空間トピックモデルによる ~~土地利用~~のモニタリング

社会経済活動集積量

広島大学工学研究科　地球環境計画学研究室

塚井　誠人

（塚野　裕太：M2　&　QUON ANN TRAN：D1）

Hedgehog

2018年6月22日
愛媛大学坊ちゃんセミナー

David Harveyの懸念@2004年夏の読書から

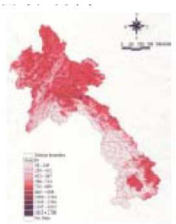
- ✓ 1935年，英国生まれ（存命）。ケンブリッジ大学より博士号取得。ブリストル大学講師、ジョンズ・ホプキンス大学准教授・教授、オックスフォード大学教授を経て、現在、ニューヨーク市立大学名誉教授。
- ✓ 『資本論』を中心とするマルクス主義を地理学に応用した批判的地理学の第一人者。社会的不公平に深い関心。地理学分野では、世界で最も多く論文が引用される学者。

wikipediaより

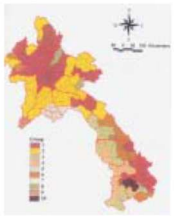
- 地理学は，地理的な共通性と相違に対する深い洞察が必要な学問.
- 然るに，地理情報データに主成分分析などを当てはめて，各地点に共通する特性にのみ着目するアプローチは，地理的な相違から目を逸らすもので，受け入れ難い…

地理情報に主成分分析を．．．

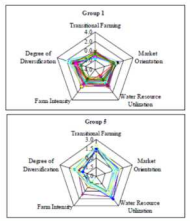
- 1972年当時では最新の研究分野だったはず：衛星画像に基づく土地被覆分類？
- google検索：N工営の国際事業部報告書がヒット



標高データ（DEM）や森林被服図などラオス政府所有の既存の地理情報システム（GIS）グリッドデータ等をフルに活用し、全国の総合的な農業現況の把握に役立てた。



地理情報システム（GIS）の応用解析および、主成分分析、クラスター分析により、ラオス全国 141 の地域を農業開発の現況に基づき 10 のグループに分割。



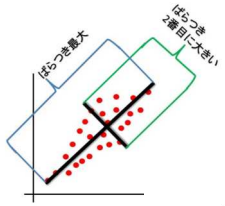
抽出された農業開発に関わる5つの主成分に関して、レーダー・チャートを作成。レーダー・チャートにより、各グループの特性を把握。農業開発阻害要因と開発に必要な重点項目を特定。

1. 主成分分析
2. クラスター分析

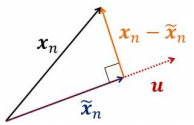
https://www.n-koei.co.jp/rd/thesis/pdf/200212/forum11_009g.pdf

技術的なおさらい１：主成分分析

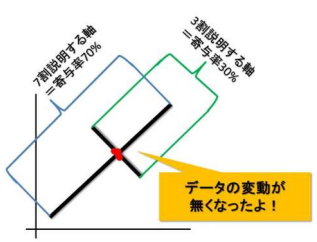
- J次元空間内のI個のベクトルxに関して，新たな軸uを発見する方法．xのu上への射影ベクトルの分散最大化問題として定式化．ただしuは単位ベクトル（長さ1）

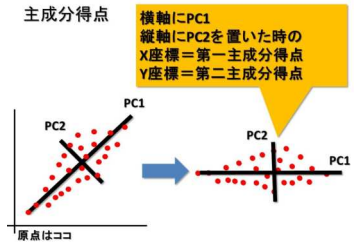


$$L(u)=\frac{1}{N}\sum_n\left(\mathbf{u}^T\mathbf{x}_n\right)^2+\lambda\left(1-\mathbf{u}^T\mathbf{u}\right)$$



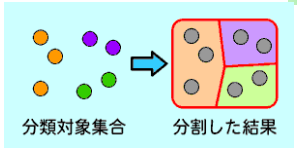
Lを最大化＝軸分散最大化



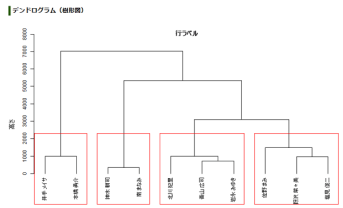


技術的なおさらい２－１：クラスター分析

- J次元空間内のI個のサンプルから，サンプル間距離に基づいてk個のサンプルグループを構成．階層的方法と非階層的方法あり．



- 階層的方法：k=Iから始めてサンプルを併合し，最後はk=1になるデンドログラム（逆樹形図）／kは，図を見て選ぶ．



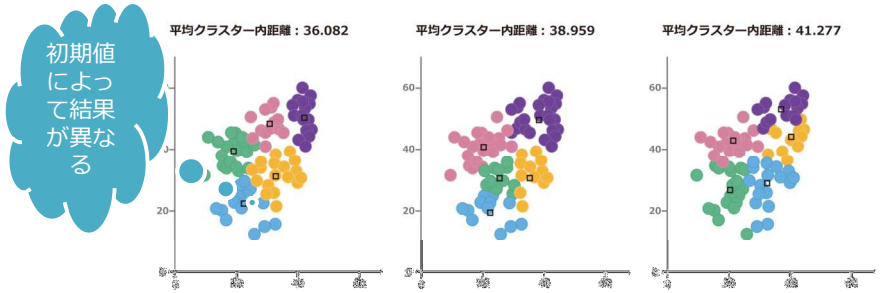
- サンプル間距離，グループ間距離の定義が複数

$$d(C_1,C_2)=\min_{\mathbf{x}_1\in C_1,\mathbf{x}_2\in C_2} d(\mathbf{x}_1,\mathbf{x}_2) \qquad d(C_1,C_2)=\frac{1}{|C_1||C_2|}\sum_{x_1\in C_1}\sum_{x_2\in C_2} d(\mathbf{x}_1,\mathbf{x}_2)$$

$$d(C_1,C_2)=\max_{\mathbf{x}_1\in C_1,\mathbf{x}_2\in C_2} d(\mathbf{x}_1,\mathbf{x}_2) \qquad d(C_1,C_2)=E[C_1\cup C_2]-E[C_1]-E[C_2]$$

技術的なおさらい２－２：クラスター分析

- 非階層的方法：k-means法／：適当に選んだ代表点（k個）と再近隣関係にある標本を併合する方法
- k，つまりグループ数を分析前に指定する方法
- もちろん，初期値依存がある．

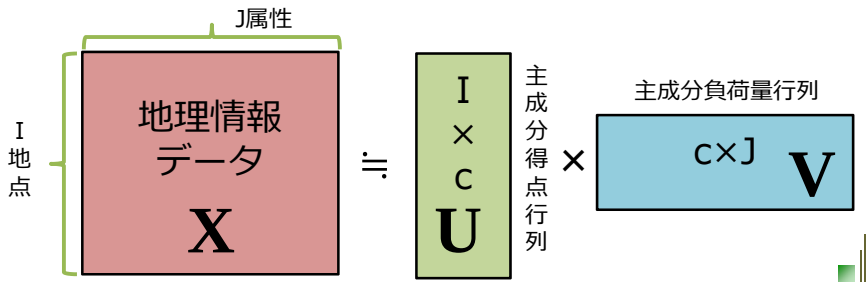


- 限界：各グループのサンプルは，超球内に均等分布する前提あり．

N工営報告書のアルゴリズム１

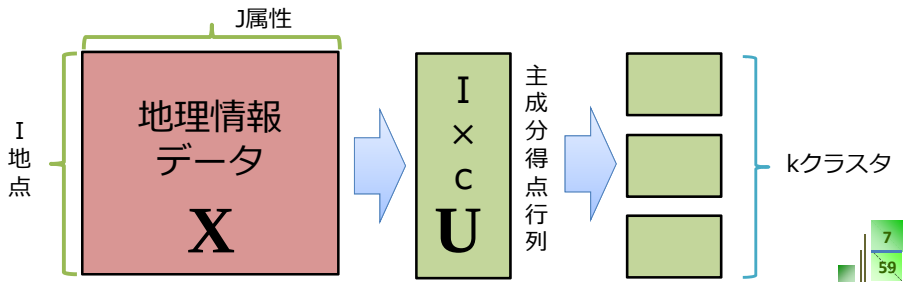
1. 対象地域の地理情報データを入手
2. 主成分分析から主成分軸を算出：c<J
3. サンプルの主成分負荷量，主成分得点行列を作成

$$\mathbf{X} \fallingdotseq \mathbf{UV}$$



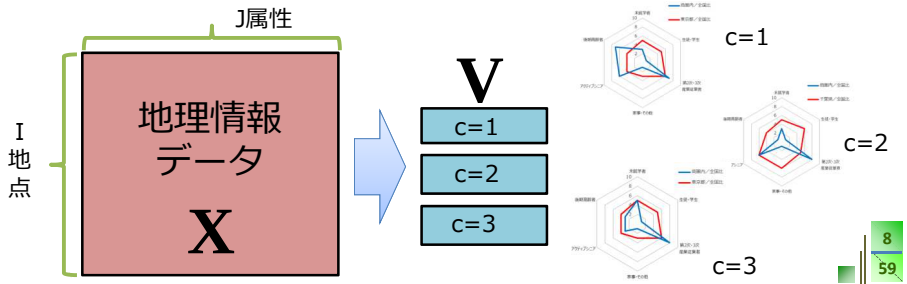
N工営報告書のアルゴリズム２

1. 対象地域の地理情報データを入手
2. 主成分分析から主成分軸を算出：c<J
3. サンプルの主成分負荷量，主成分得点行列を作成
4. 主成分得点行列から，クラスター分析でサンプルをグループ分け＝クラスターを得る：k<I

