高次元ビックデータに対する統計的な特徴抽出のための次元縮約法に関する研究

原田　魁成

2021年12月27日

博士学位論文

高次元ビックデータに対する統計的な特徴抽出のための次元縮約法に関する研究

金沢大学大学院 人間社会環境研究科 人間社会環境学

学籍番号　1921082010

氏名　原田 魁成

主任指導教員名　寒河江 雅彦

目 次

第 1 章　はじめに **4**

1.1　本研究の概要 4
1.2　研究背景と研究目的 5
1.3　本稿の構成 6
1.4　使用データと応用研究に関する先行研究 7
1.4.1　産業連関表について 7
1.4.2　産業連関表に対する多変量解析法を用いた先行研究 7
1.4.3　携帯電話の位置情報データについて 9
1.4.4　携帯電話の位置情報データを用いた先行研究 9

第 2 章　次元縮約法と正則化モデル **10**

2.1　主成分分析 10
2.2　非負値行列因子分解 14
2.3　次元縮約法と正則化モデル 18
2.3.1　正則化モデルについて 18
(i) L_1 ノルムによる正則化モデル 20
(ii) L_2 ノルムによる正則化モデル 22
(iii) その他の正則化モデル 23
2.3.2　スパース主成分分析 24
2.3.3　スパース非負値行列因子分解 25

第 3 章　非負値行列因子分解に関する応用研究 **26**

3.1　非負値行列因子分解法を用いた地域産業特性の抽出 26
3.1.1　概要 26
3.1.2　使用データと設定について 26
(i) 使用データについて 26
(ii) 分析手法の設定 27
3.1.3　分析結果 27
(i) 第 1 基底:愛知県を中心とした製造圏 28
(ii) 第 2 基底:東京都を中心とした第 3 次産業 29
(iii) 第 3 基底:神奈川県・千葉県等の製油所拠点を中核とした鉱業 30
3.2　COVID-19 流行下の石川県内滞在者移動行動変化に関する研究 31
3.2.1　概要 31
3.2.2　使用データと設定について 31
(i) 使用データについて 31
(ii) 分析手法の設定 32
3.2.3　分析結果 32
(i) 第 1 基底:滞在の行動 (就寝) 33

(ii) 第 2 基底:平日の外出行動 (就業・就学) 35
(iii) 第 3 基底:休日の外出行動 (観光・飲食・娯楽等) 37

第 4 章　スパース非負値行列因子分解に関する研究 **39**
4.1 COVID-19 流行前後における県間移動行動変化に関する研究 39
4.1.1 概要 39
4.1.2 使用データと設定について 39
(i) 使用データについて 39
(ii) 分析手法の設定 40
(iii) COVID-19 に対する諸政策 41
4.1.3 分析結果 43
(i) 業務・通勤目的変動の時系列推移 43
(ii) 居住都道府県における滞在者数変動の時系列推移 46
(iii) 長期休暇時の大都市から地方への行動の時系列推移 47
(iv) 週末の外出行動の時系列推移 48
(v) 残差の時空間分布 49
4.2 羽咋市における定常的な人口流動分析に関する研究 51
4.2.1 概要 51
4.2.2 使用データと設定について 51
(i) 使用データについて 51
(ii) 分析方法の設定 52
4.2.3 分析結果 53
(i) 非負値行列因子分解による分析結果 53
(ii) スパース非負値行列因子分解による分析結果 57

第 5 章　主成分分析に関する応用研究 **61**
5.1 地域産業構造の特徴抽出と可視化に関する研究 61
5.2 使用データと設定について 61
5.2.1 使用データについて 61
5.2.2 分析方法の設定 62
5.3 分析結果 64
5.3.1 2011 年全国産業連関表を用いた都道府県クラスタリング 64
5.3.2 福島県・東京都・石川県における産業構造の経年変化 66

第 6 章　スパース主成分分析に関する研究 **68**
6.1 COVID-19 流行前後における生活行動の変化を捉えるためのスパース主成分分析 68
6.2 使用データと設定について 68
6.2.1 使用データについて 68
6.2.2 分析手法の設定 68
6.3 分析結果 71
6.3.1 第 1 主成分「平日の行動 (就業・就学+観光)」 71
6.3.2 第 2 主成分「休日の行動 (買い物・飲食等)」 73

第 7 章　おわりに **75**
7.1　結論 75
7.2　今後の発展 77
7.2.1　基底数及びハイパーパラメータの設定 77
7.2.2　高次元テンソルデータへの拡張 80

付録 1:補足 **86**

付録 2:テンソル因子分解に関する応用研究 **88**

参考文献 **98**

第1章　はじめに

1.1　本研究の概要

情報社会の進展に伴い、様々な現象がデジタル化され、ビックデータとして蓄積されるようになってきた。内閣府 [1] がデータを利活用するための方針として掲げる「Society5.0」においても、Internet of Things(IoT) によって得られたビックデータから新たな価値を創出し、課題に悩む各現場にアウトプットとして還元することの重要性について説かれている。すなわち社会課題を解決するためのビックデータ解析は今後一層需要が増すことになるであろう。

一方でビックデータを解析する場合、大標本と高次元な構造から、データ解析でよく使用される従来型のモデルでは、解析が容易に行えない場合がある。本稿ではこうしたデータに対する解析法として、高次元データを低次元に縮約させる次元縮約法とデータの過学習[1]を防ぐ正則化モデル[2]の 2 点について研究し、それらを用いた応用研究を行っている。

次元縮約法は、行列構造を「行」と「列」に分解して解釈できる。例えば顧客の購買情報が記された POS データに関して、行方向に顧客の属性、列方向に購買した商品が記録されている場合、「顧客の属性」と「購入した商品」の 2 つの行列に分解され、これらを併せて「どの属性 (年齢・性別等) がどのような商品 (群) を購入する傾向があるか」という特徴が解析できる。また従来型の分析方法と異なり、明確な目的変数を設定しない解析手法である。したがって、特定の変数に絞って解析を行う従来型のモデルよりも、データの構造を客観的に解析できる。本稿では次元縮約法のうち、「主成分分析 (PCA)」や「非負値行列因子分解 (NMF)」に関する研究を行っている。主成分分析は 1900 年初頭から存在する解析手法 [2] であり、理論的に性質が整備されている。縮約する次元 (主成分) を、データのばらつきが最大となるように計算するため、低次元でかつデータの特徴を捉えた解析が可能であり、さらに直交性の下で主成分が構築されることから、主成分同士で異なった特徴が抽出できる利点がある。一方、非負値行列因子分解は 1999 年に提案された解析手法 [3] であり、行列分解された 2 つの行列の要素がすべて非負値で抽出されることから、正負で特徴が相殺されることなく、加法的な解釈が可能となる。特に画像データや購買データ等の非負値のデータ構造を分析するために望ましい解析手法である。

正則化モデルは、解析モデルの複雑化や過学習を防ぐことができ、分析結果の解釈を簡素化できる [4]。例えば線形回帰モデルの誤差を最小にする最小二乗法を用いた分析では、説明変数の次元数分の回帰係数が算出されるため、ビックデータを対象とした解析では高次元の回帰モデルとなり、非常に複雑となる場合がある。また同モデルは与えられたデータに対する誤差 (分散) を小さくすることは可能であるが、その一方で真のモデルとの誤差 (バイアス) を大きくする場合がある。こうした問題に対し、正則化項と呼ばれる制約条件をモデルに付与することで、回帰係数の値を平滑化させる、または 0 にすることができ、モデルの複雑化や過学習を防ぐことができる。特に本稿では、こうした性質を持つ正則化項のうち、解析結果の一部の値を 0 (この性質をスパース性と呼ぶ) にすることで、解釈を容易にできる L_1 正則化項に着目し、次元縮約法に同項を付与することで、通常の次元縮約法よりも明瞭な解釈が可能となることを示している。

本稿ではこれらの理論的な性質を整備しつつ、実際のビックデータに適用する研究を行っている。具体的には経済学の分野で使用されている、財・サービスの取引を 1 つの行列に集約した産業連関表と、携帯電話の位置情報から得られるメッシュ内滞留人口データ及び都道府県間人口移動データを解析対象としている。

[1] 与えられたデータに対してモデルがオーバーフィットすること
[2] 詳細は 2.3.1 を参照のこと。本稿では次元縮約法に正則化項を付与したモデルを正則化モデルとしている。

産業連関表は産業分類の粒度と地域間取引の組み合わせによって膨大なデータとなる。例えば、国際 (間) 地域産業連関表では 2464×2464(44ヵ国 ×56 産業の組み合わせ) の行列データとなる。また、携帯電話の位置情報データにおいても、メッシュ内滞留人口データが 500m×500m の範囲かつ 1 時間ごと 24 時間 365 日分のデータが記録されているため、解析の対象地域及び期間次第で膨大なデータとなる。特に本稿で使用したデータは 2394 時系列 × 1053 メッシュ である (「3.2 節」、「第 6 章」での分析に対応する)。

最後に今後の展開として、いずれの次元縮約法及び正則化モデルにおいても共通して未解決な課題である、縮約する次元数及び正則化項のハイパーパラメータの選択に関して、先行研究を整理しつつ、本稿の応用研究で使用した選択基準も併せて議論する。またさらなる高次元データ解析のために、高次拡張されたテンソルデータに対する解析へと展開させる。

1.2 研究背景と研究目的

本研究ではビックデータに対し、次元縮約法を用いてデータが持つ共通の構造を、事前に設定せずとも客観的に抽出できること、また次元縮約法に L_1 正則化項を付与した正則化モデルを用いて、通常の次元縮約法による特徴抽出よりもさらに明瞭な特徴が抽出できることを分析目的とし、それらの研究を行っている。次元縮約法に関する研究背景として、データ解析における従来型のモデルは、分析したい特定の変数を、あらかじめ目的変数として設定した上で解析を行っている。例えば POS(Points Of Sales) データのような、リアルタイムに購買情報が更新される時系列ビックデータを解析する場合、時系列分析でよく使用される ARIMA モデル (Auto Regressive Integrated Moving Average：自己回帰和分移動平均モデル) 等の従来型の解析手法では、ある特定の商品や年齢・性別等のカテゴリーに絞って、売上の動向を解析することとなる。こうした分析手法では、特定の商品やカテゴリーに関する時系列の動向を詳細に解析することが可能であるが、その一方でそれら以外の変数は同モデル内では考慮されない。すなわち同手法では、選択した商品やカテゴリーにおける変数間の関係性を見落とす可能性があり、その関係性が適切に解析されるか否かは、解析者の力量に依存することとなる。他方、本稿で使用する次元縮約法は、解析する目的変数を特定せず、全ての変数を含めた解析を行う手法である。そのため、複数の変数間に共通する特徴が抽出される場合がある。こうした視点による解析は本研究以外にも複数の研究において行われている。Roepke(1974) [8] では、特定の変化によって経済全体に及ぼされる影響を解析するためには産業連関表が有益な統計資料であることを主張しつつ、取引構造の複雑性から、産業クラスターの特定のためにモデルを単純化する方法が必要であると述べている。また、渡邉ら (2009) [13] は、地域の産業クラスターを特定する際に、その存在を認めたうえで特徴を明らかにするようなケーススタディ的な研究方法ではなく、「探索的」に各地域における産業クラスターの特定化を試みている。いずれの研究においても、解析者が恣意的に選択した特定の変数について分析するのではなく、分析モデルを用いて客観的にデータの解析を試みている。これらから、解析者の分析目的によるが、従来型の解析手法は、特定の変数に絞った解析においては有効であり、複数の変数間に共通する特徴を解析する場合は、本稿で議論する次元縮約法が有効であると考える。

正則化モデルは、次元縮約法に L_1 正則化項を付与することで、通常の次元縮約法よりもさらに明瞭に解釈できる特徴抽出を分析の目的とする。これは回帰モデルと同様に、回帰係数の一部を 0 にするスパース性によって達成される。研究背景として、通常の次元縮約法では、縮約後の特徴の多くが非 0 値で構成されている。こうした事象は回帰モデルにおけるデータの過学習の問題と同様 [4] であり、モデルとのバイアスを大きくするのと同時に分析結果の解釈が困難となる可能性がある。正則化モデルは通常の回帰モデルが有する課題に対し、L_2 正則化を付与する Ridge 回帰ではモデルの安定化、L_1 正則化を付与する Lasso 回帰ではモデルの簡素化させることで課題を解決できる。特に本稿では L_1 正則化を付与する Lasso 型のモデルを、一次元の回帰モデルから多次元の行列因子分解モデルへと拡張した上で、回帰係数の一部を 0 にする性質によって、分析結果の解釈の簡素化を試みている。本稿では次元縮約法の一種である主成分分析と非負値行列因子分解に正則化項を付与したモデルとしてそれぞれ、スパース主成分分析とスパース非負値

行列因子分解について議論するが、これらを適用した応用研究例は国内外含めてわずかである。そのため、応用研究を通じて、正則化項の効果と解析する上で生じる課題について検証する。

1.3 本稿の構成

本稿の構成は図 1.1 の通りである。

2 章では本稿で使用する主成分分析と非負値行列因子分解、及びモデルを簡素化するためのスパース制約について理論面の整備を行っている。2.1 節は主成分分析について、2.2 節では非負値行列因子分解について議論する。また 2.3.1 項は正則化モデルについて、2.3.2 項は主成分分析にスパースモデルを組み込んだスパース主成分分析について、2.3.3 項は非負値行列因子分解にスパースモデルを組み込んだスパース非負値行列因子分解について議論する。3 章は非負値行列因子分解に関する応用研究をまとめている。3.1 節は 47 都道府県の産業連関表データに対する研究例、3.2 節は金沢市近郊含む石川中央都市圏におけるメッシュ内滞留人口データに対する分析結果である。4 章は非負値行列因子分解にスパースモデルを組み込んだ応用研究をまとめている。4.1 節は 47 都道府県の県内滞在・県間移動人口データに対して行列分解後の片側の行列にスパース制約を課した分析結果、4.2 節は石川県羽咋市における滞在者数データに対して行列分解後の両側の行列にスパース制約を課した分析結果である。5 章は主成分分析を 47 都道府県の産業連関表データに適用した分析結果である。6 章は主成分分析にスパースモデルを組み込んだスパース主成分分析に関して、3.2 節と同様の位置情報データに適用した分析結果である。7 章は分析手法や応用研究から得られた結果をまとめる。加えて今後の展開として、(1) 縮約する次元数及び正則化項のハイパーパラメータの決定に関して、(2) 高次拡張されたテンソルデータに対するデータ解析の 2 点について述べる。また (2) に関する予備的な解析結果として中部圏地域間産業連関表を用いて解析した応用研究例を「付録 2」に記している。

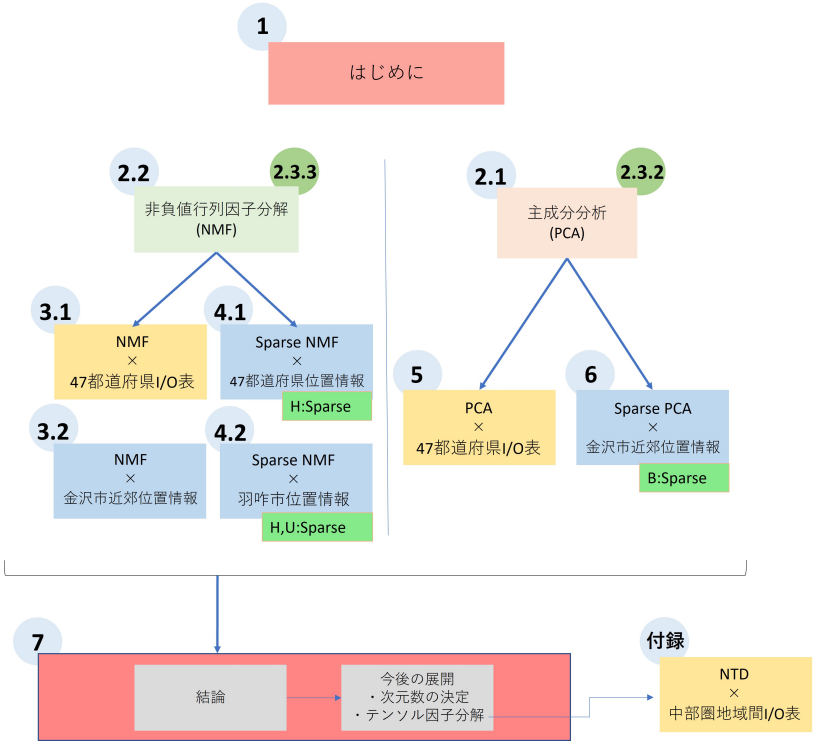


図 1.1: 本稿の構成 黄色:経済データ (産業連関表) 青色:携帯電話の位置情報データ

1.4 使用データと応用研究に関する先行研究

ここでは第 3 章から第 6 章までの応用研究で共通して使用するデータに関する概要と、それらを用いた先行研究をまとめる。

1.4.1 産業連関表について

産業連関表 (Input-Output Tables) は、国内経済において一定期間（通常 1 年間）に行われた財・サービスの産業間引を一つの行列に集約した統計表であり、1960 年より全国及び各都道府県や一部の政令指定都市において 5 年ごとに作成されている。産業連関表は、対象年次の産業構造や産業部門間の相互関係など経済の構造を総合的に把握できるため、経済波及効果の測定や各種経済統計における基準値等の研究に利用されている。

図 1.2 は産業連関表の概要図であり、取引基本表と呼ばれているデータである。これは財・サービスの取引で発生した金額を記録したものであり、他の指標作成の基となる最も基本的なデータである。産業連関表は大きく 2 方向の構造を有しており、行方向における生産物の販路構成を記録した「産出」(供給) と、列方向における原材料及び粗付加価値の費用構成を記録した「投入」(需要) によって表されている。すなわち、行方向にデータを見ることで、i 行産業が j 列産業にどれほどの財・サービスを産出 (供給) しているかが確認でき、反対に列方向にデータをみることで、j 列産業が i 行産業から財・サービスをどれほど投入 (需要) しているかが確認できる。

産業連関表の基本的な構成要素は、内生部門表、投入係数表、逆行列係数表などがある。いずれも主に図 1.2 の「中間投入」と「中間需要」が交わる範囲を指し、各部門の投入構造・産出構造が記録された産業連関表の核となるデータである。内生部門表は、「中間投入」と「中間需要」が交わる範囲の金額単位での取引を表し、産業間で発生した財・サービスの取引金額について記録されている。投入係数表は、中間需要の列部門を当該部門の総生産額で除して得た係数であり、当該部門産業 1 単位あたりの「原材料」を表す。逆行列係数表は、ある産業に対して、 1 単位の最終需要があった場合、各産業の生産が究極的にどれだけ必要となるかを示す係数となっている。逆行列係数は級数展開の性質を利用して算出されており、その展開の構造から、ある産業の最終需要が 1 単位変動した場合の経済的影響を示す経済波及効果 (又は経済効果) の算出に利用されている。

また産業連関表では、産業間の取引が様々な粒度で記録されている [5]。全国産業連関表では最も細かいものから、基本分類（行 509 部門 × 列 391 部門）、統合小分類 (187 部門 ×187 部門)、統合中分類 (107 部門 ×107 部門)、統合大分類 (37 部門 ×37 部門)、ひな形 (13 部門 ×13 部門) がある。ただし集計年度や各自治体によって部門統合数に多少のばらつきがある。産業の並びは、取引基本表内の左から (上から) 右へ (下へ) 向かうにつれ、第 1 次産業から第 3 次産業に関連する産業となるように配列されている。また基本分類を除く統合分類では、基本的に産業名の配列は対称となっている (図 1.2 参照)。

上記のことを踏まえ、本稿において産業連関表に関する分析では、地域経済の金額ベースに基づく大きな特徴を解析する目的から、「内生部門表」のデータを使用して分析する。対象とする産業部門は，経済活動の大きな特徴を捉えることを目的に、統合大分類またはひな形を用いて分析する。

1.4.2 産業連関表に対する多変量解析法を用いた先行研究

産業連関表は当該地域のマクロ的な経済取引構造が網羅的に解析できる有益な統計データである。しかし既存研究においてこのデータは、特定の産業または地域における最終需要が変化したときの経済波及効果及び経済損失の推定や、産業、粗付加価値、最終需要項目等の特定の項目またはその組み合わせによって地域比較を行う、ミクロな地域経済分析を行う際の統計資料として活用されることが多い。解析者の分析目的にもよるが、既存研究の方法では、地域の基盤産業や地域産業クラスター等を解析するためには、解析者

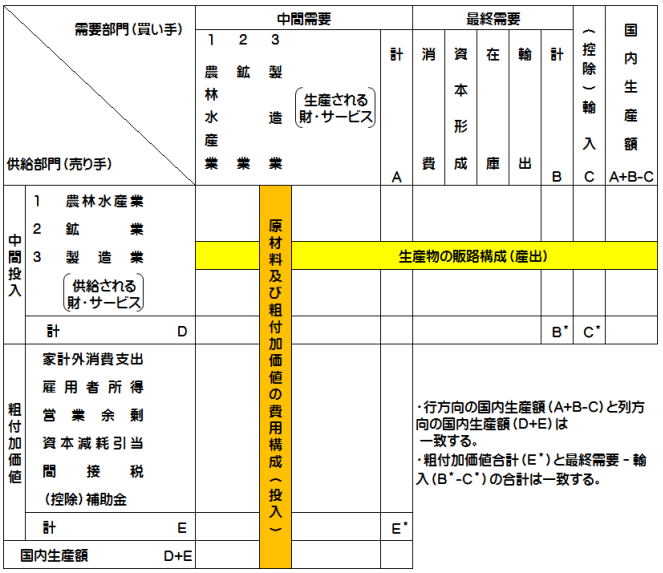


図 1.2: 産業連関表の取引基本表　(出典:総務省 [5])

が事前に基盤産業や産業クラスター等を特定しておかなければならない。また、特定の項目に絞った解析となるため、他の項目との関連性は考慮されない。こうした背景を踏まえ、産業連関表に本稿の分析手法を適用することで客観的な視点から地域経済分析が可能であることを示すため、産業連関表データを分析対象とした。

産業連関構造に関する研究は、Isard et al.(1959) の研究 [6] が産業クラスター分析の礎となり、1970 年代から 1980 年代にかけて産業クラスターの特定に関する研究が行われた。Czamanski(1974) [7] や Roepke et al.(1974) [8] は、産業クラスターを検出するために主成分分析や因子分析などの多変量分析手法が利用されている。また、産業クラスターの検出に多変量解析手法を適用する方法は Bergman and Feser (1999) [9]、Feser and Bergman (2000) [10] らによって広く知られるようになった。海外におけるこれらの研究は Hofe and Bhatta(2007) [11] が体系的にまとめている。

日本における先行研究として、長沢 (1988) [12] が日本の接続産業連関表に因子分析を適用し、製造業の投入産出構造の変化を分析している。また渡邊ら (2009) [13] は「関東」「中部」「近畿」の地域産業連関表に因子分析を適用し、各地域の産業クラスターの特定を試みている。千葉 (2019) [14] は中心都市と周辺町村との経済関係から、町村の存続可能性を解析する目的で、産業連関表の数値を一部利用して経済要因に関する指標を作成し、それに主成分分析を適用する研究を行っている。楫取 (2016) [15] は山口県と鳥取県の地域産業連関表に因子分析と非負値行列因子分解を適用し、産出行列や連携行列等の複数の産業構造分析の視点から両分析手法の比較を試みている。直近では、Mascaretti and Andrea(2019) [16] が国内産業連関表に非負値行列因子分解を適用し、ネットワーク分析の観点から経済ネットワークに影響を及ぼす産業の検出に関する研究を行っている。

産業連関表に対する多変量解析手法を用いた応用研究に関して、数学的に解析手法が早くから確立されていた主成分分析や因子分析を用いた研究は海外及び日本においても散見される。ただし、本研究で主成分分析を使用した分析例は、全国の産業構造を同一の基準で比較するために、全国産業連関表に主成分分析を用いて固有空間を形成し、その空間上に各県の特徴を射影することで比較を可能にしたものである。こうした主成分分析の使用方法は本研究が初めての分析例である。他方、非負値行列因子分解を用いた特徴抽出 (「3.1 節」の研究例) は海外及び日本においてもほとんど研究例がなく、テンソル因子分解法を適用する解析 (「付録 2」の研究例) は本稿が初めての分析例となる。

1.4.3 携帯電話の位置情報データについて

携帯電話の位置情報は、「いつ・どんな人が・どこから・どこへ」移動したかを解析できるビックデータである。具体的には日本全国を対象に、どこ (居住地) からどれだけの人が来ているか、あるエリアに住む人がどこに分布しているか等が、属性別 (性別・年代別) に記録されたデータであり、1 時間ごとの 24 時間 365 日分、最小 500 m ×500m メッシュ単位 (一部地域で 250m メッシュ単位) で提供されている。これらのデータは、電話やメールを利用するために配置された基地局がエリアごとに所在する携帯電話を測定する仕組みを利用し、その台数から各エリアの人口が推計されている。本稿では携帯電話の位置情報データとして NTT ドコモが提供する「モバイル空間統計」®[3] [17] を利用した分析を行っている。モバイル空間統計は、国内居住者 8 千万台分の携帯電話の運用データから作成されたものであり，大量なサンプルから推計される精度の高い動的人口データである。

1.4.4 携帯電話の位置情報データを用いた先行研究

携帯電話の位置情報データは、人の移動 (・滞在) を解析する上で重要なデータである。人の移動を記録したデータとして、国勢調査やパーソントリップ調査等が挙げられるが、これらは調査対象者の詳細な属性情報が獲得できる代わりに、大規模調査となるためデータの取得に多くの年数を要する。また、得られた移動に関する情報も、調査時点に限られた静的な移動情報である。すなわちこれらのデータを用いて動的な人の移動を解析することは困難である。他方、携帯電話の位置情報データに関して、本研究で使用したモバイル空間統計では、 1 時間単位 24 時間 365 日で最小 500m×500m メッシュ単位の移動データが取得できる。こうしたリアルタイムな解析が可能な点で、人の移動を記録した既存統計調査と差別化されている。

しかし、携帯電話の位置情報データはリアルタイムな移動データが取得可能である一方で、膨大な移動データが日々蓄積されることとなり、それらの解析には工夫が必要となる。特定の地域に絞った解析を行う場合は従来の解析モデルを用いた場合でも解析可能であるが、大きな人の移動の傾向を解析する場合は、特定の地域に絞らない解析手法が求められる。こうした背景を踏まえ、リアルタイムにデータが蓄積されるビックデータに対し、本稿の分析手法を適用することで、データが持つ特徴を解析できることを示すため、携帯電話の位置情報データを分析対象とした。

携帯電話の位置情報データに関する研究例は都市計画や交通計画等の分野で多数見受けられる [18, 19]。Wang et al. [20] は中国・杭州における携帯電話の位置情報データ 400 万人分のデータに対しテンソル因子分解法を適用し、人々の移動パターンを出発地・目的地・時間の 3 つに分解してそれぞれの特徴解析を行っている。奥村 (2015) [21] では東日本大震災前後の仙台市における携帯電話の位置情報データに対し、因子分析を適用することで住民の生活行動の変化について解析している。林ら (2015) [22] は、欠損値を含む位置情報に対し、可変基底非負値行列因子分解を適用し、データの傾向を考慮した欠損値の補正を行いつつ、行動パターンを抽出する手法を提案している。

位置情報の記録データから、人の意思の情報である「旅行目的」を推測する手法はいくつか提案されている [23, 24]。特に Yamaguchi and Nakayama (2020) [25] では、1461 日分の時系列情報を持つ集計データに対して非負値行列（テンソル）因子分解を適用することで、新幹線開業に対する感度の異なる旅行グループに分解する手法を提案している。分析の結果、ほぼ既存のモデルで扱われてきた「旅行目的」と対応する旅行グループに分解できた。また、原田・山口・寒河江 (2021) [67](「4.1 節」の研究例) においても、COVID-19 流行前後における県間移動データから、「自粛」の特徴など人の意思に基づく移動行動の変化の解析に成功している。またここでは携帯電話の位置情報データに初めて L_1 正則化項を付与したスパース型のモデルを使用しており、行動パターンの明瞭化に成功している。

[3]「モバイル空間統計」は株式会社 NTT ドコモの登録商標です。

第2章　次元縮約法と正則化モデル

2.1　主成分分析

主成分分析 [2] はデータの特徴を最もよく説明できるように、主成分と呼ばれる低次元の別空間に写像する解析法である。具体的にはデータに対する分散を最大化するように主成分が構成され、かつそれらの主成分はいずれも直交するように構成される。主成分分析における主成分及び固有値の導出については以下の通りである (永田 (2001) [26] を参考に一般化している)。

データ行列 $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}^{n \times m}$ に対し、列ベクトルごとに標準化したものを以下のように表す、

$$\boldsymbol{u_1} = \frac{\boldsymbol{x_1} - \bar{x_1}}{s_1}, \quad \boldsymbol{u_2} = \frac{\boldsymbol{x_2} - \bar{x_2}}{s_2}, \quad \cdots, \quad \boldsymbol{u_m} = \frac{\boldsymbol{x_m} - \bar{x_m}}{s_m}. \tag{2.1}$$

ここで $\bar{x_m}$ は列ベクトル m に関する平均、s_m は列ベクトル m に関する標準偏差を表す。

(2.1) によって定義される $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}^{n \times m}, (m < n)$ と、主成分を構成するための固有ベクトルの行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2} \cdots, \boldsymbol{a_m}] \in \mathbb{R}^{m \times m}$ とする。データ行列を標準化した行列 $\boldsymbol{U}$ 及び固有ベクトルの行列 $\boldsymbol{A}$ との積で表される行列 $\boldsymbol{Z} = [\boldsymbol{z_1}, \boldsymbol{z_2} \cdots, \boldsymbol{z_m}] \in \mathbb{R}^{m \times m}$ を一般に主成分得点と呼ぶ。第 1 主成分得点ベクトル $\boldsymbol{z_1}$ は次のように表される。

$$\begin{aligned} \boldsymbol{z_1} &= a_{11}\boldsymbol{u_1} + a_{21}\boldsymbol{u_2} + \cdots + a_{m1}\boldsymbol{u_m} \\ &= \boldsymbol{U}\boldsymbol{a_1}. \end{aligned} \tag{2.2}$$

主成分分析はデータに対する分散を最大化するように主成分を構成することから、(2.2) についての分散を考え、これの最大化を行う。すなわち、$\boldsymbol{z_1}$の分散 $\boldsymbol{V}_{z_1}$ は以下の通りとなる、

$$\begin{aligned} \boldsymbol{V}_{z_1} &= \frac{1}{n-1}\boldsymbol{z_1^T}\boldsymbol{z_1} \\ &= \frac{1}{n-1}\boldsymbol{a_1^T}\boldsymbol{U^T}\boldsymbol{U}\boldsymbol{a_1} \\ &= \boldsymbol{a_1^T}\left(\frac{1}{n-1}\boldsymbol{U^T}\boldsymbol{U}\right)\boldsymbol{a_1} \\ &= \boldsymbol{a_1^T}\boldsymbol{R}\boldsymbol{a_1}, \end{aligned} \tag{2.3}$$

ただし、$\boldsymbol{z_1^T}$は第 1 主成分得点ベクトル $\boldsymbol{z_1}$ の転置を表す。また $\boldsymbol{R}$ は相関行列を表し、以下によって表される。

$$
\begin{aligned}
\boldsymbol{R} &= \frac{1}{n-1}\boldsymbol{U^T U} \\
&= \frac{1}{n-1}\begin{bmatrix} \boldsymbol{u_1^T} \\ \boldsymbol{u_2^T} \\ \vdots \\ \boldsymbol{u_m^T} \end{bmatrix}[\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \\
&= \frac{1}{n-1}\begin{bmatrix} \boldsymbol{u_1^T u_1} & \boldsymbol{u_1^T u_2} & \cdots & \boldsymbol{u_1^T u_m} \\ \boldsymbol{u_2^T u_1} & \boldsymbol{u_2^T u_2} & \cdots & \boldsymbol{u_2^T u_m} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{u_m^T u_1} & \boldsymbol{u_m^T u_2} & \cdots & \boldsymbol{u_m^T u_m} \end{bmatrix} \\
&= \begin{bmatrix} 1 & \frac{1}{n-1}\boldsymbol{u_1^T u_2} & \cdots & \frac{1}{n-1}\boldsymbol{u_1^T u_m} \\ \frac{1}{n-1}\boldsymbol{u_2^T u_1} & 1 & \cdots & \frac{1}{n-1}\boldsymbol{u_2^T u_m} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{n-1}\boldsymbol{u_m^T u_1} & \frac{1}{n-1}\boldsymbol{u_m^T u_2} & \cdots & 1 \end{bmatrix},
\end{aligned}
\tag{2.4}
$$

ここで $\boldsymbol{u_i^T u_i}$ に関して標準化の条件より以下の結果を用いている、

$$
\boldsymbol{u_i^T u_i} = \left(\frac{\boldsymbol{x_i} - \bar{x}_i}{s_i}\right)^2 = \frac{(\boldsymbol{x_i} - \bar{x}_i)^2}{s_i^2} = \frac{(\boldsymbol{x_i} - \bar{x}_i)^2}{\frac{1}{n-1}(\boldsymbol{x_i} - \bar{x}_i)^2} = n - 1. \tag{2.5}
$$

また、(2.3) の最大化を行う上で、固有ベクトルの行列 $\boldsymbol{A} = [\boldsymbol{a_1}, \boldsymbol{a_2}\cdots, \boldsymbol{a_m}]$ の各列ベクトルの 2 乗和が 1 となる制約を設ける。すなわち第 1 主成分に関して、以下の制約条件を設ける、

$$
\boldsymbol{a_1^T a_1} = a_{11}^2 + a_{21}^2 + \cdots + a_{m1}^2 = 1. \tag{2.6}
$$

またラグランジュの未定乗数法を用いて、以下のように設定する、

$$
f(\boldsymbol{a_1}, \lambda) = \boldsymbol{a_1^T R a_1} - \lambda\left(\boldsymbol{a_1^T a_1} - 1\right), \tag{2.7}
$$

ここで λ はラグランジュ乗数である。(2.7) に関して、$\boldsymbol{a_1}$で微分して $\boldsymbol{0}$ とおく、

$$
\frac{\partial f(\boldsymbol{a_1}, \lambda)}{\partial \boldsymbol{a_1}} = 2\boldsymbol{R a_1} - 2\lambda \boldsymbol{a_1} = \boldsymbol{0}. \tag{2.8}
$$

これより、

$$
\boldsymbol{R a_1} = \lambda \boldsymbol{a_1}, \tag{2.9}
$$

が得られる。さらに (2.9) に左から $\boldsymbol{a_1^T}$ を掛けることで (2.6) より、

$$\boldsymbol{a_1^T R a_1} = \lambda \boldsymbol{a_1^T a_1} = \lambda, \tag{2.10}$$

を得る。よって、(2.3) と (2.10) より、$\boldsymbol{V_{z_1}} = \lambda$ が成り立つことから、$\boldsymbol{z_1}$ に関して分散を最大化されるためには、(2.9) の固有値問題を解いて、最大固有値に対応する固有ベクトルを求めればよいことがわかる。

第 2 主成分以降も同様にすると、以下のように表せる、

$$\begin{aligned} \boldsymbol{z_2} &= a_{12}\boldsymbol{u_1} + a_{22}\boldsymbol{u_2} + \cdots + a_{m2}\boldsymbol{u_m} \\ &= \boldsymbol{U a_2}. \end{aligned} \tag{2.11}$$

第 2 主成分の分散は以下のように表せる、

$$\begin{aligned} \boldsymbol{V_{z_2}} &= \frac{1}{n-1}\boldsymbol{z_2^T z_2} \\ &= \frac{1}{n-1}\boldsymbol{a_2^T U^T U a_2} \\ &= \boldsymbol{a_2^T}\left(\frac{1}{n-1}\boldsymbol{U^T U}\right)\boldsymbol{a_2} \\ &= \boldsymbol{a_2^T R a_2}. \end{aligned} \tag{2.12}$$

また、固有ベクトル $\boldsymbol{a}$ に関して (2.6) と同様の制約を設けると、以下のように表せる、

$$\boldsymbol{a_2^T a_2} = a_{12}^2 + a_{22}^2 + \cdots + a_{m2}^2 = 1. \tag{2.13}$$

ここで、第 1 主成分得点ベクトル $\boldsymbol{z_1}$と第 2 主成分得点ベクトル $\boldsymbol{z_2}$ は無相関となるように設定する。すなわち、互いに直交する制約を設ける。(2.2)、(2.10)、(2.11) より

$$\begin{aligned} \boldsymbol{z_1^T z_2} &= \boldsymbol{a_1^T U^T U a_2} \\ &= \boldsymbol{a_1^T R a_2} \\ &= \lambda \boldsymbol{a_1^T a_2} = 0. \end{aligned} \tag{2.14}$$

すなわち、(2.13) と (2.14) による制約条件からラグランジュの未定乗数法を用いて以下のように表せる、

$$f(\boldsymbol{a_2}, \lambda, \eta) = \boldsymbol{a_2^T R a_2} - \lambda\left(\boldsymbol{a_2^T a_2} - 1\right) - \eta \boldsymbol{a_1^T a_2}. \tag{2.15}$$

ここで λ 及び η はラグランジュ乗数である。(2.15) に関して、$\boldsymbol{a_2}$で微分して $\boldsymbol{0}$ とおくと、

$$\frac{\partial f(\boldsymbol{a_2}, \lambda, \eta)}{\partial \boldsymbol{a_2}} = 2\boldsymbol{R a_2} - 2\lambda \boldsymbol{a_2} - \eta \boldsymbol{a_1} = \boldsymbol{0}, \tag{2.16}$$

が得られる。さらに (2.16) に左から $\boldsymbol{a_1^T}$ を掛けることで以下が得られる、

$$2\boldsymbol{a_1^T R a_2} - 2\lambda \boldsymbol{a_1^T a_2} - \eta \boldsymbol{a_1^T a_1} = 0. \tag{2.17}$$

ここで、(2.14) より、$\boldsymbol{a_1^T R a_2} = 0$、$\boldsymbol{a_1^T a_2} = 0$ であるから、(2.17) に代入することで $\eta = 0$ が得られる。すなわち (2.16) より、

$$\boldsymbol{R a_2} = \lambda \boldsymbol{a_2}, \tag{2.18}$$

が得られ、(2.18) の左から $\boldsymbol{a_2^T}$ を掛けることで、(2.13) より、

$$\boldsymbol{a_2^T R a_2} = \lambda \boldsymbol{a_2^T a_2} = \lambda, \tag{2.19}$$

が得られる。よって、$\boldsymbol{V}_{z_2} = \lambda_2$ が成り立つことから、$\boldsymbol{z_2}$ に関して分散を最大化されるためには、(2.18) の固有値問題を解いて、第 2 固有値に対応する固有ベクトルを求めればよいことがわかる。第 3 主成分以降も上記と同様の手順によって、第 m 主成分まで求めることができる。

また $\boldsymbol{R}$ は相関行列を表す対称行列であるため、この固有値は実数値をとり、さらに半正定値行列であることから、固有値はすべて 0 以上となる [27]。これらの性質から

$$\lambda_1 + \lambda_2 + \cdots + \lambda_m = tr\boldsymbol{R} = m, \tag{2.20}$$

が成り立ち、一般に $\dfrac{\lambda_i}{m}$ の割合を第 i 寄与率と呼ぶ。

(2.10) 及び (2.19) とこれらを第 m 主成分まで求めた場合、以下のことが成り立つ、

$$\begin{aligned} \boldsymbol{\Sigma} &= \boldsymbol{A^T R A} \\ \boldsymbol{A \Sigma A^T} &= \boldsymbol{R} \quad ※\ \boldsymbol{A A^T} = \boldsymbol{I_m}\ . \end{aligned} \tag{2.21}$$

ただし $\boldsymbol{\Sigma}$ は以下の通りとする、

$$\boldsymbol{\Sigma} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_m \end{bmatrix}.$$

主成分分析における行列分解 (固有値分解) は、(2.21) のように対称行列を、左固有ベクトルの行列 $\boldsymbol{A}$、固有値行列 $\boldsymbol{\Sigma}$、右固有ベクトルの行列 $\boldsymbol{A^T}$ に分解することを指す。

2.2 非負値行列因子分解

非負値行列因子分解 (Non-negative Matrix Factorization,NMF) は、非負の観測データ行列 $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}_+^{n\times m}$ が得られた時、基底行列 $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{n\times k}$, 表現行列 $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}_+^{k\times m}$ の 2 つの低次元非負行列に分解することでデータの近似を行う解析手法である。ここで k は縮約化した次元数を表し、「基底数 (basis)」と呼ばれることもある。Lee and Seung(1999) [3] は行列データに対する最小二乗モデルから同分析手法を提案し、顔画像データを用いた実例を示している。そこでは他の次元縮約法である主成分分析やベクトル量子化と分析結果を比較し、正値のみで特徴が表現できる利点を挙げている。

非負値行列因子分解の計算アルゴリズムは、データとの乖離度を計測する基準によって異なる最適化問題に帰着される [28,29]。その中でもよく利用されるのが二乗誤差に基づく基準 [30]、一般化 Kullback–Leibler ダイバージェンス (I ダイバージェンス) [31] に基づく基準 [30]、板倉・齋藤擬距離 [32] に基づく基準 [33] の 3 つである (これら 3 つは β ダイバージェンス [34] を用いて一般化されている)。

本稿はデータに対する近似誤差を最小化する二乗誤差基準を採用している。以降、非負値行列因子分解による分析では、全てこの基準に基づいたアルゴリズムを使用している。非負値行列因子分解の性質及び計算アルゴリズムの導出については、亀岡 (2012,2015) [28, 29] を参照する。

$\boldsymbol{X}$ と $\boldsymbol{H}, \boldsymbol{U}$ の乖離度を $\boldsymbol{D}$ とすると、二乗誤差基準の乖離度は以下のように表せる、

$$D_{EU}(\boldsymbol{X}|\boldsymbol{HU}) = \sum_{n,m}(x_{n,m} - hu_{n,m})^2, \tag{2.22}$$

ここで D_{EU} は二乗誤差基準における $\boldsymbol{X}$ と $\boldsymbol{H}, \boldsymbol{U}$ の乖離度であり、$hu_{n,m} = \sum_k h_{n,k}u_{k,m}$ である。

非負値行列因子分解は (2.22) を最小とする $\boldsymbol{H}$ と $\boldsymbol{U}$ を推定することが目的となるが、いずれも非負制約付き非線形最適化問題であり、解析的に解くことが困難である。そこで Lee and Seung(2001) [30] は、解析困難な目的関数に代わり、補助関数を設定し、それらを反復的に降下させることで目的関数を間接的に降下させる、補助関数法 [35, 36] を用いた計算手法を採用している。

補助関数に関する定義及び原理は以下の通りである。$\theta = \{\theta_i\}_{(1\le i\le I)}$ をパラメータとする目的関数 $D(\theta)$ に対し、$D(\theta) = \min_{\alpha} G(\theta, \alpha)$ が成り立つとき、$G(\theta, \alpha)$ を $D(\theta)$ の補助関数とする。また α を補助変数と呼ぶ。補助関数 $G(\theta, \alpha)$ を α に関して最小化するステップと、$\theta_1, \cdots, \theta_I$ に関して最小化するステップ

$$\begin{aligned} \alpha &\leftarrow \arg\min_{\alpha} G(\theta, \alpha) \\ \theta_i &\leftarrow \arg\min_{\theta_i} G(\theta, \alpha) \qquad (i = 1, \cdots, I) \end{aligned} \tag{2.23}$$

を繰り返すことで、目的関数 $G(\theta)$ の値が単調に減少する。

反復計算のステップ数を l としたとき $\alpha^{(\ell+1)} = \arg\min_{\alpha} G(\theta^{(\ell)}, \alpha)$ であることから、補助関数の定義より、$D(\theta^{(\ell)}) = G(\theta^{(\ell)}, \alpha^{(\ell+1)})$ である。さらに 2.23 から、$G(\theta^{(\ell)}, \alpha^{(\ell+1)}) \ge G(\theta^{(\ell+1)}, \alpha^{(\ell+1)})$ であり、$G(\theta^{(\ell+1)}, \alpha^{(\ell+1)}) \ge D(\theta^{(\ell+1)})$ となることから、$D(\theta^{(\ell)}) \ge D(\theta^{(\ell+1)})$ となる。この性質から反復計算によって、目的関数を単調に減少させることができる。

以下では補助関数法を用いて、目的関数を単調減少させる計算アルゴリズムの導出を試みる。(2.22) に関して、$\boldsymbol{H}$、$\boldsymbol{U}$ に依存しない項を省略すると、

$$D_{EU}(\boldsymbol{X}|\boldsymbol{HU}) = \sum_{n,m}(-2x_{n,m}hu_{n,m} + hu_{n,m}^2), \tag{2.24}$$

と表せる。ただしこの時、$hu_{i,j}^2$が $\boldsymbol{H}$ と $\boldsymbol{U}$ の行列要素 $[h_{i,1}, \cdots, h_{i,k}]$、$[u_{1,j}, \cdots, u_{k,j}]^T$ を含んだ非線形関数項であり、複雑な計算処理が発生する。そこでこの項に対し、線形和に分離した上限関数を設ける。具体的にはこの項が凸関数であるため、Jensen の不等式を用いて、以下のような上限関数を設ける。$f(x)$ を実数上の凸関数とし、任意の $\lambda_i \geq 0$、$z_i, (i = 1, \cdots, n)$、$\sum_i \lambda_i = 1$ に対して

$$f\left(\sum_i z_i\right) \leq \sum_i \lambda_i f\left(\frac{z_i}{\lambda_i}\right), \text{ (等号成立は} \lambda_i = \frac{z_i}{\sum_j z_j}\text{)}. \tag{2.25}$$

を満たす関数とする。Jensen の不等式をこのモデルに適用すると、以下のように表せる、

$$\begin{aligned} hu_{n,m}^2 &= \left(\sum_k \lambda_{n,k,m} \frac{h_{n,k} u_{k,m}}{\lambda_{n,k,m}}\right)^2 \\ &\leq \sum_k \lambda_{n,k,m} \left(\frac{h_{n,k} u_{k,m}}{\lambda_{n,k,m}}\right)^2, \end{aligned} \tag{2.26}$$

ただし、$\lambda_{n,k,m} \geq 0$、$\sum_k \lambda_{n,k,m} = 1$ である。また (2.26) の等号は、

$$\lambda_{n,k,m} = \frac{h_{n,k} u_{k,m}}{hu_{n,m}}, \tag{2.27}$$

の時に成立する。(2.26) 及び (2.27) より、(2.24) に代入して以下の補助関数を得る、

$$G_{EU}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}) = \sum_{n,m} \left(-2x_{n,m} \sum_k h_{n,k} u_{k,m} + \sum_k \frac{h_{n,k}^2 u_{k,m}^2}{\lambda_{n,k,m}}\right), \tag{2.28}$$

