証拠に基づく政策立案のための潜在プロファイル分析と数法則発見法を用いた社会実情のモデル化と可視化

Modeling and Visualization of Social Reality Using Latent Profile Analysis and Number Law Discovery Methods for Evidence-Based Policy Making

長瀬 永遠 (Towa Nagase)
u255013@st.pu-toyama.ac.jp

富山県立大学大学院 工学研究科 電子・情報工学専攻
情報基盤工学講座

N212, 09:30-10:00 Tuesday, February 13, 2024.

本研究の背景 1

情報技術の発達により，社会における様々なデータを観測・収集することが可能に

→ 政策分野においても，証拠に基づく政策立案（Evidence Based Poricy Making: EBPM）に注目が集まる

EBPM とは

政策における意思決定をデータに基づいて行うという考え方

課題

大規模なデータから有用な情報を取り出すデータマイニングの技術が必要

本研究の背景 2

どんな情報を取り出すことが有効？

政策分野では，原因と結果の間に成り立つ関係性が重要

→ 複数の要因が複雑に影響しあうため，人間が把握するには限界がある

アプローチ

行政が持つ統計データを用いて分析を行い，データ間の関係性を数理モデルによって表す

数理モデルの例

$$(\text{データ A}) = 2.0 \cdot (\text{データ B}) + 1.0 \cdot (\text{データ C}) - 1.0 \cdot (\text{データ D})$$

行政が持つ統計データ

地域経済分析システム（RESAS）

経済に関する項目を中心に自治体単位でデータを公開するオープンデータサイト

Table 1: RESAS から自動取得可能なデータ

データ項目	
農地平均取引価格	農業産出額
林地平均取引価格	海面漁獲物等販売金額
住宅用地平均取引価格	林産物販売金額
商業用地平均取引価格	林作業請負収入
マンション等平均取引価格	企業数
一人あたりの固定資産税	事業所数
一人当たりの地方税	就業者数
一人当たりの法人住民税	総人口
製造品出荷額	経営耕地面積
年間商品販売額	etc

自治体区分	総数
市	792
町	743
村	183
特別区	23
合計	1741

疑問

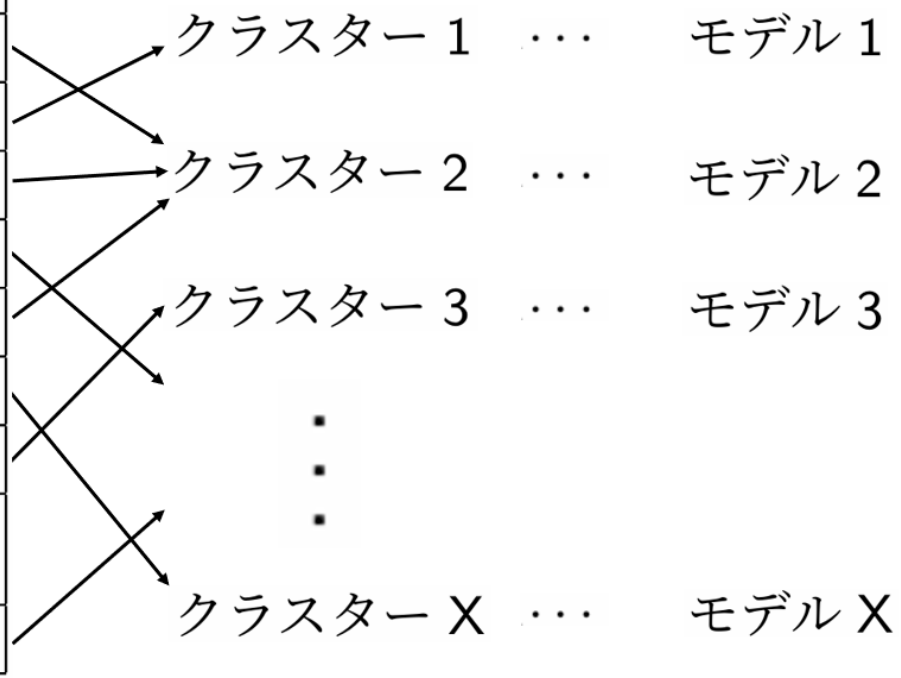
すべての市区町村を同じ尺度のサンプルとして扱うのは適切？

クラスタリングと回帰分析を用いたモデル化

目的

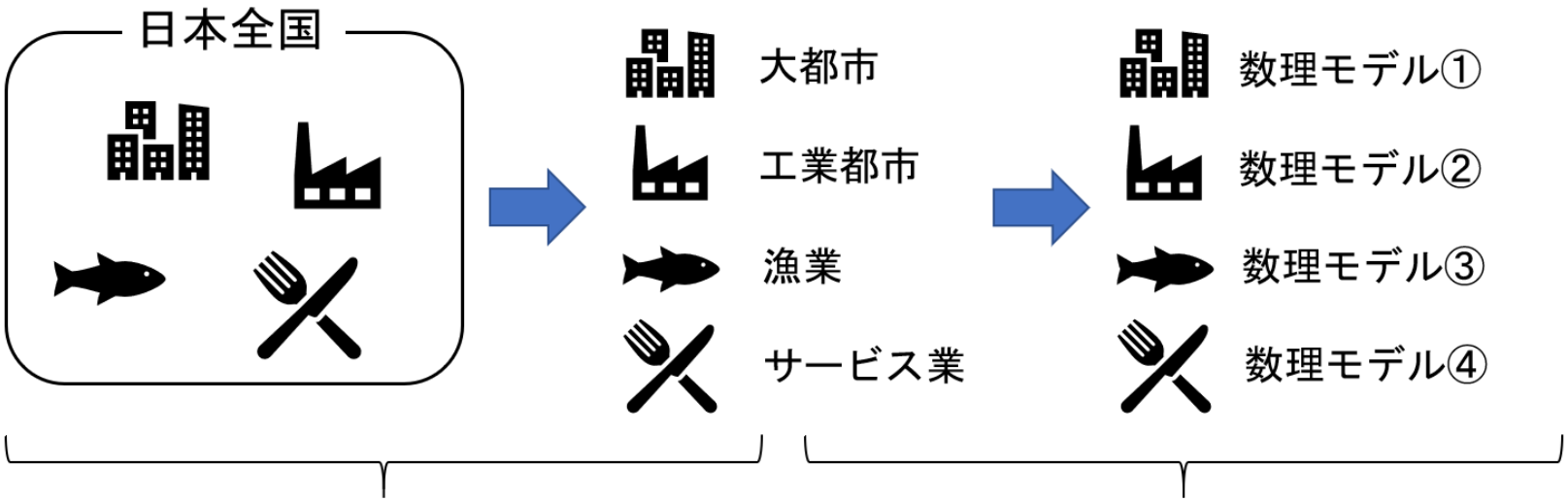
行政が持つ統計データを用いて自治体をクラスタリングし，その結果を考慮しながら統計データ間の関係性をモデル化する手法を提案する．