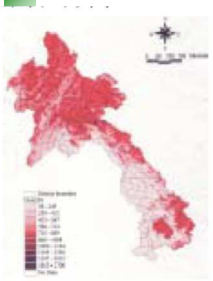


N工営報告書のアルゴリズム３

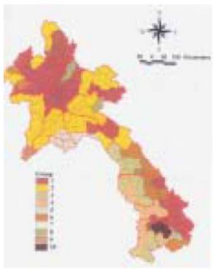
1. 対象地域の地理情報データを入手
2. 主成分分析から主成分軸を算出：c<J
3. サンプルの主成分負荷量，主成分得点行列を作成
4. 主成分得点行列から，クラスター分析でサンプルをグループ分け＝クラスターを得る．
5. 主成分負荷量行列を主成分ごとに集計して，レーダーチャートに．



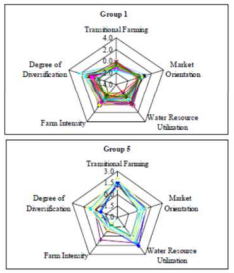
以上をまとめるとこのような結果に



標高データ（DEM）や森林被服図などラオス政府所有の既存の地理情報システム（GIS）グリッドデータ等をフルに活用し、全国の総合的な農業現況の把握に役立てた。



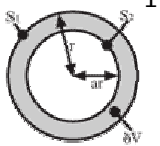
地理情報システム（GIS）の応用解析および、主成分分析、クラスター分析により、ラオス全国 141 の地域を農業開発の現況に基づき 10 のグループに分割。



抽出された農業開発に関わる5つの主成分に関して、レーダー・チャートを作成。レーダー・チャートにより、各グループの特性を把握。農業開発阻害要因と開発に必要な重点項目を特定。

数学的な問題

- 主成分分析は結局固有値問題：X'Xの固有値を求める問題は，属性次元Jが大きくなっても，高速で解ける方法がある．
- 一方，クラスタ分析には2種類の呪いがかかる．

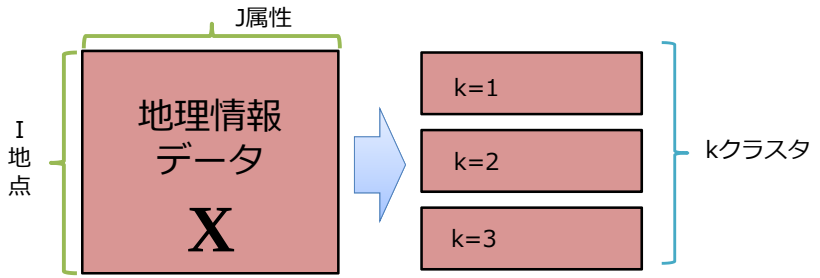


球面集中現象

1. 属性次元の呪い：半径がそれぞれ r と ar (0<a<1) の J 次元超球 S_1 と S_2 を考える． S_1 の体積 V に対する，二つの超球の体積の差 δV (図のグレー部分) の比は $\delta V/V=1-a^J$ ．サンプルが超球内に均一分布しているとすると， V内のサンプル数は超球の体積に比例．また，δV/V は Jの増加にともない1に近づく．つまり，J が大きな場合は S_1 中の対象は，ほとんど二つの超球の隙間に存在する．これは，中心のサンプルから他のサンプルまでの距離が次元Jの増加に従って急速に大きくなることを表し，どのサンプル間の類似度も低くなることを意味する．クラスタリングは，似ている対象をまとめる操作なので，有意な解が得られなくなる．
2. サンプル数の呪い：距離計算はサンプル数の二乗で増加．

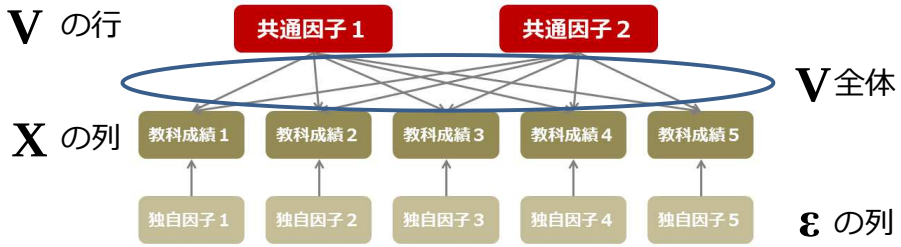
このアルゴリズムの疑問点

- アルゴリズムの概略：１）属性の主成分軸を求め，２）クラスターを主成分行列から求める
- kとcの決定に関する手掛かりがない（分析者の経験値／主観あるのみ）．
- kとcの役割の違いが不明．結構似た意味かも…
- レーダーチャートは，Xをクラスタ分けした部分行列からも書ける．どちらが正しい？



代替法１：因子分析１

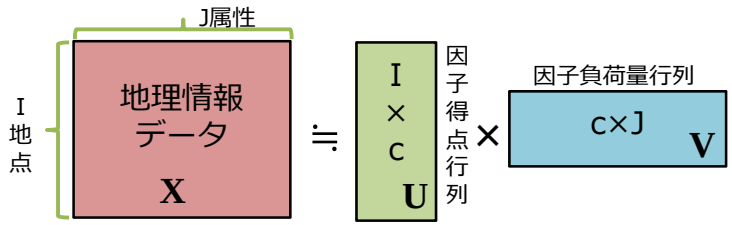
$$\mathbf{X} = \mathbf{UV} + \boldsymbol{\varepsilon}$$



よく見る上の図に騙されそうになるが，
きちんと理解すると…

代替法１：因子分析２

$$\mathbf{X} = \mathbf{UV} + \boldsymbol{\varepsilon}$$

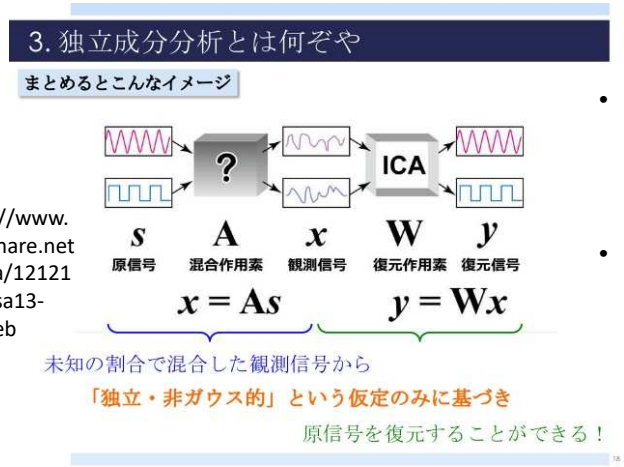


- この問題もやはり固有値問題．
- 古典的には，まずVを推定．
- Vが解釈できるように因子回転をかけて，V′を得る
- 回転後のV′を所与としてUを推定．

＋ 誤差行列（独自因子）

代替法２：独立成分分析（ICA） 1

$$\mathbf{X}'(t) \fallingdotseq \mathbf{V}'\mathbf{U}'(t) \quad \Longleftrightarrow \quad \mathbf{X} \fallingdotseq \mathbf{UV}$$



https://www.slideshare.net/zansa/121218-zansa13-for-web

- よく見る説明は音源分離（Blind Source Separation）
- 記号が違う＆行列が転置されているので混乱するが，整理するとわかる．

代替法２：独立成分分析（ICA） 2

$$\mathbf{X}'(t) \fallingdotseq \mathbf{V}'\mathbf{U}'(t) \quad \Longleftrightarrow \quad \mathbf{X} \fallingdotseq \mathbf{UV}$$

- 最尤法の導出
  - 観測信号 $\mathcal{X} = \{\mathbf{x}(1), \ldots, \mathbf{x}(T)\}$ に対する $\mathbf{W}$ の尤度
    $$\mathcal{L}(\mathbf{W}|\mathcal{X}) = \prod_{t=1}^{T} p(\mathbf{x}(t)|\mathbf{W})$$
  - 線形変換と確率密度関数
    $$\mathbf{y} = \mathbf{W}\mathbf{x} \qquad p(\mathbf{y}) = \frac{1}{|\det \mathbf{W}|} p(\mathbf{x})$$
  - 分離信号 $\mathbf{y}$ の独立性を仮定
    $$p(\mathbf{y}) = \prod_{i=1}^{I} p(y_i) \qquad p(y_i) \text{ はラプラス分布など}$$
  - 以上から導かれる対数尤度を最大化する $\mathbf{W}$ を求める
    $$\log \mathcal{L} = T \log |\det \mathbf{W}| + \sum_{t=1}^{T} \sum_{i=1}^{I} \log p(y_i(t))$$

澤田宏：独立成分分析入門，～音の分離を題材として～，部分空間法研究会チュートリアル，2010

代替法２：独立成分分析（ICA） 3

$$\mathbf{X}'(t) \fallingdotseq \mathbf{V}'\mathbf{U}'(t) \quad \Longleftrightarrow \quad \mathbf{X} \fallingdotseq \mathbf{UV}$$

- 最尤法の導出
  - サンプル数 T で割り，最大化すべき目的関数を設定

$$\mathcal{J} = \log|\det \mathbf{W}| + \textstyle\sum_{i=1}^{I} \mathrm{E}\{\log p(y_i)\}$$

» 参考: **W** をユニタリ行列に限定すればFastICAと等価

  - 勾配法により **W** を最適化

$$\mathbf{W} \leftarrow \mathbf{W} + \eta \cdot \frac{\partial \mathcal{J}}{\partial \mathbf{W}}$$

