JMRA リサーチイノベーション委員会 2023 年度研究活動

データサイエンス研究会報告書

スパース推定と機械学習

2024 年 3 月 31 日

研究代表　朝野熙彦

編集委員　森本　修

目　次

キーワード:スパース推定、正則化、CNN、k-umeyama 法

１章　問題意識

１.１　研究の背景

我が国のマーケティング・リサーチは、これまで質問紙調査を主要なデータ収集手段としてきた。そのため調査者側が用意した質問に対してデータが得られるという想定のもとにデータ解析を組み立ててきた。それに対してデジタル化社会から得られるデータは、完備した形で得られる保証がない。そのようなデータから情報を抽出するにはデータサイエンスが必要になる。その学術的な基盤は統計学と情報科学の2分野であり、中心的な方法はスパースデータの解析と機械学習と考えられる。

革新の著しいデータサイエンスをマーケティングの実務にどう活用すればよいのかは今日的な課題といえよう。

マーケティングには市場における異質性の存在を考慮しつつ市場創造を目指すという分野固有の視点がある。したがって他分野でのユースケースを模倣するのではなくマーケティングに適合した方法論を考究する意義があろう。

１.２　プロジェクトの課題

今回のプロジェクト研究において、我々は次の課題意識を持った。

(1)スパース推定

測定技術の進化にともない変数の数がケース数を上回る事態がしばしば発生するようになった。たとえばヒトには約 2 万 3 千個の DNA 遺伝子があるが、その一方で被検者はたいてい数人にとどまる。したがって変数の数がケース数を上回ることになり、従来の回帰分析では分析できない。その対策としてリッジ推定とラッソ推定が知られている。両者の特性の違いを明らかにしたい。

(2)パターン認識

質的なカテゴリーを識別する各種のモデルの関係性を整理したい。

(3) CNN と生成モデル

生成 AI は 2023 年から世界的に注目されるようになった。畳込みニューラルネットワーク（CNN）と生成 AI の基礎を理解した上でマーケティング実務に適用する留意点と限界を明らかにしたい。

(4)クラスター分析

汎距離測定を組み込んだ k-umeyama 法という新しいアルゴリズムを提案し、具体的な応用例を示したい。

２章　回帰モデルのスパース推定

２.１　重回帰分析の基礎

■重回帰分析のモデル

１ 個の目的変数 $\boldsymbol{y}$ とp個の説明変数 $\boldsymbol{X}=\left[\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_p\right]$ について、n個の観測データが得られているとする。各変数ごとにデータが中心化（平均偏差化）されているとすれば、重回帰分析のモデルは（2.1）で表される[1]。最初なのでベクトルと行列のサイズを明記した。

$$\underset{n\times 1}{\boldsymbol{y}}=\underset{n\times p}{\boldsymbol{X}}\,\underset{p\times 1}{\boldsymbol{\beta}}+\underset{n\times 1}{\boldsymbol{\varepsilon}},\qquad \boldsymbol{\varepsilon}\sim N_n\left(\boldsymbol{0},\sigma^2\boldsymbol{I}_n\right) \tag{2.1}$$

(2.1)の $\boldsymbol{\beta}$ は未知の回帰係数パラメータである。誤差（error）$\boldsymbol{\varepsilon}$ は期待値がゼロベクトルで分散共分散行列が $\sigma^2\boldsymbol{I}_n$ の多変量正規分布に従うと仮定することが多い。これは $\varepsilon_i\ \ (i=1,2,\cdots,n)$ が独立に同一の正規分布 $N\left(0,\sigma^2\right)$ に従うと言っても同じである[2]。(2.1)における確率変数は $\boldsymbol{\varepsilon}$ と $\boldsymbol{y}$ であり、説明変数行列の $\boldsymbol{X}$ は定数とみなす。$\boldsymbol{X}$ が確率変数 $\boldsymbol{\varepsilon}$ とは独立である、というのが重回帰分析の基本的な仮定である。

回帰係数の推定法には最尤法、最小二乗法、射影子による方法がある。どの解法を選んでも $\boldsymbol{\beta}$ の推定値は（2.2）で同じである。

$$\boldsymbol{b}=\left(\boldsymbol{X}'\boldsymbol{X}\right)^{-1}\boldsymbol{X}'\boldsymbol{y} \tag{2.2}$$

(2.2)の $\boldsymbol{X}'\boldsymbol{X}$ は情報行列と呼ばれる。情報行列に逆行列が存在する場合に(2.2)が利用できる。回帰係数が求められた場合には $\boldsymbol{y}$ の予測値は(2.3)で得られる。

$$\hat{\boldsymbol{y}}=\boldsymbol{X}\boldsymbol{b} \tag{2.3}$$

$\hat{\boldsymbol{y}}$ で $\boldsymbol{y}$ を予測することから生じる差を残差(residual) $\boldsymbol{e}$ と呼ぶ。$\boldsymbol{e}$ は観測値 $\boldsymbol{X}$ に依存して定まることから $\boldsymbol{\varepsilon}$ とは異なる概念である。なお予測値と残差の相関は 0 である。

$$\boldsymbol{e}=\boldsymbol{y}-\hat{\boldsymbol{y}} \tag{2.4}$$

■ 本報告書における表記の約束

ベクトルはイタリック小文字のボールド体で、行列はイタリック大文字のボールド体で表記する。それぞれの転置はプライムで表す。スカラーには小文字を用いる。たとえば平方汎距離は次のように記述される。

$$md^2 = (\boldsymbol{x} - \bar{\boldsymbol{x}})' \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}}) \qquad (2.5)$$

理論モデルを記述するときは母数にギリシャ文字の $\boldsymbol{\mu}, \boldsymbol{\Sigma}$ などを用いる。さらに次節の罰則付き回帰に出てくる l_1 ノルムと l_2 ノルムの記法を説明する。l_1 は絶対値をとるという操作だが、l_2 ノルムの意味が分かりづらい。たとえば $\|\boldsymbol{Xb}\|_2^2$ という l_2 ノルムの場合、下付きの 2 は l_2 の 2 を表し、上付きの 2 は 2 乗するので平方根をとらないことを示している。l_2 ノルムが何を意味するかを次に示す。

$$\|\boldsymbol{Xb}\|_2^2 = (\boldsymbol{Xb}, \boldsymbol{Xb}) = \boldsymbol{b'X'Xb} \qquad (2.6)$$

(2.6)はベクトル $\boldsymbol{Xb}$ どうしの内積に他ならない。(2.3)の重回帰分析の文脈でいえば(2.6)は、予測値 $\hat{y}_i \quad (i = 1, 2, \cdots, n)$ の二乗和 $(\hat{\boldsymbol{y}}, \hat{\boldsymbol{y}})$ を意味する。

2.2 正則化のアイデア

■罰則つき最小二乗推定量

$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_p$ のベクトルの組に一次従属の関係があるときは情報行列 $\boldsymbol{X'X}$ の逆行列が求められない。一次従属でなくても情報行列の行列式が 0 に近い場合は回帰係数の推定値が不安定になる。このような具合の悪い(ill condition) データへの対策として、損失関数に罰則項を加えるという提案がされている。

Hoerl と Kennard(1970)が提案した Ridge(リッジ)回帰は罰則項を回帰係数の二乗和 $(\boldsymbol{\beta}, \boldsymbol{\beta})$ とおいて回帰係数を推定する。未定乗数をλとおけば最小化すべき損失関数 L とそのリッジ推定量は

$$L(\boldsymbol{\beta}, \lambda) = (\boldsymbol{e}, \boldsymbol{e}) + \lambda (\boldsymbol{\beta}, \boldsymbol{\beta})$$
$$\boldsymbol{b} = (\boldsymbol{X'X} + \lambda I)^{-1} \boldsymbol{X'y} \qquad (2.7)$$

なおリッジ回帰も次の Tibshirani(1996)の Lasso(ラッソ)回帰も説明変数を予め平均 0 分散 1 に標準化したデータを対象に分析を行うので、$\boldsymbol{b}$ を標準偏回帰係

数と呼ぶのが正しいかもしれないが、通常は回帰係数と呼んでいる。

(2.7)を形式的に眺めれば、情報行列の主対角要素に小さな正数λを追加することによって情報行列を正則化させ、回帰係数を安定化させる操作だと解釈できる。リッジ回帰およびラッソ回帰は、情報行列を正則化させるという操作を強調して正則化回帰と呼ばれることがある。なおリッジ回帰の罰則項はl_2ノルムつまり$\|\boldsymbol{\beta}\|_2^2$であり、その罰則をどれだけ効かせるかを定めるハイパーパラメータがλである。またラッソ回帰の罰則項はβの絶対値の和$\sum_{j=1}^{p}|\beta_j|$なので、こちらはl_1ノルムである。ラッソ回帰の損失関数Lは未定乗数をλとして次の通りである。

$$L(\boldsymbol{\beta},\lambda)=(\boldsymbol{e},\boldsymbol{e})+\lambda\sum_{j=1}^{p}|\beta_j| \qquad (2.8)$$

ラッソ推定量の回帰係数は closed form で示すことができないので反復推定しなければならない。ハイパーパラメータのλを大きく指定すれば、βの絶対値の和を小さくする必要が強まり、その結果として 0 の値の回帰係数が増える。ラッソ回帰を用いることでデータ数nより変数pが多いスパースデータにおいて、変数選択と回帰推定を同時に行うことができる。

なお(2.7)と(2.8)は、どちらもハイパーパラメータを 0 に設定すれば OLS に帰着する。

2.3　ラッソ推定量の性質

本節では森本による実験結果と考察を報告する。

■ラッソ推定量の目的

ラッソ回帰によって予測モデルがデータにオーバーフィッティング(過学習)することを避けることができる。R 言語に用意されている cars データセットを用いて実験を行った。まず図 2.1 に車の速度と停車するまでの距離の散布図を示した。このデータに 9 次の多項式を OLS で適合させた結果、くねくねと曲がった破線のグラフになった。図 2.1 で多項式回帰と書いた曲線である。学習データにモデルを過剰に適合させた結果として検証用データでは決定係数が下がることがある。これを汎化性能の悪化というが車に関するデータでも確かめられた。一方、同じ9次の多項式の分析データを分析しても罰則項を加えたラッソ回帰の場合は車の速度と停車するまでの距離が単調増加する滑らかな線になった。ラッソ回帰の方が合理的な結果を導いていることは明らかである。

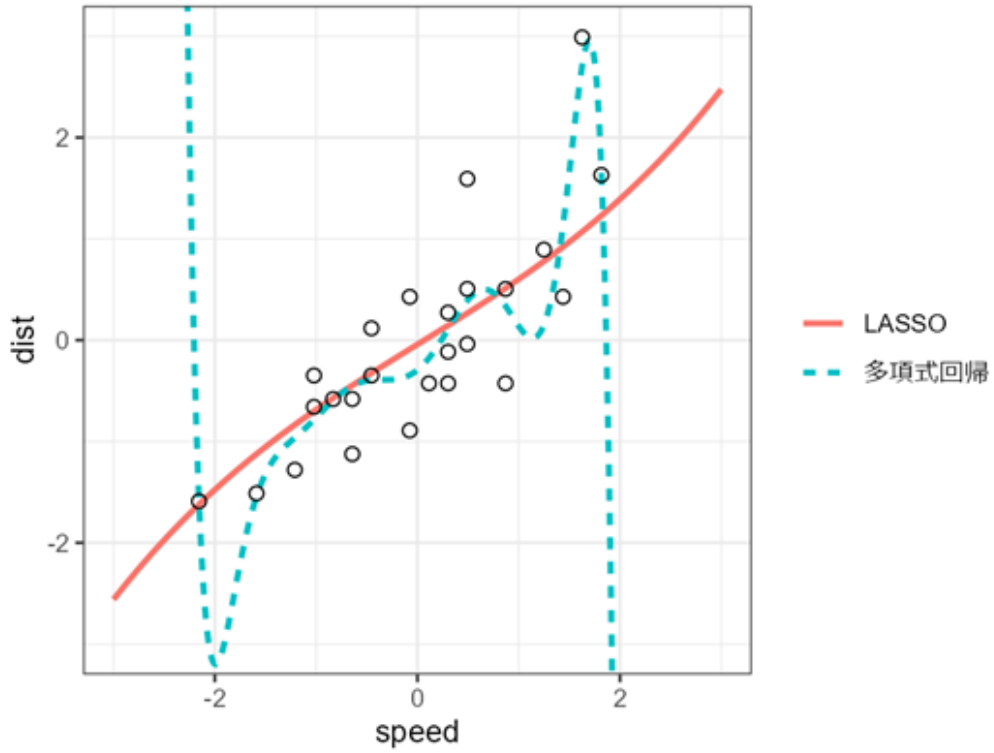


図 2.1　線型回帰とラッソ回帰の予測値の比較

リッジとラッソの推定量の相違を図 2.2 と図 2.3 で比較しよう。

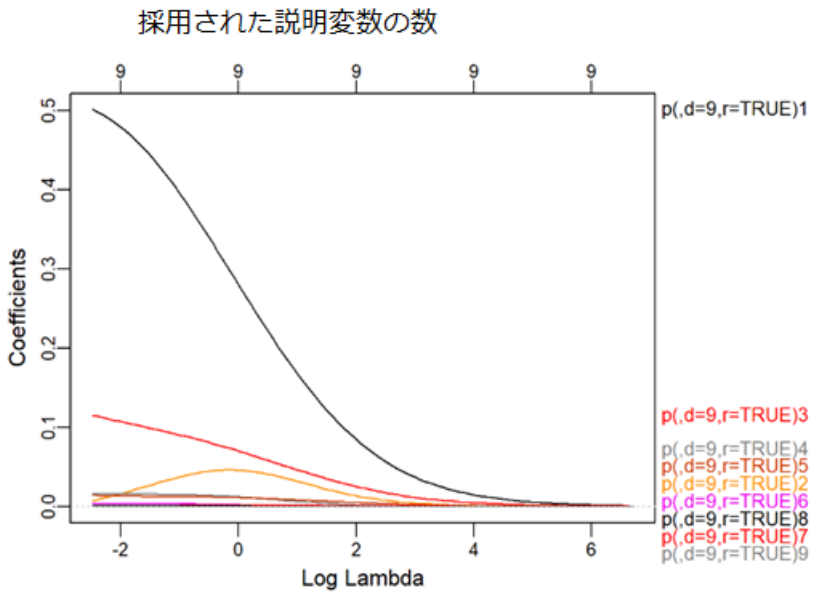


図 2.2　リッジ推定の回帰係数の変化

図 2.2 はハイパーパラメータを変化させながら、各説明変数の回帰係数の変化を描いたパスグラフである。この図の上部に書かれた 9 という数字は、リッジ推定した

結果が $\beta_j \neq 0$ となった説明変数の数を示している。リッジ回帰では変数の絞り込みは出来ていないことが分かった。もし変数選択をしたければ、さらに何らかの統計的検定を行うか主観的な変数選択に任せるしかない。

リッジ推定では説明変数についてすべて回帰係数を推定する性質がある。仮に誤って説明変数 X_1 と同じデータを X_2 にコピーした場合でも、2 つの説明変数に $0.5b_1X_1 + 0.5b_1X_2 + \cdots$ として回帰係数を 2 変数に与える結果になる。このような場合は、一方の説明変数を予め除いて、残った説明変数の回帰係数を推定する方が、解釈上も予測システムの運用上も望ましいはずである。

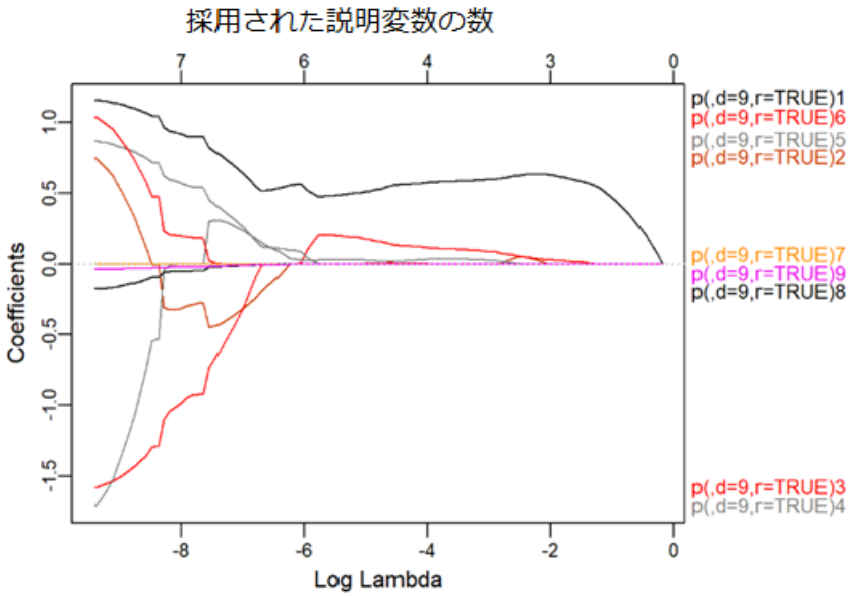


図 2.3 ラッソ推定の回帰係数の変化

一方で図 2.3 のラッソ推定は $\log\lambda = -2$ つまり $\lambda = 0.135$ の条件で3つの説明変数が $b \neq 0$ として採用され、その他の変数はカットされた。$n << p$ の事態においてはリッジ推定と比べてラッソ推定の優位性が明らかだろう。

■ ラッソ回帰のバリエーション

適応的ラッソ（Adaptive Lasso）やエラスティックネット、その他にいろいろなバリエーションが提案されている。適応的ラッソは説明変数ごとに回帰係数の絶対値の逆数を重みとして事前に与える方法である。$n << p$ の事態では重回帰分析が実行できないのだから、この方法は利用できない。便宜的な手段として単回帰分析を

p個の変数について繰り返すことは可能だが、それが全変数における正しい重みなのかについては異論があるだろう。

エラスティックネットはl_1, l_2 の両方のノルムを罰則項にするリッジとラッソのハイブリッド版である。精緻化にともなってハイパーパラメータが増えるというデメリットもあるので、これも望ましいかどうか異論があるだろう。