ここで $\boldsymbol{\lambda} = \{\lambda_{n,k,m}\}_{N \times K \times M}$ とする。また (2.28) は補助関数の要件を満たす。これにより、$\boldsymbol{H}$、$\boldsymbol{U}$、$\boldsymbol{\lambda}$ について以下を満たすように反復計算を行い、目的関数 (2.24) を減少させる。

$$\begin{aligned} \boldsymbol{\lambda} &\leftarrow \arg\min_{\boldsymbol{\lambda}} G_{EU}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}) \\ \boldsymbol{H} &\leftarrow \arg\min_{\boldsymbol{H}} G_{EU}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}) \\ \boldsymbol{U} &\leftarrow \arg\min_{\boldsymbol{U}} G_{EU}(\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{\lambda}). \end{aligned} \tag{2.29}$$

すなわち、$\boldsymbol{H}$、$\boldsymbol{U}$ について微分して 0 とおくことで、それぞれ以下の関係式が導出される、

$$\hat{h}_{n,k} = \frac{\sum_m x_{n,m} u_{k,m}}{\sum_m \frac{u_{k,m}^2}{\lambda_{n,k,m}}}, \tag{2.30}$$

$$\hat{u}_{k,m} = \frac{\sum_n x_{n,m} h_{n,k}}{\sum_n \frac{h_{n,k}^2}{\lambda_{n,k,m}}}. \tag{2.31}$$

よって、(2.27)、(2.30) 及び (2.31) から、以下の計算アルゴリズムを得る、

$$\begin{aligned} h_{n,k} &\leftarrow h_{n,k}\frac{\sum_m x_{n,m}u_{k,m}}{\sum_m hu_{n,m}u_{k,m}}, \\ u_{k,m} &\leftarrow u_{k,m}\frac{\sum_n x_{n,m}h_{n,k}}{\sum_n hu_{n,m}h_{n,k}}. \end{aligned} \tag{2.32}$$

したがって、2.32 で得られた結果を、本稿で使用する非負値行列因子分解の計算アルゴリズムとする。

非負値行列因子分解の計算アルゴリズムは行列分解の観点から以下のようにも導出できる (木村 (2011) [37] を参考とする)。非負の観測データ行列を $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}_+^{n\times m}$、基底行列を $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{n\times k}$、表現行列を $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}_+^{k\times m}$ とするとき、二乗誤差基準に基づく近似誤差は目的関数 D_{NMF} を用いて以下のように表せる、

$$D_{NMF} = \|\boldsymbol{X} - \boldsymbol{HU}\|_F^2 + tr(\boldsymbol{\Psi H}^T) + tr(\boldsymbol{\Phi U}^T), \tag{2.33}$$

ここで、$\|\ \ \|_F$ はフロベニウスノルム[1]を表す。また、$\boldsymbol{\Psi H}^T$ 及び$\boldsymbol{\Phi U}^T$ はそれぞれ $\boldsymbol{H}$、$\boldsymbol{U}$ の非負制約におけるラグランジュ乗数であり、$\boldsymbol{\Psi} \in \mathbb{R}_+^{n\times k}$、$\boldsymbol{\Phi} \in \mathbb{R}_+^{k\times m}$ はそれぞれ要素が全て ψ、ϕ である行列を表す。また (2.33) は以下のように展開できる。

$$\begin{aligned} D_{NMF} &= tr(\boldsymbol{X} - \boldsymbol{HU})(\boldsymbol{X} - \boldsymbol{HU})^T + tr(\boldsymbol{\Psi H}^T) + tr(\boldsymbol{\Phi U}^T) \\ &= tr(\boldsymbol{XX^T}) - tr(\boldsymbol{XU^TH^T}) - tr(\boldsymbol{HUX^T}) + tr(\boldsymbol{HUU^TH^T}) + tr(\boldsymbol{\Psi H}^T) + tr(\boldsymbol{\Phi U}^T) \\ &= tr(\boldsymbol{XX^T}) - 2tr(\boldsymbol{XU^TH^T}) + tr(\boldsymbol{HUU^TH^T}) + tr(\boldsymbol{\Psi H}^T) + tr(\boldsymbol{\Phi U}^T), \end{aligned} \tag{2.34}$$

2 段目から 3 段目にかけて、$\boldsymbol{XU^TH^T}$ 及び $\boldsymbol{HUX^T}$ は正方行列であることから、$tr(\boldsymbol{A}) = tr(\boldsymbol{A^T})$ の性質を用いた。また (2.34) の第 2 項及び第 3 項は、それぞれ (2.28) の第 1 項目及び第 2 項目 (λ に関して (2.27) を代入する) を行列で表記したものに対応している。この時、(2.34) に関してそれぞれ $\boldsymbol{H}$、$\boldsymbol{U}$ で微分して 0 とおくと以下のように表せる、

$$\frac{\partial D_{NMF}}{\partial \boldsymbol{H}} = -2\boldsymbol{XU^T} + 2\boldsymbol{HUU^T} + \boldsymbol{\Psi} = 0, \tag{2.35}$$

$$\frac{\partial D_{NMF}}{\partial \boldsymbol{U}} = -2\boldsymbol{H^TX} + 2\boldsymbol{H^THU} + \boldsymbol{\Phi} = 0, \tag{2.36}$$

ただし、トレースの行列の微分として、

$$\begin{aligned} \frac{\partial tr(\boldsymbol{XAX^T})}{\partial \boldsymbol{X}} &= \boldsymbol{X}(\boldsymbol{A} + \boldsymbol{A^T}), \\ \frac{\partial tr(\boldsymbol{X^TAX})}{\partial \boldsymbol{X}} &= (\boldsymbol{A} + \boldsymbol{A^T})\boldsymbol{X}, \\ tr(\boldsymbol{ABC}) &= tr(\boldsymbol{BCA}) = tr(\boldsymbol{CAB}), \end{aligned} \tag{2.37}$$

[1]フロベニウスノルムは要素ごとの二乗和を表す。行列データ $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}^{n\times m}$ に対するフロベニウスノルムは次のように計算される。$\|x\|_F = \sqrt{x_{1,1}^2 + \cdots + x_{i,j}^2} = \sum_{i=1}^n \sum_{j=1}^m \sqrt{x_{i,j}^2}$

の性質を利用している。

ここで非負制約下における Karush-Kuhn-Tucker(KKT) 条件を用いて (2.35) 及び (2.36) 以下のように表せる、[37, 38]

$$\begin{aligned} \boldsymbol{H_{i,j}} &\geq 0, \\ \boldsymbol{\Psi_{i,j}} * \boldsymbol{H_{i,j}} &= 0, \\ \frac{\partial D_{NMF}}{\partial \boldsymbol{H}_{i,j}} &= -2\boldsymbol{X}\boldsymbol{U}_{i,j}^{\boldsymbol{T}} + 2\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}_{i,j}^{\boldsymbol{T}} + \boldsymbol{\Psi}_{i,j} \geq 0, \\ \frac{\partial D_{NMF}}{\partial \boldsymbol{H}_{i,j}} * H_{i,j} &= (-2\boldsymbol{X}\boldsymbol{U}_{i,j}^{\boldsymbol{T}} + 2\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}_{i,j}^{\boldsymbol{T}} + \boldsymbol{\Psi}_{i,j})\boldsymbol{H}_{i,j} = 0, \end{aligned} \tag{2.38}$$

$$\begin{aligned} \boldsymbol{U_{i,j}} &\geq 0, \\ \boldsymbol{\Phi_{i,j}} * \boldsymbol{U_{i,j}} &= 0, \\ \frac{\partial D_{NMF}}{\partial \boldsymbol{U}_{i,j}} &= -2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X}_{i,j} + 2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U}_{i,j} + \boldsymbol{\Phi}_{i,j} \geq 0, \\ \frac{\partial D_{NMF}}{\partial \boldsymbol{U}_{i,j}} * U_{i,j} &= (-2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X}_{i,j} + 2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U}_{i,j} + \boldsymbol{\Phi}_{i,j})\boldsymbol{U}_{i,j} = 0. \end{aligned} \tag{2.39}$$

(2.38) 及び (2.39) は要素ごとの演算を表している。これらを整理して、以下の計算アルゴリズムが得られる。

$$\begin{aligned} \boldsymbol{H}_{i,j} &\leftarrow \boldsymbol{H}_{i,j}\frac{\boldsymbol{X}\boldsymbol{U}_{i,j}^{\boldsymbol{T}}}{\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}_{i,j}^{\boldsymbol{T}}}, \\ \boldsymbol{U}_{i,j} &\leftarrow \boldsymbol{U}_{i,j}\frac{\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X}_{i,j}}{\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U}_{i,j}}. \end{aligned} \tag{2.40}$$

2.3 次元縮約法と正則化モデル

2.3.1 正則化モデルについて

Tibshirani(1996) [4] は回帰分析においてよく使用される最小二乗法を用いた推定方法について、理論の観点ではデータとの分散を小さくできる一方で、モデルとのバイアスを大きくする点、説明変数が多くて解釈がしづらい場合がある点などを課題に挙げ、回帰係数の一部を 0 にすることで予測精度を向上させ、解釈を容易にするための解析手法として、L_1 正則化を付与する正則化モデルを提案した。

目的変数を $\boldsymbol{y} = [y_1, \cdots, y_n]^T \in \mathbb{R}^{n\times 1}$、説明変数であるデータ行列を $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}^{n\times m}$、回帰係数を $\boldsymbol{\beta} = [\beta_1, \cdots, \beta_m]^T \in \mathbb{R}^{m\times 1}$ とした時、線形回帰における正則化モデルは以下のように定式化できる、[39]

$$\underset{\boldsymbol{\beta}}{minimize}\ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 \qquad sub\,ject\ to\ \|\boldsymbol{\beta}\|_p \le t, \tag{2.41}$$

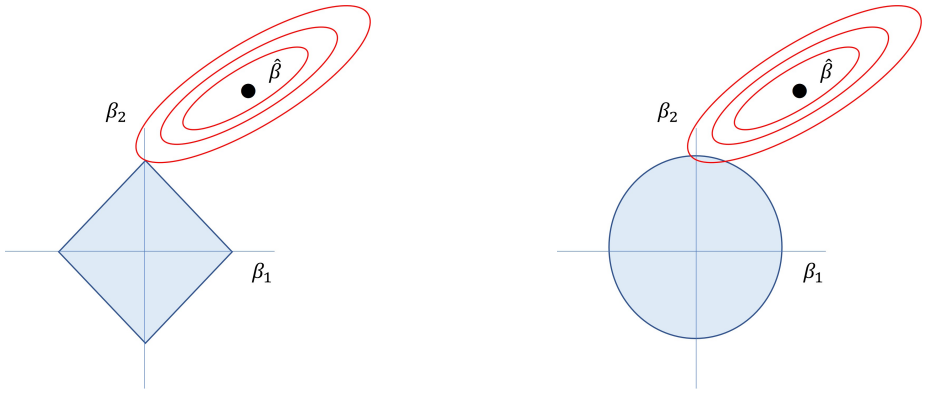
$\|\ \ \|_F$ はフロベニウスノルム、$\|\boldsymbol{\beta}\|_p$ は L_p ノルムを表す[2]。ここで $\boldsymbol{\beta}$ に対する L_p ノルムは以下のように表される、

$$\|\boldsymbol{\beta}\|_p = \sqrt[p]{|\beta_1|^p + |\beta_2|^p + \cdots + |\beta_m|^p} = \sum_j^m \sqrt[p]{|\beta_j|^p}. \tag{2.42}$$

図 2.1 は L_pノルム の中でも特に議論される L_2ノルム ($p = 2$) 及び L_1ノルム ($p = 1$) に関して、二乗誤差項によって表される凸関数との関係について示している。図 2.1a のひし形は β_1、β_2 の平面上において表される制約条件 $\|\boldsymbol{\beta}\|_1 = |\beta_1| + |\beta_2| \le t$ であり、図 2.1b の円は同条件における $\|\boldsymbol{\beta}\|_2^2 = \beta_1^2 + \beta_2^2 \le t^2$ である。また両図に共通する楕円は二乗誤差項 $\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2$ の β_1、β_2 平面上において表される解空間であり、$\hat{\boldsymbol{\beta}}$ は誤差を最小にする最小二乗推定量である。この時図 2.1a では、二乗誤差項による凸関数がひし形の制約条件下における $\beta_1 = 0$、または $\beta_2 = 0$ となる点で接しやすい、すなわち回帰係数ベクトル $\boldsymbol{\beta}$ の値を 0 にすることが可能であるため、Tibshirani(1996) が懸念する課題が解消できる。このように、ベクトルまたは行列内の係数を 0 にできる性質を「スパース性」と呼び、スパースはこのような係数が 0 である状態のことを指す。特に L_1 ノルムを適用し、スパース化を図る時の制約は一般に L_1 罰則 (L_1 penalty) または L_1 正則化 (L_1 regularization) 等と呼ばれ、これを付与した回帰モデルは Lasso 回帰 (Least absolute shrinkage and selection operator) と呼ばれている。

図 2.1b では二乗誤差項による凸関数が、円の制約条件下における $\beta_1 = 0$、または $\beta_2 = 0$ で接することはひし形の L_1 ノルムと比べて困難である。すなわち L_2 ノルムの制約条件は L_1 ノルムほど回帰係数を 0 にできるスパース効果は薄いとされる。L_2 ノルムを適用する制約は一般に L_2 罰則、L_2 正則化等とよばれ、これを付与した回帰モデルは Ridge 回帰 (Ridge regression) と呼ばれている。Ridge 回帰分析の提案者である Hoerl and Kennard(1970) [40] は Ridge 回帰の特徴として、相関が高い説明変数がある場合に特に有効であり、最小二乗推定量では不安定になる (過学習等によりノイズの影響を大きく受ける場合など）解を安定的に求めることができるとしている。

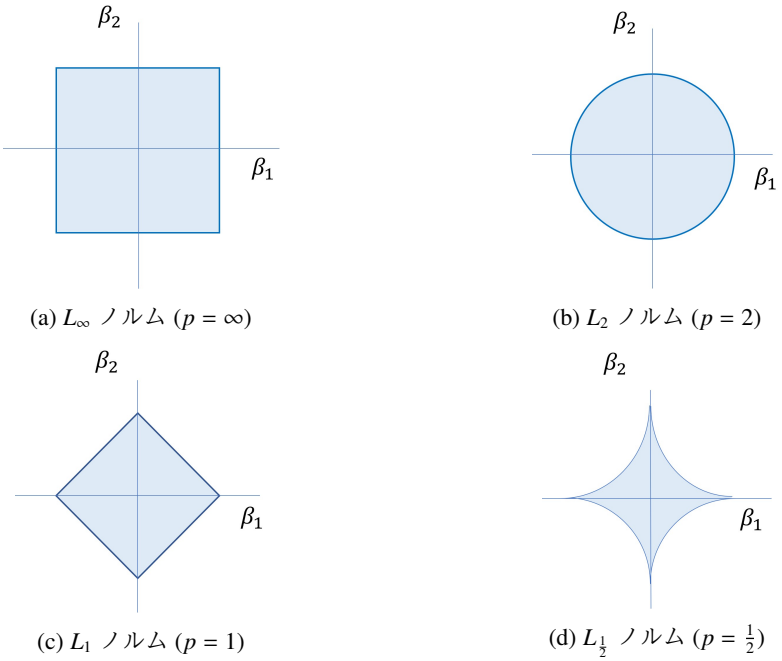
[2] $\|\ \ \|_F$ と $\|\ \ \|_2$ は同等である



(a) L_1 ノルム　　(b) L_2 ノルム

図 2.1: β_1, β_2 における正則化項のスパース効果

また図 2.2 は、L_p ノルムの p を変化させたときの正則化項を示している。図 2.1 と同様に、正方形で表される L_∞ ノルムは、二乗誤差項の凸関数が $\beta_1 = 0$ または $\beta_2 = 0$ の点と接することが困難であるため、スパース効果は極めて薄い。他方、$L_{\frac{1}{2}}$ ノルムは、$\beta_1 = 0$ または $\beta_2 = 0$ の点と接することが L_1 ノルムより容易であり、さらなるスパース効果が期待できる。



(a) L_∞ ノルム ($p = \infty$)　　(b) L_2 ノルム ($p = 2$)

(c) L_1 ノルム ($p = 1$)　　(d) $L_{\frac{1}{2}}$ ノルム ($p = \frac{1}{2}$)

図 2.2: β_1, β_2 における L_p 正則化項のイメージ

(i) L_1 ノルムによる正則化モデル

Lasso 回帰における回帰係数 $\hat{\boldsymbol{\beta}}$ の推定について議論する。以下では川野・松井・廣瀬 (2018) [41] を参考にしている。

L_1 正則化項を付与した最小二乗モデルは以下のように表される、

$$\underset{\boldsymbol{\beta}}{minimize}\, S_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{minimize}\, \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda\|\boldsymbol{\beta}\|_1, \tag{2.43}$$

ここで λ は L_1 正則化におけるラグランジュ乗数である。また、$\|\boldsymbol{\beta}\|_1 = \sum(|\beta_1| + \cdots + |\beta_m|)$ であることから、$\beta_i (i = 1, \cdots, m)$ において微分が不可能である。そのため、Lasso における $\hat{\boldsymbol{\beta}}$ の推定は

(1) データ行列が直交である場合

(2) データ行列が直交以外の場合

の 2 つのアプローチから解析されている。以下ではこれら 2 つについて議論する。

(1) データ行列が直交である場合

(2.43) より、$\boldsymbol{\beta}$ に関して最小化を行う。ここで β_j に関して劣勾配 [42](詳細は「付録 1」) を求めることで以下の式が得られる,

$$\frac{\partial S_\lambda(\boldsymbol{\beta})}{\partial \beta_j} = -\frac{1}{n}\boldsymbol{x}_j^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda d_j = 0, \tag{2.44}$$

ただし、d_j は β_j における L_1ノルム $\|\ \ \|_1$ の劣微分の要素であり、以下を満たす。

$$d_j \in \begin{cases} -1, & (\beta_j < 0), \\ [-1, 1], & (\beta_j = 0), \\ 1. & (\beta_j > 0), \end{cases} \tag{2.45}$$

また、直交制約 $\boldsymbol{X^T X} = \boldsymbol{I_m}$ より、(2.44) は β_j は以下の通りとなる、

$$\beta_j = \frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{y} - \lambda d_j, \tag{2.46}$$

すなわち、(2.45) と (2.46) から、$\hat{\beta}_j$ は以下のように表せる、

$$\hat{\beta}_j = \begin{cases} \frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{y} - \lambda. & (\boldsymbol{x}_j^T\boldsymbol{y} > n\lambda), \\ 0. & (|\boldsymbol{x}_j^T\boldsymbol{y}| < n\lambda), \\ \frac{1}{n}\boldsymbol{x}_j^T\boldsymbol{y} + \lambda. & (-\boldsymbol{x}_j^T\boldsymbol{y} > n\lambda). \end{cases} \tag{2.47}$$

(2) データ行列が直交以外の場合

データ行列が直交ではない場合、凸最適化問題 (特に 2 次計画問題) として推定量を求める必要がある。Lasso を解くための計算アルゴリズムについては複数提案されているが、ここでは様々な種類の正則化にも

汎用的に適用可能な、交互方向乗数法 (Alternating Determination Method of Multipliers:ADMM) について述べる。他の計算アルゴリズムとして最小角回帰 (Least Angle Regression:LARS) [43] や座標降下法 (Coordinate descent method) [44] などがある。

交互方向乗数法はラグランジアンの考えに基づき、最適解を求める方法であり、特に大規模な最適化問題を解く場合に有効であるとされる [45]。パラメータ $\boldsymbol{\beta}$ に関して最小化する Lasso 回帰モデルを以下のように設定する、

$$\underset{\boldsymbol{\beta},\boldsymbol{\gamma}}{minimize}\ \{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\lambda\|\gamma\|_1\},\qquad subject\ to\ \boldsymbol{\beta}=\boldsymbol{\gamma}, \tag{2.48}$$

ただし、$\boldsymbol{\gamma}\in\mathbb{R}^{m\times 1}$ である。この最小化問題に対する拡張ラグランジアンは次のように表せる、

$$L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{u})=\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_F^2+\lambda\|\gamma\|_1+\boldsymbol{u}^T(\boldsymbol{\beta}-\boldsymbol{\gamma})+\frac{\rho}{2}\|\boldsymbol{\beta}-\boldsymbol{\gamma}\|_F^2, \tag{2.49}$$

ここで $\boldsymbol{u}\in\mathbb{R}^{n\times 1}$ はラグランジュ未定乗数、ρ は正のハイパーパラメータである。交互方向乗数法では (2.49) から $\boldsymbol{\beta}$ と $\boldsymbol{\gamma}$ に関して微分して 0 とおき、$\boldsymbol{u}$ に関して $\boldsymbol{\beta}$ 及び $\boldsymbol{\gamma}$ を用いた更新を行う。これらを繰り返し行うことで、$\hat{\boldsymbol{\beta}}$ を推定する。(2.49) を整理すると以下のように表せる、

$$L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{u})=\frac{1}{2n}(\boldsymbol{y}^T\boldsymbol{y}-2\boldsymbol{y}^{\boldsymbol{T}}\boldsymbol{X}\boldsymbol{\beta}+\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{X}^{\boldsymbol{X}}\boldsymbol{\beta})+\lambda\|\gamma\|_1+\boldsymbol{u}^T(\boldsymbol{\beta}-\boldsymbol{\gamma})+\frac{\rho}{2}(\boldsymbol{\beta}^{\boldsymbol{T}}\boldsymbol{\beta}-2\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{\beta}+\boldsymbol{\gamma}^{\boldsymbol{T}}\boldsymbol{\gamma}). \tag{2.50}$$

(2.50) より、$\boldsymbol{\beta}$ に関して微分して 0 とおくと以下のように表せる、

$$\begin{aligned}\hat{\boldsymbol{\beta}}=\frac{\partial L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{u})}{\partial\boldsymbol{\beta}}&=\frac{1}{n}(-\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}+\boldsymbol{X}^{\boldsymbol{X}}\boldsymbol{\beta})+\boldsymbol{u}+\rho(\boldsymbol{\beta}-\boldsymbol{\gamma})=0,\\&=(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}+n\rho\boldsymbol{I}_{\boldsymbol{m}})\boldsymbol{\beta}=(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y})-n\boldsymbol{u}+n\rho\gamma,\\&=(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}+n\rho\boldsymbol{I}_{\boldsymbol{m}})^{-1}\{\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}+n\rho(\gamma-\frac{1}{\rho}\boldsymbol{u})\}.\end{aligned} \tag{2.51}$$

次に (2.50) より、$\boldsymbol{\gamma}$ に関して微分して 0 とおく。ただし、γ には L_1 正則化項が含まれるため、(2.44) から (2.47) を推定する手順と同様に計算する。そのため、γ_j に関して劣勾配を求めて 0 とおくと以下のように表せる、

$$\frac{\partial L_\rho(\boldsymbol{\beta},\boldsymbol{\gamma},u_j)}{\partial\gamma_j}=-u_j-\rho\beta_j+\rho\gamma_j+\lambda d_j=0, \tag{2.52}$$

ただし、d_j は γ_j における L_1ノルム $\|\ \ \|_1$ の劣微分の要素であり、以下を満たす。

$$d_j\in\begin{cases}-1, & (\gamma_j<0),\\ [-1,1], & (\gamma_j=0),\\ 1, & (\gamma_j>0),\end{cases} \tag{2.53}$$

よって、γ_j に関して以下のように表せる。

$$\hat{\gamma_j} = \begin{cases} \beta_j + \frac{1}{\rho}u_j - \frac{\lambda}{\rho}, & (\beta_j + \frac{1}{\rho}u_j > \frac{\lambda}{\rho}), \\ 0, & (\beta_j + \frac{1}{\rho}u_j = \frac{\lambda}{\rho}), \\ \beta_j + \frac{1}{\rho}u_j + \frac{\lambda}{\rho}, & (\beta_j + \frac{1}{\rho}u_j < \frac{\lambda}{\rho}). \end{cases} \tag{2.54}$$

(2.51) 及び (2.54) の結果に基づいて、交互方向乗数法の更新式は以下のようなステップで計算される。
STEP 1．初期値 $\boldsymbol{\beta^{(0)}}$、$\boldsymbol{\gamma^{(0)}}$、$\boldsymbol{u^{(0)}}$ を設定する。
STEP 2．$\boldsymbol{\beta}$ を次のように更新する。((2.51) より)

$$\boldsymbol{\beta}^{(t+1)} = (\boldsymbol{X^T X} + n\rho \boldsymbol{I_m})^{-1}\{\boldsymbol{X^T y} + n\rho(\boldsymbol{\gamma}^{(t)} - \frac{1}{\rho}\boldsymbol{u}^{(t)})\} \tag{2.55}$$

STEP 3．$\boldsymbol{\gamma}$ を次のように更新する。((2.54) より)

$$\hat{\gamma_j^{(t+1)}} = \begin{cases} \beta_j^{(t+1)} + \frac{1}{\rho}u_j^{(t)} - \frac{\lambda}{\rho}, & (\beta_j^{(t+1)} + \frac{1}{\rho}u_j^{(t)} > \frac{\lambda}{\rho}), \\ 0, & (\beta_j^{(t+1)} + \frac{1}{\rho}u_j^{(t)} = \frac{\lambda}{\rho}), \\ \beta_j^{(t+1)} + \frac{1}{\rho}u_j^{(t)} + \frac{\lambda}{\rho}, & (\beta_j^{(t+1)} + \frac{1}{\rho}u_j^{(t)} < \frac{\lambda}{\rho}), \end{cases} \tag{2.56}$$

STEP 4．$\boldsymbol{u}$ を次のように更新する。

$$\boldsymbol{u}^{(t+1)} = \boldsymbol{u}^{(t)} + \rho(\boldsymbol{\beta}^{(t+1)} - \boldsymbol{\gamma}^{(t+1)}). \tag{2.57}$$

STEP 5．STEP 2 から STEP 4 を収束するまで繰り返し、$\boldsymbol{\gamma}$ を出力する。

(ii) L_2 ノルムによる正則化モデル

Ridge 回帰における回帰係数 $\hat{\boldsymbol{\beta}}$ の推定について議論する。L_2 正則化項を付与した最小二乗モデルは以下のように表される。

$$\underset{\boldsymbol{\beta}}{minimize}\, S_\lambda(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}}{minimize}\, \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X\beta}\|_F^2 + \lambda\|\boldsymbol{\beta}\|_2^2, \tag{2.58}$$

ここで λ は L_2 正則化におけるラグランジュ未定乗数項である。Ridge 回帰では正則化項が要素の二乗和で与えられるため、$\boldsymbol{\beta}$ に関して微分が可能である。そのため、(2.58) を $\boldsymbol{\beta}$ に関して微分して 0 とおく。

$$\frac{\partial S_\lambda(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -\frac{1}{n}\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\boldsymbol{\beta} = 0. \tag{2.59}$$

(2.59) を $\boldsymbol{\beta}$ について整理すると、$\hat{\boldsymbol{\beta}}$ は以下のように表せる、

$$
\begin{aligned}
(\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + n\lambda\boldsymbol{I})\boldsymbol{\beta} &= \boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y} \\
\hat{\boldsymbol{\beta}} &= (\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + n\lambda\boldsymbol{I}_{\boldsymbol{m}})^{-1}\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{y}.
\end{aligned}
\tag{2.60}
$$

(2.60) より、最小二乗法において行列 $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X}$ の逆行列が計算できないような場合においても、Ridge 回帰では $\boldsymbol{X}^{\boldsymbol{T}}\boldsymbol{X} + n\lambda\boldsymbol{I}_{\boldsymbol{m}}$ が任意の $\lambda > 0$ に対し、正則[3]になる。

また、λ の値を大きくすると、(2.58) の右辺第 2 項が主に最小化されることから、(2.60) の推定量の全ての成分が 0 方向に縮小される。一方、λ の値を小さくすると、(2.58) の右辺第 1 項が主に最小化されるため、(2.60) の推定量は最小二乗推定量に近くなる。この推定量は縮小推定量と呼ばれている。

(iii) その他の正則化モデル

上記で正則化項が L_1 ノルムに対応する Lasso 回帰と L_2 ノルムに対応する Ridge 回帰について述べたが、これらのモデルには多くの派生系がある。例えば Lasso と Ridge の両方を混合させたエラスティックネット (Elastic net) がある。提案者の Zou and Hastie(2005) [46] は Lasso の問題点として、相関の高い 2 つの説明変数が目的変数に関係している場合、Lasso での推定ではどちらか一方の変数のみが選択される傾向があることを指摘し、相関の高いデータに対し解を安定化できる Ridge 回帰と組み合わせることで、両手法の長所を活かした解析ができるとしている。Elastic net 回帰は以下のように定式化されている、

$$
\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda_1\|\boldsymbol{\beta}\|_1 + \lambda_2\|\boldsymbol{\beta}\|_2^2, \tag{2.61}
$$

ただし λ は非負値のハイパーパラメータである。また $\lambda_1 = 0$ の時は Ridge、$\lambda_2 = 0$ の時は Lasso と同等の問題となる。これらの手法は、遺伝子データ解析のような、個々の説明変数ではなく、複数の説明変数群と目的変数との関係性に着目する分析例や、説明変数がサンプルサイズよりも大きいデータなどでの活用例がある。

Tibshirani et al. (2005) [47] は説明変数に順序関係があり、隣接する説明変数のいくつかは目的変数に同程度の寄与が想定される場合に利用できる、連結 Lasso(Fused Lasso) を提案した。連結 Lasso は以下のように定式化されている,

$$
\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_F^2 + \lambda_1\sum_{j=1}^{m}\|\beta_j\|_1 + \lambda_2\sum_{j=2}^{m}\|\beta_j - \beta_{j-1}\|_1, \tag{2.62}
$$

ただし、$\lambda_1 > 0$、$\lambda_2 > 0$ は共に正則化項のハイパーパラメータである。(2.62) より、第 3 項は隣接する係数の階差の絶対値に対して L_1 正則化を付与したものであり、隣り合う 2 つの係数の推定値が等しくなるように計算される。Tibshirani et al.(2005) [47] は健診データに基づくがん患者または健康患者の推定にこの手法と通常の Lasso 回帰を適用し、解析結果の比較を行っている。

他に、Fan and Li(2001) [48] は Lasso 回帰モデルよりさらにスパースな解が得られやすくなる正則化項として、SCAD(Smoothly Clipped Absolute Devation) を提案している。また Lasso の改良として、Zou(2006) [49] は適応型 Lasso(Adaptive Lasso) を、Tibshirani and Taylor(2011) [50] は一般化 Lasso(Genelized Lasso) を提案している。他にも連結 Lasso の拡張として、She(2010) [51] はパラメータすべての組み合わせの差分に制約を課すクラスター Lasso(Clustered Lasso) を提案している。

[3] このような罰則項の導入により、行列を正則化させることで最小二乗解を安定させる効果から、この罰則項を正則化項と呼び、このような処理を (Tikhonov の) 正則化法と呼ぶ。

2.3.2 スパース主成分分析

スパース主成分分析は、通常の主成分分析と比べて、推定される固有ベクトルの要素において 0 が多くなるように固有値分解を行う解析手法である。

スパース主成分分析に関するアルゴリズムは大きく 2 種類提案されている。Jolliffe et al.(2003) [52] は通常の主成分分析における主成分の導出方法に L_1 正則化項を付与した「SCoTLASS」(Simplified Component Technique- LASSO)) を提案した。これは第 k 主成分の分散に対し、直交制約と L_1 正則化項を付与したモデルであり、下記のように定義されている。

$$\begin{aligned} V_{z_k} &= \boldsymbol{a_k^T R a_k}, \\ \textit{subject to} \quad \boldsymbol{a_k^T a_k} &= 1\ ,\ \boldsymbol{a_h^T a_k} = 0\ (h < k) \quad \textit{and} \quad \sum_{j=1}^{p} |a_{k,j}| \leq t, \end{aligned} \tag{2.63}$$

ここで t はハイパーパラメータを表す。その他の設定は (2.3) と同様である。ただし Jolliffe et al.(2003) は (2.63) を使用し、解析を行った結果、スパース効果があまり得られなかったと結論付けている。

一方、Zou and Hastie(2006) [53] は回帰分析におけるスパースモデルを行列データに拡張するアプローチから主成分を導出する、「SPCA」(Sparse Principal Component Analysis) を提案した。このモデルでは、直交制約とスパース制約を同時に満たしつつ解析的に主成分を推定することが困難であった問題を解決し、ハイパーパラメータの設定によって大きなスパース効果を得ることができる。以下では Zou and Hastie(2006) [53] や Erichson(2020) [54] らを参考にする。

観測データ行列 $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}^{n\times m}$、固有ベクトルからなる行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_k}] \in \mathbb{R}^{m\times k}$、$\boldsymbol{A} \fallingdotseq \boldsymbol{B}$ となる行列を $\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_k}] \in \mathbb{R}^{m\times k}$,($k$ は縮約する次元数) とし、二乗誤差最小化に基づく主成分の計算アルゴリズムを以下のように設定している。

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \underset{\boldsymbol{A},\boldsymbol{B}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\boldsymbol{T}}\|_F^2 + \alpha\|\boldsymbol{B}\|_F^2 + \beta\|\boldsymbol{B}\|_1, \qquad \textit{subject to } \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \boldsymbol{I}, \tag{2.64}$$

ここで、$\|\ \ \|_F$ はフロベニウスノルム、$\|\ \ \|_1$ は L_1 ノルムとする。Zou and Hastie(2006) は、$\boldsymbol{A} \fallingdotseq \boldsymbol{B}$ となる行列を用意し、$\boldsymbol{A}$ に直交制約を、$\boldsymbol{B}$ にスパース制約を付与し、これらを交互に計算することで、$\boldsymbol{A}, \boldsymbol{B}$ の最小化する工夫を行っている。行列 $\boldsymbol{A}$,$\boldsymbol{B}$ の更新は、以下の通り交互に計算する。

STEP 1．初期値として、$\boldsymbol{X}$ に通常の主成分分析を行ったときの固有ベクトルの行列 $\boldsymbol{V}[1:k]$ を $\boldsymbol{A}$ とおく。
STEP 2．$\boldsymbol{A}$ を固定し、(2.64) を $\boldsymbol{B}$ に関して計算する。すなわち以下について解く。

$$\hat{\boldsymbol{B}} = \underset{\boldsymbol{B}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\boldsymbol{T}}\|_F^2 + \alpha\|\boldsymbol{B}\|_F^2 + \beta\|\boldsymbol{B}\|_1, \qquad \textit{subject to } \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \boldsymbol{I} \quad .$$

STEP 3．$\boldsymbol{B}$ を固定し、特異値分解 $\boldsymbol{X}^{\boldsymbol{X}}\boldsymbol{B} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^{\boldsymbol{T}}$ を計算し、$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{V}^{\boldsymbol{T}}$ とおく。
STEP 4．STEP2、STEP3 を収束するまで繰り返す。
5．収束後、$\boldsymbol{B}$ の列ベクトルに対し、$\dfrac{\boldsymbol{b_j}}{\|\boldsymbol{b_j}\|_F}$ によって標準化する。

本稿ではこうして得られた $\boldsymbol{B}$ を固有ベクトルの行列として使用している。この時、$\boldsymbol{B}$ はハイパーパラメータを大きくするほど、固有ベクトルの行列がスパースとなる一方で、直交性は失われていく。(第 6 章参考)

2.3.3 スパース非負値行列因子分解

スパース非負値行列因子分解 (Sparse Non-negative Matrix Factorization,S-NMF) は、通常の非負値行列因子分解で得られる基底ベクトルまたは表現ベクトルに対し、その要素に 0 が多くなるように行列分解を行う解析手法である。

同手法について、Hoyer(2002) [55] は非負値行列因子分解の行列分解モデルに対し、表現行列 $\boldsymbol{U}$ に L_1 の罰則を付与した、以下のモデルを提案した、

$$D_{S-NMF}(\boldsymbol{H},\boldsymbol{U}) = \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{U}\|_F^2 + \lambda\|\boldsymbol{U}\|_1. \tag{2.65}$$

以降のスパース非負値行列因子分解モデルの研究では、基底行列 $\boldsymbol{H}$ または表現行列 $\boldsymbol{U}$ 及びその両方に L_1 ノルム、L_2 ノルム、L_1 ノルムと L_2ノルム を付与するなど、正則化項の選択及びその付与の仕方について組み合わせ的にモデルが作成された [56,57]。これらは Cichocki et al.(2007) [58] や Cichocki et al.(2009) [59] によって体系化されている。

スパース非負値行列因子分解の計算アルゴリズムは、通常の非負値行列因子分解と概ね同様にして算出できる。ここでは特に基底行列 $\boldsymbol{H}$ に L_2 ノルムを、表現行列 $\boldsymbol{U}$ に L_1 ノルムを付与した場合について述べる (L_2 ノルム、L_1 ノルムを入れ替えた場合においてもアルゴリズム導出の過程に大差がないため)。

観測データ行列を $\boldsymbol{X} = [\boldsymbol{x_1},\cdots,\boldsymbol{x_m}] \in \mathbb{R}_+^{n\times m}$、基底行列を $\boldsymbol{H} = [\boldsymbol{h_1},\cdots,\boldsymbol{h_k}] \in \mathbb{R}_+^{n\times k}$, 表現行列を $\boldsymbol{U} = [\boldsymbol{u_1},\cdots,\boldsymbol{u_m}] \in \mathbb{R}_+^{k\times m}$ とするとき、目的関数を D_{S-NMF} とすると、以下のように表せる、

$$\begin{aligned} D_{S-NMF} &= \|\boldsymbol{X} - \boldsymbol{H}\boldsymbol{U}\|_F^2 + \mu\|\boldsymbol{H}\|_F^2 + \lambda\|\boldsymbol{U}\|_1 + tr(\boldsymbol{\Psi}\boldsymbol{H}^T) + tr(\boldsymbol{\Phi}\boldsymbol{U}^T) \\ &= tr(\boldsymbol{X}\boldsymbol{X}^{\boldsymbol{T}}) - 2tr(\boldsymbol{X}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{H}^{\boldsymbol{T}}) + tr(\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}}\boldsymbol{H}^{\boldsymbol{T}}) + \mu\|\boldsymbol{H}\|_F^2 + \lambda\|\boldsymbol{U}\|_1 + tr(\boldsymbol{\Psi}\boldsymbol{H}^T) + tr(\boldsymbol{\Phi}\boldsymbol{U}^T), \end{aligned} \tag{2.66}$$

ただし、μ,λ はハイパーパラメータである。$\boldsymbol{H}$ 及び $\boldsymbol{U}$ に関して微分して 0 とおく。

$$\frac{\partial D_{S-NMF}}{\partial H} = -2\boldsymbol{X}\boldsymbol{U}^{\boldsymbol{T}} + 2\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}} + 2\mu\boldsymbol{H} + \boldsymbol{\Psi}, \tag{2.67}$$

$$\frac{\partial D_{S-NMF}}{\partial U} = -2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X} + 2\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U} + \lambda(\boldsymbol{1}_{k,m}) + \boldsymbol{\Phi}, \tag{2.68}$$

ただし、$\boldsymbol{1}_{k,m}$ は $k\times m$ で要素が全て 1 の行列である。KKT 条件を用いてスパース非負値行列因子分解の計算アルゴリズムは以下のように導出できる。

$$\begin{aligned} ((-\boldsymbol{X}\boldsymbol{U}^{\boldsymbol{T}})_{i,j} + (\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}})_{i,j} + \mu\boldsymbol{H}_{i,j})\boldsymbol{H}_{i,j} = 0, \\ ((-\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H})_{i,j} + (\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U})_{i,j} + \frac{\lambda}{2}\boldsymbol{1}_{i,j})\boldsymbol{U}_{i,j} = 0, \end{aligned} \tag{2.69}$$

$$\begin{aligned} \boldsymbol{H}_{i,j} &\leftarrow \boldsymbol{H}_{i,j}\frac{(\boldsymbol{X}\boldsymbol{U}^{\boldsymbol{T}})_{i,j}}{(\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}})_{i,j} + \mu\boldsymbol{H}_{i,j}}, \\ \boldsymbol{U}_{i,j} &\leftarrow \boldsymbol{U}_{i,j}\frac{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X})_{i,j}}{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U})_{i,j} + \frac{\lambda}{2}\boldsymbol{1}_{i,j}}. \end{aligned} \tag{2.70}$$

第3章　非負値行列因子分解に関する応用研究

3.1　非負値行列因子分解法を用いた地域産業特性の抽出

3.1.1　概要

本節 [60] では非負値行列因子分解を産業連関表の内生部門表データに適用することで地域産業の特徴を解析した。非負値行列因子分解法の適用理由として、産業連関表が財・サービスの取引で発生する金額を計上したデータであり、全て非負値で表されている点や行データ及び列データが定義される産業数の数だけ存在する高次元ビックデータであるため、非負値の成分からなる 2 つの行列に分解でき、また次元縮約法としての性質を持つ非負値行列因子分解が適切であると考え、解析を行った。

本研究における使用データとして、47 都道府県分の 2011 年産業連関表内生部門表ひな形 (13 部門 ×13 部門) データから 1 つの行列データを作成し、分析を行った。具体的には産業連関表のひな形行列を 1×169 のベクトルに変換し、それを 47 都道府県分結合することで、「47 都道府県 ×169 産業間取引」の行列データを使用データとした。このデータ構造から基底行列 $\boldsymbol{H}$ では都道府県の特徴、表現行列 $\boldsymbol{U}$ では産業間取引の特徴が抽出でき、両特徴を併せて、「ある地域 (または地域群) の産業間取引構造の共通な特徴はどのようなものか」について解析する。

分析結果として、縮約する次元数に対応する基底数を 3 とした場合、日本の産業取引構造は愛知県を筆頭とした製造圏による「製造業→製造業」の取引構造、東京都をはじめとした大都市における「サービス業」を中心とした第 3 次産業に関する取引構造、神奈川県や千葉県、岡山県などの製油所機能を有する地域をはじめとした「鉱業→製造業」に関する特徴の 3 つに分解された。

3.1.2　使用データと設定について

(i) 使用データについて

2011 年産業連関表ひな形 (13×13 部門) を各都道府県の Web サイトから取得し、使用データとした。ただし、部門数が異なる地域や、ひな型を作成していない地域に関しては、部門を分割・統合し、13 部門[1]に統一した。単位は百万円である。

47 都道府県の産業連関表 13×13（ひな型）行列を 1 つの行列に集約するため、ひな形の行ごとにデータを切り取り、番号の若い行から順に横方向に結合し、1×169 ベクトルを作成した。また 47 都道府県分のデータを縦方向に結合させ、47×169 行列を作成することで、47 都道府県の産業連関表における内生部門データを 1 つの行列に集約し、使用データ $\boldsymbol{X}$ とした。このデータ構造から基底行列 $\boldsymbol{H}$ では都道府県の特徴、表現行列 $\boldsymbol{U}$ では産業間取引の特徴が抽出でき、両特徴を併せて、「ある地域 (または地域群) の産業間取引構造の共通な特徴はどのようなものか」について解析する。

[1] 13 部門の内訳は左 (または上、第 1 次産業) から順に「農林水産業」・「鉱業」・「製造業」・「建設業」・「電気・ガス・水道業」・「商業」・「金融・保険業」・「不動産業」・「運輸附帯業」・「情報通信業」・「公務」・「サービス業」・「分類不明」である。なお集計年度によって名称に違いがある。詳細は総務省 [5] を参照のこと

(ii) 分析手法の設定

上記で作成したデータに対して、非負値行列因子分解を行う。ただし、観測行列 $\boldsymbol{X}$ に零ベクトルが含まれると、計算アルゴリズムの計算途中で分母に 0 を取り、正常に起動しなくため、当該産業の取引 (公務が該当する) は分析結果に影響を及ぼさない程度に限りなく 0 に近い数字に置き換えた。

観測データ行列を $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}_+^{n \times m}$、分解する 2 つの行列を基底行列 $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{n \times k}$, 表現行列 $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}_+^{k \times m}$ とした時、以下のアルゴリズムを用いて基底行列 $\boldsymbol{H}$ 及び表現行列 $\boldsymbol{U}$ を算出した、

$$\boldsymbol{H}_{i,j} \leftarrow \boldsymbol{H}_{i,j} \frac{(\boldsymbol{X}\boldsymbol{U}^{\boldsymbol{T}})_{i,j}}{(\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}})_{i,j}}, \qquad \boldsymbol{U}_{i,j} \leftarrow \boldsymbol{U}_{i,j} \frac{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{X})_{i,j}}{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U})_{i,j}}, \quad \textit{subject to} \sum_m \boldsymbol{u_m} = \boldsymbol{1}, \tag{3.1}$$

ただし、(3.1) は要素ごとの計算を行っている。また、本データの構造から $n = 47, m = 169$ であり、基底数は日本の産業構造の大きな特徴を把握することを目的に、k を変化させて調べた結果、$k = 3$ が最も本解析に合う次元であると判断した。また (3.1) 式より、表現ベクトルの各行和を 1 とする制約を付与することで、表現ベクトル (産業間取引の特徴) を「重み」ベクトルとして解釈する。そのため、観測行列 $\boldsymbol{X}$ を近似するための量的構造は基底行列 $\boldsymbol{H}$ に集約されることになる。

計算アルゴリズムにおいて、局所解に陥ることを避けるため、初期値の行列を乱数で与えるものとし、1 から 10 の一様乱数で発生させたものを使用した。また行列 $\boldsymbol{H}$ と $\boldsymbol{U}$ の導出過程のシミュレーション回数は 500 回とし、この計算後に得られた行列と更新前の行列との差の変化率が、$\boldsymbol{H}, \boldsymbol{U}$ ともに 0.001 を下回ったところでアルゴリズムを終了し、その時の二乗誤差（RSS）を最適な基底行列 $\boldsymbol{H}$、表現行列 $\boldsymbol{U}$ とした。観測行列 $\boldsymbol{X}$ と基底行列 $\boldsymbol{H}$、表現行列 $\boldsymbol{U}$ の分解イメージは図 3.1 の通りである。

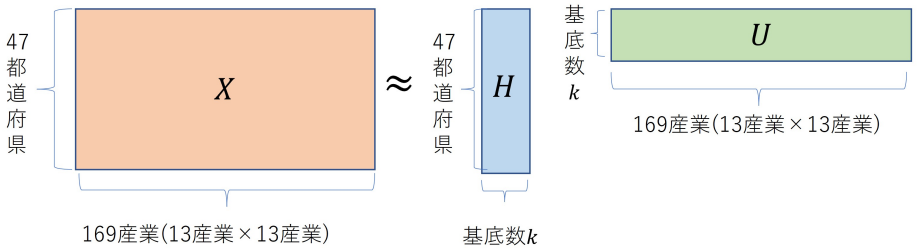


図 3.1: 産業連関表に対する NMF 分解のデザイン

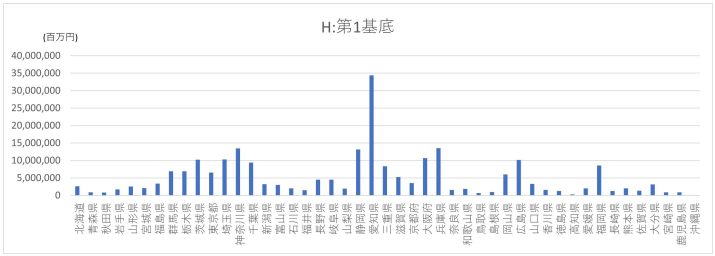
3.1.3 分析結果

図 3.2 から図 3.4 はいずれも上記のデータに非負値行列因子分解を適用した結果、3 つの基底に分解したときの、それぞれの基底行列 $\boldsymbol{H}$ と表現行列 $\boldsymbol{U}$ で抽出された特徴のペアであり、図 a が基底行列 $\boldsymbol{H}$ から得られた基底ベクトル、図 b が表現行列 $\boldsymbol{U}$ から得られた表現ベクトルである。

