情報処理学会論文誌 Vol.64 No.1 179–188 (Jan. 2023)

クラス間の距離を考慮した多変量2値分布における代表点分析法に関する研究

竹本 真悟[1,a)] 山下 遥[1,b)]

受付日 2022年5月5日, 採録日 2022年10月4日

概要：概要：多変量 2 値分布をよく表すような k 個の点として定義された 2 値型代表点分析法は，様々なデータ分析へと活用されている．ここで，2 値型代表点分析法には，値が同じような代表点が多数出てしまうと，代表点の解釈が困難となるという問題が存在する．このような場合，代表点どうしの距離が大きくなるような距離関数を定義することで，代表点間の違いに着目した分析が可能となり，有効なアプローチであると考えられる．本研究では，求める基礎としてデータと代表点との適合度の最大化，および代表点どうしの距離の最大化の双方を考慮した新たな 2 値型代表点分析を提案する．ただし，2 値型代表点を求めるための計算量が膨大となってしまうため，効率良く解を求める必要がある．また，近似アルゴリズムを適用する場合，得られる解の精度が保証されていることが実用上望ましい．そこで，提案する 2 値型代表点分析法の目的関数は，劣モジュラ性を持っていることを示し，貪欲法を用いたアルゴリズムにより，精度の下限を保証した近似アルゴリズムを提案する．さらに，本研究ではシミュレーション実験により本研究における提案手法の性質を検証しながら妥当性を示したうえで 2 種類の実際のデータに提案手法をあてはめて，分析の具体例を示すとともに実世界における適用の有効性を確認する．

キーワード：代表点分析法，劣モジュラ関数最適化，近似アルゴリズム，クラスタリング

A Study on Principal Points for a Multivariate Binary Distribution Considering the Distance between Clusters

Shingo Takemoto[1,a)] Haruka Yamashita[1,b)]

Received: May 5, 2022, Accepted: October 4, 2022

Abstract: Principal points for a binary distribution, defined as k points that well represent a multivariate binary distribution, has been used for various data analysis. The problem with the Principal points method is difficult to interpret representative points when there are many representative points with similar values. In such cases, defining a distance function that increases the distance between representative points can be an effective approach because it allows analysis that focuses on differences among principal points. This study proposes a method of principal points for a multivariate binary distribution that considers both the maximization of the goodness of fit between data and representative points and the maximization of the distance between representative points as the bias for the search. Here, the computational complexity of obtaining a binary representative point is enormous, so it is necessary to find a solution efficiently. In addition, when applying an approximation algorithm, it is desirable from a practical standpoint that the accuracy of the obtained solution is guaranteed. Therefore, in this study, we show that the objective function of the proposed an analysis method using principal points for a multivariate binary distribution is submodular and propose an approximation algorithm that guarantees a lower bound of accuracy by using a greedy algorithm. We then apply the proposed method to two types of real-world data to demonstrate the effectiveness of the method in real-world applications and to provide examples of analysis.

Keywords: principal points analysis, submodular function optimization, approximation algorithm, clustering

1. はじめに

近年，データ分析の分野において，クラスタリングに基づき，データ全体をいくつかのグループに分割して解釈するための方法が数多く展開されている．その中で，多変量2値データに対して実現値（2値）として求められる代表的な点をもとに，2値データを複数のクラスタへと分割する2値型代表点分析法 [1] が提案されている．しかしながら，この問題はNP困難な問題 [1] であるため，その性質を活用した近似アルゴリズムに関する研究 [2], [3], [4] や，応用例 [5], [6] に関する研究を通して有効性が示されてきた．

この分析手法は確率分布の代表としてFlury [7] によって提案されたPrincipal Pointsの定義を基礎としたモデルである．Principal Pointsに関する研究は，存在する部分空間を特定し，任意の分布に対するPrincipal（代表的な）点を容易に探索することを目的とした理論的な研究 [8], [9], [10], [11] や，その実データ分析への応用など [12], [13]，数多く存在する．また，機械学習の分野において有名なk-means法 [14] の理論的な性質を明らかにするための研究と位置付けることもでき，注目を集めている．

ここで，分布を多変量2値分布に限定したときのPrincipal Pointsとして定義され，その分布に従って生成されるデータとあてはまり（適合度）を最大化するような点を探索する2値型代表点分析法では，分布を分割する，すなわち分布のクラスタリングと，分布全体の代表点による簡潔な表現，すなわち，分布の量子化という2つの側面 [15] が存在する．特に代表点に興味がある場合において，2値型代表点分析を適用したとき，データの偏りが大きく偏った2値分布の形をしている場合に，値が同じような代表点が多数出てしまい，代表点の解釈が困難となる場合が指摘される [6]．このような場合，代表点どうしの距離が大きくなるような距離関数を定義することが有効であると考えられる．

そこで，本研究では2値型代表点 [1] を基礎として2値分布と代表点との適合度の最大化，および代表点どうしの距離の最大化の双方を考慮した新たな2値型代表点分析法を提案する．ただし，2値型代表点を求めるための計算はNP-困難問題として知られているため [1]，効率良く解を求める必要がある．そのため，本研究では近似解法として貪欲法を適用する．提案する目的関数は，劣モジュラ性 [1] を持っていることが示される．よって，貪欲法に基づくアルゴリズム [16] により，精度の下限を保証した近似解を得ることができる．さらに，本研究ではシミュレーション実験により本研究における提案手法の妥当性を示したうえで2種類の実際のデータに提案手法をあてはめて，分析の具体例を示すとともに有効性を確認する．

[1] 上智大学理工学部
Faculty of Science and Technology, Sophia University, Chiyoda, Tokyo 102–8554, Japan
[†1] 現在，東京工業大学情報理工学院
Presently with Presently with School of Computing, Department of Computer Science, Tokyo Institute of Technology, Kanagawa, Yokohama, 226–8501, Japan
a) takemoto.s.af@m.titech.ac.jp
b) h-yamashita-1g8@sophia.ac.jp

2. 準備

2.1 2値型代表点分析法

多変量データが2値の場合，連続値と同じ方法で代表的な値をとろうとすると，連続値になることが多く，実現値をとることは少ないので，解釈が難しくなる．そこで，Yamashitaらは2値型代表点を以下のように定義している [1]．