■ 適切な λ を決める方法

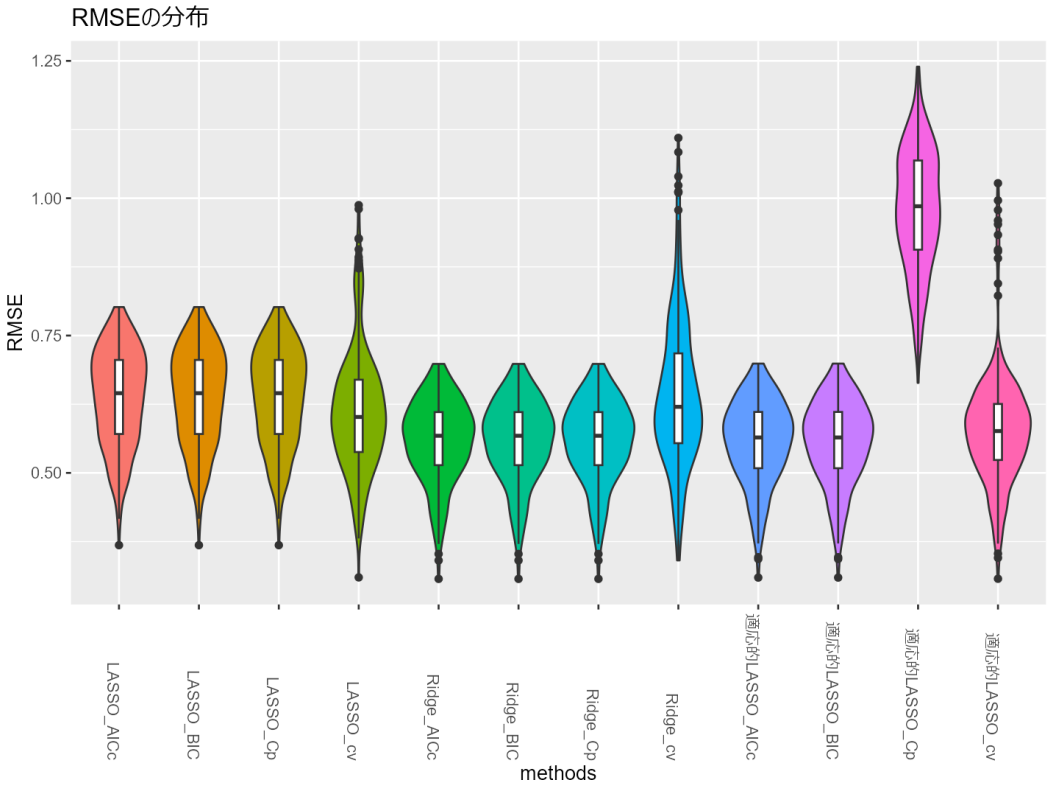
スパース推定を実際に利用する上で重要なのが λ の決め方である。λ は正則化項をどの程度効かせるかを決めるハイパーパラメータであり、より予測誤差が小さくなる λ を選択するのが望ましい。

今回は λ を決める方法として交差検証法(cross validation、 以下 cv)と情報量規準について、その安定性を確認する実験を行った。

R 言語に用意されている cars データセットを用いて、目的変数を制動距離、説明変数を速度とした 9 次の多項式について、線形回帰(多項式回帰)、リッジ回帰、ラッソ回帰、適応的ラッソ回帰でモデルを作る。

λ の選択には 10-fold cv、情報量規準(AICc、BIC、Cp)の各手法を用いる。これを学習用データと検証用データのサンプリングを変えながら 300 回試行した。

出来上がったモデルの汎化性能を RMSE(Root Mean Squared Error、二乗平均平方根誤差)で評価し、その分布から λ の選択方法の安定性を確認した。



いずれのモデルでも cv を用いた λ の選択では比較的バラツキが大きいことが分かる。また、適応的ラッソと情報量規準 Cp の組み合わせでは汎化性能が低くなっている。それ以外の情報量規準では比較的安定した λ の値が選択されていて、大外れはしない様子が分かる

３章　パターン認識

本章では、歴史的に生物統計の分野から生まれた判別分析と経済学の分野から生まれたロジットモデルを検討する。判別分析は、対象が問題にしているセグメントに該当するか否かを判定するために利用される。一方ロジットモデルは対象が各セグメントに属する確率を知るために利用される。

３.１　２群判別のタイプ分け

■ ２群判別分析の４区分

フィッシャー(1936)は判別分析において、対象の全体が排反で悉皆な ２ つの集団から成っていて、各集団が期待値 $\boldsymbol{\mu}$、分散共分散行列 $\boldsymbol{\Sigma}$ の多変量正規分布に従うことを仮定した。確率分布の母数を標本データから推定する場合はそれぞれ $\boldsymbol{m}$, $\boldsymbol{S}$ の記号を使う。これが ２ 群判別の典型的な事態といえる。本節では汎距離とベイズ判別ルールの視点を組み入れて２群判別を４区分して整理しよう。

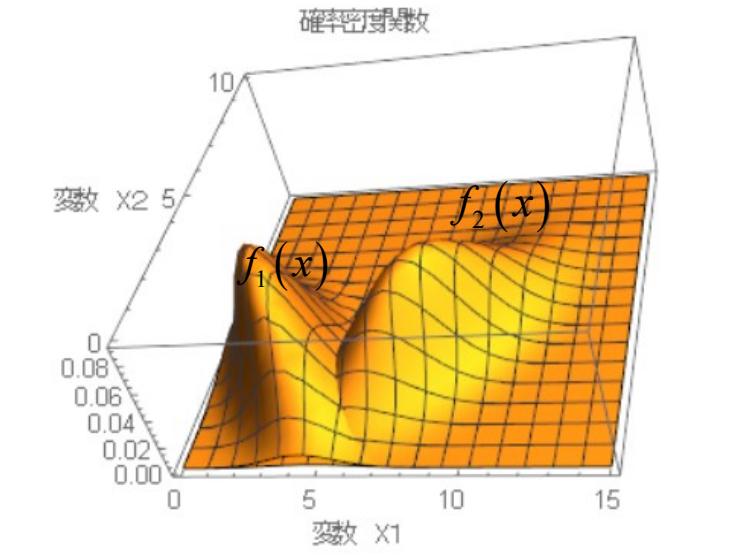


図 3.1 2群の多変量正規分布

2群 g_1, g_2 の多変量正規分布の密度関数は図 3.1 のように描かれる。図 3.1 は $p=2$ の場合を図示したものである。説明変数 X を特定の観測データ $\boldsymbol{x}$ に固定すれば、$f_1(\boldsymbol{x}), f_2(\boldsymbol{x})$ は一点 $\boldsymbol{x}$ における2つの多変量正規分布の密度関数の値を示す。次に2群判別のモデルを次の2つの基準から区分する。

1）分散共分散行列が両群で等しいか否か
2）群のサイズに関する事前確率 $\boldsymbol{\theta}$ を与えるか否か

この 2）で事前確率を与えて判別するのがベイズ判別ルールである。群1の出現確率を θ $(0<\theta<1)$ とすれば、群2の出現確率は $1-\theta$ になる。したがってサンプル $\boldsymbol{x}$ が群 1 である可能性が群2である可能性を上回るのは、$f_1(x)\theta > f_2(x)(1-\theta)$ の場合と考えられるので、次式が成り立つ場合になる。ただし分母は 0 にならないと仮定する。

$$\frac{f_1(x)\theta}{f_2(x)(1-\theta)} > 1$$

両辺の対数をとれば $$Q = \log\frac{f_1(x)}{f_2(x)} + \log\frac{\theta}{1-\theta} > \log 1 \qquad (*)$$

$\log 1 = 0$ なので、(*)の Q が正であれば、そのサンプルを群 1 に、それ以外の場合は群2に判定すればよい。次に(*)内の $\frac{f_1(x)}{f_2(x)}$ を汎距離で記述しよう。まず群によって分散共分散行列 $\boldsymbol{S}_1, \boldsymbol{S}_2$ が異なるという一般的なケースを示す。多変量正規分布の密度関数の比はg群の平均 $\boldsymbol{m}_g$ から一点 $\boldsymbol{x}$ までの汎距離 $D^2\left(\boldsymbol{x}, \boldsymbol{m}_g\right)$ の差で表すことができる。なお $|g_1$ は群で条件をつけたことを示す記法である。

$$\frac{f_1(\mathbf{x})}{f_2(\mathbf{x})}=\frac{|S_1|^{-1/2}}{|S_2|^{-1/2}}\exp\left[-\frac{1}{2}(\boldsymbol{x}-m_1)'S_1^{-1}(\boldsymbol{x}-m_1)+\frac{1}{2}(\boldsymbol{x}-m_2)'S_2^{-1}(\boldsymbol{x}-m_2)\right]$$

$$=\sqrt{\frac{|S_2|}{|S_1|}}\exp\left[\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2|g_2)-D^2(\boldsymbol{x},\boldsymbol{m}_1|g_1)\right\}\right]$$

上式を(*) に代入すれば一般的な判別式(3.1)が導かれる。

$$Q_1=\log\sqrt{\frac{|S_2|}{|S_1|}}+\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2|g_2)-D^2(\boldsymbol{x},\boldsymbol{m}_1|g_1)\right\}+\log\frac{\theta}{1-\theta}\qquad(3.1)$$

Q_1が正ならそのサンプルを群1に判定すればよい。次に(3.1)の特殊な場合を示そう。まず群サイズに関する情報がない場合は(3.1)の第 3 項を除いて(3.2)になる。

$$Q_2=\log\sqrt{\frac{|S_2|}{|S_1|}}+\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2|g_2)-D^2(\boldsymbol{x},\boldsymbol{m}_1|g_1)\right\}\qquad(3.2)$$

次に(3.1)において$S_1=S_2$ の場合は、(3.1)の第1項が 0 になり、2 群に共通したS を用いて汎距離を測ることになる。

$$Q_3=\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2)-D^2(\boldsymbol{x},\boldsymbol{m}_1)\right\}+\log\frac{\theta}{1-\theta}\qquad(3.3)$$

さらに群のサイズの情報がない場合は(3.4)になる

$$Q_4=\frac{1}{2}\left\{D^2(\boldsymbol{x},\boldsymbol{m}_2)-D^2(\boldsymbol{x},\boldsymbol{m}_1)\right\}\qquad(3.4)$$

以上 4 種類の判別分析を整理したのが表 3.1 である。

表 3.1　2 群判別分析の4区分

分析法の名称	分散共分散行列	群サイズθを指定	群サイズを指定しない
QDA: quadratic discriminant analysis	$\boldsymbol{S}_1$, $\boldsymbol{S}_2$	Q_1 (3.1)式	Q_2 (3.2)式
LDA: linear discriminant analysis	2 群に共通する$\boldsymbol{S}$	Q_3 (3.3)式	Q_4 (3.4)式

判別分析が汎距離で表せることは竹内・柳井(1972)、田中・脇本(1983)、梅津ら(2020)などでこれまで指摘されてきたことである。しかし表 3.1 による整理は本報が

はじめてである。

■ 判別境界

後藤は正規乱数に従う人工データをもとに表 3.1 のQ_1 とQ_3 を比較した。

データ数は群 1 を 300 サンプル、群 2 を 200 サンプルとした。変数は X1,X2 の 2 変数である。$\theta = 0.6$として判別した結果を図 3.2 に示す。 このデータでは 2 群が非球状に分布しており、しかも分散共分散が 2 群間で異なるので、Q_1 の QDA が妥当であると予想される。シミュレーションで比較した結果、LDA の誤判別率は 0.048 に対して、QDA の誤判別率は 0.024 であり、予想通り QDA の方が優れていた。

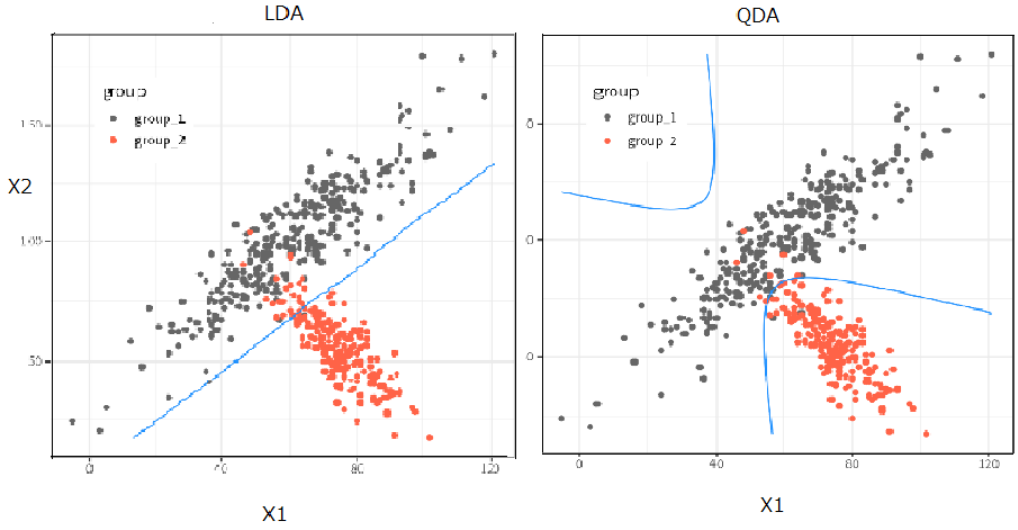


図 3.2 LDA と QDA の判別境界

図 3.2 に描いたように LDA の判別境界は直線、QDA は曲線になる。

ここで LDA で用いる分散共分散をどう計算したかを分析データを例に説明すると次の通りであった。まず 2 群のS_1, S_2 は下記のとおり共分散の正負の符号が異なっていた。

$$S_1 = \begin{bmatrix} 402.2 & 458.0 \\ 458.0 & 647.7 \end{bmatrix},\ S_2 = \begin{bmatrix} 90.8 & -113.7 \\ -113.7 & 220.9 \end{bmatrix},\ S_{total} = \begin{bmatrix} 326.3 & 76.6 \\ 76.6 & 949.5 \end{bmatrix}$$

3 番目のS_{total} は全データを一括して分散共分散を求めたものである。このように全体と部分で結論が異なることをシンプソンズパラドックスという。次のS_{pool} は、

LDA を実行するために S_1, S_2 を加重平均して共通の S を推定したものである。

$$S_{pool} = \begin{bmatrix} 277.8 & 229.5 \\ 229.5 & 477.2 \end{bmatrix}$$

R の分散共分散の cov( )関数は、平方和を(データ数-1)で割って不偏分散を求めている。したがって R の cov( )を使って分散共分散を不偏推定したければ、2 群のサイズを n_1, n_2 、全データをnとして次式で S_{pool} を推定すればよい。

$$S_{pool} = \frac{1}{n-2}\left\{(n_1-1)S_1 + (n_2-1)S_2\right\} \qquad (3.5)$$

一方 $\boldsymbol{S}$ を最尤推定する場合は(3.6)で $\boldsymbol{S}_{ML}$ を推定する。この計算法もしばしば使われている。

$$\boldsymbol{S}_{ML} = \frac{1}{n}\left(n_1\boldsymbol{S}_1 + n_2\boldsymbol{S}_2\right) \qquad (3.6)$$

■ セグメンテーション戦略と事前確率 θ の指定

Q_1, Q_3 では事前確率 θ を判別に利用する。それが望ましいかどうかは対象にしている市場に依存するだろう。市場が2つのサブ集団に分かれていることは確からしいとしても、そのどちらがドミナントな集団かはまだ定まらない、というのであれば θ を無理に指定することはない。その一方で市場が成熟化してマーケットシェアが把握できるのであれば、θ を用いることで判別精度を高めることができる。

θ を指定する際の深刻な問題は、従来のマーケティングの実務では分析データにおける群の構成比を $\boldsymbol{\theta}$ に扱うことが慣例化していた点にある。市販のプログラムによってはそれがデフォルト処理になっていることもあり、無意識的に慣例になってきたのかもしれない。なにが問題かというと、調査したデータが市場全体のよい縮図になっているとは限らないからである。データを収集する仕組みからしてデータの構成が市場全体と乖離することがある。その一例をあげれば、来店者だけを調査しても離反客と未顧客の調査データは得られない。

また意図的に偏ったデータを収集することもある。たとえば自社ブランドのユーザーを他社ブランドユーザーよりも手厚く調査することもあれば各ブランドのユーザーを同数ずつ調査することもある。

θ の指定を誤ると、学習データの判別には成功しても、現実のマーケットでは通用しないことになる。

3.2 多群判別への拡張

■ ベースラインロジットモデル

G 個の群のなかから任意の一群を baseline category に指定して、他の群と比較して「対数密度関数比」を求める方法がベースラインロジットモデルである。比較する組み合わせ数は$G-1$対になる。

表 3.1 に準じて多群判別をケース分けしたのが表 3.2 である。ここで群サイズを指定しないことと、$\theta_g = 1/G$と仮定したことは同じ効果がある。

表 3.2 ベースラインロジットの4区分

分析法の名称	分散共分散行列	群サイズ $\boldsymbol{\theta}$ を指定	群サイズを指定しない
BQDA: baseline QDA	$\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_G$	Q_1^* (3.7)式	Q_2^* (3.8)式
BLDA: baseline LDA	全群の $\boldsymbol{S}$ が等しい	Q_3^* (3.9)式	Q_4^* (3.10)式

表 3.2 の 4 ケースごとに群 G をベースラインにおいた対数密度関数比を示す。

$$Q_1^* = \log\sqrt{\frac{|S_G|}{|S_g|}} + \frac{1}{2}\left\{D^2\left(\boldsymbol{x}, \boldsymbol{m}_G \middle| G\right) - D^2\left(\boldsymbol{x}, \boldsymbol{m}_g \middle| g\right)\right\} + \log\frac{\theta_g}{\theta_G} \qquad (3.7)$$

$$Q_2^* = \log\sqrt{\frac{|S_G|}{|S_g|}} + \frac{1}{2}\left\{D^2\left(\boldsymbol{x}, \boldsymbol{m}_G \middle| G\right) - D^2\left(\boldsymbol{x}, \boldsymbol{m}_g \middle| g\right)\right\} \qquad (3.8)$$

$$Q_3^* = \frac{1}{2}\left\{D^2\left(\boldsymbol{x}, \boldsymbol{m}_G\right) - D^2\left(\boldsymbol{x}, \boldsymbol{m}_g\right)\right\} + \log\frac{\theta_g}{\theta_G} \qquad (3.9)$$

$$Q_4^* = \frac{1}{2}\left\{D^2\left(\boldsymbol{x}, \boldsymbol{m}_G\right) - D^2\left(\boldsymbol{x}, \boldsymbol{m}_g\right)\right\} \qquad (3.10)$$

(3.9)と(3.10)で必要になる分散共分散行列の不偏推定値は

$$\boldsymbol{S}_{pool} = \frac{1}{n-G}\sum_g \left(n_g - 1\right)\boldsymbol{S}_g \qquad (3.11)$$