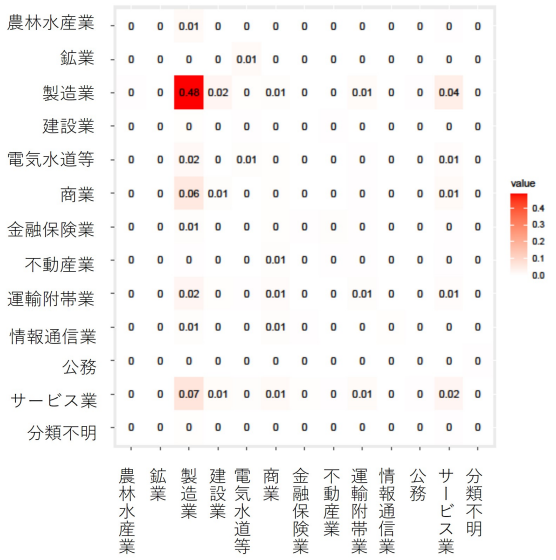
図 a の基底ベクトルでは都道府県に関する特徴が抽出され、その基底の特徴を大きく捉えた地域ほど高い値をとる。また単位は使用データと同じ百万円である。ただし行列分解近似の性質から、複数の地域の特徴が含まれているため、実際の産業連関表で記録されている数値よりは大きな値をとる、または他の基底にまたがって特徴が表れる場合は小さな値をとる場合がある。図 b の表現ベクトルでは産業間取引のペアに関する特徴が抽出されている。ただし、表現ベクトルごとに 1×169 のベクトルを、元の産業連関表ひな形と同じ構造である 13×13 となるように再変換し、ヒートマップ図で表している。(3.1) の制約から、基底ごとに図 b の総和は 1 となる。またヒートマップ内の各要素の値が 1 に近いほど濃い赤色で表現され、その基底の特徴を大きく捉えた産業間取引であることを表す。

(i) 第 1 基底:愛知県を中心とした製造圏

図 3.2a より、愛知県の特徴が約 35 兆円と最も高く、神奈川県・兵庫県・静岡県が約 14 兆円、大阪府・茨城県が約 10 兆円と続いている。他方、図 3.2b では 製造業 ⇌ 製造業 が 0.48(48 %) と最も高く、サービス業 → 製造業 が 0.07(7 %)、サービス業 → 製造業 が 0.06(6 %) と続いており、製造業に対する産出 (供給) または製造業の投入 (需要) に関する取引が上位となっている。これらの結果から、「製造業」を説明する基底であると考えられる。特に第 1 基底ベクトルで上位に抽出された東海圏（愛知、静岡、三重）、関東圏（埼玉、千葉、東京、神奈川）、大阪圏（京都、奈良、大阪、兵庫）は製造品出荷額の全国シェアが高い経済圏である。[61]



(a) 基底ベクトル (都道府県)

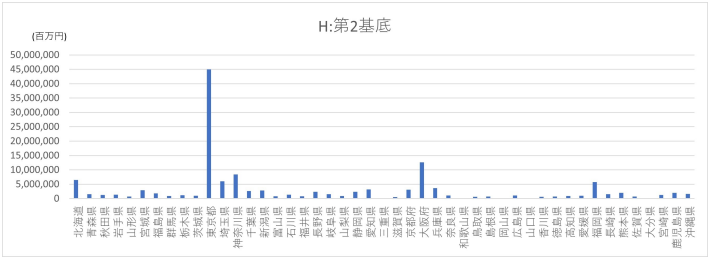


(b) 表現ベクトル (産業間取引)

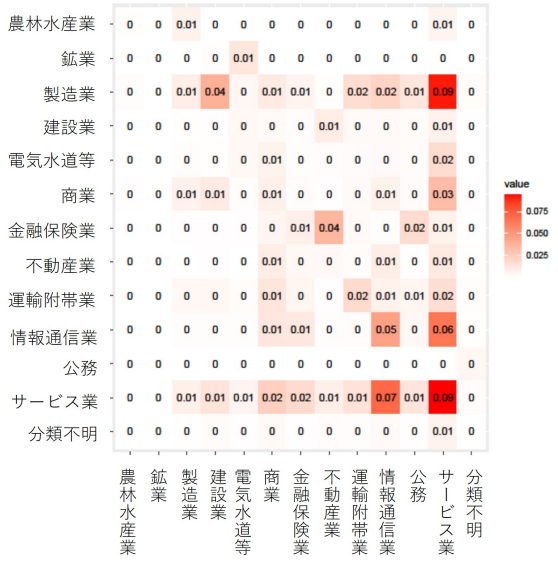
図 3.2: 第 1 基底の特徴

(ii) 第 2 基底:東京都を中心とした第 3 次産業

図 3.3a より、東京都の特徴が約 45 兆円と極めて突出しており、大阪府が約 14 兆円、神奈川県が約 8 兆円と続いている。他方、図 3.3b では サービス業 ⇌ サービス業 と 製造業 → サービス業 が 0.09(9 %) と最も高く、情報通信業 → サービス業 が 0.06(6 %)、情報通信業 ⇌ 情報通信業 が 00.5(5 %) と続いている。他にも 金融・保険業 → 商業 は 0.04(4 %) であり、図 3.2b と比べて多くの産業間で取引が活発な様子が伺える。特にサービス業の産出 (供給) 及び投入 (需要) のいずれも高い割合で特徴が抽出されている。これらの特徴から東京都などの大都市を中心としたサービス業等の第 3 次産業に関する基底であると考えられる。飲食業や宿泊業等を含んだサービス業やテレビ局等を含む情報通信業などが上位で抽出されていることからも、第 2 基底ベクトルが大都市に関連する特徴であると考えられる。



(a) 基底ベクトル (都道府県)

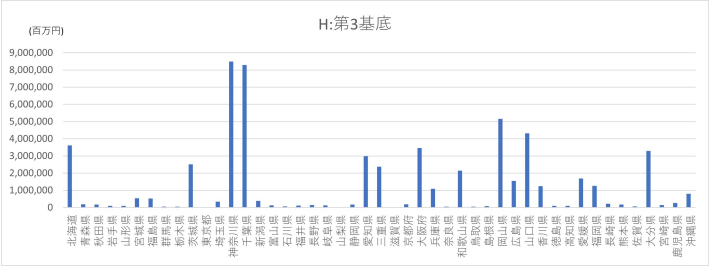


(b) 表現ベクトル (産業間取引)

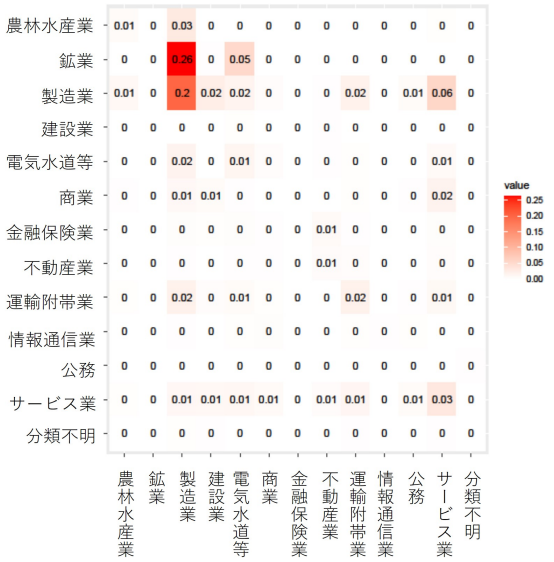
図 3.3: 第 2 基底の特徴

(iii) 第 3 基底:神奈川県・千葉県等の製油所拠点を中核とした鉱業

図 3.4a より、神奈川県の特徴が約 8.5 兆円と最も高く、千葉県が約 8.2 兆円、岡山県が約 5 兆円、山口県が約 4.2 兆円と続いている。他方、図 3.4b では 鉱業 → 製造業 が 0.26(26 %) と最も高く、製造業 ⇌ 製造業 が 0.20(20 %)、製造業 → サービス業 が 0.06(6 %) と続いている。特に製造業への産出 (供給) または製造業の投入 (需要) で高い割合で特徴が抽出されている。鉱業には「石炭・原油・天然ガス・鉄鉱石」などの、製造業において原料となる品目が含まれる産業 (サプライチェーンの川上産業) であり、これらが製造業へ産出されていることがわかる。また、第 3 基底ベクトルで抽出された都道府県はいずれも日本の石油元売が製油所を置く地域（千葉、神奈川、岡山、山口、北海道、大分、和歌山、大阪、愛媛、香川、三重、沖縄、宮城）[62] であることからも、第 3 基底は製油所機能を有する鉱業に関する基底であると考えられる。ひな形の 13 分類では、ガソリン・ナフサなどの「石油・石炭製品」が「製造業」に含まれることから、特に 鉱業 → 製造業 の特徴が強く抽出されたと推察される。



(a) 基底ベクトル (都道府県)



(b) 表現ベクトル (産業間取引)

図 3.4: 第 3 基底の特徴

3.2 COVID-19 流行下の石川県内滞在者移動行動変化に関する研究

3.2.1 概要

本節 [63] では非負値行列因子分解を携帯電話の位置情報から得られる滞留人口データに適用することで 1 日単位での人の移動行動に関する変化を分析した。携帯電話の位置情報データを利用することで、国勢調査やパーソントリップ調査等では把握できなかった人々のリアルタイムな行動分析が可能である。その一方で携帯電話の位置情報はデータ量が膨大なため、有益な特徴を抽出するための解析法が求められる。こうした研究背景から、非負値のデータ構造を非負値の特徴に分解できる、非負値行列因子分解法を適用した。

分析データは、COVID-19 流行前の 2019 年 1 月 1 日から第 3 波流行期である 2021 年 3 月 8 日までの 798 日の、それぞれ 4 時、14 時、19 時の 3 時点について、石川中央都市圏内 4 市 2 町を対象とした、2394 時系列 (798 日 ×3 時点)×1053 メッシュデータ行列を作成し、使用データとした。

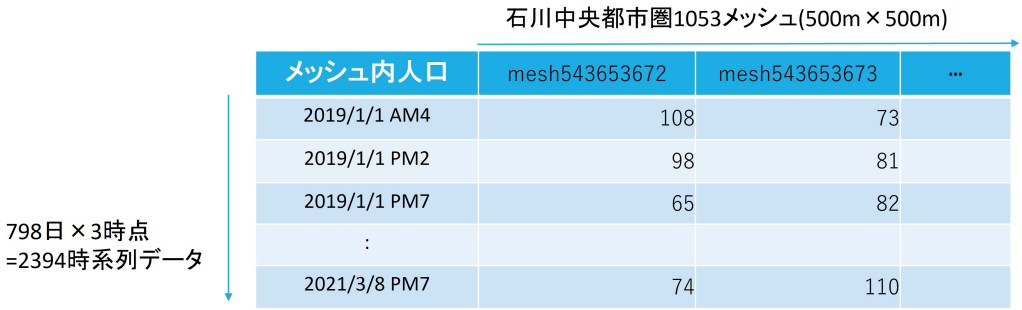
分析結果として、非負値行列因子分解の基底数を 3 とした時、「滞在の行動 (就寝)」、「平日の外出行動 (就業・就学)」、「休日の外出行動 (観光・夜間)」の基底が抽出された。いずれの基底のおいても、特に 2020 年 4 月から 5 月にかけて発生した COVID-19 の第 1 波流行期及び緊急事態宣言期間中に生活行動が大きく変化しており、緊急事態宣言解除後の 2020 年 6 月頃にはおおよそ宣言前の水準にまで生活行動が戻っていた。しかし「平日の外出行動 (就業・就学)」、「休日の外出行動 (観光・夜間)」においては、いずれも 2019 年の水準までには回復しておらず、依然として COVID-19 の影響が石川県内においても見られることが示された。

3.2.2 使用データと設定について

(i) 使用データについて

本節では石川県における COVID-19 流行前後の生活行動の変化を捉えるため、COVID-19 流行前の 2019 年 1 月 1 日から第 3 波流行期である 2021 年 3 月 8 日までの 798 日分の、それぞれ 4 時、14 時、19 時の 3 時点において、石川中央都市圏に属する 4 市 2 町 (金沢市・白山市・野々市市・かほく市・内灘町・津幡町) を対象とした 500m×500m メッシュ内人口 (1053 メッシュ) データを使用した[2]。すなわち 2394 時系列 (798 日 ×3 時点)×1053 メッシュデータの行列を作成し、本節の使用データとした。図 3.5 は本節における使用データのイメージである。

このデータ構造から基底行列 $\boldsymbol{H}$ では滞留人口の時系列変化、表現行列 $\boldsymbol{U}$ では地域メッシュデータの特徴が抽出でき、両特徴を併せて、「ある地域メッシュにおける滞留人口はどのように変化しているか」について解析する。



メッシュ内人口	mesh543653672	mesh543653673	…
2019/1/1 AM4	108	73	
2019/1/1 PM2	98	81	
2019/1/1 PM7	65	82	
:			
2021/3/8 PM7	74	110	

図 3.5: 滞留人口データの構造

[2]使用データとする年月日の期間及び対象地域の範囲は、今回取得できた最長期間・範囲である。また、4 時、14 時、19 時を選択した理由として、それぞれ就寝、就業・就学、夜間の活動に対応する特徴を抽出する目的で選択し、時間の区分は地域経済分析システムの RESAS [64] における from-to 分析で使用される時間区分を参考にした

(ii) 分析手法の設定

観測データ行列を $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}_+^{n \times m}$、分解する 2 つの行列を基底行列 $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{n \times k}$, 表現行列 $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}_+^{k \times m}$ とした時、以下のアルゴリズムを用いて基底行列 $\boldsymbol{H}$ 及び表現行列 $\boldsymbol{U}$ を算出した。

$$\boldsymbol{H}_{i,j} \leftarrow \boldsymbol{H}_{i,j} \frac{(\boldsymbol{Y}\boldsymbol{U}^{\boldsymbol{T}})_{i,j}}{(\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\boldsymbol{T}})_{i,j}}, \qquad \boldsymbol{U}_{i,j} \leftarrow \boldsymbol{U}_{i,j} \frac{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{Y})_{i,j}}{(\boldsymbol{H}^{\boldsymbol{T}}\boldsymbol{H}\boldsymbol{U})_{i,j}}, \quad \textit{subject to} \sum_m \boldsymbol{u_m} = \boldsymbol{1}, \tag{3.2}$$

ただし、(3.2) は要素ごとの計算を行っている。また、本データの構造から $n = 2394$、$m = 1053$ であり、基底数は COVID-19 による石川県への影響の大きな特徴を把握することと、朝・昼・夜の 3 時点における行動の違いに対応した特徴を把握するために $k = 3$ とした。計算アルゴリズムのその他条件は 3.1 節と同様である。また本節では、メッシュ単位の滞在人口データを地図上に可視化するため、ArcGIS Pro [65] を用いた。

3.2.3 分析結果

図 3.6 から図 3.8 はいずれも上記のデータに非負値行列因子分解を適用した結果である。

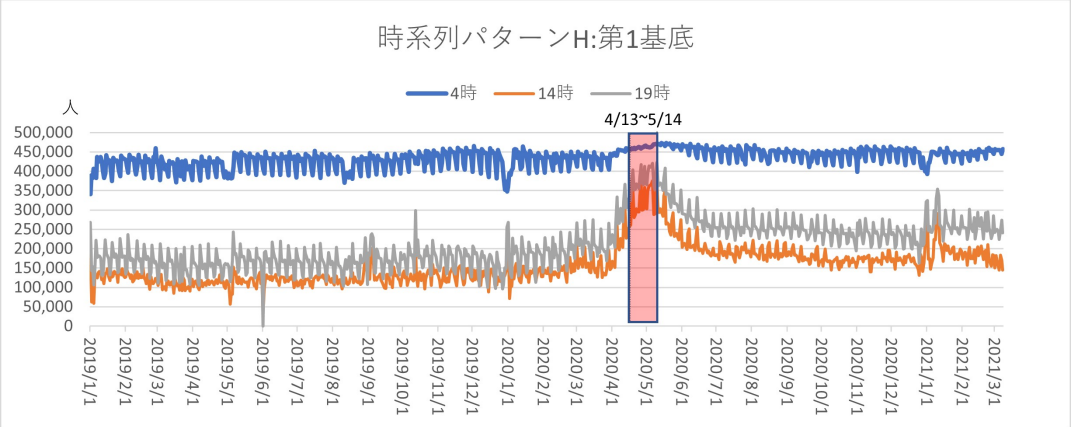
基底ごとに共通して 3 つの図がある。図 a が 2394 時系列データ分の滞留人口の推移を表す基底行列 $\boldsymbol{H}$ であり、横軸が 2019 年 1 月 1 日から 2021 年 3 月 8 日までの時系列、縦軸が基底別の滞留人口（人）である。時系列のグラフは「4 時」、「14 時」、「19 時」の 3 地点に分けてプロットしている。また、図中で赤く囲われている 2020 年 4 月 13 日から 5 月 14 日の期間は石川県において移動の自粛を強く要請する、緊急事態宣言が発令されていた期間である。図 b は、時系列変化を表す基底行列 $\boldsymbol{H}$ の内、曜日別に再集計し、1 週間単位の周期性の検出を試みている。また、COVID-19 流行前の 2019 年と 2020 年以降を区別している。ただし平常時の曜日別行動パターンを見るために祝日を除いて計算している。縦軸は滞留人口（人）を表す。また図 c は 1053 メッシュ分の地域メッシュデータを表す表現行列 $\boldsymbol{U}$ について地図上に可視化したものである。(3.2) の制約から、各基底における図 c(表現ベクトル) の総和は 1 となる。またメッシュ内における重みの値が大きいほど濃い色で表現され、その基底の特徴を大きく捉えた地域であることを示している。ただし色のグラデーションの基準値は基底ごとに異なる。

これら 3 つの図から、以降では基底行列 $\boldsymbol{H}$ に関する図 a、図 b から、COVID-19 流行前・中期におけるメッシュ内滞留人口の推移を解析し、表現行列 $\boldsymbol{U}$ に関する図 c からその時系列変動の特徴を強く示す地域を解析する。これら双方の視点から、基底ごとの COVID-19 による滞留人口へ影響と、その地域の特徴を、基底数 $k = 3$ に基づいた 3 つの基底パターンに分けて解析する。

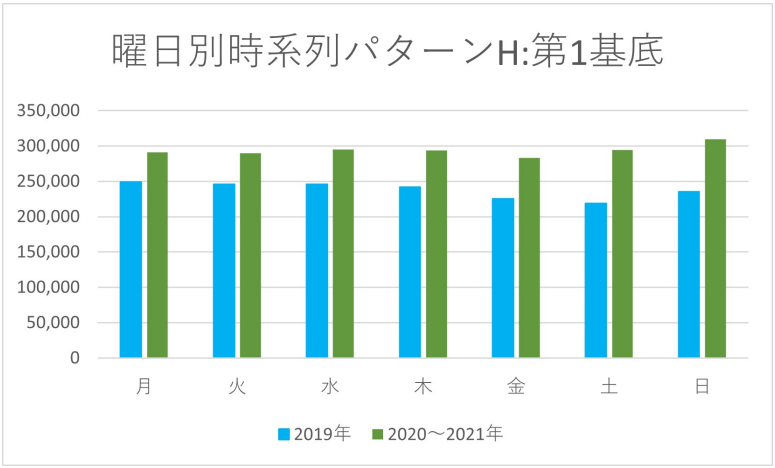
(i) 第 1 基底:滞在の行動 (就寝)

図 3.6a より、「4 時」における滞留人口が最も高く、「19 時」、「14 時」と続いている。また、COVID-19 第 1 波が流行し始めた 2020 年 4 月頃において、「4 時」は安定的な推移を示し、「19 時」及び「14 時」においては対応するメッシュでの滞留人口が大幅に増加しており、「4 時」の滞留人口に大きく近づいている。第 1 波流行後は、「4 時」において流行前より、滞留人口がやや高めに推移し、「19 時」においても流行前の 15 万人程度から 25 万人程度に滞留人口が増加し、「14 時」においても 10-15 万人程度から、15-20 万人程度に増加している。図 3.6b を見ると、曜日別で大きな差は見られないが、2019 年の COVID-19 流行前より 2020 年以降の流行後の滞留人口がいずれの曜日においても増加している。図 3.6c を見ると、地域特性がわかる。具体的には泉野・有松地区や、もりの里地区、近江町市場、彦三町周辺、西金沢地区、金石地区、野々市市、内灘町役場周辺など、多くの地域の特徴が抽出されている。

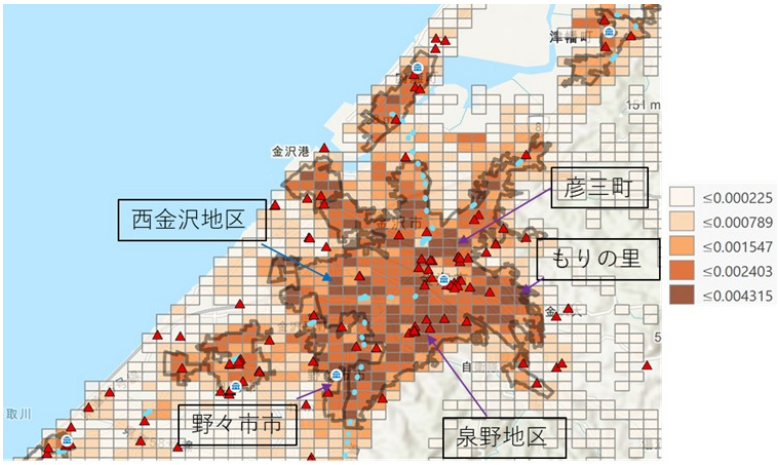
これらの結果から第 1 基底は「滞在の行動」の基底と考察される。泉野地区は金沢市における文教地区として人気の高い居住地区である。またもりの里は、約 1 万人が通う金沢大学の最寄りの地域である。他にも野々市市は東洋経済が発表する「住みよさランキング 2020」で全国 1 位を獲得した地域 [66] であり、住環境が整備された地区である。また時系列パターンにおいても、「4 時」で代表される就寝の特徴が突出することや、緊急事態宣言期間中に「14 時」「19 時」の滞留人口が増加していたことからも裏付けられる。またこれらから、石川県においては、第 1 波期においては外出自粛が強く行われていた。一方で、緊急事態宣言解除を機に昼夜の外出が増加し、2020 年 9 月から 2020 年 12 月まではその傾向が続いていたが、2021 年 1 月以降には特に「19 時」において、再び居住地に滞在する傾向が見られる。



(a) 時間別時系列パターン H:第 1 基底



(b) 曜日別時系列パターン H:第 1 基底



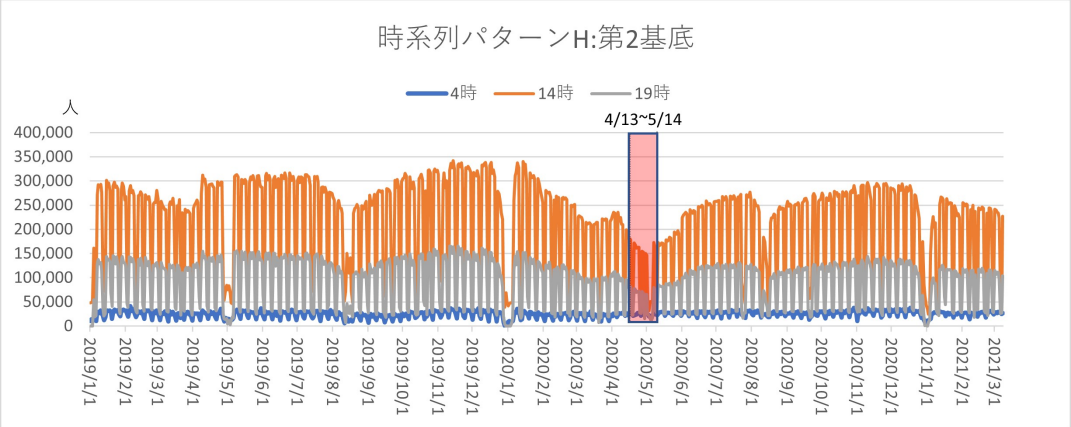
(c) 地域パターン U に対応する石川中央都市圏メッシュ:第 1 基底

図 3.6: 石川県内滞在者の移動行動パターン:第 1 基底
図中のマークは市役所、旗付きのマークは大学、楕円 (破線) は駅、丸は石川県管轄の事業所を表し、太枠は DID 地区区分を表している

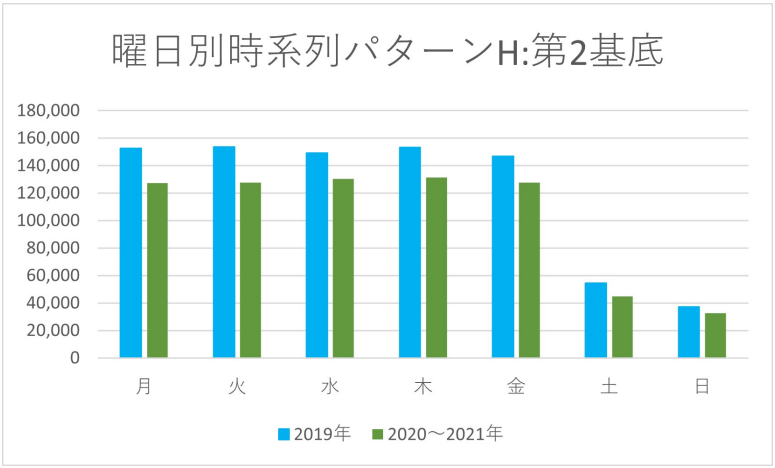
(ii) 第 2 基底:平日の外出行動 (就業・就学)

図 3.7a より、「14 時」における滞留人口が最も高く、次いで「19 時」が高い。「4 時」の滞留人口はこれらに比べてごくわずかである。また通年を総じて、ゴールデンウィークや盆期間、正月等の長期休暇中に滞留人口が大きく減少している特徴がある。COVID-19 第 1 波が流行し始めた 2020 年 4 月頃において、「14 時」及び「19 時」の滞留人口が前年にはない減少傾向を示したが、緊急事態宣言解除頃より徐々に滞留人口が増加し、2020 年 6 月には、緊急事態宣言前の水準までに回復している。しかし、特に「14 時」において、2019 年 5 月から 7 月、2019 年 10 月から 12 月にかけて 30 万人強で推移していた滞留人口が、COVID-19 第 1 波以降の 2020 年では 30 万人を一度も越えていない。図 3.7b を見ると、平日のパターンが土日よりも高い。また、2019 年の COVID-19 流行前の方が 2020 年以降の流行後の滞留人口よりいずれの曜日においても高い。図 3.7c を見ると、金沢駅から香林坊・兼六園付近、石川県庁等の就業地、金沢工業大学・金沢星稜大学・金沢学院大学等の就学地も特徴として抽出されている。

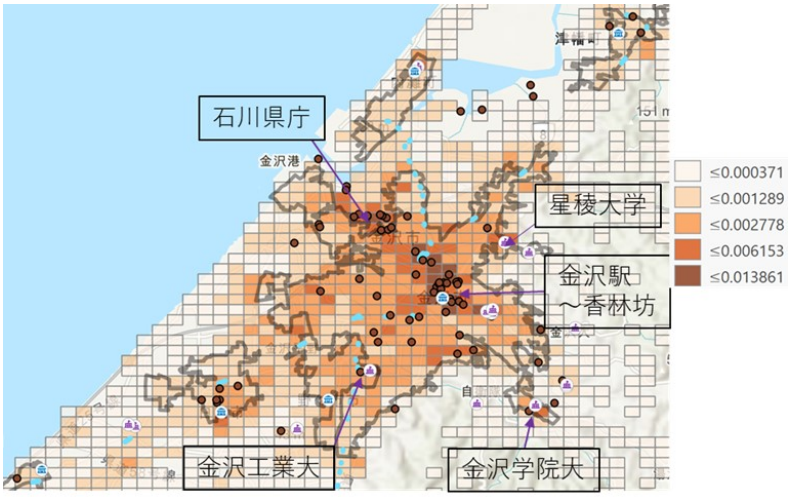
これらの結果から第 2 基底は「平日の外出行動 (就業・就学)」の基底と考察される。COVID-19 第 1 波期における石川県の就業は県内で感染者が増加した 2020 年 4 月初め頃より職場の利用が減少し、2020 年 5 月 14 日の石川県内緊急事態措置以降から回復傾向を示していたが、図 3.7a の 14 時においても類似した傾向が見られる。また図 3.7a で 2020 年 6 月頃より、14 時の滞留人口が大きく増加している。これは大学において 6 月上旬より一部の対面授業の段階的再開、校内への立ち入り制限緩和等が行われたことが一因であると考えられる。第 1 波流行期が沈静化した 6 月以降、COVID-19 流行による「就業・就学」の減少量は、流行前である 2019 年の同期間と比較して約 10 %前後である (滞留人口の推移が安定している 2019 年 6 月から 12 月と 2020 年 6 月から 12 月の平日 14 時のみの滞留人口平均を比較すると、2019 年が 296,301 人、2020 年が 263,261 人であり、2020 年/2019 年の比率は 88.8 %である)。すなわち石川県の企業・教育機関において、COVID-19 流行後も通勤・通学を控え、自宅等で勤務・勉学を行う者は平均約 10 %と考えられる。



(a) 時間別時系列パターン H:第 2 基底



(b) 曜日別時系列パターン H:第 2 基底



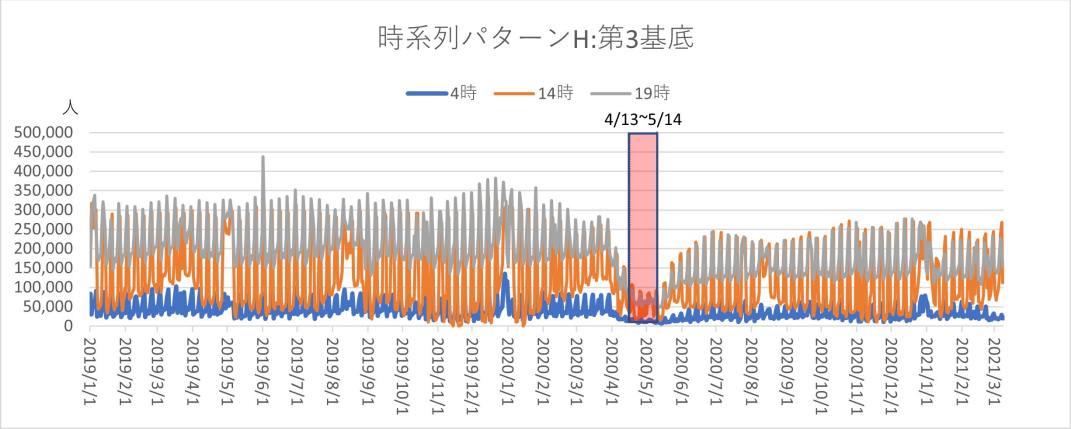
(c) 地域パターン U に対応する石川中央都市圏メッシュ:第 2 基底

図 3.7: 石川県内滞在者の移動行動パターン:第 2 基底
図中のマークは市役所、楕円 (破線) は駅、三角は美術館等の文化施設と体育館を表し、
太枠は DID 地区区分を表している

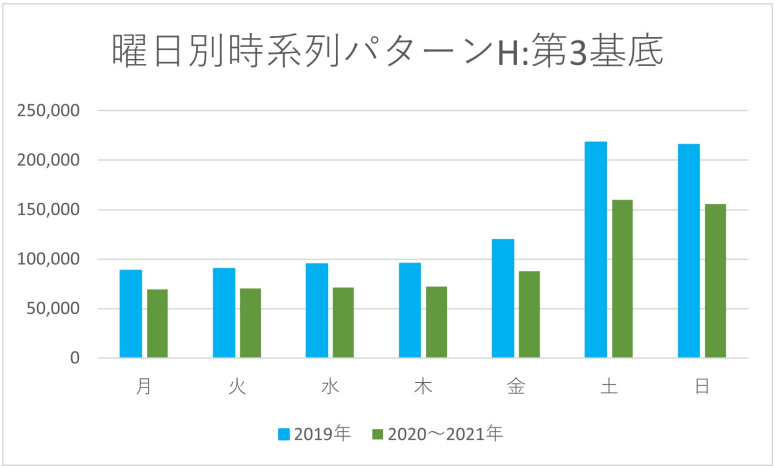
(iii) 第 3 基底:休日の外出行動 (観光・飲食・娯楽等)

図 3.8a より、「19 時」における滞留人口が最も高く、次いで「14 時」が高い。「4 時」の滞留人口はこれらに比べて少数である。また、COVID-19 第 1 波の予兆が見え始めた 2020 年 3 月下旬頃において、「14 時」及び「19 時」の滞留人口が大幅に減少したが、緊急事態宣言解除頃より徐々に滞留人口が増加し、2020 年 5 月末には、緊急事態宣言前の水準までに回復している。しかし、「19 時」において、2019 年には 15 万人から 35 万人付近で推移していた滞留人口が、COVID-19 第 1 波以降の 2020 年では 10 万人強から 25 万人程度の推移にとどまっている。図 3.8b を見ると、平日のパターンより土日が高く、平日の中でもやや金曜日が高い。また、2019 年の COVID-19 流行前の方が 2020 年以降の流行後の滞留人口よりいずれの曜日においても高い。図 3.8c を見ると、香林坊・片町や金沢駅・昭和町を中心に、ひがし茶屋街やにし茶屋街、長町武家屋敷等を含む地域で高い特徴が抽出されている。

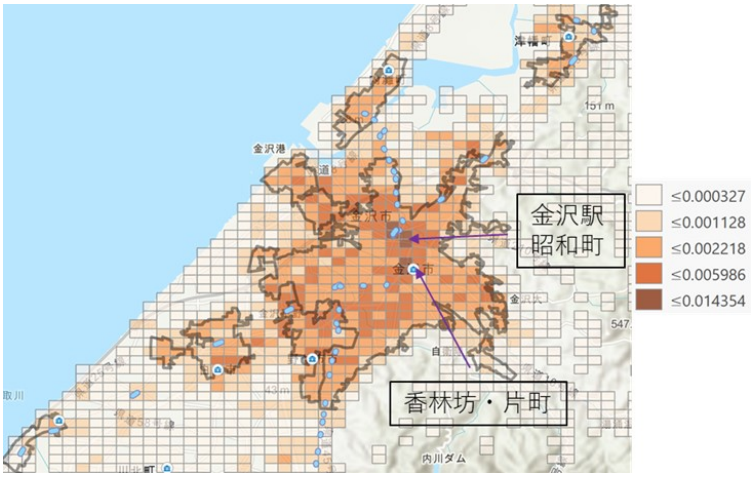
これらの結果から第 3 基底は「休日の外出行動 (観光・飲食・娯楽等)」の基底と考察される。石川県では緊急事態措置解除の翌日である 2020 年 5 月 15 日より、ホテルや旅館、商業施設等の休業要請を解除し、県内の主要な観光地も 2020 年 6 月 1 日に閉鎖が解除された。それに伴い図 3.8a より、2020 年 5 月 14 日以降に滞留人口が増加傾向となっている。また特に強く特徴が抽出された地域である香林坊・片町や金沢駅周辺は飲食店等が他の地区に比べて多く存在し、夜間に活気のある地域である。土日や金曜日に滞留人口が多いことも併せて、飲食・娯楽の行動に関する特徴であると考察される。石川県の「休日の外出行動 (観光・飲食・娯楽等)」人口は、第 1 波期以降においては安定的に推移しており、第 1 波期より感染者数が増加した第 2 波期、第 3 波期においても、第 1 波期ほどの大きな自粛行動は見られない。また、第 1 波流行期が沈静化した 6 月以降において、COVID-19 流行による「休日の外出行動 (観光・飲食・娯楽等)」の減少量は、流行前である 2019 年の同期間と比較して、14 時では約 20 %、19 時では約 30 %である (滞留人口の推移が安定している 2019 年 6 月から 12 月と 2020 年 6 月から 12 月の土日 14 時と 19 時の滞留人口平均を比較すると、14 時において 2019 年同平均が 279,006 人、2020 年同平均が 229,185 人であり、2020 年/2019 年の比率は 82.1 %である。また 19 時において 2019 年同平均が 302,399 人、2020 年同平均が 214,964 人であり、2020 年/2019 年の比率は 71.1 %である)。COVID-19 第 1 波流行後より、観光地等に活気が戻りつつあるものの、COVID-19 流行前である 2019 年ほどの水準には依然として回復していないことが考えられる。



(a) 時間別時系列パターン H:第 3 基底



(b) 曜日別時系列パターン H:第 3 基底



(c) 地域パターン U に対応する石川中央都市圏メッシュ:第 3 基底

図 3.8: 石川県内滞在者の移動行動パターン:第 3 基底
図中のマークは市役所、楕円 (破線) は駅、太枠は DID 地区区分を表している

第4章　スパース非負値行列因子分解に関する研究

4.1　COVID-19 流行前後における県間移動行動変化に関する研究

4.1.1　概要

本節 [67] では非負値行列因子分解を携帯電話の位置情報から得られる滞留人口データに適用することで人の移動行動に関する変化を分析した。ここでは 47 都道府県の県間移動・滞在データを分析対象としており、対象期間を COVID-19 流行前・中期とすることで、COVID-19 が県間移動行動に及ぼした影響について解析した。さらに解析結果を踏まえ、政府が実施した諸政策の効果について検証した。

携帯電話の位置情報は「3.2 節. COVID-19 流行下の石川県内滞在者移動行動変化に関する研究」と同様に非負値である点、ビックデータである点から非負値行列因子分解を適用する。さらに本節では分解後の行列要素に対し、 0 を多くするような罰則制約であるスパース制約 (L_1ノルム) を付与することで、微細な特徴を 0 にし、解釈に有益な特徴をより際立たせる工夫を行っている。

その結果、首都圏での通勤の抑制効果、週末の外出控え、長期休暇時の帰省の抑制及び居住地の都道府県内滞在の時系列変化を明らかにした。さらに、残差構造を解析することで、COVID-19 流行後の移動行動パターンが大型台風接近時の行動パターンと同様に、通常期と比べて特異なものであったことを示した。

4.1.2　使用データと設定について

(i) 使用データについて

本節では、47 都道府県を対象にした県間移動・滞留人口データを用いて、COVID-19 流行前・中期における移動行動の変化について解析する。分析期間は 2014 年 3 月 1 日から 2020 年 5 月 31 日まで[1]とし、対象地域は 47 都道府県間の居住地・滞在地の組み合わせデータである。すなわち時系列データ 2284 日 (2014.3.1-2020.5.31）×2209(47×47 都道府県ペア) の行列データを作成した。このデータから得られたある日付 d の都道府県間の居住地-移動先表を、下記のように定義する、

$$\boldsymbol{Q}_{d,i,j} = \begin{pmatrix} q_{1,1,1} & \cdots & q_{1,i,j} & \cdots & q_{1,j,i} & \cdots & q_{1,47,47} \\ \vdots & & \vdots & & \vdots & & \\ q_{d,1,1} & \cdots & q_{d,i,j} & \cdots & q_{d,j,i} & \cdots & q_{d,47,47} \\ \vdots & & \vdots & & \vdots & & \\ q_{2284,1,1} & \cdots & q_{2284,i,j} & \cdots & q_{2284,j,i} & \cdots & q_{2284,47,47} \end{pmatrix}. \tag{4.1}$$

.

ここで $d \in D$、$D = [2014.3.1, \cdots, 2020.5.31]$ 、$(i, j) \in (Z \times Z)$ は居住地, 移動先ペア、Z は 47 都道府県に対応し、$Z = 47$ である。つまり、$q_{d,i,j}$ の値は、都道府県 i に居住する人の中で d 日に都道府県 j に滞在している人数を、携帯電話運用データから推計したものである。ただし、$i \neq j$ のように、特定の居住地からあ

[1]使用データについて、モバイル空間統計におけるデータの記録は 2014 年 3 月 1 日から行われたため、本分析で使用するデータ期間はデータ取得時点で最長期間である。また最長期間を使用した理由として、通常期と COVID-19 流行期における差を明確化する目的から、通常期の変化を平滑化し、COVID-19 流行前・中期で比較をしやすくするためである。

る都道府県へ流出する人数、あるいは流入する人数は、いずれも居住地に滞在する人数 ($i = j$) と比較すると極めて小さく、そのままのデータをスパース非負値行列因子分解に適用した場合、基底ごとには少数の (i, j) ペアのみの特徴が抽出され、値の小さいペアのほとんどが無視される。その結果として、日本全国の大筋の変化を理解するために、より基底数を多くして解析する必要がある。そこで居住地とは異なる他の都道府県への移動が行われた属性に対し、特定の居住地からある都道府県へ流出する人数及びある都道府県から特定の居住地へ流入する人数についてそれぞれ集計したものを、「$q_{d,i,\bullet}$」、「$q_{d,\bullet,i}$」と表すことにする。ここで、「$\bullet$」は $i \neq j$ についてすべて合算したものとする．これらの処理によって使用データは以下のように表せる、

$$\boldsymbol{X} = \begin{pmatrix} \cdots & q_{1,i,i} & \cdots & q_{1,i,\bullet} & \cdots & q_{1,\bullet,i} & \cdots \\ & \vdots & & \vdots & & \vdots & \\ \cdots & q_{d,i,i} & \cdots & q_{d,i,\bullet} & \cdots & q_{d,\bullet,i} & \cdots \\ & \vdots & & \vdots & & \vdots & \end{pmatrix}. \tag{4.2}$$

$$\begin{aligned} q_{d,i,\bullet} &= \sum_{j \in [Z \setminus i]} q_{d,i,j}, \\ q_{d,\bullet,i} &= \sum_{j \in [Z \setminus i]} q_{d,j,i}, \end{aligned} \tag{4.3}$$

ここで $[q_{d,i,i}, q_{d,i,\bullet}, q_{d,\bullet,i}]$ は、自地域滞在、他地域への流出、他地域からの流入に対応する。これらの処理により、47 都道府県分の滞在・流出・流入に関する 3 パターンの合算データが得られ、これを使用データとした。すなわち 2,284 日時系列データ ×141 空間データを本節の使用データ行列 $\boldsymbol{X}$ とした。

(ii) 分析手法の設定

本節の非負値行列因子分解では、分解後の特徴を際立たせることによって、解釈しやすい分解結果を得ることを目的に、時系列パターンに L_1 正則化項を付与したモデルを使用する。2284 日分時系列データ ×141(47 都道府県 ×3 パターン) の行列 $\boldsymbol{X} = [\boldsymbol{y_1}, \cdots, \boldsymbol{y_{141}}] \in \mathbb{R}_+^{2284 \times 141}$ を、2 つの非負値行列 $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{2284 \times k}$, $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_{141}}] \in \mathbb{R}_+^{k \times 141}$ に分解するとき、スパース非負値行列因子分解は，残差行列 $\boldsymbol{X} - \boldsymbol{HU}$ のフロベニウスノルムと時系列パターン行列 $\boldsymbol{H}$ の L_1 正則化項の和である、

$$L = \|\boldsymbol{X} - \boldsymbol{HU}\|_F^2 + \mu \|\boldsymbol{H}\|_1, \tag{4.4}$$

を最小化する問題となる。ただし μ はハイパーパラメータを表す。

本節のスパース非負値行列因子分解では、行列 $\boldsymbol{H}$ のみに L_1 正則化項を付与して、スパースの性質を持つように分解する。手法としては、行列 $\boldsymbol{H}$ だけでなく行列 $\boldsymbol{U}$ にも同様に正則化項を付与することも可能であるが、行列 $\boldsymbol{U}$ に正則化項を課した場合、人口の少ない地方部の特徴を無視しやすくなる[2]。そこで、本節では時系列変化を示す行列 $\boldsymbol{H}$ のみに、解釈しやすい明瞭な結果が得られるようにスパース化を実施しつつ、その時系列変化パターンでできるだけ多くの都道府県の行動変化を解析する。

このとき、(4.4) を最小化する行列 $\boldsymbol{H}$,$\boldsymbol{U}$ は、任意の初期値を設定したうえで以下の更新式を収束するまで繰り返し適用することによって導出できる、

$$\boldsymbol{H}_{i,j} \leftarrow \boldsymbol{H}_{i,j} \frac{(\boldsymbol{YU^T})_{i,j}}{(\boldsymbol{HUU^T})_{i,j} + \mu(\boldsymbol{1}_{2284}\boldsymbol{1}_k^T)_{i,j}}, \qquad \boldsymbol{U}_{i,j} \leftarrow \boldsymbol{U}_{i,j} \frac{(\boldsymbol{H^T Y})_{i,j}}{(\boldsymbol{H^T HU})_{i,j}}, \quad \textit{subject to} \sum_m \boldsymbol{u_m} = \boldsymbol{1}, \tag{4.5}$$

[2]本分析手法によって抽出される結果を解釈する際に、「人口規模の大きな都道府県の移動をより説明しやすい基底が得られやすい」点には注意が必要である。これは、二乗誤差基準を仮定した次元縮約法が共通して持つ性質であり、本節のスパース非負値行列因子分解でも同様である。より地方部で特有の行動を抽出するためには、行動量の「変化率」を同分析に適用することで実現できると考えられる。

ここで、$\mathbf{1}_k$ は要素が全て 1 の $k \times 1$ ベクトルであり、$\boldsymbol{H}^T$ は行列 $\boldsymbol{H}$ の転置を示す。また、(4.5) は行列の要素ごとの演算である。非負値行列因子分解の収束条件は第 3 章と同様である。

ハイパーパラメータ μ の設定について、本節では 10^1 から 10^{10} の範囲で探査的に行い実施した。表 4.1 は、ハイパーパラメータごとの、$\boldsymbol{H}$、$\boldsymbol{U}$ の要素のスパース率（要素の中でゼロである割合）を示したものである。ハイパーパラメータ μ の値が大きくなればなるほど、正則化項が強く作用するために、$\boldsymbol{H}$ のスパース率は大きくなる傾向にあり、スパース率が大きいほど分解された時系列パターンの特徴に 0 が多くなるため、より解釈が容易となる。一方で、正則化項の影響を大きくすると、データとの残差 (RSS) も大きくなるため、ある程度のスパース化の効果が確認できる中で最小の μ の値を設定することが望ましいと考える。そのため、ここでは $\boldsymbol{H}$ のスパース化の効果が十分に確認できるなかで、できるだけ小さい μ を設定する。表 4.1 より、$\boldsymbol{H}$ スパース率は $\mu = 10^5$ から $\mu = 10^6$ にかけて大きく増加し、それ以降はそれほど大きく変化していない。すなわち、$\mu = 10^6$ 以上であれば、十分なスパース化の効果があり、これ以上大きくしてもその効果は大きく変化しないために、「十分な効果があると判断できる範囲内での最小の値」として $\mu = 10^6$ を設定した。

表 4.1: ハイパーパラメータ μ の変化と時系列・空間パターンのスパース率の変化

μ	10^2	10^3	10^4	10^5	10^6	10^7	10^8	10^9	10^{10}
$\boldsymbol{H}$ スパース率	1.2%	6.4%	12.0%	15.4%	25.2%	27.1%	29.4%	28.3%	27.8%
$\boldsymbol{U}$ スパース率	9.9%	19.6%	12.6%	27.9%	13.2%	12.1%	14.9%	13.2%	12.6%
RSS(10^{13})	2.5	3.5	3.7	5.8	8.2	8.6	9.9	8.3	8.6

※ 1 スパース率は H、U それぞれの全要素数に占める 0 の割合と定義している。
※ 2 RSS: Residual sum of squares （残差二乗和）

また、事前に決定するパラメータとして基底数 k も外生的に決める必要がある。その決定法には、情報量基準の一種である AIC や BIC 等を利用する方法があるが、本節では結果の解釈しやすさを優先して，一般的に利用される都市間旅行データにて想定される代表的な旅行目的（業務，観光，私用・帰省）が得られる最小の基底数を選択する。具体的には、基底数 k を 1 から順に大きくして、上述の旅行目的 3 種類の時系列面での特徴を満たすパターンがそれぞれ 1 つ以上得られた、最小の基底数 $k = 5$ を選択した。

また、上記で設定した基底数に基づく行列分解後、本モデルでは説明しきれなかった特殊な変化を、残差行列 $\boldsymbol{X} - \boldsymbol{HU}$ を用いて残差構造から解析する。

(iii) COVID-19 に対する諸政策

図 4.1 は、モバイル空間統計において推定された、2014 年 3 月 1 日から 2020 年 5 月 31 日までの「居住県外滞在人数」(県外への移動者数) の推移を示している。2019 年までの COVID-19 の影響を受けていない期間の時間推移では、概ね 2014 年-2019 年は共通の日変動であることが確認できる。具体的には、約 800 万人程度の人が 13 時台に県外滞在をしており、年間に 3 回のシーズンにおいてその量が大きくなる。それは、年末年始、5 月初旬（ゴールデンウィーク）と 8 月中旬（お盆）であり、この期間には、通常時の 1.5 倍以上、最大で 1,400 万人が居住している都道府県外に滞在している。なお、9 月末に 2015 年にだけある、居住県外滞在人数の増加はシルバーウィークに該当する。この年には祝日の配置によってカレンダー上で 5 連休（2015 年 9 月 19 日から 9 月 23 日）になっており、大きな県外滞在人数の増加が観測されている.

