CA-005

アンケートデータを対象とした傾向抽出手法と評価
Component Extraction and Evaluation for Questionnaire Data

岡本 大輝† 後藤 淳†
Hiroki Okamoto Jun Goto

1. はじめに

情報技術の進歩に伴い，大規模なデータの分析技術や機械学習による高精度の予測技術が注目を浴びて久しい．一方で，大規模で安定的なデータを取得できないような社会活動や業態も存在する．たとえば短期的な意識調査やマーケティングなどにおいては，収集されるデータは過去のデータとの連続性が乏しく小規模であることが一般的であり，昨今研究されているデータ分析の技術を活かせる場面は少ない．さらに，実際の調査やマーケティングでは分析の結果を解釈して後の活動に活かすことを目的とするため，高精度だが解釈することが困難であるような機械学習による分析・予測技術は馴染まない側面がある．NHK においても，番組制作や世論調査においてアンケート調査を行うことが少なくないが，しばしば同様の問題に直面する．

そこで本稿では，回答者(サンプル)が数千程度となるアンケートデータを対象とすることを想定して，特定の性年代や回答結果などの条件に合致する回答者(以下，対象群)とそれ以外の回答者(以下，非対象群)の回答傾向を比較し，対象群の回答者が有する特定の傾向を，重みをつけた設問のセットの形で視認性高く表現することを目指す．具体的には，アンケートの回答をもとに回答者を多次元空間へマッピングし，主成分分析，マハラノビス距離空間の生成等を組み合わせて，対象群と非対象群との回答傾向が有意に異なる設問のセットをベクトルの形で得る．さらに，統計的有意性を保ちつつ設問セットに含まれる設問の数を少なくする．抽出した設問セットと回答傾向，設問数について，尤度や誤差の観点も交えて議論する．

2. 関連研究

GPU を用いた大規模並列演算等によって，近年の機械学習はその予測精度，扱うデータの量ともに飛躍的に向上している．しかしながら，高度で複雑な計算は人間には直感的に理解できないものであり，説明可能性を求める声も数多く出ている[1]．既存の機械学習，特にディープラーニングの複雑なモデルの入出力からルールを獲得する手法や[2]，初めから可読性の高い決定木やランダムフォレストを深化させた手法などが研究されている[3].

一方，上述した複雑で大規模な計算による予測や分析は，それに見合うだけの大規模なデータを必要とする．しかし現実には，時間や労力の制約から十分な量のデータが集められない事業もあり，意識調査やマーケティングでしばしば実施されるアンケートのデータもそれにあたる．アンケートデータに対して数理的なモデルを適用し，特徴抽出をする試みもあるが[4]，大規模データに対して高度な機械学習を適用する試みと比べると圧倒的に少ない．また，アンケート調査はその結果を吟味，解釈することを目的として実施されることが多く，データのクラスタリングや欠損値の予測精度は高いものの，分析や解釈が劣後しがちな機械学習は，アンケートデータ分析との相性が悪いとも指摘されている．現在も社会学系を中心に，統計的推論によるアンケートデータ分析がしばしば行われている[5].

本稿では，機械学習を用いずに主に統計的な理論に基づいた手法を用いて，アンケートの回答者のうち特定の条件を満たす対象群の回答傾向を抽出するタスクに取り組む．タスクの内容について述べるに先立ち，関連する統計的技法や線形代数的な性質について述べる．

2.1 主成分分析(PCA)

アンケートのデータは多数の回答者(サンプル)と複数の設問を有する，ある種の多次元的なデータとみなすことができる．多次元データの相関を分析する手法として，最も一般的なものの一つに主成分分析 (PCA: Principal Component Analysis) がある．

PCA を行うにあたり，まず i 番目の観測サンプル ($1 \leq i \leq N$)，j 番目の特徴量 ($1 \leq j \leq K$) を示すデータ値 $x_{i,j}$ を用いて，a 行 b 列成分 $s_{a,b}$ を

$$s_{a,b} = \frac{1}{N}\sum_{i=1}^{N}\left(x_{i,a} - \bar{x}_a\right)\left(x_{i,b} - \bar{x}_b\right) \qquad (1)$$

とする $K \times K$ の分散共分散行列 Σ を求める．なお，$\bar{x}_a$ は a 番目の特徴量の平均値を表す.

PCA は分散共分散行列 Σ を固有値分解し，固有値と固有ベクトルを求める操作である．幾何的には，K 次元空間に分布するデータの分散が最も大きくなる方向が固有ベクトルとして求められ，分散の値が固有値と対応する．固有ベクトルの方向に強い相関をもってデータが分布しており，固有値が相関の強さを示すと言い換えることができる．PCA では最大 K 組の固有値と固有ベクトルの対が求められるが，いずれの固有ベクトルも互いに直交している．

2.2 マハラノビス距離

各特徴量がそれぞれ正規分布に従う多次元データにおいて，特徴量同士の相関を考慮しつつ，新たに観測されたサンプルの珍しさを評価する指標としてマハラノビス距離がある．マハラノビス距離は，異常値検出などで広く用いられている．

2.2.1 マハラノビス距離の定義

あらかじめ十分な数のサンプルを観測して，数式(1)によって求められる $K \times K$ の分散共分散行列 Σ から逆行列 Σ^{-1} を計算する．新たに観測されたサンプルの特徴量を成分とする行ベクトル x のマハラノビス距離 D^2 は

$$D^2 = (x - \bar{x})\Sigma^{-1}(x - \bar{x})^T \qquad (2)$$

で定義される．なお，$\bar{x}$ は全てのサンプルのデータ値の平均を成分とする K 次元の行ベクトルである．マハラノビス距離が大きいほど，観測された事象が珍しいと見做される．

† NHK 放送技術研究所
NHK Science and Technology Research Laboratories

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

2.2.2 擬似逆行列を用いたマハラノビス距離の計算

分散共分散行列 Σ は退化(階数落ち)している場合がある．退化している場合，直接的に逆行列 Σ^{-1} を求めることはできないが，2.1 節の固有値分解で求められる固有値と固有ベクトルを用いて擬似逆行列を定義することができる．

固有値分解を一般化した特異値分解によって，分散共分散行列 Σ の固有値を対角成分とする $K \times K$ の対角行列 Λ と，固有ベクトルを行にとる実ユニタリ行列 U,V によって $\Sigma = U^T \Lambda V$ と分解できる．このとき，Σ が対称行列であることから U^T, V はいずれも正規直交行列であり，$U = V$ と定めることができる．なお，Σ に退化が生じてランクが $M\ (\leq K)$ となっている場合には，Λ の右下の $K - M$ 成分が零詰めされているものとする．