任意の群間の対数密度関数比とベースラインロジットとの関係については Agresti(2002,p268)が次の関係を示している。群を a,b 、ベースラインを G として

$$\log\frac{f_a(\boldsymbol{x})}{f_b(\boldsymbol{x})}=\log\frac{f_a(\boldsymbol{x})}{f_G(\boldsymbol{x})}-\log\frac{f_b(\boldsymbol{x})}{f_G(\boldsymbol{x})} \tag{3.12}$$

この式は群間の対数オッズがベースラインの群をどこに置いても不変であることを意味している。

ベースラインとの比較対象がベースライン自身だった場合は、表 3.2 のどのケースも Q の値は 0 になる。したがってベースラインロジットの判定式は次の通り。

① G-1 個の対数密度関数比で正のものが複数あった場合は、対数密度関数比が最大の群にそのサンプルを所属させる
② G-1 個の対数密度関数比で正のものが 1 つあった場合は、その群にサンプルを所属させる
③ G-1 個の対数密度関数比がすべて負の場合は、そのサンプルを G 群に所属させる

本節では(3.7)～(3.10)に4通りの判別式を示した。BLDA は BQDA よりも計算が簡単になる。しかし等分散共分散を仮定することが無理な市場が多いだろうから、計算が楽だからといって正当な判別法になるとはいえない。

3.3　ロジットモデルによる確率予測

■ なぜ確率予測が必要か

マーケティングの課題によっては Yes か No かの判定ではなく個々の対象が群に所属する確率が必要になることがある。

近年では発生が極めて希なマーケット、たとえば超富裕層のマーケティングも必要になっている。希少ターゲットのマーケティングにおいては、5 分 5 分を閾値にして顧客を選別することは不適切かもしれない。そのような場合は、確率的な予測ができるロジットモデルを用いることに実務的な価値があるだろう。

■ ロジットモデルの系譜

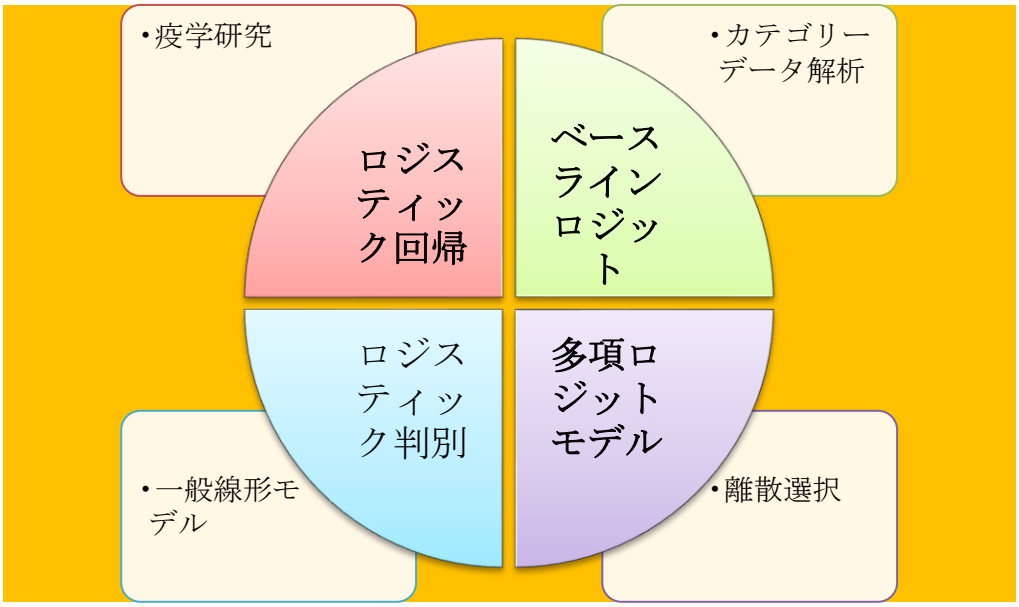


図 3.3　ロジットモデルの系譜

図 3.3 に示すようにロジットモデルについては様々な分野において異なる目的で研究が進められてきた。Truett ら(1967)の疫学研究に端を発したロジスティック回帰分析や、McFadden(1974)の多項ロジットモデルがよく知られている。

反応関数にロジスティック関数を使うという共通点はあるものの、原因系の説明変数 X については特定の確率分布を仮定するモデルもあれば仮定しないモデルもある。誤差の分布も正規分布に限らず二重指数分布を仮定するモデルもある。図 3.3 のすべてを束ねる統一理論がありえるのかは今後の研究課題としたい。

■ セミパラメトリックな推定

ベイズ判別ルールから LDA と QDA が導かれることは 3.1 節で見てきた。LDA は 2 群のサンプルが、共通の共分散行列を持つ 2 つの正規分布 $f\left(\boldsymbol{\mu}_g, \boldsymbol{\Sigma}\right)$ に従って発生していると仮定した。この場合の対数密度比(3.13)(3.14)から判別ルールが定められる。

$$\log\left\{\frac{\theta_1 f(x|\mu_1,\Sigma)}{\theta_2 f(x|\mu_2,\Sigma)}\right\} = \boldsymbol{b}'\boldsymbol{x} + c \qquad (3.13)$$

ただし

$$\boldsymbol{b} = \Sigma^{-1}\left(\mu_1 - \mu_2\right), \quad c = -\frac{1}{2}\left(\mu_1 - \mu_2\right)' \Sigma^{-1}\left(\mu_1 - \mu_2\right) + \log\frac{\theta_1}{\theta_2} \qquad (3.14)$$

ここで(3.13)の右辺は回帰分析と同じくパラメータ $\boldsymbol{b}, c$ の線形モデルになってい

る。したがって多変量正規分布の母数を推定することなく、p+1 個の回帰係数を推定するだけで判別は可能である。梅津ら(2020)はこの方法をロジスティック判別(logistic discriminant analysis)と呼んだ。(3.13)式は、左辺をリンク関数とした一般線形モデルとして理解することもできる。

線型モデルから群への所属確率を出すには、(3.14)を変形した次式を用いればよい。

$$P(Y=1)=\frac{1}{1+\exp\{-(\boldsymbol{b}'\boldsymbol{x}+c)\}},\quad P(Y=0)=\frac{1}{1+\exp\{(\boldsymbol{b}'\boldsymbol{x}+c)\}}$$

大橋はベイズ判別とロジスティック判別のシミュレーション実験を行い、次の知見を得た。

- 各群が正規分布に従い共分散行列が共通する場合は LDA が最良
- 各群が正規分布に従うが共分散行列は異なる場合は QDA が最良
- それ以外の正規性仮定から外れた場合は、ロジスティック判別が最良

あらためて考えると LDA と QDA は判別境界を引くという目的からすれば「無駄な」パラメータを推定していた。2次元空間のシミュレーション実験では、LDA ないし QDA が優れた結果になることがあっても、説明変数の数が多くなるに従いどこかでロジスティック回帰が勝るというのが大橋の予想である。

3.4　EC サイトのアップリフト効果

本節は松本による研究報告である。

■ EC サイトにおけるマーケティングの概要

マーケティングを効率化するための戦略として、1.優良顧客維持、2.既存顧客のランクアップ、および 3.新規顧客獲得がある。これら 3 つの戦略の中でも 1.と 2.に関連する戦略が費用対効果の側面から優先度が高いと言われている。一方、サービスのスタートアップ段階では、新規顧客獲得が重要な戦略となっている。

新規顧客獲得の重要性について過去の事例から考えよう。まずは共同購入型クーポン(フラッシュマーケティング)の事例から。2008 年に米国でスタートした Groupon が 2010 年に日本に参入した。それにともない、国内でもリクルートがポンパレのサービスを開始した。一見、利益率が高いビジネスに見えるため、多くの企業が参入し 2011 年には 200 サイトを超える乱立状態となった。次から次へと参入してきたサービスは統廃合を繰り返し、最終的には、Groupon とポンパレの2強状態になった。続いて、○○pay。paypay の 20%祭りをきっかけに一気に様々な企業がスマホ決済に参入してきた。乱立状態の○○pay がどうなっていったかは記憶に新しい。

これらの事例から学べることは「先行者の利益が強く、後からの参入組は辛い」ということである。いかに消費者を獲得するかというタイミングにおいては利益度外視

でとにかくスタートダッシュに成功するかどうかが重要な戦略となる。やがて、新規獲得数の伸びも限界になるにしたがい企業は既存顧客にシフトしていく。戦略として、優良顧客維持と既存顧客のランクアップを目指したクーポンやポイント施策などが考えられる。しかし、施策の効果を計量的に評価することは一般的に難度が高く、AB テストを利用した施策評価を行っている事例は多いものの、計量的に施策効果を評価している事例は少ない。

インターネットマーケティングは、オンラインマーケティングやデジタルマーケティングとも呼ばれ、E メールや SNS といったデジタルチャネルを通じて行われるマーケティングのことである。インターネットマーケティングの特徴は、次の 2 つである。

特徴 1) 消費者の行動のプロセスをトレースできる

インターネット上では、行動ログデータを利用することにより、消費者が購入した商品だけでなく、閲覧し購入した商品、閲覧したが購入に至らなかった商品、閲覧しなかった商品を把握することができる。この点は、スーパーマーケットなどでの消費者の購買行動の結果を示す ID 付き POS データと大きく異なっている。ID 付き POS データでは、来店して購買したということはデータとして獲得できるが、他の店舗への来店やどの商品を比較対象にしたのかといったことはトレースできない。

特徴 2) 個別対応のマーケティングが容易に実現できる

インターネットマーケティングでは、一人一人の違いを踏まえた施策を実現しうる仕組みになっている。そのため、一人一人の行動を精緻に捉えることができれば、従来の手法よりもコスト効率のよい効果的なマーケティング活動が実現できるのである。

■ クーポンにおけるアップリフト効果

クーポンの効果はその実施時と非実施時の成果変数の差として評価することになる。アップリフト効果の測定には、本質的に解決困難な問題が内在している。理屈上、アップリフト効果は、同じ消費者に対して施策を実施した場合と実施しなかった場合の成果をそれぞれ測定し、その差として評価しなければならない。しかし、ある個人に施策を打てば、施策を打たなかった時の成果を測定することはできず、その逆もまた同様である。実務における解析では、上記課題に対応する手段として、いくつかの顧客セグメントを構成し、そのセグメントごとにアップリフト効果を評価する。

具体的には、何らかの基準で顧客をセグメント化し、構成したセグメントごとに施策実施群(例えばクーポン付与)と施策非実施群(例えばクーポン非付与)に分ける。その上で実際に施策を実施し、セグメントごとに上記のアップリフト効果を評価する。最終的に最もアップリフト効果の大きいセグメントを見出し、以降の施策に用いる。実際に、このアプローチで実施した施策は効果がある。

より施策効果を高めるために、より粒度の細かいセグメントを構成することも考えられるが、1 セグメントあたりの対象者数が少なくなってしまい、単純なアプローチではクーポンによるアップリフト効果を正しく評価できない。細かい粒度の顧客セグメントでも精度高くアップリフト効果を評価するには、解析技術の高度化が必須である。One To One に近い状態で精緻な評価さえ実現できれば、より施策が有効に機能

する顧客を把握でき、結果としてマーケティング効率を高めることができる。アップリフトモデルについて様々な考えやモデリング方法が提唱されている。

■ 消費者の異質性を考慮したモデル

顧客ごとのアップリフト効果を計算する方法として、機械学習や統計モデルによるアプローチが考えられる。機械学習は一般的に精度が高いものの事後分析を丁寧に行わないと大きなミスを犯す可能性がある。企業がマーケティング施策として多くの金額を顧客に介入する場合は、いくらか精度を犠牲にしても説明性の高い（構造がわかりやすい）統計モデルが用いられる場合が多い。

アップリフトモデルに通常のロジスティック回帰分析を使ってもクーポンの異質性を表現することは難しい。

$$\mathrm{logit}(p) = \beta_0 + \beta_1 r + \beta_2 f + \gamma P$$

平均的なクーポン効果から、消費者の異質性を考慮した推定を行うために階層ベイズロジットモデルの枠組みで推定することが可能である。

図 3.4 は消費者が EC サイトを訪れるか否かを Y1 で表し、次に該当製品を購入するか否かを Y2 で表した階層モデルである。Y1 の確率と条件付き確率の積で同時確率が表される。

枝分かれのターミナルノードには「サイトに来てない」「購入なし」「購入あり」の3つの結末がある。それらを選択肢 3 つの中からの選択問題ととらえるのではなく、図 3.4 のように階層的な選択としてモデリングするのが妥当だろう。商業集積でのショッピングにおいても、施設の選択と店舗の選択という図 3.4 と同じメカニズムが働くと考えられる。図 3.4 は逐次 2 分割を表したものだが、一般的に逐次多肢分割でも最尤法でパラメータが推定できる。

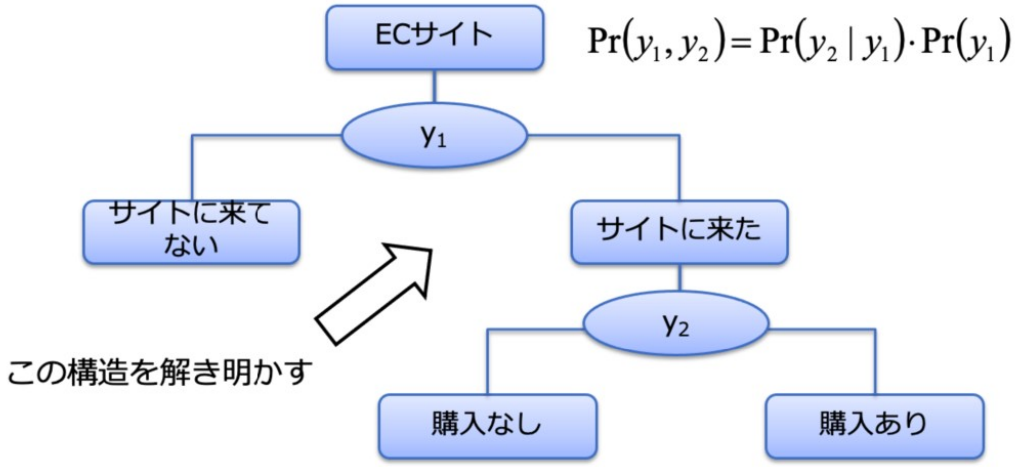


図 3.4 階層的な意思決定

ここではウェブ上の消費者の「ウェブサイトを閲覧する/しない」「商品を購入する/しない」といった消費者行動に焦点を当てよう。階層ベイズネスティッドロジットモデルの枠組みで、消費者のサイト閲覧行動と購入行動を同時にモデル化することにより、実務的な視点でクーポン付与施策を高度化することができる。

$$f(y_{i1}, y_{i2}) = f_1(y_{i2} \mid y_{i1}) f_2(y_{i1})$$
$$f_2(y_{i1}=1) = \frac{\exp(V_{i1})}{1+\exp(V_{i1})}, f_2(y_{i1}=0) = \frac{1}{1+\exp(V_{i1})}$$
$$f_1(y_{i2}=1 \mid y_{i1}=1) = \frac{\exp(V_{i2})}{1+\exp(V_{i2})}, f_1(y_{i2}=0 \mid y_{i1}=1) = \frac{1}{1+\exp(V_{i2})}$$

閲覧の効用関数

$$U_{iv} = V_{iv} + \varepsilon_{iv}$$
$$= \sum_{j=1}^{p_v} \alpha_{ji} x_{ji}^{(v)} + \alpha_{p_v+1,i} \cdot \log\{1+\exp(V_{ir})\} + \varepsilon_{iv}$$

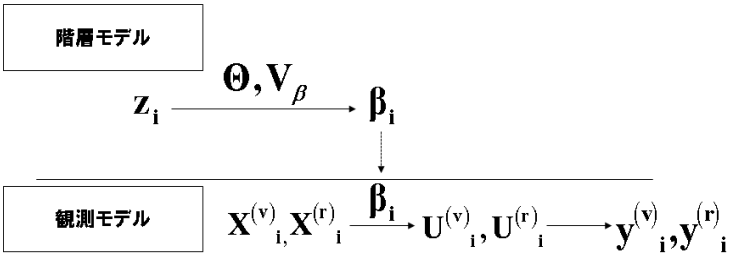
購入行動

$$U_{ir} = V_{ir} + \varepsilon_{ir}$$
$$= \sum_{j=1}^{p_r} \beta_{ji} x_{ji}^{(r)} + \varepsilon_{ir}$$

階層モデル

$$\boldsymbol{\gamma}_i = [\alpha_{1i}, \cdots, \alpha_{p_v i}, \beta_{1i}, \cdots, \beta_{p_r i}]^t$$
$$\boldsymbol{\gamma}_i = \boldsymbol{\theta} \cdot \mathbf{z}_i + \mathbf{v}_i, \mathbf{v}_i \sim \mathrm{MVN}(0, \boldsymbol{\Sigma})$$

【 階層ベイズネスティッドロジットモデルの枠組み 】



【 推定アルゴリズムのフローチャート 】

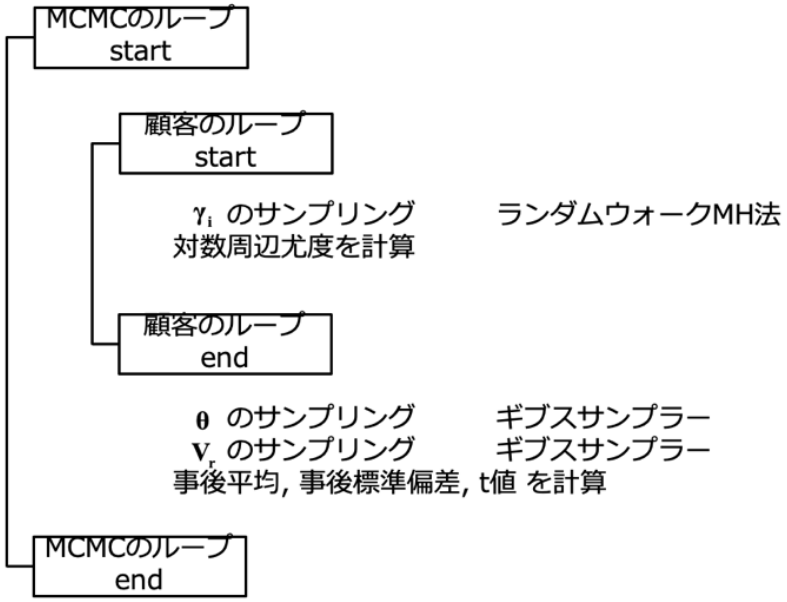


図 3.5　階層ベイズネスティッドロジットモデル

■　アップリフトモデルを用いたビジネスグロース

アップリフトモデルを用いて解いている問題は、以下の 2 点となる。

a. クーポンをもらわなくても購入する人(オーガニック)を除外する

b. クーポンをもらうことで購入する人を見つけ、クーポンを付与する

ID	クーポンを付与した場合	クーポンを付与しない場合	
1	購入する	購入する	a.のモデルで発見したい人
2	購入する	購入しない	b.のモデルで発見したい人
3	購入しない	購入しない	クーポン付与の影響はない

