

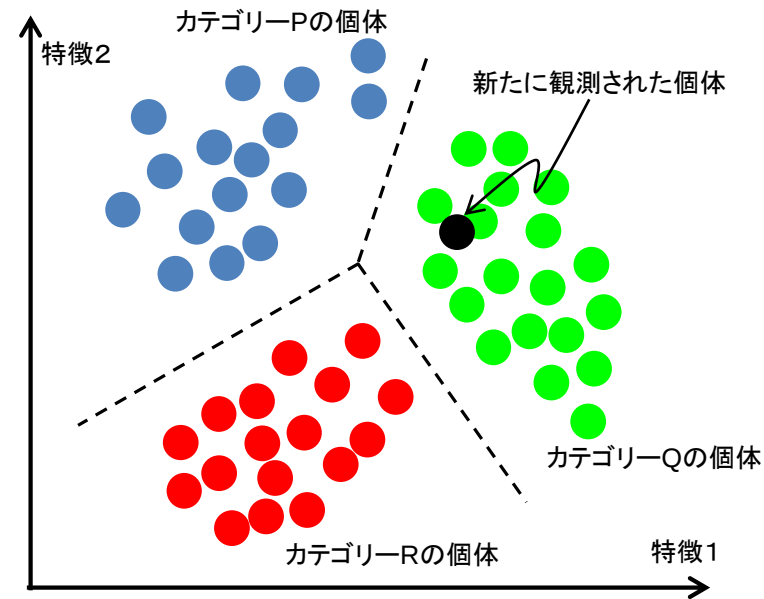
# 判別分析の手順

2021年 ver. 1.0  
倉谷 隆博

# 10-1. 判別分析とは

判別分析は、特徴ごとで分けられた複数のカテゴリー（グループ、群）があったとき、新たな個体がどのカテゴリーに属するのかを判断する方法である。

（図1）は、個体が有する2つの特徴をもとに個体をプロットしたものであるが、判別分析では、図のようにカテゴリーに分けられるとき、新たな個体がどのカテゴリーに属するのかを判断する。（図1）の例では、特徴量1と特徴量2で3つのカテゴリーに分類されており、新たに観測された個体が図の位置にプロットされるならばこの個体はカテゴリーQに属すると判別できる。このとき、すでにプロットされている個体のデータをもとに破線で示されるカテゴリー間の境界線を求めておくことが、新たに観測される個体がどのカテゴリーに属するのかを判別するために必要になる。判別分析ではこの境界線を求めることがポイントになる。



（図1）判別分析とは

## 判別分析の用途

### ①故障診断

機械の、例えば部位ごとの温度データとか振動データなどをもとに故障原因を特定する。こんな温度とこんな振動が観察されるときはこんな故障原因であるとの過去のデータをもとに、現象と故障原因の関係を明らかにしておき、故障が発生したとき温度と振動の計測データをもとに故障原因を特定することができる。

### ②画像認識

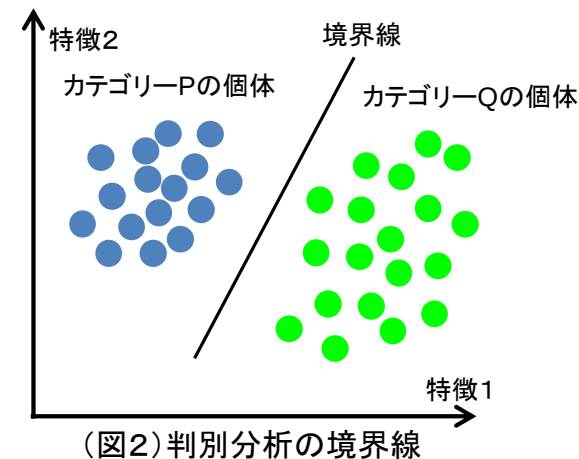
種類の異なる花の写真をそれぞれ複数枚用意し、それぞれの花を色、形などの複数の特徴で把握しておく。これらの特徴をもとに花の種類を判別できるようにしておく。すると、花の写真が新たに提示されたとき、花の特徴からどの種類の花であるかを判別できる。この方法は手書き文字認識とか顔認識とかにも活用される。