分散共分散行列 Σ は半正定値行列なので，固有値は全て非負の値となる．Σ の階数すなわち非零の固有値の数を M として，各固有値の逆数の正の平方根を対角成分とし，退化している場合に右下を零詰めした $K \times K$ の実対角行列 $\tilde{\Lambda}^{-0.5}$ と，その自乗である $\tilde{\Lambda}^{-1}$ を定義する．このとき，行列積 $\Lambda\,\tilde{\Lambda}^{-1}$ は $M \times M$ の単位行列の右下を零詰めして拡張した $K \times K$ の対角行列となる．

ここで，以下の 2 つの行列 F と $\tilde{\Sigma}$ を定める．

$$F = \tilde{\Lambda}^{-0.5} V \tag{3}$$

$$\tilde{\Sigma} = F^T F = V^T \tilde{\Lambda}^{-1} V \tag{4}$$

$\tilde{\Sigma}$ は Σ に対して Moor-Penrose の擬似逆行列の定義を満たす．そこで，式(2)の Σ^{-1} を $\tilde{\Sigma}$ に置き換えることで，退化した部分空間内でのマハラノビス距離を算出できる．これにより，分散共分散行列 Σ が退化している場合でもマハラノビス距離を正確に算出することが可能となる．また，行列 F による x_i の線形変換を y_i を

$$y_i = (x_i - \bar{x})\, F^T \tag{5}$$

とすると，L2 ノルム $||\, y_i \,||_2$ は

$$\begin{aligned} ||\, y_i \,||_2 &= y_i\, y_i{}^T \\ &= (x_i - \bar{x})\, V^T \tilde{\Lambda}^{-1} V\, (x_i - \bar{x})^T \\ &= (x_i - \bar{x})\, \tilde{\Sigma}\, (x_i - \bar{x})^T \end{aligned} \tag{6}$$

となり，マハラノビス距離と一致する．すなわち y_i は，マハラノビス距離空間における x_i の線形写像ベクトルであると捉えることができる．

2.3 多変量正規分布

それぞれ正規分布に従う確率変数 $\{X_1, X_2, \ldots, X_K\}$ を有し，その線形結合が正規分布に従う確率分布を多変量正規分布と呼ぶ．確率変数の観測値を成分とする行ベクトルを x，平均値 $\{\mu_1, \mu_2, \ldots, \mu_K\}$ を成分とする行ベクトルを μ，確率変数の分散共分散行列を Σ，その擬似逆行列を $\tilde{\Sigma}$，階数を M とすると，多変量正規分布の密度分布 $P(x, \tilde{\Sigma})$ は

$$P(x, \Sigma) = \frac{1}{\sqrt{(2\pi)^M |\Sigma|_{deg}}} exp\left(-\frac{(x - \mu)\tilde{\Sigma}(x - \mu)^T}{2}\right) \tag{7}$$

となる．なお，$|\Sigma|_{deg}$ は Σ の正の固有値の総乗である．

多変量正規分布の密度関数には，観測値 x のマハラノビス距離に相当する $(x - \mu)\tilde{\Sigma}(x - \mu)^T$ が含まれている．これは観測値 x の対数尤度 $\ln(P(x, \Sigma))$ を -2 倍して定数を調整したものである．すなわちマハラノビス距離は，多変量正規分布を前提とした確率モデルにおいて，観測値の対数尤度の正負を入れ替えた値と等価であると見做せる．

2.4 赤池情報量規準(AIC)

本稿で取り上げるタスクと直接的な関係は薄いが，確率モデルの対数尤度とパラメータ数との関係を内包する指標として赤池情報量規準(AIC)がある．確率モデルの生成に用いる開発データに含まれるサンプルの数を N_{dev}，i 番目の開発用サンプルを x_i ($\in X_{dev}$)，生成した最尤モデルに用いるパラメータ数を K，パラメータベクトルを $\hat{\theta}_K$，最尤確率モデルを $P(X_{dev}; \hat{\theta}_K)$ とすると，最大対数尤度 $L(X_{dev}, \hat{\theta}_K)$ と AIC は以下の式のように計算される．

$$L(X_{dev}, \hat{\theta}_K) = \sum_{i=1}^{N_{dev}} ln\, P(x_i; \hat{\theta}_K) \tag{8}$$

$$AIC = -2L(X_{dev}, \hat{\theta}_K) + 2K \tag{9}$$

AIC が小さいほど，すなわち対数尤度が大きくパラメータが少ないほど，生成された確率モデルは良いモデルであるとされる．なお，確率モデル $P(X_{dev}; \hat{\theta}_K)$ は開発データ X_{dev} から作られているため，最大対数尤度 $L(X_{dev}, \hat{\theta}_K)$ と，真の確率分布に従う実際の観測データ X_{obs} をモデルに当て嵌めた際に得られる対数尤度の期待値 $L(X_{obs}, \hat{\theta}_K)$ との間には誤差が生じる．すなわち AIC は，最大対数尤度と実データの対数尤度との誤差の期待値が，確率モデルのパラメータ数に依存するということを示している．ここで，観測データの集合を X_{obs}，観測データの数を N_{obs} とすると，N_{dev} と N_{obs} が十分大きいときには

$$L(X_{dev}, \hat{\theta}_K) - \frac{N_{dev}}{N_{obs}} L(X_{obs}, \hat{\theta}_K) \approx K \tag{10}$$

という関係が成り立つ．

3. 提案する設問セット生成タスク

本稿ではアンケートデータを分析し，特定の条件に合致する回答者群である対象群の回答傾向と，それ以外の非対象群の回答傾向がどのように違うのか，重みのついた設問のセットの形で提示することを目的としている．さらに，セットに含まれる設問の数を少なくすることで，分析者の視認性や解釈性を高めることも目的とする．

そこで，まずアンケートに含まれる設問の種類と回答分布に応じて，回答者，すなわちデータサンプルを多次元空間にマッピングするルールを作成する．次に，ルールに基づいて生成された多次元空間を線形変換して，非対象群の分布が標準化されたマハラノビス距離空間を生成する．さらに，マハラノビス距離空間における対象群の回答傾向を示すベクトルの統計的有意性を検定し，有意性を保つ範囲でもとのアンケートデータから設問や選択肢を減らす．