まず，p 次元2値の確率変数を $\boldsymbol{x}_i \in \{0,1\}^p$ とし，$\boldsymbol{x}_i$ を並べた確率変数ベクトルを $\boldsymbol{X}$ とする．また，2^p 個の p 次元2値ベクトルを $\boldsymbol{y}_j \in \{0,1\}^p$（$j = 1, \ldots, 2^p$）とする．ここで，$I$ を2値の確率変数の添え字集合（$\forall i \in I$），J を k 個の2値型代表点の候補ベクトルの添え字集合（$\forall j \in J$），U を2値型代表点の添え字集合（$U \subseteq J$，$|U| = k$）とする．このとき，集合 U に対する集合関数 $L(U)$ を，$\boldsymbol{x}_i$ と $\boldsymbol{y}_j$ の適合度を用いて以下のように定義する．

$$L(U) = \sum_{i \in I} P[\boldsymbol{X} = \boldsymbol{x}_i] \max_{j \in U} \frac{p - (\boldsymbol{x}_i - \boldsymbol{y}_j)^T (\boldsymbol{x}_i - \boldsymbol{y}_j)}{p} \tag{1}$$

このとき，2値型代表点の集合として，式 (2) を最適化するような k 個の $\boldsymbol{y}_j$ を2値型代表点の集合 U として与える．

$$\operatorname*{argmax}_{U \subseteq I} \{L(U) \mid |U| = k\} \tag{2}$$

2.2 2値型代表点の劣モジュラ性と貪欲法に基づく近似解法

式 (2) で求められる最適化問題はNP困難な問題であることが知られており，これに対して貪欲法によるアプローチが有効であることが理論的に，実験的に明らかになっている．台集合 V とその部分集合 S，T（$\forall S, T \subseteq V$）に対して，以下の式が成り立つ集合関数 f を劣モジュラ関数と呼ぶ．

$$f(S) + f(T) \geq f(S \cup T) + f(S \cap T) \tag{3}$$

また単調な劣モジュラ関数を貪欲法アルゴリズムに基づいて最適化した場合，$|U| = k$ とすると最適解に対して精度の下限を $\left(1 - \frac{1}{k}\right)^k$ 倍以上と保証した近似解を得られることが知られている [17]．この性質を用いて最適なパラメータの探索が難しいが劣モジュラ性は有しているような関数の最適化に関する研究も数多く行われている [18], [19], [20]．この性質を用いた2値型代表点の貪欲法に基づく近似解法 [1] が提案されており，複数の事例を通して実用上の妥当性に関する議論も行われている．

2.3 2 値型代表点分析法と確率分布

Flury [7] によって定義された代表点は，確率分布における代表点である．与えられたデータはあくまで母集団からのサンプルにすぎず，観測値の裏の確率分布から代表点を抽出すべきといった立場をとっている．よって，データが与えられた場合の代表点は，①まず確率分布を推定する②得られた確率分布の代表点を求めるといった 2 段階のアプローチがとられる [3]．すなわち，具体的な代表点を求めるフェーズは，確率分布を仮定しようとしまいと最適化問題としては同じアプローチをとることになる．

また，データから多変量 2 値の確率分布を推定する際に，対数線形モデルが用いた推定が提案されている [3]．これは，2 値変数それぞれの出現確率を線形モデルで表すための方法であり，代表点を求める際には実現値ごとの生起確率を対数線形モデルによって推定する．また，この方法はデータが十分に存在する場合にのみ有効であることが実験的に調べられており，データの数が少ない場合については，対数線形モデルにあてはめず，データから得られる頻度分布をそのまま確率分布と見なす “ノンパラメトリックな方法” を用いるべきであることが指摘されている [3]．

3. 代表点間の距離を考慮した 2 値型代表点分析法の提案

3.1 提案の着想

2 値型代表点分析には，クラスタリングする側面とクラスタごとに解釈をするための代表点を表す側面があると考えられる．そのうち代表点に興味がある場合において，従来の 2 値型代表点分析ではデータと代表点の適合度の最大化により代表点を決定しているため，データによっては複数の代表点が似たような値をとってしまうことがある [6]．代表点どうしが似たような値をとり，代表点どうしの距離が近くなってしまった場合，それらの代表点の間で差異が小さくなって代表点に特徴がなくなり，解釈がしにくくなってしまう．逆に，代表点どうしの距離が離れれば，代表点において，異なる代表点にはない特徴が生まれ，解釈しやすくなることが期待される．また，代表点どうしの距離が大きくなるとそれぞれの代表点が様々な値をとるようになり，多様なデータに対応することも期待できる．

そこで本研究では，代表点間の距離を大きくするために，式 (2) で表される 2 値型代表点における最適化問題に対してペナルティ項を加えることでデータの偏りに対応した 2 値型代表点を提案する．

3.2 代表点間の距離を考慮した 2 値型代表点分析法の提案

前節の着想に対して，求める 2 値型代表点どうしの距離が大きくなるようなペナルティ項 $R(U)$ を式 (2) に加えた代表点間の距離を考慮した 2 値型代表点を以下のように提案する．

まず，ペナルティ項に対して，代表点どうしの距離を大きくするために，変量ごとの 0 の数と 1 の数が近くなるような集合関数を設定する．ここで，$\boldsymbol{y}_j = \{y_{jh_{(1\le h\le p)}} \in \{0,1\}^p\}$ とし，そのペナルティ項を集合 U に対する集合関数 $R(U)$ として，以下のように表すことにする．

$$R(U) = \sum_{1\le h\le p}\left(\sqrt{\sum_{j\in U}(1-y_{jh})} + \sqrt{\sum_{j\in U} y_{jh}}\right) \tag{4}$$

式 (4) において，$\sum_{j\in U}(1-y_{jh})$ は，2 値型代表点の変量 h における 0 の数の和，$\sum_{j\in U} y_{jh}$ は，2 値型代表点の変量 h における 1 の数の和を表している．ここから，式 (4) の $R(U)$ を最大化させるときの右辺について考える．