次に COVID-19 の影響を受けた 2020 年の推移（青線）では、1 月と 2 月はほとんど 2019 年以前と同じ範囲を変動しているが、3 月から県外への移動者数が大幅に減少している。この図から、5 月末までの COVID-19 による居住県外での滞在者数の推移は、以下の 3 フェーズに分けることができる。

(1) 2020 年の 2 月末より、居住県外での滞在者数が減少し始めている。この減少が進んだ期間は、日本政府による「イベントの開催に関する国民の皆様へのメッセージ」(2/20)、北海道による独自の緊急事態宣言(2/28)、全国での学校一斉休校要請(3/2)など、行政が活動の抑制を発信し始めた時期である。この時の減少量は、通常時より約 100 万人(約 12%)程度少なく、この水準が 1 か月程度継続している。

(2) 3 月末から 5 月初旬(ゴールデンウィーク)にかけて、さらに県外滞在数が減少し続けた。その結果、ゴールデンウィークに記録された最小値では、県外への移動者数が約 200 万人まで減少した。これは、通常時の約 800 万人の 1/4 であり、ゴールデンウィーク期間のピークとして例年記録されていた約 1,400 万人の約 1/7 である。なお、減少し続けた期間(3 月末から 5 月初旬)において政府による緊急事態宣言が発令されている(7 都府県:4/7、全国:4/16)が、発令時点において急激に県外滞在者数が減少するといった変化は図 4.1 では確認できない。また、この居住県外滞在人数の低下は、7 都府県で先行して緊急事態が宣言される 2 週間前の首都圏 1 都 4 県知事の外出自粛要請(3/26)ごろに開始している。

このような時間推移の特徴は、全国レベルで見る限り、単一の事象に対応した変化ではなく、複数の事象や感染者数の推移に応じて、徐々に進んだと推測される。この期間には、緊急事態宣言の発令(4/7 か 4/16)だけでなく。4 月中旬まで増加し続けた感染者数とそれに関連する報道や、各自治体による発信・要請(例えば、3/20 大阪府による不要不急の外出自粛のお願い、4/12 北海道・札幌市緊急共同宣言、県独自の緊急事態宣言(4/10 愛知、4/13 石川、4/14 福井)、4/24 東北・新潟緊急共同宣言)があり、これらが重なって行動量低下が進展したものと考えられる。

(3) 5 月のゴールデンウィークから、5 月末にかけてやや回復傾向が確認できる。短期間かつ緩やかな変動であるため、より明確なトレンドを把握するためにはより長期のデータを見る必要があるが、この回復が開始した時点は、緊急事態宣言の解除(39 県:5/14、関西圏:5/21、首都圏・北海道:5/25)があった後である。なお、この時点では「県外への移動自粛」は継続した状態であり、全国的に県外への移動自粛要請が解除されたのは 6 月 19 日である。

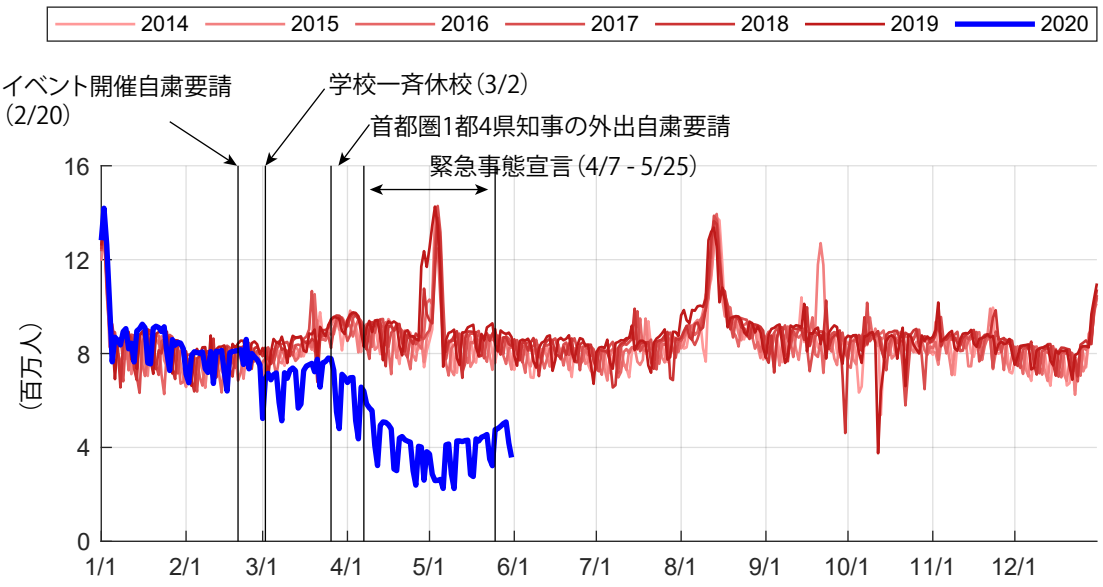


図 4.1: 居住県外滞在者数の時間推移

4.1.3 分析結果

都道府県単位の時系列データを非負値行列因子分解することで、COVID-19 による都道府県を跨ぐ滞在行動の変化を、以下の 2 点に着目して明らかにする。1 点目は、都道府県ごとの移動行動の時系列変化を、Yamaguchi and Nakayama (2020) [25] のように、複数の「行動の内容」ごとに分解することである。2 点目は、通常時の変動を示す行動分類で説明できる部分を取り除いた残差を見ることで、COVID-19 流行に伴う行動変容の地域的な偏在の有無を明らかにする。

以下ではスパース非負値行列因子分解によって得られた、時系列パターンの特徴を有する基底行列 $\boldsymbol{H}$ と空間パターンの特徴を有する表現行列 $\boldsymbol{U}$ の 2 つの非負値行列について結果を解析する。図 4.2 から図 4.7 は、それぞれ時系列パターン $\boldsymbol{H}$ と空間パターン $\boldsymbol{U}$ ごとの成分の値を、基底ごとに示したものであり、それぞれ 3 種類の図がある。図 a は時点 d において、基底成分 k によって表される居住者数及び県間移動者数（$h_{d,k}$）を示している。つまり、行列 $\boldsymbol{H}$ の 1 列分の情報を、年ごとに分けて折れ線グラフで示したものである。図 b は図 a に関して、全期間を通じた曜日ごとの平均値を示している。この図から、図 a の中で把握しにくい 7 日ごとの日変動を把握することができる。図 c は基底成分 k における空間パターン i の割合を示している。横軸の空間パターンは左から、居住地内の滞在人数（47 都道府県）、特定の居住都道府県からの流出人数（47 都道府県）、特定の都道府県への他県からの流入人数（47 都道府県）の順に並べている。ここで都道府県はコード順（北東から南西の順）に並べてある。

(i) 業務・通勤目的変動の時系列推移

図 4.2 の第 1 基底と図 4.3 の第 2 基底は、それぞれ類似性が高い成分であるため、まとめて結果を解析する。はじめに、COVID-19 の影響を受けていない期間の各基底の日変動を図 4.2a と図 4.3a から確認する。図 4.2a では 2014 年から 2019 年にかけて経年的に増加している。一方で図 4.3a では 2014 年から 2019 年にかけて経年的に減少し続けている。そして、この日変動のうちで 7 日間周期の変動のみに着目した図 4.2b と図 4.3b をみると、共通して平日に突出したパターンである。これらの基底の空間パターンを示した図 4.2c と図 4.3c からも、類似性が高い。具体的には、流入量（最も右の領域）では少数の都道府県（大きいピークから順に、東京都・大阪府・愛知県）に集中する一方で、流出量（緑色の領域）は日本全国に微量の値を取り、とくに大都市圏周辺で大きい傾向にある。これらは図 4.2c と図 4.3c のどちらでも共通の傾向であるが、第 1 基底（図 4.3c の薄い青線）では、居住都道府県内の滞在者数（最も左の領域）において東京と愛知県の居住都県内滞在が比較的大きい値をとっている違いがある。

以上の日変動の特徴から、平日に 3 大都市圏へ移動する行動パターンであることから、この第 1・2 基底は業務・通勤行動が多くを占める成分であると推測される。また、流出が関東地方で大きいことから、埼玉県・千葉県・神奈川県から東京都へ通勤する行動も、この基底に含まれると考えられる。そして、この行動が 2 つの基底として別々に導出された理由は、この行動の空間分布が 2014 年から 2019 年にかけて徐々に変化してきたことを示している。具体的には、図 4.4 より第 1 基底では、都道府県をまたいだ移動の上昇トレンドが、第 2 基底ではその下降トレンドが抽出され、COVID-19 期の特徴は第 1 基底で概ね捉えている。

図 4.4 は第 1 基底と第 2 基底の日移動平均を示したものである。この図から、基底 2 から基底 1 へ入れ替わる変化は、徐々に推移してきたわけではなく、およそ 4 回程度、短期間に急激に起こっている。それは、図 4.4 に示した矢印の 4 時点であり、2016 年から 2019 年の毎年お盆時期の前後である。この変化の原因は、転居などの居住地変化による可能性などが考えられる。

つぎに、これらの業務・通勤目的の COVID-19 による行動変化を 2020 年の情報から解析する。まず、図 4.3a に示される第 2 基底では、2020 年は継続してほぼゼロである。この基底の日変動では、COVID-19 の影響がない 2019 年末からほぼゼロで推移しており、COVID-19 に関係なく経年的に減少し続けた結果、近年はほとんど見られなくなった成分であることが確認できる。

図 4.2a より、COVID-19 の期間における推移を解析する。2020 年 3 月以降の推移を見ていくと、以下の 3 つの特徴に分けられる。

(1) 7 日周期の変動の最大値の推移を見ると、2020 年 3 月上旬（学校の休校が発表されたタイミング）に約 1,200 万人から約 1,000 万人[3]に減少し、1ヵ月程度その人数が継続している。このような変化は、2019 年以前には見られず、COVID-19 による影響であると考えられる。

(2) 7 都府県で先行して緊急事態宣言が発令される、およそ 2 週間前からさらに急激に減少を始めている。そして、全国が緊急事態宣言下にあった、4/16 から 5/7 までの期間は、この通常時の平日にある大都市圏へ集中するパターンの旅行はほぼゼロで推移していた。

(3) 緊急事態宣言が段階的に解除されるにしたがって、徐々に回復するものの、5 月末の時点で通常時に期待される量（2019 年の後半の水準が約 1,300 万人）の約 1/4 程度である。

以上から、通常時の平日に強く確認された、（業務・通勤目的と思われる）大都市へ集中する旅行行動は、緊急事態宣言中はほぼゼロになるなど大幅に減少していた。

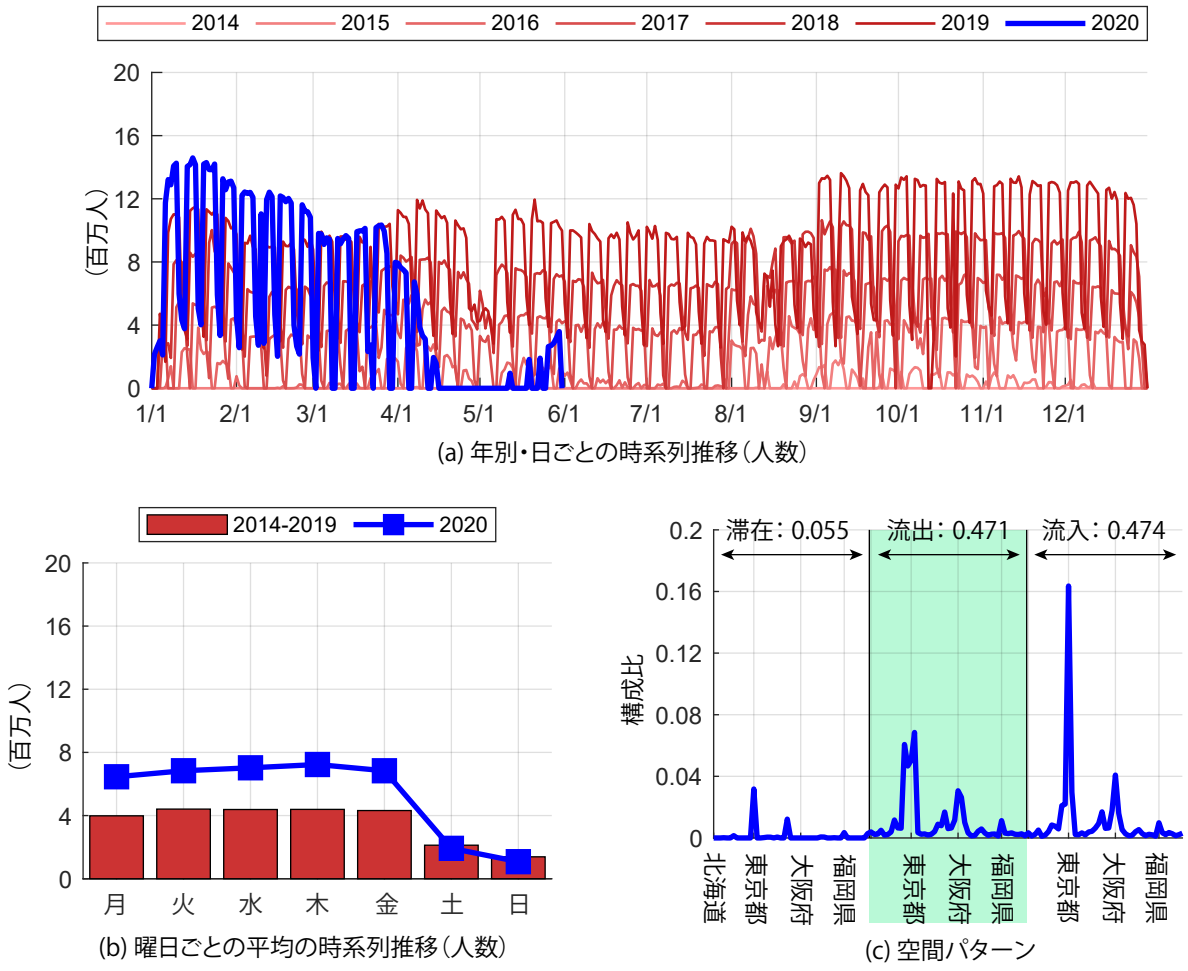


図 4.2: 第 1 基底の推定結果

[3]なお、この数値は単位が人であるが、居住県外に滞在している人は「流入人数」と「流出人数」でダブルカウントされるために、実際の行動人数の 2 倍近い数値となる点は注意が必要である。

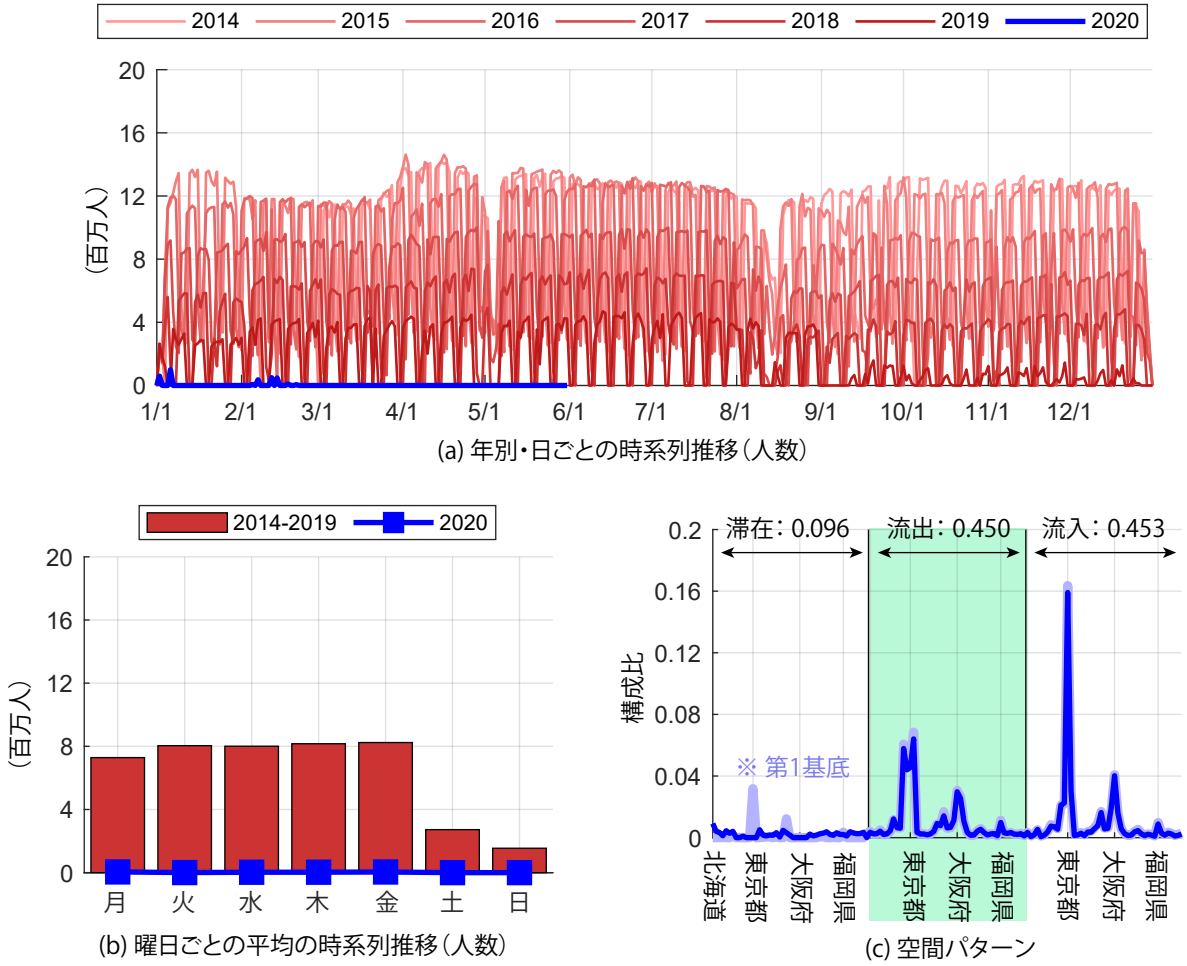


図 4.3: 第 2 基底の推定結果

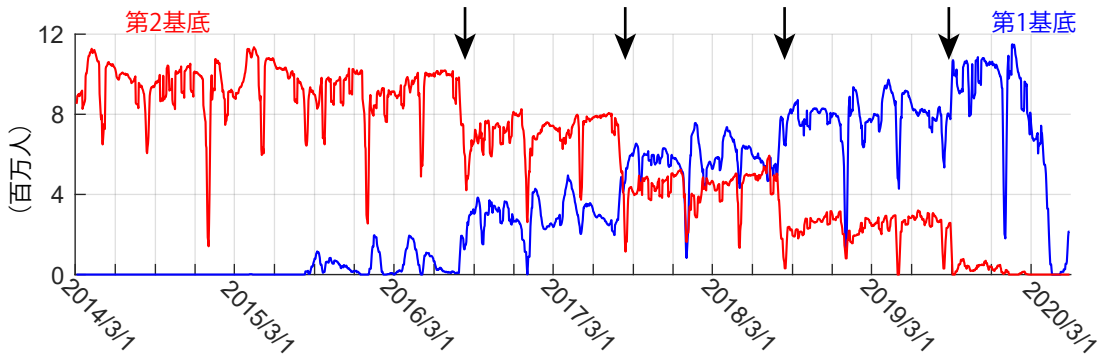


図 4.4: 第 1 基底と第 2 基底の時系列推移（7 日移動平均)

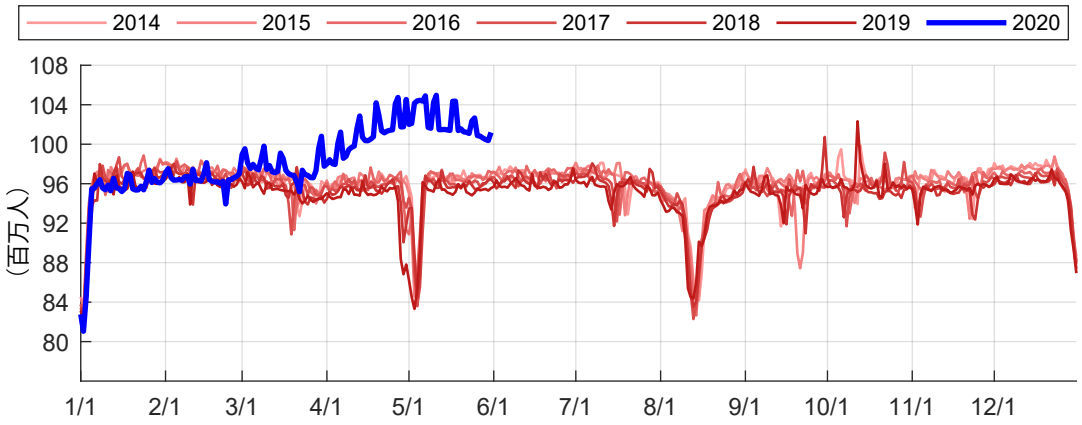
(ii) 居住都道府県における滞在者数変動の時系列推移

図 4.5 から、第 3 基底の情報を解析する。図 4.5a の数値オーダーを見ると、この基底の行動パターンを実施している人数は。他の基底と比較して大きい特徴がある。人数ベースでこの基底が占める割合を算出すると、$\frac{\mathbf{1}_{2284}^{T}\boldsymbol{h}_3}{\sum_{m=1}^{k}\mathbf{1}_{2284}^{T}\boldsymbol{h}_m} = 0.870$ であり、平均で人口の 9 割近くがこの基底が示す行動をとる。

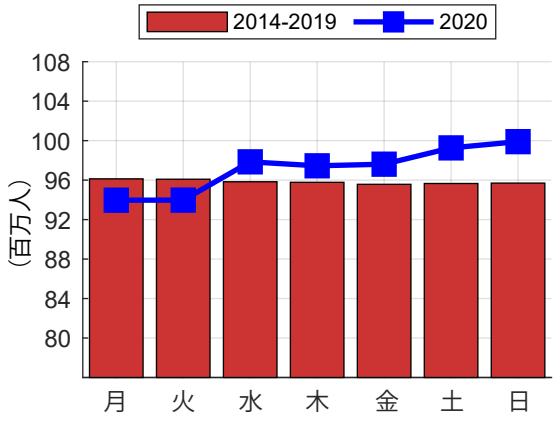
つぎに、図 4.5c をみると、この成分のうちの 95%程度が居住地県内での滞在行動であり、その都道府県内での大小関係はおおよそ人口規模に比例する形である。つまり、第 3 基底は「居住都道府県での滞在行動」を示す成分であるといえる。

この通常時（2019 年まで）の日変動に関して図 4.5a より、普段は安定した居住者の推移を示すが、特に GW や盆休み、正月休み期にこの滞在者が大きく減少する、すなわち県外への旅行行動をとる人が多くなる。また、図 4.5b の曜日ごとの変動をみると、ほぼ曜日ごとの変化が見られない。

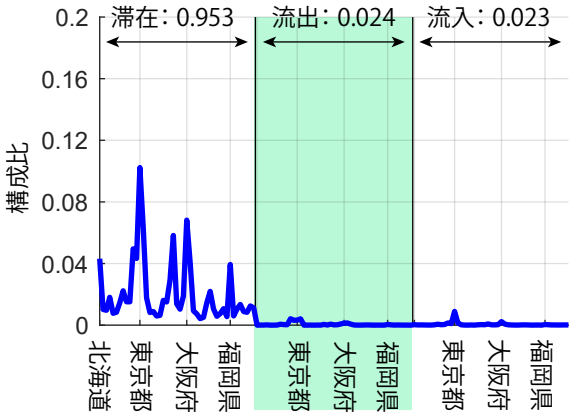
COVID-19 による行動変化は、おおよそ図 4.1 で確認できた県外旅行者と反対の特徴を示している。すなわち、学校の一斉休校が開始されたタイミングでの増加した後、さらに 3 月下旬からゴールデンウィークにかけて徐々に増加し続ける。そして、緊急事態宣言の解除後に若干の回復傾向が確認できるという推移である。これは、COVID-19 の状況下において、居住県外への移動を自粛し、都道府県内での滞在が増加していることを示す。



(a) 年別・日ごとの時系列推移（人数）



(b) 曜日ごとの平均の時系列推移（人数）



(c) 空間パターン

図 4.5: 第 3 基底の推定結果

(iii) 長期休暇時の大都市から地方への行動の時系列推移

図 4.6 から、第 4 基底の情報を解析する。COVID-19 の影響を受けていない期間の各基底の日変動は図 4.6a と図 4.6b より、通常時の休日にも特徴が抽出されているが、多くは GW や夏休み（お盆の期間がピーク）・年末年始の 3 期間に集中して発生している行動である。つぎに、図 4.6c よりこの行動パターンの空間的な特徴は、まず流入量（最も右の領域）が、地方部も含めて（東京都・大阪府を除いて）日本全国でほぼ満遍なくある。一方で、流出量（中央の緑の領域）は、東京都・大阪府・愛知県とその隣県といった大都市部が多くを占める。すなわち、大都市から地方部への旅行が多くを占める行動であると考えられる。以上の内容から、第 4 基底は「長期休暇時における大都市から地方への行動」に対応し、帰省行動が強く表れていると推測できる。なお、図 4.6c では、この成分には北海道居住者が北海道に滞在する行動も多くあることが確認できる。これは、北海道居住者が、他の県の人と比較して長期休暇の時期に道内に滞在する割合が大きいことを示している。

COVID-19 による行動変化では、2020 年では数回の微量の移動行動があるものの、ほぼ一貫してこの成分がゼロである。とくに、2014 年から 2019 年までの傾向を踏まえると膨大な量の移動が予定されていたと思われるゴールデンウィークにおいても完全にゼロであり、極めて異質な年であったと言える。

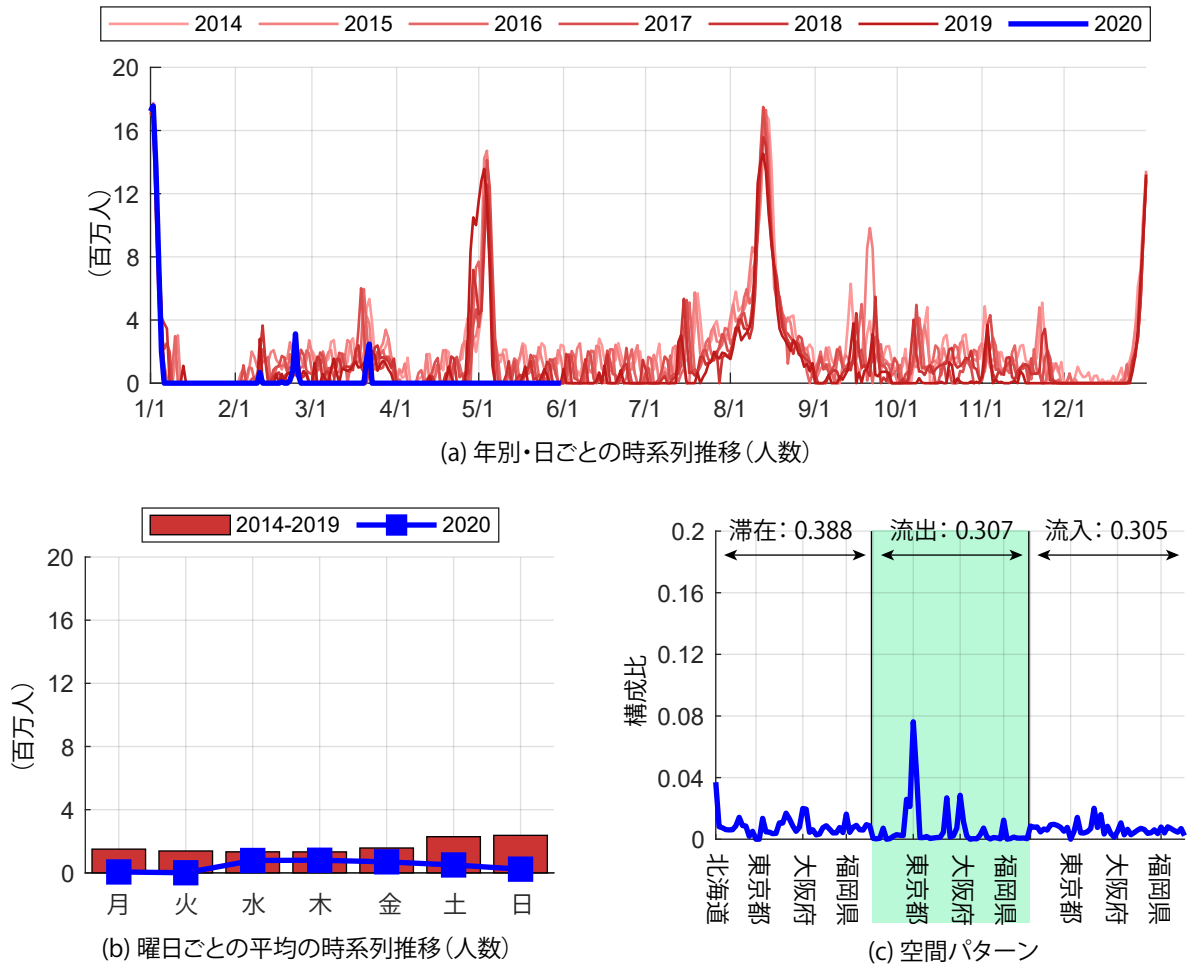


図 4.6: 第 4 基底の推定結果

(iv) 週末の外出行動の時系列推移

図 4.7 から，第 5 基底の情報を解析する。COVID-19 の影響を受けていない期間の各基底の日変動は図 4.7a より、2019 年までの間には大きな季節変動が見られない。図 4.7b より、とくに曜日ごとの変動が大きく、平日と比較して土曜日と日曜日に特に量が多い行動である。以上のような時系列変動の特徴を踏まえると、この基底が示す情報には、休日（週末）の行楽・観光行動が含まれると考えられる。

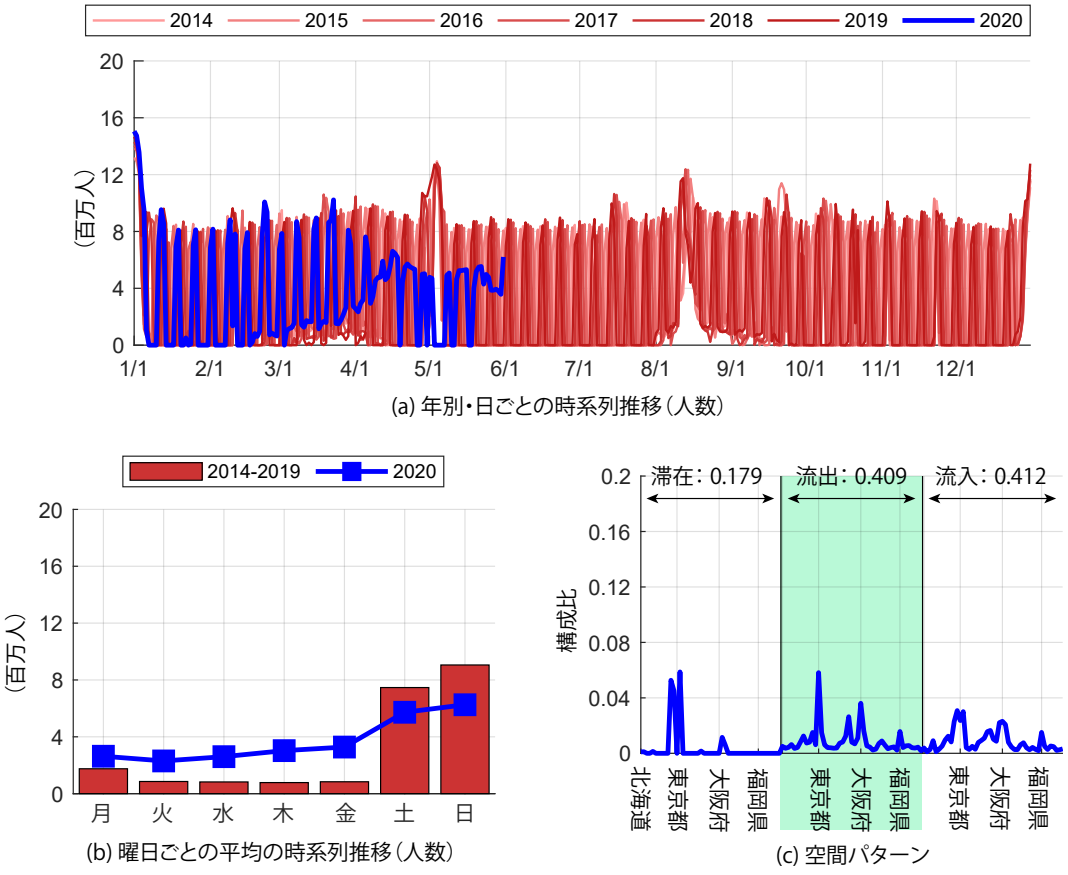
つぎに、図 4.7c より、この行動パターンの空間的な特徴は、流出量（中央の緑の領域）をみると、東京都・大阪府・愛知県といった大都市部で大きく、地方部においても一定の値をとることが分かる。一方で、流入量（最も右の領域）をみると、こちらも関東・関西地方の各県において比較的大きな値をとるが、地方部においても一定の量を占めている。これらから、東京や大阪、愛知等の大都市部からの近県への旅行行動や、地方部からの旅行及び地方部への旅行行動が考えられ、このような特徴のある「週末の外出行動」を示していると考えられる。

この行動の COVID-19 による量の変化を見ていくと、以下のような 3 つの時系列推移が確認できる。

(1) この成分では、3 月上旬の学校の一斉休校などによる行動量の減少は、土日ではあまり確認できない。一方で、通常時ではほぼゼロである、平日の行動量が徐々に増加する傾向にある。この成分の空間パターンの特徴を踏まえると、大都市居住者が平日に郊外に滞在する量が増えたことを意味する。

(2) 4 月上旬の緊急事態が宣言される前後になると、土日のこの成分の行動量が大幅に減少し、ゴールデンウィークの期間にはほぼゼロとなる。一方で，平日の滞在者数が 4 百万人を上回る値で推移している．すなわち、この行動パターンについては、COVID-19 における行動変化が大きかった時期に、土日と平日で反転するような現象が起こっている。

(3) 5 月後半の緊急事態宣言解除後には、土日の行動量について回復の兆しがやや見られる。



(a) 年別・日ごとの時系列推移（人数）

(b) 曜日ごとの平均の時系列推移（人数）

(c) 空間パターン

図 4.7: 第 5 基底の推定結果

(v) 残差の時空間分布

次に、$\boldsymbol{Y} - \boldsymbol{HU}$ で算出される残差行列の時空間分布を解析する。ここで表される残差は、上述したスパース非負値行列因子分解の 5 つの基底では十分に特徴が抽出されなかった要素であり、元データに対する近似の損失分を分析できる。COVID-19 等による滞在行動の変化が、上記で述べた 5 つの行動分類（基底）ごとに日本全体で一律に起こった場合は、それぞれの日変動成分の大小で説明可能であるため、残差は大きく変わることはない。一方で、5 つの行動内容（業務・通勤行動、県内滞在、長期休暇の大都市から地方への行動、週末の外出行動）の空間分布と大きく乖離する変化が起こっていた場合には、非負値行列因子分解で導出した基底ではその変化を説明することができずに、その時点の残差の値が大きくなる。

図 4.8 は、元データと罰則付き非負値行列因子分解における残差の 2 乗和を、日付ごとに算出したものである。図 4.8 より、2020 年では COVID-19 流行期と重なる 2020.3.29(日)（外出自粛要請後初めての週末）以降において残差が非常に大きく、例年とは極めて異質な行動パターンであった。COVID-19 の影響がない時点のなかで、次に大きい残差が確認されたのは、2019.10.12（土）と 2018.9.4（火）である。これらは、記録的暴風が日本に接近した日であり、前者は台風 19 号が静岡県や関東甲信越・東北地方に接近し、人的被害の他、土砂災害、家屋の崩壊、浸水、断水、停電等の被害をもたらした。そして、交通面を見ると関東地方を中心に大規模な計画運休が実施された日であり、北陸新幹線がこの日より長期間の運休を余儀なくされた。後者は、台風 21 号が近畿地方に接近し、大規模停電や関西国際空港の浸水に伴う閉鎖等の被害をもたらした時期である。残差として特徴が抽出されたことは、これらの時期には、平常時とは大きく異なる行動変化が起こっていたことを意味している。

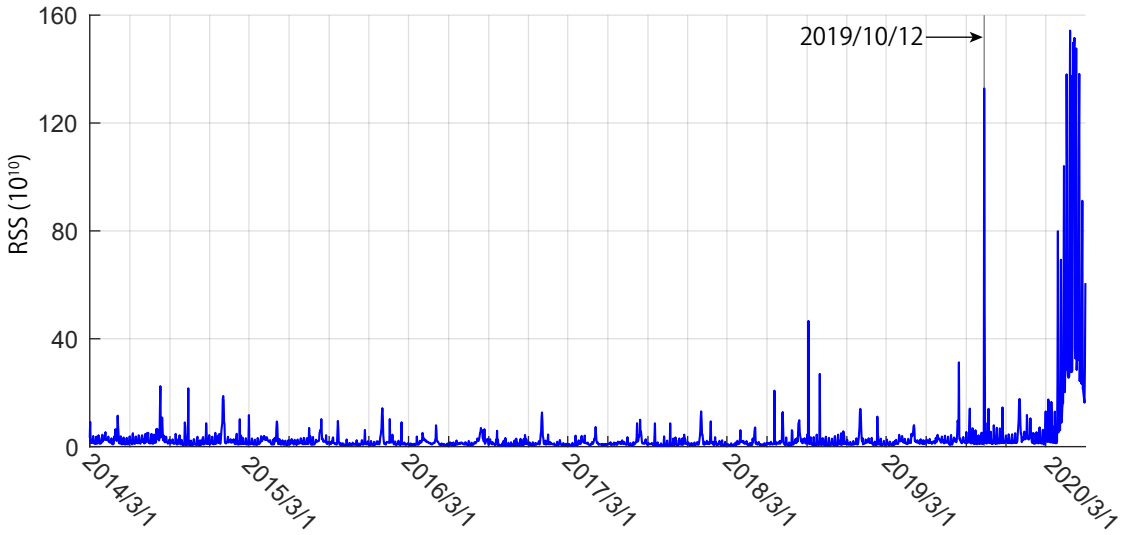


図 4.8: 元データと罰則付き非負値行列因子分解における時系列残差ベクトル 2284 日分

図 4.9 と図 4.10 は、それぞれ 2020.3.1（日）-2020.5.31（日）の COVID-19 の影響が大きいと予想される大きな残差が確認された期間と、2019.10.12（土）の 2019 年台風 19 号による大きな残差が確認された期間の、平均的な残差の空間パターンの内訳を表している。図 4.9 の COVID-19 時期の残差をみると、東京をはじめとする関東地域の居住都県内滞在者数が大きく正の残差を示しており、そのほかの地域で負の値をとっている。すなわち、COVID-19 による居住都道府県内の滞在者数増加は、関東地域の各県で特に大きい行動量変化があったことを示している。

一方で、2019 年の台風 19 号が接近していた時の残差ベクトルを示した図 4.10 を見ると、東京をはじめとする関東地域の自地域内滞在パターンが正の特徴を示しており、さらに東京への流入量が顕著に負の値をとっている。この台風による被害としては、長野県や宮城・福島県において甚大な洪水被害が発生していたが、都道府県間移動量の低下（滞在量の増加と流出・流入量の減少）は、東京都を中心に関東地方に特に偏在していた。これは、日常的に鉄道の利用率が高い地域において、計画運休による行動変化の影響が大きいことを示唆している。

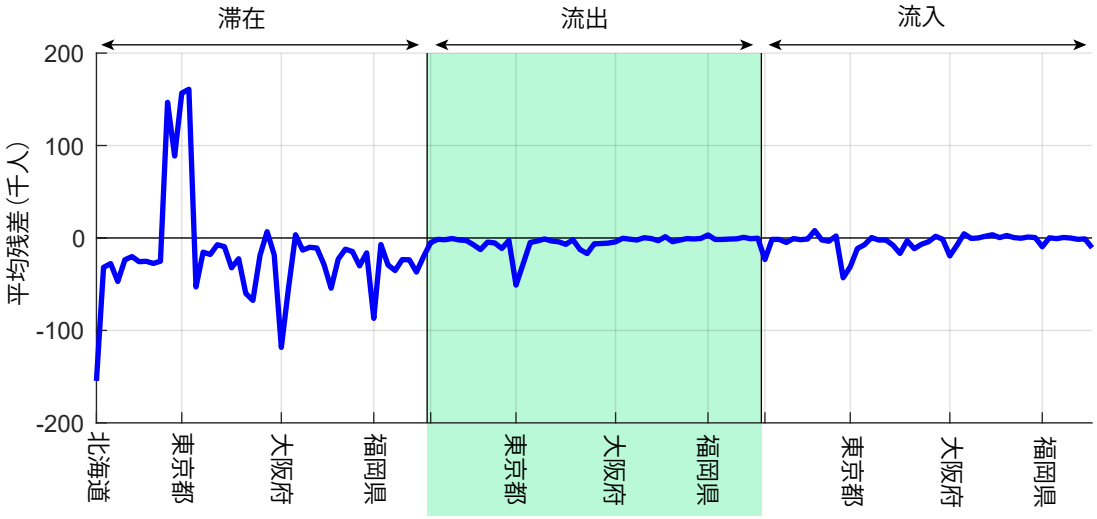


図 4.9: COVID 流行時の残差ベクトル（2020.3.1-2020.5.31 の平均値）

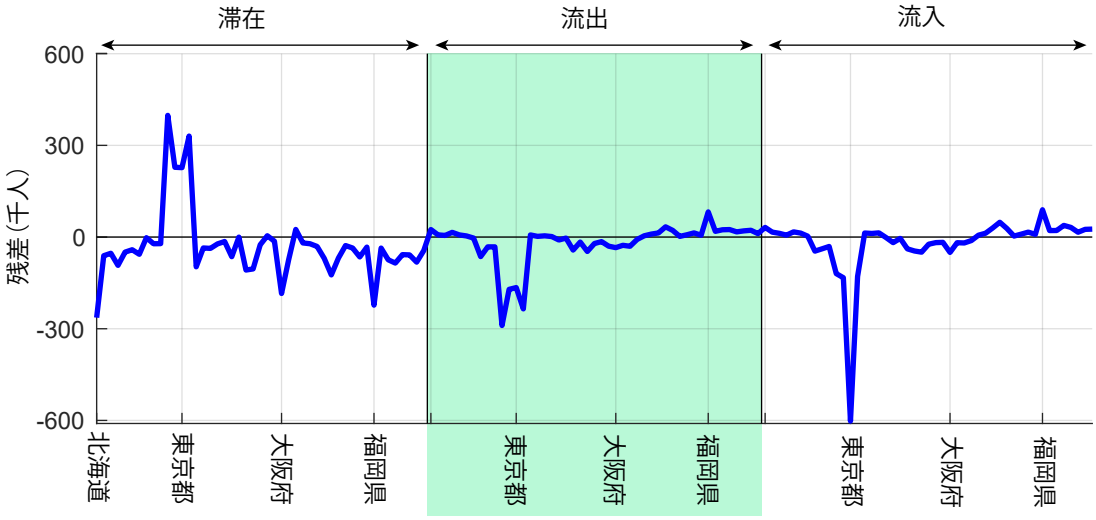


図 4.10: 令和元年台風 19 号接近時の残差ベクトル（2019.10.12）

4.2 羽咋市における定常的な人口流動分析に関する研究

4.2.1 概要

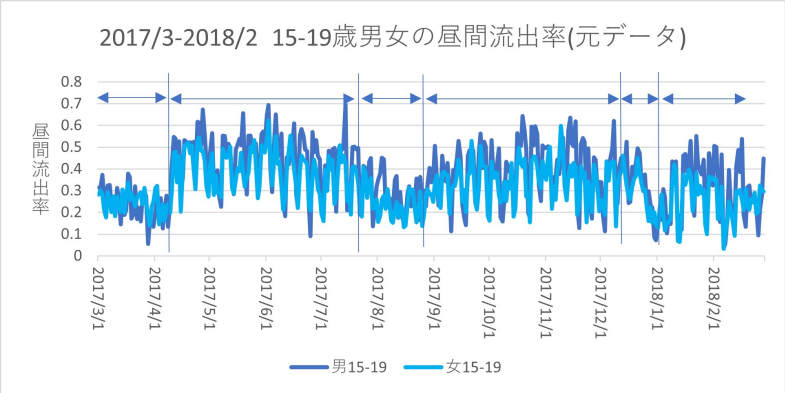
本節 [68] では非負値行列因子分解とスパース非負値行列因子分解を用いて行列分解及び抽出される特徴の違いについて検討している。また両手法において、モデルで抽出しきれなかった特徴を残差構造から解析する。具体的な使用データとして石川県羽咋市居住者の、年代・男女別（15-19 歳、20-29 歳、30-39 歳、40-49 歳、50-59 歳、60-69 歳、70 歳-79 歳の 7 年代区分 ×2 属性）1461 日分（2014 年 3 月 1 日-2018 年 2 月 28 日）の昼夜間人口データを使用した。非負値行列因子分解の適用により、基底数を 3 とした場合、概ね「15-19 歳の男性」、「15-19 歳の女性」、「20-59 歳男女」の基底と解釈される特徴が抽出され、いずれも平日の流出構造は「通勤・通学」に伴うものであった。また、スパース非負値行列因子分解では「15-19 歳の男性」、「15-19 歳の女性」、「羽咋市民」の基底が抽出されたが、「15-19 歳の男性」、「15-19 歳の女性」の特徴はいずれも他の属性がほぼ 0 と推定されており、明瞭な解釈が可能となった。残差項からは日曜日に関する特徴が抽出され、特に 20-39 歳男女における流出構造が大きく抽出された。羽咋市民の特徴を概観すると、平日は通勤・通学に伴う市外への流出パターン、休日は 20-39 歳の市外への流出パターン、盆休みや年末年始などは、昼夜間人口の差が極めて小さく、同期間中は市民の多くが羽咋市内に滞在していることが示された。

