2010 年度前期　応用統計学　第８回

判別分析とパターン認識 (1) ー マハラノビスの汎距離

パターン認識は，画像工学のなかで統計学にもっとも密接に結びついた分野です．パターン認識とは，パターンを形成するベクトルをいくつかのカテゴリーに分類することということができますが，これらのベクトルはランダム現象によるゆらぎを受けますから，統計学的枠組みで分類法を考える必要があります．このような分類法を判別分析とよび，今回はその代表的なものである「マハラノビスの汎距離」にもとづいた分類法を説明します．

パターン認識とは

パターン認識とは，対象となるパターン群をいくつかのカテゴリーに分類することです．通常，パターンの数はカテゴリーの数よりもはるかに多くなります．例えば，手書きの文字を認識することを考えてみましょう．「A」という文字を手書きで書くと無数の変形が考えられますが，これらはすべて「A」というカテゴリーに分類する必要があります．

もっとも簡単な場合として，入力される手書き文字が「A」と「B」の２通りしかないとしましょう．このとき，パターン認識装置は各々の入力手書き文字を「A」または「B」のどちらかに分類しなければなりません．このもっとも簡単なパターン認識でさえも，手書き文字は通常ゆがんでおり，またどちらのカテゴリーに分類すべきか判断に苦しむ文字が入力されることもしばしばあるので，難しい問題となります．

パターン認識では，各々のパターンから「特徴ベクトル」が抽出されます．特徴ベクトルとはパターンを少数の変量で要約したもので，例えば画像パターンの場合でいえば平均輝度，エッジ（画像中にある物体の輪郭）方向の分布，空間周波数分布などが用いられます．特徴ベクトルを用いるのは，入力パターンの次元数を小さくするためです．例えば，256×256 画素の画像は，そのままでは 65536 次元ベクトルとなります．65536 次元空間でパターンを分類すると，入力パターンの細かい部分まで分類に用いるため，入力パターンがちょっとゆがんでいるだけでも分類誤りを生じてしまいます．そこで，入力パターンの細かい違いを要約するために特徴ベクトルを導入するわけです．

特徴ベクトルの抽出の問題は，画像パターンなら画像処理の問題として，音声パターンなら音声処理の問題として研究されています．この講義では，抽出された特徴ベクトルを分類する問題にしぼって解説します．「マハラノビスの汎距離」にもとづく統計的判別分析は，分類の基礎となる手法のひとつです．

判別分析の原理ー１変量データの場合

判別分析は，２つのグループにあらかじめ分けられたデータがあるとき，新たに入ってきたデータがどちらのグループに属するデータかを判定する方法です．２つのグループに分けることはすなわち何かの行動を起こすかどうか，という意思決定に直結しますので，このような問題の例はさまざまな分野で見ることができます．例えば，病院での検査の結果からその人が病気かどうかを判別する，ある試料をしらべてそれがある物質かどうかを判定する，などいろいろ考えられます．

いま，1つの変量X（例えば血圧）についてのたくさんのデータがあり，それがA，Bの2つのグループ（例えば「健康」と「病気」，よく「健康群」「病気群」という表現をします)）に分かれているとします．このとき，新しい，どちらのグループに入るのかわからないデータxが，どちらのグループに入るかをどう判断すればよいかを考えてみましょう．

ちょっと考えると，新しいデータと，両グループのデータの平均とのへだたりを調べ，近いほうのグループに分類すればよいと思われます．しかし，これでは不十分です．図1の数直線の例を見てみましょう．「新しいデータ（◎)」は，平均とのへだたりでいえば「健康（☆)」のグループに分類されることになります．しかし，これはどう見てもおかしいでしょう．◎は★の並んでいるところに入っていますから，「病気（★)」のグループに分類されるべきです．

これは，両グループの分布のしかたが大きく異なり，「病気」のグループのほうが分散がずっと大きいことが原因です．そこで，分散をとりいれた距離を定義して，これを使って「平均とのへだたり」を定めてみましょう．つまり，分散が大きい（小さい）グループの平均との隔たりは，実際よりも小さく（大きく）するわけです．