η は適切に設定されたステップサイズ

$$\frac{\partial \mathcal{J}}{\partial \mathbf{W}} = (\mathbf{W}^{\mathrm{T}})^{-1} - \mathrm{E}\{\boldsymbol{\Phi}(\mathbf{y})\mathbf{x}^{\mathrm{T}}\}$$

$$\boldsymbol{\Phi}(\mathbf{y}) = \begin{bmatrix} \phi(y_1) \\ \vdots \\ \phi(y_I) \end{bmatrix}$$

$$\phi(y_i) = -\frac{\partial \log p(y_i)}{\partial y_i}$$

$\phi(y_i)$ の具体的な形

ラプラス分布 $\phi(y_i) = \dfrac{y_i}{|y_i|}$

非線形関数G $\phi(y_i) = \dfrac{y_i}{\sqrt{y_i^2 + \alpha}}$

澤田宏：独立成分分析入門，～音の分離を題材として～，部分空間法研究会チュートリアル，2010

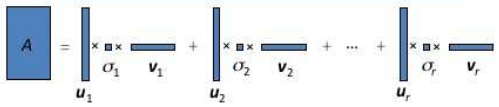
代替法３：行列の特異値分解

- 実数行列に関する特異値分解定理
  - 任意の $m \times n$ の行列Aは次の形に分解できる．

$$\mathbf{A} = \mathbf{U}\boldsymbol{\Sigma}\mathbf{V}^{\mathrm{T}}$$

    - ここで，
      - $\mathbf{U}$： $m \times m$ の直交行列
      - $\mathbf{V}$： $n \times n$ の直交行列
      - $\boldsymbol{\Sigma}$： $m \times n$ の対角行列 $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots, \sigma_n)$
        （ただし $\sigma_1 \geqq \sigma_2 \geqq \ldots \geqq \sigma_r > \sigma_{r+1} = \ldots = \sigma_n = 0$）
  - これらを使うと，A は以下のように表すことが出来る．

$$\mathbf{A} = \sigma_1 \mathbf{u}_1 \mathbf{v}_1^{T} + \sigma_2 \mathbf{u}_2 \mathbf{v}_2^{T} + \ldots + \sigma_r \mathbf{u}_r \mathbf{v}_r^{T}$$



https://www.slideshare.net/HAL9801/20141204journal-club

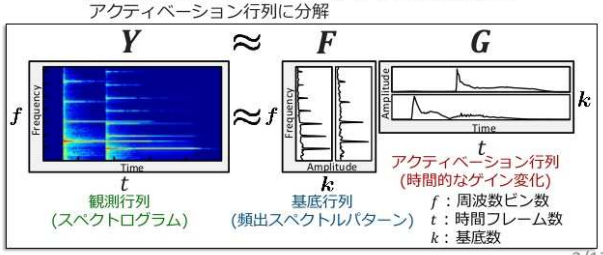
HIROSH

代替法４：非負値行列因子分解（NMF） 1

$$\mathbf{X}' \fallingdotseq \mathbf{V}'\mathbf{U}' \quad \Longleftrightarrow \quad \mathbf{X} \fallingdotseq \mathbf{UV}$$

非負値行列因子分解（NMF）

- NMF [Lee & Seung, 1999]
  - 非負値行列を非負値行列の積に低ランク近似
  - 画像処理、自動採譜など応用先は様々
  - 音源分離の場合，音源のスペクトログラムを基底行列とアクティベーション行列に分解



2/17

- ここでも音源分離が例に
- 記号が違う＆行列が転置されているので混乱するが，整理するとわかる

https://www.slideshare.net/DaichiKitamura/ss-66384298

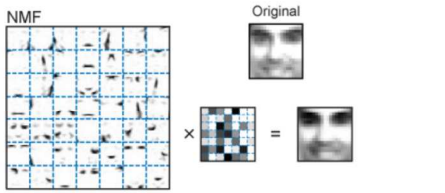
代替法４：非負値行列因子分解（NMF） 2

NMFの生まれた背景

- 画像処理分野で生まれた技術 [Lee1999]
  - 顔画像から顔パーツを抽出するのが目的
  - 概念自体は90年代前半に登場 [Paatero1994]

VQ: vector quantization

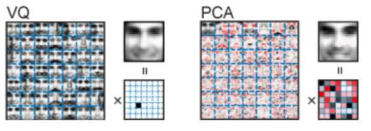
- 音のスペクトルを画像と見なして適用（後述）
  → 音声分離・自動採譜等
  [Smaragdis2003]以降極めて多数...



- 効率的な反復アルゴリズム [Lee2000]

$$H_{\omega,k} \leftarrow H_{\omega,k} \frac{[\boldsymbol{Y}\boldsymbol{U}^{\mathrm{T}}]_{\omega,k}}{[\boldsymbol{H}\boldsymbol{U}\boldsymbol{U}^{\mathrm{T}}]_{\omega,k}}$$

$$U_{k,t} \leftarrow U_{k,t} \frac{[\boldsymbol{H}^{\mathrm{T}}\boldsymbol{Y}]_{k,t}}{[\boldsymbol{H}^{\mathrm{T}}\boldsymbol{H}\boldsymbol{U}]_{k,t}}$$



亀岡弘和：非負値行列因子分解―チュートリアル，音楽情報科学研究会資料

代替法４：非負値行列因子分解（NMF） 3

NMFの定式化

- T 個の観測データ(非負値ベクトル) $\boldsymbol{y}_1, \ldots, \boldsymbol{y}_T$
- K 個の非負値基底ベクトル $\boldsymbol{h}_1, \ldots, \boldsymbol{h}_K$ の非負結合で
  どの観測データも良く表現できる基底セットを求めたい

$$\boldsymbol{y}_t \simeq \sum_{k=1}^{K} \boldsymbol{h}_k U_{k,t}, \ U_{k,t} \geq 0 \quad (t = 1, \ldots, T)$$

$$\left.\begin{array}{l} \boldsymbol{Y} := (\boldsymbol{y}_1, \cdots, \boldsymbol{y}_T) \\ \boldsymbol{H} := (\boldsymbol{h}_1, \cdots, \boldsymbol{h}_K) \\ \boldsymbol{U} := (U_{k,t})_{K \times T} \end{array}\right\} \ \boldsymbol{Y} \simeq \boldsymbol{HU} \quad H_{\omega,k} \geq 0, \ U_{k,t} \geq 0$$

$\boldsymbol{Y}$ と $\boldsymbol{HU}$ の乖離度を表す規準

minimize	$\mathcal{D}(\boldsymbol{Y}\|\boldsymbol{HU})$
subject to	$H_{\omega,k} \geq 0, \ U_{k,t} \geq 0$

代替法４：非負値行列因子分解（NMF） 4

非負値であることのメリット

- データ行列の非負性
  - 実世界には非負値データが多い
    （例）パワースペクトル，画素値，度数，

OD表も，モバ空も

- 基底行列の非負性
  - 「非負値データの構成要素もまた非負値データであるべき(でないと物理的に意味をなさない！)」という考え方
    （例）負の値をもったパワースペクトルなんて解釈のしようがない

$\mathbf{X}' = \mathbf{V}'\mathbf{U}'$

- 基底とは，V'の列＝Vの行のこと

- 係数行列の非負性
  - 構成要素の混ざり方は「足し算」のみ
  - 係数行列をスパースに誘導 → 基底の情報量をアップ

代替法４：非負値行列因子分解（NMF） 5

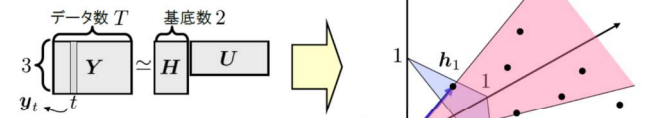
非負値で基底を張るとは…？

$\mathbf{X}' = \mathbf{V}'\mathbf{U}'$

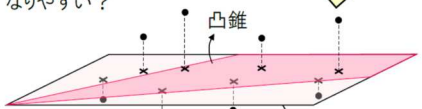
- Frobeniusノルム規準のNMF

minimize $\|\boldsymbol{Y} - \boldsymbol{HU}\|_F^2 = \sum_t \left\| \boldsymbol{y}_t - \sum_k \boldsymbol{h}_k U_{k,t} \right\|_2^2$

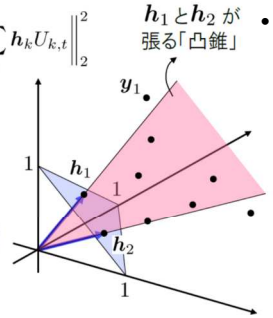
subject to $H_{\omega,k} \geq 0, \ U_{k,t} \geq 0$



