

3.1 観測行列の作成

本稿ではユーザ ID，年月日時刻，場所の情報を含むチェックインログを想定する．特定のユーザの位置情報履歴に対して可変基底 NMF を適用する場合の前処理として，場所 $\omega\times$ 時刻 t の観測行列 $Y_{\omega,t}$ を作成する．時刻は日付単位，時間帯単位など任意の単位時間で区切り，単位時間内に各場所へのチェックイン履歴があれば $Y_{\omega,t} = 1$，ない場合には $Y_{\omega,t} = 0$ とする．

3.2 補間ステップの導入による欠損の考慮

チェックインはユーザの能動的操作により記録されるため，$Y_{\omega,t} = 0$ となっていても，その場所への滞在や他の場所への到着・滞在があった可能性がある．本稿では $Y_{\omega,t} = 0$ となっている箇所を欠損と呼ぶ．$Y_{\omega,t}$ における非ゼロ要素率は手元のデータで 1%を下回り，欠損が多い．提案手法では，可変基底 NMF で計算を繰り返していく際に，抽出されたパターンを考慮して欠損箇所を補間するステップを導入する．

入力データにおいて 0 以外の値が入っている箇所を Γ とする．提案手法では，入力データで 0 以外の値が入っている箇所では $Y_{\omega,t}$ と $X_{\omega,t}$ が近づくように，0 の値が入っている欠損箇所では $S_{\omega,t}$ すなわち更新前の $X_{\omega,t}$ の値と $X_{\omega,t}$ が近づくように，最小化させる目的関数を変更する．欠損箇所の補間を行うと，補間箇所に強い影響を受けて，各時刻 t におけるアクティベーション $U_{k,t}$ が多くの基底 k で類似した値になってしまう可能性がある．そこで，基底 k 毎の成分が独立になるように誘導するため，$U_{k,t}$ をスパース化させる．ここでは，$U_{k,t}$ の値が 0 に一定程度近づいたら強制的に 0 にするイメージで目的関数に $|U_{k,t}|^P$ を最小化させる項を加える．

上記の議論から提案法の目的関数を定式化する．まず，欠損していない箇所だけを考えると，提案手法による目的関数は以下のように表現される．

$$\begin{aligned}&\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z})\\ &= \sum_{(\omega,t)\in\Gamma} I(Y_{\omega,t}|X_{\omega,t}) + w_p \sum_{k,t} |U_{k,t}|^p - \log P(Z)\end{aligned}$$

しかしながら，2 章のように $\boldsymbol{Z}$ の更新を Viterbi アルゴリズムにより行うためには，全ての (ω, t) において観測値が必要となる．そこで，目的関数に含まれる ω, t が全範囲になるように，欠損領域における仮想データを表す $\boldsymbol{S}$（未知変数）を導入し，以下の補助関数 $\mathcal{I}$‘ を考える．

$$\mathcal{I}'(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z},\boldsymbol{S}) = \mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}) + w_s \sum_{(\omega,t)\notin\Gamma} I(S_{\omega,t}|X_{\omega,t})$$

$I(x|y)$ は $x = y$ のときにのみ 0 となる非負値関数であるため，$\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}) = \min_S \mathcal{I}'(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z},\boldsymbol{S})$ となり，$\mathcal{I}'$ が $\mathcal{I}$ の補助関数としての要件を満たす．従って，$\mathcal{I}'$ を最小化するように $\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}$ を更新するステップと，$\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z}) = \mathcal{I}'(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z},\boldsymbol{S})$ となるように $\boldsymbol{S}$ を更新するステップ（すなわち $\boldsymbol{S} = \boldsymbol{X}$）を繰り返せば，$\mathcal{I}(\boldsymbol{H},\boldsymbol{U},\boldsymbol{Z})$ を単調減少させることができる．この際，$\boldsymbol{S}$ を欠損箇所における観測データとみなして，2 章と同様の方法で $\mathcal{I}'$ を最小化できる．なお，$\mathcal{I}'$ において，w_s は補間箇所における I-divergence を考慮する重みを示す定数である．補間された値に $X_{\omega,t}$ が近づき過ぎないように重みづけを行う．w_s の値として，例えば非ゼロ要素の割合を利用できる．w_p は全体の目的関数に対して，どの程度の重みでアクティベーションのスパース化を考慮するかを示す定数である．P は $0 < P < 2$ を満たす定数であり，0 に近いほど強い制約となる．