任意の変量 h における $\sqrt{\sum_{j\in U}(1-y_{jh})} + \sqrt{\sum_{j\in U} y_{jh}}$ について考える．関数 $f(x)$ を式 (5) のようにおく．

$$f(x) = \sqrt{k-x} + \sqrt{x}\,(0 \le x \le k) \tag{5}$$

この関数 $f(x)$ は，$0 \le x \le k$ において，つねに $f(x) \ge 0$ であるので，2 乗しても大小関係は変わらない．$f(x)$ を 2 乗すると，以下のように変換される．

$$\begin{aligned} f^2(x) &= (\sqrt{k-x} + \sqrt{x})^2 \\ &= k + 2\sqrt{-x^2 + kx} \\ &= k + 2\sqrt{-\left(x - \frac{k}{2}\right)^2 + \frac{k^2}{4}} \end{aligned} \tag{6}$$

式 (6) より，$x = k/2$，つまり，$k - x = x$ に近づくほど $f(x)$ は大きくなる．よって，任意の変量 h における $\sqrt{\sum_{j\in U}(1-y_{jh})} + \sqrt{\sum_{j\in U} y_{jh}}$ は，$\sum_{j\in U}(1-y_{jh}) = \sum_{j\in U} y_{jh}$ に近づくとき，つまり 2 値型代表点の変量 h における 0 の数の和と 1 の数の和が近づくほど，大きな値をとる．また，$R(U)$ は変量 $1 \le h \le p$ における $\sqrt{\sum_{j\in U}(1-y_{jh})} + \sqrt{\sum_{j\in U} y_{jh}}$ の線形結合和である．よって，$R(U)$ を最大化するとき，各変量において 0 の数の和と 1 の数の和が近づく．これにより，変量ごとに代表点の値が散らばるようになり，代表点をばらつかせることが可能となる．

このように求めた $R(U)$ を 2 値型代表点の定義式に加えた関数 $Z(U)$ を最大化する集合 U^* の要素に対応する k 個の $\boldsymbol{y}_j$ を代表点間の距離を考慮した 2 値型代表点分析法として定義する．ただし，λ はハイパーパラメータである．

$$U^* = \underset{U\subseteq J}{\operatorname{argmax}}\{Z(U) = L(U) + \lambda R(U) \mid |U| = k\} \tag{7}$$

ここで，式 (7) における第 1 項 $L(U)$ を最大化させる集合 U を求める問題は，NP 困難な問題であることが知られている．$Z(U)$ を最大化させるときに関しても，ペナルティ項の $R(U)$ を加えても計算量は減少しないため，NP 困難な問題である．そのため，近似解法を用いて計算量を

抑えることが望ましい．一方で，式 (7) で表される目的関数は劣モジュラ性を指摘することができるため（詳細は付録に示す）貪欲法によるアルゴリズムが有効であると考えられる．

3.3 貪欲法に基づく近似解法の提案

本研究では，以下の近似アルゴリズムを提案する．

貪欲法に基づく最適化アルゴリズム

初期値：$U_0 = \emptyset$，$t = 1$
STEP1：$|U_t| = k$ になるまで繰り返す．
(1) $j_t = \underset{j \in J - U_{i-1}}{\operatorname{argmax}} \{Z(U_{t-1} \cup j_t) - Z(U_{t-1})\}$
(2) $U_t = U_{t-1} \cup j_t$
(3) $t = t + 1$ に更新する．
STEP2：$U_t(t = k)$ を出力する．

劣モジュラ関数に対して貪欲法を用いて近似解を求めた場合，その精度の下限は最適解を求めた場合の精度の $\left(1 - \frac{1}{k}\right)^k$ 倍になることが示されている [2]．また，計算量に着目すると，全探索により最適解を求める場合の解の候補の組合せは $O(2^{pk})$ 通りあり，変数が増加すると計算量が爆発的に増加する．一方，貪欲法を用いて近似解を求める場合の解の候補の組合せは $O(k2^p)$ 通りとなり，計算量を 2^p の累乗倍から 2^p の定数倍に減少させることができる．

4. 数値例による実験

4.1 実験設定

提案手法の性質を調べるために，数値例による実験を行う．ここでは，あらかじめ 3 つの代表点を決定した（3 つのクラスタを想定した）下で 4 変量の 2 値データを発生させ，2 値型代表点の貪欲法に基づく近似解法 [1] と提案モデルの貪欲法に基づく近似解法の結果を比較し，推定された 2 値型代表点の分布に対する適合度（$L(U)$）と推定された 2 値型代表点どうしの距離（d^2）

$$d^2 = \frac{1}{2} \sum_{j \in U} \sum_{j' \in U - \{j\}} (\boldsymbol{y}_j - \boldsymbol{y}_{j'})(\boldsymbol{y}_j - \boldsymbol{y}_{j'})^T \qquad (8)$$

によって比較する．以下に詳細な設定を示す．なお，本来であれば，Yamashita ら [3] の提案しているような生成過程に則りデータを発生させることが望ましいが，以下に示すように複雑なクラスタを想定し，クラスタ内でのばらつきを変化させていくような設定を上記の過程で表現することは難しい．そこで，以下に示すような生成の過程に基づき，シミュレーションデータを発生させることにする．

発生されたデータの中で最も大きなサイズのクラスタ（クラスタ 1）の代表点の座標を $(1, 1, 0, 0)$ として，100 個のデータを確率的に発生させる．このとき，1 番目と 2 番目の変数が 1 となる確率を a，3 番目と 4 番目の変数が 1 となる確率を b とし，$(a, b) = (0.9, 0.1), (0.8, 0.2), (0.7, 0.3)$ と設定する．これにより，クラスタ 1 の内部での分布がばらついているようなデータを生成することができる．次に，残りの 2 つの代表点のうち，一方を $(1, 1, 1, 1)$ として固定しすべての変数が 1 になる確率を 0.9 として 10 個のデータを確率的に発生させる（クラスタ 2）．また，他方の代表点として $(1, 0, 0, 0)$，$(0, 0, 0, 0)$，$(0, 0, 0, 1)$，$(0, 0, 1, 1)$ の 4 種類を設定し（クラスタ 1 との代表点の距離がそれぞれ 1，2，3，4 となる），変数が 1 の値をとる確率を 0.9，0 をとる確率を 0.1 として 10 個のデータを発生させる（クラスタ 3）．これにより，クラスタ 3 と 1 との距離を変化させることができ，分布の偏りを調整することができる．この操作を 10000 回繰り返し，代表点を算出した．ただし 2 値分布の推定はノンパラメトリックな方法 [3] を採用した．このとき，d^2，$L(U)$ の平均値を用いて比較する．