4.2.2 使用データと設定について

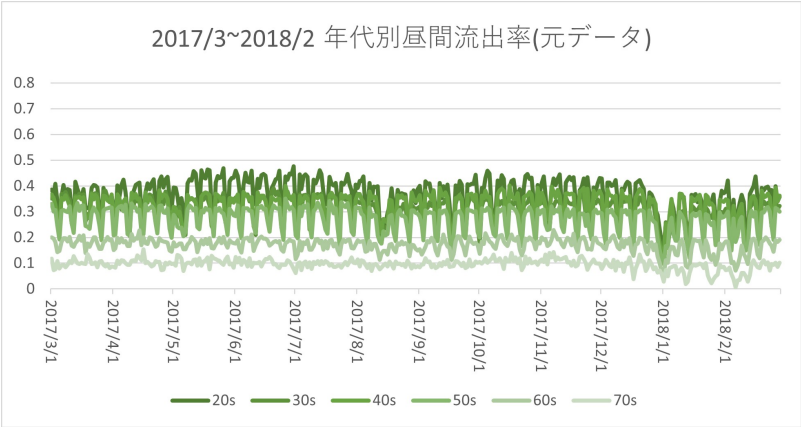
(i) 使用データについて

本研究の使用データとして石川県羽咋市居住者の、年代・男女別（15-19 歳、20-29 歳、30-39 歳、40-49 歳、50-59 歳、60-69 歳、70 歳-79 歳の 7 年代区分 ×2 属性）1461 日分（2014 年 3 月 1 日-2018 年 2 月 28 日）の昼夜間人口データを用いた。ここで羽咋市居住者とは、羽咋市内にて契約を行った携帯電話を有する者であり、羽咋市居住者以外の他市への流出及び他市からの流入は含まれない。また、夜間人口は 3 時、昼間人口は 13 時における羽咋市内滞留者数を表す。

本研究では羽咋市民の昼夜間時における滞留・流出パターンの解析を目的に、属性別に夜間人口から昼間人口の差分を取り、さらにデータを基準化するため、夜間人口の平均で除した、夜間人口 1 人当たりの「昼間流出率」を算出し、これを使用データとした。このデータの特徴として、値が 0(0%) に近いほど昼夜間の人口差が小さい、すなわち市内に滞留していることを表し、1(100%) に近いほど昼間人口が他市へ流出していることを表す。図 4.11 は 2017 年 3 月-2018 年 2 月における性別・年代別昼間流出率を可視化したものである。図 4.11a は 15-19 歳男女、図 4.11b は 20-79 歳男女平均に着目した図である。図 4.11 は共通して、横軸は上記期間の時系列、縦軸は昼間流出率であり、縦軸のスケールは統一している。図 4.11b では年代別に色分けがなされており、色が濃いほど若年層、色が薄いほど高齢層を表す。図 4.11a より、15-19 歳男性と 15-19 歳女性の昼間流出率はおおよそ類似した特徴であり、3 月や 8 月、年末年始等は昼間流出率が、それぞれ 24 %、26 %、18 %と、他の区間と比べて低い。その他の区間として 4 月-7 月や 9 月-12 月下旬、1 月上旬-2 月等は相対的に昼間流出率が高く、それぞれ 41 %、37 %、29 %である。これらから 15-19 歳男女は主に通学で市外に流出していると推察され、学校の開校時と閉校時に関して周期性が見られる。他方、図 4.11b より、20-79 歳男女は 5 月上旬（ゴールデンウィーク）、8 月中旬（盆期間）、年末年始等の部分的に昼間流出構造が普段と異なり、市内に滞在している様子が伺えるが、その他期間は昼間流出率が安定的に推移している。また流出の構造は若年層ほどその傾向が高く、高齢層ほど昼間時の流出が少ないことがわかる。また図 4.11a 及び図 4.11b より、15-19 歳男女は周期性を有する行動パターンであり、他の属性と流出構造が異なっていることが言える。



(a) 昼間流出率:15-19 歳男女



(b) 昼間流出率:20-79 歳男女

図 4.11: 携帯電話の位置情報による羽咋市民の年代別昼間流出率

(ii) 分析方法の設定

本節では、基底行列 $\boldsymbol{H}$ 及び表現行列 $\boldsymbol{U}$ の両側に L_1 正則化項を付与したモデルを使用している。[59] 観測データ行列を $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}_+^{n \times m}$、基底行列を $\boldsymbol{H} = [\boldsymbol{h_1}, \cdots, \boldsymbol{h_k}] \in \mathbb{R}_+^{n \times k}$, 表現行列を $\boldsymbol{U} = [\boldsymbol{u_1}, \cdots, \boldsymbol{u_m}] \in \mathbb{R}_+^{k \times m}$ とするとき、目的関数を D_{S-NMF} とすると、以下のように表せる、

$$D_{S-NMF} = \|\boldsymbol{X} - \boldsymbol{HU}\|_F^2 + \mu\|\boldsymbol{H}\|_1 + \lambda\|\boldsymbol{U}\|_1 + tr(\boldsymbol{\Psi H}^T) + tr(\boldsymbol{\Phi U}^T), \tag{4.6}$$

ただし、μ, λ はハイパーパラメータ、$\boldsymbol{\Psi}, \boldsymbol{\Phi}$ はそれぞれ基底行列 $\boldsymbol{H}$、表現行列 $\boldsymbol{U}$ のラグランジュ乗数である。本節の使用データに併せて $n = 1461$(日)、$m = 14$(属性) である。また基底数は大きな人の流れを解析することを目的に、解釈を優先して基底数 $k = 3$ とした。L_1 正則化による制約付き非負値行列因子分解の計算アルゴリズムは以下の通りである、

$$\boldsymbol{H}_{i,j} \leftarrow \boldsymbol{H}_{i,j} \frac{(\boldsymbol{XU^T})_{i,j}}{(\boldsymbol{HUU^T})_{i,j} + \mu \boldsymbol{1}_{i,j}}, \qquad \boldsymbol{U}_{i,j} \leftarrow \boldsymbol{U}_{i,j} \frac{(\boldsymbol{H^T X})_{i,j}}{(\boldsymbol{H^T HU})_{i,j} + \lambda \boldsymbol{1}_{i,j}}, \qquad subject\ to \sum_m \boldsymbol{u_m} = \boldsymbol{1}, \tag{4.7}$$

ここで (4.7) は要素ごとの計算を表している。また $\boldsymbol{1}_{i,j}$ は全ての要素が 1 の行列である。

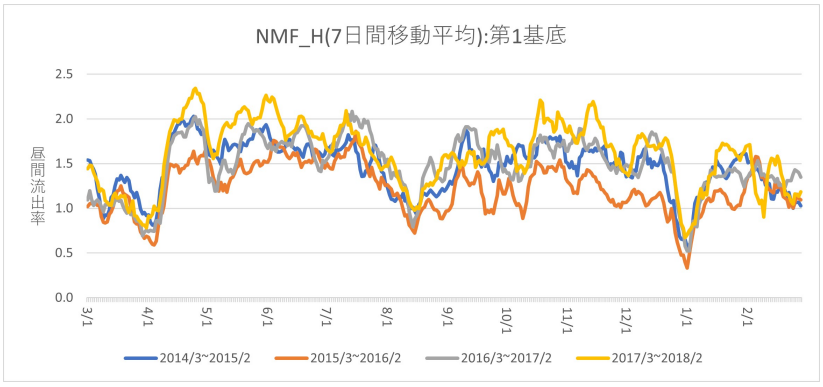
L_1 正則化項のハイパーパラメータの決定として、本節では二乗誤差項のオーダーと概ね等しい値となる $\mu = 10^{-3}, \lambda = 10^{1}$ の値を選択した。このような値を設定した場合、(4.6) において、非負制約に基づくラグランジュの未定乗数項を除き、いずれも概ね等しいオーダーとなる。また、(4.7) において、$\boldsymbol{H}$、$\boldsymbol{U}$ の両更新式ともに、ハイパーパラメータの影響により、一部の要素において分母が分子よりも大きくなるため、$\boldsymbol{H}$ 及び $\boldsymbol{U}$ の値を 0 に近づけている、すなわちスパースの効果が表れると考え、上記の設定を行った。

4.2.3 分析結果

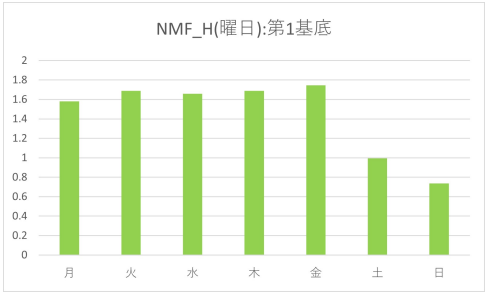
(i) 非負値行列因子分解による分析結果

基底数を 3 とした時の非負値行列因子分解による分析結果をそれぞれの基底ごとに述べる。基底行列 $\boldsymbol{H}$ に関して時系列パターンと曜日周期パターンの 2 種類の図があり、時系列パターン (図 4.12a から図 4.14a) では横軸が 1 年分の時系列、縦軸が昼間流出率を表し、7 日間移動平均をとった 4 年分の時系列パターンをそれぞれ図中にプロットしている。またその時系列パターンのうち、曜日に関して平均を算出したものが曜日周期のパターン (図 4.12b から図 4.14b) であり、横軸は月曜日から日曜日までの曜日周期、縦軸は曜日ごとの昼間流出率を表している。表現行列 $\boldsymbol{U}$(図 4.12c から図 4.14c) は横軸が性別・年代別の属性であり、縦軸は総和が 1 の重みを表している。

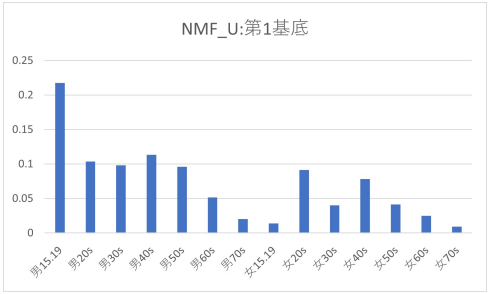
1 つ目の特徴として、「15-19 歳男性」が抽出された。図 4.12a より、時系列パターンでは、3 月や 8 月、年末年始などに他市への流出が少なく、その他の期間は流出が多い。また、図 4.12b より、曜日周期パターンでは、平日の方が土日より流出が多い。図 4.12c より、15-19 歳男性の特徴が 0.22 と最も高く抽出されており、20-59 歳男性や 20-29 歳女性など、多くの属性に共通した特徴が抽出されている。これらの結果から、概ね「15-19 歳男性」の特徴と解釈され、3 月や 8 月、年末年始等はそれぞれ春休み、夏休み、冬休みに該当し、平日は市外の学校や会社に通勤・通学し、土日は市内に滞留している特徴がある。



(a) 基底ベクトル (時系列パターン H):第 1 基底



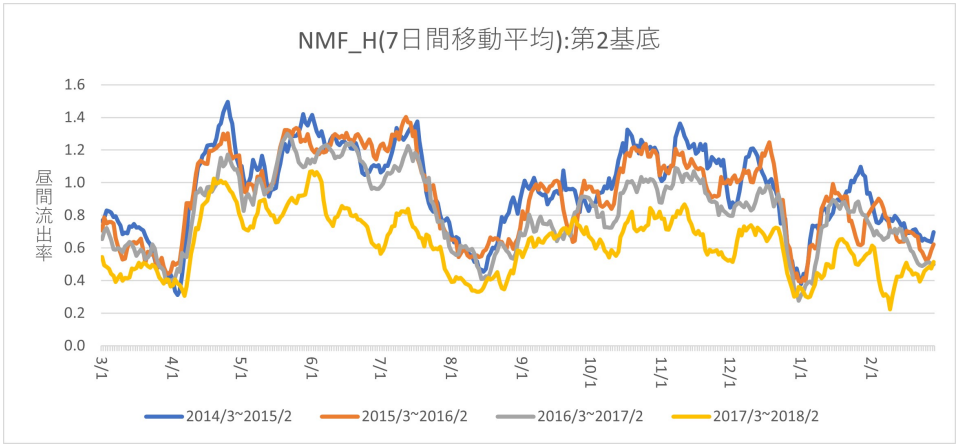
(b) 基底ベクトル (曜日別集計 H):第 1 基底



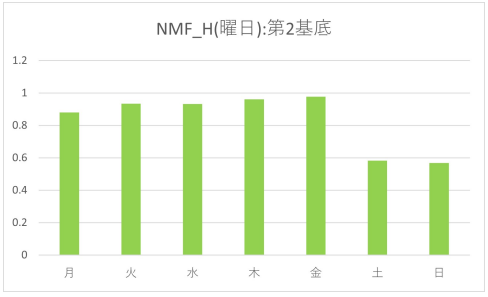
(c) 表現ベクトル (属性パターン U):第 1 基底

図 4.12: 非負値行列因子分解:第 1 基底

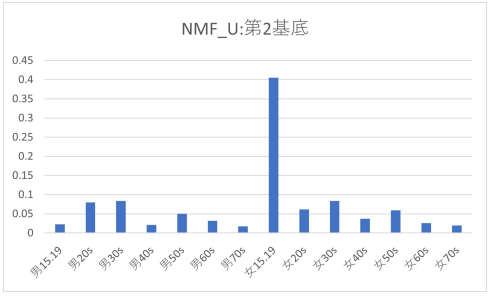
2 つ目の特徴として、「15-19 歳女性」が抽出された。図 4.13a より、時系列パターンでは、3 月や 7 月中旬-8 月、年末年始などに他市への流出が少なく、その他の期間は流出が多い。図 4.13b より、曜日周期パターンでは、平日の方が土日より流出が多い。図 4.13c より、15-19 歳女性の特徴が 0.41 と最も高く抽出されている。これらの結果から、概ね「15-19 歳女性」の特徴と解釈され、「15-19 歳男性」と類似した特徴を示している。すなわち 3 月や 7 月中旬-8 月、年末年始はそれぞれ春休み、夏休み、冬休みに該当し、平日は市外の学校や会社に通勤・通学し、土日は市内に滞留している特徴がある。また、「15-19 歳女性」の方が、U 側で特徴が突出しているため、より「学生らしい」行動パターンとなっている。特に「夏休み」について石川県の公立高校は 7 月中旬頃から 8 月末日までを長期休暇とする特徴があり、その影響で学校が休みとなり、羽咋市内に滞留する時系列パターンとなっていることが推察される。



(a) 基底ベクトル (時系列パターン H):第 2 基底



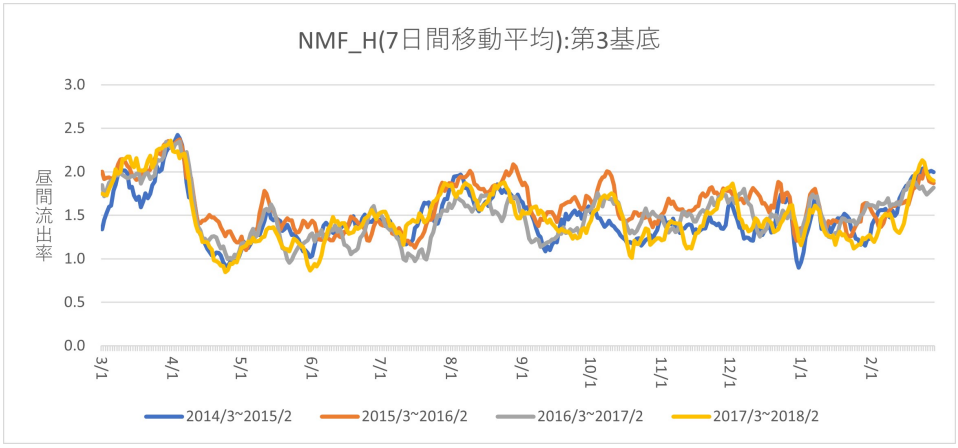
(b) 基底ベクトル (曜日別集計 H):第 2 基底



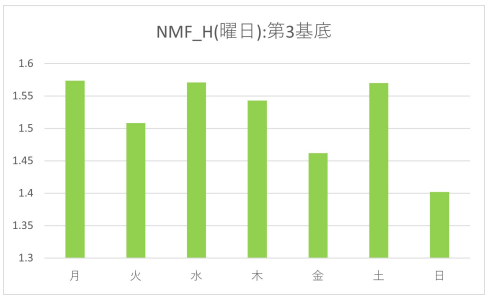
(c) 表現ベクトル (属性パターン U):第 2 基底

図 4.13: 非負値行列因子分解:第 2 基底

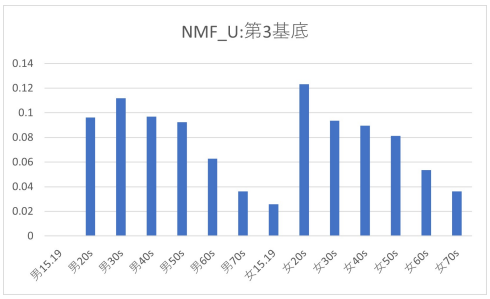
3 つ目の特徴として、「20～59 歳男女」が抽出された。図 4.14a より、時系列パターンでは、2 月末-3 月末、7 月中旬-8 月末に昼間流出率が高い。その他期間では、同比率は概ね安定的に推移している。図 4.14b より、曜日周期パターンでは、日曜日の昼間流出率は相対的に低いが土曜日は高い。図 4.14c より、20-29 歳女性が 0.12 であることを筆頭に 20-59 歳男性や 20-59 歳女性の特徴が高く抽出されている。これらの結果から、概ね「20-59 歳男女」の特徴と解釈される。平日は通勤パターンであるが、2 月末-3 月末や 7 月中旬-8 月末あたりに昼間流出率が高いことや、土曜日も同比率が高いことから、市外での宿泊を含む旅行に関する流出パターンも推察される。また 4 月はこれらの属性において転出が発生していると考えられる。例えば今まで羽咋市内に在住しており、就労で市外に流出していた者が転出を行うと、羽咋市内の昼間人口は変化しないが、夜間人口は減少し、それによって昼夜間の人口差が減少するため昼間流出率は減少することとなる。そのため 4 月は他の期間よりも昼間流出率が低いことが推察される。



(a) 基底ベクトル (時系列パターン H):第 3 基底



(b) 基底ベクトル (曜日別集計 H):第 3 基底



(c) 表現ベクトル (属性パターン U):第 3 基底

図 4.14: 非負値行列因子分解:第 3 基底

以下では非負値行列因子分解の残差解析について示す。非負値行列因子分解を適用することで、上記で示された 3 つの特徴が抽出されたが、他方で元データを 3 次元に低次元近似した際の特徴空間上では捉えられなかった構造が残差項として存在する。これは、上述の 3 つの特徴パターンの増減だけでは表せない（表せる場合は残差が限りなく 0 に近くなるため）特徴であり、この残差構造を分析することで他の特徴が捉えられる場合がある。

図 4.15 は元データに対する非負値行列因子分解モデルでの近似残差の 2 乗構造を表しており、属性パターンごと（列方向）に曜日別で平均をとった「属性残差二乗ベクトル」である。縦軸は残差を表しており、縦軸の値が大きいほど、モデルと元データの残差が大きいことを示す。図 4.15 より、曜日別残差平均の総和では日曜日の残差が 3.47 と最も大きく、次いで土曜日が 2.11、平日が 1.52 であることから、上記の基底で特に平日の特徴を抽出していたと言える。属性別の残差二乗ベクトル総和は、20-29 歳男性の残差が 1.43 と最も高く、次いで 20-29 歳女性が 1.03、30-39 歳男性が 0.78、30-39 歳女性が 0.59 と続いている。その一方で、第 1 基底（図 4.12）や第 2 基底（図 4.13）で抽出された、それぞれ 15-19 歳男性、15-19 歳女性の残差は小さく、モデル内で特徴が抽出できていることがわかる。20-39 歳男女の属性は第 3 基底（図 4.14）でも大きく抽出されていたが、図 4.14b より、それは平日や土曜日に関する特徴であった。対して残差においてこれらの属性は、特に日曜日に関して大きな値をとっている。3 つの基底では日曜日に関する特徴を大きく抽出されていないことを踏まえると、非負値行列因子分解モデルでは日曜日に関する近似を過小評価していることが推察される。これらを総括して、羽咋市の 20〜39 歳男女は非負値行列因子分解モデル及び残差分析を併せて、平日においても、また土日においても他の属性より大きな流出構造を持っていることが推察される。

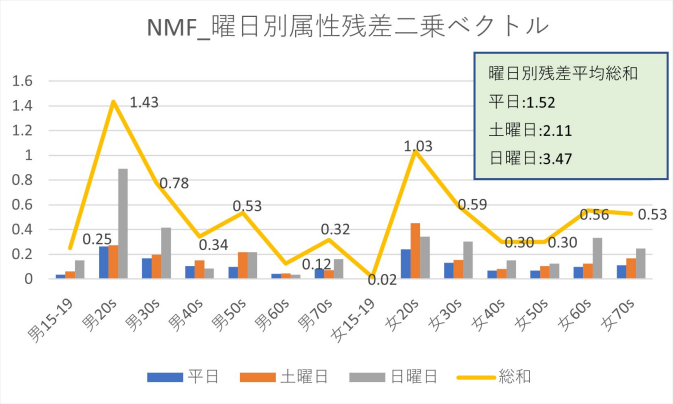
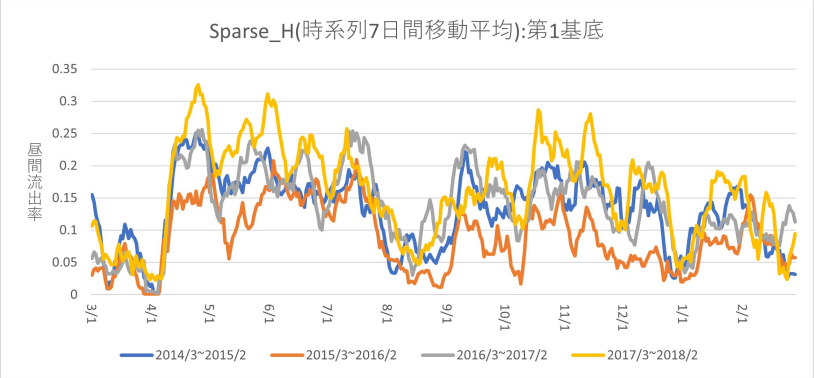


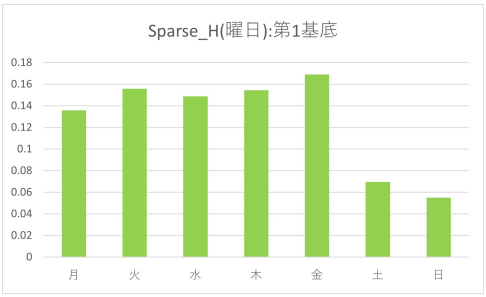
図 4.15: 非負値行列因子分解モデルにおける曜日別残差分析

(ii) スパース非負値行列因子分解による分析結果

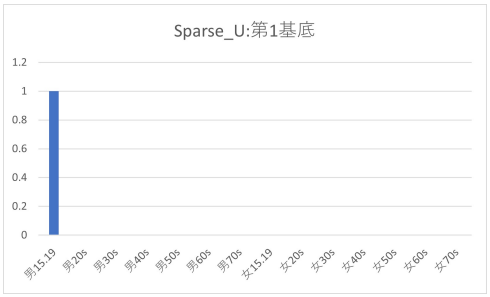
1 つ目の特徴として、「15-19 歳男性」が抽出された。図 4.16a より、時系列パターンでは、3 月や 8 月、年末年始等に他市への流出が少なく、その他の期間は流出が多い。図 4.16b より、曜日周期パターンでは、平日の方が土日より流出が多い。図 4.16c より、15〜19 歳男性の特徴のみが抽出されている。これらの結果から、「15-19 歳男性」の特徴であり、3 月や 8 月、年末年始はそれぞれ春休み、夏休み、冬休みに該当し、平日は市外の学校や会社に通勤・通学し、土日は市内に滞留している特徴がある。L_1 正則化項の影響により、U 側の主要な要素以外の特徴を 0 にすることができ、属性が 15-19 歳男性のみとなった。したがって時系列パターンも 15-19 歳男性の行動パターンを抽出したものであると推察できる。



(a) 基底ベクトル (時系列パターン H):第 1 基底



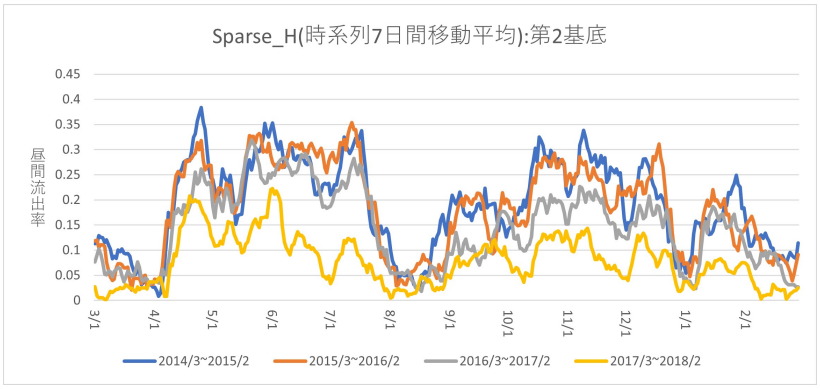
(b) 基底ベクトル (曜日別集計 H):第 1 基底



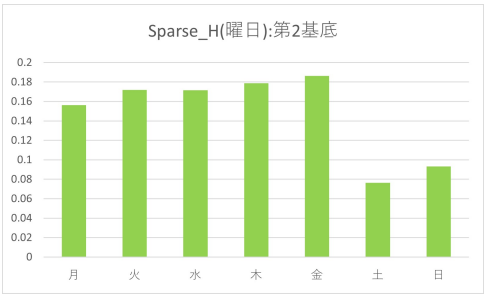
(c) 表現ベクトル (属性パターン U):第 1 基底

図 4.16: スパース非負値行列因子分解:第 1 基底

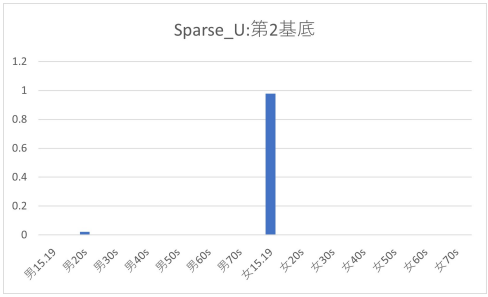
2 つ目の特徴として、「15-19 歳女性」が抽出された。図 4.17a より、時系列パターンでは、3 月や 7 月中旬から 8 月、年末年始などに他市への流出が少なく、その他の期間は流出が多い。図 4.17b より、曜日周期パターンでは、平日の方が土日より流出が多い。図 4.17c より、15-19 歳女性の特徴が 0.98 と極めて高く抽出されている。これらの結果から、概ね「15-19 歳女性」の特徴と解釈される。図 4.16c の「15-19 歳男性」と同様に L_1 正則化項の影響によりほぼ 15-19 歳女性以外の特徴が 0 となったため、同属性の特徴が際立ったものとなっており、それに伴って時系列パターンの解釈を容易なものとしている。



(a) 基底ベクトル (時系列パターン H):第 2 基底



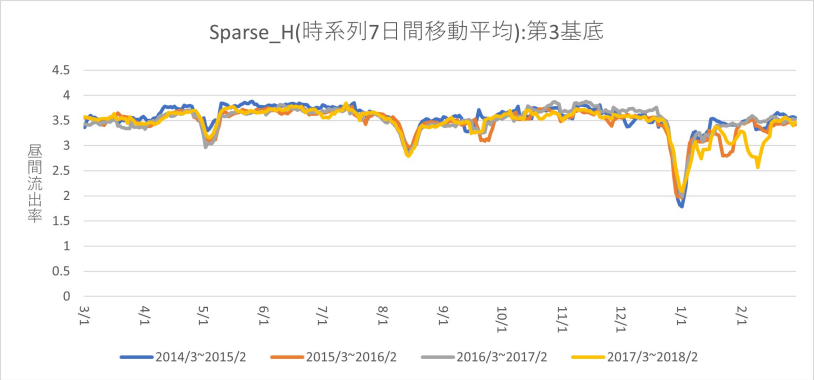
(b) 基底ベクトル (曜日別集計 H):第 2 基底



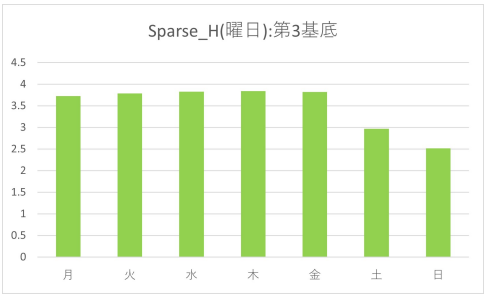
(c) 表現ベクトル (属性パターン U):第 2 基底

図 4.17: スパース非負値行列因子分解:第 2 基底

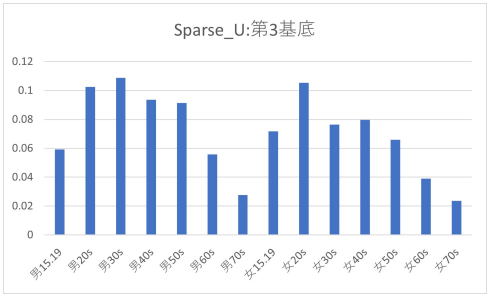
3 つ目の特徴として、「羽咋市民全体」が抽出された。図 4.18a より、時系列パターンでは、GW、盆休み、年末年始に昼間流出率が低くなっており、その他は安定的に推移している。図 4.18b より、曜日周期パターンでは、平日の方が土日より流出が多い。図 4.18c より、20-29 歳女性や 20-39 歳男性の特徴が 0.11 ずつ占めていることを筆頭に全ての属性で特徴が抽出されている。これらの結果から、概ね「羽咋市民全体」の特徴と解釈される。羽咋市民の基本的な行動パターンとして、平日の方が土日よりも昼間流出が多く、安定した周期的行動をとり、また GW や盆休み、年末年始は市内に滞留している特徴があることが推察される。



(a) 基底ベクトル (時系列パターン H):第 3 基底



(b) 基底ベクトル (曜日別集計 H):第 3 基底



(c) 表現ベクトル (属性パターン U):第 3 基底

図 4.18: スパース非負値行列因子分解:第 3 基底

以下ではスパース非負値行列因子分解近似の残差項について示す。図 4.19 は図 4.15 と同様、属性パターンごと（列方向）に曜日別で平均をとった「属性残差二乗ベクトル」である。曜日別残差平均の総和では日曜日の残差が 4.47 と最も大きく、次いで土曜日が 2.61、平日が 1.72 であることから、上記の基底で特に平日の特徴を抽出していたと言える。属性別の残差二乗ベクトル総和より、20-29 歳男性の残差が 1.43 と最も高く、次いで 20-29 歳女性が 1.31、30-39 歳女性が 0.92、60-69 歳女性が 0.78 と続いている。これらの属性は第 3 基底（図 4.18）でも大きく抽出されていたが、図 4.18b より、それは平日に関する特徴であった。加えて第 1 基底（図 4.16）、第 2 基底（図 4.17）の基底がそれぞれ 15-19 歳男性、15-19 歳女性を説明する基底であったことから、第 1-第 3 基底までで大きく抽出されなかった土日に関する特徴が残差として表れ、非負値行列因子分解モデルの残差 (図 4.15) より大きいことが推察される。

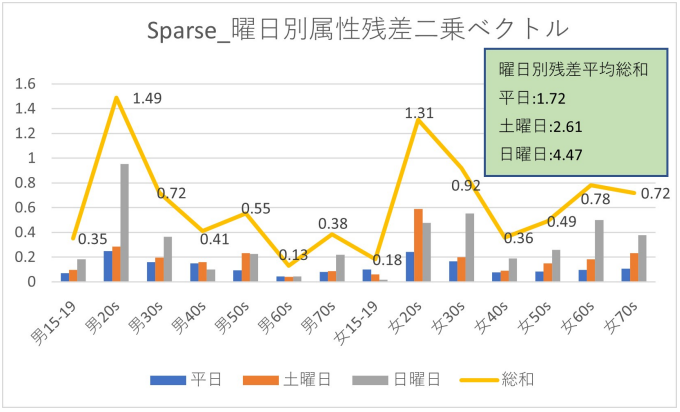


図 4.19: スパース非負値行列因子分解モデルにおける曜日別残差分析

第5章　主成分分析に関する応用研究

5.1　地域産業構造の特徴抽出と可視化に関する研究

本節 [69] では産業連関表に主成分分析を適用することで、各都道府県における産業構造の特徴抽出を行った。産業連関表から特徴を抽出する例は「3.1 節 非負値行列因子分解法を用いた産業特性の抽出 [原田・寒河江 (2019)]」で分析が行われているが、ここでは主成分分析を用いて 2 つの観点から産業構造の特徴について解析している。

1 つ目は 47 都道府県の産業構造の特徴を全国産業連関表の固有空間上に射影し、同一空間上での比較を可能とした上で、近い特徴同士でクラスターを作成し、類似度を可視化する手法であるクラスター分析を適用した。全国産業連関表 (34 産業) に主成分分析を適用した結果、第 1 主成分として「輸送機械業」、第 2 主成分として対事業所サービスや商業等の「第 3 次産業」に関する特徴が抽出された。これら 2 つが形成する固有空間に 47 都道府県の特徴を射影した結果、3 つのクラスターに分類した場合、第 1 主成分では「愛知県」、神奈川・静岡等を含む「製造圏」とその他に対応するクラスターが形成された。同様に第 2 主成分では「東京都」、愛知県・神奈川県・大阪府等の「大都市」とその他に対応するクラスターが形成された。

2 つ目は都道府県の 2005 年及び 2011 年産業連関表に対し、それぞれ主成分分析を適用することで、経年変化による産業構造の違いについて可視化を行っている。特に本節では対象地域を「東京都」「福島県」「石川県」の 3 県に絞って解析している。福島県は東日本大震災の影響を受け、県内での財・サービスの取引が減少した結果、電気・ガス等が県の基盤産業の特徴として抽出された。また「医療・保健・介護」の特徴や輸送業含む「運輸」の特徴など、震災及びその復興支援として発生した特需効果と考えられる特徴も併せて抽出された [70]。東京都では特に不動産業において、産業連関表の観測時の 2005 年から 2011 年の間に特に商業地において「ミニバブル」[72] が起こっており、それによって 2011 年の産業連関表から不動産業に関する特徴が抽出された。石川県ではリーマンショックの影響 [73] から特に 2005 年では石川県の基幹産業であった商業や情報通信業等に関して 2011 年では生産規模が縮小していた。

これらの結果から、主成分分析と可視化の工夫により、地域の産業構造の類似性及び経年変化を客観的に解析することが出来た。

5.2　使用データと設定について

5.2.1　使用データについて

本節で用いる産業連関表は、都道府県単位で作成された 2005 年 (平成 17 年) と 2011 年 (平成 23 年) の内生部門表データ（大分類）及び 2011 年の全国産業連関表 (大分類) を使用データとしている。2005 年版は 34 分類、2011 年版は 37 分類であったため、2011 年版を 34 分類になるよう次のように再構成した。

具体的な再構成の方法として、総務省 [5]「平成 23 年産業連関表作成基本要綱」の平成 17 年表との相違点 [別表 4] 部門分類対応表 (4) 統合大分類を参考に、2011 年産業連関表にある「はん用機械」「生産用機械」は「一般機械」、「水道」「廃棄物処理」は「水道・廃棄物処理」、「業務用機械」は「一般機械」と「精密機械」、「プラスチック・ゴム」は「その他の製造工業製品」とし、2005 年産業連関表の 34 分類に合わせた。

統一的に作成されている産業連関表であるが、都道府県によって産業部門の設定に相違がみられたため、できる限り全国産業連関表の産業部門に合わせる形で修正を加えた。具体的には、全国産業連関表では 1 つの産業部門であるが都道府県版では複数の部門に分割されているものは値を合算し、一方で全国産業連関表では複数の産業部門となっているものが都道府県版では 1 つの部門として表されているものは、部門数に応じて値を等配分した。また、都道府県独自の産業部門（例えば、愛知県では「自動車」、「航空機」、「その他輸送機械」がある）を有する場合は、全国表にある最も近いと思われる部門（前述の例では、「輸送機械」）へ集約した。内生部門データは、34 に分類された産業（行方向：供給部門）がどの産業（列方向：需要部門）にいくら販売したか、逆に言えば、34 の産業（列方向：需要部門）は、どの産業（行方向：供給部門）からいくら購入したかを表している。

内生部門表は、都道府県によって金額の単位が大きく異なることに加えて、産業部門によっても取引金額（生産額）に大きな差が見られるため、分析の際に内生部門表の値をそのまま使用することは、金額に応じた影響が反映されると考えてよい。一方で、産業間の取引構造が類似しているにもかかわらず、取引金額が小規模なためにそれらの情報が表面化しないという状況も考えられる。その場合、データを列ごとに平均 0、分散 1 の処理によって標準化し、金額の大小（単位）の影響を排除するような分析も可能である。他にも投入係数表を利用する方法や、行列総和で除する方法、行和で除する方法、平均 0 分散 1 で正規化するなどの様々な基準化の方法が考えられる。

5.2.2 分析方法の設定

本節では、(1) 全国産業連関表と 47 都道府県産業連関表を用いた都道府県クラスタリング、(2) 都道府県ごとの産業連関表を用いた産業構造の経年変化の 2 つの解析を行っている。以下ではそれぞれの分析手法の設定を示す。

(1) 全国産業連関表と 47 都道府県産業連関表を用いた都道府県クラスタリング

1 つ目の分析では、全国産業連関表の主成分によって形成される固有空間に 47 都道府県の産業連関表データを射影し、共通する固有空間上で 47 都道府県における産業構造の類似性をクラスター分析によって可視化する。全国産業連関表を $\boldsymbol{X_0} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}] \in \mathbb{R}^{n \times n}$、47 都道府県産業連関表の行列を $\boldsymbol{X_I} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}] \in \mathbb{R}^{n \times n}$ $(I = 1, \cdots, 47)$、固有ベクトルの行列を $\boldsymbol{A_k} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_k}] \in \mathbb{R}^{n \times k}$ とする。ただし、n は使用する産業連関表大分類の産業数に対応する、$n = 34$ である。また、k は縮約する次元数である。この時、以下の 4 つのステップで分析を行った。

STEP 1．全国産業連関表 $\boldsymbol{X_0}$ から固有ベクトルを計算する、

$$\boldsymbol{A \Sigma A^T} = \boldsymbol{C} \quad ※ \boldsymbol{C} = \frac{1}{n-1} \boldsymbol{X_0^T X_0}, \tag{5.1}$$

ただし Σ は以下の通りとする、

$$\boldsymbol{\Sigma} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix}.$$

STEP 2．47 都道府県の産業連関表データ $\boldsymbol{X_I}(I = 1, \cdots, 47)$ を全国産業連関表の固有空間 $\boldsymbol{A}$ へ射影する。還元すると全国産業連関表の固有ベクトルを用いて 47 都道府県ごとの主成分得点 $\boldsymbol{Z_I}(I = 1, \cdots, 47)$ を計算する、

$$\boldsymbol{Z_I} = \boldsymbol{X_I A_k}. \tag{5.2}$$

STEP 3.「STEP 2.」で得られたデータをクラスター分析に適用する。ここでクラスター分析は階層型クラスタリング[1]を使用する。類似度 $d(\boldsymbol{Z_I}, \boldsymbol{Z_J})$ の測定はユークリッド距離によって行い、ウォード法によってクラスターを結合させる[2]。ここで $d(\boldsymbol{z_{p,I}}, \boldsymbol{z_{p,J}})$ を I 県と J 県の第 p 主成分得点の乖離度とすると、以下のように計算される、

$$d(\boldsymbol{z_{p,I}}, \boldsymbol{z_{p,J}}) = \|\boldsymbol{z_{p,I}} - \boldsymbol{z_{p,J}}\|_F^2 \quad (p = 1, \cdots, k), \tag{5.3}$$

ここで $\| \ \|_F$ はフロベニウスノルムを表す。すなわち、(5.3) は第 p 主成分得点における I 県と J 県の要素ごとの 2 乗和を計算している。本節では $k = 2$ と設定した。すなわち第 1 主成分得点及び第 2 主成分得点における類似度を測定し、クラスタリングを行っている。

STEP 4.「STEP 3.」で測定した距離が近い地域同士を結合させ、クラスターを作成する。ここで結合の方法としてウォード法を採用した。ウォード法は 2 つのクラスターを統合する際に、クラスター内の平方和が最小となるようにクラスターを形成する手法である。ある 2 つのクラスター u と v を合併して新しいクラスター w をつくるとき、クラスター w と別の任意のクラスター t との間の非類似度 D_wt は、D_ut、D_vt、D_uv を用いて以下の式で表される。ここで n_u、n_v、n_t はそれぞれクラスター u、v、w に含まれるデータ数としている。

$$D_{wt} = \frac{1}{n_u + n_v + n_t}\{(n_u + n_t){D_{ut}}^2 + (n_v + n_t){D_{vt}}^2 - n_t{D_{uv}}^2\}. \tag{5.4}$$

これらの条件から、47 都道府県における産業構造の類似度を測定する。

(2) 都道府県ごとの産業連関表を用いた産業構造の経年変化

2 つ目の分析では、2005 年及び 2011 年産業連関表に対してそれぞれ主成分分析を適用し、得られた固有ベクトルから産業構造の変化について解析している。特に本節では分析する都道府県を福島県、東京都、石川県の 3 都県としている。2005 年及び 2011 年における 3 都県産業連関表の行列を $\boldsymbol{X_{P,year}} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_n}] \in \mathbb{R}^{n \times n}$ (P = 福島県, 東京都, 石川県 : $year$ = 2005, 2011) とすると、主成分分析による固有値分解は以下のように表せる、

$$\boldsymbol{A_{k,P,year}}\boldsymbol{\Sigma_{k,P,year}}\boldsymbol{A^T_{k,P,year}} = \boldsymbol{C_{k,P,year}}, \quad \boldsymbol{C} = \frac{1}{n-1}\boldsymbol{X^T_{P,year}}\boldsymbol{X_{P,year}}. \tag{5.5}$$

(5.5) では 2005 年及び 2011 年の 47 都道県産業連関表それぞれに固有値分解を行っている。本節では $k = 2$ と設定した。すなわち、年・都道府県ごとの第 1 固有ベクトル $\boldsymbol{a_{1,P,year}}$ 及び第 2 固有ベクトル $\boldsymbol{a_{2,P,year}}$ を可視化し、産業構造の経年変化を解析した。

[1]クラスター分析は大きく 2 種類があり、階層型クラスタリングと非階層型クラスタリングがある。非階層型クラスタリングには代表的な解析法として k-means 法などが挙げられ、性質の近しい特徴同士をクラスターとして類型化する解析手法がある。ただし、事前にクラスター数を設定する必要がある。

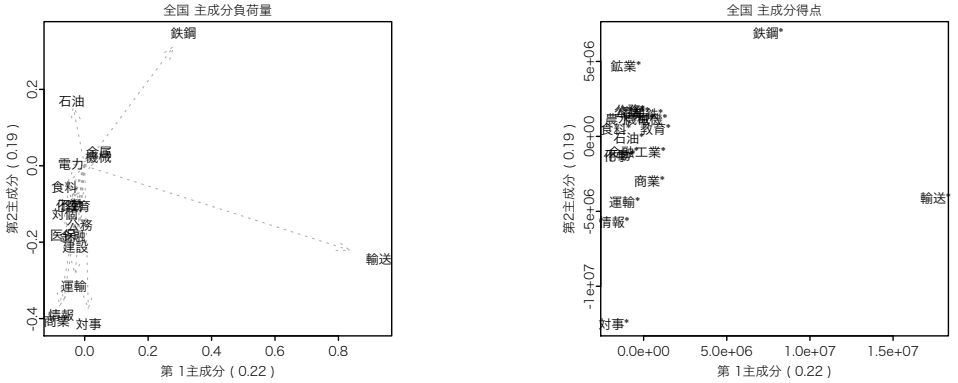
[2]類似度の指標としてユークリッド距離の他に、マンハッタン距離、マハラノビス距離、コサイン類似度などがある。クラスター間の結合方法としてウォード法の他に、群平均法、最短距離法、最長距離法などがある。

5.3 分析結果

5.3.1 2011 年全国産業連関表を用いた都道府県クラスタリング

本節では、全国産業連関表と 47 都道府県分の産業連関表から、産業構造の類似性を測定するためのクラスター分析を行った。図 5.1 は各都道府県データを変換するための 2011 年全国産業連関表に対する主成分分析に関するプロットである。第 1 主成分と第 2 主成分を軸として、図 5.1a が主成分負荷量（固有ベクトル）、図 5.1b が主成分得点の散布図を示したものになっている。また各軸の ( ) 内の数値は寄与率である。例えば、図 5.1 に示した第 1 固有ベクトル及び第 2 固有ベクトルのプロットについて、第 1 固有ベクトルと第 2 固有ベクトルの寄与率の合計が 0.41 であることから、縮約した第 1・第 2 主成分によって、元の 34 次元データの約 41 %の情報を保持しているとみなすことができる。矢印の長さはその主成分における特徴量の大きさを表しており、矢印が長いほど大きな特徴であると言える。また矢印の始点は第 1 主成分、第 2 主成分共に 0 となる地点 (原点) としている。

抽出された産業の特徴に着目すると、図 5.1 より、第 1 主成分 (横軸方向) では輸送機械が主成分負荷量（固有ベクトル）および主成分得点ともに大きな値として特出し、次いで鉄鋼が抽出されている。第 2 主成分 (縦軸方向) では、固有ベクトルとして負の方向に商業や対事業所サービス、情報通信などが抽出されており、正の方向に鉄鋼や石油・石炭製品の特徴が抽出されている。主成分得点では負の方向に対事業所サービスが特出している[3]。すなわち日本全国の産業構造の特徴として第 1 主成分では輸送機械や鉄鋼等の第 2 次産業、第 2 主成分では対事業所サービスを中心とした第 3 次産業に関する特徴などが大きく抽出された。