### ③マーケティング

- ・倒産した企業の経営データをもとに、ある会社が倒産の危機に直面しているか否かを判別する。
- ・ある商品を購入した人のアンケート結果をもとに、おすすめ商品を提示する。

# 10-2. 判別分析の方法(1/3)

(図2)に示すように、2つの特徴量を軸とする平面上に個体をプロットしたとき、2つのカテゴリに分かれているときは、2つのカテゴリを区別するための境界線を2つのカテゴリの間に引くことができる。  
判別分析では、この境界線を使って新しい個体がどちら側のカテゴリに属しているのかを判定する。



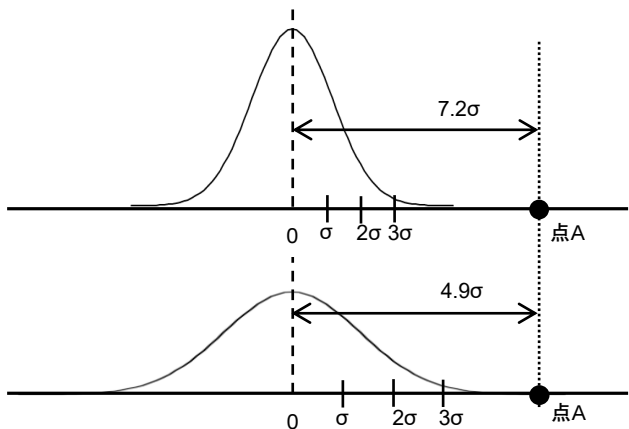
## (1) マハラノビス距離を使って境界線を引く方法

マハラノビス距離とは、データ分布の広がりを加味した距離であり、1次元データの場合、式(1)で定義される。

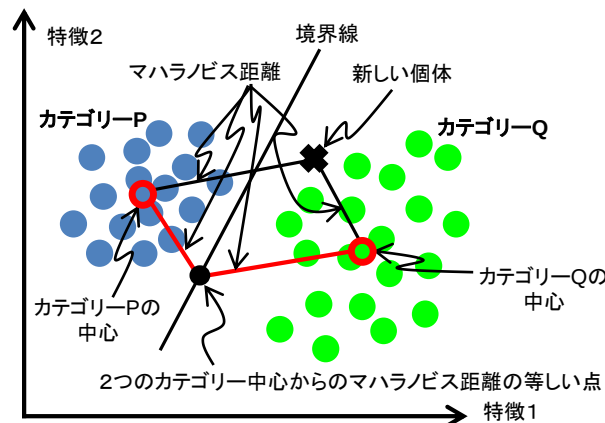
データ  $x$  の分布の中心(平均  $\mu$ )からの距離を標準偏差  $\sigma$  で割っており、標準偏差を1単位とした距離になる。

$$d = |x - \mu| / \sigma \quad \dots (1)$$

(図3)の上の分布では点Aは分布の中心から  $7.2\sigma$  離れているが、同じ点Aであっても、下の分布では分布の中心から  $4.9\sigma$  離れていると考えるのがマハラノビス距離である。分布が広がっている方の距離が短くなる。  
マハラノビス距離を境界線に使う方法は、(図4)に示すように2つのカテゴリの間にそれぞれのカテゴリの中心(平均)からのマハラノビス距離が等しいところに境界線を引く方法である。新しい個体(図では、◆印)が**どちらのカテゴリに属するか**の判別は、**カテゴリの中心からのマハラノビス距離が近いほう**が選ばれる(図では、カテゴリQが選ばれる)。