なぜ、アップリフトモデルを使うと ROI が良くなるか？

a.のモデル

$$ROI = \frac{\sum gain}{\sum cost}$$

仮に、クーポンを付与しなくても購入する人を見つけることができたとする。その人にクーポンを付与しなくて済むので、その数だけコストが減る。

b.のモデル

$$ROI = \frac{\sum gain + \sum gain'}{\sum cost + \sum cost'}$$

仮に、クーポンを付与することで購入する人を見つけることができたとする。利得が増えるが、同時にコストも増える。

3.5　非線形写像による SVM

Cortes と Vapnik (1995)が提唱した SVM は、ディープラーニングが普及する以前の 1990 年代には最も強力なパターン分類とされてきた。SVM は今日でも学習データ数が少なく、データの分布が重なっている場合に有用な方法である。

■　カーネルによる違い

入力空間(input space)のままでは群を線形分離できない散布状況であっても、高次元の特徴空間(feature space)に原データを射影することで線形の分離平面を求めることができる[3]。これは回帰分析でいえば変数をデータ数-1 まで増やせば決定係数が 1 になって完全に予測できるのと同じ理屈である。

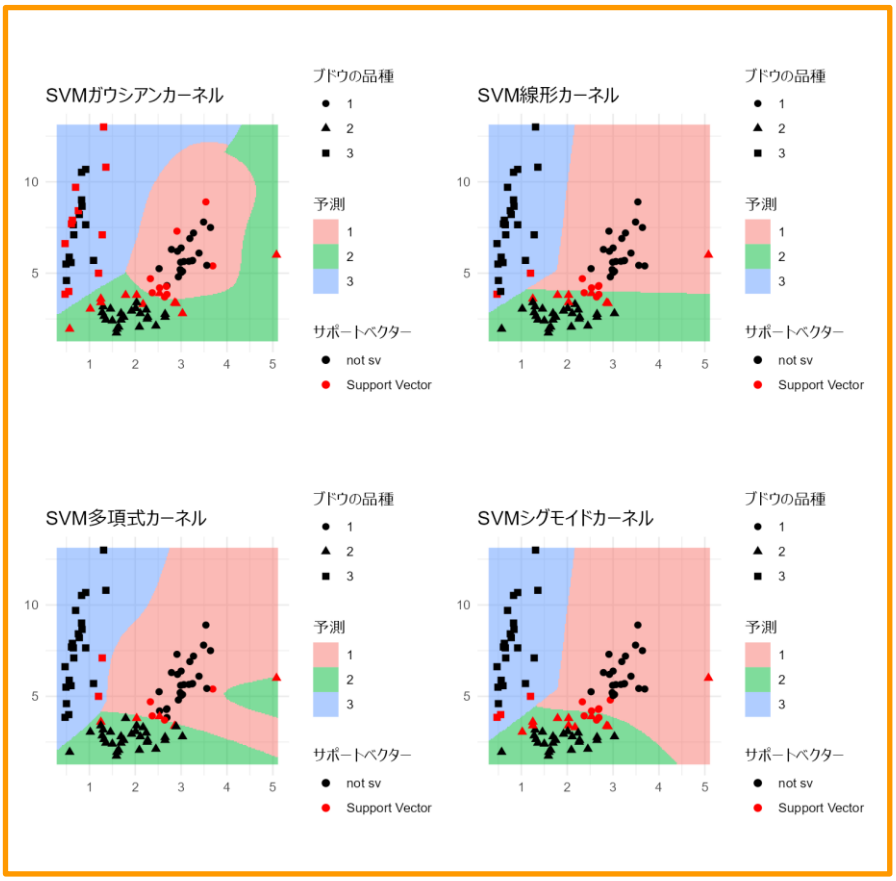


図 3.6 カーネルによる分離の違い

森本は3品種のブドウをフラボノイド含有量, 色強度の2変数を入力変数として分類を行った。入力空間では入り混じっていた3品種のブドウが高次元空間へ射影することにより分類が可能となった。入力変数の空間に逆写像した図を見ると、利用したカーネル関数によって分類の境界が変化することが分かる。この実験から得られるSVM利用上のヒントは次の4点であろう。

1）判別の精度は学習用データではなく、検証用のデータで評価するのがよい。つまり汎化性能の良さで比較するのである。

2）誤分類率が一番小さいカーネルを採用して運用に用いる。図3.6では多項式カーネルだった。

3）図3.6では群の境界を越えてデータがプロットされているようにみえるが、それは高次元の散布図を入力空間に逆写像したことによる錯覚である。

たとえば上空から都市を俯瞰すればビルがあってその地下には地下鉄が通っていることがある。しかし平面だけでなく高度という第 3 の次元を加えれば、建物と地下鉄は分離されているはずだ。高次元への射影というのはそういう意味なのである。

4）図3.6では境界を定めるのに用いられたサポートベクターを赤色で区別した。判別分析では全てのデータを使って判別境界を求めていたが、SVM では判別境界付近のサンプルだけで判別境界を決めている。判別境界から遠く離れた領域においてデータがどう分布していようとも、判別境界の決定には関係しないという意味である。すべてのサンプルを使うことは意味がなく、ボーダーラインに近いサンプルを重視して線形判別空間を定めるというのがSVMの着眼点だった。大学合格を目指した受験予備校の教育方針もSVM的な性格があるかもしれない。

■他の機械学習との比較

大屋は次の2要因について大型のシミュレーション実験を行った。学習データの選別の偏りをなくすために50%の学習データを100回リサンプリングした。汎化性能を比較するために、テストデータにおける誤判別率を求めた。結果は100回出るが、その中央値で評価した。実験要因は次の2つである。

1）手法の相違：線形カーネルSVM、ガウスカーネルSVM、シグモイドカーネルSVM、決定木、ランダムフォレスト、アダブースト、バギングの7水準

2）学習データの割合：25%、50%、75%の3水準

以上の7×3水準を組み合わせで実験を行った。

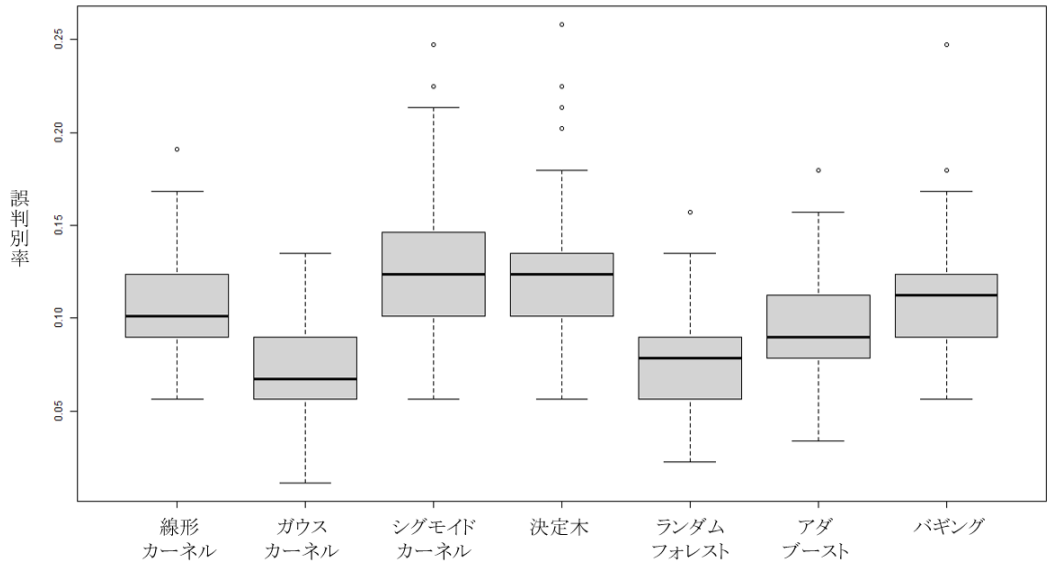


図 3.7　誤判定率の箱ヒゲ図

今回の実験で誤判別率が最も低かったのは、ガウスカーネル SVM を使った場合で 1.12%であった。なおカーネル法については高野（2020）の研究がある。

機械学習の研究課題として次の 2 点があげられる。

①チューニング後のアンサンブル学習の精度検証、

②スパース・データ（n≪p）での精度検証

４章　ディープラーニング

CNN（畳み込みニューラルネットワーク）や VAE（変分オートエンコーダ）は深層学習の手法であり、特に画像認識やパターン認識などにおいて優れた性能を示す。CNN は教師あり学習の一形態であり、分類タスクにおいてはラベル付きのトレーニングデータを用いて特徴抽出とクラス分類を同時に行うことができる。しかしながら、ラベル付きデータの入手は一般的に困難であることから、データのラベリングに関連するコストが課題となる。また CNN そのものは学習データに含まれていない新たな画像の生成は行えない。

一方、VAE は教師なし学習の手法で、データ内の潜在的な構造を明示的にラベル付けすることなく捉えることができる。VAE はエンコーディングと呼ばれる手続きの中で、データの構造を抽出する際に CNN の手法を用いて特徴を圧縮し、潜在変数を確率分布化することでその特徴を規定する。この潜在変数の確率分布から、画像の類似性や近接性をクラスタリングすることができる。また、任意の確率変数を与えることで、学習データには含まれていない画像の生成や再構築もデコーディングと呼ばれる手続きを通じて可能となる。

マーケティングの領域においては、今後、画像を介した相対取引やマーケット分野では、顧客の属性と画像データを潜在空間にマッピングし、そこで顧客の位置付けやクラスタリングを行うことで、顧客の嗜好や行動パターンを理解することのほか、不正取引などの異常検知への応用が期待される。

4.1　CNN

1980 年に福島は階層的なニューロンネットワークモデルであるネオコグニトロンを発表した。これは、視覚神経科学（Visual Neuroscience）における単純細胞と複雑細胞という 2 種類のニューロンで構成されているモデルに着想を得たもので、単純細胞は局所的な特徴を抽出、複雑細胞はそれらの特徴を統合する役割をもつもので、このネオコグニトロンにより、手書き文字認識が可能となることが示された。

1990 年に LeCun らは手書き数字の低解像度画像を分類するためにバックプロパゲーションで訓練された畳み込みネットワークに関する技術を発表した。これらの研究成果をもとに、1998 年に同じく LeCun らのグループによって CNN が発表された。CNN は、畳み込み層とプーリング層という 2 種類の層を組み合わせることで、画像の特徴を効率的に抽出する。CNN は、ネオコグニトロンとの構成は類似しているが、ネオコグニトロンモデルにはない誤差逆伝播法（バックプロパゲーション）など教師あり学習アルゴリズムを導入することでネットワークの重み共有を図り、画像認識の精度向上の進展が図られた点に特徴がある。

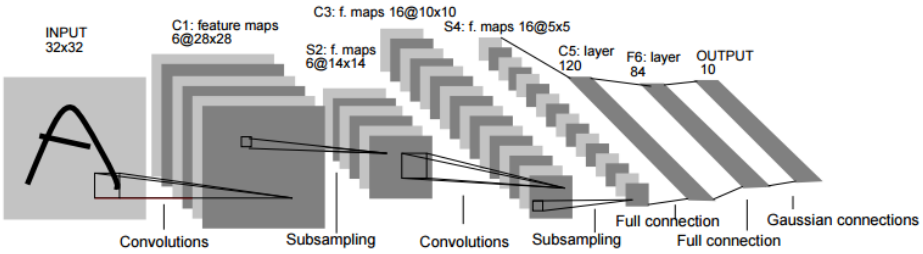


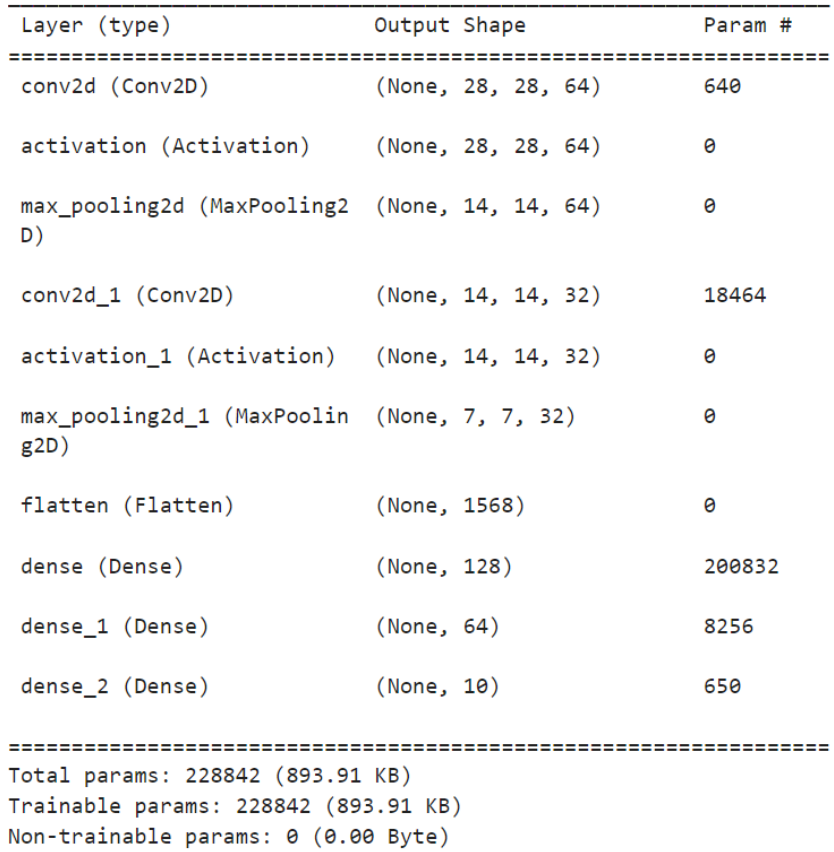
図 4.1　LeCun らによって発表された畳み込みニューラルネットワークのアーキテクチャ（LeNet-5）。出所は LeCun ら(1990)

■Keras による LeNet-5 の追試

1998 年に提案された畳み込みニューラルネットワーク（CNN）のアーキテクチャは LeNet-5 と呼ばれるもので図 4.1 となっている。当時の手書き数字認識のためのベンチマークデータセットである MNIST を対象として設計された。LeNet-5 のアーキテクチャは以下のようになっている。

① 入力層（Input Layer）：32×32 ピクセルのグレースケール画像を受け取る。
② 畳み込み層（Convolutional Layer）：畳み込みフィルター（カーネルとも呼ばれる）を使用し、畳み込み演算を実行する。それぞれのフィルターによって特徴マップが生成される。
③ サブサンプリング層（Subsampling Layer）：発表当時はサブサンプリング層として表記されていたが、現在は「プーリング層」と呼ばれることが一般的になっている。ここでは、2×2 のマックスプーリングを使用して、特徴マップから特徴値をサンプリングする。特徴の位置感度が低下し、計算量が削減される。
④ 2 番目の畳み込み層とサブサンプリング層：同様の構造で、畳み込みフィルターとそれに続くサブサンプリング層が構成される。
⑤ 全結合層（Fully Connected Layer）：全結合層で畳み込み層とサブサンプリング層からの特徴を結合し、それを使用して出力を生成する。段階的に縮約し、最終的な出力層は 10 個の変数からなるように絞り込まれる。これは、0 から 9 までの 10 個のクラスに対応している。

実用上の LuNet-5 のアーキテクチャの構築と性能を確認するため、MNIST の 60,000 枚の画像データをもとに Python の Keras ライブラリを用いて追試を行った。ここでは教師データ（Training Data）を 50,000 枚、検証データ（Test Data）を 10,000 枚として分割して実施した。



```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 conv2d (Conv2D)             (None, 28, 28, 64)        640

 activation (Activation)     (None, 28, 28, 64)        0

 max_pooling2d (MaxPooling2  (None, 14, 14, 64)        0
 D)

 conv2d_1 (Conv2D)           (None, 14, 14, 32)        18464

 activation_1 (Activation)   (None, 14, 14, 32)        0

 max_pooling2d_1 (MaxPoolin  (None, 7, 7, 32)          0
 g2D)

 flatten (Flatten)           (None, 1568)              0

 dense (Dense)               (None, 128)               200832

 dense_1 (Dense)             (None, 64)                8256

 dense_2 (Dense)             (None, 10)                650

=================================================================
Total params: 228842 (893.91 KB)
Trainable params: 228842 (893.91 KB)
Non-trainable params: 0 (0.00 Byte)
```

図 4.2　Keras を用いた LuNet-5 のアーキテクチャ構成。パラメータ数は 228,842。

また、追試には Keras の開発者である Chollet の教本 (2018)とそのサンプルコード[1]を参照した。

LuNet-5 の Keras におけるアーキテクチャは図 4.2 の通りで、パラメータ総数は約 23 万に達した。損失関数は多分類判別であることからカテゴリー交差エントロピーを用い、オプティマイザーは rmsprop 法を選択した。また、最適化にあたりサブマッチ数は 128 とし、30 エポックのイテレーションを実行した。

追試の結果、通常の CPU（インテル® Core™ i5-1155G7 プロセッサー）上での計算速度は 1 エポックあたり 20 秒程度であり、30 エポックの処理は約 10 分で完了した。

学習データ（黒線）および検証データ（赤線）の損失値および精度のエポックに対する依存性を図 4.3 に示す。これより、学習データからはおおよそ 5 回

[1] https://github.com/fchollet/deep-learning-with-python-notebooks

目のイテレーションで精度は 99%に達し、その後も漸増的に精度が向上していることが確認された。一方で検証データでは同じく 5 回目のイテレーションで 99%の精度に達するが、それ以上はイテレーションを回しても精度向上はみてとれない。このことは損失値でも同様の傾向にある。これらのことから、LuNet-5 では約 5 回のイテレーション（2 分程度）で 99%以上の分類精度を確保する性能が得られることが確かめられた。