自治体名	統計データ
A 市	$(A_1, A_2, \cdots)$
B 町	$(B_1, B_2, \cdots)$
C 村	$(C_1, C_2, \cdots)$
D 区	$(D_1, D_2, \cdots)$
E 町	$(E_1, E_2, \cdots)$
F 村	$(F_1, F_2, \cdots)$
G 市	$(G_1, G_2, \cdots)$
⋮	⋮
Z 市	$(Z_1, Z_2, \cdots)$



本研究におけるクラスタリングと回帰分析の概要

使用する手法の概要



潜在プロファイル分析（LPA）

- データの潜在的な特徴を考慮
- ソフトクラスタリング

RF6.4法
（Rule extraction method from Fact 6.4）

- 表現力が高く複雑なデータに対応
- 式の可読性も比較的高い

潜在変数を考慮したクラスタリング 1

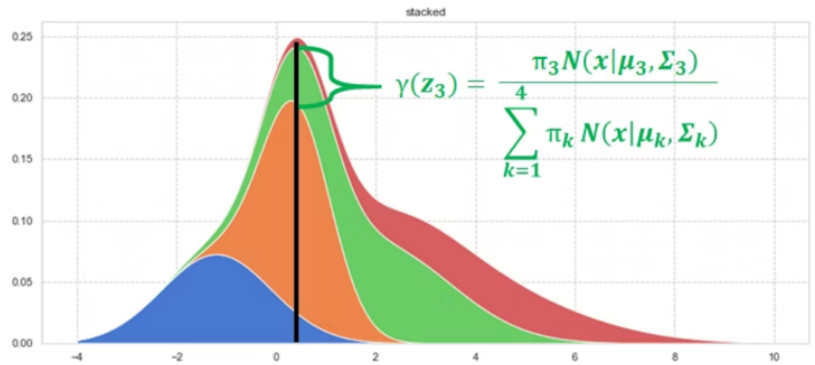
LPA

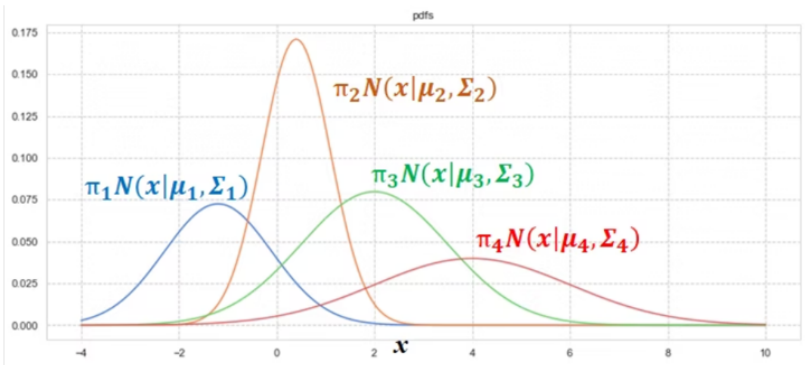
観測されたデータを混合分布モデルとして扱い，モデルの背後に存在する潜在変数を分析．その結果を考慮しながらデータをクラスタリングする手法である．

（混合ガウス分布）

$$N(\mathbf{x}_i|\pi,\mu,\mathbf{\Sigma}) = \sum_{k=1}^{K} \pi_k N_k(\mathbf{x}_i|\mu_k,\mathbf{\Sigma}_{rk}) \tag{1}$$

K：潜在変数の数，π_k：母集団における潜在変数の構成割合，
$\mathbf{x}_i = (x_1,\cdots,x_r)_i \quad i = 1,\cdots,I$：観測変数，
$\mu_k = (\mu_{1k},\cdots,\mu_{rk})$：観測変数の平均，$\mathbf{\Sigma}_{rk}$：各観測変数の共分散行列，





"(機械学習) 混合ガウス分布における EM アルゴリズムを実装を交えて丁寧に理解しようとした",
https : qiita.comFumio-eisanitems6862222fda418544f7e0, （参照 2024-02-12）

潜在プロファイル分析 2

分かってない情報

- 各分布のパラメータ：$\pi_k, \mu_k, \mathbf{\Sigma}_{rk}$
- 潜在変数 z_k の事後分布の期待値 = 各サンプルの存在確率：$E_{z_{ik}}[z_{ik}]$
- クラスター数 = 潜在変数の数：K

（EM アルゴリズム）

- **E ステップ**
  その時点で最適と考えれるパラメータを式 (2) に代入することによって $E_{z_{ik}}[z_{ik}]$ の値を求める．

$$E_{z_{ik}}[z_{ik}] = \frac{\pi_k N_k(\mathbf{x}_i|\mu_k, \Sigma_{rk})}{\sum_k \pi_k N_k(\mathbf{x}_i|\mu_k, \Sigma_{rk})} \tag{2}$$

- **M ステップ**
  E ステップで求めた $E_{z_{ik}}[z_{ik}]$ を式 (3) の右辺に代入し，式 (3) が最大となるようなパラメータを求める．求めたパラメータをその時点で最適なパラメータとし，再度 E ステップに戻る．

$$E_z[\ln p(\mathbf{x}, z|\pi, \mu, \sigma)] = \sum_i \sum_k E_{z_{ik}}[z_{ik}](\ln \pi_k + \ln N_k(\mathbf{x}_i|\mu_k, \Sigma_{rk})) \tag{3}$$

潜在プロファイル分析 3

モデル選択

EM アルゴリズムでは，クラスター数 K を仮置きしているので，ベイズ情報量基準 (Bayesian information criterion: BIC) が最小となるクラスター数を最適とする．

$$BIC = -2L + K \ln n \tag{4}$$

L: M ステップで最大化した値，K: クラスター数，n: サンプル数

各サンプルにおける存在確率の算出

最適なモデルが決定した後，各サンプルにおける観測変数ベクトル $\mathbf{y_i}$ を用いて存在確率を算出．

$$\pi_{k|\mathbf{x}_i} = \frac{\pi_k N_k(\mathbf{x_i}|\mu_k, \mathbf{\Sigma}_{rk})}{\sum_k \pi_k N_k(\mathbf{x}_i|\mu_k, \mathbf{\Sigma}_{rk})} \tag{5}$$