世の中の人を全て検査したわけではありませんから，★や☆が病気の人や健康な人の全てではありません．ですから，「病気」「健康」それぞれのグループには母集団があって，その分布は図2のようになっていると考えます．したがって，図1の★や☆のデータは，母集団分布と同じ確率分布にしたがって得られる標本と考えます．そこで，「病気」グループの母平均をμ_A，母分散をσ_A^2とし，「健康」グループの母平均をμ_B，母分散をσ_B^2としましょう．このとき，「新しいデータ」xとμ_A, μ_Bとの隔たりD_A^2, D_B^2を，次のように定義します．

$$D_A^2 = \frac{(X-\mu_A)^2}{\sigma_A^2},\quad D_B^2 = \frac{(X-\mu_B)^2}{\sigma_B^2} \tag{1}$$

上の式を見てわかるように，この隔たりは，グループの分散が大きい（小さい）ほど小さく（大きく）なることがわかります．この距離を用いて「新しいデータ」と両グループのそれぞれの平均との隔たりを求め，隔たりが小さいほうのグループに分類すればよいことになります．実際には母平均や母分散はわかりませんから，標本から求められる量で代用します．すなわち，母平均は標本平均で，母分散は不偏分散で代用します．

2変量データの場合は？ ー多次元確率分布

では，「血圧」というひとつの変量だけではなく，2つの項目がある，つまり2変量の判別分析を考えてみましょう．例によって散布図上で考えます．2つの変量X_1, X_2の組によるデータがあり，それがA，Bの2つのグループに別れているとします．このとき，ある値の組からなるデータは，2変量の確率分布にしたがって得られる標本と考えます．

2つの変量の組の確率分布とは，どのように考えればよいのでしょうか？ それは，図3のような散布図上の濃淡の「雲」で表されます．この「雲」は，X_1, X_2の値のある組み合わせが現れる確率密度を，雲

図1: 「新しいデータ」は「健康」「病気」のどちらに分類すべきか？

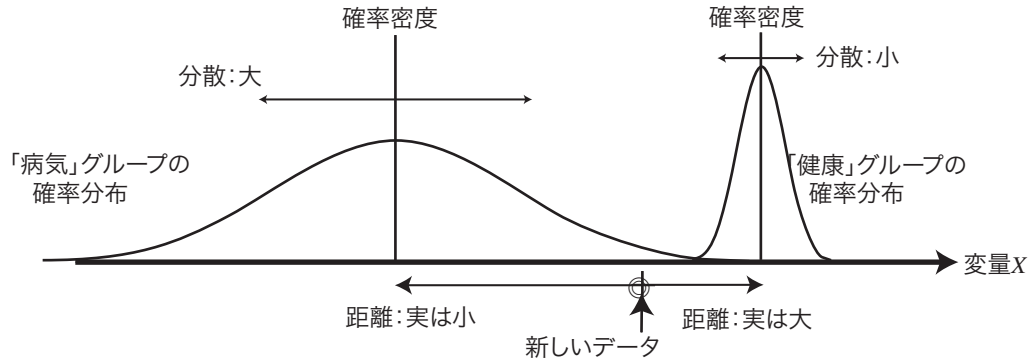

図 2: 両グループの確率分布と「距離」

のグレーの濃さで表したものです．このように「いくつかの値の組み合わせの，出現のしやすさ」を表す確率分布を，多次元確率分布といいます．このような多次元確率分布を使って母集団分布を考えると，観測されたデータはこの確率分布にしたがって得られる標本と考えられます．

X_1, X_2 の組み合わせについての多次元確率分布においても，以前相関を説明したときのように，X_1, X_2 それぞれについての平均と分散，および X_1, X_2 の組の共分散・相関係数を考えることができます．X_1, X_2 の組の相関とは，一言で言えば「X_1 の現れやすい値と X_2 の現れやすい値の関係」で，例えば正の相関があるときには，X_1, X_2 がどちらも大きな値の組は現れやすく，X_1 が大きな値で X_2 が小さな値の組は現れにくいことを示しています．

マハラノビスの汎距離

さて，A，B の 2 つのグループの母集団分布が図 4 のように表されるとしましょう．図 2 の場合と同様に，それぞれのグループに属するデータは，それぞれの母集団の母集団分布と同じ確率分布にしたがって得られた標本と考えます．このとき，新しい（どちらのグループに入るかわからない）データ $x = (x_1, x_2)$ が，どちらのグループに入るべきかを考えてみましょう．