これらのマッピング・多次元空間生成・検定・設問削減の操作を繰り返すことで，対象群に特徴的かつ統計的に有意な傾向を示す設問セットを，設問数が少ない形で求める．

3.1 設問の種類に応じたマッピングルール生成

アンケートの回答を数値で表現するために，まずアンケートの設問を以下のように分類する．

- 選択肢から 1 つを選んで回答する設問(択一式)

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

- 選択肢から複数選んで回答する設問(複数回答式)
- 数値を入力する設問(数値入力式)
- 回答者の裁量で記述する設問(記述式)

本稿では上記のうち，択一式と複数回答式の設問に限って分析を行う．いずれも回答者の自由度が低く，回答傾向や回答の分布にノイズが含まれにくいと考えられる．

3.1.1 程度を問う形式の択一式設問

ある事項について，回答者の状態や考えが合致するか否かを問い，「とてもあてはまる」「ややあてはまる」「あまりあてはまらない」「全くあてはまらない」等，程度を示す選択肢の中から回答者の裁量で一つを選択する形式の設問である．アンケート調査においては，程度の順に並べ替えられて回答者に提示されることが一般的である．

回答者の真の状態や客観的事実に基づき，ノイズなく最も妥当な選択肢を回答者が選択することが理想であるが，設問の掲載順序や選択肢の文面，回答者の言語感覚や性格等によって，回答にノイズが生じることが知られている．そこで，ノイズを含めた真の確率変数に従って回答者の分布が正規分布的になり，累積分布と回答分布が一致していると仮定して，1 設問につき 1 次元を割り当てて回答者の最尤推定値を計算し，その次元における変位とする．

程度を示す R 種類の選択肢に対して，程度の順に R 以下の自然数を割り当てる．整数 r ($0 \leq r \leq R$) に対応する選択肢を選択した回答者の数を h_r とする．なお，対応する選択肢が存在しないため $h_0 = 0$ とする．このとき，全ての回答者の数を N，整数 r 以下の数値が割り当てられた選択肢を選んだ回答者の割合を H_r とすると，

$$H_r = \sum_{s=0}^{r} \frac{h_s}{N} \tag{11}$$

$$H_R = \sum_{s=0}^{R} \frac{h_s}{N} = 1 \tag{12}$$

となる．さらに，誤差逆関数 erf^{-1} を用いて

$$B_r = erf^{-1}(2H_r - 1) \tag{13}$$

として，整数 r が割り当てられた選択肢を選択した回答者の座標 x_r ($1 \leq r \leq R$) を

$$x_r = \int_{\sqrt{2}B_{r-1}}^{\sqrt{2}B_r} z e^{-z^2/2}\, dz \Bigg/ \int_{\sqrt{2}B_{r-1}}^{\sqrt{2}B_r} e^{-z^2/2}\, dz = \frac{N\left(e^{-B_{r-1}{}^2} - e^{-B_r{}^2}\right)}{\sqrt{2\pi}\, h_r} \tag{14}$$

とする．これは，1 つの確率変数 Z に従って回答者が標準正規分布していると仮定した際に，整数 r が割り当てられた選択肢を選択した回答者の Z の期待値，すなわち最尤推定値を表す(図 1).

3.1.2 選択肢が連続的でない択一式または複数回答式設問

回答者の状態や考えについて，「最も当てはまるものを一つ / 全て答える」ように問う形式の設問である．選択肢は， 3.1.1 節で示したように程度を問う連続的な内容ではなく，選択肢ごとに独立した内容となることが多い．たとえば，居住地を問う設問等が該当する．

1 つの設問に与えられた選択肢のそれぞれについて，選ばれたか否かの 2 値に変換し，選択肢の数と同じ数の次元にマッピングする．たとえば，居住地を問う設問で 47 都道府県の選択肢が示されるのであれば，47 次元の部分空間が生成される．各次元の変位の計算には式 (14) を用いて，選択肢の数 R をそれぞれ 2 として計算する．

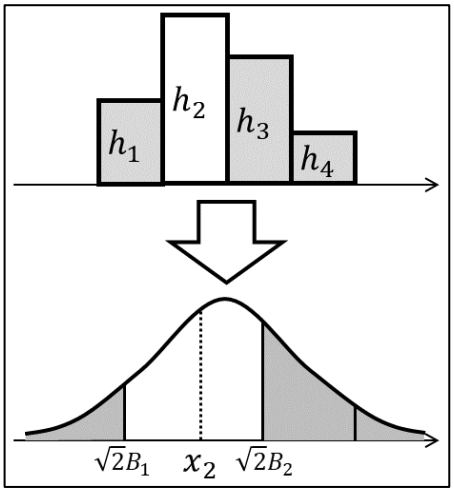

図 1　回答分布と座標の関係

3.2 多次元空間の生成

3.1 節で，設問の種類に応じて次元を生成し，回答分布をもとに回答者データの座標を定めるマッピングルールを定めた．本稿では，対象群と非対象群の回答傾向の違いを多次元空間中のベクトルの形で抽出することを目的とするが，本節ではアンケートデータ分析に用いる 2 つの多次元空間の生成方法について述べる．

3.2.1 設問空間の生成

3.1 節で定めた手順に従って，データ集合 A を用いてデータサンプルをマッピングするルールを作り，データ集合 B を多次元空間にマッピングする操作を $M(B;\, A)$ と表記する．ここで，対象群のデータ集合を $D^{(pos)}$，非対象群のデータ集合を $D^{(neg)}$ とする．$D^{(neg)}$ のみを用いてマッピングルールを作り，対象群と非対象群をそれぞれマッピングした後のデータ集合 $X^{(pos)}$，$X^{(neg)}$ を

$$X^{(pos)} = M\left(D^{(pos)};\, D^{(neg)}\right) \tag{15}$$

$$X^{(neg)} = M\left(D^{(neg)};\, D^{(neg)}\right) \tag{16}$$

とする．これにより，確率変数がとり得る値が離散的であるものの，非対象群が多変量正規分布様に分布する多次元空間(以下，設問空間)が作られる．本稿では一貫して，上記の式(15)と式(16)のように非対象群のデータ $D^{(neg)}$ のみを用いて，設問空間へのマッピングルールを作るとする．