(図3) 1次元でのマハラノビス距離



この図の2次元のマハラノビス距離は、式(2)で定義される。

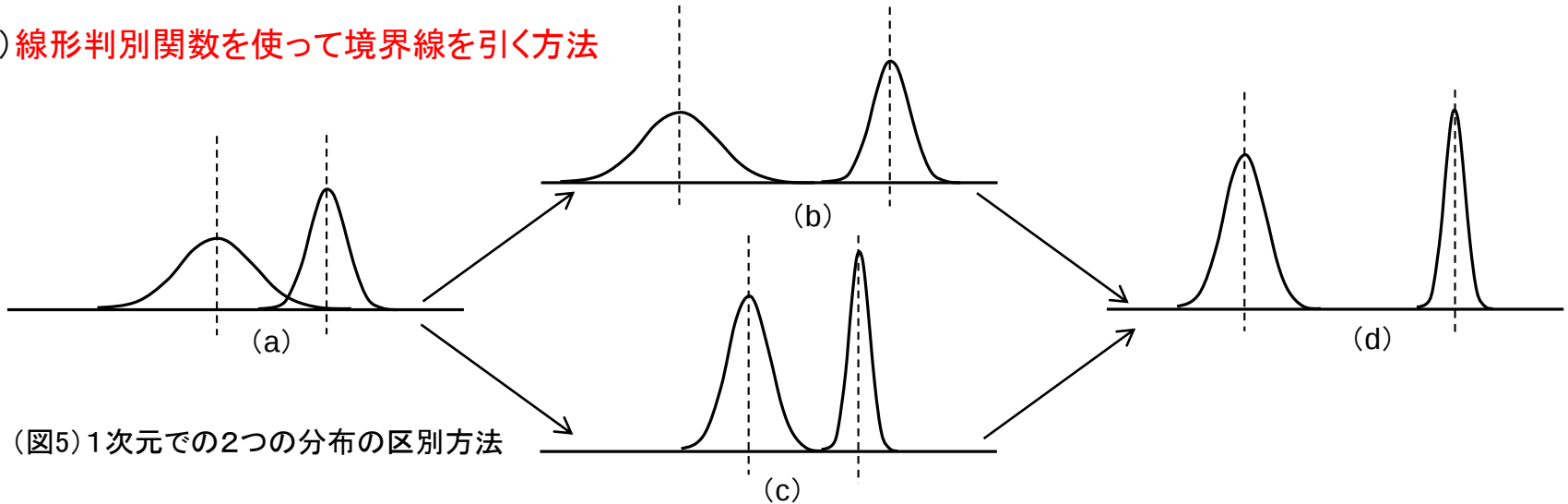
$$d = [(x - \mu)^T S^{-1} (x - \mu)]^{1/2} \quad \dots (2)$$

