年齢別人口移動NMFモデルのベイズ推定

元井　初音[1]・奥村　誠[2]・水谷　大二郎[3]

[1]非会員　東北大学　大学院工学研究科土木工学専攻（〒980-8579 仙台市青葉区荒巻字青葉6-6）
E-mail:hatsune.motoi.t7@dc.tohoku.ac.jp

[2]正会員　東北大学教授　災害科学国際研究所（〒980-0845 仙台市青葉区荒巻字青葉468-1）
E-mail:makoto.okumura.b6@tohoku.ac.jp

[3]正会員　東北大学助教　災害科学国際研究所（〒980-0845 仙台市青葉区荒巻字青葉468-1）
E-mail:mizutani@irides.tohoku.ac.jp

人々の転出入選択の多くは人生の節目に当たる就職や進学，退職と言ったライフイベントによるものである．非負値行列因子分解（NMF; Non-negative Matrix Factorization）法を非負値データである年齢階層別転出入データに適用することにより，複数のライフイベントの発生年齢分布と地域におけるライフイベント発生率をそれぞれ非負の形で計算できる．このNMFモデルに対するベイズ推定により，得られる行列の意味の明確化と信頼性の向上が期待できる．本稿は都道府県間のデータに関する計算例に基づきベイズ推定による効果を考察する．

Key Words : *demography, non-negative matrix factorization, Bayesian statistics, life events, immigrants and emigrants*

1. はじめに

地域社会の年齢別人口移動が変化すると，公共サービスや施設への新しいニーズが発生するなどし，公共が提供する必要があるものに変化が起こる．また，今後の縮小社会の中で経済状況が変化することで公共が住民に提供できる質や量が限られたときに，どこを取捨選択するかを決めなくてはならない場合が有り得る．そこで，地域社会の変化や現状について，他地域や過去に観測された変化や現状との類似性や共通性を部分的に把握できれば，その後の推移の傾向や発生しうる社会的な問題とその対応策，重要となりそうな公共サービスのニーズを考える際に有用だと考えられる．

地域の変化は居住者の転居に影響を与えることが多く，変化の原因によって異なる年齢階層に影響する．また，年齢階層ごとに人々の公共サービスに対するニーズも異なる．したがって，年齢構成比を通して地域社会の特徴を観察し，他地域や過去の年齢構成と比較の上で把握することでその地域で起こっている変動の原因を推測するとともに，今後発生する公共サービスのニーズを把握できる可能性がある．特に，転出と転入を差し引きした純転入ではなく，転出，転入のそれぞれを別に分析することでその地域内の人口の年齢構成比の入れ替わりを議論することができる．更に残留者数にも考慮しながら分析を行うことで，各地域の人口規模の差による結果のズレを最小限に抑えることができる．

多くの観測データの中に共通する変動パターンを，少数の要因の加算として把握する手法としてパターン認識が挙げられる．中でも，すべての要因が非負の要素を持ち，その加算の程度も非負の数値として把握する分析手法として，非負値行列因子分解（Non-negative Matrix Factorization; NMF）がある．地域の転出，転入，残留者数は，測定誤差や記録時の誤差を考えなければ必ず非負の値を取ることから，筆者らはNMFに基づいた人口移動特性の把握のための方法論を提案し，解釈が可能な複数の要因を抽出して地域別の分布の違いを明らかにした[1]．その一方で，各要因が多くの年齢階層に正の値をもち，明瞭な解釈が難しいこと，計算の収束状況が不安定で，計算をどこで打ち切るべきかの基準がないことなどの問題点が残された．

本研究では，ベイズ統計学の考え方を取り入れたNMF法[2]を用いてこれらの問題点の解決を図ることを試みる．先行研究と同じデータに対して分析を行い，結果を比較して考察を行った．

2. 分析手法

NMFについては数多くの既往研究や成書が存在し[2]，近年土木計画学における利用事例が増加して来ている．以下一般的なNMFの概要について説明したあとに，ベイズ統計学の考え方を用いたNMF推定法[2]を説明する．

(1) 非負値行列因子分解

主成分分析や因子分析など，与えられたデータを複数の成分に分解する手法は様々あるが，特に NMF は非負値のデータを対象とした多変量解析手法である [2]．現実世界に存在するデータの多くは非負値で表されるため，NMF によって合奏曲や顔画像，文書内の単語の出現回数といった非負値のデータの構成成分を抽出し，音源分離や顔画像処理，文書の自動分類をする研究が進められている．