3.2.2 マハラノビス距離空間の生成

設問空間は，1 つの次元がもとのアンケートの 1 つの設問，または 1 つの選択肢(以下，設問等)に対応している．さらに，最少で 2 値しかとらない極端な離散的分布ではあるものの， $X^{(neg)}$ は各次元について正規分布的な分布となっているため，2.3 節で述べた多変量正規分布としての表現を取り入れる．すなわち，$X^{(neg)}$ に含まれる i 番目のデータを表す行ベクトルを $x_i^{(neg)}$ ($1 \leq i \leq N^{(neg)}$) として，分散共分散行列 $\Sigma^{(neg)}$ を

$$\Sigma^{(neg)} = \frac{1}{N^{(neg)}} \sum_{i=1}^{N^{(neg)}} \left\{x_i^{(neg)}\right\}^T x_i^{(neg)} \tag{17}$$

とする．なお，式(14)および式(16)の定義から，$X^{(neg)}$ は全ての次元で平均値が 0 となるため，式(17)では平均値の減算を省略している．

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

さらに，2.2.2 節で示した行列の特異値分解および擬似逆行列の生成に倣い，$\Sigma^{(neg)}$ からマハラノビス距離空間への線形写像行列 $F^{(neg)}$ と，自身の転置との積である擬似逆行列 $\tilde{\Sigma}^{(neg)}$ を計算する．線形写像行列 $F^{(neg)}$ によってマハラノビス距離空間に写像された非対象群のデータ $D^{(neg)}$ は，離散的な分布ではあるものの，あらゆる方向に標準正規分布様に分布している．

3.3 設問等の選別と削減

3.2 節で設問等が 1 つの次元に対応した設問空間と，その線形変換で非対象群が多変量標準正規分布様に分布するマハラノビス距離空間の生成方法について述べた．本節では，マハラノビス距離空間に対象群をマッピングし，対象群の分布特徴ベクトルを元の設問空間に逆変換して設問等の重みを評価し，特徴に寄与しない設問等を削減する手順について述べる．

3.3.1 削減する設問の選択

式(15)によって得られた対象群のマッピング後データ $X^{(pos)}$ のうち，j 番目のサンプル ($1 \leq j \leq N^{(pos)}$) のデータを表す行ベクトル $x_j^{(pos)}$ を線形写像行列 $F^{(neg)}$ によってマハラノビス距離空間に写像した位置ベクトル $y_j^{(pos)}$ は，

$$y_j^{(pos)} = x_j^{(pos)}\left\{F^{(neg)}\right\}^T \tag{18}$$

となる．このマハラノビス距離空間中で，何らかの手法により対象群の分布特徴を示す正規化された行ベクトル c を得るとする．マハラノビス距離空間中の位置ベクトル $y_j^{(pos)}$ を，ベクトル c とその直交成分に分解する際に，ベクトル c に乗じるべき係数は位置ベクトル $y_j^{(pos)}$ との内積に相当するので

$$y_j^{(pos)}c^T = x_j^{(pos)}\left\{c\,F^{(neg)}\right\}^T \tag{19}$$

と計算できる．この値は，位置ベクトル $y_j^{(pos)}$ における分布特徴ベクトル c の含有量であると換言できる．

ここで行列積 $c\,F^{(neg)}$ は，設問空間と同じ次元数の成分を持つ行ベクトルとなる．このベクトルは，任意のサンプルに対して分布特徴ベクトル c の含有量を計算するために，設問空間に属する特徴量にそれぞれ乗じるべき重みのベクトルであると解釈できる．すなわち，行列積 $c\,F^{(neg)}$ で絶対値の大きな成分の次元に対応する設問等が，分布特徴に大きく作用する設問等であり，成分の符号が設問等の回答傾向に対応していると見做すことができる．

翻って，行列積 $c\,F^{(neg)}$ の成分のうち絶対値の小さなものは，成分の次元と対応する設問等が分布特徴に影響しないと見做せる．そこで，その設問等を削減対象とする．

3.3.2 統計的有意性の検定による削減停止

式(17)で述べたマハラノビス距離空間では，非対象群のデータ $D^{(neg)}$ が多変量標準正規分布様に分布するが，同じ空間にマッピングされた対象群のデータ集合 $D^{(pos)}$ の分布との間に有意な差があるか，統計的検定により検証する．具体的には，3.3.1 節で定めたベクトル c の方向に限り，すなわち特徴成分量 $y_j^{(pos)}c^T$ について，全ての対象群のデータの平均値が標準正規分布の平均値である 0 と有意に異なるかを検定する．有意差があれば，ベクトル c は非対象群と比較して有意な対象群の特徴を示していることになる．

平均値の差の検定には Z 検定を用いる．非対象群のデータ $D^{(neg)}$ を母集団，対象群のデータ $D^{(pos)}$ を標本として扱って検定する．非対象群の平均と標準偏差をそれぞれ $\mu^{(neg)}$ ， $\sigma^{(neg)}$ とすると，検定統計量 z は

$$z = \frac{1}{\sigma^{(neg)}\sqrt{N^{(pos)}}}\sum_{j=1}^{N^{(pos)}}\left(y_j^{(pos)}c^T - \mu^{(neg)}\right)$$
$$= \frac{1}{\sqrt{N^{(pos)}}}\sum_{j=1}^{N^{(pos)}} y_j^{(pos)}c^T \tag{20}$$

となる．この検定統計量 z を標準正規分布に適用し，両側確率によって『母集団と標本は平均値が等しい』とする仮説を棄却できるか検定する．棄却できる場合は，3.3.1 節で述べた選択方法によって削減する設問等を選び，もとのデータから該当する設問のデータを削除する．その後，設問等が削減されたアンケートデータに対して，3.1 節で述べたマッピング以降の手順を再び適用し，特徴ベクトルを生成する．

なお，多変量正規分布が想定される環境において，平均ではなく分散の差の有意性を検定する際には χ^2 検定を用いることが一般的である．しかしながら，本タスクで検定の対象となる特徴成分量 $y_j^{(pos)}c^T$ はマハラノビス距離空間の各次元の重み付き和であり，各次元について正規分布していることが想定されるものの，重み付き和の分散値の分布は χ^2 分布にはならない．したがって，本稿のタスクでは Z 検定のみを用いてベクトルの有意性を検定し，設問等の削減の継続と終了を判断する．

4. 実験設定

3 章では，アンケートデータの設問と回答から多次元空間を生成し，対象群の特徴を示す多次元ベクトルの統計的有意性を保ちつつ，設問等を削減して視認性の高い設問セットを生成するタスクについて述べた．本章では，実際にタスクを適用して行った実験の設定について述べる．