$\mu$  : カテゴリの中心(平均)  
 $S$  : 分散共分散行列  
 $T$  : 転置行列  
 $^{-1}$  : 逆行列

(図4) マハラノビス距離を使った境界線

# 10-2. 判別分析の方法(2/3)

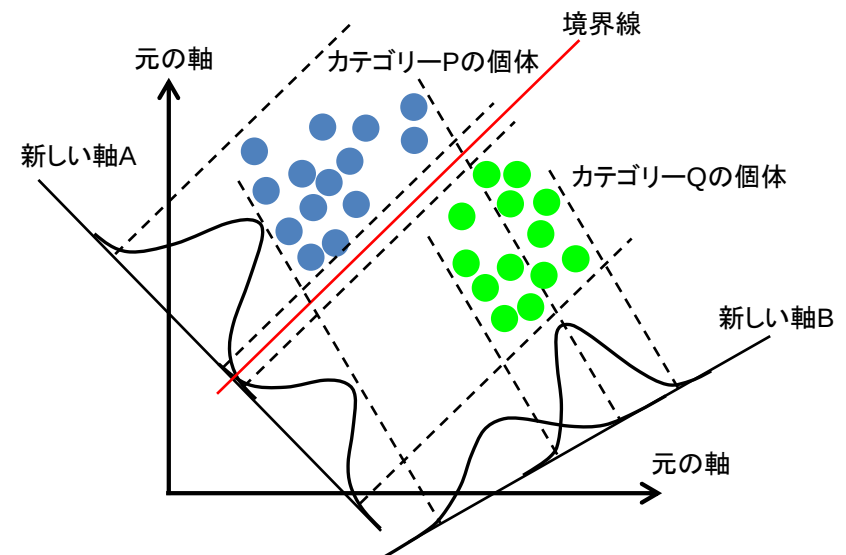
## (2) 線形判別関数を使って境界線を引く方法



(図5) 1次元での2つの分布の区別方法

2つのカテゴリーの分布があった場合、2つのカテゴリーを区別しやすくする方法を(図5)で示す。(a)のように分布が重なっていると重なり部分ではカテゴリー区分は難しい。区別しやすくする方法には2つある。1つは(b)のように2つのカテゴリーの分布を離すことである。もう一つの方法は、(c)のように分布の幅を狭くすることである。そして、この2つを組み合わせるのが(d)であり、区別がとてもしやすくなる。

(図6)は、2つのカテゴリーの分布の形が見る方向によって変わることを示したものである。新しい軸Aへの分布の射影では2つのカテゴリー分布が分かれて見えるが、新しい軸Bへの分布の射影では2つのカテゴリー分布が重なって見えてしまう。判別分析では、軸Aのような分布が分かれて見えるような新しい軸を、(図5)に示す(b)と(c)の視点から作り出し、2つの分布の間に判別のための境界線を設定する。



(図6) 1次元での2つの分布の区別方法

# 10-2. 判別分析の方法(3/3)

## (2) 線形判別関数を使って境界線を引く方法(続き)

判別分析で、カテゴリーが分けしやすく見える新しい軸を求める。そのために、(図7)に示すカテゴリー間分散とカテゴリー内分散を使用して新しい軸を定める。この軸は式(3)で定義される相関比を最大にする軸である。

$$\text{相関比} = \text{カテゴリー間分散} / \text{カテゴリー内分散} \cdots (3)$$

$$\text{カテゴリー間分散} = \text{新しい軸での、カテゴリーPの中心とカテゴリーQの中心との距離の2乗} \cdots (4)$$

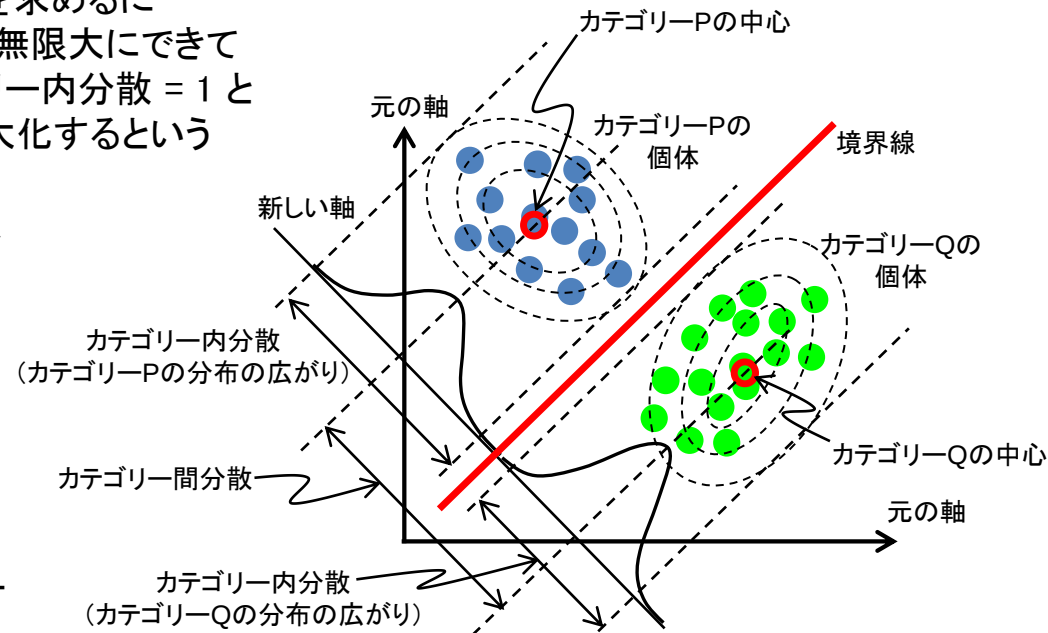
$$\text{カテゴリー内分散} = \text{新しい軸での、カテゴリーPの分散} + \text{カテゴリーQの分散} \cdots (5)$$