斜交基底だから係数はスパースになりやすい？



$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ が張る部分空間



$\boldsymbol{h}_1$ と $\boldsymbol{h}_2$ のなす角を大きくした方がお得！

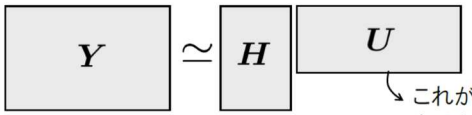
陽に制約を入れていないのに基底が直交化される傾向に

- 基底とは，V'の列＝Vの行のこと

代替法４：非負値行列因子分解（NMF） 6

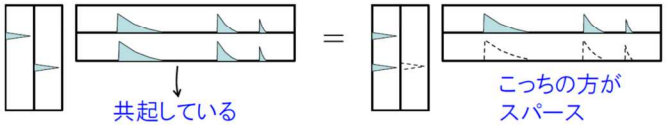
負荷量行列がスパースとはどんな状態？

$\mathbf{X}' = \mathbf{V}'\mathbf{U}'$



これがスパースになるよう誘導される

- Hがどうなればスパースになる？
  - 共起するパーツがあればひとまとめにした方がスパースに！

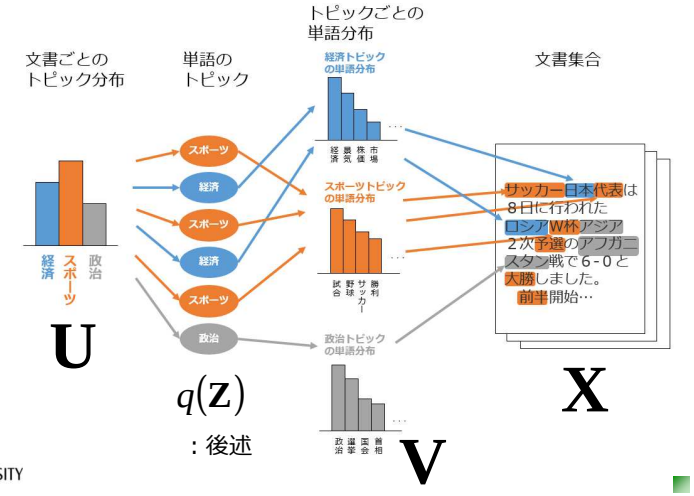


  - 共起頻度が高いパーツのペアをひとまとめにした方がスパースに！
    （上のひとまとめ操作では を3つ分減らせている）

頻出するひとまとまりのパーツが
各基底ベクトルとして得られる傾向になる

トピックモデルへ

$$\mathbf{X} \fallingdotseq \mathbf{UV}$$



$\mathbf{U}$

$q(\mathbf{z})$

：後述

$\mathbf{X}$

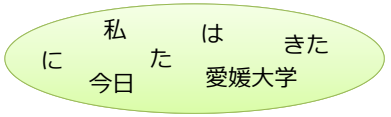
$\mathbf{V}$

文書データの作成 1

- One-hot表現の作成

「今日私は愛媛大学にきた」

形態素解析



単語順序

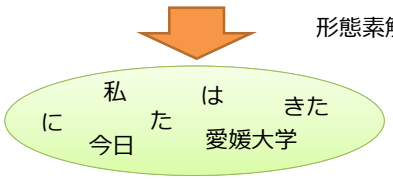
	昨日	今日	明日	私	君	は	が	松山	愛媛大学	に	くる	帰る
1		1										
2				1								
3						1						
4									1			
5										1		
6											1	

文書データの作成 2

- 単語順序のOne-hot表現を文書単位で集計すると，Bag-of-Wordsに

「今日私は愛媛大学にきた」

形態素解析



文書

	昨日	今日	明日	私	君	は	が	松山	愛媛大学	に	くる	帰る
A		1		1		1			1	1	1	

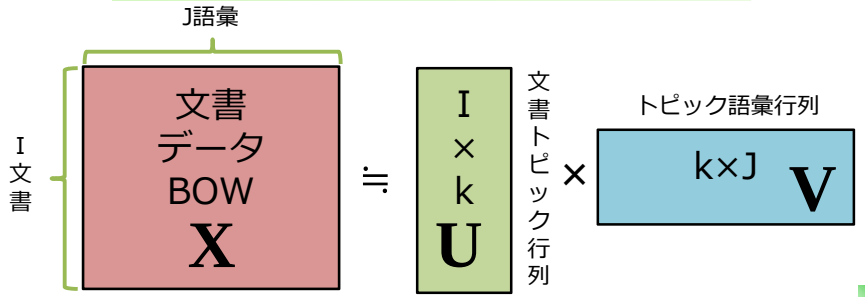
全文書に現れる全語彙（＝属性）

- Bag-of-Wordsでは，文書内の語の順序は無視する．
- Bag-of-Wordsの全語彙集合からは，低頻度語を除く処理を加えることがある．

行列分解法としてのトピックモデル

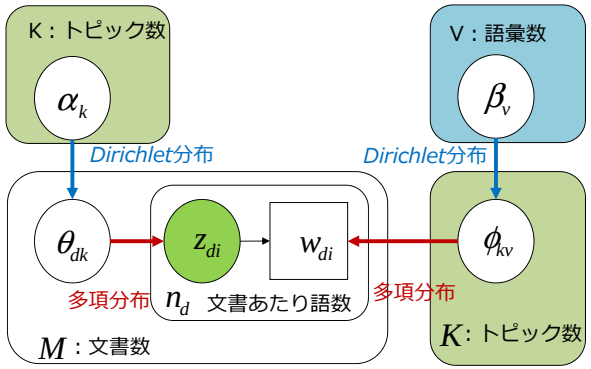
トピックモデルのはたらき

J個の属性をJより少ないk個のトピックにまとめる
階層ベイズモデル：斜交成分の抽出
確率構造の明示：尤度基準によるモデル探索Vからトピックを命名
Uから文書ごとのトピックを推定



トピックモデルの概要

- 観測された語：w_{di}
- 語とトピックの関係：z_{di}
- トピックと文書の関係：θ_{dk}
- トピックの分布（集中度）：α_k
- 語彙とトピックの関係：Φ_{kv}
- 語彙の分布（集中度）：β_v



トピックモデルの定式化

多項分布：

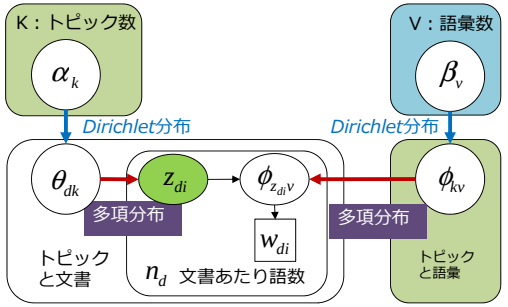
$$p(\{n_k\}_{k=1}^K|\boldsymbol{\pi}) = \frac{n!}{\prod_{k=1}^K n_k!}\prod_{k=1}^K \pi_k^{n_k}$$

Dirichlet分布：

$$p(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)}\prod_{k=1}^K \pi_k^{\alpha_k - 1}$$

$$\begin{aligned} z_{di} &\sim \mathrm{Multi}(\theta_d),\ k=1,\ldots,K \\ w_{di} &\sim \mathrm{Multi}(\phi_{z\_di,v}),\ v=1,\ldots,V \\ \theta_{dk} &\sim \mathrm{Dir}(\boldsymbol{\alpha}),\ k=1,\ldots,K \\ \phi_{kv} &\sim \mathrm{Dir}(\boldsymbol{\beta}),\ v=1,\ldots,V \end{aligned}$$

$$\begin{aligned} &p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi}|\boldsymbol{\alpha},\boldsymbol{\beta}) \\ &= p(\mathbf{w}|\mathbf{z},\boldsymbol{\varphi})p(\mathbf{z}|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha})p(\boldsymbol{\varphi}|\boldsymbol{\beta}) \end{aligned}$$



同時分布の因数分解近似

HIROSHIMA UNIVERSITY

周辺化変分ベイズ法：CVB0の導出

周辺化変分下限の導出

$$p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{\mathbf{z}} p(\mathbf{w},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})$$ 潜在変数zの導入

$$= \int \sum_{\mathbf{z}} p(\mathbf{w},\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi}|\boldsymbol{\alpha},\boldsymbol{\beta})\,d\boldsymbol{\theta}\,d\boldsymbol{\phi}$$

$$= \int \sum_{\mathbf{z}} p(\mathbf{w}|\boldsymbol{\varphi},\mathbf{z})\,p(\mathbf{z}|\boldsymbol{\theta})\,p(\boldsymbol{\theta},\boldsymbol{\varphi}|\boldsymbol{\alpha},\boldsymbol{\beta})\,d\boldsymbol{\theta}\,d\boldsymbol{\phi}$$

$\log p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\beta})$ 事後分布pと近似事後分布qの距離KLの最小化

$$= F[q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi})] + KL\left[q(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi}) \,\|\, p(\mathbf{z},\boldsymbol{\theta},\boldsymbol{\varphi}|\mathbf{w},\boldsymbol{\alpha},\boldsymbol{\beta})\right]$$

⇔変分下限Fの最大化

$$\log p(\mathbf{w}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \sum_{\mathbf{z}} \log p(\mathbf{w},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})$$

周辺化によるθ，φの消去（qの因数分解近似より）

$$= \sum_{\mathbf{z}} \log q(\mathbf{z}) \frac{p(\mathbf{w},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})}{q(\mathbf{z})}$$

$$\geq \sum_{\mathbf{z}} q(\mathbf{z}) \log \frac{p(\mathbf{w},\mathbf{z}|\boldsymbol{\alpha},\boldsymbol{\beta})}{q(\mathbf{z})}$$

$$= F_{CVB}[q(\mathbf{z})]$$

周辺化変分下限更新式の導出

$\mathbf{z} = (z_{di}, \mathbf{z}^{\backslash di})$ 逐次更新のための分解

$$\begin{aligned} &q^*(z_{di}=k \,\&\, w_{di}=v) \\ &\propto \exp\left(\begin{array}{l}\sum_{\mathbf{z}} q(\mathbf{z}^{\backslash di}=k) \\ \times \log p(z_{di}=k, w_{di}=v|\mathbf{w}^{\backslash di},\mathbf{z}^{\backslash di},\boldsymbol{\alpha},\boldsymbol{\beta})\end{array}\right) \\ &= \exp E_{q(\mathbf{z}^{\backslash di})}\left[\log p(z_{di}=k, w_{di}=v|\mathbf{w}^{\backslash di},\mathbf{z}^{\backslash di},\boldsymbol{\alpha},\boldsymbol{\beta})\right] \\ &= \exp E_{q(\mathbf{z}^{\backslash di})}\left[\begin{array}{l}\log p(w_{di}=v|z_{di}=k,\mathbf{w}^{\backslash di},\mathbf{z}^{\backslash di},\boldsymbol{\alpha},\boldsymbol{\beta}) \\ +\log p(z_{di}=k|\mathbf{z}^{\backslash di},\boldsymbol{\alpha})\end{array}\right] \\ &= \exp E_{q(\mathbf{z}^{\backslash di})}\left[\log \frac{n_{kv}^{\backslash di}+\beta_v}{n_k^{\backslash di}+\sum_v \beta_v} + \log \frac{n_{dk}^{\backslash di}+\alpha_k}{n_d^{\backslash di}+\sum_k \alpha_k}\right] \\ &\propto \frac{\exp E_{q(\mathbf{z}^{\backslash di})}\left[\log\left(n_{kv}^{\backslash di}+\beta_v\right)\right]}{\exp E_{q(\mathbf{z}^{\backslash di})}\left[\log\left(n_k^{\backslash di}+\sum_v \beta_v\right)\right]} \times \exp E_{q(\mathbf{z}^{\backslash di})}\left[\log\left(n_{dk}^{\backslash di}+\alpha_k\right)\right] \\ &\approx \frac{E[n_{kv}^{\backslash di}]+\beta_v}{E[n_k^{\backslash di}]+\sum_v \beta_v} E[n_{dk}^{\backslash di}]+\alpha_k \end{aligned}$$