4.1 データセット

アンケートの回答データを模して人工的に作成した疑似データと，インターネット上で実際に実施したアンケートの回答データを用いる．疑似データによってベクトル生成手法ごとの特性を分析し，アンケートデータで検証する．

4.1.1 疑似データ

対象群と非対象群の回答者がそれぞれ 1,500 サンプル，設問等が 100 項目あるアンケートを想定する．対象群の 1,500 サンプルはさらに 3 つの回答者グループに分割し，[対象群 1]，[対象群 2]，[対象群 3]として，それぞれ回答傾向が異なるものとする．設問等は全て ”0” または ”1” をとる 2 択として，設問および各回答者グループのそれぞれに回答確率分布を設定する．回答確率分布は 5 つの類型に分けられ，それぞれ LMNQR のアルファベット 1 文字ずつを割り当てる．類型ごとの確率分布を表 1 に示す．

まず，対象群全体と非対象群で回答分布が異なるデータ(SIM_BIAS)を生成する．また，3 つの対象群グループのそれぞれで回答傾向が異なるものの，対象群全体で回答確率分布が非対象群と等しくなるように調整したデータ(SIM_ROUND)も生成する．100 項目の設問等のうち 12 項目で回答確率分布を調整する．回答確率分布の類型を回答者グループの種類ごとに組み合わせて設問等を構成し，4 文字のアルファベットで表記する例を表 2 に示す．

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

なお，SIM_ROUND は全体の回答分布が対象群と非対象群で等しいため，3.3 節で定めた設問削減の手順の中の Z 検定では有意差を示さない．すなわち，設問削減がすぐに停止してしまうため，設問等の数が 50 を下回るまでは有意性の有無によらず設問削減を繰り返すとする．

表 1：疑似データの設問等の回答確率分布の類型

	L	M	N	Q	R
“0”をとる確率	0.9	0.7	0.5	0.3	0.1
“1”をとる確率	0.1	0.3	0.5	0.7	0.9

表 2：疑似データの設問構成の例

疑似データ共通の設問 (88 項目)	LLLL, MMMM, NNNN, … (※枝番をつけて区別する)
SIM_BIAS 固有の設問 (12 項目)	LRRR, RLRR, RRLR, RLLL, LRLL, LLRL, …
SIM_ROUND 固有の設問 (12 項目)	LQQN, QLQN, QQLN, MRRQ, RMRQ, RRMQ, …

(左から [対象群 1][対象群 2][対象群 3][非対象群] の 4 文字表記)

4.1.2 アンケートデータ

2020 年 12 月 10 日から 11 日にインターネットを用いて実施したアンケートデータ(QUESTIONNAIRE)を用いる．設問のテーマは食生活，オリンピックやスポーツとの接触，普段利用するメディア，環境問題に対する意識など多岐に渡る．回答者は 3,094 名，対象群の条件は『東京オリンピックは楽しみですか。』の問いに『とても楽しみ』または『どちらかというと楽しみ』と答えた 1,240 名とする．択一式と複数回答式のすべての設問を特徴量として，88 特徴量でタスクを適用する．

4.2 回答傾向ベクトルの生成手法

3.3.1 節で定めた対象群の回答分布特徴を示す行ベクトル c の生成方法について述べる．ベクトルの生成にはいずれも開発セットを用いて，生成後には 3.3 節で述べたタスクに従って設問等の削減を行う．さらに，最終的に設問等が削減されたベクトルの成分を元の設問空間から削減して空間を退化させ，退化した設問空間をもとに再びマハラノビス距離空間を生成し，回答傾向ベクトルの生成を行う．

4.2.1 L2 ノルム最大化ベクトル

3.2.2 節で述べたマハラノビス距離空間への変換行列 $F^{(neg)}$ を開発セットの非対象群のデータから生成し，サンプル数 $N^{(pos)}$ の対象群のデータ $x_i^{(pos)}$ をマハラノビス距離空間中の位置ベクトル $y_i^{(pos)}$ を得て，対称行列 $\Sigma^{(pos)}$ を以下のように生成する．

$$\Sigma^{(pos)} = \frac{1}{N^{(pos)}} \sum_{i=1}^{N^{(pos)}} \left\{ y_i^{(pos)} \right\}^T y_i^{(pos)} \qquad (21)$$

ここで，開発セットに含まれる全ての対象群のサンプルで位置ベクトル $y_i^{(pos)}$ を平均したベクトルは零ベクトルとはならないため，対称行列 $\Sigma^{(pos)}$ は分散共分散行列ではないことに注意する．$\Sigma^{(pos)}$ を特異値分解することで得られる基底ベクトルは，マハラノビス距離空間の原点から見た対象群の L2 ノルムの総和が極大になる方向を示している．この基底ベクトルのうち固有値が最も大きなもの，すなわち L2 ノルムの総和が最大になるベクトルを，回答分布特徴を示すベクトル(L2MAX)とする．

4.2.2 設問空間での L2 ノルム最大化ベクトル

3 章のタスクにおいてマハラノビス距離空間への線形変換を行うが，その手順を省略し，設問空間中で式(21)と同様の計算で L2 ノルムを最大化するベクトルを生成し，回答分布特徴を示すベクトル(No-STAN)とする．

4.2.3 対象群と非対象群の平均差ベクトル

式(21)に現れるマハラノビス距離空間中の対象群の位置ベクトル $y_i^{(pos)}$ は，対象群の全てのサンプルの平均をとると零ベクトルとならないことを述べた．そこで，マハラノビス距離空間中の対象群と非対象群の位置ベクトルの平均の差を，回答分布特徴を示すベクトル(MEAN)とする．このベクトルは，マハラノビス距離空間への線形変換前後で本質的に同じものが得られる．

4.2.4 削減後の設問等の下限

3.3 節で述べた設問等の削減において，データに残す設問等の数の下限を 4 項目とする．アンケートにおいて複数の設問等の相関を分析する際には，クロス集計が行われることが一般的である．クロス集計では，集計表の行と列にそれぞれ 1 項目ずつの設問等が配置されることが基本であり，行または列に 2 つ以上の項目を盛り込むと可読性が落ちる．そこで，相関分析が複雑になる 4 項目を，本稿のタスクにおける設問等の削減後の下限値とする．