式(4)は、(図3)の(a)から(b)、つまり、2つのカテゴリーの分布ができるだけ離れて見えるように、式(5)は、(図3)の(a)から(c)、つまり、2つのカテゴリーの分布の広がりができるだけ小さく見えるように、新しい軸を選ぶことを示している。式(4)と式(5)の組合せの式(3)は(図3)の(d)に相当する。

なお、式(3)の相関比を最大化する新しい軸を求めるにあたっては、分母を0に近づければ相関比を無限大にできてしまうという問題がある。そこで分母のカテゴリー内分散 = 1 という制約条件を設けてカテゴリー間分散を最大化するという方法を採用する。

この問題を解くには、主成分分析のときと同じように、条件付きの最適化問題の解法であるラグランジュの未定乗数法が用いられる。判別分析への適用にあたっては、**相関比と制約条件から成るラグランジュ関数の偏微方程式 = 0 から導かれる固有方程式を解くことで、求めた固有値を使って新しい座標軸(線形判別式)を決めることができる。**

そして、この新しい座標軸をもとに、カテゴリーを区分する境界線を設定することができる。



(図7) 線形判別関数を使用した方法

# 付録 座標変換の手順 1/3

線形判別関数を使用した判別分析では、(図7)に示すような座標変換を行う。ここでは、例を使って座標変換の手順を示す。

(付表1)のデータ、つまり(付図1)に示す座標軸  $x_1$  と座標軸  $x_2$  でプロットされる 10 個のデータを新しい座標軸  $z$  へ移す。この座標変換では、新しい座標軸  $z$  として2つのカテゴリーの相関比を最大にする座標軸を求める。そのためには、相関比と制約条件を織り込んだラグランジュ関数の偏微分 = 0 から求まる下記の固有方程式を解くことになる。

$$S_W^{-1} S_B w = \lambda w \cdots (\text{付式1})$$

ただし、

$S_W$  : 元の座標軸での、カテゴリーQの中心とカテゴリーPの中心との距離2乗(行列)

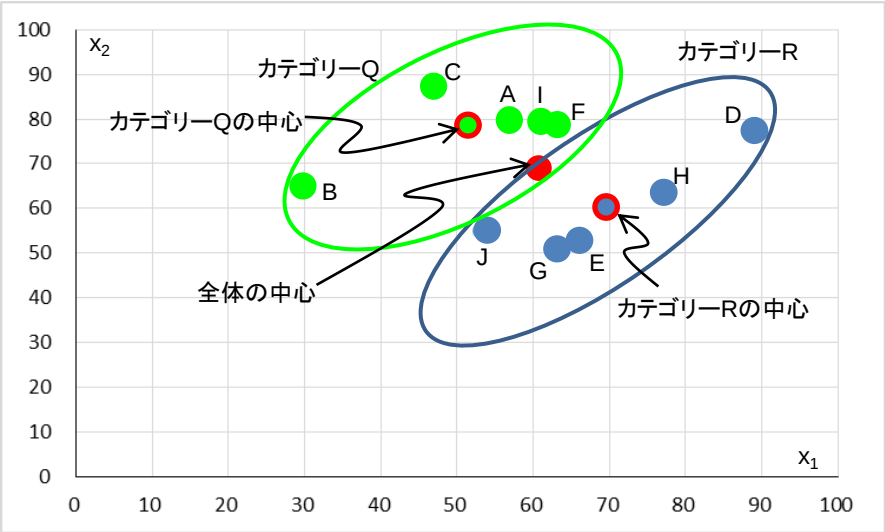
$S_B$  : 元の座標軸での、カテゴリーPの分散+カテゴリーQの分散(行列)

$w$  : 座標変換に使う係数ベクトル(固有ベクトル)