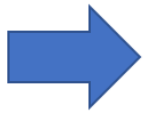
パーセプトロンを用いた数法則発見法 1

各変数各変数に対するデータ項目

目的変数 y: 自治体の財源
量的説明変数 x_1, x_2: 自主財源，依存財源
質的説明変数 q_1, q_2, q_3, q_4: 大都市，工業都市，漁業，サービス業

大都市

工業都市



漁業

サービス業

$$if \quad q_1 \quad y = v_0 + 5.5x_1^{1.5}x_2^{-0.5} + 0.1x_1^0x_2^{2.0}$$

$$if \quad q_2 \quad y = v_0 + 4.0x_1^{1.5}x_2^{-0.5} + 0.5x_1^0x_2^{2.0}$$

$$if \quad q_3 \quad y = v_0 + 0.5x_1^{1.5}x_2^{-0.5} + 3.5x_1^0x_2^{2.0}$$

$$if \quad q_4 \quad y = v_0 + 0.7x_1^{1.5}x_2^{-0.5} + 2.0x_1^0x_2^{2.0}$$

パーセプトロンを用いた数法則発見法 2

RF6.4 法

目的変数 y と M 個の量的説明変数 x_m，K 個の質的説明変数 q_{kl} からなるデータセットを与え，4 層パーセプトロンの学習を用いてパラメータ $c_{0d}, c_{gd}, c_{dkl}, v_{gm}$ を最適化することで質的条件付き多変量多項式を発見する．

（質的条件付き多変量多項式）

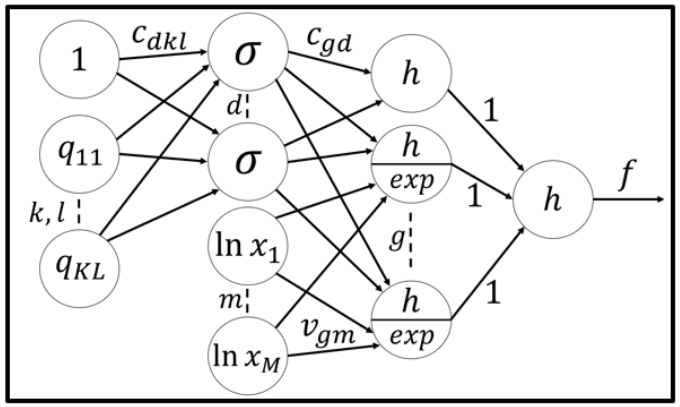
$$if \quad \bigwedge_{k} \bigvee_{q_{kl} \in Q_k^i} q_{kl} \quad then \quad y = v_0^i + \sum_{g=1}^{G} v_g^i \prod_{m=1}^{M} x_m^{v_{gm}^i} + \epsilon, \qquad i = 1, \cdots, I \quad (6)$$

RF6.4 法の 4 層パーセプトロン

$$f(\mathbf{q}, \mathbf{x}; \phi) = v_0 + \sum_{g=1}^{G} v_g s_g,$$

$$v_0 = \sum_{d=1}^{D} c_{0d}\sigma_d, \quad v_g = \sum_{d=1}^{D} c_{gd}\sigma_d,$$

$$\sigma_d = \sigma\left(\sum_{k=1}^{K}\sum_{l=1}^{L_k} c_{dkl} q_{kl}\right), \quad s_g = \exp\left(\sum_{m=1}^{M} v_{gm} \ln x_m\right)$$



パーセプトロンを用いた数法則発見法 3

分かってない情報

- モデルのパラメータ：$\phi = (c_0, c_{gd}, c_{dkl}, v_{gm})$
- パーセプトロンの中間層の数：G, D

最終的に求めたい数理モデル

$$f(\mathbf{q}, \mathbf{x}; \phi) = w_0 + \sum_{g=1}^{G} w_g s_g,$$

$$w_0 = \sum_{d=1}^{D} c_{0d}\sigma_d, \quad w_g = \sum_{d=1}^{D} c_{gd}\sigma_d, \tag{7}$$

$$\sigma_d = \sigma\left(\sum_{k=1}^{K}\sum_{l=1}^{L_k} c_{dkl} q_{kl}\right), \quad s_g = \exp\left(\sum_{m=1}^{M} v_{gm} \ln x_m\right)$$

下記の目的関数を BFGS 法によって最小化することでパラメータを求める

目的関数

$$E(\mathbf{q}, \mathbf{x}; \phi) = \frac{1}{2}\sum_{n=1}^{N}(f(\mathbf{q}, \mathbf{x}; \phi)^n - y^n)^2 \tag{8}$$

パーセプトロンを用いた数法則発見法 4

パーセプトロンの最適中間層数

パーセプトロンの学習では，中間層の数 G, D を仮置きしているので，BIC が最小となるときのモデルを採用．本研究では G, D ともに 1〜5 まで変化させて学習した．

$$BIC(G,D) = \frac{N}{2} \ln \left(\frac{1}{N} \sum_{n=1}^{N} (f(\mathbf{q}^n, f(\mathbf{x}^n; \hat{\phi}_{G,D}) - y^n)^2 \right) + \frac{Z}{2} \ln N \quad (9)$$

N：サンプル数，Z：パラメータ数

ルールの復元

パーセプトロンの学習結果では，各サンプルに対するパラメータベクトルが得られるが，より一般的な数法則として記述するために，以下の処理を行う．

1. 各サンプルに対する数法則の係数値ベクトル $\mathbf{c}^\mu = (c_0^\mu, \cdots, c_J^\mu)$ を求める
2. k-means 法を用いて I 個のクラスターを求め，それぞれの重心ベクトル $\mathbf{a}^I = (a_0^i, \cdots, c_J^i)$ を数法則の係数値とする

潜在的クラスターを考慮した数理モデル化手法

提案手法

自治体名	統計データ
A 市	$(A_1, A_2, \cdots, Y_A)$
B 町	$(B_1, B_2, \cdots, Y_B)$
C 区	$(C_1, C_2, \cdots, Y_C)$
⋮	⋮
Z 村	$(Z_1, Z_2, \cdots, Y_Z)$

LPA

所属確率 1	所属確率 2	⋯	所属確率 X
0.9	0.1	⋯	0
0.3	0	⋯	0.1
0.5	0.5	⋯	0
⋮	⋮	⋮	⋮
0	0	⋯	1.0

$$\tilde{y} = \frac{y - mean(y)}{std(y)} \qquad \tilde{x} = \frac{x}{max(x)}$$

量子化

目的変数	量的説明変数
$\widetilde{Y_A}$	$(\widetilde{A_1}, \widetilde{A_2}, \cdots)$
$\widetilde{Y_B}$	$(\widetilde{B_1}, \widetilde{B_2}, \cdots)$
$\widetilde{Y_C}$	$(\widetilde{C_1}, \widetilde{C_2}, \cdots)$
⋮	⋮
$\widetilde{Y_Z}$	$(\widetilde{Z_1}, \widetilde{Z_2}, \cdots)$