また，$U_{k,t}$ をスパース化させるための関数 $|U_{k,t}|^P$ の上界関数として，$|U_{k,t}|^P$ に接する放物線 $\frac{1}{2}|\nu_{k,t}|^{P-2}U_{k,t}^2$ を考えて $\mathcal{I}'^{+}$ を設計する．ここで補助変数 $\nu_{k,t}$ が，更新前の $U_{k,t}$ の値となる時に等号が成立する．

これまでの拡張により，提案する欠損と周期性を考慮した可変基底 NMF を適用する際の処理 **(iv)** における $\boldsymbol{H}$，$\boldsymbol{U}$ の更新式は以下の通りになる．

$$H_{\omega,k,j} \leftarrow \frac{\sum_t \mathbf{1}[z_t = j]\ \lambda_{\omega,k,t} A_{\omega,t}}{\sum_t \mathbf{1}[z_t = j]\ B_{k,t}} \quad (1)$$

$$U_{k,t} \leftarrow \frac{\sum_\omega \mathbf{1}[\omega, t \in \Gamma]\left\{-H_{\omega,k,z_t^k} + \sqrt{C_{\omega,k,t}}\right\}}{\sum_\omega 2\lambda P|\overline{U_{k,t}}|^{p-2}} + \frac{\sum_\omega w_s \mathbf{1}[\omega, t \notin \Gamma]\left\{-H_{\omega,k,z_t^k} + \sqrt{D_{\omega,k,t}}\right\}}{\sum_\omega 2\lambda P|\overline{U_{k,t}}|^{p-2}} \quad (2)$$

$$A_{\omega,t} = \{\mathbf{1}[\omega, t \in \Gamma]\ Y_{\omega,t} + w_s\ \mathbf{1}[\omega, t \notin \Gamma]\ S_{\omega,t}\}$$

$$B_{k,t} = \{\mathbf{1}[\omega, t \in \Gamma]\ U_{k,t} + w_s\ \mathbf{1}[\omega, t \notin \Gamma]\ U_{k,t}\}$$

$$C_{\omega,k,t} = H_{\omega,k,z_t^k}^2 + 4\lambda P|\nu_{k,t}|^{p-2} Y_{\omega,t}\lambda_{\omega,k,t}$$

$$D_{\omega,k,t} = H_{\omega,k,z_t^k}^2 + 4\lambda P|\nu_{k,t}|^{p-2} S_{\omega,t}\lambda_{\omega,k,t}$$

ここで，$\mathbf{1}[\cdot]$ は，$[\cdot]$ 内の条件を満たす場合には 1，そうでない場合には 0 を取る．

3.3 制約付状態推定による周期性の考慮

位置情報ログでは，行動パターンに時間帯，曜日，月，季節などに応じた周期性が想定される．従来の可変基底 NMF でも，NMF の各基底に left-to-right の HMM を仮定し，周期的パターンの抽出を試みたが，時間帯や曜日など外部情報から得られる絶対時間の考慮は行わず，抽出パターンを絶対時間周期と対応付けて解釈できるとは限らなかった．提案手法では，絶対時間軸上の周期性を考慮することにより，従来よりも解釈が容易で，情報配信などへの応用にも効果的なパターンの抽出を試みる．

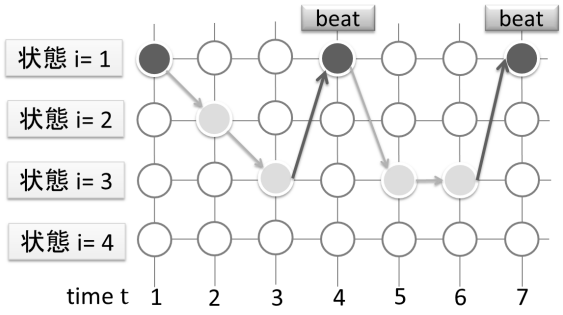

図 2 周期性制約のイメージ

提案手法では，各基底に異なる絶対時間周期を仮定した上で，HMM の状態推定時に，各周期の最初は特定の状態に固定されるよう制約を与える．周期性による制約のイメージを図 2 に示す．基底毎に異なる周期の周期性を仮定し，各周期の開始時には状態 i が $i = 0$ に固定されるようにする．すなわち，Viterbi Algorithm による状態推定時に，$t \in \{beat_k\}$ かつ $i \neq 0$ ならば $\delta_t^k(s) = \infty$ とする．$\{beat_k\}$ は基底 k に与えた周期の開始時刻 t の集合を示し，図 2 の例では，$\{beat_k\} = \{4, 7\}$ である．