テイラー展開＆0次近似

計算法：Collapsed Variational Bayes 0

- 周辺化変分ベイズ法：決定論アルゴリズム
  MCMC法と異なり，初期乱数の後は不動点探索を行う

0-1. Give the number of **K**　　0-3. Randomly Initialize : $\mathbf{q}(z_{di})$
0-2. Initialize parameters : $\boldsymbol{\alpha},\boldsymbol{\beta}$　　0-4. Calculate $E[N_{kv}], E[N_{dk}]$ by $\mathbf{q}(z_{di})$

1. $\mathbf{q}(z_{di}) = f(E[N_{kv}], E[N_{dk}], \boldsymbol{\alpha},\boldsymbol{\beta})$
2. $E[N_{kv}]$, $E[N_{dk}]$ Update by $\mathbf{q}(z_{di})$
3. $\boldsymbol{\alpha},\boldsymbol{\beta}$ Update by $E[N_{kv}]$, $E[N_{dk}]$

4. Check the convergence in α, β

$$\alpha_k^{new} = \frac{\sum_d [\Psi(E[n_{dk}]+\alpha_k) - \Psi(\alpha_k)]}{\sum_d \left[\Psi\left(n_d + \sum_k \alpha_k\right) - \Psi\left(\sum_k \alpha_k\right)\right]} \cdot \alpha_k$$

$$\beta_v^{new} = \frac{\sum_k \Psi(E[n_{kv}]+\beta_v) - \Psi(\beta_v)}{\sum_k \left[\Psi\left(\sum_v E[n_{kv}]+\beta_v\right) - \Psi\left(\sum_v \beta_v\right)\right]} \cdot \beta_v$$

Converged

V $\boldsymbol{\varphi}_{kv} = \frac{E[N_{kv}]}{N_k}$ $E[N_{kv}]$ トピック別の語彙分布

U $\boldsymbol{\theta}_{dk} = \frac{E[N_{dk}]}{N_d}$ $E[N_{dk}]$ 文書別のトピック分布

CVB0アルゴリズムの入出力１

語彙数：V

$$D(\boldsymbol{x}_d)=\begin{pmatrix} x_{11} & \cdots & x_{1V} \\ \vdots & x_{dv} & \vdots \\ x_{D1} & \cdots & x_{DV} \end{pmatrix}$$

文書数・D

文書ごとに集計

$\boldsymbol{x}_d(x_{d1}, x_{d2}, \ldots, x_{dv})$

BOW : Bag of words表現

文書ごとに語彙をカウント

$$\boldsymbol{w}_{di}=\begin{pmatrix} w_{11} \\ \vdots \\ w_{di} \\ \vdots \\ w_{DN} \end{pmatrix}$$

総語数・N

$w_{di}=v_l \quad l=1,\ldots,V$

トピック潜在変数を導入

$$\boldsymbol{z}_{di}=\begin{pmatrix} z_{11} \\ \vdots \\ z_{di} \\ \vdots \\ z_{DN} \end{pmatrix}$$

総語数・N

$z_{di}=k \quad k=1,\ldots,K$

- それぞれの語とトピックの関係を指し示す，潜在変数Ｚを導入．

CVB0アルゴリズムの入出力２

- 語とトピックの関係を示す潜在変数Ｚを，トピック帰属確率に分解する．

$$\mathbf{q}(z_{di})=\begin{pmatrix} q_{z11} & \cdots & q_{z1K} \\ \vdots & q_{zik} & \vdots \\ q_{zN1} & \cdots & q_{zNK} \end{pmatrix}$$

総語数・N

トピック数：K

$$\mathbf{q}(z_{di})=\sum_{k=1}^{K} q_k(z_{di})=1 \quad \text{for all } z_{di}$$

- それぞれの語とトピックの関係を指し示す，潜在変数Ｚを導入．

$$\boldsymbol{z}_{di}=\begin{pmatrix} z_{11} \\ \vdots \\ z_{di} \\ \vdots \\ z_{DN} \end{pmatrix}$$

総語数・N

$z_{di}=k \quad k=1,\ldots,K$

HIROSHIMA UNIVERSITY

CVB0アルゴリズムの入出力３

- 推定パラメータが，語数に関する加法条件を満たしている

N ：全文書の総語数　N_d：文書d（1 to D）の語数

N_v：語彙v（1 to V）の語数

V ：全文書の語彙数　V_k：トピックk（1 to K）の語彙数

N_{dk}：文書dのトピックkの語数　N_{kv}：トピックkの語彙vの語数

$$\sum_{d=1}^{D} N_d = N \qquad \sum_{k=1}^{K} N_{dk} = N_d \ \text{for all } d \qquad \sum_{k=1}^{K} N_{kv} = N_v \ \text{for all } v$$

$$\sum_{v=1}^{V} N_v = N \qquad \sum_{v=1}^{V} N_{kv} = \sum_{d=1}^{D} N_{dk} = N_k \ \text{for all } k$$

N_k：トピックｋの語数＝トピックkの大きさ

HIROSHIMA UNIVERSITY

各手法の特徴まとめ

	データ	基底順序	基底回転	基底直斜	基底数探索	解の正値性	係数の疎性	解の唯一性	解の初期値依存	興味は？
PCA 主成分分析	連続	有	無	直交	累積寄与	無	無	有	無	$V_?$
FA 因子分析	連続	無	有	直交→斜交	累積寄与？	無	無	無	無	
ICA 独立成分分析	連続	無	無	斜交	尤度	無	無	無	無	$U_?$
SVD 行列特異値分解	連続	有	無	直交	×	無	無	無	無	V, U
NMF 非負値行列因子分解	連続	無	無	斜交	残差基準	有	有	無	有	
トピックモデル	離散	無	無	斜交	尤度	有	有	無	有	

HIROSHIMA UNIVERSITY

地理情報へのトピックモデルの適用

GISの進歩に伴い，
地理情報データ整備が進む．

↓

土地利用や人口動態など，
多種類の情報が詳細な地点単位で入手できる．

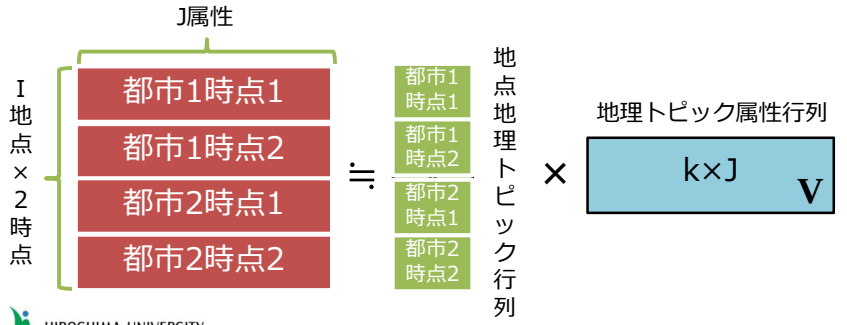
↓

そのままでは扱いづらい，膨大な情報量の適切な圧縮手法が求められるように．

出典) esriジャパン, https://www.esrij.com/gis-guide/gis-datamodel/gis-datamodel/

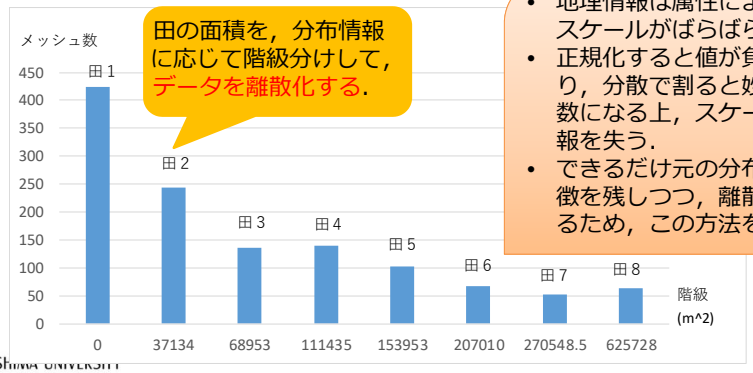
入力データの構成

- 2都市2時点のデータをプール
- 共通する＝都市や時点によって変わらない地理トピック（V）を抽出
- トピック重み（U）の違いが，年次別の土地利用特性の違いに現れるようにデータ作成



データ加工：離散化

- 属性ごとに8段階に自然階級分類によって階級分け．
- 階級別に該当/非該当をダミー変数で与える．
- 下図において田1とは，1メッシュ内の田の面積がもっとも小さい階級を示す．
- データ欠損や秘匿のメッシュは，NA階級．



田の面積を，分布情報に応じて階級分けして，データを離散化する．

- 地理情報は属性によってスケールがばらばら．
- 正規化すると値が負になり，分散で割ると妙な実数になる上，スケール情報を失う．
- できるだけ元の分布の特徴を残しつつ，離散化するため，この方法を採用．

データの概要

- 対象地域：福岡都市圏，北九州都市圏
- データソース：国勢調査地域メッシュ統計，経済センサス，事業所企業統計調査，国土数値情報土地利用
- メッシュの種類：3次メッシュ
- メッシュ数：2113×2時点＝4226メッシュ
- 属性数：36変数×9階級→324属性：データの離散化

土地	田　他農用地
世帯	1人世帯数，2人世帯数，3人世帯数．4人世帯数，5人以上世帯数，1戸建て世帯数．共同住宅世帯数，核家族世帯数
就業者	第１次産業就業者総数　第２次産業就業者総数　第３次産業就業者総数
事業所・従従業者	第2次産業事業所数，第2次産業従業者数，卸売・小売業事業所数　金融・保険業事業所数　不動産・物品賃貸業事業所数　宿泊業・飲食サービス業事務所数，教育・学習支援事業所数．医療・福祉事業所数，卸売・小売業従業者数，金融・保険従業syは数，不動産業従業者数，宿泊業・飲食サービス業従業者数，教育・学習支援従業者数．医療・福祉従業者数，公務事業所数，公務従業者数
人口	0～14歳人口総数．15～64歳人口総数，65歳以上人口総数，居住期間10年未満総数，居住期間出走時から総数