この場合も，新しいデータと，両グループの分布の中心との距離を求め，近いほうのグループに分類するのでは不十分です．図 5 のように，A グループと B グループとで分散が大きく違う場合を考えてみましょう．データ x から各グループの分布の中心への（ユークリッド）距離は同じです．しかし，B グループの確率分布は分布の中心の周りに集まっているため，データ x が B グループに属している可能性はほとんどゼロです．これに対して，A グループの確率分布は広くひろがっているので，このデータ x が A グループから得られる可能性はいくらかはあります．したがって，データ x は A グループに分類されるべきです．そうするために，データ x は B グループよりも A グループに「近い」と評価されるような「距離」を定義する必要があります．このような距離は，1 次元の場合と同様，ユークリッド距離を分散で標準化することで定義できます．

ところが，2 次元（以上）の場合は，さらに考慮すべき問題があります．図 6 の分布では b-d 方向の分散が大きく，a-c 方向の分散は小さくなっています．したがって，a, b, c, d の各点にあるデータは，分布の中心からの「距離」はいずれも同じになるように定義する必要があります．このように定義するためには，分布の形，すなわち変量 X_1, X_2 の相関係数（あるいは共分散）をも考慮する必要があります．

上で述べたような要求を満たす，「散布図上の点と分布の中心との『距離』」を定義してみましょう．あるグループの確率分布が，図 7 のように表されているとします．変量 X_1, X_2 の平均をそれぞれ μ_1, μ_2，分

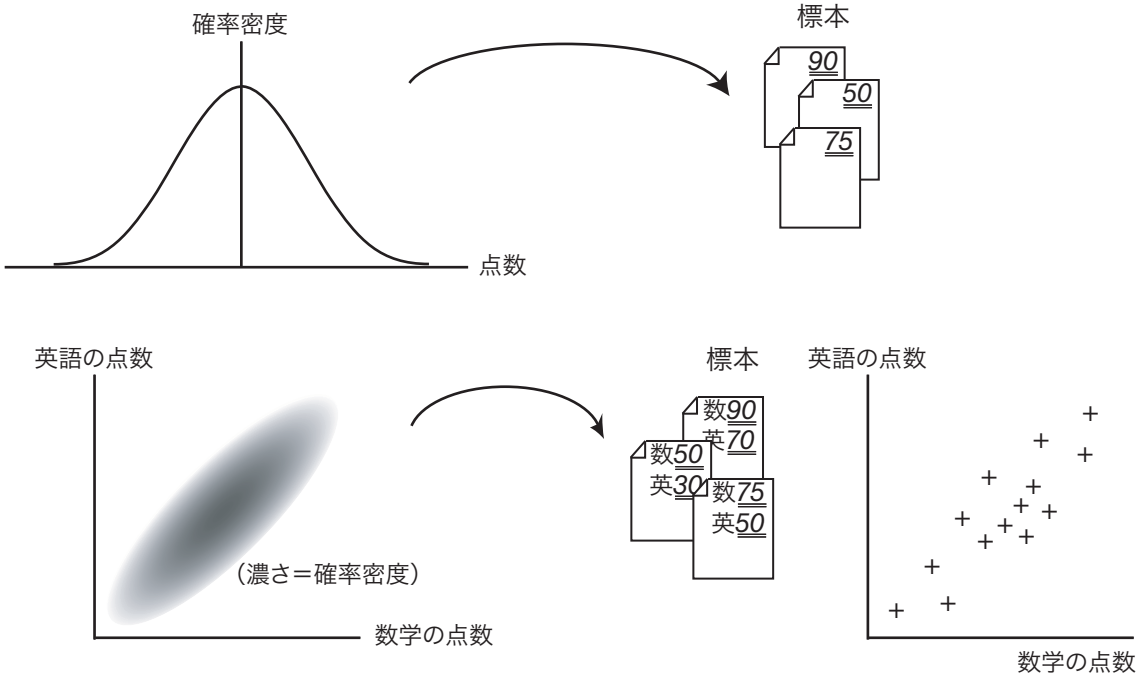

図 3: 多次元確率分布と標本

散をそれぞれ σ_1^2, σ_2^2 とし，X_1, X_2 の相関係数を ρ とします．ここで，簡単のために，変量 X_1, X_2 のデータ x_1, x_2 を

$$u_1 = \frac{x_1 - \mu_1}{\sigma_1}, \quad u_2 = \frac{X_2 - \mu_2}{\sigma_2} \tag{2}$$