$\lambda$  : 固有値

(付表1)準備したデータ

|       | A    | B    | C    | D    | E    | F    | G    | H    | I    | J    |
|-------|------|------|------|------|------|------|------|------|------|------|
| $x_1$ | 57.0 | 30.0 | 47.0 | 89.0 | 66.0 | 63.0 | 63.0 | 77.0 | 61.0 | 53.0 |
| $x_2$ | 80.0 | 65.0 | 87.0 | 77.0 | 53.0 | 79.0 | 51.0 | 64.0 | 80.0 | 55.0 |
| カテゴリー | Q    | Q    | Q    | R    | R    | Q    | R    | R    | Q    | R    |



(付式1)のベクトル  $w$  は  $0$  ではないので、(付式1)から、(付式2)の行列式が求まる。

この行列式を解いて固有値  $\lambda$  を求め、座標変換に使う係数である固有ベクトル  $w$  を求める。この固有ベクトルが新しい軸を定めることになる。

$$| S_W^{-1} S_B - \lambda I | = 0 \cdots (\text{付式2})$$

ただし、 $I = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$  : 単位行列

カテゴリーQの中心(平均値):  $x_1 = 51.6$   $x_2 = 78.2$   
カテゴリーRの中心(平均値):  $x_1 = 69.6$   $x_2 = 60.0$   
全体の中心(平均値):  $x_1 = 60.6$   $x_2 = 69.1$

(付図1) (付表1)のデータのプロット

# 付録 座標変換の手順 2/3

手順1) カテゴリーQとカテゴリーR、そして全体の中心(平均値)を求める(結果は、(付図1)に記載)。

手順2) (付式2)の行列  $\mathbf{S}_B$  と逆行列  $\mathbf{S}_W^{-1}$  の積  $\mathbf{S}_B \mathbf{S}_W^{-1}$  を求める。

$$\mathbf{S}_B = \begin{bmatrix} 51.6 - 69.6 \\ 78.2 - 60.0 \end{bmatrix} \begin{bmatrix} 51.6 - 69.6 & 78.2 - 60.0 \end{bmatrix} = \begin{bmatrix} -18.0 \\ 18.2 \end{bmatrix} \begin{bmatrix} -18.0 & 18.2 \end{bmatrix} = \begin{bmatrix} 324.0 & -327.6 \\ -327.6 & -331.2 \end{bmatrix}$$

$$\begin{aligned} \mathbf{S}_W &= \begin{bmatrix} 57.0 - 51.6 \\ 80.0 - 78.2 \end{bmatrix} \begin{bmatrix} 57.0 - 51.6 & 80.0 - 78.2 \end{bmatrix} + \begin{bmatrix} 30.0 - 51.6 \\ 65.0 - 78.2 \end{bmatrix} \begin{bmatrix} 30.0 - 51.6 & 65.0 - 78.2 \end{bmatrix} \\ &+ \begin{bmatrix} 47.0 - 51.6 \\ 87.0 - 78.2 \end{bmatrix} \begin{bmatrix} 47.0 - 51.6 & 87.0 - 78.2 \end{bmatrix} + \begin{bmatrix} 63.0 - 51.6 \\ 79.0 - 78.2 \end{bmatrix} \begin{bmatrix} 63.0 - 51.6 & 79.0 - 78.2 \end{bmatrix} \\ &+ \begin{bmatrix} 61.0 - 51.6 \\ 80.0 - 78.2 \end{bmatrix} \begin{bmatrix} 61.0 - 51.6 & 80.0 - 78.2 \end{bmatrix} + \begin{bmatrix} 89.0 - 69.6 \\ 77.0 - 60.0 \end{bmatrix} \begin{bmatrix} 89.0 - 69.6 & 77.0 - 60.0 \end{bmatrix} \\ &+ \begin{bmatrix} 66.0 - 69.6 \\ 53.0 - 60.0 \end{bmatrix} \begin{bmatrix} 66.0 - 69.6 & 53.0 - 60.0 \end{bmatrix} + \begin{bmatrix} 63.0 - 69.6 \\ 51.0 - 60.0 \end{bmatrix} \begin{bmatrix} 63.0 - 69.6 & 51.0 - 60.0 \end{bmatrix} \\ &+ \begin{bmatrix} 77.0 - 69.6 \\ 64.0 - 60.0 \end{bmatrix} \begin{bmatrix} 77.0 - 69.6 & 64.0 - 60.0 \end{bmatrix} + \begin{bmatrix} 53.0 - 69.6 \\ 55.0 - 60.0 \end{bmatrix} \begin{bmatrix} 53.0 - 69.6 & 55.0 - 60.0 \end{bmatrix} \end{aligned}$$