質的説明変数 1	質的説明変数 2	⋯	質的説明変数 X
[1, 0, ⋯,0]	[0, 0, ⋯,1]	⋯	[0, 0, ⋯,1]
[0, 0, ⋯,0]	[0, 0, ⋯,1]	⋯	[0, 0, ⋯,1]
[0, 1, ⋯,0]	[0, 1, ⋯,0]	⋯	[0, 0, ⋯,1]
⋮	⋮	⋮	⋮
[0, 0, ⋯,1]	[0, 0, ⋯,1]	⋯	[1, 0, ⋯,0]

RF6.4法

$$Y_1 = v_{1\_0} + \sum_{g=1}^{G} v_{1\_g} \prod_{m=1}^{M} x_m^{v_{gm}} \qquad Y_2 = v_{2\_0} + \sum_{g=1}^{G} v_{2\_g} \prod_{m=1}^{M} x_m^{v_{gm}} \quad \cdots \quad Y_X = v_{X\_0} + \sum_{g=1}^{G} v_{X\_g} \prod_{m=1}^{M} x_m^{v_{gm}}$$

GIS による分析結果の可視化 1

各自治体が所属するクラスターと対応する数理モデルを視覚的に提示するため，GIS を用いたシステムを作成

GIS とは

地理情報システム（Geographic Information System）の略称．地図上にデータを表示し，可視化や重ね合わせなどを行うことが可能

GIS を用いる狙い

- 分析結果全体を俯瞰的に把握
- クラスターの分布と地理的特徴の関係を可視化

GIS による分析結果の可視化 2

動画

提案システムのながれを動画でお見せします.

数値実験の概要 1

使用したデータ

項目：11 種類の量的データ
サンプル数： データ欠損のない 650 の自治体におけるオープンデータ
対象年：2020 年
参照元： 地域経済分析システム-RESAS-

Table 2: 数値実験に用いたデータ

データ項目	単位	データ項目	単位
1 人あたりの固定資産税	千円/人	総人口	人
1 人あたりの地方税	千円/人	住宅用地平均取引価格	円 $/m^2$
1 人当たりの法人住民税	千円/人	商業用地平均取引価格	円 $/m^2$
経営耕地面積	畝 / 経営体	農地平均取引価格	円 $/m^2$
製造品出荷額	万円	林地平均取引価格	円 $/m^2$
年間商品販売額	百万円		

数値実験の概要 2

目的

以下の 3 つの事柄について，検証するためそれぞれ数値実験を行った

- LPA は自治体の特徴をどのように捉えているか
- 自治体のクラスタリング結果を考慮することが数理モデルの精度向上に有効か
- クラスタリング手法として LPA を用いることは適切か

実験 1

総務省から発行されている市町村類型の内容と本研究で得られたクラスタリング結果を比較

実験 2

クラスタリングを用いない手法と提案手法それぞれにおいて得られた数理モデルの精度を比較

実験 3

クラスタリングに k-means を用いた手法と提案手法それぞれにおいて得られた数理モデルの精度を比較

実験１の概要１

市町村類型とは

「人口」と「産業構造」の二軸から市町村を分類したもの.

類型の数

- 政令指定都市，特別区，中核市，特例市 ： それぞれ 1 類型
- 一般市 ：16 類型
- 町村 ：15 類型

Table 3: 市の類型

産業構造		Ⅱ次,Ⅲ次90%以上		Ⅱ次,Ⅲ次90%未満	
		Ⅲ次65%以上	Ⅲ次65%未満	Ⅲ次55%以上	Ⅲ次55%未満
人口	0以上～50,000未満	Ⅰ-3	Ⅰ-2	Ⅰ-1	Ⅰ-0
	50,000以上～100,000未満	Ⅱ-3	Ⅱ-2	Ⅱ-1	Ⅱ-0
	100,000以上～150,000未満	Ⅲ-3	Ⅲ-2	Ⅲ-1	Ⅲ-0
	150,000以上～	Ⅳ-3	Ⅳ-2	Ⅳ-1	Ⅳ-0

Table 4: 町村の類型

産業構造		Ⅱ次,Ⅲ次80%以上		Ⅱ次、Ⅲ次80%未満
		Ⅲ次60%以上	Ⅲ次60%未満	
人口	0以上～5,000未満	Ⅰ-2	Ⅰ-1	Ⅰ-0
	5,000以上～10,000未満	Ⅱ-2	Ⅱ-1	Ⅱ-0
	10,000以上～15,000未満	Ⅲ-2	Ⅲ-1	Ⅲ-0
	15,000以上～20,000未満	Ⅳ-2	Ⅳ-1	Ⅳ-0
	20,000以上～	Ⅴ-2	Ⅴ-1	Ⅴ-0

実験 1 の概要 2

分析対象

年代：2020 年
クラスター数：15 クラスター
範囲： 九州地方
サンプル数：112 の自治体

分析方法

15 個のクラスターそれぞれに所属する自治体の類型を調査し，クラスターごとの内訳を分析

Table 5: 数値実験に用いたデータ

データ項目	単位	データ項目	単位
1 人あたりの固定資産税	千円/人	総人口	人
1 人あたりの地方税	千円/人	住宅用地平均取引価格	円 $/m^2$
1 人当たりの法人住民税	千円/人	商業用地平均取引価格	円 $/m^2$
経営耕地面積	畝 / 経営体	農地平均取引価格	円 $/m^2$
製造品出荷額	万円	林地平均取引価格	円 $/m^2$
年間商品販売額	百万円		