ここで，分析したいデータをサイズ $K\times N$ の観測行列 X とする．観測行列 X を

$$\boldsymbol{V}\cdot\boldsymbol{W}+\boldsymbol{E}\simeq\boldsymbol{X} \tag{1}$$

を満たすようなサイズ $K\times M$ の因子得点の行列 $\boldsymbol{V}$ とサイズ $M\times N$ の因子負荷量の行列 $\boldsymbol{W}$ とサイズ $K\times N$ 残差行列 $\boldsymbol{E}$ に分解する．なお，分解後の因子数 M は解析する者が事前に決めておき，

$$M\le\min(K,N) \tag{2}$$

を満たす必要がある．一般的に M の数が多ければ多いほど残差行列 $\boldsymbol{E}$ が小さくなり，内積 $\boldsymbol{V}\cdot\boldsymbol{W}$ は観測行列 $\boldsymbol{X}$ に近づいていく．

(2) ベイズ推定による非負値行列因子分解

Schmidt et al.[4]では，NMF にベイズ統計学の考え方を取り入れることで，得られる結果の不確実性を考慮したモデル評価が可能になり，説明力が向上することが示されている．以下ベイズ統計学の考え方を用いた NMF の手順を説明しておく．

まず，残差行列 $\boldsymbol{E}$ の生起確率密度を尤度関数とし，行列 $\boldsymbol{V}$，$\boldsymbol{W}$ のそれぞれの行列値について事前分布を仮定する．尤度関数や事前分布は，データの発生の仕方などを考慮し，適切なものを設定する．ここでは，人口移動データの性質に加え，サンプリングの簡便性も考慮し，事後分布が調整済み正規分布となるよう，尤度関数と事前分布の組合せとして正規分布と指数分布を設定した．具体的には，まず，誤差行列 $\boldsymbol{E}$ の個々の要素の生起確率密度を平均 0，分散 σ^2 を持つ正規分布と仮定して，

$$p(\boldsymbol{X}|\boldsymbol{V},\boldsymbol{W},\sigma^2)=\prod_{k,n}\mathcal{N}\left(\boldsymbol{X}_{k,n};(\boldsymbol{VW})_{k,n},\sigma^2\right) \tag{3}$$

のように尤度関数を設定する．また，行列 $\boldsymbol{V}$，$\boldsymbol{W}$ のそれぞれの行列値については以下の事前分布を仮定する．

$$\begin{aligned}p(\boldsymbol{V})&=\prod_{k,m}\mathcal{E}\left(\boldsymbol{V}_{k,m};\varphi_{k,m}\right)\\ p(\boldsymbol{W})&=\prod_{m,n}\mathcal{E}\left(\boldsymbol{W}_{m,n};\omega_{m,n}\right)\end{aligned} \tag{4}$$

ここで，

$$\mathcal{N}(x;\mu,\sigma^2)=\frac{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}{\sqrt{2\pi\sigma^2}} \tag{5}$$

は正規分布の確率密度関数であり，

$$\mathcal{E}(x;\lambda)=\lambda e^{-\lambda x}\cdot u(x) \tag{6}$$

は指数分布の確率密度関数を表す．なお，

$$u(x)=\begin{cases}0(x<0)\\1(x\ge0)\end{cases} \tag{7}$$

は，単位ステップ関数であり，非負条件を満たすために設定する．非負の値を持つ x が表れた場合，0 を返す．

また，残差行列の分散 σ^2 は，形状パラメータ k と規模パラメータ θ によって記述される逆ガンマ分布に従って発生すると仮定し，以下のように事前分布を設定する．

$$\begin{aligned}&p(\sigma^2)\\&=\mathcal{G}^{-1}(\sigma^2;k,\theta)=\frac{\theta^k}{\Gamma(k)}(\sigma^2)^{-k-1}\exp\left(-\frac{\theta}{\sigma^2}\right)\end{aligned} \tag{8}$$

以上の設定により，行列 $\boldsymbol{V}$ および $\boldsymbol{W}$ の事後分布が得られる．事後分布からのサンプリングを行うためにマルコフ連鎖モンテカルロ法を用いる．