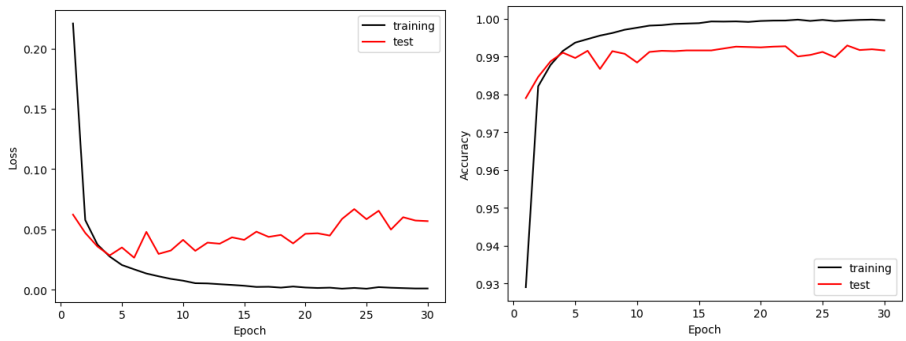


図 4.3　LuNet-5 の損失値（左）および精度（右）のエポック数の依存性

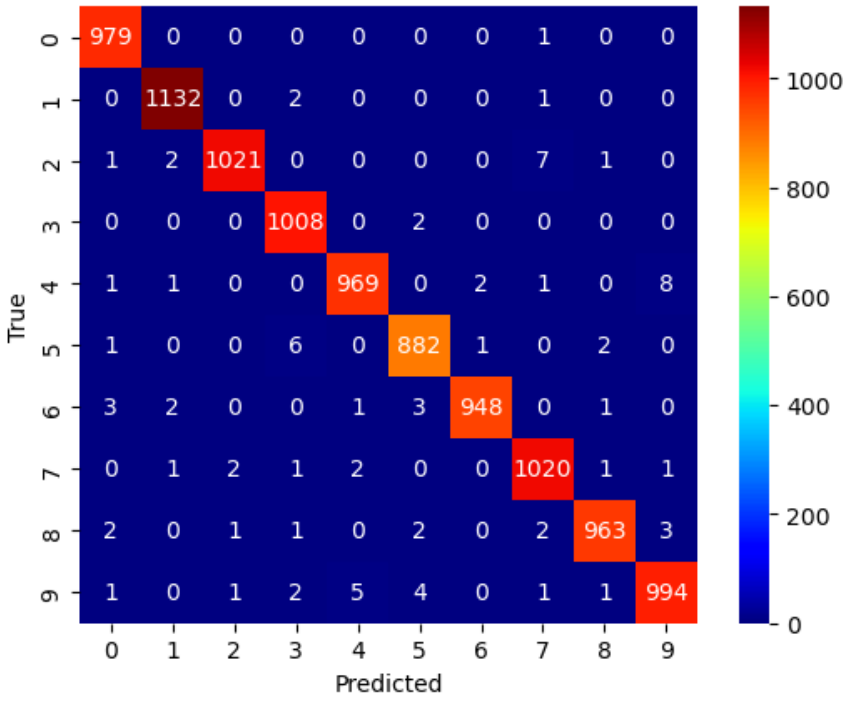


図 4.4　LuNet-5 による MNIST の識別能力

実際の検証データ 10,000 枚の識別能力を図 4.4 に示す。誤認識が多いパタ

ーンについて、5 回以上の画像を個別に確認すると、「２」の画像が「７」と誤って識別されるケースが 7 件あり、また、「４」が「９」と誤って識別されるケースが 8 件あった。さらに、「５」の画像が「３」と誤って識別されるケースが 6 件あり、そして「９」の画像が「４」と誤って識別されるケースが 5 件あった。しかし、これ以外の誤判別は数枚のレベルにとどまっていることがわかる。

このように、LeNet-5 の基本構造だけでも優れた画像の認識と識別が行えることが確認された。

4.2　生成モデル

2006 年に Hinton らが提案したオートエンコーダ（AE）の手法は、ニューラルネットワークの手法を用いた非線形の次元削減の方法である。このことは、当該研究が発表された論文の書き出しに次のように記されていることでも理解される。

『高次元データの分類、可視化、情報伝達、保存を容易にする手法が次元削減である。その中でも広く使われている手法の一つが主成分分析（PCA）である。PCA はデータセット内の分散が最も大きい方向を見つけ、各データポイントをそれらの各方向に沿った座標で表現する。今回、私たちは PCA の非線形一般化を提案する。この手法では、多層の適応型「エンコーダ」ネットワークを使用して高次元データを低次元のコードに変換し、同様の「デコーダ」ネットワークでコードからデータを復元する』

図式的には図 4.5 に示されるような入力データをニューラルネットワークの手法で隠れ層（Hinton らの論文では Code 層と呼んでいる）へ次元削減し、その隠れ層から再びニューラルネットワークの手続きから出力データを復元するという 2 段階のモデルである。

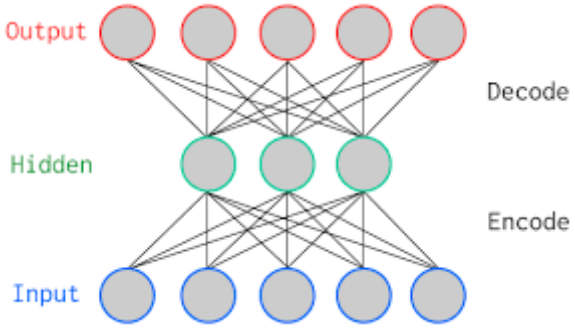


図 4.5 オートエンコーダの模式図。（図は CC-BY-SA のもと Wikipedia より参照）https://ja.wikipedia.org/wiki/オートエンコーダ#/media/ファイル:AutoEncoder.png

これにより入力データよりも低次元な隠れ層（中間層）をデータの特徴が圧縮されている潜在空間として情報を保持させるだけでなく、デコーダという中間層から再構成の処理を行うことで非線形の次元圧縮情報でありながらデータの復元や再生が可能となることが示され、かつ画像データにおいては PCA と比較をしてもその復元における再現性が高いことが示された。

よく知られているように、非線形の次元削減は情報損失を伴い、一旦、次元削減された方法から逆変換をかけて元の情報を復元することは困難である。例えば代表的な Hinton ら（2002）の t-SNE（t-distributed Stochastic Neighbor Embedding）や McInnes ら（2018）の UMAP（Uniform Manifold Approximation and Projection）は、高次のデー タ空間の情報を低次元に非線形的に削減する手法では、低次元におけるクラスタリングに行いやすさ（分離性能）が PCA よりも高いことで知られているが、元の高次の情報への復元はできない。

このようにオートエンコーダは非線形の次元削減の手法でありながら、次元圧縮と同時に、その情報から元の情報に近い形で復元を可能にする手法として画期的であったが、その潜在空間の意味付けは、特に画像のようなデータを入力データとした場合にはその解釈が困難となる課題が認められた。

2013 年に発表された Kingma らの変分オートエンコーダ（VAE）は、これらの課題を克服する。VAE は、エンコーダの出力層に平均と分散のベクトル値を出力して中間層（潜在空間）（z）に保持する。続いて画像再現においては、これらの潜在空間に保持されている確率分布のパラメータからサンプリングを行い、デコーダにおいて元のデータを再構築する。一般に、その確率分布はガウス分布を仮定し、平均 0 を中心とした標準正規分布に付置する。

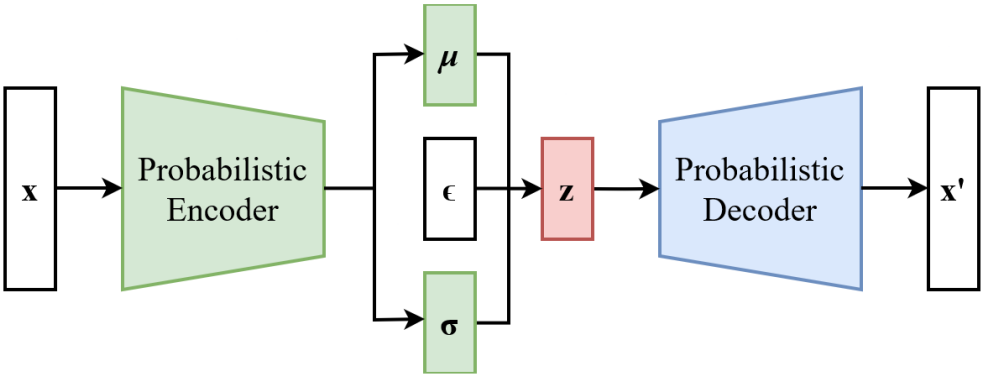


図 4.6 変分オートエンコーダの模式図（図は CC-BY-SA のもと Wikipedia より参照）
https://en.wikipedia.org/wiki/Variational_autoencoder#/media/File:Reparameterized_V

ariational_Autoencoder.png

VAE の主な利点は、潜在変数に確率分布を導入することで、潜在空間の連続性や解釈性を向上させることが可能になることにある。また、潜在空間内での操作によって入力データにはない新しい画像を生成することが可能であることが示されている。

まとめると、オートエンコーダ（AE）と変分オートエンコーダ VAE は、データの次元削減と復元というプロセスにおいて共通しているが、潜在空間におけるモデリングや、モデル評価（損失関数）においては以下のような差異がある。

	AE	VAE
潜在空間のモデリング方法	潜在空間の分布を明示的にモデル化しない。エンコーダは入力データを潜在表現に変換し、デコーダはその潜在表現を元のデータにデコードするが、潜在空間の構造や分布については何も仮定しない。	潜在空間の分布を明示的にモデル化する。エンコーダは入力データを潜在空間の確率分布（μ、Σ）のパラメータに変換し、デコーダはその確率分布からサンプリングされた潜在変数を使用してデータを生成する。
潜在表現の性質	データの圧縮や特徴抽出のために次元削減されるが、その表現は連続的であるかどうか、また意味のある変動を保証せず。	確率分布からサンプリングされるため、通常、連続的で意味のある変動を保証する。
損失関数の構成	再構築誤差（入力とデコードされた出力の差）を最小化するように学習	再構築誤差に加えて、KL（Kullback-Leibler）情報量も損失関数に含む。この項は、エンコーダの出力する潜在空間の分布と事前分布（通常はガウス分布）との間の一致の度合いを測る。

■Keras による VAE の追試

VAE における潜在表現の性質や表現を検証するため、MNIST の 60,000 枚の画像データを用いて Python の Keras ライブラリを用いた計算を実行した。AE の概念自体がいわゆる教師なし学習型の PCA の次元削減と類似していることから、VAE におけるモデル構築においても、CNN で行ったような学習データと検証データに分割するようなことはせず、MNIST の全画像データを使用した。なお、CNN の追試と同じように、ここでも Keras の開発者である F.

Chollet の教本とサンプルコードを参照している。

VAE の Keras におけるエンコーダおよびデコーダのアーキテクチャは、それぞれ図 4.7 と図 4.8 に示す配置とパラメータで設定が行われた。VAE でも畳み込み層は CNN と同様に使用されているが、プーリング層は配置されない。エンコーダを通過した後、平均値と分散値がそれぞれ計算され潜在空間に保持されるが、ここでの潜在空間の次元数は２次元としている。

```
Layer (type)                Output Shape              Param #   Connected to
==================================================================================================
input_1 (InputLayer)        [(None, 28, 28, 1)]       0         []

conv2d (Conv2D)             (None, 14, 14, 32)        320       ['input_1[0][0]']

conv2d_1 (Conv2D)           (None, 7, 7, 64)          18496     ['conv2d[0][0]']

flatten (Flatten)           (None, 3136)              0         ['conv2d_1[0][0]']

dense (Dense)               (None, 16)                50192     ['flatten[0][0]']

z_mean (Dense)              (None, 2)                 34        ['dense[0][0]']

z_log_var (Dense)           (None, 2)                 34        ['dense[0][0]']

sampling (Sampling)         (None, 2)                 0         ['z_mean[0][0]',
                                                                 'z_log_var[0][0]']

==================================================================================================
Total params: 69076 (269.83 KB)
Trainable params: 69076 (269.83 KB)
Non-trainable params: 0 (0.00 Byte)
```

図 4.7

図 4.7　Keras を用いた VAE のエンコーダのアーキテクチャ構成。二段の畳み込み層を経て全結合されたあと、その平均値と分散値が計算される。パラメータ数は 69076 にのぼる。

```
_________________________________________________________________
 Layer (type)                Output Shape              Param #
=================================================================
 input_2 (InputLayer)        [(None, 2)]               0

 dense_1 (Dense)             (None, 3136)              9408

 reshape (Reshape)           (None, 7, 7, 64)          0

 conv2d_transpose (Conv2DTr  (None, 14, 14, 64)        36928
 anspose)

 conv2d_transpose_1 (Conv2D  (None, 28, 28, 32)        18464
 Transpose)

 conv2d_transpose_2 (Conv2D  (None, 28, 28, 1)         289
 Transpose)

=================================================================
Total params: 65089 (254.25 KB)
Trainable params: 65089 (254.25 KB)
Non-trainable params: 0 (0.00 Byte)
```

図 4.8 :　Keras を用いた VAE のデコーダのアーキテクチャ構成。中間層からデータを受け取り、二段の逆の畳み込み層を経て復元される。デコーダに要するパラメータ数は 65,089 にのぼる。

損失関数は、再構築誤差である実際のデータと生成されたデータの差をバイナリ交差エントロピーとし、加えて KL 情報量を用いて潜在変数の確率分布と事前分布との間の類似性を評価した。オプティマイザーには adam 法を選択し、サブマッチ数は 128 とし、30 エポックのイテレーションを実行した。その結果、実行速度は 1 エポックあたり約 30 秒であり、30 エポックの処理は約 15 分で完了した。

図 4.9 は、MNIST の VCA によって次元圧縮された潜在空間の表現となる。非線形の次元削減であるため、横軸や縦軸には PCA のような意味を持たせることはできない。MNIST の 0〜9 の画像に関する分布は、カラーバーのコントラストを用いて表現されている。0〜9 の数字の分布は明確に 10 個の集合に分かれているわけではないが、例えば「0」の特徴は左下の象限に集中し、「1」の集合は右上の象限に固まっていることが観察される。他の数字のパターンは、「0」のように環を描く数字が「0」側に近い分布を示し、一方で非環の数字の「1」へとシフトするように配置されている。

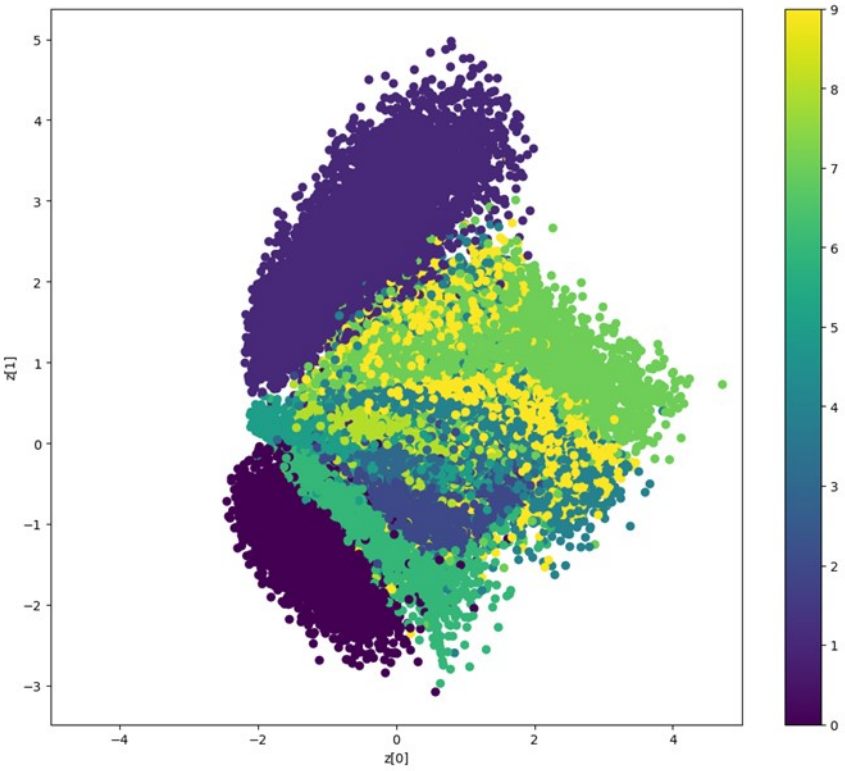


図 4.9　MNIST の VCA によって次元圧縮された潜在空間の表現

図 4.10 は上記の分布を 30×30 のグリッドで分割したときの画像を可視化して配置したものである。左下から y 軸に沿って移動すると、「0」→「6」→「5」→「8」→「7」→「1」と画像が遷移していることが観察される。なお、この図に示されている 900 枚の画像は MNIST の画像そのものではないことに留意されたい。これらの画像は、MNIST の入力データを潜在空間の確率分布（μ、Σ）のパラメータに変換し、デコーダはその確率分布から 30×30 の代表点でサンプリングされた潜在変数を使用して「生成」された手書き風の画像データとなっている。

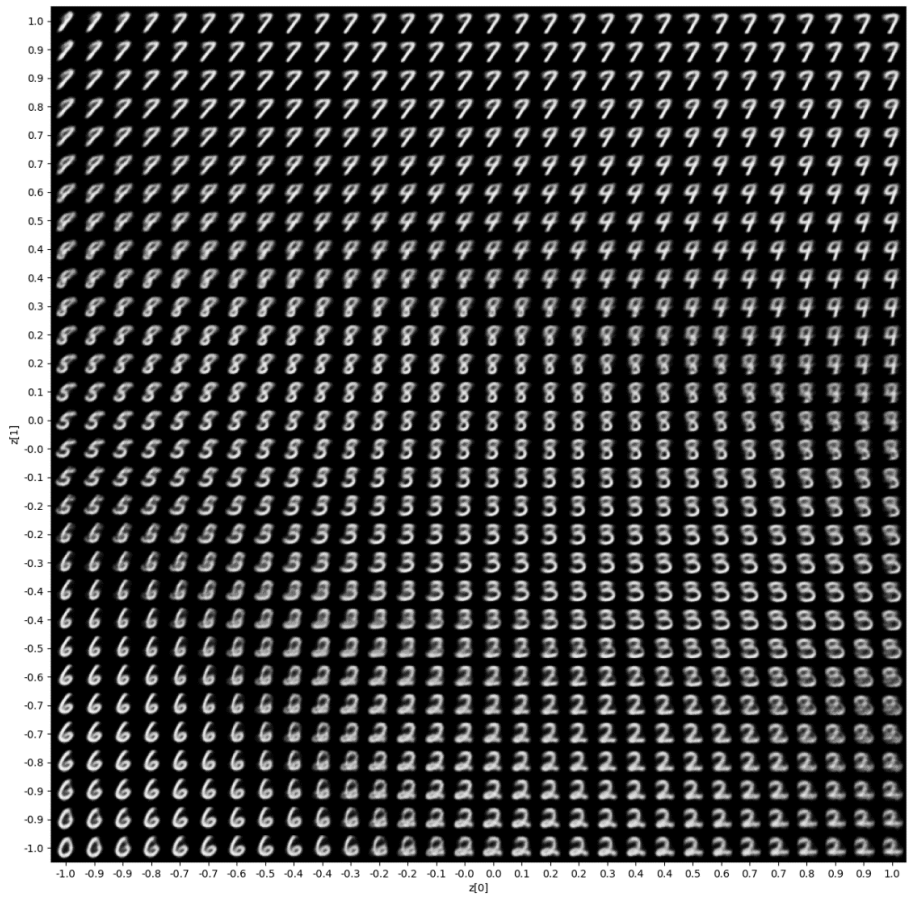


図 4.10　図 4.9 の分布を 30×30 のグリッドで分割した画像の可視化図

最後に、VAE のマーケティングにおける活用を意識して、MNIST のデータセットに代わり、Fashion-MNIST[2]のデータセットを使って同様のコード

[2] Xiao, H., Rasul, K., Vollgraf, R., Fashion-MNIST: a Novel Image Dataset for Benchmarking Machine Learning Algorithms. arXiv:1708.07747 (2017)