トピック数の決定方法

● コサイン類似度（V行列／U行列）

$$\cos(\mathbf{e}_i, \mathbf{e}_{i'}) = \frac{\mathbf{e}_i \cdot \mathbf{e}_{i'}}{|\mathbf{e}_i||\mathbf{e}_{i'}|}$$

- トピック比較の場合，i'は別のトピック・属性ベクトル
- 同地点の時点間比較の場合，i'は同地点異時点の地点・地理トピックベクトル

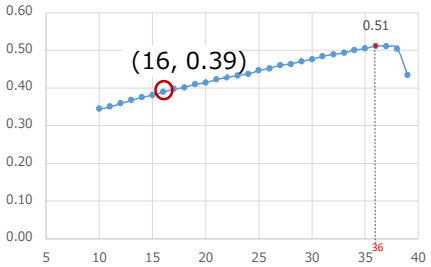
● 対数尤度（U・V行列）

$$\log L(u, v|x) = \sum_{d=1}^{D}\sum_{i=1}^{N_d}\sum_{k=1}^{K} \log u_{dk} v_{k,j=x}$$

トピック間類似度が低く，対数尤度の高いトピック数を選択

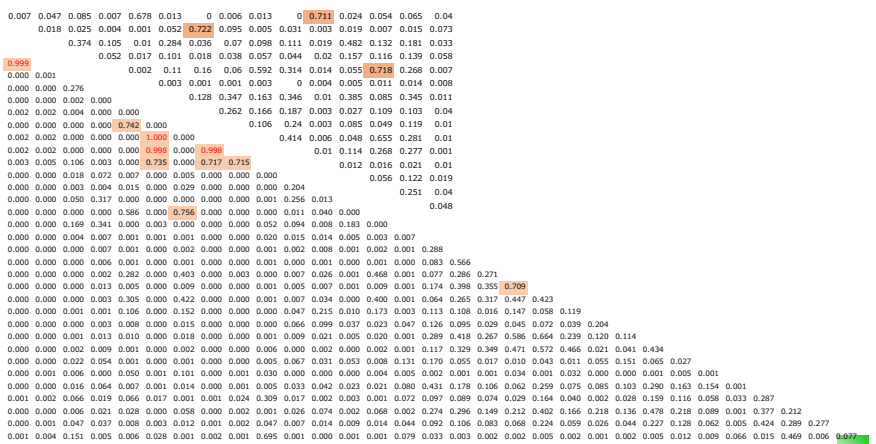
尤度・トピック類似度

- 尤度はトピック数36で極大
- k=36で，得られたトピック内容の重複を確認
- kを減らしながらトピック類似度を確認したところ，k=16で，トピック類似度が全て0.8を下回ることを確認→16トピックスを採用



トピック類似度：補足

- k=16のトピック類似度



- k=32のトピック類似度

V行列：トピックの命名 1

寄与の大きな属性の並びから，トピックを命名

高集積		準高集積	
他農用地0	0.030	金融業.保険業従業者数1	0.032
四人世帯数7	0.029	卸売業.小売業事務所数1	0.026
田0	0.027	卸売業.小売業従業者数2	0.025
一戸建世帯数7	0.026	不動産業事務所数2	0.025
第一次産業就業者総数7	0.024	第2次産業従業者数2	0.024
共同住宅世帯数7	0.023	医療.福祉事務所数0	0.023
６５歳以上人口総数7	0.021	公務従業者数0	0.023
二人世帯数7	0.021	三人世帯数5	0.023
核家族世帯数7	0.020	不動産業従業者数2	0.021
居住期間出生時から総数7	0.020	他農用地0	0.021
１５.６４歳人口総数7	0.019	第二次産業就業者総数7	0.020
三人世帯数7	0.018	１５.６４歳人口総数5	0.020
０.１４歳人口総数7	0.018	教育.学習支援業従業者数2	0.019
卸売業.小売業従業者数3	0.017	一戸建世帯数5	0.019
不動産業事務所数3	0.017	第一次産業就業者総数5	0.019
第2次産業従業者数3	0.016	五人以上世帯数4	0.018
三人世帯数6	0.015	宿泊業.飲食サービス業従業者数2	0.018
不動産業従業者数3	0.015	金融業.保険業事務所数2	0.018
１５.６４歳人口総数6	0.014	共同住宅世帯数5	0.018
五人以上世帯数7	0.014	教育.学習支援業事務所数3	0.017

- トピックごとに総重みが1になるように基準化
- 表は，20位まで or 重みが0.01以上のみ作成

V行列：トピックの命名２

居住（密度高）		居住（密度中）		居住（密度中低）	
五人以上世帯数3	0.036	第一次産業就業者総数3	0.039	１５.６４歳人口総数2	0.060
第一次産業就業者総数4	0.035	一戸建世帯数3	0.039	三人世帯数2	0.058
一戸建世帯数4	0.034	金融業,保険業事務所数1	0.036	一戸建世帯数2	0.055
居住期間１０年未満総数2	0.033	不動産業従業者数1	0.035	共同住宅世帯数2	0.053
共同住宅世帯数4	0.032	１５.６４歳人口総数3	0.033	第一次産業就業者総数2	0.053
１５.６４歳人口総数4	0.030	五人以上世帯数2	0.033	居住期間出生時から総数2	0.049
第三次産業就業者総数2	0.030	共同住宅世帯数3	0.032	居住期間１０年未満総数1	0.040
三人世帯数4	0.029	卸売業,小売業事務所数1	0.029	０.１４歳人口総数2	0.039
二人世帯数4	0.024	金融業,保険業従業者数1	0.028	第三次産業就業者総数1	0.037
居住期間出生時から総数4	0.023	第二次産業就業者総数4	0.028	６５歳以上人口総数2	0.037
第2次産業従業者数2	0.022	三人世帯数3	0.028	二人世帯数2	0.036
０.１４歳人口総数4	0.021	第三次産業就業者総数1	0.027	第二次産業就業者総数3	0.035
核家族世帯数4	0.021	居住期間出生時から総数3	0.026	核家族世帯数2	0.027
四人世帯数4	0.021	６５歳以上人口総数3	0.025	四人世帯数3	0.026
教育,学習支援業事務所数2	0.021	二人世帯数3	0.024	五人以上世帯数2	0.025
０.１４歳人口総数3	0.020	医療,福祉事務所数0	0.023	核家族世帯数3	0.024
第二次産業就業者総数5	0.019	公務従業者数0	0.023	不動産業従業者数1	0.021
宿泊業,飲食サービス業事務所数2	0.019	居住期間１０年未満総数1	0.023	四人世帯数2	0.021
四人世帯数5	0.018	核家族世帯数3	0.021	金融業,保険業事務所数1	0.020
他農用地0	0.018	教育,学習支援業従業者数1	0.021	五人以上世帯数1	0.017

V行列：トピックの命名３

低密度居住		低密度居住＋農業		低密度居住+低密度商業	
第二次産業就業者総数1	0.071	０.１４歳人口総数1	0.060	教育,学習支援業事務所数1	0.099
核家族世帯数1	0.067	五人以上世帯数1	0.058	宿泊業,飲食サービス業事務所数1	0.088
６５歳以上人口総数1	0.065	居住期間出生時から総数1	0.055	医療,福祉従業者数1	0.087
１５.６４歳人口総数1	0.065	共同住宅世帯数1	0.054	教育,学習支援業従業者数1	0.087
第一次産業就業者総数1	0.064	三人世帯数1	0.053	卸売業,小売業事務所数0	0.053
三人世帯数1	0.063	居住期間１０年未満総数1	0.049	第2次産業従業者数1	0.053
二人世帯数1	0.063	１５.６４歳人口総数1	0.048	金融業,保険業従業者数0	0.053
居住期間出生時から総数1	0.063	第一次産業就業者総数1	0.045	卸売業,小売業従業者数1	0.053
四人世帯数1	0.062	四人世帯数2	0.045	公務従業者数0	0.039
五人以上世帯数1	0.061	第二次産業就業者総数2	0.041	医療,福祉事務所数0	0.039
一戸建世帯数1	0.061	一戸建世帯数1	0.040	不動産業従業者数0	0.034
共同住宅世帯数1	0.059	核家族世帯数2	0.034	金融業,保険業事務所数0	0.034
０.１４歳人口総数1	0.058	第三次産業就業者総数0	0.027	第三次産業就業者総数1	0.030
居住期間１０年未満総数1	0.055	６５歳以上人口総数1	0.026	他農用地1	0.026
第三次産業就業者総数0	0.048	二人世帯数1	0.025	公務事務所数1	0.026
一人世帯数1	0.025	公務事務所数1	0.023	居住期間１０年未満総数1	0.025
一人世帯数2	0.014	田7	0.019	６５歳以上人口総数2	0.023
		一人世帯数4	0.018	第2次産業事務所数1	0.021
		二人世帯数2	0.017	一人世帯数3	0.015
		一人世帯数3	0.016		

