マーケティング・データ解析
—Excel/Access による—

—付録—

木島正明／中川慶一郎／生田目崇編著

まえがき

本書は木島正明，中川慶一郎，生田目崇 (編著) 「マーケティング・データ解析」(朝倉書店, 2003) で紹介されている分析手法について，紙面の都合上書くことができなかった事項ならびに分析手法の詳細についてまとめたものである．

本書は大きく 3 つの部分に分かれている．付録 A では，Excel の操作のうち，本書と密接な関係がある「ピボットテーブル」と「ソルバー」の利用方法について概説する．付録 B では統計の基礎として，基本統計量と確率分布について簡単に触れ，さらに最尤推定法について説明する．また行列に関連してニュートン法と固有値問題について論ずる．付録 C では，本書で取り上げた各統計手法の数理的側面について詳細に説明する．記述を簡略化するために，行列・ベクトルを用いて説明している部分も多くある．分析手法の詳細について深く知ることは，その手法を自身で自在に応用するためには必要不可欠であるので，ぜひ内容を理解していただきたい．

2003 年 11 月

木島正明
中川慶一郎
生田目崇

目　次

A

Excel/Access の操作

本章では，Excel の機能のうち本書と特に関係の深い「ピボットテーブル」と「ソルバー」について説明する．また，Access によるデータベース構築方法について説明する．さらに，Access のデータベースを元にした Excel によるクエリの実行方法について解説する．本章で説明していない機能や操作方法については専門書を参照いただきたい．

A.1　ピボットテーブル

ピボットテーブルは集計もしくはクロス集計をするために Excel にあらかじめ含まれている機能である．ピボットテーブルを利用することにより，Excel 上でデータ項目間の集計を簡易に行うことができる．

ピボットテーブルを作成するためには,「ツール」メニューの「ピボットテーブルとピボットグラフレポート」を選択する．そして，以下の手順で集計値や集計軸となる項目を指定する．

1）「ピボットテーブル/ピボットグラフ ウィザード - 1/3」(図 A.1) において，入力として分析対象とするデータの場所と出力形式を指定する．入力に関する選択肢は次の 3 つである．

- Excel のリスト/データベース：Excel のワークシート上の範囲を指定
- 外部データソース：Access などの外部ファイルを指定
- 複数のワークシート範囲：Excel のワークシートを複数指定

また，出力形式については，次のいずれかを選択する．

- ピボットテーブル：集計結果を表形式で出力する．
- ピボットグラフ (ピボットテーブル付き)：ピボットテーブルとともにテーブルとリンクしたグラフも出力する．

ここでは，例として,「Excel のリスト/データベース」と「ピボットテーブル」を選択し,「次へ」をクリックする．

2) 「ピボットテーブル/ピボットグラフ ウィザード - 2/3」(図 A.2) では，Excel のワークシートからデータの範囲を指定し,「次へ」をクリックする．

ここで指定するデータは先頭行が各列の項目ラベルとして扱われるため，重複や空欄があってはならない．

図 A.1 ピボットテーブル/ピボットグラフ ウィザード - 1/3

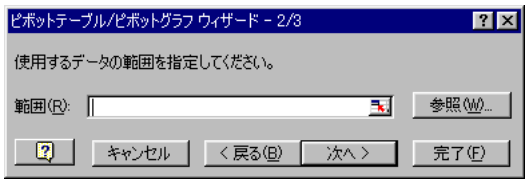

図 A.2 ピボットテーブル/ピボットグラフ ウィザード - 2/3

3）「ピボットテーブル/ピボットグラフ ウィザード - 3/3」(図 A.3) において，表の出力先を指定する．出力先に関する選択肢は次の 2 つである．

- 新規ワークシート：新たにワークシートを作成する場合
- 既存のワークシート：既存のワークシートに出力する場合

4）「レイアウト」をクリックすることで，集計値や集計軸を選択するためのメニューが表示される (図 A.4)．

「ピボットテーブル/ピボットグラフ ウィザード - レイアウト」は以下のフィールドから構成されており，右側にある項目ラベルを各フィールドへドラッグ・アンド・ドロップすることで設定できる．

- 行フィールド：表側の集計軸となる項目
- 列フィールド：表頭の集計軸となる項目
- ページ・フィールド：集計する対象を絞り込む項目
- データ・フィールド：集計値となる項目

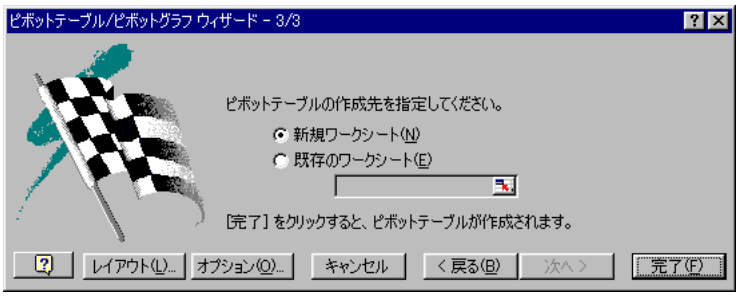

図 A.3　ピボットテーブル/ピボットグラフ ウィザード - 3/3

5）「OK」をクリックし,「ピボットテーブル/ピボットグラフ ウィザード - レイアウト」を閉じた後,「ピボットテーブル/ピボットグラフ ウィザード - 3/3」で「完了」をクリックすることで，指定したセルを起点としてテーブルが作成される．

テーブルとともにピボットテーブル・ツールバーが表示され，これは上側に書式設定などのコマンド・ボタン，下側に項目ラベルの一覧

図 A.4 ピボットテーブル/ピボットグラフ ウィザード - レイアウト

であるフィールド・ボタンで構成されている.

6）図 A.4 のような空欄のピボットテーブルのレイアウト図と，ピボットテーブル・ツールバーが表示される.

たとえば，日別商品別の売上額を表すクロス集計表を作成するためには，行フィールドに「日付」，列フィールドに「商品名」，データ・フィールドに「総額」を指定する．さらに，店舗別に集計する場合にはページ・フィールドに「店舗」を指定する.

一度テーブルを作成した後，各フィールドの項目を入れ替えるには，ピボットテーブル・ツールバーのフィールド・ボタンから選択する[*1)]．なお，集計軸をピボットテーブルから削除したい場合には，各フィールド中の項目ラベルを表の外へドラッグする.

また，集計方法を変更する場合には，データ・フィールド中の項目ラベルをダブル・クリックすることで，図 A.5 が表示され,「データの個数」,「平均」などに変更することができる.

なお，外部データを用いる場合は以下のようにすればよい.

(1) 「ウィザード - 1/3」でデータ元のタイプを指定するときに「外部デー

[*1)] ピボットテーブル・ツールバーが表示されないときは,「表示」メニューの「ツールバー」から「ピボットテーブル」を選択すると表示される.

図 A.5　ピボットテーブル フィールド

タソース」を選択する．

(2) 「ウィザード - 2/3」で「データの取り出し」をクリックし，Access のデータベースを指定する．

(3) 「データの列」を選択する画面が表示されるので，テーブルまたはクエリの全体あるいは特定の列を選択する．

(4) 「ウィザード - 3/3」では通常のピボットテーブルと同様にレイアウトなどを設定する．

A.2　ソルバー

ソルバーは，最適化問題を解くための Excel の Add-In マクロである．ソルバーを使うことにより，複数の制約条件を満たしつつ複数のセルの値を変化させることで，特定のセルの値を最適解として求めることができる．

ソルバーを使用するためには，あらかじめソルバー・アドインを組み込んでおく必要がある．ソルバー・アドインを組み込むためには,「ツール」メニューの「アドイン」を選択し,「アドイン」ダイアログ・ボックスの一覧にソルバー・アドインが表示されていない場合には，Excel のセットアップ・プログラムを実行し，ソルバーを組み込む．ソルバーを組み込んだ後,「ツール」メニューの「ソルバー」を選択すると,「ソルバー：パラメータ設定」(図 A.6) が表示され，各項目について入力する．

- 目的セル：目的式のセルを指定する．この値を最適化する．

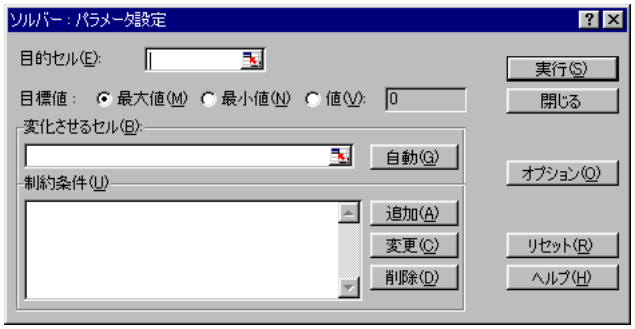

図 A.6 ソルバー：パラメータ設定

- 目標値：「目的セル」の値が最適となる条件を選択する.「最大値」,「最小値」,「特定の値」から選択できる.
- 変化させるセル:「目的セル」の値に影響する変数の値のセルを指定する.
- 制約条件：最適解を求める際の制約条件を指定する．制約の対象となるセルと制約条件の基準値との間の関係 (<=，=，>=，整数，バイナリ) を選択できる.

ソルバーは「制約条件」の下で「変化させるセル」の値を動かし，最適解となる「目的セル」の値を探索する．さらに,「ソルバー：パラメータ設定」の「オプション」をクリックすると「ソルバー：オプション設定」(図 A.7) が表示され，分析方法を詳細に指定できる.

ソルバーにより分析した結果，最適解を求めることができれば,「ソルバー：探索結果」(図 A.8) が表示される.

「ソルバー：探索結果」で「解を記入する」を選択することにより，求められた最適解はワークシート上に反映される．このとき同時に以下のレポートを作成することができる.

- 解答レポート：「目的セル」と「変化させるセル」に関して，初期値及び最適解を出力する．さらに，それぞれの「制約条件」について，条件を満たしているかどうかを表示する.
- 感度レポート：「目的セル」の式および「制約条件」の式の変化に対して，最適解がどの程度敏感に反応しているかを示す.

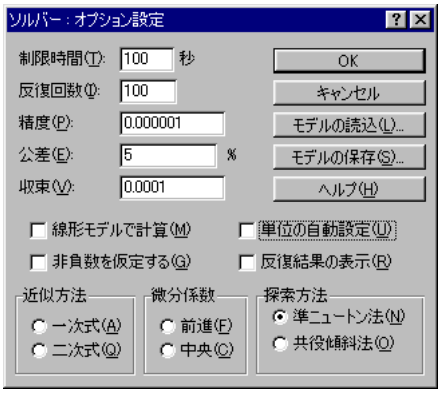

図 A.7　ソルバー：オプション設定

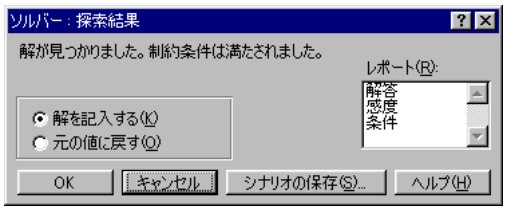

図 A.8　ソルバー：探索結果

- 条件レポート：「変化させるセル」の「制約条件」における上限または下限に対応する「目的セル」の値を表示する．

このソルバー機能を使うことによって，線形計画法やその他の最適化問題を Excel 上で簡易に解くことができる．

A.3　Access によるデータベース構築方法

本節では，Access によるデータベースの構築法を述べる．ここでは，CSV 形式のテキスト・ファイルを元データとして，Access を用いたデータベースの構築を行う手順を以下に示す．

1）Access を起動すると「Microsoft Access」(図 A.9) が表示されるので，「空のデータベース」を選択し，「OK」ボタンをクリックする．

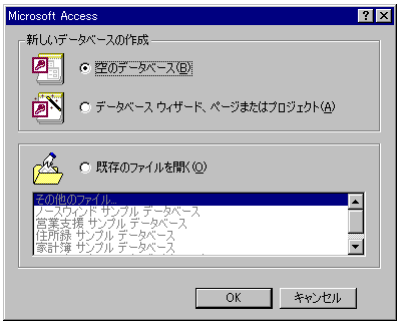

図 A.9 Access の起動

2) データベースの保存場所を聞いてくるので，ファイル名を入力し,「作成」ボタンをクリックする．

3) 「データベースでは，各種オブジェクトの作成・変更などを行う．今回はテーブルを作成するので,「オブジェクト」から「テーブル」を選択し,「新規作成」をクリックする」(図 A.10).

図 A.10 データベース

4) 「テーブルの新規作成」では，テーブルの作成方法を指定する．ここでは，CSV ファイルから作成することを想定し,「テーブルのインポート」を選択し,「OK」ボタンをクリックする (図 A.11).

5) 「インポート」では，ソースデータのファイル名を指定する.「ファイルの種類」で「テキストファイル」をメニューから選び，ファイル名

図 A.11　テーブルの新規作成

を指定し「インポート」ボタンをクリックする．

6）5) に続き,「テキスト インポート ウィザード」が現れるので，テキストファイルの形式を指定する (図 A.12)．今回は CSV 形式であるので,「区切り記号付き」を選択し,「次へ」をクリックする．

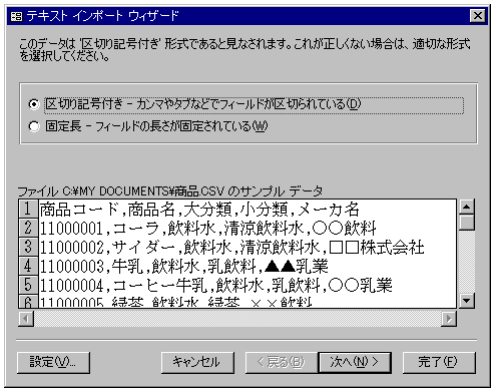

図 A.12　テキスト インポート ウィザード

7）フィールドの区切り記号とフィールド名の有無を指定する (図 A.13)．CSV 形式の区切り記号である「カンマ」を選択する．ファイルの先頭行がデータではなくフィールド名になっている場合には,「先頭行をフィールド名として使う」も選択し,「次へ」をクリックする．

8）データを保存する場所を聞かれるので,「新規テーブルに保存」を選択し,「次へ」をクリックする．

図 A.13 テキスト インポート ウィザード

9）各フィールドのフィールド名 (テーブルの列名) とデータ型を設定する (図 A.14)．なお，インデックスを設定すると，そのカラムによる条件検索が速くなる．

図 A.14 テキスト インポート ウィザード

10）次に，主キーの設定を行うが (図 A.15)，あらかじめ主キーがわかっている場合には「次のフィールドに主キーを設定する」を選択し，メニューからフィールド名を選択する．顧客テーブルでは顧客 ID が，商品テーブルでは商品 ID がこれに該当する．

図 A.15　テキスト インポート ウィザード

11）最後にテーブル名を聞いてくるので，適当な名前を入力し,「完了」ボタンをクリックする．

12）これで,「データベース」(図 A.16) のテーブル・オブジェクトに，作成したテーブルが追加される．なお，テーブルの中身を確認したい場合には，テーブル名をダブルクリックすればよい．

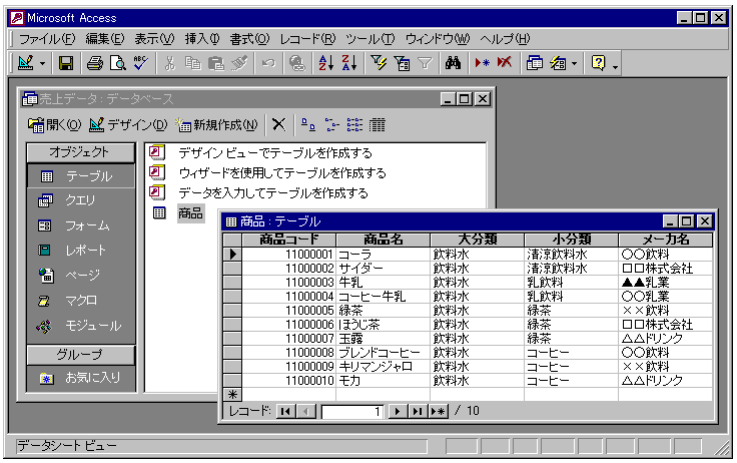

図 A.16　インポート終了

13）他の CSV ファイルも同様にインポートすることで，データベース内に複数のテーブルを作成することができる．

A.4 Excel によるクエリ実行方法

SQL 言語でデータの抽出を行う場合は，SQL 言語の文法を理解していることはもちろん，テーブル名やカラム名を設定しなければならないため，SQL 言語に日ごろから慣れていないとなにかと面倒である．そうでない場合は，GUI により簡単に SQL 文を作成できるツールを使うのが便利である．Excel では，Microsoft Query[*1)] というツールがこれに該当する．Microsoft Query を利用すると，一覧の中からデータを抽出したいテーブル名やカラム名を選択するだけでデータの抽出ができるため，非常に簡単である．

以下に，Microsoft Query を用いて Access データベースからデータを抽出する手順を示す[*2)]．

1）Excel の「データ」メニューの「外部データの取り込み」の「新しいデータベースクエリ」を選択する．すると「データソースの選択」が表示されるので (図 A.17)，ここでデータ抽出先を指定する．本節では Access データベースを利用することとし，[MS Access Database*] を選択する．また，[クエリ ウィザードを使ってクエリを作成/編集する] をチェックし，OK ボタンをクリックする[*3)]．

2）「データベースの選択」が表示されるので，ここで Access データベースのファイルを選択し，[OK] ボタンをクリックする．

3）「クエリ ウィザード – 列の選択」(図 A.18) では，抽出したいデータがあるテーブルとカラムを選択する．[利用可能なテーブルと列] の欄には，テーブルの一覧が表示されるが，テーブル名をダブルクリック

---

*1) Microsoft Query を利用するには，Microsoft Office の中にある「ODBC アドイン」と「Microsoft Query」をインストールしている必要がある．

*2) その他の方法として，Access を使いデータベース中にクエリ・オブジェクトを作成し，Excel から参照する方法もある．この場合，Excel からはクエリが一つのテーブルのように見える．

*3) ウィザードを使って簡単にクエリを作成するときに必要である．ウィザードを使わない時には，チェックを外しておく．

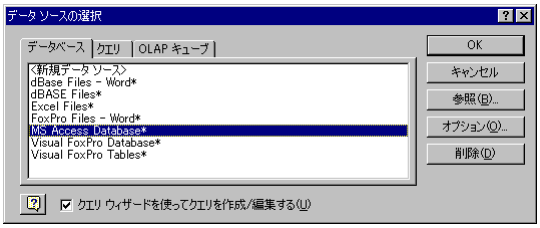

図 A.17　データソースの選択

(または，[+] 記号をクリック) するとテーブルのカラム名一覧が表示されるようになる．ここから必要なカラム名を選択し中央の [>] ボタンをクリックすると，[クエリの列] の欄に取り出されるカラムが設定される[*4]．設定がすべて完了したならば，[次へ] をクリックする．

図 A.18　列の選択

4) 「クエリ ウィザード – データの抽出」(図 A.19) では，取り出すデータの条件設定を行う．条件はカラムごとに設定するが，まず，[抽出する列] からカラムを選択し，次に比較演算子と値をメニューから選択することで行う[*1]．図 A.19 は,「都道府県名＝東京都」の条件を設定した例である．

[*4] テーブルのすべてのカラムを取り出したいときは，テーブル名を選択し，中央の [>] ボタンをクリックする．すると，そのテーブルのカラム全部が [クエリの列] 欄へ設定される．

[*1] 複数の条件を設定したいときは「AND」と「OR」を選択しながら，2 段目以降の個所を埋めていけばよい．

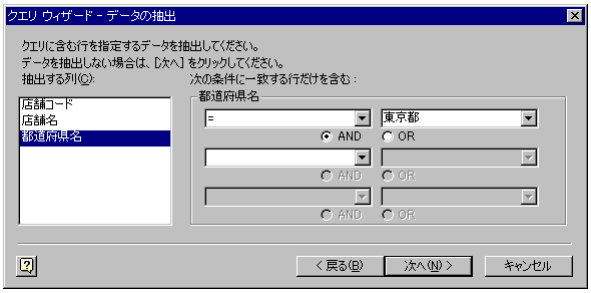

図 A.19 データの抽出

5) 「クエリ ウィザード – 並び替え順序の設定」(図 A.20) では，どのカラムでデータを並べ替えるかを指定する．並べ替える必要がなければ，そのままでよい[*2)]．

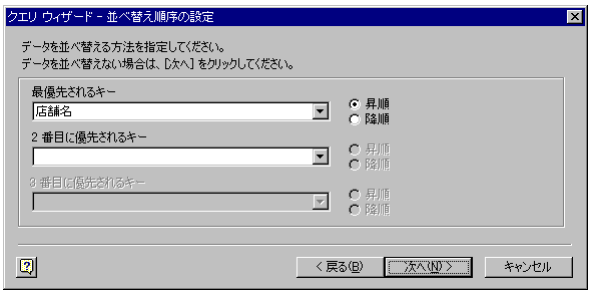

図 A.20 並べ替え順序の設定

6) 「クエリ ウィザード – 完了」では，データの出力先の種類を指定する．今回は Excel ワークシート上なので,「Microsoft Excel にデータを返す」を選択し,「完了」ボタンをクリックする[*1)]．

7) 「Microsoft Excel への外部データの取り出し」(図 A.21) では，取り出したデータの出力先を指定する．特定の場所に出力したいときは,「既存のワークシート」を選択し，出力先の左上の場所にあたるセルを選

[*2)] ここで並べ替えなくても，Excel シート上にデータを抽出すれば Excel 上で自由に並べ替えられる．
[*1)] ここまでの作業をファイルに保存したいときは,「クエリの保存」を実行すればよい．

択し「OK」ボタンをクリックする．新たにワークシートを作成し，そこへ出力したい場合は「新規ワークシート」を選択する．ワークシートにデータを出力するのではなく，ピボットテーブルとして出力したいときは,「ピボットテーブル レポート」を選択する．

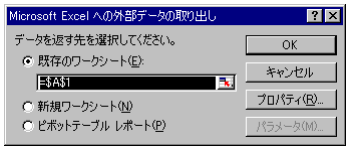

図 A.21　外部データの取り出し

以上の作業を完了すると，データベースに対してクエリが実行される．クエリ終了後，データがワークシート上に出力される (図 A.22).