4.2 実験結果

得られた結果を**表 1** に示す．まず $k = 3$，すなわち，クラスタのサイズを正しく設定したときに関して，考察を行う．

このとき，以下の傾向を読み取ることができる．

① (a, b) の値が近くなるにつれ，そしてクラスタ 1 とクラスタ 3 の間の距離が大きくなるにつれて従来でも提案でも $L(U)$ の値は小さくなることが分かる．ただし，その値の減少の仕方は提案手法の方が小さいということが読み取れる．これは，クラス内の分布のばらつきが大きいほど従来手法と提案手法のパフォーマンスの差が小さくなることを示唆しているものと考えられる．

② d^2 の値は従来手法では，クラスタ 1 とクラスタ 3 との距離が大きいほど大きな値をとる一方で，提案手法ではつねに最大値をとっていることが分かる．すなわち，提案手法により，距離が大きな代表点を抽出していることが分かる．

上記より，$k = 3$ のとき，かつクラス内の分布のばらつきが大きいときに従来手法と提案手法の $L(U)$ は近い値をとり，さらに d^2 はつねに提案手法により違いが明確な代表点を求めることができていることが分かる．

次に $k = 4$，すなわちクラスタの個数を正しく設定できなかったときに関する考察を行う．

③ (a, b) の値が近くなるほど，そしてクラスタ 1 とクラスタ 3 との距離が大きくなるほど提案手法と従来手法から得られる $L(U)$ の値が近づく．

④ d^2 の値は従来手法が提案手法の結果に近づいていくものの，提案法の d^2 はほとんど変わらない．

③・④の結果より，提案手法によりクラス数を正しく設定したときも，多少異なるクラス数に設定したときも，好ましい結果を得られていることが示唆される．

以上の結果より提案手法の有効性および安定性が指摘さ

表 1 3つのクラスタを想定したシミュレーションデータにおける従来手法と提案手法の性能
Table 1 Performance of the conventional and proposed method for analyzing the simulation data assuming the three clusters.

		$k=3$				$k=4$			
		従来		提案		従来		提案	
(a,b)	クラスタ1との距離	d^2	$L(U)$	d^2	$L(U)$	d^2	$L(U)$	d^2	$L(U)$
(0.9,0.1)	1	4.4644	0.9322	7.9024	0.9105	12.1118	0.9531	18.7578	0.9223
	2	5.0806	0.9201	7.9988	0.9054	13.4657	0.9411	18.4742	0.9214
	3	6.8876	0.9155	8.0000	0.9100	15.5983	0.9374	18.1020	0.9283
	4	7.0302	0.9093	8.0000	0.9155	15.7143	0.9306	17.7627	0.9343
(0.8,0.2)	1	4.2366	0.8845	7.9984	0.8623	9.5936	0.9146	15.5132	0.8734
	2	4.6422	0.8727	7.9998	0.8570	10.3079	0.9027	15.1441	0.8744
	3	6.4724	0.8652	8.0000	0.8629	12.4660	0.8956	15.3130	0.8820
	4	7.2574	0.8574	8.0000	0.8605	13.0557	0.8862	15.6606	0.8832
(0.7,0.3)	1	5.3538	0.8348	8.0000	0.8237	11.0786	0.8704	15.6322	0.8421
	2	6.2198	0.8253	8.0000	0.8187	12.2497	0.8616	15.4019	0.8425
	3	7.5146	0.8241	8.0000	0.8221	13.5237	0.8588	15.6099	0.8480
	4	7.7156	0.8165	8.0000	0.8176	14.3660	0.8505	15.9517	0.8482

れ，提案手法を用いた分析の妥当性が示唆される．

5. 実データによる分析

5.1 マンゴーの官能評価データの分析

5.1.1 分析対象データ [21]

まずマンゴーを購入するにあたって何を重要視するかについてのアンケートデータを使用して，変数が少ない場合のデータにおける提案モデルの適用例を示す．高級フルーツとして知られるマンゴーは，その保存可能期間の短さや大量生産の難しさから，高価な値段で取引がされており，マンゴーのシェア拡大に向けて品種改良や保存方法，輸送方法の開発が行われている [22]．その一連の中で，消費者がどのようなマンゴーを好むのかに関する定量的な調査が必要になり，行ったアンケートデータである．対象は大学生 53 名である．アンケートデータの使用した項目は，「①酸味がある」「②噛み応えがある」「③濃厚」「④みずみずしい」「⑤やわらかい」の 5 個があり，それぞれ 5 段階で評価している（評価 5 が最高評価）．

ここでは提案手法への適用のため，データを 2 値化した．このデータを用いて「マンゴーの好み」を，顧客の傾向ごとに把握するためには，「違いをよく表す」ような代表点に基づく解釈が好ましい．そこで，このデータに対して距離が離れる代表点を推定する．このときの実現値ごとの頻度分布は，**表 2** のとおりである．なお，2 値分布の推定はノンパラメトリックな方法 [3] を採用した．

表 2 2 値のアンケートの頻度分布
Table 2 A distribution of the real binary questionnaire data.

	0	1
①酸味がある	24	30
②噛み応えがある	20	34
③濃厚	20	34
④みずみずしい	24	30
⑤やわらかい	34	20

表 3 得られた 2 値型代表点
Table 3 Result of the analysis with principal points for a binary distribution.