$$= \begin{bmatrix} 1498.4 & 807.4 \\ 807.4 & 718.8 \end{bmatrix} \Longrightarrow \mathbf{S}_W^{-1} = \frac{1}{1498.4 \times 718.8 - 807.4 \times 807.4} \begin{bmatrix} 718.8 & -807.4 \\ -807.4 & 1498.4 \end{bmatrix} = \begin{bmatrix} 0.00169 & 0.00190 \\ -0.00190 & 0.00352 \end{bmatrix}$$

$$\text{よって、} \mathbf{S}_B \mathbf{S}_W^{-1} = \begin{bmatrix} 324.0 & -327.6 \\ -327.6 & -331.2 \end{bmatrix} \begin{bmatrix} 0.00169 & 0.00190 \\ -0.00190 & 0.00352 \end{bmatrix} = \begin{bmatrix} 1.170 & -1.183 \\ -1.770 & 1.789 \end{bmatrix}$$

手順3) (付式2)に、 $\mathbf{S}_B \mathbf{S}_W^{-1}$  を代入すると、 $\lambda$  の2次方程式(付式3)ができる。

$$\begin{vmatrix} 1.170 - \lambda & -1.183 \\ -1.770 & 1.789 - \lambda \end{vmatrix} = 0 \quad \cdots \quad (\text{付式3})$$

よって、 $(1.170 - \lambda)(1.789 - \lambda) - (-1.183)(-1.770) = \lambda^2 - 2.959\lambda - 0.00000305 = 0 \quad \cdots \quad (\text{付式4})$   
この2次方程式を解いて、固有値  $\lambda_1 = 2.959$ ,  $\lambda_2 = -0.00001 \doteq 0$  が求まる。

# 付録 座標変換の手順 3/3

手順4) 元の座標から新しい座標軸へ変換するとき使用する係数のベクトル(固有ベクトル)  $\mathbf{w}$  を求める。  
固有値  $\lambda_1 = 2.959$  の固有ベクトルを求める(なお、 $\lambda_2$ の固有値は 0 であり、固有ベクトルは求まらない)。

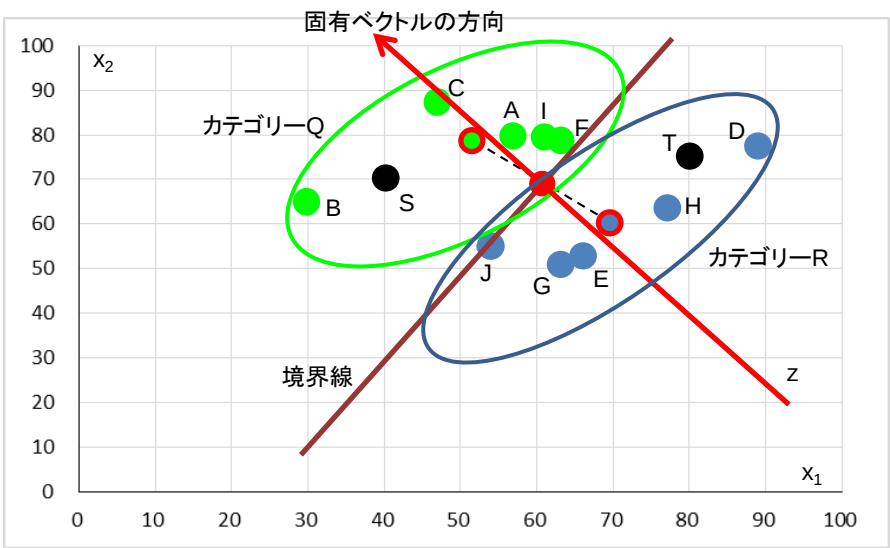
$$\begin{pmatrix} 1.170 & -1.183 \\ -1.770 & 1.789 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = 2.959 \begin{pmatrix} w_1 \\ w_2 \end{pmatrix}$$