	A	B	C
1	**店舗コード**	**店舗名**	**都道府県名**
2	10014	渋谷店	東京都
3	10016	上野店	東京都
4	10013	新宿東口店	東京都
5	10012	新宿南口店	東京都
6	10015	池袋店	東京都

図 A.22　クエリ結果

B

統計の基礎

B.1　基本統計量

多変量データを解析する第 1 のステップは，それぞれの変量の傾向やばらつき，変量間の傾向を知るために，データ全体を眺めることである．これは，はずれ値を検出したり，このために，ヒストグラムや散布図，またこれらを組み合わせた多変量連関図などのグラフにより視覚的にデータの様子を把握することが行われる．これとは別に，各変数の平均や分散といった**基本統計量**を求めたり，2 変量間の線形の増加減の関係を見る共分散行列もしくは相関係数行列により，データ全体の様子を数量的に把握することができる．

グラフの作成に関しては他の専門書に譲ることにし，ここでは本書内で必要なデータととその基本統計量について記述する．

次の行列は，各行が回答者などの各データ取得機会を表し，各列がそれぞれ変量を表してた多変量データ行列である．したがって，下記は p 変量で n サンプルの場合を示しており，右辺はそのベクトル表示である．

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1j} & \cdots & x_{1p} \\ x_{11} & x_{12} & \cdots & x_{2j} & \cdots & x_{1p} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{i1} & x_{i2} & \cdots & x_{ij} & \cdots & x_{ip} \\ \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nj} & \cdots & x_{np} \end{bmatrix} = [\boldsymbol{x}_{\cdot 1}, \boldsymbol{x}_{\cdot 2}, \cdots, \boldsymbol{x}_{\cdot j}, \cdots, \boldsymbol{x}_{\cdot p}]$$

このデータから，次のような統計量を計算することができる．ただし，I は

単位行列，つまり対角要素のみが 1 であり，その他の要素が 0 である適当な大きさの正方行列，J はすべての要素が 1 である適当な大きさの正方行列とする．また，$\boldsymbol{e}$ はすべての要素が 1 であるような適当な大きさのベクトルとする．

B.1.1　合計・平均

行列 X の要素の総合計および総平均は，

$$\text{総合計} \ : \ \boldsymbol{e}^\top X\boldsymbol{e}, \qquad \text{総平均} \ : \ \frac{1}{np}\boldsymbol{e}^\top X\boldsymbol{e}$$

で与えられる．しかし多くの場合，それぞれの変量の単位は異なるため，それぞれの変量に関する合計・平均が興味の対象となる．各変量の合計・平均ベクトル (p 次元たてベクトル) はそれぞれ以下の式で求められる．

$$\text{各変量の合計} \ : \ \boldsymbol{e}^\top X, \qquad \text{各変量の平均} \ : \ \frac{1}{n}\boldsymbol{e}^\top X = [\bar{x}_{\cdot 1}, \cdots, \bar{x}_{\cdot p}]^\top = \bar{\boldsymbol{x}}_{\cdot j}$$

B.1.2　分散・共分散

分散は観測された各変量のバラツキの度合を示す尺度であり，偏差平方和を自由度で除して与えられる．また，変量間のバラツキの大きさと傾向を示す指標として，共分散がある．共分散は偏差積和を自由度で除して与えられる．行列 X に対する共分散行列 Σ_X は次のように与えられる．

$$\begin{aligned}\Sigma_X &= \begin{bmatrix} Var(\boldsymbol{x}_{\cdot 1}) & \cdots & Cov(\boldsymbol{x}_{\cdot 1}, \boldsymbol{x}_{\cdot j}) & \cdots & Cov(\boldsymbol{x}_{\cdot 1}, \boldsymbol{x}_{\cdot n}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Cov(\boldsymbol{x}_{\cdot j}, \boldsymbol{x}_{\cdot 1}) & \cdots & Var(\boldsymbol{x}_{\cdot j}) & \cdots & Cov(\boldsymbol{x}_{\cdot 2}, \boldsymbol{x}_{\cdot n}) \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ Cov(\boldsymbol{x}_{\cdot p}, \boldsymbol{x}_{\cdot 1}) & \cdots & Cov(\boldsymbol{x}_{\cdot p}, \boldsymbol{x}_{\cdot j}) & \cdots & Var(\boldsymbol{x}_{\cdot p}) \end{bmatrix} \\ &= \frac{1}{n-1}(X - \boldsymbol{e}\bar{\boldsymbol{x}}_j^\top)^\top (X - \boldsymbol{e}\bar{\boldsymbol{x}}_j^\top) \end{aligned} \tag{B.1}$$

(B.1) 式の，対角項である分散は $Var(\boldsymbol{x}_{\cdot j}) \geq 0$ であり，すべての $i, j (i \neq j)$ について $Cov(\boldsymbol{x}_{\cdot i}, \boldsymbol{x}_{\cdot j}) = Cov(\boldsymbol{x}_{\cdot j}, \boldsymbol{x}_{\cdot i})$ であるので，共分散行列は対称行列である．

分散の単位は元の各変量の単位の 2 乗となっているため，元の単位と揃えるためには分散の平方根である標準偏差 σ_j を用いる．

$$\sigma_j = \sqrt{Var(\boldsymbol{x}_{\cdot j})} \tag{B.2}$$

また，各変量の標準偏差をたてに並べたベクトルを以下のように記述することにする．

$$\boldsymbol{\sigma}_x = \begin{bmatrix} \sigma_1 \\ \vdots \\ \sigma_p \end{bmatrix} \tag{B.3}$$

B.1.3 データの標準化

一般には各変量の単位は異なるため，変量同士を直接比較することは難しい．そこで，変量間の単位を無次元化し，さらに，バラツキの尺度である分散の大きさを統一することで変量間の振舞いを比較することを考える．このような操作を標準化という．この場合，各変量の平均を 0，分散を 1 とする．そのために，各変量の各サンプルについて，その変量の平均を引き標準偏差で除する．第 j 変量は以下のように標準化される．

$$\boldsymbol{z}_{\cdot j} = \frac{\boldsymbol{x}_{\cdot j} - \boldsymbol{e}\bar{x}_{\cdot j}}{\sigma_j}$$

標準化されたデータ行列 (標準得点行列)Z は以下のように求められる[*1)]．

$$Z = [\boldsymbol{z}_{\cdot 1}, \cdots, \boldsymbol{z}_{\cdot j}, \cdots, \boldsymbol{z}_{\cdot p}] = \frac{X - \boldsymbol{e}\bar{\boldsymbol{x}}_j^{\top}}{\boldsymbol{e}\boldsymbol{\sigma}_x^{\top}} \tag{B.4}$$

B.1.4 相関係数

2 変量間の変量の線形関係の方向を見る指標として，相関係数がある．相関係数はその値の取りうる範囲は $[-1, 1]$ であり，1 に近いほど正の相関 (どちらかの変量の値が大きくなるほど，もう一方の変量も線形に大きくなる傾向があるという関係) があり，-1 に近いほど負の相関 (どちらかの変量の値

[*1)] 本付録では，ベクトル同士もしくは行列同士の割り算は，対応する要素同士を割ったものとする．

が大きくなるほど，もう一方の変量も線形に小さくなる傾向がある関係) がある[*2)]．

相関係数行列 P は以下のように求められる．

$$P = \begin{bmatrix} 1 & \cdots & \rho_{1j} & \cdots & \rho_{1p} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \rho_{j1} & \cdots & 1 & \cdots & \rho_{jp} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \rho_{p1} & \cdots & \rho_{pj} & \cdots & 1 \end{bmatrix} = \frac{\Sigma_X}{\boldsymbol{\sigma}_X \boldsymbol{\sigma}_X^\top} \tag{B.5}$$

共分散の性質より，すべての j, k について $\rho_{jk} = \rho_{kj}$ であるので，相関係数行列は対称行列である．別の見方をすると，相関係数は標準得点行列の共分散行列である．したがって変量 j と変量 k の相関係数 ρ_{jk} はそれぞれの標準得点ベクトルを用いて，

$$\rho_{jk} = \frac{\langle \boldsymbol{z}_j, \boldsymbol{z}_k \rangle}{\|\boldsymbol{z}_j\| \ \|\boldsymbol{z}_k\|} \tag{B.6}$$

と書くことができる．このように，相関係数は変量 j と変量 k に関する余弦の値を求めているということができる．

B.2 さまざまな分布

本節では本書に登場する分布の性質について簡単に触れる．確率変数や確率分布の性質などのより詳しい説明については専門書を参照されたい (たとえば岡太ら, 2001)．

a. 正規分布

正規分布は連続分布の中で最も基本的な分布であり，多くの分布が正規分布と関係づけられる．

[*2)] このように，相関係数はサンプル全体に関する線形関係を 1 つの指標で示したものでしかないので，相関係数の絶対値が小さくても必ずしも 2 変量間に関係がないとは言い切れない．たとえば，変数変換をすることで，相関関係を見出すこともできる場合もある．

確率変数 X[*1)]が平均 μ，分散 σ^2 の正規分布に従う場合，X の分布関数は以下のように与えられる．

$$\Pr\{X \leq x\} = F(x) = \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(u-\mu)^2}{2\sigma^2}\right\} du. \tag{B.7}$$

したがって，正規分布の密度関数は以下のように与えられる．

$$f(x) = \frac{dF(x)}{dx} = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}. \tag{B.8}$$

確率変数 X が平均 μ，分散 σ^2 の正規分布に従うとき，$X \sim N(\mu, \sigma^2)$ と表記する．また，確率変数 X は前節で述べた標準化をおこなうことにより，平均 0，分散 1 の標準正規分布に従う確率変数に変換することができる．

$$Z = \frac{X-\mu}{\sigma}. \tag{B.9}$$

任意の平均 μ と分散 σ^2 を持つ正規分布に従う確率変数は，(B.9) 式を X について解き，

$$X = \mu + \sigma Z, \tag{B.10}$$

というように標準正規分布に従う確率変数から得ることができる．

変数 $X_1, X_2, \cdots, X_m$ が独立で同一の正規分布 $N(\mu, \sigma^2)$ に従うとき，X_i の線形結合，

$$Y = a_1 X_1 + a_2 X_2 + \cdots + a_m X_m, \tag{B.11}$$

は，平均が $\sum_{i=1}^{m} a_i X_i$，分散が $\sum_{i=1}^{m} a_i^2 \sigma_i^2$ の正規分布に従う．$a_i (i = 1, 2, \cdots, m)$ は実定数である．したがって，$Y \sim N(\sum_{i=1}^{m} a_i x_i, \sum_{i=1}^{m} a_i^2 \sigma_i^2)$ となる．

b. カイ 2 乗分布

確率変数 $X_i, X_2, \cdots, X_n$ がそれぞれ独立の標準正規分布に従うとき，

$$Z = \sum_{i=1}^{n} X_i^2 \tag{B.12}$$

は自由度 n のカイ 2 乗分布に従う．

[*1)] 本節では X は確率変数を表す．

正規分布の性質から，変数 $X_1, X_2, \cdots, X_n$ がそれぞれ独立で同一の正規分布 $N(\mu, \sigma^2)$ に従うとき，

$$\frac{1}{\sigma^2}\sum_{i=1}^{n}(X_i - \bar{X})^2, \tag{B.13}$$

は自由度 $n-1$ のカイ 2 乗分布に従う．ただし $\bar{X}$ は X_i の平均 $\bar{X} = \sum_{i=1}^{n} X_i/n$ である．これを平方和の性質という．

c. F 分布

確率変数 X_1 と X_2 が互いに独立でそれぞれが自由度 n_1, n_2 のカイ 2 乗分布に従うとき，それぞれの確率変数を互いの自由度で割った比，

$$\frac{X_1/n_1}{X_2/n_2}, \tag{B.14}$$

は自由度 n_1 と n_2 の F 分布に従う．

F 分布については，以下の定理が成り立つ．

定理 B.1. 共通の分散 σ^2 をもつ 2 つの母集団 $N(\mu_X, \sigma^2)$, $N(\mu_Y, \sigma^2)$ のそれぞれから，n_1, n_2 個のサンプルを抽出する．サンプルの分散を $Var(X)$, $Var(Y)$ とすると，その分散比 $Var(X)/Var(Y)$ は自由度 $(n_1 - 1, n_2 - 1)$ の F 分布に従う．

d. t 分布

標準正規分布に従う確率変数 X と自由度 n のカイ 2 乗分布に従う Y が独立ならば，$Z = X/\sqrt{Y/n}$ は自由度 n の t 分布に従う．

t 分布については検定で用いる以下の重要な定理が知られている．

定理 B.2. 互いに独立で同一の正規分布 $N(\mu, \sigma^2)$ に従う確率変数 X_1, X_2, $\cdots$,X_n について，統計量

$$t = \frac{\bar{X} - \mu}{Var(X)/\sqrt{n}}, \tag{B.15}$$

は自由度 $n-1$ の t 分布に従う．

この定理より，t 分布は分散が未知の正規分布に関する検定に用いられる．

e. 二重指数分布

分布関数が

$$F(x) = \exp\left\{-\mathrm{e}^{-bx}\right\}, \quad x \in I\!R, \tag{B.16}$$

で与えられる分布を二重指数分布 (もしくは第 1 種極値分布) とよぶ (b は分散に関するパラメータであり，この分布の平均は 0，分散は $\pi^2/(6b^2)$ である)．確率密度関数は,

$$f(x) = b\mathrm{e}^{-bx} \exp\left\{\mathrm{e}^{-bx}\right\}, \tag{B.17}$$

となる．この分布は単峰であるが左右対称ではない．しかし，数学的な取り扱いやすさから尤度計算などで広く用いられており，ロジット・モデルなどで利用されている．

f. 指数分布

分布関数が,

$$F(x) = 1 - \exp\{-\lambda x\}, \tag{B.18}$$

で与えられる分布を指数分布という．指数分布の密度関数は,

$$f(x) = \frac{dF(x)}{dx} = \lambda \exp\{-\lambda x\}, \tag{B.19}$$

で与えられる．市場における普及が指数分布に従うとき，条件付購買発生率つまりハザード率は,

$$h(x) = \frac{f(x)}{1 - F(x)} = \lambda, \tag{B.20}$$

で与えられ，一定である．これは，ある商品の市場普及率が指数分布に従うとし，顧客は一度だけ購買行動を起こすと仮定すると，顧客の購買は時点に依存することなく，常に同じ割合で発生することを表している．この性質を**無記憶性**という．指数分布の平均と分散はそれぞれ $1/\lambda$, $1/\lambda^2$ で与えられる．指数分布の関数形はロジスティック関数と並び，市場普及過程を表す関数として頻繁に利用されている．

B.3　最尤推定法

本節では，尤度について触れ，最尤推定法について述べる．

$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)$ をある母集団から抽出された n 個の観測データとする．個々の観測データの確率分布について，離散の場合には確率分布関数を $P(y_1|\boldsymbol{\theta}), P(y_2|\boldsymbol{\theta}), \cdots, P(y_n|\boldsymbol{\theta})$，連続の場合には密度関数を $f(y_1|\boldsymbol{\theta}), f(y_2|\boldsymbol{\theta}), \cdots, f(y_n|\boldsymbol{\theta})$ とする．ただし，$\boldsymbol{\theta} = (\theta_1, \theta_2, \cdots, \theta_q)$ は確率分布の構造を支配する q 個の未知のパラメータ (母数) である．ここで，注意すべきことは $P(y|\boldsymbol{\theta})$ あるいは $f(y|\boldsymbol{\theta})$ は，関数形が既知の確率分布であり，$\boldsymbol{\theta}$が与えられれば一意に定まるということである．

確率分布の関数形を所与として実際に抽出された観測データ系列から，パラメータを推定することを考える．このとき，以下の関数を定義する．

(離散の場合)

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} P(y_i|\boldsymbol{\theta}) = P(y_1|\boldsymbol{\theta})P(y_2|\boldsymbol{\theta})\cdots P(y_n|\boldsymbol{\theta})$$

(連続の場合)

$$L(\boldsymbol{\theta}|\boldsymbol{y}) = \prod_{i=1}^{n} f(y_i|\boldsymbol{\theta}) = f(y_1|\boldsymbol{\theta})f(y_2|\boldsymbol{\theta})\cdots f(y_n|\boldsymbol{\theta})$$

$L(\boldsymbol{\theta}|\boldsymbol{y})$ は尤度関数と呼ばれ，これを最大にするパラメータ $\hat{\boldsymbol{\theta}}^* = \hat{\boldsymbol{\theta}}(\boldsymbol{y})$ を最尤推定値という．最尤推定値は観測データ y の関数である．このようにしてパラメータを推定する方法を最尤推定法と呼ぶ．

簡単な例として，第 3.3 節の重回帰分析のデータを取り上げる．少し見方を変えて既知の変数である $\boldsymbol{x}$ と $N(0, \sigma^2)$ に従う確率変数である誤差 ε によって，確率変数 Y が以下のように説明されるとする．

$$Y = a_0 + \boldsymbol{a}^\top \boldsymbol{x} + \varepsilon$$

ここで，y_i を $\boldsymbol{x} = \boldsymbol{x}_i$ のときに観測されたデータとすると，その密度関

数は,

$$f(y_i|a_0,\boldsymbol{a}) = \left(\frac{1}{2\pi\sigma^2}\right)^{1/2} \exp\left[-\frac{1}{2\sigma^2}\left\{y_i - (a_0 + \boldsymbol{a}^\top \boldsymbol{x}_i)\right\}^2\right]$$

したがって，尤度関数 $L(a_0,\boldsymbol{a}|\boldsymbol{y})$ は以下のように表される．

$$L(a_0,\boldsymbol{a}|\boldsymbol{y}) = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{y_i - (a_0 + \boldsymbol{a}^\top \boldsymbol{x}_i)\right\}^2\right] \quad \text{(B.21)}$$

$L(a_0,\boldsymbol{a}|\boldsymbol{y})$ を最大にすることはすなわち，$\sum_{i=1}^{n}\left\{y_i - (a_0 + \boldsymbol{a}^\top \boldsymbol{x}_i)\right\}^2$ を最小にすることと等しく，重回帰分析において誤差の 2 乗和を最小にすることは誤差に正規分布を仮定した場合の尤度を最大にすることに対応していることがわかる．

一般に，尤度関数は単峰であるという保証はない．したがって，数値計算によってパラメータを推定する場合には，事前に単峰性が保証されている場合に限るということに注意して頂きたい．

Excel を用いてパラメータを推定する場合は，ソルバーを用いることになる．離散，連続いずれの場合でも尤度関数が単峰であることが保証され，個々の観測データの確率分布あるいは確率密度が閉じた関数として与えられているのならば，ソルバーによってパラメータを求められる．

B. 4 多変数関数と行列

本節では，行列表現による多変数関数に関する基本的な事項に触れ，最適化手法の基本であるニュートン法について述べる．

B.4.1　ヘッセ行列

n 次元変数ベクトル $\boldsymbol{x} = (x_1, \cdots, x_j, \cdots, x_n)^\top$ に関する多変数関数 $f(\boldsymbol{x})$ に関して，

$$\nabla f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f}{\partial x_1} \\ \vdots \\ \frac{\partial f}{\partial x_j} \\ \vdots \\ \frac{\partial f}{\partial x_n} \end{pmatrix}, \quad \nabla^2 f(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_j \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_j^2} & \cdots & \frac{\partial^2 f}{\partial x_j \partial x_n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n \partial x_j} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix},$$

をそれぞれ，関数 f の勾配ベクトルおよびヘッセ行列とよぶ．関数 f が 2 回微分可能かつ 2 階の偏導関数がすべて連続ならば，ヘッセ行列は対称行列となる．これらを用いて，関数 f を座標 $\boldsymbol{a}$ の周りで Taylor 展開すると，以下のようになる．

$$f(\boldsymbol{x}) = f(\boldsymbol{a}) + \nabla f(\boldsymbol{a})^\top (\boldsymbol{x} - \boldsymbol{a}) + \frac{1}{2} (\boldsymbol{x} - \boldsymbol{a})^\top \nabla^2 f(\boldsymbol{a}) (\boldsymbol{x} - \boldsymbol{a}) + O(\boldsymbol{x}^2) \quad \text{(B.22)}$$

B.4.2　ニュートン法

最適化問題を解くためには，ある点から出発し評価関数を改善するように反復的に解を探索するのが一般的である．そのもっとも代表的な方法がニュートン法である．また他のほとんどの方法も，基本的にはニュートン法の考え方を元にしているといっても過言ではないであろう．反復の際に解を改善する方向を決定するために (B.22) 式による展開を用いる．今，2 階微分可能多変数関数 $f(\boldsymbol{x})$ について，i 回目の反復によって得られた解を $\boldsymbol{x}_i$ とする．$\boldsymbol{x}_i$ の周りで 2 次の項まで Taylor 展開すると (B.22) 式より次の式を得る[*1)]．

$$f(\boldsymbol{x}_i + \boldsymbol{d}) \approx f(\boldsymbol{x}_i) + \nabla f(\boldsymbol{x}_i)^\top \boldsymbol{d} + \frac{1}{2} \boldsymbol{d}^\top \nabla^2 f(\boldsymbol{x}_i) \boldsymbol{d} \quad \text{(B.23)}$$

(B.23) 式を最小にするには，(B.23) 式を変数ベクトル $\boldsymbol{d}$ で微分した，

$$\nabla f(\boldsymbol{x}_i + \boldsymbol{d}) = \nabla f(\boldsymbol{x}_i) + \nabla^2 f(\boldsymbol{x}_i) \boldsymbol{d}$$

[*1)] Taylor 展開は差分の近似である．したがって (B.22) 式では $\boldsymbol{x}$ が $\boldsymbol{a}$ の近傍であると考え，(B.23) 式では $\boldsymbol{d}$ がゼロ・ベクトルの近傍であると考えればこれら 2 つの式の対応がつくであろう．

の各要素が 0 となればよい．したがって，

$$\nabla^2 f(\boldsymbol{x}_i)\boldsymbol{d} = -\nabla f(\boldsymbol{x}_i)$$

を変数ベクトル $\boldsymbol{d}$ について解けばよい．そして，適当な方法で $\boldsymbol{d}$ の幅を決めることにより更新された解 $\boldsymbol{x}_{i+1}$ を求めることができる．これを繰り返しおこなうことで，解を次々と改善していく．

ニュートン法は局所的には 2 次収束するので，非常に速い方法として知られている．しかしニュートン法の場合，大域的な最適解を求めるためにはヘッセ行列が正定値行列[*1)]である必要がある．しかし，関数によってはヘッセ行列が正定値行列であるとは限らないので，ニュートン法のアルゴリズムにより得られる探索方向が関数の改善方向になるという保証はない．そこで，ヘッセ行列を適当な正定値行列に近似することを考える．この方法は準ニュートン法として知られている．この近似に関する更新ルールにはさまざまな方法[*2)]があるが，本書の範囲を逸脱するので興味がある読者はたとえば八巻・矢部 (1999) を参照いただきたい．

B.5 固有値問題

多変量解析手法は最小 2 乗法に帰着できるもの，もしくは固有値問題に帰着できるものの 2 つに大別できると言っても過言ではないだろう．したがって，固有値問題は多変量解析諸手法の重要なエンジンとなるものであり，主成分分析をはじめ多くの手法が固有値問題に帰着される．本節では固有値問題について述べる．

固有値問題とは以下のようなものである．n 次の正方行列 A に対して，

$$A\boldsymbol{x} = \lambda\boldsymbol{x}, \quad \boldsymbol{x} \neq \boldsymbol{0}$$

を満たすようなベクトル $\boldsymbol{x}$ が存在するとき，この λ を行列 A の固有値，$\boldsymbol{x}$

[*1)] n 次の正方行列 A について n 次の任意の実数ベクトル $\boldsymbol{x}$ に対して，$\boldsymbol{x}^\top A\boldsymbol{x} > 0$ が成り立つならば行列 A は正定値行列であるという．

[*2)] BFGS 公式などがある．

を λ に対する固有ベクトルという．これは，連立方程式，

$$\lambda I\boldsymbol{x} - A\boldsymbol{x} = (\lambda I - A)\boldsymbol{x} = \boldsymbol{0}$$

が $\boldsymbol{x} \neq \boldsymbol{0}$ となる解を持つことになるので，行列 A の固有値 λ は，方程式

$$|\lambda E - A| = 0$$

を満足する．なお，x についての n 次多項式 $\varphi_A(x) = |xI - A|$ を A の固有多項式，$\varphi_A(x) = |xI - A| = 0$ を A の固有方程式という．

固有値に関する詳細は本書の範囲を越えるので他書に譲るが，多変量解析の理論で必要となる特徴を以下にまとめておく．

1）n 次正方行列 A の固有値の数は複素数の範囲で考えると，重複も含めて n 個となる．
2）正方行列 A が実行列であっても固有値は実数とは限らない．
3）任意の固有ベクトルは定数倍しても固有ベクトルである．
4）n 次正方行列 A が正則ならば，固有値はゼロではない．

また，特に正方行列 A が実対称行列のとき，以下のことが知られている．

1）A の固有値はすべて実数である．
2）A の相違なる固有値に対応する固有ベクトルの内積はゼロ，すなわち直交する．
3）適当な直交行列 L により $L^\top AL$ を対角行列にすることができる．

一般に実対称行列 A の固有値問題を解く方法としては，べき乗法，Jacobi 法，QR 法といったものがあるが，詳細についてはシャトラン (2003) を参照されたい．