を実行した結果について示す。Fashion-MNIST は、MNIST の手書き風の画像に代わりファッションの商品イメージに類した画像集となっている[3]。Fashion-MNIST のデータ構造も MNIST でのコードがそのまま使えるように 10 種類の商品からなるラベルが付されていることのほか、6 万枚のトレーニングデータと 1 万枚のテストデータとすることでも MNIST とデータセットの規模を同じくしている。なお、ラベルと商品は次のように設定されている。

Label	Description	Label	Description
0	T-shirt/top	5	Sandal
1	Trouser	6	Shirt
2	Pullover	7	Sneaker
3	Dress	8	Bag
4	Coat	9	Ankle boot

図 4.11 は Fashon-MNIST を VAE で最適化された 15×15 のグリッドで配置した潜在空間における画像分布である。コードの構成はグリッドの配置を変えた以外は上記の MNIST と全く同一にしている。左下には「Ankle boot（アンクルブーツ）」が配置され y 軸に沿って移動すると、「Coat（コート）」から「T-shirt/top（T シャツ-トップス）」、そして最後は「Dress（ドレス）」に近い画像へと遷移している様子がわかる。

図 4.12 は、Fashon-MNIST の VCA によって次元圧縮された潜在空間の表現となる。カラーバーのコントラストは商品と対応するように色付けされている。ここでも MNIST と同様に明確に 10 個の商品の集合に分かれているわけではないが、例えば靴系の商品は下半分に、トップス系の商品は上半分に分布していることがわかる。

[3] https://github.com/zalandoresearch/fashion-mnist

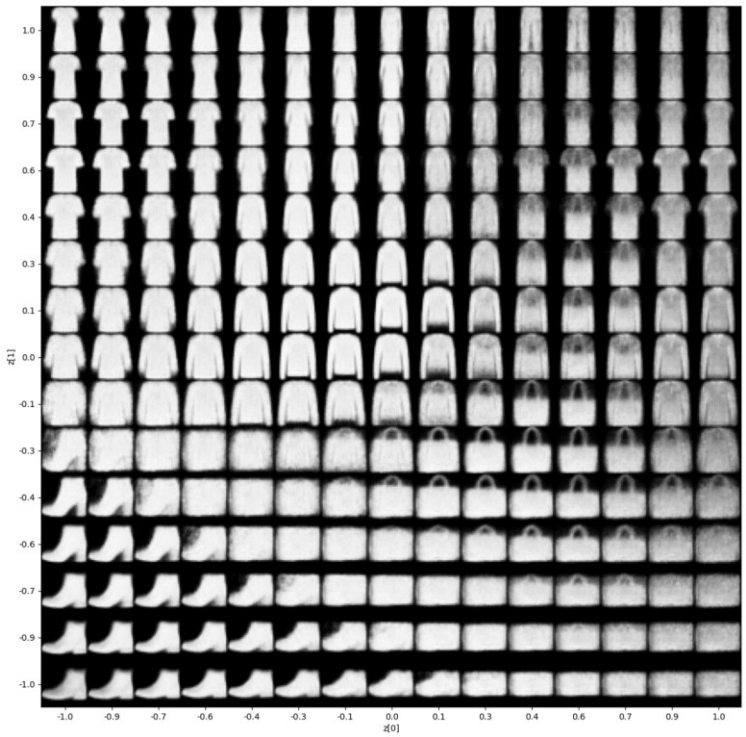


図 4.11： Fashion-MNIST の VAE による画像分布を 15×15 のグリッドで分割した画像の可視化図

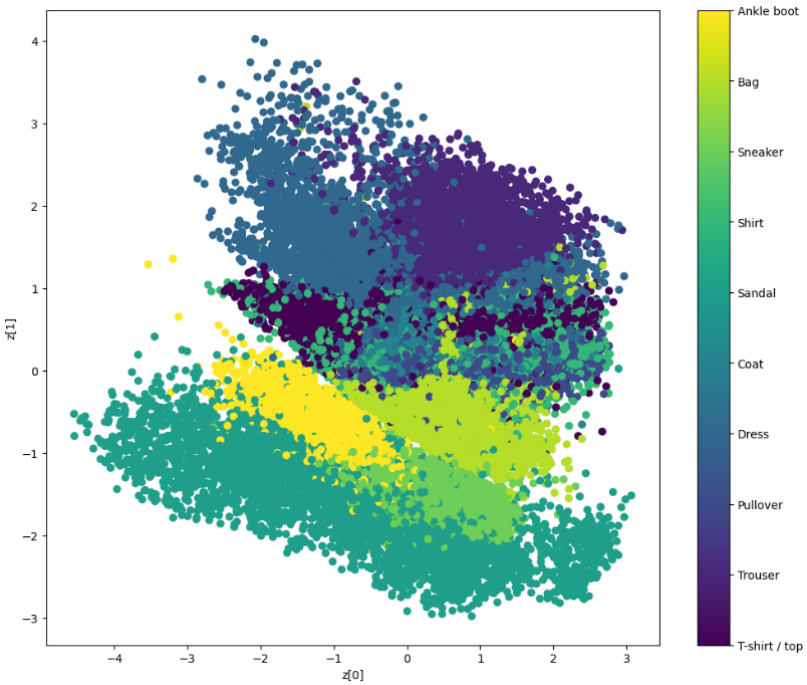


図 4.12： Fashon-MNIST の VCA によって次元圧縮された潜在空間の表現

５章　クラスター分析

クラスター分析は大きく分ければ、類似したサンプルどうしを段階的に併合して大きなクラスターにまとめる階層クラスター分析と、N 個のサンプルを K 個のクラスターに分割する非階層クラスター分析に分けられる。本章では大規模データの分析でよく利用されている非階層型クラスター分析を検討する。まず 5.1 節では非階層型のクラスター分析で代表的な k-means 法の問題点を指摘し、欠点を改良した k-umeyama 法を提案する。5.2 節ではクラスターの最適性基準を比べ、5.3 節でクラスター分析の今後の発展を論じる。

５.１　k+means 法の改訂

■　品質管理とマーケティングの違い

マーケティング・リサーチの実務では、消費者をセグメントする際に、まず原データを因子分析あるいは主成分分析にかけてからクラスター分析することが慣習のように行われてきた。因子分析の因子および主成分はいずれも直交するのだから、ユークリッド距離で距離を測ることが正当化されるはずだ、というのが実務の上では通説だったようである。

しかしデータを正規直交化しさえすれば変数間の相関情報は無視して構わないと言えるのは、データが単一の母集団から発生するという前提が成り立つ場合に限る。

たとえば品質管理においては、一つの生産ラインには唯一の目標仕様があって、そこを中心にしながら製品がほぼ正規分布にしたがって生産されると想定することが多い。しかしマーケティングで扱う市場は品質管理とは異なる。

マーケティングでは市場は複数の集団から成り、しかも集団間に異質性があることを想定する。もしそう想定するのであれば、各クラスターは平均の位置だけでなく分散共分散行列もクラスターごとに異なると想定するのは自然だろう。つまり因子分析や主成分分析をすれば変数間の相関は考慮する必要がなくなるというのは誤解である。データ全体を直交化したところで、それぞれのクラスターに所属するサンプルまで無相関に再配置されることにはならないからである。

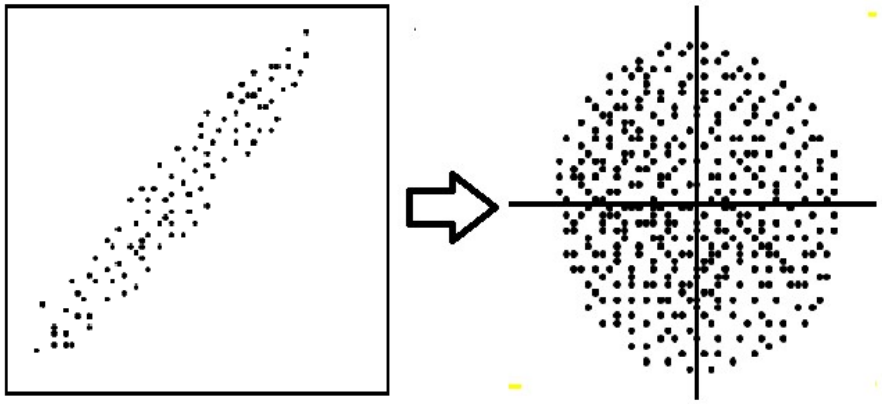


図 5.1　品質管理データの正規直交化

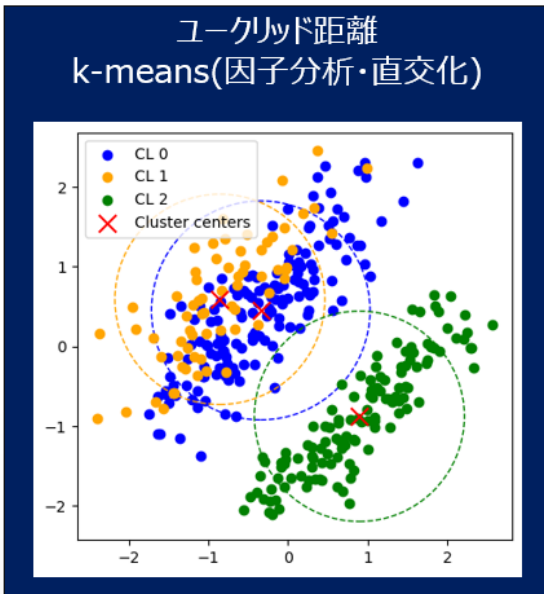


図 5.2　因子得点の散布図

図 5.2 の散布図をみれば、因子分析をしても集団ごとに spherical にデータが散布するわけではないのは明らかである。因子得点はクラスターごとに異なる相関をもつことが見てとれる。

本来クラスターとは異質集団の存在を前提にした概念だったはずだが、マーケティング実務における分析の進め方は論理的な矛盾があった。本章では複数のクラスターの中心からサンプルまでの汎距離を測りつつクラスタリングを進める新しい方法を提案する。具体的にはて Mahalanobis(1936)が提唱した汎距離ではなく朝野（2023）がいうところの汎距離ケース 6 を用いる。

■　従来のクラスター分析

k-means 法は非階層クラスター分析の中ではよく知られた方法である。k-means 法に関連して、これまでに表 5.1 に示す提案がなされてきた。まず MacQueen(1967)による原案は、変数間の相関を考慮せずに距離を測るために、相関の高い変数群に影響されてクラスター所属が決まるという問題があった。

Dunn(1974)のファジークラスタリングは、所属するクラスターへのオーバーラッピングを許容する分析法であった。クラスターへの所属度合いはメンバーシップ値によって表わされる。

Cerioli(2005)は各クラスターの分散共分散が異なる汎距離を提案していて、我々と問題意識を共有する。ただし表 5.1 に示すようにセリオリは初期シードの選択については提案ができなかった。最後の Arthur ら(2007)は k-means++ で初期シードを逐次選択する提案を行った。彼らにも欠点があったことを次に指摘する。

表 5.1　従来の非階層型 k-means 法

発表	名称	初期シードの選択	クラスター中心とサンプルの距離	各クラスターの分散共分散の利用
MacQueen(1967)	k-means	乱数で一括選択／ユーザー指定	ユークリッド距離	利用せず
Dunn(1974)	FuzzyC-Means	乱数で一括選択／ユーザー指定	ユークリッド距離の逆数をメンバーシップ値に	利用せず
Cerioli(2005)	修正k-means	提案なし	汎距離	異なるS g
Arthurら(2007)	k++	距離の2乗に比例した逐次選択	ユークリッド距離	利用せず

■　k-means 法の欠点

k-means 法には、まず第 1 に初期シードというクラスターの核を選択する方法に問題があった。

ユーザーにシードの位置を指定させるというオプションはユーザーにとって困難なタスクであった。とくに次元の高い空間においてシードを指定させるのは現実性が低い。もう一つのオプションである一括選択法は、独立な乱数でサンプルを選んでシードにする方法である。この一括選択は個々の抽出が独立に行われるために多次元空間の偏った領域にシードが集中する可能性が排除できない。偏ったシードから反復計算をスタートすると局所的最適値に収束する危険がある。実際に乱数を変えて一括選択してみれば、クラスタリングの再現性が低いことが容易に確認できる。

第 2 の欠点は、k-means 法に適したデータの分布が限定的だったことにある。セリオリは、k-means 法はデータが超球状に散布しており、各クラスターのサイズがほぼ等しいことを暗黙に仮定していると指摘した。つまりクラスターが非球状 non-spherical に分布し、しかもサイズが異なるというマーケティングで起きそうな市場に対して k-means 法を使うのは不適切だということになる。

■　初期シードに関する先行提案

アーサーらは、最初の 1 つのシードを乱数で選び、2 つ目のシードは、第 1 シードからの距離の二乗に比例した確率で選び、それ以降も逐次的に次のシードを確率抽出することを提案した。シード選定の段階ではまだ集団が形成されていないため、集団の中心からの汎距離が測定できない。そのためアーサーらはユークリッド距離を使っている。

k-means++のシーディングでは各サンプルと既存のシードとの距離 d(x)を測る必要がある。シードが 1 つしかない段階では、そのシードとの距離が d(x)である。シードが複数になると、既存の各シードとサンプルの距離を測って最短距離を d (x)に使う。そして（5.1）のウェイトに比例させて次のシードを確率抽出する。

$$w_i = \frac{d\left(x_i\right)^2}{\sum_{i=1}^{n} d\left(x_i\right)^2} \qquad (5.1)$$

k-means++法では、遠く離れたサンプルほど距離の二乗に比例して選ばれやすくなる。そのため最遠点の外れ値がシードに選ばれる可能性が最も高くなる。外れ値の近くにはデータポイントが何もない事態も起きるだろう。これがアーサーらの提案の欠点であった。

■　k-umeyama 法によるシード選択法

次に後藤・梅山(2023)が提唱した k-umeyama の詳細を示す。まず第iサンプルから最近隣のシードまでの距離d_i を求める。次に全サンプルについて距離の平均値$\bar{d}$ を求める。その後d_i をシグモイド関数で変換する。

$$y_i = \frac{1}{1+exp\{-a(d_i-\bar{d})\}} \qquad (5.2)$$

上式は項目反応理論における 1 母数ロジスティックモデルでありaは曲線の傾きを調整するパラメータである。aが大きくなるほど関数はステップ関数（階段関数）に近づく。最後に(5.3)に比例した確率で新たなシードを抽出する。

$$w_i = \frac{y_i}{\Sigma_{j=1}^{n} y_j} \qquad (5.3)$$

ユーザーが指定したクラスター数 K に達するまで以上の手順を繰り返して初期シードのセットを決める。

k -means++と k-umeyama 法で、シード選択がどう変わるかを図示したのが図 5.3 である。k-umeyama 法では外れ値をシードに選ぶ確率を相対的に抑制できることが分かる。図 5.3 右の k-umeyama 法なら 11 番目のサンプルが選ばれる可能性は 4，7，10 番のサンプルが選ばれる可能性と同等になる。

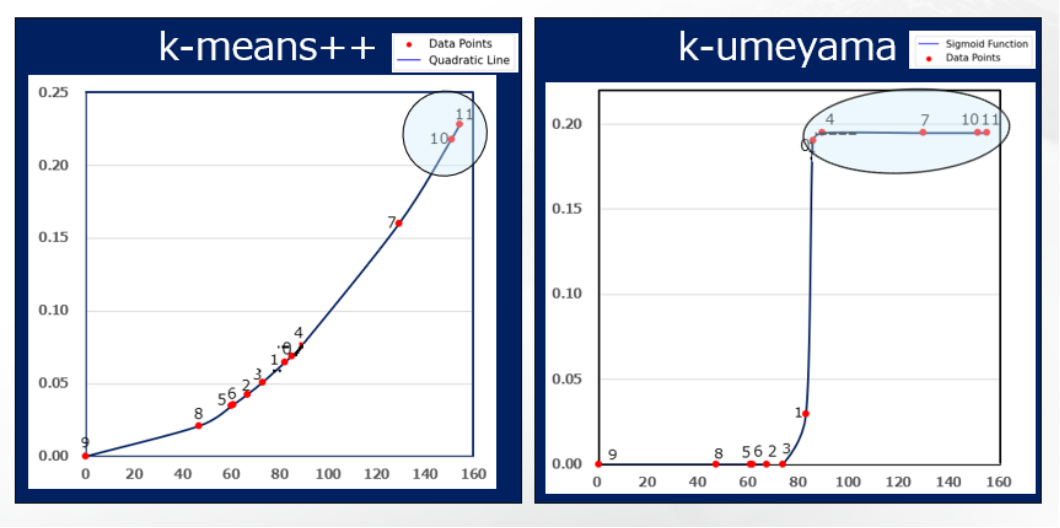


図 5.3　次のシードに選択されやすいサンプルの所在

■　汎距離計算の改良

汎距離を測れるのは 1 回目のクラスターが作られてから以降の繰り返し段階である。サンプル $\boldsymbol{x}_i$ と第 g クラスターの平均ベクトル $\boldsymbol{m}_g$ との汎距離は次式で測定する。二乗してもしなくても遠近の順序は変わらないので平方汎距離のまま測定する。

$$D^2\left(\boldsymbol{x}_i, \boldsymbol{m}_g\right) = \left(\boldsymbol{x}_i - \boldsymbol{m}_g\right)' S_g^{-1} \left(\boldsymbol{x}_i - \boldsymbol{m}_g\right) \qquad (5.4)$$