実験１の結果と考察

Figure 1: クラスター１における市の類型割合

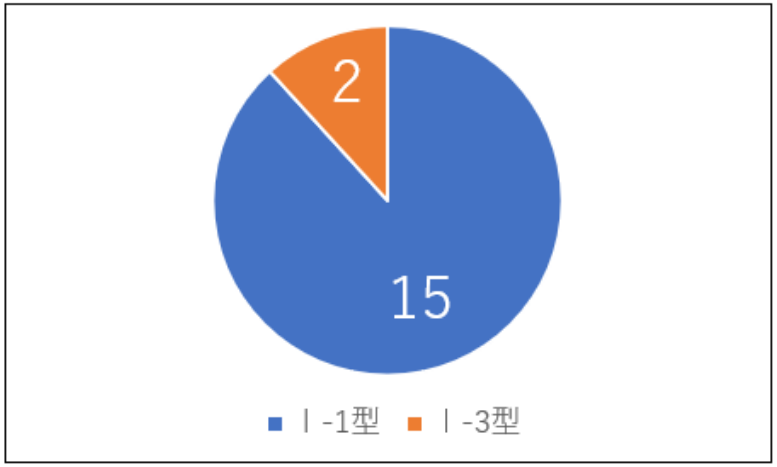
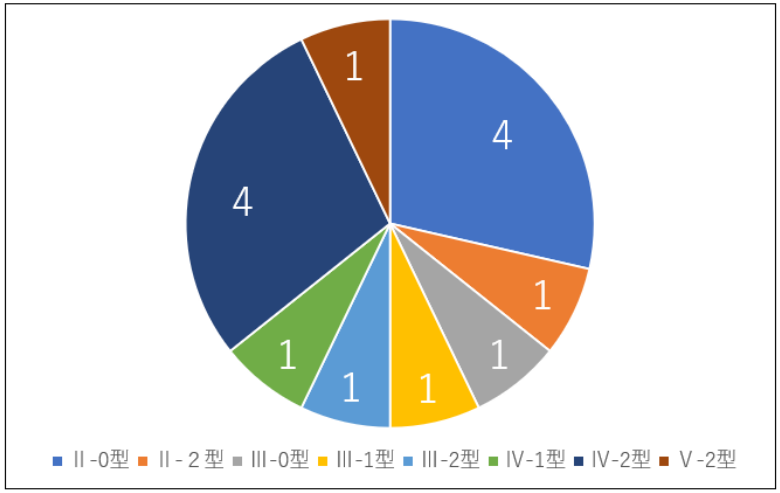


Figure 2: クラスター１における町村の類型割合



結果

- 市においては全体的に人口が近い自治体が同じクラスターとなった
- 町村は比較的そうでない場合が多かった

考察

- クラスタリングに用いたデータが産業面で不十分であった
- 町村には人口とは別の潜在的共通要素が存在する

実験２の概要

検証内容

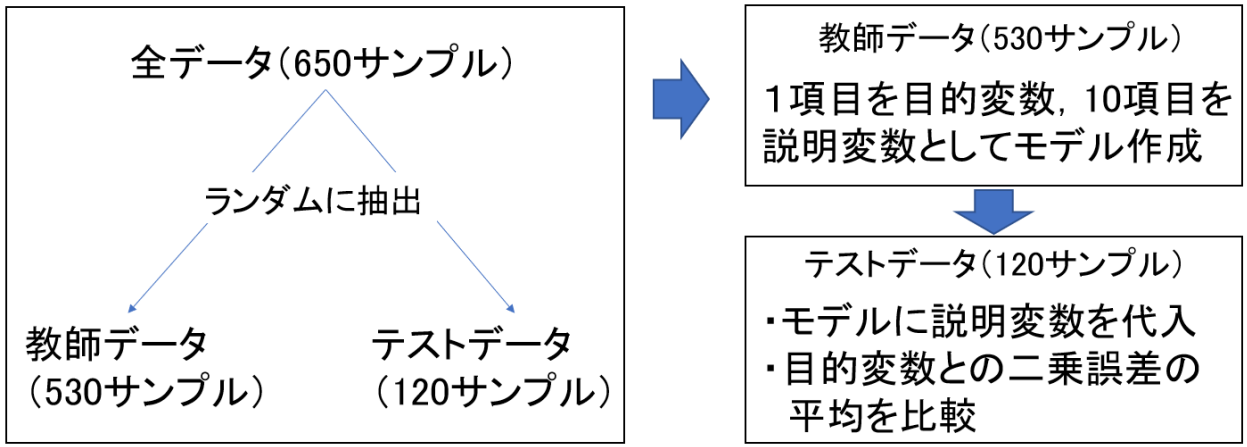
自治体のクラスタリング結果を考慮することが数理モデルの精度向上に有効か

比較対象

- クラスタリングを用いない手法（RF5 法）
- 提案手法（LPA ＋ RF6.4 法）

データ

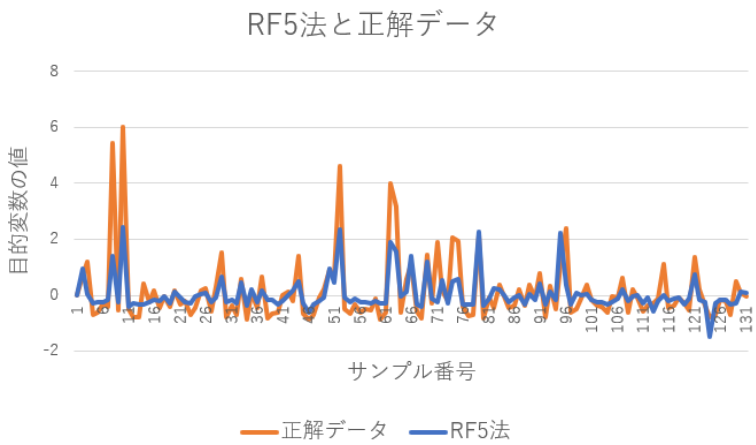
データ：11 項目のオープンデータ
サンプル： データ欠損のない 650 の自治体

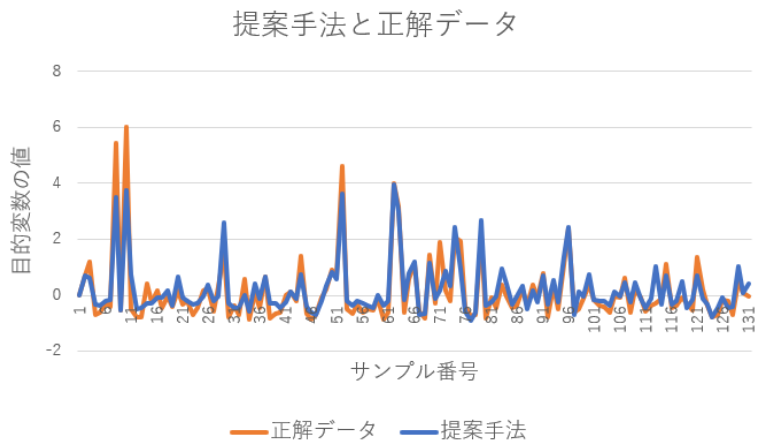


実験 2 の結果と考察

Table 6: 最小の BIC と二乗誤差の平均値

手法	BIC	二乗誤差
RF5 法	-234.7011	0.5543
提案手法	-254.8814	0.2262





考察

- BIC，二乗誤差ともに提案手法が優れていることから自治体のクラスターを考慮することは数理モデルの精度向上に有効である
- 提案手法においても正解データに対して誤差が大きいため，改善の余地が残される