C

分析手法の詳細

C. 1　　分散分析

多変量データを統計的に分析する場合，設定したモデルが統計的な視点からみて意味のあるものであるかどうかを考えなければならない．こういった決定は通常「検定」を通じておこなわれる．さまざまな分析手法の詳細を見る前に，もっとも基本的な検定の 1 つである分散分析について述べる．

a. 分散分析とは

分散分析は，多群の標本を比較することで，それらが同じ平均値を持つ母集団から抽出されたものであるかどうかという仮説に関する検定をおこなう分析手法である．

分散分析はその名前が示す通り，各標本群の「分散」を通して群を規定する因子が各群の反応に影響を与えているかどうかを統計的に検証することを目的とする．以下では，もっとも単純な一元配置分散分析について説明する．

b. 一元配置分散分析のモデル式

表 C.1 のように取り上げた因子に関して a 個の水準を考え，各水準を 1 つの群としてそれぞれ n 個のサンプルが得られている場合を考える[*1)]．第 i 水準の第 j 番目のサンプルを y_{ij} と表す．

一元配置分散分析では，1 つの因子の水準が反応であるサンプルに影響を及ぼすかどうかを検証する．そのために，サンプルに対して次のようなモデル式を考える．

$$y_{ij} = \mu + \eta_i + \varepsilon_{ij} \tag{C.1}$$

[*1)] 各水準でサンプルの数が異なる場合も以下の手順を踏めば分析可能である．

表 C.1　分散分析のデータ例

群	サンプル					
	1	2	$\cdots$	j	$\cdots$	n
1	y_{11}	y_{12}	$\cdots$	y_{1j}	$\cdots$	y_{1n}
2	y_{21}	y_{22}	$\cdots$	y_{2j}	$\cdots$	y_{2n}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
i	y_{i1}	y_{i2}	$\cdots$	y_{ij}	$\cdots$	y_{in}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
a	y_{a1}	y_{a2}	$\cdots$	y_{aj}	$\cdots$	y_{an}

このモデル式では，第 i 水準の j 番目のサンプル y_{ij} はすべての水準における共通のパラメータ μ に各水準の効果を示すパラメータ η_i と観測誤差を示す ε_{ij} を加えられている．パラメータ μ, η_i は定数であり，誤差 ε_{ij} は互いに独立で平均 0，分散 σ^2 の正規分布に従うとする．

しかし，このモデル式のパラメータ μ, η_i の真の値は分からない．したがって，これらを観測データ y_{ij} から推定することを考える．η_i の平均を 0，つまり $\sum_i \eta_i = 0$ とすれば，μ は測定値全体の母平均であるので，サンプルの総平均 $\bar{\bar{y}} = \dfrac{\sum_i \sum_j y_{ij}}{na}$ をその推定値として採用する．また，η_i は総平均 μ と各水準の母平均の差であるので各水準の平均値から総平均を引いた $\bar{y}_i = \dfrac{\sum_j y_{ij}}{n}$ を推定値とする[*1)]．

今，水準間の反応に差があるかどうかを確かめたい．水準間に差がない，つまり各水準の反応の平均が等しいならば $\eta_1 = \cdots = \eta_a = 0$ となるはずである．ここで，以下のような帰無仮説 H_0 と対立仮説 H_1 を設定する．

$$H_0 : \quad \eta_1 = \cdots = \eta_a = 0. \tag{C.2}$$

$$H_1 : \quad H_0\text{でない}. \tag{C.3}$$

c. 一元配置分散分析のパラメータの評価

上記のように，η_i を評価することが最終的な目的であるが，実際にはそれらの値を直接比較することはできない．そこで，サンプルと μ の差の平方和を，η_i を導入して分解する．μ, η_i の推定値 $\bar{\bar{y}}, \bar{y}_i$ を用いると，サンプルか

[*1)] パラメータのこれらの推定値は誤差の 2 乗和を最小にするようなラグランジュ未定乗数法を解くことによっても同様の結果が得られる．

ら総平均を引いた差の 2 乗和は次式のように分解できる．

$$\underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{\bar{y}})^2}_{\text{総平方和}} = \underbrace{n\sum_{i=1}^{a}(\bar{y}_i-\bar{\bar{y}})^2}_{\text{モデルの平方和}} + \underbrace{\sum_{i=1}^{a}\sum_{j=1}^{n}(y_{ij}-\bar{y}_i)^2}_{\text{誤差の平方和}} \qquad \text{(C.4)}$$

(C.4) 式の左辺を総平方和とよび，右辺の第 1, 2 項をそれぞれモデルの平方和，誤差の平方和とよぶ．

分散分析では，モデルの平方和と誤差の平方和をそれぞれ，水準間のばらつき，水準内のばらつきとして比較する．(C.4) 式の右辺のモデルの平方和は水準間のばらつきを，誤差の平方和は水準内のばらつきを表している．これら 2 つのばらつきの大きさを比較したいが，直接比較することはできない．

そこで，(C.4) 式を σ^2 で割り，平均平方として表現する．すると，カイ 2 乗分布の平方和の性質より，$n\sum_i(\bar{y}_i-\bar{\bar{y}})^2/\sigma^2$, $\sum_i\sum_j(y_{ij}-\bar{y}_i)^2/\sigma^2$ はそれぞれ自由度 $a-1$, $na-a$ のカイ 2 乗分布に従う．2 つのカイ 2 乗分布の比は F 分布であるので，これら 2 つの平均平方の比を考える．これを F 比 (もしくは分散比) という．もしも水準間に違いがなければ平均平方の比は F 分布に従う．したがって，F 検定により有意差があると結論付けられれば，水準間に違いがあるとはいえない[*1)]という帰無仮説 H_0 が棄却され，水準間に違いがあると結論づけられる．F 比は次の式で与えられる．

$$F = \frac{n\sum_i(\bar{y}_i-\bar{\bar{y}})^2/(a-1)}{\sum_i\sum_j(\bar{y}_{ij}-\bar{\bar{y}})^2/(na-a)} \qquad \text{(C.5)}$$

分散分析では以上の流れを，分散分析表にまとめて表すことが多い．分散分析表は特に書式が決まっているわけではないが，多くのものは表 C.2 に示されるようなものである．

ただし表中の F は，モデルと誤差の平均平方の比 (C.5) 式である．

たとえば，有意水準 5%で検定をしたい場合は表 C.2 の F の値と $F(a-1, na-a, 0.05)$ の値を比較して，F の方が大きければ帰無仮説は棄却される．

また因子を 2 つ考えた場合の分散分析が二元配置分散分析である．この場合は，2 つの因子それぞれのの効果とともに，2 つの因子に関する同時効果

[*1)] 「違いがない」というように断定的な記述をしないのが一般的である．

表 C.2　一元配置分散分析の分散分析表

	平方和	自由度	平均平方	F 比
モデル	$\sum_i n(\bar{y}_i - \bar{\bar{y}})^2$	$a-1$	$\sum_i n(\bar{y}_i - \bar{\bar{y}})^2/(a-1)$	F
誤差	$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2$	$na-a$	$\sum_i \sum_j (y_{ij} - \bar{y}_i)^2/(na-a)$	
計	$\sum_i \sum_j (y_{ij} - \bar{\bar{y}})^2$	$na-1$		

(これを交互作用という) を考慮する必要がある．詳しくは専門書 (たとえば 河口, 1978) を参照いただきたい．

C. 2　重回帰分析

C. 2. 1　パラメータの推定

パラメータの推定値を求めるためには，以下のように理論値 $\hat{y}$ と実測値 y との誤差の 2 乗和 Q が最小になるようにする．

$$Q \equiv \sum_{i=1}^{n} \{y_i - (\hat{\alpha}_0 + \hat{\alpha}_1 x_{1i} + \hat{\alpha}_2 x_{2i} + \cdots + \hat{\alpha}_p x_{pi})\}^2 \to \min$$

ここで，誤差の 2 乗和 Q を最小にする理由ついては付録 B. 3 の最尤推定法, Q を最小にする $\hat{\alpha}_0$, $\hat{\alpha}_1$, $\hat{\alpha}_2$, $\cdots$, $\hat{\alpha}_p$ の求め方については付録 C. 2. 1 を参照されたい．

そこで，図 3.9 のようなデータを一般的な形式で記述すると表 C.3 のようになる．

サンプル i $(i = 1, 2, \cdots, n)$ について説明変数を x_{ij}，目的変数 y_i としたとき，重回帰分析のモデル式は以下のようになる．

$$y_i = \alpha_0 + \sum_{j=1}^{m} \alpha_j x_{ij} + \varepsilon_i, \quad i = 1, 2, \cdots, n$$

このモデル式に従って表 C.3 のデータを記述すると以下のようになる．

$$\boldsymbol{y} = \left[\begin{array}{cc} \boldsymbol{e} & X \end{array} \right] \left[\begin{array}{c} \alpha_0 \\ \boldsymbol{\alpha} \end{array} \right] + \boldsymbol{\varepsilon} = \boldsymbol{e}\alpha_0 + X\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \qquad \text{(C.6)}$$

表 C.3 重回帰分析のデータ

目的変数	説明変数					
	1	2	$\cdots$	j	$\cdots$	m
y_1	x_{11}	x_{12}	$\cdots$	x_{1j}	$\cdots$	x_{1m}
y_2	x_{21}	x_{22}	$\cdots$	x_{2j}	$\cdots$	x_{2m}
y_3	x_{31}	x_{32}	$\cdots$	x_{3j}	$\cdots$	x_{3m}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_i	x_{i1}	x_{i2}	$\cdots$	x_{ij}	$\cdots$	x_{im}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
y_n	x_{n1}	x_{n2}	$\cdots$	x_{nj}	$\cdots$	x_{nm}
$\boldsymbol{y}$	$\boldsymbol{x}_1$	$\boldsymbol{x}_2$	$\cdots$	$\boldsymbol{x}_j$	$\cdots$	$\boldsymbol{x}_m$
パラメータ	α_1	α_2	$\cdots$	α_j	$\cdots$	α_m

ただし，$\boldsymbol{y}=(y_1,y_2,\cdots,y_n)^\top$，$X=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_m)$，$\boldsymbol{x}_j=(x_{1j},x_{2j},\cdots,x_{nj})^\top$，$\boldsymbol{e}=(1,1,\cdots,1)^\top$[*1)]，$\boldsymbol{\alpha}=(\alpha_1,\alpha_2,\cdots,\alpha_m)^\top$，$\boldsymbol{\varepsilon}=(\varepsilon_1,\varepsilon_2,\cdots,\varepsilon_n)^\top$ である．

ここで，以下のように誤差の二乗和が最小になるパラメータ α_j を定める．なお，誤差の二乗和最小とは誤差のベクトル $\boldsymbol{\varepsilon}$ のノルム (大きさ) 最小を意味することに注意されたい．

$$
\begin{aligned}
Q &= \sum_{i=1}^{n} \varepsilon_i^2 \to \min \\
&= \|\boldsymbol{\varepsilon}\|^2 = \langle \boldsymbol{\varepsilon}, \boldsymbol{\varepsilon} \rangle \\
&= \|\boldsymbol{y} - \boldsymbol{e}\alpha_0 - X\boldsymbol{\alpha}\|^2 = \langle \boldsymbol{y} - \boldsymbol{e}\alpha_0 - X\boldsymbol{\alpha}, \boldsymbol{y} - \boldsymbol{e}\alpha_0 - X\boldsymbol{\alpha} \rangle \\
&= \langle \boldsymbol{y}, \boldsymbol{y} \rangle + \alpha_0^2 \langle \boldsymbol{e}, \boldsymbol{e} \rangle + \langle X\boldsymbol{\alpha}, X\boldsymbol{\alpha} \rangle \\
&\qquad\qquad -2\alpha_0 \langle \boldsymbol{y}, \boldsymbol{e} \rangle - 2 \langle \boldsymbol{y}, X\boldsymbol{\alpha} \rangle + 2\alpha_0 \langle \boldsymbol{e}, X\boldsymbol{\alpha} \rangle \\
&= \langle \boldsymbol{y}, \boldsymbol{y} \rangle + \alpha_0^2 \langle \boldsymbol{e}, \boldsymbol{e} \rangle + \boldsymbol{\alpha}^\top X^\top X\boldsymbol{\alpha} \\
&\qquad\qquad -2\alpha_0 \langle \boldsymbol{y}, \boldsymbol{e} \rangle - 2\boldsymbol{y}^\top X\boldsymbol{\alpha} + 2\alpha_0 \boldsymbol{e}^\top X\boldsymbol{\alpha} \qquad \text{(C.7)}
\end{aligned}
$$

α_j を求めるためには (C.7) 式を α_0，$\boldsymbol{\alpha}$ で偏微分し，それぞれを 0 として解けばよい．したがって，

$$
\frac{\partial Q}{\partial \alpha_0} = 2\alpha_0 \langle \boldsymbol{e}, \boldsymbol{e} \rangle - 2 \langle \boldsymbol{y}, \boldsymbol{e} \rangle + 2 \langle \boldsymbol{e}, X\boldsymbol{\alpha} \rangle = 0 \qquad \text{(C.8)}
$$

[*1)] 要素数は n 個である．

$$\frac{\partial Q}{\partial \boldsymbol{\alpha}} = 2X^\top X\boldsymbol{\alpha} - 2\boldsymbol{y}^\top X + 2\alpha_0 \boldsymbol{e}^\top X = \boldsymbol{0} \tag{C.9}$$

となる．(C.8) 式より，α_0 は容易に求められる．

$$\alpha_0 = \bar{y} - \langle \bar{\boldsymbol{x}}, \boldsymbol{\alpha} \rangle = \bar{y} - \bar{\boldsymbol{x}}^\top \boldsymbol{\alpha} \tag{C.10}$$

ただし，

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\langle \boldsymbol{y}, \boldsymbol{e} \rangle = \frac{1}{n}\boldsymbol{y}^\top \boldsymbol{e}$$

$$\bar{\boldsymbol{x}}^\top = (\bar{\boldsymbol{x}}_1, \bar{\boldsymbol{x}}_2, \cdots, \bar{\boldsymbol{x}}_m) = \frac{1}{n}\left(\sum_{i=1}^{n} x_{i1}, \sum_{i=1}^{n} x_{i2}, \cdots, \sum_{i=1}^{n} x_{im}\right) = \frac{1}{n}\boldsymbol{e}^\top X$$

$$\langle \boldsymbol{e}, X\boldsymbol{\alpha} \rangle = \boldsymbol{e}^\top X\boldsymbol{\alpha} = n\bar{\boldsymbol{x}}^\top \boldsymbol{\alpha} = n\langle \bar{\boldsymbol{x}}, \boldsymbol{\alpha} \rangle$$

したがって，重回帰のモデル式では $\bar{\boldsymbol{x}}$ すなわち説明変数の平均に対して，$\bar{y}$ すなわち目的変数の平均が与えられるということがわかる．

(C.10) 式を (C.9) 式に代入すると以下の式が得られる．

$$X^\top X\boldsymbol{\alpha} - \boldsymbol{y}^\top X + (\bar{y} - \bar{\boldsymbol{x}}^\top \boldsymbol{\alpha})\boldsymbol{e}^\top X = \boldsymbol{0}$$

ここで，$(\boldsymbol{e}\bar{\boldsymbol{x}}^\top)^\top (X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top) = \boldsymbol{0}$，$X\boldsymbol{e}\bar{y} - (\boldsymbol{e}\bar{\boldsymbol{x}}^\top)^\top \boldsymbol{y} = \boldsymbol{0}$ に注意すると以下の式が得られる．

$$(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)^\top (X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha} = (X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)^\top (\boldsymbol{y} - \boldsymbol{e}\bar{y}) \tag{C.11}$$

(C.11) 式は**正規方程式**とよばれ，これを解くと $\boldsymbol{\alpha}$ を求めることができる．また，$\boldsymbol{y}$ の予測値 $\hat{\boldsymbol{y}}$ は以下のようにして与えられる．

$$\hat{\boldsymbol{y}} = \boldsymbol{e}\alpha_0 + X\boldsymbol{\alpha}$$

C.2.2　重回帰分析の幾何的な解釈

前項で回帰係数を求める方法を説明した．ここでは重回帰分析の幾何学的な解釈について説明する．(C.10) 式より，重回帰分析では $\bar{\boldsymbol{x}}$ に対して $\bar{y}$ が与えられるということを説明した．したがって，(C.6) 式は以下のようになる．

$$(\boldsymbol{y} - \boldsymbol{e}\bar{y}) = (X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha} + \boldsymbol{\varepsilon}$$

ここで，$\boldsymbol{y}-\boldsymbol{e}\bar{y}$ は $\boldsymbol{y}$ の各要素から $\bar{y}$ を引いたベクトルであり，$(X-\boldsymbol{e}\bar{\boldsymbol{x}}^\top)$ は X から列の平均 $\bar{\boldsymbol{x}}^\top$ を引いた行列である．そこで，これらを $\boldsymbol{w}$，$V=(\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m)$ と書き換えると以下のようになる．

$$\boldsymbol{w}=V\boldsymbol{\alpha}+\boldsymbol{\varepsilon}=\alpha_1\boldsymbol{v}_1+\alpha_2\boldsymbol{v}_2+\cdots\alpha_m\boldsymbol{v}_m+\boldsymbol{\varepsilon}$$

ここで，各サンプルを軸とする空間に変量ベクトル $\boldsymbol{w},\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m$ を付置した図を考える．

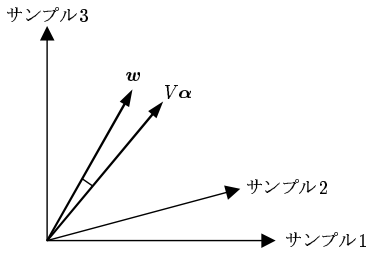

図 C.1 変量ベクトルの付置

重回帰分析では $\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_m$ を $\alpha_1,\alpha_2,\cdots,\alpha_m$ によって合成したベクトル $\hat{\boldsymbol{w}}$ が平面 $\mathcal{V}$ 上で作られる．そのとき，$\hat{\boldsymbol{w}}$ がなるべく $\boldsymbol{w}$ と一致するように $\boldsymbol{\alpha}$ を定める．また，$\hat{\boldsymbol{w}}$ と $\boldsymbol{w}$ のずれが $\boldsymbol{\varepsilon}$ となるので，なるべく一致されるということは $\boldsymbol{\varepsilon}$ のノルムを最小にすることである．$\|\boldsymbol{\varepsilon}\|^2$ が最小になるのは，$\hat{\boldsymbol{w}}$ が $\boldsymbol{w}$ の平面 $\mathcal{V}$ への射影となるときであり，$\boldsymbol{\varepsilon}$ は $\hat{\boldsymbol{w}}$ の垂線に一致する．これより重回帰分析に関するいくつのの性質をまとめる．

- 重回帰分析では 2 つのベクトル $\hat{\boldsymbol{w}}$，$\boldsymbol{\varepsilon}$ が直交するように $\boldsymbol{\alpha}$ を定めている．

$$\langle\hat{\boldsymbol{w}},\boldsymbol{\varepsilon}\rangle=\big\langle(X-\boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha},(\boldsymbol{y}-\boldsymbol{e}\bar{y})-(X-\boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\big\rangle=0$$

- 重回帰分析では 2 つのベクトル $\boldsymbol{w}$ と $\hat{\boldsymbol{w}}$ のなす角 θ を最小にする，すなわち $\cos\theta$ が最大になるように $\boldsymbol{\alpha}$ を定めている．

$$\cos\theta=\frac{\langle\boldsymbol{w},\hat{\boldsymbol{w}}\rangle}{\|\boldsymbol{w}\|\,\|\hat{\boldsymbol{w}}\|}=\frac{\big\langle\boldsymbol{y}-\boldsymbol{e}\bar{y},(X-\boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\big\rangle}{\|\boldsymbol{y}-\boldsymbol{e}\bar{y}\|\,\big\|(X-\boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\big\|}\to\max\qquad\text{(C.12)}$$

このとき，$\cos\theta$ は重相関係数 (R)，$\cos^2\theta$ は決定係数 (R^2) とよばれる．重回帰分析では R の値が高い程モデルのあてはまりがよいと考える[*1)]．

- ベクトル $\boldsymbol{w}$，$\hat{\boldsymbol{w}}$，$\boldsymbol{\varepsilon}$ の間には三平方の定理が成り立つ．これより，全変動 S_T は回帰による変動 S_R と誤差変動 S_e の和に等しくなる．

$$\|\boldsymbol{w}\|^2 = \|\hat{\boldsymbol{w}}\|^2 + \|\boldsymbol{\varepsilon}\|^2$$

$$\|\boldsymbol{y} - \boldsymbol{e}\bar{y}\|^2 = \left\|(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\right\|^2 + \|\boldsymbol{\varepsilon}\|^2$$

$$\underbrace{\|\boldsymbol{y} - \boldsymbol{e}\bar{y}\|^2}_{\text{全変動 } S_T} = \underbrace{\|\hat{\boldsymbol{y}} - \boldsymbol{e}\bar{y}\|^2}_{\text{回帰による変動 } S_R} + \underbrace{\|\boldsymbol{y} - \hat{\boldsymbol{y}}\|^2}_{\text{誤差変動 } S_e}$$

a. 分析結果の検討

多変量解析の手法はデータを入力すれば，何らかの分析結果が出力される．したがって，分析結果を鵜呑みにするのではなく，分析結果の妥当性を検討しなければならない．

Excel の出力結果をみるとパラメータの推定値以外にも様々な値が出力される．これらは主に分析結果の妥当性を検討するために利用される．そこで，以下では分析結果の検討にあたって最低限考慮すべき 3 つの側面について説明する．

1) モデルの説明力　説明力のあるモデルとは，説明変数と推定されたパラメータによって目的変数を忠実に再現できるモデルのことである．重回帰分析では誤差の 2 乗和 Q を最小にするようにパラメータを定めるので，この値 (残差平方和[*1)]) がどの程度小さくなったかということを調べればよいことになる．ところが，この値は目的変数の単位の取り方によって大きく変わってしまうので，一概には判断できない．そこで，目的変数の単位の取り方に依存しない方法を考える必要がある．

重回帰分析では，目的変数の理論値 $\hat{y}$ と実測値 y の間に以下の関係があ

[*1)] $\boldsymbol{w}$ とその $\mathcal{V}$ への射影である $\hat{\boldsymbol{w}}$ のなす角度は 0° 以上であり，また，90° を超えることはない．したがって，$0 \leq R \leq 1$ が成り立つ．

[*1)] 最小化された Q はモデル式によって説明がつかない部分であり，残差と呼ばれる．

ることが知られている[*2)](付録 C. 2. 2 参照).

$$\underbrace{\sum_{i=1}^{n}(y_i-\bar{y})^2}_{\text{全変動の偏差平方和 } S_T} = \underbrace{\sum_{i=1}^{n}(\hat{y}_i-\bar{\hat{y}})^2}_{\text{回帰モデルによる変動の偏差平方和 } S_R} + \underbrace{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}_{\text{残差変動の偏差平方和 } S_e}$$

この式は目的変数の変動 S_T が回帰モデルにより説明される理論値の変動 S_R と残差の変動 S_e に分解されることを意味しており，S_e に比べて S_R の割合が高いほどモデル式が説明力を持っていると解釈される．そこで，その比率を以下のように定める．

$$R^2 = \frac{S_R}{S_T} = 1 - \frac{S_e}{S_T} \tag{C.13}$$

この比率は**決定係数** (coefficient of determination) と呼ばれる．また，決定係数の平方根は理論値 $\hat{y}$ と実測値 y の相関係数に等しく，**重相関係数** (multiple correlation coefficient) と呼ばれる．

決定係数 R^2 には，説明変数の数を増やしていくと，S_T は一定のまま S_e が小さくなるという性質がある．サンプル数を一定として説明変数を増やしていくと R^2 は 1 に近づくことに注意されたい．

さらに，モデルの説明力を統計的に検定したいという場合には，表 C.4 のような分散分析を行う．このとき，$F_0 \geq F^p_{n-p-1}(\beta)$ ならば，有意水準 β でこの回帰は有意であるということになる．ここで，n はサンプル数，p は説明変数の数，F^p_{n-p-1} は自由度 $p, n-p-1$ の F 値である．

表 C.4 分散分析表

変動要因	自由度	偏差平方和	不偏分散	分散比
モデル式による変動	p	S_R	$V_R = S_R/p$	$F_0 = V_R/V_e$
残差変動	$n-p-1$	S_e	$V_e = S_e/(n-p-1)$	
全変動	$n-1$	S_T		

[*2)] $\bar{y} = \bar{\hat{y}}$ であることに注意されたい (付録 C. 2. 1 の (C.10) 式参照).