分散共分散行列 $\boldsymbol{S}_g$ が正則でない場合は $\boldsymbol{S}_g$ の逆行列が求められない。従って汎距離が計算できない。もしクラスター規模 n_k と変数の数 p が $n_k > p$ であったとしても $\boldsymbol{S}_g$ が正則になるとは限らない。たとえば評定尺度に対して「どちらともいえない」と回答する人が何人もいればデータ行列は退化することになる。セリオリはクラスター分析に汎距離を用いることを提案したものの、非正則な場合の対処法までは示さなかった。

k-umeyama 法ではムーアペンローズの一般逆行列をもちいて $\boldsymbol{S}_g$ の逆行列を計算する[4]。クラスター分析の反復計算の過程で、はじめは小規模のクラスターから始めて次第にクラスター規模を大きくする、というアルゴリズムを用いることもあるだろうし、そもそもサンプル数が多くないのにクラスター数Kが多い場合もある。そうした場合に一般逆行列を用いる必要が出てくる。

■ k-umeyama 法の計算手続き

分析のフロー	分析内容
①ランダムに1シードを選択	・ランダムにシードを一つ選択 ・シードとサンプルiの最短距離d_iを求める・全サンプルについて距離の平均値$\bar{d}$
②シグモイド関数で変換	$y_i = \frac{1}{1+exp\{-a(d_i-\bar{d})\}}$
③次のシードを抽出する確率のウェイト付け	$w_i = \frac{y_i}{\sum_{j=1}^{n} y_j}$
④クラスターの特性を計算	・初回だけユークリッド距離を測って近いサンプルをクラスターに所属させる ・クラスターごとに平均と分散共分散行列の一般逆行列
⑤平均からのマハラノビス距離を測ってクラスター所属を更新⇒④	$D^2(x_i, m_g) = (x_i - m_g)' {S_g}^{-1} (x_i - m_g)$
⑥収束判定・更新・終了	クラスターの平均値が変化しなくなったら更新を終了

図 5.4　k-umeyama 法の計算手続き

図 5.4 でユーザーが指定するハイパーパラメーターはクラスター数 K とシグモイド関数の形状を決める a の 2 つである。

■ 実データでの検証

ライフスタイルに関する自主調査の実データを使って、k-means 法と k-umeyama 法を比較した。具体的な手続きは次の通りである。

k-means	k-umeyama
◆使用データ ライフスタイル調査：生活分野への力の入れ具合14変数、1050サンプル	
◆ライフスタイル14変数（1～5段階）の因子分析（主因子解、バリマックス回転）	-
◆3因子での因子得点を算出	-
◆3因子の因子得点を使って k-meansで非階層クラスター分析	◆ライフスタイル調査の14変数（1～5段階）の原データのまま、k-umeyamaで分析
◆最適性基準比較	
◆ライフスタイル調査の14変数のクロス集計比較	

■ k-umeyama 法の精度検証

図 5.5 の通り、望大指標であるフリードマン・ルービンの基準では k-means が 5.27、k-umeyama 法が 7.05 と k-umeyama 法の方が優れていた。その他のウィルクスのラムダ、ピライのトレース、ホテリング・ローリーのトレースでも、k-umeyama 法が k-means 法よりも良い結果を示している。各指標の定義については 5.2 節で述べる。結論として、この分析事例では k-umeyama 法が k-means よりもクラスタリングの精度が優れていることが確認できた。

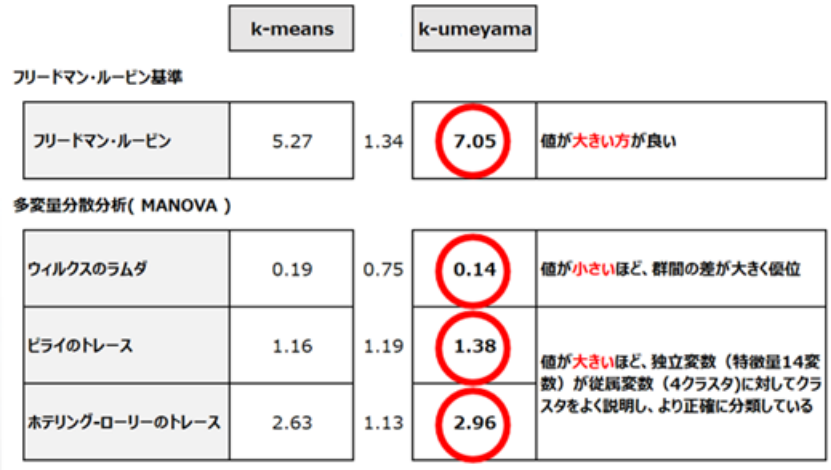


図 5.5　クラスターの最適性基準

■　画像修復への応用例

パターン認識の課題であるが、汚れた画像の修復に k-umeyama 法を適用した例を図 5.6 に示す。この左の図が欠損のある画像で、右が K=1024 で k-umeyama 法で修復した結果である。

図 5.6　画像修復への応用例　著作権： 国（文部科学省所管）

図 5.6 の画像修復の具体的なフローは下記の通り。

① 画像の分割：画像から$m \times n$ サイズのパッチを抽出し、これをベクトル化する。今回は8×8サイズのパッチを取り出したので画像のピクセルは 64 ピクセルとなる。 i 番目のパッチのデータをベクトルx_iで表現する。

② シーディング：k-umeyama 法に従ってランダムにパッチを選んでシードを選択する。シードの選択はシグモイド変換を加えた逐次選択法。

③ クラスタリング：画像の特徴が属性と扱われて汎距離でクラスタリングする。各パッチx_iに最も近いクラスター中心m_gを、汎距離を用いて割り当てる。

$$D^2(x_i, m_g) = (x_i - m_g)' S_g^{-1}(x_i - m_g)$$

画像再構築：クラスターの平均値が動かなくなったところで反復を終了。最終的に、パッチx_iを、それが属するクラスタ中心m_gで置き換える。全てのパッチがそれぞれのクラスタ中心に置き換えられた後、これらの更新されたパッチを元の画像の格子状の配置に従って再配置され、再構築された画像が形成される。この再構築された画像は、元の画像と比較して重要な視覚情報や特徴を維持することになる。

k-umeyama 法はマーケティングでの顧客のセグメンテーションに限らず、図 5.6 のような画像の修復、メディカル分野での患者のグループ化、その他経済分野、工学分野などでも応用できるのではないかと考えられる。

■　k-umeyama 法の意義

k-umeyama 法の特徴として、初期シードの選択にシグモイド関数を採用したことが挙げられる。シグモイド関数は、近接したサンプルがシードに選ばれることを避けると同時に、外れ値をシードにすることが k-means 法より抑制できる。その意味で、既存のクラスター分析よりも優れた方法といえよう。さ

らに、k-umeyama 法ではクラスターの中心とサンプル間の距離を汎距離で測定している。汎距離は、変数の相関関係の情報を加味して距離を評価できるため、クラスタリングに用いる情報がより豊かになる。しかもクラスター単位でのデータ行列が特異になっても汎距離が計算できるように、一般逆行列を採用した。以上の手法を組み合わせることで、k-umeyama 法は従来の非階層クラスター方法よりも高い精度でクラスタリングが実行できるものと期待される。特に次の３点は本提案の長所である。

１）クラスターごとに分散共分散が異なることを許容しているので、消費者の異質性を考慮したマーケティングに適合している。

２）クラスター分析の入力データに因子得点や主成分得点を用いると、累積寄与率分の情報しかクラスタリングに活用されないことになる。一方 k-umeyama 法は、原データの持つ情報を捨てることなくサンプルをクラスタリングする方法である。より多くの情報を使ってセグメンテーションできることは望ましい性質といえよう。

３）シグモイド関数を用いて逐次的にシードを抽出する方法なので、k-means++法とくらべて外れ値をシードに選ぶ可能性を抑制できる。近接したシードを選ぶ可能性はゼロに近いので、クラスター分析の結果の再現性を高めることができる。表 5.2 に k-umeyama 法と既存の非階層クラスター分析の特徴を一覧した。

表 5.2　k-umeyama 法の特徴

発表	名称	初期シードの選択	クラスター中心とサンプルの距離	各クラスターの分散共分散の利用
MacQueen(1967)	k-means	乱数で一括選択／ユーザー指定	ユークリッド距離	利用せず
Dunn(1973)	FuzzyC-Means	乱数で一括選択／ユーザー指定	ユークリッド距離の逆数をメンバーシップ値に	利用せず
Cerioli(2005)	修正k-means	提案なし	汎距離	異なるS g
Arthurら(2007)	k++	距離の2乗に比例した逐次選択	ユークリッド距離	利用せず
梅山(2023)	k-umeyama	Sigmoidによる逐次選択	汎距離	異なるSg

5.2　クラスターの最適性基準

クラスター分析は目的変数が存在しないために正解率のような単純な評価指標がない。そこでクラスター間はセパレートしていてクラスター内は似ている状態を示す何らかの基準を用いて判断するしかない。ここでは5つの基準とそれらの相互関係を検討する。

■ 本節での記法

調査サンプルを$i=1,2,\cdots,N$、測定変数の数をpとする。本節ではクラスターの数をGとし、第gクラスターのサンプル数をn_g $(g=1,2,\cdots,G)$ と書く[5]。また全サンプルを対象にしたp次の平均値ベクトルを$\boldsymbol{m}$、グループgの平均値ベクトルを$\boldsymbol{m}_g$と書く。

一般性を失うことなく分析データは列に関して平均偏差化した行列$\underset{N\times p}{\boldsymbol{X}}=\left(x_{ij}\right)$とする。クラスターgに所属するサンプルiのデータを$\boldsymbol{x}_i^g$ と書く。これはp次のベクトルである。

全分散とクラスター間分散とクラスター内分散の行列を次式で定義する。

$$\underset{p\times p}{\boldsymbol{T}}=\frac{1}{N}\boldsymbol{X}'\boldsymbol{X} \tag{5.5}$$

$$\underset{p\times p}{\boldsymbol{B}}=\frac{1}{N}\sum_g n_g\left(\boldsymbol{m}_g-\boldsymbol{m}\right)\left(\boldsymbol{m}_g-\boldsymbol{m}\right)' \tag{5.6}$$

$$\underset{p\times p}{\boldsymbol{W}}=\frac{1}{N}\sum_g\sum_{i\in g}\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)' \tag{5.7}$$

$\boldsymbol{X}$が平均偏差化されていることから(5.6)の全体平均ベクトルは$\boldsymbol{m}=\boldsymbol{0}$ である。(5.7)の$i\in g$ はクラスターgに属するデータだけで総和をとることを示している。クラスターgにおける分散共分散行列を$V_g=\frac{1}{n_g}\sum_{i\in g}\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)$ として$\boldsymbol{W}$ を求めると $\underset{p\times p}{\boldsymbol{W}}=\sum_g\frac{n_g}{N}V_g=\frac{1}{N}\sum_g\sum_{i\in g}\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)\left(\boldsymbol{x}_i^g-\boldsymbol{m}_g\right)'$ となり(5.7)が求められる。またクラスターサイズn_gで$\boldsymbol{V}_g$を加重合計する式になっている。もともと(5.7)の 2 重の総和記号が N サンプル全体について和をとっているので、サンプル数で加重合計することは合理的である。(6.5)～(5.7)の間には次の関係式が成り立つ。

$$\boldsymbol{T}=\boldsymbol{B}+\boldsymbol{W} \tag{5.8}$$

したがって3つの分散共分散行列のうち 2 つが分かれば、残りは(5.8)から計算できる。(5.5)と(6.6)の計算が簡単なのでふつうは$\boldsymbol{T}-\boldsymbol{B}$で$\boldsymbol{W}$ を求める。なお(5.5)～(5.7)のようにデータ数Nで割る計算は省くことが多い。その理由は、分散分析なら平方和 SS(Sum of Squares)を計算するという実験計画法の伝統があるからである。

また尺度不変性がある基準では N で割っても割らなくても基準値は変わらない。この点を次に確認する。

■ 最適性基準と尺度の不変性

水野(1996)は下記の3つの基準を紹介している。いずれも同一データを同一のクラスター数で分割した結果を相対評価するのに利用できる。異なる調査の間やクラスター数が異なるケース間で比較しても意味がない。

$tr(\boldsymbol{W})$:Edwards ら(1965)が提唱した基準で小さい値の方が望ましい。

$|\boldsymbol{T}|/|\boldsymbol{W}|$:これが Friedman-Rubin(1967)の基準で望大指標である。分散共分散行列の行列式$|\boldsymbol{T}|,|\boldsymbol{W}|$ を一般化分散と呼ぶ。

$tr\left(\boldsymbol{W}^{-1}\boldsymbol{B}\right)$:Lawley-Hotelling(1951)のトレース基準と呼ばれる。これも望大指標である。

本節では以上3つの基準を ED、FR、LH と略記しよう。まず尺度の不変性を確かめる。N=500 のデータで比較したところ ED 基準では平方和行列を用いると分散共分散行列の場合の N 倍になる。残り2つの基準は線形変換に関して不変な基準(invariant criteria)であることが分かった。

表 5.3 不変性の確認

	分散共分散行列	平方和行列
Edwards	1.9718	985.923
Friedman-Rubin	2.3667	2.3667
Lawley-Hotelling	1.3541	1.3541

FR は固有値を用いても定義できる。$\boldsymbol{W}^{-1}\boldsymbol{B}$ の正の固有値を大きい順に $\lambda_k\left(k=1,2,\cdots\lambda_q\right)$ と書けば(5.9)式が導出される。λ_k に 1 を加える根拠が分かりづらいが、それは(5.13)以降で詳述する。

$$FR=\frac{|\boldsymbol{T}|}{|\boldsymbol{W}|}=\left|\boldsymbol{W}^{-1}\right|\left|\boldsymbol{T}\right|=\left|\boldsymbol{W}^{-1}\boldsymbol{T}\right|=\left|\boldsymbol{W}^{-1}\left(\boldsymbol{W}+\boldsymbol{B}\right)\right|=\left|\boldsymbol{I}+\boldsymbol{W}^{-1}\boldsymbol{B}\right|=\prod_k\left(1+\lambda_k\right) \quad (5.9)$$

■ MANOVA(多変量分散分析)において使われる基準

MANOVA の研究は Wilks(1932)に始まった。Wilks の $\Lambda=\dfrac{|\boldsymbol{W}|}{|\boldsymbol{T}|}$ は集団間の平均ベクトルの同等性を検定するための一般化尤度比の中で登場した。

$$\Lambda = \frac{|\boldsymbol{W}|}{|\boldsymbol{B}+\boldsymbol{W}|} = \frac{1}{|\boldsymbol{W}^{-1}||\boldsymbol{B}+\boldsymbol{W}|} = \frac{1}{|\boldsymbol{W}^{-1}\boldsymbol{B}+\boldsymbol{I}|} = \frac{1}{\prod_k (1+\lambda_k)} \qquad (5.10)$$

(5.10)は(5.9)の逆数になる。MANOVA の検定論ではグループ間で分散共分散行列$\boldsymbol{V}_g$が等しいと仮定している。しかしマーケティングでは、この仮定は受け入れ難い。したがって、本研究ではΛをあくまでも記述的な指標としてとりあげる。

Lawley-Hotelling のトレース基準

$$LH = tr\left(\boldsymbol{W}^{-1}\boldsymbol{B}\right) = \sum_k \lambda_k \qquad (5.11)$$

Pillai(1955)のトレース基準

$$Pillai = tr\left(\boldsymbol{B}\boldsymbol{T}^{-1}\right) = \sum_k \frac{\lambda_k}{1+\lambda_k} \qquad (5.12)$$

Morrison(1990、240 頁)の 21 サンプル、2 変数、4 クラスターのデータを使って 5 種類の基準値を表 5.4 で比較した。なおエドワーズの基準は$\boldsymbol{W}$の固有値の和を計算するので他の基準とは異なる[6]。固有値から計算した結果と、トレースまたは行列式から計算した結果はすべて一致することが確認できた。

表 5.4 Morrison のデータにもとづく各種の基準値

	トレースまたは det	固有値から計算
Edwards	112.7	112.7
Friedman-Rubin	1.775152	1.775152
Wilks Λ	0.5633319	0.5633319
Lawley-Hotelling	0.6853434	0.6853434
Pillai	0.4872604	0.4872604

■ 固有値による最適性基準の導出

クラスターの最適性基準を求めるのに固有値がなぜ必要なのだろうか。この素朴な疑問への回答が見あたらない。表 5.4 で取りあげた5つの基準のうちエドワーズの基準は$\boldsymbol{W}$のトレースを求めれば済む。LH と Pillai も行列のトレースから計算できる。

残る FR とΛは、2つの行列式の比をとる基準であるが、行列が特異な場合は行列式が 0 になる。特に FR では比の分母がゼロになるので問題が起きる。したがって LH とその逆数であるΛでは行列の正の固有値を求めて、それらの積から基準値を求めなければならない場合がある。ここまでを整理したのが表 5.5 である。

表 5.5　各種の基準の相互関係

提唱者	定義	$W^{-1}B$の正の固有値による表現	固有値の必要性	尺度不変性
Edwards	$tr(\boldsymbol{W})$	×	×	×
Friedman-Rubin	$\lvert\boldsymbol{T}\rvert/\lvert\boldsymbol{W}\rvert$	$\prod_k(1+\lambda_k)$	○	○
Wilks	$\lvert\boldsymbol{W}\rvert/\lvert\boldsymbol{T}\rvert$	$\left[\prod_k(1+\lambda_k)\right]^{-1}$	○	○
Lawley-Hotelling	$tr(\boldsymbol{W}^{-1}\boldsymbol{B})$	$\sum_k\lambda_k$	×	○
Pillai	$tr(\boldsymbol{B}\boldsymbol{T}^{-1})$	$\sum_k\frac{\lambda_k}{1+\lambda_k}$	×	○

分割による分散は$\boldsymbol{T},\boldsymbol{B},\boldsymbol{W}$　の3つしかなく、そのどれに着目して基準が作られたかを一覧したのが表 5.6 である。一般には$\boldsymbol{T},\boldsymbol{B},\boldsymbol{W}$ の中から2つを選んで基準を作っているがエドワーズは$\boldsymbol{W}$ しか利用していないのがユニークである。

表 5.6　各基準が利用している分散

	T	B	W
T		Pillai	FR,Wilks
B			Lawley-Hotelling
W			Edwards

エドワーズ以外の4つの基準はすべて$\boldsymbol{W}^{-1}\boldsymbol{B}$の固有値をもとにして表現できる。

$$\lambda_k>0\quad(k=1,2,\cdots\lambda_q),\qquad q=rank\left(\boldsymbol{W}^{-1}\boldsymbol{B}\right)\ \le\ \min(p,G-1)$$