各基底に対する周期は，欠損値補間の拡張のみを実施して，アクティベーションのスパース化と周期性の考慮は実施せずに拡張可変基底 NMF 適用を適用した際に，各基底のアクティベーションの時間変化が季節，曜日，時間帯周期を持つ可能性が高そうかどうかを元に決定する．例えば，毎日複数回アクティベーションが高い場合には，時間帯変化を表現するための 24 時間周期を与える．また，1 日 1 回または特定の曜日にアクティベーションが高い場合には，曜日変化を表現するために 1 週間周期を与える．同様にして，特定の季節のみでアクティベーションが高い場合には，季節変化を表現するために，1 年の周期を与える．この時，全ての想定周期に対する集計値が不均一な場合には，周期性を考慮しない．この周期は，周期性を考慮した可変基底 NMF を適用しながら，再度変更していくことも考えられる．決定した周期を用いて拡張可変基底 NMF を適用した際に，特定の状態のみでしかアクティベーションが高くならない場合には，周期を 1 週間周期から 24 時間周期など，より小さいものに変更する．一方，アクティベーションが高くなる状態の偏りが少ない場合は，周期の変更を停止する．また，周期が最小となっても状態に偏りが生じた場合は，その基底への周期性付与は行わない．周期の最小値については，データ期間に応じて 1 時間周期，6 時間周期などに決定する．

3.4 拡張可変基底 NMF の処理フロー

提案手法による拡張可変基底 NMF の処理の流れは以下のようになる．

(I) 時刻 t・場所 ω 毎の訪問有無を集計し，$Y_{\omega,t}$ を作成する．

(II) $Y_{\omega,t}$ に NMF を適用し，基底 k 毎の場所分布 $H_{\omega,k}$ とアクティベーション $U_{k,t}$ を出力する．

(III) 周期性の考慮とアクティベーション $U_{k.t}$ のスパース化を行わずに，欠損値補間の拡張のみを実施し，アクティベーション $U_{k,t}$ の出力結果を用いて各基底に与える周期を決定する．

(IV) **(III)** で決定した周期を用いて，すべての拡張部分を採用した可変基底 NMF を適用する．

(V) **(IV)** で出力された各基底 k の各状態毎の場所分布 H_{ω,k,z_t^k} とアクティベーション $U_{k,t}$ を描画しながら検出されたパターンの観察を行い，行動の分析を実施する．

なお，**(IV)** では，2 章に記載した可変基底 NMF の適用手順のうち，**(iii)** の Viterbi Algorithm で用いる $\mathcal{J}_{k,t}^{+}(s)$ を補助関数 $\mathcal{I}'^{+}$ に対応するよう変更した上で，$\delta_t^k(s)$ の定義も 3.3 章で論じたように変更する．また，**(iv)** の $H_{\omega,k,z_t^k}, U_{k,t}$ の更新式を式 (1) と (2) に変更する．なお，与える周期について，**(V)** でアクティベーションが高くなる状態に偏りが生じている場合には，周期をより短い周期に変更して **(IV)** に戻ることも考えられる．

4 実験結果

実際の位置情報データを用いて，提案手法を検証する．検証実験については，Stanford 大学が以下で公開している Gowalla[1]のチェックインデータを利用する．チェックインデータでは，ユーザの能動的操作により，訪問履歴が記録される．基本的に，各地点に到着した際のみに訪問履歴の記録が実施され，各地点から離れる際の記録は行われないことが多い．チェックインデータでは，各訪問に対してユーザ ID と，日付時刻，緯度経度に加え，同一の緯度経度に割り振られる場所 ID が格納されている．

ここでは，ある特定のユーザのデータ (102 日分）のみを抽出して検証実験を行った．場所 ω として場所 ID を用いた．時刻 t については，データに記載された時刻を 0 時台，1 時台など 1 時間毎の時間窓に分割した上で，最も古いチェックインログが記録された時間窓から数えた時間窓数を用いた．当該データを集計して $Y_{\omega,t}$ を作成したところ，非ゼロ要素数は少なく，非ゼロ要素率は 0.269%であった．

図 3 に，初回の可変基底 NMF 適用時（欠損値補間は行うがアクティベーション $U_{k,t}$ のスパース化と周期性の考慮は未適用）の，アクティベーション $U_{k,t}$ と，決定した初期付与周期を示す．基底数は $K = 5$ とした．各折れ線グラフの横軸は時間窓 t を示し，縦軸は $U_{k,t}$ の