手法＼項目	v1	v2	v3	v4	v5	個数	d^2
提案手法	0	0	0	0	1	7.5	20
	0	0	1	1	1	13.5	
	1	1	0	0	0	14.5	
	1	1	1	1	0	18.5	
従来手法	1	1	1	0	0	7.5	16
	0	0	1	1	1	16.5	
	1	1	0	0	0	15	
	1	1	1	1	0	15	

5.1.2 分析方法

このデータに対し，提案手法を適用してクラスタごとに 2 値の代表点を求める．このとき，ハイパーパラメータである λ は 1 で固定した．また，代表点の個数は 5 個とした．上記のデータに対して提案手法と，従来手法（2 値型代表点分析 [1]）を適用した．その結果として，それぞれの代表点，以下の式 (4.1) で表される代表点どうしの距離の和 d^2，代表点に含まれるデータの個数を以下の**表 3** に記載している．ここで，代表点ごとの個数に関して，データによってはデータからの距離が最短となる代表点が複数個ある場合が存在する．その場合は，その個数を m 個として，$1/m$ 個をそれぞれの代表点の個数にカウントした．

5.1.3 分析結果

提案法を用いて分析した結果を表 3 に示す．表 3 におい

て，それぞれの手法の一番上の行にある代表点（代表点 1）がそれぞれの手法で異なる代表点である．従来の 2 値型代表点分析では，代表点 1 と 3，代表点 1 と 4 の代表点どうしの距離が 1 しか変わらず，似た代表点が求められている．そのため，それらの代表点においてほかの代表点と異なるような特徴が少なくなっている．それに対し，提案手法では，代表点どうしの距離は少なくとも 2 は離れていて，代表点どうしの距離の総和も，提案手法のほうが大きくなっている．

ここで，実際に提案手法によって得られた解釈を以下に示す．まず，提案手法から得られた代表点の解釈を示す．
代表点 1：「やわらかさ」を重視
代表点 2：「濃厚」，「みずみずしい」，「やわらかさ」を重視
代表点 3：「酸味がある」，「噛み応えがある」を重視
代表点 4：「やわらかさ」以外すべて重視

一方の従来手法では，代表点 1 が「酸味がある」，「噛み応えがある」，「濃厚」を重視となっており，代表点 3，4 と似た解釈となっている．以上の結果より，提案手法のほうが代表点それぞれに明確な特徴の違いが出ていて，解釈がしやすいことが分かる．また，クラスタごとにマーケティング施策を考えようとする場合，代表点に明確な特徴の差があると，たとえば「酸味がある」，「噛み応えがある」と答えた人に合うようなマンゴーのブランディングができ，ターゲットを明確に絞った施策ができる．

また，代表点間の距離を考慮した 2 値型代表点分析において，貪欲法を用いた近似アルゴリズムを適用し，近似解を求めた場合と全探索により最適解を求めた場合において，それぞれでかかった CPU 時間と適合度 $L(U)$ の値を以下の**表 4** にまとめた．

表 4 より，貪欲法を用いた近似アルゴリズムを適用した場合，CPU 時間が全探索に比べて $5{,}612 \times 10^{-3}$ 倍になり，また近似解の精度は，最適解の精度と同じ値が得られた．これらのことから，提案手法の妥当性が示唆される．

5.2 新聞におけるアンケートデータの分析

5.2.1 分析対象データ [23]

次に，変数の数が 17 個のアンケートデータの解析例を紹介する．解析するデータは，日刊工業新聞 2010 年 9 月 28 日に掲載されたアンケートデータである．これは，日本 100 社の社長をアンケート対象としており，「今後の日本経済に影響があると考えられる項目」を 15 項目の中から 5 つ選択するものである．17 項目の内容は以下のとおりである．

(1) 個人消費，(2) 為替，(3) 規制緩和，(4) 株価，(5) 税制改革，(6) 設備投資動向，(7) 中国景気，(8) 米国景気，(9) 新興国の景気，(10) 国内の政局，(11) 新技術・新製品開発，(12) 鉄鋼などの素材価格，(13) 原油・石油製品価格，(14) 国際情勢，(15) 製品価格動向．

表 4 CPU 時間と精度の比較
Table 4 Comparison of CPU time and Accuracy.

項目	最適解	近似解
CPU 時間(s)	6.898(s)	0.03871(s)
$L(U)$	0.8481	0.8481

表 5 新聞データの 0 と 1 の分布
Table 5 A distribution of binary variables of newspaper data.

変数	1 の数	0 の数	変数	1 の数	0 の数
(1)	32	65	(9)	19	78
(2)	67	30	(10)	15	82
(3)	12	85	(11)	44	53
(4)	14	83	(12)	28	69
(5)	15	82	(13)	23	74
(6)	59	38	(14)	14	83
(7)	56	41	(15)	27	70
(8)	43	54			

このときの有効回答者数は 94 人であった．また，このアンケートデータでは，選択する項目の数が 5 つと定められているため，とりうる実現値の集合は，1 の数が 5 つ，すなわち $|x| = 5$ となる有限集合を 2 値型代表点の候補とした．このデータでは，従来法（貪欲法による近似解）と提案法（提案法による近似解）の結果を比較し，それぞれの性能について評価した．ここで与えられたデータの変数ごとの分布を**表 5** に示す．なお，2 値分布の推定はノンパラメトリックな方法 [3] を採用した．

5.2.2 分析条件

まず，k の数を決定するために，このときの λ の値を 1 と固定し，$k = 2, \ldots, 10$ のすべての場合において提案法および従来法による代表点を算出し，d^2 および $L(U)$ を計算した．ただし，d^2 は代表点どうしの距離の和として計算される．よって，代表点の数が増加すればするほど，その値が大きく増加してしまう．そこで，この分析では，d^2 に対して，代表点の組合せの数で割り算をした値を修正 d^2 として，$L(U)$ および修正 d^2 の 2 つの観点から最適な k の数を決定する．以下に詳細を示す．

図 1 の結果より，提案手法の $L(U)$ は $k = 2, \ldots, 5$ までは k の数が増加するにつれて大きく改善されているが，$k = 5$ を境に増加が緩やかになっていることが分かる．さらに，従来手法と提案手法の違いが最も小さくなっていることが分かる．また，修正 d^2 の値は $k = 5$ のときに最も良い値が得られている．そこで最も代表点間の距離が大きくなり（違いが明確になり），かつ分布に対するあてはまりが従来手法に最も近い $k = 5$ のときの結果について以降で詳しく検討する．

5.2.3 分析結果

分析結果を**表 6** に示す．まず，従来法では，$v4$，$v5$，$v14$ がすべて同じ値となっており，これらの変数は代表点とし

表 6 従来手法および提案手法で得られた代表点 ($k = 5$)
Table 6 The result of analysis with conventional and proposed method for principal points ($k = 5$)