(3) ギブスサンプリングのための条件付確率設定

本研究では，マルコフ連鎖モンテカルロ法のうちギブスサンプリングを用いて行列 $\boldsymbol{V}$，$\boldsymbol{W}$ のサンプリングを行う．(2)で設定した尤度関数と事前分布を用いて，ベイズの定理に基づき，パラメータである行列 $\boldsymbol{V}$，$\boldsymbol{W}$ の事後分布を定義する．

正規分布と指数分布を掛け合わせたものに比例した，調整済み正規分布を

$$\mathcal{R}(x;\mu,\sigma^2,\lambda)\propto\mathcal{N}(x;\mu,\sigma^2)\mathcal{E}(x;\lambda) \tag{9}$$

と定義する．行列 $\boldsymbol{V}$ のある要素の事後確率確率密度関数は調整済み正規分布を用いて式(10)のように表すことができる．

$$\begin{aligned}&p\left(\boldsymbol{V}_{k,m}|\boldsymbol{X},\boldsymbol{V}_{\setminus(k,m)},\boldsymbol{W},\sigma^2\right)\\&=\mathcal{R}\left(\boldsymbol{V}_{k,m};\mu_{\boldsymbol{V}_{k,m}},\sigma^2{}_{\boldsymbol{V}_{k,m}},\varphi_{k,m}\right)\end{aligned} \tag{10}$$

$$\mu_{\boldsymbol{V}_{k,m}}=\frac{\sum_n\left(\boldsymbol{X}_{k,n}-\sum_{m'\neq m}\boldsymbol{V}_{k,m'}\boldsymbol{W}_{m',n}\right)\boldsymbol{W}_{m,n}}{\sum_n\boldsymbol{W}_{m,n}^2} \tag{11}$$

$$\sigma^2_{\boldsymbol{V}_{k,m}}=\frac{\sigma^2}{\sum_n\boldsymbol{W}_{m,n}^2} \tag{12}$$

$$p(\sigma^2|\boldsymbol{X},\boldsymbol{V},\boldsymbol{W})=\mathcal{G}^{-1}(\sigma^2;k_{\sigma^2},\theta_{\sigma^2}) \tag{13}$$

$$k_{\sigma^2} = \frac{KN}{2} + 1 + k \tag{14}$$

$$\theta_{\sigma^2} = \frac{1}{2}\sum_{i,j}(\boldsymbol{X} - \boldsymbol{VW})^2_{k,n} + \theta \tag{15}$$

なお，記号$\boldsymbol{V}_{\backslash(k,m)}$は行列$\boldsymbol{V}$の$k$行$m$列以外の要素全てを指す．対称性を考慮し，$\boldsymbol{W}_{m,n}$についても同様に，$\boldsymbol{V}_{k,m}$と対称となるように調整済み正規分布を設定できる．また，分散σ^2の事前分布は式(11)のようにパラメータ$k_{\sigma^2}, \theta_{\sigma^2}$を持つ逆ガンマ分布を仮定する．調整済み正規分布の正規化定数が解析的に求まらないため，パラメータの事後確率密度を解析的に求めることができない．そのため，ギブスサンプリングを用いた繰り返し計算により，パラメータの事後分布からのサンプルを獲得する．得られたパラメータのサンプルを用いて，事後分布の統計量を算出することができ，ここでは，サンプルの期待値をパラメータの推定値として用いる．事後分布からのサンプルを用いることにより，パラメータの不確実性を考慮しながらモデル評価を行うことができ，NMFの因子数の選定にも有用となる．また，NMFのパラメータの事後分布が多峰性を持つ場合においても，ギブスサンプリングで十分な数のサンプリングを繰り返すことにより，パラメータの定義域を大域的に探索でき，局所解への収束を回避できる可能性がある．

3. 分析対象データと解釈の方法

(1) 分析対象データ

先行研究での分析結果との比較を容易にするため，先行研究と同じデータを分析する．すなわち，転出，転入，残留の実態を表している国勢調査を採用した[1]．国勢調査では西暦の末尾が0の年に大規模調査，末尾が5の年に簡易調査を行うが，2015年度調査では東日本大震災の影響を把握するため従来大規模調査で調査されていた「5年前の住居の所在地」項目が加わっている[5]．

分析対象期間は，15年であればライフイベントが起こる年齢が時代によって変化する影響が大きくないと考え，1995年から震災前後年までの期間である1995-2000年，2005-2010年，2010-2015年の3期間とした．