4.2.5 退化した空間でのベクトル生成

後述するベクトルの評価実験において交差検証を行うが，回答傾向を示す設問セットを探すという本稿が目指すタスクの性質上，開発データに含まれるサンプルがわずかに異なることで，筆頭となるべき回答傾向と次点の回答傾向の強さの順位が入れ替わって抽出されることが起こり得る．そのため，1 種類の開発データと 1 種類のベクトル生成手法のペアから，次元縮退を経て複数のベクトルを生成し，開発データの組成によって筆頭と次点の順位が入れ替わっても取りこぼさないように実験設定を行う．

設問等の数と等しい次元を有する行列積 $c\,F^{(neg)}$ は，設問空間において回答傾向を表すベクトルと見做せる．そこから設問等の削減を経て得た回答傾向ベクトルを c'，もとの設問空間の任意の位置ベクトルを x とすると，退化した空間での位置ベクトル x' は，

$$x' = x - \frac{x \cdot c'}{|c'|^2} c' \qquad (22)$$

となる．こうして得られる縮退されたデータから，3.2.2 節で述べたマハラノビス距離空間を再び生成し，回答傾向ベクトルを生成する操作を繰り返す．線形代数学的には，先に見つかった回答傾向ベクトル c' による直交補空間内で，別の回答傾向ベクトルを再び探索する操作であると言い換えることができる．

4.3 回答傾向ベクトルの評価方法

4.1 節で述べた各データのそれぞれで，80%を開発セット，20%をテストセットとする 5 分割交差検証を 10 ループ行う．開発セットを 3 章で述べた回答傾向ベクトル生成タスクに適用し，設問等を削減した回答傾向ベクトルを求める．さらに，4.2.5 節で述べた空間縮退を生成手法ごとに 2 回ずつ実施して，計 3 種類のベクトルを得る．

すなわち，1 種類のデータと 1 種類のベクトル生成手法のペアに対して，3 種類のベクトルを生成する試行が 50 回実施され，150 のベクトルを得る．それによって得られたベクトルの扱いと評価について述べる．

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

4.3.1 テストセットの対象群を正例とした AUC

計 50 回の交差検証のそれぞれで，開発セットから得られる 3 種類の回答傾向ベクトルをテストセットに適用する．テストセットに含まれる対象群のサンプルを正例，非対象群のサンプルを負例として，各サンプルの座標とベクトルの内積をスコアとしたランキングを作成し，ランキングの妥当性を AUC (Area Under the Curve) によって評価する．後述するベクトルのグルーピングを経て，各ベクトルに紐づいた AUC をグループ内で平均して評価する．

4.3.2 ベクトルのグルーピングと内容評価

1 種類のデータと 1 種類のベクトル生成手法のペアから得られる 150 のベクトル同士でコサイン類似度を取り，他のベクトルとの類似度の絶対値総和が最大となるベクトルを中心として，類似度の高いベクトルを 49 集め，計 50 のベクトルでグループを作る．さらに，残り 100 のベクトルからも同様の手順で 50 のベクトルからなるグループを作る．それぞれのグループに属するベクトルの平均を，グループを作った順に枝番を付けて L2MAX_1，L2MAX_2 のように表記する．得られた平均ベクトルの成分のうち，絶対値の大きなものと対応する設問等の内容と回答傾向を調べる．また，グループに属する各ベクトルをテストデータに適用した際の AUC の平均も，グループごとに評価する．

4.3.3 アンケートデータの L2 ノルム誤差の検証

2.3 節で述べたように，マハラノビス距離空間におけるデータの L2 ノルムは，多変量正規分布を前提とした確率モデルにおける観測値の対数尤度と等価の関係にある．また 2.4 節で述べたように，モデルの最大対数尤度と真の対数尤度の間には誤差があり，誤差はパラメータ数に依存する．そこで，テストセットを真の分布に従うデータと見做し，設問を削減するごとに開発セットとテストセットの両方で L2MAX_1 の方向の L2 ノルムを計算し，式(10)に倣ってサンプル数による調整を行い，L2 ノルムの誤差として設問数との関係を調べる．その結果を受け，最良の設問削減数について議論する．

以上，3 章と 4 章で述べたタスクと実験の手順についてまとめたものを図 2 に示す．

5. 実験結果と考察

各データから抽出した回答傾向ベクトルの結果を表 3 と表 4 に示す．以下，結果について考察する．

5.1 疑似データの結果

表 3 に，疑似データに対してタスクを適用して抽出された対象群の特徴ベクトルの成分と，ベクトルをテストデータに適用した際の平均 AUC をまとめる．ベクトルの成分が正の場合は ”1”，負の場合は ”0” の回答が多いという傾向を示しているため，回答者グループごとの回答確率類型を示している設問等の名称の 4 文字のうち，ベクトルの符号と回答分布が合致する回答者グループに対応する文字を太字にしている．

5.1.1 SIM_BIAS の結果

L2MAX と No-STAN の両手法によって抽出された特徴ベクトルは，各ベクトルの主要な成分，すなわち特徴として有力とされた設問等の回答傾向が，いずれのベクトルにおいても同一の回答者グループの傾向を示している．また，いずれも上位 2 成分で各ベクトルの L2 ノルムの 50% 以上を占めており，特定の回答者グループを強く指示するベクトルとなっている．実際のデータではこれらのベクトルを見ることで，対象群の中に個別に複数存在する傾向を弁別したうえで傾向を閲覧できることが期待される．

一方，MEAN によって抽出された特徴ベクトルは，上位の各成分が示す回答傾向と対応する回答者グループが分散している．AUC は高く，対象群全体の傾向を総合することで対象群と非対象群を分けるという機能が高くなっているものの，回答傾向が異なる 3 つの対象群をそれぞれ弁別する性能は高くないといえる．

5.1.2 SIM_ROUND の結果

一般的に線形分析を適用することが困難なデータであり，いずれの手法でもベクトルに従って作成したランキングの AUC が 0.5 程度となる．これは，対象群と非対象群をランダムにランキングした場合と変わらない結果である．しかしながら，各ベクトルの成分に着目すると， L2MAX や No-STAN ではベクトル内の上位成分が示す回答傾向と合致する回答者グループが，ベクトル内でほぼ共通していることがわかる．MEAN では対象群と非対象群が完全に同質であるはずの設問が上位に入っており，データの偶然の偏りによる影響だと考えられる．

5.2 アンケートデータの結果

表 4 に，アンケートデータにタスクを適用して『東京オリンピックを楽しみにしている人』の回答傾向ベクトルを抽出した結果を示す．また，開発セットとテストセットの L2 ノルムの誤差と設問等の数との関係を図 3 に示す．