2) 個々の説明変数の妥当性　変数 x_j を説明変数とすることの妥当性について，以下の 2 つの視点から検討する必要がある．

(1) 説明変数が原因を示す変数，目的変数が結果を示す変数という関係になっているのか．

(2) 各説明変数がどの程度目的変数を説明するのに役立っているのか．

ここで，(1) については対象とする問題固有の定性的な要素が含まれる[*1)]．重回帰分析はあくまでも変数間の相関関係を分析するものなので，分析結果から (1) を結論づけることはできない．(2) については，パラメータの検定，偏相関係数などいくつか検討方法があるが，ここではパラメータの検定について説明する．

いま仮に，真のモデルでは説明変数 x_j と y はまったく無関係 (独立) であったとする．このとき，モデル式では $\alpha_j = 0$ となるが，与えられたデータを用いて推定値 $\hat{\alpha}_j$ を求めると何らかの値が算出される．この値はよほどの偶然でもない限り $\hat{\alpha}_j = 0$ となることはないが，ある確率 (危険率または有意水準 β) で 0 を中心とした特定の範囲に納まるはずである．また，サンプルを多く取れば推定値は真の値に近づくことが期待されるので，この範囲は狭くなることが期待される．

この考え方に基づいて，パラメータの検定では各パラメータについて $\alpha_j = 0$ という帰無仮説 H_0 を考える．そして，以下の不等式が成立するならば，有意水準 β で帰無仮説 H_0 は棄却される．

$$|t_0| = \frac{|\hat{\alpha}_j|}{SE(\hat{\alpha}_j)} \geq t_{n-p-1}(\beta), \quad SE(\hat{\alpha}_j) = \sqrt{s^{jj} V_e/(n-1)}$$

この不等式の左辺は t 値と呼ばれ，推定値 $\hat{\alpha}_j$ が単位の取り方に依存しないように $SE(\hat{\alpha}_j)$ で基準化されている[*2)]．また，V_e は表 C.4 の分散分析表にある誤差の不偏分散である．

[*1)] 具体例としては，いわゆる「コウノトリの繁殖率と赤ん坊の出生率」がある．ある都市でコウノトリの繁殖率と赤ん坊の出生率に正の相関が認められた．そこで，コウノトリの繁殖率を説明変数，赤ん坊の出生率を目的変数として重回帰分析を行ったら，結果は有為であったらしい．この分析の大きな誤りはコウノトリの繁殖率と赤ん坊の出生率という「結果のデータ」同士で分析を行っていることである．実際に背後にあった原因は，産業の発達に伴う都市化の進展であった．

[*2)] $SE(\hat{\alpha}_j)$ は $\hat{\alpha}_j$ の標準誤差であり，s^{jj} は $x_1, x_2, \cdots, x_p$ に関する共分散行列の逆行列における第 j 番目の対角要素である．

重回帰分析では，誤差 ε_i が $N(0,\sigma^2)$ に従うと仮定して $\hat{\alpha}_j$ を推定すると，推定値は α_j を中心とした t 分布に従うことが知られている．特に，この分布の標準偏差 $SE(\hat{\alpha}_j)$ は標準誤差と呼ばれる．

3) モデルの良さ　多変量解析では説明力があり，かつ単純な構造をもつモデルを良いモデルと考える．しかし，説明力と単純な構造はトレード・オフの関係にある．そこで，両者を勘案してモデルの当てはまりのよさを測る尺度として，自由度調整済み決定係数や AIC (Akaike Information Criteria) などが提案されている．以下では自由度調整済み決定係数についてのみ説明する．

決定係数 R^2 には，説明変数の数を増やしていくと，S_T は一定のまま S_e が小さくなるという性質がある．この欠点を改善するために，以下のように (C.13) 式の総平方和 S_T と残差平方和 S_e をそれぞれ不変分散 V_T, V_e で置き換える．

$$\bar{R}^2 = 1 - \frac{V_e}{V_T}$$

この値は**自由度調整済み決定係数** (coefficient of determination adjusted for the degrees of freedom) とよばれる．なお，決定係数 R^2 は $0 < R^2 < 1$ であるが，自由度調整済み決定係数 $\bar{R}^2$ は

$$\bar{R}^2 = 1 - \frac{S_e/(n-p-1)}{S_T/(n-1)}$$

であることより，n もしくは $n-p$ が小さい (すなわち p が大きい) ときにマイナスになることもあるので注意されたい．

C. 3　正準相関分析

重回帰分析では 1 つの基準変量に対する複数の説明変数の関係を求めた．それに対して，基準変数も複数に拡張したものが正準相関分析である．

正準相関分析 (canonical correlation analysis) は，2 組の変数群 X および Y の関係を知りたいという場合を考える．X, Y は以下のように与えられて

いるとする．

$$[X|Y] = \left[\begin{array}{ccc|ccc} x_{11} & \cdots & x_{1p} & y_{11} & \cdots & y_{1q} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & \cdots & x_{np} & y_{n1} & \cdots & y_{nq} \end{array}\right] \tag{C.14}$$

ここで，各群による合成変数ベクトル $\boldsymbol{f} = X\boldsymbol{a}$, $\boldsymbol{g} = Y\boldsymbol{b}$ を考える．$\boldsymbol{f}$ と $\boldsymbol{g}$ の相関係数は，

$$\rho = \frac{\langle \boldsymbol{f}, \boldsymbol{g} \rangle}{\|\boldsymbol{f}\|\|\boldsymbol{g}\|} = \frac{\boldsymbol{a}^\top \Sigma_{XY} \boldsymbol{b}}{\sqrt{\boldsymbol{a}^\top \Sigma_X \boldsymbol{a}}\sqrt{\boldsymbol{b}^\top \Sigma_Y \boldsymbol{b}}} \tag{C.15}$$

となる．Σ_{XY} は X と Y の共分散行列であり，

$$\Sigma_{XY} = \frac{1}{n-1}\left\{(X - \boldsymbol{e}\bar{\boldsymbol{x}}_{\cdot j}^\top)^\top (X - \boldsymbol{e}\bar{\boldsymbol{x}}_{\cdot j}^\top)\right\} \tag{C.16}$$

となる．重回帰分析と同様に，合成変数ベクトル $\boldsymbol{f}$ と $\boldsymbol{g}$ の相関係数を最大にすることで $\boldsymbol{f}$ と $\boldsymbol{g}$ の関係がもっとも良く表されるものと考えると，(C.16) 式の分母について $\boldsymbol{a}^\top \Sigma_X \boldsymbol{a} = 1$, $\boldsymbol{b}^\top \Sigma_Y \boldsymbol{b} = 1$ という条件をおいて，分子を最大化すればよい．したがって，このときラグランジュ関数は，

$$L(\boldsymbol{a}, \boldsymbol{b}, \boldsymbol{\lambda}) = \boldsymbol{a}^\top \Sigma_{XY} \boldsymbol{b} - \lambda_a(\boldsymbol{a}^\top \Sigma_X \boldsymbol{a} - 1) - \lambda_b(\boldsymbol{b}^\top \Sigma_Y \boldsymbol{b} - 1) \tag{C.17}$$

となる．したがって，最適性の条件は，

$$\frac{\partial L}{\partial \boldsymbol{a}} = \Sigma_{XY}\boldsymbol{b} - 2\lambda_a \Sigma_X \boldsymbol{a} = \boldsymbol{0} \tag{C.18}$$

$$\frac{\partial L}{\partial \boldsymbol{b}} = \Sigma_{YX}\boldsymbol{a} - 2\lambda_b \Sigma_Y \boldsymbol{b} = \boldsymbol{0} \tag{C.19}$$

となる．(C.18), (C.19) 式にそれぞれ $\boldsymbol{a}^\top$, $\boldsymbol{b}^\top$ を左から乗じてまとめると，

$$2\lambda_a = \boldsymbol{a}^\top \Sigma_{XY} \boldsymbol{b}$$

$$2\lambda_b = \boldsymbol{b}^\top \Sigma_{XY} \boldsymbol{a}$$

となる．ここでこれら 2 式の右辺は等しいので，$\lambda_a = \lambda_b$ となる．そこで $\lambda_a = \lambda_b = \sqrt{\lambda}/2$ とし，(C.18), (C.19) 式に代入する．これを $\boldsymbol{b}$ について解くと，

$$(\Sigma_{XY}\Sigma_Y^{-1}\Sigma_{YX} - \lambda\Sigma_X)\boldsymbol{a} = \boldsymbol{0} \tag{C.20}$$

という固有値問題が得られる．(C.20) 式から得られる固有値は合成変量 $\boldsymbol{f}$ と $\boldsymbol{g}$ の相関係数を表す．(C.20) 式は正の固有値を $r = \min(p, q)$ 個もつ．それらの固有値を大きい順に $\lambda_1, \lambda_2, \cdots, \lambda_r$ とする．このとき λ_1 に対する固有ベクトル $\boldsymbol{a}$(および，その $\boldsymbol{a}$ に対応する $\boldsymbol{b}$) により得られる合成変量 $\boldsymbol{f}, \boldsymbol{g}$ を**第 1 正準変量**という．また，λ_1 を**第 1 正準係数**という．第 1 正準変量だけでは元の変量群の関係をうまく表せ切れていない場合には，以下順に λ_2, $\lambda_3 \cdots$ と採用し，これらにより得られる合成変量を用いる．それぞれを第 2 正準変量，第 3 正準変量 $\cdots$ と呼ぶ．

実際の計算では，各変量はあらかじめ平均を 0，分散を 1 に標準化しておくことが一般である．

C. 4 判別分析

C. 4. 1 判別問題

二群の判別問題を例として，判別問題のモデルと判別分析の考え方について説明する．

はじめに判別問題の前提を整理すると以下のようになる．

(1) 2 つの群の母集団 G_1, G_2 から，それぞれ大きさ n_1, n_2 個のサンプルが与えられている．

$$\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)}; \boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)}$$

ただし，$\mathcal{S}$ を標本空間としたとき，$\mathcal{S} \subset \mathbb{R}^p$ であり，$\boldsymbol{x} \in \mathcal{S}$ である．

(2) 各群の母集団は既知の確率密度関数 $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ に従っている．

(3) 未知のサンプルが G_1 から発生する事前確率，G_2 から発生する事前確率はそれぞれ $\Pr(1)$, $\Pr(2)$ であり，既知である．

このとき判別問題のプロセスは次のようになる．

(1) 未知のサンプルが $\Pr(1)$, $\Pr(2)$ に従って発生する．この時点でどちらの群から発生したのかは決まっているが，観測者は知ることができない．

(2) 未知のサンプルは，属する群の $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ に従って $\boldsymbol{x}$ の値をとる．

(3) 観測者は $\boldsymbol{x}$ の値よりどちらの群に属するか判別する．

図 C.2 は判別問題のプロセスを図示したものである．判別分析では標本空間 $\mathcal{S}$ を第 1 群，第 2 群の領域 $\mathcal{R}_1$, $\mathcal{R}_2$ に分割し，どちらの領域に属するかによって判別を行う．

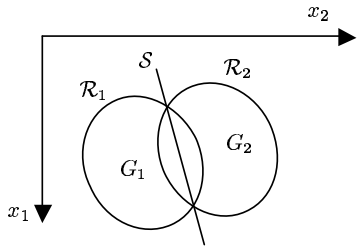

図 C.2　判別問題のイメージ

ある観測値 $\boldsymbol{x}$ が得られたとき，これが G_1 からの観測値であるときに $\mathcal{R}_1$ に含まれる確率 $\Pr(1|1)$，G_1 からの観測値であるにも関わらず $\mathcal{R}_2$ に含まれる確率 $\Pr(2|1)$ は

$$\Pr(1|1) = \int_{\mathcal{R}_1} f_1(\boldsymbol{x})d\boldsymbol{x}, \quad \Pr(2|1) = \int_{\mathcal{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x}$$

である．同様に，G_2 からの観測値が $\mathcal{R}_2$ に含まれる確率 $\Pr(2|2)$，$\mathcal{R}_1$ に含まれる確率 $\Pr(1|2)$ は

$$\Pr(2|2) = \int_{\mathcal{R}_2} f_2(\boldsymbol{x})d\boldsymbol{x}, \quad \Pr(1|2) = \int_{\mathcal{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x}$$

である．ただし，$d\boldsymbol{x} = dx_1dx_2\cdots dx_n$ である．

G_1 からの観測値を G_2 と誤判別による損失を $C(2|1)$，G_2 からの観測値を G_1 と誤判別による損失を $C(1|2)$ としたとき，誤判別による損失の期待値は

$$\begin{aligned} &C(2|1)\Pr(1)\Pr(2|1) + C(1|2)\Pr(2)\Pr(1|2) \\ &\quad = C(2|1)\Pr(1)\int_{\mathcal{R}_2} f_1(\boldsymbol{x})d\boldsymbol{x} + C(1|2)\Pr(2)\int_{\mathcal{R}_1} f_2(\boldsymbol{x})d\boldsymbol{x} \end{aligned} \tag{C.21}$$

となり，特に $C(2|1)$=$C(1|2)$ であるならば誤判別確率となる．判別分析とはこれを最小にするような空間を分割する問題と考える．

このとき上式は

$$\int_{\mathcal{R}_1} \{C(1|2)\Pr(2)f_2(\boldsymbol{x}) - C(2|1)\Pr(1)f_1(\boldsymbol{x})\}\, d\boldsymbol{x} + C(2|1)\Pr(1)\int_{\mathcal{S}} f_1(\boldsymbol{x})d\boldsymbol{x}$$

となり[*1)]，第 2 項が定数であることに注意すると[*2)]

$$\mathcal{R}_1 = \{\boldsymbol{x}\,|\,C(2|1)\Pr(1)f_1(\boldsymbol{x}) > C(1|2)\Pr(2)f_2(\boldsymbol{x})\}$$

を満たす点 $\boldsymbol{x}$ の集合を $\mathcal{R}_1$ に取れば，第 1 項をが最小になることがわかる．また，$\mathcal{R}_2$ は

$$\mathcal{R}_2 = \{\boldsymbol{x}\,|\,C(2|1)\Pr(1)f_1(\boldsymbol{x}) < C(1|2)\Pr(2)f_2(\boldsymbol{x})\}$$

となり，以下の式を満たす点 $\boldsymbol{x}$ の集合は判別境界となる．

$$C(2|1)\Pr(1)f_1(\boldsymbol{x}) = C(1|2)\Pr(2)f_2(\boldsymbol{x})$$

一方，事前確率 $\Pr(1), \Pr(2)$ が既知であるので，観測値 $\boldsymbol{x}$ が第 k 群から発生したデータである確率 $\Pr(k\,|\,\boldsymbol{x})$ は Bayes の公式より

$$\Pr(k\,|\,\boldsymbol{x}) = \frac{\Pr(k)f_k(\boldsymbol{x})}{\Pr(1)f_1(\boldsymbol{x}) + \Pr(2)f_2(\boldsymbol{x})}, \quad k = 1, 2$$

となる．したがって，$C(2|1) = C(1|2)$ のときに誤判別による損失の期待値を最小にするためには $\Pr(k\,|\,\boldsymbol{x})$ を比較して判別すればよいということがわかる．一般にこのような判別法は **Bayes 決定法**と呼ばれる．

さらに，

(1) $C(2|1) = C(1|2)$ かつ $\Pr(1) = \Pr(2)$ である．

(2) $f_1(\boldsymbol{x})$ と $f_2(\boldsymbol{x})$ が $N(\boldsymbol{\mu}^{(1)}, \Sigma^{(1)})$, $N(\boldsymbol{\mu}^{(2)}, \Sigma^{(2)})$ に従う (**正規性**).

(3) $\Sigma^{(1)} = \Sigma^{(2)}$ である (**等分散性**).

[*1)] $\mathcal{R}_2 \cap \mathcal{R}_1 = \mathcal{S}$, $\mathcal{R}_2 \cup \mathcal{R}_1 = \emptyset$ であることに注意せよ．

[*2)] $C(2|1)$, $\Pr(1)$ は既知であり，$\int_{\mathcal{S}} f_1(\boldsymbol{x})d\boldsymbol{x} = 1$ である．

という条件が満たされるとき，Bayes 決定法はマハラノビス汎距離に帰着する（第 C. 4. 3 項を参照）.

ここで，(3) の条件が満たされないとき，判別境界は二次曲線となる (**二次判別分析**)．また，(2)，(3) の条件が満たされないときは，直接 $f_1(\boldsymbol{x})$ と $f_2(\boldsymbol{x})$ の大きさを比較することになるが，このときの判別境界は非線形の曲線となる.

なお，正規性，等分散性の検定については木島ら (木島・小守林, 1999) を参照されたい.

C. 4. 2　相関比の最大化

ここでは相関比の最大化について，その考え方を 2 変量の二群判別問題で説明する．図 C.3 はサンプル分布を山に見立てたイメージ図である[*1)]．このとき，2 つの山を様々な方角から見るとそれぞれの方角で山の重なり具合が異なる．そこで，2 つの山がはっきりと見分けられる方角を定め，山の尾根に沿って判別境界を引けば誤判別確率が最小になることが “期待できる”.

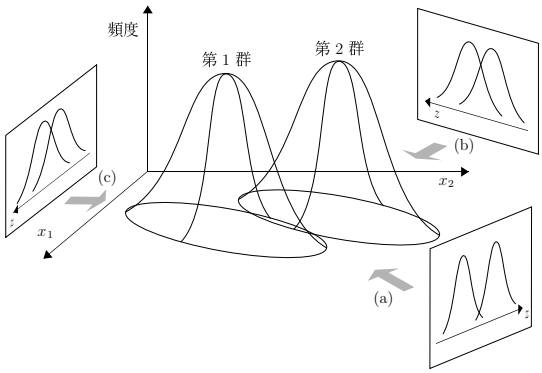

図 C.3　各グループの分布イメージ

いま，(a), (b), (c) の 3 つの方角を考え，そこから見た山の写像をそれぞ

[*1)] この山は正規分布と同じ形状をしているとは限らず，また，2 つの山の形状，大きさも等しいとは限らないものとする．したがって，両群の共分散行列も特に等しいわけではない.

れ平面上に図示する．このように特定の方角を決めるとそれに応じて判別境界と平面が得られるが，この平面上の横軸が合成変量 z となる．

図 C.3 では，2 つの山がはっきりと区別できる方角として方角 (a) を定めることは簡単である．しかし，これ多変量データとなった場合，どのような手順で方角 (a) を見つけだすかが問題となる．

図 C.4 で示したように，総平方和 S_T は全体のばらつき，群間平方和 S_B は群の離れ具合，群内平方和 S_W は群内のばらつきに対応している．また，これらの間には $S_T = S_B + S_W$ の関係があり，(a), (b), (c) をはじめどのような方角から見ても成り立つ．

2 つの山を見る方角を変えることによって総平方和 S_T に占める群間平方和 S_B の割合が大きくなれば，相対的に S_W が小さくなる．したがって，群間平方和 S_B の割合が最も大きいところでは，2 つの山が最も離れており，両方の山も幅が狭く見える．そこで，**相関比** (correlation ratio) $\eta^2 = S_B/S_W$ の値を最大化するパラメータを求め，2 つの山が最も区別される方角を確定する．

二群判別問題のデータを一般的な形式で記述すると表 C.5 のようになる．

どちらに属するかわからない新しいサンプルを $(x_1, x_2, \cdots, x_m)$ としたとき，どちらに属するか判別するルールとして以下の線形判別関数を考える．

$$z = \sum_{j=1}^{m} \alpha_j x_j, \quad i = 1, 2, \cdots, n \tag{C.22}$$

判別分析では線形結合によって作られた z と基準となる値の大小比較によって判別を行う．

パラメータ α_j は (C.22) 式に表 C.5 を当てはめたとき “最もよく” 判別されるよう定める．

$$\underbrace{\begin{bmatrix} \boldsymbol{z}^{(1)} \\ \boldsymbol{z}^{(2)} \end{bmatrix}}_{\boldsymbol{z}} = \underbrace{\begin{bmatrix} X^{(1)} \\ X^{(2)} \end{bmatrix}}_{X} \boldsymbol{\alpha}, \quad k = 1, 2$$

ただし，$\boldsymbol{z}^{(k)} = (z_1^{(k)}, z_2^{(k)}, \cdots, z_{n_k}^{(k)})^\top$，$X^{(k)} = (\boldsymbol{x}_1^{(k)}, \boldsymbol{x}_2^{(k)}, \cdots, \boldsymbol{x}_m^{(k)})$，$\boldsymbol{x}_j^{(k)} = (x_{1j}^{(k)}, x_{2j}^{(k)}, \cdots, x_{n_k j}^{(k)})^\top$，$\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \cdots, \alpha_m)^\top$ である．

表 C.5　判別分析のデータ形式

	サンプル No.	目的変数		説明変数					
		1 群	2 群	1	2	$\cdots$	j	$\cdots$	m
第	1	1	0	$x_{11}^{(1)}$	$x_{12}^{(1)}$	$x_{1\cdots}^{(1)}$	$x_{1j}^{(1)}$	$x_{1\cdots}^{(1)}$	$x_{1m}^{(1)}$
	2	1	0	$x_{21}^{(1)}$	$x_{22}^{(1)}$	$x_{2\cdots}^{(1)}$	$x_{2j}^{(1)}$	$x_{2\cdots}^{(1)}$	$x_{2m}^{(1)}$
1	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	i	1	0	$x_{i1}^{(1)}$	$x_{i2}^{(1)}$	$x_{i\cdots}^{(1)}$	$x_{ij}^{(1)}$	$x_{i\cdots}^{(1)}$	$x_{im}^{(1)}$
群	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	n_1	1	0	$x_{n_1 1}^{(1)}$	$x_{n_1 2}^{(1)}$	$x_{n_1\cdots}^{(1)}$	$x_{n_1 j}^{(1)}$	$x_{n_1\cdots}^{(1)}$	$x_{n_1 m}^{(1)}$
第	1	0	1	$x_{11}^{(2)}$	$x_{12}^{(2)}$	$x_{1\cdots}^{(2)}$	$x_{1j}^{(2)}$	$x_{1\cdots}^{(2)}$	$x_{1m}^{(2)}$
	2	0	1	$x_{21}^{(2)}$	$x_{22}^{(2)}$	$x_{2\cdots}^{(2)}$	$x_{2j}^{(2)}$	$x_{2\cdots}^{(2)}$	$x_{2m}^{(2)}$
2	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	i	0	1	$x_{i1}^{(2)}$	$x_{i2}^{(2)}$	$x_{i\cdots}^{(2)}$	$x_{ij}^{(2)}$	$x_{i\cdots}^{(2)}$	$x_{im}^{(2)}$
群	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\ddots$	$\vdots$
	n_2	0	1	$x_{n_2 1}^{(2)}$	$x_{n_2 2}^{(2)}$	$x_{n_2\cdots}^{(2)}$	$x_{n_2 j}^{(2)}$	$x_{n_2\cdots}^{(2)}$	$x_{n_2 m}^{(2)}$
パラメータ				α_1	α_2	$\cdots$	α_j	$\cdots$	α_m

ここで，$\boldsymbol{z}$ は，分散分析と同様に全体のばらつきである総平方和 S_T は，群の離れ具合である群間平方和 S_B と群内のばらつきである群内平方和 S_W に分解される (図 C.4).

$$\underbrace{\|\boldsymbol{z}-\boldsymbol{e}\bar{z}\|^2}_{\text{総平方和 } S_T} = \underbrace{\|\bar{\boldsymbol{z}}_W-\boldsymbol{e}\bar{z}\|^2}_{\text{群間平方和 } S_B} + \underbrace{\|\boldsymbol{z}-\bar{\boldsymbol{z}}_W\|^2}_{\text{群内平方和 } S_W}$$

ただし，$\bar{\boldsymbol{z}}_W = (\underbrace{\bar{z}^{(1)},\cdots,\bar{z}^{(1)}}_{n_1\text{ 個}},\underbrace{\bar{z}^{(2)},\cdots,\bar{z}^{(2)}}_{n_2\text{ 個}})^\top$ である.