[1]https://snap.stanford.edu/data/loc-gowalla.html

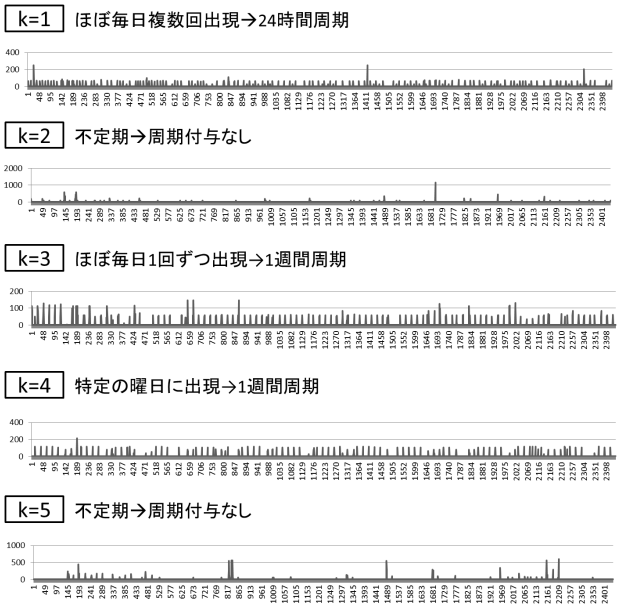

図 3 初期の $U_{k,t}$ と付与する周期の検討例

大きさを示す．ほぼ毎日複数回のパターンが出現している $k = 1$ では，1 日の中での行動パターンの把握を行うために 24 時間の周期を与えた．また，ほぼ毎日，1 日 1 回ずつアクティベーションが高くなっている $k = 3$ では，曜日によって異なる行動パターンの把握を目的として 1 週間の周期を与えた．同様に，特定の曜日（主に平日）のみにパターンが出現している $k = 4$ についても，各曜日で異なるパターンの抽出を試みるために 1 週間の周期を与えた．アクティベーションが不定期に高くなっている基底 $k = 2$ と $k = 5$ については，周期性の付与は行わなかった．

周期性付与の効果を検証するために，図 4 に周期性の付与を行った場合と行わなかった場合のアクティベーション $U_{k,t}$ の出力をそれぞれ示す．各ヒートマップの横軸は時間窓 t を示し，縦軸は各状態を示す．$k = 1$ については，24 時間周期の周期性の付与を行うと，ほぼ 1 つの状態が選択され続けていることが分かる．従って，基底 $k = 1$ には単一の行動パターンのみしか存在しないと考えられる．周期を与えないと，状態 1，2 の複数の状態が出現しているが，2 つの状態の H_{ω,k,z_t^k} を比較すると，パターンが酷似しており，周期性の導入により類似した状態を統合することができたと捉えることができる．$k = 3$ については，周期性の考慮により，状態 1 と状態 2，状態 2 と状態 3 を短期間に含むパターンが複数回抽出された．以上のように，提案手法では，例えば 1 週間周期と 24 時間周期など異なる階層的な周期を持つ繰り返しパターンを抽出できることが分かった．