5.2.1 回答特徴ベクトルの生成結果

L2MAX_1 の上位成分には，オリンピックの観戦方法と

表 3：疑似データから抽出した回答傾向ベクトルの AUC と上位 4 成分

	SIM_BIAS					SIM_ROUND				
	AUC	設問等(上)と成分量(下)				AUC	設問等(上)と成分量(下)			
L2MAX_1	0.682	RR**L**R -0.615	LL**R**L 0.592	QQ**M**Q -0.170	MM**Q**M 0.136	0.494	MM**R**N 0.239	RR**M**Q -0.216	Q**LL**M -0.187	LL**Q**M 0.187
L2MAX_2	0.626	**R**LLL 0.529	**L**RRR -0.521	**Q**MMM 0.151	**M**QQQ -0.131	0.498	R**M**RQ -0.321	L**Q**LM 0.315	Q**L**QN -0.243	M**R**MN 0.227
No-STAN_1	0.653	RR**L**R -0.546	LL**R**L 0.541	QQ**M**Q -0.168	MM**Q**M 0.152	0.486	Q**LL**M -0.105	M**RR**Q 0.102	L**QQ**N 0.094	R**MM**N -0.092
No-STAN_2	0.671	**L**RRR -0.676	**R**LLL 0.667	**Q**MMM 0.192	**M**QQQ -0.173	0.521	Q**LL**M -0.349	M**RR**Q 0.339	L**QQ**N 0.327	R**MM**N -0.321
MEAN_1	0.811	**R**LLL 0.524	RR**L**R -0.387	LL**R**L 0.310	R**L**RR -0.292	0.488	QQQQ -0.373	RRRR -0.301	MMMM 0.116	MMMM -0.100
MEAN_2	0.778	L**R**LL 0.554	**L**RRR -0.324	R**L**RR -0.289	LL**R**L 0.191	0.490	MMMM 0.235	MMMM -0.209	QQQQ -0.165	MMMM 0.131

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

表 4：アンケートデータから抽出した回答傾向ベクトルの成分(左)と設問内容一覧(右)

	AUC	設問等(上)と成分量(下)			
L2MAX_1	0.743	Q19_2 0.551	Q19_1 0.245	Q17_3 0.187	Q21S2 0.112
L2MAX_2	0.796	Q19_1 0.642	Q19_3 0.362	Q21S2 0.214	Q21S1 0.093
No-STAN_1	0.791	Q19_2 0.677	Q21S2 0.471	Q19_1 0.454	Q17_3 0.223
No-STAN_2	0.732	Q19_2 - 0.542	Q19_1 0.404	Q21S2 0.394	Q21S1 0.260
MEAN_1	0.829	Q19_3 0.747	Q19_1 0.460	Q21S1 0.277	Q17_4 0.122
MEAN_2	0.774	Q21S2 0.634	Q17_2 0.262	Q19_2 0.236	Q17_4 0.177

Q17_2: 普段から誘われたらスポーツ観戦に出かける
Q17_3: 普段から PV や居酒屋でスポーツ観戦する
Q17_4: 普段から TV やネットでスポーツを視聴する
Q19_1: 東京オリンピックを生で観戦したい
Q19_2: PV があればオリンピックを観戦したい
Q19_3: TV やネットでオリンピックを観たい
Q21S1: 東京オリンピックで経済に良い影響があると思う
Q21S2: 東京オリンピックで暮らしに良い影響があると思う
※　PV: パブリックビューイング

して PV(パブリックビューイング)を希望するかを問う Q19_2 が含まれ，希望するという回答傾向が得られた．さらに，普段のスポーツ観戦の方法として居酒屋や PV を利用するか問う Q17_3 が，同じように肯定的な回答傾向を示すとして含まれている．オリンピックの観戦方法と普段のスポーツ観戦の方法との相関が示されていることになる．

No-STAN_1 で得られたベクトルは，L2MAX_1 で得られたベクトルと似ているが，オリンピックによる暮らしへの良い影響を問う Q21S2 が大きな成分を有している．スポーツの観戦方法に関する設問が上位に集中する L2MAX_1 と比べ，傾向の統一性が薄らいでいることになる．また，No-STAN_1 の直交補空間から抽出されたベクトルである No-STAN_2 は，Q19_2 に対応する成分が大きくマイナスの値をとっており，これはオリンピック観戦に PV を利用しない意向が強いという回答傾向を示している．しかしながら，AUC の値は大きく低下していないため，PV を利用する意向がなくてもオリンピックを楽しみにしている回答者の傾向を補足できていると考えられる．

5.2.2 L2 ノルム誤差と設問等の数の関係

各試行で得られたベクトル方向の L2 ノルムを計算し，開発セットの L2 ノルムからテストセットの L2 ノルムを減じた差(L2norm)と，設問数との関係を図 3 に示す．参考として，ランダムに設問を削減した場合 (L2norm_random) も併せて図示する．L2norm_random がほぼ直線状に増減していることから，設問数と L2 ノルム，すなわち対数尤度の誤差は比例の関係にあり，AIC の理論における対数尤度誤差とパラメータ数との関係に類似している．

設問数が多い段階では L2norm の増減が少ない．これは 3.3.1 節で定めた選択方法によって選択・削減された設問が，回答傾向ベクトル方向の L2 ノルムの値に大きな影響を与えていないことを意味する．

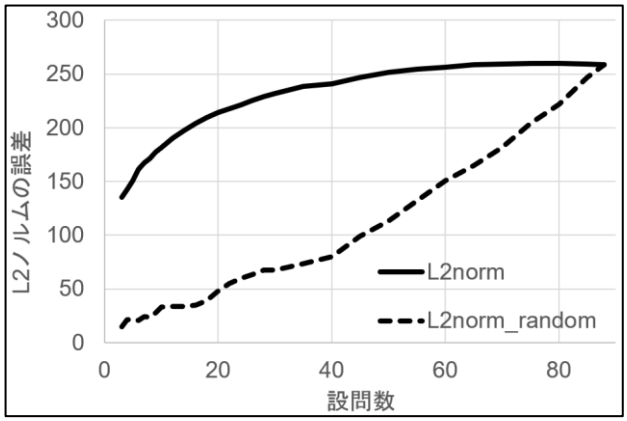

図 3　回答傾向ベクトル方向の L2 ノルム誤差と設問数

6. まとめと課題

アンケートデータの回答分布を正規分布に近似して回答者の座標を定義し，特定の回答者である対象群の傾向を多次元空間中のベクトルの形で抽出するタスクに取り組んだ．疑似逆行列を用いてマハラノビス距離空間を生成し，L2 ノルム最大化ベクトルによって対象群が有する複数の傾向を個別に純度高く提示できる効果を示した．