二群判別分析では以下に示す相関比を最大にすることを “2 つの群が最もよく判別された” と考える.

$$\eta^2 = \frac{S_B}{S_T} \to \max$$

これを $\boldsymbol{\alpha}$ で偏微分して 0 とおくと以下のようになる.

$$\frac{\partial\eta^2}{\partial\boldsymbol{\alpha}} = \frac{1}{S_T^2}\left(\frac{\partial S_B}{\partial\boldsymbol{\alpha}}S_T - S_B\frac{\partial S_T}{\partial\boldsymbol{\alpha}}\right) = \frac{1}{S_T}\left(\frac{\partial S_B}{\partial\boldsymbol{\alpha}} - \eta^2\frac{\partial S_T}{\partial\boldsymbol{\alpha}}\right) = 0$$

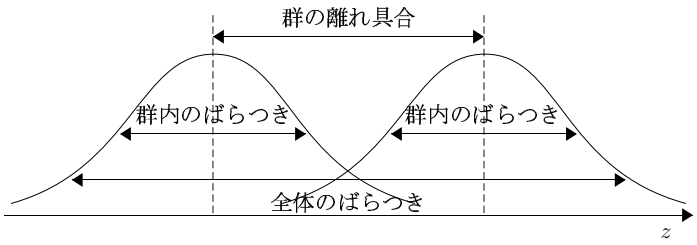

図 C.4 全平方和，群間平方和，群内平方和の関係

ここで，

$$
\begin{aligned}
S_T &= \|\boldsymbol{z} - \boldsymbol{e}\bar{z}\|^2 = \left\|X\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\boldsymbol{\alpha}\right\|^2 = \left\|(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\right\|^2 \\
&= \boldsymbol{\alpha}^\top \underbrace{\left(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right)^\top \left(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right)}_{T} \boldsymbol{\alpha} \\
S_B &= \|\bar{\boldsymbol{z}}_W - \boldsymbol{e}\bar{z}\|^2 = \left\|\bar{X}_W\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\boldsymbol{\alpha}\right\|^2 = \left\|(\bar{X}_W - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha}\right\|^2 \\
&= \boldsymbol{\alpha}^\top \underbrace{\left(\bar{X}_W - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right)^\top \left(\bar{X}_W - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right)}_{B} \boldsymbol{\alpha}
\end{aligned}
$$

に着目すると以下の一般固有値問題が得られる．

$$B\boldsymbol{\alpha} - \eta^2 T\boldsymbol{\alpha} = 0$$

ただし，

$$\bar{X}_W = \boldsymbol{y}^{(1)}(\boldsymbol{x}^{(1)})^\top + \boldsymbol{y}^{(2)}(\boldsymbol{x}^{(2)})^\top$$

である．このとき，最大固有値が相関比，固有ベクトルが $\boldsymbol{\alpha}$ となる．しかし，2 群の判別分析の場合，より簡単に $\boldsymbol{\alpha}$ を求めることができる．

はじめに

$$S_B = \frac{n_1 n_2}{n}\left\langle \frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}, X\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^\top \right\rangle^2$$

である．ここで，S_B は次のように求められる．

$$S_B = \left[\bar{\boldsymbol{z}}_W - \boldsymbol{e}\bar{z}\right]^\top \left[\bar{\boldsymbol{z}}_W - \boldsymbol{e}\bar{z}\right]$$

$$
\begin{aligned}
&= n_1(\bar{z}^{(1)})^2 + n_2(\bar{z}^{(2)})^2 - n\bar{z}^2, \quad (n = n_1 + n_2) \\
&= \frac{(n_1+n_2)n_1(\bar{z}^{(1)})^2}{n} + \frac{(n_1+n_2)n_2(\bar{z}^{(2)})^2}{n} - \frac{(n_1\bar{z}^{(1)} + n_2\bar{z}^{(2)})^2}{n} \\
&= \frac{(n_1+n_2)n_1(\bar{z}^{(1)})^2 + (n_1+n_2)n_2(\bar{z}^{(2)})^2 - (n_1\bar{z}^{(1)} + n_2\bar{z}^{(2)})^2}{n} \\
&= \frac{n_1n_2(\bar{z}^{(1)})^2 + n_1n_2(\bar{z}^{(2)})^2 - 2n_1n_2\bar{z}^{(1)}\bar{z}^{(2)}}{n} \\
&= \frac{n_1n_2}{n}(\bar{z}^{(1)} - \bar{z}^{(2)})^2
\end{aligned}
$$

ここで,

$$
\begin{aligned}
\bar{z}^{(1)} - \bar{z}^{(2)} &= \left[(\bar{\boldsymbol{x}}^{(1)})^\top - (\bar{\boldsymbol{x}}^{(2)})^\top\right]\boldsymbol{\alpha}, \quad \bar{z}^{(1)} = (\bar{\boldsymbol{x}}^{(1)})^\top\boldsymbol{\alpha} \\
&= \left[\frac{1}{n_1}(\boldsymbol{y}^{(1)})^\top X - \frac{1}{n_2}(\boldsymbol{y}^{(2)})^\top X\right]\boldsymbol{\alpha} \\
&= \frac{n}{n_1n_2}\left[\frac{n_2}{n}(\boldsymbol{y}^{(1)})^\top - \frac{n_1}{n}(\boldsymbol{y}^{(2)})^\top\right]X\boldsymbol{\alpha} \\
&= \frac{n}{n_1n_2}\left[\frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}\right]^\top (X\boldsymbol{\alpha})
\end{aligned}
$$

となり，以下の点に注意すると,

$$
\boldsymbol{e}^\top\left[\frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}\right] = \frac{n_2}{n}\boldsymbol{e}^\top\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{e}^\top\boldsymbol{y}^{(2)}\frac{n_1n_2}{n} - \frac{n_1n_2}{n} = 0
$$

さらに以下のように書くことができる.

$$
\begin{aligned}
\bar{z}^{(1)} - \bar{z}^{(2)} &= \frac{n}{n_1n_2}\left[\frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}\right]^\top (X\boldsymbol{\alpha}) \\
&\quad - \frac{n}{n_1n_2}\left[\frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}\right]^\top \boldsymbol{e}(\bar{\boldsymbol{x}}^\top\boldsymbol{\alpha}) \\
&= \frac{n}{n_1n_2}\left[\frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}\right]^\top (X\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^\top)
\end{aligned}
$$

このとき，相関比は以下のようになる.

$$
\eta^2 = \frac{S_B}{S_T} = \frac{n_1n_2}{n}\frac{\left\langle \frac{n_2}{n}\boldsymbol{y}^{(1)} - \frac{n_1}{n}\boldsymbol{y}^{(2)}, X\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right\rangle^2}{\left\|\left(X - \boldsymbol{e}\bar{\boldsymbol{x}}^\top\right)\boldsymbol{\alpha}\right\|^2} \to \max
$$

すなわち，相関比を最大にすることは

$$\frac{\left\langle \dfrac{n_2}{n}\boldsymbol{y}^{(1)} - \dfrac{n_1}{n}\boldsymbol{y}^{(2)}, X\boldsymbol{\alpha} - \boldsymbol{e}\bar{\boldsymbol{x}}^{\top} \right\rangle}{\left\| \dfrac{n_2}{n}\boldsymbol{y}^{(1)} - \dfrac{n_1}{n}\boldsymbol{y}^{(2)} \right\| \left\| \left(X - \boldsymbol{e}\bar{\boldsymbol{x}}^{\top}\right)\boldsymbol{\alpha} \right\|} \to \max$$

を最大化する $\boldsymbol{\alpha}$ を求めることに他ならない．

ここで (C.12) 式と比較して考えると，これは第 1 群には $n_2/(n_1+n_2)$，第 2 群には $-n_1/(n_1+n_2)$ を付与し，これを目的変数とした重回帰分析のパラメータを求めることと同じである[*1)]．

C.4.3 マハラノビス汎距離

$g \geq 2$ である g 個の群があり，大きさ $n_1, n_2, \cdots, n_g$ の p 変量データ $(x_1, x_2, \cdots, x_p)$ がそれぞれで与えられているとする．

$$\boldsymbol{x}_1^{(1)}, \boldsymbol{x}_2^{(1)}, \cdots, \boldsymbol{x}_{n_1}^{(1)};\ \boldsymbol{x}_1^{(2)}, \boldsymbol{x}_2^{(2)}, \cdots, \boldsymbol{x}_{n_2}^{(2)};\ \cdots;\ \boldsymbol{x}_1^{(g)}, \boldsymbol{x}_2^{(g)}, \cdots, \boldsymbol{x}_{n_g}^{(g)}$$

ただし，各群の母集団の平均と共分散行列は，それぞれ

$$\boldsymbol{\mu}^{(k)} = \left[\mu_1^{(k)}, \mu_2^{(k)}, \cdots, \mu_p^{(k)}\right]^{\top}, \quad k = 1, 2, \cdots, g$$

$$\Sigma^{(k)} = \left(\sigma_{jj'}^{(k)}\right), \quad j, j' = 1, 2, \cdots, p, \quad k = 1, 2, \cdots, g$$

であり，何らかの分布に従っているものとする．

本項では，母集団分布における各群の共分散行列が

$$\Sigma^{(1)} = \Sigma^{(2)} = \cdots = \Sigma^{(k)} = \Sigma$$

のように共通である場合の多群判別問題を考え，そのときの判別ルールの 1 つであるマハラノビス汎距離について説明する．

この判別ルールでは，未知のサンプル $\boldsymbol{x}$ が与えられたとき，各群の平均 $\boldsymbol{\mu}^{(1)}$, $\boldsymbol{\mu}^{(2)}$, $\cdots$, $\boldsymbol{\mu}^{(g)}$ からの距離を計算して，一番近い群に属すると判定す

[*1)] $n_2/n\boldsymbol{y}^{(1)} - n_1/n\boldsymbol{y}^{(2)}$ より目的変数の平均は 0 であることに注意されたい．

る．ただし，ここでの距離はユークリッド距離ではなく，以下のように定義する距離 $d_{(k)}$ を用いる．

$$d^2_{(k)} = \left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right)^\top \Sigma^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right), \quad k = 1, 2, \cdots, g$$

この距離 $d_{(k)}$ は**マハラノビス汎距離** (Mahalanobis generalized distance) と呼ばれ，各変量の分散や変量間の相関が考慮されている[*2)]．

a. 線形判別関数の導出

第 k 群と第 ℓ 群を判別する判別関数 $z(\boldsymbol{x})$ は以下のように 1 次式として導出される．

$$\begin{aligned}
&z_{k\ell}(\boldsymbol{x}) = d^2_{(\ell)} - d^2_{(k)} \\
&= \left(\boldsymbol{x} - \boldsymbol{\mu}^{(\ell)}\right)^\top \Sigma^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(\ell)}\right) - \left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right)^\top \Sigma^{-1} \left(\boldsymbol{x} - \boldsymbol{\mu}^{(k)}\right) \\
&= -2\boldsymbol{x}^\top \Sigma^{-1} \left(\boldsymbol{\mu}^{(\ell)} - \boldsymbol{\mu}^{(k)}\right) + \left(\boldsymbol{\mu}^{(\ell)} + \boldsymbol{\mu}^{(k)}\right)^\top \Sigma^{-1} \left(\boldsymbol{\mu}^{(\ell)} - \boldsymbol{\mu}^{(k)}\right) \\
&= 0, \quad k, \ell = 1, 2, \cdots, g
\end{aligned}$$

実際に，この判別ルールを適用するためには両群の母集団の平均 $\boldsymbol{\mu}^{(k)}$ や共分散行列 $\Sigma = (\sigma_{jj'})$ を知る必要があるが，これらは未知である．そこで，これらの代わりに平均と共分散行列の不偏推定量 $\bar{\Sigma} = \bar{\boldsymbol{x}}^{(k)}, (s_{jj'})$ を用いる．

$$\begin{aligned}
&\boldsymbol{\mu}^{(k)}:\ \bar{\boldsymbol{x}}^{(k)} = \left[\bar{x}^{(k)}_1, \bar{x}^{(k)}_2, \cdots, \bar{x}^{(k)}_p\right], \\
&\sigma_{jj'}:\ s_{jj'} = \frac{1}{n-g} \sum_{k=1}^{2} \sum_{i=1}^{n_k} \left(x^{(k)}_{ji} - \bar{x}^{(k)}_j\right) \left(x^{(k)}_{j'i} - \bar{x}^{\prime(k)}_j\right), \\
&\qquad k = 1, 2, \cdots, g, \quad j, j' = 1, 2, \cdots, p
\end{aligned}$$

ただし，$n = \sum_{k=1}^{g} n_k$ である．

b. マハラノビス汎距離の意味

マハラノビス汎距離は各群の母集団分布が正規分布 $N(\boldsymbol{\mu}^{(k)}, \Sigma)$ に従うとき，はっきりとした意味を持つ．そこで，以下では変量の二群判別問題でその意味を説明する．

[*2)] マハラノビス汎距離は，共分散行列が単位行列のとき，すなわち各変量が分散 1 で無相関のときユークリッド距離に帰着される．

図 C.5 は両群の母集団分布を図示したものであるが，ここでは両群の母集団分布はともに正規分布に従い，共分散行列も共通であるので，分布の形状，大きさは同じになる．

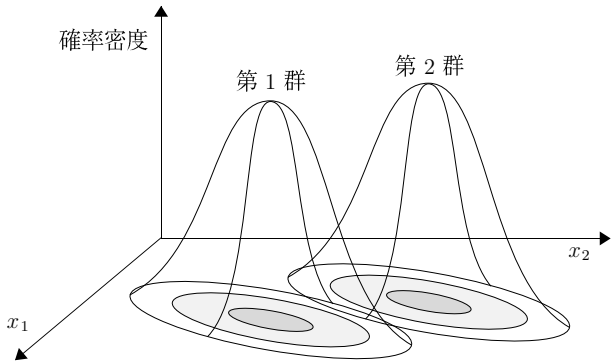

図 C.5 各グループの分布イメージ

未知のサンプル $\boldsymbol{x} = (x_1, x_2)$ が与えられたとき，$\boldsymbol{x}$ が第 1 群のサンプルである確率を $\Pr(1|\boldsymbol{x})$，第 2 群のサンプルである確率を $\Pr(2|\boldsymbol{x})$ とする．また，任意の点 $\boldsymbol{x}$ における両群の確率密度を $f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ とする．一般に，$\boldsymbol{x}$ の誤判定確率を最小にするには $\Pr(1|\boldsymbol{x})$, $\Pr(2|\boldsymbol{x})$ を比較して大きい方を選べばよいということが知られている．Bayes の公式を用いると，両群のデータが同じ確率で発生するときの $\Pr(1|\boldsymbol{x})$, $\Pr(2|\boldsymbol{x})$ は

$$\Pr(1|\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})}, \quad \Pr(2|\boldsymbol{x}) = \frac{f_1(\boldsymbol{x})}{f_1(\boldsymbol{x}) + f_2(\boldsymbol{x})}$$

となる．したがって，未知のサンプル $\boldsymbol{x}$ の判別を行うには，$f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ を比較すればよい．

図 C.5 を見ると，$f_1(\boldsymbol{x})$, $f_2(\boldsymbol{x})$ は点 $\boldsymbol{x}$ におけるそれぞれの山が高さになる．そこで，$\boldsymbol{x}$ を x_1–x_2 平面上に付置し，その点における山の高さを比較すると第 1 群と判別される．

一方，各群の確率密度関数は，

$$f_k(\boldsymbol{x}) = \frac{1}{2\pi \left|\Sigma^{(k)}\right|^{1/2}} \exp\left\{-\frac{1}{2} d^2_{(k)}\right\}, \quad k = 1, 2 \qquad \text{(C.23)}$$

と表わされる[*1)]．したがって，母集団が正規分布に従うのであれば，マハラノビス汎距離 $d^2_{(k)}$ を比較すれば山の高さを比較することができる．

地図などでは山の高さを表わすときに等高線が用いられる．それと同様に，図 C.5 において 2 つの山の等高線を x_1–x_2 平面 に表示したものが図 C.6 である．なお，母集団は正規分布に従うと仮定したので，等高線は楕円形になり，楕円の中心は各群の平均となる．

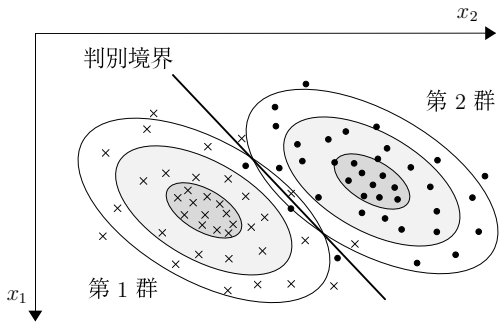

図 C.6　各群のサンプルの分布と判別境界

このようにマハラノビス汎距離による方法では，x_1–x_2 平面上で各群の分布に基づく等高線を考え，これを比較することにより，誤判別率を最小となるような判別を行っている．

C. 4. 4　多群判別分析

a. 多群判別のパラメータの推定

第 C. 4. 2 項の相関比による線形判別分析では，全体変動の偏差平方和 S_T は群平均の偏差平方和 S_B と群内変動の偏差平方和 S_W の和に等しいということを述べた．この性質は群の数が 2 以上になっても成り立つ．これを要素

[*1)] p 変量正規分布の確率密度関数が

$$f_k(\boldsymbol{x}) = \frac{1}{(2\pi)^{p/2}\,|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}\,(\boldsymbol{x}-\boldsymbol{\mu})^{\top}\,\Sigma^{-1}\,(x\boldsymbol{x}-\boldsymbol{\mu})\right\}, \quad k = 1, 2$$

であることに注意せよ．

にで表現すると，

$$\underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_i^{(k)}-\bar{z}\right)^2}_{\text{総平方和 } S_T\ \text{(全体変動の偏差平方和)}} = \underbrace{\sum_{k=1}^{g} n_k\left(\bar{z}^{(k)}-\bar{z}\right)^2}_{\text{群間平方和 } S_B\ \text{(群平均の偏差平方和)}} + \underbrace{\sum_{k=1}^{g}\sum_{i=1}^{n_k}\left(z_i^{(k)}-\bar{z}^{(k)}\right)^2}_{\text{群内平方和 } S_W\ \text{(群内変動の偏差平方和)}}$$

となる．ただし，k は群を表し，i はサンプルを表す．

2 群のときと同様に，正準判別分析においても相関比 $\eta^2 = S_B/S_T$ が最大になるとき，g 個の群がよく判別されていると考える．そこで，次の η^2 を最大にする $\alpha_1, \alpha_2, \cdots, \alpha_p$ を求めるという制約なし最大化問題を考える．

$$\max \quad S_B/S_T$$

ここで，求める係数ベクトルを $\boldsymbol{\alpha}$=$(\alpha_1, \alpha_2, \cdots, \alpha_p)^\top$ とし，以下のような行列 B, T を定義する．

$$B = (b_{jj'}), \quad b_{jj'} = \sum_{k=1}^{g} n_k(\bar{x}_j^{(k)} - \bar{x}_j)(\bar{x}_{j'}^{(k)} - \bar{x}_{j'}),$$
$$T = (t_{jj'}), \quad t_{jj'} = \sum_{k=1}^{g}\sum_{i=1}^{n_k}(x_{ji}^{(k)} - \bar{x}_j)(x_{j'i}^{(k)} - \bar{x}_{j'}),$$
$$j, j' = 1, 2, \cdots, p$$

このとき，上述の最適化問題は次の一般固有値問題に帰着される．

$$(B - \lambda T)\boldsymbol{\alpha} = \mathbf{0}$$

ここで，得られる固有値の数を r 個とすると，$r = \min(g-1, p)$ となることが知られている．さらに，$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_r \geq 0$ の中で，最大の固有値と対応する固有ベクトルが，それぞれ最大化された η^2 とそのときの係数ベクトル $\boldsymbol{\alpha}$ になる．

C. 5　因子分析

C. 5. 1　因子分析のパラメータ推定

直交因子の場合，α, f, e がモデル式の条件を満たすならば，観測データ x_1, x_2,$\cdots$, x_p の共分散行列 Σ は α, f, e によって以下のように表される．

$$\Sigma = AA^{\top} + D \tag{C.24}$$

ただし，D は $(d_1^2, d_2^2,\cdots, d_p^2)$ を対角成分とする対角行列であり，独自因子の分散である．

因子分析では (C.24) 式の関係をたよりに因子付加行列 A，独自因子の分散行列 D を求め，これを用いて f と e を推定する．以下では因子付加行列の代表的な推定法である主因子法について説明する．

C. 5. 2　主因子法

(C.24) 式における行列 $AA^{\top}$ の対角成分は

$$\underbrace{(a_{j1}^2 + a_{j2}^2 + \cdots + a_{jq}^2)}_{\text{共通性}} + \underbrace{d_j^2}_{\text{独自性}} = h_j^2 + d_j^2 = \sigma_{ii}, \quad j = 1, 2, \cdots, p$$

と表される．このとき，h_j^2 と d_j^2 はそれぞれ**共通性**，**独自性**と呼ばれる．ただし，σ_{ij} は共分散行列 Σ の第 $i \times j$ 番目の対角要素である．

ここで，(C.24) 式から，共分散行列の対角要素を共通性 h_j^2 で置き換えると，

$$\Sigma - D = \begin{bmatrix} h_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & h_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & h_p^2 \end{bmatrix}$$

となる．このとき，もし当初仮定した q 個の因子を持つモデルが妥当ならば，この行列は $AA^{\top}$ に分解できる．$AA^{\top}$ のランクは q なので，このとき，

$\Sigma - D$ の固有値は

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_q > \lambda_{q+1} = \cdots = \lambda_p$$

となり，q 個の正の固有値を持つ．したがって，λ_i に対応する固有ベクトルを $\boldsymbol{a}_i$ とすれば[*1)]，

$$\begin{aligned}\Sigma - D &= \sum_{i=1}^{q} \lambda_i \boldsymbol{a}_i \boldsymbol{a}_i^\top \\ &= \left[\sqrt{\lambda_1}\boldsymbol{a}_1, \sqrt{\lambda_2}\boldsymbol{a}_2, \cdots \sqrt{\lambda_q}\boldsymbol{a}_q\right]\left[\sqrt{\lambda_1}\boldsymbol{a}_1, \sqrt{\lambda_2}\boldsymbol{a}_2, \cdots \sqrt{\lambda_q}\boldsymbol{a}_q\right]^\top\end{aligned}$$

と分解することができる．(C.24) 式と対応をとると，

$$A = \left[\sqrt{\lambda_1}\boldsymbol{a}_1, \sqrt{\lambda_2}\boldsymbol{a}_2, \cdots \sqrt{\lambda_q}\boldsymbol{a}_q\right] \tag{C.25}$$

となり，因子負荷行列を求めることができる[*1)]．

この行列を推定するために用いられるの代表的な方法が主因子法である．主因子法では，共通性 h_j^2 の推定値として，変量 j と他の変量との重相関係数の 2 乗が用いられる[*2)]．そして，サンプルの相関係数行列の対角要素をこの h_j^2 の推定値で置き換え，固有値問題を解く．固有値の大きい方から q 個の固有値とその固有ベクトルから (C.25) 式により行列 A を求める．

この主因子法を一度解くだけでは当てはまりがよくない場合，得られる行列 A の要素 α_{ij} により $\sum_k \alpha_{jk}^2$ を j 番目の対角要素に代入し，再度問題を解く．そして，元のサンプルの相関係数行列の対角要素に値が十分近づくまで，この操作を繰り返す．この方法は反復主因子法と呼ばれ，因子分析のための多くのソフトウェアで用いられている[*3)]．

C. 5. 3 軸の回転

一般に (C.24) 式の関係から得られる因子負荷行列 A は一意に定まらないことが知られている．因子をうまく解釈するためには，少数の因子負荷量の

[*1)] ただし $||\boldsymbol{a}_i|| = 1$ と仮定する．

[*1)] ただし，得られる行列は単位に依存してしまうため，計算上は各変量の分散を 1，つまり相関係数行列を共分散行列の代わりに利用することが多い．以下でも相関係数行列を用いる．