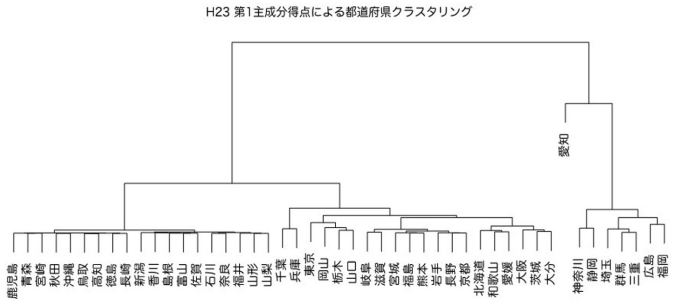
(a) 主成分負荷量 (固有ベクトル)　　(b) 主成分得点

図 5.1: 2011 年全国産業連関表に対する主成分分析の結果

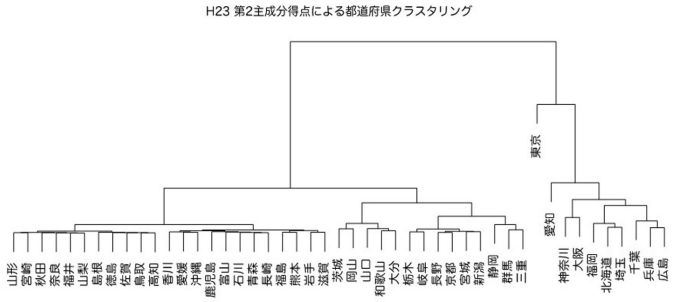
[3]主成分分析では主成分ごとに直交制約がある ((2.14) より) ため、他の主成分との関係から、強い特徴を有する産業においても、負値で抽出されることがある。例えば第 1 主成分で正に突出した輸送機械業は、第 2 主成分では負の値をとっている。そこで本節では特徴の大きさを絶対値の観点から解析している。

図 5.2 は上記で得られた全国産業連関表の固有空間上に 2011 年 47 都道府県分の産業連関表データを射影し、第 1 主成分得点、第 2 主成分得点ごとにクラスター分析を行った結果である。図 5.2a より、愛知県が 1 つのクラスターを形成している。これは、全国産業連関表の第 1 主成分（図 5.1）に影響ある産業部門として「輸送機械」がみられることから、大手自動車産業を有する愛知県が特出していると考えられる。このことは、元の産業連関表の数値においても確認することができる。大手自動車産業を有する愛知県では、輸送機械部門の供給側総額が 6.1 兆円（供給側全国同比 26.7 %)、需要側総額が 9.5 兆円 (需要側全国同比 28.2 %) であり、これは愛知県に次いで輸送機械が突出する神奈川県の供給側総額 1.9 兆円 (8.5 %)、需要側総額 2.9 兆円 (8.5 %) や静岡県の供給側総額 1.8 兆円 (7.6 %)、需要側総額 2.7 兆円（8.0 %）などと比較しても極めて突出していることがわかる。また、このクラスターを 3 つに分ける (3 つに分かれるように横線を引いて分ける) とすると、「愛知県」、神奈川県・静岡県等の「製造圏」、「その他」のクラスターに大分できる。この時、神奈川県・静岡県等の製造圏はその他の地域と異なる産業構造または産業規模を有していると考えられる。

図 5.2b より、第 2 主成分得点では東京都がクラスターとして出現しているが、図 5.1 で確認すると、第 2 主成分は、「商業」、「運輸」、「対事業所サービス」などが特出しており，それらの特徴は東京都の産業として関連づけることができる。またこのクラスターを 3 つに分けるとすると、「東京都」、愛知県・神奈川県・大阪府等の「大都市圏」、「その他」のクラスターに大分できる。



(a) 第 1 主成分得点



(b) 第 2 主成分得点

図 5.2: 47 都道府県の 2011 年版産業連関表を用いたクラスター分析

5.3.2 福島県・東京都・石川県における産業構造の経年変化

ここでは福島県、東京都、石川県の 3 都県の 2005 年及び 2011 年産業連関表に主成分分析を用いて、当該都県における産業間取引の全体構造の把握と可視化を行う。

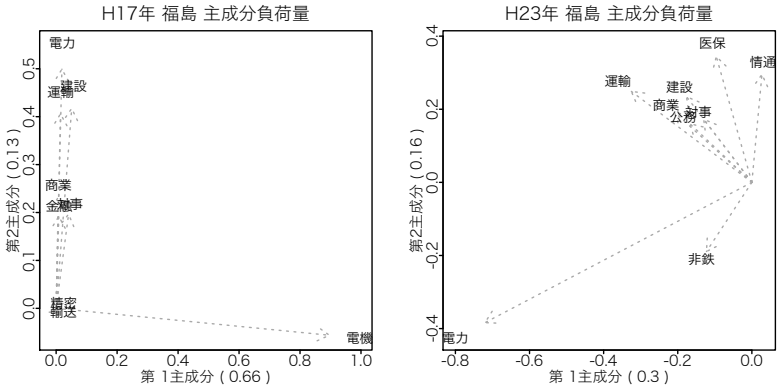
図 5.3 は、2005 年及び 2011 年産業連関表に主成分分析を適用した結果得られた、第 1・第 2 主成分負荷量（固有ベクトル）と第 1・第 2 主成分得点を 3 都県を図示している。これらの図は、ベクトルの方向や長さが、特徴空間に縮約された産業間取引の大きさを表現していると考えられることから、同一地域の 2 時点におけるこれらの構図を視覚的に比較することで、産業構造の推移としてみなすことができると考えている。ここでは、取引金額における、第 1 主成分、第 2 主成分について絶対値の大きい順に 6 産業部門についてのみプロットしている。各軸の ( ) 内の数値は寄与率である。

図 5.3a より、福島県の場合、2005 年に比べて 2011 年では「電力ガス」の影響が大きい。特に福島県は 2011 年に発生した東日本大震災の影響により、産業構造が大きく変化したと考えられる。局所的に特需があったと考えられる産業の特徴も抽出されており、例えば 2005 年には表れなかった「医療・保健・介護」の特徴が 2011 年に表れている。輸送業含む「運輸」の特徴も震災時に発生した運搬に関する特需による効果として特徴が表れたと考えられる [70]。

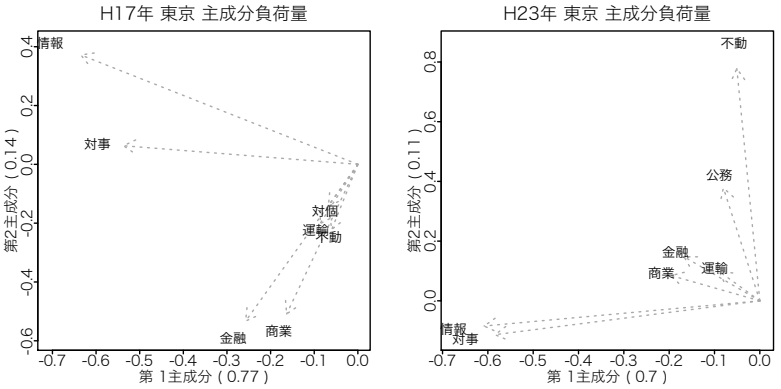
図 5.3b より、東京都の場合、2005 年では「情報通信」と「金融」、「サービス業」に特徴的な傾向がみられたが、2011 年では「サービス業」の増加に加えて、さらに「不動産」も一つの大きな特徴として現れている。東京では 2008 年頃より商業地に関する地価が急上昇した「ミニバブル」[4]が起こっていたことから、その余韻として特徴が抽出されたと考えられる [71,72]。

図 5.3c より、石川県の場合、基盤産業の「機械」に加えて,「電子部品」や「電力ガス」の長さが増している。石川県ではリーマンショックの影響から全体的に経済活動が縮退しており、2005 年と 2011 年の県内総生産を比較すると、4.73 兆円から 4.40 兆円へとおよそ 7 %ほど景気が落ち込んでいた [73]。特に 2005 年において突出していた「商業」や「情報通信」などはその影響を受けていると考えられる。

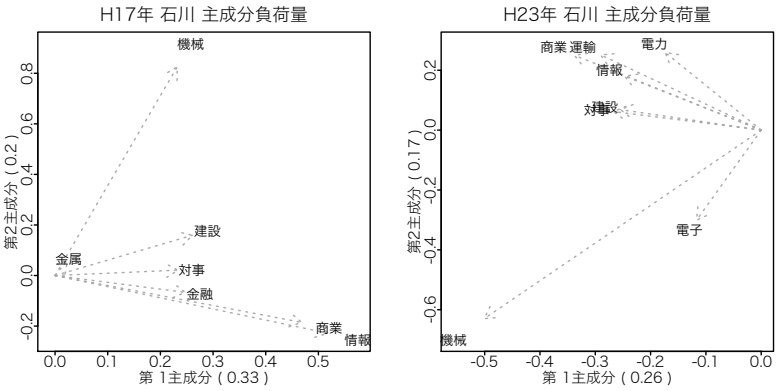
[4] ミニバブル期には、2008 年 (平成 20 年) においてバブル経済期に記録した商業地地価額の最高値を更新するほど、地価が高騰した。その要因として、景気回復が続く中での企業のオフィス需要の増大や不動産投資の拡大等を背景に、利便性・収益性が向上したためであるとされている。[71]2010 年には地価額が大きく減少したが、2013 年以降にさらに「ミニバブル期」を上回る土地価格の上昇が起こっている [72]。



(a) 福島県:第 1・第 2 主成分負荷量 (左:2005 年 右:2011 年)



(b) 東京都:第 1・第 2 主成分負荷量 (左:2005 年 右:2011 年)



(c) 石川県:第 1・第 2 主成分負荷量 (左:2005 年 右:2011 年)

図 5.3: 福島県・東京都・石川県における産業構造の経年変化 (固有ベクトル)

第6章　スパース主成分分析に関する研究

6.1　COVID-19流行前後における生活行動の変化を捉えるためのスパース主成分分析

本節 [74] では、COVID-19 流行前後における石川県内に滞在する人々の生活行動変化を分析する目的で、携帯電話の位置情報データにスパース主成分分析を適用し、特徴抽出を試みている。「3.2 節 COVID-19 流行下の石川県内滞在者移動行動変化に関する研究」とは使用データ及び構造は同じであるが、分析手法が異なっている。特に非負制約の有無、スパース制約の有無などの観点で異なった分析となり、その結果においても異なる特徴が抽出された。具体的には寄与率に基づき、縮約次元数を 2 と設定した場合、「就業・就学・観光」や「買い物・消費等」が抽出された。いずれも COVID-19 の影響を大きく受け、観光業や飲食業等では 2021 年 2 月時点において COVID-19 流行以前ほど人流が回復していないことが解析された。

6.2　使用データと設定について

6.2.1　使用データについて

石川県における COVID-19 流行前後の生活行動の変化を捉えるため、COVID-19 流行前の 2019 年 1 月 1 日から石川県第 3 波期中である 2021 年 3 月 8 日までの 798 日分の、それぞれ 4 時、14 時、19 時の 3 時点において、石川中央都市圏に属する 4 市 2 町 (金沢市・白山市・野々市市・かほく市・内灘町・津幡町) を対象とした 500m×500m メッシュ内人口 (1053 メッシュ) データを使用した。すなわち、2394 時系列 (798 日×3 時点)×1053 メッシュデータの行列を作成し、本節の使用データとした。

6.2.2　分析手法の設定

本節では Zou ら (2006) [53] のスパース主成分分析を参考に解析を行った。観測データ行列 $\boldsymbol{X} = [\boldsymbol{x_1}, \cdots, \boldsymbol{x_m}] \in \mathbb{R}^{n \times m}$、固有ベクトルからなる行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_k}] \in \mathbb{R}^{m \times k}$、$\boldsymbol{A} \fallingdotseq \boldsymbol{B}$ となる行列を $\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_k}] \in \mathbb{R}^{m \times k}$ とし、最小二乗誤差基準に基づく主成分の計算アルゴリズムを以下のように設定する、

$$(\hat{\boldsymbol{A}}, \hat{\boldsymbol{B}}) = \underset{\boldsymbol{A},\boldsymbol{B}}{\arg\min} \|\boldsymbol{X} - \boldsymbol{X}\boldsymbol{B}\boldsymbol{A}^{\boldsymbol{T}}\|_F^2 + \lambda\|\boldsymbol{B}\|_1 \qquad subject\ to\ \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} = \boldsymbol{I}, \tag{6.1}$$

ここで、$\| \ \|_F$ はフロベニウスノルム、$\| \ \|_1$ は L_1 ノルムとする。本節の使用データに合わせて、$n = 2394$、$m = 1053$ である。本稿の分析ではスパース効果の大きい L_1 正則化項のみのモデルを採用した。

L_1 正則化項におけるハイパーパラメータ λ の選択法として交差検証法 (Cross Validation method:以降 CV 法とする) を用いた (ハイパーパラメータの選択方法は 7.2 節 で後述する)。CV 法とは、過学習を抑え、汎

化性能を高めることを目的に、分析データを test データと training データに分け、training データを用いて解析を行い、test データとの誤差を評価することでモデルの妥当性を測る手法である[1]。

CV 法は多くのデータ構造の推定に用いられているが、行列分解モデルにそのまま同手法を適用するのは困難である。理由として、行列分解モデル (行列データを 2 つ以上の行列データに分解し、近似を測るモデル、例えば主成分分析、非負値行列因子分解など) では、データの 1 要素を hold-out した状態で行列分解することは不可能である。列または行を hold-out する場合、分解された行列は hold-out した列や行の分だけランク落ちしているため、hold-out した箇所を復元することは不可能である。

これらを背景に、行列分解モデルに対して CV 法を用いて縮約する次元数を決定する方法に関して、Owen and Perry(2009) [75] は test データをブロック型で hold-out する方法と評価方法について提案している (図 7.1、図 7.2 参照)。また、Williams et al.(2020) [76] はデータを「まだら」に hold-out する方法を提案し、その箇所を欠損値と見なし、「0」とおいて CV の評価を行っている。本節では、欠損箇所の選択方法として、BIBD 法 (Balanced Incomplete Block Design : つり合い不完備ブロック計画) [77] を採用した。これは、実験計画の分野で使用されている手法であり、行列データの列または行から、一定のルールに基づいてデータが選択される手法である。例えば本節の場合、2394 時系列 ×1053 メッシュデータの行列データに対し、列に関して 120 個の固定数だけデータをランダムに選択し、それを 1053 メッシュ分行うことで、120×1053 分 (データの 5%分) が選択対象となった。

他方、本節のようなデータ選択を行い、これらを test データとした場合、残りの training データで test データ箇所を復元するのは困難である。そこで使用データが非負値であることから、Williams et al.(2020) を参考に対象箇所を「0」とする処理を行い、training データとした。この training データに対し、ハイパーパラメータごとにスパース主成分分析モデルを 50 回ほど適用し、test データ箇所と training データにおける BIBD 法選択箇所について誤差を測定することで、CV 値を測定した。これらの手順を図 6.1 として示した。また CV 値は以下のように計算した、

$$CV\ 値 = \frac{1}{n}\sum (\boldsymbol{X}_{test,i,j} - \boldsymbol{X}\hat{\boldsymbol{B}}\hat{\boldsymbol{A}}^{T}_{train,i,j})^2, \tag{6.2}$$

ここで i, j は BIBD 法による選択箇所、n は試行を繰り返した回数を表す。

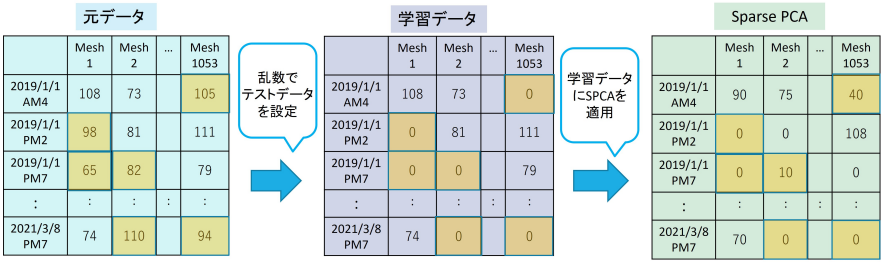


図 6.1: スパース PCACV
test データ:「元データ」の黄色枠　training データ:「Sparse PCA(スパース主成分分析)」の黄色枠

[1]CV 法には複数の種類があり、データをある適当なサイズに分割し、一方を test データ、残りを training データとして、モデルを評価する「hold-out 法」や、データを k 分割し、そのうち 1 つを test データ、残り $k-1$ を training データとして解析し、それを k 回繰り返した平均値を評価値とする「k-分割交差検証法」、1 つのデータのみを test データ、残りを training データとし、それをデータ数分繰り返す「Leave one out 法」などがある。

表 6.1 はハイパーパラメータごとに上記の条件で CV 値を測定した結果と、スパース主成分分析適用時のスパース効果、及び寄与率を表している。ハイパーパラメータは $\lambda = 10^{-5}$から 0 の範囲で変動させている。特に $\lambda = 0$ の時は、正則化項が 0 となるため、通常の主成分分析と一致する。

CV 値を計算した結果、ハイパーパラメータ $\lambda = 10^{-3}$ となる点が CV 値が最も低くなった、すなわち test データに対する誤差が最も小さくなった。ハイパーパラメータの値が大きいほど、スパース主成分分析の結果に 0 が多くなり、非負値で与えられている test データとの誤差が大きくなる。他方、ハイパーパラメータの値が小さいほど、直交性が強くなり、スパース主成分分析の結果が正負でばらつくことになる。(6.2) から、training データが負となる場合、CV 値が大きくなるため、適度に 0 となる方が CV 値を下げる可能性がある。これらのバランスから、$\lambda = 10^{-3}$ が選択された。そのため本節では $\lambda = 10^{-3}$ を最適なハイパーパラメータ値として分析を行った。

また、スパース主成分分析のスパース効果として、固有ベクトルの行列 $\boldsymbol{B}$ の全要素 ($1053 \times 1053 = 1,108,809$) 数に占める 0 の数及びその割合 (スパース率) を測定している。ハイパーパラメータが大きいほど、L_1 正則化の罰則が強く、多くの値を 0 にする効果がある。例えば、$\lambda = 10^{-1}$ の時、固有ベクトルの行列 $\boldsymbol{B}$ の全要素中 2 つのみが値を有しており、残りはすべて 0 となった。他方ハイパーパラメータが小さいほど、L_1 正則化の罰則が弱く、あまり要素を 0 にする効果が見られない。例えば、$\lambda = 10^{-5}$ の時、スパースの効果として 0 となった要素数は固有ベクトルの行列 $\boldsymbol{B}$ 全体の 3.33 %程度であった。本節で解析する $\lambda = 10^{-3}$ の時は 0 となった要素数が固有ベクトルの行列 $\boldsymbol{B}$ 全体の 99.4 %であった。

表 6.1 中段以降に関して、スパース主成分分析を適用した結果の第 1 寄与率から第 3 寄与率までとその累積値を示している。ここで第 i 寄与率は固有値の総和に占める第 i 固有値の比率としている。固有値は (2.3.2) におけるスパース主成分分析の「STEP3. $\boldsymbol{X^T X B} = \boldsymbol{U \Sigma V^T}$」の Σ に対応している。ハイパーパラメータ λ の値が大きい場合、$\boldsymbol{X^T X B}$ の多くの値が 0、または列、行が 0 になることによるランク落ちが発生しており、そうしたデータを固有値分解した結果得られる固有値もその影響を受け、第 1 主成分や第 2 主成分に値が集約される。例えば $\lambda = 10^{-1}$ の場合、固有ベクトルの行列 $\boldsymbol{B}$ は非 0 値が 2 つしかないため、$\boldsymbol{X^T X B}$ で表される行列のほとんどがランク落ちしており、それを特異値分解した結果算出される固有値は第 1 固有値のみとなり、寄与率 100 %となる。他方、ハイパーパラメータの値が小さい場合、特に $\lambda = 0$ の場合は通常の主成分分析と同様に直交性を有しているため、$\boldsymbol{X^T X B}$ 行列及び固有値分解後の固有値はフルランクであるため、第 1 主成分から第 m 主成分まで固有値を有している。本節で解析する $\lambda = 10^{-3}$ の時、第 1 主成分で 81.1 %、第 2 主成分で 11.5 %であり、合計で 90 %を越える結果となった。そのため本節では第 2 主成分までで元データの特徴を十分に捉えていると考え、縮約する次元数 $k = 2$ とした。

表 6.1: スパース主成分分析のハイパーパラメータごとの CV 値、スパース率、寄与率

λ	10^{-1}	10^{-2}	10^{-3}	10^{-4}	10^{-5}	0
CV	20235.81	31941.1	7006.781	14795.87	20549.0	21420.4
固有ベクトル要素「0」の個数 (スパース率)	1108807 (99.99%)	1108791 (99.99%)	1102231 (99.4%)	344664 (31.08%)	37004 (3.33%)	0 (0%)
第1寄与率	1	0.894	0.811	0.733	0.717	0.715
第2寄与率	0	0.092	0.115	0.108	0.106	0.106
第3寄与率	0	0.008	0.020	0.023	0.023	0.023
累積寄与率(1~3)	1	0.994	0.946	0.863	0.846	0.844

6.3 分析結果

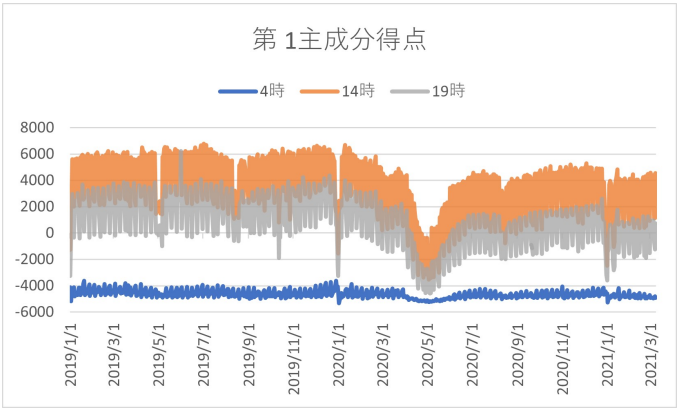
2394 時系列 ×1053 メッシュデータに対し、スパース主成分分析によって解析された 2 つの特徴について結果を示す。各主成分ごとに共通して、3 つの図があり、それぞれ a:時間別時系列パターン (主成分得点)、b:曜日別時系列パターン (主成分得点)、c:地域パターン (固有ベクトル) に対応する石川中央都市圏メッシュとしている。また地域パターンについて正値はグラデーションが濃い色ほど、時系列データの特徴を大きく捉えた地域であり、負値は一律で水色としている。また黄色は値が 0 の地域であり、スパース効果が表れている地域である。その他の設定は 3.2.3 項と同様である。

6.3.1 第 1 主成分「平日の行動 (就業・就学+観光)」

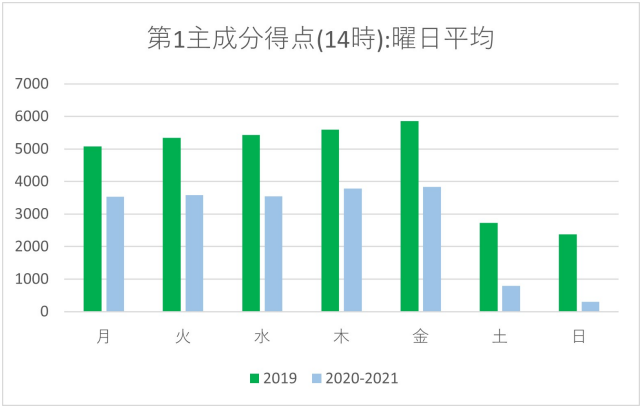
図 6.2a より、第 1 主成分得点の時系列パターンを見ると、「14 時」における滞留人口が最も高く、次いで「19 時」が高い。「4 時」の滞留人口は「14 時」や「19 時」とは反対に、マイナスの特徴として表れている。また通年を総じて、ゴールデンウィークや盆期間、正月等の長期休暇中に滞留人口が大きく減少している特徴がある。COVID-19 第 1 波が流行し始めた 2020 年 4 月頃において、「14 時」及び「19 時」の滞留人口が前年にはない減少傾向を示したが、緊急事態宣言解除頃より徐々に滞留人口が増加し、2020 年 6 月には、緊急事態宣言前の水準までに回復している。しかし、特に「14 時」において、2019 年 5 月から 7 月、2019 年 10 月から 12 月にかけて 6,000 人付近で推移していた滞留人口が、COVID-19 第 1 波以降の 2020 年では 6,000 人を一度も越えていない。

図 6.2b を見ると、平日のパターンが土日よりも高い。また、2019 年の COVID-19 流行前の方が 2020 年以降の流行後の滞留人口よりいずれの曜日においても高い。

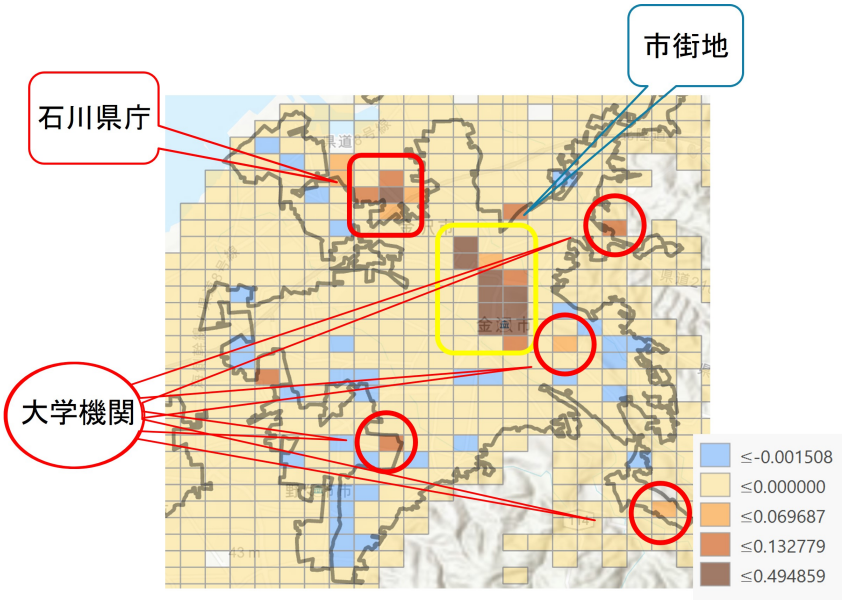
図 6.2c を見ると、金沢市の有名な観光地である金沢駅から香林坊・兼六園付近、石川県庁等の就業地、金沢工業大学・金沢星稜大学・金沢学院大学等の就学地が正の特徴として抽出されている。他方、負の特徴では居住が整った野々市市や、金沢大学最寄りの住宅地である、もりの里地区などが抽出されている。これらの結果から第 1 主成分の特徴は「平日の外出行動 (就業・就学+観光)」の特徴であると考察される。同データに非負値行列因子分解を適用した結果 (3.2 節 第 2 基底) と比較しても、概ね類似した特徴が抽出されている。



(a) 時間別時系列パターン:第 1 主成分



(b) 曜日別時系列パターン:第 1 主成分



(c) 地域パターンに対応する石川中央都市圏メッシュ:第 1 主成分

図 6.2: スパース主成分分析 第 1 主成分の主成分得点と固有ベクトル

図 6.3 は同データに対するスパース主成分分析 (左図) と通常の主成分分析 (右図) で第 1 主成分の特徴を比較したものである。なお色分けのスケールは統一している。この時、スパース主成分分析によって得られた地域パターンは局所的に正値及び負値の特徴を持ち、多くの地域が黄色、すなわち値が 0 となっている。一方、通常の主成分分析は各地で色の濃い値及び負値である水色が表れている。これらからスパース主成分分析の方が明瞭な特徴の解析が可能であると言える[2]。

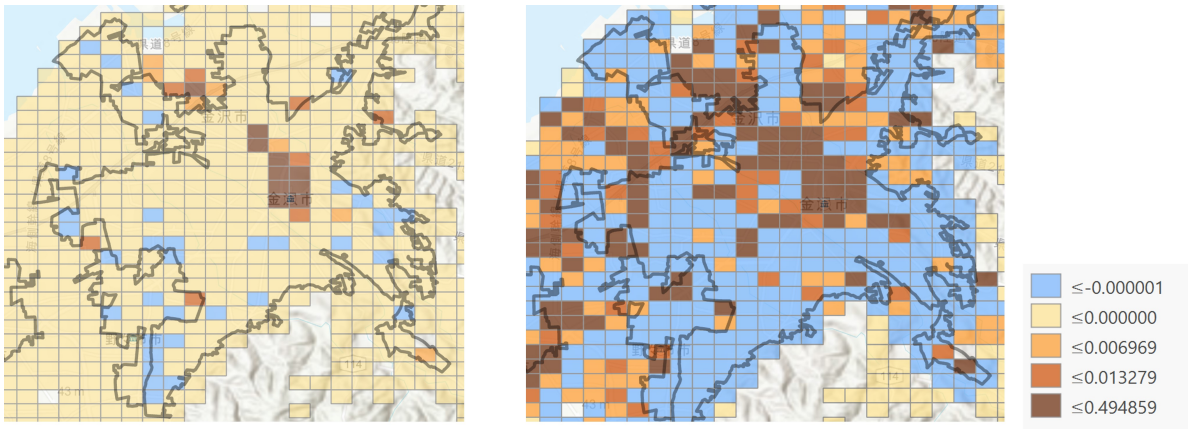
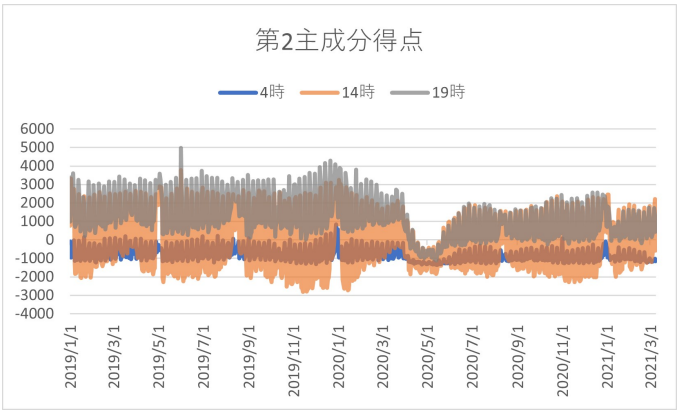


図 6.3: 第 1 主成分におけるスパース効果の可視化
左図:スパース主成分分析 ($\lambda = 10^{-3}$)　右図 L 通常の主成分分析 ($\lambda = 0$)

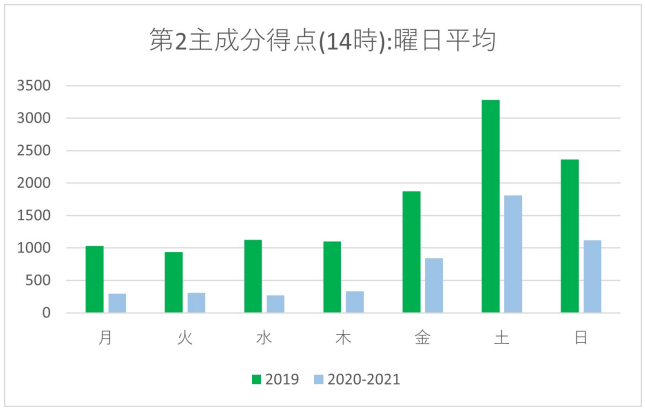
6.3.2　第 2 主成分「休日の行動 (買い物・飲食等)」

図 6.4a より、「19 時」における滞留人口が最も高く、次いで「14 時」が高い。ただし「14 時」は正にも負にも大きく振れている。「4 時」の滞留人口は負値で推移している。また、COVID-19 第 1 波の予兆が見え始めた 2020 年 3 月下旬頃において、「14 時」及び「19 時」の滞留人口が大幅に減少したが、緊急事態宣言解除頃より徐々に滞留人口が増加し、2020 年 5 月末には、緊急事態宣言前の水準までに回復している。しかし、「19 時」において、2019 年の水準には未だ回復していない。図 6.4b を見ると、平日のパターンより土日が高く、平日の中でもやや金曜日が高い。また、2019 年の COVID-19 流行前の方が 2020 年以降の流行後の滞留人口よりいずれの曜日においても高い。図 6.4c を見ると、香林坊・片町や金沢駅・昭和町、買い物・飲食施設と映画館等の娯楽施設を備えた高柳地区、石川県内 4 位の敷地面積を持つ大型ショッピングセンターがある御経塚地区、同じく大型ショッピング機能を持つコストコがある野々市市などの特徴が抽出された。一方で負値として、第 1 主成分で正の特徴として抽出された、兼六園周辺の観光地、石川県庁等の就業地、大学機関等は負の特徴として抽出されている。これらの結果から第 2 主成分は「休日の行動 (買い物・飲食等) の基底と考察される。地域の特徴と時系列の特徴を併せると、大型ショッピングセンター等への移動は 2020 年の COVID-19 第 1 波流行期には特に控えられていた。すなわち COVID-19 流行期における石川県内滞留者の買い物事情は、近場のスーパーやコンビニ、ドラッグストア等、またはインターネット経由での購買活動が優先されていたと考えられる。また、香林坊・片町等の石川県の繁華街をはじめとする飲食業界は、2021 年 2 月末時点において「14 時」、「19 時」の昼、夜ともに人流が 2019 年ほどには回復しておらず、COVID-19 のこれら業界に対する影響が続いていることがわかる。

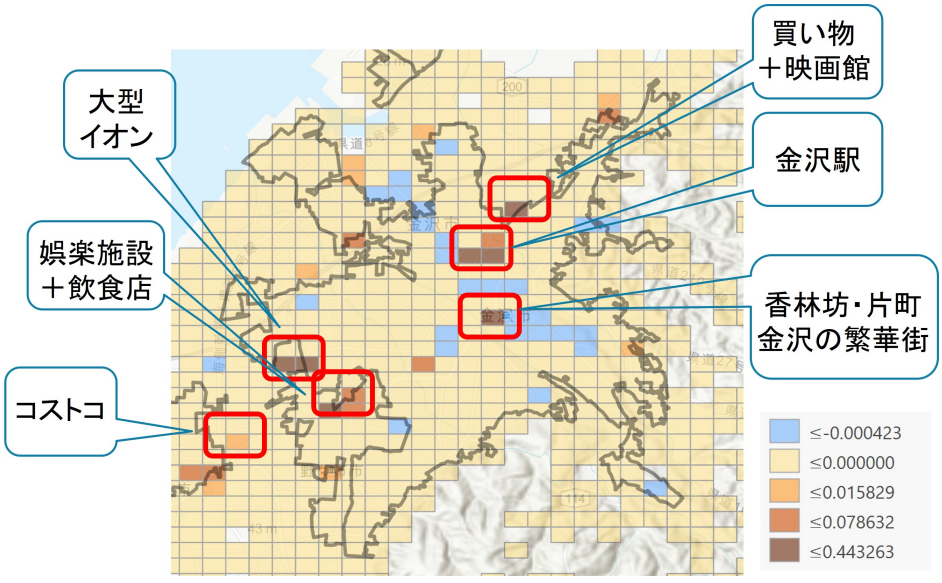
[2]通常の主成分分析の図中に存在する黄色のメッシュはモバイル空間統計において秘匿性の観点から欠損値とされている箇所である。通常の主成分分析では図 6.1 にあるように、固有ベクトルの行列に 0 はもたない。



(a) 時間別時系列パターン:第 2 主成分



(b) 曜日別時系列パターン:第 2 主成分



(c) 地域パターンに対応する石川中央都市圏メッシュ:第 2 主成分

図 6.4: スパース主成分分析 第 2 主成分の主成分得点と固有ベクトル

第7章　おわりに

7.1　結論

本稿では主成分分析、非負値行列因子分解等の次元縮約法に着目し、理論面の整備と実例研究に関する考察を行った。また、これらの次元縮約法に対し、抽出される特徴の要素を 0 にし、解釈を容易にできる L_1 正則化項を付与したモデルについても言及した。

第 3 章では非負値行列因子分解を非負値で定義される産業連関表及び携帯電話の位置情報データに適用した結果について考察した。3.1 節の産業連関表に対する分析では愛知県を中心とした「製造圏」、東京都などの大都市を中心とした「第 3 次産業」、神奈川県や千葉県などの製油所機能を有する「鉱業」の特徴が抽出でき、日本の大規模産業について客観的に解析できた。本分析の発展として、産業間の取引に着目した内生部門以外へのデータにも本分析手法は適用できる。例えば地域経済分析を行う上で重要な指標である、需要増加に伴う地域経済全体への波及効果を示す逆行列係数表に本分析手法を適用した場合、他産業の生産に正の総乗効果をもたらす産業及び産業クラスターの特定が可能となる。他にも、行方向 (供給) で内生部門取引を除した「産出係数行列」やその波及効果を表す「ゴーシュ逆行列」[78]、産業間の距離を表す「APL(Average Propagation Length:平均波及世代数)」[79] など、分析目的に応じて様々な観点から解析可能である。3.2 節の携帯電話の位置情報データでは朝、昼、夜の 3 時点において石川県内滞留人口が密集 (または居住) する地域について解析できた。人的流動状態を分析することでまちづくりや観光等に関する政策にも活用できるであろう。第 4 章では非負値行列因子分解にスパース制約を付与した L_1 正則化項の有効性について考察した。4.1 節の都道府県間を跨いだ移動に関する解析では、選ばれた基底から COVID-19 流行前の特徴を抽出でき、他方で COVID-19 流行期の特徴の多くが 0 とできたことから、COVID-19 流行期の特徴が通常期と比較して極めて特異なものであることが確認できた。さらにデータとモデルとの残差を分析すると、COVID-19 の特徴が大きく抽出されたことから、モデルと残差の両面から解析することの有効性が示された。また、4.2 節の石川県羽咋市を対象とした昼間流出率を用いた解析では、通常の非負値行列因子分解で抽出された、やや他の属性と行動パターンが異なる「15 歳-19 歳男女」の特徴 (学校及び長期休み等で行動が周期的な属性) が、スパース制約により他の属性とほぼ完全に分離できた。第 5 章では、47 都道府県産業連関表データを用いて産業構造の類似度及び経年変化について可視化を行った。産業構造の類似度に関する解析では全国産業連関表の固有空間上に 47 都道府県の産業連関表データを射影することで、同一空間上において地域比較が行う工夫を行った。また産業構造の経年変化に関する研究では福島県、東京都、石川県の産業構造がいずれも外部要因の影響によって基幹産業の特徴が 2005 年から 2011 年の間で変化していたことが視覚的に解析できた。第 6 章では、携帯電話の位置情報で表されるメッシュ人口情報をスパース化し、解析の簡素化を試みた。地図情報にメッシュデータを投影した分析では通常の主成分分析よりもスパース主成分分析の方が、スパース効果により解釈が明確になっており、「就業・就学・観光」や「買い物・飲食等」の特徴が抽出された。

本稿において研究目的としていた、次元縮約法によってデータが持つ共通な構造を客観的に抽出すること、また正則化モデルによって特徴をより明瞭に抽出することに関して、応用研究を踏まえてそれらの成果が実証された。次元縮約法に関して例えば、3.1 節における産業連関表への非負値行列因子分解法の適用によって、製油所機能を持つ共通の地域が抽出され、特に神奈川県や千葉県など、日本最大級の製油所を有する地域の特徴が突出して抽出できたことを示した。また、正則化モデルに関して、第 6 章における携帯電話の位置情報データへのスパース主成分分析法の適用によって、石川県内滞留者の COVID-19 流行前後に

おいて大きく変化した行動パターンが抽出できたとともに、その行動パターンが顕著に見られる地域の特徴が、L_1 正則化項を付与しない通常の主成分分析と比較して明瞭に抽出されることを示した。

本稿で使用した分析手法と分析結果の違いについて表 7.1 にまとめる。通常の主成分分析は、各主成分が直交するように、かつデータを最も効率よく縮約できるように分散を最大化しながら分解する手法であり、その直交性から解の一意性が得られる。一方で、非負性はなく、また主成分の要素に完全な 0 を取る場合が極めて少ないため、解析時には正と負の符号を考慮した解釈が求められ、かつ非 0 である微小な値をとることがあるため、その解析は複雑なものとなることがある。また L_1 正則化項を付与した場合、各主成分は完全には直交せず、解の一意性は保証されないが、主成分の要素を 0 にする効果から、通常の主成分分析よりも明瞭な特徴抽出及び解釈が可能となる。ただし、依然として正と負を考慮した上での解釈の困難さがある。非負値行列因子分解は、2 つの非負値の行列に分解できることから、分析結果を「積み上げ式」に解釈できる。そのため、主成分分析のように正と負で特徴が相殺されることがない。また、データを非負値の空間に射影する性質から、一部の値を 0 にする効果もある。一方、直交性が制約条件に含まれていないため、各基底で相関性のある、類似した特徴が現れる場合がある。また、解の一意性が保証されず、解は計算アルゴリズムの初期値に大きく依存している。L_1 正則化項を付与した場合、通常の非負値行列因子分解と比べて、基底ベクトル内の値を 0 にする効果から、通常の非負値行列因子分解よりも明瞭な特徴抽出及び解釈が可能となる。上記にて、本稿で使用した次元縮約法及び正則化モデルに関して議論したが、これらの使い分けは解析者に委ねられる。例えばビックデータの構造をなるべく少ない次元数で解析したい場合は主成分分析が有効であろう。他方、非負値の分解でデータを解析したい場合、またはメッシュ内滞留者データや画像データ等の正値のみでの解釈が望まれるデータに対しては非負値行列因子分解が有効であろう。正則化モデルは次元縮約法によって得られる特徴をより明瞭に際立たせたい場合に有効であろう。

表 7.1: 分析方法と分析結果の違い

	主成分分析		非負値行列因子分解	
L1正則化項	無	有	無	有
非負性	×	×	○	○
直交性	○	△	×	△
0の数	×	○	△	○
解の一意性	○	○	×	×
解釈のしやすさ	×	△	△	○

○はその特性を持つ、×はその特性を持たない、△はそれらの中間を意味する

本研究で議論した分析手法は、正値の行列構造を持つデータ全般に適用可能である。POS データを用いた場合は、「どの属性 (年齢・性別等) がどのような商品をよく購入する傾向がある」という特徴が解析でき、携帯電話の位置情報データを用いた場合は「どの属性がどの地点に多く滞留する傾向があるか」という特徴が解析できる。他にも医療レセプトデータを用いた場合は、「どの属性がどの病気を多く罹患している傾向があるか」などが解析できるであろう。さらにテンソル因子分解 (7.2 節) を使用する場合、上記の例のいずれにも変数を 1 つ追加して解析ができる。POS データや携帯電話の位置情報データの場合、「時系列」別に購買、滞留情報を解析でき、医療レセプトの場合は「初診」の情報を加えて、「どの属性がどの病気を何歳頃から多く罹患し始めている傾向がある」というような解析が可能である。また、使用データの構成として、本稿の応用研究ではいずれも正値かつ金額・人ベースのデータを使用したが、分析目的によって 0 から 1 の間に正規化されたデータや、0 と 1 の 2 値に変換したデータ等でも分析可能である。負値を含むデータを使用する場合、主成分分析を適用することで解析可能である。

本分析手法で得られた分析結果に対し、さらに新たな知見を得るための方法として、縮約する次元数を多

くすることや、データとモデルとの残差に関してさらに解析を行うことなどが挙げられる。縮約する次元数に関して、主成分分析であれば解析する主成分数を増やすことで、非負値行列因子分解であれば基底数を増やすことでより詳細に解析できる。その場合、産業連関表であれば地方小都市または小規模産業の特徴を、携帯電話の位置情報データであれば中・小規模地域における人流の特徴を解析できるであろう。また、残差について、特に非負値行列因子分解では最小二乗法を基にしたアルゴリズムであるため、モデル内では大都市や大規模産業等の大きな特徴を持った地域や産業が優先して抽出される。一方、モデル内で取り切れなかった特徴は残差として表れるため、これに対してさらに解析を行うことで、中・小地域や産業の特徴を抽出できる可能性がある。

本稿で使用した次元縮約法及び正則化項を付与した次元縮約法は、いずれも統計解析プログラミング言語の R でパッケージが提供されている。具体的には、非負値行列因子分解及びスパース非負値行列因子分解は「”NMF”」(“NMF”を使用する際には「Biobase」も併せてインストールする必要がある)、主成分分析は「”prcomp”」(R のデフォルトにて内蔵済み)、スパース主成分分析は「”sparsepca”」などのパッケージを使用することで解析が可能である 。ここでは、縮約する次元数や正則化項のハイパーパラメータなどを解析者側で任意の値を設定し、データを解析することが可能である。ただし、主成分分析 (prcomp) 以外は、初期値によって解析結果が変化することに注意が必要である。また、データのサイズによって計算処理速度が大きく異なり、本稿 3.2 で使用した 2394×1053 のデータを使用した場合、プログラミング言語の R を用いて、Intel(R) Core(TM) i9-9980HK CPU @ 2.40GHz の PC スペックで 31 分を要した[1]。今後はより効率よくデータ解析を行うために、GPU を用いた高速な並列計算処理を行うことが望まれる。

本稿での理論展開における新規性として、縮約する次元数及び正則化項のハイパーパラメータの決定に対する一つの選択基準を検証したことである。主成分分析や非負値行列因子分解、それらに正則化項を付与したスパース主成分分析やスパース非負値行列因子分解は、いずれも計算アルゴリズムが整備されており、本稿でもそれらに基づいた分析アルゴリズムを構成している。ただし、いずれにおいても縮約する次元数や正則化項のハイパーパラメータの決定法には言及しておらず、未解決な問題とされている。そのため、次元縮約法を用いた研究例の多くは、縮約する次元数や正則化項のハイパーパラメータ等に関して、解析者の想定する任意の値を設定して解析を行っている。こうした問題に対し、本稿では BIBD 法を用いた CV 法によるハイパーパラメータの決定法について、応用研究を交えて実証した (詳細は 6 章及び 7 章を参考のこと)。特に行列因子分解モデルにおいて、縮約する次元数及び正則化項のハイパーパラメータの決定法に CV 法を用いた研究例は少なく、さらに実験計画の分野で利用される BIBD 法をこれに応用した解析は本稿が初めてとなる。縮約する次元数及び正則化項のハイパーパラメータの決定に関しては、今後の展開にて議論する。

また応用研究の発展として、行列に対する次元縮約法を拡張した、テンソル因子分解に関する研究を行っている。テンソルデータに対する解析法の理論的な整備を今後の課題に記し、産業連関表を用いた予備的な分析結果を付録 2 に記する。

7.2 今後の発展

7.2.1 基底数及びハイパーパラメータの設定

次元縮約法に Lasso 等の正則化項を付与するモデルでは、縮約する次元数 (または基底数) やハイパーパラメータの選択方法が問題となる。主成分分析は目安として寄与率を使用する方法があるが、直交制約のない非負値行列因子分解ではこうした目安が適用できず、主成分分析以上に縮約する次元数の客観的な決定が未解決な問題である。そこで以降では非負値行列因子分解における基底数の選択問題について、近年の先行研究と選択方法に関する方針から 2 つに分けて議論する。