また，ベクトルの有意性を検定によって評価し，もとのデータから設問を削減する試みも行った．マハラノビス距離空間における L2 ノルムが対数尤度と等価であることから，設問数と L2 ノルムの誤差には基本的に比例の関係があり，削減する設問の選び方によって L2 ノルムおよび誤差を大きく保てる効果が認められた．

適切な設問削減数については引き続き検討が必要である．本稿の実験では，設問数と L2 ノルム誤差は整数倍の関係ではなかった．本稿のタスクは AIC で行う最尤推定とは逆で，尤度が最も低くなる(L2 ノルムが最大になる)パラメータを探索しているためであり，データの特性によって比の値が変化することが分かっている．一方で，L2 ノルムの絶対値のスケールと誤差との相関は認められなかった．

また，誤差を設問削減の参考にするにあたり，マハラノビス距離空間への変換行列の生成にも工夫が必要であることが分かった．設問空間でのデータの分布は離散的であり，設問を削減することでさらに”疎”となることで，極端に小さな固有値を有する基底が出現し得る．その基底と対応するマハラノビス距離空間の次元では，データが外れ値的な挙動をすることになるため，その対策が課題となる．

参考文献

[1] R. Guidotti, et al, “A Survey of Methods for Explaining Black Box Models,” ACM computing surveys, Vol.51, (2018).
[2] M. T. Ribeiro, et al, ““Why Should I You?”: Explaining the Predictions of Any Classifier,” Proceedings of the 22nd ACM International Conference on Knowledge Discovery and Data Mining (KDD), 1135-1144, (2016).
[3] E. Angelino, et al, “Learning Certifiably Optimal Rule Lists,” Proceedings of the 23rd ACM International Conference on Knowledge Discovery and Data Mining (KDD), 35-44, (2018).
[4] 稲垣 他, “アンケートデータにおける局所的に類似した回答者グループの抽出手法に関する検討”, 日本感性工学会論文誌, 14 巻, 3 号, 425-431, (2015)
[5] 岩崎, “統計的因果推論の視点による重回帰分析”, 日本統計学会誌, 50 巻, 2 号, 363-379, (2021)

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.

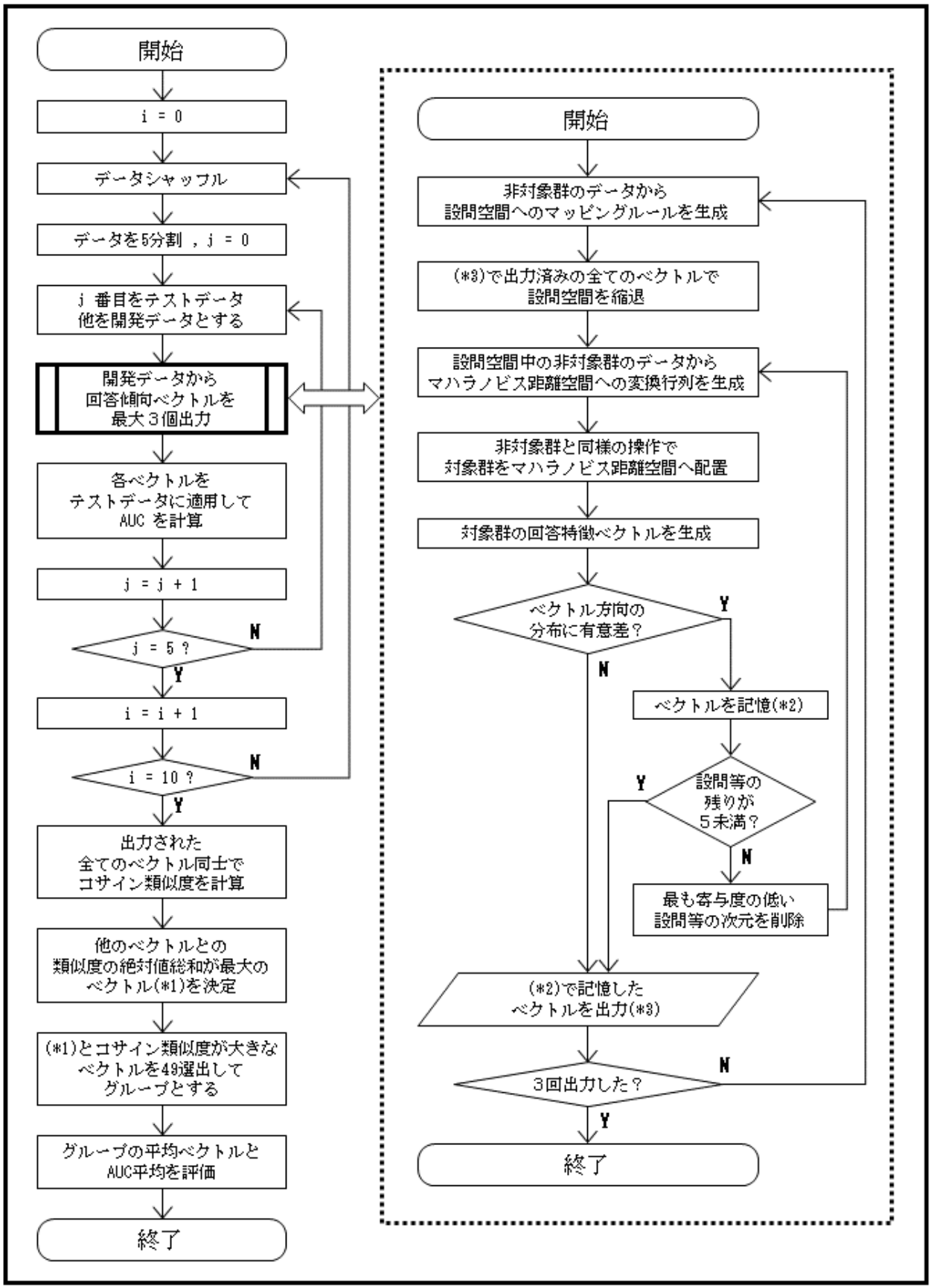

図２　実験手順(左)と回答傾向ベクトル生成タスク(右)のフローチャート

Copyright © 2021 by
The Institute of Electronics, Information and Communication Engineers and
Information Processing Society of Japan All rights reserved.