代表点（従来）	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	個数	割合
1	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	23.33	0.29
2	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	17.75	0.22
3	1	1	0	0	0	1	1	1	0	0	0	0	0	0	0	18.42	0.23
4	0	0	1	0	0	0	1	0	0	1	0	1	1	0	0	7.08	0.09
5	0	1	0	0	0	1	1	1	1	0	0	0	0	0	0	13.42	0.17
代表点（提案）	v1	v2	v3	v4	v5	v6	v7	v8	v9	v10	v11	v12	v13	v14	v15	個数	割合
1	0	1	0	0	0	1	1	1	0	0	1	0	0	0	0	42.83	0.54
2	1	1	0	0	0	0	0	0	0	0	1	0	1	0	1	15.50	0.19
3	1	0	0	0	1	0	0	0	0	0	0	1	1	0	1	7.50	0.09
4	0	0	1	1	0	1	0	0	0	1	0	0	0	0	1	5.33	0.07
5	0	1	0	0	0	0	1	0	1	1	0	0	0	1	0	5.83	0.11

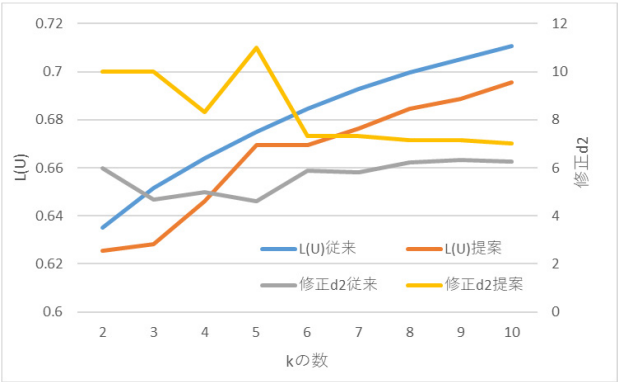

図 1 k の値の変化による $L(U)$ の変化および修正 d^2 の値の変化
Fig. 1 The value of the adjusted d^2 for each number of k.

て解釈されない．一方で，提案手法ではすべての代表点が同じ値をとることがないため，すべての変数が代表点として解釈に含むことができる．1 番目の代表点と 2 番目の代表点は従来手法および提案手法ともに共通となっている．異なる代表点として抽出された 3，4，5 番目の従来法および提案法の結果に着目すると，その代表点に含まれるデータの数は従来手法で得られた代表点に含まれるデータの数よりも少ないことが分かる．これが従来手法と比較して $Z(U)$ を低下させる要因となっているものの，一方で代表点間の距離を増加させる役割を持っている．

次に 1 つ 1 つの代表点に対する解釈を行う．従来手法の代表点を具体的に記述する．

代表点 1：（為替，設備投資動向，中国景気，米国景気，新技術・新製品開発）【高い技術を有し，輸出することに重きをおいている企業】

代表点 2：（個人消費，税制改革，新技術・新製品開発，原油・石油製品価格，製品価格動向）【国内での製造業に力を入れている企業】

代表点 3：（個人消費，税制改革，鉄鋼などの素材価格，原油・石油製品価格，製品価格動向）【資源を扱う国内の製造業に力を入れている企業】

代表点 4：（規制緩和，株価，設備投資動向，国内の政局，製品価格動向）【日本における金融に力を入れている企業】

代表点 5：（為替，中国景気，新興国の景気，国内の政局，国際情勢）【高い技術を有し，輸出に力を入れている企業】

となり，それぞれの企業の名前を参照したものと解釈を比較すると，よく説明ができていることが分かる．すなわち．提案手法により，解釈性に優れ，確率分布へのあてはまりが大きい代表点を抽出することでき，提案法の妥当性が示唆される．

6. 解釈

6.1 データに対する提案手法の適用

本研究で提案する手法は，多変量の 2 値データ，かつ変数ごとにデータの偏りが大きいものがある（0 に偏っている変数や 1 に偏っている変数がある）分布に対して有効な手法である．逆に，分布に偏りがないものに関しては従来手法と提案手法のパフォーマンスは変わらないことが懸念される．

そこで例として，4 章の数値例による実験を基に，代表点を $(1,0,1,1)$，$(1,1,0,0)$，$(0,1,1,0)$，$(0,0,0,1)$ として，各変数に対して代表点が持つ 1 の数と 0 の数が等しくなるように設定し，代表点が 1 となっている変数では 1 をとる確率を 0.9，代表点が 0 となっている変数では 0 をとる確率を 0.9 として代表点ごとに 2 値データを 100 個ずつ発生させる．この状態で得られた 400 個のデータから 4 つの代表点を求め，これを 10,000 回繰り返した場合の確率分布に対するあてはまり $L(U)$ および代表点どうしの距離 d^2 の平均値と分散値を**表 7** に示す．

表 7 従来法および提案法の $L(U)$ および d^2 の平均値と分散値
Table 7 Average and variance values of $L(U)$ and d^2 of conventional and proposed methods.

平均（分散）	$L(U)$	d^2
従来	0.867(0.001)	15.183(0.299)
提案	0.849(0.001)	15.401(0.667)

これらの結果を比較すると，$L(U)$ の値は従来手法の方が値が大きくなった一方で d^2 の値に着目すると，より違いをよく表すような代表点を選択していることが分かる．よってこのような場合においても提案法を用いるメリットがあることが分かる．ただし，計算コストが大きくかかるため分析の際には生データの様子を十分に確認し，必要性を確認したうえで提案手法を用いることが望ましい．

6.2 高次元データに対する適用について

本研究で提案する 2 値型代表点は，データが高次元である場合（高次元の確率分布を仮定する場合）にも定義することができる．ただし，多変量 2 値分布のとりうるすべての値の中から最適な代表点の組合せを探索する必要があるため，目的変数である式 (7) の集合 J の要素数が爆発的に大きくなってしまい，計算量コストの少ない貪欲法であっても最適化のための計算量が膨大になってしまう．よって，このような場合に関しては，データとして発生しているパターンのみに J の要素を限定する，または，対象となるデータの特徴を調べ，変数の数を減らしたうえで最適化を行う必要がある．1 の数や 0 の数の偏りによって変数を限定することは，本来，本研究において対象としているデータの偏りを解消することになる点には注意が必要であると考えられる．