と標準化すると，平均は u_1, u_2 とも 0，分散はともに 1，相関係数は同様に ρ となります．

さらに，u_1, u_2 を互いに相関のない変量 $z_{(1)}, z_{(2)}$ に変換します．「互いに相関のない変量」とは，すなわち第１２回で取り上げた主成分です．標準化された２変量の主成分は相関係数によらず常に同じ[1]で，

$$z_{(1)} = \frac{u_1 + u_2}{\sqrt{2}}, \quad z_{(2)} = \frac{u_1 - u_2}{\sqrt{2}} \tag{3}$$

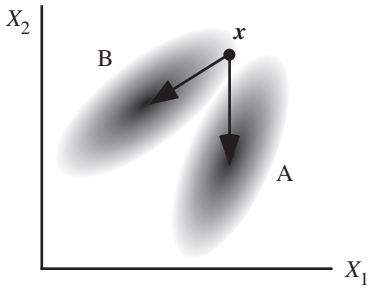

図 4: x は A, B どちらのグループに近いか？

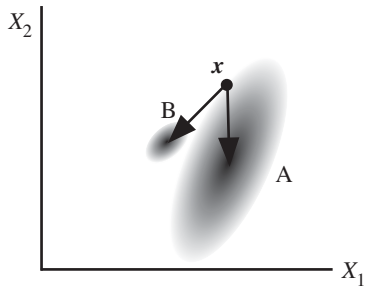

図 5: x は A に「近い」

[1]付録参照．「相関係数によらず常に同じ」になるのは２次元の時だけで，３次元以上の場合はこうはなりません．

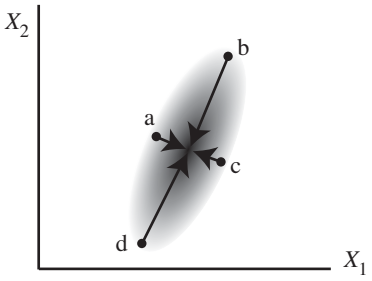

図 6: 分布の中心から「等距離」の点

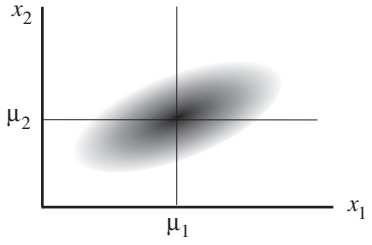

図 7: ２次元の確率分布

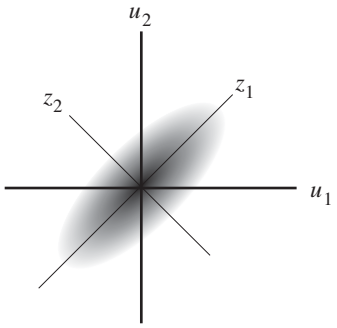

図 8: 主成分に変換

となります（図 8）.

このように変換してしまうと，２つの変量の相関，すなわち「分布が散布図上でどちら向きに傾いているか」はもう考える必要がありません．そこで，散布図上のある 1 点 $(z_{(1)}, z_{(2)})$ と分布の中心（x_1，x_2 軸上では (μ_1, μ_2)，$z_{(1)}, z_{(2)}$ 軸上では $(0, 0)$）との「分散で標準化した」ユークリッド距離（平方距離）を，三平方の定理により，

$$D^2 = \frac{z_{(1)}^2}{V(z_{(1)})} + \frac{z_{(2)}^2}{V(z_{(2)})} \tag{4}$$

のように，$z_{(1)}, z_{(2)}$ 軸上でのそれぞれの分散 $V(z_{(1)})$，$V(z_{(2)})$ で標準化した平方距離の和で表します．この D^2 を**マハラノビスの汎距離**といいます.

$V(z_{(1)})$，$V(z_{(2)})$ は，標準化した場合の主成分分析における $z_{(1)}, z_{(2)}$ の固有値，すなわち $z_{(1)}, z_{(2)}$ の各軸上の分散で，付録にあるように

$$V(z_{(1)}) = 1 + \rho, \quad V(z_{(2)}) = 1 - \rho, \tag{5}$$

です．これと (2) 式を使って $z_{(1)}, z_{(2)}$ を u_1，u_2 に戻すと，

$$\begin{aligned} D^2 &= \frac{z_{(1)}^2}{V(z_{(1)})} + \frac{z_{(2)}^2}{V(z_{(2)})} = \frac{\left(\frac{u_1+u_2}{\sqrt{2}}\right)^2}{1+\rho} + \frac{\left(\frac{u_1-u_2}{\sqrt{2}}\right)^2}{1-\rho} \\ &= \frac{(1-\rho)(u_1+u_2)^2 + (1+\rho)(u_1-u_2)^2}{2(1+\rho)(1-\rho)} \\ &= \frac{\{(1-\rho)+(1+\rho)\}(u_1^2+u_2^2) + \{(1-\rho)-(1+\rho)\}2u_1u_2}{2(1+\rho)(1-\rho)} \\ &= \frac{2(u_1^2+u_2^2) - 2\rho\cdot 2u_1u_2}{2(1-\rho^2)} = \frac{u_1^2+u_2^2-2\rho u_1u_2}{1-\rho^2} \end{aligned} \tag{6}$$