実験 3 の概要

検証内容

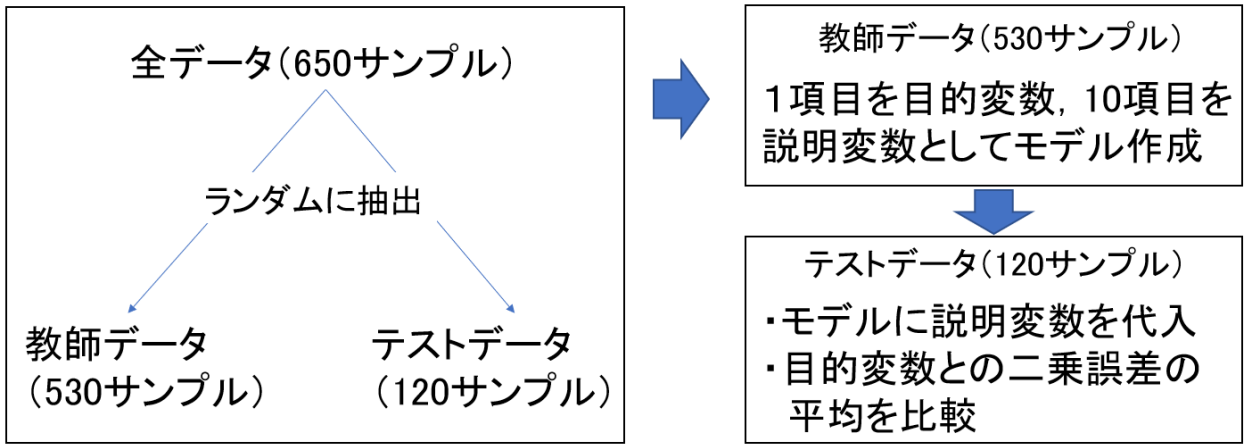
クラスタリング手法として LPA を用いることは適切か

比較対象

- クラスタリングに k-means を用いた手法（k-means ＋ RF6.4 法）
- 提案手法（LPA ＋ RF6.4 法）

データ

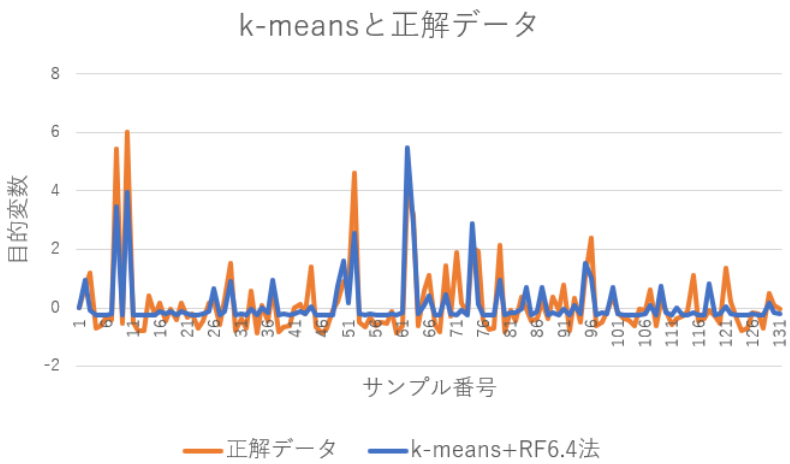
データ：11 項目のオープンデータ
サンプル： データ欠損のない 650 の自治体

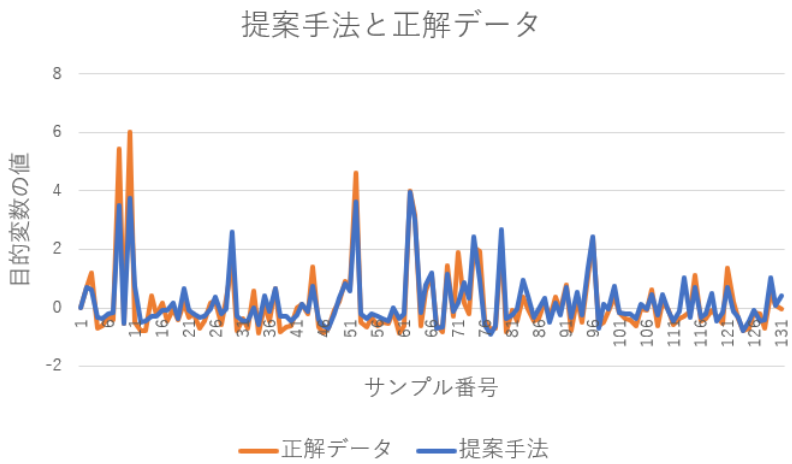


実験 3 の結果と考察

Table 7: 最小の BIC と二乗誤差の平均値

手法	BIC	二乗誤差
k-means+RF6.4 法	-244.8024	0.3929
提案手法	-254.8814	0.2262





考察

- BIC，二乗誤差ともに提案手法が優れていることからクラスタリング手法に LPA を用いることは適切である
- グラフから k-means は小さい値への適合に難がある可能性が示唆された

おわりに

まとめ

行政が持つ統計データを用いて，自治体をクラスタリングし，データの関係を数理モデルで表す手法を提案した．提案手法の結果に対して，クラスタリングの傾向とモデルの精度に関する検証を行った．

- 既存の市町村分類と比較して，人口の観点では概ね合致したが，産業構造では異なる点が見られた
- クラスタリングを用いない場合と比較してモデルの精度が向上した
-
- モデルの精度自体には改善の余地がある

今後の展望

- 項目数，サンプル数ともにより大規模なデータを用いた議論が必要
- 専門的な知見をもって検証と考察することが必要
- データの関係性が既知な対象への適用も検討すべき

差分

以下，差分

類団の表

1	s-Ⅰ1	15
総数	s-Ⅰ3	2
31	m-Ⅱ0	4
	m-Ⅱ2	1
	m-Ⅲ0	1
	m-Ⅲ1	1
	m-Ⅲ2	1
	m-Ⅳ1	1
	m-Ⅳ2	4
	m-Ⅴ2	1