続いて，欠損値補間の効果を確認するために，図 5 に基底 $k = 3$ における各状態の H_{ω,k,z_t^k} の出力結果を示す．

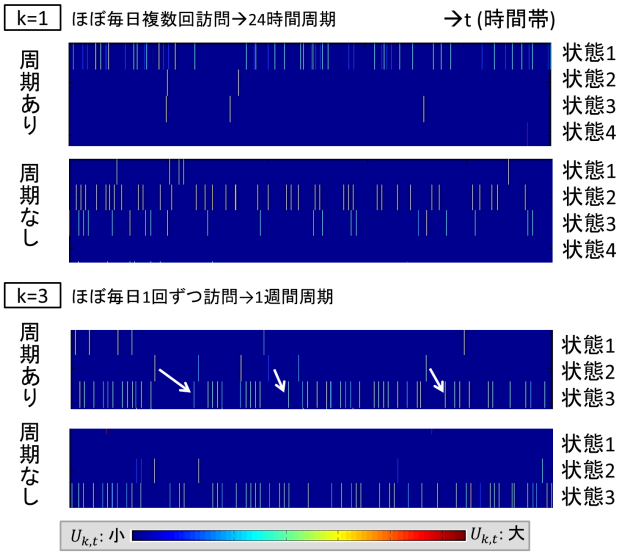

図 4 従来法と拡張可変基底 NMF の $U_{k,t}$ の出力比較

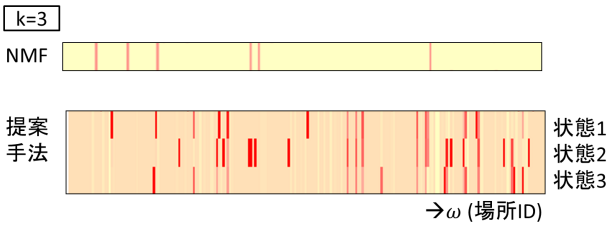

図 5 NMF と拡張可変基底 NMF の H_{ω,t,z_t^k} の出力比較

上は NMF によって抽出された基底 $k=3$ の分布である．欠損値補間を行わない NMF の結果に比べ，拡張可変基底 NMF の結果では訪問可能性がある場所が薄く検出されている．欠損箇所の補間を行わずに可変基底 NMF を適用すると，状態が遷移しない場合や，すべての場所への訪問可能性が 0 の分布が抽出される場合がある．欠損箇所の補間を行った上で，同時に複数の基底がアクティベートされず，基底毎の成分が独立になるように，アクティベーションのスパース化を行うことにより，図 5 の下部に示されているような行動パターンを抽出することが可能になった．加えて，薄く色がついている訪問場所には，実際に訪問した場所だけでなく，訪問可能性が高い場所群が表現されており，各状態が，場所群への「訪問気分」に該当すると期待される．

抽出されたパターンの中身を分析してみると，状態 1 には非習慣的に訪問される店や住宅地が，状態 2 には特定地域の娯楽施設やカジュアルな飲食店が，状態 3 には状態 2 と同じ地域のバーや居酒屋と隣接県の駅や住宅街が分類されていることが分かった．状態 2 には主に昼間に訪問する場所群が，状態 3 には自宅を含め，夜間によく訪問される場所群が分類されていると捉えることができ，状態 2 から状態 3 へと続く行動パターンが特定の曜日に複数回に渡って起こることを検出できた．加えて，図 4 の $k=3$ の基底を見ると，状態 1 から状態 3 への遷移よりも，状態 2 から状態 3 への遷移の方が間に長い時間が空いていることが分かった．

5 まとめ

本稿では，可変基底 NMF を位置情報ログの欠損や周期性を考慮して拡張し，位置情報ログに周期的に出現する，階層的な周期をもつ行動パターンを抽出する手法を提案した．提案手法では，欠損の多いデータにおいても，各時刻で訪問可能性の高い場所群を予測できると期待される．また，場所に応じた訪問周期，滞在時間，次に訪問する場所の傾向など，複数のユーザに共通する行動のパターンを把握して，都市計画などに役立てたり，ユーザ毎に特有な行動パターンを抽出することにより，情報配信のタイミングや内容に関する満足度を向上させたりといった効果が期待される．今後は，より大規模なデータへの適用を試み，より柔軟かつ多様な周期性を持つ階層的な行動パターンの抽出を試みたい．また，各行動が習慣に沿っているかどうかを数値化する習慣度算出手法 [4] や，NMF において $\boldsymbol{U}$ に周期性を仮定して，その周期を自動的に推定する手法 [3] との融合を行いたい．

参考文献

[1] Z. Fan, X. Song, and R. Shibasaki. Cityspectrum: a non-negative tensor factorization approach. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 213–223. ACM, 2014.

[2] Z. Ghahramani and M. I. Jordan. Factorial hidden markov models. *Machine learning*, 29(2-3):245–273, 1997.

[3] A. Hayashi, H. Kameoka, T. Matsubayashi, and H. Sawada. Non-negative periodic component analysis for music source separation. In *Acoustics, Speech and Signal Processing (ICASSP), 2015 IEEE International Conference on*. IEEE, 2015.

[4] A. Hayashi, T. Matsubayashi, and H. Sawada. Regular behavior measure for location based services. In *Proceedings of the 2014 ACM conference on Web science*, pages 299–300. ACM, 2014.

[5] D. D. Lee and H. S. Seung. Algorithms for non-negative matrix factorization. In *Advances in neural information processing systems*, pages 556–562, 2001.

[6] A. Shashua and T. Hazan. Non-negative tensor factorization with applications to statistics and computer vision. In *Proceedings of the 22nd international conference on Machine learning*, pages 792–799. ACM, 2005.

[7] Y. Wang, Y. Zheng, and Y. Xue. Travel time estimation of a path using sparse trajectories. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 25–34. ACM, 2014.

[8] 中野允裕, 北野佑, ルルージョナトン, 亀岡弘和, 小野順貴, 嵯峨山茂樹. 可変基底 NMF に基づく音楽音響信号の解析. **研究報告音楽情報科学**, 2010(10):1–6, 2010.