次に(5.9)の導出根拠を論じよう。$\boldsymbol{W}^{-1}\boldsymbol{B}$　の正の固有値を大きい順に並べた対角行列を$\boldsymbol{\Lambda}$　とする。$q=2,p=4$　という特異(singular)なケースを例示する。

$$\underset{p\times p}{\boldsymbol{\Lambda}}=\begin{bmatrix}\lambda_1&0&0&0\\0&\lambda_q&0&0\\0&0&0&0\\0&0&0&0\end{bmatrix}\qquad(5.13)$$

奥野ら(1981,312 頁)には次の証明でウィルクスの Λ が λ_k+1の積で表されるとし

た。$\boldsymbol{W}^{-1}\boldsymbol{B}$ の固有値を大きい順に $\lambda_1,\lambda_2,\cdots,\lambda_q>0$ として対応した q 個の固有ベクトルを最初のq列に並べ、残りのブロックにゼロ行列と単位行列を配置した正方行列を構成する。$q=2,p=4$ の場合を書けば、

$$\underset{4\times4}{\boldsymbol{P}}=\begin{bmatrix} & & 0 & 0\\ \boldsymbol{u}_1 & \boldsymbol{u}_2 & 0 & 0\\ & & 1 & 0\\ & & 0 & 1\end{bmatrix} \qquad (5.14)$$

奥野らの解説では固有値と固有ベクトルの関係から $\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{P}=\boldsymbol{P}\Lambda$ であり、$\boldsymbol{P}$ は $\boldsymbol{P}'\boldsymbol{P}=\boldsymbol{I}$ の直交行列だとして上式の左から $\boldsymbol{P}'$ を掛けて $\boldsymbol{P}'\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{P}=\boldsymbol{P}'\boldsymbol{P}\boldsymbol{\Lambda}=\boldsymbol{\Lambda}$ を導いている。しかし $\boldsymbol{W}^{-1}\boldsymbol{B}$ は対称行列ではないから $\boldsymbol{P}'\boldsymbol{P}\neq\boldsymbol{I}$ であり、(5.14)の行列を用いても $\boldsymbol{W}^{-1}\boldsymbol{B}$ を対角化することはできない。

そこで平岡ら(2004,203 頁)の相似変換による対角化を行う。$\boldsymbol{W}^{-1}\boldsymbol{B}$ の p 個の固有ベクトル $\boldsymbol{u}_1,\boldsymbol{u}_2,\cdots,\boldsymbol{u}_p$ を並べて $p\times p$ の $\boldsymbol{P}$ を構成する。この $\boldsymbol{P}$ も直交行列ではない。$\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{P}=\boldsymbol{P}\Lambda$ という関係を用いて

$$\boldsymbol{P}^{-1}\boldsymbol{W}^{-1}\left(\boldsymbol{B}+\boldsymbol{W}\right)\boldsymbol{P}=\boldsymbol{P}^{-1}\boldsymbol{W}^{-1}\boldsymbol{B}\boldsymbol{P}+\boldsymbol{P}^{-1}\boldsymbol{W}^{-1}\boldsymbol{W}\boldsymbol{P}=\boldsymbol{\Lambda}+\boldsymbol{I}$$

ここで両辺の行列式をとると $\boldsymbol{\Lambda}+\boldsymbol{I}$ が対角行列であるから $k=1,2,\cdots,q$ の範囲で積をとれば $\left|\boldsymbol{P}^{-1}\right|\left|\boldsymbol{W}^{-1}\right|\left|\boldsymbol{B}+\boldsymbol{W}\right|\left|\boldsymbol{P}\right|=\dfrac{\left|\boldsymbol{T}\right|}{\left|\boldsymbol{W}\right|}=\prod_{k=1}^{q}\left(\lambda_k+1\right)$

これで FR の基準値(5.9)が導出された。ウィルクスの Λ はその逆数である[7]。

■ Morrison のデータで確認

ここでは表 5.4 の Morrison のデータの変数 A と B を X1,X2 とし、それに X3＝X1−X2, X4=X1+X2 を追加して 4 変数のデータ行列を作った。p＝4 で G＝4, $q=rank(X)=2$ という特異な行列である。$\lambda_1=12.319,\lambda_2=0.053$

表 5.7　各基準値

	トレース	固有値から計算
Edwards	220.8	220.8
Friedman-Rubin	-	14.024
Wilks Λ	-	0.071
Lawley-Hotelling	12.3717	12.3717
Pillai V	0.9749	0.9749

表中の-は、トレースを使わない基準である。固有値計算が必要になるのは FR と Wilks の Λ だけである。その他の基準はトレースでも固有値でも計算できるしどちらでも数値は等しい。

5.3　クラスター分析の今後の発展

クラスター分析では伝統的に、市場を排反(exclusive)かつ悉皆(exhaustive)に分割してきた。排反で悉皆とは、市場のすべてをカバーすると同時にクラスターに重複がないという意味である。図 5.7a)がその典型例である。図中の X は外れ値なのだがグリーンのクラスターに所属させ、中央の Y はどのクラスターに入れればよいか困る境界領域に位置している。

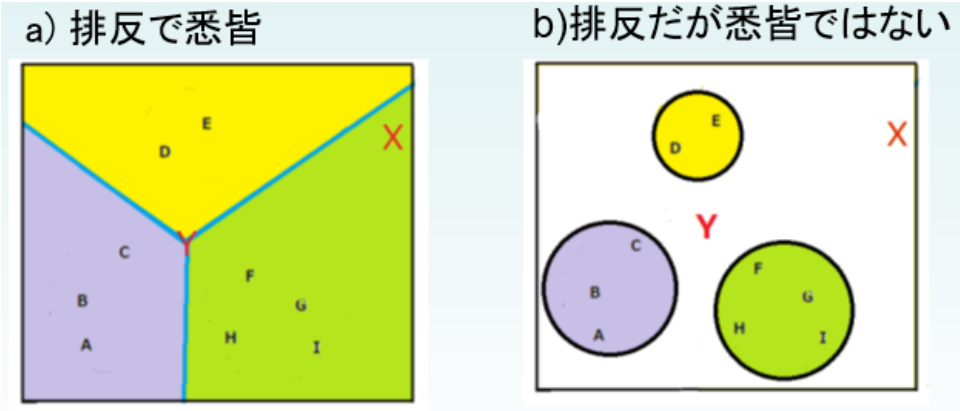


図 5.7　伝統的なクラスタリングは自然なのか

排反かつ悉皆という固定観念にとらわれないクラスタリングも考えられる。図 5.7 の b) は排反だが悉皆ではない例である。X や Y は無理にクラスターに所属させないという対処ができる。

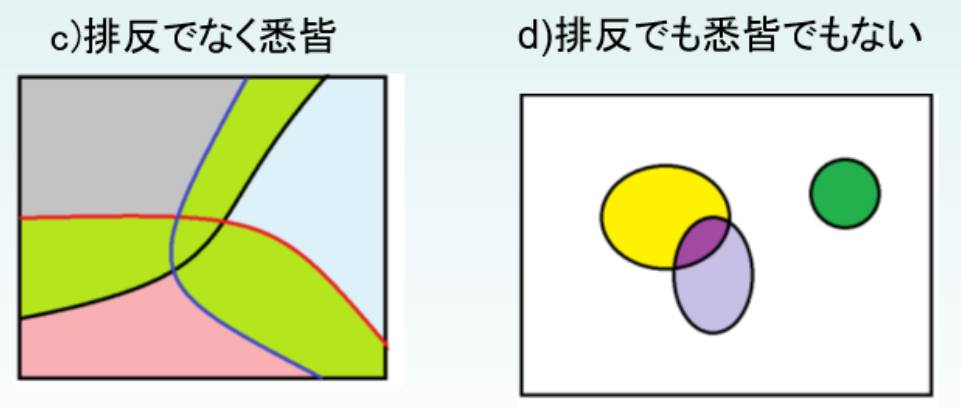


図 5.8　さまざまなクラスタリング

また図 5.8 c）のようクラスターがオーバーラップすることを想定する方が適した市場もあるだろう。この論点に関しては Dunn(1974)によるファジークラスタリングが早い時期の提唱であった。また新規に開拓された市場は潜在顧客の一部しか顕在化していないし、クラスター同士も未分化で重なっているかもしれない。図 5.8 の d）がその図解である。何が望ましいクラスタリングなのかは統計学が決める問題ではなくマーケターが決める問題である。

6 章　討論

6.1　本研究の意義

(1) 2.3 節では正則化回帰のリッジ推定とラッソ推定を比較した。回帰係数の推定と分析変数の選択を同時に行うのがラッソ推定の特徴である。スパースデータの解析には過学習の危険性を伴うので、決定係数が高いからといって良い予測モデルだと判断してはならないことを実験データで示した。
(2) 正則化回帰のハイパーパラメータ λ を決定するために交差検証法が従来よく使われてきた。しかし情報量規準を用いることでより安定した推定が行えることを示した。
(3) 正則化回帰の各手法について、汎化性能を基準とした場合の使い分けの一例を示した。
(4) 判別分析は群が 2 群か多群か、分散共分散が群間で等しいか否か、群の規模の事前確率を利用するか否かで 2×2×2 の 8 通りに分類できる。各ケースの判別法の相違を整理した。
(5) ベイズ判別（LDA、QDA）とロジスティック判別のシミュレーション実験を行った。各群の分布と共分散行列しだいで最良の判別結果が得られる手法が異なることを示した。
(6) SVM のカーネルによる分離の違いを整理した。

(7) CNN と生成モデル（VAE）についてその基礎を確認し応用可能性を指摘した。
(8) クラスター分析の最適性基準を５種類とりあげ、それらの関係を論じた。また Friedman-Rubin の基準については新しい導出の仕方を示した。

６.２　今後の研究課題

■　正則化回帰の解法
2022 年度の研究会では説明変数に多重共線性がある場合の対策として一般逆行列の利用を提案した。しかしスパース回帰分析で一般逆行列は利用されていない。その理由は何だろうか。理論上あるいは計算上の理由を追求するのは興味深い。
■　解の一意性
正則化回帰によって情報行列が正則でなくてもパラメータが推定できる。しかし正則化回帰の解の一意性があるかどうかに関しては疑問がある。
■　正則化回帰における偏回帰係数のバイアス問題への対処
正則化回帰では偏回係数が真の値よりも原点方向へ過小推定されるバイアスがある。これに対してマーケティングミックスモデリング（MMM）などの実用上、どのように対処すべきかという実務上の課題がある。
■　KL divergence（KL 情報量）と尤度（likeliohood）、フィッシャー情報量（Fisher information）との関係の整理
■　クラスター内分散共分散行列に 0 が生じる場合
ある変数のデータがすべて 0 ないし定数の場合はその変数の分散は 0 になり逆行列が求められない。例えば、ＥＣサイトのある商品にだれもアクセスしていない、などでこの事態が発生する。
その対処方法としてリッジ推定で汎距離を計算することが考えられる。しかしその測定法が正しいという論拠はまだ明らかでない。

$$D^2{}_{\text{ridge}}(x_i, m_g) = (\boldsymbol{x_i} - \boldsymbol{m_g})' (S_g + \lambda I)^{-1} (\boldsymbol{x_i} - \boldsymbol{m_g})$$

引用文献

Agresti,A.（2002）*"Categorical Data Analysis,Second Edition."* John Wiley.

Arthur, D. and Vassilvitskii,S. (2007) k-means++: the advantages of careful seeding. *SODA '07: Proceedings of the eighteenth annual ACM-SIAM symposium on discrete algorithms* , 1027-1035.

朝野熙彦(2023) マハラノビスの汎距離と消費者の異質性, 日本行動計量学会第 51 回大会抄録集,70-73.

Cerioli,A. (2005) K-means cluster analysis and Mahalanobis metrics：A problematic match or an overlooked opportunity? *Statistica Applicata*, **17**,(1),61-73.

Chollet, F. 著、巣籠 悠輔他・(株)クイープ 翻訳 (2018)「Python と Keras によるディープラーニング」マイナビ出版

Cortes,C. and Vapnik,V.N. (1995) Support-vector networks, *Machine Learning*, 20,273-297.

Dunn, J.C. (1974) A fuzzy relative of the ISODATA Process and Its use in detecting compact well-separated clusters. *Journal of Cybernetics* **3**, 32-57.

Edwards,A.W.F. and Cavalli-Sforza,L.L.(1965) A method for cluster analysis. *Biometrics*,Vol.21,362-375.

Fisher,R.A. (1936) The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, **7**,179-188.

Friedman,H.P. and Rubin,J. (1967) On some invariant criteria for grouping data. *Journal of American Statistical Association*, Vol.62. 1159-1178.

Fukushima, K. (1980) Neocognitron: A self-organizing neural network model for a mechanism of pattern recognition. *Biological Cybernetics*, 36(4), 193-202.

後藤太郎・梅山貴彦(2023) マーケティングにおける各種の汎距離の応用, 日本行動計量学会第 51 回大会抄録集,74-77.

Hinton, G. and Roweis, S. (2002) Stochastic neighbor embedding. *In Advances in Neural Information Processing Systems*, 15, 833–840, Cambridge, MA, USA. The MIT Press.

Hinton, G. E. and Salakhutdinov, R. R. (2006) Reducing the dimensionality of data with neural networks, *Science*, 313(5786), p504-507.

平岡和幸・堀玄(2004)「プログラミングのための線形代数」オーム社

Hoerl ,A.E. and Kennard, R.W. (1970) Ridge regression：biased estimation for nonorthogonal problems, *Technometrics,* **12, No1,**55-67.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, 28(3/4), 321-377.

Hotelling, H. (1951) A generalized T test and measure of multivariate dispersion, *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability*, University of California Press, Berkeley,23-41.
Kingma, D. P. and Welling, M. (2023) Auto-encoding variational bayes, arXiv:1312.6114 .
LeCun, Y. et al. (1990) Handwritten digit recognition with a back-propagation network. In *Proc. Advances in Neural Information Processing Systems*, 396–404.
LeCun, Y., Bottou, L., Bengio, Y. and Haffner, P. (1998) Gradient-based learning applied to document recognition. *Proc. IEEE 86*, 2278–2324.
LeCun, Y., Bengio, Y., & Hinton, G. (2015) Deep learning. *Nature*, 521(7553), 436-444.
MacQueen,J. (1967) Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on Mathematical Statistics and Probability*, 281-297.
Mahalanobis,P.C. (1936) On the generalized distance in statistics. *Proceedings of the National Institute of Sciences of India*, **2**(1),49-55.
McFadden,D.(1974) Conditional logit analysis of qualitative choice behavior,In Zarembka,P.(ed.) "*Frontiers in Econometrics*." New York:Academic Press,105-142.
McInnes,L., Healy, J., and Melville, J. (2018) UMAP: Uniform manifold approximation and projection for dimension reduction, arXiv:1802.03426.
水野欽司（1996）「多変量データ解析講義」朝倉書店
Morrison,D.F.(1990) "*Multivariate Statistical Methods,*3rd edition", McGraw-Hill.Inc.
奥野忠一・他（1981）「多変量解析法《改訂版》」日科技連
Pillai, K. C. S. (1955). Some new test criteria in multivariate analysis. *Annals of Mathematical Statistics*, Vol.26(1), 117-121.
高根芳雄 (1995)「制約つき主成分分析法」朝倉書店
高野祐一 (2020)　サポートベクトルマシンとカーネル法,オペレーションズ・リサーチ, *Communications of the Operations Research Society of Japan*: 経営の科学, 65(6), 304-309.
竹村彰通（2020）「現代数理統計学」学術図書出版社
竹内啓・柳井晴夫（1972）「多変量解析の基礎」東洋経済新報社
田中豊・脇本和昌（1983）「多変量統計解析法」現代数学社
Tibshirani,R.（1996）Regression shrinkage and selection via the lasso. *J. Royal Statistical Society*, *Ser B* ,**58** (1),267-288.
Truett,J.,Cornfield,J. Kannel,W. (1967) A multivariate analysis of the risk of coronary heart disease in Framingham. *Journal of Chronic*

Disease, **20**, 511-524.
梅津佑太・西村龍映・上田勇祐(2020)「スパース回帰分析とパターン認識」
講談社
Wilks, S.S. (1932). Certain generalizations in the analysis of variance.
Biometrika,24, 471-494.

【研究会の実施概容】

産学協同研究会のメンバーは次の通りである。

研究代表者　朝野熙彦　東京都立大学元教授
研究メンバー
奥瀬喜之　専修大学
大屋伸彦　愛国学園大学
松波成行　物質・材料研究機構
後藤太郎　CCCMK ホールディングス
松本　健　メルカリ
大橋耕也　メルカリ
梅山貴彦　リサーチイノベーション委員会委員長
* 森本　修　DeNA
* 編集委員
研究期間：2023 年 9 月 15 日～2024 年 3 月 21 日（月 1 回）
研究会場：専修大学神田キャンパス

---

[1] 重回帰分析は定数項を明示した$a\mathbf{1}+\boldsymbol{X\beta}+\boldsymbol{\varepsilon}$でモデルを記述することが多い。しかし竹村（2020，241 頁）も指摘しているように、定数項は設けず$\boldsymbol{X}$の任意の 1 列を 1 の値とする記述にしてもモデルは同じである。

[2] 誤差に特定の確率分布を仮定しない定式化もある。

[3] 主成分分析が典型例であるが、多変量解析は次元の縮小を狙ったものが多い。SVM はそれと正反対のモデルである。

[4] 高根（1995)は 6 種類の一般逆行列を示した。その中でもムーア・ペンローズの一般逆行列は最小ノルム、最小 2 乗でありかつ一意に定まる。

[5] 判別分析では判別グループの数を G と記述することが多いので、本節ではグループ数を G で記述した。5.1 節で書いた K と同じ意味である。

[6] 表 5.4 は$rank(\boldsymbol{W})=p$であるから$\boldsymbol{W}$は正則である。

[7] ここでは p 次の正方行列の固有値と行列式についての次の関係式を用いた。
$tr(\boldsymbol{T})=\lambda_1+\lambda_2+\cdots+\lambda_p$，$|\boldsymbol{T}|=\prod_k \lambda_k$，$|\boldsymbol{AB}|=|\boldsymbol{A}||\boldsymbol{B}|=|\boldsymbol{BA}|$，$|\boldsymbol{A}^{-1}|=1/|\boldsymbol{A}|$，$|\boldsymbol{A}|=|\boldsymbol{A}'|$