7. おわりに

本研究では，多変量 2 値分布に対するあてはまりの最大化の観点から探索される 2 値型代表点におけるクラスタリングと解釈の 2 つの視点を同時に考慮した新たな 2 値型代表点分析を提案した．さらに，2 値型代表点を求めるための計算量が膨大となってしまうため，効率良く解を求めるための近似解法として解の精度の下限を保証したアルゴリズムを提案した．この目的関数は，劣モジュラ性を持っていることが分かるため，貪欲法を用いたアルゴリズムにより，精度の下限を保証した近似解を得ることができることを示した．さらに，本研究ではシミュレーション実験により本研究における提案手法の妥当性を示したうえで 2 種類の実際のデータに提案手法をあてはめ，分析の具体例を示すとともに有効性を確認した．

今後の課題として 2 点を指摘する．1 点目はハイパーパラメータの設定方法である．今回のモデルにおけるクラス数 k やハイパーパラメータである λ は，現状，結果の比較をすることで，最適な値を決めている．その結果，解釈性やデータへのあてはまりといった評価指標が複数存在し，それらを総合的に考えるという時点で，恣意性が含まれている．よって，それぞれの指標をどのように評価するのかに関する汎用的かつ客観的な方法を開発することが 1 つ目の課題である．

2 つ目は，モデルの拡張である．今回はクラス間の距離を大きくするというポリシのもと定式化を行った．しかしながら，現実のデータ分析に関する問題では，なるべくクラスに含まれる分布の出現確率を均一にしたい，選択される 1 の数を一定にしたい，というようなニーズが考えられる．これに対し，適切な目的変数を設定し，さらに性能の良い最適化アルゴリズムを提案することによって，データ分析における数多くの問題を解決することができるようになるものと思われる．そこで，これらの問題設定とその定式化について，様々な実データの分析をもとに提案していくことも課題としてあげられる．

謝辞 本研究の一部は JSPS 科研費 21K14369 の助成を受けたものです．

参考文献

[1] Yamashita, H. and Suzuki, H.: Heuristic approximation methods for principal points for binary distributions, *Journal of Japan Industrial Management Association*, Vol.65, No.2E, pp.131–141 (2014).
[2] Yamashita, H. and Kawahara, Y.: Principal points analysis via p-median problem for binary data, *Journal of Applied Statistics*, Vol.47, No.7, pp.1282–1297 (2020).
[3] Yamashita, H., Matsuura, S. and Suzuki, H.: Estimation of principal points for a multivariate binary distribution using a log-linear model, *Communications in Statistics-Simulation and Computation*, Vol.46, No.2, pp.1136–1147 (2017).
[4] Yamashita, H. and Suzuki, H.: Heuristic approximation methods for principal points for binary distributions, 日本経営工学会論文誌，Vol.65, No.2E, pp.131–141 (2014).
[5] 山下 遥，鈴木秀男：セール品に注目した顧客の購買行動の解析：2 値データのクラスタリングを考慮したロジスティック回帰分析，オペレーションズ・リサーチ，Vol.60, No.2, pp.81–88 (2015).
[6] Yamashita, H.: An Algorithm for Principal Points Considering External Criterion for Multivariate Binary Distributions, *International Journal of Japan Association for Management Systems*, Vol.10, No.1, pp.75–80 (2018).
[7] Flury, B.: Principal points, *Biometrika*, Vol.77, No.1, pp.33–41 (1990).
[8] Flury, B.: Estimation of principal points, *Journal of the Royal Statistical Society, Series C* (*Applied Statistics*), Vol.42, No.1, pp.139–151 (1993).
[9] Kurata, H.: On principal points for location mixtures of spherically symmetric distributions, *Journal of Statistical Planning and Inference*, Vol.138, No.11, pp.3405–3418 (2008).
[10] Matsuura, S. and Kurata, H.: Principal points of a multivariate mixture distribution, *Journal of Multivariate Analysis*, Vol.102, No.2, pp.213–224 (2011).

[11] Tarpey, T., Li, L. and Flury, B.: Principal points and self-consistent points of elliptical distributions, *Annals of Statistics*, Vol.23, No.1, pp.103–112 (1995).
[12] Shimizu, N. and Mizuta, M.: Functional principal points and functional cluster analysis, *Computational Intelligence Paradigms, Studies in Computational Intelligence*, Jain, L.C. et al. (Eds.), Vol.137, pp.149–165 (2008).
[13] Mease, D., Nair, V.N. and Sudjianto, A.: Selective assembly in manufacturing: statistical issues and optimal binning strategies, *Technometrics*, Vol.46, No.2, pp.165–175 (2004).
[14] Stampfer, E. amd Stadlober, E.: Methods for estimating principal points, *Communications in Statistics—Simulation and Computation*, Vol.31, No.2, pp.261–277 (2002).
[15] Yamashita, H.: A study on principal points for a multivariate binay distribution, 慶応義塾大学 2014 年度博士論文 (2015).
[16] Lovász, L.: Submodular functions and convexity, *Mathematical Programming the State of the Art*, pp.235–257, Springer, Berlin, Heidelberg (1983).
[17] Nemhauser, G.L., Wolsey, L.A. and Fisher, M.L.: An analysis of approximations for maximizing submodular set functions-I, *Mathematical Programming*, Vol.14, No.1, pp.265–294 (1978).
[18] Kawahara, Y., Nagano, K., Tsuda, K. and Bilmes, J.A.: Submodularity cuts and applications, *Advances in Neural Information Processing Systems*, Vol.22 (2009).
[19] Nagano, K., Kawahara, Y. and Iwata, S.: Minimum average cost clustering, *Advances in Neural Information Processing Systems*, Vol.23 (2010).
[20] Shioura, A., Shakhlevich, N.V. and Strusevich, V.A.: Application of submodular optimization to single machine scheduling with controllable processing times subject to release dates and deadlines, *INFORMS Journal on Computing*, Vol.28, No.1, pp.148–161 (2016).
[21] 井上 彩，山下 遥，菊野日出：被験者の消費価値観の違いを考慮した宮古島産マンゴーの官能評価，日本経営工学会論文誌，Vol.71, No.4, pp229–232 (2021).
[22] Siyling Zhang, 山下 遥，菊野日出彦：少人数の官能試験に基づく宮古島産マンゴーの経日劣化の評価に関する研究，日本経営システム学会論文誌，Vol.38, No.3, pp.163–169 (2022).
[23] 河原吉伸，永野清仁：劣モジュラ最適化と機械学習，講談社 (2015).