よって、 $w_1 = -0.661 w_2 \cdots$ (付式5)

(付式5)から、ベクトルの長さを 1 の単位ベクトル(固有ベクトル)を求めると、 $\mathbf{w} = \begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} -0.551 \\ 0.834 \end{pmatrix} \cdots$ (付式6)

手順5) (付式6)から、新しい座標の軸  $z$  を求める。軸の方向は固有ベクトル  $\mathbf{w}$  と同じ方向である。  
 $z = w_1(x_1 - x_1 \text{の平均値}) + w_2(x_2 - x_2 \text{の平均値}) = -0.551(x_1 - 60.6) + 0.834(x_2 - 69.1) \cdots$  (付式7)

手順6) (付式7)を使って、(付図1)に、全体の中を通る新しい座標軸  $z$  と、全体の中を通り新しい座標軸に直交する境界線を書き入れる(付図2)。この境界線を使ってカテゴリーQとカテゴリーRを区分する。新しい個体がカテゴリーQに属するのか、カテゴリーRに属するのかは、(付式6)の値がプラスであればカテゴリーQに属する、マイナスであればカテゴリーRに属すると判別する。  
なお、(付図2)の例でも分かるが、カテゴリーの分布が重複しているところにプロットされる場合は、判別が難しくなる。



(付図2) 線形判別式と判別のための境界線

例えば、新しい点 S ( $x_1 = 40.0, x_2 = 70.0$ ) のときは、  
(付式7)より、 $z = 12.1 > 0$   
よって、点 S は、カテゴリーQ に属すると判別できる。  
また、新しい点 T ( $x_1 = 80.0, x_2 = 75.0$ ) のときは、  
(付式7)より、 $z = -5.8 < 0$   
よって、点 T は、カテゴリーR に属すると判別できる。

# 参考にした書籍

河本薫：「会社を変える分析の力」、講談社現代新書

小山昇：「数字は人格」、ダイヤモンド社

永野裕之：「ビジネス×数学＝最強」、すばる舎

竹内薫：「数学×思考＝ざっくりといかにして問題をとくか」、丸善出版

中西達夫：「統計データをすぐに分析できる本」、アニモ出版

中西達夫：「すぐれた判断が統計データ分析から生まれる」、実務教育出版

豊田裕貴：「マンガでわかる ビジネスを成功に導くデータ分析」、ナツメ社

向後千春、富永敦子：「統計学がわかる」、技術評論社

石井俊全：「意味がわかる統計学」、ベレ出版

涌井良幸、涌井貞美：「中学数学でわかる統計の授業」、日本実業出版

涌井良幸、涌井貞美：「統計学の図鑑」、技術評論社

西内啓：「統計学が最強の学問である」、ダイヤモンド社

西内啓：「統計学が最強の学問である(実践編)」、ダイヤモンド社

西内啓：「統計学が最強の学問である(ビジネス編)」、ダイヤモンド社

森岡毅、今西聖貴：「確率思考の戦略論」、角川書店

デビッド・マクアダムス：「世界一流企業はゲーム理論で決めている」、ダイヤモンド社

河村真一ほか：「本物のデータ分析力が身に付く本」、日経BPムック

末吉正成、末吉美貴：「Excel ビジネス統計分析 この分析できますか?」、翔泳社

谷岡一郎：「社会調査のウソ リサーチ・リテラシーのすすめ」、文藝春秋

林知己夫：「調査の科学」、ちくま学芸文庫

八谷大岳：「データ解析」シリーズ 全15回、(例)[データ解析 第1回 ベクトルの復習 – YouTube](#)