が得られます.

判別の方法

マハラノビスの汎距離を使うと，あるデータがグループ A，B のどちらに属するかは「A，B それぞれに対応する確率分布の中心とのマハラノビスの汎距離が短いほうのグループに属する」という基準で判別されます．ただし実際の判別では，グループ A，B の確率分布は通常わかりません．そこで，グループ A，B の母平均・母分散を，すでにわかっているグループ A，B に属するデータから得られる推定量

で代用します．すなわち，n をデータ数（標本サイズ）として（データ数はグループ A，B で違っていてもかまいません），i 番目のデータの変量 x_1 の値を $x_{1,i}$，変量 x_2 の値を $x_{2,i}$ とするとき，母平均 μ_1, μ_2 を

$$\mu_1 \leftarrow \bar{x_1} = \frac{1}{n}\sum_{i=1}^{n} x_{1,i},\ \ \mu_2 \leftarrow \bar{x_2} = \frac{1}{n}\sum_{i=1}^{n} x_{2,i} \tag{7}$$

のように標本平均で代用します．また，分散 σ_1^2, σ_2^2 や共分散 σ_{12}，相関係数 ρ を

$$\begin{aligned} \sigma_1^2 \ &\leftarrow\ s_{11} = \frac{1}{n-1}\sum_{1} n(x_{1,i}-\mu_1)^2,\ \ \sigma_2^2 \leftarrow s_{22} = \frac{1}{n-1}\sum_{1} n(x_{2,i}-\mu_1)^2 \\ \sigma_{12} \ &\leftarrow\ s_{12} = \frac{1}{n-1}\sum_{1} n(x_{1,i}-\mu_1)(x_{2,i}-\mu_1),\ \ \rho = \frac{\sigma_{12}}{\sigma_1\sigma_2} \leftarrow r_{12} = \frac{s_12}{\sqrt{s_{11}}\sqrt{s_{22}}} \end{aligned} \tag{8}$$

のように，標本から求められる不偏分散・共分散を使って代用します（不偏分散を用いるので $n-1$ で割ることに注意して下さい）．

グループ A, B それぞれについて，すでにわかっているデータからこれらの値を事前に計算しておけば，A, B どちらに属するかわからない「新しいデータ」が A, B どちらのグループに判別されるかは次のようにして求められます．すなわち，A, B それぞれのグループについて，(7)(8) 式で計算される μ_1, μ_2, σ_1, σ_2 を用いて，(2) 式で新しいデータの x_1, x_2 の値を使って u_1, u_2 を求めます．さらに，(6) 式によってマハラノビスの汎距離を求めます．グループ A，グループ B それぞれについて求めた汎距離をそれぞれ D_A^2, D_B^2 とすると，「新しいデータ」は

$$\begin{aligned} D_A^2 - D_B^2 < 0 &\Rightarrow \text{グループ } A \text{ に判別} \\ D_A^2 - D_B^2 > 0 &\Rightarrow \text{グループ } B \text{ に判別} \end{aligned} \tag{9}$$

のように，A，B のうち汎距離が小さいほうのグループに属すると判定します．

「次元のわな」

マハラノビスの汎距離は，相関係数や分散共分散行列から求められます．これらの値は母集団分布，すなわち，現れる可能性のあるすべての特徴ベクトルの分布によって決まります．しかし，この母集団分布は実際には不明で，わかっているのはすでに分類されている標本だけです．ですから，前節のように，この母集団分布を標本から推定する必要があります．

パターン認識問題の場合，特徴ベクトルを用いたとしても，パターンを分類するためにさまざまな特徴を調べる必要があるので，特徴ベクトルの次元は高くなりがちです．一方，すでに分類されている標本ベクトルは，それほど多くは用意できません．高次元空間で少ない標本ベクトルから母集団分布を推定するのは，標本がすき間だらけに配置されていることになるので難しく，推定は不正確になります．これを「次元のわな」(curse of dimensionarity) とよびます．第 10 回の講義で説明するサポートベクタマシンとカーネル法は，次元のわなを回避する方法のひとつです．