対象データは**表-1**のように，5歳階級の年齢階層別，性別の人口移動数を，各都道府県の転出，転入，残留ごとに各期間ごと並べた形式を採用している．ここで「転出」は5年前に当該都道府県に居住していたが調査時点では都道府県外に居住地を移していた人数である．また「転入」は調査時点で当該都道府県に居住する者のうち5年前には当該都道府県外に居住していた人数である．更に「残留」は調査時点と5年前とで同一の都道府県に居住している人数であるが，住所を変えなかった者だけでなく当該都道府県内で転居した人数の合計である．なおこれらの定義により，調査時点で5歳未満の者や調査時点より5年前までの期間に死亡した者は含まれない．

(2) 結果の標準化

結果を標準化するために，先行研究と同様因子負荷量行列$\boldsymbol{W}$と因子得点行列$\boldsymbol{V}$の要素について以下の操作を行い，標準化因子負荷量行列$\boldsymbol{W}'$，標準化因子得点行列$\boldsymbol{V}'$

$$N_m = \sum_n \boldsymbol{W}_{m,n} \tag{16}$$

$$\boldsymbol{W}'_{m,n} = \frac{\boldsymbol{W}_{m,n}}{N_m} \tag{17}$$

$$\boldsymbol{V}'_{k,m} = N_m \cdot \boldsymbol{V}_{k,m} \tag{18}$$

とする．

標準化因子負荷量行列$\boldsymbol{W}'$は，行ごとの和が1になっており，それぞれの因子が含んでいる年齢階層の比率を表している．また，標準化因子得点行列$\boldsymbol{V}'$は，因子ごとの実際の人数と同じオーダーの要素を持っている．これらの内積は，

$$\sum_m \boldsymbol{V}_{k,m}\boldsymbol{W}_{m,n} = \sum_m \boldsymbol{V}'_{k,m}\boldsymbol{W}'_{m,n} \tag{19}$$

のように元の行列同士の内積と等しく，標準化因子負荷量行列$\boldsymbol{W}'$と標準化因子得点行列$\boldsymbol{V}'$は観測行列$\boldsymbol{X}$を分解した一つの形である．

(3) 結果の意味づけ

都道府県外へ転居するかどうかの選択は，基本的に就職，進学，職場の変化，結婚・離婚や住み替えと言った重大なイベントごとに起こる．この人生の節目を「ライフイベント」とする．本研究は，移動人口と残留人口をライフイベントの加算として把握することを目的としている．

以下，本研究での因子得点行列$\boldsymbol{V}$および因子負荷量行列$\boldsymbol{W}$の意味づけについて説明する．因子負荷量行列$\boldsymbol{W}$の各行は，M個ある因子それぞれの年齢階層の発生度合を表す．また因子得点行列$\boldsymbol{V}$の各行は観測行列Xの各都道府県の転出，転入，残留を表現する因子の加算の度合いを表している．因子負荷量行列$\boldsymbol{W}$は，NMFの条件の一つである非負制約により疎になる傾向があるため，因子は互いに独立になるように求まる傾向にある．したがって，各ライフイベントがそれぞれ異なる年齢階層をカバーする傾向にあるため，分析結果の解釈はより容易になる．

4. 考察

以下，標準化後の結果について考察していく．各因子の因子負荷量は**図-1**のように得られた．各因子は**表-2**のようなライフイベントとして解釈し，それに基づいて各