付　　録

A.1 劣モジュラ性の証明

近似解法を提案する際に式 (7) の劣モジュラ性を示すことができれば Yamashita ら [1] と同様に貪欲法に基づく精度の下限値を保証した近似解を現実的な計算量によって求めることができる．

A.1.1 命題

U を集合，λ を実数とする．このとき，下式で表される集合 U に対する集合関数 $Z(U) = L(U) + \lambda R(U)$ は劣モジュラ関数である．

以下，この命題を示すために 1 つの補題を示す．

A.1.2 補題

集合 U に対し，その要素 j に対応する 2 値ベクトルを $\boldsymbol{y}_j$ とする．また，そのベクトルの要素を $\boldsymbol{y}_j = \{y_{jh_{(1\leq h\leq p)}} \in \{0,1\}^p\}$ とする．このとき，集合関数 $\sqrt{\sum_{j\in U}(1-y_{jh})}$, $\sqrt{\sum_{j\in U} y_{jh}}$ はいずれも劣モジュラ関数である．

〈証明〉

$h(x) = \sqrt{x}$ $(x \in \mathbb{R} \mid x \geq 0)$ は単調増加関数であり，かつ補凹関数であるので，$0 \leq a_h < b_h$，$t_h \geq 0$ となる $a_h, b_h, t_h \in \mathbb{R}$ に対して，以下の式が成り立つ．

$$\sqrt{a_h + t_h} - \sqrt{a}_h \geq \sqrt{b_h + t_h} - \sqrt{b_h} \tag{A.1}$$

ここで，$T \subseteq U \subseteq J$，$i \in J - U$ となる集合，要素に対し，a_h，b_h，t_h を以下の式のようにおく．

$$a_h = \sum_{j\in T}(1-y_{jh}) \tag{A.2}$$

$$b_h = \sum_{j\in U}(1-y_{jh}) \tag{A.3}$$

$$t_h = 1 - y_{ih} \tag{A.4}$$

このとき，$\sum_{j\in T}(1-y_{jh})$ と $\sum_{j\in U}(1-y_{jh})$ に対して，下式が成り立つことに注意する．

$$\sum_{j\in T}(1-y_{jh}) \leq \sum_{j\in U}(1-y_{jh}) \tag{A.5}$$

この式 (A.2) から式 (A.5) を式 (A.1) に代入すると，以下の式が成り立つ．

$$\sqrt{\sum_{j\in T}(1-y_{jh})+1-y_{ih}} - \sqrt{\sum_{j\in T}(1-y_{jh})} \geq \sqrt{\sum_{j\in U}(1-y_{jh})+1-y_{ih}} - \sqrt{\sum_{j\in U}(1-y_{jh})} \tag{A.6}$$

$$\sqrt{\sum_{j\in T\cup\{i\}}(1-y_{jh})} - \sqrt{\sum_{j\in T}(1-y_{jh})} \geq \sqrt{\sum_{j\in U\cup\{i\}}(1-y_{jh})} - \sqrt{\sum_{j\in U}(1-y_{jh})} \tag{A.7}$$

この式 (A.7) は式 (3) より，集合関数 $\sqrt{\sum_{j\in U}(1-y_{jh})}$ が劣モジュラ関数であることを示している．

同様に，$\sqrt{\sum_{j\in U} y_{jh}}$ についても劣モジュラ関数であることが分かる．よって，補題は示された．

以下 1 つの補題より，命題が成立することを示す．まず，式 (4) の $R(U)$ に関して考える．任意の変量 h における $\sqrt{\sum_{j\in U}(1-y_{jh})}$，$\sqrt{\sum_{j\in U} y_{jh}}$ は，補題よりどちらも劣モジュラ関数である．ここで，劣モジュラ関数の和や実数倍からできた集合関数は劣モジュラ関数であることが知られている [23]．よって，$\sqrt{\sum_{j\in U}(1-y_{jh})} + \sqrt{\sum_{j\in U} y_{jh}}$ も劣モジュラ関数である．また，$R(U)$ は変量を $1 \leq h \leq p$

とした $\sqrt{\sum_{j\in U}(1-y_{jh})}+\sqrt{\sum_{j\in U}y_{jh}}$ の線形結合和であるので，$R(U)$ は劣モジュラ関数である．

次に，$Z(U)=L(U)+\lambda R(U)$ について考える．$\lambda R(U)$ は劣モジュラ関数 $R(U)$ の実数倍となっているので，劣モジュラ関数である．また，$L(U)$ は劣モジュラ関数である．よって，$Z(U)$ は，2 つの劣モジュラ関数 $L(U)$ と $\lambda R(U)$ の和でできた集合関数であるので，$Z(U)$ は劣モジュラ関数である．これにより，命題は示された．

竹本 真悟

2020 年上智大学理工学部情報理工学科卒業．現在，東京工業大学情報理工学部在学中．グラフを取り入れたハイパースペクトル画像復元に関する研究に従事．

山下 遥

2010 年東京理科大学理工学部経営工学科卒業．2012 年慶應義塾大学大学院理工学研究科解放環境科学専攻博士前期課程修了．博士（工学）取得．2015 年同大学大学院理工学研究科解放環境科学専攻博士後期課程修了．2015 年早稲田大学理工学術院創造理工学部経営システム工学科助手．2017 年上智大学理工学部情報理工学科助教．2021 年同大学同学部准教授．機械学習のアプローチに基づくビジネスアナリティクスモデルの構築，統計的手法に基づく農業データの分析，深層学習に基づくスポーツアナリティクスなどに興味を持つ．経営工学会，品質管理学会，応用統計学会の各会員．