今日の演習

「血圧」と「コレステロール」の２つの検査項目で，ある病気の診断をするとします．「健康群」「病気群」のそれぞれの検査項目の平均，不偏分散，相関係数は次の通りでした．

健康群：血圧の平均 100・分散 100，コレステロールの平均 150・分散 144，相関係数 0.7
病気群：血圧の平均 150・分散 169，コレステロールの平均 200・分散 225，相関係数 0.8

さて，ある人を検査すると血圧 120，コレステロール 160 でした．この人はどちらの群に判別すべきかを答えてください．［「健康群」「病気群」が上のグループ A, B に，「血圧」「コレステロール」が上の変量 1, 2 に相当します］

付録：標準化された２変量の主成分

２変量のとき，標準化された変量 u_1, u_2 については，分散はどちらも 1 であり，共分散はすなわち相関係数 ρ となります．このとき，固有値問題

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\begin{pmatrix} u_1 \\ u_2 \end{pmatrix} = \lambda \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \tag{A1}$$

を解くと，特性方程式は

$$(1-\lambda)^2 - \rho = 0 \tag{A2}$$

となり，固有値は

$$\begin{aligned} \lambda &= 1 \pm \sqrt{1-(1-\rho^2)} \\ &= 1 \pm \rho \end{aligned} \tag{A3}$$

となります．すなわち両主成分 $z_{(1)}, z_{(2)}$ の分散 $V(z_{(1)}), V(z_{(2)})$ は $1 \pm \rho$ となります．また，(A1) から導かれる方程式

$$(1-\lambda)u_1 + \rho u_2 = 0 \tag{A4}$$

に (A3) 式の結果を代入すると，

$$\begin{aligned} (1-\lambda)^2 - \rho &= 0 \\ (1-(1\pm\rho))u_1 + \rho u_2 &= 0 \\ \rho(\mp u_1 + u_2) &= 0 \ \text{ゆえに}\ u_2 = \pm u_1 \end{aligned} \tag{A5}$$

となります．$u_1^2 + u_2^2 = 1$ の関係を考慮すると，求められる主成分 z_1, z_2 は常に

$$\frac{u_1+u_2}{\sqrt{2}},\ \frac{u_1-u_2}{\sqrt{2}} \tag{A6}$$

の２つになります．どちらが第 1 主成分か，および寄与率は，相関係数の正負・値によります．

これは当然といえば当然で，元の変量が標準化された結果どちらも同じ分散（すなわち 1）に圧縮（伸長）されていますから，図 A1 のように主成分は常に右上 45 度・左上 45 度の２つの直線になります．ただし，このようになるのは２変量の場合だけで，変量の個数が３つ以上の場合は，それぞれの変量間の相関が問題になるため，こんなに簡単にはなりません．

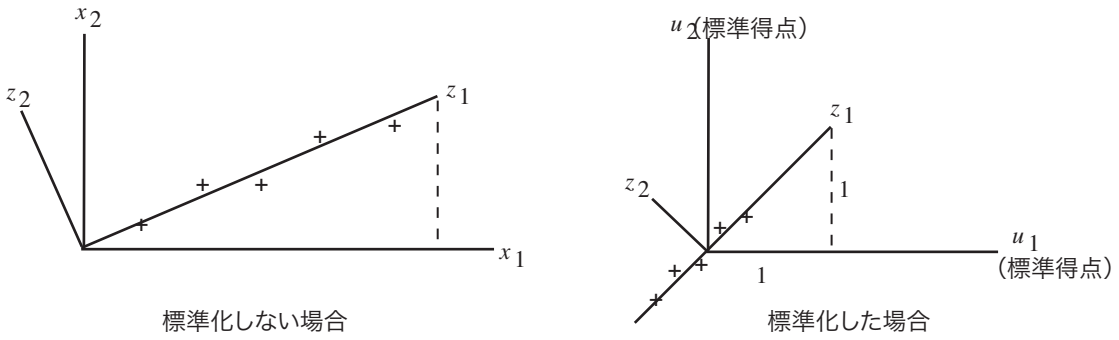

図 A1: 標準化されたデータの主成分分析