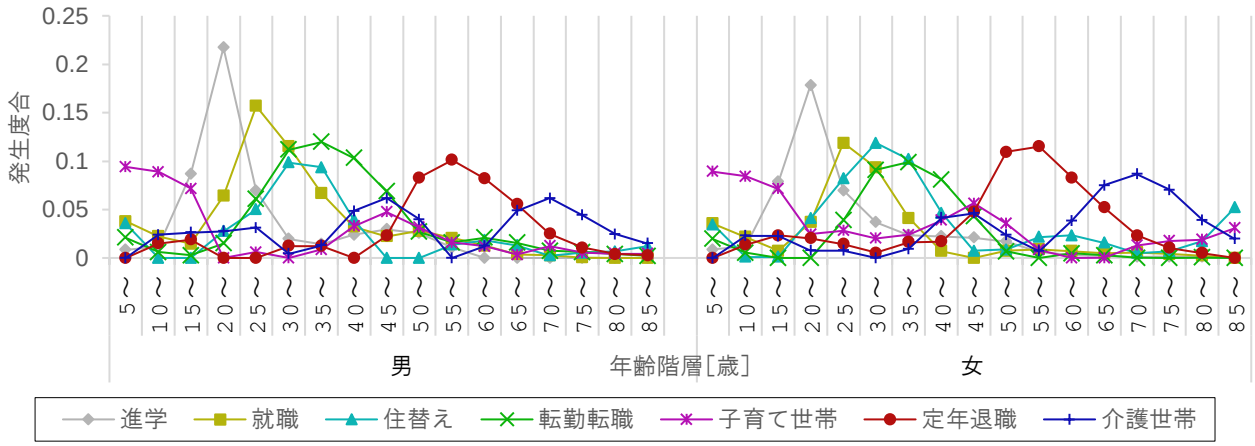

図-1　ライフイベント発生年齢

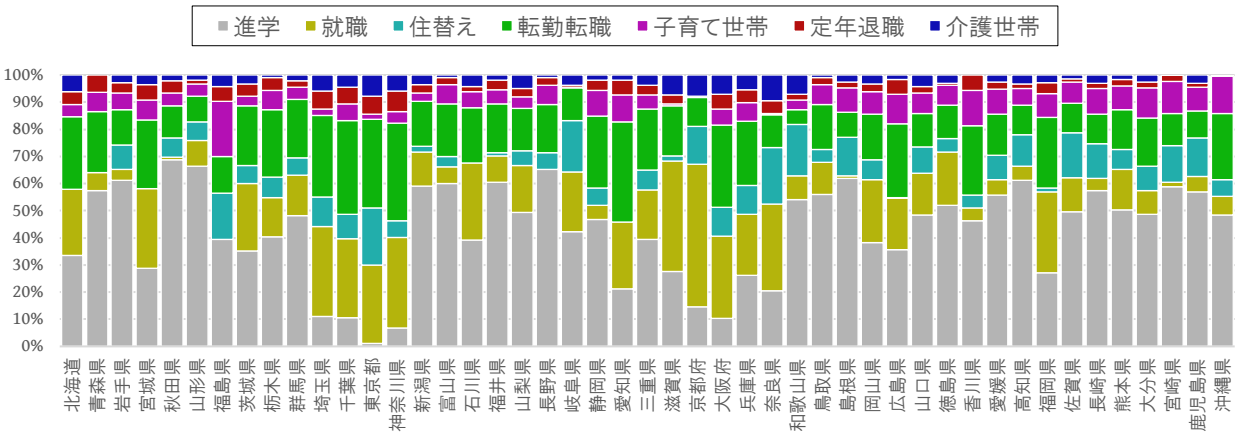

図-2　2010-2015 年　転出　ライフイベント発生割合

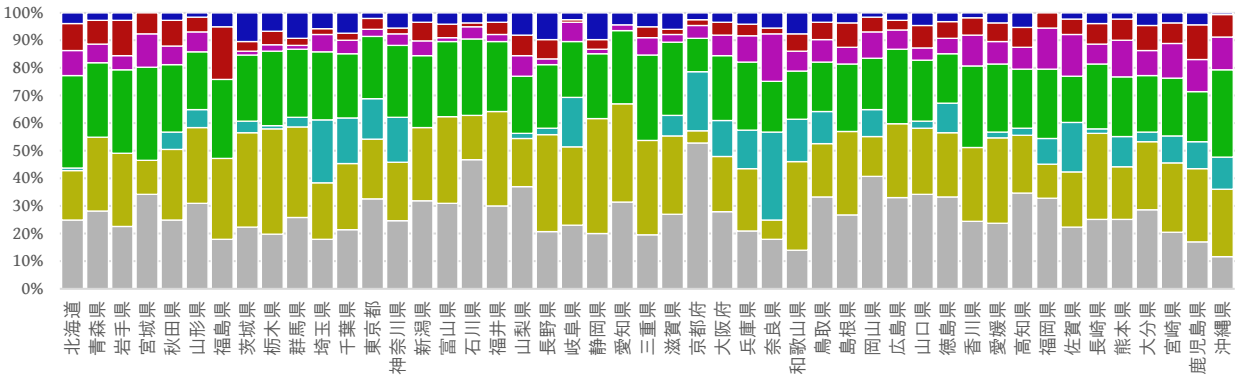

図-3　2010-2015 年　転入　ライフイベント発生割合

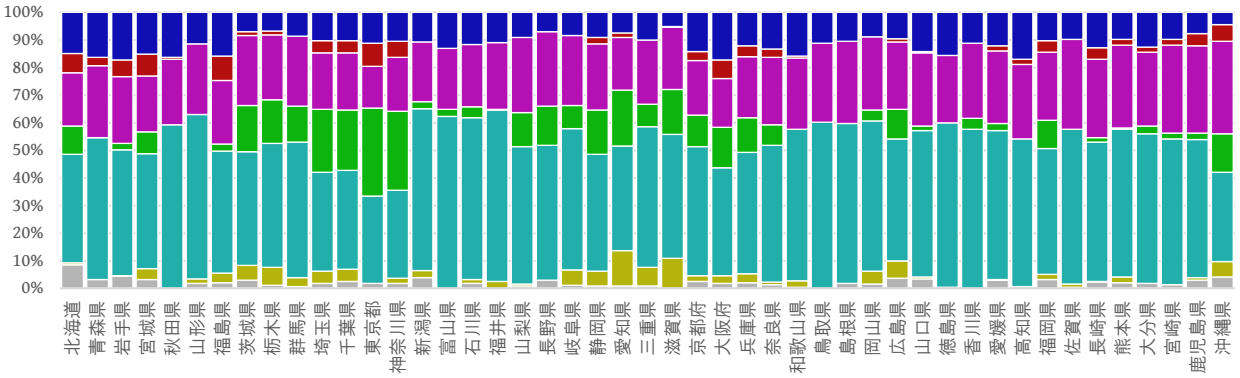

図-4　2010-2015 年　残留　ライフイベント発生割合

表-1　分析データの形式

[人]	5~10 歳 男性	…	80~85 歳 女性	85 歳以上 女性
2010-2015 年 北海道転出		…		
:	:		:	:
2010-2015 年 沖縄県転出		…		
2010-2015 年 北海道転入		…		
:	:		:	:
2010-2015 年 沖縄県転入		…		
2010-2015 年 北海道残留				
:	:	…	:	:
2010-2015 年 沖縄県残留		…		
2000-2005 年 北海道転出		…		
:	:		:	:
2000-2005 年 沖縄県残留		…		
1995-2000 年 北海道転出		…		
:	:		:	:
1995-2000 年 沖縄県残留		…		

表-2　各因子の解釈

因子名	ライフイベントとしての解釈
進学	高校卒業後の進学.
就職	高校，大学，短大，専門，大学院等からの卒業後の就職.
住替え	比較的若い層の男女．結婚，子供が産まれ広い住居へ移る他，社会人になって何年か経過し落ち着いた成人の住み替えや独立など.