[1] R は ver.4.0.2 を使用した。

(1) 情報量基準の適用

客観的に縮約する次元数を決定する方法として、AIC(Akaike's Information Criterion) [80] や BIC(Bayesian information criterion) [81] 等の情報量基準を用いた研究例がある。これらは統計モデルの良さを評価する指標であり、AIC 及び BIC は以下のように表せる、

$$\begin{aligned} AIC &= -2lnL + 2K, \\ BIC &= -2lnL + K \cdot ln(n), \end{aligned} \tag{7.1}$$

ここで、L は尤度関数、K はパラメータ数、n はデータ数を表す。実際に Bai and Ng(2002) [82] によって BIC の正則化項 ((7.1) 右辺の第 2 項) が改良されたモデルを、非負値行列因子分解の基底数決定における評価基準に採用している例 [83] がある。

AIC や BIC 等の情報量基準は尤度関数を評価指標としていることから、非負値行列因子分解モデルに尤度関数の仮定を定義できた場合に有効に作用すると考えれられる。亀岡 (2015) [29] は、非負値行列因子分解法における 3 種類の乖離度である、二乗誤差基準、I ダイバージェンス基準、板倉・齊藤擬距離がそれぞれ正規分布、ポアソン分布、指数分布に従って独立に生成されたと仮定した場合の行列 $\boldsymbol{H}, \boldsymbol{U}$ の最尤推定問題と等価であるとしている[2]。特に Kullback–Leibler ダイバージェンスに由来する I ダイバージェンス基準では、Cemgil(2008) [84] や Schmidt et al.(2009) [85] によって、非負値行列因子分解における基底行列 $\boldsymbol{H}$ 及び表現行列 $\boldsymbol{U}$ について周辺尤度に基づくモデル選択の問題として定式化されている。具体的にはベイズモデリングを用いて、$\boldsymbol{H}, \boldsymbol{U}$ にガンマ分布、これら $\boldsymbol{H}, \boldsymbol{U}$ をパラメータとする行列 $\boldsymbol{S}$ を設定して、以下のように定式化している、

$$p(\boldsymbol{X}|\boldsymbol{\Theta}) = \int dHdU \sum_{s} p(\boldsymbol{X}|\boldsymbol{S}) p(\boldsymbol{S}|\boldsymbol{H}, \boldsymbol{U}) p(\boldsymbol{H}, \boldsymbol{U}|\boldsymbol{\Theta}), \tag{7.2}$$

ただし、Θ はガンマ分布のハイパーパラメータである。Cemgil(2008) [84] では、パラメータ $\boldsymbol{H}, \boldsymbol{U}, \boldsymbol{S}$ の推定方法として、変分ベイズ法及び MCMC 法を用いている。

上記のように、非負値行列因子分解をベイズモデリングの仮定の下で定義できる場合、情報量基準だけでなく、ベイズ統計学の分野で理論的に整備されているモデルを適用できる可能性がある。例えば、田邊・寒河江 (2000) [86] ではベイズモデリングの仮定の下で、線形回帰モデルにおける分散と罰則付き線形回帰モデルの分散、及び事前分布とした罰則項の分散との関係性から、ハイパーパラメータの決定後のモデルと罰則項のトレードオフを示す尺度として、Relative Credibility Criteria(RCC) を提案した。また、ベイズモデリングの性質は応用研究における解釈面においても活用できる。例えば非負値行列因子分解を用いて各種基底において、小規模地域及び産業等の解析を行いたい場合、事前分布にこれら地域の特徴を抽出する構造を組み込む、または大規模地域及び産業が 1 つの基底にまとまり、他の基底で小規模地域及び産業の特徴を抽出する構造を組み込むなどの工夫によって、客観的にかつ多様な分野での解析が可能となるであろう。

(2) Cross Validation(CV) 法の適用

第 6 章ではスパース主成分分析におけるハイパーパラメータの選択として、CV 法を適用した。この手法はモデルに依存しない簡便な計算法であり、統計学及び機械学習における過学習抑制のためのハイパーパラメータ (Lasso, Ridge 等) の調整等で使用されている。主成分分析や非負値行列因子分解などの行列分解モデルに対し CV 法を用いて縮約する次元数を決定する方法に関して、Owen and Perry(2009) [75] は test デー

[2] しかし、例えば二乗誤差基準の場合、正規分布の仮定 ($x_{i,j} \sim N(x_{i,j} : hu_{i,j}, \sigma^2)$) とするため、$hu_{i,j} = 0$ となるような場合、等分散 σ^2 によって負値の推定量 $\hat{x_{i,j}}$ が算出される可能性がある。

タを行列型で hold-out する方法とその時の評価方法について提案している (図 7.1、図 7.2 参照)。図 7.1 では N×M で表される全体の行列に対し、$A \in \mathbb{R}^{n \times m}$ に対応する箇所について hold-out し、その他 $B \in \mathbb{R}^{n \times M-m}$、$C \in \mathbb{R}^{N-n \times m}$、$D \in \mathbb{R}^{N-n \times M-m}$ を用いて $\hat{A}$ を復元し、CV の評価を行っている。具体的には以下のように表される、

$$\begin{aligned} \hat{A} &= B(H_D U_D)^{\dagger} C, \\ \hat{A} &= B H_D^{\dagger} U_D^{\dagger} C, \\ CV_A &= \|A - \hat{A}\|_F^2, \end{aligned} \tag{7.3}$$

ここで H_D, U_D は行列 D に対する行列分解によって得られる 2 つの行列 $H \in \mathbb{R}^{N-n \times k}$、$U \in \mathbb{R}^{k \times M-m}$ であり、Owen and Perry(2009) は非負値行列因子分解による行列分解を行っている。また、「$\dagger$」はムーア-ペンローズの擬似逆行列 (pseudo-inverse matrix、一般化逆行列:generalized inverse ともいう) であり、柳井・竹内 (1983) [87] ではこの逆行列を以下のように定義している。

定義 1. $\boldsymbol{A}$ を (n, m) 型逆行列とする線型方程式 $\boldsymbol{Ax} = \boldsymbol{y}$ が解 $\boldsymbol{x}$ をもつような $\boldsymbol{y}$ に対して、$\boldsymbol{x} = \boldsymbol{A^{-1}y}$ がこの方程式の一つの解となる場合、(m, n) 型行列 $\boldsymbol{A^{-1}}$ を $\boldsymbol{A}$ の一般化逆行列という。

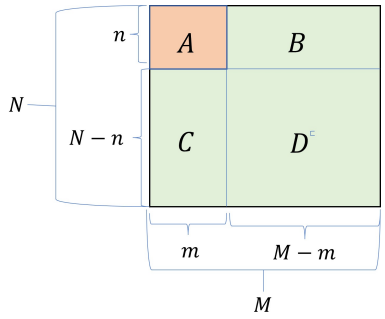


図 7.1: Bi-cross validation 法による hold-out 法 Owen and Perry(2009)

通常の逆行列は $\boldsymbol{A}$ が正則 (n, n) 型 であることと同値条件であるが、一般化逆行列では正則でない場合でも定義される。すなわち、(7.3) において、$H_D^{\dagger} \in \mathbb{R}^{k \times N-n}$、$U_D^{\dagger} \in \mathbb{R}^{M-m \times k}$ であり行列の積の条件を満たす。また、図 7.1 において 2×2 のブロック構造に分解したことに対し、図 7.2 は 3×3 に分解した場合の例を表している。Owen and Perry(2009) [75] は行列をブロック単位での置換に応用し、$A \in \mathbb{R}^{n \times m}$ に対応する箇所について hold-out した場合、図 7.2 右のように行列 B、C、D を構築することで (7.3) 式と同様の計算が可能であることを示した。Kanagal and Sindhwani(2010) [88] は重み行列を二乗誤差項に導入した Weighted NMF モデルに Owen and Perry(2009) の CV 法を適用している。

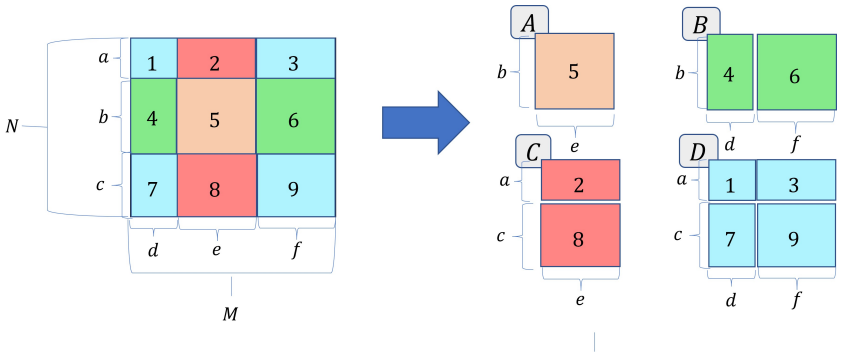


図 7.2: Bi-cross validation 法による hold-out 法 3 × 3 の場合

Owen and Perry(2009) ではブロック構造で hold-out し、CV を評価しているが、データ行列からランダムに hold-out する箇所を選択した場合、行列分解モデルの近似では該当箇所の復元は困難である。Williams et al.(2020) [76] はデータを「まだら」に hold-out する方法について提案し、その箇所を欠損値と見なし、0 とおいた処理を行っている。本稿の第 6 章で使用した CV 法は Williams(2020) のアイデアを参考に、BIBD (balanced incomplete block design : つり合い不完備ブロック計画) 法 [77] を採用し、それに基づいた hold-out を行っている。また該当箇所は 0 とする処理を行っている。BIBD 法は、行列データの列または行から、一定のルールに基づいてデータが選択される手法である。本稿で BIBD を採用した理由として、データ構造が一部の列や行に情報が偏っていた[3]ため、ランダムに hold-out する箇所を選択する場合、偏った情報を含む箇所が選択されるか否かで CV 値が大きく変動する可能性があり、それを平準化するためには多くの試行回数をこなす必要がある。そのため、いずれの列からもバランスよく hold-out できるモデルとして BIBD 法を用いた計算を行った。

今後の研究として、hold-out 箇所の処理の仕方及び非負値行列因子分解やそのハイパーパラメータの選択などを検討し、客観的な基底数及びハイパーパラメータの選択に基づいたデータ解析を行いたい。

7.2.2 高次元テンソルデータへの拡張

テンソル因子分解は高次元テンソルのデータ構造を持つデータに対する分解方法である。分解方法については、(1) コアテンソルを持たない場合、(2) コアテンソルを持つ場合、で大きく 2 種類に分けられる。

(1) コアテンソルを持たない場合

コアテンソルを持たないテンソル因子分解法は CP 分解[4] [92] を用いて、n 個の行列データに分解される。以下では最も基本的な 3 次元テンソルについて述べる ([59, 89] を参考にする)。図 7.3 は 3 次元テンソルにおける CP 分解のイメージである。

[3]具体的には使用データとした石川中央都市圏のメッシュ内人口データでは比較的都市部に分類される金沢市とその近郊地域では、500m メッシュ単位で 10 倍から最大 100 倍程度の人口差がある。

[4]Canonical decomposition and Parallel factors, 前者から CANDECOMP [90], 後者から PARAFAC [91] とも呼ばれる

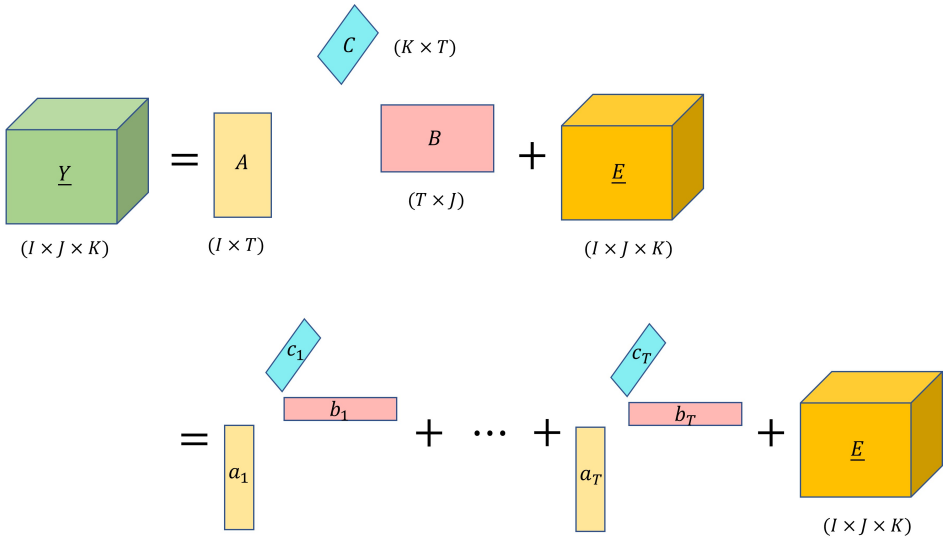


図 7.3: CP 分解のイメージ

テンソルデータ $\underline{\boldsymbol{Y}} \in \mathbb{R}^{\boldsymbol{I}\times\boldsymbol{J}\times\boldsymbol{K}}$ に対し、分解する行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_t}] \in \mathbb{R}^{I\times T}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_t}] \in \mathbb{R}^{J\times T}$、$\boldsymbol{C} = [\boldsymbol{c_1}, \cdots, \boldsymbol{c_t}] \in \mathbb{R}^{K\times T}$、誤差テンソルを $\underline{\boldsymbol{E}} \in \mathbb{R}^{\boldsymbol{I}\times\boldsymbol{J}\times\boldsymbol{K}}$ とするとき、CP 分解モデルは以下のように表される、

$$\begin{aligned} \underline{\boldsymbol{Y}} &= \sum_{t=1}^{T} \boldsymbol{a_t} \circ \boldsymbol{b_t} \circ \boldsymbol{c_t} + \underline{\boldsymbol{E}}, \\ &= [\![\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}]\!] + \underline{\boldsymbol{E}}, \end{aligned} \tag{7.4}$$

ここで、「$\circ$」は外積 (outer product) を表す。また CP 分解における外積計算を $[\![\quad]\!]$ と表す。(7.4) は要素ごとにすると以下のように表される、

$$y_{i,j,k} = \sum_{t=1}^{T} a_{i,t} b_{j,t} c_{k,t} + e_{i,j,k}. \tag{7.5}$$

CP 分解における $\boldsymbol{A}$、$\boldsymbol{B}$、$\boldsymbol{C}$ の算出は交互最小二乗法 (Alternating Least Square (ALS)) による計算アルゴリズムによって推定される。CP 分解における損失関数 D_{CP} を以下のように設定する。

$$D_{CP}(\underline{\boldsymbol{Y}} | [\![\boldsymbol{A}, \boldsymbol{B}.\boldsymbol{C}]\!]) = \|\underline{\boldsymbol{Y}} - [\![\boldsymbol{A}, \boldsymbol{B}.\boldsymbol{C}]\!]\|_F^2. \tag{7.6}$$

(7.6) の右辺より、$\boldsymbol{A}$ について最小化する。

$$\begin{aligned} D_{CP} &= \|\boldsymbol{Y_{(1)}} - \boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T\|_F^2, \\ &= tr(\boldsymbol{Y_{(1)}} \boldsymbol{X_{(1)}^T}) - 2tr(\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T) \boldsymbol{Y_{(1)}}) + tr(\boldsymbol{A}(\boldsymbol{C^T C} \circledast \boldsymbol{B^T B}) \boldsymbol{A^T}), \end{aligned} \tag{7.7}$$

ただし、$\boldsymbol{Y_{(1)}}$ はモード 1 においてテンソルデータの行列化を行った $I \times JK$ 行列である (図 7.4 はテンソルデータの行列化の例である)。また $\sum_{t=1}^{T} \boldsymbol{a_t} \circ \boldsymbol{b_t} \circ \boldsymbol{c_t}$ を Khatri-Rao Product を用いて、$\boldsymbol{A}(\boldsymbol{C} \odot \boldsymbol{B})^T$ としており、$\boldsymbol{C^T C} \circledast \boldsymbol{B^T B}$ は「付録 1」の性質を利用している。

これらの展開から、$\boldsymbol{A}$ に関して微分して 0 とすると、$\boldsymbol{A}$ の更新条件が得られる。

$$\frac{\partial \boldsymbol{L}}{\partial \boldsymbol{A}} = -2\boldsymbol{Y}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B}) + 2\boldsymbol{A}(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B}) = 0, \tag{7.8}$$

$$\boldsymbol{A} \leftarrow \boldsymbol{Y}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B})^{-1}, \tag{7.9}$$

$\boldsymbol{B}, \boldsymbol{C}$ についても同様の手順により、CP 分解に関して以下の更新式が得られる。

$$\begin{aligned} \boldsymbol{A} &\leftarrow \boldsymbol{Y}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B})^{-1}, \\ \boldsymbol{B} &\leftarrow \boldsymbol{Y}_{(2)}(\boldsymbol{C} \odot \boldsymbol{A})(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})^{-1}, \\ \boldsymbol{C} &\leftarrow \boldsymbol{Y}_{(3)}(\boldsymbol{B} \odot \boldsymbol{A})(\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B} \circledast \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A})^{-1}. \end{aligned} \tag{7.10}$$

他方、CP 分解を非負値の制約下で実行するアルゴリズムとして、非負値テンソル因子分解 (Non-negative Tensor Factorization, NTF) がある。非負値テンソル因子分解モデルの場合、CP 分解の損失関数に非負制約を付与したものと定義する。すなわち、損失関数 D_{NTF} は以下の通りである、

$$D_{NTF}(\underline{\boldsymbol{Y}} \| [\![\boldsymbol{A}, \boldsymbol{B}.\boldsymbol{C}]\!]) = \|\underline{\boldsymbol{Y}} - [\![\boldsymbol{A}, \boldsymbol{B}.\boldsymbol{C}]\!]\|_F^2 + \alpha_A\|\boldsymbol{A}\|_F^2 + \alpha_B\|\boldsymbol{B}\|_F^2 + \alpha_C\|\boldsymbol{C}\|_F^2, \tag{7.11}$$

ただし、$\alpha_A, \alpha_B, \alpha_C$ は非負値のハイパーパラメータである。また、非負値テンソル因子分解の更新アルゴリズムは CP 分解における導出と同様に ALS 法を用いた計算により、以下のアルゴリズムが得られる。

$$\begin{aligned} \boldsymbol{A} &\leftarrow \boldsymbol{Y}_{(1)}(\boldsymbol{C} \odot \boldsymbol{B})(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B} + \alpha_A\boldsymbol{1})^{-1}, \\ \boldsymbol{B} &\leftarrow \boldsymbol{Y}_{(2)}(\boldsymbol{C} \odot \boldsymbol{A})(\boldsymbol{C}^{\boldsymbol{T}}\boldsymbol{C} \circledast \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} + \alpha_B\boldsymbol{1})^{-1}, \\ \boldsymbol{C} &\leftarrow \boldsymbol{Y}_{(3)}(\boldsymbol{B} \odot \boldsymbol{A})(\boldsymbol{B}^{\boldsymbol{T}}\boldsymbol{B} \circledast \boldsymbol{A}^{\boldsymbol{T}}\boldsymbol{A} + \alpha_C\boldsymbol{1})^{-1}, \end{aligned} \tag{7.12}$$

ただし、$\boldsymbol{1}$ のサイズに関して、$\alpha_A\boldsymbol{1}$ は $\boldsymbol{A}$、$\alpha_B\boldsymbol{1}$ は $\boldsymbol{B}$、$\alpha_C\boldsymbol{1}$ は $\boldsymbol{C}$ のサイズに対応し、要素が全て 1 の行列である。

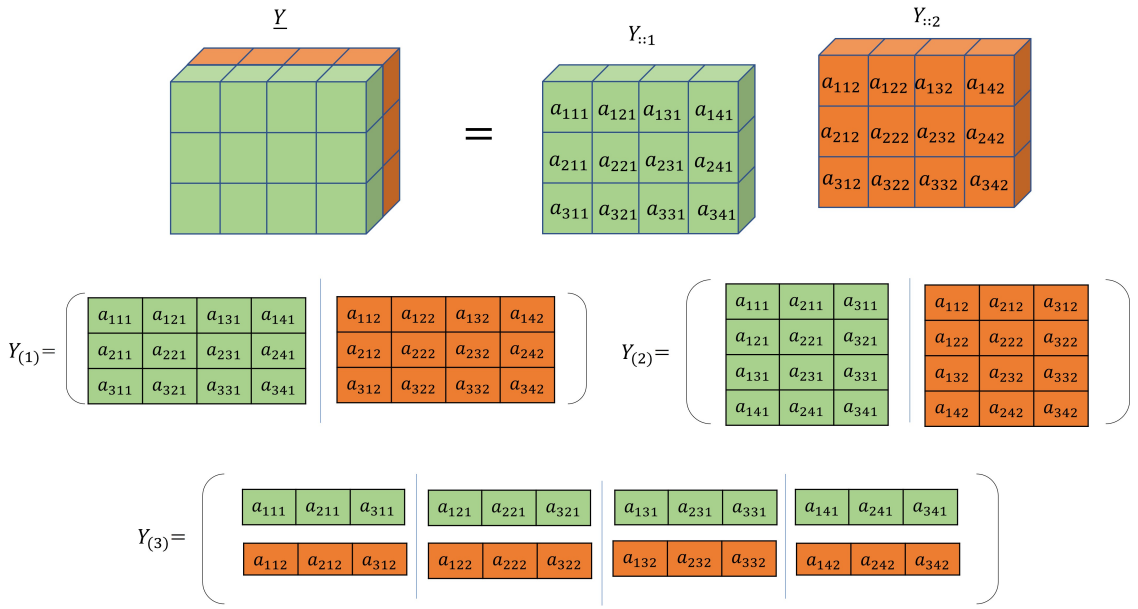


図 7.4: 3 次元テンソルの行列化の例 $\boldsymbol{Y_{(1)}} \in \mathbb{R}^{I \times JK}$ $\boldsymbol{Y_{(2)}} \in \mathbb{R}^{J \times IK}$ $\boldsymbol{Y_{(3)}} \in \mathbb{R}^{K \times IJ}$

(2) コアテンソルを有する場合

他方、コアテンソルを有したテンソル因子分解法は Tucker 分解 (Tucker Decomposition) と呼ばれ、n 次元テンソルデータを n 種類の行列と n 次元のコアテンソルに分解する (図 7.5 参照)。Tucker 分解は、分解する条件として各行列内の列ベクトルが直交する制約のもとで行われる。すなわち $\boldsymbol{A^T A = I}$、$\boldsymbol{B^T B = I}$、$\boldsymbol{C^T C = I}$ であり、負を許す条件下では主成分分析の高次拡張に対応する。

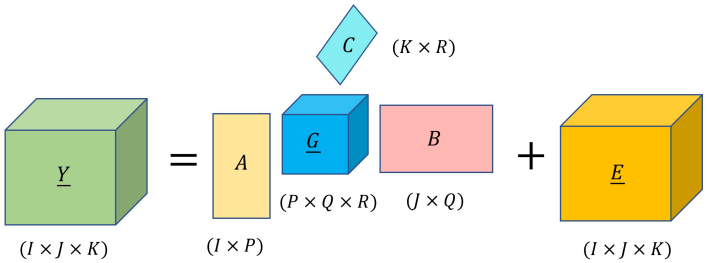


図 7.5: Tucker 分解のイメージ

テンソルデータ $\underline{\boldsymbol{Y}} \in \mathbb{R}^{\boldsymbol{I \times J \times K}}$ に対し、分解する行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_p}] \in \mathbb{R}^{I \times P}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_q}] \in \mathbb{R}^{J \times Q}$、$\boldsymbol{C} = [\boldsymbol{c_1}, \cdots, \boldsymbol{c_r}] \in \mathbb{R}^{K \times R}$、それらを連結するコアテンソルを $\underline{\boldsymbol{G}} \in \mathbb{R}^{\boldsymbol{P \times Q \times R}}$、誤差テンソルを $\underline{\boldsymbol{E}} \in \mathbb{R}^{\boldsymbol{I \times J \times K}}$ とするとき、Tucker 分解のモデルは以下のように表される、

$$\underline{\boldsymbol{Y}} = \underline{\boldsymbol{G}} \times \boldsymbol{A} \times \boldsymbol{B} \times \boldsymbol{C} + \underline{\boldsymbol{E}}, \tag{7.13}$$

ここで、「$\times$」はモード積を表す。モード積の計算例は図 7.6 の通りである。また (7.13) は以下のようにも表される、

$$y_{i,j,k} = \sum_{p=1}^{P} \sum_{q=1}^{Q} \sum_{r=1}^{R} g_{p,q,r} a_{i,p} b_{j,q} c_{k,r} + e_{i,j,k}, \tag{7.14}$$

また Tucker 分解は、コアテンソルが $\underline{\boldsymbol{G}} \in \mathbb{R}^{T \times T \times T}$ であり、テンソル内の要素が対角のみの場合、Harshman'S CP 分解と呼ばれ、図 7.7 のように表せる。

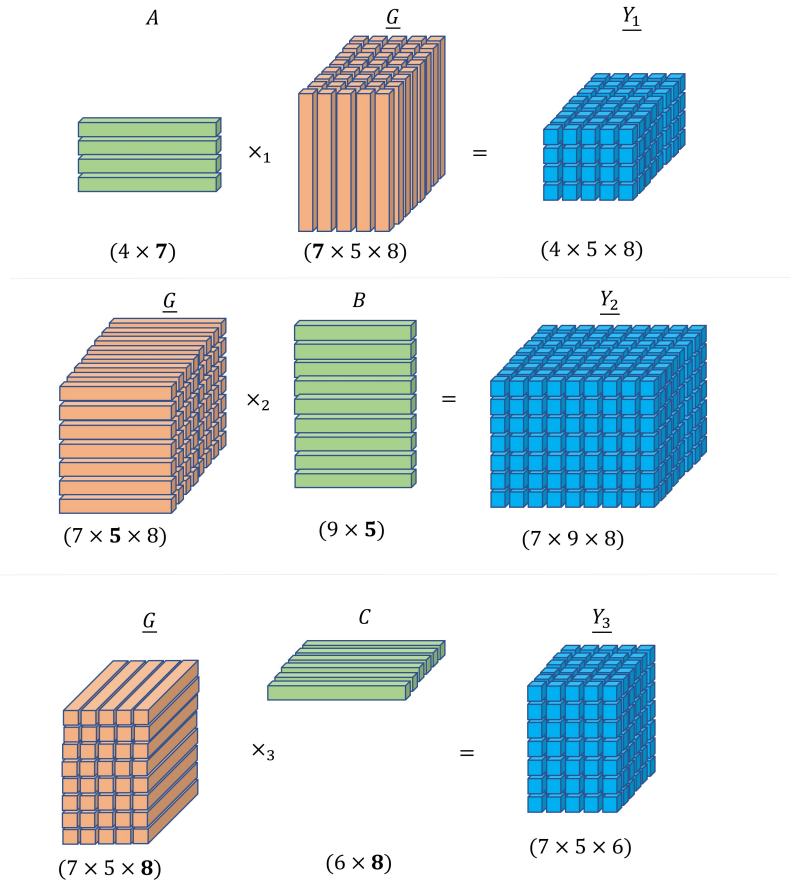


図 7.6: モード積の計算イメージ

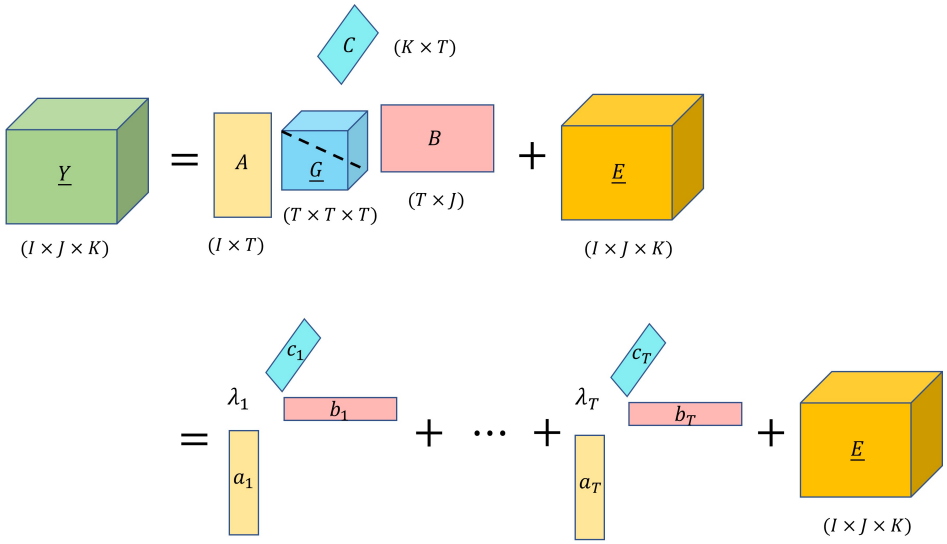


図 7.7: Harshman's CP 分解のイメージ

Tucker 分解の計算アルゴリズムの導出には、HOSVD (Higher Order Singular Value Decomposition) 等 [93,94] の高次元における直交条件付きの計算アルゴリズムが採用される。これは特異値分解 (SVD) を高次元に拡張したものであり、各モードごとに特異値分解を適用することで、直交制約下でのテンソル分解を行う解析手法である。すなわち以下の 2 つのステップでテンソルデータを分解する。[95]

STEP 1．行列 $\boldsymbol{Y}_{(n)}$ の特異値分解を行い、特異ベクトルを求める。

$$\boldsymbol{Y}_{(n)} = \boldsymbol{U}^{(n)}\boldsymbol{\Sigma}^{(n)}\boldsymbol{V}^{(n)T}. \tag{7.15}$$

STEP 2．元のテンソル $\underline{\boldsymbol{Y}}$ と (7.15) で得られた左特異ベクトル $\boldsymbol{U}^{(n)}$ から、以下のようにコアテンソル $\underline{\boldsymbol{G}}$ を計算する。

$$\underline{\boldsymbol{G}} = \underline{\boldsymbol{A}} \times \boldsymbol{U}^{(1)T} \times \boldsymbol{U}^{(2)T} \times \boldsymbol{U}^{(3)T}. \tag{7.16}$$

また Tucker 分解において分解される全ての行列及びコアテンソルに非負制約を付与した非負値 Tucker 分解 (Non-negative Tucker Decomposition,NTD) は HOSVD における Tucker 分解の導出条件における左特異ベクトルに非負の制約を付与することで導出できる。すなわち、上記に非負の制約を加えた 3 つのステップでテンソルデータを分解する [59]。

STEP1．行列 $\boldsymbol{Y}_{(n)}$ の特異値分解を行い、特異ベクトルを求める。

$$\boldsymbol{Y}_{(n)} = \boldsymbol{U}^{(n)}\boldsymbol{\Sigma}^{(n)}\boldsymbol{V}^{(n)T}. \tag{7.17}$$

STEP 2．$\boldsymbol{U}^{(n)}$ が負値を取った場合、0 に限りなく近い値に変換する。すなわち以下の条件を付す。

$$\boldsymbol{U}^{(n)} = max(\epsilon, \boldsymbol{U}^{(n)}),\ \epsilon = 2^{-52}. \tag{7.18}$$

STEP 3．元のテンソル $\underline{\boldsymbol{Y}}$ と (7.15) で得られた左特異ベクトル $\boldsymbol{U}^{(n)}$ から、以下のようにコアテンソル $\underline{\boldsymbol{G}}$ を計算する。

$$\underline{\boldsymbol{G}}_{+} = \underline{\boldsymbol{A}} \times \boldsymbol{U}_{+}^{(1)T} \times \boldsymbol{U}_{+}^{(2)T} \times \boldsymbol{U}_{+}^{(3)T}. \tag{7.19}$$

付録

付録 1：補足

劣勾配・劣微分について

点 $\boldsymbol{x} \in \mathbb{R}^n$と凸関数 $f : \mathbb{R}^n \to \mathbb{R}$ に関し、条件

$$f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \boldsymbol{s}^T(\boldsymbol{y} - \boldsymbol{x}),\ \forall \boldsymbol{y} \in \mathbb{R}^n,$$

を満たす $\boldsymbol{s}$ を点 $\boldsymbol{x}$ における f の劣勾配と呼ぶ。関数 f が点 $\boldsymbol{x}$ で一回微分可能であるならば、勾配 $\nabla f(\boldsymbol{x})$ が存在し、$\boldsymbol{s} = \nabla f(\boldsymbol{x})$ が成立する。また $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$ の時、点 $\boldsymbol{x} = 0$ における劣勾配 $\boldsymbol{s}$ は、n 次元閉区間 $[-1, 1]^n$ の任意の点となる。

また、劣勾配全体の集合を劣微分と呼び、$\partial f(\boldsymbol{x})$ とする。このとき以下のように表せる、

$$\partial f(\boldsymbol{x}) = \{\boldsymbol{s} \in \mathbb{R}^n | f(\boldsymbol{y}) \geq + f(\boldsymbol{x}) + \boldsymbol{s}^T(\boldsymbol{y} - \boldsymbol{x}),\ \forall \boldsymbol{y} \in \mathbb{R}^n\}.$$

関数 f が点 $\boldsymbol{x}$ で一回微分可能であるならば、$\partial f(\boldsymbol{x}) = \nabla f\boldsymbol{x}$ であり、点 $\boldsymbol{x} = 0$ における関数 $f(\boldsymbol{x}) = \|\boldsymbol{x}\|_1$に対して$\partial f(\boldsymbol{x}) = [-1, 1]^n$ である。[41, 42]

テンソルデータの計算について

(1)Kronecker Product

$\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_P}] \in \mathbb{R}^{I \times P}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_Q}] \in \mathbb{R}^{J \times Q}$ とするとき、Kronecker Product$\boldsymbol{A} \otimes \boldsymbol{B} \in \mathbb{R}^{IJ \times PQ}$ は以下のように定義される。

$$\begin{aligned}\boldsymbol{A} \otimes \boldsymbol{B} &= \begin{bmatrix} a_{11}\boldsymbol{B} & a_{12}\boldsymbol{B} & \cdots & a_{1P}\boldsymbol{B} \\ a_{21}\boldsymbol{B} & a_{22}\boldsymbol{B} & \cdots & a_{2P}\boldsymbol{B} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}\boldsymbol{B} & a_{I2}\boldsymbol{B} & \cdots & a_{IP}\boldsymbol{B} \end{bmatrix}, \\ &= \begin{bmatrix} \boldsymbol{a_1} \otimes \boldsymbol{b_1} & \boldsymbol{a_1} \otimes \boldsymbol{b_2} & \boldsymbol{a_1} \otimes \boldsymbol{b_3} & \cdots & \boldsymbol{a_P} \otimes \boldsymbol{b_{Q-1}} & \boldsymbol{a_P} \otimes \boldsymbol{b_Q} \end{bmatrix}.\end{aligned}$$

また、新たに $\boldsymbol{C} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_Q}] \in \mathbb{R}^{J \times Q}$ とした時、以下のような性質がある。

$$
\begin{aligned}
(\boldsymbol{A} \otimes \boldsymbol{B})^T &= \boldsymbol{A}^T \otimes \boldsymbol{B}^T, \\
\boldsymbol{A} \otimes (\boldsymbol{B} + \boldsymbol{C}) &= (\boldsymbol{A} \otimes \boldsymbol{B}) + (\boldsymbol{A} \otimes \boldsymbol{C}), \\
(\boldsymbol{B} + \boldsymbol{C}) \otimes \boldsymbol{A} &= (\boldsymbol{B} \otimes \boldsymbol{A}) + (\boldsymbol{C} \otimes \boldsymbol{A}), \\
c(\boldsymbol{A} \otimes \boldsymbol{B}) &= (c\boldsymbol{A}) \otimes \boldsymbol{B} = \boldsymbol{A} \otimes (c\boldsymbol{B}),
\end{aligned}
$$

(2)Hadamard Product

$\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_P}] \in \mathbb{R}^{I \times P}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_P}] \in \mathbb{R}^{I \times P}$ とするとき、Hadamard Product$\boldsymbol{A} \circledast \boldsymbol{B}$ は以下のように表せる。

$$
\boldsymbol{A} \circledast \boldsymbol{B} = \begin{bmatrix} a_{11}b_{11} & a_{12}b_{12} & \cdots & a_{1P}b_{1P} \\ a_{21}b_{21} & a_{22}b_{22} & \cdots & a_{2P}b_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ a_{I1}b_{I1} & a_{I2}b_{I2} & \cdots & a_{IP}b_{IP} \end{bmatrix}.
$$

(3)Khatri-Rao Product [96]

$\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_t}] \in \mathbb{R}^{I \times T}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_t}] \in \mathbb{R}^{J \times T}$ とするとき、$\boldsymbol{A} \odot \boldsymbol{B}$ は以下のように表せる。

$$
\begin{aligned}
\boldsymbol{A} \odot \boldsymbol{B} &= [\boldsymbol{a_1} \odot \boldsymbol{b_1} \quad \boldsymbol{a_2} \odot \boldsymbol{b_2} \quad \cdots \quad \boldsymbol{a_T} \odot \boldsymbol{b_T}], \\
&= \left[vec(\boldsymbol{b_1}\boldsymbol{a_1}^T) \quad vec(\boldsymbol{b_2}\boldsymbol{a_2}^T) \quad \cdots \quad vec(\boldsymbol{b_T}\boldsymbol{a_T}^T) \right], \quad \in \mathbb{R}^{IJ \times T}.
\end{aligned}
$$

また、新たに $\boldsymbol{C} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_T}] \in \mathbb{R}^{K \times T}$ とした時、Khatri-Rao Product は以下のような性質がある。

$$
\begin{aligned}
\boldsymbol{A} \odot (\boldsymbol{B} \odot \boldsymbol{C}) &= (\boldsymbol{A} \odot \boldsymbol{B}) \odot \boldsymbol{C}, \\
\boldsymbol{A} \odot \boldsymbol{B} &\neq \boldsymbol{B} \odot \boldsymbol{A}, \\
(\boldsymbol{A} \odot \boldsymbol{B})^T (\boldsymbol{A} \odot \boldsymbol{B}) &= \boldsymbol{A}^T \boldsymbol{A} \circledast \boldsymbol{B}^T \boldsymbol{B}.
\end{aligned}
$$

付録 2：テンソル因子分解に関する応用研究

地域間産業連関表を用いた県間取引構造の解析に関する研究 [原田・寒河江 (2021)]

本節 [97] では次元縮約法を用いて、地域間産業連関表から地域・産業間取引構造の特徴抽出を試みる。地域間産業連関表では、地域をまたいだ産業間取引構造が記録されていることから、地域経済活動を分析できる有益な指標である。一方、そのデータ構造は地域数 × 産業数の組み合わせで構成されるため、データ量が膨大であり、取引構造を客観的に解析するためには工夫が求められる。本節では、地域間産業連関表が地域内・地域間ごとに産業部門 × 産業部門の取引構造を有していることから、産業数 (供給)× 産業数 (需要)× 地域数の高次元データに配列し、テンソル因子分解を用いて特徴抽出を行った。同手法では、地域間取引の特徴を「供給」、「需要」、「地域」の 3 つの側面に分解でき、「ある産業とある産業の取引がどの地域間で活発に行われているのか」が解析できる。分析の結果、愛知県及びその周辺地域による輸送機械業を中核とした経済圏が中部圏の特徴であることが、同分析手法によって客観的に示された。

使用データと設定について

使用データについて

中部圏地域間産業連関表は、中部圏に属する 9 つの都道府県 (富山県・石川県・福井県・長野県・岐阜県・静岡県・愛知県・三重県・滋賀県) 及びその他全国間で生じる取引を 34 分類の生産者価格ベースで記録したデータである。ただし本節では中部圏間のみの取引を解析するため、その他全国との間で生じる取引を使用データに含めない。すなわち、34 産業 (供給)×34 産業 (需要) の取引のペアが 9 つの都道府県の組み合わせ分 (81 パターン) 存在する表であり、全体では 306×306(行、列ともに 34 産業 ×9 都道府県) の行列データとなっている。

中部圏の経済構造として、愛知県を中心とした製造業、特に輸送機械工業が全国より突出している特徴がある [99]。また中部圏社会経済研究所 (2015) [100] では、中部圏地域間産業連関表を用いて 9 つの都道府県間の産業取引に基づく空間的相互依存関係の可視化 [79, 101] を試みており、愛知県が石川県を除いた残り 7 つの都道府県に対するコアの役割を担っており、特に岐阜県・静岡県・三重県との繋がりが強いことが示されている。さらに輸送機械に絞った分析では、岐阜県・静岡県・三重県の輸送機械業と強いつながりがあることが示されている。

分析手法の設定

テンソル因子分解はテンソルデータを 3 つの行列データと 1 つのコアテンソルに分解し、データを縮約しながら特徴を解析できる分析手法である。本節では産業連関表が非負値であることから、非負値でのテンソル分解が可能である非負値テンソル因子分解 (Nonnegative Tucker Decomposition,NTD) [102] を適用する。分析手法のイメージは図 7.8 の通りである。

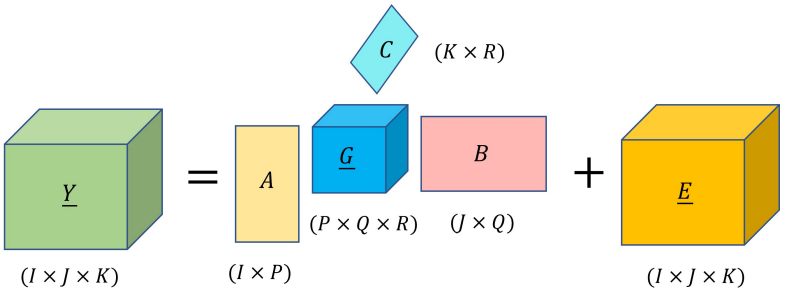


図 7.8: 非負値テンソル因子分解 (Nonnegative Tucker Decomposition,NTD) のイメージ図

テンソルデータ $\underline{\boldsymbol{Y}} \in \mathbb{R}^{\boldsymbol{I}\times\boldsymbol{J}\times\boldsymbol{K}}$ に対し、分解する行列を $\boldsymbol{A} = [\boldsymbol{a_1}, \cdots, \boldsymbol{a_p}] \in \mathbb{R}^{I\times P}$、$\boldsymbol{B} = [\boldsymbol{b_1}, \cdots, \boldsymbol{b_q}] \in \mathbb{R}^{J\times Q}$、$\boldsymbol{C} = [\boldsymbol{c_1}, \cdots, \boldsymbol{c_r}] \in \mathbb{R}^{K\times R}$、それらを連結するコアテンソルを $\underline{\boldsymbol{G}} \in \mathbb{R}^{\boldsymbol{P}\times\boldsymbol{Q}\times\boldsymbol{R}}$、誤差テンソルを $\underline{\boldsymbol{E}} \in \mathbb{R}^{\boldsymbol{I}\times\boldsymbol{J}\times\boldsymbol{K}}$ とするとき、非負値テンソル因子分解モデルは以下のように表される、

$$\underline{\boldsymbol{Y}} = \underline{\boldsymbol{G}} \times \boldsymbol{A} \times \boldsymbol{B} \times \boldsymbol{C} + \underline{\boldsymbol{E}}, \tag{7.20}$$

ここで、「×」はモード積を表す。また (7.20) は以下のようにも表される、

$$y_{i,j,k} = \sum_{p=1}^{P}\sum_{q=1}^{Q}\sum_{r=1}^{R} g_{p,q,r} a_{i,p} b_{j,q} c_{k,r} + e_{i,j,k}, \tag{7.21}$$

なお $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{R}$ は、データを縮約する観点から、$\{\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{R}\} << \{\boldsymbol{I}, \boldsymbol{J}, \boldsymbol{K}\}$ を満たすものとする。

データの設定

地域間産業連関表が地域内・地域間ごとに産業部門×産業部門の取引構造を有していることに着目し、図 7.9 のように「産業数 (供給)× 産業数 (需要)× 地域間ペア」のテンソル構造となるようデータを配列した。本データの構造に合わせると、$\boldsymbol{I} = 34$(供給), $\boldsymbol{J} = 34$(需要), $\boldsymbol{K} = 81$(都道府県の組み合わせ) に対応する。また $\boldsymbol{P}, \boldsymbol{Q}, \boldsymbol{R}$ は縮約させる次元数に対応するが、本節では解釈を優先し、$\boldsymbol{P} = 3$(供給)、$\boldsymbol{Q} = 3$(需要)、$\boldsymbol{R} = 3$(地域間ペア) と設定した。すなわち $\underline{\boldsymbol{Y}} \in \mathbb{R}^{34\times34\times81}$、$\boldsymbol{A} \in \mathbb{R}^{34\times3}$、$\boldsymbol{B} \in \mathbb{R}^{34\times3}$、$\boldsymbol{C} \in \mathbb{R}^{81\times3}$、$\underline{\boldsymbol{G}} \in \mathbb{R}^{3\times3\times3}$ となる設定の下で分析を行った。なお以降では 行列 $\boldsymbol{A}$、行列 $\boldsymbol{B}$、行列 $\boldsymbol{C}$ をそれぞれ「供給の基底」、「需要の基底」、「地域間ペアの基底」と呼び、各行列の列ベクトルを「供給の第 n 基底」…と呼ぶことにする。計算アルゴリズムは誤差を最小化する手法である最小二乗法を採用し、フロベニウスノルムを用いたデータの近似度を測定すると、$\boldsymbol{P} = 3$、$\boldsymbol{Q} = 3$、$\boldsymbol{R} = 3$ までの縮約で 71.2 %の近似精度であった。また分析は統計解析ソフトの「R」を使用した。