[*2)] もしくは最大の相関係数が用いられる場合もある．

[*3)] 因子負荷量を求める方法としては，主因子法以外に最尤法や最小 2 乗法も用いられることがある．

絶対値が大きく，その他の変量の因子負荷量の絶対値が小さい方が望ましい．こういった構造は**単純構造**と呼ばれる．そこで，因子負荷行列を回転することにより，この単純構造を実現し，因子負荷行列を一意に定める．回転の方法は大きく分けて，軸同士が直交する（つまり無相関）の直交回転と，軸同士の相関関係を考慮した斜交回転がある．直交回転には，バリマックス回転やオーソマックス回転，斜交回転にはオブリミン回転，オブリマックス回転，プロマックス回転などがある．詳しくは専門書を参考いただきたい．

ここでは，**バリマックス回転**について詳述する．因子負荷行列 A を回転させたものを B とする．このとき適当な行列 R を用いて，

$$B = AT$$

と表すことができる．ここで回転後の因子負荷量 $B = [\beta_{jk}]_{p\times q}$ は前述した単純構造が望ましい．そのため，各因子の因子負荷量の 2 乗の分散を最大にするように回転する．つまり，全因子の分散，

$$V_\beta = \sum_{k=1}^{q}\left\{\sum_{j=1}^{p}\beta_{jk}^4 - \frac{1}{p}\left(\sum_{j=1}^{p}\beta_{jk}^2\right)^2\right\}$$

を最大にするようにもとの因子を回転させる．すべての因子を同時に回転させるのではなく，2 つの因子を取り出しその 2 つの軸で構成される平面上で軸を回転させる．その回転を任意の因子の組み合わせで繰り返し，収束するまで続ける．2 つの因子 k, ℓ を，

$$\begin{aligned}\beta_{jk} &= \alpha_{jk}\cos\theta_{k\ell} + \alpha_{j\ell}\sin\theta_{k\ell}\\ \beta_{j\ell} &= -\alpha_{jk}\sin\theta_{k\ell} + \alpha_{j\ell}\cos\theta_{k\ell}\end{aligned}$$

と回転する．回転角度 $\theta_{k\ell}$ は，

$$\theta_{k\ell} = \frac{d - 2\frac{ab}{p}}{c - \frac{a^2-b^2}{p}}$$

である．ここで，$\theta_{k\ell}$ は $(d-2ab/p)\sin 4\theta_{k\ell} > 0$ を満たし，$a = \sum_j(\alpha_{jk}-\alpha_{j\ell})^2$, $b = 2\sum_j \alpha_{jk}\alpha_{j\ell}$, $c = \sum_j\{(\alpha_{jk}^2 - \alpha_{j\ell}^2)^2 - 4\alpha_{jk}^2\alpha_{j\ell}^2\}$, $d = 4\sum_j \alpha_{jk}\alpha_{j\ell}(\alpha_{jk}^2 -$

$\alpha_{j\ell}^2$) である．ただし，この方法で得られる回転後の因子負荷量は共通性の大きさを考慮していない．その問題を回避するために，β_{jk}^2 の分散ではなく共通性で除した β_{jk}^2/h_j^2 の分散を最大にするという方法がとられる．この場合は，上記 a から d の α_{jk}, $\alpha_{j\ell}$ をそれぞれ α_{jk}/h_j, $\alpha_{j\ell}/h_j$ で置き換えることによって求めることができる．この方法を**規準化バリマックス法**という．

C.5.4 因子得点の推定

因子負荷量が適切に推定されると，次に各サンプルの各因子における得点を推定する．因子得点は，因子負荷行列を元に推定されるが，行列のランクの性質により，一意に定まらないという問題を含んでいる．このため，因子得点の推定には重み付最小 2 乗法や回帰推定などのいくつかの方法があるが，因子得点の推定については本書では割愛する．

C.6 主成分分析

C.6.1 集約指標の考え方

2 変量の場合を例にとって集約についての考え方を説明する．図 C.7 は，x_1–x_2 平面に複数のサンプルが分布している様子を表している．

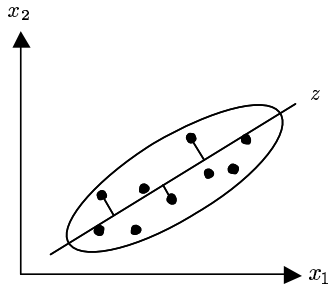

図 C.7 サンプル分布と集約指標

この図を見ると変量 x_1 と変量 x_2 の間には強い正の相関があることがわかる．そこで，データ全体を囲むような楕円を描き，長軸方向に新しく z 軸

を引く．そして，z 軸に各サンプルを射影する．z 軸はサンプル間の差をもっとも大きく表現するものになっているので，z 軸上の値を比較することで，1 つの変量でサンプルのバラツキの様子を大まかに把握することができる．

主成分分析では，このように相関の強い変量を合成することによって，新しい変量 (集約指標) の軸を生成する．このとき，バラツキに関する情報をできるだけ失うことなく軸を決めるために，各サンプルを射影したときの分散を最大するという基準で z 軸を定めることを考える．一つの軸では元のデータのバラツキの様子を十分に表せていないと判断される場合には，この軸と無相関な軸[*1)]を新たに引くことで，別の見方をすることができる．

一般に，観測データが p 変量の場合に r 個 $(r \leq p)$ の軸を引き，集約指標を生成する．観測データの各変量を $x_1, x_2, \cdots, x_p$，新しく生成した変量を $z_1, z_2, \cdots, z_r$ としたとき，$x_1, x_2, \cdots, x_p$ のバラツキに関する情報をできるだけ失うことなく z_1 を生成し，残りの情報をできるだけ失うことなく z_2 を生成するというように繰り返し，適当な z_r まで変量を生成する．このとき，z_1 軸，z_2 軸，$\cdots$，z_r 軸は互いに垂直であるものとする．

主成分分析では，新しく生成する z 軸は最大 p 個生成することができる．したがって，直交座標系 $x_1, x_2, \cdots, x_p$ を回転して，新しい直交座標系 $z_1, \cdots, z_r, \cdots, z_p$ を作り，そこから第 r 番目までの軸を採用していると考えることもできる．

C.6.2　主成分分析の係数推定

新たに生成される軸 z_i は，サンプルの観測値の各項の線形和で与えられる．第 1 主成分 z_1 の分散 $V(z_1)$ は

$$V(z_1) = \frac{1}{n-1}\sum_{i=1}^{n}(z_{1i} - \bar{z}_1)^2 = \sum_{j=1}^{p}\sum_{k=1}^{p}\sigma_{jk}\beta_{1j}\beta_{1k} = \boldsymbol{\beta}_1^{\top}\Sigma\boldsymbol{\beta}_1 \quad \text{(C.26)}$$

である．ただし，$\boldsymbol{\beta}_1$ は第 1 主成分の係数ベクトルである．

係数ベクトルの大きさが 1 であるとすると，第 1 主成分の係数ベクトルを

[*1)] 2 次元データの場合は z 軸に垂直な軸が相当する．

求める問題は次の等号制約付き最適化問題となる．

$$\max \quad \boldsymbol{\beta}_1^\top \Sigma \boldsymbol{\beta}_1 \qquad \text{s.t.} \quad \boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 = 1$$

同様にして，第 2 主成分以降の係数ベクトルを求める問題は以下のようになる．

$$\begin{aligned}
&\max \quad \boldsymbol{\beta}_2^\top \Sigma \boldsymbol{\beta}_2 \qquad \text{s.t.} \quad \boldsymbol{\beta}_2^\top \boldsymbol{\beta}_2 = 1, \quad \boldsymbol{\beta}_2^\top \boldsymbol{\beta}_1 = 0 \\
&\max \quad \boldsymbol{\beta}_3^\top \Sigma \boldsymbol{\beta}_3 \qquad \text{s.t.} \quad \boldsymbol{\beta}_3^\top \boldsymbol{\beta}_3 = 1, \quad \boldsymbol{\beta}_3^\top \boldsymbol{\beta}_1 = 0, \quad \boldsymbol{\beta}_3^\top \boldsymbol{\beta}_2 = 0 \\
&\quad \cdots
\end{aligned}$$

ただし，第 2 主成分以降はそれまで求めた主成分と直交するように，すなわち $\boldsymbol{\beta}_2^\top \boldsymbol{\beta}_1 = 0$ という制約条件が加得られる．

主成分分析では，これらの等号制約付き最適化問題が，以下の固有値問題に帰着される．

$$(\Sigma - \lambda I)\boldsymbol{\beta} = \mathbf{0}$$

ここで，得られた固有値 $\lambda_1 \geq \cdots \geq \lambda_r \geq \cdots \geq \lambda_p \geq 0$ の中で，大きいほうから順に r 個の固有値と対応する固有ベクトルが第 1 主成分，第 2 主成分，$\cdots$，第 r 主成分の分散と係数ベクトルになることが知られている．

なお，固有値問題の詳細については付録 B.5 を参照されたい．

(C.26) 式はラグランジュの未定乗数法により，

$$L(\boldsymbol{\beta}, \lambda) = \boldsymbol{\beta}_1^\top \Sigma \boldsymbol{\beta}_1 + \lambda(\boldsymbol{\beta}_1^\top \boldsymbol{\beta}_1 - 1)$$

となる．そこで，$\boldsymbol{\beta}_1$ で偏微分して 0 とおけば，

$$\frac{1}{2}\frac{\partial L(\boldsymbol{\beta}, \lambda)}{\partial \boldsymbol{\beta}_1} = (\Sigma - \lambda I) = \mathbf{0}$$

$\boldsymbol{z}_1$ の分散は次のようにして求められる．

$$\left(\boldsymbol{z}_1 - \frac{1}{n} I \boldsymbol{z}_1\right)^\top \left(\boldsymbol{z}_1 - \frac{1}{n} I \boldsymbol{z}_1\right)$$

$$
\begin{aligned}
&= \left\{ \left(I - \frac{1}{n} J \right) X \boldsymbol{\ell}_1 \right\}^\top \left\{ \left(I - \frac{1}{n} J \right) X \boldsymbol{\ell}_1 \right\} \\
&= \boldsymbol{\ell}_1^\top X^\top \left(I - \frac{1}{n} J \right)^\top \left(I - \frac{1}{n} J \right) X \boldsymbol{\ell}_1 \\
&= \boldsymbol{\ell}_1^\top \left\{ X^\top \left(I - \frac{1}{n} J \right) X \right\} \boldsymbol{\ell}_1 \qquad \text{(C.27)}
\end{aligned}
$$

C.7　数量化 I 類

一般的にデータ形式を記述すると表 C.6 のようになる．

表 C.6　数量化 I 類のデータ例

被説明変数	説明変数								
	第 1 アイテム				⋯	第 m アイテム			
	1	2	⋯	c_1	⋯	1	2	⋯	c_m
y_1	$\delta_{1,11}$	$\delta_{1,12}$	⋯	$\delta_{1,1c_1}$	⋯	$\delta_{1,m1}$	$\delta_{1,m2}$	⋯	δ_{1,mc_m}
y_2	$\delta_{2,11}$	$\delta_{2,12}$	⋯	$\delta_{2,1c_1}$	⋯	$\delta_{2,m1}$	$\delta_{2,m2}$	⋯	δ_{2,mc_m}
y_3	$\delta_{3,11}$	$\delta_{3,12}$	⋯	$\delta_{3,1c_1}$	⋯	$\delta_{3,m1}$	$\delta_{3,m2}$	⋯	δ_{3,mc_m}
⋮	⋮	⋮	⋱	⋮	⋱	⋮	⋮	⋱	⋮
y_n	$\delta_{n,11}$	$\delta_{n,12}$	⋯	$\delta_{n,1c_1}$	⋯	$\delta_{n,m1}$	$\delta_{n,m2}$	⋯	δ_{n,mc_m}
パラメータ	α_{11}	α_{12}	⋯	α_{1c_1}	⋯	α_{m1}	α_{m2}	⋯	α_{mc_m}

このとき，サンプル i $(i = 1, 2, \cdots, n)$ について説明変数を $x_{i,jk}$，目的変数 y_i とする．ここで，$x_{i,jk}$ はサンプル i がカテゴリ j のアイテム k にあてはまるとき 1 となり，そうでないとき 0 となる．このとき，数量化 I 類のモデル式は以下のようになる．

$$
y_i = \sum_{j=1}^{m} \sum_{k=1}^{c_j} \alpha_{jk} x_{i,jk} + \varepsilon_i
$$

このモデル式に従って表 C.6 のデータを記述すると以下のようになる．

$$
\boldsymbol{y} = \begin{pmatrix} X_1 & X_2 & \cdots & X_m \end{pmatrix} \begin{pmatrix} \boldsymbol{\alpha}_1 \\ \boldsymbol{\alpha}_2 \\ \vdots \\ \boldsymbol{\alpha}_m \end{pmatrix} + \boldsymbol{\varepsilon} \qquad \text{(C.28)}
$$

$$= \sum_{j=1}^{m} X_j \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon} \tag{C.29}$$

$$= X\boldsymbol{\alpha} + \boldsymbol{\varepsilon} \tag{C.30}$$

ただし，$\boldsymbol{y} = (y_1, y_2, \cdots, y_n)^\top$，$X_j = (\boldsymbol{x}_{j1}, \boldsymbol{x}_{j2}, \cdots, \boldsymbol{x}_{jc_j})$，$\boldsymbol{x}_{jk} = (x_{1,jk}, x_{2,jk}, \cdots, x_{n,jk})^\top$，$\boldsymbol{\alpha}_j = (\alpha_{j1}, \alpha_{j2}, \cdots, \alpha_{jc_j})^\top$

ここで，重回帰分析と同じように誤差の二乗和が最小になるパラメータ α_{jk} を定める．

$$Q = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}^\top \boldsymbol{\varepsilon} = \left(\boldsymbol{y} - \sum_{j=1}^{m} X_j \boldsymbol{\alpha}_j \right)^\top \left(\boldsymbol{y} - \sum_{j=1}^{m} X_j \boldsymbol{\alpha}_j \right) \to \min \tag{C.31}$$

重回帰分析に従うと (C.31) 式を $\boldsymbol{\alpha}_j$ で偏微分して 0 とおくと以下のようになる．

$$\frac{\partial Q}{\partial \boldsymbol{\alpha}} = 2X^\top X\boldsymbol{\alpha} - 2\boldsymbol{y}^\top X + 2\alpha_0 \mathbf{1}^\top X = 0 \tag{C.32}$$

(C.8) 式より，α_0 は以下のように容易に求められる．

$$X^\top X\boldsymbol{\alpha} - \boldsymbol{y}^\top X + (\bar{y} - \bar{\boldsymbol{x}}^\top \boldsymbol{\alpha})\mathbf{1}^\top X = \mathbf{0}$$

正規方程式を求めると以下のようになる．

$$(X - \mathbf{1}\bar{\boldsymbol{x}}^\top)^\top (X - \mathbf{1}\bar{\boldsymbol{x}}^\top)\boldsymbol{\alpha} = (X - \mathbf{1}\bar{\boldsymbol{x}}^\top)^\top (\boldsymbol{y} - \mathbf{1}\bar{y}) \tag{C.33}$$

しかし，正規方程式における行列の階数は $\sum_{j=1}^{m}(c_j - 1) + 1$ なので，逆行列が存在しない．すなわち，$\boldsymbol{\alpha}_j$ は一意的には定まらない．そこで，解を一意に定めるために $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \boldsymbol{\alpha}_m$ の 1 番目の要素を 0 として，新たに $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \boldsymbol{\beta}_m$ を考える．ただし，$\boldsymbol{\beta}_j = (0, \alpha_{j2}, \cdots, \alpha_{jc_j})^\top$ である．

このようにして $\boldsymbol{\beta}_j$ を求めてから各アイテム内のカテゴリ数量を固体数で重み付けした平均が 0 になるように調整する．

$$\boldsymbol{y} = \alpha_0 + \sum_{j=1}^{m} X_j \boldsymbol{\alpha}_j + \boldsymbol{\varepsilon}$$

C.7.1　カテゴリ数量の推定

数量化 I 類では，重回帰分析と同じように，誤差の 2 乗和が最小になるカテゴリ数量の推定値 $\hat{\alpha}_0, \hat{\alpha}_{jk}$ を定める．

$$Q \equiv \sum_{i=1}^{n} \left\{ y_i - \left(\hat{\alpha}_0 + \sum_{j=1}^{m} \sum_{k=1}^{c_j} \hat{\alpha}_{jk} \delta_{i,jk} \right) \right\}^2 \to \min$$

このとき注意しなければならないのは，各アイテム内でカテゴリのダミー変数のどれかが必ず 1 になるということである．したがって，同一アイテム内のダミー変数は従属の関係にあり，$\hat{\alpha}_0, \hat{\alpha}_{jk}$ を一意に定めることはできない[*1)]．

そこで，数量化 I 類では次の手順でカテゴリ数量の推定値を一意に定める．

(1) 各アイテムのどれか 1 つのカテゴリ (例えば，1 番目のカテゴリ) を削除して，定数項を含む重回帰分析を行う．

(2) カテゴリ数量に定数を加減し，カテゴリ数量を調整する．また，重回帰分析の定数項も含めて定数部分をまとめることにより新たな定数項とする．

ここで，(1) の操作は削除されたカテゴリの数量を 0 にすることと同じである．したがって，得られるカテゴリ数量 $\hat{\alpha}_{jk}$ は，削除したカテゴリをベースに考えたとき，アイテム内での条件の変化が予測値の増減に与える影響度を示している．

カテゴリ数量の解釈をする上で特定のカテゴリ数量を 0 にすることが不都

[*1)] 具体的には，本文中のアイテム「天気」に着目すると (晴れ) + (曇り) + (雨) = 1 なので，(雨) というダミー変数を消去しても残りのダミー変数から求めることができる．したがって，任意の値 γ に対して以下のような関係式が成り立つ．

$$\begin{aligned} &10.81 \times (\text{晴れ}) + (-8.33) \times (\text{曇り}) + (0.53) \times (\text{雨}) \\ &= (10.81 + \gamma) \times (\text{晴れ}) + (-8.33 + \gamma) \times (\text{曇り}) + (0.53 + \gamma) \times (\text{雨}) - \gamma \end{aligned}$$

これより，同一アイテム内の各カテゴリ数量に同じ値を加え，最後の定数項として同じ値を引く操作を行うことでカテゴリ数量を調整しても理論値は変化しないことがわかる．また，特に $\gamma = -0.53$ とすると以下のように (雨) というダミー変数を消去することができる．

$$(10.28 + \gamma) \times (\text{晴れ}) + (-8.86 + \gamma) \times (\text{曇り}) - \gamma$$

合な場合，(2) の操作でカテゴリ数量を調整する．このとき，各アイテムのカテゴリ数量を固体数で重み付けした平均について，以下の関係が成り立つことが知られている．

$$\bar{y} = \hat{\alpha}_0 + \frac{1}{n}\sum_{j=1}^{m}\sum_{\ell=1}^{c_j} n_{j\ell}\hat{\alpha}_{j\ell}$$

そこで，(1) で得られたカテゴリ数量 $\hat{\alpha}_{jk}$ を調整して $\hat{\beta}_{jk}$ とおくと，

$$\hat{\beta}_{jk} = \hat{\alpha}_{jk} - \frac{1}{n}\sum_{\ell=1}^{c_j} n_{j\ell}\hat{\alpha}_{j\ell}$$

これに応じて $\hat{\alpha}_0$ は $\bar{y}$ と置き換えられるので，モデル式は以下のようになる．

$$y_i = \bar{y} + \sum_{j=1}^{m}\sum_{k=1}^{c_j} \hat{\beta}_{jk}\delta_{i,jk} + \boldsymbol{\varepsilon}_i$$

このモデル式では，y の平均を基準として，アイテム内でのカテゴリにより y の増減が決まり，これらを合わせたものが y の理論値となっている．

C. 8 数量化 II 類

表 C.7 は，数量化 II 類のデータ例である．数量化本文の例では，群ごとではなく，サンプルごとにデータが並んでいるが，ここでは群ごとに反応するものをまとめている．

数量化 II 類でも，数量化 I 類と同様にダミー変数を用いて説明変数を質的データから数値データに変換する．つまり，各群に対して列を一つ当てはめ，その列に該当する群に反応しているならば 1，そうでなければ 0 とするダミー変数により表現する．

次にこのデータに対して多群の線形判別分析と同じ操作を施す．前節で説明したように多群の線形判別分析は実際には正準相関分析と計算手順で行う．しかし，ここでも数量化 I 類と同様に各アイテム変量については行列のランクの問題があるため，$\boldsymbol{\alpha}_j$ は一意的に定めることはできない．そこで，今度

表 C.7　数量化 II 類のデータ例

群		アイテム 1			⋯	アイテム m		
		1	⋯	c_1		1	⋯	c_i
1	1	$\delta_{11,11}$		$\delta_{11,1c_1}$		$\delta_{11,m1}$		δ_{11,mc_m}
	2	$\delta_{12,11}$		$\delta_{12,1c_1}$		$\delta_{12,m1}$		δ_{12,mc_m}
	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮
	n_1	$\delta_{1n_1,11}$		$\delta_{1n_1,1c_1}$		$\delta_{1n_1,m1}$		δ_{1n_1,mc_m}
2	1	$\delta_{21,11}$		$\delta_{21,1c_1}$		$\delta_{21,m1}$		δ_{11,mc_m}
	2	$\delta_{22,11}$		$\delta_{22,1c_1}$		$\delta_{22,m1}$		δ_{12,mc_m}
	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮
	n_2	$\delta_{2n_2,11}$		$\delta_{2n_2,1c_1}$		$\delta_{2n_2,m1}$		δ_{1n_1,mc_m}
⋮	⋮	⋮		⋮		⋮		⋮
g	1	$\delta_{g1,11}$		$\delta_{g1,1c_1}$		$\delta_{g1,m1}$		δ_{g1,mc_m}
	2	$\delta_{g2,11}$		$\delta_{g2,1c_1}$		$\delta_{g2,m1}$		δ_{g2,mc_m}
	⋮	⋮	⋯	⋮	⋯	⋮	⋯	⋮
	n_g	$\delta_{gn_g,11}$		$\delta_{gn_g,1c_1}$		$\delta_{gn_g,m1}$		δ_{gn_g,mc_m}

は $\boldsymbol{\alpha}_1, \boldsymbol{\alpha}_2, \cdots, \alpha_m$ の 1 番目の要素を 0 として，新たに $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2, \cdots, \beta_m$ を考える．ただし，$\boldsymbol{\beta}_j = (0, \alpha_{j2}, \cdots, \alpha_{jc_j})^\top$ である．

このようにして $\boldsymbol{\beta}_j$ を求めてから，各アイテム内のカテゴリ数量を固体数で重み付けした平均が 0 になるように調整する．

したがって，数量化 II 類では，正準判別分析と同様に，合成変量 z についての相関比が最大になるカテゴリ数量を定める．k 群に属するサンプルを

$$z_i^{(k)} = \sum_{j=1}^{m} \sum_{\ell=1}^{c_\ell - 1} \beta_{j\ell} \delta_{ki,j\ell}$$

として表すと[*1)]，$z_i^{(k)}$ の総平方和は次のように表すことができる．

$$\underbrace{\sum_{k=1}^{g} \sum_{i=1}^{n_k} (z_i^{(k)} - \bar{z})^2}_{\text{総平方和 } S_T} - \underbrace{\sum_{k=1}^{g} n_k (\overline{z^{(k)}} - \bar{z})^2}_{\text{群間平方和 } S_B} + \underbrace{\sum_{k=1}^{g} \sum_{i=1}^{n_k} (z_i^{(k)} - \overline{z^{(k)}})^2}_{\text{群内平方和 } S_W}$$

ただし $\overline{z^{(k)}}$ は k 群の合成変量の平均，n_k は k 群のサンプル数である．この

[*1)] 各アイテムからカテゴリを一つ削除していることに注意.