転勤転職	生産年齢人口の若年層で占められる．やや男性に偏っている.
子育て世帯	成人前後までの年齢の子供とその両親.
定年退職	退職に伴う住み替え.
介護世帯	後期高齢者とその介護を行う世帯．女性の方が平均寿命が長いため女性が多い.

期間の各都道府県の転出，転入，残留の因子得点について考察できる.

(1)　今回得られた結果への考察

分析結果では，採用した3期間の中でも特徴的な人口移動の起こっている[1)]2015年に着目して考察する.

2011年に発生した東日本大震災の原発事故により，子どもの放射線リスクを回避したい母子が原発避難をする一方，父親は被災地の働き手として残留する母子避難が起こった[6)]．これが2015年の福島県の転出者数に子育て世帯の増加という形で表れている．また，被災地で急増した復旧・復興事業の多くが男性中心の仕事内容であったため男性労働者が福島県へ他都道府県から流入した[7)]. 特に，原発作業員の年齢層は4割が50代以上であり[8)]，「定年退職」の因子が同じ年齢階層であることから，2015年の福島県の転入数の増加に当たると考えられる.

更に，都道府県別残留数を見ると，首都圏では転勤や転職に伴って残留を選択する者が多い．首都圏の一都三県では，他都道府県に比べ越県通勤が容易であるためだと考えられる．特に埼玉県，千葉県及び神奈川県は，県外に通勤・通学している者を合わせると人口の１割を上回っている[9)]．

(2)　従来手法による分析結果との違い

a)　スパース性の向上

従来手法と比べ，本研究の手法では因子の年齢階層別の発生度合にあたる因子得点の行列にスパース性が増し，解釈しやすい明確な結果が得られている.