2	s-Ⅲ3	1
総数	中核市	1
2		

3	大都市	1
総数		
1		

5	s-Ⅲ1	3
総数	s-Ⅲ3	2
6	s-Ⅳ1	1

3	大都市	1
総数		
1		

6	s-Ⅰ0	1
総数	s-Ⅰ1	6
14	s-Ⅰ3	3
	m-Ⅱ2	1
	m-Ⅳ0	1
	m-Ⅳ1	1
	m-Ⅳ2	1

7	s-Ⅰ0	2
総数	s-Ⅰ1	4
23	s-Ⅰ2	1
	s-Ⅰ3	1
	s-Ⅱ1	8
	s-Ⅱ2	3
	s-Ⅱ3	2
	m-Ⅴ2	2

8	sⅠ-1	1
総数	s-Ⅰ2	1
10	s-Ⅰ3	1
	s-Ⅱ2	1
	s-Ⅱ3	4
	m-Ⅴ2	2

9	s-Ⅱ3	1
総数	m-Ⅱ2	1
3	m-Ⅴ2	1

10	s-Ⅱ1	1
総数	s-Ⅱ2	1
5	s-Ⅱ3	1
	m-Ⅳ2	1
	m-Ⅴ1	1

11	政令指定都市	1
総数	大都市	1
3	中核市	1

12	中核市	1
総数		
1		

提案手法の数式

$$y = 1.0504 + 0.0915h_1 - 0.4932h_2 - 1.5345h_3$$
$$y = 1.5630 - 0.2676h_1 + 0.4047h_2 + 0.2488h_3$$
$$y = 2.6819 - 0.1679h_1 - 0.1266h_2 - 1.3990h_3$$
$$y = 3.7813 - 0.6816h_1 + 1.0757h_2 + 0.8165h_3$$
$$y = 0.9802 - 0.1457h_1 + 0.1915h_2 + 0.0174h_3$$
$$h_1 = x_1^{-0.3317} x_2^{0.8776} x_3^{0.4363} x_4^{0.2339} x_5^{0.8738} x_6^{0.5663} x_7^{-0.1736} x_8^{0.6950} x_9^{1.1073} x_{10}^{0.2303}$$
$$h_2 = x_1^{-0.0301} x_2^{-1.3162} x_3^{0.4784} x_4^{0.1623} x_5^{0.0419} x_6^{-0.2382} x_7^{-0.1257} x_8^{-0.0275} x_9^{-0.0340} x_{10}^{-0.}$$
$$h_3 = x_1^{0.3044} x_2^{-0.0357} x_3^{0.1758} x_4^{0.5694} x_5^{1.2188} x_6^{1.2825} x_7^{-0.5933} x_8^{-0.7382} x_9^{0.3629} x_{10}^{0.4825}$$

(10)

(目的変数)
住宅地取引平均価格
(説明変数)
固定資産税，地方税，住民税，経営耕地面，製造品出荷額，商業地取引平均価格，年間商品販売額，農地取引平均価格，林地取引平均価格，総人口

3.1 因果探索によるデータ間の関係性

因果探索とは，観測データを用いて，データ群の因果グラフ（複数の観測データにおいて，それぞれの値がお互いに及ぼしあっている影響の度合いを構造的に示したもの）を導出するための教師なし学習である．

LiNGAM

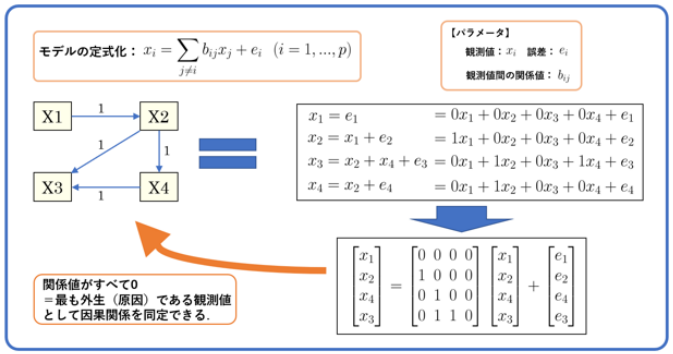


図15　LiNGAMのアルゴリズム

LiNGAMの制約

1. 内生変数と外生変数をつなぐ関数は線形．
2. 外生変数の分布は非ガウス連続分布．
3. 因果グラフは非巡回
4. 外線変数同士は互いに独立．

内生・・・観測済み　外生・・・未観測

Direct-LiNGAM

Direct-LiNGAMのアプローチ

・・・回帰分析を用いる手法

・内生変数群から2変数を取り出しそれらの変数間に成り立つ因果関係を同定することを繰り返して因果グラフの始点を探索．
・その変数を内生変数群から除外し，残った変数のみで内生変数群を再形成．

適当な2変数

$$\begin{cases} x_1 = e_1 \\ x_2 = b_{21} x_1 + e_2 \end{cases}$$

目的変数をx_1，説明変数をx_2としたときの回帰残差(r_2)．

$$r_2 = \left\{1 - \frac{b_{21} cov(x_1, x_2)}{var(x_2)}\right\} e_1 - \frac{b_{21} var(x_1)}{var(x_2)} e_2$$

ダルモア・スキットビッチの定理

2つの確率変数y_1, y_2が互いに独立な確率変数$s_i (i = 1, \ldots, q)$を用いて下記のように表されるとき，y_1, y_2が独立なら，$\alpha_j, \beta_j \neq 0$となるような変数s_jはガウス分布に従う．

$$y_1 = \sum_{i=1}^{q} \alpha_i s_i \qquad y_2 = \sum_{i=1}^{q} \beta_i s_i$$

3.2 DEA による効率値と入力・出力改善値の導出

DEA とは，ある分野における組織の集合において，対象の組織の業績を評価するために生み出されたノンパラメトリックなアプローチである．組織とは，その活動においていくつかの種類の入力（投入）をいくつかの出力（産出）に変換することに携わる生産体（Decision Making Unit: DMU）を指す．

<CCR モデルの主問題 >

$$\text{maximize} \quad \frac{u^T y_o}{v^T x_o} = z$$

対象の効率性＝対象の仮想出力 / 対象の仮想入力
を可能な限り最大化する重みを定める．

$$\text{subject to} \quad -v^T X + u^T Y \leq 0$$

他のDMUに対しても，効率値は1を超えないように制約．

$$u \geq 0 \quad v \geq 0$$

重みが負の値をとらないように制約．

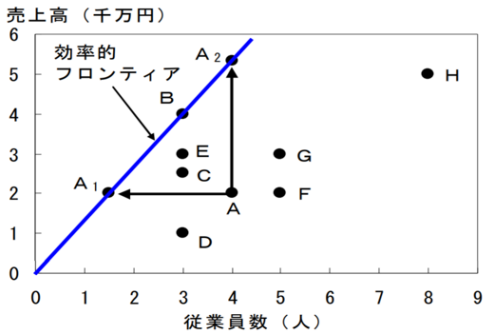


図16　1入力・1出力のDMU群

双対問題

< 入力指向モデル >

$$\text{minimize} \quad w = \theta$$
$$\text{subject to} \quad Y\lambda \geq y_o$$
$$-X\lambda + x_o\theta \geq 0$$
$$\lambda \geq 0$$