関係より，相関比を最大にすることは，

$$\max \quad S_B/S_T$$

を解けばよい．この問題の解は，正準相関分析と同様に，一般固有値問題を解くことによって求められる[*1)]．

カテゴリ数量の解釈をする上で特定のカテゴリ数量を 0 とすることが不都合な場合，数量化 I 類と同様の手順で，カテゴリ数量を調整する．あるカテゴリ変数を 0 とおき，上記問題の解として得られるカテゴリ数量 $\beta_{j\ell}$ を次のように調整し，元の問題のカテゴリ数量 $\alpha_{j\ell}$ を求める．

$$\alpha_{j\ell} = \beta_{j\ell} - \frac{1}{n}\sum_{k=1}^{g}\sum_{\ell=1}^{c_j-1} n_{c_k}\beta_{j\ell}$$

C. 9 数量化 III 類

C. 9. 1 サンプル・スコア，カテゴリ・スコアの推定

サンプル・スコアとカテゴリ・スコアを推定するにあたり，図??のデータ形式を表 C.8 のように変換する．

表 C.8 について，サンプルとカテゴリの関係をもっともよく表すために，これらの間の相関係数を最大にすることを考える．相関係数は次のように与えられる．

$$r = \frac{(1/N)\sum_{i=1}^{n}\sum_{j=1}^{m}\{(x_i-\bar{x})\delta_{ij}\}\{(y_j-\bar{y})\delta_{ij}\}}{\sqrt{\sum_{i=1}^{n}\sum_{j=1}^{m}\{(1/N)(x_i-\bar{x})^2\delta_{ij}\}\{(1/N)(y_j-\bar{y})^2\delta_{ij}\}}} \qquad \text{(C.34)}$$

ここで，各サンプルと各カテゴリの平均を 0，分散を 1 という制約条件をおくと，(C.34) 式の最大化問題は，

$$\max \quad r = \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \delta_{ij}$$

[*1)] 第 1 固有値，固有ベクトルのみでうまく解釈できない場合は，第 2 固有値，第 3 固有値... を用いて解釈する．

表 C.8　数量化 III 類のデータ形式 (2)

		サンプル				カテゴリ			
サンプル	カテゴリ	1	2	$\cdots$	n	1	2	$\cdots$	m
1	1	δ_{11}	0	$\cdots$	0	δ_{11}	0	$\cdots$	0
1	2	δ_{12}	0	$\cdots$	0	0	δ_{12}	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
1	m	δ_{1m}	0	$\cdots$	0	0	0	$\cdots$	δ_{1m}
2	1	0	δ_{21}	$\cdots$	0	δ_{21}	0	$\cdots$	0
2	2	0	δ_{22}	$\cdots$	0	0	δ_{22}	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
2	m	0	δ_{2m}	$\cdots$	0	0	0	$\cdots$	δ_{2m}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	1	0	0	$\cdots$	δ_{n1}	δ_{n1}	0	$\cdots$	0
n	2	0	0	$\cdots$	δ_{n2}	0	δ_{n2}	$\cdots$	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
n	m	0	0	$\cdots$	δ_{nm}	0	0	$\cdots$	δ_{nm}
		サンプル				カテゴリ			
付与する値		x_1	x_2	$\cdots$	x_n	y_1	y_2	$\cdots$	y_m

$$\text{s.t.}\quad \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} x_i\delta_{ij} = 0, \qquad \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}(x_i\delta_{ij})^2 = 1$$
$$\frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m} y_j\delta_{ij} = 0, \qquad \frac{1}{N}\sum_{i=1}^{n}\sum_{j=1}^{m}(y_j\delta_{ij})^2 = 1$$

を解けばよいことになる[*1)].

この問題を解くために，目的関数にサンプルとカテゴリの分散の制約をラグランジュ乗数を乗じてラグランジュ関数を考えると，最適性の条件は，

$$\sum_{j=1}^{m}\frac{\delta_{ij}}{\sum_{j=1}^{m}\delta_{ij}}y_j - \lambda x_i = 0, \qquad i = 1, 2, \cdots, n$$
$$\sum_{i=1}^{n}\frac{\delta_{ij}}{\sum_{i=1}^{n}\delta_{ij}}x_i - \lambda y_j = 0, \qquad j = 1, 2, \cdots, m$$

となる．このとき，ラグランジュ乗数 λ は相関係数 r と等しくなる．これら

[*1)] 本書の最適化問題の分母は 1 となるため，このように簡略に記述することができる

より，x_i もしくは y_j を消去すると，

$$\sum_{k=1}^{m}\left(\sum_{i=1}^{n}\frac{a_i\sqrt{b_jb_k}}{\delta_{ij}\delta_{ik}}\right)\sqrt{b_k}y_k-\lambda^2\sqrt{b_j}y_j=0,\quad j=1,2,\cdots,m$$
$$\sum_{\ell=1}^{n}\left(\sum_{j=1}^{m}\frac{b_j\sqrt{a_ia_\ell}}{\delta_{ij}\delta_{\ell j}}\right)\sqrt{a_\ell}x_\ell-\lambda^2\sqrt{a_i}x_i=0,\quad i=1,2,\cdots,n$$

という共通の固有値 λ^2 を持つ固有値問題となる．ただし，$a_i=\sum_{j=1}^{m}\delta_{ij}$，$b_j=\sum_{i=1}^{n}\delta_{ij}$ である．この固有値問題の第 1 固有値はいずれも 1 となり，このとき対応する固有ベクトルはサンプル，カテゴリについてそれぞれ $(\sqrt{b_1},\sqrt{b_2},\cdots,\sqrt{b_m})^\top$，$(\sqrt{a_1},\sqrt{a_2},\cdots,\sqrt{a_n})^\top$ となる．しかしこの解はサンプル，カテゴリにそれぞれ同じ数量を与えることになり，条件式を満たさない．したがって，1 以外の最大固有値（第 2 固有値）λ^2 に対する固有ベクトル $(c_1,c_2,\cdots,c_m)$，$(d_1,d_2,\cdots,d_n)$ を求め，

$$x_i=\sqrt{N}\frac{d_i}{\sqrt{a_i}},\quad i=1,2,\cdots,n$$
$$y_j=\sqrt{N}\frac{c_j}{\sqrt{b_j}},\quad j=1,2,\cdots,m$$

としてサンプルとカテゴリに与える数量を求めることができる．もし，一つの軸（第 2 固有値）のみで解釈を十分に行うことができないならば，第 3 固有値以降を用いて，同様にサンプルとカテゴリに対する数量を求めていく．

ここでは，サンプルの数量とカテゴリの数量の相関を最大にするように定式化し，固有値問題に帰着させたが，各サンプルを一つの層と考えて，相関比を最大にするという基準 n より定式化しても同じ数量が求まる．このとき，相関比 η^2 は λ^2 と等しくなる．

C. 10 主座標分析

複数の対象について，対となる対象間の類似度が与えられたときに，その距離に応じて対象を配置する手法に主座標分析がある．主座標分析は類似度行列を元にして，対象の 2 次元配置を求めることを目的する．したがって，

類似度が高い場合はより近くに，また類似度が低い場合にはより遠くに配置される．今，n 個の対象間の距離行列 (非類似度行列) が次のように与えられているとしよう．一般には，対象同士が似ていないほど大きな負の値をとるとする．ただし，類似度行列はすべての要素が非正で対角要素 s_{ii} は 0 とし，さらに対称行列，つまり $d_{ij} = d_{ji}$ とする．

$$S = \begin{pmatrix} s_{11} & \cdots & s_{1j} & \cdots & s_{1n} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{i1} & \cdots & s_{ij} & \cdots & s_{in} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ s_{n1} & \cdots & s_{nj} & \cdots & s_{nn} \end{pmatrix}$$

このとき行列 S の固有値問題を考え，その固有値を大きい順に $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq 0$ とし，固有値 λ_i に対応する．固有ベクトルを $\boldsymbol{c}_i = (c_{1i}, c_{2i}, \cdots, c_{ni})^\top$ とする．ここで，固有ベクトルを $\boldsymbol{c}_i^\top \boldsymbol{c}_i = \lambda_i$ となるように調整すると，距離行列 S は

$$S = \boldsymbol{c}_i \boldsymbol{c}_i^\top$$

のようにスペクトル分解することができる．ここで，対象 i の座標を各固有ベクトルの第 i 要素を固有値の大きい順に並べた $(c_{i1}, c_{i2}, \cdots, c_{in})$ とすると，対象 i と対象 j 間のユークリッド距離の 2 乗 d_{ij}^2 と類似度行列 S の (i, j) 要素 s_{ij} の関係は

$$d_{ij}^2 = \sum_{k=1}^{n} (c_{ik} - c_{jk})^2 = -2s_{ij} \tag{C.35}$$

となる．したがって，(C.35) 式より，$s_{ij} = -d_{ij}^2/2$ という関係があることがわかる．

このように，固有ベクトルにより類似度行列 S の要素間関係を表現できるが，各対象の座標が固有値の順であることから，前の座標の方が後ろの座標に比べて行列 S に関する情報を多く含んでいる．$\boldsymbol{c}_i$ の大きさは λ_i により決まるため，λ_i が小さくなるにつれ $\boldsymbol{c}_i$ の大きさも小さくなる．もし，$m + 1$

番目の固有値 λ_{m+1} 以降がほぼ 0 であるならば，$\boldsymbol{c}_{m+1}$ の大きさもほぼ 0 になる．したがって，このとき $m+1$ 番目以降の座標を無視し m 番目までの座標だけを利用すると，対象間の距離 d_{ij}^2 について

$$-2s_{ij} = d_{ij}^2 \approx \sum_{k=1}^{m} (c_{ik} - c_{jk})^2$$

という近似式が得られ，あまり情報が失われることなく，n 次元類似度行列の要素間関係をより少ない m 次元で表現することができる．次元を減らしたときの説明力は主成分分析と同様，第 m 固有値までの累積寄与率を求めればよい．

C. 11 多項ロジット・モデル

C. 11. 1 魅力型モデル

ある選択対象 $i \in \mathcal{N}$ の魅力度を $A_i > 0$ としたとき，選択対象 i の選択確率が

$$\Pr(i; \mathcal{N}) = \frac{A_i}{\sum_{k \in \mathcal{N}} A_k}, \quad i \in \mathcal{N}$$

となるモデルは**魅力型モデル** (atraction model) と呼ばれ，以下の 4 つの公理を満たす確率的選択モデルと同値であることが知られている．

公理 C.1. 選択対象の集合を $\mathcal{N}$ とし，その部分集合を $\mathcal{N}'$ とする．すべての選択対象 $i \in \mathcal{N}$ に対して非負の魅力度 A_i が定義される． □

公理 C.2. 魅力度 A_i は有限であり，少なくとも $\mathcal{N}$ の 1 つの要素について 0 でない． □

公理 C.3. 任意の部分集合 $\mathcal{N}' \subset \mathcal{N}$ の魅力度は，それに含まれる要素の魅力度の和に等しい． □

公理 C.4. $\mathcal{N}$ の 2 つの部分集合 $\mathcal{N}'_1$ と $\mathcal{N}'_2$ が同じ魅力度を持つとき，その選択確率は等しい．□

魅力型モデルの持つ特徴的な性質としては，以下に示す Luce(1959) の個人選択公理を満たすことが挙げられる．

公理 C.5 (個人選択公理) 選択対象の集合を $\mathcal{N}$ とし，その部分集合を $\mathcal{N}'$ とする．選択対象 $i \in \mathcal{N}'$ を $\mathcal{N}$ から選択する確率は，$\mathcal{N}'$ の選択とは無関係に

$$\Pr(i \,|\, i \in \mathcal{N}) = \Pr(i \,|\, i \in \mathcal{N}') \Pr(\mathcal{N}' \,|\, \mathcal{N}' \subset \mathcal{N})$$

が成立する．□

また，Luce の個人選択公理は以下に示す「無関係な代替案からの独立 (Independence from Irrelevant Alternatives : I.I.A)」という特性を持つ．

定理 C.1 (無関係な代替案からの独立) 任意の 2 つの代替案 $(i, j \in \mathcal{N}')$ のそれぞれが選択される確率の比はそれら以外の第三の代替案の存在いかんに依存しない．□

なお，魅力型モデルの代表的なものとしては McFadden(1974) による「多項ロジット・モデル (MultiNomial Logit Model: MNL モデル)」，Nakanishi and Cooper(1974, 1988b) による「積乗型競合相互作用モデル (Multiplicative Competitive Interaction Model: MCI モデル」が挙げられる．

C.11.2　多項ロジット・モデルによる選択確率

図 4.5 のようなデータを一般的な形式で記述すると表 C.9 のようになる．なお，目的変数についてはダミー変数を用いて表現していることに注意されたい．

個人 $i(i = 1, 2, \cdots, n)$ における選択肢 $k(k = 1, 2, \cdots, \ell)$ に対する選好度を $U_k^{(i)}$ とする．選好度 $U_k^{(i)}$ はモデルによって説明される確定的選好度 $V_k^{(i)}$ と

表 C.9 多項ロジット・モデルのデータ例 (3)

No.	目的変数					説明変数						
	選択結果	ダミー変数				選択肢 1			選択肢 2			···
		選択肢				特性			特性			···
		1	2	···		1	2	···	1	2	···	···
1	1	1	0	···		$x_{11}^{(1)}$	$x_{12}^{(1)}$	···	$x_{21}^{(1)}$	$x_{22}^{(1)}$	···	···
2	2	0	1	···		$x_{11}^{(2)}$	$x_{12}^{(2)}$	···	$x_{21}^{(2)}$	$x_{22}^{(2)}$	···	···
3	4	0	0	···		$x_{11}^{(3)}$	$x_{12}^{(3)}$	···	$x_{21}^{(3)}$	$x_{22}^{(3)}$	···	···
4	2	0	1	···		$x_{11}^{(4)}$	$x_{12}^{(4)}$	···	$x_{21}^{(4)}$	$x_{22}^{(4)}$	···	···
5	3	0	0	···		$x_{11}^{(5)}$	$x_{12}^{(5)}$	···	$x_{21}^{(5)}$	$x_{22}^{(5)}$	···	···
⋮	⋮	⋮	⋮	⋱		⋮	⋮	⋱	⋮	⋮	⋱	⋱
i	2	0	1	···		$x_{11}^{(i)}$	$x_{12}^{(i)}$	···	$x_{21}^{(i)}$	$x_{22}^{(i)}$	···	···
⋮	⋮	⋮	⋮	⋱		⋮	⋮	⋱	⋮	⋮	⋱	⋱
n	1	1	0	···		$x_{11}^{(n)}$	$x_{12}^{(n)}$	···	$x_{21}^{(n)}$	$x_{22}^{(n)}$	···	···

確率的選好度 $\varepsilon_k^{(i)}$ から成り，以下のように表わされるものとする．

$$U_k^{(i)} = V_k^{(i)} + \varepsilon_k^{(i)}, \quad i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, \ell$$

各選択肢は m 個の共通した特性項目を持ち，個人 i が選択肢 k の特性 $j(j = 1, 2, \cdots, m)$ に対して下す評価値を $x_{kj}^{(i)}$ とする．このとき，本文 p.64 に示したように個人 i における選択肢 k の確定的選好度 $V_k^{(i)}$ は以下のように表わされるものとする．

$$V_k^{(i)} = \sum_{j=1}^{m} \alpha_j x_{kj}^{(i)}, \quad i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, \ell$$

ただし，α_j はパラメータである．

一方，$\varepsilon_k^{(i)}$ は b を尺度パラメータとする独立で同一の二重指数分布で表されるものとする．

ここで，選好度が最大となる選択肢が選択されるとすると，個人 i が選択肢 k を選択する確率 $p_k^{(i)}$ は，

$$p_k^{(i)} = \Pr\{U_k > U_h, \quad \forall h \neq k\}$$

この確率を計算すると次のようになる．

$$\begin{aligned} p_k^{(i)} &= \int_{-\infty}^{\infty} \left[\prod_{h \neq k} \Pr\left\{ \varepsilon_h^{(i)} \leq V_k^{(i)} - V_h^{(i)} + x \right\} \right] f(x) dx \\ &= \int_{-\infty}^{\infty} \left[\prod_{h \neq k} \exp\left(-\mathrm{e}^{-b(V_k^{(i)} - V_h^{(i)} + x)} \right\} \right] b\, \mathrm{e}^{-b\,x} \exp\left(-\mathrm{e}^{-b\,x} \right) dx \\ &= \int_{-\infty}^{\infty} \exp\left[-\sum_{h \neq k} \mathrm{e}^{-b(V_k - V_h + x)} \right] b\, \mathrm{e}^{-b\,x} \exp\left(-\mathrm{e}^{-b\,x} \right) dx. \end{aligned}$$

ここで，$y = x + V_k$ と置き換え，整理すると以下のようになる．

$$\begin{aligned} p_k^{(i)} &= \exp\left(b\, V_k \right) \int_{-\infty}^{\infty} b \exp\left[-\mathrm{e}^{-b\,y} \sum_{h=1}^{\ell} \mathrm{e}^{b\, V_j^{(i)}} \right] \mathrm{e}^{-b\,y} dy \\ &= \frac{\exp\left(b\, V_k^{(i)} \right)}{\sum_{h=1}^{\ell} \exp\left(b\, V_h^{(i)} \right)} \int_{-\infty}^{\infty} b \sum_{h=1}^{\ell} \mathrm{e}^{b\, V_h^{(i)}} \exp\left[-\mathrm{e}^{-by} \sum_{h=1}^{\ell} \mathrm{e}^{b\, V_h^{(i)}} \right] \mathrm{e}^{-b\,y} dy. \end{aligned}$$

ここで，a,b が定数のときに

$$\int a\, b \exp\left(-a \mathrm{e}^{-b\,y} \right) \mathrm{e}^{-b\,y} dy = \exp\left(-a \mathrm{e}^{-by} \right)$$

であるので，選択確率は以下のように計算できる．

$$\begin{aligned} p_k^{(i)} &= \frac{\exp\left(b\, V_k^{(i)} \right)}{\sum_{h=1}^{\ell} \exp\left(b\, V_h^{(i)} \right)} \left[\exp\left(-\mathrm{e}^{-b\,y} \sum_{h=1}^{\ell} \mathrm{e}^{b\, V_h^{(i)}} \right) \right]_{-\infty}^{\infty} \\ &= \frac{\exp\left(b\, V_k^{(i)} \right)}{\sum_{h=1}^{\ell} \exp\left(b\, V_h^{(i)} \right)}. \end{aligned}$$

したがって，表 C.9 のデータを用いると，多項ロジット・モデルにおける選択確率は次のように表すことができる．

$$P_k^{(i)} = \Pr\left[U_k^{(i)} = \max_h U_h^{(i)} \right]$$

$$= \frac{\exp\left[b\, V_k^{(i)}\right]}{\sum_{h=1}^{\ell} \exp\left[b\, V_h^{(i)}\right]} = \frac{\exp\left[b \sum_{j=1}^{m} \alpha_j x_{kj}^{(i)}\right]}{\sum_{h=1}^{\ell} \exp\left[b \sum_{j=1}^{m} \alpha_j x_{hj}^{(i)}\right]}, \quad \text{(C.36)}$$

$$i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, \ell$$

C. 11. 3 パラメータの推定

多項ロジット・モデルでは最尤推定法を用いてパラメータを推定する．本来，パラメータは個々人で異なるはずであるが，推定すべきパラメータの数が多数になるので，多くの場合，すべての個人においてパラメータは共通であると仮定する．

このときの対数尤度関数は以下のように表わされる．

$$\ln L(b, \alpha_1, \cdots, \alpha_\ell) = \sum_{k=1}^{m} \sum_{i=1}^{n} \delta_i^{(k)} p_i^{(k)}$$

ただし，$\delta_i^{(k)}$ は個人 k が選択肢 i を選択したとき 1，そうでなければ 0 となるダミー変数である．

この関数は単峰性が保証されているが[*1)]，求めるパラメータの次元数が 1 つ退化しているので，このままでは b と α_j を一意に推定することはできない[*2)]．そこで，以下の 3 通りのいずれかの制約条件を設け，制約なし非線形計画法としてパラメータを推定する．

- $\beta_j = b\,\alpha_j$ とおいて，β_j を推定する．
  (計算上は $b = 1$ という制約条件を置くのと同じである．)
- α_j のどれか 1 つを 1 として，残りのパラメータを推定する．
- $\sum_{j=1}^{\ell} \alpha_j = 1$ という制約条件をおいて b と α_j を推定する．

---

*1) 例えば，McFadden(1974) を参照せよ．

*2) b と α_j が与えられたもと各 $U_k^{(i)}$ の大小比較により選択が決定されるが，すべての $U_k^{(i)}$ を定数倍してもこの大小関係は変わらない．このとき，パラメータはそれぞれ b/(定数)，α_j×(定数) と調整される．(C.36) 式をこのように調整しても選択確率は変わらないことからも明らかである．

C. 12　コンジョイント分析と LINMAP

C. 12. 1　コンジョイント分析の考え方

コンジョイント分析 (conjoint analysis) は，計量心理学において発展してきた「コンジョイントと測定法」の理論体系を基礎にして，それをマーケティングにおける消費者選好の測定に応用しようという試み全体を総称したものである．コンジョイント分析では，選択候補となる商品の集合，すなわち想起集合をあらかじめ与えた上で，商品の選好を分析モデルである．消費者の商品に対する選好は補償型モデルと捉えられることが一般的であり，属性への分解型アプローチによって消費者のもつ選好構造を捉えるモデルである．コンジョイント分析の特徴は「プロフィール (profile)」と呼ばれる仮想的商品・サービスを選好の対象として，物理的，機能的属性と選好の関係を測定，分析しようとするところにある．

C. 12. 2　コンジョイント分析のモデル

今，比較しようとしている各プロフィール $(i = 1, 2, \cdots, m)$ に関して，製品特性ベクトル $\mathbf{x} = (x_1, x_2, \cdots, x_\ell)^\top$ が与えられているとする．属性としては価格のような連続的なもの，および付帯機能のような離散的なものが考えられる．前者の場合は数値をそのまま用い，後者の場合はダミー変数に変換する．プロフィール i の属性ベクトルを $\mathbf{x}_i = (x_{i1}, x_{i2}, \cdots, x_{i\ell})$ とする．このとき，回答者 k のプロフィール i の選好度 V_i は多項ロジット・モデルと同様，以下のように線形モデルとして表されるとする．

$$V_i^{(k)} = \alpha_1 x_{i1}^{(k)} + \alpha_2 x_{i2}^{(k)} + \cdots + \alpha_\ell x_{i\ell}^{(k)},$$
$$i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, m \tag{C.37}$$

コンジョイント分析では一般には，プロフィールに対し回答者が選好の全順序をつけたデータを観測値として用いる．各回答者に対してこの順序にした

がって並び替えた第 i 位のプロフィールを $\{i\}$ で表すと，選好については，

$$V_{\{1\}}^{(k)} > V_{\{2\}}^{(k)} > \cdots > V_{\{\ell\}}^{(k)}, \quad k = 1, 2, . \cdots, m$$

となるはずである．コンジョイント分析では，その順序を満たすようにパラメータ $\alpha_i^{(k)}, i = 1, 2, \cdots \ell$ を求める．求められたパラメータの値より，属性の重要性の度合や感応度を測定することができる．

C. 12. 3 パラメータの推定

コンジョイント分析において，プロフィール属性のパラメータ推定についてはいくつかの方法が示されている．

最も古くから用いられてきた方法は，Kruskal 流のコンジョイント測定法 (MONANOVA) であり，回帰分析の考え方からパラメータを推定している．

また，Green and Srinivasan は線形計画法を応用した LINMAP を提案している．

LINMAP は各プロフィールの選好が得られた順序を満たさない量の和を最小にするような問題を線形計画問題としてモデル化したものである．まず，順序プロフィール $\{i\} = 1, 2, \cdots, m$ の選好序列データを前提とする．順序 $\{i\}$ のプロフィールの選好は (C.37) 式で与えられる．回答者 k の第 $\{i\}$ 順序と第 $\{i+1\}$ 順序のプロフィールの選好が逆転する量を $d_{\{i\}}^{(k)}$ とし，その総和を最小にするような問題を考える．次のような問題として定式化される．

$$\begin{aligned}
\min \quad & \sum_{k=1}^{n} \sum_{i=1}^{m-1} d_{\{i\}}^{(k)} \\
\text{s.t.} \quad & V_{\{i\}}^{(k)} - V_{\{i+1\}}^{(k)} + d_{\{i\}}^{(k)} \geq 0, \quad i = 1, 2, \cdots, n-1, \quad k = 1, 2, \cdots, m \\
& V_1^{(k)} + V_n^{(k)} = 1, \quad k = 1, 2, \cdots, n \\
& d_{\{i\}}^{(k)} \geq 0, \quad i = 1, 2, \cdots, m-1, \quad k = 1, 2, \cdots, n
\end{aligned}$$