従来手法である一般的なNMFは，EMアルゴリズムにより誤差行列$\boldsymbol{E}$を元に誤差基準値となる値を計算し，それを最小化することで行われる[2)]．誤差基準値が小さくなればなるほど，得られる結果はスパースとなり，特に年齢別人口移動モデルにおいて同一因子内で高い発生度合を持つ年齢階層の組合せがより理解しやすい形となる. 結果のスパース性が小さく，一つの因子内で多数の年齢階層が共起してしまうと，異なる都道府県や異なる期間どうしのライフイベント発生割合の性質の違いが読み取りにくくなってしまう.

ベイズ統計学の考え方を取り入れた本研究のNMF法では，0に近い値や負値が得られた際に，サンプルの別の要素の力により，0に収束させるような仕組みが働くため，スパース性が高まる.

b)　収束判定の設定

そもそもNMFは収束判定の研究が少なく，音響データや画像データと言った他の性質のデータを用いたNMFの研究でも，分解された行列の妥当性を分析者が各自様々な方法で観測して判断し収束判定を設定している場合が多い．目や耳で分解のクオリティを確認しやすい音響や画像のモデリングとは異なり，人口移動モデルは妥当性を確認しにくく，収束判定の設定が問題となる.

本研究ではベイズの手法により不確実性の評価が可能

であるため，適切な結果であるかどうかをベイズの観点から議論することができる．Schmidt et al.[4]が採用した周辺尤度の他にも，BIC基準などを採用できると考えられる．

実際に図-1を見ると，理由の不明瞭な住み替えの因子が1つに絞られたなど，先行研究[1]よりも解釈が明確な結果が得られたことがわかる．

5. 今後の方針

今後の発展方向として，外部情報を取り入れた事前分布をパラメータに与えることで，外部情報が分析結果に反映されるようにすることが考えられる．

また，ベイズの考え方で収束判定の条件に用いられる周辺尤度などの基準値は，適切な因子数選択の基準にもなる[4]．同様に年齢別人口移動モデルでも因子数選択に解釈のしやすさだけでなく具体的根拠を与えられると考えられる．

6. おわりに

本研究では，年齢構成ごとの転出者数と転入者数を他期間や他の都道府県と比較し人口移動の質的特質を明らかにするNMF手法における幾つかの課題に対して，ベイズの考え方に基づくNMFモデルを適用することで解決することを考えた．震災前後年の国勢調査都道府県別データに適用した結果，分析結果のスパース性が向上するなど，より妥当な結果を得ることができた．

今後は，外部情報を取り入れた事前分布や，因子数選択の基準などを設定し，分析手法を発展させていきたい．

参考文献

1) 元井初音，奥村誠，水谷大二郎：年齢構成の共通性に着目した地域社会の比較分析法，土木計画学研究・講演集，Vol.57，No.17-09，2018.
2) 亀岡弘和：非負値行列因子分解，計測と制御，第 51 巻第 9 号，pp.835-844，2012.
3) 澤田宏：非負値行列因子分解 NMF の基礎とデータ／信号解析への応用，電気情報通信学会誌，Vol.95，No.9，pp.829-833，2012.
4) Schmidt, M. N., Winther, O. and Hansen, L. K.: Bayesian non-negative matrix factorization, *Proceedings of ICA 2009*, LNCS 5441, pp.540-547, 2009.
5) 総務省統計局：統計局ホームページ/国勢調査の基本に関する Q&A，2019/10/03 参照 http://www.stat.go.jp/data/kokusei/qa-6.htm#f1
6) 宝田惇史：福島第一原発事故に伴う避難・移住における交通関連の課題―避難者の二重生活と支援者の全国ネットワーク化を中心として―，交通権，No.33，pp.53-62，2016.
7) 日本学術会議東日本大震災復興支援委員会福島復興支援分科会：東京電力福島第一原子力発電所事故による長期避難者の暮らしと住まいの再建に関する提言，第 13 回東日本大震災復興支援委員会，2014.
8) 産経ニュース：【福島第 1 原発事故　5 年目の真実（4）】今もピンハネある、口を出すと解雇される…作業員は弱い立場だ，2019/10/03 参照 https://www.sankei.com/affairs/news/160226/afr1602260010-n3.html
9) 総務省統計局：平成 27 年国勢調査　従業地・通学地による人口・就業状態等集計結果　結果の概要，2019/10/03 参照 https://www.stat.go.jp/data/kokusei/2015/kekka/jyutsu1/pdf/gaiyou.pdf

（2019. 10. 4 受付）