< 出力指向モデル >

$$\text{maximize} \quad w = \eta$$
$$\text{subject to} \quad X\mu \leq x_o$$
$$-Y\mu + y_o\eta \leq 0$$
$$\mu \geq 0$$

< 入力改善案 >

$$\hat{x}_i = \sum_{k=1}^{K} x_{ik}\lambda_k \qquad i = 1, 2, \cdots m$$

< 出力改善案 >

$$\hat{y}_j = \sum_{k=1}^{K} y_{jk}\mu_k \qquad j = 1, 2, \cdots n$$

それぞれのDMUに対して各入力をどれだけ減少，各出力をどれだけ増加させれば評価値が1になるかが算出できる．
また，その際に参考としたDMUもわかる．

潜在変数を考慮したクラスタリング1

LPA

観測されたデータを混合分布モデルとして扱い，モデルの背後に存在する潜在変数を分析．その結果を考慮しながらデータをクラスタリングする手法である．

（混合ガウス分布）

$$N(\mathbf{x}_i|\pi,\mu,\mathbf{\Sigma}) = \sum_{k=1}^{K} \pi_k N_k(\mathbf{x}_i|\mu_k,\mathbf{\Sigma}_{rk}) \tag{11}$$

K：潜在変数の数，π_k：母集団における潜在変数の構成割合，
$\mathbf{x}_i = (x_1,\cdots,x_r)_i \quad i = 1,\cdots,I$：観測変数，
$\mu_k = (\mu_{1k},\cdots,\mu_{rk})$：観測変数の平均，$\mathbf{\Sigma}_{rk}$：各観測変数の共分散行列，

Σ_{rk}

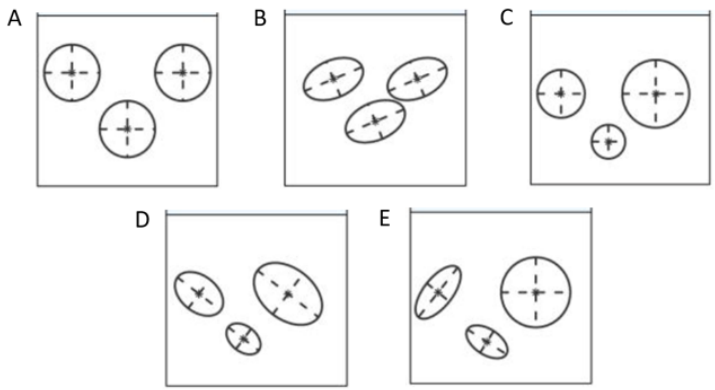
A
$$\begin{bmatrix} \sigma_1^2 & & & \\ 0 & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_r^2 \end{bmatrix}$$

B
$$\begin{bmatrix} \sigma_1^2 & & & \\ \sigma_{21} & \sigma_2^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_r^2 \end{bmatrix}$$

C
$$\begin{bmatrix} \sigma_{1k}^2 & & & \\ 0 & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ 0 & 0 & \cdots & \sigma_{rk}^2 \end{bmatrix}$$

D
$$\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1} & \sigma_{r2} & \cdots & \sigma_{rk}^2 \end{bmatrix}$$

E
$$\begin{bmatrix} \sigma_{1k}^2 & & & \\ \sigma_{21k} & \sigma_{2k}^2 & & \\ \vdots & \vdots & \ddots & \\ \sigma_{r1k} & \sigma_{r2k} & \cdots & \sigma_{rk}^2 \end{bmatrix}$$



潜在プロファイル分析 2

パラメータ推定

観測変数 $\mathbf{y}$ と潜在変数 z の同時確率の尤度を目的関数とし，その最大値を求める．値の最大化には EM アルゴリズムを用いる．

（EM アルゴリズム）

- **E ステップ**
  その時点で最適と考えれるパラメータを式 (2) に代入することによって $E_{z_{ik}}[z_{ik}]$ の値を求める．

$$E_{z_{ik}}[z_{ik}] = \frac{\pi_k f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})}{\sum_k \pi_k f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})} \tag{12}$$

- **M ステップ**
  E ステップで求めた $E_{z_{ik}}[z_{ik}]$ を式 (3) の右辺に代入し，式 (3) が最大となるようなパラメータを求める．求めたパラメータをその時点で最適なパラメータとし，再度 E ステップに戻る．

$$E_z[\ln p(\mathbf{y}, z|\pi, \mu, \sigma)] = \sum_i \sum_k E_{z_{ik}}[z_{ik}](\ln \pi_k + \ln f_k(\mathbf{y}_i|\mu_k, \Sigma_{rk})) \tag{13}$$

パーセプトロンを用いた数法則発見法 3

パーセプトロンの学習方法

RF6.4 法の学習には，準ニュートン法を基本枠組みとする学習アルゴリズムである BPQ（Back propagation based on Partial Quasi-Newton）法を用いる．

（最小化する関数）

$$E(\mathbf{q}, \mathbf{x}; \phi) = \frac{1}{2} \sum_{n=1}^{N} (f(\mathbf{q}, \mathbf{x}; \phi)^n - y^n)^2 \tag{14}$$

Step 1:パラメータの初期化

ϕ_1 を初期化．$\mathbf{H}_1 = \mathbf{I}, b = 1$ とおく．$\mathbf{I}$ は ϕ と次元の等しい単位行列

Step 2: 探索方向 $\Delta\phi_s$ の計算

探索方向 $\Delta\phi_s = -\mathbf{H}_s \nabla E(\mathbf{q}, \mathbf{x}; \phi_s)$ を計算．$\nabla E(\mathbf{q}, \mathbf{x}; \phi_s)$ の全要素が 0.0001 未満の場合，反復を終了．

Step 3: 最適探索幅 α_s の計算

$E(\mathbf{x}; \phi_b + \alpha_b \Delta\phi_b)$ を最小にする最適探索幅 α_b を計算．

Step 4: 結合重み ϕ_s の更新

$\boldsymbol{\phi_{b+1}} = \boldsymbol{\phi_b} + \boldsymbol{\alpha_s \Delta \phi_b}$ に従って結合重み ϕ_b を更新．

Step 5: 二次微分の逆行列の近似値 $\mathbf{H}$ の更新

$b \equiv 0 \pmod{Z}, Z$：全パラメータ数) のとき，$\mathbf{H}_{b+1} = \mathbf{I}$ とし，それ以外のとき，BFGS 公式で $\mathbf{H}_{b+1}$ を更新．$b \leftarrow b + 1$ として，Step2 の戻る．