$V_1^{(k)} + V_n^{(k)} = 1$ という制約条件は最適解を 0 としないためのものである．これを $\alpha_{\{i\}}^{(k)}$ を変数として解くことにより，部分選好を求めることができる．一般にはパラメータは個人ごとではなく，消費者に共通のものとして推定さ

れる．次にランク・ロジット・モデル (Rank Logit Model) について説明する．ランク・ロジット・モデルは，すでに紹介した多項ロジット分析を用い，確率選択問題の枠組みによりコンジョイント分析をおこなうものである．多項ロジット・モデルと同様に，回答者 k のプロフィール i の選好 $U_i^{(k)}$ を以下のように考える．

$$U_i^{(k)} = V_i^{(k)} + \varepsilon_i^{(k)}, \quad i = 1, 2, \cdots, n, \quad k = 1, 2, \cdots, m \qquad \text{(C.38)}$$

ただし多項ロジット・モデルと同様に，$V_i^{(k)}$ は (C.37) 式で与えられる確定的効用であり，$\varepsilon_i^{(k)}$ は尺度パラメータ b をもつ独立で同一の二重指数分布に従うとする．

多項ロジット・モデルでは，複数の製品群の中から選択される一つの製品に着目し，その製品の選択確率に関する尤度を目的関数とした．ランク・ロジット・モデルでは，プロフィールが順位付けされていることから，ある順位のプロフィールはそのプロフィールの順位以下のプロフィールの集合の中から選択される確率を尤度の対象とする．つまり，

- 1 位のプロフィールについては，すべてのプロフィールの中から選択される確率
- 2 位のプロフィールについては，2 位以下のプロフィールの中から選択される確率 item $\cdots$
- i 位のプロフィールについては，i 位以下のプロフィールの中から選択される確率

をそれぞれ考え，これらすべてを同時に満たす確率を尤度とする．

具体的には次のように表すことができる．まず，すべてのプロフィールから第 $\{1\}$ 位のプロフィールが選択される確率は

$$p_{\{1\}}^{(k)} = \Pr\left\{U_{\{1\}} > U_{\{i\}}, \quad i = 2, 3, \cdots, n\right\}$$

となるが，ロジットモデルに従いその確率は以下のように求められる．

$$p_{\{1\}}^{(k)} = \frac{\exp\left(bV_{\{1\}}^{(k)}\right)}{\sum_{j=1}^{n} \exp\left(bV_{\{j\}}^{(k)}\right)}$$

次に第 $\{1\}$ 位のプロフィールを除いた中で第 $\{2\}$ 位のプロフィールが選ばれる確率は,

$$p_{\{2\}}^{(k)} = \Pr\left\{U_{\{2\}} > U_{\{i\}}, \quad i = 3, 4, \cdots, n\right\}$$

であり,

$$p_{\{2\}}^{(k)} = \frac{\exp\left(bV_{\{i\}}^{(k)}\right)}{\sum_{j=2}^{n} \exp\left(bV_{\{j\}}^{(k)}\right)}$$

となる．以下同様に第 $\{i\}$ 番目に第 $\{i\}$ 位のプロフィールが選ばれる確率は次のように第 $\{i\}$ 位以下のすべてのプロフィールよりも第 $\{i\}$ 位のプロフィールの選好が大きい場合である．

$$p_{\{i\}}^{(k)} = \Pr\left\{U_{\{i\}} > U_{\{j\}}, \quad j = i+1, i+2, \cdots, n\right\}$$

したがって，一般に i 位のプロフィールが i 位以下から選択される確率は以下のように与えられる．

$$p_{\{i\}}^{(k)} = \frac{\exp\left(bV_{\{i\}}^{(k)}\right)}{\sum_{j=i}^{n} \exp\left(bV_{\{j\}}^{(k)}\right)}$$

これらより，回答者 k に関して，プロフィールの選好順序 $(\{1\}, \{2\}, \cdots, \{n\})$ に関する尤度は,

$$L^{(k)} = \prod_{i=1}^{m} p_{\{i\}}^{(k)}$$

となる．したがって，全回答者に関する同時確率を考えると結局尤度は,

$$L = \prod_{k=1}^{m} L^{(k)} = \prod_{k=1}^{n} \prod_{i=1}^{m} p_{\{i\}}^{(k)} = \prod_{k=1}^{n} \prod_{i=1}^{m} \frac{\exp\left(bV_{\{i\}}^{(k)}\right)}{\sum_{j=i}^{n} \exp\left(bV_{\{j\}}^{(k)}\right)}$$

となるので，その対数尤度,

$$\ln L = \sum_{k=1}^{m} \sum_{i=1}^{n} \ln p_{\{i\}}^{(k)}$$

を最大化するようなパラメータを求めればよい．多項ロジット・モデルと同様，一般には消費者ごとにパラメータを推定するのではなく，全体を一つも

しくは少数のセグメントに分け，セグメントごとにパラメータを推定する．またパラメータ推定に関する注意は前節を参照されたい．

C. 12. 4　プロフィール属性の水準の組み合わせ

コンジョイント分析では，部分属性を積み上げることで商品が形成されるとしている．ここで，例えば属性数が 7 つありそれぞれについて 2 つの水準ある場合，すべての水準の組み合わせは $2^7 = 128$ 個にもなる．これらのプロフィールを比較し，順序をつけなければならなくなる．しかし，多数のプロフィールの順序付けは被験者にとっても大変負担であるし，また得られた順序の信頼性は低くなってしまうであろう．そこで，提示するプロフィール数はなるべく少ないほうがよいだろうが，ただ闇雲に少なくすると，偏った情報しか得られなくなる可能性もある．そこで，少数で偏りのないプロフィール作成をするために，実験計画法の分野で研究されている**直交配列**が広く用いられている．直交配列は取り上げる属性の数や水準の数によって様々なタイプがある．本書の例の場合は $L_4(2^3)$ という直交表を用いる．L は直交表を表す記号であり，$L_a(b^c)$ は a がプロフィール数，b が水準数，c が列数[*1)]となる．各列に属性を割り当てる．$L_4(2^3)$ の直交表を表 C.10 に示す．

表 C.10　$L_4(2^3)$ 直交表

No. ＼列番	1	2	3
1	1	1	1
2	1	2	1
3	2	1	2
4	2	2	1

表の各要素が属性の水準を示している．どの列も水準 1 と水準 2 が同数（2 つずつ）あることが分かる．また，各列をベクトルと考えると任意の 2 列の内積は 0 になる．これが直交表と言われている理由である．本書で示した例は属性が 3 種類であったが，それ以上の場合（7 属性まで）の場合，$L_8(2^7)$ を用いる．$L_8(2^7)$ は表 C.11 のようになる．これを用いると，例えば 7 属性の場合，水準のすべての組み合わせでは $2^7 = 128$ 個のプロフィールが必要

[*1)] 列数は行数より 1 少ない

なところ，わずか 8 つのプロフィールによって偏りのない実験を行うことができる．

表 C.11 $L_4(2^3)$ 直交表

No. ＼列番	1	2	3	4	5	6	7
1	1	1	1	1	1	1	1
2	1	1	1	2	2	2	2
3	1	2	2	1	1	2	2
4	1	2	2	2	2	1	1
5	2	1	2	1	2	1	2
6	2	1	2	2	1	2	1
7	2	2	1	1	2	2	1
8	2	2	1	2	1	1	2

もしも，属性数が 5 つの場合は，$L_4(2^3)$ の任意の列に 5 つの属性を割り付けてプロフィールを作成する．

これまでは属性同士はそれぞれ関係のない加法モデルで議論したが，属性間の相互関係[*1)]を考慮する場合の直交配列もある．詳しくは専門書を参照されたい．

C. 13 線形計画問題の双対問題

制約つき最適化問題においては双対問題という，別角度から見た問題により定式化することができることが知られている．同次座標によって表現された空間において双対性が成立することから支持される．ここでいう双対性とは，幾何学的な命題において，線（平面）と点を交換しても成立する性質をいう．たとえば，2 つの直線によって 1 つの点が作られる，という命題を考えよう．するとこの双対は 2 つの点によって 1 つの直線が作られる，となる．

数理計画問題における双対問題（dual problem）とは，上記の双対性を数理計画問題の各数式に当てはめたものである．一般の数理計画問題の双対問題はラグランジュ緩和 (Lagrange relaxation) により導かれるが，その厳密な議論は本書の範囲を逸脱するので，もっとも広く知られている線形計画問

*1) これを交互作用という．

題の場合における双対問題を証明なしに示すことにする．

今下記のような線形計画問題 (これを主問題と呼ぶ) が与えられているとする．

$$\max \quad \boldsymbol{c}^\top \boldsymbol{x} \quad \text{s.t.} \quad A\boldsymbol{x} \leq \boldsymbol{b} \quad \mathbf{x} \geq \mathbf{0}$$

この主問題に対する双対問題は，

$$\max \quad \boldsymbol{b}^\top \boldsymbol{y} \quad \text{s.t.} \quad A^\top \boldsymbol{y} \geq \boldsymbol{c} \quad \boldsymbol{y} \geq \mathbf{0}$$

となる．この式を比べてみると，制約条件のそれぞれの式が双対変数 $\boldsymbol{y}$ のそれぞれの要素に対応していることが見てとれよう．このことが上記に述べた直線と点の関係に対応する．詳しくは説明しないが，主問題で最大化を目指す場合その双対問題は最小化問題となる．これは，ラグランジュ緩和した問題の中でもっとも良いラグランジュ乗数を求める問題を解こうとしたときの最適な問題となっている．また双対問題の双対問題は主問題になる．主問題と双対問題の関係を見ると，主問題の制約条件式それぞれに対して双対問題の変数（これを双対変数と呼ぶ）を対応させている．つまり，ある面（もしくはその面からなる制約条件）に対して一つの双対変数を対応させる．さらに，双対変数からなる（双対問題の）制約条件式について，双対問題を考えると，元の問題に戻る[*1)]．

線形計画問題の主問題が等号制約を持つ場合，$\boldsymbol{a}_i^\top \boldsymbol{x} \leq b_i$ と $-\boldsymbol{a}_i^\top \boldsymbol{x} \leq -b_i$ という 2 つの制約条件とすることで，不等式の場合と同様の枠組みで扱うことができる．その際，別々の双対変数がそれぞれの式に対応するので，それらを組み合わせたると，等号制約に対する双対変数には非負制約がないことになる．また，線形計画問題の双対問題についてはいくつもの重要な定理が示されており，システム解析の上で非常に大きな役割をもつ．証明なしに示しておく．詳しくは今野 (1987) を参照されたい．

定理 C.2 (弱双対定理) $\boldsymbol{x}, \boldsymbol{y}$をそれぞれ上記主問題と双対問題の実行可能解

[*1)] この関係を，立方体について考えてみる．立方体の 6 つの面について双対変数一つを対応させる．つまり，立方体の各面の中心の点を双対変数と考えると，これらの双対変数で囲まれる領域は，8 つの三角形でできる 8 面体となる．同様に，この 8 面体の各三角形それぞれに双対変数を考えこれらにより囲まれる領域は立方体となる．

とすると，

$$\boldsymbol{c}^\top \boldsymbol{x} \leq \boldsymbol{b}^\top \boldsymbol{y}$$

が成立する． □

定理 C.3 (双対定理) 主問題が実行可能領域が存在する場合その双対問題も実行可能領域が存在し，それらの最適目的関数値は一致する． □

もし主問題の実行可能領域が存在しない場合は双対問題は無限解を持ち，主問題が無限解をもつ場合双対問題の実行可能領域は存在しない．

主問題および双対問題の最適解には以下の関係がある．

定理 C.4 (相補スラック定理) 主問題および双対問題の実行可能解 $\boldsymbol{x}, \boldsymbol{y}$ がそれぞれの最適解であるための必要十分条件は

$$\boldsymbol{x}^\top (A^\top \boldsymbol{y} - \boldsymbol{c}) = \boldsymbol{0} \tag{C.39}$$

$$\boldsymbol{y}^\top (\boldsymbol{b} - A\boldsymbol{x}) = \boldsymbol{0} \tag{C.40}$$

が同時に成り立つことである． □

したがって，主問題の最適基底の単体乗数が双対問題の最適解となる．

C. 14 データ包絡分析

データ包絡分析 (Data Envelopment Analysis; DEA) は多入力多出力系の相対的効率性判定のための手法であり，投入 (入力) で産出 (出力) を割るという従来の効率性評価の式を多入力多出力の場合に拡張した方法である．

DEA のためのデータは一般に表 C.12 のような多入力，多出力のデータとなる．ただし各要素は正と仮定する．表下の u_r, v_i および，表右の λ_j は以降で説明する評価モデルのパラメータである．

多入力多出力系システムを 1 つの値で評価するためには，入力および出力に対する重要度を検討する必要がある．評価者が絶対的な価値判断基準を

表 C.12　DEA のデータ形式

評価対象	入力 1	入力 2	$\cdots$	入力 m	出力 1	出力 2	$\cdots$	出力 s	非負結合係数
DMU1	x_{11}	x_{12}	$\cdots$	x_{1m}	y_{11}	y_{12}	$\cdots$	y_{1s}	λ_1
DMU2	x_{211}	x_{22}	$\cdots$	x_{2m}	y_{21}	y_{22}	$\cdots$	y_{2s}	λ_2
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$	$\vdots$
DMUn	x_{n1}	x_{n2}	$\cdots$	x_{nm}	y_{n1}	y_{n2}	$\cdots$	y_{ns}	λ_n
ウェイト	v_1	v_2	$\cdots$	v_m	u_1	u_2	$\cdots$	u_s	

持っている場合や，客観的な判断基準がある場合には，それを用いることもできるが，DEA ではデータ自身に語らせウェイトは DMU の活動に応じて求める．

また，DEA の大きな特徴として，回帰分析などの平均的な振る舞いとの比較ではなく，あくまでも活動の中の「優れもの」に着目し，それとの比較を通して自己の評価・改善を図るという方法であるということが挙げられる．

C. 14. 1　生産可能集合

DEA における**生産可能集合**は，DMU の活動可能な入出力値の水準の集合を表したものである．DEA では観測された DMU の活動をもとに生産可能集合を定義しているが，その際に以下のような仮定を設けている．

- 観測された活動 $(\boldsymbol{x}_j, \boldsymbol{y}_j)$ は生産可能な活動であり生産可能集合に属する．
- 活動の各要素は正とする．
- 生産可能な活動を定数倍した活動は生産可能集合に属する．
- 生産可能な活動の非負結合は生産可能集合に属する．
- 生産可能な活動 $(\boldsymbol{x}, \boldsymbol{y})$ に対して，入力余剰もしくは出力不足の活動は生産可能集合に属する．

以上をまとめると，生産可能集合 P は，

$$P = \{(\boldsymbol{x}, \boldsymbol{y}) | \boldsymbol{x} \geq X\boldsymbol{\lambda}, \boldsymbol{y} \leq Y\boldsymbol{\lambda}, \boldsymbol{\lambda} \geq \boldsymbol{0}\}$$

となる．

C.14.2 効率的フロンティア

DEA は生産可能集合内での相対的な効率判定をおこなう．したがって，効率的と判定される活動を定義する必要がある．効率の定義から，入力は小さく出力は大きいほうが望ましい．入力と出力がそれぞれ 1 変量ずつの場合を考えよう．効率は “出力／入力” として定義できるので，各活動の効率は原点をから活動の座標を通る直線の傾きとなる．逆に，傾きが等しければ，同じ効率の活動となる．そのうち，もっとも傾きの大きい活動が観測された活動の中でもっとも効率的な活動をしているといえよう．効率的な活動を通るような直線が生産可能集合の境界線となり，それより下の領域が生還可能集合となる．境界線上の活動は，もし生産可能集合内で入力を小さくするためには，出力も小さくしなければならない．DEA では入力は小さくする方向，出力は大きくする方向にみたときに生産可能集合の境界にある活動の集合を**効率的フロンティア**とよび，すべて効率的な活動とみなす．DEA では効率的フロンティアを基準に効率性を評価する．また，上記の生産可能集合を仮定する場合，効率的フロンティアは原点を通り放射状に構成される．効率的フロンティアにちゃくもくすると，出力を k 倍するためには入力も k 倍しなければならない．このような場合を規模の収穫が一定であるという．規模の収穫に対する設定を変えるためには，生産可能集合の非負結合係数 $\boldsymbol{\lambda}$ にさらに制約をつけなければならない．

C.14.3 入力指向モデルと出力指向モデル

本文では [FP] の分子を固定し分母を最大化する問題として変換したが，以下の [LPDI] のように分母を固定し分子を最大化する線形計画問題として変換することもできる．

$$\begin{aligned}[\text{LPDI}]\quad &\max \quad z = \boldsymbol{u}^\top \boldsymbol{y}_a, \\ &\text{s.t.} \quad \boldsymbol{v}^\top \boldsymbol{x}_a = 1, \quad \boldsymbol{u}^\top Y^\top - \boldsymbol{v}^\top X^\top \leq \boldsymbol{0}, \quad \boldsymbol{v}, \boldsymbol{u} \geq \boldsymbol{0}.\end{aligned}$$

[LPDI] の双対問題は次の [LPI] となる．

$$\text{[LPI]} \quad \min \quad \omega,$$
$$\text{s.t.} \quad X\boldsymbol{\lambda} - \omega \boldsymbol{x}_a \leq \mathbf{0}, \quad Y\boldsymbol{\lambda} \geq \boldsymbol{y}_a, \quad \boldsymbol{\lambda} \geq \mathbf{0}.$$

すでに述べたように，[FP] からの変形は解くための技術的な変形であり，[LPDO], [LPDI] は線形計画問題の双対の関係にあるので，どちらで解いても同じ効率値を得る．したがって，[FP], [LPDO], [LPDI] の最適目的関数値は以下の関係を持つ．

$$\theta^* = \frac{1}{h^*} = z^*$$

次に [LPO] と [LPI] を比較すると，生産可能集合内において当該の DMU の活動の入力もしくは出力を固定した上で，他方を拡大もしくは縮小しようという問題として定式化されている．[LPO] の場合は，入力を現状の観測値として固定した上で出力を生産可能集合内でどこまで拡大できるかを解く問題となっている．また，[LPI] の場合は，出力を現状の観測値として固定した上で入力を生産可能集合内でどこまで縮小できるかを解く問題である．このように，DEA では線形計画問題に変換するときに，暗に入力もしくは出力のどちらかに着目した定式化が行われていることになる．[LPO] を出力指向モデルといい，[LPI] を入力指向モデルという．効率的な DMU については入力方向に削減することも出力方向に拡大することもできないので，最適解は 1 となる．これらのモデルは考えている問題が入力を操作すべき問題なのか出力を操作すべき問題なのかによって使い分けるべきである．

[LPDO], [LPDI] はそれぞれ [LPO], [LPI] の双対 (dual) 問題から命名されている．

C.14.4 ウェイトと非負結合係数

1 入力 2 出力の場合について出力指向モデルを例に，[LPO] における非負結合係数 $\boldsymbol{\lambda}$ と [LPDI] におけるウェイト $\boldsymbol{v}, \boldsymbol{u}$ の関係および効率値の考え方を図示する (図 C.8).

図 C.8 は各出力を入力で割ったものをプロットしたものである．生産可能集合は原点から点 A を垂直軸におろした点，A, B, C, D および D を水平軸

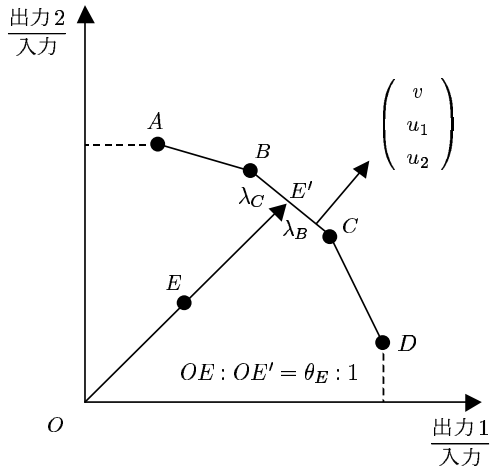

図 C.8 DEA における解の関係

におろした点を順位結び原点を結んだ領域である．5 つの DMU のうち DMU A, B, C, D の 4 つは効率的となる．したがって，点 A, B, C, D を順に結んだ区分線が効率的フロンティアとなる．なぜなら，ある傾きの直線によりそれぞれの DMU がもっとも高い (原点から遠い) 位置にくることができるからである．どんな傾きでももっとも高い位置になることのない DMU E に着目してみよう．すべての DMU の効率値が 1 以下という制約を満たすウェイト$\boldsymbol{v}$, $\boldsymbol{u}$の組み合わせは，効率的な DMU に接するときの支持超平面の法線ベクトルである[*1)]．点 E の効率性は，原点から点 E を通り支持超平面までの距離に対する原点から点 E の距離の比として与えられる．DEA では，もっとも高い効率値を目指すので，結局点 B と点 C を通る直線が求めるべき支持超平面となる．したがって，[LPDO] の最適なウェイトは点 B, C を通る支持超平面の法線ベクトルとして与えられる．そのとき，効率値は OE/OE' として与えられる．逆に，E の活動は生産可能集合内で入力を保ったまま出力を OE'/OE 倍することができる．この値が [LPO] の最適目的関数値として得られる．そして活動 E' が活動 E の改善案として与えられる．

[*1)] 図の点線は一例であるがもちろん連続的に変化させることができる

この図より，DMU B と C が DMU E の参照集合になっていることがわかるが，さらに初等的な幾何学から [LPO] の非負結合係数は，

$$BE' : E'C = \lambda_C : \lambda_B$$

となる．また，$\lambda_A, \lambda_D, \lambda_E$ の最適目的関数値は 0 となる．

相補スラック定理からもわかるように，非負結合係数が正の値をとるときは，そのウェイトに関してその非負結合係数に関する DMU の効率値は 1 となる．逆に，あるウェイトによってある DMU の効率値が 1 とならない場合は双対問題の当該の非負結合係数の値は 0 となる．

DEA ではウェイトの決定に柔軟性を持たせることで各 DMU の活動状態に見合った評価をおこなうことができる．したがって，点 E の活動が変化すると参照集合が変わり，その結果ウェイトも変わる可能性がある．

これまで述べてきたように，DEA では双対関係にある 2 つの問題から得られる解に明確な意味があり，両者の解を得ることは非常に重要な情報をなりうる．双対問題の解は主問題を解けば求めることができるのであるが，一般には単体法における最適基底行列が必要となる．残念ながらソルバーで基底行列を保存することはできないので，Excel を使って DEA の解のすべての情報を得るためには，2 つの問題を解かねばならない．もちろん効率値のみが必要な場合は，一方の問題だけを解けばよい[*1)]．

ここで述べた DEA モデルはもっとも基本的なモデルであり，最初のモデルが提唱されて以来 DEA に関する多くのモデルが提案されている．たとえば，規模の収穫を考慮したモデルや，ウェイトの取りうる範囲を限定した領域限定法，複数期間の効率性の推移をみようとする WINDOW 分析などがある．これらのモデルの詳細を述べるのは本書の範囲を超えるので，興味のある方は専門書 (刀根・上田編, 2000) を参照いただきたい．

また DEA を解くためのソフトウェアもいくつか公開されており，それらを利用することにより手軽に DEA による分析をおこなうこともできる．こういった DEA のソフトウェアにはさまざまな応用モデルも含まれており便利である．

[*1)] 計算量の観点からみると [LPI] もしくは [LPO] を解く方がよい．