V行列：トピックの命名４

低密度商業		低密度工業		低密度工業+農業	
宿泊業,飲食サービス業従業者数1	0.175	教育,学習支援業従業者数0	0.097	不動産業事務所数0	0.100
不動産業事務所数1	0.171	宿泊業,飲食サービス業事務所数0	0.097	宿泊業,飲食サービス業従業者数0	0.100
公務従業者数0	0.068	第2次産業従業者数1	0.091	公務事務所数1	0.072
医療,福祉事務所数0	0.067	卸売業,小売業従業者数1	0.086	第2次産業事務所数1	0.070
他農用地0	0.066	医療,福祉従業者数0	0.080	医療,福祉事務所数0	0.058
公務事務所数1	0.058	教育,学習支援業事務所数0	0.080	公務従業者数0	0.058
金融業,保険業従業者数0	0.054	不動産業従業者数0	0.066	卸売業,小売業事務所数0	0.055
卸売業,小売業事務所数0	0.053	金融業,保険業事務所数0	0.065	金融業,保険業従業者数0	0.055
田0	0.048	金融業,保険業従業者数0	0.059	金融業,保険業事務所数0	0.050
卸売業,小売業従業者数1	0.041	卸売業,小売業事務所数0	0.059	不動産業従業者数0	0.050
第2次産業従業者数1	0.034	医療,福祉事務所数0	0.056	教育,学習支援業従業者数0	0.045
不動産業従業者数1	0.029	公務従業者数0	0.056	宿泊業,飲食サービス業事務所数0	0.045
金融業,保険業事務所数1	0.029	第2次産業事務所数1	0.029	医療,福祉従業者数0	0.045
田1	0.024	田2	0.018	教育,学習支援業事務所数0	0.044
第2次産業事務所数2	0.022	他農用地0	0.018	第2次産業従業者数0	0.033
第三次産業就業者総数1	0.017			卸売業,小売業従業者数0	0.033
一人世帯数1	0.015			田1	0.017
				他農用地1	0.015
				田3	0.010
				他農用地0	0.010

V行列：トピックの命名５

未利用1		未利用2	
公務事務所数0	0.072	核家族世帯数0	0.071
第2次産業事務所数0	0.072	０.１４歳人口総数0	0.070
不動産業従業者数0	0.061	一人世帯数0	0.069
金融業,保険業事務所数0	0.061	共同住宅世帯数0	0.066
医療,福祉従業者数0	0.059	一戸建世帯数0	0.065
教育,学習支援業事務所数0	0.058	二人世帯数0	0.063
卸売業,小売業従業者数0	0.057	第三次産業就業者総数0	0.062
第2次産業従業者数0	0.057	四人世帯数0	0.058
金融業,保険業従業者数0	0.057	居住期間１０年未満総数0	0.057
卸売業,小売業事務所数0	0.056	五人以上世帯数0	0.055
公務従業者数0	0.055	居住期間出生時から総数0	0.054
医療,福祉事務所数0	0.055	第二次産業就業者総数0	0.053
教育,学習支援業従業者数0	0.051	第一次産業就業者総数0	0.052
宿泊業,飲食サービス業事務所数0	0.050	６５歳以上人口総数0	0.050
宿泊業,飲食サービス業従業者数0	0.041	三人世帯数0	0.049
田0	0.041	１５.６４歳人口総数0	0.045
不動産業事務所数0	0.040	他農用地0	0.032
他農用地0	0.033	田0	0.023

V行列：トピックの命名６

秘匿1		秘匿2		秘匿3	
公務従業者数NA	0.058	第2次産業事務所数NA	0.030	第三次産業就業者総数NA	0.061
医療.福祉従業者数NA	0.058	第2次産業従業者数NA	0.030	第二次産業就業者総数NA	0.061
教育.学習支援業従業者数NA	0.058	卸売業.小売業事務所数NA	0.030	第一次産業就業者総数NA	0.061
宿泊業.飲食サービス業従業者数NA	0.058	金融業.保険業事務所数NA	0.030	核家族世帯数NA	0.061
不動産業従業者数NA	0.058	不動産業事務所数NA	0.030	共同住宅世帯数NA	0.061
金融業.保険業従業者数NA	0.057	宿泊業.飲食サービス業事務所数NA	0.030	一戸建世帯数NA	0.061
卸売業.小売業従業者数NA	0.057	教育.学習支援業事務所数NA	0.030	居住期間出生時から総数NA	0.061
公務事務所数NA	0.057	医療.福祉事務所数NA	0.030	居住期間１０年未満総数NA	0.061
医療.福祉事務所数NA	0.057	公務事務所数NA	0.030	五人以上世帯数NA	0.061
教育.学習支援業事務所数NA	0.057	卸売業.小売業従業者数NA	0.030	四人世帯数NA	0.061
宿泊業.飲食サービス業事務所数NA	0.057	金融業.保険業従業者数NA	0.030	三人世帯数NA	0.061
不動産業事務所数NA	0.057	不動産業従業者数NA	0.030	二人世帯数NA	0.061
金融業.保険業事務所数NA	0.057	宿泊業.飲食サービス業従業者数NA	0.030	一人世帯数NA	0.061
卸売業.小売業事務所数NA	0.057	教育.学習支援業従業者数NA	0.030	６５歳以上人口総数NA	0.061
第2次産業従業者数NA	0.057	医療.福祉従業者数NA	0.030	１５.６４歳人口総数NA	0.061
第2次産業事務所数NA	0.057	公務従業者数NA	0.030	０.１４歳人口総数NA	0.061
田1	0.021	他農用地0	0.030		
他農用地1	0.015	田0	0.029		
		０.１４歳人口総数NA	0.029		
		１５.６４歳人口総数NA	0.029		

土地利用数値化指標1

● 優勢トピック（U行列）

$$\hat{k}_i = \left\{ k \,\middle|\, \begin{array}{l} k\text{は}, \max(u_{i1}, ..., u_{ik}.., u_{iK}) \\ \text{を満たすトピック} \end{array} \right\}$$

- トピックモデルは，本来ソフトクラスタリングモデル＝どれか一つのトピックに分類されない．
- 地図を描くための方便

● 都市化推移指標（U行列）

$$C_i = \sum_{k=1}^{K} w_k \left(u_{ik} - u_{i'k} \right)$$

- i'は同地点異時点の地点・地理トピックベクトル
- w_kは，集積程度に応じたトピックkの重み：高集積＝大きな値→低集積＝小さな値．

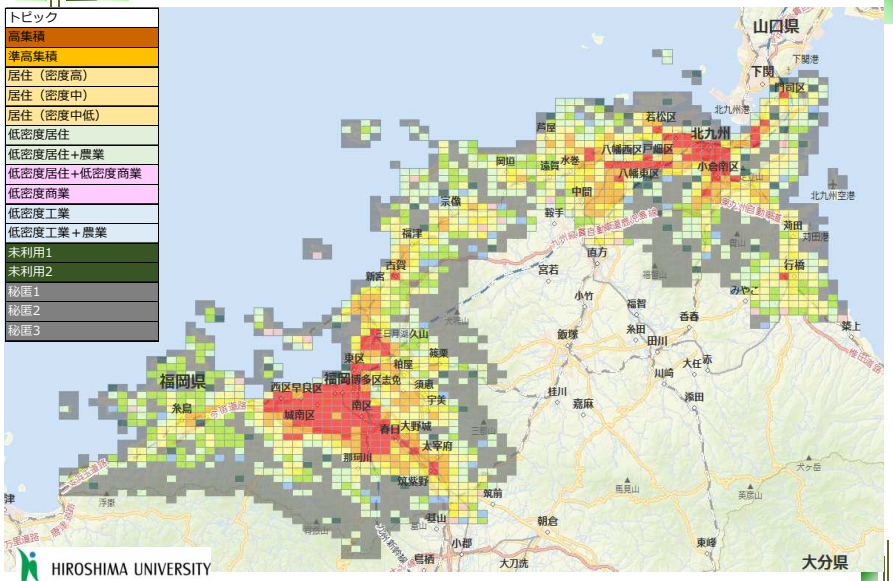
土地利用数値化指標2

● 建物開発率（U行列）

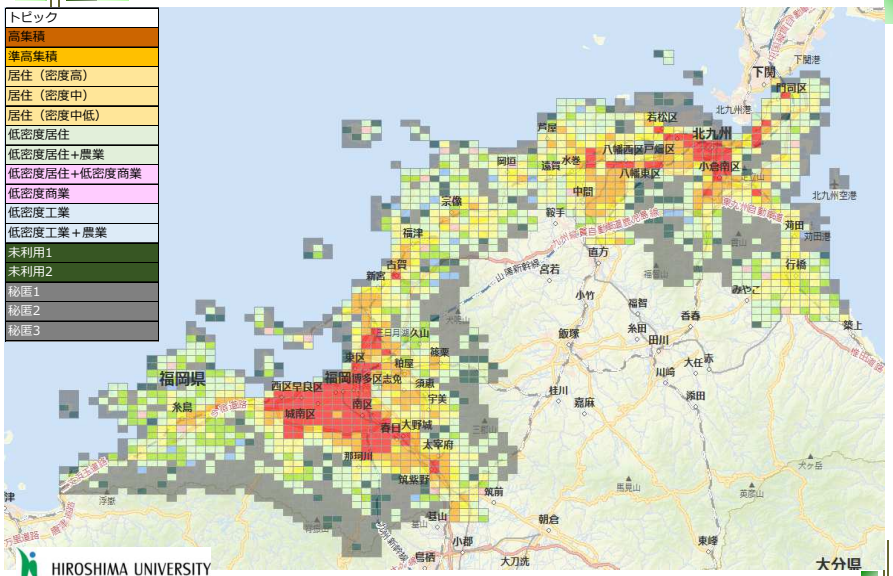
$$B_k = \frac{\sum_{i \in R} p_{ki}^{t+1} b_i^{t+1}}{\sum_{i \in R} p_{ki}^{t} b_i^{t}} \qquad p_{ki}^{t} = \frac{u_{ik}^{t}}{\sum_{k} u_{ik}^{t}}$$