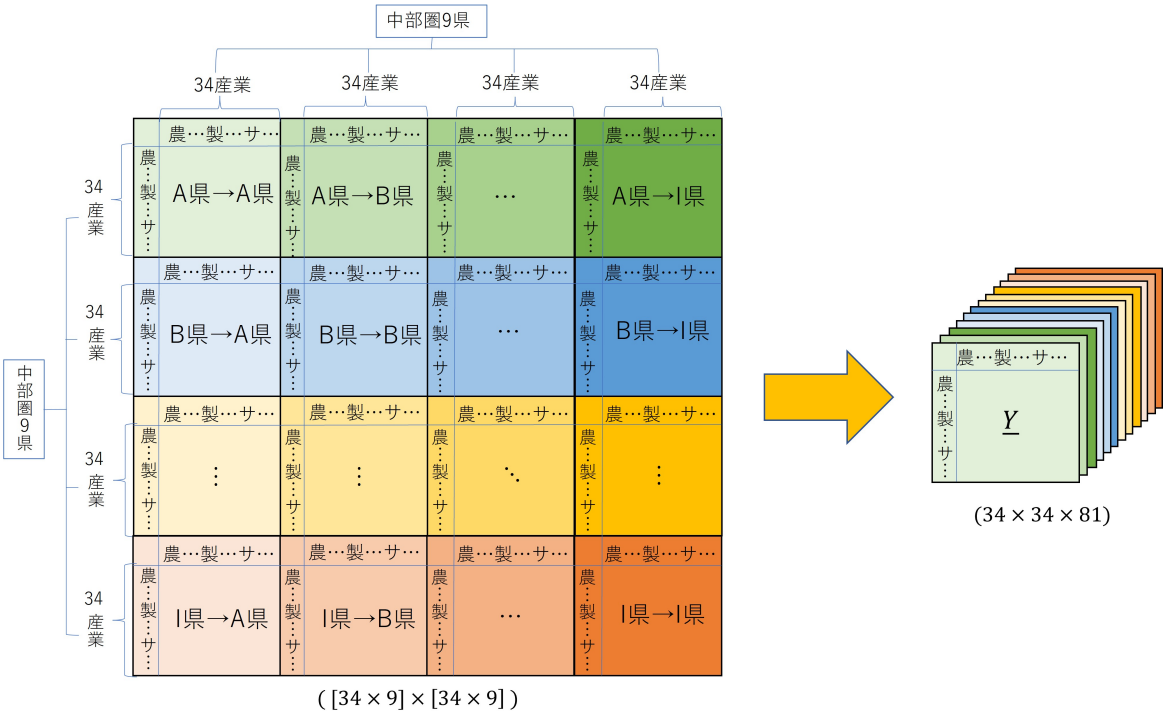


図 7.9: 地域間産業連関表からテンソルデータの作成方法

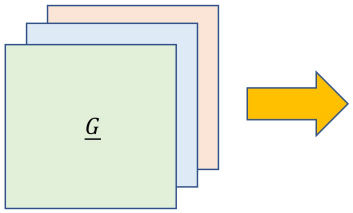
分析の結果

結果の概要

本稿で使用した計算アルゴリズムでは分解後の特徴を簡素化し、解釈を容易なものとするため、$\boldsymbol{A}$ 行列、$\boldsymbol{B}$ 行列、$\boldsymbol{C}$ 行列 のいずれの列ベクトルの 2 乗和が 1 となる制約で計算を行った。そのためデータテンソル $\underline{\boldsymbol{Y}}$ を近似するための量的要素は全てコアテンソル $\underline{\boldsymbol{G}}$ に集約されている。すなわちコアテンソル $\underline{\boldsymbol{G}}$ を確認することで、データテンソルに対する近似の度合いが解析できる。図 7.10 は非負値テンソル因子分解をデータテンソル $\underline{\boldsymbol{Y}}$ に適用し、3 つの行列と 1 つのコアテンソルに分解したうちの、コアテンソル $\underline{\boldsymbol{G}}$ を表している。本節では $\boldsymbol{P}=3$、$\boldsymbol{Q}=3$、$\boldsymbol{R}=3$ としたため、コアテンソルは $\underline{\boldsymbol{G}} \in \mathbb{R}^{3\times3\times3}$ であり、供給の第 1 基底から第 3 基底まで及び需要の第 1 基底から第 3 基底までの組み合わせの行列が地域間ペアの第 1 基底から第 3 基底数の分だけ存在している[5]。コアテンソル内の数値はデータテンソルを近似するための係数 (定数倍) で、単位は中部圏地域間産業連関表で使用されている単位と同じ、「百万円」である (ただしデータを縮約して近似する性質上、複数の産業の値も含めた数値となっているため、実際の産業連関表上の数値より大きくなっている)。またカッコ内で表される割合はコアテンソル $\underline{\boldsymbol{G}}$ の全要素の総和に占める各要素の割合を示しており、図 7.10 のカッコ内の値が大きいほど、データの近似を介して間接的に中部圏地域間産業連関表の特徴を強く説明していることになる。

これらから図 7.10 のカッコ内のパーセントに着目すると、「地域間ペア 1・供給 3・需要 3」の組み合わせが 45.5 %と最も高く、次いで「地域間ペア 2・供給 3・需要 3」の組み合わせが 44.1 %であり、これら 2 つで近似されたデータに対する寄与分が約 90 %である。その他に、「地域間ペア 2・供給 2・需要 1」の組み合わせが 4.53 %、「地域間ペア 3・供給 3・需要 3」の組み合わせが 2.16 %と続いている。

地域間ペアの特徴に関して、「地域間ペア 2」及び「地域間ペア 3」で表される特徴の寄与分が高いため、これらの地域間ペアで中部圏の中核を担う地域の特徴が抽出されていると想定される。また産業の特徴に関して、「供給 3・需要 3」の組み合わせで表される特徴の寄与分が高いため、中部圏地域間産業連関表の特徴を象徴する取引であることが想定される。本節では近似に対する寄与分が大きいこれら上位 4 つの組み合わせについて、それぞれの基底ごとの特徴を詳細に確認する。



コアテンソルG
(3 × 3 × 3)
(供給 × 需要 × 地域間ペア)

地域間ペア1	需要1	需要2	需要3
供給1	7.22E+08(0.01%)	5.92E+09(0.08%)	1.1E+10(0.15%)
供給2	1.2E+11(1.66%)	1.29E+09(0.02%)	8.02E+08(0.01%)
供給3	0(0%)	3.08E+10(0.43%)	3.29E+12(45.5%)
地域間ペア2	需要1	需要2	需要3
供給1	2.16E+09(0.03%)	1.58E+10(0.22%)	3.36E+10(0.47%)
供給2	3.28E+11(4.53%)	1.82E+09(0.03%)	1.23E+09(0.02%)
供給3	0(0%)	2.88E+10(0.40%)	3.18E+12(44.1%)
地域間ペア3	需要1	需要2	需要3
供給1	1.79E+3(0.0%)	0(0%)	0(0%)
供給2	8.55E+09(0.12%)	4.11E+08(0.05%)	4.17E+06(0.00%)
供給3	0(0%)	3.95E+09(0.05%)	1.56E+11(2.16%)

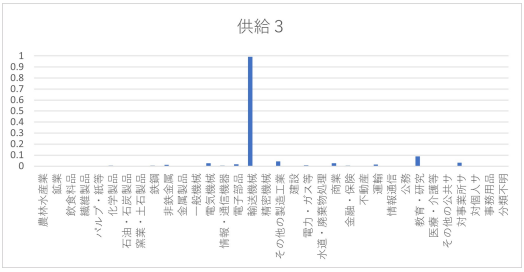
図 7.10: コアテンソルの構造とコアテンソル内の要素の値

[5]ただしここで表される供給 1 から供給 3 及び需要 1 から需要 3 は、いずれも単一の産業ではなく、複数の産業の情報が含まれている (図 7.11 から図 7.14 の a,b 参照)。また、例えば供給 1 と需要 1 など、基底数の番号が等しい場合でも同じ産業構造を有するとは限らないことに注意されたい。

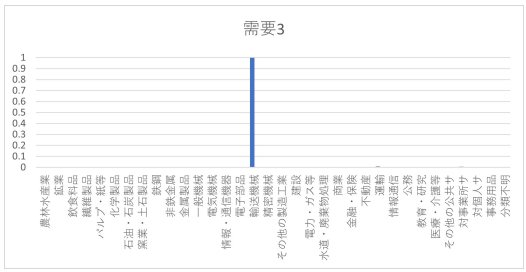
輸送機械に関する愛知県と周辺地域の取引関係

図 7.11 はコアテンソルにおける近似データに対する寄与分が最も高かった、「地域間ペア 1・供給 3・需要 3」の組み合わせに着目したものである。図 7.11a はある産業からの供給の特徴、図 7.11b はある産業への需要の特徴を示しており、いずれの基底も列ベクトルの 2 乗和が 1 となる。また図 7.11c はどの地域からどの地域への取引が活発であるかの組み合わせを表にしており、中部圏地域間産業連関表データに属する 9 県を列挙している。図中の数値は 2 乗和が 1 となる、かつ 0 から 1 の間の値を取り、1 に近いほど (色が赤に近い、または濃いほど)、対応する地域間の取引が活発であることを示している。図の見方に関しては以降の図 7.11 から図 7.14 も同様とする。また、供給の基底と需要の基底、地域間ペアの基底に関する詳細な数値は図 7.15 の通りとする。

これらの特徴から供給側を見ると、「輸送機械」が強く突出しており、「教育・研究」や「その他の製造工業」、「電気機械」等の特徴もわずかに見られる。また需要側を見ると「輸送機械」が強く突出しており、他の特徴はほとんど見られない。さらに地域間ペアの特徴では、「愛知県から愛知県」への取引が 0.91 と最も高い値を記録しており、次いで「三重県から三重県」が 0.23、「静岡県から静岡県」が 0.21 と、自地域内の取引が続く。他方で「愛知県から静岡県」が 0.17、「岐阜県から愛知県」が 0.15、「静岡県から愛知県」が 0.13 など、愛知県を中心に周辺地域との取引も本基底では見受けられる。これらの特徴から図 7.11 は「輸送機械」に関する愛知県及び周辺地域との取引関係を表した特徴であると考えられる。愛知県の「輸送機械」とその周辺地域の取引関係は中部圏社会経済研究所 [100] が示した結果と類似しており、輸送機械において愛知県がコアの役割を果たしていることがわかる。



(a) 供給パターン 3:第 1 基底



(b) 需要パターン 3:第 1 基底

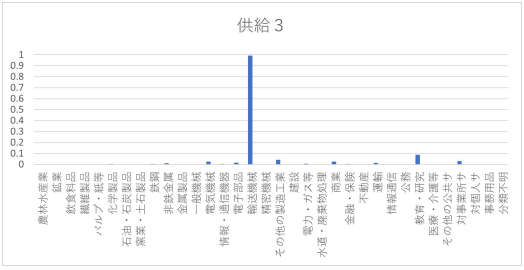
地域間ペア1	富山県	石川県	福井県	長野県	岐阜県	静岡県	愛知県	三重県	滋賀県
富山県	0	0	0	0	0	0	0	0	0
石川県	0	0	0	0	0	0	0	0	0
福井県	0	0	0	0	0	0	0	0	0
長野県	0	0	0	0	0.01	0	0.02	0	0
岐阜県	0	0	0	0	0.02	0	0.15	0	0
静岡県	0	0	0	0	0	0.21	0.13	0.01	0
愛知県	0	0	0	0	0.07	0.17	0.91	0.06	0.01
三重県	0	0	0	0	0.02	0.01	0.05	0.23	0
滋賀県	0	0	0	0	0	0	0	0	0.05

(c) 地域間パターン 1:第 1 基底

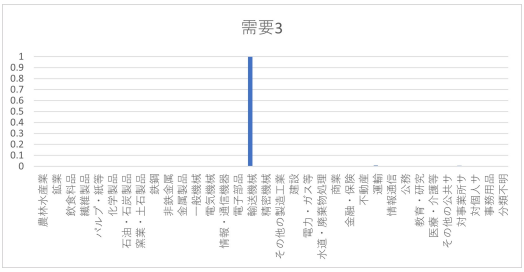
図 7.11: 「地域 1・供給 3・需要 3」の基底

輸送機械に関する愛知県内での取引

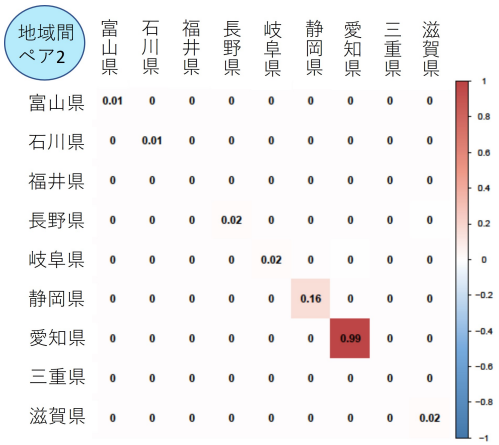
図 7.12 はコアテンソルにおける近似データに対する寄与分が 2 番目に高かった、「地域間ペア 2・供給 3・需要 3」の組み合わせに着目したものである。「供給 3・需要 3」の組み合わせは上述の「4.2 輸送機械に関する愛知県と周辺地域の取引関係」と同じであるが、地域間ペアの特徴が異なる基底である。図 7.12c の地域間ペアの特徴を確認すると、「愛知県から愛知県」への取引が 0.99 と極めて突出した値を記録している。これらの特徴から図 7.12 は「輸送機械」に関する愛知県内の取引を表した特徴であると考えられる。図 7.11 に続いて愛知県が突出して表れるのは、「輸送機械」に関して、愛知県の生産額が他の中部圏の地域のどの産業よりも極めて大きいためと考えられる。実際に愛知県の「輸送機械」は、使用した中部圏地域間産業連関表の生産者価格評価表の中で最も値が大きく (輸送機械→輸送機械の値)、域内総生産額が愛知県に次いで 2 番目に大きい静岡県で最も大きい値の約 3.5 倍 (輸送機械→輸送機械の値)、同 3 番目に大きい三重県で最も大きい値の 5 倍 (輸送機械→輸送機械の値)、その他中部圏地域においても同様に最も大きい値の 10 倍から 40 倍以上の生産額である。



(a) 供給パターン 3:第 2 基底



(b) 需要パターン 3:第 2 基底

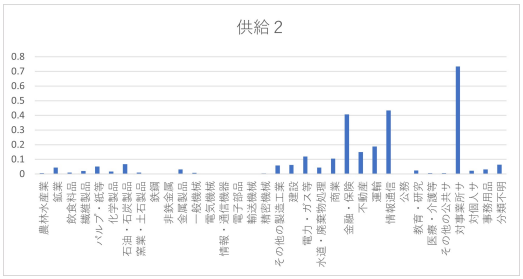


(c) 地域間パターン 1:第 2 基底

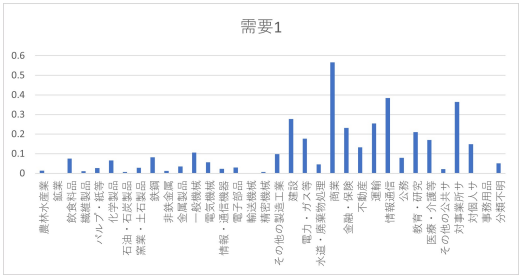
図 7.12: 「地域 2・供給 3・需要 3」の基底

第 3 次産業に関する愛知県内での取引

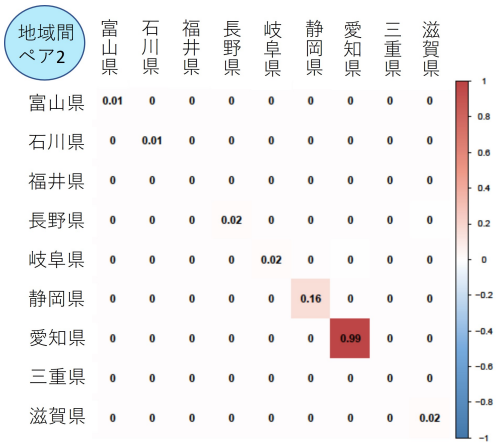
図 7.13 はコアテンソルにおける近似データに対する寄与分が 3 番目に高かった、「地域間ペア 2・供給 2・需要 1」の組み合わせに着目したものである。「地域間ペア 2」は上述の「4.3 輸送機械に関する愛知県内での取引」と同じであるため、供給と需要に関する特徴に着目する。図 7.13a の供給の特徴を確認すると、「対事業所サービス」が最も大きく抽出されており、次いで「金融・保険」や「情報通信」等の特徴が抽出されている。また需要側を見ると「商業」が最も大きく抽出されており、次いで「情報通信」や「対事業所サービス」、「建設」等の特徴が抽出されている。これらの特徴から図 7.13 は「第 3 次産業に関する愛知県内での取引」に関する特徴であると考えられる。この基底では愛知県の商業地区としての特徴が抽出され、第 3 次産業は愛知県内を中心に様々な産業と取引関係があることを示している。



(a) 供給パターン 2:第 3 基底



(b) 需要パターン 1:第 3 基底

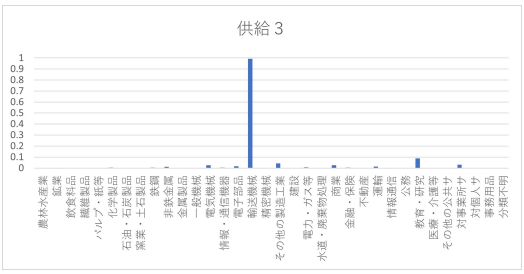


(c) 地域間パターン 2:第 3 基底

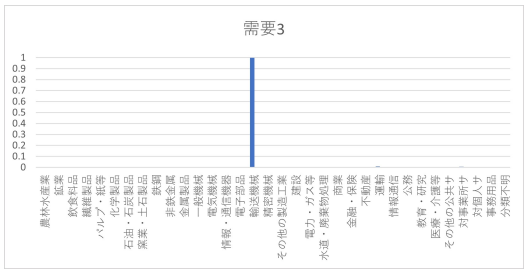
図 7.13: 「地域 2・供給 2・需要 1」の基底

輸送機械に関する愛知県以外の取引

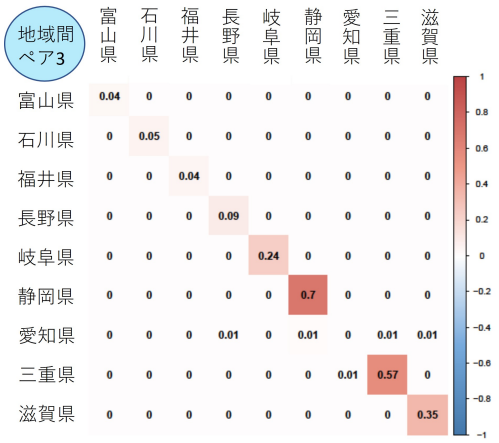
図 7.14 はコアテンソルにおける近似データに対する寄与分が 4 番目に高かった、「地域間ペア 3・供給 3・需要 3」の組み合わせに着目したものである。「供給 3・需要 3」の組み合わせは上述の「4.2 輸送機械に関する愛知県と周辺地域の取引関係」や「4.3 輸送機械に関する愛知県内での取引」と同じであるが、地域間ペアの特徴が異なっている。図 7.14c の地域間ペアの特徴を確認すると、「静岡県から静岡県」への取引が 0.70 と最も高く、「三重県から三重県」が 0.57、「滋賀県から滋賀県」が 0.35、「岐阜県から岐阜県」が 0.24 などと続いている。これらの特徴から図 7.14 は「輸送機械」に関する愛知県以外の取引を表した特徴であると考えられる。この基底では愛知県の特徴が全く含まれていない。すなわち図 7.10 の「地域間ペア 3」に対応する行列要素はいずれも愛知県以外の取引構造に着目した特徴であることが言える。ただし、「地域間ペア 3」の近似データに対する寄与分は本基底の 2.16 %が最大であり、全体の総和を計算しても、約 2.4 %程度に留まる。このことからも愛知県が中部圏に影響する規模の大きさが伺える。



(a) 供給パターン 3:第 4 基底



(b) 需要パターン 3:第 4 基底



(c) 地域間パターン 3:第 4 基底

図 7.14: 「地域 3・供給 3・需要 3」の基底

おわりに

本節では非負値テンソル因子分解を用いた地域間産業連関表からの産業間取引に関する特徴抽出を試みた。分析結果として中部圏では愛知県の存在が極めて大きく、テンソル因子分解後における地域間ペア基底 $\boldsymbol{C}$ の特徴を見ると、「愛知県」、「愛知県とその周辺地域」、「愛知県以外の周辺地域」の 3 つに分解でき、いずれも愛知県に関連する特徴が抽出された。「輸送機械」に関しても同様のことが言え、中部圏は愛知県及び輸送機械業を中核とした経済圏であることが分析された。

一方、愛知県や輸送機械など、大規模な地域・産業の特徴が除かれた組み合わせも特徴として抽出された。これは本節で採用したモデルの非負性による効果と考えられる。非負性からデータの近似が全て足し算によって表されるため、ある組み合わせで特徴が取り切れた場合、他の組み合わせ以降では先の組み合わせの影響を除いた特徴が抽出される。例えば「地域間ペア 3」は愛知県の特徴が全く抽出されなかったが、愛知県は「地域間ペア 1」や「地域間ペア 2」で抽出されている、かつ寄与分も高いため、これらで愛知県の特徴が概ね「取り切れた」ことによる影響であろう。

また本節の分析結果では、愛知県や輸送機械業に関する特徴が多く抽出されたが、他地域間との取引や産業間取引をさらに細かく分析したい場合、縮約する次元の条件を緩くする、具体的には 供給 $\boldsymbol{P} = 3$、需要 $\boldsymbol{Q} = 3$、地域間ペア $\boldsymbol{R} = 3$ の設定より大きい値にすることで解析可能となるだろう。

今回の分析結果に関連して、産業連関表では自地域内での取引が他地域間と比べて多いため、自地域間取引以外の「非対角ブロック」の特徴を客観的に抽出することが既存の次元縮約法では困難とされていた [68] が、本節における「テンソル化」の工夫によって「愛知県とその周辺地域」のような、他地域間との取引構造の特徴抽出に成功した。非対角ブロック要素の抽出は他の産業連関データ及び企業間取引データ等にも活用できる。例えば総務省が提供するような地域ブロックごとの地域間産業連関表や、WIOT(World Input-Output Database) が提供する国際産業連関表に対しても分析可能である。他にもある地域の産業構造の時系列変化の解析にも応用でき、例えば 5 年ごとの地域産業連関表をテンソル化することで年度変化を捉えた分析が可能となるだろう。

本節では地域間産業連関表をテンソル化して分析を行ったが、行列データのまま分析可能な非負値行列因子分解 [15, 60] 等の次元縮約法との使い分けは分析目的によるだろう。行列データのまま分析を行う場合、供給・需要共に都道府県と産業が紐づいているため、「ある地域のある産業が、どの地域のどの産業と取引が多いか」が解析できる。ただし、テンソル化する場合と比べてデータ数及び次元数が大きくなりやすく、特徴の解釈が複雑になる場合がある。例えば本節のデータを採用する場合、供給・需要の行列はいずれも「$306 \times k$」(k は縮約する次元数) で表されるため、複雑な解釈が求められる可能性がある。他方でテンソル因子分解の場合、産業と地域の直接的な紐づけはなくなるが、地域間の取引が紐づけられるため、「都道府県間のつながり」を解析したい場合に適している。非負値行列因子分解とテンソル因子分解のいずれの手法においても、地域間及び産業間取引の大きな傾向を把握するために有益な解析手法であろう。

	供給 1	供給 2	供給 3
農林水産業	0.000	0.006	0.000
鉱業	0.000	0.044	0.000
飲食料品	0.000	0.011	0.000
繊維製品	0.004	0.020	0.002
パルプ・紙等	0.001	0.051	0.000
化学製品	0.016	0.016	0.004
石油・石炭製品	0.004	0.066	0.001
窯業・土石製品	0.063	0.009	0.001
鉄鋼	0.549	0.000	0.005
非鉄金属	0.223	0.000	0.012
金属製品	0.159	0.031	0.002
一般機械	0.029	0.008	0.001
電気機械	0.078	0.000	0.026
情報・通信機器	0.003	0.001	0.005
電子部品	0.016	0.001	0.017
輸送機械	0.000	0.000	0.993
精密機械	0.003	0.005	0.000
その他の製造工業	0.131	0.058	0.042
建設	0.002	0.061	0.001
電力・ガス等	0.152	0.120	0.006
水道・廃棄物処理	0.008	0.045	0.001
商業	0.211	0.106	0.027
金融・保険	0.001	0.406	0.004
不動産	0.000	0.149	0.000
運輸	0.050	0.188	0.014
情報通信	0.000	0.433	0.000
公務	0.000	0.000	0.000
教育・研究	0.652	0.024	0.090
医療・介護等	0.000	0.006	0.000
その他の公共サ	0.004	0.006	0.000
対事業所サ	0.316	0.734	0.032
対個人サ	0.000	0.022	0.000
事務用品	0.000	0.032	0.001
分類不明	0.002	0.064	0.000

	需要1	需要2	需要3
農林水産業	0.014	0.000	0.000
鉱業	0.002	0.001	0.000
飲食料品	0.076	0.046	0.000
繊維製品	0.011	0.002	0.000
パルプ・紙等	0.027	0.013	0.000
化学製品	0.066	0.093	0.000
石油・石炭製品	0.007	0.000	0.000
窯業・土石製品	0.029	0.024	0.000
鉄鋼	0.082	0.361	0.000
非鉄金属	0.012	0.014	0.000
金属製品	0.035	0.039	0.000
一般機械	0.106	0.119	0.000
電気機械	0.057	0.090	0.000
情報・通信機器	0.023	0.031	0.000
電子部品	0.030	0.053	0.000
輸送機械	0.000	0.891	1.000
精密機械	0.007	0.008	0.000
その他の製造工業	0.098	0.081	0.000
建設	0.277	0.112	0.000
電力・ガス等	0.177	0.001	0.004
水道・廃棄物処理	0.046	0.014	0.000
商業	0.567	0.001	0.000
金融・保険	0.232	0.000	0.000
不動産	0.133	0.000	0.000
運輸	0.254	0.000	0.008
情報通信	0.385	0.046	0.000
公務	0.079	0.025	0.000
教育・研究	0.211	0.048	0.000
医療・介護等	0.171	0.051	0.000
その他の公共サ	0.022	0.001	0.000
対事業所サ	0.365	0.091	0.005
対個人サ	0.149	0.031	0.000
事務用品	0.001	0.001	0.000
分類不明	0.051	0.000	0.000

	地域間1	地域間2	地域間3
富山,富山	0.000	0.007	0.039
富山,石川	0.000	0.000	0.000
富山,福井	0.000	0.000	0.000
富山,長野	0.000	0.000	0.000
富山,岐阜	0.000	0.000	0.001
富山,静岡	0.001	0.000	0.003
富山,愛知	0.000	0.000	0.001
富山,三重	0.000	0.000	0.000
富山,滋賀	0.000	0.000	0.000
石川,富山	0.000	0.000	0.000
石川,石川	0.000	0.009	0.053
石川,福井	0.000	0.000	0.000
石川,長野	0.000	0.000	0.000
石川,岐阜	0.000	0.000	0.000
石川,静岡	0.000	0.000	0.000
石川,愛知	0.000	0.000	0.001
石川,三重	0.000	0.000	0.000
石川,滋賀	0.000	0.000	0.000
福井,富山	0.000	0.000	0.000
福井,石川	0.000	0.000	0.000
福井,福井	0.000	0.004	0.045
福井,長野	0.000	0.000	0.000
福井,岐阜	0.000	0.000	0.000
福井,静岡	0.000	-0.000	0.000
福井,愛知	0.002	0.000	0.005
福井,三重	0.000	0.000	0.000
福井,滋賀	0.000	0.000	0.000
長野,富山	0.000	0.000	0.000
長野,石川	0.000	0.000	0.000
長野,福井	0.000	-0.000	0.000
長野,長野	0.000	0.020	0.092
長野,岐阜	0.000	0.000	0.000
長野,静岡	0.003	0.000	0.001
長野,愛知	0.002	0.000	0.005
長野,三重	0.000	0.000	0.000
長野,滋賀	0.000	0.000	0.000
岐阜,富山	0.000	0.000	0.000
岐阜,石川	0.000	0.000	0.000
岐阜,福井	0.000	0.000	0.000
岐阜,長野	0.005	0.000	0.000
岐阜,岐阜	0.022	0.017	0.237
岐阜,静岡	0.000	0.000	0.001
岐阜,愛知	0.066	0.000	0.003
岐阜,三重	0.023	0.000	0.000
岐阜,滋賀	0.001	0.000	0.000
静岡,富山	0.000	0.000	0.000
静岡,石川	0.000	0.000	0.000
静岡,福井	0.000	0.000	0.000
静岡,長野	0.005	0.000	0.001
静岡,岐阜	0.004	0.000	0.001
静岡,静岡	0.208	0.161	0.698
静岡,愛知	0.171	0.000	0.011
静岡,三重	0.006	0.000	0.002
静岡,滋賀	0.001	0.000	0.000
愛知,富山	0.003	0.000	0.003
愛知,石川	0.000	0.000	0.000
愛知,福井	0.001	0.000	0.000
愛知,長野	0.015	0.001	0.000
愛知,岐阜	0.145	-0.000	0.003
愛知,静岡	0.130	0.000	0.004
愛知,愛知	0.908	0.986	0.000
愛知,三重	0.049	-0.000	0.005
愛知,滋賀	0.004	0.000	0.003
三重,富山	0.000	0.000	0.000
三重,石川	0.000	0.000	0.000
三重,福井	0.000	0.000	0.000
三重,長野	0.003	0.000	0.000
三重,岐阜	0.003	0.000	0.001
三重,静岡	0.010	0.000	0.001
三重,愛知	0.057	0.000	0.007
三重,三重	0.225	0.002	0.568
三重,滋賀	0.001	0.000	0.000
滋賀,富山	0.000	0.000	0.000
滋賀,石川	0.000	0.000	0.000
滋賀,福井	0.000	-0.000	0.000
滋賀,長野	0.000	-0.000	0.001
滋賀,岐阜	0.000	0.000	0.001
滋賀,静岡	0.001	0.000	0.001
滋賀,愛知	0.007	0.000	0.009
滋賀,三重	0.001	0.000	0.000
滋賀,滋賀	0.047	0.019	0.345

図 7.15: 供給の基底 (左)、需要の基底 (中)、地域の基底 (右) の各列ベクトルに占める重み (2 乗和が 1)
※供給と需要は各列上位 5 項目、地域は各列上位 10 項目に色を付けている。

参考文献

[1] 内閣府. society5.0 https://www8.cao.go.jp/ 2021 年 9 月 20 日閲覧

[2] K, Pearson.“On Lines and Planes of Closest Fit to Systems of Points in Space”, Philosophical Magazine 2 (11), 559–572,1901.

[3] D.D. Lee. and H. S. Seung. ”Learning the parts of objects by non-negative matrix factorization”, Nature, Vol.401, pp. 788-791, 1999.

[4] R, Tibshirani.”Regression Shrinkage and Selection via the Lasso”Journal of the Royal Statistical Society. Series B (Methodological) ,Vol. 58,No. 1, pp. 267-288,1996.

[5] 総務省.　産業連関表 2021 年 9 月 25 日閲覧

[6] W, Isard., E.W. Schooler. and T, Vietorisz. ”Industrial Complex Analysis and Regional Development: A Case Study of Refinery-Petrochemical-synthetic-Fiber Complexes and Puerto Rico”Technology Press of the Massachusetts Institute of Technology,1959.

[7] S, Czamanski. ”Study of Clustering of Industries”. Halifax, Nova Scotia, Canada, Institute of Public Affairs, Dalhousie University,1974.

[8] H.D.Roepke., D, Adams. and R, Wiseman. ”A New Approach to the Identification of Industrial Complexes Using Input-Output Data” Journal of Regional Science 14,1,15-29, 1974.

[9] E. M. Bergman. and E. J. Feser. ”Industrial and Regional Clusters: Concept and Comparative Applications”. Web Book in Regional Science, Regional Research Institute, West Virginia University ,1999.

[10] E. J. Feser. and E. M. Bergman. ”National Industry Templates: A Framework for Applied Regional Cluster Analysis” Regional Studies 34.1, pp.1-19, 2000.

[11] R. V. Hofe. and S. D. Bhatta. ”Method for identifying local and domestic industrial clusters using interregional commodity trade data”,The Industial Geographer Vol.4 , 2 ,2007.

[12] 長沢克重.「因子分析による投入産出構造変動の分析一昭和 45-50-55 年接続産業連関表による一」経済統計学会 , 第 55 号 , pp.52-63 ,1988.

[13] 渡邊隆俊, 下田充, 藤川清史.「投入構造と産出構造からみた産業クラスターの地域別特性 -2000 年の関東・中部・近畿を例にとって一」経営経済 , 第 44 号 , pp.39-64, 2009.

[14] 千葉雄二.「市町村間産業連関表の作成と町村の存続」, 日本地域学会年次大会学術発表論文集 (web) , 第 56 巻 , 2019.

[15] 楫取和明.「非負値行列因子分解による産業クラスター検出の試み」 ,Journal of National Fisheries University ,64(4) ,pp.227-239 , 2016.

[16] A, Mascaretti.“Non-negative matrix factorization and compositional clustering of national input-output tables”, politesi , 2019.

[17] NTT ドコモ, モバイル空間統計

[18] 細江美欧，桑野将司，谷本圭志 .「非負値テンソル因子分解を用いた交通系 IC カードデータからの移動パターンの抽出に関する研究」, 都市計画論文集, Vol.53, No.3, pp.1320-1326, 2018.

[19] D,Yao., C, Yu. and Q, Ding. ”Human mobility synthesis using matrix and tensor factorizations”, Information Fusion, Vol.23, pp.25-32, 2015.

[20] D ,Wang., Z, Cai., Y, Cui. and X, Chen. ”Nonnegative tensor decomposition for urban mobility analysis and applications with mobile phone data”,Transportmetrica A: Transport Science,2019.

[21] 奥村誠.「都市内災害復旧家庭の自空間パターンの把握 携帯電話位置情報集計データの活用」, 土木計画論文集, Vol50 , No.3 ,2015.

[22] 林亜紀, 亀岡弘和, 松林達史, 澤田宏. 「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」, ARG WI2 No.7, 2015.

[23] L, Alexander., S, Jiang., M, Murga. and M. C. González. ”Origin-destination trips by purpose and time of day inferred from mobile phone data”, Transportation Research Part C: Emerging Technologies, Vol. 58, pp.240–250, 2015.

[24] A, Alsger., A, Tavassoli., M,Mesbah., L, Ferreira. and M, Hickman. ”Public transport trip purpose inference using smart card fare data”,Transportation Research Part C: Emerging Technologies, Vol. 87, pp.123–137, 2018.

[25] H,Yamaguchi., and S, Nakayama.” Detection of base travel groups with different sensitivities to new high-speed rail services: Non-negative tensor decomposition approach”,Transport Policy, Vol. 97, pp.37–46, 2020.

[26] 永田靖, 棟近雅彦.「多変量解析法入門」, サイエンス社

[27] 三宅敏恒,「入門線形代数」, 培風館, 1991

[28] 亀岡弘和 .「非負値行列因子分解」, 計測と制御 51(9), pp.835–844, 2012.

[29] 亀岡弘和. 「非負値行列因子分解とその音響信号処理への応用」, 日本統計学会誌 第 44 巻 第 2 号, pp.383–407, 2015.

[30] D. D.Lee. and H. S. Seung. ”Algorithms for nonnegative matrix factorization”, Advances in Neural Information Processing Systems 13 (NIPS 2000), pp.556–562, 2000.

[31] I, Csiszár. ”I-divergence geometry of probability distributions and minimization problems”, The Annals of Probability, 3(1), pp.146–158, 1975

[32] 板倉文忠 .「統計的手法による音声分析合成系に関する研究」博士論文, 名古屋大学大学院工学研究科, 1972.

[33] C, Févotte., N, Bertin. and J. L. Durrieu. ”Nonnegative matrix factorization with the Itakura-Saito divergence”, With application to music analysis, Neural Computation, 21(3), pp.793–830, 2009.

[34] S, Eguchi. and Y, Kano.”Robustifying maximum likelihood estimation”, Technical report, Institute of Statistical Mathematics, Research Memo, 802, 2001

[35] J. M. Ortega. and W. C. Rheinboldt.”Iterative Solutions of Nonlinear Equations in Several Variables”, Academic Press, New York ,1970.

[36] D. R. Hunter. and K, Lange. ”Quantile regression via an MM algorithm”, Journal of Computational and Graphical Statistics, 9, pp.60–77, 2000.

[37] 木村圭吾, 吉田哲也.「特徴表現のスパース性を考慮した NMF」, 情報処理学会研究報告,Vol2011-MPS-85 ,No.2 2011

[38] 福島雅夫.「非線形最適化の基礎」, 朝倉書店,2001

[39] T, Hastie., R, Tibshirani. and M,Wainwright. ”Statistical Learning with Sparsity :The Lasso and Generalization”.Chapman & Hall/CRC Monographs on Statistics and Applied Probability 143,2015.

[40] A. E. Hoerl. and R. W. Kennard.”Ridge Regression: Biased Estimation for Nonorthogonal Problems”,Technometrics ,Vol. 12, No. 1 pp. 55-67 , 1970.

[41] 川野秀一, 松井秀俊, 廣瀬慧. 「スパース推定法による統計モデリング」, 共立出版, 2018

[42] 寒野善博, 土谷隆. 「最適化と変分法」, 丸善出版

[43] B, Efron., T, Hastie., I, Johnstone. and R, Tibshirani.”Least angle regression (with discusion)”, Annals of statistics, 32, pp.407-499, 2004.

[44] J, Friedman., T, Hastie., H, Höfling. and R, Tibshirani. ”Pathwise Coordinate Optimization”, Annals of Applied Statistics, Vol. 1, No. 2, pp. 302–332, 2007.

[45] S, Boyed., N, Parikh., E, Chu., B, Peleato. and J, Eckstein.”Distributed optimization and statistical learning via the alternating direction method of multipliers”, Foundations and trends in Machine Learning, 3 , pp.1-122 , 2011.

[46] H, Zou. and T, Hastie. ”Regularization and Variable Selection via the Elastic Net”, Journal of the Royal Statistical Society, Series B, pp.301-320 , 2005.

[47] R, Tibshirani., M, Saunders., S, Rosset., J, Zhu. and K, Knightm.”Sparsity and smoothness via the fused lasso”, Journal of the Royal Statistical Society Series B, 67, pp.91-108, 2005.

[48] J, Fan., R, Li.”Variable selection via nonconcave penalized likelihood and its oracle properties”, Journal of the American Statistical Association, 96, pp.1348-1360, 2001.

[49] H, Zou.”The adaptive lasso and its oracle properties”,Journal of the American Statistical Association, 101, pp.1418-1429, 2006.

[50] R, Tibshirani. and J, Taylor.”The solution path of the generalized lasso”, Annals of Statistics, 39, pp.1335-1371, 2011.

[51] Y, She., ”Sparse regression with exact clustering”, Electronic Journal of Statistics, 4, pp.1055-1096 , 2010.

[52] I. T. Jolliffe, N. T. Trendafilov. and M, Uddin.“ A Modified Principal Component Technique Based on the LASSO ”, Journal of Computational and Graphical Statistics, Vol. 12, No. 3, pp. 531-547, 2003.

[53] H, Zou., T, Hastie. and R, Tibshirani. ”Sparse Principal Component Analysis”, Journal of Computational and Graphical Statistics, Volume 15, No. 2, pp. 265–286, 2006

[54] N. B. Erichson., P, Zheng., K, Manohar., S, Brunton., J. N. Kutz. and A. Y. Aravkin.”Sparse Principal Component Analysis via Variable Projection”, SIAM Journal on Applied Mathematics 80:2, pp.977-1002, 2020.

[55] P. O. Hoyer. ”Non-negative sparse coding”, In Neural Networks for Signal Processing XII (Proc. IEEE Workshop on Neural Networks for Signal Processing),Martigny, Switzerland, pp.557–565, 2002.

[56] V. P. Pauca., J, Piper. and R. J. Plemmons. ”Nonnegative matrix factorization for spectral data analysis”, Linear Algebra and Applications,Vol.416, 1, 1, pp. 29-47, 2006.

[57] H, Kim. and H, Park ”Sparse non-negative matrix factorizations via alternating non-negativity-constrained least squares for microarray data analysis”, Bioinformatics. 15;23(12), pp.1495-1502, 2007.

[58] A, Cichocki., R, Zdunek. and S, Amari. ”Hierarchical ALS Algorithms for Nonnegative Matrix and 3D Tensor Factorization”. In: Davies M.E., James C.J., Abdallah S.A., Plumbley M.D. (eds) Independent Component Analysis and Signal Separation, ICA 2007m Lecture Notes in Computer Science, vol 4666, Springer, Berlin, Heidelberg , 2007

[59] A, Cichocki., R, Zdunek.. A. H. Phan. and S, Amari.”Nonnegative Matrix and Tensor Factorizations : Applications to Exploratory Multi-way Data Analysis and Blind Source Separation”,WILEY 2009

[60] 原田魁成, 寒河江雅彦.「非負値行列因子分解法を用いた地域産業特性の抽出」, 環太平洋産業連関分析学会第 30 回大会, 2019.

[61] 中部経済産業局, 東海経済のポイント 2018（３大経済圏比較）
http://www.chubu.meti.go.jp/a51chosa/point.html 2018 年 12 月 20 日閲覧

[62] 石油連盟, 製油所の所在地と原油処理能力
http://www.paj.gr.jp/statis/statis/ 2018 年 12 月 20 日閲覧

[63] 原田魁成・寒河江雅彦・山口裕通「COVID-19 下での石川県における移動行動分析」人間社会環境研究第 42 号 pp.183-197 2021

[64] RESAS, https://resas.go.jp/

[65] esri ジャパン ArcGIS https://www.esrij.com/

[66] 住みよさランキング 2020, 東洋経済, https://toyokeizai.net/ 2021 年 3 月 22 日閲覧

[67] 原田魁成, 山口裕通, 寒河江雅彦.「スパース非負値行列因子分解を用いた COVID-19 流行期の県間旅行行動の変容分析」, 土木学会論文集 D3 , 77 巻 2 号, pp. 160-173, 2021.

[68] 原田魁成, 寒河江雅彦.「羽咋市におけるモバイル空間統計データによる人口流動分析～非負値行列因子分解法を用いて～」, 人間社会環境研究, 第 41 号, pp.63-74, 2021.

[69] 山本けい子, 原田魁成, 寒河江雅彦.「主成分分析に基づく地域クラスタリングと産業構造の可視化」人間社会環境研究第 42 号, pp.199-211 2021.

[70] 福島県,「平成 23 年度福島県産業連関表」, https://www.pref.fukushima.lg.jp/ 2021 年 9 月 28 日閲覧

[71] 一般財団法人 土地総合研究所,「リサーチ・メモ ミニバブル期の状況を振り返って（2）」
https://www.lij.jp/ 2021 年 9 月 28 日閲覧

[72] 国土交通省,「地価公示結果の概要」 https://www.mlit.go.jp/ 2021 年 9 月 28 日閲覧

[73] 石川県, 石川県統計指標ランド https://www.pref.ishikawa.lg.jp/ 2021 年 9 月 28 日閲覧

[74] 原田魁成, 寒河江雅彦.「COVID-19 流行前後における生活行動の変化を捉えるためのスパース PCA 分析」 , 2021 年度統計関連学会連合大会, 2021.

[75] P. O. Perry. and A. B. Owen. ”Bi-cross-validation of the svd and the non-negative matrix factorization”. InAnnals of Applied Statistics, pp.564–594, 2009

[76] AH, Williams., A, Degleris., Y, Wang., and SW, Linderman. ”Point process models for sequence detection in high-dimensional neural spike trains”, Neural Information Processing Systems, Vancouver, CA, 2020.

[77] F, YATES. ” INCOMPLETE RANDOMIZED BLOCKS”, Ann.Eugen, 7, pp.121-140, 1936.

[78] A.Ghosh.,”Input-Output Approach in an Allocation System”,Economica,25,pp.58-64,1958

[79] 猪俣哲史.「産業間の「距離」を計るーアジア国際産業連関表を用いた平均波及世代数の計測」, 産業連関 Vol.16, No.1, pp.45-56, 2008.

[80] H, Akaike. ”Information theory and an extension of the maximum likelihood principle”, Proceedings of the 2nd International Symposium on Information Theory, Petrov, B. N., and Caski, F. (eds.), Akadimiai Kiado, Budapest: pp.267-281, 1973.

[81] G, Schwarz. ”Estimating the Dimension of a Model”,Ann. Statist. 6(2): pp.461-464 ,1978.

[82] J, Bai. and S, Ng., ”Determing the number of factors in approximate factor models”, Econometrica, Vol70, No.1, pp.191-221, 2002.

[83] 安川武彦.「非負値行列因子分解を用いたテキストデータ解析」, 計量器統計学,28 巻 1 号, pp.41-55, 2015.

[84] A. T. Cemgil. ”Bayesian Inference for NonnegativeMatrix FactorisationModels”, Hindawi Publishing Corporation Computational Intelligence and Neuroscience Volume 2009, Article ID 785152, 2009.

[85] M. N. Schmidt., O, Winther. and L. K. Hansen,”Bayesian Non-negative Matrix Factorization”, International Conference on Independent Component Analysis and Signal Separation ICA 2009, Independent Component Analysis and Signal Separation ,pp.540-547, 2009.

[86] K, Tanabe., M, Sagae., ”Improper priors and model selection criteria in misspecified bayes linear models”, ノンパラメトリック・ファンクショナル推定の理論と応用, 統計数理研究所共同研究リポート 134, pp.1-26 2000.

[87] 柳井晴夫, 竹内啓. 「射影行列・一般逆行列・特異値分解」, 東京大学出版会, 1983.

[88] B, Kanagal. and V, Sindhwani. ”Rank Selection in Low-rank Matrix Approximations:A Study of Cross-Validation for NMFs”,Low-rank Methods for Large-Scale Machine Learning Workshop, NIPS 2010.

[89] T. G. Kolda. and B. W. Bader.”Tensor Decompositions and Applications”, SIAM Review,Vol. 51, No. 3, pp. 455–500, 2009.

[90] J. D. Carroll and J. J. Chang. ”Analysis of individual differences in multidimensional scaling via an N-way generalization of “ Eckart-Young ” decomposition”, Psychometrika, 35, pp.283–319,1970

[91] R. A. Harshman. ”Foundations of the PARAFAC procedure: Models and conditions for an “ explanatory ” multi-modal factor analysis”, UCLA Working Papers in Phonetics, 16, pp.1–84, 1970.

[92] H. A. L. Kiers, ”Towards a standardized notation and terminology in multiway analysis,” J. Chemometrics, 14, pp.105–122, 2000

[93] L. D. Lathauwer., B. D. Moor. and J, Vandewalle. ”A multilinear singular value decomposition”, SIAM journal on Matrix Analysis and Applications, 21, 4 ,pp.1253-1278, 2000.

[94] L. D. Lathauwer.”Decompositions of a higher-order tensor in block terms—Part I: Lemmas for partitioned matrices”,SIAM journal on Matrix Analysis and Applications, 30, 3, pp.1022-1032, 2008

[95] 山本直樹, 村上純, 石田明男.「テンソル分解プログラミングの理解支援のための立体パズルの利用」ソフトウェアシンポジウム 2019 pp.115-124, 2019.

[96] C. G. Khatri. and R. C. Rao. ”Solutions to Some Functional Equations and Their Applications to Characterization of Probability Distributions”. Sankhya: Indian J. Statistics, Series A 30, pp.167–180, 1968.

[97] 原田魁成, 寒河江雅彦.「非負値テンソル因子分解を用いた地域間産業連関構造の特徴抽出」環太平洋産業連関分析学会第 32 回大会, 2021.

[98] 公益財団法人中部圏社会経済研究所.「中部圏地域間産業連関表（延長表 2010 年版）」

[99] 経済産業省中部経済産業局.「中部経済のポイント 2020」
https://www.chubu.meti.go.jp/a51chosa/point.html 2021 年 9 月 18 日閲覧

[100] 公益財団法人中部圏社会経済研究所.「中部圏の地域経済構造 ～一極集中型から多極分散型へ～」
https://www.criser.jp/research/documents/2015_ cyubuken.pdf 2021 年 9 月 18 日閲覧

[101] E, Dietzenbacher., I, Romero., N. S. Bosma. ”Using Average Propagation Lengths to Identify Production Chains in the Andalusian Economy”, Estudios de Economia APLicada,23(2), pp.405-422, 2005.

[102] Y. D. Kim. and S. Choi., ”Nonnegative Tucker Decomposition”, IEEE Conference on Computer Vision and Pattern Recognition, 2007