- pは，時点t，メッシュiの地理トピックkのシェア
- bは，時点t，メッシュiの建物用地面積

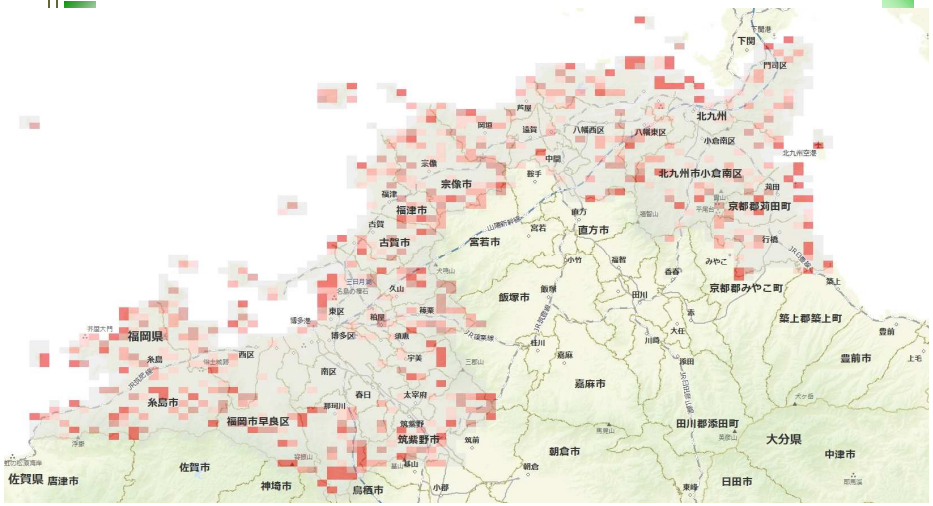
U行列：2000年の優勢トピック



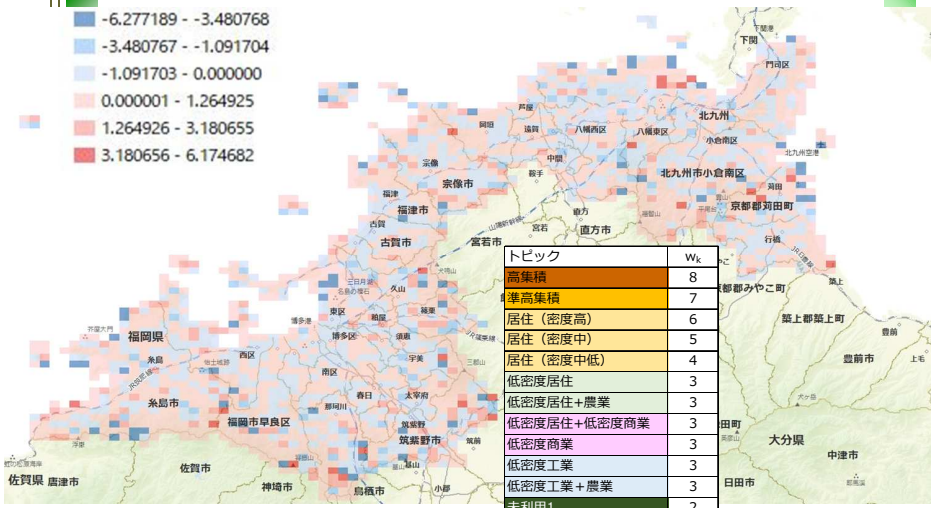
U行列：2010年の優勢トピック



地点別トピックのコサイン類似度

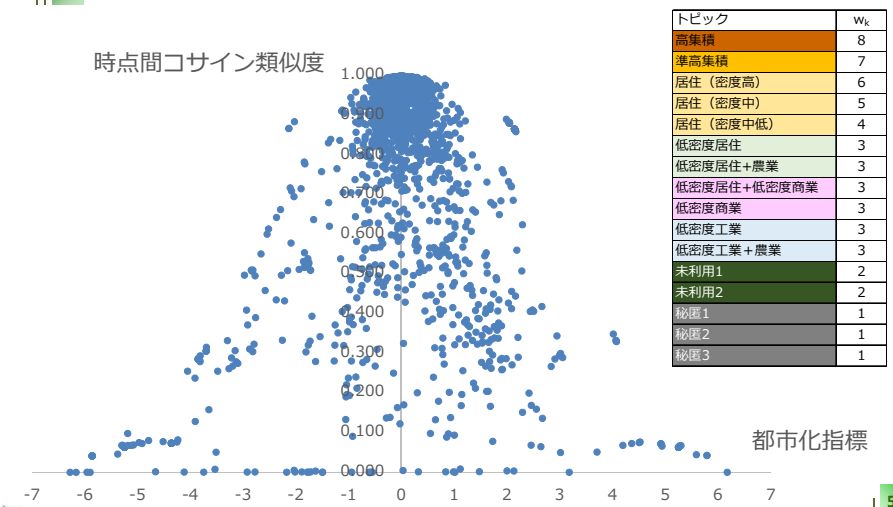


都市化推移指標



トピック	w_k
高集積	8
準高集積	7
居住（密度高）	6
居住（密度中）	5
居住（密度中低）	4
低密度居住	3
低密度居住＋農業	3
低密度居住＋低密度商業	3
低密度商業	3
低密度工業	3
低密度工業＋農業	3
未利用1	2
未利用2	2
秘匿1	1
秘匿2	1
秘匿3	1

都市化指標と時点間コサイン類似度



トピック	w_k
高集積	8
準高集積	7
居住（密度高）	6
居住（密度中）	5
居住（密度中低）	4
低密度居住	3
低密度居住＋農業	3
低密度居住＋低密度商業	3
低密度商業	3
低密度工業	3
低密度工業＋農業	3
未利用1	2
未利用2	2
秘匿1	1
秘匿2	1
秘匿3	1

各都市圏のトピック構成比（%）

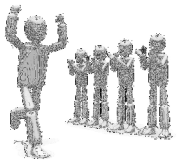
トピック	福岡都市圏			北九州都市圏		
	2000年	2010年	差分	2000年	2010年	差分
高集積	6.6	7.2	0.6	5.4	4.7	-0.7
準高集積	6.6	6.9	0.3	7.5	7.9	0.4
居住（密度高）	4	4.2	0.2	5.6	6.5	0.9
居住（密度中）	4.9	5.2	0.3	6.6	6.6	0
居住（密度中低）	4.2	4	-0.2	6.9	7.1	0.2
低密度居住	8.9	11.7	2.8	10.2	12.1	1.9
低密度居住+農業	6.7	5.6	-1.1	5.3	3.8	-1.5
低密度居住+低密度商業	4.9	4.7	-0.2	5.3	5.4	0.1
低密度商業	3.9	4.4	0.5	5.1	5.7	0.6
低密度工業	4.8	4.9	0.1	5.3	5.3	0
低密度工業＋農業	6	5.6	-0.4	6.9	6	-0.9
未利用1	4.5	5	0.5	4.4	4.9	0.5
未利用2	1	3.8	2.8	1.7	2.7	1
秘匿1	5.8	5.4	-0.4	3.6	3.3	-0.3
秘匿2	18.8	17.2	-1.6	14.2	13.3	-0.9
秘匿3	8.4	4.2	-4.2	6	4.7	-1.3
	100	100		100	100	

建物用地の伸び（km²）

トピック	福岡都市圏			北九州都市圏		
	2000年	2010年	伸び率	2000年	2010年	伸び率
高集積	69.8	87.2	1.25	31.7	31.5	0.99
準高集積	54.5	72.1	1.32	38.9	48.5	1.25
居住（密度高）	24.1	36.2	1.50	22.6	33.9	1.50
居住（密度中）	22.9	33.7	1.47	19.8	25.5	1.29
居住（密度中低）	1.4	15.6	11.14	14.3	20.2	1.41
低密度居住	7.1	13.6	1.92	7.1	11.2	1.58
低密度居住+農業	9.4	11.6	1.23	5.9	6.2	1.05
低密度居住+低密度商業	14.4	16.7	1.16	9.8	12.1	1.23
低密度商業	11.4	17.8	1.56	12.9	17.2	1.33
低密度工業	6.7	9.3	1.39	6.6	10.2	1.55
低密度工業＋農業	7	8.1	1.16	5.9	6.6	1.12
未利用1	2.5	3.2	1.28	2.7	3.3	1.22
未利用2	0.4	1.2	3.00	1.3	2.4	1.85
秘匿1	1.2	1.2	1.00	1.1	0.8	0.73
秘匿2	0.8	0.6	0.75	1.1	3.8	3.45
秘匿3	2	0.6	0.30	3.3	5.4	1.64
	235.6	328.7	1.40	185	238.8	1.29

地理トピックモデルの成果

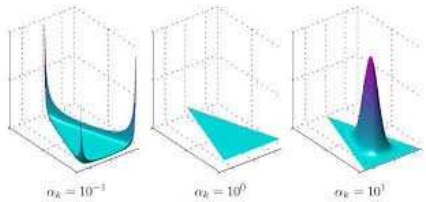
- 地理情報＋主成分分析に関する長年の違和感が解消した.
- 独立成分分析への苛立ち（分解の面白さ ⇔ 成分と負荷量の負値）も解消した.



- トピック分析（50周年幹事の宿題）に答えつつ，作成したアルゴリズムの転用に成功した.
- 教科書を首ったけで読み解いてアルゴルズムを自作する，しんどい作業が報われた.
- 連続データの離散化による属性数インフレという暴挙を試みたが，見込み通り計算ができた.
- ~~土地利用~~→社会経済活動集積量の推移を可視化できて満足だった.
- 都市計画方面への応用を考えられそう.

サンプル分類問題の”解呪法”

- データの属性に基づいてサンプル分類するとき，全ての属性を同じ重みで評価することに疑問があった.



- トピックモデルによる”解呪”の仕組み
  - Dirichlet分布＝単体の模式図によると，この分布でパラメータをうまく選べば，いくつかの属性の寄与を，ほぼ完全に無視できる.
  - つまり分類グループを構成する上で重要な＝都合の良い属性だけを眺めて特徴を把握している＝V行列.
  - 隠れ変数qの集計から合成されるU行列は，特徴と整合的なソフトクラスタリングを与える.
- 「おっさん」でも，数学の勉強は大事.

代替案の続き：非負値テンソル分解